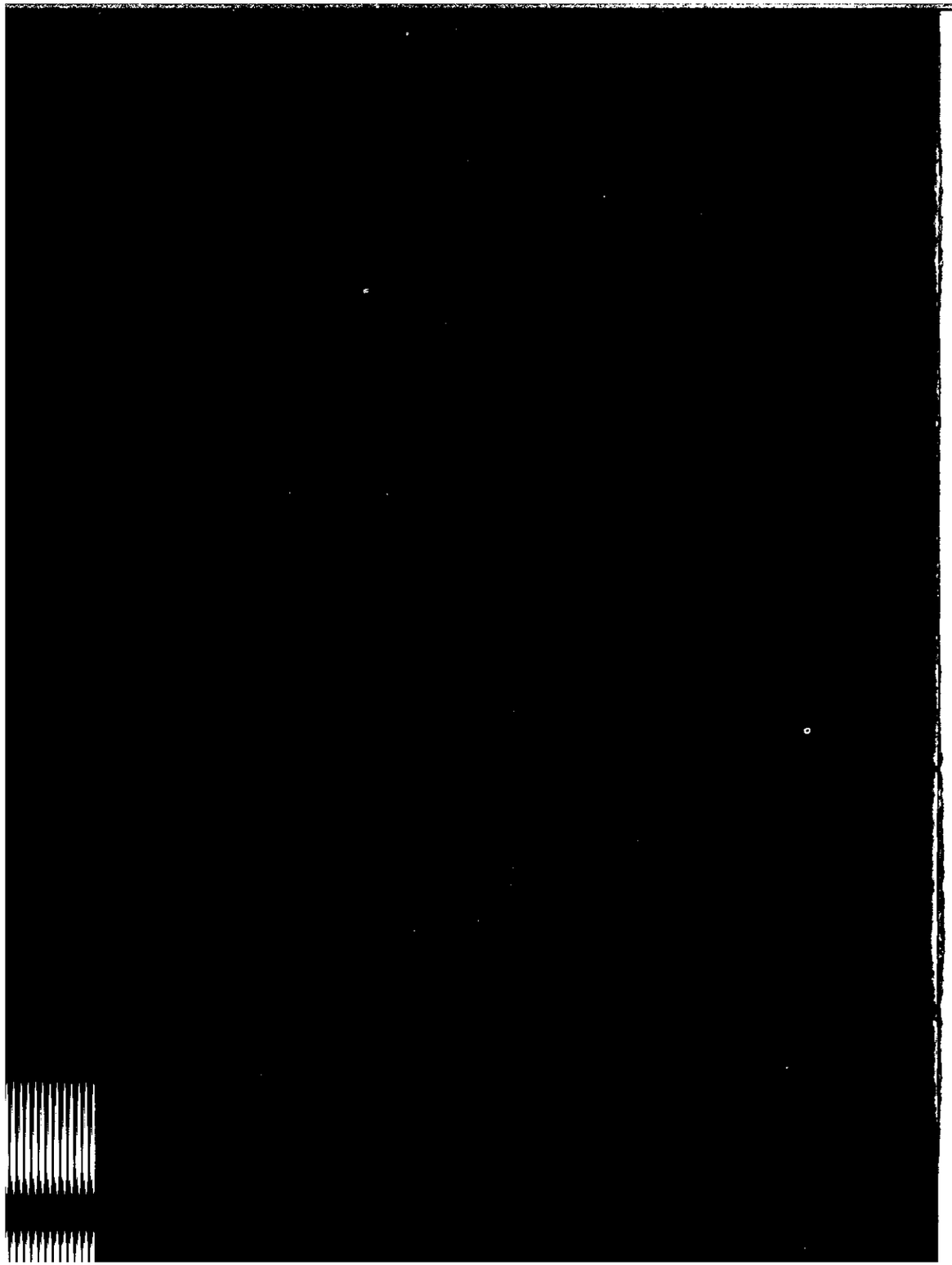


COMSAT Technical Review

Volume 3 Number 1, Spring 1973

The COMSAT TECHNICAL REVIEW is published twice a year by the Communications Satellite Corporation (COMSAT). Subscriptions including postage, $5.00 U. S. per year; single copies, $3.00 U. S. Bulk rates are available upon request. Remittances for subscriptions should be made payable to The Treasurer, COMSAT, and addressed to COMSAT TECHNICAL REVIEW, Room 5273, 950 L'Enfant Plaza, S. W., Washington, D. C. 20024, U. S. A.

COMSAT Technical Review

Volume 3 Number 1, Spring 1973

COMSAT Technical Review

Volume 3 1973

Spring

Index: spacecraft, solar cells, photovoltaic efficiency, quantum efficiency, near ultraviolet efficiency.

*The violet cell: An improved silicon solar cell**

J. LINDMAYER AND J. F. ALLISON

Abstract

State-of-the-art silicon solar cells exhibit a poor quantum yield at short wavelengths; below 0.5 μm the typical response drops sharply. Extensive work has resulted in an extension of the response to wavelengths as short as 0.3 μm, significantly improving the solar cell current. The conversion efficiency has been further improved by an increased fill factor. The combination of a short wavelength response and a sharper I-V curve has produced a conversion efficiency which is about 30 percent higher than that of state-of-the-art cells for space applications. The improved solar cell is called the "violet cell."

Introduction

State-of-the-art silicon solar cells perform considerably below predicted conversion efficiency limits. The projected efficiency of converting solar energy to electrical energy varies widely, depending on the parameter values assumed. For relatively long-lived minority carriers and thick layers, an upper bound of about 20 percent has been projected [1]–[3]. By contrast, we find that real space-quality n^+-p junction solar cells 300 μm thick with a base resistivity of 10 Ω-cm exhibit a typical efficiency somewhat above 10 percent outside of the atmosphere. After these cells are irradiated

*This material was presented in part at the *Ninth IEEE Photovoltaic Specialists Conference*, Silver Spring, Md., May 2–4, 1972.

by 1-MeV electrons to a level of 3×10^{14} e/cm^2, the conversion efficiency falls to the neighborhood of 8.5 percent.

Conventional cells are very limited in the short wavelength region [4] and their diode characteristics are far from ideal. In a recent review paper, Wolf [5] calculated that, with a junction depth of 2,000 Å and a front surface recombination velocity reduced to the order of 10^2 cm/s, the quantum yield could be significantly improved in the 0.45- to 0.6-μm range, resulting in a 17-percent improvement in the photocurrent.

In actuality, it can be shown that recombination in the silicon crystal assumes controlling importance. There are four basic regions in which recombination mechanisms can be identified:

a. In the diffusion layer, recombination of photocarriers generated by short wavelength light limits the blue-violet response of the cell.

b. In the space charge layer of the junction, recombination primarily affects the sharpness of the junction (fill factor).

c. In the bulk region, recombination of photocarriers generated by penetrating light affects the red response and limits photovoltage.

d. At the rear contact, interface recombination limits the infrared response; the significance of this effect depends on cell thickness.

This paper reports that, as a result of an independent study concerning the first two recombination mechanisms, the short wavelength response can be extended and the diode characteristics improved to near-theoretical values. During this study it was found that recombination of photocarriers generated by blue-violet light is not controlled by front surface recombination. Instead, the front regime should be broken up into three regions: a shallow region with an extremely short lifetime, called the "dead layer"; a high-field region maintained by the impurity profile; and the actual space charge region. This model indicates the importance of minimizing the dead layer thickness and the recombination states appearing in the space charge layer so that the short wavelength response can be extended over the entire solar spectral range and the diode characteristics can be made nearly ideal. The associated gains in current and fill factor are not subject to degradation from high-energy electron irradiation.

The average conversion efficiency, specified with respect to total area, has been raised above 13 percent. (With respect to active area, the corresponding efficiency is about 14 percent.) This increased efficiency has not been obtained by increasing the minority carrier lifetime with associated radiation-sensitive red response (effective lifetime less than 10 μs). The operating efficiency after irradiation by 1-MeV electrons to 3×10^{14} e/cm^2 is 11.5 percent. This latter figure is quite independent of silicon thickness, at least down to thicknesses of 150 μm.

In view of the increased blue-violet response, proper simulation of the solar spectrum required special attention. In addition, since conventional antireflective coatings of SiO_x and TiO_x show absorption at short wavelengths, it was necessary to develop fully transparent coatings with appropriate refractive indexes. In this respect, the dead layer model shows that the problem of varying surface recombination velocities associated with various coatings can be ignored. Instead, stress and optical requirements are the main considerations. The cells described here employ tantalum oxide antireflective coatings, which have the required transparency and refractive index.

The reduced depth of impurity diffusion employed here to minimize the dead layer thickness significantly increases the lateral resistance of the cell's n-diffused layer. In addition, the better basic diode requires a lower than usual series resistance. To respond to these demands, the collecting metal grid was changed profoundly to result in a new pattern containing about 60 fine lines over a 2-cm length. The new grid configuration is called fine geometry. Since the short wavelength response of the present cell is much better than that of conventional cells, it is called the "violet cell."

Fine geometry and diffusion

The benefits offered by the violet cell cannot be realized without a major change in the grid collection pattern because of the high lateral resistance of the thin diffusion layer. The degradation of efficiency with series resistance can be estimated readily from the I-V characteristic:

$$I = I_0 \left(e^{V/V_0} - 1\right) - I_{sc} \tag{1}$$

where I_0 is the theoretical reverse current (a constant truly applicable in the forward direction only), V is the photovoltage, I_{sc} is the short-circuit photocurrent, and V_0 is the thermal voltage, which is equal to kT/q for the ideal diode. Note that, in the power-producing quadrant, I is negative, whereas I_{sc} is positive by definition. From equation (1), the internal conductance of the cell is

$$G_i = \frac{dI}{dV} = \frac{I_{sc} + I + I_0}{V_0} \cong \frac{I_{sc} - |I|}{V_0} . \tag{2}$$

Note also that the highest conductance occurs near the open-circuit voltage ($I = 0$). For high fill factors the current at the maximum power point is about $I_m \sim 0.95\, I_{sc}$ and $V_0 = 26$ mV at room temperature. Hence, the conductance at the maximum power point is

$$G_m \cong \frac{I_{sc} - |I_m|}{V_0} \simeq \frac{0.05\, I_{sc}}{V_0} = \frac{I_{sc}\,(mA)}{520} . \tag{3}$$

The relative power loss, as opposed to the useful power in the load, is $R_s G_m$ (for small losses). By restricting the loss to 1 percent, we find an approximate series resistance maximum of 0.032 Ω. This figure indicates a degradation rate of |2.5 %|/0.1 Ω in the fill factor. Hence, it can be seen that a very low series resistance is needed if the ideal diode curve is to be approached.

In order to optimize the fine geometry pattern, a simple parallel bar pickup pattern, shown in Figure 1, will be analyzed. The bars are spaced at a distance d on a cell having linear dimensions of d_0. It is assumed that the width of a line is much smaller than the spacing between lines. Figure 1 shows two neighboring lines and the equivalent current source lines. If the resistance of the surface layer is denoted as $R_\square$ (ohms/square), the series resistance becomes

$$R_s = R_\square \frac{d}{4} \frac{1}{d_0} \frac{1}{2m} \tag{4}$$

where m is the number of lines (having two edges). Since $d_0/d = m$ for equally spaced lines,

$$R_s = \frac{R_\square}{8} \frac{1}{m^2} . \tag{5}$$

The important conclusion is that the series resistance decreases with the square of the number of lines used. In the conventional technology, six pickup bars are used (on the 2-cm × 2-cm cell) and the usual diffusion creates a resistance of about 50 ohms/square in the n^+-p cell. From equation (5), the series resistance would be about 0.17 Ω. In reality, however, experimental measurements indicate a resistance of about 0.25 Ω, suggesting additional contact resistance. Clearly, the conventional geometry cannot be used when, for example, the sheet resistance is increased

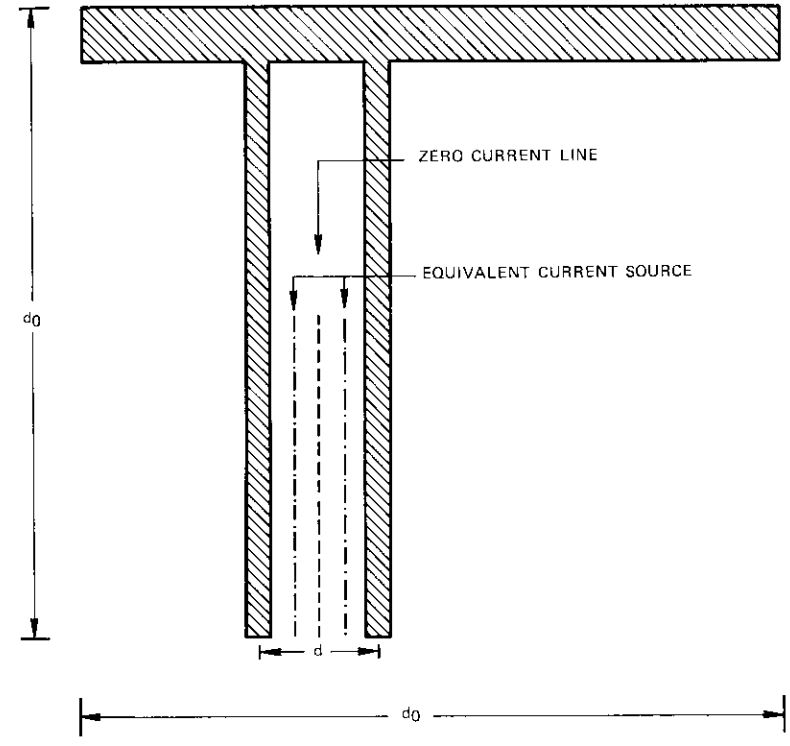


Figure 1. *Simple Illustrative Aid for Calculation of Series Resistance as a Function of Grid Spacing*

by an order of magnitude; in such cases the series resistance would be prohibitively high.

This work employed a geometry consisting of about 60 lines. An order-of-magnitude increase in the number of lines would provide a new degree of freedom. If, for example, the lateral resistance were an order of magnitude higher (about 500 ohms/square), this structure would allow for a series resistance in the hundredths of an ohm. With the pattern shown in Figure 2, it was possible to hold the grid area obstruction loss to only 5 to 7 percent. In addition, it was possible to maintain a negligible series resistance for the metal pattern itself.

The n^+ layer was formed by diffusion of phosphorus into the silicon. Diffusion time and temperature were mapped and correlated with the

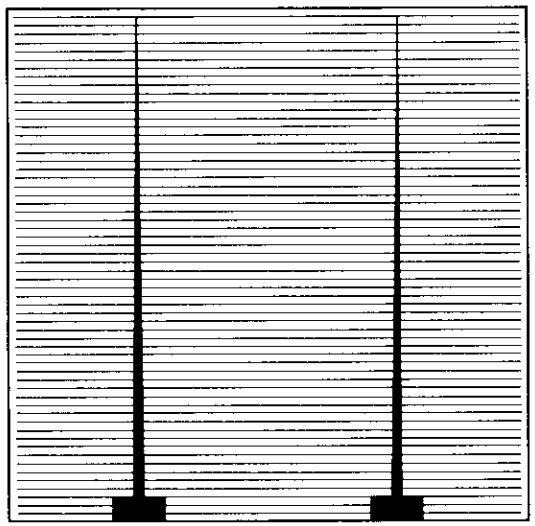


Figure 2. *Fine Geometry Contact Pattern*

resulting junction depth. First, the usual junction depth of 4,000 Å was reproduced, and then cells were made with progressively decreasing junction depths of 2,900 Å, 1,500 Å, and approximately 1,000 Å. The majority of the cells finally employed this latter thickness, and it was found that with proper precautions a 1,000-Å junction depth is practicable and results in stable cells.

The diffusion studies showed that the lateral conductivity of the diffused layer drops much faster with junction depth than linearly. At a depth of 1,500 Å, the sheet resistance is about 500 ohms/square. Another parameter monitored carefully during these studies was the sharpness of the I-V curve in the forward direction. Stresses and defects originating at the silicon surface propagate into the crystal and create recombination sites in the space charge layer, causing deviation from the ideal diode characteristics with an attendant reduction in the fill factor [6].

It is well-known [7], [8] that, in the usual diffusion process, the distribution of phosphorus does not follow a complementary error function distribution characteristic of simple diffusion processes. Instead, a nearly constant impurity concentration regime arises near the surface at the solid solubility level. Actually, the constant concentration regime may be

characterized by a mixture of substitutional and interstitial phosphorus, with an accompanying very short minority carrier lifetime.

Figure 3 shows typical phosphorus distributions for junction depths of 4,000 Å, 2,900 Å, and 1,200 Å. It can be seen that the 4,000 Å diffusion has an attendant heavily damaged layer of about 1,500 Å, but this "dead layer" diminishes in thickness as the diffusion depth becomes more shallow. The

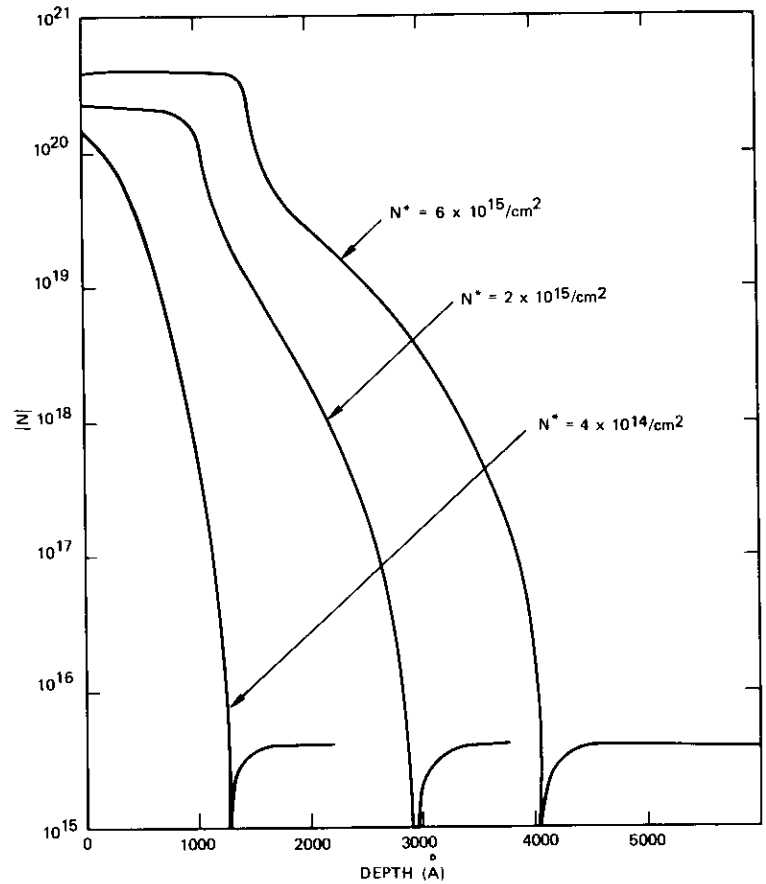


Figure 3. *Diffusion Profiles for Phosphorus in Silicon for Three Junction Depths (N* Denotes the Integrated Impurity Concentration)*

critical concentration at which the dead layer begins to develop is the same order of magnitude as the surface density of silicon atoms, i.e., on the order of 10^{15} atoms/cm^2.

Shockley has shown that dislocation generation is not only dependent on surface concentration, but is also affected by the total number of impurities (N*) found in a unit surface area [9]. A model advanced by Van Der Merwe demonstrates that above a critical misfit dislocations are created; the critical value for phosphorus in silicon appears to be $N^* = 1.2 \times 10^{15}$ atoms/cm^2 [10]. As indicated in Figure 3, a great decrease in dislocation density will occur for the shallower junctions. While it is generally true that the dislocation density decreases sharply toward the diffusion front, the total density of such dislocations is minimized by shallow junctions.

Fine geometry allows for shallow junctions, thus improving the blue-violet response and also increasing the perfection of the diode by reducing the dislocations.

Model for short wavelength response

To compute the short wavelength response, it is usually assumed that a lifetime can be assigned to minority carriers in the diffused layer and that the presence of the surface can be taken into account by a surface recombination velocity. For the defects introduced by diffusion, the situation is somewhat different. Figure 4 is an energy diagram for an n$^+$-p junction with a distance-dependent recombination state density. (The density of recombination states decreases with increasing depth into the bulk beyond the junction.) There are three easily distinguishable areas of importance:

a. *Dead Layer.* In a conventional solar cell with a junction depth of 3,500–4,000 Å, the width of the interstitial phosphorus layer is at least 1,000 Å. Photocarriers cannot be collected from this region where the diffusion transit time is longer than the lifetime

$$\frac{x_0^2}{D_p} > \tau_p \ (n^+\text{-}p \ cell) \ , \tag{6}$$

where x_0 is the width of the dead layer, D_p is the diffusion constant of holes, and τ_p is the lifetime of holes. If it is assumed that the diffusion constant is about equal to one, the transit time is on the order of 10^{-10} s. This layer is dead for lifetimes less than 100 ps. Such a short lifetime is

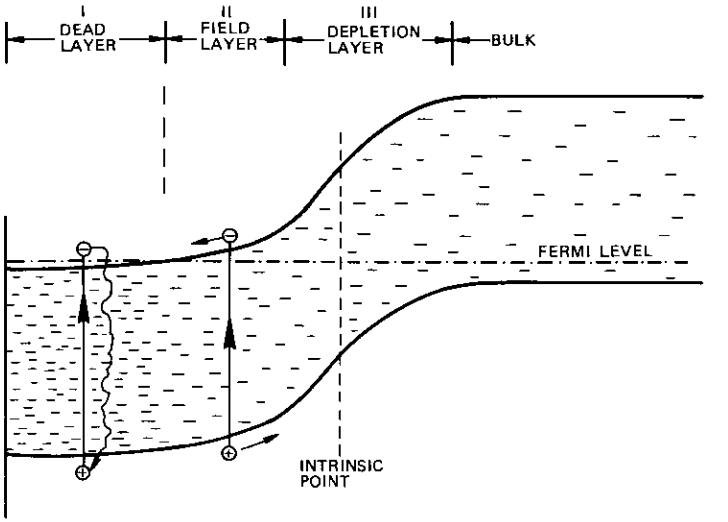


Figure 4. *Energy Band Diagram for n$^+$-p Junction*

possible if we remember that this is a degenerate region with a recombination state density on the order of the atomic density.

b. *Field Layer.* In the region where the impurity concentration decreases with distance, the electric field is well above 10 kV/cm. At such fields, the carriers approach saturation drift velocities [11] (10^7 cm/s for electrons and somewhat less for holes). Now the drift transit time may be compared with the lifetime. For the collection of carriers one needs

$$\frac{x_F}{v} < \tau_p \ (n^+\text{-}p \ cell) \tag{7}$$

where x_F is the width of the field region and v is the drift velocity. The width of this region is about 2,000 Å in the conventional cell so that the drift transit time is on the order of 10^{-12} s. In addition, the lifetime in this region is rapidly increasing as the associated state density drops sharply with decreasing impurity concentration. The conditions for collection are easily met, particularly if we realize that at the lower impurity concentration the lifetime may be on the order of 10^{-6} s.

c. *Depletion Layer.* In this region the field is very large, the drift

transit is again on the order of 10^{-12} s, and the lifetime may be as long as 10^{-6} s. While the conditions for photocarrier collection are clearly met, it must be recognized that the dislocations always advance ahead of the diffusion profile. This means that there are more recombination states in the depletion layer than in the bulk. The effect of such states is to increase the space charge recombination current of the diode; while their effect on the short-circuit current is insignificant, the fill factor is reduced [6].

The model was tested experimentally by matching the predictable dead layer quantum yield behavior with actual measurements and by enforcing different surface recombination velocities externally. Let us first describe some of the studies related to the surface recombination velocity.

In many of the initial experiments an antireflective coating was grown by thermal oxidation of silicon. Thermally grown SiO_2 has an extremely good transparency (that of fused quartz); however, its index of refraction is low (n = 1.46) and therefore a 9- to 10-percent reflection remains at the quarter-wave minimum point. It is known from MOS field effect transistor studies that the surface state density changes with crystal orientation; it is highest on the <111> plane and lowest on the <100> plane. Surface recombination velocities as low as 100 cm/s have been reported for oxidized silicon surfaces. If surface recombination is an important minority carrier loss mechanism in the silicon cell, thermal oxidation and different surface treatments should affect the blue-violet response.

Figure 5 shows some results obtained on the <100> plane. This figure indicates that the thinned oxide results in about 10-percent reflection at the matching wavelength and an improved short wavelength response. After the SiO_2 was completely removed, the quantum yield dropped to a level controlled by the reflection coefficient of silicon. Exposure of the bare surface to moisture temporarily raised the quantum yield for all wavelengths with no change in the short wavelength characteristics.

Apparently moisture changes only the optical properties. Exposing the bare silicon surface to a variety of coatings caused no significant change in the short wavelength response. Similar results were obtained on the <111> plane. Because these results were obtained under very different surface conditions, it appears that the magnitude of the surface recombination velocity must have changed greatly without clearly affecting the violet response. Such behavior is expected from the dead layer model.

The quantum yield of a cell can be predicted readily from the dead layer model. We will make the simple assumption that all carriers generated beyond x_0 are collected, while those generated between 0 and x_0 are lost.

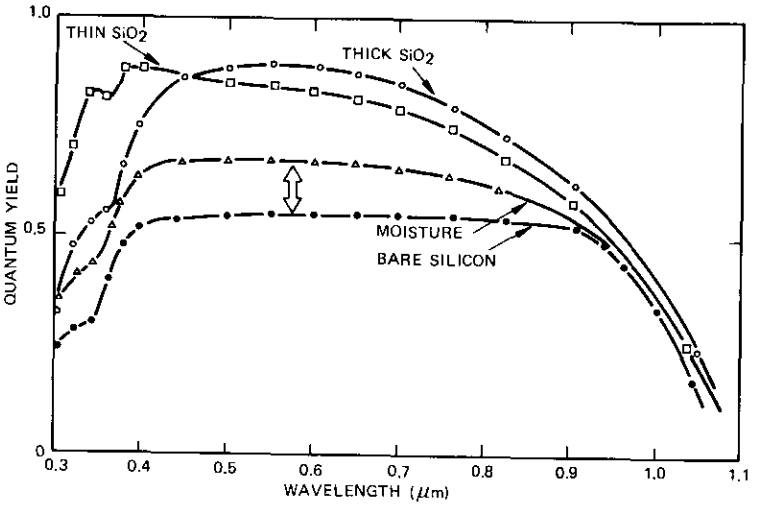


Figure 5. *Quantum Yield per Incident Photon for Two Thicknesses of Thermal SiO_2 and Bare Silicon Surface in Dry and Moist Ambients*

Then the yield for a particular wavelength is

$$Y = \frac{\int_{x_0}^{\infty} e^{-\alpha x}\, dx}{\int_{0}^{\infty} e^{-\alpha x}\, dx} = e^{-\alpha x_0} \tag{8}$$

where α is the absorption coefficient for that wavelength.

Figure 6 shows computed quantum yield curves for three different dead layer thicknesses (dashed lines) and the characteristics of a good conventional cell. The infrared response was computed by assuming that no recombination occurs at the back contact of the 200-μm cell. The curve associated with 1,500 Å of the dead layer was corrected for reflections arising from the antireflective coating. The solid line is the actual measurement of the cell, clearly indicating that the limited blue-violet response is the result of a dead layer thickness of 1,500 Å. It is interesting to note that fairly long wavelengths also suffer some losses. The SiO_x coating matched

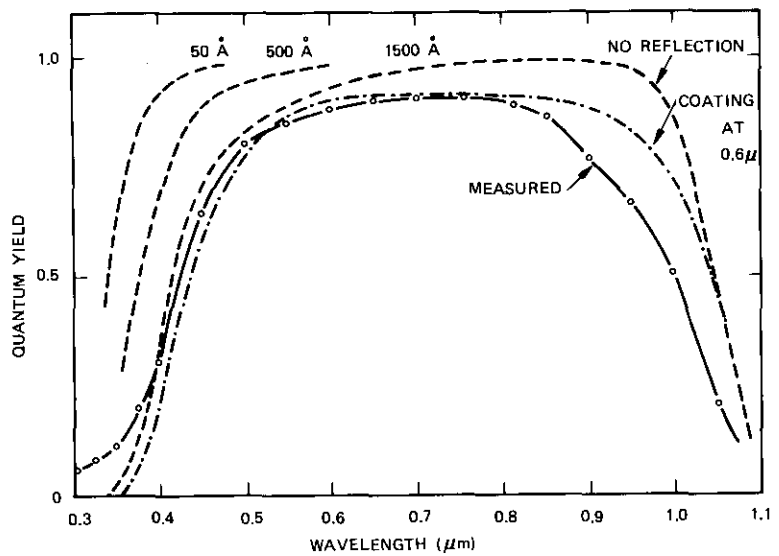


Figure 6. *Plot of Calculated Short Wavelength Response Using Dead Layer Model*

at 0.6 μm holds the quantum yield at about 0.9, even for longer wavelengths. Figure 6 also shows that extension of sensitivity into the short wavelengths (e.g., below 0.4 μm) requires a very shallow dead layer and, accordingly, a very shallow junction.

Figure 7 shows the allowable dead layer thickness for a given cutoff wavelength. The cutoff is defined at 0.71 collection efficiency and the absorption coefficients are those available from the literature [12],[13]. Figure 7 indicates that an extended short wave response requires a rapidly decreasing dead layer thickness or a very shallow junction.

Quantum yield of violet cell

When it became apparent that the short wavelength response could be extended significantly, the question of antireflective coating had to be reviewed in terms of blue-violet transparency. The widely used SiO_x coating is quite absorbent at short wavelengths. When $x \rightarrow 2$, it becomes less absorbent, but its index of reflection is low ($n \rightarrow 1.46$). On the other

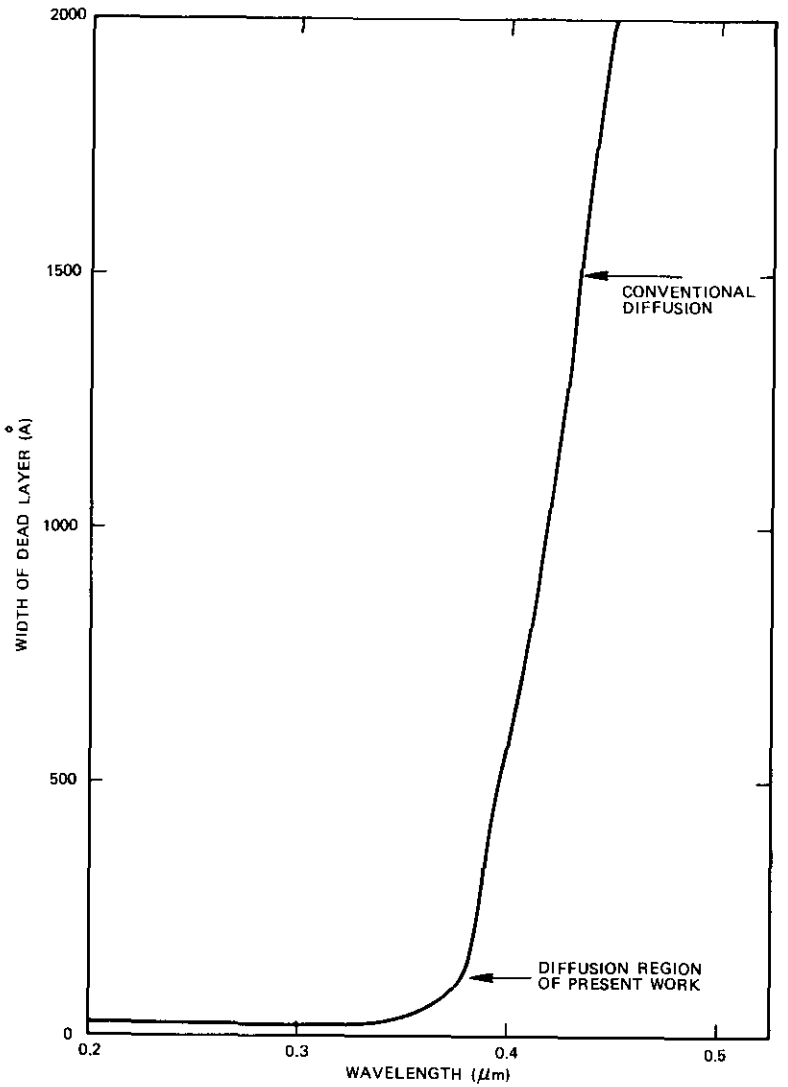


Figure 7. *Maximum Dead Layer Thickness as a Function of Cutoff Wavelength for 71-percent Response*

hand, its refractive index can be increased by decreasing x (as $x \rightarrow 1$, the index approaches 2, which is still a marginal match), but its absorption is significant [14], as shown in Figure 8. Although TiO_2 is a far better coating (it has a higher refractive index and less absorption), it also has a band gap of about 3.1 eV so that a sharp absorption sets in at about 0.4 μm, which limits quantum yield measurements below this wavelength. To provide the necessary bandpass, tantalum oxide was used on the violet cells reported in this work. The reflection plus transmission (R + T) of these three oxides is given in Figure 8, together with the reflection curve for tantalum oxide covered with quartz.

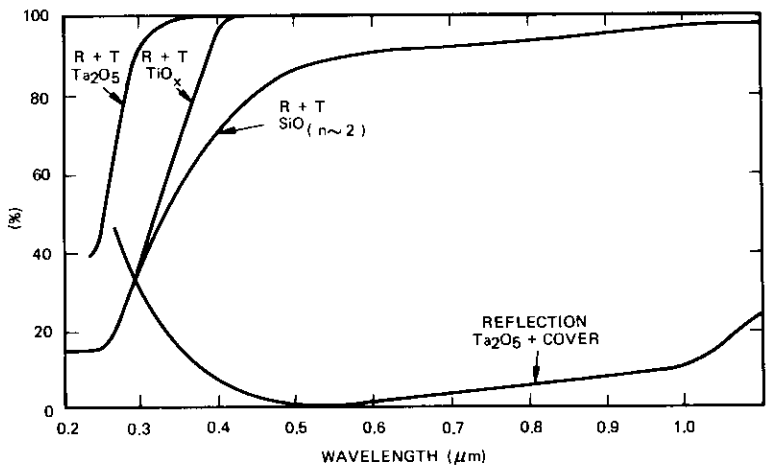


Figure 8. *Reflection Plus Transmission for Three Oxides*

The quantum yield of a typical violet cell is shown in Figure 9. The antireflective coating is matched around 0.5 μm, and as shown in the figure, the quantum yield is nearly equal to one in this region. A second curve representing the quantum yield after irradiation by 1-MeV electrons to 3×10^{14} e/cm^2 indicates that the losses are restricted to the longer wavelengths. The short-circuit current adds up to 160 mA (2-cm × 2-cm cell). The bars with points indicate the current not collected for each 0.05-μm segment of the solar spectrum. Most of the loss occurs in the red and infrared regions, portions of the spectrum which are most readily damaged by ionizing radiation.

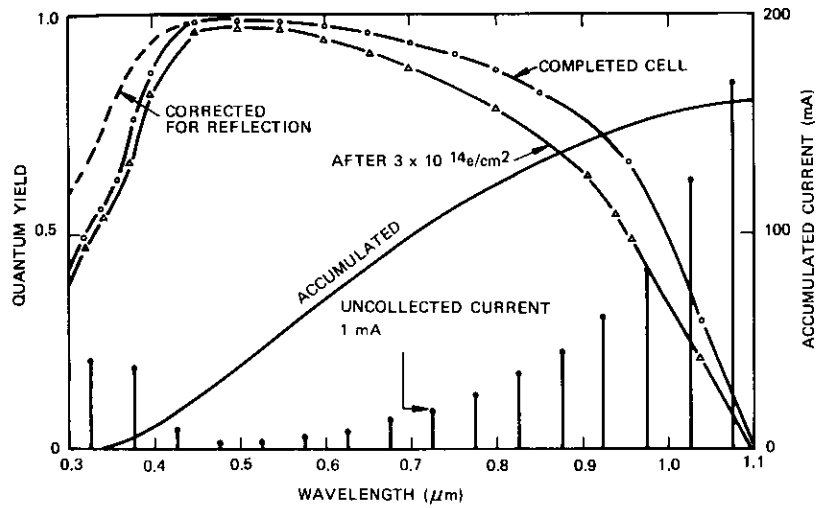


Figure 9. *Quantum Yield of Typical Violet Cell Before and After Irradiation by 3×10^{14} e/cm^2 and Accumulated Cell Current and Current Lost per 0.05 μm as Functions of Wavelength*

Characteristics of completed cell

The cell characteristics reported here relate to 2-cm × 2-cm cells. Figure 10 shows the current-voltage characteristics of a recent cell. The AM0 radiation efficiency of this cell is 13.5 percent (over 14 percent in terms of the active area). After irradiation by 1-MeV electrons to a fluence of 3×10^{14} e/cm^2, its actual efficiency is 11.5 percent. Such radiation affects the cell parameters as follows:

change in current: −10%

change in voltage: −5%

change in fill factor: 0%

The fill factor for such a cell is in the neighborhood of 80 percent and is somewhat dependent on the spectrum. It can be seen that direct illumination of the depletion layer by energetic photons (at AM0) tends to desensitize the recombination levels, thus reducing the space charge recombination current.

Figure 10 illustrates the improvement in the overall power output of an experimental violet cell (solid curves) when compared to a typical space-qualified commercial solar cell (dashed curves). The I-V curves were taken under illumination simulating AM0 conditions with a balloon flown cell used to standardize the simulator (Spectrolab X-25). The violet cell illuminated under terrestrial conditions exhibits an efficiency of 15.5–16 percent.

Figure 11 shows a series resistance of only 0.05 Ω deduced from I-V measurements at several intensities [15]. Figure 12 shows the efficiency as a function of increasing fluence of 1-MeV electron irradiation. The violet cell is again compared with the cell used widely in space applications and shows an improvement of over 30 percent for higher fluences of irradiation.

Conclusions

A major change in grid geometry, coupled with very shallow junctions, has produced a major improvement in the short wavelength response and the fill factor. Neither improvement is susceptible to degradation from ionizing radiation. The new cell is called the violet cell.

A new model has been developed for the front of such cells, explaining the short wavelength cutoff in terms of a dead layer. It has been shown that front surface recombination is not an important factor limiting the response. The improved diode characteristics (fill factor) have been explained qualitatively in terms of a critical integrated impurity concentration at the cell surface.

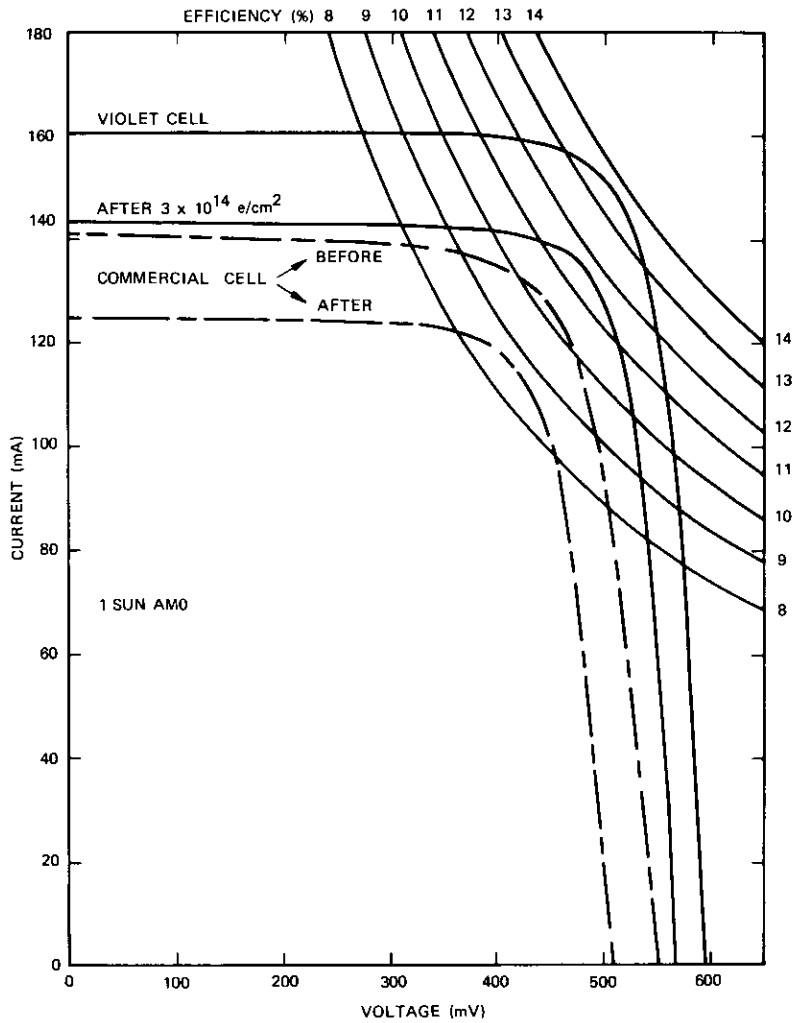


Figure 10. *Typical I-V Curves for a Violet Cell and a Commercial Cell Before and After Irradiation by 3 × 10¹⁴ e/cm²*

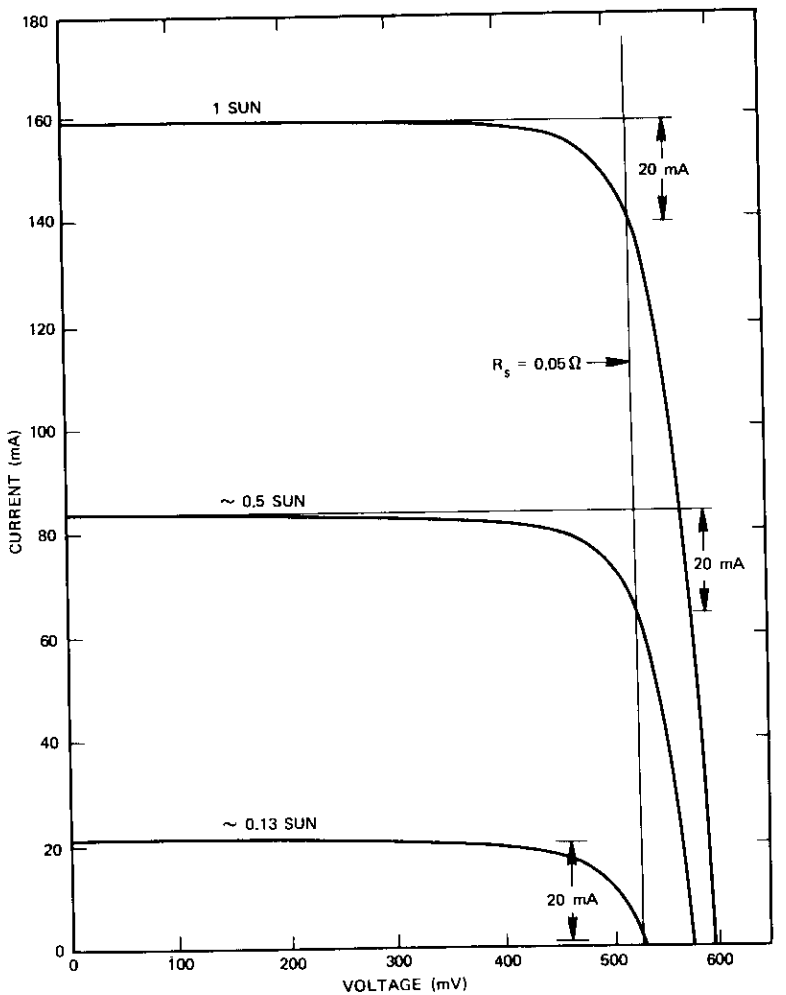


Figure 11. *Measurement of Total Series Resistance under AM0 Illumination Using Neutral Density Filters*

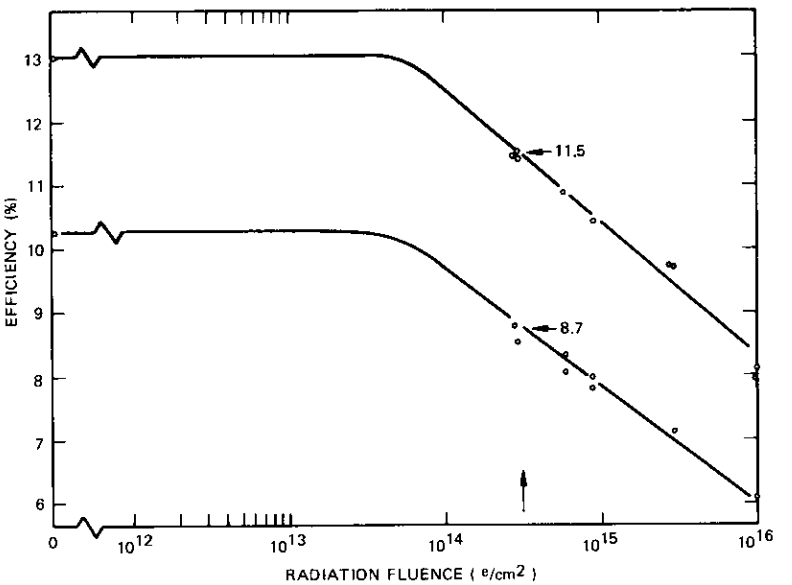


Figure 12. *Violet Cell Compared with Commercial Cell for Radiation Damage with 1-MeV Electrons*

Acknowledgments

This achievement could not have been accomplished without the invaluable assistance of many persons. The authors are indebted to A. G. Revesz and J. H. Reynolds for many helpful discussions and experiments. Special thanks are due to R. J. Dendall for contributions to the technology. The solar cells were fabricated with the assistance of F. Bland, A. Busch, and D. Martin, while the extensive measurements were carried out by I. Szabo. Appreciable contributions were made by C. Maag in terms of many discussions and measurements of the optical properties of the violet cell. The measurements of the impurity profile were done by M. Croset of Sescosem on the CAMECA ion probe, and the cell irradiations were carried out by A. Meulenberg. Special effort was expended by D. Curtin in providing an AM0 solar spectrum. Appreciation is due to E. S. Rittner for continuing discussions and encouragement, as well as to W. L. Pritchard, former Director of COMSAT Laboratories, for support throughout the research and development program.

