COMSAT Technical Review

Volume 6 Number 1, Spring 1976

Advisory Board
Joseph V. Charyk
William W. Hagerty
John V. Harrington
Sidney Metzger

Editorial Board
Pier L. Bargellini, Chairman
Robert D. Briskman
S. J. Campanella
William L. Cook
Denis J. Curtin
Jorge D. Fuenzalida
Joachim Kaiser
R. W. Kreutel
Terrence P. McGarty
H. J. Meyerhoff
Akos G. Revesz
Robert Strauss

Editorial Staff
Stephen D. Smoke
MANAGING EDITOR
Leonard F. Smith
Margaret B. Jacocks
TECHNICAL EDITORS
Edgar Bolen
PRODUCTION
Michael K. Glasby
CIRCULATION

COMSAT TECHNICAL REVIEW is published twice a year by Communications Satellite Corporation (COMSAT). Subscriptions, which include the two issues published within a calendar year, are: one year, $7 U.S.; two years, $12; three years, $15; single copies, $5; article reprints, $1. Make checks payable to COMSAT and address to Treasurer's Office, Communications Satellite Corporation, 950 L'Enfant Plaza, S. W., Washington, D.C. 20024, U.S.A.

© COMMUNICATIONS SATELLITE CORPORATION 1976

COMSAT TECHNICAL REVIEW

Volume 6 Number 1, Spring 1976

[illegible] Paul [illegible]

[illegible] microwave switch network (MSN

[illegible] MMIC dual switch [illegible]

[illegible] head-to-head [illegible]

8/18 2 CH

Index: attitude control, transfer orbits, spacecraft maneuvers, torque motors.

Attitude acquisition maneuver for bias momentum satellites

M. H. Kaplan and T. C. Patterson

(Manuscript received September 4, 1975)

Abstract

Certain 3-axis-stabilized geostationary spacecraft spin about a principal axis during the transfer orbit interval to provide attitude stability, ensure solar power for telemetry and command, and minimize thermal gradients. The use of bias momentum for on-orbit control and an apogee rocket motor for synchronous injection leads to a transition problem because initially the spacecraft will probably be spinning about its yaw axis while the wheel will be mounted about the pitch axis. Thus, the spacecraft must be despun and reoriented, and the momentum wheel must be spun up.

This paper presents an attitude acquisition sequence which minimizes the use of sensors, logic, and propellant. The technique is based upon an open-loop maneuver using the torque motor to simultaneously reorient the satellite and spin up the wheel. A physical interpretation of this dynamic sequence is of primary concern because this maneuver has not been well understood, nor have practical quantitative performance equations or stability conditions for the spacecraft designer been forthcoming. Results indicate that this maneuver is practical only when the initial spin axis is the axis of maximum inertia. Otherwise, it results in unstable motion or very large values of nutation angle with respect to the wheel axis. A simple nutational performance equation is also presented.

This paper is based upon work performed at COMSAT Laboratories under the sponsorship of the International Telecommunications Satellite Organization (INTELSAT). Views expressed in this paper are not necessarily those of INTELSAT.

Introduction

Many new and planned geostationary communications satellites and other high-altitude spacecraft have body-stabilized configurations which employ bias momentum devices and apogee rocket motors. Novel acquisition maneuvers are necessary to establish on-orbit attitude with minimum propellant and sensor penalties. During the transfer orbit phase, spin stabilization about the yaw axis is typically employed to maintain orientation prior to apogee burn, to minimize apogee motor thrust misalignment effects, to ensure solar power, and to minimize thermal gradients. Upon completion of orbit injection, it is then necessary to despin the vehicle, reorient it, and spin up the momentum wheel. Since this wheel is expected to be mounted along the pitch axis, which is transverse to that of the apogee motor, this sequence may be complicated. An attitude acquisition scheme has been recently considered in which turning on the wheel spin-up motor simultaneously transfers momentum from the yaw axis to the pitch-axis-mounted wheel [1].

A special case of the maneuver, known as "flat-spin recovery," has been reported in the literature [2]–[4]. Barba and Aubrun have applied the conditions of a flat-spin case to the acquisition problem of a bias momentum satellite, dealing primarily with computer simulations and qualitative arguments based on the application of energy and momentum considerations. They have also developed a computer-generated movie of a complete acquisition sequence. Gebman and Mingori have considered a flat-spin recovery through the application of multiple-time-scale asymptotic parameter expansion techniques, a sophisticated mathematical approach leading to limited quantitative results in terms of maneuver performance.

This paper presents a unique physical interpretation of the more general problem of acquisition using the momentum wheel. Flat-spin recovery situations require that the spacecraft initially spin about its major axis of inertia. The treatment presented herein, which considers all principal axis cases and excludes cross products of inertia, leads directly to simple, but apparently accurate, expressions for acquisition performance. Conditions for successful execution of the maneuvers have also been developed and verified by computer simulations. Although limited results obtained from this type of approach have been reported [5], more general and complete conclusions are now offered. The approach consists of developing the appropriate equations of motion and representative simulations to emphasize critical ranges of parameters. Effects of varying inertia ratios, initial conditions, motor torque, and wheel momentum have been studied.

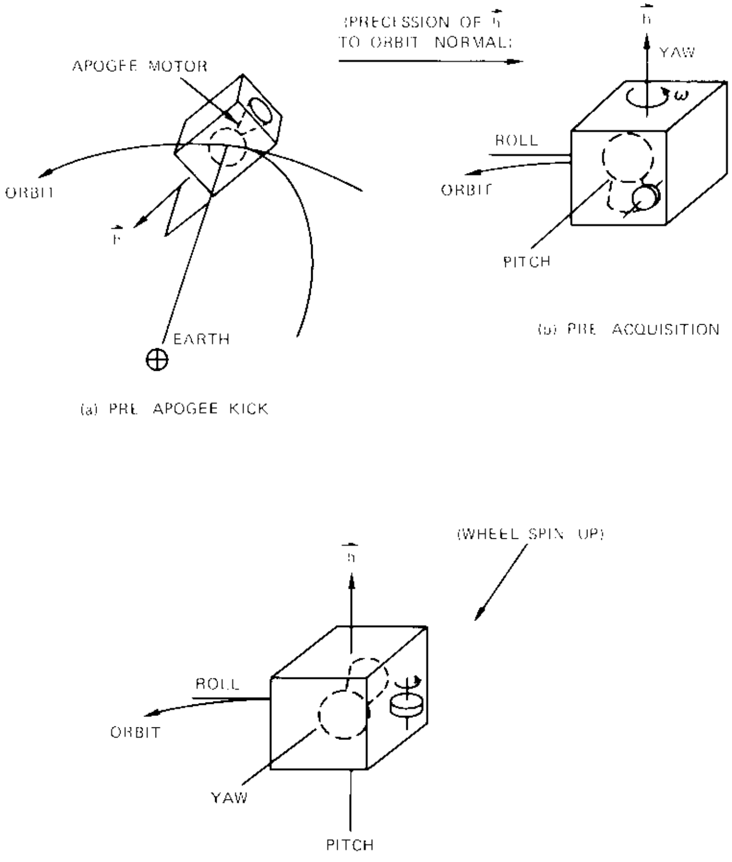
To understand the need for such a variety of approaches to this maneuver, consider a simple first attempt at testing feasibility. If the wheel spins up to a speed corresponding to the entire original angular momentum of the spacecraft, then the law of conservation of momentum would appear to guarantee success. However, closer examination of this principle indicates that it is the sum of individual angular momentum vectors which is conserved. It is entirely possible that, as the wheel spins up, the platform may be spinning up further in the opposite direction so that the satellite will reorient itself into an "upside down" position. As demonstrated herein, this situation is unstable and tends to generate large nutation angles. Thus, a more sophisticated approach is required to determine the distribution of angular momentum during execution of the maneuver.

In summary, this paper deals with the several aspects of a new attitude acquisition maneuver for bias momentum satellites. Of particular interest are high-altitude, initially spinning configurations which employ a single momentum wheel for on-orbit attitude control. Results indicate that any such satellite whose initial spin axis is its major principal (maximum inertia) axis can successfully use this maneuver. Furthermore, a simple expression can give a quantitative measure of this success.

Acquisition sequence

For a typical acquisition sequence, consider that of a body-stabilized geostationary satellite. The three steps shown in Figure 1 comprise the transverse wheel acquisition sequence. At an appropriate apogee, the circularization rocket is aligned with the proper thrust direction, followed by execution of the insertion or apogee motor burn. An entire attitude acquisition sequence is presented in Table 1, beginning with this event. Steps 2 through 5 ensure that spacecraft angular momentum is decreased to a value close to that of the final bias momentum of the wheel. Some excess momentum should be included to provide a slewing rate about the pitch axis after wheel spin-up which will permit earth acquisition and minimize thermal gradients before solar array deployment. Step 6 is a precession maneuver required to orient the momentum vector along the orbit normal. The satellite is still spinning about its yaw axis with the wheel aligned along the pitch axis.

The wheel motor is turned on in step 8 and reorientation begins. Spacecraft momentum is absorbed by the wheel as the pitch axis aligns itself with the orbit normal. As the wheel approaches its final speed, nutation begins and dampers are used passively to decrease this motion in a reason-



(a) PRE APOGEE KICK

(b) PRE ACQUISITION

(c) POST ACQUISITION

Figure 1. *Transverse Wheel Acquisition Sequence*

able time. No attitude sensors are required during the wheel spin-up interval. Only acceleration of the wheel with respect to the spacecraft should be limited. Otherwise, the maneuver is done in an open-loop manner. Steps 9 through 11 complete the sequence. In principle, the concept of the transverse wheel maneuver is simple, but conditions for success and resulting nutational motion have not been well understood.

TABLE 1. ATTITUDE ACQUISITION SEQUENCE FOR A SYNCHRONOUS, BIAS MOMENTUM SATELLITE

Event	Equipment	Comments
1. Apogee Motor Burn	Solid or Liquid Apogee Rocket	Injection into synchronous orbit with inclination near zero.
2. Attitude Determination	Sun and Earth Sensors	To check post-apogee kick orientation.
3. Spin Rate Determination	Sun Sensors	To calculate despin impulse.
4. Despin to a Preselected Rate	Thrusters	To establish initial rate for transverse wheel maneuver.
5. Spin Rate Check	Sun Sensors	To verify despin maneuver.
6. Attitude Precession	Sun Sensors and Thrusters	To align yaw axis to orbit normal.
7. Attitude Check	Sun and Earth Sensors	Last check before acquisition begins.
8. Execution of Transverse Wheel Maneuver	Torque Motor and Passive Damper	To bring pitch axis to orbit normal orientation and despin body to earth capture slew rate.
9. Earth Acquisition and Momentum Dumping	Earth Sensor and Thrusters	
10. Solar Array Deployment	Deployment Mechanisms	
11. Sun Acquisition	Sun Sensors, Drive Motor, and Thrusters	Full power available.

Equations of motion

Equations describing the motion of a spacecraft with moving internal and external parts have appeared frequently in recent literature. Of concern here are the particular equations associated with an asymmetrical platform, P, containing a symmetric wheel, w. Without loss of generality, a linear mass-spring-dashpot type damper is included to allow nutation damping and to ensure stability where possible. This device is easily modeled and the qualitative aspects of conclusions based on the presence of damping should be independent of the type of damper. Products of inertia with the damper undeflected are assumed to be zero, and principal axes are labeled y, r, and p to represent yaw, roll, and pitch, respectively. Figure 2 shows the configuration, and the nomenclature follows Likins [6].

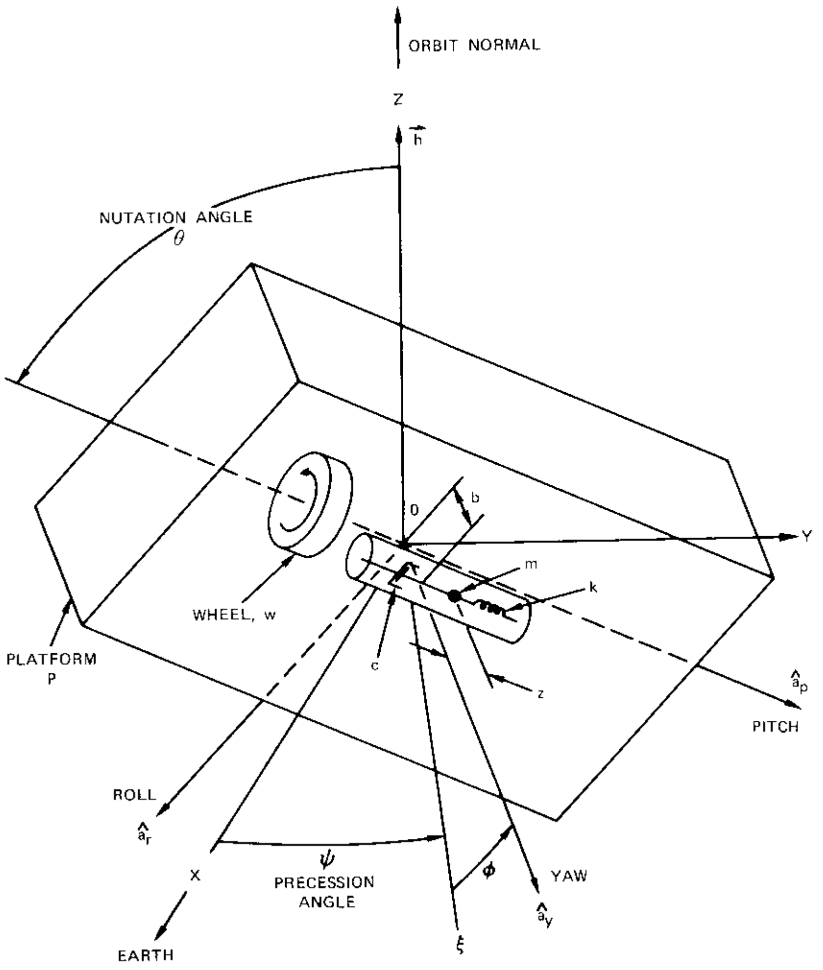


Figure 2. *Nomenclature and Configuration Considered*

The damper is located to provide maximum sensitivity to nutational oscillations with its center on the yaw axis and mass movement parallel to the pitch axis at distance b with displacement z. The vehicle center of

mass is at point 0. With the damper in motion, the inertia tensor, written in matrix form, is

$$\bar{\bar{I}} = \begin{bmatrix} I_y + mz^2 & 0 & -mbz \\ 0 & I_r + mz^2 & 0 \\ -mbz & 0 & I_p \end{bmatrix} .$$

The general vector equation of motion for a body with moving parts is given by [7]

$$\vec{M} = \dot{\vec{h}} + \vec{s} \times \vec{a} \quad . \tag{1}$$

In this case,

$$\vec{h} = \bar{\bar{I}} \cdot \vec{\omega} + I_w \Omega \hat{a}_p - mb\dot{z}a_r \quad .$$

The vector product $\vec{s} \times \vec{a}$ accounts for center of mass motion as the damper mass oscillates, i.e., $z \neq 0$. Thus, for the case at hand, $\vec{s} = mz\hat{a}_p$ and $\vec{a} = -\mu\, d^2/dt^2\, (z\hat{a}_p)$, where $\mu = m/(m + M_p + M_w)$ and time derivatives refer to inertial space. Since cases of interest exclude outside torques, $\vec{M} = 0$. Carrying out the indicated operations in equation (1) yields

$$\begin{aligned} I_y\dot{\omega}_y - \omega_r\omega_p(I_r - I_p) + I_w\Omega\omega_r + m(1 - \mu)\, \dot{\omega}_y z^2 - m(1 - \mu)\, \omega_r\omega_p z^2 \\ + 2m(1 - \mu)\, \omega_y \dot{z} z - mb\dot{\omega}_p z - mb\omega_y\omega_r z = 0 \end{aligned} \tag{2}$$

$$\begin{aligned} I_r\dot{\omega}_r - \omega_p\omega_y(I_p - I_y) - I_w\Omega\omega_y + m(1 - \mu)\, \dot{\omega}_r z^2 + m(1 - \mu)\, \omega_y\omega_p z^2 \\ + 2m(1 - \mu)\, \omega_r z\dot{z} - mb\ddot{z} + mb\omega_y^2 z - mb\omega_p^2 z = 0 \end{aligned} \tag{3}$$

$$I_p\dot{\omega}_p - \omega_y\omega_r(I_y - I_r) + I_w\dot{\Omega} + mb\omega_r\omega_p z - 2mb\omega_y\dot{z} - mb\dot{\omega}_y z = 0 \quad . \tag{4}$$

These three equations in five unknowns, ω_y, ω_r, ω_p, Ω, and z, require two additional relationships associated with wheel torque and damper force balance. Since wheel torque is countered by inertial torque,

$$I_w(\dot{\omega}_p + \dot{\Omega}) = T \tag{5}$$

where bearing friction is ignored. Acceleration-force balance of the damper mass gives

$$\begin{aligned} m(1 - \mu)\, \ddot{z} + c\dot{z} + kz - m(1 - \mu)(\omega_y^2 + \omega_r^2)\, z \\ + mb\omega_y\omega_p - mb\dot{\omega}_r = 0 \quad . \end{aligned} \tag{6}$$

Equations (2)-(6) constitute a complete description of attitude motion for a body containing a symmetric rotor and damper as specified above.

One special solution of interest here corresponds to nominal orientation and operation on orbit. Thus, $\omega_p = \omega_P =$ constant, $\Omega = \Omega_w =$ constant, and $\omega_y = \omega_r = z = T = 0$ is the solution which represents normal attitude motion when ω_P is the orbital rate. Another basic and mathematically trivial solution represents the post-apogee-kick situation: $\omega_y =$ constant, and $\omega_r = \omega_p = \Omega = z = T = 0$.

Stability considerations

Stability of motion is sometimes tested by perturbing special solutions. This technique produces linearized equations of motion which can be studied by using linear methods. Rigorous and approximate methods have been considered in the literature for dual-spin spacecraft [6]-[8]. These methods have been surveyed for application to the situation at hand, and they have provided rigorous stability conditions for the final state, i.e., normal attitude motion, where ω_p is the orbital rate.

Routh's method, which is applied to the final state situation, requires linearization about a known solution of the equations of motion and at least limited dissipation of energy. A perturbed form of motion about this solution is generated through a variation of coordinates: $\omega_y = \omega_y$, $\omega_r = \omega_r$, $\omega_p = \omega_P + \omega_P{}^*$, $z = z$, and $\Omega = \Omega_w + \Omega^*$, where ω_P and Ω_w are constants and ω_y, ω_r, $\omega_P{}^*$, Ω^*, and z are small quantities. Equations (2), (3), and (6) can then be specialized. Equations (4) and (5) can be eliminated because they represent perturbed motion about the $\hat{a}_p$ axis, and only motion about $\hat{a}_y$ and $\hat{a}_r$ need be considered for stability. Hence, equations (2), (3), and (6) become

$$\dot{\omega}_y + \lambda_1\omega_r = 0$$

$$\dot{\omega}_r - \lambda_2\omega_y - \delta\ddot{\zeta} - \delta\omega_P^2\zeta = 0$$

$$(1 - \mu)\,\ddot{\xi} + \beta\dot{\zeta} + p^2\zeta + \omega_P\omega_y - \dot{\omega}_r = 0$$

where
$$\lambda_1 = \frac{h - I_r(\omega_P)}{I_y}, \qquad \lambda_2 = \frac{h - I_y(\omega_P)}{I_r} \tag{7}$$

and $p = \sqrt{k/m}$, $\beta = c/m$, $\zeta = z/b$, and $\delta = mb^2/I_r$.

Note that angular momentum magnitude, which is nominally

$$h = I_p^P\,\omega_P + I_w\Omega_w$$

where I_p^P is the platform moment of inertia about its pitch axis, is assumed to be unchanged by the perturbation. The difference between relative and absolute wheel momenta is negligible because body rates are small relative to the wheel speed. The associated characteristic equation is

$$s^4(1 - \mu - \delta) + s^3\beta + s^2[p^2 - \delta\omega_P^2 + \lambda_1\lambda_2(1 - \mu) - \lambda_1\delta\omega_P]$$
$$+ s\beta\lambda_1\lambda_2 + (\lambda_1\lambda_2 p^2 - \lambda_1\delta\omega_P^3) = 0 \quad .$$

A necessary condition for asymptotic stability is that all coefficients of this equation have the same sign. This type of stability occurs when perturbed motion, which has been superimposed on some steady-state situation, dissipates with time and motion returns to its original state. For typical systems $\delta < < 1$ and $\mu < < 1$, resulting in the condition $\lambda_1\lambda_2 > 0$. The Routhian array is used to establish necessary and sufficient conditions. Thus, if and only if all signs of the first array column are the same (positive here), motion is asymptotically stable. This column reduces to a list of inequality conditions [6]:

$$1 - \mu - \delta > 0 \tag{8a}$$

$$\beta > 0 \tag{8b}$$

$$p^2 - \delta(\omega_P^2 + \omega_P\lambda_1 - \lambda_1\lambda_2) > 0 \tag{8c}$$

$$\lambda_1(\omega_P - \lambda_2)(\omega_P^2 - \lambda_1\lambda_2) > 0 \tag{8d}$$

where $\delta > 0$ by definition and

$$\lambda_1(\lambda_2 p^2 - \delta\omega_P^3) > 0 \quad . \tag{8e}$$

In cases of practical interest, the damper mass is small, and in general, $p^2 > 0$, $\beta > 0$. As $m \rightarrow 0$ (and also as $\mu \rightarrow 0$ and $\delta \rightarrow 0$), the only remaining non-trivial conditions for stability are $\lambda_1(\omega_P - \lambda_2)(\omega_P^2 - \lambda_1\lambda_2) > 0$ and $\lambda_1\lambda_2 > 0$. The first of these conditions can be rewritten using equation (7):

$$\frac{\lambda_1 h[\omega_P(I_y + I_r) - h]^2}{I_y I_r^2} > 0$$

or

$$\lambda_1 h > 0 \quad .$$

If $h > 0$ is required by sign convention, then $\lambda_1 > 0$, and the only other required condition is $\lambda_2 > 0$. Again from equation (7), these become

$$h - I_r\omega_P > 0 \tag{9a}$$

$$h - I_y\omega_P > 0 \quad . \tag{9b}$$

For practical cases the wheel has a momentum less than h; i.e., $h > I_w\Omega_w$. Thus, $\omega_P > 0$, and an unstable condition can exist. These results will be applied later to several examples.

Simulations

Digital computer simulations were used to generate body rates and nutation angle histories. A large number of cases were run to demonstrate the feasibility of the process and to confirm the validity of analytic results. In all cases the initial spin axis was nominally along the positive yaw axis. A constant net torque was applied to the momentum wheel, which was initially at rest relative to the vehicle body. In some cases torque was applied until the pitch rate of the vehicle reached zero. In others, torque was applied until the wheel acquired a desired angular momentum relative to the vehicle. Moments of inertia, motor torque, initial yaw rate, and final wheel momentum were varied.

The maneuver phase of primary interest is initiation of momentum transfer. Some simplification of the equations is possible because damping has proven to be a minor factor in this part of the sequence. A comparison of damped and undamped cases using two digital computer programs has permitted the elimination of a damper model for the results presented here. A typical case is shown in Figure 3, in which angular rates about the pitch and yaw axes and nutation angle are plotted. This case corresponds to a vehicle with inertia ratios representative of future communications satellites. Pitch rate starts upward in a positive manner as a direct response to wheel torque turn-on, but gyro-torquing and centrifugal effects quickly reverse the sign of ω_p. The nutation angle decreases monotonically until $\theta_{\min}$ is reached. At this point the period of almost steady precession, decreasing yaw rate, and negatively increasing pitch rate ends. Beyond this point motion is essentially spin about the pitch axis with nutation and precession. Note that $\theta_{\min}$ is reached as ω_y changes sign for the first time. Thereafter θ oscillates between $\theta_{\min}$ and some higher value due to inertia asymmetries. The mean value is θ_f, which is about 23° for the case shown in Figure 3.

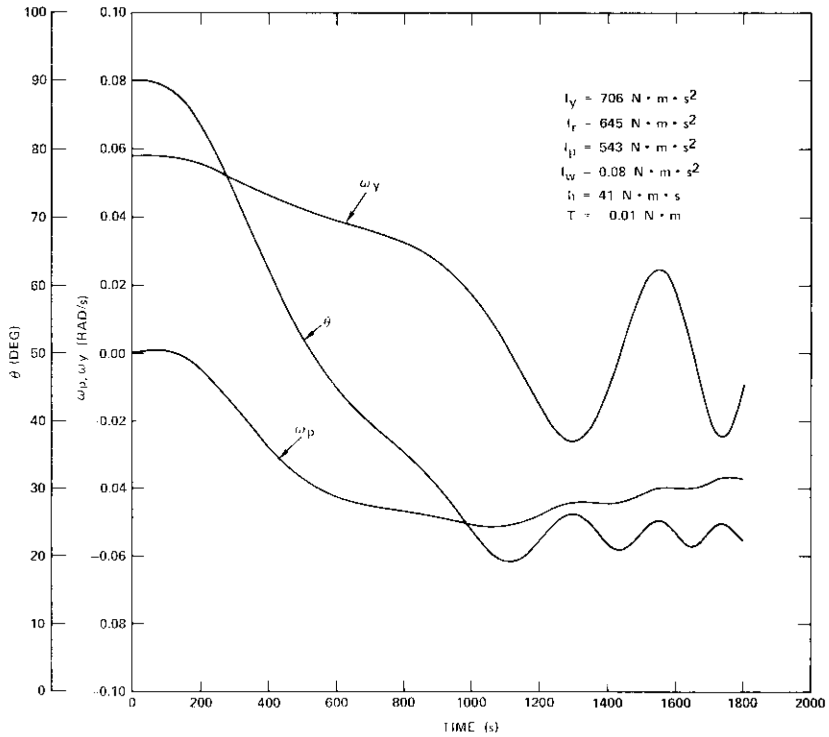


Figure 3. *Typical Simulation Profile with I_y as the Maximum Inertia*

Physical arguments

Dynamics of initial wheel start-up will now be considered by using two heuristic arguments. First, an interpretation of the equations of motion as the wheel speeds up is employed to derive a basic condition required of the inertia ratios. Then a direct interpretation of internal torques and momentum balance using a dumbbell inertia model is discussed. This technique permits the derivation of a very basic and useful performance equation for the wheel spin-up portion of the acquisition maneuver. Since dampers will probably be tuned to be most effective near the final state, very little dissipation will occur initially. Thus, equation (4) for pitch axis motion gives

$$I_p\dot{\omega}_p + I_w\dot{\Omega} - \omega_y\omega_r(I_y - I_r) = 0$$

which may be rewritten as

$$I_p^P \dot{\omega}_p = (I_y - I_r)\, \omega_y \omega_r - T$$

where $I_p^P = I_p - I_w$. Since T is taken as a negative quantity during reorientation,

$$I_p^P \dot{\omega}_p = (I_y - I_r)\, \omega_y \omega_r + |T| \quad . \tag{10}$$

Initial wheel start-up conditions are developed by noting that the first term on the right-hand side is the inertial torque. It must have a negative value large enough to exceed the motor torque for a period of time near the beginning of the momentum wheel spin-up to prevent platform spin-up in the opposite direction. It is assumed that the vehicle is initially spinning about the yaw axis with $\omega_y > 0$. The roll rate is zero when the motor torque is first applied. Therefore, for at least a short period of time, the motor torque will dominate and ω_p will be positive. Once the wheel starts spinning, ω_r will become negative because pitch motion induces a negative component of $\vec{\omega}$ along the roll axis. The sign of the inertial torque thus depends on the sign of $(I_y - I_r)$. If the initial spin axis, $\hat{a}_y$, has a greater inertia than the other axis transverse to the wheel, $\hat{a}_r$, inertial torque will have the required negative sign. This appears to be a necessary condition for proper initiation of the maneuver. The presence of products of inertia would modify this argument.

In addition to being a negative quantity, the inertial torque must have a sufficiently large magnitude to exceed the motor torque. It might also be expected that parameter variations which tend to increase the magnitude of the inertial torque relative to the motor torque will result in the buildup of a greater net angular momentum along the pitch axis and therefore a smaller final nutation angle. Computer simulations have confirmed that decreases in $|T|$ and increases in $(I_y - I_r)$ and ω_y produce smaller final nutations.

This interpretation can be visualized through the use of dumbbell models, originally applied at the Space and Communications Group of Hughes Aircraft Company by D. B. Krimgold and G. J. Adams to gain insight into processes occurring during flat-spin recovery of a dual-spin satellite. Since the transverse wheel spin-up maneuver is similar to flat-spin recovery, a similar approach has been quite helpful in pointing out physical explanations of the effects of different parameters on successful reorientation and final nutation angle. Although the arguments presented are heuristic, they have been confirmed by simulations in the cases studied.

The dumbbell-model approach depicts platform and rotor asymmetries as pairs of dumbbells oriented to represent dynamic imbalances (products of inertia) and mass asymmetries (differences in inertias). Figure 4 is a basic dumbbell representation. For a dual-spin satellite both rotor and platform can be modeled in this way. Dynamic torques produced by centrifugal forces acting when the motor is engaged generally result in an upper and lower bound on motor torque to achieve despin. The lower limit is associated with an asymmetrical rotor and establishes a minimum required torque to overcome dynamic moments due to centrifugal forces acting on the rotor dumbbell pairs during flat-spin recovery of a dual-spin satellite. The upper limit is associated with platform spin-up. To transfer momentum from a transverse axis to the wheel axis, the platform must resist being spun up while the rotor or wheel increases its momentum. Thus, an upper bound on motor torque is just the dynamic torque on the platform due to initial spin rate and inertia asymmetries of the platform itself.

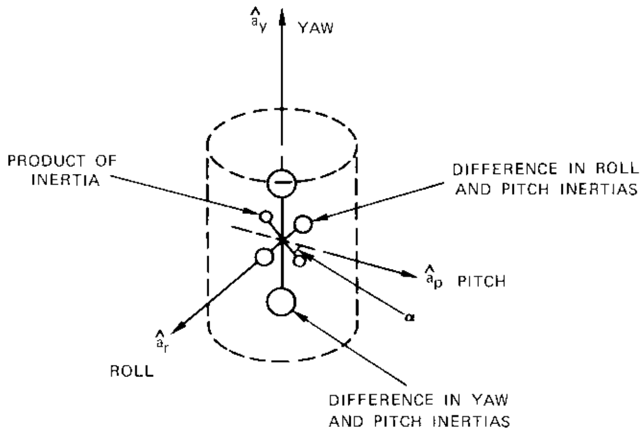

$\hat{a}_y$
YAW
PRODUCT OF INERTIA
DIFFERENCE IN ROLL AND PITCH INERTIAS
$\hat{a}_p$ PITCH
α
$\hat{a}_r$
ROLL
DIFFERENCE IN YAW AND PITCH INERTIAS

Figure 4. *Dumbbell Representation of Inertia Distribution*

The situation of interest here dictates an axially symmetric wheel, which implies that there is no dynamic torque to restrain it from spinning up under any motor torque. Therefore, there is no lower bound on the applied torque. However, the platform asymmetries still dictate an upper bound. In fact, this bound does not necessarily determine the success of the reorientation, but only the nutation angle from which active or passive damping must bring the vehicle to the final orientation. This nutation angle generally increases with increased torque. For a given target nutation angle to be reached as the wheel attains the final spin rate, there is an upper bound established by initial yaw rate and platform inertia asymmetry. Thus, a platform which is dynamically balanced and symmetrical about the pitch axis has an upper torque bound of zero and any applied torque will leave the nutation angle near 90°.

The process of momentum transfer and the nature of restraining torques may be described with reference to a balanced asymmetrical platform such as that depicted in Figure 5. This configuration corresponds to an inertia distribution in which $I_y > I_r > I_p$. Initially the spacecraft is spinning about the yaw axis and the wheel is stationary with respect to the body. When torque is applied to the wheel, its speed steadily increases. This torque is experienced by the platform in an opposite direction and a pitching motion begins. However, this motion is restrained by a restoring torque due to centrifugal forces (from ω_y) acting on the small dumbbell pair. Thus, if the motor torque is small, pitch motion is essentially stopped at an angle which produces an equal and opposite centrifugal torque. Meanwhile, the wheel continues to collect momentum.

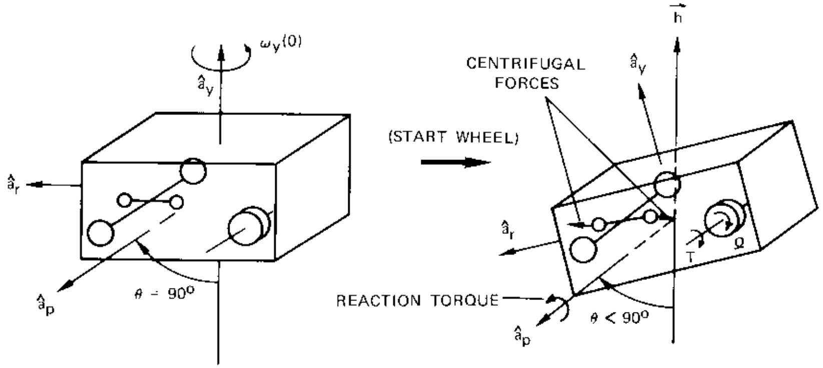


Figure 5. *Wheel Start-up Situation*

To obey the law of conservation of momentum, the wheel axis must tend to line up with the original yaw axis direction and the platform rates must be decreasing. Thus, a rolling motion also takes place. The large dumbbells are deflected, but the yaw angular rate produces a centrifugal torque to restrain them. However, gyrocompassing torque from the wheel counters this restraining torque, and roll motion continues ($\omega_r < 0$) as the wheel spins up. As the body rates decrease, centrifugal torques diminish until they cannot match the motor torque. At this point nutation reaches a minimum and thereafter oscillates between two limits until damping dissipates this motion. Thus, pitch continuously adjusts until the orientation associated with maximum centrifugal torque is reached, at which time this torque just equals that of the wheel spin-up motor. The best situation is one in which initial yaw rate is high and the platform is highly asymmetrical about the pitch axis. Both of these factors contribute to increased dynamic torque, thus permitting higher motor torque for a given nutation angle.

Development of performance equations

Now that the physical basis for the transition maneuver has been established, expressions for dynamic torque may be derived and performance expressions established. The momentum wheel is assumed to be axially symmetric and parallel to the body pitch axis. Equations of motion during wheel spin-up neglect damping in the platform because of low damper frequencies and the short time interval involved. It should be noted that damping in the wheel has not been considered in the past, but that such damping is possible and may be a necessary consideration when working with actual hardware.

The equations of motion are easily obtained if the wheel is considered to be a producer of applied torques. Equations (2)–(4) are rewritten without damping as follows:

$$I_y\dot{\omega}_y - \omega_r\omega_p(I_r - I_p) = - I_w\Omega\omega_r$$

$$I_r\dot{\omega}_r - \omega_p\omega_y(I_p - I_y) = I_w\Omega\omega_y$$

$$I_p\omega_p - \omega_y\omega_r(I_y - I_r) = - I_w\dot{\Omega} \quad .$$

The pitch equation relates motor torque to dynamic torque directly when rewritten similar to equation (10):

$$I_p\dot{\omega}_p - \omega_y\omega_r(I_y - I_r) = -T \tag{11}$$

where the wheel inertia is assumed to be negligible relative to I_p. A relationship between inertia ratios, motor torque, and $\theta_{\min}$ can be derived by using this expression. First, it is rewritten in terms of Euler angles and rates using the transformation given by

$$\begin{bmatrix} \omega_y \\ \omega_r \\ \omega_p \end{bmatrix} = \begin{bmatrix} S_\theta S_\phi & C_\phi & 0 \\ S_\theta C_\phi & -S_\phi & 0 \\ C_\theta & 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{\psi} \\ \dot{\theta} \\ \dot{\phi} \end{bmatrix}$$

where $S = \sin$, $C = \cos$, $\dot{\theta}$ is the nutation rate, $\dot{\psi}$ is the precession rate, and $\dot{\phi}$ is the body rate about the pitch axis, referred to the line of nodes, ξ, as shown in Figure 2. Equation (11), the pitch equation, now becomes

$$-T = I_p (\ddot{\phi} + \ddot{\psi} \cos\theta - \dot{\psi}\dot{\theta} \sin\theta) - \frac{1}{2} (I_y - I_r) [\dot{\psi}^2 \sin^2\theta \sin 2\phi + 2\dot{\psi}\dot{\theta} \sin\theta \cos 2\phi - \dot{\theta}^2 \sin 2\phi] \quad . \qquad (12)$$

Consider the physical situation as θ decreases for a given motor torque. The following events occur simultaneously: θ approaches a minimum value, i.e., $\dot{\theta} \rightarrow 0$; $\dot{\psi}$ approaches a stationary value to maintain overall momentum; and $\dot{\phi}$ is essentially zero until $\dot{\theta}$ comes to zero due to torque balances. Thus, if θ stops at the maximum allowable value beyond which damping can take over, $T = T_{\max}$. Equation (12) gives $T_{\max}$ by setting $\dot{\theta} = \ddot{\psi} = \ddot{\phi} = 0$:

$$T_{\max} \cong \frac{1}{2} \dot{\psi}^2 [(I_y - I_r) \sin^2\theta \sin 2\phi] \quad . \qquad (13)$$

The appropriate value of ϕ at $\theta_{\min}$ is the one which maximizes the right side of this equation. This angle changes slowly from 90° as θ decreases and continues to approach either 45° or 135° as θ stops. As $\dot{\theta}$ reaches zero, $\dot{\phi}$ changes significantly. Thus, $\sin 2\phi$ is replaced by 1. Now only $\dot{\psi}$ must be related to the physical situation at $\theta_{\min}$. Since the value of θ is generally determined by

$$\sin^2\theta = \frac{h_y^2 + h_r^2}{h^2} \qquad (14)$$

where $h_y = I_y\omega_y$ and $h_r = I_r\omega_r$, the situation at $\theta_{\min}$ leads to

$$\dot{\psi} = \frac{\sqrt{2}h}{\sqrt{I_y^2 + I_r^2}} \quad . \qquad (15)$$

Torque, momentum, and nutation may now be related through a single expression. Equations (13) and (15) are combined to obtain the "transverse wheel performance equation":

$$T^* = \sin^2 \theta_{\min} \qquad (16)$$

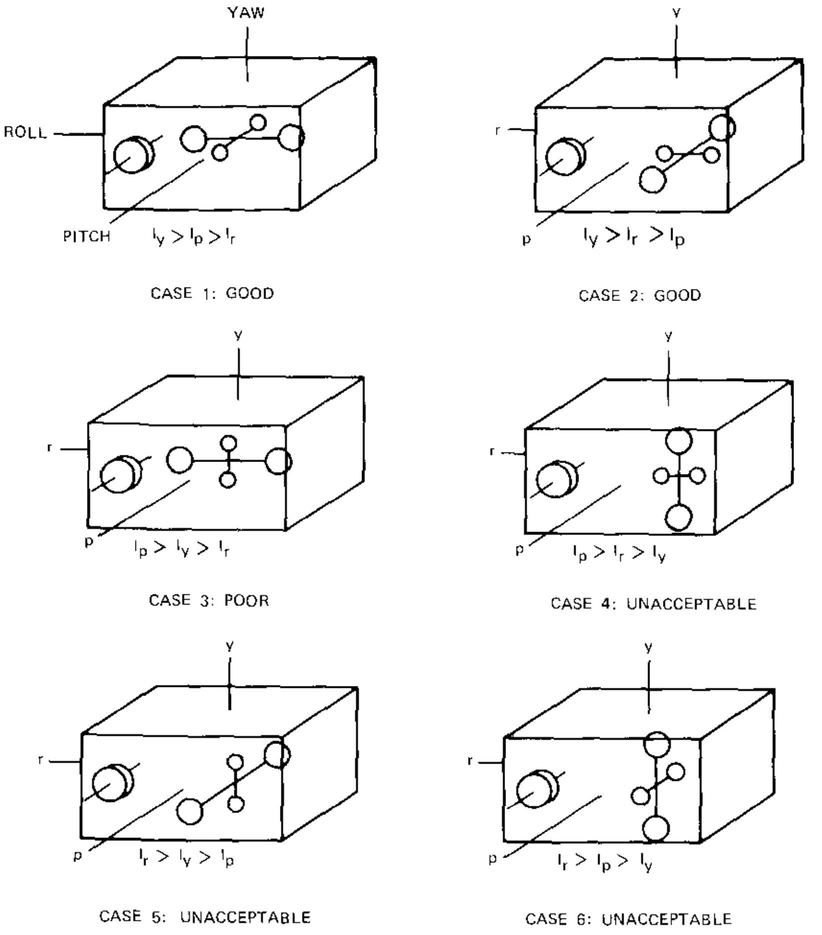
where

$$T^* = \frac{T_{\max} (I_y^2 + I_r^2)}{h^2 (I_y - I_r)} \qquad (17)$$

This result has proven consistent with simulations. Thus, $\theta_{\min}$ is simply related to a nondimensionalized torque. Higher values of torque are permitted for higher values of h if $\theta_{\min}$ is specified. Note that, if $I_r > I_y$, equation (17) does not make sense. Thus, the validity of this result and the success of a maneuver where $I_r > I_y$ are questionable. Initial starting conditions require that $I_y > I_r$ to prevent the body from spinning up in the opposite direction. Simulations confirm that the only acceptable inertia ratios are those for which $I_y > I_r$. Furthermore, the best situations correspond to $I_y > I_p > I_r$ and $I_y > I_r > I_p$.

With the benefit of insight gained from the dumbbell model concept, results may be physically explained. The case in which the pitch axis is the intermediate principal axis with yaw as the major axis appears to give the lowest values of $\theta_{\min}$ for the same motor torque and values of inertia. This has been confirmed by simulation. A worst case may be anticipated by considering other dumbbell arrangements. In Figure 6, case 6, yaw is the minor principal axis and roll the major axis. The restraining pitch torque is not available, but there is a pitching torque to help speed up the pitch rate in the wrong direction. Figure 7 shows a simulation profile for this case. The pitch rate immediately becomes positive and remains positive with increasing magnitude. Nevertheless, simulations indicate that a low value of $\theta_{\min}$ is reached, but that the spacecraft has in fact turned itself upside down so that all wheel momentum is balanced by excess body momentum. The final pitch rate is sufficient to contain initial momentum plus wheel momentum. To summarize all the possible combinations, cases 1, 2, and 3 do maneuver to an upright position, but case 3 results in very large values of $\theta_{\min}$. Cases 4, 5, and 6 turn upside down.

To illustrate the stability problem of an upside down maneuver, equation (9) is applied to the final state stability of case 4, corresponding to $I_p > I_r > I_y$. It is assumed that the wheel momentum magnitude is just



CASE 1: GOOD

CASE 2: GOOD

CASE 3: POOR

CASE 4: UNACCEPTABLE

CASE 5: UNACCEPTABLE

CASE 6: UNACCEPTABLE

Figure 6. *All Possible Asymmetrical Inertia Cases*

equal to the initial body momentum about yaw. After wheel spin-up the distribution of momentum is

$$h_P = 2h_w = 2h \quad .$$

Thus, the net momentum, h, is

$$h = \frac{I_p \omega_p (f)}{2}$$

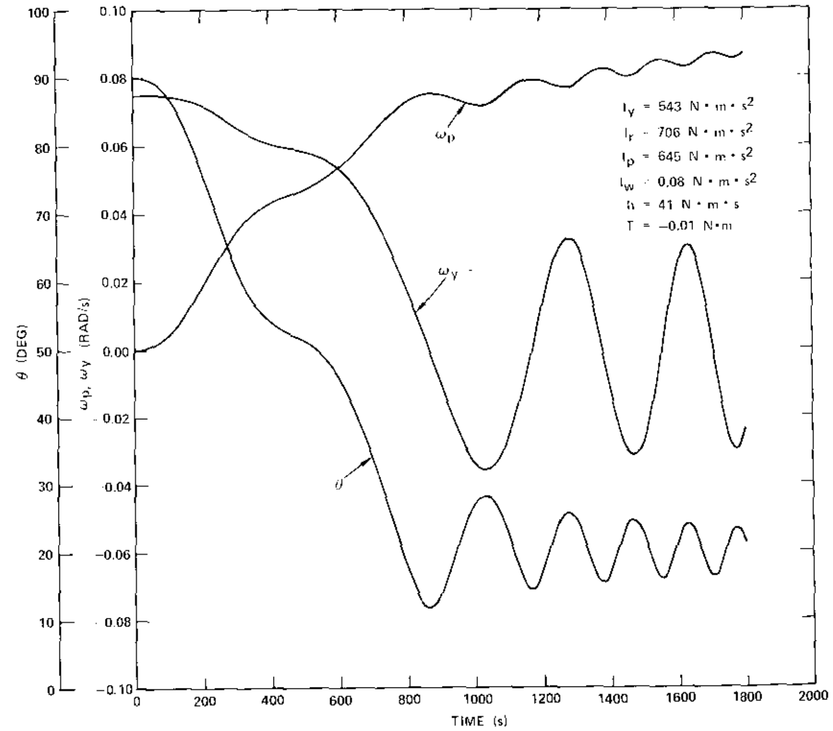


Figure 7. *Simulation of an Upside Down Transfer*

where $\omega_p(f)$ is the final platform rate. The stability conditions become

$$\left(\frac{I_p}{2} - I_r\right) \omega_p (f) > 0, \quad \left(\frac{I_p}{2} - I_y\right) \omega_p (f) > 0$$

or $I_p > 2I_r$, $I_p > 2I_y$. This implies that stability about the pitch axis is possible for this inertia distribution only when the pitch inertia is at least twice that of roll and yaw, a physical impossibility. Wheel cross-coupling torques cause this instability, which, in fact, cannot exist without the wheel because pitch is the major axis.

To anticipate damping times required to dissipate nutation after wheel spin-up, limiting values of θ must be calculated. Typically, the profile of nutation over time, after θ_{min} is reached, includes a peak, θ_{max}, just after θ_{min}. The average of θ_{min} and θ_{max} can be thought of as the nutation angle from which damping must take over after the wheel reaches its final speed. Simulations indicate that this is a valid assumption although wheel spin-up beyond the θ_{min} point results in some convergence of the peaks. Thus, an estimate of θ_{max} is important to anticipate overall reorientation performance. Based on simulations, observations of θ_{max} values, and the form of equation (16), a log-log plot of T^* versus $\sin^2 \theta_{max}$ has been constructed and combined with plots of θ_{min} in Figure 8. These results are only for cases in which yaw is the major axis and pitch is either the minor or intermediate axis. The corresponding equation for θ_{max} is $T^* = 0.8474 (\sin^2 \theta_{max})^{1.347}$, which gives good agreement with simulation results for values of θ_{max} up to about 20°.

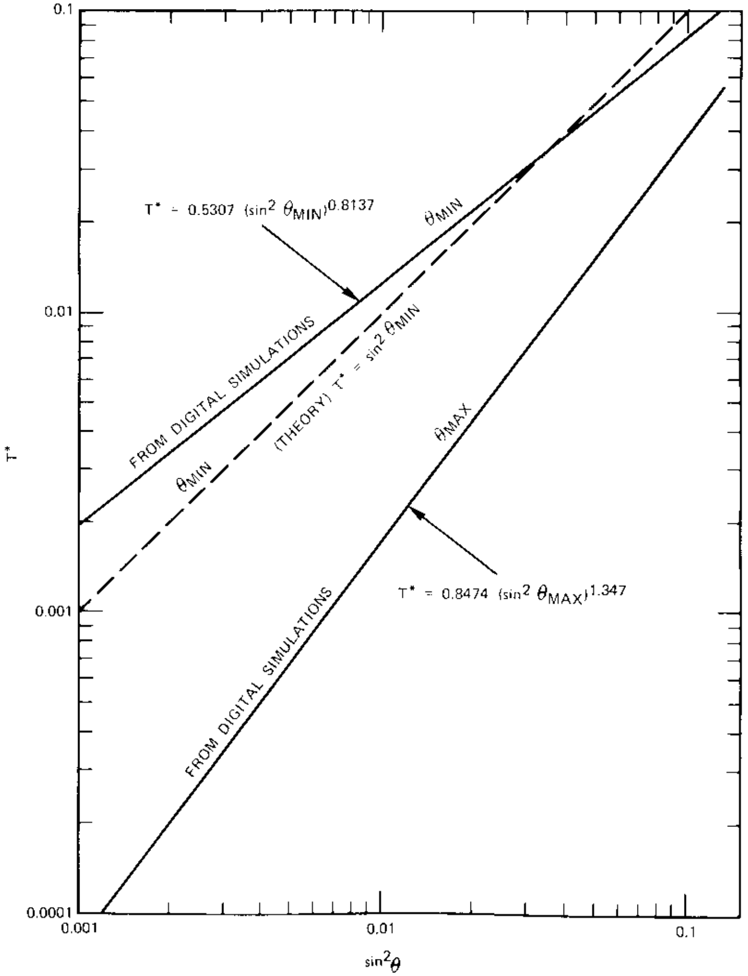


Figure 8. *Summary of Maneuver Performance from Simulations and Physical Interpretation*

Conclusions

A controversial new attitude acquisition maneuver has been described by physical arguments, and results have been verified by computer simulations. The concept of simultaneous despin, reorientation, and momentum wheel spin-up appears practical only for cases in which the initial spin axis is the major principal axis of inertia. In Figure 6, which may be used as a selection chart, only cases 1 and 2 are acceptable. Fortunately, to maintain attitude stability during the transfer orbit, spin about the axis of maximum inertia is highly desirable. Spin about the minor axis is possible with active damping, but this would preclude the acquisition scheme proposed here.

The physical interpretation of dynamic phenomena has not only permitted identification of conditions for a successful maneuver, but has also led to a simple performance equation relating motor torque and inertia asymmetries to the minimum nutation angle reached during wheel spin-up. This is a critical parameter because it allows an estimate of total acquisition time, including nutation damping. Due to the unique relationship between reorientation performance and parameter values, there is an optimum combination of spin-up motor torque and initial vehicle momentum for a given spacecraft and damper. This combination should be determined for each vehicle under consideration. Effects of non-zero products of inertia should also be treated in depth. Reacquisition (flat-spin recovery) of a deployed satellite, which is a very important maneuver, will probably involve such inertia products.

Acknowledgment

The authors would like to thank A. Ramos of COMSAT *Laboratories for the technical suggestion which initiated the concept several years ago. Also, thanks go to G. J. Adams of Hughes Aircraft Company, Space and Communications Group, for his suggestions regarding physical interpretations.*

Glossary

$\dot{\vec{a}}$	Inertial acceleration of coordinate reference point 0
$\hat{a}_y, \hat{a}_r, \hat{a}_p$	Unit vectors along the yaw, roll, and pitch axes, respectively
b	Distance between the damper axis and reference point 0
c	Damping constant of the damper
$\vec{h}$	Total system angular momentum about point 0
$\bar{\bar{I}}$	Instantaneous system inertia dyadic
I_w	Moment of inertia of the wheel about its bearing axis
I_y, I_r, I_p	Moments of inertia of the system about the yaw, roll, and pitch axes, respectively
k	Spring constant of the damper
m	Mass of the damper
$\vec{M}$	External applied torque
M_P	Mass of the platform excluding damper
M_w	Mass of the rotor or wheel
p, r, y	Subscripts referring to pitch, roll, and yaw, respectively
P, w	Subscripts referring to the platform and wheel, respectively
$\vec{s}$	First moment of system mass about point 0
T	Motor torque applied to the wheel
z	Damper displacement from equilibrium
θ	Nutation angle
θ_f	Mean nutation angle after wheel spin-up
$\theta_{\max}$	Peak value of the nutation angle after $\theta_{\min}$ is reached
$\theta_{\min}$	Minimum value of the nutation angle during wheel spin-up
μ	Ratio of damper mass to system mass
$\vec{\omega}$	Angular velocity of the vehicle
ω_p	Pitch rate of the platform
ω_r	Roll rate of the vehicle
ω_y	Yaw rate of the vehicle
Ω	Angular speed of the wheel with respect to the platform

References

[1] L. Muhlfelder, RCA Astro-Electronics Division, private communications.

[2] P. M. Barba, N. Furumoto, and I. P. Leliakov, "Techniques for Flat-Spin Recovery of Spinning Satellites," AIAA Guidance and Control Conference, Key Biscayne, Florida, August 1973, AIAA Paper No. 73-859.

[3] P. M. Barba and J. N. Aubrun, "Satellite Attitude Acquisition by Momentum Transfer," AAS/AIAA Astrodynamics Specialist Conference, Nassau, Bahamas, July 1975, AAS Paper No. 75-043.

[4] J. R. Gebman and D. L. Mingori, "Perturbation Solution for the Flat Spin Recovery of a Dual-Spin Spacecraft," AAS/AIAA Astrodynamics Specialist Conference, Nassau, Bahamas, July 1975, AAS Paper No. 75-044.

[5] M. H. Kaplan and T. C. Patterson, "Transverse Wheel Spin-Up Maneuver for Attitude Acquisition," *EASCON '74 Record*, October 1974, pp. 151–158.

[6] P. W. Likins, "Attitude Stability Criteria for Dual Spin Spacecraft," *Journal of Spacecraft and Rockets*, Vol. 4, No. 12, December 1967, pp. 1638–1643.

[7] G. Grubin, "Dynamics of a Vehicle Containing Moving Parts," *Transactions of ASME, Journal of Applied Mechanics*, Vol. 29, September 1962, pp. 486–488.

[8] R. Pringle, "On the Stability of a Body with Connected Moving Parts,' *AIAA Journal*, Vol. 4, No. 8, August 1966, pp. 1395–1404.

