The Telecommunications and Data Acquisition Progress Report 42-111

July–September 1992

E. C. Posner
Editor

November 15, 1992

NASA
National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The Telecommunications and Data Acquisition Progress Report 42-111

July–September 1992

E. C. Posner
Editor

November 15, 1992

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The research described in this publication was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

PRIMARY

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 and radar astronomy, it reports on activities of the Deep Space Network (DSN) in planning, supporting research and technology, implementation, and operations. Also included are standards activity at JPL for space 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 Communications (OSC). The TDA Office also peforms work funded by two other NASA program offices through and with the cooperation of OSC. These are the Orbital Debris Radar Program (with the Office of Space Systems Development) and 21st Century Communication Studies (with the Office of Exploration).

For the High-Resolution Microwave Survey (HRMS), *The TDA Progress Report* reports on implementation and operations for searching the microwave spectrum. In solar system radar, it reports on the uses of the Goldstone Solar System Radar for scientific exploration of the planets, their rings and satellites, asteroids, and comets. In radio astronomy, the areas of support include spectroscopy, very long baseline interferometry, and astrometry. These three programs are performed for NASA's Office of Space Science and Applications (OSSA), with OSC funding DSN operational support.

Finally, tasks funded under the JPL Director's Discretionary Fund and the Caltech President's Fund that 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:

OSC 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 Existing Projects
  - Capabilities for New Projects
  - New Initiatives
  - Network Upgrade and Sustaining
- DSN Operations
  - Network Operations and Operations Support
  - Mission Interface and Support
  - TDA Program Management and Analysis
- Ground Communications Implementation and Operations
- Data and Information Systems
- Flight–Ground Advanced Engineering
- Long-Range Program Planning

OSC Cooperative Tasks:
- Orbital Debris Radar Program
- 21st Century Communication Studies

Contents

OSC TASKS
DSN Advanced Systems
TRACKING AND GROUND-BASED NAVIGATION

COMMUNICATIONS, SPACECRAFT–GROUND

STATION CONTROL AND SYSTEM TECHNOLOGY

DSN Systems Implementation
CAPABILITIES FOR EXISTING PROJECTS

CAPABILITIES FOR NEW PROJECTS

NETWORK UPGRADE AND SUSTAINING

DSN Operations

NETWORK OPERATIONS AND OPERATIONS SUPPORT

GROUND COMMUNICATION IMPLEMENTATION AND OPERATIONS

FLIGHT–GROUND ADVANCED ENGINEERING

TDA PROGRAM MANAGEMENT AND ANALYSIS

OSSA Tasks

HIGH RESOLUTION MICROWAVE SURVEY

Subnanosecond GPS-Based Clock Synchronization and Precision Deep-Space Tracking

C. E. Dunn, S. M. Lichten, D. C. Jefferson, and J. S. Border
Tracking Systems and Applications Section

Interferometric spacecraft tracking is accomplished by the DSN by comparing the arrival time of electromagnetic spacecraft signals at ground antennas separated by baselines on the order of 8000 km. Clock synchronization errors within and between DSN stations directly impact the attainable tracking accuracy, with a 0.3-nsec error in clock synchronization resulting in an 11-nrad angular position error. This level of synchronization is currently achieved by observing a quasar which is angularly close to the spacecraft just after the spacecraft observations. By determining the differential arrival times of the random quasar signal at the stations, clock offsets and propagation delays within the atmosphere and within the DSN stations are calibrated.

Recent developments in time transfer techniques may allow medium accuracy (50–100 nrad) spacecraft tracking without near-simultaneous quasar-based calibrations. Solutions are presented for a worldwide network of Global Positioning System (GPS) receivers in which the formal errors for DSN clock offset parameters are less than 0.5 nsec. Comparisons of clock rate offsets derived from GPS measurements and from very long baseline interferometry, as well as the examination of clock closure, suggest that these formal errors are a realistic measure of GPS-based clock offset precision and accuracy.

Incorporating GPS-based clock synchronization measurements into a spacecraft differential ranging system would allow tracking without near-simultaneous quasar observations. The impact on individual spacecraft navigation-error sources due to elimination of quasar-based calibrations is presented. System implementation, including calibration of station electronic delays, is discussed.

I. Introduction

The DSN supports spacecraft navigation for an international community of users. In order to complete most missions successfully, the location of the spacecraft must be determined with very high accuracy. This is done by comparing the arrival time of a signal broadcast by the spacecraft as it is received at two widely separated DSN stations. The delay observable thus formed provides some of the data from which the spacecraft's orbit is determined.

Most of the observed time delay between two stations is due to the geometry of the spacecraft and receiving stations; however, delays due to solar plasma, the atmosphere of the Earth, ground station instrumentation, and general relativity also play a role. This measurement technique is shown in Fig. 1 and is called very long baseline interferometry (VLBI). In order to determine the angle, θ, giving the direction to the spacecraft to an accuracy of $\delta\theta$, the error in determining the delay, δt, can be no more than

$$\delta t \leq \frac{D}{c}\delta\theta \tag{1}$$

where D is the separation of the stations and c is the speed of light. Clearly, it is advantageous to use the longest baselines possible. Currently there are three DSN complexes: at Goldstone, California; Madrid, Spain; and Canberra, Australia. Thus, a typical DSN baseline is 8000 km. A typical medium-accuracy tracking requirement is 50 nrad. Using Eq. (1), one arrives at a maximum error of 1.33 nsec.

In order to keep the total error in delay within this limit, the effective VLBI clock synchronization must be better than 1 nsec. Subnanosecond time transfer is a difficult problem, yet 50-nrad accuracy of spacecraft angular position in the radio reference frame is routinely obtained, and 5-nrad accuracy is achieved in special cases. This high level of performance is accomplished by using the signals from an extragalactic radio source (quasar) to calibrate spacecraft observations. The radio signal from the quasar is essentially wideband random noise, so when the recorded signals from the two stations are cross-correlated, significant correlation amplitude arises only when the differential quasar-to-station delay is precisely matched by the offset in tape playback times. This has the effect of measuring the station clock offset as well as differencing out many of the errors which are common to the quasar and spacecraft observations. The spacecraft signal spectrum contains tones which are used to measure a one-way range from the spacecraft to the Earth station. An observable formed by subtracting the quasar delay from the spacecraft one-way range difference between stations determines one component of the geocentric angle between the spacecraft and quasar. Measurements must be made on two baselines to determine both components of angular position.

An error budget for spacecraft–quasar differential VLBI delay measurements is given in Table 1. This error estimate is based on expected DSN receiver performance and calibration systems' capabilities in the late 1990s [1]. The root-sum-square error is 0.22 nsec. By contrast, systems operating today provide an accuracy of about 0.67 nsec.[1]

Although this method provides accuracy sufficient to carry out deep-space missions, it has some drawbacks. In order to be useful, the quasar observations must be made as close in time as possible to the spacecraft observations. In order to view the quasar, the antenna must be physically pointed at the quasar and is thus unable to receive signals from the spacecraft. As a result, the phase data from the spacecraft are not continuous, which results in a weaker orbit solution and a gap in spacecraft telemetry while the quasar is being observed.

Quasar observations also complicate the hardware otherwise required to track a spacecraft. The quasar is a wideband radio source, and so a wide bandwidth is required to record and process the quasar data. If the quasar could be dispensed with, only the phases of the received spacecraft signals at each time point would need to be recorded. In the quasarless system proposed here, 50-nrad observables could be available in near-real time. The number of bits resulting from a ~10-min observation would be reduced from 10^9 to 10^5.

This article will examine medium-accuracy deep space tracking as an application of subnanosecond clock synchronization. It will begin by defining the requirements that 50-nrad tracking accuracy places on the clock synchronization system. It will then discuss recent results obtained using Global Positioning System (GPS) satellites for clock synchronization, which indicate this level of accuracy may be possible on an operational basis. Finally, it will discuss the hurdles remaining before this technology can be implemented.

II. Clock Synchronization Requirements

Time transfer–aided spacecraft tracking will never be able to achieve the accuracy possible using quasar-based differential VLBI. This is because in differencing the quasar signals many of the media errors affecting the signal are differenced out as well. Without quasar differencing, the errors due to station location, Earth orientation, troposphere, and ionosphere would increase by a factor of 2 to 4. However, the GPS solution from which the VLBI clock offset will be derived can also be used to provide calibrations for Earth orientation and troposphere and ionosphere

[1] J. S. Border, *Analysis of DDOR and DDOD Measurement Errors for Mars Observer Using the DSN Narrow Channel Bandwidth VLBI System*, JPL Interoffice Memorandum 335.1-90-026 (internal document), Jet Propulsion Laboratory, Pasadena, California, 1990.

delays. A conservative assumption is that these errors increase by a factor of 3 compared to quasar-calibrated VLBI. Of the remaining errors, those pertaining to quasar signal-to-noise ratio (SNR) and quasar location are eliminated with the elimination of the quasar. The errors due to spacecraft SNR, phase ripple, and solar plasma would remain unchanged. If GPS is used to estimate clock offsets every 6 min during the spacecraft pass, the clock instability error remains roughly the same. The rss of all errors excluding the clock is 0.534 nsec. This leaves a 1.2-nsec maximum-allowable clock synchronization error, which includes instrumental errors incurred in tying GPS time to VLBI time. A reasonable and conservative system allocation is 0.5 nsec for clock synchronization errors, which would allow for a margin of safety of about a factor of 2.

III. Achieving Subnanosecond Clock Synchronization

Since the inception of the GPS in 1978, techniques for using it for high-accuracy clock synchronization have matured rapidly. The number of GPS satellites has recently reached 16, and, in addition, capable p-code receivers have proliferated. For this reason, the possibility of achieving nsec and better clock synchronization using the GPS system has been studied [2–6].

In order to investigate the feasibility of meeting the requirements posed in Section II, the authors investigated clock offset solutions for a global network of Rogue [7] GPS receivers which was assembled for the GPS International Earth Rotation Service (IERS) and Geodynamics (GIG'91) campaign in January and February of 1991 [8]. Clock offsets were calculated for the three DSN sites and compared with clock data derived from VLBI quasar measurements. A closure test was made to verify the internal consistency of the method.

A. Comparison With VLBI

The GIG'91 data from the 21 global Rogue sites were processed with the Jet Propulsion Laboratory's GPS Inferred Positioning System (GIPSY) software. A general description of the square-root Kalman filtering algorithms used for determination of timing and geodynamical parameters simultaneously with GPS orbits is described in detail in [9,10] and references therein. Estimated parameters included GPS positions and velocities, three solar pressure coefficients per satellite, GPS carrier phase biases, nonfiducial station coordinates, variations in Earth rotation (UT1–UTC), random-walk zenith troposphere delays for each site, and white-noise transmitter and receiver clocks. The only significant constraint imposed on the estimated parameters was the random-walk constraint for the tropospheric delay, 1.2 cm/$\sqrt{\text{hr}}$ (the random-walk model adds process noise to the system such that, in the absence of data, the uncertainty for the parameter increases as the square root of time). All other estimated parameters, including the clocks, were essentially unconstrained.

The white noise [10,11] clock model for the station and satellite clocks corresponds to the estimation of a new and independent clock offset for each receiver and transmitter (one ground clock was held fixed as a reference for all the other clocks in the system) at each measurement time (every 6 min in this case). This approach is very conservative, since most of the GPS clocks and many of the receiver clocks were running off atomic standards (high-quality hydrogen masers for the three DSN sites), and it would be quite reasonable to apply constraints based on known stable behavior of such clocks. However, the authors wished to test the capability of the GPS to independently and completely characterize all the clocks in the system without a priori knowledge and therefore used the white noise model. Coordinates for two fiducial sites, Goldstone, California, and Kootwijk, Netherlands, were held fixed (not estimated) to their SV5 values. SV5 is a reference frame defined primarily by VLBI measurements of baselines and satellite laser ranging determination of the geocenter [12]. Three geocenter parameters were also estimated, representing a translation estimated from the GPS data for the Earth's center of mass relative to the nominal SV5 origin.

The GPS data were initially filtered in 24-hr increments, with new solutions for the orbits determined for each day. Since the computed formal errors for the estimated clock offsets appeared to be well below 1 nsec (typically several tenths of a nsec), in some cases 12-hr solution arcs were used in order to shorten the processing time.

The nominal time series for both polar motion and UT1–UTC was from the *International Earth Rotation Service (IERS) Bulletins* B37 and B38 [13,14], which contain a smoothed time series from VLBI measurements separated by 5 days. The GPS data were used to estimate variations in UT1–UTC twice per day relative to this nominal time series. These Earth rotation estimates had only an insignificant effect on the clock estimates.

In order to tie GPS data to the DSN station clocks, the Rogue GPS receiver was fed a 5-MHz reference signal generated by the station hydrogen-maser frequency standard. The time tags of the data are derived from this reference, subject to delays within the interconnection and the receiver. The highly digital nature of the Rogue receiver eliminates most of the delay variations which arise

from variability of analog components [15]. The remaining receiver instrumental delays have been shown to remain constant on a day-to-day basis to within 0.7 nsec.[2] This insures that the receiver clock and the station clock run at the same rate within ~0.7 nsec/day.

Currently there exists no system to measure the offset between the station clock and the receiver clock. As a result, a constant offset is assumed and clock rates derived from GPS are compared with those derived from VLBI. This problem will be discussed further in Section IV.

For this experiment, GPS selective availability (SA) was not turned on. SA involves a dithering of the GPS frequencies, which has the effect of making the GPS clocks look very noisy to users who are not authorized with the classified keys. When SA is on, the GPS broadcast information about GPS clocks and ephemerides is also altered to degrade the accuracy of point positioning by users without the keys. However, since the GPS estimates in this experiment involve only simultaneous data involving many receivers and satellites with dynamically estimated orbits and spacecraft clocks, this type of GPS processing is insensitive to both components (dithering and broadcast ephemerides) of SA. The estimated GPS clocks and orbits remove the SA effects, while the ground (DSN) clock estimates are unaffected. This statement is supported by the authors' limited experience from the past two years. When SA was active, the accuracy and precision of ground-station position estimates were insensitive to SA. A second type of GPS user restriction can result from antispoofing (AS), which encrypts the p-code with a Y-code that can be used directly only by certain authorized users. By effectively restricting most users to single (L1) frequency operation, antispoofing can degrade the accuracy of some GPS measurements. If the p-code were so encrypted, these results would probably be somewhat noisier. However, the Rogue GPS receiver is still able to provide an ionospheric calibration by cross-correlating the P1 and P2 signals. Hence, even in the event of AS, the authors expect to maintain subnanosecond clock synchronization.

Table 2 presents a comparison of the GPS clock rates with VLBI clock rates on days when VLBI solutions were available (GPS solutions were available nearly continuously during the 3-week experiment). The VLBI column presents the clock frequency offset between the specified DSN sites as determined by the DSN Time and Earth Motion Precision Observations (TEMPO) [16] service. These measurements are made by observing a set of quasars over a 3-hr interval centered on an epoch. The GPS column was produced by decimating clock estimates originally computed at a 6-min interval to a 60-min interval and then fitting 6 or 7 of these points to a line. The hourly points are selected to be centered as closely as possible on the VLBI epoch. An example of one of these fits is shown in Fig. 2. The typical rms scatter in these GPS clock fits was 0.1–0.3 nsec.

VLBI is probably the most accurate established, independent technique for measuring clock differences between tracking sites separated by intercontinental distances. Yet the formal errors for the GPS fits are similar and, in most cases, lower than the VLBI formal errors. The reduced χ^2 statistic for the GPS fits, which basically measures the ratio of the postfit scatter to the formal clock estimate errors, was generally 0.5–1.0. The agreement between the GPS and VLBI clock rate estimates shown in Table 2 is at the ~nsec/day level and can be explained by the VLBI formal errors in estimating the clock rates. Note that some aspects of the conservative GPS fitting procedure (decimation of data by a factor of 30 and fitting a line to the clock-offset time series instead of solving explicitly for the rate parameter with the original data) tend to make the GPS formal errors (and presumably the actual errors) larger. A more aggressive analysis strategy could easily be devised to further reduce the GPS clock-rate estimation errors. The results suggest, in any case, that the GPS observations can be straightforwardly used to faithfully track clock variations at time and frequency standards separated by thousands of km. The comparison made by the authors with the independent VLBI technique appears to be limited by the uncertainties in the VLBI data, not by uncertainties in the GPS data.

B. Clock Closure

To investigate the internal consistency of the estimates, a typical day (February 2, 1991) was chosen, and the clock offset between each pair of DSN stations was estimated, with the data from the remaining station excluded for a 12-hr period. The sum of these numbers was then formed, which is referred to as the clock closure. If the receiver clocks have the same estimate in each of the runs for which they are included, the clock closure will be zero. The clock closure is shown as a function of time in Fig. 3. The formal errors presented are calculated assuming each of the six estimates has an independent random error, summed in quadrature, and are thus probably on the conservative side.

Removing the data from a single station should have a very small effect on the result. Nonzero clock closure

[2] L. E. Young, *Rogue Clock Synch Data*, JPL Interoffice Memorandum (internal document), Jet Propulsion Laboratory, Pasadena, California, January 28, 1991.

is an indication of systematic errors in the calculation of the offset of the remaining two clocks. By examining the clock closure, the authors hope to identify error sources as well as verify that nothing is seriously wrong with their estimates.

The curve in Fig. 3 has several interesting features. It is believed that the large, slow variation in the clock-closure result is due to errors in the estimated GPS satellite orbits. To test this, the orbital parameters were initialized with more precise values by using data collected in the previous 24-hr period, including all stations. The resulting clock closure is shown in Fig. 4. The large-scale variation is absent, although it has been replaced with a 0.2-nsec bias. This can be explained because the orbits are constrained by the previous 24 hr of data and therefore are less sensitive to data noise. On the other hand, systematic orbit errors due to dynamic models are more important for longer data arcs and may be causing the 0.2-nsec bias.

Another of these features is the small (0.1-nsec) jump occurring just before 2:00 a.m., as shown in Fig. 3. This is thought to be due to an abrupt change in the satellite geometry, since relatively few satellites were visible at that time and most were either rising or setting. The jump near noon may be due to similar satellite geometry effects, as well as a short data outage. The GPS constellation in early 1991 consisted of only 15 operational satellites. There were short periods of time when GPS visibility was poor from a given ground site, thus leading to high sensitivity to scheduling of observations and data gaps such as described above. It is expected that, in the future, complete ground coverage with the fully operational 21-satellite constellation will make such episodes infrequent and of less consequence.

In another test for data consistency, the reference clock was changed. For January 23, 1991, the clock solutions with Goldstone, California, as the reference clock were compared with the solutions with Kokee, Hawaii, as the reference clock. Because both clocks were hydrogen masers, no appreciable difference in solutions should occur. The clock rate solutions differed by 0.03 nsec/day and 0.01 nsec/day, respectively, for the two cases, a statistically insignificant difference.

C. Data Availability

Another important issue to be addressed is whether high-quality GPS clock estimates can be produced reliably enough to be used operationally. During the IERS GIG'91 campaign, it was possible to form a clock offset with formal errors of less than 1 nsec for 74 percent of the hourly estimates on the Goldstone–Madrid baseline and 79 percent of the time on the Goldstone–Canberra baseline. This includes times in which there was insufficient common view to obtain an accurate clock offset, as well as periods in which one of the two receivers was not tracking satellites.

With the full 21-satellite GPS constellation and with data from approximately six to nine globally distributed stations in addition to the DSN stations, it should be possible to continuously provide subnanosecond clock-offset estimates for the DSN complexes. It is currently possible to service such a network and provide one-day turnaround of clock estimates.[3] For example, with recently installed Hewlett Packard computers in the supermini class, full filtering and smoothing for 24 hr of GPS data from a global tracking network require approximately 1 hr of computing time, including computation of postfit data residuals, some data editing, and automated outlier removal and correction. With the development of forward-running Kalman filters and a real-time data-retrieval system, it conceivably would be possible to provide 1-nsec, hydrogen-maser clock offsets in as little as 5 min by incrementally adding small amounts of data to a continuously running Kalman filter. Somewhat better offsets could be provided within a few hours. More study is needed to determine the minimal configuration necessary to provide near-real-time clock estimates.

Note that because the DSN stations are equipped with hydrogen masers, errors in estimated clock offsets grow gradually, so that short GPS data outages are more likely to result in degraded system performance rather than catastrophic system failure.

IV. Implementation

In order to implement an operational GPS-aided VLBI system, the receiver clock synchronization discussed above must be transferred to the VLBI clock. This can be done by using a time interval counter (TIC) to measure the difference in the 1-pulse-per-second (pps) signals generated by Rogue and VLBI time. It is not difficult to obtain time interval counters accurate to 100 psec, which would not severely impact the level of accuracy. This TIC could be

[3] G. Blewitt, personal communication, Technical Group Supervisor, Tracking Systems and Applications Section, Jet Propulsion Laboratory, Pasadena, California, November 1, 1991.

machine-readable to allow real-time calibration at a rate similar to the frequency of clock offset estimates available from the GPS solution.

The calculated clock offset between receiver time and GPS time includes the delays and phase shifts introduced by the analog electronics between the GPS antenna and the Rogue receiver. The only remaining uncalibrated delays are between internal receiver time and the resultant 1-pps signal, and the corresponding delay in the VLBI system. JPL experiments have shown that these combined delays in the receiver remain constant over a 4-day period to within 0.3 to 0.7 nsec.[4] These delays can be calibrated by making occasional quasar VLBI observations and comparing the clock synchronizations determined by this method with GPS clock synchronization. Two-way satellite time transfer is also approaching the accuracy necessary to calibrate GPS instrumental errors. Quasar observations are currently performed weekly on each baseline to determine Earth orientation and clock offsets and rates. If receiver delays can be held constant to within the 0.5-nsec limit given in Section II, it appears that instrumental errors in GPS clock synchronization can be dealt with by weekly quasar calibration. If this is not possible, two-way satellite time transfers on a more frequent basis may be necessary.

[4] L. E. Young, op. cit.

V. Conclusion

A subnanosecond clock synchronization capability would be of great benefit to deep space tracking. This level of synchronization would allow spacecraft angular position to be measured to an accuracy of 50 nrad by differential ranging at two widely separated tracking stations with the potential for near-real-time capability. Quasar-based differential VLBI, which is currently used for this purpose, might be reserved for only the most demanding navigation challenges.

Clock offsets between DSN stations with formal errors of approximately 0.5 nsec have been determined from GPS measurements. Comparisons of VLBI and GPS clock rates and analysis of clock closure suggest that these formal errors are a realistic measure of the precision of the GPS clock solutions. The calibration of absolute station instrumental delays and of the offset between VLBI time (used to time tag differential spacecraft range measurements) and GPS receiver time appear to be tractable implementation tasks.

Acknowledgments

The authors thank Susan G. Finley, Susan Oliveau, Lawrence Young, Thomas Meehan, and Ruth Neilan for their contributions to this work. The GIG'91 GPS data were collected by dozens of technical institutions worldwide, and the participation of these collaborators was essential for the success of the experiment and the clock synchronization analysis presented here.

References

[1] J. S. Border and E. R. Kursinski, "Deep Space Tracking and Frequency Standards," *Proceedings of the 45th Annual Symposium on Frequency Control*, Los Angeles, California, pp. 594–607, May 29–31, 1991.

[2] L. A. Buennagel, D. J. Spitzmesser, and L. E. Young, "One Nanosecond Time Synchronization Using Series and GPS," *Proceedings of the 24th Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, NASA Conference Publication 2265, Greenbelt, Maryland, pp. 605–620, December 1–2, 1982.

[3] D. W. Allen and M. Weiss, "Accurate Time and Frequency Transfer During Common-View of a GPS Satellite," *Proceedings of the 34th Annual Symposium on Frequency Control*, Philadelphia, Pennsylvania, pp. 334–346, 1980.

[4] W. Lewandowski, C. Petit, C. Thomas, and M. Weiss, "The Use of Precise Ephemerides, Ionospheric Data and Corrected Antenna Coordinates in a Long-Distance GPS Time Transfer," *Proceedings of the 22nd Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*," NASA Conference Publication 3116, Vienna, Virginia, pp. 547–548, December 4–6, 1990.

[5] C. Veillet, D. Feraudy, J. M. Torre, J. F. Mangin, P. Gruder, F. Baumont, J. Gaignebet, J. L. Hatat, W. Hanson, A. Clements, J. Jesperson, M. Lombardi, D. Kirchner, H. Ressler, S. S. Fassi, W. J. Klepczynski, P. J. Wheeler, W. Powell, A. Davis, M. Granveaud, and D. Uhrich, "LASSO, Two-Way and GPS Time Comparisons: A (Very) Preliminary Status Report," *Proceedings of the 22nd Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, NASA Conference Publication 3116, Vienna, Virginia, pp. 575–582, 1990.

[6] W. J. Klepczynski, M. Miranian, J. Jeffries, P. J. Wheeler, and W. Powell, "Comparison of Two-Way Satellite Time Transfers and GPS Common View Time Transfers," 43rd Annual Symposium on Frequency Control, Denver, Colorado, pp. 199–201, 1989.

[7] T. K. Meehan, T. Munson, J. Srinivasan, J. B. Thomas, L. E. Young, and R. Neilan, "ROGUE: A New High Accuracy Digital GPS Receiver," paper presented at the 19th General Assembly, Int. Union of Geod. and Geophys., Vancouver, Canada, August 9–22, 1987.

[8] W. G. Melbourne, S. S. Fisher, R. E. Neilan, T. T. Yunck, B. Engen, Ch. Reigber, and S. Patevjan, "The First GPS IERS and Geodynamics Experiment—1991," XX General Assembly International Union of Geodesy and Geophysics, Symposium G2.2, International Association of Geodesy, Vienna, Austria (in press), 1991.

[9] S. M. Lichten, "Towards GPS Orbit Accuracy of Tens of Centimeters," *Geophys. Res. Lett.*, vol. 17, pp. 215–218, 1990.

[10] S. M. Lichten, "Estimation and Filtering for High-Precision GPS Positioning Applications," *Manuscripta Geodaetica*, vol. 15, pp. 159–176, 1990.

[11] G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, New York: Academic Press, 1977.

[12] M. H. Murray, R. W. King, and P. J. Morgan, "SV5: A Terrestrial Reference Frame for Monitoring Crustal Deformation with the Global Positioning System," 1990 Fall AGU Meeting, *EOS Trans. AGU*, vol. 71, p. 1274, 1990.

[13] *IERS Monthly Bulletin*, B37, IERS Central Bureau, Observatoire de Paris, Paris, France, 1991.

[14] *IERS Monthly Bulletin*, B38, IERS Central Bureau, Observatoire de Paris, Paris, France, 1991.

[15] J. B. Thomas, *Functional Description of Signal Processing in the Rogue GPS Receiver*, JPL Publication 88-15, Jet Propulsion Laboratory, Pasadena, California, June 1, 1988.

[16] J. A. Steppe, S. H. Oliveau, and O. J. Sovers, "Earth Rotation Parameters From DSN VLBI: 1991," JPL Geodesy and Geophysics Preprint no. 210, April, 1991 (submitted to the International Earth Rotation Service Annual Report for 1991).

Table 1. Late 1990s spacecraft–quasar differential VLBI error budget.

Error source	Magnitude, nsec
Quasar SNR	0.11
Spacecraft SNR	0.033
Quasar position	0.066
Clock offset	0.030
Clock instability	0.003
Phase ripple	0.10
Station location	0.033
Earth orientation	0.056
Troposphere	0.080
Ionosphere	0.010
Solar plasma	0.017
RSS	0.22

Table 2. Clock rate estimates.

Date, 1991	Baseline	VLBI, nsec/day	GPS, nsec/day
January 23	Canberra–Goldstone	−10.0 ±1.6	−9.0 ±0.9
January 27	Madrid–Goldstone	−4.5 ±2.2	−2.7 ±0.5
January 30	Canberra–Goldstone	−7.9 ±3.1	−9.1 ±0.9
February 6	Canberra–Goldstone	46.6 ±1.3	46.4 ±1.9
February 10	Madrid–Goldstone	5.9 ±1.8	2.8 ±0.4

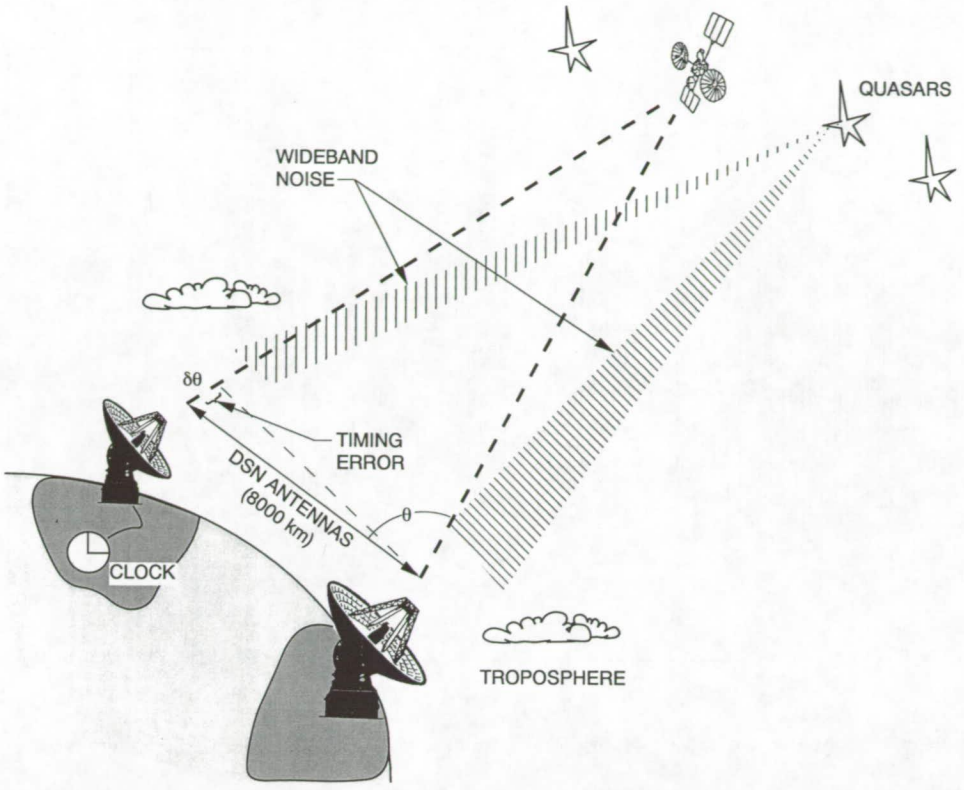


Fig. 1. Spacecraft tracking using quasar-differenced differential one-way range.

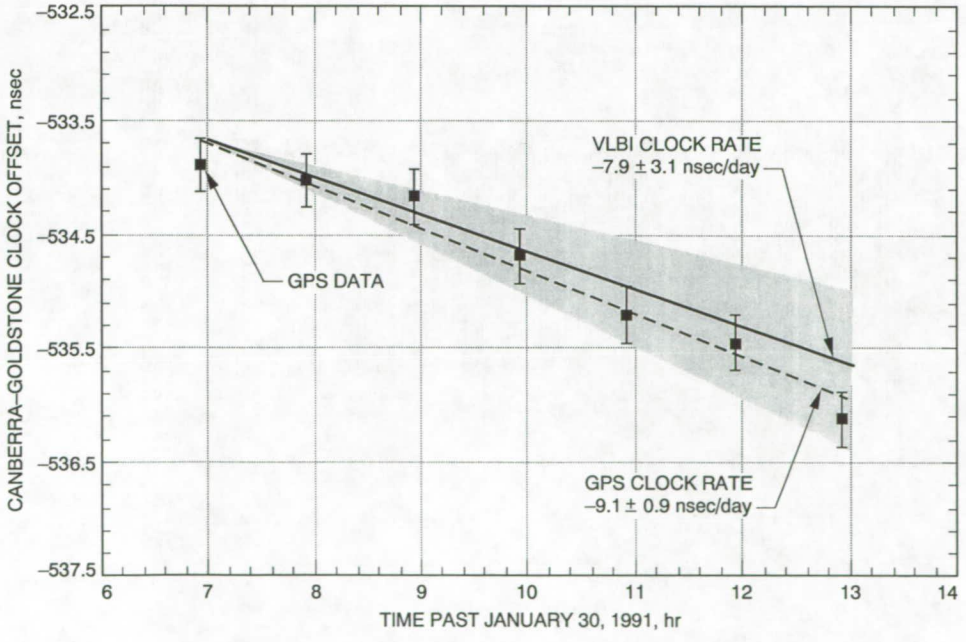


Fig. 2. GPS-based clock estimates and linear fit.

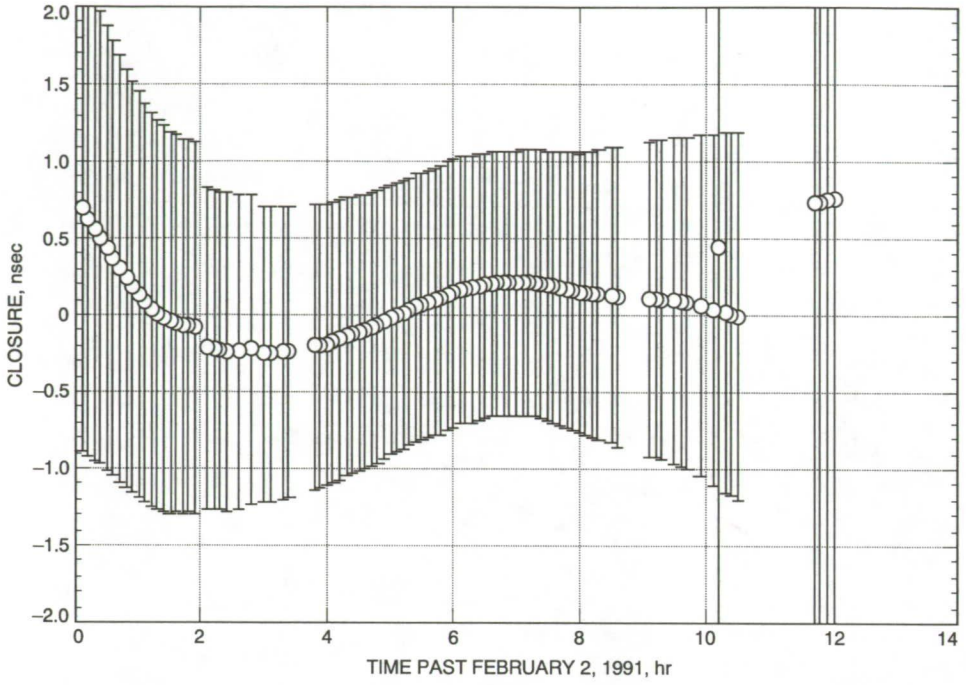


Fig. 3. Canberra–Goldstone–Madrid clock closure with unconstrained GPS satellite orbits.

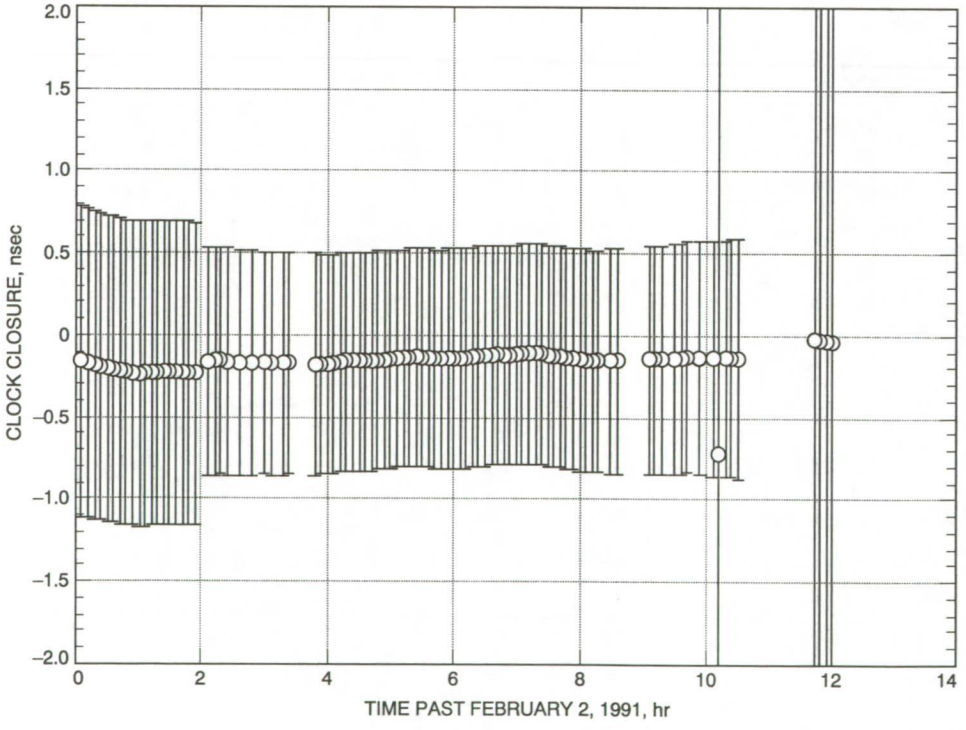


Fig. 4. Clock closure with GPS satellite orbits initialized from previous day.

Precision X-band Radio Doppler and Ranging Navigation: Mars Observer Interplanetary Cruise Scenario

J. A. Estefan and S. W. Thurman
Navigation Systems Section

This article describes an error covariance analysis based on a Mars Observer mission scenario; the study was performed to establish the navigation performance that can potentially be achieved in a demonstration of precision two-way X-band (8.4-GHz) Doppler and ranging with the Mars Observer spacecraft planned for next year, and to evaluate the sensitivity of the predicted performance to variations in ground system error modeling assumptions. Orbit determination error statistics computed for a 182-day Doppler and ranging data arc predicted Mars approach orbit determination accuracies of about 0.45 μrad in an angular sense, using a conservative ground system error model as a baseline. When less-conservative error model assumptions were employed, it was found that orbit determination accuracies of 0.19 to 0.30 μrad could be obtained; the level of accuracy of the assumed Mars ephemeris is about 0.11 μrad. In comparison, Doppler-only performance with the baseline error model was predicted to be about 1.30 to 1.51 μrad, although it was found that when improved station location accuracies and Global Positioning System–based tropospheric calibration accuracies were assumed, accuracies of 0.44 to 0.52 μrad were predicted. In the Doppler plus ranging cases, the results were relatively insensitive to variations in ranging system and station delay calibration uncertainties of a few meters and tropospheric zenith delay calibration uncertainties of a few centimeters.

I. Introduction

The Mars Observer (MO) spacecraft was launched on September 25 of this year and is the first of a series of planetary observer missions to be flown to the inner planets and small bodies of the Solar System, using modified versions of existing Earth-orbiting spacecraft [1]. MO carries an X-/X-band (7.2-GHz uplink/8.4-GHz downlink) transponder and is the first interplanetary spacecraft to rely solely on a single-frequency X-band telecommunications system.[1] The MO mission will provide another opportunity to test the range data filtering technique that proved to be very successful in recent demonstrations utilizing S-/S-band

[1] P. B. Esposito, S. W. Demcak, D. C. Roth, W. E. Bollman, and C. A. Halsell, *Mars Observer Project Navigation Plan*, Project Document 642-312, Rev. C (internal document), Jet Propulsion Laboratory, Pasadena, California, June 15, 1990.

(2.1-GHz uplink/2.3-GHz downlink) ranging data from Galileo [2] and S-/X-band (2.1-GHz uplink/8.4-GHz downlink) ranging data from Ulysses [3,4].

Two-way ranging has been an operational data type for interplanetary spacecraft navigation for many years. Past mission experience has been that ranging data cannot be utilized at their inherent accuracy due to the influence of small unmodeled spacecraft nongravitational forces caused by attitude control thruster firings and other spacecraft activity. In addition, inconsistent and unreliable station delay calibrations often precluded the effective use of precise ranging. In this article, it will be shown that for the MO spacecraft's approach to Mars, notable orbit determination accuracies can be achieved with X-band Doppler and ranging by taking advantage of relatively recent improvements in the consistency and accuracy of station delay calibrations. These improvements, coupled with explicit modeling of spacecraft nongravitational forces and residual station delay calibration errors, should make it possible to utilize MO ranging data at or near their inherent accuracy, even for data arcs much longer than those used in the recent demonstrations.

II. Orbit Determination Error Analysis

The Orbit Analysis and SImulation Software set (OASIS/314) was used to perform the error covariance analysis. The numeric qualifier represents the JPL Navigation Systems Section (314) version of this software, which was modified extensively by A. S. Konopliv to support interplanetary work. The original version of OASIS was designed for Earth orbiter systems error analysis.

A. Earth-to-Mars Interplanetary Cruise

MO's interplanetary cruise has been segmented by the MO Navigation Team into five independent phases, each ending prior to a planned trajectory correction maneuver. The strategy assumed for analysis purposes, as described in the MO Navigation Plan,[2] was to simulate n days of data for each individual phase and determine the spacecraft state vector at the initial epoch of each phase, along with the remaining error model parameters and their associated uncertainties.[3] The error statistics were then propagated, or mapped, to the time of encounter and displayed in a target-centered aiming plane ("B-plane") coordinate system, which is defined in the Appendix.

The MO trajectory segment selected for this analysis was the fourth phase (184-day interval), which represents the longest leg of the interplanetary cruise and has the most stringent orbit determination accuracy requirements, in order to support the final trajectory correction maneuver prior to Mars orbit insertion. The analysis corresponds to an injection epoch of September 16, 1992, the first date of the 20-day launch window. The focus of more recent orbit determination analyses has been the last date of the launch window, October 5, 1992; however, the first date was chosen for ease of reference to results presented by Esposito, et al.[4] in which initial conditions at the earlier date of the launch window are provided.

B. Radio Metric Data Acquisition Strategy

Using a nomenclature in which I = injection, E = encounter, and T_0 = epoch time, two-way Doppler and range data were simulated from T_0 = February 7, 1993, 00:00:00.000 UTC (I + 143) to I + 325 days (E − 12 days) and were assumed to have been acquired from the Deep Space Network's 34-m high-efficiency (HEF) subnet. From I + 143 to E − 90 days, Doppler and range data were acquired during a daily tracking pass. From E − 90 to E − 30 days, two passes of Doppler and range data were acquired on a daily basis, and from E − 30 to E − 12 days, Doppler and range data were acquired continuously. In all cases, the Doppler and range data were sampled at a rate of 1 point every 10 min, or 6 points/hr.

To account for data noise, the Doppler were weighted (an assumed random measurement uncertainty was chosen) at a 1-σ uncertainty of 0.032 mm/sec over a 600-sec integration time (equivalent to a 0.1-mm/sec weight for a 60-sec integration time) or "deweighted" to 0.32 mm/sec in some cases. A 60-sec Doppler weight of 0.1 mm/sec is believed to be near the true inherent accuracy of the data at X-band (8.4 GHz). It is not expected that the Doppler data quality will significantly degrade during the cruise phase, as the first solar conjunction does not occur until December 12, 1993, several months after encounter.[5] The Doppler data weight was adjusted by an elevation-dependent function (to reduce the weight of the low-elevation data) at each station, and data points below a 10-deg elevation were omitted. A similar approach was used for the range data, which were weighted at a 1-σ uncertainty of 10 m or, in some cases, 1 m.

C. Navigation Error Modeling

In this subsection, the modeled parameters are broken down into two categories: estimated and considered; random data noise characteristics, which were described

[2] Ibid.

[3] Ibid.

[4] Ibid.

[5] Ibid.

previously, are summarized in Table 1 along with all the estimated and considered parameters and their associated a priori uncertainties. Modeling assumptions are described below.

1. Estimated Parameters. The estimated parameters include state variables to account for mismodeling of spacecraft nongravitational (NG) accelerations induced by environmental disturbances, such as solar radiation pressure (SRP), as well as spacecraft self-induced perturbing forces, such as gas leaks from valves and pressurized tanks, attitude thruster misalignments, and/or angular momentum desaturation burn mismodeling. These parameters, together with random biases representing station delay calibration errors and other uncalibrated delays in the range measurements, were estimated along with the spacecraft trajectory by using a batch-sequential, factorized Kalman filter.

For the MO SRP model, an area-to-mass ratio of 17.0 $m^2/2328.8$ kg was assumed. Three nondimensional coefficients representing a simple spacecraft bus model (G_r, G_x, G_y) were estimated, with a priori uncertainties equivalent to 10 percent of their maximum attainable values (see Table 1). In a recently published MO interplanetary cruise error analysis,[6] a much more sophisticated SRP model was employed and will probably be used during the operational phase of the mission. Although the model used in this analysis is simplistic, it does provide a reasonable standard of comparison with the earlier studies.[7] Spacecraft nongravitational accelerations in each spacecraft body-fixed axis were modeled as estimated stochastic parameters. A three-parameter, first-order Markov colored noise model (exponentially correlated process) was used, with steady-state sigmas of 3×10^{-12} km/sec^2, and time constants (correlation times) of 1 day, with a 1-day batch size. The random bias parameters representing ranging system calibration errors on the data were also estimated as stochastic parameters with a correlation time of zero and a steady-state sigma equal to an a priori offset calibration uncertainty of 5 m or 2 m in some cases. A separate range bias parameter was modeled for each station pass and a 1-day batch size assumed.

2. Considered Parameters. Recall that a consider parameter is treated by the filter as an unmodeled systematic error and may significantly affect the error statistics of the estimated parameter set. The total error covariance, or full-consider covariance, accounts for the effects of consider parameter uncertainties as well as the formal covariance computed by the filter, so as not to understate the predicted navigation performance. The considered parameters used in this study account for systematic errors in the Earth–Mars ephemerides, Mars' Newtonian gravitational parameter (GM), DSN station locations, tropospheric calibration data, and ionospheric calibration data.

The error covariance from planetary ephemeris DE234 was chosen to account for the Earth–Moon barycenter and Mars ephemeris errors. The 12 × 12 formal error covariance matrix gives errors in the six Brouwer and Clemence Set III orbital elements of the Earth–Moon system and the six elements of Mars. The formal covariance was scaled by a factor of 2, as suggested by its creators, to reflect a more realistic assessment of the errors.[8] Uncertainty of the Mars GM value was taken to be 0.15 km^3/sec^2, or about 3 parts in 10^6.

Station location uncertainties include both a relative component and an absolute (geocentric) component. The relative component refers to DSN site-to-site uncertainty, which is measured accurately by very long baseline interferometry (VLBI); the geocentric component refers to a common error in locating the DSN sites with respect to the Earth's center of mass (VLBI is insensitive to this component). Two station location covariances for the DSN HEF stations (DSS's 15, 45, and 65) were used for this analysis: Moyer's LS234IP station location covariance,[9] and Folkner's MO station location covariance.[10] In the case of LS234IP, the formal covariance representing spin radius and longitude errors were scaled by a factor of 5 and 2.5 (in sigma), respectively, as suggested by Moyer.[11] For Folkner's set, which is based on VLBI and lunar laser ranging (LLR) data, the formal covariance was scaled by a factor of 2 (in sigma) to give realistic station-to-station uncertainties comparable with differences derived independently from satellite laser ranging (SLR) and VLBI. The formal covariance was also inflated to account for

[6] D. C. Roth, "Orbit Determination Results for Planetary Protection Analysis (5 OCT 92 Launch Date)," JPL Interoffice Memorandum NAV-92-007 (internal document), Jet Propulsion Laboratory, Pasadena, California, April 15, 1992.

[7] Esposito, Demcak, Roth, Bollman, and Halsell, op. cit.

[8] E. M. Standish, "The JPL Planetary Ephemerides, DE234/LE234," JPL Interoffice Memorandum 314.6-1348 (internal document), Jet Propulsion Laboratory, Pasadena, California, October 8, 1991.

[9] T. D. Moyer, "Station Location Set LS234IP for Planetary Ephemeris DE234," JPL Interoffice Memorandum 314.5-1588 (internal document), Jet Propulsion Laboratory, Pasadena, California, November 20, 1991.

[10] W. M. Folkner, "DE234 Station Locations and Covariance for Mars Observer," JPL Interoffice Memorandum 335.1-92-013 (internal document), Jet Propulsion Laboratory, Pasadena, California, May 26, 1992.

[11] Moyer, op. cit.

geocenter offset between LLR and SLR measurements, radio-planetary frame-tie uncertainty, and Earth orientation mismodeling.[12] Table 1 provides the absolute and relative uncertainties in station cylindrical coordinates for both station location sets.

Wet and dry components of the tropospheric path delay calibration error were considered with 1-σ zenith uncertainties of 4 cm and 1 cm, respectively, based on present-day accuracy of adjustments to a seasonal model. In some cases, Global Positioning System (GPS)–based calibrations were assumed to be available, with a total 1-σ zenith delay uncertainty of 2 cm. The model used for ionospheric refraction is described by Wu [5]. For this analysis, a 1-σ zenith electron content uncertainty of 5×10^{16} elec/m^2 was assumed and is representative of current DSN calibration capability using ground-based observations of GPS satellites.

III. Accuracy Assessment

The filter-generated computed covariance was combined with consider parameter sensitivities in order to construct the full-consider covariance, which was then mapped to encounter (August 19, 1993, 12:48:08.000 ET) and displayed in Mars-centered aiming plane ("B-plane") coordinates, referred to the Mars Mean Equator of Date. In some instances, the mapped computed-only statistics are tabulated along with the full-consider statistics for comparison of "filter-world" versus "real-world" results.

A. Doppler-Only Performance

The baseline case for Doppler-only analysis, utilizing the conservative error model described in the previous section (LS234IP station location covariance and present tropospheric calibration accuracy), and assuming a Doppler weight of 0.1 mm/sec (60-sec count time), yields aiming plane dispersions equivalent to about 1.51 μrad in an angular sense (by comparison, the accuracy requirement for the fourth phase of the interplanetary cruise is about 0.5 μrad[13]). The statistical results for this case are given in Table 2, where SMAA and SMIA denote the semi-major and semi-minor axes of the dispersion ellipse, respectively; θ denotes the orientation of the dispersion ellipse, measured clockwise from the horizontal axis of the aiming plane; and LTOF denotes the linearized time-of-flight uncertainty. The Doppler data in this case exhibited a very high sensitivity to station location uncertainty, which was about 1.5 m (Doppler data are nearly insensitive to z-height errors). In an initial attempt to reduce this sensitivity, the Doppler data were "deweighted" by an order of magnitude, to a value of 1.0 mm/sec. This reduced the aiming plane dispersions by about 30 percent, and the time-of-flight uncertainty by about 20 percent; the angular accuracy for this case was about 1.10 μrad. When the baseline Doppler-only case (0.1-mm/sec weight) was repeated with the Folkner station location covariance, the aiming plane dispersions and time-of-flight uncertainty were reduced by about a factor of 2 relative to the baseline case, which corresponded to an angular accuracy of about 0.67 μrad. If it was assumed that GPS-based calibrations of the troposphere were used in addition to the Folkner station covariance, the aiming plane dispersions and time-of-flight uncertainty were further reduced by about 22 percent, yielding an angular accuracy on the order of 0.52 μrad. These results suggest that the full benefit of X-band Doppler can only be realized with Earth platform (station locations and Earth orientation) calibration accuracies of about 15 cm, and centimeter-level zenith tropospheric calibration accuracies.

B. Doppler Plus Range

When ranging data were included along with the Doppler data, the results improved substantially over those obtained in the Doppler-only cases. Statistical results for these cases are given in Table 3. Use of the conservative baseline error model (10-m range weight and a 5-m a priori range bias uncertainty for each station pass) resulted in an angular accuracy of about 0.50 μrad. As with the Doppler-only results, the use of an improved zenith tropospheric delay calibration value did not significantly improve performance in this case; however, when Folkner's improved station location covariance was utilized, much better performance was obtained, with improvements in the aiming plane dispersions on the order of 40 percent and time-of-flight uncertainty of about a factor of 2. The use of the improved GPS calibrations of the troposphere together with Folkner's station covariance resulted in only about an additional 7- to 9-percent improvement.

In an attempt to further reduce the data sensitivity to systematic error sources, the Doppler measurements were deweighted by an order of magnitude, with mixed results (see Table 3). Another case was evaluated in which the a priori uncertainties for the range bias parameters were reduced to 2 m, which should be achievable for X-band ranging data. Here, only a slight improvement in overall performance was observed; a different result from that witnessed in the Galileo and Ulysses precision ranging demonstrations, in which shorter data arcs were assumed [2–4].

[12] Folkner, op. cit.

[13] P. Esposito, personal communication, Mars Observer Navigation Team Chief, Navigation Systems Section, Jet Propulsion Laboratory, Pasadena, California, March 1992.

C. Doppler Plus High-Precision Range

Recent experiments by T. P. McElrath using S-/X-band two-way ranging data obtained from Ulysses near the spacecraft's recent Jupiter encounter have met with remarkable success using range data weights of 2 m [4]. With a 2-m range weight, a substantial improvement in orbit determination performance was obtained for Ulysses over orbit solutions computed with 10-m and 5-m range weights; an earlier analysis indicated that it should be possible to weight X-/X-band ranging data at 1 m or better [3]. To predict the potential performance of high-precision ranging for MO, an additional set of cases was calculated using a range weight of 1 m and assumed a priori range bias uncertainties of 2 m. The results obtained for these cases are provided in Table 4. For the case in which the Folkner station location covariance was used, orbit determination accuracies of 0.19 to 0.30 μrad were predicted; by comparison, the level of accuracy of the assumed Mars ephemeris is about 0.11 μrad. The use of an improved GPS-based tropospheric calibration accuracy produced no noticeable change, as shown in Table 4. One final case was computed (not shown in the table) using the Folkner station location covariance, GPS-based tropospheric calibration accuracy, and a priori range bias uncertainties of 5 m; once again, no noticeable change in the aiming plane dispersions was observed. It is interesting to observe that the use of an order-of-magnitude improvement in the range data noise value did not result in a dramatic improvement over the the 10-m range case, except when the Doppler data were deweighted to 1.0 mm/sec. This is consistent with the fact that the a priori range measurement biases, used at their current level, impose a limit on the ability of the range data to provide a greater share of the information content over the "tight" (0.1-mm/sec) Doppler.

Figures 1 and 2 provide a graphical comparison of the aiming plane statistics (dispersion ellipses and uncertainty in time-of-flight) for the three primary data strategies addressed in this article: Doppler-only navigation, Doppler plus 10-m range navigation, and Doppler plus high-precision (1-m) range navigation. Each result is representative of the improved orbit determination error model which incorporates Folkner's station location covariance and GPS-based tropospheric calibrations. The improvement of the Doppler plus range orbit solutions over the Doppler-only result is clearly evident in the illustrations.

IV. Discussion

There are some subtle differences in terms of modeling and data acquisition strategies between the analysis described in this article and the recent interplanetary cruise error analysis performed by the MO Navigation Team.[14] As alluded to earlier, a more sophisticated SRP model was mechanized in the Navigation Team's analysis, which will be a good candidate for use in mission operations. For completeness, the analysis provided here could be revisited by using the more realistic model. Also, the first day of the 20-day launch window was used in this analysis versus the last day of the 20-day launch window. This is such a short interval of time relative to the interplanetary cruise phase selected for this study that the differences are considered insignificant for covariance analysis purposes. Additionally, the data cutoff for the navigation team's study was E–25 days versus E–12 days assumed herein.

V. Conclusions

The results of the error covariance analysis predicts that Doppler-based orbit solutions for MO will be much more sensitive to Earth platform and transmission media calibration errors than solutions derived from both Doppler and ranging data. The Doppler and ranging cases analyzed herein also showed that orbit solutions using high-precision (meter-level) ranging data appear to be nearly insensitive to variations in ranging system and station delay calibration errors of a few meters, and tropospheric calibration uncertainties of a few centimeters at zenith. Not studied here was the potential performance of precise to highly precise radio Doppler and ranging navigation resulting from improved data noise characteristics and a priori uncertainties of the range measurement biases. Whether these findings will be consistent with the results obtained with shorter (one-and-a-half to three-month) data arcs in the Galileo and Ulysses precision ranging demonstrations is yet to be determined.

---

[14] D. C. Roth, op. cit.

Acknowledgment

The authors would like the thank V. M. Pollmeier for his helpful comments and suggestions in review of this article.

References

[1] J. F. Jordan and L. J. Wood, "Navigation, Space Mission," *Encyclopedia of Physical Science and Technology*, New York: Academic Press, vol. 8, pp. 744–767, 1987.

[2] V. M. Pollmeier and S. W. Thurman, "Application of High-Precision Two-Way Ranging to Galileo Earth-1 Encounter Navigation," *TDA Progress Report 42-110*, vol. April–June 1992, Jet Propulsion Laboratory, Pasadena, California, pp. 21–32, August 15, 1992.

[3] S. W. Thurman, T. P. McElrath, and V. M. Pollmeier, "Short-Arc Orbit Determination Using Coherent X-band Ranging Data," paper AAS 92-109, AAS/AIAA Spaceflight Mechanics Meeting, Colorado Springs, Colorado, February 24–26, 1992.

[4] T. P. McElrath, B. Tucker, K. E. Criddle, P. R. Menon, and E. S. Higa, "Ulysses Navigation at Jupiter Encounter," paper AIAA 92-4524 (in press), AIAA/AAS Astrodynamics Conference, Hilton Head, South Carolina, August 10–12, 1992.

[5] S. C. Wu, "Atmospheric Media Effects on ARIES Baseline Determination," *TDA Progress Report 42-103*, vol. July–September 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 61–69, November 15, 1990.

Table 1. Assumed orbit determination modeling errors.

Source	A priori uncertainty, 1σ	Remarks
Random data noise		
Two-way Doppler (60-sec average)	Baseline, 1.0 mm/sec Precise, 0.1 mm/sec	
Two-way range	Baseline, 10 m Precise, 1 m	
Estimated parameters		
Spacecraft state vector	Position, 10^5 km Velocity, 1 km/sec	Loose knowledge
SRP constant coefficients (percent of nominal value)	G_r = 10 percent (= 0.13) G_x = 10 percent (= 0.01) G_y = 10 percent (= 0.01)	Radial and lateral components
Estimated stochastic parameters		
Spacecraft self-induced accelerations	Radial, 3×10^{-12} km/sec^2 Lateral, 3×10^{-12} km/sec^2	τ (correlaton time) = 1 day
Range biases (1 per station pass)	Baseline, 5 m Precise, 2 m	τ (correlaton time) = 0
Considered parameters		
Mars planetary ephemeris (geocentric)	Radial, 0.4 km Down track, 30 km Normal, 48 km	JPL ephemeris DE234 (scaled by 2-in. sigma)
Mars GM	0.15 km^3/sec^2	Conservative
DSN station coordinates baseline (LS234IP)	Spin radius, 0.84 m Longitude, 0.47 m z-height, 9.99 m	Relative uncertainties between stations $\approx$ 10 cm
Precise (Folkner)	Spin radius, 0.18 m Longitude, 0.19 m z-height, 0.22 m	Relative uncertainties between stations $\approx$ 3–5 cm
Tropospheric zenith delay calibration error	4 cm (wet) 1 cm (dry)	
GPS-based tropospheric calibration error	2 cm (total)	
Ionospheric zenith electron content calibration error (at "ionospheric bulge")	5×10^{16}e/m^2	

Table 2. 1-σ aiming plane statistics for DSN Doppler-only navigation.

SMAA,[a] km	SMIA,[b] km	θ, deg	LTOF[c] sec
Baseline case (LS234IP station covariance)			
64	38	164	14[d]
515	119	25	187
Folkner station covariance			
226	81	25	82
Folkner station covariance with GPS tropospheric calibrations			
177	77	28	64
Baseline case with GPS tropospheric calibrations			
495	117	25	180
Baseline case with deweighted Doppler			
279	104	50	96[d]
372	136	43	144

[a] Semi-major areas.
[b] Semi-minor areas.
[c] Linearized time-of-flight.
[d] Computed-only results.

Table 3. 1-σ aiming plane statistics for DSN Doppler plus range navigation.

SMAA, km	SMIA, km	θ, deg	LTOF,[a] sec
Baseline case (LS234IP station covariance)			
27	0.06	64	5.8[a]
169	1.9	64	65
Folkner station covariance			
99	1.2	64	32
Folkner station covariance with GPS tropospheric calibrations			
92	1.1	64	29
Baseline case with GPS tropospheric calibrations			
165	1.9	64	64
Baseline case with deweighted Doppler			
167	0.29	64	35[a]
187	1.7	64	43

[a] Computed-only results.

Table 4. 1-σ aiming plane statistics for DSN Doppler plus high-precision range navigation.

SMAA, km	SMIA, km	θ, deg	LTOF,[a] sec
Baseline case (LS234IP station covariance)			
26	0.06	64	5.6[a]
154	2.0	64	60
Folkner station covariance			
92	1.2	64	30
Folkner station covariance with GPS tropospheric calibrations			
87	1.1	64	27
Baseline case with GPS tropospheric calibrations			
150	1.9	64	59
Baseline case with deweighted Doppler			
81	0.15	64	17[a]
118	1.7	63	29

[a] Computed-only results.

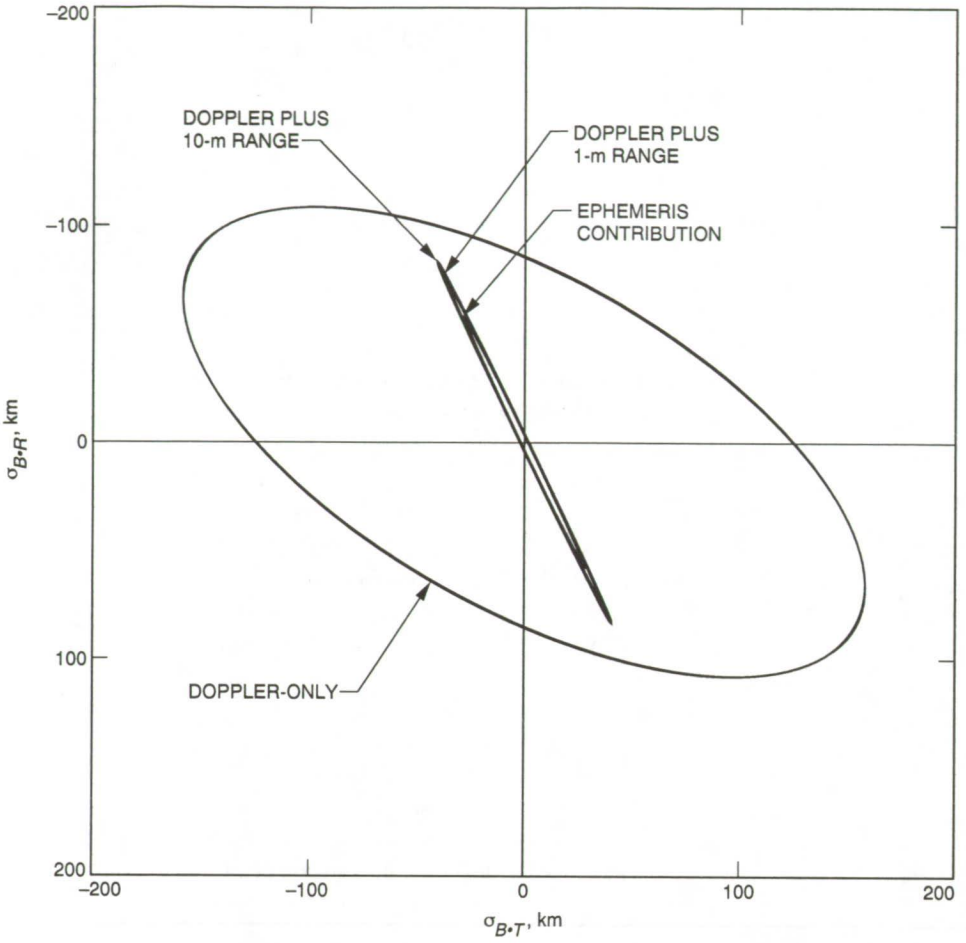


Fig. 1. Aiming plane dispersion ellipses for different data strategies.

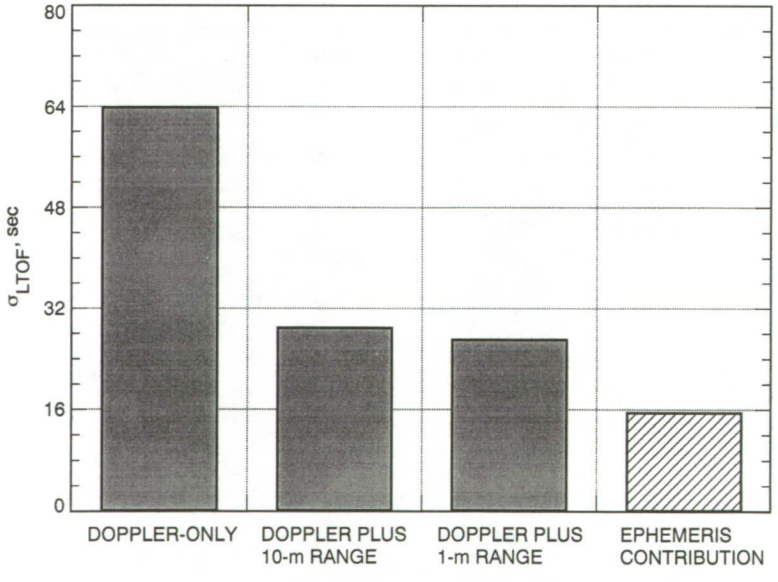


Fig. 2. Uncertainty in linearized time-of-flight (LTOF) for different data strategies.

Appendix

Planetary approach trajectories are typically described in aiming plane coordinates, often referred to as "B-plane" coordinates (see Fig. A-1). The origin of this system is the center of the target body. The encounter parameters are defined in terms of the orthogonal unit vectors $\hat{S}$, $\hat{T}$, and $\hat{R}$. The $\hat{S}$ vector is parallel to the incoming asymptote of the approach hyperbola, while $\hat{T}$ is normally specified to lie in the ecliptic plane (the mean plane of the Earth's orbit); $\hat{R}$ completes the triad. (In this particular analysis, the $\hat{T}$ unit vector was defined to lie in the Martian equatorial plane.) The aim point is specified by the miss vector, $\vec{B}$, which locates where the point of closest approach would be if the target planet had no mass and did not deflect the flight path; the time from encounter (point of closest approach) is defined by the LTOF, which specifies what the time of flight to encounter would be if the magnitude of the miss vector were zero. Orbit determination errors are characterized by a 1-σ or 3-σ B-plane dispersion ellipse, also shown in Fig. A-1, and the 1-σ or 3-σ uncertainty in LTOF. In Fig. A-1, SMAA and SMIA denote the semi-major and semi-minor axes of the dispersion ellipse, respectively.

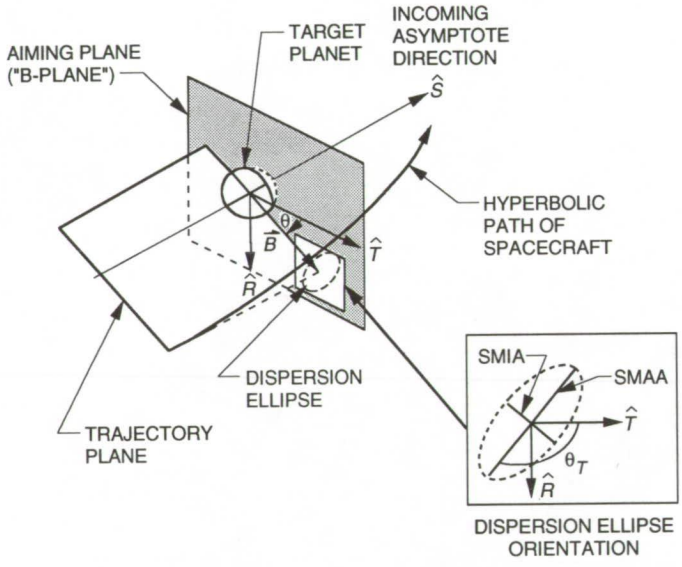


Fig. A-1. Aiming plane coordinate system definition.

Systems Analysis for Ground-Based Optical Navigation

G. W. Null, W. M. Owen, Jr., and S. P. Synnott
Navigation Systems Section

Deep-space telecommunications systems will eventually operate at visible or near-infrared regions to provide increased information return from interplanetary spacecraft. This would require an onboard laser transponder in place of (or in addition to) the usual microwave transponder, as well as a network of ground-based and/or space-based optical observing stations. This article examines the expected navigation requirements for future missions, as well as possible ground-based optical observing systems to meet these requirements. Special emphasis is given to optical astrometric (angular) measurements of stars, solar system target bodies, and (when available) laser-bearing spacecraft, since these observations can potentially provide the locations of both spacecraft and target bodies. The role of astrometry in the navigation system and the development options for astrometric observing systems are also discussed.

I. Introduction

Eventually, deep-space telecommunications will operate at optical or near-infrared wavelengths to provide increased information return; development plans and current progress toward a prototype optical system are described in [1]. The Deep Space Optical Reception Antenna (DSORA) ground system would employ a large, segmented 10-m mirror for reception and a smaller (perhaps 1-m) transmitter telescope. Of course, deep-space communications would require a transmitter/receiver capability on each interplanetary spacecraft.

These laser-based communications links can also be used for spacecraft navigation. This is analogous to the current microwave situation with a very important difference: in addition to the possibility of ranging, Doppler, and interferometric observations to a spacecraft, the optical system will allow ground-based observations of target bodies relative to each other, to a laser-bearing spacecraft, or to the star background. These "ground-based optical navigation" observations can potentially provide information comparable to the existing onboard optical navigation (OPNAV) system. This will be a new capability for the DSN, opening up the possibility that, for some missions, the entire navigation task can be performed with DSN data alone.

For instrument development purposes, the instrument characteristics required for accurate angular observations of a laser-bearing spacecraft with respect to target bodies or stars are similar to the instrument requirements for observing faint solar system objects against a star background. Therefore, as will be discussed, instrument development will make extensive use of target-body observations.

The realization of optical communications systems for interplanetary missions is still a number of years off. However, preliminary versions of the required astrometric ob-

serving systems can be developed and used today to improve mission navigation by reducing target ephemeris errors. This target-body observation capability can provide important navigation benefits for current missions, such as Galileo and Cassini, and it will also be required for future missions which employ laser transmitters. In summary, target-body tracking is of primary importance, both for instrument development and for mission navigation support.

Examples of requirements on ground-based optical navigation for current and future missions are discussed below, as are the possible observing systems to meet those requirements. Both near-term and long-term prospects will be discussed, but the emphasis here is on long-term technology development trends and prospects for the next 10 to 15 years.

Some cost-effective development options were identified for narrow-field telescopes with large format charge-coupled-device (CCD) detectors, and these instruments can potentially provide significant navigation benefits. These near-term options will not be described in detail here, because a full discussion is already available in [2], including a description of ongoing CCD observation programs being conducted as a target location technology demonstration. This work is a collaborative effort with the U. S. Naval Observatory (USNO) Flagstaff Station. For readers unfamiliar with CCDs, a CCD is a silicon-based two-dimensional array of "pixels"; pixels are small individual detectors which very efficiently convert visible light photons to countable electrons.

A very high-level overview of the possible ground-based optical observing scenarios is shown in Fig. 1, which depicts the possible observed objects and types of observatories. The term "filled aperture" is used here to denote conventional single-telescope observing systems. "Stare-mode" systems are guided so that stars are essentially fixed in the field of view; "scan-mode" systems are guided so that stars appear to move at a controlled rate across the field [3]. The challenge to the designer of the ground-based optical navigation system is to find an effective mix of observables (bottom of Fig. 1) to provide accurate astrometric positions for the observed objects (top of Fig. 1) so that the navigation requirements can be met.

This article is divided into seven major sections: Introduction, Observing Strategy, Navigation Requirements, Astrometric Observing System Overview, Filled-Aperture Instruments, Interferometric Instruments, and Summary and Conclusions. Section II examines the observational roles of target-relative astrometry and of optical counterparts to conventional radiometric observations. Section III demonstrates that there are significant navigation requirements for 25-nrad observational accuracy. A generic introduction to observational systems is provided in Section IV, including an explanation of the important relationship between instrument development plans and star-catalog positional accuracy and density. Then, filled-aperture and interferometric instrument development choices are discussed, including several candidate observation systems which potentially could provide the required 25-nrad accuracy. Finally, the most important results are briefly reviewed in Section VII.

II. Observation Strategy

Optical counterparts (optical ranging, Doppler, and differenced range) to current microwave radiometric observables have been investigated by Folkner and Finger [4]. As discussed in [4], these optical methods potentially can provide spacecraft positional accuracy roughly comparable to radio metric techniques, but this must be verified by actually building and testing a prototype system. A key part of this concept is that the angular information from differenced ranging is referenced to the well-known Earth orientation, and so does not require direct observations relative to optical stars. However, these techniques require a spacecraft transponder and are not suitable for stars and target-body observations. For this reason, there will be a navigation requirement for astrometric observing, so that the target bodies can be located relative to the spacecraft.

Although observational systems for optical ranging, Doppler, and differenced range can employ narrow filters to observe laser-bearing spacecraft within a few degrees of the Sun, astrometric systems cannot, because this would usually eliminate too much of the optical signal from stars and target bodies. Therefore, ground-based astrometric systems currently can obtain accurate observations only at night, at geocentric angles of roughly 50 deg or more from the Sun, and it appears likely that this will also be the situation for future systems. This implies that accurate astrometric observations of a given object will be limited to a roughly 260-day span per year. Space-based optical astrometric systems will also have a solar-exclusion constraint, which, depending on sun-shade design and other factors, may be less than 50 deg.

Fortunately, accurate target-body orbits can be determined from observations obtained during these 260-day intervals, and then these deterministic orbits can be accurately extrapolated into the solar exclusion regions. As

discussed in [5, pp. 33–34], a 200-day extrapolation can be accomplished with acceptable accuracy. Spacecraft, however, are subject to unpredictable nongravitational forces which preclude a sufficiently accurate trajectory extrapolation; therefore laser-bearing spacecraft must be observed to within a few degrees of the Sun, as is routinely accomplished for spacecraft with microwave transponders.

In summary, an acceptable ground-based optical observing strategy is to obtain astrometric observations of stars, target bodies, and laser-bearing spacecraft at night, but, during solar exclusion periods, obtain angular positions for the laser-bearing spacecraft with differenced range or other suitable narrow-filter techniques.

III. Navigation Requirements

A complete survey of all possible requirements for ground-based optical measurement systems is beyond the scope of this article. However, two examples of the most important applications of ground-based optical navigation will be discussed and are shown in Fig. 2.

A. Planet or Asteroid Approach

For the planetary approach case shown in the left panel of Fig. 2, the typical radio-only delivery into orbit will be limited primarily by the a priori planetary ephemeris error, which can be assumed to be about 100 to 150 km for Jupiter during the Galileo approach and several hundred kilometers for Saturn. These planetary ephemeris errors cannot be significantly improved with conventional ground-based optical observations, which have errors of 1000 nrad (roughly 750 km at Jupiter; 1500 km at Saturn) or more.

Onboard optical observations will reduce the angular two-dimensional (2-D) error in the plane of the sky to a few tens of kilometers or smaller (i.e., roughly 25 geocentric nrad at Jupiter). If the ground-based optical data are to provide the same level of Jupiter navigation accuracy as onboard imaging, the ground-based optical system must also achieve 25-nrad 2-D measurement accuracies. Note that since the onboard and Earth-based observations usually have different lines of sight, combining these two complementary 2-D observation types will provide a quick three-dimensional (3-D) target-body position fix at roughly the same 25-nrad accuracy.

If the spacecraft carries a laser transmitter and if observing conditions (e.g., instrument field, spacecraft–target angular separation, and Sun angle) are suitable, the ground-based observation will directly measure the angular coordinates of the spacecraft relative to the target. Otherwise, this information can be obtained by an indirect technique, using separate observations of the spacecraft and target body.

As shown in the right panel of Fig. 2, the planet or asteroid target can be observed relative to stars. Observation over at least one-half of the target orbital period will enable an accurate extrapolated orbit determination for the target body. As previously discussed, this can be accomplished without any requirement for daytime observations. Spacecraft observations with radiometric techniques (or optical counterparts for laser-bearing spacecraft) can then provide accurate 3-D spacecraft orbits. Finally, a combination of spacecraft and target-body information yields an accurate 3-D spacecraft–target encounter position to roughly the same accuracy as the direct method.

As will be discussed, the indirect method is dependent on the availability of suitably accurate star catalog positions, with sufficient numbers of stars for a given instrument field, and there are several possible near-term sources for this catalog. Direct target-relative observations of a laser-bearing spacecraft are much less dependent on an accurate star catalog, since only the instrument scale and orientation may depend on this information; also, it may be possible to develop suitable scale and orientation techniques [2] which do not require accurate catalog positions.

B. Intersatellite Observations for Planetary Orbit Phase Navigation

In Fig. 2 (center panel), direct measurements of the intersatellite positions (relative angular measurements between satellites) are indicated. As discussed in later sections and in [2], very near-term technology developments in optical detector systems are the goal, and these potentially can provide ground-based intersatellite measurements of the Jovian or Saturnian systems with about an order of magnitude greater accuracy than can be achieved with the usual photographic techniques. Measurements with these systems would be acquired for several years, ending just prior to spacecraft orbit insertion at the planet.

These observations can potentially reduce the intersatellite and planet-centered ephemeris errors from the 100-km level to the 15-km level, competitive with errors found in onboard data. For both Galileo and Cassini, this satellite ephemeris improvement can have a major impact on the number of onboard optical navigation frames required, thereby saving more downlink capacity science data and reducing navigation tracking coverage requirements.

Note that this major benefit accrues without a laser-bearing spacecraft and with the same type of optical imaging system that would be needed for direct target-relative astrometric measurements of a laser-bearing spacecraft in the optical communication era. This would appear to be a nearly ideal situation, one in which development and extensive testing occur under essentially the same conditions as those that will apply when a new component is added (the laser source on the sky).

Achievable optical astrometric measurement accuracy for natural satellites will probably be limited by the difficulty in accurately modelling satellite shape, and, most importantly, brightness distribution and reflection properties. As discussed in [2], the JPL Optical Systems Analysis Group is currently analyzing Voyager images of the Galilean satellites and expects to calibrate brightness variations to support 25-nrad ground-based astrometry. Further improvements may be possible, but this will require overcoming the effects of incomplete spatial and spectral coverage and other systematic errors.

IV. Astrometric Observing System Overview

This section presents background information on optical astrometric observing systems, preparatory to later, more detailed descriptions of filled-aperture and interferometric observing options. The emphasis here is on generic considerations, which are applicable to all observing systems. Star-catalog positional accuracy and star-relative astrometric accuracy for current systems are reviewed and shown to be inadequate for navigation purposes.

Potential sources for accurate global (all-sky) catalogs are reviewed, and several good candidates are identified. These include the European Space Agency (ESA)/ Hipparcos star catalog (available in the mid-1990s), as well as catalog improvements from ground-based observing systems, such as optical interferometry or optical observation of radio sources tied to the quasar catalog. Catalog densification techniques and atmospheric limitations to astrometric accuracy are also discussed. An instrument development strategy is presented, taking into account expected catalog improvements.

There is a vast literature about ground-based optical astrometry, but much of this is not current or does not supply an adequate error description for moving-body astrometry. A good background for the present article is provided in a recent (1988) survey of astrometric techniques and instruments by Monet [3], but, as will be discussed, Monet's accuracy estimates are usually given in the context of parallax or proper motion solutions; for moving-body astrometry, these accuracy estimates should be interpreted as reproducibility (precision) estimates. The present article does not attempt to duplicate this survey, but instead concentrates on the instruments and techniques required for accurate astrometric observations of moving bodies.

A. Current Astrometric Accuracy for Moving Bodies

Optical astrometry of moving objects (targets and laser-bearing spacecraft) currently has very limited accuracy compared to the 25-nrad observational accuracy goal. At present, the best star-relative observations have errors of at least 500 nrad and the best satellite–satellite observations [2] have errors of about 200 nrad; typical performance is usually much worse.

B. Current Star-Catalog Positional Accuracy

The importance to navigation of an accurate global star catalog is that it allows the connection of high-accuracy observations through large angles on the sky or, in the inner solar system, around the whole orbit of an object such as an asteroid. Also, a sufficiently accurate, dense catalog facilitates an easy, accurate determination of instrument scale and orientation.

Currently available star catalogs provide relatively poor positional accuracy for individual stars; the most accurate catalog (the fundamental FK5 global catalog of 1535 stars) has positional standard errors at a 1990 epoch of about 200 to 250 nrad [6] and provides only about 0.035 stars per square deg. Other catalogs usually provide more stars, but often with nearly an order of magnitude reduction in accuracy.

Of course, these inaccuracies enter directly into the absolute angular coordinates (right ascension and declination) computed from star-relative observations of moving bodies. Catalog positions are also usually required to determine the instrument scale (conversion from linear units on the detector to angular units on the sky) and orientation; potential catalog-independent methods are being tested in a near-term development effort [2], but these are difficult and the outcome is still uncertain.

C. Potential Near-Term Global-Catalog Improvements

Although the positional accuracy of current global catalogs is definitely unsatisfactory, there are excellent prospects for a dramatic improvement in catalog accuracy within the next few years. The ESA Hipparcos mission, currently flying, will produce a global catalog with roughly

2.5 stars per square deg; catalog delivery is scheduled for the mid-1990s. Astrometric errors for an individual star at an early 1990s epoch are expected to be about 10 nrad, but these errors will grow as the stellar proper-motion errors (about 10 nrad/yr) integrate over time [7]. However, it may be possible to significantly reduce the effect of individual star errors by observing target bodies or spacecraft relative to many Hipparcos stars.

By early in the next century, there will be a need for accurate star positions at a second epoch; the long interval between catalog epochs will then provide improved accuracy for star proper motions. The second-epoch positions could be obtained by a second Hipparcos mission or some other method with comparable accuracy, possibly ground based.

Two other techniques may also provide accurate, but sparse, global star catalogs. The first of these is optical filled-aperture observations of radio sources that have accurate positions in the quasar catalog. Some radio stars and quasars are bright enough (roughly 17th magnitude) to be observed with optical instruments [3]. The second technique is optical interferometry (to be discussed later) which can potentially provide an accurate (sub-50-nrad) global catalog of a few hundred stars.

Thus, there are several potential methods for constructing a global catalog of accurate star positions, and some of these methods do not require the Hipparcos catalog. However, these other methods produce sparse catalogs which must be densified with other wide-field instruments.

D. Catalog Densification Methods

As discussed in [2], accurate astrometry of moving bodies usually requires simultaneous observation of two or more stars whose angular positions are accurately known. Since even the Hipparcos global catalog is too sparse for easy use with existing narrow-field instruments, it will usually be beneficial to observe faint stars relative to stars from a global catalog and construct a suitably dense, accurate local star catalog. Other, more sparse global catalogs must be densified for navigation use.

All these densification techniques require a wide-field instrument. As discussed later, two filled-aperture techniques (scanning CCDs and wide-field stare-mode instruments) appear to be most suitable for this task. However, scanning CCD performance is still relatively modest compared to the required accuracy of 25 nrad or better, and suitable stare-mode instruments would require a major development effort. Obviously, assessment of the suitability of the global catalog and of the need for development of a wide-field instrument will be strongly influenced by the availability of a suitable catalog densification technique.

E. Observational Reproducibility Versus Accuracy

At this point, it is useful to briefly explore the effect that current catalog limitations have had on today's astrometric observing systems and procedures. For this purpose, the key concept is the distinction between astrometric reproducibility (i.e., precision) and accuracy.

Reproducibility can be measured by repeated observations of the same star field, taking care that each star is always positioned approximately at the same position on the detector and that the hour angle for each observation is approximately the same. These observations are reduced to the same scale and orientation by using star-catalog positions, which need only be known to a few arcsec. As will be discussed, reproducibility at the 25- to 50-nrad level has been demonstrated for some instruments with modern detectors.

Night-to-night reproducibility is sensitive to the signal-to-noise ratio, to image jitter caused by atmospheric fluctuations, and to star properties that change between observations. The best-known examples of changing star properties are star parallax and proper motion, which can be accurately determined from the small changes in positions of the target star.

Astrometry for moving targets or for star-catalog generation cannot be performed with the restrictive assumptions used for reproducibility observations, and, therefore, there are many additional error sources that affect astrometric accuracy, but not reproducibility. These include all the astrometric errors induced by star properties that do not change between observations (i.e., star catalog position errors, star brightness, color, and image position in the detector field).

These properties are different for each star and also there are errors in their numerical values. For both reasons, these effects can induce errors in such calibrations as instrument aberration (distortion), instrument scale-value and orientation, atmospheric dispersion, differential refraction, and detector defects. Also, moving-body observations require significant changes in zenith and hour angles, which cause additional errors for calibration of atmospheric effects and instrumental gravity flexure.

Thus, errors affecting accuracy (but not reproducibility) have had a significant effect for astrometry. Limi-

tations caused by star-catalog position have seriously affected both the ability to calibrate these other errors and the motivation to do so. Fortunately, this star-catalog limitation may soon be effectively removed.

F. Effect of Earth's Atmosphere on Astrometric Accuracy

The Earth's atmosphere affects astrometric accuracy primarily from the effects of atmospheric dispersion (i.e., changes in index of refraction with incident wavelength [8,9]) and image jitter caused by fluctuations in the atmosphere. However, the astrometric effect of the atmosphere can be reduced through proper observational procedures, consistent with a 15-nrad atmospheric contribution to a 25-nrad error budget.

Atmospheric dispersion can be reduced by the use of narrow filters and by obtaining adequate knowledge of the spectral characteristics of each observed object. Narrow filters create a requirement for larger aperture telescopes and/or higher efficiency detectors, especially for the observation of faint objects. Further discussion is beyond the scope of this article.

Image jitter can be averaged down by increasing the integration (exposure) time T; for relative astrometry (i.e., for simultaneous observations of two objects in the same instrument field), astrometric error caused by jitter is proportional to $T^{-1/2}$ [10,11]. Equations relating time T, angular accuracy (considering only atmospheric effects), and angular separation S were experimentally verified by Lindegren [10] and Han [11], using different data sets; their results were roughly comparable. Lindegren's theoretical expressions indicate a significant advantage for very small values of S, but there are no observational results to check this.

Figure 3 displays curves of integration time T versus required angular accuracy σ, based on Han's results. Each curve is for a different angular separation S between observed sources. These results are for stare mode; there are no comparable results for scan mode. Results at $S > 0.5$ deg are extrapolated and may be increasingly in error at larger separations. As discussed, Fig. 3 may be too pessimistic for $S \leq 0.01$ to 0.02 deg.

As can be seen, T increases significantly for larger values of S or smaller values of σ. For example, if $S = 5.0$ deg and $\sigma = 25$ nrad, then $T \approx 2$ to 3 hours. Integration times of a few hours are tolerable for target-body observations, but, for certain time-critical observations of laser-bearing spacecraft, it may be necessary to reduce the effective integration time by combining observations from many sites and instruments. For spacecraft observations, this would probably be necessary in any case for weather-related reasons.

For very small angular separations $S \leq 0.01$ deg, such as might exist between two laser-emitting spacecraft at Mars, the curves in Fig. 3 predict that 5-nrad astrometric accuracy can be achieved within 1 hr or less. Since most of the nonatmospheric contributors to the astrometric error budget will also decrease for small separations, there could eventually be very accurate (few nrad or better) optical navigation using multispacecraft observations. In terms of navigation capability, this could possibly be analogous to recently developed "single-beam" radio interferometry techniques [12]. Further discussion is beyond the scope of this article.

G. Instrument Development Strategy

Assuming that an accurate star catalog will be available by the mid-1990s, one may ask, "What is an appropriate instrument development strategy to support accurate observations of moving bodies?" Several possible themes can be described. First, as improved catalogs become available, the complete set of astrometric errors would be calibrated; this may not an easy task with some of these calibrations. Instrument design would have to provide the capability for adequate minimization or calibration of these effects.

Second, improved methods of star-catalog construction and/or densification would be developed. The developments for densification will trade off to some extent with efforts to widen the fields for narrow-field instruments, since wider fields require less catalog densification. The optimum trade-off can be found only through actual observational system development and testing. A wider field is desirable for another reason, namely to directly observe laser-bearing spacecraft relative to target bodies.

Finally, at least for near-term purposes, methods of determining scale and orientation without an accurate catalog could be tested; this is currently underway for intersatellite observations [2]. This may provide some insurance against possible problems with catalog densification and would be applicable to observations of laser-bearing spacecraft relative to target bodies.

V. Filled-Aperture Instruments

Two possible filled-aperture observing techniques are listed at the bottom of Fig. 1. As discussed, stare-mode operation moves the telescope in angle to keep stars fixed

in the field of view. In the scanning mode, the telescope is moved at a fixed angular rate so that the images move across the detector. For some systems, image motion in right ascension is achieved by turning off the telescope drive, but declination is not changed.

Image motion (smear) is a problem that must eventually be addressed for all these systems, since stars and moving bodies have different angular rates. Astrometric error induced by image motion is typically much less than the image motion itself, because the image motion is essentially symmetric about the center of the image. As a rough rule, the resulting astrometric errors can be held at acceptable levels if the image motion is less than one point-spread radius (i.e., roughly 0.5 to 1.0 arcsec for ground-based observing).

There are several potentially viable ways to obtain accurate astrometry in the presence of smear. For CCDs, these typically involve implementing instrumental and procedural changes to limit image motion to less than 0.5 arcsec (for example, by expanding the point spread by dithering or defocussing the telescope [13], by obtaining many short exposures, or by developing new data reduction and analysis methods to process smeared images). Further discussion of image smear for CCDs is beyond the scope of this article. As will be discussed, for Ronchi instruments the best approach appears to be movable photometers, under automatic computer control.

A. Stare-Mode Systems

1. Wide-field Instruments. A stare-mode instrument with a wide field (for example, a 5-deg field) would make it possible to avoid the requirement for Hipparcos catalog densification, to observe enough Hipparcos stars to "average down" the proper motion error, and to simultaneously observe a laser-bearing spacecraft and a target body, at desirably large encounter distances. Also, this instrument could produce a densified local star catalog for narrow-field use.

As will be discussed, it appears that suitable wide-field optical telescopes can be found, but a major problem arises in finding suitable visible light detectors to cover a wide field for stare-mode astrometry of moving bodies. A narrow-field instrument would avoid many of these detector difficulties, but would probably require some type of star-catalog densification, possibly with scan-mode observations of star fields.

Although the tentative goal is a 5-deg field, high-accuracy 25-nrad astrometry has never been accomplished over even a 1-deg field. To assess whether there could be an optical telescope capable of achieving such accuracy for a 5-deg field (again, assuming a perfect detector), Owen and Shaklan [14] performed a ray and wave trace analysis of a wide-field astrograph, based on an optical prescription for a USNO telescope that was designed for accurate astrometry, but not constructed.

This 36-cm refractor (a multielement lens system) is shown in Fig. 4. Owen and Shaklan's analysis has shown that this optical system is capable of producing an essentially symmetric point-spread function (analogous to a beam pattern in the radio) over the whole 5-deg field. An asymmetric point-spread function can cause significant amplitude-dependent image-centroid errors, which are difficult to calibrate. In addition, it was found that the position shifts due to optical system aberrations (distortions in the mapping from the sky to the detector) can be calibrated to the 5-nrad level. These aberrations were found to be essentially temperature invariant.

Assuming a perfect detector, a few-hour integration time, and (for dispersion calibrations) a narrow 0.02-μm filter, this telescope is large enough to allow accurate observation of 13th-magnitude objects, such as a small asteroid, a spacecraft at Saturn, or faint stars for a densified catalog. It might be possible to design an astrograph with a slightly larger aperture (say 0.5 m). However, adequate wide-field astrometry with low-efficiency detectors probably would require an even larger telescope. If this becomes necessary, the best choices appear to be reflectors with a modified Schmidt or Ritchey-Chrétien design [15,16,17]. However, these designs lead to much more severe aberration problems than those for the 36-cm wide-field astrograph, and this probably would significantly increase the difficulty of telescope design and construction.

Narrow-field Instruments. As discussed in Section V.C., the size, expense, and technological difficulty of making wide-field detectors forces consideration of narrow fields, perhaps as small as 0.5 deg. With a narrower field, the number of CCD chips would be dramatically reduced and the aberration problem incurred in using a larger telescope aperture for a Ronchi detector would also be reduced.

In this case, the best alternatives are to use a densified catalog or to acquire observations during those time-critical opportunities when the moving body can be imaged with two or more stars from the Hipparcos catalog, which will probably be the only accurate global star catalog with sufficient density for this time-critical option.

B. Scan-Mode Systems

An approach to reducing the wide-field detector problem while at the same time achieving fields of many degrees employs a telescope which can be slewed at a nonsidereal rate, as indicated in Fig. 5. In this case, the images of the observed objects are not fixed on the detector. In a stare-mode CCD-based observation, the images are fixed to allow signal integration in a few pixels before the image is electronically read out. In the scanning mode with a CCD detector, the telescope slew rate is set to allow the images to move across the CCD at the same rate at which the signal is transferred between CCD pixels during a normal readout. In effect, this allows the signal to be built up during readout. The sky is scanned with a relatively small array of CCDs. High-accuracy reference star positions, shown as images labeled with an H in Fig. 5, can then be used to accurately refer one scan position to another across a wider field than that covered instantaneously by the CCD array.

Currently, reproducibility for these systems is at the 200- to 500-nrad level [18,19]. Monet [3, p. 432] indicates that the ultimate performance for the USNO 0.2-m CCD scanning instrument could be as good as 50 nrad, and, of course, larger telescopes and larger format CCDs could potentially provide significant performance improvements.

CCD scanning instruments typically observe each object for only a few minutes per night, so that adequate averaging of the image jitter caused by the atmosphere may require many (perhaps 10) nights. This mode of operation is well suited to star-catalog densification, but may not be suitable for time-critical observations of moving bodies, particularly of laser-bearing spacecraft.

C. Detector Considerations

Photographic detectors are commonly used for both wide- and narrow-field applications, but these detectors do not meet the accuracy requirements [2] and will not be discussed further in this article.

The two major classes of modern detectors for optical filled-aperture astrometry are charge-coupled devices (CCDs) and Ronchi rulings coupled with photometers. As discussed, a CCD is a silicon-based two-dimensional array of "pixels," that is, an array of small (10 or 20 μm) individual detectors which very efficiently convert visible light photons to countable electrons.

A Ronchi ruling is essentially a mask with alternating opaque and transparent parallel bars. Two or more photon detectors (photometers), usually photomultiplier tubes, are placed behind the ruling to measure the oscillations in the visible signals as the ruling is scanned across the objects in the astronomical field of interest. Figure 6 is a diagram of a Ronchi focal plane for observation of an asteroid and a single star (both shown in open circles, to represent photometers). Since the a priori position of observed objects is almost always known to better than one ruling line, the difference in modulation phase can be transformed into a differential angle between the asteroid and star.

The literature describing both of these astrometric devices is extensive and no attempt will be made to duplicate the full content of those descriptions. Night-to-night reproducibility of roughly 20 to 25 nrad has been achieved both for CCDs [20] and Ronchi devices [21,22].

1. CCDs. Because CCDs are an increasingly popular detector for astrometry and the technology is improving at an impressive rate, the present description will describe the technology trends, but not attempt a detailed prediction of its future progress. However, if current trends are extrapolated, significant future progress is likely both for CCD chip arraying and for fabrication of larger format devices.

For CCD instruments in stare mode, the difficulty for wide fields is primarily one of arraying large numbers of CCD chips. This can possibly be accomplished by either a brute force method ("tiling" the desired field with CCD chips) or by more selective techniques, such as placing a CCD chip under each observed object and accurately measuring the relative position and orientation of these chips.

Optimistically assuming CCD astrometric precision of about 1/100 pixel (slightly beyond the current state of the art [3]), then centroid accuracy of 15 nrad (to meet a total error budget of 25 nrad) requires a pixel angular dimension of 1500 nrad (about 0.3 arcsec) or smaller. The largest format CCD made today is a 4096 × 4096 chip [23], which could provide roughly a 0.34- × 0.34-deg field. Under these assumptions, about 225 chips would be required to tile a 5-deg field. Each chip currently costs many tens of thousands of dollars. Even with cost reductions over the next decade, the acquisition of 225 CCD chips would be a great expense. A narrower 1-deg field would require only about nine 4096 × 4096 chips.

Larger format CCDs require increasingly long readout times, which may be impractically long for extremely large formats. Chip formats of 8192 × 8192 pixels are potentially feasible, but these may take about 43 minutes to read out [23]. Therefore, at some point, it may be preferable to

array chips rather than increase the number of pixels per chip.

At present, verification of astrometric stability has not yet been demonstrated with even two chips! However, work is in progress at the USNO Flagstaff Station [20, pp. 663–664] and the University of Hawaii [24] to array several CCD chips.

Placing a chip under the focal plane image of each object would require movable chips whose positions and orientations must then be very accurately determined to about 0.1 μm, perhaps via a complex laser-metrology system. In either case, full tiling or movable CCDs, the next step in wide-field CCD detector development would be a major, and very expensive, effort.

However, for narrow fields, there are some very significant potential applications of CCD technology. For near-term development, the most cost-effective options appear to be large-format single CCD chips and eventually, with more capability and cost, an array of several chips. As discussed, these CCD systems are being developed by astronomers for their own purposes. If cooperative arrangements can be made with these observers, then technology development questions for moving-body astrometry can be addressed by actually acquiring and analyzing the appropriate ground-based observations. As discussed, JPL has an ongoing cooperative arrangement with the U. S. Naval Observatory Flagstaff Station [2].

2. Ronchi Ruling Devices. The wide-field Ronchi situation is depicted in Fig. 6 for a stare-mode instrument. For each object to be observed in the field, a photodetector must be positioned behind the ruling to record the modulation of the light caused by the ruling. Again, some implications of extending the field of these devices to a 5-deg field will be examined.

Although the ruling can be made to cover a 5-deg field, it must be very precisely ruled so that false frequencies are not artificially embedded in the data. For a typical wide-field telescope, with a 50-arcsec/mm scale value, a ruling precision of roughly 0.04 μm is required to achieve 10-nrad astrometric accuracy; this requirement primarily refers to long-period variations across the ruling.

Since this is approximately the ruling precision required for a proposed space mission, ground-based astrometry may eventually be able to take advantage of the high precision required for space. The required ruling precision has already been demonstrated for small (few-cm) gratings as part of a development effort for the Astronomical Imaging Telescope (AIT) [25]; work is continuing to extend this precision to meet flight requirements of 10 by 25 cm. Since the present ground-based application (5-deg field with 50-arcsec/mm scale) implies a grating size of roughly 36 cm in height and perhaps twice that in the direction of grating motion, the necessary grating precision probably is more difficult to obtain than in the AIT case.

Ronchi systems are inherently less efficient light collectors than CCD-based systems, in part because the half-opaque ruling throws half the light away, and in part because the photodetector probes are less efficient than CCDs (by a factor of 3 to 5). This inefficiency implies that a larger telescope must be used to capture more light or that dramatically longer integration times will be necessary than for a CCD-based system. Finally, the Ronchi device must be scanned separately in right ascension and declination to measure both sky coordinates, thereby requiring yet additional telescope resources.

All these considerations tend to drive Ronchi instrument design to larger telescope apertures. As discussed, this type of low-efficiency detector increases the technical difficulty for telescope design and fabrication.

As with CCDs, it may be possible to observe faint moving objects in a narrow-field instrument (either CCD or Ronchi) and then densify the star catalog with a wide-field instrument (again, Ronchi instrumentation is one of the choices). This would ease the requirements on the wide-field instrument by reducing the requirements for faint-object observations from about 16th magnitude to about 13th magnitude.

However, even for catalog densification, there are many other wide-field development problems. For catalog densification, it would be desirable to observe all the Hipparcos stars in the field (to reduce the proper-motion errors by averaging errors for many stars) and also to perform the densification by observing many fainter stars. The large number of probes required, 50 to 100 of them at tens of thousand of dollars each, guarantees that a wide-field Ronchi approach will be expensive as well. For observations of moving bodies, the probe or probes for the target objects must be movable during an observation to account for the object's motion relative to the stars. This technology would require a significant development.

Technology development of a narrow-field Ronchi instrument would be significantly easier than for the wide field, because the narrow field would require only a small subset of the probes required for a 5-deg field and would significantly ease the requirements for ruling precision.

VI. Interferometric Instruments

Interferometric navigation observables are possible in the visible and infrared (IR) regions as they are in the radio. The basic idea is the same. The extremely narrow field of an interferometer allows it to observe, in general, only one object at a time. In the radio regime, these objects are quasars and spacecraft. In the visible or IR, the objects would be stars, spacecraft, and some solar system bodies. However, observations of stars and solar system objects may be limited to nighttime hours, just as for filled-aperture astrometry.

In the radio regime, the position of the "fringe" is measured electronically as the value of a time shift in a cross-correlation device. In the visible or IR, the position of the fringe could be measured from the locations of the bright and dark bands of the sinusoidal interference pattern on a CCD or some other array detector.

Fringe ambiguity resolution for the optical or IR case will be complicated by the essentially monochromatic character of a spacecraft laser transmission. Although this problem can possibly be circumvented by using either multiple wavelengths or multiple baselines, there are potentially serious development difficulties. First, the use of multiple laser wavelengths increases the complexity of both the spacecraft and ground stations, and may not be suitable for the communications development. Second, use of multiple baselines for ambiguity resolution of a monochromatic signal is an untried concept, whose implementation may require major changes in hardware and data processing.

Solar system objects pose a special problem for interferometry because usually they are not point sources; typically these objects subtend angles equivalent to many fringe cycles. (As an example, a 10-m optical baseline will result in more than 100 fringes across Io.) As shown in Fig. 7, because the fringe patterns from many incoherent points on the object's surface overlap, the fringe contrast (visibility) will be extremely low on the detector and can become totally lost in the noise. Visible light interferometry can be used with only a few small solar system objects.[1]

Fringe spacing is given by wavelength divided by projected baseline length, so that increasing the wavelength from the visible (roughly 0.5 μm) to IR (about 10 μm) will decrease the number of fringes per body radius by about a factor of 20. Since this increases the pool of observable solar system objects by about the same factor, IR interferometric observations of solar system bodies, particularly in the 10-μm atmospheric window, may be found to be more generally useful for navigation than those in the visible.

If, as seems likely, the laser communications are at visible wavelengths, then an IR interferometer would not be able to observe the spacecraft. However, as discussed in the IR interferometry subsection, the instrument builders expect to achieve enough sensitivity for observation of a large set of target bodies.

As will be discussed, the present optical and IR interferometry instruments do not provide the sensitivity (and in the IR case, the astrometric capabilities) to observe either target bodies or laser-bearing spacecraft. New systems with significantly improved sensitivity and astrometric capabilities are being proposed by the astronomical community, and may be built during the 1990s, at a multimillion dollar cost for each system.

Filled-aperture instruments can also potentially achieve these goals, so there is no navigation requirement for construction and successful operation of an optical or IR interferometric system. However, from a systems viewpoint, it is important to understand the potential development challenges and observational capabilities of these systems, since they can still potentially play a role in a future navigational system.

A. Optical Interferometry

For navigation, optical interferometry is of some interest as a possible observing capability for angular observation of laser-bearing spacecraft relative to the star background and for construction of a sparse global catalog of bright stars.

Representative current and expected future optical interferometric capabilities will now be examined. The current best optical interferometric system at Mt. Wilson provides a reproducibility of about 50 nrad for repeated observations of 6th-magnitude stars over angles as large as 90 deg [26], but extensive new observations will be necessary to assess the astrometric accuracy.

The USNO optical interferometer, consisting of several telescopes (initially 0.5-m apertures, with possible future upgrade to 1-m apertures) [19,27], may become operational in the mid-1990s. This instrument may have the necessary sensitivity (13th magnitude or better for 0.5-m apertures) for useful astrometric experiments involving

[1] G. W. Null, "Astrometric Optical-Interferometry for Solar-System Bodies," JPL Interoffice Memorandum 314.5–1309 (Revised) (internal document), Jet Propulsion Laboratory, Pasadena, California, November 17, 1988.

small solar system objects, such as asteroids or the Martian satellites. The Infrared-Optical Telescope Array is a similar system being developed at Mt. Hopkins, Arizona, by a consortium of universities and research laboratories [28]. This system initially will operate with two 0.45-m telescopes with observations both at visible (0.8- to 1.0-μm) and IR (1.0- to 2.4-μm) wavelengths. The eventual goal is operation with several larger elements, perhaps the seven 1.8-m mirrors scheduled for removal from the Multi-Mirror Telescope (MMT).

As previously discussed, these systems could also potentially construct a sparse global catalog to 25-nrad or better accuracy, which then could be densified with suitable techniques (perhaps with a scanning CCD instrument).

B. Infrared Interferometry

For navigation, infrared interferometry could potentially provide accurate angular observations of most target bodies relative to bright stars from the Hipparcos star catalog or other suitable global catalogs.

The present infrared interferometry discussion is restricted here to small target bodies, such that β, the number of fringes per body radius, is less than 0.6 (i.e., on the main lobe). As shown by experience with microwave interferometry, this will potentially provide suitably high fringe visibility for accurate angular measurements.

For example, Muhleman et al. [29] obtained interferometric observations of Jupiter's Galilean satellites at microwave wavelengths (λ = 2 and 6 cm) at the range of β values above. These satellites have nearly spherical shapes but have significant albedo variations. Muhleman et al. were able to obtain 150-nrad accuracy for the satellite positions and their analysis indicated that effects of albedo variations were fairly minor.

Asteroids appear to be especially suitable objects for infrared observations. Although the solar radiation at the asteroid is much less at infrared than at visible wavelengths, there are compensating effects that ensure that the energy flux is roughly constant from the visible out to about 20 μm, as discussed by Lebofsky and Spencer [30]. Specifically, asteroids tend to be dark bodies (visible albedo < 0.3), and so the energy received in the visible band is mostly converted to heat and re-emitted in the infrared. Obviously, this is very favorable for IR observations of low-albedo solar system bodies.

M. Shao is collaborating with C. Townes, U. C. Berkeley, to update Townes' IR interferometer at Mt. Wilson by adding laser metrology and other improvements.[2] This array currently uses two 1.65-m telescopes. For the updated system, Shao indicates that a single night's observations could potentially reach S/N = 10 (sufficient to accurately measure fringes) for a 12.4 magnitude star at 2.2 μm, a 7.8-magnitude star at 5 μm, or a 3.7-magnitude star at 10 μm. These magnitudes (m) are given for spectral classification A0; by convention, the A0 magnitude of a given star is the same at all wavelengths. Eventual IR astrometric accuracy is difficult to predict, but might be comparable to that at visible wavelengths, i.e., 25 nrad or better. Plans call for operation of a prototype two-element version of this IR system by the mid-1990s.

Figure 8 provides a rough plot of energy flux above the Earth's atmosphere versus IR wavelength. Flux values are shown as dashed lines for asteroids [30] and as black lines for A0 stars [31]; Shao's S/N = 10 sensitivity limits expressed in flux units are marked by large ×'s. The plotted asteroid fluxes have not been reduced to compensate for finite-body visibility losses, which are insignificant for most portions of these flux curves. The single exception is for the 2.2-μm window for a 100-km asteroid at 5 AU; for this case, $\beta \approx 0.14$ and the squared visibility is slightly reduced to roughly 0.8. As can be seen, asteroid flux is relatively constant as wavelength increases, but the stellar flux decreases significantly. However, since IR interferometry can be performed at large angular separations, bright stars from a sparse catalog are acceptable background sources.

As shown, 10-km asteroids at 1 AU would be marginally observable at 2.2 μm and easily observable at 5 and 10 μm. A more distant asteroid (or low-albedo satellite) at 5 AU from the Sun with radius R = 100 km would have a signal-to-noise ratio of almost 10 (marginally observable) at 2.2 and 10 μm. Some of Jupiter's Galilean satellites may also be observable at 10 μm, since, for a 2.5-m baseline, $\beta \approx 0.50$ to 0.81. Their large radii (roughly 1550 to 2650 km) may provide enough flux to sufficiently compensate for fringe visibility losses and for reductions in IR emissions caused by higher visible albedos. If not, then accurate Galilean satellite observations would require improved sensitivity and/or longer wavelengths ($\geq$20 μm).

VII. Summary and Conclusions

This article has examined both the long-term and near-term development prospects for the astrometry of stars

[2] M. Shao, personal communication, Optical Sciences and Applications Section, Jet Propulsion Laboratory, Pasadena, California, August 1991.

and moving bodies. For the near-term, large-format CCD detectors have been identified as the most cost-effective means to obtain early, useful narrow-field observations of solar system bodies. CCD technology is being actively developed by the astronomical community for their own purposes, and there are strong, continuing trends toward larger format devices and chip arraying. Therefore, collaborative arrangements with astronomers can potentially provide accurate observations of moving bodies, which can provide valuable information for navigation observing system design. Of course, CCD technology will also be important for long-term navigation development, both for conventional spacecraft and those with laser transponders.

One near-term activity, currently in progress, involves a cooperative effort with the U. S. Naval Observatory Flagstaff Station. The Flagstaff instrument, a 1.55-m astrometric reflector with a 2048 × 2048 CCD chip, provides an 11-arcmin field that is large enough for efficient observation of Jupiter's Galilean satellites. Details of this effort and its potential navigation benefits for the Galileo project are provided in [2].

For long-term purposes, the primary goal has been to identify important development trends and their potential applications for accurate astrometry of spacecraft and target bodies. What conclusions can be drawn from the material just presented? First, no single instrument is likely to be superior for all the different observational requirements (such as catalog densification, observation of very faint 16th- to 17th-magnitude objects, observation of moving bodies, and time-critical observations of spacecraft). This suggests that the best strategy is to employ several instruments for complementary purposes.

Second, although some of these instruments have provided reproducibility consistent with our 25-nrad navigation accuracy goal, none has demonstrated suitable accuracy. Typical demonstrated observational accuracy is roughly 10 to 20 times worse, primarily because of errors in available global star catalogs.

For the near-term navigation observation program, the focus has been on intersatellite observations, since accurate astrometry with these observations may not require an accurate catalog. However, star-relative observations will always require accurate catalog positions.

Fortunately, the ESA/Hipparcos Earth-orbiting observatory (now in orbit since November 1989) is expected to provide an accurate global catalog of roughly 2.5 stars per square deg by the mid-1990s, and ground-based scanning CCD instruments may be able to accurately densify this catalog or possibly to densify directly from optically observable radio sources. Thus, by the end of this decade, it may be possible to obtain an adequate observational verification of astrometric accuracy for each candidate observational system.

For navigation development, the long-term task is to cooperate with instrument builders and observers to investigate, improve, and observationally verify the performance of these systems for navigation purposes, including moving body observations and catalog densification. As appropriate, there could also be instrument development of dedicated navigation instruments, possibly for use in the DSORA facility.

Turning now to specific instruments, what were some highlights described in this article? First, as just discussed, CCD development is giving promising results and is being energetically pursued by astronomers. Use of narrow-field CCD or Ronchi instruments with a densified star catalog is an attractive development option.

Second, wide-field (5-deg) astrometry with filled-aperture stare-mode instruments appears to require a major CCD or Ronchi detector development; this would probably require navigation support, since there are no other plans for such an instrument. Studying this capability is important, since it potentially would provide direct angular observations of laser-bearing spacecraft relative to target bodies and it could also densify a global catalog. CCD scanning instruments, now being upgraded by astronomers, could possibly perform these tasks, but these instruments may be less well suited to time-critical navigation observations.

Third, optical and IR interferometry instruments with suitable sensitivity and sufficient baselines to support moving-body astrometry appear to be major, high-cost developments that are probably 5 to 10 years away. Optical interferometry may provide accurate observations of laser-bearing spacecraft, if adequate methods of fringe-ambiguity resolution can be devised. However, these ambiguity-resolution methods present potentially serious development difficulties. IR interferometry could possibly observe many asteroids and satellites with the desired 25-nrad accuracy. However, neither of these systems will usually be capable of directly observing a spacecraft relative to a target body.

As with CCDs, optical and IR interferometry techniques are being actively developed by astronomers for scientific purposes, not for spacecraft navigation. However, at the appropriate times, it may be possible to arrange

cooperative arrangements to investigate the navigation-related development questions.

Finally, what is the answer to the bottom-line question, namely, "Can suitable observational systems be developed for 25-nrad astrometry of laser-bearing spacecraft, target bodies, and stars?" At this point, it is difficult to be sure of the answer, but the availability of many promising development paths and the progress being made by astrometric instrument builders gives credence to the prognosis that the desired capabilities will eventually be attained, possibly within the next 10 to 15 years. However, this will require navigation support and involvement, so that the particular problems for moving-body observations are adequately addressed, both to ensure timely development and to provide an adequate operational capability for mission navigation. Reliable acquisition of time-critical observations involving spacecraft and target bodies will probably require construction of DSN astrometric observing facilities.

Acknowledgments

The authors gratefully acknowledge helpful conversations about Ronchi instruments (G. Gatewood, J. Stein, and M. Castelaz at Allegheny Observatory, Pittsburgh, and A. Buffington, UCSD), about CCD instruments (D. Monet, U. S. Naval Observatory Flagstaff Station), and about optical and IR interferometry (M. Shao, JPL). They also thank J. Ulvestad of JPL for helpful suggestions about this article.

References

[1] J. R. Lesh, L. J. Deutsch, and W. J. Weber, "A Plan for the Development and Demonstration of Optical Communications for Deep Space," *TDA Progress Report 42-103*, vol. July–September 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 97–109, November 15, 1990.

[2] G. W. Null, W. M. Owen, Jr., and S. P. Synnott, "Deep Space Navigation Applications of Improved Ground-Based Optical Astrometry," *TDA Progress Report 42-110*, vol. April–June 1992, Jet Propulsion Laboratory, Pasadena, California, pp. 118–127, August 15, 1992.

[3] D. G. Monet, "Recent Advances in Optical Astrometry," *Ann. Rev. Astron. Astrophys.*, edited by G. Burbidge, D. Layzer, and J. G. Phillips, vol. 26, pp. 413–440, 1988.

[4] W. M. Folkner and M. H. Finger, "Preliminary Error Budget for an Optical Ranging System: Range, Range Rate, and Differenced Range Observables," *TDA Progress Report 42-101*, vol. January–March 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 121–135, May 15, 1990.

[5] G. W. Null, *Deep Space Target Location with Hubble Space Telescope and Hipparcos Data*, JPL Publication 88–4, Jet Propulsion Laboratory, Pasadena, California, February 15, 1988.

[6] H. Schwan, "The FK5: Present Status and Some Derived Results," in *Inertial Coordinates on the Sky*, edited by J. H. Lieske and V. K. Abalakin, Dordrecht: Kluwer Academic Publishers, pp. 371–381, 1990.

[7] M. A. C. Perryman, E. Høg, J. Kovalevsky, L. Lindegren, C. Turon, P. L. Bernacca, M. Crézé, F. Donati, M. Grenon, M. Grewing, F. van Leeuwen, H. van der Marel, C. A. Murray, R. S. Le Poole, and H. Schrijver, "In-orbit performance of the Hipparcos Astrometry Satellite," *Astron. Astrophys.*, vol. 258, no. 1, pp. 1–6, 1992.

[8] B. Edlen, "The Refractive Index of Air," *Metrologia*, vol. 2, pp. 71–80, 1966.

[9] R. C. Stone, "The Effect of Differential Color Refraction on Declinations Determined in Meridian Circle Programs," *Astron. Astrophys.*, vol. 138, pp. 275–284, 1984.

[10] L. Lindegren, "Atmospheric Limitations of Narrow-Field Optical Astrometry," *Astron. Astrophys.*, vol. 89, pp. 41–47, 1980.

[11] I. Han, "The Accuracy of Differential Astrometry Limited by the Atmosphere," *Astron. J.*, vol. 97, no. 2, pp. 607–610, February 1989.

[12] W. M. Folkner and J. S. Border, "Orbiter–Orbiter and Orbiter–Lander Tracking Using Same-Beam Interferometry," *TDA Progress Report 42-109*, vol. January–March 1992, Jet Propulsion Laboratory, Pasadena, California, pp. 74–86, May 15, 1992.

[13] A. Buffington, C. H. Booth, and H. S. Hudson, "Using Image Area to Control CCD Systematic Errors in Spaceborne Photometric and Astrometric Time-Series Measurements," *Publ. Astron. Soc. Pacific*, vol. 103, pp. 685–693, July 1991.

[14] W. M. Owen, Jr., and S. B. Shaklan, "Geometric Distortion Analysis of a Wide-Field Astrograph," *TDA Progress Report 42-109*, vol. January–March 1992, Jet Propulsion Laboratory, Pasadena, California, pp. 87–93, May 15, 1992.

[15] S. C. B. Gascoigne, "Recent Advances in Astronomical Optics," *Applied Optics*, vol. 12, no. 7, pp. 1419–1429, July 1973.

[16] I. H. Bowen and A. H. Vaughn, Jr., "The Optical Design of the 40-in. Telescope and of the Irenée Dupont Telescope at Las Campañas Observatory, Chile," *Applied Optics*, vol. 12, no. 7, pp. 1430–1435, July 1973.

[17] D. J. Schroeder, *Astronomical Optics*, New York: Academic Press, Inc., 1987.

[18] G. F. Benedict, J. T. McGraw, T. R. Hess, M. G. M. Cawson, and M. J. Keane, "Relative Astrometry with the Steward CCD/Transit Instrument," *Astron. J.*, vol. 101, no. 1, pp. 279–289, January 1991.

[19] G. Westerhout and W. Donat III, "Annual Observatory Report for United States Naval Observatory," *Bull. Amer. Astron. Soc.*, vol. 24, no. 1, pp. 596–598, 1992.

[20] D. G. Monet et al., "U. S. Naval Observatory CCD Parallaxes of Faint Stars. 1. Program Description and First Results," *Astron. J.*, vol. 102, no. 2, pp. 638–665, February 1992.

[21] A. Buffington and M. R. Geller, "A Photoelectric Astrometric Telescope using a Ronchi Ruling," *Publ. Astron. Soc. Pacific*, vol. 102, pp. 200–211, February 1990.

[22] G. D. Gatewood, "The Multichannel Astrometric Photometer and Atmospheric Limitations in the Measurement of Relative Positions," *Astron J.*, vol. 94, no. 1, pp. 213–224, 1987.

[23] J. R. Janesick and S. T. Elliot, "History and Advancements of Large Area Array Scientific CCD Imagers," to be published in *Astronomical Society of Pacific Conference Series '91*, Tucson, Arizona, 1992.

[24] "Annual Observatory Report for University of Hawaii," *Bull. Amer. Astron. Soc.*, vol. 24, no. 1, pp. 220–222, 1992.

[25] S. Pravdo, *Astrometric Imaging Telescope 1991 Final Report*, JPL D–9651, Jet Propulsion Laboratory, Pasadena, California, May 1992.

[26] M. Shao, M. M. Colavita, B. E. Hines, J. L. Hershey, J. A. Hughes, D. J. Hutter, G. H. Kaplan, K. J. Johnston, D. Mozurkewich, R. S. Simon, and X. P. Pan, "Wide Angle Astrometry With the Mark III Stellar Interferometer," *Astron. J.*, vol. 100, no. 5, pp. 1701–1711, 1990.

[27] J. A. Hughes, G. H. Caplan, and M. Shao, "Design Considerations for USNO Astrometric Optical Interferometer," *Astrophys. Space Science*, vol. 177, pp. 151–159, 1991.

[28] R. D. Reasenberg, "IOTA Interferometer Project: Plans, Engineering, and Laboratory Results," in *Amplitude and Intensity Spatial Interferometry*, edited by J. B. Breckinridge, *Proc. SPIE*, vol. 1237, pp. 128–137, 1990.

[29] D. O. Muhleman, G. L. Berge, D. Rudy, and A. E. Niell, "Precise VLA Positions and Flux-Density Measurements of the Jupiter System," *Astron. J.*, vol. 92, no. 6, pp. 1428–1435, December 1986.

[30] L. A. Lebofsky and J. R. Spencer, "Radiometry and Thermal Modeling of Asteroids," in *Asteroids II*, edited by R. P. Binzel, T. Gehrels, and M. S. Matthews, Tucson: University of Arizona Press, pp. 128–147, 1989.

[31] C. W. Allen, *Astrophysical Quantities*, 3rd Edition, London: University of London, Athelone Press, p. 202, 1973.

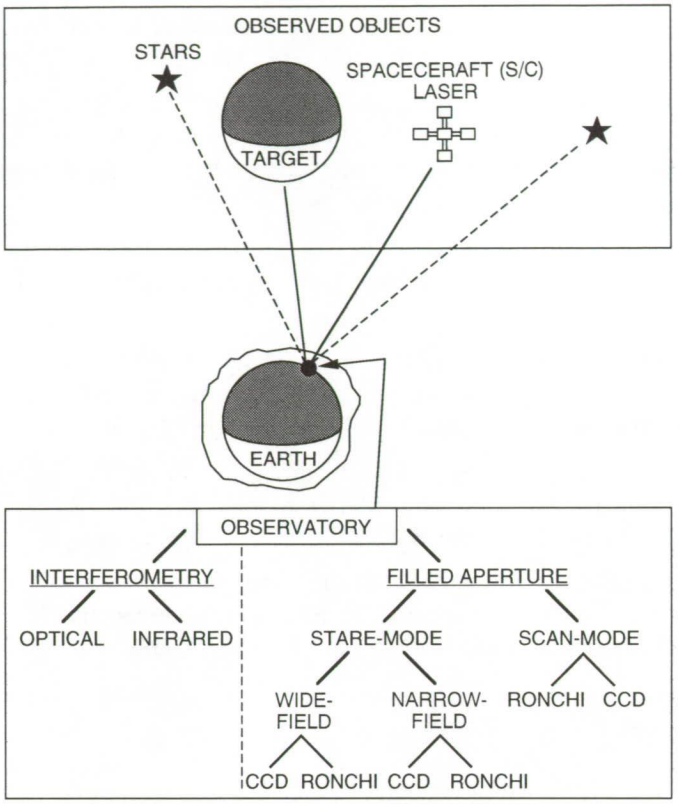


Fig. 1. Systems overview of ground-based optical astrometry.

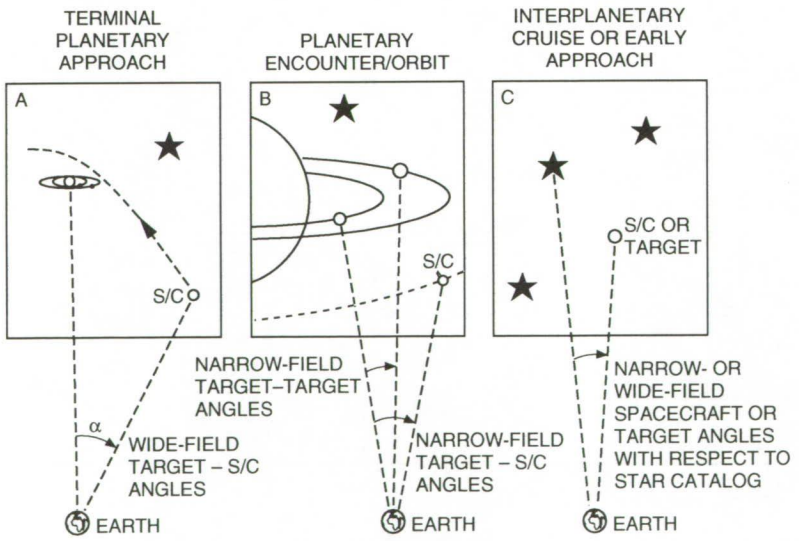


Fig. 2. Possible optical astrometric observations.

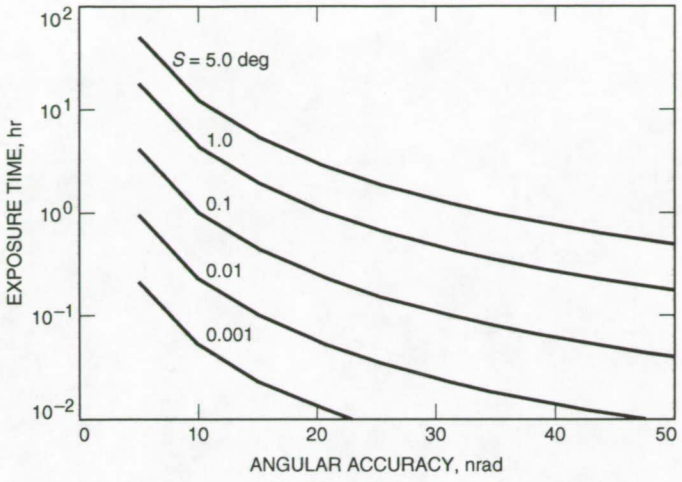


Fig. 3. Exposure time versus angular accuracy and angular separation. (Astrometric effect of atmospheric jitter based on Han [11].)

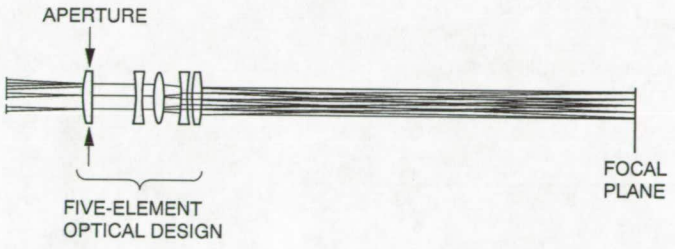


Fig. 4. Schematic diagram of a 5-deg wide-field astrograph.

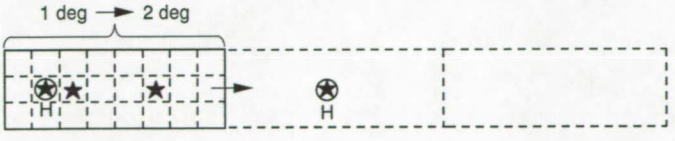


Fig. 5. Star observations with a scanning telescope and CCD detector. (Hipparcos stars are marked with an H.)

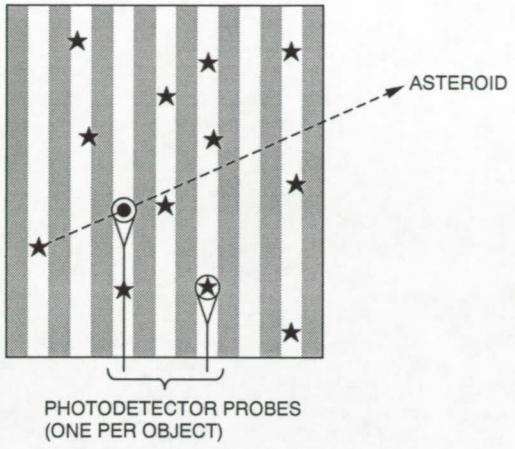


Fig. 6. Focal plane of Ronchi instrument (for stare mode). (Open circles represent photometers under asteroid and one star.)

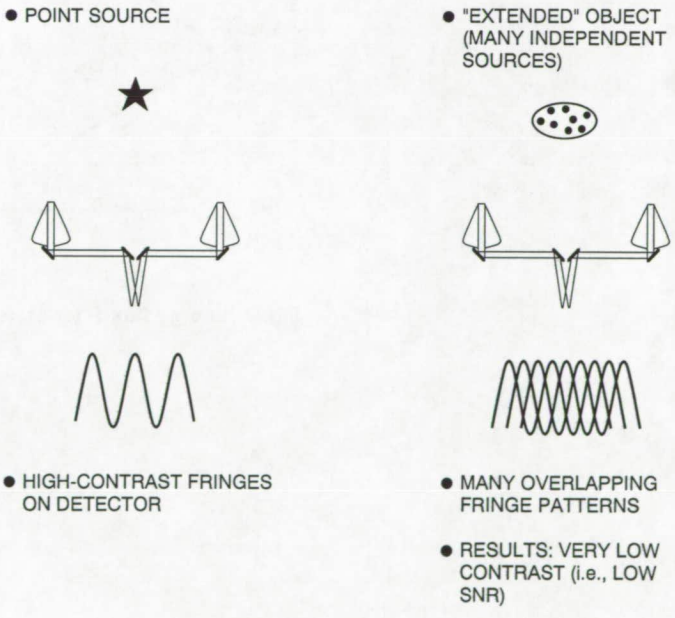


Fig. 7. Optical interferometry for solar system target objects.

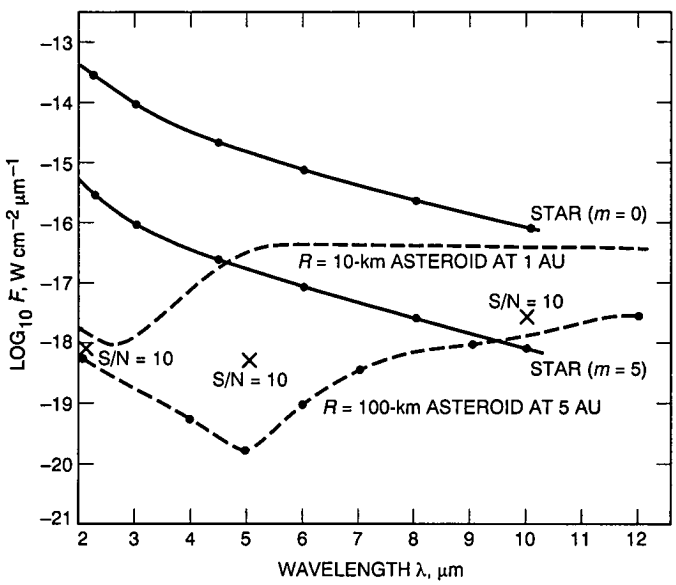


Fig. 8. Energy flux *F* for stars and asteroids versus wavelength.

Study of Narrow-Band Dichroic Plates With Circular, Rectangular, or Pyleguide Apertures

J. C. Chen
Ground Antennas and Facilities Engineering Section

The dichroic plate considered in this article is a metal plate perforated with arrays of apertures such that electromagnetic waves of certain frequencies pass through the plate and other frequencies are reflected. The shape of the apertures is an important contributor to the performance of the dichroic plate. An S/X dichroic plate, which passes X-band (8.4–8.45 GHz) and reflects S-band (2.0–2.32 GHz), is chosen for a case study. A comparison of the performance of dichroic plates of the same thickness and array pattern is presented for three different aperture shapes: circular, Pyleguide, and rectangular.

I. Introduction

Consider an S/X dichroic plate illuminated by a circularly polarized wave; this wave is formed as a combination of two orthogonal linearly polarized waves (TE and TM polarizations). If the apertures on the dichroic plate are symmetric, such as circular or square shapes, and the circularly polarized wave travels in the direction normal to the plate, then the relative phase difference of the two linearly polarized waves remains the same after the wave hits the plate. This means that the transmitted and the reflected waves are still circularly polarized. On the other hand, for oblique incidence, the relative phase difference between TE and TM polarizations will be changed after the wave leaves the plate, resulting in the transmitted and reflected waves being elliptically rather than circularly polarized (depolarization). Since oblique incidence is required in the Deep Space Network, an incident angle of 30 deg from the normal direction of the plate was chosen for this case study (Fig. 1).

The S/X dichroic plate on the DSS-14 70-m antenna was originally designed with circular apertures. Then the depolarization was improved by replacing the circular apertures with Pyleguide apertures by P. D. Potter in 1973 [1]. The results of the circular- and Pyleguide-aperture computations and measurements by Potter are discussed below. Furthermore, results of a theoretical design study of a dichroic plate with rectangular apertures are also included.

II. Circular Apertures

Circular apertures were originally used because they were easy to machine and the characteristics of the circu-

lar waveguide modes were well known. The first design of an S/X dichroic plate was a 3.576-cm-thick metal plate perforated with 2.273-cm-diameter circular apertures on a 60-deg skew grid with 2.388-cm spacing [Fig. 2(a)] [1]. A depolarization problem was found in this dichroic plate design. Calculations, using the C. C. Chen Holey Plate Computer Program, showed that the plate resonant frequencies were 8.481 and 8.363 GHz for TE and TM polarizations, respectively. A differential phase shift of 11.3 deg and an ellipticity of 1.75 dB were computed at the operating frequency of 8.415 GHz. The actual measured ellipticity was 1.84 dB. The reflected energy level was measured to be −18.6 dB at 8.415 GHz.

III. Pyleguide Apertures

A Pyleguide, a circular waveguide with a pair of flats on opposite sides, was introduced to solve the depolarization problem. Because one dimension is shorter than the other for a Pyleguide aperture, the cutoff frequencies are different for TE and TM polarizations. Since there was no computer program to analyze a dichroic plate with Pyleguide apertures, the size of the apertures was determined from the cutoff wavelengths of the Pyleguides, which corresponded to the resonant frequencies (8.481 GHz and 8.363 GHz) of the original circular-aperture plate. The Pyleguide aperture was 0.013 cm larger than the original circular aperture, and the flat depth was 0.043 cm [Fig. 2(b)]. The Pyleguide apertures were only a small change from the original circular apertures. The Pyleguide holes were made with a tolerance of ±0.005 cm. The phase shift was 3 deg and the ellipticity was 0.4 dB. The reflected power average level for the Pyleguide dichroic plate was −24.5 dB at 8.415 GHz [1]. The performance was improved by using Pyleguide rather than circular apertures.

IV. Rectangular Apertures

A dichroic plate with rectangular apertures has enough degrees of freedom to minimize the depolarization of the circularly polarized incident wave. A computer program based on a modal matching method was written to calculate the transmission and reflection coefficients of a dichroic plate with rectangular apertures [2]. The experiments on the test dichroic plates showed that calculated and measured resonant frequencies of the dichroic plate were within 0.3 percent, which indicates good accuracy of the software. The study of the dichroic plate with rectangular apertures in this article was based on this computer program, while the dichroic plates with circular or Pyleguide apertures discussed above were built and tested by Potter [1]. An optimization program was integrated with the rectangular-aperture dichroic plate program to accelerate the time-consuming process of optimizing the size of the rectangular apertures.

The optimization program finds the minimum of a user-defined cost function of several variables. Each variable is constrained between an upper bound and a lower bound. The function is first evaluated at the initial values of variables given by the users. Then the function is recalculated at different values of the variables. The program continues to search for a new set of variables until the local minimum of the function is found or until any one of the variable boundary constraints is broken. In this case, the variables were the dimensions of the apertures and the cost function was a function of the reflection loss of both the TE and TM polarizations, and the relative phase shift between them. Since several local minima may exit, the optimization process should be repeated with different initial values in order to find the absolute minimum.

The dimensions of the rectangular apertures were determined using the following method. First, the boundary constraints for the dimensions of rectangular apertures A_x and A_y were found. The aperture should be larger than the cutoff size for the lowest frequency of the passband, and smaller than the unit cell:

$$\frac{c}{2f_l} < A_x \leq D_x - w_{\min}$$

$$\frac{c}{2f_l} < A_y \leq D_{y'} \sin\Omega - w_{\min}$$

where c is the speed of light, f_l is the lower frequency of the passband, and $w_{\min}$ is the minimum wall thickness of the apertures. The wall of the aperture has to be thick enough to support the plate. Ω is the skew angle, and D_x and $D_{y'}$ are the dimensions of the unit cell (Fig. 1).

Second, the aperture was set to be square, $A_x = A_y$, so only one variable was optimized. The program searched for the optimized square aperture $A_1 = A_x = A_y$.

Third, A_x was fixed to the optimized square aperture dimension A_1, and A_y was varied as the optimization variable. The optimized value of A_y was A_2.

Finally, using A_1 and A_2 as the initial values, both A_x and A_y were optimized (Fig. 3). The optimization program searched for only the local minimum; therefore it was necessary to repeat the procedure (starting from the second step) to assure that optimum aperture size was achieved.

A 1.925-cm by 1.951-cm rectangular aperture was found using the optimization program for a 3.576-cm-thick dichroic plate on a 60-deg skew grid with 2.388-cm spacing [Fig. 2(c)]. Figures 4 through 6 show the transmission coefficients for TE and TM polarizations, the relative phase shift between TE and TM polarizations, and the ellipticity from 7 to 9 GHz. The calculations show a reflection loss of 0.000 dB for TE polarization and 0.001 dB for TM polarization, a relative phase shift of 2.0 deg, and ellipticity of 0.3 dB at 8.415 GHz. The conductivity loss was estimated to be 0.01 dB based on a 3.576-cm-long, 1.925-cm by 1.951-cm rectangular waveguide.

V. Summary and Conclusion

The results of this case study show that the apertures of unequal dimensions (rectangular or Pyleguide) produced a better axial ratio than the symmetric apertures (circular) for the case of oblique incidence with a circularly polarized wave. The performance of a dichroic plate with rectangular apertures was easier to optimize than that with Pyleguide apertures since an accurate computer program for a dichroic plate with rectangular apertures was available. Table 1 shows insertion loss, relative phase shift, ellipticity, and maximum reflected power of the dichroic plate with circular, Pyleguide, and rectangular apertures at 8.415 GHz. The maximum reflected power level at 8.415 GHz for the dichroic plate with rectangular apertures is expected to be slightly better than −24.5 dB. The grating lobes depend on the array pattern of the dichroic plate. Since in this case study the array pattern remains the same, the grating lobes are the same for these three configurations.

The reflected power from the dichroic plate resulted in increased temperature in the antenna system. The maximum noise temperature contribution from ground radiation was 0.7 K at 8.415 GHz for the 70-m antenna. While in the beam-waveguide-antenna pedestal room (300 K for DSS 13), the maximum noise temperature contribution of the dichroic plate was 1.06 K. The horn radiation pattern is not a perfect plane wave. Part of the horn radiation strikes the plate at angles other than the design angle; as a consequence extra reflections are created which result in an increased noise temperature. Further study using a plane wave expansion of the horn pattern would allow the reflected power to be predicted theoretically.

References

[1] P. D. Potter, "Improved Dichroic Reflector Design for 64-m Antenna S- and X-Band Feed Systems," *The Deep Space Network Progress Report*, JPL Technical Report 32-1526, vol. XIX, Jet Propulsion Laboratory, Pasadena, California, pp. 55–62, February 15, 1974.

[2] J. C. Chen, "Analysis of a Thick Dichroic Plate with Rectangular Holes at Arbitrary Angle of Incidence," *TDA Progress Report 42-104*, vol. October–December 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 9–16, February 15, 1991.

[3] T. Y. Otoshi and M. M. Franco, "Dual Passband Dichroic Plate for X-band," *TDA Progress Report 42-94*, vol. April–June 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 110–134, August 15, 1988.

Table 1. The performance of S/X dichroic plates at 8.415 GHz.

Parameter	Aperture		
	Circular	Pyleguide	Rectangular
Insertion loss, dB			
TE	—	0.009[a]	0.010[b]
TM	—	0.004	0.011
Relative phase shift, deg	11.3[c]	3.0	2.0
Ellipticity, dB	1.84	0.4	0.3
Reflected power level,[d] dB (max)	−18.6	−25.4	—

[a] The insertion loss of the dichroic plate with a Pyleguide aperture is achieved by experiments [3].

[b] The calculated insertion loss of the dichroic plate with rectangular apertures includes the reflection loss and conductivity loss, but not the loss due to surface roughness.

[c] The relative phase shift and ellipticity of the dichroic plate with circular or Pyleguide apertures are determined experimentally [1].

[d] The reflected power level is achieved by experiments.

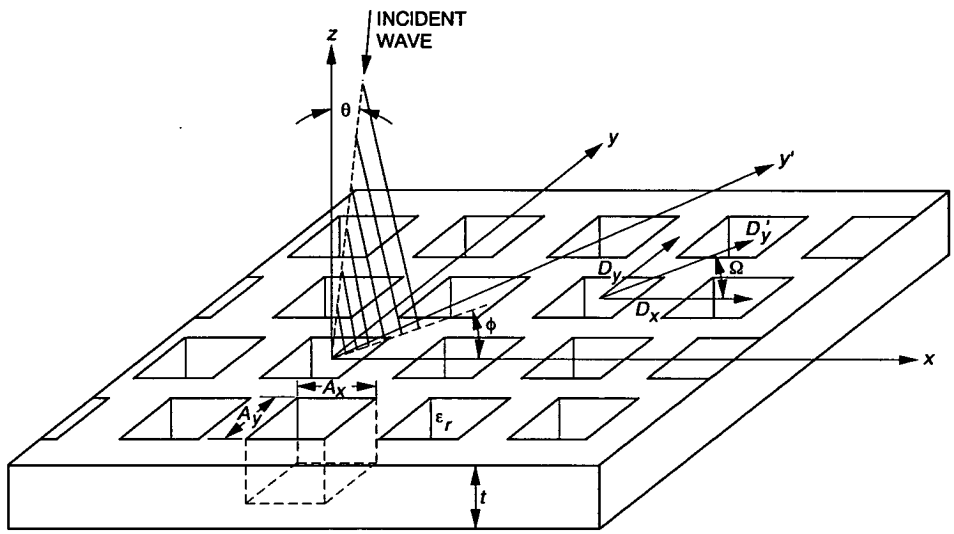


Fig. 1. Geometry of X/Ka dichroic plate.

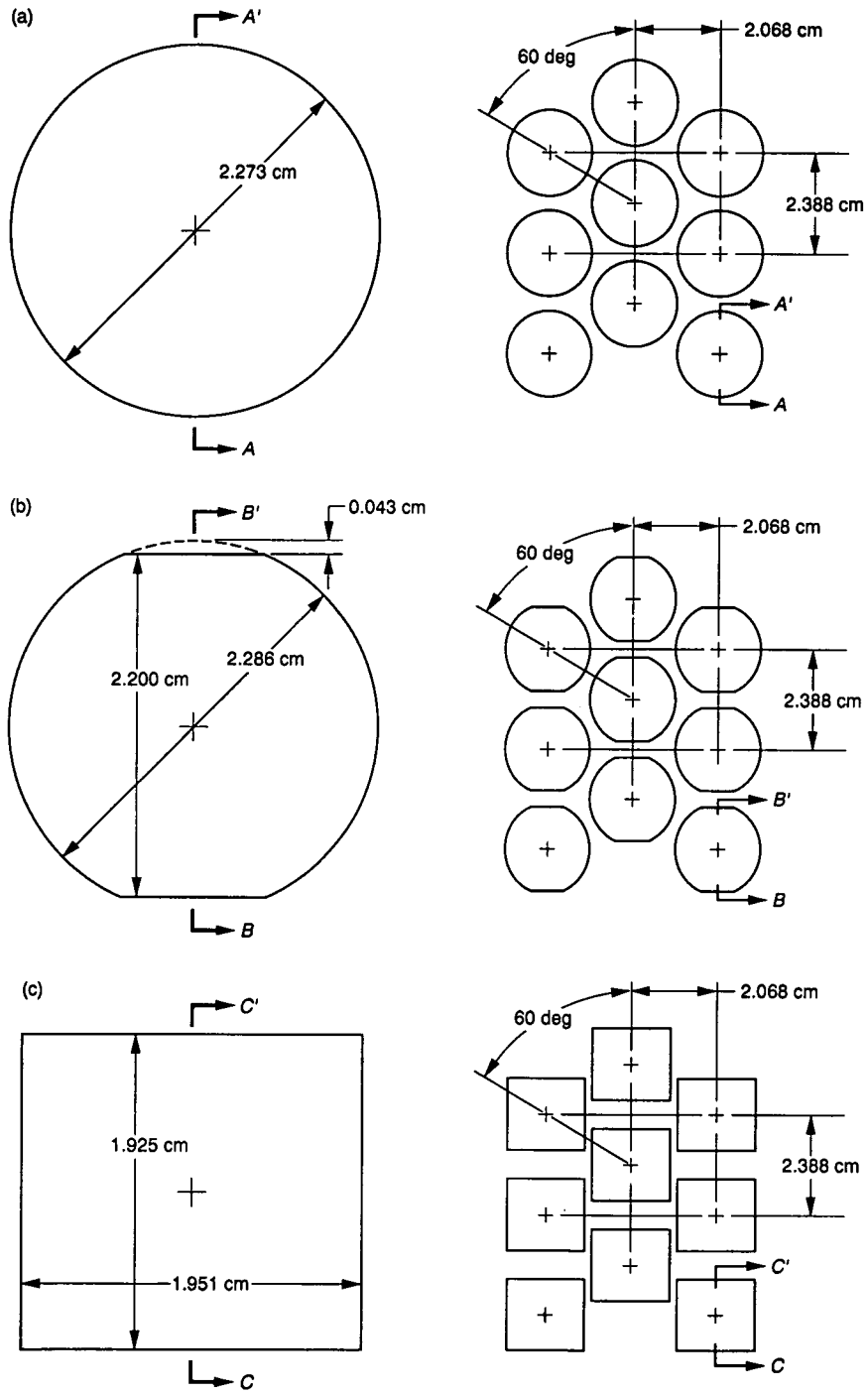


Fig. 2. Apertures: (a) circular; (b) Pyleguide; and (c) rectangular.