References

[1] E. S. Rittner, "Use of p-n Junctions for Solar Energy Conversion," *Physical Review*, Vol. 96, No. 6, December 15, 1954, p. 1708.

[2] J. J. Loferski, "Theoretical Considerations Governing the Choice of the Optimum Semiconductor for Photovoltaic Solar Energy Conversion," *Journal of Applied Physics*, Vol. 27, 1956, pp. 777–784.

[3] P. Rappaport, "The Photovoltaic Effect and Its Utilization," *RCA Review*, Vol. 20, No. 3, September 1959, pp. 373–397.

[4] F. M. Smits, K. D. Smith, and W. L. Brown, "Solar Cells for Communication Satellites in Van Allen Belt," *Journal of the British Institute of Radio Engineers*, Vol. 22, No. 2, August 1961, pp. 161–169.

[5] M. Wolf, "A New Look at Silicon Solar Cell Performance," *Conference Record of the Eighth IEEE Photovoltaic Specialists Conference*, 1970, pp. 360–371.

[6] J. Lindmayer, "Theoretical and Practical Fill Factors in Solar Cells," *COMSAT Technical Review*, Vol. 2, No. 1, Spring 1972, pp. 105–121.

[7] J. C. C. Tsai, "Shallow Phosphorus Diffusion Profiles in Silicon," *Proc. IEEE*, Vol. 57, No. 9, September 1969, pp. 1499–1506.

[8] R. A. McDonald et al., "Control of Diffusion Induced Dislocations in Phosphorus Diffused Silicon," *Solid State Electronics*, Vol. 9, No. 8, August 1966, p. 807.

[9] H. J. Queisser, "Slip Patterns on Boron-Doped Silicon Surfaces," *Journal of Applied Physics*, Vol. 32, No. 9, 1961, pp. 1776–1780.

[10] S. Dash and M. L. Joshi, "Diffusion-Induced Defects and Diffusion Kinetics in Silicon," Soc. AIME, *Proc. Defects in Electronics Materials for Devices Conference*, Boston, Massachusetts, August 1969, pp. 202–222.

[11] C. B. Norris and J. F. Gibbons, "Measurement of High-Field Carrier Drift Velocities in Silicon via Time-of-Flight Technique," *IEEE Transactions on Electron Devices*, ED-14, No. 1, January 1967, pp. 38–43.

[12] W. C. Dash and R. Newman, "Intrinsic Optical Absorption in Single-Crystal Germanium and Silicon at 77°K and 300°K," *Physical Review*, Vol. 99, August 15, 1955, pp. 1151–1156.

[13] H. R. Philipp and E. A. Taft, "Optical Constants of Silicon in the Region of 1 to 10 eV," *Physical Review*, Vol. 120, October 1, 1960, pp. 37–38.

[14] A. P. Bradford et al., "Solar Absorptivity and Thermal Emissivity of Aluminum Coated with Silicon Oxide Films Prepared by Evaporation of Silicon Monoxide," *Applied Optics*, Vol. 9, No. 2, February 1970, pp. 339–344.

[15] R. J. Handy, "Theoretical Analysis of the Series Resistance of a Solar Cell," *Solid State Electronics*, Vol. 10, August 1967, pp. 765–775.

Addendum

Recent results indicate that the efficiency can be raised further. The actual efficiency is now 14 percent for outer space conditions (AM0), when a solar constant of 140 mW/cm^2 is used and when the actual area of the cell is taken into account. It must be pointed out, however, that at present there is no universally accepted quantity for the sun's radiant power or agreement on the cell area to be used in the computation of efficiency. Early measurements of the solar constant by F. S. Johnson [N1] resulted in a value of 139.5 mW/cm^2, while more recently M. P. Thekaekara [N2] has measured 135.3 mW/cm^2. Additional confusion is caused by the frequent use of only the exposed solar cell area in the computation of efficiency (active area). In view of this situation, our recent results are summarized as follows:

AM0 spectrum, 140 mW/cm^2, actual area (4 cm^2)	14.0%
AM0 spectrum, 140 mW/cm^2, active area	15.0%
AM0 spectrum, 135 mW/cm^2, active area	15.5%

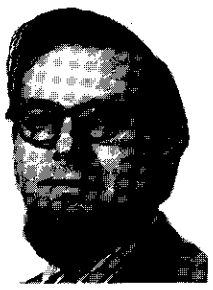
At sea level the solar spectrum is shifted and on a clear day the solar input power is approximately 100 mW/cm^2. Use of a pyrheliometer to measure the terrestrial solar input power resulted in a conversion efficiency of nearly 18 percent based on actual area and 19 percent based on active area.

References

[N1] F. S. Johnson, "The Solar Constant," *Journal of Meteorology*, Vol. II, No. 6, December 1954, pp. 431–439.

[N2] M. P. Thekaekara, "The Solar Constant and the Solar Spectrum Measured from a Research Aircraft," NASA-TR-R-351, October 1970.

Dr. Joseph Lindmayer was educated as an E. E. in Hungary. He received his M. S. degree in physics in 1963 at Williams College and his Ph.D. in Aachen, Germany, in 1968. He is Director of the Physics Lab and Acting Manager of the Solid State Physics Department at COMSAT Laboratories. He has contributed to the general field of solid state electronics and is the author or co-author of numerous scientific articles in this field, as well as a textbook, "Fundamentals of Semiconductor Devices."



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 Semiconductor Technology at COMSAT Laboratories.

Index: solar cells, communications satellites, aerospace environment, space environment simulation, radiation effects.

Influence of the space radiation environment on the Intelsat IV design

R. W. ROSTRON

Abstract

An extensive study was carried out to specify the anticipated radiation environment for INTELSAT IV and to predict the effects of this environment on silicon solar cells. The results of the study, which were derived from latest satellite and laboratory data, were presented to the spacecraft contractor in the form of a working engineering model designed for use in sizing the INTELSAT IV solar array, determining solar cell shielding requirements, and determining the shielding requirements for other radiation-sensitive electronic components. The model is presented here, both graphically and analytically, in the form of electron and proton fluences as functions of particle energy. Curves showing the equivalent 1.0-MeV electron fluences as functions of solar cell cover slide thickness and solar cell output as a function of time in orbit are also presented. In addition, other considerations, such as cover slide darkening, low-energy proton damage, and penetration of solar flare protons into the magnetosphere are discussed.

Introduction

The operational lifetime of an INTELSAT IV satellite is directly related to the ability of its photovoltaic prime-power source to endure the effects of

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

the space radiation environment. An accurate prediction of this environment is essential to the satellite designer for sizing the solar array and predicting array performance. Until the launch of ATS-1 into the geostationary orbit late in 1966, data depicting the synchronous orbit radiation environment were sparse and unreliable. This uncertainty resulted in the incorporation of extremely high safety margins into the design of solar array power sources to allow for radiation degradation. While these designs were unduly conservative in many cases, they proved to be inadequate in other instances because of the lack of information regarding the geostationary environment.

In formulating the performance specifications for INTELSAT IV, COMSAT presented the spacecraft designer with a working engineering model of the geostationary radiation environment. This model, incorporating the results of an in-house analysis of ATS-1 and other data, is presented in the form of curves and analytical approximations of the time-integrated radiation fluxes to be encountered by INTELSAT IV. Emphasis was placed on the solar flare and trapped electron and proton fluences encountered in synchronous orbit. This paper also delineates the effects of the environment on solar cells of the type utilized on INTELSAT IV.

Intelsat IV radiation environment

In the design of INTELSAT IV, emphasis was placed on assessing the hostile proton and electron environments at synchronous altitude, since these particles will inflict nearly all of the radiation damage to be sustained by INTELSAT IV. Other forms of radiation exist at synchronous altitude, but they were found to have negligible effects on satellite performance.

Trapped electrons

The power output of silicon solar cells, which are the INTELSAT IV prime-power source, suffers a substantial degradation because these cells are continually bombarded in orbit by Van Allen electrons. Data concerning the energy and intensity of these trapped electrons at synchronous altitude have been collected by detectors aboard ATS-1 and are shown in Figure 1. The solid curve (labeled electrons) depicts the time-averaged integral electron flux as a function of electron energy, while the broken curve represents the softer spectrum obtained by using earlier data [1].

The ATS-1 data have been curve fitted by two approximating expressions; one follows a power law and the other is exponential:

$$0.05 \leq E_e \leq 0.5$$

$$\Phi_e(>E_e) = 7.96 \times 10^5 \, E_e^{-1.56} \tag{1a}$$

$$0.5 \leq E_e$$

$$\Phi_e(>E_e) = 1.00 \times 10^7 \, exp\,(-2.94\, E_e) \tag{1b}$$

where E_e = electron energy, in MeV

$\Phi_e(>E_e)$ = electron flux, in electrons/(cm^2·s), with energy greater than E (integral flux).

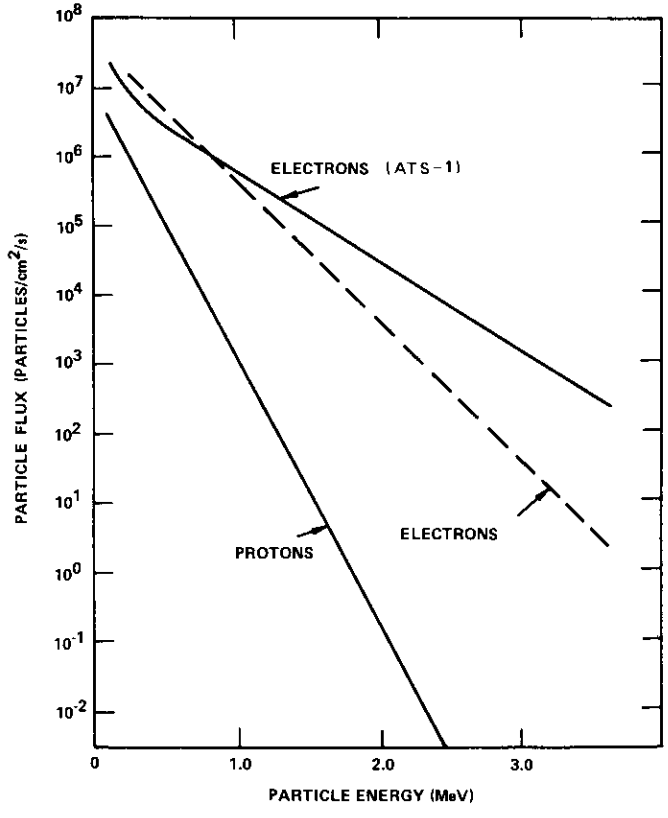


Figure 1. *Time-Averaged Trapped Flux at Synchronous Altitude*

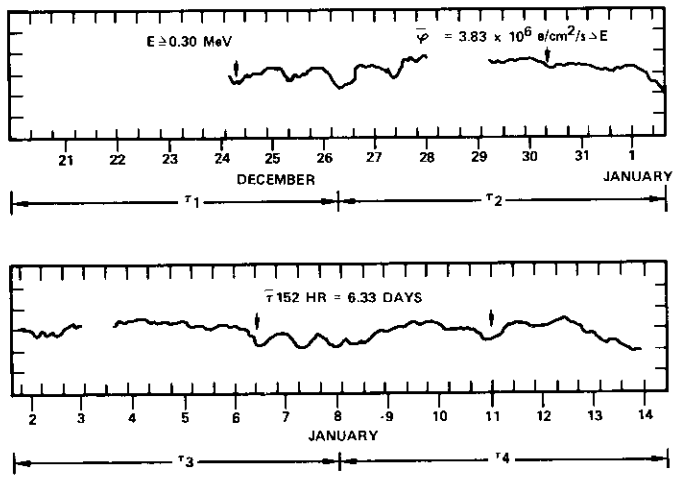


Figure 2. *Time Variation of Electron Flux at Synchronous Altitude*

The analysis of the ATS-1 data indicated that there is a periodic variation of electron flux intensity. This variation is shown in Figure 2. It can be seen that the electron flux remains relatively constant for 6.33 days and then drops several orders of magnitude.

Trapped protons

Since the intensity of the trapped proton flux at synchronous altitude is negligible for energies greater than 1.0 MeV, the solar cells can be shielded from damaging effects of these particles by 25 to 50 μm of glass. However, the INTELSAT II F-4 and ATS-1 solar cells were not completely covered by the cover glass material. The resulting gaps allowed low-energy (0- to 1.0-MeV) protons to enter the surface and junction regions of the solar cells and produce severe degradation of the power output. Various methods of fully shielding solar cells have now been implemented. The synchronous proton environment data required to accomplish this implementation are presented here.

The time-averaged trapped or Van Allen proton flux over the energy range from 0.1 to 4.0 MeV is shown in Figure 1 [2] and may be fitted by the following exponential approximating equation:

$$0.1 \leq E_p \leq 4.0$$

$$\Phi_p(>E_p) = 1.80 \times 10^7 \, exp\,(-9.0\, E_p) \qquad (2)$$

where E_p = proton energy, in MeV

$\Phi_p(>E_p)$ = proton flux, in protons/(cm²·s), with energy greater than E_p.

Solar flare protons

Although many theories have been advanced to predict solar flare activity, it is still virtually impossible to accurately predict the total solar flare proton fluence to be encountered by INTELSAT IV. In the absence of reliable prediction techniques, it becomes advantageous to examine solar activity over previous cycles and to correlate this information with current solar flare activity data.

The last complete solar cycle (cycle 19), which was the first cycle during which solar flare protons were monitored by satellite-borne detectors, began in April 1954 and ended in October 1964. Data from satellite measurements of solar flare activity during cycle 19, as well as rocket and balloon data, were collected and analyzed at COMSAT Labs to yield the total solar flare proton fluence for cycle 19 as a function of proton energy. These data, which are plotted in Figure 3, may be curve fitted by the following expression:

$$\Phi_s(>E_s) = 1.50 \times 10^{12}\, E^{-1.53} \qquad (3)$$

where E_s = solar proton energy, in MeV

$\Phi_s(>E_s)$ = integral solar flare proton fluence, in protons/cm².

The year of maximum solar proton intensity during cycle 19 was 1959. Approximately 87 percent of the total number of protons were encountered during that year. The fluence data for 1959 are also shown in Figure 3 and may be curve fitted by the expression

$$\Phi_s(>E_s) = 1.30 \times 10^{12}\, E^{-1.73}. \qquad (4)$$

Prior to the launch of ATS-1 in late 1966, it was assumed that the geomagnetic field would prevent solar protons with energies below about 30 MeV from penetrating to synchronous altitude. This assumption was

reflected in the design of satellites such as ATS-1 and INTELSAT I and II. However, ATS-1 data indicated that solar flare protons of all energies could arrive at synchronous altitude and that the low-energy protons appeared to be trapped. Thus, the full spectrum shown in Figure 3 was assumed as a design guideline for INTELSAT IV.

The solar flare proton fluence spectrum through 1968 (cycle 20), when design criteria for INTELSAT IV had to be determined, is also shown in Figure 3. Data from cycle 20 indicated that the proton spectrum would be softer; however, only one conclusion concerning its intensity was reached, i.e., that the intensity of cycle 20 should not exceed the intensity of cycle 19. Thus, the cycle 19 solar proton spectrum was used in the design of INTELSAT IV.

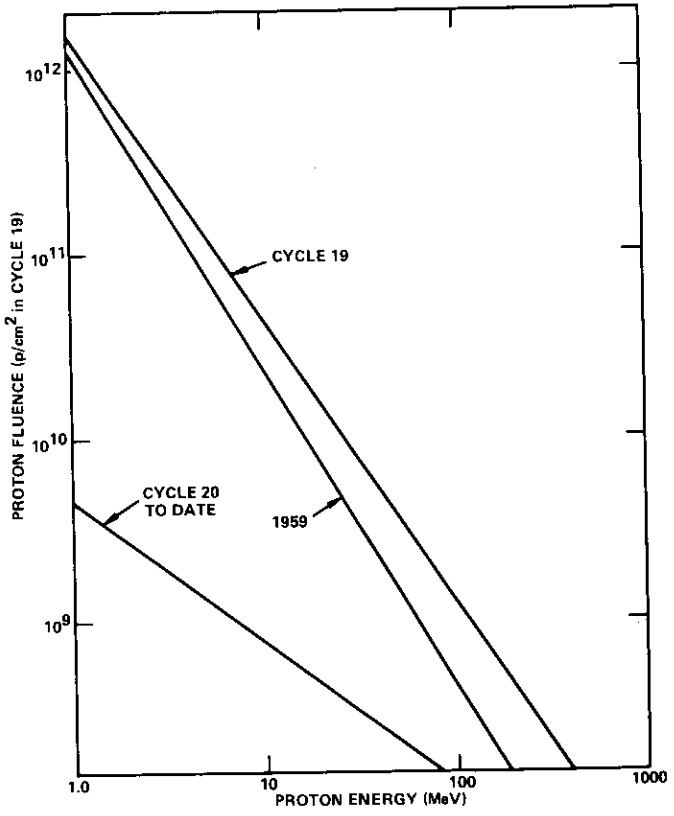


Figure 3. *Unattenuated Solar Proton Fluence*

Other radiation

The integral flux of Van Allen electrons encountered by INTELSAT IV in the Hohmann transfer orbit is several orders of magnitude lower than it is in synchronous orbit. For this reason, the effects of these electrons on INTELSAT IV were considered negligible.

Solar cells are readily shielded from the most damaging trapped protons encountered by INTELSAT IV in the Hohmann transfer ellipse by 25 to 50 μm of cover glass. Thus, the effects of those protons which do reach the solar cells are negligible compared to the effects of the trapped electrons and solar flare protons encountered at synchronous altitude. The only hazardous effect of lower energy protons on solar cells, either at synchronous altitude or in the transfer orbit, is surface and junction damage to cells which are not fully covered.

Ultraviolet radiation incident on INTELSAT IV can produce color changes in thermal coatings and darkening of solar cell cover assemblies, thus reducing their efficiency. The Johnson curves [3] for solar ultraviolet radiation incident on the earth's upper atmosphere were used to select ultraviolet-resistant thermal coatings and cover assemblies for INTELSAT IV.

Radiation effects

Silicon solar cells, which constitute the prime-power source for INTELSAT IV satellites, and certain MOS devices used in the telemetry and command systems are the spacecraft components which are most susceptible to performance degradation through exposure to the space radiation environment. The least penetrating charged particle radiation, composed of low-energy protons, constitutes the greatest hazard to solar cell performance. However, the higher energy protons and electrons also produce significant degradation of solar cell output and can degrade MOS device performance. Fortunately, glass shields can be used to prevent low-energy protons from reaching the solar cells, and the MOS devices can be shielded from penetrating charged-particle radiation by metal housings.

Solar cell degradation

Solar cells exposed to charged-particle radiation in space will suffer performance degradation caused primarily by two mechanisms. The first is produced by low-energy, high-mass charged particles such as protons, which just penetrate the surface of the solar cell, but do not penetrate into the base region.* These particles produce generation-recombination centers

* It is assumed that protons are fully shielded from the rear of the cell. If this were not the case, another type of damage would materialize.

close to the junction which cause enhanced thermal generation of carriers and increased leakage current, thus drastically reducing the output voltage.

Other secondary effects may also be produced by these low-energy particles. It is these effects which can be prevented by shielding the solar cell with 25 to 50 μm of cover glass. However, care must be taken to fully cover the photosensitive surface of the solar cell. In most cases the contact bar must also be covered, since the contact thickness may be less than 25 μm and therefore allow charged particle penetration to the solar cell junction beneath the contact bar. The decrease in output voltage, and thus power output, is a nonlinear function of exposed surface area; hence, a small uncovered area may result in a large decrease in output power [4].

The second damage mechanism which cannot be fully eliminated by shielding (because of spacecraft weight considerations) is a decrease in the base region minority carrier lifetime. This is a result of displacements, caused by the penetrating charged-particle radiation (protons and electrons), which lead to an increase in recombination center density. Thus, the carriers produced by the light entering the cell are less likely to reach the junction before recombination occurs. This results in a reduction in the current output of the cell and, to a lesser degree, in the output voltage.

This effect can be (and is) reduced substantially by glass cover shields, but there is a tradeoff between the thickness of the shield and the spacecraft weight allowance and power requirements. To perform this tradeoff for INTELSAT IV, it was necessary to predict the solar cell performance degradation caused by penetrating charged-particle radiation over the lifetime of the mission.

It is possible to reduce any space spectrum of protons and electrons to an equivalent 1.0-MeV electron spectrum. This equivalency relates the solar cell damage caused by a given spectrum of penetrating charged particles to that produced by a certain equivalent number of 1.0-MeV electrons. Figure 4 is a curve showing this equivalency for various charged-particle environments encountered by INTELSAT IV as a function of cover slide thickness. (Because the equivalency of trapped protons at synchronous orbit and electrons at transfer orbit is negligible, it is not presented.) From Figure 5 it can be seen that, after seven years in orbit, INTELSAT IV solar cells which encounter a solar flare cycle equivalent to that of cycle 19 will be exposed to an equivalent 1.0-MeV fluence of about 3.0×10^{14} e/cm^2 if 12-mil cover slides are used.

The degradation of solar cells is predicted by using the curves of Figure 4. If the INTELSAT IV cell mentioned previously were irradiated without a

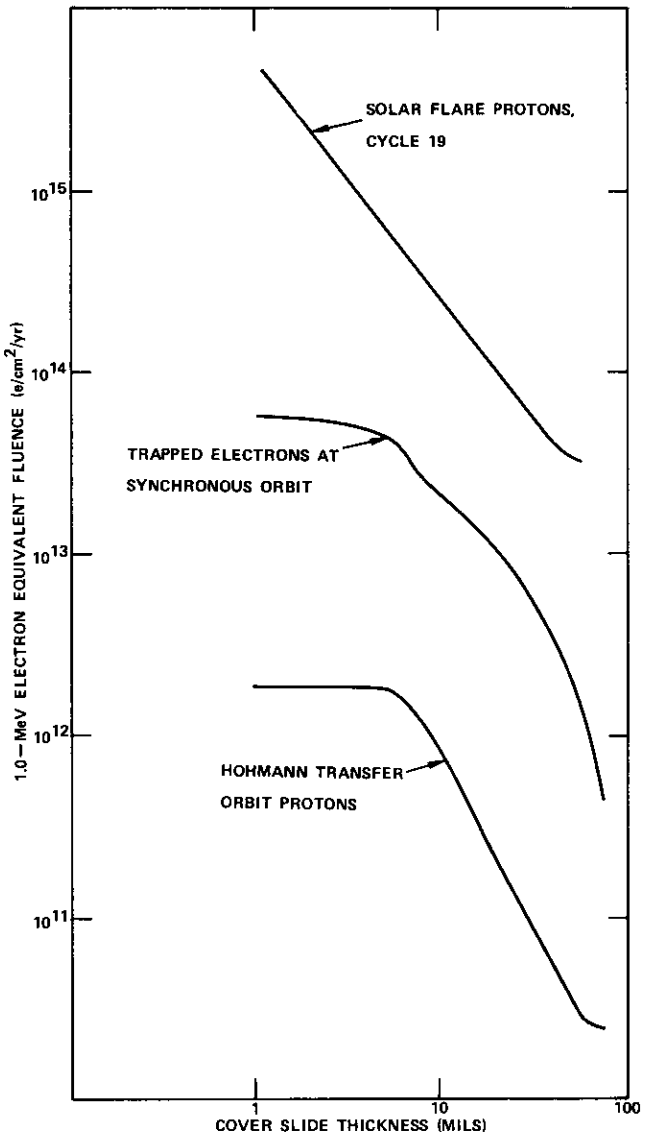


Figure 4. *Equivalent 1.0-MeV Fluences*

cover slide by a perpendicular beam of 1.0-MeV electrons of 3.0×10^{14} e/cm^2, the damage would be the same as the cell would sustain with a 12-mil cover slide after seven years at synchronous altitude. Laboratory data on silicon solar cells are shown in Figure 5, where solar cell outputs are plotted as a function of 1.0-MeV electron fluence. From these curves it is found that a 1.0-MeV fluence of 3.0×10^{14} e/cm^2 would reduce the open-circuit voltage (V_o) by about 7 percent, the short-circuit current (I_o) by about 12 percent, and the maximum power (P_o) by about 15 percent. These values were then used to size the solar array of INTELSAT IV.

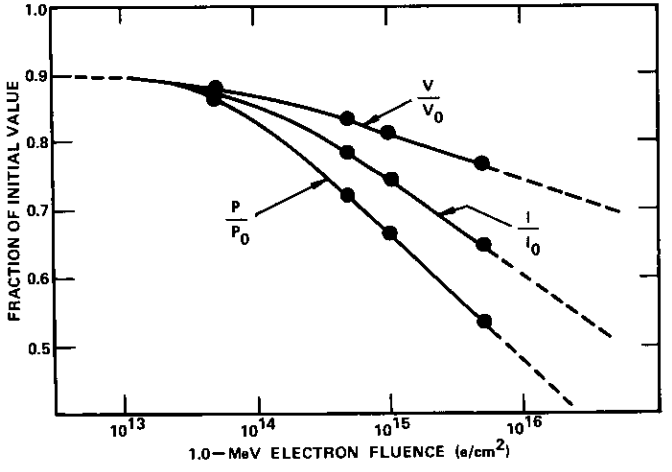


Figure 5. *Silicon Solar Cell Output Degradation*

Solar cell cover assemblies

When specifications were written for INTELSAT IV, it was known from ATS-1 data that the current and power outputs of solar cells could suffer a severe degradation caused by what appeared to be a "darkening" of the cover slide and/or cover slide adhesive. Although this effect was observed in space, it could not be duplicated in the laboratory. It was surmised that this effect might be caused by the synergistic effects of ultraviolet and charged-particle radiation. A number of other theories were advanced and laboratory tests were initiated by COMSAT Labs and others to try to reproduce the "darkening." These tests are still under way and the problem

is not yet solved. For design purposes, based on ATS-1 data, INTELSAT IV solar cells were assumed to experience a 7.5-percent power degradation because of the "darkening" effect.

Other components

Many INTELSAT IV semiconductor devices other than solar cells would suffer performance degradation if exposed to the radiation environment at synchronous altitude for prolonged periods. Adequate shielding of most of these components is provided by the spacecraft structure and component housings. In the case of MOS devices on INTELSAT IV, however, additional shielding was required. Figure 6 shows the radiation dose

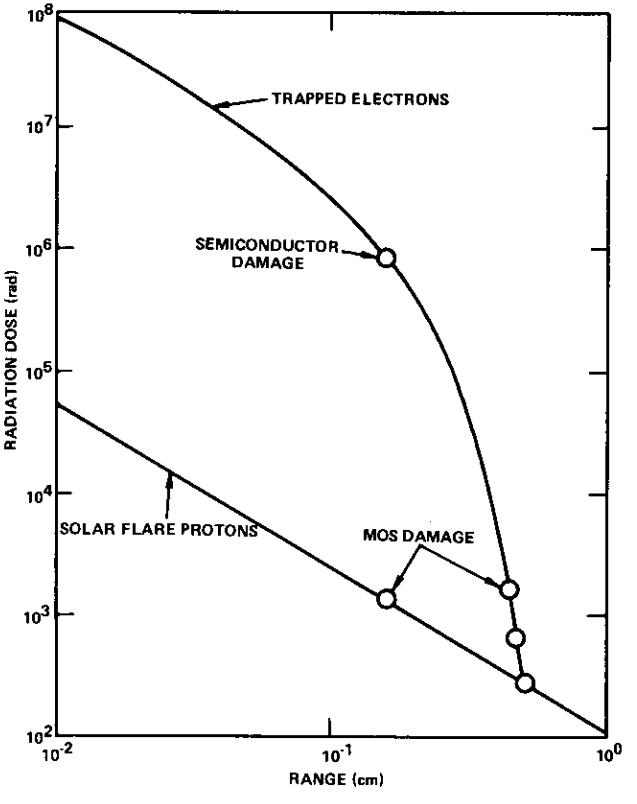


Figure 6. *Internal Component Dosage*

received by INTELSAT IV components as a function of shield thickness. It can be seen that the dose resulting from solar flare protons is negligible compared to that sustained from electrons.

Since most semiconductor components will operate satisfactorily at accumulated doses of 10^6 rad (Si), a nominal 1/16-inch aluminum housing or its equivalent will produce adequate shielding for the INTELSAT IV mission. However, the MOS devices may malfunction after accumulating a dose of 10^3 rad (Si) and require about 3/16 inch of shielding. Such shielding requirements present no particular design problem, but must be taken into account.

References

[1] J. I. Vette, "Models of the Trapped Radiation Environment," NASA AP-3024, 1966.

[2] J. H. King, "Models of the Trapped Radiation Environment," NASA AP-3024, 1967.

[3] F. S. Johnson, *Satellite Environment*, Stanford, Calif.: Stanford University Press, 1965, pp. 97–100.

[4] R. W. Rostron, "Effects of 10- to 150-KeV Proton Bombardment on Silicon Solar Cells," *Energy Conversion*, Vol. 12, No. 4, 1972, pp. 125–129.

Robert W. Rostron received B.S.E.E. and B.S.E.P. degrees from Washington University, St. Louis (1959), and a Ph.D. from the University of Arizona (1964). Before coming to COMSAT Laboratories in 1967 to establish the Space Physics Branch, he was Assistant Professor of Electrical Engineering at St. Louis University and Consultant to the Manager of the Advanced Electronics Systems Department of McDonnell-Douglas Corporation. Dr. Rostron is now Senior Scientist and Manager of the Satellite Projects Department of the Special Projects Division of COMSAT Laboratories.

Index: tunnel diode, electron probe, electron microscopy, stereoscopy, microanalysis, X-ray spectroscopy.

Physical and chemical analysis of germanium tunnel diodes

P. F. VÁRADI AND T. D. KIRKENDALL

Abstract

Scanning Electron Microscope (SEM) and Electron Microprobe (EM) studies were performed on commercially available germanium tunnel diodes of the type used in communications satellite receivers to delineate the physical structure and spatial distribution of the constituent chemical elements. The SEM images revealed the mushroom-like structural characteristic of the ball alloy and pinnacle enclosing the n-p junction. Examination of the physical structure of several diodes showed a large variation from diode to diode caused by the fabrication procedure. These findings may help to illuminate the wide variation in the electrical characteristics and reliability of the unscreened device.

Chemical composition studies were conducted in which electron probe techniques were used to measure the elemental distributions of Sn, As, Ge, Ga, and Ni in cross-sections of several tunnel diodes. Quantitative analyses of the As and Ga dopants revealed that, while the Ga dopant in the p-type Ge was evenly distributed, the As was segregated in arsenic-rich regions in the Sn "ball" after the alloying. The use of a computerized electron microprobe made it possible to delineate the position of the n-p junction through direct "chemical" analysis and to demonstrate that it typically lies a few micrometers beyond the metallurgical junction on the Ge side.

Introduction

Ball-alloy-type germanium tunnel diodes have been used in the 6-GHz and the 4-GHz amplifiers of INTELSAT III and IV satellites, as well as in

the 4-GHz amplifier of the transponder destined for the ATS-F Propagation Experiment. Since the reliability of a satellite system depends to a large extent on the satisfactory performance of these tiny structures, a study of their physical and chemical nature was undertaken. The physical dimension of the germanium tunnel diode permits the entire structure to be set into a cylindrical enclosure which is 0.13 cm (0.05 in.) in diameter and 0.13 cm high. The cross-sectional diameter of the actual device may vary from a minimum of about 0.0005 cm (0.0002 in.) to a maximum of 0.025 cm (0.010 in.).

Several years ago when these devices were first used, it was attempted to view them with a high-powered optical microscope. This required sectioning and polishing of the device because of the low depth of field of such an instrument; hence, the study revealed only some basic information about the size and shape of the tunnel diode without producing an actual picture. With the advent of the scanning electron microscope, which has a very high depth of field and a variable magnification between 10,000 and 50,000X, it became feasible to obtain a picture of the tunnel diode [1]. Correlation of the physical structure of the tunnel diode with the spatial distribution of the constituent chemical elements on the same scale became possible only with the recent development of the combination of a scanning electron microscope and an electron microprobe (SEM/EM)*, however. This paper reports a study of ball alloy germanium tunnel diodes in which combined SEM/EM techniques were used to obtain a highly magnified picture of the device and also to analyze its elemental composition.

Tunnel diode fabrication technology

The assembly and fabrication of tunnel diodes also became a part of this study, since they are closely related to the structure and chemistry of the final device. Figure 1 is a schematic of a tunnel diode structure. The major components are a p-type germanium chip doped with Ga, a tin ball doped with As to provide the n-type impurity in the Ge needed to form the n-p junction, the connecting electrode (wire mesh), and the enclosure.

* The Scanning Electron Microscope/Electron Microprobe employs a finely focused electron beam which is directed onto the object to be analyzed. The interaction of the high-voltage electrons with the sample creates several effects: it generates secondary (low-energy) and backscattered (high-energy) electrons, as well as X-rays which are characteristic of each of the elements present in the sample.

If the electron beam is held stationary, the generated X-rays can be utilized to perform a qualitative and quantitative chemical analysis of a spot with a diameter of about 1 μm. Scanning the electron beam in a raster fashion and displaying the synchronized X-ray output on an oscilloscope makes it possible to display the spatial distribution of the elements. Similarly, the secondary and/or backscattered electron output signals may be displayed to create a quasi-optical image showing the topography of the sample.

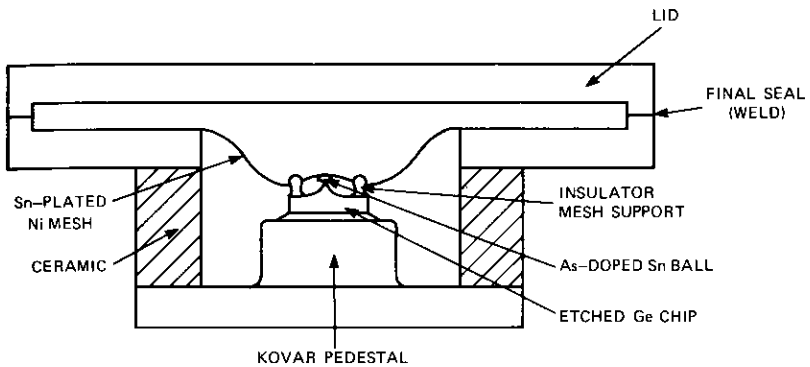


Figure 1. *Schematic Cross-section of Ball Alloy Ge Tunnel Diode*

The first step in the fabrication process is an alloying step which consists of heating the Ge chip while in contact with the Sn ball to a temperature of 600°C–800°C (1090°F–1454.5°F), well above the melting point of Sn. The temperature chosen, its duration, and the rate of ensuing cooling determine the concentration profile of the arsenic in the Ge and the location and sharpness of the resulting n-p junction on which the diode characteristics are critically dependent.

The tin ball utilized in this process must be prepared so that the arsenic and tin are uniformly mixed. (The arsenic concentration in the tin is 2 percent or higher.) This uniform distribution cannot be achieved by a regular alloying process, since the arsenic will precipitate in the form of an Sn_3As_2 phase [2]. The SEM/EM technique was used to verify the uniform distribution of arsenic in an actual tin ball before its utilization.

Figure 2 is an SEM/EM picture of an As-doped Sn ball. Figure 2a is a BSE (backscattered electron) picture which gives the general shape and dimensions of the ball. Figures 2b and 2c, which are EM pictures representing the tin and arsenic distributions, show that these two elements are uniformly mixed on a micrometer scale.

After the alloying process, the elements are no longer uniformly distributed. Figure 3a is a BSE image of a Ge chip and the solidly attached Sn

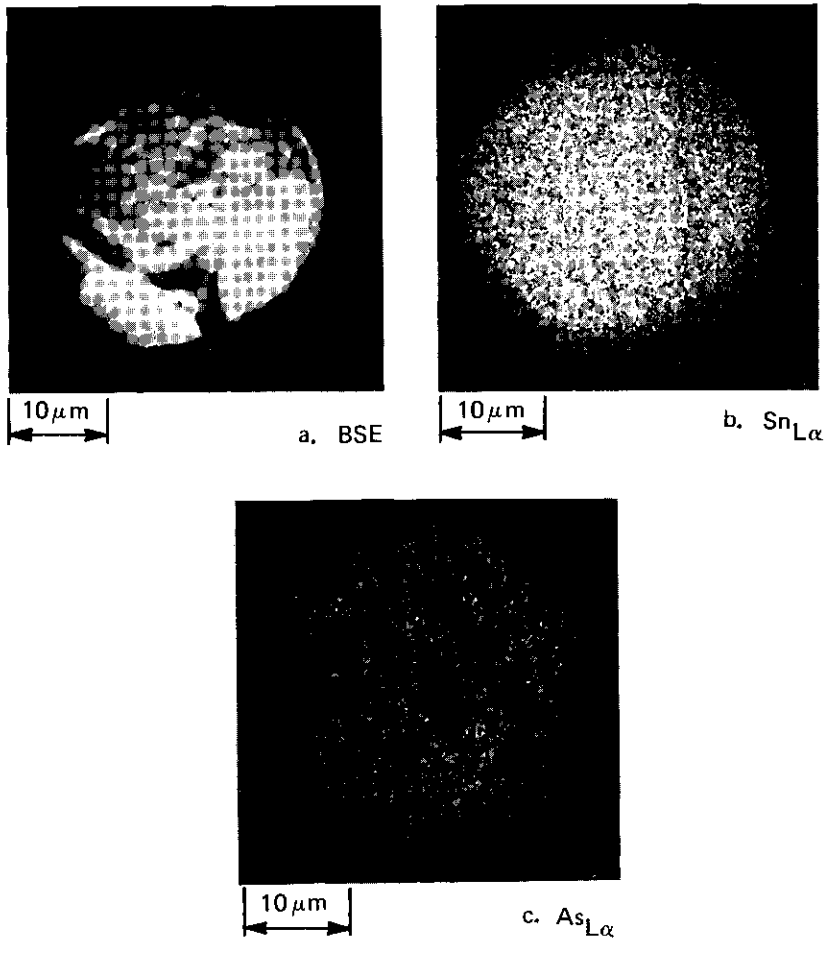


Figure 2. *Sn-Doped As Ball Before Alloying*

ball after completion of the alloying process. Figures 3b, 3c, and 3d are a closeup BSE picture of the Sn ball and EM pictures of the Sn and As distributions, respectively. The majority of the As is concentrated in only a few spots. This segregation of the As after the heat cycle can be explained on the basis of the known phase diagram [2] of the Sn/As binary alloy system.

In the subsequent assembly steps, the alloyed chip is mounted in the enclosure via a tin soldering process, the center of the mesh strip electrode is soldered to the tin ball, and the ends are welded to the enclosure. In these soldering steps, it is necessary to ensure that the entire tin ball does not melt again, thereby causing changes in the established junction. The importance of properly attaching the mesh to the tin ball will be discussed further in the following section.

The last step before sealing the enclosure is an etching process. In this operation the Ge is gradually dissolved in KOH, which does not attack the Sn. Hence, a very small connection is left between the Ge body and the Sn ball, thereby reducing the capacitance of the n-p junction. The etching is stopped when the desired peak current and capacitance characteristics are reached. At this point the device is rinsed, dried, and sealed.

SEM observations

To obtain a picture of the physical structure of the diodes, devices were examined after their electrical characteristics had stabilized. It was necessary to open the sealed tunnel diodes by carefully lapping off the Kovar base plate. When the Kovar-ceramic seal was weakened and the ends of the wire mesh carefully severed from the enclosure, the entire device could be removed intact, ready for mounting on the tilting stage of the SEM/EM instrument. The SEM/EM instrument utilized for the investigation was an Applied Research Laboratory EMX/SM, which was implemented with computer control at COMSAT Laboratories [3].

To obtain a better understanding of the device, most of the SEM pictures have been taken in a stereoscopic mode. The stereoscopic effect is achieved by taking two pictures of the device so that the second picture is taken with the specimen tilted between 4° and 8° relative to the first one, referenced to the electron beam axis. To provide an example of the actual shape and size of a ball alloy tunnel diode, Plates I and II are presented in a stereoscopic mode. The two pictures obtained by the SEM in the secondary electron (SE) mode are separately printed in two colors on one

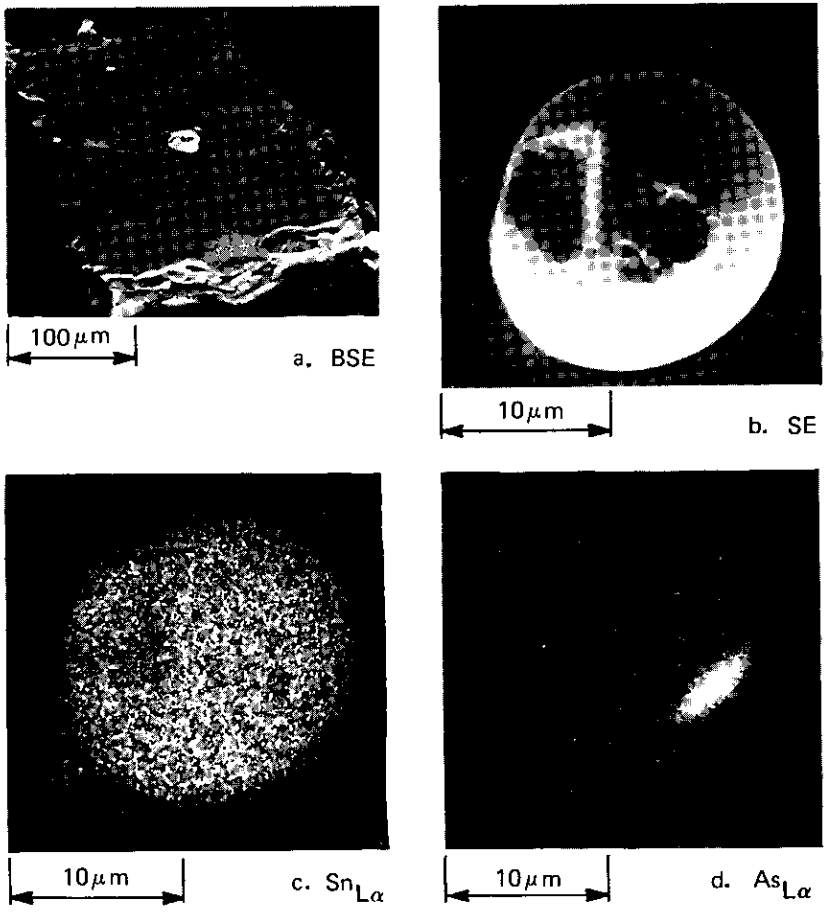


Figure 3. *Sn-Doped As Ball Alloyed on Ge Chip*

sheet. The stereoscopic impression of the SEM picture may then be perceived by viewing the picture with colored "glasses", each eyepiece of which allows perception of only one image.