Marshall H. Kaplan is Associate Professor of Aerospace Engineering at the Pennsylvania State University and a consultant to COMSAT *Laboratories and Princeton University on satellite dynamics and control. His teaching and research interests include astrodynamics, spacecraft autopilots, propulsion, and planetary gravity analysis. Recently, he has completed a textbook entitled* Modern Spacecraft Dynamics and Control, *to be published in 1976 by John Wiley and Sons. Dr. Kaplan received a B.S. in aeronautical engineering from Wayne State University, an S.M. in aeronautics and astronautics from M.I.T., and a Ph.D. in aeronautics and astronautics from Stanford University. He is an Associate Fellow of AIAA, a member of Sigma Xi, and a Founder Member of the American Academy of Mechanics.*

Thomas C. Patterson received an A.B. degree in physics from the University of Pennsylvania in 1966 and an A.M. degree in physics from Harvard University in 1968. In February 1968 he joined COMSAT, *where he is presently a member of the technical staff in the Stabilization and Structures Department of the Spacecraft Laboratory. He is concerned with problems of satellite attitude dynamics and is currently studying the interaction of structurally flexible satellite appendages with attitude control systems.*

8/08: 2 CH [illegible] Paul Watt FET
[illegible] microwave switch matrix (MSM)
~~[illegible]~~ SW8 MMIC dual gate FET;
[illegible] Ku-band 24-horn phased array SSPA
[illegible]

Index: communications satellites, orbital rendezvous, maintenance

Repairing a communications satellite on orbit

G. D. GORDON AND W. L. DEROCHER, JR.

(Manuscript received September 29, 1975)

Abstract

Communications satellite systems have matured to the point at which a satellite must operate for many years. The development of the shuttle and tug will soon make the geostationary orbit more accessible. This paper describes satellite repair in orbit and the benefits which it can provide. A serviceable satellite can be built with replaceable modules, and an unmanned servicer can rendezvous with the satellite and replace any failed modules. Body-stabilized communications satellites can be designed to be repairable with an estimated 25-percent increase in weight and an 8-percent increase in production costs.

Benefits of servicing or repair would include increased satellite availability (up to 99.99 percent), increased reliability, decreased life cycle costs, replacement of worn-out items, installation of updated equipment, and correction of design failures. The savings for a representative group of communications satellites (41) would amount to 15 percent. The use of servicing instead of replacement could result in cost savings in excess of 40 percent. Much remains to be done in terms of designing, developing, and proving the effectiveness and reli-

This paper is based on a joint study performed at COMSAT Laboratories under NASA Contract NAS 8-30849 and at Martin Marietta, Denver Division, under NASA Contract NAS 8-30820. This paper does not necessarily reflect the opinions of the Communications Satellite Corporation or of Martin Marietta.

ability of such a servicing system, but the advantages are there and will increase with time. By the end of the next decade on-orbit servicing of satellites may well be established as the preferred method of operating a communications satellite system.

Introduction

A number of studies [1]-[5] have investigated the idea of sending an unmanned servicer (repair system) to geostationary orbit to exchange modules on a communications satellite. This may at first appear impractical, but it is generally agreed that the procedure is technically feasible. The more difficult question is whether such a procedure should be implemented. To answer this question, many factors must be examined, including the transportation available, the length of time before the satellite becomes obsolete, the repair approach, the costs incurred, and the effect on system performance. This paper describes methods of servicing and the resulting benefits.

Repair or servicing of satellites will become more attractive in the next 10 or 20 years. The development of the shuttle and tug will offer new possibilities for transportation into space. The full-capability tug, planned for the next decade, can lift 3,180 kg into geostationary orbit, with capabilities for rendezvous and docking with spacecraft. The number of communications satellites in geostationary orbit is increasing so that over 40 are expected in the mid-1980's. The design and use of communications satellites are maturing so that operations for 10 and perhaps 15 years will be desirable before the satellites become obsolete.

The initial step in the implementation of the servicing concept is the design of a satellite with most components in replaceable modules. The major nonreplaceable components are the structure, electrical harness, antennas, and solar array. The replacement modules are loaded into a stowage rack, which is mounted on the tug and carried into low orbit by the orbiter and into high orbit by the tug. The tug docks with the body-stabilized satellite, and the servicer exchanges modules between the stowage rack and the satellite. Alternatively, a free-flying servicer can service a number of satellites and be resupplied periodically by the tug.

The advantages of on-orbit servicing are greatest when there are many similar spacecraft in orbit, when the program time is long relative to the satellite mean time to failure, and when the satellite availability requirement is high.

Need for servicing

Since failures have occurred on communications satellites in the past, it is reasonable to assume that similar failures will occur in the future. Table 1, which lists some past failures and anomalies, includes failures that resulted in mission failure, subsystem failures that did not affect the main mission, and unexpected anomalies. In the column labeled "type", events are classified in terms of reliability. A "design" failure occurs early in life; its identification shows that the reliability was not as high as had been planned. (This may result from an actual error in design or it may be associated with quality control.) A "random" failure occurs at any time, and a single occurrence does not change the estimate of future reliability. A "wear-out" failure is one that is expected late in the planned life of the satellite.

A striking feature of Table 1 is the large number of "design" failures. Any study of the reliability of communications satellites should deal with possible design failures, yet the usual treatment is an elaborate reliability model based on random failure rates and the assumption that no design failures will occur.

Communications satellites in geostationary orbit will represent a significant part of the space shuttle missions of the 1980's. Table 2 provides an

TABLE 1. POSSIBLE NEED FOR SERVICING IN COMMUNICATIONS SATELLITES

Subsystem	Component	Failure Type
Communications	Receiver	Design
	Transponder	Random
Electric Power	Solar Array Bearings	Design
	Battery	Random
	Power Conditioning	Random
Positioning	Low Orbit (launch vehicle)	Random
	Fuel Depletion	Wear-Out
	Thruster	Design
Orientation	Initial Erection	Design
	Propellant	Design
	Propellant Relief Valves	Design
	Earth Sensor	Design
Telemetry and Command	Decoder	Design
	Encoder	Random
	Telemetry	Random

TABLE 2. AVERAGE NUMBER OF COMMUNICATIONS SATELLITES IN THE 1980's (excluding Russian and DOD satellites)

Type of Service	Total	Atlantic Ocean	North America	Pacific Ocean	Indian Ocean
International Communications	9	4		2	3
Domestic Communications	7		7		
Foreign Communications	12	3	3	5	1
Civil Transoceanic Aviation	3	1		1	1
Maritime Carriers	4	2		1	1
Disaster Warning	2		2		
Tracking Data Relay (TDRS)	3		3		
Communications R&D	1	1			
TOTAL	41	11	15	9	6

estimate of the average number of communications satellites (operational or on-orbit spares) expected, excluding Russian and U.S. military satellites. Other estimates [6], [7] differ in detail, but the total is still substantial. In addition to these satellites, a half dozen earth observation satellites which will have similar requirements are expected in geostationary orbit.

Design of a serviceable communications satellite

The serviceable communications satellite will be designed with replaceable modules to ease the requirements imposed on the unmanned servicer. The major components that will not be modularized (and hence will not be replaceable) are the structure, harness, solar array and drive, and antennas. The subsystems described herein are those used in present designs because they are more familiar to most readers; pursuant to later developments some of these will be changed by the time a serviceable communications satellite is built.

To design a serviceable satellite, many choices must be made. In the design presented herein there are 20 modules; other studies have used fewer but heavier modules. In addition, it has been decided to have the servicer dock on the anti-earth side. In another design of a serviceable communications satellite [5], one of the two solar arrays has been deleted so that the servicer can dock on the south side. Finally, all modules in the design described herein have identical dimensions; it is also possible to use modules of various sizes.

The communications subsystem is based on 48 transponders divided into eight modules with six transponders per module. Traveling wave tubes have been assumed to have an RF output of 6 W, a heat dissipation of 14.8 W, and a maximum allowable collector temperature of 45°C. The radiator requirement for this module has been the determining factor for the area of the module facing north or south. A newer type of traveling wave tube, a dual collector unit, would result in a smaller module.

Other modules for which dimensions are critical are the attitude control module and the propulsion module. The attitude control module contains an externally gimbaled momentum wheel which just fits into the module dimensions. Future developments, such as skewed reaction wheels, an internally gimbaled momentum wheel, or a much higher speed momentum wheel, would allow a smaller module. The four propulsion modules each contain about 50 kg of hydrazine to support a 7-year mission ($\Delta V \approx 400$ m/s), including stationkeeping and attitude control.

On the basis of these factors, the module size has been chosen as 40 x 60 x 90 cm, which allows for latch and attach mechanisms and results in a satellite that fits into the shuttle cargo bay. The module weights are given in Table 3, and the weight of the entire spacecraft is shown in Table 4. The weight penalty for a modularized satellite, as opposed to an expendable non-modularized satellite, is estimated at 20 to 30 percent.

TABLE 3. SERVICEABLE SATELLITE MODULE WEIGHTS (present technology)

Module	Component Weight (kg)	Structure, Harness, and Connectors (kg)	Latch/Attach Mechanisms (kg)	Total Weight (kg)	Number of Modules	System Total (kg)
TWT	28	5	6	39	8	312
Receiver	30	6	6	42	1	42
Attitude Control	28	5	6	39	2	78
Battery and T&C	35	6	6	47	2	94
Battery and Converter	28	6	6	40	2	80
Propulsion	55	10	6	71	4	284
						890

The thermal design of the satellite assumes little, if any, heat transfer between the module and the structure. The temperature of each module is determined by its internal heat dissipation and the area of the side that

TABLE 4. WEIGHT OF A SERVICEABLE SATELLITE*

Modules	890
Structure and Harness	145
Temperature Control	25
Solar Array	55
Antenna and Feeds	110
Total	1,225 kg

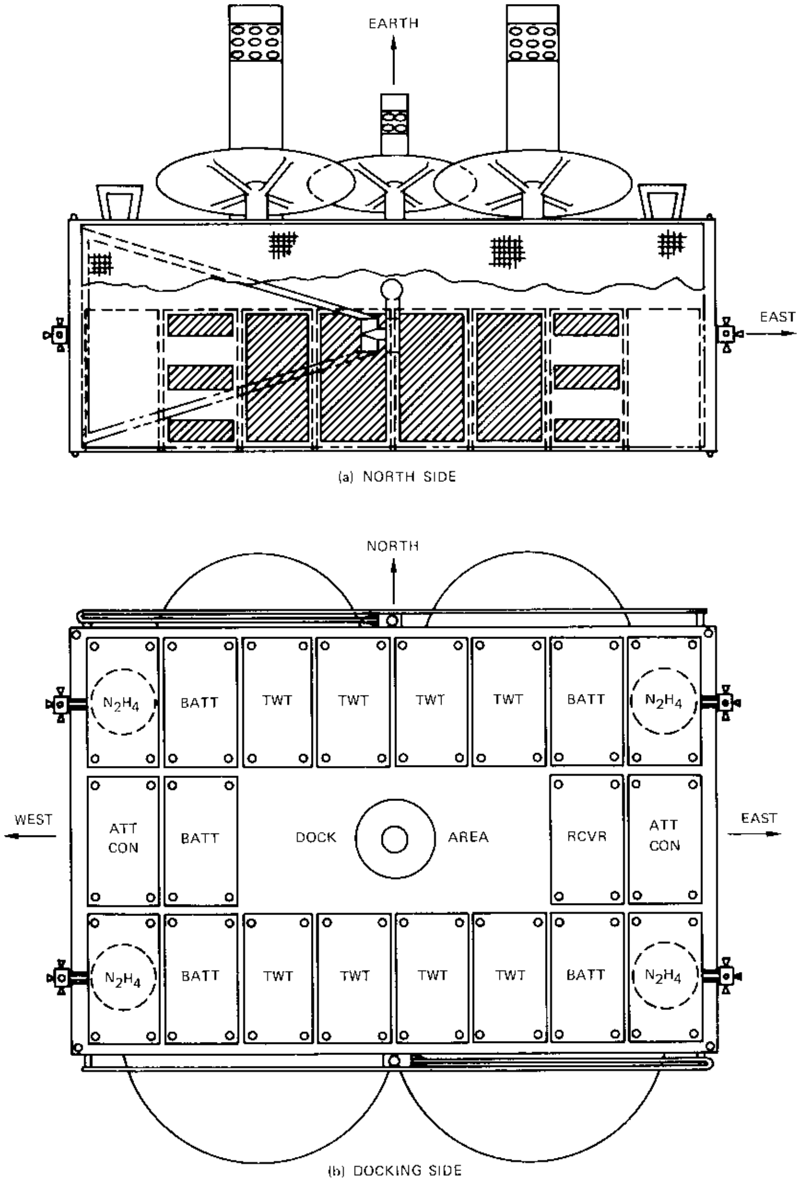
*The weight penalty for this satellite as opposed to an expendable non-modularized satellite (950 kg) is 29 percent.

faces north (or south). To minimize temperature gradients within the module, components that dissipate a large amount of heat should be mounted on the inside surface of the radiator plate. This plate should be fairly thick to diffuse the heat from the components uniformly over the radiator surface. The latch/attach side of the module should also be reasonably thick to take the loads required to mate the connectors. These two requirements can be integrated by mounting the latch/attach mechanism on one of the east (or west) sides of the module. Components can be mounted on both the inside of the radiator and on the latch/attach side; these two sides will then form the backbone of the module structure.

Figure 1 is a configuration of a serviceable satellite. The modules with more power dissipation (especially the transponder or TWT modules) can radiate heat directly to outer space in a north or south direction. The other five faces can be insulated; however, if a neighboring module is to operate at the same temperature, insulation may be unnecessary.

The propulsion modules are located in the four corners for maximum effectiveness of the thruster assemblies. Each module contains one hydrazine tank, valves, filters, and a thruster assembly of five thrusters. There are no fuel lines to be connected when replacing modules. The total of 20 thrusters on the satellite provides full functional redundancy. All attitude control and stationkeeping modes are possible with one of the four propulsion modules inoperative.

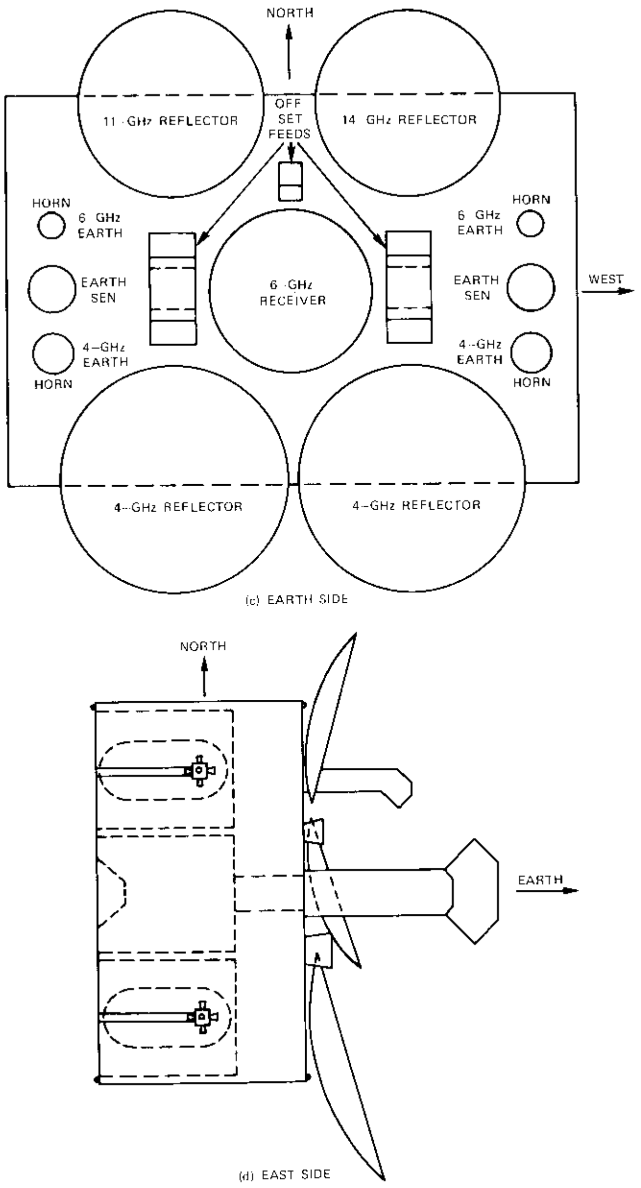
For compactness, the attitude control modules and receiver module have no direct thermal radiation into space. These low-power modules can transfer their heat load to the satellite structure and to other modules. The attitude control modules contain sensors and a processor as well as momentum storage devices. The earth sensor must be able to look through the satellite structure to view the earth. Either attitude control module will be able to control the satellite attitude.



(a) NORTH SIDE

(b) DOCKING SIDE

Figure 1a. *Serviceable Communications Satellite Configuration*



(c) EARTH SIDE

(d) EAST SIDE

Figure 1b. *Serviceable Communications Satellite Configuration*

The satellite configuration represents a compromise among many factors. A communications satellite requires several antennas pointed toward the earth. Three-axis stabilization has been chosen for ease of servicing. The deployed solar arrays must face the sun so that the drive axis must be north-south. The anti-earth face has been chosen as the docking face because there is no interference with the earth-pointed antennas or with the solar arrays on the north and south faces, and because all modules can be reached from a single docking. The north and south faces are used for radiator surfaces because the maximum heat input occurs when the sun is shining 66.5° off the normal, and during half the year the sun never shines on the surface.

The telemetry and command signals will be distributed on a few data buses to decrease the number of electrical connections to the modules. All the connectors will be grouped together, and the latch/attach mechanism will provide the forces needed to connect or disconnect. There may be as many as eight RF connectors to the receiver module and four to each of the eight TWT modules. Where possible, coaxial connectors will be used, but, if necessary, waveguide connectors are feasible. A possible waveguide connector, shown in Figure 2, includes a short length of flexible

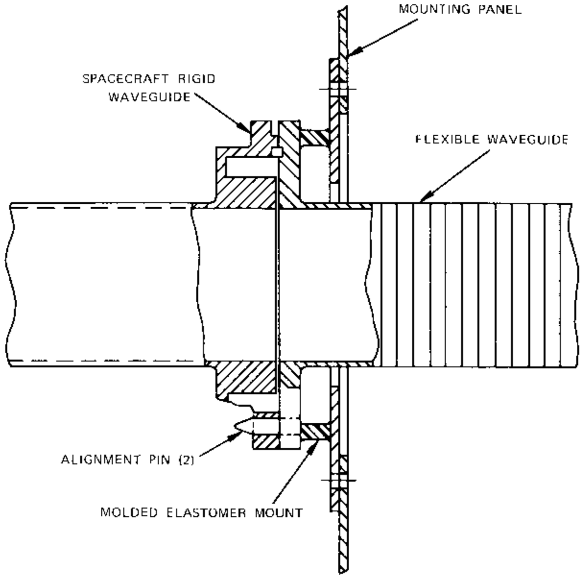


Figure 2. *Waveguide Alignment Compensation Device*

waveguide, two alignment pins, and an RF choke coupling with a crushable gasket to prevent RF leakage.

Design of the on-orbit servicer

The purpose of an on-orbit servicer is to remove failed modules from a spacecraft and to replace them with good modules. Many different concepts have been advanced for performing these mechanical operations. The servicer shown in Figure 3 is simple, versatile, and lightweight and has a short stowage length in the orbiter cargo bay. It can be used on tug as well as orbiter-only missions so that its development and operations costs can be spread over a wide range of satellites. Although Figure 3 also shows a docking mechanism for reference and to permit easier visualization of the mechanical interface aspects, the servicer itself has only two major components: a pivoting arm servicer mechanism, and a stowage rack for module transport. The servicer mechanism and the stowage rack have been designed separately with interfaces for individual removal and replacement. This permits simple removal for maintenance and for quick ground reconfiguration. Stowage racks can be configured and loaded for particular flights prior to attachment to the carrier vehicle.

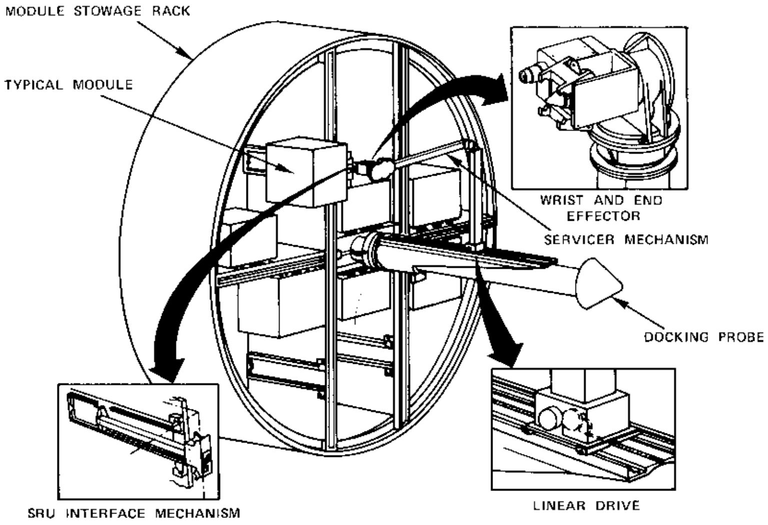


Figure 3. *Pivoting Arm On-Orbit Servicer*

A simple 4-degree-of-freedom approach is used for the servicer mechanism. The servicer has an operating length of 1.5 m and a stowed length of 53 cm forward of the module stowage rack. Its overall stowed length is thus 1.55 m. These short stowed lengths are possible because the arm and docking probe can be folded flat against the front of the stowage rack. The servicer mechanism can handle modules up to 320 kg and 1 m^3 in dimension. The four degrees of freedom permit any module to be placed anywhere on the end of a 4.5-m-diameter satellite and in any roll orientation.

The number of modules exchanged during the servicing of a single satellite will generally be less than six. The stowage rack has been configured for a mix of 18 modules in three representative sizes. The specific configuration can be selected for each mission. Thus, space is available in the stowage rack for the modules needed to service up to three spacecraft on a single mission. The stowage rack is a continuation of the 4.5-m-diameter tug outer skin.

The servicer mechanism has a tip force greater than 90 N in the worst configuration. This force level is representative of good engineering practice for a servicer-type space application. It is significant to note that the tip force requirement interrelates with the attach/latch actuator located in the end effector. The large forces required for making and breaking connectors are generated by the end effector actuator and do not impose forces on the servicer mechanism arm. A simulation has shown that this tip force level, which is compatible with that for the shuttle remote manipulator system, is reasonable for module exchange. The servicer can operate at 1 g with an empty module to permit ground check-out both on a subsystem level after fabrication and at the launch facility.

The servicer mechanism weighs 62 kg and a representative empty stowage rack weighs 220 kg. The stowage rack has been designed to take the launch and crash loads of a 1,500-kg satellite mounted on the front of a fully loaded rack. If these requirements are decreased, the stowage rack weight can be significantly reduced.

A 2-position pitch drive, located at the wrist end of the outer arm, turns the modules end for end so that they may be placed in the satellite or stowage rack. The drive is thus an indexing rather than a servo-controlled drive. The end effector, designed to mate with the interface mechanism, attaches the servicer mechanism to the module and operates the latching mechanism. End effector attachment is accomplished by two closing jaws grasping the rectangular baseplate grip. The closing force is supplied by a motor-driven ball screw drive which applies a low initial closing force during radial alignment and a very high final clamping force during module handling.

The servicer mechanism must attach to each module so that it can transport the module between the stowage rack and the satellite. Each module must be latched into both the satellite and the stowage rack. Associated with these operations are status indications and mating/demating of several types of connectors. Since there is one interface mechanism for each module, the cumulative weight and volume effect is important. There is also an important interaction among the interface mechanism "capture volume," the servicer mechanism accuracy, and the servicer control system.

One form of module interface mechanism is shown in Figure 4. The light-colored part is the baseplate that fastens to the module, while the gray part is the baseplate receptacle (module track) that is mounted in the satellite and the stowage rack. The cams, links, and rollers guide the baseplate into the receptacle and hold it there. The end effector locates via the round hole in the right-hand end and grips via the vee edges adjacent to the hole. The links, bell cranks, and rollers are driven via the screwdriver-like slot adjacent to the end effector attachment. The baseplate latch drive mechanism is an integral part of the end effector attach drive.

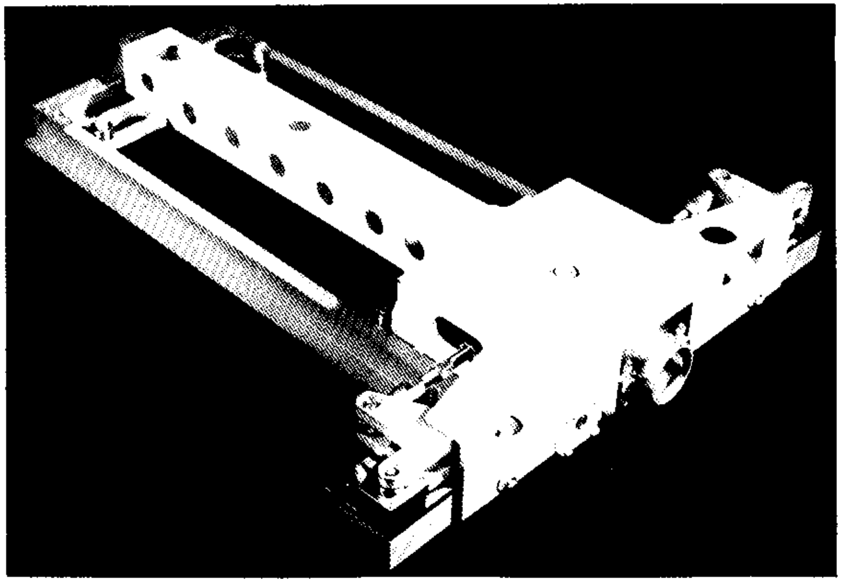


Figure 4. *Side-Mounting Module Interface Mechanism*

It is useful to think of the interface mechanism as a 2-part kit, one part to be mounted on the satellite and the other to be mounted on the module. Thus, each satellite/module designer will know the type of interface and will not need to develop a new latch concept for each satellite. Instead, one kit concept may be used across all satellites. The variety of module sizes may dictate latch kits of several sizes, but they can all use the same basic design and the same attach interface.

The control system selected for use with the on-orbit servicer mechanism can strongly affect the servicer's operational utility and versatility. If the control system is too limited in its capability, then the on-orbit servicer system will also be limited. Conversely, the control system need not have a comparatively greater capability than the servicer mechanism.

The suggested control system [3], defined in Table 5, combines the best qualities of each of two modes. Thus, it overcomes each of their deficiencies by using automatic control as the primary mode and remotely manned control to provide backup operation for failures and operational contingencies. Because the remotely manned control is only a backup mode and will not be used frequently, longer operating times can be accepted. This permits the use of one or more simplified TV cameras with very low frame rates (3 per minute) as well as the use of the TV system instead of proximity sensors for the alternative hazard avoidance system in this backup mode. Tolerance compensation can be handled by the operator using his ground-based computer. The major advantage of this combined mode is the availability of different and completely separate

TABLE 5. CONTROL SYSTEM CHARACTERISTICS

- Primary Mode: Automatic Control
- Backup Mode: Remotely Manned Control
- Stored/Interpolated Module Trajectories
- Hazard Avoidance
    - Supervisory: Precalculated
    - Remotely Manned: TV and Ground Computer Graphics
- System Errors Measured Manually and Biased Onboard
- Separate Translation and Rotation Hand Controllers
- TV and Mechanism Position Displays
- Mechanism Joint Control
    - Supervisory: Position
    - Remotely Manned: Rate
- TV Refresh Rate: 3 per minute

backup functions to obtain the highest probability of successful module exchange over the widest range of operating conditions.

Another servicer mechanism and satellite configuration, shown in Figure 5, is one of the 15 discussed in Reference 1. This is a full 6-degree-of-freedom servicer mechanism in which the initial degree of freedom is a track running around the periphery of the stowage rack. The stowage rack is mounted to the tug at the right of the figure. The satellite is a serviceable version of the Synchronous Earth Observation Satellite. The modules are removed/replaced radially from the exterior cylindrical surfaces of both the satellite and the stowage rack. Failed modules are temporarily stowed by using a double-ended end effector. The docking system is of the peripheral type, although the concept can be used with a central docking mechanism. As opposed to the pivoting arm on-orbit servicer, the general-purpose manipulator on-orbit servicer is heavier, longer, harder to use in the orbiter cargo bay, and more expensive.

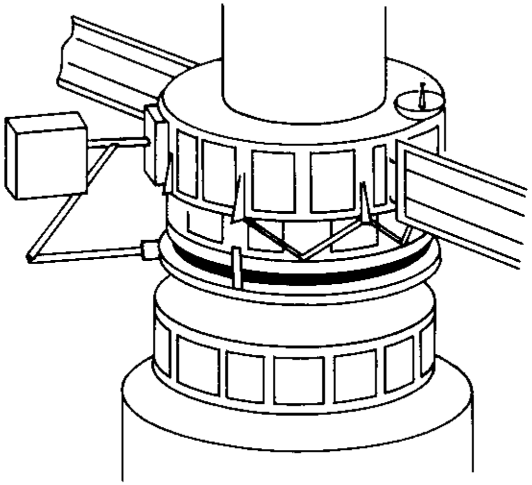


Figure 5. *General-Purpose Manipulator On-Orbit Servicer*

The operation of module exchange to repair satellites and the design details of an on-orbit servicer are generally independent of the carrier vehicle. While the space tug is the usual carrier vehicle for geostationary operations, and the orbiter is the usual vehicle for low-earth-orbit operations, one other alternative is the geosynchronous free-flyer. It consists of an on-orbit servicer, additional module stowage capability, electrical power system, 2-way communications, attitude control, rendezvous and docking system, and propulsion capability. The free-flyer is designed to remain in orbit for three years and to be resupplied by the space tug once each year. With its ability to remain in orbit for such a long period of time, it can service many satellites as it works around the earth, taking most of the year to do so and using a modest amount of propellant.

Figure 6 shows how the free-flyer may progress through its servicing activities for 30 satellites over a 3-year period. The velocity budget is based on the assumption that it will visit 10 spacecraft in a 4-month period. The locations of these spacecraft are based on the probable distribution of communications satellites in the late 1980's. Velocities have been selected to minimize the fuel used on this tour (i.e., higher velocities have been used for long jumps and smaller velocities for short jumps). Hence, a velocity increment of only 100 m/s is required to service all of the satellites. (An allowance of 10 percent of the free-flyer dry weight for propellants would provide a 200-m/s capability. Thus, a reserve for demand servicing and return to the resupply tug longitude would be available.) The free-flyer can also be resupplied by using excess capacity on space tugs scheduled to geostationary orbit for other reasons. The differences in longitude of free-flyer operations and tug targets will require proper scheduling so that the tug and free-flyer can rendezvous when the tug's primary mission is complete.

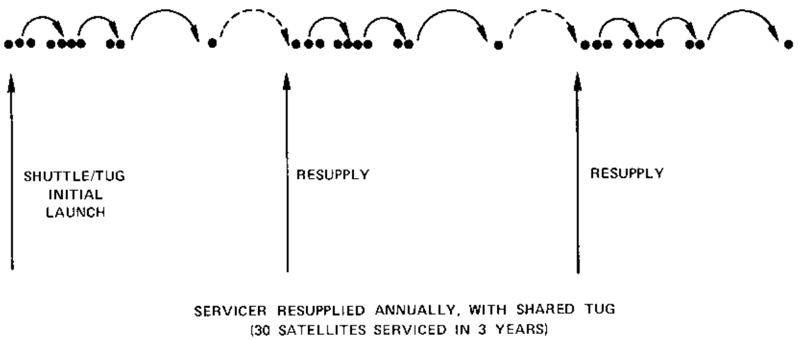

SHUTTLE/TUG
INITIAL
LAUNCH
RESUPPLY
RESUPPLY
SERVICER RESUPPLIED ANNUALLY, WITH SHARED TUG
(30 SATELLITES SERVICED IN 3 YEARS)

Figure 6. *Free-Flying Servicer*

Servicing operations

While on-orbit servicing operations can be developed in conjunction with expendable launch vehicles, they are discussed here in terms of the reusable launch vehicles of the space transportation system (STS). The design and development of the orbiter and its launch and operations support have reached the stage at which firm planning for its use has commenced. The spaceborne parts of the STS which are required for geostationary orbit servicing operations are the orbiter and the full-capability tug.

The space shuttle flight system is composed of the orbiter, an external tank containing the ascent propellants to be used by the orbiter main engines, and two solid rocket boosters. The solid rocket boosters and the orbiter main engines fire in parallel at lift-off. The boosters are jettisoned and the cases recovered for reuse, while the external tank is jettisoned and not reused. The orbital maneuvering system is used to put the orbiter in the desired orbit. The orbiter, shown in Figure 7, delivers and retrieves payloads, conducts orbital operations, and returns to a land base similar to that of a high-performance aircraft. It is reusable and designed to operate in low-earth orbit for missions of up to 7 days. The basic crew consists of a commander and a pilot; other crewmen such as mission and payload specialists are used as mission requirements dictate.

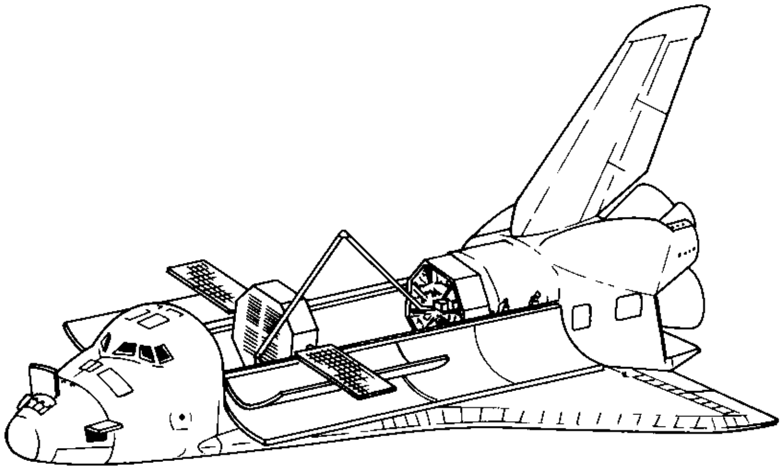


Figure 7. *On-Orbit Configuration of the Orbiter*

A number of orbiter characteristics related to orbital servicing are listed in Table 6. All payloads must be designed to withstand crash loads while maintaining structural integrity because of the possibility of a launch abort. The orbiter is also capable of deploying and retrieving payloads by means of the shuttle remote manipulator system. Shrouds over the payloads provide contamination protection.

TABLE 6. ORBITER CHARACTERISTICS

Scheduled Initial Operating Capability	1980
Mission Duration	7 days
Cargo Bay Length	18.3 m
Cargo Bay Diameter	4.6 m
Launch Performance	29,500 kg into 400-km orbit
Reentry Performance	14,500 kg
Crash Loads	10 g
Random Vibration	0.09 g^2/Hz
Acoustic Noise	145 dB re 20 μN/m^2
Caution and Warning Data Required from Payloads	

The full-capability tug is shown in Figure 8 as it might be configured for a servicing mission. The tug is 9.1 m long and 4.5 m in diameter. It holds 22,800 kg of cryogenic propellants, has a vacuum thrust of 6,800

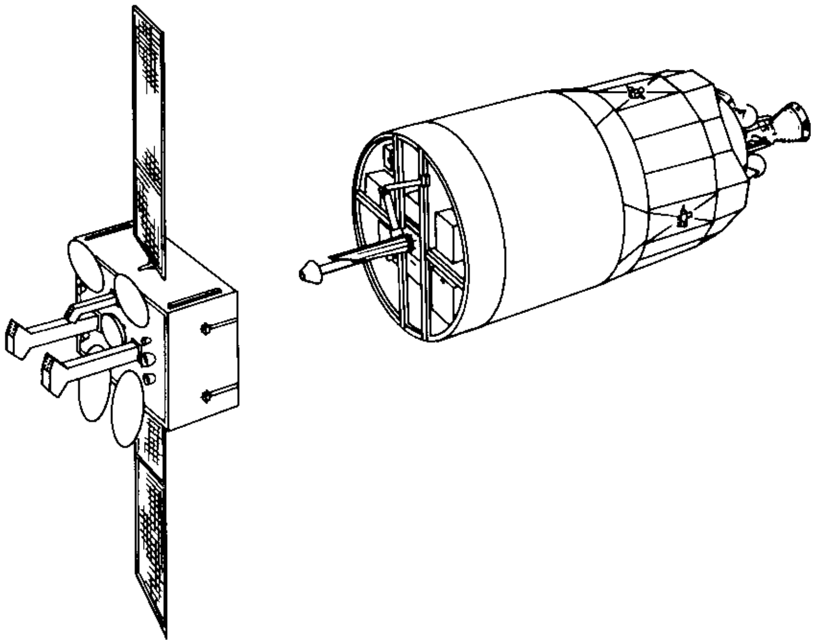


Figure 8. *Full-Capability Tug on Servicing Mission*

kg, and is completely reusable. It has a rendezvous and docking capability with payloads and with the orbiter. It can provide attitude control, electrical power, and communications capability to the attached payloads. The full-capability tug can take 3,180 kg to geostationary orbit or return 1,360 kg, and has a round trip capability of 950 kg. Its initial operating capability is scheduled for 1983.

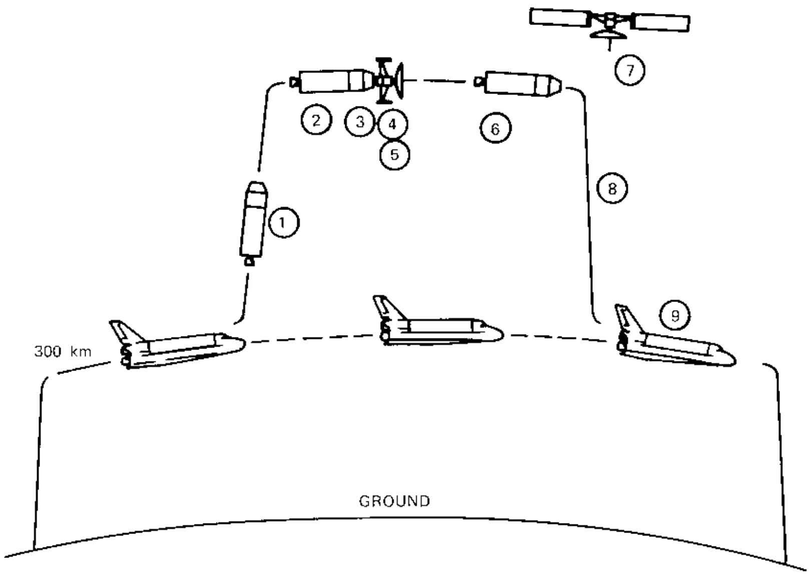
A servicing operation is initiated with the decision to service a satellite and the selection of specific satellites to be serviced and modules to be replaced. The health of each satellite in orbit is continuously monitored by the user. It is the user's responsibility to identify the modules causing satellite failures. Loss of one of the parallel paths in a redundant system may not be sufficient cause to initiate a servicing mission. However, each of the partial failures is tracked until the number of partial failures is sufficient or a total failure occurs. The satellites to be serviced and the modules to be replaced are selected on the basis of a set of priorities called a servicing strategy. Considerations include the following:

a. failed modules,
b. partial failures,
c. consumables remaining,
d. satellite age,
e. program time remaining,
f. satellite locations in orbit,
g. total weight of modules,
h. module availability,
i. orbiter and tug schedules,
j. mission equipment updating desired.

The cost effectiveness of the mission increases with the number of functions which can be performed. Before a launch, simulations of module exchange and verifications that module replacement will correct the failure can be conducted.

Once the complement of modules has been selected, the stowage rack loaded, and the on-orbit servicer mounted on the tug and checked out, the scenario shown in Figure 9 is conducted. The actual time intervals will depend on the orbit phasing required to match longitudes on the way to geostationary orbit and on return to the orbiter. The times to go from one satellite to another will depend on longitudinal spacing and tug propellant allocations. The 164-hour total is a constraint based on allowable tug and orbiter mission durations.

Docking of the on-orbit servicer to the satellite can be performed automatically, but the operation is more likely to be monitored and partially controlled by a ground operator. Rendezvous techniques have been developed as part of the Apollo program, while unmanned docking is being developed by NASA. The docking operation requires that the satellite be 3-axis body stabilized, which implies a highly reliable satellite attitude control system. Docking with spinning satellites is possible, but seems to be an unnecessary requirement. After docking, the satellite attitude control system is shut down and the tug attitude control system controls both the tug and the satellite.



	EST. TIME (HR)
1. Tug separates from orbiter and transfers servicer to high orbit	22
2. Circularize at geosynchronous altitude and trim orbit	11
3. Tug/servicer rendezvous and docks with satellite	8
4. Servicer performs maintenance activities	100
5. Satellite orientation, activation and preliminary checkout	1
6. Tug/servicer separation from satellite	1
7. Final satellite checkout	2
8. Transfer of servicer to 300 km orbit	15
9. Orbiter rendezvous and retrieval of tug/servicer	4
	164

Figure 9. *Geostationary Orbit Servicing Mission Scenario*

The servicing operation includes the following steps:

a. unstowing of the servicer mechanism,
b. removal of a failed module from the satellite,
c. stowage of the failed module (temporarily or permanently),
d. location of the replacement module in the stowage rack,
e. placement of the replacement module in the satellite,
f. verification that the replaced module is latched in place,
g. permanent stowage of the failed module,
h. repetition as necessary,
i. restowage of the servicer mechanism.

The satellite is then powered up and checks are made to ensure that the replaced modules have been properly installed. The tug orients the satellite, undocks, and backs away; the satellite attitude control is turned on; and the satellite stabilizes itself. Further brief checks of the satellite are made, the tug continues its mission, and the satellite check-out continues until the satellite is operational again. The time required to exchange an average of six modules per satellite is one hour.

The preceding discussion has assumed that the satellite is not operating during module exchange. It may be possible to keep communications satellites operating at a reduced level during module exchange. The docking and undocking impacts can be made low enough that the satellite attitude returns to within limits in a few seconds and is then maintained by the tug. Each module can be turned off individually before it is replaced. While a module is off, its functions are not available but the other modules can continue to operate.

The tug can then go to another satellite or initiate module disposition. When the cost of new modules is compared with module refurbishment costs and transportation costs, it is more economical to leave the modules in geostationary orbit. However, if the need can be justified, it is possible to return modules to earth.

Development of the on-orbit servicing system appears to be the responsibility of NASA. However, there are several options for operating such a system. First, NASA could provide the service for any satellite; this would imply a long-term (20-year) commitment by NASA. Secondly, servicer operations might be limited to a single class of satellite such as all geostationary satellites built by all manufacturers. Such a limitation could reduce on-orbit servicer development and production costs, but would increase the per-service operation costs. Thirdly, a satellite manufacturer could operate an orbital servicing system exclusively for his satellites. He could use orbital maintenance to guarantee satellite availability at a lower cost and thus sell more satellites. The last approach would be a customer-exclusive system in which a satellite operator, such as COMSAT, would operate an orbital servicing system for his own satellites. He could then ensure that the maintenance system would be available for as long as he desired, and it could be updated over the years as his requirements changed.

Benefits of servicing

Availability

When a satellite fails in orbit, it can usually be repaired by module exchange. However, failures of communications satellites are rarely a simple transition from "now it's working" to "now it isn't." The benefits of servicing to a communications satellite involve considerably more than simply extending its lifetime after a satellite failure. In fact, if the capabilities of servicing are fully exploited, a satellite will rarely fail, and most servicing operations will be performed on satellites that are still capable of fulfilling most of their mission objectives.

Efforts to increase the availability of a satellite usually focus on decreasing the unavailability or down-time of the satellite, and more properly, of the satellite system. A goal [8] for unavailability of 1×10^{-4} corresponds to a down-time of less than one hour per year or five hours over an expected satellite lifetime. This level of down-time can be obtained only by providing redundancy in orbit, spare satellites, spare equipment, and redundant elements. Even this is not enough unless the failed redundant paths are replaced. It is through repair of these redundant elements that on-orbit servicing can help provide high availability at reduced cost. While the cost of providing on-orbit servicing can be determined, the value of increased availability and confidence that the availability will be obtained and retained is more difficult to determine.

The conventional uses of module exchange include the following:

a. repair of failed satellites,
b. repair of redundant elements,
c. replacement of expendables and worn-out items.
d. repair of design failures,
e. replacement of equipment with newer models.

There may be other ways in which the satellite designer can take advantage of on-orbit servicing to obtain the most cost-effective system.

Replacement of expendables and worn-out items may be planned as part of a serviceable system. For example, it is possible to load a 4- or 5-year supply of hydrazine fuel in the satellite and then resupply it periodically instead of loading all the fuel initially. Several components such as batteries, bearings, and traveling wave tubes have limited lifetimes; when they are approaching the end of their lifetimes they can be replaced. Thus, the lifetime of a serviceable satellite may be extended beyond that possible with expendable satellites.

Servicing becomes more attractive if the ratio of program life to satellite life is increased. If satellites can be built with an average life of 7 years, and if they become obsolete after 10 years, servicing is not particularly useful. On the other hand, if a satellite is built with a 5-year lifetime (at reduced cost) and does not become obsolete for 15 years, then servicing becomes more attractive. At present, the program life of communications satellites is only 5 to 7 years; in the future, program lifetimes will become longer as the business matures. For most of the communications satellite programs in the 1980's, a program lifetime of over 10 years may be optimistic. However, for the 1990's, it becomes a more realistic assumption. There is a trend toward increasing program lifetimes, and it is only a matter of time before program lifetimes of more than 10 years will be realized.

Most communications satellites in use are not in perfect condition. Design failures may have been discovered on a similar satellite so that the corresponding subsystems are suspect. Some components may have given telemetry indications that something is amiss, or redundant components may have failed. Replacing all these components will increase the probability that the satellite will operate without a failure in the following year. However, because replacement of these components will be costly, a cost-benefit tradeoff will be needed. The repair of design failures makes servicing attractive for two reasons:

a. repairing a satellite early in its design life provides years of additional service,

b. often such a failure suggests servicing of similar satellites in which failures have not yet occurred.

Servicing provides an excellent way of increasing the availability of a satellite. In fact, it can probably improve the reliability of a satellite to the point at which repairing a satellite is far better than replacing it with a new satellite. Preliminary simulation studies indicate that an availability of 0.9999 can be achieved.

Servicing has been compared with alternative methods of improving system availability. A summary of this comparison is provided here; more details are available elsewhere [7]. Various methods of improving the availability of a communications satellite system are shown in Figure 10.

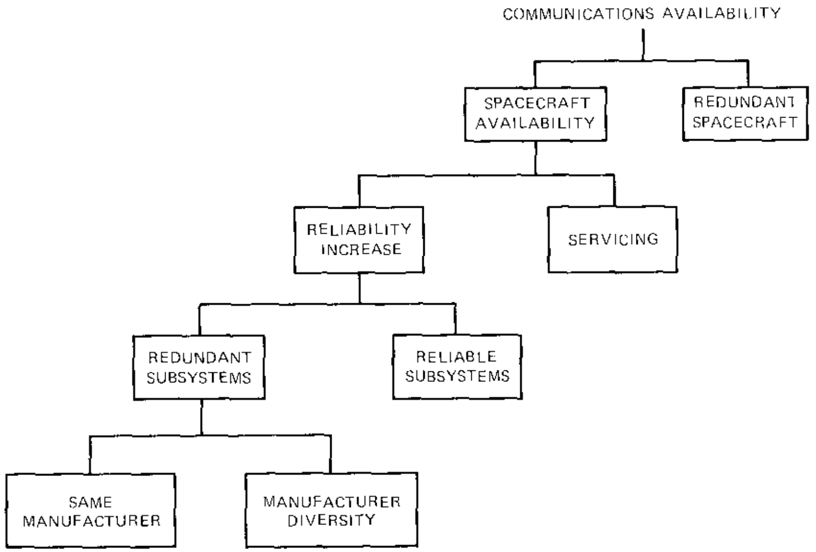


Figure 10. *Approaches to Improving Spacecraft Program Availability*

Those used most extensively are redundant spacecraft, reliable subsystems, and redundant subsystems from the same manufacturer. As described earlier, these methods have failed to eliminate a number of design failures. Adding redundant subsystems beyond the "2" level has only a minor effect on system availability. To reduce significantly the number of design failures, the viable alternatives are servicing, subsystem improvement, or additional redundant subsystems from different manufacturers. Of these, basic improvement of the subsystem is probably the best in the long run and the least expensive, but as the state-of-the-art continues to change, there is a limitation on subsystem reliability. Manufacturer diversity and servicing are probably comparable in cost; a more detailed study would be needed to choose between them. Most expensive is the use of a redundant spacecraft, but this method will be used for a number of years until confidence increases in the use of alternatives.

The allowable repair time is subjective. Usually the urgency is less when the satellite is first injected into orbit because the user is not dependent on the satellite, and traffic requirements (if any) are smaller. On the other hand, when a communications satellite is in operation and a critical subsystem fails, a repair is desirable immediately, preferably within minutes or seconds. However, there are often alternative modes of operation which permit repairs to be delayed. Batteries, for example, are needed only during the eclipse season. Furthermore, battery failures are usually preceded by about a year's warning so that replacement can be scheduled in advance. One conclusion from considerations such as these is that servicing may be delayed for periods of a year or so without excessive availability penalties. This will depend on the degree of redundancy that is built into the satellite; without redundancy, it is not meaningful to service a satellite before failure. With servicing, some redundancy will still be required in the satellite, especially when high availability is required.

In addition to repairing satellites and maintaining them to provide high availability, servicing also makes it possible to update and modify equipment. Within a module, changes can be made as long as the interface remains the same. In some cases, more extensive changes that affect several modules may be made.

The concept of a serviceable satellite can be described for a system with a steady traffic growth. Initially the satellite would be designed for full capacity, but some module spaces would be empty, with a savings of cost and weight. A few years later the satellite would be brought up to full capacity with the additional modules. After several more years the satellite would be replaced by a newer model but would still be maintained as an in-orbit spare. When the satellite was no longer useful for its primary mission, it could be used for another mission whose requirements were less stringent. The satellite could be moved to another location and limited changes in equipment made by using module exchange.

In summary, if servicing is available, most communications satellites can benefit from it. In the first three years of a satellite system, it can be expected that every satellite may require module replacement due to design failures. During the life of a satellite, the probability that servicing will be required to correct a random failure is somewhere between 0.5 and 1.0. If the satellite lifetime is extended beyond present design lifetimes, additional satellite servicing will be required to repair wear-out and additional random failures. Thus, it can be seen that on-orbit servicing can increase program flexibility and satellite reliability, lifetime, and availability.

Cost savings

In addition to the availability and operational benefits discussed in the preceding subsection, significant cost benefits can result from the use of on-orbit servicing [9]-[11]. This is true even when the satellites have not been designed to obtain the maximum benefits from servicing, as in the following example. In this example, satellites have been designed to be reliable and to have a long lifetime relative to the program length. On-orbit servicing is more effective when the satellite lifetime is one-third to one-fifth of the program duration.

The technique, rationale, and details of the cost analysis for these satellites are presented elsewhere [1]. In addition, however, the cost analysis involved a series of conditions. The first was that satellite program duration and availability are held constant for each of the maintenance modes. For the expendable mode, this meant that the number of satellites successfully injected was equal to the number of operating cycles. For the on-orbit maintainable mode, the number of satellites purchased represented the desired on-orbit fleet size, and the number of repair activities was equal to the number of operating cycles less the on-orbit fleet size. All costs were presented in 1975 dollars, and the satellite non-recurring and recurring costs were provided by NASA.

The orbiter and tug basic launch costs were $12 million and $1.1 million, respectively.* However, the launch charges to the satellite programs were based on that portion of the launch vehicle capability actually used, and the basic costs were increased by applying average load factors which raised the basic costs to $17.1 million and $1.3 million, respectively. For the expendable mode, the launch costs were generated for each spacecraft launched, while for the on-orbit maintainable mode, launch costs were generated for launching the on-orbit fleet and for conducting the repair flights. The on-orbit servicer was taken to and from orbit, while the replacement modules were taken only to orbit and not returned.

The satellite non-recurring costs were increased by 4 percent for the serviceable satellites, while the recurring costs were increased by 8 percent. The major savings resulted from the smaller number of production satellites that had to be purchased. Satellite operation costs included satellite launch check-out costs (9 percent of satellite production costs) and the

*These NASA estimates were obtained at the start of the study (1974); they have already increased substantially. Increases in transportation costs tend to favor servicing.

costs of sustaining a cadre of knowledgeable people between the end of production and the last launch or repair activity. The mission model consisted of the set of geostationary orbit communications satellites from the NASA 1974 SSPD. These seven spacecraft programs involved 83 missions, an on-orbit fleet of 41 satellites, and an average operational time of 5.6 years. Generally one satellite replacement or repair activity was planned.

For the on-orbit maintainable mode it was necessary to produce replacement modules which were all expended. Prelaunch check-out of these modules was also included, and the check-out costs were based on the value of the replacement modules launched. The value of the replacement modules was based on the parts factor, or the cost of the replaced modules divided by the satellite unit cost. This ratio was assumed to be equal to the weight of the replaced modules divided by the spacecraft total weight. Although the weight of propellants replaced was not included in this calculation because of the relatively low cost of the propellants, it was used in the launch cost calculations. Both random failures and wear-out effects were included. The random failures were based on detailed reliability models. The wear-out effects included propellant resupply and battery replacement. The parts factors varied from 0.24 to 0.31, with a weighted average of 0.27. The range of parts factors was relatively small because of the similarity in the types of satellites considered.

Loss factors were included in the analysis to represent additional satellites that must be built, launched, or serviced due to potential losses such as shuttle system failures, nonreplaceable unit failures, and on-orbit servicer failures. The average loss factor was 0.26 spacecraft for the expendable mode and 0.54 spacecraft for the on-orbit maintainable mode.

Development of the on-orbit servicer would not be as economical for a small number of satellites; thus, its cost data were based on a larger mission model [1]. The total life cycle costs included design, development, test, evaluation, production, and operations. Operations for 12 years in the eastern test range and 9 years in the western test range were included. The resulting costs for 247 repair missions were as follows:

On-orbit servicer development costs (DDT&E):	$ 29 million
Production:	17 million
Operations:	57 million
Total life cycle:	$103 million

The direct on-orbit servicer costs were $0.42 million per mission or $17.5 million for all the repair missions for the seven spacecraft programs considered. Thus, it can be seen that the direct cost per servicing is small.