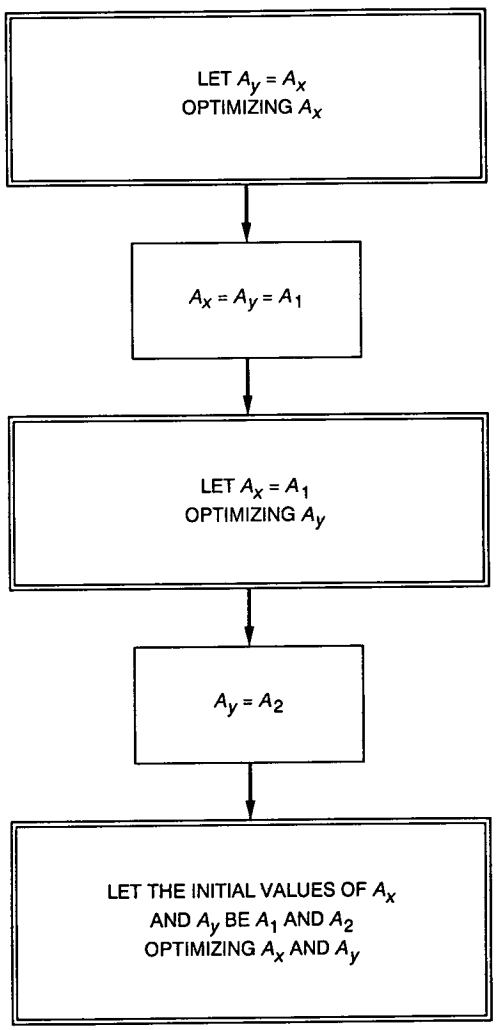


Fig. 3. Method for finding an optimized rectangular aperture for a dichroic plate.

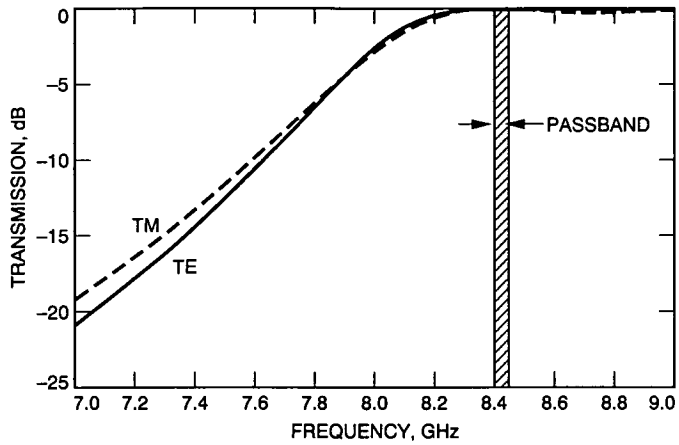


Fig. 4. Transmission coefficient versus frequency for X/S dichroic plate with rectangular apertures for TE and TM polarizations.

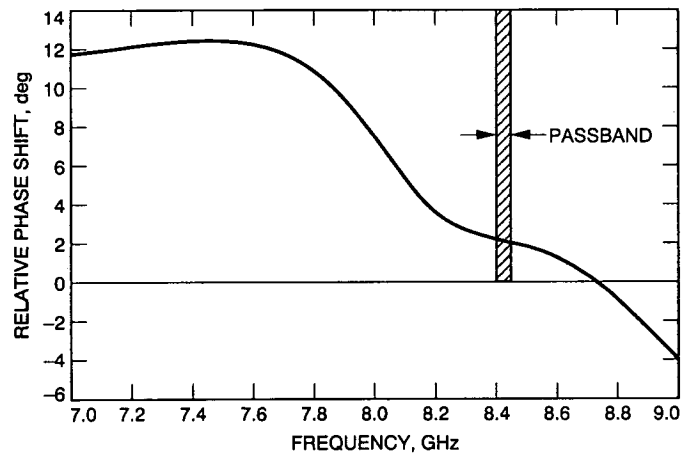


Fig. 5. Relative phase shift between TE and TM polarizations versus frequency for X/S dichroic plate with rectangular apertures.

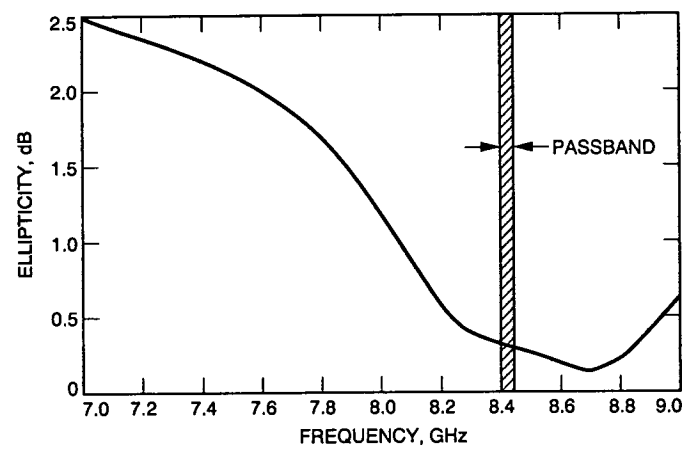


Fig. 6. Ellipticity versus frequency for X/S dichroic plate with rectangular apertures.

Analysis and Applications of a General Boresight Algorithm for the DSS-13 Beam Waveguide Antenna

L. S. Alvarez
Ground Antennas and Facilities Engineering Section

A general antenna beam boresight algorithm is presented. Equations for axial pointing error, peak received signal level, and antenna half-power beamwidth are given. A pointing error variance equation is derived that illustrates the dependence of the measurement estimation performance on the various algorithm inputs, including RF signal level uncertainty. Plots showing pointing error uncertainty as a function of algorithm inputs are presented. Insight gained from the performance analysis is discussed in terms of its application to the areas of antenna controller and receiver interfacing, pointing error compensation, and antenna calibrations. Current and planned applications of the boresight algorithm, including its role in the upcoming Ka-band downlink experiment (KABLE), are highlighted.

I. Introduction

Antenna beam boresigthing algorithms were developed in 1990 for gain and pointing calibrations carried out during the DSS-13 beam waveguide (BWG) Antenna Phase 1 Project[1]. The three- and five-point algorithms are currently implemented on a personal computer (PC) that interfaces with both the antenna radiometer systems and the antenna mechanical system to close an RF loop around the pointing system. At the present time, these algorithms are being moved to the antenna mechanical system as part of a planned DSS-13 controller upgrade that is to be a prototype for future DSN antenna controllers (e.g., the new DSS-24 BWG antenna). The boresighting program will obtain the input signal level from receivers over the station local area network (LAN). Specifically, system noise temperature estimates will be received from the total power radiometer subsystem and coherent SNR estimates will be obtained from the TP-13 and Advanced Receiver II (ARX) subsytem.

This article will present the general boresight algorithm. As part of the new implementation, an equation describing the propagation of signal level input measurement uncertainty into axial pointing errors was derived and is presented here. The signal level uncertainty can be the sum of numerous ground antenna sources as noise and/or errors from the antenna front-end electronics and measurement devices, antenna servo errors, as well as incident (coherent) signal dynamics due to spacecraft attitude controls and varying downlink modulation indices. It will be shown that the pointing error variance equation, also to be coded in the DSS-13 upgrade, allows prediction of bore-

[1] M. J. Britcliffe, L. S. Alvarez, D. A. Bathker, P. W. Cramer, T. Y. Otoshi, D. J. Rochblatt, B. L. Seidel, S. D. Slobin, S. R. Stewart, W. Veruttipong, and G. E. Wood, *DSS 13 Beam Waveguide Antenna Project: Phase 1 Final Report*, JPL D-8451 (internal document), Jet Propulsion Laboratory, Pasadena, California, May 15, 1991.

sighting performance based on the (assumed random and independent) RF SNR measurements and also on other system parameters (e.g., antenna half-power beamwidth, pointing error magnitudes, etc.). Plots that illustrate the dependence of the algorithm performance on the numerous input variables will be given. Insight gained from the performance analysis will be discussed in terms of antenna controller and receiver interfaces, pointing error compensation, and pointing calibrations. The article will also highlight some of the current and planned applications of the algorithm, including its utilization in the upcoming Ka-band downlink experiment (KABLE).

II. Boresight Algorithm

Three- and five-point boresight algorithms applicable to tracking extragalactic radio sources were developed and programmed in 1990 by R. Riggs. The algorithms map measured on- and off-source noise temperature measurements into estimates of axial (cross-elevation and elevation) pointing error, peak temperature, and antenna half-power beamwidth (full-width). The algorithms continue to be successfully applied at DSS 13 to boresight the BWG antenna, that is, maintain the peak of the antenna beam aligned with the target during gain and pointing calibration sessions. Details of the current PC-based implementation and usage can be found in Footnote 1 and [1,2]. The general algorithm will be presented below in such a manner as to not be dependent on any specific receiver for signal level input.

This particular boresighting scheme is actually a step scan that measures signal level at small angles off-boresight along a single axis at a time. A five- (or three-) point boresight implies that five (or three) measurements on-source along an axis are input into the algorithm. In the case of noncoherent signal inputs, the on-source measurements are obtained by using two off-source measurements to negate contribution of the baseline cold-sky system noise. The data points are then fit to a linearized exponential to yield desired pointing and peak received power estimates. Five- and three-point step scans will be implemented in the DSS-13 antenna pointing system, but in general any finite number (greater than two) of signal level measurements can comprise the input. The default offsets where measurements are taken are points that correspond to the 3-dB, 1-dB, and 0-dB points on both sides of the beam relative to the assumed true direction of the target. The three-point version has default offsets corresponding to measurements at the 1-dB (both sides) and peak-signal levels. Arbitrary offsets can be used but their magnitudes are bound by the validity of the Gaussian model used to approximate the antenna beam pattern.

Let the signal level inputs (in linear units) be $y_i, i = 1, \ldots, n$, sensed at the angular offsets $x_i, i = 1, \ldots, n$ superimposed on the predicted target angles, then the received signal level model is

$$y_i(x_i) = y_{peak} \exp\left(-\frac{4\ln(2)}{H^2}(x_i - \varepsilon)^2\right) \tag{1}$$

where y_{peak} is the peak signal level at boresight, H is the antenna (full-width) half-power beamwidth, and ε is the pointing error. These three variables are calculated by the algorithm. The solution involves solving for the coefficients c_1, c_2, and c_3 of the best-fit parabola satisfying the set of n equations

$$\ln\left(y_i(x_i)\right) = c_1 + c_2 x_i + c_3 x_i^2 \tag{2}$$

or

$$y_i(x_i) = \exp\left(c_1 + c_2 x_i + c_3 x_i^2\right) \tag{3}$$

By differentiating Eq. (3) with respect to x and equating to zero, the offset corresponding to peak signal level, that is, the pointing error ε, is determined to be

$$\varepsilon = -\frac{c_2}{2c_3} \tag{4}$$

Now, since y_{peak} is defined to be $y(\varepsilon)$ by Eq. (3), the peak signal level is found to be

$$y_{peak} = \exp\left(c_1 - \frac{c_2^2}{4c_3}\right) \tag{5}$$

Finally, by equating Eqs. (1) and (3), and setting the angular offset x arbitrarily to zero, the solution equation for the half-power beamwdith can be solved to be

$$H = \sqrt{-\frac{4\ln(2)}{c_3}} \tag{6}$$

In general, the coefficients c_1, c_2 and c_3 are calculated via the method of least squares from the set of n equations defined by Eq. (2). However, in the case where $n = 3$, the three coefficients can be solved for directly from the set of three equations obtained from Eq. (2).

III. Variance Propagation

The computed values of ε, y_{peak}, and H will have uncertainties σ_ε, $\sigma_{y_{peak}}$, and σ_H due to the measurement errors in the signal level inputs y_i, and perhaps due to significant antenna position errors in the offsets x_i. An equation describing the propagation of the uncertainties in the signal level inputs into axial pointing errors, assuming negligible position offset errors, will be presented below. If needed in the future, similar variance equations can be derived for the efficiency variables y_{peak} and H.

The measurement function in Eq. (2) is ln (y_i). Let $y_{\ln_i} = \ln(y_i)$, and let the uncertainty in the signal level inputs y_i be σ_{y_i}, then the measurement uncertainty in $y_{\ln_i}$ is

$$\sigma_{y_{\ln_i}} = \frac{\sigma_{y_i}}{y_i} \tag{7}$$

and will be propagated through the boresight algorithm. The σ_{y_i}'s are also presumed to encompass all uncertainty that was introduced in negating cold-sky background noise from the y_i. For example, in the classic two-point method of removing the background sky noise contribution in order to isolate the target signal levels, the resultant variance on y_i is increased as described in Appendix B.

Now, assuming negligible uncertainty in the x_i and that the $\sigma_{y_{\ln_i}}$ are random and independent, then the variance of the computed pointing error ε is

$$\sigma_\varepsilon^2 = \sum_{i=1}^{n} \left(\frac{\partial \varepsilon}{\partial y_{\ln_i}} \sigma_{y_{\ln_i}} \right)^2 \tag{8}$$

From the pointing error solution Eq. (4), the following partial derivative equation is obtained:

$$\frac{\partial \varepsilon}{\partial y_{\ln_i}} = \frac{1}{2c_3^2} \left(c_2 \frac{\partial c_3}{\partial y_{\ln_i}} - c_3 \frac{\partial c_2}{\partial y_{\ln_i}} \right) \tag{9}$$

which can be inserted into Eq. (8) to yield

$$\sigma_\varepsilon^2 = \left(\frac{1}{2c_3^2} \right)^2 \sum_{i=1}^{n} \left(c_2 \frac{\partial c_3}{\partial y_{\ln_i}} - c_3 \frac{\partial c_2}{\partial y_{\ln_i}} \right)^2 \sigma_{y_{\ln_i}}^2 \tag{10}$$

where c_1, c_2 and c_3 are the best fit coefficients from Eq. (2). From Eq. (6), c_3 can be written in terms of the half-power beamwidth H as

$$c_3 = -\frac{4 \ln (2)}{H^2} \tag{11}$$

and then substituting c_3 into the pointing error Eq. (4) yields

$$c_2 = (2\varepsilon) \frac{4 \ln (2)}{H^2} \tag{12}$$

Both of the above two equations along with the noise to signal Eq. (7) can now be inserted into Eq. (10) to write the pointing error variance in terms of the boresight algorithm input variables. Inserting and simplifying yield

$$\sigma_\varepsilon^2 = \left(\frac{H^2}{8 \ln (2)} \right)^2 \sum_{i=1}^{n} \left(2\varepsilon \frac{\partial c_3}{\partial y_{\ln_i}} + \frac{\partial c_2}{\partial y_{\ln_i}} \right)^2 \left(\frac{\sigma_{y_i}}{y_i} \right)^2 \tag{13}$$

Further expansion of the partial derivatives of c_2 and c_3 is given in Appendix A. As seen from Eq. (13) the accuracy of ε depends, in general, on a number of factors: the noise-to-signal ratio (NSR) of the target, the antenna half-power beamwidth H, and the magnitude of the actual pointing error ε being estimated. As illustrated in Appendix A, the magnitude of the axial offsets x_i also impacts σ_ε through the partial derivative equations of c_2 and c_3. Performance in terms of these parameters will be illustrated in the following section.

Two subtle effects that can also degrade estimation accuracy of ε are large, simultaneous pointing errors in each axis and inordinate signal integration time. Large errors in the axis perpendicular to the one where ε is being computed will increase σ_ε through a decrease in signal level y_i (which increases the NSR) in Eq. (13). This may or may not be a major concern, depending on the particular application of the boresight algorithm and the quality of the pointing calibration of the antenna. For example, gain and pointing calibration applications minimize this effect by continuously measuring and correcting ε sequentially in each axis.

Long signal integration times combined with the significant antenna servo move delays may result in axial step scan measurement periods on the order of minutes or longer. The sidereal motion of the antenna can then result in a smearing-type degradation on ε. This is especially true if the antenna is in a poor calibration state where pointing errors can quickly build up as a function of antenna orientation.

Careful interpretation should be applied to computed values of σ_ε during practical DSN tracking operations.

Specifically, the pointing error variance equation was derived assuming that the input noise to signal measurements σ_{y_i}/y_i are independent and random. This assumption does not hold, for example, when the received signal error sources include drifts and/or biases induced by incident signal dynamics from spacecraft attitude controls and varying downlink modulation indices.

IV. Performance Analysis and Applications

A. Performance

Performance of the boresight algorithm can be expressed in terms of the output pointing error standard deviation σ_ϵ with respect to the various inputs. In the following, the performance of the algorithm will be quantified with respect to a 34-m antenna at both X-band (8.45 GHz) and Ka-band (32 GHz) frequencies (i.e., the half-power beamwidths are 65 and 17 mdeg, respectively). An important statistic is the target NSR, or standard deviation σ_{yln_i} defined by Eq. (7). Figure 1 shows a plot of σ_ϵ versus σ_{yln_i} for both three- and five-point boresight algorithms. The uncertainties were calculated with offsets corresponding to a zero pointing error. The signal losses corresponding to both five-point and three-point algorithms are the same for X- and Ka-bands, and are shown in Fig. 2 for the zero pointing error case. As seen, irrespective of frequency, the five-point algorithm is less sensitive to input noise.

Estimation uncertainty in a single axis also increases with the magnitude of the pointing error ε. Figures 3 and 4 illustrate this effect for the five-point algorithm at the Ka- and X-band frequencies, respectively. As seen, σ_ϵ increases dramatically with ε at Ka-band, while with the same error magnitudes at X-band it is essentially insensitive. The antenna beam-pointing error vector is defined as the rss of the simultaneous errors sensed in the elevation and cross-elevation axes. Thus, the true performance of the algorithm is two-dimensional. Figure 5 shows how ε varies while estimating zero error with an increasing pointing error in the opposite axis. Here σ_y/y is chosen to be 0.03 for all five signal inputs. This is a typical NSR observed during low wind (less than 16.7 km/hr) radio source measurement periods at DSS 13. As seen, the Ka-band two-axes performance can be significantly degraded if large errors are present in the perpendicular axis. The X-band performance is basically insensitive to the cross-axis error due to the larger beamwidth. Figure 6 shows the cross-offset Ka-band effect on the five-point algorithm as a function of NSR for the zero pointing error case and Fig. 7 summarizes its two-axes Ka-band performance with equally increasing axial and cross-axis errors.

B. Antenna Controller and Receiver Interfaces

The antenna pointing system controller interfaces with radiometers and receivers. With respect to the implementation of the boresight algorithm, signal level integration times can be specified based on required pointing estimate accuracy, as shown in the plots of σ_ϵ against input noise to signal levels. The design can be based on worst-case expected receiver SNR, or the signal levels can be integrated (at the receiver or the antenna controller) with an adaptive integration time until a specified input σ_{yln} is achieved. Estimates of the signal level variances need to be processed simultaneously with the magnitude estimates for input into the boresight algorithm. The equations that process the measurement variance estimates from the DSS-13 total power radiometer and TP-13 and ARX II receiver subsystems are presented in Appendix B.

C. Pointing Error Compensation

The boresight algorithm is typically used solely for antenna calibration and alignment tasks. However, it will be utilized as a beam-pointing error compensation mechanism for the upcoming KABLE at DSS 13. The plan is to blind point the antenna at the Mars Observer Spacecraft and use three- or five-point boresights to periodically peak the received signal by negating any residual pointing error buildup. The goal is to have the antenna in the best possible pointing calibration state so as to maximize time between boresights and thus avoid the received signal losses associated with the off-peak axial measurements taken at the offsets x_i. The pointing error variances will allow real-time assessment of the computed axial corrections before their application.

During KABLE, the spacecraft will provide a simultaneous X-/Ka-band downlink. If the antenna X- and Ka-band beams are coaligned, it is possible to utilize either of the two signals for the pointing error compensation. With respect to the input NSR, Fig. 1 indicates that the X-band signal needs to be considerably stronger than the Ka-band in order to achieve the same pointing estimation performance. This is especially true with the expectation that DSS 13 will be in a highly calibrated state for KABLE, since Fig. 7 illustrates that the Ka-band performance is still superior when estimating small pointing errors.

D. Pointing Calibrations

The formal computation of a pointing error variance during pointing measurements is novel in the DSN. In addition to adding a real-time performance assessment capability during calibration tracks, the pointing uncertainties can be used as a measurement quality index in off-line

analysis and pointing error modeling. It is anticipated that a more accurate antenna pointing calibration can be obtained from a now-possible weighted least-squares estimate of the pointing model coefficients.

Figures 3 through 7 also give some pointing calibration insight for Ka-band applications. When possible, it is best to first calibrate the antenna (or a particular focal point) at a lower frequency (e.g., X-band) to take out the large pointing errors. As illustrated, the larger X-band 34-m antenna beamwidth yields an increased insensitivity of the boresight uncertainty with respect to pointing error magnitudes. Figure 7 suggests, without considering the actual SNR conditions, that calibrating with the 34-m Ka-band signal should commence when the pointing errors are brought down to the 5.0- to 6.0-mdeg level. For comparison, Figs. 8 and 9 show the boresight estimation performance of the 34- and 70-m antennas with a 0.0- and 5.0-mdeg error in both the measurement axis and cross-axis, respectively.

V. Summary

A general antenna beam boresight algorithm has been presented. Although three- and five-point algorithms were analyzed for the current DSS-13 antenna controller upgrade, the equations derived for pointing error, peak received signal and antenna half-power beamwidth are general. A variance equation, which maps signal level measurement uncertainty into axial pointing errors, was derived. It is also being implemented in the DSS-13 upgrade. This is the first time that pointing error measurement uncertainties will be formally computed in a DSN or DSN research antenna. Plots were presented that illustrate the dependence of the pointing error estimation performance on the numerous input variables. In general, for the 34-m antenna, the algorithm with Ka-band input will outperform the X-band input case with respect to input SNR, assuming that the errors being estimated, as well as cross-axis errors, are at or below 5.0 mdeg. The Ka-band performance is dramatically degraded as these pointing errors increase.

Current and planned applications of the boresight algorithm and pointing error variance equation were discussed. Signal level input variance equations specific to the DSS-13 antenna controller interface with the total power radiometer and TP-13 and ARX II receiver subsystems were given in Appendix B. The boresight performance analysis presented will aid in future antenna controller/receiver interface and pointing error compensation designs. The availability of pointing error measurement uncertainties will enhance off-line antenna performance analysis and calibration efforts. The planned application of the algorithm for pointing error compensation in the upcoming KABLE was also highlighted.

Acknowledgments

Christian Sanelli is acknowledged for his assistance in deriving the pointing error variance equation. Hemali Vyas, who is currently implementing the boresight algorithm in the DSS-13 antenna controller, provided many insightful discussions on a practical level. A previous uncertainty analysis of a six-point parabolic cross-scan algorithm by J. C. Breidenthal provided useful technical direction.

References

[1] S. D. Slobin, T. Y. Otoshi, M. J. Britcliffe, L. S. Alvarez, S. R. Stewart, and M. M. Franco, "Efficiencey Calibration of the DSS 13 34-Meter Beam Waveguide Antenna at 8.45 and 32 GHz," *TDA Progress Report 42-106*, vol. April–June 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 283–297, August 15, 1991.

[2] L. S. Alvarez, "Initial Pointing Calibrations for the DSS 13 34-Meter Beam-Waveguide Antenna," *TDA Progress Report 42-106*, vol. April–June 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 188–204, August 15, 1991.

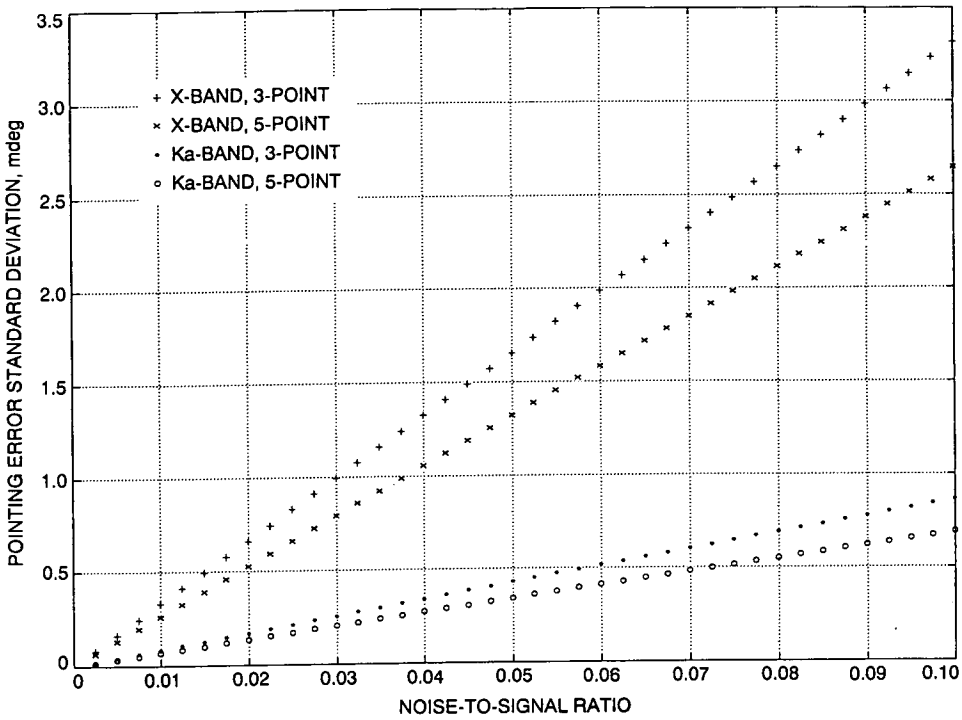


Fig. 1. Boresight uncertainty analysis for zero pointing error case.

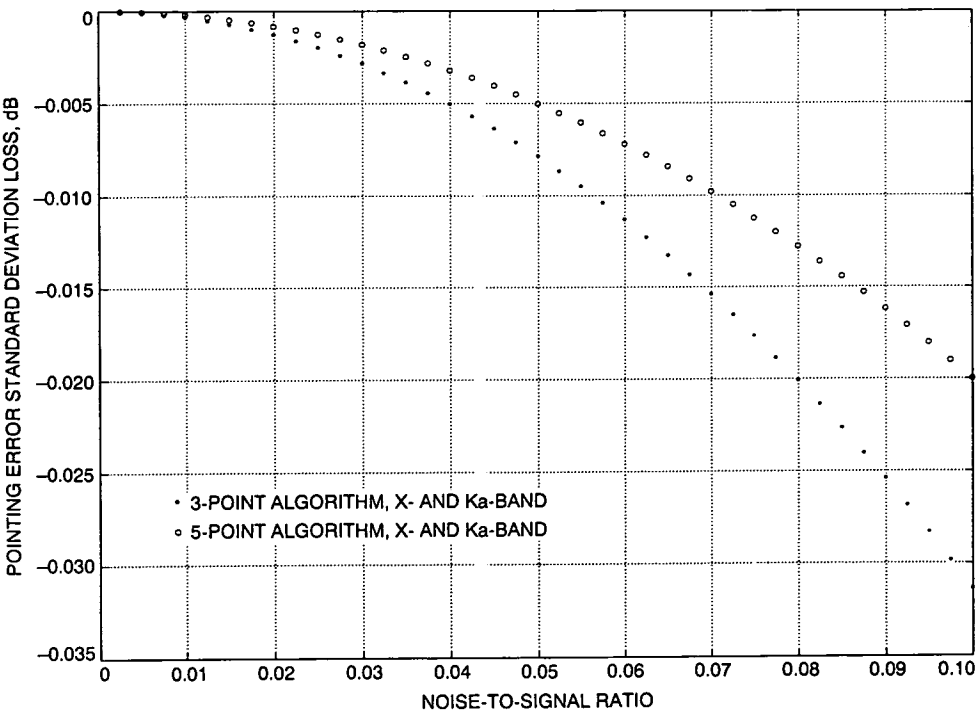


Fig. 2. Boresight error uncertainty loss for zero pointing error case.

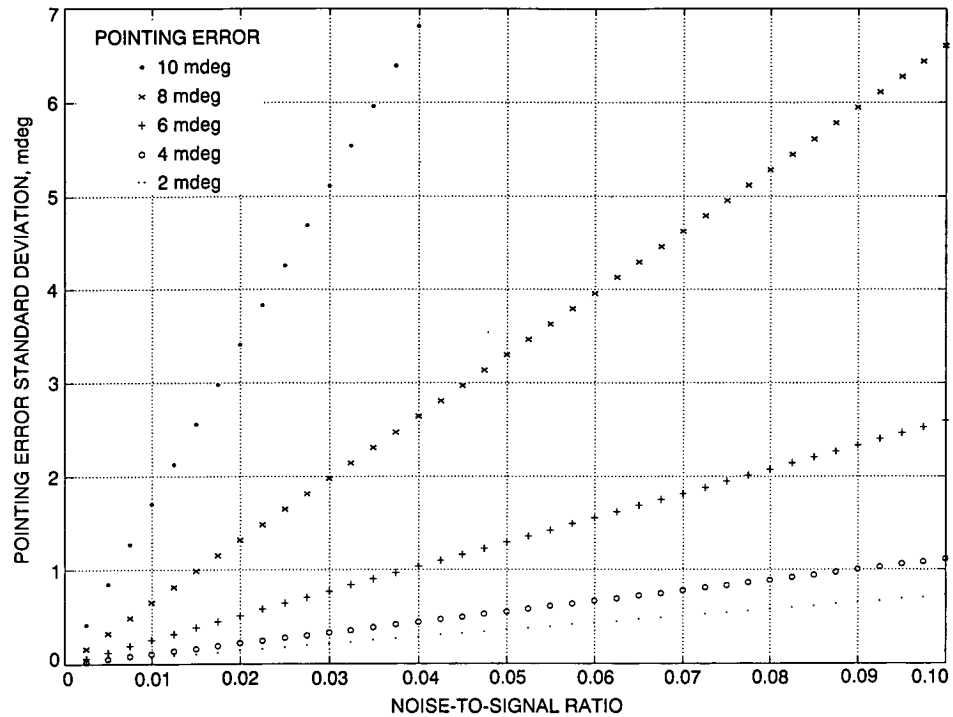


Fig. 3. Ka-band 5-point boresight uncertainty for increasing pointing errors.

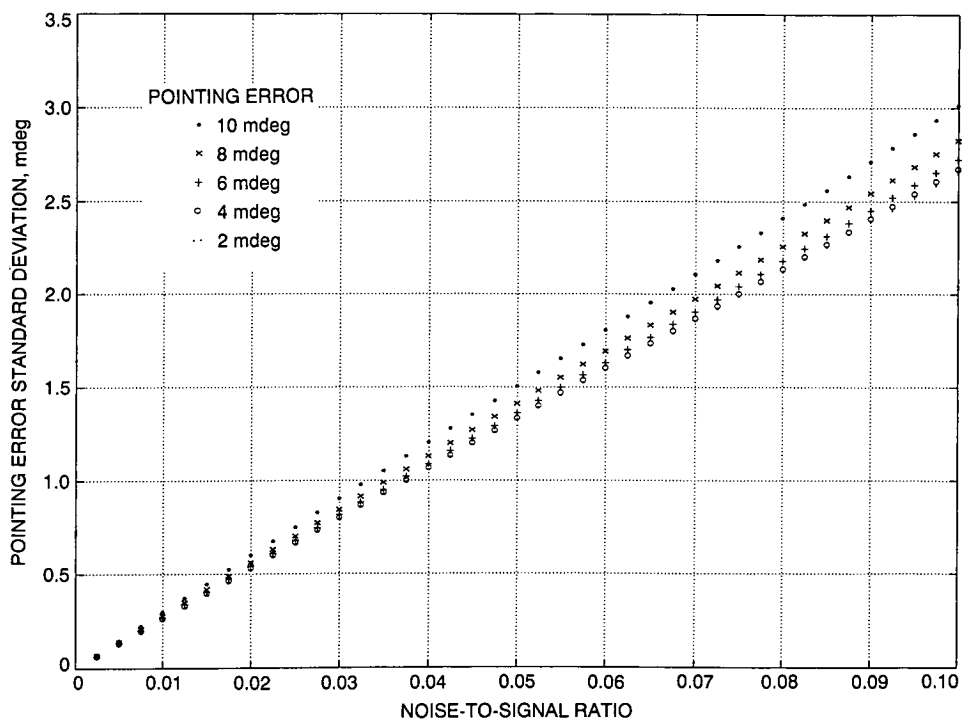


Fig. 4. X-Band 5-point boresight uncertainty for increasing pointing errors.

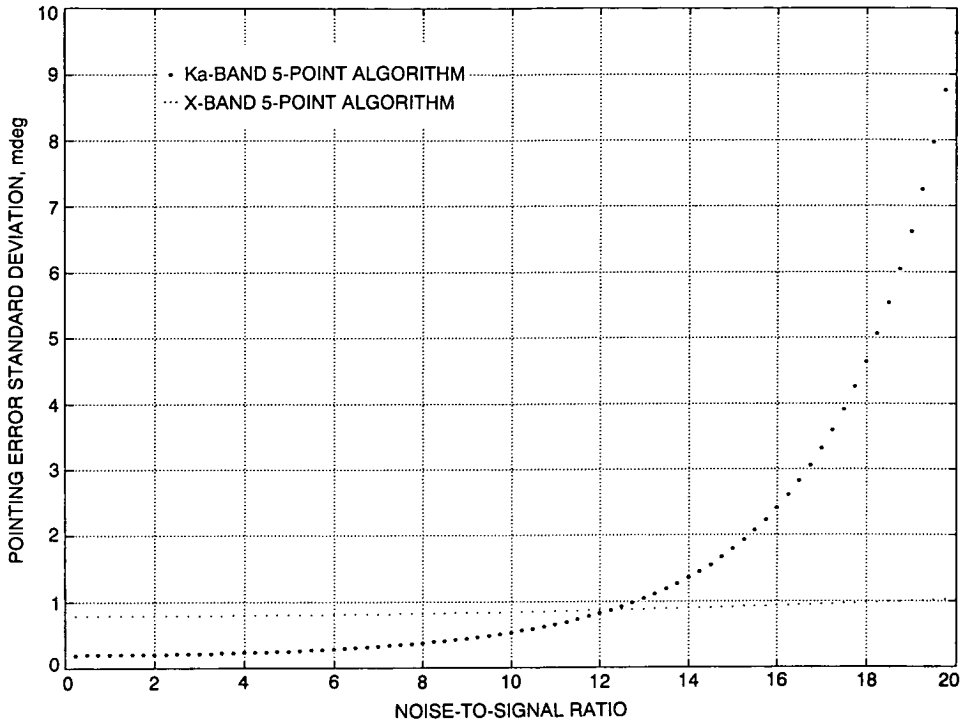


Fig. 5. Five-point boresight uncertainty for increasing cross-axis errors with NSR of 0.03, zero pointing error case.

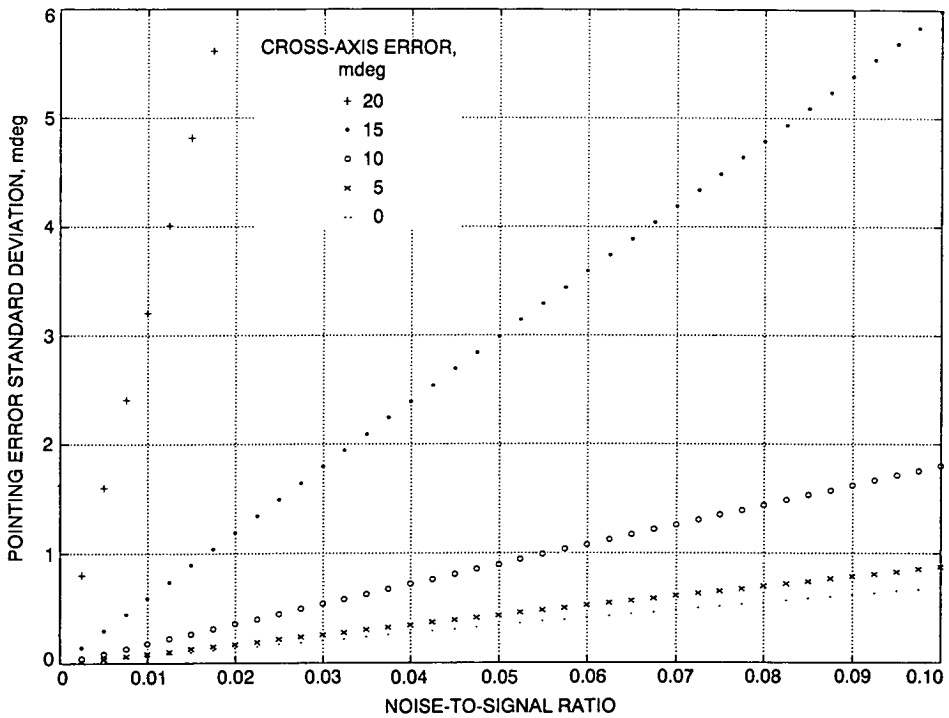


Fig. 6. Ka-band 5-point boresight uncertainty for increasing cross-axis errors, zero pointing error case.

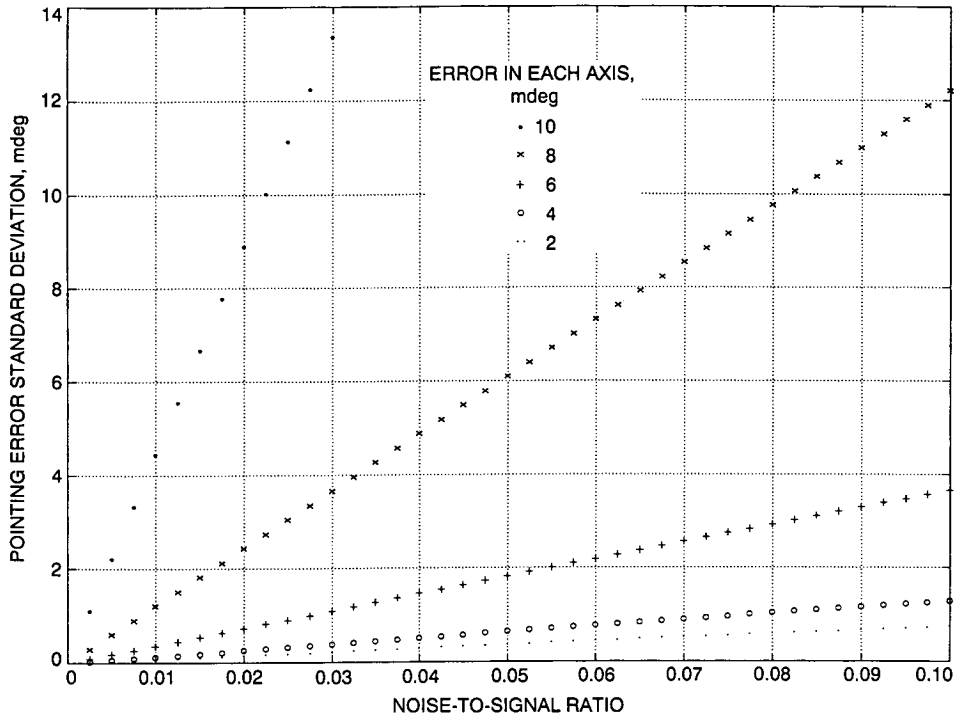


Fig. 7. Ka-band 5-point boresight uncertainty for equally increasing pointing and cross-axis errors.

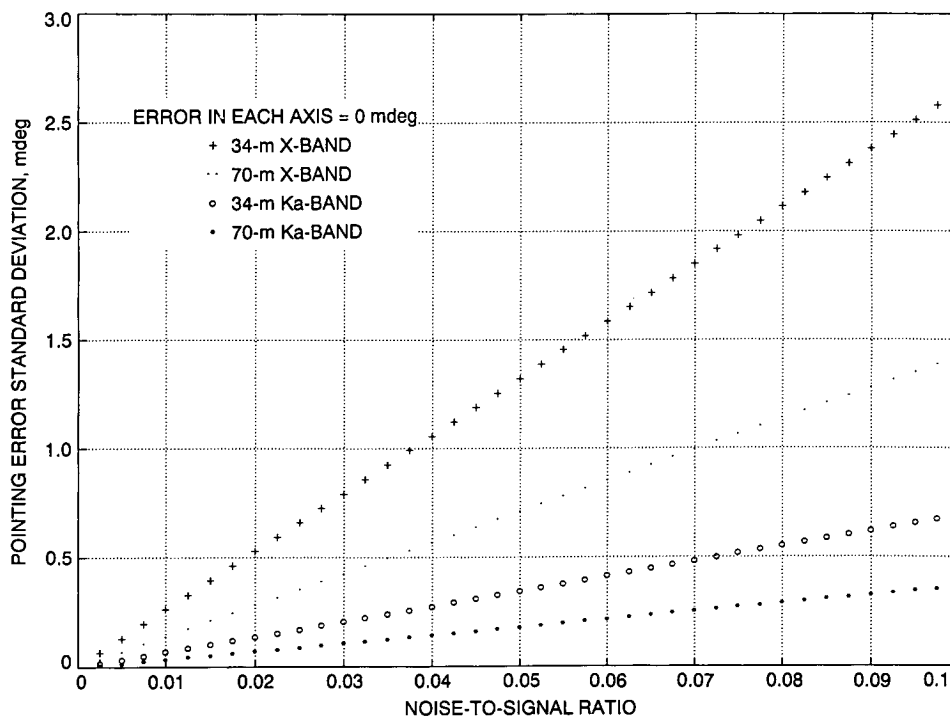


Fig. 8. 5-point boresight uncertainty for zero pointing and cross-axis errors.

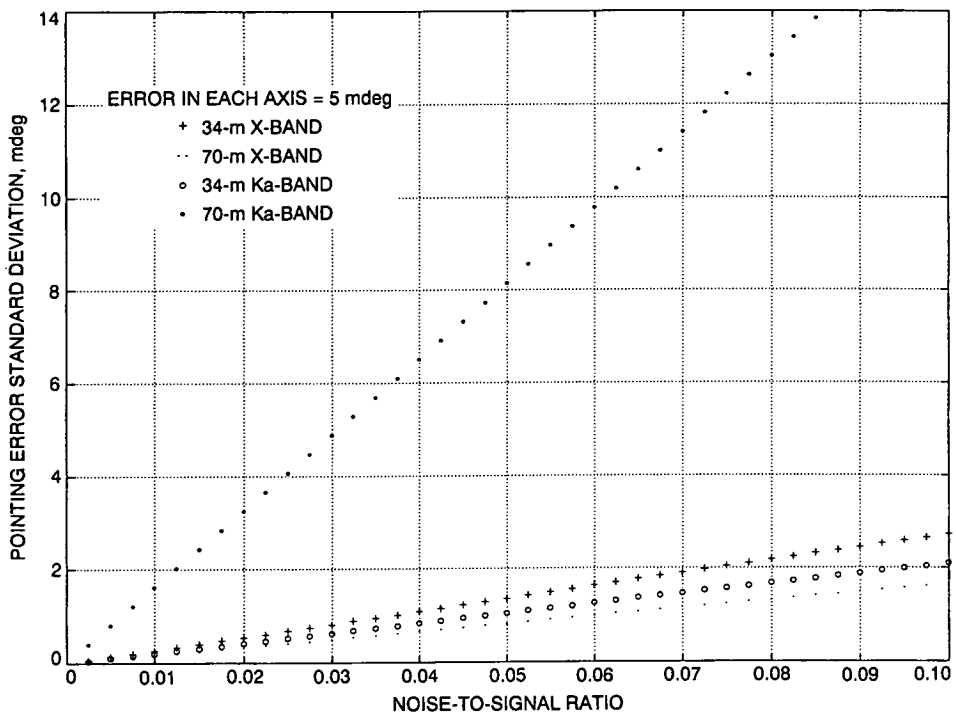


Fig. 9. 5-point boresight uncertainty for equal pointing and cross-axis errors of 5.0 mdeg.

Appendix A

Pointing Error Variance Equation

Let the signal level inputs be $y_i, i = 1, \ldots, n$, that are sensed at the angular offsets $x_i, i = 1, \ldots, n$ superimposed on the predicted target angles, then the received signal level model can be written in the following exponential form:

$$y_i(x_i) = \exp\left(c_1 + c_2 x_i + c_3 x_i^2\right) \tag{A-1}$$

It can be shown that the offset where the signal level is maximum, i.e., where $x = \varepsilon$ and ε is the beam pointing error, is

$$\varepsilon = -\frac{c_2}{2c_3} \tag{A-2}$$

where c_i, $i = 1, \ldots, 3$ are the coefficients of the best fit parabola satisfying the set of n equations

$$\ln\left(y_i\left(x_i\right)\right) = c_1 + c_2 x_i + c_3 x_i^2 \tag{A-3}$$

or in matrix notation

$$\begin{bmatrix} \ln\left(y_1\right) \\ \vdots \\ \ln\left(y_n\right) \end{bmatrix} = \begin{bmatrix} 1 & x(1) & x(1)^2 \\ & \vdots & \\ 1 & x(n) & x(n)^2 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \tag{A-4}$$

which can be written as

$$\mathbf{Y_{ln}} = \mathbf{AC} \tag{A-5}$$

where the measurement vector $\mathbf{Y_{ln}}$ is $n \times 1$, the measurement distribution matrix $\mathbf{A}$ is $n \times 3$, and the parameter vector $\mathbf{C}$ is 3×1. The computed value of the pointing error from Eq. (A-2) will have an uncertainty σ_ε due to measurement errors in the signal level inputs y_i, and perhaps due to significant antenna position errors in the offsets x_i.

The measurement function in Eq. (A-3) is $\ln\left(y_i\right)$. Let $y_{ln_i} = \ln\left(y_i\right)$, and the uncertainty in the signal level inputs y_i be σ_{y_i}, then the measurement uncertainty in y_{ln_i} is

$$\sigma_{yln_i} = \frac{\sigma_{y_i}}{y_i} \tag{A-6}$$

and will be propagated through the boresight algorithm. The σ_{y_i} are also presumed to encompass all uncertainty that was introduced in negating cold sky background noise from the y_i. Now, assuming negligible uncertainty in the x_i and that the σ_{yln_i} are random and independent, then the variance of the computed pointing error ε is

$$\sigma_\varepsilon^2 = \sum_{i=1}^{n} \left(\frac{\partial \varepsilon}{\partial y_{ln_i}} \sigma_{yln_i}\right)^2 \tag{A-7}$$

From the pointing error solution Eq. (A-2), the following partial derivative equation is obtained:

$$\frac{\partial \varepsilon}{\partial y_{ln_i}} = \frac{1}{2c_3^2}\left(c_2 \frac{\partial c_3}{\partial y_{ln_i}} - c_3 \frac{\partial c_2}{\partial y_{ln_i}}\right) \tag{A-8}$$

which can be inserted into Eq. (A-7) to yield

$$\sigma_\varepsilon^2 = \left(\frac{1}{2c_3^2}\right)^2 \sum_{i=1}^{n} \left(c_2 \frac{\partial c_3}{\partial y_{ln_i}} - c_3 \frac{\partial c_2}{\partial y_{ln_i}}\right)^2 \sigma_{yln_i}^2 \tag{A-9}$$

Analytic expressions for $\partial c_2 / \partial y_{ln_i}$ and $\partial c_3 / \partial y_{ln_i}$ are needed to completely describe the variance Eq. (A-9). The least squares solution for the vector $\mathbf{C}$ satisfies the following matrix equation:

$$\mathbf{C} = \left(\mathbf{A}^t \mathbf{A}\right)^{-1} \mathbf{A}^t \mathbf{Y_{ln}} \tag{A-10}$$

where $\mathbf{A}^t$ is the transpose of $\mathbf{A}$. Let the the matrix product $\left(\mathbf{A}^t \mathbf{A}\right)$ be denoted as $\mathbf{D}$ and be expanded [from Eq. (A-4)] as

$$\begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{bmatrix} = \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \tag{A-11}$$

where all summations (from here on) are from $i = 1$ to n. Let $\mathbf{D}'$ be the inverse of the matrix $\mathbf{D}$, or $\mathbf{D}' = \mathbf{D}^{-1}$ where

$$\mathbf{D}' = \begin{bmatrix} d'_{11} & d'_{12} & d'_{13} \\ d'_{21} & d'_{22} & d'_{23} \\ d'_{31} & d'_{32} & d'_{33} \end{bmatrix} \tag{A-12}$$

From Eq. (A-10) it can be shown that the solutions for c_2 and c_3 are

$$c_2 = d'_{21} \sum y_{ln_i} + d'_{22} \sum x_i y_{ln_i} + d'_{23} \sum x_i^2 y_{ln_i} \tag{A-13}$$

and

$$c_3 = d'_{31} \sum y_{ln_i} + d'_{32} \sum x_i y_{ln_i} + d'_{33} \sum x_i^2 y_{ln_i} \tag{A-14}$$

Finally, from the last two solution equations the needed partial derivative equations are calculated to be

$$\frac{\partial c_2}{\partial y_{ln_i}} = d'_{21} + d'_{22} x_i + d'_{23} x_i^2 \tag{A-15}$$

and

$$\frac{\partial c_3}{\partial y_{ln_i}} = d'_{31} + d'_{32} x_i + d'_{33} x_i^2 \tag{A-16}$$

and can be inserted in Eq. (A-9) to complete the pointing error variance equation.

Appendix B

DSS-13 Signal Level Input Variance Propagation Equations

The new DSS-13 antenna mechanical controller, called the antenna monitor and control computer (AMC), will interface with both the total power radiometer and TP-13 and ARX II receiver subsystems for noncoherent and coherent RF signal input into the boresight algorithm. This appendix will summarize the signal level variance propagation equations needed for each of the interfaces. It is noted that the boresight algorithm input is required to be in linear units, as opposed to logarithmic units proportional to power typically used to describe RF power measurments.

I. Noncoherent Signal Input

AMC inputs from the total power radiometer subsystem will be estimates of system noise temperature T_i with their standard deviations σ_{T_i} measured at the on-source angular offsets x_i, $i = 1, \ldots, n$, and the two off-source measurements T_{off1} and T_{off2} with their standard deviations $\sigma_{T_{off1}}$ and $\sigma_{T_{off2}}$ sensed at the offsets x_{off1} and x_{off2}. This axial measurement scheme is depicticed in Fig. B-1 for the case of $n = 3$, i.e., the three-point boresight algorithm. The required input to the boresight algorithm are the source noise temperature estimates T_{source_i} with variances $\sigma^2_{T_{source_i}}$ corresponding to the x_i. The classic straight-line sky background noise approximation is used to isolate the target signal levels. The following equation is applied to estimate the off-source background sky temperatures T_{sky_i} along the line shown in Fig. B-1:

$$T_{sky_i} = a_i(T_{off2} - T_{off1}) + T_{off1} \quad for\ i = 1, \ldots, n \quad \text{(B-1)}$$

where

$$a_i = \frac{abs(x_i - x_{off1})}{abs(x_{off2} - x_{off1})} \quad \text{(B-2)}$$

Then, the source temperature estimates are obtained through the subtraction

$$T_{source_i} = T_i - T_{sky_i} \quad \text{(B-3)}$$

The variances $\sigma^2_{T_i}$ are mapped through these equations to obtain the source variances $\sigma^2_{T_{source_i}}$. Equation (B-1) can be rewritten as

$$T_{sky_i} = (1 - a_i)T_{off1} + a_i T_{off2} \quad for\ i = 1, \ldots, n \quad \text{(B-4)}$$

and then, assuming that all of the temperature variances are independent and random, the estimated sky variances from (B-4) are

$$\sigma^2_{T_{sky_i}} = (1 - a_i)^2\sigma^2_{T_{off1}} + a_i^2\sigma^2_{T_{off2}} \quad \text{(B-5)}$$

where a_i is defined above. Finally, from Eq. (B-3) the source signal level variances are obtained as

$$\sigma^2_{T_{source_i}} = \sigma^2_{T_i} + \sigma^2_{T_{sky_i}} \quad \text{(B-6)}$$

and T_{source_i} and $\sigma^2_{T_{source_i}}$ for $i = 1, \ldots, n$ are the required inputs for the boresight algorithm and pointing error variance equation.

II. Coherent Signal Input

AMC inputs from the TP-13 and ARX II receiver subsystems will be estimates of SNR's $(P_c/N_o)_i$ and variances $\sigma^2_{(P_c/N_o)_i}$ sensed at the on-source angular offsets x_i. Estimates of noise power $(N_o)_i$ and variances $\sigma^2_{(N_o)_i}$ will be derived from simultaneous radiometer subsystem inputs of system noise temperatures T_i and standard deviations σ_{T_i} measured at the on-source angular offsets x_i. The noise power estimates over a 1-Hz bandwidth are defined as $(\hat{N}_o)_i = \kappa T_i$, where κ is Boltzmann's Constant. Then the estimated noise power variances are then $\sigma^2_{(\hat{N}_o)_i} = \kappa^2\sigma^2_{T_i}$. Multiplication negates the noise power from the target signal power, as follows:

$$(\hat{P}_c)_i = \left(\frac{P_c}{N_o}\right)_i (\hat{N}_o)_i \quad \text{(B-7)}$$

Now, assuming that the above equation is the product of two independent random variables, the signal power variance can then be shown to be

$$\sigma^2_{(\hat{P}_c)_i} = \sigma^2_{(\hat{N}_o)_i}\left(\frac{P_c}{N_o}\right)^2_i + \sigma^2_{(P_c/N_o)_i}(\hat{N}_o)^2_i + \sigma^2_{(\hat{N}_o)_i}\sigma^2_{(P_c/N_o)_i} \quad \text{(B-8)}$$

and $(\hat{P}_c)_i$ and $\sigma^2_{(\hat{P}_c)_i}$ for $i = 1, \ldots, n$ are the required inputs for the boresight algorithm and pointing error variance equation.

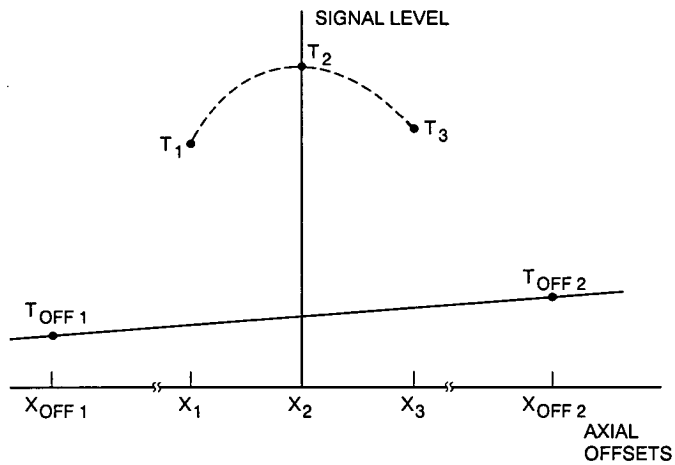


Fig. 10. Noncoherent signal measurement scheme.

Atmospheric Refraction Correction for Ka-Band Blind Pointing on the DSS-13 Beam Waveguide Antenna

I. M. Perez-Borroto and L. S. Alvarez
Ground Antennas and Facilities Engineering Section

An analysis of the atmospheric refraction corrections at the DSS-13 34-m diameter beam waveguide antenna for the period covering July through December 1990 is presented. The current DSN atmospheric refraction model and its sensitivity with respect to sensor accuracy are reviewed. Refraction corrections based on actual atmospheric parameters are compared with the DSS-13 station default corrections for the six-month period. Average blind-pointing improvement during the worst month would have amounted to 5 mdeg at 10 deg elevation using actual surface weather values. This would have resulted in an average gain improvement of 1.1 dB.

I. Introduction

The bending of radio waves as they pass obliquely through the atmosphere can produce large pointing errors. Currently in the DSN, a refraction model (referred to here as the Lanyi model[1] is used to compute a correction which reduces antenna beam-pointing errors in elevation. The inputs to the refraction model are predicted elevation angle, surface atmospheric pressure, surface temperature, and surface relative humidity. The use of actual weather parameter inputs from antenna station instrumentation, instead of from default values, will significantly improve blind-pointing accuracy. This article investigates the magnitude of the pointing improvement by analyzing real atmospheric conditions at the DSS-13 beam waveguide (BWG) antenna.

During the phase 1 testing period [1] at DSS 13, gain and pointing calibrations were conducted from June 1990 through January 1991. Throughout this period the Lanyi refraction model was used to compute and apply refraction corrections based on station default weather parameters coded in the Antenna Controller Subassembly (ACS). Station weather instrumentation readings, not interfaced with the antenna-pointing system at the time, were also logged at ten-minute intervals. These atmospheric readings allow analysis and comparison of actual weather-based refraction corrections with the fixed-parameter corrections during the experiment period.

It was anticipated that actual beam-pointing error measurements made on stellar radio sources during the phase 1

[1] G. Lanyi, "Atmospheric Refraction Corrections to Antenna Pointing at 1 Millidegree Accuracy," JPL Interoffice Memorandum 335.3-89-026, Jet Propulsion Laboratory, Pasadena, California, March 24, 1989.

testing would enable assessment of the reduction of actual elevation pointing errors. Unfortunately, only a limited number of measurements were made at the lower elevation angles, and they appeared to be of only marginal quality. Due to the nonrepeatable nature and poor quality of these pointing measurements, they were deemed unreliable for the purpose of analysis. In this article, estimated improvement in the refraction pointing error correction at DSS 13 is based strictly on the deviation of the logged meteorological parameters (pressure, temperature, and relative humidity) from the default DSS-13 weather parameters.

The data analysis to be presented will span the real-time weather input measurement period between July 1 and December 31, 1990. The data acquisition and reduction techniques are described. The Lanyi refraction model, along with its sensitivity to input weather parameters, is reviewed. It is shown that the magnitude of the refraction pointing error correction based on actual atmospheric parameters deviates significantly from the default, on the average, as the elevation tracking angle decreases. The Ka-band gain loss corresponding to deviations in the corrections emphasizes the need for accurate, real-time weather inputs to the DSN refraction correction algorithm.

II. Refraction Correction in the DSN

A. Background

Until 1990, the DSN used the Berman–Rockwell refraction model defined in [2,3]. In 1990, the Lanyi model was implemented in the DSN, as well as at the DSS-13 34-m BWG antenna. A detailed comparison of the models can be obtained through the references. The main motivation for implementing the Lanyi model was its estimated improvement in accuracy, especially at the low-elevation angles. The estimated Lanyi model accuracy is stated to be better than 0.5 mdeg at elevation angles greater than 6 deg, assuming that accurate and current measurements of the atmosphere are provided to the antenna controller. In terms of the current DSN Media Calibration Subsystem (DMD) weather instrumentation specifications, this could produce an error of 1.5 to 4.0 mdeg at 6 deg elevation for average and extreme DSN water vapor content, respectively. Such pointing errors may seem small but are significant when contemplating error budgets for Ka-band beamwidths (17 mdeg for 34-m antennas and 9 mdeg for 70-m antennas).

B. The Lanyi Refraction Model

The Lanyi refraction model has the basic functional form of

$$\Delta EL = f(P, T, RH, EL) \tag{1}$$

where

ΔEL = change in elevation pointing

P = surface pressure

T = surface temperature

RH = surface relative humidity

EL = predicted elevation angle

The DSN default surface weather parameters, which were used at DSS 13 are

P = 901.09 mbar

T = 22.07 deg C

RH = 31.57 percent

Figure 1 shows the magnitude of the refraction corrections at DSS 13 as a function of elevation angle, using the default weather parameters.

III. DSS-13 Surface Atmosphere Measurements

A. Weather Instrumentation and Measurement Analysis

Daily records of surface atmospheric conditions at DSS 13 were collected from July through December 1990. These time-tagged measurements provide the basis for computation of actual refraction corrections, which were used for comparison with the default corrections. It is thus necessary to try to assess the quality of the measurements.

Two sensors for each weather parameter were operational during the period of interest, and their observations were recorded every 10 min by the station weather data acquisition computer. The accuracy specifications of the instruments are stated to be

Pressure: 800 to 1100 mbar, ±0.3 mbar

Temperature: −50 to +80 deg C, ±0.1 deg C

Relative humidity: 0 to 80 percent, ±2 percent, and 80 to 100 percent, ±3 percent

The differences between the two sensors for each measurement (25,329 data points) were computed. The average

difference between the two sensors for all days was computed to be

Pressure: 0.0493 mbar (σ = 0.0791 mbar)

Temperature: 0.0710 deg C (σ = 0.1428 deg C)

Relative humidity: 3.9 percent (σ = 3.7 percent)

Both the pressure and temperature sensor pairs agreed extremely well during this period, while at least one relative humidity probe appeared to be biased. The differences also display large variations. To illustrate the variability in the humidity sensors, Fig. 2 shows the average daily difference between the humidity observations for days 267 through 321. The readings from sensor 2 are always larger than those registered by sensor 1. The true relative humidity value may lie between the sensor readings, or one or both of the sensors may be biased high or low. These discrepancies in humidity can map into significant differences in refraction correction, especially at low elevation angles. To minimize the impact on the refraction correction analysis, the atmospheric measurements were filtered in the following manner: Points were eliminated when the difference between the two measurements from each of the sensors was greater than twice the stated DSS-13 accuracy specification. Approximately twenty percent of the data points were removed in this manner.

B. Sensor Error Propagation

As noted, uncertainties in all three surface weather measurements will propagate into errors in the computed refraction corrections. To quantify the correction degradation, the sum of the squares of the partial derivatives of the refraction model with respect to each input parameter were computed.

A first-order approximation to the Lanyi model, which is adequate for sensitivity analysis, is given by

$$\Delta EL = \frac{\chi_0 - Z_{tot}/\left(R\sin^2(EL)\right)}{\tan(EL)} \qquad (2)$$

where

$\chi_0 = \chi_{dry} + \chi_{wet}$ = total surface refractivity

$Z_{tot} = Z_{dry} + Z_{wet}$ = total zenith path delay

R = Earth radius

EL = the uncorrected elevation angle

The dry and wet surface refractivities and dry and wet zenith path delays can be determined from surface measurements of pressure, temperature, and relative humidity. The DSS-13 weather instrumentation specifications were then input as uncertainties and yielded an error of 0.71 mdeg in the computed refraction correction at 10 deg elevation. Thus, the sensor error propagation would not be a major problem in this current study if all the sensors were within their accuracy limits.

Figure 3 shows the DSS-13 rss refraction correction error at 10 deg elevation due to relative humidity uncertainty, using the default weather parameters and the given pressure and temperature sensor specifications. It is seen that when the sensor error increases above 3 percent, refraction error correction on the millidegree level is unachievable.

IV. Refraction Correction Analysis

A. Computed Refraction Corrections

Refraction corrections were computed at 1-hr intervals for the atmospheric measurement set spanning the six months (3446 points). In order to examine the variability of the computed corrections over this period, the extreme ranges of the sensor readings are considered. By setting two of the three input variables (pressure, temperature, and relative humidity) to the DSS-13 default values and entering the extreme points listed below into the refraction model, the correction ranges listed in Table 1, in millidegrees, are computed.

As seen in that table, the change in refraction correction due to relative humidity is about 17 times greater than that due to pressure and about 3 times greater than that due to temperature.

B. Effect on Gain

The absolute differences between the default refraction corrections and those corresponding to actual weather parameters were computed. The resultant values are assumed to be improvements in the beam-pointing accuracy for blind pointing if the real-time surface weather observations were used in the refraction correction for the six-month period. For all the hour-interval atmospheric measurements, the absolute difference is computed at elevation increments of 5 deg. Figure 4 illustrates the differences for the month of October 1990, which had the highest average refraction difference from the default refraction values.

Note that for this particular month, very few actual refraction corrections equaled the default corrections (absolute difference = 0). Thus, rarely would good blind pointing be achieved, and the average pointing errors at low elevation angles would be rather large (4.8 millidegrees). Expected beam-pointing improvement (using real weather inputs) increases significantly as the elevation angle is decreased.

To summarize the whole six-month period, statistics were computed for the entire data set at 5-deg increments. Figure 5 shows the means and standard deviations of the absolute differences. At 10 deg elevation, the refraction pointing error should, on the average, be reduced by 4 mdeg, with a 2.4-mdeg 1-σ variation. The expected DSS-13 Ka-band gain degradation corresponding to the average differences is shown in Fig. 6. X-band gain loss would be less than 0.1 dB. The large magnitude of the average gain loss at the lower elevations stresses the need for accurate, real-time weather inputs for refraction correction during Ka-band tracking operations.

Figure 7 shows the means of the absolute correction differences for the months of August and October, when the smallest and largest average differences were observed, respectively. Figure 8 illustrates the corresponding Ka-band gain loss. The plots indicate that on the average, the DSS-13 default weather parameters matched the actual atmospheric conditions best in August and worst in October.

V. Conclusions

An analysis of the atmospheric refraction correction at the DSS-13 BWG antenna for the period covering July through December 1990 has been presented. The Lanyi refraction model and its sensitivity with respect to sensor error were reviewed. It was shown that the present specifications on the DSS-13 weather instrumentation are sufficient to provide submillidegree refraction correction, however, performance will sharply degrade when the relative humidity sensors fail to meet their specified accuracy.

Refraction corrections based on actual atmospheric parameters from the six-month period were computed and compared with the DSS-13 station default corrections. The average worst-month differences between the corrections was 5 mdeg at 10 deg elevation (Fig. 7). The corresponding average Ka-band gain loss expected using the DSS-13 default weather parameters during this period was thus 1.1 dB at that elevation (Fig. 8). The X-band gain loss would be approximately 0.1 dB. This significant gain loss mandates the need for accurate, real-time weather-parameter-based refraction correction for future Ka-band tracking operations.

Acknowledgment

Steve Slobin is acknowledged for providing many helpful technical suggestions and encouragement.

References

[1] S. D. Slobin, T. Y. Otoshi, L. S. Alvarez, M. J. Britcliffe, S. R. Stewart, and M. M. Franco, "Efficiency Calibration of the DSS 13 34-Meter Beam Waveguide Antenna at 8.45 and 32 GHz," *TDA Progress Report 42-106*, vol. April–June 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 283–297, August 15, 1991.

[2] A. L. Berman and S. T. Rockwell, *New Optical and Radio Frequency Angular Tropospheric Refraction Models for Deep Space Applications*, Technical Report 32-1601, Jet Propulsion Laboratory, Pasadena, California, November 1, 1975.

[3] A. L. Berman, "Modification of the DSN Radio Frequency Angular Tropospheric Refraction Model," *DSN Progress Report 42-38*, vol. January–February 1977, Jet Propulsion Laboratory, Pasadena, California, pp. 184–186, April 15, 1977.

Table 1. Effect of measured weather extremes on calculated refraction correction.

Parameter	Refraction correction, mdeg		
	10-deg elevation	20-deg elevation	30-deg elevation
Pressure, 883 to 907 mbar	83.6–85.6	41.7–42.7	26.5–27.1
Temperature, −9.7 to 37.7 deg C	84.3–95.1	42.1–47.4	26.7–30.1
Relative humidity, 4 to 99 percent	75.2–109.7	37.6–54.6	23.9–34.6

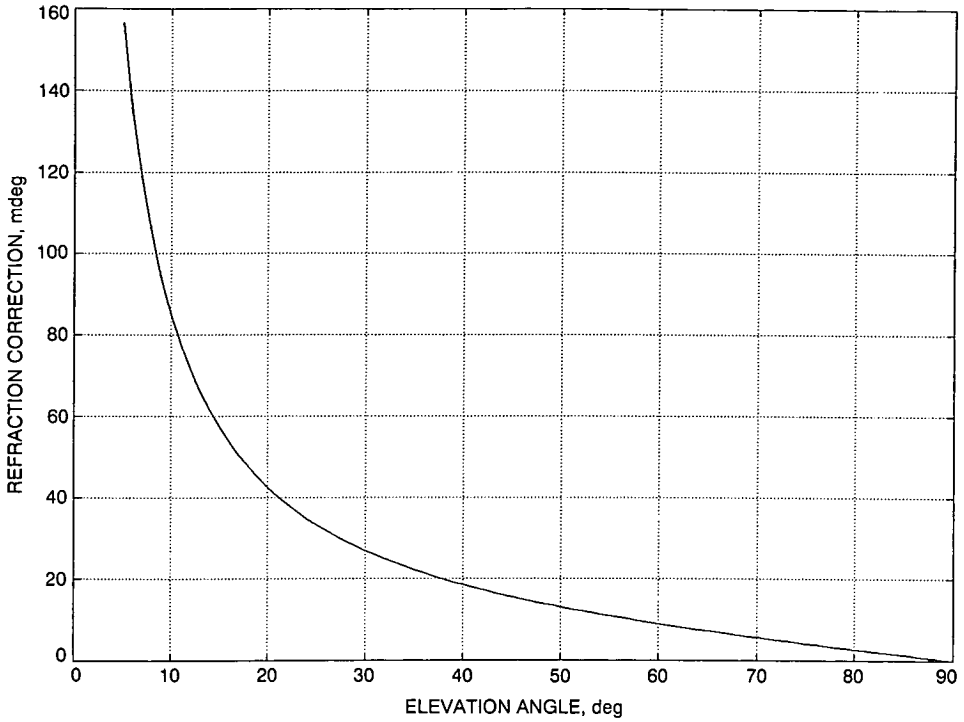


Fig. 1. Lanyi angular refraction correction model for DSS-13 default atmospheric parameters.

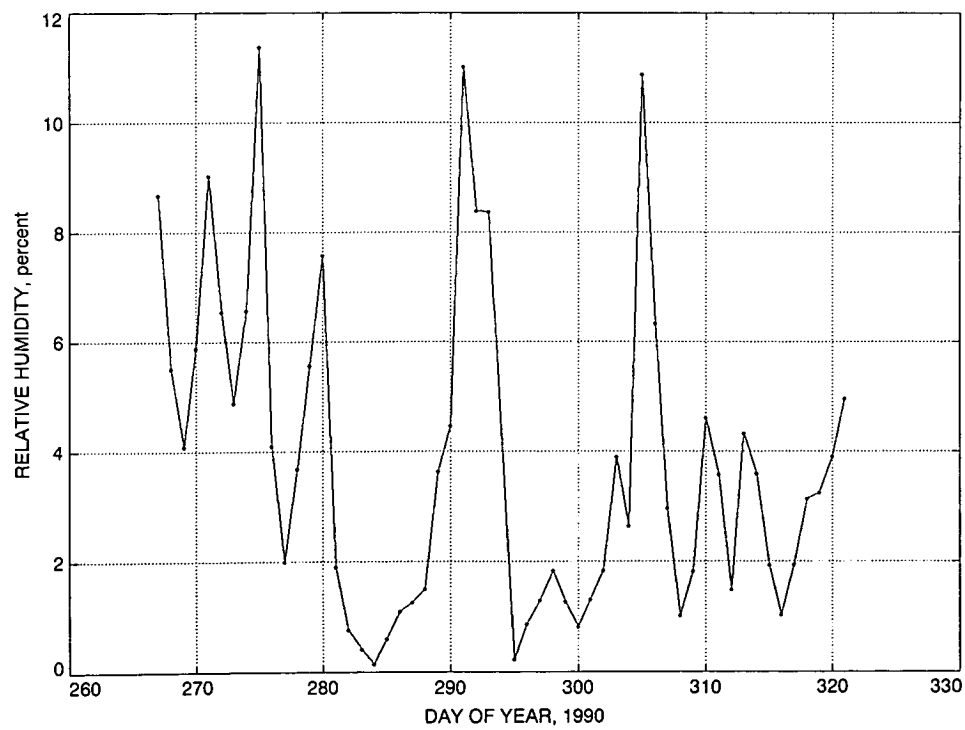


Fig. 2. Daily average difference between relative humidity probe readings.

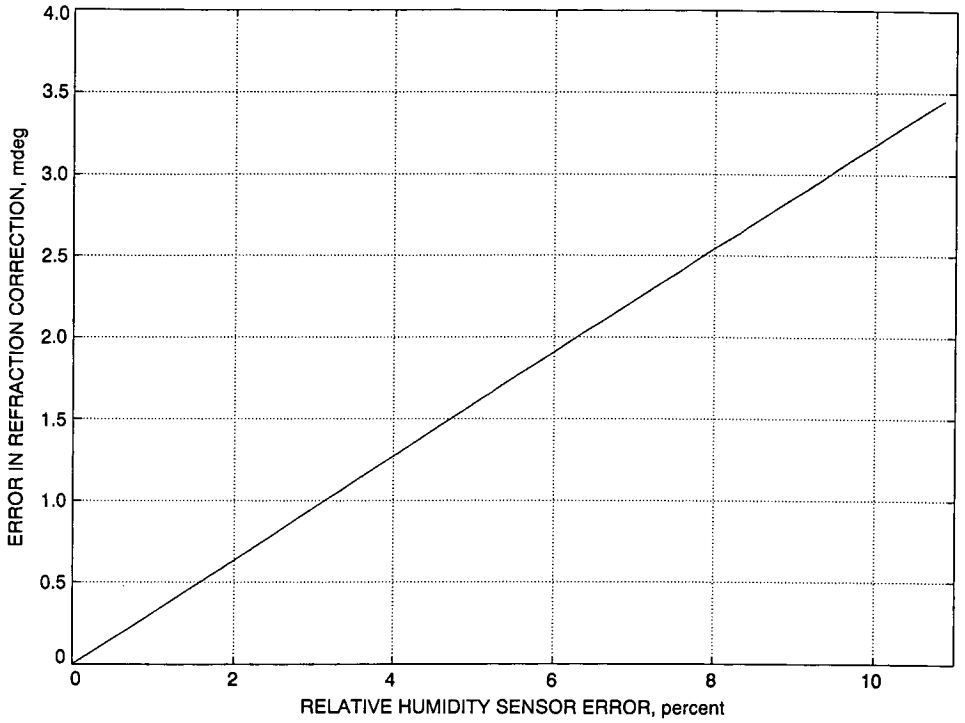


Fig. 3. Error in computed correction at 10 deg elevation, due to relative humidity instrument error, using default parameters.

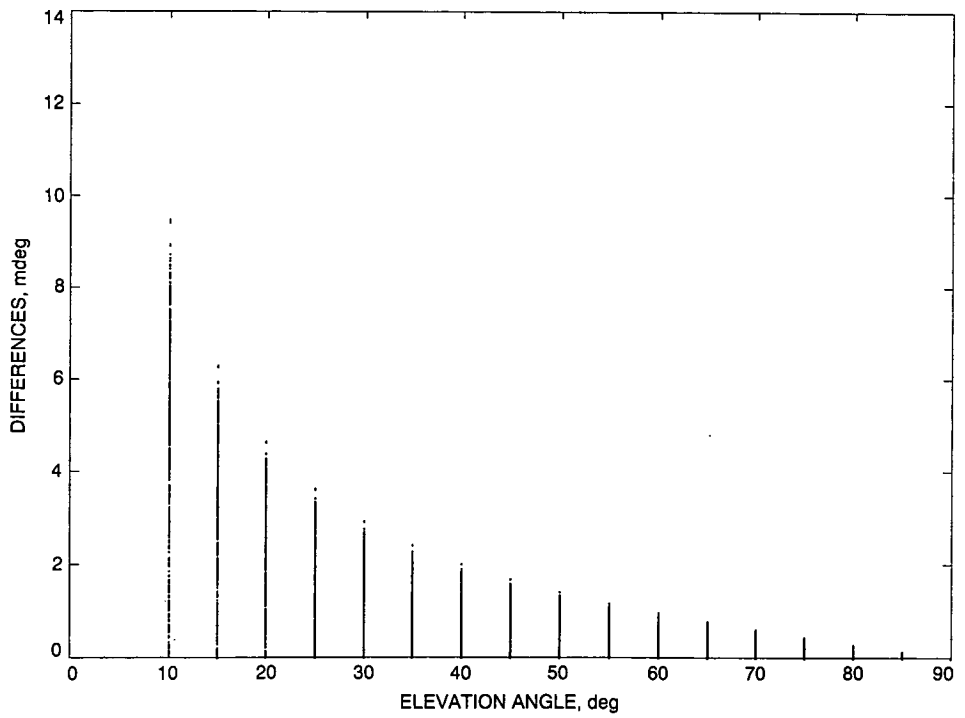


Fig. 4. Absolute difference between actual weather-based refraction corrections and default corrections for October 1990.

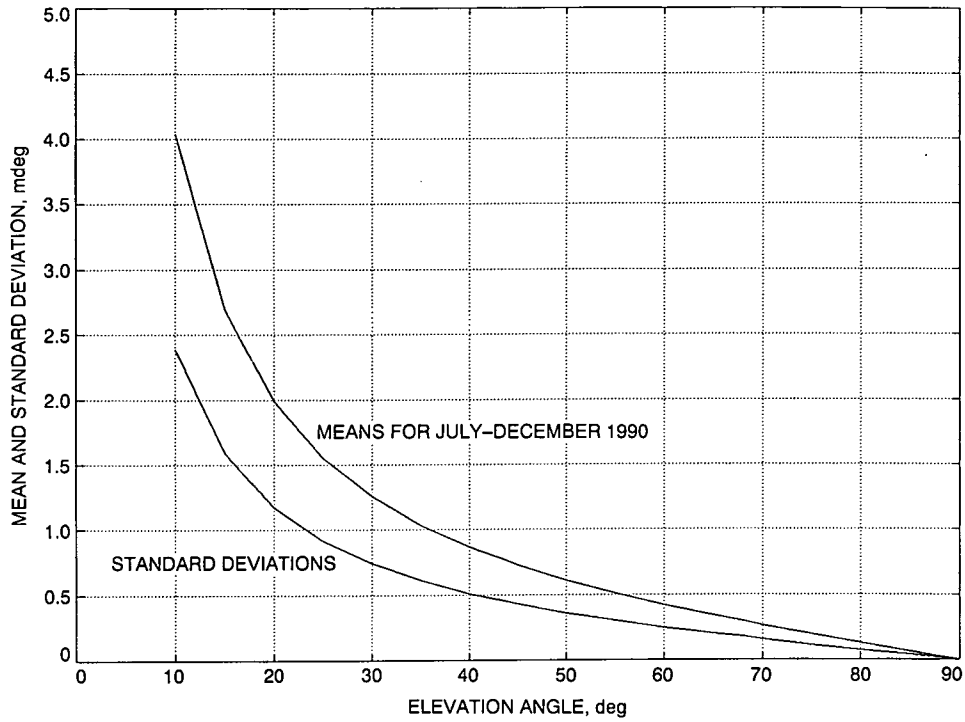


Fig. 5. Mean and standard deviation of the absolute difference between default and actual refraction corrections during July–December 1990.

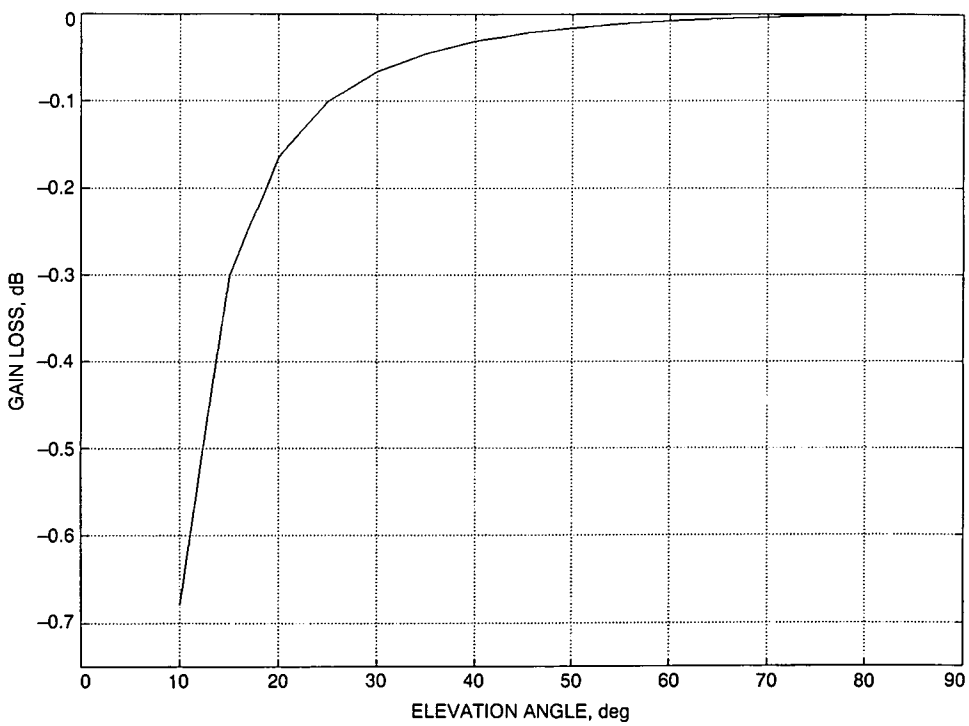


Fig. 6. Gain loss for refraction pointing error, July–December 1990.

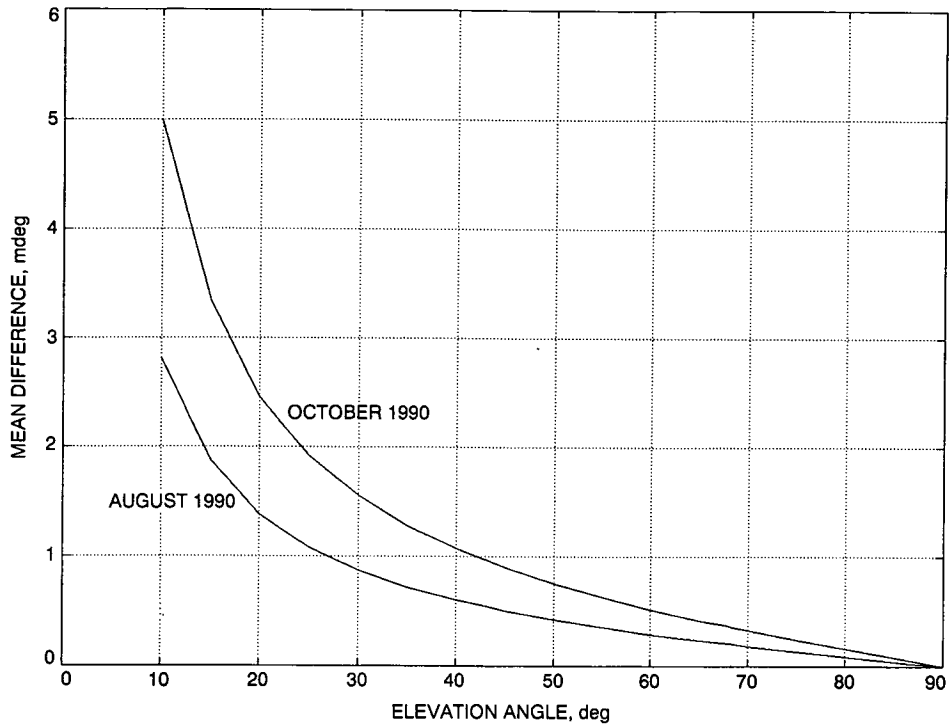


Fig. 7. Mean of the absolute difference between default and absolute refraction corrections, August and October 1990.

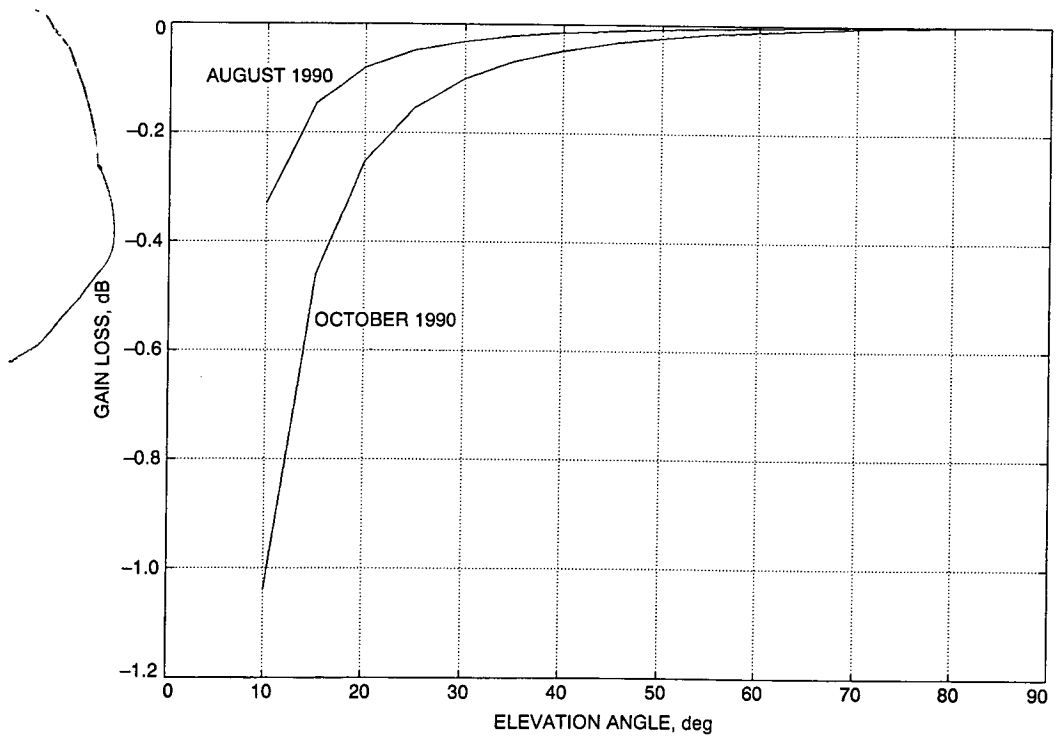


Fig. 8. Gain loss for refraction pointing error, August and October 1990.

Determination of the Dissipative Loss of a Two-Port Network From Noise Temperature Measurements

T. Y. Otoshi
Ground Antennas and Facilities Engineering Section

When radiometric equipment is available, noise temperature measurement techniques provide a convenient and accurate means for determining the dissipative component of the insertion loss of a two-port network. It is increasingly becoming the practice to ignore mismatch errors caused by multiple reflections between the source, the receiver, and the component whose dissipative loss is being measured. Mismatch errors are difficult to determine in practice because of the requirements of having full knowledge of the magnitudes of reflection coefficients and S-parameters. This article shows it is permissible to neglect the effects of mismatch errors if special conditions are met. These special conditions only require that the reflection coefficients of the source and load be made negligibly small and that the two-port network being evaluated has reciprocal and symmetrical properties.

I. Introduction

Noise temperature techniques have been employed successfully in the past for determining the dissipative losses of low-loss, two-port networks, including antenna waveguide components [1], dichroic plates [2], and radome materials.[1] Accuracies on the order of 0.002 dB have been achieved on a measurement of dissipative loss of 0.007 dB for a mode generator and quarterwave plate polarizer waveguide section [1].

Rigorous treatments of general case mismatch errors associated with noise temperature measurement techniques have been presented elsewhere [1,3]. In practice, reflection coefficient and S-parameter information required for mismatch error corrections are difficult to obtain. As a result, the matched case formulas are often employed without knowledge of the magnitude of mismatch errors involved. It is fortuitous that, in practice, attempts are still being made to reduce the reflection coefficients of the source and load (horn-receiver) to negligibly small values.

In the following section, it will be shown that for a special case where the source and load are Z_0-matched, it becomes possible to obtain a very simple mismatch error correction term that can be applied to the dissipative loss value calculated from matched case formulas. There is no restriction on the magnitude of the reflection coefficient of the two-port network, as long as the network is reciprocal and symmetrical. It will further be demonstrated that for most practical cases where the magnitude of the reflec-

[1] T. Y. Otoshi and M. M. Franco, "Results of Dielectric Constant and Radiometric Measurements on Radome Samples," JPL Memorandum 3328-92-041 (internal document), Jet Propulsion Laboratory, Pasadena, California, March 31, 1992.

tion coefficient is less than 0.1 [corresponding to a voltage standing-wave ratio (VSWR) of 1.2], the correction term is small and may be neglected.

II. Derivation

Insertion loss of a microwave component consists of both reflective and dissipative components of loss. Low dissipative loss in the passband is one of the more important properties that a component must have when it is used as part of the front end of a deep-space communications antenna receiving system.

For the purposes of this article, it will be assumed that the two-port network being considered is part of a basic receiving system, such as the one shown in Fig. 1. Applying the special case conditions that $\Gamma_g = \Gamma_r = 0$, $|S_{11}| = |S_{22}|$, and $S_{12} = S_{21}$ to the general case equations in [3], the total input noise temperature *delivered*[2] to the receiver (load) is

$$T_g' = T_g|S_{21}|^2 + T_n \tag{1}$$

where S_{11}, S_{12}, S_{21}, and S_{22} are the scattering parameters of the two-port network [4]; Γ_g and Γ_r are the voltage reflection coefficients of the generator and receiver, respectively; T_g is the source noise temperature *available* at port 1, in kelvins; and T_n is the noise temperature that is generated by the two-port network and *delivered* to the receiver at port 2, in kelvins. For the stated special case conditions,

$$T_n = \left(1 - |S_{11}|^2 - |S_{21}|^2\right) T_p \tag{2}$$

where T_p is the physical temperature of the network in kelvins.

Examples of two-port networks that meet the requirements of reciprocity and $|S_{11}| = |S_{22}|$ are dichroic plates and radome materials. It is of interest to express Eq. (1) in terms of power loss ratios associated with the reflective and dissipative attenuation of the signal. Attenuation is defined as the insertion loss of a network when the source and load reflection coefficients are zero [4]. The total attenuation of a network can be expressed as

$$A_T = -20\log_{10}|S_{21}| = -10\log_{10}\left(1 - |S_{11}|^2\right) + L_{\mathrm{dB}} \tag{3}$$

where L_{dB} is the dissipative component of attenuation expressed as

$$L_{\mathrm{dB}} = 10\log_{10} L \tag{4}$$

and

$$L = \frac{1 - |S_{11}|^2}{|S_{21}|^2} \tag{5}$$

The term L is the familiar dissipative power loss ratio (≥ 1), but differs from the matched load case formula in that $|S_{11}|$ (equal to $|S_{22}|$) is not zero. Use of Eq. (5) in Eq. (1) leads to

$$T_g' = T_g\left(1 - |S_{11}|^2\right) L^{-1} + T_n \tag{6}$$

where the expression for the delivered network noise temperature given by Eq. (2) now becomes

$$T_n = \left(1 - |S_{11}|^2\right)\left(1 - L^{-1}\right) T_p \tag{7}$$

Note that when $|S_{11}| = 0$, Eq. (6) becomes the familiar Z_0-matched case formula

$$T_g' = T_g L^{-1} + T_n \tag{8}$$

where the network generated noise temperature due to dissipative loss becomes

$$T_n = \left(1 - L^{-1}\right) T_p \tag{9}$$

It will now be assumed that the noise temperature T_n given by Eq. (7) has been measured radiometrically through the use of techniques similar to those described in [1] and [2]. It is now of interest to determine the associated dissipative component of attenuation. The procedure involves solving for L in Eq. (7), and then substituting the resulting expression into Eq. (4). These steps yield the desired result of

$$L_{\mathrm{dB}} = -10\log_{10}\left(1 - \frac{T_n}{T_p}\right) + C_{\mathrm{dB}} \tag{10}$$

where C_{dB} is a correction term derived to be

$$C_{\mathrm{dB}} = 10\log_{10}\left[1 + (L-1)|S_{11}|^2\right] \tag{11}$$

For low-loss cases, L is close to unity, so that the above correction term simplifies to

[2] Delivered and available noise temperatures at a particular port differ and are related by a mismatch factor [3].

$$C_{\mathrm{dB}} \approx L_{\mathrm{dB}}|S_{11}|^2 \qquad (12)$$

when L_{dB} was defined by Eq. (4). It can now can be seen that for many practical cases involving low-loss components, the correction term will be very small. For example, if values $|S_{11}| = 0.1$ (corresponding to a VSWR = 1.2) and $L_{\mathrm{dB}} = 0.1$ dB (corresponding to $L = 1.023$), the correction term from Eq. (12) is 0.001 dB. A second example is that when $L_{\mathrm{dB}} = 0.1$ dB, but $|S_{11}| = 0.3$ (corresponding to a VSWR = 1.86), the correction term is still only 0.0091 dB [from Eq. (11)] and 0.009 dB [from Eq. (12)]. A third example is when the dissipative component of attenuation is only 0.01 dB. It can be seen from Eq. (12) that the correction term will still be negligibly small if $|S_{11}| < 0.3$.

It should be reemphasized that the above is a special case, where it is assumed that the generator and receiver reflection coefficients are zero. An attempt should be made in practice to achieve this condition as nearly as possible. The exact equations for the general mismatched case, with none of the above restrictions, can be found in [3].

III. Conclusions

For the special case conditions stated in this article, a simple correction term has been derived for applying it to the Z_0-matched case formula for determining the dissipative loss of two-port networks. It has been found that when the special case conditions are met, the correction term will typically be negligibly small.

References

[1] C. T. Stelzried and T. Y. Otoshi, "Radiometric Evaluation of Antenna-Feed Component Losses," *IEEE Transactions of Instrumentation and Measurement*, vol. IM-18, no. 3, pp. 172–183, September 1969.

[2] T. Y. Otoshi and M. M. Franco, "Dual passband dichroic plate for X-band," *TDA Progress Report 42-94*, vol. April–June 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 110–134, August 15, 1988.

[3] T. Y. Otoshi, "The Effect of Mismatched Components on Microwave Noise Temperature Calibrations," *IEEE Transactions on Microwave Theory and Techniques*, (Special Issue on Noise), vol. MTT-16, no. 9, pp. 675–686, September 1968.

[4] D. M. Kerns and R. W. Beatty, *Basic Theory of Waveguide and Introductory Microwave Network Analysis*, New York:Pergamon, pp. 41, 61, 63, 1967.

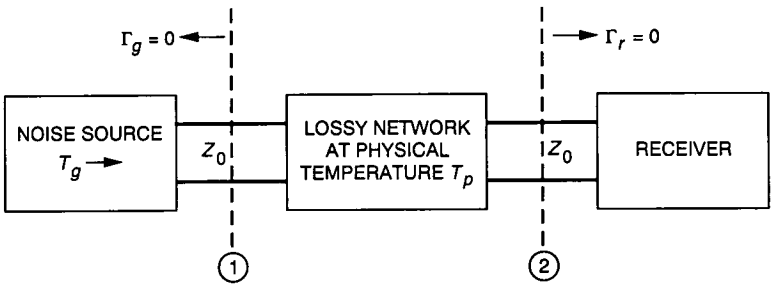


Fig. 1. A two-port lossy network in a receiving system having a matched generator and matched receiver ($\Gamma_g = \Gamma_r = 0$).

Self-Dual (48,24;12) Codes

G. Solomon
Communications Systems Research Section

Two self-dual (48,24;12) codes are constructed as 6 × 8 matrices whose columns add up to form an extended BCH-Hamming (8,4;4) code and whose rows sum to odd or even parity. The codes constructed have the identical weight structure of the extended quadratic residue code of length 48. Algebraic isomorphisms may exist between pairs of these three codes. However, because of their matrix form, the newly constructed codes are easily correctable for all five-error and many six-error patterns. The first code comes from restricting a binary cyclic (63,18;36) code to a 6 × 7 matrix and then adjoining six dimensions to the extended 6 × 8 matrix. These six dimensions are generated by linear combinations of row permutations of a 6 × 8 matrix of weight 12, whose sums of rows and columns add to one. The second code comes from a slight modification in the parity (eighth) dimension of the Reed-Solomon (8,4;5) code over GF(64). Error correction in both codes uses the row sum parity information to detect errors in the correction algorithm.

I. The Code Constructions

The two constructed codes are similar in their final six dimensions. The first 18 dimensions of the constructed (48,24;12) codes are basically (42,18;12) codes represented as 6 × 7 matrices. An additional eighth row of all zeros is adjoined to give a (48,18;12) code. The last six dimensions are constructed the same way, although the encoding algorithm for the second (modified Reed-Solomon) code is direct and systematic. Decoding the two codes uses basically the same techniques as discussed in the second code description.

II. A Self-Dual (48,24;12) Code

Consider the BCH (63,18;24) code of length 63 generated by the recursion polynomial $f_1(z)f_3(z)f_{-1}(z)$, where $f_1(z) = z^6+z+1$, where β is a root that is a primitive generator of the 63 roots of unity in $GF(64)$. The polynomials $f_3(z)$ and $f_{-1}(z)$ contain β^3 and β^{-1} as roots, respectively. Restrict the values of the code to the coordinates $9i + 7j$ for $0 \leq i \leq 6$, $j = 1, 2, 4, 5, 7, 8$. The constructed code is a (42,18;12) code. To prove this, one examines the matrix in a Mattson-Solomon (MS) polynomial formulation over the rows.

Let $z = xy$, where $x^7 = 1$, $y^9 = 1$, $x = \beta^{9i}$, and $y = \beta^{7j}$. Indexing the rows by y, the MS polynomial for each row y is $P_y(x) = \text{Tr}\,(C_1y + (C_1y)^8)x + (C_3y^3 + C_3^8y^6)x^3 + (C_1y^1 + C_1^8y^1)x^5$, which becomes, in the Solomon-McEliece Γ_2 formulation, $P_y(x) = \text{Tr}\,(C_1y + (C_1y)^8)x + (C_3^2y^6 + C_3^{16}y^3 + C_1^4y^1 + C_1^8y^8)x^6$. Thus, the coefficient of x is seen to be a (6,2;5) code over $GF(8)$, while the coefficient of x^6 is a (6,4;3) code over $GF(8)$. The minimum binary weight of the six rows is ≥ 10. The sum Γ_2 over the rows gives zero, showing that the weights are multiples of 4 and proving that the minimum distance of the code is ≥ 12. Note that the sum over the rows is the (7,3;4) codeword given by Tr $(C_3 + C_3^8)x^6$.

Adjoin an eighth column of all zeros to this 7×6 matrix. Then, add six more dimensions by forming linear combinations of all cyclic row permutations of the single matrix

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

The newly constructed code of length 48 and dimension 24 has a minimum distance of 12. The code of dimension 23 coming from sums of pairs of rows with weights of 24 is easily seen to have a distance of 12 and has row sums equal to zero. For the 24th dimension, whose row sums are odd, one need only check that certain weight patterns in the dimension 18 code do not exist.

To verify the results, one notes that if $\sum_j \Gamma_2 = 0$, $j = 1, 2, 4, 5, 7, 8$, one need only investigate the weight forms in any row permutation (6, 6, 6, 6, 6, 6), (6, 6, 6, 6, 2, 2), (6, 6, 6, 6, 4, 0), (2, 2, 2, 2, 2, 2), etc., to verify that this addition does not alter the basic minimum distance and self-orthogonality. This works. This is Code A.

III. A Modified Reed-Solomon (8,4;5) Code Over $GF(64)$

This code, Code B, while resembling Code A above, is easier to encode and decode. The idea emerged from the well-known result that the Reed-Solomon extended (8,4;5) code over $GF(8)$ represented in binary form in a normal basis is isomorphic to the extended Golay (24,12;8) code. The idea in its simpler form was first presented in [1]. This article presents a newer form with a proof; the proof relies heavily on [1].

IV. The Reed-Solomon (8,4;5) Code Over $GF(8)$ is the Golay Code

Let β be a root of the polynomial $g(x) = x^3 + x^2 + 1$. Express the elements of $GF(8)$ in terms of the root β^i, $i = 1, 2, 4$, of $g(x)$. This is a normal basis of the field. Consider the extended Reed-Solomon (8,4;5) code over $GF(8)$ given by the check recursion polynomial $f(x) = \prod_{i=0}^{3}(x + \beta^i)$. The MS polynomial for codeword a is given by

$$P_a(x) = C_0 + C_1x + C_2x^2 + C_3x^3$$

$$x \in GF(8), \quad x = 1,\ \beta,\ \beta^2,\ \beta^3,\ \beta^4,\ \beta^5,\ \beta^6,\ 0$$

$$C_i \in GF(8), 0 \leq i \leq 3$$

Writing the codewords in binary, using the normal basis above, three codewords of length eight emerge as Tr $P(x)\beta^j$, $j = 1, 2, 4$, where Tr denotes the value in $GF(2)$ given by the Trace of an element $a \in GF(8)$, where Tr $a = a + a^2 + a^4$. Note that Tr $\beta^i = 1$, $i = 0, 1, 2, 4$, and Tr $\beta^i = 0$, $i = 3, 5, 6$.

It is easy to show, using the Γ_2 formula of [1], that the total weight of the three binary codewords is a multiple of 4, with a minimum weight equal to 8.

Proof. Set $C_0 = 0$. The binary polynomials may be written as Tr $P(x)\beta^j = \text{Tr}\,(C_1\beta^j + (C_2\beta^j)^4)x + (C_3\beta^j)^2x^6$. The expressions for $\Gamma_2(\text{Tr}\ P(x)\beta^j)$, $j = 1, 2, 4$, are

$$\Gamma_2(\text{Tr}\ P(x)\beta) = \text{Tr}\,(C_1C_3\beta^3 + C_2^4\beta^6)$$

$$= \text{Tr}\,(C_1C_3\beta^3 + C_2^4\beta^6)$$

$$\Gamma_2(\text{Tr}\ P(x)\beta^2) = \text{Tr}\,(C_1C_3\beta^6 + C_2^4\beta^5)$$

$$\Gamma_2(\text{Tr}\ P(x)\beta^4 = \text{Tr}\,(C_1C_3\beta^5 + C_2^4\beta^3)$$

Summing over all three binary codewords Tr $P(x)\beta^j$, $j = 1, 2, 4$, one obtains $\sum_j \Gamma_2(\text{Tr}\ P(x)\beta^j) = 0$. This implies that the binary weight of the above Reed-Solomon codeword is a multiple of 4. Since the minimum symbol weight is 5, the binary code is a (24,12;8) code. Furthermore, this stamps the code as a Golay code, by virtue of the uniqueness of the Golay code.

To decode this, simply use the decomposition of Tr $P(x)\beta^j$, $j = 1, 2, 4$, and test eight values for C_3 and decode the (8,4;4) BCH-Hamming code that remains. Note

that the parity of each Tr $P(x)\beta^j$, $j = 1, 2, 4$, is zero, which yields additional information to detect single errors and thus reduce the search for C_3. Use each of these eight trials to soft decode or maximum-likely decode three BCH-Hamming codes.

V. Code B via the Reed-Solomon Code

Using techniques similar to those above, if one starts with a Reed-Solomon (8,4;5) code over $GF(64)$, and represents the code in binary using a particular normal basis with the special property defined below, one can generate a code of length 48 and dimension 24 with weights that are multiples of 4.

The binary representation of the usual Reed-Solomon (8,4;5) code over $GF(64)$, yields six (8,7;2) codewords whose decomposition into two cyclic code components and a constant component looks like Reed-Solomon (6,4;3) and (6,2;5) codes over $GF(8)$ and a binary (6,6;1) code, respectively. However, Code B is constructed by modifying the extended coding rule for the parity symbol.

In particular, let γ be a root of the polynomial $f(x) = x^6 + x^5 + x^4 + x + 1$, where γ is a primitive generator of the 63 roots of unity. Represent the elements of $GF(64)$ in the normal representation using the roots of $f(x)$. The roots are γ^j, $j \in J$, $J = \{1, 2, 4, 8, 16, 32\}$.

Note that for this particular choice of $f(x)$,

$$\text{Tr } \gamma^j = 1; \quad j \in J; \quad J = 1,2,4,8,16,32$$

$$\text{Tr } \gamma^i\gamma^k = 0; \quad i \neq k; \quad i,k \in J$$

Let β be a root of the polynomial $g(x) = x^3 + x^2 + 1$. Then, β is an element of $GF(8)$, a subfield of $GF(64)$, and $\beta = \gamma^9$.

Now use the recursion or check polynomial $h(x) = \prod_{i=0}^{3}(x + \beta^i)$ to generate a Reed-Solomon (7,4;4) code over $GF(64)$. This means that the initial shift register contains four elements in $GF(64)$ expressed as coefficients in the normal representation above. The cyclic portion of the code is of length seven, but the overall parity symbol, the eighth dimension, is defined differently. Representing the binary code as components Tr $P(x)\gamma^i$, $i = 1, 2, 4, 8, 16, 32$, extends the codes to the eighth coordinates by the rules. The binary value at the row indexed by the ith coordinate is achieved by Tr $C_0\gamma^i$ + Tr $\sum_{j\in J} C_0\gamma^i$.

Thus, for the constant term C_0 with Tr $C_0 = 0$, this behaves like the normal parity symbol, which is a sum over the values of the cyclic code coordinates. Note that this is equivalent to the way the additional six coordinates were defined in Code A above.

The general Mattson-Solomon polynomial of codeword **a**, similar to the Golay codeword over $GF(8)$, is $P_{\mathbf{a}}(x) = C_0 + C_1x + C_2x^2 + C_3x^3$, where $C_i \in GF(64)$ for $0 \leq i \leq 3$ and $x \in GF(8)$. Encode the codeword in its cyclic portion. The extended codeword **a** expressed in terms of the MS polynomial, is

$$\mathbf{a} = P_{\mathbf{a}}(\beta^i); \quad 0 \leq i \leq 6, \quad (P_{\mathbf{a}}(0))$$

Writing the codewords in binary, using the normal basis γ^j, where $j \in J$ above, there are six binary codewords of length eight

$$\text{Tr } P(x)\gamma^j; \quad j = 1, 2, 4, 8, 16, 32$$

where Tr a denotes the value in $GF(2)$ given by the Trace of an element $a \in GF(64)$

$$\text{Tr } a = a + a^2 + a^4 + a^8 + a^{16} + a^{32}$$

Consider one of the six binary words in its Mattson-Solomon setting,

$$\text{Tr } P_{\mathbf{a}}(x)\gamma^j = \text{Tr } (C_0 + C_1x + C_2x^2 + C_3x^3)\gamma^j$$

$$= \text{Tr } C_0\gamma^j + \text{Tr }'[(C_1x + C_2x^2 + C_3x^3)\gamma^j$$

$$+ ((C_1x + C_2x^2 + C_3x^3)\gamma^j)^8]$$

$$\text{Tr }'a = a + a^2 + a^4; \quad a \in GF(8)$$

Set $C_0 = 0$ temporarily, as this does not affect the arguments to follow, and

$$\text{Tr } P(x)\gamma^j = \text{Tr }'(C_1\gamma^j + (C_1\gamma^j)^8 + (C_2\gamma^j)^{32}$$

$$+ (C_2\gamma^{32j})^8)x + \text{Tr }'((C_3\gamma^j)^2 + (C_3\gamma^j)^{16})x^6$$

Lemma. The coefficient of x is a (6,4;3) code over $GF(8)$. The coefficient of x^3 is a (6,2;5) code over $GF(8)$.

The code is indexed by the values of γ^j, $j \in J$, $J = \{1, 2, 4, 8, 16, 32\}$.

Proof. The set γ^j, $j \in J$, $J = \{1, 2, 4, 8, 16, 32\}$ is a linear independent set, and thus can only take zero values one less than the number of terms in the coefficients of x and x^6. The components Tr $C_0\gamma^j$ for Tr $C_0 = 0$ form a binary (6,5;2) code.

Theorem. The Reed-Solomon code determined by codewords with MS polynomials $P_a(x)$, when Tr $C_0 = 0$, forms a binary (48,23;12) code with weights that are multiples of 4.

Proof. The property of the weights that are multiples of 4 follows from using the Solomon-McEliece Γ_2 formula

$$\text{Tr } P(x)\gamma = \text{Tr } ((C_1\gamma + (C_2\gamma)^4 + (C_3\gamma^2 x^6))$$

where Tr is defined in $GF(64)$. Now

$$\begin{aligned}\Gamma_2(\text{Tr } P(x)\gamma) = \text{Tr } [&C_1C_3^2\gamma^3 + C_1^8C_3^2\gamma^{10} + C_1C_3^{16}\gamma^{17} \\ &+ C_1^8C_3^{16}\gamma^{24} + C_2^{32}C_3^2\gamma^{34} + C_2^4C_3^2\gamma^6 \\ &+ C_2^{32}C_3^{16}\gamma^{48} + C_2^4C_3^{16}\gamma^{20}]\end{aligned}$$

for

$$\begin{aligned}\Gamma_2(\text{Tr } P(x)\gamma^2) = \text{Tr } [&C_1C_3^2\gamma^6 + C_1^8C_3^2\gamma^{20} + C_1C_3^{16}\gamma^{34} \\ &+ C_1^8C_3^{16}\gamma^{48} + C_2^{32}C_3^2\gamma^5 + C_2^4C_3^2\gamma^{12} \\ &+ C_2^{32}C_3^{16}\gamma^{33} + C_2^4C_3^{16}\gamma^{40}]\end{aligned}$$

Similarly,

$$\begin{aligned}\Gamma_2(\text{Tr } P(x)\gamma^4) = \text{Tr } [&C_1C_3^2\gamma^{12} + C_1^8C_3^2\gamma^{40} + C_1C_3^{16}\gamma^5 \\ &+ C_1^8C_3^{16}\gamma^{33} + C_2^{32}C_3^2\gamma^{10} + C_2^4C_3^2\gamma^{24} \\ &+ C_2^{32}C_3^{16}\gamma^3 + C_2^4C_3^{16}\gamma^{17}]\end{aligned}$$

and, therefore, $\sum_{j\in J} \Gamma_2(\text{Tr } P(x)\gamma^j) = 0$. Recall that the normal basis was chosen so that Tr $\gamma^j = 1$, $j \in J$, and Tr $\gamma^3 =$ Tr $\gamma^5 =$ Tr $\gamma^9 = 0$.

It has been demonstrated that the binary weight of any codeword in the Reed-Solomon code above is a multiple of 4. Since the symbol distance of the code is greater than 5, the weights have been narrowed down to 8, 12, 16, 20, ..., 40.

VI. Structure of the Code

An examination of the Reed-Solomon code in its cyclic components reveals that most code symbols weights are 3, 5, and 6. The code has weight 3 when $C_3 \neq 0$ and C_1 and $C_2 = 0$. Clearly, the binary weight of the code is ≥ 12, independent of the value of C_0. From above, it is clear that the dimension 18 portion of the code has a minimum distance of 12. In fact, the codewords have weights 12 through 36 in multiples of 4. Now, consider C_0 by itself when Tr $C_0 = 0$. Adding this to the dimension 18 code addressed above, the minimum distance is kept at 12. If the symbol weight of the Reed-Solomon code is 5 or 6, then the binary weight of the dimension 18 code is ≥ 12, and no complementing can reduce this weight. For symbols of weight 3, one must have three codewords of weight 4 each; again, complementing does not reduce the weight.

The extension parity rule for C_0 has been changed as follows: if, say, for $i = 1$, Tr $P(x)\gamma^i = 1$ and Tr $P(x)\gamma^i = 0$, $i \neq 1$, an eighth row of weight 5 is adjoined, e.g., 011111. The argument invoked in the first code above may be used to prove that the minimum code distance is 12 and all words have weight multiples of 4, i.e., the code is self-dual.

VII. Decoding Binary

To correct five or more errors, begin by trying 64 different values of C_3. This eliminates the cyclic component attached to x^6. Since each of the six binary codes is now an odd-error-detecting, single-error-correcting code, the parity information is used to correct single errors when they occur. In the case of a five binary error pattern, at least four correct values of C_1 and C_2 are available. Therefore, in $\binom{6}{2}$ trials, a total of at most $15 \times 64 = 1024 - 64 = 960$ candidates is available to find the most likely error pattern.

Note that five-error patterns occur in rows as 5, 4 1, 2 3, 1 2 2, 1 3 1, 1 1 2, and 1 1 1 1 1. Six-error patterns occur in rows as 6, 5 1, 4 2, 3 2 1, 3 3, 3 1 1 1, 2 2 2, 2 2 1 1, and 2 1 1 1 1. Using this technique, the only error pattern that would not emerge as a candidate would be the 2 2 2 case. This occurs in 560 patterns out of the

possible $\binom{48}{6}$ error patterns. The number of uncorrectable, but detectable, six-error patterns is given by $\binom{12}{6}$ times the number of codewords of weight 12.

A soft decoding using these techniques would perhaps add a few decibels to the hard-decision decoding performance.

Reference

[1] G. Solomon, "Golay and Other Box Codes," *TDA Progress Report 42-109*, vol. January–March 1992, pp. 130–135, May 15, 1992.

More Box Codes

G. Solomon
Communications Systems Research Section

A new investigation shows that, starting from the BCH (21,15;3) code represented as a 7 × 3 matrix and adding a row and column to add even parity, one obtains an 8 × 4 matrix (32,15;8) code. An additional dimension is obtained by specifying odd parity on the rows and even parity on the columns, i.e., adjoining to the 8 × 4 matrix, the matrix, which is zero except for the fourth column (of all ones). Furthermore, any seven rows and three columns will form the BCH (21,15;3) code. This box code has the same weight structure as the quadratic residue and BCH codes of the same dimensions. Whether there exists an algebraic isomorphism to either code is as yet unknown.

I. Constructions

A box (32,16;8) code with the weight distribution of the BCH or quadratic codes of the same length is constructed here in a manner similar to the way in which the extended Golay (24,12;8) code was constructed [1]. The extended Golay (24,12;8) code was generated as a 6 × 4 binary matrix from the BCH-Hamming (15,11;3) code, represented as a 5 × 3 matrix, by adding a row and a column that are both of odd or even parity. The odd-parity case gave the additional twelfth dimension. Furthermore, any three columns and five rows of the 6 × 4 Golay code form a BCH-Hamming (15,11;3) code. The construction parallels [1] without going through the details. The original article contains the proofs.

The (32,16;8) code in 8 × 4 matrix form is obtained from the BCH-Hamming (21,15;3) code by adjoining row and column even parity. The BCH code is expressed here as a 7 × 3 matrix with entries in the $(i,\ j)$ positions, $0 \leq i \leq 6,\ 0 \leq j \leq 2$, corresponding to the coordinates $7i + 3j$ mod 21 of the code.

Let $\mathbf{A}$ be the BCH-Hamming (21,15;3) code. The Mattson-Solomon polynomial for a codeword $\mathbf{a} \in \mathbf{A} = (a_i,\ i = 0 \ldots 20)$ is given by

$$P_{\mathbf{a}}(z) = C_0 + \mathrm{Tr}\ Cz + \mathrm{Tr}'\ Dz^3 + \mathrm{Tr}'\ Gz^{-3} + Ez^7 + E^2 z^{14}$$

where $C \in GF(64)$, G and $D \in GF(8)$, $E \in GF(4)$, and $C_0 \in GF(2)$.

Here, $P_{\mathbf{a}}(\beta^i) = a_i$, for β as a primitive 21st root of unity. Tr denotes the linear operator Trace in $GF(64)$, i.e., $\mathrm{Tr}\ a = a + a^2 + a^4 + a^8 + a^{16} + a^{32}$. Tr′ denotes the linear operator Trace in $GF(8)$, i.e., $\mathrm{Tr}'\ a = a + a^2 + a^4$.

The parity check polynomial for the code is $(z+1)f_1(z)$ $f_3(z)f_7(z)f_{-3}(z)$, where $f_i(z)$ is the irreducible polynomial over $GF(2)$ with β^i as a root

$$f_1(z) = z^6 + z^4 + z^2 + z + 1$$

$$f_3(z) = z^3 + z^2 + 1$$

$$f_{-3}(z) = z^3 + z + 1$$

$$f_7(z) = z^2 + z + 1$$

The weight, $w(\mathbf{a})$ mod 4 for even-weight words $\mathbf{a}(C_0 = 0)$, is given by $w \bmod 4 = 2\Gamma_2(P_{\mathbf{a}}(x))$, where $\Gamma_2(P_{\mathbf{a}}(x)) = DG + (DG)^2 + (DG)^4 + E^3$ [2].

Now place the codewords in 7×3 matrices (b_{ij}), $0 \leq i \leq 6$ and $0 \leq j \leq 2$, corresponding to their values $7i + 3j$ mod 21. The ith coordinate is entered thusly,

$$\begin{pmatrix} 0 & 7 & 14 \\ 3 & 10 & 17 \\ 6 & 13 & 20 \\ 9 & 16 & 2 \\ 12 & 19 & 5 \\ 15 & 1 & 8 \\ 18 & 4 & 11 \end{pmatrix}$$

The MS polynomial expressed in the 7×3 setting, indexing each row by y in terms of the independent variable x, becomes

$$\text{Tr } Dy^3 + \text{Tr } Gy^{-3} + \text{Tr}' (E' + Cy + C^4y^4 + c^{16}y^2)x,\ E' = E^2$$

Note again that in the MS polynomials of each row, the Trace is defined over $GF(4)$ as $\text{Tr}' a = a + a^2$ for $a \in GF(4)$, and is defined over $GF(8)$ as $\text{Tr } Dy^3 = (Dy^3) + (Dy^3)^2 + (Dy^3)^4$ and $\text{Tr } Gy^{-3} = (Gy^{-3}) + (Gy^{-3})^2 + (Gy^{-3})^4$.

Form the sum over the rows to give an eighth row with an MS polynomial of $\text{Tr}' E'x$. Form the parity sum over the columns to obtain an 8×1 column that is, of course, $\text{Tr } Dy^3 + \text{Tr } Gy^{-3}; y^8 = y$. The bottom row is indexed by $y = 0$, and the parity column corresponds to $x = 0$.

Note that the coefficient of x, $E' + Cy + (Cy)^4 + (Cy)^{16}$, is the MS polynomial for an (8,4;4) code indexed by $y^8 = y$ over $GF(4)$. Note that the constant term in each row varies and is a binary (7,6;2) code. The constant term contributes the same values to the fourth parity column. Thus, if one started with a BCH subcode of dimension 14 of even-weight w with $w \bmod 4 = 2\Gamma_2$, where $\Gamma_2 = TrDG + E^3$, when one adjoins the parity rows and columns, one is adjoining row and column codewords whose weight modulo 4, $w \bmod 4 = TrDG + E^3$. So the total new weight $w' = 0 \bmod 4$.

This proves that $w' \geq 8$. For if $w = 4$ originally, now either $E^3 = 1$ and Tr $DG = 1$, adding weight 4, or $E = 0$ and TR $DG = 0$, adding a column of weight 4.

One could also show easily that $w' \geq 8$ by noting that the coefficient of x is now an (8,4;4) code over $GF(4)$, having adjoined an even-parity row. Thus, there are at least four rows each of weight 2. The addition of the even-parity column ensures that $w \geq 8$ when the coefficient of $x = 0$. The new codewords have weights of 8, 12, 16, 20, 24, and 32 in the 8×4 matrix code that was generated. Complementing these new codewords still gives words with weights of 8, 12, and 16, which accounts for the odd-weight BCH codewords adding up to dimension 15.

The 16th dimension of the constructed code is achieved by adding an odd-parity row and an even-parity column to the BCH words.

II. A Startling Property

Theorem. Consider the 8×4 binary matrix and consider any 7×3 submatrix obtained by removing one column and one row. This is the BCH (21,15;3) code.

Proof. Consider the 8×4 matrix with the top row deleted. Using the bottom parity check row and considering the first three columns, one can now show a permuted 7×3 BCH code where the rows have been interchanged.

The coefficient of x, the (8,4;4) code over $GF(4)$, gives rise to the (32,8;8) portion of the code. The map of $y \rightarrow (1 + y)$ is a permutation of this code that interchanges the top and bottom rows, corresponding to $y = \beta^0 = 1$ and $y = 0$.

The remaining five dimensions, which are functions of C_0 and D in the BCH-Hamming code, are easily manipulated so that the weights stay the same. Since the code is clearly invariant under cyclic row permutations, this takes care of all subcodes with the first three columns fixed.

Now interchange the first column with the fourth rightmost parity column, and the second column with the third

column, in order to obtain a BCH code like the one above in the top seven rows. This interchange of columns is given by $x \rightarrow x + 1$.

This map takes the row indexed by y, Tr Dy+Tr' $(E'+Cy+(Cy)^4)x$ into a permuted row indexed by y, where $D \in GF(16)$ has been augmented. Tr' $(E'+Cy+(Cy)^4)+$ Tr Dy + Tr' $(E' + Cy + (Cy)^4)x$. There clearly exists a value of D', such that $D' =$ Tr' $(E'+Cy+(Cy)^4)+$Tr Dy for all values of y. Now, clearly, every three columns that occurred in the leftmost 5×3 matrix now occur in the newly formed 5×3 matrix. As the code is invariant under cyclic column permutation, the proof is completed.

References

[1] G. Solomon, "Golay and Other Box Codes," *TDA Progress Report 42-109*, vol. January–March 1992, pp. 130–135, May 15, 1992.

[2] G. Solomon and R. J. McEliece, "Weights of Cyclic Codes," *Journal of Combinatorial Theory*, vol. 1, no. 4, pp. 459–475, December 1966.

Fault Detection Using a Two-Model Test for Changes in the Parameters of an Autoregressive Time Series

P. Scholtz and P. Smyth
Communications Systems Research

This article describes an investigation of a statistical hypothesis testing method for detecting changes in the characteristics of an observed time series. The work is motivated by the need for practical automated methods for on-line monitoring of DSN equipment to detect failures and changes in behavior. In particular, on-line monitoring of the motor current in a DSN 34-m beam waveguide (BWG) antenna is used as an example. The algorithm is based on a measure of the information theoretic distance between two autoregressive models: one estimated with data from a dynamic reference window and one estimated with data from a sliding reference window. The Hinkley cumulative sum stopping rule is utilized to detect a change in the mean of this distance measure, corresponding to the detection of a change in the underlying process. The basic theory behind this two-model test is presented, and the problem of practical implementation is addressed, examining windowing methods, model estimation, and detection parameter assignment. Results from the five fault-transition simulations are presented to show the possible limitations of the detection method, and suggestions for future implementation are given.

I. Introduction

The motivation for this study is on-line performance monitoring of DSN electromechanical and hydraulic equipment, particularly the pointing system components of the DSN 70-m and 34-m antennas. Previous articles [1–3] have described in detail the motivation behind on-line monitoring: Essentially as the antennas get older and deep space missions are of longer duration, early detection of component failures becomes more critical.

Simple thresholding methods, whereby detection of change occurs when the magnitude of the observed signal exceeds prespecified alarm limits are available in commercial off-the-shelf products and widely used for on-line monitoring. However, the simple thresholding approach is nonadaptive and may be susceptible to false alarms in the presence of noise. In this article more sophisticated methods, which not only account for changes in the signal amplitude, but can also detect changes in the underlying statistical characteristics of the signal in cases where no amplitude change is observable, are investigated.

Various methods which detect changes in the mean of a signal directly by utilizing statistical cumulative sum

(cusum) schemes have been thoroughly examined in [4–7]. These methods generally work well when there is sufficient prior knowledge about the magnitude of change. Mean-change detection algorithms have been well-defined for observations which consist of independent, identically distributed Gaussian random variables; however, when there is a significant time correlation in the observations (as is almost always the case in applications of interest), the usefulness of these methods is diminished, and other techniques must be utilized.

One promising technique for detecting parametric changes in the model of the process is well documented in [8,9]. This method computes a cumulative sum based on the information-theoretic distance between two estimated autoregressive (AR) models of the process. The focus of this article is to summarize the theory underlying this cusum algorithm and to examine the problems and necessary adaptations for practical implementation in a DSN application.

In choosing the two-model algorithm for implementation, a test was desired with the following properties:

(1) Few (not necessarily minimal) false alarms.

(2) Robustness with respect to variance in model estimates.

(3) Few a priori assignments of detection parameters and the use of on-line estimated parameters.

(4) Symmetry with respect to transitions from both low-to-high and high-to-low signal variance, and vice versa, and with respect to changes in the AR parameters.

The two-model algorithm described in this article utilizes two filters, a growing memory AR model and a local memory AR model, which are implemented by using two data windows: a reference window of growing length and a fixed-length window. A statistic based on conditional Kullback information measures the distance between the two models based on the innovations from both filters. The crossing of a threshold by the cumulative sum of this statistic detects the distribution change. An advantage of this technique is that it will detect a change in the actual AR parameters of the process or a change in the energy of the signal. Section II of this article presents the AR model change hypothesis. Section III focuses on the theoretical derivation of the cumulative sum, while Section IV examines the problems for implementing the algorithm for DSN applications. Results of simulated fault detection are discussed in Section V. Limitations of the method are examined in Section VI, and Section VII compares the method with an alternate approach using hidden Markov models.

II. Autoregressive Model Change Hypothesis

Suppose the observed signal (Y_n) can be modeled by an autoregressive process of order p, that is

$$Y_n = \theta_p \overline{Y}_{n-p}^{n-1} + \epsilon_n \tag{1}$$

where $\overline{Y}_{n-p}^{n-1}$ is the column vector of the past p values of the observation signal, θ_p is the AR(p) parameter vector, and the first term on the right is a vector dot product. It is well established in time series analysis that any stationary signal (or any piecewise stationary signal) can be modeled as an AR model of sufficient (finite) order, plus a deterministic term. To test for change, assume there exists a time r such that, for $n \leq r$

$$\theta_p = \theta_p^0 \quad \text{and} \quad \sigma_{\epsilon_n}^2 = \sigma_{\epsilon_n^0}^2 \tag{2}$$

and for $n > r$

$$\theta_p = \theta_p^1 \quad \text{and} \quad \sigma_{\epsilon_n}^2 = \sigma_{\epsilon_n^1}^2 \tag{3}$$

where $\sigma_{\epsilon_n}^2$ is the variance of the prediction error ϵ_n at time n. These changes in AR parameters reflect an underlying change in the probability laws governing the process. The problem is to test sequentially at each time n, the null hypothesis H_n^0 of no change in the probability law of the process against the alternate hypothesis H_n^1 that the above change in probability laws occurs at a time $r < n$.

III. Cumulative Sum Detection Statistic

The most obvious test for difference between two models is the likelihood ratio test. This tests the null hypothesis H_n^0 that all observations up to time n follow the *joint* probability law $p^0\left(Y_n, \overline{Y}^{n-1}\right)$ against the hypothesis that all observations up to time n follow $p^1\left(Y_n, \overline{Y}^{n-1}\right)$ [8]. However, the sequential test desired for application is the *conditional* hypothesis test, which tests the null hypothesis H_n^0 against the alternate hypothesis H_n^1 that the probability distribution changes from p^0 to p^1 at time $r < n$ (as stated in the introduction). Hence, letting y denote the value of the variable Y_k, the detection test is the sequential cumulative sum test:

undergoes a decrease from an initial mean of zero, reflecting the increasing difference in the two models after the change in distributions; hence a one-sided Hinkley test is used. Detection occurs at the first time when $D > h$ for a preset threshold h. For recursive computation on line, the value

$$U_n = U_{n-1} + T_n - \delta \tag{11}$$

can be used [9]. Implementation of this scheme will be discussed in the next section.

IV. Implementation of the Two-Model Method

A. Model Order

Prior work modeling the elevation pointing system motor current from the DSS-13 34-m BWG antenna led to an autoregressive-exogenous input (ARX)(5,3,1) model as a first approximation to the proper model [1]; this model is defined by 3 poles, 5 zeros, and 1 delay in the transfer function. The order of this model was determined using Akaike's information criterion. Further work has shown that comparing the model parameters of the ARX(5,3,1) to AR(5) yields no significant difference in the estimated autoregressive coefficients. Hence, although the model underlying the process has yet to be properly identified, it appears that an AR(5) model may be sufficient for change detection purposes. However, it should be noted that since the detector relies directly on the goodness-of-fit of the model, the opportunity exists to potentially improve the model identification method in order to reduce false alarms due to incorrectly estimated prediction errors.

B. Window Implementation

There are two well-documented approaches to block window implementation of the two-model detection algorithm. The first relies on a fixed-length reference window which is allowed initially to stabilize in a controlled period of normal operation in order to estimate the AR(p) reference model. This model is then compared with the estimated model in a moving fixed-length window, with detection when the distance between these two models becomes sufficiently large. In the past, the difference between these two models has been measured by the mean quadratic difference between the two spectra [14]. Basseville and Benveniste [8] report that the disadvantages of this approach are a large variance in the metric and an asymmetrical test for increases or decreases in signal noise.

A more robust, dynamical window system employs a growing-memory window for reference along with the fixed-length sliding window. First a growing reference window is allowed to attain a stable model. As soon as the window stabilizes, a shorter fixed-length sliding window begins to move along the time series with the reference window. At each time n, a model is estimated in each data window. When a transition occurs in the spectrum of the signal, the abrupt change is reflected in the local window, while the reference model remains relatively unchanged due to its long memory. The information metric between these two windows is measured by the statistic T of Eq. (9), which is integrated in the cumulative sum U of Eq. (8); then the Hinkley decision rule given by Eq. (10) is invoked. The use of this growing reference window instead of a static reference window greatly reduces the rate of false alarms by adapting to the dynamic nature of the system [8].

C. Model Estimation

The first attempt at implementing the algorithm employed a dynamic *block* window scheme for model estimation. Sequentially, at each time n an AR(5) model was estimated for both the growing window and the sliding window. Treating each window as a batch, each model was estimated using the "forward–backward" AR estimation algorithm and then the cusum was computed based on the model errors. As expected, choosing an appppropriate window size for the fixed-length window was critical in the implementation of the algorithm. An undersized window leads to unstable model estimation, creating largely varying prediction errors which lead to false alarms. On the other hand, oversized local windows lead to longer delays until detection and more computation. A window size of 200 to 400 data points (approximately 4 to 8 seconds at the sampling rate of 50 Hz) yielded a reasonable fit. Below a window size of 200, the AR model was highly unstable and produced unreliable results.

From a practical standpoint, this nonrecursive algorithm was computationally intensive and infeasible to implement on-line.

A second attempt utilized a recursive algorithm for model estimation. The Normalized Gradient approach of the recursive least mean squares (LMS) parameter estimation was used to fit the AR(5) model [15]. For the linear regression

$$\hat{Y}_n = \overline{Y}^{n-1}\hat{\theta} \tag{12}$$

the parameter estimate $\hat{\theta}_n$ is

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \gamma \overline{Y}^{n-1} \epsilon_n \tag{13}$$

where γ is the (constant) gain. For the normalized approach, γ is replaced by

$$\gamma' = \frac{\gamma}{|\overline{Y}^{n-1}|^2} \tag{14}$$

This is the same method employed by Eggers and Khuon [9] with their recursive LMS learning-parameter method, which they showed will converge to give the true AR coefficients. The dynamic windowing method described above is implemented by choosing the gains γ_1, γ_2 corresponding to the reference window and the fixed-size window, respectively, such that the model is weighted to reflect the information content of the most recent observations. Hence, the gains are chosen such that $0 < \gamma_1 < \gamma_2$, reflecting the abrupt change in spectral characteristics in the local model while leaving the reference model relatively unchanged [9]. The gain is usually chosen in the range 0.001 to 0.02 [15]. In the work reported here, γ_1 was chosen to be 0.001, while γ_2 was set to 0.02, for maximum distinction between the two models. The gains must be bounded by $1/tr(\overline{R}) = 1/pr_0$, where $\overline{R}$ is the autocorrelation matrix of the last $p-1$ observations, p is the AR(p) model order, and r_0 is the autocorrelation sequence at lag 0. For example, one can obtain an estimate of this bound on-line by estimating the autocorrelation sequence r_0 after n samples by

$$\hat{r}_0^{YY} = \frac{1}{n} \sum_{k=1}^{n} |Y_k|^2 \tag{15}$$

as shown in Marple [16]. However, in the work presented here, γ_1 and γ_2 were fixed as described above for simplicity. Further investigation is required to determine a reasonable way to preset the gains for this approach.

Other recursive methods were considered, such as the recursive AR algorithm utilizing a Kalman filter scheme. However, these methods require more prior information, such as a knowledgeable guess of the covariance of the innovations. The gradient approach utilized here, as shown in Eqs. (12) through (14), is a logical choice for the case where there is little a priori information. A possible alternative estimation method is the fast recursive least-squares algorithm of Ljung, as described by Marple [16]. This algorithm employs a time-varying gain for the normalized gradient approach resulting in lower MSE than the recusive LMS algorithm, with approximately the same computational efficiency.

D. Choice of the A Priori Detection Parameters

In on-line implementation, it was difficult to choose appropriately the a priori parameters h (threshold) and δ (drift) for Eqs. (8), (9), and (10). The most difficult detection parameter to assign is the drift δ of T_n. In cases of small changes in the signal energy or AR parameters, the cusum is particularly sensitive to the choice of δ. Hinkley [4] determined the value of δ to be

$$\delta = \frac{1}{2}(\mu_1^* - \hat{\mu}_0) \tag{16}$$

where $\hat{\mu}_0$ is the estimated initial mean, and μ_1^* is the minimum expected final mean of the process. In order to assign δ, there must be prior knowledge of the expected behavior of T_n, and therefore prior knowledge of the expected faults.

It should be noted that the cumulative sum can be run with an assigned drift bias equal to zero. Instead of using the Hinkley stopping rule of Eq. (10) to detect a large deviation from the maximum of the cusum, the decision to stop would occur when the cusum U given by Eq. (8) passes a set threshold value h'. However, this merely further complicates the problem of setting the decision threshold.

When the Hinkley stopping rule of Eq. (10) is utilized, the threshold h must be assigned a priori. In this experiment, a constant value for h was chosen, specific to the fault transition under examination. However, for practical implementation when multiple unknown changes can occur, a constant threshold does not appear feasible. Certainly a dynamic threshold h based on the variance of T seems appropriate. In fact, in another application Basseville and Benveniste [17] suggest a threshold

$$h = c\frac{\hat{\sigma}_n^2}{\delta} \tag{17}$$

where $c > 0$ is a constant, $\hat{\sigma}_n^2$ is the variance of the desired random variable (in this case T_n) estimated on-line, and δ is assigned the value of one-half the minimum expected magnitude of change in the mean. The usefulness of this threshold has not yet been examined in this context.

E. Signal Energy Change Considerations

As can be seen in the next section, low- to high-energy transitions in the signal are easily detected with the use of the detection statistics of Eqs. (8), (9), and (10). However, when the signal energy drops at the transition, detection is more difficult. By running a second test in parallel this obstacle is overcome. This alternate test requires switching the innovations and variances of model 0 in Eq. (9) with those of model 1 [9]. The drift in T_n is more delineated in this alternate test, allowing for better behavior of the cusum, and thus better detection.

V. Results

A. Simulated Fault Data

As mentioned previously, the motivation in examining various cumulative sum tests was to find a failure detection test which was feasible for on-line implementation. In a previous experiment [1], on-line readings from various sensors on the DSS-13 34-m BWG elevation pointing system were collected, both under normal conditions and under five hardware faults which were introduced in a controlled manner. The proposed change detection method (as described above) was tested on this data set. The introduced hardware faults were

(1) Noise in the tachometer feedback path.

(2) Tachometer failure.

(3) Torque share/bias loss to motor.

(4) Integrator short circuit.

(5) Rate loop compensation short circuit.

The simulation of these failures is described in greater detail in [1]. Of the 12 sensors on the control system, the most easily modeled with a time series was the motor current, and this was the sensor data selected for this study. In the past, a pattern recognition approach utilizing a hidden Markov model had been used to attempt detection of these five simulated fault transitions, with very accurate results [3]. However, this method requires training data for each fault a priori. In contrast, the purpose of the study reported here was to investigate alternative techniques which do not necessarily require training data for a set of faults which are known a priori. Nonetheless, as pointed out earlier, the two-model approach investigated here does in fact require some prior knowledge of the fault characteristics in terms of setting the drift parameters and choosing the order of the model.

B. Data Preprocessing

The raw data from the motor current sensor had a significant amount of sensor noise and outliers due to sensor faults. First, linear detrending was performed over the entire sample to remove low-level linear trends. Then the raw data was bandpass filtered with a tenth-order Butterworth filter to remove any further outliers. The passband of the filter was from 0.5 to 10 Hz for effective smoothing.

C. Fault Transition Detection

In the current detection experiment, transitions from normal to fault conditions were examined for all five faults. Data were compiled with normal conditions for the first 1000 samples, and with the desired fault conditions for the remaining 1000 data samples. At a 50-Hz sampling rate, this represents only approximately 2/3 of one minute of real data since the data records used were shortened for computational reasons. Transitions to faults 1 and 2 increased the signal energy, transitions to faults 3 and 4 preserved the signal energy, and transitions to fault 5 decreased the signal energy. Also, fault 1 conditions were scaled to equal energy with normal conditions to ascertain whether the test could detect a change in AR parameters without a signal energy change.

A summary of the important detection parameters is presented in Table 1. For this table, the AR parameters of each process were estimated using a block window AR model algorithm with window size 200. Figure 1 shows a typical fault transition statistic T and cusum U for a low to high energy transition. Note the decrease in mean of T at the time of jump, sample number 1001, and the drop in cusum U from its maximum near transition. Figures 2 through 7 present the fault transitions examined and the behavior of the cusum U prior to detection. The filtered signal shown in the top view of these figures is proportional to motor current. The cusum U (dimensionless) shown is equal to zero for the first 200 points, corresponding to the initial model stabilization period, and then varies with time until detection, when it is reinitialized to a value of zero.

1. Faults 1 and 2. Faults 1 and 2 are the simplest to detect; they have both large increases in signal energy and distinct changes in the AR parameters θ. Figure 2 shows the behavior of the filtered signal and cusum U for a transition to fault 1. Detection occurs at sample 1093, a delay of 93 sample points (or about 2 seconds), which should be acceptable for DSN operational requirements. Similarly, the transition to fault 2 conditions is detected with a delay of only 11 points (0.3 seconds) (Fig. 3).

2. Faults 3 and 4. Faults 3 and 4, the bias loss and integrator short circuit, show little deviation from nominal conditions in the AR parameters or signal energy, as seen in Table 1. Although components were physically removed or altered, no significant change in sensor data or pointing performance was observed due to the redundancy in the system. Figures 4 and 5 show the signal and appropriate cusum for these transition detectors. Recall that the fault transition occurs at sample point 1001; clearly there is no visible change in the signal. As might be expected, the transition was undetected for both faults.

3. Fault 5. The compensation short circuit, fault 5, is a transition from high to low signal energy, with changing AR parameters. For this type of transition, the alternate T_n statistic described in Section IV.E is used, which increases the model distance measure for better detection. Figure 6 shows the signal under consideration and the alternate cusum U'. In this case, detection is possible and occurs almost instantaneously at transition, with a delay of only 8 sample points. However small this delay may be, it should be noted that in this case detection is highly sensitive to the choice of the threshold h; increasing the threshold by a small amount may lead to a missed detection, while decreasing the threshold by a small amount may lead to multiple false alarms. Thus, it is clear that the algorithm does detect faults that cause decreases in signal energy, but with some limitations.

4. Scaled Fault 1 Case. A simulation was run using a transition to fault 1 conditions which were scaled by 0.25 to match the signal energy of the nominal data. This test was performed to determine whether a change in AR parameters *without* a change in signal noise could be detected. Figure 7 shows the signal transitions and cusum behavior for this case; detection occurred at sample 1151, a reasonable delay of 150 points (3 seconds). Again there is some doubt as to exactly how small a transition can be detected, but at least the test works with the appropriate choice of parameters.

VI. Limitations of the Approach

A. Prior Knowledge Requirements

Although not conclusive, the results of the five fault detection simulations point out some limitations of the two-model method. First, some knowledge of the manner in which faults will affect the AR parameters describing the data is required. Since this detector relies on the prediction errors of the AR model, prior knowledge of the model order for both normal and fault conditions is required. Moreover, this method is not designed to detect changes in the model order, which may occur for fault transitions. Such order changes could be detected by a more sophisticated algorithm which dynamically tries to fit multiple-order models to the data—however, this would be both computationally intensive and potentially difficult to stabilize.

B. Parameter Assignment

A more detailed examination of the preassigned parameters h (threshold) and δ (drift bias), which are fault-specific parameters, would have to be conducted to determine a systematic method of assignment. A global time-varying threshold for all the parallel tests based on the variance of T_n, as given by Eq. (17), is the logical candidate for improving the parameter h. The drift bias δ, corresponding to a change in the mean of the statistic T, can be difficult to assign. Furthermore, the detection test is highly sensitive to the choice of δ, so proper assignment is critical. However, with a larger record of available data, a small number of trials for each expected fault type is expected to yield feasible values of δ for on-line implementation.

C. False-Alarm Rate

The false-alarm rate has not been derived analytically for the recursive AR estimation approach for the two-model approach. For the block window implementation as presented in Section IV.C, the rate of false alarms was derived by Basseville in [8] and is equal to the inverse of the expected value of the detection time D_h

$$E(D_h) = \frac{1}{\delta}\left[\frac{2}{\delta N}(e^{2Nh} - 1) - h\right] \tag{18}$$

where N is the size of the fixed-length block window and δ is chosen as a function of the distance of the two probability laws describing the process before and after the change. However, the exact relationship between the block window model estimation and the recursive normalized-gradient model estimation is not yet clear, so an analytic expression for the false-alarm rate has not yet been derived. In any case, derivation of false-alarm rates for these types of models requires a complete model of the fault to be detected and, thus, is of questionable utility for practical purposes.

VII. Comparison to the Hidden Markov model method

In [2,3] a hidden Markov model (HMM) method for fault detection was reported. The HMM method assumes that the set of faults is known in advance and training data are required for each fault. However, once trained the model is quite robust and does not require any additional parameters to be set or calibrated. On the other hand, the two-model method described here does not specifically require training data in advance; rather, some prior knowledge about the possible faults is required. It would appear that if training data are available the HMM method is more robust and accurate as a detector than the two-model hypothesis testing approach. However, since training data for specific faults are unlikely to be available in many applications of interest (particularly at the 70-m antennas where experimentation is difficult due to operational commitments), methods such as the two-model approach may be a more practical alternative in the long run. Hybrid models which combine the better features of each approach are worth further investigation.

VIII. Conclusions

The use of a two-model cumulative sum detection scheme has been investigated for on-line detection of faults in a dynamic system. This method involves detecting the change in the mean of a function defined on the prediction errors of two recursive AR models estimated sequentially on-line. The algorithm detects changes in the AR model parameters or the signal energy. Experimental results were presented for this two-model method for five controlled hardware faults at the DSS 13 34-m BWG antenna control assembly. The following conclusions can be made in summary:

(1) This method is feasible if prior knowledge of faults is available.

(2) This method can be sensitive to parameter choices. Thus, making more robust detectors which require fewer parameter choices and prior assumptions would be useful.

References

[1] J. Mellstrom and P. Smyth, "Pattern Recognition Techniques Applied to Performance Monitoring of the DSS-13 34-m Antenna Control Assembly," *TDA Progress Report 42-106*, vol. April–June 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 30–51, August 15, 1991.