Plate I is a side view of the tunnel diode. The dimensions may be measured by using the attached scale. For this type of diode, designed to operate at 6 GHz, the diameter at the junction is only 5 μm. Plate II is a stereoscopic view of the same tunnel diode viewed from the top. This picture provides insight into the entire tunnel diode structure.

The smooth etched Ge chip can be seen through the Sn-plated Ni mesh electrode to which the top of the Sn ball (or rather "mushroom cap") is soldered. The plastic material holding the grid is just outside of the picture. The Sn solder holding the chip to the base and the rough surface of the Kovar can be easily distinguished in the background. These pictures indicate the large depth of field of SEM pictures.

Observation of these stereoscopic pictures will provide a better understanding of Figure 4, which compares the SE images of four different and electrically satisfactory tunnel diodes. Figure 4a shows the same diode which was depicted in Plates I and II.

Figures 4b, 4c, and 4d are different samples, all designed to operate at 4 GHz. There are large variations in the shape of the pinnacle and in the results of soldering the mushroom cap to the mesh electrode. Figure 4a shows a mushroom-shaped tunnel diode. However, this particular diode does not represent the best possible execution, since the mesh is not solidly soldered to the mushroom ball, but is making contact at only a few points. This is better observed in the stereo images. Examples of intimate and mechanically secure mesh-to-ball connections are shown in Figures 4b, 4c, and 4d. Figure 4c shows the extreme case in which the mesh electrode was pressed almost entirely through the tin ball during the soldering cycle. (The debris seen in Figure 4c are a result of particles trapped during the opening of the device.)

The variations in the shape of the pinnacle are also interesting. While the pinnacles shown in Figures 4a and 4b are mushroom-type pinnacles, Figure 4c shows a wide, stubby pinnacle, and the pinnacle shown in Figure 4d is cylindrical. (A stereo view of this diode establishes that the protrusion seen above the mesh is not connected to the diode.) The differences in the mechanical structure of the tunnel diode and in particular the very small area of that portion of the pinnacle which supports the mushroom cap corroborate the theory [4] that failure in the germanium tunnel diodes is primarily a result of plastic deformation, caused by inherent stresses in the

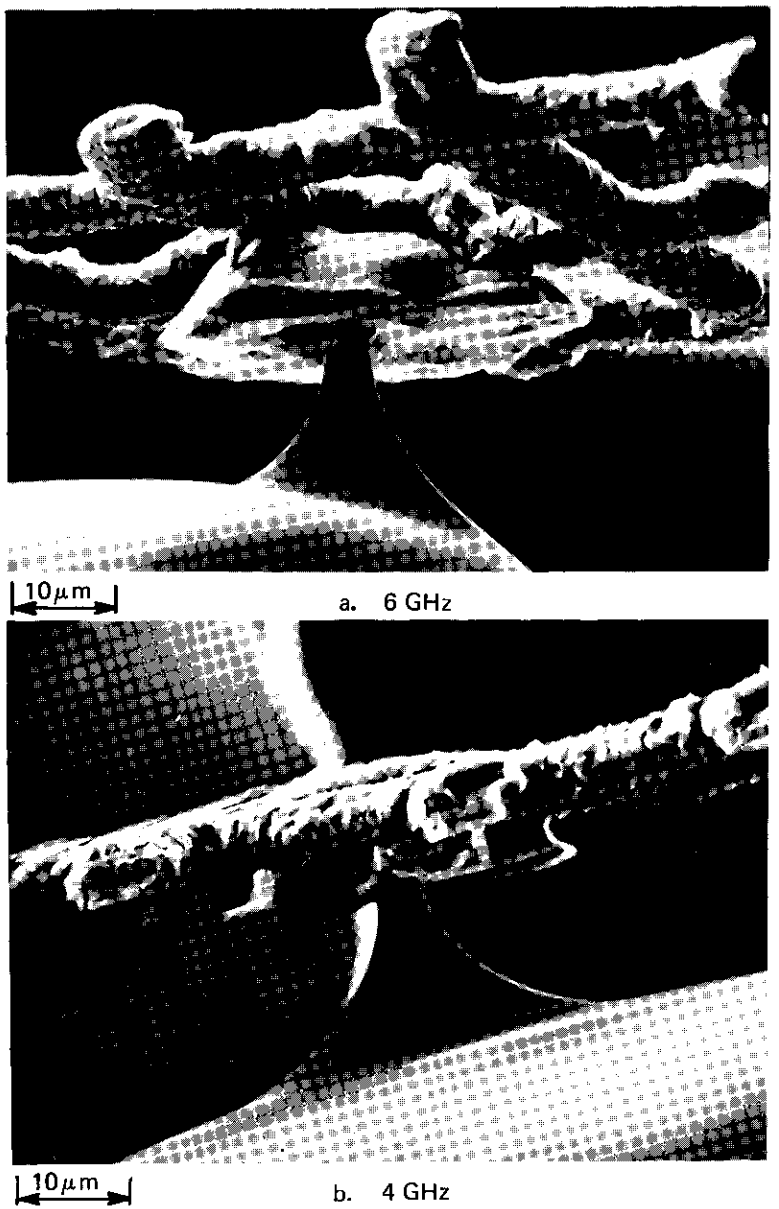


Figure 4. *SE Side View Images of Four Tunnel Diodes*

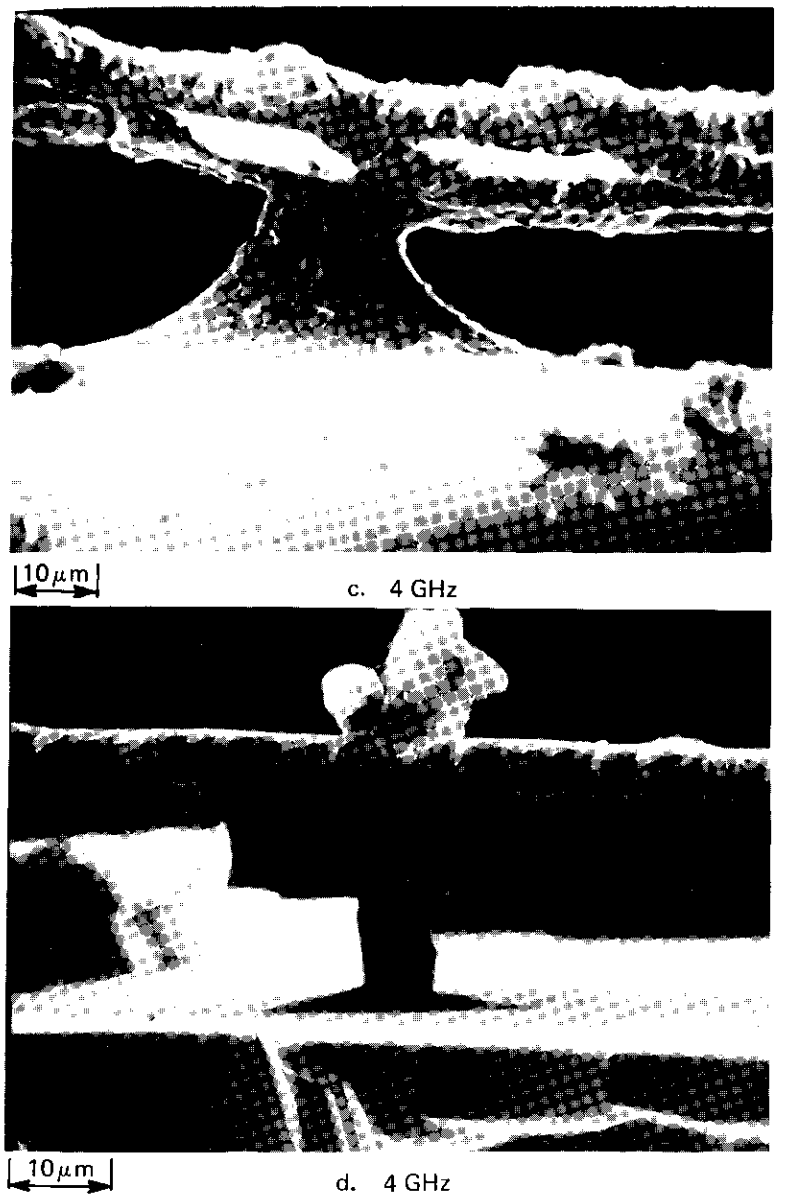


Figure 4. *SE Side View Images of Four Tunnel Diodes (continued)*

device, which increases the valley current in those diodes in which the stresses are sufficiently high. A more detailed but brief discussion of this potential failure mechanism may be found in Reference 5.

Figures 5a through 5d show the top view of the same diodes whose side views were shown in Figures 4a through 4d. These SE pictures were taken at about the same magnification as Figure 4. The poor soldering of the mesh to the cap mentioned previously is evident in Figure 5a. The attachment of the cap to the mesh is better on the diode shown in Figure 5b. In both cases, however, it is evident that the mesh grid was not heated sufficiently to make its tin coating flow. In Figure 5a, the original crystalline structure of the tin coating on the nickel mesh grid has not been disrupted by melting. The tin has just barely melted at the soldering point in Figure 5b.

Figure 5c shows a diode in which the mesh grid melted the tin mushroom extensively, causing noticeable deformation of the ball. (The debris in this picture, similar to those in Figure 4c, are artifacts.) Proper melting of the tin mushroom cap and its adhesion to the mesh are quite evident in Figure 5d. It can also be seen that the tin coating of the nickel mesh grid has been well melted. (Comparison of this picture with Figure 4d shows that the protrusions above the grid are not connected to the diode.)

Various other diodes were investigated in the same way. SEM pictures reveal several structural variations which may exist in tunnel diodes:

a. If the tin ball is not carefully positioned near the center of the chip, then the pinnacle will appear off center under the tin ball (see Plates I and II).

b. The tin ball may not be soldered to the center of the grid, but may be soldered into a hole or to the side of a grid wire (see Figure 5).

c. The tin ball may be barely attached to the grid or the grid may be pushed very deeply into the tin ball (compare Figures 5a through 5d).

Chemical analysis

To determine the distribution of the various elements throughout the entire device, an Electron Microprobe analysis was performed. The tunnel diode was first removed from its metal-ceramic housing and potted in epoxy. The potted device was then carefully sectioned and lapped in the middle of the pinnacle, perpendicular to the junction. Hence, it was

Plates I and II, on the following pages, will be perceived in three dimensions when the anaglyphoscope attached to the inside back cover is used with the red eyepiece over the right eye.

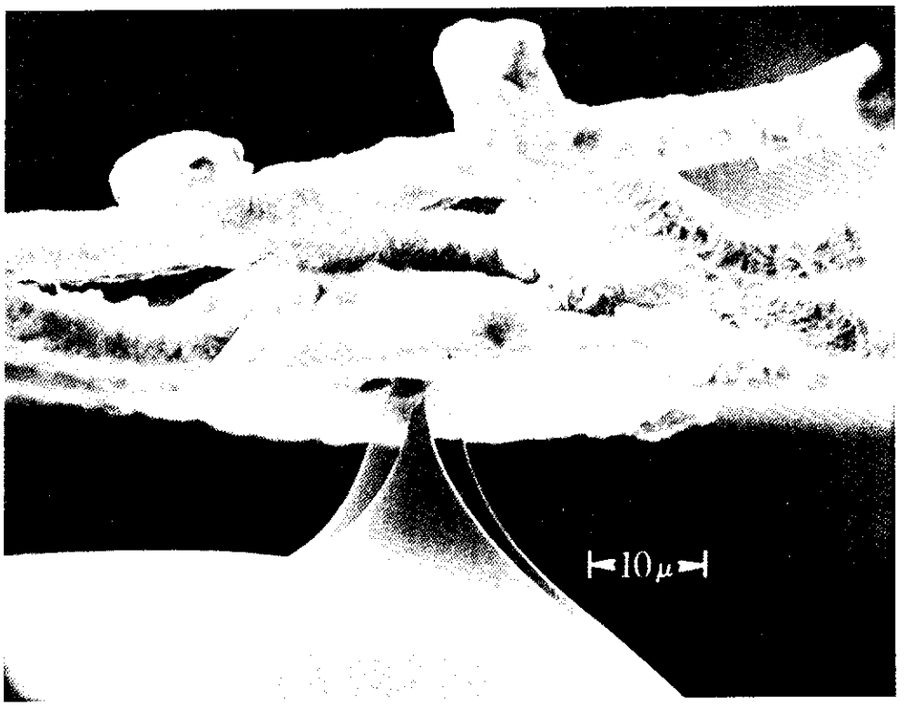


PLATE I. *Stereoscopic SE Image of a Germanium Tunnel Diode (side view)*

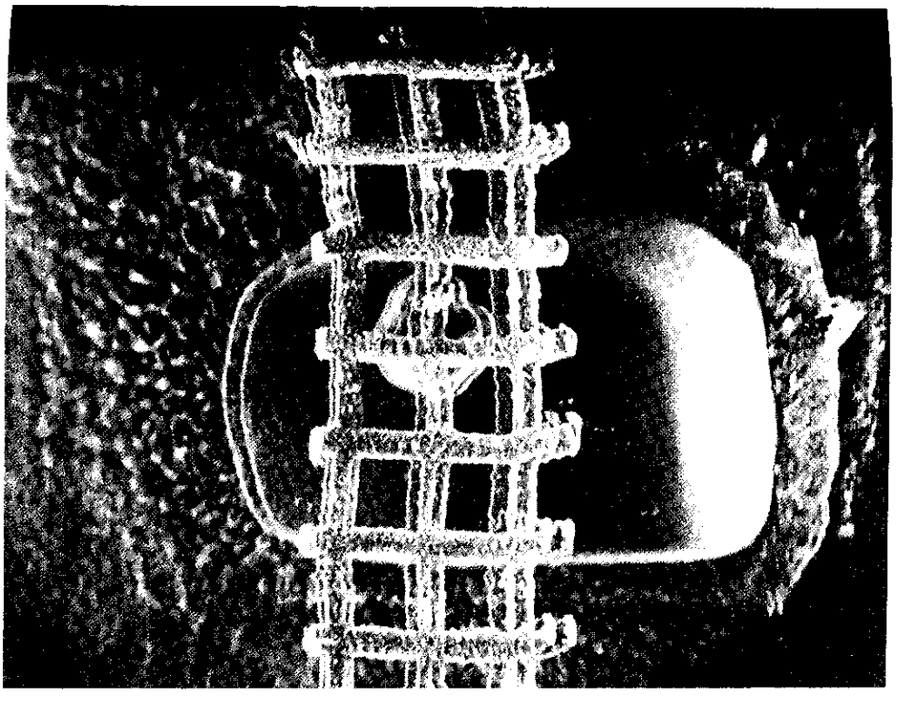
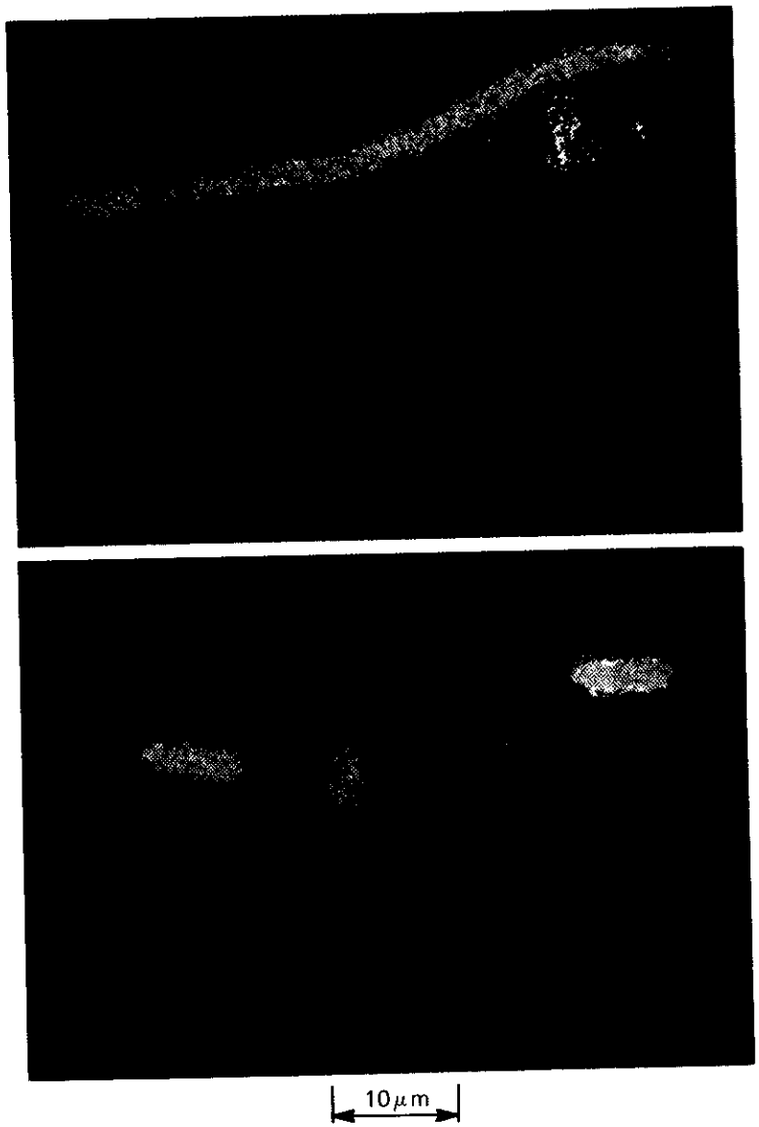


PLATE II. *Stereoscopic SE Image of a Germanium Tunnel Diode (top view)*



KEY: VIOLET - Ge, ORANGE - As, LT. BLUE - Ni, GREEN - Sn

PLATE III. *Elemental Distributions in Two Cross-sectioned Tunnel Diodes*

possible to determine the distribution of the Ge, As, Sn, Ga, and Ni elements over the entire device.

Plate III shows the distribution of the Ge, Sn, As, and Ni elements in two separate tunnel diodes. Each color picture is a montage of individual pictures of the elemental distribution. Violet represents the distribution of the Ge and green represents the distribution of Sn, while Ni is shown as blue, and As as orange. In both of these pictures, the Ni mesh is securely attached to the Sn mushroom cap. It is also evident that the As has become segregated in the Sn, as was shown in Figure 3. The As segregation should correspond to an Sn_3As_2 phase, while after alloying, the As is practically depleted in the remainder of the Sn mushroom cap.

Although the scanning pictures of the elemental distribution qualitatively revealed the segregation of As and the position of the Ni grid, this technique is not suitable to establish the exact composition or the position and chemistry of the junction. To achieve these ends, a point-by-point quantitative elemental analysis had to be performed. However, the sectioned devices were found to exhibit excessive thermally induced motion because of the longer electron beam exposure required for quantitative analysis.

To avoid this limitation, a quantitative analysis was performed on a germanium chip which had undergone all steps of the fabrication process with the exception of etching. The unetched but alloyed chip was potted and sectioned perpendicular to the junction area and lapped to a point representing the middle of the tin ball. A point-by-point microprobe analysis, in which 15 points, 1.4 μm apart, were sampled across the "junction," was performed on the specimen. The size of the electron beam permitted X-rays to be generated from an area approximately 1 μm in diameter at the surface of the sample; hence, the measuring points did not overlap. At each point a quantitative analysis for Ge, Sn, As, and Ga was performed.

In this mode the finely focused electron beam was successfully positioned under computer control for one second at each of the 15 analytical points until 30 seconds of data for each point were accumulated. This ensured an accurate mapping of the elemental concentration over the extremely small area to be analyzed. The analysis included peak and background counts for all elements and their standards. The weight percentages were then calculated on an IBM 360/65 computer by applying the Magic IV electron probe analysis program [6] to correct the raw X-ray intensity data for the effects of X-ray absorption and fluorescence and electron backscatter. The results, presented in Table 1 and Figure 6, are based

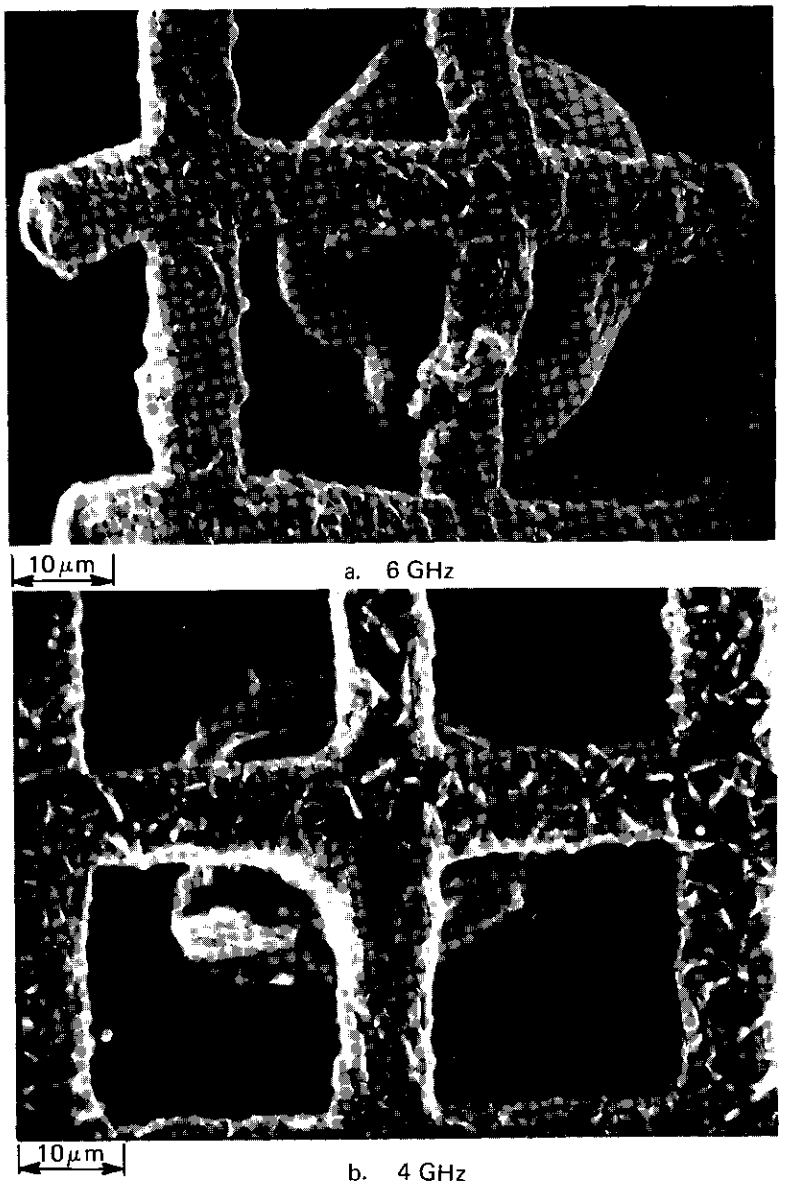


Figure 5. *SE Top View Images of Four Tunnel Diodes*

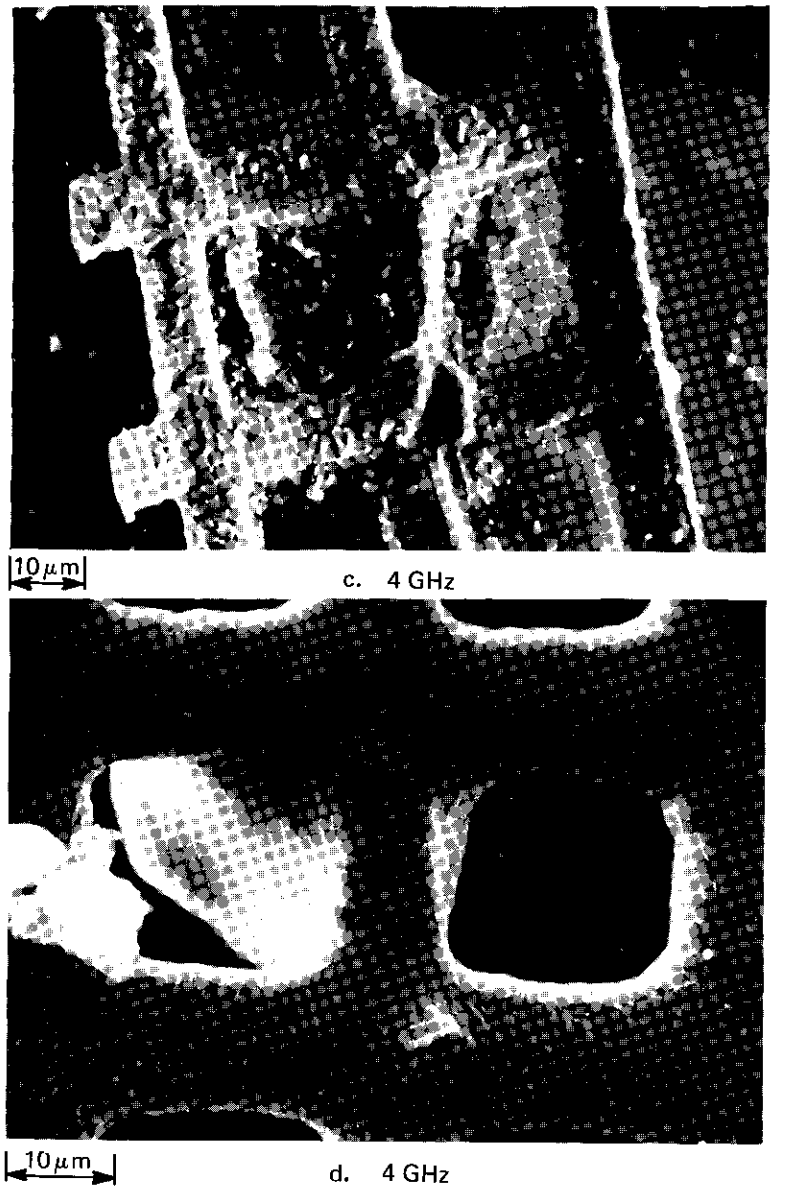


Figure 5. *SE Top View Images of Four Tunnel Diodes* (*continued*)

on the average of three complete sets of data obtained according to this procedure.

Table 1 indicates that a sizable quantity of germanium, i.e., ~4 percent, remains in the tin phase. This germanium appears to be accompanied by some Ga. Another interesting point is the amount of arsenic, 8 weight percent, in the enriched area resulting from the segregation. This amount is lower than is to be expected from an Sn_3As_2 phase, probably because the volume from which X-rays are obtained may be larger than that of the inclusion. Analysis of the germanium chip revealed only that the Ga was evenly distributed in the germanium at a concentration of 0.1 weight percent, even after the heat treatment.

TABLE 1. MEASURED WEIGHT PERCENT CONCENTRATIONS OF SN, AS, GE, AND GA IN A BALL ALLOY GERMANIUM DIODE

	Measured Weight Percent and 2-Sigma Limits (normalized to 100 percent*)			
Location	Sn	As	Ge	Ga
Sn Ball				
Average (excluding As enriched area)	95.33 ± 0.8	0.43 ± 0.35	4.20 ± 0.45	0.03 ± 0.01
As Enriched Area	88.85 ± 1.74	8.05 ± 1.12	3.08 ± 0.07	—
Ge Chip				
Average	0.07 ± 0.04	0.02 ± 0.02	99.75 ± 0.73	0.15 ± 0.04

*Data were taken at 25-kV, 1.5-nA sample current using the following X-ray emission lines: $Sn_{L\alpha}$, $As_{L\alpha}$, $Ge_{K\alpha}$, and $Ga_{L\alpha}$.

The chemistry at the junction was interpolated from the data plotted in Figure 6. It can be seen that there exists a region about 3 μm wide across the metallurgical interface (line II′ to JJ′) in which the tin concentration increases from left to right. In this region the arsenic concentration decreases rapidly. The Ga concentration is fairly constant, but falls off rapidly in the Sn phase. It can be seen that the n-p junction, where the As concentration equals that of the Ga (~0.1 weight percent), line JJ′, is confined to a narrow zone at the interface of the recrystallized germanium with the original germanium material and resides a few micrometers from

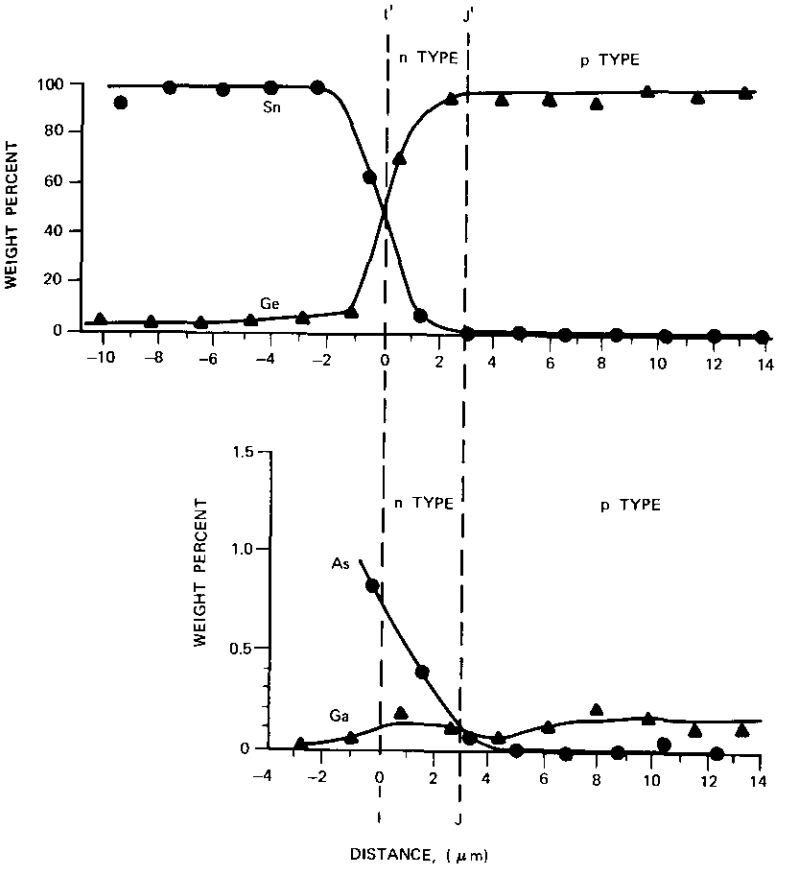


Figure 6. *Sn, As, Ge, and Ga Distributions Measured in Ball Alloy Diode Cross-section*

the metallurgical interface. The magnitude of the As and Ga concentrations in the neighborhood of the n-p junction are consistent with those derived from measurements of the I-V characteristics of actual diodes and interpretation thereof according to standard device theory such as that presented in Reference 7.

Conclusion

The SEM/EM investigation of ball alloy tunnel diodes has revealed a striking mechanical precariousness in their structure. That is, a mushroom

cap of tin, 25 μm in diameter, is perched upon a pinnacle of Ge about 5 μm in diameter at the recrystallized tip. It has also revealed that improper assembly procedures may result in poor grid connection and in the formation of an asymmetrical structure. The pictures obtained from this study support a recently advanced theory [4] on the failure mechanism. This theory is based upon plastic deformation of the Ge in the junction area caused by inherent mechanical stresses within the diodes and resulting in higher valley current in those diodes in which the stresses are sufficiently high.

Microprobe analysis of the tunnel diodes has revealed by direct "chemical" analytical means the exact location of the n-p junction. It has also revealed the segregation of arsenic in the tin after the alloying process and has elucidated the chemistry of the recrystallized zone.

Acknowledgment

The authors wish to express their thanks to J. Hannsen for his collaboration in the experimental work, to P. Fleming for selection of the tunnel diodes, and to E. S. Rittner for constant encouragement in the course of their work, as well as for valuable comments on the manuscript. They would also like to thank C. Blain of KMC Semiconductor Corporation, Long Valley, New Jersey, who provided the unetched Ge chip for chemical analysis.

References

[1] S. Virk, "State of Tunnel Diode Technology," *Electronic Products*, Vol. 12, No. 6, November 1969, pp. 30–32.

[2] M. Hansen, *Constitution of Binary Alloys*, New York: McGraw-Hill, 1958.

[3] T. D. Kirkendall and P. F. Váradi, "An Automated Microprobe Under PDP-8 Control Using an IBM 360/65 for Program and Data Storage," *Sixth National Conference on Electron Probe Analysis*, Pittsburgh, Pennsylvania, July 1971.

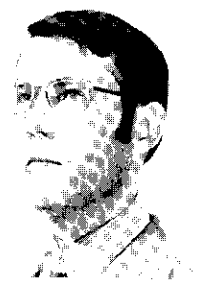
[4] A. G. Revesz, J. Reynolds, and J. Lindmayer, "New Aspects of Failure Mechanism in Germanium Tunnel Diodes," *Solid State Electronics*, Vol. 14, No. 11, November 1971, pp. 1137–1147.

[5] P. L. Bargellini and E. S. Rittner, *Advances in Satellite Communications: Advances in Electronics and Electron Physics*, Vol. 31, edited by L. Marton, New York: Academic Press, 1972.

[6] J. W. Colby, "Magic IV-A Computer Program for Quantitative Electron Microprobe Analysis," Bell Telephone Laboratories, Allentown, Pennsylvania.

[7] D. R. Fredkin and G. H. Wannier, "Theory of Electron Tunneling in Semiconductor Junctions," *Physics Review*, Vol. 128, No. 5, December 1, 1962, p. 2054.

Peter F. Váradi studied chemistry at the University of Szeged, Hungary, and received a Ph.D. in physico-chemistry. Since 1968 he has been Manager of the Materials Sciences Department of COMSAT Laboratories. He is author and co-author of numerous scientific papers and patents in the field of electronic materials and vacuum technology, and in the various branches of analytical chemistry instrumentations. Dr. Váradi is a fellow of the American Institute of Chemists, senior member of the American Vacuum Society and he is presently chairman of the Baltimore-Washington Section of the Society for Applied Spectroscopy.



T. D. Kirkendall received a B.A. degree in physics from Colby College (1961). He has seven years experience at the Machlett Laboratories, Raytheon Co., in the analysis and application of materials in vacuum tubes and opto-electronic devices. Mr. Kirkendall joined COMSAT Laboratories in 1969. He is a Member of the Technical Staff, Materials Sciences Department, COMSAT Laboratories, working on the physico-chemical analysis of spacecraft and earth station materials and components with special interests in electron microprobe and X-ray spectroscopy techniques.

Index: communications satellites, electric propulsion, thrust vector control, alignment, error analysis.

Estimation and correction of electric thruster misalignment effects on a geostationary satellite

M. H. KAPLAN

Abstract

An investigation of the effects of thruster misalignment on a communications satellite employing a double-gimballed momentum wheel attitude control system is presented. Electric north-south thrusters are the primary source of misalignment torques. These misalignment effects are particularly harmful when continuous yaw sensing is not provided and narrow-beam antennas are incorporated into the satellite. Thrust misalignment cannot be accurately predicted; thus, actual flight data must be used to estimate errors.

Determination of the characteristics of ground-based and onboard control techniques is the major objective of this investigation. All necessary east-west drift corrections can be made with longitude control units, while attitude perturbations may be eliminated through thrust vectoring of the exhaust beam. Misalignment limits and specific vehicle properties are assumed, and consequent sensor requirements and control system characteristics are discussed.

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

One of the more promising configurations for a family of advanced communications satellites employs an earth-oriented central body with sun-oriented panels as illustrated in Figure 1. Several orbit and attitude control systems are currently being investigated for application to this configuration in geostationary orbit. The study reported here deals with a vehicle containing a double-gimballed momentum wheel for attitude control and electric thrusters for north-south (N-S) stationkeeping.

There is always some finite misalignment of the thrust vector of an orbit control engine; this misalignment may be thought of as a combination of angular and position errors. Misalignment effects are of no consequence to low-impulse maneuvers such as east-west (E–W) holding, but for maneuvers with large impulse requirements, such as N–S stationkeeping and large station change maneuvers, they can cause significant attitude errors if they are not detected and corrected. When narrow-beam antennas are employed on a satellite without a yaw sensor, the large yaw errors which may result during N-S thrusting will lead to excessive beam-pointing errors. Such yaw errors are not so critical during station change maneuvers because high-density communications functions are not required. Thus, N–S thrusting represents a particularly important source of attitude error in this situation. Furthermore, misalignment may have an adverse effect on longitude holding which must be countered with direct E-W thrusting.

Such misalignments cannot be predicted accurately because of uncertainties in structural deformations and propellant movement. Thus, actual flight data must be used to estimate errors and generate corrections. The vehicle center of mass experiences slight shifts as a result of changing propellant mass, structural response to attitude control torques, and thermal bending; hence, thrust alignment and offset measurements and corrections should be performed at the beginning of each thrusting interval, and possibly more often.

Investigation and determination of ground-based and onboard estimation and control techniques are the major objectives of this work. Vectoring of N–S thruster exhausts is assumed to be the correction mechanism for misalignment torques. This technique may also be useful as a mode for nominal attitude control. In-plane drift can be controlled with E–W stationkeeping engines.

Sensor requirements and control system characteristics are discussed in terms of the misalignment limits and specific vehicle properties assumed. Analyses of E–W motion and attitude dynamics are also included.

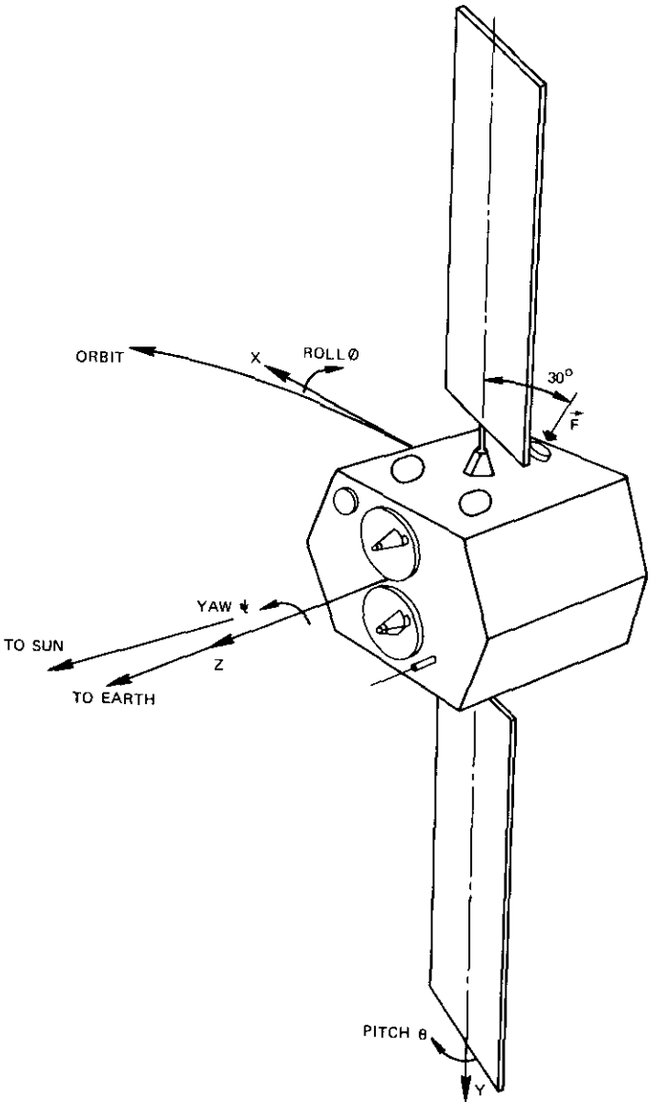


Figure 1. *Assumed Satellite Configuration*

Satellite specifications

The satellite configuration employed in this investigation is illustrated in Figure 1. Since the launch vehicle is assumed to be an Atlas-Centaur, in-orbit mass is limited to about 716 kg (1575 lb). Satellite mass and power properties are essentially those of the vehicle studied in Reference 1. Specific spacecraft characteristics of interest are listed in Table 1 and are based on realistic estimates [2]. Of course, the addition of a continuous yaw sensor would greatly simplify misalignment estimates and corrections. However, it is desirable to achieve high beam-pointing accuracy without the added complexity of a star tracker. This capability can also be thought of as a backup mode in case of yaw sensor failure.

TABLE 1. ASSUMED SATELLITE SPECIFICATIONS

Initial In-Orbit Mass*	716 kg (1575 lb)
Moments of Inertia	
$I_X = I_Z$	2,000 kg-m² (1475 slug-ft²)
I_Y	440 kg-m² (325 slug-ft²)
Solar Array Area	20 m² (215 ft²)
Useful In-Orbit Life	7 years (1979–1985)
Attitude Accuracy	
Pitch/Roll	0.05°
Yaw	0.40°
Onboard Sensors	
Earth Sensor	pitch and roll
Sun Sensors	pitch and part time yaw
Stationkeeping Accuracy	
N–S	0.1°
E–W	0.05°
Attitude Control	double-gimballed momentum wheel and dumping jets

*For an Atlas-Centaur launch with incorporated apogee kick motor.

The N–S thruster selected for consideration is an electric device with thrust vectoring capacity. Conventional high-thrust devices such as hydrazine jets create significant misalignment torques, but short firing intervals may permit easy measurement of torque because N–S corrections may be made at times favorable to yaw measurement by sun sensors. Torque effects are eliminated with attitude jets. Low-thrust devices introduce unique torque estimation problems and correction techniques which

are discussed in detail for electric thrusters whose properties are listed in Table 2. Selection of an engine of this size was based on duty cycle, satellite mass, and perturbation considerations [1].

TABLE 2. NORTH-SOUTH THRUSTER PROPERTIES

Nominal Thrust Level	6.7 mN (1.5 mlb)
Orientation	30° cant angle away from pitch line into radial direction
Nominal Duty Cycle	6 hours symmetrical about each node (25 percent per engine when correcting)
Misalignment	$\leq$5 cm (1.96 in.) in center of mass offset, 0.1° to 1.0° into E–W
Thrust Vectoring	$\leq$10° away from nominal direction

The use of a momentum wheel system for a vehicle of this size has not previously been considered. Selection of the nominal wheel momentum depends on steady-state yaw error magnitudes resulting from constant disturbance torques and specified antenna pointing accuracy [3]. A discussion of nominal momentum will be presented in a later section.

Required data for ground-based measurements

To eliminate misalignment thrusts it is necessary to relate measurable quantities to the thrust vector geometry. A unique description of this vector will require evaluation of four small unknown quantities plus the magnitude of thrust. Two of these quantities are required to define "offset," i.e., displacement of the thrust vector from the center of gravity without misalignment of the thrust orientation. The other two quantities define "angular misalignment" and uniquely orient the thrust vector with respect to the reference orientation.

Figure 2 illustrates the coordinates and nomenclature used to formulate transformations relating thruster offset to torque components. Force perturbations which affect the orbit occur only when there is an angular misalignment or attitude error. For this analysis it is assumed that attitude errors are nominally zero in the case of an active control system. Figure 3 illustrates misalignment nomenclature. In summary, perturbing torques have two distinct contributors: center of mass offset and angular misalignment.