The cost analysis is summarized in Table 7. The satellite production costs represent 55 percent of the total satellite program costs in the expendable mode and only 37 percent in the on-orbit maintainable mode. The launch costs are reduced by only 10 percent in the on-orbit maintainable mode. The repair costs are 14 percent of the on-orbit maintainable satellite program costs and the on-orbit servicer costs are less than 1 percent. Overall the on-orbit maintainable mode saves $450 million or 15 percent of the expendable mode costs. This saving varies between $20 million and $160 million or 10 to 20 percent for each of the seven satellite programs.

TABLE 7. COST ANALYSIS SUMMARY

	Program Cost ($million)	
	Expendable	On-Orbit Maintainable
Launch Costs	420	380
Satellite Program		
DDT&E	690	720
Production	1,600	910
Operations	200	80
Repair		
Modules	—	230
Operations	—	120
On-Orbit Servicer	—	20
TOTAL	2,910	2,460

The launch costs are respectively $420 million and $380 million when the orbiter and tug are used for the expendable and on-orbit maintainable modes. If the expendable Atlas-Centaur is considered instead, then the comparable figures are $2,320 million and $1,850 million. The assumptions involved include one spacecraft or two servicings per launch, expended servicer and equipment for rendezvous and docking, a rendezvous and docking capability DDT&E of $30 million with a recurring cost of $2 million, an Atlas-Centaur launch cost of $28 million, and no launch cost sharing. The on-orbit maintainable mode thus saves 20 percent of the expendable Atlas-Centaur launch vehicle costs.

Another expendable launch vehicle alternative is the Delta, which has a launch cost of about $16 million. It is assumed that the weights of these low-cost design spacecraft can be reduced so that they can be singly launched, and that the weights of the on-orbit servicer system and the replacement modules can be similarly reduced so that two servicings per launch can be obtained. On the basis of these assumptions and those listed above, the launch vehicle costs are $1,330 million and $1,100 million for the expendable and on-orbit maintainable modes, respectively, using the Delta launch vehicle. The saving for on-orbit servicing is thus 17 percent of the expendable mode vehicle launch costs. The major difference is in going from the expendable launch vehicles to the shuttle, where the savings are 80 percent for the Atlas-Centaur and 65 percent for the Delta. If expendable spacecraft on expendable launch vehicles (Delta) are compared with repairable spacecraft using the shuttle, then the launch vehicle saving is 71 percent.

The sensitivity of the cost analysis results to variations in the launch vehicle charges has been investigated [1]. The major factors are the launch costs and the variations in load factors. Generally, it can be anticipated that the launch costs will go up and load factors will go down, both of which will result in a higher launch vehicle charge. However, launch charges for both maintenance modes can be expected to vary in the same way. Hence, the launch costs for the on-orbit maintainable mode will still be less, but the savings will be greater. For example, a 25-percent increase in launch vehicle costs and a 25-percent decrease in load factor will result in an increase in launch charge savings from $40 million to $67 million, or 67 percent.

The cost figures can also be thought of in terms of the cost of conducting an additional satellite operating cycle. This operating cycle can be obtained by launching a replacement expendable satellite or by on-orbit servicing of a failed satellite. For the seven communications satellite programs in the expendable mode, Table 8 shows that there are 83 satellites, and that the average recurring costs are $27 million per satellite. For the on-orbit maintainable mode, the recurring cost of servicing a failed satellite is $15 million (including the additional costs to make the satellite serviceable and the on-orbit servicer costs). The satellite development costs have been excluded from both of these figures. The user must choose between launching an expendable replacement satellite for $27 million or repairing a failed satellite for $15 million. The potential saving is $12 million or 44 percent. This significant cost saving is in addition to the other benefits outlined above.

All of these cost estimates must be viewed with caution, in terms of both reliability and interpretation. While they are the result of a year-long study, it is obviously impossible to determine the cost of satellites with unknown specifications to be built 10 years in the future. The percentage saving depends on the area under consideration. As shown in Table 7, the satellite production costs are reduced by over 40 percent for servicing, while the total program costs show only a 15-percent saving. When the cost of a replacement is compared with the cost of servicing, Table 8 shows a 40-percent savings. The results of the study have been written so that the reader can modify the assumptions and calculate any resulting changes [1], [9]. However, the results of any study that covers all communications satellites will not be nearly as accurate as those of a study that concentrates on a single program [10].

TABLE 8. RECURRING COSTS PER COMMUNICATIONS SATELLITE [a]

On-Orbit Fleet Size	41 satellites
Total Number of Expendable Satellites	83
Expendable (per satellite)	
Launch Cost [b] ($million)	5
Satellite ($million)	22
	—
Total ($million)	27
On-Orbit maintainable (per service)	
Launch Cost ($million)	4
Replacement modules ($million)	5
Operations ($million)	3
Initial Changes in Satellite ($million)	3
	—
Total ($million)	15

[a] Excluding satellite development costs (DDT&E).

[b] Fraction of shuttle costs with load sharing.

Conclusions

A review of past failures proves that module exchange can satisfy most repair requirements of communications satellites. The relatively few failures that do not occur in the modules will require replacement of the satellite. Design failures are especially suited to servicing, since they occur early in the life of a satellite.

Communications satellites can be designed to be serviceable. Body-stabilized spacecraft have been assumed. The effects of required changes are acceptable, and the additional costs are reasonable. Estimates for the initial spacecraft are a 25-percent increase in weight and an 8-percent increase in production cost.

Designs of effective on-orbit servicers are available. One design has been presented in this paper, and more details on this and other designs are available in various references. One servicer can be developed to service all communications satellites, and yet considerable design flexibility will still be available to each satellite designer.

On-orbit servicing can increase both the lifetime and the availability of communications satellites. By repairing failed satellites, servicing can increase the satellite lifetime to 10 or 15 years or to the appropriate program life. Servicing an operating satellite can increase the reliability over the next year or two (and hence the satellite availability) by repairing design and redundant failures, or by replacing worn-out items near the end of their expected lifetimes.

Significant savings are possible when the program lifetime is long relative to the satellite lifetime. The average recurring cost of injecting a new communications satellite is $27 million; the average cost of servicing such a satellite is only $15 million. Including the non-recurring costs, a 15-percent savings in total program costs is estimated if one service operation is used to double the satellite lifetime.

Acknowledgment

The authors are grateful to James R. Owens and Chester Pentlicki of COMSAT, who designed the basic serviceable communications satellite; Richard D. Sosnay of Martin Marietta, who calculated the cost analysis; and Gil M. Kyrias of Martin Marietta, who designed the latch-attach mechanism.

References

[1] W. L. DeRocher, Jr., "Integrated Orbital Servicing Study for Low-Cost Payload Programs," Final Report, Vols. I and II, Contract NAS 8-30820, Martin Marietta Corporation, September 1975.

[2] G. D. Gordon, "Integrated Orbital Servicing and Payloads Study," Final Reports, Vols. I and II, Contract NAS 8-30849, COMSAT, September 1975.

[3] G. Smith and W. DeRocher, Jr., "Orbital Servicing and Remotely Manned Systems," Second Conference on Remotely Manned Systems, Los Angeles, California, June 1975.

[4] G. D. Gordon, "A User Assessment of Servicing at Geostationary Orbit," Second Conference on Remotely Manned Systems, Los Angeles, California, June 1975.

[5] A. Fiul, "On-Orbit Serviceable Satellite Design," Second Conference on Remotely Manned Systems, Los Angeles, California, June 1975.

[6] W. L. Morgan, "Satellite Utilization of the Geosynchronous Orbit," *COMSAT Technical Review*, Vol. 6, No. 1, Spring 1976, pp. 195–206.

[7] G. D. Gordon, "Availability of a Communications Satellite, Requirement and Feasibility," Conference on Space Shuttle Missions of the 80's, Denver, Colorado, August 1975, Paper No. AAS 75-254.

[8] B. I. Edelson, R. Strauss, and P. L. Bargellini, "INTELSAT System Reliability," International Astronautic Federation XXVth Congress, Amsterdam, Holland, October 1974, Paper No. 74-054.

[9] W. L. DeRocher, Jr., and R. S. Sosnay, "The Economics of Satellite Maintenance," Conference on Space Shuttle Missions of the 80's, Denver, Colorado, August 1975, Paper No. AAS 75-150.

[10] J. Mosich and A. Fiul, "Economics of Space Servicing," Conference on Space Shuttle Missions of the 80's, Denver, Colorado, August 1975, Paper No. AAS 75-148.

[11] F. J. Cepollina and J. Mansfield, "In-Orbit Servicing," *Astronautics and Aeronautics*, Vol. 13, No. 2, February 1975, pp. 48–56.

Gary D. Gordon received a B.A. from Wesleyan University in 1950 and a Ph.D. in physics from Harvard University in 1954. In 1958 he began working at the RCA Space Center, where he contributed to the thermal design of the TIROS weather satellite and was responsible for the thermal design of the RELAY communications satellite. He has taught courses on spacecraft thermal design, satellite orbits, satellite reliability, and FORTRAN programming. Dr. Gordon joined COMSAT *Laboratories in 1969 and has worked on spacecraft spin stabilization, mechanical bearings, and system reliability analysis. He is presently Senior Staff Scientist in the Spacecraft Laboratory. He is listed in* American Men of Science *and in* Who's Who in the East. *He is a member of AIAA, AAPT, Sigma Xi, and Phi Beta Kappa.*

Wilfred L. DeRocher is a departmental staff engineer at the Denver Division of Martin Marietta Corporation, where he is program manager of the Integrated Orbital Servicing Study for the Marshall Space Flight Center. For Martin Marietta he has been engaged in research and development of manned and unmanned orbital operations, space operations simulations, electromechanical systems, and control systems. He received a B.S. in mechanical engineering from the Worcester Polytechnic Institute (1947), an M.S. in mechanical engineering from the University of Michigan (1948), and an M.S. in applied mathematics from the University of Colorado (1968). Before joining Martin Marietta in 1961, Mr. DeRocher was employed by Republic Aviation (1957–1961), Bell Aircraft Company (1951–1957), and the NACA (1948–1950). He is an Associate Fellow of the AIAA and a member of IEEE and AAS.

Index: tantalum oxides, niobium oxides, antireflection films, silicon solar cells.

Tantalum oxide and niobium oxide antireflection films in silicon solar cells

A. G. REVESZ, J. F. ALLISON, AND J. H. REYNOLDS

(Manuscript received October 31, 1975)

Abstract

Thermal oxidation of Ta and Nb films deposited in vacuum on Si substrate results in noncrystalline Ta_2O_5 and Nb_2O_5 films. The refractive index of these films is very close to the optimum value of 2.3 required for an antireflection (AR) film in silicon solar cells. Losses caused by light scattering and/or optical absorption are minimized due to the lack of grain boundaries and to the high degree of short-range order in the noncrystalline structure. The silicon/oxide interface is of high perfection; hence, carriers generated in its vicinity reach the n-p junction without recombination and the quantum yield at short wavelengths is high. Also, the good quality of the shallow n-p junction is maintained so that the fill factor is close to its theoretical value. Proper integration of the AR film preparation into the overall solar cell technology has been an essential step in achieving high conversion efficiency for the Violet cell (corresponding to a power output ≥ 18 mW/cm^2) and even higher efficiency (≥ 21 mW/cm^2) for the COMSAT Non-reflecting cell. The Ta_2O_5 films have exhibited excellent stability and adherence under various severe optical, mechanical, thermal, and humidity tests.

Introduction

Silicon solar cells with significantly increased conversion efficiency, the so-called Violet [1] and the COMSAT Non-reflecting (CNR) [2] cells,*

* Both cells have been licensed for manufacture by the Optical Coating Laboratory, Inc., Photoelectronic Group, 15251 E. Don Julian Way, Industry, California 91746.

have been developed by using, among other improvements, a new technique for fabricating the antireflection (AR) film. The fundamental considerations underlying the application of vitreous oxides, especially Ta_2O_5 and Nb_2O_5, as AR films have already been described in a CTR Note [3] which emphasizes the following properties of the film: noncrystallinity to avoid grain boundary effects and short-range order in the structure to minimize light absorption and to maximize stability and reproducibility.

It has also been pointed out that, to achieve the desired properties, the oxide films should be *grown* rather than deposited. Since both Ta_2O_5 and Nb_2O_5 films can be prepared by anodic oxidation in a well-characterized noncrystalline form [4] and thermal oxidation of bulk tantalum and niobium usually results in polycrystalline oxides with poorly defined properties [5], [6], anodic oxidation was originally considered preferable. However, it was discovered that, under proper conditions, thermal oxidation of Ta and Nb films vacuum-deposited on a silicon substrate could also result in noncrystalline oxides with well-defined properties and, in addition, in a silicon/oxide interface of high perfection.

After the implementation of these concepts, the authors became aware of the recent application of noncrystalline Ta_2O_5 films in the formation of optical waveguides [7]-[9]. These films were prepared by thermal oxidation of sputtered Ta films on various insulator substrates (e.g., glass and silica). In some respects, their properties are similar to those of Ta_2O_5 films on silicon, but there are also significant differences.

The possibility of using Ta_2O_5 as an AR film in silicon solar cells has been mentioned during the past few years, but no explicit results have been published; furthermore, the important structural aspects and the effects of preparation conditions upon these films have not been considered [10], [11]. Subsequent to the introduction of Ta_2O_5 and Nb_2O_5 as AR films in the Violet cell, reflectivity calculations have demonstrated that, theoretically, a material with the refractive index of Ta_2O_5 is the best single-layer AR film providing maximum short-circuit current [12].*

This paper will provide a detailed descr:ption of the preparation and properties of noncrystalline Ta_2O_5 and, to a lesser degree, Nb_2O_5 films on silicon in terms of their application as AR coating for silicon solar cells.

* It should be pointed out that the calculations in Reference 12 are based on the assumption that the dielectric films are nonabsorbing. This assumption is generally not justified, especially for SiO_x. Therefore, it is very unlikely that the combination of SiO_x and TiO_2, which gave a better calculated value for the short-circuit current than the single-layer Ta_2O_5, can be experimentally achieved.

Experimental

Tantalum and niobium films were deposited on silicon substrates by electron beam evaporation from a high-purity source in a minimum vacuum of 10^{-6} torr. In a few cases Ta films were also deposited by RF sputtering. It was found experimentally that, to produce a quarter-wavelength AR film "bloomed" at a wavelength of 540 nm, it was necessary to deposit ~28 nm of tantalum or ~48 nm of niobium. After oxidation this yields a film ~60 nm thick in the case of Ta_2O_5 and ~56 nm thick in the case of Nb_2O_5. These thickness values together with their respective refractive indices (see below) correspond to an effective optical path of 134 nm.

The deposited metal films were oxidized in oxygen using a resistance-heated silica tube. The temperature range was 350°C to 700°C, but practically all the AR films were obtained at 525°C for 5 minutes in the case of tantalum, and at 450°C for 5 minutes in the case of niobium. Under these conditions, the metal films were fully oxidized as revealed by their optical properties. For comparison, tantalum oxide films were also obtained by vacuum deposition using electron beam heating of Ta_2O_5 pressed powder.

The refractive index and thickness of the oxide films were determined with ellipsometry at 546.1-nm wavelength. The light absorption as a function of wavelength was determined by measuring the reflected and transmitted light using an integrating sphere attached to a spectrometer. For this purpose, the oxide films were produced on fused silica substrates.

Several methods are available* for producing an oxide film which does not interfere with the current collector metal grid contact:

a. The Ta or Nb metal is selectively deposited or etched after deposition so that it does not cover those areas where the contact metal will be deposited.

b. The uniformly deposited Ta or Nb film is oxidized; then a pattern is etched in the oxide film.

c. The uniformly deposited Ta or Nb film is selectively covered with the contact metal and then oxidized; the contact metal serves as a mask during oxidation so that oxide film is produced only in the exposed areas.

According to the authors' experience, all of these methods are feasible. The important point is that preparation of this type of AR film must be

* See, for example, Reference 13. Other patents are pending in the U.S. Patent Office and elsewhere.

considered as an integral part of the overall solar cell fabrication process rather than an independent process as in the case of conventionally deposited SiO_x and other films.

Results

Both transmission and reflection electron diffraction analyses of tantalum oxide films on silicon showed that the films are noncrystalline [14]. No discernible morphological features could be observed with either optical or electron microscopy, indicating that the oxide films are very uniform. They also exhibited bright interference colors. However, contamination during metal deposition and/or subsequent oxidation might initiate localized crystallization in the oxide film, resulting, for instance, in a loss of their brightness. It has generally been observed that the properties of solar cells are impaired when morphologically imperfect oxides are used as AR films.

It should be noted that oxidation of a tantalum foil under identical conditions resulted in a polycrystalline oxide film that was so nonuniform that ellipsometric measurements could not be performed. This observation was in accordance with original expectations regarding the thermal oxidation of tantalum as mentioned above. It underlines the importance of the properties of the metal in determining whether the oxide film would be noncrystalline or crystalline.

It has been demonstrated by Rutherford backscattering analysis that tantalum oxide films on silicon substrate are, to a first approximation, stoichiometric Ta_2O_5 [15]. In addition, the infrared absorption spectrum has been found to be very similar to that of crystalline Ta_2O_5 (broad absorption at ~16-μm wavelength [16]), indicating a similarity in the short-range order.

Closely related to stoichiometry is the absorption of light in the visible part of the spectrum. This is particularly important in terms of applying these oxides as AR film in solar cells. The large difference in absorption between SiO_x and Ta or Nb oxides is shown in Figure 1. This difference is due to the well-known fact that SiO_x absorbs significantly at shorter wavelengths. This absorption is mainly due to the presence of Si-Si bonds in amorphous SiO_x that are also responsible for the increased refractive index (relative to SiO_2); as a result, the light absorption of SiO_x increases as its refractive index increases [17].

It is very instructive to observe in Figure 1 the significant difference in absorption between thermally grown (curve 1) and vacuum-deposited (curve 3) tantalum oxides. From the absorption spectrum of thermally grown Ta_2O_5 film, its band gap can be roughly estimated as 4.1 eV. This value is comparable to 4.2 eV reported for chemically deposited polycrystalline β-Ta_2O_5 [18] and noncrystalline Ta_2O_5 films [19]. On the other hand, the band gap of vacuum-deposited tantalum oxide is estimated to be less than 3 eV. This indicates that, as expected, the grown oxide has a higher degree of short-range order and hence less unsaturated valences and/or Ta-Ta bonds than the deposited one.

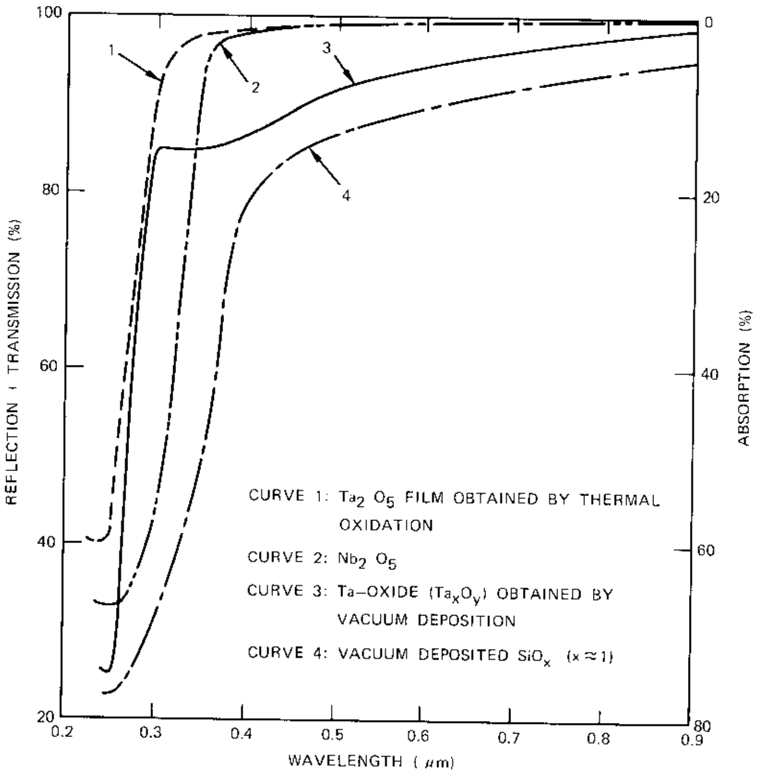


Figure 1. *Absorption of Various Dielectric Films (The total of the reflected and transmitted light was measured; the absorption A, was determined from the relationship $R + T + A = 100\%$. The substrate in all cases was fused silica.)*

The refractive indices of the oxide films, as determined from ellipsometric measurements, are 2.23 for ~60-nm-thick Ta_2O_5 and 2.40–2.55 for ~55-nm-thick Nb_2O_5 films. This value for Ta_2O_5 can be compared with the refractive index of anodic Ta_2O_5 film (n = 2.22) [20] and with the range of values ($2.14 \leq n \leq 2.25$) reported for thermal oxidation of sputtered Ta films on glass or other insulator substrates [7]–[9]. The refractive index of thermally grown Nb_2O_5 is somewhat higher than that of anodic oxide [20] (n = 2.37). The refractive index of vacuum-deposited tantalum oxide is not well characterized, but generally it is lower than that of thermally grown Ta_2O_5 (e.g., it has been found that n = 2.10) and therefore represents a poorer optical match to silicon.

It has been demonstrated by stepwise etching that the refractive index of Ta_2O_5 films on silicon is characterized by a gradient across the film [14]. Its value increases from the silicon/oxide interface to the outer surface; therefore, the value given above is, in fact, the refractive index of the film averaged through its total thickness. Furthermore, the refractive index of ~60-nm-thick Ta_2O_5 films on silicon decreases from 2.26 to 2.20 as the oxidation temperature increases from 350°C to 700°C; it also decreases during post-oxidation heat treatments in oxidizing ambient [14]. Full oxidation of Ta films of varying thickness at 525°C results in oxide films whose average refractive index varies from 1.94 to 2.34 as the thickness increases from 12 to 112 nm [21]. It is significant to note that, despite the large variation in the refractive index, the ratio of tantalum and oxygen atoms in these films is roughly maintained at 1:2.5.

These effects have been attributed to an interaction in the Ta (Ta_2O_5)-Si structure which results in the incorporation of silicon in oxidized form into the noncrystalline tantalum oxide. Niobium oxide films have not been investigated in such detail as Ta_2O_5 films, but it can be safely assumed that their behavior is similar. The interaction between the oxide film and the silicon substrate plays an important role in determining the characteristics of the solar cell; this will be discussed further in the following paragraphs.

In view of these observations, the refractive index of Ta_2O_5 films prepared under given conditions for use as an AR film cannot be considered to be a general property of these films; instead, it must be considered to be determined by the specific preparation conditions. However, the particular refractive index values of the AR films (2.23 and 2.37 for Ta_2O_5 and Nb_2O_5, respectively) prepared under the conditions described above are very close to the optimum value at which light reflection (at the matching wavelength) from the silicon surface in an assembled solar cell is practically nil. It can be easily shown that this optimum value is about 2.3 under the following conditions: the silicon is covered with a fused silica cover slide (n = 1.45) and the film thickness is such that the effective optical path is a quarter wavelength in the film for λ = 540 nm.

The nearly ideal value of n renders thermally grown Ta_2O_5 a more efficient AR film than the conventional SiO_x. This can also be seen from Figure 2, which shows the relative amount of reflected light for a bare silicon surface as well as for completed solar cell surfaces coated with SiO_x, Ta_2O_5, and Nb_2O_5. The reflectance combined with the absorption shown in Figure 1 results in an increased transmission of light into silicon for thermally grown Ta_2O_5 and Nb_2O_5 AR films relative to SiO_x. It should be noted that application of SiO_x in the Violet cell would wipe out most of the gain in efficiency resulting from the use of a shallow diffusion; this is mostly due to the relatively high absorption of SiO_x at short wavelengths.

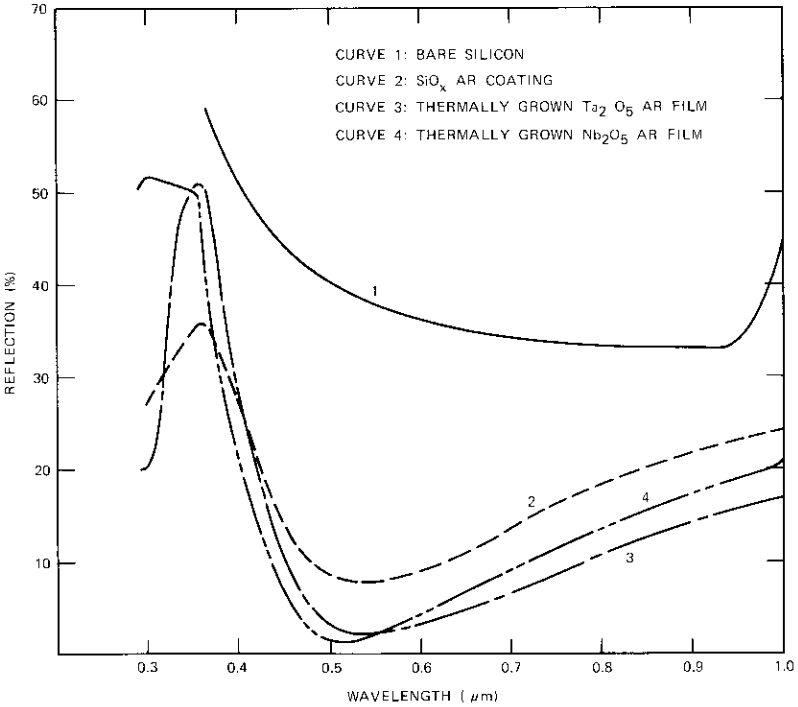


Figure 2. *Reflectance Spectra of Various Silicon Surfaces*

In addition to the lack of absorption and near-perfect value of the refractive index, there are some other properties of an AR film which contribute to the overall conversion efficiency of the solar cell. These properties play a more important role in the Violet cell than in the conventional cell because, in the former, the junction is very near the surface (100-150 nm). One of these properties is the effect of AR film on the quality of the n-p junction as reflected in the sharpness of the I-V characteristics of the solar cell; this is usually indicated by the fill factor. The experimental values of this fill factor are within the range of 78-80 percent for properly prepared Ta_2O_5 and Nb_2O_5 films as opposed to the theoretical value of 83 percent at room temperature for the substrate doping level employed. It has been observed, however, that, when contamination occurs during the metal deposition and/or oxidation, the fill factor decreases. A very striking decrease to 50-60 percent was observed when the Ta film was deposited by sputtering rather than by vacuum deposition. Apparently the sputtering process damaged the n-p junction. This observation also indicates the importance of preparation conditions for the AR film and demonstrates that they form an integral part of the overall fabrication process.

Another important property of the AR film is the quality of its interface with the Si substrate. This is illustrated by the following considerations. The basis of the increased conversion efficiency of the Violet cell is the extension of the spectral response to shorter wavelengths [1]. However, this is the very region where the optical absorption in silicon is quite strong; for example, the reciprocal absorption coefficient at 0.36-μm wavelength is 10 nm. Carriers at this and shorter wavelengths are generated virtually at the Si/Ta_2O_5 interface. If the surface recombination velocity were not sufficiently low, these carriers could not be collected at the junction. Consequently, the quantum yield of the Violet cell would be significantly less than the observed range of 0.6 to 0.9 in the short-wavelength (0.31- to 0.45-μm) regime of the spectrum.

The low surface recombination velocity is an indication of the low density of interface states at the Si/Ta_2O_5 interface. This has also been demonstrated by MOS capacitance measurements which have resulted in interface state densities of the order of 10^{12} cm^{-2} (eV)$^{-1}$. It is very probable that the reason for the relatively low interface state density is the incorporation of silicon into the oxide film during its growth [14]. As a result, the silicon/oxide interface does not coincide with the original Si/Ta(Nb) interface. It is also very probable that the structural flexibility of noncrystalline Ta_2O_5 (Nb_2O_5) plays a role in accommodating the oxide film to the silicon surface without generating too many interface states (defects) there. In both respects, the situation is somewhat similar to the thermal oxidation of silicon, which is well known to result in a very perfect silicon/oxide interface (see, for example, Reference 22).

The Ta_2O_5 (Nb_2O_5) AR films not only have excellent optical and interface properties, but they also adhere well to the surface and are very stable. This has been demonstrated by a 1,000-hour humidity, tape peel, thermal shock, and thermal soak test during which the average maximum power degradation of bare and covered (with fused silica cover slide) Violet cells was 4 and 0.3 percent, respectively [23]. Even this small degradation has been attributed to contact problems rather than to the AR film.

Another aspect of the stability of the AR film is its behavior during UV irradiation. Investigations carried out for 1,500 hours at one solar UV constant have shown that both bare and covered (with a UV filter of 0.35-μm cutoff wavelength) Violet cells using Ta_2O_5 AR film degraded by about 1.5 percent, whereas conventional cells using SiO_x AR coating degrade by about 3 percent [24].*

Application of Ta_2O_5 (Nb_2O_5) AR film reduces the reflection not only from a flat silicon surface (see Figure 2), but also from an etched surface which exhibits much lower reflection than the bare flat surface. This is illustrated by the following example: At 0.5-μm wavelength, the reflectance of an etched silicon surface (consisting of pyramids with a base length of several micrometers) is about 14 percent, whereas it is zero when Ta_2O_5 (Nb_2O_5) AR film is applied; corresponding values of integrated reflectance, weighted according to the solar spectrum, are 13 percent and 2.5 percent, respectively [25].

The good optical and interface properties of Ta_2O_5 and Nb_2O_5 films, their stability, and the possibility of fabricating these films without damaging the n-p junction made it possible to utilize the increased short wave spectral response of the shallow n-p junction in the Violet and CNR cells. The net result was a significant improvement in the power output per unit area (conversion efficiency) relative to that of the conventional solar cell; typical values exceeded 18 mW/cm^2 for Violet cells and 21 mW/cm^2 for CNR cells, respectively. A detailed comparison of the properties of these advanced types of silicon solar cells is given in Reference 24.

Conclusions

This paper has shown that thermal oxidation of vacuum-deposited Ta and Nb films on Si substrate results in noncrystalline Ta_2O_5 and Nb_2O_5 films with well-defined properties. The refractive index of these films is

*It should be noted that, since SiO_x is not a well-defined compound, this degradation behavior is characteristic of only the particular film used in the cells investigated rather than a general property of SiO_x.

very close to the optimum value of 2.3 required for the antireflection film in silicon solar cells. Contrary to that of the conventional SiO_x AR films, this high refractive index is not associated with relatively large optical absorption; therefore, the increased short-wavelength response of shallowly diffused silicon n-p solar cells can be effectively utilized. Interaction with the Si substrate during oxidation of the Ta (Nb) film and the flexibility of the noncrystalline structure of these thermally grown oxide films lead to the formation of a good silicon/oxide interface. As a result, carriers generated in the vicinity of the interface reach the junction without recombining, thus providing for a high quantum yield at short wavelengths.

Under proper process conditions, the preparation of these oxide films does not impair the characteristics of the nearby junction, and thus the fill factor is close to its theoretical value. The fabrication process is reproducible and the properties of the Ta_2O_5 (Nb_2O_5) films are very stable. Careful incorporation of metal deposition and oxidation techniques into the overall technology has been an essential factor in achieving the significantly increased conversion efficiency of the Violet and Non-reflecting solar cells.

Acknowledgment

Thanks are due to R. Arndt, D. Curtin, and A. Meulenberg for helpful discussions as well as to F. Bland, A. Busch, R. Dendall, and D. Perry for their assistance in the experimental work. The authors are indebted to E. S. Rittner for constructive comments on the manuscript.

References

[1] J. Lindmayer and J. Allison, "The Violet Cell: An Improved Silicon Solar Cell," *COMSAT Technical Review*, Vol. 3, No. 1, Spring 1973, pp. 1–21; "An Improved Silicon Solar Cell—The Violet Cell," *Conference Record of the Ninth IEEE Photovoltaic Specialists Conference*, Silver Spring, Maryland, 1972, p. 83.

[2] J. Haynos et al., "The COMSAT Non-reflective Silicon Solar Cell: A Second Generation Improved Cell," *Proc. of the International Conference of Photovoltaic Power Generation*, Hamburg, Germany, 1974, pp. 487–500.

[3] A. G. Revesz, "Vitreous Oxide Antireflection Films in High-Efficiency Solar Cells," *COMSAT Technical Review*, Vol. 3, No. 2, Fall 1973, pp. 449–452; *Conference Record of the Tenth IEEE Photovoltaic Specialists Conference*, Palo Alto, California, 1973, pp. 180–181.

[4] L. Young, *Anodic Oxide Films*, New York: Academic Press, 1961.

[5] P. Kofstad, *High Temperature Oxidation of Metals*, New York: John Wiley & Sons, 1966.

[6] K. R. Lawless, "The Oxidation of Metals," *Reports on Progress in Physics*, Vol. 37, 1974, p. 231.

[7] D. H. Hensler et al., "Optical Propagation in Sheet and Pattern Generated Films of Ta_2O_5," *Applied Optics*, Vol. 10, No. 5, May 1971, pp. 1037–1042.

[8] M. Fujimori, A. Okamoto, and Y. Nishimura, "Tantalum Pentoxide Thin Films for Light Guide," *Fujitsu Scientific and Technical Journal*, Vol. 8, No. 3, September 1972, pp. 177–189.

[9] Y. K. Lee and S. Wang, "Tantalum Oxide Light Guide on Lithium Tantalate," *Applied Physics Letters*, Vol. 25, No. 3, August 1, 1974, pp. 164–166.

[10] K. S. Tarneja and W. R. Harding, U.S. Patent 3,533,850, 1970.

[11] R. L. Crabb and D. Basnett, "Environmental Assessment of Thin Silicon Solar Cells from Pilot Production," *IEEE Transactions on Electron Devices*, ED-18, No. 8, August 1971, pp. 491–506.

[12] E. Y. Wang et al., "Optimum Design of Antireflection Coating for Silicon Solar Cells," *Conference Record of the Tenth IEEE Photovoltaic Specialists Conference*, Palo Alto, California, 1973, pp. 168–173.

[13] A. G. Revesz and R. J. Dendall, U.S. Patent 3,904,453, 1975.

[14] A. G. Revesz et al., "Oxidation of Tantalum Film on Silicon," *Thin Solid Films*, Vol. 23, No. 3, October 1974, pp. S63–S66.

[15] J. Hirvonen, A. G. Revesz, and T. Kirkendall, "Rutherford Backscattering Investigation of Thermally Oxidized Tantalum on Silicon," *Thin Solid Films*, to be published.

[16] N. T. McDevitt and W. L. Baun, "Infrared Absorption Study of Metal Oxides in the Low Frequency Region (700-240 cm^{-1})," *Spectrochim. Acta*, Vol. 20, 1964, p. 799.

[17] H. R. Philipp, "Optical Properties of Non-crystalline Si, SiO, SiO_x, and SiO_2," *Journal of Physics and Chemistry of Solids*, Vol. 32, No. 8, August 1971, pp. 1935–1946.

[18] W. H. Knausenberger and R. N. Tauber, "Selected Properties of Pyrolytic Ta_2O_5 Films," *Journal of the Electrochemical Society*, Vol. 120, No. 7, July 1973, pp. 927–931.

[19] C. C. Wang, K. H. Zaininger, and M. T. Duffy, "Metal Oxide Thin Films," *RCA Review*, Vol. 31, No. 4, December 1970, pp. 728–741.

[20] L. Young and F. G. R. Zobel, "Ellipsometric Study of Steady-State High Field Ionic Conduction in Anodic Oxide Films on Tantalum, Niobium, and Silicon," *Journal of the Electrochemical Society*, Vol. 113, No. 3, March 1966, pp. 277–284.

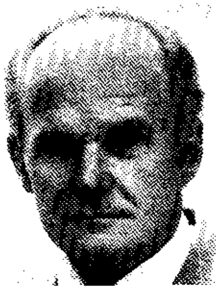
[21] A. G. Revesz, J. Reynolds, and J. Allison, "Optical Properties of Tantalum Oxide Films on Silicon," *Journal of the Electrochemical Society*, to be published.

[22] A. G. Revesz, "Noncrystalline Silicon Dioxide Films on Silicon: A Review," *Journal of Non-crystalline Solids*, Vol. 11, 1973, pp. 309–330.

[23] D. J. Curtin and R. W. Cool, "Qualification Testing of Laboratory Produced Violet Solar Cells," *Conference Record of the Tenth IEEE Photovoltaic Specialists Conference*, Palo Alto, California, 1973, pp. 139–152.

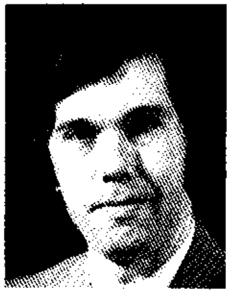
[24] J. F. Allison, R. A. Arndt, and A. Meulenberg, "A Comparison of the COMSAT Violet and Non-reflective Solar Cells," *COMSAT Technical Review*, Vol. 5, No. 2, Fall 1975, pp. 211–223.

[25] R. A. Arndt et al., "Optical Properties of the COMSAT Nonreflective Cell," *Conference Record of the 11th IEEE Photovoltaic Specialists Conference*, Scottsdale, Arizona, 1975, p. 40.



A. G. Revesz received Diploma of Engineering and Ph.D. degrees from the Technical University of Budapest, Hungary. As Senior Staff Scientist in the Applied Sciences Division at COMSAT *Laboratories, he has been the principal investigator in identifying the failure mechanism of tunnel diodes and devising a better selection procedure for these devices. His work on semiconductor-insulator interfaces and noncrystalline oxides has resulted in a new antireflection film essential for the development of silicon solar cells of increased efficiency. Prior to joining* COMSAT, *he was with RCA David Sarnoff Research Center, where he received an Outstanding Achievement Award for his work on the Si-SiO$_2$ interface. He is a member of the editorial board of the* COMSAT TECHNICAL REVIEW *and a Fellow of the American Institute of Chemists.*

James F. Allison received a B.S.E.E. degree from Carnegie Mellon University in 1959 and an M.S.E.E. from Princeton University in 1962. He worked at RCA Laboratories for a period of ten years, conducting research in the areas of thin films and solid-state devices, for which he received an RCA Laboratories' Achievement Award. In 1969 he joined COMSAT, *where he has been active in conducting solid-state device technology research. He is presently Manager of the Semiconductor Technology Department at* COMSAT *Laboratories.*



John Reynolds received a B.S. degree in physics from the Worchester Polytechnic Institute in 1962 and an M.S. degree in physics from Purdue University in 1964. Prior to joining COMSAT, *he was a staff member at the Sprague Electric Research Center in North Adams, Massachusetts, where he was engaged in research on MOS transistors and other solid-state devices.*

Mr. Reynolds joined COMSAT *in 1968 and has been engaged in the study and analysis of solar cells and more recently solid-state microwave devices. He is a member of IEEE and Sigma Xi.*

Index: Cassegrain antennas, measurement, sidelobe envelopes

Wide-angle sidelobe envelopes of a Cassegrain antenna

R. W. Kreutel

(Manuscript received November 5, 1975)

Abstract

The wide-angle sidelobe envelope of a high-efficiency Cassegrain antenna is computed on the basis of a relatively simple analysis. The model selected for analysis typifies the antenna designs presently employed as satellite communications earth terminals. Individual contributions to the sidelobe envelope, such as subreflector support blocking, are identified and characterized. The composite wide-angle envelope of the peak sidelobe levels is then calculated by summing the individual contributors. The results are compared with experimental data and the agreement is shown to be good.

Introduction

An important consideration in the design of microwave communications systems which reuse a common frequency band is the interference between systems [1]–[3]. For example, the interference between a satellite communications network and terrestrial radio relays sharing the 4/6-GHz bands is fundamentally limited by the radiation characteristics of the antenna employed at the network terminals.

In this paper a Cassegrain antenna, typical of the antennas presently deployed for satellite communications, is modeled and its wide-angle radiation envelope is derived theoretically. The elements of the antenna which influence the sidelobe envelope are treated individually and their relative contributions to the wide-angle antenna characteristics are given. The results of the analysis are compared with experimental data and the agreement is shown to be good.

Analytical model

Figure 1 is a diagram of the model used for the analysis. The model, which is intended to represent a high-efficiency shaped Cassegrain antenna, is a reflector configuration in which the subreflector and main reflector are "shaped" to provide a uniform aperture field distribution [4]. This configuration is typical of that used for satellite communications earth terminal antennas. The notation indicated in Figure 1 will be used in the analysis which follows.

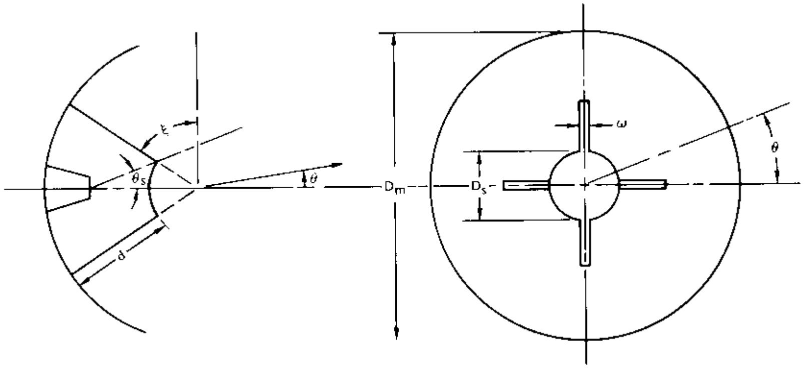


Figure 1. *Model of the Cassegrain Antenna* ($D_s/D = 0.1$, $\omega/D = 0.01$, $d/D_m = 0.39$, $\xi = 45°$)

The antenna gain function

The voltage gain function of the antenna modeled in Figure 1 can be written as

$$G_v(\theta, \phi) \cong \eta \left[g_m(\theta, \phi) - \sum_{k=1}^{n} \overline{g_k(\theta, \phi)} \right] + g_f(\theta, \phi)\, u_{-1}(\theta - \theta_s) \qquad (1)$$

where $g_m(\theta, \phi)$ = voltage gain function of the main reflector
$g_k(\theta, \phi)$ = voltage scattering function of the kth blocking obstacle
$g_f(\theta, \phi)$ = voltage gain function of the feed
u_{-1} = unit step function
η = square root of the feed spillover efficiency factor

and θ_s is defined in Figure 1.

Equation (1) represents the complete antenna pattern in the forward hemisphere; its evaluation would be involved and lengthy. Since it is of interest to provide bounds on the sidelobe levels, only the envelope of equation (1) must be calculated. Thus, equation (1) can be rewritten as

$$\overline{G_v(\theta, \phi)} \cong \eta \left[\overline{g_m(\theta, \phi)} + \sum_{k} \overline{g_k(\theta, \phi)} \right] + \overline{g_f(\theta, \phi)}\, u_{-1}(\theta - \theta_s) \qquad (2)$$

where the bar is used to indicate the envelope of a function. The individual quantities of equation (2) will be defined in the following subsection.

Main reflector and subreflector

The voltage gain function of the main reflector is simply the Fourier transform of the uniform aperture field distribution [5]. Thus,

$$g_m(\theta, \phi) = \frac{\pi D_m}{\lambda} \frac{J_1\left(\frac{\pi D_m}{\lambda} \sin \theta\right)}{\frac{\pi D_m}{\lambda} \sin \theta} (1 + \cos \theta) \qquad (3)$$

where $J_1(x)$ is the first-order Bessel function of the first kind.

The subreflector generates a forward-scattered field which precisely cancels the aperture field which would have been present in the absence of subreflector blocking. Its voltage gain function is given by

$$g_1(\theta, \phi) = -\frac{\pi D_m}{\lambda} \left(\frac{D_s}{D_m}\right)^2 \frac{J_1\left(\frac{D_s}{D_m} \frac{\pi D_m}{\lambda} \sin \theta\right)}{\left(\frac{D_s}{D_m}\right)\left(\frac{\pi D_m}{\lambda} \sin \theta\right)} (1 + \cos \theta) \quad . \qquad (4)$$

The first two terms of equation (2) are the envelopes of equations (3) and

(4). For $\theta > \sin^{-1}(4\lambda)/(\pi D_s)$ the asymptotic form of the Bessel function can be used to obtain

$$\overline{g_m(\theta, \phi)} + \overline{g_1(\theta, \phi)} = \frac{\sqrt{2}}{\pi}\left[1 + \sqrt{\frac{D_s}{D_m}}\right]\left[\frac{D_m}{\lambda}\sin\theta\right]^{-1/2}\cot\frac{\theta}{2} \qquad (5)$$

for the envelope sum of equations (3) and (4). Equation (5) is plotted in Figure 2 in dB relative to isotropic for $D_s/D_m = 0.1$ and $D_m/\lambda = 400$.

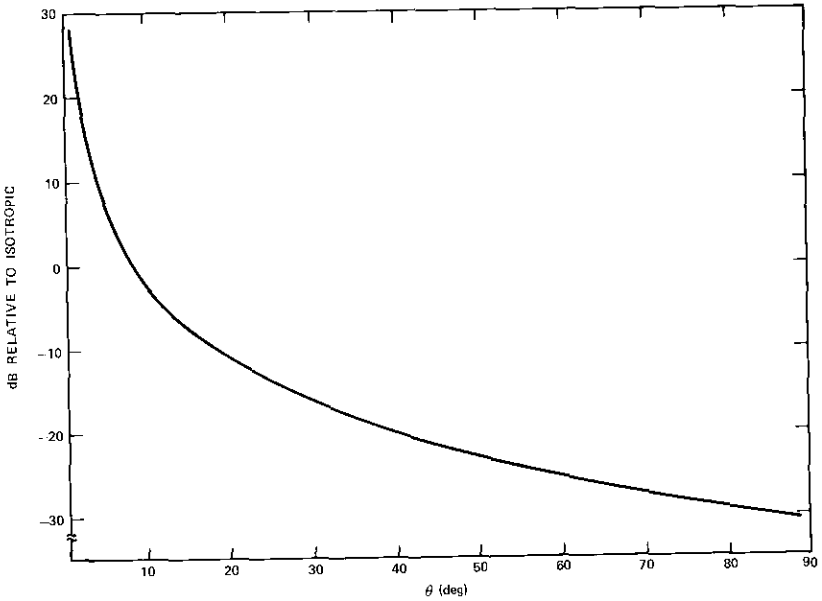


Figure 2. *Radiation Envelope of the Main Reflector and Subreflector (uniform illumination, D/λ = 400, D_s/D = 0.1)*

Plane wave quadripod blocking

The quadripod structure blocks a portion of the incident plane wave. The plane wave field induces currents on the quadripod members which generate two scattered fields, a forward-scattered field and a back-scattered field. For a quadripod member inclined to the aperture plane, the scattered field resembles that of a phased line source and maximizes on a conical surface, as shown in Figure 3.

The quadripod structure is obviously not a circularly symmetric structure. Therefore, the scattering patterns of the structure will not be circularly

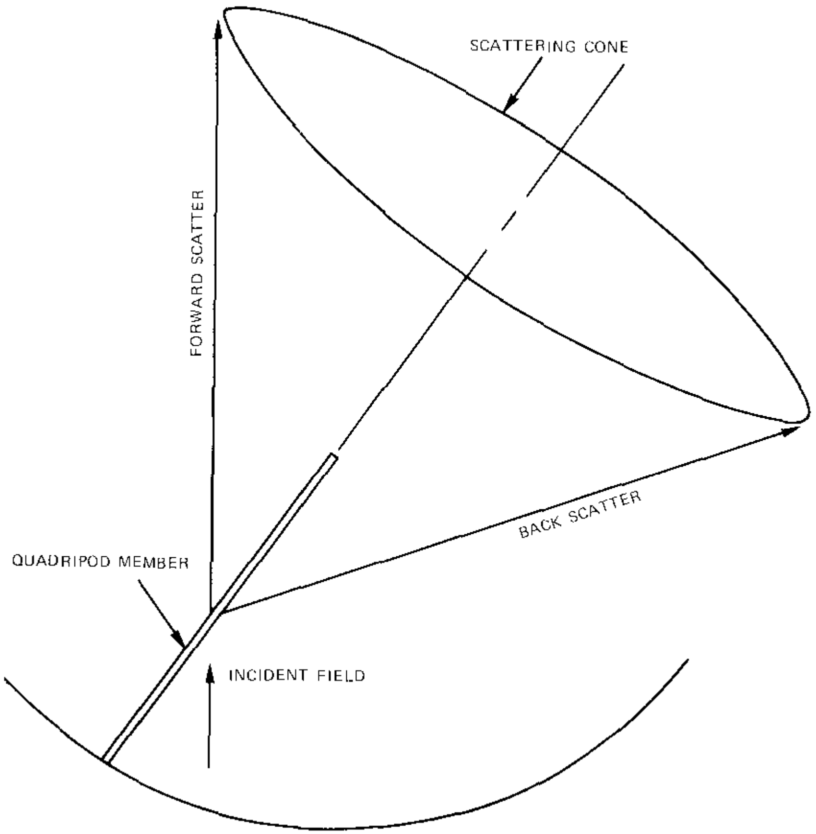


Figure 3. *Plane Wave Scattering by a Quadripod Member*

symmetric, and scattering patterns must be calculated in at least two planes to provide a reasonably complete representation. In this case the patterns have been calculated in the 0° and 45° planes (see Figure 1).

The procedure for computing the scattering pattern is based on the assumption that a quadripod member can be approximated by a plane strip of width ω. The current distribution is determined by applying the field boundary conditions to the surface of the strip. Finally, the scattering pattern is obtained by integrating the currents over the area of this strip.

In the 0° plane the contribution of the vertical pair of quadripod members can be written as

$$g_2(\theta, 0^\circ) \cong \frac{\pi D_m}{\lambda}\left[\frac{8\omega d}{\pi D_m}\right][\sin^2\theta + \cos^2\theta\cos^2\xi]^{1/2}$$

$$\cdot\left[\frac{\sin\left(\frac{\pi\omega}{\lambda}\sin\theta\right)}{\frac{\pi\omega}{\lambda}\sin\theta}\right]\left\{\frac{\sin\frac{\pi d}{\lambda}[\sin\xi(1-\cos\theta)]}{\frac{\pi d}{\lambda}[\sin\xi(1-\cos\theta)]}\right\} \qquad (6)$$

where the square root term is the radiation pattern of an infinitesimal current element. The 0° plane contribution of the horizontal quadripod members is

$$g_3(\theta, 0^\circ) \cong \frac{\pi D_m}{\lambda}\left(\frac{4\omega d}{\pi D_m^2}\right)\cos\xi\left\{\frac{\sin\frac{\pi d}{\lambda}[\sin\xi + \sin(\theta-\xi)]}{\frac{\pi d}{\lambda}[\sin\xi + \sin(\theta-\xi)]}\right.$$

$$\left.+\frac{\sin\frac{\pi d}{\lambda}[\sin\xi - \sin(\theta+\xi)]}{\frac{\pi d}{\lambda}[\sin\xi - \sin(\theta+\xi)]}\right\} \qquad (7)$$

where, since only the envelope of g_3 is of interest, the phase term relating the two $\sin x/x$ functions has been discarded.

The envelopes of equations (6) and (7) are plotted in Figure 4 for $\omega = 0.01D_m$, $d = 0.39D_m$, $\xi = 45°$, and $D_m/\lambda = 400$. It should be noted that the contribution of the horizontal members begins to increase for θ greater than about 50°. Actually, if the plot were continued, a secondary maximum would appear (for $\xi = 45°$) at $\theta = 90°$, corresponding to a peak of the back-scatter pattern of a horizontal member. In practice the back scatter will be intercepted by an adjacent quadripod member as well as by the edge of the main reflector and scattered diffusely, as shown in Figure 5. The presence of this back-scatter field is therefore neglected.

In the 45° plane the entire scattering pattern of the quadripod structure can be written as

$$g_2(\theta, 45^\circ) + g_1(\theta, 45^\circ) \cong \frac{\pi D_m}{\lambda}\left[\frac{8\omega d}{\pi D_m^2}\right]$$

$$\cdot\left\{\frac{\sqrt{2}}{2}\left[1+\frac{1}{2}\sin^2\theta+\frac{\sqrt{2}}{2}\sin 2\theta\right]^{1/2}\right.$$

$$\cdot\left[\frac{\sin\left\{\frac{\sqrt{2}}{2}\frac{\pi d}{\lambda}\left[1+\frac{\sqrt{2}}{2}\sin\theta-\cos\theta\right]\right\}}{\frac{\sqrt{2}}{2}\frac{\pi d}{\lambda}\left[1+\frac{\sqrt{2}}{2}\sin\theta-\cos\theta\right]}\right]$$

$$+\frac{\sqrt{2}}{2}\left[1+\frac{1}{2}\sin^2\theta-\frac{\sqrt{2}}{2}\sin 2\theta\right]^{1/2}$$

$$\left.\cdot\frac{\sin\left[\frac{\sqrt{2}}{2}\frac{\pi d}{\lambda}\left\{1-\frac{\sqrt{2}}{2}\sin\theta-\cos\theta\right\}\right]}{\frac{\sqrt{2}}{2}\frac{\pi d}{\lambda}\left[1-\frac{\sqrt{2}}{2}\sin\theta-\cos\theta\right]}\right\}$$

$$\cdot\left\{\frac{\sin\frac{\sqrt{2}}{2}\frac{\pi\omega}{\lambda}\sin\theta}{\frac{\sqrt{2}}{2}\frac{\pi\omega}{\lambda}\sin\theta}\right\} \qquad (8)$$

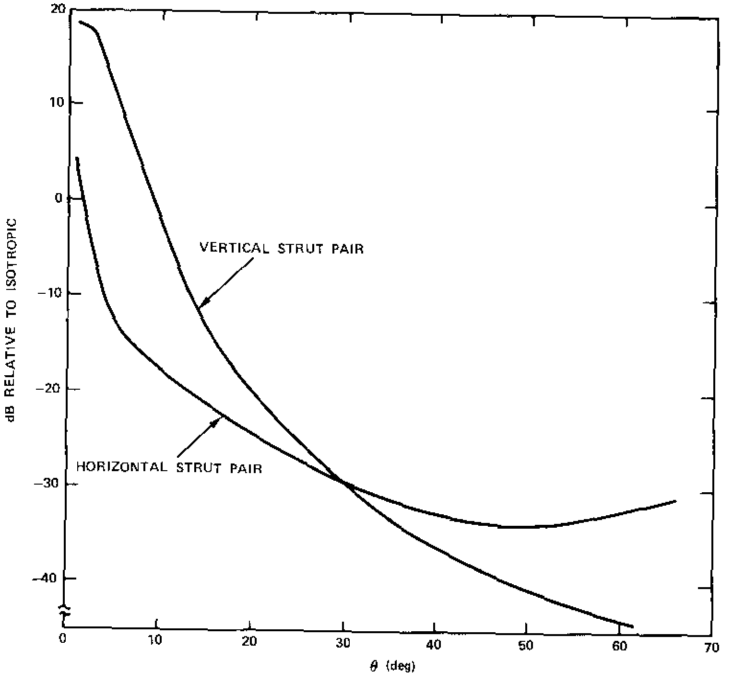


Figure 4. *Plane Wave Radiation Envelope with Quadripod Blocking* ($\phi = 0°$, $D/\lambda = 400$)

where the square root terms are the appropriate element patterns. It has been assumed that $\xi = 45°$ to simplify the expressions, and phase terms have been neglected for the same reason as before.