[2] J. Mellstrom, C. Pierson and P. Smyth, "Real-Time Antenna Fault Diagnosis Experiments at DSS-13," *TDA Progress Report 42-108*, vol. October–December 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 96–108, February 15, 1992.

[3] P. Smyth, "Hidden Markov Models for Fault Detection in Dynamic Systems," submitted to *Pattern Recognition*, May 1992.

[4] D. Hinkley, "Inference About the Change Point From Cumulative Sum Tests," *Biometrika*, vol. 58, no. 3, pp. 509–523, 1971.

[5] M. Basseville, "On-Line Detection of Jumps in Mean," *Detection of Abrupt Changes in Signals and Dynamical Systems*, edited by M. Basseville and A. Benveniste, Berlin: Springer-Verlag, 1986.

[6] M. Basseville, "Detecting Changes in Signals and Systems—A Survey," *Automatica*, vol. 24, no. 3, pp. 309–326, 1988.

[7] M. Basseville, "Edge Detection Using Sequential Methods for Change in Level—Part II: Sequential Detection of Change in Mean," *IEEE Trans. Acoust. Speech Sig. Process.*, vol. ASSP-29, no. 1, pp. 32–50, February 1981.

[8] M. Basseville and A. Benveniste, "Sequential Detection of Abrupt Changes in Spectral Characteristics of Digital Signals," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 5, p. 709–723, September 1983.

[9] M. Eggers and T. Khuon, *Adaptive Preprocessing of Nonstationary Signals*, Technical Report 849, MIT Lincoln Laboratory, Cambridge, Massachusetts, 1989.

[10] S. Kullback, *Information Theory and Statistics*, New York: Wiley and Sons, 1959.

[11] J. Segen and A. Sanderson, "Detecting Changes in a Time-Series," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 2, pp. 249–254, March 1980.

[12] M. Basseville, "The Two Models Approach for the On-Line Detection of Changes in AR Processes," *Detection of Abrupt Changes in Signals and Dynamical Systems*, edited by M. Basseville and A. Benveniste, Berlin: Springer-Verlag, 1986.

[13] E. Page, "Continuous Inspection Schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.

[14] G. Bodenstein and H. Praetorious, "Feature Extraction from the Encephalogram by Adaptive Segmentation," *Proc. IEEE*, vol. 65, pp. 642–652, 1977.

[15] L. Ljung, *System Identification*, Englewood Cliffs, New Jersey: Prentice-Hall, 1987.

[16] S. Marple, *Digital Spectral Analysis With Applications*, Englewood Cliffs, New Jersey: Prentice-Hall, 1987.

[17] M. Basseville and A. Benveniste, "Design and Comparative Study of Some Sequential Jump Detection Algorithms for Digital Signals," *IEEE Trans. Acoust. Speech. Sig. Process.*, vol. ASSP-31, no. 3, pp. 521–535, June 1983.

Table 1. Detection parameters.

Conditions	Threshold h	Drift δ	Detection time D_h units of τ, $\tau = 20$ msec	AR parameters $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5$	Signal energy[a]
Nominal	—	—	—	−0.41, −0.21, −0.12, −0.11, −0.05	0.9897
Fault 1	500	−0.04	1093	−1.61, +0.97, −0.27, −0.13, −0.10	3.9677
Fault 2	500	−8.0	1011	−0.85, −0.37, −0.10, +0.07, +0.29	3.4507
Fault 3	50	−0.01	N/A	−0.40, −0.17, −0.13, −0.14, −0.05	1.1861
Fault 4	50	−0.01	N/A	−0.37, −0.18, −0.14, −0.14, −0.06	0.9333
Fault 5	35	−0.10	1009	−0.02, +0.03, −0.02, 0.0004, −0.06	0.9891

[a] Data units are proportional to one Joule; the value of the proportionality constant is unknown.

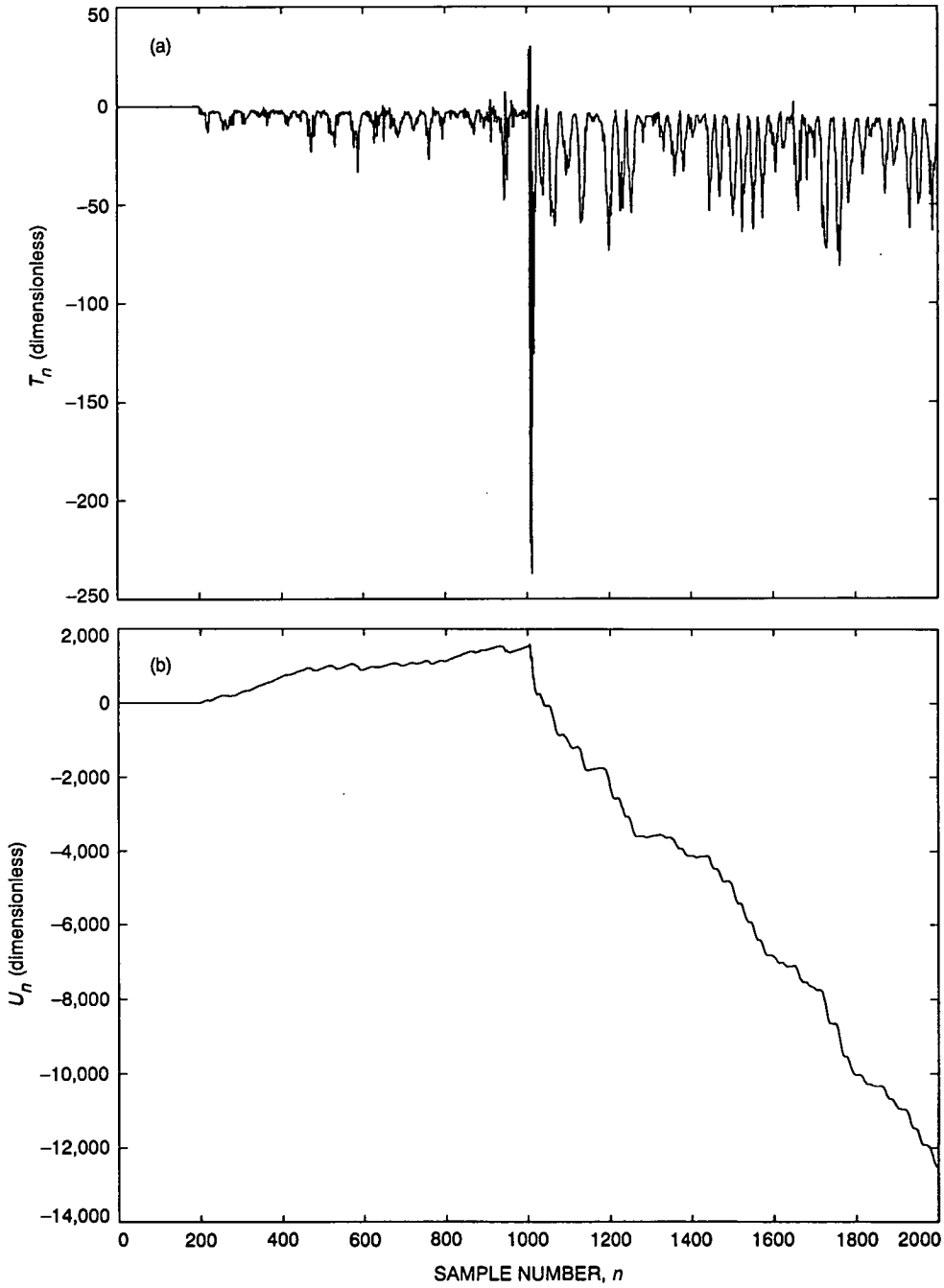


Fig. 1. Low-to-high signal energy transition: (a) typical *T* statistic and (b) cumulative sum *U*.

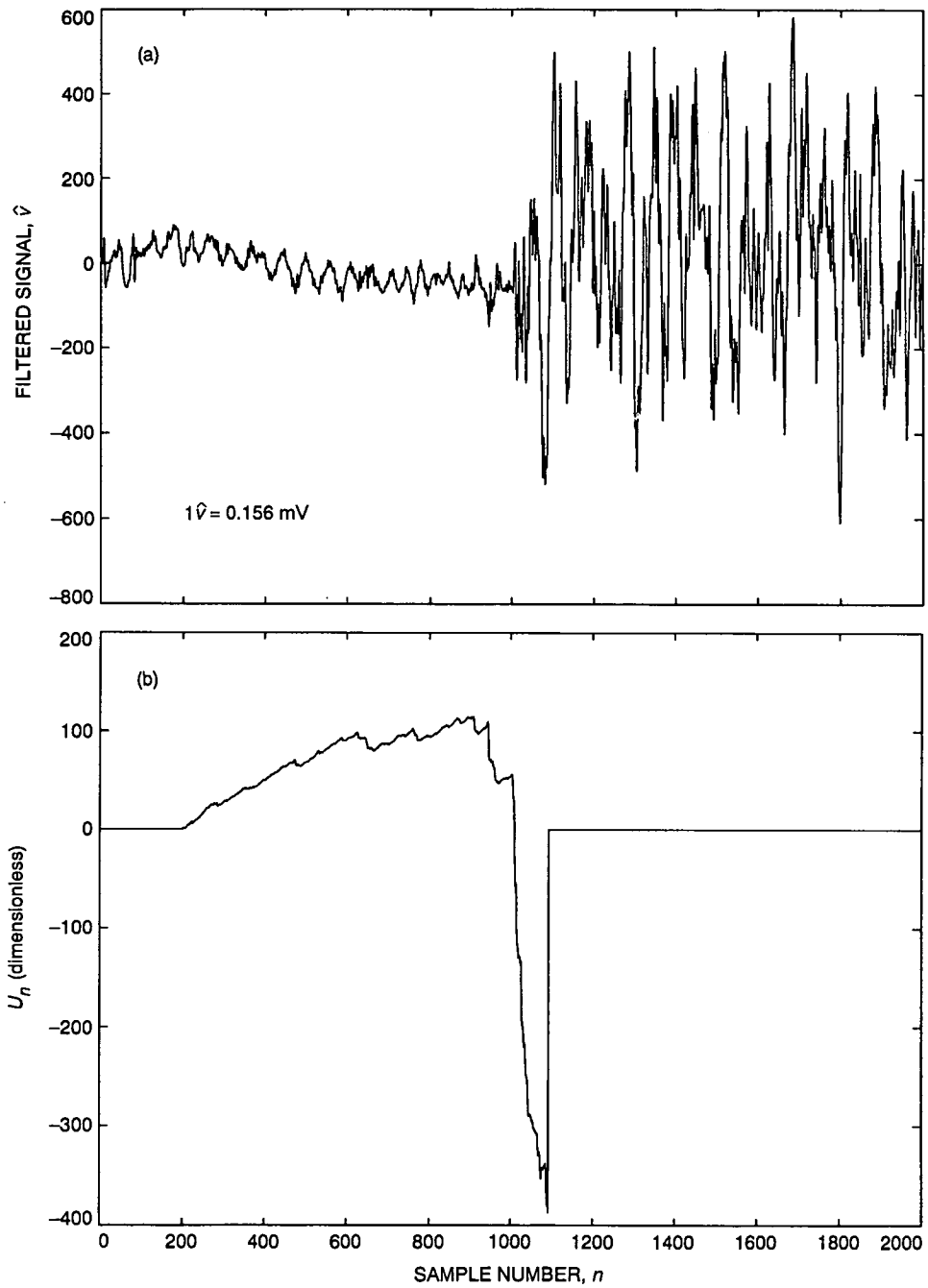


Fig. 2. Normal to fault 1 transition: (a) a filtered signal and (b) cumulative sum *U*.

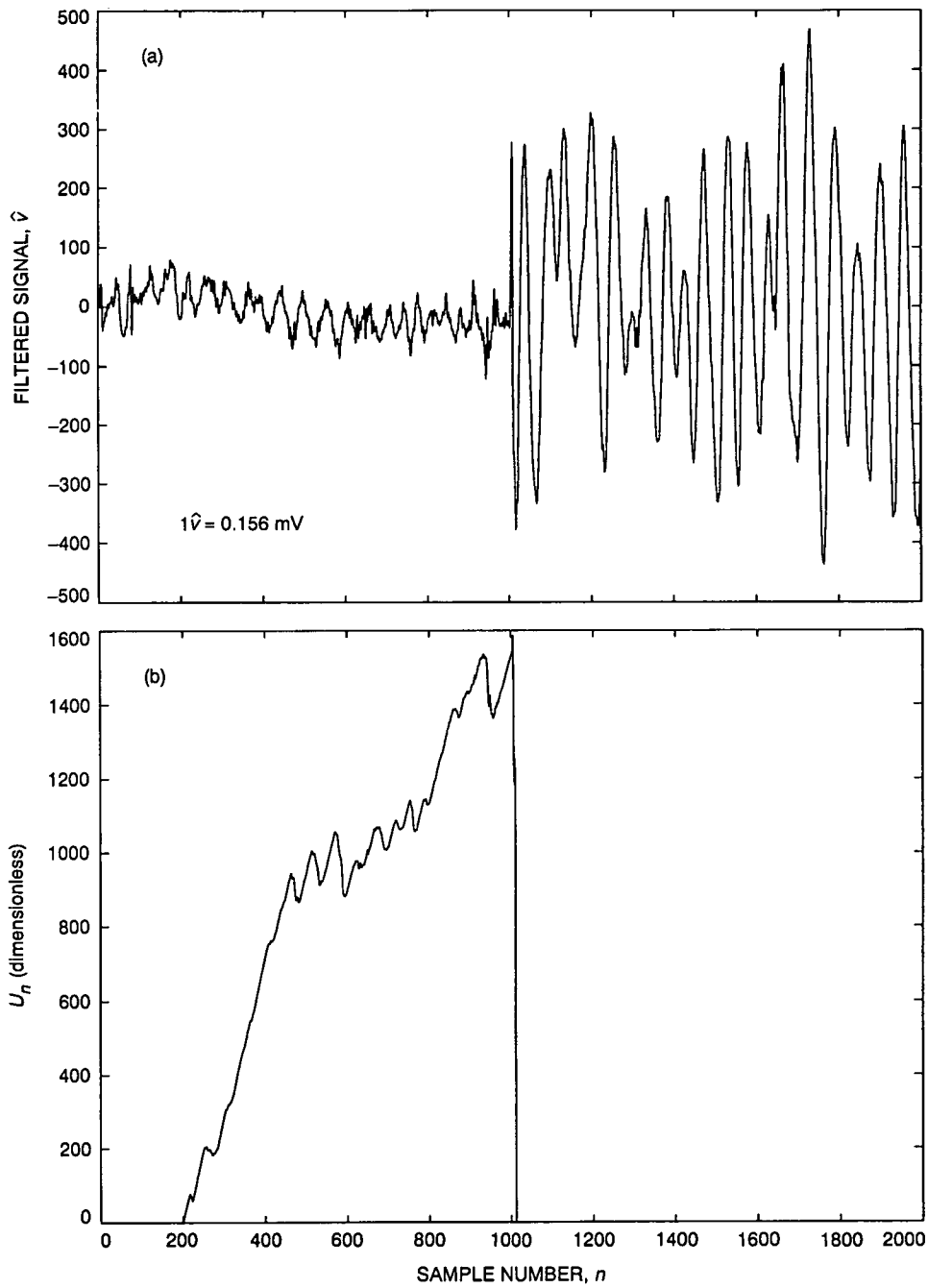


Fig. 3. Normal to fault 2 transition: (a) filtered signal and (b) cumulative sum *U*.

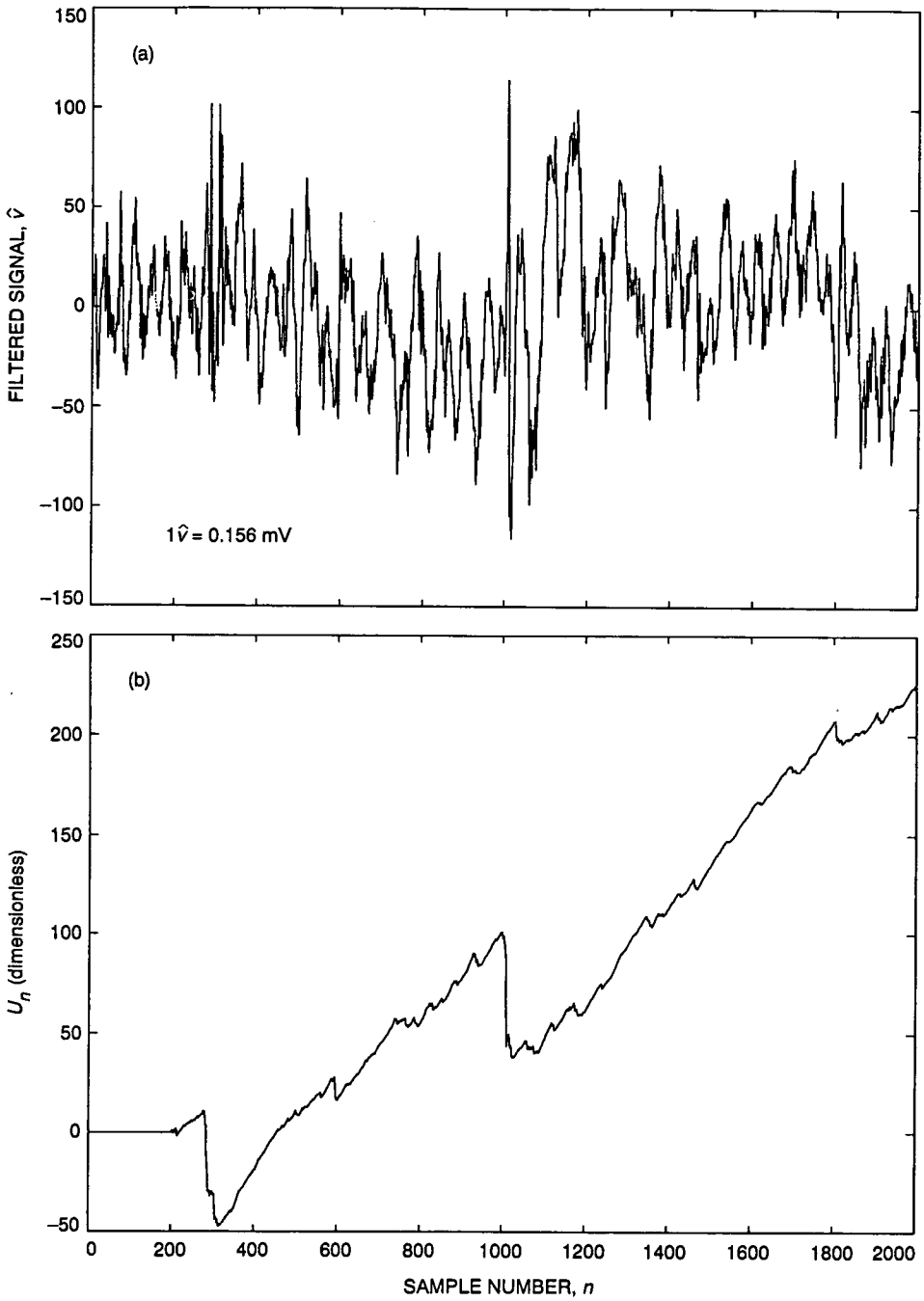


Fig. 4. Normal to fault 3 transition: (a) filtered signal and (b) cumulative sum *U*.

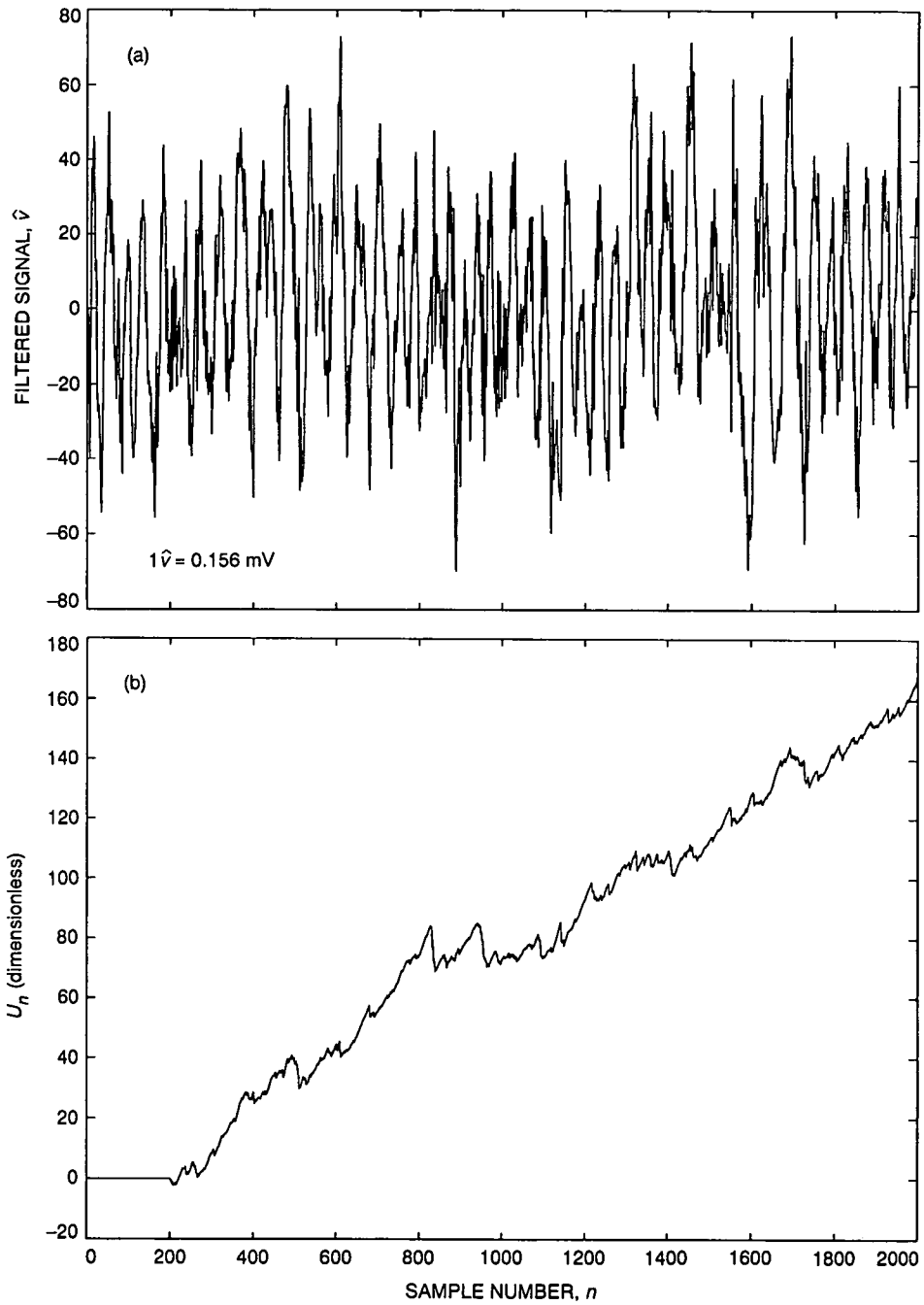


Fig. 5. Normal to fault 4 transition: (a) filtered signal and (b) cumulative sum *U*.

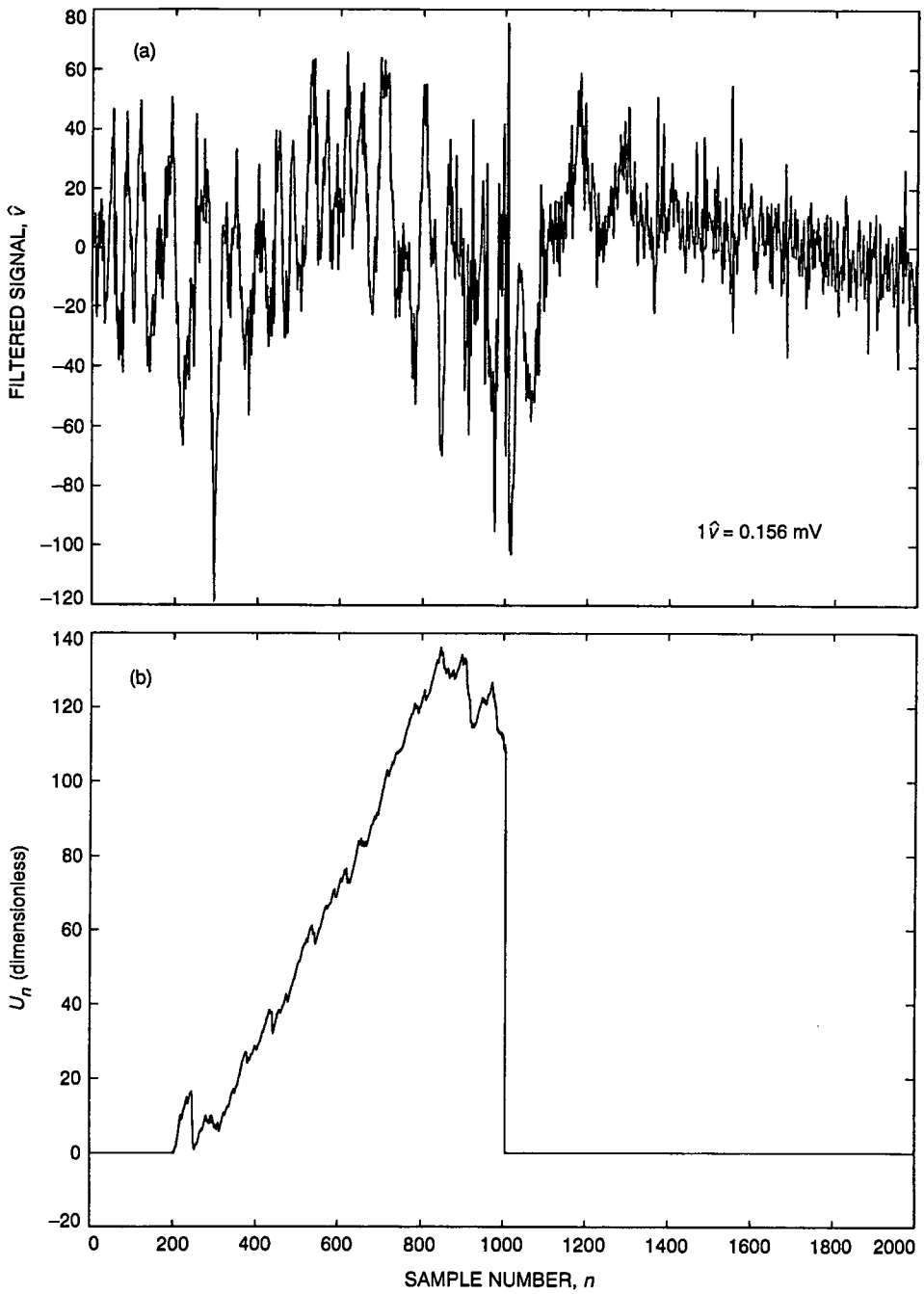


Fig. 6. Normal to fault 5 transition: (a) filtered signal and (b) cumulative sum *U*.

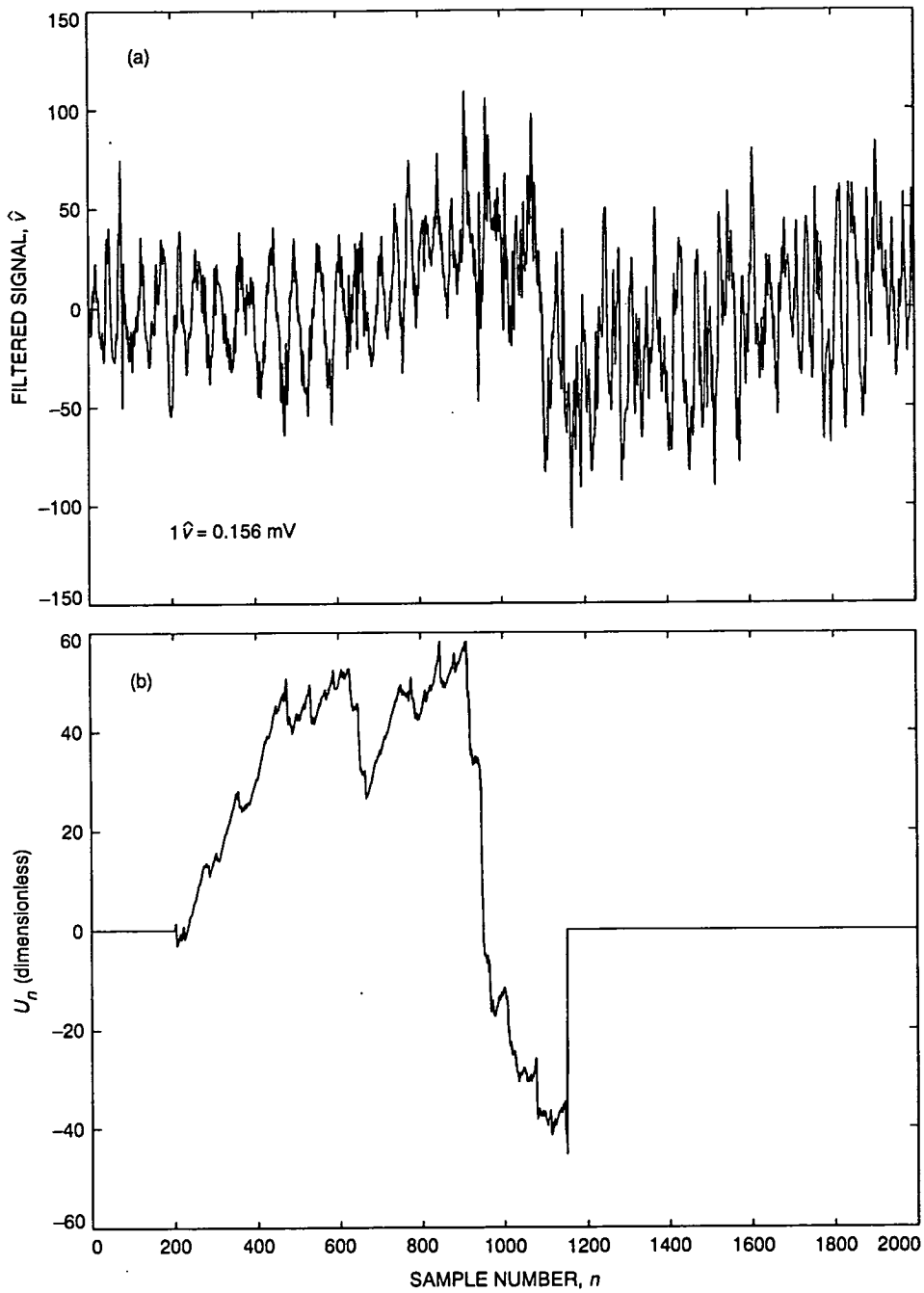


Fig. 7. Normal to scaled fault 1 transition: (a) filtered signal and (b) cumulative sum U.

Appendix

Derivation of Detection Statistic *T*

The derivation of the two-model distance metric T, as presented by Basseville [12], is replicated for the reader's convenience here.

Recall Eq. (5) which gives the distance metric T_k at time k, based on the Kullback conditional information-theoretic distance between probability laws $p^0\left(Y_k|\overline{Y}^{k-1}\right)$ and $p^1\left(Y_k|\overline{Y}^{k-1}\right)$

$$T_k = \int p^0\left(y|\overline{Y}^{k-1}\right)\log\left(\frac{p^1\left(y|\overline{Y}^{k-1}\right)}{p^0\left(y|\overline{Y}^{k-1}\right)}\right)dy$$

$$-\log\left(\frac{p^1\left(Y_k|\overline{Y}^{k-1}\right)}{p^0\left(Y_k|\overline{Y}^{k-1}\right)}\right) \tag{A-1}$$

When the observations Y_k are independent identically distributed Gaussian random variables, the conditional probabilities are given by Eq. (7):

$$p^0\left(Y_k|\overline{Y}^{k-1}\right) = \frac{1}{\sqrt{2\pi}\sigma_{\epsilon^0}}e^{-\left[\left(Y_k-\theta^0\overline{Y}^{k-1}\right)^2/2\sigma^2_{\epsilon^0}\right]}$$

$$p^1\left(Y_k|\overline{Y}^{k-1}\right) = \frac{1}{\sqrt{2\pi}\sigma_{\epsilon^1}}e^{-\left[\left(Y_k-\theta^1\overline{Y}^{k-1}\right)^2/2\sigma^2_{\epsilon^1}\right]} \tag{A-2}$$

With direct substitution, T_k becomes

$$T_k = -\frac{1}{2}\log\frac{\sigma^2_{\epsilon^0}}{\sigma^2\epsilon^1} + \frac{\left(y-\theta^0\overline{Y}^{k-1}\right)^2}{2\sigma^2_{\epsilon^0}}$$

$$-\frac{\left(y-\theta^1\overline{Y}^{k-1}\right)^2}{2\sigma^2_{\epsilon^1}} + I_{k-1} \tag{A-3}$$

where I_{k-1} is given by

$$I_{k-1} = \int\frac{1}{\sigma_{\epsilon^0}\sqrt{2\pi}}\exp\left(\frac{-\left(y-\theta^0\overline{Y}^{k-1}\right)^2}{2\sigma^2_{\epsilon^0}}\right)$$

$$\times\left[\frac{1}{2}\log\frac{\sigma^2_{\epsilon^0}}{\sigma^2_{\epsilon^1}} + \frac{\left(y-\theta^0\overline{Y}^{k-1}\right)^2}{2\sigma^2_{\epsilon^0}}\right.$$

$$\left.-\frac{\left(y-\theta^1\overline{Y}^{k-1}\right)^2}{2\sigma^2_{\epsilon^1}}\right]dy \tag{A-4}$$

Integrating,

$$I_{k-1} = \frac{1}{2}\log\frac{\sigma^2_{\epsilon^0}}{\sigma^2_{\epsilon^1}} + \frac{1}{2} - \frac{1}{2\sigma^2_{\epsilon^1}}I\left(\theta^0\overline{Y}^{k-1}, \theta^1\overline{Y}^{k-1}\right) \tag{A-5}$$

where

$$I(A,B) = \int\frac{1}{\sigma_{\epsilon^0}\sqrt{2\pi}}e^{\left[-(y-A)^2\right]/2\sigma^2_{\epsilon^0}}$$

$$\times(y-B)^2\,dy \tag{A-6}$$

$$= \sigma^2_{\epsilon^0} + (B-A)^2 \tag{A-7}$$

Moreover, by observation,

$$\theta^1\overline{Y}^{k-1} - \theta^0\overline{Y}^{k-1} = \epsilon^0 - \epsilon^1 \tag{A-8}$$

Equations (A-5) through (A-8) show that Eq. (9) holds

$$T_k = \frac{1}{2}\left[1 - \frac{\sigma^2_{\epsilon^0}}{\sigma^2_{\epsilon^1}} + \frac{\left(\epsilon^1\right)^2}{\sigma^2_{\epsilon^1}} - \frac{\left(\epsilon^0\right)^2}{\sigma^2_{\epsilon^0}} - \frac{\left(\epsilon^1-\epsilon^0\right)^2}{\sigma^2_{\epsilon^1}}\right] \tag{A-9}$$

Application of Multirate Digital Filter Banks to Wideband All-Digital Phase-Locked Loops Design

R. Sadr, B. Shah, and S. Hinedi
Communications Systems Research Section

A new class of architecture for all-digital phase-locked loops (DPLL's) is presented in this article. These architectures, referred to as parallel DPLL (PDPLL), employ multirate digital filter banks (DFB's) to track signals with a lower processing rate than the Nyquist rate, without reducing the input (Nyquist) bandwidth. The PDPLL basically trades complexity for hardware-processing speed by introducing parallel processing in the receiver. It is demonstrated here that the DPLL performance is identical to that of a PDPLL for both steady-state and transient behavior. A test signal with a time-varying Doppler characteristic is used to compare the performance of both the DPLL and the PDPLL.

I. Introduction and Background

Implementation of wideband phase-locked loops (PLL's) has various applications in the areas of ranging, navigation, communications, and many other fields, where it is desirable to coherently track a continuous waveform (CW) signal with a particular Doppler profile. Such scenarios arise in Earth-orbiting satellites or in deep space links where a satellite or a probe is capable of transmitting signals in various channels (all within the same band) spanning a few hundred megahertz. Currently, superwide PLL's with a front-end bandwidth in the neighborhood of a gigahertz are implemented using analog devices as the digital technology is not mature enough to operate at these high clock rates. Digital PLL's, implemented with complementary metal-oxide semiconductor (CMOS) or gallium arsenide (GaAs) technologies, can operate at 100 MHz, hence covering a 50-MHz bandwidth at best.

For example, consider the Block V receiver[1] which samples the signal at 160 MHz and then processes these samples at 80 MHz. In order to cover the 100-MHz allocated bandwidth at X-band, preselect filters (see Fig. 1) are used to downconvert a portion of the spectrum to the appropriate intermediate frequency (IF) for digitization. If the Doppler rate were to span a bandwidth larger than the

[1] *Block V Receiver, Subsystem Design Review*, JPL D-7420 (internal document), Jet Propulsion Laboratory, Pasadena, California, June 1992.

preselect filter, the receiver would have to be reset in the middle of a track and then restarted with the output of a different preselect filter. In the future when the DSN operates at Ka-band (33 GHz), the allocated bandwidth will be about 500 MHz and more preselect filters will be required. The Doppler rates at Ka-band can easily reach 1 kHz/sec (the Mars Observer spacecraft experiences a two-way Doppler of 800 Hz/sec in the Ka-band Link Experiment) requiring Doppler tuning using predicts to center the signal in the passband of the appropriate preselect filter.

Advancements in high-speed digital integrated circuit technology enable high-end data acquisition systems to sample signals in excess of hundreds of megahertz. However, digital signal-processing operations (such as filtering, mixing, etc.) of these samples are still not cost-effective and sometimes not even feasible at these rates. A survey of technology (as of December 1991) was performed for high-speed state-of-the-art signal-processing hardware, as shown in Table 1. The total power consumed by these high-speed circuits is excessive, thereby requiring large heat sinks and specially designed boards for proper heat conduction and dissipation. Furthermore, to develop digital hardware at speeds around 100 MHz, it becomes necessary to use extremely accurate scopes and test equipment that could be another major cost-bearing investment on the part of the developer. Naturally one wonders whether it is possible to devise architectures that can employ a low-power CMOS that, at its best, can be clocked at a speed of 75 MHz or lower. This would significantly reduce the cost of development, the cost of the components, and that of the end product as a whole.

This article focuses on a new class of wideband all-digital PLL (DPLL) architectures that employ parallel signal-processing techniques to reduce the processing rate to below the Nyquist rate, while still maintaining a super-wide front-end bandwidth. As an example, a CW signal in principle can be sampled at 1 GHz (with a corresponding front-end bandwidth of 500 MHz) and processed in parallel using 20 channels, each with a 50-MHz clock, to provide a continuous coherent reference, even if the signal's Doppler profile spans the full 500-MHz bandwidth. The new architectures presented in this article, referred to as parallel DPLL (PDPLL), employ multirate digital filter banks (DFB's) to process the signal. There are many possible approaches [1–3] for designing DPLL's for CW tones. A discussion of the approaches and the merits of each approach is beyond the scope of this article. Here, the application of DFB to the DPLL is studied. In Section II, the DPLL is briefly described, and in Section III, the PDPLL is introduced and discussed. The simulation results and the realization of a particular PDPLL are outlined in Section IV. The design methodology and the characterisitics of the DFB are outlined in the Appendix. The discussion begins with a brief background of DPLL theory.

II. DPLL Operation

A general DPPL block diagram is depicted in Fig. 2. The received signal $r(t)$ is assumed to be a single tone embedded in noise, i.e.,

$$r(t) = \sqrt{2P_c}\sin(\Theta_c(t)) + n(t) \tag{1}$$

where

P_c = the carrier power in watts (W)

$\Theta_c(t) = \omega_c t + \theta_c$ is the total carrier phase in radians

$n(t)$ = the additive white Gaussian noise (AWGN) process with two-sided power spectral density, $N_0/2$ W/Hz

The received signal is first band-limited by an anti-aliasing filter, then bandpass sampled to form the DPLL input. The sampled input, $r(nT_s)$ in Fig. 2, is given as (the time between samples T_s is omitted in the following equations to allow a simpler notation)

$$r(n) = \sqrt{2P_c}\sin(\Theta_c(n)) + n_{bp}(n) \tag{2}$$

where the discrete-time bandpass noise process is given by

$$n_{bp}(n) = \sqrt{2}\,n_c(n)\cos(\Theta_c(n)) - \sqrt{2}n_s(n)\sin(\Theta_c(n)) \tag{3}$$

and $n_c(n)$ and $n_s(n)$ are statistically independent, band-limited, with a two-sided bandwidth of $2Bn$ Hz and a two-sided power spectral density level of $N_0/2$ W/Hz.

The DPLL error signal is obtained by mixing $r(n)$ with $\cos(\hat{\Theta}_c(n))$ and then lowpass filtering the mixer output to simultaneously retain the resulting dc term and suppress the double frequency term. The term $\hat{\Theta}_c(n)$ is an estimate of the incoming carrier phase $\Theta_c(n)$. Assuming an ideal lowpass filter and $N = 1$, the input to the loop filter in Fig. 3 is given as

$$\hat{\phi}_c(n) = \sqrt{P_c}\sin(\phi_c(n)) + n_c(n)\cos(\phi_c(n)) - n_s(n)\sin(\phi_c(n)) \tag{4}$$

where $\phi_e(n) = \Theta_e(n) - \hat{\Theta}_e(n)$ is the actual total phase error to be estimated. The loop filter output, $\hat{f}(n)$, is used to update the incoming carrier phase estimate, as follows (assuming $N = 1$ in Fig. 3):

$$\hat{\Theta}_e(n+1) = \hat{\Theta}_e(n) + 2\pi\hat{f}(n)T_u \tag{5}$$

The loop filter transfer function for a third-order loop is given by

$$F(z) = G_1 + G_2\frac{z}{z-1} + G_3\left(\frac{z}{z-1}\right)^2 \tag{6}$$

where $G_1 = \dfrac{rd}{2\pi T_u}$, $G_2 = \dfrac{rd^2}{2\pi T_u}$, $G_3 = \dfrac{krd^3}{2\pi T_u}$, and

$$d = \frac{4B_L T_u(r-k)}{r(r-k+1)} \tag{7}$$

The parameter B_L in Eq. (7) denotes the design or equivalent one-sided analog loop bandwidth in hertz, $T_u = NT_s$ is the loop update interval, r is typically 2 or 4 and is equal to 4ξ where ξ is the analog damping ratio and k is a loop gain parameter for a third-order loop with typical values ranging from 1/4 to 1/2. The actual loop bandwidth, B_L^*, might be larger than B_L, depending on the product $B_L T_u$. Generally for $B_L T_u < 0.05$, the actual loop bandwidth is very close to the analog loop bandwidth B_L.

The tracking performance of the DPLL is well known [1] to be related to the loop bandwidth and carrier-to-noise density ratio as follows:

$$\sigma_{\phi_c}^2 = \frac{N_o B_L}{P_c} \tag{8}$$

where $\sigma_{\phi_c}^2$ is the variance of the phase error $\phi_e(n) = \Theta_e(n) - \hat{\Theta}_e(n)$.

The DPLL described in this section requires that the analog-to-digital (A/D) output $r(n)$ be downconverted and filtered at the sampling rate. Hence, with the exception of the loop filter, which can be implemented at the lower update rate, all the DPLL components must operate at the higher sampling rate. This is undesirable because the implementation cost of post-sampling operations, such as downconversion and filtering, limit the A/D conversion rate, and consequently, the Nyquist bandwidth. One way to circumvent this problem is by introducing a multirate DFB between the A/D and DPLL. As shown in the next section, a DPLL together with a DFB, or PDPLL, can track a signal over the Nyquist band but with a DPLL that operates at a much lower rate than the conventional DPLL described in this section.

III. Multirate DFB Implementation of DPLL

Multirate DFB's have been studied extensively in the past [4,5]. Let $H(z)$ denote the transfer function of an arbitrary digital filter, i.e.,

$$H(z) = \sum_{k=-\infty}^{\infty} z^{-k}h_k \tag{9}$$

where $\{\ldots h_2, h_{-1}, h_0, h_1, h_2 \ldots\}$ is the impulse response of the filter. It is possible to represent $H(z)$ in terms of its M-component polyphase form

$$H(z) = \sum_{k=0}^{M-1} z^{-k}E_k(z^M) \tag{10}$$

where the coefficients

$$e_k(n) = h(nM+k),\ \forall\ 0 \le k \le M-1 \tag{11}$$

and $E_k(z)$ is the z-transform of $e_k(n)$ and is called the kth polyphase component of $H(z)$. The expansion in Eq. (10) is simply the decomposition of $\{h(n)\}$ into M-subsequence $e_k(n)$. For example, by grouping the impulse response coefficients $h(n)$ into even- and odd-numbered samples, i.e., $e_0(n) = h(2n)$ and $e_1(n) = h(2n+1)$, the transfer function $H(z)$ may be represented as

$$H(z) = E_0(z^2) + z^{-1}E_1(z^2) \tag{12}$$

where

$$E_0(z) = \sum_{n=-\infty}^{\infty} h(2n)z^{-n} \tag{13a}$$

and

$$E_1(z) = \sum_{n=-\infty}^{\infty} h(2n+1)z^{-n} \tag{13b}$$

An important property of this representation is that if the filter is followed by a decimation operation, then the filtering operation and the decimation can be commuted.

This property, known as the Noble identity, is depicted in Fig. 4. With the application of the Noble identity to the polyphase representation of Eq. (10), the filter $H(z)$ can be redrawn, as shown in Fig. 5. This representation is also referred to as the blocked version of the prototype filter $H(z)$. The key advantage of using the model shown in Fig. 5(b) is that the processing rate in each filter bank is a factor of M slower than the sampling clock. The polyphase representation results in an efficient rearrangement of the computations of the filtering operation. This effectively distributes the computations into a set of parallel filter banks operating at a lower speed. This, in turn, reduces the speed constraints on the digital signal processing hardware, thereby enabling it to process samples at a rate much lower than the sampling rate.

In application to a DPLL, as shown in Fig. 2, a full band filter $G(z)$ is inserted at the output of the A/D before the input to the mixer, as depicted in Fig. 6(a). For all practical purposes, the insertion of this digital filter does not alter the performance of the loop as it is a bandpass filter, with a bandwidth identical to that of the bandpass filter prior to the A/D conversion in Fig. 2. Consider decomposing the filter $G(z)$ into M sub-band parallel filters, $G_0(z)$, $G_1(z), \ldots, G_{M-1}(z)$, where $G_i(z)$ is a bandpass filter that passes a portion of the spectrum, as shown in Fig. 6(b). Since the input to the DPLL is a CW tone, it occupies a single filter at any given time and, therefore, only the output of that specific filter needs to be processed by the DPLL. In this case, the adder can be replaced by a multiplexer, which only passes the output of the appropriate filter to the DPLL. These filters are implemented at the sampling rate and the DPLL is operating (mixing and phase estimating) also at the sampling rate. Note that since each sub-band filter $G_i(z)$ has a bandwidth equal to $1/M$ of the bandwidth of $G(z)$, the output rate of the sub-band filters can therefore be decimated by a factor M while still satisfying the bandpass sampling theorem [6]. In this case, the DPLL can thus operate at $1/M$ of the sampling rate without any loss of information, as depicted in Fig. 6(c). Each branch now consists of a bandpass filter followed by a decimator. Each filter can be decomposed in terms of its polyphase components, as given by Eq. (10), i.e.,

$$G_i(z) = \sum_{k=0}^{M-1} z^{-k} E_{k,i}(z^M) \tag{14}$$

and since the filter is followed by a decimation, the Noble identity can be invoked to commute the filtering and the decimation, resulting in the structure shown in Fig. 6(d). Only one sub-band filter has been decomposed in the figure, even though all filters should be decomposed in any practical implementation to allow for a lower processing rate. The combination of the filter banks, their respective polyphase decomposition, the multiplexer, and the lower rate DPLL is referred to as the PDPLL. As can be seen from Fig. 6, the PDPLL processes the samples at a significantly reduced rate (depending on M) while still being able to track a signal spanning the full input bandwidth (which is at most half the sampling rate).

When the signal frequency at the DFB input changes due to spacecraft acceleration or jerk, the signal could pass from one filter to the next. In this case, the output of the filter with a signal present in its passband is multiplexed and used to drive the DPLL. Formally, denote the total DFB bandwidth as Ω. Let $\Omega = [F_1, F_2]$ and let $I_i = [F_i^l, F_i^u]$ such that $\Omega = \bigcup_{i=0}^{M-1} I_i$ where the passband of the ith filter corresponds to the interval I_i. Furthermore, assume that a uniform filter bank is employed, i.e., $F_i^u - F_i^l = F_2 - F_1/M = \Delta F, \forall i$. Note that $I_0 = [F_1, F_1^u]$ and $I_{M-1} = [F_{M-1}^1, F_2]$. In order to properly select the output of the filter bank, let m denote the mth filter whose output is tuned to the received signal. By monitoring the estimated instantaneous frequency of the numerically controlled oscillator (NCO), the multiplexing control operation becomes simply that of selecting the appropriate filter within the bank where the signal lies. Denoting the instantaneous frequency $\hat{f}_{IF}$ as the output from the NCO, the mth filter is found to be

$$m : \hat{f}_{IF} \in I_m \quad \text{and} \quad m \in [0,\ M-1] \tag{15}$$

The algorithm for selecting the proper filter is also shown in the flowchart of Fig. 7. The IF estimate, $\hat{f}_{IF}$, of the input signal is found by adding the output of the loop filter to the nominal frequency of the NCO, i.e, $\hat{f}_{IF} = f_{NCO} + \hat{f}$. The multiplexer position is then set to the mth filter in the DFB according to Eq. (15). The control logic can be added as shown in Fig. 6(d) to provide the control signal for the multiplexer. A more elaborate algorithm can be used to control the multiplexer by taking into account the states of the loop filter, which are related to the Doppler rate and its derivatives.

For the PDPLL application, the DFB's in Fig. 6(d) are required to have a constant group delay and continuous phase as a function of frequency. These requirements become critical when the incoming tone is likely to span more than one filter due to a time-varying Doppler characteristic. In this case, DFB's with a constant group delay and continuous phase would not introduce a phase jump to the DPLL input when the multiplexer selects a different filter

output. DFB's without these properties would introduce a phase jump that would cause the DPLL to temporarily loose lock.

In the following section, an example of five filter banks is considered, and simulations are performed to verify both the steady-state and the transient behavior of the PDPLL.

IV. Example of PDPLL and Simulation Results

The tracking and acquisition performance of the DPLL and the PDPLL is characterized by simulation. The DPLL simulation is based on the block diagram shown in Fig. 6(a), and PDPLL simulation is based on Fig. 6(d) with $M = 5$. The DFB's used in the simulation are described in the Appendix. The passband and center frequencies, normalized by the sampling rate, are summarized in Table A-1. The "cross-over frequency," the frequency at which the DFB output is selected to be from a different filter, is also shown in Table A-1. The multiplexing algorithm in the simulation is the same as that described in Section III.

The theoretical tracking variance of the DPLL [Eq. (8) in Section II] versus the tracking variance of the PDPLL obtained via simulation is shown in Fig. 8. The input to the DFB is a 10-KHz sine wave sampled at 40 kHz, the DFB output is decimated by 5 so that the input to the loop is a 2-kHz sine wave. Consequently, the loop was simulated to be in lock by operating the loop NCO at a frequency of 2 kHz. The loop bandwidth was kept constant at 100 Hz and the simulated loop SNR's of 10, 20, and 30 dB were obtained by generating noisy sine waves with P_c/N_0 values of 30, 40, and 50 dB, respectively. The simulations were performed for $100/B_L$ or 1 sec and are seen to agree very well with theory.

The remaining figures (Figs. 9, 10, and 11), which show the acquisition performance of the PDPLL and DPLL in the absence of noise, indicate that both loops have identical transient responses. In the results that follow, the curves for the PDPLL case are deliberately offset to differentiate them from the DPLL curves. Figure 9(a) depicts the transient phase error response to a 0.1-rad phase step when the input to the DFB is a 10-kHz sine wave. Clearly, both cases have the same transient response to a phase step. The next figure, Fig. 9(b), depicts the transient behavior when the frequency error is 10 Hz and the phase error is 0.1 rad. Here, the A/D output is a 10.010-kHz sine wave. Once again, the two cases are seen to be identical. The acquisition performance when the input signal frequency is linearly changing at a rate of 200 Hz/sec is shown in Fig. 10(a). In this case, the initial signal frequency was set to 10.5 kHz and the simulation was run for 1.25 sec at a sampling rate of 40 kHz. From Table A-1, with the sampling rate of 40 KHz, it is seen that the signal passes from $G_3(z)$ to $G_4(z)$ during the simulation. Figure 10(b) shows the filter selected by the multiplexer versus the frequency of the incoming signal. As expected, the multiplexer selects $G_3(z)$ when the input signal frequency is lower than 10.625 kHz (the cross-over frequency), and it selects $G_4(z)$ when the signal frequency is higher than 10.625 kHz. Note from Fig. 10 that the loop does not lose lock when the multiplexer changes its output. That is, there are no transients in the phase error response at $t = 0.625$ sec, the time when the multiplexer changes its position.

The frequency and phase error responses when the input signal frequency is changing at a rate of 5145 Hz/sec^2 [as depicted in Fig. 11(a)] are shown in Figs. 11(b) and 11(c). In this case, the initial signal frequency was set to 10.425 kHz and the simulation was run for 1 sec. The frequency was kept constant for the first 0.5 sec of the simulation, after which it changed according to (5145 Hz/sec^2)$(t - 0.5)^2/2$ [see Fig. 11(a)], where t is the actual simulation time. Hence, the signal frequency at the end of the simulation is the initial frequency 10.425 kHz plus 5145 $(0.5^2/2)$ or 11.068 kHz. Once again the signal crosses over from $G_3(z)$ to $G_4(z)$, and the multiplexer, as shown in Fig. 11(d), selects the proper filter output. The steady-state phase error in Fig. 11(c) for these dynamics and loop bandwidth matches the theoretical steady-state error of 0.235 rad [2].

V. Conclusion

It is concluded that the PDPLL provides a viable solution to fabricating low-cost wideband DPLL's. By employing multirate digital filters, it is possible to use a much lower processing rate than the sampling rate. Various input signals with different Doppler profiles were used to demonstrate the utility of the PDPLL. The simulation results indicate that the tracking and acquisition performance of the PDPLL is essentially equivalent to that of the conventional PLL. In order to establish the dynamic range of the tracking loop with the multirate DFB, the simulation was performed with a frequency ramp of 200 Hz/sec, and in the presence of very high dynamics, namely a jerk of 5145 Hz/sec^2. The steady-state phase error in each case agreed with the theoretically predicted steady-state phase error.

References

[1] W. C. Lindsey and C. M. Chie, "A Survey of Digital Phase Locked Loops," *IEEE Proceeding*, vol. 69, no. 4, pp. 410–431, April 1981.

[2] S. Aguire and W. Hurd, "Design and Performance of Sampled Data Loops for Carrier and Subcarrier Tracking," *TDA Progress Report 42-79*, vol. July–September 1984, pp. 81–85, Jet Propulsion Laboratory, Pasadena, California, November 15, 1984.

[3] S. Aguire, "A Comparison of Methods for DPLL Loop Filter Design," *TDA Progress Report 42-87*, vol. July–September, 1986, pp. 114–124, Jet Propulsion Laboratory, Pasadena, California, August 15, 1986.

[4] P. P. Vaidyanathan,"Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications," *IEEE Proceedings*, vol. 78, no. 1, pp. 56–93, January 1990.

[5] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*, Englewood Clifs, New Jersey: Prentice Hall, 1986.

[6] R. Sadr and M. Shahshahani, "On Sampling Band-Pass Signals," *TDA Progress Report 42-96*, vol. October–December 1988, pp. 14–20, Jet Propulsion Laboratory, Pasadena, California, February 15, 1985.

Table 1. Survey of high-speed signal processing hardware (December 1991).

Technology	Device	Speed	Company	Power, W
Acoustic charged transport	FIR	360 MHz	Electronic Design, Inc.	6
Bipolar GaAs	Arithmetic logic unit	250 MHz	Texas Instruments	1.5
Emitter-coupled logic (ECL)	316-NCO	1.4 GHz	Plessy	5
GaAs	10-bit digital-to-analog converter	1 GHz	ITT Avionics	5
Enhanced depletion mode GaAs	Logic gates	1 GHz	Harris	–[a]
ECL	8-bit analog-to-digital converter	500 MSPS[b]	Tektronics	5
Surface acoustic wave	Voltage-controlled oscillator (VCO)	100 MHz	Ericsson Fatme	1
High-electron mobility transistor	Logic gates at 77 K	40 GHz	(See footnote c)	–
GaAs	4 K x 4 RAM	200 MHz	Mitsubishi Electric	2
GaAs	ROM	650 MHz	Gigabit Logic	4
Heterojunction bipolar transistors	VCO-divider	12 GHz	TRW	–

[a] 50 times less than ECL.

[b] Millions of samples per second.

[c] Six Japanese manufacturing companies, including Fujitsu and Mitsubishi, among others.

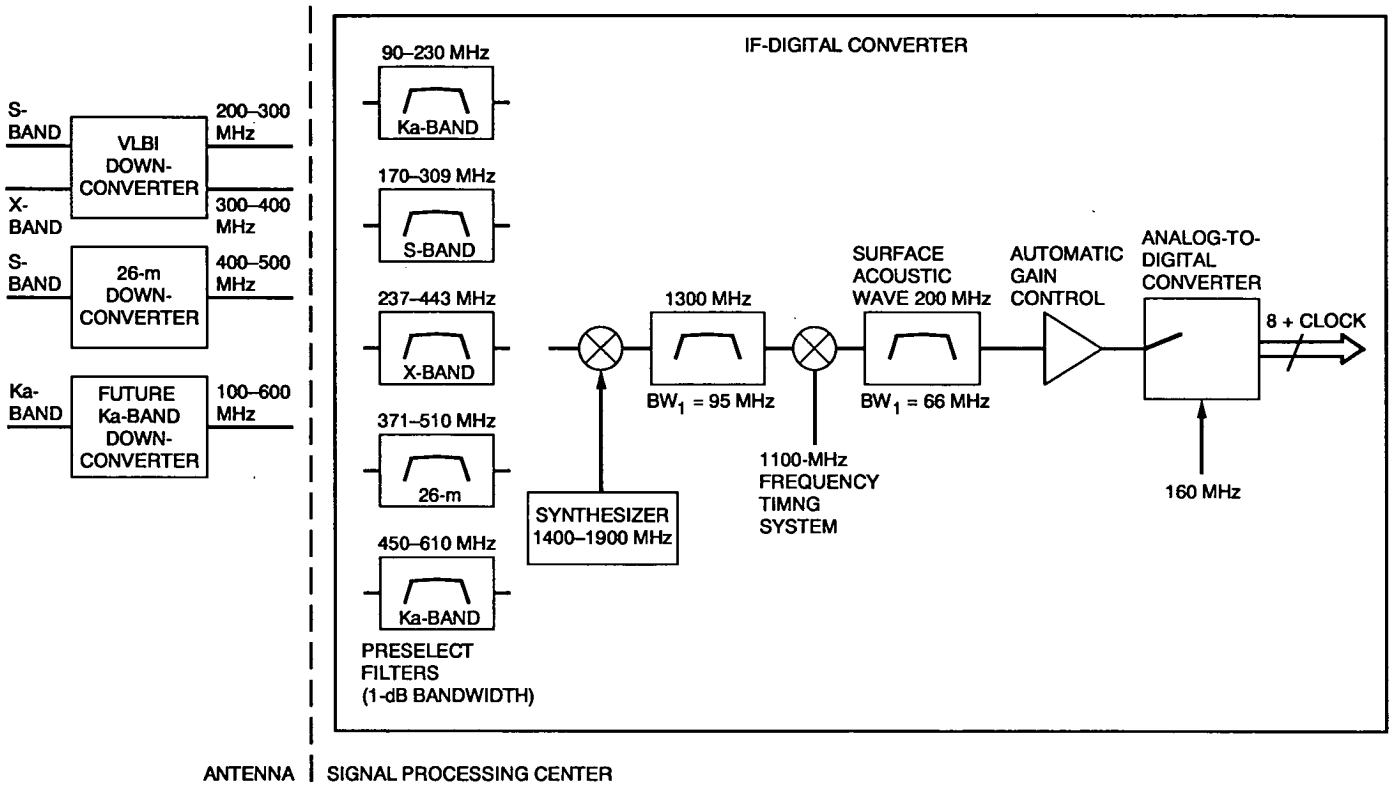


Fig 1. Block V receiver architecture.

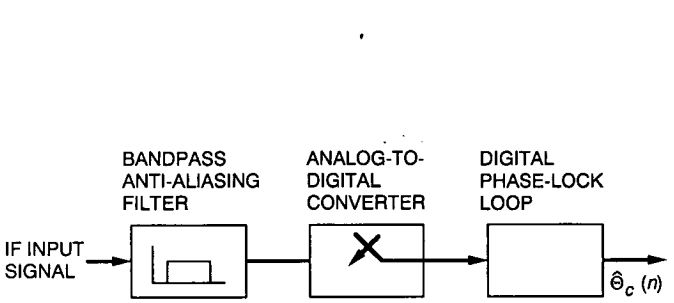


Fig. 2. Digital phase-lock loop preceded by an A/D converter.

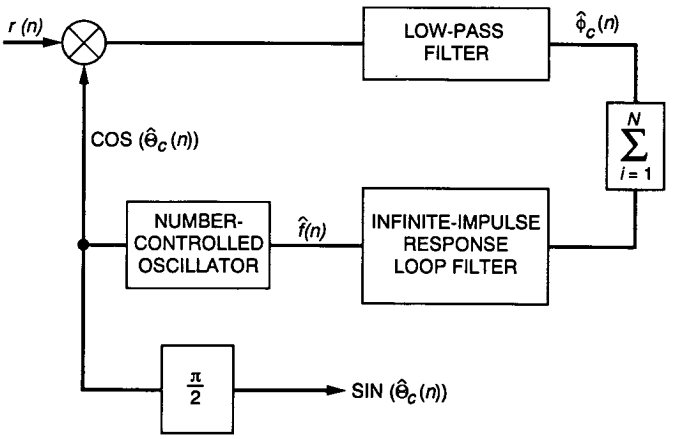


Fig. 3. DPLL block diagram.

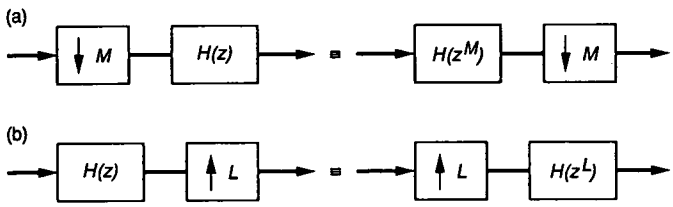


Fig. 4. Noble identity for multirate systems: (a) decimation and (b) interpolation.

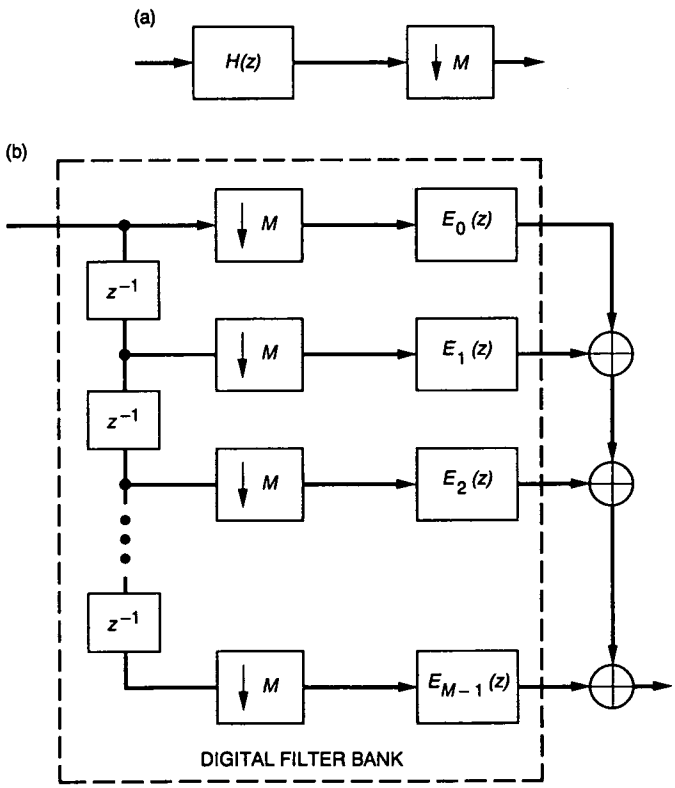


Fig. 5. Polyphase representation: (a) prototype filter and (b) blocked version.

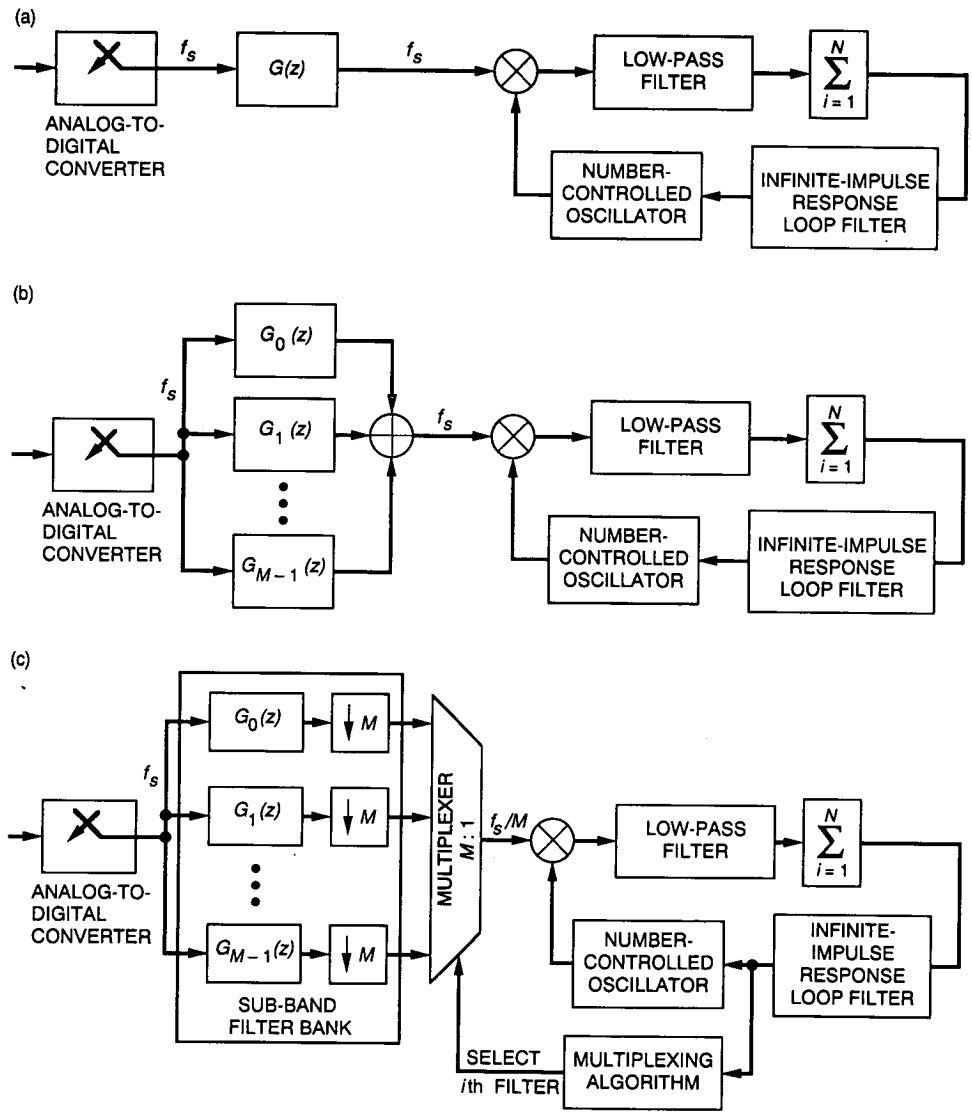


Fig. 6. Development of PDPLL: (a) bandpass filter insertion; (b) sub-band filter decompositions; (c) decimation; and (d) polyphase decomposition.

(d)

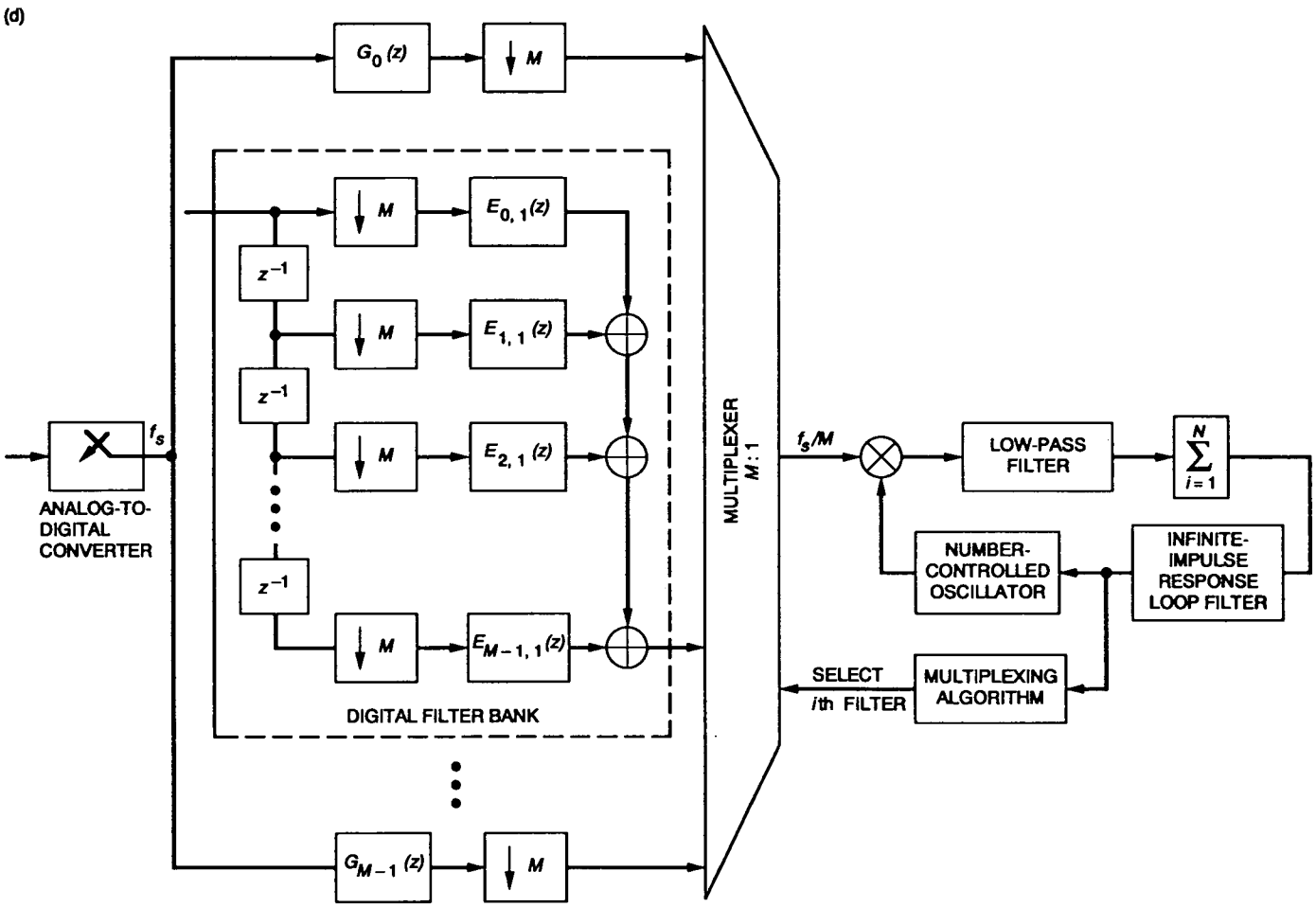
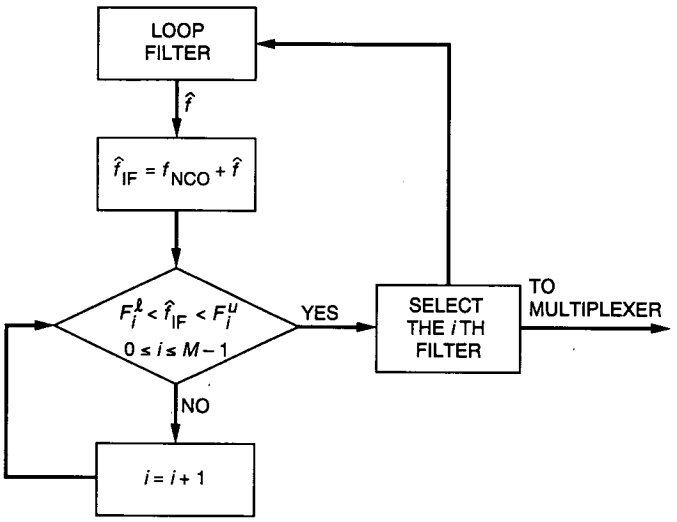

$G_0(z)$
M
$E_{0,1}(z)$
z^{-1}
$E_{1,1}(z)$
$E_{2,1}(z)$
$E_{M-1,1}(z)$
DIGITAL FILTER BANK
$G_{M-1}(z)$
f_s
ANALOG-TO-DIGITAL CONVERTER
MULTIPLEXER M:1
f_s/M
LOW-PASS FILTER
$\sum_{i=1}^{N}$
NUMBER-CONTROLLED OSCILLATOR
INFINITE-IMPULSE RESPONSE LOOP FILTER
SELECT ith FILTER
MULTIPLEXING ALGORITHM

Fig. 6 (contd).



$\hat{f}$ DOPPLER FREQUENCY ESTIMATE

f_{NCO} NUMBER-CONTROLLED OSCILLATOR FREE-RUNNING FREQUENCY

$\hat{f}_{IF}$ DIGITAL FILTER BANK INPUT FREQUENCY ESTIMATE

F_i^{ℓ} LOWER CUTOFF FREQUENCY, iTH FILTER

F_i^{u} UPPER CUTOFF FREQUENCY, iTH FILTER

Fig. 7. Flowchart for the filter selection (multiplexing) algorithm.

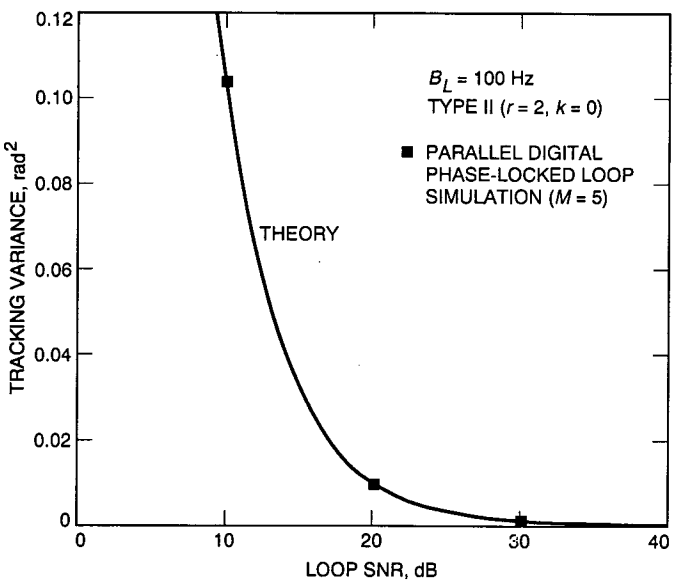


Fig. 8. The simulated tracking variance of the PDPLL (M = 5) and the theoretical tracking variance as a function of loop SNR.

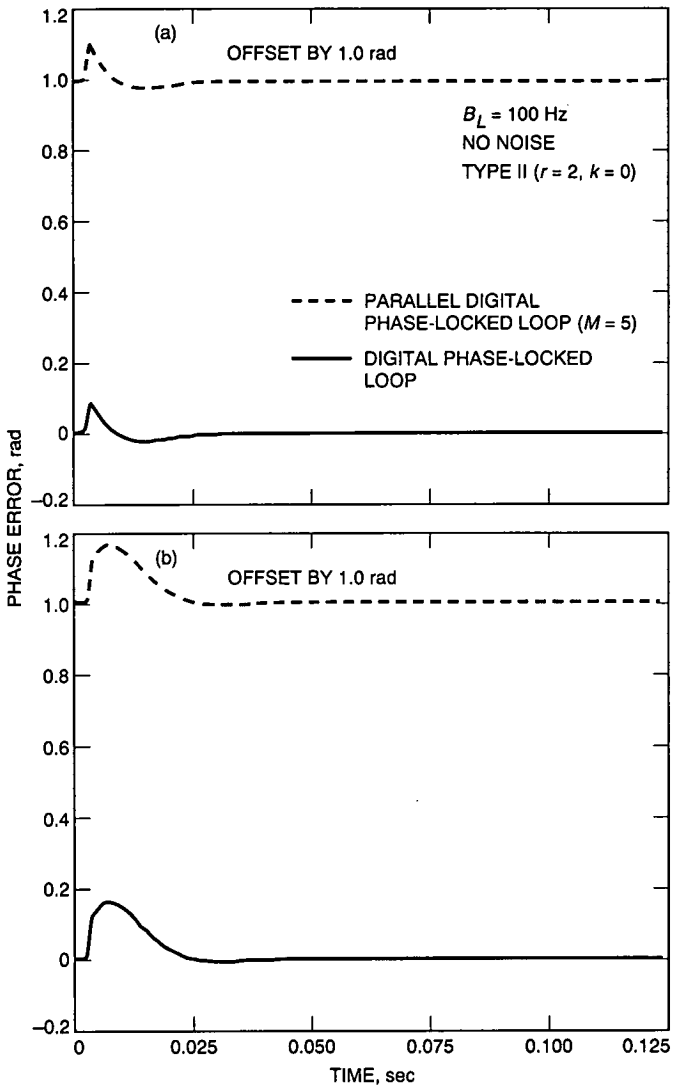


Fig. 9. The DPLL and PDPLL transient response to: (a) 0.1-rad phase step and (b) 10-Hz frequency offset plus a 0.1-rad phase step.

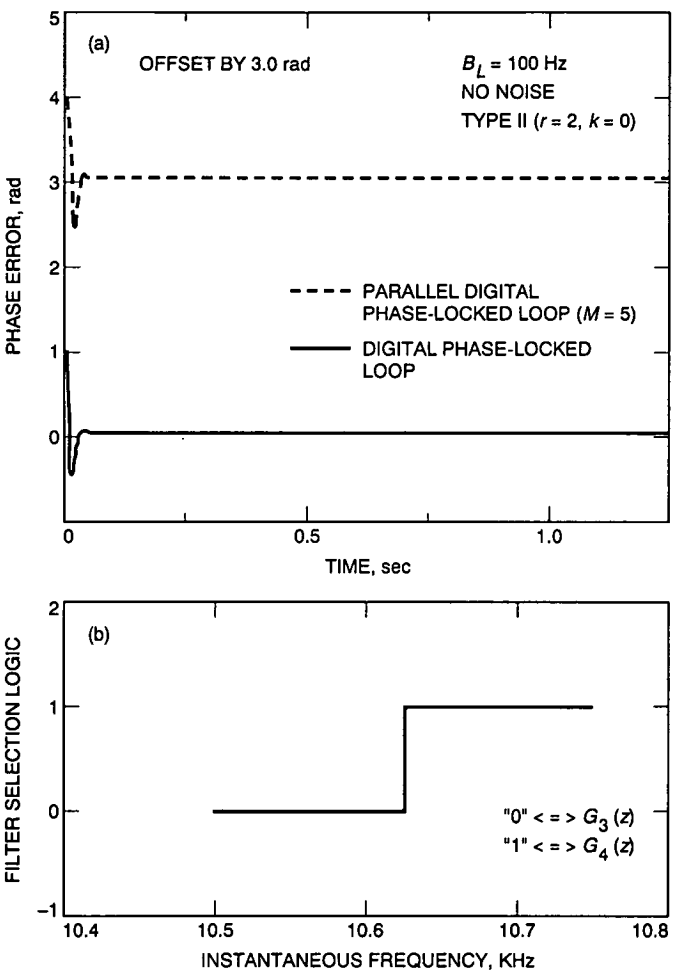


Fig. 10. PDPLL and DPLL simulation results for a 200-Hz/sec Doppler rate: (a) phase-error as a function of time and (b) multiplex control signal as a function of frequency.

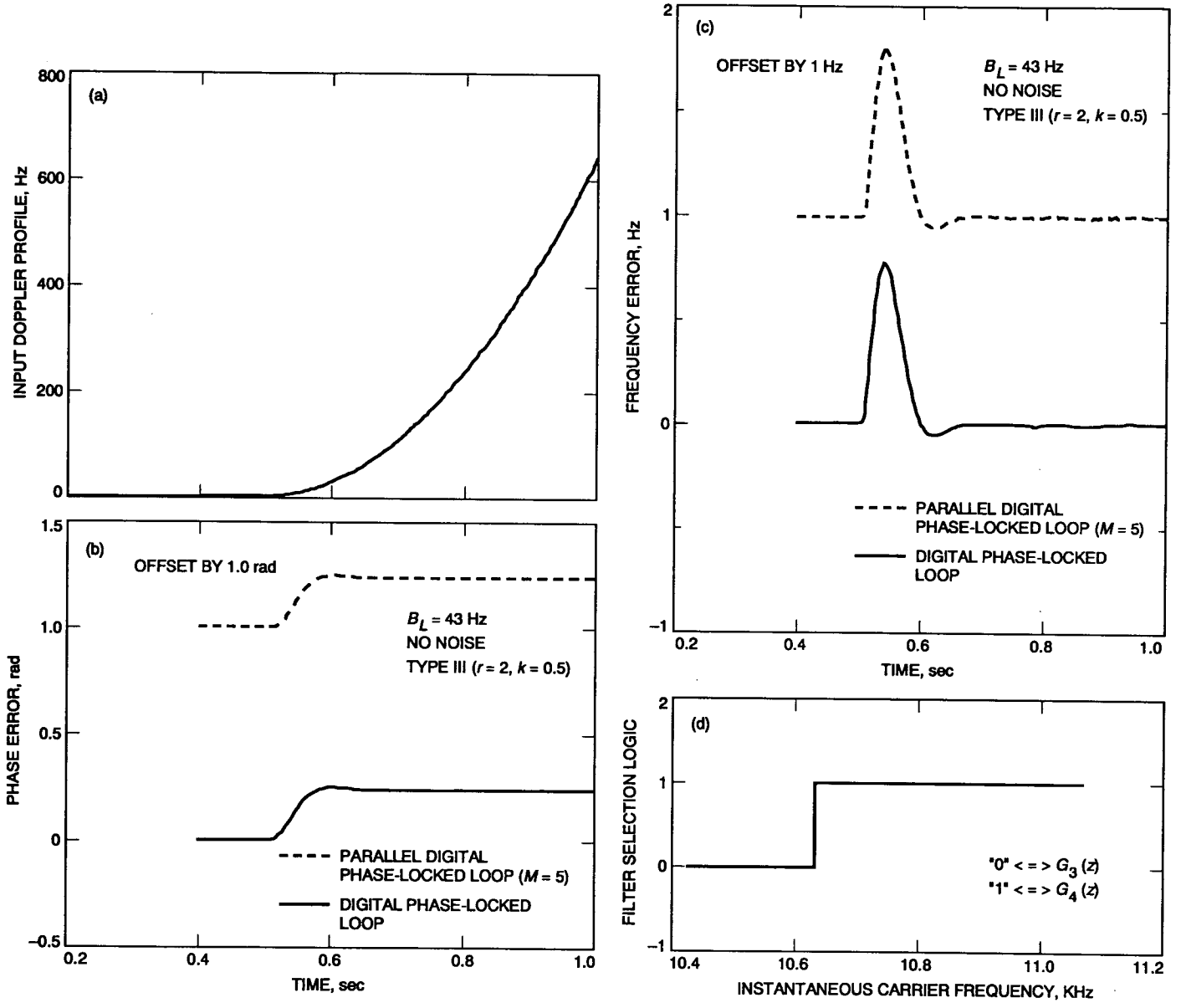


Fig. 11. PDPLL and DPLL simulation results for a jerk of 5145 Hz/sec^2: (a) input Doppler profile versus time; (b) phase-error versus time; (c) frequency error versus time; and (d) multiplex control signal as a function of frequency.

Appendix

Typically, the bandpass filters in a uniform DFB are obtained by frequency shifting the response of a low-pass prototype filter. Suppose that the prototype filter is an $N+1$ tap low-pass finite impulse response (FIR) filter with a passband in the frequency interval $[0, \omega_p^u]$ rad, where the upper passband cut-off frequency, ω_p^u, is given as

$$\omega_p^u = \frac{2\pi F_p^u}{f_s} \tag{A-1}$$

The parameters F_p^u and f_s are the analog cut-off frequency of the prototype filter and the sampling frequency, respectively, in hertz. The prototype filter transfer function, $H_p(\omega)$, is represented as

$$H_p(\omega) = |H_p(\omega)|e^{-j\omega N/2} \tag{A-2}$$

where $|H_p(\omega)|$ is the magnitude response and $-\omega N/2$ is the phase response of $H_p(\omega)$. The ith filter of a uniform DFB is obtained by multiplying the prototype filter impulse response, $h_p(n)$, by $2\cos(\omega_i n T_s)$. Hence, the transfer function of the ith filter is given as

$$G_i(\omega) = H_p(\omega - \omega_i) + H_p(\omega + \omega_i) \tag{A-3}$$

The center frequency of the ith filter was chosen for simulation purposes according to

$$\omega_i = F_1 + \frac{\Delta F}{2} i \tag{A-4}$$

where, using the same notation as in Section III, F_1 is the lower passband cutoff frequency of the DFB and ΔF is the length of the sub-bands.

As noted in the main text, the filters of the DFB should have continuous phase as a function of frequency. That is, the phase angle of $G_i(\omega)$ and $G_{i+1}(\omega)$ must be a multiple of 2π, i.e.,

$$\angle G_i(\omega) - \angle G_{i+1}(\omega) = 2k\pi \quad \forall \omega \in I_i \cap I_{i+1} \tag{A-5}$$

Substituting for the angles in Eq. (A-5) yields

$$\frac{|\omega_{i+1} - \omega_i| N}{2} = 2k\pi, \quad \forall \omega \in I_i \cap I_{i+1} \tag{A-6}$$

Using Eqs. (A-4) and (A-6), the following constraint must be satisfied when designing a phase continuous DFB

$$\frac{\Delta f N}{4} = 2k\pi \tag{A-7}$$

The prototype filter, $H_p(\omega)$, used in the simulations is shown in Fig. A-1(a). The digital cut-off frequency for this filter, ω_p^u, is equal to 0.03125 rad, and the parameter N is 256. The DFB was arbitrarily chosen to have five filters, which are shown in Fig. A-1(b) and summarized in Table 2. These filters have continuous phase because their center frequency separation $(\omega_{i+1} + \omega_i)/2 = (0.03125)2\pi \times N/2 = 128$ satisfies the constraint expressed in Eq. (A-6). The frequency response of the polyphase components of $G_4(z)$ is shown in Fig. A-2. The polyphase decomposition was obtained by applying Eq. (14) with $M = 5$ to the transfer function of $G_4(z)$.

Table A-1. The normalized center frequency, passband, and crossover frequency of the simulated DFB's.

Filter	ω_i, rad	Passband, rad	Crossover frequency, rad
$G_0(z)$	0.1875	[0.15625, 0.21875)	0.203125
$G_1(z)$	0.21875	[0.18755, 0.25)	0.234375
$G_2(z)$	0.25	[0.21875, 0.28125)	0.265625
$G_3(z)$	0.28125	[0.25, 0.3125)	0.296875
$G_4(z)$	0.3125	[0.28125, 0.34375)	—

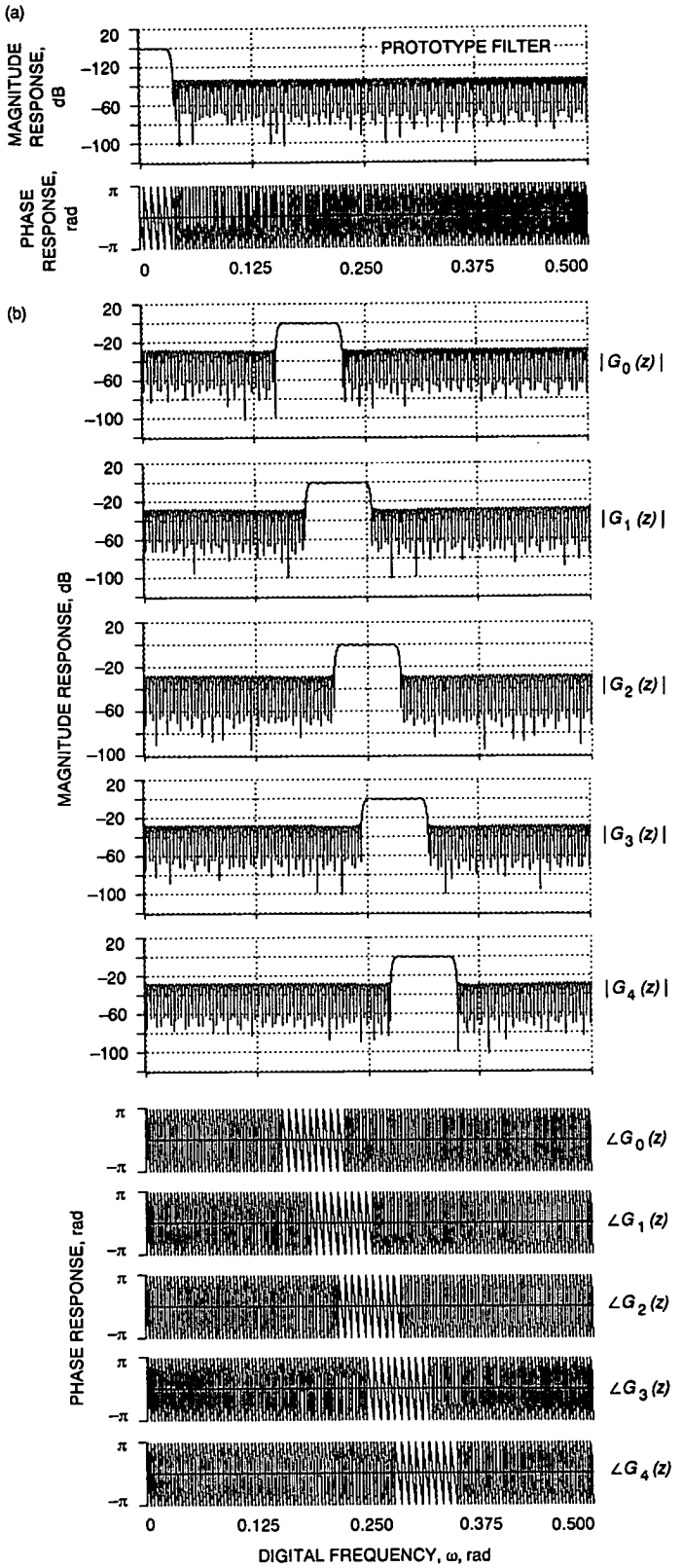


Fig. A-1. Magnitude and phase response as a function of the digital frequency for (a) the prototype filter and (b) the simulated DFB's.

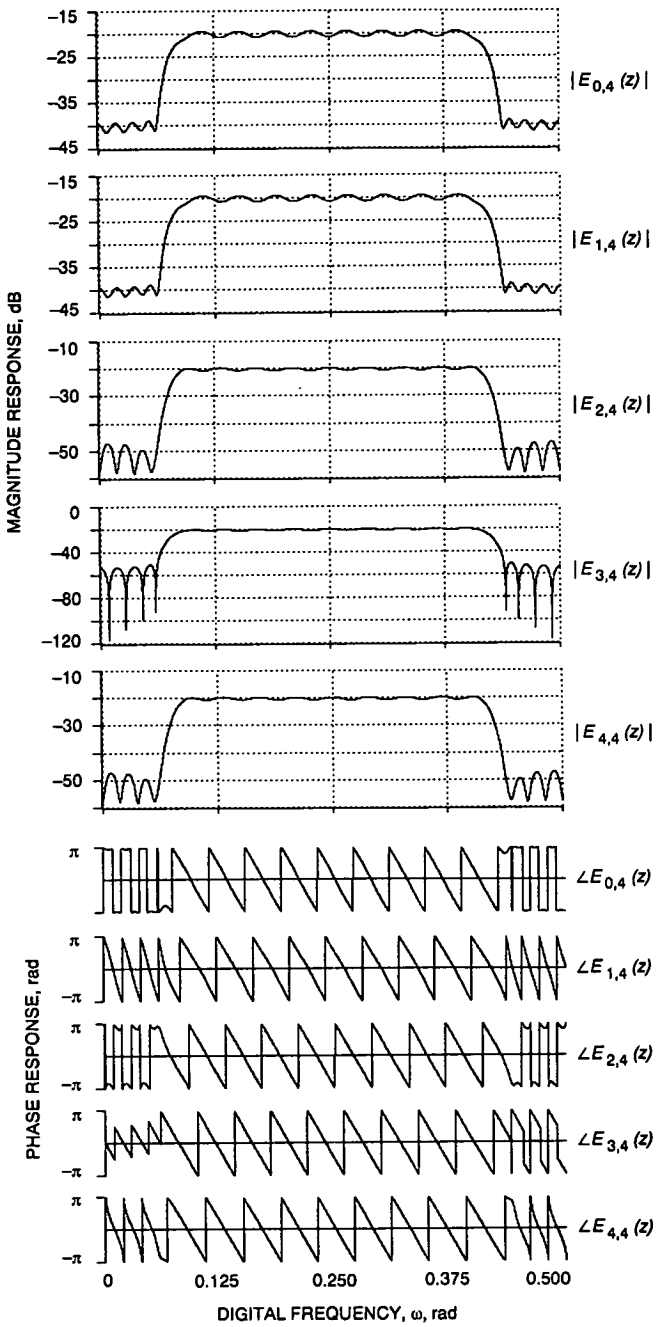


Fig. A-2. Magnitude and phase response of the polyphase components of $G_4(z)$.

Correlated Flux Densities From VLBI Observations With the DSN

R. F. Coker
Tracking Systems and Applications Section

Correlated flux densities of extragalactic radio sources in the very long baseline interferometry (VLBI) astrometric catalog are required for the VLBI tracking of Galileo, Mars Observer, and future missions. A system to produce correlated and total flux density catalogs has been developed to meet these requirements. A correlated flux density catalog of 274 sources, accurate to about 20 percent, has been derived from more than 5000 DSN VLBI observations at 2.3 GHz (S-band) and 8.4 GHz (X-band) using 43 VLBI radio reference frame experiments during the period 1989–1992. Various consistency checks have been carried out to ensure the accuracy of the correlated flux densities. All observations were made on the California–Spain and California–Australia DSN baselines using the Mark III wideband data acquisition system. A total flux density catalog, accurate to about 20 percent, with data on 150 sources, has also been created. Together, these catalogs can be used to predict source strengths to assist in the scheduling of VLBI tracking passes. In addition, for those sources with sufficient observations, a rough estimate of source structure parameters can be made.

I. The Need for Correlated Flux Densities

An astrometric catalog of compact extragalactic radio sources (generically referred to as quasars) is produced and maintained using VLBI observations with the DSN [1]. The primary purpose of this catalog is to provide a reference frame for angular tracking measurements of interplanetary spacecraft. Such tracking requires a catalog with a moderately high angular density of observable reference quasars along a spacecraft trajectory. In the standard tracking mode, it is generally desirable to have at least one observable reference quasar within about 10 deg of the spacecraft position at all times, since errors in the interferometric delay measurements grow with quasar–spacecraft separation.[1] In an alternate tracking mode, several quasars within a 30- to 50-deg field are observed to form a "local reference frame" in which the position of the spacecraft can be measured with extremely high accuracy [2].

Navigational scheduling requires a priori knowledge regarding which antennas are needed to observe, with an SNR adequate for detection, a specific quasar on a specific

[1] J. B. Thomas, "An Error Analysis for Galileo Angular Position Measurements with the Block II ΔDOR System," JPL Engineering Memorandum 335-26 (internal document), Jet Propulsion Laboratory, Pasadena, California, November 11, 1981.

baseline. Nominally, the correlated flux density detection limits, using the DSN's operational 250-kHz bandwidth navigation VLBI system (the Block I system), are about 0.2 Jy (1 Jy is 10^{-26} W/m^2/Hz) for two 70-m antennas, 0.4 Jy for one 70-m and one 34-m antenna, and 0.8 Jy for two 34-m antennas.[2] These limits apply at 8.4 GHz (X-band); at 2.3 GHz (S-band) the limits are somewhat higher at 0.2, 0.5, and 1.0 Jy, respectively, due to higher system temperatures. Since the correlated flux density of a given quasar can change from month to month by as much as 50 percent in extreme cases (especially at higher observing frequencies, such as X-band) [3], it is necessary to make correlated flux density estimates a routine part of reference frame data processing. A method has been developed which easily and quickly produces a correlated and total flux density catalog that is accurate enough (~20 percent) to determine which antenna pairs will be able to detect specified reference sources.

Flux density monitoring will also assist in quickly identifying sources that are inappropriate for reference frame determinations due to large variations in correlated flux density with baseline angle (as a result of extensive structure). It will also provide valuable time and source structure variability information. Prior to this work, no method existed for producing correlated flux density data from DSN Mark III Block II observations (see [4] for Mark III system details). In addition, the production of correlated flux density data from DSN Mark II observations was tedious and the results were often highly uncertain.[3] The new flux density estimates (with a detection limit of approximately 30 mJy), as well as the developed methodology, will be useful to anyone using the DSN to observe natural radio sources.

This article is organized in the following manner: Section II presents the expression for correlated flux density and gives details of the production and contents of the flux density catalogs. Section III describes the consistency checks and integrity tests carried out to verify the flux density results. The error budget is discussed in Section IV. Section V summarizes the results.

[2] J. S. Border, "An Error Analysis for Magellan Differential Delay Rate Measurements," JPL Engineering Memorandum 335-96 (internal document), Jet Propulsion Laboratory, Pasadena, California, February 23, 1987.

[3] I. Chen and J. S. Ulvestad, "Determination of Correlated Flux Densities from Mark II VLBI Experiments," JPL Interoffice Memorandum 335.3-88-102 (internal document), Jet Propulsion Laboratory, Pasadena, California, September 22, 1988.

II. Estimating Correlated Flux Densities and Total Flux Densities

For a radio source with nonzero angular size, the correlated flux density (i.e., that measured with an interferometer) will always be less than the total flux density. Roughly speaking, the correlated flux density is the strength of that part of the source which is angularly small as compared with an interference fringe on the sky. The angular width of an interference fringe is λ/B, where λ is the observing wavelength and B is the projected length of the interferometer baseline in the plane perpendicular to the source direction. The correlated flux density of a resolved source will vary as a function of baseline length and orientation because source structure is two-dimensional. In the absence of a complete, current description of the structure of a source, there is no way to specify a priori the source visibility (defined as the ratio of correlated flux density to total flux density) [5].

The correlated flux density, S_{cor}, of a source is given by[4]

$$S_{cor} = \frac{\sin(\rho\,\pi/2)}{b}\sqrt{\frac{T_{sys_1}T_{sys_2}}{G_1G_2}} \qquad (1)$$

where ρ is the correlation amplitude for data recorded with one-bit sampling, b is a normalization factor due to correlator and bandpass effects (with this definition, b for DSN Mark III data is greater than 1; see Section III.A for details), T_{sys_i} is the total on-source system temperature at station i in kelvins (K), and G_i is the antenna gain at station i in K/Jy [6]. In addition, the total flux density, S_{tot}, as measured by a single antenna with gain G, can be written as

$$S_{tot} = \frac{T_{on} - T_{off}}{G} = \frac{T_{ant}}{G} \qquad (2)$$

Here T_{on} and T_{off} are on-source and off-source system temperatures, respectively, and T_{ant} is the antenna temperature.

Using the new procedure reported here, correlated flux densities are obtained from DSN Mark III VLBI data that have been correlated at the JPL/CIT Block II correlator, with observable (e.g., delay and amplitude) extraction by

[4] R. F. Coker, "Mark III Correlated Flux Densities on the DSN," JPL Interoffice Memorandum, 335.6-91-028 (internal document), Jet Propulsion Laboratory, Pasadena, California, December 31, 1991.

the program FIT [7]. A master program takes the schedule file for the DSN antennas, correlation amplitudes, and observing frequencies from FIT, the measured system temperatures from antenna log files, and the modeled antenna gain, and produces a correlated flux density and a total flux density catalog. Details of this procedure are given below.

A. Amplitudes, Frequencies, and Hour Angle

The observing frequency and the correlation amplitude for each observation are read in from the FIT output file. The FIT data are assumed to be DSN Mark III Block II dual band (S- and X-bands), Mark III X-band only, or Mark II X-band only. One method of specifying baseline orientation is the interferometer hour angle (IHA) [6]; therefore, the correlated flux density of a resolved source will vary with the IHA. The source positions (right ascension and declination), observation times, and station locations are taken from the schedule file and used to calculate the IHA for each observation.

B. System Temperatures

System temperatures are obtained from strip-chart data files, data files produced by power meters, or from a mapping function using an estimated off-source zenith system temperature.

1. Strip charts. Strip-chart data files are lists of sources with system temperatures and error estimates read from strip charts. The independently measured zenith system temperature (recorded in the station logs) can be used to check for mistakes in the labeling, reading, or editing of the strip-chart file. In general, however, due to operator errors and instrument inaccuracies, strip-chart system temperature values remain unreliable. Also, the receiver used for strip-chart measurements is not the same receiver as that used for reference frame experiments. Therefore, other methods are preferred.

2. Power Meter Data. Power meter data are machine-readable total power values measured at the intermediate frequency expansion ports at a station. These time-tagged power readings are used to determine system temperature as a function of time.

For a given bandpass, system temperature is proportional to observed power [6]. Therefore, a single simultaneous power and system temperature measurement is adequate to convert the power readings to system temperatures. Using zenith system temperature and zenith power values that are in the header of each data file, the system temperature, T_{sys}, for a given observation is

$$T_{sys} = T_{zen} \frac{10^{(P/10)}}{10^{(P_{zen}/10)}} \tag{3}$$

where P is a power reading (in dBm) for that observation, P_{zen} is a power reading (in dBm) taken at zenith on cold sky, and T_{zen} (in K) is the zenith system temperature. There are usually about 40 power readings per 2- to 5-min observation; an average temperature is used for an observation's final system temperature. Errors are estimated from the scatter of the power data about a linear fit of the observation's power readings versus time, and the estimated error in P_{zen} and T_{zen}.

3. Modeling. If no system temperature data are available for a given frequency band and observing station, they are estimated using a temperature mapping function which relates the off-source system temperature at elevation angle *elev* to that at the zenith. An experimentally determined mapping function is used[5]

$$T_{sys} = A + N_1 \, e^{-N_2/(90^\circ - elev)} + \frac{B}{\sin(elev)} \tag{4}$$

where N_1 and N_2 depend on band and station, and *elev* is the elevation of the observation in degrees [8,9]. For stations without N_1 and N_2 values, a temperature version of the Chao mapping function is used [10].

If the total flux density of a source is known, the resulting antenna temperature is added to the modeled off-source system temperature. Antenna temperatures are usually less than 5 percent of the total system temperature, but for strong sources their contribution to the total system temperature is more significant.

4. Zenith System Temperatures. T_{zen} is calculated at the beginning of an experiment by taking a power reading of an ambient load and then of the sky; the difference, the Y factor (in dB), is a measurement of the receiver gain. Then T_{zen} is calculated

$$T_{zen} = \frac{T_{LNA} + T_{FO} + T_{amb}}{10^{(Y/10)}} \tag{5}$$

T_{amb} is the temperature of the ambient load. The noise temperatures used for the low-noise amplifier, T_{LNA}, and for the follow-on hardware (e.g., cables and waveguide),

[5] S. D. Slobin, *DSN Telecommunications Interfaces 34-Meter Antenna Subnets*, Document 810-5, Rev. D, vol. I, TCI-30, Jet Propulsion Laboratory, Pasadena, California, June 1, 1990.

T_{FO}, were incorrect before April 1991.[6] This introduced an error of a few K in T_{zen}. Formal precision measurement errors for T_{zen} appear small (~0.1 K), but the stability of T_{LNA}, T_{FO}, and T_{amb} values (as well as operator errors and weather effects) suggest that the overall accuracy of T_{zen}, as calculated above, is ~5 percent.[7,8]

The method for converting observed power to temperature, as described by Eqs. (3) and (5), implicitly assumes that for a given bandpass, power is directly proportional to the system temperature. For very large values of total power (corresponding to system temperatures of ~300 K), this assumption of detector linearity breaks down (at the 1- to 2-percent level).

C. Antenna Gains

The antenna gain for each station is modeled using experimentally determined functions; only a summary of the gain estimation procedure is given here (see [8] for details). The desired quantity is gain, G, in units of K/Jy (to use in Eq. (1)):

$$G = \frac{\eta\ Area}{2\ k} \tag{6}$$

where k is Boltzmann's constant, η is the antenna aperture efficiency, and $Area$ is the antenna's total physical collecting area. Variations in η across a given band are assumed to be negligible. The antenna aperture efficiency depends on the deformation of the antenna from its optimal shape, on weather conditions, and on the basic characteristics of the antenna surface. Thus, it can be written as

$$\eta = \eta_{max}\eta_{def}\eta_{atm} \tag{7}$$

where η_{max} is the aperture efficiency in a vacuum with no deviations from a perfectly shaped surface; η_{def} accounts for efficiency losses due to antenna gravitational deformation; and η_{atm} accounts for losses due to atmospheric attenuation. The value η_{max} is constant for a given antenna while η_{def} and η_{atm} depend on antenna elevation and weather conditions. The overall accuracy of the antenna gains modeled in this fashion (exclusive of antenna pointing effects; see Subsection III.C) is estimated at ~5 percent.

[6] Personal communication from T. Anderson, TDA Mission Support and DSN Operations Section, Jet Propulsion Laboratory, Pasadena, California, to C. Jacobs, member of technical staff, Tracking Systems and Applications Section, Jet Propulsion Laboratory, Pasadena, California, April 1, 1991.

[7] Slobin, op. cit.

[8] K. Clark and C. Cho, "Front-End Characteristics for NCB VLBI," JPL Interoffice Memorandum OEA.14/12.07.e.0, Jet Propulsion Laboratory, Pasadena, California, December 7, 1990.

III. Consistency Checks and Integrity Testing

A number of tests were carried out to verify the validity of the flux density catalogs. Since standard flux monitoring techniques (such as antenna nodding, repeated zenith calibrations, and standard source comparisons) are impractical for reference frame experiments due to time constraints, the internal and absolute accuracy of the DSN Mark III flux densities had to be determined in other ways.

A. Correlation Amplitudes and Their Normalization

The b-factor in Eq. (1) normalizes the correlation amplitude so that a signal with 100-percent correlation produces an amplitude of unity. Assuming identical bandpasses and a flat power spectrum and ignoring Doppler and polarization effects, one has, for some frequency ω [11],

$$b_\omega = a_1 \frac{(BW_0)\ B^2}{D_N} \tag{8}$$

where

a_1 = 1.176 for a three-level lobe rotator, one path [6]

BW_0 = nominal or Nyquist bandwidth (2 MHz for a single Mark III channel)

B^2 = bandpass amplitude at ω

$D_N = \int_0^\infty B^2\ d\omega$

Note that the correction for single-bit quantization ($\approx 2/\pi$) is included in the expression for S_{cor} in Eq. (1).

The value B^2 is included in FIT's calculation of the correlation amplitude, ρ. This results in a b-factor that is independent of frequency. The value of B^2 was empirically calculated from over 200 Mark III and Mark II observations; the results, which agree reasonably well with theoretical bandpass shapes, are listed in Table 1. These values are used in FIT to weight the amplitude of each bin to produce one composite amplitude for each band. Also included in Table 1 are values for D_N. The value D_N is constant at the 1- to 2-percent level for any given correlator configuration (e.g., number of delay lags per channel). The resulting b-factor for DSN Mark III 8 lag data is 1.34.

Variable Doppler shifting due to Earth rotation causes the two station bandpasses to be mismatched by up to 20 kHz at X-band; this has been ignored. The frequencies of the local oscillators are not adjusted to correct for this Doppler shift. In addition, there will be some small changes in b due to mismatched bandpasses as a result of polarization effects (the actual size of this effect is unknown but DSN specifications suggest it is ~1 percent). These effects, combined with aliasing of frequencies above 2 MHz, correlation of neighboring bits due to oversampling, and incomplete sideband rejection imply that b will vary at the ~2-percent level. These effects, plus errors in modeling, give an estimated maximum error of approximately 5 percent in the b-factor.

Two errors in the standard data path were uncovered during this work. The first involved the fractional bit-shift (FBS) correction at the correlator. If the FBS correction is not done, the upper and lower 10 percent of the bandpass lose ~10 to 15 percent in amplitude [12]. This causes a composite loss in band amplitude of ~4 percent. Before June 10, 1991, the Block II correlator was not doing this correction properly [13].

The second error occurred in FIT. Instead of summing the correlator sine and cosine counts, FIT averaged the amplitudes over time and frequency; this resulted in too large an amplitude for observations with low SNR. This has been corrected. FIT now also makes a correction for a nonzero residual fringe rate.[9] FIT amplitudes are normalized with respect to the number of lags, the number of time points, and the number of frequency bins containing good data.

To verify the above calculations for correlated flux densities, a short baseline experiment was conducted. On a baseline of 500 m, the correlated flux densities of the compact extragalactic radio sources used for VLBI astrometry are almost the same (within 10 percent) as the total flux densities. Figure 1 is a plot of the experiment's final visibilities (S_{cor}/S_{total}) versus scan number. Fitting a constant to the visibilities gave a ratio of 0.92. This suggests one or more of the following: The sources are all partially resolved, the antenna gains used in the data analysis are too high, the system temperatures are too low, or the b-factor is too large. However, the chi-square of a fit to unity is only 0.72, so changing the b-factor or antenna gain models is unjustified. *If the sources are resolved by a few percent on average, as expected, the* b*-factor is accurate to about 5 percent.*

[9] S. C. Unwin, "Amplitude Errors in AIPS Global Fringe-Fitting," Caltech Interoffice Memorandum (internal document), California Institute of Technology, Pasadena, California, October 26, 1988.

B. System Temperatures

The power meter program, QMVLBI, has a zenith system temperature calibration procedure, as described above. The calibration is performed at the start of an experiment. As time passes, the accuracy of the calibration degrades due to weather changes and receiver gain variations. The receiver gain was checked for variability over a 24-hour experiment and was found to be extremely stable (less than 1-percent variance in the power-to-temperature conversion factor). However, changes in the outside temperature over 24 hours can cause a significant (up to 5-percent) change in the zenith system temperature. Weather changes (such as a passing cloud) can cause a more substantial change in system temperature at low elevations; these are measured by the power meter.

The power meters measure power across the entire high electron mobility transistor (HEMT) bandwidth (400 MHz at X-band and 100 MHz at S-band). The 14 recorded 2-MHz channels (which are not contiguous) are scattered throughout this range. Therefore, any radio interference falling outside the observed channels (but within the HEMT bandwidth) will cause a spuriously high system temperature reading. In the next version of station software, it will be possible to take system temperature measurements for each channel,[10] thereby eliminating this potentially serious problem. Until then, it is assumed that variations in system temperature (as well as receiver gain) across an entire band are negligible.

System temperature measurements using the power meters, the strip charts, and results from modeling all agree within the expected errors. Figure 2 is an example of temperature versus air mass for each observation for an experiment with measurements by all three methods. Note that even in the absence of any station system temperature measurements during an experiment, models can be used to get system temperatures that agree with the power meter values to about 20 percent.

On- and off-source system temperature measurements are estimated in a somewhat subjective fashion. For strong sources, inspection of the power meter data clearly reveals when the antenna gets on source and when it moves off. The magnitudes of the observed sharp rise and fall of the system temperature values are used to estimate the antenna temperature (that part of the system temperature due to the source). For the case of weak sources, it is not always possible to determine the antenna temperature. In addition, the error in antenna temperature is

[10] *DSN Systems Requirements*, Document 820-13, Module VLB-13-5, Jet Propulsion Laboratory, Pasadena, California, 1990.

difficult to determine since the bulk of the error in system temperature is not statistical in nature. Repeatability suggests an error of about 15 percent in antenna temperature (down to the detection limit of 0.1 K). Note that variations in antenna temperature across the 400-MHz bandwidth spanned by the wideband amplifier at X-band due to a nonzero spectral index will be less than 3 percent for a typical VLBI source. Also, the next version of station software will be able to take power data with finer time and power resolution than the present power meters; this will lower the error in total flux density for weak sources.

C. Antenna Pointing and Gain

Total flux density measurements are determined only from antenna temperature and antenna gain. Therefore, using a strong source (so as to minimize antenna temperature errors) with a known total flux density, it is possible to estimate the antenna gain. Some observations of the Very Large Array (VLA) calibrator source 3C 286 were used to compare the modeled antenna gains with gains estimated in this fashion. The resulting total flux densities were consistent at the 20-percent level.

An error in antenna pointing will yield reduced antenna temperatures and correlation amplitudes, giving spuriously low values of both total and correlated flux densities. Reference frame experiments use blind pointing; pointing correction tables, squint corrections, and refraction correction calculations are estimated to yield a blind pointing accuracy of ~10 mdeg for the 34-m antennas.[11] The refraction correction model used at the antennas was verified to agree, to within 1 mdeg, with refraction calculations using real-time meteorological parameters.[12] However, under similar circumstances (observations of the same source closely spaced in time and at approximately the same elevation, system temperature, and IHA), different X-band observations are seen to produce different correlated flux densities, sometimes by as much as a factor of 2. This is probably due to pointing errors; a pointing error of 30 mdeg by one station will reduce that antenna's X-band gain by ~50 percent (S-band gain is much less affected by pointing errors of this magnitude). This in turn yields a correlated flux density that is ~30 percent too small. The cause of this pointing problem is as yet unknown.

[11] D. Girdner, personal communication, Antenna Performance Engineer, TDA Mission Support and DSN Operations Section, February 15, 1992.

[12] G. E. Lanyi, "Atmospheric Refraction Corrections to Antenna Pointing at 1 Millidegree Accuracy," JPL Interoffice Memorandum 335.3-89-026 (internal document), Jet Propulsion Laboratory, Pasadena, California, March 24, 1989.

D. Comparison of Results With a Known Structure

The source 4C 39.25 has been a fairly well-studied source at high frequencies (see, for example, [14]). It is known to be extended in the east–west direction with knots in its structure. At X-band it is also highly time-variable [15]. Therefore, X-band correlated flux densities on two baselines on a monthly basis can give little source structure detail; indeed, it is found that measurements vary by more than a factor of 2 from month to month. The picture at S-band is much clearer; Fig. 3 is a plot of measured correlated flux density versus the projected east–west baseline length. The smoothness of the curve implies a smooth underlying source structure (to the limit of the resolution of the interferometer) and little time variability. The fact that the correlated flux density depends solely on the absolute value of the east–west projected baseline length suggests that the source is not resolved in the north–south direction. This is completely consistent with maps of 4C 39.25 that are in the literature.

IV. Error Budget

The variance, σ^2, in measured correlated flux density (for $\rho\pi/2 \ll 1$), derived from Eq. (1), is:

$$\sigma^2 = {S_{cor}}^2 \left(\frac{{\sigma_\rho}^2}{\rho^2} + \frac{{\sigma_b}^2}{b^2} + \frac{{\sigma_{T_1}}^2}{4\,{T_1}^2} + \frac{{\sigma_{T_2}}^2}{4\,{T_2}^2} + \frac{{\sigma_{G_1}}^2}{4\,{G_1}^2} + \frac{{\sigma_{G_2}}^2}{4\,{G_2}^2} \right) \qquad (9)$$

where T_i is the system temperature for an observation at station i and G_i is the antenna gain for station i. Note that the error is proportional to the correlated flux density.

The total error in correlated flux density comes from a number of sources. Principal sources are the errors in the determination of system temperatures (about 5 percent when using the power meters), errors in gain due to weather and frequency variations within an experiment (about 5 percent), errors in ρ (typically less than 2 percent, determined by the SNR of the observation), and errors in b (approximately 5 percent).

A summary of total estimated errors from all sources, assuming uncorrelated errors as shown in Eq. (9), is given in Table 2. Note that in addition to the errors listed, time and structure variability can, in extreme cases, cause up to a factor of 2 error in correlated flux density values extrapolated from measurements at previous epochs.

V. Summary

With the methods described here, it is possible to provide correlated and total flux density measurements accurate to ~20 percent. This accuracy is sufficient to determine which antennas are necessary to observe a given source on a given baseline at a given time. A complete correlated and total flux density catalog can now be built. However, to get better than 20-percent accuracy, the pointing problems of the DSN antennas must be corrected.

Acknowledgments

Dave Fort was of great help in explaining the Block II correlator. Steve Lowe assisted with many discussions on correlated amplitudes. Roger Linfield provided invaluable comments and critiques.

References

[1] O. J. Sovers, C. D. Edwards, C. S. Jacobs, G. E. Lanyi, K. M. Liewer, and R. N. Treuhaft, "Astrometric Results of 1978–1985 Deep Space Network Radio Interferometry: The JPL 1987-1 Extragalactic Source Catalog," *Astron. J.*, vol. 95, pp. 1647–1658, 1988.

[2] R. N. Treuhaft and S. T. Lowe, "A Nanoradian Differential VLBI Tracking Demonstration," *TDA Progress Report 42-109*, vol. January–March 1992, Jet Propulsion Laboratory, Pasadena, California, pp. 1–15, May 15, 1992.

[3] T. F. Haddock, H. D. Aller, and M. F. Aller, "Frequent Observations of Extragalactic Compact Sources at 24 GHz," *Astron. J.*, vol. 93, pp. 1356–1367, June 10, 1987.

[4] A. E. Rogers and T. A. Clark, *Mark III Data Acquisition Terminal,* Crustal Dynamics Project, Goddard Space Flight Center, Greenbelt, Maryland, January 1982.

[5] J. S. Ulvestad and R. P. Linfield, "The Search for Reference Sources for ΔVLBI Navigation of the Galileo Spacecraft," *TDA Progress Report 42-84*, vol. October–December 1985, Jet Propulsion Laboratory, Pasadena, California, pp. 152–163, February 15, 1986.

[6] A. R. Thompson, J. M. Moran, and G. W. Swenson, Jr., *Interferometry and Synthesis in Radio Astronomy*, New York: John Wiley and Sons, pp. 10, 88, 248, and 305, 1986.

[7] S. T. Lowe, *Theory of Post Block II VLBI Observable Extraction*, JPL Publication 92-7, Jet Propulsion Laboratory, Pasadena, California, July 15, 1992.

[8] P. H. Richter and S. D. Slobin, "DSN 70-Meter Antenna X- and S-Band Calibration Part I: Gain Measurements," *TDA Progress Report 42-97*, vol. January–March 1989, Jet Propulsion Laboratory, Pasadena, California, pp. 314–346, May 15, 1989.

[9] S. D. Slobin and P. H. Richter, "DSN 70-Meter Antenna X- and S-Band Calibration Part II: System Noise Temperature Measurements and Telecommunications Link Evaluation," *TDA Progress Report 42-97*, vol. January–March 1989, Jet Propulsion Laboratory, Pasadena, California, pp. 352–366, May 15, 1989.

[10] C. C. Chao, *The Tropospheric Calibration Model for Mariner Mars 1971*, Technical Report 32-1587, Jet Propulsion Laboratory, Pasadena, California, pp. 61–76, March 1974.

[11] J. B. Thomas, *Interferometry Theory for the Block II Processor*, JPL Publication 87-29, Jet Propulsion Laboratory, Pasadena, California, October 15, 1987.

[12] M. L. Meeks, ed., *Methods of Experimental Physics*, vol. 12, Part C, Astrophysics, Radio Observations, New York: Academic Press, 1989.

[13] S. C. Unwin, ed., *JPL/Caltech Block II VLBI Correlator Newsletter*, vol. 2, no. 3, California Institute of Technology, Pasadena, California, July 1991.

[14] D. Shaffer and A. Marscher, "4C 39.25: Superluminal Motion Between Stationary Components," in *Superluminal Radio Sources*, edited by J. A. Zensus and T. J. Pearson, New York: Cambridge University Press, pp. 67–71, 1987.

[15] J. M. Marcaide, A. Alberdi, P. Elosegui, A. P. Marscher, Y. F. Zhang, D. B. Shaffer, C. J. Schalinski, A. Witzel, N. Jackson, and G. Sandell, "Detection of a New Component in the Peculiar Superluminal Quasar 4C 39.25," in *Parsec-scale Radio Jets: Proceedings of a Workshop Held at the National Radio Astronomy Observatory, Socorro, New Mexico, October 17–18, 1989*, edited by J. A. Zensus and T. J. Pearson, New York: Cambridge University Press, pp. 59–65, 1990.

Table 1. *b*-factor bandpass parameters for the Block II correlator.

	Mark III				Mark II			
	8 lags		16 lags		8 lags		16 lags	
	Empirical	Theory	Empirical	Theory	Empirical	Theory	Empirical	Theory
B^2 for bin 1	0.840	0.899	0.830	0.889	–	0.907	1.050	0.892
bin 2	1.002	1.002	0.992	1.015	–	0.993	1.122	1.015
bin 3	1.023	0.997	0.987	0.994	–	1.019	1.020	0.994
bin 4	0.643	0.604	1.003	1.003	–	0.523	1.004	1.004
bin 5	–	–	1.011	0.997	–	–	0.978	0.996
bin 6	–	–	1.001	0.977	–	–	0.935	0.999
bin 7	–	–	0.877	0.813	–	–	0.913	0.804
bin 8	–	–	0.467	0.408	–	–	0.337	0.244
D_N	0.877	0.876	0.896	0.887	–	0.871	0.921	0.869

Notes:

B^2 and D_N are relative to a square bandpass of amplitude unity.

No S- or X-band Mark II 8-lag data were available for analysis.

In 8-lag mode, bins 5 through 8 contain no signal.

Empirical fits were made relative to bin 2 (for 8 lags) or bin 4 (16 lags).

Mark II theoretical calculations based on an 11-pole Butterworth filter.

Table 2. Error sources.

Error source	Resulting fractional error, percent	
	X-band	S-band
Gain (each station)		
Weather and modeling	5	5
Pointing errors	0 to 50	0 to 10
b-factor	5	5
Amplitudes		
Signal to noise	0 to 15	0 to 15
FIT approximations	5	5
System temperatures (each station)		
Using model	10 to 15	15 to 20
Using strip chart	10	15
Using power meters	5	10
Estimated total (using power meters with T_{zen} at both stations)		
Clear weather, high elevation, high SNR source	10	10
Poor weather, low elevation, low SNR source	30	20

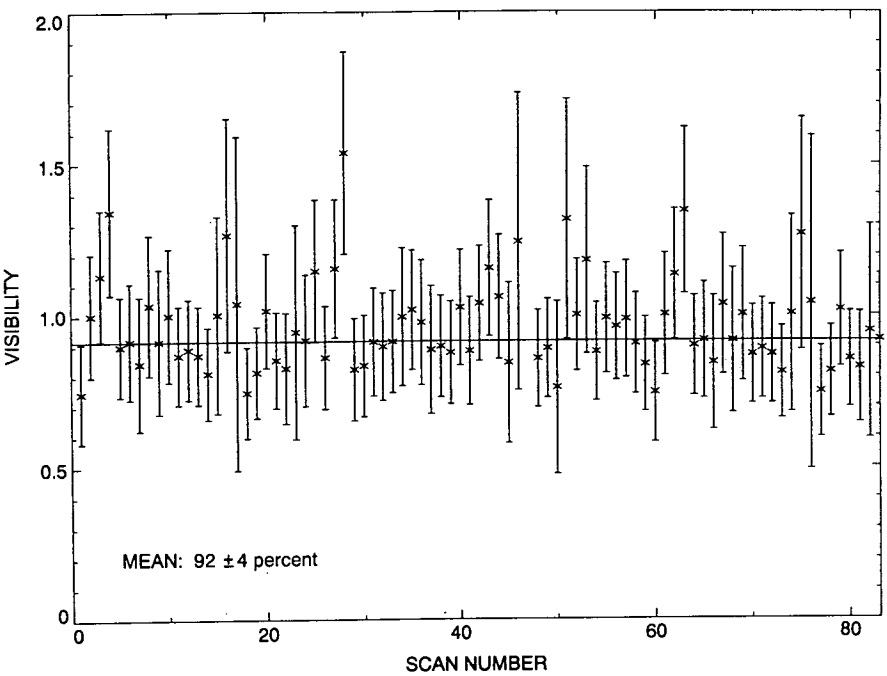


Fig. 1. The final visibilities (S_{cor}/S_{total}) for the short-baseline experiment 92CC124 using DSS 14 and DSS 15. The values for the total flux density, S_{total}, were taken from antenna temperature measurements at DSS 15. A weighted fit of zero slope was made to the data. The error bars are calculated from each observation's error in total flux density and in correlated flux density. For some scans, the antenna temperature was comparable to the resolution limit of 0.1 K; hence the large error bars.

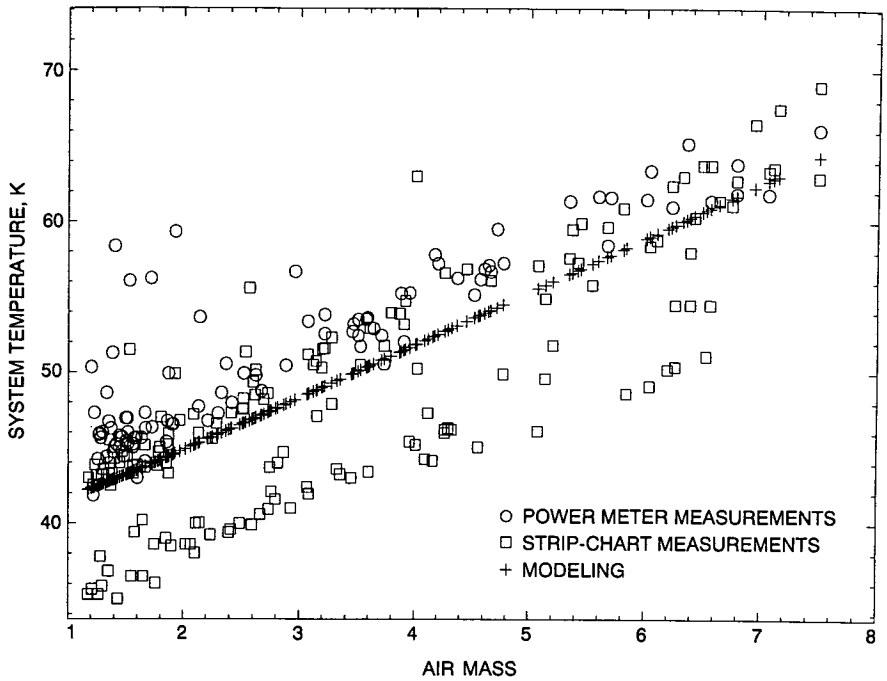


Fig. 2. A plot of X-band (8.4-GHz) system temperature versus air mass (1/sin (elevation)) for the three different methods for the experiment 90CA049. In most cases, agreement of the model with the power meter data is within 20 percent. The strip-chart data show an unexplained (and probably unreal) drop in system temperature halfway through the experiment.

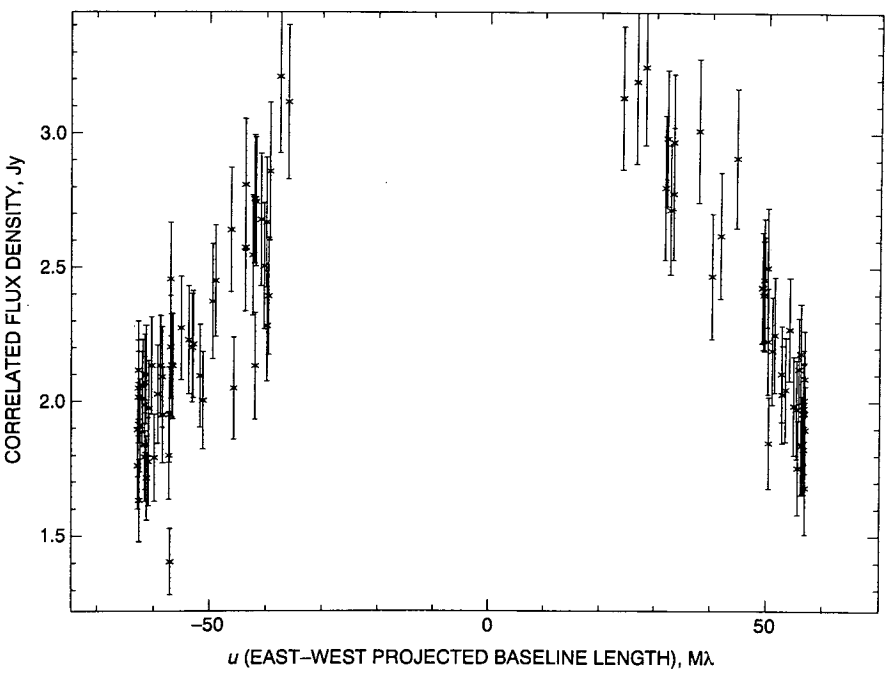


Fig. 3. A plot of correlated flux density versus *u* for the source 4C 39.25 at 2.3 GHz. The smoothness of the curve suggests smooth source structure. The dependence of correlated flux density solely on the absolute value of *u* suggests that the source is not resolved in the north–south direction but that it is extended in the east–west direction.

The Effects of Correlated Noise in Intra-Complex DSN Arrays for S-Band Galileo Telemetry Reception

R. J. Dewey
Navigation Systems Section

A number of the proposals for supporting a Galileo S-band (2.3-GHz) mission involve arraying several antennas to maximize the signal-to-noise ratio (and bit rate) obtainable from a given set of antennas. Arraying is no longer a new idea, having been used successfully during the Voyager encounters with Uranus and Neptune. However, arraying for Galileo's tour of Jupiter is complicated by Jupiter's strong radio emission, which produces correlated noise effects. This article discusses the general problem of correlated noise due to a planet, or other radio source, and applies the results to the specific case of an array of antennas at the DSN's Tidbinbilla, Australia, complex (DSS 42, DSS 43, DSS 45, and the yet-to-be built DSS 34). The effects of correlated noise are highly dependent on the specific geometry of the array and on the spacecraft–planet configuration; in some cases, correlated noise effects produce an enhancement, rather than a degradation, of the signal-to-noise ratio. For the case considered here—an array of the DSN's Australian antennas observing Galileo and Jupiter—there are three regimes of interest. If the spacecraft–planet separation is $\lesssim$75 arcsec, the average effect of correlated noise is a loss of signal to noise ($\sim$0.2 dB as the spacecraft–planet separation approaches zero). For spacecraft–planet separations $\gtrsim$75 arcsec, but $\lesssim$400 arcsec, the effects of correlated noise cause signal-to-noise variations as large as several tenths of a decibel over time scales of hours or changes in spacecraft–planet separation of tens of arcseconds; however, on average its effects are small (<0.01 dB). When the spacecraft is more than 400 arcsec from Jupiter (as is the case for about half of Galileo's tour), correlated noise is a <0.05-dB effect.

I. Introduction

It is now becoming common practice to array a number of antennas for telemetry reception in order to increase the effective aperture [1,2]. The signal-to-noise ratio (SNR) improvement obtained by arraying is straightforward to calculate when a pointlike spacecraft is the only source in the array's field of view. However, the calculation becomes complicated when an additional hot body (such as a radio-loud planet) is in the beam since, unlike system noise, the incoming radiation from this hot body is correlated at different antennas. This article presents a calculation

of the effect of correlated noise on the SNR achieved by a given array configuration. Other aspects of this topic have been discussed in [3] and in references therein.

This article is organized as follows: Sections II and III present a general formulation of the problem, Section IV considers the special case of an array of N identical antennas, and Section V looks at the specific case of arraying the DSN complex in Australia for Galileo S-band (2.3-GHz) telemetry reception. To make the article more readable, many of the detailed calculations have been relegated to the appendices. Much of the discussion parallels Hjellming's treatment of the Very Large Array (VLA) [4].

II. Outline of the Problem

Consider an array of N (not necessarily identical) antennas. The on-axis gain[1] of the ith antenna is denoted by G_i and its system temperature (looking at an empty field) is denoted by T_i. The direction toward a point on the sky can be described by a unit vector $\hat{\mathbf{s}}$, with $\hat{\mathbf{s}}_0$ pointing toward a source at the array's field center.[2] The shape of the ith antenna beam can be described by the field pattern,[3] a dimensionless function $f_i(\hat{\mathbf{s}} - \hat{\mathbf{s}}_0)$ such that $f_i^2(\hat{\mathbf{s}} - \hat{\mathbf{s}}_0)G_i$ describes the effective gain of the ith antenna in a direction $\hat{\mathbf{s}} - \hat{\mathbf{s}}_0$ from the beam center. In general, G_i, T_i, and f_i may all be elevation dependent. This dependence is not considered here; it can easily be included in quantitative calculations. The geometry of the array of N antennas can be described by the set of baseline vectors between each antenna pair $\mathbf{B}_{ik} = -\mathbf{B}_{ki}$, where $\mathbf{B}_{ik}$ denotes the vector from the ith antenna to the kth antenna. The basic geometry is shown in Fig. 1.

It is assumed that the spacecraft is transmitting a narrow-band, polarized signal with a total Earth-received power per unit area of $\mathcal{P}_{s/c}$ and that the planet has a total, unpolarized flux density (power per unit area per unit bandwidth) S_P. The angular distribution of the planet's radio emission is described by the brightness distribution (flux density per unit solid angle) $I_P(\hat{\mathbf{s}})$, which is assumed to be constant in time and only slowly varying with frequency (i.e., constant across the observing bandwidth).

Fairly simple assumptions are made about the method of arraying: The incoming RF signal at the ith antenna is mixed to baseband, after which a phase shift is inserted. This phase shift has a component ϕ_i^m (the model phase) that can be used to compensate, in real time, for phase differences between antennas (see Appendices A and C for details). The baseband signal is delayed by a model delay τ_i^m, and finally, the delayed signals from the N antennas are summed with appropriate weighting (see Appendix D). It is assumed that the signal is filtered through a rectangular baseband filter of width $\Delta\nu$; for an upper sideband system (which is assumed throughout), the sky frequency at the center of the passband is $\nu_{sky}^c = \nu_{lo} + \Delta\nu/2$, where ν_{lo} is the mixer frequency. It is also assumed that the model delays are chosen so as to compensate for the delay that a signal from the direction $\hat{\mathbf{s}}_0$ (the field center) suffers between an arbitrary reference point and the ith antenna.

$$\tau_i^m = \frac{\mathbf{r}_i \cdot \hat{\mathbf{s}}_0}{c} \tag{1}$$

where $\mathbf{r}_i$ is the vector from the reference point to antenna i. A simple block diagram of the signal path is shown in Fig. 2.

The total arrayed power can be written as the sum of the arrayed power from each of the three sources $P_\Sigma = P_\Sigma^{s/c} + P_\Sigma^P + P_\Sigma^N$, where $P_\Sigma^{s/c}, P_\Sigma^P, P_\Sigma^N$, represent, respectively, the arrayed power from the spacecraft, from the planet, and from system noise.[4] The effective SNR of the array is proportional to the ratio of the arrayed spacecraft power to the total arrayed power; in the low signal limit, this reduces to the ratio of $P_\Sigma^{s/c}$ to the power received from all sources other than the spacecraft (in this case, the planet and system noise). A figure of merit β can be defined for an array; β is analogous to G/T for a single dish and quantifies how, for a given spacecraft power ($\mathcal{P}_{s/c}$) an observing bandwidth ($\Delta\nu$), the SNR varies with the following array configuration:

$$\beta = \frac{\Delta\nu}{2\mathcal{P}_{s/c}} \frac{P_\Sigma^{s/c}}{P_\Sigma} \approx \frac{\Delta\nu}{2\mathcal{P}_{s/c}} \frac{P_\Sigma^{s/c}}{P_\Sigma^P + P_\Sigma^N} \tag{2}$$

[1] Throughout this article, an astronomer's definition of antenna gain is used (commonly measured in units of K per Jy) of $G = \epsilon A/2k_B$, where ϵ is the dimensionless antenna efficiency, A is the physical antenna area, and k_B is Boltzmann's constant.

[2] Here, $\hat{\mathbf{s}}_0$ is defined as the nominal direction in which the antennas in the array are pointing, which is the direction assumed when model delays are calculated (see Appendix A). As discussed in Appendix C, it is not necessarily the phase center of the array.

[3] In general, the field pattern is considered to be a complex quantity (e.g., [5, p. 27]), but for the purposes of this article, it is sufficient to assume that it is real.

[4] System noise includes receiver noise, sky background, and ground pickup—all sources of unwanted noise other than individual, identifiable sources (such as a planet) in the field of view.

The term $P_\Sigma^{s/c}/P_\Sigma$ is the SNR; the term $\Delta\nu/2\ P_\Sigma^{s/c}$ is a normalization factor that removes from β any dependence on quantities other than array and source geometry.

III. General Expression for β

As discussed in Appendix A, the summed baseband voltage can be written in the form

$$v_\Sigma(\nu,t) = \sum_{i=1}^{N} W_i e^{-i\phi_i^m} v_i(\nu, t+\tau_i^m) \tag{3}$$

where $v_i(\nu,t)$ is the baseband voltage given by Eq. (A-4), ν is the baseband frequency, and W_i is the weighting factor.

The arrayed power is then

$$\begin{aligned} P_\Sigma &= \left\langle \int_0^{\Delta\nu} d\nu\, |v_\Sigma(\nu,t)|^2 \right\rangle_{\mathcal{T}} \\ &= \left\langle \int_0^{\Delta\nu} d\nu \left| \sum_{i=1}^{N} W_i e^{-i\phi_i^m} v_i(\nu, t+\tau_i^m) \right|^2 \right\rangle_{\mathcal{T}} \\ &= \sum_{i=1}^{N} W_i^2 P_i + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \rho_{ik} \end{aligned} \tag{4}$$

where the angle brackets denote time averaging and $\mathcal{T}$ is the interval over which the time average is taken; P_i is the single dish power from the ith antenna,

$$P_i = \left\langle \int_0^{\Delta\nu} d\nu\, |v_i(\nu,t)|^2 \right\rangle_{\mathcal{T}} \tag{5}$$

and ρ_{ik} is the unnormalized correlation between the ith and kth antennas,

$$\rho_{ik} = \left\langle \int_0^{\Delta\nu} d\nu\, e^{i(\phi_k^m - \phi_i^m)} v_i(\nu, t+\tau_i^m) v_k^*(\nu, t+\tau_k^m) \right\rangle_{\mathcal{T}} \tag{6}$$

As can be seen in Eqs. (A-11) and (A-18), both P_i and ρ_{ik} can be written as the sum of contributions from the spacecraft ($P_i^{s/c}$, $\rho_{ik}^{s/c}$), the planet (P_i^P, ρ_{ik}^P), and random noise (P_i^N, ρ_{ik}^N), so Eq. (4) can be written

$$\begin{aligned} P_\Sigma = &\sum_{i=1}^{N} W_i^2 (P_i^{s/c} + P_i^P + P_i^N) \\ &+ \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k (\rho_{ik}^{s/c} + \rho_{ik}^P + \rho_{ik}^N) \end{aligned} \tag{7}$$

Thus, P_Σ, the total power in the arrayed signal, can be written as the sum of contributions from the three sources (the spacecraft, the planet, and system noise), $P_\Sigma = P_\Sigma^{s/c} + P_\Sigma^P + P_\Sigma^N$, with

$$P_\Sigma^{s/c} = \sum_{i=1}^{N} W_i^2 P_i^{s/c} + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \rho_{ik}^{s/c} \tag{8}$$

$$P_\Sigma^P = \sum_{i=1}^{N} W_i^2 P_i^P + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \rho_{ik}^P \tag{9}$$

$$\begin{aligned} P_\Sigma^N &= \sum_{i=1}^{N} W_i^2 P_i^N + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \rho_{ik}^N \\ &\approx \sum_{i=1}^{N} W_i^2 P_i^N \end{aligned} \tag{10}$$

Because the phases of the individual ρ_{ik}^N terms in Eq. (10) are random, their sum is always small (on average, zero) and was dropped. However, the sums over ρ_{ik}^P and $\rho_{ik}^{s/c}$ are not necessarily small; in fact, a key aspect of arraying is choosing the inserted model phases to ensure that the sum over $\rho_{ik}^{s/c}$ is maximized (see Appendix C). Expressions for $P_i^{s/c}$, P_i^P, and P_i^N are derived in Appendix A, Section II, and given by Eqs. (A-13), (A-14), and (A-15), respectively. Expressions for $\rho_{ik}^{s/c}$ and ρ_{ik}^P are derived in Appendix A, Section C, and given by Eqs. (A-25) and (A-28). Substituting Eqs. (8)–(10) into Eq. (2) gives

$$\beta = \frac{\Delta\nu}{2\mathcal{P}_{s/c}} \times \frac{\displaystyle\sum_{i=1}^{N} W_i^2 P_i^{s/c} + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \rho_{ik}^{s/c}}{\displaystyle\sum_{i=1}^{N} W_i^2 (P_i^N + P_i^P) + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \rho_{ik}^P} \tag{11}$$

The phases of $\rho_{ik}^{s/c}$ and ρ_{ik}^{P} depend on the quantity $\delta\phi_i - \delta\phi_k = \phi_i - \phi_i^m - \phi_k + \phi_k^m$, where ϕ_i, ϕ_k are the actual antenna-based phases, which include both hardware and atmospheric effects, and ϕ_i^m, ϕ_k^m are the model phases. If no attempt is made to compensate for the individual antenna phases (i.e., if $\phi_i^m = \phi_k^m = 0$), the phases of $\rho_{ik}^{s/c}$ and ρ_{ik}^{P} are uncorrelated from baseline to baseline, and in both Eqs. (8) and (9) the sums over the cross-correlation will be small (zero on average). However, as discussed in Appendix C, the total phase of $\rho_{ik}^{s/c}$ will be zero for all i, k, if the values of ϕ_i^m, ϕ_k^m are chosen to satisfy Eq. (C-1):

$$\begin{aligned}\delta\phi_i - \delta\phi_k &= \phi_i - \phi_i^m - \phi_k + \phi_k^m \\ &= -2\pi\nu^{s/c}\frac{\mathbf{B}_{ik}}{c}\cdot[\hat{\mathbf{s}}_{s/c} - \hat{\mathbf{s}}_0]\end{aligned}$$

When ϕ_i^m, ϕ_k^m are chosen to satisfy the above expression, all the terms in the double sum in Eq. (8) add in phase, and for large N, the sum over $\rho_{ik}^{s/c}$ contributes significantly more to $P_\Sigma^{s/c}$ than does the sum over $P_i^{s/c}$ (since $\rho_{ik} \sim \sqrt{P_i P_k}$). Phases chosen to satisfy Eq. (C-1) maximize $P_\Sigma^{s/c}$, and except in pathological[5] cases this choice maximizes β. The process of determining and inserting these values of ϕ_i^m, ϕ_k^m is referred to as phasing the array. In the following, the subscript ϕ is used to denote expressions that assume that the array is phased on the spacecraft.