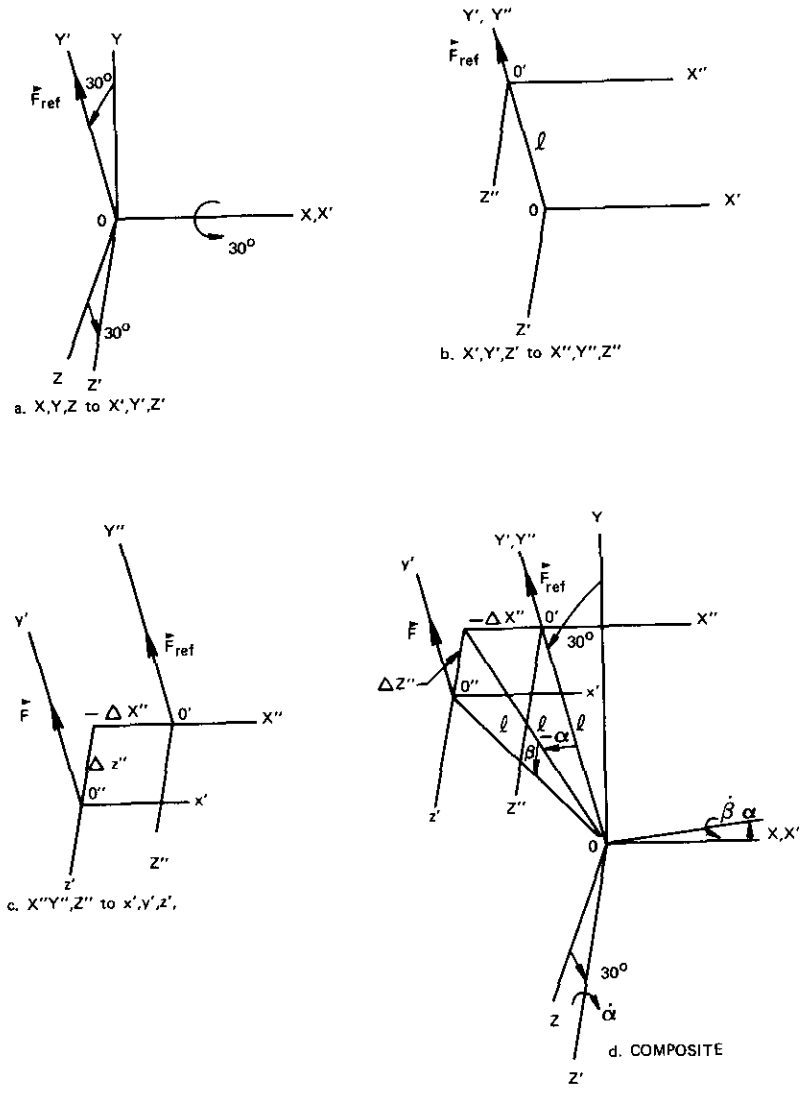


Figure 2. *Nomenclature for Thruster Offset*

Initially, the thrust direction of each N–S engine relative to the vehicle center of mass and axes is assumed to be unknown and arbitrary to within

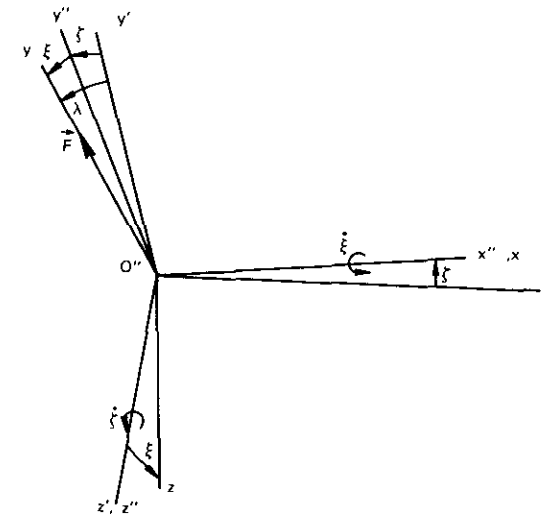


Figure 3. *Nomenclature for Thrust Vector Angular Misalignment*

the specified error limits [center of mass offset of 5 cm (1.96 in.) and angular error of 1°]. For this analysis only one N–S thruster must be considered. A reference point, 0″, on the thruster centerline is selected to locate the unit on a spherical surface with a radius, l, equal to 0.91 m (3.0 ft). The location of this thruster is uniquely specified by $\Delta X''$ and $\Delta Z''$, which together satisfy the offset limit

$$\sqrt{(\Delta X'')^2 + (\Delta Z'')^2} \leq 5 \text{ cm } (1.96 \text{ in.}) \quad .$$

Small variations in l will not significantly affect motion since they will cause only "second-order" errors.

The sequence of transformations used to compute torques caused by offset only is illustrated in Figure 2 and can be described as follows:

a. 30° rotation from body axes X,Y,Z to ideal thrust axes X',Y',Z';
b. displacement of X',Y',Z' axes from the center of mass, 0, to thruster-centered axes X'',Y'',Z''; and
c. displacement of the thrust vector from the ideal position 0′ to the offset position 0″, which is the origin of the x',y',z' axes.

Since $F_{Y'} = F$, the force components when there is no misalignment are

$F_X = 0$, $F_Y = F \cos 30°$, and $F_Z = F \sin 30°$. An offset of this thruster produces moment arms about the center of mass, 0, with magnitudes of $\Delta Z''$ about the X axis and $\Delta X''$ about the Y and Z axes. (Movement of 0 will cause the same moments.) Thus, the torque components caused by $\Delta X''$ and $\Delta Z''$ are

$$T_X = -F\Delta Z'' \tag{1a}$$

$$T_Y = -F\Delta X'' \sin 30° \tag{1b}$$

$$T_Z = F\Delta X'' \cos 30° \quad . \tag{1c}$$

In addition to these torques, angular misalignment of the thrust vector produces a complete set of perturbing forces and torques. This misalignment is most conveniently handled as shown in Figure 3, in which the y' axis is the reference direction of the thrust vector. Two small angles, ζ and ξ, uniquely orient this vector relative to its ideal direction, and the angle λ is the total angular misalignment of $\vec{F}$. Thus, the limit on this misalignment is expressed as $\lambda \leq 1°$. Spherical trigonometry provides the relationship between Euler angles ζ and ξ and misalignment λ:

$$\cos \lambda = \cos \zeta \cos \xi \quad .$$

The transformation from x,y,z to x',y',z' is easily accomplished with an orthogonal matrix. Thus,

$$\begin{bmatrix} F_{x'} \\ F_{y'} \\ F_{z'} \end{bmatrix} = \begin{bmatrix} (\cos \zeta) & (-\sin \zeta \cos \xi) & (\sin \zeta \sin \xi) \\ (\sin \zeta) & (\cos \zeta \cos \xi) & (-\cos \zeta \sin \xi) \\ 0 & (\sin \xi) & (\cos \xi) \end{bmatrix} \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} \quad .$$

Since the components of $\vec{F}$ are $F_x = F_z = 0$, $F_y = F$,

$$\begin{bmatrix} F_{x'} \\ F_{y'} \\ F_{z'} \end{bmatrix} = \begin{bmatrix} -\sin \zeta \cos \xi \\ \cos \zeta \cos \xi \\ \sin \xi \end{bmatrix} F \quad .$$

To calculate F_X,F_Y,F_Z, the transformation of the force, F, from x',y',z' to a set parallel to X,Y,Z must be made:

$$\begin{bmatrix} F_X \\ F_Y \\ F_Z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos 30° & -\sin 30° \\ 0 & \sin 30° & \cos 30° \end{bmatrix} \begin{bmatrix} F_{x'} \\ F_{y'} \\ F_{z'} \end{bmatrix}$$

or

$$\begin{aligned} F_X &= -(\sin \zeta \cos \xi)\, F && \text{(E–W component)} \\ F_Y &= (\cos 30° \cos \zeta \cos \xi - \sin 30° \sin \xi)\, F && \text{(N–S component)} \\ F_Z &= (\sin 30° \cos \zeta \cos \xi + \cos 30° \sin \xi)\, F && \text{(radial component).} \end{aligned}$$

Since the maximum value of ζ or ξ must be less than λ, these angles are very small and only first-order effects must be considered. Thus,

$$\begin{bmatrix} F_X \\ F_Y \\ F_Z \end{bmatrix} = \begin{bmatrix} -\zeta \\ \cos 30° - \xi \sin 30° \\ \sin 30° + \xi \cos 30° \end{bmatrix} F \tag{2}$$

where ζ and ξ are in radians. Perturbing torques caused by angular misalignment are obtained from the set in equation (2) and the geometry of Figure 2:

$$T_X \simeq Fl\xi \tag{3a}$$

$$T_Y \simeq -Fl\zeta \sin 30° \tag{3b}$$

$$T_Z \simeq Fl\zeta \cos 30° \quad . \tag{3c}$$

Combining offset and misalignment torques, equations (1) and (3), yields

$$T_X = F(l\xi - \Delta Z'') \quad \text{(roll)} \tag{4}$$

$$T_Y = -F \sin 30° \, (l\zeta + \Delta X'') \quad \text{(pitch)} \tag{5}$$

$$T_Z = F \cos 30° \, (l\zeta + \Delta X'') \quad \text{(yaw)} \quad . \tag{6}$$

Thrust vectoring can eliminate perturbing torques, while direct E–W impulses will correct longitudinal drift rates and position errors. The unknowns $\Delta X''$, $\Delta Z''$, ζ, and ξ must be determined from observed responses during thrusting intervals. It is assumed that the thrust magnitude can be accurately measured from telemetry. Thus, an E–W drift attributable to N–S thrusting results from a force $-\zeta F$ plus the orbital coupling effect of F_Z. Subtracting the anticipated longitudinal displacement caused by F_Z yields the value of ζ, as discussed in the next section.

The yaw torque caused by the offset and misalignment expressed in equation (6) is of particular interest because yaw errors cannot always be measured if no direct yaw sensing is used. However, only two unknowns (ζ, $\Delta X''$) are required to obtain a first-order estimation of T_Z. Momentum

wheel stiffness will limit second-order effects to the specified accuracy. The value of ζ may be obtained from observations of E–W motion, while $\Delta X''$ may then be obtained from observations of pitch errors and from equation (5). Adjustment of the angle ζ will eliminate $\Delta X''$ effects and possibly increase E–W drift. Hence, it is necessary to determine only the sensitivity of this motion to ζ. This is also discussed in the next section.

A unique set of values for ξ and $\Delta Z''$ seems to be very difficult to obtain because N–S and radial motions are relatively insensitive to small thrust component errors. Fortunately, determination of these two quantities is not critical, since they result in directly observable and correctable motions. In the worst case, attitude jets may be directly employed to control roll motion. However, a few attempts at varying ξ (and thus controlling $\Delta Z''$) should quickly improve the situation.

Effects of misalignment on E-W drift

Since $F_X = -\zeta F$, the value of ζ may be obtained through observations of E–W drift. The relationship between ζ and longitudinal drift can be derived by using the equations for in-plane relative motion [1]. In the notation of Figure 4, these become

$$\ddot{\rho}_r - 2\omega_0 \dot{\rho}_t - 3\omega_0^2 \rho_r = p_r + f_r$$

$$\ddot{\rho}_t + 2\omega_0 \dot{\rho}_r = p_t + f_t$$

where p_r,p_t and f_r,f_t are perturbative and thrust acceleration components, respectively. Since a 30° cant angle is assumed,

$$|f_r| = \frac{1}{2} F$$

$$|f_t|_{\max} = F \sin 1° = (0.0175)\ F$$

where F is the magnitude of N–S thruster acceleration on the vehicle. For constant values of f_r, f_t, p_r, and p_t, the particular solution which satisfies both direct E–W effects and drift resulting from the radial component of thrust is

$$\rho_r(t) = \frac{(p_r + f_r)}{\omega_0^2} + \frac{2(p_t + f_t)}{\omega_0}\, t$$

$$\rho_t(t) = -\frac{3}{2}(p_t + f_t)\, t^2 - \frac{2}{\omega_0}(p_r + f_r)\ t$$

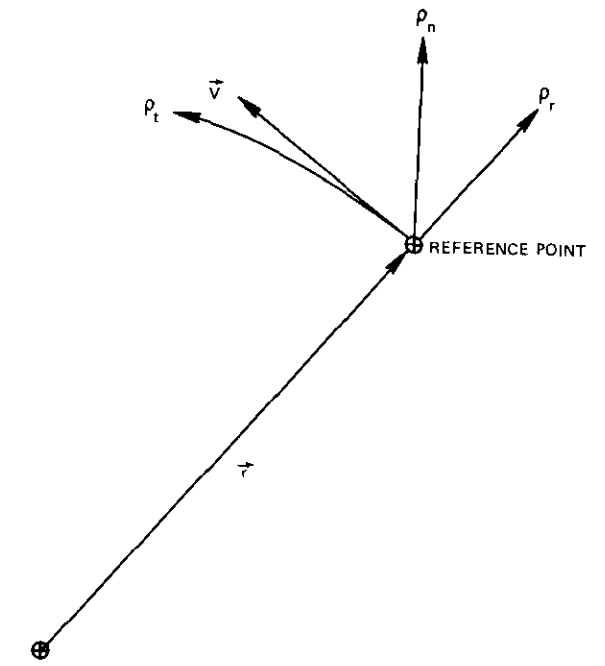


Figure 4. *Nomenclature for Longitudinal Drift Analysis*

The effect of positive $(p_r + f_r)$ on longitude is a constant backward drift in ρ_t of magnitude $(2/\omega_0)\ (p_r + f_r)$. Acceleration caused by $(p_t + f_t)$ would cause a parabolic in-plane drift as described in Reference 1.

Consider the case of interest:

$F = 9.34 \times 10^{-6}$ m/s^2 (3.06×10^{-5} ft/s^2) for 6.7 mN

$|f_t|_{\max} = 0.163 \times 10^{-6}$ m/s^2 (0.053×10^{-5} ft/s^2) max misalignment

$f_r = 4.67 \times 10^{-6}$ m/s^2 (1.53×10^{-5} ft/s^2)

$\omega_0 = 7.28 \times 10^{-5}$ rad/s

Since misalignment effects can be superimposed on normally perturbed

motion, perturbations p_r and p_t can be treated separately. If initial conditions are all zero, the in-plane motion caused by this thruster is

$$\rho_{r_F} = \frac{f_r}{\omega_0^2} + \frac{2f_t}{\omega_0} t$$

$$\rho_{t_F} = -\frac{3}{2} f_t t^2 - \frac{2}{\omega_0} f_r t$$

However, misalignment effects on radial motion do not exceed the order of thrust magnitude uncertainty. Longitudinal motion is most sensitive to thrust misalignment, and observations of E–W drift can yield values of ζ. In this case,

$$\rho_{t_F} = -0.244 \times 10^{-6} t^2 - 0.128 t$$

where ρ_{t_F} is in meters when t is in seconds. After a 6-hour thrusting interval,

$$\rho_{t_F} \text{ (6hr)} = -114 \text{ m} - 2770 \text{ m} = -2884 \text{ m } (-9461.9 \text{ ft})$$

Therefore, the effect of misalignment on E–W motion is about four percent of that caused by the radial component of N–S thrust, and misalignment produces a maximum longitudinal error of $1.8° \times 10^{-4}$. Present technology will permit a minimum measurement of $5° \times 10^{-2}$ based on observations obtained from ground-based equipment, and only after several orbital periods. Hence, it appears that ground-based techniques do not provide sufficiently sensitive or fast measurements for thrust misalignment evaluations.

Onboard estimation and automatic correction

Before automatic correction schemes are considered, the uncontrolled response of the spacecraft to the steady outside torque of an N–S thruster should be determined. This may provide some insight into momentum wheel sizing and backup modes in case of gimbal actuator failure. Of course, uncontrolled pitching motion results in pitch angle acceleration, which can be corrected with direct reaction jet applications. The linearized roll/yaw equations in the case of a fixed actuator (uncontrolled roll/yaw) are [4]

$$T_X = I_X \ddot{\phi} + \omega_0 H_N \phi + H_N \dot{\psi}$$

$$T_Z = I_Z \ddot{\psi} + \omega_0 H_N \psi - H_N \dot{\phi}$$

where H_N = nominal wheel momentum

T_X, T_Z = external roll and yaw disturbance torques, respectively.

If it is assumed that initial conditions are all zero, Laplace transforms may be used to solve these equations for the yaw response:

$$\Psi(s) = \frac{T_X(s)\ H_N s + T_Z(s)\ (I_X s^2 + \omega_0 H_N)}{I_X I_Z s^4 + [H_N^2 + \omega_0 H_N (I_X + I_Z)]\ s^2 + \omega_0 H_N^2}$$

It should be noted at this point that steady-state values of $\psi(t)$ cannot be obtained because yaw motion is oscillatory and without natural damping. Therefore, the Final Value Theorem cannot be applied. The yaw response to a step input about the Z-axis is obtained by setting $T_X = 0$ and $T_Z(s) = T_Z/s$ and noting that $I_X = I_Z$ in this case. Then,

$$\psi(t) = \frac{T_Z}{\omega_0 H_N} (1 - \cos \omega_0 t) + \frac{I_X T_Z}{H_N^2} \left[1 - \cos \left(\frac{H_N}{I_X} \right) t \right]$$

This solution represents two separate phenomena. The first periodic term is associated with roll coupling into yaw at the orbital rate, ω_0, while the second represents the precessional effect caused by the momentum wheel. This latter motion is a short-period oscillation whose frequency and magnitude depend on the value of the nominal momentum, H_N. For example, if $H_N = 138$ N·m-s (100 lb-ft-s), since the maximum torque caused by an offset of 5 cm (1.96 in.) and an angular misalignment of 1° is $T_Z = 3.8 \times 10^{-4}$ N·m (2.8×10^{-4} ft-lb), then ψ becomes

$$\psi(t) = 2.2(1 - \cos \omega_0 t) + 2.36 \times 10^{-3} (1 - \cos 0.0678\ t)$$

in degrees. Thus, the short-period oscillation produces a very small yaw error, while orbit rate coupling results in a maximum allowable yaw error (0.4°) in two hours and 21 minutes. To passively control this maximum N–S thruster torque, H_N must be at least 1,520 N·m-s (1,100 lb-ft-s).

With an active control system in operation, the amplitude of the yaw error will differ from that obtained in the previous equation because artificial damping can be provided through the selection of an actuator control law. Thus, a steady-state yaw error whose magnitude is a function of the bias momentum, H_N, will result. This error can be determined by con-

sidering the linearized roll/yaw equations and selected control laws. These equations are written with actuator inputs as

$$T_X = I_X\ddot{\phi} + \omega_0 H_N \phi + H_N\dot{\psi} + \dot{H}_X - \omega_0 H_Z$$
$$T_Z = I_Z\ddot{\psi} + \omega_0 H_N \psi - H_N\dot{\phi} + \dot{H}_Z + \omega_0 H_X$$

where it has been assumed that

$$H_N \gg \max\,[I_X\omega_0,\ I_Y\omega_0,\ I_Z\omega_0]$$

to ensure that the yaw rate is coupled into the roll output. The selected roll and yaw control moment laws are

$$M_{X_c} = \dot{H}_X - \omega_0 H_Z = K\tau\dot{\phi} + K\phi - \omega_0 H_N \phi$$
$$M_{Z_c} = \dot{H}_Z + \omega_0 H_X - H_N\dot{\phi} = -kK(\tau\dot{\phi} + \phi)$$

where
K = roll autopilot gain
k = yaw-to-roll gain ratio
τ = roll time constant.

It is assumed that the horizon sensor output provides exact values of ϕ and that a pseudorate modulator generates exact values of $\dot{\phi}$. The roll/yaw equations then become

$$T_X = I_X\ddot{\phi} + H_N\dot{\psi} + K\tau\dot{\phi} + K\phi$$
$$T_Z = I_Z\ddot{\psi} + \omega_0 H_N \psi - kK(\tau\dot{\phi} + \phi)$$

In the Laplace transform notation, these are written as follows:

$$\begin{bmatrix} I_X s^2 + K(\tau s + 1) & H_N s \\ -kK(\tau s + 1) & I_Z s^2 + \omega_0 H_N \end{bmatrix} \begin{bmatrix} \Phi \\ \Psi \end{bmatrix} = \begin{bmatrix} T_X(s) \\ T_Z(s) \end{bmatrix} \tag{7}$$

This partially factored characteristic equation is approximated by

$$(I_X s^2 + K\tau s + K)\ (I_Z s^2 + kH_N s + \omega_0 H_N) = 0$$

provided that

$$K\tau I_Z \gg kH_N I_X \tag{8}$$

The associated natural frequencies and damping ratios are

$$\omega_1 = \sqrt{\frac{K}{I_X}} \tag{9a}$$
$$\omega_2 = \sqrt{\frac{\omega_0 H_N}{I_Z}} \tag{9b}$$
$$\zeta_1 = \frac{\tau}{2}\sqrt{\frac{K}{I_X}} \tag{9c}$$
$$\zeta_2 = \frac{k}{2}\sqrt{\frac{H_N}{\omega_0 I_Z}} \tag{9d}$$

where ω_1 and ζ_1 are related to roll dynamics, while ω_2 and ζ_2 correspond to yaw error correction sequences.

Solving equation (7) for the yaw response yields

$$\Psi = \frac{kK(\tau s + 1)\ T_X + [I_X s^2 + K(\tau s + 1)]\ T_Z}{(I_X s^2 + K\tau s + K)\ (I_Z s^2 + kH_N s + \omega_0 H_N)} \tag{10}$$

The steady-state yaw error, ψ_{ss}, resulting from constant roll and yaw torques is obtained by applying the Final Value Theorem to equation (10):

$$\psi_{ss} = \frac{kT_X + T_Z}{\omega_0 H_N}$$

To estimate the time required to reach steady-state yaw, some numerical values must be assumed. To avoid "overshoot" of the response, critical damping is assumed. Values of H_N and K are selected through consideration of steady-state errors and hardware properties. Thus,

$$\zeta_1 = \zeta_2 = 1.0$$
$$H_N = 276\ \text{N} \cdot \text{m-s}\ (200\ \text{lb-ft-s})$$
$$K = 4.5\ \text{N} \cdot \text{m/rad}\ (3.33\ \text{lb-ft/rad})$$

Equation (9) then yields

$$k = 0.0464$$
$$\tau = 42\ \text{s}$$
$$\omega_1 = 0.0476\ \text{rad/s}$$
$$\omega_2 = 3.14 \times 10^{-3}\ \text{rad/s}$$

These values satisfy equation (8). Thus, yaw exponentially approaches its steady-state error within the first hour of N–S thrusting.

Since the value of K is only about 0.05, the criterion for bias momentum selection becomes

$$H_N \geq \frac{T_Z}{\omega_0 \psi_{\max}} = 743 \text{ N} \cdot \text{m-s (550 lb-ft-s)}$$

for the maximum torque caused by N–S thrusting. Of course, for a given value of H_N, the steady-state yaw error will be

$$\psi_{ss} = \frac{T_Z}{\omega_0 H_N}$$

If the wheel is sized according to a solar torque of magnitude $|T_Z| = 5 \times 10^{-5}$ N·m (3.7×10^{-5} lb-ft), then

$$H_N \geq 98.0 \text{ N} \cdot \text{m-s (72.5 lb-ft-s)}$$

For $H_N = 276$ N·m-s (200 lb-ft-s), the corresponding steady-state yaw offset can be as much as 1.1°. Thus, it is essential to determine torque components as quickly as possible after each initiation of N–S thrusting.

Onboard estimation of roll and yaw torque components is possible because of coupling phenomena in the attitude control system [4]. The number of actuator command pulses required for the orbit rate interchange of roll and yaw momentum (assuming orbit rate decoupling) can be subtracted from the total number of pulses; those in excess yield a measure of the roll gimbal response caused by external torques. At about five minutes ($1/\omega_2$) after initiation of thrusting, the roll actuator reaches a maximum sensitivity to yaw torque and is insensitive to roll torque if perturbing moments of the N–S thruster are constant. Thus, an estimate of yaw unbalanced torque can be obtained by counting the number of roll gimbal actuator pulses occurring in a fixed time interval around the 5-minute point. Since roll actuator sensitivity is reversed beyond the 25-minute ($5/\omega_2$) point, the roll component of torque can be estimated by using the same technique. This method claims to provide thrust vector correction to within about 0.5 cm (0.2 in.) of the center of mass by simply varying the misalignment angles ζ and ξ according to the roll actuator output.

Conclusions

The following basic conclusions regarding the feasibility of ground-based and onboard estimation and control of N–S thruster torque effects are based upon assumptions and analyses presented in this paper:

a. Ground-based observations of E–W drift can yield the yaw component of N–S thruster torque. However, the time required to obtain accurate data is excessive, thus rendering this technique impractical.

b. Onboard estimation of N–S thruster torques is possible because of coupling phenomena in the attitude control system. This technique is practical because the time to make observations is shorter than the time of yaw error buildup. This scheme also assumes that torques do not change during observation intervals.

c. The inherent stiffness of the spacecraft depends on the nominal momentum of the wheel, H_N. A large value of H_N is very desirable for both controlled and gimbal-fixed operations.

d. Excessive E–W drift caused by N–S thruster misalignment is most easily and effectively controlled with longitudinal stationkeeping thrusters. Such drifts may be induced as a byproduct of thrust vectoring to eliminate misalignments or generate attitude control torques.

In summary, it appears that estimation and correction of torques caused by N–S thrusting are possible and practical even without direct yaw sensing. Further analyses and development of such schemes should, however, await a comprehensive investigation of the overall vehicle control system. This should include realistic selections of thruster sizes and locations based on consideration of both attitude control operations and stationkeeping strategies.

Acknowledgment

The author would like to thank Messrs. A. Ramos and B. Free and Dr. G. Gordon of COMSAT Labs for their assistance in this study.

References

[1] M. H. Kaplan, "All-Electric Thruster Control of a Geostationary Communications Satellite which Employs Narrow-Beam Antennas," *AIAA 9th Electric Propulsion Conference*, Bethesda, Maryland, April 1972, AIAA Paper No. 72-436.

[2] H. L. Mork, "Synthesis and Design of a Gimballed Reaction Wheel Attitude Stabilization Package (GRASP)," *AIAA Guidance, Control and Flight Mechanics Conference*, Hempstead, New York, August 1971, AIAA Paper No. 71–950.

[3] H. J. Dougherty et al., "Attitude Stabilization of Synchronous Communications Satellites Employing Narrow-Beam Antennas," *Journal of Spacecraft and Rockets*, Vol. 8, No. 8, August 1971, pp. 834-841.

[4] M. G. Lyons et al., "Double Gimballed Reaction Wheel Attitude Control System for High Altitude Communications Satellites," *AIAA Guidance, Control and Flight Mechanics Conference*, Hempstead, New York, August 1971, AIAA Paper No. 71–949.

Marshall H. Kaplan is Associate Professor of Aerospace Engineering at the Pennsylvania State University and a consultant to COMSAT Laboratories on satellite dynamics and control. His teaching and research interests include astrodynamics, spacecraft autopilots, propulsion, and planetary gravity analysis. Recently, he has published papers on the use of electric thrusters for complete satellite control, retrieval of spinning objects from orbit, and modeling of the lunar gravity field. 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.

Index: Intelsat IV, communications satellites, spacecraft bearings, stability.

Investigations of the Intelsat IV bearing and power transfer assembly

C. J. PENTLICKI

Abstract

The bearing and power transfer assembly (BAPTA) on the INTELSAT IV spacecraft is a functional element of the structure, the power distribution system, and the telemetry and attitude control systems. It despins the communications antenna system so that the antennas are continuously earth oriented. Its reliable performance is crucial to the spacecraft mission.

Orbital experience with the BAPTA on the first INTELSAT IV, particularly the anomalous platform pointing error and the means by which the error source was identified, is described. The results of the experimental program undertaken to explain the causative phenomenon, bearing retainer instability, are discussed, and means of moderating the retainer's propensity for instability are suggested.

Introduction

INTELSAT IV is generically a dual-spin spacecraft. A large part of the spacecraft mass spins at approximately 50 rpm, thus developing angular momentum that stabilizes the attitude of the spacecraft. A means of continuously pointing the antennas at the earth while the rest of the spacecraft

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.

rotates has been devised and demonstrated. This technique requires a device to service the interface between the spinning body and the earth pointing antenna system. This rotary interface is the bearing and power transfer assembly (BAPTA) on INTELSAT IV.

The BAPTA is a functional part of several systems. As a part of the control system, it accurately points the antenna platform toward the earth by counterspinning it at a rate identical to that of the drum. In addition, since the electric power generation and storage system is located on the spinning drum, while the primary power consumers are located on the other side of the rotating interface on a despun platform, the BAPTA transfers power across the interface by using slip rings from the drum to the platform. Signals, including telemetry and command, must also cross this interface by rotary transformer because all communications equipment is located on the despun platform, while sensors, telemetry data sources, and commandable elements are located on the spinning drum. Finally, the BAPTA is a series element in the spacecraft structure and is required to accept the launch and orbital loads of the entire despun section.

Relevant design features of the BAPTA are shown in Figure 1; a detailed discussion of the various elements of the BAPTA can be found in Reference 1. This paper is primarily concerned with the anomalous flight performance of the BAPTA as a part of the control system.

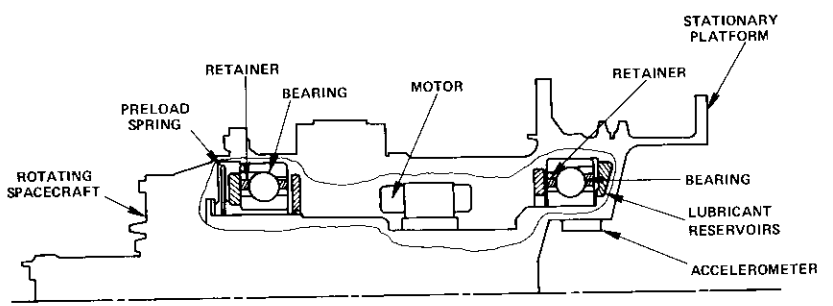


Figure 1. INTELSAT IV *BAPTA Configuration*

Platform pointing

The BAPTA, as a functional element of the attitude control system, points the despun platform and antenna farm toward the earth by adjusting the rate of platform counterspin. The earth reference with which the platform reference is compared may be derived from several sources including the spacecraft earth sensors. The BAPTA uses a brushless DC motor with a rate pulse generator and a resolver in conjunction with the despin control electronics (DCE) to accomplish this function.

The BAPTA's maximum deliverable torque at 50 rpm is about 1.74 N·m (250 oz-in). The average BAPTA running torque, which consists principally of bearing running torque and a small amount of torque from the slip rings, is on the order of 0.134 N·m (19 oz-in); hence, there is a 13 to 1 torque margin. Although such torque margins define the tolerance of the device to slowly increasing torque requirements, such as those arising from bearing wear, they do not characterize its response to rapid torque variations such as bearing torque noise.

Despun platform errors induced by torque noise can be evaluated only by studying the servoloop response and the effects of platform inertia. (The BAPTA's response is shown in Figure 30 of Reference 2.) A rate pulse generator, consisting of two permanent magnets mounted 180° apart on the BAPTA shaft and two coils placed on the housing, produces a voltage from which the housing position and rotation rate are derived. The actual platform direction is compared with the required direction, and appropriate motor torque commands are generated in the DCE so that the platform rate can be adjusted to reduce the error. A complete discussion of the DCE is also included in Reference 2.

Bearing and lubricant

The BAPTA has a design life requirement of seven years at approximately 51 rpm, or 1.9×10^8 revolutions. The bearing performance is critical to the BAPTA lifetime. The type of bearing application necessary for despun antennas is unusual in that a large-bore bearing with a high load capacity is operated at very low loads and must have low torque and torque noise for the mission life. In addition, the bearing operates in a vacuum with a limited supply of a special liquid lubricant.

The bearing system operates with a thin oil film on the order of 0.1 μm between balls and raceway. This lubricant film is sufficient to separate the metal parts under actual load conditions. The adequacy of the film can be expressed in terms of the ratio of the film thickness to the height of surface roughness of the balls and races. Low surface distress will occur when the metal asperities are separated by thick films. However, as the ratio of film thickness to asperity height approaches unity, surface distress may increase; as it nears zero, gross wear may be encountered.

Under thin film conditions, the high-pressure additive in the lubricant becomes important. The thickness of the lubricant film is determined by

bearing geometry, speed of the rolling elements, and viscosity of the lubricant. The BAPTA bearings operate at 51 rpm with a film ratio on the order of 1.6 on the inner raceway at 90°. The bearing surface finish is closely controlled so that a finish in the region of 0.05 0.07 μm is achieved. The finish permits adequate film protection at the BAPTA operating speeds and temperatures. (A 1-year life test conducted on a BAPTA showed no discernible bearing wear.) The orbital bearing loads are primarily caused by the preload (about 27 kg or 59.4 lb); thus, the fatigue life of the BAPTA bearing far exceeds the mission requirements.

An accelerometer mounted on the shaft very near the inner race of the upper bearing provides information which is telemetered to the ground for analysis. This accelerometer provides prelaunch BAPTA quality control, and it also permits monitoring of the BAPTA in orbit. In addition, data from four temperature sensors and one motor current sensor serve as indicators of BAPTA operation.

Flight experience

In the first INTELSAT IV, launched on January 25, 1971, the BAPTA performed perfectly. Motor current, platform pointing error, and BAPTA accelerometer noise were all as expected. The spin rate was 51 rpm.

However, during the first eclipse in March, the platform pointing error increased. For several reasons, the BAPTA was one of the suspected sources. First, transient torque increases of about 40 percent would be consistent with the pointing error experienced. Second, if the BAPTA temperature is not restricted by using heaters, it will vary significantly during eclipse, and it was determined that the peak pointing error appeared to vary inversely with temperature.

The use of heater elements mounted on the BAPTA made it possible to derive the temperature/error correlation shown in Figure 2. Telemetered accelerometer data from the flight unit were acquired by the Andover, Maine, earth station and analyzed at COMSAT Laboratories. These data showed intermittent periods of increased noise activity. The occurrence of the noise appeared to be consistent with the occurrence of the pointing anomalies. A real-time analysis performed shortly thereafter confirmed that the increased noise activity began just prior to increased pointing error. The most active region of the accelerometer frequency spectrum during pointing error increases was found to be centered near 800 Hz. Figure 3 shows the amplitude of the 800-Hz content of the accelerometer noise and the pointing error. This phenomenon was called "groan," which is generally an apt description of the sound.

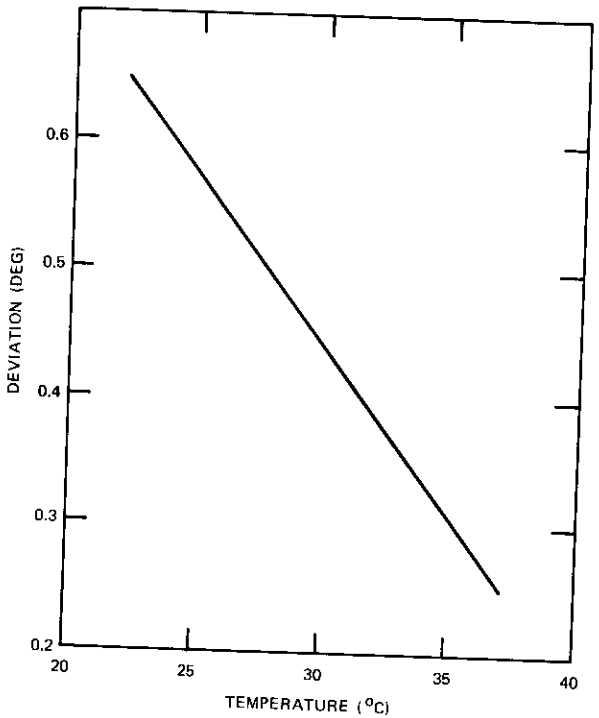


Figure 2. *Peak Platform Deviation vs Temperature*

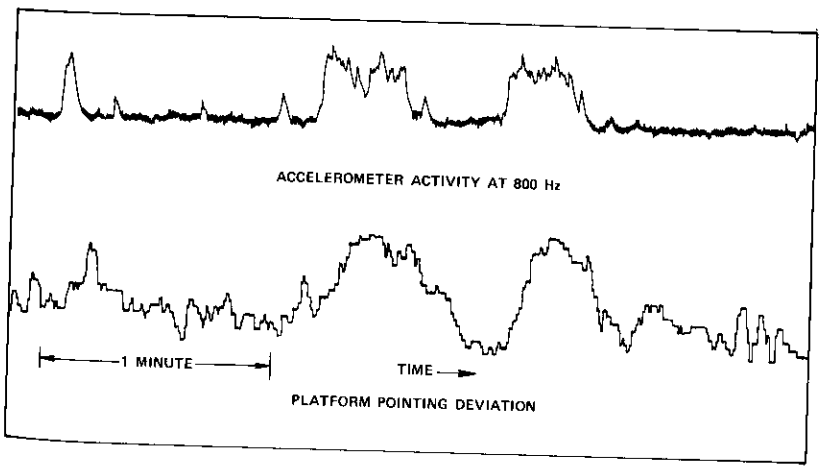


Figure 3. *800-Hz Accelerometer Noise Platform Pointing Deviation*

Test program

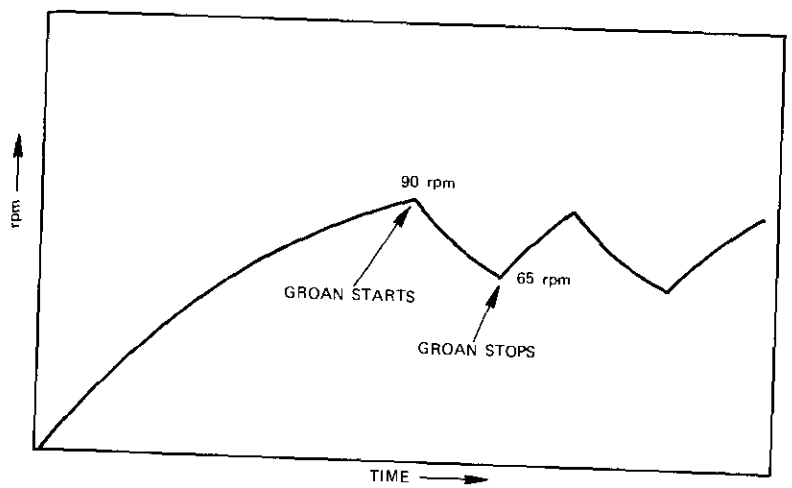
A test program intended to identify the noise source in the BAPTA and its relationship to torque increases which were consistent with pointing errors in the flight model was set up. The laboratory BAPTA operated normally throughout its nominal speed range of 60 rpm ± 15, but during one test up to 90 rpm, a noise was heard and the speed dropped off. The BAPTA was being run with constant motor current; hence the speed decrease was caused by a torque increase. As the speed decreased, the noise stopped (at about 65 rpm), whereupon the speed increased until noise developed; the cycle then repeated in a continuous fashion.

Figure 4 illustrates the speed and torque characteristic. Note the bi-valued torque characteristics at speeds within the range from 65 to 90 rpm; the particular speed at which these characteristics occur is determined by the presence or absence of pre-existing groan. Listening comparisons with the noise from the flight unit BAPTA supported the contention that the laboratory BAPTA was experiencing the same problem experienced by the flight system. Although the noise spectra of each device did not agree precisely, the discrepancies were attributed to differences in the accelerometer mountings. The flight unit was most active around 800 Hz, while the laboratory unit's noise activity occurred between 250 and 400 Hz. These differences can be attributed to various structural resonances excited by the source of the torque anomalies. Hence, they do not represent a fundamental characteristic of the source itself.

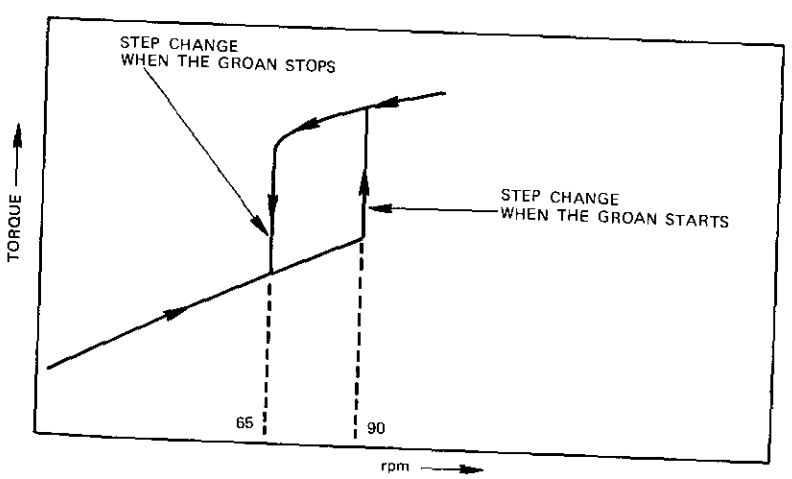
During simultaneous tests conducted on a BAPTA bearing, it was found that, at certain operating speeds, a noise was produced whose characteristics were similar to those of the BAPTA's noise. The noise output coincided with a bearing torque increase. The value of this torque increase was consistent with that attributed to the laboratory BAPTA and the flight unit during groan. This torque increase was accompanied by large amplitude motions of the ball riding retainer.

Temperature chamber tests of the laboratory BAPTA and the individual bearing showed that the speed threshold for the occurrence of groan varied with temperature, thus implicating the lubricant and furthering the correlation with flight experience.

The retainer motions were reminiscent of the gyro bearing instabilities reported by Kingsbury [3] at much higher speeds. To define the retainer motions, the test bearing's retainer was equipped with a thin aluminum ring that would permit proximeters to discern displacements of the retainer. The test arrangement is shown in Figure 5. At very low speeds,



A. SPEED VARIATIONS – CONSTANT MOTOR CURRENT



B. RUNNING TORQUE HYSTERESIS

Figure 4. *Characteristics of Laboratory BAPTA*

the motions showed low-frequency characteristics and were attributed to "walking" of the retainer on the balls. As the speed increased, a relatively pure high-frequency motion developed, but there was no discernible acoustic output or significant torque increase. At even higher speeds, the retainer "broke" into high-frequency oscillations coupled with a step torque increase and audible noise. This latter condition, identified as groan, is clearly preceded by an abnormal retainer motion that was called "wobble."

The retainer's gross motion has superimposed upon it a reverse coning motion. Figure 6 shows data describing the three states of the retainer as they affect torque and retainer displacement. Since the principal frequency components of the retainer's motion during groan were predominant in the bearing torque noise, it became clear that the retainer movement was responsible for the torque increase.

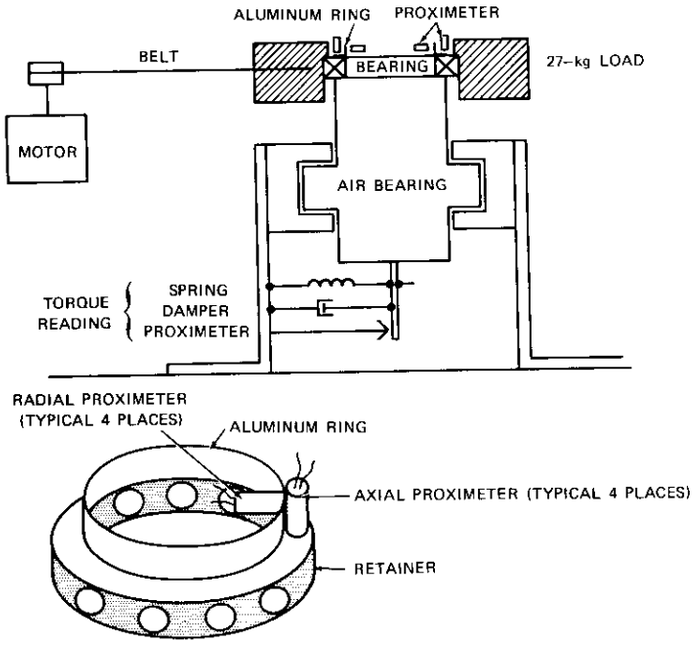


Figure 5. *Bearing Test Setup*

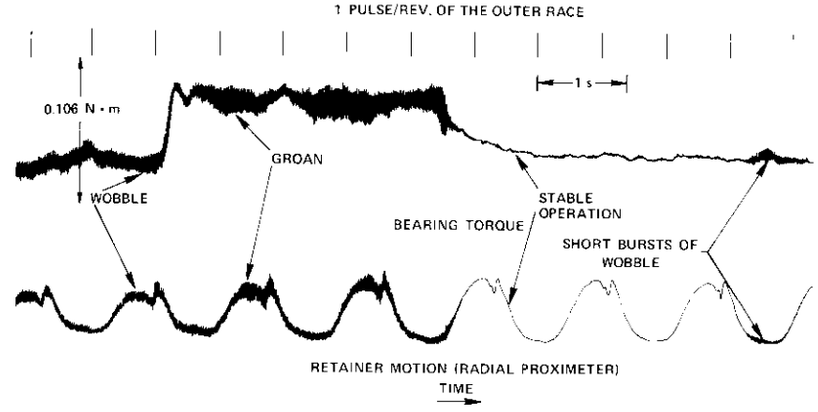


Figure 6. *Wobble, Groan, and Stable Conditions*

Factors influencing retainer stability

It should be noted that retainer instability was not easily reproduced. Relatively large numbers of flight bearings underwent acceptance tests during which no retainer instability was reported, and several flight retainers were laboratory tested without evidence of instability. This phenomenon, which obviously required some special conditions, apparently resulted from a combination of several factors.

The lubricant viscosity was an important factor because a small increase in temperature (~6°C or ~10°F) resulted in a reduction of the groan. Other tests showed that copious quantities of lubricant inhibited groan, while sparse amounts increased instability. Operating speed was also found to be a significant parameter; below a certain speed threshold, instability did not occur, and above that threshold, the amount of time during which the retainer was unstable increased with speed. Small manufacturing differences in the geometry of the retainer were possibly influential in producing the large differences in performance encountered with various retainers. Moreover, running time was apparently relevant because INTELSAT IV F-2 operated uneventfully for several weeks before the instability developed. The retainer was of course being driven by the friction force developed at the ball/pocket interface, and all of these factors affected that friction force. The use of much greater amounts of lubricant than were present in the flight system appeared to provide sufficient viscous damping to inhibit instability.