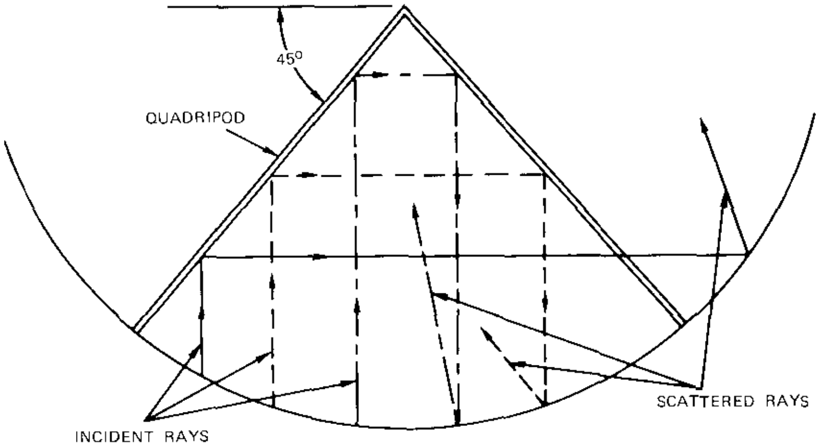


Figure 5. *Diffusion of Wide-Angle Back Scatter by the Quadripod and Reflector*

Unfortunately, simplifying equation (8) by assuming that $\xi = 45°$ results in the loss of an important characteristic of the expression. In general form the angular-dependent argument of the second sin x/x function in equation (8) is proportional to

$$\sin \xi - \frac{\sqrt{2}}{2} \cos \xi \sin \theta - \sin \xi \cos \theta$$

which is zero at $\theta = 0$ and for θ given by

$$\theta = 2 \cot^{-1}(\sqrt{2} \tan \xi) \quad . \tag{9}$$

Thus, as in the case of equation (7), there is a secondary maximum in the envelope. In this case, however, this effect cannot be neglected; the maximum is allowed to form and is present in the sidelobe envelope of the antenna.

For $\xi = 45°$, equation (9) indicates a maximum at $\theta = 70°$. This is evidenced in Figure 6, in which the envelope of equation (8) is plotted for $\omega = 0.01D_m$, $d = 0.39D_m$, and $D_m/\lambda = 400$.

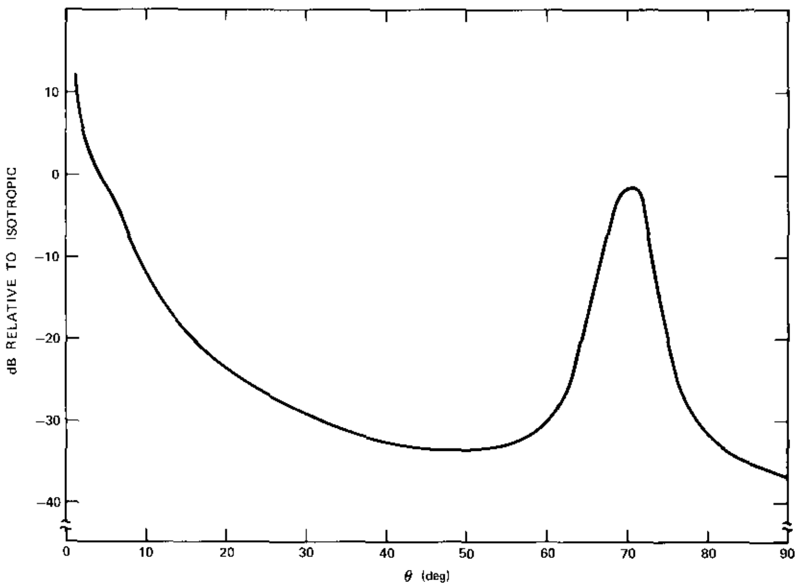


Figure 6. *Plane Wave Radiation Envelope with Quadripod Blocking* ($\phi = 45°$)

Spherical wave quadripod blocking

Because the quadripod members do not extend to the edge of the main reflector, a portion of the spherical wave emanating from the subreflector toward the main reflector is blocked. As shown in Figure 7a, this results in a shadowed area between the base of a strut and the edge of the main reflector. In the $\phi = 0°$ plane, the shadowed area is approximated by a trapezoid and the voltage gain function is given by

$$G_4(\theta, 0°) \cong \frac{\pi D_m}{\lambda} \left[\frac{16a}{\pi D_m^2}\right]\left[\frac{1 + \cos \theta}{2}\right] \left\{ \left(b + \frac{\omega}{2}\right) \frac{\sin \left[\left(b + \frac{\omega}{2}\right) \frac{\pi}{\lambda} \sin \theta\right]}{\left(b + \frac{\omega}{2}\right) \frac{\pi}{\lambda} \sin \theta} \right.$$

$$\cdot \frac{\sin\left[\left(b - \frac{\omega}{2}\right) \frac{\pi}{\lambda} \sin \theta\right]}{\left(b - \frac{\omega}{2}\right) \frac{\pi}{\lambda} \sin \theta}$$

$$+\left(b+\frac{\omega}{2}\right)\frac{\sin\left(\frac{2\pi a}{\lambda}\sin\theta\right)}{\frac{2\pi a}{\lambda}\sin\theta}\pm j\left(b-\frac{\omega}{2}\right)$$

$$\cdot\left[\frac{\sin\left(\frac{2\pi a}{\lambda}\sin\theta\right)-\left(\frac{2\pi a}{\lambda}\sin\theta\right)\cos\left(\frac{2\pi a}{\lambda}\sin\theta\right)}{\left(\frac{2\pi a}{\lambda}\sin\theta\right)^2}\right]\Bigg\} \qquad (10)$$

where a and b are as defined in Figure 7a and phase terms have been suppressed. The envelope of equation (10) is plotted in Figure 8 for $D_m/\lambda = 400$.

Calculation in the $\phi = 45°$ plane is simplified by approximating the shadowed area as shown in Figure 7b. With this approximation, the voltage gain function can be written as

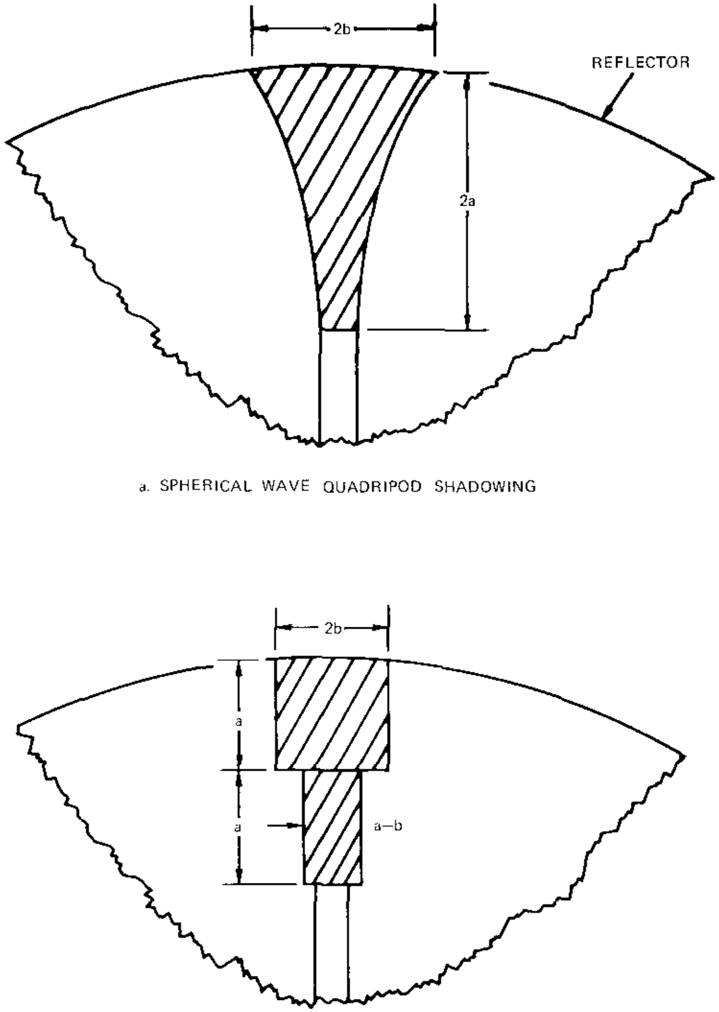
$$g_4(\theta, 45°) \cong \frac{\pi D_m}{\lambda}\left[\frac{16ab}{\pi D_m^2}\right]\left[\frac{1+\cos\theta}{2}\right]\left[\frac{\sin\left(\frac{\sqrt{2}}{2}\frac{\pi a}{\lambda}\sin\theta\right)}{\frac{\sqrt{2}}{2}\frac{\pi a}{\lambda}\sin\theta}\right]$$

$$\cdot\left\{\frac{2\sin\left(\frac{\sqrt{2}\pi b}{\lambda}\sin\theta\right)}{\frac{\sqrt{2}\pi b}{\lambda}\sin\theta}+\frac{\sin\left(\frac{\sqrt{2}}{2}\frac{\pi b}{\lambda}\sin\theta\right)}{\frac{\sqrt{2}}{2}\frac{\pi b}{\lambda}\sin\theta}\right\} \qquad (11)$$

where, as before, phase terms are neglected. The envelope of equation (11) is shown in Figure 9 for $D_m/\lambda = 400$.

Surface tolerance effect

Formulations for the average sidelobe envelope of a randomly deformed reflector antenna have been developed and could be applied here to calculate the typical effect of random errors [6]. However, recent photogrammetric measurements of 97-ft-diameter reflector surfaces have clearly indicated that deformations are largely systematic; i.e., the surface errors are correlated over large portions of the surface. Application of the random error theory for the case of large correlation intervals results in a predicted scattering pattern which is directive and largely contained in the region



a. SPHERICAL WAVE QUADRIPOD SHADOWING

b. APPROXIMATION TO SPHERICAL WAVE SHADOWING
$a = 0.112\ D_m$
$b = 0.0226\ D_m$

Figure 7. *Spherical Wave Quadripod Shadowing*

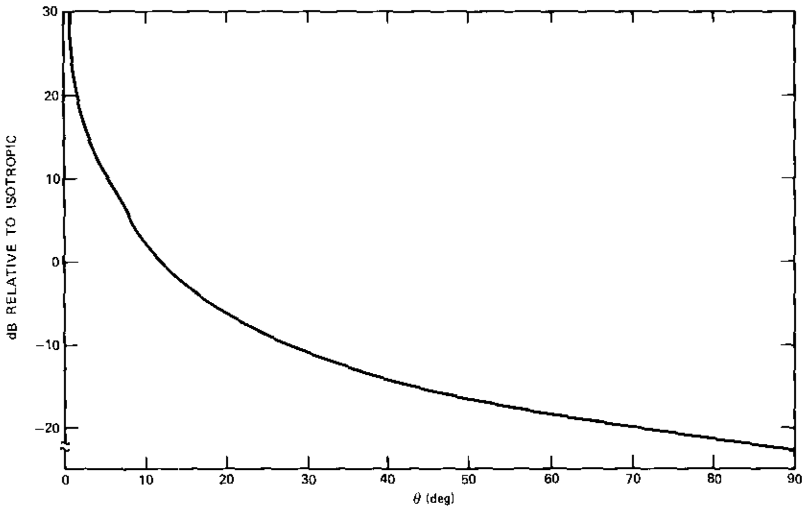


Figure 8. *Spherical Wave Scattering Envelope* (ϕ = *0°*, D/λ = *400*)

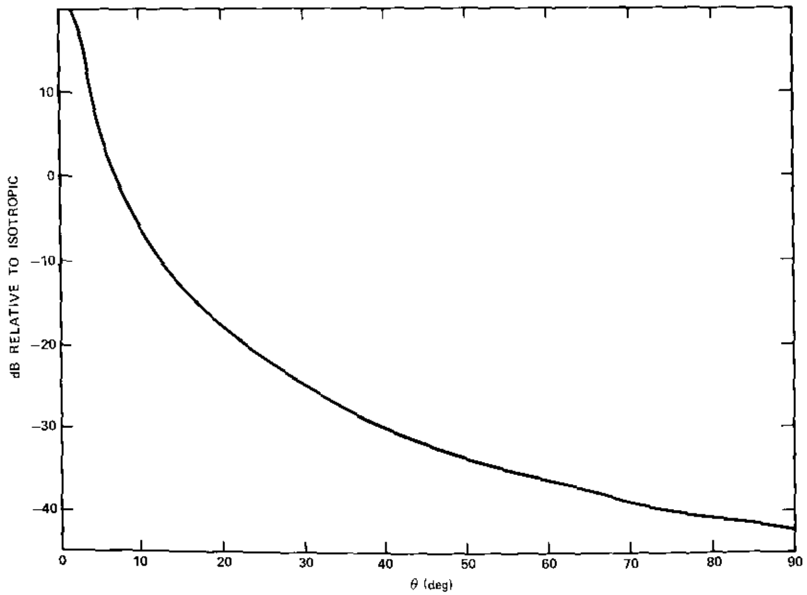


Figure 9. *Spherical Wave Scattering Envelope* (ϕ = *45°*, D/λ = *400*)

of the main lobe and first few sidelobes. Furthermore, the magnitude of the deformation is small and gain measurements have shown that the resulting gain loss is also small. In addition, pattern measurements have shown a well-defined sidelobe structure, suggesting that some of the systematic error is offset by experimentally optimizing the position of the subreflector and feed [7]. These arguments justify neglecting the effects of random reflector deformation on the wide-angle sidelobe envelope.

The feed system

A reasonable prediction of the effects of the feed system on the wide-angle sidelobe envelope can be achieved by assuming a representative feed geometry. However, it is difficult to predict the feed effects in the vicinity of the subreflector edge ($\theta = \theta_s$). The scattered field close to $\theta = \theta_s$ is a complicated function involving the feed radiation intensity at $\theta = \theta_s$, its angular derivatives evaluated at $\theta = \theta_s$, the diameter of the subreflector, and the depth or curvature of the subreflector. Since an exact formulation of a typical feed-subreflector system is well beyond the scope of the present work, the limitations of a simplified analysis are acknowledged and errors in predictions close to $\theta = \theta_s$ are anticipated.

For present purposes, a uniformly illuminated circular aperture, 5 λ in diameter, is taken as the feed element. This feed generates a pattern which closely represents the scatter diagram of a high-efficiency feed and subreflector, at least in the terms of the envelope of the peaks for $\theta > \theta_s$ [8]. Mathematically, the feed pattern envelope is written as

$$\overline{g_5(\theta, \phi)} = 0.2 \cot\frac{\theta}{2} \sin^{-1/2}\theta, \qquad \theta > 18.5° \tag{12a}$$

$$\overline{g_5(\theta, \phi)} = 2.21 \left(\frac{\theta}{18.5}\right)^2, \qquad \theta < 18.5° \tag{12b}$$

where it has been assumed that $\theta_s = 12°$. Equation (12a) is simply the envelope of the wide-angle radian field of the uniform circular aperture. Its resemblance to equation (5) is evident. Equation (12b) is an empirically derived function chosen to approximate the field behavior in the vicinity of θ_s. Choosing the feed pattern envelopes in this manner makes it possible to omit the step function in equation (1) and at the same time to achieve a more accurate representation of feed radiation. Equation (12) is plotted in Figure 10.

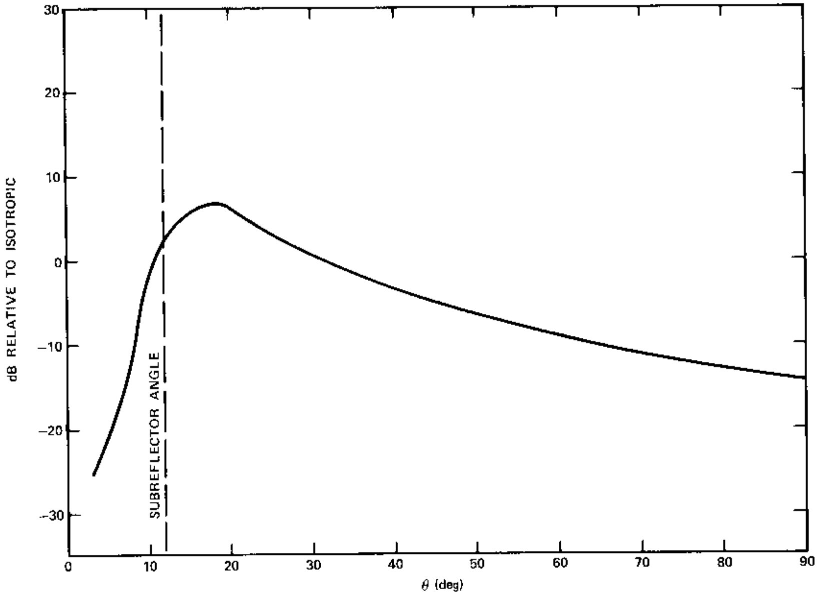


Figure 10. *Wide-Angle Radiation Envelope of the Feed System*

The composite sidelobe envelope

From equation (1) the composite radiation envelope is given by the sum of the envelopes developed in the previous section. This sum has been evaluated for the 0° and 45° planes and for $D_m/\lambda = 400$; the results are plotted in Figures 11 and 12.

Comparison of experimental and theoretical data

The antenna for which experimental data were available was a 97-ft-diameter shaped Cassegrain antenna at a test frequency of 4 GHz. With the exception of the quadripod angle ξ, which was 54° rather than 45°, the model used herein was a close representation of this antenna. The calculations were modified to account for this difference and the resulting composite envelope is shown in Figure 13. The agreement between this envelope and the measured envelope also shown in Figure 13 is quite good. The most significant difference appears in the vicinity of the subreflector angle and, as indicated earlier, disparity in this region is not surprising.

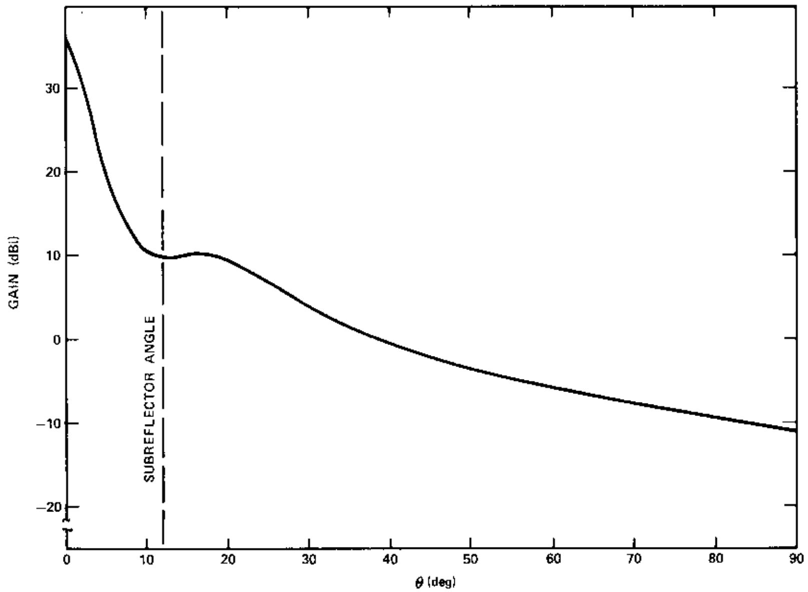


Figure 11. *Composite Radiation Envelope* ($\phi = 0°$, $D/\lambda = 400$)

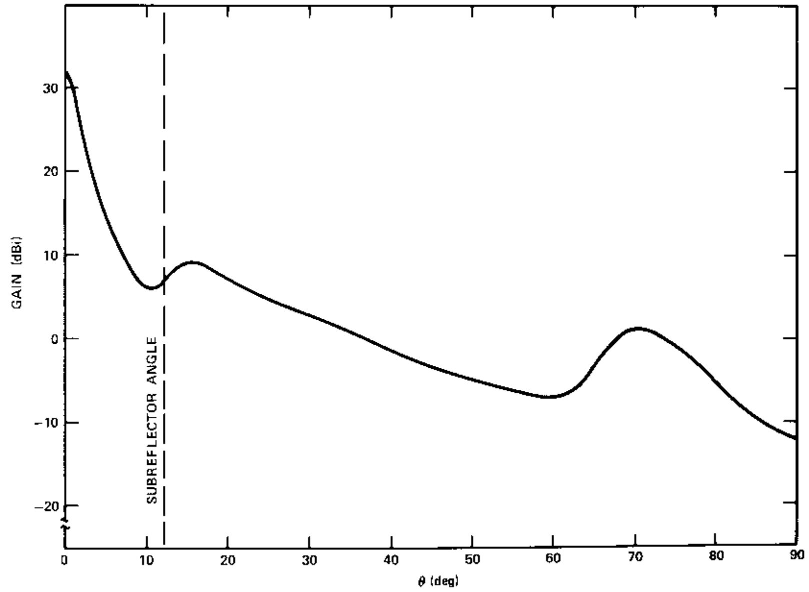


Figure 12. *Composite Radiation Envelope* ($\phi = 45°$, $D/\lambda = 400$)

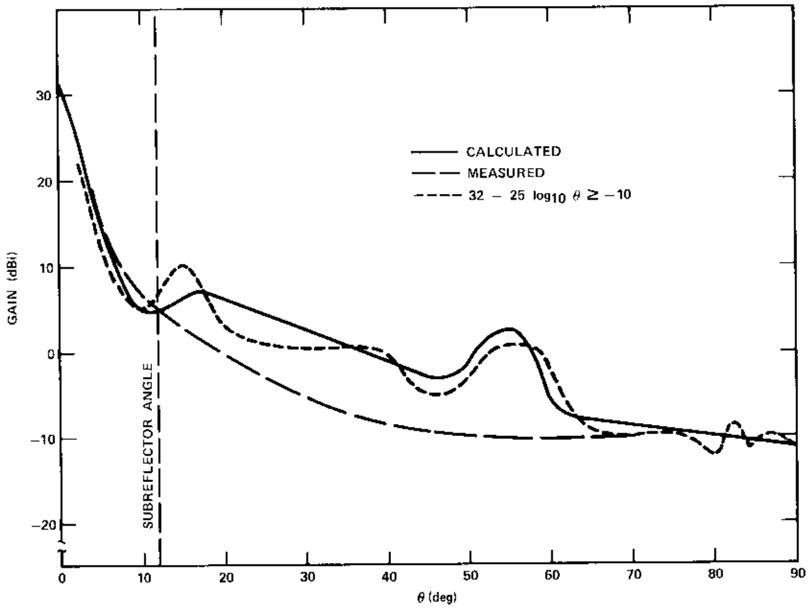


Figure 13. *Comparison of Measured and Calculated Sidelobe Envelopes (54° quadripod angle, $\phi = 45°$)*

The analysis presented in this paper was primarily intended for electrically large reflector configurations. However, as a matter of interest, the wide-angle peak sidelobe envelope of a Cassegrain antenna with $D_m/\lambda = 90$ was calculated using equations (5), (8), (11), and (12) and compared with measured data derived recently. The results are shown in Figure 14 along with the relevant geometric parameters of the antenna. The agreement is reasonably good, suggesting that the envelope equations derived in this paper have a fairly wide range of applicability.

Conclusion

Envelope expressions for the wide-angle peak sidelobes of a typical electrically large Cassegrain antenna have been presented. The envelopes have been compared with available measured data and the agreement is quite good, even for a reflector configuration for which the "electrically large" condition is only marginally satisfied.

The analytic approach presented for sidelobe envelope computation is relatively straightforward and simple. For example, the main reflector and subreflector diffraction is computed using aperture field methods rather than current integration techniques. This simplification results in negligible error, primarily because these diffraction terms are not major contributors to the wide-angle radiation envelope. The best evidence of the validity of the computational techniques used herein is that they predict to within acceptable errors the experimentally derived data.

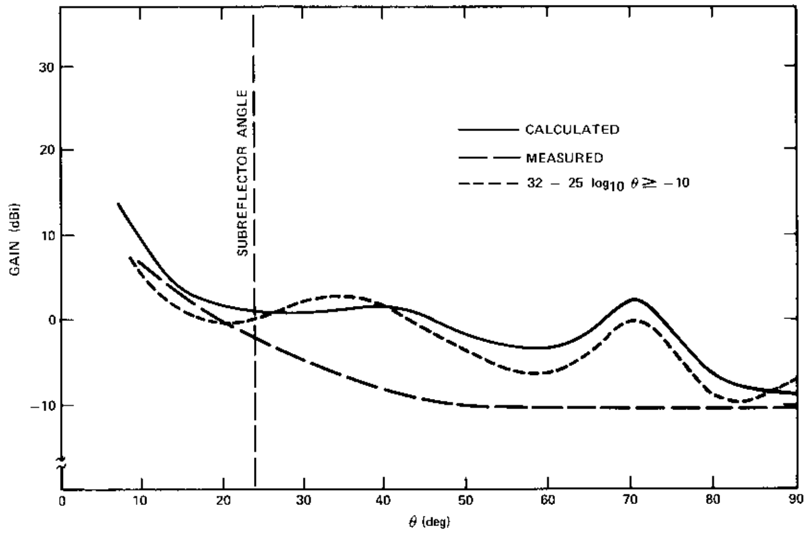


Figure 14. *Sidelobe Envelope for the Cassegrain Antenna ($\phi = 45°$, $D_m/D = 90$, $\theta_s = 24°$, $D_s/D_m = 0.139$, 45° quadripod angle)*

In Figures 13 and 14 the relationship $G = 32 - 25 \log_{10} \theta$ has been plotted for comparison. This expression has been recommended as a basis for interference calculations [9]. It should be noted that the predicted and measured sidelobe envelopes in both figures differ substantially from the levels predicted by this relationship. It should be added, however, that the $32 - 25 \log_{10} \theta$ relationship is based on "smoothed" sidelobe levels as opposed to the peak sidelobe levels presented herein. This difference accounts at least in part for the noted discrepancies.

Acknowledgment

Mr. Fred Frey assisted in performing some of the calculations presented. The experimental data on the 90-λ antenna were derived by Dr. W. English and Mr. R. W. Gruner. The experimental data on the 97-ft antenna were supplied by Mr. E. J. Wilkinson.

References

[1] W. Korvin and R. W. Kreutel, "Earth Station Radiation Diagram with Respect to Interference Isolation Capability: A Comparative Evaluation," *Communication Satellites for the 70's: Technology*, Cambridge, Mass.: M.I.T. Press, 1971.

[2] R. W. Kreutel, W. Korvin, and D. DiFonzo, "Wide Angle Radiation Characteristics of Antennas for Frequency Shared Communication Services," Antennas and Propagation International Symposium, September 1971.

[3] G. Hyde and R. W. Kreutel, "Earth Station Antenna Sidelobe Envelope Analysis," International Communications Conference, June 1969.

[4] V. Galindo, "Design of Dual-Reflector Antennas with Arbitrary Phase and Amplitude Distribution," *IEEE Transactions on Antennas and Propagation*, AP-14, No. 4, July 1964.

[5] *Microwave Antenna Theory and Design*, S. Silver ed., New York: McGraw-Hill Book Company, 1949, p. 192f.

[6] J. Ruze, "Antenna Tolerance Theory—A Review," *Proc. IEEE*, Vol. 54, No. 4, April 1966.

[7] P. R. Cowles and E. A. Parker, "Reflector Surface Error Compensation in Cassegrain Antennas," *AP-S Transactions*, AP-23, No. 3, May 1975.

[8] W. V. T. Rusch, "Scattering from a Hyperboloidal Reflector in a Cassegrain Feed System," *IEEE Transactions on Antennas and Propagation*, AP-13, No. 4, July 1963.

[9] International Radio Consultative Committee (C.C.I.R.) "Radiation Diagram of Antennae at Communications-Satellite Earth Stations for Use in Interference Studies," Report 391, *XIth Plenary Assembly, Oslo, 1966*, Geneva: International Telecommunication Union, 1967.

Randall W. Kreutel, Jr., received B.S.E.E. and M.S.E.E. degrees from Northeastern University. He served in the U.S. Navy from 1952 to 1956. Prior to joining COMSAT *he was a research engineer with the Antenna and Microwave Lab of Sylvania Electronic Systems, working on phased array, frequency-independent, and low-noise antennas. He joined* COMSAT *in 1966, and is presently Manager of the Antennas Department, responsible for research and development of antennas and antenna-related equipment.*

Mr. Kreutel is a senior member of IEEE, a member of Eta Kappa Nu, and a member of Commission I of URSI.

Index: unattended earth terminals, low-noise amplifier

The unattended earth terminal low-noise amplifier

S. CHOU, P. KOSKOS, AND W. GETSINGER

(Manuscript received October 4, 1975)

Abstract

This paper describes a low-noise amplifier designed and built for use in unattended earth terminals (UETs). The unrefrigerated amplifier provides 50-dB minimum gain and 75-Kelvin maximum noise temperature over the 3.7- to 4.2-GHz frequency band. The synchronously tuned paramp stages can be interchanged and replaced without interrupting the signal flow and without appreciable change in performance characteristics. The fail-soft aspects of the design preclude the need for a redundant amplifier and a low-loss front-end switch. Test results are presented, and design and construction features are described in detail.

Introduction

Parametric amplifiers (paramps) have been used as the low-noise front end in satellite communications earth station receivers for many years. To ensure service availability, redundant amplifiers and a low-loss switch system are presently used in most earth stations. As satellite communications continue to expand, many more earth terminals will be required, making it highly desirable to reduce both initial and operating costs. Elimination of the parametric amplifier redundancy is a significant step in this direction.

In late 1970 a research project to develop an unattended earth terminal

(UET) was initiated at COMSAT Laboratories. The general guidelines for designing such a system were as follows:

a. Reliability. The terminal should maintain overall high-quality service while unattended for relatively long periods of time. Service interruption due to component failures should be minimized.

b. Low cost. The initial and maintenance costs should be kept as low as possible.

c. Simplicity. A service specialist should be able to perform routine inspection and maintenance easily once every few months.

These objectives led to the design of a unique low-noise amplifier which operates at room temperature (uncooled) with 50-dB gain and an effective input noise temperature less than 75 Kelvin.

General Description

Figure 1 is a block diagram of the low-noise amplifier (LNA) unit. It consists of three 10-dB-gain paramp stages followed by a 20-dB-gain transistor amplifier. Interstage isolation is provided by a low-loss 10-port circulator.

Each paramp stage is self-contained in its individual enclosure and consists of a solid-state pump source, paramp varactor module, and temperature stabilization circuits. Figure 2 is a single-stage block diagram. The individual paramp stages and the circulator are pretuned so that any paramp stage module can be quickly replaced at any amplifier port of the circulator and the new module will provide essentially unchanged performance without retuning. The circulator port connectors are designed so that removal of a paramp stage creates a length of cutoff circular waveguide which reflects signal power losslessly from that port to the next amplifier stage. Thus, the signal flow is not interrupted by the removal of a stage, and only the gain of that stage is lost. The logic behind such an arrangement is as follows:

a. The failure of any single stage does not interrupt communications and at worst the system temperature is degraded by no more than 1 or 2 dB.

b. The failed stage can be replaced with another pretuned paramp plug-in stage by inexperienced personnel very easily and quickly without interrupting the signal flow. All the plug-in modules are fully interchangeable.

The paramp stage plug-in unit is guided in and out by two well-aligned guide rails on the main chassis. A large 2-turn lock nut is used to secure the RF connection between circulator and paramp stage.

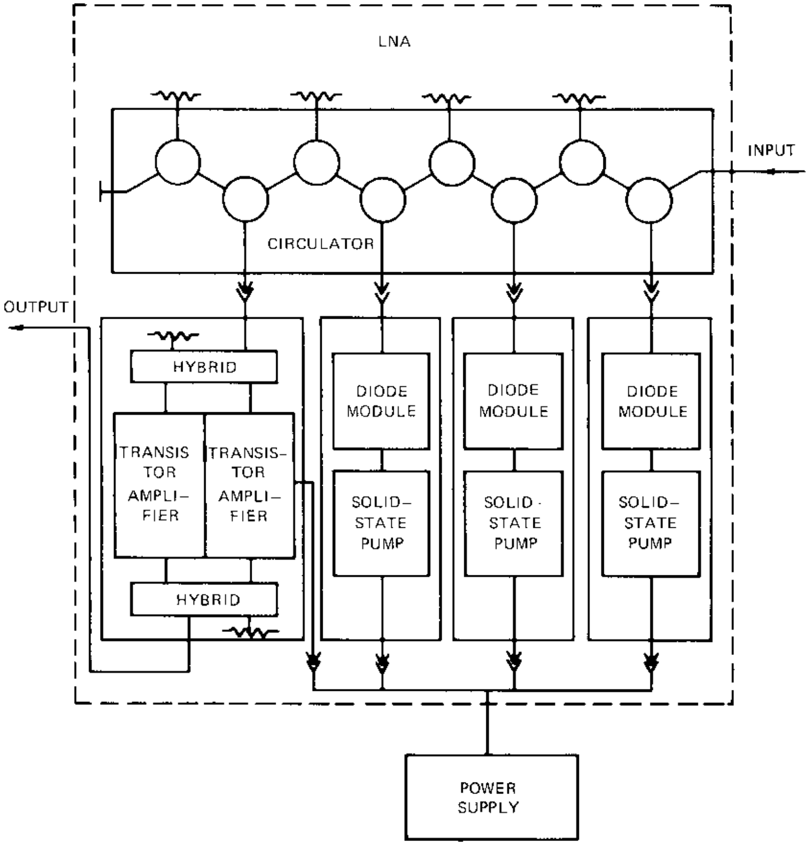


Figure 1. *Low-Noise Amplifier Block Diagram*

Finally, two 3-stage transistor amplifiers are connected in parallel between two microstrip hybrids to provide additional low-cost system gain with built-in redundancy. Figure 3 shows the front of the complete LNA unit. Each paramp stage has a test point select switch. A voltmeter can be plugged in to check the varactor bias and several monitor points in the temperature control circuit. In addition, a bias off/on switch is installed so that the performance of each individual stage can be checked without removing stages from the amplifier. The bias voltage and the operating temperatures of the diode module and pump source can be adjusted from the front panel behind a swing-open cover.

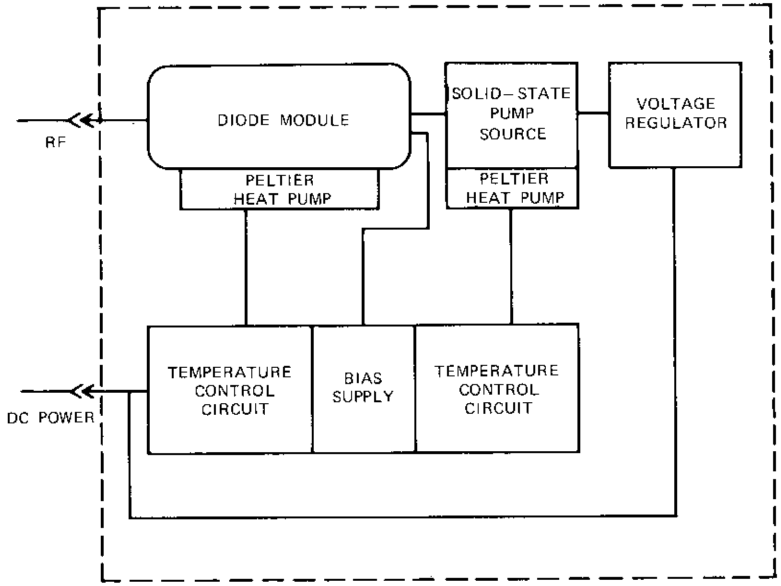


Figure 2. *Block Diagram of Paramp Plug-in Unit*

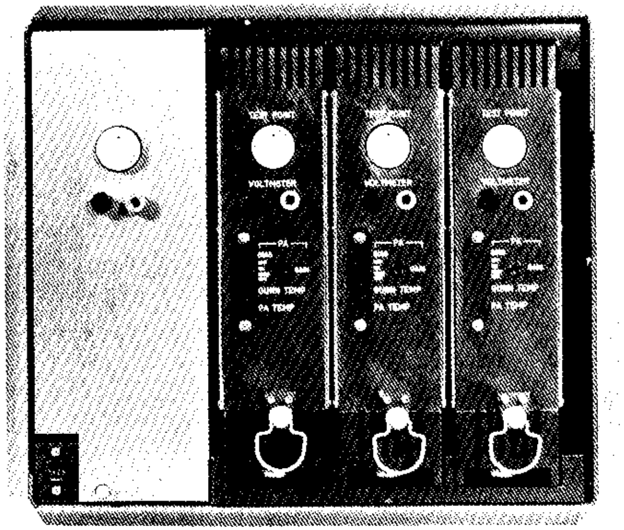


Figure 3. *Front View of the Low-Noise Amplifier*

Circulator

The multiport circulator, which is an extension of a circulator developed previously at COMSAT Laboratories, consists of eight 3-port units. Four of these units are terminated with a 50-ohm load at one port and behave as isolators to provide isolation between the amplifier ports.

Figure 4 is an interior view of the circulator. The ferrite cylinders are held in properly located holes in two flat dielectric plates. A shaped center conductor is sandwiched between the dielectric plates to connect the circulator junctions. Low-loss ground planes are provided by two 0.002-inch-thick copper foil sheets which are evenly pressed on the outer surfaces of the dielectric plates by a thin flat layer of pure indium acting as a soft cushion between the foil and the aluminum outer case. The center conductor is made of 0.020-inch-thick copper shim stock with a chemical machining process used to cut the pattern from both sides. A single-step impedance transformer is used to match the junction impedance to 50-ohm

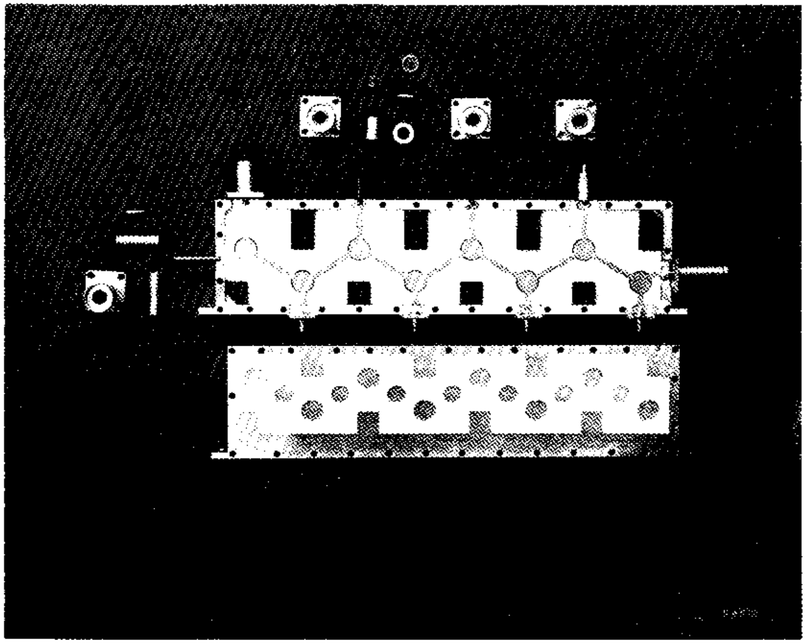


Figure 4. *Interior View of the Circulator*

line. To provide sufficient physical separation between the amplifier ports to accommodate the amplifier modules, an extra quarter-wavelength 50-ohm stripline is inserted between each pair of circulator junctions. The broadband performance is improved by the conjugate impedance matching action of the quarter-wave line.

DC isolation at the paramp ports, which is necessary to permit individual biasing of each varactor stage, is provided by breaking the center conductor of the circulator 50-ohm coaxial line section and inserting a piston and cylinder-type capacitor as shown in Figure 5. The piston has a length of about 90 electrical degrees at the band center frequency and is wrapped with 0.001-inch-thick mylar. This DC block causes no observable performance degradation and very little insertion loss (less than 0.05 dB). The RF connectors are modified APC-7 precision connectors with an add-on adapter which is used to quickly engage or disengage the paramp plug-in unit from the circulator.

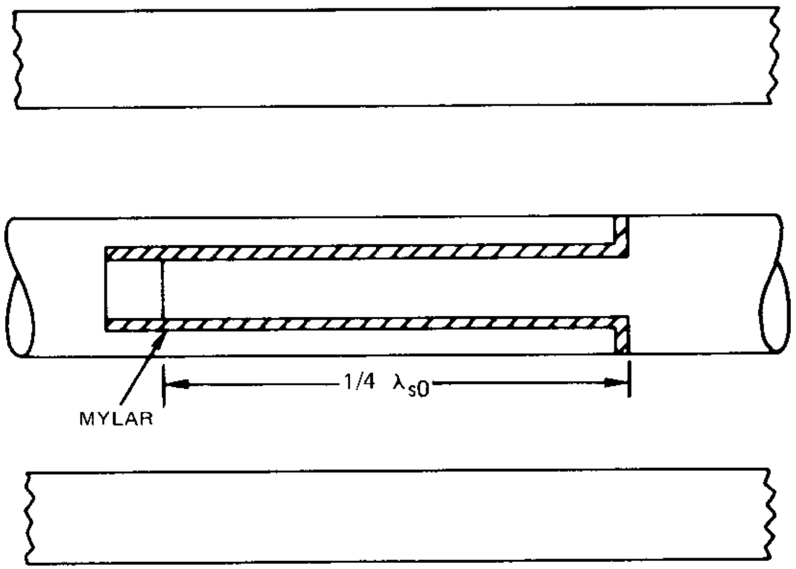


Figure 5. *DC Block in Signal Circuit* (*λ_{s0} is the wavelength at signal frequency band center*)

Within the circulator, eight RF absorbent blocks of the same height as the main dielectric are spread along the zig-zag transmission line to suppress the propagation of spurious modes and hence improve isolation and reduce loss.

The major performance parameters of the circulator over the 3.7- to 4.2-GHz passband are summarized in Table 1.

TABLE 1. CIRCULATOR PERFORMANCE

Insertion Loss	<0.1 dB per pass
VSWR	
Input	<1.25
Paramp Ports	<1.05
Output	<1.15
Isolation	
First Amplifier Port to Input	>25 dB
Between Amplifier Ports	>55 dB
Junction Impedance	16.6 Ω

Paramp plug-in unit

Figure 6 is a photograph of the interior of a paramp plug-in unit. The

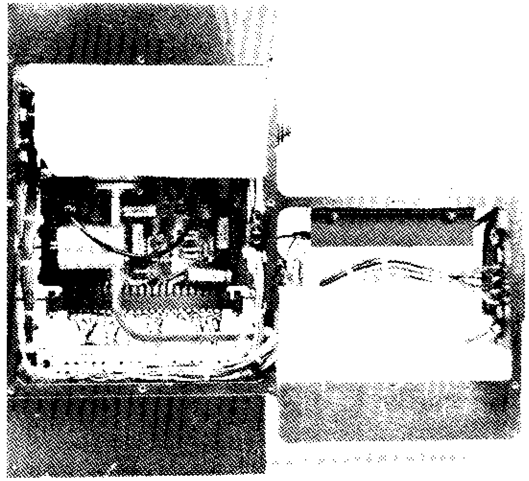


Figure 6. *Paramp Plug-in Unit*

unit consists of three major parts: the paramp varactor module, the solid-state pump source, and a printed circuit board with temperature stabilization and bias circuitry.

Paramp diode module

Figure 7 is a cutaway view of the amplifier mount. This configuration has been adopted for the following reasons:

a. it provides high isolation between pump and idler circuits,

b. the ready accessibility of the varactor idler circuit greatly simplifies tuning adjustments,

c. no reduced height waveguide or stepped transformer is needed for pump matching and no stub tuners are needed for signal matching so that construction is simplified.

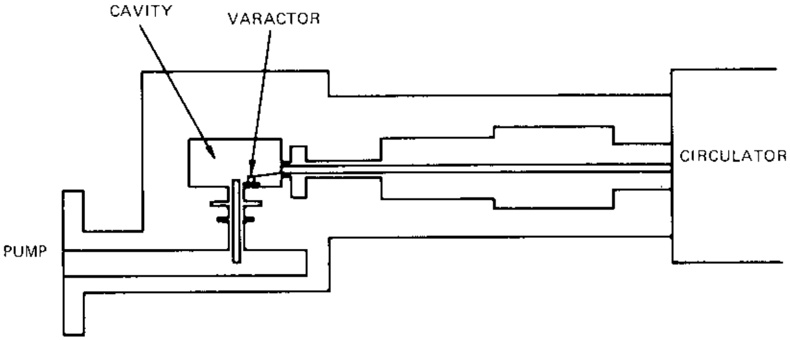


Figure 7. *Cutaway View of the Paramp Module*

The paramp is pumped at 45 GHz. The pump power is coupled to the cavity from standard WR-19 waveguide through a section of coaxial line in which two rejection filters, one for the idler and one for the upper sideband frequency, are cascaded to isolate the idler circuit from the pump circuit. The cavity is resonant at the pump frequency in the TM_{011} mode. A magnetic coupling loop, formed by a thin ribbon connecting the diode and signal line, is used for coupling pump power to the diode; the diameter of the cavity is determined by resonating the diode at the idler frequency in the TM_{010} mode. A small capacitive cap on top of the diode is used as a fine adjustment for this resonance. It has been found that varactors having a wide range of junction capacitance can be used without significant changes in the main circuit.

The length of the ribbon which joins the diode and signal line is designed to be a quarter wavelength at the idler frequency. One end of this ribbon is connected to a very-low-impedance coaxial section followed immediately by a high-impedance coaxial section. Therefore, in terms of the idler frequency, the diode is completely decoupled from the signal circuit. The equivalent circuit at the idler frequency is shown in Figure 8.

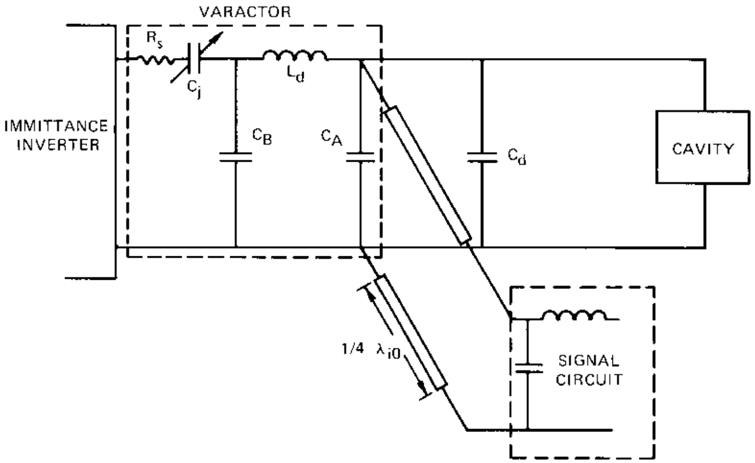


Figure 8. *Idler Equivalent Circuit* (λ_{i0} *is the wavelength at the center of the idler frequency band*)

The varactor is a GaAs diffused junction device in a min-dot package with package capacitance, $C_p = C_B + C_A$, of 0.12 pF and series inductance, L_d, of 0.13 nH [1]. The junction capacitance is 0.39 pF at zero bias.

The signal circuit design in a paramp is the most critical item in terms of obtaining the desired gain, gain flatness, and low noise temperature. The equivalent signal circuit shown in Figure 9 consists of four major sections: the varactor, the signal resonant circuit, the impedance transformer, and the broadbanding circuit.

The immittance inverter, shown at the left of the signal equivalent circuit in Figure 9, represents the coupling mechanism between the idler and the signal circuit. Through the inverter the nonlinear time-varying capacitance of the pumped varactor presents to the signal circuit an impedance with a negative real part and an imaginary part having a negative slope parameter. Therefore, the signal circuit must include an appropriate impedance transformation to obtain the desired gain level

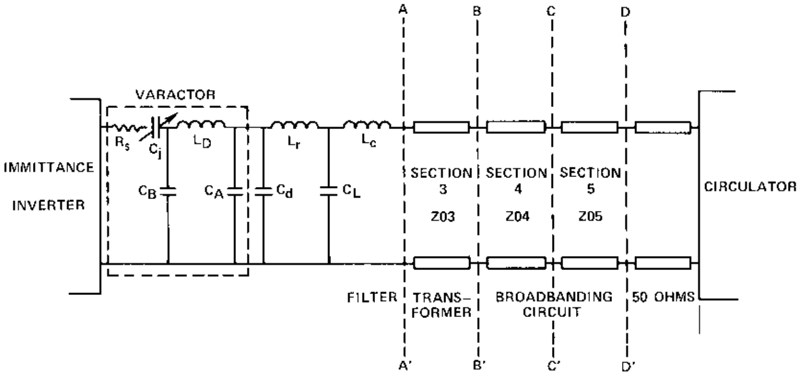


Figure 9. *Signal Equivalent Circuit*

and a circuit with a suitable positive slope parameter to achieve broad bandwidth with flat gain.

Three quarter-wavelength stepped impedance transformer sections (sections 3–5 in Figure 9) are used to fulfill the basic signal circuit requirements described above. They immediately follow plane A-A', where the impedance of the pumped varactor is resonated at center frequency using series inductances L_r and L_c, which also function as a high-frequency (idler and pump) rejection filter (together with the shunt capacitance, C_L).

A computer program [2] that simulates the paramp circuits has been used to facilitate the design work, and a series of impedance plots on a Smith chart at various reference planes has been used to reach the optimum design. Figures 10 through 13 are typical impedance plots of different reference planes. Figure 14 shows the computer-predicted gain performance together with actual measured paramp gain performance for a single stage.

Pump source

A Gunn effect oscillator is used as the pump source. It is mechanically tunable and delivers a minimum output power of 100 mW at the pump frequency, 45 GHz.

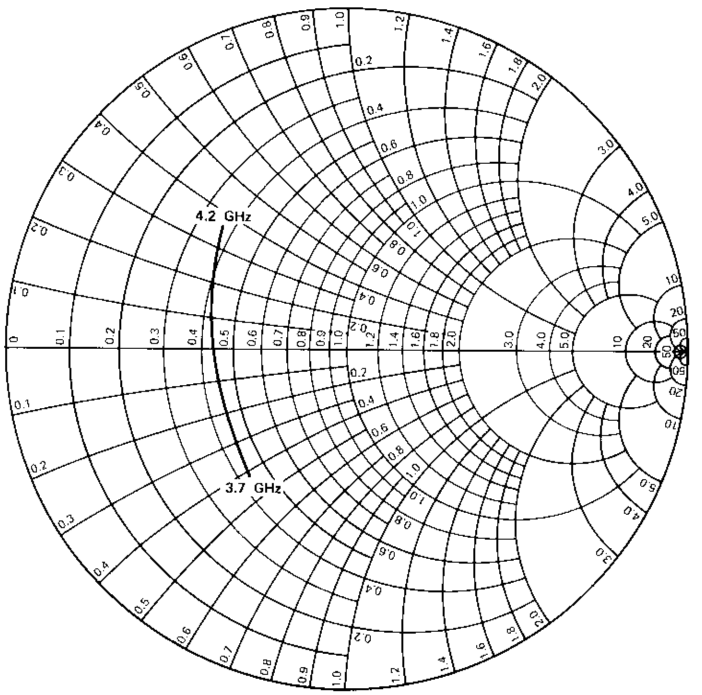


Figure 10. *Normalized Impedance Plot at Plane A-A'* $(-R_a + jX_a)/Z03$, *Z03: Characteristic Impedance of Section 3*

Temperature stabilization circuit

The paramp diode module and the pump source are temperature stabilized using Peltier thermoelectric heat pumps with solid-state feedback controllers. They operate at 30°C and 35°C, respectively, with the temperature held to within 2°C over an ambient range of 10°C to 50°C. The paramp and pump source are thermally insulated with styrofoam.

At normal ambient conditions, the temperature stabilization circuits require about 2 watts per stage. The power increases to 35 watts per stage at both high and low ambient temperature limits. The circuit of the temperature controller and the varactor bias supply are built on a 2.5- × 4.5-inch printed circuit board for each module. A bias off/on switch and all the adjustment controls can be reached from the front panel.

Test results and unique features

Three LNAS have been built and tested. Each amplifier consists of three paramp stages followed by a transistor amplifier as described above.