Unfortunately, the values of ϕ_i^m, ϕ_k^m, which maximize the sum over $\rho_{ik}^{s/c}$, may also de-randomize the phases of ρ_{ik}^{P}; as a consequence, sums over ρ_{ik}^{P} in Eqs. (4) and (11) can become significantly nonzero. This is the source of the correlated noise contribution from the planet.

Equations (A-13), (A-14), (A-15), (A-25), and (A-28) can be used to expand Eqs. (8), (9), and (10), and if the array is phased on the spacecraft [Eq. (C-1) is satisfied], the arrayed power from the spacecraft, the planet, and system noise can be written, respectively, as

$$P_{\Sigma\phi}^{s/c} = 2k_B\mathcal{P}_{s/c}\left[\sum_{i=1}^{N} W_i^2 G_i + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \sqrt{G_i G_k}\right] \quad (12)$$

$$P_{\Sigma\phi}^{P} = \Delta\nu k_B S_P\left[\sum_{i=1}^{N} W_i^2 \bar{f}_{i_P}^2 G_i + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \bar{f}_{i_P}\bar{f}_{k_P}\sqrt{G_i G_k}\,\mathcal{F}_{ik}^{P}\right] \quad (13)$$

$$P_{\Sigma\phi}^{N} = \Delta\nu k_B \sum_{i=1}^{N} W_i^2 T_i \quad (14)$$

In the above expressions, $\bar{f}_{i_P}$, $\bar{f}_{k_P}$ represent the average value of the beam pattern in the direction of the planet, and $\mathcal{F}_{ik}^{P}$ [defined by Eq. (A-27)] is a dimensionless, complex quantity, with magnitude less than or equal to unity, which depends only on the array geometry and the structure of the planet (or other background source).

By inserting Eqs. (12), (13), and (14) into Eq. (2), the expression for β reduces to

$$\beta_\phi = \frac{\left[\sum_{i=1}^{N} W_i^2 G_i + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \sqrt{G_i G_k}\right]}{S_P\left[\sum_{i=1}^{N} W_i^2 \bar{f}_{i_P}^2 G_i + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \bar{f}_{i_P}\bar{f}_{k_P}\sqrt{G_i G_k}\,\mathcal{F}_{ik}^{P}\right] + \sum_{i=1}^{N} W_i^2 T_i} \quad (15)$$

IV. An Array of Identical Antennas

The major effects of correlated noise are easily seen by examining an array of identical antennas. Consider an array of N antennas, each with a gain G, a system temperature T, and an average field pattern in the direction of the planet $\bar{f}_P$. Since the antennas are identical, $W_i = 1$ for all

[5] One can imagine cases where the increase in $P_\Sigma^{s/c}$ obtained by phasing the array is more than counteracted by the resulting increase in P_Σ^{P}, but even attempting to array in such cases would be difficult.

i. When this array is properly phased on the spacecraft, Eqs. (12), (13), and (14) become

$$P_{\Sigma\phi}^{s/c} = 2k_B N^2 G \mathcal{P}_{s/c} \tag{16}$$

$$P_{\Sigma\phi}^{P} = \Delta\nu k_B \bar{f}_P^2 G S_P \left[N + \sum_{i=1}^{N} \sum_{k\neq i}^{N} \mathcal{F}_{ik}^{P} \right] \tag{17}$$

$$P_{\Sigma\phi}^{N} = \Delta\nu k_B N T \tag{18}$$

and Eq. (15) becomes

$$\beta_\phi = \frac{NG}{\bar{f}_P^2 G S_P \left[1 + N^{-1} \sum_{i=1}^{N} \sum_{k\neq i}^{N} \mathcal{F}_{ik}^{P} \right] + T} \tag{19}$$

Since $|\mathcal{F}_{ik}^{P}| \leq 1$ for all baselines, it is always true that

$$1 + N^{-1} \sum_{i=1}^{N} \sum_{k\neq i}^{N} \mathcal{F}_{ik}^{P} \leq N \tag{20}$$

From Eq. (A-27) [see also Eqs. (B-1), (B-3), and (B-5)], it can be seen that, in the limit of an extremely compact array [i.e., $B_{ik} << c/(\nu_{sky}^{c} R_P)$, for all i,k; R_P being the characteristic angular radius of the planet], $\mathcal{F}_{ik}^{P} \rightarrow 1$ on all baselines and

$$\beta_{\phi_0} \rightarrow \frac{NG}{NGS_P + T} \tag{21}$$

where the substitution $\bar{f}_P = 1$, appropriate in this limit,[6] has been made. This is equivalent to observing the spacecraft with a single dish of gain NG and an effective system temperature of $NGS_P + T$. Not surprisingly, observing the spacecraft and planet with a compact array is analogous to observing the two objects with an antenna with N times the area of a single array element.

In the limit where $B_{ik} >> c/(\nu_{sky}^{c} R_P)$, for all i,k (i.e., an extremely extended array), $|\mathcal{F}_{ik}^{P}| \rightarrow 0$ on all baselines and

$$\beta_{\phi_\infty} \rightarrow \frac{NG}{\bar{f}_P^2 G S_P + T} \tag{22}$$

This is equivalent to observing the spacecraft with a single dish of gain NG and an effective system temperature of $\bar{f}_P^2 G S_P + T$.

In both cases, the effective gain is NG; it is independent of baseline length and depends only on the gains of the individual antennas. However, the effective system temperature is not the same in the two extreme configurations. The planet contributes $N\bar{f}_P^2 G S_P$ to the system temperature in the compact array limit but only $\bar{f}_P^2 G S_P$ in the extended array limit. In the extended array limit, the noise contribution from the planet is simply the sum of its (uncorrelated) contributions to the individual system temperature. In the compact array limit, its contribution is N times larger due to the correlated noise effects.

Though, as discussed below, the effects of correlated noise are a complicated function of array geometry. In general, *the more extended the array configuration, the smaller the correlated noise contribution* of a planet or other hot body. As a rule of thumb, correlated-noise contributions are significant on baselines where $\nu_{sky}^{c} B_{ik} R_P / c \lesssim 1$ (see Appendix A, Section III).

It should be noted that in the intermediate cases where $B_{ik} \sim C/(\nu_{sky}^{c} R_P)$, the sum over $\mathcal{F}_{ik}^{P}$ may, for some geometries, be negative. In such cases, the performance of the array is actually enhanced over that of the extended array limit.

V. DSN Complexes at S-Band

One of the proposals for support of the Galileo S-band mission is the arraying of antennas at each DSN complex. This section describes the performance of an array at the Canberra complex that includes DSS 42, 43, 45, and the soon-to-be-built DSS 34. Table 1 lists each antenna, its S-band gain and system temperature at zenith, its S-band beamwidth at full-width half-power (FWHP), and its station coordinates (east, north, and vertical) relative to DSS 43.

With the model of Jupiter given in Appendix B, Eq. (15) can be used to calculate β_ϕ for the array. The improvement that the array would provide relative to the stand-alone use of DSS 43, if correlated noise effects could be neglected, is given by the ratio

[6] If the baseline is small compared to $c/(\nu_{sky}^{c} R_P)$, the diameters of the individual antennas would be smaller still.

$$\frac{\beta_{\phi_\infty}}{\beta_{43}} = \frac{\left[\sum_{i=1}^{N} W_i^2 G_i + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \sqrt{G_i G_k}\right]}{S_P \sum_{i=1}^{N} W_i^2 \bar{f}_{i_P}^2 G_i + \sum_{i=1}^{N} W_i^2 T_i}$$

$$\times \frac{T_{43} + \bar{f}_{43_P}^2 G_{43} S_P}{G_{43}} \qquad (23)$$

When correlated noise is neglected, this ratio is independent of hour angle, array geometry, or source structure, but it depends on S_P—the flux of Jupiter—and, therefore, on the Earth–Jupiter distance. At opposition, this distance is 4.2 AU, the S-band flux of Jupiter is approximately 5.8 Jy and, with the parameters given in Table 1, $\beta_{\phi_\infty}/\beta_{43} = 1.41$. At conjunction (an Earth–Jupiter distance of 6.2 AU), $S_P \approx 2.6$ Jy and $\beta_{\phi_\infty}/\beta_{43} = 1.38$. The improvement is not as great at large Earth–Jupiter distances because Jupiter's contribution to the system temperature is less; for comparison, if $S_P = 0$, $\beta_{\phi_\infty}/\beta_{43} = 1.34$. It is useful to note that despite the effects of correlated noise, arraying is most useful when extended background sources are present, particularly if the baselines are long. In the above calculations, it is assumed that $\bar{f}_P = 1$ for all antennas and, as in all the calculations in this section, $W_i = \sqrt{G_i}/T_i$ [see Eq. (D-4)]. Throughout this section, it is also assumed that the gains G_i and system temperatures T_i do not vary with elevation, which leads to a slight overestimate of correlated noise effects at low elevations.

To assess the effects of correlated noise, the SNR provided by the actual array is compared with that provided by the same antennas if arrayed in a configuration with infinitely long baselines, examining the ratio $\beta_\phi/\beta_{\phi_\infty}$. This ratio depends not only on Earth–Jupiter distance, but, in a complicated way, on the array and source geometries and on hour angle. Figure 3 plots this ratio for a number of different geometries. In each plot the declination of Jupiter,[7] the Earth–Jupiter distance, and the angular separation between the spacecraft and the center of Jupiter are held fixed. Each point on the plot then represents the value of $\beta_\phi/\beta_{\phi_\infty}$ for a randomly chosen value of the hour angle, the orientation on the sky of Jupiter's radiation belts, and the orientation of the spacecraft–Jupiter separation. Thus, the density of points on a portion of a plot provides an estimate of the likelihood, as a function of the hour angle, of obtaining a particular value of $\beta_\phi/\beta_{\phi_\infty}$ for the given parameters.

Figures 4(a) and (b) summarize the results of the calculations shown in Fig. 3 plotting, as a function of spacecraft–Jupiter separation, the average (over hour angle and orientation) value of $\beta_\phi/\beta_{\phi_\infty}$, as well as its minimum and maximum values. Thus, each plot in Fig. 3 is reduced to three points in Fig. 4—one on an average curve, one on a maximum curve, and one on a minimum curve. The following conclusions can be drawn from Fig. 4:

(1) When Galileo is within ~75 arcsec of Jupiter, the average effect of correlated noise is a loss of SNR, which may be as large as 5 percent at zero separation.

(2) When Galileo is more than 400 arcsec from Jupiter, the effects of correlated noise are negligible.

(3) For separations in the range of 75 to 400 arcsec, the effects of correlated noise are small *on average*. However, the loss may be significant for certain geometries (as may be the enhancement). In this range of planet–spacecraft separations, careful calculations of correlated noise are necessary in a case where a 5-percent SNR loss would be critical.

The above conclusions refer specifically to an array consisting of the antennas listed in Table 1. Correlated noise effects are likely to be similar for arrays of similar antennas with similar baseline geometries, but quantitative calculations for other arrays have not been carried out. Because correlated noise causes significant SNR variations, and because these variations are highly geometry dependent, detailed calculations of these effects should be done for any situations in which a few tenths of a decibel of SNR are crucial. For most purposes involving Galileo and Jupiter, Eqs. (15), (B-6), and (B-9) should be suitable for such calculations.

[7] A declination of −21 deg, corresponding to that of the Galileo encounter in December 1995, has been used for Fig. 3. The results for other Jupiter declinations are different in detail but qualitatively very similar.

Acknowledgments

Many thanks to George Resch for numerous informative discussions of interferometry, reading the early drafts of this article, and help with Figs. 1 and 2; and to Jim Ulvestad for his careful refereeing of the article.

Table 1. S-band parameters of antennas in proposed Canberra array.

Antenna	Gain, K/Jy	System temperature, K	S-band beamwidth (FWHM), deg	Station coordinates[a]		
				East, m	North, m	Vertical, m
DSS 43	0.95	18.5	0.11	0.0	0.0	0.0
DSS 42	0.21	22.0	0.28	0.0003	194.1921	−13.6414
DSS 45	0.16	38.0	0.23	−325.3907	440.1822	−13.1378
DSS 34[b]	0.16	30.0	0.23	68.8	440.2	0.0

[a] Relative to DSS 43.

[b] Values for DSS 34 are approximate.

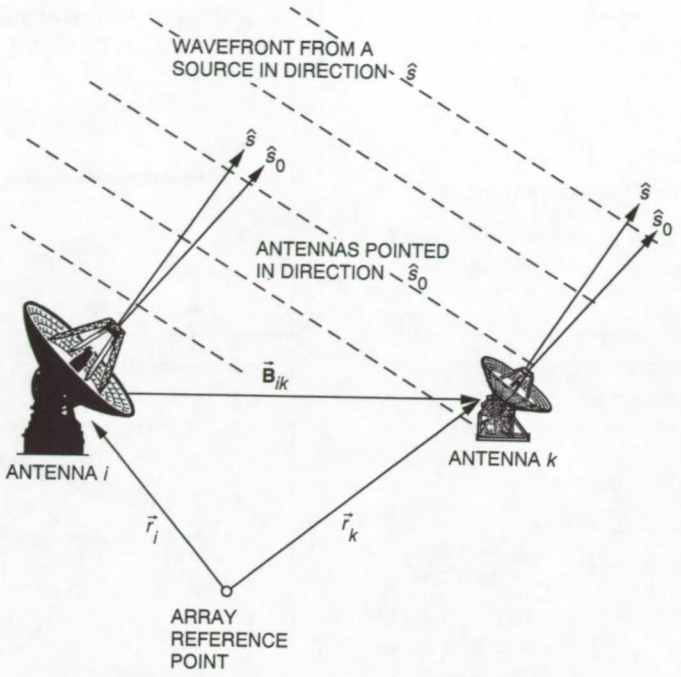


Fig. 1. The basic geometry of an array. Antennas i and k are located at, respectively, r_i and r_k with respect to the array reference point and are pointed toward the direction $\hat{s}_0$. It is clear from the diagram that the delay suffered by a wavefront from a direction $\hat{s}$ between antenna i and antenna k is proportional to $B_{ik} \cdot \hat{s}$.

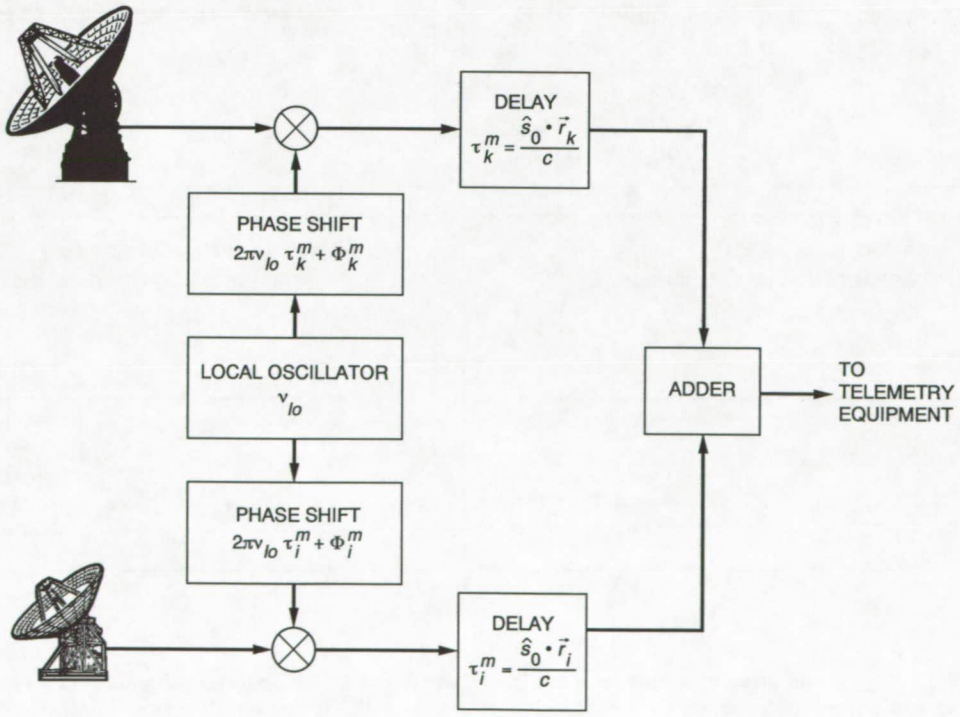


Fig. 2. A simplified block diagram of an array signal path.

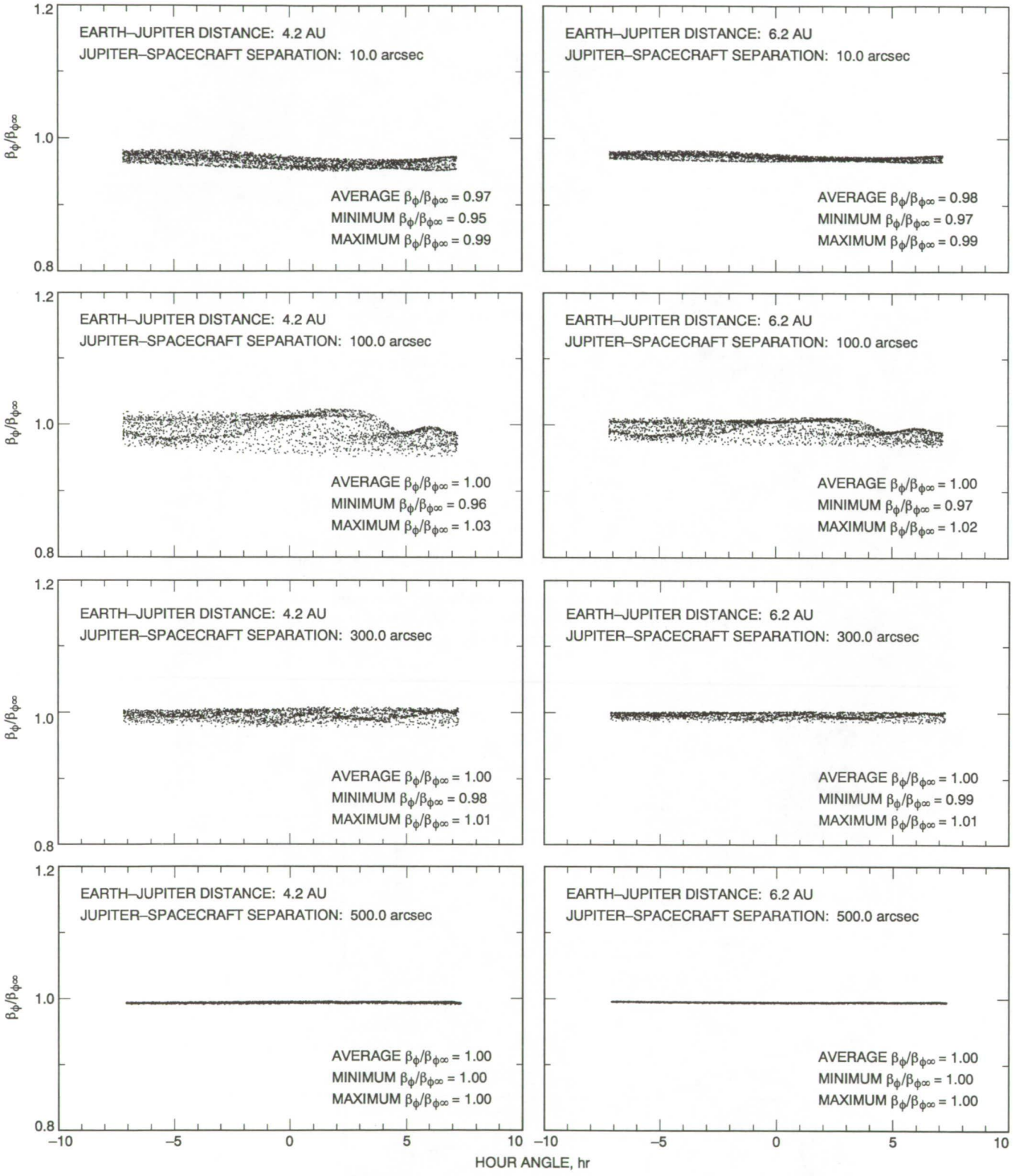


Fig. 3. The ratio of $\beta_\phi / \beta_{\phi\infty}$ for an array of antennas at the DSN Canberra complex observing Jupiter and a spacecraft at S-band, for a variety of array and source geometries. The declination of Jupiter = –21.0 deg and the array consists of DSS 43, DSS 42, DSS 45, and DSS 34.

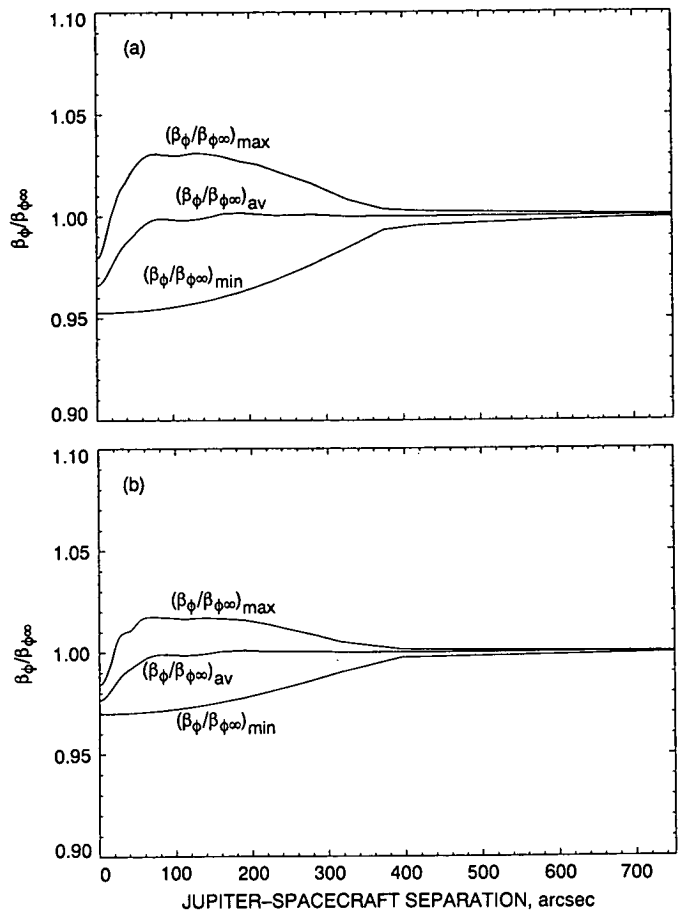


Fig. 4. The average, minimum, and maximum values of $\beta_\phi/\beta_{\phi\infty}$ obtained as source orientation is varied, plotted as a function of the separation between the spacecraft and the center of Jupiter, with Jupiter declination = –21.0 deg: (a) Earth–Jupiter distance = 4.2 AU and (b) Earth–Jupiter distance = 6.2 AU.

Appendix A

Basic Expressions

I. Incoming Signals

The voltage (as a function of time t and sky frequency ν_{sky}) induced in a single polarization channel at $\mathbf{r}_i$, the focal point of the ith antenna, by a distant[8] source (e.g., a planet or spacecraft) can be written

$$V_i(\nu_{sky},t) = e^{i\phi_i}\sqrt{k_B G_i}\iint d\hat{s}\; f_i(\hat{s}-\hat{s}_0)\mathcal{I}(\nu_{sky},t,\hat{s})e^{i(2\pi\nu_{sky}[t-(\hat{s}\cdot\mathbf{r}_i/c)]+\theta(\nu_{sky},\hat{s}))} \tag{A-1}$$

In this expression, k_B represents Boltzmann's constant and c is the speed of light; G_i is the on-axis antenna gain, f_i is the magnitude of the antenna's field pattern, $\hat{s}$ is a unit vector in the direction of the source, $\mathcal{I}(\nu_{sky},t,\hat{s})$ is the amplitude of the electric field due to the source, $\theta(\nu_{sky},\hat{s})$ is a phase term (which for most astronomical sources can be assumed to be uncorrelated over ν_{sky} and $\hat{s}$), and ϕ_i is an antenna-based phase shift (including atmospheric effects).

The voltage induced by system noise can be written

$$V_i^N(\nu_{sky},t) = \sqrt{k_B T_i}\,e^{i\left(2\pi\nu_{sky}t+\theta_i^N(\nu_{sky},t)\right)} \tag{A-2}$$

where T_i is the noise temperature and $\theta_i^N(\nu_{sky},t)$ is a random phase that is uncorrelated over time and frequency intervals satisfying $\Delta\nu_{sky}\Delta t \gtrsim 1$. The total RF voltage can be written as the sum of terms due to the spacecraft, planet, and system noise:

$$V_i(\nu_{sky},t) = V_i^{s/c}(\nu_{sky},t) + V_i^P(\nu_{sky},t) + V_i^N(\nu_{sky},t) \tag{A-3}$$

When the RF signal is mixed to baseband, it is phase shifted by $2\pi\nu_{lo}\tau_i^m$; this stops the fringes, allowing time averaging over a longer interval. An additional phase shift $-\phi_i^m$ can be inserted to (attempt to) compensate for the antenna-based phase shifts ϕ_i. Without this additional phase shift, the baseband voltage can be written

$$\begin{aligned}
v_i(\nu,t) &= e^{i2\pi\nu_{lo}\tau_i^m}e^{-i2\pi\nu_{lo}t}V_i(\nu_{sky},t)\\
&= e^{i2\pi\nu_{lo}[\tau_i^m-t]}\left[V_i^{s/c}(\nu_{sky},t)+V_i^P(\nu_{sky},t)+V_i^N(\nu_{sky},t)\right]\\
&= e^{i(2\pi\nu_{lo}[\tau_i^m-t]+\phi_i)}\sqrt{k_B G_i}\iint d\hat{s}\; f_i(\hat{s}-\hat{s}_0)e^{i2\pi\nu_{sky}[t-(\hat{s}\cdot\mathbf{r}_i/c)]}\Big[\mathcal{I}_{s/c}(\nu_{sky},t,\hat{s})e^{i\theta^{s/c}(\nu_{sky},\hat{s})}\\
&\quad+\mathcal{I}_P(\nu_{sky},t,\hat{s})e^{i\theta^P(\nu_{sky},\hat{s})}\Big]+e^{i2\pi\nu_{lo}[\tau_i^m-t]}\sqrt{k_B T_i}\,e^{i\left(2\pi\nu_{sky}t+\theta_i^N(\nu_{sky},t)\right)}\\
&= e^{i\phi_i}\sqrt{k_B G_i}\iint d\hat{s}\; f_i(\hat{s}-\hat{s}_0)e^{i2\pi(\nu[t-(\hat{s}\cdot\mathbf{r}_i/c)]+\nu_{lo}[\tau_i^m-(\hat{s}\cdot\mathbf{r}_i/c)])}\Big[\mathcal{I}_{s/c}(\nu_{sky},t,\hat{s})e^{i\theta^{s/c}(\nu_{sky},\hat{s})}\\
&\quad+\mathcal{I}_P(\nu_{sky},t,\hat{s})e^{i\theta^P(\nu_{sky},\hat{s})}\Big]+\sqrt{k_B T_i}\,e^{i\left(2\pi[\nu t+\nu_{lo}\tau_i^m]+\theta_i^N(\nu_{sky},t)\right)}
\end{aligned} \tag{A-4}$$

[8] Here, "distant" means a source sufficiently far away that, over the extent of the array, a plane wave approximation is valid.

where $\nu = \nu_{sky} - \nu_{lo}$ is the baseband frequency. Like the RF voltage, the baseband voltage can be written as the sum of terms due to the spacecraft, planet, and system noise:

$$v_i(\nu,t) = v_i^{s/c}(\nu,t) + v_i^P(\nu,t) + v_i^N(\nu,t) \tag{A-5}$$

If the additional phase shift is inserted, the baseband signal has the form $e^{-i\phi_i^m} v_i(\nu,t)$.

II. Single-Dish Power

For the ith antenna, the single-dish power can be written

$$\begin{aligned}
P_i &= \left\langle \int_0^{\Delta\nu} d\nu \, |v_i(\nu,t)|^2 \right\rangle_{\mathcal{T}} \\
&= \left\langle \int_0^{\Delta\nu} d\nu \, |v_i^{s/c}(\nu,t) + v_i^P(\nu,t) + v_i^N(\nu,t)|^2 \right\rangle_{\mathcal{T}} \\
&= \left\langle \int_0^{\Delta\nu} d\nu \left[v_i^{s/c}(\nu,t) v_i^{s/c*}(\nu,t) + v_i^P(\nu,t) v_i^{s/c*}(\nu,t) + v_i^N(\nu,t) v_i^{s/c*}(\nu,t) \right.\right. \\
&\quad + v_i^{s/c}(\nu,t) v_i^{P*}(\nu,t) + v_i^P(\nu,t) v_i^{P*}(\nu,t) + v_i^N(\nu,t) v_i^{P*}(\nu,t) \\
&\quad \left.\left. + v_i^{s/c}(\nu,t) v_i^{N*}(\nu,t) + v_i^P(\nu,t) v_i^{N*}(\nu,t) + v_i^N(\nu,t) v_i^{N*}(\nu,t) \right] \right\rangle_{\mathcal{T}}
\end{aligned} \tag{A-6}$$

where the angle brackets denote a time average and $\mathcal{T}$ is the interval over which the time average is taken. Looking at each term separately and considering the planet term first,

$$\begin{aligned}
v_i^P(\nu,t) v_i^{P*}(\nu,t) &= e^{i\phi_i} \sqrt{k_B G_i} \iint d\hat{s} \; f_i(\hat{s} - \hat{s}_0) e^{i2\pi(\nu[t - (\hat{s}\cdot\mathbf{r}_i/c)] + \nu_{lo}[\tau_i^m - (\hat{s}\cdot\mathbf{r}_i/c)])} \mathcal{I}_P(\nu + \nu_{lo}, t, \hat{s}) e^{i\theta^P(\nu+\nu_{lo},\hat{s})} \\
&\quad \times e^{-i\phi_i} \sqrt{k_B G_i} \iint d\hat{s}' \; f_i(\hat{s}' - \hat{s}_0) e^{-i2\pi(\nu[t - (\hat{s}'\cdot\mathbf{r}_i/c)] + \nu_{lo}[\tau_i^m - (\hat{s}'\cdot\mathbf{r}_i/c)])} \mathcal{I}_P(\nu + \nu_{lo}, t, \hat{s}') e^{-i\theta^P(\nu+\nu_{lo},\hat{s}')} \\
&= k_B G_i \iint d\hat{s} \; f_i(\hat{s} - \hat{s}_0) \mathcal{I}_P(\nu + \nu_{lo}, t, \hat{s}) \iint d\hat{s}' \left\{ e^{-i(2\pi(\nu[\hat{s}-\hat{s}']\cdot\mathbf{r}_i + \nu_{lo}[\hat{s}-\hat{s}']\cdot\mathbf{r}_i)/c)} \right. \\
&\quad \left. \times \; e^{i[\theta^P(\nu+\nu_{lo},\hat{s}) - \theta^P(\nu+\nu_{lo},\hat{s}')]} f_i(\hat{s}' - \hat{s}_0) \mathcal{I}_P(\nu + \nu_{lo}, t, \hat{s}') \right\} \\
&= k_B G_i \iint d\hat{s} \; f_i^2(\hat{s} - \hat{s}_0) \mathcal{I}_P^2(\nu + \nu_{lo}, t, \hat{s}) \\
&= k_B G_i \iint d\hat{s} \; f_i^2(\hat{s} - \hat{s}_0) I_P(\nu + \nu_{lo}, t, \hat{s})
\end{aligned} \tag{A-7}$$

where the integral over $\hat{s}'$ is nonzero only when $\hat{s}' = \hat{s}$ because the radiation from the planet is spatially incoherent, and where the substitution $\mathcal{I}_P^2(\nu+\nu_{lo},t,\hat{s}) = I_P(\nu+\nu_{lo},t,\hat{s})$ was made; I_P is the brightness distribution (flux per unit solid angle) of the planet. Similarly

$$v_i^{s/c}(\nu,t)v_i^{s/c*}(\nu,t) = k_B G_i \iint d\hat{s}\; f_i^2(\hat{s}-\hat{s}_0) I_{s/c}(\nu+\nu_{lo},t,\hat{s}) \tag{A-8}$$

with $I_{s/c}$ being the brightness distribution of the spacecraft (usually assumed to be pointlike), and

$$v_i^N(\nu,t)v_i^{N,*}(\nu,t) = k_B T_i \tag{A-9}$$

Compared to these terms, the cross terms are all small, as can be seen by considering

$$\begin{aligned} v_i^{s/c}(\nu,t)v_i^{P*}(\nu,t) \;=\; & k_B G_i \iint_{s/c} d\hat{s}\; f_i(\hat{s}-\hat{s}_0)\mathcal{I}_{s/c}(\nu+\nu_{lo},t,\hat{s}) \iint_P d\hat{s}' \left\{ e^{-i(2\pi/c)(\nu[\hat{s}-\hat{s}']\cdot\mathbf{r}_i+\nu_{lo}[\hat{s}-\hat{s}']\cdot\mathbf{r}_i)} \right. \\ & \left. \times\, e^{i[\theta^{s/c}(\nu+\nu_{lo},\hat{s})-\theta^P(\nu+\nu_{lo},\hat{s}')]} f_i(\hat{s}'-\hat{s}_0)\mathcal{I}_P(\nu+\nu_{lo},t,\hat{s}') \right\} \\ \approx\; & 0 \end{aligned} \tag{A-10}$$

where the integral over $\hat{s}'$ is small because $\theta^{s/c}$ and θ^P are uncorrelated [unlike in Eqs. (A-7) and (A-8), this integral is, on average, zero, even when $\hat{s}' = \hat{s}$].

One can therefore write

$$\begin{aligned} P_i &= \left\langle \int_0^{\Delta\nu} d\nu\, |v_i(\nu,t)|^2 \right\rangle_{\mathcal{T}} \\ &= \left\langle \int_0^{\Delta\nu} d\nu\, |v_i^{s/c}(\nu,t)|^2 + |v_i^P(\nu,t)|^2 + |v_i^N(\nu,t)|^2 \right\rangle_{\mathcal{T}} \\ &= P_i^{s/c} + P_i^P + P_i^N \end{aligned} \tag{A-11}$$

It is assumed that the spacecraft is a point source located at $\hat{s}_{s/c}$, transmitting a narrowband, polarized signal at a sky frequency $\nu_{sky}^{s/c}$ with a total power $\mathcal{P}_{s/c}$. The brightness distribution seen by a receiver matched to the signal's polarization[9] is

$$I_{s/c}(\nu_{sky},t,\hat{s}) = 2\mathcal{P}_{s/c}\delta\left(\nu_{sky}-\nu_{sky}^{s/c}\right)\delta(\hat{s}-\hat{s}_{s/c}) \tag{A-12}$$

(with δ representing a Dirac delta function), so

[9] The factor of 2 in Eq. (A-12) arises from the assumption of matched polarizations.

$$P_i^{s/c} = 2k_B f_i^2(\hat{\mathbf{s}}_{s/c} - \hat{\mathbf{s}}_0) G_i \mathcal{P}_{s/c} \approx 2k_B G_i \mathcal{P}_{s/c} \tag{A-13}$$

where it is assumed that $f_i^2(\hat{\mathbf{s}}_{s/c} - \hat{\mathbf{s}}_0) \approx 1$ (i.e., the spacecraft is close to center of the array's field of view).

It is assumed that the planet is an extended, unpolarized, broadband source of total flux at the frequencies of interest of S_P, which is constant in time and varies only slowly with frequency (i.e., it can be considered constant across the observing bandwidth $\Delta\nu$), so $I_P(\nu_{sky}, t, \hat{\mathbf{s}}) = I_P(\hat{\mathbf{s}})$ and

$$P_i^P = k_B G_i \Delta\nu \iint_P d\hat{\mathbf{s}}\; f_{i_P}^2(\hat{\mathbf{s}} - \hat{\mathbf{s}}_0) I_P(\hat{\mathbf{s}}) = k_B \bar{f}_{i_P}^2 G_i \Delta\nu S_P \tag{A-14}$$

where $\bar{f}_{i_P}^2$ is a weighted average of $f_{i_P}^2$ over the planet.

Finally, from Eq. (A-9)

$$P_i^N = k_B T_i \Delta\nu \tag{A-15}$$

III. Cross-Correlation

The cross-correlation of the voltages from the ith and kth antennas is given by

$$\rho_{ik} = \left\langle \int_0^{\Delta\nu} d\nu\, e^{i(\phi_k^m - \phi_i^m)} v_i(\nu, t + \tau_i^m) v_k^*(\nu, t + \tau_k^m) \right\rangle_{\mathcal{T}} \tag{A-16}$$

By using the general form of the baseband signal [Eq. (A-4)] and because the signal is spatially incoherent, this becomes

$$\begin{aligned}
\rho_{ik} &= \left\langle \int_0^{\Delta\nu} d\nu\, e^{i(\phi_k^m - \phi_i^m)} \left[v_i^{s/c}(\nu, t + \tau_i^m) + v_i^P(\nu, t + \tau_i^m) + v_i^N(\nu, t + \tau_i^m) \right] \right. \\
&\quad \left. \times \left[v_k^{*s/c}(\nu, t + \tau_k^m) + v_k^{*p}(\nu, t + \tau_k^m) + v_k^{*N}(\nu, t + \tau_k^m) \right] \right\rangle_{\mathcal{T}} \\
&= \left\langle \int_0^{\Delta\nu} d\nu\, e^{i(\phi_k^m - \phi_i^m)} \left[v_i^{s/c}(\nu, t + \tau_i^m) v_k^{*s/c}(\nu, t + \tau_k^m) + v_i^P(\nu, t + \tau_i^m) v_k^{*s/c}(\nu, t + \tau_k^m) \right. \right. \\
&\quad + v_i^N(\nu, t + \tau_i^m) v_k^{*s/c}(\nu, t + \tau_k^m) + v_i^{s/c}(\nu, t + \tau_i^m) v_k^{*P}(\nu, t + \tau_k^m) \\
&\quad + v_i^P(\nu, t + \tau_i^m) v_k^{*P}(\nu, t + \tau_k^m) + v_i^N(\nu, t + \tau_i^m) v_k^{*P}(\nu, t + \tau_k^m) \\
&\quad + v_i^{s/c}(\nu, t + \tau_i^m) v_k^{*N}(\nu, t + \tau_k^m) + v_i^P(\nu, t + \tau_i^m) v_k^{*N}(\nu, t + \tau_k^m) \\
&\quad \left. \left. + v_i^N(\nu, t + \tau_i^m) v_k^{*N}(\nu, t + \tau_k^m) \right] \right\rangle_{\mathcal{T}}
\end{aligned} \tag{A-17}$$

This can be written as

$$\rho_{ik} = \rho_{ik}^{s/c} + \rho_{ik}^{P} + \rho_{ik}^{N} \tag{A-18}$$

where

$$\rho_{ik}^{s/c} = \left\langle \int_0^{\Delta\nu} d\nu\, e^{i(\phi_k^m - \phi_i^m)} v_i^{s/c}(\nu, t+\tau_i^m) v_k^{*s/c}(\nu, t+\tau_k^m) \right\rangle_{\mathcal{T}} \tag{A-19}$$

$$\rho_{ik}^{P} = \left\langle \int_0^{\Delta\nu} d\nu\, e^{i(\phi_k^m - \phi_i^m)} v_i^{P}(\nu, t+\tau_i^m) v_{kP}^{*}(\nu, t+\tau_k^m) \right\rangle_{\mathcal{T}} \tag{A-20}$$

$$\begin{aligned}
\rho_{ik}^{N} = \Bigg\langle \int_0^{\Delta\nu} d\nu\, e^{i(\phi_k^m - \phi_i^m)} \Big[& v_i^N(\nu, t+\tau_i^m) v_k^{*N}(\nu, t+\tau_k^m) + v_i^P(\nu, t+\tau_i^m) v_k^{*s/c}(\nu, t+\tau_k^m) \\
& + v_i^N(\nu, t+\tau_i^m) v_k^{*s/c}(\nu, t+\tau_k^m) + v_i^{s/c}(\nu, t+\tau_i^m) v_k^{*P}(\nu, t+\tau_k^m) \\
& + v_i^N(\nu, t+\tau_i^m) v_k^{*P}(\nu, t+\tau_k^m) + v_i^{s/c}(\nu, t+\tau_i^m) v_k^{*N}(\nu, t+\tau_k^m) \\
& + v_i^P(\nu, t+\tau_i^m) v_k^{*N}(\nu, t+\tau_k^m) \Big] \Bigg\rangle_{\mathcal{T}}
\end{aligned} \tag{A-21}$$

These expressions are similar in structure to the expression for single-dish power [Eq. (A-6)]. Looking at ρ_{ik}^{P}, one finds

$$\begin{aligned}
v_i^P(\nu, t+\tau_i^m) v_k^{*P}(\nu, t+\tau_k^m) &= e^{i(\phi_i - \phi_k)} k_B \sqrt{G_i G_k}\, e^{i2\pi[\nu+\nu_{lo}][\tau_i^m - \tau_k^m]} \Bigg[\iint d\hat{s} \Big\{ e^{-i(2\pi[\nu+\nu_{lo}]\hat{s}\cdot \mathbf{r}_i/c)} f_i(\hat{s}-\hat{s}_0) \mathcal{I}_P(\hat{s}) \\
&\quad \times \iint d\hat{s}'\; f_k(\hat{s}'-\hat{s}_0) \mathcal{I}_P(\hat{s}') e^{i\left(2\pi[\nu+\nu_{lo}]\hat{s}'\cdot \mathbf{r}_k/c\right)} e^{i\left(\theta^P(\nu+\nu_{lo},\hat{s}) - \theta^P(\nu+\nu_{lo},\hat{s}')\right)} \Big\} \Bigg] \\
&= e^{i(\phi_i - \phi_k)} k_B \sqrt{G_i G_k}\, e^{i(2\pi[\nu+\nu_{lo}][\hat{s}_0\cdot \mathbf{r}_i - \hat{s}_0\cdot \mathbf{r}_k]/c)} \\
&\quad \times \iint d\hat{s}\; e^{i(2\pi[\nu+\nu_{lo}]\hat{s}\cdot[\mathbf{r}_k - \mathbf{r}_i]/c)} f_i(\hat{s}-\hat{s}_0) f_k(\hat{s}-\hat{s}_0) I_P(\hat{s}) \\
&= e^{i(\phi_i - \phi_k)} k_B \sqrt{G_i G_k} \iint d\hat{s}\; e^{i(2\pi[\nu+\nu_{lo}][\hat{s}-\hat{s}_0]\cdot[\mathbf{r}_k - \mathbf{r}_i]/c)} f_i(\hat{s}-\hat{s}_0) f_k(\hat{s}-\hat{s}_0) I_P(\hat{s}) \\
&= e^{i(\phi_i - \phi_k)} k_B \sqrt{G_i G_k} \iint d\hat{s}\; e^{i(2\pi[\nu+\nu_{lo}][\hat{s}-\hat{s}_0]\cdot \mathbf{B}_{ik}/c)} f_i(\hat{s}-\hat{s}_0) f_k(\hat{s}-\hat{s}_0) I_P(\hat{s})
\end{aligned} \tag{A-22}$$

where the substitutions $\tau_i^m = \hat{\mathbf{s}}_0 \cdot \mathbf{r}_i/c$, $\tau_k^m = \hat{\mathbf{s}}_0 \cdot \mathbf{r}_k/c$ and $\mathbf{B}_{ik} = \mathbf{r}_k - \mathbf{r}_i$ have been made.

Similarly

$$v_i^{s/c}(\nu, t+\tau_i^m) v_k^{*s/c}(\nu, t+\tau_k^m) = e^{i(\phi_i - \phi_k)} k_B \sqrt{G_i G_k} \times \iint d\hat{\mathbf{s}}\, e^{i(2\pi[\nu+\nu_{lo}][\hat{\mathbf{s}}-\hat{\mathbf{s}}_0]\cdot \mathbf{B}_{ik}/c)} f_i(\hat{\mathbf{s}}-\hat{\mathbf{s}}_0) f_k(\hat{\mathbf{s}}-\hat{\mathbf{s}}_0) I_{s/c}(\hat{\mathbf{s}}) \tag{A-23}$$

In considering the effects of noise, it is assumed[10] that Eq. (A-21) can be approximated as

$$\begin{aligned}
\rho_{ik}^N &\approx \left\langle \int_0^{\Delta\nu} d\nu\, e^{i(\phi_k^m - \phi_i^m)} v_i^N(\nu, t+\tau_i^m)\, v_k^{*N}(\nu, t+\tau_k^m) \right\rangle_{\mathcal{T}} \\
&\approx k_B \sqrt{T_i T_k} \left\langle \int_0^{\Delta\nu} d\nu\, e^{i(\phi_k^m - \phi_i^m)} e^{i(2\pi\nu[t+\tau_i^m] + 2\pi\nu_{lo}\tau_i^m + \theta_i^N(\nu_{sky}, t+\tau_i^m))} e^{-i(2\pi\nu[t+\tau_k^m] + 2\pi\nu_{lo}\tau_k^m + \theta_k^N(\nu_{sky}, t+\tau_k^m))} \right\rangle_{\mathcal{T}} \\
&\approx k_B \sqrt{T_i T_k} \left\langle \int_0^{\Delta\nu} d\nu\, e^{i(2\pi\nu[\tau_i^m - \tau_k^m] + 2\pi\nu_{lo}[\tau_i^m - \tau_k^m] + \phi_k^m - \phi_i^m + \theta_i^N(\nu_{sky}, t+\tau_i^m) - \theta_k^N(\nu_{sky}, t+\tau_k^m))} \right\rangle_{\mathcal{T}} \\
&\approx k_B \sqrt{T_i T_k} \frac{\sqrt{\Delta\nu \mathcal{T}}}{\mathcal{T}} e^{i\theta_{ik}^N} \\
&\approx k_B \sqrt{T_i T_k} \sqrt{\frac{\Delta\nu}{\mathcal{T}}} e^{i\theta_{ik}^N}
\end{aligned} \tag{A-24}$$

where θ_{ik}^N is a phase randomly distributed between 0 and 2π, and where use has been made of the fact that $\theta_i^N(\nu_{sky}, t)$ is uncorrelated between antennas and over time and frequency intervals where $\Delta\nu\mathcal{T} > 1$. Because the phases of ρ_{ik}^N are random, the double sum in Eq. (10) will be small. However, as discussed in Appendix C, noise terms become important when the process of phasing the array is considered.

Using Eq. (A-12) to substitute for $I_{s/c}$,

$$\begin{aligned}
\rho_{ik}^{s/c} &= 2\mathcal{P}_{s/c} \left\langle \int_0^{\Delta\nu} d\nu\, \delta(\nu - [\nu_{sky}^{s/c} - \nu_{lo}]) e^{i(\phi_k^m - \phi_i^m)} e^{i(\phi_i - \phi_k)} k_B \sqrt{G_i G_k} \right. \\
&\qquad \left. \times \iint d\hat{\mathbf{s}}\, e^{i(2\pi[\nu+\nu_{lo}][\hat{\mathbf{s}}-\hat{\mathbf{s}}_0]\cdot \mathbf{B}_{ik}]/c)} f_i(\hat{\mathbf{s}}-\hat{\mathbf{s}}_0) f_k(\hat{\mathbf{s}}-\hat{\mathbf{s}}_0) \delta(\hat{\mathbf{s}}-\hat{\mathbf{s}}_{s/c}) \right\rangle_{\mathcal{T}} \\
&= 2\mathcal{P}_{s/c} k_B \sqrt{G_i G_k} e^{i(\delta\phi_i - \delta\phi_k)} e^{i(2\pi\nu_{sky}^{s/c}[\hat{\mathbf{s}}_{s/c}-\hat{\mathbf{s}}_0]\cdot \mathbf{B}_{ik}/c)} f_i(\hat{\mathbf{s}}_{s/c}-\hat{\mathbf{s}}_0) f_k(\hat{\mathbf{s}}_{s/c}-\hat{\mathbf{s}}_0)
\end{aligned} \tag{A-25}$$

[10] This assumption is not valid if the planet or the spacecraft contributes significantly to the system temperature; in that case, the cross terms in Eq. (A-21) will be non-negligible. However, they can be treated in a manner similar to the $v^N v^{*N}$ term.

Note that $\hat{\mathbf{s}}_{s/c} - \hat{\mathbf{s}}_0$ must be small enough that the quantity $\nu_{sky}^{s/c}[\hat{\mathbf{s}}_{s/c} - \hat{\mathbf{s}}_0]\cdot\mathbf{B}_{ik}/c$ does not change significantly over the interval of the time averaging.

Similarly,

$$\rho_{ik}^P = k_B\sqrt{G_iG_k}\,e^{i(\delta\phi_i - \delta\phi_k)}\left\langle \iint d\hat{\mathbf{s}}\; f_i(\hat{\mathbf{s}} - \hat{\mathbf{s}}_0)f_k(\hat{\mathbf{s}} - \hat{\mathbf{s}}_0)I_P(\hat{\mathbf{s}}) \int_0^{\Delta\nu} d\nu\, e^{i(2\pi[\nu+\nu_{lo}][\hat{\mathbf{s}}-\hat{\mathbf{s}}_0]\cdot\mathbf{B}_{ik}/c)} \right\rangle_{\mathcal{T}} \tag{A-26}$$

At this point, it is useful to define the quantity

$$\mathcal{F}_{ik}^P \equiv \frac{1}{\Delta\nu S_P}\iint d\hat{\mathbf{s}}\; e^{i(2\pi\nu_{lo}\mathbf{B}_{ik}\cdot(\hat{\mathbf{s}}-\hat{\mathbf{s}}_0)/c)} f_i(\hat{\mathbf{s}} - \hat{\mathbf{s}}_0)f_k(\hat{\mathbf{s}} - \hat{\mathbf{s}}_0)I_P(\hat{\mathbf{s}}) \int_0^{\Delta\nu} d\nu\, e^{i(2\pi\nu\mathbf{B}_{ik}\cdot(\hat{\mathbf{s}}-\hat{\mathbf{s}}_0)/c)} \tag{A-27}$$

so that

$$\rho_{ik}^P = \Delta\nu k_B S_P\sqrt{G_iG_k}\,e^{i(\delta\phi_i - \delta\phi_k)}\mathcal{F}_{ik}^P \tag{A-28}$$

$\mathcal{F}_{ik}^P$ is a dimensionless complex quantity that depends only on the array geometry and the source structure and whose magnitude is always less than or equal to unity ($|\mathcal{F}_{ik}^P| \to 1$ in the short baseline limit, and $|\mathcal{F}_{ik}^P| \to 0$ in the long baseline limit).

It has been implicitly assumed that the bandpass is rectangular, in which case the integral over ν can be simplified:

$$\int_0^{\Delta\nu} d\nu\, e^{i(2\pi\nu\mathbf{B}_{ik}\cdot(\hat{\mathbf{s}}-\hat{\mathbf{s}}_0)/c)} = e^{i(\pi\Delta\nu\mathbf{B}_{ik}\cdot[\hat{\mathbf{s}}-\hat{\mathbf{s}}_0]/c)}\mathrm{sinc}\left(\frac{\Delta\nu}{c}\mathbf{B}_{ik}\cdot[\hat{\mathbf{s}} - \hat{\mathbf{s}}_0]\right) \tag{A-29}$$

where $\mathrm{sinc}(x) = \sin(\pi x)/\pi x$. This term, often referred to as the delay beam, introduces a phase shift and lowers the correlation amplitude for $\hat{\mathbf{s}} \neq \hat{\mathbf{s}}_0$; both these effects are more pronounced for larger bandwidths $\Delta\nu$.

One can therefore write

$$\mathcal{F}_{ik}^P = \frac{1}{\Delta\nu S_P}\iint d\hat{\mathbf{s}}\; e^{i(2\pi\nu_{sky}^c\mathbf{B}_{ik}\cdot(\hat{\mathbf{s}}-\hat{\mathbf{s}}_0)/c)} f_i(\hat{\mathbf{s}} - \hat{\mathbf{s}}_0)f_k(\hat{\mathbf{s}} - \hat{\mathbf{s}}_0)I_P(\hat{\mathbf{s}})\,\mathrm{sinc}\left(\frac{\Delta\nu}{c}\mathbf{B}_{ik}\cdot[\hat{\mathbf{s}} - \hat{\mathbf{s}}_0]\right) \tag{A-30}$$

where $\nu_{sky}^c = \nu_{lo} + \Delta\nu/2$.

If the quantities f_i, f_k and $\mathrm{sinc}(\Delta\nu\mathbf{B}_{ik}\cdot[\hat{\mathbf{s}} - \hat{\mathbf{s}}_0]/c)$ do not vary greatly over the extent of the planet, one can make the further simplification

$$\mathcal{F}_{ik}^P = \frac{1}{\Delta\nu S_P}\bar{f}_{i_P}\bar{f}_{k_P}\bar{f}_{\Delta\nu\,ik}\iint d\hat{\mathbf{s}}\; e^{i(2\pi\nu_{sky}^c\mathbf{B}_{ik}\cdot(\hat{\mathbf{s}}-\hat{\mathbf{s}}_0)/c)} I_P(\hat{\mathbf{s}}) \tag{A-31}$$

where $\bar{f}_{i_P}$, $\bar{f}_{k_P}$, $\bar{f}_{\Delta\nu\,ik}$ are suitably weighted averages of f_i, f_k and $\mathrm{sinc}(\Delta\nu\mathbf{B}_{ik}\cdot[\hat{\mathbf{s}} - \hat{\mathbf{s}}_0]/c)$, respectively.

Appendix B

Jupiter Model

At centimeter wavelengths, Jupiter is not a simple thermal disk; there is significant synchrotron emission from the radiation belts and the resulting flux distribution is quite complicated [6,7]. For the purposes of this article, Jupiter can be modelled as the sum of three components: two circular (2-dimensional) Gaussian components (representing the radiation belts) and a uniform central disk. This is a simple model to work with because Eq. (A-27) can be integrated analytically for each of these components.

If the brightness distribution of a source is radially symmetric about the point $\hat{\mathbf{s}}_C$, small compared to the delay beam, and small compared to the primary beam of any antenna, Eq. (A-27) can be written

$$\mathcal{F}_{ik} = \frac{2\bar{f}_i\bar{f}_k}{S}\bar{f}_{\Delta\nu_{ik}}e^{i(2\pi\nu^c_{sky}\mathbf{B}_{ik}\cdot[\hat{\mathbf{s}}_C-\hat{\mathbf{s}}_0]/c)}\int_0^\infty r\,dr\,I(r)J_0\left(\frac{2\pi\nu^c_{sky}B_{ik}r}{c}\right) \tag{B-1}$$

where $r = |\hat{\mathbf{s}} - \hat{\mathbf{s}}_C|$; $\bar{f}_{\Delta\nu_{ik}}$ is the value of the delay beam at the source; $\bar{f}_i$, $\bar{f}_k$ are the average values of the antenna; field pattern at the source; for the ith and kth antenna; S is the total flux of the source; $I(r)$ is the source's radially symmetric brightness distribution; and J_0 is a zeroth-order Bessel function.

For a uniform disk of angular radius R_D centered at $\hat{\mathbf{s}}_C$,

$$\begin{aligned} I(r) &= \frac{S}{\pi R_D^2} \qquad r \le R_D \\ &= 0 \qquad r > R_D \end{aligned} \tag{B-2}$$

$$\mathcal{F}^D_{ik} = \bar{f}_i\bar{f}_k\bar{f}_{\Delta\nu_{ik}}e^{i(2\pi\nu^c_{sky}/c)\mathbf{B}_{ik}\cdot(\hat{\mathbf{s}}_D-\hat{\mathbf{s}}_0)}\frac{cJ_1(2\pi\nu^c_{sky}B_{ik}R_D/c)}{\pi\nu^c_{sky}B_{ik}R_D} \tag{B-3}$$

For a circular Gaussian of total flux S and characteristic size (1/e radius) R_G, centered at $\hat{\mathbf{s}}_C$,

$$I(r) = \frac{S}{\pi R_G^2}e^{-r^2/R_G^2} \tag{B-4}$$

and Eq. (B-1) becomes

$$\mathcal{F}^G_{ik} = \bar{f}_i\bar{f}_k\bar{f}_{\Delta\nu_{ik}}e^{i(2\pi\nu^c_{sky}\mathbf{B}_{ik}\cdot[\hat{\mathbf{s}}_C-\hat{\mathbf{s}}_0]/c)}\,\frac{e^{-(\pi\nu^c_{sky}B_{ik}R_G/c)^2}}{\pi} \tag{B-5}$$

Jupiter's brightness distribution is modelled as the sum of a uniform central disk of radius R_J, centered at $\hat{\mathbf{s}}_J$, and two circular Gaussians of characteristic size R_B, representing the radiation belts, which are offset from the center of the disk by $\pm\Delta\mathbf{s}_B$. The flux of the central disk is written as $F_D S_J$, where S_J is the total flux of Jupiter and F_D is the fraction of that flux in the disk; the integrated flux of each wing component can be written as $(1 - F_D)S_J/2$. Using this model with Eqs. (B-3) and (B-5),

$$\mathcal{F}_{ik}^{J} = \bar{f}_i^J \bar{f}_k^J \bar{f}_{\Delta\nu_{ik}}^J e^{i(2\pi\nu_{sky}^c \mathbf{B}_{ik}\cdot[\hat{\mathbf{s}}_J - \hat{\mathbf{s}}_0]/c)} \left[F_D \frac{cJ_1(2\pi\nu_{sky}^c B_{ik} R_J/c)}{\pi\nu_{sky}^c B_{ik} R_J} \right.$$

$$\left. + (1 - F_D) \frac{e^{-(\pi\nu_{sky}^c B_{ik} R_B/c)^2}}{\pi} \cos\left(\frac{2\pi\nu_{sky}^c}{c} \mathbf{B}_{ik}\cdot\Delta\mathbf{s}_B\right) \right] \tag{B-6}$$

If the above expression and Eq. (C-1), the condition that the array is properly phased on the spacecraft, are substituted into Eq. (A-28), one gets, for the cross-correlation due to Jupiter in a phased array,

$$\rho_{ik}^J = \Delta\nu k_B S_J \sqrt{G_i G_k}\, \bar{f}_i^J \bar{f}_k^J \bar{f}_{\Delta\nu_{ik}}^J e^{i(2\pi \mathbf{B}_{ik}\cdot(\nu_{sky}^c[\hat{\mathbf{s}}_J - \hat{\mathbf{s}}_0] - \nu_{sky}^{s/c}[\hat{\mathbf{s}}_{s/c} - \hat{\mathbf{s}}_0])/c)}$$

$$\times \left[F_D \frac{cJ_1(2\pi\nu_{sky}^c B_{ik} R_J/c)}{\pi\nu_{sky}^c B_{ik} R_J} + (1 - F_D) \frac{e^{-(\pi\nu_{sky}^c B_{ik} R_B/c)^2}}{\pi} \cos\left(\frac{2\pi\nu_{sky}^c}{c} \mathbf{B}_{ik}\cdot\Delta\mathbf{s}_B\right) \right] \tag{B-7}$$

In the case where $\hat{\mathbf{s}}_{s/c} = \hat{\mathbf{s}}_0$ and/or $\nu_{sky}^{s/c} = \nu_{sky}^c$, this simplifies to

$$\rho_{ik}^J = \Delta\nu k_B S_J \sqrt{G_i G_k}\, \bar{f}_i^J \bar{f}_k^J \bar{f}_{\Delta\nu_{ik}}^J e^{i(2\pi\nu_{sky}^c \mathbf{B}_{ik}\cdot[\hat{\mathbf{s}}_J - \hat{\mathbf{s}}_{s/c}]/c)}$$

$$\times \left[F_D \frac{cJ_1(2\pi\nu_{sky}^c B_{ik} R_J/c)}{\pi\nu_{sky}^c B_{ik} R_J} + (1 - F_D) \frac{e^{-(\pi\nu_{sky}^c B_{ik} R_B/c)^2}}{\pi} \cos\left(\frac{2\pi\nu_{sky}^c}{c} \mathbf{B}_{ik}\cdot\Delta\mathbf{s}_B\right) \right] \tag{B-8}$$

Both the total flux S_J and the flux distribution are functions of frequency and Earth–Jupiter distance. From the data given in [5], the model at S-band (2.3 GHz) used in this article is

$$S_J = 6.3 \left(\frac{4.04\,\text{AU}}{d}\right)^2 \text{Jy}$$

$$F_D = 0.3$$

$$R_J = 24.3\, \frac{4.04\,\text{AU}}{d} \text{ arcsec}$$

$$|\Delta\mathbf{s}_B| \approx 2R_J$$

$$R_B \approx 1.3 R_J \tag{B-9}$$

At conjunction, d = 6.2 AU, S_J = 2.7 Jy, and R_J = 15.8 arcsec; at opposition, d = 4.2 AU, S_J = 5.8 Jy, and R_J = 23.4 arcsec.

Figure B-1(a) shows a map of Jupiter from [6] at 2.9 GHz (10.4 cm), while Fig. B-1(b) shows a higher resolution map from [7] at 1.4 GHz. Note that in Fig. B-1(b), the disk has not been removed. Figure (B-2) shows the model of Jupiter used in this article in the same format as Fig. B-1(a).

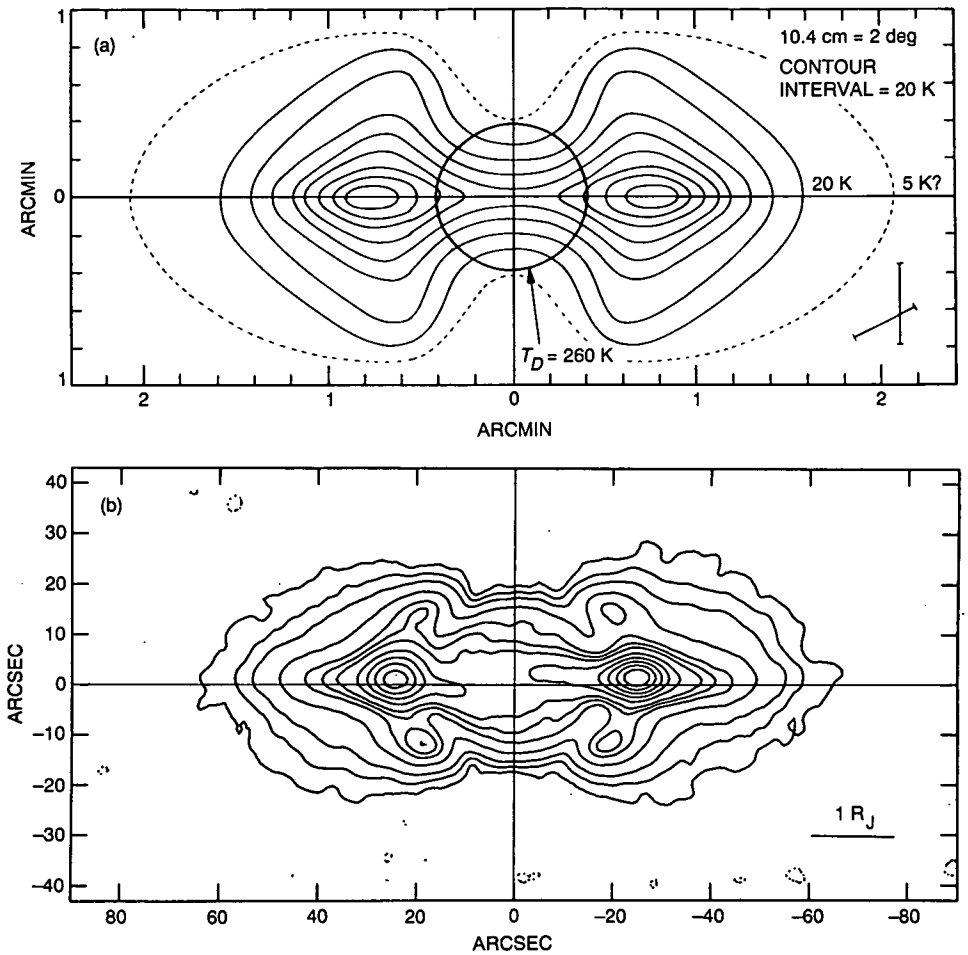


Fig. B-1. A map of the brightness distribution of Jupiter: (a) at 2.9 GHz with a 260-K disk component removed and 20-K contour intervals (from [6]) and (b) at 1.4 GHz, including disk component with contour intervals at 2, 5, 10, 20, 30, 35, 40, 50, 60, 70, 80, and 90 percent (from [7]).

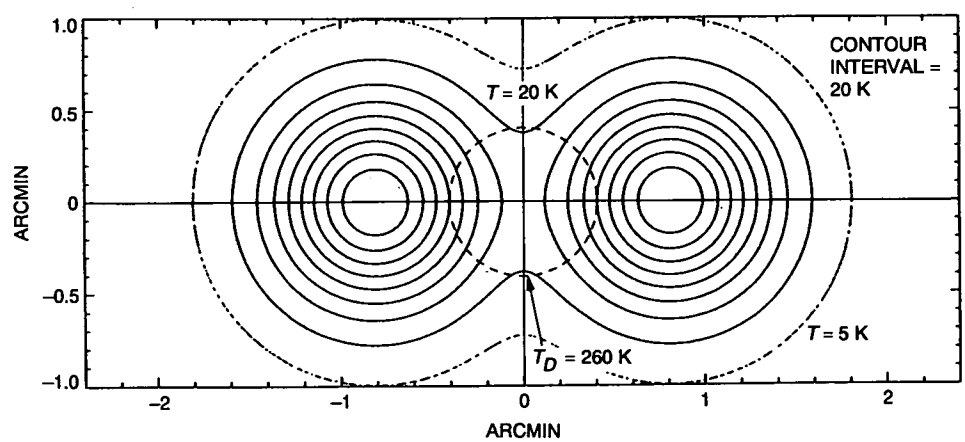


Fig. B-2. A model of Jupiter's brightness distribution at S-band [see Eq. (B-9)], plotted in the same format as Fig. B-1(a) with disk component removed.

Appendix C

Phasing

I. The Ideal Case

The SNR on the spacecraft is maximized when the terms in the double sum in Eq. (8) add in phase. This can be (approximately) arranged with the appropriate choice of model phases ϕ_i^m. By using Eq. (A-25), $\rho_{ik}^{s/c}$ will have zero phase for all i,k when

$$\delta\phi_k - \delta\phi_i = \phi_k - \phi_k^m - \phi_i + \phi_i^m = \frac{2\pi\nu_{sky}^{s/c}}{c}\mathbf{B}_{ik}\cdot(\hat{\mathbf{s}}_{s/c} - \hat{\mathbf{s}}_0) \tag{C-1}$$

The above relation can be satisfied if, for all $i = 1,...N$,

$$\phi_i^m = \phi_i + \frac{2\pi\nu_{sky}^{s/c}}{c}\mathbf{B}_{ir}\cdot(\hat{\mathbf{s}}_{s/c} - \hat{\mathbf{s}}_0) \tag{C-2}$$

where r represents an arbitrarily chosen reference antenna.

The values of ϕ_i are not known a priori, but there are $N(N-1)/2$ independent measurements of the phase of the correlation function that can be used to fit for the N model phases ϕ_i^m (see [8]). Note that since $\delta\phi_i$, $\delta\phi_k$ include media effects as well as phase shifts in each antenna's signal path, they vary on timescales of seconds to minutes, so that the model phases must be recalculated at least that frequently to prevent degradation of the signal.

II. The Effects of Random Noise and a Background Planet

The above discussion ignores the contributions of random noise and a background planet to the observed phases. These complicate the process of solving for model phases, since what actually can be observed is the phase of the total correlation ρ_{ik}, not the phase of $\rho_{ik}^{s/c}$, the correlation due to the spacecraft signal.

The phase Φ_{ik}^{ρ} of the correlation function is given by

$$\begin{aligned}\tan\Phi_{ik}^{\rho} &= \frac{\mathrm{Im}(\rho_{ik})}{\mathrm{Re}(\rho_{ik})} \\ &= \frac{\mathrm{Im}(\rho_{ik}^{s/c}) + \mathrm{Im}(\rho_{ik}^{P}) + \mathrm{Im}(\rho_{ik}^{N})}{\mathrm{Re}(\rho_{ik}^{s/c}) + \mathrm{Re}(\rho_{ik}^{P}) + \mathrm{Re}(\rho_{ik}^{N})}\end{aligned} \tag{C-3}$$

where Re and Im represent, respectively, the real and imaginary parts of the complex function. The simplest phasing algorithms make the approximation that the spacecraft signal is the dominant contribution to ρ_{ik}, i.e.,

$$\tan\Phi_{ik}^{\rho} = \frac{\mathrm{Im}(\rho_{ik})}{\mathrm{Re}(\rho_{ik})} \approx \frac{\mathrm{Im}(\rho_{ik}^{s/c})}{\mathrm{Re}(\rho_{ik}^{s/c})} \tag{C-4}$$

This approximation requires that $|\rho_{ik}^{s/c}| >> |\rho_{ik}^{P}|$ and $|\rho_{ik}^{s/c}| >> |\rho_{ik}^{N}|$ for all baselines $\mathbf{B}_{ik}$. Using Eqs. (A-24), (A-25), and (A-28), this implies

$$\mathcal{P}_{s/c} >> \frac{\Delta\nu S_P}{2}|\mathcal{F}_{ik}^{P}| \tag{C-5}$$

and

$$\mathcal{P}_{s/c} >> \frac{1}{2}\sqrt{\frac{T_i}{G_i}\frac{T_k}{G_k}}\sqrt{\frac{\Delta\nu}{\mathcal{T}}} \tag{C-6}$$

In theory, the first of these conditions can be circumvented by modelling the contribution of ρ_{ik}^{P} to the phase of ρ_{ik} in the phasing algorithm. The second condition is more fundamental and can only be dealt with by decreasing the observing bandwidth (if the spacecraft is a narrowband source) or increasing the time over which the correlation is averaged.

Appendix D

Weighting Factors

The weighting factors W_i should be chosen so as to maximize β_ϕ, which is given by Eq. (15). Such W_i will satisfy the condition

$$\frac{d\beta_\phi}{dW_j} = 0 \tag{D-1}$$

for all $j = 1, ...N$. Since one of the weighting factors can be chosen arbitrarily, this condition amounts to $N - 1$ equations for $N - 1$ variables.

If the effects of the planet can be ignored (i.e., $S_p \approx 0$), Eq. (15) reduces to

$$\beta_\phi = \frac{\left[\sum_{i=1}^{N} W_i^2 G_i + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \sqrt{G_i G_k}\right]}{\sum_{i=1}^{N} W_i^2 T_i} \tag{D-2}$$

and Eq. (D-1) becomes

$$\frac{d\beta_\phi}{dW_j} = 0 = \frac{\left[\sum_{i=1}^{N} W_i^2 T_i\right]\left[2W_j G_j + \sum_{k\neq j}^{N} W_k \sqrt{G_j G_k}\right] - 2W_j T_j \left[\sum_{i=1}^{N} W_i^2 G_i + \sum_{i=1}^{N}\sum_{k\neq i}^{N} W_i W_k \sqrt{G_i G_k}\right]}{\left[\sum_{i=1}^{N} W_i^2 T_i\right]^2} \tag{D-3}$$

It is not terribly difficult to show that this is satisfied for

$$W_i = \frac{\sqrt{G_i}}{T_i} \tag{D-4}$$

If the contribution of the planet to the system temperature at each antenna is accounted for, but the correlated noise terms are ignored, the optimal weighting is

$$W_i = \frac{\sqrt{G_i}}{T_i + G_i S_P} \tag{D-5}$$

This is appropriate for an extremely extended array. If correlated noise terms are included, it is difficult (perhaps impossible) to solve for W_i in closed form, but it can be done numerically. For the array configuration at the DSN complex in Australia, the weights obtained, including the effects of correlated noise, differ by only a few percent from those given by Eq. (D-5), and they have a negligible effect ($\lesssim$0.5 percent) on the values of β_ϕ. In fact, though the values of W_i obtained, including the effects of correlated noise, often differ by $\sim$10 percent from those obtained from Eq. (D-4), the resulting values of β_ϕ differ by $\lesssim$1 percent. Therefore, throughout this article, Eq. (D-4) is used to calculate weighting factors.

References

[1] R. Stevens, "Applications of Telemetry Arraying in the DSN," *TDA Progress Report 42-72*, vol. October–December 1982, Jet Propulsion Laboratory, Pasadena, California, pp. 78–82, February 15, 1983.

[2] D. W. Brown, W. D. Brundage, J. S. Ulvestad, S. S. Kent, and K. P. Bartos, "Interagency Telemetry Arraying for Voyager–Neptune Encounter," *TDA Progress Report 42-102*, vol. April–June 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 91–118, August 15, 1990.

[3] M. H. Brockman, "The Effect of Partial Coherence in Receiving System Noise Temperature on Array Gain for Telemetry and Radio Frequency Carrier Reception for Receiving Systems with Unequal Predetection Signal-to-Noise Ratios," *TDA Progress Report 42-72*, vol. October–December 1982, Jet Propulsion Laboratory, Pasadena, California, pp. 91–118, February 15, 1983.

[4] R. M. Hjellming, ed., *An Introduction to the NRAO Very Large National Radio Astronomy Observatory*, Socorro, New Mexico, 1983.

[5] W. N. Christiansen and J. A. Hogbom, *Radio Telescopes*, New York: Cambridge University Press, 1985.

[6] G. L. Berge and S. Gulkis, "Earth-Based Observations of Jupiter: Millimeter to Meter Wavelengths," in *Jupiter*, T. Geherls, ed., Tucson: University of Arizona Press, pp. 621–691, 1976.

[7] I. de Pater, "Radio Images of Jupiter's Synchrotron Radiation at 6, 20 and 90 cm," *Astronomical Journal*, vol. 102, pp. 795–805, August 1991.

[8] J. S. Ulvestad, "Phasing the Antennas of the Very Large Array for Reception of Telemetry From Voyager 2 at Neptune Encounter," *TDA Progress Report 42-94*, vol. April–June 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 257–267, August 15, 1988.

Demonstration of a Joint U.S.–Russian Very Long Baseline Interferometry Tracking Capability

P. M. Kroger, B. A. Iijima, and C. D. Edwards
Tracking Systems and Applications Section

V. Altunin, V. Alexeev, B. Lipatov, and E. Molotov
Astro Space Center, Radiophysical Research Institute,
Space Device Corporation, Russia

This article discusses results of the first very long baseline interferometric (VLBI) measurements between antennas of the NASA DSN and the Russian three-station spacecraft tracking network. The VLBI systems of the U.S. and Russian tracking networks are described, and their compatibility for joint U.S.–Russian measurements is discussed. The results of a series of VLBI measurements involving Deep Space Stations and Russian tracking antennas are presented. The purpose of these first observations is to establish the compatibility of the two VLBI recording systems and verify that data recorded on these systems can be successfully correlated. The delay and delay rate observables produced by correlation of the recorded data are then used to estimate the locations of the Russian tracking stations relative to the Deep Space Stations. These first experiments, carried out at 1.7 GHz, are precursors to a future series of observations at 2 and 8 GHz, which will provide far more accurate station location estimates. The capability of the VLBI systems for joint U.S.–Russian spacecraft navigation measurements is also discussed.

I. Introduction

NASA's DSN has cooperated with Soviet (Russian) tracking stations in performing spacecraft navigation for several missions, including the Soviet VEGA (Pathfinder mission) and PHOBOS missions. Because of incompatibilities in the antenna receiving systems and uncertainties in the locations of the Soviet tracking stations relative to the DSN stations, the Soviet and DSN tracking data were analyzed independently. Recent upgrades to three tracking stations in Russia, at Evpatoria, Ussuriisk, and Bear Lakes, have now made it possible to carry out joint VLBI measurements with antennas of the DSN.

There are several potential benefits that will result from joint spacecraft VLBI measurements between DSN and Russian antennas. The increased number of stations and their geographical locations relative to the DSN stations (see Fig. 1) greatly increase the common spacecraft visibility window. An example of this is illustrated in Fig. 2 where the common visibility period of a source at various

declinations is plotted for several baselines. As is evident from this figure, the addition of the Ussuriisk station significantly increases the visibility period for sources within 20 deg of the ecliptic plane, and with DSS 14 and DSS 43 would allow observations with a three-station network. The addition of more and longer baselines would also improve the accuracy of the VLBI measurements. From an operational viewpoint, the additional antennas would decrease the demands on any single antenna for tracking time and provide greater flexibility in scheduling interferometric observations. These same benefits would also apply to differenced range and Doppler observations. Several preliminary observations have now been completed to verify the compatibility of receiving and recording systems at the Russian and DSN stations.

II. VLBI System Parameters

This section provides a brief description of the VLBI system in place at the DSN sites and the equivalent system under development at the Russian sites.

A. NASA DSN VLBI System[1]

The NASA DSN VLBI system consists of 70-m and 34-m antennas at three Deep Space Communication Complexes located at Goldstone, California; Canberra, Australia; and Madrid, Spain (see Table 1). In addition to the antennas and their receiving systems, there are two VLBI correlators, a tracking network control center, and a system for direct data transmission to JPL. The characteristics of the receiving systems are given in Table 2.

Figure 3 shows a block diagram of the DSN VLBI receiving system. Each terminal includes two subsystems: narrow-channel bandwidth (NCB) and wide-channel bandwidth (WCB). The NCB system is used for operational spacecraft tracking and for rapid calibration of Earth orientation. The limited data rate (500 kbits/sec) of the NCB system allows rapid transfer of the data to JPL via a ground communications system where the data can be processed in 12 hours or less for such time-critical applications as planetary approach maneuvers. The WCB system, on the other hand, takes advantage of greatly increased data rates and spanned bandwidths to provide much higher accuracy and sensitivity. Data rates of up to 112 Mbits/sec are recorded to wide-band video tape, which is then shipped to JPL for further processing. The WCB system is used for developing an inertial reference frame defined by the angular positions of distant, extragalactic radio sources (e.g., quasars) and for monitoring long-term motions of the DSN stations in this reference frame. The NCB and WCB systems each have their own VLBI correlator. The Block I correlator, located in JPL's Space Flight Operations Facility (SFOF), processes the NCB data transmitted to JPL via a ground communications system. The Block II correlator, located on the Caltech campus, is used to process data recorded with the WCB system. It can correlate data recorded in either of two standard radio astronomy formats: Mark III, used with the WCB system, or Mark II, an older format with a narrower bandwidth, which is still in common use at many sites. Table 3 summarizes the characteristics of the NCB and WCB subsystems.

[1] *Deep Space Network/Flight Interface Design Handbook, Vol. 1, Existing DSN Capabilities*, Document 810-5 (internal document), Rev. D, Jet Propulsion Laboratory, Pasadena, California, July 15, 1991.

B. Russian VLBI Navigation System [1]

The Russian VLBI navigation system, known as "ORION," consists of three deep space tracking stations located in the territory of the former Soviet Union: Evpatoria (Ukraine), Ussuriisk (Russia), and Bear Lakes (Russia) (see Table 1). As in the case of the DSN, there is also a VLBI correlator, a tracking network control center, a data distribution and delivery system, and an orbit determination center.

During 1992, the 70-m antennas at Evpatoria and Ussuriisk and the 64-m antenna at Bear Lakes will be equipped with 2- and 8-GHz receivers. The 1.7-GHz receivers will continue to be available in antennas at Ussuriisk and Bear Lakes. The parameters of the receiving systems are given in Table 4. Figure 4 shows a block diagram of the Russian VLBI navigation receiving terminal. Each terminal includes two subsystems: navigation and calibration. The navigation subsystem is designed exclusively for spacecraft navigation and is based on the principle of frequency bandwidth synthesis (BWS) with time-multiplexed channels [2]. The calibration subsystem is intended for calibration of the interferometer, including measurement of baseline lengths and Earth orientation parameters, positions of reference radio sources, and clock synchronization.

The 2- and 8-GHz systems will have bandwidths of 2215–2375 MHz and 8370–8530 MHz, respectively. The intermediate frequency (IF) signal, with a bandwidth of 160 MHz (375–535 MHz), is split into two signals with frequency bandwidths of 375–465 MHz and 445–535 MHz. These two signal paths are then downconverted separately into a bandwidth of 23–91 MHz with two synthesizers. The output frequency of the synthesizers can be changed through a range of ±12 MHz about the nominal frequen-

cies of 478 and 432 MHz in steps of 10 Hz. A summary of the navigation subsystem is given in Table 5.

When in the navigation mode, signals from the spacecraft and reference sources will be recorded in the Mark II format using a 3-channel, time-multiplexed BWS system. Measurements made in the calibration mode will use the Very Long Baseline Array (VLBA) recording system provided by NASA for use in the Radioastron and other missions. Tapes recorded with this system are compatible with the Mark III system tapes and can be correlated on the JPL/Caltech Block II VLBI correlator.

Each tracking station in the ORION network will use a high-quality hydrogen maser frequency standard with a stability of 10^{-14} to 10^{-15} for integration times up to 1000 sec. Clock synchronization will be performed with the Global Positioning System (GPS) or Global Orbiting Navigation Satellite System (GLONASS) receivers located at each station.

Data recorded during the calibration observing sessions in the VLBA format will be processed at the JPL/Caltech Block II correlator. Data recorded in the navigational mode will use either the Block II correlator or a correlator for the ORION system now under development in Russia. The Block II correlator output from the three frequency channels will be combined using the bandwidth synthesis technique.

III. Compatibility Issues for Joint U.S.–Russian VLBI Observations

If future joint U.S.–Russian spacecraft observations are to be possible, the compatibility of the DSN VLBI navigation system and the Russian ORION system must be resolved in several important areas. The current state of each system is discussed in this section.

A. Receiving System Bandwidth

At the beginning of 1992, 2-GHz and 8-GHz receivers were installed in the 70-m antennas at Evpatoria and Ussuriisk. Eventually, the 64-m antenna at Bear Lakes will have similar receivers. This will allow joint U.S.–Russian tracking of spacecraft broadcasting at these frequency bands. At the present time these will be only U.S. spacecraft, but the next generation of Russian spacecraft will implement 2- and 8-GHz transmitters. The Russian Mars '94 spacecraft, for example, will transmit VLBI navigation tones at 8 GHz, and the Radioastron spacecraft is planned to have an 8-GHz radio link.

For the Soviet VEGA and PHOBOS missions, the Soviet and U.S. antennas used 1.7-GHz receivers. However, there are no plans to put 1.7-GHz transmitters on future spacecraft, and this frequency will no longer be used for VLBI navigation. For this reason the 1.7-GHz receiver at the Evpatoria antenna was removed in the Fall of 1991.

B. VLBI Navigation Systems

The DSN uses the NCB system (see Sec. II.A) for operational spacecraft VLBI measurements. Data recorded with this system, however, are not compatible with the Russian VLBI systems now in place. At the present time, only VLBI data recorded with the DSN WCB system can be correlated with data recorded at Russian sites in either the Mark II or VLBA format.

C. Data Delivery

The means of transmitting the radio metric data to a central processing facility will depend upon the requirements for mission navigation. During planetary encounters, navigation data must be processed rapidly (1–2 days) in order to provide mission controllers with the information needed for orbit corrections. At JPL this is accomplished by direct transmission of the NCB VLBI data through a dedicated communications link. No means currently exists for rapid transfer of Russian VLBI data to JPL for processing.

Processing time for the calibration measurements is much less critical, and the current system of tape shipment to JPL is adequate for these observations.

D. VLBI Data Processing Systems

The JPL/Caltech Block II correlator can correlate data recorded in either the Mark II or Mark III format. This is currently the only available means of correlating VLBI data from joint U.S.–Russian VLBI measurements. The Russian tracking network will have its own correlator for processing the data from "navigation" measurements recorded in the Mark II format. For future joint U.S.–Russian spacecraft VLBI measurements, it is important that the algorithms and software of the two correlation systems be compatible.

E. Reference Systems

Processing of post-correlation VLBI delay and delay rate observables should proceed with a common set of VLBI station locations, radio source coordinates, and Earth orientation time series (polar motion and UT1–UTC). During 1991, five VLBI observing sessions involving

Russian and DSN stations were completed for the purpose of measuring the locations of the Russian stations in the reference frame of the DSN (see Section IV). An antenna located at Hobart, Tasmania, also participated in several of these measurements.

F. Coordination of Joint U.S.–Russian VLBI Measurements

By its very nature, VLBI requires close cooperation between the often widely separated antennas participating in the measurements. Scheduling of joint U.S.–Russian spacecraft VLBI observations will require coordination between the DSN and ORION network control groups. A common set of experimental parameters must be generated and transmitted to all stations participating in a spacecraft VLBI measurement. These parameters include the sequence of spacecraft and radio source observations, source coordinates for both the spacecraft and radio sources to be observed, and the configuration of the frequency channels of the VLBI recording systems. Generation of the frequency and spacecraft antenna pointing information requires knowledge of the spacecraft trajectory from the Russian or U.S. orbit determination centers. All observations would have to be scheduled well in advance to assure that the network facilities would be available and to avoid conflicts with the requirements of other missions. A reliable and rapid means of communication between Russian and DSN network control centers should be established.

IV. The First Joint DSN–Russian VLBI Experiments

During 1991, five VLBI experiments were performed to test the compatibility of the DSN and Russian VLBI systems and to obtain improved estimates of the Russian station locations relative to the DSN sites. A summary of these measurements is contained in Table 6. In these first experiments, data were recorded at 1.7 GHz in a single 2-MHz channel at each station. Because of the narrow bandwidth in these measurements, the expected accuracy of the estimated station locations was several meters. Additionally, the effect of the ionosphere on the 1.7-GHz signals was expected to introduce a significant systematic error into these results.

The compatibility of the recording systems (bandwidths, sampling rates, recorders, recording media, and data formats) is fundamental to any VLBI measurement. During these first measurements, the Russian stations recorded data on VHS tapes in the Mark II format. Data at the DSN stations were recorded with the Mark III system operating in mode D (single-channel recording). One of the most important results of these first tests was to demonstrate that data recorded in these configurations could be successfully cross-correlated at the JPL/Caltech Block II VLBI correlator.

The second major result of these tests was the estimation of the locations of the Russian tracking antennas from the delay and delay rate observables obtained from correlation of the VLBI data. This information is critical to the success of future joint U.S.–Russian spacecraft VLBI measurements and orbiting VLBI (OVLBI) where station location accuracy is an important component of the error budget. Accurate station locations for the Russian antennas in the DSN reference frame will also benefit the navigation accuracy for Doppler tracking. Future experiments at 8 GHz will significantly improve on the station location estimates provided by these initial 1.7-GHz measurements.

A. VLBI Data Processing

Processing of VLBI data proceeds in several steps beginning with the correlation of the recorded data streams from each station to produce the complex fringes, followed by extraction of the delay and delay rate observables from the fringe phase and amplitude, and finally estimation of the VLBI delay model parameters from the delay and delay rate observables. The following sections describe in some detail each of these steps for the five U.S.–Russian VLBI measurements completed in 1991.

B. Correlation of Recorded Data Streams

Each of the five VLBI measurement sessions consists of a series of repeated observations of several different radio sources. The typical duration of each observation ranged from 20 to 40 minutes. Except for the DSN sites, the data were recorded in a Mark II–compatible format. At the DSN sites, the data were recorded with a Mark III system operating in mode D, in which a single 2-MHz channel of data was recorded in successive tracks of the 28-track Mark III video tape recorder. The correlation of the raw VLBI data was performed with the JPL/Caltech Block II VLBI Correlator.

The magnitude of the correlation SNR of the interferometric fringes is a function of several experimental parameters. A correlation SNR of at least 4.5 was considered necessary in order to estimate meaningful values for delay and delay rate observables from the complex fringes. The correlation SNR is related to the experimental parameters by

$$SNR = 2.05 \times 10^{-4} \gamma S D_1 D_2 \sqrt{\frac{e_1 e_2 BT}{T_{sys_1} T_{sys_2}}} \qquad (1)$$

where γ is the ratio of the correlated flux density to the total source flux density, S is the source strength in janskys, D_1 and D_2 are the antenna diameters in meters, e_1 and e_2 are the antenna efficiencies (0–1), B is the bandwidth in hertz (2 MHz), T is the integration time in seconds, and T_{sys_1} and T_{sys_2} are the antenna system temperatures in kelvins. A low value for the correlation SNR could result from a weak source (S), resolution of an extended source ($\gamma < 1$), an insufficient integration time (T) or some problem in the recording system hardware.

In general, the quality of the recording systems was adequate to allow fringe detection ($SNR > 4.5$) on most baselines. Where fringes could not be detected, the causes were usually obvious problems in the recording hardware or instabilities in the station frequency standards of such a magnitude that adequate signal averaging was not possible during data correlation. In certain cases, however, there was no apparent reason for not detecting fringes on a given baseline. Especially peculiar was the lack of fringe detection on the Ussuriisk–DSS-43 baseline in the October 21 experiment coupled with the lower than expected correlation SNR on the Hobart–Ussuriisk and DSS-43–Hobart baselines in the same experiment. Simply based upon the antenna diameters, one would expect the correlation SNR on the DSS-43–Ussuriisk baseline to be higher than the Hobart–Ussuriisk baseline [see Eq. (1)]. Also, the short Hobart–DSS-43 baseline was expected to have a much higher than observed correlation SNR due to a high visibility factor [γ in Eq. (1)]. These facts indicate that a significant attenuation of the recorded signal occurred at DSS 43, possibly due to an error in the configuration of the seldom used L-band receiving system. Table 6 summarizes the results of the correlation for all baselines and all five experiments.

C. Extraction of Delay and Delay Rate Observables

For those baselines where correlation is successful, the output of the Block II correlator contains the complex fringes in the form of their real and imaginary parts at 2-sec averaging intervals over the length of each scan. These fringes are then processed by the post-correlation software, "FIT," [3] to provide estimates of the VLBI delay and delay rate observables and their uncertainties for each baseline and each scan. Due to the narrow 2-MHz bandwidth used in these measurements, the delay observables had relatively large random errors of 1.0×10^{-9} to 6.0×10^{-8} sec. The delay rate observables exhibited formal errors of 1.0×10^{-15} to 1.0×10^{-13} sec/sec.

D. Estimation of Station Locations

Once the delay and delay rate observables have been extracted from the correlator output, they are input to the parameter estimation software, "MODEST" [4], which can be used to estimate all parameters of the interferometric delay model. For these purposes, the interest is primarily in estimating the locations of the antennas at Bear Lakes, Evpatoria, and Ussuriisk. The other parameters of the delay model are held fixed at values obtained from independent sources. Radio source coordinates and DSN and Hobart station locations were obtained from the International Earth Rotation Service (IERS) 1990 Annual Report [5], and Earth orientation parameters (UT1–UTC, polar motion, and nutation corrections) were obtained from the IERS series 90-C-04.[2] This ensured that the Russian station location estimates would be in the IERS terrestrial reference frame. The particular reference frame used is somewhat arbitrary, but the IERS provides a well-defined, documented set of conventions for Earth orientation parameters, station locations, and radio source positions.

The best available weather data were used to estimate zenith tropospheric delays at each station. Hobart and the Russian stations did not have weather data available for the dates of these measurements so seasonal weather models were used for these sites.

The Bent model of the ionosphere [6] was used to calibrate the ionospheric delays at all the stations. This semi-empirical model uses the smoothed sunspot number, the smoothed 10-cm radio flux, station location, and the observation time to estimate the ionospheric delay at each station. The error in the Bent model calibrations is around 35 percent: The Bent model typically reproduces the diurnal shape of the ionospheric delay at a station fairly well, the 35-percent error being mainly due to a difference in overall magnitudes.[3] In addition to the diurnal behavior, there are also smaller short-term ionospheric delay fluctuations that will have a large effect on the delay rate observable.

In the parameter estimation process, the random errors on the delays and delay rates were increased to account for unmodeled tropospheric and ionospheric fluctuations. Observations with large tropospheric or ionospheric calibrations were weighted less heavily in the estimation procedure. The errors on the observables were adjusted to yield

[2] N. Essaifi, personal communication, Observatoire de Paris, May 11, 1992.

[3] H. N. Royden, personal communication, Tracking Systems and Applications Section, Jet Propulsion Laboratory, Pasadena, California, October 16, 1991.

a χ^2 per degree of freedom equal to 1. This required that the random errors on the delay observable be increased by an additional 1.0×10^{-9} to 6.0×10^{-8} sec and the delay rate observables increased by 1.0×10^{-12} sec/sec.

Simultaneous analysis of all five VLBI measurements yielded station location estimates with formal errors of 1 to 2 m in the X and Y (equatorial) components, and about 5 m in the Z (Earth rotation axis) direction (see Table 7). The much larger formal errors in the Z components are due to the fact that this coordinate is not sensitive to the delay rate observable and is, therefore, determined solely from the less-precise single-channel delay observable.

Because these observations were carried out at 1.7 GHz, the results are expected to be quite sensitive to ionospheric effects. Indeed, it was found that the values of the estimated station locations changed by several meters when the Bent model ionospheric calibrations were scaled by 35 percent about the nominal values in the various experiments. The systematic errors listed in Table 7 are the changes in the station location estimates resulting from these scalings. Varying the tropospheric calibrations was found to have little effect on the station location estimates.

The dominant error sources in the station location estimates are random and instrumental phase errors in the delay observable and ionospheric calibration errors in both the delay and delay rate observables. Both random and instrumental phase errors will be much improved in future multiple-frequency channel observations, and the ionospheric errors will be much reduced in future 8-GHz observations, or effectively eliminated in dual-frequency band (2- and 8-GHz) observations.

V. Plans for Future Observations

In 1992, U.S.–Russian VLBI tests emphasized observations at 2 and 8 GHz using both the 3-channel bandwidth synthesis system (see Section II.B) of the Russian ORION system and Mark III compatible systems. The main purposes of these observations are

(1) to complete a full test of the compatibility of the time multiplexed bandwidth synthesis system at the Russian stations with the Mark III system at the DSN stations.

(2) to use the results of these 2- and 8-GHz measurements to improve the estimates of the Russian station locations to an accuracy of 5 cm.

On March 12, 1992, a VLBI measurement was completed that involved Ussuriisk, Canberra (DSS 43), the 9-m antenna at Kauai, Hawaii, and the 26-m antenna at Gilmore Creek, Alaska. In this measurement, data were recorded in a Mark III–compatible format at 8 GHz at all sites. At Ussuriisk this was accomplished through the use of a U.S.–supplied VLBA recording system on temporary loan from the National Radio Astronomy Observatory (NRAO). Data from this measurement will be correlated at the Mark III Haystack Correlator with the post-correlation analysis and geodetic parameter estimation completed at the Goddard Space Flight Center (GSFC) and JPL. The results of this first 8-GHz experiment will improve the estimates of the Ussuriisk location by an order of magnitude over the 1.7-GHz results. Future experiments with the two other Russian sites await the installation of 2- and 8-GHz receivers and the completion of the other components of the ORION VLBI recording system.

VI. Conclusions

Five VLBI measurements involving three Russian and two DSN tracking stations have been completed and the data have been processed to provide estimates of the locations of the Russian antennas in the DSN reference frame with an accuracy of 5 to 15 m in each coordinate. These first experiments have demonstrated the compatibility of the receiving and recording systems of the two tracking networks and have shown that VLBI data from the two networks can be successfully correlated to produce delay and delay rate observables. Future experiments at other wavelengths will improve the accuracy of the station locations estimated from the data in the first series of experiments and may also include VLBI measurements of U.S. spacecraft.

Acknowledgments

The authors thank Dave Jauncey for valuable assistance in conducting the DSN observations in Australia, K. Pobedonoscev and Y. Gorshenkov at the Moscow Power Institute, Peter McCulloch and Edward King from Hobart University (Tasmania) for cooperation in the joint VLBI measurements, and Don Green and Herb Royden of JPL for supplying and explaining the ionospheric calibrations.

References

[1] *VLBI Navigation System ORION*, Space Device Corp. document, (in Russian).

[2] A. E. E. Rogers, "Very long baseline interferometry with large effective bandwidth for phase-delay measurements," *Radio Sci.*, vol. 5, pp. 1239–1247, 1970.

[3] S. T. Lowe, *Theory of Post-Block II VLBI Observable Extraction*, JPL Publication 92-7, Jet Propulsion Laboratory, Pasadena, California, July 15, 1992.

[4] O. J. Sovers, *Observation Model and Parameter Partials for the JPL VLBI Parameter Estimation Software 'MODEST'—1991*, JPL Publication 83-39, Rev. 4, Jet Propulsion Laboratory, Pasadena, California, August 1, 1991.

[5] *International Earth Rotation Service, Annual Report for 1990*, Observatoire de Paris, France, 1991.

[6] S. K. Llewellyn and R. B. Bent, *Documentation and description of the Bent ionospheric model*, Rep. AD-772 733, available from National Technical Information Service, Springfield, Virginia, 1973.

Table 1. Locations of Deep Space Station Antennas for U.S. and Russian networks.

Site	Antenna	Latitude	E. longitude	Diameter, m
Goldstone	DSS 14	35 25 33.3	243 06 40.6	70
	DSS 15	35 25 18.8	243 06 49.1	34
Canberra	DSS 43	−35 24 14.4	148 58 58.1	70
	DSS 45	−35 24 00.1	148 58 35.2	34
Madrid	DSS 63	40 25 56.6	355 45 11.0	70
	DSS 65	40 25 42.1	355 44 59.7	34
Evpatoria	DSS 52	45 11 22.0	33 11 19.0	70
Ussuriisk	DSS 47	44 00 57.0	131 45 22.0	70
Moscow	Bear Lakes	55 51 57.0	37 57 17.0	64
Hobart[a]	–	−42 48 13.0	147 26 26.0	26

[a] The 26-m antenna at Hobart, operated by the University of Tasmania, participated in these measurements, but is not part of either the Russian or the U.S. tracking network.

Table 2. Characteristics of receiving systems at DSN stations.

Antenna	Receiver	Bandwidth, MHz	System temperature, K	Polarization	Amplifier[a]
70-m	X-band	8400–8500	21	RCP[b]/LCP[c]	TWM
	S-band	2265–2305	23	RCP/LCP	TWM
	L-band	1628–1708	35	LCP[d]	FET
34-m	X-band	8400–8500	20	RCP/LCP	TWM
		8200–8600	36	RCP/LCP	HEMT
	S-band	2200–2305	38	RCP/LCP	HEMT

[a] TWM = traveling wave maser, HEMT = high-electron mobility transistor, FET = field effect transistor.

[b] RCP = right circular polarization.

[c] LCP = left circular polarization.

[d] RCP is available by performing a mechanical adjustment in the L-band receiver.

Table 3. Characteristics of DSN narrowband and wideband VLBI recording systems.

System parameter	NCB system	WCB system
Channel configuration	4 S-band, 8 X-band	14 independent channels
Channel bandwidth	250, 125, 62.5, 31.25 KHz	4, 2 MHz
Channel local oscillator	10-Hz resolution	10-KHz resolution
Sampling mode	Time-multiplexed sampling of channels, 0 to 60-sec dwell time	All channels recorded continuously
Sampling rate	500, 250, 125, 62.5 KHz	8, 4 MHz
Quantization	1 bit	1 bit
Recording medium	Disk or 9-track tape	1-in. 28-track video tape
Data transmission	Direct, via satellite relay	Shipment of tapes to JPL
Data correlation	JPL Block I VLBI correlator	Caltech/JPL Block II VLBI processor

Table 4. Characteristics of receiving systems at Russian stations.

Receiver	Bandwidth, MHz	System temperature, K	Polarization	Amplifier[a]
X-band	8370–8530[b]	45	RCP/LCP	FET
	8395–8445		RCP/LCP	FET
S-band	2270–2300[b]		RCP/LCP	FET
	2215–2375[b]		RCP/LCP	FET
L-band	1662-1692	55	LCP	

[a] FET = field effect transistor.

[b] To be installed in 1992.

Table 5. Characteristics of Russian VLBI navigation system.

System parameter	Specification
Channel configuration	3 independent channels
Channel bandwidth	0–2 MHz
Channel local oscillator	10-Hz resolution
Sampling mode	Multiplexed, 0.2-sec dwell
Sampling rate	4 MHz
Quantization	1 bit
Recording medium	VHS tapes, Mark II format
Data transmission	Shipment of tapes to correlator
Data correlation	JPL/Caltech Block II VLBI processor/ Russian ORION processor

Table 6. Summary of 1.7-GHz VLBI measurements.

Date	Start time, hr:min	Stop time, hr:min	Baseline	Length, km	Number of sources detected	Number of sources observed
March 27, 1991	00:00	12:00	Hobart–Evpatoria	11,715.9	0	5
			Hobart–Ussuriisk	8813.0	9	9
			Evpatoria–Ussuriisk	6896.2	0	5
June 19–20, 1991	20:55	08:45	Hobart–Evpatoria	11,715.9	4	6
			Hobart–Ussuriisk	8813.0	3	9
			Hobart–Bear Lakes	11,730.1	0	6
			Evpatoria–Ussuriisk	6896.2	3	6
			Evpatoria–Bear Lakes	1232.3	0	6
			Ussuriisk–Bear Lakes	6072.9	0	6
June 22, 1991	06:40	10:00	DSS 63–Evpatoria	3049.2	5	5
			DSS 63–Ussuriisk	8773.9	0	4
			DSS 63–Bear Lakes	3459.9	5	5
			Evpatoria–Ussuriisk	6896.2	4	4
			Evpatoria–Bear Lakes	1232.2	5	5
			Ussuriisk–Bear Lakes	6072.9	4	4
October 17, 1991	12:30	18:50	DSS 43–Ussuriisk	8238.5	0	6
October 21, 1991	13:15	22:05	DSS 43–Ussuriisk	8238.5	0	8
			DSS 43–Hobart	831.8	7	8
			Ussuriisk–Hobart	8813.0	5	8

Table 7. Estimates of Russian tracking station locations.

Station	Estimates of Cartesian station locations and uncertainties, m								
	X	σ_f[a]	σ_{ion}[b]	Y	σ_f^a	σ_{ion}^b	Z	σ_f^a	σ_{ion}^b
Bear Lakes	2828548.1	1.4	0.4	2206063.7	1.4	2.8	5256401.7	5.4	3.2
Evpatoria	3768306.9	1.7	3.5	2464683.0	1.3	2.6	4502254.5	4.7	3.6
Ussuriisk	−3059724.4	1.6	0.3	3427253.3	1.5	0.7	4409476.2	4.8	6.4

[a] Formal errors from parameter estimation software.

[b] Systematic error from ionospheric effects.

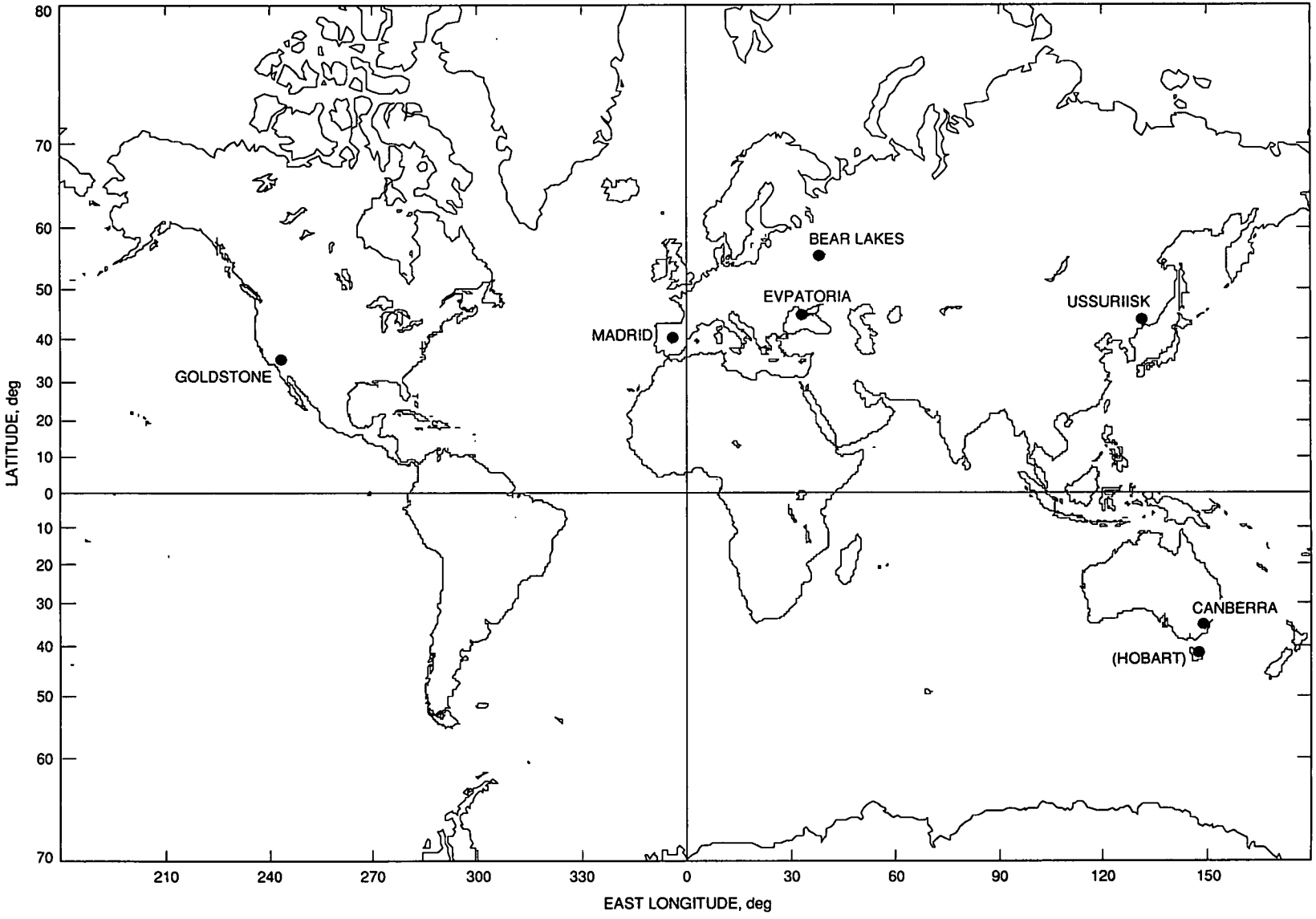


Fig. 1. Location of Deep Space Stations and Russian tracking stations. The 26-m antenna at Hobart, Tasmania, participated in the VLBI measurements described in this article, but it is not part of either the Russian or U.S. tracking network.

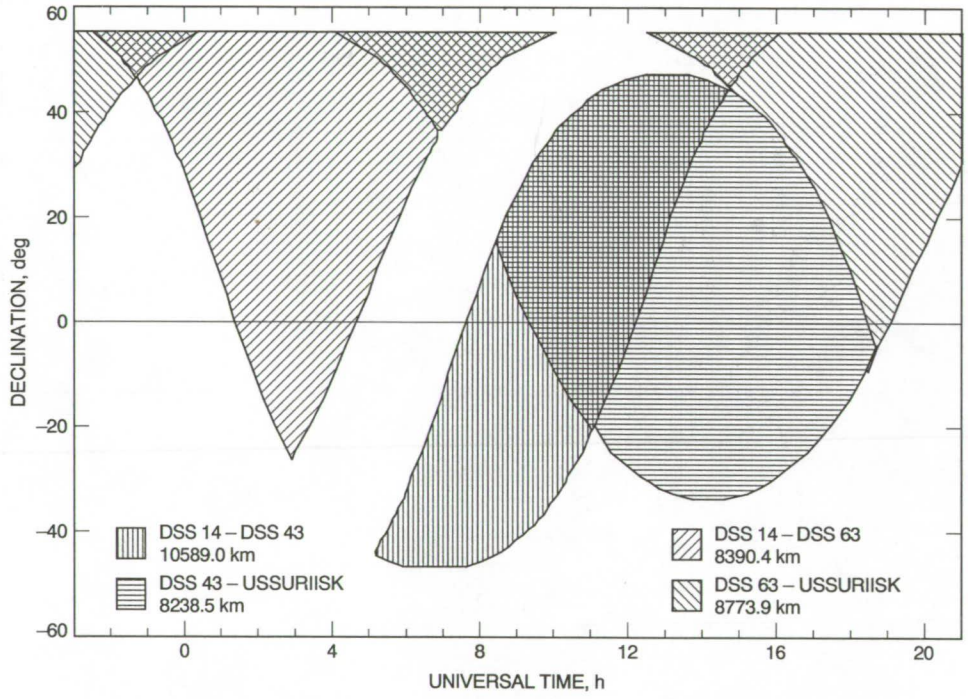


Fig. 2. Common visibility periods of a source with right ascension 0 h 0 min 0 sec, at a range of declinations on 4 different baselines.

FOLDOUT FRAME 1

FOLDOUT FRAME 2

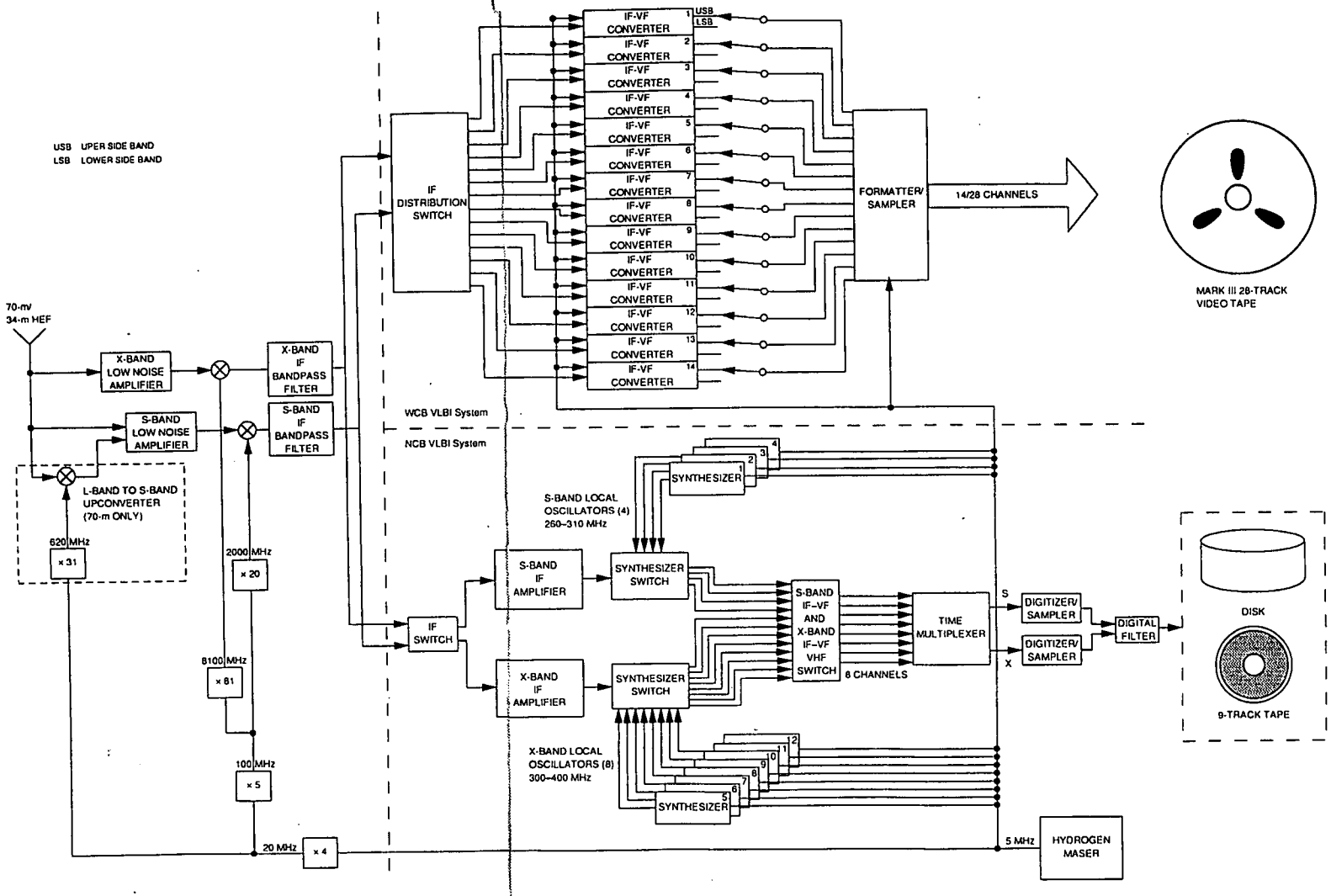


Fig. 3. Block diagram of the DSN VLBI system.

FOLDOUT FRAME 1

FOLDOUT FRAME 2

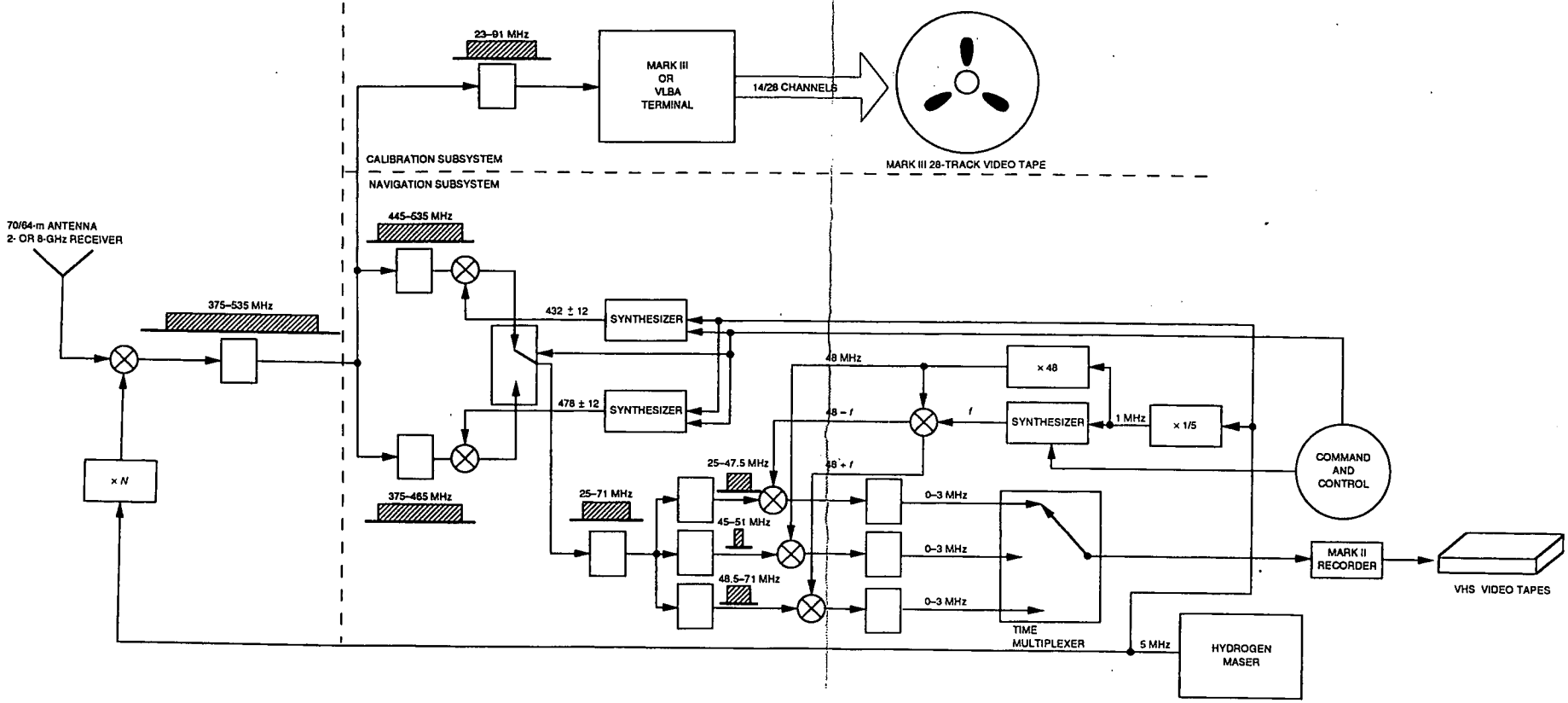


Fig. 4. Block diagram of the Russian ORION VLBI system.

Adaptive Low-Bandwidth Tracking of Galileo and Pioneer 10 Carriers

D. A. Watola
Communications Systems Research Section

In the Deep Space Network, tracking of residual carrier phase typically occurs with a fixed-bandwidth phase-locked loop using a bandwidth sufficiently wide to prevent loss of lock under worst-case conditions of signal dynamics, received signal phase noise, and receiver phase noise. Much of the time, however, such a high bandwidth is not required and may inflict unnecessarily heavy penalties on loop signal-to-noise ratios. This article describes a technique for improving tracking performance by permitting initial tracking at narrow bandwidths and gradually widening the loop as needed. The cost is a requirement for signal buffering, which is relatively inexpensive for low data rate applications. Results based on off-line processing of recorded carrier data from Galileo and Pioneer 10 are presented, and show potential 10–16 dB gains in loop SNR over worst-case fixed-bandwidth tracking.

I. Introduction

The Block IV receiver phase-locked loops (PLL's) used by the DSN to track the carrier signals of the Galileo and Pioneer 10 spacecraft have a minimum loop bandwidth of around 1.0 Hz. The next generation receiver, the Block V, allows much narrower digital loops with potential improvement in loop signal-to-noise ratio (SNR). However, the carrier loop bandwidth is usually selected conservatively to be wide enough to prevent loss of lock even under the worst-case tracking conditions of received signal phase noise, receiver phase noise, and signal dynamics. This is because carrier phase recovery is performed in real time, making loss of carrier lock extremely undesirable since it causes unlocking and loss of data in symbol detectors and error-correcting decoders in the downstream system; resynchronization of symbols takes much longer than carrier acquisition, and can result in the loss of tens of minutes of data. In the case of Pioneer, this consideration dictates that loops as wide as 3.0 Hz be used to prevent loss of lock. Worst-case conditions are relatively rare, however, and carrier recovery can potentially be achieved at much lower bandwidths most of the time, resulting in better carrier phase estimates.

This article describes an algorithm that exploits this fact without compromising the ability to recover carrier phase in the presence of significant noise transients. In the approach described here, the carrier is sampled and recorded before tracking. However, the algorithm can be modified to operate on a delayed (rather than off-line) basis. Because buffering of data is required in either approach, loss of lock at a given bandwidth is not a catastrophic occurrence since another attempt can be made at a more suitable bandwidth. Off-line or delayed computation also permits noncausal tracking. Unfortunately, this requirement also limits applicability to low data rate mis-

sions where low sampling rates permit a practical data volume; obvious examples are Pioneer 10, Voyager 1 and 2, and the Galileo S-band missions. The results presented here, though not optimum, show significant improvement over fixed-bandwidth PLL tracking for an experimental data set taken from Galileo and Pioneer 10.

II. Experimental Setup and Data

One hour of carrier from Galileo and four hours from Pioneer 10 were recorded onto a large hard disk on an IBM PC-AT compatible platform. These data were taken at the output of the Goldstone Solar System Radar (GSSR) at DSS 14 on December 19, 1991. Using the GSSR programmable local oscillator to remove predicted Doppler, the carrier was mixed down to near-baseband (100 Hz nominal) and lowpass filtered to 512 Hz before sampling at 1.024 kHz per open loop I/Q channel with 12 bits of resolution. This experimental setup is depicted in Fig. 1.

Typical spectra from these recordings are presented in Figs. 2 and 3. Here, the carriers have been shifted to near DC, lowpass filtered to 5 Hz, and subsampled by a factor of 50. This allowed better than 0.005-Hz resolution from a 2048-point fast Fourier transform, based on subsampling a 100-second (102,400-point) segment of data. These figures indicate carrier spectral line widths much smaller than the minimum fixed-bandwidth tracking loops currently used in the DSN.

III. Basic Tracking Algorithm

The adaptive carrier-tracking algorithm is built around the digital phase-locked loop (DPLL) depicted in Fig. 4. In this system, the residual carrier is assumed to have the form

$$r(t) = \sqrt{2P_c}\sin(\omega_c t + \Phi_c) + n(t) \qquad (1)$$

where P_c is the power, ω_c is the frequency, and Φ_c is the phase of the downconverted residual carrier. The noise, $n(t)$, will be ignored in this discussion since its effects on PLL behavior are thoroughly documented elsewhere e.g., [1,2]; it is only assumed that $n(t)$ is an additive white Gaussian noise (AWGN) process with power spectral density $N_0/2$. For simplicity, this treatment only considers the case where ω_c and Φ_c are not time varying.

Noisy samples recorded at rate T_s^{-1} are first processed by an automatic gain control (AGC) to normalize the residual carrier sinusoid to unit amplitude based on estimates of the sample SNR. The output from the AGC,

$$x[n] = \sin(\omega_c n T_s + \Phi_c) \qquad (2)$$

drives the DPLL, which mixes it with the numerically controlled oscillator (NCO) output waveform. Assuming that the DPLL is operating in the locked condition, the NCO output is a cosine with frequency ω_c and phase Φ_{nco}. As long as the loop is operating in lock with $\Phi_{nco} \approx \Phi_c$, the mixer output $\Phi_e[n]$ contains a DC component nearly proportional to the phase difference:

$$\Phi_e[n] = \sin(\Phi_c - \Phi_{nco}) + \sin(2\omega_c n T_s + \Phi_c + \Phi_{nco})$$

$$\approx \Phi_c - \Phi_{nco} + \sin(2\omega_c n T_s + \Phi_c + \Phi_{nco}) \qquad (3)$$

An integrate-and-dump filter, shown in Fig. 4 as a moving average followed by a decimation by N, smooths this phase error and removes the high-frequency term from Eq. (3).

The smoothed phase error passes through two filters to form Φ_{nco}. The first is a loop filter having system response function

$$H(z) = \frac{C_2 - C_1 z^{-1}}{1 - z^{-1}} \qquad (4)$$

where

$$C_1 = 4B_A \frac{r}{r+1} \qquad (5)$$

$$C_2 = C_1 + rNT_s \left(\frac{4B_A}{r+1}\right)^2 \qquad (6)$$

Here, r is a damping parameter and B_A is the noise bandwidth of the corresponding analog filter prototype [2,3]; as long as $B_A N T_s << 0.1$, $H(z)$ should provide nearly the same loop bandwidth as B_A. The second filter is the NCO, which can be modelled by a single system function having the form

$$N(z) = \frac{\Phi_{nco}(z)}{E(z)} = \frac{NT_s}{2}\frac{z+1}{z^2(z-1)} \qquad (7)$$

as described in [2] for the DSN advanced receiver. Note that use of an integrate-and-dump filter causes the NCO

output to be updated N times slower than the input sample rate.

In addition to the mixing output, the NCO also provides a second output that is phase shifted by 90 deg. Mixing this quadrature output with the DPLL input signal produces

$$\hat{\lambda}[n] = \cos(\Phi_c - \Phi_{nco}) + \cos(2\omega_c nT_s + \Phi_c + \Phi_{nco}) \quad (8)$$

which is lowpass filtered by $L(z)$ to remove the term at $2\omega_c$, yielding a lock-detection signal

$$\lambda[n] = \cos(\Phi_c - \Phi_{nco}) \quad (9)$$

The Maclaurin series expansion of Eq. (9) shows that

$$\lambda[n] = 1 - \frac{1}{2}(\Phi_c - \Phi_{nco})^2 + \cdots \approx 1 \quad (10)$$

for small phase errors. Thus, the lock detection filter output is near unity when the loop is operating in lock as long as the AGC accurately estimates P_c. Passing $\lambda[n]$ through a hysteretic comparator with thresholds L_- and L_+ to form the final Boolean lock indication signal lends improved reliability and reduces clatter [4] in the final Boolean lock indication signal, LOCK.

IV. Adaptive Bandwidth Tracking

While a software implementation of the adaptation algorithm is complex, the method is conceptually simple. Thus, a brief description suffices here.

To make the above DPLL adaptive, a supervisory loop is added to the system shown in Fig. 4. This outer loop processes recorded carrier samples at a fixed initial bandwidth, B_0, using the LOCK signal to determine which portions of the data are successfully tracked. The segment boundaries are noted, and outputs generated during a tracked section are recorded.

After tracking attempts at this bandwidth are complete, the loop bandwidth is increased and DPLL filter coefficients are recalculated. Portions of the data that remain unlocked are reprocessed at the new bandwidth; this process continues iteratively until the carrier is tracked for the duration of the recorded samples or until some maximum acceptable bandwidth is exceeded. If the filters are implemented as direct form 1 structures (Fig. 5) [5], *only* the filter coefficients need to be recomputed; filter initial conditions are completely specified by input samples and output samples generated at lower bandwidths. Alternately, direct form 2 transposed structures (Fig. 6) can be used for better numerical stability [5] at the expense of greater computational complexity.

The untracked carrier samples are actually processed by the DPLL twice at each bandwidth—once in each direction. The forward pass uses causal filtering, and the reverse pass is anti-causal. Processing carrier data backward through time in this fashion has several advantages. First, it allows the initial segment of data to be tracked if the loop does not start in the locked state. It also allows minor transients that cause the lock detector output to drop below L_- in one direction to be processed in the other direction without a change in bandwidth. More importantly, though, it allows the effects of cycle slips to be eliminated from the carrier phase.

Traditional noncausal filtering [6] cancels phase shifts by filtering data in both directions. However, in this case once carrier phase has been successfully recovered, that segment of data is no longer filtered. This implies the existence of a phase discontinuity between adjacent segments processed in opposite directions. Fortunately, because the carrier is an extremely narrowband signal located essentially at the center of the DPLL passband, the phase discontinuity between such adjacent segments is insignificant.

A demonstration of the algorithm is presented in Fig. 7. At time 0, DPLL tracking attempts begin at loop bandwidth B_1. The loop acquires lock at time t_1 and drops lock at t_2. Lock is reacquired at t_3 and is held until the end of the data, t_{end}. Thus, after the first forward pass there are two segments, s_1 and s_2, that no longer need to be processed. Subsequently, noncausal tracking begins at t_3, where the loop is able to use conditions for $t > t_3$ to start out in lock at t_3. In this example, lock is dropped again at t_4 and not regained before t_2. Segment s_1 is skipped and tracking begins at t_1, where initial conditions allow the loop to track the carrier all the way back to time 0. Without resorting to anti-causal tracking, segments s_3 and s_4 would require a higher bandwidth. Finally, after s_4 is locked, the bandwidth is increased to B_2 and tracking is tried beginning at t_2; here, B_2 is sufficient to follow the carrier for the rest of the data. This example is slightly simplified since there is a delay through $L(z)$ as the quadrature lock detector output is smoothed. However, it is adequate for summarizing the mechanics of the bandwidth adaptation algorithm.

Figure 7 is also useful in revealing a weakness of the algorithm. It is not truly adaptive because bandwidth is

never lowered once it is increased. It is certainly conceivable that some segment between t_2 and t_4 in Fig. 7 could be tracked at B_1 even though at the instants t_2 and t_4, a width of B_2 is needed. The algorithm described above will not discover this situation. In fact, no attempt to optimize the loop bandwidth is made at all; optimum bandwidth is never computed or examined as it is in [7]. Despite this apparent shortcoming, this method does effectively respond to transient disturbances in the data and provides good results without incurring a large computational burden. The algorithm can also be modified to attempt reduced bandwidths at the expense of a greater computational burden.

V. Measured Performance (Galileo Data)

Figures 8 through 11 give simulated results of applying the adaptive bandwidth tracking algorithm to one hour of Galileo carrier data recorded beginning at 22:41:03 UTC on DOY 353, 1991. For these simulations, $T_s = 1024^{-1}$ sec, $N = 100$, and $r = 2$. Switching thresholds in the lock detection comparator were selected at $L_- = 0.70$ and $L_+ = 0.75$ for 5 percent hysteresis. The simulations for Fig. 8 were made with various initial bandwidths ranging from 0.1 Hz to 1.0 Hz. Bandwidth increases within each simulation are in steps of 10 percent. Figure 9 is a similar plot for a starting bandwidth $B_0 = 0.025$ Hz; this proved to be a lower limit for successful carrier recovery using the current software implementation with a first-order loop filter.

These results indicate that for most of the hour, the carrier phase can be tracked with a 0.05- to 0.055-Hz loop. For nearly the entire hour, 0.1 Hz is sufficient. Only for a few brief transients is it necessary to resort to loops as wide as 0.6 to 0.7 Hz. By contrast, the minimum loop bandwidth for carrier tracking with current DSN equipment is 1.0 Hz. Since the phase estimation error variance in the linearized PLL model follows

$$\sigma^2_{\Phi_e} = \frac{P_c}{2B_A N_0} = \frac{1}{2SNR_L} \qquad (11)$$

a 13-dB decrease in phase error variance can potentially result over much of the tracking period. This translates directly to a potential 13-dB increase in loop SNR. In rare instances, as much as a 16-dB gain can be seen.

Figures 8 and 9 also show that successful tracking at lower B_0 does not necessarily outperform higher initial bandwidths over the complete span. This is reasonable since more noise is allowed into the tracking loop as the bandwidth is widened, so transients that appear at higher bandwidths may not affect narrower loops.

Figures 10 and 11 demonstrate respectively the recovered phase $\Phi[n]$ and the output from the lock detector $\lambda[n]$ for the first half hour of the experimental data. Note that each point on Fig. 9 where the bandwidth changes indicates a point where $\lambda[n]$ dipped below $L_- = 0.7$ and lock was lost. However, in Figs. 10 and 11 the phase is continuous and $\lambda[n] \geq L_-$ for all n since they are composed of results from multiple tracking passes at different bandwidths.

VI. Measured Performance (Pioneer 10 Data)

Figure 12 presents the results of applying the same process to a 700-sec segment of Pioneer 10 carrier data recorded beginning at 07:27:10 UTC on DOY 354, 1991. All simulation parameters are identical to those given above except that $r = 4$ and $B_0 = 0.5$ Hz. For Pioneer, the algorithm does not perform nearly as well as it did for Galileo. Although it does track the carrier with loops as low as 0.5 Hz, it is unable to recover the carrier over the entire time span; this is denoted in the plot by a bandwidth of 0 Hz.

The inability to completely recover the carrier from this data has not been completely explained. However, it should be noted that the quality of the data is suspect. Estimated P_c/N_0 in the data is consistently lower than 10 dB-Hz; the DSN, which successfully recovered data from Pioneer 10 during this period with $B_{LO} \approx 3$ Hz, requires approximately $P_c/2N_0B_{LO} \geq 6-8$ dB[1] or $P_c/N_0 \geq 13.8 - 15.8$ dB-Hz. It should also be noted that the software implementation of the tracking algorithm assumed that the input samples were white and Gaussian; both of these assumptions proved true for the Galileo data but neither applied well to the recordings from the Pioneer 10 spacecraft. Both of these observations suggest that the data recorded from Pioneer 10 are not as good as that used by the DSN.

VII. Design Considerations

Several design trade-offs were revealed during the implementation and testing of this algorithm. Most of them involve the selection of B_0 and its interaction with other

[1] Personal communication with T. Peng, Group Supervisor, Telecommunications Systems Section, Jet Propulsion Laboratory, Pasadena, California, June 9, 1992.

C-3

parameters. Intuitively, B_0 should be selected to be as low as possible in order to minimize $\sigma^2_{\Phi_e}$ and maximize SNR_L in Eq. (11). However, this choice can conflict with speed and reliability concerns.

One problem with choosing B_0 too low is that much time will be wasted processing untrackable carrier segments at this bandwidth and the loop will have to reprocess nearly the whole time span again at the next bandwidth; small bandwidth increments exacerbate this problem. Wiser choices for B_0 do not fragment the data into many short locked and unlocked partitions that require significant processor time to compute filter state variables and to resynchronize the loop with the data. In general, a more successful initial pass through the data will speed the overall tracking process.

As a practical example, a Sun Sparc II workstation requires approximately 250 seconds to process one hour of contiguous data at any bandwidth. At an initial bandwidth of 0.1 Hz, tracking the carrier over the entire Galileo data set involved 16 forward and backward passes through the data, for a total time of approximately 550 seconds. For $B_0 = 0.05$ Hz, nearly 2400 sec are required for 18 passes due to the computational overhead involved in resolving this fragmentation. Simulations show that total processing time is roughly proportional to the length of data. However, its variation with B_0 depends strongly on how much of the carrier can be recovered at B_0. With excessively low bandwidths, less time is consumed by restarting the entire process at a higher bandwidth instead of retaining carrier phase information gathered at B_0 because the loop rarely holds lock for segments significantly longer than the delay through $L(z)$. Once again, judicious selection of B_0 and ΔB can greatly speed processing.

Low initial bandwidths can also impact the effectiveness of the lock detection circuit. For very narrow loops, the lock detector output tends to hover just above L_- for long intervals. It is natural for $\lambda[n]$ to waver between L_- and unity while in lock, but an affinity for values near L_- suggests that performance may be marginal even though the loop is considered locked according to the LOCK indicator criterion. This is especially a problem if B_0 is just at the threshold of trackability and bandwidth steps are very small. Under these conditions, SNR_L computed by Eq. 11 is likely to be an overestimate, and phase error variance will be underestimated.

These two phenomena suggest that some criterion be specified for determining if a given B_0 is acceptable. The current implementation requires that a single locked region of at least 100 seconds be found in the data. Not only is such a length helpful in preventing these problems, but the long segment also acts as a "seed" providing initial conditions for tracking in either direction with wider loops. Additionally, it may be helpful to require that the initial pass through the data result in a certain minimum percentage of the recorded span being tracked. This method would be particularly useful in a near-real-time implementation where the loop can be widened when this running average drops below the threshold.

VIII. Summary

For applications in which data recording and postprocessing are viable, an adaptive, noncausal tracking algorithm that increases loop bandwidth to accommodate transients in the data can provide significantly higher loop SNR's than currently achieved in the DSN. Simulations of the algorithm described here show that the Galileo carrier phase can be recovered with DPLL's as narrow as 0.025 Hz by using only a simple first-order loop filter. This is a potential 16-dB gain in loop SNR over the DSN Block IV receiver. Most of the time, however, improvements in the range of 10 to 13 dB are possible. Results for Pioneer 10 are inconclusive but may be useful when more experimental data become available.

References

[1] F. M. Gardner, *Phaselock Techniques*, 2nd ed., New York: John Wiley and Sons, 1979.

[2] S. Aguirre and W. J. Hurd, "Design and Performance of Sampled Data Loops for Subcarrier and Carrier Tracking," *TDA Progress Report 42-79*, vol. July–September 1984, Jet Propulsion Laboratory, Pasadena, California, pp. 81–95, November 15, 1984.

[3] S. Aguirre, W. J. Hurd, R. Kumar, and J. Statman, "A Comparison of Methods for DPLL Loop Filter Design," *TDA Progress Report 42-87*, vol. July–September 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 114–124, November 15, 1986.

[4] R. C. Tausworthe, "Design of Lock Detectors," *JPL Space Programs Summary 37-43*, vol. 3, Pasadena, California: Jet Propulsion Laboratory, pp. 71–75, January 31, 1967.

[5] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*, Englewood Cliffs, New Jersey: Prentice-Hall, 1989.

[6] J. Kormylo and V. Jain, "Two Pass Recursive Digital Filter with Zero Phase Shift," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-22, pp. 384–387, October 1974.

[7] V. A. Vilnrotter, W. J. Hurd, and D. H. Brown, "Optimized Tracking of RF Carriers with Phase Noise, Including Pioneer 10 Results," *TDA Progress Report 42-91*, vol. July–September 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 141–157, November 15, 1987.

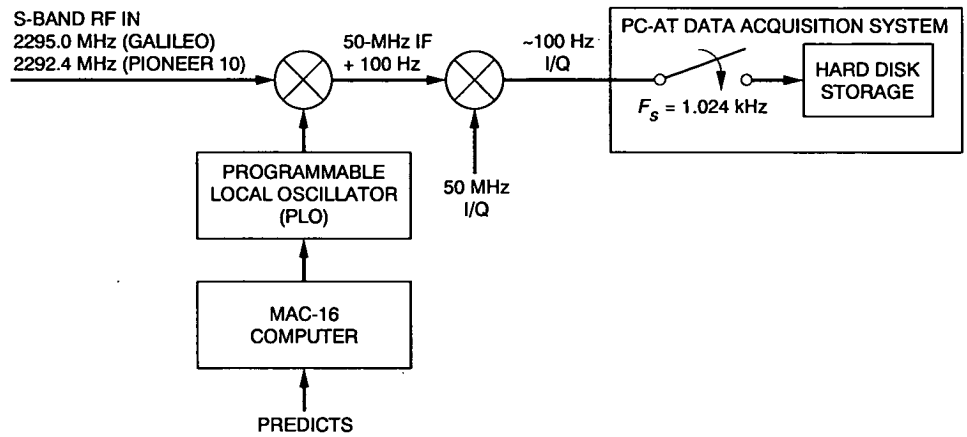


Fig. 1. Block diagram of the experimental setup.

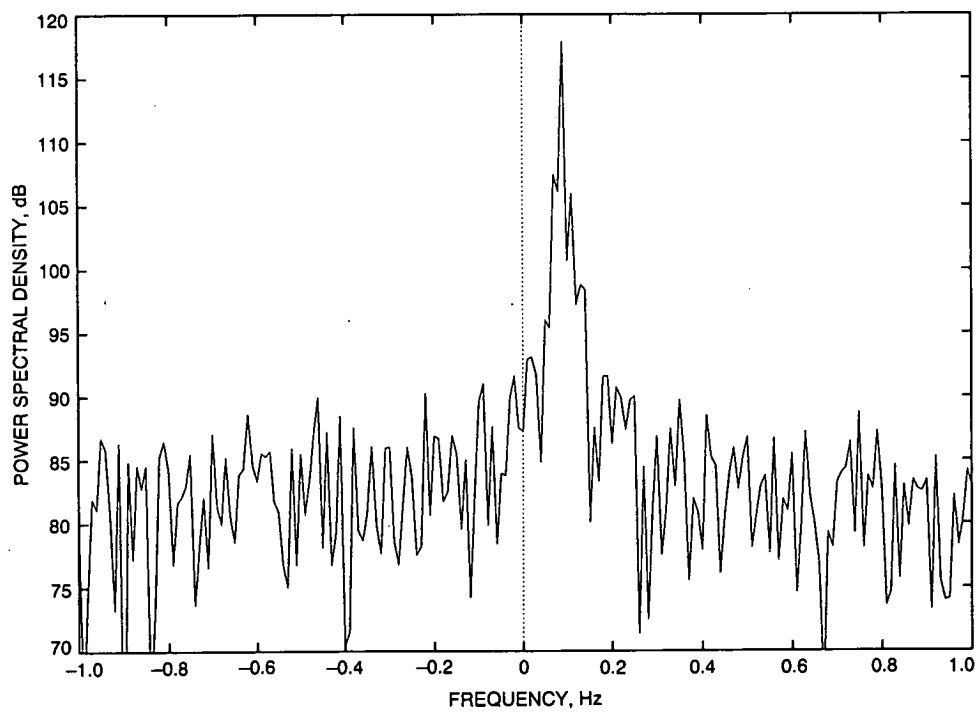


Fig. 2. Observed typical power spectrum (at baseband) for the Galileo carrier.

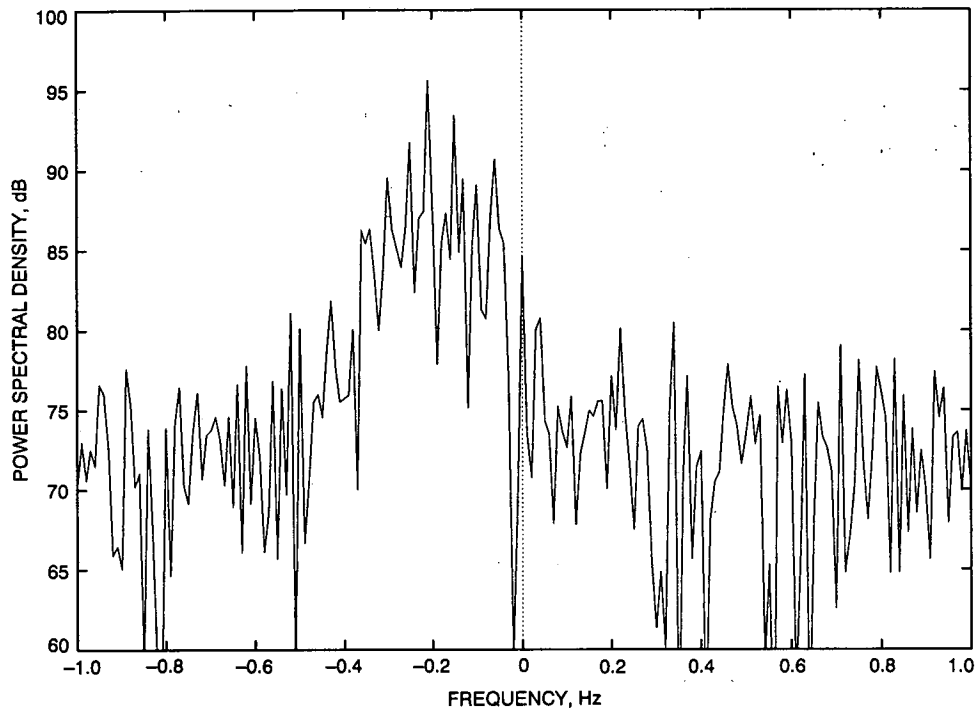


Fig. 3. Observed typical power spectrum (at baseband) for the Pioneer 10 carrier.

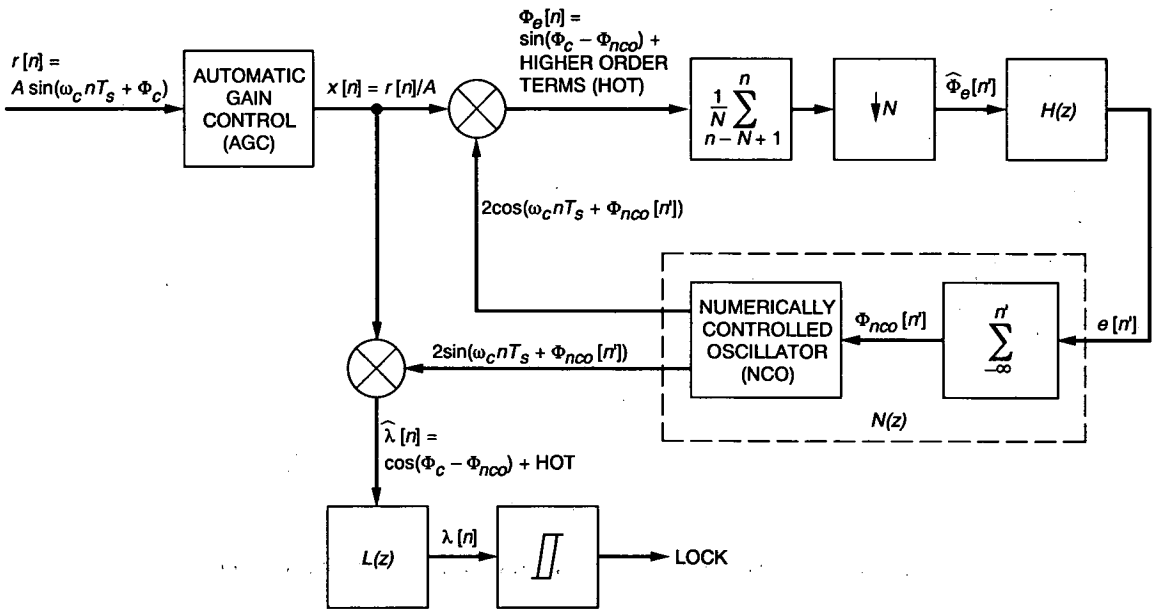


Fig. 4. Digital phase-locked loop (DPLL) with lock detector.