Adjustable retainer tests

The manufacturing drawing showed tolerances in the geometry of the retainer that could conceivably affect performance. A series of phenolic retainers, manufactured with controlled variations in geometry, was intended to reproduce the instability and thereby identify the principal contributor. Yet this approach was not conclusive, since no retainer developed sufficient instability to be considered a reliable indicator. Reference 3 indicates that symmetrical perfection in retainer geometry may induce instability; therefore, it was decided to use an "adjustable" retainer, which would permit better control over the geometry and make study of the problem easier. Figure 7 is a drawing of the actual retainer, while Figure 8 is a photograph of the adjustable retainer. The phenolic ball pockets are mounted in a threaded carrier that allows the over-the-ball dimension to be established at will. The basic retainer structure is magnesium and its total weight is slightly more than that of the original phenolic retainer.

Tests to evaluate the effect of a single retainer geometry parameter were undertaken. As shown in Figure 7, the only dimension varied was the over-the-ball dimension, which was varied within the tolerance permitted by the manufacturing drawing. This dimension was expected to be the most influential item, and it also incorporated the effect of the ball pocket cone angle on retainer clearance. The retainer configurations were all symmetrical; i.e., the pockets for the eight ball pairs were adjusted so that variation in over-the-ball dimensions would not exceed 2.54×10^{-5} m (0.001 in.).

The performance of each retainer configuration at 50, 100, and 250 rpm was observed by recording histograms of the bearing torque and the frequency of retainer movement. Speed was used as a variable because past experience had shown that a retainer which is apparently stable at a particular speed can become unstable at a higher speed. Although several tests of symmetrical retainers were conducted to survey the clearance band, only three examples will be reported here. For these examples, the three geometrical conditions were maximum, minimum, and intermediate clearance, as permitted by the manufacturing drawing.

The retainer with the maximum clearance showed the lowest torque of those tested, i.e., less than 1.4×10^{-2} N·m (2 oz-in) at 50 rpm. The frequency of the motions exhibited infrequent excursions to 50 Hz at 250 rpm. This geometry qualifies as stable. As the retainer's clearance was reduced (over-the-ball dimensions increased), the motion of the retainer developed the characteristic ~50-Hz wobble frequency. For the retainer having an intermediate clearance, there was significant groan at 50 rpm,

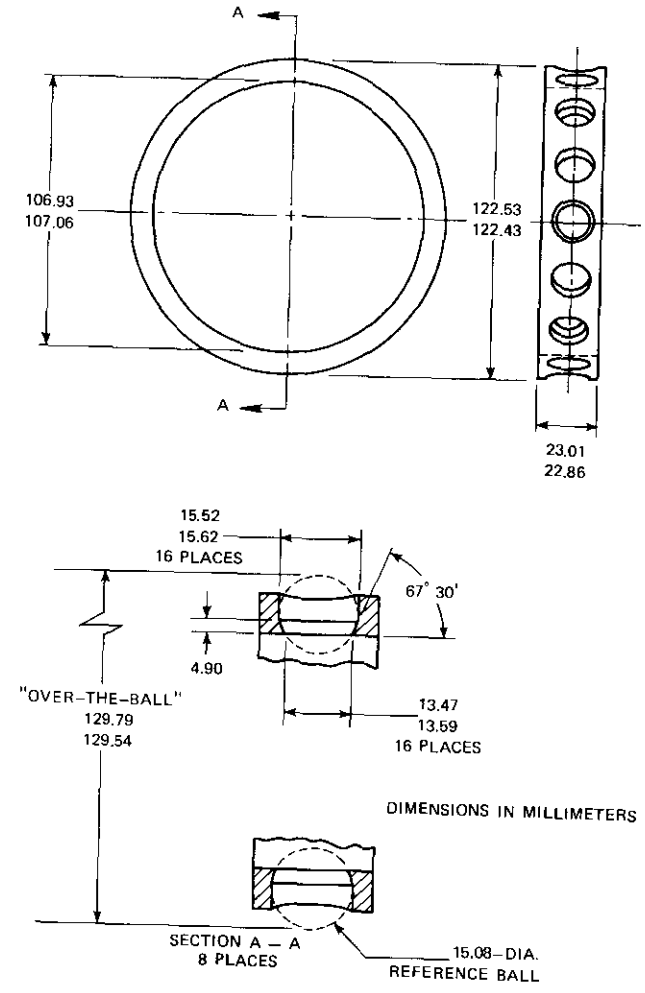


Figure 7. *Flight Retainer*

and the percentage of time during which the retainer was unstable increased with increasing speed. At 250 rpm, the bearing torque varied from 4×10^{-2} to 10.6×10^{-2} N·m (5.6–15 oz-in), and the frequency of the retainer motions reached 120 Hz.

The retainer with the minimum clearance resulted in high bearing torque (4.6×10^{-2} N·m or 6.5 oz-in at 50 rpm) but with so little clearance that

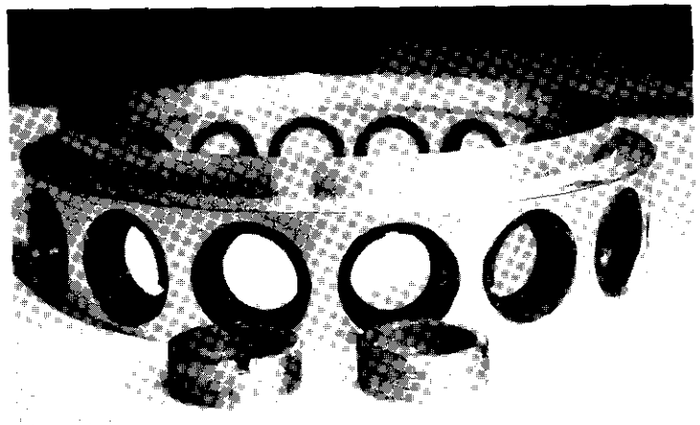


Figure 8. *Adjustable Retainer*

the retainer was stable at 50 rpm; in fact, the frequency of the motion never exceeded 18 Hz. Only at 250 rpm did groan develop. Therefore, although this retainer is stable, it requires substantial driving torque, resulting in the possibility of excessive ball pocket wear. The tests showed that the propensity to groan was dependent upon the retainer clearance. These retainers were, in practical terms, perfectly symmetrical and therefore did not simulate the greater geometrical imperfections occurring in flight models.

The effect of variations in the clearance of ball pairs in the retainer was also evaluated so that the dimensional scatter encountered in flight retainers could be modeled more representatively. The baseline dimensions used were those of the retainer with the maximum clearance, but the clearance for two ball pairs, 90° apart, was reduced by 7.5×10^{-5} m (0.003 in.). This clearance reduction is approximately one-third of the way toward the minimum clearance.

It should be recalled that the retainer with the maximum clearance is stable, but when it undergoes clearance reduction in two ball pairs, it becomes bistable; that is, it may show stable and unstable states while operating at the same speed. At 50 rpm this retainer was stable, but at 100 rpm two states were possible, as shown in Figure 9.

This performance is reminiscent of the laboratory BAPTA's bistable states shown in Figure 4. This retainer has stable and unstable states at 100 rpm, but at 250 rpm it demonstrates different states of instability that

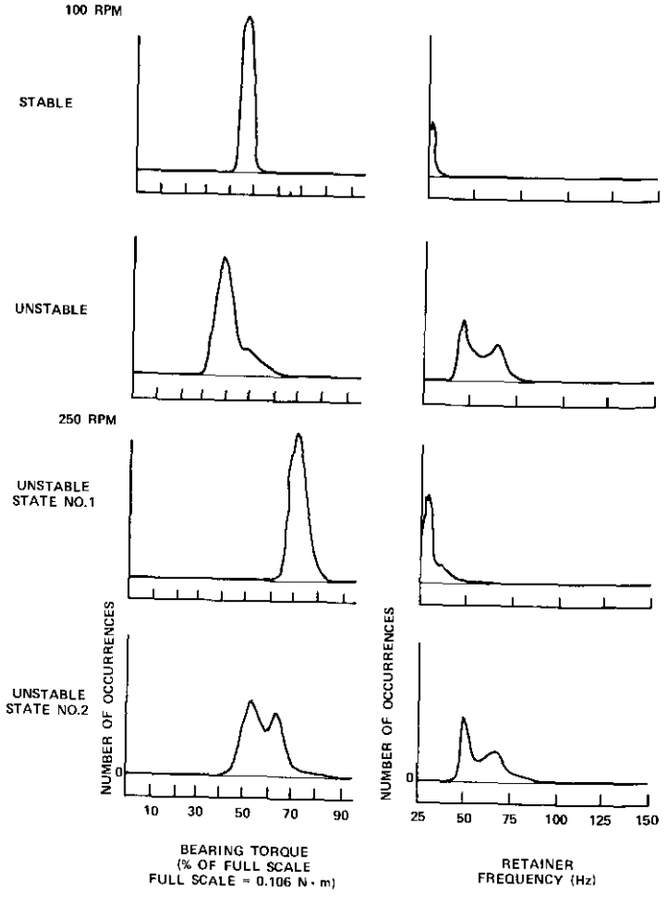


Figure 9. *Bistable Retainer*

differ markedly in character. The principal torque value is lower for the unstable operating condition than for the stable condition, but the spread between minimum and maximum torques is larger.

To validate the results of the testing performed with the adjustable retainer, it was considered prudent to manufacture a phenolic retainer that would groan. The dimensions for this retainer were predicated upon the test data, and tolerances were very carefully controlled in manufacture. The retainer was unstable at and above 50 rpm.

Conclusion

Some of the early flight experience with the first INTELSAT IV BAPTA has been described, and the highlights of a test program that identified the source of a torque anomaly in the unit have been presented. This torque anomaly was caused by a bearing retainer instability. The degree of instability is affected by lubricant viscosity, lubricant quantity, and bearing speed. It has been demonstrated that the occurrence of retainer instability is strongly dependent upon the clearance that is available for retainer motion. The divergence in the degree of stability experienced with various retainers is primarily attributable to variations in the geometrical symmetry and clearance of the retainers. The laboratory data have provided some guidance for the manufacture of retainers having little propensity toward instability.

References

[1] J. Ouellette, Bearing and Power Transfer Assembly Section of "The INTELSAT IV Spacecraft," E. T. Jilg, editor, *COMSAT Technical Review*, Vol. 2, No. 2, Fall 1972, p. 313.

[2] L. B. Ricks, Despin Control System Section of "The INTELSAT IV Spacecraft," E. T. Jilg, editor, *COMSAT Technical Review*, Vol. 2, No. 2, Fall 1972, p. 295.

[3] E. P. Kingsbury, "Torque Variations in Instrument Ball Bearings," *American Society of Lubrication Engineers Transactions*, Vol. 8, No. 4, October 1965.

Acknowledgment

The author would like to thank François Grassot, who performed much of the early and definitive experimental work, and D. M. Van Der Loo, who conducted the bearing tests. Advice and helpful suggestions were provided by Drs. Harrington and Curtis of Harrington, Davenport and Curtis Inc.

Mr. Pentlicki, after graduating from the Newark College of Engineering, worked at Bell Telephone Laboratories from 1961 to 1967. He participated in the mechanical design, development, and fabrication of the Telstar communications satellites. In 1965 he was engaged in design and manufacture of portions of the Nike X control computer. He joined COMSAT Laboratories in 1967 and is manager of the Mechanical Devices Branch working on advanced mechanical systems for communications satellites.

Index: communications satellites, nonlinear systems, intermodulation, Fourier analysis, computer programs.

Time-domain analysis of intermodulation effects caused by nonlinear amplifiers

J. C. FUENZALIDA, O. SHIMBO, AND W. L. COOK

Abstract

Nonlinear amplifiers are widely used in microwave radio communications. If the input signal to such a device exhibits a time-varying envelope, then the output signal is distorted. This situation arises in communications satellites when a number of angle-modulated carriers share a common nonlinear amplifier so that the output contains intermodulation products in addition to the carriers.

This paper uses a time-domain analysis to express the baseband distortion caused by intermodulation in angle-modulated carriers. The combined effects of amplitude nonlinearity and AM-PM conversion are considered, and the particular case of an input consisting of angle-modulated carriers and a band of Gaussian noise is treated in detail.

The physical interpretation of the analytical results and the computational method used to implement these results are described. The computational method is the basis for a flexible, user-oriented computer program which provides considerable generality while minimizing computational requirements. The computer program is used to investigate several problems of practical interest for which analytical results have been previously unobtainable.

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

When a nonlinear amplifier is used simultaneously by a number of carriers, intermodulation products are generated which cause distortion in the signals. Many studies dealing with calculation of the levels of intermodulation products at the output of the amplifier have been reported [1]–[6]; the most complete to date is the general analysis by Shimbo [5]. Chitre and Fuenzalida [7] have shown that the baseband distortion caused by such intermodulation products cannot always be calculated from the power spectral density of the signal at the output. The analysis of Reference 5 has therefore been extended to permit the evaluation of baseband distortion in the time-domain representation.

This paper presents a unified theory which combines the effects of amplitude nonlinearity and AM/PM conversion for the memoryless case. The analysis is performed initially for an input consisting of arbitrary bandpass signals. The general expressions are then applied to obtain both the RF and the angle-demodulated power spectra for an input consisting of a band of Gaussian noise in addition to multiple angle-modulated carriers.

The steps in the analytical derivation may be summarized as follows. First, the nonlinearity is modeled. Then, an analysis is performed to obtain a time-domain representation of the output of the nonlinear device, as well as general expressions for the power spectrum of the angle-demodulated output. It is shown that the expression for the output power spectrum reported in Reference 5 can be derived from these results.

The theoretical development is followed by a description of a computer program, based on these results, which has been found to be a useful tool in the determination of intermodulation spectra and baseband distortion in multicarrier systems. The design objectives of the program are discussed, with emphasis on the important computational techniques. Results of the calculations are then given for several interesting problems for which a rigorous solution was not previously available.

Modeling and Analysis

The essential features of the model used in the analysis and the important results generated by the analysis are summarized in the following subsections. Detailed mathematical derivations of the time-domain representation of the output of a nonlinear device and the resulting baseband and RF power spectra are included in Appendix A.

Mathematical model

The nonlinear device is characterized in terms of the input and output envelopes, following the mathematical model of Reference 5. Two nonlinear functions are necessary to completely describe the properties of such a nonlinearity: the nonlinear amplitude and phase functions. For an input consisting of a single unmodulated carrier, the output of the nonlinear amplifier contains the fundamental frequency as well as higher harmonics. The laboratory measurements necessary to characterize the nonlinearity involve only the fundamental component. Therefore, the envelope characterization is ideally suited for the purposes of this analysis and physically more meaningful than instantaneous voltage characterizations.

Assume that the general representation of the input bandpass signal is

$$\begin{aligned} e_i(t) &= Re\ \{V(t)\ \exp\ (j\omega_o t)\} \\ &= Re\ \{\rho(t)\ \exp\ [j\omega_o t + j\Theta(t)]\} \end{aligned} \tag{1}$$

where

$$\begin{aligned} V(t) &= \rho(t)\ \exp\ [j\Theta(t)] \\ \rho(t) &= \text{amplitude of } V(t) \\ \Theta(t) &= \text{angle of } V(t) \\ \omega_o &= \text{midband angular frequency.} \end{aligned}$$

The fundamental signal output in a single-carrier test can be represented by

$$e_o(t) = Re\ \{g(A)\ \exp\ [j\omega_o t + jf(A)]\} \tag{2}$$

for the cosine wave input $Re\ \{A\ \exp\ [j\omega_o t]\}$, where $g(A)$ and $f(A)$ are the output amplitude and phase functions, respectively. It is assumed that $g(A)$ and $f(A)$ are independent of frequency.* Therefore, the bandpass output caused by an arbitrary bandpass input signal [as given in equation (1)] can be represented by

$$e_o(t) = Re[g\{\rho(t)\}\ \exp\ \{j\omega_o t + jf[\rho(t)] + j\Theta(t)\}]\ . \tag{3}$$

*This case is sometimes referred to as "memoryless."

The output can also be written in terms of a complex envelope gain function:

$$e_o(t) = Re\,\{G[\rho(t)]\,V(t)\exp[j\omega_o t]\}$$

where

$$G[\rho(t)] = \frac{g[\rho(t)]}{\rho(t)}\exp\{jf[\rho(t)]\} \tag{4}$$

In this paper a bandpass envelope characterization of the nonlinear device is used since the functions $g(\rho)$ and $f(\rho)$ can be directly measured in the laboratory. The instantaneous voltage transfer function (including all harmonics) cannot be uniquely defined on the basis of envelope measurements. However, this transfer function is related to the equivalent bandpass envelope transfer function by a Chebychev transformation [8]. In the absence of AM/PM conversion, Chebychev transformation of the real transfer function $g(\rho)$ is required; otherwise the bandpass envelope transfer function is complex. The real and imaginary components can be transformed separately, resulting in two instantaneous voltage transfer functions in quadrature, as reported in Reference 6.

Time-domain output of the nonlinear device

INPUT CONSISTING OF NARROWBAND BANDPASS SIGNALS

Consider the case in which the input to the nonlinear device is formed by the sum of m narrowband bandpass signals (as defined in Reference 9):

$$\begin{aligned} e_i(t) &= Re\sum_{l=1}^{m} A_l(t)\exp[j\omega_o t + j\theta_l(t)] \\ &= Re\,\{\rho(t)\exp[j\omega_o t + j\Theta(t)]\} \end{aligned} \tag{5}$$

If the bandpass output is represented by

$$e_o(t) = Re\,[g\{\rho(t)\}\exp\{j\omega_o t + j\Theta(t) + jf[\rho(t)]\}] \tag{6}$$

it is shown in Appendix A [equation (A10)] that $e_o(t)$ can also be expressed as

$$\begin{aligned} e_o(t) = Re\Bigg\{&\exp[j\omega_o t]\sum_{\substack{k_1,k_2,\ldots,k_m=-\infty \\ (k_1+k_2+\ldots+k_m=1)}}^{\infty}\exp\left[j\sum_{l=1}^{m}k_l\theta_l(t)\right] \\ &\cdot M(k_1, k_2, \ldots, k_m)\Bigg\} \end{aligned} \tag{7}$$

The condition $\sum_{l=1}^{m} k_l = 1$ is a consequence of the bandpass representation of the output. The complex amplitude $M(k_1, k_2, \ldots, k_m)$, which can be expressed as [equation (A12)]

$$\begin{aligned} &M(k_1, k_2, \ldots, k_m) \\ &= \int_0^\infty \gamma\left\{\prod_{l=1}^{m} J_{k_l}[A_l(t)\,\gamma]\right\} d\gamma \int_0^\infty \rho g(\rho)\exp[jf(\rho)]\,J_1(\gamma\rho)\,d\rho \end{aligned} \tag{8}$$

can be calculated for a number of series expansions of $g(\rho)\exp[jf(\rho)]$, as described in Reference 5. The Bessel function expansion

$$g(\rho)\exp[jf(\rho)] = \sum_{s=1}^{L} b_s J_1(\alpha s\rho) \tag{9}$$

where b_s are complex coefficients, leads to a particularly simple result.

Berman and Podraczky [3] used a Fourier series expansion with real coefficients of the instantaneous voltage transfer function. The corresponding single-carrier bandpass output is given by a Bessel function series. More recently Kaye et al. have introduced complex coefficients to account for the effects of AM/PM conversion. Then, as shown in Appendix A, introducing equation (9) into equation (8) yields

$$M(k_1, k_2, \ldots, k_m) = \sum_{s=1}^{L} b_s \prod_{l=1}^{m} J_{k_l}[\alpha s A_l(t)] \quad . \tag{10}$$

In the particular case of an input consisting solely of angle-modulated carriers, the bandpass output of the nonlinearity is also formed solely by angle-modulated signals. Each component (carriers as well as intermodulation products) is described by a set of integers $k_1, k_2, \ldots, k_m$. The amplitude (which is constant) is $M(k_1, k_2, \ldots, k_m)$, and its angle modulation is given by $\sum_{l=1}^{m} k_l\theta_l(t)$.

In the general case of an input consisting of narrowband bandpass signals, $M(k_1, k_2, \ldots, k_m)$ is a function of time. Consequently, the set of integers k_l is no longer sufficient to describe the intermodulation products. In the next subsection the case of an input consisting of angle-modulated signals and a band of noise is analyzed, and it is shown that an additional parameter is required to account for the effects of the amplitude variations of the noise signals.

INPUT CONSISTING OF ANGLE-MODULATED CARRIERS AND NOISE

Consider the case of $m-1$ angle-modulated carriers plus noise. The input to the nonlinear device is

$$\begin{aligned} e_i(t) &= Re\left\{\sum_{l=1}^{m-1} A_l \exp\left[j\omega_o t + j\theta_l(t)\right] + \left[N_c(t) + jN_s(t)\right]\right. \\ &\quad \left.\cdot \exp\left[j\omega_o t + j\omega_m t\right]\right\} \\ &= Re\left\{\sum_{l=1}^{m-1} A_l \exp\left[j\omega_o t + j\theta_l(t)\right]\right. \\ &\quad \left. + A_m(t) \exp\left[j\omega_o t + j\theta_m(t)\right]\right\} \end{aligned} \tag{11}$$

where $\theta_l(t) = \omega_l t + \phi_l(t) + \lambda_l, \quad l = 1, 2, \ldots, m-1$

$$\theta_m(t) = \omega_m t + \tan^{-1}\frac{N_s(t)}{N_c(t)}$$

$$A_l = \text{constant}, \quad l = 1, 2, \ldots, m-1$$

$$A_m(t) = \sqrt{N_c^2(t) + N_s^2(t)} \quad . \tag{12}$$

The output, $e_o(t)$, is therefore given by

$$e_o(t) = Re\left\{\exp\left[j\omega_o t\right] \sum_{\substack{k_1,k_2,\ldots,k_m=-\infty \\ (k_1+k_2+\ldots+k_m=1)}}^{\infty} \exp\left[j\sum_{l=1}^{m-1} k_l\theta_l(t)\right]\right.$$

$$\left.\cdot\, M(k_1, k_2, \ldots, k_m) \exp\left[jk_m\theta_m(t)\right]\right\} \quad . \tag{13}$$

In the absence of a noise signal at the input of the device, the output, $e_o(t)$, consists of the angle-modulated carriers and intermodulation products which also have the properties of angle-modulated carriers. With the introduction of noise at the input, the output may be divided into two categories:

a. the original output components with modified complex amplitudes, and

b. additional intermodulation components caused by the introduction of noise.

These two classes of output are represented by $e_S(t)$ and $e_N(t)$, respectively:

$$e_o(t) = e_S(t) + e_N(t) \quad . \tag{14}$$

To obtain $e_S(t)$, it is necessary to take the expected value of $e_o(t)$ on $N_c(t)$ and $N_s(t)$. For the particular case of Gaussian noise whose rms power is $R(0)$, this yields

$$e_S(t) = Re\left\{\exp\left[j\omega_o t\right] \sum_{\substack{k_1,k_2,\ldots,k_{m-1}=-\infty \\ (k_1+k_2+\ldots+k_{m-1}=1)}}^{\infty} \exp\left[j\sum_{l=1}^{m-1} k_l\theta_l(t)\right]\right.$$

$$\left.\cdot\, M_S(k_1, k_2, \ldots, k_{m-1})\right\} \tag{15}$$

where, in genera',

$$M_S(k_1, k_2, \ldots, k_{m-1}) = \int_0^\infty \gamma \prod_{l=1}^{m-1} J_{k_l}(\gamma A_l) \exp\left[-\frac{\gamma^2}{2} R(0)\right] d\gamma$$

$$\cdot \int_0^\infty \rho g(\rho) \exp\left[jf(\rho)\right] J_1(\gamma\rho)\, d\rho \quad . \tag{16}$$

For the Bessel function characterization, the carrier and intermodulation product levels are given by

$$M_S(k_1, k_2, \ldots, k_{m-1}) = \sum_{s=1}^{L} b_s \exp\left[-\frac{\alpha^2 s^2}{2} R(0)\right] \prod_{l=1}^{m-1} \mathrm{J}_{k_l}(\alpha s A_l) \tag{17}$$

where $R(\tau)$ is the autocorrelation function of $N_c(t)$ or $N_s(t)$. As suggested in equation (17), the effects of noise in this case may be introduced merely by modifying each complex coefficient, b_s, with the factor

$$\exp\left[-\frac{\alpha^2 s^2}{2} R(0)\right] \quad .$$

An equivalent expression for $e_N(t)$ in which all intermodulation products can be identified was not derived here. Nevertheless, in the following expressions for the *RF* and baseband power spectra caused by intermodulation, the individual products can be identified. Identification of these products is necessary to incorporate the effects of noise in the computational algorithm.

RF intermodulation power spectrum

The general expression for the *RF* intermodulation power spectrum is given in Reference 5. For the particular case of a Bessel function expansion of the nonlinearity, the autocorrelation function of $e_o(t)$ is

$$\begin{aligned}&\text{avg}\,[e_o(t)\,e_o(t+\tau)]\\&=\frac{1}{2}\,Re\sum_{k_1,k_2,\ldots,k_{m-1}=-\infty}^{\infty}\exp\left(j\omega_o\tau+j\sum_{l=1}^{m}k_l\omega_l\tau\right)\\&\cdot\ \text{avg}\left[\exp\sum_{l=1}^{m-1}jk_l\psi_l\right]H(k_1,k_2,\ldots,k_{m-1})\end{aligned}\tag{18}$$

where

$$\psi_l=\phi_l(t)-\phi_l(t+\tau)\quad,$$

and

$$\begin{aligned}&H(k_1,k_2,\ldots,k_{m-1})\\&=\sum_{s=1}^{L}\sum_{p=1}^{L}b_sb_p^*\left[\prod_{l=1}^{m-1}J_{k_l}(\alpha sA_l)\right]\left[\prod_{l=1}^{m-1}J_{k_l}(\alpha pA_l)\right]\\&\cdot\ \exp\left[-\frac{\alpha^2(s^2+p^2)}{2}R(0)\right]I_{k_m}[R(\tau)\,\alpha^2sp]\end{aligned}\tag{19}$$

where $I_n(x)$ is the modified Bessel function of the first kind and nth order. In equation (18), the average on the left-hand side is taken on t, $\phi_l(t)$, $\phi_l(t+\tau)$, $N_c(t)$, $N_s(t)$, $N_c(t+\tau)$, and $N_s(t+\tau)$. The average on the right-hand side is taken on $\phi_l(t)$ and $\phi_l(t+\tau)$. In all subsequent equations, it will be assumed that $k_m = 1 - \sum_{l=1}^{m-1}k_l$, since this selection ensures that $\sum_{l=1}^{m}k_l = 1$, as required.

A series expansion of $I_n(x)$ leads to

$$I_{k_m}[R(\tau)\,\alpha^2sp]=\sum_{q=0}^{\infty}\frac{\left[\frac{1}{2}R(\tau)\,\alpha^2sp\right]^{2q+|k_m|}}{q!(|k_m|+q)!}\quad.\tag{20}$$

The term $H(k_1, k_2, \ldots, k_{m-1})$ therefore includes the noise contributions to the spectrum through $R(\tau)$ in the argument of the modified Bessel function. Rewriting makes it possible to obtain

$$\begin{aligned}H(k_1,k_2,\ldots,k_{m-1})&=\sum_{q=0}^{\infty}|N(k_1,k_2,\ldots,k_{m-1};q)|^2\\&\cdot\ [\rho_o(\tau)]^{2q+|km|}\end{aligned}\tag{21}$$

where

$$\rho_o(\tau)=\frac{R(\tau)}{R(0)}\tag{22}$$

$$\begin{aligned}N(k_1,k_2,\ldots,k_{m-1};q)&=\sum_{s=1}^{L}b_s\exp\left[-\frac{\alpha^2s^2}{2}R(0)\right]\\&\cdot\ \prod_{l=1}^{m-1}J_{k_l}(\alpha sA_l)\,T(q,|k_m|,s)\end{aligned}\tag{23}$$

and

$$T(q,|k_m|,s)=\left\{\frac{\left[\frac{1}{2}\,\mathrm{R}(0)\,\alpha^2s^2\right]^{2q+|k_m|}}{q!(|k_m|+q)!}\right\}^{1/2}\quad.\tag{24}$$

When equation (21) is introduced into equation (18), the output autocorrelation function becomes

$$\begin{aligned}&\text{avg}\,[e_o(t)\,e_o(t+\tau)]\\&=\frac{1}{2}\,Re\sum_{k_1,k_2,\ldots,k_{m-1}=-\infty}^{\infty}\exp\left(j\omega_o\tau+j\sum_{l=1}^{m}k_l\omega_l\tau\right)\\&\cdot\ \text{avg}\left\{\exp\left[\sum_{l=1}^{m-1}jk_l\psi_l\right]\right\}\sum_{q=0}^{\infty}[\rho_o(\tau)]^{2q+|k_m|}\\&\cdot\ |N(k_1,k_2,\ldots,k_{m-1};q)|^2\quad.\end{aligned}\tag{25}$$

In the absence of noise, this autocorrelation function reduces to

$$\begin{aligned}&\text{avg}\,[e_o(t)\,e_o(t+\tau)]\\&=\frac{1}{2}\,Re\sum_{\substack{k_1,k_2,\ldots,k_{m-1}=-\infty\\(k_1+k_2+\ldots+k_{m-1}=1)}}^{\infty}\exp\left(j\omega_o\tau+j\sum_{l=1}^{m-1}k_l\omega_l\tau\right)\\&\cdot\ \text{avg}\left\{\exp\left[\sum_{l=1}^{m-1}jk_l\psi_l\right]\right\}|M(k_1,k_2,\ldots,k_{m-1})|^2\end{aligned}\tag{26}$$

since

$$N(k_1,k_2,\ldots,k_{m-1};0)=M(k_1,k_2,\ldots,k_{m-1})\tag{27}$$

when

$$k_m=1-\sum_{l=1}^{m-1}k_l=0.$$

The power spectral density of the output is given by the Fourier transform of equation (18). After a term-by-term transformation is performed, the following expression for the output power spectrum is obtained:

$$S(f) = \frac{1}{2} Re \sum_{k_1,k_2,\ldots,k_{m-1}=-\infty}^{\infty} \sum_{q=0}^{\infty} |N(k_1, k_2, \ldots, k_{m-1}); q|^2 \cdot \Omega\left(k_1, k_2, \ldots, k_{m-1}; q; f - f_o - \sum_{l=1}^{m} k_l f_l\right) \quad (28)$$

where $\Omega(k_1, k_2, \ldots, k_{m-1}; q; f) = \left(\sigma_{\Sigma k_l \phi_l} \overset{2q+|k_m|}{\otimes} \sigma_n\right)(f)$

and $\sigma_n(f)$ = power spectral density of $N_c(t)$ or $N_s(t)$ normalized to unit power

$\sigma_{\Sigma k_l \phi_l}$ = low-pass equivalent power spectral density, normalized to unit power, of a carrier angle modulated by $\sum_{l=1}^{m-1} k_l \phi_l(t)$.

The convolution operator $\overset{i}{\otimes}$ is defined as follows:

$$\sigma_n(f) \overset{0}{\otimes} \sigma_n(f) = \delta(f)$$

$$\sigma_n(f) \overset{1}{\otimes} \sigma_n(f) = \sigma_n(f)$$

and $\sigma_n(f) \overset{i}{\otimes} \sigma_n(f)$ represents $i - 1$ convolutions of $\sigma_n(f)$.

It has been shown that the output power spectrum is given by the summation of components indentified by the integers k_l and q. Analogous to the noise-free case, each component can be regarded as an intermodulaiont product. The total power of each product is $\frac{1}{2} |N(k_1, k_2, \ldots, k_{m-1}; q)|^2$, the corresponding normalized spectrum is $\Omega(k_1, k_2, \ldots, k_{m-1}; q; f)$, and its center frequency is $f_o + \sum_{l=1}^{m} k_l f_l$.

The normalized spectrum of a carrier which is angle modulated by a series of baseband functions can be calculated by convolving the normalized spectra of carriers modulated by each of the baseband functions; i.e.,

$$\sigma_{\Sigma k_l \phi_l} = \sigma_{k_1 \phi_1} \otimes \sigma_{k_2 \phi_2} \otimes \cdots \otimes \sigma_{k_{m-1} \phi_{m-1}} \quad . \quad (29)$$

Baseband intermodulation power spectrum

In the case of an interfering signal which is statistically independent of the angle-modulated desired carrier, the RF spectra of both the interfering signal and the desired carrier determine the spectrum of the baseband signal generated by the interference. Most intermodulation products fall into this category, with the exception of intermodulation products that contain the desired carrier [7]. Of particular interest are those products that are modulation sidebands of the desired carrier.

The mathematical derivation of the baseband power spectrum generated by the intermodulation process is included in Appendix A. In this section, only the most significant results will be discussed.

Again consider the case of $m - 1$ angle-modulated carriers and a band of noise, as shown in equation (11). The signal at the input to the angle demodulator can be represented by three terms:

a. the angle-modulated carrier (undistorted),
b. the modulation sidebands caused by the noise, and
c. the other intermodulation components.

Without loss of generality, consider the demodulation of the first carrier. The three corresponding terms are shown in the following equation:

$$\begin{aligned} e_{o1}(t) &= Re\,\{\exp[j\omega_o + j\theta_1(t)]\, M_o\} \\ &\quad + Re\,\{\exp[j\omega_o t + j\theta_1(t)]\,[M(1, 0, \ldots, 0; t) - M_o]\} \\ &\quad + Re\left\{\exp[j\omega_o t] \sum_{\substack{k_1,k_2,\ldots,k_m=-\infty \\ (k_1+k_2+\ldots+k_m=1)}}^{\infty} \exp\left[j \sum_{l=1}^{m-1} k_l \theta_l(t)\right] \right. \\ &\quad \left. \cdot\, M(k_1, k_2, \ldots, k_m; t) \exp[jk_m \theta_m(t)]\right\} \\ &= Re\,\{\exp[j\omega_o t + j\theta_1(t)]\, M_o[1 + R(t) + jI(t)]\} \end{aligned} \quad (30)$$

where, in the case of a Bessel function expansion,

$$M_o = \sum_{s=1}^{L} b_s \exp\left[-\frac{\alpha^2 s^2}{2} R(0)\right] J_1(\alpha s A_1) \prod_{l=2}^{m-1} J_0(\alpha s A_l)$$

and

$$M(k_1, k_2, \ldots, k_m; t) = \sum_{s=1}^{L} b_s J_{km}\left[\alpha s \sqrt{N_c^2(t) + N_s^2(t)}\right] \cdot \prod_{l=1}^{m-1} J_{k_l}(\alpha s A_l) \quad . \quad (31)$$

For small intermodulation levels relative to the demodulated carrier, the detected angle can be approximated by*

*This approximation is good in all practical cases since the distortion objectives for intermodulation are low.

$$\phi_1(t) + \tan^{-1}\left[\frac{I(t)}{1 + R(t)}\right] \approx \phi_1(t) + I(t) \quad . \tag{32}$$

To evaluate the power spectral density of $I(t)$, its autocorrelation function is first determined. The two types of components for which the cross-correlation terms do not vanish are identified as $M(2 - k_1, -k_2, -k_3, \ldots, -k_m)$ and $M(k_1, k_2, \ldots, k_m)$. Particular components satisfying this condition are the $A + B - C$ and $A + C - B$ intermodulation products, which are modulation sidebands of carrier A.

The autocorrelation function can be written as

$$\text{avg}\,[I(t)\,I(t+\tau)] = R_1(\tau) + \sum_{k_1,k_2,\ldots,k_{m-1}=-\infty}^{\infty\,\prime\prime} \cdot R(k_1, k_2, \ldots, k_{m-1}; \tau) \tag{33}$$

where the components for which the cross-correlation terms do not vanish have been combined in the summation $\sum''$ [see equation (A40)], and $R_1(\tau)$ represents the effects of noise components falling on the demodulated carrier, i.e., intermodulation products of the form $N + A - N$ or $2N + A - 2N$. The other terms of particular interest are those for which $k_1 = 1$, since these terms are modulation sidebands of the demodulated carrier.

After a detailed derivation, which can be found in Appendix A, $R_1(\tau)$ and $R(1, k_2, \ldots, k_{m-1}; \tau)$ are expressed as

$$R_1(\tau) = \sum_{q=1}^{\infty} \left\{ Im\left[\frac{1}{M_o} N(1, 0, \ldots, 0; q)\right]\right\}^2 [\rho_o(\tau)]^{2q} \tag{34}$$

and

$$R(1, k_2, \ldots, k_{m-1}; \tau) = 2\sum_{q=0}^{\infty} \left\{ Im\left[\frac{1}{M_o} N(1, k_2, \ldots, k_{m-1}; q)\right]\right\}^2 [\rho_o(\tau)]^{2q+|k_m|} \cdot \exp\left[j\sum_{l=2}^{m} k_l\omega_l\tau\right] \text{avg}\left\{\exp\left[j\sum_{l=2}^{m-1} k_l\psi_l\right]\right\} \tag{35}$$

respectively. For all other components, the phase between the carrier and the intermodulation products can be ignored, and the correlation is given by

$$R(k_1, k_2, \ldots, k_{m-1}; \tau) = \sum_{q=0}^{\infty} \frac{|N(k_1, k_2, \ldots, k_{m-1}; q)|^2}{|M_o|^2} [\rho_o(\tau)]^{2q+|k_m|} \exp\left[\sum_{l=2}^{m} k_l\omega_l\tau\right] \cdot \text{avg}\left\{\exp\left[j\sum_{l=2}^{m-1} k_l\psi_l\right]\exp[j(k_1 - 1)\psi_1]\right\} \cdot \exp[j(k_1 - 1)\omega_1\tau] \quad . \tag{36}$$

The corresponding baseband power spectrum after angle demodulation is

$$S_\phi(f) = \sum_{q=1}^{\infty}\left\{Im\left[\frac{N(1, 0, \ldots, 0; q)}{M_o}\right]\right\}^2 \Omega(0, 0, \ldots, 0; q; F) + \sum_{k_2,\ldots,k_{m-1}=-\infty}^{\infty}\sum_{q=0}^{\infty} 2\left\{Im\left[\frac{N(1, k_2, \ldots, k_{m-1}; q)}{M_o}\right]\right\}^2 \cdot \Omega(0, k_2, \ldots, k_{m-1}; q; F) + \sum_{k_1,k_2,\ldots,k_{m-1}=-\infty}^{\infty}\sum_{q=0}^{\infty} \frac{|N(k_1 \neq 1, k_2, \ldots, k_{m-1}; q|^2}{|M_o|^2} \cdot \Omega(k_1 - 1, k_2, \ldots, k_{m-1}; q; F) \tag{37}$$

where

$$F = f - (k_1 - 1)f_1 - \sum_{l=2}^{m} k_l f_l \quad . \tag{38}$$

The corresponding baseband power spectrum after frequency demodulation is

$$S_f(f) = f^2 S_\phi(f) \quad . \tag{39}$$

The power spectrum of the demodulated output is formed by components which can be easily related to those appearing in the RF power spectrum [equation (28)]. However, in the power spectrum of the angle-modulated output, three distinct categories of terms can be identified:

a. *Noise intermodulation components centered on the demodulated carrier A*: $N(1, 0, \ldots, 0; q)$. These components are of the type which includes $N + A - N$ and $2N + A - 2N$ products. The spectrum shape Ω is determined by convolutions of the input noise spectrum. Only the component in phase quadrature appears in the demodulator.

b. *Modulation sidebands of the demodulated carrier A*: $N(1, k_2, \ldots, k_{m-1}; q)$. These components are of the type which includes $A + B - C$

and $A + C - B$ products. The spectrum shape Ω is determined by convolutions of the spectra of the other signals generating the product. Only the component in phase quadrature appears in the demodulator.

c. *All other components*: $N(k_1 \neq 1, k_2, \ldots, k_{m-1}; q)$. These components are of the type which includes $B + C - D$ and $2B - C$ products. Such products may also contain the desired carrier, in which case the index corresponding to the demodulated carrier in the spectrum shape Ω is $k_1 - 1$. Both the in-phase and the in-quadrature components of the product appear in the demodulator.

Computational methods

In previous sections, the pertinent aspects of the theory have been discussed. Attention is now directed to the manner in which these techniques have been implemented in a versatile, user-oriented computer program for intermodulation analysis. The design goals, as well as the means by which these goals have been achieved, will be discussed.

Program overview

The Intermodulation Analyzer Program has been designed with the following objectives in mind:

a. *Arbitary nonlinearity*. The user may enter any arbitrary set of transfer curves to determine the real and imaginary coefficients of the Bessel function expansion.

b. *Choice of input signals*. The program accepts three distinct carrier types (FM telephony, FM television, and PCM/PSK) in addition to bands of equal carriers or thermal noise.

c. *Analysis options*. Both 3rd- and 5th-order intermodulation products, as well as the resulting intermodulation spectra and baseband distortion for all carrier types may be calculated.

d. *Output options*. Intermodulation spectra and intermodulation product frequencies may be generated in printed, plotted, or punched form. A listing of intermodulation products which includes all such products, only those in a specified range, or only those exceeding a specified threshold value may be generated.

e. *Ease of use*. A large number of user convenience features, including a mnemonic-controlled, free-field input format and the capability for rerun with modified parameters, are provided.

Figure 1 is an idealized flow chart of the program. After the input data deck is processed and the nonlinear coefficients are determined, the intermodulation products are selected and processed one at a time. The only

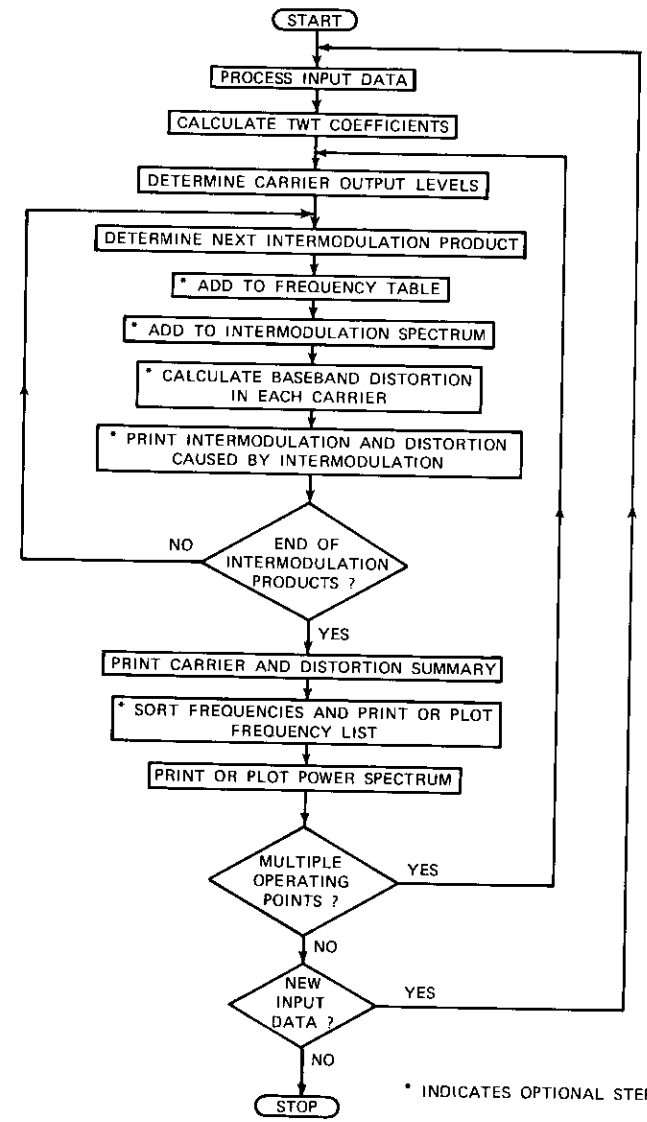


Figure 1. *Program Flow Chart*

information retained from this process is a list of distinct product frequencies, a table of power spectrum values, and the total baseband distortion in each carrier. If data pertaining to particular intermodulation products are desired, these data are printed as each intermodulation product is generated.

Following the processing of the intermodulation products, an output processor is called to print a summary of carriers and baseband distortion, and to print, plot, or punch the power spectrum and intermodulation product frequency data. Provisions are included in the program for considering multiple operating points, or for modifying any parameters in the problem in multiple executions.