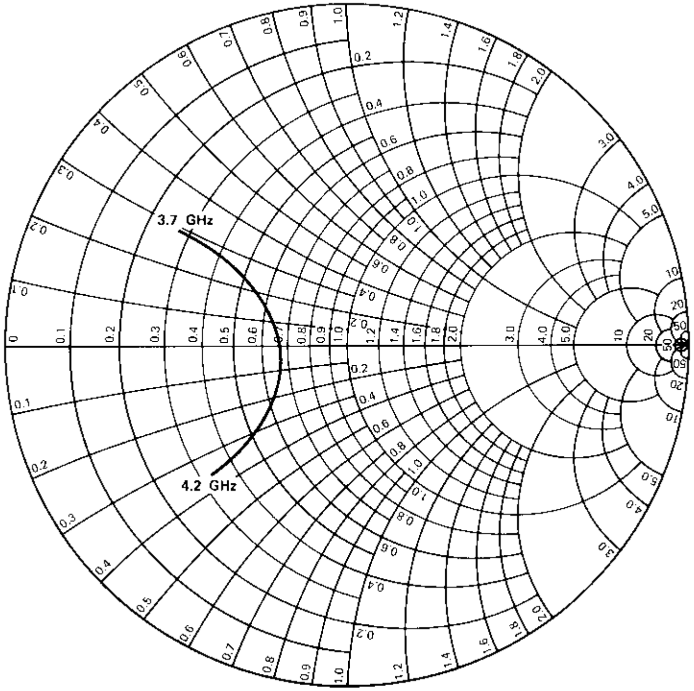


Figure 11. *Normalized Impedance Plot at Plane B-B′* $(-R_b + jX_b)/Z04$, *Z04: Characteristic Impedance of Section 4*

Typical measured characteristics of the amplifier are summarized in Table 2.

Three unique features characterize the design of the amplifier. First, a flat gain response can be achieved without complex tuning or adjustment in the signal line. No tuning mechanisms, such as open-end stubs, are needed in the signal circuit. This design permits the paramp modules to be electroformed if desired for high-volume low-cost applications. Figure 15 shows a typical 3-stage paramp gain response.

TABLE 2. MEASURED AMPLIFIER CHARACTERISTICS

Gain	50-dB min. ± 0.35 dB over 3.7–4.2 GHz
Gain Stability	±0.75 dB over 10°C–50°C
Dynamic Range	Gain compression < 0.5 dB at input levels ≤ −44 dBm
Amplitude Linearity	C/I > 40 dB with two equal carriers at −70 dBm each at minimum separation of 5 MHz
Noise Temperature	<75 K
Input and Output VSWR	1.25:1 max.

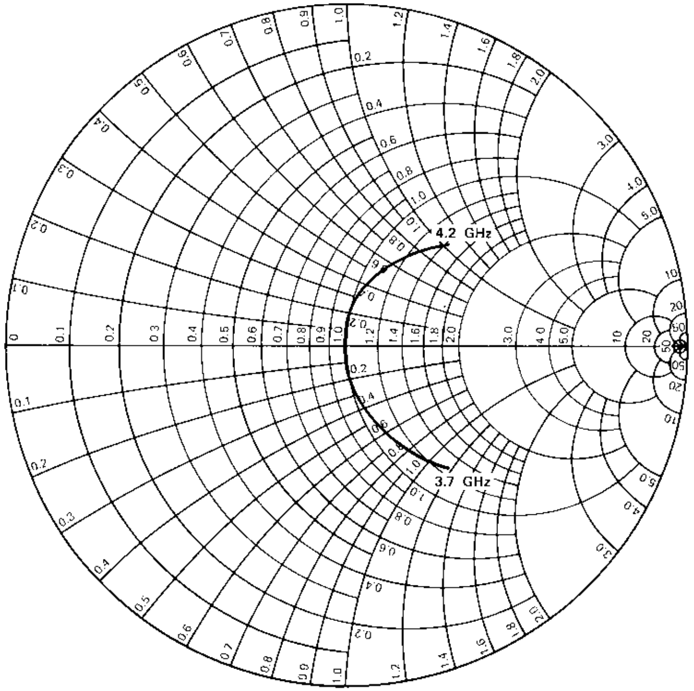


Figure 12. *Normalized Impedance Plot at Plane C-C′* $(-R_c + jX_c)/Z05$, *Z05: Characteristic Impedance of Section 5*

Secondly, the paramp plug-in units are fully interchangeable. They can be adjusted on a test circulator and connected to any amplifier port with no change in the gain response. Figure 16 presents a series of gain responses after such interchanges with three paramp plug-in modules on one receiver.

Finally, a malfunctioning paramp stage can be easily and quickly replaced, without breaking the signal line or interrupting service, by inexperienced personnel using specially designed quick connect/disconnect RF connectors between the paramp stages and circulator ports.

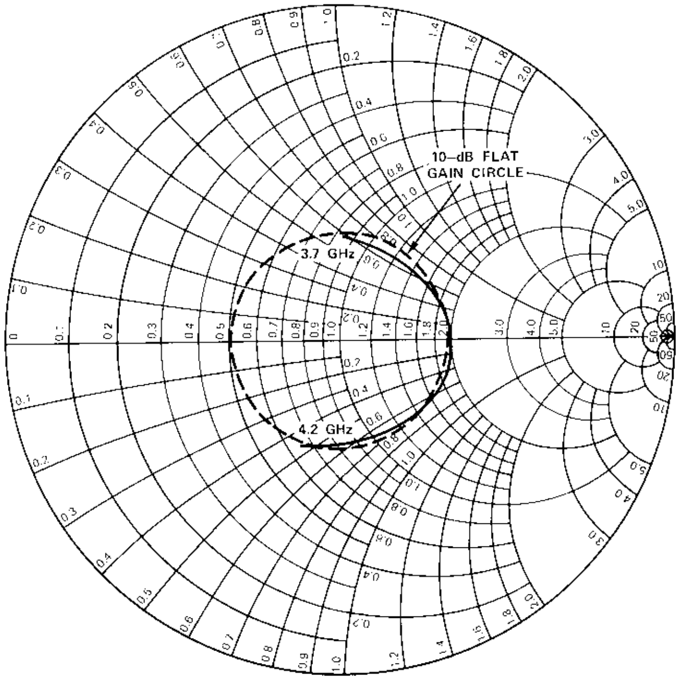


Figure 13. *Normalized Impedance Plot at Plane D-D′ $(-R_d + jX_d)/50$*

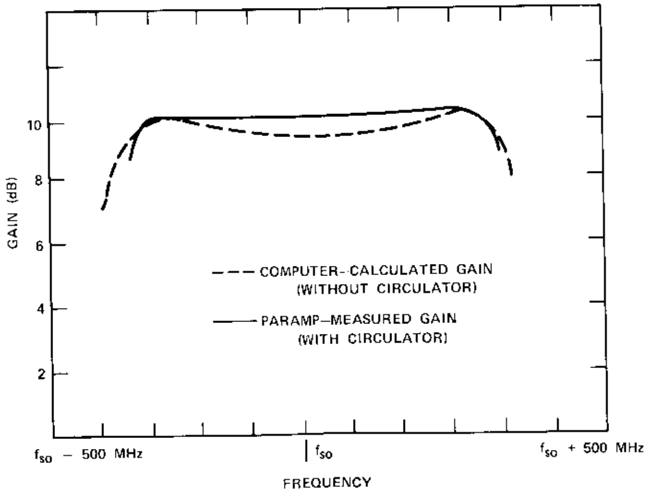


Figure 14. *Comparison of Calculated and Measured Gain of Single-Stage Paramp*

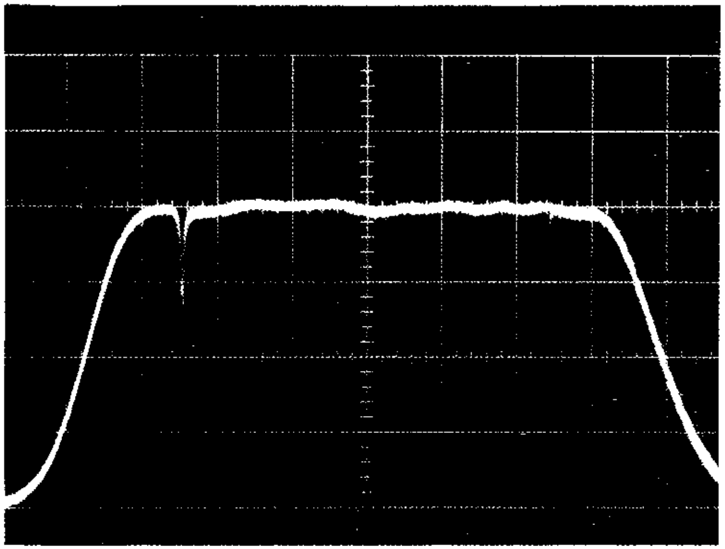


Figure 15. *Gain Response of 3-Stage Paramp (gain: 30 ± 0.15 dB, markers: 3.7 and 4.2 GHz, vertical scale: linear in power, horizontal scale: 100 MHz/cm)*

Acknowledgment

The authors wish to express their appreciation to many co-workers at COMSAT *Labs who contributed to the successful completion of this program. To list them all would require more space than is available, but a partial list must include J. Molz, J. Jerome, L. Sparrow, R. Kessler, N. Miller, W. Sones, and L. Pollack, who have provided technical support and encouragement.*

References

[1] W. Getsinger, "The Packaged and Mounted Diode as a Microwave Circuit," *IEEE Transactions on Microwave Theory and Techniques*, MTT-14, No. 2, February 1966, pp. 58–59.

[2] W. Getsinger and A. Kessler, "Computer-Design of Diode Using Microwave Components, and A Computer-Dimensioned, X-Band Parametric Amplifier," *Microwave Journal*, Vol. 12, No. 3, March 1969, pp. 119–123.

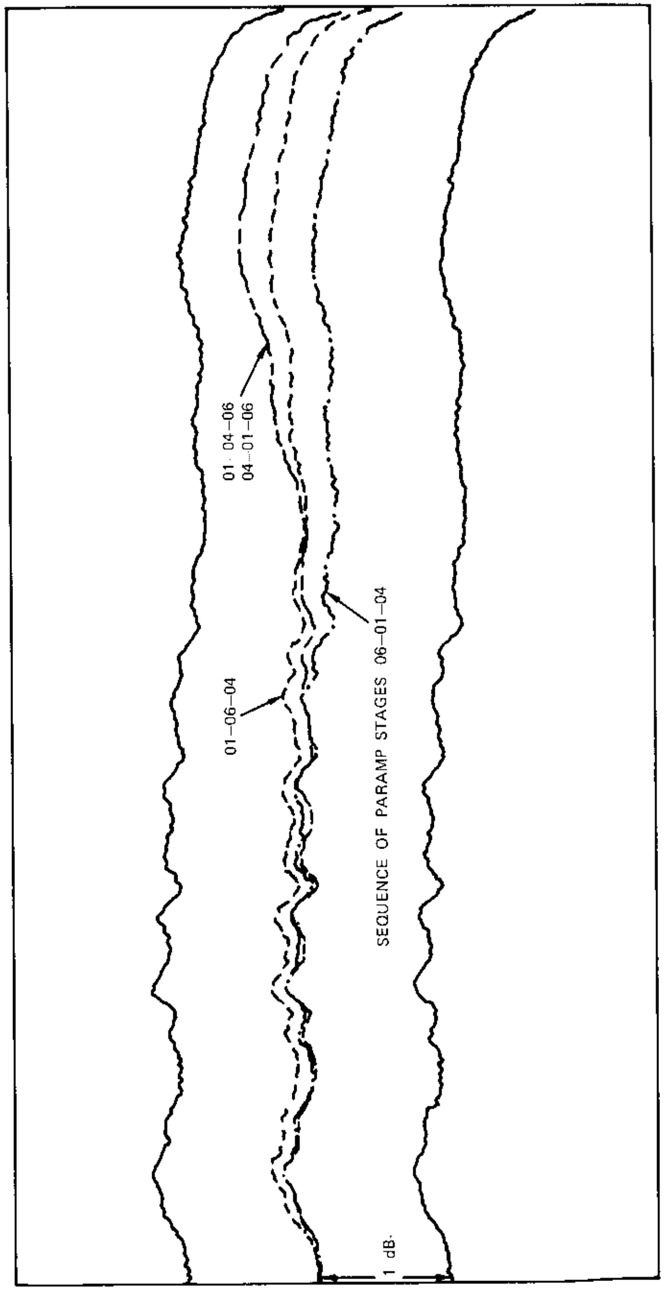


Figure 16. *Interchangeability Test*

Su Min Chou received a B.S.E.E. from the Chinese Naval College of Technology in 1954, an M.S. from Chiao Tung University in 1961, and a Ph.D. from the University of Utah in 1967. Prior to joining COMSAT *Labs, he was an engineer in a Chinese shipyard. He is presently a member of the technical staff of the RF Transmission Laboratory at* COMSAT *Laboratories working on low-noise amplifier development for advanced communications systems.*

Paul Koskos received a B.S.E.E. from Columbia University and an M.S.E.E. from New York University; he has continued his postgraduate studies at George Washington University. Prior to joining COMSAT, *he taught briefly at Brooklyn Polytechnic University. At Western Electric, Sylvania Laboratories, and RCA, he was responsible for product engineering, design, and development of vacuum tubes and semiconductor devices. At RCA, his primary responsibilities were traveling wave tubes and parametric amplifiers.*

After joining COMSAT, *he worked in the RF Laboratory on low-noise receivers. He has been Assistant Director, Reliability and Quality Assurance, since 1973. He is a senior member of IEEE and holds five patents on electronic devices.*

William J. Getsinger was born in Waterbury, Connecticut, on January 24, 1924. He received a B.S.E.E. from the University of Connecticut in 1949 and an M.S.E.E. and the degree of Engineer in electrical engineering from Stanford University in 1959 and 1961, respectively. Since 1950 he has worked on microwave components at Technicraft Laboratories, the Westinghouse Electric Company, Stanford Research Institute, and M.I.T. Lincoln Laboratories. In 1969 he joined COMSAT *Laboratories, where he is presently Manager of the Microwave Circuits Department.*

8108: 2 CH

[illegible] Paul [illegible] FET

[illegible] microwave switch network (MSN

~~[illegible]~~ SV8 MMIC dual gate FET,

SSPA

[illegible] Ku-band 24-horn phased array

[illegible]

Index: phase shift keying, demodulation, oscillator phase noise

Effects of oscillator phase noise on PSK demodulation

C. J. Wolejsza, Jr.

(Manuscript received October 28, 1975)

Abstract

A considerable interest has recently evolved in the feasibility of digital transmission involving bit rates not exceeding 9,600 bps via satellite for teletype and other applications. At these rates the inherent phase noise of up- and down-converters and satellite local oscillators becomes a significant factor in the design of carrier tracking loops for direct coherent digital modulation. This paper presents a characterization of oscillator phase noise which can be related to measurable parameters. A set of equations is then derived which allows the measured phase noise parameters to be related to the expected performance of a biphase PSK demodulator. Finally, the results of this analysis are compared with measured performance and a typical application to satellite system design is described.

Introduction

The expansion of satellite communications has generated strong interest in new types of services that can be provided via satellite. In particular,

This paper is based upon work performed in COMSAT Laboratories under the joint sponsorship of the Communications Satellite Corporation (COMSAT) and the COMSAT GENERAL Corporation.

the feasibility of digital transmission involving bit rates not exceeding 9,600 bps for teletype and other applications is under study. At these rates the inherent phase noise of up- and down-converters and satellite local oscillators becomes a significant factor in the design of systems which use direct, coherent, digital modulation techniques. A number of authors have addressed the phase noise problem, including the effects of this oscillator noise on a binary PSK (BPSK) system [1]-[4]. This paper derives a simple closed form set of equations which makes it possible to calculate the expected performance of a BPSK demodulator in terms of the signal-to-noise (S/N) ratio from measurable parameters of the oscillator phase noise.

Channel model

Figure 1 is a general channel model for the system under consideration. It can be seen that three oscillators contribute to the phase noise impressed on the carrier as it passes through the system. The output of these oscillators is at microwave frequencies derived by frequency multiplication from crystal oscillators operating at much lower frequencies. As a consequence of the frequency multiplication process, any phase noise on the original signal is increased in direct proportion to the multiplication factor, thereby causing spectrum spreading. At bit rates less than about 9,600 bps, the "bandwidth" resulting from the frequency multiplication process for normal C-band oscillators becomes a significant fraction of the signal bandwidth. Therefore, the degradation due to phase noise must be included in the calculation of overall performance.

This paper will deal with the specification of oscillator noise and the use of this specification to derive the spectral parameters of the phase noise. The combined effects of phase and thermal noise on coherent carrier recovery can be derived from the phase noise spectrum to yield a quantitative expression for the rate of cycle skipping and the effect of increased cycle skipping on the average bit-error rate. This expression can then be used to calculate bit-error rate performance as a function of oscillator noise and carrier-to-noise power density (C/N_o) ratio.

For a 1,200-bps BPSK demodulator in the presence of both thermal and phase noise, the theoretical performance, obtained by using the method described herein, is compared with the measured performance. The test procedure is described and application to a typical satellite system design is discussed.

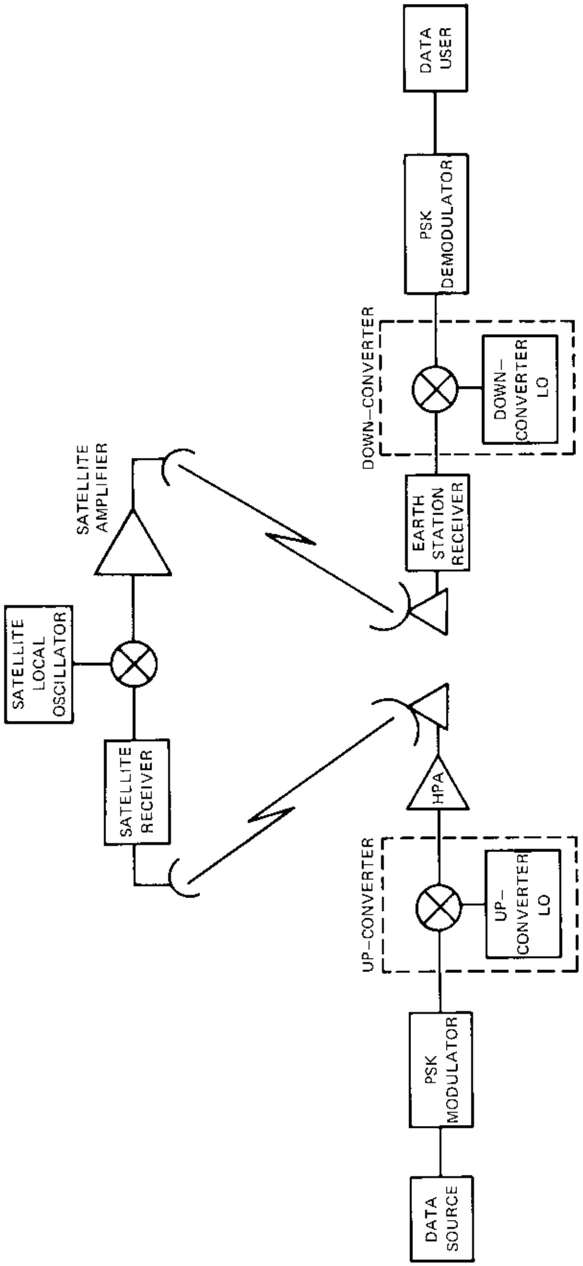


Figure 1. *General Channel Model*

Effects of phase noise

The phase noise in the received signal is specified in terms of the rms frequency deviations in two bandwidths, from ω_1 to ω_2 and from ω_2 to ω_3. This method of specifying phase noise has been chosen since it is commonly used and since it also allows easy measurements. An alternative technique for determining the phase noise characteristics will be discussed in the following section.

In general, the output of an oscillator can be modeled in the form

$$s(t) = A \sin \{\omega_c t + \phi(t)\} \tag{1}$$

where $\phi(t)$ is the phase noise by definition of ω_c and therefore $\overline{\phi(t)} = 0$. The phase power spectrum is assumed to have the following form [1], [3]:

$$S_\phi(\omega) = \frac{C_1}{\omega^3} + \frac{C_2}{\omega^2} + \frac{C_3}{\omega} + C_4 \tag{2}$$

where C_1 through C_4 are suitably chosen constants.

The first two terms, which normally dominate at low frequencies, will be retained.* Hence, the power spectrum of the frequency modulation is given by

$$S_\omega(\omega) = S_{\dot{\phi}}(\omega) = \omega^2 S_\phi(\omega) = \frac{C_1}{\omega} + C_2 \quad . \tag{3}$$

The mean square frequency deviation in a band from ω_1 to ω_2 rad/s is

$$\omega_{rms}^2(\omega_1, \omega_2) = \int_{\omega_1}^{\omega_2} S_\omega(\omega)\, d\omega \quad . \tag{4}$$

From equation (3),

$$\omega_{rms}^2(\omega_1, \omega_2) = \int_{\omega_1}^{\omega_2} \frac{C_1}{\omega}\, d\omega + \int_{\omega_1}^{\omega_2} C_2\, d\omega$$

$$\omega_{rms}^2(\omega_1, \omega_2) = C_1 \ln \left\{\frac{\omega_2}{\omega_1}\right\} + C_2(\omega_2 - \omega_1) \quad . \tag{5}$$

*In Reference 3 only the first term is retained. It can be shown that this is an inadequate description. Otherwise, the analysis in this subsection and in the subsection entitled "Calculation of total phase jitter" closely parallels that of Reference 3.

Application of equation (5) to the specifications makes it possible to find constants C_1 and C_2 so that the specification is just satisfied.

Cycle skipping rate

It has been found that the mean time to unlock for a second-order phase-locked loop is approximated by the empirical formula [5]

$$T_{AV} \cong \frac{2}{\omega_n} \exp \{\pi (S/N)_L\} \tag{6}$$

where ω_n is the loop natural frequency and $(S/N)_L$ is the signal-to-noise ratio in the loop bandwidth. If the high S/N ratio approximation is used for $(S/N)_L$ $[(S/N)_L \cong 1/2\sigma_{\phi L}^2]$ and the loop damping factor is 0.7071,

$$T_{AV} \cong \frac{1.06}{B_L} \exp \left\{\frac{\pi}{2\sigma_{\phi L}^2}\right\} \tag{7}$$

where B_L is the loop bandwidth, and $\sigma_{\phi L}$ is the rms phase jitter in the loop.

Some experimental results have shown [6] that the mean time during which the loop remains unlocked may be approximated by

$$T_{UL} \cong \frac{\pi}{\omega_n} \cong \frac{0.53\pi}{B_L} \quad . \tag{8}$$

Using equations (6) and (8) and assuming an error rate of 0.5 during an out-of-lock condition yields the following average probability of error contributed by cycle skipping:

$$P_{ecs} = P[\epsilon/cs]\, P[cs]$$

$$= \left(\frac{1}{2}\right) \frac{T_{UL}R}{T_{AV}R} \cong \frac{\pi}{4} \exp \left\{\frac{-\pi}{2\sigma_{\phi L}^2}\right\} \tag{9}$$

where R = bit rate

$P[\epsilon/cs] = \frac{1}{2}$

$$P[cs] = \frac{\text{number of bits during cycle skip}}{\text{total number of bits}} = \frac{T_{UL}R}{T_{AV}R} \quad .$$

Calculation of total phase jitter

The mean square phase noise for a "small" amount of phase jitter in a phase-locked loop may be related to the rms frequency deviation as follows:

$$\sigma^2_\phi(\omega_1, \omega_2) = \int_{\omega_1}^{\omega_2} S_\phi(\omega) \,|1 - H(\omega)|^2 \, d\omega \tag{10}$$

where $S_\phi(\omega)$ is defined in equation (3), $H(\omega)$ is the phase-locked loop linearized transfer function, and ω_1 and ω_2 define the bandwidth of measurement.

For a high-gain second-order loop [3], [5] with damping factor, $\zeta = 0.7071$:

$$H(\omega) = \frac{1 + j\omega\tau_p}{1 + j\omega\tau_p - (\omega^2\tau_p^2)/2} \tag{11}$$

where $\tau_p = \dfrac{3}{4B_L}$.

Then

$$\sigma^2_\phi(\omega_1, \omega_2) = \int_{\omega_1}^{\omega_2} \frac{C_1\omega \, d\omega}{\omega^4 + 4/\tau_p^4} + \int_{\omega_1}^{\omega_2} \frac{C_2\omega^2 \, d\omega}{\omega^4 + 4/\tau_p^4} \quad . \tag{12}$$

A damping factor of 0.7071 has been chosen because it simplifies the resultant integrals in equation (12) and because it is a commonly used value. Results for other damping factors may be found in a similar manner.

The first integral is calculated by substituting $Z = \omega^2$ and the second by partial fraction expansion. Since both integrals converge, the upper limit is taken as ∞ and the lower as 0, yielding an upper bound on σ^2_ϕ. Then, the contribution due to the local oscillator phase noise is

$$\sigma^2_\phi = \frac{9C_1\pi}{128B_L^2} + \frac{3C_2\pi}{16B_L} \quad . \tag{13}$$

The thermal noise contribution is given by

$$\sigma^2_{\phi th} = \frac{B_L}{\nu M(\mathrm{C/N_o})} \tag{14}$$

where ν = correction factor to account for the loss in S/N ratio due to frequency doubling or remodulation in the carrier recovery loop

M = correction factor to account for S/N ratio loss in the receive filter

$\mathrm{C/N_o}$ = carrier-to-noise density ratio at the down-converter output.

Therefore, summing (in a mean square sense) the total phase jitter yields

$$\sigma^2_{\phi T} = \frac{9C_1\pi}{128B_L^2} + \frac{3C_2\pi}{16B_L} + \frac{B_L}{\nu M(\mathrm{C/N_o})} \quad . \tag{15}$$

Because a squaring or Costas loop is used, the effective phase jitter which contributes to cycle skipping is

$$\sigma_{\phi L} = 2\sigma_{\phi T} \quad . \tag{16}$$

Hence, equation (9) becomes

$$P_{ecs} \cong \frac{\pi}{4} \exp\left\{-\frac{\pi}{8\sigma^2_{\phi T}}\right\} \quad . \tag{17}$$

Calculation of bit-error rate

To obtain the overall bit-error rate, equation (17) must be combined with the probability of error due to thermal noise. For cases of infrequent cycle skipping, errors due to thermal noise occur between cycle skips and are independent thereof. The thermal noise contribution is therefore combined as an independent variable. For both BPSK and QPSK (quadriphase PSK),

$$P_{eth} = Q\left\{\sqrt{\frac{2E_b}{N_o}}\right\} \tag{18}$$

where $Q(\alpha) \triangleq \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_\alpha^\infty e^{-\gamma^2/2} \, d\gamma = \dfrac{1}{2} \operatorname{erfc}\left(\dfrac{\alpha}{\sqrt{2}}\right)$

E_b = energy per bit
N_o = noise power density.

For large S/N ratios,

$$P_{eth} \cong \frac{\exp(-E_b/N_o)}{2\sqrt{\pi(E_b/N_o)}} \quad . \tag{19}$$

Hence, the combined probability of error is

$$P_e = P[\epsilon/cs]\,P[cs] + P[\epsilon/no\ cs]\,P[no\ cs] \tag{20}$$

where $P[\epsilon/cs]\,P[cs]$ is given by equation (17), and

$$P[\epsilon/no\ cs] \cong P_{eth}$$

$$P[no\ cs] = \left[1 - \frac{\pi}{2}\exp\left\{\frac{-\pi}{2\sigma^2_{\phi L}}\right\}\right] \cong 1$$

for cases of interest. Therefore,

$$P_e \cong P_{ecs} + P_{eth} \cong \frac{\pi}{4}\exp\left\{\frac{\pi}{8\sigma^2_{\phi T}}\right\} + \frac{\exp(-E_b/N_o)}{2\sqrt{\pi(E_b/N_o)}} \quad . \tag{21}$$

Experimental verification

To verify the assumptions made in the previous analysis and to demonstrate the applicability of the equations presented in the previous section, bit-error rate performance in the presence of phase noise has been measured using a 1,200-bps BPSK modem built as part of an experimental maritime terminal. The results of these measurements will be compared with calculated performance from equations (15), (17), and (21).

Measurement Technique

Figure 2 is a block diagram of the test setup used to measure both the phase noise spectral density and the bit-error rate. A Frederick Electronics model 600A test set used as a pseudo-random data source drives the PSK modulator through a level converter and differential encoder. After translation to 70 MHz by the transmit IF subsystem, the signal is connected directly to the receive IF subsystem.

The receive IF subsystem normally provides automatic frequency control (AFC) by means of a voltage-controlled crystal oscillator (VCXO) and loop feedback. To simulate phase noise conditions the AFC loop is

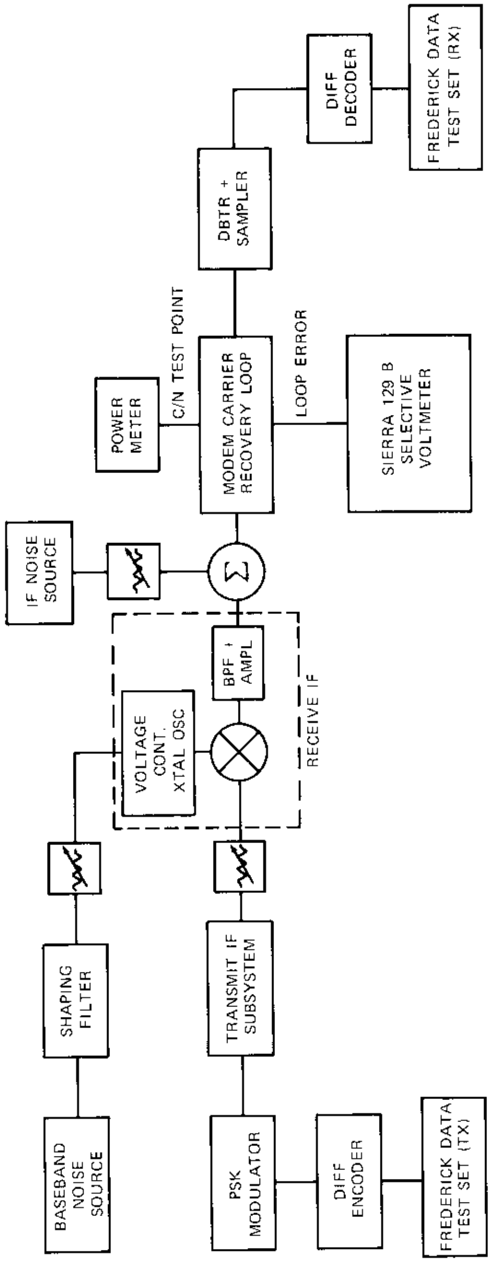


Figure 2. Phase Noise Bit-Error Rate Measurement Setup

opened and the control voltage to the VCXO is driven by a filtered baseband noise source to approximate the desired spectrum. Manual frequency control is also inserted into the circuit at this point.

Following the IF subsystem, thermal noise is added to the resultant 30-MHz signal and then connected to the PSK demodulator input. The nominal data output of the demodulator is differentially decoded and sent to a second Frederick Electronics model 600A test set. In addition, a second signal from the loop phase error detector is brought out to a frequency selective voltmeter. It is this output which is used to measure the phase noise power spectrum.

For small phase error (i.e., when $\sin\theta \cong \theta$, where θ is the phase difference at the loop), the power spectral density of the loop error voltage is given by

$$K_d^2 S_\theta^2(\omega) = K_d^2 S_\phi(\omega) \left|1 - H(\omega)\right|^2 \tag{22}$$

where

K_d = gain of the detector
$H(\omega)$ = loop transfer function
ϕ = absolute modulation angle.

From the form of equation (22) it is clear that, relative to the phase error, the phase-locked loop acts as a high-pass filter. Therefore, a narrowband tracking loop is required to obtain an accurate phase power spectral density in the low-frequency range. For these measurements, a loop filter whose parameters result in a tracking loop bandwidth of 16.6 Hz has been chosen. However, for normal operation and measurement of bit-error rate, the loop filter parameters have been chosen to provide a nominal bandwidth of 416 Hz and hence better acquisition time and tracking range.

It has been noted that equation (22) holds only when $\theta \cong \sin\theta$. However, for cases of interest the phase jitter is large enough to invalidate this approximation. The error voltage power spectrum for larger phase errors will be analyzed in the following subsection.

Analysis for large phase jitter

When the total phase error is large, the approximation $\theta \cong \sin\theta$ is invalid. To estimate the effect of this nonlinearity some assumptions have been made. The first is that θ is Gaussian. This is a reasonable assumption since ϕ is Gaussian (i.e., ϕ has been generated by modulating a linear VCXO with Gaussian noise) and θ is a high-pass version of ϕ. Therefore, for a narrow loop, $\theta \approx \phi$. The second assumption follows from the observed behavior of the measured power spectrum. In Figure 3 it can be seen that, as the noise power is increased from the linear region (case B) to a region where compression is occurring (cases C, D, and E), the shape of the power spectrum does not change significantly. Hence, it is assumed that, for any value of frequency ω, the power density of $S_{\sin\theta}(\omega)$ is proportional to $S_\theta(\omega)$ in the same ratio as the ratio of the total power in $\sin\theta$ to that in θ. The latter ratio can be calculated from the Gaussian assumption. Therefore,

$$\frac{S_\theta(\omega)}{S_{\sin\theta}(\omega)} \cong \frac{\sigma_\theta^2}{(\overline{\sin^2\theta})} \quad . \tag{23}$$

From the Gaussian assumption,

$$\overline{\sin^2\theta} = \int_{-\infty}^{\infty} \sin^2\theta \exp\left\{\frac{-\theta^2}{2\sigma_\theta^2}\right\} \frac{d\theta}{\sqrt{2\pi}\,\sigma_\theta} = \frac{1}{2}\left(1 - e^{-2\sigma_\theta^2}\right) \tag{24}$$

where σ_θ^2 is the variance of the total oscillator phase jitter given by equation (13). Rearranging yields

$$\frac{\sigma_\theta^2}{\overline{\sin^2\theta}} = -\frac{1}{2}\,\frac{\ln\left[1 - 2\,\overline{\sin^2\theta}\right]}{\overline{\sin^2\theta}} \quad . \tag{25}$$

To obtain $\overline{\sin^2\theta}$ from measured data it is necessary only to integrate (numerically or by using suitable straight line approximations) the measured power spectrum of $\sin\theta$ from the frequency selective voltmeter readings, as shown in Figure 3, where curves C, D, and E are in the saturation region. It should be noted that the measured curves in this figure have an uncertainty of at least ±0.5 dB because of the low-frequency random noise remaining in the selective voltmeter indication. Curves modified according to equations (24) and (25) are shown in Figure 4.

Comparison of computed and measured bit-error rate

To evaluate the bit-error rate predictions given by equation (21), the bit-error rate of the 1,200-bps modem was measured as a function of the C/N ratio for various levels of oscillator phase noise. (The oscillator phase noise was adjusted by varying the noise power level at the VCXO

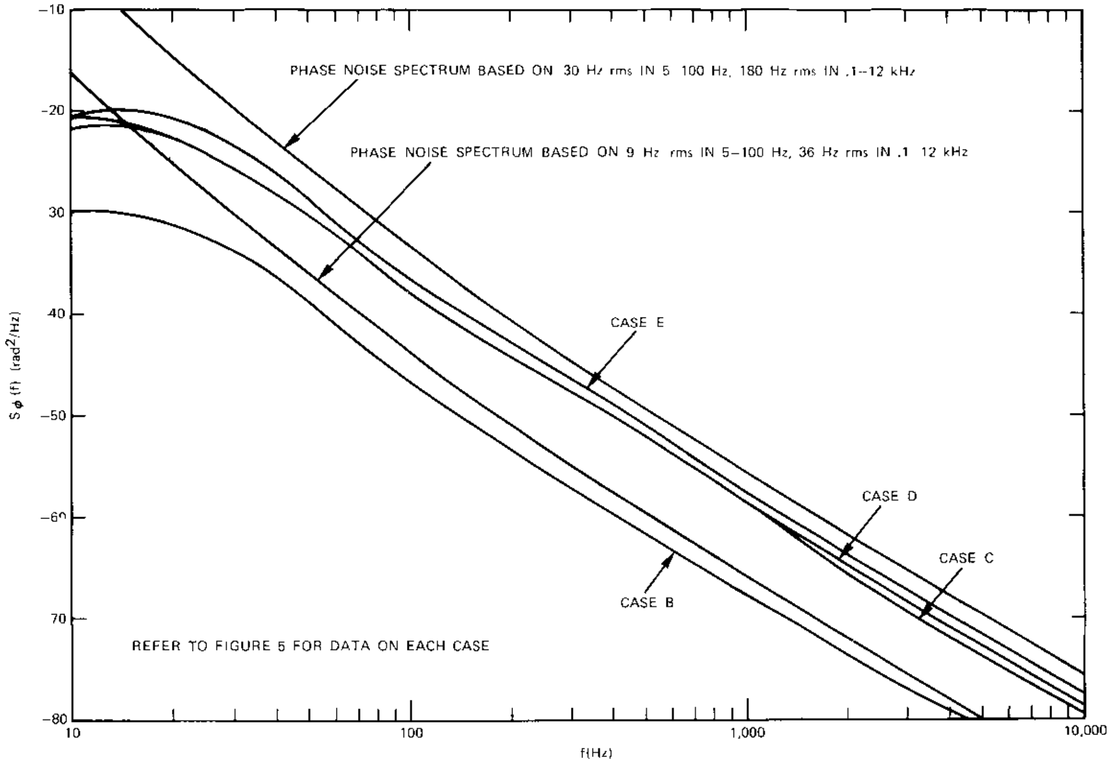


Figure 3. *Measured Phase Noise Spectrum*

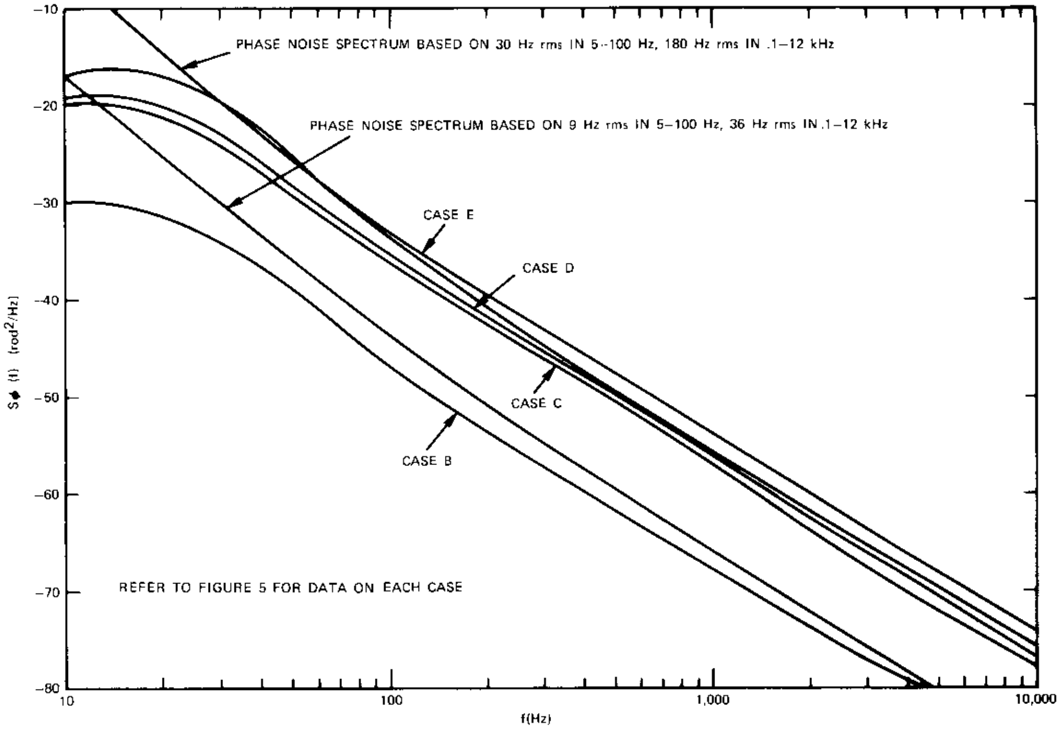


Figure 4. *Phase Noise Spectrum "Corrected" for Compression Effects*

input.) For each oscillator noise level the phase power spectrum was measured as described above; the results are shown in Figure 3. The correction factor given by equation (25) was used to obtain the constants required for equation (15). Finally, equation (21) was used to generate the "calculated" bit-error rate. The results of these calculations are shown in Figure 5, with the measured bit-error rate indicated on the same scale. It can be seen that the calculated bit-error rate is very close to the measurement, especially in view of the uncertainty in the phase noise measurements.

Application to a typical satellite system

The equations given in the section entitled "Effects of phase noise" can be applied to the determination of an allowable phase jitter specification for a satellite transponder oscillator. To determine the overall performance of a satellite link, it is necessary to include the effects of up- and down-converter phase noise as well as satellite and high-power amplifier (HPA) noise. In the following subsections, the effects of phase noise on a 1,200-bps BPSK time-division-multiplexed (TDM) demodulator will be considered.

Maximum allowable phase jitter

The following assumptions have been made concerning the phase noise characteristics:

a. all contributions to the phase noise have the same spectral shape, i.e., the same break point between $1/f$ noise and flat noise;

b. the contributions from all oscillators add in an rss (root square sum) manner;

c. the satellite oscillator and the transmit HPA are the principal variable contributions to the phase noise;

d. the threshold bit-error rate for the 1,200-bps TDM demodulator is 10^{-5} at a C/N$_o$ ratio of 43.4 dB-Hz.

At threshold, if half the errors (5×10^{-6}) are allocated to the cycle skipping errors, equation (17) may be used to estimate the maximum allowable phase jitter. For example,

$$5 \times 10^{-6} = \frac{\pi}{4} \exp \left\{ \frac{-\pi}{8\sigma^2_{\phi\max}} \right\} \tag{26}$$

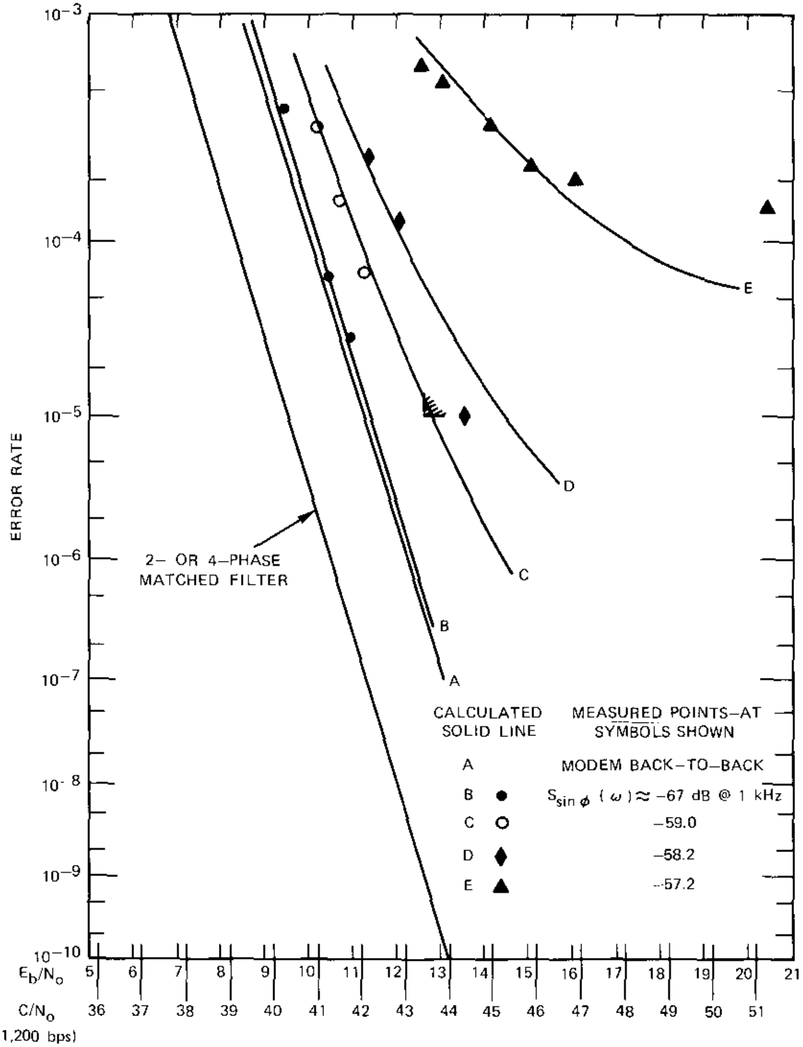


Figure 5. *Comparison of Calculated and Measured Bit-Error Rate with Various Phase Noise Levels*

or

$$\sigma^2_{\phi\ \max} = 0.0328 \text{ rad}^2$$

$$\sigma_{\phi\ \max} = 0.18 \text{ rad} = 10.4^\circ \quad .$$

rms frequency deviation

Because there are three parameters, C_1, C_2, and B_L, it is necessary to apply the other assumptions in the previous subsection and to relate the phase noise constants to a new satellite rms frequency deviation specification. On the basis of the first two assumptions, equation (5), and the method described in the preceding section, it can be shown by solving equation (12) for two sets of values for ω_1 and ω_2 that

$$C_1 \propto C_2 \propto f_T^2 \tag{27}$$

where f_T^2 is the total rms frequency deviation in a 5- to 100-Hz bandwidth. Therefore, equation (15) can be written in the form

$$\sigma^2_{\phi T} = \frac{K_1 f_T^2}{B_L^2} + \frac{K_2 f_T^2}{B_L} + \frac{B_L}{\nu M(\mathrm{C/N_o})} \tag{28}$$

where $K_1 = 2.1011$ and $K_2 = 0.01084$. (These values of K_1 and K_2 correspond to the ratio of $1/f$ to flat noise shown in Figures 3 and 4.)

On the basis of equation (28), a loop bandwidth assumed to be between one-fifth and one-tenth the bit rate, and a 1.56-dB loss for additional doubler and filter effects, a trial and error procedure yields values for f_T and B_L which meet the conditions of equation (23) (i.e., $f_T = 9$ Hz, $B_L = 200$ Hz, and $\sigma^2_{\phi T} = 0.021736$ rad^2). Note that $\sigma^2_{\phi T}$ has been chosen so that it is less than $\sigma^2_{\phi\ \max}$ to allow a 2-dB cycle skipping margin when the carrier recovery loop is stressed, i.e., when it is operating at the extreme frequency offset. These parameters have been selected as the design point for the demodulator.

Effects of varying the rms frequency deviation

This subsection will investigate the effect of constraining B_L to 200 Hz and allowing only the satellite oscillator (and HPA) noise to vary. To relate the total noise to the satellite noise, the following assumptions are necessary:

a. the rms deviation of the up-converter is fixed at 2 Hz in a 5- to 100-Hz bandwidth;
b. the rms deviation of the down-converter is 2.5 Hz in 5 to 100 Hz;
c. 2-dB degradation is allowed for AFC operation.

These contributions are combined with the satellite deviation in a root square sum which is then increased by 2 dB to provide a margin for AFC operation. The total rms deviation is then

$$f_T^2 = \{(2.0)^2 + (2.5)^2 + f^2_{sat}\}\ 10^{0.2} \quad . \tag{29}$$

From equation (28),

$$\sigma^2_{\phi T} = \frac{2.1011 f_T^2}{(200)^2} + \frac{(0.01084)\ f_T^2}{(200)} + \frac{200}{0.698(\mathrm{C/N_o})} \quad . \tag{30}$$

Modifying equation (17) to include 2-dB degradation for loop stress yields

$$P_{ecs} = \frac{\pi}{4} \exp \left\{\frac{-\pi 10^{-0.2}}{8\sigma^2_{\theta T}}\right\} = \frac{\pi}{4} \exp \left\{\frac{-0.2478}{\sigma^2_{\phi T}}\right\} \quad . \tag{31}$$

Equations (29)–(31) have been used to calculate P_{ecs} vs f_{sat} over the range of f_{sat} from 5 to 9 Hz. The total P_e over the same parameter range is determined by adding the thermal noise probability of error to the cycle skipping probability. To calculate the thermal P_e, the theoretical matched filter performance has been modified to provide a 2-dB implementation margin. Thus, equation (21) becomes

$$P_e = P_{eth} + P_{ecs} \cong \frac{\pi}{4} \exp \left\{\frac{-\pi 10^{-0.2}}{8\sigma^2_{\phi T}}\right\} + \frac{\exp\left\{\dfrac{-\ 10^{-0.2}C}{N_o R}\right\}}{2\sqrt{\dfrac{10^{-0.2}C\pi}{N_o R}}} \quad . \tag{32}$$

The resulting total error rate is shown in Figure 6 with f_{sat} as a parameter. These curves can be used to estimate the required phase jitter when various implementation margins are assumed.

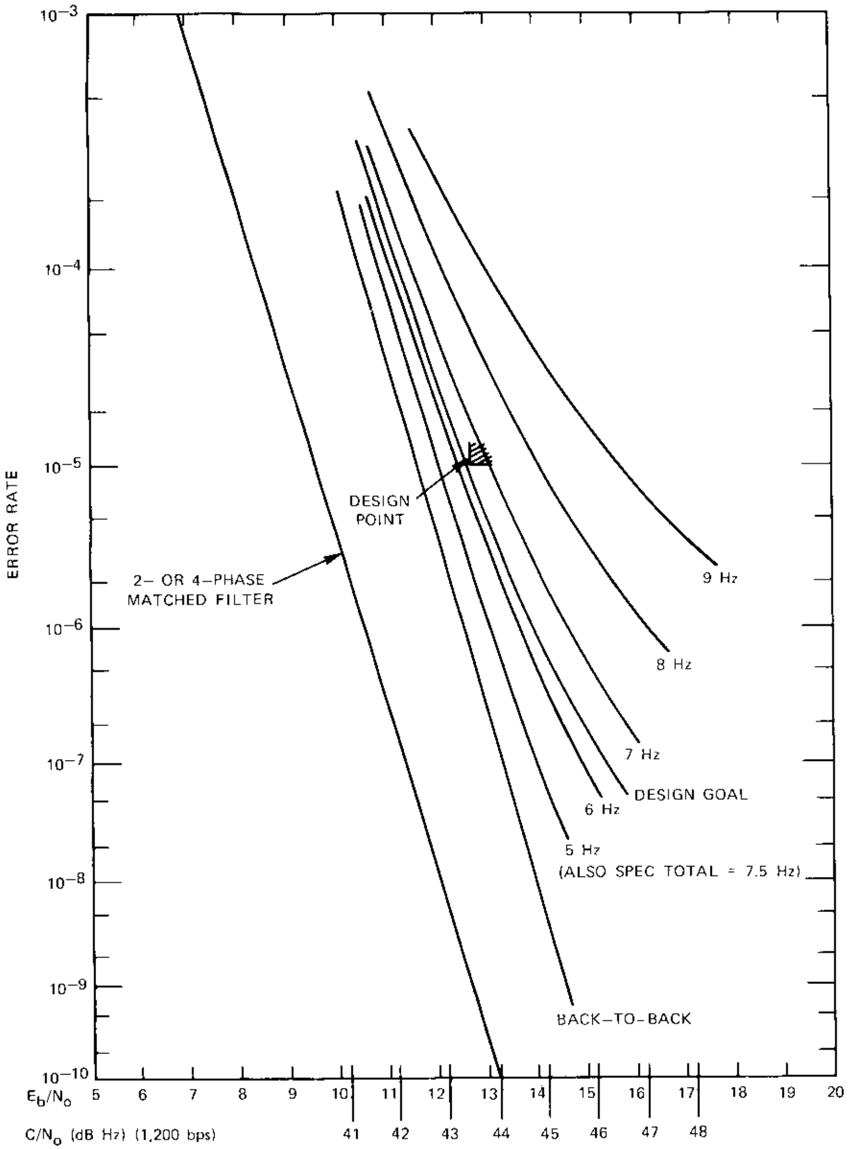


Figure 6. *Probability of Error vs C/N_o Ratio with Spurious FM Noise*

Conclusions

It has been shown that a relatively simple expression for bit-error rate in the presence of phase and thermal noise can be derived for phase noise with combined $1/f$ and flat FM spectral characteristics. The resulting expression has been shown to agree with measured results.

Acknowledgment

The author wishes to thank Dr. L. Clayton of Scientific Atlanta *and Mr. D. Swearingen of* COMSAT GENERAL Corporation *for their contributions and help. He also wishes to thank Drs. W. Sones and T. McGarty for their careful review of the paper.*

References

[1] T. L. Grant and R. L. Cramer, "Effect of Oscillator Instability on Telemetry Signals," NASA Technical Note NASA TN D-6424, July 1971.

[2] R. Snydar et al., "Frequency Stability Requirements for Space Communications and Tracking," *Proc. IEEE*, Vol. 54, No. 2, pp. 231–236.

[3] K. Park, "Phase Error and Loop Noise Bandwidth Limitation in View of COMSAT Specification," Private communication, Magnavox Corporation, August 23, 1973.

[4] K. Park, "Effects of Phase Jitter on Squaring Loop," Private communication, Magnavox Corporation, September 20, 1973.

[5] F. Gardner, *Phase Lock Techniques*, New York: John Wiley, 1966.

[6] L. C. Palmer and S. A. Klein, "Phase Slipping Phase Locked Loop Configurations That Track BiPhase or Quadri-Phase Modulated Carriers," *IEEE Transactions on Communications Technology*, COM-20, October 1972, p. 984.

Chester J. Wolejsza, Jr., received a B.E.E.E. degree from Cooper Union in 1966 and M.S. and E.E. degrees from M.I.T. in 1967. Immediately thereafter, he joined COMSAT *Laboratories, where he is currently Manager of the Modulation Techniques Department of the Communications Processing Laboratory. His responsibilities include supervision of the design and development of PSK modulation systems which perform optimally in the band- and power-limited environment of a satellite communications channel.*

Mr. Wolejsza is a member of Tau Beta Pi and Eta Kappa Nu and an associate member of Sigma Xi.

Index: SPEC (speech predictive encoding communications) system, digital speech interpolation, telephone channels, time-assigned speech interpolation.