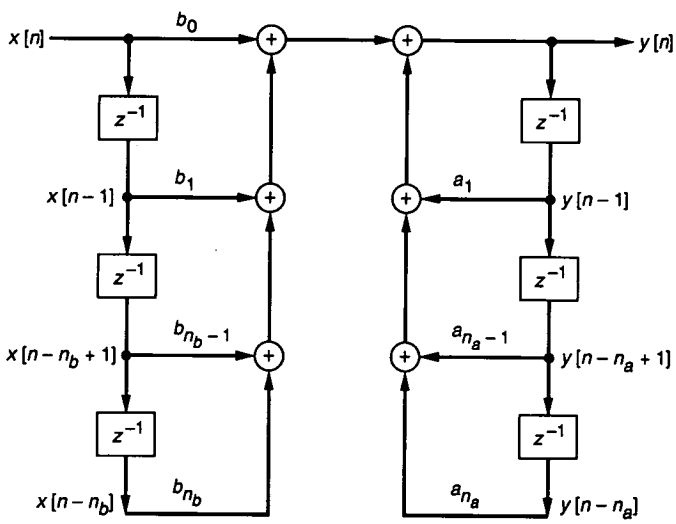


Fig. 5. Direct form 1 filter structure.

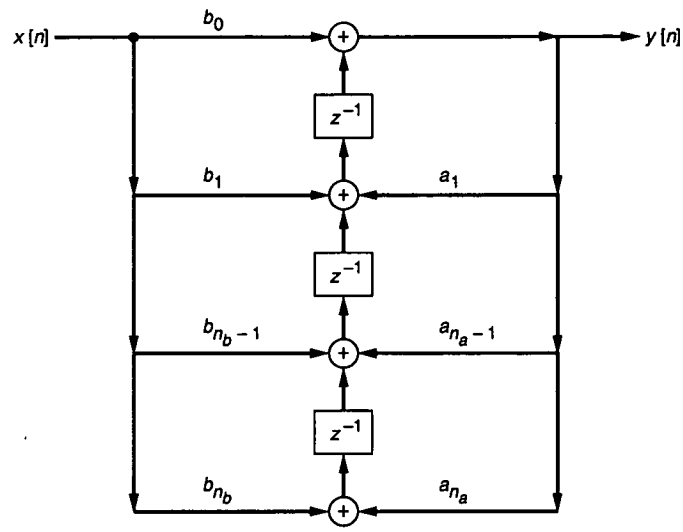


Fig. 6. Direct form 2 transposed filter structure.

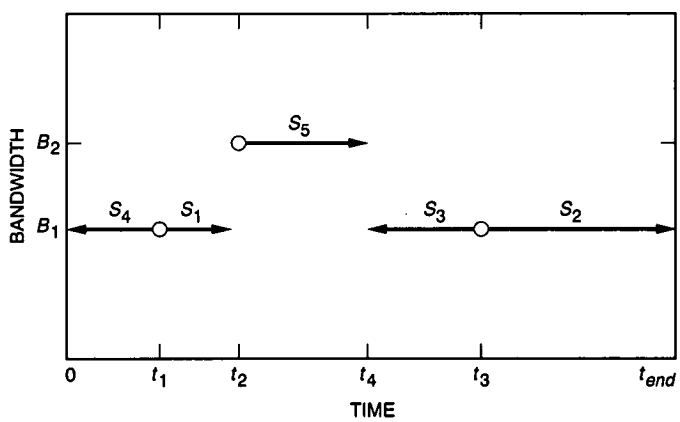


Fig. 7. Demonstration of algorithm operation.

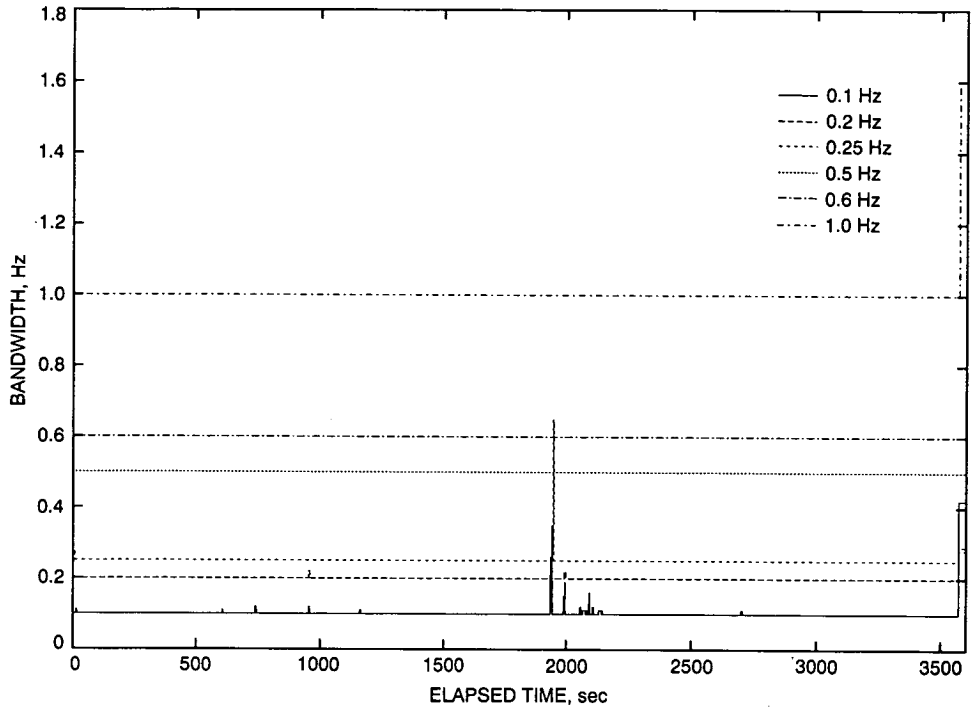


Fig. 8. Loop bandwidth for Galileo carrier phase recovery at various B_0. The events near ~3570 sec are due to a switch from one-way to two-way tracking.

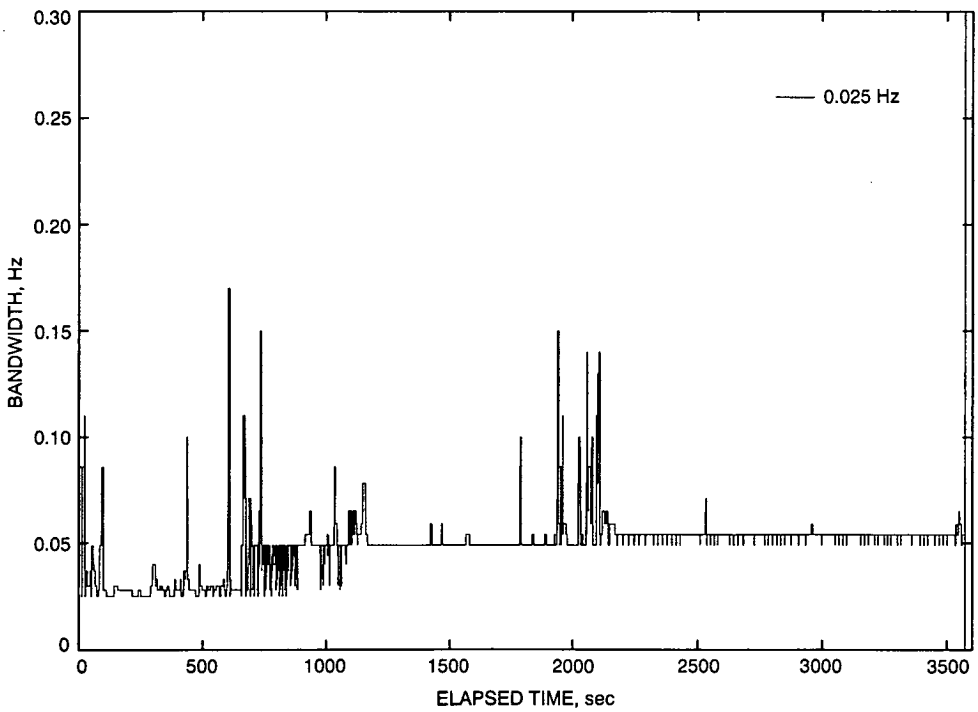


Fig. 9. Loop bandwidth for Galileo carrier phase recovery, B_0 = 0.025 Hz.

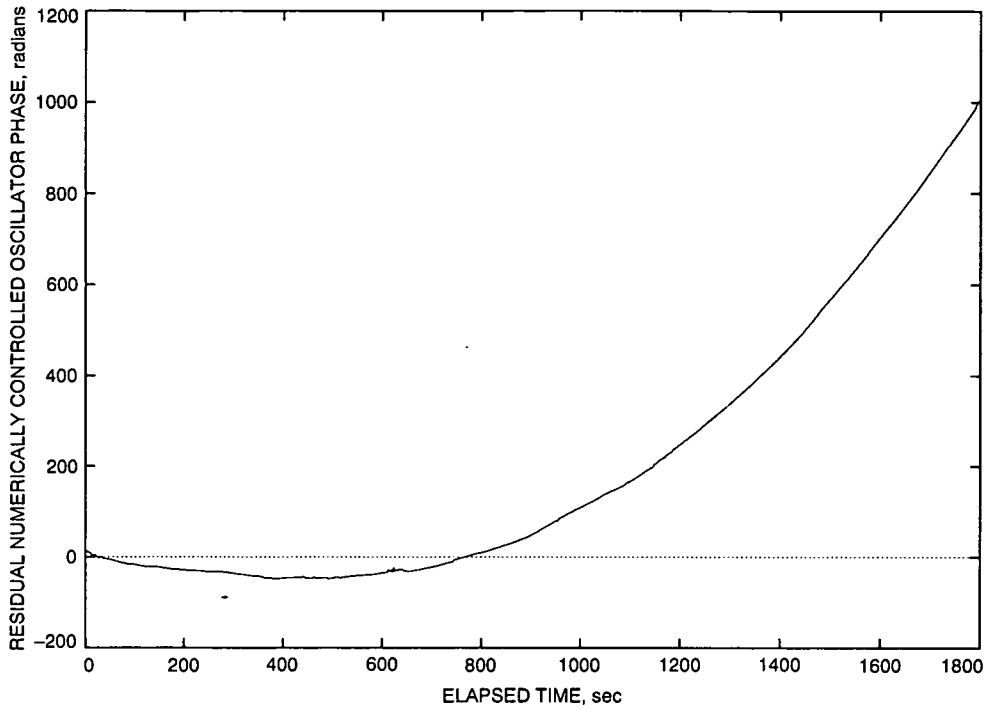


Fig. 10. Recovered phase from the first half hour of the experimental Galileo data set with initial bandwidth B_0 = 0.1 Hz.

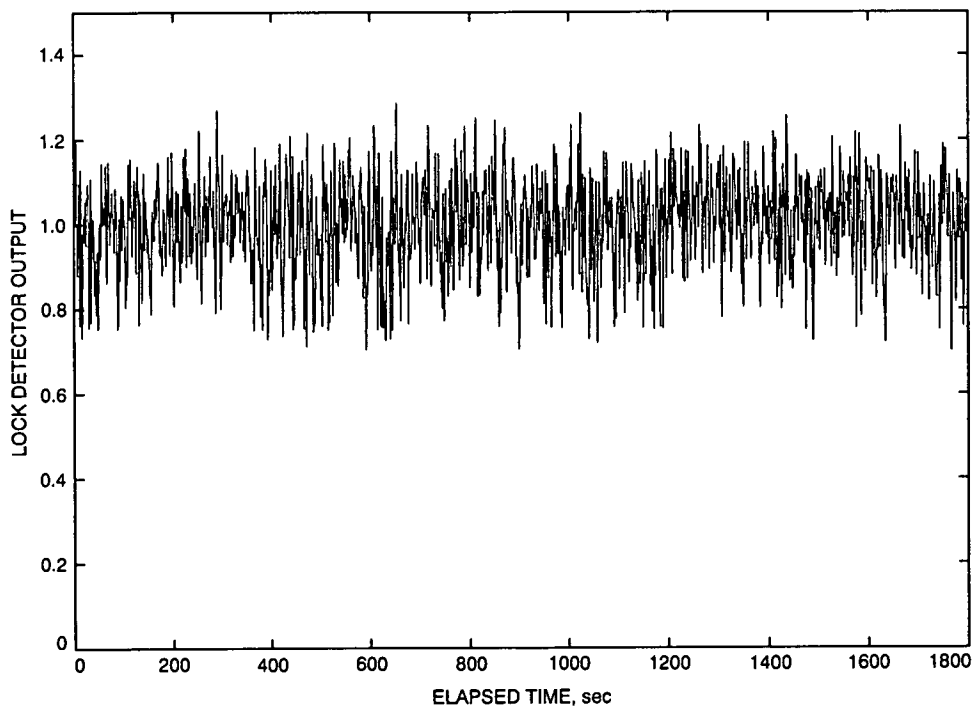


Fig. 11. Output from lock detection filter for the Fig. 10 simulation. Lower and upper hysteresis thresholds are 0.70 and 0.75.

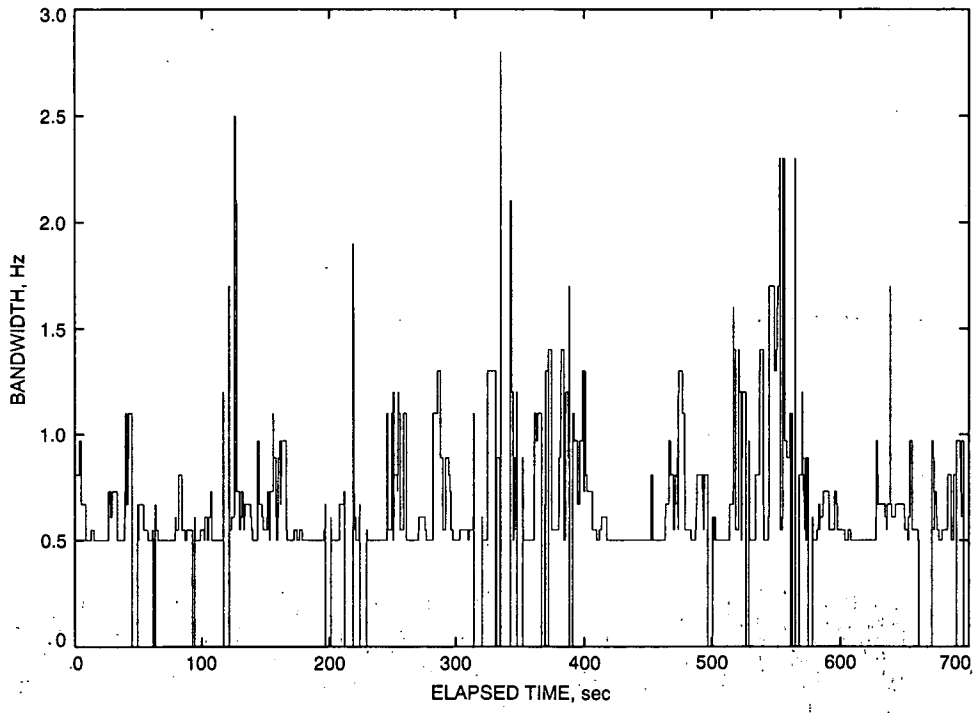


Fig. 12. Loop bandwidth for the Pioneer 10 carrier phase recovery, B_0 = 0.5 Hz.

Tracking Performance and Cycle Slipping in the All-Digital Symbol Synchronizer Loop of the Block V Receiver

M. Aung
Radio Frequency and Microwave Subsystems Section

Computer simulated noise performance of the symbol synchronizer loop (SSL) in the Block V receiver is compared with the theoretical noise performance. Good agreement is seen at the higher loop SNR's (SNR_L's), with gradual degradation as the SNR_L is decreased. For the different cases simulated, cycle slipping is observed (within the simulation time of 10^4 seconds) at SNR_L's below different thresholds, ranging from 6 to 8.5 dB, comparable to that of a classical phase-locked loop. An important point, however, is that to achieve the desired loop SNR above the seemingly low threshold to avoid cycle slipping, a large data-to-loop-noise power ratio, $P_D/(N_0B_L)$, is necessary (at least 13 dB larger than the desired SNR_L in the optimum case and larger otherwise). This is due to the large squaring loss ($\geq$13 dB) inherent in the SSL. For the special case of symbol rates approximately equaling the loop update rate, a more accurate equivalent model accounting for an extra loop update period delay (characteristic of the SSL phase detector design) is derived. This model results in a more accurate estimation of the noise-equivalent bandwidth of the loop.

I. Introduction

In the Block V receiver, an estimate of the instantaneous symbol phase is generated by the symbol synchronizer loop (SSL). An accurate estimate of the instantaneous symbol phase is necessary for sum-and-dump accumulations over a symbol period of the data, which is done in various parts of the receiver, such as in the biphase-shift-keying (BPSK) and the quadriphase-shift-keying (QPSK) Costas loops, the subcarrier loop, and the symbol signal-to-noise ratio (SNR) estimator.

The Block V symbol synchronizer loop will be an *all-digital* implementation of the data-transition tracking loop (DTTL), which has been studied in depth [1,2,3]. *All-digital* is emphasized to indicate that the entire loop, including the phase detector, is implemented digitally, as opposed to the analog phase detection used in most references of the digital DTTL.

When the number of samples per symbol is large, the behavior of the all-digital loop implementation is expected to be comparable to the equivalent analog loop as long as the loop is updated fast enough; i.e., when the loop bandwidth-update time product, B_LT_u, is much less than one ($B_LT_u \ll 1$). Computer simulations were run to make the comparison of the digital versus analog loop noise per-

formance. Results presented in this article show the level of agreement between the simulation and the equivalent model assumed for analysis.

In Section II of this article, the SSL implementation is described and its noise-equivalent bandwidth is derived. The expected phase error variance of the loop[1] is cited in Section III and compared with the simulated values in Section IV. The normalized phase error is monitored to detect cycle slipping in the simulations. Results are discussed in Section V.

II. Analysis of the Digital SSL

In this section, the digital SSL implementation is described and the linear equivalent models are derived for two cases: (1) for loop update rates approximately equaling the symbol rates ($f_u \approx R_{sym}$) and (2) for loop update rates much less than the symbol rates ($f_u \ll R_{sym}$). From the equivalent models, the actual noise-equivalent bandwidths of the loops are calculated.

A. Description of the SSL

The digital SSL design is shown in Fig. 1. The input to the loop is assumed to be data plus noise sampled at $f_s = 1/T_s$ Hz

$$r(k) = Ad(k) + n(k)$$

$$= (Ad(t) + n(t))\,|_{t=kT_s}$$

$n(k)$'s are white, Gaussian noise samples with variance $N_0/(2T_s)$ where N_0 is the one-sided spectral density of the noise. The $d(k)$'s are samples of a non-return-to-zero (NRZ) random data of data rate $R_{sym} = 1/T_{sym}$ Hz and amplitude A, shown in Fig. 2(a). The loop's estimated end-of-bit (EOB) indicator signal (which goes low when the EOB is detected) with a timing error of τ seconds is sketched in Fig. 2(b) with respect to the true data stream in Fig. 2(a). Basically, the SSL drives τ to zero.

From Fig. 1, it can be seen that the SSL is a form of the classical phase-locked loop (PLL) with the difference in phase detection. The SSL phase detector (PD) estimates the timing error τ as $\hat{\tau}$ in seconds (instead of the phase error as in the classical PLL). The estimated timing error is then converted to the symbol phase error, $\hat{\phi}$, which is filtered and used to adjust the numerically controlled oscillator (NCO). Instead of feeding back phase as in a classical PLL, symbol timing is fed back to the PD. For timing feedback, the NCO phase output is converted to pertinent symbol timing signals which control the phase detector timing. The conversion is achieved in the timing logic.

1. SSL Phase Detector. The PD design is shown in Fig. 3. The mid-phase accumulation, $M(n)$, is the sum of samples across a window width W about the estimated symbol transition [3] [the accumulation interval is shown in Fig. 2(b)]; i.e., accumulation of samples from $(1 - W/2)$ of the first symbol through $(W/2)$ of the second symbol. (W is $2^{-\ell}$ for any integer $\ell = 0, 1, 2, \ldots$.) Defining d_j to be the jth symbol,

$$M(j) \doteq \begin{cases} 2AN_s d_j \hat{\tau}/T_{sym} & \text{if a data transition is detected } (d_{j-1} \neq d_j) \\ AW d_j N_s & \text{if no data transition is detected } (d_{j-1} = d_j) \end{cases}$$

where $\hat{\tau}$ is the estimated timing error and the index j corresponds to time jT_{sym}. Due to the discrete nature of the accumulation, $\hat{\tau}$ is always quantized to an integer multiple of the sampling period; i.e., $\hat{\tau} = LT_s$, where L is an integer. However, the presence of noise makes the quantization effect negligible.

In-phase accumulation, $I(j)$, is the sum of samples over the estimated symbol period [the accumulation interval is shown in Fig. 2(b)]. For $\tau < T_{sym}/2$,

$$\text{sgn}(I(j)) = \text{sgn}(d_j) \tag{1}$$

where sgn(·) = signum function. Hence,

$$V(j) \triangleq M(j)\frac{[\text{sgn}(I(j)) - \text{sgn}(I(j-1))]}{2} \tag{2}$$

$$= \begin{cases} 2AN_s\hat{\tau}/T_{sym} & \text{if there is data transition} \\ 0 & \text{if there is no data transition} \end{cases}$$

In the presence of noise, the S-curve of the SSL has a slope K_g about the origin [1] (more details are given in Section III). In this case,

[1] M. K. Simon, personal communication, Telecommunications Systems Section, Jet Propulsion Laboratory, Pasadena, California.

$$V(j) = \begin{cases} 2AN_s K_g \hat{\tau}/T_{sym} & \text{if there is data transition} \\ 0 & \text{if there is no data transition} \end{cases} \quad (3)$$

The value $V(j)$ is accumulated over T_u seconds. Since $V(j)$ is updated at every detected EOB (which is at irregular intervals) and T_u is fixed, some asynchrony exists in the averaging process. The asynchrony is especially noticeable for symbol rates approximately equaling the loop update rate. Nevertheless, the averaged phase error estimate, $\bar{V}(m)$, is approximately

$$\bar{V}(m) \approx 2AN_s K_g M P_t \frac{\hat{\tau}}{T_{sym}} \quad (4)$$

where the time index m corresponds to time mT_u and

$$M = \frac{T_u}{T_{sym}}$$

$P_t = 1/2$ (probability of symbol transition in the data stream)

$K_g =$ Slope of the S-curve about the origin

The estimated timing error is then

$$\hat{\tau}(m) = \frac{\bar{V}(m) T_{sym}}{2AN_s K_g M P_t} \quad (5)$$

which is converted to a symbol phase error estimate as

$$\hat{\phi}(m) = 2\pi \frac{\hat{\tau}(m)}{T_{sym}} \quad (6)$$

2. Loop filter. The phase error $\hat{\phi}(m)$ is filtered in the loop filter (the same as for the standard DPLL loop filter [4,5] and is included for reference in Appendix A).

$$F(z) = \left[\alpha_1 + \alpha_2 \frac{z}{z-1} + \alpha_3 \frac{z^2}{(z-1)^2}\right] \quad (7)$$

and the loop filter output is $\Delta\omega(m)$.

3. NCO and the timing logic. The phase accumulation in the NCO is adjusted by $\Delta\omega(m)$ as

$$\hat{\theta}(k) = [G_{NCO}(\Delta\omega(k) + \omega_{init})T_s + \hat{\theta}(k-1)]_{\text{mod} G_{NCO}} \quad (8)$$

The value G_{NCO} is the NCO gain, and ω_{init} is the initial NCO frequency. Note that the NCO accumulates at the sample rate f_s Hz, a faster rate than the loop update rate, and the time index k corresponds to time kT_s.

From $\hat{\theta}(k)$, the timing logic block generates the EOB, the $(W/2)$-of-bit, and the $(1 - W/2)$-of-bit pulses, which are fed back to the PD (instead of phase, as in the classical PLL). The EOB is indicated at the time $\hat{\theta}(k)$ (before the modulo G_{NCO} operation) first equals or exceeds G_{NCO}. Similarly, $(W/2)$- and $(1 - W/2)$-of-bit are indicated when $\hat{\theta}(k)$ (before the modulo G_{NCO} operation) first equals or exceeds $(G_{NCO} \times W/2)$ and $(G_{NCO} \times (1 - W/2))$, respectively. Note that $\hat{\theta}(k)$ is also an estimate of the instantaneous symbol phase. The loop is closed as the EOB, the $(W/2)$- and $(1 - W/2)$-of-bit indicators are fed back to the in-phase and mid-phase accumulators in the phase detector.

B. Equivalent Linear Model of the SSL

A linear equivalent model of the loop is derived in this section. From the equivalent model, the noise-equivalent bandwidth of the loop, B_L^*, is estimated for calculation of the theoretical phase error variance. The derivation is made for two cases: (1) $f_u \approx R_{sym}$; and (2) $f_u \ll R_{sym}$.

1. Equivalent model for $f_u \approx R_{sym}$. For $f_u \approx R_{sym}$, an equivalent model can be derived at the symbol rate as shown in Fig. 4. The first delay models the fact that the estimated phase error at a given symbol transition is available at the PD output *after* a one-symbol period delay beyond the transition. This delay is characteristic of the SSL phase detector design (and, for example, not present in the residual carrier tracking PLL phase detector). This delay is significant when $f_u \approx R_{sym}$, and it is important that it be modeled for calculating B_L^*.

The second delay in Fig. 4 models the transport lag in the loop, and the third delay models the time delay prior to when the NCO output phase is corrected by the entire amount prescribed by the loop filter output.

The closed-loop transfer function of the model is

$$H(z) \triangleq \frac{\hat{\Theta}(z)}{\Theta(z)}$$

$$= \frac{z^{-3} F_{\theta_i}(z)}{(1 - z^{-1}) + z^{-3} F_{\theta_i}(z)} \quad (9)$$

where $F_{\theta_1}(z) = F(z)T_{sym}$. The loop filter transfer function $F(z)$ is as defined in Eq. (7).

2. Equivalent model for $f_u \ll R_{sym}$. For $f_u \ll R_{sym}$, an approximately equivalent model at the loop update rate f_u is shown in Fig. 5. The first delay approximates the delay between the PD output and the loop filter update corresponding to the accumulation time over the update period. Since in this case, several phase error estimates are accumulated over an update period, the one-symbol delay due to the phase detector is neglected. The second delay models the transport lag in the loop, and the third delay models the time delay prior to when the NCO output phase is corrected by the entire amount prescribed by the loop filter output. It should be noted that for $f_u \ll R_{sym}$, the total of three delays estimated in the equivalent model is conservative and results in an overestimated B_L^* in the linear region. The true number of delays in the loop is between 2.5 and 3. The approximate closed-loop transfer function of the loop for $f_u \ll R_{sym}$ is

$$H(z) \triangleq \frac{\hat{\Theta}(z)}{\Theta(z)}$$

$$= \frac{z^{-3}F_{\theta_2}(z)}{(1-z^{-1}) + z^{-3}F_{\theta_2}(z)} \tag{10}$$

where $F_{\theta_2}(z) = F(z)T_u = F(z)/f_u$. The loop filter transfer function $F(z)$ is as defined in Eq. (7).

C. The Noise-Equivalent Loop Bandwidth, B_L^*, of the Digital SSL

Using the linear equivalent model, the actual noise-equivalent loop bandwidth of the loop is calculated from the closed-loop transfer function as [5]

$$B_L^* = \frac{1}{2T_u H^2(1)} I_n \tag{11}$$

where

$$I_n \triangleq \frac{1}{2\pi j} \oint H(z)H(z^{-1})\frac{dz}{z}$$

The value I_n can be evaluated using methods described in [6,7].

For the parameters used in the simulations, B_L^* can be quite different from the loop bandwidth parameter, B_L, chosen for the loop.

III. Theoretical Noise Performance

The variance of the normalized phase error (λ) of the SSL was derived in [1,2] based on the assumption that the SSL is equivalent to an analog phase-locked loop (Fig. 6) when the symbol phase error is approximately constant over many symbols *and* when the loop response is much slower than a symbol period ($2B_L T_{sym} \ll 1$). The normalized phase error is defined as

$$\lambda \triangleq \frac{\tau - \hat{\tau}}{T_{sym}} \tag{12}$$

in unitless fractional cycles, and where $\tau - \hat{\tau}$ is the time offset between the true and the estimated symbol times in seconds. For uncorrupted NRZ data input with additive white Gaussian noise (AWGN) of one-sided power spectral density N_0, and a high data to noise power ratio in the loop, $P_D/(N_0 B_L)$, the normalized phase error variance, σ_λ^2, in cycles squared, is[2]

$$\sigma_\lambda^2 \triangleq \mathrm{Var}[\lambda]$$

$$= \frac{h(0)WB_L}{2R_{sym}SNR_{sym}K_g^2[1-2B_L T_{sym}]} \tag{13}$$

For comparison with the simulated all-digital SSL, the estimated noise-equivalent bandwidth, B_L^* [Eq. (11)], is used instead of the loop bandwidth parameter, B_L

$$\sigma_\lambda^2 = \frac{h(0)WB_L^*}{2R_{sym}SNR_{sym}K_g^2[1-2B_L^* T_{sym}]} \tag{14}$$

where

$$R_{sym} = 1/T_{sym} \text{ (symbol rate)}$$

$$SNR_{sym} = A^2 T_{sym}/N_0 \text{ (symbol SNR)}$$

$$h(0) \triangleq S(0,0)/W(N_0 T_{sym}/4)$$

$$= 1 + \frac{W}{2}SNR_{sym} - \frac{W}{2}\left[\frac{1}{\sqrt{\pi}}e^{-SNR_{sym}} + \sqrt{SNR_{sym}}Erf\left[\sqrt{SNR_{sym}}\right]\right]^2$$

[2] Ibid.

W = Window width of the mid-phase accumulation in the PD

$S(0,0)$ = Spectral density of the equivalent additive noise $n_\lambda(t)$ at $\omega = 0,\ \lambda = 0$

$K_g \triangleq \dfrac{\partial g(\lambda)}{\partial \lambda}\Big|_{\lambda=0}$ ($g(\lambda)$ is the normalized S-curve)

$= Erf[\sqrt{SNR_{sym}}] - \dfrac{W}{2}\sqrt{SNR_{sym}/\pi}\, e^{-SNR_{sym}}$

B_L = loop bandwidth parameter

B_L^* = loop noise-equivalent bandwidth from Eq. (11)

The phase error variance in radians squared can be expressed as

$$\sigma_{\phi_e}^2 = (2\pi)^2\sigma_\lambda^2 \tag{15}$$

$$= (2\pi)^2 \frac{\dot{h}(0) W B_L^*}{2R_{sym} SNR_{sym} K_g^2 [1 - 2B_L^* T_{sym}]} \tag{16}$$

where $\phi_e \triangleq 2\pi(\tau - \hat{\tau})/T_{sym}$. The loop SNR, SNR_L, is defined as

$$SNR_L \triangleq \frac{1}{\sigma_{\phi_e}^2} \tag{17}$$

$$= S_L \frac{P_D}{N_0 B_L^*} \tag{18}$$

where

$$P_D = A^2 \quad \text{(data power)}$$

$$S_L = \frac{2}{(2\pi)^2} \frac{K_g^2[1 - 2B_L^* T_{sym}]}{\dot{h}(0) W} \quad \text{(squaring loss)} \tag{19}$$

Note that the squaring loss S_L is less than $2/(2\pi)^2$ and approaches this value when $B_L^* T_{sym} \ll 1$, $W = 1$, and the symbol SNR is large [when $\dot{h}(0)$ and K_g equal 1]. Hence, to achieve a given loop SNR of SNR_L', the data-to-loop noise power ratio must be

$$\frac{P_D}{N_0 B_L^*} = SNR_L' S_L^{-1}$$

$$\geq SNR_L' \frac{(2\pi)^2}{2} \tag{20}$$

i.e., $P_D/(N_0 B_L^*)$ must be at least 13 dB higher than the desired loop SNR. For low SNR_{sym}, e.g., -10 dB, S_L approaches $(2/\pi^3)(SNR_{sym}/W)$; then for $W = 1$, $P_D/(N_0 B_L^*)$ must be at least [13 dB $-$ SNR_{sym}(dB)] higher than the required loop SNR.

IV. Simulation Results

For noise evaluation, the variance of the normalized phase error, $\hat{\sigma}_\lambda^2$, is measured via computer simulations and compared with the theoretical value of Eq. (14). Agreement between the two is expressed through the percentage error, $\Delta e\%$, defined as

$$\Delta e\% \triangleq \frac{\hat{\sigma}_\lambda^2 - \sigma_\lambda^2}{\sigma_\lambda^2} \times 100\% \tag{21}$$

Also, $\hat{\lambda}$, defined as the normalized difference between the true symbol phase and the NCO estimated symbol phase, $(\tau - \hat{\tau})/T_{sym}$, is monitored in the simulations to detect the occurrence of cycle slipping during the simulation.

All simulations were made for 10^4 seconds unless otherwise stated. Since the loop bandwidths ranged from 0.2 to 20 Hz, there were at least 2000 inverse loop bandwidths in each simulation, which is enough to identify the cycle-slipping threshold SNR_L. Simulations were made for the two cases, $f_u \approx R_{sym}$ and $f_u \ll R_{sym}$, for decreasing values of loop SNR's.

A. First Case: $f_u \approx R_{sym}$

For $f_u \approx R_{sym}$, simulations were made for the first- and the second-order loops with the asynchronous implementation where the loop is updated at a fixed period of T_u seconds. Simulations were made with both integer and noninteger number of samples per symbol. The parameters used in the simulations are

$f_s = 10^5$ Hz (sampling frequency)

$R_{sym} = f_s/N_s$ (symbol rate)

N_s = number of samples per symbol

f_u = 1000 Hz

SNR_{sym} = $(A^2T_{sym})/N_0$ (symbol SNR)

B_L = loop bandwidth parameter in Hz

W = 1 (window width of the mid-phase accumulation)

For the first set of simulations, $SNR_{sym} = 5$ dB, and N_s was set to 100.0 and 100.001, which results in $R_{sym} \approx 1000$ Hz. Results are shown in Table 1.

The unreasonably large $\Delta e\%$'s are due to cycle slipping. This is confirmed by observation of $\hat{\lambda}(t)$. In Figs. 7(a) and (b), plots of $\hat{\lambda}(t)$ for $t = 0.9 \times 10^4$ seconds to 10^4 seconds of the simulations are shown for $N_s = 100.001$, and $B_L = 13$ Hz and $B_L = 17$ Hz, respectively. No cycle slipping is seen for $B_L = 13$ Hz, whereas cycle slipping is apparent for $B_L = 17$ Hz. It is seen that cycle slipping occurs at $SNR_L \approx 8.5$ dB and below. For SNR_L's above 8.5 dB, gradual degradation of $\Delta e\%$ is observed with the gradual decrease in SNR_L.

For the same parameters as above, additional simulation results for the second-order loop are shown in Table 2.

For the second-order loop, cycle slipping is observed at a higher loop SNR than that of the first-order loop, which is characteristic of a PLL. Plots of $\hat{\lambda}(t)$ for $N_s = 100.001$, and $B_L = 3$ Hz and $B_L = 15$ Hz, respectively, are shown in Figs. 8(a) and (b) where no cycle slipping is present for $B_L = 3$ Hz and cycle slipping is present for $B_L = 15$ Hz. Extensive simulations with gradually decreasing SNR_L's must be made to determine the actual cycle-slipping threshold for this case.

An additional set of simulation results for the first-order loop and the same parameters as above, but for a lower $SNR_{sym} = -1$ dB, is shown in Table 3.

B. Second Case: $f_u \ll R_{sym}$

For $f_u \ll R_{sym}$, cycle slipping was observed to start occurring at lower loop SNR's in the following simulations of the first-order loop. The parameters used in the simulations are

B_L = 3 Hz $(B_L^* = 4.08$ Hz)

f_s = 10^5 Hz

R_{sym} = f_s/N_s

N_s = 100.001 (number of samples per symbol)

f_u = 100 Hz

W = 1 (window width of the mid-phase accumulation)

and decreasing values of SNR_{sym}'s. Results are summarized in Table 4.

Cycle slipping is seen to start at an SNR_L between 6.0 and 6.8 dB. Note, however, that even though this minimum required loop SNR is small, squaring loss is large for these parameters and a $P_D/(N_0B_L^*)$ greater than 22.4 dB is necessary to achieve the minimum loop SNR to avoid cycle slipping. Extensive simulations with gradually decreasing SNR_L's must be made to determine the exact cycle-slipping threshold for this case.

To check the agreement between σ_λ^2 and $\hat{\sigma}_\lambda^2$ in the absence of cycle slipping, "quick" simulations (of simulation time = 30 seconds) were run for decreasing SNR_L for the following parameters. Results are shown in Table 5.

f_s = 10^5 Hz

R_{sym} = f_s/N_s

R_u = 50 Hz

N_s = 100.0 (number of samples per symbol)

SNR_{sym} = 5 dB

The difference $\Delta e\%$ is seen to grow as SNR_L is decreased. The very large negative values just show the invalidity of the equivalent model when B_LT_u is large (note that $B_LT_u = 0.1$ at this point).

V. Discussion of the Simulation Results

From the simulation results, it can be seen that $\Delta e\%$ increases with the decrease in the SNR_L until a loop SNR threshold is reached below which cycle slipping occurs (within the simulation time of 10^4 seconds). Cycle slipping is confirmed from the observation of $\hat{\lambda}(t)$, as shown in Figs. 7 and 8. The cycle-slipping threshold SNR_L varies for different cases. For $f_u \approx R_{sym}$, in a first-order loop, the cycle-slipping loop SNR threshold is approximately 8.5 dB; for the second-order loop, the threshold is higher, above 9.3 dB. For $f_u \ll R_{sym}$, the cycle-slipping threshold in the first-order loop is seen to

be between 6.0 and 6.8 dB; further simulations need to be executed to find a more accurate threshold. Additional simulations are also necessary to study the different cycle-slipping thresholds. In all cases, the cycle-slipping threshold loop SNR's are reasonably small. However, the difference in the SSL is that to achieve the small minimum SNR_L, a large $P_D/(N_0B_L^*)$ which is at least 13 dB higher than the desired SNR_L is necessary; a larger $P_D/(N_0B_L^*)$ is required when the symbol SNR is smaller. For example, to achieve the minimum required 8.5 dB for $f_u \approx R_{sym}$, $SNR_{sym} = 5$ dB in the first-order loop, $P_D/(N_0B_L^*) > 22$ dB is required. Similarly, to achieve the minimum required loop SNR (which is around 6 dB) in the simulations of $f_u \ll R_{sym}$, $SNR_{sym} = -1.5$ dB in the first-order loop, $P_D/(N_0B_L^*) > 22.4$ dB is required. This is due to the large squaring loss that exists in the SSL [Eq. (19)].

Prior to the loop breakdown due to cycle slipping, gradually increasing $\Delta e\%$ between the model and the simulated variances was observed for decreasing SNR_L. (Note that in the simulations, SNR_L was decreased by increasing B_L at a fixed symbol SNR, or by decreasing the symbol SNR at a fixed loop bandwidth.) One possible cause may be the self-noise due to the randomness that exists within the loop; i.e., the random phase detector output of 0 or $\hat{\phi}$ with transition probability P_t. In the simulations, the randomness in the phase detector output is compensated for by dividing the PD output by P_t, assuming then that the expected value of the PD output is $P_t\hat{\phi}$ and the expected value after the compensation is $\hat{\phi}$. In reality, the averaged value varies about $\hat{\phi}$ from one update time to another. This variation can cause additional variations in the NCO phase values which are used to estimate $\hat{\sigma}_\lambda^2$. As B_L is increased, variations are increased resulting in a larger additional variance.

For $f_u \approx R_{sym}$, another contribution to the increasing $\Delta e\%$ may be the nature of the PD design which detects the phase error after one symbol period delay, adding one delay to the equivalent digital loop (Fig. 4). As the number of transport delays in a digital PLL is increased, the equivalence between the digital and its equivalent analog loops collapses for smaller values of B_LT_u.

VI. Conclusion

The all-digital SSL design was computer simulated and the performance was compared with the theoretical model performance. Results show good agreement for higher loop SNR's, and gradual degradation in the agreement as SNR_L was decreased. Eventually, cycle slipping starts to occur (within the simulation time of 10^4 seconds) at SNR_L's below the cycle-slipping threshold. The cycle slipping thresholds appear to be comparable to that of the classical PLL. However, a very large $P_D/(N_0B_L^*)$ is necessary to achieve the seemingly small required minimum loop SNR, due to the large squaring loss that is inherent in the SSL.

Acknowledgments

The author thanks W. J. Hurd, M. K. Simon, S. Hinedi, A. Mileant, S. Million, J. B. Thomas, J. B. Berner, P. K. Kinman, and J. M. Layland for their helpful discussions and comments.

References

[1] W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering*, Englewood Cliffs, New Jersey: Prentice-Hall, pp. 442–457, 1973.

[2] M. K. Simon, "An Analysis of the Steady-State Phase Noise Performance of a Digital Data Transition Tracking Loop," *JPL Space Programs Summary 37-55*, vol. 3, pp. 54–62, February 28, 1969.

[3] W. J. Hurd and T. O. Anderson, "Digital Transition Tracking Symbol Synchronizer for Low SNR Coded Systems," *IEEE Transactions on Communication Technology*, vol. COM-18, no. 2, pp. 141–147, April 1970.

[4] S. Aguirre, "Acquisition Times of Carrier Tracking Sampled Data Phase-Locked Loops," *TDA Progress Report 42-84*, vol. October–December 1985, Jet Propulsion Laboratory, Pasadena, California, pp. 88–93, February 15, 1986.

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

[6] E. I. Jury, *Theory and Application of the z-Transform Method*, Malabar, Florida: Robert E. Krieger Publishing Company, pp. 297–299, 1986.

[7] R. Winkelstein, "Closed Form Evaluation of Symmetric Two-Sided Complex Integrals," *TDA Progress Report 42-65*, vol. July and August 1981, Jet Propulsion Laboratory, Pasadena, California, pp. 133–141, October 15, 1981.

Table 1. Percentage error Δe % in a first-order SSL for $f_U \approx R_{sym}$, SNR_{sym} = 5 dB, increasing B_L.

B_L, Hz	B_L^*, Hz	$B_L T_u$	$P_D/(N_0 B_L^*)$, dB	SNR_L, dB	$N_s = 100.0$, Δe%, percent	$N_s = 100.001$, Δe%, percent
0.2	0.2004	0.0002	42	28.7	2.18	–
0.7	0.704	0.0007	36.5	23.3	2.49	–
1.5	1.52	0.0015	33.2	19.9	3.08	–
3	3.09	0.003	30.1	16.8	4.5	2.19
5	5.26	0.005	27.8	14.5	6.2	–
7	7.5	0.007	26.2	13.8	8.9	3.32
8	8.67	0.008	25.6	12.3	–	3.68
9	9.85	0.009	25.1	11.8	9.21	–
10	11.06	0.01	24.6	11.2	9.95	4.5
13	14.8	0.013	23.3	9.9	–	5.7
14	16.1	0.014	22.9	9.6	–	6.2
15	17.4	0.015	22.6	9.2	13.7	6.7
16	18.8	0.016	22.3	8.7	14.35	7.15
17	20.2	0.017	22.0	8.5	14.7	962
18	21.59	0.018	21.7	8.2	1.5×10^4	–
20	24.48	0.02	21.1	7.7	2×10^5	1790

Table 2. Percentage error Δe % in a second-order SSL for $f_U \approx R_{sym}$, SNR_{sym} = 5 dB, increasing B_L.

B_L, Hz	B_L^*, Hz	$B_L T_u$	$P_D/(N_0 B_L^*)$, dB	SNR_L, dB	$N_s = 100.0$, Δe%, percent	$N_s = 100.001$, Δe%, percent
3	3.07	0.003	30.1	16.9	–	3.4
10	10.85	0.01	24.6	11.1	9.9	–
15	16.99	0.015	22.7	9.3	290	3300
16	18.27	0.016	22.4	9.0	–	8600

Table 3. Percentage error Δe % in a first-order SSL for $f_u \approx R_{sym}$, SNR_{sym} = 1 dB, increasing B_L.

B_L, Hz	B_L^*, Hz	$B_L T_u$	$P_D/(N_0 B_L^*)$, dB	SNR_L, dB	$N_s = 100.001$, Δe%, percent
2	2.04	0.002	26	9.9	15.8
3	3.09	0.003	24.1	8.0	Cycle slipping occurs
3.5	3.62	0.0035	23.4	7.3	Cycle slipping occurs
4	4.16	0.004	22.8	6.7	Cycle slipping occurs

Table 4. Percentage error Δe % in a first-order SSL for $f_u \ll R_{sym}$, decreasing SNR_{sym}.

SNR_{sym}, dB	$P_D/(N_0 B_L^*)$, dB	SNR_L, dB	Δe %, percent	Comments
3	26.9	13.1	−5.6	No cycle slipping
0	23.9	8.5	1.97	No cycle slipping
−0.9	23	7	7.8	No cycle slipping
−1	22.9	6.8	8.4	No cycle slipping
−1.5	22.4	6.0	–	Cycle slipping occurs
−2.0	21.9	5.1	–	Cycle slipping occurs
−2.53	21.4	4.1	–	Cycle slipping occurs

Table 5. Percentage error Δe % in first-order SSL for $f_u \ll R_{sym}$ for a short simulation time = 30 seconds, SNR_{sym} = 5 dB, increasing B_L.

B_L, Hz	B_L^*, Hz	$P_D/(N_0 B_L^*)$, dB	SNR_L, dB	σ_λ^2 (cycles)^2	$\hat{\sigma}^2$ (cycle	Δe%, percent
1.5	2.04	32	18.7	3.4448×10^{-4}	3.4753×10^{-4}	0.88
2.0	3.03	30.2	16.9	5.128×10^{-4}	5.02×10^{-4}	−2.1
2.5	4.24	28.7	15.5	7.198×10^{-4}	6.89×10^{-4}	−4.3
3.0	5.7	27.4	14.1	9.772×10^{-4}	9.22×10^{-4}	−5.6
5.0	17.6	22.5	9.2	3.07×10^{-3}	2.23774×10^{-3}	−27

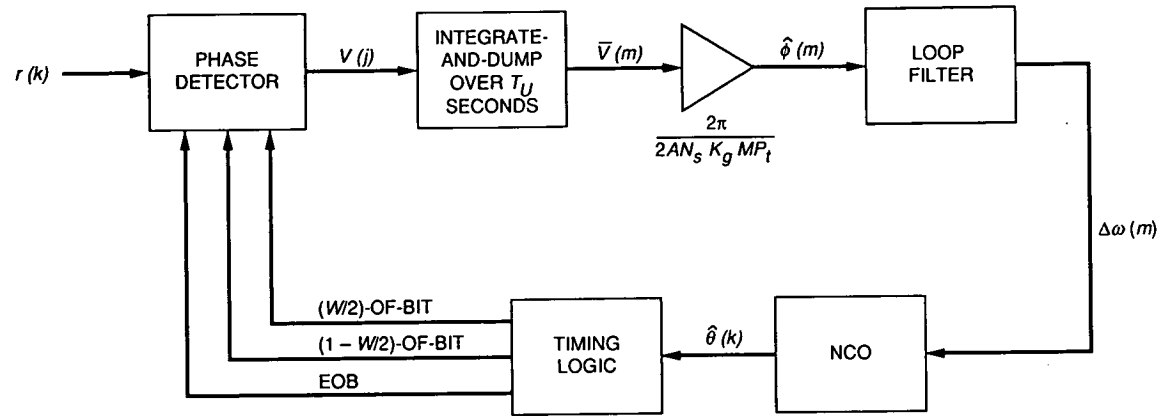


Fig. 1. The Block V digital symbol synchronizer loop.

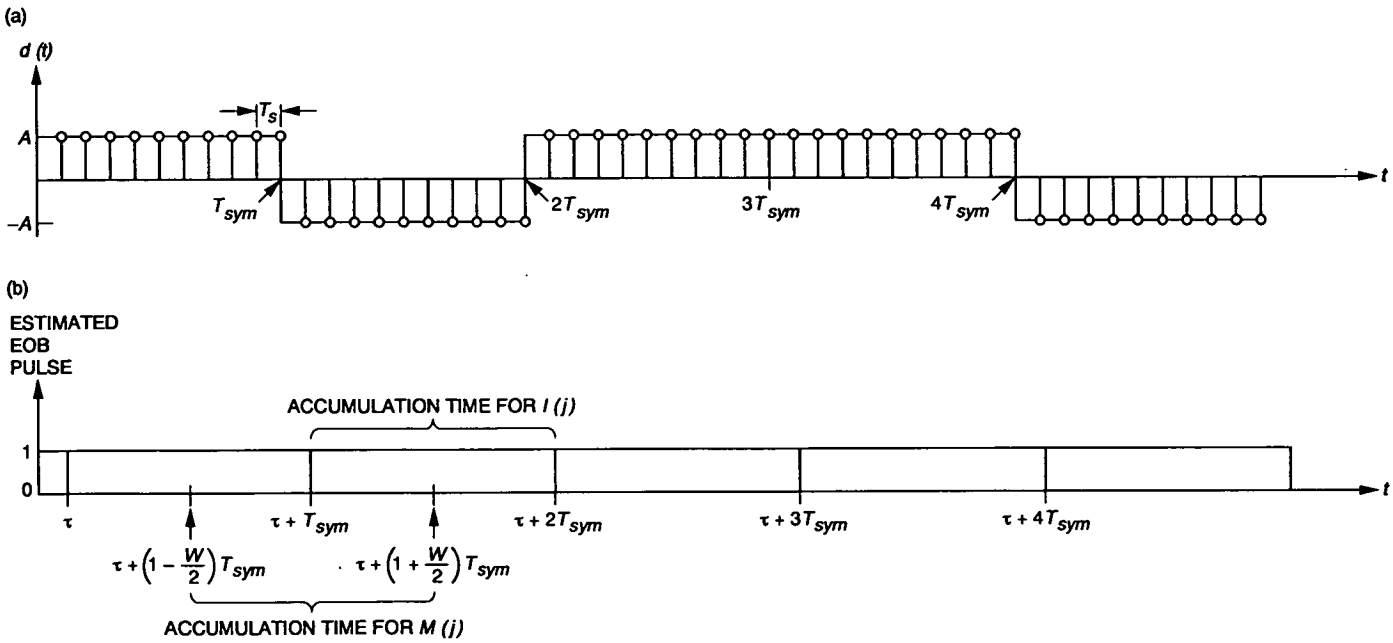


Fig. 2. Input data stream and loop estimate of the EOB-time: (a) input data to the SSL and (b) estimated EOB timing with an error of τ seconds. Also shown are integration times for the in-phase and mid-phase integrations.

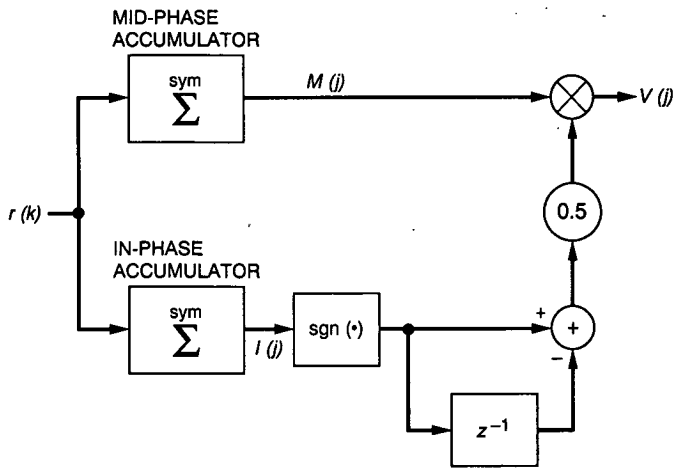


Fig. 3. The SSL phase detector design.

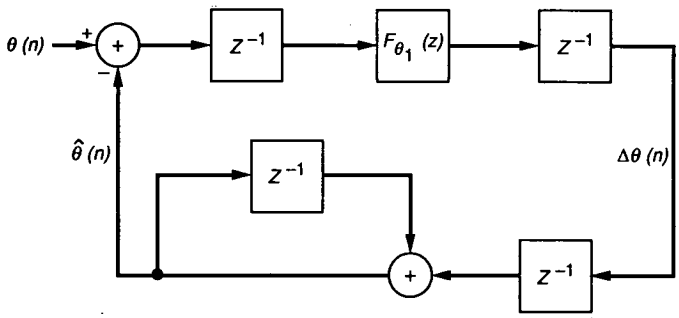


Fig. 4. Equivalent digital PLL of the SSL for $f_u \approx R_{sym}$.

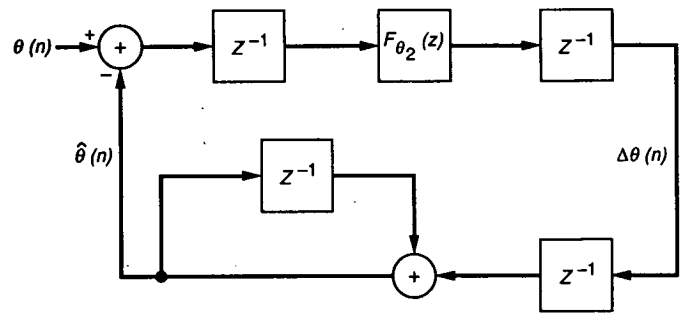


Fig. 5. An approximately equivalent digital PLL of the SSL for $f_u << R_{sym}$.

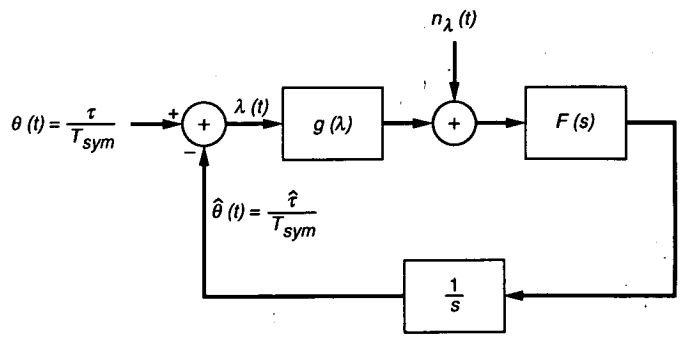


Fig. 6. Equivalent analog PLL of the SSL [1].

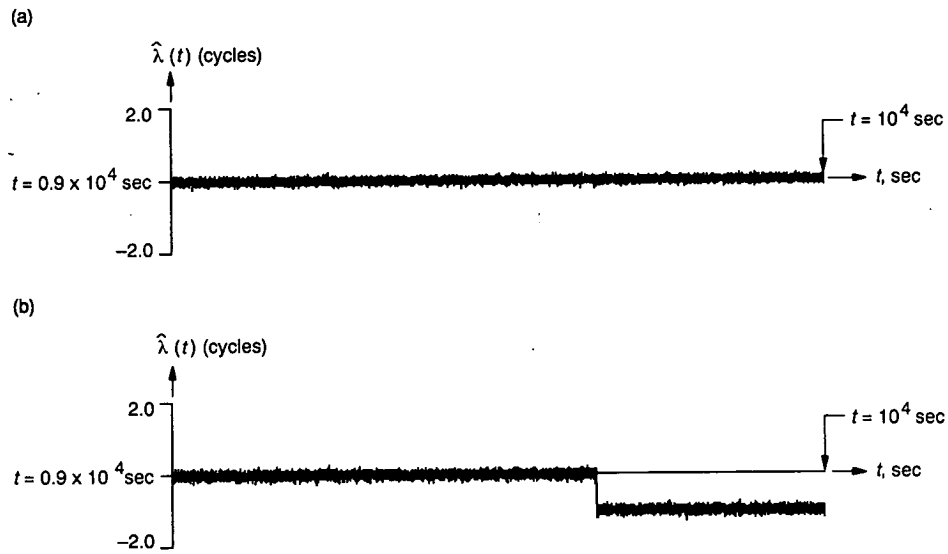


Fig. 7. $\hat{\lambda}(t)$ for a first-order SSL, $R_{sym} \approx f_U$: (a) $SNR_L = 9.9$ dB and $B_L T_u = 0.013$ and (b) $SNR_L = 8.5$ dB and $B_L T_u = 0.017$.

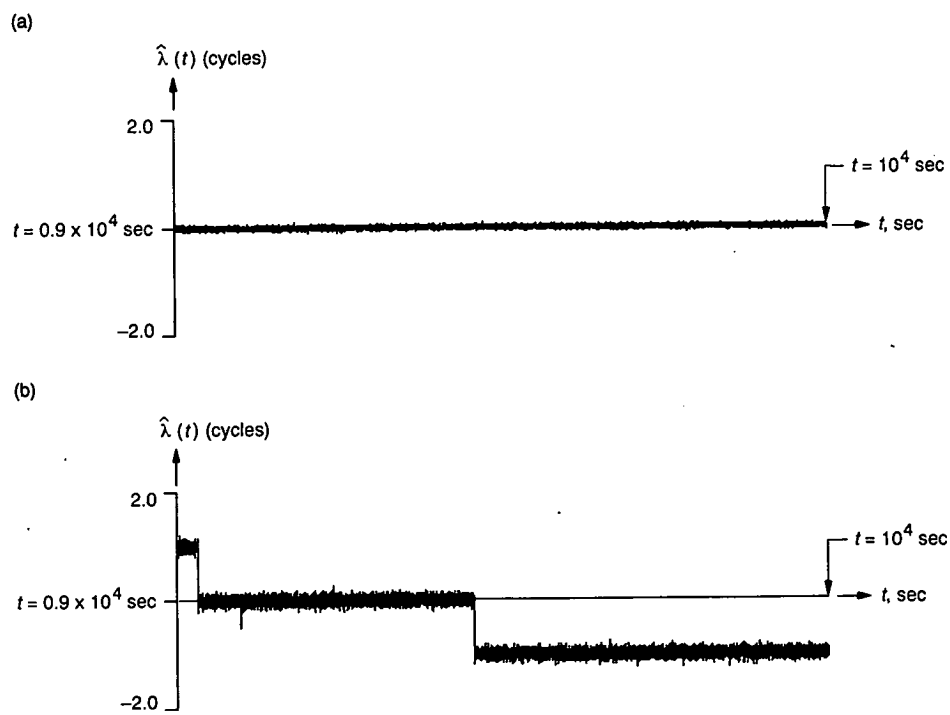


Fig. 8. $\hat{\lambda}(t)$ for a second-order SSL, $R_{sym} \approx f_U$: (a) $SNR_L = 16.9$ dB and $B_L T_u = 0.003$ and (b) $SNR_L = 9.3$ dB and $B_L T_u = 0.015$.

SNR Degradation in Square-Wave Subcarrier Downconversion

Y. Feria and J. Statman
Communications Systems Research Section

This article presents a study of signal-to-noise ratio (SNR) degradation in the process of square-wave subcarrier downconversion. The study shows three factors that contribute to the SNR degradation: the cutoff of the higher frequency components in the data, the approximation of a square wave with a finite number of harmonics, and nonideal filtering. Both analytical and simulation results are presented.

I. Introduction

A square-wave subcarrier can be downconverted by using a method such as the one presented in [1]. The downconversion procedure is illustrated in Fig. 1. However, the study in [1] was done under the following assumptions: (1) the filters are ideal, (2) the data signal is band limited, and (3) the power in the higher harmonics of the square wave is negligible. These assumptions are not practical. The relaxation of these conditions will quantitatively change the output results. A measurement of this change is the signal-to-noise ratio (SNR) degradation. This article presents a study of the SNR degradation in the square-wave subcarrier downconversion using realizable filters. A pseudorandom sequence, which modulates a square-wave subcarrier approximated by the first, third, and fifth harmonics, is used as the input data. The definition of SNR degradation used here is

$$\text{SNR degradation (dB)} = \text{SNR}_{\text{ideal}} - \text{SNR}_{\text{real}}$$

where $\text{SNR}_{\text{ideal}}$ and SNR_{real} are the signal-to-noise ratios in decibels at the output of the downconverter using ideal and realizable filters, respectively.

Two types of realizable filters will be considered here: the infinite-duration impulse response (IIR) filters and the finite-duration impulse response (FIR) filters. The IIR filters are easy to implement but do not provide linear phase. The output suffers distortion due to different group delays for different frequency components. The FIR filters, on the other hand, have linear phase but need a high filter tap number, thus a long processing time.

Since the noise is additive and the downconversion system is linear, the noise and the signal can be studied separately. Then the total SNR degradation can be obtained:

$$\text{total SNR degradation (dB)} = \text{signal power loss (dB)} + \text{noise power gain (dB)}$$

II. Noise Power Change

Unlike ideal filters, a realizable bandpass filter does not have a sharp frequency cutoff; therefore, some of the noise in the stop band may be passed. On the other hand, a

realizable filter does not have a flat magnitude in the pass band; so, some of the noise in this region will be weakened, assuming unity filter gain at the center frequency. Hence, the output noise power may change as compared to the ideal case. This change can be measured as the noise power gain in decibels, defined as follows:

$$\text{noise power gain} = 10 \log \frac{\text{noise power using real filter}}{\text{noise power using ideal filter}}$$

When the value of the noise power gain is negative, the implication is that the noise power is decreased, as compared with the ideal case.

The noise power after filtering can be computed as follows: Assuming additive white noise with a power spectral density of $S(\omega) = N_0/2$, the noise power theoretically becomes [2]

$$\text{noise power} = \frac{N_0}{2} \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 d\omega \qquad (1)$$

where $H(e^{j\omega})$ is the filter transfer function. In the down-conversion process, both bandpass and lowpass filters are involved, and each of them will be analyzed separately in the following subsections.

A. Effect of Bandpass Filtering

For an ideal bandpass filter (BPF), the noise power becomes

$$\text{noise power using ideal filter} = |H|^2 f_B/(f_s/2)$$

where $|H| = 1$ is the filter magnitude, f_B is the BPF bandwidth, and f_s is the sample rate.

For the IIR filter, the Butterworth bandpass filters are chosen as an example since they have a maximally flat magnitude. The integral can be computed through an equivalent lowpass filter (LPF) transfer function [3, p. 421]:

$$\int_{-\pi}^{\pi} |H(e^{j\omega})|^2 d\omega = \int_{-\pi}^{\pi} \frac{1}{1 + \left[\dfrac{\tan(\omega/2)}{\tan(\omega_c/2)}\right]^{2N}} d\omega$$

where ω_c is the cutoff angular frequency that takes values from 0 to π, π corresponds to half of the sample frequency, and N is the filter order.

If the FIR bandpass filters are used, the integrals can be evaluated by using Parseval's theorem [3, p. 187]:

$$\sum_{n=-\infty}^{\infty} |h[n]|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 d\omega$$

Since $h[n]$ has a finite duration, the sum is over a finite number of terms.

The IIR filters are obtained by converting the analog Butterworth filters, using bilinear transformation with frequency prewarping. The cutoff frequency is the half-power or 3-dB cutoff frequency. The FIR filters are designed using the classical method of windowed linear-phase FIR filter design. A Hamming window is used in this case.

By varying the IIR filter order and the cutoff frequency, a series of integrals is numerically evaluated, and the noise power gain in decibels is computed at a sample rate of 260 kHz as shown in Table 1. For comparison, a simulation of the bandpass filtering process is realized with white Gaussian noise as the input. For the output variance to be less than 0.02 dB, the simulation was run with 5×10^6 samples (see Appendix A) at a sample rate of 260 kHz. The results are also shown in Table 1.

A similar series of results is obtained for the FIR filters. The designed FIR filters, however, may not have the half-power cutoff frequency matching the goal cutoff frequency. For a fair comparison, the half-power cutoff frequencies are obtained. The results are shown in Table 2.

Taking the difference of the analytical and simulation results in Tables 1 and 2, it is found that the largest difference in decibels is about 0.03 dB.

B. Effect of Down-Mixing

Unlike the bandpass filtering, whose effect on the noise power can be computed analytically, the effect of the down-mixing process on the noise power can be studied more easily through simulation since the noise is no longer white. The BPF's are assumed to have the same order and bandwidth at different center frequencies. It is assumed that the lowpass filter has the same order as the bandpass filters and that its bandwidth is three times that of the BPF. The filter gain is assumed to be 1 at the center frequency. The simulation was done with 5×10^6 samples at a sample rate of 260 kHz for filter bandwidths less than 5 kHz, and at 270 kHz otherwise. The results are shown in Table 3.

III. Signal Power Change

There are three factors that cause the degradation in signal power: the cutoff of the data bandwidth, the finite number of harmonics considered, and the nonideal filtering. These are next discussed individually.

A. Data Bandwidth Cutoff

The bandwidth of the BPF should be greater than or equal to the data bandwidth if the data are band limited. However, this may not be satisfied in practice.

In the simulation, the signal consists of a square-wave subcarrier modulated by a 1-kHz pseudorandom sequence as data. A pseudorandom-noise (PN) shift register generator of length 10 associated with $\{3, 10\}$ is used to generate the data [4, p. 342]. The power spectral density of the PN sequence is [4, p. 380]

$$P_{PN}(f) = -\frac{1}{N}\delta(f) + \frac{N+1}{N^2}\text{sinc}^2(\pi f T_c)\sum_{m=-\infty}^{\infty}\delta\left(f+\frac{m}{NT_c}\right)$$

where N is the sequence length, and T_c is the chip time. Clearly, a PN sequence is not band limited. Hence, when the BPF's are applied to a data modulated square-wave subcarrier, some of the signal power carried by the higher frequency components is filtered out, which leads to a signal power degradation. In general, the wider the filter bandwidth is, the lower the signal power degradation is. This relationship is shown in Fig. 2. However, there are upper limits on the filter bandwidth so that the undesirable harmonic terms can easily be separated from the desired ones [1]. For instance, in the simulation, the restriction is $22.5\text{ kHz} - f_B/2 \geq 3f_B$, which implies that the filter bandwidth, f_B, should not exceed 6.43 kHz because aliasing may occur beyond this point. An alternative downconversion procedure (see Appendix B) may be used where no upper limit on the filter bandwidth will apply. However, the goal is to keep the bandwidth narrow for a low sample rate.

B. Finite Number of Harmonics Considered

Another factor that causes the signal power degradation is that only the first, third, and fifth harmonics are downconverted. The first three harmonics of a square-wave subcarrier carries only 93.3 percent of the total power of the square wave. So, there is about a 0.3-dB signal power degradation in neglecting the higher order harmonics. Table 4 shows the relationship between the power loss and the harmonic number up to which the downconversion is carried out.

C. Nonideal Filtering

Finally, the signal degradation, due to the nonideal filtering, is analyzed through simulation. To separate the nonideal filtering effect from the other two factors, the input signal is considered to have the first three harmonics only:

$$\text{subcarrier} = \frac{4}{\pi}\left[\sin(2\pi ft) + \frac{1}{3}\sin(2\pi \times 3ft) + \frac{1}{5}\sin(2\pi \times 5ft)\right]$$

where $f = 22.5$ kHz for the Galileo signal. The simulation results can be subtracted from the signal power degradation due to the data-bandwidth cutoff computed in Section III.A. The results obtained directly from the simulation are shown in Table 5 and Figs. 3 and 4. These results reflect the signal power degradation due to both data-bandwidth cutoff and nonideal filtering. The signal degradation due to only nonideal filtering is shown in Table 6 and Figs. 5 and 6. This simulation was done with 5×10^6 samples at a sample rate of 260 kHz for filter bandwidths less than 5 kHz, and at 270 kHz otherwise.

Elliptic lowpass and bandpass filters have been simulated as well. The results in signal power degradation are lower than those using the same order Butterworth filters, but the output is severely distorted.

IV. SNR Degradation

Finally, the total SNR degradation can be obtained by adding the noise power gain in Table 3 to the signal power loss in Table 5. The results are shown in Table 7 and Figs. 7 and 8.

Note that the results in Table 7 exclude the effect of the consideration of a finite number of harmonics for the square wave.

V. Conclusion

This article presented an analysis on the SNR degradation in the square-wave subcarrier downconversion pro-

cess, as may be used in the Galileo S-band mission. There are three factors that affect the SNR degradation: the data-bandwidth cutoff, the approximation of a square wave with a finite number of harmonics, and nonideal filtering. The three factors were analyzed separately, and the analytical and simulated results were presented. The distortion effects on the detection were not considered.

Acknowledgment

The authors thank Edgar Satorius for his many helpful suggestions and contributions to this work.

Table 1. Noise power gain (dB) due to IIR BPF.

Filter order	Filter bandwidth, kHz				
	2	3	4	5	6
6 (analytical)	0.1999	0.1993	0.1986	0.1977	0.1965
8 (analytical)	0.1120	0.1117	0.1114	0.1109	0.1104
10 (analytical)	0.0716	0.0714	0.0712	0.0709	0.0706
6 (simulated)	0.1760	0.1915	0.1899	0.2046	0.1957
8 (simulated)	0.1169	0.1004	0.1262	0.1108	0.1080
10 (simulated)	0.0926	0.0611	0.0785	0.1080	0.0732

Table 2. Noise power gain (dB) due to FIR BPF.

Filter order	Intended filter bandwidth, kHz				
	2	3	4	5	6
	Obtained filter bandwidth, kHz				
160	2.2820	2.5420	3.0620	3.7640	4.6740
320	1.5280	2.3340	3.3480	4.3360	5.3500
	Noise power gain, dB				
160 (analytical)	0.2447	0.2678	0.1513	0.0915	0.0336
320 (analytical)	0.1583	0.0386	0.0129	0.0155	0.0266
160 (simulated)	0.2304	0.2646	0.1303	0.0998	0.0123
320 (simulated)	0.1560	0.0739	−0.0012	−0.0032	0.0198

Table 3. Noise power gain (dB) in downconversion.

Filter order	Filter bandwidth, kHz				
	2	3	4	5	6
IIR 6	−0.0646	−0.0801	−0.0801	−0.0801	−0.0795
IIR 8	−0.0752	−0.0882	−0.0827	−0.0960	−0.0711
IIR 10	−0.0743	−0.0943	−0.0748	−0.0736	−0.0998
FIR 160	–	−0.1429	−0.1610	−0.1422	−0.1501
FIR 320	−0.1328	−0.1009	−0.1210	−0.0451	−0.0664

Table 4. Signal power loss due to finite number of harmonics.

Harmonic number	1st	3rd	5th	7th	9th
Power loss, dB	0.9121	0.4545	0.3009	0.2246	0.1719
Harmonic number	11th	13th	15th	17th	19th
Power Loss, dB	0.1489	0.1274	0.1113	0.0988	0.0888

Table 5. Signal power loss (dB) from simulation.

Filter order	Filter bandwidth, kHz				
	2	3	4	5	6
IIR 6	0.5069	0.3554	0.2639	0.2105	0.1780
IIR 8	0.4816	0.3422	0.2484	0.1999	0.1661
IIR 10	0.4708	0.3351	0.2403	0.1946	0.1604
FIR 160	–	0.5851	0.4807	0.3857	0.2772
FIR 320	1.0134	0.5556	0.3120	0.2253	0.1813

Table 6. Signal power degraded (dB) due to nonideal filtering.

Filter order	Filter bandwidth, kHz				
	2	3	4	5	6
IIR 6	0.0517	0.0413	0.0393	0.0284	0.0295
IIR 8	0.0318	0.0281	0.0238	0.0178	0.0176
IIR 10	0.0210	0.0210	0.0157	0.0125	0.0119
FIR 160	–	0.1653	0.1643	0.1366	0.0462
FIR 320	0.4936	0.1134	0.0464	0.0063	−0.0136

Table 7. Total SNR degradation (dB) in downconversion.

Filter order	Filter bandwidth, kHz				
	2	3	4	5	6
IIR 6	0.4413	0.2753	0.1837	0.1304	0.0985
IIR 8	0.4064	0.2541	0.1657	0.1039	0.0950
IIR 10	0.3964	0.2407	0.1655	0.1210	0.0606
FIR 160	–	0.4422	0.3197	0.2435	0.1271
FIR 320	0.8806	0.4547	0.1910	0.1802	0.1149

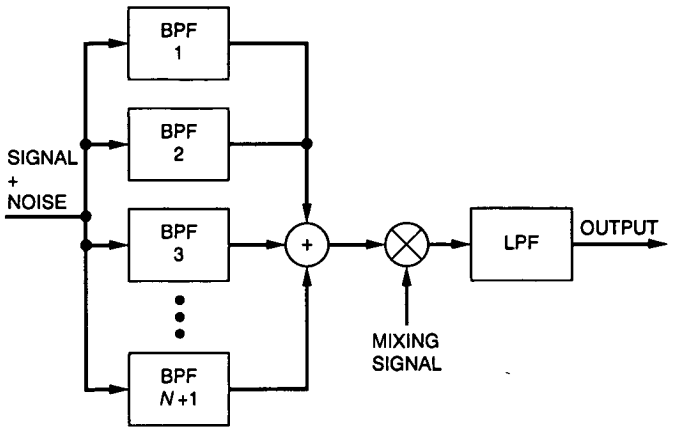


Fig. 1. Square-wave subcarrier down-conversion.

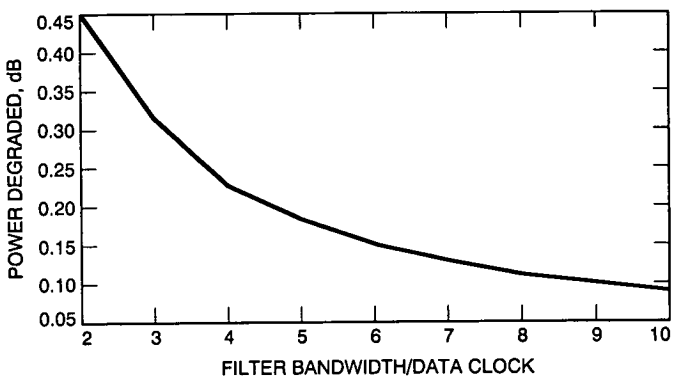


Fig. 2. Signal power degradation due to the PN bandwidth cutoff.

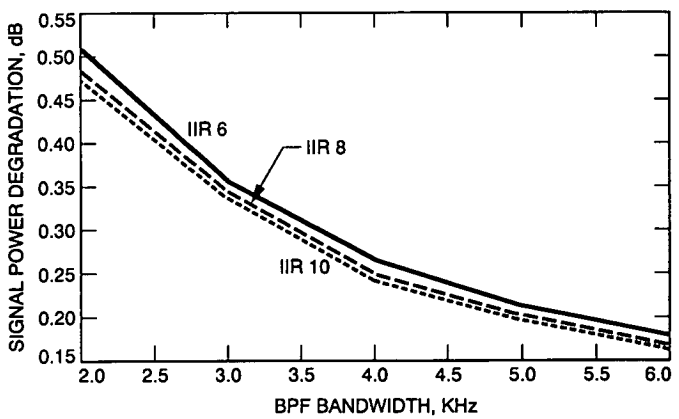


Fig. 3. Signal power degradation in downconversion using IIR filters.

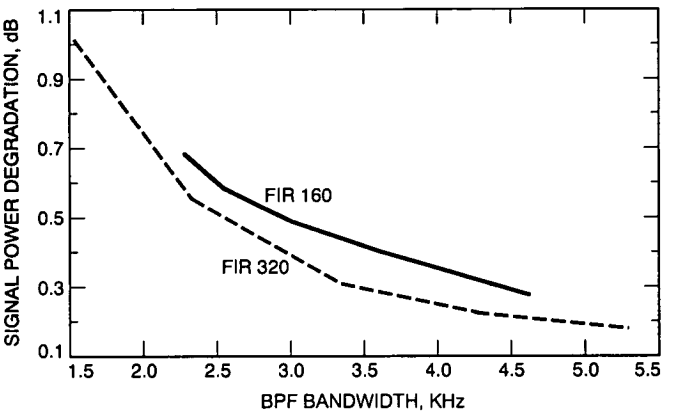


Fig. 4. Signal power degradation in downconversion using FIR filters.

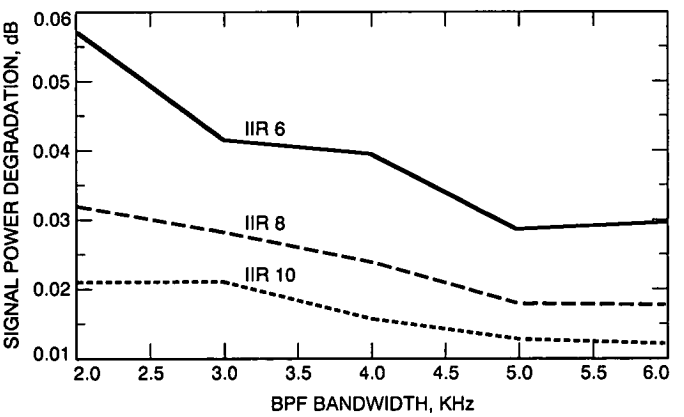


Fig. 5. Signal power degradation due to IIR filtering.

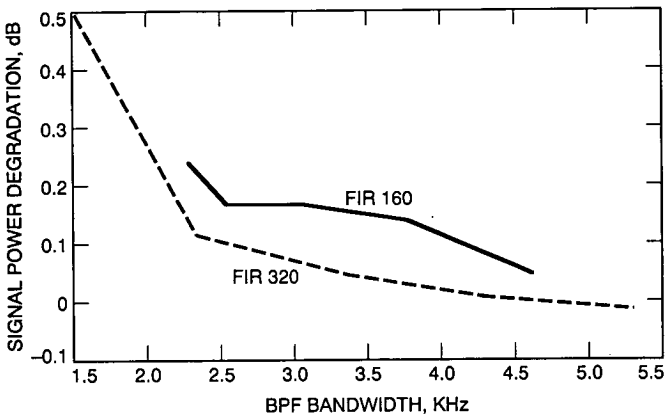


Fig. 6. Signal power degradation due to nonideal FIR filtering.

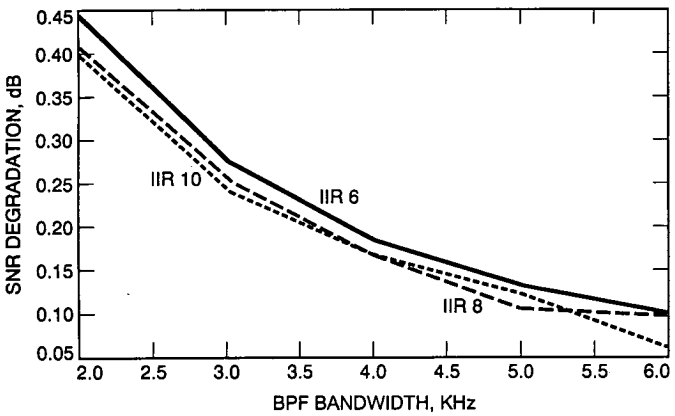


Fig. 7. SNR degradation in downconversion using IIR filters.

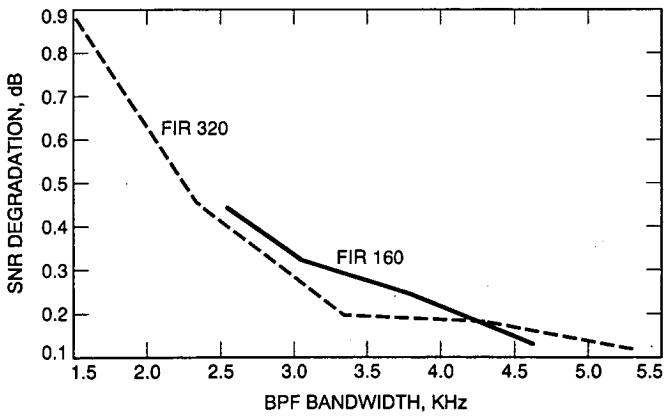


Fig. 8. SNR degradation in downconversion using FIR filters.

Appendix A

Number of Samples Needed in the Simulation

For N independent samples, x_i, of a random variable, x, the variance can be estimated by using an asymptotically unbiased and consistent estimate [5], namely,

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{M})^2 \qquad \text{(A-1)}$$

where $\hat{M}$ is the mean estimate

$$\hat{M} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The mean of the variance estimate is

$$E\{\hat{\sigma}^2\} = \frac{N-1}{N}\sigma^2$$

where σ^2 is the true variance, and the variance of the estimate is

$$\text{Var}\{\hat{\sigma}^2\} = \frac{1}{N}(\text{E}\{x^4\} - \text{E}^2\{x^2\}) \qquad \text{(A-2)}$$

Assuming that the noise has a normal distribution, $\text{E}\{x^4\} = 3\text{E}^2\{x^2\}$ [6]. Substituting the last expression in Eq. (A-2),

$$\text{Var}\{\hat{\sigma}^2\} = \frac{2}{N}(\sigma^2)^2 \qquad \text{(A-3)}$$

or

$$\text{standard deviation}\{\hat{\sigma}^2\} = \sqrt{\frac{2}{N}}\sigma^2$$

where N is the number of independent samples. In this case, at the BPF output, the number of independent samples reduces to approximately $N_{in}f_B/(f_s/2)$, with N_{in} being the number of independent samples at the input, f_B being the BPF bandwidth, and f_s being the sample rate. When $N_{in} = 5 \times 10^6$ and $f_s = 260$ kHz, the deviations (in decibels) of the estimated variance, $\hat{\sigma}^2$, from the true variance, σ^2, in terms of f_B are shown in Table A-1.

Table A-1. Deviation of the estimated variance (dB) at output of BPF.

BPF bandwidth (kHz)	2	3	4	5	6
10 log $(\hat{\sigma}^2/\sigma^2)$	0.0210	0.0180	0.0156	0.0140	0.0128

Appendix B

An Alternative Downconversion Procedure

An alternative procedure for the square-wave subcarrier downconversion is shown in Fig. B-1. This procedure differs from the one shown in Fig. 1 in the following ways: The undesirable terms will reside in a farther region in the frequency domain. For instance, assuming that the subcarrier frequency is smaller than the mixing signal frequency, $f_s < f_m$, then the lowest frequency of the undesirable terms in the procedure of Fig. 1 is

$$7f_s - 5f_m = 2f_s + 5(f_s - f_m)$$

whereas in the alternative procedure, this frequency is

$$f_s + f_m = 2f_s + (f_m - f_s)$$

The second term of the last two expressions are negative and positive, respectively. This fact makes the lowest frequency of the undesirable terms in the second case much larger than the first case, which is desired since this may lead to lesser interference from the undesirable terms when nonideal filters are realized. From the hardware perspective, this procedure may need four multipliers and three adders versus six adders and one multiplier in the procedure shown in Fig. 1.

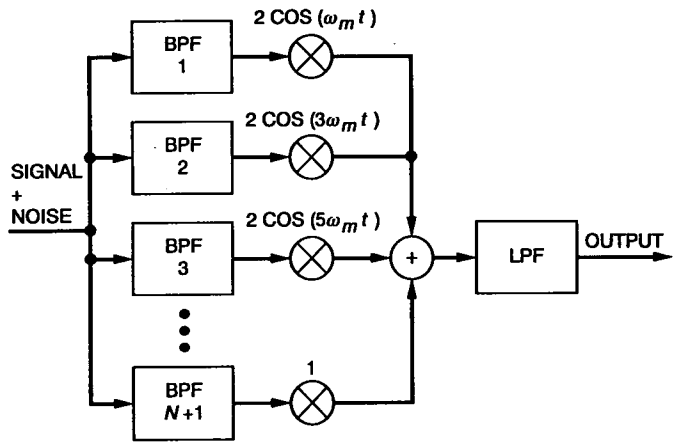


Fig. B-1. Alternative procedure for square-wave subcarrier downconversion.

References

[1] Y. Feria and J. Statman, "Bandwidth Compression of Noisy Signal With Square-Wave Subcarrier," *TDA Progress Report 42-109*, vol. January–March 1992, Jet Propulsion Laboratory, Pasadena, California, pp. 170–178, May 15, 1992.

[2] A. B. Carlson, *Communication Systems—An Introduction to Signals and Noise in Electrical Communication*, 3rd ed., New York: McGraw-Hill, pp. 174–176, 1986.

[3] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Englewood Cliffs, New Jersey: Prentice Hall, 1989.

[4] J. K. Holmes, *Coherent Spread Spectrum Systems*, Malabar, Florida: Robert E. Krieger Publishing Company, 1990.

[5] R. L. Fante, *Signal Analysis and Estimation*, New York: John Wiley and Sons, pp. 353–354, 1988.

[6] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, p. 148, 1965.

The Use of Subcarriers in Future DSN Missions

M. M. Shihabi, T. M. Nguyen, and S. M. Hinedi
Communications Systems Research Section

The performance of deep space telemetry signals that employ a residual carrier modulation technique is compared in the presence and absence of a subcarrier. When the subcarrier is present, the performance for the resulting pulse-coded modulation/phase-shift keyed/phase-modulated (PCM/PSK/PM) scheme is evaluated for both sine wave and square wave subcarriers and non-return-to zero (NRZ) data. When the subcarrier is absent, the performance for the resulting PCM/PM technique is evaluated for both the NRZ and the bi-phase data format. The comparison is based on telemetry performance as well as bandwidth efficiency. The first criterion is characterized in terms of the symbol error rate (SER) as a function of symbol SNR, loop bandwidth-to-data rate ratio, and modulation index. The bandwidth efficiency is characterized by the occupancy factor. The results of both the analysis and measurements show that when the interference-to-carrier ratio (ICR) is less than −15 dB, the performance degradation in the absence of a subcarrier is negligible. Various combinations of loop bandwidth-to-data rate ratios and modulation indices that achieve this performance are derived and listed. Bandwidth occupancy comparison indicates that PCM/PM/NRZ is the most efficient in this regard. Therefore, by eliminating the subcarrier and using the PCM/PM/NRZ scheme, many advantages can be realized without any sacrifice in performance.

I. Introduction

In the past when the DSN operated at low data rates using residual carrier modulation, it was necessary to separate the data from the residual carrier to avoid interference. This was achieved by placing the data modulation on a subcarrier since direct modulation of the data on the carrier would cause most of the data power to fall within the bandwidth of the carrier phase-lock loop (PLL) and, as a consequence, interfere with its operation [1, Chap. 3]. The scheme in which the data is phase-shift keyed (PSK) onto a subcarrier and then phase-modulated (PM) onto a sinusoidal carrier is called the PCM/PSK/PM modulation scheme. However, the DSN has evolved over the years and will soon be capable of supporting very high data rates, on the order of tens of megabits per second [1, Chap. 3].

At these higher data rates, the data signal spectrum is broad, and therefore, if the subcarrier were to be eliminated, the part of the spectrum that would fall within the carrier loop bandwidth would be flat and appear as white noise to a narrow loop. Since the ratio of loop bandwidth to data rate is very small at high data rates, it would seem that this additional white noise component would degrade the tracking performance very little. This later scheme in which the telemetry data are phase modulated directly onto the carrier is known as the PCM/PM modulation scheme.

In this article, the use of residual carrier techniques with no subcarrier (PCM/PM) will be advocated. It will be shown that the use of the PCM/PM technique under appropriate conditions will reap many advantages without sacrificing performance. Foremost among these advantages is the lower bandwidth, which can be extremely useful when antenna arraying is employed or when tracking multiple spacecraft signals within the field of view of a single antenna. In the latter case, the various signals can be separated in the frequency domain and a single receiver front end can be used to downconvert them to baseband. This idea is very similar to frequency division multiplexing (FDM) except that the carriers are generated by different oscillators. Secondly, PCM/PM has the advantage of eliminating subcarrier loss in the receiver. Finally, the presence of the residual carrier provides useful data for navigation purposes and also safeguards the investment made in PLL receivers by the DSN and other space agencies. In light of the above, the goal of this article is the assessment and comparison of performance degradation in DSN receivers due to the interference between the data and the residual carrier in the above two schemes. For the PCM/PM modulation scheme, two kinds of data formats are considered. The non-return-to-zero (NRZ) format and the Manchester (or bi-phase) data format. For the PCM/PSK/PM modulation scheme, the performance is determined for the two most commonly used subcarriers: the square wave, which is used for deep space missions (category B) and the sine wave, which is used for near-Earth missions (category A), as recommended by the Consultative Committee for Space Data Systems (CCSDS) [2]. The comparison is based on the telemetry symbol error rate (SER) and the bandwidth (BW) occupancy. Throughout this article, it is assumed that the data are not encoded. Hence, the SER is equivalent to the bit error rate (BER), and the two terms are used interchangeably.

The article is organized as follows: A brief description of the various residual carrier modulation schemes is presented in Section II. The general analysis and summary of the theoretical results are presented in Section III. The discussion and comparison of theoretical and measured performance are carried out in Section IV, followed by the conclusion in Section V.

II. Description of the Residual Carrier Modulation Techniques

The telemetry signal in the two schemes can be represented mathematically by

$$s(t) = \sqrt{2P_T}\sin\left(\omega_c t + mP(t)d(t)\right) \tag{1}$$

where P_T is the total power; ω_c is the angular carrier frequency in rad/sec; m is the modulation index in radians ($0 < m < \pi/2$); $d(t)$ is the binary data sequence with symbol rate $R_s = 1/T_s$; and

$$P(t) = \begin{cases} 1 & ; \text{PCM/PM} \\ \mathrm{Sin}\left(\omega_{sc}t + \theta_{sc}\right) & ; \text{PCM/PSK/PM (square wave)} \\ \sin\left(\omega_{sc}t + \theta_{sc}\right) & ; \text{PCM/PSK/PM (sine wave)} \end{cases}$$

A. The PCM/PM Technique

There is no subcarrier used in this scheme, hence one substitutes 1 for $P(t)$ in Eq. (1), obtaining

$$s(t) = \sqrt{2P_T}\sin\left(\omega_c t + md(t)\right) \tag{2}$$

The received signal, $r(t)$, is corrupted by additive white Gaussian noise (AWGN), $n(t)$, with one-sided power spectral density N_0(W/Hz). By using simple trigonometric identities, the received signal can be expanded as

$$r(t) = \sqrt{2P_T}\left[\cos\left(m\right)\sin\left(\omega_c t + \theta_c\right) + d(t)\sin\left(m\right)\cos\left(\omega_c t + \theta_c\right)\right] + n(t) \tag{3}$$

where θ_c is the carrier phase. The first and second terms of Eq. (3) are the residual carrier components, $C(t)$, and the data components, $I_c(t)$, respectively. Explicitly,

$$C(t) = \sqrt{2P_T}\cos(m)\sin(\omega_c t + \theta_c) \quad (4a)$$

$$I_c(t) = \sqrt{2P_T}d(t)\sin(m)\cos(\omega_c t + \theta_c) \quad (4b)$$

The data component is denoted by $I_c(t)$ to explicitly indicate its role as the interfering component to the carrier PLL. It is seen that the modulation index, m, has allocated the total transmitted power to the carrier and to the data channel, where the carrier power and data power are respectively given by

$$P_C = P_T \cos^2 m \quad (5a)$$

$$P_D = P_T \sin^2 m \quad (5b)$$

Another important parameter is the power spectral density (PSD) of the transmitted signal. From Eq. (2), it is easy to determine that the one-sided PSD of the transmitted signal in the PCM/PM scheme is equal to

$$S(f) = P_T\left[\cos^2(m)\delta(f - f_c) + \sin^2(m)S_D(f - f_c)\right] \quad (6)$$

where $S_D(f)$ is the PSD of the data sequence and is given by [1,4]

$$S_D(f) = \begin{cases} 1/R_s \dfrac{\sin^2(\pi f/R_s)}{(\pi f/R_s)^2}; & \text{NRZ data} \\ 1/R_s \dfrac{\sin^4(\pi f/2R_s)}{(\pi f/2R_s)^2}; & \text{bi-phase data} \end{cases} \quad (7)$$

By substituting in Eq. (6), one obtains expressions for the spectrum of the transmitted signal for these two cases, as shown in Fig. 1(a) and 1(b), respectively. Because of the symmetrical nature of the spectrum around the origin, only the positive frequency portion is shown in the graph.

B. The PCM/PSK/PM Technique

This scheme is the traditional residual carrier technique, where a subcarrier is used to separate the data from the carrier. The telemetry signal is given in Eq. (1) and results in

$$s(t) = \sqrt{2P_T}\left[\cos(mP(t))\sin(\omega_c t) + d(t)\sin(mP(t))\cos(\omega_c t)\right] \quad (8a)$$

where $d(t)$ is the NRZ binary data sequence, and $P(t)$ is the subcarrier waveform. Using Eq. (8), an expression for the PSD of the transmitted signal was derived in [5] for the two subcarrier waveforms used in space applications. When $P(t)$ is a unit-power square-wave subcarrier of frequency f_{sc}, Eq. (8a) reduces to

$$s(t) = \sqrt{2P_T}[\cos(m)\sin(\omega_c t) + d(t)P(t)\sin(m)\cos(\omega_c t)] \quad (8b)$$

The power spectral density for this case is given by

$$S(f) = P_T\left\{\cos^2(m)\delta(f - f_c) + \left(\frac{4}{\pi^2}\right)\sin^2(m)\sum_{k\geq 1}\left[\left[S_D(f - f_c - (2k-1)f_{sc}) + S_D(f - f_c + (2k-1)f_{sc})\right]/(2K-1)^2\right]\right\} \quad (9)$$

where $S_D(f)$ is the PSD of the NRZ binary data sequence which was defined earlier in Eq. (7). The first term in Eq. (9) is the residual carrier spectral component, and the second term is the data component. The plot of Eq. (9) is shown in Fig. 1(c). The CCSDS recommends that the subcarrier frequency-to-bit rate, $n = f_{sc}/R_s$, be an integer [2] where R_s is the data rate. On the other hand, when $P(t)$ is a sine wave subcarrier, the power spectral density of the telemetry signal is given by

$$S(f) = P_T\left[J_0^2(m)\delta(f - f_c) + \sum_{i \text{ even}} J_i^2(m)\left[\delta(f - f_c - if_{sc}) + \delta(f - f_c + if_{sc})\right] + \sum_{k \text{ odd}} J_k^2(m)\left[S_D(f - f_c - kf_{sc}) + S_D(f - f_c + kf_{sc})\right]\right] \quad (10)$$

where $J_k(\cdot)$ is the kth-order Bessel function. The first term in Eq. (10) is the residual carrier spectral component, the second term is the intermodulation loss component, and the third term is the data component.

III. Performance Analysis

On the ground, the carrier component, $C(t)$, of the received signal, $s(t)$, is tracked by the PLL. The tracking performance of the PLL depends on the modulation index m and the tracking loop noise bandwidth-to-data rate ratio (B_L/R_S). Figures 1(a) and 1(b) illustrate the spectrum of the PCM/PM signals and the portion (shaded area) of the data spectrum that lies within the carrier loop bandwidth B_L. From these figures, it is clear that the interference power is a function of both B_L and R_S. The performance of the carrier tracking loop in the presence of the data is characterized by the interference-to-carrier signal power ratio (ICR). This ratio is determined for all schemes, and the effect of this interference on the phase error process is assessed. Consequently, the SER is calculated for all cases. In order to proceed, a few parameters must be identified. First, the one-sided loop bandwidth of the PLL is defined in terms of the carrier tracking loop transfer function, $H(j2\pi f)$, as follows:

$$B_L = \int_0^\infty |H(j2\pi f)|^2 \, df \tag{11}$$

Current DSN receivers typically use passive second-order PLL filters resulting in [5]

$$|H(f)|^2 = \frac{1 + 2(f/f_n)^2}{1 + (f/f_n)^4} \tag{12}$$

where f_n is the loop natural frequency and is related to the one-sided loop noise bandwidth B_L through

$$B_L = \pi f_n \left(\xi + \frac{1}{4\xi}\right) \tag{13}$$

with ξ denoting the damping factor (typically $\xi = 0.707$ is used). Next, the ICR is determined for the various schemes.

A. Determination of the ICR

From Eq. (1) and Fig. (1), the interference power is that portion of the data spectrum that falls within the carrier loop bandwidth and interferes with the carrier power, as given by Eq. (5). Thus, the ICR becomes

$$ICR = \tan^2(m) \int_0^\infty |H(j2\pi f)|^2 S_D(f) \, df \tag{14}$$

Depending on the data format and rate, Eq. (14) can be simplified further.

1. PCM/PM with Ideal NRZ data Format. By substituting Eqs. (7), (12), and (13) into Eq. (14) and evaluating the integral, one obtains

$$ICR = \tan^2(m)\left[\frac{1}{2} + \frac{1}{4\gamma\sqrt{2}}\left(1 - e^{-\gamma\sqrt{2}}\left(\cos\left(\gamma\sqrt{2}\right) + 3\sin\left(\gamma\sqrt{2}\right)\right)\right)\right] \tag{15}$$

where

$$\gamma \triangleq \frac{B_L}{R_S} \tag{16}$$

The upper and lower bound for the ICR can be found from Eq. (15). If one lets γ approach zero (i.e., the data rate R_S approaches infinity) and by using L'Hospital's rule, one gets the lower bound

$$ICR = 0 \tag{17}$$

On the other hand, if one lets γ approach infinity (R_S approaches zero), then one obtains the upper bound

$$ICR = \frac{\tan^2(m)}{2} \tag{18}$$

From these bounds, it is clear that as the loop bandwidth-to-data rate ratio, γ, increases, the ICR increases. These results are obvious from Figs. 1(a) and (b), because as B_L increases, the shaded area representing the interference due to the data increases while the residual carrier power P_c remains unchanged. Also from Eq. (15), one notices that for a fixed γ, the ICR increases as m increases, since for a fixed total power P_T, P_c decreases as m increases (remember $0 < m < \pi/2$) while P_D increases. Using Eq. (15), the ICR is plotted in Fig. 2(a) as a function

of γ for different values of m. The appropriate values of γ can be determined for a given modulation index m so that the ICR does not exceed the maximum allowable value. For a deep space mission, the maximum allowable interference for the carrier tracking is $(ICR)_{\max} = -15$ dB, as recommended by the International Radio Consultative Committee (CCIR) [3]. Another important parameter is the critical value of γ (for a given m) that will cause the ICR to reach its maximum allowable value. These critical values for γ as a function of m are plotted in Fig. 2(b). The region on this graph corresponding to values of γ that yield an ICR ratio below the maximum allowable value is called the operating region (OR). From Fig. 2(b), one can see that γ decreases as m increases in order to keep the ICR fixed.

For the NRZ data format case, determining the performance is of interest in the following two special cases. For the high data rate case, the data power spectrum that falls into the carrier tracking loop bandwidth is essentially constant over the tracking loop bandwidth. Thus, for $\gamma < 0.1$, the interference can be considered as additional additive white noise and Eq. (14) can be approximated by

$$ICR \approx \tan^2(m) S_D(0) \int_0^\infty |H(f)|^2 \, df \qquad (19)$$

or

$$ICR \approx \left(\frac{B_L}{R_S}\right) \tan^2(m) \quad \text{when} \quad \frac{B_L}{R_S} < 0.1 \qquad (20)$$

For the low data rate case, all the power of the interference component falls within the carrier tracking loop bandwidth. Thus, the ICR can be approximated by

$$ICR \approx \tan^2(m) \frac{1}{R_S} \int_0^{B_L} \frac{\sin^2(\pi f/R_S)}{(\pi f/R_S)^2} \, df \qquad (21)$$

When $B_L/R_S > 10$, the above integral is easily evaluated:

$$ICR \approx \left(\frac{1}{2}\right) \tan^2(m) \qquad (22)$$

In practice, the above case is possible when a data rate of 32 bps is transmitted and a wide carrier tracking mode at $2B_L = 300.3$ Hz is selected by the ground station receiver.

2. PCM/PM With Bi-Phase Data Format. By substituting Eqs. (7) for the bi-phase spectrum and Eq. (12) into Eq. (14) and evaluating the integral, one obtains

$$ICR = \tan^2(m) \left[\frac{1}{2} + \frac{9}{16\gamma} - \frac{3}{4\gamma} e^{-2\gamma/3} \times \left(\cos\left(\frac{2\gamma}{3}\right) + 3\sin\left(\frac{2\gamma}{3}\right)\right) + \frac{3}{16\gamma} e^{-4\gamma/3} \times \left(\cos\left(\frac{4\gamma}{3}\right) + 3\sin\left(\frac{4\gamma}{3}\right)\right)\right] \qquad (23)$$

The above ICR is plotted in Fig. 3(a) as a function of γ for various values of m. The corresponding operating region for the ratio γ is plotted in Fig. 3(b).

3. PCM/PSK/PM. This scheme is used to separate the residual carrier and the data components so that the interference caused by the data to the residual carrier is minimized. The higher the subcarrier-frequency-to-data rate ratio is, the less the interference. The ICR will be derived for the case of the square wave subcarrier, and the results can be extended to the sine wave case very easily. The data component spectrum can be extracted from the transmitted signal spectrum, as given by Eq. (9), and is equal to

$$S_{\text{Data}}(f) = \left(\frac{4}{\pi^2}\right) \sum_{k \geq 1} \Big[[S_D(f - f_c - (2k-1)f_{sc}) + S_D(f - f_c + (2k-1)f_{sc})]/(2k-1)^2 \Big] \qquad (24)$$

where $S_D(f)$ is given in Eq. (7), and $f_{sc} = n \times R_s$. By substituting Eqs. (13) and (24) in Eq. (14) and evaluating the integral numerically for the case when $f_{sc} = 3 \times R_s (n = 3)$, one obtains values for the ICR as a function of γ when the modulation index, m, is chosen to have the values 0.9 and 1.3 rad, respectively. The results are plotted in Fig. 4 where it is found that when $m = 0.9$ rad, the ICR never reaches the maximum allowable value when $\gamma \leq 1$. However, a critical value of $(\gamma)_{CR} = 0.56$ is found when the modulation index is increased to $m = 1.3$ rad. In Fig. 4(a), it is shown that the ICR decreases rapidly for the two modulation indices when $\gamma \leq 0.2$. The above procedure was repeated for $n \geq 3$ and it was found that the interference is negligible. Note that for the PCM/PSK/PM scheme,

the critical value for γ is quite large even at the largest modulation index chosen, $m = 1.3$ rad. The same cannot be said about the PCM/PM scheme where one must operate at values of γ much below 1, in order to keep the ICR below -15 dB.

B. The Effect of the Interference on Carrier Tracking Phase Error

When the received signal is tracked by a PLL, the phase error process, defined as

$$\phi(t) = \theta(t) - \hat{\theta}(t) \tag{25}$$

has contributions from both the thermal noise and the interference caused by the data. The accuracy of the estimate $\hat{\theta}(t)$ is dependent on the power allocated to the residual carrier component. Under the high data rate assumption, the data interference component has a spectrum over the PLL bandwidth, and hence it can be considered as an additional additive white noise. In this case, the variance of the phase error due to the combined additive Gaussian noise is given by [6]

$$\sigma^2 = \frac{NB_L}{P_c} \tag{26}$$

where N is the effective noise spectral density resulting from thermal noise and data interference, and is determined as follows:

$$N = \frac{1}{B_L} \int_0^\infty |H(j2\pi f)|^2 \left[N_0 + P_D S_D(f)\right] \, df \tag{27}$$

Alternatively, one has

$$N = N_0 + \frac{P_D}{B_L}\alpha \tag{28}$$

where

$$\alpha \triangleq \int_0^\infty S_D(f)|H(j2\pi f)|^2 \, df \tag{29}$$

Rather than expressing the variance of the phase error in terms of the modified noise spectral density N, it will be expressed in terms of the ICR to show the interdependence explicitly. Thus, under the high data rate assumption, the variance of the carrier tracking phase error becomes [7]

$$\sigma^2 = \frac{1}{\rho} = \frac{1}{\rho_0} + ICR \tag{30}$$

where ρ is the effective loop SNR, and ρ_0 is the carrier loop signal-to-thermal-noise power ratio defined as

$$\rho_0 = \frac{P_C}{N_0 B_L} = \frac{E_S/N_0}{(B_L/R_S) \times \tan^2(m)} \tag{31}$$

Under the linear operation assumption, the probability density function (pdf) of the phase error process may be approximated by the Tikhonov pdf [1]. Hence the pdf of $\phi(t)$ is completely determined, once ρ is known, from

$$p(\phi) = \frac{\exp(\rho \cos\phi)}{2\pi I_0(\rho)} \text{ when } -\pi \le \phi \le \pi \tag{32}$$

where $I_0(\cdot)$ is the modified Bessel function of the first kind, and ρ is the effective loop SNR, which is the inverse of the variance σ^2 of the carrier tracking phase error. Now, this variance for all cases must be found.

1. PCM/PM With Ideal NRZ Data Format. For the NRZ data format, the performance will be determined for the low-data-rate case as well as the high-data-rate case.

a. High-Data-Rate Case ($B_L/R_S < 0.1$). The expression for the ICR for this case is found in Eq. (20), which when substituted in Eq. (30), gives the phase error variance

$$\sigma^2 = \frac{1}{\rho} = \frac{1}{\rho_0} + \left(\frac{B_L}{R_S}\right) \tan^2(m) \tag{33}$$

The inverse of the phase error variance, i.e., the effective loop SNR as a function of symbol SNR (SSNR) is plotted in Figs. 5(a) and 5(b) for this case. The region corresponding to values of $\rho < 8$ dB is called the PLL nonoperating region (PLLNOR). This is the region where the loop is slipping cycles and should be avoided.

b. Very Low-Data-Rate Case ($B_L/R_S > 10$). Here one cannot treat the data interference as just additional white noise, but rather the procedure chosen by Youn and Lindsey [8] will be followed. When the data rate is very low, the carrier interference component, $I_C(t)$, can be treated as a CW interference with its phase switching slowly and randomly between 0 and π with respect to the phase of the residual carrier. Thus, if one lets $\Delta\theta$ be the phase

difference between the interference component $I_C(t)$ and the carrier component $C(t)$, then within the tracking loop bandwidth, $\Delta\theta$ will take on the values 0 and π with equal probability. The Fokker–Plank method can be used to derive the pdf for the carrier tracking phase error. Let ϕ denote the carrier tracking phase error induced by $I_C(t)$ and the AWGN, $n(t)$. Then the conditional pdf, $p(\phi|\Delta\theta)$, in the presence of CW interference can be approximated as

$$P_W(\phi|\Delta\theta) \cong \left[\frac{\exp\left[-(\phi - M)^2/2\sigma^2\right]}{\sqrt{2\pi\sigma^2}\ \mathrm{erf}\left(\pi/\sqrt{2\sigma^2}\right)}\right] \text{ for } |\phi| \leq \pi \tag{34}$$

where M and σ^2 are the conditional mean and variance of the phase error ϕ, respectively, and they are defined as

$$M = -\frac{\sqrt{ICR}\ \sin(\Delta\theta)}{1 + \left(\sqrt{ICR}\ \cos(\Delta\theta)\right)} \tag{35}$$

$$\sigma^2 = \frac{1}{\rho_0\left[1 + \sqrt{ICR}\ \cos(\Delta\theta)\right)} \tag{36}$$

By using Bayes' rule, the conditioning on $\Delta\theta$ can be removed to obtain the pdf of the phase error process:

$$p(\phi) = \frac{1}{2}\left[\frac{\exp\left[-\phi^2/2\sigma_+^2\right]}{\sqrt{2\pi\sigma_+^2}\ \mathrm{erf}\left(\pi/\sqrt{2\sigma_+^2}\right)} + \frac{\exp\left[-\phi^2/2\sigma_-^2\right]}{\sqrt{2\pi\sigma_-^2}\ \mathrm{erf}\left(\pi/\sqrt{2\sigma_-^2}\right)}\right] \tag{37}$$

where

$$\sigma_\pm^2 = \frac{1}{\rho_0\left(1 \pm \sqrt{ICR}\right)} \tag{38}$$

The above approximation is valid only when the ICR < 1, which is the case of interest here. The effective loop SNR is depicted in Fig. (6) as a function of the SSNR when $\gamma = 10$. The above pdf will be used to derive the SER for the low data rate case instead of the Tikhonov pdf that is used in the high data rate case.

2. PCM/PM With Ideal Bi-Phase Data Format. Under the high data rate assumption, the variance of the phase error process is given earlier in Eq. (30). By substituting the expression for the ICR derived in Eq. (23), the variance is determined, and consequently the Tikhonov pdf is now completely characterized. As mentioned before, in order to specify the PLL operating region, the inverse of the phase error variance is plotted as a function of SSNR in Figs. 7(a) and 7(b) for different modulation indices.

3. PCM/PSK/PM. By substituting the numerical values for the ICR obtained in Section (III.A.3) when $m = 1.3$ rad and the ratio $B_L/R_S = 0.05$ in Eq. (30), one obtains the phase error variance for the square-wave subcarrier. The RMS of this variance is plotted as a function of loop SNR, ρ_0, in Fig. 8 for $n = 1$ and $n = 3$ where n is the subcarrier frequency to bit rate, as defined earlier. The figure indicates that if one chooses the integer $n \geq 3$, then the performance of this scheme is almost identical to that of the ideal case. The ideal case refers to the binary phase-shift keying (BPSK) modulation scheme over an AWGN channel.

C. Effect of Interference on the SER Performance

The symbol error probability for uncoded PSK transmission over a Gaussian channel disturbed by additive white noise of one-sided spectral thermal noise density N_0 is given by [1, Chap. 5]

$$P_S = \frac{1}{2}\ \mathrm{erfc}\left(\sqrt{\frac{E_S}{N_0}}\right) \tag{39}$$

The conditional symbol error probability that takes into account the phase error process is given by [6]

$$P_S(\phi) = \frac{1}{2}\ \mathrm{erfc}\left(\sqrt{\frac{E_S}{N_0}}\ Y(\phi)\right) \tag{40}$$

where

$$Y(\phi) = \frac{1}{T_S}\int_0^{T_S} \cos[\phi(t)]\ dt \tag{41}$$

Then the unconditional symbol error probability can be obtained by averaging $P_S(\phi)$ over ϕ, i.e.,

$$P_S = \int_{-\pi}^{\pi} P_S(\phi)\, p(\phi) d\phi \qquad (42)$$

1. PCM/PM With Ideal NRZ Data Format

a. High-Data-Rate Case. When the data rate is high with respect to the receiver tracking loop bandwidth (i.e., $\gamma < 0.1$), the phase error process $\phi(t)$ varies slowly and is essentially constant over a symbol interval T_S. Then from Eq. (41), it is concluded that for this case

$$Y(\phi) \approx \cos\phi \qquad (43)$$

and the average symbol error probability is obtained by substituting Eq. (43) in Eq. (42) to give

$$P_S = \frac{1}{2}\int_{-\pi}^{\pi} \operatorname{erfc}\left(\sqrt{\frac{E_S}{N_0}}\cos\phi\right) \frac{\exp(\rho\cos\phi)}{2\pi I_0(\rho)} d\phi \qquad (44)$$

This SER, P_S, is plotted in Fig. 9(a) as a function of SSNR (E_S/N_0) for different values of m when $\gamma = 0.005$. From this figure, one sees that when $m = 1.3$ rad, the degradation becomes considerable. Figure 9(b) depicts that even when $m = 1.3$ rad, the performance can be improved by decreasing γ, i.e., by either increasing the data rate or by decreasing the loop noise bandwidth. For Figs. 9(a) and 9(b) to be useful, they must be used in the operating region of the PLL, as given earlier in Figs. 5(a) and 5(b). For example, when $\gamma = 0.01$ and $m = 1.3$ rad, Fig. 5(b) indicates that the PLL locks, provided that the SSNR is at least 8 dB. Therefore, when using the performance curve corresponding to the above operating conditions [Fig. 9(b)], only the region where the SSNR $\geq$ 8 dB is useful.

b. Very Low-Data-Rate Case. When the data rate is low relative to the receiver tracking loop bandwidth, the phase process $\phi(t)$ varies rapidly over the symbol interval T_S. Hence, the random variable $Y(\phi)$, as given in Eq. (41), is a good approximation of the true time average of the function $\cos[\phi(t)]$. When $\phi(t)$ can be modeled as an ergodic process, the time average may be replaced by the statistical mean. Thus,

$$Y(\phi) \simeq E\{\cos\phi\} = \int_{-\pi}^{\pi} \cos\phi \times p(\phi) d\phi = C_0 \qquad (45)$$

The constant C_0 can be computed by substituting the expression for the pdf of the phase error derived from Eq. (37) in Eq. (45). Then, by substituting in Eqs. (41) and (43) one obtains the average symbol error probability

$$P_S = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{E_S}{N_0}}\, C_0\right) \qquad (46)$$

In Fig. 10, P_S is plotted as a function of SSNR for different values of m when $\gamma = 10$. The figure shows that the degradation is substantial and becomes worse as m increases. It should be noted that the approximation used is valid when $ICR < 1$, which corresponds to $m \leq 0.955$ rad. However, Fig. 6 indicates that in order for the PLL to maintain lock, one must operate at an SSNR higher than 20 dB, even for the smallest modulation index chosen (0.7 rad), an impractical requirement to say the least. Therefore, the PCM/PM modulation technique with the NRZ data format is not a viable choice when $B_L \geq R_S$.

2. PCM/PM With Bi-Phase Data Format. Assuming that the data rate is high, i.e., $\gamma < 0.1$, the random variable $Y(\phi)$ is approximated as in Eq. (43). By substituting the result and the appropriate Tikhonov pdf for this case, as derived in Eq. (42), one obtains an expression for the SER. This average SER is calculated as a function of the SSNR E_S/N_0 for different values of m when $\gamma = 0.05$, and the results are plotted in Fig. 11(a). One sees that a noticeable degradation is observed only for the case when $m = 1.3$ rad. The effect of varying γ is shown in Fig. 11(b) for $m = 1.3$ rad. When $\gamma = 0.005$, the performance of the loop SNR approaches that of the ideal case.

3. PCM/PSK/PM. By substituting the calculated phase error variance obtained in Section III.B.3 and Eq. (42), it is found that when the integer $n \geq 3$, the performance approaches that of the ideal case. Whereas when $n = 1$, the performance is identical to the PCM/PM scheme with the bi-phase data format case, as it should be.

IV. Discussion and Summary of the Results

A. Theoretical Results

The critical values of γ for a given modulation index for which the ICR exceeds the CCIR recommended threshold value of -15 dB are obtained. These values are given in Table 1 for all the schemes when $m = 0.9$ and 1.3 rad.

From Table 1, one notes that for a given modulation index, the PCM/PM scheme with the NRZ data format requires a value of γ smaller than the other two schemes to operate at the threshold interference level. The

PCM/PSK/PM scheme performs just as well at a much larger γ ratio when $n = 3$ or higher.

The SER and the effective loop SNR of the PCM/PM scheme with both NRZ and bi-phase data formats, as well as the ideal case, are superimposed in Figs. 12(a) and 12(b) when $m = 1.3$ rad for different values of γ, in order to compare their performance. The criterion is the degradation in SSNR for the particular scheme considered relative to the ideal case, i.e., the required increase in SSNR in order to achieve the same performance as the ideal case. This degradation is due to both the thermal noise and the data interference. The SSNR degradation of the three schemes considered is obtained for different values of m and γ from the figures and is summarized in Tables 2–5.

In Table 2, the chosen value for γ is very small and is below the critical value (see Table 1) for all the schemes. One notes a very small degradation for the PCM/PM/NRZ scheme and negligible degradation for the other two schemes. Table 3 shows the effects of increasing γ.

In Table 4, the dependence of the performance on the modulation index is demonstrated. The loop bandwidth-to-data rate ratio is unchanged from the previous table, but the modulation index is smaller. It is observed that the performance of the PCM/PM/NRZ scheme improves as the modulation index decreases.

In Table 5, both PCM/PM schemes are in the not-recommended region (NRR) (see Table 1). The two PCM/PM schemes can be compared in terms of the effective loop SNR for the same values of SSNR, m and γ. The scheme with the higher effective loop SNR performs better. Using Figs. 5(b) and 7(b), the values are tabulated in Table 6. It is observed that for the same operating parameters, the PCM/PM/bi-phase scheme always has a higher effective loop SNR than the PCM/PM/NRZ scheme.

The unfiltered occupied bandwidth with 99-percent power containment as a function of the modulation index for both PCM/PSK/PM and PCM/PM signals was derived in [9]. The results are summarized here in Fig. 13. As expected, it is shown that the PCM/PM scheme with the NRZ data format requires the least occupied bandwidth and the PCM/PSK/PM scheme requires the most. The bandwidth occupancy increases linearly with the modulation index for all the schemes.

B. Measured Performance

The measured SER values for all schemes were obtained by using the Advanced Receiver (ARX) II. A functional description of the ARX II is presented in [10]. Figures 14(a), 14(b) and 14(c) depict the measured performance for the PCM/PM/NRZ signal when the ICR = −24 dB, −18.6 dB and −15.8 dB. It is clear that the model predicts the measurements for ICR < −15 dB. As the ICR increases, the measurements diverge from the model, as shown in Fig. 14(c). The same observations are reported for the PCM/PM/bi-phase signal type, as evident in Figs. 15(a), 15(b), and 15(c). However, for this signal type the deviation of the measured values from the theoretical values is less than that of the NRZ signal type for the same m and γ. The above deviations indicate that the assumptions made by regarding the data interference as additional white noise are valid as long as γ is less than a threshold value, as specified in Table 1. Figure 16 indicates that for the PCM/PSK/PM case, the measured performance is identical to the theoretical performance when $m = 1.1$ rad and $\gamma = 0.001$.

V. Conclusion

In this article, the performance of the PCM/PM/NRZ scheme was considered in the limiting two cases of high and very low data rates and the performance of PCM/PM/bi-phase scheme was considered for the high-data-rate case and compared with the PCM/PSK/PM scheme. As long as the PLL implementation by the DSN is used, the following conclusions are valid. In terms of SER performance, the PCM/PSK/PM scheme is better than the PCM/PM scheme (both NRZ and bi-phase) when the loop bandwidth-to-data rate ratio is not small enough, as discussed in the previous section. However, the PCM/PSK/PM scheme requires a great deal of bandwidth. In terms of bandwidth efficiency, the PCM/PM/NRZ scheme is the best but has higher degradation for $\gamma \geq 0.001$ and $m = 1.3$ rad. As a compromise, the PCM/PM/bi-phase scheme can be used with twice the bandwidth of the PCM/PM/NRZ scheme but with much less degradation. However, at a very small loop bandwidth-to-data rate ratio, which at present the DSN receivers are capable of supporting, the performance of the PCM/PM schemes coincides with that of the PCM/PSK/PM scheme. In addition, if one takes into account the subcarrier loss in the receiver when the PCM/PSK/PM scheme is employed, the performance of PCM/PM schemes becomes as good if not better than that of the PCM/PSK/PM scheme. Moreover, under these conditions the PCM/PM/NRZ scheme becomes the most attractive because it requires the least of bandwidth occupancy. If PLL-based receiver implementation is not required, then the BPSK scheme without a residual carrier is most attractive.

Acknowledgments

The authors thank Dr. V. Vilnrotter for all of his valuable comments and Brenda Goforth for her help in preparing this article for submission.

References

[1] J. Yuen, Ed., *Deep Space Telecommunications Systems Engineering*, New York: Plenum Press, 1983.

[2] Consultative Committee for Space Data Systems, *Recommendations for Space Data System Standards, Radio Frequency and Modulation Systems, Part I, Earth Stations and Spacecraft*, CCSDS 401.0-B, Blue Book, Washington, DC: NASA, CCSDS Secretariat, Communications and Data Systems Division (Code OS), 1989.

[3] International Radio Consultative Committee (CCIR), *Recommendations and Reports of the CCIR, XVIth Plenary Assembly, Volume II: Space Research and Radio Astronomy*, "Protection Criteria and Sharing Considerations Relating to Deep Space Research," Rep. 685-2, Dubrovnik, pp. 278–289, 1986.

[4] J. K. Holmes, *Coherent Spread Spectrum Systems*, New York: John Wiley and Sons, 1982.

[5] T. M. Nguyen, "Closed Form Expressions for Computing the Occupied Bandwidth of PCM/PSK/PM Signals," *Proceedings of the IEEE International Symposium on EMC*, Cherry Hills, New Jersey, August 1991.

[6] W. C. Lindsey and M. K. Simon, *Telecommunications Systems Engineering*, Englewood Cliffs, New Jersey: Prentice-Hall, 1973.

[7] T. M. Nguyen, "Space Telemetry Degradation Due to Manchester Data Asymmetry Induced Carrier Tracking Phase Error," *IEEE Transactions on EMC*, vol. 33, no. 3, pp. 262–268, August 1991.

[8] C. Y. Youn and W. C. Lindsey, "Phase-Locked Loop Performance in the Presence of CW Interference and Additive Noise," *IEEE Transactions on Communications*, vol. COM-30, no. 10, pp. 2305–2311, October 1982.

[9] T. M. Nguyen, "Occupied Bandwidths for PCM/PSK/PM and PCM/PM Signals—A Comparative Study," CCSDS Subpanel 1E, RF and Modulation Meeting, Salzburg, Austria, May 1992.

[10] S. Hinedi, "A Functional Description of the Advanced Receiver," *TDA Progress Report 42-100*, vol. October–December 1989, Jet Propulsion Laboratory, Pasadena, California, pp. 139–149, February 15, 1990.

Table 1. Critical values of $(\gamma)_{CR}$.

Modulation index, rad	γ		
	PCM/PM/NRZ	PCM/PM/bi-phase	PCM/PSK/PM, $n = 3$
0.9	0.010	0.235	Any value
1.3	0.002	0.035	0.56

Table 2. SSNR degradation when $\gamma = 0.001$ and $m = 1.3$ rad.

SER	SSNR degradation, dB		
	PCM/PM/NRZ	PCM/PM/bi-phase	PCM/PSK/PM, $n = 3$
10^{-3}	0.04	Negligible	Negligible
10^{-4}	0.06	Negligible	Negligible
10^{-5}	0.09	Negligible	Negligible

Table 3. SSNR degradation when $\gamma = 0.005$ and $m = 1.3$ rad.

SER	SSNR degradation, dB		
	PCM/PM/NRZ, NRR[a]	PCM/PM/bi-phase	PCM/PSK/PM, $n = 3$
10^{-3}	0.5	Negligible	Negligible
10^{-4}	0.7	Negligible	Negligible
10^{-5}	1.2	Negligible	Negligible

[a] NRR in the column corresponding to the PCM/PM/NRZ scheme refers to not recommended region, because the chosen value for γ is greater than the critical value for this scheme (see Table 1). Therefore, serious degradation for that particular scheme is observed.

Table 4. SSNR degradation when $\gamma = 0.005$ and $m = 1.1$ rad.

SER	SSNR degradation, dB		
	PCM/PM/NRZ, NRR[a]	PCM/PM/bi-phase	PCM/PSK/PM, $n = 3$
10^{-3}	0.10	Negligible	Negligible
10^{-4}	0.15	Negligible	Negligible
10^{-5}	0.20	Negligible	Negligible

[a] NRR = not recommended region.

Table 5. SSNR degradation when $\gamma = 0.05$ and $m = 1.3$ rad.

SER	SSNR degradation, dB		
	PCM/PM/NRZ	PCM/PM/bi-phase	PCM/PSK/PM, $n = 3$
10^{-3}	NRR[a] PLLNOR[b]	1.2 (NRR)	Negligible
10^{-4}	NRR PLLNOR	1.1 (NRR)	Negligible
10^{-5}	NRR PLLNOR	1.05 (NRR)	Negligible

[a] NRR = not recommended region.
[b] PLLNOR = PLL nonoperating region ($\rho < 8$ dB).

Table 6. Effective loop SNR when $m = 1.3$ rad and SSNR = 7 dB.

γ	Effective loop SNR, ρ, dB	
	PCM/PM/NRZ	PCM/PM/bi-phase
0.050	1.0 (NRR[a], PLLNOR[b])	9.0 (NRR)
0.010	8.0 (NRR),	16.0
0.005	11.5 (NRR)	19.0
0.001	18.0	25.9

[a] NRR = not recommended region.
[b] PLLNOR = PLL nonoperating region ($\rho < 8$ dB).

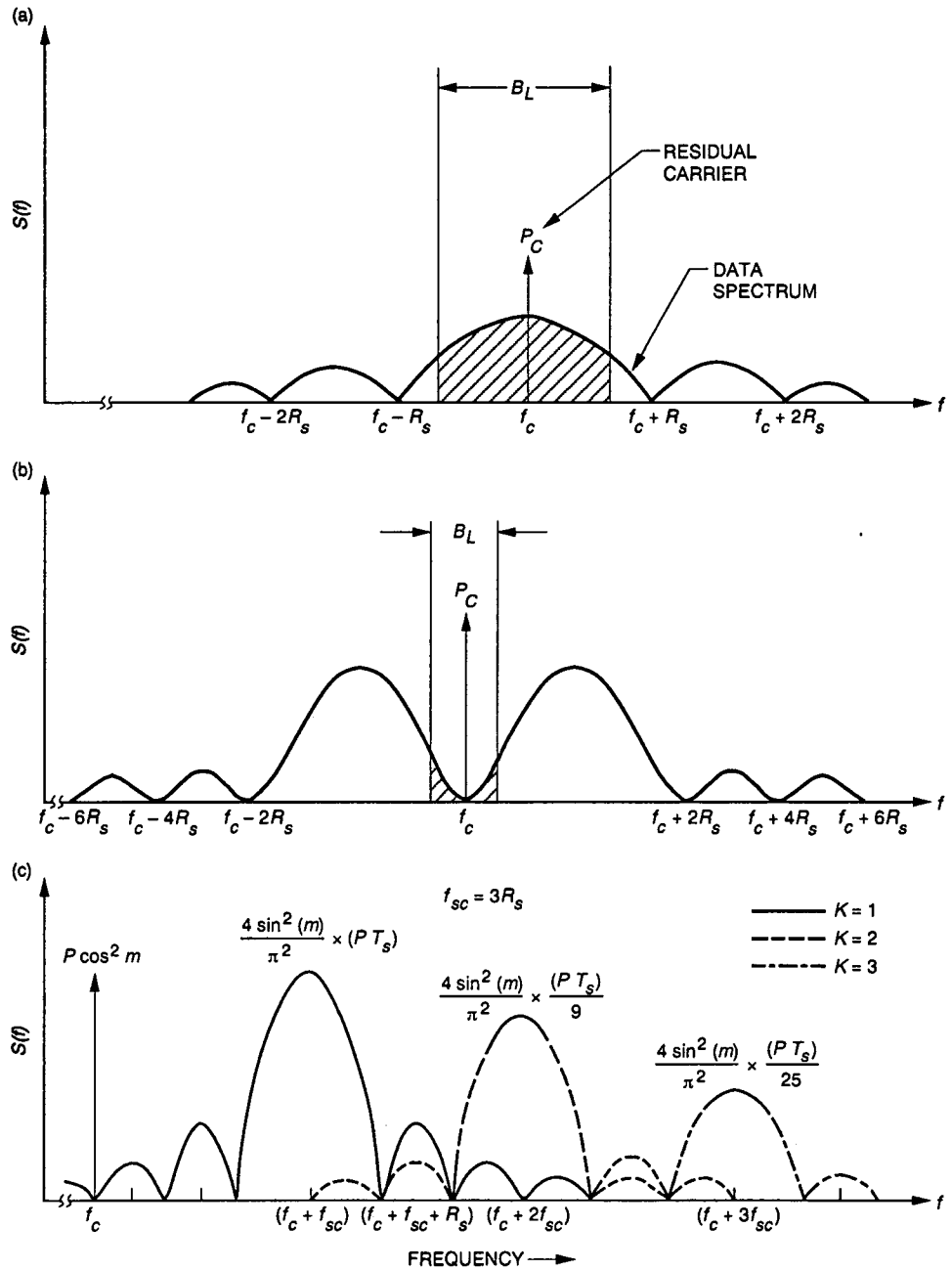


Fig. 1. Interaction between the residual carrier and the data spectrum: (a) interaction between the residual carrier and the NRZ data spectrum when $B_L < R_S$; (b) interaction between the residual carrier and the bi-phase data spectrum when $B_L < R_S$; and (c) power spectrum of the residual carrier with square waveform as its subcarrier.

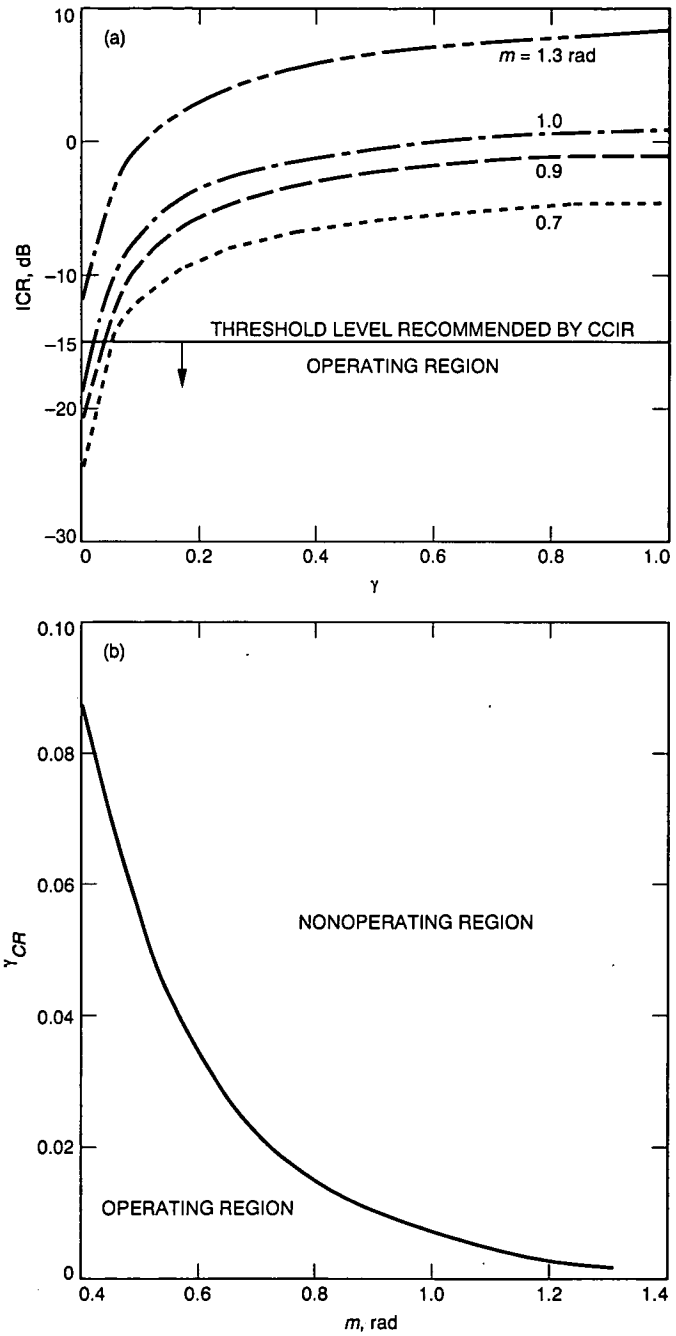


Fig. 2. Effects of the interference due to the NRZ data format: (a) ICR versus γ for the PCM/PM/NRZ scheme and (b) γ versus modulation index m when ICR = threshold value of –15 dB for the PCM/PM/NRZ case.

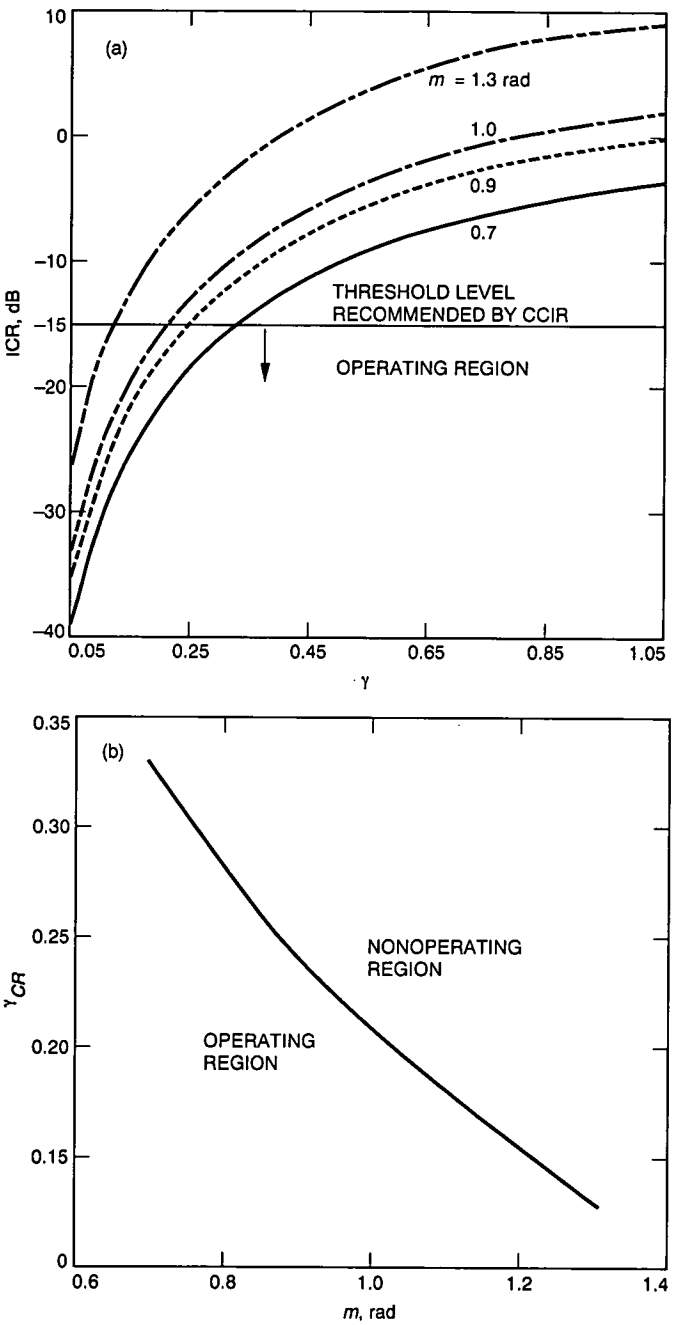


Fig. 3. Effects of the interference due to the bi-phase data format: (a) ICR versus γ for PCM/PM/bi-phase scheme and (b) γ versus modulation index m when ICR = –15 dB for the PCM/PM/bi-phase case.

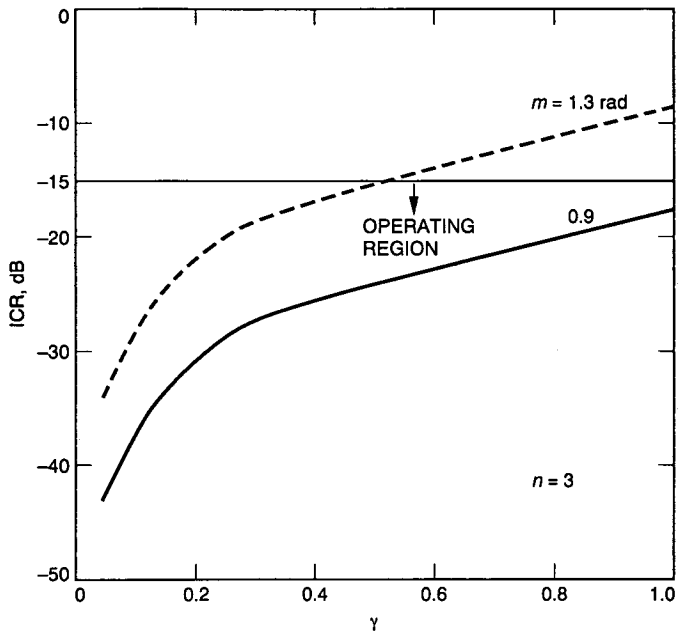


Fig. 4. ICR versus γ for the PCM/PSK/PM scheme with square-wave subcarrier.

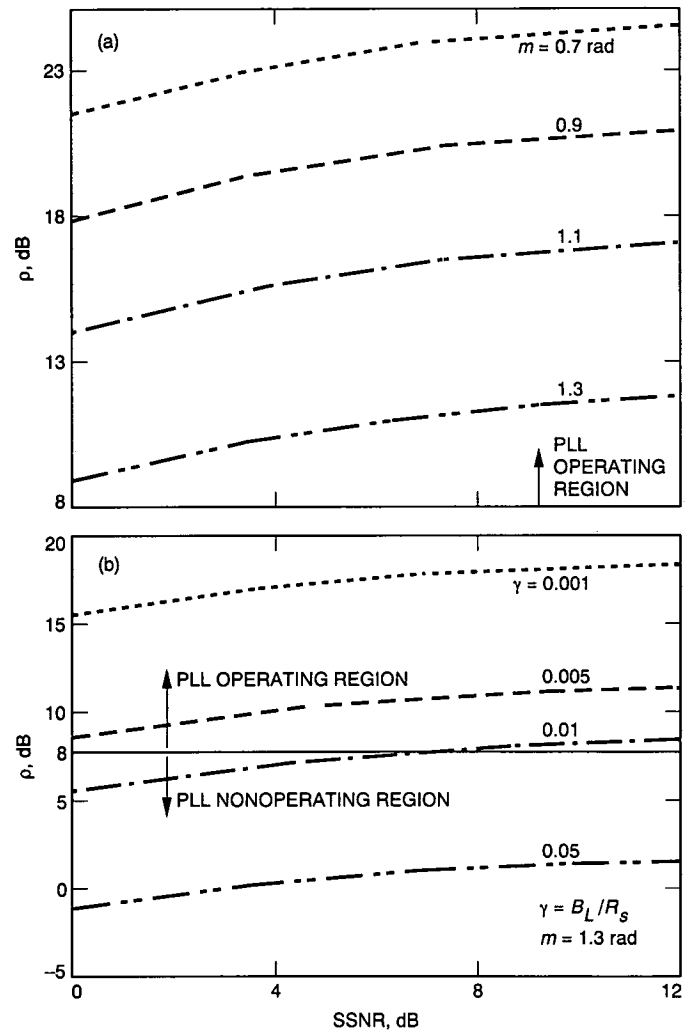


Fig. 5. Effective loop SNR as a function of SSNR and *m* for PCM/PM/ NRZ: (a) effective loop SNR versus SSNR for PCM/PM/ NRZ (γ = 0.005) and (b) effective loop SNR versus symbol SNR for PCM/PM/ NRZ (*m*=1.3 rad).

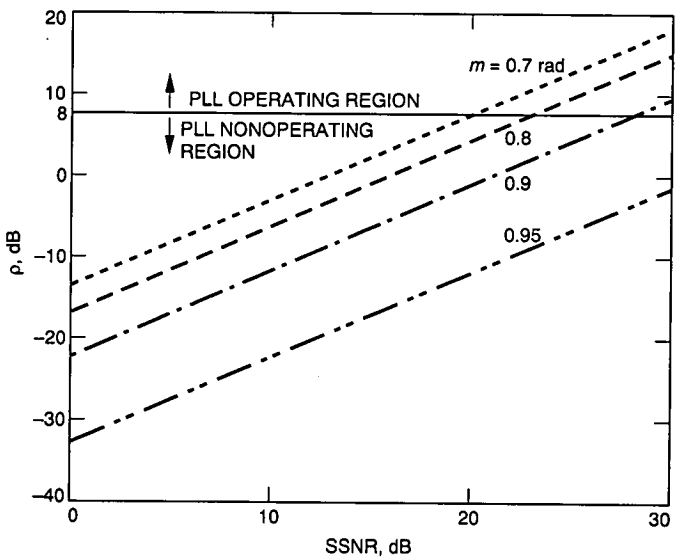


Fig. 6. Effective loop SNR versus SSNR for PCM/PM/NRZ (γ = 10), low data rate case.

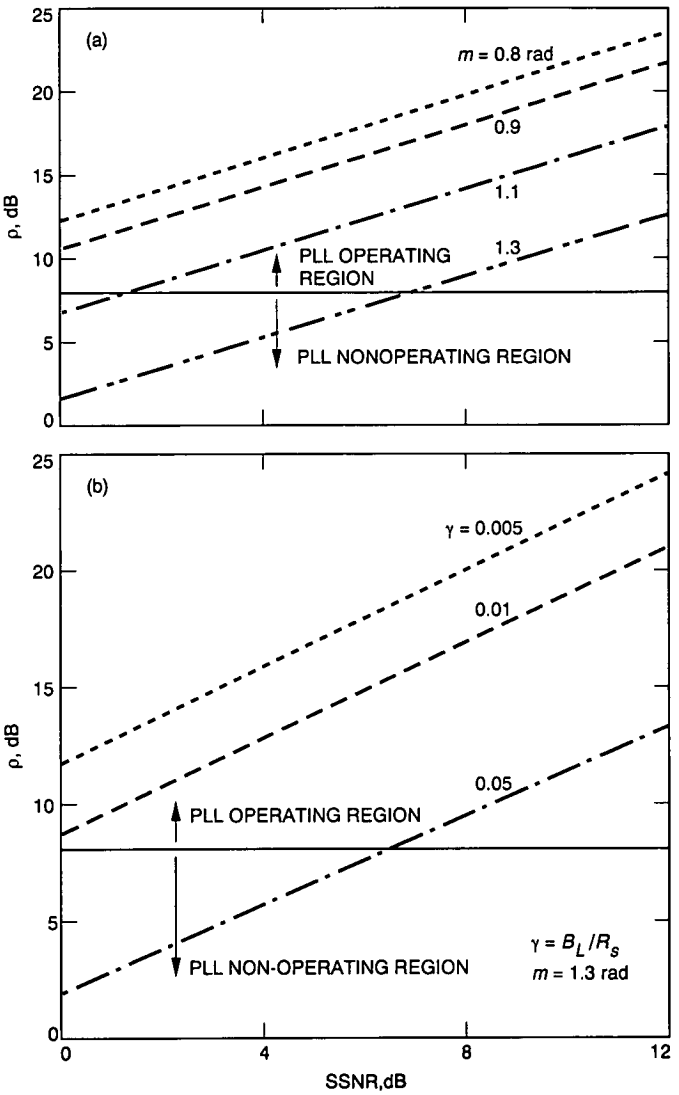


Fig 7. Effective loop SNR as a function of SNR for PCM/PM/bi-phase: (a) effective loop SNR versus SSNR for PCM/PM/bi-phase (γ = 0.05) and (b) effective loop SNR versus SSNR for PCM/PM/bi-phase (m = 1.3 rad).

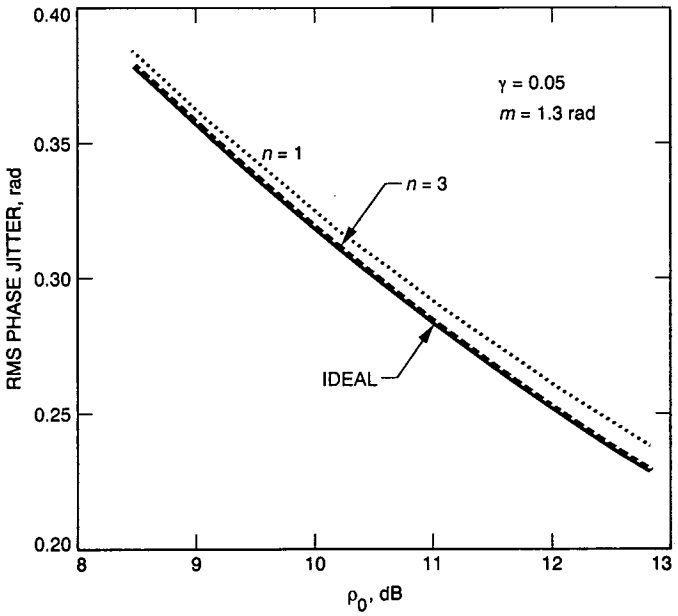


Fig. 8. RMS of phase jitter for the PCM/PM/PSK square-wave case.