Characterization of the nonlinear device

The characteristics of the nonlinear device are derived from the single-carrier transfer curves relating the output power, P_o, and phase, ψ_o, to the input power, P_i. Assume that the complex output envelope may be approximated by the Bessel function expansion shown in equation (9). Then the real and imaginary components of equation (9) are

$$g(\rho) \cos f(\rho) = \sum_{s=1}^{L} b_{sr} J_1(\alpha s \rho) \tag{40a}$$

$$g(\rho) \sin f(\rho) = \sum_{s=1}^{L} b_{si} J_1(\alpha s \rho) \tag{40b}$$

respectively. It has been found that 10 terms are sufficient for typical nonlinear characteristics. A procedure has been included in the program for determining appropriate values of b_{sr} and b_{si} from a set of measured values for P_i, P_o, and ψ_o. The normalized envelope levels corresponding to the measured input and output power levels may be found relative to the levels at saturation:

$$\bar{\rho} = \sqrt{\frac{2P_i}{P_{is}}} \tag{41a}$$

$$\bar{g} = \sqrt{\frac{2P_o}{P_{os}}} \tag{41b}$$

$$\bar{f} = \psi_o, \text{ in radians} \tag{41c}$$

where P_{is} and P_{os} are the power levels at saturation.

Error functions E_1 and E_2, which represent the squared difference between the envelope quantities in equation (41) and the expansion in equation (40), summed over all N measured points on the transfer curves, are defined as follows:

$$E_1(b_{sr}) = \sum_{l=1}^{N} \left[\bar{g}_l \cos(\bar{f}_l) - \sum_{s=1}^{10} b_{sr} J_1(\alpha s \bar{\rho}_l) \right]^2 \tag{42a}$$

$$E_2(b_{si}) = \sum_{l=1}^{N} \left[\bar{g}_l \sin(\bar{f}_l) - \sum_{s=1}^{10} b_{si} J_1(\alpha s \bar{\rho}_l) \right]^2 . \tag{42b}$$

The limit of the output phase angle as the input power approaches zero is given by

$$\lim_{\rho \to 0} \psi_o = \frac{\sum_{s=1}^{10} s b_{si}}{\sum_{s=1}^{10} s b_{sr}} \tag{43}$$

To ensure that this limit is zero, it has been found necessary to impose the following constraint:

$$\sum_{s=1}^{10} s b_{si} \equiv 0 . \tag{44}$$

The Fletcher-Powell optimization scheme [10] is used to determine values of the coefficients b_{sr} and b_{si} which minimize the error functions $E_1(b_{sr})$ and $E_2(b_{si})$, subject to the constraint given by equation (44). As an example of the application of this technique, a typical traveling wave tube (TWT) amplifier on the INTELSAT IV satellite will be considered. The measured transfer curves for this device are shown in Figure 2, where the values presented to the program are indicated by dots. The resulting real and imaginary coefficients generated by the program for $\alpha = 0.6$ are shown in Table 1. The contributions of selected terms to the total real and imaginary components of the envelope are indicated in Figures 3 and 4.

The difference between the measured data and the calculated values based on these coefficients is smaller than the measurement error. For low input levels, the calculated real output component is characterized by a linear behavior, whereas the imaginary component follows the third power of the input level. Furthermore, the calculated output angle is

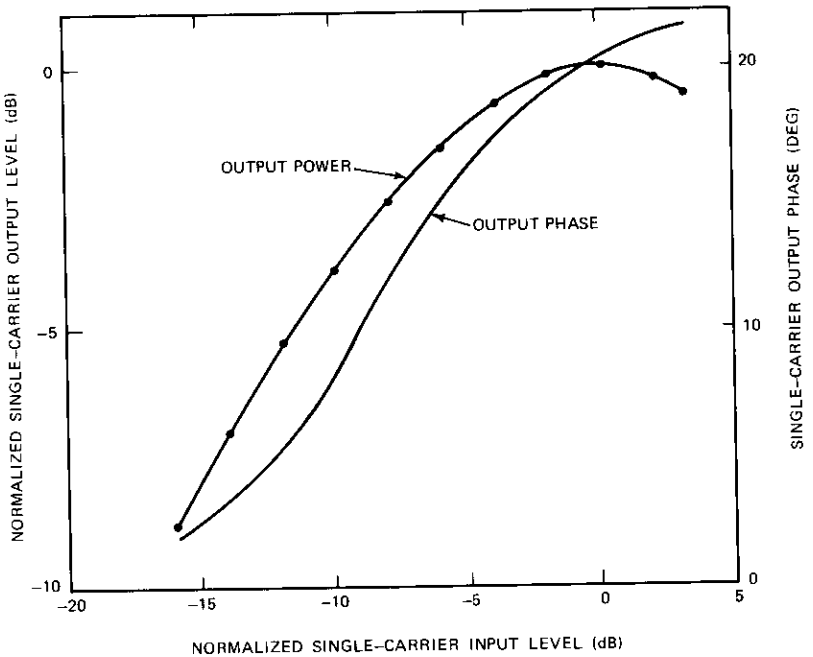


Figure 2. *Measured Nonlinear Transfer Curves*

TABLE 1. CALCULATED TRANSFER CURVE COEFFICIENTS (α = 0.6)

Index, s	1	2	3	4	5
b_{sr}	3.089	−0.09465	−0.2075	1.399	−0.1674
b_{si}	1.045	−1.034	1.992	−0.900	−0.6464
Index, s	6	7	8	9	10
b_{sr}	−0.4258	0.3040	0.4548	−0.5160	0.2435
b_{si}	0.6189	1.017	−2.342	1.837	−0.6750

linear with input power, indicating a convergence to zero as the input level approaches zero.

Since the Bessel function expansions for the real and imaginary components are obtained independently, a different α might have been used

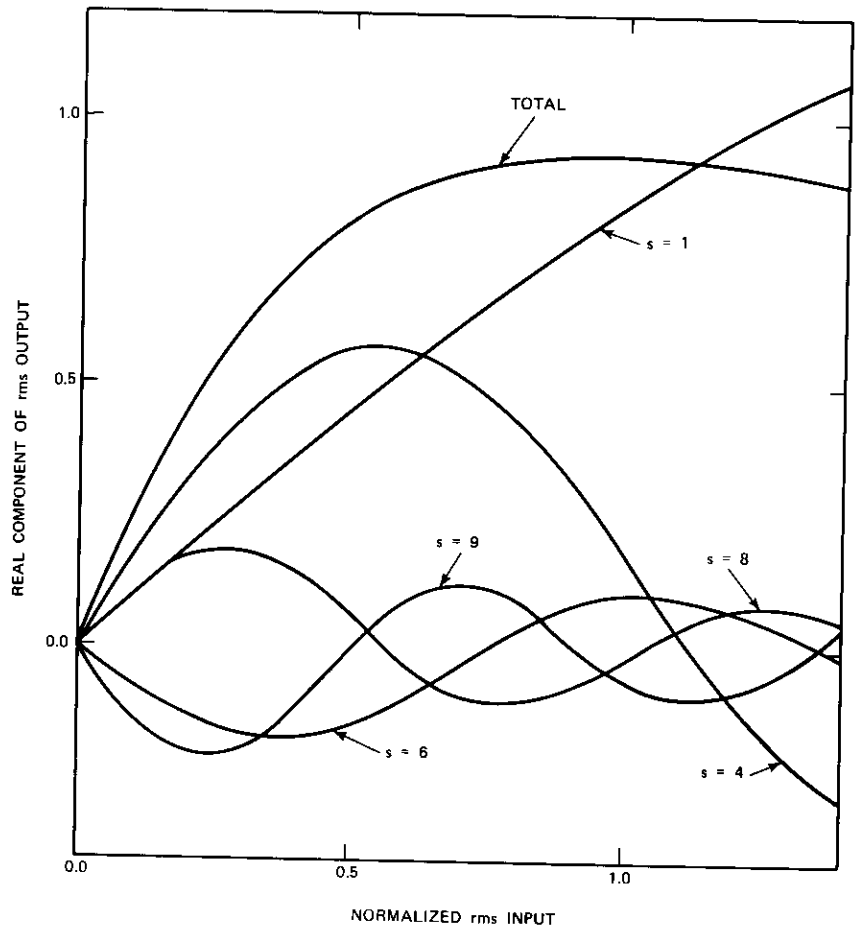


Figure 3. *Bessel Function Expansion of the Real Component of the Single-Carrier Transfer Function*

in each expansion. Furthermore, an optimum α could be found for each component by using the same optimization scheme. For the typical nonlinear transfer curves encountered, this has not been found to be necessary, since an adequate fit may be obtained for almost any reasonable value of α.

The coefficients b_s might also be determined from measurements of intermodulation product output power and phase as functions of input

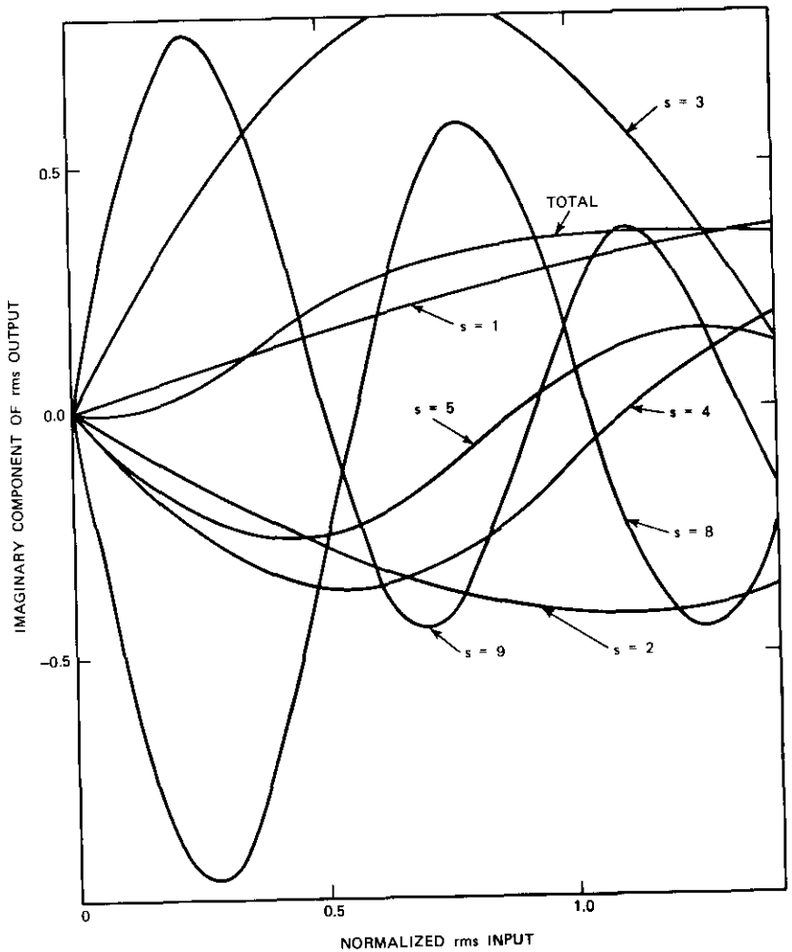


Figure 4. *Bessel Function Expansion of the Imaginary Component of the Single-Carrier Transfer Function*

level. This approach, although mathematically attractive, has practical disadvantages.

Characterization of input signals

The input signals are defined by a series of parameters specifying the input power level, the center frequency, and other information necessary

to calculate the RF spectrum and baseband distortion caused by intermodulation. The power spectral density of all input signals is assumed to be either Gaussian, rectangular (i.e., uniform over a finite range), or a convolution of both. This restriction has been imposed to eliminate the need for time-consuming numerical convolution techniques; all convolutions in the program are based on closed-form solutions.

The spectral shape assumed for each signal type is summarized in the following:

a. *FM telephony carriers.* All fully loaded telephony carriers are represented by a Gaussian spectrum whose rms deviation is specified by the user. For partially loaded carriers, the rms deviation of the signal is reduced accordingly. If the addition of a triangular spectrum-spreading waveform is desired, the resulting spectrum is the convolution of the spectrum of the partially loaded carrier and a uniform power density spectrum. This uniform spectrum represents the RF spectrum of the carrier modulated by the spreading waveform.

b. *FM television carriers.* No attempt was made to model the TV spectrum in the presence of a baseband signal, since each baseband signal generates a different RF spectrum. Instead, the case of a black picture, in which the RF spectrum is uniform and its peak-to-peak frequency deviation is determined by the spreading waveform, was considered. Unlike the multichannel telephony case, the spreading waveform power was not changed during transmission.

c. *Digital PCM/PSK carriers.* The RF spectrum (normalized to unit power) of a PCM/PSK carrier is given by

$$S_D(f) = T\left[\frac{\sin(\pi fT)}{(\pi fT)}\right]^2 \tag{45}$$

where $1/T$ is the baud rate and f is the frequency measured from the carrier frequency. However, because of the self-imposed restriction on Gaussian and rectangular spectra, an alternate representation had to be derived. It can be shown that the convolution of a Gaussian function with a rectangle of appropriate width closely approximates the function $S_D(f)$.

If an rms deviation equal to $0.246/T$ and a uniform spectrum of width $0.826/T$ are chosen, then the mean squared error between $S_D(f)$ and the convolved function, summed over a representative number of points on the main lobe of $S_D(f)$, is minimized. The effect of

filtering the PCM/PSK carrier and the resulting time-varying envelope has not been considered.

d. *Noise or carrier band.* The noise or carrier band spectrum is assumed to be rectangular.

Calculation of intermodulation product levels, spectra, and baseband distortion intermodulation products

Intermodulation products are generated by the program one at a time, according to generic form. Table 2 shows all possible 3rd- and 5th-order intermodulation products, including those arising in the presence of noise.

The center frequency of each intermodulation product is given by

$$F(k_1, k_2, \ldots, k_m; q) = \sum_{k=1}^{m} k_l f_l \tag{46}$$

where f_l is the center frequency of the lth input signal. The indices k_l can take integer values between -2 and $+3$. Furthermore, the condition for an in-band center frequency imposes the constraint that $\sum_{l=1}^{m} k_l \equiv 1$.

Note that the intermodulation products which include the noise term are characterized by $k_m \neq 0$ and/or $q \neq 0$. The index q affects the order of the intermodulation product, the spectrum, and the level, but does not alter the center frequency. Analogous to the noise-free case, the order of an intermodulation product may be defined as

$$\text{order} = \sum_{l=1}^{m} |k_l| + 2q \quad . \tag{47}$$

The selection of intermodulation products in this manner may be extended to odd orders higher than the fifth order, although the computation time will become prohibitive for a reasonable number of carriers. However, since the intermodulation product level decreases rapidly with increasing order, the inclusion of only 3rd-order products has been found to be adequate for most analyses.

INTERMODULATION PRODUCT LEVELS

The complex amplitude of carrier and intermodulation products, $M_S(k_1, k_2, \ldots, k_{m-1})$, is determined from equation (17). The corresponding complex amplitude of intermodulation products which include noise, $N(k_1, k_2, \ldots, k_{m-1}; q)$, is found from equation (23).

TABLE 2. SUMMARY OF INTERMODULATION PRODUCT TYPES

Generic Name*	Order	k_A	k_B	k_C	k_D	k_E	k_m	$2q$
$A + B - C$	3	1	1	−1	0	0	0	0
$2A - B$	3	2	−1	0	0	0	0	0
$A + B + C - D - E$	5	1	1	1	−1	−1	0	0
$2A + B - C - D$	5	2	1	−1	−1	0	0	0
$3A - B - C$	5	3	−1	−1	0	0	0	0
$A + B + C - 2D$	5	1	1	1	−2	0	0	0
$3A - 2B$	5	3	−2	0	0	0	0	0
$A + N - B$	3	1	−1	0	0	0	1	0
$A + B - N$	3	1	1	0	0	0	−1	0
$2A - N$	3	2	0	0	0	0	−1	0
$2N - A$	3	−1	0	0	0	0	2	0
$N + A - N$	3	1	0	0	0	0	0	2
$N + N - N$	3	0	0	0	0	0	1	2
$A + B + N - C - D$	5	1	1	−1	−1	0	1	0
$A + B + C - D - N$	5	1	1	1	−1	0	−1	0
$2A + N - B - C$	5	2	−1	−1	0	0	1	0
$2A + B - C - N$	5	2	1	−1	0	0	−1	0
$2N + A - B - C$	5	1	−1	−1	0	0	2	0
$3N - A - B$	5	−1	−1	0	0	0	3	0
$3A - B - N$	5	3	−1	0	0	0	−1	0
$A + B + N - 2C$	5	1	1	−2	0	0	1	0
$A + B + C - 2N$	5	1	1	1	0	0	−2	0
$3N - 2A$	5	−2	0	0	0	0	3	0
$3A - 2N$	5	3	0	0	0	0	−2	0
$2N + A - 2N$	5	1	0	0	0	0	0	4
$2N + N - 2N$	5	0	0	0	0	0	1	4
$N + A + B - C - N$	5	1	1	−1	0	0	0	2
$N + 2A - B - N$	5	2	−1	0	0	0	0	2
$N + A + N - B - N$	5	1	−1	0	0	0	1	2
$N + A + B - N - N$	5	1	1	0	0	0	−1	2
$N + 2A - N - N$	5	2	0	0	0	0	−1	2
$N + 2N - A - N$	5	−1	0	0	0	0	2	2

*A, B, C, D, and E are angle-modulated carriers; N is a noise or carrier band.

In the calculation of the intermodulation product levels, it is found that a small number of factors recur in various combinations. These factors, which are calculated only once at the beginning of the execution, are $J_{k_l}(\alpha s A_l)$ and $T(0, |k_m|, s)$ for k_l and k_m between 0 and 3.

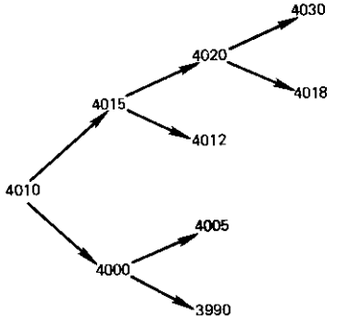
SORTING OF INTERMODULATION PRODUCT FREQUENCIES

Recall that the intermodulation products are calculated one at a time, on the basis of generic form; hence, sorting is required if an ordered listing of distinct intermodulation frequencies and the corresponding total power is desired. Because of the large number of intermodulation products generated in most problems of practical interest, improper choice of a sorting algorithm can drastically affect the program running time.

Throughout the execution of the program, an array of intermodulation product frequencies and the total level at each frequency are maintained in core. As each new intermodulation product frequency and level are calculated, it must be determined whether an intermodulation product at that frequency has been previously calculated. If so, the levels are added; if not, the new intermodulation product is added to the array and will appear in the sorted list generated by the output processor.

A technique which has been found to be ideally suited to this purpose is a binary distributive sort [11]. Figure 5 is an example of a tree structure demonstrating the sorting algorithm. Each node in the structure corresponds to a distinct intermodulation product frequency, which, in turn, points to at most two other frequencies in the list, one higher in value and one lower in value. These points take the form of integer arrays UP and DWN.

As each new frequency is generated, it must be compared with no more than $\log_2 n$ frequencies in the list, where n is the total length of the list,



LOC	FREQ	ARRAY	
		UP	DWN
1	4010	2	3
2	4015	6	8
3	4000	4	5
4	4005	—	—
5	3990	—	—
6	4020	9	7
7	4018	—	—
8	4012	—	—
9	4030	—	—

Figure 5. *Tree Structure for Sorting Algorithm*

to determine whether an identical frequency has been previously entered. If not, the new frequency is entered in the list and an appropriate pointer is set in the UP or DWN array. After all frequencies have been entered in the table, the arrays UP and DWN contain sufficient information to generate a sorted list of frequencies without referencing the actual frequency values.

CONVOLUTION TECHNIQUE

Convolutions of spectra are necessary to calculate both RF spectra and baseband distortion. Hence, the process of convolution is an important aspect of the program design. The convolution algorithm is restricted to accept only Gaussian or rectangular spectra. This restriction was established to avoid both the excessive computation times associated with numerical convolution and the complexity involved in accommodating a wider range of spectrum types.

A further limitation of the algorithm is that an exact closed-form solution is obtained for only three or fewer rectangles. If additional rectangles are present, the narrowest ones are approximated by Gaussian functions with equivalent rms deviations. Since convolving two or more Gaussian functions results in another Gaussian function, the convolution algorithm inputs are limited to a single Gaussian function and up to three rectangular functions.

BASEBAND DISTORTION

The distortion measures of interest depend on the modulation and baseband signals used. Three cases have been considered:

a. frequency modulation with multichannel telephony signals,
b. frequency modulation with various TV signals, and
c. PCM/PSK.

Although an exact analysis is available for the first two cases, for PCM/PSK, approximations are used which correspond to treating the intermodulation effect as white Gaussian additive noise.

Telephony Carriers. The *NPR* (noise power ratio) is defined as the ratio of signal power to noise power in a specified frequency band. The *NPR* caused by intermodulation is calculated by using the formula

$$NPR = \frac{P(f_r) \cdot f_{rms}^2}{(1-\epsilon) f_r^2 S_f} \tag{48}$$

where S_f is the frequency demodulated power spectral density, f_{rms} is the *rms* frequency deviation, ϵ is the ratio of minimum to maximum baseband frequencies, f_r is the frequency for which the *NPR* is calculated, and $P(f_r)$ is the pre-emphasis weighting factor.

TV Carriers. In TV transmission the signal-to-noise ratio, S/N, is defined as the peak-to-peak luminance power divided by the noise power (either weighted or unweighted). In the baseband, the peak-to-peak luminance voltage is 0.714 volt. The peak test-tone deviation, f_d, referenced to a 1-volt peak-to-peak luminance power, L_{p-p}, in Hz^2, is given by

$$L_{p-p} = (2 \times 0.714 f_d)^2 \tag{49}$$

and the resulting S/N is

$$S/N = \frac{(2 \times 0.714 f_d)^2}{\int_{f_1}^{f_m} S_f(f)\, df} \quad . \tag{50}$$

The effects of the de-emphasis network and noise weighting are introduced into the S/N equation as follows:

$$(S/N)_D = \frac{(2 \times 0.714 f_d)^2}{\int_{f_1}^{f_m} S_f(f)\, P_D(f)\, df} \tag{51}$$

$$(S/N)_{D,W} = \frac{(2 \times 0.714 f_d)^2}{\int_{f_1}^{f_m} S_f(f)\, P_D(f)\, P_W(f)\, df} \tag{52}$$

where $P_D(f)$ and $P_W(f)$ are respectively the network responses of the de-emphasis and noise weighting networks.

Digital Carriers. The baseband impairment in a carrier with digital traffic is measured in terms of the bit-error probability, P_{be}. To estimate this parameter, it has been necessary to make two assumptions:

a. That P_{be} can be calculated from the following formula, which is valid for Gaussian noise and 4-phase PSK:

$$P_{be} = \frac{1}{2}\, erfc(\sqrt{\gamma}) \tag{53}$$

where γ is the S/N at the filter output at the sampling instant. For the carrier-to-IF noise power, the total power must be divided between the two quadrature components; hence,

$$\gamma = \frac{C}{2N}$$

where C/N is the IF carrier-to-noise power.

b. That the noise N is strictly the IF noise, and that no special consideration is made to incorporate time-domain results since no rigorous mathematical solutions are available.

Sample calculations

This section presents some results obtained from the computer program. Two separate examples have been selected, each illustrating a different aspect of the analysis. The first example deals exclusively with RF calculations. A number of RF spectra are included to demonstrate the validity of approximating a band of carriers by an equivalent noise input.

In a second example, an input consisting of three carriers is assumed, and the resulting intermodulation spectrum and demodulated output are determined. This case is especially significant since it involves the generation of intermodulation products which are modulation sidebands of the input carriers.

RF spectrum calculations

Three cases have been considered here. All three are characterized by an input signal made up of an angle-modulated carrier (FM with a Gaussian baseband signal) and a second signal which can take one of three forms, each having a uniform RF spectrum:

case 1: 10 small, equally spaced carriers at a uniform level;
case 2: a band of Gaussian noise of constant level; and
case 3: a single carrier, frequency modulated by a triangular waveform. A large modulation index is assumed.

It is assumed that both signals are at the same level and that they saturate the amplifier. The FM carrier is located at 4,000 MHz and the other signal in the range between 4,015 MHz and 4,025 MHz. The resulting intermodulation power spectra are calculated for all three cases.

CASE 1

The unmodulated intermodulation products resulting from the single carrier and the band of 10 small carriers are shown in Figure 6. If both 3rd- and 5th-order products are calculated, the plot shown in Figure 7 is obtained. The effect of the 5th-order products, which may be observed by comparing both figures, is to fill in the valleys between 3rd-order intermodulation products and to change slightly the level at frequencies where 3rd-order products fall.

The corresponding power spectra are shown in Figure 8 for the case in which all 10 carriers are modulated with 500-MHz rms deviation. Note that, if the level of the input is decreased, the 5th-order products

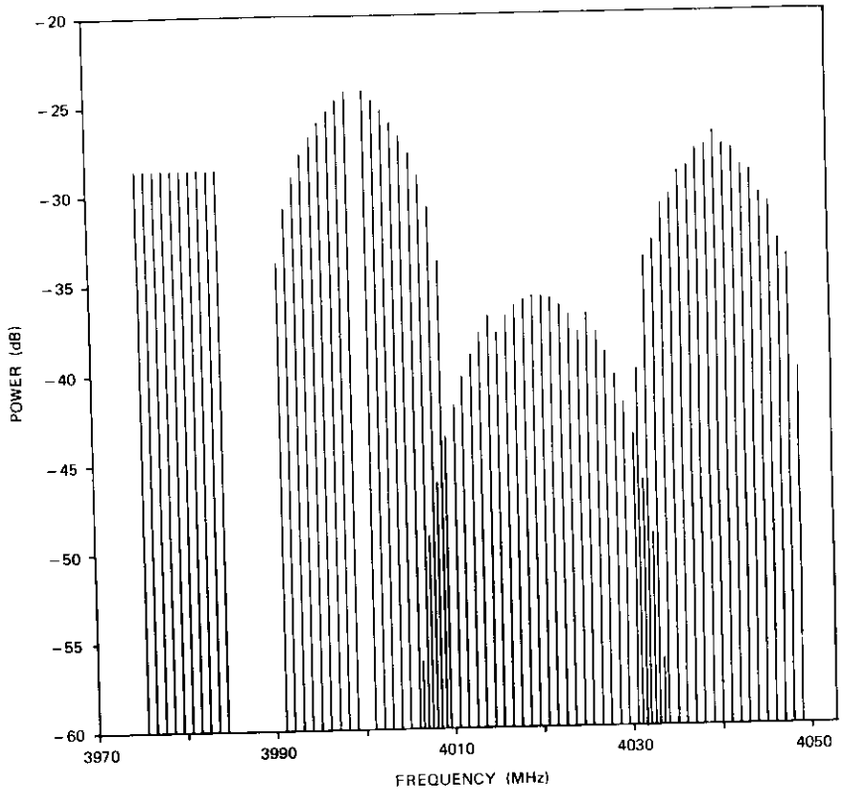


Figure 6. *3rd-Order RF Intermodulation Products Generated by 10 Equal Carriers and a Single Carrier*

decrease very rapidly and can be ignored in most cases. Also note that there are many intermodulation products clustered about the single carrier at 4,000 MHz. (The spectrum is triangular in a linear scale.) There is no product falling exactly on the single carrier, which accounts for the dip in the spectrum at 4,000 MHz.

CASE 2

The input is characterized by a single carrier at 4,000 MHz and a band of noise (or carriers) centered at 4,020 MHz. The intermodulation power spectrum plots generated by the program are shown in Figure 9. Note that the spectrum is strikingly similar to that of case 1, except that the dip at 4,000 MHz does not occur, since calculating the power spectrum of a band of noise is equivalent to taking the limit as the interval between carriers approaches zero.

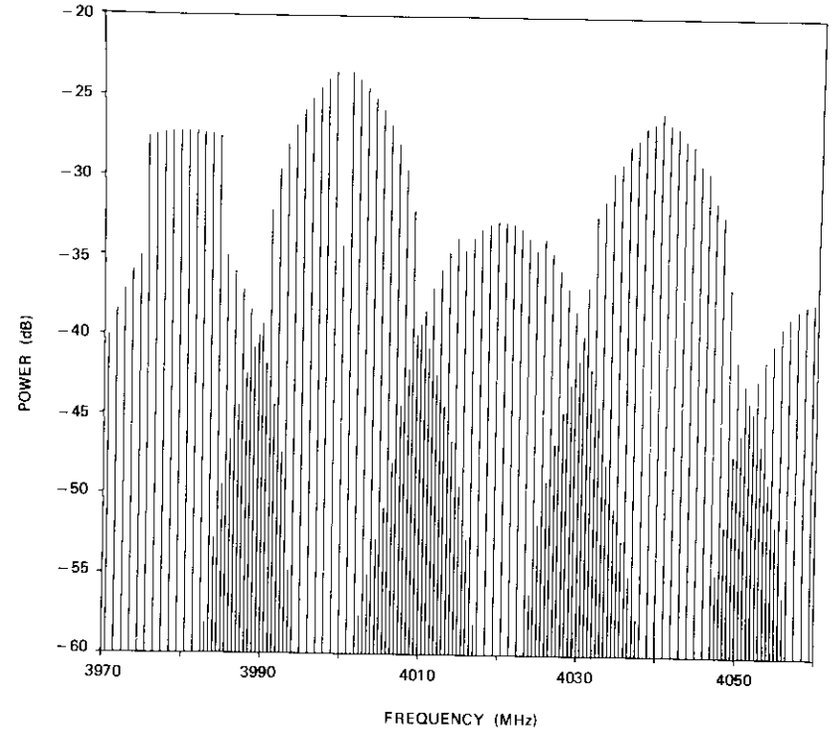


Figure 7. *3rd- and 5th-Order RF Intermodulation Products Generated by 10 Equal Carriers and a Single Carrier*

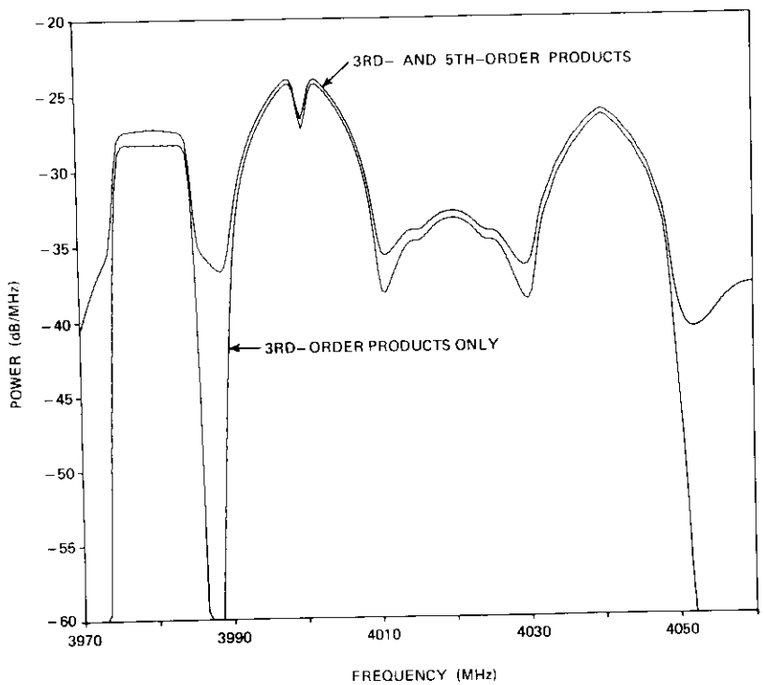


Figure 8. *RF Intermodulation Power Spectral Density Generated by 10 Equal Carriers and a Single Carrier*

CASE 3

The same single carrier is now assumed to share an amplifier with another carrier which is frequency modulated with a triangular waveform so that the input spectrum is identical to that of case 2. The resulting intermodulation spectrum is shown in Figure 10. There is a significant difference between this spectrum and the previous spectra because, although the input spectra are identical in all cases, in this last case, both input signals have a constant envelope and hence generate considerably fewer intermodulation products.

3-Carrier input

A 3-carrier input to a nonlinear device is an interesting configuration, since there are intermodulation products which are modulation sidebands

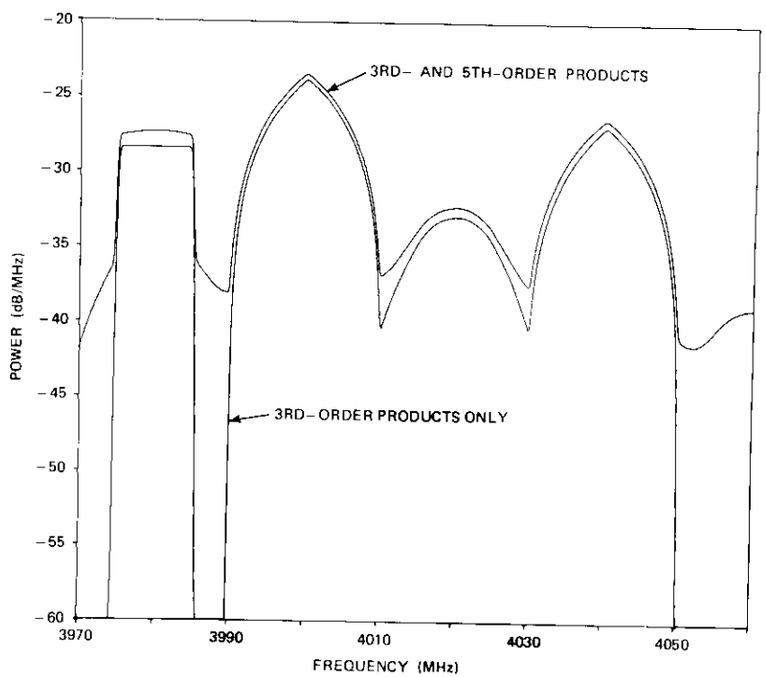


Figure 9. *RF Intermodulation Power Spectral Density Generated by a Carrier Band and a Single Carrier*

and consequently cannot be treated as an independent interference entry. The three carriers considered are listed in Table 3. Changes in the relative carrier level, Δ, were found to alter dramatically the character of the carrier-to-intermodulation curves and the resulting demodulated output of reference carrier A.

TABLE 3. INPUT CARRIER PARAMETERS

Carrier	Frequency (MHz)	Level (dB)	Modulation
A	4,020.0	0.0	TV
B	4,005.0	Δ	Telephony
C	4,007.5	Δ	Telephony

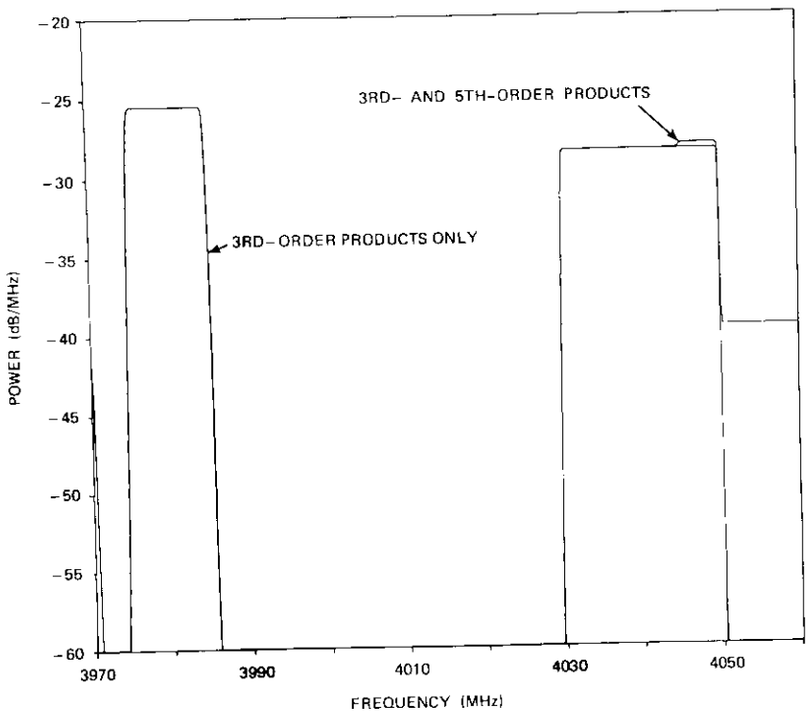


Figure 10. *RF Intermodulation Power Spectral Density Generated by a Carrier with a Rectangular Spectrum and a Single Carrier*

RF INTERMODULATION CALCULATIONS FOR $\Delta = -12$

Initially the levels of carriers B and C are assumed to be 12 dB below that of carrier A ($\Delta = -12$ dB), which is the typical level for program channels sharing a satellite transponder with a TV carrier. The RF intermodulation power spectrum in the vicinity of carrier A is shown in Figure 11 for input backoffs varying from 0 to −14 dB. Third-order A + B − C and A + C − B products occur at 4,017.5 MHz and 4,022.5 MHz, respectively. Fifth-order A + 2B − 2C and A + 2C − 2B products occur at 4,015 MHz and 4,025 MHz, respectively.

It appears that the 3rd-order intermodulation products decrease monotonically as the tube is backed off, while in the regions where 5th-order intermodulation products dominate, this is no longer the case. This is

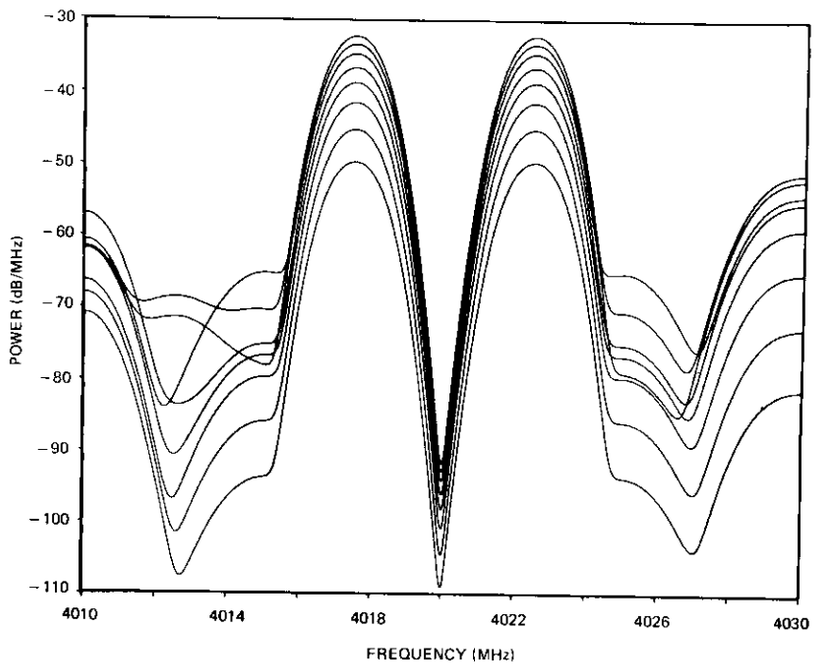


Figure 11. *RF Intermodulation Power Spectral Density Generated by a TV Carrier (spreading only) and Two Program Channels*

confirmed in Figure 12, where the levels of the A + B − C and A + 2B − 2C intermodulation products are plotted as functions of input backoff. It is apparent that, at some points, the levels of 5th-order products may actually increase as the input level is decreased. This effect has also been reported in Reference 12.

BASEBAND INTERMODULATION CALCULATIONS FOR $\Delta = -12$

The major distortion of carrier A is caused by the 3rd-order A + B − C and A + C − B intermodulation products located symmetrically about A at 4,017.5 MHz and 4,022.5 MHz, respectively. Although the RF power of these products decreases with increasing backoff, the baseband impairment (measured as signal-to-weighted noise) does not behave in the same fashion. Figure 13 shows the calculated S/N ratio, as well as the ratios between the carrier and both the in-phase and in-quadrature components of the intermodulation products. It is apparent that the signal-to-noise ratio is considerably higher than the estimate based on the RF

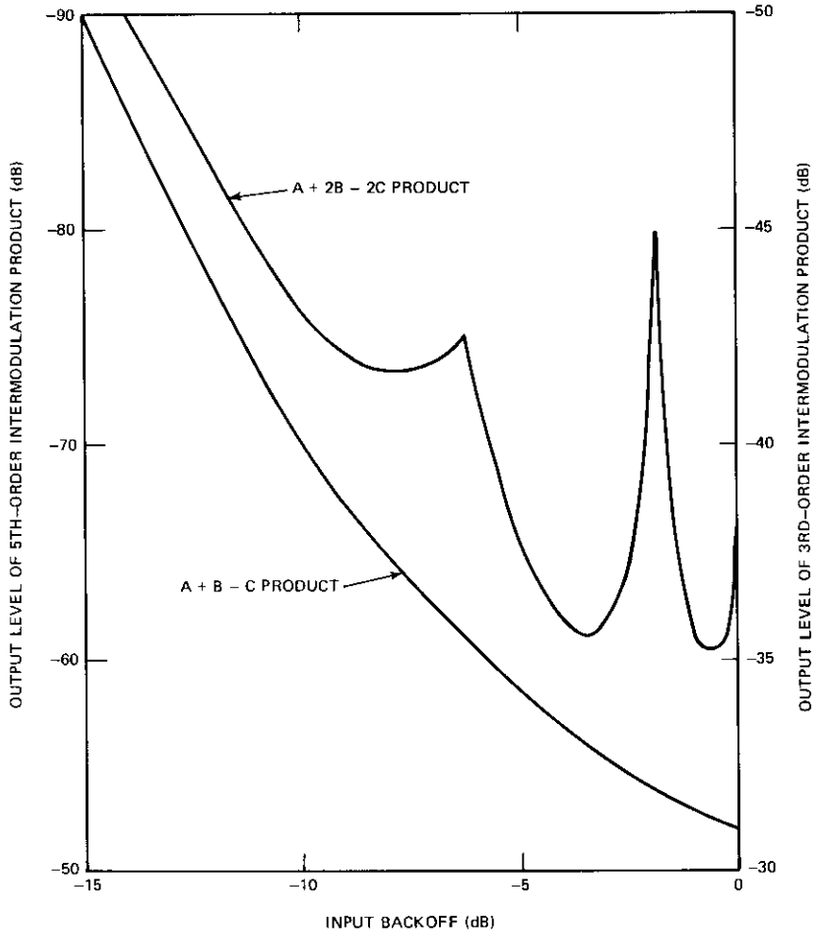


Figure 12. *Intermodulation Product Level as a Function of Input Backoff*

intermodulation spectral density. This phenomenon was observed experimentally during the INTELSAT IV simulation and is reported in Reference 7.

EFFECT OF THE RELATIVE CARRIER LEVELS, Δ

The angle between the A + B − C intermodulation product and the demodulated carrier is shown in Figure 14. Initially the angle is positive, but it becomes negative with increasing backoffs. For $\Delta = -12$ dB, the

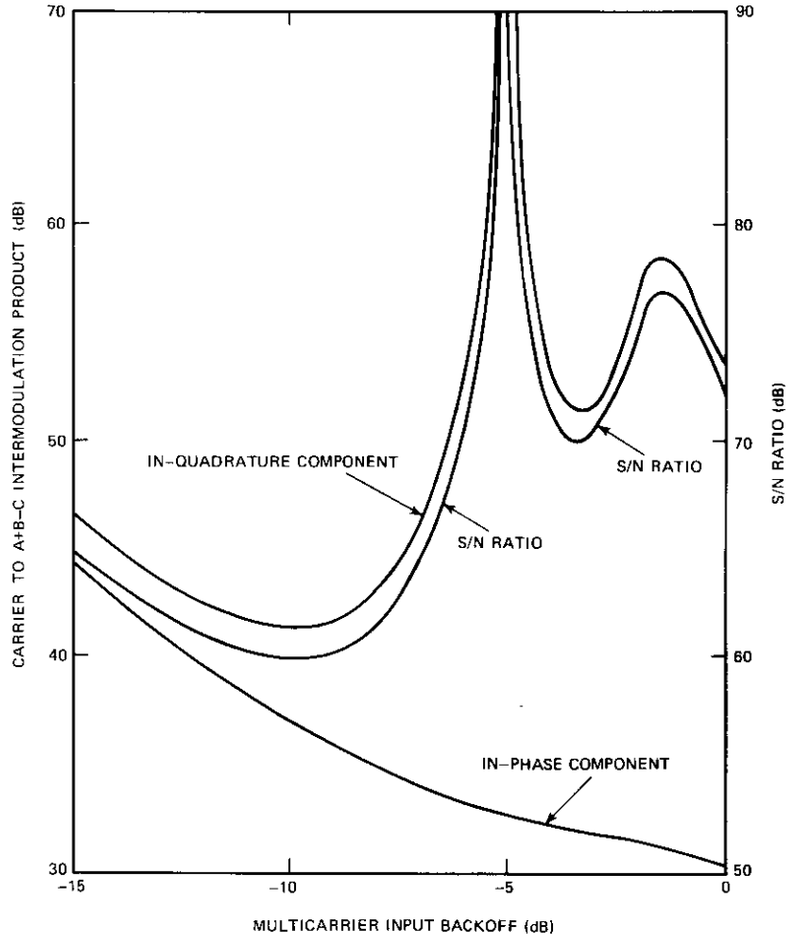


Figure 13. *Carrier-to-Intermodulation Ratio and S/N vs Multicarrier Input Backoff*

angle is zero at the backoff level for which S/N becomes infinite. Also shown in Figure 14 are the angles obtained for other relative levels of the carriers.