Digital speech interpolation

S. J. CAMPANELLA

(Manuscript received August 29, 1975)

Abstract

This paper discusses two different methods of digital speech interpolation (DSI). One uses a digitized version of time-assigned speech interpolation (TASI) and the other a technique known as SPEC (speech predictive encoded communications). The digitized TASI system employs the low speech activity on incoming telephone channels to interpolate speech spurts onto a smaller number of transmission channels. A technique known as bit reduction makes it possible to generate additional channels for TASI transmission during overload conditions so that the probability of disruptive speech clips can be significantly reduced. The SPEC method incorporates PCM sample prediction to achieve a substantial reduction in the number of samples needed for transmission and inherently avoids the initial clips of speech spurts which are encountered in TASI.

The operation of each method is analyzed to evaluate its performance as a function of the number of channels processed. TASI is considered in terms of degradation due to initial clips of speech spurts whose duration is sufficient to damage the intelligibility of initial consonants, while SPEC is evaluated in terms

This paper is based upon work performed in COMSAT Laboratories under the sponsorship of the International Telecommunications Satellite Organization (INTELSAT). Views expressed in this paper are not necessarily those of INTELSAT.

A condensed version of this paper was presented at the INTELSAT/IECE/ITE 3rd International Conference on Digital Satellite Communications in Kyoto, Japan, November 11–13, 1975. The SPEC system is a COMSAT corporate development.

of degradation due to added prediction distortion. In each case the ratio of the number of incoming channels to the number of transmission channels (the interpolation advantage) as a function of the total number of transmission channels is calculated for amounts of degradation that are considered subjectively insignificant. Both methods are competitive in terms of interpolation advantage, although SPEC achieves slightly better performance, especially when the number of channels is small.

Introduction

Since it is customary during 2-way conversation for one talker to pause while the other speaks, active speech signals are present on a transmission channel for only a fraction of the total conversation time. In addition, even when only one talker is speaking, pauses occur between utterances and there are times when the circuit is idle. On the average, speech is present for less than 50 percent of the time; in fact, actual measurements show that speech is present on a typical telephone channel approximately 40 percent of the time [1]. This paper discusses digital time-assigned speech interpolation (TASI) [2]–[5] and speech predictive encoded communications (SPEC) [6]–[8] techniques which, by exploiting these activity properties, lead to a reduction of the total information rate needed to handle a multiplicity of telephone channels.

Time-assigned speech interpolation

General considerations

TASI [9]–[11] is a technique in which the idle time between calls and the conversation pauses during calls are used to accommodate additional calls. With a sufficiently large number of channels, most of the idle time on the transmission link can be filled, giving an enhancement in transmission capacity greater than two.

Observations of transatlantic cable circuits have indicated that speech is actively present on a busy channel about 40 percent of the time. Each period of time occupied by a caller's speech is called a speech spurt. The low activity is largely due to the facts that a subscriber speaks less than half the time during a conversation and that each subscriber's speech is carried on a separate channel. If all connections supplied to a TASI terminal are busy, the speech spurt activity will average 40 percent. If, on the other hand, all circuits are not busy, the average speech activity experienced by the TASI terminal will be further decreased. The percentage of busy circuits

is called the incoming channel activity and the percentage of time that speech spurts occupy a channel is the speech spurt activity, or simply speech activity. Thus, if the incoming channel activity is 85 percent, the speech spurt activity will be reduced to 34 percent.

Quality aspects

COMPETITIVE CLIPPING AND FREEZE-OUT

TASI exploits the low speech spurt activity by assigning transmission channels only when a speech spurt is present. It is evident that the process becomes more efficient as the number of channels increases. An attempt to interpolate two independent conversations on a single transmission channel will cause a large percentage of the speech to be lost due to competition for simultaneous channel occupancy. Until it is terminated, the conversation occupying the transmission channel will "freeze out" any attempt by the other conversations to occupy the channel.

When a large number of independent conversations compete for some smaller number of transmission channels, the same type of competition prevails and there is always a finite probability that the number of conversations demanding service will exceed the number of available transmission channels. This competition manifests itself in the form of clipping of the initial portion of a speech spurt, or competitive clipping. The percentage of time that speech is lost due to such competition is called the percent freeze-out or freeze-out fraction. Provided that the population of incoming channels processed is sufficiently large and the ratio of the number of incoming channels to the number of channels available for transmission (the interpolation advantage, or in a digital implementation, the DSI advantage) is sufficiently small, the fraction of speech lost to freeze-out can be rendered acceptably small. TASI systems have been designed so that the freeze-out fraction is 0.5 percent.

The freeze-outs consist principally of short clips of the initial portions of speech spurts ranging from zero to a few hundred milliseconds. Speech clips longer than 50 ms cause perceptible mutilation of initial plosive, stop, fricative, and nasal consonants [12]. It is important to minimize the frequency of occurrence of speech clips longer than 50 ms if high quality is to be maintained. In this paper the performance of TASI is analyzed in terms of the criterion that the percentage of clips longer than 50 ms is less than 2 percent. This criterion is considered to be preferable to the constant freeze-out fraction criterion traditionally used by designers of TASI systems [6] because it relates more directly to the cause of degradation, viz, clipping of initial consonants.

CONNECT CLIPPING

At each TASI terminal the presence of speech on a telephone channel is sensed by a speech detector which initiates a request for a transmission channel. A channel assignment processor assigns an idle transmission channel to the incoming channel in response to the request and also sends a connect signal to the TASI terminal at the far end specifying the outgoing channel to which the transmission channel is to be connected. During the time required to make the channel assignment and to connect the listener and talker, the speech can be clipped. This phenomenon is called a connect clip and is to be distinguished from the competitive clip discussed previously.

The connect clip is constrained to a very short duration to minimize subjective degradation consistent with reliable connect signaling. The statistics of this type of clip are controlled by the competition of connect messages for space on the assignment channel. During periods of light loading on the incoming channels, transmission channels may remain assigned even during idle intervals so that the impact of connect clipping is absent. During periods of heavy loading on incoming channels, the connect clip may occur on every speech spurt. Thus, the degradation caused by connect clipping will vary with trunk activity.

SPEECH DETECTOR CLIPPING

In addition to competitive and connect clipping, some time is required to accomplish speech detection and a detector clip can occur. Since it is possible to virtually eliminate detector clipping by appropriate voice detector design, detector clipping can for all practical purposes be ignored.

Digital TASI implementation

A functional implementation of a digital TASI system is shown in Figure 1. The incoming telephone channels are digitized into conventional 8-bit-per-sample, 8,000-sample-per-second, pulse code modulated (PCM) time-division-multiplexed (TDM) format as part of the existing telephone plant or specifically to interface with the DSI system. The digitized signals are then processed by the transmit assignment processor, which is sequentially shared among all the incoming telephone channels.

The transmission channels consist of time slots arrayed consecutively in a TDM frame. The PCM samples selected for transmission are assigned to slots which become available after the assignment processor demands a transmission channel to service a particular incoming channel. The control channel needed to carry the channel assignment information to the far end is included in the same TDM frame.

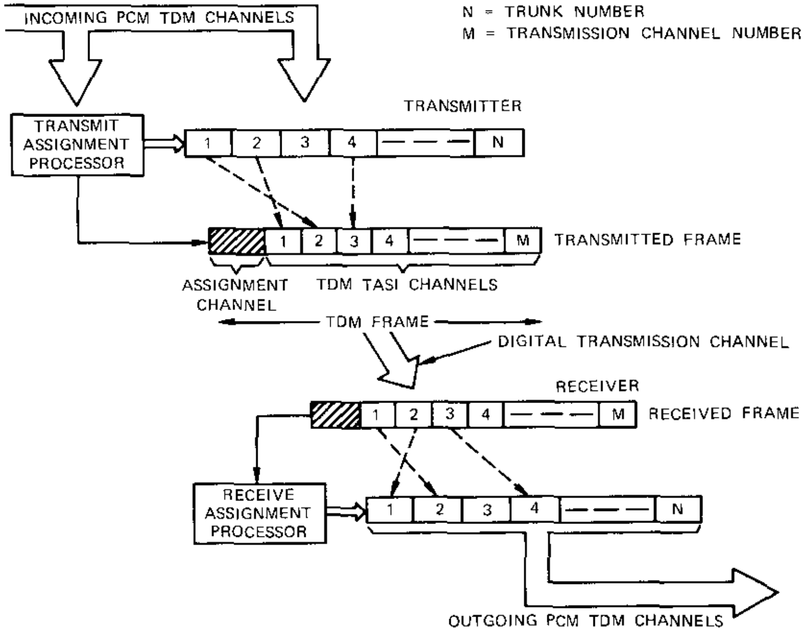

INCOMING PCM TDM CHANNELS
N = TRUNK NUMBER
M = TRANSMISSION CHANNEL NUMBER
TRANSMITTER
TRANSMIT ASSIGNMENT PROCESSOR
1
2
3
4
N
1
2
3
4
M
TRANSMITTED FRAME
ASSIGNMENT CHANNEL
TDM TASI CHANNELS
TDM FRAME
DIGITAL TRANSMISSION CHANNEL
RECEIVER
1
2
3
4
M
RECEIVED FRAME
RECEIVE ASSIGNMENT PROCESSOR
1
2
3
4
N
OUTGOING PCM TDM CHANNELS

Figure 1. *TASI-Type DSI System*

Digital TASI has a number of advantages over analog TASI. For example, digital voice detectors perform better than their analog counterparts.* In addition, more precise and efficient switching of digital speech samples among channel slots of the TDM time frame is inherent to digital techniques. An all-digital approach is also compatible with the use of a digital channel assignment processor for recording the channel assignments at any instant and communicating this information to a companion digital channel assignment processor at the receiver.

*In a DSI system, the digital voice detector is part of a central assignment processor that is time shared among all the telephone channels. Therefore, a more sophisticated design can be incorporated. An equivalent analog design may be much more costly since it must essentially be repeated in every detail for each incoming telephone channel.

An additional advantage of the digital implementation of TASI is the ability to expand the number of transmission channels to avoid freeze-out during instants of overload by reappropriating the least significant bits of the digital transmission time slots. This technique, known as channel augmentation by bit reduction, can be invoked during overload conditions to avoid excessive competitive speech clipping. Although reducing the number of quantizing levels per slot from 256 (8 bits) to 128 (7 bits) produces a 6-dB increase in quantization noise, the fraction of time that bit reduction is required is very low. Hence, its presence is not apparent.

Analysis of digital TASI performance

STATISTICS

As mentioned previously, a principal factor governing TASI performance from the subjective point of view is the probability of occurrence of voice spurt clips of 50 ms or more. A 2-percent probability of occurrence of clips with durations equal to or greater than 50 ms is used as a threshold of acceptability in the following analysis. Hence, on the basis of an average voice spurt duration of 1.5 seconds and an activity (average percentage of time during which the speech is present on a channel) of 40 percent, a clip of this type will occur once every 1.5 minutes in a telephone conversation to one of the subscribers.

At any particular instant, the probability that the number of simultaneous talkers on N incoming channels with activity α will equal or exceed c (where c is the number of transmission channels) is given by the binomial distribution

$$B_{c,N,\alpha} = \sum_{x=c}^{N} \frac{N!}{x!\,(n-x)\,!}\,\alpha^{x}(1-p)^{N-x} \tag{1}$$

If it is further assumed that the voice spurts have durations that are exponentially distributed with mean L, then the probability that a spurt is frozen out (i.e., clipped) for longer than time t is given by $B_{c,n,\theta}$ [9], where

$$\theta = \alpha\epsilon^{-t/L} \tag{2}$$

From these expressions, the probability of occurrence of clips with $t > 50$ ms and $L = 1.5$ seconds as a function of the number of transmission channels, c, for $N = 15, 30, 60, 120, 180$, and 240 incoming channels has been calculated. The results are shown in Figure 2, which also indicates the 2-percent boundary and the number of transmission channels needed to accommodate a given number of incoming channels, N. The ratio of N/c, or the DSI advantage, is plotted in Figure 3 as a function of N. It

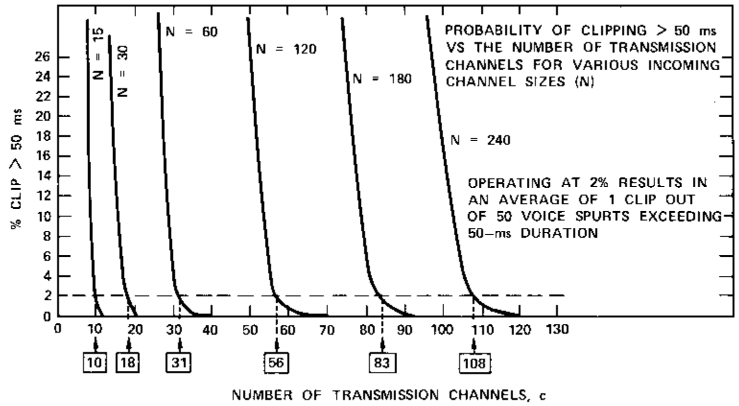
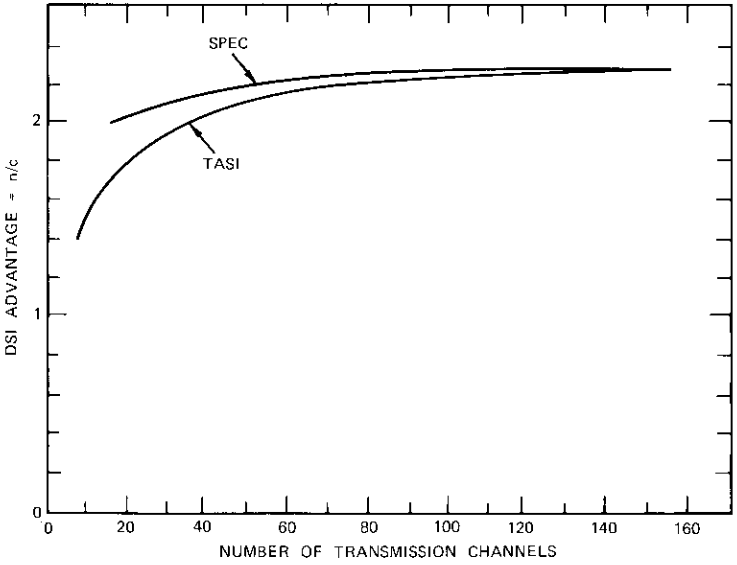


Figure 2. *TASI Performance in Terms of Competitive Clipping*



can be seen that the TASI advantage, as defined here, reaches a value of 2.2 for 240 incoming channels and exceeds 2 for all cases in which the number of available transmission channels exceeds 38. For comparison, the advantage of SPEC for the same number of transmission channels is also shown in Figure 3. The curve for SPEC is not based on clipping occurrence, since SPEC totally avoids the clipping problem. Instead, the SPEC performance is based on the probability that degradation due to distortion produced by the predictor exceeds 0.5 dB for 25 percent of the time, a point which will be discussed in detail later.

The distribution of competitive clip durations is a function of the number of incoming channels, as shown in Figure 4, which indicates the distributions of competitive clip duration for $N = 60$ and 240. Both distributions have about the same probability of competitive clip duration greater than 50 ms, but longer clips are more likely to occur for $N = 60$ than for $N = 240$.

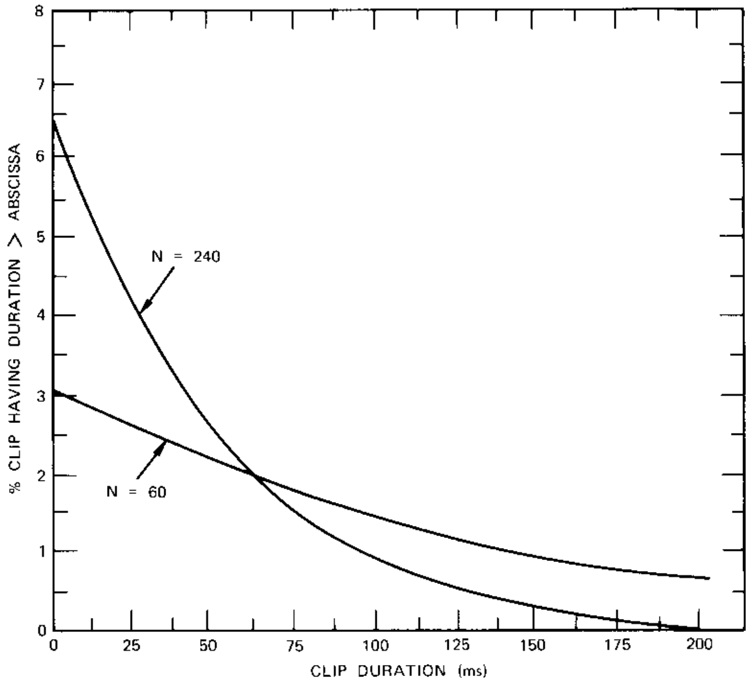


Figure 4. *Distribution of Competitive Clip Duration for $N = 60$ and 240 Channels*

INFLUENCE OF BIT REDUCTION

As mentioned previously, bit reduction can be used to absorb overloads that would otherwise result in clipping. Bit reduction is required when the number of telephone channels demanding service exceeds the number of normal transmission channels. For a TASI system, if 240 channels are interpolated for the DSI gain shown in Figure 3, bit reduction will be required 5 percent of the time.

Additional 7-bit PCM words are generated without increasing the number of bits allocated to PCM samples in the DSI subburst by reappropriating the least significant bit of all 8-bit PCM words and hence reducing their resolution to 7 bits. Thus, for a DSI subburst composed of c 8-bit PCM words, bit reduction will augment the number of transmission PCM words to yield

$$c_{BR} = c + \left\lfloor \frac{c}{7} \right\rfloor \qquad (3)$$

where $\lfloor\ \rfloor$ indicates the next lower integer. When bit reduction is used, the probability of occurrence of clips longer than 50 ms drops significantly. This is shown in Table 1, where the probability that the clip duration exceeds 50 ms for 40-percent speech spurt activity is compared for systems with and without bit reduction.

TABLE 1. PROBABILITY OF CLIP DURATION >50 ms WITH AND WITHOUT BIT REDUCTION ($\alpha = 40\%$)

Number of Incoming Channels, N	Probability of Clip Duration >50 ms (%)	
	With Bit Reduction	Without Bit Reduction
60	0.13	2
120	0.07	2
180	0.01	2
240	0.004	2

Table 1 indicates that bit reduction can effectively reduce the clipping problem encountered in TASI. Although the introduction of bit reduction causes a 6-dB increase in quantization noise on all channels affected, as long as the system is not heavily overloaded, the subjective impact of this increased quantization noise is negligible.

ASSIGNMENT MESSAGE FREEZE-OUT

In a version of DSI being considered for TDMA use, assignment messages are transmitted once every 750 μs as part of the DSI subburst. Each assignment message consists of 24 bits of assignment information which is rate one-half coded for transmission error protection and can accommodate one connect or disconnect operation. Use of the bit reduction method requires that both a connect and a disconnect message be sent to accommodate each voice spurt.

The probability that the assignment messages will be frozen out can be analyzed in terms of the binomial probability expression $B_{c,N,q}$ introduced in equation (1). Since there is one assignment transmission channel, c is equal to one. This channel must serve the randomly occurring assignment messages arriving on the N parallel incoming channels. The activity, q, is the ratio τ/l, where τ is the length of the TDMA frame and l the average time interval between assignment messages. If both connect and disconnect messages are needed, then

$$l = \frac{L}{2\alpha} \quad . \tag{4}$$

If only connect messages are needed, then l is twice this value.

For the TDMA equipment, in which both connect and disconnect messages are needed, l = 1.875 seconds and $q = 4 \times 10^{-4}$ for L = 1.5 seconds and α = 40 percent. In this case, the freeze-out probability of an assignment message for N = 240 channels, where c = 1 and $q = 4 \times 10^{-4}$, is $B_{c,N,q} = 10^{-1}$. This is in effect the probability of having to wait more than one TDMA frame period, i.e., 750 μs. This situation will occur once every five voice spurts or once every 18.75 seconds in each direction of conversation. If statistical independence is assumed, the probability of having to wait s TDMA frames for assignment is simply $(B_{c,N,q})^s$. For s = 4, the probability of having to wait longer than 3 ms is 10^{-4} and the time between occurrences is 5.2 hours. Hence, it is obvious that speech clips caused by assignment message channel congestion are of negligible consequence for the case of interest here.

Speech predictive encoded communications

General considerations

SPEC is a form of digital speech interpolation which differs significantly from digital TASI. One of its principal merits is total avoidance of the competitive clip problem experienced by TASI. Its operation protocol does not require recordkeeping for connections from incoming channel to transmission channel to outgoing channel since all channel assignment information is contained in each frame. Hence, channel assignment memory and assignment message channel implementation is unnecessary. Its adaptive processing method exhibits only a slight increase in quantizing noise when confronted with overload. This event occurs so infrequently at the channel augmentation ratios shown in Figure 3 that it is of negligible subjective consequence.

Quality aspects

PREDICTOR DISTORTION

SPEC requires the signal to pass a speech detector before samples are admitted to the predictor. The signals passed by the speech detector exhibit an activity similar to that experienced with TASI. The predictor algorithm reduces this activity by eliminating unnecessary samples in the instantaneous speech waveform or in the short intersyllabic pauses not sensed by the voice detector. Under average load conditions the SPEC predictor removes more than 25 percent of the PCM samples during voice spurts 25 percent of the time. This results in a signal-to-distortion (S/D) ratio decrease of only 0.5 dB. The achievable DSI advantage ratios are shown in Figure 3 as functions of the number of incoming channels.

COMPETITIVE CLIPPING

SPEC cannot cause clips such as those encountered in TASI since samples would have to be frozen out for a succession of 80 SPEC frames to produce a 10-ms clip. The probability of occurrence of such an event is very small.

SPEECH DETECTOR CLIPPING

SPEC incorporates a speech detector and, as in the case of digital TASI, the inclusion of adaptive threshold features in the design results in negligible clipping.

CONNECT CLIPPING

Since there is one bit in the sample assignment word (SAW) for each incoming channel, the SPEC SAW allows fully flexible connectivity for all incoming channels. Hence, there cannot be any connect clipping caused by waiting for channel assignment.

Implementation

In the SPEC system, shown functionally in Figure 5, the speech information contained on N incoming 8-bit-per-sample PCM telephone channels is transmitted in the space of $(N/8) + c$ 8-bit-per-sample transmission channels, where c is the number of PCM sample channels used for transmission and is typically equal to slightly more than $N/3$. SPEC processes all incoming trunks and transmits a frame once per period of the PCM Nyquist rate, which corresponds to a 125-μs sampling interval in typical commercial telephone usage.

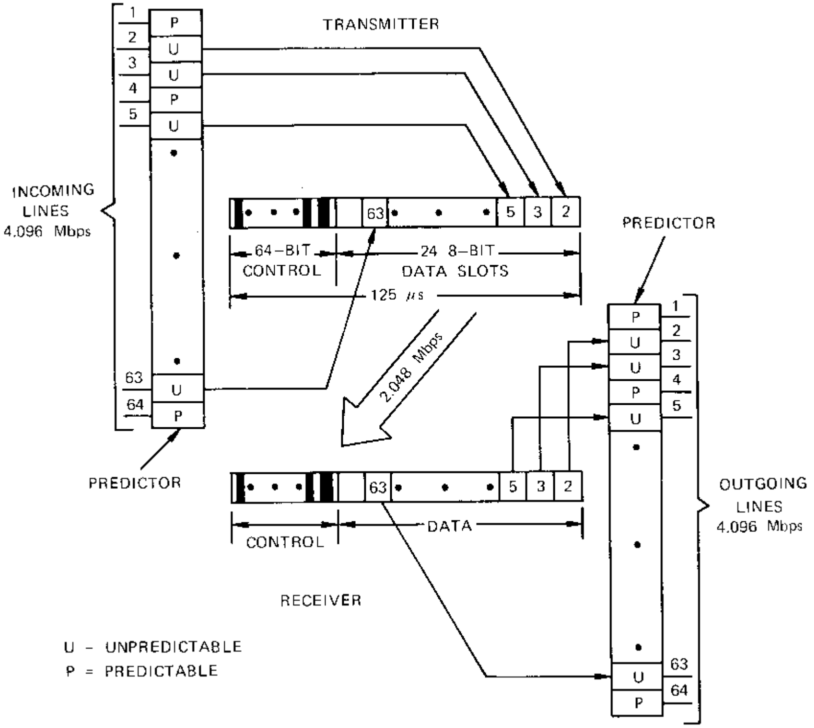


Figure 5. *SPEC-Type DSI System*

SPEC operation is conveniently described by using an actual example of its implementation* for accommodating 64 incoming telephone channels (corresponding to a bit rate of 4.096 Mbps) in the transmission space normally allotted to 32 PCM channels (corresponding to a bit rate of 2.048 Mbps). The PCM samples derived during each sample period from the 64 incoming channels are compared with the samples previously sent to the receiver and stored in a memory at the transmitter. Any that differ by an amount equal to or less than some given number of quantizing steps, called the aperture, are discarded and not sent to the receiver. These are referred to as "predictable" samples. The remaining "unpredictable" samples are transmitted to the receiver and replace the values formerly stored in memories at both the transmitter and receiver. The aperture is automatically adjusted as a function of activity observed over the 64 incoming channels on each frame so that the number of samples transmitted is nearly constant.

The SPEC transmission frame is composed of an initial SAW followed by a number of 8-bit time slots that carry the individual PCM samples judged unpredictable by the transmitter's prediction algorithm. The SAW contains one bit for each of the incoming telephone channels. Thus, for the 64-terrestrial-channel system, it contains 64 bits. The bit corresponding to a given channel is a "1" if the frame contains a sample for that particular channel and a "0" if it does not. Thus, the SAW contains all of the information needed to distribute the samples among the 64 outgoing channels at the receive end.

At the receiver the unpredictable samples received in the SPEC frame replace previously stored samples in the receiver's 64-channel memory as directed by the SAW. The samples in the memory, in the form of a conventional PCM/TDM frame, are strobed into the outgoing channels at the proper rate. The most recent frame thus contains new samples on the channels that have been updated by the most recent SPEC frame and repetitions of the samples that have not been updated.

Analysis of SPEC performance

ACTIVITY REDUCTION BY PREDICTION

The PCM samples transmitted by SPEC are determined by selecting that set in which the differences between the samples previously sent and those currently being sent are equal to or greater than some number of quantizing steps called the aperture, a. The value of the aperture is adjusted so that the number of samples transmitted does not exceed the number of available transmission channel slots.

If the incoming channel activity for a given frame is such that all new samples can be transmitted with zero aperture, i.e., if the number of samples to be transmitted with differences greater than zero is equal to or

*A system such as that described has been built and tested in demonstration service between the U.S. mainland and Hawaii on the INTELSAT IV satellite [8].

less than the number of transmission slots, no error is made in the transmission. If the incoming channel activity for a given frame is such that the aperture $a = 1$, samples differing from the previously transmitted samples by a number of quantizing steps greater than 1 are transmitted. Hence, the number of samples requiring transmission is reduced by culling out those with differences less than or equal to 1. In this case errors equal to one quantizing step will be made for those samples with differences of zero. The activity resulting from this action is

$$\alpha_1 = \alpha[1 - P(\Delta = 0) - P(\Delta = 1)] \qquad (5)$$

where α = activity experienced on the ensemble of incoming terrestrial channels with no prediction
α_1 = activity after culling out samples with differences less than or equal to 1
$P(\Delta = k)$ = probability of samples differing from those previously sampled by k quantizing steps.

In this case, errors of the magnitude of one quantizing step will be made for a fraction of the samples equal to $P(\Delta = 0)$.

In the general case, if a set of SPEC processed values is transmitted with an aperture of value a, then the resulting activity for that set is

$$\alpha_a = \alpha\left[1 - \sum_{k=0}^{a} P(\Delta = k)\right] = \alpha(1 - P_a) \quad . \qquad (6)$$

The amount given by the summation and designated as P_a is known as sample predictability for aperture a. On the average, the value of incoming channel activity, α, is approximately 40 percent.

SPEC is designed to operate at an average transmit channel activity (given by the ratio of the number of transmission channels available for samples, c, to the number of incoming channels, N) of $c/N = 0.3$. The aperture must be such that

$$\alpha_a \leqq \frac{c}{N} \qquad (7)$$

and the predictability must be such that

$$P_a \geqq 1 - \frac{c}{N\alpha} \quad . \qquad (8)$$

These rules establish the average aperture at which SPEC operates.

DISTRIBUTION OF PREDICTION

For a given incoming channel activity, α, the number of incoming channels, n, requiring transmission due to the presence of voice spurts is a random variable. The probability that the number of incoming channels requiring service at a given instant will exceed the number of transmission channels is given by the binomial distribution

$$P(n > c) = B_{c,n,\alpha} = \sum_{x=c}^{n} \frac{n!}{x!\,(n - x)\,!}\,\alpha^x(1 - \alpha)^{n-x} \quad . \qquad (9)$$

The system is constrained so that the number of samples needing transmission never exceeds c. This is accomplished by using prediction to reduce the activity in the entire set of incoming channels to a new activity, α_a, described previously. The average number of channels requiring service under these circumstances is then

$$n_a = \alpha_a N \quad . \qquad (10)$$

Substituting the new value of activity into the binomial distribution yields

$$P(n > c) = B_{c,n,\alpha_a} \qquad (11)$$

where $\alpha_a = \alpha(1 - P_a)$, which expresses the modification in activity resulting from the introduction of the predictor. Since the average number of transmission channels used with prediction is

$$c = N\alpha(1 - P_a) \qquad (12)$$

the binomial distribution can be redesignated as

$$P[n > N\alpha(1 - P_a)] = B_{c,n,\alpha_a} \qquad (13)$$

which can be further modified as follows:

$$\begin{aligned} P\left[\alpha(1 - P_a) < \frac{n}{N}\right] &= P\left[\alpha < \frac{n}{N(1 - P_a)}\right] \\ &= P\left[P_a > 1 - \frac{n}{N\alpha}\right] \\ &= P\left[P_a > \left(1 - \frac{\alpha_a}{\alpha}\right)\right] \\ &= B_{c,n,\alpha_a} \quad . \qquad (14) \end{aligned}$$

Equation (14), which gives the probability that the prediction will exceed the value $(1 - \alpha_a)/\alpha$, is very useful in assessing the level of degradation introduced by the predictor.

Figure 6 shows the performance of SPEC in terms of the predictability, $P_a = 0.25$, calculated from equation (14) for $\alpha = 0.40$ and $N = 30$, 60, 120, 180, and 240 incoming channels plotted against the number of transmission channels, c. The ordinate is the probability that the indicated value of P_a is exceeded for the number of transmission channels on the abscissa, which includes those channels needed to accommodate the SAW. Each transmission channel is assumed to be 8 bits wide. A predictability of $P_a = 0.25$ means that, on the average, one sample out of four is predicted during a voice spurt. In the case of $N = 120$, it can be seen that, for $c = 54$ transmission channels, the probability that the predictability will exceed $P_a = 0.25$ is 25 percent. This can also be interpreted to mean that for 75 percent of the time the predictability will be such that less than one in four samples will be predicted during a voice spurt when 120 incoming channels are carried in the space of 54 transmission channels. In this case, the SPEC DSI advantage is $120/54 = 2.22$.

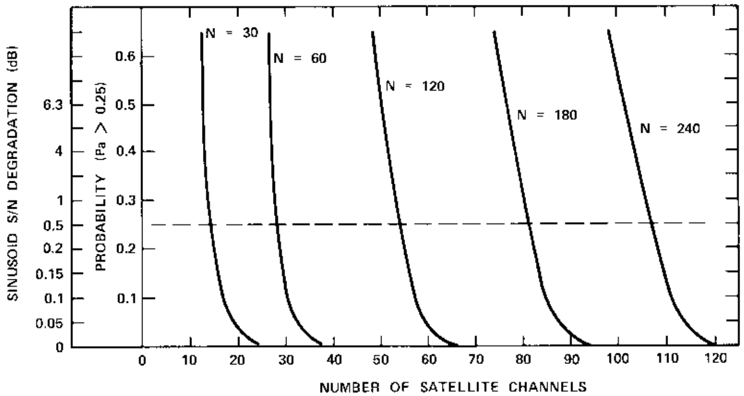


Figure 6. *SPEC Predictor Performance*

SIGNAL-TO-DISTORTION RATIO

It is possible to relate the probability that $P_a > 0.25$ to degradation in the S/D ratio over SPEC channels. This is done by using experimental data regarding the S/D ratio experienced by a 0-dBm0 sinusoidal signal passed through one channel of an experimental SPEC terminal built to accommodate 64 incoming channels in 32 PCM transmission channels plus 64 SAW bits [8]. The experimental data, given in Figure 7, show the degradation of the sinusoidal S/D ratio as the average activity on the 64 incoming channels varies over a wide range of values. The overall degradation is due to a combination of PCM quantization noise and prediction distortion. As the incoming channel activity increases, the predictor must handle a greater number of samples by increasing its aperture, thus causing the total distortion to increase. Figure 7 also shows the signal-to-Gaussian-noise ratio (S/N) for a 0-dBm0 sinusoidal signal on a Gaussian-noise-perturbed PCM channel with speech quality equivalent to that experienced with prediction distortion. This curve indicates that a given level of prediction distortion corresponds to a lower level of Gaussian noise.

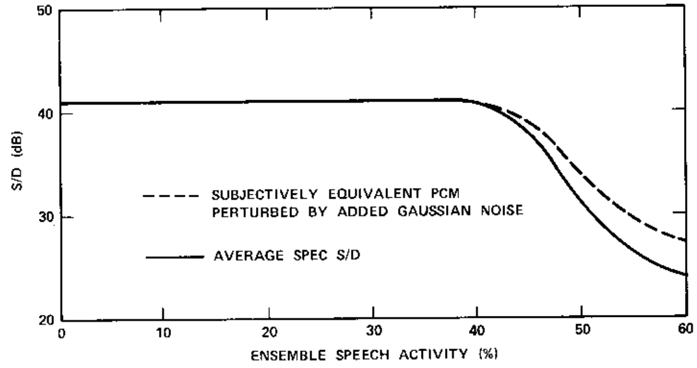


Figure 7. *Measured and Subjectively Equivalent S/D as Functions of Ensemble Speech Activity*

The data presented in Figure 7 can be directly related to predictability by using equation (14). The results are found in Figure 8, which shows the dB degradation in sinusoidal S/D ratio incurred by prediction as a function of the probability that $P_a > 0.25$.

Also shown in Figure 8 is a curve giving the equivalent performance of the PCM channel perturbed by additive Gaussian noise. This curve indicates that, for conditions such that $P(P_a > 0.25) = 25$ percent, the dB degradation in the sinusoidal S/D is only 0.75 dB. The dB degradation in the sinusoidal S/N on a subjectively equivalent channel perturbed by Gaussian noise is only 0.5 dB. Thus, it appears that $P(P_a > 0.25) = 25$ percent is an acceptable threshold for determining the advantage offered by the SPEC system. The ordinate of Figure 6 also contains a scale giving the dB deg-

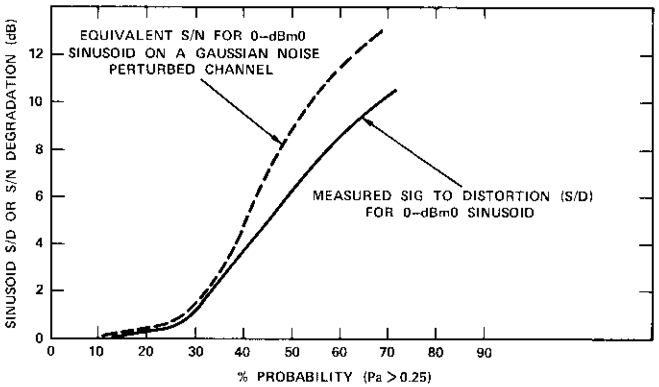


Figure 8. *Sinusoid S/D or Equivalent S/N versus $P(P_a > 0.25)$*

radation in the subjectively equivalent S/N obtained by using the data provided in Figure 8.

The number of transmission channels needed to accommodate each number of incoming channels is determined from the intersection of each curve shown in Figure 6 with the 0.5-dB degradation boundary. The plot of SPEC advantage versus the number of transmission channels (including SAW bits) shown in Figure 3 has been obtained from this result.

Digital speech interpolation in multidestinational satellite use

Digital TASI

In multidestinational satellite DSI, voice spurts originating on incoming terrestrial lines must be connected to appropriate outgoing terrestrial lines at the desired destinations. For digital TASI, these connections are accomplished in two steps, as shown by the connectivity maps of Figure 9. First, voice spurts occurring on the incoming terrestrial lines must be connected to satellite channel slots in the DSI burst. Then the DSI burst channels must be connected to the appropriate outgoing terrestrial lines at the designated receive terminal.

As shown in Figure 9, there are two maps at transmit terminal A. Map I indicates the connections between the incoming terrestrial lines on which voice spurts are occurring and the individual satellite transmission channel numbers that appear in the DSI burst. The connections at all receive terminals with which A communicates are indicated in map II, which shows that terminal A has distributed its satellite channels among destination terminals B, C, and D. Three satellite channels are assigned to destination B, two to destination C, and three to destination D. The assignments made at terminal A are communicated to the various destinations by the assignment message portion of the DSI burst.

Consider next the outgoing connectivity map at a typical receive terminal such as the 3-section map at terminal B. One section corresponds to channels in the DSI burst from terminal A, and the other two correspond to channels in bursts received from terminals C and D. In that portion of the map corresponding to reception from A, it can be seen that satellite channel slots 0, 2, and 3 are connected to outgoing terrestrial lines 1, 2, and 4. This is of course a duplicate of the map which has been set up at terminal A. The other sections of terminal B's map duplicate the assignments at terminals C and D and direct the connections at B to receive the voice spurts from terminals C and D, respectively.

To make the appropriate destination line assignments in multidestinational operation, a transmit terminal must be familiar with the connectivity maps at each destination with which it communicates. Thus, at terminal A there are replicas of the connectivity maps which designate the assignments of satellite channels to outgoing terrestrial lines for each destination with which terminal A is to communicate. If it is established in advance that terminal A has exclusive use of a preassigned set of outgoing lines at each destination, then the maps which are part of map II can be established at terminal A without knowledge of assignments on outgoing lines at the destination terminal made by other terminals communicating with the same destination. If, however, the outgoing lines at each of the destinations with which A communicates are to be shared among all other terminals communicating with the same destinations, then all terminals must possess the complete connectivity maps for all destinations and these must be updated to show the most recent connectivities. This will result in a fully variable demand-assignment DSI communications network. In either case, regardless of destination, the interpolation process occurs over the entire pool of incoming lines since satellite channels are fully variable in terms of both source and destination assignment.

SPEC

The method of channel assignment used in the SPEC system is different from that used in a digital TASI system. The SPEC frame, which is transmitted

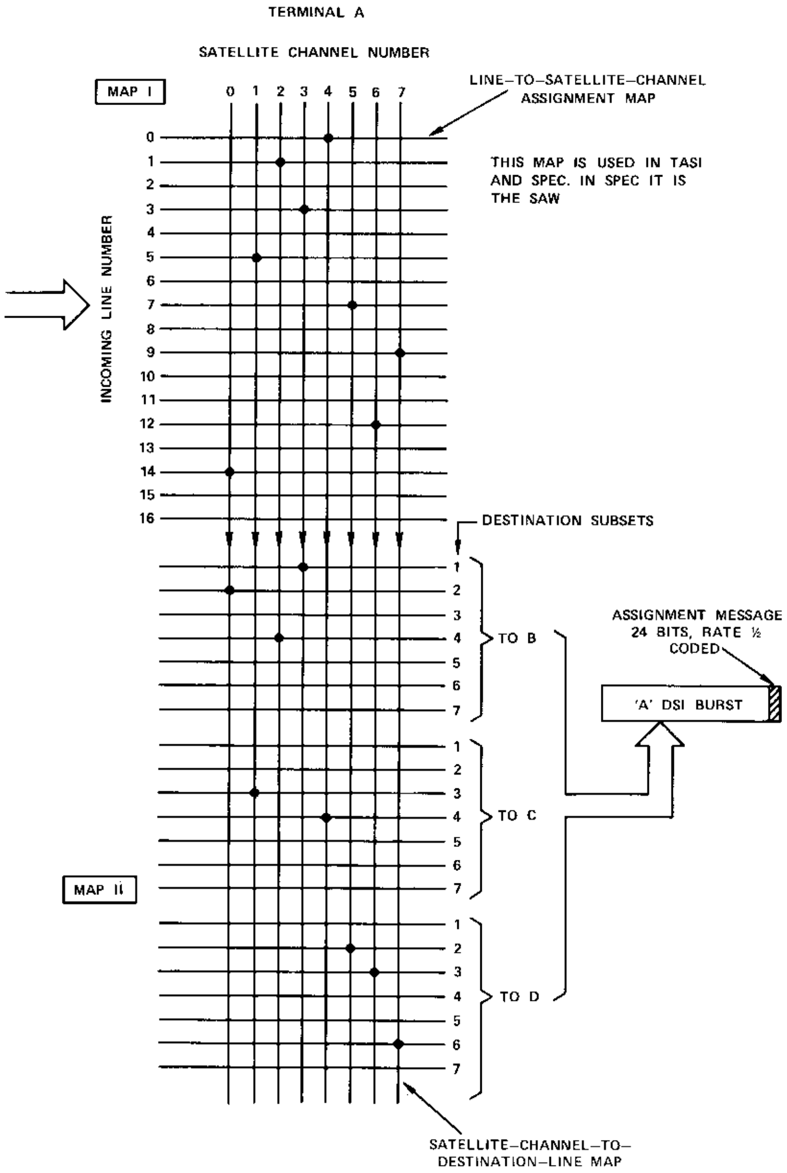

TERMINAL A
SATELLITE CHANNEL NUMBER
MAP I
0 1 2 3 4 5 6 7
LINE-TO-SATELLITE-CHANNEL ASSIGNMENT MAP
THIS MAP IS USED IN TASI AND SPEC. IN SPEC IT IS THE SAW
INCOMING LINE NUMBER
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
DESTINATION SUBSETS
1 2 3 4 5 6 7
TO B
TO C
TO D
ASSIGNMENT MESSAGE 24 BITS, RATE ½ CODED
'A' DSI BURST
MAP II
SATELLITE-CHANNEL-TO-DESTINATION-LINE MAP

Figure 9a. *TDMA/DSI Connectivity Maps from Terminal A to Terminals B, C, and D*

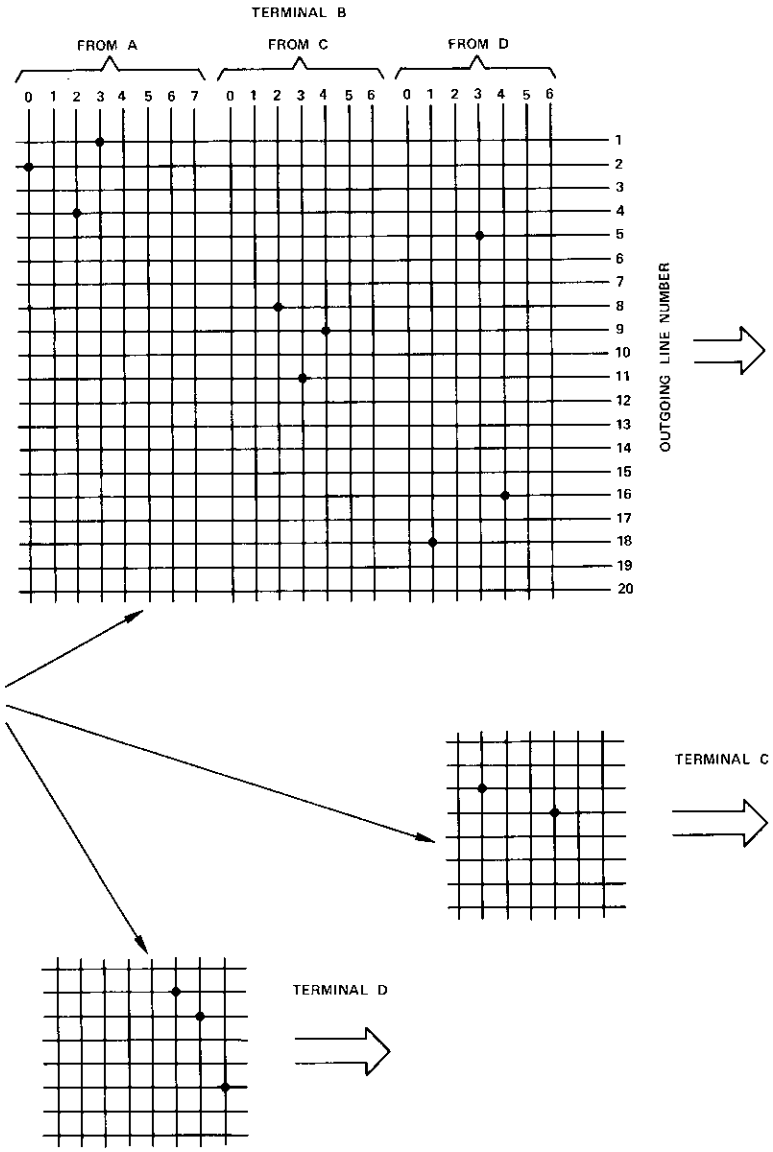


Figure 9b. *TDMA/DSI Connectivity Maps from Terminal A to Terminals B, C, and D*

once every 125 μs, consists of a SAW followed by unpredictable PCM samples of the speech channels which are to be transmitted to a set of destinations. The number of bits in the SAW corresponds to the number of destination lines to which the burst is directed. If a channel is to receive information, a "1" is transmitted in the SAW in the position corresponding to that channel; otherwise a "0" is transmitted. The samples that are transmitted in the PCM sample part of the SPEC frame correspond one-for-one to the "1's" in the SAW part of the frame. Thus, if the first terrestrial channel to which a sample is to be transmitted is number 3, then a "1" appearing in the third bit of the SAW directs the first sample of the sample portion of the SPEC frame to terrestrial channel number 3. If the second channel to which a sample is to be transmitted is channel number 7, then the next "1" appearing in the seventh bit of the SAW directs the second sample in the PCM part of the frame to terrestrial channel number 7. The number of samples transmitted in the PCM part of the frame will correspond to the number of "1's" present in the SAW part of the frame. The SAW may thus be considered to be a sample activity mask.

The SAW is a replica of map I in Figure 9. If the SPEC frame is to be communicated to one destination, then all of the channel numbers appearing in the SAW part of the frame correspond to channel numbers at that destination. If SPEC is to be operated in a multidestinational mode, then the channel numbers appearing in the SAW part of the frame are grouped according to destination. Thus, channel numbers 0 through 30 may go to destination A, 31 to 60 to destination B, 61 to 90 destination C, and so on. This constitutes a preassignment multidestinational mode of operation.

At the receiver, each burst received from a given source is stored in two parts, the SAW part and the PCM sample part. Only samples designated for a given destination are transferred from the SPEC frame into the appropriate outgoing lines as dictated by the SAW part of the frame. SPEC transmits a new connection map every frame, thus permitting highly flexible reassignment of the transmitted PCM samples among the outgoing lines at the destinations as a function of the instantaneous activity on all incoming lines at each transmit terminal. Hence, the SPEC system interpolates over the entire pool of channels going to all destinations.

It is also possible to reassign the traffic capacities among various destinations in the SPEC system by redesignating the transmission channels as traffic patterns change. This process, which can be accomplished by a low-data-rate channel transmitted as part of the SPEC frame, makes it possible to adapt to variations in traffic among destinations.

DSI implementation for TDMA

Methods for implementing the TASI and SPEC DSI systems for TDMA operation will now be described. Although actual equipment designs may differ in specific details, the structures presented are basic and should illustrate the functions needed to realize operational equipment.

TASI/TDMA implementation

TRANSMIT TERMINAL

The transmit terminal of a digital TASI system for possible use in TDMA equipment is shown in Figure 10. The terminal is configured to accommodate 240 terrestrial channels. Analog terrestrial input lines in groups of 30 are supplied to CEPT-32 PCM analog-to-digital (A/D) converters. Each of these converters delivers a PCM/TDM digital stream at a bit rate of 2.048 Mbps. These outputs are then multiplexed by a 60-channel multiplexer into a single stream at a bit rate of 4.096 Mbps. Modular grouping into 60 channels is used as the basic building block of the system. The system can handle additional channels in groups of 60 by adding more building blocks of this type in parallel.

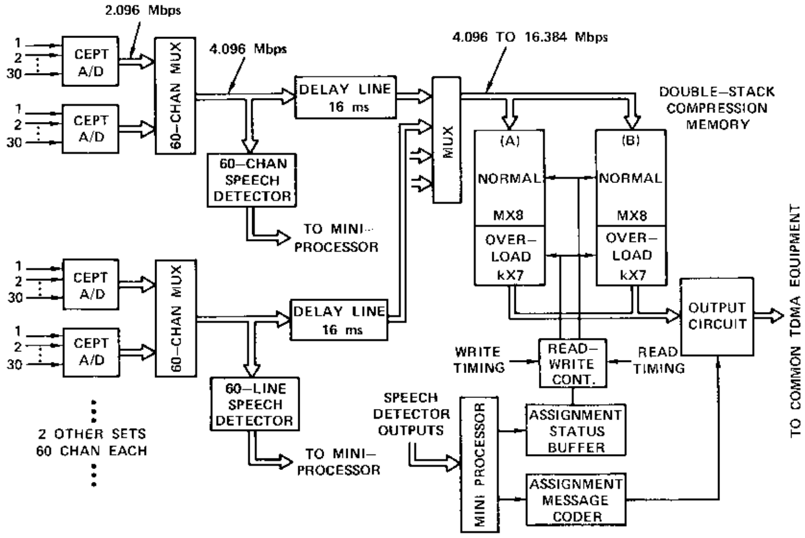


Figure 10. *TASI/TDMA Transmit Terminal*

The digital output of each 60-channel multiplexer is next fed to a digital delay line having a delay of 16 ms to offset any delay in the voice spurt detection circuits. The same digital stream is also fed to a voice detector, which simultaneously processes 60 channels in digital form. The voice detector decisions are supplied as input to the miniprocessor section of the DSI equipment. The miniprocessor then assigns active voice spurts to appropriate satellite channels as they become available in accordance with a satellite channel map stored in its memory. It also assigns the relationship, in the form of a destination map, between the satellite channel and an outgoing terrestrial channel at the desired destination and formulates assignment messages to inform the destination of the connection for each new voice spurt or the disconnection for each terminated voice spurt. [Disconnects are needed in overload (bit reduction) operation only.] Finally, the miniprocessor updates the transmit assignment status buffer so that selected TDM time slots coming from the multiplexer are assigned to their appropriate channel slot positions in the TDMA compression buffer.

The TDMA compression buffer is a double-stack design in which one stack is filled while the other is emptied to prevent conflict due to read/write function overlapping. It stores as many PCM samples as there are time slots available in the DSI burst. Each DSI burst contains six Nyquist frame sections, and each section can hold up to 108 eight-bit samples, which is the number required for 240 input terrestrial channels. In addition, each section has an overload memory to accomplish bit reduction. The bit reduction strategy essentially increases the number of time slots available in the DSI burst by a factor of 8/7 rounded to the next lower integer. Thus, if the system is operating with 108 eight-bit time slots, the capacity is increased to 123 seven-bit time slots when bit reduction is used. The miniprocessor controls the onset of the bit reduction mode and establishes the necessary signaling to inform the receiver that bit reduction is being used.

RECEIVE TERMINAL

The receive side of the DSI subsystem, shown in Figure 11, must be capable of accepting bursts from multiple sources and appropriately distributing the information contained in these bursts to the terrestrial channels. Assignment messages from all sources are stored in the miniprocessor in the form of maps which associate channel slots in the various received DSI subbursts with outgoing terrestrial channels. The expansion buffer is a dual-stack structure. Each stack is capable of storing the contents of all DSI frames destined to the receive terminal and stores as many PCM samples as there are outgoing terrestrial channels. Under control of the assignment status buffer at the receive terminal, the samples stored in the expansion buffer are read out to the appropriate terrestrial channels.

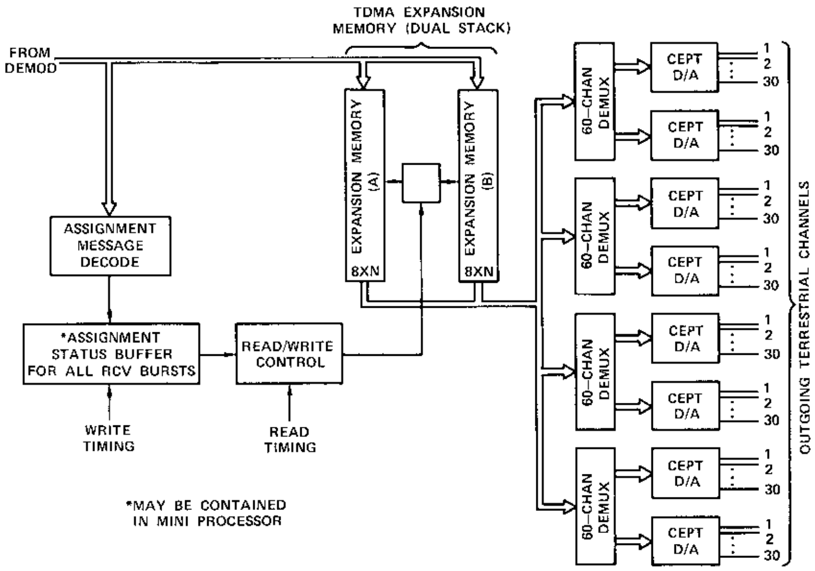

TDMA EXPANSION MEMORY (DUAL STACK)
FROM DEMOD
EXPANSION MEMORY (A)
8XN
EXPANSION MEMORY (B)
8XN
ASSIGNMENT MESSAGE DECODE
*ASSIGNMENT STATUS BUFFER FOR ALL RCV BURSTS
READ/WRITE CONTROL
WRITE TIMING
READ TIMING
*MAY BE CONTAINED IN MINI PROCESSOR
60-CHAN DEMUX
CEPT D/A
1
2
30
OUTGOING TERRESTRIAL CHANNELS

Figure 11. *TASI/TDMA Receive Terminal*

SPEC/TDMA implementation

TRANSMIT TERMINAL

Figure 12 is a block diagram of a SPEC transmit terminal configured for 240 incoming terrestrial channels handled in modules of 60 channels each. Each group of 60 channels is processed through a pair of CEPT-32 A/D converters. The outputs of all CEPT units are multiplexed in a single multiplexer to a bit rate of 16.384 Mbps. The bit stream is then supplied to two devices: a 5-ms digital delay line and the 240-channel speech detector. The speech detector detects the presence of voice spurts on each of the 240 incoming terrestrial channels. Whenever a voice spurt is present it permits the PCM samples to pass to the intermediate frame memory (IFM) and the zero-order predictor (ZOP).