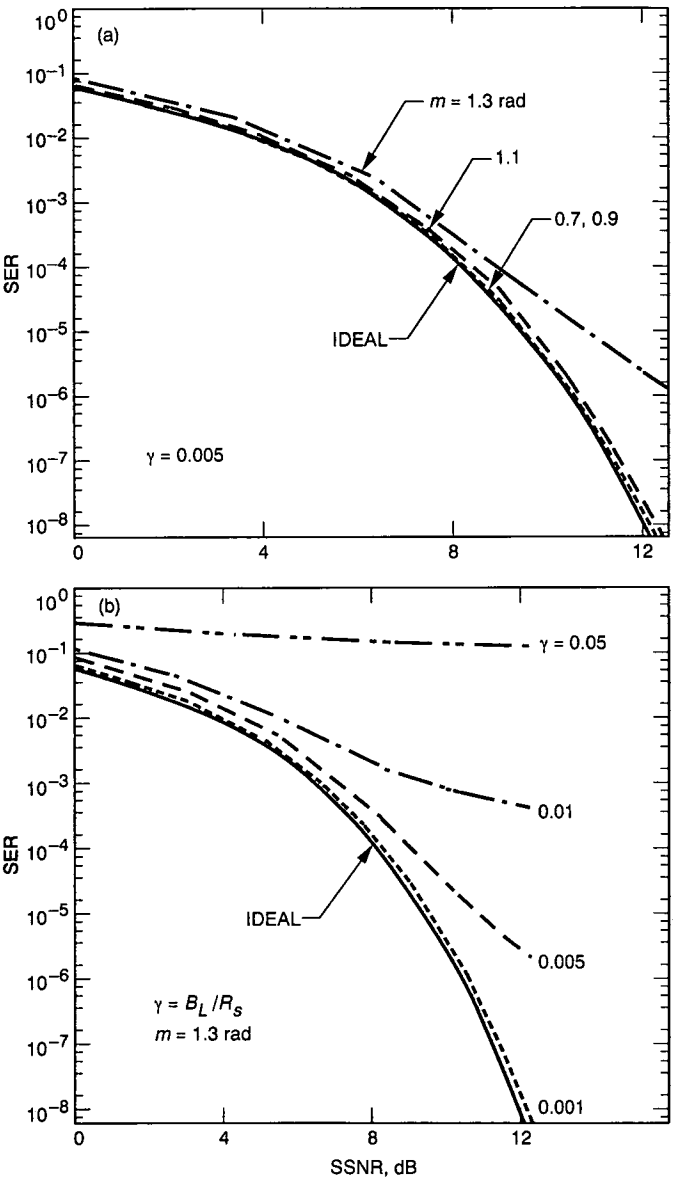


Fig. 9. SER for PCM/PM/NRZ: (a) SER versus SSNR for PCM/PM/NRZ (γ = 0.005) and (b) SER versus SSNR for PCM/PM/NRZ (m = 1.3 rad).

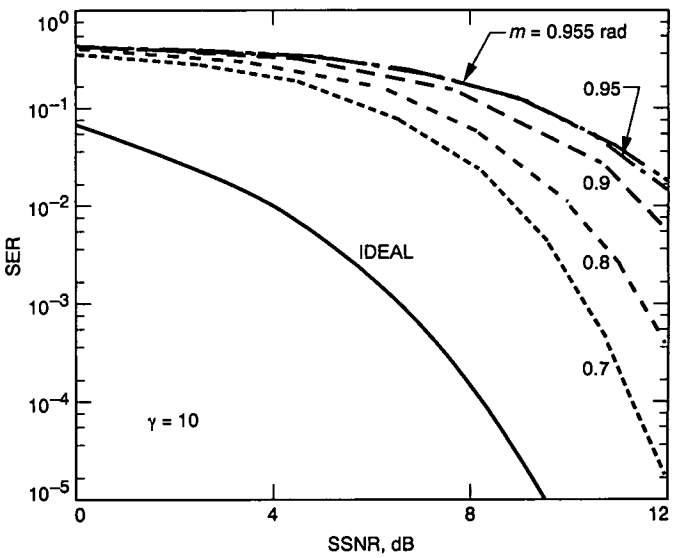


Fig. 10. SER versus SSNR for PCM/PM/NRZ (γ = 10), low-data rate case.

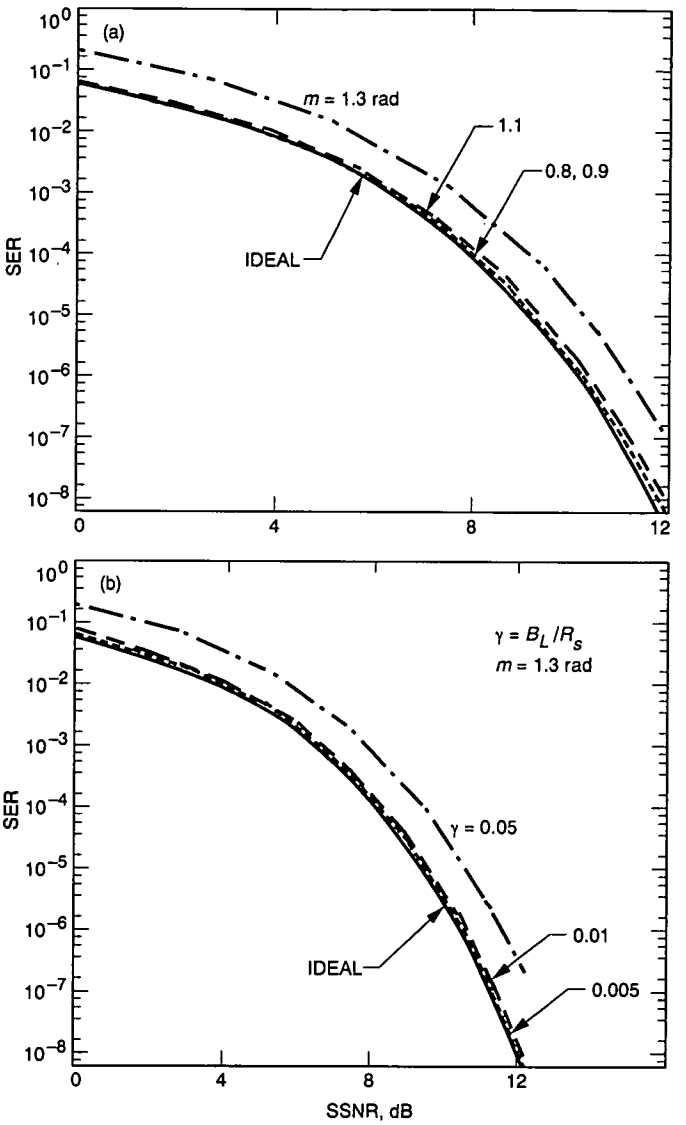


Fig. 11. SER for PCM/PM/bi-phase: (a) SER versus SSNR for PCM/PM/bi-phase (γ = 0.05) and (b) SER versus SSNR for PCM/PM/bi-phase (m = 1.3 rad).

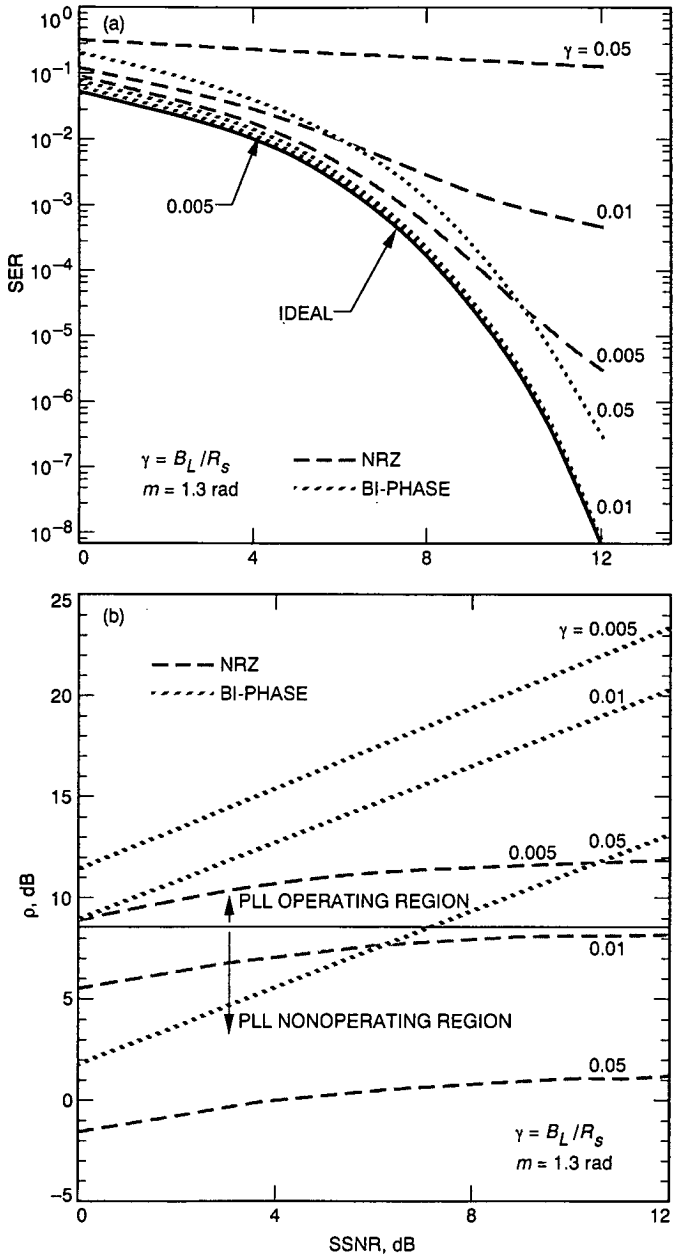


Fig. 12. Performance composition of PCM/PM/NRZ and PCM/PM/bi-phase: (a) SER versus SSNR for PCM/PM when m = 1.3 rad and (b) effective loop SNR versus SSNR for PCM/PM when m = 1.3 rad.

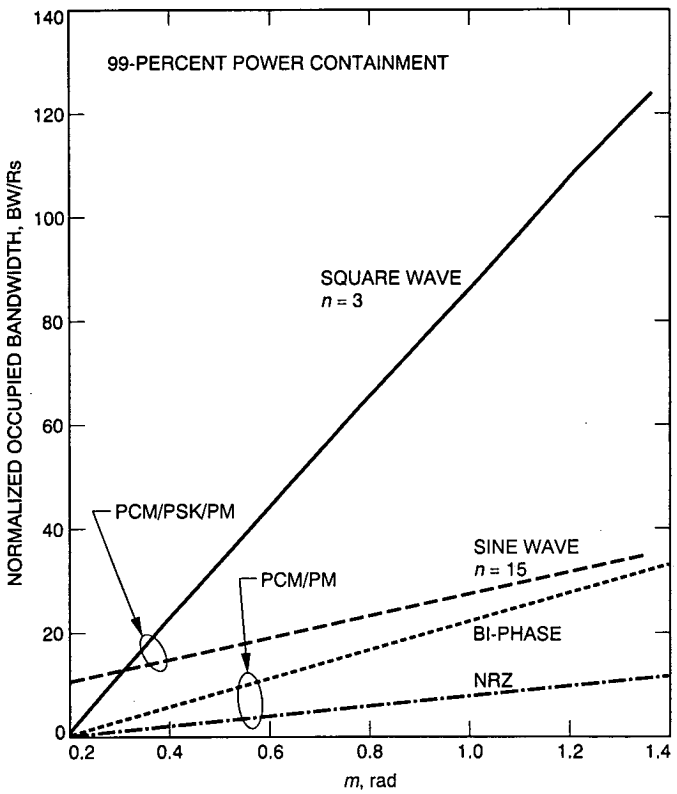


Fig. 13. Bandwidth occupancy of PCM/PSK/PM versus PCM/PM signals.

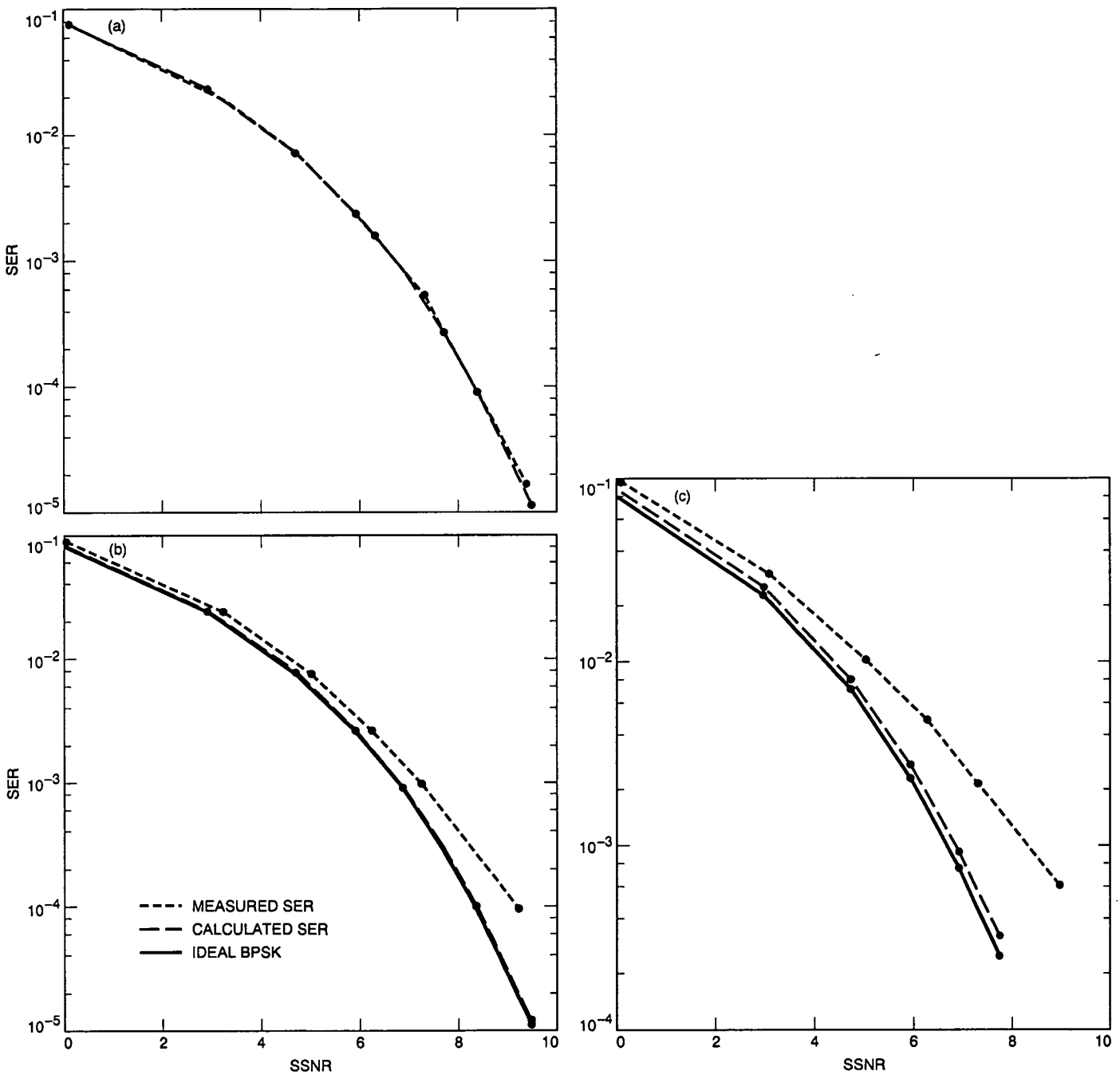


Fig. 14. Measured performance for PCM/PM/NRZ: (a) ICR = –24 dB (m = 1.1 rad, γ = 0.001); (b) ICR = –18.6 dB (m = 1.1 rad, γ = 0.001); and (c) ICR = –15.8 dB (m = 1.3 rad, γ = 0.002).

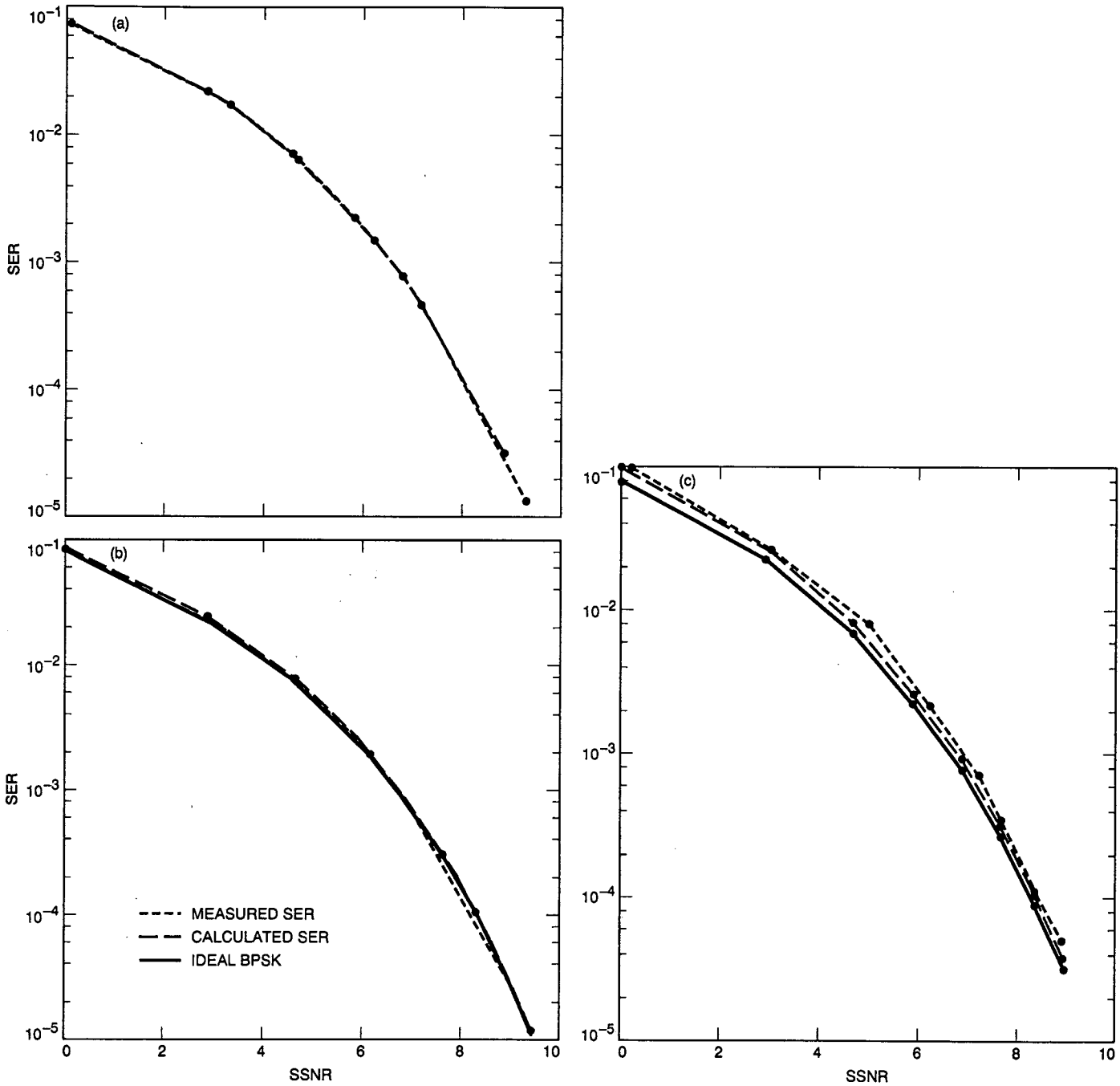


Fig. 15. Measured performance for PCM/PM/bi-phase: (a) ICR = –59.5 dB (m = 1.1 rad, γ = 0.001); (b) ICR = –54.2 dB (m = 1.3 rad, γ = 0.001); and (c) ICR = –40.2 dB (m = 1.3 rad, γ = 0.005).

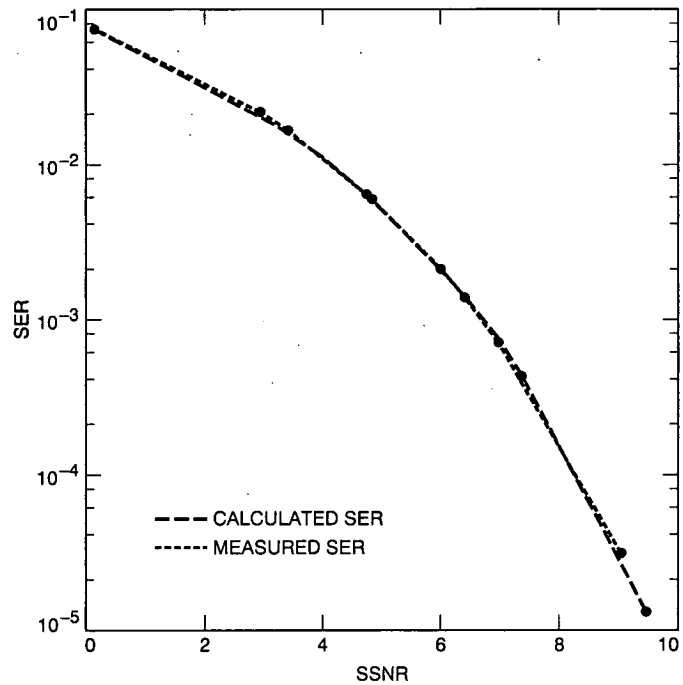


Fig. 16. Measured SER for PCM/PSK/PM with square-wave subcarrier, m = 1.1 rad and γ = 0.001.

Mathematical Modeling of Bent-Axis Hydraulic Piston Motors

R. D. Bartos
Ground Antennas and Facilities Engineering Section

Each of the DSN 70-m antennas uses 16 bent-axis hydraulic piston motors as part of the antenna drive system. On each of the two antenna axes, four motors are used to drive the antenna and four motors provide counter torque to remove the backlash in the antenna drive train. This article presents a mathematical model for bent-axis hydraulic piston motors. The model was developed to understand the influence of the hydraulic motors on the performance of the DSN 70-m antennas' servo control system.

I. Introduction

Hydraulic motors are often selected to drive mechanical devices because they are able to provide large output torque over a wide range of speeds while being relatively compact and lightweight. When a hydraulic motor is selected to drive a device under precision control conditions, it is necessary to understand how a particular motor will affect the overall system performance so that system components and control algorithms can be chosen to achieve optimum performance. The effect of a particular motor on system performance is best evaluated through computer simulations. This article develops a mathematical model, which can be implemented within any standard computer simulation package, for a bent-axis hydraulic piston motor. A glossary is provided in the Appendix.

II. Theoretical Model Development

The configuration of a typical bent-axis hydraulic piston motor has been provided in Figs. 1 [1], 2 [1], 3, and 4. This section develops mathematical models which describe the output torque and fluid flow rate through a bent-axis hydraulic piston motor.

A. Hydraulic Motor Torque

Based upon the geometrical relationships indicated in Fig. 3 and Fig. 4, the instantaneous torque generated by the fluid pressure forces is found to be given by

$$T_P = \sum_{i=1}^{n} ARP_i \sin\gamma \cos\theta_i \tag{1}$$

where

$$\theta_i = \theta_1 + \frac{2\pi(i-1)}{n} \tag{2}$$

$$P_i = P_A \qquad \text{if } \cos\theta_i > 0 \tag{3}$$

$$P_i = P_B \quad \text{if } \cos\theta_i < 0 \tag{4}$$

$$k = 0, \pm 1, \pm 2, \pm 3, \ldots \tag{5}$$

The variables used in Eqs. (1) through (5) are defined as

$A =$ area of one piston

$n =$ the number of pistons

$P_A =$ pressure at motor port A

$P_B =$ pressure at motor port B

$P_i =$ pressure acting on piston i

$R =$ pitch circle radius of the cylinder block bores

$\gamma =$ motor housing angle

$\theta_i =$ angle of the ith piston with respect to the valve plate, as shown in Fig. 4

Note that Eq. (1) only expresses the torque generated by the pressure forces and does not model the net torque output of the motor since losses due to viscous forces and Coulomb friction are not considered. Substituting the motor displacement per shaft revolution given by

$$D = 2nAR\sin\gamma \tag{6}$$

into Eq. (1) yields

$$T_P = \frac{D}{2n}\sum_{i=1}^{n} P_i \cos\theta_i \tag{7}$$

where θ_i and P_i are as defined in Eqs. (2) through (5). Equation (7) is a more useful form of Eq. (1) because manufacturers of hydraulic motors specify the size of their motors in terms of the displacement per revolution.

It is clear that the use of Eq. (7) involves very intensive computations which make it difficult to implement. Additionally, the equation gives no indication of the magnitude of the torque ripple relative to the mean value. Simulating the generated torque by using Eq. (7), it is evident that the generated torque can be approximated by the relations [2]

$$T_P \approx \frac{D\Delta P}{2\pi}\left[C_1 + C_2\left|\cos\left(n\theta_1\right)\right|\right] \quad \text{if } n = 2k+1 \tag{8}$$

$$T_P \approx \frac{D\Delta P}{2\pi}\left[C_1 + C_2\left|\cos\left(\frac{n\theta_1}{2}\right)\right|\right] \quad \text{if } n = 4k-2 \tag{9}$$

$$T_P \approx \frac{D\Delta P}{2\pi}\left[C_1 + C_2\left|\sin\left(\frac{n\theta_1}{2}\right)\right|\right] \quad \text{if } n = 4k \tag{10}$$

$$\Delta P = P_A - P_B \tag{11}$$

$$k = 1, 2, 3, \ldots \tag{12}$$

The nondimensional constants C_1 and C_2 are computed by means of least-squares estimation using Eq. (7) to obtain the actual motor torque. Parameter values are presented in Table 1 through Table 3, along with normalized standard deviation of the error between the approximations given in Eqs. (8) through (12) and the actual values given by Eq. (7). The estimation of the parameters was performed using 500 data points evenly spaced over a single ripple. The constant C_2 physically represents the fraction of the mean torque equivalent to the peak to peak amplitude of the ripple because

$$\int_0^{2\pi} \left[C_1 + C_2\left|\cos\left(n\theta_1\right)\right|\right] d\theta_1 = 1 \tag{13}$$

$$\int_0^{2\pi} \left[C_1 + C_2\left|\cos\left(\frac{n\theta_1}{2}\right)\right|\right] d\theta_1 = 1 \tag{14}$$

$$\int_0^{2\pi} \left[C_1 + C_2\left|\sin\left(\frac{n\theta_1}{2}\right)\right|\right] d\theta_1 = 1 \tag{15}$$

$$\text{mean torque} = \frac{D\Delta P}{2\pi} \tag{16}$$

Some of the torque generated by the pressure forces acting on the motor pistons is used to accelerate the internal motor parts. The torque expended to accelerate the motor parts is given by the equation

$$T_J = -J\alpha \tag{17}$$

where J is the inertia of the motor and α is the angular acceleration of the shaft. The negative sign is included in Eq. (17) because the inertial torque opposes the torque generated by the hydraulic pressure forces.

During the operation of the hydraulic motor, relative motion exists between the cylinder block and the valve

plate used to regulate the fluid pressure in the valve block bores. Figures 1 and 2 illustrate the interaction of the valve plate and cylinder block. The shear stress of the fluid in the clearance between the cylinder block and the valve plate is given by

$$\tau = \frac{\mu \omega r}{h} \tag{18}$$

where the variables are defined as

$h =$ clearance between the cylinder block and the valve plate

$r =$ radius from the center of the cylinder block

$\mu =$ absolute fluid viscosity

$\omega =$ angular velocity of the cylinder block

The torque due to viscous forces is found by integrating the shear stress over the area of the cylinder block to valve plate interface shown in Fig. 5 by using the equation

$$T = -\int_{R_1}^{R_2} \int_{\psi_1}^{\psi_2} (r\tau)\,(r\partial\psi\partial r) \tag{19}$$

Combining Eqs. (18) and (19) gives

$$T = -\int_{R_1}^{R_2} \int_{\psi_1}^{\psi_2} \frac{\mu\omega r^3}{h} \partial\psi\partial r \tag{20}$$

which after integration yields

$$T = -\frac{\mu\omega}{4h}\,(\psi_2 - \psi_1)\,(R_2^4 - R_1^4) \tag{21}$$

To compute the total viscous torque of the valve plate to cylinder block interface, the results gathered from Eq. (21) for all surfaces where viscous forces exist must be added together. Viscous forces also exist in the clearance between the cylinder block and the motor housing. The resisting torque created by these forces can be found by integrating the shear stress given by Eq. (18) over the lateral area of the cylinder block. By approximating the lateral surface of the cylinder block as a cylinder and integrating the shear stress over the lateral area of the cylinder, the torque created by viscous forces is found to be

$$T = -\frac{2\pi\mu L r^3 \omega}{h} \tag{22}$$

where

$$h << r \tag{23}$$

and where the variables are defined as

$h =$ clearance

$L =$ length of the cylinder block

$r =$ outside diameter of the cylinder block

From Eqs. (21) and (22), it is evident that the torque due to viscous friction is proportional to both viscosity and angular velocity. Hence, the torque due to viscous friction can be expressed as

$$T_V = -K_V D\mu\omega \tag{24}$$

where K_V is a constant of proportionality for a given motor. The motor displacement, D, is included as part of the proportionality constant in Eq. (24) for mathematical convenience when one is experimentally estimating model parameters.

An examination of Fig. 3 reveals that the five bearings of a hydraulic motor are loaded by the pressure forces acting upon each of the motor pistons. A forces and moment analysis on the shaft and piston subassembly shows that the bearings will be loaded proportionally to the hydraulic pressures in motor ports A and B, assuming that the pressure in the motor case is negligible. Since the motor bearings are not ideal and have Coulomb friction, there exists an average frictional torque acting on the motor shaft according to the equation

$$T_{F1} = -D\,(P_A + P_B)\,(K_F + K_{FS}\,|\mathrm{sgn}\,(\omega)|)\,\mathrm{sgn}\,(\omega) \tag{25}$$

where

$$\mathrm{sgn}\,(x) = \left\{ \begin{array}{ll} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{array} \right\} \tag{26}$$

and where the variables are defined as

$K_F =$ a proportionality constant under non-stick-slip conditions

$K_{FS} =$ a proportionality constant under stick-slip conditions

Shaft torques resulting from Coulomb friction are produced from the internal oil seals and from the piston motion within the barrel of the hydraulic motor. This internal friction is not related to the fluid pressures existing within the motor and is given by the relationship

$$T_{F2} = -T_{FC}\left(1 + K_S\left|\operatorname{sgn}(\omega)\right|\right)\operatorname{sgn}(\omega) \tag{27}$$

where T_{FC} is the constant friction torque at high speed and K_S is the proportionality constant under stick-slip conditions.

The torque output available to the hydraulic motor load is the sum of the torque effects due to the piston forces, motor inertia, viscous damping, and Coulomb friction. The net motor output torque to the load is therefore given by the equation

$$T_L = T_P + T_J + T_V + T_{F1} + T_{F2} \tag{28}$$

B. Hydraulic Motor Flow

Hydraulic motors are actuated by a control element such as a valve or pump which provides fluid flow to the motor. Hence, consideration of the fluid flow rate through the hydraulic motor is important from a system-level analysis point of view. By computing the rate at which volume is swept by the pistons using the geometrical relationships shown in Fig. 3 and Fig. 4, the theoretical flow rate through the motor, ignoring leakage effects, is found to be

$$Q_A = \sum_{i=1}^{n} K_{Ai} AR \sin\gamma \cos\theta_i \omega \tag{29}$$

$$Q_B = \sum_{i=1}^{n} K_{Bi} AR \sin\gamma \cos\theta_i \omega \tag{30}$$

$$K_{Ai} = 1 \quad \text{if } \cos\theta_i > 0 \tag{31}$$

$$K_{Ai} = 0 \quad \text{if } \cos\theta_i < 0 \tag{32}$$

$$K_{Bi} = 1 \quad \text{if } \cos\theta_i < 0 \tag{33}$$

$$K_{Bi} = 0 \quad \text{if } \cos\theta_i > 0 \tag{34}$$

$$\theta_i = \theta_1 + \frac{2\pi(i-1)}{n} \tag{35}$$

The variables are defined as

A = area of one piston

K_{Ai} = a constant

K_{Bi} = a constant

Q_A = flow rate into motor port A

Q_B = flow rate into motor port B

R = pitch circle radius of the cylinder block bores

γ = motor housing angle

ω = angular velocity of the motor shaft

Substituting the motor displacement given by Eq. (6) into Eqs. (29) and (30) yields

$$Q_A = \frac{D\omega}{2n} \sum_{i=1}^{n} K_{Ai} \cos\theta_i \tag{36}$$

$$Q_B = \frac{D\omega}{2n} \sum_{i=1}^{n} K_{Bi} \cos\theta_i \tag{37}$$

It is clear that Eqs. (36) and (37) are difficult to implement within a computer simulation package because they involve very intensive computations. These equations also do not give any indication of the magnitude of the flow ripple relative to the mean value. By simulating the fluid flow through the motors using Eqs. (36) and (37), it is evident that the fluid flow rate through the motor can be approximated by the relations [2]

$$Q_A \approx \frac{D\omega}{2\pi}\left[C_1 + C_2\left|\cos(n\theta_1)\right|\right] \quad \text{if } n = 2k+1 \tag{38}$$

$$Q_A \approx \frac{D\omega}{2\pi}\left[C_1 + C_2\left|\cos\left(\frac{n\theta_1}{2}\right)\right|\right] \quad \text{if } n = 4k-2 \tag{39}$$

$$Q_A \approx \frac{D\omega}{2\pi}\left[C_1 + C_2\left|\sin\left(\frac{n\theta_1}{2}\right)\right|\right] \quad \text{if } n = 4k \tag{40}$$

$$Q_B = -Q_A \tag{41}$$

$$k = 1, 2, 3, \ldots \tag{42}$$

The nondimensional constants C_1 and C_2 are computed by means of least-squares estimation using Eqs. (36) and (37) to obtain the actual motor flow rates. It turns out that the nondimensional constants have the same values as those for the torque expressed in Eqs. (8) through (10). As a result of Eqs. (13), (14), and (15), the constant C_2 physically represents the fraction of the mean flow equivalent to the peak to peak amplitude of the ripple where

$$\text{mean flow rate} = \frac{D\omega}{2\pi} \tag{43}$$

An evaluation of the flow rate through a hydraulic motor must consider the internal leakage of the motor. The leakage past the hydraulic motor pistons from the pressure and return ports to the case drain can be modeled as laminar flow in an annulus between a circular shaft and a cylinder. The leakage flow rate is given by the equation [3]

$$Q = \frac{\pi r c^3}{6\mu L}\left[1 + \frac{3}{2}\left(\frac{e}{c}\right)^2\right]\Delta P \tag{44}$$

where

$$c << r \tag{45}$$

The variables are defined as

$c =$ radial clearance

$e =$ eccentricity of the piston

$L =$ length of the leakage path

$\Delta P =$ the pressure difference between the motor chamber and the case

$r =$ radius of the cylinder bore

During the operation of a hydraulic motor, fluid leaks between the valve plate and the cylinder block. The leakage occurs from the case drain as shown by the curved arrows in Fig. 6. Also shown in Fig. 6 are the variables for one of the four leakage paths from the motor ports to the case drain. The leakage flow rate is derived as follows [4]: The leakage flow rate through a finite slot is given by

$$Q = \frac{\Delta P b h^3}{12\mu l} \tag{46}$$

where the variables are defined as

$b =$ slot width

$h =$ slot thickness

$l =$ slot length

$\Delta P =$ differential pressure

$\mu =$ absolute fluid viscosity

Applying Eq. (46) to a differential slot yields

$$Q = -\frac{2\pi r h^3}{12\mu}\frac{\partial P}{\partial r} \tag{47}$$

where the negative sign is used because flow occurs in the opposite direction of $\partial P/\partial r$. By rearranging Eq. (47) and integrating, the following expressions are obtained:

$$P = \int -\frac{6\mu Q}{\pi h^3 r}\partial r \tag{48}$$

$$P = -\frac{6\mu Q}{\pi h^3}\ln r + C \tag{49}$$

Applying boundary conditions and solving for the constant of integration, C, one obtains the relationship

$$Q = \frac{(P_1 - P_2)\pi h^3}{6\mu \ln\left(\frac{R_2}{R_1}\right)} \tag{50}$$

Since Eq. (50) gives the radial leakage over an angle of 2π radians while the motor leakage only occurs over an angle ϕ, Eq. (50) is adjusted by a factor of $\phi/2\pi$ to express the leakage from the motor port to the case drain as

$$Q = \frac{(P_1 - P_2)h^3\phi}{12\mu \ln\left(\frac{R_2}{R_1}\right)} \tag{51}$$

where the variables are defined as

$P_1 =$ pressure at radius 1

$P_2 =$ pressure at radius 2

$R_1 =$ radius 1

$R_2 =$ radius 2

$\phi =$ angular length of kidney-shaped motor port in radians

Using Eqs. (44) and (51), it is clear that the internal leakage from the port to the case is proportional to the pressure difference between the port and the case and inversely proportional to the fluid viscosity. Because the length of the leakage path found in Eq. (44) changes with the angular shaft position, the leakage proportionality constant will be a function of the motor shaft angle. The internal leakage variation with shaft angle has been observed by Johnson [5]. Hence, the internal leakage from the motor ports to the case drain is expressed as

$$Q_{LA} = \frac{K(\theta)_{LA}(P_A - P_C)}{\mu} \tag{52}$$

$$Q_{LB} = \frac{K(\theta)_{LB}(P_B - P_C)}{\mu} \tag{53}$$

where

$K(\theta)_{LA} =$ constant of proportionality for leakage from port A to case

$K(\theta)_{LB} =$ constant of proportionality for leakage from port B to case

$P_C =$ case pressure

Because the valve plate is symmetrical across the line dividing the A and B motor chambers, it is expected that

$$K(\theta)_{LA} = K(\theta - \varphi)_{LB} \tag{54}$$

where φ is a phase angle given by

$$\varphi = \frac{2\pi}{n} \tag{55}$$

and n is the number of pistons.

Internal hydraulic motor leakage exists between the two motor ports. The leakage flow between motor ports occurs across the valve plate to cylinder block sealing surface as illustrated in Fig. 7. Assume that the leakage path between the motor ports is modeled as a rectangular leakage path. The leakage flow rate for a single cross port leakage path is given by [6]

$$Q = \frac{bh^3}{12\mu l}\Delta P \tag{56}$$

Since there are two identical cross port leakage paths, the total cross port leakage is expressed as

$$Q = \frac{bh^3}{6\mu l}\Delta P \tag{57}$$

where the variables are defined as

$b =$ width of leakage path

$h =$ clearance

$l =$ length of leakage path

$\Delta P =$ pressure difference across the motor

Equation (57) shows that the cross port leakage is proportional to the pressure difference between the motor ports and inversely proportional to the viscosity of the fluid. The leakage from port A to port B can be expressed as

$$Q_{AB} = \frac{K_{AB}(P_A - P_B)}{\mu} \tag{58}$$

where K_{AB} is a constant of proportionality determined through experiments.

The net flow passing into port A, port B, and the case of a hydraulic motor is determined by summing the effects of the rate at which volume is swept by the pistons and the effects of internal leakage. The net flow to port A and port B is

$$Q_{AN} = Q_A + Q_{LA} + Q_{AB} \tag{59}$$

$$Q_{BN} = Q_B + Q_{LB} - Q_{AB} \tag{60}$$

respectively, and the net leakage to the case is

$$Q_{CN} = Q_{LA} + Q_{LB} \tag{61}$$

Positive flow is defined as flow into the motor port or case.

III. Conclusion

This article has presented a mathematical model for a bent-axis hydraulic piston motor; the motor can be used to perform dynamic system simulations which predict system performance under various conditions. The model parameters must be experimentally determined for a particular motor before the model can be used effectively because motor manufacturers do not possess this type of information or are unwilling to provide it. A discussion of the experimental test procedures required to determine the model parameters is beyond the scope of this article but can be obtained: see Bartos,[1] Merrit [3], and the *Hydraulic Fluid Power* documents [7,8].

[1] R. D. Bartos, *Statement of Work 332-SW-1015: Hydraulic Motor and Relief Valve Package Experimental Evaluation* (internal document), Jet Propulsion Laboratory, Pasadena, California, January 7, 1992.

References

[1] *Operations and Maintenance Information*, Vickers Hydraulic Piston Type Motors for Industrial Equipment Operation and Maintenance Information, I-1424-S, Vickers Incorporated, a TRINOVA company, Troy, Michigan, January 1, 1955.

[2] J. Watton, *Fluid Power Systems: Modeling, Simulation, Analog, and Microcomputer Control*, New York: Prentice Hall, 1989.

[3] H. E. Merritt, *Hydraulic Control Systems*, New York: John Wiley and Sons, 1967.

[4] D. D. Fuller, *Theory and Practice of Lubrication for Engineers*, New York: John Wiley and Sons, 1956.

[5] J. L. Johnson, "Predicting Low Speed Performance Of Hydraulic Motors By Digital Computer Simulation," *Proceedings of the National Conference on Fluid Power*, vol. 34, pp. 247–261, Cleveland, Ohio, October 28–30, 1980.

[6] J. Halling, *Principles of Tribology*, London: Macmillan Education Ltd., 1978.

[7] *Hydraulic Fluid Power: Determination of Characteristics of Motors, Part 1: At Constant Low Speed and At Constant Pressure*, ISO 4392-1, Global Engineering Documents, Irvine, California, August 15, 1989.

[8] *Hydraulic Fluid Power: Determination of Characteristics of Motors, Part 2: Startability*, ISO 4392-2, Global Engineering Documents, Irvine, California, August 15, 1989.

Table 1. Nondimensional parameter values for $n = 2k + 1$.

Number of pistons	C_1	C_2	$2\pi\sigma/D\Delta P$ or $2\pi\sigma/D\omega$
3	9.1331e–01	1.3617e–01	2.1080e–03
5	9.6932e–01	4.8183e–02	8.0199e–04
7	9.8442e–01	2.4468e–02	4.1507e–04
9	9.9059e–01	1.4773e–02	2.5254e–04
11	9.9371e–01	9.8797e–03	1.6955e–04
13	9.9550e–01	7.0697e–03	1.2159e–04
15	9.9662e–01	5.3083e–03	9.1425e–05
17	9.9737e–01	4.1318e–03	7.1228e–05
19	9.9789e–01	3.3072e–03	5.7050e–05

C_1 = a dimensionless constant.

C_2 = a dimensionless constant.

σ = standard deviation of the approximation error.

$2\pi\sigma/D\Delta P$ = standard deviation of the approximation error normalized with respect to the average torque.

$2\pi\sigma/D\omega$ = standard deviation of the approximation error normalized with respect to the average flow rate.

Table 2. Nondimensional parameter values for $n = 4k - 2$.

Number of pistons	C_1	C_2	$2\pi\sigma/D\Delta P$ or $2\pi\sigma/D\omega$
2	2.9905e–16	1.5708e+00	6.5742e–16
6	9.1331e–01	1.3617e–01	2.1080e–03
10	9.6932e–01	4.8183e–02	8.0199e–04
14	9.8442e–01	2.4468e–02	4.1507e–04
18	9.9059e–01	1.4773e–02	2.5254e–04

C_1 = a dimensionless constant.

C_2 = a dimensionless constant.

σ = standard deviation of the approximation error.

$2\pi\sigma/D\Delta P$ = standard deviation of the approximation error normalized with respect to the average torque.

$2\pi\sigma/D\omega$ = standard deviation of the approximation error normalized with respect to the average flow rate.

Table 3. Nondimensional parameter values for $n = 4k$.

Number of pistons	C_1	C_2	$2\pi\sigma/D\Delta P$ or $2\pi\sigma/D\omega$
4	7.9791e–01	3.1740e–01	4.2129e–03
8	9.5176e–01	7.5756e–02	1.2421e–03
12	9.7874e–01	3.3390e–02	5.6651e–04
16	9.8808e–01	1.8727e–02	3.2147e–04
20	9.9238e–01	1.1970e–02	2.0657e–04

C_1 = a dimensionless constant.

C_2 = a dimensionless constant.

σ = standard deviation of the approximation error.

$2\pi\sigma/D\Delta P$ = standard deviation of the approximation error normalized with respect to the average torque.

$2\pi\sigma/D\omega$ = standard deviation of the approximation error normalized with respect to the average flow rate.

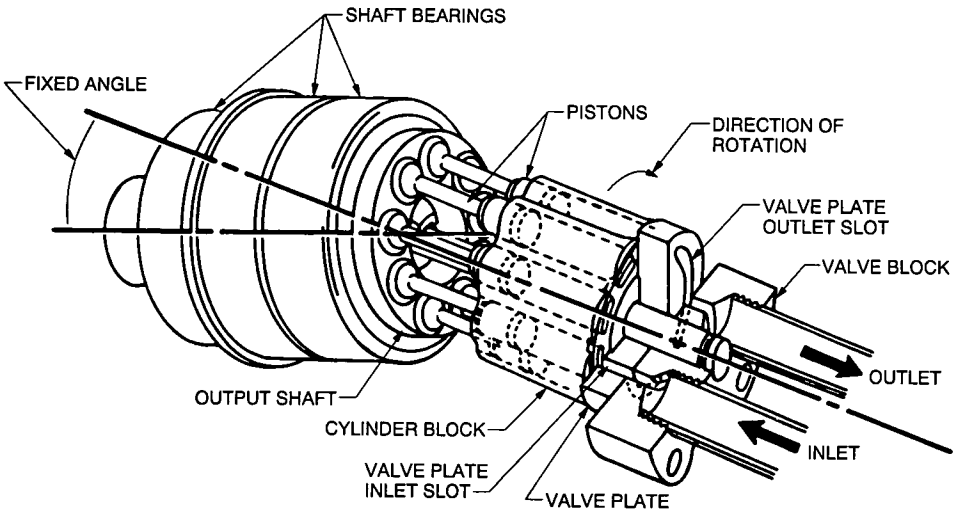


Fig. 1. Basic bent-axis hydraulic piston motor operation [1].

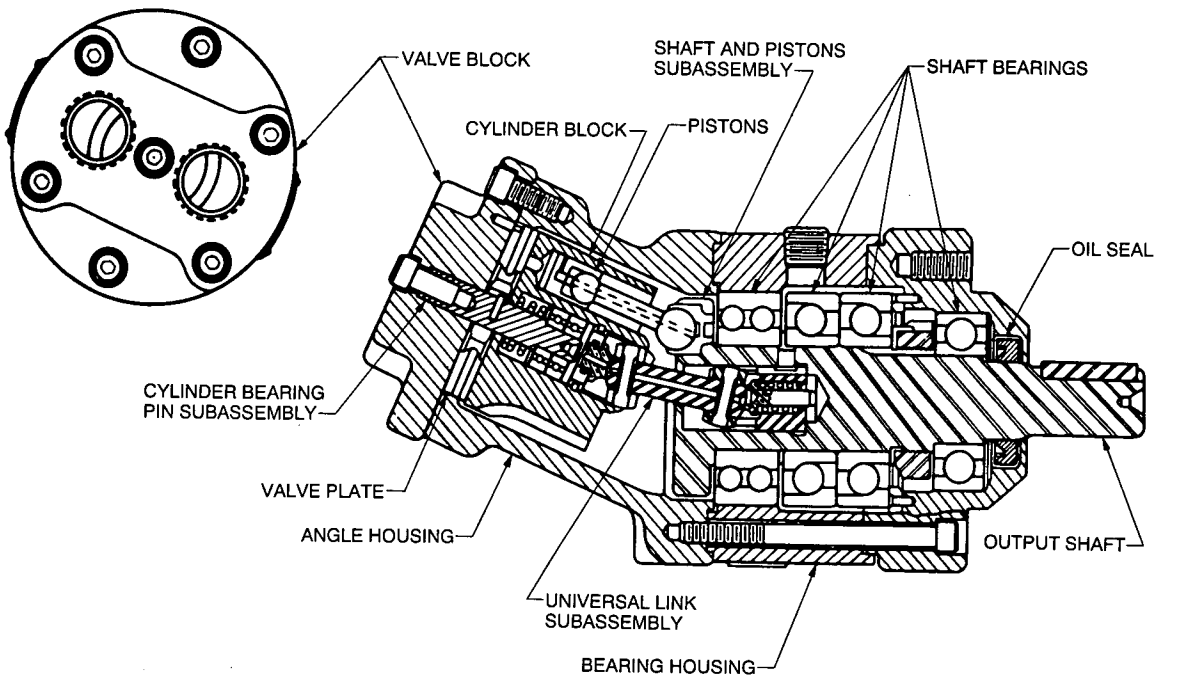
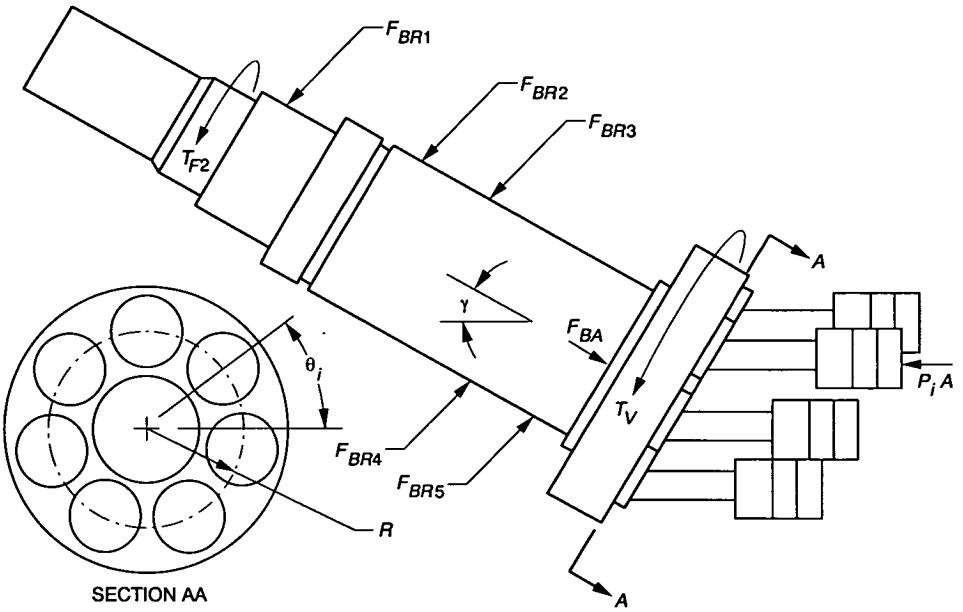


Fig. 2. Cross-sectional view of a bent-axis hydraulic piston motor [1].



A	PISTON AREA	T_{F2}	TORQUE DUE TO OIL SEAL AND PISTON FRICTION
F_{BA}	AXIAL REACTION FORCE OF BEARINGS	T_V	TORQUE DUE TO VISCOUS FORCES
F_{BRi}	RADIAL REACTION FORCE OF BEARING i	γ	MOTOR HOUSING ANGLE
P_i	PRESSURE ACTING ON PISTON i	θ_i	PISTON ANGLE FROM REFERENCE OF PISTON i
R	PITCH CIRCLE RADIUS		

Fig. 3. Shaft and piston subassembly of a bent-axis hydraulic piston motor.

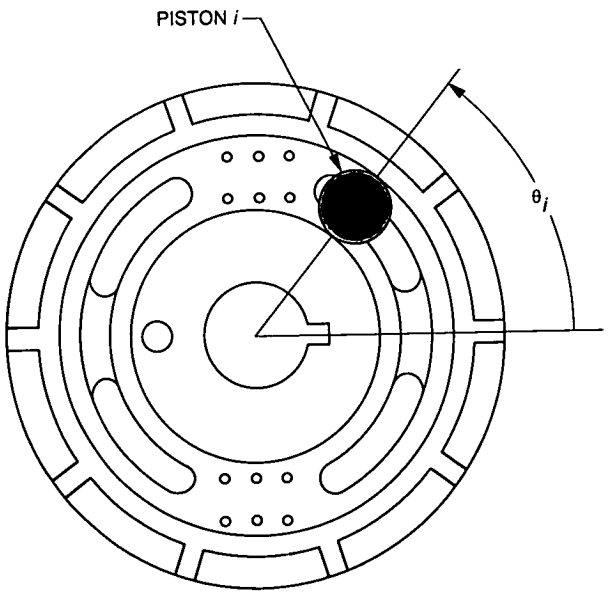


Fig. 4. Valve plate sealing surface of a bent-axis hydraulic piston motor and piston angle relative to the valve plate.

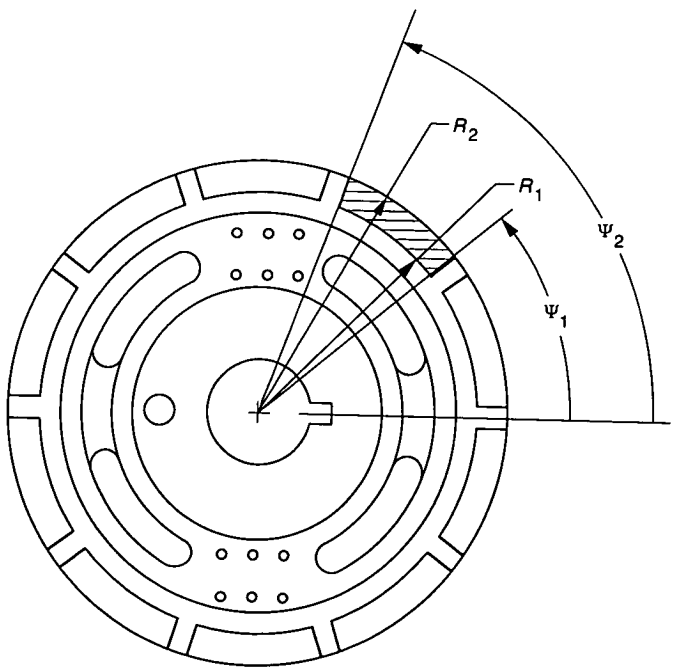


Fig. 5. One region of fluid shear stress acting between the valve plate to motor cylinder block interface.

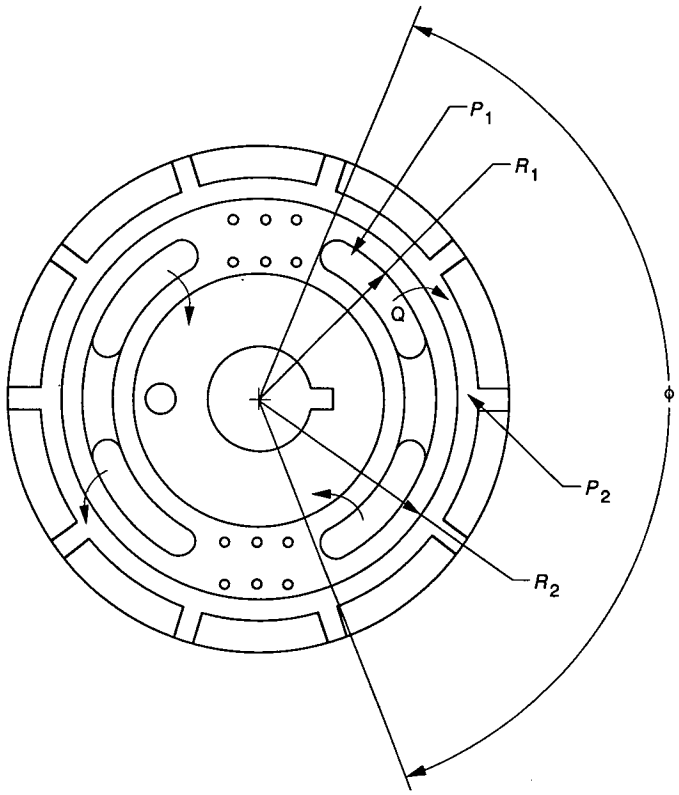


Fig. 6. Leakage flow paths across the valve plate sealin surfaces to the motor case.

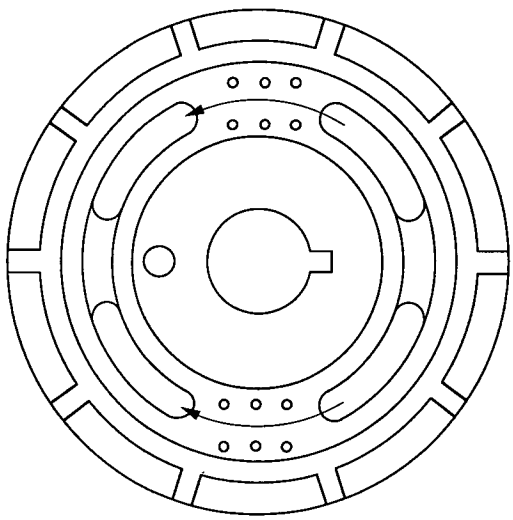


Fig. 7. Leakage flow path across the valve plate sealing surfaces between the two motor ports.

Appendix

A = area of one piston

b = slot width, width of leakage path

c = radial clearance

C = a constant of integration

C_1 = a dimensionless constant

C_2 = a dimensionless constant

D = motor displacement per revolution

e = eccentricity of the piston

h = clearance, slot thickness

i = piston number

J = motor inertia

k = integer

K_{Ai} = a constant

K_{Bi} = a constant

$K(\theta)_{AB}$ = a constant of proportionality related to the leakage between motor ports

K_F = a proportionality constant under kinetic friction conditions

K_{FS} = a proportionality constant under static friction conditions

$K(\theta)_{LA}$ = a constant of proportionality for leakage from port A to case

$K(\theta)_{LB}$ = a constant of proportionality for leakage from port B to case

K_S = a proportionality constant under static friction conditions

K_V = a constant of proportionality related to viscous friction

l = slot length, length of leakage path

L = cylinder block length

n = number of pistons

P = pressure

P_A = pressure of motor port A

P_B = pressure of motor port B

P_C = case pressure

P_i = pressure acting on piston i, pressure at boundary i

ΔP = differential pressure across the motor

Q_A = theoretical fluid flow rate into motor port A

Q_{AB} = leakage flow rate from port A to port B

Q_{AN} = net fluid flow rate into motor port A

Q_B = theoretical fluid flow rate into motor port B

Q_{BN} = net fluid flow rate into motor port B

Q_{CN} = net leakage flow to the case drain

Q_{LA} = internal leakage from port A to case

Q_{LB} = internal leakage from port B to case

r = radius

R = pitch circle radius of the cylinder block

R_i = radius at boundary i

T = torque

T_{FC} = constant friction torque at high speed independent of port pressures

T_{F1} = torque due to Coulomb friction originating from motor bearings

T_{F2} = torque due to Coulomb friction originating from oil seals and piston motion

T_J = torque to accelerate the inertial load

T_L = torque delivered to the load

T_P = torque generated by fluid pressure

T_V = torque required to overcome viscous friction

x = a variable

α = angular acceleration of the motor shaft

γ = motor housing angle

θ_i = angle of piston i

μ = absolute fluid viscosity

ρ = fluid density

τ = shear stress

φ = phase angle

ϕ = angular length of kidney-shaped port

Ψ = angle

ω = angular velocity of the motor shaft

Reducing the Net Torque and Flow Ripple Effects of Multiple Hydraulic Piston Motor Drives

R. D. Bartos
Ground Antennas and Facilities Engineering Section

The torque and flow ripple effects which result when multiple hydraulic motors are used to drive a single motion of a mechanical device can significantly affect the way in which the device performs. This article presents a mathematical model describing the torque and flow ripple effects of a bent-axis hydraulic piston motor. The model is used to show how the ripple magnitude can be reduced when multiple motors are used to drive a motion. A discussion of the hydraulic servo system of the 70-m antennas located within the Deep Space Network is included to demonstrate the application of the concepts presented.

I. Introduction

Multiple bent-axis hydraulic piston motors are frequently used to drive mechanical devices because they are capable of providing high torque while being physically small and lightweight. The output torque and flow rate of a bent-axis hydraulic piston motor vary with the angular position of a hydraulic motor shaft. Because most controllers are designed without regard for these variations in output torque and fluid flow rates, the performance of some machines may be significantly less than expected. This article presents a mathematical model for the output torque and fluid flow rate of a bent-axis hydraulic piston motor. This model is subsequently used to derive a method of reducing the torque and flow rate variations in hydraulic systems where multiple hydraulic motors are used to drive a single device. A glossary is provided in the Appendix.

II. Reduction of Torque and Flow Ripple Effects

When multiple bent-axis hydraulic piston motors are used to drive a single motion of a mechanical device, the resultant magnitude of the torque and flow ripple experienced by the device is significantly affected by how the relative angle between the reference piston and the valve plate within each motor is phased with the corresponding relative angles of the other hydraulic motors. The convention in this article for defining the angle between the reference motor piston and the valve plate is shown in Fig. 1. This section describes how the resultant torque and flow ripple can be reduced by properly phasing the hydraulic motors when multiple motors are used to drive a device.

A. Motor Torque

The analysis of the effects of motor phasing on the net torque ripple delivered to a device begins with the equation for the torque generated by the pressure forces acting on the pistons of a single bent-axis hydraulic piston motor, which is given by [1],

$$T_P = \frac{D}{2n} \sum_{i=1}^{n} P_i \cos\theta_i \qquad (1)$$

where

$$\theta_i = \theta_1 + \frac{2\pi(i-1)}{n} \tag{2}$$

$$P_i = P_A \qquad \text{if } \cos\theta_i > 0 \tag{3}$$

$$P_i = P_B \qquad \text{if } \cos\theta_i < 0 \tag{4}$$

$$k = 0, \pm 1, \pm 2, \pm 3, \ldots \tag{5}$$

The variables are defined as

D = motor displacement per revolution

k = integer

n = the number of pistons

P_A = pressure at motor port A

P_B = pressure at motor port B

P_i = pressure acting on piston i

θ_i = angle of the ith piston with respect to the valve plate, as shown in Fig. 1.

ω = angular velocity of the motor shaft

Based on Eq. (1), the total resultant torque generated by m motors driving a single device while operating at the same speed is given by

$$T_P = \sum_{j=1}^{m} \frac{Dm}{2n} \sum_{i=1}^{n} P_{ij} \cos\theta_{ij} \tag{6}$$

where

$$\theta_{ij} = \theta_{11} + \frac{2\pi(i-1)}{n} + \phi_{1j} \tag{7}$$

$$P_{ij} = P_A \qquad \text{if } \cos\theta_{ij} > 0 \tag{8}$$

$$P_{ij} = P_B \qquad \text{if } \cos\theta_{ij} < 0 \tag{9}$$

$$k = 0, \pm 1, \pm 2, \pm 3, \ldots \tag{10}$$

and where the variables are defined as

j = motor index number

P_{ij} = the pressure acting on piston i in motor j

ϕ_{1j} = the phase angle of the first piston in motor j

θ_{ij} = the angle of piston i in motor j

By using Eqs. (6) through (10), the maximum torque ripple is determined to occur when

$$\phi_{1j} = 0 \qquad j = 1, 2, 3, \ldots m \tag{11}$$

while the minimum torque ripple is found to occur when

$$\phi_{1j} = \frac{\pi(j-1)}{mn} \qquad \text{if } n = 2k+1 \tag{12}$$

$$\phi_{1j} = \frac{2\pi(j-1)}{mn} \qquad \text{if } n = 2k \tag{13}$$

$$k = 1, 2, 3, \ldots \tag{14}$$

If the motor phasing is given by Eqs. (12) through (14), the total instantaneous torque generated by the hydraulic motors is equivalent to the instantaneous torque generated by a single motor with a displacement of mD and $2mn$ pistons if the motor has an odd number of pistons, or a displacement of mD and mn pistons if the motor has an even number of pistons. The approximations for the generated torque given by [1],

$$T_P \approx \frac{D_e \Delta P}{2\pi} \left[C_1 + C_2 \left|\cos\left(n_e\theta_1\right)\right|\right] \qquad \text{if } n_e = 2k+1 \tag{15}$$

$$T_P \approx \frac{D_e \Delta P}{2\pi} \left[C_1 + C_2 \left|\cos\left(\frac{n_e\theta_1}{2}\right)\right|\right] \qquad \text{if } n_e = 4k-2 \tag{16}$$

$$T_P \approx \frac{D_e \Delta P}{2\pi} \left[C_1 + C_2 \left|\sin\left(\frac{n_e\theta_1}{2}\right)\right|\right] \qquad \text{if } n_e = 4k \tag{17}$$

$$\Delta P = P_A - P_B \tag{18}$$

$$k = 1, 2, 3, \ldots \tag{19}$$

can be used to analyze systems in which the ripple magnitude is either a maximum or a minimum in relationship to the motor phasing. In these equations, D_e represents the equivalent motor displacement, n_e represents the equivalent number of pistons, and C_1 and C_2 are nondimensional constants presented in Tables 1 through 3. The variables D_e and n_e are given by

$$D_e = mD \tag{20}$$

$$n_e = n \tag{21}$$

when the torque ripple magnitude is a maximum and

$$D_e = mD \tag{22}$$

$$n_e = mn \quad \text{if } n = 2k \tag{23}$$

$$n_e = 2mn \quad \text{if } n = 2k + 1 \tag{24}$$

when the torque ripple is a minimum. Notice that

$$n_e = 2n \quad \text{when } m = 1 \quad \text{and } n = 2k + 1 \tag{25}$$

Equation (25) shows why a manufacturer will not produce a motor with an even number of pistons, since a ripple of lower magnitude can be achieved using fewer pistons if there are an odd number of them.

B. Motor Flow

In applications where multiple hydraulic motors are actuated using a single control element, flow variations through the control element and pressure variations in the hydraulic plumbing can be reduced by phasing the motors properly. The flow rate derived from the rate at which volume is swept by the motor pistons is given by [1]

$$Q_A = \frac{D\omega}{2n} \sum_{i=1}^{n} K_{Ai} \cos\theta_i \tag{26}$$

$$Q_B = \frac{D\omega}{2n} \sum_{i=1}^{n} K_{Bi} \cos\theta_i \tag{27}$$

$$K_{Ai} = 1 \qquad \text{if } \cos\theta_i > 0 \tag{28}$$

$$K_{Ai} = 0 \qquad \text{if } \cos\theta_i < 0 \tag{29}$$

$$K_{Bi} = 1 \qquad \text{if } \cos\theta_i < 0 \tag{30}$$

$$K_{Bi} = 0 \qquad \text{if } \cos\theta_i > 0 \tag{31}$$

$$\theta_i = \theta_1 + \frac{2\pi(i-1)}{n} \tag{32}$$

where

D = motor displacement per revolution

K_{Ai} = a constant

K_{Bi} = a constant

n = the number of pistons

Q_A = flow rate into motor port A

Q_B = flow rate into motor port B

θ_i = angle of the ith piston with respect to the valve plate, as shown in Fig. 1

ω = angular velocity of the motor shaft

Under the conditions when multiple hydraulic motors are operated, the total instantaneous flow rate through the motors is found, using Eqs. (26) and (27), to be

$$Q_A = \sum_{j=1}^{m} \frac{D\omega}{2n} \sum_{i=1}^{n} K_{Aij} \cos\theta_{ij} \tag{33}$$

$$Q_B = \sum_{j=1}^{m} \frac{D\omega}{2n} \sum_{i=1}^{n} K_{Bij} \cos\theta_{ij} \tag{34}$$

where

$$\theta_{ij} = \theta_{11} + \frac{2\pi(i-1)}{n} + \phi_{1j} \tag{35}$$

$$K_{Aij} = 1 \qquad \text{if } \cos\theta_{ij} > 0 \tag{36}$$

$$K_{Aij} = 0 \qquad \text{if } \cos\theta_{ij} < 0 \tag{37}$$

$$K_{Bij} = 1 \qquad \text{if } \cos\theta_{ij} < 0 \tag{38}$$

$$K_{Bij} = 0 \qquad \text{if } \cos\theta_{ij} > 0 \tag{39}$$

The variables are defined as

j = motor index number

K_{Aij} = a constant

K_{Bij} = a constant

ϕ_{1j} = the phase angle of the first piston in motor j

θ_{ij} = the angle of piston i in motor j

Based upon Eqs. (33) through (39), the maximum flow ripple through the control element occurs when

$$\phi_{1j} = 0 \qquad \text{if } j = 1,2,3,\ldots,m \tag{40}$$

and the minimum occurs when

$$\phi_{1j} = \frac{\pi(j-1)}{mn} \qquad \text{if } n = 2k+1 \tag{41}$$

$$\phi_{1j} = \frac{2\pi(j-1)}{mn} \qquad \text{if } n = 2k \tag{42}$$

$$k = 1,2,3,\ldots \tag{43}$$

It is important to note that the motor phasing at which the maximum and minimum flow ripples occur is the same as the phasing at which the maximum and minimum torque ripples occur. If the motor phasing is given by Eqs. (41) through (43), the total instantaneous flow rate through the control element is equivalent to the instantaneous flow rate of a single motor with a displacement of mD and $2mn$ pistons if the motor has an odd number of pistons, or a displacement of mD and mn pistons if the motor has an even number of pistons. The approximations for the instantaneous flow rate given by [1],

$$Q_A \approx \frac{D_e\omega}{2\pi}\left[C_1 + C_2\left|\cos\left(n_e\theta_1\right)\right|\right] \qquad \text{if } n_e = 2k+1 \tag{44}$$

$$Q_A \approx \frac{D_e\omega}{2\pi}\left[C_1 + C_2\left|\cos\left(\frac{n_e\theta_1}{2}\right)\right|\right] \qquad \text{if } n_e = 4k-2 \tag{45}$$

$$Q_A \approx \frac{D_e\omega}{2\pi}\left[C_1 + C_2\left|\sin\left(\frac{n_e\theta_1}{2}\right)\right|\right] \qquad \text{if } n_e = 4k \tag{46}$$

$$Q_B = -Q_A \tag{47}$$

$$k = 1,2,3,\ldots \tag{48}$$

can be used to analyze systems in which the ripple magnitude is either a maximum or a minimum in relationship to the motor phasing. In these equations D_e represents the equivalent motor displacement, n_e represents the equivalent number of pistons, and C_1 and C_2 are nondimensional constants presented in Tables 1 through 3. The variables D_e and n_e are given in Eqs. (20) through (24) since they take on the same values as they do in the torque equations.

III. An Application

The Deep Space Network 70-m antenna hydraulic drive system is considered here in order to apply the analysis techniques presented. The drive system of the 70-m antennas for each axis of motion consists of four hydraulic bent-axis piston motors which are simultaneously controlled by a single servovalve. In addition, there are four counter-torque motors which have a constant differential pressure maintained across the ports. The configuration of the system for a single axis is shown in Fig. 2. The objective of this analysis is to reduce the net torque variations experienced by the antenna at a given differential pressure across the motors in order to improve the performance of the position controller. On the 70-m antennas the following parameter values apply:

$$D = 3.949 \times 10^{-5}\text{m}^3/\text{revolution} \quad (2.41\ \text{in.}^3/\text{revolution}) \tag{49}$$

$$n = 7 \text{ pistons} \tag{50}$$

$$m = 4 \text{ motors} \tag{51}$$

Transmission line length between the servovalve and motors 1 and 3 = 22.352 m (880 in.) (52)

Transmission line length between the servovalve and motors 2 and 4 = 11.430 m (450 in.) (53)

The analysis begins by first considering the phasing of the counter-torque motors. It is evident from Fig. 2 that two of the counter-torque motors create a torque on the antenna opposite the direction of the other two counter-torque motors. Since the differential pressure across all of the counter-torque motors is maintained at 4136.88 kilopascals (kPa), or 600 pounds per square inch differential (PSID), the instantaneous torque delivered to the antenna can be maintained at zero independent of the angular position of the antenna, if the counter-torque motors are all in

phase. This occurs because the instantaneous torque generated by two of the counter-torque motors exactly cancels the instantaneous torque generated by the other two counter-torque motors. The second consideration in reducing the torque variations with respect to the antenna angular position is the phasing of the four drive motors. All of the hydraulic motors have approximately the same pressure drop across the motor ports at any instant in time since they are connected in parallel to the same servovalve. Any differential pressure variations that exist among them are due to the pressure drops across the various lengths of pipe connecting the servovalve to each of the motors. These pressure drops are generally assumed to be negligible since the flow rates through the pipes under tracking conditions are very small in relation to the diameter of the pipes. Hence, the optimal phasing for the four motors with seven pistons each, as given by Eqs. (12) and (41), is

$$\phi_{11} = 0 \text{ radians} \tag{54}$$

$$\phi_{12} = \frac{\pi}{28} \text{ radians} \tag{55}$$

$$\phi_{13} = \frac{\pi}{14} \text{ radians} \tag{56}$$

$$\phi_{14} = \frac{3\pi}{28} \text{ radians} \tag{57}$$

These phases were selected such that the motor pairs (motor 1 and motor 3) and (motor 2 and motor 4) had the optimal phasing of a two-motor drive system because the length of the hydraulic line from the servovalve is the same for each motor of a given pair. [See Eqs. (52) and (53).] The total torque exerted on each antenna axis for both the best and worst phasing of the motors and countertorque motors for a differential motor pressure of 4826.36 kPa (700 PSID) is presented in Fig. 3, and the total flow passing through the motors for the best and worst phasing of the motors and counter-torque motors for a rotational speed of one radian per second is presented in Fig. 4. Upon examination of Fig. 3, it is clear that the peak-to-peak magnitude of the torque ripple experienced by the antenna is approximately 1.5818×10^5 newton-meters (N-m) under the worst phasing of the hydraulic motors and countertorque motors, while the peak-to-peak magnitude is approximately 5.3017×10^3 N-m under the best phasing conditions. The 1.5818×10^5 N-m peak-to-peak amplitude which occurs under the worst phasing conditions is considered significant since the amplitude is equivalent to the axis torque created by a 2.69 m/sec wind load [2]. Figure 4 shows that the flow variations experienced by the servovalve are $6.1416 \times 10^{-7} m^3/sec$ and $3.8243 \times 10^{-8} m^3/sec$ under the worst and best phasing conditions, respectively. The results obtained through this analysis without consideration of friction, leakage, or line capacitance indicate that the performance of the 70-m antennas may be improved by properly phasing the motors, thus reducing the magnitude of system nonlinearities.

IV. Conclusion

Frequently multiple bent-axis piston type hydraulic motors are used to drive a single motion of a mechanical device. With this type of system configuration, the phase of the pistons within each motor relative to the pistons of the other motors can significantly affect the magnitude of the torque ripple experienced by the device and the flow variations experienced by the control element which provides fluid to the motors. This article has described how the minimum torque and flow variations can be obtained. An example related to the hydraulic system of the 70-m antennas within the Deep Space Network illustrated how the concepts presented can be applied to improve system performance.

References

[1] R. D. Bartos, "Mathematical Modeling of Bent-Axis Hydraulic Piston Motors," *TDA Progress Report 42-111*, vol. July–September, 1992, Jet Propulsion Laboratory, Pasadena, California, pp. 224–235, November 15, 1992.

[2] H. McGinness, "Antenna Axis Drive Torques for the 70-Meter Antenna," *TDA Progress Report 42-80*, vol. October–December, 1984, Jet Propulsion Laboratory, Pasadena, California, pp. 121–126, February 15, 1985.

Table 1. Nondimensional parameter values for $n_\theta = 2k + 1$.

Number of pistons	C_1	C_2	$2\pi\sigma/D\Delta P$ or $2\pi\sigma/D\omega$	Number of pistons	C_1	C_2	$2\pi\sigma/D\Delta P$ or $2\pi\sigma/D\omega$
3	9.1331e–01	1.3617e–01	2.1080e–03	53	9.9973e–01	4.2478e–04	7.3442e–06
5	9.6932e–01	4.8183e–02	8.0199e–04	55	9.9975e–01	3.9444e–04	6.8199e–06
7	9.8442e–01	2.4468e–02	4.1507e–04	57	9.9977e–01	3.6725e–04	6.3498e–06
9	9.9059e–01	1.4773e–02	2.5254e–04	59	9.9978e–01	3.4277e–04	5.9267e–06
11	9.9371e–01	9.8797e–03	1.6955e–04	61	9.9980e–01	3.2066e–04	5.5445e–06
13	9.9550e–01	7.0697e–03	1.2159e–04	63	9.9981e–01	3.0062e–04	5.1981e–06
15	9.9662e–01	5.3083e–03	9.1425e–05	65	9.9982e–01	2.8241e–04	4.8832e–06
17	9.9737e–01	4.1318e–03	7.1228e–05	67	9.9983e–01	2.6580e–04	4.5961e–06
19	9.9789e–01	3.3072e–03	5.7050e–05	69	9.9984e–01	2.5061e–04	4.3335e–06
21	9.9828e–01	2.7069e–03	4.6717e–05	71	9.9985e–01	2.3669e–04	4.0929e–06
23	9.9856e–01	2.2564e–03	3.8956e–05	73	9.9986e–01	2.2390e–04	3.8717e–06
25	9.9878e–01	1.9097e–03	3.2979e–05	75	9.9986e–01	2.1212e–04	3.6680e–06
27	9.9896e–01	1.6372e–03	2.8279e–05	77	9.9987e–01	2.0124e–04	3.4799e–06
29	9.9910e–01	1.4191e–03	2.4516e–05	79	9.9988e–01	1.9118e–04	3.3060e–06
31	9.9921e–01	1.2418e–03	2.1457e–05	81	9.9988e–01	1.8185e–04	3.1448e–06
33	9.9930e–01	1.0958e–03	1.8936e–05	83	9.9989e–01	1.7319e–04	2.9950e–06
35	9.9938e–01	9.7415e–04	1.6835e–05	85	9.9989e–01	1.6514e–04	2.8558e–06
37	9.9945e–01	8.7166e–04	1.5065e–05	87	9.9990e–01	1.5763e–04	2.7260e–06
39	9.9950e–01	7.8454e–04	1.3560e–05	89	9.9990e–01	1.5063e–04	2.6049e–06
41	9.9955e–01	7.0985e–04	1.2270e–05	91	9.9991e–01	1.4408e–04	2.4916e–06
43	9.9959e–01	6.4535e–04	1.1156e–05	93	9.9991e–01	1.3795e–04	2.3856e–06
45	9.9962e–01	5.8925e–04	1.0187e–05	95	9.9992e–01	1.3220e–04	2.2862e–06
47	9.9966e–01	5.4017e–04	9.3383e–06	97	9.9992e–01	1.2681e–04	2.1929e–06
49	9.9968e–01	4.9697e–04	8.5918e–06	99	9.9992e–01	1.2174e–04	2.1052e–06
51	9.9971e–01	4.5875e–04	7.9313e–06				

C_1 = a dimensionless constant.

C_2 = a dimensionless constant.

σ = standard deviation of the approximation error.

$2\pi\sigma/D\Delta P$ = standard deviation of the approximation error normalized with respect to the average torque.

$2\pi\sigma/D\omega$ = standard deviation of the approximation error normalized with respect to the average flow rate.

Table 2. Nondimensional parameter values for $n_e = 4k - 2$.

Number of pistons	C_1	C_2	$2\pi\sigma/D\Delta P$ or $2\pi\sigma/D\omega$
2	2.9905e–16	1.5708e+00	6.5742e–16
6	9.1331e–01	1.3617e–01	2.1080e–03
10	9.6932e–01	4.8183e–02	8.0199e–04
14	9.8442e–01	2.4468e–02	4.1507e–04
18	9.9059e–01	1.4773e–02	2.5254e–04
22	9.9371e–01	9.8797e–03	1.6955e–04
26	9.9550e–01	7.0697e–03	1.2159e–04
30	9.9662e–01	5.3083e–03	9.1425e–05
34	9.9737e–01	4.1318e–03	7.1228e–05
38	9.9789e–01	3.3072e–03	5.7050e–05
42	9.9828e–01	2.7069e–03	4.6717e–05
46	9.9856e–01	2.2564e–03	3.8956e–05
50	9.9878e–01	1.9097e–03	3.2979e–05
54	9.9896e–01	1.6372e–03	2.8279e–05
58	9.9910e–01	1.4191e–03	2.4516e–05
62	9.9921e–01	1.2418e–03	2.1457e–05
66	9.9930e–01	1.0958e–03	1.8936e–05
70	9.9938e–01	9.7415e–04	1.6835e–05
74	9.9945e–01	8.7166e–04	1.5065e–05
78	9.9950e–01	7.8454e–04	1.3560e–05
82	9.9955e–01	7.0985e–04	1.2270e–05
86	9.9959e–01	6.4535e–04	1.1156e–05
90	9.9962e–01	5.8925e–04	1.0187e–05
94	9.9966e–01	5.4017e–04	9.3383e–06
98	9.9968e–01	4.9697e–04	8.5918e–06

C_1 = a dimensionless constant.

C_2 = a dimensionless constant.

σ = standard deviation of the approximation error.

$2\pi\sigma/D\Delta P$ = standard deviation of the approximation error normalized with respect to the average torque.

$2\pi\sigma/D\omega$ = standard deviation of the approximation error normalized with respect to the average flow rate.

Table 3. Nondimensional parameter values for $n_e = 4k$.

Number of pistons	C_1	C_2	$2\pi\sigma/D\Delta P$ or $2\pi\sigma/D\omega$
4	7.9791e–01	3.1740e–01	4.2129e–03
8	9.5176e–01	7.5756e–02	1.2421e–03
12	9.7874e–01	3.3390e–02	5.6651e–04
16	9.8808e–01	1.8727e–02	3.2147e–04
20	9.9238e–01	1.1970e–02	2.0657e–04
24	9.9471e–01	8.3062e–03	1.4376e–04
28	9.9612e–01	6.0998e–03	1.0576e–04
32	9.9703e–01	4.6688e–03	8.1039e–05
36	9.9765e–01	3.6882e–03	6.4067e–05
40	9.9810e–01	2.9871e–03	5.1916e–05
44	9.9843e–01	2.4684e–03	4.2919e–05
48	9.9868e–01	2.0740e–03	3.6072e–05
52	9.9887e–01	1.7671e–03	3.0742e–05
56	9.9903e–01	1.5236e–03	2.6511e–05
60	9.9915e–01	1.3271e–03	2.3096e–05
64	9.9926e–01	1.1664e–03	2.0301e–05
68	9.9934e–01	1.0332e–03	1.7985e–05
72	9.9941e–01	9.2155e–04	1.6043e–05
76	9.9947e–01	8.2709e–04	1.4399e–05
80	9.9952e–01	7.4643e–04	1.2996e–05
84	9.9957e–01	6.7703e–04	1.1788e–05
88	9.9961e–01	6.1687e–04	1.0741e–05
92	9.9964e–01	5.6439e–04	9.8280e–06
96	9.9967e–01	5.1833e–04	9.0263e–06
100	9.9970e–01	4.7769e–04	8.3188e–06

C_1 = a dimensionless constant.

C_2 = a dimensionless constant.

σ = standard deviation of the approximation error.

$2\pi\sigma/D\Delta P$ = standard deviation of the approximation error normalized with respect to the average torque.

$2\pi\sigma/D\omega$ = standard deviation of the approximation error normalized with respect to the average flow rate.

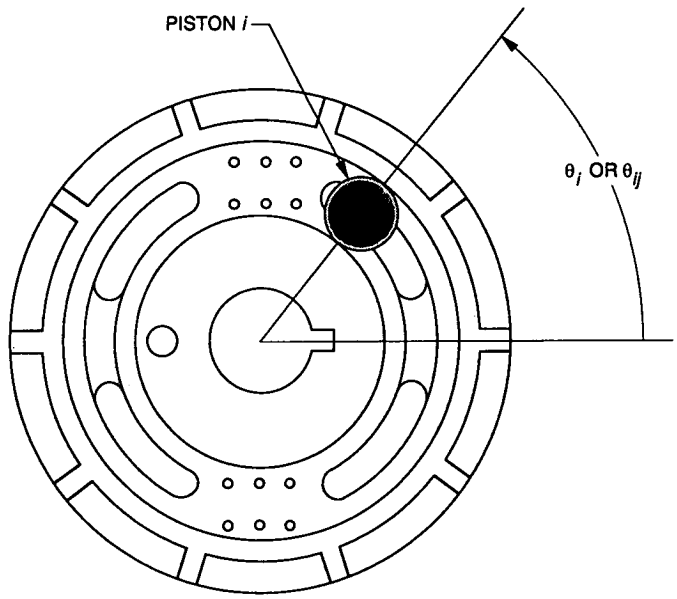


Fig. 1. The valve plate sealing surface of a bent-axis hydraulic piston motor and the piston angle relative to the valve plate.

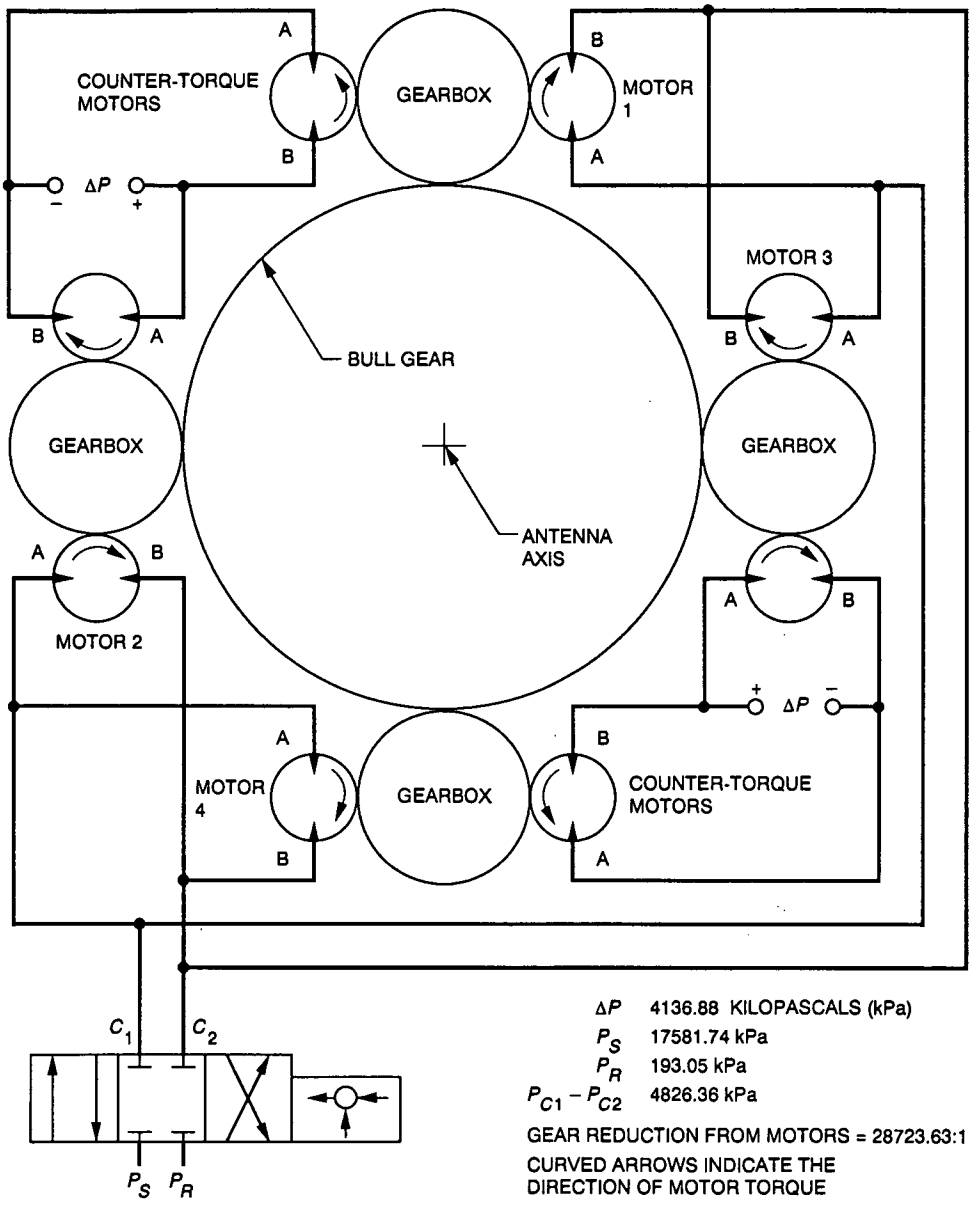


Fig. 2. System configuration of the 70-m antenna azimuth axis.

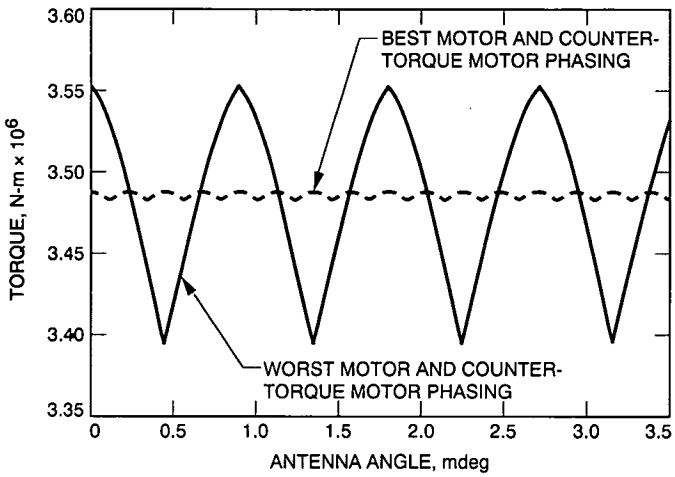


Fig. 3. The 70-m antenna azimuth axis torque as a function of the antenna angular position for the best and worst phasing of the hydraulic motors and counter-torque motors. The differential pressure is 4826.36 kilopascals (kPa) for the drive motors and 4136.88 kPa for the counter-torque motors.

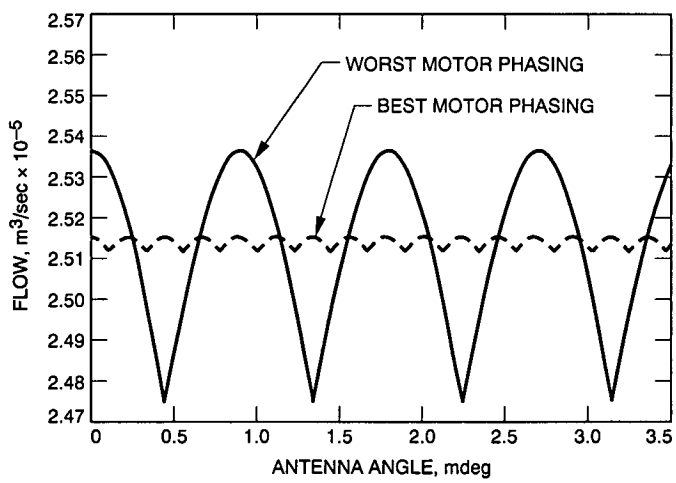


Fig. 4. Total instantaneous motor flow rate through the 70-m antenna azimuth hydraulic drive motors as a function of the antenna angular position for the best and worst phasing of the hydraulic motors. The angular velocity of the antenna is two millidegrees per second, which corresponds to a one radian per second angular velocity of the hydraulic motor shafts.

Helium Compressors for Closed-Cycle, 4.5-Kelvin Refrigerators

T. R. Hanson
Radio Frequency and Microwave Subsystems Section

An improved helium compressor for traveling-wave maser and closed-cycle refrigerator systems was developed and is currently being supplied to the DSN. This new 5-hp compressor package is designed to replace the current 3-hp DSN compressors. The new compressor package was designed to retrofit into the existing 3-hp compressor frame and reuse many of the same components, therefore saving the cost of documenting and fabricating these components when implementing a new 5-hp compressor.

I. Introduction

The DSN currently uses 30 1-W 4.5-K cryogenic systems to cool traveling-wave masers (TWM's). The systems use 3-hp reciprocating compressors to supply helium gas at the required pressures. The compressors exhibit a mean time between failure (MTBF) of approximately 5000 hours and require approximately 18 work hours to repair. Although this performance has been considered adequate, the DSN has desired that the design be upgraded to (1) reduce overhaul labor, (2) increase MTBF, and (3) increase helium mass flow to accommodate larger closed-cycle refrigerators (CCR's). Future maser systems will be based on a new 2-W capacity refrigerator that requires more mass flow than the current DSN compressor can deliver.

The improved DSN helium compressor (Fig. 1) is a 5-hp, two-stage, reciprocating, oil-lubricated unit, 58 in. long, 24 in. wide, 32.25 in. high, and 720 lb in weight. It is designed to supply high-pressure helium gas to 4.5-K Gifford–McMahon/Joule Thompson (GM/JT) CCR's. The compressor is a modified Dunham Bush model 50 PCF 5-hp unit with a free air displacement of 10.7 l/sec (22.6 standard ft^3/min). This increase in displacement over the current 3-hp DSN units of 4 l/sec (8 standard ft^3/min) allows the GM engine to operate at a higher pressure ratio (2.0 MPa supply, 0.5 MPa return) and a higher mass flow, resulting in a 40-percent average increase in maser/refrigerator third-stage (4.5-K stage) cooling capacity. In spite of this increase in mass flow, the new compressor operates at lower temperatures and requires no more AC input power than the 3-hp unit it is designed to replace. Improvements in the oiling system, heat exchanger, electrical system, and hardware are also designed to increase reliability and improve maintainability of the system. Table 1 shows a performance comparison between the 3-hp and 5-hp compressors. The operation of the 5-hp

compressor will be described by presenting the flow diagram of a simple one-stage compressor, then a simple two-stage compressor, and finally a simplified flow diagram of the 5-hp compressor.

II. Description

A simplified block diagram of a typical single-stage compressor is shown in Fig. 2. The purpose of this is to compress gas from low pressure to high pressure while maintaining required purity. Oil injection to the compressor is required to lubricate and control temperatures of moving parts. The high-pressure gas and oil stream flows through a heat exchanger to remove the heat of compression. From the heat exchanger, the gas and oil stream flows through an oil separator assembly that returns the oil to the compressor and supplies the refrigerator with high-purity, high-pressure gas.

A two-stage compressor is required to operate a GM/JT system. A simplified block diagram of a two-stage compressor is shown in Fig. 3. A two-stage compressor cascades a single-stage compressor with another to provide two stages of compression. The first stage compresses gas from the low-pressure Joule-Thompson return circuit to an intermediate pressure that corresponds to the return pressure from the Gifford-McMahan expansion engine. Compression heat is removed in the first-stage heat exchanger. The oil and gas stream output from the first stage is combined with the gas returning from the expansion engine. Both streams are then recompressed in the second compressor stage to the required supply pressure. The heat of compression is then removed by the second-stage heat exchanger. Oil from both stages is removed by a common oil separator and returned to the compressor crankcase. High-pressure gas is routed to the supply (input) sides of both the GM expansion engine and the JT circuit in the closed-cycle refrigerator.

A schematic diagram of the 5-hp DSN two-stage compressor is shown in Fig. 4. Additional components have been added to the basic two-stage system discussed above to control and monitor system pressures, meter oil injection to both compressor stages, and improve compression heat removal. The compressor used for the DSN configuration is a Dunham Bush V-4, air-cooled, reciprocating, hermetic R-22 refrigeration unit. The unit is then modified at JPL for a two-stage helium closed-cycle application. Operating pressures are controlled by a low-pressure regulator, a high-pressure regulator, and a storage tank. The low-pressure regulator governs the pressure at the first-stage intake of the compressor. The high-pressure regulator governs the second-stage supply pressure by returning the gas not required by the CCR system to the storage tank. The storage tank contains a reserve supply of helium gas. Gas is withdrawn from the storage tank during refrigerator cooldown. The heat exchanger is a fan-cooled gas-to-air type with two sections, one for each compressor stage. System lubrication is supplied by an oil pump for the second stage and by crankcase pressure for the first stage. Metering orifices at both compressor stages govern the amount of oil injected.

III. Reliability and Maintainability

The new helium compressor design has evolved from the experience gained with past prototype and test compressors. The low MTBF (5000 hours) of current 3-hp compressors is primarily attributed to compressor component overheating and reed valve failure. A larger heat exchanger and the placement of internal oil-injection nozzles just above each reed valve have reduced operating temperatures in the compressor. Both first- and second-stage valve plates and reed valves have also been redesigned and have proven to be stronger than earlier designs.

Three major design changes were incorporated into the new model to improve compressor maintainability. The heat exchanger is mounted externally on the compressor frame and can be removed and repaired or replaced without interfering with other compressor components. All major electrical components are also externally mounted at the opposite end of the compressor frame in an electrical box. Grouping and relocating these electrical components have reduced maintenance time when troubleshooting and repairing the electrical system. The main compressor motor is now mounted to the compressor frame by means of a caster and rail system. A compressor motor handling fixture attaches to existing screw holes in the compressor frame and allows the compressor motor to be rolled out and back into the compressor frame without manual lifting or other handling apparatus. This new compressor motor handling system (Fig. 5) eliminates the necessity of removing other compressor components and helium piping to gain access for overhead hoisting systems.

IV. Performance

The 5-hp helium compressor significantly increases the 4.5-K cooling capacity of the existing DSN refrigerator. This increase is due to an increase in the GM engine pressure ratio and therefore the capacity of the refrigerator's first and second stages. The increase in the 4.5-K refrigerator capacity using the 5-hp compressor is

even more substantial when the first stage of the refrigerator is loaded heavily due to large radiation shields and RF input lines, such as in the DSN Block V 2.3-GHz (S-Band) TWM/CCR, the DSN Block II A 8.4-GHz (X-Band) TWM, and the 23-GHz Maser/CCR.

Figure 6 shows a comparison of a standard DSN refrigerator 4.5-K-stage capacity when using a standard DSN 3-hp compressor or the new 5-hp compressor. The 5-hp compressor provides a 230-percent increase in 4.5-K capacity when an additional 20-W load is applied to the refrigerator first stage.

The additional JT mass flow available from the larger 5-hp compressor is shown in Fig. 7. The 3-hp compressor is capable of providing the refrigerator with a JT maximum mass flow of 3.5 standard ft^3/min, (at a maximum desired JT return pressure of 3.5 standard ft^3/min) while the 5-hp compressor can provide the refrigerator with a JT mass flow of 5.5 standard ft^3/min.

The first of these newly designed 5-hp compressors was built in 1990. This unit is being used for continuous-operation "life testing" and is completely disassembled and thoroughly inspected for signs of wear and deterioration at 5000-hour intervals. The compressor has accumulated over 20,000 hours of operation and has exhibited no failures or significant symptoms of wear or deterioration to date.

V. Conclusion

The new DSN 5-hp helium compressor, compared with the existing 3-hp design, improves the cooling capacity at all three stages of maser refrigerators. This results in an increase in the refrigerator's tolerance to first- and second-stage GM expander performance degradation and external heat loads (which have been increasing with the recent development of higher performance masers, some of which are incorporating cryogenic feed components). In addition, the increased helium mass flow that the new compressor can provide will be required for operation of future 2-W DSN refrigerators. This compressor has been designed to retrofit into the current 3-hp compressor frames and share many existing 3-hp compressor components, therefore saving the cost of documenting and fabricating these components when implementing a new 5-hp compressor.

The new compressor requires no more AC input power to operate and runs at lower operating temperatures than the 3-hp units. The design improvements to the heat exchanger, valve plates, and oil-injection system should result in an improvement in both DSN closed-cycle refrigerator and compressor MTBF. Packaging improvement has significantly reduced the labor required to repair and replace major components. The 5-hp life-test compressor has already surpassed 20,000 hours of operation with virtually no measurable signs of wear or failures.

Acknowledgments

The author thanks Mike Britcliffe and Robert Clauss for providing technical assistance, guidance, and constructive criticism during design and development. Both deserve equal credit for sharing their combined years of experience with closed-cycle helium refrigeration systems.

The author also thanks Steve Montanez of Planning Research Corporation for providing superb craftsmanship that went into the fabrication of the prototype compressor.

Table 1. Performance comparison of 3-hp and 5-hp compressors.

Specification	3-hp compressor	5-hp compressor
Bore, mm (in.)	48 (1 7/8)	56 (2 3/16)
Stroke, mm (in.)	35 (1 3/8)	38 (1 1/2)
Cycles	2	4
RPM	1750	1725
Amps	11.0	9.5
Compressor rating (air), l/sec (scfm[a])	3.6 (7.66)	10.7 (22.6)
Average motor temp., deg C (deg F)	119 (246)	59 (138)
Average compression temp., deg C (deg F)	77 (170)	72 (162)
Maximum JT mass flow (at 300 psig supply) l/sec (scfm)	1.4 (3)	2.6 (5.5)

[a] scfm = standard cubic feet per min.

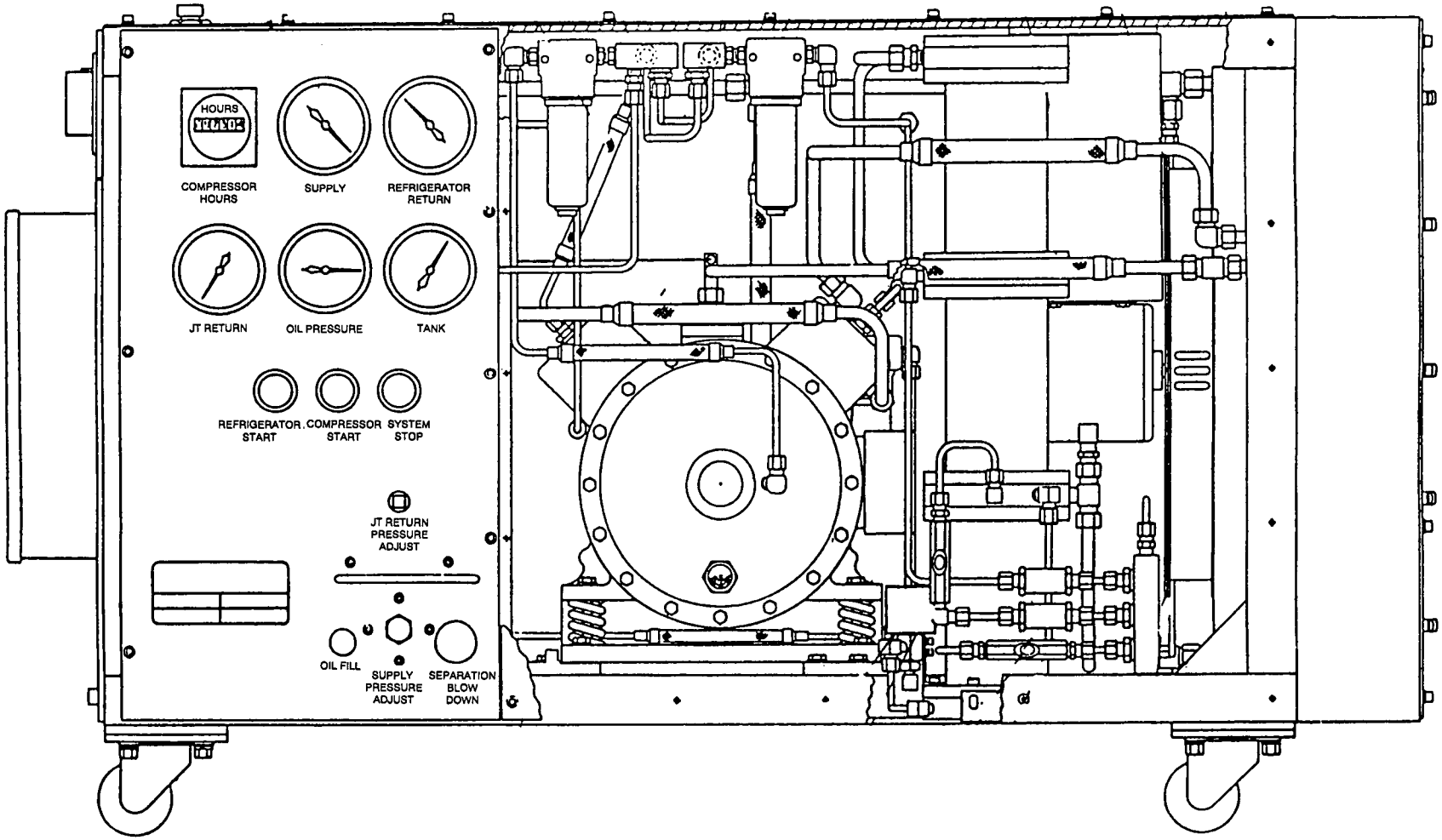


Fig. 1. 5-hp DSN helium compressor.

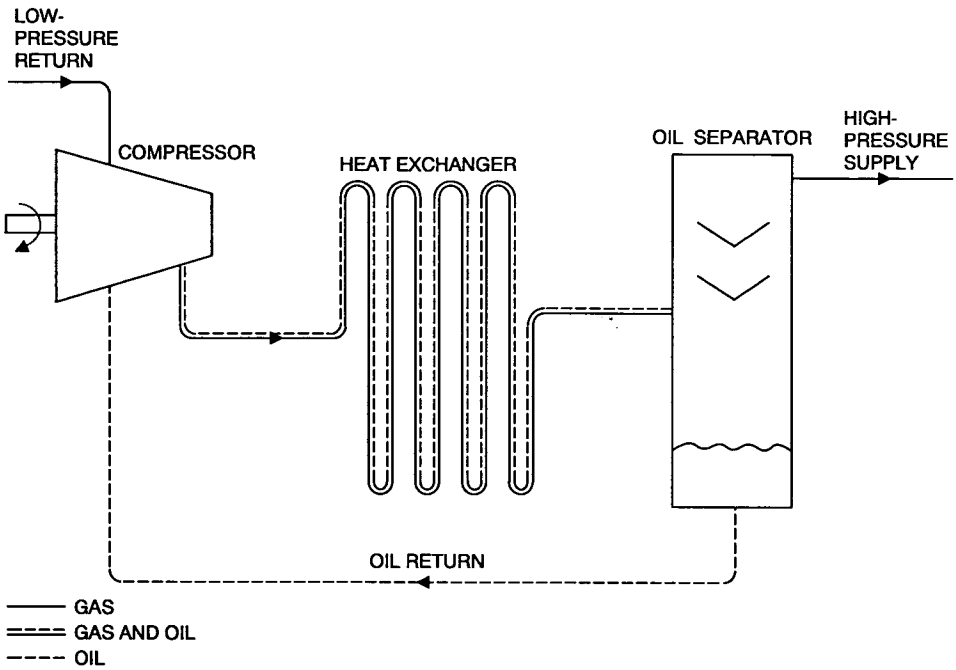


Fig. 2. Simplified single-stage compressor schematic.

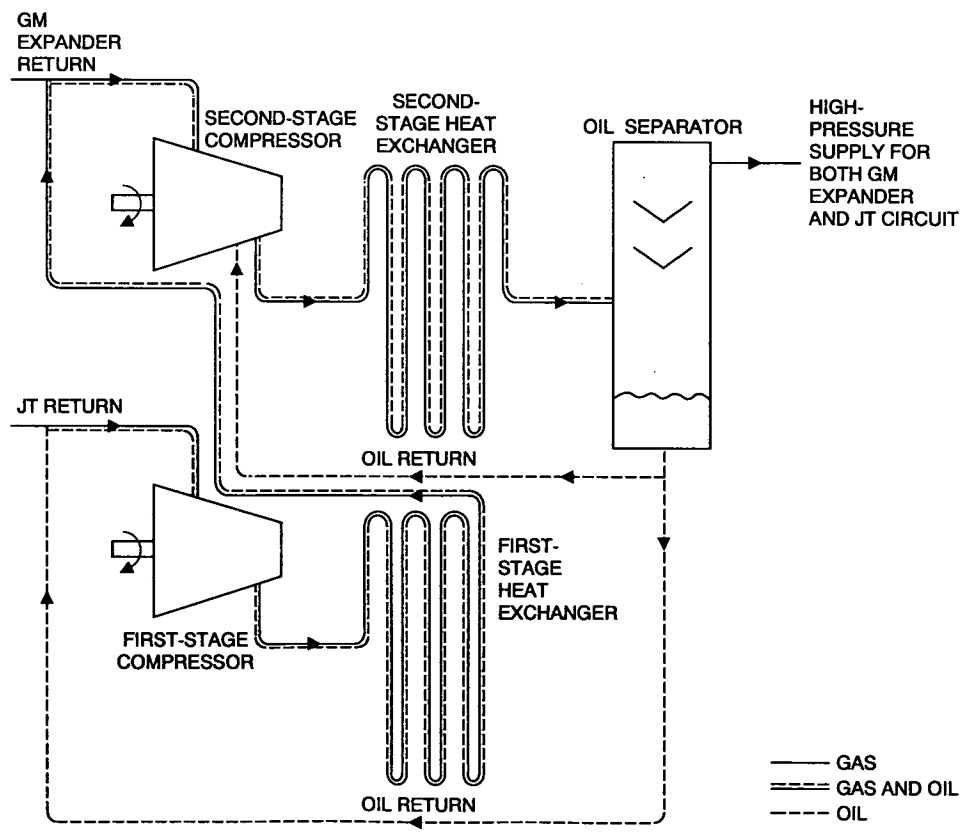


Fig. 3. Simplified two-stage compressor schematic.

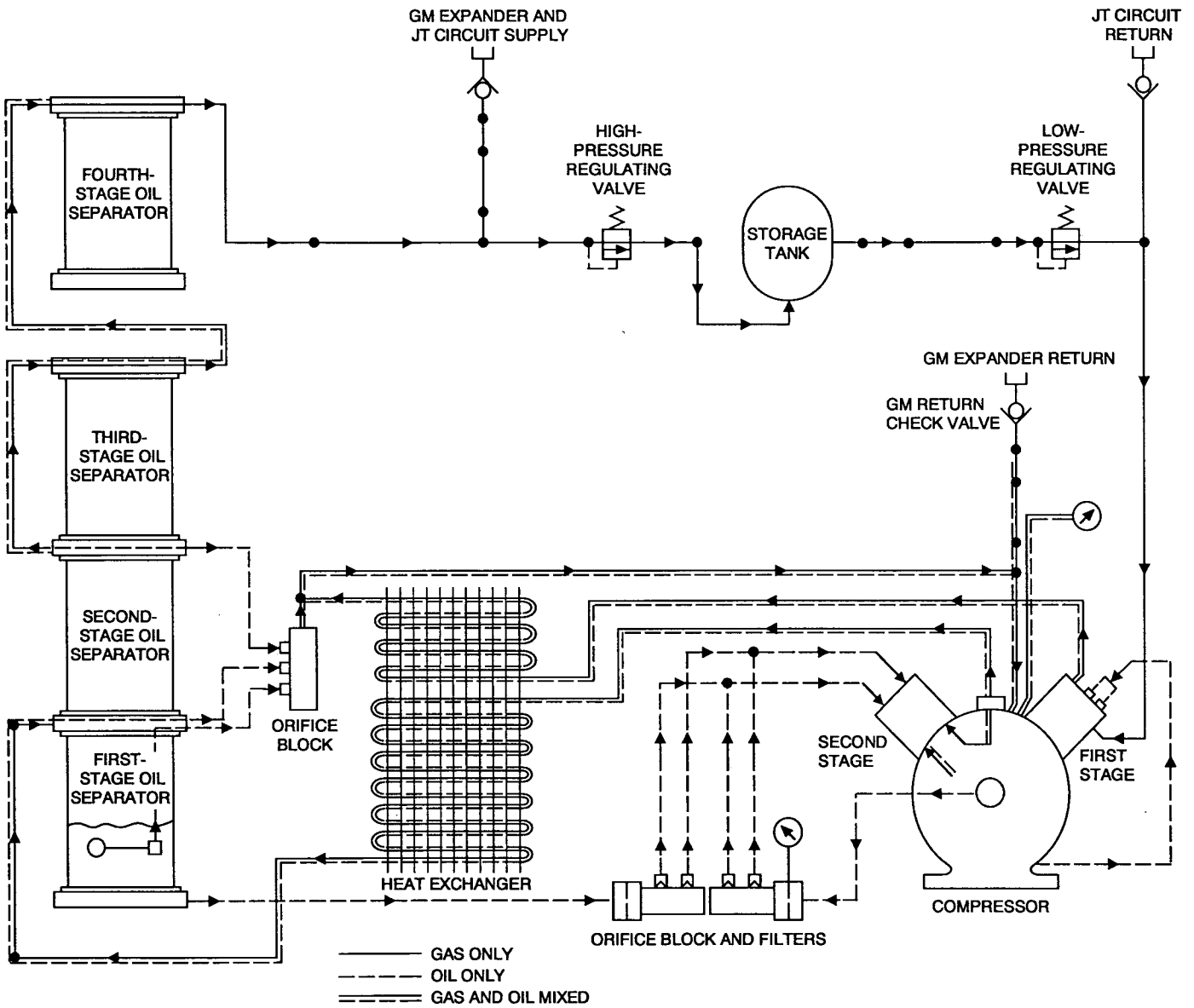


Fig. 4. Simplified 5-hp DSN two-stage compressor schematic.

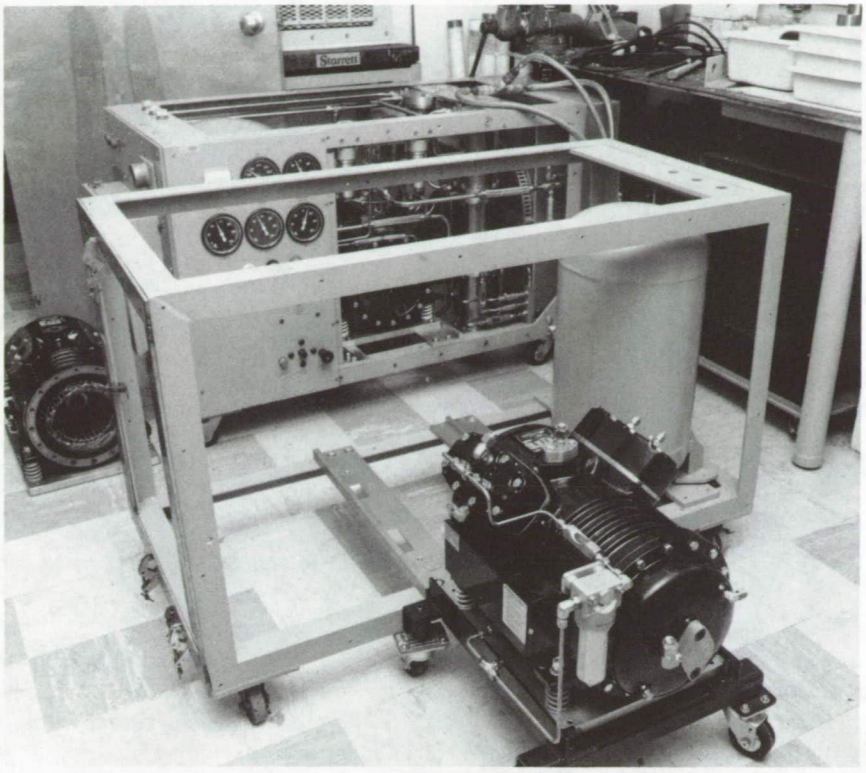


Fig. 5. The new compressor motor handling system.

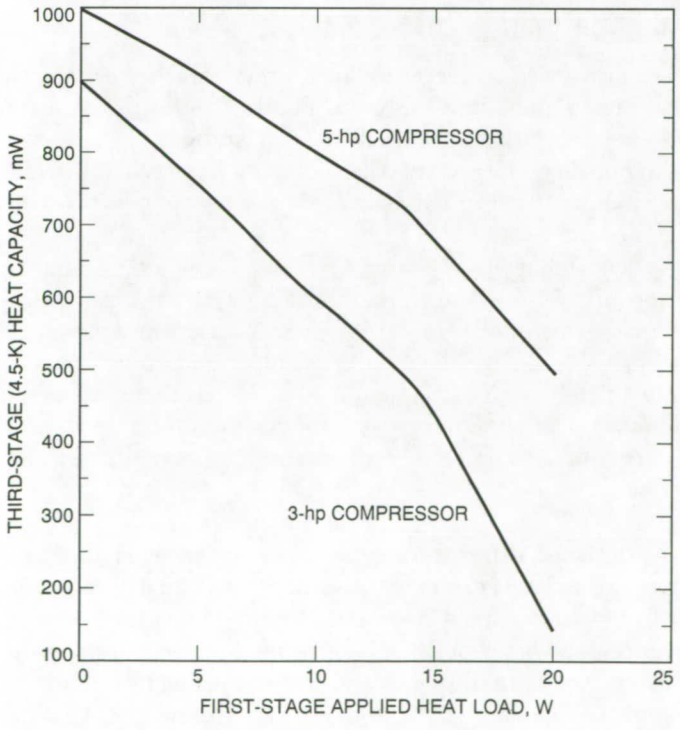


Fig. 6. DSN refrigerator third-stage (4.5-K) heat capacity with load applied to refrigerator first stage.

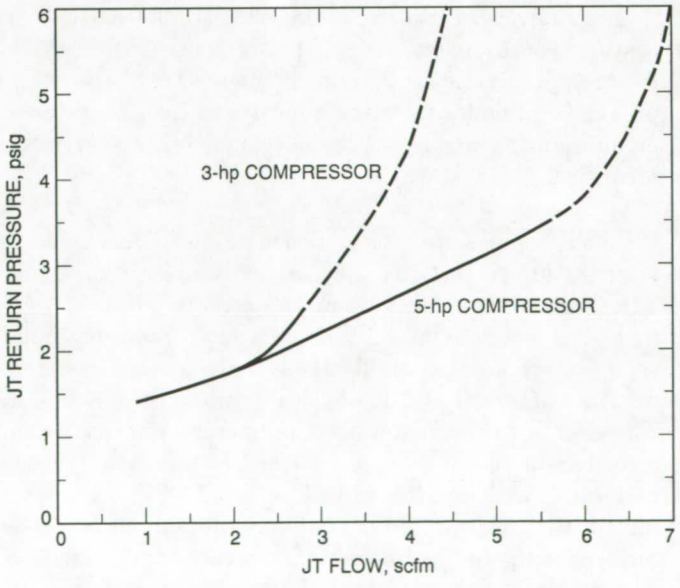


Fig. 7. Refrigerator JT return flow comparison.

Minimizing Noise-Temperature Measurement Errors

C. T. Stelzried
TDA Technology Development Section

An analysis of noise-temperature measurement errors of low-noise amplifiers has been performed. Results of this analysis can be used to optimize measurement schemes for minimum errors. For the cases evaluated, the effective noise temperature (Te) of a Ka-band maser can be measured most accurately by switching between an ambient and a 2-K cooled load without an isolation attenuator. A measurement accuracy of 0.3 K was obtained for this example.

I. Introduction

An analysis of low-noise amplifier (LNA) noise-temperature measurement errors is performed based on the Y-factor (power ratio measurement) method [1]. The results provide a guide for selecting a measurement scheme based on minimizing errors. The individual error sources are identified.

Load (input source termination) high-frequency noise-temperature corrections [2,3] are not addressed in this article because the effect upon the error analysis is minimal. For example, at 32 GHz, the high-frequency noise-temperature correction effectively reduces the noise temperature of a load at 2-K physical temperature by 0.67 K. The effect of this on the noise-temperature measurement error for the optimum case analyzed in this article, case 4 in Table 1, with no attenuator loss, is 0.03 K. This is determined using the equations and computer program as described, with the load noise temperature reduced 0.67 K. This small error is acceptable for the purpose of this article.

The effects of microwave component mismatch errors and receiver nonlinearity are not analyzed in this article.

II. Analysis

Two noise-temperature measurement schemes, shown in Figs. 1 and 2, are analyzed. Figure 1 shows the Y factor measured with the noise source on and off. The scheme in Fig. 2 requires that the Y factor be measured while switching between two loads at different temperatures.

The following equations were used in a Supercalc 4 spreadsheet computer program to analyze the noise-temperature measurement errors as a function of attenuator setting L, in dB. The measurement errors are determined by differencing Te using nominal and perturbed parameter values. The amplifier noise-temperature differences (measurement errors) are shown as summed (worst-case) and rms.

Both the Y-factor measurement and the attenuator values are assigned errors of the form dB, A, and dB, B, where A represents a fixed error and B represents a fractional or percentage error per dB. The terms used in the program associated with Eqs. (1) and (2) are defined at the bottoms of Figs. 3, 5, 7, 9, and 11. (Figures 4, 6, 8, 10, and 12 plot the data presented in each preceding figure.) For Fig. 1, using the noise source,

$$Y = \frac{(T/L) + (Tn/L) + TL + Te}{(T/L) + TL + Te} \quad (1)$$

and

$$Te = \frac{Te}{L(Y-1)} - (T/L) - Tp(1 - 1/L) \quad (2)$$

For Fig. 2, using two loads at different temperatures,

$$Y = \frac{(T2/L) + TL + Te}{(T1/L) + TL + Te} \quad (3)$$

and

$$Te = \frac{T2 - T1Y}{L(Y-1)} - Tp(1 - 1/L) \quad (4)$$

where for both cases,

$$TL = Tp(1 - (1/L))$$

III. Results

Table 1 shows selected results for Eq. (2), Figs. 3 and 5 (cases 1 and 2), and Eq. (4), Figs. 7, 9, and 11 (cases 3, 4, and 5). Assumed parameter and error values are indicated in the figures. A 1-sec measurement integration time, 50-MHz amplifier bandwidth, designated amplifier gain change (the higher gain change due to a different load match effect for $L = 0$ dB), and 5-percent noise-source accuracy parameters are used. The noise source might be calibrated more accurately by a standards laboratory. The approximate value of L, dB, resulting in the minimum Te-rms measurement error, is indicated for each case, except case 5.

A comparison of cases 1 and 2 with cases 3, 4, and 5 shows that with the accuracies assumed, two loads are preferable to one load and a noise source. A comparison of cases 3 and 5 shows that the cold load should be as cold as possible. A comparison of cases 4 and 5 shows that reducing the amplifier gain change (DYG) obtained with an attenuator between the amplifier and loads does not result in a net advantage.

Case 4, Fig. 9, using an ambient and a 2-K cold load with no attenuator, provides the most accurate noise-temperature measurement result of 0.3-K rms.

Averaged measurements with several configurations, including different attenuations, $L = 10$ dB (case 5) for example, to reduce the amplifier gain change due to a different input load match, will reduce bias errors and indicate the magnitude of the error.

IV. Conclusion

For the schemes analyzed in this article with assumed parameter and measurement accuracies, it is preferable to use two loads as opposed to one load and a noise source. A fixed attenuator between the amplifier and input loads has no advantage except as a measurement consistency check. The most accurate case evaluated, with input loads at an ambient and a 2-K temperature and no attenuator, provides an LNA noise-temperature measurement accuracy result of 0.3-K rms.

Acknowledgments

R. Clauss and J. Shell contributed numerous useful suggestions for this article.

References

[1] C. Stelzried, *The Deep Space Network—Noise Temperature Concepts, Measurements, and Performance*, JPL Publication 82-33, Jet Propulsion Laboratory, Pasadena, California, September 15, 1982.

[2] C. Stelzried, R. Clauss, W. Rafferty, and S. Petty, "DSN G/T_{op} and Telecommunications System Performance," *TDA Progress Report 42-108*, vol. October–December 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 271–278, February 15, 1991.

[3] C. Stelzried, "Corrections of High-Frequency Noise-Temperature Inaccuracies," *TDA Progress Report 42-111*, vol. July–September 1992, Jet Propulsion Laboratory, Pasadena, California, pp. 169–277, November 15, 1992.

Table 1. Summary of Supercalc 4 computer programs NOISE1ND and NOISE1LD analysis of an LNA noise-temperature measurement delta or error (*DTe*).

Case	Figure	Method	Configuration	L, dB	DTe (rms), K
1	3, 4	1 input load and noise source	$T = 300$ K $Tn = 1000$ K $DYG = 0.01$	20	1.0
2	5, 6	1 input load and noise source	$T = 300$ K $Tn = 60{,}000$ K $DYG = 0.02$	20	1.1
3	7, 8	2 input loads	$T1 = 80$ K $T2 = 300$ K $DYG = 0.01$	10	0.7
4	9, 10	2 input loads	$T1 = 2$ K $T2 = 300$ K $DYG = 0.02$	0	0.3
5	11, 12	2 input loads	$T1 = 2$ K $T2 = 300$ K $DYG = 0.01$	10	0.5

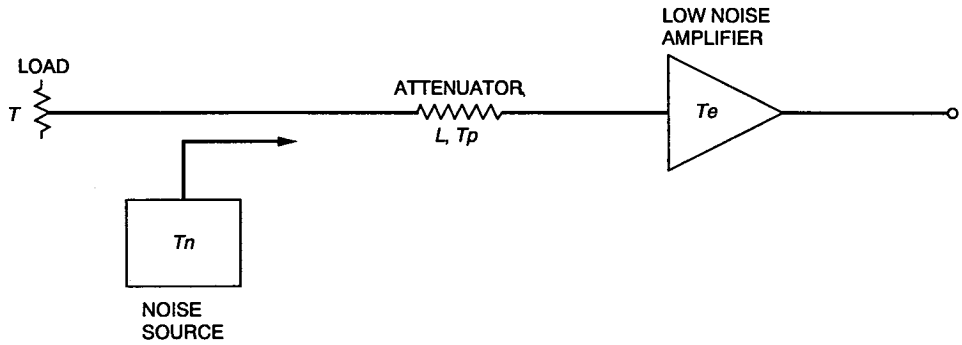


Fig. 1. Low-noise amplifier measurement scheme using a load at temperature *T*, a noise source, and a fixed attenuator.

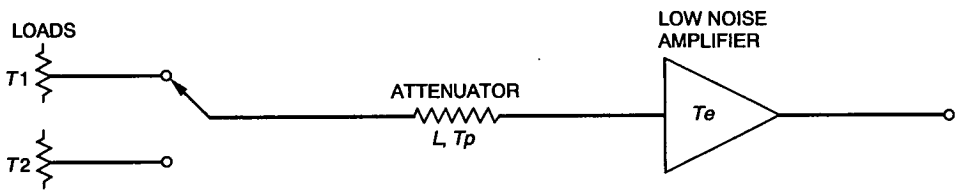


Fig. 2. Low-noise amplifier measurement scheme using two loads at temperatures *T*1 and *T*2 and a fixed attenuator.

```
INPUT
      T=  300      Tn=  1000      Tp=     2
     DT=   .1     DTn=    50     DTp=    .1      Te=     4
DYLDB,A=  .01 DLDB,A=   .01   B,MHZ=    50     DYG=   .01
DYLDB,B=  .01 DLDB,B=   .03   T,SEC=     1
    L,DB           0     3      10     15      20     23     30
       L           1 1.995      10  31.62     100 199.53   1000
RESULTS      ----------   Te ERROR,K    ----------
      DL     .013800 .1366 .4133527 .6030 .7862374 .89317 1.1342
      DT  .1      .1 .0501     .01 .0032    .001 .00050 .00010
     DTn   50    15.2 7.768     1.79 .7712    .449 .37468  .3149
     DTp   .1       0 .0499     .09 .0968    .099 .09950  .0999
     DYL    6.59566 3.356 .7500375 .3047 .1645480 .13678 .16922
     DYN    .112082 .0575 .0137455 .0065 .0048182 .00528 .01297
     DYG    7.72680 3.966 .9466179 .4457 .3284142 .35618 .80217
     SUM    29.7483 15.38 4.013754 2.231 1.833018 1.8661 2.5335
     RMS    18.2830 9.347 2.200452 1.122 .9821036 1.0458 1.4380
NOMINAL CALC
      TL           0 .9976      1.8 1.937     1.98 1.9900  1.998
       Y     4.28947 4.226 3.793296 3.050 2.113586 1.6688 1.1588
     YDB     6.32404 6.259 5.790167 4.843 3.250199 2.2241 .64001
      Te           4     4        4     4        4     4      4
ERROR CALC
 L+DL,DB         .01   3.1    10.31 15.46    20.61   23.7  30.91
    L+DL     1.00231 2.042 10.73989 35.16 115.0800 234.42 1233.1
      TL     .004600 1.020 1.813778 1.943 1.982621 1.9915 1.9984
      Te     3.98620 3.863 3.586647 3.397 3.213763 3.1068 2.8658
    T+DT       300.1 300.1    300.1 300.1    300.1  300.1  300.1
      Te         3.9 3.950     3.99 3.997    3.999 3.9995 3.9999
  Tn+DTn        1050  1050     1050  1050     1050   1050   1050
      Te        19.2 11.77     5.79 4.771    4.449 4.3747 4.3149
  Tp+DTp         2.1   2.1      2.1   2.1      2.1    2.1    2.1
      TL           0 1.048     1.89 2.034    2.079 2.0895 2.0979
      Te           4 3.950     3.91 3.903    3.901 3.9005 3.9001
Y+DYL,DB     6.39728 6.332 5.858069 4.902 3.292701 2.2564 .65641
   Y+DYL     4.36243 4.297 3.853070 3.092 2.134372 1.6813 1.1632
      Te     -2.5957 .6438 3.249962 3.695 3.835452 3.8632 3.8308
   Y+DYN     4.29069 4.227 3.794369 3.051 2.114184 1.6693 1.1591
      Te     3.88792 3.942 3.986254 3.994 3.995182 3.9947 3.9870
   Y+DYG     4.37526 4.311 3.869162 3.111 2.155857 1.7022 1.1820
      Te     -3.7268 .0337 3.053382 3.554 3.671586 3.6438 3.1978
DEFINITIONS
L=ATTEN LOSS                Y=Y FACTOR
DL=DELTA L                  DYL=DELTA Y NON-LINEARITY
TL=TEMP CONTRI OF L         DYN=DELTA Y, RADIOMETER NOISE (T,B)
T=INPUT LOAD TEMP           DYG=DELTA Y, RADIOMETER GAIN CHANGE G
DT=DELTA T                  Te=LNA NOISE TEMPERATURE
Tp=PHY TEMP OF L            Tn=NOISE SOURCE CONTRIBUTION
DTp=DELTA Tp
```

Fig. 3. Supercalc 4 computer program NOISE1ND printout of the analysis of Fig. 1, showing the measurement error as a function of attenuator loss *L*, dB, and assumed input parameter errors; noise source = 1000 K.

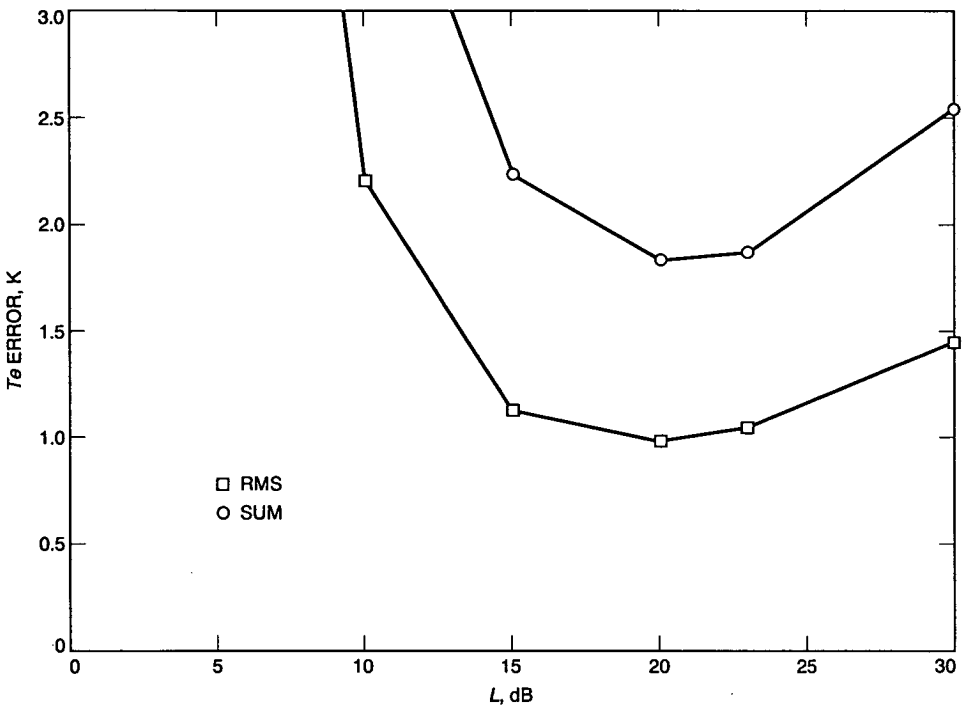


Fig. 4. Plot of the data in Fig. 3.

C-4

```
INPUT
      T=  300      Tn= 60000      Tp=     2
     DT=   .1     DTn=  3000     DTp=    .1      Te=      4
DYLDB,A=  .01 DLDB,A=   .01   B,MHZ=    50     DYG=    .02
DYLDB,B=  .01 DLDB,B=   .03   T,SEC=     1
    L,DB           0     3      10     15      20      23      30
       L           1 1.995      10 31.62     100 199.53    1000
RESULTS       -----------   Te ERROR,K   ---------------------
      DL      .013800 .1366 .4133527 .6030 .7862374 .89317 1.1342
      DT   .1      .1 .0501      .01 .0032     .001 .00050 .00010
     DTn 3000    15.2 7.768     1.79 .7712     .449 .37468  .3149
      DTp   .1      0 .0499      .09 .0968     .099 .09950  .0999
      DYL      16.4054 8.356 1.878164 .7655 .3961832 .29695 .17706
      DYN      .086395 .0442 .0101833 .0044 .0025772 .00217 .00197
      DYG      11.7493 6.005 1.384821 .5979 .3503522 .29511 .26658
      SUM      43.5549 22.41 5.576521 2.842 2.084350 1.9621 1.9947
      RMS      25.2634 12.89 2.971278 1.382 1.053227 1.0599 1.2239
NOMINAL CALC
       TL           0 .9976       1.8 1.937      1.98 1.9900  1.998
        Y      198.368 194.6 168.5978 124.0 67.81514 41.130 10.527
      YDB      22.9747 22.89 22.26852 20.93 18.31327 16.142 10.223
       Te           4     4         4     4        4      4      4
ERROR CALC
 L+DL,DB          .01   3.1     10.31 15.46    20.61   23.7  30.91
    L+DL      1.00231 2.042 10.73989 35.16 115.0800 234.42 1233.1
      TL      .004600 1.020 1.813778 1.943 1.982621 1.9915 1.9984
      Te      3.98620 3.863 3.586647 3.397 3.213763 3.1068 2.8658
    T+DT        300.1 300.1     300.1 300.1    300.1  300.1  300.1
      Te          3.9 3.950      3.99 3.997    3.999 3.9995 3.9999
  Tn+DTn        63000 63000     63000 63000    63000  63000  63000
      Te         19.2 11.77      5.79 4.771    4.449 4.3747 4.3149
  Tp+DTp          2.1   2.1       2.1   2.1      2.1    2.1    2.1
      TL            0 1.048       1.89 2.034    2.079 2.0895 2.0979
      Te            4 3.950       3.91 3.903    3.901 3.9005 3.9001
Y+DYL,DB      23.2145 23.13 22.50120 21.15 18.50640 16.313 10.335
   Y+DYL      209.627 205.6 177.8772 130.4 70.89898 42.785 10.802
      Te      -12.405 -4.36 2.121836 3.234 3.603817 3.7031 3.8229
   Y+DYN      198.425 194.6 168.6455 124.1 67.83433 41.141 10.530
      Te      3.91360 3.956 3.989817 3.996 3.997423 3.9978 3.9980
   Y+DYG      206.303 202.3 175.3417 129.0 70.52775 42.775 10.948
      Te      -7.7493 -2.00 2.615179 3.402 3.649648 3.7049 3.7334
DEFINITIONS
L=ATTEN LOSS                Y=Y FACTOR
DL=DELTA L                  DYL=DELTA Y NON-LINEARITY
TL=TEMP CONTRI OF L         DYN=DELTA Y, RADIOMETER NOISE (T,B)
T=INPUT LOAD TEMP           DYG=DELTA Y, RADIOMETER GAIN CHANGE G
DT=DELTA T                  Te=LNA NOISE TEMPERATURE
Tp=PHY TEMP OF L            Tn=NOISE SOURCE CONTRIBUTION
DTp=DELTA Tp
```

Fig. 5. Supercalc 4 computer program NOISE1ND printout of the analysis of Fig. 1, showing the measurement error as a function of attenuator loss *L*, dB, and assumed input parameter errors; noise source = 60,000 K. The higher gain change (*DYG*) than that in Fig. 3 is appropriate for *L* = small attenuation.

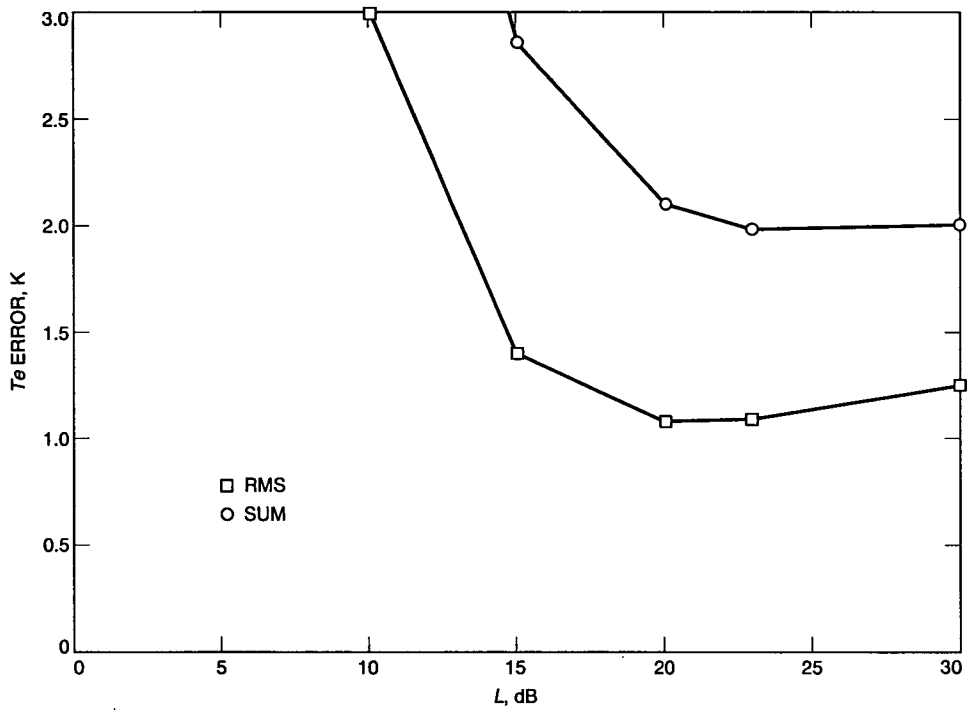


Fig. 6. Plot of the data in Fig. 5.

```
INPUT
     T2=    300      T1=      80      Tp=       2
    DT2=     .1     DT1=       1     DTp=     .01      Te=       4
DYLDB,A=     .01 DLDB,A=     .01   B,MHZ=      50     DYG=     .01
DYLDB,B=     .01 DLDB,B=     .03   T,SEC=       1
    L,DB              0       3      10      15      20      23      30
       L              1 1.99526      10 31.6228     100 199.526    1000
RESULTS         ------------   Te ERROR ,K      ----------------------
      DL         .013800 .136577 .4133527 .603015 .7862374 .893172 1.13423
     DT2     .1 .038182 .020497 .0062727 .003848 .0030818 .002905 .002763
     DT1      1 1.38182 .706154 .1627273 .070107 .0408182 .034062 .028627
     DTP     .01       0 .004988    .009 .009684    .0099 .009950  .00999
     DYL         1.73695 .919911 .2622773 .153609 .1389508 .165092 .430334
     DYN .00014 .032817 .017963 .0063487 .005306 .0078186 .012261 .048818
     DYG     .01 2.25902 1.23585 .4349709 .359467 .5117196 .764731 2.21293
     SUM         5.46259 3.04194 1.294950 1.20504 1.498526 1.88217 3.86769
     RMS         3.16738 1.70048 .6749003 .722143 .9492989 1.18796 2.52429

 NOMINAL
       L    NOM       1 1.99526      10 31.6228     100 199.526    1000
      TL    NOM       0 .997626     1.8 1.93675    1.98 1.98998   1.998
       Y    NOM 3.61905 3.44522 2.594203 1.82170 1.324484 1.17253 1.03620
     YDB    NOM 5.58594 5.37217 4.140039 2.60478 1.220466 .691231 .154420
      Te    NOM       4       4       4       4       4       4       4

ERROR CALC
 L+DL,DB            .01     3.1    10.31   15.46   20.61    23.7   30.91
    L+DL        1.00231 2.04174 10.73989 35.1560 115.0800 234.423 1233.10
      TL        .004600 1.02044 1.813778 1.94311 1.982621 1.99147 1.99838
      Te        3.98620 3.86342 3.586647 3.39699 3.213763 3.10683 2.86577
  T2+DT2  300.1
      Te        4.03818 4.02050 4.006273 4.00385 4.003082 4.00290 4.00276
  T1+DT1     81
      Te        2.61818 3.29385 3.837273 3.92989 3.959182 3.96594 3.97137
  TP+DTP   2.01
      TL              0 1.00261   1.809 1.94644   1.9899 1.99993 2.00799
      Te              4 3.99501   3.991 3.99032   3.9901 3.99005 3.99001
Y+DYL,DB        5.65180 5.43589 4.191440 2.64082 1.242671 .708144 .165964
   Y+DYL        3.67435 3.49614 2.625089 1.83689 1.331273 1.17710 1.03895
      Te        2.26305 3.08009 3.737723 3.84639 3.861049 3.83491 3.56967
   Y+DYN        3.62007 3.44619 2.594937 1.82222 1.324858 1.17286 1.03649
      Te        3.96718 3.98204 3.993651 3.99469 3.992181 3.98774 3.95118
   Y+DYG        3.69143 3.51412 2.646087 1.85814 1.350973 1.19598 1.05692
      Te        1.74098 2.76415 3.565029 3.64053 3.488280 3.23527 1.78707
DEFINITIONS
L=ATTEN LOSS                                    Tp=PHY TEMP OF L
DL=DELTA L                                      DTp=DELTA Tp
TL=TEMP CONTR OF L                              Te=LNA NOISE TEMP
Y=Y FACTOR                                      T1=COLD LOAD TEMP
DYL=DELTA Y FACTOR NON-LINEARITY                DT1=DELTA T1
DYN=DELTA Y, RADIOMETER NOISE (T,B)             T2=HOT LOAD TEMP
DYG=DELTA Y, RADIOMETER GAIN DELTA G            DT2=DELTA T2
```

Fig. 7. Supercalc 4 computer program NOISE1LD printout of the analysis of Fig. 2, showing the measurement error as a function of attenuator loss *L*, dB, and assumed input parameter errors; *T*1 = 80 K and *T*2 = 300 K.