Figures 15 and 16 show the corresponding ratios between the carrier and the two intermodulation product components. The improvement in the ratio of carrier-to-in-quadrature components with increasing input level becomes pronounced as the level of the reference carrier is increased.

This explains why the measurement reported in Reference 4 for $\Delta = 20$ dB does not show this effect.

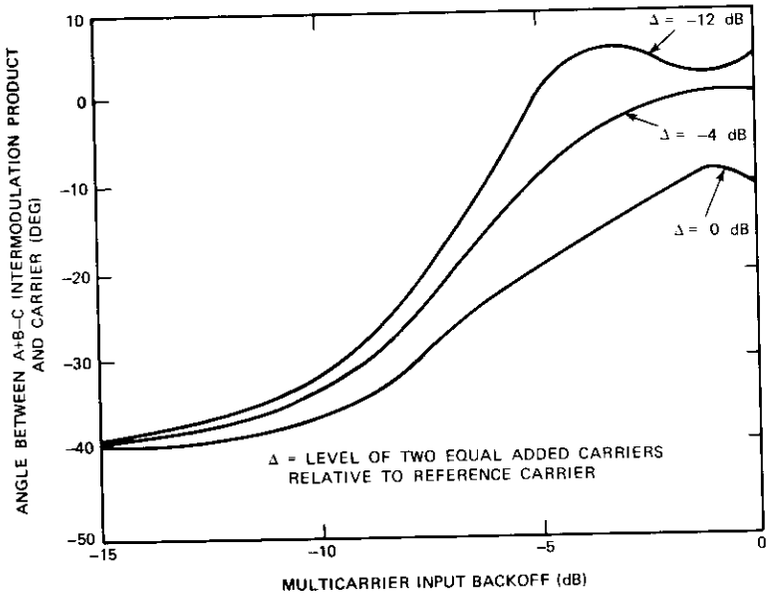


Figure 14. *Relative Phase Angle Between Intermodulation Product and Carrier vs Multicarrier Input Backoff*

Conclusions

A unified approach to the calculation of intermodulation effects in a "memoryless" nonlinear amplifier has been presented. Both amplitude nonlinearity and AM/PM conversion effects have been considered. An expression has been derived for the output of the device when the input consists of a number of bandpass band-limited signals. The particular case of an input formed by angle-modulated carriers and a band of Gaussian noise has been treated in detail.

The power spectral density of the baseband signal caused by intermodulation has been derived by using a time-domain approach. This technique permits an exact calculation of baseband distortion since the statistical dependence between the intermodulation noise and carriers is elucidated.

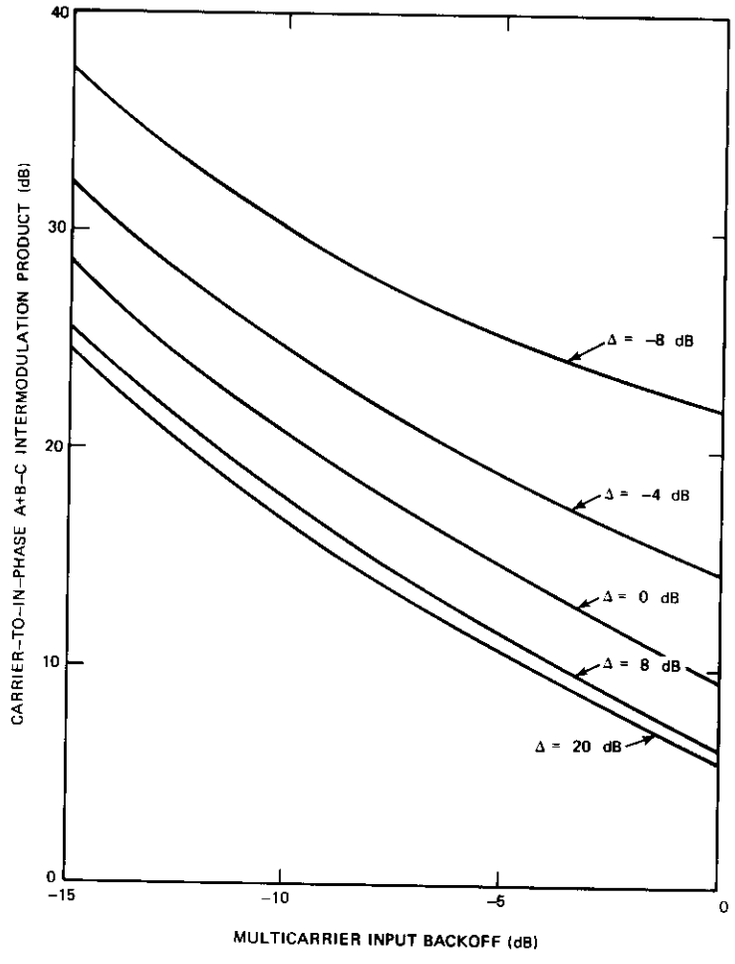


Figure 15. *Carrier-to-In-Phase Intermodulation Ratio vs Multicarrier Input Backoff*

A flexible computer program has been developed and exercised on a wide range of intermodulation problems. Comparison of calculations with reported measurements indicates that the memoryless assumption is valid for helix-type TWTs such as those presently used in communications satellites.

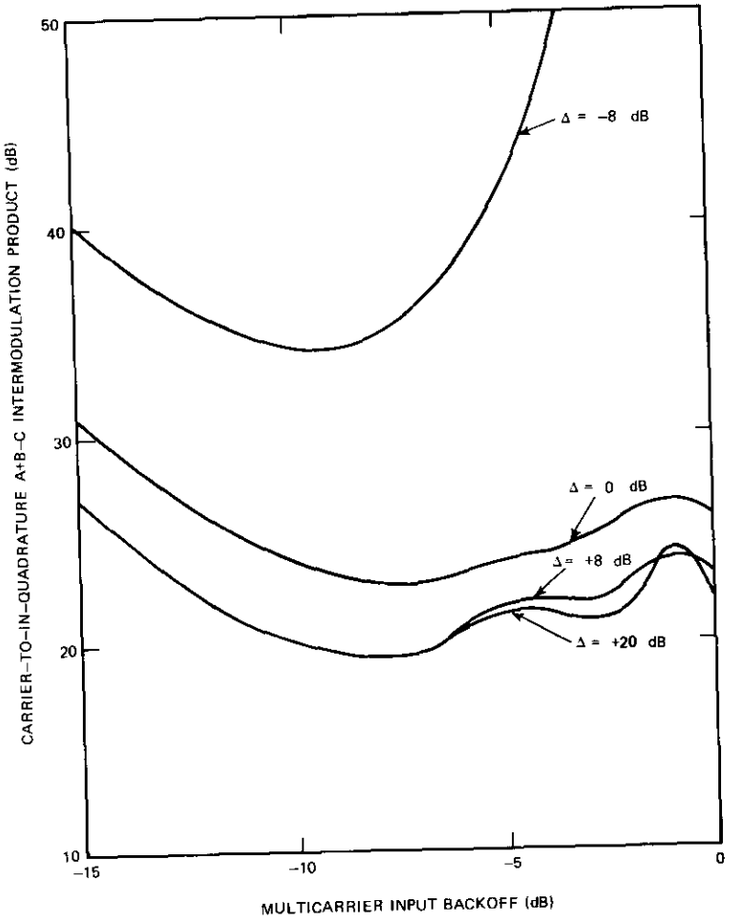


Figure 16. *Carrier-to-In-Quadrature Intermodulation Ratio vs Multicarrier Input Backoff*

For want of a more rigorous model, the program has also been used to evaluate the performance of high-power cavity TWTs using a representative nonlinear transfer function. Comparison with measurements has indicated that such nonlinearities can be accurately described as memoryless over a narrow bandwidth.

References

[1] E. D. Sunde, "Intermodulation Distortion in Multicarrier FM Systems," IEEE International Convention, New York, March 1965, *Convention Record*, Vol. 13, Part 2, pp. 130–146.

[2] R. J. Westcott, "Investigation of Multiple FM/FDM Carriers Through a Satellite TWT Operating Near to Saturation," *Proceedings of the IEE*, Vol. 114, No. 6, June 1967, pp. 726–750.

[3] A. L. Berman and E. Podraczky, "Experimental Determination of Intermodulation Distortion Produced in a Wide-Band Communication Repeater," IEEE International Convention, New York, March 1967, *Convention Record*, Vol. 15, Part 2, pp. 69–88.

[4] A. L. Berman and C. E. Mahle, "Nonlinear Phase Shift in Traveling Wave Tubes as Applied to Multiple-Access Communications Satellites," *IEEE Transactions on Communications Technology*, COM-18, No. 1, February 1970, pp. 37–48.

[5] O. Shimbo, "Effects of Intermodulation, AM-PM Conversion, and Additive Noise in Multi-Carrier TWT Systems," *Proceedings of the IEEE*, Vol. 59, No. 2, February 1971, pp. 230–238.

[6] A. R. Kaye, D. A. George, and M. J. Eric, "Analysis and Compensation of Bandpass Nonlinearities for Communications," *IEEE Transactions on Communications*, COM-20, No. 5, October 1972, pp. 965–972.

[7] N. K. M. Chitre and J. C. Fuenzalida, "Baseband Distortion Caused by Intermodulation in Multicarrier FM Systems," *COMSAT Technical Review*, Vol. 2, No. 1, Spring 1972, pp. 147–172.

[8] N. M. Blachman, "Detectors, Bandpass Nonlinearities and Their Optimization: Inversion of the Chebyshev Transform," *IEEE Transactions on Information Theory*, IT-17, No. 4, July 1971, pp. 398–404.

[9] S. Stein and J. J. Jones, *Modern Communication Principles*, New York: McGraw-Hill, 1967, p. 75.

[10] R. Fletcher and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization," *The Computer Journal*, Vol. 6, 1963–1964, pp. 163–168.

[11] T. N. Hibbard, "Some Combinatorial Properties of Certain Trees with Application to Searching and Sorting," *Journal of the Association for Computing Machinery*, January 1962, pp. 13–28.

[12] O. Shimbo, "Nonlinear Distortion of Frequency Division Multiplexed Signals," *Journal of the Institute of Electrical Communications Engineers of Japan*, February 1961 (in Japanese).

Appendix A. Mathematical Derivations

Time-domain analysis

The input signal is represented by

$$e_i(t) = Re\left\{\sum_{l=1}^{m} A_l(t) \exp\left[j\omega_o t + j\theta_l(t)\right]\right\} \tag{A1}$$

where $A_l(t)$ and $\theta_l(t)$ are arbitrary baseband time functions. Equation (A1) can be rewritten as

$$e_i(t) = Re\left\{\hat{e}_i(t)\right\} \tag{A2a}$$

$$\hat{e}_i(t) = \rho(t) \exp\left[j\omega_o t + j\Theta(t)\right] \tag{A2b}$$

where

$$\rho = \sqrt{x^2 + y^2} \tag{A3a}$$

$$\Theta = \tan^{-1}\frac{y}{x} \tag{A3b}$$

$$x = \sum_{l=1}^{m} A_l(t) \cos\theta_l(t) \tag{A3c}$$

$$y = \sum_{l=1}^{m} A_l(t) \sin\theta_l(t) \quad . \tag{A3d}$$

Cartesian coordinates will be introduced during the following steps in the derivation, since a double Fourier transformation is required.

The fundamental component of the output is represented by

$$\begin{aligned} e_o(t) &= Re\left\{g\left(\sqrt{x^2+y^2}\right) \exp\left[j\omega_o t + j\tan^{-1}\frac{y}{x} + jf\left(\sqrt{x^2+y^2}\right)\right]\right\} \\ &= Re\left\{\frac{g\left(\sqrt{x^2+y^2}\right)}{\sqrt{x^2+y^2}} \exp\left[jf\left(\sqrt{x^2+y^2}\right)\right](x+jy)\exp\left(j\omega_o t\right)\right\} \\ &= Re\left\{\frac{g\left(\sqrt{x^2+y^2}\right)}{\sqrt{x^2+y^2}} \exp\left[jf\left(\sqrt{x^2+y^2}\right)\right]\hat{e}_i(t)\right\} \quad . \end{aligned} \tag{A4}$$

For the sake of convenience, the following double Fourier transformation is defined as

$$\begin{aligned} L(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} &\frac{g\left(\sqrt{x^2+y^2}\right)}{\sqrt{x^2+y^2}} \exp\left[jf\left(\sqrt{x^2+y^2}\right)\right] \\ &\cdot (x+jy)\exp\left(-jux - jvy\right) dx\, dy \end{aligned} \tag{A5}$$

or, alternatively,

$$\begin{aligned} &\frac{g\left(\sqrt{x^2+y^2}\right)}{\sqrt{x^2+y^2}} \exp\left[jf\left(\sqrt{x^2+y^2}\right)\right](x+jy) \\ &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L(u,v)\exp\left(jux + jvy\right) du\, dv \quad . \end{aligned} \tag{A6}$$

Substituting the function $L(u, v)$ into equation (A4) makes it possible to represent the output by

$$\begin{aligned} e_o(t) = \frac{1}{4\pi^2} Re\Bigg\{ &\exp\left[j\omega_o t\right]\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L(u,v) \\ &\cdot \exp\left[jux + jvy\right] du\, dv\Bigg\} \quad . \end{aligned} \tag{A7}$$

Substituting the expressions for x and y from equation (A3) into equation (A7) results in

$$\begin{aligned} e_o(t) = \frac{1}{4\pi^2} Re\Bigg\{ &\exp\left[j\omega_o t\right]\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L(u,v)\prod_{l=1}^{m}\exp\Bigg[jA_l\sqrt{u^2+v^2} \\ &\cdot \sin\left\{\theta_l(t) + \tan^{-1}\frac{u}{v}\right\}\Bigg] du\, dv\Bigg\} \quad . \end{aligned} \tag{A8}$$

Expanding the exponential as a Bessel function series, i.e.,

$$\exp\left[jz\sin\theta\right] = \sum_{k=-\infty}^{\infty} J_k(z)\exp\left[jk\theta\right] \tag{A9}$$

makes it possible to represent equation (A8) by

$$\begin{aligned} e_o(t) = Re\Bigg\{ &\exp\left[j\omega_o t\right] \sum_{\substack{k_1,k_2,\ldots,k_m=-\infty \\ (k_1+k_2+\ldots+k_m=1)}}^{\infty} \exp\left[j\sum_{l=1}^{m} k_l\theta_l(t)\right] \\ &\cdot M(k_1, k_2, \ldots, k_m)\Bigg\} \end{aligned} \tag{A10}$$

where

$$M(k_1, k_2, \ldots, k_m) = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} L(u,v) \left[\prod_{l=1}^{m} J_{k_l}(A_l \sqrt{u^2 + v^2}) \right]$$
$$\cdot \exp\left[j \sum_{l=1}^{m} k_l \tan^{-1} \frac{u}{v} \right] du\, dv$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{g(\sqrt{x^2 + y^2})}{\sqrt{x^2 + y^2}}$$
$$\cdot \exp\left[jf(\sqrt{x^2 + y^2}) \right] (x + jy)$$
$$\cdot \left[\prod_{l=1}^{m} J_{k_l}(A_l \sqrt{u^2 + v^2}) \right]$$
$$\cdot \exp\left[j \sum_{l=1}^{m} k_l \tan^{-1} \frac{u}{v} \right]$$
$$\cdot \exp(-jux - jvy)\, du\, dv\, dx\, dy \quad . \tag{A11}$$

Using the polar coordinate transformations,

$$x = \rho \cos \theta \qquad u = \gamma \sin \eta$$
$$y = \rho \sin \theta \qquad v = \gamma \cos \eta$$

and performing the integrations on θ and η makes it possible to simplify equation (A11):

$$M(k_1, k_2, \ldots, k_m) = \int_0^{\infty} \gamma \left[\prod_{l=1}^{m} J_{k_l}(\gamma A_l) \right] d\gamma$$
$$\cdot \int_0^{\infty} \rho g(\rho) \exp[jf(\rho)] J_1(\gamma\rho)\, d\rho \quad . \tag{A12}$$

In the summation in equation (A10), it is assumed that $M(k_1, k_2, \ldots, k_m) \equiv 0$ if $k_1 + k_2 + \ldots + k_m \neq 1$ and that M is given by equation (A12) if $k_1 + k_2 + \ldots + k_m = 1$. Thus,

$$e_o(t) = Re\left\{ \exp[j\omega_o t] \sum_{\substack{k_1, k_2, \ldots, k_m = -\infty \\ (k_1 + k_2 + \ldots + k_m = 1)}}^{\infty} \exp\left[j \sum_{l=1}^{m} k_l \theta_l(t) \right] \right.$$
$$\left. \cdot M(k_1, k_2, \ldots, k_m) \right\} \tag{A13}$$

where $k_l (l = 1, 2, \ldots, m)$ can take negative, positive, and zero values under the constraint that

$$\sum_{l=1}^{m} k_l = 1 \quad . \tag{A14}$$

If

$$\theta_l(t) = \omega_l t + \phi_l(t) + \lambda_l, \qquad l = 1, 2, \ldots, m - 1$$
$$\theta_m(t) = \omega_m t + \tan^{-1}\left[\frac{N_s(t)}{N_c(t)} \right] = \omega_m t + \phi_m(t) \tag{A15}$$

the input, $e_i(t)$, represents the $m - 1$ carriers, modulated in amplitude and phase, plus narrowband noise. In the output, $e_o(t)$, the component for which $k_p = 1$ and all of the other $k_l = 0$ represents the pure output signal corresponding to the pth pure signal of the input, $e_i(t)$. The pure output signals and intermodulation products which are not disturbed by the noise (N_c or N_s) may be obtained by taking the average of equation (A13) on N_c and N_s.

In the following, it is assumed that the noise angle, $\phi_m(t)$, is uniformly distributed on $[0, 2\pi]$ and is independent of the noise amplitude,

$$\xi = \sqrt{N_c^2(t) + N_s^2(t)} \quad . \tag{A16}$$

Furthermore,

$$\exp\left[jk_m \tan^{-1} \frac{N_s(t)}{N_c(t)} \right] = \left[\frac{N_c(t) + jN_s(t)}{\sqrt{N_c^2(t) + N_s^2(t)}} \right]^{k_m} \quad . \tag{A17}$$

Thus, the average of the factors including $N_c(t)$ and $N_s(t)$ is

$$C(\gamma) = \underset{N_c N_s}{E} \left\{ J_{k_m}\left[\gamma \sqrt{N_c^2(t) + N_s^2(t)} \right] \exp\left[jk_m \tan^{-1} \frac{N_s(t)}{N_c(t)} \right] \right\}$$
$$= 0 \quad , \qquad k_m \neq 0$$
$$= \int_0^{\infty} J_0(\xi\gamma)\, p(\xi)\, d\xi \quad , \qquad k_m = 0 \tag{A18}$$

where $p(\xi)$ is the probability density function of the noise amplitude, ξ.

Now, when the average of equation (A18) is represented by

$$C(\gamma) = \int_0^{\infty} J_0(\xi\gamma)\, p(\xi)\, d\xi \tag{A19}$$

and the noise is Gaussian, $p(\xi)$ is Rayleigh distributed; hence,

$$C(\gamma) = \exp\left[-\frac{\gamma^2}{2} R(0) \right] \tag{A20}$$

where $R(0)$ represents the RF noise power or the average of N_c^2 or N_s^2. Thus, the pure signals and intermodulation products are given by

$$e_S(t) = Re\left\{\exp[j\omega_o t] \sum_{\substack{k_1,k_2,\ldots,k_{m-1}=-\infty \\ (k_1+k_2+\ldots+k_{m-1}=1)}}^{\infty} \exp\left[j\sum_{l=1}^{m-1} k_l\theta_l(t)\right] \cdot M_S(k_1, k_2, \ldots, k_{m-1})\right\} \quad \text{(A21)}$$

where

$$M_S(k_1, k_2, \ldots, k_{m-1}) = \int_0^\infty \gamma\left[\prod_{l=1}^{m-1} J_{kl}(\gamma A_l)\right] C(\gamma)\, d\gamma \int_0^\infty \rho g(\rho) \cdot \exp[jf(\rho)]\, J_1(\gamma\rho)\, d\rho \quad \text{(A22)}$$

and the components at the output disturbed by noise, which represent new noise at the output, are given by

$$e_N(t) = e_o(t) - e_S(t) \quad . \quad \text{(A23)}$$

Reference A1 describes a method for evaluating $M(k_1, k_2, \ldots, k_m)$ when the nonlinear characteristic $g(\rho)\exp[jf(\rho)]$ is approximated by a power or Fourier series. The method used here is that suggested in Reference A2.

First, $g(\rho)\exp[jf(\rho)]$ is approximated by

$$g(\rho)\exp[jf(\rho)] = \sum_{s=1}^{L} b_s J_1(\alpha s\rho) \quad . \quad \text{(A24)}$$

Substituting this expression into equation (A12) and using the formula

$$\int_0^\infty G(\gamma)\, d\gamma \int_0^\infty \rho J_1(B\rho)\, J_1(\gamma\rho)\, d\rho = \frac{G(B)}{B} \quad \text{(A25)}$$

results in

$$M(k_1, k_2, \ldots, k_m) = \sum_{s-1}^{L} b_s\left[\prod_{l=1}^{m} J_{kl}(\alpha s A_l)\right] \quad . \quad \text{(A26)}$$

In a similar fashion, M_S of equation (A22) is found to be

$$M_S(k_1, k_2, \ldots, k_{m-1}) = \sum_{s=1}^{L} b_s C(\alpha s)\left[\prod_{l=1}^{m-1} J_{kl}(\alpha s A_l)\right] \quad . \quad \text{(A27)}$$

If there is no noise and the input consists of $m-1$ modulated carriers,

$$M = M_S = \sum_{s=1}^{L} b_s\left[\prod_{l=1}^{m-1} J_{kl}(\alpha s A_l)\right] \quad . \quad \text{(A28)}$$

Similarly, for pure signals and intermodulation products, if noise is present, the nonlinear characteristic $g(\rho)\exp[jf(\rho)]$ becomes

$$g(\rho)\exp[jf(\rho)] = \sum_{s=1}^{L} b_s C(\alpha s)\, J_1(\alpha s\rho) \quad . \quad \text{(A29)}$$

Demodulated output

In equation (A1), it has been assumed that all of the amplitudes $A_l (l = 1,2, \ldots, m-1)$ are constants and that the last component ($l = m$) represents thermal noise. The problem is to demodulate an angle $\theta_p(t)$ at the output. It can be assumed that $p = 1$ without loss of generality. The output can then be represented by the following three categories:

a. the main carrier to be demodulated;

b. the intermodulation products and noise falling in the receiver filter of this main carrier; and

c. the other carriers, intermodulation products, and noise falling away from the main carrier, which can be filtered out.

Only categories a and b are important in demodulating the main carrier. The components of b and c are represented as follows:

$$e_{o1}(t) = Re\,\{\exp[j\omega_o t + j\theta_1(t)]\, M_o\} + Re\,\{\exp[j\omega_o t + j\theta_1(t)]\,[M(1, 0, \ldots, 0; t) - M_o]\} + Re\left\{\exp[j\omega_o t] \sum_{\substack{k_1,k_2,\ldots,k_m=-\infty \\ (k_1+k_2+\ldots+k_m=1)}}^{\infty}{}' \exp\left[j\sum_{l=1}^{m-1} k_l\theta_l(t)\right] \cdot M(k_1, k_2, \ldots, k_m; t)\exp[jk_m\theta_m(t)]\right\} \quad \text{(A30)}$$

where

$$M_o = \int_0^\infty \gamma\left[\prod_{l=2}^{m-1} J_o(\gamma A_l)\right] J_1(\gamma A_1)\, C(\gamma)\, d\gamma \cdot \int_0^\infty \rho g(\rho)\exp[jg(\rho)]\, J_1(\gamma\rho)\, d\rho \quad \text{(A31)}$$

$$M(1, 0, \ldots, 0; t) = \int_0^\infty \gamma \left[\prod_{l=2}^{m-1} J_0(\gamma A_l) \right] J_1(\gamma A_1) \cdot J_0\left[\gamma \sqrt{N_c^2(t) + N_s^2(t)}\right] d\gamma \cdot \int_0^\infty \rho g(\rho) \exp\left[jf(\rho)\right] J_1(\gamma\rho)\, d\rho \qquad \text{(A32)}$$

$$M(k_1, k_2, \ldots, k_m) \exp\left[jk_m\theta_m(t)\right] = \int_0^\infty \gamma \left[\prod_{l=1}^{m-1} J_{k_l}(\gamma A_l) \right] J_{k_m}\left[\gamma \sqrt{N_c^2(t) + N_s^2(t)}\right] d\gamma \exp\left[jk_m\omega_m t\right] \cdot \int_0^\infty \rho g(\rho) \exp\left[jf(\rho)\right] J_1(\gamma\rho)\, d\rho \left[\frac{N_c(t) + jN_s(t)}{\sqrt{N_c^2(t) + N_s^2(t)}} \right]^{k_m} . \qquad \text{(A33)}$$

The prime following the summation indicates that categories a and c are excluded. Equation (A30) is represented as follows:

$$e_{o1}(t) = Re \exp\left[j\omega_o t + j\theta_1(t)\right] M_o[1 + R(t) + jI(t)] . \qquad \text{(A34)}$$

The demodulated output can now be represented by

$$\phi_1(t) + \tan^{-1}\left[\frac{I(t)}{1 + R(t)} \right] . \qquad \text{(A35)}$$

Since, in the normal situation, $R(t)$ and $I(t)$ are small, equation (A35) can be approximated as

$$\phi_1(t) + \tan^{-1}\left[\frac{I(t)}{1 + R(t)} \right] \approx \phi_1(t) + I(t) . \qquad \text{(A36)}$$

The demodulated intermodulation products and noise are given by

$$I(t) = Im\left\{M_o^{-1}\left[M(1, 0, \ldots, 0; t) - M_o\right]\right\} + Im\left\{M_o^{-1} \sum_{\substack{k_1, k_2, \ldots, k_m = -\infty \\ (k_1 + k_2 + \ldots + k_m = 1)}}^{\infty\prime} M(k_1, k_2, \ldots, k_m; t) \cdot \exp\left[jk_m \left\{ \omega_m t + \tan^{-1}\frac{N_s(t)}{N_c(t)} \right\} \right] \cdot \exp\left[j(k_1 - 1)\,\theta_1(t) + j \sum_{l=2}^{m-1} k_l\theta_l(t) \right] \right\} \qquad \text{(A37)}$$

where, for the particular case of the Bessel function expansion,

$$M_o = \sum_{s=1}^{L} b_s \exp\left[-\frac{\alpha^2 s^2}{2} R(0) \right] J_1(\alpha s A_1) \prod_{l=2}^{m-1} J_0(\alpha s A_l) \qquad \text{(A38)}$$

and

$$M(k_1, k_2, \ldots, k_m; t) = \sum_{s=1}^{L} b_s J_{k_m}\left[\alpha s \sqrt{N_c^2(t) + N_s^2(t)}\right] \prod_{l=1}^{m-1} J_{k_l}(\alpha s A_l) . \qquad \text{(A39)}$$

The autocorrelation function of $I(t)$, i.e., avg $\{I(t) \cdot I(t + \tau)\}$, may now be obtained in the following manner. In the average of $I(t) \cdot I(t + \tau)$, the cross-terms are not all zero; i.e., the cross-correlation functions of the two components $(k_1, k_2, \ldots, k_m)$ and $(2 - k_1, -k_2, -k_3, \ldots, -k_m)$ are not zero. (All other terms in the cross-correlation function are zero.) Then $I(t)$ of equation (A37) can be expressed as

$$I(t) = Im\left\{ M_o^{-1}\left[M(1, 0, \ldots, 0; t) - M_o\right] + M_o^{-1} \sum_{\substack{k_1, k_2, \ldots, k_m = -\infty \\ (k_1 + k_2 + \ldots + k_m = 1)}}^{\infty\prime\prime} \exp\left[j(k_1 - 1)\,\theta_1(t) + j \sum_{l=2}^{m-1} k_l\theta_l(t) \right] M(k_1, k_2, \ldots, k_m; t) \cdot \exp\left[jk_m \left\{ \omega_m t + \tan^{-1}\frac{N_s(t)}{N_c(t)} \right\} \right] + M(2 - k_1, -k_2, \ldots, -k_m; t) \cdot \exp\left[-j(k_1 - 1)\,\theta_1(t) - j \sum_{l=2}^{m-1} k_l\theta_l(t) \right] \cdot \exp\left[-jk_m \left\{ \omega_m t + \tan^{-1}\frac{N_s(t)}{N_c(t)} \right\} \right] \right\} \qquad \text{(A40)}$$

where Σ' is rewritten as Σ'' by combining the components $(k_1, k_2, \ldots, k_m)$ and $(2 - k_1, -k_2, \ldots, -k_m)$.

The relationship between arbitrary complex numbers a and b,

$$[Im\ a]\,[Im\ b] = \frac{1}{2}\,Re\,\{ab^* - ab\} \tag{A41}$$

and

$$\begin{aligned}&\text{avg}\left\{J_{km}\left[\alpha s\sqrt{N_c^2(t)+N_s^2(t)}\right]J_{km}\left[\alpha p\sqrt{N_c^2(t+\tau)+N_s^2(t+\tau)}\right]\right.\\&\left.\cdot\ \exp\left[\pm k_m j\tan^{-1}\frac{N_s(t)}{N_c(t)}\right]\exp\left[\mp k_m j\tan^{-1}\frac{N_s(t+\tau)}{N_c(t+\tau)}\right]\right\}\\&=\exp\left[-\frac{\alpha^2(s^2+p^2)}{2}R(0)\right]I_{km}[\alpha^2 spR(\tau)]\end{aligned} \tag{A42}$$

yield

$$\begin{aligned}&\text{avg}\,\{I(t)\,I(t+\tau)\}\\&=R_1(\tau)+\sum_{k_1,k_2,\ldots,k_{m-1}=-\infty}^{\infty}{}''\,R(k_1,k_2,\ldots,k_{m-1};\tau)\end{aligned} \tag{A43}$$

where

$$\begin{aligned}R_1(\tau) &= \frac{1}{2}\,Re\left[\sum_{s=1}^{L}\sum_{p=1}^{L}\frac{b_s}{M_o}\left(\frac{b_p}{M_o}\right)^* J_1(\alpha sA_1)\,J_1(\alpha pA_1)\right.\\&\cdot\ \exp\left\{-\frac{\alpha^2(s^2+p^2)}{2}R(0)\right\}\prod_{l=2}^{m-1}J_0(\alpha sA_l)\prod_{l=2}^{m-1}J_0(\alpha pA_l)\\&\cdot\ \{I_0[\alpha^2 spR(\tau)]-1\}-\sum_{s=1}^{L}\sum_{p=1}^{L}\frac{b_s}{M_o}\frac{b_p}{M_o}J_1(\alpha sA_1)\,J_1(\alpha pA_1)\\&\cdot\ \exp\left\{-\frac{\alpha^2(s^2+p^2)}{2}R(0)\right\}\prod_{l=2}^{m-1}J_0(\alpha sA_l)\prod_{l=2}^{m-1}J_0(\alpha pA_l)\\&\left.\cdot\ \{I_0[\alpha^2 spR(\tau)]-1\}\right]\\&=\sum_{q=1}^{\infty}\left\{Im\left[\frac{N(1,0,\ldots,0;q)}{M_o}\right]\right\}^2[\rho_o(\tau)]^{2q}\ .\end{aligned} \tag{A44}$$

In equation (A44),

$$\begin{aligned}N(k_1,k_2,\ldots,k_{m-1};q) &= \sum_{s=1}^{L}b_s\exp\left[-\frac{\alpha^2 s^2}{2}R(0)\right]\prod_{l=1}^{m-1}J_{kl}(\alpha sA_l)\\&\cdot\left\{\frac{\left[\frac{1}{2}R(0)\,\alpha^2 s^2\right]^{2q+|k_m|}}{q!(|k_m|+q)!}\right\}^{1/2}\\\rho_o(\tau) &= \frac{R(\tau)}{R(0)}\end{aligned}$$

$$\begin{aligned}&R(k_1,k_2,\ldots,k_{m-1};\tau)\\&=\frac{1}{2}\,Re\,|M_o|^{-2}\sum_{s=1}^{L}\sum_{p=1}^{L}b_s b_p^*\exp\left[-\frac{\alpha^2(s^2+p^2)}{2}R(0)\right]J_{k1}(\alpha sA_1)\\&\cdot\ J_{k1}(\alpha pA_1)\prod_{l=2}^{m-1}J_{kl}(\alpha sA_l)\prod_{l=2}^{m-1}J_{kl}(\alpha pA_l)\,I_{km}[\alpha^2 spR(\tau)]\\&\cdot\ \text{avg}\left\{\exp\left[j(k_1-1)(\omega_1\tau+\psi_1)+\sum_{l=2}^{m-1}jk_l\psi_l+\sum_{l=2}^{m}j\omega_l\tau\right]\right\}\\&+\frac{1}{2}\,Re\,|M_o|^{-2}\sum_{s=1}^{L}\sum_{p=1}^{L}b_s b_p^*\exp\left[-\frac{\alpha^2(s^2+p^2)}{2}R(0)\right]J_{2-k1}(\alpha sA_1)\\&\cdot\ J_{2-k1}(\alpha pA_1)\prod_{l=2}^{m-1}J_{-kl}(\alpha sA_l)\prod_{l=2}^{m-1}J_{-kl}(\alpha sA_l)\,I_{-km}[\alpha^2 spR(\tau)]\\&\cdot\ \text{avg}\left\{\exp\left[-j(k_1-1)(\omega_1\tau+\psi_1)-\sum_{l=2}^{m-1}jk_l\psi_l-\sum_{l=2}^{m}j\omega_l\tau\right]\right\}\\&-\frac{(-1)^{k_m}}{2}\,Re\left\{M_o^{-2}\sum_{s=1}^{L}\sum_{p=1}^{L}b_s b_p\exp\left[-\frac{\alpha^2(s^2+p^2)}{2}R(0)\right]\right.\\&\left.\cdot\ J_{k1}(\alpha sA_1)\,J_{2-k1}(\alpha pA_1)\right\}\\&\cdot\ \prod_{l=2}^{m-1}J_{kl}(\alpha sA_l)\prod_{l=2}^{m-1}J_{-kl}(\alpha pA_l)\,I_{km}[\alpha^2 spR(\tau)]\\&\cdot\ \text{avg}\left\{\exp\left[j(k_1-1)(\omega_1\tau+\psi_1)+\sum_{l=2}^{m-1}jk_l\psi_l+\sum_{l=2}^{m}j\omega_l\tau\right]\right\}\\&-\frac{(-1)^{k_m}}{2}\,Re\left\{M_o^{-2}\sum_{s=1}^{L}\sum_{p=1}^{L}b_s b_p\exp\left[-\frac{\alpha^2(s^2+p^2)}{2}R(0)\right]\right.\\&\left.\cdot\ J_{2-k1}(\alpha sA_1)\,J_{k1}(\alpha pA_1)\right\}\\&\cdot\ \prod_{l=2}^{m-1}J_{-kl}(\alpha sA_l)\prod_{l=2}^{m-1}J_{kl}(\alpha pA_l)\,I_{km}[\alpha^2 spR(\tau)]\\&\cdot\ \text{avg}\left\{\exp\left[-j(k_1-1)(\omega_1\tau+\psi_1)-\sum_{l=2}^{m-1}jk_l\psi_l-\sum_{l=2}^{m}j\omega_l\tau\right]\right\}\\&=\frac{1}{2}\,Re\sum_{q=0}^{\infty}\left\{\left|\frac{N(k_1,k_2,\ldots,k_{m-1};q)}{M_o}\right|^2\right.\end{aligned}$$

(continued on next page)

$$\cdot\ \Delta(k_1 - 1, k_2, \ldots, k_{m-1}; q; \tau)$$

$$+ \left| \frac{N(k_1 - 2, k_2, \ldots, k_{m-1}; q)}{M_o} \right|^2$$

$$\cdot\ \Delta^*(k_1 - 1, k_2, \ldots, k_{m-1}; q; \tau)$$

$$+ \frac{N(k_1, k_2, \ldots, k_{m-1}; q)\ N(k_1 - 2, k_2, \ldots, k_{m-1}; q)}{M_o^2}$$

$$\cdot\ \Delta(k_1 - 1, k_2, \ldots, k_{m-1}; q; \tau)$$

$$+ \frac{N(k_1 - 2, k_2, \ldots, k_{m-1}; q)\ N(k_1, k_2, \ldots, k_{m-1}; q)}{M_o^2}$$

$$\cdot\ \Delta^*(k_1 - 1, k_2, \ldots, k_{m-1}; q; \tau) \Bigg\} \qquad \text{(A45)}$$

where

$$\psi_l = \phi_l(t) - \phi_l(t + \tau) \qquad \text{(A46)}$$

and

$$\Delta(k_1, k_2, \ldots, k_{m-1}; q; \tau)$$

$$= \text{avg} \left\{ \exp\left[j \sum_{l=1}^{m-1} k_l(\omega_l \tau + \psi_l) + j k_m \omega_m \tau \right] \right\} [\rho_o]^{2q + |k_m|} \quad . \qquad \text{(A47)}$$

If Δ is real, equation (A45) becomes

$$R(k_1, k_2, \ldots, k_{m-1}; \tau)$$

$$= \frac{1}{2} \sum_{q=0}^{\infty} \frac{|N(k_1, k_2, \ldots, k_{m-1}; q) + N^*(k_1 - 2, k_2, \ldots, k_{m-1}; q)|^2}{|M_o|^2}$$

$$\cdot\ \Delta(k_1 - 1, k_2, \ldots, k_{m-1}; q; \tau) \quad . \qquad \text{(A48)}$$

For the special case of $k_1 = 1$, equation (A45) can be reduced to

$$R(k_1, k_2, \ldots, k_{m-1}; \tau) = 2 \sum_{q=0}^{\infty} \left\{ \left[Im \frac{N(1, k_2, \ldots, k_{m-1}; q)}{M_o} \right]^2 \right\}$$

$$\cdot\ Re\ \Delta(0, k_2, \ldots, k_{m-1}; q; \tau) \qquad \text{(A49)}$$

where

$$N(1, k_2, \ldots, k_{m-1}; q) = -N(-1, k_2, \ldots, k_{m-1}; q) \quad . \qquad \text{(A50)}$$

Note that, in the preceding derivations, the power series expansion for $I_{k_m}(z)$, namely,

$$I_{k_m}(z) = \sum_{q=0}^{\infty} \frac{(1/2)^{|k_m| + 2q}}{q!(|k_m| + q)!} z^{|k_m| + 2q} \qquad \text{(A51)}$$

was used. Note also that, in the case of $k_1 = 1$, if there is no AM/PM conversion, equation (A49) becomes zero. However, in all other cases ($k_1 \neq 1$), even if there is no AM/PM conversion, equation (A49) does not become zero.

The order of the intermodulation products is defined as

$$|k_1| + |k_2| + \ldots + |k_{m-1}| + |k_m| + 2q \quad . \qquad \text{(A52)}$$

In the expressions for $R(k_1, k_2, \ldots, k_{m-1}; \tau)$, the magnitude and spread of the power spectrum are represented separately; i.e., the Fourier transform of Δ gives the power spectrum spread whose total power is unity, and the other coefficient gives the magnitude.

Power spectrum analysis

In this section, it is assumed that the input consists of $m - 1$ angle-modulated carriers and a Gaussian noise signal. As in Reference A1, the autocorrelation function of the output is given by*

$$\text{avg}\ [e_o(t)\ e_o(t + \tau)] = \frac{1}{2} Re \sum_{k_1, k_2, \ldots, k_{m-1} = -\infty}^{\infty} \exp\left[j\omega_o \tau + j \sum_{l=1}^{m} \omega_l \tau \right] \text{avg} \left\{ \exp\left[j \sum_{l=1}^{m-1} k_l \psi_l \right] \right\}$$

$$\cdot \int_0^{\infty} t \exp\left[-\frac{t^2}{2} \right]$$

$$\cdot\ |Q(k_1, k_2, \ldots, k_m; t)|^2\ dt \qquad \text{(A53)}$$

where the averages are taken over the values

$$\phi_1(t), \phi_2(t), \ldots, \phi_{m-1}(t) \quad .$$

Then, rewriting equation (13) in [A1] yields

$$Q(k_1, k_2, \ldots, k_{m-1}; t)$$

$$= \int_0^{\infty} \gamma \prod_{l=1}^{m-1} J_{k_l}(\gamma A_l)\ J_{k_m}\left[\sqrt{R(\tau)}\ \gamma t \right] \exp\left\{ -[R(0) - R(\tau)] \frac{\gamma^2}{2} \right\} d\gamma$$

$$\cdot \int_0^{\infty} \rho g(\rho) \exp\left[j f(\rho) \right] J_1(\gamma\rho)\ d\rho \qquad \text{(A54)}$$

*The result obtained in this section can also be obtained directly from equations (A12), (A13), and (A26) if it is assumed that all of the cross-terms in avg $\{e_i(t)\ e_i(t + \tau)\}$ are zero.

where $$k_m = 1 - \sum_{l=1}^{m-1} k_l.$$

The method used to evaluate $Q(k_1, k_2, \ldots, k_m; t)$ is described in Reference A1. The case in which the nonlinearity characteristics are approximated by a sum of Bessel functions, as in equation (A24), is analyzed here, since this case is not analyzed in Reference A1. Expansion of equation (A24) and use of equation (A25) yield

$$|Q(k_1, k_2, \ldots, k_{m-1}; t)|^2 = \left| \sum_{s=1}^{L} b_s \left[\prod_{l=1}^{m-1} J_{k_l}(\alpha s A_l) \right] \cdot \exp\left\{ -\frac{\alpha^2 s^2}{2} [R(0) - R(\tau)] \right\} \cdot J_{k_m}[\sqrt{R(\tau)}\, \alpha s t] \right|^2 . \quad \text{(A55)}$$

Therefore, the integral with respect to t in equation (A53) is given by

$$\sum_{s=1}^{L} \sum_{p=1}^{L} b_s b_p^* \left[\prod_{l=1}^{m-1} J_{k_l}(\alpha s A_l) \right] \left[\prod_{l=1}^{m-1} J_{k_l}(\alpha p A_l) \right] \cdot \exp\left\{ -\frac{\alpha^2(s^2 + p^2)}{2} [R(0) - R(\tau)] \right\} \int_0^\infty \exp\left[-\frac{t^2}{2} \right] t \cdot J_{k_m}[\sqrt{R(\tau)}\, \alpha p t]\, J_{k_m}[\sqrt{R(\tau)}\, \alpha s t]\, dt . \quad \text{(A56)}$$

The formula given in Reference A3 [equation (1), p. 395] makes it possible to evaluate the integral and reduce equation (A56):

$$H(k_1, k_2, \ldots, k_{m-1}) = \sum_{s=1}^{L} \sum_{p=1}^{L} b_s b_p^* \left[\prod_{l=1}^{m-1} J_{k_l}(\alpha s A_l) \right] \cdot \left[\prod_{l=1}^{m-1} J_{k_l}(\alpha p A_l) \right] \exp\left[-\frac{\alpha^2(s^2 + p^2)}{2} R(0) \right] \cdot I_{k_m}[R(\tau)\, \alpha^2 s p] . \quad \text{(A57)}$$

Expanding the Bessel function as a power series

$$H(k_1, k_2, \ldots, k_{m-1}) = \sum_{q=0}^{\infty} |N(k_1, k_2, \ldots, k_{m-1}; q)|^2 \cdot [\rho_o(\tau)]^{2q + |k_m|} \quad \text{(A58)}$$

yields the autocorrelation function of the output:

$$\begin{aligned} &\text{avg}\,[e_o(t)\, e_o(t + \tau)] \\ &= \frac{1}{2} Re \sum_{k_1, k_2, \ldots, k_{m-1} = -\infty}^{\infty} \exp\left[j\omega_o \tau + j \sum_{l=1}^{m-1} k_l \omega_l \tau \right] \\ &\quad \cdot \text{avg}\left\{ \exp\left[j \sum_{l=1}^{m-1} k_l \psi_l \right] \right\} H(k_1, k_2, \ldots, k_{m-1}) \\ &= \frac{1}{2} Re \sum_{k_1, k_2, \ldots, k_{m-1} = -\infty}^{\infty} \sum_{q=0}^{\infty} |N(k_1, k_2, \ldots, k_{m-1}; q)|^2 \\ &\quad \cdot \Delta(k_1, k_2, \ldots, k_{m-1}; q; \tau) . \end{aligned} \quad \text{(A59)}$$

Acknowledgment

The authors wish to acknowledge the helpful suggestions of numerous colleagues concerning the assumptions inherent in the mathematical model and the validity of the results. In particular, the insight and encouragement offered by Dr. N. K. M. Chitre and the contributions of Dr. R. Fang have been invaluable.