The IFM is capatle of storing 240 eight-bit PCM samples. Depending on decisions made by the ZOP, the samples stored in the IFM will be transferred to the predictor frame memory (PFM). Specifically, the ZOP calculates the difference between values stored in the PFM and the most recent set of values which has been supplied to the IFM. For those PCM values whose difference is greater than the aperture, the values in the IFM are transferred to the PFM to replace the old values and are stored in the transmit frame memory (TFM) for transmission to the appropriate set of destinations.

Within the predictor there are a number of 240-bit-long storage units in which SAWs are stored for several aperture values. The SAW used in a given frame which controls the transfer of values from the IFM to PFM and TFM is the one corresponding to an aperture such that the number of samples transmitted is just less than the number of sample slots available in the TFM. For a 240-channel system the number of sample slots available in the TFM is 90. With low activity on the incoming channels, it will be found that all samples transmitted are the result of applying a small aperture value to yield a low quantization noise. If, on the other hand, the activity on the incoming terrestrial channels is high, the aperture needed to reduce the population to a number below 90 will be greater and the quantization noise will be correspondingly greater. Thus, this implementation of the SPEC system permits the aperture occurring from each SPEC burst to vary over a large range to accommodate wide fluctuations in activity.

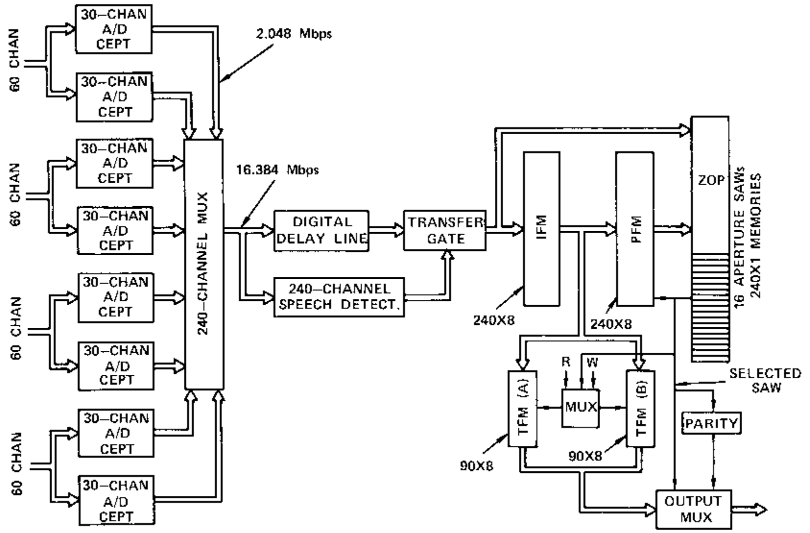


Figure 12. *SPEC Transmit Terminal*

All values which are updated in the PFM each time the ZOP executes its function are called unpredictable values and are also transferred to the TFM where they await transmission to their destinations. The TFM is a dual-stack buffer configuration consisting of two 90 × 8 sections, each able to store an entire TFM. While one half is being filled the other half is being emptied, thus permitting a continuous flow of signal information to the outgoing communications link. An output multiplexer combines the samples of the TFM with the appropriate SAW to constitute the transmission burst. The frame duration upon which the SPEC transmitter operates is 125 μs and is synchronized to the Nyquist sampling frame used in the A/D PCM converters at the input.

For the SPEC system implementation to correspond to the 750-μs frame specified by INTELSAT, the frames generated by the equipment shown in Figure 12 must be accumulated in a TDMA compression buffer. This compression buffer consists of a dual-stack configuration which stores six SPEC frames and outputs them at the burst rate of the TDMA terminal equipment.

RECEIVE TERMINAL

A SPEC receive terminal must select from each received DSI burst only those samples destined to its outgoing terrestrial channels by examining each source's SAW and selecting and storing only the corresponding samples contained in the PCM sample part of each SPEC DSI frame. A possible implementation is shown in Figure 13. Each SPEC frame obtained from the TDMA demodulator is supplied to a demultiplexer that separates the SAW and PCM sample parts. The SAW is supplied to a PARITY CHECK and a SAW SUBSET SELECT unit. If the parity check is successful, that subset of SAW bits [designated as $(\text{SAW})_x$] which identifies the samples destined to the particular terminal of interest is passed on to the SAW memory. The SAW SUBSET SELECTOR then admits to the TFM only the appropriate subset of PCM samples, designated as $(\text{TFM})_x$. For any DSI subburst on which the SAW parity check fails, none of the received samples are stored in the TFM and all samples for that subburst are treated as predictable. The $(\text{SAW})_x$ and corresponding $(\text{TFM})_x$ samples are passed into the PFM, which reconstitutes the conventional PCM/TDM format with all predicted samples appropriately filled in. The conventional PCM/TDM stream is then supplied to the digital-to-analog (D/A) sections of CEPT-21 PCM units and hence to the outgoing terrestrial lines.

Comparison of TASI and SPEC implementations

The TASI and SPEC DSI implementation schemes which have been described herein have certain common features. In particular, each implementation requires the same kinds of A/D and D/A conversion interfaces between terrestrial links and satellite channels. Each also requires speech detectors and digital delay lines to appropriately implement these detectors. There are essentially no differences in the voice detector performance requirements of the two systems.

There are differences in the manner in which the samples are prepared for transmission, however. In digital TASI, incoming terrestrial channels on which speech spurts occur are connected to and disconnected from satellite channels and the satellite channels are connected to and disconnected from destination terrestrial channels on the basis of voice spurt demand. PCM samples on active channels are carried in TDM form. TASI uses a separate assignment message channel to make the appropriate connects and disconnects at various destinations.

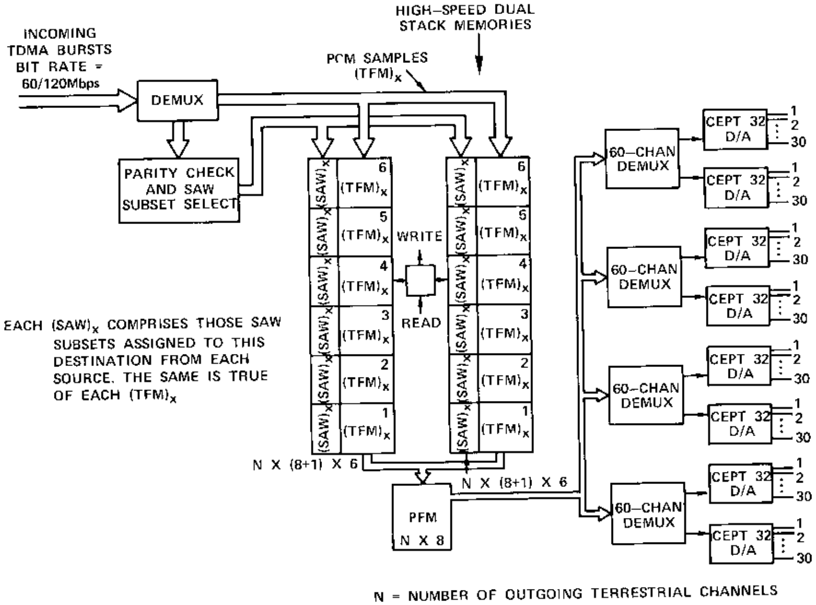


Figure 13. *SPEC Receive Terminal for TDMA Operation*

By comparison, SPEC PCM samples occurring at the Nyquist rate on each terrestrial channel are compared in a ZOP and those determined to be unpredictable are sent to appropriate destinations on a sample-by-sample demand basis. Similarly to digital TASI, the SPEC samples are carried in TDM form. Unlike the TASI system, however, the SPEC system does not use a separate assignment message channel for connect and disconnect information, but instead incorporates a set of bits called a SAW, which is included as part of each SPEC frame, to direct individual samples to their destination. TASI hardware also uses a miniprocessor which retains maps giving all of the connectivities needed to steer speech spurts to satellite channels and satellite channels to all destinations. Storage of these maps is not required in SPEC, since all the destination information needed to distribute the PCM samples for each frame is transmitted in the SAW.

SPEC and TASI also differ with respect to operation during high peaks of activity. In SPEC, because of its operating characteristics, the predictor works harder, removing a greater fraction of the incoming speech samples by increasing its aperture value. Although this action does increase the quantization noise, this increase is not subjectively significant under operational conditions. In TASI, unless special precautions are taken, high peaks of activity will produce perceptible (>50-ms) clips with a frequency that may be unacceptable when operating at a DSI advantage ratio of two. This deficiency of TASI is overcome by adopting the bit reduction strategy, which has been shown to be very effective. However, introduction of the bit reduction strategy complicates voice spurt channel assignment protocol and raises the implementation cost.

The salient features of the SPEC and TDMA techniques are compared in Table 2.

Conclusions

Both the TASI and SPEC methods of digital speech interpolation offer significant enhancements in the capacity of digital transmission for speech communications. Both methods achieve interpolation advantages greater than two, with SPEC attaining slightly higher values than TASI when the number of channels processed is small.

The principal cause of degradation in the TASI method is initial clips of speech spurts. The frequency of occurrence of destructive initial clips can

TABLE 2. COMPARISON OF SPEC AND TDMA TECHNIQUES

	SPEC	TASI
DSI Advantage for 240 Incoming Channels*	2.2	2.2
DSI Advantage for 60 Incoming Channels*	2.1	2.0
Susceptible to Speech Spurt Clipping without Special Precautions	No	Yes
Requires Changes in Channel Assignment Protocol During Overload	No	Yes
Uses Miniprocessor for Channel Assignment Control Connectivity Map	No	Yes
Susceptible to Connect Message Channel Overload	No	Yes
Relative Hardware Complexity	Lower	Higher
Relative Cost of Transmit Terminal	Lower	Higher
Relative Cost of Receive Terminal	Same	Same
Efficient Application in Multidestinational Service	Same	Same

*Based on an assumed voice spurt activity of 40 percent.

be kept low enough to produce little degradation by properly adjusting the interpolation advantage. The threshold of acceptability is based on the condition that the probability of initial clips longer than 50 ms must be no greater than 2 percent. Under this condition, clips longer than 50 ms will occur once every 1.5 minutes. The technique of bit reduction permits additional transmission channels to be made available during occasions of overload, which reduces the probability of occurrence of initial clips. If the least significant bit of each 8-bit PCM sample in the transmit frame is used to generate 7-bit PCM sample slots, the probability of occurrence of initial clips is reduced by more than an order of magnitude.

The principal cause of degradation in SPEC is the production of prediction distortion. The amount of distortion varies as the fraction of samples predicted varies in response to changes in the ensemble average of voice spurt activity. The fraction of samples predicted and hence the amount of prediction noise produced is controlled by the interpolation advantage, as expected. The design objective is to adjust the interpolation advantage so that the probability that more than 25 percent of the samples are predicted during speech spurts is 0.25. It is shown that on the average this results in approximately 0.5-dB degradation in the subjectively assessed speech power-to-quantization-noise power ratio. SPEC is an adaptive system that yields to occasions of higher than average voice spurt activity by automatically increasing the fraction of samples predicted. This is an inherent feature of this technique and at no time does SPEC produce damaging initial clips.

Both TASI and SPEC methods are suited to transmission systems using digital modulation with each carrier capable of operating in preassignment multidestinational service. In terms of implementation complexity SPEC is simpler because it does not require an assignment map memory nor does it need the bit reduction option suggested for use in TASI. Each method can also be adapted to demand-assignment multidestinational service.

Acknowledgment

The author would like to acknowledge the valuable contributions made by Messrs. H. Suyderhoud and R. Ridings in providing information needed to quantify the SPEC *system performance.*

References

[1] P. T. Brady, "A Statistical Analysis of On-Off Patterns in 16 Conversations," *Bell System Technical Journal*, Vol. 47, No. 1, January 1968, pp. 73–91.

[2] F. D. Daynonnet, A. Jowsset, and A. Profit, "LeCELTIC: Concentrateur Exploitant les Temps d'Inactivete des Circuits," *L'Onde Electrique*, Vol. 42, No. 426, September 1962.

[3] M. Hashimoto et al., "An Application of the Digital Speech Interpolation Technique to a PCM-TDMA Demand Assignment System," International Conference on Space and Communications, Paris, France, March 1971.

[4] E. Lyghounis, "Il Sistema A.T.I.C.," *Telecommunicazioni*, No. 26, March 1968, pp. 21–30.

[5] I. Poretti, G. Monty, and A. Bagnoli, "Speech Interpolation Systems and Their Applications in TDM/TDMA Systems," International Conference on Digital Satellite Communications, Paris, France, 1972, p. 348.

[6] S. J. Campanella and J. A. Sciulli, "Speech Predictive Encoded Communications," International Conference on Digital Satellite Communications, Paris, France, 1972, p. 342.

[7] J. Sciulli and S. J. Campanella, "A Speech Predictive Encoding Communication System for Multichannel Telephony," *IEEE Transactions on Communications*, COM-21, No. 7, July 1973.

[8] H. Suyderhoud, J. A. Jankowski, and R. P. Ridings, "Results and Analysis of the Speech Predictive Encoding Communications System Field Trial," *COMSAT Technical Review*, Vol. 4, No. 2, Fall 1974, pp. 371–393.

[9] K. Bullington and J. M. Fraser, "Engineering Aspects of TASI," *Bell System Technical Journal*, Vol. 38, No. 2, March 1959.

[10] J. M. Fraser, D. B. Bullock, and H. G. Long, "Overall Characteristics of a TASI System," *Bell System Technical Journal*, Vol. 41, No. 4, July 1962.

[11] H. Midema and M. G. Schachtman, "TASI Quality-Effect of Speech Detectors and Interpolation," *Bell System Technical Journal*, Vol. 41, No. 4, July 1962.

[12] R. Ahmend and R. Fatechand, "Effect of Sample Duration on the Articulation of Sounds in Normal and Clipped Speech," *Journal of the Acoustical Society of America*, Vol. 31, No. 7, July 1959, p. 1022.

S. J. Campanella received a B.S.E.E. from the Catholic University of America, an M.S.E.E. from the University of Maryland, and a Ph.D. in electrical engineering from the Catholic University of America. Before joining COMSAT, *he was Manager of the Electronics Research Center at Melpar, Inc.*

He is presently Director of the Communications Processing Laboratory of COMSAT *Laboratories where he is responsible for the technical orientation of the efforts pursued in the Signal Processing and Multiplexing, Image Processing, Digital Control, and Modulation Techniques Departments.*

Dr. Campanella is a member of AIAA, AAAS, IEEE, Sigma Xi, Phi Eta Sigma, and the Acoustical Society of America. He is also a member of the board of directors of EASCON.

Index: SPEC system (speech predictive encoding communications system), computer simulation, speech recognition, voice-activated switch.

A new digital voice-activated switch

J. A. JANKOWSKI, JR.

(Manuscript received November 4, 1975)

Abstract

A digital voice-activated switch using real-time computer simulations has been developed for application in the speech predictive encoding communications (SPEC) system. This switch can also be applied to other voice-activated systems such as SPADE and SCPC. It employs a level-detection technique with adaptive thresholds to detect speech in the presence of noise on telephone circuits.

This paper discusses the design criteria and objective tests involved in determining the operating characteristics of the voice detector. The objective testing confirms the advantages of the new design: rapid fine grain threshold adjustment to varying noise, improved noise immunity at all noise levels, and enhanced switching characteristics on speech bursts. A limited critical-listener subjective evaluation has yielded excellent results.

Introduction

Certain voice communications systems such as SCPC (*s*ingle-*c*hannel-*p*er *c*arrier), SPADE (*s*ingle-channel-per-carrier, *p*ulse code modulation, multiple-*a*ccess, *d*emand-assignment *e*quipment), and DSI (*d*igital *s*peech *i*nterpolation) are likely to become increasingly important in telecommunications. The performance of these systems is heavily dependent upon the ability of a voice detector to recognize speech in the presence of noise. A failure to detect speech signals may result in excessively long clipping of speech utterances, causing user dissatisfaction. It is equally important that noise alone be prevented from activating the communications channel to minimize system loading. This is particularly true of speech interpolation systems, which utilize silence intervals along with intersyllabic speech pauses to reduce the transmission bandwidth or bit rate requirements of multichannel systems.

Digital voice detectors which operate in a similar manner on PCM-encoded telephone signals are used in these systems. In general, the voice switches use a detection scheme which distinguishes the difference between noise and speech levels. A threshold level is set and all signals exceeding it activate the switch. In the SCPC [1] and SPADE [2] systems the voice detector employs fixed threshold levels which have been chosen according to the typical noise levels. In a design previously devised for the speech predictive encoding communications (SPEC) DSI [3], [4] system, the threshold is allowed to vary according to the actual noise level on the circuit.

The voice switch discussed in this paper has been developed as an improvement of the previous SPEC voice switch. It includes significant design modifications which greatly enhance its performance.

The problem of detecting voice signals in the presence of noise has been attacked by employing various techniques which exploit the differences between the characteristics of noise and speech. Telephone noise can generally be assumed to be characterized by a Gaussian amplitude distribution and a uniform frequency spectrum. On the other hand, speech is more complex. Its amplitude distribution can be approximated by a Laplacian function showing a peak-to-rms ratio approximately 6–8 dB higher than that of noise. Although the frequency spectrum varies from talker to talker, in general it shows a peak between 300 and 500 Hz and rolls off at 6 dB per octave or more at the higher frequencies [5]. Furthermore, speech signals occur in bursts, while noise is continuous. Speech detection algorithms have been proposed which use pitch contours, compare higher order moments, count zero crossings, or employ other similar strategies.

Level detection, one of the simplest and most reliable methods of detection, is based upon the facts that speech signals generally occur in bursts at a higher level than noise and that speech contains a greater percentage of lower frequency components than noise. Hence, several consecutive speech samples will have a much higher probability of exceeding a given threshold than noise.

The performance of this detection technique is therefore entirely dependent upon the level of the threshold. An optimum threshold should be far enough above the noise to screen out most of the noise peaks, yet low enough to detect low-level speech signals. Adjusting the threshold to the correct level would be a simple task if the noise level remained constant. However, since noise on a typical telephone channel varies over a considerable range of levels and since step changes or impulses may occur, threshold adjustment suffers from certain limitations.

The voice detectors in the SCPC [1] and SPADE [2] systems use fixed thresholds. The value of the threshold is chosen as a compromise between a high level to maximize noise immunity and a low level to yield maximum sensitivity to speech. As mentioned previously, the resulting design yields marginal speech detection and noise immunity performance.

The earlier SPEC voice-activated switch design attempted to overcome these shortcomings by allowing the threshold to vary according to the noise level on the circuit. Since the design of the new switch presented in this paper is based on the earlier SPEC design, the earlier design will be presented first.

Figure 1, which is a block diagram of the early SPEC voice switch, reveals two major sections. One, the voice detection portion, detects speech when four consecutive samples, spaced 125 μs apart, exceed a threshold T_H. The through path of the samples incorporates 4-ms delay, thus providing a buffer of 3.5 ms (the delay minus the detection time) to allow detection and thereby reduce clipping of the front of the speech burst. The switch is set to the ON state for 150 ms each time four consecutive samples exceed the threshold. Thus, it compensates for portions within the speech burst which may drop below the threshold and also provides a hangover period after the last detection to prevent the trailing portion of the speech burst from being clipped. This type of operation is common to the voice detectors in the SCPC and SPADE systems.

The earlier SPEC design uses a second threshold, T_L, in addition to the voice detection threshold, T_H, where T_H equals T_L plus one quantizing step, to adjust the threshold to the noise. The number of samples exceeding each threshold, T_H and T_L, is accumulated over a one-half-second interval or over N samples, where N equals 4,000 for 8-kHz sampling. If the number of samples exceeding T_L is less than 5 percent of N for two successive intervals, noise is considered to be below the thresholds and both are decreased by a single quantizing

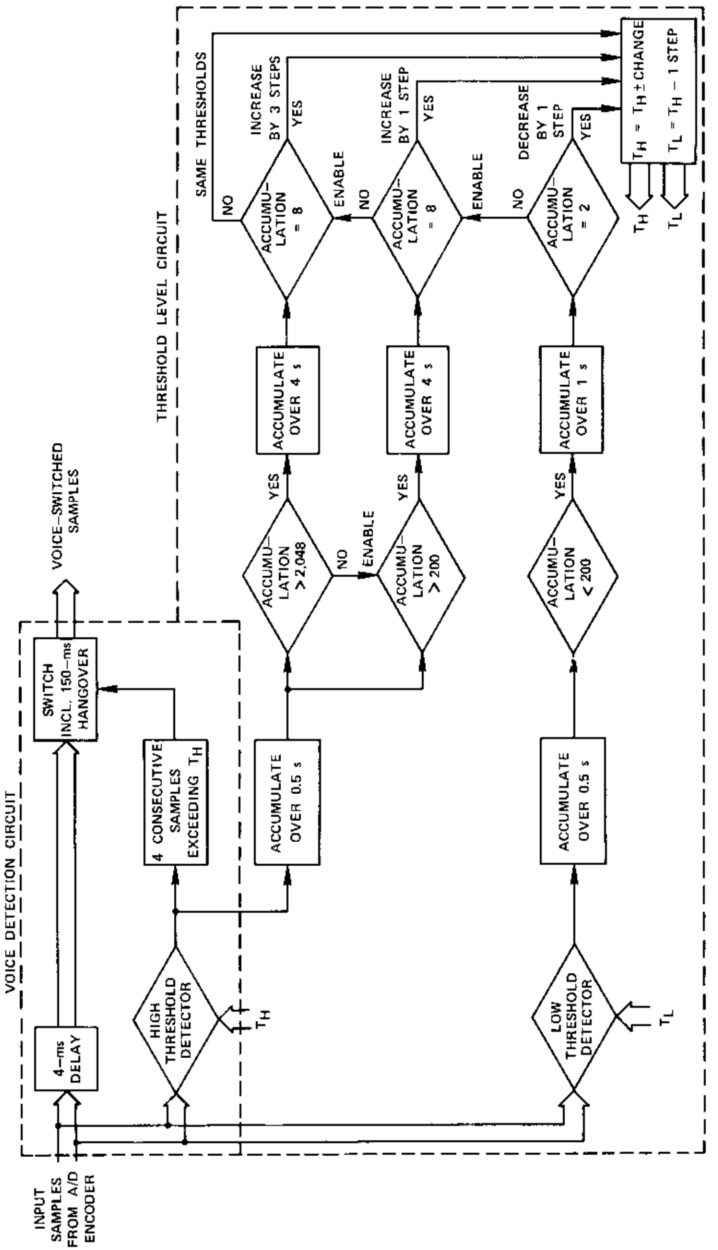


Figure 1. *Block Diagram of Previous SPEC Voice Switch*

step. If the number of samples exceeding T_H is greater than 5 percent but less than 50 percent of N for eight consecutive intervals, the noise is considered to be slightly above the thresholds and both are increased one quantizing step. If during eight successive intervals the number of samples exceeding T_H is greater than 50 percent of N, the noise level is considered to be substantially above the thresholds and both are increased by three quantizing steps. If none of these criteria are met, that is, if a continuous pattern is not detected, speech is assumed to be present and the thresholds are not adjusted.

The adjustment intervals used in the early design were determined by typical speech characteristics. Since threshold adjustment was a continuous process, it required a safeguard to prevent the speech signals from incorrectly adjusting the thresholds, and the design relied on the amplitude irregularities of speech as well as the typical speech burst durations. With a mean speech talk spurt duration of approximately 1.3 s [6] and a peak-to-rms ratio approximately 6–8 dB higher than that of Gaussian noise, speech generally would not meet the threshold adjustment criterion.

The threshold adjustment process is updated every half second. However, the initial decision to decrease the threshold levels to a low noise level will occur after one second with further reduction possible each half second thereafter. The first decision to increase the thresholds to a high noise level can occur after four seconds with subsequent increases every half second thereafter. These adjustment limitations have yielded unsatisfactory switch performance in the presence of step increases in noise level. When a step increase in noise occurs, the noise is transmitted for a minimum of four seconds before the threshold begins to adjust.

A further problem associated with this previous SPEC voice switch design was that the threshold (T_H), when adjusted to stationary noise, was positioned too close to the noise level. Consequently peaks present in normal telephone line noise quite often exceeded the voice detection threshold (T_H), hence triggering the voice switch.

Design objectives of the new switch

The design objectives of the new voice switch are as follows:

a. faster threshold adjustment to a variable noise level on the telephone circuit,

b. greater immunity to false detection of noise,
c. no noticeable clipping of speech,
d. simplification of the design where possible.

The primary consideration in developing the new design is rapid adjustment of the threshold to the optimum level above the noise on the telephone circuit. The optimum level is defined as one which simultaneously maximizes speech sensitivity and noise immunity. To adjust the threshold to this optimum point it is first necessary to decide that noise, not speech, exists on the circuit and secondly to determine the noise level.

As in the earlier SPEC design, the number of samples exceeding a threshold during a given interval of time has been used to determine the noise level. Telephone line noise can be considered to be Gaussian and the noise measuring threshold in the new design is adjusted to correspond to the upper 96-percentile point of the noise distribution. It can be shown theoretically that the upper 96-percentile point for a normal distribution with zero mean can be estimated within 10-percent accuracy by taking a sample size, N, approximately equal to 1,200. This result has been confirmed by computer simulations. Therefore, it has been decided that an interval of 150 ms* is sufficient for determining the threshold adjustments.

A real-time computer simulation has been used to optimize the remaining parameters of the voice switch. The parameters have been made variable inputs to the program, and have been optimized to yield high noise immunity with minimal speech clipping.

Design implementation

The new voice switch design employs three threshold levels: one (T_H) to detect the presence of speech, a second (T_L) to set the first threshold to the optimum point with respect to noise level, and a third (T_M) operating in conjunction with the speech detection circuit to disable the threshold adjustment circuitry while speech is present. Figure 2 is a block diagram of the switch; the relative positions of these thresholds are shown in Figure 3.

* An interval of 150 ms yields 1,200 samples for signals sampled at 8 kHz according to present C.C.I.T.T. standards.

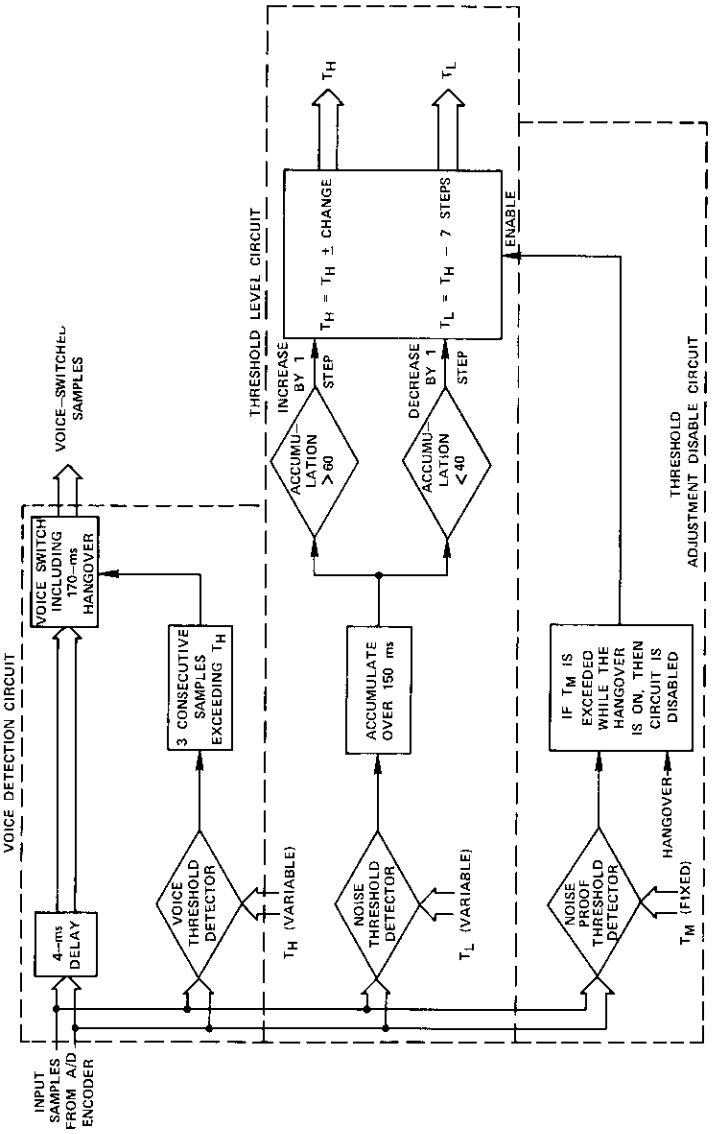


Figure 2. *Block Diagram of New Voice Switch Design*

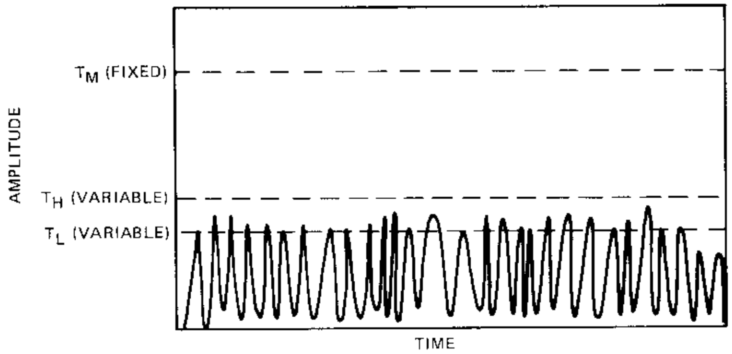


Figure 3. *Position of the Thresholds Relative to the Noise*

The first threshold, T_H, is used to detect the presence of speech using the criterion that three consecutive samples must exceed the threshold. This reduction from the four consecutive samples used in previous designs enables the threshold to be positioned farther above the noise with no reduction in detection capability.* Upon detecting speech, the switch remains in the ON state for the hangover period, which has been increased to 170 ms (as opposed to 150 ms for the previous designs) to minimize chopping on stop consonants such as p, t, and k and clipping of the end of the speech burst. The hangover is reset each time speech is detected. The input samples are delayed 4 ms in the through path before being switched to provide a buffer to reduce clipping of the front of the speech burst. (The net buffer time is 3.625 ms.)

The second threshold, T_L, is used to position the speech detection threshold, T_H, according to the noise level present. Both thresholds, where $T_H = T_L + 7$ quantizing steps, are adjusted to yield approximately 4 percent of the noise samples above T_L. The speech detection threshold is thus positioned slightly above the noise peaks, yielding good speech sensitivity with good noise immunity. Threshold adjustment is accomplished by accumulating the number of samples exceeding T_L in 150 ms. If the accumulation is greater than 5 percent the thresholds are increased by one quantizing step, and if it is less than 3.3 percent they are decreased by one step. However, the input to the voice switch contains noise punctuated by talk spurts. If threshold adjustments are allowed to continue while speech is present, the thresholds will become incorrectly adjusted and speech clipping may occur because the threshold adjustment decision interval (150 ms) is considerably less than the mean talk spurt length (1.3 s) [6].

A third threshold, T_M, is used to disable the threshold adjustments while speech is present. The threshold adjustment circuitry is disabled when T_M is exceeded and the hangover is ON and remains disabled for the entire duration of the hangover. This threshold is fixed at −23.0 dBm0, a level which is high enough so that it is not exceeded by typical telephone line noise. That is, the peaks in 54 dBrnc0 of line noise* are not likely to exceed T_M. However, the peaks of low-level speech are substantially above T_M (as will be seen in Figure 7b). When T_M is exceeded it can be assumed that a noise impulse, a continuous high-noise level, or speech is present. If the threshold adjustments are disabled by a noise impulse, the result is inconsequential since the thresholds will remain fixed for only one hangover period (0.17 s). As an override condition, noise above 54 dBrnc0 will disable the threshold adjustments and probably be continuously transmitted. This maximum noise level has been chosen because voice switching at higher noise levels will cause changes in the noise on the circuit that are so drastic that they will actually impair the ability of the listener to recognize speech [7], [8].

Performance impairment

There are two kinds of degradation introduced into a telephone channel by improper voice switch operation: false triggering on noise and speech segment clipping or simply speech clipping. False triggering on noise causes bursts of noise to be transmitted, which can be annoying to listeners. More importantly, this false triggering also contributes unnecessary system loading in speech interpolation systems and can possibly generate additional degradation. Therefore, it is necessary for voice switch designs to maintain a high degree of noise immunity.

There are several conditions which cause false triggering on noise in voice switches. Specifically, these are step increases in noise, impulses found in normal line noise, sine wave or speech-type crosstalk, the change from no noise to normal line noise when calls are connected,

* See Appendix A.

* This level is 14 dB above the Bell System long-distance noise objective.

and other conditions in which the noise signal contains any of the characteristics common to speech. Three tests have been devised to demonstrate the amount of noise immunity offered by this voice switch design. These tests concentrate on the dynamic qualities of line noise which have the greatest effect on a level-detection algorithm.

The first test, the initial adjustment test, is designed to determine the adjustment time required by the voice switch when an idle channel* becomes active. It is a good indicator of the switch's adaptation capabilities because it tests the transition from the idle channel condition, which generates the minimum threshold levels in the voice switch, to the operating condition. All signals are removed from the input to the voice switch for a sufficient time to ensure that the minimum thresholds are present. Then noise is introduced and the time necessary for the thresholds to reach equilibrium is recorded. The test is performed for 10 noise levels from 34 to 54 dBrnc0. The average adjustment times for both the previous SPEC voice switch and the new design are plotted in Figure 4. These results indicate that the new design provides a substantial reduction in the amount of noise which will be transmitted when idle channels become active.

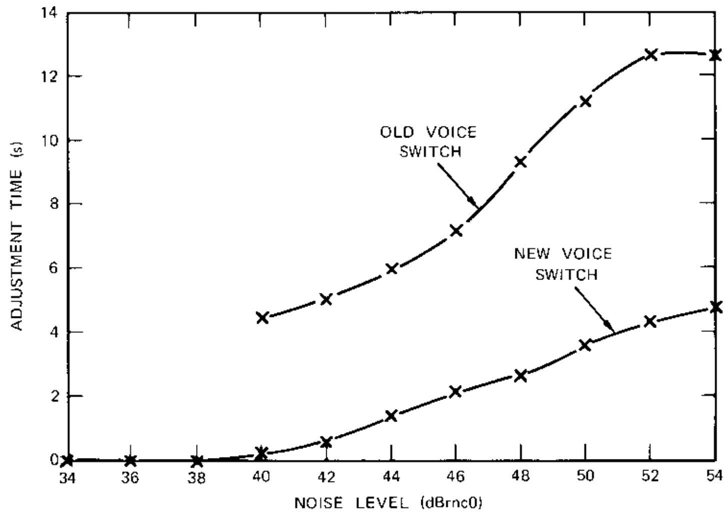


Figure 4. *Average Adjustment Time of Both SPEC Voice Switches to Noise from No Noise Condition*

The second test, the step response test, is designed to indicate the time needed for the threshold levels to react to a step change in noise. Step changes in the line noise may result from changes in crosstalk or other similar switching-type occurrences on the telephone circuit. To react to a change in noise level the thresholds start at some value and increase (or decrease) to a new level. The adjustment time for both the new and the old switch designs is measured for a 10-dB increase in noise, which is thought to be a worst-case condition. The noise is stepped from 34 to 44 dBrnc0 and from 44 to 54 dBrnc0. The test is repeated six times for each step; the average adjustment time is recorded in Table 1. These data show the substantial improvement offered by the new design in an area where the previous SPEC design needs modification.

TABLE 1. AVERAGE ADJUSTMENT TIME TO A STEP INCREASE IN NOISE FOR BOTH SPEC VOICE SWITCHES

Noise Step (dBrnc0)	Average Adjustment Time (s)	
	Old Voice Switch	New Voice Switch
44–54	10.9	3.2
34–44	5.2	0.6

The third test, the false triggering test, is perhaps the most revealing test of voice switch noise performance. It measures the amount of false triggering in a 2-minute interval and is designed to determine the degree of immunity offered by the voice switch against noise peaks normally found in line noise. In this test a constant noise level is applied to the input of the switch and the number of times the hangover period is set is accumulated over a 2-minute interval and recorded. The test is performed on both the old and new designs for noise in the range of 34 to 54 dBrnc0. The results, plotted in Figure 5, show a substantial improvement for the new design relative to the previous SPEC design. The results also indicate that this new design exhibits a maximum false triggering percentage of 0.7 percent on continuous noise.

In addition to false triggering on noise, a voice switch may also introduce speech clipping into a telephone channel. In this case, a portion of the talk spurt is chopped or clipped by the voice switch. The leading

* That is, when no signal is present, a condition occurring between calls.

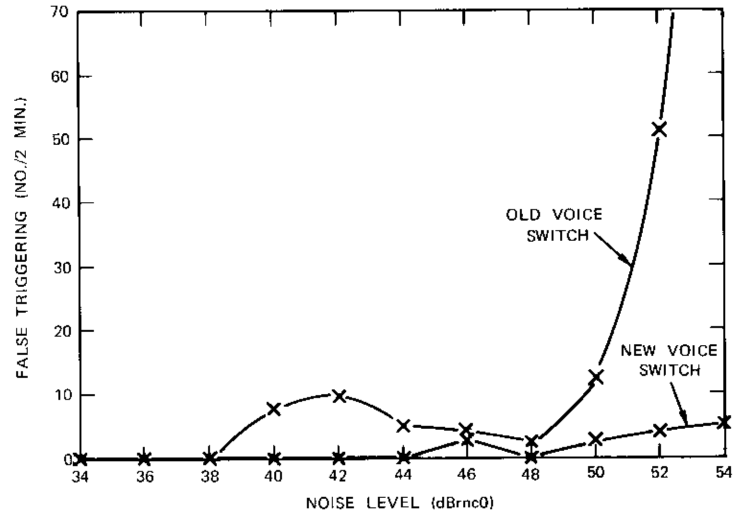


Figure 5. *Amount of False Triggering in Both SPEC Voice Switches in a 2-Minute Interval*

edge of the talk spurt is subject to a higher degree of clipping than the remainder of the talk spurt because the signal must be detected while rising out of the noise. The remainder of the talk spurt is generally well out of the noise and is therefore more easily detected.

The perceptibility of this initial clipping is a function of the duration of the clip. Listener recognition versus talk spurt clipping duration has been subjectively tested [9]–[11]. Ahmed and Fatechand's [10] testing of the effects of speech segment clipping on the articulation of consonants has revealed that front-end clips for plosives and fricatives up to 30 ms and for semi-vowels up to 40 ms do not affect the articulation. Their results indicate that clips longer than 50 ms significantly reduce articulation scores and must be minimized. Laboratory tests to determine the threshold of perception of clipping on speech indicate that clips shorter than 15 ms can be considered imperceptible to the average listener.

Standards, such as C.C.I.T.T. recommendations, which specify the duration or the rate of occurrence of speech clipping considered tolerable for telephony are not presently available. However, clips longer than 50 ms are known to destroy part of the information content of the talk spurts. Therefore, any future requirements will probably deal with the rate of clips in excess of 50 ms. An additional problem at this time is that no test procedures have been established to provide a uniform means of measuring the duration of a clip. If actual speech signals are used for testing, repetition of such tests will be very difficult as will determination of the measurement points. Therefore, several simple waveforms are generated which model speech signals to objectively evaluate the voice switch performance.

In an attempt to characterize the onset envelope of common speech waveforms, three types of signals are generated. A burst sinusoid is used to simulate the onset of stop consonant sounds such as p, t, ch, and k. A linearly ramped sinusoid is generated to simulate the onset of semi-vowel and transitional sounds such as l, m, n, w, and h. An exponentially ramped sinusoid is used to simulate the onset of pure vowel sounds such as ū, ē, a, ō, and i. The three frequencies (600, 1,200, and 1,800 Hz) of the sinusoids used with each waveform are chosen to represent formants commonly encountered in speech [5]. Each waveform is 25 ms long as a compromise of the onset times of the different sounds [12]. The maximum signal level of each waveform is varied from −6 to −23 dBm0 to simulate several of the different talker levels present in the telephone system. Each signal is individually summed with Gaussian noise and applied to the input of the voice switch. The noise level is then varied from 45 to 55 dBrnc0 in 2-dB steps and the detection time recorded. The clipping introduced by the voice switch never exceeded 20 ms in the test. Figure 6 comprises photographs of the input and the output* of the voice switch for several test signals.

These tests are intended to objectively characterize the speech detection capability of the voice switch. The photographs in Figure 6 reveal that the waveform is detected as soon as it exceeds the noise level. To show that these results may be applied to the actual speech detection capability of the voice switch, Figures 7a–7c are rectified† signals of the phrase "Joe took Father's . . ." alone, with 54 dBrnc0 of noise added, and as switched by the voice switch,‡ respectively. As can be

* The voice switch output is offset in the photographs due to the 4-ms throughput delay incorporated in the switch.

† Rectified signals are used because the actual switch uses only the magnitude of the signals.

‡ This phrase has been taken from a group of phonetically balanced sentences and represents a portion of speech known to be susceptible to clipping when detected in the presence of noise.

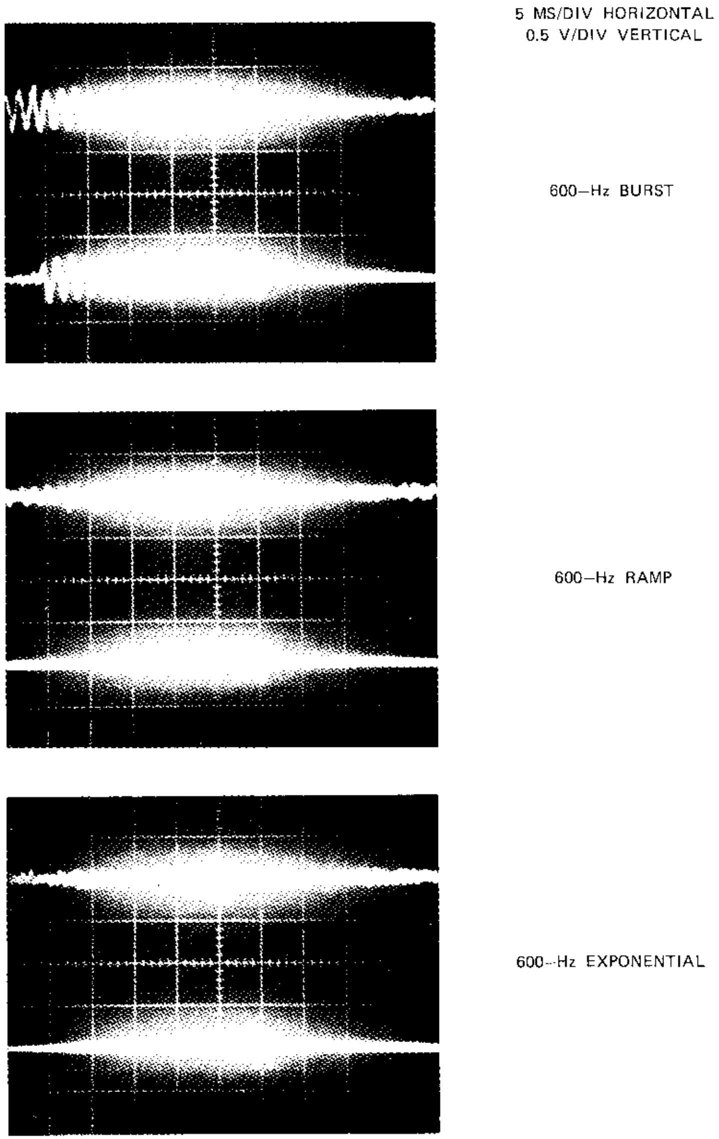


Figure 6. *Switched Sinusoidal Waveforms (25-ms, 54-dBrnc0 noise level; – 16-dBm0 maximum signal level; input to voice switch on top, output below)*

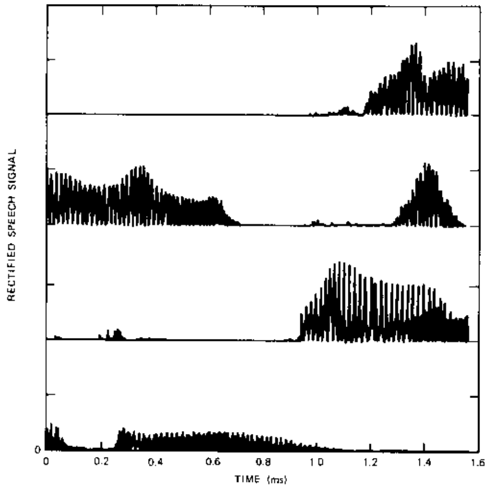


Figure 7a. *Computer-Generated Plot of the Phrase "Joe took Father's . . ."*

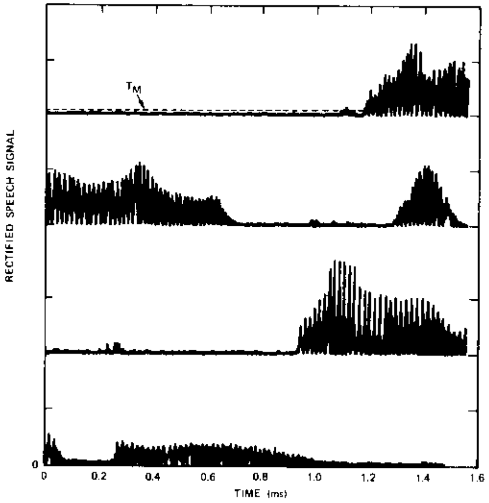


Figure 7b. *Computer-Generated Plot of the Phrase "Joe took Father's . . ." with 54 dBrnc0 of Noise Added*

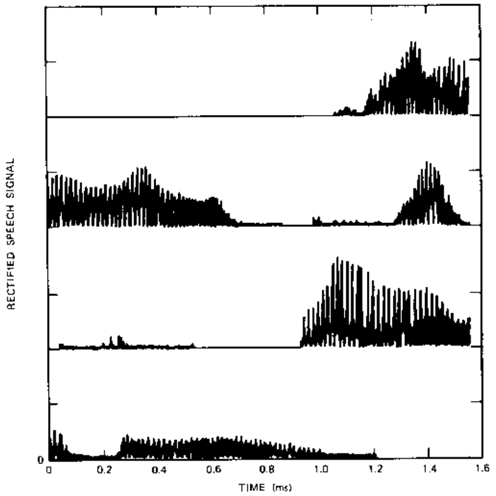


Figure 7c. *Computer-Generated Plot of the Phrase "Joe took Father's . . ." with 54 dBrnc0 of Noise as Detected by the Voice Switch*

seen in these computer-generated plots, only the very lowest level signals which have been masked by the noise are clipped.

Critical-listener subjective assessment

To assess the perceived performance of the voice switch and consequently indicate the validity of the objective tests, a limited subjective evaluation by critical listeners was performed. The listeners were telecommunications experts whose experience included talking over voice-switched circuits. Thus, they could anticipate the specific type of degradation to be detected. To provide a basis for comparison, the previous SPEC voice switch was arranged in a parallel circuit which could be selected by the listener. Noise was added to the circuits both before and after the voice switches to simulate the operating environment. All listeners were easily able to detect the operation of the previous SPEC switch, whereas a large majority indicated that they were able to detect the operation of the new design only after careful scrutiny. This result indicated good correlation between the measured maximum clip of 20 ms and the threshold of clipping perception of 15 ms. It was the unanimous opinion that the new switch functioned considerably better than the previous design and that the speech quality was very good.

Conclusions

This paper has discussed a new voice switch design and several tests involved in determining the operating characteristics of the switch. This switch offers improvements over the designs currently being used in SCPC, SPADE, and SPEC systems. Specifically its advantages include rapid threshold adjustment to varying noise levels, improved noise immunity, and enhanced switching characteristics on speech bursts. The improved noise immunity offered by the new voice switch will greatly enhance the operation of DSI systems.

Testing of the switch has involved new procedures which may provide a basis for future voice switch testing. The simple sinusoidal models used to represent speech envelopes appear to be useful indicators of the actual voice switch performance.

The need for standard tests to gauge acceptable voice switch performance is a problem which will increase in importance as the use of voice-switched systems increases. Such standard tests should ensure high-quality telephone channel performance, but they should accommodate variations in voice switch algorithms. The voice switch tests described in this paper should be useful in moving toward such a set of standard tests.

Acknowledgment

The author wishes to thank H. G. Suyderhoud for his assistance in preparing this paper; M. Onufry, R. Ridings, and F. Corcoran for their guidance and assistance in the design and development of the test procedures; Dr. J. Su for the speech plots and useful discussions on speech detection; W. Scott and R. Kuenzli for hardware design and testing of the actual SPEC voice switch circuit; P. Lutz for the original SPEC variable threshold voice switch design; and A. Speare for her patience in typing the many drafts of this paper.

References

[1] "PSK SCPC System Specification," BG-9-21E, H/5/74 (Rev. 1), 15 May 1974, Contribution of the Manager.

[2] "SPADE System Specification," ISCS-39-30E, W/6/69, 20 June 1969, SPADE Working Group Contribution.

[3] J. A. Sciulli and S. J. Campanella, "A Speech Predictive Encoding Communications (SPEC) System for Multichannel Telephony," *IEEE Transactions on Communications*, COM–21, No. 7, July 1973.

[4] H. G. Suyderhoud, J. Jankowski, and R. Ridings, "The Results and Analysis of a Field Trial of the Speech Predictive Encoding Communications System," *COMSAT Technical Review*, Vol. 4, No. 2, Fall 1974, pp. 371–393.

[5] H. Fletcher, *Speech and Hearing in Communication*, New York: Van Nostrand, 1953.

[6] Paul T. Brady, "A Technique for Investigating On-Off Patterns of Speech," *Bell System Technical Journal*, Vol. 44, No. 1, January 1965.

[7] D. W. Massaro, "A Comparison of Forward Versus Backward Recognition Masking," *Journal of Experimental Psychology*, Vol. 100, No. 2, 1973.

[8] D. W. Massaro, "Perceptual Images, Processing Time, and Perceptual Units in Auditory Perception," *Psychological Review*, Vol. 79, No. 2, 1972.

[9] G. J. Barnes, "Voice Switching Parameters in Telephony," *Electrical Communication*, Vol. 47, No. 3, 1972.

[10] R. Ahmed and R. Fatechand, "Effect of Sample Duration on Articulation of Sounds in Normal and Clipped Speech," *Journal of the Acoustical Society of America*, Vol. 31, 1959, pp. 1022–1029.

[11] H. Miedema and M. G. Schachtman, "TASI Quality—Effect of Speech Detectors and Interpolation," *Bell System Technical Journal*, Vol. 51, No. 4, July 1962.

[12] D. H. Klatt, "Voice-Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters," *Speech Communication*, Vol. VII, Cambridge, Mass.: M.I.T. Press, April 1973.

Appendix A. Optimum combination of speech sensitivity and noise immunity

A study was conducted to determine the optimum number of consecutive samples above a threshold to be used for speech detection. Threshold levels for 1, 2, 3, 4, and 5 consecutive samples yielding equivalent detection (see Figure A-1) were determined by using a recorded 2-way conversation and varying the thresholds until the "OFF" duration of the switches was equivalent. The "OFF" duration was determined by using a counter to detect the zero crossings of a 1-kHz sine wave which was controlled by the "SWITCH NOT" output of the voice detectors. A fixed level of Gaussian noise was then input to each detector and the false triggering rate measured. Figure A-2 shows the false detection rate of each voice switch. These results indicate that a detection scheme employing three consecutive samples above a threshold yields the optimum combination of speech sensitivity and noise immunity.

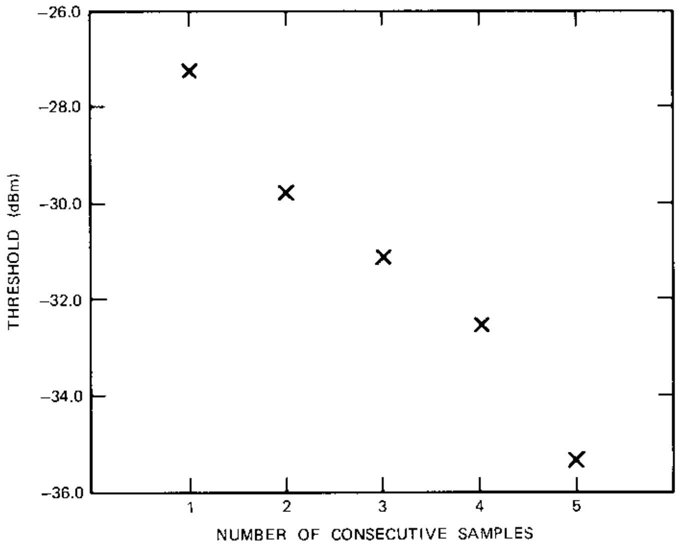


Figure A-1. *Threshold Levels Yielding Equivalent Detection for 1, 2, 3, 4, and 5 Consecutive Samples*

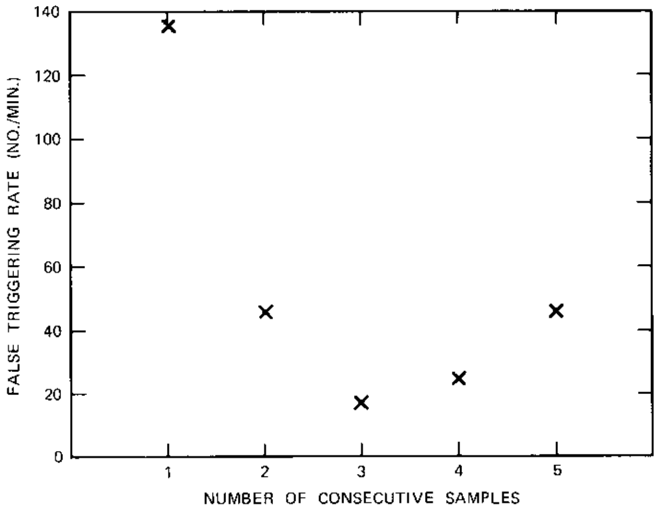
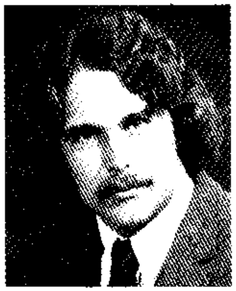


Figure A-2. *False Detection Rate at Threshold Levels Yielding Equivalent Speech Detection*



Joseph A. Jankowski, Jr., received a B.S. degree in Electrical Engineering from Northeastern University in 1973. In July 1973 he joined COMSAT *Laboratories, where he is presently a member of the technical staff of the Communications Processing Laboratory. He is a member of Tau Beta Pi.*

CTR Notes

Error control on satellite channels using ARQ techniques

A. GATFIELD

(Manuscript received November 24, 1975)

Introduction

A new form of error control system designated as selective repeat ARQ has been developed for data circuits established by means of a satellite link. The complete end-to-end circuit has the combined characteristics of the satellite link and the user-connected terrestrial circuit sections [1], [2].