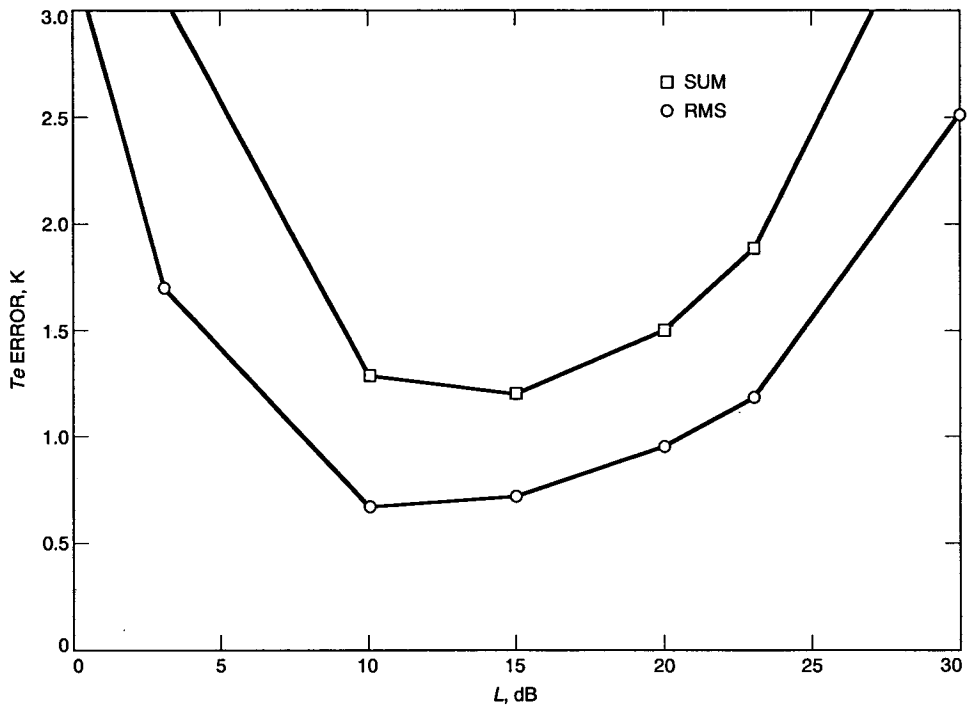


Fig. 8. Plot of the data in Fig. 7.

```
INPUT
     T2=    300     T1=       2      Tp=       2
    DT2=     .1    DT1=     .01     DTp=     .01      Te=       4
DYLDB,A=     .01 DLDB,A=     .01   B,MHZ=      50     DYG=     .02
DYLDB,B=     .01 DLDB,B=     .03   T,SEC=       1
    L,DB              0       3      10       15       20       23       30
       L              1 1.99526      10 31.6228      100 199.526     1000
RESULTS          -------------   Te ERROR ,K      ----------------------
      DL         .013800 .136577 .4133527 .603015 .7862374 .893172 1.13423
     DT2     .1 .002013 .002013 .0020134 .002013 .0020134 .002013 .002013
     DT1    .01 .010201 .005213 .0012013 .000518 .0003013 .000251 .000211
     DTP    .01       0 .004988     .009 .009684    .0099 .009950  .00999
     DYL         .248935 .213412 .1433149 .113808 .1127445 .133503 .334223
     DYN .00014 .001731 .001765 .0020381 .002776 .0051096 .008503 .035653
     DYG    .02 .235233 .239670 .2751025 .368672 .6454241 1.00288 2.74862
     SUM         .511914 .603638 .8460230 1.10048 1.561730 2.05027 4.26494
     RMS         .342936 .348854 .5168865 .715963 1.023514 1.34964 2.99240

 NOMINAL
       L    NOM       1 1.99526      10 31.6228      100 199.526     1000
      TL    NOM       0 .997626     1.8 1.93675     1.98 1.98998    1.998
       Y    NOM 50.6667 25.8923 5.966667 2.57060 1.496667 1.24892 1.04967
     YDB    NOM 17.0472 14.1317 7.757318 4.10034 1.751251 .965357 .210514
      Te    NOM       4       4       4       4       4       4       4

ERROR CALC
 L+DL,DB            .01     3.1    10.31   15.46    20.61    23.7   30.91
    L+DL        1.00231 2.04174 10.73989 35.1560 115.0800 234.423 1233.10
      TL        .004600 1.02044 1.813778 1.94311 1.982621 1.99147 1.99838
      Te        3.98620 3.86342 3.586647 3.39699 3.213763 3.10683 2.86577
  T2+DT2  300.1
      Te        4.00201 4.00201 4.002013 4.00201 4.002013 4.00201 4.00201
  T1+DT1   2.01
      Te        3.98980 3.99479 3.998799 3.99948 3.999699 3.99975 3.99979
  TP+DTP   2.01
      TL              0 1.00261    1.809 1.94644   1.9899 1.99993 2.00799
      Te              4 3.99501    3.991 3.99032   3.9901 3.99005 3.99001
Y+DYL,DB        17.2277 14.2830 7.844891 4.15134 1.778763 .985010 .222619
   Y+DYL        52.8165 26.8103 6.088203 2.60096 1.506178 1.25459 1.05260
      Te        3.75106 3.78659 3.856685 3.88619 3.887256 3.86650 3.66578
   Y+DYN        50.6810 25.8996 5.968354 2.57132 1.497090 1.24928 1.04996
      Te        3.99827 3.99824 3.997962 3.99722 3.994890 3.99150 3.96435
   Y+DYG        52.6933 26.9280 6.205333 2.67342 1.556533 1.29888 1.09165
      Te        3.76477 3.76033 3.724898 3.63133 3.354576 2.99712 1.25138
DEFINITIONS
L=ATTEN LOSS                                    Tp=PHY TEMP OF L
DL=DELTA L                                      DTp=DELTA Tp
TL=TEMP CONTR OF L                              Te=LNA NOISE TEMP
Y=Y FACTOR                                      T1=COLD LOAD TEMP
DYL=DELTA Y FACTOR NON-LINEARITY                DT1=DELTA T1
DYN=DELTA Y, RADIOMETER NOISE (T,B)             T2=HOT LOAD TEMP
DYG=DELTA Y, RADIOMETER GAIN DELTA G            DT2=DELTA T2
```

Fig. 9. Supercalc 4 computer program NOISE1LD printout of the analysis of Fig. 2, showing the measurement error as a function of attenuator loss *L*, dB, and assumed input parameter errors; *T*1 = 2 K and *T*2 = 300 K.

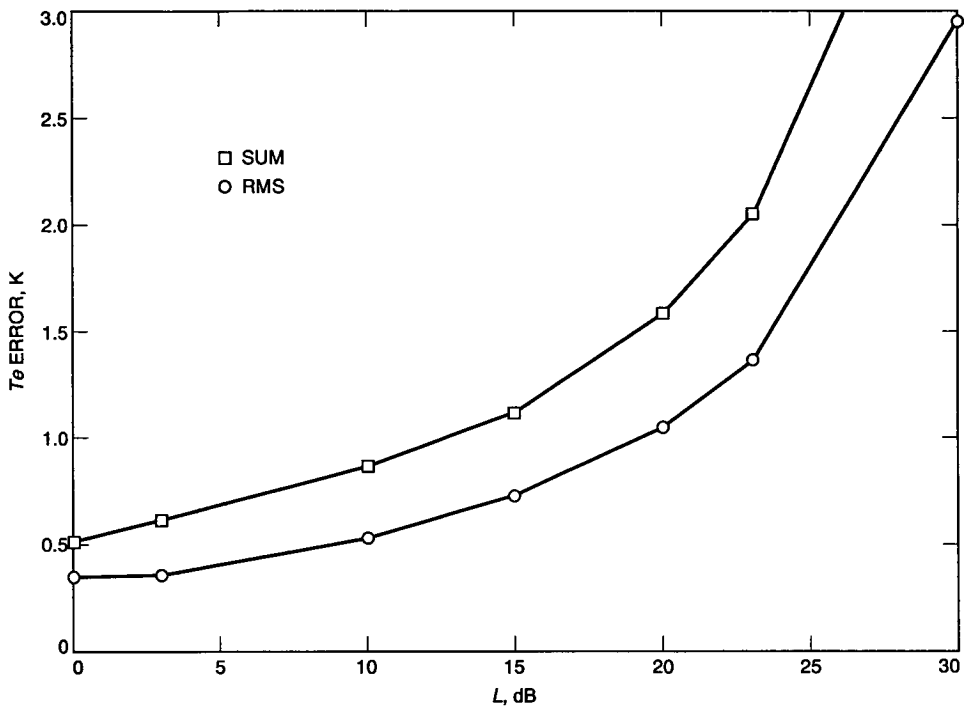
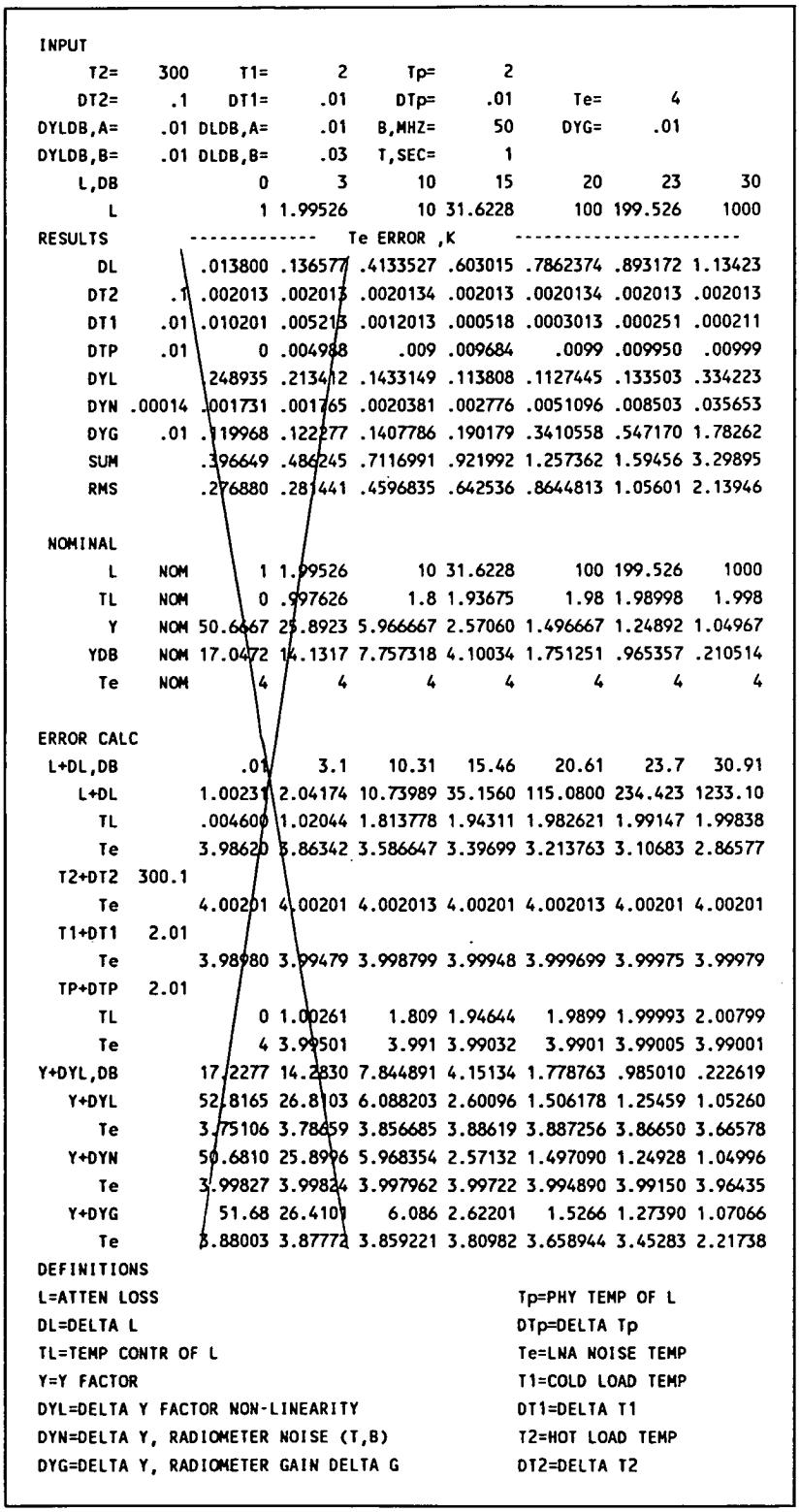


Fig. 10. Plot of the data in Fig. 9.



INPUT

T2=	300	T1=	2	Tp=	2			
DT2=	.1	DT1=	.01	DTp=	.01	Te=	4	
DYLDB,A=	.01	DLDB,A=	.01	B,MHZ=	50	DYG=	.01	
DYLDB,B=	.01	DLDB,B=	.03	T,SEC=	1			
L,DB		0	3	10	15	20	23	30
L		1	1.99526	10	31.6228	100	199.526	1000

RESULTS — Te ERROR ,K

DL		.013800	.136577	.4133527	.603015	.7862374	.893172	1.13423
DT2	.1	.002013	.002013	.0020134	.002013	.0020134	.002013	.002013
DT1	.01	.010201	.005213	.0012013	.000518	.0003013	.000251	.000211
DTP	.01	0	.004988	.009	.009684	.0099	.009950	.00999
DYL		248935	.213412	.1433149	.113808	.1127445	.133503	.334223
DYN	.00014	.001731	.001765	.0020381	.002776	.0051096	.008503	.035653
DYG	.01	.119968	.122277	.1407786	.190179	.3410558	.547170	1.78262
SUM		.396649	.486245	.7116991	.921992	1.257362	1.59456	3.29895
RMS		.276880	.281441	.4596835	.642536	.8644813	1.05601	2.13946

NOMINAL

L	NOM	1	1.99526	10	31.6228	100	199.526	1000
TL	NOM	0	.997626	1.8	1.93675	1.98	1.98998	1.998
Y	NOM	50.6667	25.8923	5.966667	2.57060	1.496667	1.24892	1.04967
YDB	NOM	17.0472	14.1317	7.757318	4.10034	1.751251	.965357	.210514
Te	NOM	4	4	4	4	4	4	4

ERROR CALC

L+DL,DB		.01	3.1	10.31	15.46	20.61	23.7	30.91
L+DL		1.00231	2.04174	10.73989	35.1560	115.0800	234.423	1233.10
TL		.004600	1.02044	1.813778	1.94311	1.982621	1.99147	1.99838
Te		3.98620	3.86342	3.586647	3.39699	3.213763	3.10683	2.86577
T2+DT2	300.1							
Te		4.00201	4.00201	4.002013	4.00201	4.002013	4.00201	4.00201
T1+DT1	2.01							
Te		3.98980	3.99479	3.998799	3.99948	3.999699	3.99975	3.99979
TP+DTP	2.01							
TL		0	1.00261	1.809	1.94644	1.9899	1.99993	2.00799
Te		4	3.99501	3.991	3.99032	3.9901	3.99005	3.99001
Y+DYL,DB		17.2277	14.2830	7.844891	4.15134	1.778763	.985010	.222619
Y+DYL		52.8165	26.8103	6.088203	2.60096	1.506178	1.25459	1.05260
Te		3.75106	3.78659	3.856685	3.88619	3.887256	3.86650	3.66578
Y+DYN		50.6810	25.8996	5.968354	2.57132	1.497090	1.24928	1.04996
Te		3.99827	3.99824	3.997962	3.99722	3.994890	3.99150	3.96435
Y+DYG		51.68	26.4101	6.086	2.62201	1.5266	1.27390	1.07066
Te		3.88003	3.87772	3.859221	3.80982	3.658944	3.45283	2.21738

DEFINITIONS

L=ATTEN LOSS	Tp=PHY TEMP OF L
DL=DELTA L	DTp=DELTA Tp
TL=TEMP CONTR OF L	Te=LNA NOISE TEMP
Y=Y FACTOR	T1=COLD LOAD TEMP
DYL=DELTA Y FACTOR NON-LINEARITY	DT1=DELTA T1
DYN=DELTA Y, RADIOMETER NOISE (T,B)	T2=HOT LOAD TEMP
DYG=DELTA Y, RADIOMETER GAIN DELTA G	DT2=DELTA T2

Fig. 11. Supercalc 4 computer program NOISE1LD printout of the analysis of Fig. 2, showing the measurement error as a function of attenuator loss *L*, dB, and assumed input parameter errors; *T*1 = 2 K and *T*2 = 300 K. The lower gain change (*DYG*) than that in Fig. 9 is appropriate for case 5, *L* = 10 dB.

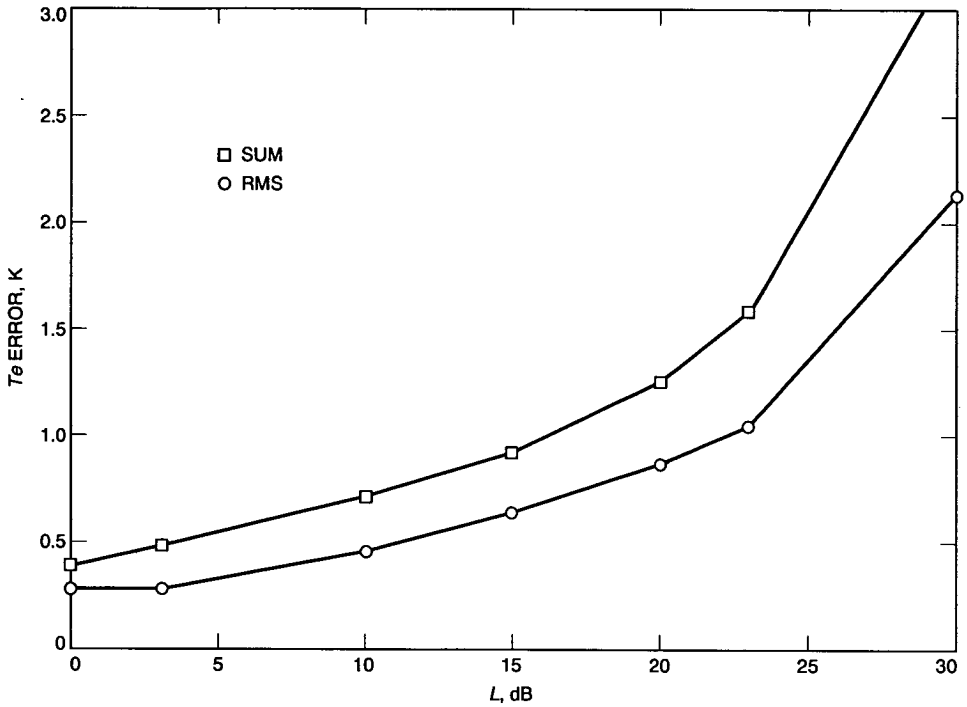


Fig. 12. Plot of the data in Fig. 11.

Correction of High-Frequency Noise-Temperature Inaccuracies

C. T. Stelzried
TDA Technology Development Section

Deep-space mission data rates to Earth are limited by the system operating noise-temperature (T_{op}) performance of the DSN. This article addresses some of the techniques and definitions used for measuring and reporting the effective noise temperature of receivers (Te) *and* T_{op} *of the DSN's ground receiving systems. Calibration loads are used to measure* T_{op} *of the DSN antennas. At 32 GHz, a calibration load cooled to 2-K physical temperature requires a correction of 0.67 K to determine the noise temperature. Using corrected noise temperature for the calibration loads results in the correct values for* T_{op} *such that the total system noise power can be defined by* $Pn = kT_{op}B$, *as required for DSN telecommunications design control tables.* T_{op} *and* Te *should not be converted to equivalent physical temperatures.*

I. Introduction

System operating noise temperature (T_{op}) is very important to the DSN; deep-space mission data rates to Earth are limited by the DSN's T_{op} performance. The DSN uses design control tables to document parameters of the spacecraft-to-ground end-to-end telecommunications system. A key parameter affecting the data quality is the signal-to-noise ratio (SNR) of the signal received by the DSN. The received SNR is proportional to DSN antenna gain divided by the system operating noise temperature (G/T_{op}).

This article addresses some of the techniques and definitions used for measuring and reporting the effective noise temperature of receivers (Te) and T_{op} of the DSN's ground receiving systems. Proper evaluation of the noise-temperature performance of high-frequency, low-noise amplifiers (LNA's), such as Ka-band (32-GHz) masers currently under development in the DSN Advanced Systems Program, requires the use of frequency-dependent corrections for the noise power available from the calibration loads. At 32 GHz, a calibration load cooled to 2 K has an available noise power equivalent to 1.33 K; a correction of 0.67 K is needed.

An analysis for optimizing the testing configuration for LNA noise temperature is provided in [1]; frequency-dependent corrections are not used. Frequency-dependent

corrections for the calibration loads are required for accurate and correct evaluation of the measured noise temperatures Te and T_{op}. Using calibration loads with frequency-dependent corrections results in the proper measured and reportable values for Te and T_{op}. No further correction is needed. These corrected values for Te and T_{op} are the values needed for the DSN telecommunications-link design control tables, as tabulated from data in the *DSN/Flight Project Interface Design Handbook*.[1]

The values of Te and T_{op} should not be converted to equivalent physical temperatures for DSN applications.

II. Analysis

The purpose of the following is to apply frequency-dependent corrections to calibration loads used for LNA and system noise-temperature measurements and to clarify the results for application in the DSN. The available noise power, Pn, from a load [2] is given by

$$Pn = \frac{hfB}{e^{hf/kT}-1} \tag{1}$$

where

h = Planck's constant = 6.6262×10^{-34} J

k = Boltzmann's constant = 1.3806×10^{-23} J/K

f = operating frequency, Hz

T = physical temperature, K

B = bandwidth, Hz

hf/k = $0.0480f$, GHz

For $(hf/kT) << 1$

$$Pn = kTB \tag{2}$$

Equation (2) can be used to accurately determine the noise power available from calibration loads at any frequency by substituting an equivalent noise temperature, Tn, for the physical temperature, T. With this substitution and Eq. (1)

$$Tn = \frac{hf/k}{e^{hf/kT}-1} \tag{3}$$

The value of Tn approaches T as f/T approaches 0, and $Tn = T$ in the limit, when $f/T = 0$. For calibration loads at physical temperatures above 1 K and frequencies below 1 MHz, letting $Tn = T$ results in an error of less than 0.000024 K. At 32 GHz, a calibration load cooled to 2 K has an available noise power equivalent to 1.33 K, i.e., a correction (Tc) of 0.67 K is needed.

Rewriting Eq. (3) in terms of the physical temperature, T, and the correction, Tc ,

$$Tn = T - Tc \tag{3a}$$

where

$$T_c = T - \frac{hf/k}{e^{hf/kT}-1} \tag{3b}$$

where T_c = a frequency-dependent correction term to physical temperature to obtain the equivalent noise temperature of a calibration load, K.

The frequency-dependent correction term, Tc, is small below 10 GHz and is significant at 32 GHz and above, as shown below. Without this correction, calibration loads would generate infinite power over the entire frequency spectrum, according to Eq. (2).

Either Eq. (3) or (3a) can be used to obtain the corrected equivalent noise temperature of calibration loads. As shown in Table 1, for a fixed microwave frequency, Tc can be treated as a constant over a wide range of temperatures above 2 K with small error. By measuring T, the physical temperature, and subtracting Tc, treated as a constant, Eq. (3a) is useful for obtaining the equivalent noise temperature, Tn. This provides better than 0.0001-K accuracy for ambient calibration loads, such as used in the DSN at the present S- (2.295-GHz), X- (8.42-GHz), and future Ka-band microwave frequencies over the range of expected ambient temperatures. Note that Tc cannot be treated as a constant with operational frequency changes.

Other important corrections for calibration loads not discussed here, such as mismatch, must be accounted for. Minimizing mismatch is important for ambient-temperature calibration loads [3]. Receiver nonlinearities

[1] *DSN/Flight Project Interface Design Handbook*, Vol. I, 810-5, Rev. D (internal document), Jet Propulsion Laboratory, Pasadena, California), August 1, 1992.

[4] can give large errors. Using Eq. (2) with the equivalent corrected noise temperature results in

$$Pn = kTnB \tag{4}$$

Use of the corrected equivalent noise temperature is appropriate for telecommunications design control tables, such as used in the DSN [5].

Calibration of the corrected equivalent noise temperature of an LNA using a load with physical T and a noise source (usually a noise diode connected to the LNA through a directional coupler between the LNA and the load) with temperature TND, all referred to the amplifier input, requires solution of

$$Te = \frac{TND}{Y-1} - Tn \tag{5}$$

where

Y = power ratio at the output of follow-up amplifiers with the noise source turned on and off

TND = noise source excess noise at amplifier input, K

Tn = equivalent noise temperature of source at physical temperature T, K

Measurement configurations using cooled attenuators located between the load and the LNA are evaluated using Eq. (5) by analyzing an equivalent TND and Tn defined at the LNA input. Similarly, using two calibration loads and a cooled attenuator requires the evaluation of corrected equivalent noise temperatures $T1$ and $T2$ for the loads, defined at the LNA input, and the solution of

$$Te = \frac{T2 - T1Y}{Y-1} \tag{6}$$

where Y = power ratio obtained at output of follow-up amplifiers switching between $T1$ and $T2$.

The value Te, as measured with Eqs. (5) and (6), contains the follow-up amplifier noise temperature Tf. The LNA noise temperature, $TLNA$, requires the correction

$$TLNA = Te - Tf \tag{7}$$

For most system applications, especially when the first amplifier has more than a 30-dB gain, Tf is small compared with Te. For system applications, Te is the significant parameter.

III. Results

Equations (5) and (6) have been programmed in Supercalc 4. Assuming the load and noise source are separated from the amplifier by an attenuator with loss L at physical temperature Tp, Figs. 1 and 2 show the solutions using Eq. (5). Using Eq. (6), Figs. 3 and 4 assume the loads are separated from the amplifier by the attenuator. The values TND, L, Tp, $T1$, $T2$, and Y are considered known. Equation (3) is used to correct $T1$, $T2$, and TL for frequency. The equation $TL = Tp\ (1 - 1/L)$ represents the attenuator noise-temperature contribution. The Eq. (3) correction is applied to Tp, not TL.

For the purposes of this article, $f = 0.001$ GHz is used as the dc ($f = 0$) case. The results shown in Fig. 1, at dc, assume $Y = 2.1136$, appropriate for $Te = 4.0$ K and the other input parameters assumed and used in [2]. Figure 2 shows the result of operating at $f = 32$ GHz with all other inputs unchanged. The errors in Te due to various parameter changes are virtually unchanged with frequency and also agree with [2] (Fig. 3 for $L = 20$ dB). However, Te increases from 4.00 to 4.67 K at $f = 32$ GHz relative to dc. Figures 3 and 4 have similar results, with Te increasing from 4.00 K at dc to 4.68 K at 32 GHz.

IV. Conclusion

Using loads with corrected equivalent noise temperatures results in the proper value for the amplifier noise temperature, Te. The value Te in this case is the equivalent noise temperature, not the physical temperature. From Eq. (3), the physical temperature is given by

$$T = \frac{hf/k}{ln((hf/kTn) + 1)} \tag{8}$$

Equation (8) is useful for converting a measured noise temperature, Tn, to a physical temperature, T. An example of this is using T, the physical or thermodynamic temperature, for reporting the cosmic background radiation temperature. The cosmic physical or thermodynamic temperature obtained with Eq. (8) after measuring the noise temperature is independent of frequency.

The physical or thermodynamic temperature is not appropriate for reporting measurements of amplifier noise

temperature, Te, for such purposes as the DSN telecommunications-link design tables. In addition, quantum noise is inherent to low-noise amplifiers and is included in the measured value.

System operating noise temperature, T_{op}, in the DSN is typically evaluated using an ambient load [4] and a previously measured Te (Te is equivalent noise temperature, not the physical temperature). Application of Eq. (3) to the ambient load physical temperature provides the needed frequency-dependent correction. This correction is small at S- and X-bands presently used in the DSN (T reduced by 0.2 K at 8.4 GHz for a 300-K load) but is important for future Ka-band operation (T reduced by 0.77 K at 32 GHz for a 300-K load).

This article provides an overall consistent approach using corrected equivalent noise temperatures for both T_e for the LNA and T_{op} for the system, as obtained using calibration loads. The DSN telecommunications-link design control tables should use the simplified Eq. (4) for noise power.

Acknowledgments

Extensive discussions with R. Clauss are appreciated for this article. M. Sue verified that the DSN telecommunications-link design tables compute noise power using Eq. (4) with an equivalent noise temperature Tn, consistent with this article.

References

[1] C. Stelzried, "Minimizing Noise-Temperature Measurement Errors," *TDA Progress Report 42-111*, vol. July–September 1992, Jet Propulsion Laboratory, Pasadena, California, pp. 254–268, November 15, 1992.

[2] C. Stelzried, *The Deep Space Network—Noise Temperature Concepts, Measurements, and Performance*, JPL Publication 82-33, Jet Propulsion Laboratory, Pasadena, California, September 15, 1982.

[3] C. Stelzried, "Operating Noise-Temperature Calibrations of Low-Noise Receiving Systems," *Microwave Journal*, vol. 14, no. 6, pp. 41–47, June 1971.

[4] C. Stelzried, "Non-Linearity in Measurement Systems: Evaluation Method and Application to Microwave Radiometers," *TDA Progress Report 42-91*, vol. July–September 1987, Jet Propulsion Laboratory, Pasadena, California, p. 57, November 15, 1987.

[5] J. Yuen, *Deep Space Telecommunications Systems Engineering*, New York: Plenum Press, p. 12, 1983.

Table 1. Examples of errors in treating *Tc* as a constant with changes of frequency, *f*, and the calibration load physical temperature, *T*.

f, GHz	T, K	Tc, K	Error in Tc, K due to 10-percent change in T
8	2	0.19	0.0006
8	80	0.19	<0.0001
8	300	0.19	<0.0001
32	2	0.67	0.0087
32	80	0.77	0.0002
32	300	0.77	<0.0001

```
INPUT ------------------------------------------------
      T=    300     TND=   1000     Tp=      2  f,GHz=    .001
     DT=     .1    DTND=     50   DTp=     .1      Y=  2.1136
DYLDB,A=    .01  DLDB,A=    .01 B,MHZ=     50    DYG=     .01
DYLDB,B=    .01  DLDB,B=    .03 T,SEC=      1      L=     100

RESULTS --------------"Te error (DTe)-----------------------
      DL .78622             DTp .09900            DYG .328408
      DT .00100             DYL .16455            SUM 1.83299
    DTND .44899             DYN .00482            RMS .982087

CALCULATIONS ----------------------------------------------
   L,DB=     20    Y,DB= 3.2502   DYN= .00014   hf/k= .000048
        NOMINAL         DELTA CALCULATIONS ---------------
                        L+DL,DB=      T+DT=          TN+DTN
                          20.61        300.1            1050
                         L+DL=
                        115.08
     TL=   1.98          1.9826         1.98            1.98
    TLn= 1.9800          1.9826       1.9800          1.97998
    TnR= 3.0000          2.6069       3.0010          3.00000
      T= 4.9800          4.5895       4.9810          4.97998
    TND=     10          8.6896           10            10.5
     Te= 3.9999          3.2137       3.9989          4.44890

DELTA CALCULATIONS,CONT  ----------------------------------
        Tp+DTp=         Y+DYL,DB=    Y(1+2*DYN)=    Y(1+2*DYG)=
            2.1          3.2927         2.1142         2.15587
                         Y+DY=
                        2.1344
     TL=  2.079            1.98           1.98            1.98
    TLn= 2.0790          1.9800         1.9800          1.97998
    TnR= 3.0000          3.0000         3.0000          3.00000
      T= 5.0790          4.9800         4.9800          4.97998
    TND=     10              10             10              10
     Te= 3.9009          3.8354         3.9951          3.67150

DEFINITIONS
Te=LNA NOISE TEMP                    Tp=PHY TEMP OF L
L=ATTEN LOSS                         DTp=DELTA Tp
DL=DELTA L                           TND=NOISE DIODE
TL=TEMP CONTR OF L                   DTND=DELTA NOISE DIODE
Y=Y FACTOR                           T=LOAD TEMP
DYL=DELTA Y FACTOR,LINEARITY         DT=DELTA T
DYN=DELTA Y, RADIOMETER NOISE (T,B)  Tn=T CORRECTED FOR FREQ
DYG=DELTA Y, RADIOMETER GAIN DELTA G  TnR=Tn AT REF
```

Fig. 1. Supercalc 4 computer program NOISE2ND printout of the measured noise temperature and errors of a low-noise amplifier using a load, noise diode, and cooled attenuator at 0.001 GHz (dc).

```
INPUT  ------------------------------------------------
      T=    300     TND=   1000    Tp=      2  f,GHz=      32
     DT=     .1    DTND=     50   DTp=     .1      Y=  2.1136
DYLDB,A=    .01  DLDB,A=    .01 B,MHZ=     50    DYG=     .01
DYLDB,B=    .01  DLDB,B=    .03 T,SEC=      1      L=     100

RESULTS  --------"Te error (DTe) ----------------------------
      DL .78635             DTp .10794            DYG .328408
      DT .00100             DYL .16455            SUM 1.84206
    DTND .44899             DYN .00482            RMS .983130

CALCULATIONS  ---------------------------------------------
   L,DB=     20    Y,DB= 3.2502   DYN= .00014   hf/k=   1.536
        NOMINAL         DELTA CALCULATIONS ---------------
                       L+DL,DB=       T+DT=          TN+DTN
                          20.61        300.1            1050
                         L+DL=
                        115.08
    Tpn= 1.3294         1.3294        1.3294         1.32935
    TLn= 1.3161         1.3178        1.3161         1.31606
    TnR= 2.9923         2.6002        2.9933         2.99233
      T= 4.3084         3.9180        4.3094         4.30838
    TND=     10         8.6896           10            10.5
     Te= 4.6715         3.8852        4.6705         5.12050

DELTA CALCULATIONS,CONT  ----------------------------------
        Tp+DTp=         Y+DYL,DB=     Y(1+2*DYN)=    Y(1+2*DYG)=
            2.1           3.2927          2.1142         2.15587
                         Y+DY=
                        2.1344
    Tpn= 1.4384         1.3294        1.3294         1.32935
    TLn= 1.4240         1.3161        1.3161         1.31606
    TnR= 2.9923         2.9923        2.9923         2.99233
      T= 4.4163         4.3084        4.3084         4.30838
    TND=     10            10           10              10
     Te= 4.5636         4.5070        4.6667         4.34309

DEFINITIONS                          Tpn=Tp CORRECTED FOR FREQ
Te=LNA NOISE TEMP                    Tp=PHY TEMP OF L
L=ATTEN LOSS                         DTp=DELTA Tp
DL=DELTA L                           TND=NOISE DIODE
TL=TEMP CONTR OF L                   DTND=DELTA NOISE DIODE
Y=Y FACTOR                           T=LOAD TEMP
DYL=DELTA Y FACTOR,LINEARITY         DT=DELTA T
DYN=DELTA Y, RADIOMETER NOISE (T,B)  Tn=T CORRECTED FOR FREQ
DYG=DELTA Y, RADIOMETER GAIN DELTA G  TnR=Tn AT REF
```

Fig. 2. Supercalc 4 computer program NOISE2ND printout of the measured noise temperature and errors of a low-noise amplifier using a load, noise diode, and cooled attenuator at 32 GHz.

```
INPUT  ------------------------------------------------
     T2=    300      T1=     80    Tp=      2  f,Ghz=    .001
    DT2=     .1     DT1=      1   DTp=    .01      Y=  2.5942
DYLDB,A=    .01  DLDB,A=    .01 B,MHZ=     50    DYG=     .01
DYLDB,B=    .01  DLDB,B=    .03 T,SEC=      1      L=      10

RESULTS ---------------- TE error (DTE) ---------------------
      DL .41335             DTP .00900            DYG .434972
     DT2 .00627             DYL .26228            SUM 1.29495
     DT1 .16273             DYN .00635            RMS .674902

CALCULATIONS ----------------------------------------------
   L,DB=     10    Y,DB= 4.1400   DYN= .00014   hf/k= .000048
        NOMINAL         DELTA CALCULATIONS ---------------
                       L+DL,DB=      T2+DT2=        T1+DT1=
                          10.31        300.1            81
                          L+DL=
                         10.740
     TL=    1.8          1.8138          1.8           1.8
    TLn= 1.8000          1.8138       1.8000       1.79998
   T2nR= 30.000          27.933       30.010       30.0000
     T2 31.800          29.747       31.810       31.8000
   T1nR= 8.0000          7.4489       8.0000       8.10000
     T1 9.8000          9.2626       9.8000       9.89997
     Te= 4.0001          3.5867       4.0063       3.83732

DELTA CALCULATIONS,CONT  ----------------------------------
       Tp+DTp=         Y+DYL,DB=     Y(1+2*DYN)=    Y(1+2*DYG)=
          2.01          4.1914          2.5949         2.64608
                         Y+DY=
                        2.6251
     TL=  1.809            1.8             1.8            1.8
    TLn= 1.8090          1.8000          1.8000       1.79998
   T2nR= 30.000          30.000          30.000       30.0000
     T2 31.809          31.800          31.800       31.8000
   T1nR= 8.0000          8.0000          8.0000       8.00000
     T1 9.8090          9.8000          9.8000       9.79997
     Te= 3.9911          3.7378          3.9937       3.56508

DEFINITIONS
Te=LNA NOISE TEMP                    Tp=PHY TEMP OF L
L=ATTEN LOSS                         DTp=DELTA Tp
DL=DELTA L                           T1=COLD LOAD TEMP
TL=TEMP CONTR OF L                   DT1=DELTA T1
Y=Y FACTOR                           T2=HOT LOAD TEMP
DYL=DELTA Y FACTOR,LINEARITY         DT2=DELTA T2
DYN=DELTA Y, RADIOMETER NOISE (T,B)  Tn=T CORRECTED FOR FREQ
DYG=DELTA Y, RADIOMETER GAIN DELTA G TnR=Tn AT REF
```

Fig. 3. Supercalc 4 computer program NOISE2LD printout of the measured noise temperature and errors of a low-noise amplifier using two loads and a cooled attenuator at 0.001 GHz (dc).

```
INPUT ------------------------------------------------
     T2=    300      T1=     80    Tp=      2  f,Ghz=      32
    DT2=     .1     DT1=      1   DTp=    .01      Y=  2.5942
DYLDB,A=    .01  DLDB,A=    .01 B,MHZ=     50    DYG=     .01
DYLDB,B=    .01  DLDB,B=    .03 T,SEC=      1     L=      10

RESULTS ------------------ TE error (DTE) ------------------
      DL .41400             DTP .00857            DYG .434968
     DT2 .00627             DYL .26228            SUM 1.29516
     DT1 .16272             DYN .00635            RMS .675288

CALCULATIONS ---------------------------------------------
   L,DB=     10    Y,DB= 4.1400   DYN= .00014   hf/k=   1.536
        NOMINAL         DELTA CALCULATIONS ---------------
                        L+DL,DB=      T2+DT2=        T1+DT1=
                          10.31        300.1             81
                          L+DL=
                         10.740
    Tpn= 1.3294          1.3294        1.3294        1.32935
    TLn= 1.1964          1.2056        1.1964        1.19642
   T2nR= 29.923          27.862        29.933        29.9233
     T2= 31.120          29.067        31.130        31.1197
   T1nR= 7.9234          7.3776        7.9234        8.02344
     T1= 9.1199          8.5832        9.1199        9.21986
     Te= 4.6801          4.2660        4.6863        4.51733

DELTA CALCULATIONS,CONT  ---------------------------------
        Tp+DTp=         Y+DYL,DB=     Y(1+2*DYN)=    Y(1+2*DYG)=
           2.01          4.1914         2.5949         2.64608
                          Y+DY=
                         2.6251
    Tpn= 1.3389          1.3294        1.3294        1.32935
    TLn= 1.2050          1.1964        1.1964        1.19642
   T2nR= 29.923          29.923        29.923        29.9233
     T2= 31.128          31.120        31.120        31.1197
   T1nR= 7.9234          7.9234        7.9234        7.92345
     T1= 9.1284          9.1199        9.1199        9.11986
     Te= 4.6715          4.4178        4.6737        4.24508

DEFINITIONS                          Tpn=Tp CORRECTED FOR FREQ
Te=LNA NOISE TEMP                    Tp=PHY TEMP OF L
L=ATTEN LOSS                         DTp=DELTA Tp
DL=DELTA L                           T1=COLD LOAD TEMP
TL=TEMP CONTR OF L                   DT1=DELTA T1
Y=Y FACTOR                           T2=HOT LOAD TEMP
DYL=DELTA Y FACTOR,LINEARITY         DT2=DELTA T2
DYN=DELTA Y, RADIOMETER NOISE (T,B)  Tn=T CORRECTED FOR FREQ
DYG=DELTA Y, RADIOMETER GAIN DELTA G  TnR=Tn AT REF
```

Fig. 4. Supercalc 4 computer program NOISE2LD printout of the measured noise temperature and errors of a low-noise amplifier using two loads and a cooled attenuator at 32 GHz.

Fault Identification Using Multidisciplinary Techniques at the Mars/Uranus Station Antenna Sites

D. S. Santo[1]

M. B. Schluter[2]

R. J. Shlemon[3]

A fault investigation was performed at the Mars and Uranus antenna sites at the Goldstone Deep Space Communications Complex in the Mojave desert. The Mars/Uranus Station consists of two large-diameter reflector antennas used for communication and control of deep-space probes and other missions. The investigation included interpretation of Landsat thematic mapper scenes, side-looking airborne radar transparencies, and both color-infrared and black-and-white aerial photography. Four photolineaments suggestive of previously undocumented faults were identified. Three generally discrete morphostratigraphic alluvial-fan deposits were also recognized and dated using geomorphic and soil stratigraphic techniques. Fourteen trenches were excavated across the four lineaments; the trenches show that three of the photolineaments coincide with faults. The last displacement of two of the faults occurred between about 12,000 and about 35,000 years ago. The third fault was judged to be older than 12,000 years before present (ybp), although uncertainty remains. None of the surface traces of the three faults crosses under existing antennas or structures; however, their potential activity necessitates appropriate seismic retrofit designs and loss-prevention measures to mitigate potential earthquake damage to facilities and structures.

I. Introduction

The Mars/Uranus Station consists of two antennas and surrounding support buildings (Fig. 1). The DSN 70-m antennas are the largest fully steerable communications antennas in the world. The Mars antenna at DSS 14 was constructed as a 64-m antenna in 1966 and later enlarged to 70 m in 1988. Standing more than 71 m high, this antenna is one of the more striking features of the Goldstone complex.

The Uranus antenna (DSS 15) is a 34-m, high-efficiency, precision-shaped antenna located approximately 487 m southeast of the Mars antenna (Fig. 1). Built in 1984, this

[1,2] Consultant to the Ground Antennas and Facilities Engineering Section from Converse Consultants West.

[3] Independent consultant to the Ground Antennas and Facilities Engineering Section.

latest antenna addition was first used in January 1986 to support the Voyager 2 spacecraft encounter with the planet Uranus, which is more than 3 billion km from Earth.

In the early 1980s, plans were developed to construct additional antennas and support buildings at the facility. Geotechnical investigations performed for the proposed expansion revealed several photolineaments within the Mars/Uranus Station valley [1]. The lineaments were thought to be associated with faulting. Refraction seismic lines run through the center of the valley were inconclusive regarding the origin of the photolineaments, and doubt remained about the possible impact of these features and the potential hazard they posed to the facilities. The late 1980s brought a renewed interest in expansion of the facility and a heightened concern among facility designers about the lineaments and the possible presence of faults within the Mars/Uranus Station valley. A multidisciplinary fault investigation was undertaken to determine the origin of the lineaments and to assess their potential impact on existing and proposed structures.

The purpose of this investigation was twofold: (1) to evaluate the Mars/Uranus Station valley for the presence of faults, including the previously identified lineaments, and (2) to determine the relative age and activity of any identified faults.

II. Location and Regional Geologic Setting

The Mars/Uranus Station is located in the northwest corner of the Goldstone Deep Space Communications Complex (DSCC) approximately 90 km north of Barstow, in San Bernardino County, California (Fig. 2). This is a remote area, about a 2- to 3-hr drive from major population centers of Southern California. Public access is restricted because the DSCC is located within the U.S. Army's Fort Irwin National Training Center and east of the China Lake Naval Weapons Center, Mojave Range B.

The Goldstone DSCC lies in the north-central Mojave Block geomorphic province. The Mojave Block is bounded on the southwest by the San Andreas Fault and Transverse Ranges; on the north and northeast by the Garlock Fault, Tehachapi Mountains, and the Basin and Ranges; and on the east by the arbitrary boundary of the Nevada state line and the Colorado River. The central portion of the block is characterized by potentially active Quaternary and locally active Holocene faults [2]. The faults trend northwest and southeast and are predominently strike-slip faults (Fig. 3, after [3]). The faults cut Tertiary extensional basins separated by low hills of Tertiary volcanic and sedimentary rock or pre-Tertiary intrusive rock [4]. Many basins contain Pleistocene pluvial dry lakes.

III. Technical Approach

Geologic information within the Goldstone DSCC is limited and original work was required during the course of the investigation. To cope with the remote location and lack of information in a timely and economical way, analyses were made of remote-sensing imagery and aerial photography. Results of the imagery analyses were then checked in the field with ground-truth verification, reconnaissance geologic mapping, geomorphic observations, fault trenching, and soil-stratigraphic analyses. A summary of the technical approach is presented in Fig. 4.

IV. Imagery Analysis

Remote-sensing imagery and aerial photography that encompassed 15,540 square km were reviewed for evidence of possible Quaternary faulting. The imagery included a Landsat thematic mapper scene, side-looking airborne radar (SLAR) transparencies, and both black-and-white and color-infrared aerial photography (Table 1).

The imagery was selected in accordance with the multi-approach outlined in [5]. This approach optimizes the results of an imagery analysis by reviewing multiscale, multistation, multiband, multidate, and multi-enhanced imagery. The imagery is then reviewed and interpreted by multiple observers.

Geomorphic and photographic features recognized on the imagery and considered to be evidence of faulting included:

(1) topographic and vegetation alignments

(2) linear breaks in slope

(3) rectilinear closed depressions

(4) linear or uphill facing scarps on alluvial fans

(5) linear change in drainage or erosional texture

(6) offset stream channels

(7) linear change in photographic tone or color

(8) linear spectral contrasts on Landsat imagery

A. Field Reconnaissance

After the imagery analysis, photolineaments were checked in the field. The purpose of the field reconnaissance was to differentiate fault-related geomorphic features

from those formed by man-made, natural erosional, or depositional processes. In addition, geologic mapping and geomorphic observations were made. Geomorphic observations included relative dissection of fan surfaces, relative development of desert pavement and patina (desert varnish), and fan superposition relationships. Early morning, low-angle-sun-illumination, 35mm slides were taken of photolineaments.

B. Results of Imagery Analysis

Based on the photographic evidence and observations made during the field reconnaissance, four principal lineaments were identified and suspected to be associated with Quaternary faulting (Fig. 5). Figure 6 is a composite aerial photograph showing these lineaments. The characteristics of each lineament are compiled in Table 2.

V. Exploration

Field exploration included reconnaissance geologic mapping, siting exploratory trenches, and detailed logging of the trenches, including soil-stratigraphic analyses. The trenches served two purposes: (1) to determine if the photolineaments were fault related, and (2) to expose sediments amenable to relative dating by soil stratigraphic techniques.

Fourteen trenches were excavated with a backhoe equipped with a 61-cm shovel; Fig. 7 shows the trench locations.

VI. Site Geology

The Mars/Uranus Station is in a linear interior-draining valley bounded on the south by steep volcanic ridges and on the north by low-lying alluvial fan–capped hills. Southwest of the antennas is a small dry lake. As shown in Fig. 7, the major geologic units are three different-age alluvial-fan deposits, playa deposits, and extrusive volcanic rocks.

A. Undifferentiated Volcanic Rocks

Andesite forms the steep hills south of the antennas and crops out in two isolated locations in the hills to the north. The rock is light gray to black, massive, and slightly vesicular. It is resistant to weathering and forms steep slopes.

Latitic to dacitic tuffs underlie the low hills north of the main valley where they are capped by Unit 3 alluvium. The rock is white to pink, massive to thickly bedded, lithic tuff, and tuff breccia. It is easily eroded and exposures are limited to small stream canyons.

The volcanic rocks are probably equivalent to the Pliocene, Lane Mountain Andesite of Dibblee [6] or the lower Miocene, Lane Mountain Quartz Latite of Burke [7].

B. Alluvial-Fan Deposits

The alluvial-fan deposits were subdivided into three mapping units based mainly on geomorphic and soil-stratigraphic evidence. Diagnostic criteria included elevation above base level, superposition relationships, fan surface morphology (relative dissection, etc.), development of desert pavement and patina, and soil-profile development. These criteria and the application of soil-profile development for fault-activity assessments are summarized in [8], [9], [10], and [11]. Relative ages for site-specific surfaces and alluvial fan sediments are given in [12].

The nested geomorphic expression of the three alluvial-fan deposits and their diagnostic surface and soil-stratigraphic characteristics allow them to be mapped as generally discrete morphostratigraphic units, which were designated as Units 1 (youngest), 2, and 3, respectively (Fig. 7).

1. Unit 3 Alluvial Fans. Unit 3 fan deposits cap the low hills north of the antennas and form the highest level geomorphic surfaces flanking the Mars/Uranus Station valley. Remnant fan surfaces are heavily dissected, but locally well-developed desert pavements with dark patinas are still visible on drainage divides. The deposits are coarse gravel/cobble fanglomerates interbedded with finer grained fluvial lenses. On the undissected surfaces, these deposits support strongly developed, relict paleosols characterized by truncated argillic horizons and multiple, superimposed, stage III to IV calcic horizons (Btk).[4] Their relative elevation, surface characteristics, and pedologic development suggest that these fans are well in excess of 100,000 years old [12].

2. Unit 2 Alluvial Fans. Unit 2 fan deposits form mid-level geomorphic surfaces surrounding the antennas on the west and east. Fan surfaces are moderately dissected with poorly to moderately developed desert pavements. Pavements display a relatively light-colored patina.

[4] The term horizon refers to an interface indicative of a position in a stratigraphic sequence; the A horizon is the uppermost soil horizon and the B horizon is below. Bt horizon refers to an accumulation of silicate clay that has either formed in situ or is illuvial; hence, it has more clay than the assumed parent material and/or the overlying horizon. Btk indicates an accumulation of alkaline earth carbonates within the Bt horizon. See Appendix for additional definitions.

The material is finer grained than Unit 3 but still contains coarse gravel/cobble layers. The soils capping these surfaces are of moderately to strongly developed relict paleosols characterized by argillic horizons and multiple stage II to III calcic horizons. Remnants of moderately developed, near-surface buried soils are also present. Geomorphic and soil-stratigraphic evidence suggest that the Unit 2 fan complexes are at least 35,000 to 40,000 years old, but may be as old as 100,000 years before present (ybp).

3. Unit 1 Alluvial Fans. Unit 1 fans form the lowest level geomorphic surfaces in the Mars/Uranus Station valley. Fan surfaces are smooth, undissected, and represent areas of active deposition. Where desert pavement is present, it is thin and patchy. The material consists of relatively fine-grained sands and gravels. Pedogenic profiles are generally absent to weakly developed. Where present, soils are characterized by cumulic profiles with incipient argillic horizons (Btjk)[5] containing disseminated carbonate. Geomorphic and soil-stratigraphic evidence suggests that the older portions of these fans are approximately 7,000 to 12,000 years old, but local portions may be as old as 35,000 ybp.

Recent stream channel and valley-fill sediments were also mapped as Unit 1. These deposits are inherently young, and at the scale of the mapping are difficult to separate from the older portions of the fans. These deposits are estimated to be less than about 5,000 to 7,000 years old.

C. Other Surficial Deposits

1. Playa Deposits. Playa deposits consist of unconsolidated, well-sorted clay, silt, and fine sand. This material underlies the surface of the dry lake located west of the Uranus antenna. The morphostratigraphic relationship with Unit 1 fans is not entirely clear, but the two are probably intercalated at depth.

2. Talus. Talus blankets the base of the volcanic ridges south of the main valley. It consists of loose gravel to cobble-sized clasts of volcanic rock.

3. Landslides. One small landslide was mapped on the ridge south of the Uranus antenna. The slide was mapped based solely on geomorphic evidence; hence, its age and thickness are as yet unknown.

[5] Btjk: See preceding footnote; the "j" is used in combination with other horizon designations to denote incipient development of that particular feature or property.

VII. Fault Activity Analysis

The origin of Photolineaments 1, 2, and 3 were assessed by trenching. Faults were encountered in several trenches placed across the lineaments (Table 3).

A. Photolineament 1

The three segments that comprise Photolineament 1 were explored at trenches emplaced at 10 locations within the Mars/Uranus Station valley (Fig. 7). Nine of those trenches exposed faults that, near the surface, trend about N75E and comprise a through-going, north-stepping, en echelon system. The faults form low-angle, subdued scarps in fan Units 2 and 3, but are not surficially expressed in alluvial fan Unit 1 (Fig. 7).

The Photolineament 1 fault system is characterized by a zone of dislocation several feet wide. Movement has been typically accommodated along multiple slip surfaces and fractures. The slip surfaces generally dip steeply 60 to 90 deg to the south with the north side up. Minimum offset was in each case greater than the maximum trench depth, which varied from 1.8 to 3 m. Further, some fault-associated fractures extend into a colluvial wedge, presumably deposited on a scarp formed by an earlier faulting event (Fig. 8). This fact, combined with a minimum offset observed of least 1.8 to 3 m, suggests that the multiple movements have occurred on the Photolineament 1 fault system. However, no site-specific information is presently available from which to judge the recurrence of individual tectonic events.

In Trench 11, the fault offsets a strongly developed buried argillic horizon in fan Unit 2 (Fig. 9). Based on relative profile development, this soil required at least 35,000 to 40,000 years to form; however, it may be much older [12]. Immediately above the fault at this location is a buried cumulic soil profile within fan Unit 1. The cumulic profile is characterized by a weakly developed argillic horizon (Btjk) with disseminated carbonate filaments. This soil is an estimated 7,000 to 12,000 years old [12] and is not offset by the Photolineament 1 fault. Accordingly, while not wholly definitive, the nondisplaced Unit 1 soil horizon exposed in Trench 11, the lack of any fault-controlled geomorphic expression in Unit 1 fans, and the subdued scarps characteristic of the older fans suggest that the last fault displacement most likely occurred in pre-Holocene time.

B. Photolineament 2

Photolineament 2 was trenched at two locations near the intersection with the fault that forms Photolineament 1

(Fig. 7). Both trenches exposed faults generally coincident with lineament trend. This is exemplified in Trench 6 by a N75E fault trend and in Trench 5 where the faults trend about N80W. The Photolineament 2 fault has only minor surface expression in the Unit 3 fans and no expression in the Unit 1 fans.

In Trench 5, the fault offsets Unit 2 alluvium. These sediments were judged to be in the range of 35,000 to 100,000 years old, based on the presence of moderately developed soil profiles. Trench exposures were not sufficient to bracket the most recent offset or to determine the structural relationship between Faults 1 and 2; however, geomorphic evidence suggests that Fault 1 truncates and therefore is younger than Fault 2.

C. Photolineament 3

Three trenches were excavated across the trace of Photolineament 3 (Fig. 7). Trench 12 exposed a fault that displaces Unit 2 and Unit 3 alluvium (Fig. 10).

Trench 13, located east of Trench 12, exposed both unbroken Unit 2 alluvium and a buried paleosol estimated to be at least 35,000 to 45,000 years old. These fault–paleosol relationships suggest that the last displacement of the Photolineament 3 fault occurred between about 35,000 to 100,000 years ago.

D. Photolineament 4

Trench 3 was excavated across the trace of Photolineament 4 (Fig. 7). Unit 2 alluvium was exposed along the entire length of the trench. No faults were observed. It is possible that a fault exists below the maximum trenched depth. However, if a fault exists, the soil stratigraphic age assessments of the unbroken Unit 2 alluvium indicate that the last displacement occurred at least 35,000 years ago and probably well before that time.

VIII. Summary and Conclusions

Multidisciplinary techniques were used to determine the presence and age of faults near the Mars and Uranus antenna sites at the Goldstone DSCC. Photolineament analyses identified four linear trends suggestive of fault control. Site-specific trenching shows that Photolineaments 1, 2, and 3 are fault controlled, but that Photolineament 4 is not. The three, previously unreported, fault systems are dated mainly by geomorphic and soil-stratigraphic techniques. Based on the presence of buried and relict paleosols, with relative development of argillic and calcic horizons varying from slightly to strongly developed, three morphostratigraphic alluvial-fan units were identified: Unit 1, less than about 7,000 to 12,000 years old; Unit 2, approximately 35,000 to 40,000 years old; and Unit 3, estimated to be in excess of 100,000 years old.

The last displacement of Photolineament Faults 1 and 3 occurred between about 12,000 to about 35,000 years ago. Subsurface evidence for the existence of Fault 2, which corresponds with Photolineament 2, is less conclusive; however, when subsurface and geomorphic evidence are combined, faulting is clearly suggested. The structural relationship between Faults 1 and 2 is not known. The most recent displacement of Fault 2 is not known, but is believed to be older than Fault 1, based on geomorphic evidence. A review of Landsat and SLAR imagery indicated that Faults 2 and 3 may be the northwest and southeast extensions of the same fault.

None of the surface traces of the three faults crosses under or projects into existing Mars/Uranus Station structures. Seismic risk assessment of local and regional faults utilizing statistical analysis of earthquake data records and computer fault models based on geologic evidence indicated the maximum credible earthquake (MCE) for the site is a magnitude 7.5 event on the Garlock Fault, which is located about 19 km north of the study area. The new fault location and activity assessment will now permit design engineers to proceed with seismic retrofit designs and loss-prevention measures for mitigation of potential earthquake damage to support facilities and structures.

Acknowledgments

Valuable site geologic information, on-site liaison, and technical contributions provided by Tony Riewe of JPL are appreciated. Also appreciated are the critical reviews of this paper by Howard "Buzz" Spellman, Jr., Chief Geologist of Converse Consultants West, and Perry L. Ehlig, of California State University, Los Angeles.

References

[1] "Geologic and Seismic Refraction Study Proposed Antenna Site Goldstone Tracking Station, Goldstone, California," report prepared by Dames and Moore, Inc., for the Jet Propulsion Laboratory, Pasadena, California, 1982.

[2] E. W. Hart et al., "Summary Report: Fault Evaluation Program, 1986–1987, Mojave Desert Region and Other Areas," California Division of Mines and Geology Open File Report 88-1, Sacramento, California, 1987.

[3] C. W. Jennings et al., "Fault Map of California, Map Scale 1:750,000," California Division of Mines and Geology, Sacramento, California, 1975.

[4] R. K. Dokka and C. J. Travis, "Late Cenozoic Strike-Slip Faulting in the Mojave Desert, California," *Tectonics*, vol. 9, no. 2, pp. 311–340, 1990.

[5] C. E. Glass and D. B. Slemmons, "State-of-the-Art for Assessing Earthquake Hazards in the United States," in Report 11, *Imagery in Earthquake Analysis: Waterways Experiment Station*, U.S. Army Corp of Engineers, Vicksburg, Mississippi, 1978.

[6] T. W. Dibblee, Jr., "Geology of the Fremont Peak and Opal Mountain Quadrangles, California," California Division of Mines and Geology Bulletin 188, Sacramento, California, 1968.

[7] D. B. Burke et al., "Cenozoic Rocks in the Barstow Basin Area of Southern California: Stratigraphic Relations, Radiometric Ages, and Peleomagnetism," U.S. Geologic Survey Professional Paper 1529-E, 1982.

[8] G. E. Christenson and C. Purcell, "Correlation and Age of Quaternary Alluvial-Fan Sequences, Basin and Range Province, Southwestern United States," Geological Society of America, Special Paper 203, pp. 115–122, 1985.

[9] M. N. Machette, "Dating Quaternary Faults in the Southwestern United States by Using Buried Calcic Paleosols," *Journal of Research, U.S. Geological Survey*, vol. 6, no. 3, pp. 369–381, 1978.

[10] R. J. Shlemon, "Quaternary Soil-Geomorphic Relationship, Southeastern Mojave Desert, California and Arizona," in *Geology and Mineral Wealth of the California Desert* by T. W. Dibblee, Jr., South Coast Geological Society, Santa Ana, California, pp. 388–402, 1980.

[11] R. J. Shlemon, "Application of Soil-Stratigraphic Techniques to Engineering Geology," *Bulletin of the Association of Engineering Geologists*, vol. 22, pp. 129–142, 1985.

[12] R. J. Shlemon, "Geomorphic and Soil-Stratigraphic Assessment of Fault Activity, JPL–Goldstone Deep Space Communications Complex, Mars Station, San Bernardino County, California," appendix report prepared for Converse Consultants for the Jet Propulsion Laboratory, Pasadena, California, 1989.

Table 1. Aerial photography, Landsat, and SLAR imagery reviewed.

Type	Date	Project	Scale	Roll/Frame	Source
Landsat TM[a]	07/28/85	–	1:155,000	50514-17515	JPL (EOSAT)
Side-looking airborne radar (SLAR)	12/85	TRPA[b]	1:250,000	1/4-7	USGS[c]
Low-altitude aerial photography					
Color infrared	06/25/83	NHAP 83[d]	1:58,000	89/122, 123 95/136, 137	USGS
Black and white	06/25/83	NHAP 83	1:80,000	62/23, 124 66/100, 101	USGS
Black and white	10/09/87	87156[e]	1:18,000	1/8-10 2/8-10 3/11-13	JPL
Black and white	01/14/82	–	1:6,000	–/1-8	JPL

[a] Thematic mapper.

[b] Trona Project Area.

[c] U.S. Geological Survey.

[d] National High-Altitude Photography, flown in 1983.

[e] JPL photo number.

Table 2. Characteristics of photolineaments.

Photo-lineament	Length, km	Imagery in which lineament is visible	Geomorphic evidence suggestive of faulting	Projects close to existing structures	Remarks
1 (1, 1a, 1b)	>4.8	SLAR, aerial photos	Strong	Yes	Strong lineament defined by a combination of features, including obvious topographic alignments, linear scarps on alluvial fans, rectilinear depressions, and linear changes in erosional texture.
2	>16 (regional)	Landsat, SLAR, aerial photos	Moderate	Yes	Regional lineament visible on high-altitude imagery, locally defined by linear changes in drainage/ erosional texture and photographic tone.
3	>16 (regional)	Landsat, SLAR, aerial photos	Moderate	Yes	Regional lineament visible on high-altitude imagery, probably represents southeast extension of Lineament 2.
4	0.64±	Aerial photos	Weak	Yes	Local lineament (<609 m long) defined by a linear break in slope and an apparently offset stream channel.

Table 3. Summary of fault trenching.

Photo-lineament	Identified as a fault during exploration	Estimated minimum age of most recent activity, ybp	Remarks
1	Yes	~12,000	Through-going fault system dated by geomorphic and soil-stratigraphic methods. Fault trends roughly N75E and is characterized by north-stepping segments up to 2.4 km long. Minimum offset is 1.8 to 3 m, north side up. Some evidence of multiple events.
2	Yes	>12,000	Coincident with regional lineament identified on Landsat imagery. Trenching and geomorphic evidence suggest fault origin. Geomorphically it appears to be truncated by Fault 1. However, this was not demonstrated during exploration. The structural relationship with Fault 1 and the approximate age of this fault are yet unclear.
3	Yes	35,000	Fault is coincident with regional lineament identified on Landsat imagery. Dated by geomorphic and soil-stratigraphic methods. Probably forms southeast extension of Lineament 2.
4	No	–	Trenching across the trace of this lineament exposed unbroken 35,000 to 100,000-year-old deposits. Photolineament 4 is not fault related.

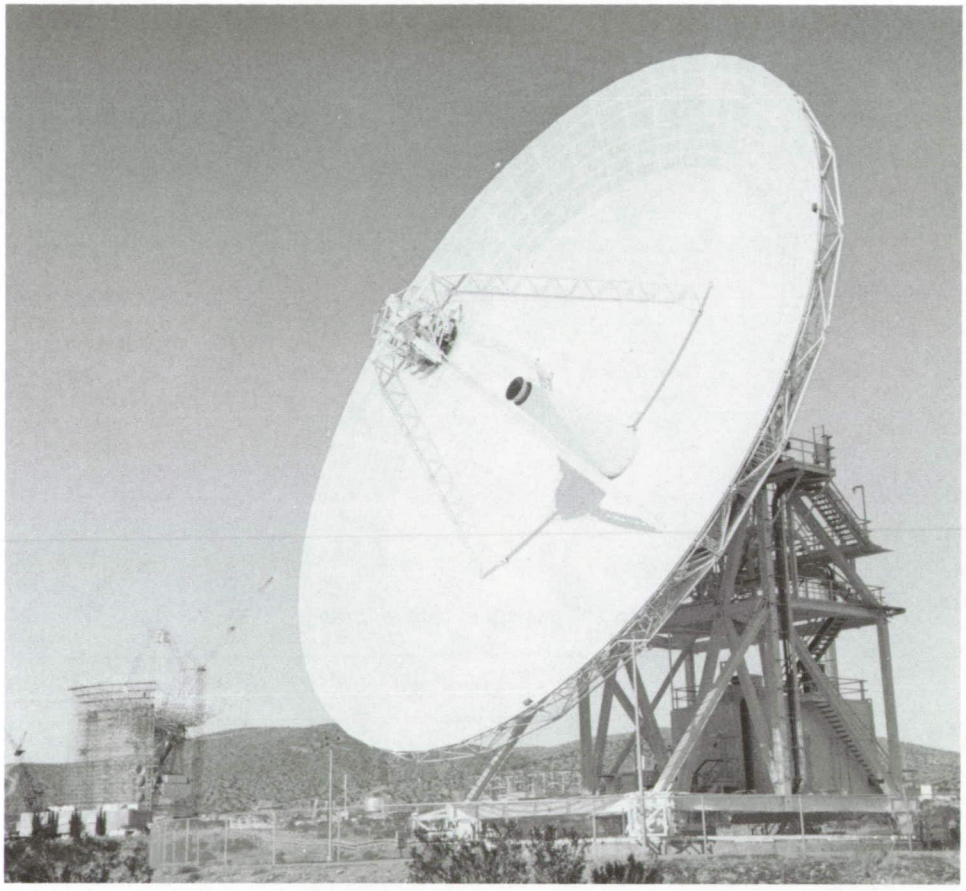


Fig. 1. The 34-m Uranus antenna (DSS 15). The 70-m Mars antenna (DSS 14) is in the background.

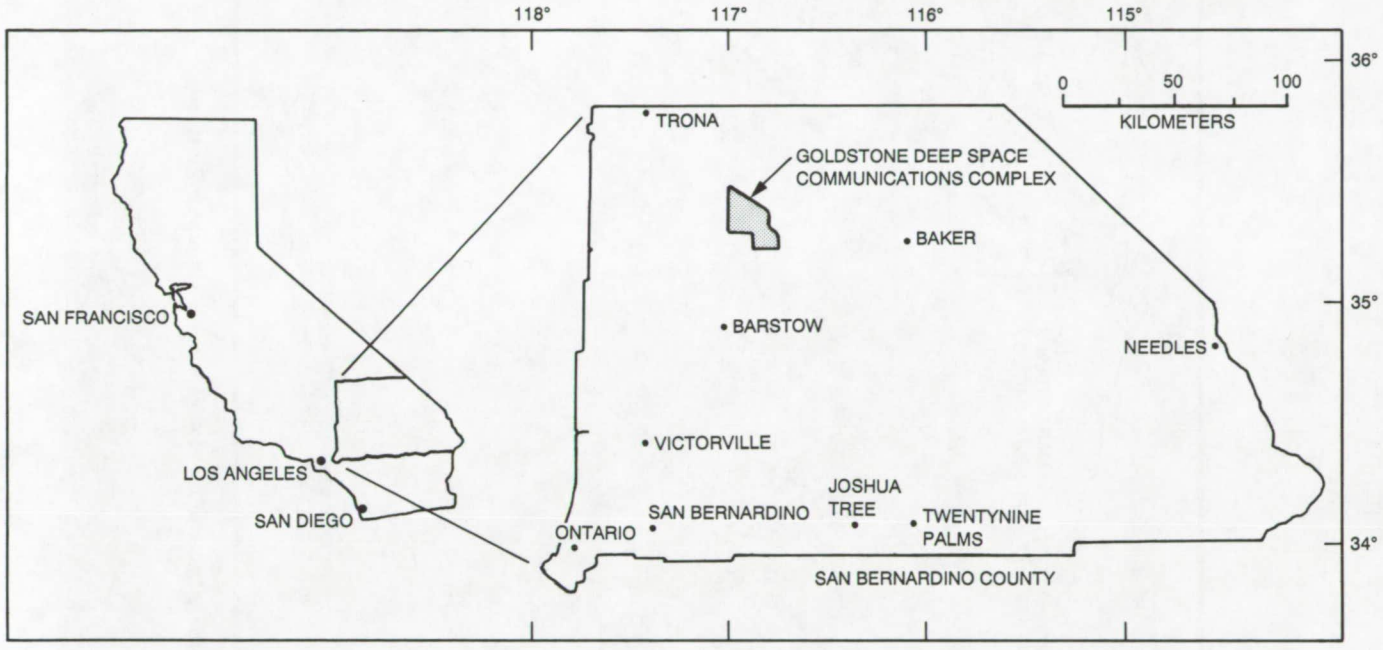


Fig. 2. Map showing the Goldstone Deep Space Communications Complex in the Mojave Desert, California.

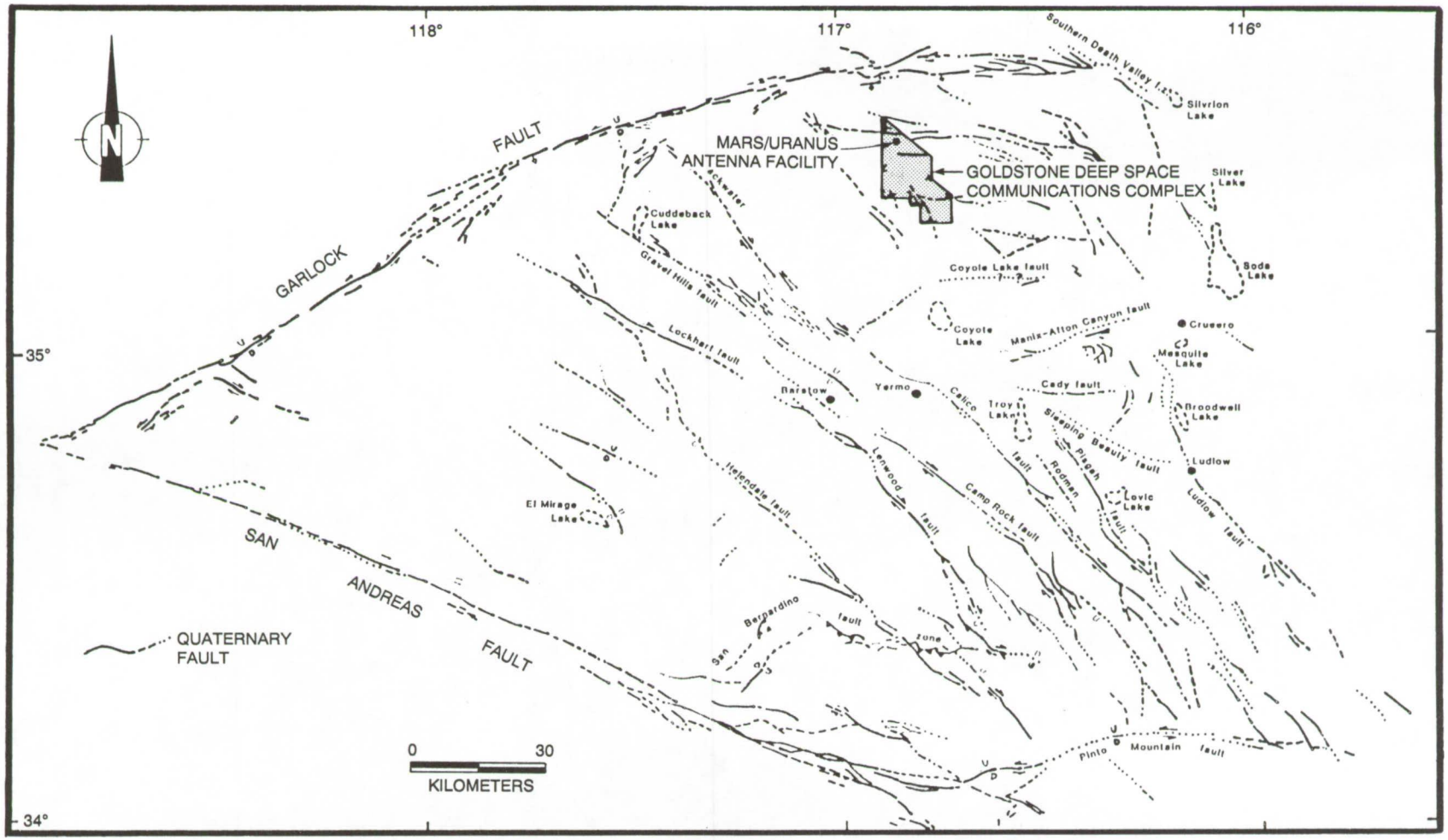


Fig. 3. Major Quaternary faults of the Mojave Block Geomorphic Province.

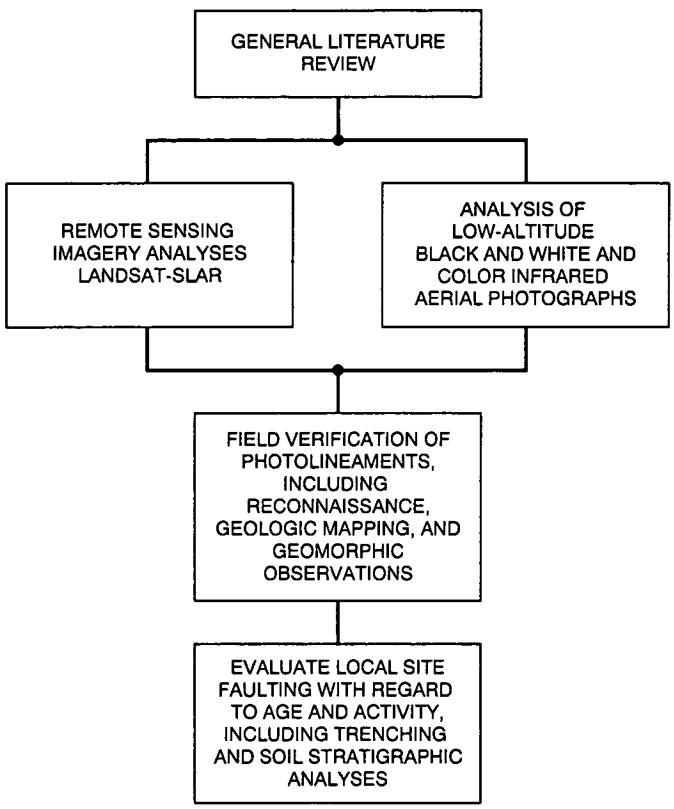
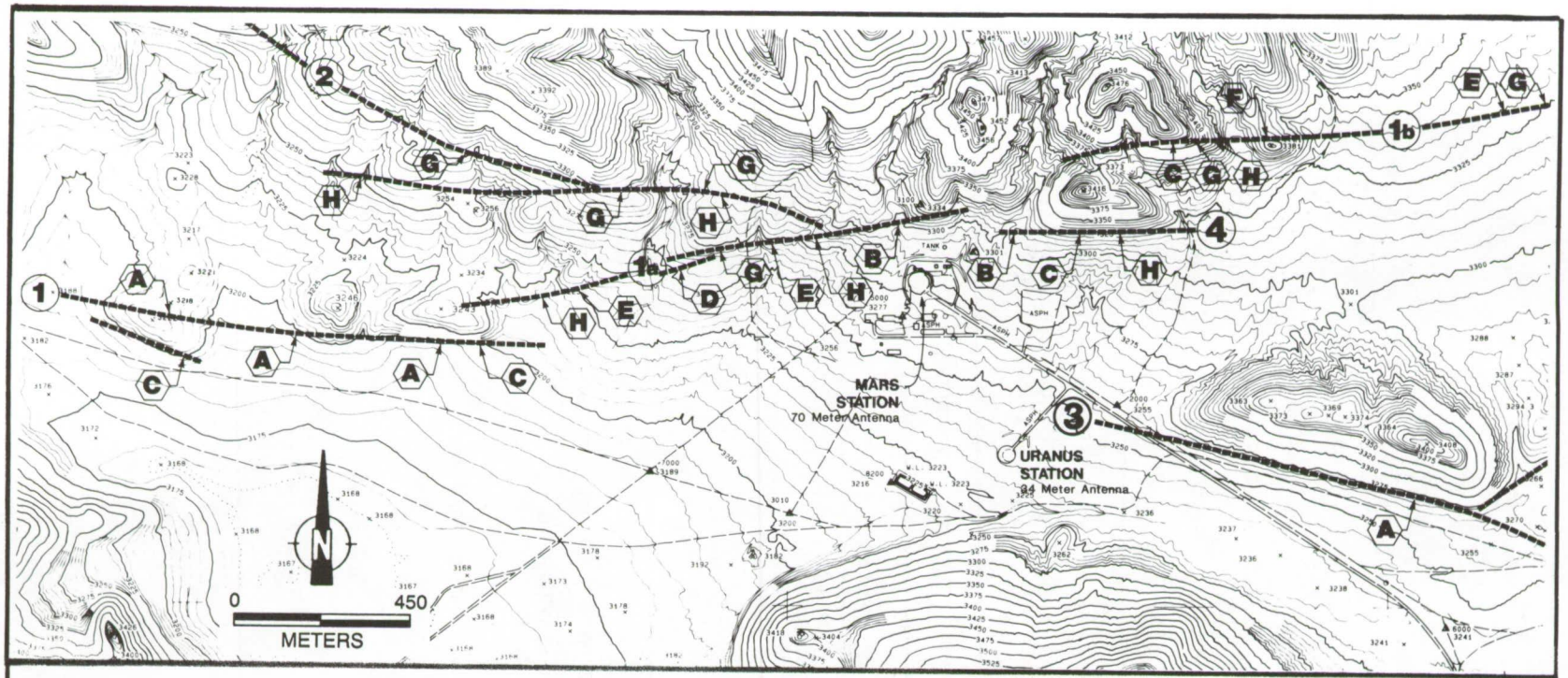
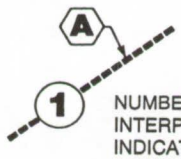


Fig. 4. Summary of technical approach.



GEOMORPHIC FEATURES RECOGNIZED DURING IMAGERY ANALYSIS: EVIDENCE OF PROBABLE QUATERNARY-AGE FAULTING

- A VEGETATION AND/OR TOPOGRAPHIC ALIGNMENT
- B POSSIBLE OFFSET STREAM CHANNEL
- C LINEAR BREAK IN SLOPE
- D RECTILINEAR AND/OR CLOSED DEPRESSION
- E LINEAR SCARP ON ALLUVIAL FAN
- F UPHILL FACING SCARP
- G LINEAR CHANGE IN PHOTOGRAPHIC TONE (COLOR)
- H LINEAR CHANGE IN DRAINAGE OR EROSIONAL TEXTURE



NUMBER AND APPROXIMATE LOCATION OF INTERPRETED PHOTOLINEAMENT, HEXAGON INDICATES LOCATION AND TYPE OF GEOMORPHIC FEATURE

Fig. 5. Photolineament map of the Mars/Uranus Station valley.

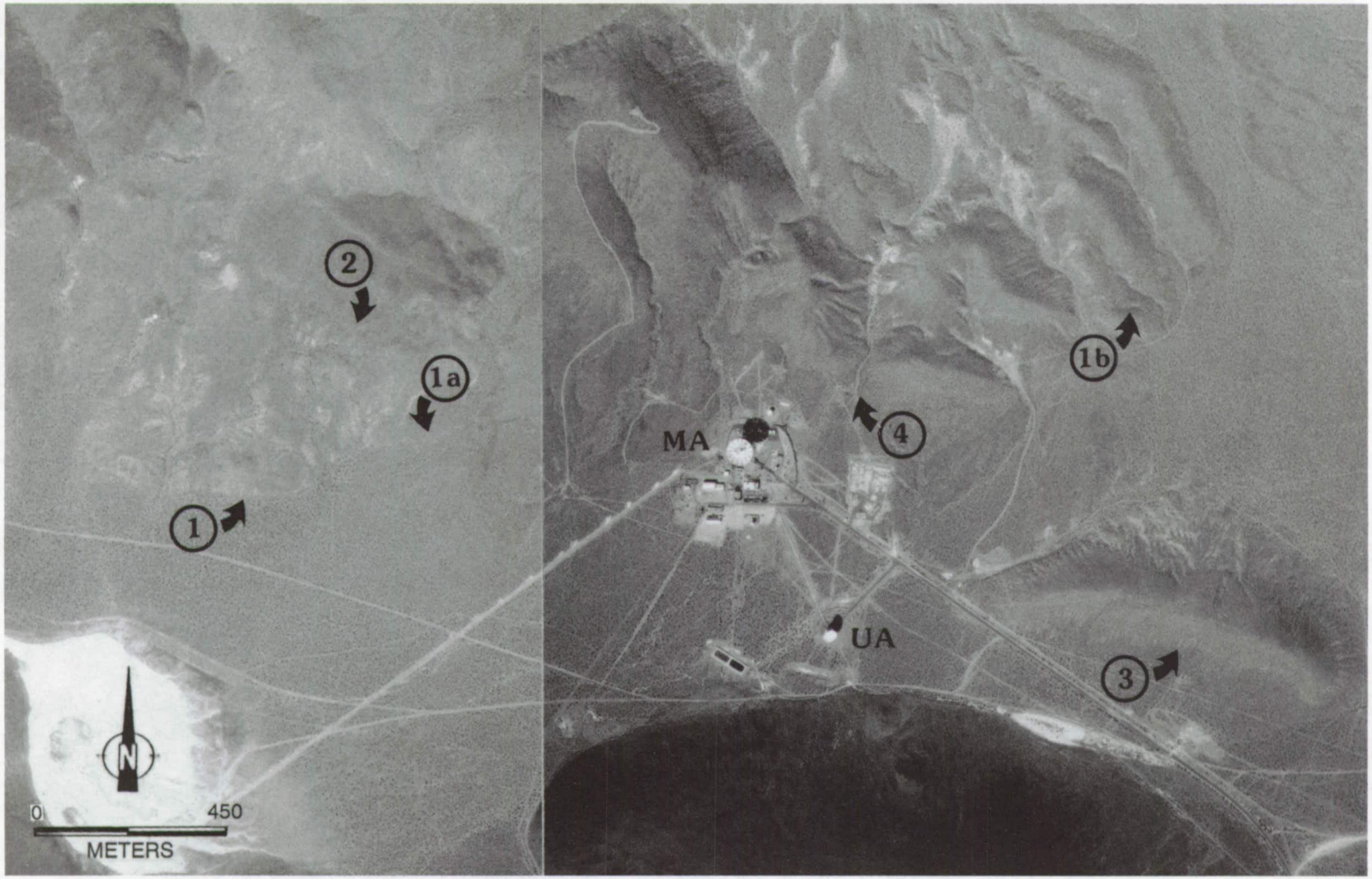


Fig. 6. Composite aerial photo of Mars/Uranus Station valley showing photographic expression of lineaments, with MA = Mars antenna and UA = Uranus antenna. Note the change in morphology and darker shade of alluvial-fan surfaces.

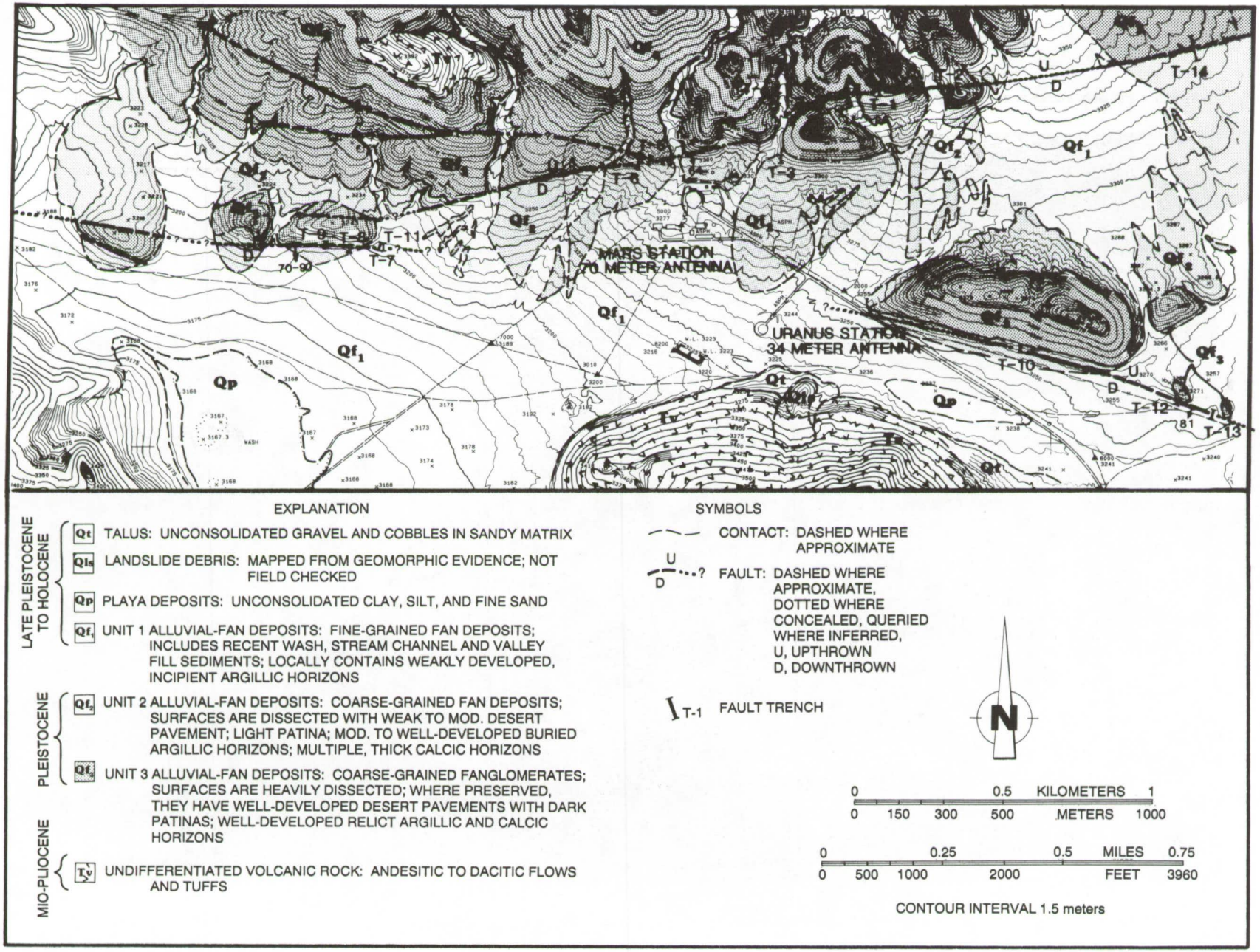


Fig. 7. Generalized geologic map of the Mars/Uranus Station valley, Goldstone DSCC.

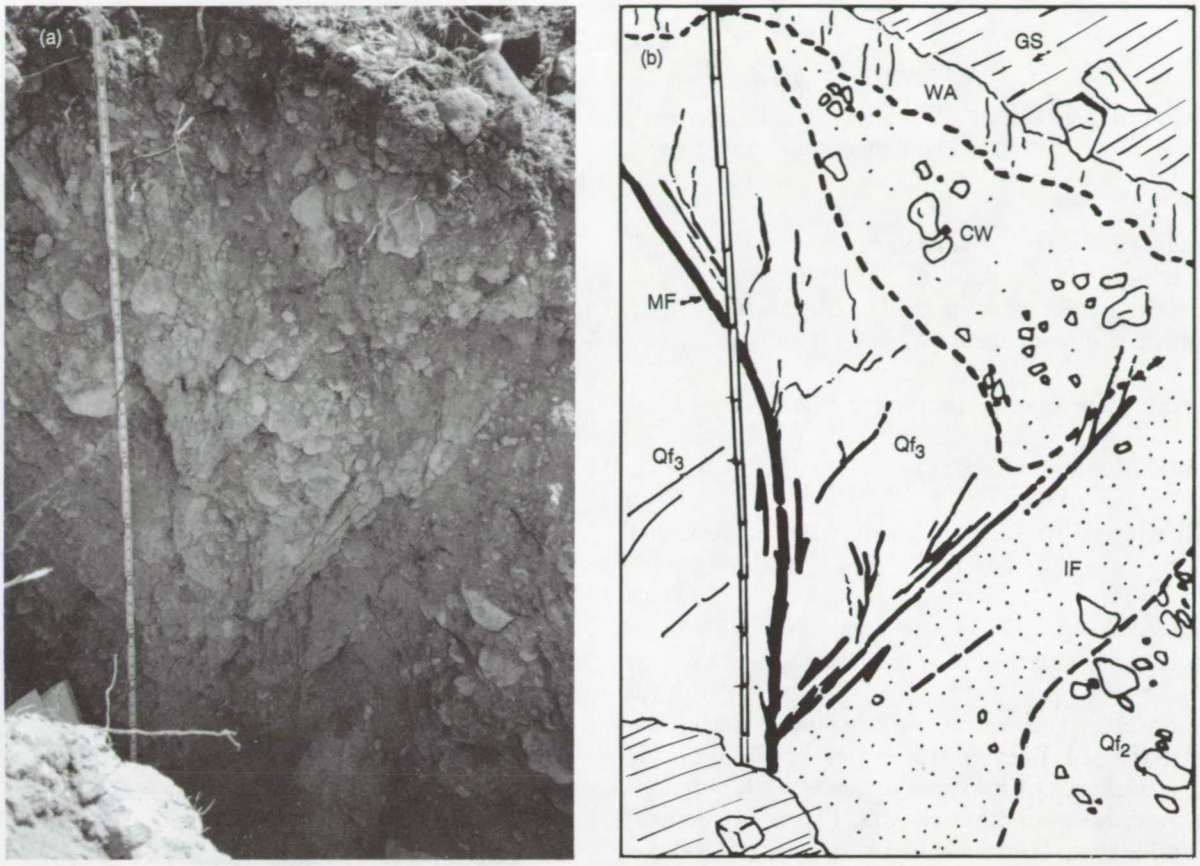


Fig. 8. Photo (a) and sketch (b) of fault observed in Trench 8 showing offset alluvial units and possible offset of scarp derived colluvial wedge. Qf_3 = Unit 3 fanglomerate, Qf_2 = Unit 2 alluvium, CW = colluvial wedge, WA = weathered alluvium below active geomorphic surface, IF = loose, fine-grained infilling, MF = main fault splay, and GS = ground surface. The folding rule is 1.8 m long.

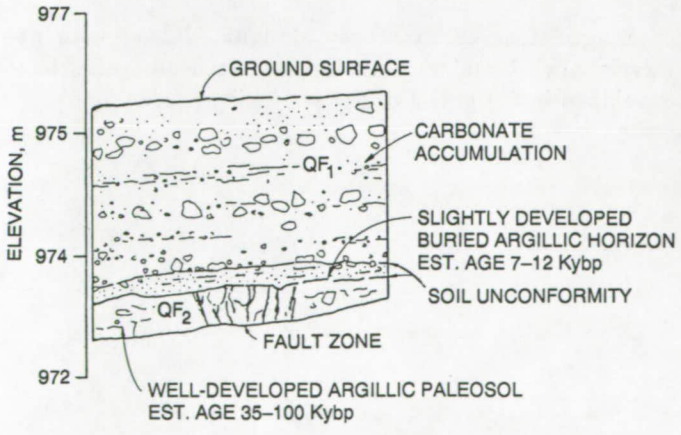


Fig. 9. Sketch of fault observed in Trench 9 showing diagnostic soil horizons used to bracket the age of most recent fault activity.

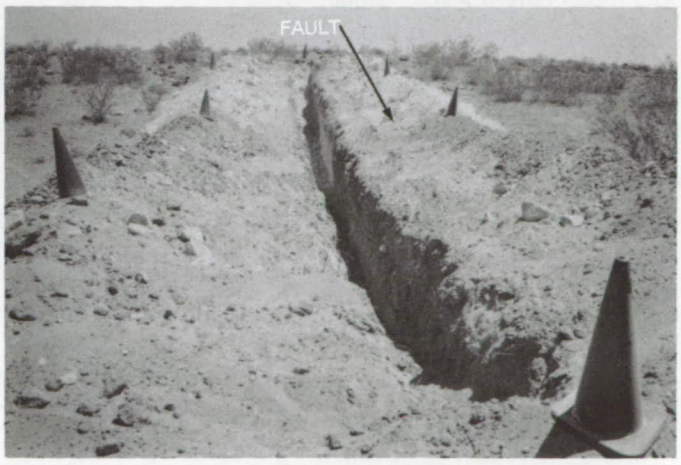


Fig. 10. Fault exposed in Trench 12 that juxtaposes Unit 2 and Unit 3 alluvium. Note color change at fault: Unit 3 is light colored, Unit 2 is dark colored.

Appendix

Glossary

A horizon: Uppermost soil horizon (loosely, topsoil); usually has roots and a mixture of organic and mineral matter.

B horizon: Below the A horizon; characterized by illuvial or in situ accumulation of clay, iron, or aluminum; residual accumulation of sesquioxides; darker, redder color due to presence of sesquioxides. Loosely: subsoil.

Alluvial: Deposited by running water.

Argillic: Pertaining to clay.

Argillic horizon: Diagnostic subsurface horizon characterized by an accumulation of clay.

Calcic horizon: Diagnostic subsurface horizon, at least 15 cm thick; enriched in secondary carbonates.

Developed: The terms weakly, moderately, and strongly developed refer to categories of paleosols described by thickness of horizons, types of materials built up in the horizons, and other criteria. (There are actually five categories altogether, including very weakly developed and very strongly developed).

Diagnostic subsurface horizon: Soil horizon that forms below the surface and is used to classify soils. Usually a B horizon (e.g., argillic, spodic, cambic) but may be A (albic) or C (calcic, duripan). B and C horizons occur at surface only if soil is truncated.

Eluviation: Downward movement of soluble or suspended material in soil from A (eluvial) to B (illuvial) horizon.

En echelon: Stepped series of parallel faults or other structures.

Fanglomerate: Conglomerate formed as part of an alluvial fan.

Fluvial lens: Buried stream or river bed deposits seen in subsurface cross section; typically lens-shaped.

Horizon: Interface indicative of a particular position in a stratigraphic sequence.

Illuviation: Accumulation in a lower soil horizon of soluble or suspended material that was exported from an upper horizon by eluviation.

Lineament: Linear topographical feature thought to reflect subsurface structure.

Paleosol: Buried, ancient (fossil) soil horizon, especially one formed during an interglacial period.

Pedogenic: Pertaining to soil formation.

Photolineament: Linear topographical feature in an aerial photograph.

Relict: Remnant of topographic feature that remains after other parts have been removed.

Stage: One of six stages, designated by roman numerals, of carbonate accumulations in paleosols that have developed in (1) gravel or (2) sand, silt, or clay.

A Method of Optimal Radio Frequency Assignment for Deep Space Missions

C. J. Ruggier, J. M. Gevargiz, L. H. Truong, and K. S. Suwitra
Telecommunications Systems Section

A method for determining optimal radio frequency channels for the Deep Space Network is described. Computer automated routines calculate interference-to-signal ratios over a given mission period and provide a quantitative assessment of the channels which could then be assigned to a new mission. This automated procedure reduces the analysis time considerably and effectively improves upon the accuracy of existing channel assignment techniques.

I. Introduction

The continuously increasing demand for communications channels by the Deep Space Network (DSN) has necessitated the development of more extensive methods of selecting channel frequencies which best minimize the overall potential of mutual interference. Communications channels must be assigned judiciously to new DSN missions with the objective of achieving and maintaining an optimal level of intra-system compatibility. Transmission link and dynamic geometrical parameters which pertain to the spacecraft and tracking station are used in the computations to determine the most suitable channel frequency for both the uplink and downlink transmission modes.

This article presents a method of determining the optimal channel frequencies for new DSN missions.[1,2] A computer model has been developed which calculates the mutual signal degradation between systems as it relates to any given channel assignment. The model utilizes parameters which include the effects of spacecraft position and pointing angles, the modulation schemes, data rates and formats, Doppler shift, and discrete ranging components. An algorithm calculates the optimal frequency for the new mission, a frequency that will have a minimal interference impact on the overall DSN system. The computer program, called the Interference Analysis Program (IAP), is for the most part database driven to provide a high level of automation in the computation.

In the past, there were only a few missions to contend with, and the channel assignment procedure relied heavily on qualitative evaluation techniques. Now with the rapid growth of space research missions, more definitive analyses are needed to ensure efficient use of the DSN frequency spectrum. As the complexity in the analysis increases, there is a concomitant need to improve the accuracy of the analysis and, at the same time, lessen the analyst's dependence on qualitative assessment. The method described in this article effectively reduces the potential for human

[1] J. Gevargiz and C. Ruggier, *DSN Intra-System Spectral Compatibility Analysis, Mars '94 Channel Assignment*, JPL D-7663 (internal document), Jet Propulsion Laboratory, Pasadena, California, July 12, 1990.

[2] D. Bishop, *DSN Inter-System Spectral Compatibility Analysis: CRAF/Cassini Channel Assignment*, JPL D-8797 (internal document), Jet Propulsion Laboratory, Pasadena, California, August 6, 1991.

error and provides a basis for more objective, standardized analysis techniques.

II. Interference Geometry

The criteria for selecting optimal frequencies for a new DSN mission are dependent on the potential for link performance degradation caused within the DSN missions as a whole; that is, performance degradation caused to the existing missions with the introduction of the new DSN mission and, alternately, performance degradation to the new mission caused by the existing missions. This applies to both the uplink and downlink transmission modes.

Figures 1 and 2 show how the interference signal couples to the receiver antenna of a tracked mission for the uplink and downlink modes, respectively. The value R_1 is the range between the Earth station and the tracked spacecraft, SC1, while R_2 is the range between the Earth station and the interfering or interfered-with spacecraft, SC2. The antenna off-axis angle ϕ determines the gain of the interfering signal for the uplink and downlink modes. These trajectory-dependent parameters are used to calculate the mutual interference power spectral density between SC1 and SC2 for any given mission period. The interference-to-signal ratio (ISR) can also be calculated and used as a simple performance factor for evaluating the interference degradation to the DSN.

A. Uplink Interference Mode

From Fig. 1, it can be noted that the uplink signal to SC1 can also couple through the sidelobes of the SC2 antenna and interfere with its received command signal. The absolute interference power level is dependent on the antenna off-axis angle relative to SC2 and the range of SC2.

Typical DSN uplink command signals require only a relatively narrow transmission bandwidth, usually on the order of a few kilohertz. With the narrow-bandwidth requirement and a sufficient guard band between channels, the uplink mode is not considered to be a major factor in determining new channel assignments; however, in some cases, it can significantly impact the accuracy of the overall performance assessment. For this reason, the uplink interference mode is included as an integral part of the channel assignment analysis.

B. Downlink Interference Mode

Referring to Fig. 2, the downlink signal from SC2 is shown to couple through the sidelobes of the ground station receiving antenna which tracks SC1. The level of absolute interference power from SC2 is dependent on the antenna off-axis angle of SC2 with SC1 and the range of SC2.

Unlike the uplink command signals, the downlink signals require a wider bandwidth to accommodate high data rate telemetry and, in some instances, ranging tones which are several megahertz apart. Due to the downlink's wider transmission bandwidth, the signal's power spectrum can spread over a large segment of the frequency band. Consequently, a more stringent approach is needed for the assignment of new channels in the downlink band.

III. Description of the Interference Models

The basic configuration of the IAP analysis models is shown in Fig. 3; this configuration shows a cochannel interference model, an adjacent channel interference model, a Doppler shift model, a frequency optimization algorithm, and a discrete tone analysis model. These models, in turn, are driven by the spacecraft trajectory model and from parameters stored in the mission database. The primary functions of each model are described in the following sections.

A. Spacecraft Trajectory Model

This algorithm calculates the spacecraft range, downlink antenna off-axis angle, and range rate. The trajectory model is driven from a mission database, which contains the spacecraft state vectors and timing data, significant mission event profiles, and the telecommunications parameters required for the uplink and downlink mode analysis.[3]

B. Cochannel Interference Model

Figure 4 illustrates the flow diagram for the cochannel interference analysis between SC1 and SC2. The model calculates the absolute power, power spectral density, and ISR for the case in which the interference signal frequency is coincident with that of the interfered-with signal over the interfered-with spacecraft mission period. This applies to both the uplink and downlink interference mode.

As a first step in the analysis, the total sample of active DSN missions is culled and limited to a sample consisting of only those missions which cause, or are susceptible to, interference in the cochannel mode.

As an example, if isotropic gain of 0 dB is assumed for the spacecraft antenna, from Fig. 2 (downlink mode) the ISR at the ground station is given by

[3] K. Suwitra, "Source Code for the Automated Interference Analysis Program (IAP)" Interoffice Memorandum 3396-92-061, (internal document), Jet Propulsion Laboratory, Pasadena, California, October 19, 1992.

$$ISR = \frac{P(\text{interfering spacecraft})}{P(\text{tracked spacecraft})}$$

$$= \frac{P^d_{C_2} \cdot G(\phi)/L(R_2)}{P^d_{C_1} \cdot G_{\max}/L(R_1)} \quad (1)$$

where

$P^d_{C_1}$ = downlink (d) cochannel transmitted power of SC1 (watts)

$P^d_{C_2}$ = downlink cochannel transmitted power of SC2 (watts)

$G(\phi)$ = off-axis gain of SC2 link, ground station antenna, in the direction of SC2 (ratio)

$G_{\max}$ = maximum ground station antenna gain (ratio)

$L(R_2)$ = $(4\pi R_2/\lambda)^2$ = free space loss for distance R_2 (ratio)

$L(R_1)$ = free space loss for distance R_1 (ratio)

R_1 = range between ground station and SC1 (meters)

R_2 = range between ground station and SC2 (meters)

λ = wavelength of signal (meters)

With the interfering and interfered-with signals in the same frequency band, this expression reduces to

$$ISR = \frac{P^d_{C_2} \cdot G(\phi) \cdot (R_1)^2}{P^d_{C_1} \cdot G_{\max} \cdot (R_2)^2} \quad (2)$$

The antenna off-axis gain $G(\phi)$ is given by the International Radio Consultative Committee (CCIR) antenna reference pattern [1]:

$$\left.\begin{array}{l} \text{for } \dfrac{D}{\lambda} \geq 100 \\ G^*(\phi) = G^*_{\max} - (2.50 \times 10^{-3})\left[\dfrac{D\phi}{\lambda}\right]^2 \text{ (dB)} \\ \qquad \text{for } 0 < \phi < \phi_m \\ G^*(\phi) = G_1 \text{ (dB) for } \phi_m \leq \phi < \phi_r \\ G^*(\phi) = 32 - 25 \log_{10}(\phi) \text{ (dB)} \\ \qquad \text{for } \phi_r \leq \phi < 48 \text{ deg} \\ G^*(\phi) = 10 \text{ (dB) for 48 deg} < \phi \end{array}\right\} \quad (3)$$

where

D = antenna diameter (meters)

λ = wavelength (meters)

$G^*_{\max}$ = $10 \log_{10}(G_{\max})$ = antenna off-axis gain (dB)

$G^*(\phi)$ = $10 \log_{10} G(\phi)$ = antenna off-axis gain (dB)

ϕ = off-axis angle of the first sidelobe (deg)

G_1 = gain of the first sidelobe = $2 + 15 \log\left(\frac{D}{\lambda}\right)$ (decibels)

$$\phi_m = \left[\frac{20\lambda}{D}\right]\sqrt{G^*_{\max} - G_1} \text{ (degrees)}$$

$$\phi_r = 15.85\left[\frac{D}{\lambda}\right] \text{ (degrees)}$$

The carrier-to-carrier, data-to-data, and total-power ISR's are calculated and compared to the given interference threshold power ratio. For the carrier-to-carrier mode, the adjusted (after modulation) interfering and interfered-with carriers are compared to a user-defined threshold in decibels. If the threshold is exceeded over the mission period of the interfered-with system, then that mission pair is considered for further analysis. If the threshold is not exceeded, then the data-to-data mode is examined. Similarly, for this mode, the data power of the interfering and interfered-with systems are compared to a user-defined threshold. In the event that user-defined threshold levels are not given, default values are used. Figure 5 illustrates the downlink total power cochannel ISR from Galileo to Cassini versus the days past the launch date of the Cassini mission. This figure also shows the ISR threshold of -20 dB employed in this analysis.

The uplink interference analysis is similar to the downlink analysis with the ISR substituted for the absolute interference power. The received uplink interference power at SC2 is given by

$$\left.\begin{array}{ll} P^u_{C_2} = \dfrac{P^{GS}_1 \cdot G_2(\phi)}{L(R_2)} & \text{uplink } (u) \text{ cochannel interference power to SC2} \\ P^u_{C_1} = \dfrac{P^{GS}_2 \cdot G_1(\phi)}{L(R_1)} & \text{uplink } (u) \text{ cochannel interference power to SC1} \end{array}\right\} \quad (4)$$

where

P^{GS}_1 = ground station (GS) transmitted power to SC1

$G_2(\phi)$ = off-axis transmitted gain of SC1 link antenna, ground station antenna, in the direction of SC2

$G_1(\phi)$ = off-axis transmitted gain of SC2 link antenna, ground station antenna, in the direction of SC1

$L(R_2) = (4\pi R_2/\lambda)^2$ = free space loss for distance R_2 (ratio)

For example, Fig. 6 illustrates the uplink cochannel interference power from Galileo to Cassini versus the days past the launch date of the Cassini mission. The results of the cochannel analysis are then used to eliminate the potential sources of interference from further analysis when for a given interference source,

(1) The absolute power of the uplink interference does not exceed the spacecraft's interference power threshold.

(2) The downlink carrier-to-carrier interference power ratio does not exceed the interference threshold.

(3) The data-to-data interference power ratio does not exceed the interference threshold.

(4) The total-power ISR does not exceed the interference threshold during the entire mission period.

C. Doppler Shift Model

Signals are subjected to Doppler frequency shifts, which although occurring periodically, could cause intolerable interference to the system. The Doppler shift of the interference signal relative to the interfered-with signal is used to calculate the instantaneous changes in the interference power throughout the mission period being analyzed. Maximum Doppler shift is derived from the Doppler rate for each day in the mission period under analysis. The Doppler shift data are then applied as an adjustment factor in the time- and frequency-dependent interference calculations.

The Doppler shift is an important consideration in the channel assignment analysis in that, for some period of time, it can either increase or decrease the degree of isolation between channels.

D. Adjacent Channel Interference Model

Figure 7 illustrates the adjacent channel analysis for a pair of missions (e.g., SC1 and SC2). This model calculates the absolute power, power spectral density, and ISR for the case in which the new DSN link (i.e., SC1), operates on a channel other than that used by the existing DSN mission (i.e., SC2). This procedure is applied to both the uplink and downlink modes.

The adjacent channel interference model is used to calculate the total interference power which couples to the ground station and spacecraft receiver of the interfered-with DSN mission. It can generate plots of the interference power versus time and also provides the basis for the frequency optimization procedure.

The adjacent channel interference analysis constitutes the core of the channel assignment process. Due to the sideband products of a DSN signal, portions of the transmitted power of the interference source will overlap into an adjacent channel user. This interference signal couples spatially, through the antenna sidelobes, and also spectrally, between channels. Interference caused by inadequate frequency spectrum isolation between the band channels is generally referred to as "adjacent channel interference."

In contrast to cochannel interference, the level of the incident adjacent channel interference depends greatly on the rejection properties of the receiver. The problem is that a typical DSN telemetry signal occupies a spectral bandwidth in excess of the channel bandwidth limitation. Sideband products spill over into adjacent channels and can still cause interference, even though separated from other users by several channels.

A method of calculating the interference power incident on a system operating in an adjacent channel is inherently complex. As an example, the typical composite DSN telemetry signal can react to the interference with a loss of carrier lock, a loss of telemetry lock, or a degradation in the output signal-to-noise ratio (increase in the bit error rate). These various forms of link degradation are not necessarily correlated and will depend on the spectral characteristics of the composite interference signal. To overcome these difficulties, a simplified method is needed for modeling the interference components and their effect on the interfered-with system.

1. Spectral Power Envelope Technique. A practical and simple method of calculating adjacent channel interference utilizes the spectral power envelope technique. This technique is used as a worst-case representative model of the interfering signal characteristics.

The spectral power envelope of the interfering signal is constructed using a simple procedure. The procedure de-

fines the signal's power envelope as the upper bound of its spectral power density, limited by the peaks of its spectral components. In the case of the DSN uplink and downlink signals, the power envelope is constrained by the carrier and sideband peak power levels. When the signal's power level, data format, and modulation scheme are specified, a representative model of the spectral power envelope of the composite signal can then be implemented for any particular link.

The spectral power envelope technique is a simplified, first-order approach for the assessment of adjacent channel interference. From the interference geometries and the specified link parameters, the in-band interference power and the ISR are easily calculated. These parameters are applied in the analysis model, as a first step in the evaluation of link performance degradation. Figure 8(a) illustrates an example of the power spectral envelope for missions corresponding to SC1 and SC2. Shown is the spectral power envelope of the interference, $P(f)$, and the harmonics of the interfered-with signal, with the harmonic number labeled $i = 1$, 3, 5, and 7. Also shown in this figure is the interfered-with frequency band, presented as a shaded area labeled $P_I(i)$ for $i = 1$, 3, 5, and 7. The adjacent channel algorithm calculates the interference from mission SC1 to mission SC2 by determining the interference power that falls within the data bandwidth of the subcarrier components of mission SC2, shown as a shaded area in Fig. 8(a). The total adjacent channel interference power P_A, is then given by

$$P^d_{A_1}(i, f_c) = \sum_{j=-N}^{N} \frac{1}{j^2} P^d_{A_1}(i, j, f_c) \qquad (5)$$

where

j = subcarrier harmonic number

i = day number

f_c = channel frequency of mission SC1

$P^d_{A_1}(i, f_c)$ = total downlink adjacent channel interference power from SC1 when operating on channel 18 (f_c)

$P^d_{A_1}(i, j, f_c)$ = adjacent channel interference to the jth subcarrier harmonic of SC2 when SC1 operating on channel 18 (f_c)

Similarly, as shown in Fig. 8(b), the interference to mission SC1 can be calculated using the same algorithm.

An analysis using relative signal power levels is sufficient for conducting the frequency optimization procedure, thereby significantly reducing the complexity of the analysis and the required computer execution time. Although these interference parameters are not, in themselves, sufficient to characterize the actual link performance degradation, they provide a first-order assessment, which satisfies the basic requirements of the analysis.

2. Adjacent Channel Interference Power Calculations. In general, calculation of the interference power $P^d_{A_1}(i, f_c)$ at the interfered-with receiver involves an integration of the total interference power within the receiver's bandwidth. The interference spectral power envelope, P_{spd}, is multiplied with the magnitude squared of the receiver transfer function and then integrated over its spectral bandwidth.

$$P^d_{A_1}(i, j, f_c) = \int_{f_1(j)}^{f_2(j)} P^d_{spd_1}(i, f_c, f) \cdot |H(f)|^2 \, df \qquad (6)$$

where $f_1(j)$ and $f_2(j)$ are the frequency limits of the interference signal within the interfered-with receiver's bandwidth for the jth subcarrier harmonic of mission SC2; $P^d_{spd_1}(i, f_c, f)$, is the spectral power envelope of the interfering signal; and $|H(f)|$ is the magnitude of the receiver transfer function.

For the downlink case, employing the Block III and IV receiver, the receiver bandwidth is referenced to the final IF stage. Similarly, for the uplink case, the typical transponder receiver bandwidth is limited by the predetector filter.

A particularly useful parameter for assessing interference is the ISR. One of the outputs of the adjacent channel interference model is a plot of the aggregate ISR of the new mission and another mission operating in the same frequency band, for some given period of time. As an example, Fig. 9 shows a plot of the downlink adjacent channel ISR for Cassini versus Galileo[4] in the 8-GHz band. This plot shows the periods where the interference levels are expected to exceed the recommended limit.

E. Frequency Optimization Model

An acceptable procedure for determining the optimal channel for a new mission requires an interference analysis that evaluates the overall effect in the DSN as a system,

[4] See footnote 2.

considering both the uplink and downlink telecommunication. This involves the application of some basic analytical calculations, formulation of realistic assumptions, and conclusive evaluation. The automation of the procedure requires implementation of a method which provides a quick and accurate assessment, by which the overall interference impact to the DSN system can be demonstrated parametrically as a function of frequency. Consequently, an optimal channel can be determined without the need for labor-intensive analysis.

The power spectral envelope of the interfering signal within the tracked receiver bandwidth is integrated as a measure of degradation due to interference. The procedure is performed for the uplink and downlink modes, and it takes into account the interference at the SC1 link and the SC2 link (see Figs. 1 and 2), yielding a degradation factor (F_{dgr})

$$F_{dgr} = P^d_{T_1} + P^u_{T_1} + P^d_{T_2} + P^u_{T_2} \qquad (7)$$

where

$$P^d_{T_1} = \sum_{i=1}^{N} P(i)$$

for $P(i) = P^d_{A_1}(i, f_c)$ when $P^d_{A_1}(i, f_c) \geq$ interference criterion.

$$P(i) = 0 \quad \text{otherwise} \qquad (8)$$

where F_{dgr}, $P^d_{A_1}(i, f_c)$, and $P^d_{T_1}$ are, respectively, the degradation factor, the adjacent interference power given by Eq. 5, and the total interference power over the mission period of N days when exceeding the interference criterion.

This process is repeated with the interfering signal placed on all the possible channels to produce a complete set of degradation factors which are interpolated and plotted as a performance degradation curve. The performance degradation curve is calculated for all the DSN mission pairs (SC1/SC2, SC1/SC3, ..., SC1/SCN), and Fig. 10 illustrates an example of the performance degradation curves computed for the Cassini channel assignment. The performance degradation is shown between Cassini (SC1) and some of the existing DSN missions, Galileo, Ulysses, Mars Observer, and the Voyagers.

When all of the degradation curves are assembled into one plot, the result is a frequency optimization curve which indicates the overall trend in degradation to the DSN system across the channel frequency spectrum. The optimization curve is characterized by the outermost peaks, which indicate regions of high levels of interference, and the null values, defined at the intersection of the curves, which indicate regions of relatively low interference. For example, Fig. 10 illustrates the frequency optimization curve for the Cassini channel assignment. The Y-axis represents the daily mutual interference potential, given in "watts·days," whereas the X-axis represents the DSN channel number.

In this example, channel numbers with the lower value on the curve can be considered as viable candidate channels for the Cassini mission. Channels 14, 15, 18, and 19 show corresponding peak values and should be avoided. The candidate channels can then be further assessed for optimization by considering other spectral components, such as ranging and differential-one-way-ranging (DOR) tones. The impact of discrete tones is analyzed using the discrete tone interference algorithm, discussed in the next section. This will ensure that the use of this optimal channel will not cause interference with the introduction of other discrete modulation components.

F. Discrete Tone Interference Analysis and Algorithm

Discrete signal components are known to be other sources of interference that can impact the choice of channel assignment. An analysis of their impact on the interfered-with signal is performed using procedures different from those used for continuous signal component analysis. Among the discrete components of a DSN signal are the special continuous wave (CW) tone signals and the modulation products of the ranging and DOR tones, whose emission spectrum is characterized as a set of harmonically related tones. The frequencies and power levels of these sinusoidal components are easily computed to provide further evaluation of the spectral isolation between channels.

Interference caused by the discrete tones is evaluated in terms of the power ratio of the interfering tone and the interfered-with signal component, and the frequency offset (including Doppler shift) between them. Predefined limits for these two interference criteria will give an indication of whether interference exists.

Figure 11 illustrates the flow diagram for analyzing discrete component interference for DSN uplinks and downlinks. The discrete tone algorithm (DTA) computes the complete set of interference tone frequencies and power levels and tabulates those signal components which lie near the interference tones. Given this information, the analyst can then determine if an intolerable interference situation

exists and whether another channel frequency ought to be examined.

IV. Interference Criteria

A prerequisite for conducting an effective interference evaluation is the establishment of suitable interference criteria. The interference criteria represent thresholds in link parameters and, when exceeded, give an indication of intolerable performance degradation. It can be specified as a numerical value of interference power level, ISR, bit error rate, or another suitable parameter which gives a meaningful indication of the expected degradation in link performance.

The interference criteria used in the analysis are defined as spacecraft-specific interference power limits for the uplink case, and the standard CCIR spectral power density limits for the downlink case. Table 1 shows a sample of the interference criteria used for the DSN tracking station receivers and spacecraft transponder receivers operating in the 2-GHz (S-) and 8-GHz (X-) bands.

The tracking station receiver interference criteria were established through the CCIR to protect the DSN from harmful interference. If the interference signal is assumed to be broadband, its power spectral density received at the threshold level shown in Table 1 will cause an increase in carrier tracking loop noise and, consequently, a degradation of about 1 dB in the output signal-to-noise ratio (SNR). Limitations on the duration of the interference event, as percentages of time, are also specified in conjunction with the interference criteria threshold levels.

The downlink power spectral density limits listed above can be converted to absolute interference power threshold levels with the multiplication of a suitable bandwidth factor. For the Block III and Block IV receivers, a multiplication factor of 10 Hz is used, corresponding to the receiver's most probable tracking loop bandwidth.

Interference criteria threshold levels for the spacecraft are given as absolute power levels and are specified from hardware and link performance considerations. The spacecraft receiver is generally less sensitive to interference than the tracking station receiver. Furthermore, a typical uplink signal occupies far less bandwidth than the downlink signal.

Reference

[1] *Radio Regulations, Appendices 25–44*, Oslo, Norway: Norwegian Telecommunications Administration, pp. AP29-12 – AP29-15, 1982.

Table 1. Interference criteria for DSN tracking station receivers.

Band, GHz	Maximum allowable interference spectral power density, dBW/Hz	Maximum allowable interference spectral power flux density,[a] $dBW/m^2 \cdot Hz$
2.3	−222.5	−255.5
8.4	−220.9	−253.2
13.0	−220.5	−251.7
32.0	−217.3	−239.1

[a] For 70-meter antenna.

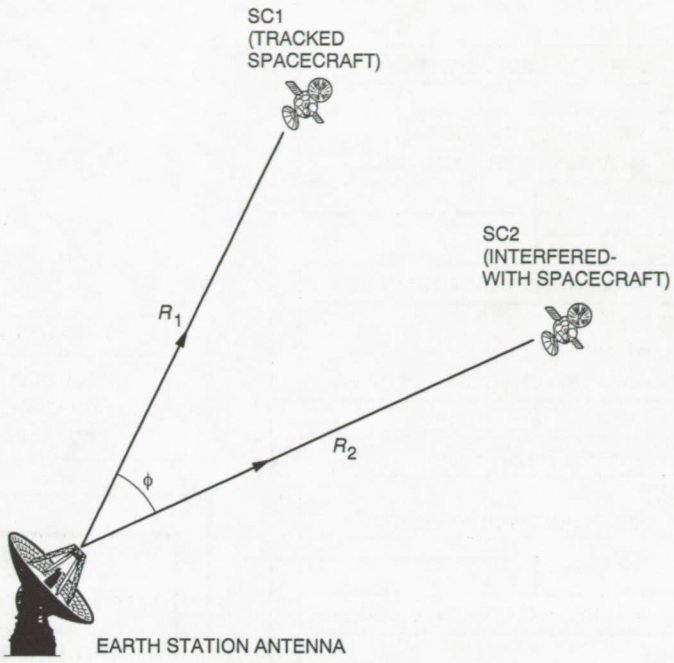


Fig. 1. Interference geometry for uplink.

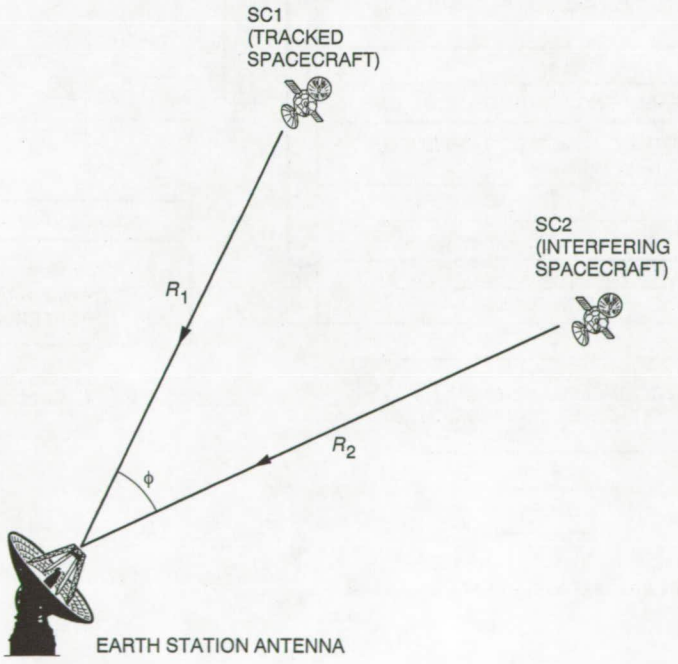


Fig. 2. Interference geometry for downlink.

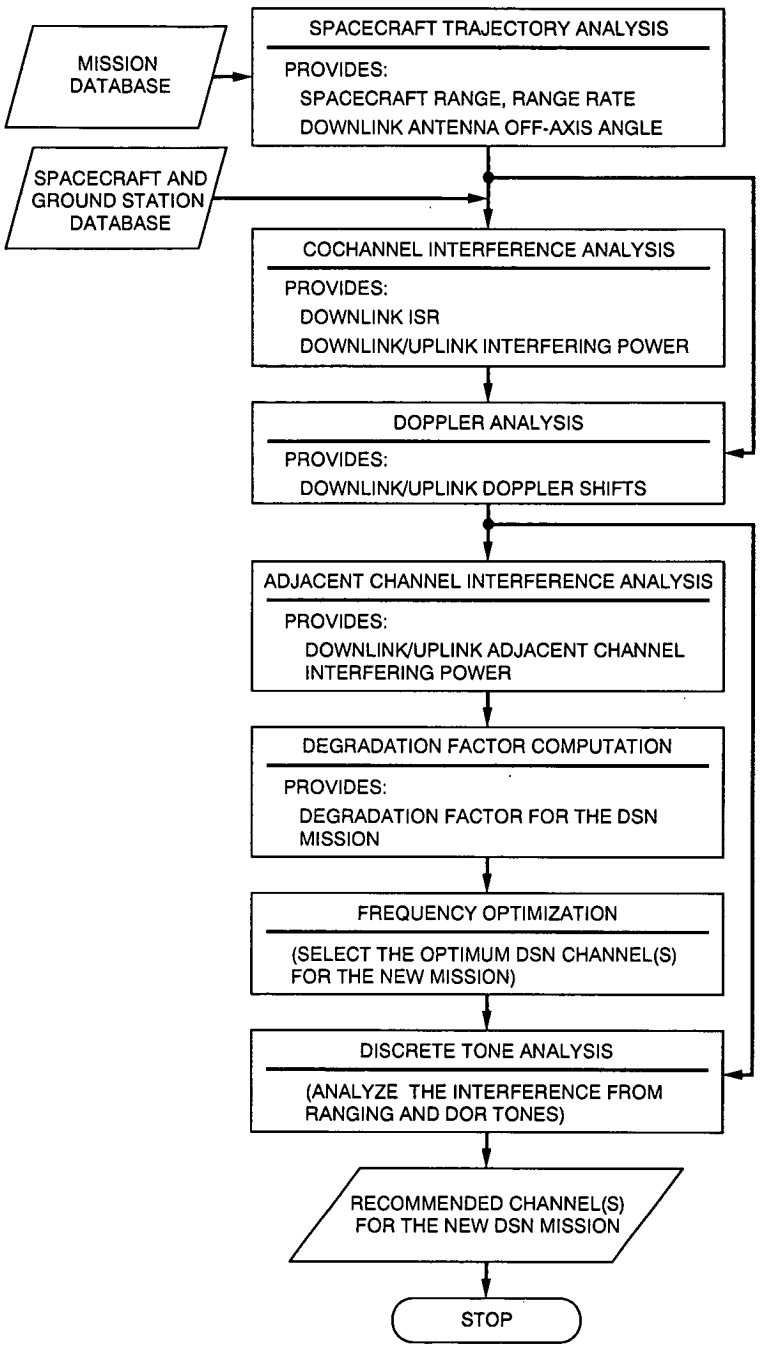


Fig. 3. Block diagram of DSN channel assignment using IAP.

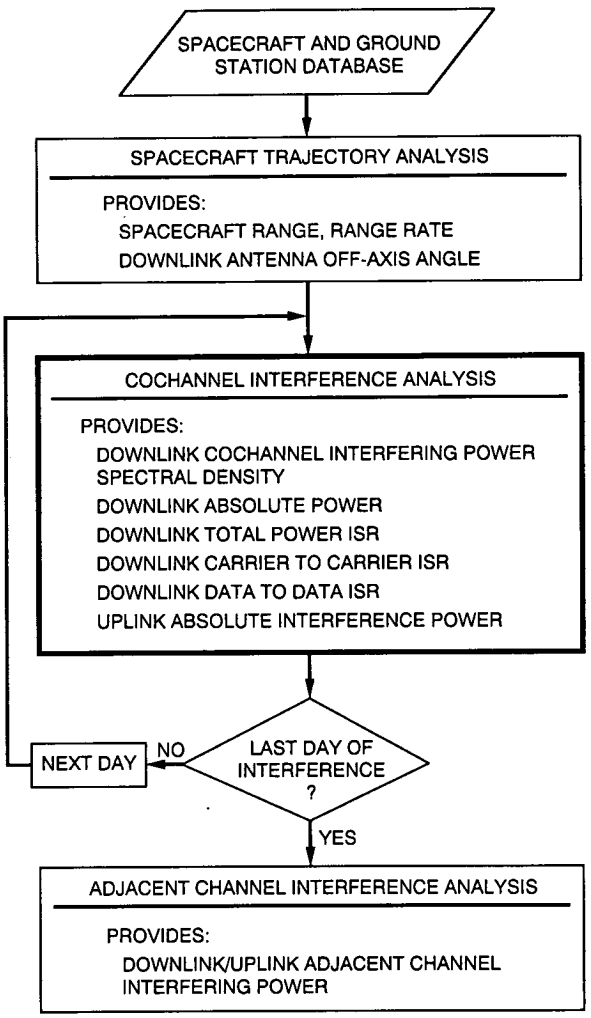


Fig. 4. Cochannel interference analysis.

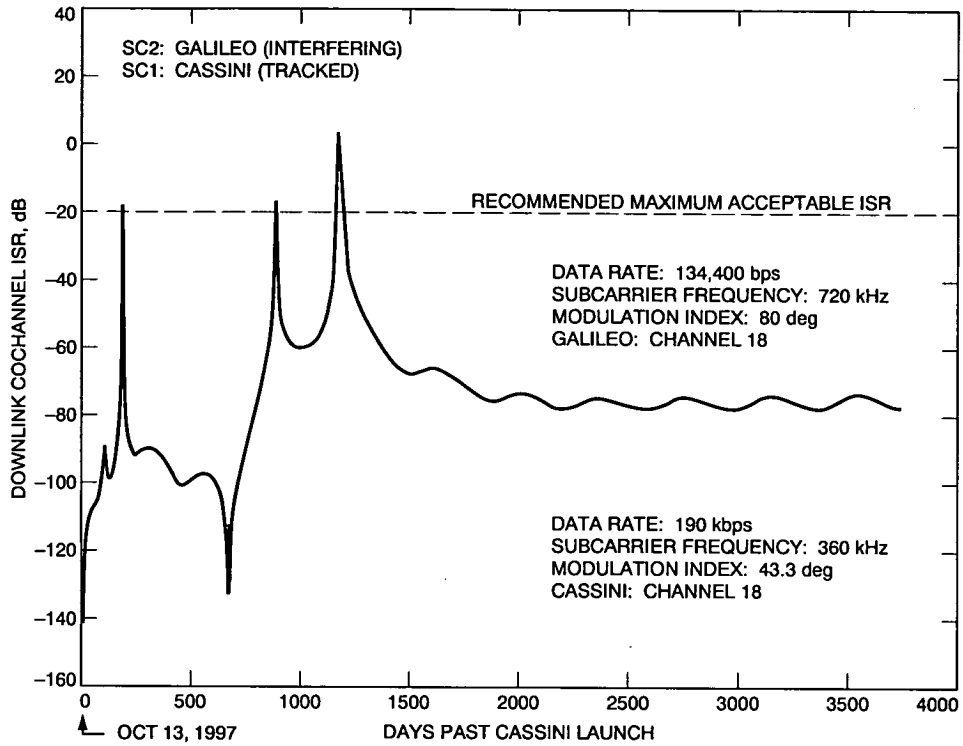


Fig. 5. Downlink total power cochannel ISR from Galileo to Cassini.

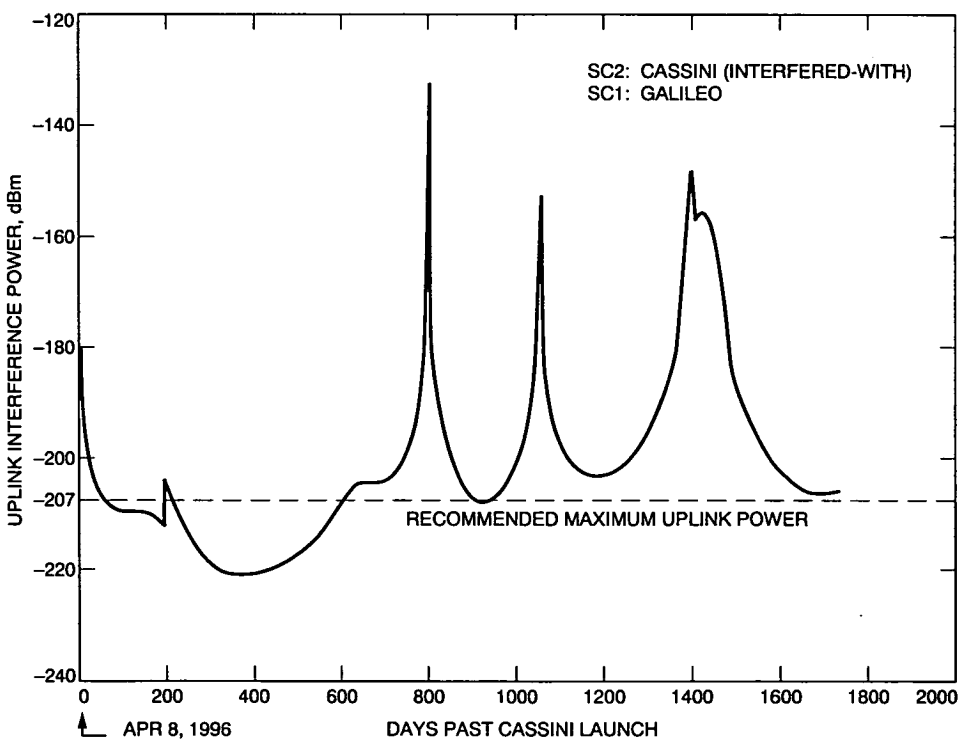


Fig. 6. Uplink cochannel interference power from Galileo to Cassini.

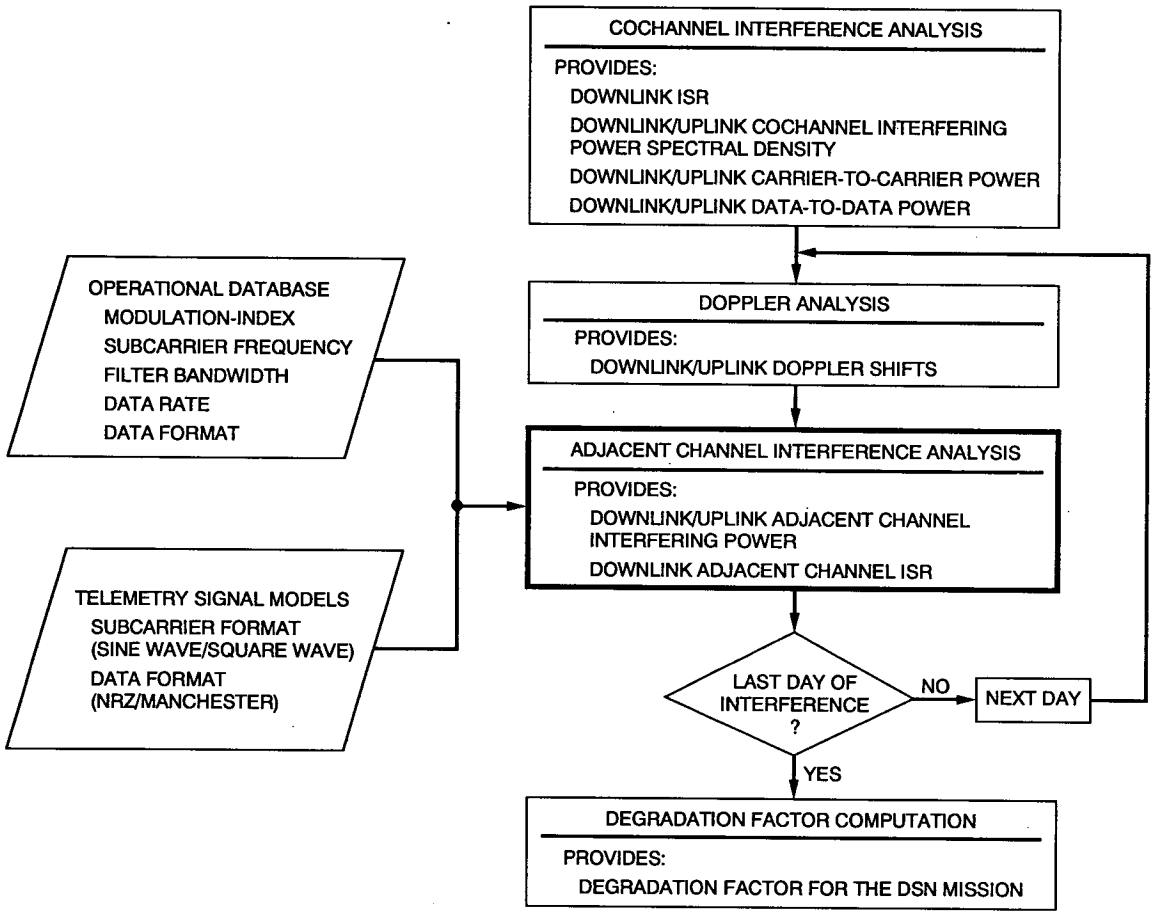


Fig. 7. Adjacent channel interference analysis.

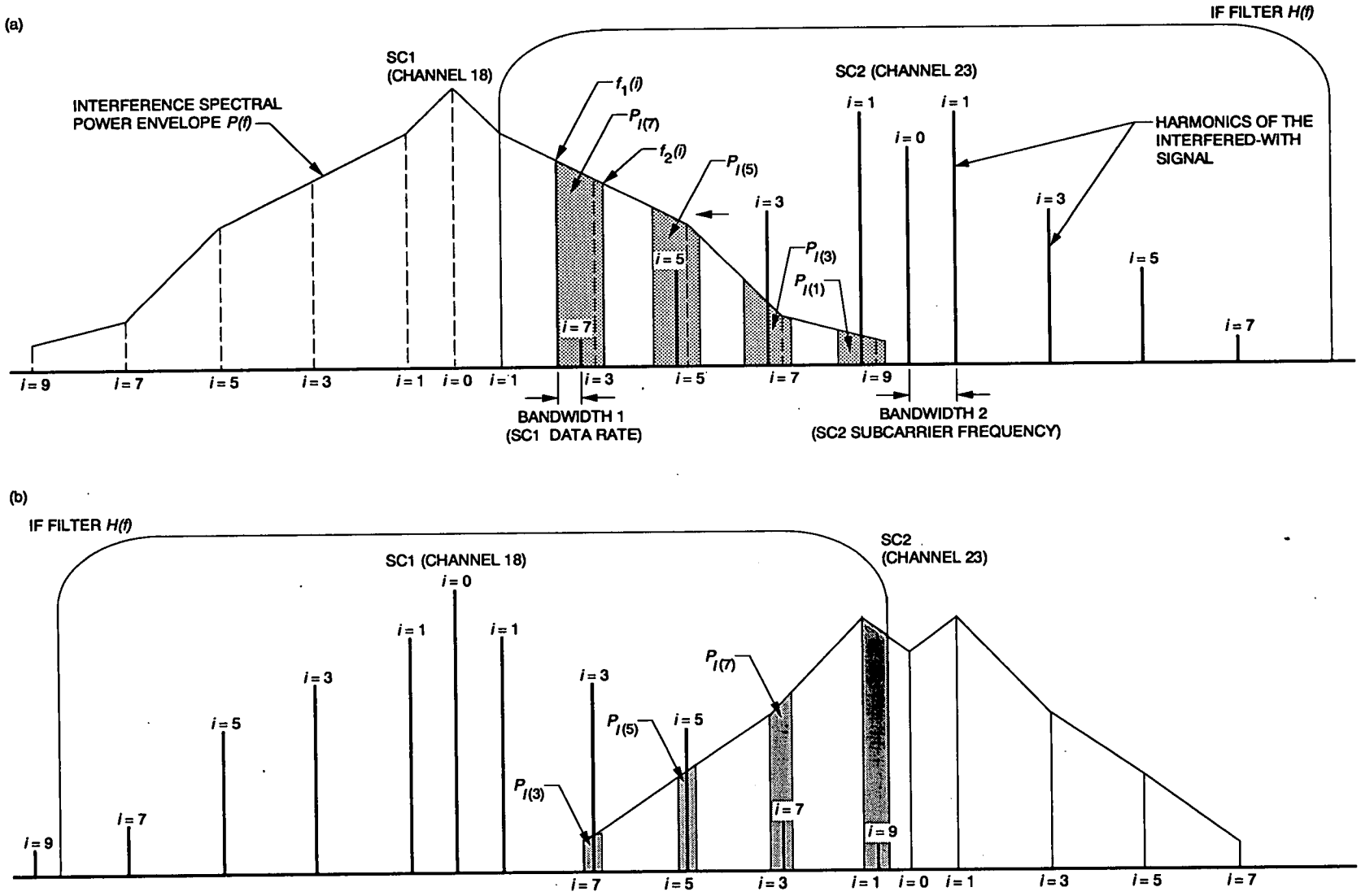


Fig. 8. Interference: (a) from SC1 to SC2 and (b) from SC2 to SC1.

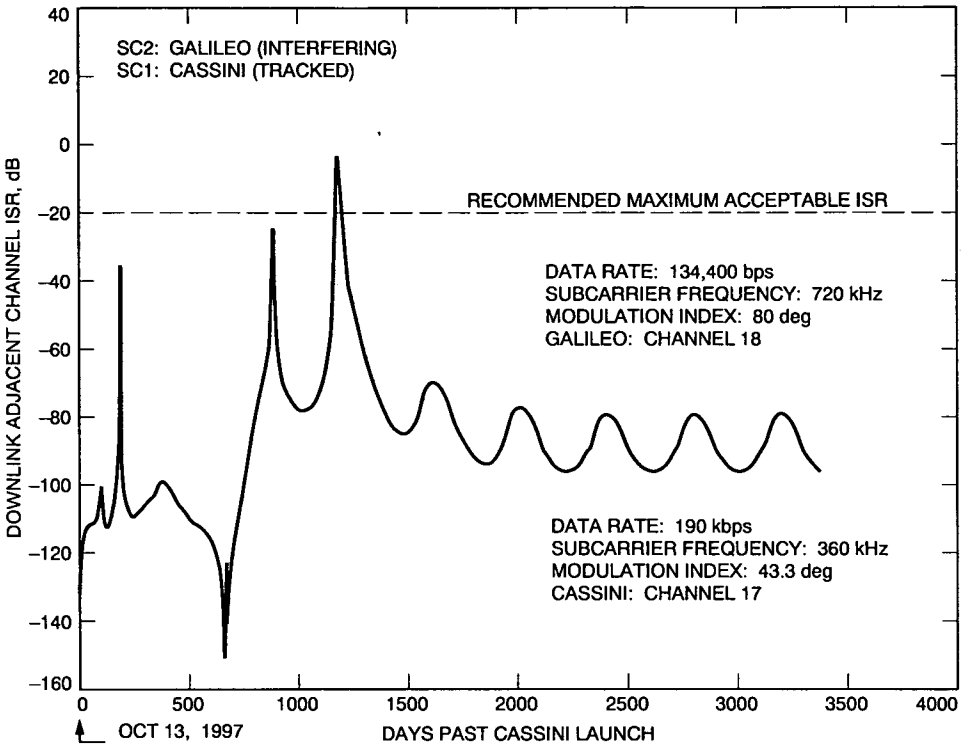


Fig. 9. Downlink adjacent channel ISR for Cassini versus Galileo in the 8-GHz band.

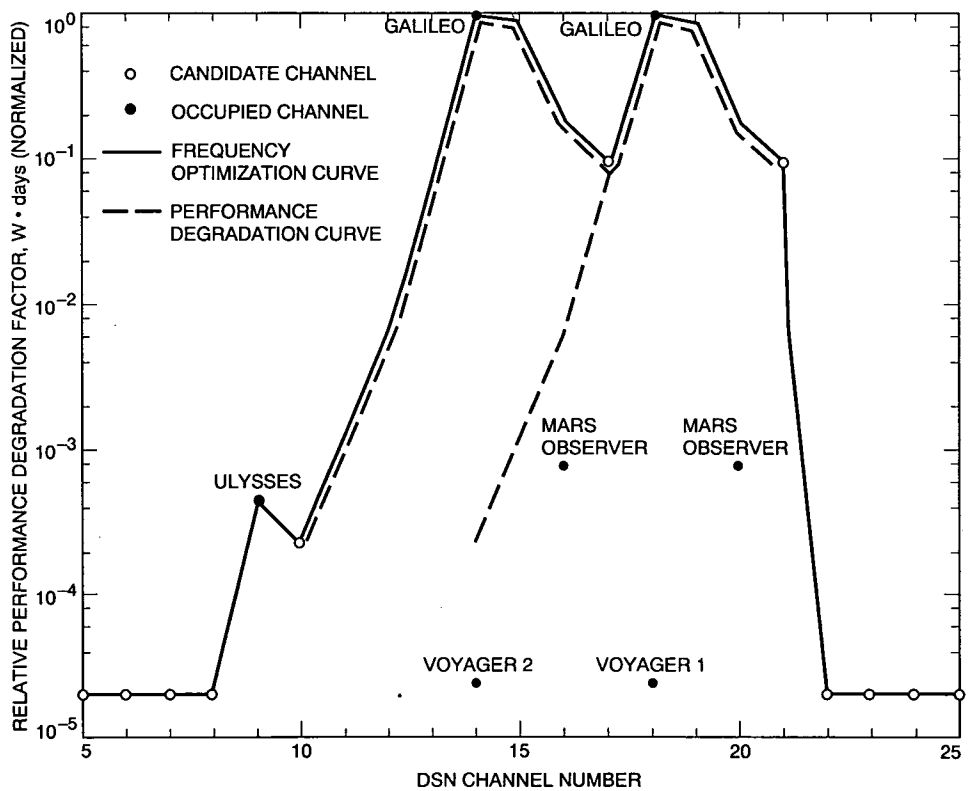


Fig. 10. Frequency optimization curve.

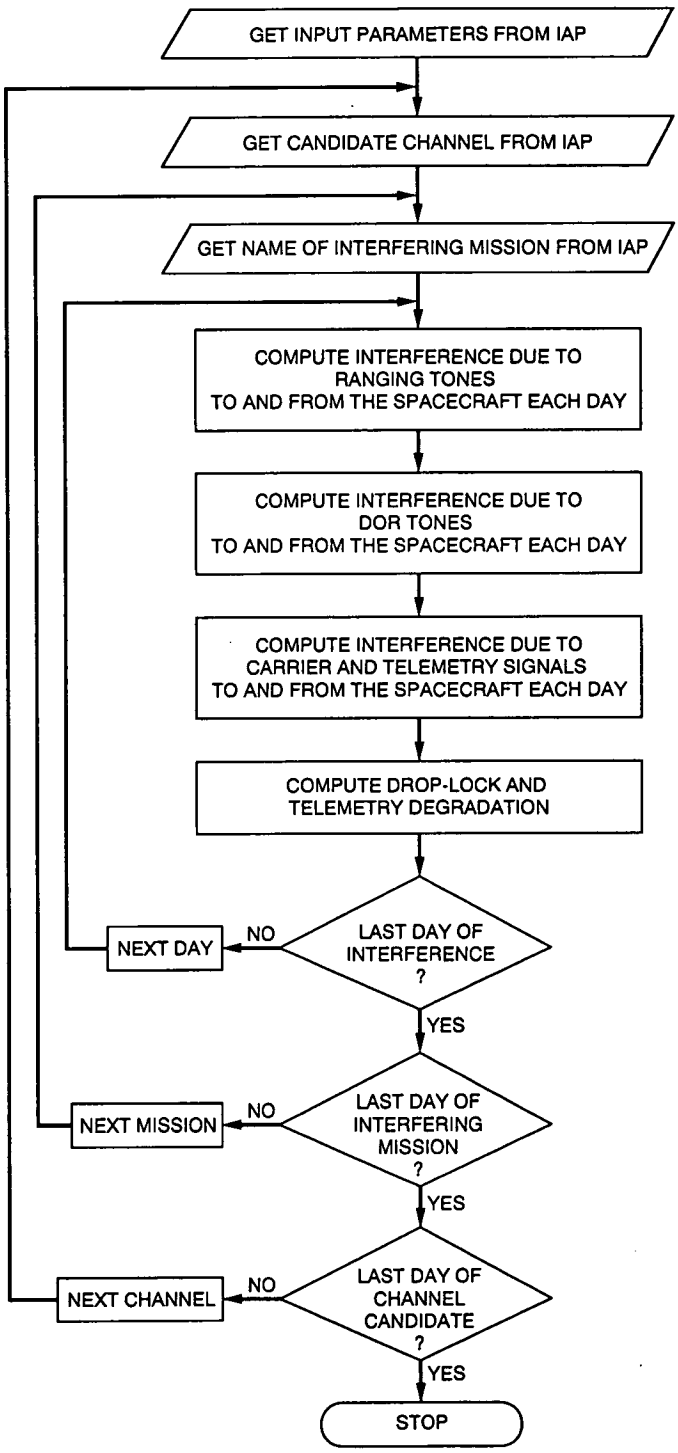


Fig. 11. Block diagram for the discrete tone analysis.

A 32-GHz Phased Array Transmit Feed for Spacecraft Telecommunications

K. A. Lee, D. L. Rascoe, R. A. Crist, J. Huang, P. D. Wamhof, and F. S. Lansing
Spacecraft Telecommunications Equipment Section

A 21-element phased array transmit feed has been demonstrated as part of an effort to develop and evaluate state-of-the-art transmitter and receiver components at 32 and 34 GHz for future deep-space missions. Antenna pattern measurements demonstrating electronic beam steering of the two-dimensional array are reported and compared with predictions based on measured performance of MMIC-based phase shifter and amplifier modules and Vivaldi slotline radiating elements.

I. Introduction

The DSN is implementing telecommunication capability at 32 GHz (Ka-band) with expected improvement of as much as 8 dB over the current capability at 8 GHz (X-band). In preparation for this new capability on the ground, a proof-of-concept 32-GHz phased array transmitter to feed a near-field dual-reflector system has been designed and implemented for future space application. The work described herein builds on the earlier design of a smaller active linear array [1] and a passive two-dimensional array [2,3].

The objective was to demonstrate useful RF power (1 to 10 W) and end-to-end efficiency (15 to 30 percent) at 32 GHz through array combining of low-power devices. One benefit of array combining of solid-state devices is improved reliability through graceful degradation of the RF output if individual devices fail. In addition to solid-state amplification, vernier electronic beam-steering capability (±10 deg) was also a goal with potential relief to pointing required of the spacecraft attitude control system.

The array development to be described has provided practical hands-on experience in the incorporation of monolithic microwave integrated circuits (MMIC) into arrays, mechanical and electronic system integration, and electronic beam-steering calibration and test. The completed hardware from this effort is currently used as a testbed for evaluating more efficient active devices, antenna elements, and signal distribution circuits as they are developed. Future microspacecraft missions to asteroids and the outer planets are the missions most likely to benefit from the Ka-band technology being demonstrated through this work.

II. Subsystem Design

The transmit phased array was designed to feed a spacecraft's dual-reflector antenna typically from the Cassegrain feed point. In this configuration, limited scan capability is traded for higher system gain. The feed, capable of ±10-deg scanning, would provide ±1-deg scanning in a system with a magnification factor of 10 [3]. The pointing resolu-

tion of the feed is also reduced by a factor of 10, suitable for satisfying pointing requirements on the order of hundredths of a degree.

The array consists of 21 active antenna elements fed by a beam-forming network. A modular architecture was selected to facilitate testing and substitution of individual components. Five linear subarray layers comprise the full two-dimensional array. The middle three layers contain five elements each and the outer two layers contain only three. Each subarray slides horizontally into an aluminum structure, providing support and alignment, and then plugs into a vertical RF signal distribution strip at the back of the structure (Fig. 1). Each subarray layer is divided into four modules: an input power divider, a phase shifter module, an amplifier module, and an antenna module. A photograph of a five-element subarray layer is shown in Fig. 2. Each path of a module is tested before integration into the subarray layer, and each subarray layer can be tested before integration into the full array.

III. Array Performance

The array was measured on a 7-m indoor antenna range capable of operation over the 26.5- to 40-GHz band. Amplitude and phase of the received signal were measured in a calibration sequence that steps through the 4-bit phase shifters' 16 states for each of the 21 elements. Figure 3 shows a typical calibration measurement giving the measured phase and amplitude in the far field. These data are contained in the lookup table used to set the beam-steered position of the array. In order to improve pattern measurements, paths with higher gain were attenuated, which provided a more uniform array aperture. The amplitudes for 17 elements vary within ± 3 dB, 3 of the elements have amplitudes ~ 5 dB lower, and 2 of the elements were not operating due to device failures during preliminary testing on the antenna range.

A. Antenna Patterns

Three-dimensional antenna patterns for the array are measured by sweeping in 1-deg increments in azimuth and then rotating in 10-deg increments, as indicated in Fig. 4(a). The result is that each azimuth cut measures a 10-deg sector of the azimuth–elevation plane in the contour plots that appear in Fig. 4(b) and (c). The half-power beamwidth is 9 deg, and the peak side-lobe level measured -13 dB pointing boresight. When steered off boresight by 10 deg, the scan loss is -1.5 dB and the peak side-lobe level is -10 dB.

The measured and modeled E- and H-plane patterns for a boresight beam and the principal cuts for beams scanned in the E- and H-planes are shown in Fig. 5. The 3-dB beamwidth is 9 deg for the boresight and scanned beams. The peak side-lobe level is -13 dB for the boresight and -10 dB for the scanned beams. The measured patterns are compared with a model that assumes an element pattern with a $(\cos\theta)^{10}$ variation for the main beam shape and a constant -15.7-dB side-lobe level. In addition, the modeled patterns include the actual phase and amplitude measured for each element. Differences between the measured and modeled patterns are a result of using a simplified element pattern and not including mutual coupling between elements in the model.

B. Gain and Efficiency

The gain of the 21-element array was predicted to be 26.7 dB. This was scaled from the measured gain of 22 dB for the best five-element linear subarray. The actual gain of the 21-element array was measured to be 25.6 dB where 2 of the elements had failed, 6 of the elements were 4 to 7 dB lower in gain than expected, and 10 elements were 1 dB higher than expected because they were not operating in compression.

The power-added efficiency (PAE) for the array, defined as

$$PAE = \frac{RF_{\text{power out}} - RF_{\text{power in}}}{DC_{\text{power in}}}$$

is predicted to be 8.7 percent based on measured RF output power for the best five-element tray. Efficiency will increase to 16 percent with a reduction in extraneous circuit losses. In future arrays, 30-percent array efficiency is feasible with an increase in device efficiency to 35 percent.

C. EIRP

A common figure of merit for a transmitter array is the effective isotropic radiated power (EIRP). The measured EIRP for the array, based on results for the best five-element subarray, is 56 dBm with 1 W of input power. For the 21-element phased array, the power into the coaxial connector feeding the power divider is the transmit power and the gain of the array includes all the losses from that point. Table 1 summarizes the gain and loss of the array components and predicts an EIRP of 57 dBm.

IV. Array Modules

A. Antenna Module

The Vivaldi slotline antenna element design and passive array analysis was performed under contract with the Uni-

versity of Massachusetts [3,4]. The slot antenna is fed with a microstrip-to-slotline transition from the back side of the substrate for easy integration in a planar microstrip configuration with MMIC amplifiers and phase shifters. The two sides of the antenna substrate, containing five active and two parasitic elements, are shown in Fig. 6.

The optimum element spacing takes into consideration the reflector system, packing density, and grating lobes. For an isotropic element beam, as the spacing increases to greater than 1.0λ, the directivity decreases due to losses in the grating lobes; however, by using elements with higher gain, such as Vivaldi slots (12-dB gain), this loss is reduced while allowing more space between elements to accommodate the MMICs. The University of Massachusetts and JPL performed a study and concluded that an element spacing of 1.22λ could be tolerated while maintaining high (>95 percent) radiative efficiency. This element spacing was selected for the 21-element array.

B. Amplifier Module

Two types of amplifer MMICs were integrated into the array. The devices used are pseudomorphic high-electron-mobility transistor (PHEMT) MMICs designed by Texas Instruments' (TI) Central Research Laboratory [5]. The first device is a three-stage amplifier with transistors of gate widths 50 μm, 100 μm, and 250 μm in cascade. Gate and drain bias circuitry are fabricated on the chip. The best device of this design is capable of 30-percent power added efficiency, 190-mW output power, and 23-dB gain at 31 GHz. Typical devices produce 20-percent efficiency, 100-mW output power, and 20-dB gain at 32 GHz. A photograph of the chip appears in Fig. 7(a).

Due to lack of availability of the first MMIC, a second three-stage device [Fig. 7(b)] was obtained from TI and integrated into the two three-channel modules. This device had lower gain (12 dB) and required a different bias voltage than the others, and therefore produced only 25 mW of RF input at 32 GHz under the fixed RF input and dc bias conditions of the array. However, under proper operating conditions, TI reports that this device is capable of >30-percent PAE, 250-mW output power, and 18-dB gain at 31 GHz.

Five amplifier MMICs were integrated on a single carrier module. All devices share common drain and gate supplies so that only two dc lines are required to interface to the module. In this approach, not all the devices are optimized at the same bias point, so RF power and efficiency are traded for array-biasing simplicity. The printed circuitry for a module is combined on one substrate, which simplifies assembly since all the channels are automatically aligned; however, the substrate becomes more complicated and reworking individual channels is less convenient than if each amplifier were first integrated into a submodule.

The original design contained a cascade of two amplifiers: the 100-mW driver amplifier followed by a power amplifier capable of 5-dB gain and 350-mW output power. The 350-mW device consisted of a single stage with 800-μm gate periphery and no on-chip bias or decoupling circuitry. The two devices were integrated on the module and tuned in a coaxial test fixture. Upon integrating the cascade of amplifiers with the antennas and phase shifters, destructive oscillations occurred and the circuit design was reevaluated. Modifications to the bias circuit were made in an attempt to prevent oscillation; however, schedule and resource constraints did not permit the extent of rework that was required. The approach taken was to use only the 100-mW amplifier, which has proved to be stable by itself in the module. A 50-Ω transmission line fabricated on quartz was placed in the second substrate cutout where a suitable high-power device could be inserted in the future. A photograph of the modified amplifier module is shown in Fig. 8.

The amplifier modules were measured on a scalar network analyzer setup. The best amplifier module produced 560 mW with 17-percent power-added efficiency at 32 GHz. Frequency response and efficiency measurements for an amplifier module are shown in Fig. 9. Crosstalk between channels was observed to be as high as −20 dB at 32 GHz. A channelized cover machined from RF-absorbing material was placed over the module, which reduced the crosstalk by >10 dB.

Molybdenum was the selected carrier material because of its relatively high thermal conductivity as compared to other materials with a temperature coefficient of expansion matching the alumina substrate. A trade study was performed to analyze the thermal performance of various die-attach methods with the result that gallium arsenide (GaAs) chips were gold-eutectic soldered to a copper subcarrier and then silver epoxied to the molybdenum carrier. A thermal model of the 21-element array was developed, and the predicted temperature rise from the carrier to the chip was calculated to be 20 deg C, with a channel temperature of 61 deg C. Actual temperatures were measured using thermocouples placed on the array. Predicted and measured temperatures agree to within 5 deg C.

C. Phase Shifter Module

The phase shifter module contains the 4-bit switched-line phase shift MMICs designed by Honeywell under con-

tract with NASA Lewis Research Center. Associated with each phase shift MMIC is an addressable digital serial-to-parallel converter application-specific integrated circuit (ASIC) designed at JPL. Use of this ASIC reduces the number of signal lines to each module from 35 to 6. A photograph of a single path appears in Fig. 10.

The power dividers and phase shifters were integrated together and measured on a vector network analyzer. Typical insertion loss was around 20.5 dB, measured without the vertically oriented power divider, which provides an additional 9-dB loss. These measurements are consistent with previous measurements of the separate components, as recorded in Table 1. Phase errors were less than 10 deg. Plots of insertion loss and phase at 32 GHz as a function of phase state are shown in Figs. 11(a) and (b). The load impedance presented to the amplifiers was also measured, and plots of the output return loss $|S_{22}|$ for different phase states are shown in Fig. 11(c). Lower loss phase shifters are becoming available from industry and will be implemented in future arrays.

V. Summary

A phased array transmitter operating at 32 GHz with 2-W total RF power and 9-percent total efficiency, with a capability of 16-percent efficiency, has been demonstrated using existing devices. This is based on the performance of the best five-element subarray. The method of spatially combining low-power devices to increase transmit power has proved successful. Beam steering in a 10-deg cone by the phase shifters was also demonstrated, with a scan loss of −1.5 dB and a peak side-lobe level of −10 dB.

Fabrication of the hardware provided valuable practical experience in the area of MMIC insertion. System-level issues for arrays were also addressed, including accommodation for bias and control for many individual elements and reduction of crosstalk within the array. By integrating the full array, procedures and control software for operating and evaluating the array were developed. The existing array will be used as a testbed for more efficient antennas, power dividers, and devices.

Several areas for future work have been identified. MMIC technology is the key to realizing economical large arrays, and a fast and universal method for evaluating state-of-the-art devices is being developed. Advanced MMIC power amplifiers are under development with improved RF output and efficiency. Packaging of individual devices and modules at 32 GHz is the next major step from a testbed to a breadboard phased array. Device packaging will address hermeticity, crosstalk, thermal management, and radiation as well as facilitate device handling. Advances in these areas will enable smaller, more efficient array designs needed for future microspacecraft missions.

Acknowledgments

The authors gratefully acknowledge Dr. G. Birur for the thermal analysis, C. Jones and C. Cruzan for assembly assistance, and R. Thomas and R. Beckon for assistance in measuring antenna patterns.

References

[1] D. L. Rascoe, A. L. Riley, J. Huang, V. Lubecke, and L. Duffy, "Ka-band MMIC Beam Steered Transmitter Array," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-37, pp. 2165–2168, December 1989; also, *TDA Progress Report 42-98*, vol. April–June 1989, pp. 207–217, Jet Propulsion Laboratory, Pasadena, California, August 15, 1989.

[2] A. L. Riley, D. L. Rascoe, T. Cooley, L. Duffy, V. Jamnejad, R. Thomas, and K. S. Yngvesson, "A Ka-band MMIC Array Feed Transmitter for Space Application," *IEEE 1991 Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 11–14, June 1991.

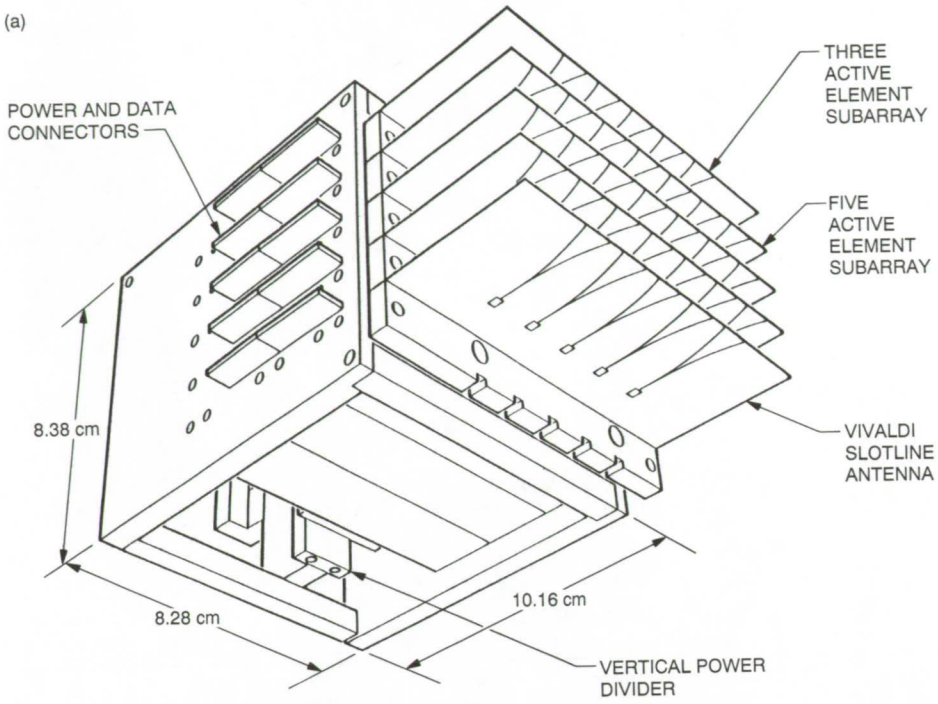
[3] V. Jamnejad, "Ka-band Feed Arrays for Spacecraft Reflector Antennas With Limited Scan Capability—An Overview," *TDA Progress Report 42-102*, vol. April–June 1990, pp. 367–376, Jet Propulsion Laboratory, Pasadena, California, May 15, 1992.

[4] K. S. Yngvesson, Y.-S. Kim, T. L. Korzeniowski, E. L. Kollberg, and J. F. Johansson, "The Tapered Slot Antenna—A New Integrated Element for Millimeter Wave Applications," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-S 37, pp. 365–374, February 1989.

[5] P. Saunier, H. Q. Tserng, and Y. C. Kao, "A High Efficiency Ka-band Monolithic Pseudomorphic HEMT Amplifier," *Proceedings of the Monolithic Microwave Integrated Circuits for Sensors, Radar, and Communications Systems Conference*, Orlando, Florida, pp. 86–90, April 1991.

Table 1. Gain summary.

Parameter		Value
Input power, dBm		31.5
Power division (÷21), dB	−13.2	
Power divider insertion loss, dB	−4.3	
Phase shifter insertion loss, dB	−12.2	
Amplifier gain, dB	18.0	
Power combination (×21), dB	13.2	
Antenna gain, dB	24	
Total, dB	25.5	
EIRP, dBm		57.0
Total radiated power, W		2.0



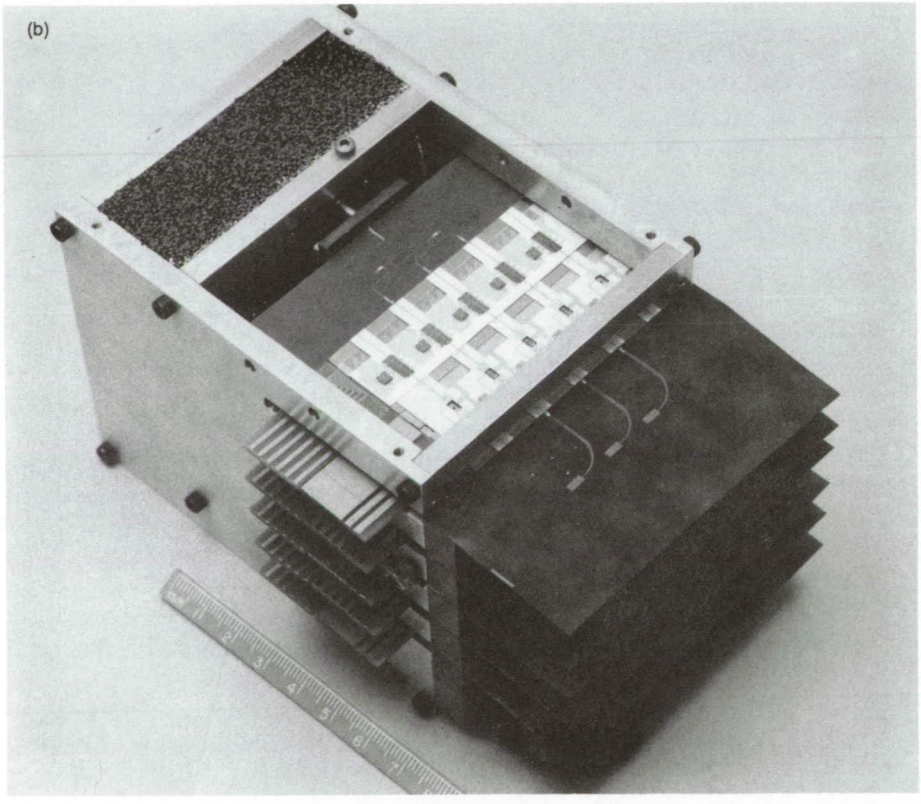


Fig. 1. The 32-GHz phased array transmitter feed: (a) subarray modules integrated into a two-dimensional array and (b) the completed testbed array. The use of highly integrated solid-state devices reduces the overall volume and simplifies array construction. MMICs provide phase shifting and amplification at 32 GHz. Custom designed digital ASICs provide phase shift control and reduce the number of signal lines required for the array.

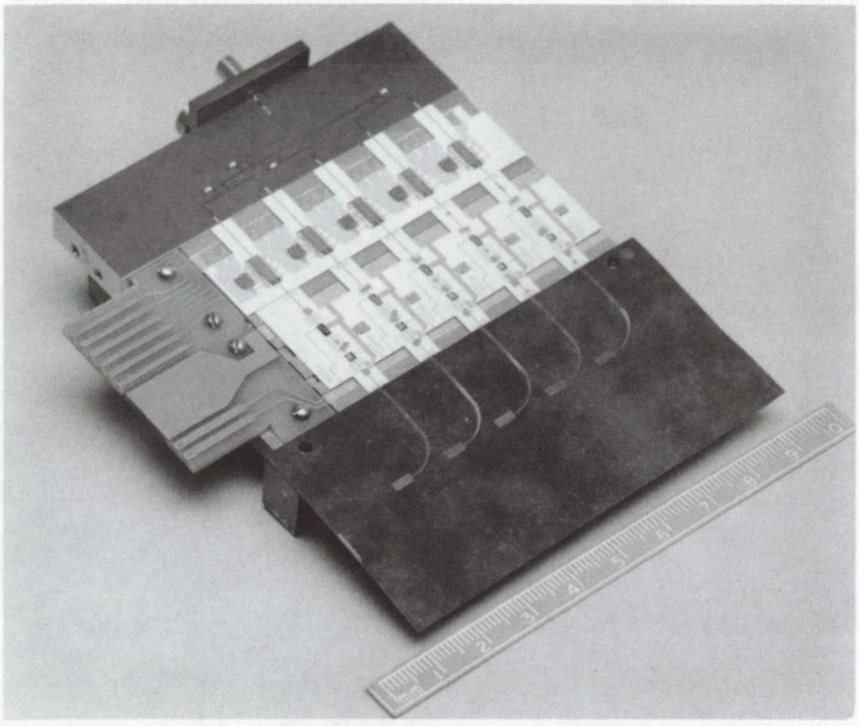


Fig. 2. Power divider, phase shifter, amplifier, and antenna modules integrated into a five-element subarray tray.

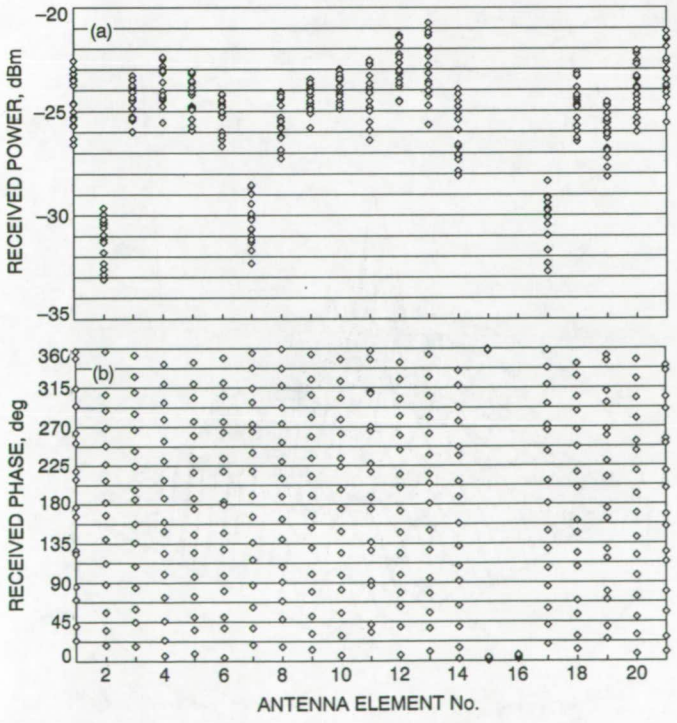


Fig. 3. Received amplitude (a) and phase (b) for each of 21 elements in each of 16 phase shifter states measured in the far field of the array at 32 GHz.

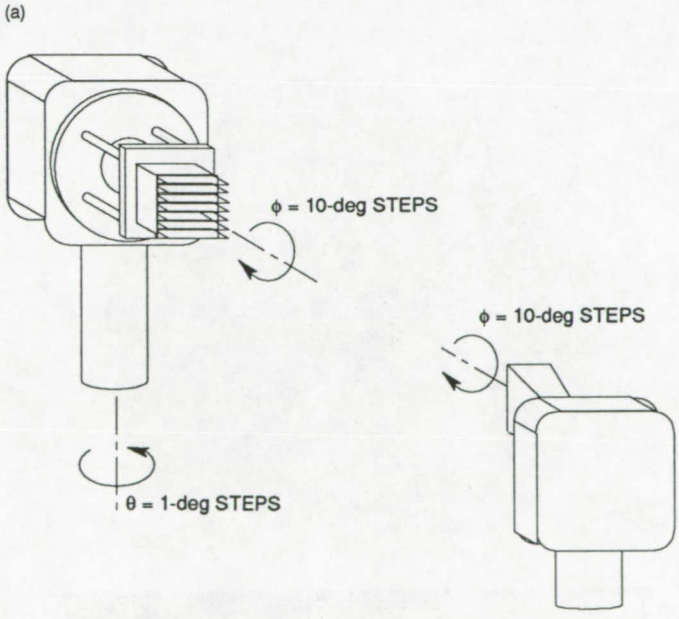
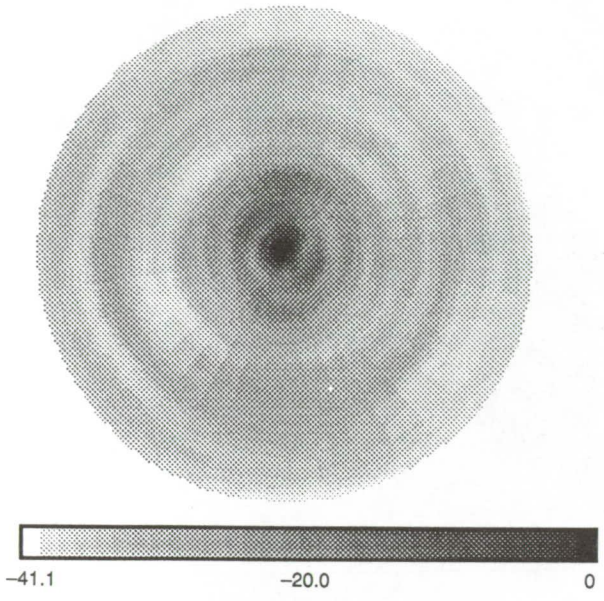
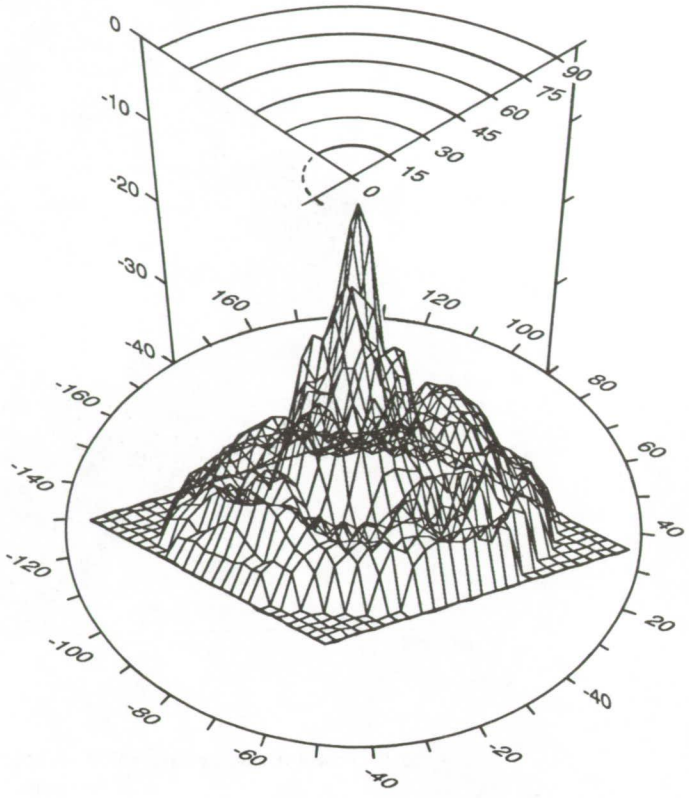


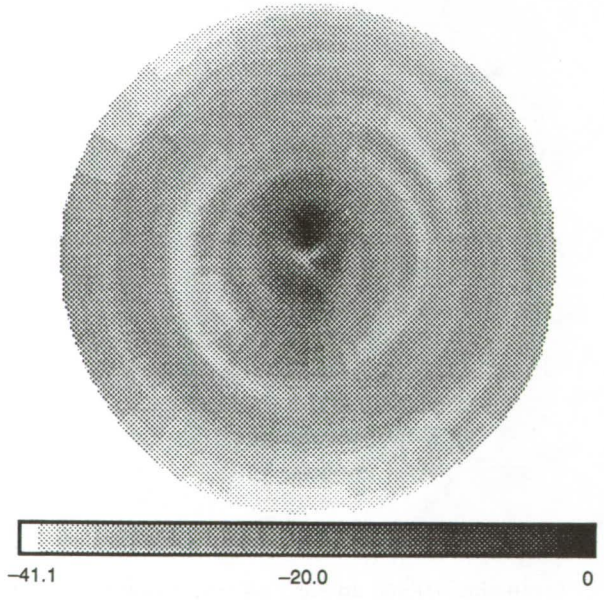
Fig. 4. Three-dimensional antenna pattern generation for the array: (a) pattern ϕ cuts generated by sweeping in azimuth and three-dimensional patterns formed from cuts with ϕ = 0 to 180 deg, with receive horn rotated to maintain copolarization; (b) measured far-field antenna patterns of the array feed pointed boresight; and (c) scanned 10 deg. Half-power beamwidths are 9 deg, and peak side lobe levels are –13 dB and –10 dB for the boresight and scanned beams, respectively.

(b)





(c)



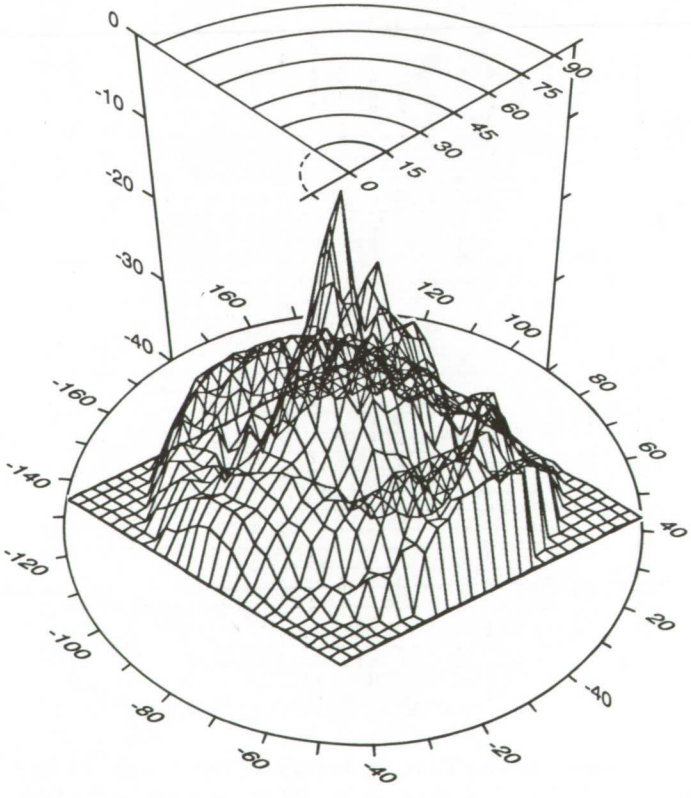


Fig. 4 (contd).

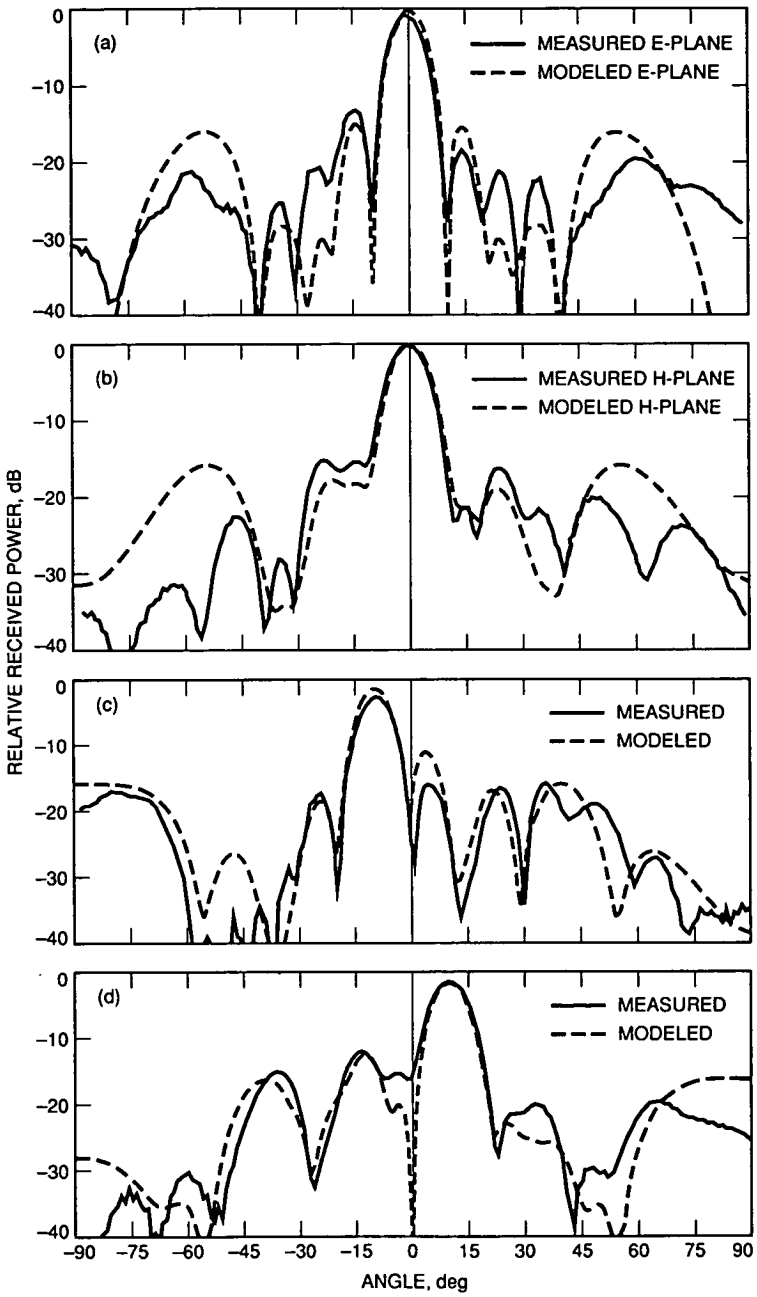


Fig. 5. Comparison of measured and modeled patterns of the phased array: (a) E-plane at boresight; (b) H-plane at boresight; (c) scanned 10 deg in the E-plane; and (d) scanned 10 deg in the H-plane. (The model assumes an element main beam shape of $(\cos \theta)^{10}$ with a constant side lobe level of −15.7 dB.)

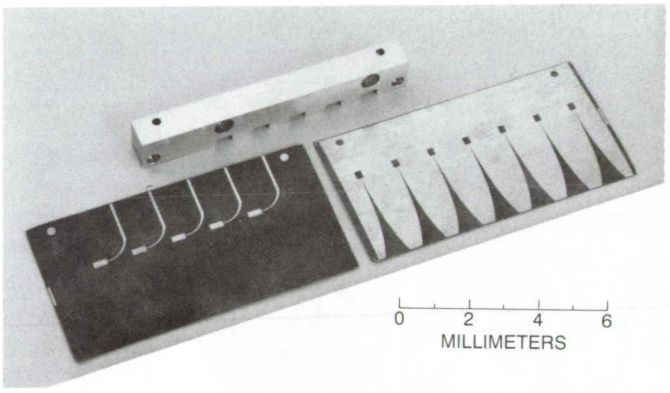


Fig. 6. The Vivaldi antenna substrate. Five active and two parasitic printed antennas are etched on one side (right) and microstrip-slotline transitions are etched on the opposite side (left).

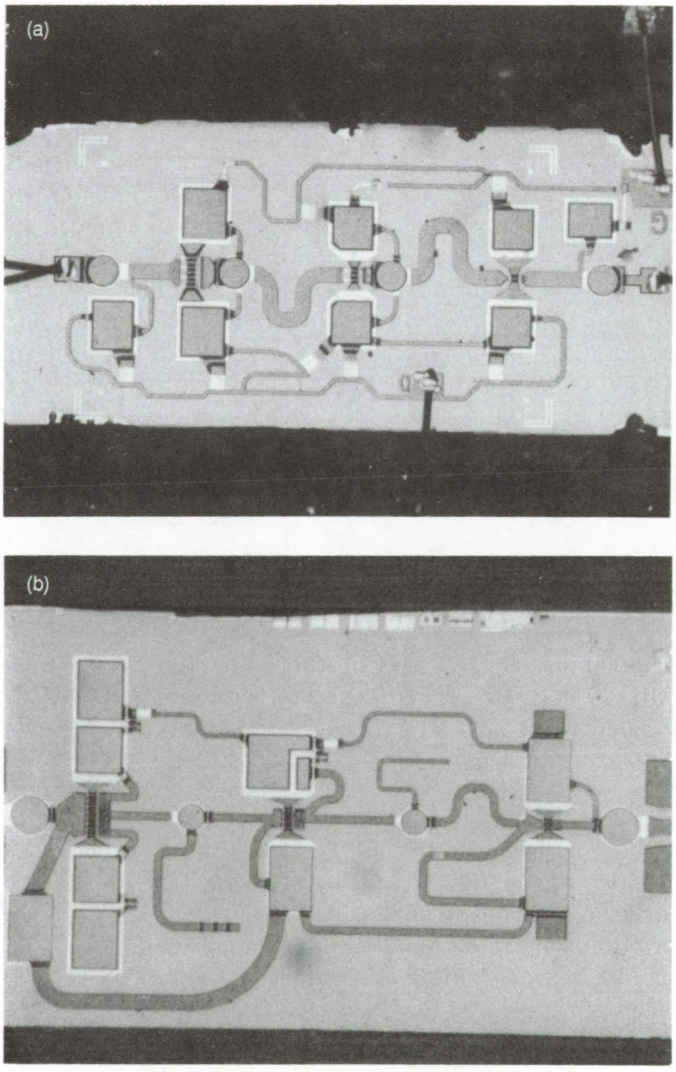


Fig. 7. Amplifier MMICs: (a) the 50-, 100-, and 250-μm amplifier and (b) the 100-, 200-, and 400-μm amplifier.

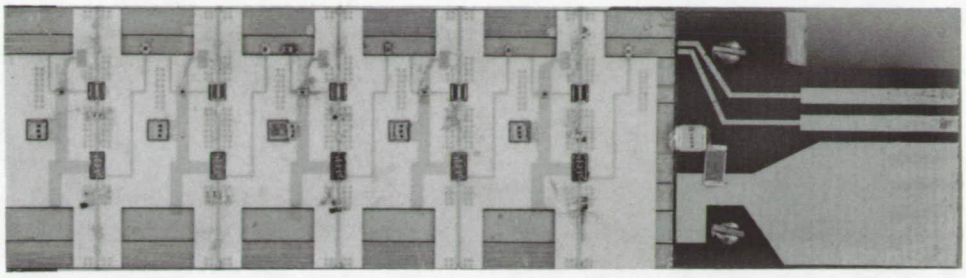


Fig. 8. Amplifier module containing the 100-mW 32-GHz amplifier and 50-Ω through line.

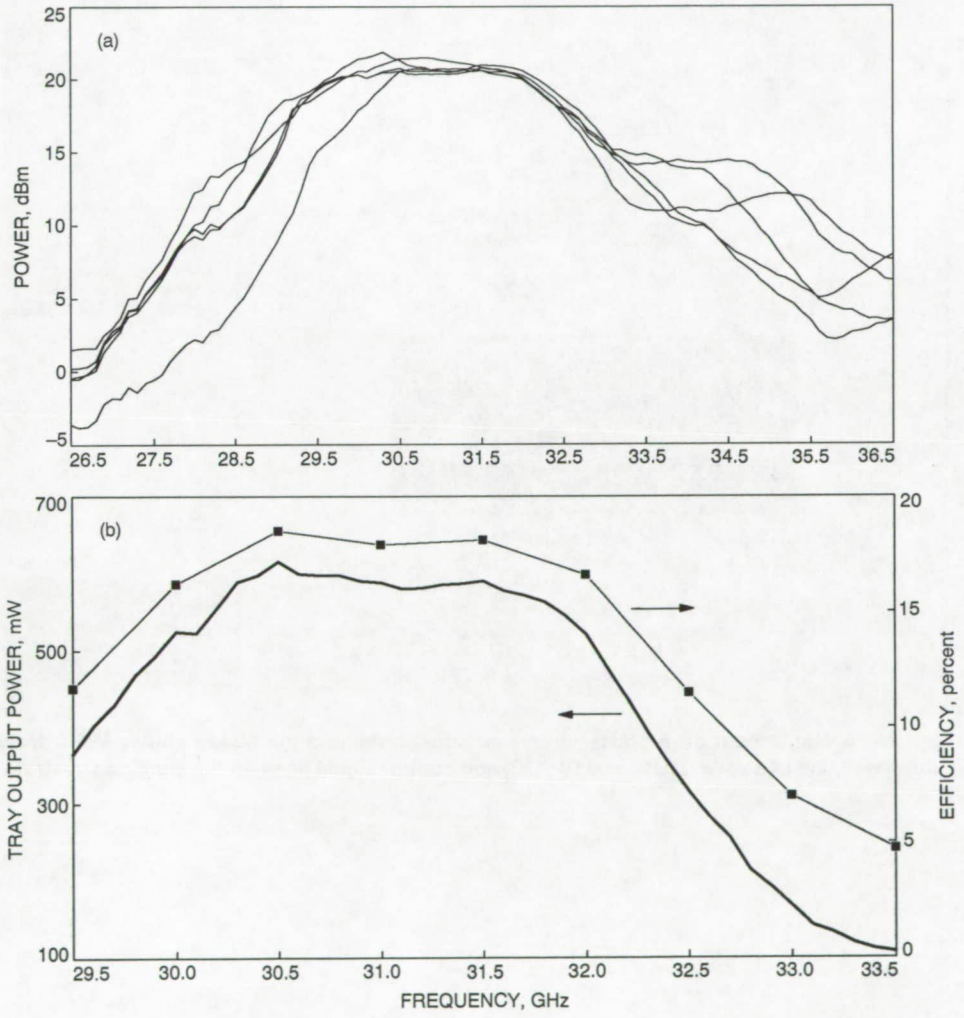


Fig. 9. Amplifier module measurements: (a) output power of each of five paths on an amplifier module over the 26.5- to 36.5-GHz range with 2-dBm input power and (b) total power and efficiency of the amplifier module.

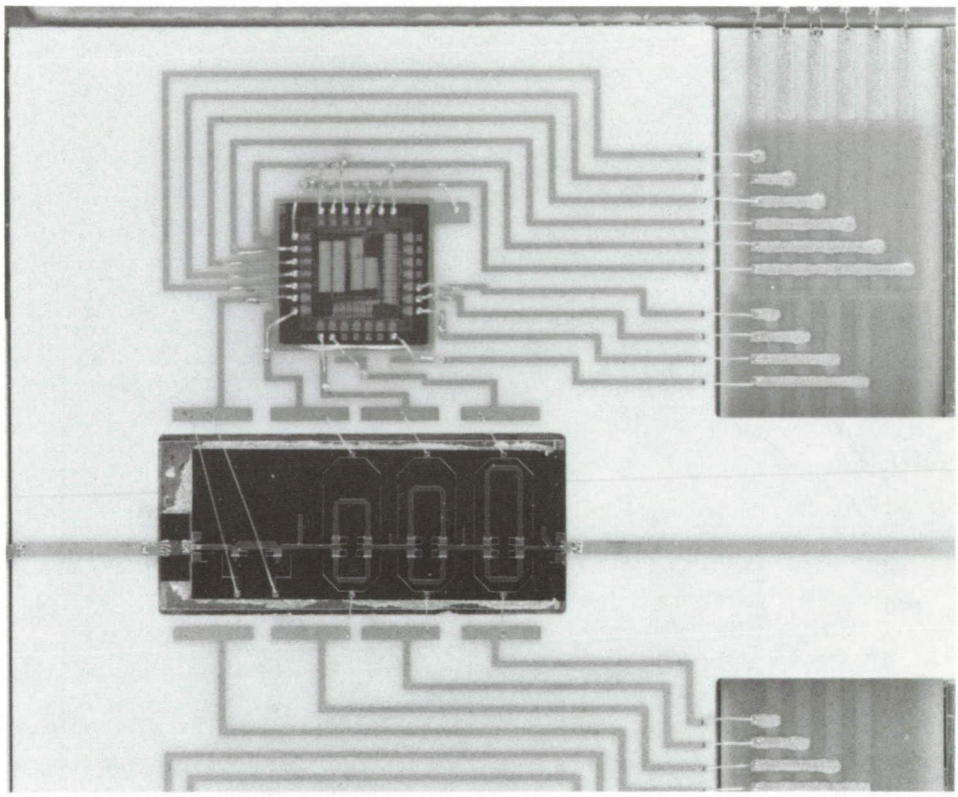


Fig. 10. A single path on a phase shifter module containing the phase shifter MMIC from Honeywell, the controller ASIC, and RF, DC, and control signal lines on the alumina substrate.

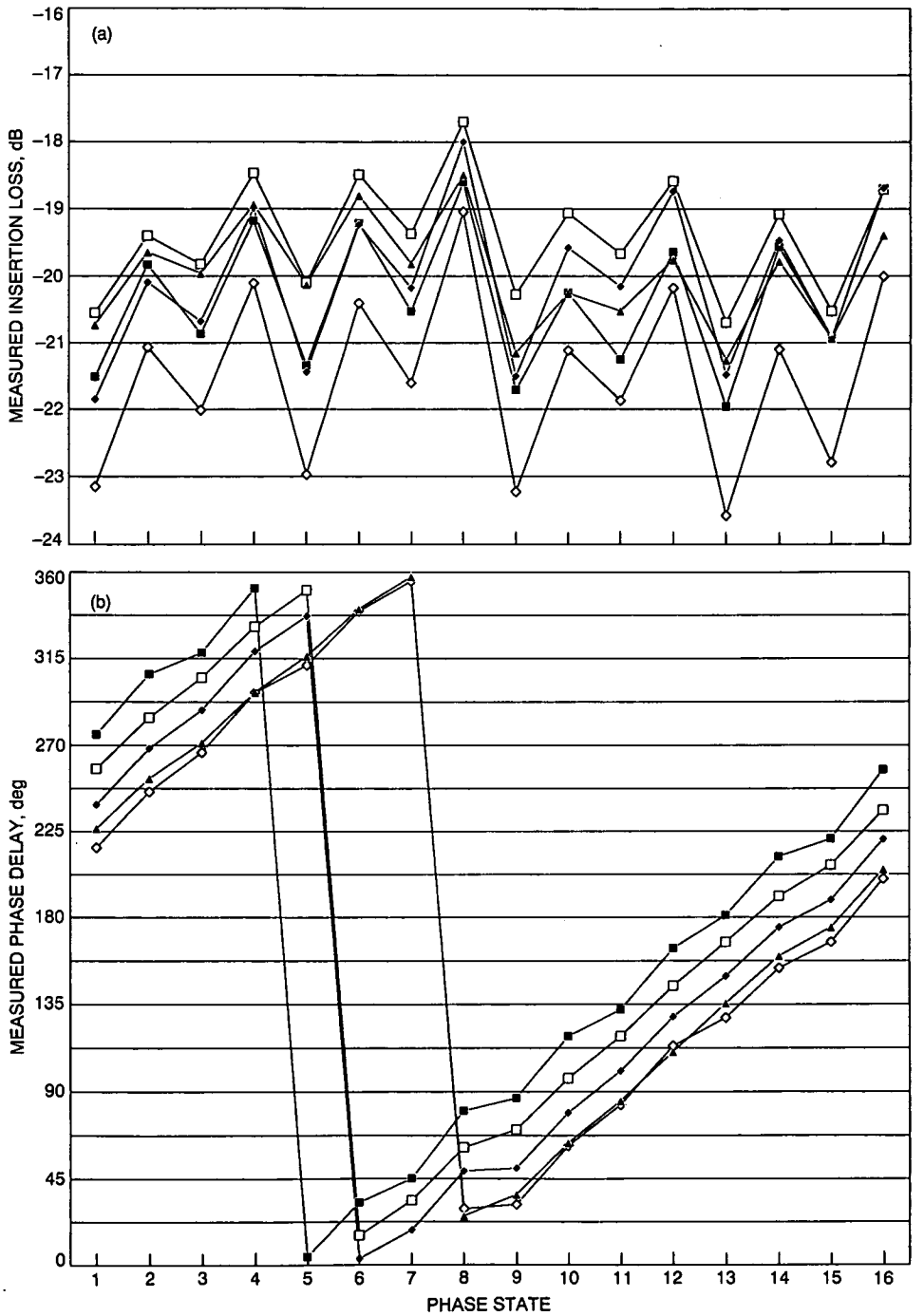


Fig. 11. Phase shifter module measurements: (a) insertion loss and (b) phase for each of five paths in each of 16 states at 32 GHz; (c) output return loss of a single path in each of 16 states over the 26.5- to 36.5-GHz range.

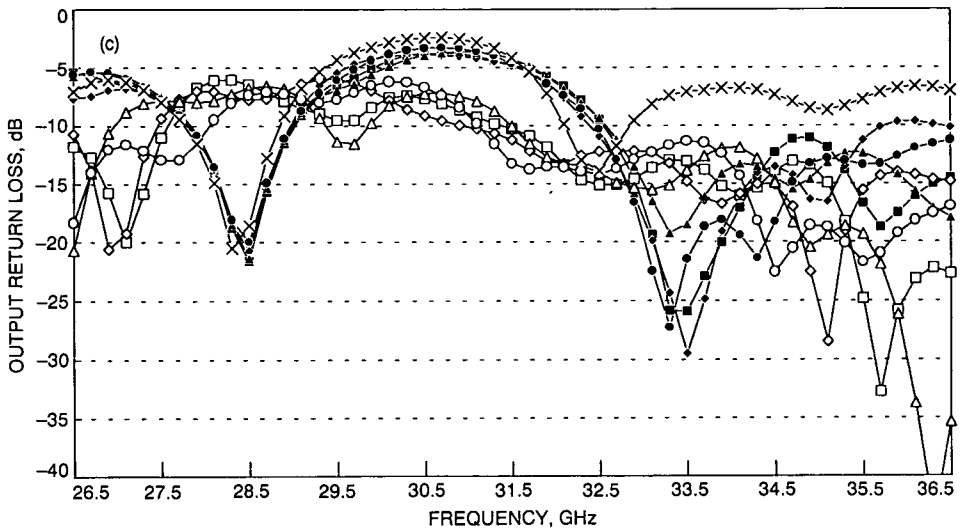

(c)
0
-5
-10
-15
-20
-25
-30
-35
-40
OUTPUT RETURN LOSS, dB
26.5
27.5
28.5
29.5
30.5
31.5
32.5
33.5
34.5
35.5
36.5
FREQUENCY, GHz

Fig. 11 (contd).

A Statistical Model for Evaluating GOPEX Uplink Performance

K. Kiesaleh and T.-Y. Yan
Communications Systems Research Section

This article describes a statistical model to analyze the signal intensity received at the solid-state imaging (SSI) camera of the Galileo optical communications system from an Earth-based transmitter (GOPEX) demonstration. The analytical model assumes that the optical beam possesses a Gaussian profile and the communication channel has a log-normal scattering characteristic. The atmospheric-induced jitter is modelled as two independent zero mean Gaussian random variables. By modelling the system parameters as a set of independent and identically distributed (iid) random variables, the combined impact of uncertainties due to system parameters and the turbulent atmosphere can be approximated by a log-normal distributed signal intensity at the spacecraft. A Monte-Carlo software simulation package has also been developed to compute the confidence interval probabilities for general optical beam profiles. Numerical results show that the approximation is valid for a wide range of operation scenarios.

I. Introduction

In this article, analytical expressions for the probability density functions (pdf's) of the optical field intensity and the number of observed photoelectrons per pixel of Galileo's solid-state camera for a laser pulse interval in the Galileo optical communication from an Earth-based transmitter (GOPEX) experiment are derived. These pdf's play a critical role in assessing the performance of this optical uplink and provide a means of acquiring meaningful design predicts. Furthermore, the aforementioned pdf's shed light on the characteristics of the underlying processes involved in this experiment and the impacts of various parameters on the detection probability.

It is assumed that the main sources of disturbance are the atmosphere-induced jitter, for the short-exposure model at hand, and the log-normal large-particle channel scattering, which proves to be significant for the operating characteristics of this experiment. In general, there are a number of parameters, described below, that directly impact the observed signal intensity at the spacecraft. These parameters are assumed to be independent and identically

distributed (iid) random variables. It is imperative to note the importance of this observation, since as a result of this assumption the signal intensity in the absence of atmospheric effects may be approximated by a log-normal random variable as well. In this article, the log-normal channel scattering is shown to be the dominant effect for the operating characteristics of this experiment. Thus, it is concluded that the combined impact of random system parameters and turbulent atmosphere would yield a log-normal distributed signal intensity at the spacecraft. Before going any further, a description of the system parameters that directly impact the observed signal intensity and which are needed to conduct the ensuing analytical study of the GOPEX uplink is presented.

The optical field intensity at the spacecraft is a function of transmitter optics efficiency, η_t, receiver optics efficiency, η_r, atmospheric transmittance, η_{atm}, transmitter laser energy (in joules), W_t, distance from the spacecraft, Z, and the predefined angular beam diameter, θ_s. Since the number of photoelectrons per pixel, observed over a laser pulse duration, is of importance here, one must also include the effective detector area (pixel area), A_{rec}, and the quantum efficiency of the photodetector, η_c, in determining the overall statistical modelling. In the event of a uniform beam pattern, and for $Z >> 1$, one can describe the optical field intensity in joules/m^2 at the spacecraft as

$$< I(x,y,t) >_t = \frac{4W_t}{\pi Z^2 \theta_s^2} \eta_t \eta_r \eta_{atm} \qquad (1)$$

where $< I(x,y,t) >_t$ describes the intensity in the plane of observation at coordinates (x,y), and $< . >_t$ signifies short-term intensity averaging over time, since the observation interval is limited to only 6 to 12 nsec, the laser pulse duration. The number of observed photoelectrons in the laser pulse duration is then

$$N_{electron} = < I(x,y,t) >_t A_{rec} \eta_c \qquad (2)$$

These expressions will be modified in the next section to include the Gaussian beam profile and the log-normal large-particle scattering effect.

II. Uplink Analysis: Gaussian Beam Profile

In this section, a transmitted beam with a Gaussian profile is assumed. Therefore, for a short exposure model, one may assume that the Gaussian beam profile maintains its integrity in the plane of observation. In this case, the optical field intensity at (x,y) in the plane of observation may be approximated by

$$< I(x,y,t) >_t \approx \frac{4W_t}{\pi Z^2 \theta_s^2} \eta_t \eta_r \eta_0^{\sec(\theta)} I_l \exp\left[-\frac{4(x^2+y^2)}{\theta_s^2 Z^2}\right] \qquad (3)$$

where now I_l accounts for the log-normal atmospheric scattering effect, η_0 is the atmospheric loss at zenith, and θ is the spacecraft's zenith angle. This implies that $\eta_{atm} = \eta_0^{\sec(\theta)}$, which agrees with most experimental and theoretical studies of atmospheric absorption. In this analysis, assume a negligible pointing error, and thus the field intensity must be evaluated for $(x,y) = (0,0)$. Furthermore, since the receiving area of the charged-coupled device (CCD) camera is significantly smaller than the beam footprint at the spacecraft location, the CCD camera may be considered a point detector. Therefore, Eq. (3) at $(x,y) = (0,0)$ accurately describes the field intensity for all the detectors of the CCD camera.

The pdf of I_l is given by the well-known log-normal density

$$f_{I_l}(i_l) = \frac{1}{\sqrt{2\pi\sigma_l^2} i_l} \times \exp\left\{-\frac{[\ln(i_l) + \frac{\sigma_l^2}{2}]^2}{2\sigma_l^2}\right\}; \qquad i_l \geq 0 \qquad (4)$$

where σ_l is, in turn, given by [1]

$$\sigma_l^2 \approx 2.24 (\frac{2\pi}{\lambda})^{\frac{7}{6}} \sec(\theta)^{\frac{11}{6}} \times \int_0^H C_n^2(h) h^{\frac{5}{6}} \, dh \qquad (5)$$

In the above equation, λ is the wavelength of the laser in meters, H is the height of the atmosphere, and $C_n(h)$ is the medium index of the refraction structure constant. Major stumbling blocks in determining an accurate estimate of σ_l^2 are the dependency of C_n on various random channel parameters and the unavailability of an accurate model for C_n. However, in the literature a number of approximate models for this structure constant are available.

For a complete list, refer to [2]. Among those references, the following expression, originally proposed by Hufnagel, approximates this structure constant:

$$C_n^2(h) = 8.2 \times 10^{-56} V^2 h^{10} e^{-h/1000} + 2.7 \times 10^{-16} e^{-h/1500} \quad (6)$$

where h is the altitude in meters above sea level and V is the rms wind velocity, averaged over 5 to 20 km altitude, and is considered to be Gaussian distributed about a mean of 27 m/sec with a standard deviation of 9 m/sec. This model, however, is valid only for altitudes in excess of 5 km. For altitudes below 5 km, an approximate expression from Hufnagel and Stanely for the structure constant is given by

$$C_n^2(h) \approx 1.5 \times 10^{-13}/h \quad (7)$$

For the parameters of the GOPEX project, and based upon the above approximations, $\sigma_l \leq 0.9$.

In a turbulent medium, a Gaussian beam, as described above, experiences deflection in various directions due to gaseous blobs and other turbulent particles flowing through the path of the beam. For a short exposure model (6- to 12-nsec observation interval), the atmospherically induced deflections may accurately be modelled as atmospherically induced pointing "jitter." This implies that x and y in Eq. (3) may now be viewed as two independent and zero mean Gaussian random variables with standard deviation σZ. Thus, σ represents the rms angular (half-beam) jitter due to turbulence. When pointing error is present, its standard deviation may be added directly to σ. However, caution must be exercised in applying the ensuing results to a model with a constant pointing error. In this event, x and y are nonzero mean Gaussian random variables. For this analysis, however, the constant pointing error is considered to be negligible, and thus one may consider x and y as zero mean Gaussian random variables.

The standard deviation σ can be estimated for a given index of refraction profile by using the following expression [3]:

$$\sigma = \frac{2}{k}\left[1.46k^2 \sec(\theta) \times \int_0^{20000} d\zeta C_n^2(\zeta)\left(1 - \frac{\zeta}{20000}\right)^{5/3}\right]^{3/5} \quad (8)$$

where k is the wave number. For the parameters of this experiment, σ is estimated to be 9 μrad for near zenith and 14 μrad for a zenith angle of 60 deg. Define the random variable I_g as follows:

$$I_g(x,y) = \exp\left[-\frac{4(x^2+y^2)}{\theta_s^2 Z^2}\right] \quad (9)$$

It can readily be shown that for jointly Gaussian x and y, I_g is a special case of *beta* distributed random variables with pdf

$$f_{I_g}(i_g) = \beta i_g^{(\beta-1)}; \qquad 0 \leq i_g \leq 1 \quad (10)$$

where $\beta = \theta_s^2/8\sigma^2$. In the following, β ranges from 5 to 19 ($\theta_s = 110\,\mu$rad and $\sigma \approx 9$ to 14 μrad) for all practical purposes. It is interesting to note that for $\beta = 1$ this density reduces to a uniform density, signifying the detrimental impact of a turbulent medium.

The pdf of $I_t = I_g I_l$ can now be found. It represents the combined impact of log-normal medium scattering and atmospherically induced pointing jitter. This pdf is expressed as

$$f_{I_t}(i_t) = \beta \exp\left[\frac{\sigma_l^2}{2}\beta(\beta+1)\right] i_t^{(\beta-1)} \times Q\left[\frac{\ln(i_t) + \sigma_l^2(\beta+1/2)}{\sigma_l}\right]; \quad I_t \geq 0 \quad (11)$$

where $Q(x) = 1/(\sqrt{2\pi}) \int_x^\infty \exp(-s^2/2)\, ds$. For a reasonable range of system parameters, this density is depicted in Figs. 1 through 4. It is important to note that σ_l plays a critical role in determining the behavior of this random variable. Unfortunately, for zenith angles in the range of 40 to 55 deg, $\lambda = 0.532$ μm, and a rms wind velocity of 27 m/s, σ_l proves to be in excess of 0.2, resulting in a pdf that can accurately be approximated by a log-normal density function. This implies that for all practical purposes,

$$f_{I_t}(i_t) \approx \frac{1}{\sqrt{2\pi\sigma_t^2 i_t}} \times \exp\left\{-\frac{[\ln(i_t) - m_t]^2}{2\sigma_t^2}\right\}; \quad i_t \geq 0 \quad (12)$$

where

$$m_t = \int_0^\infty \ln(s)\beta \exp\left[\frac{\sigma_l^2}{2}\beta(\beta+1)\right] s^{(\beta-1)} \times Q\left[\frac{\ln(s)+\sigma_l^2(\beta+1/2)}{\sigma_l}\right] ds \quad (13)$$

$$\sigma_t^2 = \int_0^\infty \ln^2(s)\beta \exp\left[\frac{\sigma_l^2}{2}\beta(\beta+1)\right] s^{(\beta-1)} \times Q\left[\frac{\ln(s)+\sigma_l^2(\beta+1/2)}{\sigma_l}\right] ds - m_t^2 \quad (14)$$

These expressions can be numerically evaluated for a desired set of system parameters. The validity of approximating the characteristics of I_t with that of a log-normal random variable is examined in a number of plots (see Figs. 5 through 8). As noted, for a wide range of system parameters the penalty for this approximation is comfortably small.

One can now express the number of observed photoelectrons in a pulse interval as

$$N_{electron} = \frac{4W_t}{\pi Z^2 \theta_s^2} \eta_{tot} I_t \quad (15)$$

where $\eta_{tot} = \prod_{i=1}^5 \eta_i$, with $\eta_1 = \eta_r$, $\eta_2 = \eta_t$, $\eta_3 = \eta_0^{\sec(\theta)}$, $\eta_4 = \eta_c$, and $\eta_5 = A_{rec}$. If one considers η_i's as iid random variables, it is possible to approximate η_{tot} with a log-normal random variable. Due to a lack of sufficient data for accurate characterization of these parameters or for considering the worst-case scenario, one may assume that η_i is uniform over $(\eta_i^{min}, \eta_i^{max})$ for all i. In this event,

$$f_{\eta_{tot}}(\eta_{tot}) \approx \frac{1}{\sqrt{2\pi\sigma_\eta^2}\eta_{tot}} \times \exp\left\{-\frac{[\ln(\eta_{tot}) - m_\eta]^2}{2\sigma_\eta^2}\right\}; \eta_{tot} \geq 0 \quad (16)$$

where $m_\eta = \sum_{i=1}^5 m_{\eta_i}$ and $\sigma_\eta^2 = \sum_{i=1}^5 \sigma_{\eta_i}^2$ with

$$m_{\eta_i} = \frac{1}{\eta_i^{max} - \eta_i^{min}} \int_{\eta_i^{min}}^{\eta_i^{max}} \ln(s)\, ds = \frac{1}{\eta_i^{max} - \eta_i^{min}} \times \left[\eta_i^{max}\ln(\eta_i^{max}) - \eta_i^{min}\ln(\eta_i^{min})\right] - 1 \quad (17)$$

$$\sigma_{\eta_i}^2 = \frac{\eta_i^{max}}{\eta_i^{max} - \eta_i^{min}} \times \left[\ln^2(\eta_i^{max}) - 2\ln\left(\frac{\eta_i^{max}}{e}\right)\right] - \frac{\eta_i^{min}}{\eta_i^{max} - \eta_i^{min}} \times \left[\ln^2(\eta_i^{min}) - 2\ln\left(\frac{\eta_i^{min}}{e}\right)\right] - m_{\eta_i}^2 \quad (18)$$

Because the product of two log-normal random variables yields log-normal statistics

$$f_{N_{tot}}(n_{tot}) \approx \frac{1}{\sqrt{2\pi\sigma_N^2} n_{tot}} \times \exp\left\{-\frac{[\ln(n_{tot}) - m_N]^2}{2\sigma_N^2}\right\}; \; n_{tot} \geq 0 \quad (19)$$

where $N_{tot} = \eta_{tot} I_t$, $m_N = m_\eta + m_t$, and $\sigma_N^2 = \sigma_\eta^2 + \sigma_t^2$. Finally, the total number of photoelectrons observed over a pulse interval is given by

$$N_{electron} = \frac{4W_t}{\pi Z^2 \theta_s^2} N_{tot} \quad (20)$$

Note that an alternate means of computing σ_N and m_N is to first compute the mean and mean-square of $\eta_{tot} I_t$, which are given by

$$E\{N_{tot}\} = \frac{\beta}{1+\beta} \prod_{i=1}^5 \frac{(\eta_i^{max} + \eta_i^{min})}{2} \quad (21)$$

$$E\{N_{tot}^2\} = \exp(\sigma_l^2) \frac{\beta}{2+\beta} \times \prod_{i=1}^5 \frac{[(\eta_i^{max})^2 + (\eta_i^{min})^2 + \eta_i^{max}\eta_i^{min}]}{3} \quad (22)$$

and use the following expressions:

$$\sigma_N^2 = \log\left[E\{N_{tot}^2\}/E^2\{N_{tot}\} + 1\right] \quad (23)$$

$$m_N = \log\left[E\{N_{tot}\}\right] - \frac{1}{2}\sigma_N^2 \quad (24)$$

It is imperative to note that the accuracy of the above model is highly sensitive to σ_l and other system parameters. In particular, caution must be exercised in assuming a log-normal scattering model for values of $\sigma_l > 0.75$, since for such values of σ_l the variance of the normalized intensity fluctuation due to scattering (i.e., $e^{\sigma_l^2} - 1$) exceeds 0.75. In this event, log-normal statistics are no longer valid in describing the characteristics of the scattering channel. Instead, an exponential pdf must be employed to characterize the scattering channel [3,4]. For the experiment at hand, σ_l for $V \leq 27$ m/sec and $\theta \leq 55$ deg is less than 0.9. This, therefore, makes the log-normal assumption rather suspect for some portion of this experiment where the spacecraft takes on large zenith angles. However, since the rms wind velocity is a random variable, and σ_l is highly sensitive to this parameter, one may approximate the channel scattering effect with the log-normal statistic.

III. Uplink Analysis: General Beam Profile

This section provides a brief description of a software simulation package that was developed for the analysis of the GOPEX uplink when the optical beam possesses a general profile. To compute confidence interval probabilities, one has to resort to a Monte Carlo simulation to include the impacts of atmospherically induced jitter, log-normal channel scattering, and uncertainties in other system parameters that are modelled as uniformly distributed random variables (see above).

This package reads the beam matrix profile, recorded at a known distance, and evaluates the beam profile at the location of the spacecraft. The turbulent medium is, once again, modelled as a zero mean Gaussian pointing jitter in both x and y coordinates. This program provides the mean, standard deviation, and the pdf of the number of observed photoelectrons in a pulse duration for a specified interval. The confidence probability for a given confidence interval is also computed. Since no theoretical results were available to test the validity of the results of this program for a general beam profile, a Gaussian beam was specified as the input beam profile. The system parameters were set at the following values: $W_t = 0.25$ joules, $\lambda = 0.532\,\mu$m, $\theta_s = 110\,\mu$rad, $\eta_t = 0.705$, $\eta_r = 0.24$, $\eta_c = 0.37$, $\eta_0 = 0.715$ (atmospheric attenuation at zenith and station altitude of 2.286 km [5]), $A_{rec} = 0.01825$ m^2, $Z = 0.6 \times 10^6$ km, $V = 27$ m/sec, and $\theta = 55$ deg. From the above analytical results for the Gaussian beam profile, the mean and standard deviation of the observed photoelectrons were estimated at 111.991×10^3 and 126.08×10^3, respectively. The large standard deviation is a clear indication of log-normal statistics due to dominant channel scattering. The simulation for 10,000 samples yielded the following mean and standard deviation: 110.834×10^3 and 119.535×10^3. These numbers improved to 111.950×10^3 and 124.361×10^3 for 50,000 samples. It is quite clear that a reasonable accuracy may be achieved with a relatively small number of samples. However, with the aid of stratified or importance sampling, the required number of samples may be drastically reduced. The above calculation was repeated for a 60-deg elevation, i.e., $\theta = 30$ deg, with the remaining parameters fixed at the above values. From theoretical results, the mean and standard deviation were found to be 142.025×10^3 and 97.166×10^3, respectively. For 10,000 samples, simulation yielded 141.27×10^3 and 94.877×10^3 as the mean and standard deviation of the observed photoelectrons, respectively. Once again, these numbers improved to 142.095×10^3 and 96.536×10^3 for 50,000 samples, clearly indicating the consistency of the simulation.

IV. Conclusion

This article described a statistical model to evaluate the signal intensity received at the solid-state imaging camera for the GOPEX demonstration. The model includes the effect of log-normal channel scattering, atmospherically induced jitter, and uncertain system parameters. It has been shown that the resulting probability density function can be well approximated by a log-normal distribution. A Monte Carlo simulation software package has been developed to analyze the uplink performance when the optical beam profile is non-Gaussian. The confidence-interval probability can be computed to analyze the GOPEX experimental data during demonstration.

References

[1] V. I. Tatarskii, *Wave Propagation in a Turbulent Medium*, New York: McGraw Hill, 1961.

[2] W. L. Wolf and G. J. Zissis, Chapter 6 in *The Infrared Handbook*, Detroit: The Infrared Information Analysis (IRIA) Center, Environmental Research Institute of Michigan, 1989.

[3] S. Karp, R. Gagliardi, S. E. Moran, and L. B. Stotts, *Optical Channels*, New York: Plenum Press, pp. 178–180, 1988.

[4] R. Fante, Electromagnetic Beam Propagation in Turbulent Medium, an Update, *Proc. IEEE*, vol. 68, pp. 1424–1443, 1980.

[5] K. S. Shaik, "Atmospheric Transmission Calculations for Optical Frequencies," *Proceedings of the Thirteenth NASA Propagation Experiments Meeting (NAPEX XIII)*, pp. 158–162, June 29-30, 1989.

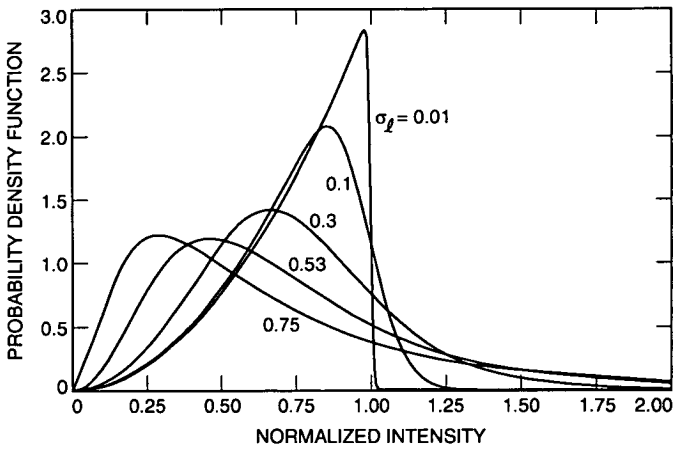


Fig. 1. Probability density function of the normalized intensity, $\beta = 3$.

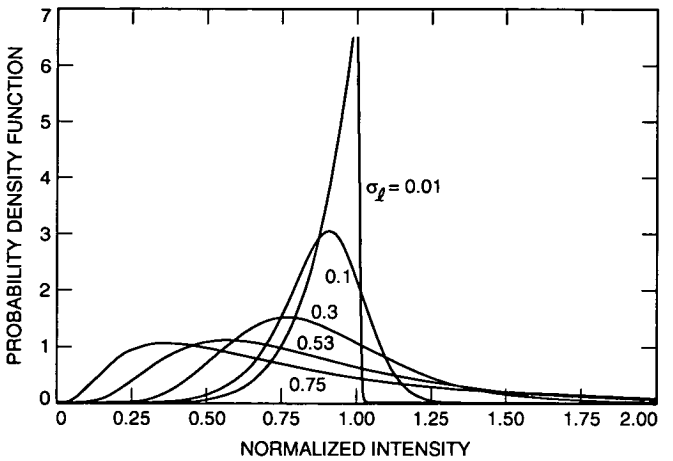


Fig. 3. Probability density function of the normalized intensity, $\beta = 7$.

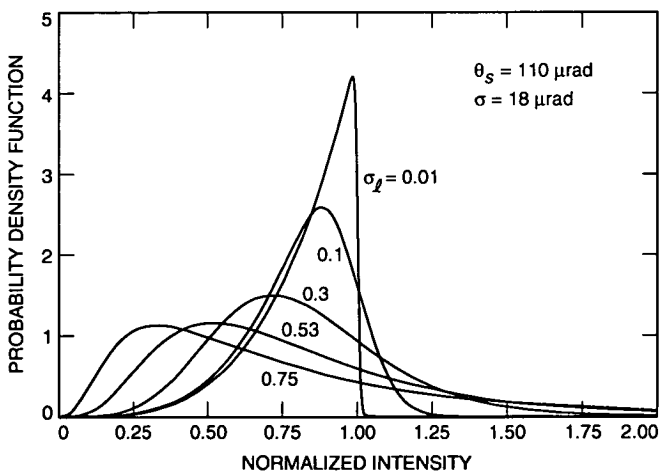


Fig. 2. Probability density function of the normalized intensity, $\beta = 4.6$.

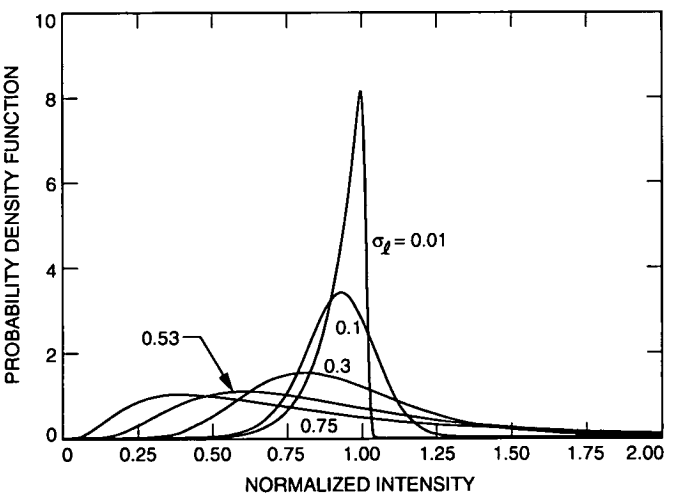


Fig. 4. Probability density function of the normalized intensity, $\beta = 10$.

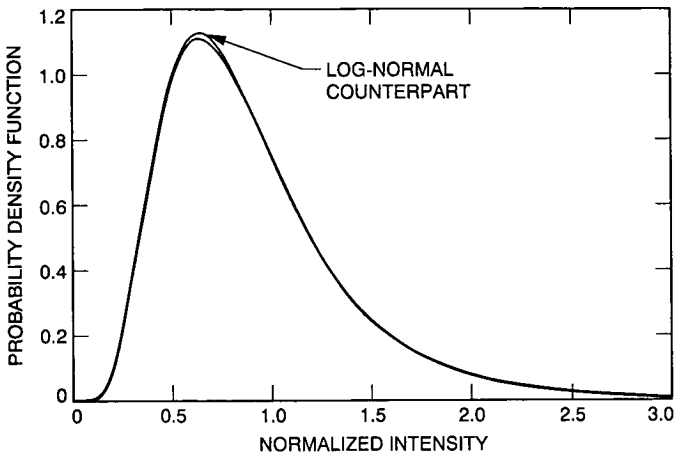


Fig. 5. Probability density function of the normalized intensity, β = 5 and σ_ℓ = 0.5, and its approximated log-normal counterpart.

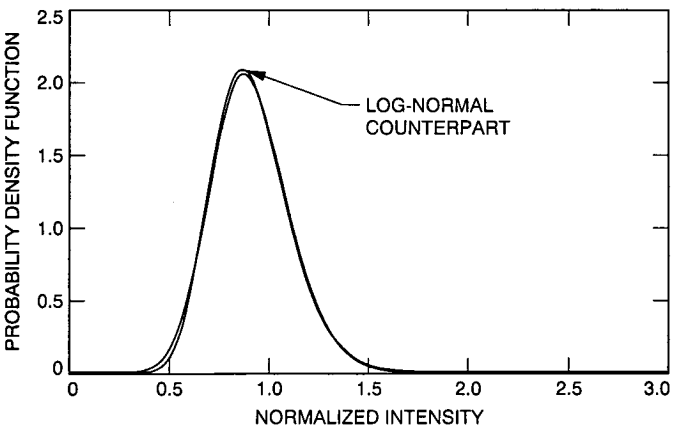


Fig. 6. Probability density function of the normalized intensity, β = 10 and σ_ℓ = 0.2, and its approximated log-normal counterpart.

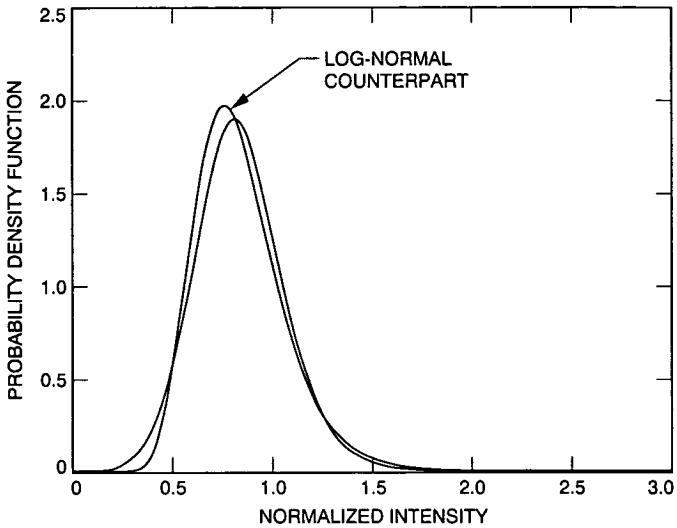


Fig. 7. Probability density function of the normalized intensity, β = 5 and σ_ℓ = 0.2, and its approximated log-normal counterpart.

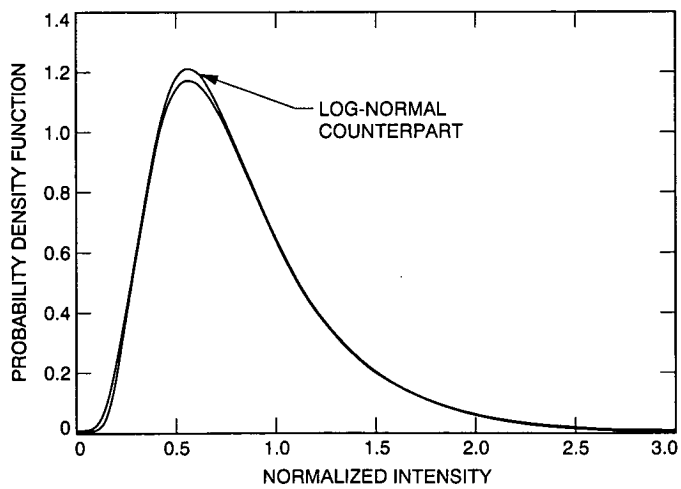


Fig. 8. Probability density function of the normalized intensity, β = 5 and σ_ℓ = 0.5, and its approximated log-normal counterpart.

A Model to Compare Performance of Space and Ground Network Support of Low-Earth Orbiters

E. C. Posner
Office of Telecommunications and Data Acquisition

This article compares the downlink performance in a gross average sense between space and ground network support of low-Earth orbiters. The purpose is to assess what the demand for DSN support of future small low-cost missions might be, if data storage for spacecraft becomes reliable enough and small enough to support the storage requirements needed to enable support only a fraction of the time. It is shown that the link advantage of the DSN over space reception in an average sense is enormous for low-Earth orbiters. The much shorter distances needed to communicate with the ground network more than make up for the speedup in data rate needed to compensate for the short contact times with the DSN that low-Earth orbiters have. The result is that more and more requests for DSN-only support of low-Earth orbiters can be expected.

I. Introduction

NASA's Tracking and Data Relay Satellite System (TDRSS) can support low-Earth orbiters, providing continuous coverage when necessary, for example, for piloted missions. Due to the synchronous altitude of TDRSS, the communication path length to a low-Earth orbiter is about equal to the synchronous altitude, or 36,000 km. The communication path from an Earth station such as the DSN to a low-Earth orbiter is much less, resulting in a huge reduction in path loss. On the other hand, the visibility (fraction) of a low-Earth orbiter reduction to a ground network of only three locations, such as the DSN, is much less than 1, i.e., the fraction of the time a low-Earth orbiter is visible to the DSN is small. This means that the data must be stored and then dumped over the station during the orbiter's passes or times of visibility. This means an increase in the required data rate, to partially offset the gain from the shorter communication path.

For example, in a typical case of a 400-km circular orbit at a random inclination, ignoring station masks, the network sees the orbiter 8.8 percent of the time, for a required increase in image data rate of $1/0.088 = 11.36$, a 10.6-dB loss as accounted for in this article. This loss is much more than made up for by the gain due to shorter communication paths. The distance at rise or set for a 400-km orbit is 2298 km. This gives a gain at rise or set relative to the 36,000-km synchronous altitude of $(36000/2298)^2 = 245.5$ or 23.9 dB. The data rate can be varied during a pass, i.e., increased as the spacecraft rises and decreased as it sets. Although continuous rate variation is assumed below, it turns out that most of the improvement can be realized

with only two rate changes. On average, data volume for 400-km orbits improves an average increased factor of 3.61, or 5.6 dB, as is shown below. The total average improvement for using an Earth station instead of one at a synchronous altitude for a random 400-km circular orbit is then (23.9 + 5.6 − 10.6) dB = 18.9 dB.

The 18.9-dB average improvement does not consider that a ground station will typically have a greater aperture than a space relay spacecraft, and can also have a much lower noise temperature than that which could be provided by any space low-noise reception technology forseeable over the next decade. For example, a 26-m ground station such as in the DSN has an advantage in gain of 14.3 dB over a 5-m TDRSS-style space-based antenna, assuming equal antenna efficiencies, yet ignoring any noise temperature differences, which tend to favor ground-based antennas. Also, the space system may not use its highest gain antenna to support simpler missions, but rather its multiple-access capability. The multiple-access capability allows several spacecraft to be supported at once, which may allow the continuous support assumed here, but at a still greater advantage to the ground than the 14.3 dB quoted above.

The effect of frequency selection needs to be considered. Here, roughly the same frequency is assumed for the space network as for the ground network. The space network could ultimately make use of a relay frequency of 60 GHz, or even of an optical frequency. These choices could change some of the considerations. On the other hand, receiver improvements are much easier to make in a ground station than in a space relay station, once launched. In fact, servicing a synchronous spacecraft will be very difficult and is certainly not foreseen for at least 20 years. The rest of the article, therefore, assumes roughly the same frequencies for the space network as for the ground network.

Other factors ignored in this article, which is not intended to substitute for a system engineering trade-off study, are any costs to the using spacecraft of the access protocol to use a multiple-access capability on the space-based relay platform, implementation cost differences between space-based and ground-based reception capabilities and likewise any operations cost differences, the ability to upgrade a ground station during its lifetime, the number of spacecraft practically supportable by a space or ground network capability, the differential costs of onboard data storage by using the ground-based capability, and finally the costs of the using spacecraft communications package to produce the required effective isotropic radiated power (EIRP) in the two cases. The assumption of random circular orbits would also have to be replaced by realistic orbits and mission requirements in a more detailed analysis, which would also include realistic station masks.

Nevertheless, the potential gain of (18.9 + 14.3) dB = 33.2 dB or more (a factor of 2100) will be attractive to using spacecraft as a cost, size, and weight reduction, assuming the availability of data storage that can overcome the low visibility from the ground. This will apply then to robotic missions that do not have very high average data acquisition rates. Many missions may be included in the set of potential users in coming years as spacecraft are downsized. Such spacecraft may use the 33.2-dB or more average decrease in required spacecraft EIRP in several ways, or use all these ways at once: lower spacecraft power, a lower gain or even an omnidirectional spacecraft antenna, less or no restrictions on spacecraft attitude, and a general relaxation of the constraints on the spacecraft radio. The mass reduction allows a smaller launch vehicle, which could open up whole new mission opportunities.

There are corresponding uplink and navigation improvements as well, which are not considered in this article. Overall, the result of support of low-Earth orbiters by the ground will be the provision of a communications capability that will result in smaller, cheaper, and lighter spacecraft of the same capability. Thus, it can be expected that in coming years there will be more and more low-Earth orbiters that are designed for support exclusively by the DSN.

The remainder of this article is organized as follows. Section II presents the simple model of random circular orbits that is being considered. Section III calculates the average visibility of a low-Earth orbiter to the DSN, and thus the amount of data storage that must be provided on board the using spacecraft. Section IV finds the average improvement in the data volume received when using the ground network as opposed to the space network. Finally, Section V summarizes the results and computes the average advantage of ground versus space network support for low-Earth orbiters, considering both communication distance and spacecraft visibility.

II. Model of Random Circular Orbits

Here random circular orbits are assumed. Real orbits are only sometimes circular. Moreover, certain inclinations are often preferred, due for example to the requirement of keeping the line of apsides fixed or to provide Sun synchronism. The location of a single tracking station on Earth may as well be assumed fixed, given random orbits. The rotation of the Earth can also then be ignored, since

one is in effect assuming a random phase of the orbiter in its orbit. The DSN has its communication complexes far enough apart so that there is never simultaneous visibility of low-Earth orbiters. Thus, visibilities from a single station, when multiplied by 3, the number of complexes, give the long-term average visibility to the entire DSN with its present complex locations.

The radius of the Earth, r, is taken in calculations as 6400 km. The altitude above Earth of the circular orbit is h; in calculations, h is taken as 400 km. The fraction of time that a spacecraft in random circular orbit at distance h above the Earth is visible to a given ground station on a spherical Earth of radius r is found in the next section.

III. Average Visibility of a Low-Earth Orbiter

Since the spacecraft is in a random circular orbit, random phase and inclination, it is at a given time at a random (uniform) point on a sphere of radius $r + h$ centered at the Earth's center. A given tracking station sees a certain fraction of the surface area of the larger sphere, and that fraction is clearly the average visibility from a single station. The single station may as well be random too, since the average visibility is clearly independent of the station's location. By the same reasoning, the spacecraft may as well be considered fixed. Thus, the visibility of a spacecraft from a single Earth station is just the fraction of the Earth's area seen by a fixed spacecraft.

The fraction of the area on the Earth seen by a spacecraft at height h above the Earth is now easily found (see Fig. 1). A tangent is drawn from the spacecraft to the Earth, and the area of the spherical cap is sought. Let θ be the angle subtended by the spacecraft tangent, so that

$$\theta = \arccos\left(\frac{r}{r+h}\right)$$

Let ϕ be the angle of integration between a point on the cap and the center of the Earth. This angle goes from 0, the angle corresponding to the spacecraft location at zenith, to the edge of Earth visibility at angle θ. The perpendicular radius of the circle at ϕ is $r\sin\phi$, and its circumference is $2\pi r\sin\phi$. The slant length of the differential ring corresponding to angle $d\phi$ around ϕ is $rd\phi$. Thus, the area dA of the differential ring is

$$dA = 2\pi r\sin\phi \cdot rd\phi = 2\pi r^2 \sin\phi d\phi$$

The area of the cap A is then

$$A = \int_{\phi=0}^{\theta} 2\pi r^2 \sin\phi d\phi$$

$$= 2\pi r^2 \left(-\cos\phi\right)_{\phi=0}^{\theta}$$

$$= 2\pi r^2 \left(1 - \cos\theta\right)$$

$$= 2\pi r^2 \left(1 - \frac{r}{r+h}\right)$$

so that

$$A = 2\pi r^2 \left(\frac{h}{r+h}\right)$$

The area of the surface of the Earth is $4\pi r^2$, so the visibility V from a single station is

$$V = 2\pi r^2 \left(\frac{h}{r+h}\right) / 4\pi r^2$$

$$V = \frac{h}{2(r+h)}$$

With h = 400 km and r = 6400 km, V = 1/34. The visibility from the entire network of three complexes is 3 times this, or 3/34 = 8.8 percent. This is a long-term average. It does not of course mean that the DSN will see 8.8 percent of every or any orbit.

Assuming the orbiter is gathering data even while transmitting, it must transmit 34/3 = 11.33 times as fast (10.5 dB more link is needed) as a spacecraft that can always be seen and is continually serviced by a synchronous relay satellite system. Such a data rate is supportable even by present technology for satellites whose data acquisition rate is up to several megabits per second. This is because all the data can be dumped if during a pass there is an average data rate of tens of megabits per second, which is compatible with current technology. Future data acquisition digital equipment in the signal processing centers of the DSN is expected to be able to accommodate even higher peak rates.

What about data storage requirements? Maximum storage requirements cannot be found from the averaging techniques used here. Nevertheless, the following discussion is of some interest. The orbital period for a low-Earth

orbiter is about 90 minutes (5400 sec). This means, on average, a pass every 1800 sec or so, with a three-complex network, assuming every pass is supported as in the assumption here. If the spacecraft transmits while also acquiring and recording more data, then 1800 sec, say 2000 sec, of data must be stored. If the storage on board is limited to 10 gigabits, today's technology, a maximum spacecraft data acquisition rate of 5 megabits per second can then be supported. Future terabit onboard storage could support 100 times this data acquisition rate, or 500 megabits/sec downlink. The link, not the storage, would then become the limiting factor. The storage and dump mode is clearly feasible for small low-Earth orbiters from the standpoint of data storage technology.

Now that the average visibility is known, one can find the link improvements, using the method given in the next section.

IV. Average Link Improvement

The average link improvement is readily calculated from the average inverse squared distance to the Earth tracking station from the low-Earth orbiter. Recall that it is being assumed that the distance to a space relay station is essentially constant at 36,000 km. Thus, the decibel improvement due to the distance effect is correctly calculated in terms of data volume received as the product of 36,000 km squared times the average inverse squared distance. This gain must be de-rated by the visibility factor, i.e., the fraction of time that the low-Earth orbiter is visible to the ground tracking network. The resulting product is the average link improvement taking both communication distance and visibility into account.

The calculation to follow assumes that the data rate is varied by an onboard stored sequence during a pass as the spacecraft distance to the Earth station changes. The goal is to make the data rate as large as possible at all times. This is similar to deep space support, where the rate can be varied during a pass, but with more frequent changes during a near-Earth pass. Rate changing results in a maximum to minimum data rate in an overhead 400-km orbit of $(2298/400)^2 = 33.0$ or 15.2 dB. If the minimum data rate is 2 megabits/sec, the maximum rate when the spacecraft is directly overhead would be 66 megabits/sec. So the high maximum–minimum ratio is not an insurmountable problem, except perhaps at the highest average data acquisition rates.

The calculation shall proceed by calculating the average improvement factor F relative to distance squared at spacecraft rise or set:

$$F = [(r+h)^2 - r^2] \times \text{average inverse squared distance}$$

This is used for ease of correctness testing later. Referring to Fig. 2, it is seen, as in the previous section, that with

$$\theta = \arccos\left(r/(r+h)\right)$$

as before, and normalizing over the area of the cap,

$$F = h(2r+h)\int_{\phi=0}^{\theta} \frac{2\pi r^2 \sin\phi d\phi / \left(2\pi r^2 h/(r+h)\right)}{l^2} \quad (1)$$

Here l is the distance from the spacecraft to what is now thought of as the randomly situated tracking station, found from a vector diagram (Fig. 3) to be

$$l^2 = (r+h)^2 + r^2 - 2r(r+h)\cos\phi \quad (2)$$

$$l^2 = 2r^2 + 2rh + r^2 - 2r(r+h)\cos\phi \quad (3)$$

Substituting Eq. (2) into Eq. (1) and using $\sin\phi d\phi = -d\cos\phi$, Eq. (1) becomes

$$F = (2r+h)(r+h)\int_{\phi=0}^{\theta} \frac{-d\cos\phi}{2r^2 + 2rh + h^2 - 2r(r+h)\cos\phi} \quad (4)$$

Using $\cos\theta = r/(r+h)$, Eq. (4) becomes, with $u = \cos\phi$,

$$F = (2r+h)(r+h)\int_{u=\frac{r}{(r+h)}}^{1} \frac{du}{2r^2 + 2rh + h^2 - 2r(r+h)u} \quad (5)$$

and so

$$F = \frac{(2r+h)(r+h)}{2r^2+2rh+h^2} \times \int_{u=\frac{r}{(r+h)}}^{1} \frac{du}{1 - 2r(r+h)u/\left(2r^2+2rh+h^2\right)} \quad (6)$$

Make the substitution

$$v = 2r(r+h)u/\left(2r^2+2rh+h^2\right) \tag{7}$$

so that Eq. (6) becomes

$$F = \frac{(2r+h)(r+h)}{(2r^2+2rh+h^2)} \frac{(2r^2+2rh+h^2)}{2r(r+h)}$$

$$\times \int_{v=2r^2/(2r^2+2rh+h^2)}^{2r(r+h)/(2r^2+2rh+h^2)} \frac{dv}{1-v} \tag{8}$$

Using

$$\int \frac{dv}{1-v} = -\ln(1-v) \tag{9}$$

Eq. (8) becomes

$$F = \left(1+\frac{h}{2r}\right) \ln(1-v)\Big|_{v=2r(r+h)/(2r^2+2rh+h^2)}^{2r^2/(2r^2+2rh+h^2)} \tag{10}$$

or

$$F = \left(1+\frac{h}{2r}\right) \ln\left(\frac{2rh+h^2}{h^2}\right) \tag{11}$$

or

$$F = \left(1+\frac{h}{2r}\right) \ln\left(1+\frac{2r}{h}\right) \tag{12}$$

This is the desired result on the average improvement factor F.

Several checks are in order. First F is always greater than 1, since

$$\left(1+\frac{1}{x}\right) \ln(1+x) > 1 \quad \text{for } x > 0$$

i.e., $(1+x)\ln(1+x) > x$. This is because the derivative of the left-hand side, $1+\ln(1+x)$, is greater than 1, the derivative of the right-hand side, if $x > 0$. Second, it is clear from the geometry that F ought to be decreasing in h, i.e., less improvement for higher orbits. This amounts to $(1+1/x)\ln(1+x)$ increasing in x as x increases from 0. The derivative of this with respect to x is

$$\frac{x - \ln(x+1)}{x^2}$$

whose numerator is obviously and well-known to be positive for $x > 0$, as required.

Another check is that F clearly approaches 1 from above as h approaches ∞, from the geometry. This easily follows algebraically from Eq. (12):

$$F \doteq \left(1+\frac{h}{2r}\right)\left(\frac{2r}{h}\right) \text{ as } h \to \infty$$

$$F \doteq 1+\frac{2r}{h} \downarrow 1 \text{ as } h \to \infty$$

It is interesting to note that, as $h \to 0$, the improvement F becomes infinite as $\ln(2r/h)$.

For orbits conditioned to pass directly overhead, it is clear that the improvement factor is even greater than that given by Eq. (12). For example, for 400-km orbits conditioned to go overhead, the improvement factor is 10 dB instead of the 5.6 dB to be derived below. This aspect will not be pursued.

V. Summary

For 400-km orbits, the average improvement factor F is found from Eq. (13) to be

$$F = \frac{33}{32} \ln 33 = 3.605 = 5.6 \text{ dB}$$

The distance at rise or set is 2298 km, so the gain relative to a 36,000-km path, ignoring temporarily the visibility loss, is

$$\left(\frac{36000}{2298}\right)^2 F = \left(\frac{36000}{2298}\right)^2 (3.605) = 884.7 = 29.5 \text{ dB}$$

This is to be de-rated by the visibility loss of Section III, 0.088-dB or 10.5-dB loss, to get the average gain of 19.0 dB reported in Section I.

To recapitulate, the average link gain factor G for Earth reception over full-time synchronous space reception, considering distances, variable data rates, and visibilities, is given for random circular orbits and a three-station ground network by

$$G = \frac{d^2}{h(2r+h)} \left(1 + \frac{h}{2r}\right) \ln \left(1 + \frac{2r}{h}\right) \frac{3h}{2(r+h)} \quad (13)$$

using Eq. (12) and the equation for the visibility V from a single Earth station derived in Section II. The factor of 3 occurs because expected values add together, and no two DSN stations ever see the same low-Earth orbiter at the same time. Equation (13) simplifies to

$$G = \frac{3}{4} \frac{d^2}{r(r+h)} \ln \left(1 + \frac{2r}{h}\right) \quad (14)$$

Here d is synchronous altitude, 36,000 km, r is the radius of the Earth, 6400 km, and h is the height of the circular orbit in kilometers. This G is easily seen to be decreasing as h increases, as is reasonable.

In summary, small Earth orbiters will more and more find it advantageous to be supported by the ground network rather than by the space network. The advantage of lowered communications package mass, size, and power consumption, resulting from smaller communication distances capitalized upon by available onboard data storage, is amplified by the greater figure of merit (gain divided by noise temperature) of ground stations compared to space relay stations. Overall, NASA costs are expected to be lower as well, but this article does not attempt a formal trade-off study. Nevertheless, the DSN can expect to be supporting more and more low-Earth orbiters.

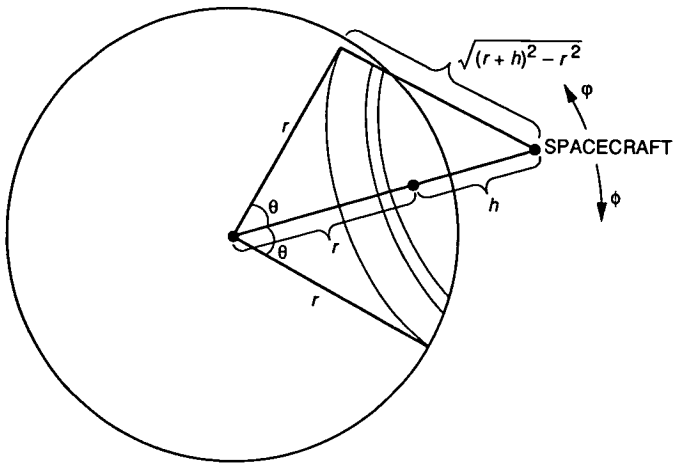


Fig. 1. The geometry of visibility.

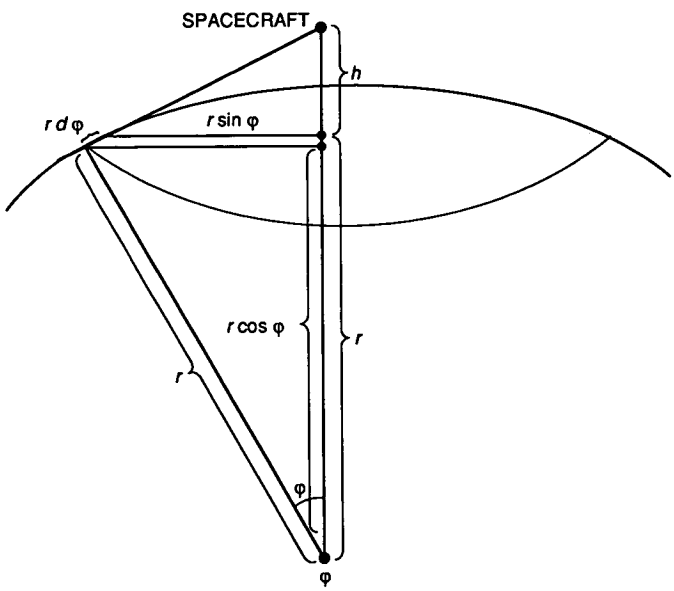


Fig. 2. The average inverse squared distance.

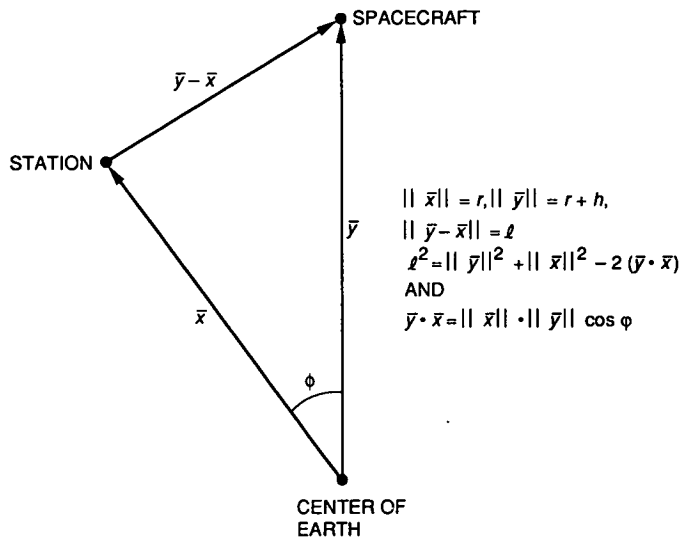


Fig. 3. The distance between the spacecraft and the tracking station.

An Analysis of I/O Efficient Order-Statistic-Based Techniques for Noise Power Estimation in the HRMS Sky Survey's Operational System

G. A. Zimmerman
Communications Systems Research Section

E. T. Olsen
Space Physics and Astrophysics Section

Noise power estimation in the High-Resolution Microwave Survey (HRMS) sky survey element is considered as an example of a constant false alarm rate (CFAR) signal detection problem. Order-statistic-based noise power estimators for CFAR detection are considered in terms of required estimator accuracy and estimator dynamic range. By limiting the dynamic range of the value to be estimated, the performance of an order-statistic estimator can be achieved by simpler techniques requiring only a single pass of the data. Simple threshold-and-count techniques are examined, and it is shown how several parallel threshold-and-count estimation devices can be used to expand the dynamic range to meet HRMS system requirements with minimal hardware complexity. An input/output (I/O) efficient limited-precision order-statistic estimator with wide but limited dynamic range is also examined.

I. Introduction

The purpose of this article is to examine I/O efficient methods for dynamically estimating noise power in the High Resolution Microwave Survey (HRMS) sky survey element. One I/O efficient estimation method was suggested during the development of the HRMS sky survey prototype system (SSPS) signal-processing subsystem by Dr. Bernard Oliver at NASA Ames Research Center.[1] At that time this technique, now called the "threshold-and-count" estimator, was labeled the "truncated data" method. It was examined during the SSPS development;[2,3,4] however, several relevant system param-

[1] B. Oliver, private communication, Deputy Chief of HRMS Office, NASA Ames Research Center, Moffett Field, California, summer 1991.

[2] R. Brady, "An Alternative to Using Order Statistics for Determining the Mean Noise Power Estimate in the EDM," Interoffice Memorandum, Jet Propulsion Laboratory, Pasadena, California, April 14, 1988.

[3] W. Deich, "Truncated Data and Background Estimation," Interoffice Memorandum, Jet Propulsion Laboratory, Pasadena, California, May 3, 1988.

[4] M. F. Garyantes, "Search For Extraterrestrial Intelligence Microwave Observing Project Sky Survey Element Subsystem Functional Requirements and Design BECAT 1 Processor (Prototype)," JPL D-7116, (internal document), Jet Propulsion Laboratory, Pasadena, California, December 1, 1989.

eters have since changed, and in an effort to save special-purpose hardware complexity and cost, a new analysis of this technique was undertaken. This article describes the analysis and results.

HRMS sky survey signal detection will operate on the output of a real-time, 640-MHz, 32-megachannel polyphase discrete Fourier transform (DFT) spectrum analyzer [1]. In normal operation, the spectrum analyzer will accumulate the power of from 2 to 10 spectra for input to the signal-detection assembly. The signal-detection assembly applies gain normalization and a five-tap finite impulse response (FIR) matched filter to each accumulated spectral channel. The matched filter outputs are then thresholded at a probability of false alarm, P_{FA}, of approximately 10^{-5} to reduce the data rate for data input to a general-purpose computer. Using adjacent scans across the sky, the computer excises interference and combines data. After the excision of interference, the data are passed through a second threshold at a much lower P_{FA}, and a relatively small number of sky positions passing the threshold are re-observed. Because of the large number of channels and limited re-observation time, the false alarm rate due to noise must be tightly controlled. False-alarm-rate control is accomplished by normalizing the accumulated power spectra with estimates of the noise power as a function of frequency.

Dynamic estimation of the noise environment is required for many signal-processing algorithms. Noise power estimates obtained from order statistics are robust in the presence of interference, and they can be adapted to variations in the statistical distribution of the noise samples [2,3,4,5]. In general, computation of a fixed order statistic requires that a sample population be sorted. Sorting techniques require multiple passes on a data set and can be I/O intensive. The advantages of the order-statistic estimation approach can be obtained by fixing the order statistic's value rather than fixing the rank of the order statistic, provided that the value remains within certain bounds. This article presents an analysis of the estimator requirements and introduces an additional performance requirement called the estimator dynamic range. A simple order-statistic-based estimator is analyzed, and simple modifications to extend the limited dynamic range of the estimator are presented. An efficient estimator with a wide dynamic range is formed by combining this simple estimator with a fixed-order-statistic technique. The requirements for noise power estimation and the solutions are discussed in the context of NASA's HRMS sky survey, although the approach is applicable to other constant false-alarm-rate (CFAR) detection applications.

II. Summary of Requirements

There are two performance requirements on the estimator. The first requirement is the obvious one of estimate accuracy. For interference immunity, methods under consideration in this article use an order statistic to estimate the noise power. As a result, in the limit of large population sizes (populations >1000), the error in the estimate will be normally distributed and zero mean [4]. When the estimate is used to control detection thresholds, as in HRMS, errors in noise power estimates change the P_{FA} for the detection algorithms. By specifying the allowable variation in the P_{FA}, the required estimator accuracy can be determined. A new requirement considered here is the estimator dynamic range. This is defined as the variation in noise power over which the estimator will perform within the specified error. For example, if the estimator can meet the accuracy requirement as the noise power varies over a factor-of-two range, the estimator has a dynamic range of 3 dB. This requirement depends on the variability in the detection environment and the underlying probability distribution. Thus, determination of the dynamic range requirement is dependent both on the observations to be performed and on the estimate accuracy requirement. Dynamic range is discussed following the discussion of estimate accuracy.

A. Estimate Accuracy

The noise power estimator in the HRMS system is used to control the data rate by setting detection thresholds. It is not intended to be used for radio-astronomy-continuum-measurement calibration. As a result, errors in the estimate will only change the data rate out of the sky survey signal-processing hardware, and will ultimately change the P_{FA} for the HRMS detection algorithms. The effect of errors in the noise power estimate, $\hat{T}$, on the data rate at the output of the hardware has been previously addressed.[5] However, because an error in $\hat{T}$ changes where the data rate–reducing threshold is located relative to the distribution function of the noise, the sensitivity of the data rate to estimate error is dependent on the position of the threshold. At the time of W. Deich's memorandum,[5] the hardware threshold was to be positioned to provide a P_{FA} of 10^{-3}; however, more recent analysis[6] has indicated that this threshold should be set at a P_{FA} ranging from 10^{-5} to 10^{-6}. Therefore, a new look at the accuracy requirement is needed.

---

[5] W. Deich, "Baseline Ripple, Estimation Error in $\hat{T}$, and False Alarms," Interoffice Memorandum, Jet Propulsion Laboratory, Pasadena, California, February 14, 1989.

[6] S. Levin, personal communication, Member of Technical Staff, Space Physics and Astrophysics Section, Jet Propulsion Laboratory, Pasadena, California, spring 1991.

When the data can come from more than one statistical distribution, it is useful to identify the limiting distribution as a case for analysis. Because the number of spectra per accumulated power spectral output is selectable in HRMS sky survey operations, the power spectral samples for white Gaussian noise input will come from χ^2_ν distributions with varying degrees of freedom (ν). For χ^2_ν data, the limiting case to consider in the analysis of the effect of estimate error on P_{FA} is that of the maximum number of degrees of freedom expected. Normal-rate operations in the HRMS sky survey require the greatest control over the P_{FA}, and these will have a maximum of 10 spectra per accumulated power spectral output.

At 10 spectra per accumulated output, each spectrum will consist of samples from the χ^2_{20} distribution. The value to be thresholded is a weighted sum of five of these samples weighted approximately by $\{0.64, 0.89, 1.00, 0.89, 0.64\}$ in a five-tap FIR filter matched to the antenna beam passing over a point source. The weighted sum can be approximated as samples of the χ^2_{81} distribution. For such a large number of degrees of freedom, it is appropriate to use the following approximation [Eq. (1)] for the probability that a sample is greater than a value x (see [6], Eq. 26.4.14).

$$\begin{aligned} \text{Prob}[\chi^2_\nu > x] &= Q(x|\nu) \approx Q(x_2) \\ x_2 &= \frac{(x/\nu)^{1/3} - (1 - 2/(9\nu))}{\sqrt{2/(9\nu)}} \end{aligned} \quad (1)$$

Note that in the region of interest where $P_{FA} \leq 10^{-5}$, the following approximation [Eq. (2)] is valid:

$$Q(x_2) \approx \frac{e^{-x_2^2/2}}{x_2\sqrt{2\pi}} \quad (2)$$

Let the estimated threshold be related to the ideal threshold by $\hat{T} = (1+\alpha)T$; i.e., the fractional error is α. The probability of false alarm, given a fractional estimate error α, is:

$$\begin{aligned} \text{Prob}[\chi^2_{81} > (1+\alpha)T] &\approx Q(\hat{x}_2) \approx \frac{e^{-\hat{x}_2{}^2/2}}{\hat{x}_2\sqrt{2\pi}} \\ \hat{x}_2 &\approx 19.092[(1+\alpha)^{1/3}(T/81)^{1/3} \\ &\quad + 0.99726] \end{aligned} \quad (3)$$

Figure 1 shows the effect of errors in the estimate on the probability of false alarm for 10 spectra per accumulation.

It is important to note that the noise power estimates are obtained on a per-spectrum basis, and it is the effective combined error of the five estimates that must meet the accuracy requirement. The errors in the methods under consideration are normally distributed, zero mean, and independent for each spectrum. It can be shown that the combined fractional error is approximately normally distributed, zero mean, with standard deviation

$$\sigma\sqrt{\frac{0.64^2 + 0.89^2 + 1 + 0.89^2 + 0.64^2}{0.64 + 0.89 + 1 + 0.89 + 0.64}} = 0.454\sigma \quad (4)$$

where σ is the standard error of an individual estimate. This results in a reduction of the fractional error by 54.6 percent. For example, a 1-percent error in the final effective estimate corresponds to an error of 2.2 percent in the noise power estimates for the individual spectra.

Conservative estimates of false alarms due to interference suggest that the data rate out of the special-purpose hardware will be a factor of 10 greater than the target P_{FA} of 10^{-5}. This implies that maintaining the data rate out of the special-purpose hardware requires that the noise power estimate be effectively accurate to within 6 percent. If such an error were specified as a 3-standard-error event, the effective estimator accuracy requirement for one standard error would be only 2 percent.

The significantly smaller P_{FA} required to select candidates for re-observation would tend to imply a stricter requirement on the noise power than that driven by the hardware data rate. Additional candidates translate directly into additional re-observation time, and as a result, it is unlikely to think that any more than a factor-of-2 variation would be tolerable. A 3-standard-error event corresponding to a factor-of-2 change would imply an effective standard error for the five estimates of about 0.37 percent. This corresponds to a standard error of 0.82 percent in the individual estimates. It is important to note, however, that determination of the P_{FA} for re-observation candidate selection uses many individual noise power estimates. Improvement of the noise power estimates or rejection of obviously bad data may be possible at this stage. Taking this possibility into account, and considering that a 3-standard-error event occurs with approximately 10^{-3} probability, the required error in the noise power estimate is taken to be less than 1 percent.[7]

[7] The error in the false alarm rate, defined as the difference between the target and the actual false alarm rate, will not be zero mean and will not, in general, be symmetrically distributed about the mean. However, this can be easily compensated for.

B. Dynamic Range of the Estimator

Once an estimate accuracy requirement has been defined, a dynamic range within which the estimator meets its accuracy requirement can be determined. In almost all estimation applications, the dynamic range of the estimated parameter can be limited. It is shown below that by limiting the dynamic range, one can construct simple, efficient order-statistic-based estimators.

A noise power estimator must be flexible enough to cover natural variations in received noise power. Unnatural variations that would result in unusable data, e.g., significant drops in receiver gain, need not be within the performance limits of the estimator. For HRMS or radio astronomy purposes, natural variations are mainly caused by four sources: changes in air mass as a scan changes elevation angle, increased water vapor as the antenna beam passes through a cloud or water vapor bubble in the atmosphere, the contribution from galactic or extragalactic radio sources at $\lambda = 21$ cm due to Doppler-broadened atomic hydrogen hyperfine emission, and the contributions from common strong astrophysical radio sources. For all of these sources, increases in system temperature will be a function of frequency. Compensating for the variations due to the passage of the beam across rare strong astrophysical sources is not considered a requirement, but is a goal.

Since the underlying statistical distribution affects the dynamic range of the estimator, the statistics under which the dynamic range must be met also need to be defined. The HRMS sky survey will vary the number of spectra per accumulation inversely as the sky frequency of the observations changes from 1 to 10 GHz. The nominal range is from 10 to 2 spectra per accumulated power spectral output. To accommodate slower scan rates, the nominal accumulation may be increased by a factor of 10. For this application, the system should also be capable of allowing the number of spectra per accumulation at 1 GHz to vary from 10 to 100 and the number of spectra per accumulation at 10 GHz to vary from 2 to 20.

The observing environment ultimately defines the required dynamic range. Conditions at Canberra, Australia, will be worse than those at Goldstone, California.[8] The HRMS system will have a nominal (i.e., cold, dry sky at zenith) system temperature of 25 K. The atmospheric contribution to the system temperature scales as the square of the frequency.

[8] Based on the models for the atmospheric contribution to system temperature in *DSN Standard Flight Project Interface Design Project Handbook* 810-5 (internal document), Revision D, Jet Propulsion Laboratory, Pasadena, California, September 15, 1991.

Dynamic range requirements for from 20 to 40, 40 to 60, 60 to 80, and 80 to 100 spectra per accumulation can be derived by using the translations from number of spectra per accumulation to RF center frequency that are given by Olsen,[9] and the planned slowdown factor of 10. Dynamic range calculations due to atmosphere alone are given in Table 1, and consideration of strong astrophysical sources is presented in Table 2. As shown in Table 3, the dynamic range requirement is for 7.1 dB of dynamic range for from 2 to 20 spectra per accumulated output, with a goal of 9.0 dB, and 3.8 dB of dynamic range for from 20 to 100 spectra, with a goal of from 4.2 to 4.8 dB.

III. Description of the Threshold-and-Count Estimator

The threshold-and-count estimator consists of a threshold applied to a population and a count of the number of points in the population that do not exceed the threshold.[10] Modifications may be made to this method to dynamically adjust the threshold, if necessary, or to apply multiple thresholds and pick the best one, but the basic method remains the same. Provided that the population is sufficiently dense around the threshold, this method is equivalent to choosing an order statistic for the noise power estimate.[11] However, the value of the order statistic is now fixed, as it is the threshold value, and the rank of the order statistic is the random variable. The result is an estimate of the value of the cumulative distribution function at the threshold, i.e., the probability that a sample is less than the threshold value. The noise power estimate can then be obtained from a look-up table, interpolating from the cumulative distribution function to recover the mean noise power of the population. A block diagram of the threshold-and-count estimator is shown in Fig. 2.

IV. Estimator Performance

As expected, and verified by simulation,[12] the performance of the threshold-and-count estimator is equivalent to choosing a true order statistic, given a sufficiently dense population of points around the threshold. Simulations

[9] E. T. Olsen, "Time Required to Complete the All Sky Survey Campaign," Interoffice Memorandum 1720-6025-3280, Jet Propulsion Laboratory, Pasadena, California, February 27, 1991.

[10] B. Oliver, personal communication, Deputy Chief of HRMS Office, NASA Ames Research Center, Moffett Field, California, summer 1991.

[11] R. Brady, op. cit.

[12] Ibid.

demonstrated[13] that 1 percent of 8192, or 82 points below the threshold, was a sufficient number for the variance of the order statistic to mask the threshold error due to finite sample density around the threshold. In the limit of large population size, order statistics are normally distributed, with a mean and variance for the rth smallest order statistic

$$E[x_{(r)}] = \xi = F^{-1}(r/N) \quad \mathrm{Var}[x_{(r)}] = \frac{r/N\,(1-\frac{r}{N})}{N f(\xi)^2} \tag{5}$$

where $F()$ is the cumulative distribution function of the population, $f()$ is the population's probability density function, and N is the number of samples in the population. For accumulated power spectra of white Gaussian input voltage samples, the populations will be drawn from the χ^2_ν distribution with an even number of degrees of freedom with ν = 2×the number of spectra per accumulated output ($\nu = 2n$), with the probability density function

$$f(\xi) = \frac{\xi^{n-1} e^{-\xi/T_s}}{T_s^n\,(n-1)!} \tag{6}$$

Since the mean value of the order statistic is proportional to the mean noise power, and

$$E\left[\frac{r}{N}\right] = \mathrm{Prob}\left[\chi^2_{2n} \le \xi\right] = 1 - e^{-\beta} \sum_{i=0}^{n-1} \frac{\beta^i}{i!} \tag{7}$$

the fractional standard error is:

$$\frac{\sigma}{\xi} = \sqrt{\frac{r/N\,(1-r/N)}{N}}\,\frac{T_s^n e^{\xi/T_s}(n-1)!}{\xi^n}$$

$$= \sqrt{\frac{r/N\,(1-r/N)}{N}}\,\frac{e^{\beta}(n-1)!}{\beta^n} \quad ; \beta = \frac{\xi}{T_s} \tag{8}$$

In a true order-statistic estimate, the operating point, r/N, is fixed. As a result, establishment of a given error at a chosen operating point requires a minimum population size, $N_{\min}$:

$$N_{\min} = \left(\frac{r}{N}\right)\left(1-\frac{r}{N}\right)\left[\frac{e^{\beta}(n-1)!}{\sigma\beta^n}\right]^2 \tag{9}$$

[13] Ibid.

Table 4 gives $N_{\min}$ for some likely operating points. Order-statistic-based methods similarly require populations larger than the $N_{\min}$ at their operating points. Table 4 shows that an order-statistic-based approach can achieve a 1-percent standard error with a population of 16K (16,384) samples, a 0.5-percent standard error with 64K (65,536) samples, and a 0.33-percent standard error with 128K (131,072) samples. Since the threshold-and-count estimator is equivalent to a true order-statistic estimate at the given values of r/N, it achieves the same standard errors with the same population sizes.

Because r/N remains fixed in a true order-statistic algorithm, it will always produce the same standard error, provided that the statistics of the data do not change. Thus, if it can operate within the estimate accuracy requirement, its dynamic range will only be limited by the dynamic range of the numerical representation. Approximate order-statistic methods, like the threshold-and-count method, have more limited dynamic range. The dynamic range of a single, fixed-threshold threshold-and-count estimator is examined first below. Modifications may be made to the single, fixed-threshold estimator to increase its dynamic range, such as automatically adjusting the threshold over time to track slow variations, or using multiple thresholds based on different a priori values of the noise power, but the performance of these modifications can be analyzed easily through the single fixed-threshold case.

Consider the factor β in the probability density and distribution functions. The factor β reflects the relative position of the threshold within the probability distribution of the samples. Increasing the noise power by a factor α is equivalent to dividing β by α. When the noise temperature changes, the new value of r/N must be calculated using the cumulative distribution function with the new value of β. The accuracy of the estimator can then be computed as before. Figure 3 shows the standard error as a function of system temperature change.

The shape of the probability density function for the χ^2 distribution and the quadratic form in r/N in the expression for the estimate variance guarantee that if the population size is greater than $N_{\min}$, the estimator will perform within the accuracy requirement within an isolated, continuous range of noise power values. The ratio of the maximum noise power value at which the estimator accuracy is within specification to the minimum such value is the estimator's dynamic range.

The single, fixed-threshold threshold-and-count estimator was evaluated with 64K samples for the dynamic range defined by a standard error of less than 1 percent. Table 5

shows that the dynamic range of the single, fixed-threshold threshold-and-count estimator is limited by the behavior of the estimator at high numbers of spectra per accumulation. As the number of degrees of freedom, ν, increases, the χ^2_ν distribution of the samples approaches the Gaussian distribution, with the mean approaching ν, and the standard deviation approaching $\sqrt{\nu}$. This indicates that the probability density is being concentrated about the mean, collapsing the range over which it has non-negligible values. As a result, the threshold-and-count estimator becomes more sensitive to changes in the noise power. Hence, the dynamic range of the threshold-and-count estimator is limited by the performance with the maximum specified number of degrees of freedom.

It can be observed from Table 5 that a 1-percent error usually corresponds to the fairly loose requirement of the threshold, falling between the 0.1 and the 99.9 percentage points of the distribution. However, in order to preserve the assumption of a dense population of points near the threshold, required for the equivalence of the threshold-and-count method to a true order-statistic noise power estimator, the operating range will be limited to the 1-percent to the 99-percent points of the distribution.

It has been asserted,[14] perhaps incorrectly, that as the upper limit of the range increases past the median of the distribution, the interference robustness of the algorithm is lost. This is based on a generalization of the effect of interference on high-level order statistics in a true order-statistic algorithm. Since the threshold is fixed in the threshold-and-count algorithm, as opposed to the threshold in a true order-statistic technique, depletion of the population due to interference can always be corrected for. This is different from using a high-level order statistic, e.g., the 99-percent point, where high levels of interference might give an erroneously high value for the order statistic. With a fixed threshold, high levels of interference do not change the threshold. The interference reduces the effective population size, the same effect that it has on order statistics from below the median of the distribution [4]. In both a true order statistic and the threshold-and-count methods, this reduction of the population biases the estimate. Correction for this bias can easily be performed in the threshold-and-count estimator by adjusting the contents of the lookup table.

Therefore, consider the dynamic range of the threshold-and-count estimator to be the ratio of the 99-percent to the 1-percent point of the distribution, guaranteeing both a dense population around the threshold and a standard error less than 1 percent for all but 2 or 4 spectra per accumulation. For 2 or 4 spectra per accumulation, a 95- to 5-percent range achieves the required accuracy, and demonstrates dynamic ranges still in excess of the 99- to 1-percent ranges achieved when there are more than 8 spectra per accumulation. Standard errors at the 1-, 5-, 50-, 95-, and 99-percent points of the distributions are shown in Table 6. The dynamic range for the 1- to 99-percent points is shown in Table 7, as well as the dynamic range defined by the more conservative 5- to 95-percent and 1- to 50-percent levels.

None of the estimators will cover the 7.1-dB dynamic range required at 20 spectra per accumulated output with 1 percent of the points remaining above or below the threshold. This problem is not insurmountable. One previously explored solution [15,16] is to adapt the threshold to track a slowly time-varying noise power. This assumes that the noise power will not vary outside the dynamic range of the estimator from accumulation to accumulation. This assumption can be difficult to maintain at high numbers of spectra per accumulation, suggesting adaptation of the threshold during the accumulation. Such a solution would have high dynamic range but might be difficult to implement, since the number of degrees of freedom in the data increases with each spectrum added to the accumulation. Such a scheme would require a look-up table to invert several different cumulative distribution functions, at least until the distribution becomes sufficiently Gaussian. Evaluation of adaptation methods requires knowledge of the dynamics of the noise power, which could be obtained from field experiments with the SSPS. Time-varying thresholds will not be considered further here.

Another, simpler modification to increase the dynamic range of the estimator would be to have multiple thresholds arranged so that one would always be within the estimator's dynamic range. An estimator with a dynamic range of γ dB can be extended to cover a dynamic range of μ dB by replicating it μ/γ times, placing the operating points such that the lower edge of one threshold's range was at the upper edge of the previous threshold's range. Table 7 shows that only two such combined thresholds produce a dynamic range of 9.15 dB at 20 spectra per accumulated output and a dynamic range of 4.0 dB at 100 spectra, exceeding the dynamic range requirements and meeting the dynamic range goals at 20 spectra. The dynamic range

[14] Ibid.

[15] Ibid.

[16] W. Deich, "Truncated Data and Background Estimator," Interoffice Memorandum, Jet Propulsion Laboratory, Pasadena, California, May 3, 1988.

requirements at all numbers of spectra per accumulation up to 100 can be met with from 2 to 4 threshold estimators, depending on how conservative one is in choosing the percentage points defining the dynamic range. All the dynamic range goals can be met with from 4 to 6 estimators, with the number again depending on how conservatively one chooses the percentage points. Since the threshold does not need to be settable to a high precision, e.g., 1 percent is sufficient, a floating-point threshold comparison can be performed with fewer than 16 bits compared. Given the simplicity of the algorithm, 6 threshold units is not an unwieldy number, and the additional threshold counts could be used as a check on how well the data fit the assumed distribution. Furthermore, all the threshold-and-count units operate on the same input data, making this architecture an ideal candidate for implementation as integrated logic, e.g., a field-programmable gate array or application-specific integrated circuit (ASIC). The drawback of this approach is that the dynamic range of the noise power estimator is limited at design time by the number of thresholds built into special-purpose hardware. This can limit flexibility for possible long accumulation studies.

V. A Hybrid Threshold-Order-Statistic Estimator

By combining a threshold with the current order-statistic technique, it is possible to produce an estimator that is significantly more hardware-efficient than the two-pass order-statistic estimator in the SSPS, yet with a much wider dynamic range than the threshold method alone. The key to this method is to use an a priori threshold to determine a range of numeric values to histogram to obtain an order statistic of a desired rank.

First, consider that the required estimate accuracy is 1 percent. One should then make the fractional numeric precision of the estimate small relative to the required estimate accuracy. In floating-point arithmetic, the fractional numeric precision defines the number of bits required to represent the mantissa. A precision of ± 0.1 percent will be achieved with 9 bits of mantissa. Hence, if one were to perform a one-pass histogram, one would require 9 bits of mantissa plus 8 bits of exponent for a total of 17 bits, or 128K histogram bin values. An efficient hardware implementation used in the SSPS requires that the population size be at least twice the histogram length, and hence, the current approach could not support an expansion to 17 bits with a population size of 64K points. As a result, a one-pass reduced-precision histogram, while attractive, would either require a more complex hardware implementation, or a significantly larger minimum population size of 256K points.

If only 3 bits of exponent were supported in addition to the 9 bits of mantissa required for 0.1-percent precision, a dynamic range of almost a factor of 2^8, or 24 dB, would be obtained. This far exceeds the dynamic range requirements for the estimator, and, if desired, bits of exponent can be traded for greater precision. Two bits of exponent give a dynamic range of a factor of almost 16, or 12 dB, while 1 bit of exponent reduces the dynamic range to a factor of almost 4, or 6 dB. Two bits of exponent meet all the HRMS requirements and goals. Such a hybrid noise power estimator would consist of the following: (1) a threshold value determined by the most significant 6 exponent bits of the desired order statistic, (2) a comparator that determines if the 6 most significant exponent bits of a data point are less than, greater than, or equal to the threshold, (3) a counter for the data points less than the threshold, and (4) an 11-bit histogram unit, as in the SSPS, for data points with the most significant 6 bits of exponent equal to the threshold. The desired order statistic is then obtained by adding the count below the threshold to the ascending values in the histogram until the desired level is reached, producing an order statistic of the desired rank. This procedure is of approximately the same complexity as that implemented in the SSPS on a single 14- by 14-in. wire-wrap board.

VI. Conclusions

At the cost of limited dynamic range, estimation equivalent to sampling fixed order statistics can be performed with a single pass of the data. The requirements and goals for the HRMS sky survey's signal detection are met with a noise-power-estimator accuracy requirement of 1 percent and a dynamic range of up to 7.1 dB for most observations.

Under normal operating conditions, a fixed-threshold threshold-and-count estimator is equivalent to a true order-statistic estimator in accuracy, with a limited dynamic range. If a population size of 64K is used, the estimator will have better than 1-percent accuracy, but less than the required 7.1 dB dynamic range at 20 spectra per accumulation. Multiple thresholds may be used to increase the dynamic range. The HRMS requirements are met with from 2 to 4 thresholds, and the goals are met with from 4 to 6 thresholds. This technique provides a single-pass noise power estimator, as opposed to the current two-pass order-statistic technique, at the cost of a fixed dynamic range. In summary, an acceptable noise power estimator

could be constructed with the threshold-and-count technique, and this estimator might occupy less than half of a single board in the sky survey operational system. In contrast, the current true order-statistic estimator implementation would occupy from 2 to 3 boards.

A combination of the threshold-and-count estimator and the current implementation of the true order-statistic estimator also produces a single-pass, single-board noise power estimator for floating-point systems with superior dynamic range. This new, hybrid threshold-histogram estimator would produce an order statistic of a desired rank, provided that the desired order statistic is within its dynamic range. This new, hybrid estimator would require only one threshold, would be based on the design in the SSPS, and while slightly more complicated than the raw threshold-and-count, offers the advantage of constant estimate accuracy.

Based on this analysis, it is recommended that the following research activities be performed with the SSPS, currently deployed at the Goldstone Deep Space Communications Complex, California: (1) monitoring the dynamic range of the noise power estimates to confirm the requirements derived in this article, (2) testing the interference robustness of the true order-statistic estimator in the SSPS, (3) testing the performance and interference robustness of the threshold-and-count estimator, and (4) analyzing the estimator bias induced by interference in all the candidate estimators, and if necessary, (5) analyzing and testing an algorithm for automatically correcting significant estimator biases.

References

[1] G. Zimmerman, M. Garyantes, and M. Grimm, "A 640 MHz, 32 Megachannel Real-Time Polyphase FFT Spectrum Analyzer," *Conference Record of the 25th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, pp. 106–110, November 1991.

[2] K. Wong and C. Shuang, "Detection of Narrow-Band Sonar Signals Using Order Statistical Filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35, pp. 597–613, May 1987.

[3] E. Satorius, R. Brady, W. Deich, S. Gulkis, and E. Olsen, "On the Performance of an Order Statistic Threshold Strategy for Detection of Narrowband Signals," *Conference Record: Papers Presented at the 21st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, pp. 793–796, November 1987.

[4] E. Satorius and R. Brady, "SETI Signal Processing," *Conference Record: Papers Presented at the 22nd Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, pp. 194–198, November 1988.

[5] R. Rifkin, "Performance Analysis of an Order Statistic-Based Radar Threshold Normalizer in Weibull Noise," *Conference Record of the 25th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, pp. 830–834, November 1991.

[6] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards Applied Mathematics Series, tenth printing, Washington, D.C.: U.S. Government Printing Office, December 1972.

Table 1. Dynamic range constraints due to atmosphere, calculated by using the Slobin model, a 99-percent cumulative, 20-deg elevation, 25-K system, for the Canberra Deep Space Communications Complex.

RF, GHz	Maximum system temperature, K	Dynamic range needed, dB
1–10	128	7.1
1–3.8	49	2.9
1–2.25	38	1.8
1–1.63	35	1.5
1–1.32	34	1.3

Table 2. High-temperature radio sources considered in conjunction with atmosphere.

Atmosphere 20-deg elevation, K	Source temperature, K	Dynamic range needed, dB	Source and frequency
35	10	2.6	Galactic plane background (λ 21 cm, 1.4 GHz)
35	25	3.8	Common strong sources (L-band, 1–1.55 GHz)
34 to 128	70	9.0	Rare strong sources (all frequencies)

Table 3. Dynamic range requirements and goals.

Number of accumulations	Dynamic range requirement, dB	Dynamic range goal, dB
2–20	7.1	9.0
20–40	3.8	6.8
40–60	3.8	6.4
60–80	3.8	6.2
80–100	3.7	6.2

Table 4. Minimum population size versus fractional error.

Error, σ, %	$r/N = 0.25$	$r/N = 0.375$	$r/N = 0.5$
1.00	15,058	10,931	9035
0.50	60,231	43,726	36,140
0.33	135,520	98,383	81,316

Table 5. Fixed threshold, threshold-and-count estimator performance with system temperature change to T'_{sys} from T_{sys}. N = 65536 samples in the population.

Number of accumulations	β factor	T_{sys}/T'_{sys} 0.5		T_{sys}/T'_{sys} 0.75		T_{sys}/T'_{sys} 1		T_{sys}/T'_{sys} 1.25		T_{sys}/T'_{sys} 1.5		T_{sys}/T'_{sys} 1.75		T_{sys}/T'_{sys} 2		T_{sys}/T'_{sys} 2.25	
Operating point, r/N, = 0.25		r/N, %	Standard error, %	r/N, %	Standard error, %	r/N, %	Standard error, %	r/N, %	Standard error, %	r/N, %	Standard error, %	r/N, %	Standard error, %	r/N, %	Standard error, %	r/N, %	Standard error, %
2	0.96	57.19	0.358	36.61	0.413	24.95	0.479	17.98	0.548	13.52	0.619	10.53	0.690	8.42	0.761	6.88	0.832
4	2.54	74.60	0.246	43.87	0.261	25.10	0.310	14.87	0.373	9.22	0.448	5.97	0.534	4.02	0.630	2.79	0.735
6	4.22	84.58	0.217	49.27	0.205	25.01	0.245	12.64	0.308	6.63	0.392	3.64	0.498	2.09	0.626	1.25	0.780
8	5.96	90.69	0.211	53.96	0.174	25.05	0.208	11.03	0.272	4.96	0.366	2.33	0.494	1.15	0.664	0.59	0.882
10	7.73	94.37	0.217	57.98	0.155	25.05	0.183	9.72	0.249	3.77	0.354	1.53	0.509	0.65	0.730	0.29	1.035
20	16.83	99.56	0.375	72.54	0.113	25.00	0.126	5.67	0.202	1.13	0.375	0.23	0.738	0.05	1.472	0.01	2.897
MAX error			0.375		0.413		0.479		0.548		0.619		0.738		1.472		2.897
Operating point, r/N, = 0.375																	
2	1.31	73.65	0.344	52.11	0.367	37.67	0.409	28.19	0.456	21.78	0.506	17.29	0.557	14.03	0.609	11.61	0.661
4	3.1	86.58	0.266	59.21	0.246	37.52	0.273	23.82	0.315	15.51	0.367	10.42	0.427	7.21	0.495	5.13	0.569
6	4.95	92.90	0.255	64.53	0.199	37.53	0.218	20.86	0.259	11.71	0.316	6.76	0.389	4.04	0.477	2.49	0.582
8	6.82	96.15	0.265	68.69	0.174	37.45	0.186	18.51	0.228	9.05	0.292	4.52	0.379	2.34	0.493	1.26	0.638
10	8.71	97.90	0.290	72.22	0.157	37.44	0.166	16.63	0.208	7.13	0.278	3.10	0.382	1.40	0.528	0.66	0.727
20	18.29	99.89	0.647	83.88	0.123	37.47	0.115	10.52	0.163	2.45	0.275	0.55	0.504	0.13	0.948	0.03	1.786
MAX error			0.647		0.367		0.409		0.456		0.506		0.557		0.948		1.786
Operating point, r/N, = 0.5																	
2	1.6	84.86	0.357	65.51	0.348	50.05	0.371	38.87	0.404	30.83	0.441	24.95	0.479	20.57	0.518	17.22	0.558
4	3.67	93.43	0.308	71.97	0.245	49.96	0.254	33.84	0.281	23.11	0.318	16.08	0.362	11.44	0.412	8.32	0.467
6	5.67	96.94	0.319	76.51	0.204	50.00	0.205	30.32	0.231	18.15	0.271	11.00	0.324	6.82	0.388	4.34	0.463
8	7.67	98.52	0.356	79.95	0.182	50.01	0.176	27.49	0.202	14.55	0.247	7.73	0.309	4.19	0.390	2.34	0.493
10	9.67	99.27	0.413	82.70	0.168	50.02	0.157	25.12	0.183	11.81	0.233	5.51	0.306	2.62	0.408	1.29	0.546
20	19.67	99.97	1.163	91.02	0.141	50.00	0.110	16.96	0.140	4.58	0.218	1.15	0.372	0.29	0.664	0.07	1.198
MAX error			1.163		0.348		0.371		0.404		0.441		0.479		0.664		1.198

Table 6. Standard error versus percentage points of the distribution functions.

Number of accumulations	Standard error at percentage point, percent				
	1	5	50	95	99
2	2.01	0.96	0.37	0.43	0.67
4	1.15	0.58	0.25	0.33	0.53
6	0.86	0.44	0.20	0.28	0.45
8	0.71	0.36	0.18	0.25	0.41
10	0.61	0.32	0.16	0.22	0.37
20	0.39	0.21	0.11	0.16	0.28
50	0.23	0.13	0.07	0.11	0.19
100	0.17	0.09	0.05	0.08	0.12

Table 7. Estimator dynamic range.

Accumulations	Percentage point					Dynamic range percentage, dB		
	1	5	50	95	99	1–99	5–95	1–50
2	0.30	0.71	3.36	9.49	13.28	16.5	11.3	10.5
4	1.65	2.73	7.34	15.51	20.09	10.9	7.54	6.48
6	3.57	5.23	11.34	21.02	26.22	8.66	6.04	5.02
8	5.81	7.96	15.34	26.30	32.00	7.41	5.19	4.22
10	8.26	10.85	19.34	31.41	37.57	6.58	4.61	3.69
20	22.16	26.51	39.34	55.76	63.69	4.58	3.22	2.50
50	70.06	77.93	99.33	124.3	135.8	2.88	2.04	1.52
100	156.0	167.9	199.5	233.8	248.3	2.01	1.43	1.07

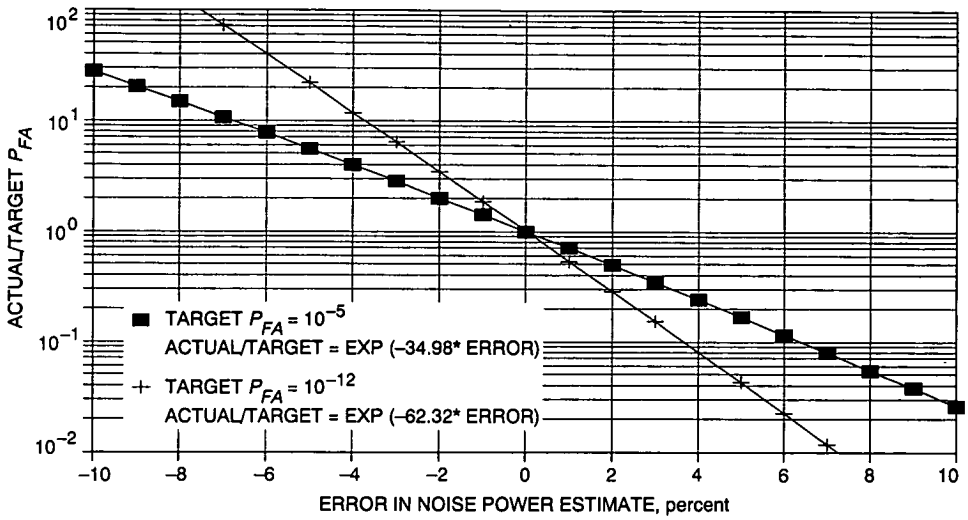


Fig. 1. Effect of noise power estimate error on P_{FA} 10 accumulations.

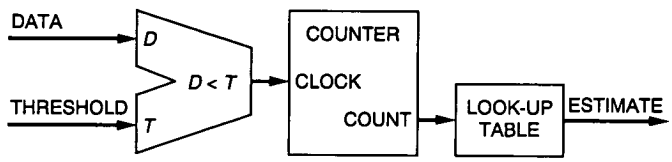


Fig. 2. Threshold-and-count estimator.

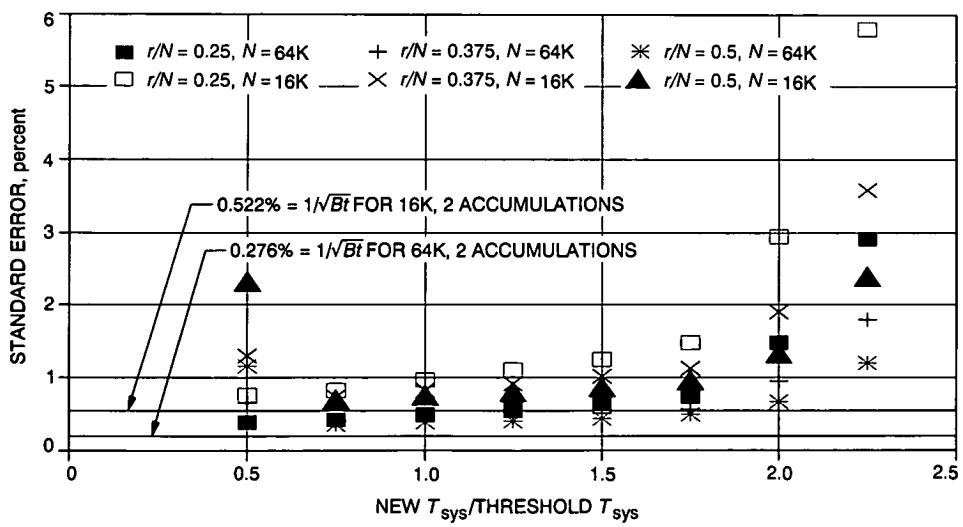


Fig. 3. Standard error versus system temperature change for a fixed-threshold estimator, showing the maximum error over 2 to 20 accumulations.