Although the quality of the satellite link is excellent in terms of frequency response, relative envelope delay, noise, and phase jitter, a circuit established by a satellite link is characterized by an end-to-end transmission delay, which for geosynchronous orbit satellites is in the 300- to 350-ms range. Conventional start-stop ARQ error control systems are generally unsuitable for satellite circuits because of this long transmission delay.

This note will discuss briefly the conventional stop-and-wait ARQ, the previously developed forms of continuous ARQ, and the newly developed selective repeat ARQ system. Test data for the new system will be presented and its performance compared with that of the system described in C.C.I.T.T. Recommendation V.41. The factor of primary concern is the transmission efficiency, which is defined as the ratio of the number of correctly received information bits to the product of the transmitted bit rate and the elapsed time.

Types of error control systems

STOP-AND-WAIT ARQ

Stop-and-wait ARQ, or simple ARQ, is widely used on terrestrial circuits. The transmitter sends one block of data to the far end of the circuit to be

Allen G. Gatfield is Manager of the Image Processing Department, Communications Processing Lab, COMSAT *Laboratories.*

checked for error and then waits for a positive (ACK) or negative (NAK) response. The transmission efficiency of this technique may be poor even on terrestrial circuits if the block size is small. As the transmission loop delay becomes significant relative to the block duration, the transmission efficiency will fall rapidly.

Systems of this type may operate as full- or half-duplex systems in terms of the return or acknowledgment channel. In addition, the acknowledgment message may be protected with error detection coding.

CONTINUOUS ARQ

Continuous ARQ systems transmit data without waiting for an acknowledgment between blocks [3]. When a NAK response is received, the transmitter must back up and repeat all blocks up to and including the one in which the error has been detected.

C.C.I.T.T. Recommendation V.41 [4] describes a form of continuous ARQ known as go-back-2 ARQ. When an error occurs, the transmitter completes transmission of the current block and then goes back and repeats two blocks. The block duration must be adequate to allow for the round trip transmission time. Recommendation V.41 has been amended to add an optional block length of 3,860 bits. This amendment permits the V.41 ARQ system to operate over single-hop satellite circuits at bit rates up to 4,800 bps. The V.41 system, which uses a narrowband "backward" channel in the modem, is suited for use on 2-wire lines.

A more advanced continuous ARQ system called adaptive go-back-N ARQ has been developed. This system measures the loop delay of the circuit during its initial synchronization procedure. The number of data blocks, N, which must be repeated after each error is then adjusted to the minimum required by that particular circuit. While this system is well suited to applications which may use satellite and terrestrial circuits alternately, it provides only a slight increase in transfer efficiency on satellite circuits.

Continuous ARQ systems may use a narrowband acknowledgment channel for 2-wire operation or they may be designed for 1- or 2-way transmission of data on a 4-wire circuit. Their transmission efficiency is high for low error rates, but falls off rapidly as the error rate increases.

SELECTIVE REPEAT ARQ

The selective repeat ARQ system is specifically intended to provide high transmission efficiency for circuits having both high error rates and long transmission delays. Transmission efficiency is not a direct function of bit rate or transmission delay as in the go-back-2 or go-back-N systems. It can be increased by employing relatively short blocks and repeating only the block containing the error. The use of shorter blocks will yield a slightly higher coding overhead to maintain a given error detectability, resulting in a tradeoff between transmission efficiency at low and high error rates. This tradeoff is also encountered in go-back-N systems.

The selective repeat ARQ system implemented by COMSAT is a 4-wire full-duplex system. Acknowledgment messages for one direction of data transmission are combined with forward data for the opposite direction of transmission and protected by the block error detection code. Acknowledgment messages are repeated three times in successive blocks to ensure reception [5].

Although the selective repeat ARQ system requires more complex logic and more storage than the continuous ARQ systems, the availability of low-cost random-access-memory integrated circuits has made it practical.

Transmission efficiency

An expression for transmission efficiency, E, is [6]

$$E = \frac{k}{n} \frac{(1/B) - 1}{(1/B) + N - 1} E_w \, 100$$

where k = message bits per block
n = total bits per block
B = block-error rate
$= \dfrac{\text{blocks with one or more errors}}{\text{total blocks}}$
$= 1 - (1 - p)^n$ (for random error distribution)
p = bit-error probability
E_w = waiting factor
$= 1/(d + 1)$, where d is the round trip delay time in blocks, for stop-and-wait ARQ
$= 1$ for continuous systems
N = blocks repeated for each block error detected.

This expression encompasses first-order effects. Second-order effects associated with synchronization or special function blocks, errors in the acknowledgment channel, or limited storage in the selective repeat system are excluded.

A more complete expression, which includes the effects of limited storage and errors in the acknowledgment messages, has been developed [5] for the selective repeat ARQ system. This equation has been used to compute the selective repeat system efficiency curves in Figures 1–3.

Figure 1 shows transmission efficiency curves at a bit rate of 4,800 bps for different ARQ error control systems with a random distribution of errors and a round trip time delay of 800 ms. In the case of the selective repeat ARQ system, substantial improvement is evident for bit-error rates from 10^{-3} to 10^{-4}. However, higher coding overhead causes lower efficiency for very low error rates. One factor which is most important in this efficiency comparison is not evident in Figure 1. The selective repeat ARQ efficiency curve is independent of transmitted bit rates. The efficiency curves for both the go-back-2 and go-back-*N* systems will move to the left roughly in proportion to bit rate. The block length must be increased for the go-back-2 system, while either the block length or the number of blocks repeated must increase for the go-back-*N* system. The efficiency curves of the stop-and-wait system will drop at low bit rates if the block length is held constant and will move to the left if the block length is increased.

Laboratory tests and analysis

Laboratory tests with random data and error distribution were performed to verify the theoretical curves discussed above and to ensure that the implementation was correct. The results for the V.41 system and the selective repeat system are shown in Figure 2. In both cases the measured data include the effect of errors in the return channel. However, in the V.41 system, which uses a narrowband "backward" channel for the return path, the effect of return channel errors on efficiency is insignificant.

The tests with random error distribution show that both systems have been implemented correctly and that the selective repeat system provides the expected improvement in performance for random error distributions. However, when data are transmitted over voice band telephone circuits, the error distribution is usually not random, but is instead bunched or "bursty". Figure 3 indicates the performance of the two error control systems with the error distribution found on a test circuit combining terrestrial and satellite links. The solid lines show the expected performance for random errors, while the data points with circles and triangles represent calculated performance based on the actual error distribution found on the test circuit. Each vertical pair of test points represents 12 hours of data. The point pairs marked with an asterisk represent worst-case 1-hour periods, and the circled point pair represents the average for the entire test.

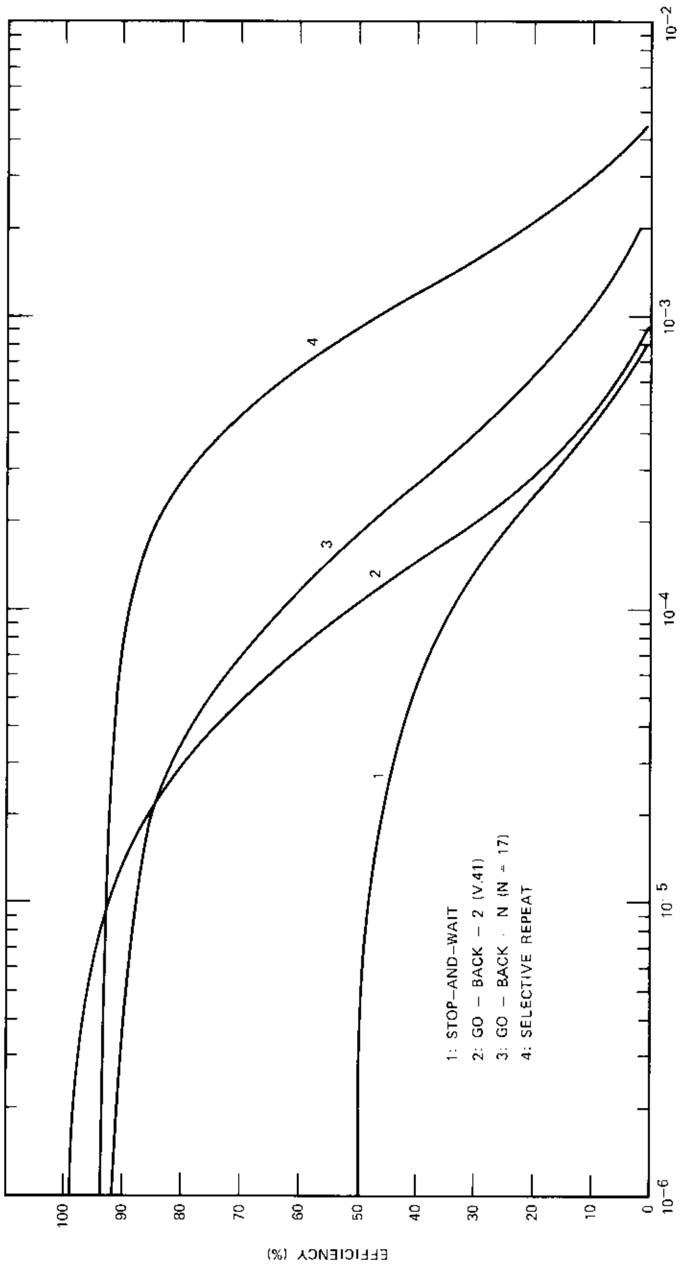


Figure 1. *Comparison of Transmission Efficiencies at 4,800 bps*

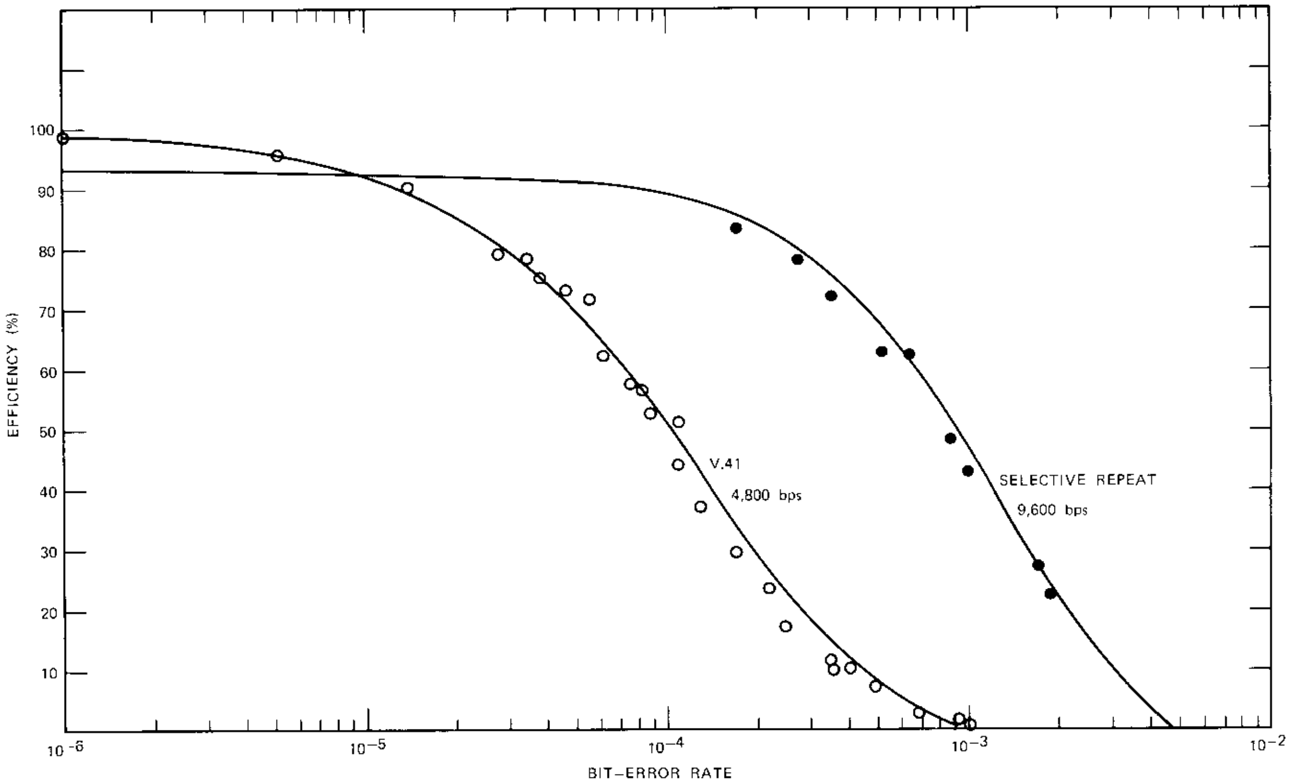


Figure 2. *Theoretical vs Measured Transmission Efficiency*

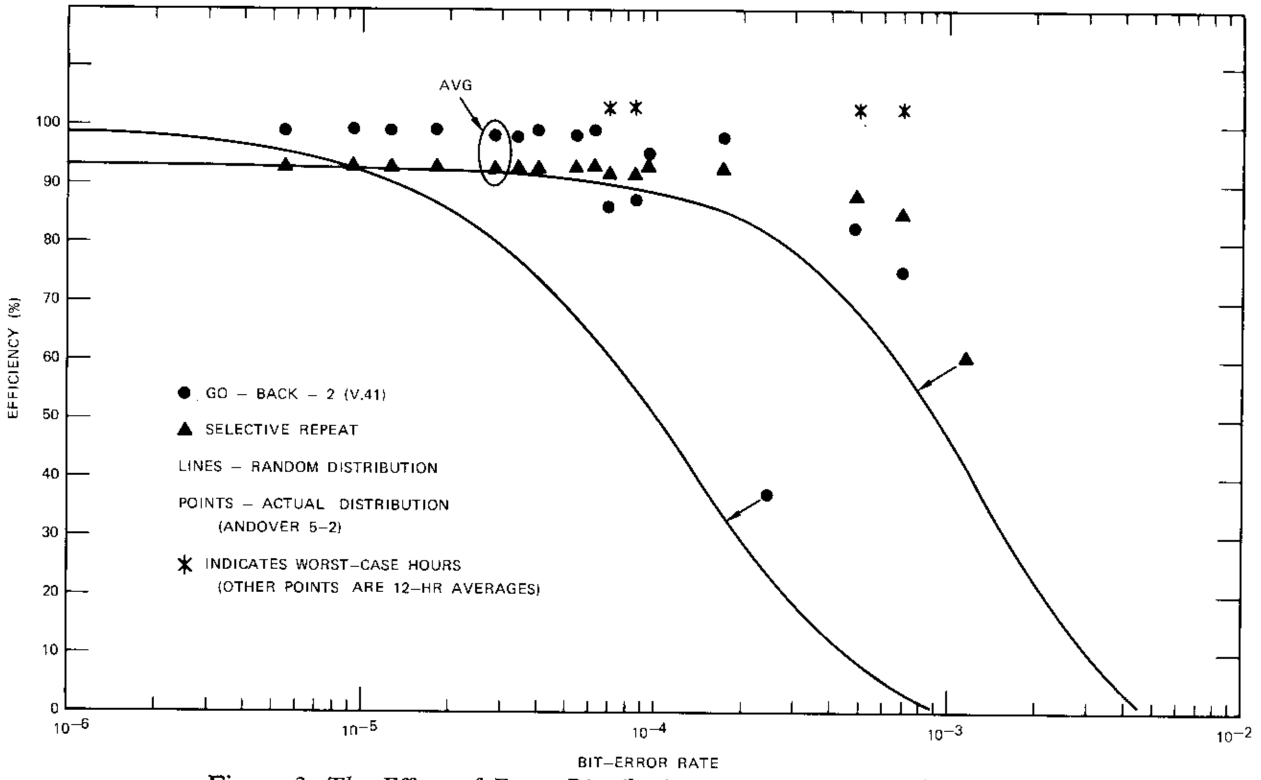


Figure 3. *The Effect of Error Distribution on Transmission Efficiency at 4,800 bps*

The data which have been collected to establish the distribution of errors on this test circuit are similar to those collected from several different circuits. However, the number of circuits which have been examined is insufficient to indicate that this distribution is typical of all circuits.

The bursty distribution of errors has caused both systems to operate at an efficiency which is higher than that predicted for random errors. This is to be expected, since bursty error distributions result in lower block error rates and allow better performance for any ARQ system. The average efficiency of the simple go-back-2 ARQ system is better than that of the selective repeat ARQ system because of the lower coding overhead. However, the increased efficiency of the selective repeat system is evident during the four worst-case hours. If the bit rate or the error rate had been higher or the error distribution had been more nearly random, the efficiency of the selective repeat ARQ system might have greatly exceeded that of the go-back-2 ARQ system.

Field tests

Field tests were conducted with both error control terminals at COMSAT Laboratories in Clarksburg, Maryland. Data grade leased lines were obtained from each error control terminal to the COMSAT earth station at Andover, Maine. The terrestrial sections at each end of the circuit were approximately 600 miles long. Spare channels in an operational carrier were used to transmit the signal from the Andover earth station to an INTELSAT satellite. Special equipment was used to permit the Andover earth station to receive its own transmitted carrier and complete a single-hop satellite link.

This test arrangement provided a 4-wire test circuit with both ends at COMSAT Laboratories. This circuit accurately represented a real intercontinental data circuit utilizing a single-hop satellite link plus leased terrestrial sections at each end.

The measured transmission delay of the circuit including the modems was 600 ms. Amplitude and delay distortion were normal and well within the range of the automatic equalizer of the modem. Signal-to-noise ratio measurements averaged about 31 dB, and long periods of error-free operation were experienced. The modem eye pattern suggested that errors, which occurred in bursts, were primarily the result of pronounced phase and amplitude hits. No effort was made to determine the cause of the errors or to improve the circuit since errors were desirable in terms of testing the error control equipment.

Tests of the two error control systems were performed in the same manner and over similar circuits. The V.41 tests were performed at 4,800 bps and the selective repeat tests at 4,800 and 9,600 bps. The test data are summarized in Table 1.

TABLE 1. ERROR CONTROL SYSTEM TEST RESULTS

Parameter	V.41 Go-Back-2	Selective Repeat	
Bit Rate (bps)	4,800	9,600	4,800
Block Length (bits)	3,860	512	512
Total Test Duration (hr)	519.5	101	75
Total Number of Blocks Transmitted	2,327,677	6,878,540	2,531,250
Total Error-Free Blocks Released to Data Sink	2,310,075	6,661,280	2,526,517
Average Transfer Efficiency (%)	98.7	89.8	92.6
Average Error Rate	3.4×10^{-6}	8.7×10^{-6}	1×10^{-6}

Conclusions

Data can be transmitted at high transmission efficiency and with the error protection afforded by ARQ techniques over circuits established by satellite links. Both go-back-2 and selective repeat ARQ systems have been shown to provide excellent transmission efficiency for voice band bit rates when operating over these circuits. For circuits having high error rates with random distributions of errors or for transmission rates above 9,600 bps, the selective repeat system provides superior results.

References

[1] A. A. Alexander, R. M. Gryb, and D. W. Nast, "Capabilities of the Telephone Network for Data Transmission," *Bell System Technical Journal*, Vol. 39, No. 3, May 1960, pp. 431–476.

[2] H. O. Burton and D. D. Sullivan, "Errors and Error Control," *Proc. IEEE*, Vol. 60, No. 11, November 1972.

[3] B. Reiffen, W. G. Schmidt, and H. L. Yudkin, "The Design of an Error Free Data Transmission System for Telephone Circuits," *AIEE Transactions on Communication Electronics*, Pt. 1, Vol. 80, July 1961, pp. 224–231.

[4] International Telegraph and Telephone Consultative Committee (C.C.I.T.T.), "Code-Independent Error Control System," *Vth Plenary Assembly, Green Book*, Vol. VIII, Rec. V.41, Geneva: International Telecommunication Union, 1972.

[5] J. A. Lockitt, A. G. Gatfield, and T. R. Dobyns, "A Selective Repeat ARQ System," Third International Conference on Digital Satellite Communications, Kyoto, Japan, November 1975.

[6] A. G. Gatfield, "ARQ Error Control on the Satellite Channel," *ICC Conference Record*, June 1974.

Electrical discharges on spacecraft at synchronous altitude

A. MEULENBERG

(Manuscript received October 25, 1975)

In recent years numerous outages and anomalous events, some quite serious, have been noted among spacecraft in synchronous orbit [1]. Of these events, it has been found that certain types have a high correlation with magnetic substorm activity and spacecraft position in the local morning time quadrant. This correlation indicates that the source of these anomalies is a spacecraft charging and discharging mechanism.

During periods of unusual solar magnetic activity (measured on earth as magnetic substorms), higher energy protons and electrons from the magnetosphere tail replace the normal low-temperature plasma at synchronous altitude. With no other effects present, a body immersed in a plasma will assume a negative potential relative to the plasma. This negative potential, whose magnitude is equivalent to the average temperature of the plasma, results from the higher velocity and hence greater collision probability of electrons relative to that of protons in the plasma. The "hot" plasma of the morning quadrant ($\geq$10-keV average energy) associated with magnetic substorms can therefore raise the spacecraft potential far above its normal range. Data from ATS-5 and ATS-6 indicate spacecraft potentials in excess of -10 keV during eclipse under these conditions [2], [3].

On INTELSAT spacecraft, outages attributable to charge-discharge phenomena have not been observed during eclipse; therefore, high spacecraft potentials alone cannot be the source of the problem. Under illumination the exposed spacecraft components will discharge by photoemission, thereby setting up potential differentials within the spacecraft itself.

This note is based upon work performed in COMSAT Laboratories under the sponsorship of the International Telecommunications Satellite Organization (INTELSAT). Views expressed in this note are not necessarily those of INTELSAT.

Andrew Meulenberg is a member of the Semiconductor Technology Department of the Applied Sciences Lab, COMSAT *Laboratories.*

These potential differences can exist between isolated conductors or across dielectrics; discharges can occur in either case. It is these discharges which are credited with causing the outages. Isolated conductors are considered to be the source of the largest discharges; discharges from dielectrics are expected to be slower and therefore less intense.

Measurement equipment on some spacecraft [4] has provided evidence of electrical discharges occurring during eclipse and in regions of reduced magnetic substorm activity. These data indicate a discharge mechanism which is different from that described previously in that neither sunlight nor magnetic storms are required. Based on results obtained from electron irradiation of dielectrics in the laboratory, a mechanism which may utilize but does not require these conditions has been proposed [5]. This new mechanism requires only a dielectric having a sufficiently high resistivity and a secondary electron emission coefficient in excess of 0.5.

Electrons entering a material will excite secondaries, some of which will have sufficient energy to escape the material, thus depleting the surface region of electrons. Since most of the incident electron beam will penetrate much deeper into the material, a bilayer, consisting of the electron-depleted surface layer and the deeper region with its excess of electrons, will develop. If the secondary emission coefficient is too small, the depletion layer will not form, or if the bulk resistivity of the material is too low, the depletion layer will be washed out by the back flow of electrons from the deeper layer.

Figure 1 shows the electron charge density and resulting fields for these situations in a dielectric with a grounded back conductive layer. The field direction is indicated by the positive or negative amplitude. The negative field indicates that electrons will flow to the right. The slight positive field at the left results from field lines terminating in the dielectric from external positive charges. It can be seen that the dielectric, with charge accumulating in the front layer and a grounded conductive layer on the back surface, will act as a capacitor being charged. The electric field from electrons being deposited in the front layer will draw a current of positive charges from ground to the back layer which is adequate to terminate all the field lines at the dielectric-metal interface and to fulfill the requirement of zero electric field within the conductor. As a result of the equilibrium charge distribution within the dielectric (Figure 1), the most intense field will exist in the back portion of the dielectric. If the stress due to this electric field is sufficiently high, dielectric discharge will originate in this region and propagate throughout the dielectric.

Figure 2, which is an enlargement of the front region of the dielectric in Figure 1, represents the equilibrium condition for a more realistic value of the secondary emission coefficient (0.5). It can be seen that the positive charge on the front surface, as a result of secondary emission, has shifted the field curve upward, causing a large field in the opposite direction near the surface and reducing the field inside the sample. Electrons are now as likely to move toward the front as to the rear. Breakdown of the dielectric will thus occur near the front surface, with electrons going to and through this surface.

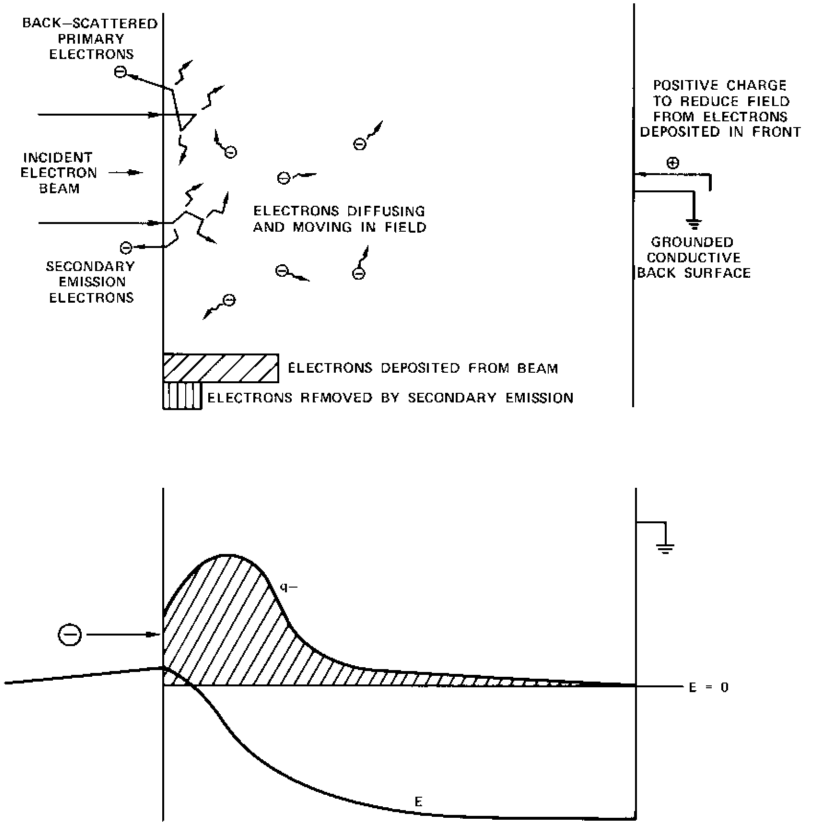

BACK–SCATTERED PRIMARY ELECTRONS
INCIDENT ELECTRON BEAM
SECONDARY EMISSION ELECTRONS
ELECTRONS DIFFUSING AND MOVING IN FIELD
POSITIVE CHARGE TO REDUCE FIELD FROM ELECTRONS DEPOSITED IN FRONT
GROUNDED CONDUCTIVE BACK SURFACE
ELECTRONS DEPOSITED FROM BEAM
ELECTRONS REMOVED BY SECONDARY EMISSION
q–
E = 0
E

Figure 1. *Electron Processes, Charge Density, and Electric Field Resulting from an Electron Beam Incident on the Left*

The relative probability of these two breakdown mechanisms in dielectrics depends on many factors such as material secondary emission characteristics, bulk and surface resistivities, and temperature; incident electron

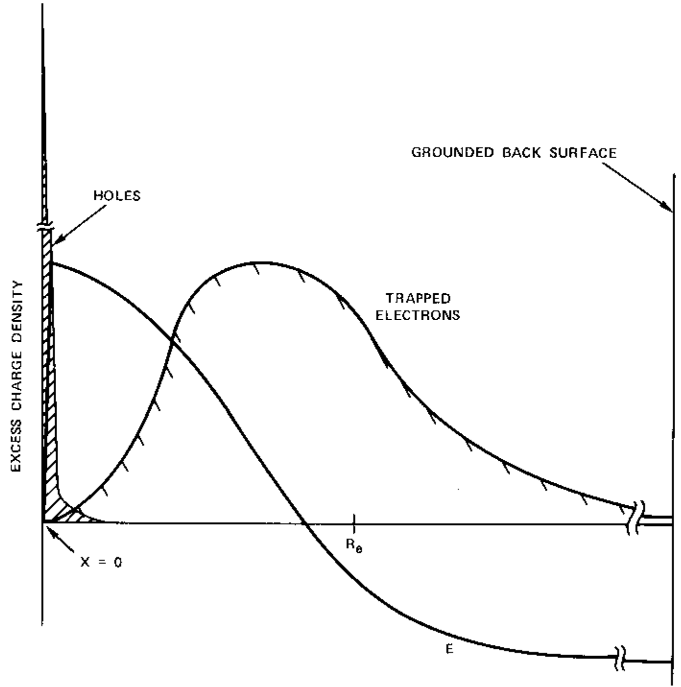


Figure 2. *Detail of Figure* 1 *with a Secondary Emission Coefficient* $\gtrsim$ *0.5* (*from Reference 5*)

flux, direction, and energy; local and general illumination; spacecraft potential; and ground connections. It is these parameters which must be manipulated to prevent discharges from occurring on or in satellite dielectrics.

Methods of preventing discharges on spacecraft must include several techniques to encompass all three possible discharge mechanisms. First, the isolated conductors can be grounded to prevent charge buildup there. Secondly, discharge through dielectrics can be prevented by adding a grounded conductive surface layer or by using an onboard ion source which will provide a plasma to neutralize any charge accumulations. Finally, bilayer formation can be prevented by reducing the dielectric resistivity (by chemical or radiation treatment) or by using a conductive surface layer thick enough (~2 μm) to absorb most of the incident electrons.

Each method of discharge prevention has disadvantages which must also be considered. For example, grounding of isolated conductors is not easy or certain when the thin aluminized layers of the thermal blankets or the silvered backs of the thousands of optical solar reflectors are considered. In addition, conductive coatings which do not degrade the thermal and optical properties of the spacecraft surface components are expensive and have not yet been shown to be applicable to large flexible substrates. Treatment of dielectrics to lower bulk resistivity, either during or after production, has been found to be effective, but the results have not yet been space qualified. Finally, an onboard ion source is an active system that requires power and fuel but can protect parts of the satellite that cannot otherwise be treated. It appears that at least two or three of these methods must be used in concert to protect the spacecraft.

Further studies must be conducted to determine material parameters which are not yet adequately understood so that the relative susceptibility of different spacecraft dielectrics and surface components to charge-discharge phenomena can be predicted. The various proposed corrective measures must then be developed and tested. Finally, the materials and measures must be evaluated individually and in various combinations to determine the optimum configuration to be used in preventing charge and/or discharge on satellites in synchronous orbit.

References

[1] A. Rosen, "Spacecraft Charging: Environment Induced Anomalies," AIAA 13th Aerospace Sciences Meeting, Pasadena, California, January 1975, AIAA Paper No. 75-91.

[2] S. E. Deforest, "Electrostatic Potentials Developed by ATS-5," *Photon and Particle Interaction with Surfaces in Space*, R. J. L. Grard, ed., Dordrecht, Holland: D. Reidel Publishing Co., 1973.

[3] S. E. Deforest, "Spacecraft Charging at Synchronous Orbit," *Journal of Geophysical Research*, Vol. 75, No. 4, February 1, 1972.

[4] J. E. Nanevicz, R. C. Adamo, and W. E. Scharfman, "Satellite Lifetime Monitoring," SRI Project 2611, Stanford Research Institute, Menlo Park, California, 1974.

[5] A. Meulenberg, "Evidence for a New Discharge Mechanism for Dielectrics in a Plasma," *AIAA Proceedings of the American Geophysical Meeting*, June 1975.

Satellite utilization of the geosynchronous orbit

W. L. MORGAN

(Manuscript received January 14, 1976)

This note is intended to show the dynamic growth in the use of the geosynchronous orbit which has occurred during the past decade and will continue in the near future. Figures 1–3, which indicate the increasing utilization of the geosynchronous orbit, give the approximate shape and longitudinal position of satellites already in orbit, as well as the nominal longitudinal positions of satellites announced for launch through about 1980. The longitudinal positions of satellites presently in orbit are based on data available circa year-end 1975. Satellites of undisclosed configurations are designated by a question mark. For simplicity, all satellites are shown on the equatorial plane and their orbit inclination is not given.

Table 1, which lists the satellites according to the key numbers indicated in the figures, provides other pertinent data. It should be noted that not all of the satellites in orbit are operated on a full-time or full-frequency-band basis, and that some satellites, having exhausted their stationkeeping fuel supply, are drifting in orbit. Most of the earlier satellites are, of course, electrically inactive. In addition, it should be noted that the increased number of satellites in orbit has already resulted in potential "parking" problems, especially in certain preferred sectors of the orbit.

New uses for communications satellites, including domestic services, TV distribution, and mobile services, are rapidly emerging. Furthermore, new users, in addition to current users such as NASA, INTELSAT, and the military, are eager to occupy orbit slots. In view of these increasing demands for satellite services, more efficient utilization of both the orbit and the spectrum could be achieved through frequency reuse via orthogonal polarization and/or multiple beams, the use of new frequency bands, and improved stationkeeping techniques.

Walter L. Morgan is a Senior Staff Scientist on the Project Staff of the Assistant Director, Technical, at COMSAT *Laboratories.*

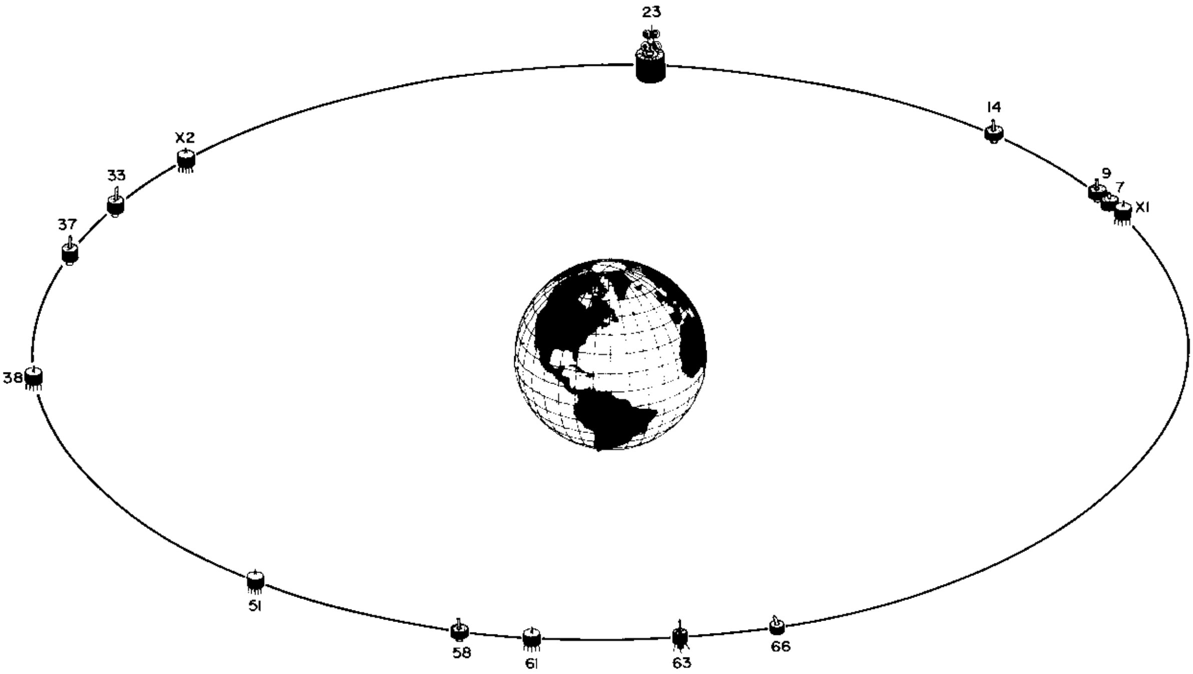


Figure 1. *Geosynchronous Satellites Launched Prior to 1970*

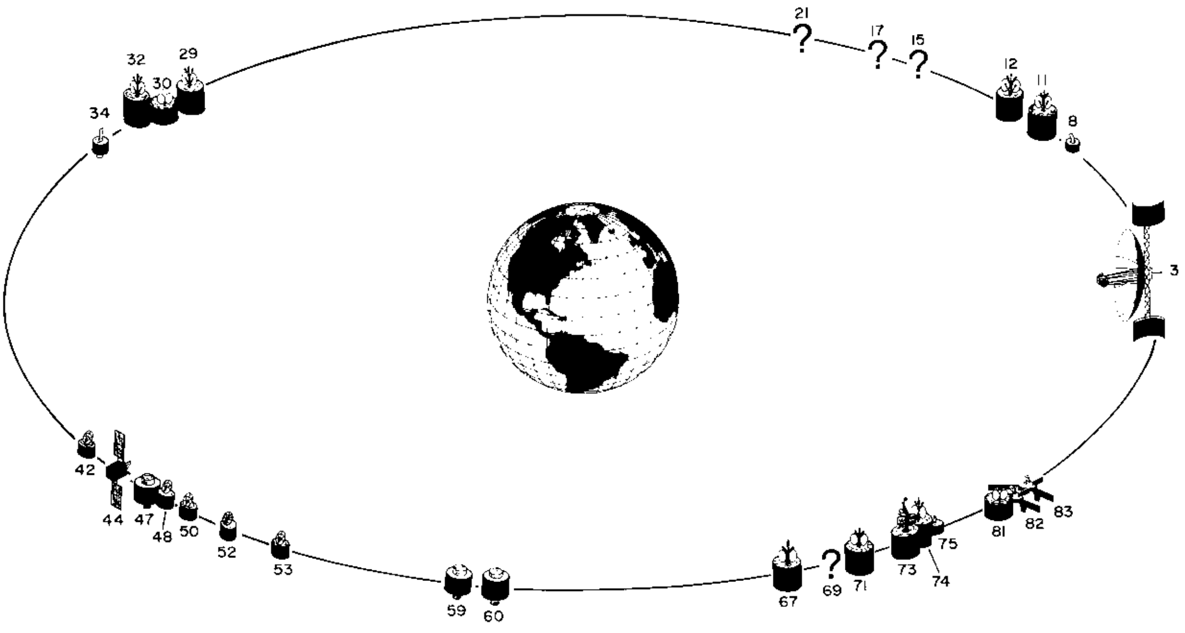


Figure 2. *Geosynchronous Satellites Launched Between 1970 and December 31, 1975*

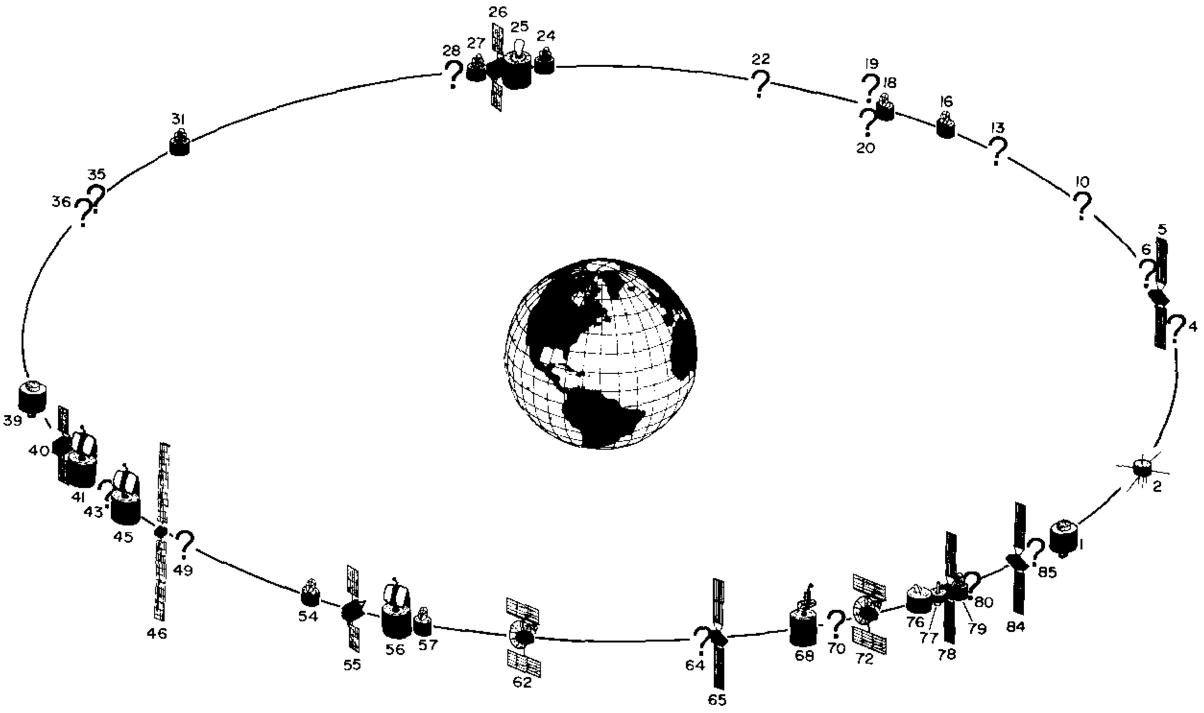


Figure 3. *Geosynchronous Satellites Launched or to be Launched after January 1, 1976*

TABLE 1. SYNCHRONOUS SATELLITE DATA

Key No.	Longitude (°E)	Satellite Name	Sponsor	Launch Year	Function	Up/Down Bands (GHz)
1	0	Meteosat	European Space Agency	1977	Meteorological	Below 4 GHz
2	0–35	GEOS	European Space Agency	1976	Experimental	
3	35	ATS-6	NASA	1974	Experimental	0.15, 1.6, 6, 13, 18/0.039–0.361, 0.86, 1.5, 2, 4, 20, 30
4	35	Statsionar 2	USSR	c. 1976	Domestic Communications	6/4
5	40	MAROTS	European Space Agency	1977	Maritime & Ship Communications	1.6, 14/1.5, 11
6	45	Statsionar 9	USSR	1980	Domestic Communications	6/4
7	55	Skynet 1	UK	1969	Military Communications	
8	55	Skynet 2B	UK	1974	Military Communications	7/8
9	56.5	INTELSAT III F3	INTELSAT	1969	International Communications	6/4
10	58	Statsionar 5	USSR	1978–79	Domestic Communications	6/4
11	60.3	INTELSAT IV F5	INTELSAT	1972	International Communications	6/4
12	63.9	INTELSAT IV F1	INTELSAT	1975	International Communications	6/4
13	70	Future Meteorological Satellite	USSR	— [a]	Meteorological	

TABLE 1. SYNCHRONOUS SATELLITE DATA (Continued)

Key No.	Longitude (°E)	Satellite Name	Sponsor	Launch Year	Function	Up/Down Bands (GHz)
14	70	INTELSAT II F2	INTELSAT	1967	International Communications	6/4
15	75–85	COSMOS 637	USSR	1975	Experimental	
16	77	(Spare)	Indonesia	—[a]	Domestic Communications	6/4
17	80	Statsionar 1 (Raduga)	USSR	1975	Domestic Communications	6/4
18	83	Palapa	Indonesia	1976	Domestic Communications	6/4
19	85	Statsionar 3	USSR	c. 1976	Domestic Communications	6/4
20	85	Statsionar 6	USSR	1979–80	Domestic Communications	6/4
21	~90	Molniya 1S	USSR	1974	Domestic Communications	
22	99	Statsionar T	USSR	c. 1976	TV Broadcast	6/0.714
23	115	TACSAT	USAF	1969	Military Communications	0.225–0.4, 8/7
24	130	ETS II	Japan	—[a]	Experimental	0.148, 2.1/0.136, 1.7, 11.5, 34.5
25	135	CS	Japan	—[a]	Domestic Communications	2, 6, 30/2, 4, 20
26	~137	BS	Japan	—[a]	TV Broadcast	14/12
27	140	GMS	Japan	—[a]	Meteorological	2
28	140	Statsionar 7	USSR	1979–80	Domestic Communications	6/4

TABLE 1. SYNCHRONOUS SATELLITE DATA (Continued)

Key No.	Longitude (°E)	Satellite Name	Sponsor	Launch Year	Function	Up/Down Bands (GHz)
29	173.6	INTELSAT IV F8	INTELSAT	1974	International Communications	6/4
30	175	DCSC–II	Defense Communications Agency (U.S.)	1971 & 1973	Military Communications	7/8
31	176.5	MARISAT	COMSAT General Corp.[b]	1976	Maritime & Ship Communications	0.3–0.312, 1.6, 6/0.248–0.26, 1.5, 4
32	177.2	INTELSAT IV F4	INTELSAT	1972	International Communications	6/4
33	180.3	INTELSAT II F3	INTELSAT	1967	International Communications	6/4
34	182.0	INTELSAT III F6	INTELSAT	1970	International Communications	6/4
35	189	TDRSS	NASA	—[a]	Experimental	6/4
36	190	Statsionar 10	USSR	1980	Domestic Communications	6/4
37	190.6	INTELSAT III F4	INTELSAT	1969	International Communications	6/4
38	211	ATS–1	NASA	1966	Experimental	Below 4, 6/below 4
39	220–225	Future SMS	NASA/NOAA[c]	—[a]	Meteorological	2/1.7
40	231	SATCOM-B	RCA[d]	1976	Domestic Communications	6/4
41	233	COMSTAR-B	COMSAT General Corp.	1976	Domestic Communications	6/4, 20, 30
42	236.5	WESTAR 2	Western Union Telegraph Co.	1974	Domestic Communications	6/4

TABLE 1. SYNCHRONOUS SATELLITE DATA (Continued

Key No.	Longitude (°E)	Satellite Name	Sponsor	Launch Year	Function	Up/Down Bands (GHz)
43	238	SBS-A	Satellite Business Systems	1979	Domestic Communications	14/12
44	241	SATCOM 1	RCA [d]	1975	Domestic Communications	6/4
45	241	COMSTAR-A	COMSAT General Corp.	1976	Domestic Communications	6/4, 30, 20
46	244	CTS	US/Canada	1976	Experimental	14/12
47	245	SMS-2	NASA/NOAA [c]	1975	Meteorological	2/0.136, 0.468, 1.7
48	246	ANIK-1	TELESAT Canada	1972	Domestic Communications	6/4
49	250	SBS-B	Satellite Business Systems	1979	Domestic Communications	14/12
50	251	ANIK-2	TELESAT Canada	1973	Domestic Communications	6/4
51	255.3	ATS-5	NASA	1969	Experimental	6/4
52	256	ANIK-3	TELESAT	1975	Domestic Communications	6/4
53	261	WESTAR 1	Western Union Telegraph Co.	1974	Domestic Communications	6/4
54	265	NSS-A	National Satellite Systems [e]	— [a]	Domestic Communications	6/4
55	269	SATCOM-C	RCA [d]	— [a]	Domestic Communications	6/4
56	270.3	COMSTAR-C	COMSAT General Corp.	— [a]	Domestic Communications	6/4, 20, 30

TABLE 1. SYNCHRONOUS SATELLITE DATA (Continued)

Key No.	Longitude (°E)	Satellite Name	Sponsor	Launch Year	Function	Up/Down Bands (GHz)
57	274.3	NSS-B	National Satellite Systems [e]	— [a]	Domestic Communications	6/4
58	280.4	INTELSAT III F2	INTELSAT	1968	International Communications	6/4
59	285	SMS-1	NASA/NOAA [c]	1974	Meteorological	2/0.136, 0.468, 1.7
60	290 [f]	GOES-1	NASA/NOAA [c]	1975	Meteorological	2/0.136, 0.468, 1.7
61	290	ATS-3	NASA	1967	Experimental	6/4
62	290	FLTSATCOM	US Navy	— [a]	Military Communications	0.29-0.32, S/ 0.24-0.27, S
63	312.4	INTELSAT I F1	INTELSAT	1965	International Communications	6/4
64	319	TDRSS	NASA	— [a]	Experimental	6/4
65	320	AEROSAT	— [g]	— [a]	Aeronautical Communications	VHF, 1.6, 5/VHF, 1.5, 5
66	325.8	INTELSAT II F4	INTELSAT	1967	International Communications	6/4
67	329.8	INTELSAT IV F7	INTELSAT	1973	International Communications	6/4
68	330.5	INTELSAT IV-A F2	INTELSAT	1976	International Communications	6/4
69	336	COSMOS 775	USSR	1975	?	
70	335	Statsionar 8	USSR	1980	Domestic Communications	6/4
71	336.9	INTELSAT IV F3	INTELSAT	1971	International Communications	6/4
72	337	FLTSATCOM	US Navy	— [a]	Military Communications	0.29-0.32, S/ 0.24-0.27, S

TABLE 1. SYNCHRONOUS SATELLITE DATA (Continued)

Key No.	Longitude (°E)	Satellite Name	Sponsor	Launch Year	Function	Up/Down Bands (GHz)
73	338.5	INTELSAT IV-A F1	INTELSAT	1975	International Communications	6/4
74	341	INTELSAT IV F2	INTELSAT	1971	International Communications	6/4
75	342	NATO-2	NATO	1971	Military Communications	8/7
76	342	NATO-3	NATO	1976	Military Communications	8/7
77	345	SIRIO	Italy	1977	Experimental	18/12
78	345	AEROSAT	—[g]	—[a]	Aeronautical Communications	VHF, 1.6, 5/VHF, 1.5, 5
79	345	MARISAT	COMSAT General Corp.[b]	1976	Maritime & Ship Communications	0.3–0.312, 1.6, 6/0.248–0.26, 1.5, 4
80	346	Statsionar 4	USSR	1978–79	Domestic Communications	6/4
81	347	DCSC–II	Defense Communications Agency (U.S.)	1971 & 1973	Military Communications	8/7
82	348.5	Symphonie 1	France & West Germany	1974	Experimental	6/4
83	348.5	Symphonie 2	France & West Germany	1975	Experimental	6/4
84	350	OTS	European Space Agency	c. 1977	Domestic Communications	14/12
85	350	(TV Broadcast)	West Germany	—[a]	TV Broadcast	14/12
X1	—[a]	SYNCOM-2	NASA	1963	Experimental	8/2

TABLE 1. SYNCHRONOUS SATELLITE DATA (Continued)

Key No.	Longitude (°E)	Satellite Name	Sponsor	Launch Year	Function	Up/Down Bands (GHz)
X2	—[a]	SYNCOM-3	NASA	1964	Experimental	8/2
—	—[a]	NATO-1	NATO	1970	Military Communications	8/7
—	—[a]	TELESAT F-4	TELESAT Canada	1978	Domestic Communications	6, 14/4, 12
—	—[a]	ECS	Japan	—[a]	Experimental	6, 35/4, 32
—	—[a]	DSCS III (2)	Defense Communications Agency (U.S.)	—[a]	Military Communications	8/7
—	—[a]	INTELSAT V	INTELSAT	c. 1979	International Communications	6, 14/4, 11

[a] Undetermined.

[b] Leader of the consortium.

[c] NOAA is the National Oceanographic and Atmospheric Administration.

[d] RCA is RCA Corporation and subsidiaries.

[e] National Satellite Systems is a subsidiary of Hughes Aircraft.

[f] Initial location.

[g] This consortium consists of the European Space Agency, COMSAT General Corporation, and Canada.

Translations of Abstracts in this issue

Manoeuvre d'acquisition d'attitude pour satellites stabilisés par volant d'inertie

M. H. KAPLAN ET T. C. PATTERSON

Sommaire

Certains satellites géostationnaires stabilisés selon les trois axes sont mis en rotation autour de l'un de leurs axes principaux pendant l'orbite de transfert afin de stabiliser leur orientation, obtenir l'énergie solaire nécessaire aux fonctions de télémesure et de télécommande et minimiser les gradients thermiques. L'emploi d'un volant d'inertie pour le pilotage sur orbite et celui d'un moteur d'apogée pour l'injection sur l'orbite finale causent un problème de transition, car au début, le satellite sera probablement en rotation autour de son axe de lacet, alors que le volant d'inertie sera monté sur l'axe de tangage. Il faut donc décelerer la rotation du satellite et le réorienter, et il faut aussi lancer le volant d'inertie.

Cet article présente une séquence d'acquisition d'orientation qui utilise au minimum les détecteurs, la logique et le propergol. La technique est basée sur une manoeuvre en boucle ouverte qui fait appel au moteur couple pour réorienter le satellite tout en mettant en marche le volant d'inertie. Il importe au plus haut point d'obtenir une interprétation physique de cette séquence dynamique parce que cette manoeuvre n'est pas bien comprise et l'ingénieur travaillant à la conception des satellites manque encore de données quantitatives pratiques relatives à la performance et aux conditions de stabilité. Les résultats indiquent que cette manoeuvre n'est pratique que lorsque l'axe initial de rotation est l'axe d'inertie maximale. Sinon, elle donne lieu à un mouvement instable ou à des valeurs considérables d'angle de nutation par rapport à l'axe du volant. On présente également une équation simple relative à la performance de nutation.