References

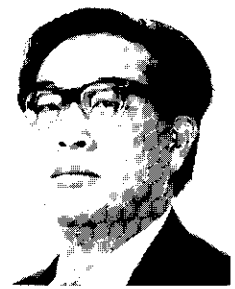
[A1] O. Shimbo, "Effects of Intermodulation, AM-PM Conversion, and Additive Noise in Multicarrier TWT Systems," *Proceedings of the IEEE*, Vol. 59, No. 2, February 1971, pp. 230–238.

[A2] A. R. Kaye, D. A. George, and N. J. Eric, "Analysis and Compensation of Bandpass Nonlinearities for Communications," *IEEE Transactions on Communications*, COM-20, No. 5, October 1972, pp. 965–972.

[A3] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd Edition, Cambridge, U.K.: University Press, 1966.

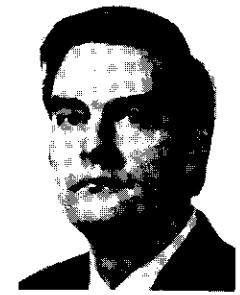
Jorge C. Fuenzalida was born in Santiago, Chile. After finishing engineering school, he emigrated to the United States in 1965. He received the M.S.E.E. degree from Columbia University in 1966, and subsequently was a Ph.D. candidate in systems engineering at the University of Pennsylvania. Mr. Fuenzalida was a Member of the Technical Staff in the Systems Laboratory at COMSAT Laboratories, working on interference and nonlinear intermodulation in FM systems, on satellite capacity optimization problems, on frequency sharing between satellites and radio-relay systems, and on orbit and spectrum sharing. He has been active in the Study Group 4 of the CCIR.

He is now with the Systems Engineering Division, American Satellite Corporation.



Osamu Shimbo is Senior Scientist, Advanced Studies Laboratory, COMSAT Laboratories. Before joining COMSAT, he was a Senior Member of the Scientific Staff in the Research and Development Laboratories, Northern Electric Co., Ltd.; Senior Scientific Staff Member at Hirst Research Centre, General Electric Co., Ltd., of England; an exchange visitor in the Department of Electrical Engineering, Columbia University; and a Manager in the Transmission Laboratory of Oki Electric Industrial Co., Ltd., of Japan.

He holds a Bachelor of Engineering degree from Tohoku University, Japan (1956), and a Doctor of Engineering degree from Hokkaido University of Japan (1965) and has received awards for a paper on "Synchronization of PCM Systems" and for achievement in FM and PCM systems analysis.



William L. Cook was born in Baltimore, Maryland. He received a B.S. degree in engineering mechanics from Lehigh University (1964), an M.S. degree in engineering sciences from Purdue University (1966), and the D.Sc. in computer sciences from The George Washington University (1973). He is currently a Member of the Technical Staff of the Scientific Computer Applications Department of COMSAT Laboratories.

Dr. Cook is a member of Tau Beta Pi, Sigma Xi, and the Association for Computing Machinery.

Index: communications satellites, earth terminals, ionospheric propagation, scintillation, statistical distributions.

Ionospheric scintillation at 4 and 6 GHz

R. R. TAUR

Abstract

Based on approximately 15 months of observation made at satellite earth stations around the world, the probability distribution of the amplitude fluctuation of ionospheric scintillation at 12 geomagnetic equatorial locations has been obtained. The scintillation activity, which is found between 30°N and 30°S geomagnetic latitudes, with a higher occurrence rate near the geomagnetic equator, shows a very strong diurnal peak at about one hour after local sunset and seasonal peaks near the vernal and autumnal equinoxes. It is hypothesized that the observed scintillation is caused by the very dense and thick irregular layers in the F-region, which may exist only in the early part of the evening.

Introduction

At radio frequencies near 4 and 6 GHz, the ionosphere is essentially nonabsorbent. However, rapid fluctuations of the signal amplitude with time have been reported by many earth stations of the INTELSAT network. After the other possible causes were carefully eliminated, it appeared

---

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.

that these fluctuations were caused by irregularities in the ionosphere. Based upon a short period of observation, a preliminary report on ionospheric scintillation at 4 and 6 GHz was prepared [1]. The amplitude fluctuation of the scintillation was generally found to be less than ± 4 dB, with periods of about four to six seconds between two adjacent fades. The scintillation was observed mainly at the stations in the geomagnetic equatorial region.

Beginning in August 1970, continuous recordings of ionospheric scintillation were made at satellite earth stations at Andover, Maine; Etam, West Virginia; Jamesburg, California; Paumalu, Hawaii; Goonhilly, United Kingdom; Raisting, Germany; and Ras Abu Jarjur, Bahrain. Each station monitored the amplitudes of several carriers from selected distant transmitting stations via satellite. The locations of the recording and transmitting stations are shown in Figure 1, and the appropriate geomagnetic latitude and elevation angle of each station are listed in Table 1. The up- and down-link frequencies were 6 and 4 GHz, respectively. Each recording station monitored four to seven carriers of the distant transmitting stations and the beacon of the satellite. If all carriers plus the beacon

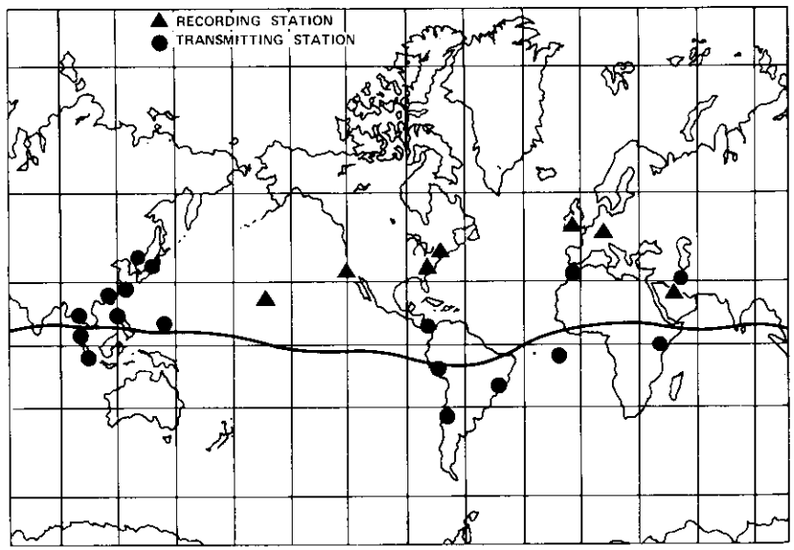


Figure 1. *Recording and Transmitting Stations*

received at a recording station were fluctuating in an identical pattern, scintillation was assumed to have occurred at the down-link, i.e., at 4 GHz; otherwise, it was assumed that ionospheric irregularities were associated with the 6-GHz up-link between the transmitting station and the satellite.

Approximately 15 months of data were collected at each station. These data have been reduced and analyzed, and the probability distributions of amplitude fluctuation, diurnal and seasonal variations, and geomagnetic latitudinal dependence are presented in this paper. A possible mechanism is hypothesized to explain the observed scintillation.

Theoretical background

Although the normal ionosphere is essentially nonabsorbent at 4 and 6 GHz, the ionosphere contains scattering irregularities which distort the phase of the wave front. As these irregularities drift through the propagation path, the received signal at the earth station fluctuates with time; this fluctuation is called scintillation. The modulation index associated with scintillation in the case of weak scattering is approximately proportional to $\lambda Z/r_0^2$ and ϕ_0, where λ is the wavelength in free space, Z is the height of the irregularity, r_0 is the size of a single irregularity, and ϕ_0 is the rms phase deviation as the wave passes through the irregularity [2], [3]. Since ϕ_0 is proportional to λ, it would be expected that the modulation index

TABLE 1. PARTICIPATING EARTH STATIONS

Station	Latitude	Longitude	Approximate Geomagnetic Latitude	Approximate Elevation Angle
Ras Abu Jarjur, Bahrain	26°04′N	50°37′E	19°N	57°
Tangua, Brazil	22°44′S	42°46′W	12°S	60°
Longovilo, Chile	33°57′S	71°24′W	20°S	32°
Taipei, Republic of China	25°10′N	121°34′E	15°N	26°

TABLE 1. PARTICIPATING EARTH STATIONS (CONTINUED)

Station	Latitude	Longitude	Approximate Geomagnetic Latitude	Approximate Elevation Angle
Mt. Margaret, East Africa	01°01′S	26°30′E	8°S	60°
Raisting 1, 2,* Germany	47°54′N	11°07′E	50°N	25°, 16°
Djatiluhur, Indonesia	06°31′S	107°25′E	18°S	38°
Asadabad, Iran	34°45′N	48°06′E	27°N	6°
Yamaguchi, Japan	34°12′N	131°33′E	23°N	9°
Kum San, Korea	36°07′N	127°29′E	25°N	26°
Kuantan, Malaysia	03°00′N	101°30′E	5°S	43°
Utibe, Panama	09°99′N	79°30′W	20°N	34°
Lurin, Peru	12°17′S	76°51′W	0°N	36°
Si Racha, 1, 2, Thailand	13°06′N	100°56′E	5°N	8°, 43°
Goonhilly 1,* U.K.	50°03′N	05°10′W	53°N	5°
Ascension Island, U.K.	07°57′S	14°23′W	3°S	75°
Hong Kong 1, U.K.	22°12′N	114°13′E	13°N	20°
Andover,* U.S.	44°39′N	70°43′W	55°N	21°
Etam,* U.S.	39°17′N	79°45′W	50°N	23°

*Control station.

TABLE 1. PARTICIPATING EARTH STATIONS (CONTINUED)

Station	Latitude	Longitude	Approximate Geomagnetic Latitude	Approximate Elevation Angle
Jamesburg,* U.S.	36°24′N	121°39′W	43°N	12°
Guam, U.S.	13°24′N	144°42′E	9°N	53°
Paumalu 1, 2,* U.S.	21°40′N	158°02′W	20°N	66°, 48°

*Control station.

would have a λ^2 dependence. This λ^2 dependence at 4 and 6 GHz has been observed in Reference 1.

The transverse radius of the irregularity that can cause scintillation at 4 and 6 GHz is in the range of 150–300 m. The fluctuation of the electron density of the irregularity depends upon L, the thickness of the disturbed layer. According to Reference 1, for a modulation index of about 0.15 (peak-to-peak deviation of about 4 dB) at 4 GHz, the product $(\Delta N)^2 \cdot L$ proves to be $6 \cdot 10^6$ electrons^2/cm^5, where ΔN is the average fluctuation in electron density from the ambient density. When ΔN is equal to 10 percent of the typical equatorial F-region ambient densities during the evening, the corresponding value of L is approximately 100 km. Such a thick disturbed layer can probably be formed only by large-scale instability in the ionosphere, such as that which would occur after the rapid disappearance of the major ion-producting mechanism, the solar radiation. Therefore, one would expect to see more scintillation of the signal near the local sunset hour. This will be discussed in a later section entitled "Diurnal Variation."

Data reduction

Data available

All of the data were recorded on multichannel stripcharts at a relatively slow speed. The signal bandwidths were relatively narrow and ranged from 2.5 to 36 MHz (less than 1 percent of the operating frequency). The details of each fluctuation cannot be seen; however, the peak amplitudes are

clearly marked on the stripchart and can easily be identified. The recordings were made continuously during the 15-month period, except when there were equipment malfunctions or rearrangements. Figure 2 shows the periods of availability of the scintillation data. A break in the horizontal line indicates a lack of data.

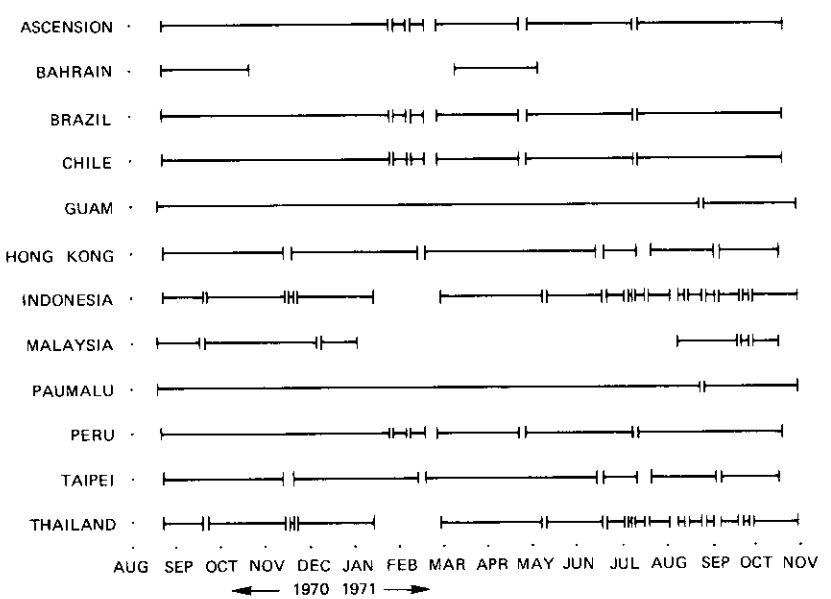


Figure 2. *Scintillation Data Available*

Data reduction procedures

The original data recordings were carefully calibrated and labeled. Slow scintillation caused by atmospheric anomalies at low elevation angles, and the small, rapid type of signal variation caused by rain were both excluded. Ionospheric scintillations with fluctuations greater than ± 0.25 dB were recorded. The amplitude fluctuation distributions were obtained by summing the total hours of the approximate peak magnitudes of scintillation at 0.25 dB and at each 0.5-dB increment starting from 0.5 dB. The diurnal, seasonal, and geomagnetic latitudinal variations were obtained through similar procedures; the results will be presented later in this paper.

Results and discussion

Amplitude distribution

The yearly cumulative amplitude distribution of each station is shown in Figures 3–8. It should be noted that the results shown here represent the probability of scintillation with a peak-to-peak variation greater than the value of the ordinate. Therefore, if one is interested in only the fading of the ionospheric scintillation, both the abscissas and the ordinates in these figures should be approximately halved.

For frequencies other than 6 GHz, the λ^2 dependence can be used to obtain the approximate corresponding amplitude distribution. However, at frequencies below 2 or 3 GHz, the λ^2 dependence has not been demonstrated. Therefore, extrapolation of the results for frequencies below 3 GHz is not recommended.

Seasonal variation

The duration of each measurable scintillation event is registered at each station. The corresponding monthly scintillation activity of each earth station is given in Figures 9–11. This activity is shown as an average monthly number of minutes of measurable scintillation, i.e., signal fluctuation greater than ± 0.5 dB for most of the earth stations at 6 GHz (up-link), per day. One interesting fact concerning the monthly variation is that the ionospheric scintillation in equatorial regions definitely has a rather strong seasonal dependence. It can be seen that most of the scintillation events occur near the vernal and autumnal equinoxes and that the autumnal peak is generally stronger than the vernal peak.

The observed seasonal variation of the ionospheric scintillation is compared with that of the equatorial Spread F,* as shown in Figure 12 [4]. Since Djibouti's coordinates are 11°N, 42°E, only the Ras Abu Jarjur (Bahrain) data are used in the comparison of the diurnal variations of Spread F and ionospheric scintillation. The comparison indicates that the F-region irregularities which are responsible for scintillation at 4 and 6 GHz quite possibly differ from the irregularities that cause Spread F. The difference is believed to be in the relative density and size of the irregularities.

* Spread F refers to the spreading of the returned signal transmitted by the ionospheric sounder and indicates the presence of ionospheric irregularities.

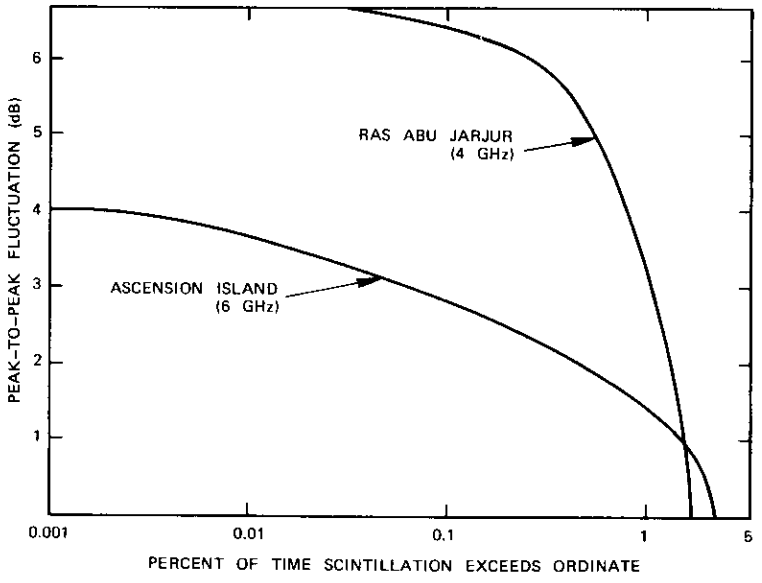


Figure 3. *Cumulative Amplitude Distributions at Ascension Island (6 GHz) and Ras Abu Jarjur (4 GHz)*

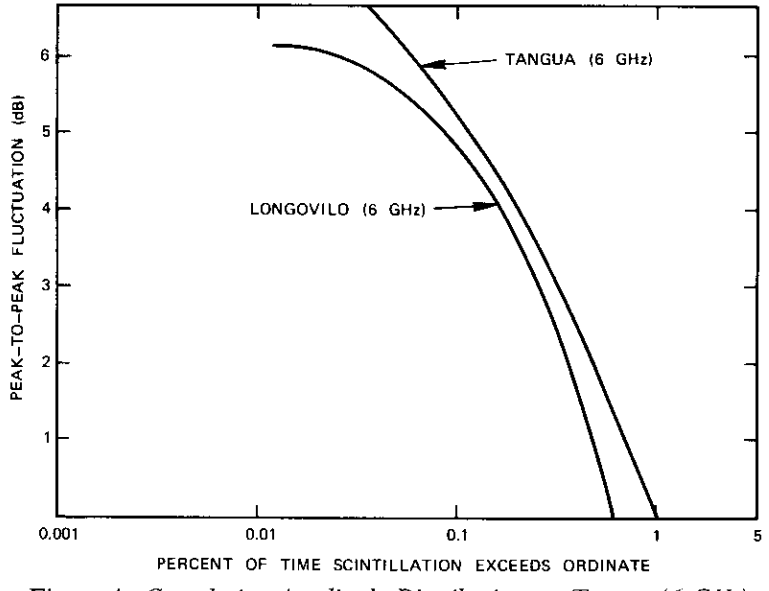


Figure 4. *Cumulative Amplitude Distributions at Tangua (6 GHz) and Longovilo (6 GHz)*

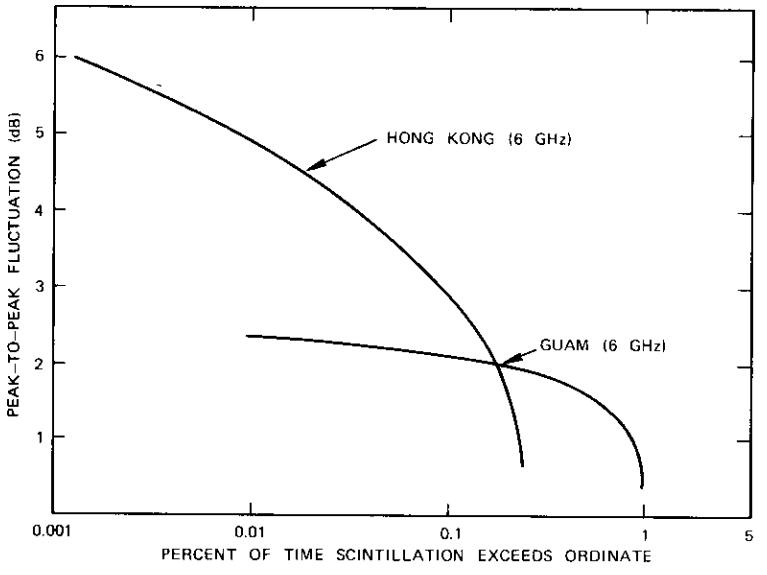


Figure 5. *Cumulative Amplitude Distributions at Guam (6 GHz) and Hong Kong (6 GHz)*

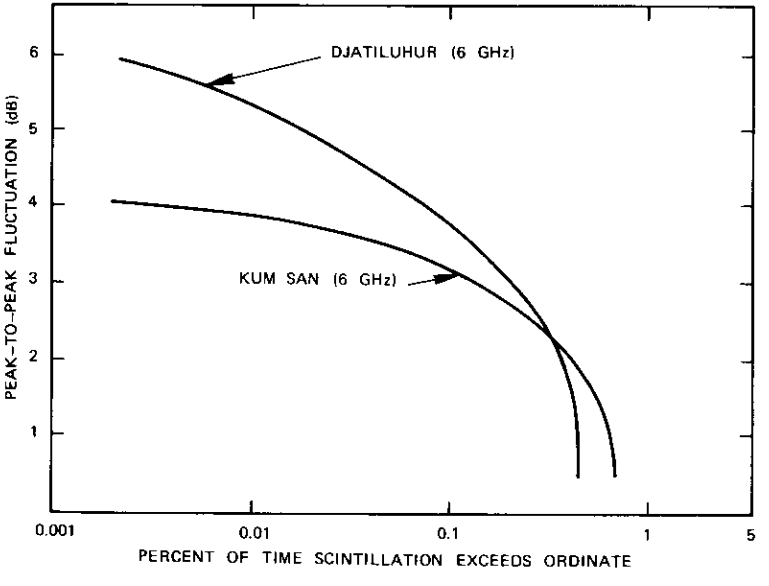


Figure 6. *Cumulative Amplitude Distributions at Djatiluhur (6 GHz) and Kum San (6 GHz)*

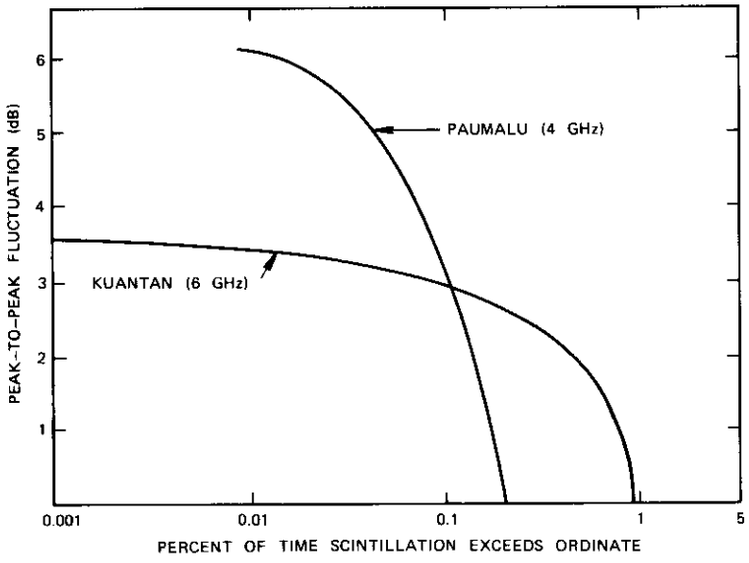


Figure 7. *Cumulative Amplitude Distributions at Kuantan (6 GHz) and Paumalu (4 GHz)*

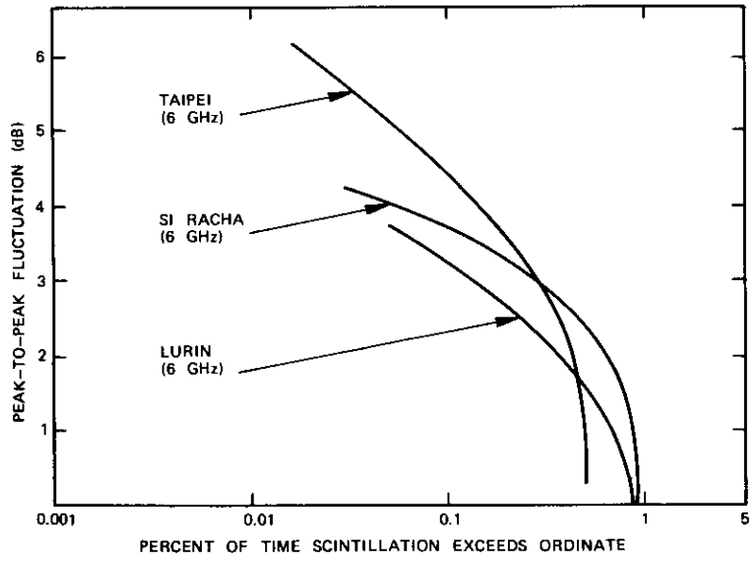


Figure 8. *Cumulative Amplitude Distributions at Lurin (6 GHz), Taipei (6 GHz), and Si Racha (6 GHz)*

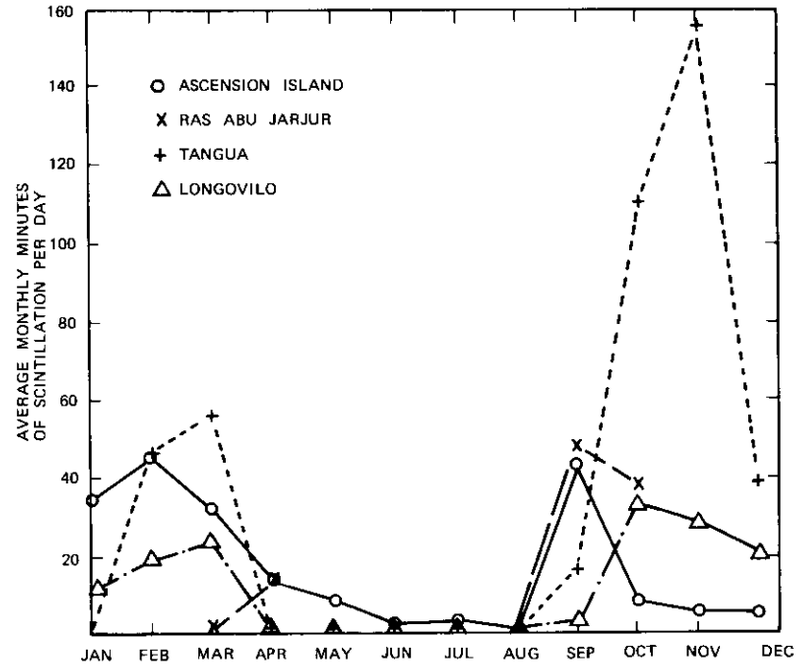


Figure 9. *Seasonal Variation of Ionospheric Scintillation Greater than 0.5 dB at Ascension Island, Ras Abu Jarjur, Tangua, and Longovilo*

Diurnal variation

The percent of total scintillation for each hour (local time) at each earth station is shown in Figures 13 and 14. It can be seen that the scintillation activity peaks at around 2000 hours local time for every station. Since scintillation occurs mainly in the vernal and autumnal seasons, 2000 hours is approximately sunset in the F-region. As the sun goes down, the main ionization source disappears. The disappearance of the solar flux will make the ionosphere more suitable for the production of irregularities caused by the inhomogeneities. The recombination of ions and electrons causes the ion density to decrease, and because of the inhomogeneities, the ionosphere becomes "patchy." A crude analogy is the breaking of the ice on a lake as the ambient temperature rises. Because of the existing inhomogeneities the ice layer will melt into large pieces at first, and then the patches will gradually become smaller until they all disappear.

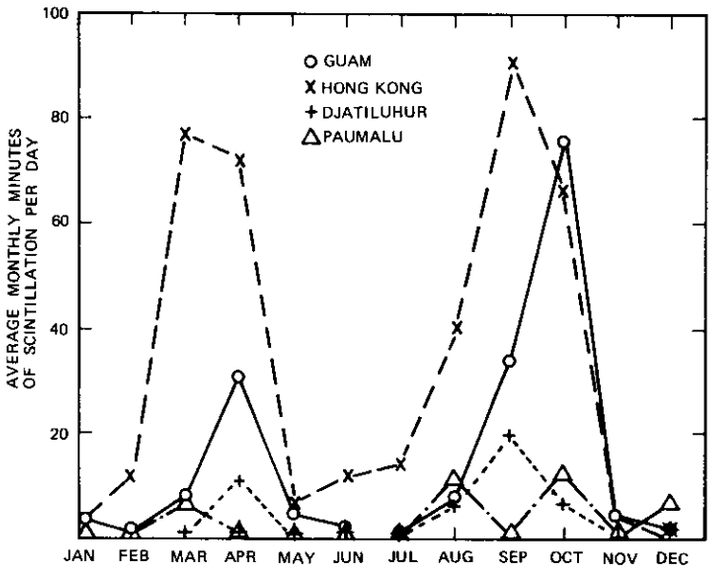


Figure 10. *Seasonal Variation of Ionospheric Scintillation Greater than 0.5 dB at Guam, Hong Kong, Djatiluhur, and Paumalu*

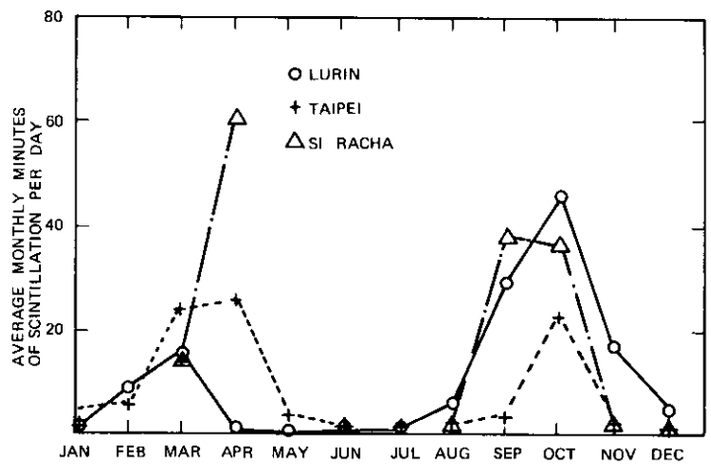


Figure 11. *Seasonal Variation of Ionospheric Scintillation Greater than 0.5 dB at Lurin, Taipei, and Si Racha*

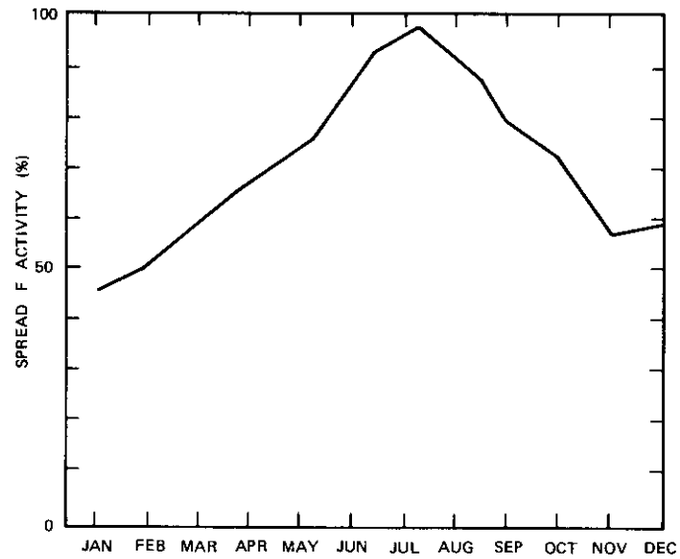


Figure 12. *Seasonal Variation of Spread F Observed at Djibouti* [4]

The data collected in the geomagnetic equatorial region indicate that Spread F is essentially a nighttime phenomenon that begins approximately one hour after local sunset. A typical diurnal variation of the percentage of occurrence of Spread F is reproduced from Reference 4 and shown in Figure 15. It can be seen that, as shown in the previous subsection, the diurnal variation of the scintillation is submerged in that of the Spread F. Therefore, it is hypothesized that the larger ion patches formed at the beginning of the evening are sufficiently dense and thick to cause the ionospheric scintillation. As these patches and the disturbed layer become smaller and thinner, their effect upon the radio wave at SHF will decrease until it finally becomes unnoticeable; yet the irregularities will still be dense enough to be detected by the ionospheric sounder.

According to Reference 2, the smallest value of ΔN of the irregularity which is detectable as Spread F is about 5×10^3 cm^{-3}. Based on the discussion in the "Theoretical Background" section of this paper, it will take a layer approximately 24,000 km thick to produce a scintillation with a peak-to-peak deviation of about 4 dB at 4 GHz. This clearly indicates that the

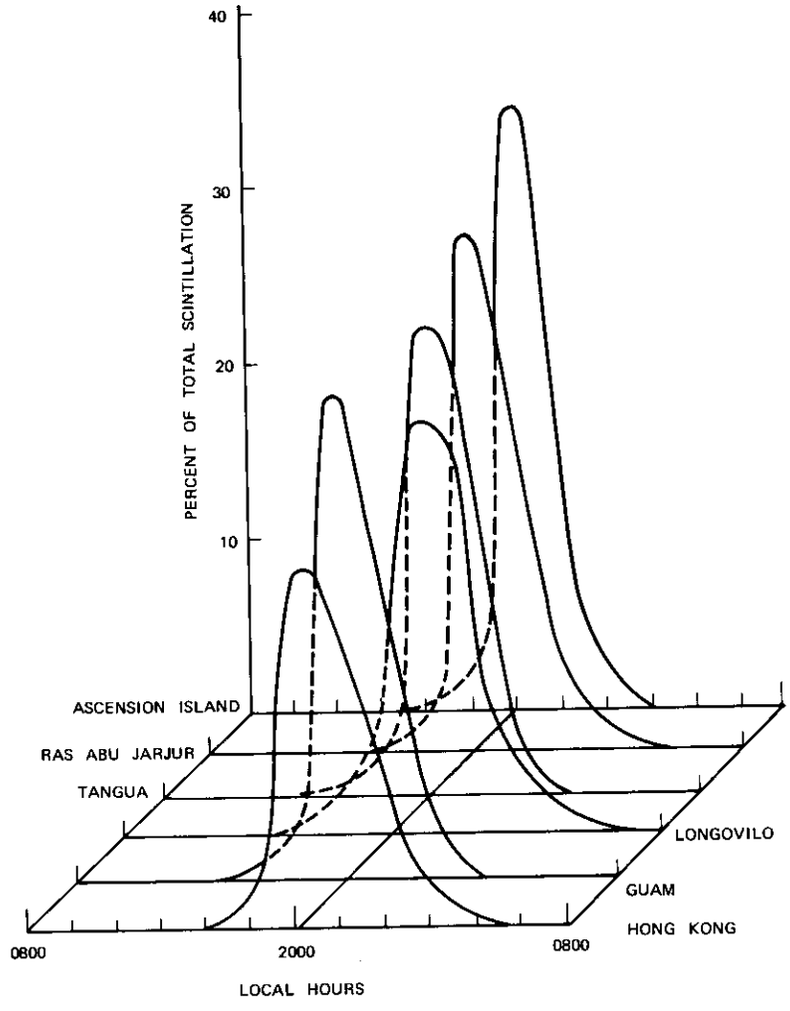


Figure 13. *Diurnal Variation of Ionospheric Scintillation Greater than 0.5 dB at Ascension Island, Ras Abu Jarjur, Tangua, Longovilo, Guam, and Hong Kong*

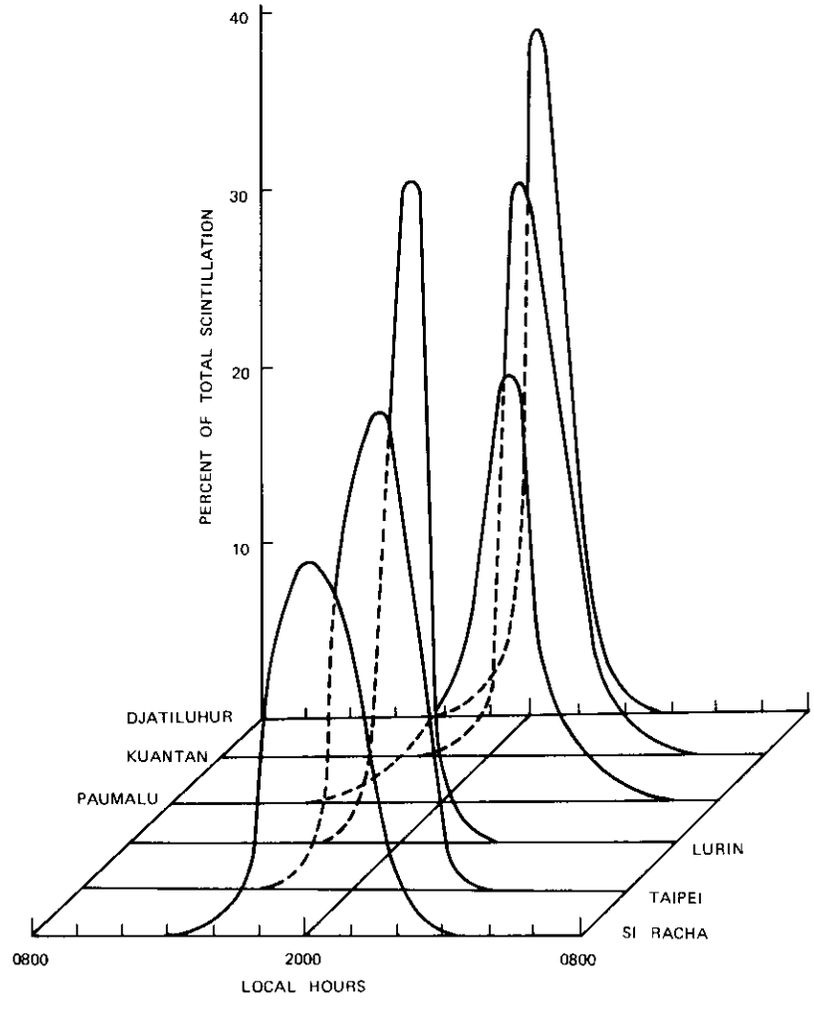


Figure 14. *Diurnal Variation of Ionospheric Scintillation Greater than 0.5 dB at Djatiluhur, Kuantan, Paumalu, Lurin, Taipei, and Si Racha*

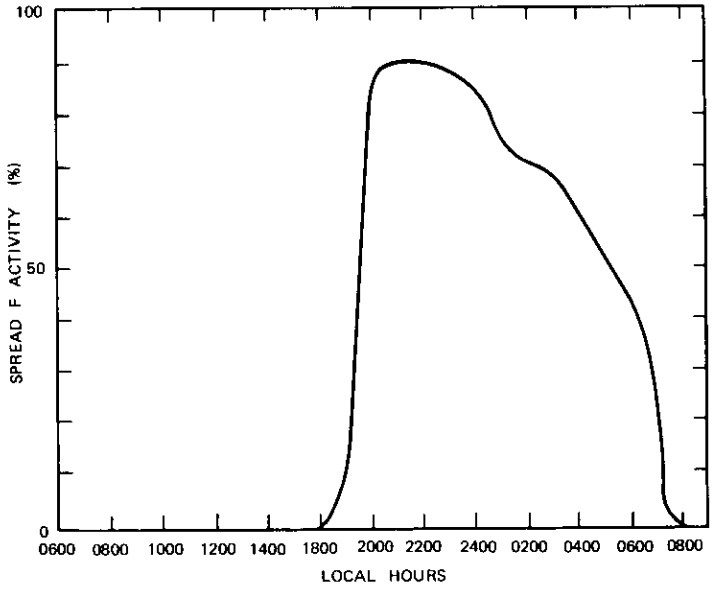


Figure 15. *Diurnal Variation of Spread F at Djibouti on Quiet Days with High Solar Activity During Equinox Months* [4]

thick, dense irregularities exist only about one or two hours after sunset in the F-region. The irregularities observed during the rest of the night are smaller patches which can be detected only by the ionospheric sounder or scintillation at a much lower frequency.

Geomagnetic activity dependence

A reasonably good negative correlation between magnetic activity and occurrence of scintillation has been given in Reference 1. This negative correlation again points toward the scintillation mechanism proposed here. As an extension of the analogy given in the previous subsection, if there are some disturbances in the lake, then the large ice patches will be broken into smaller pieces and dissolve much faster. Similarly, under the influence of magnetic disturbances, the ionosphere is "stirred," and therefore the recombination process is accelerated. Observation of Spread F also shows the negative correlation between magnetic activity and the occurrence of Spread F.

Geomagnetic latitudinal variation

The dependence of the occurrence of scintillation on geomagnetic latitude is shown in Figure 16. It can be seen that scintillation at 4 and

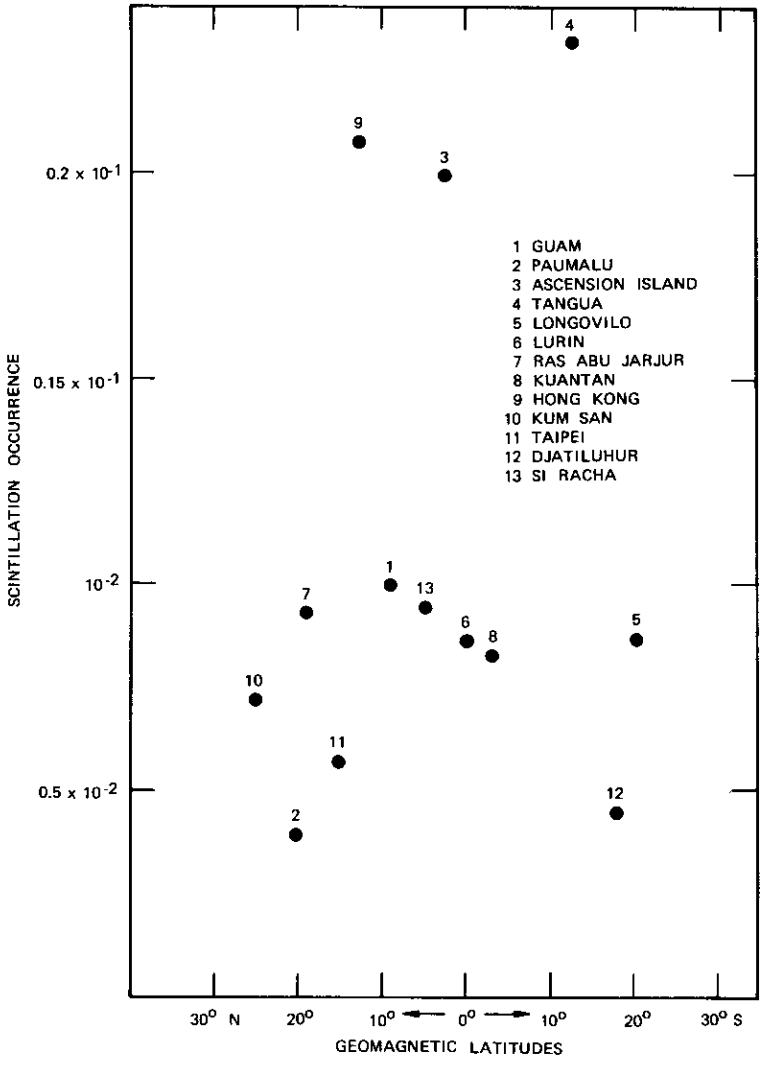


Figure 16. *Geomagnetic Latitudinal Variation*

6 GHz is limited to the region of $\pm 30°$ geomagnetic latitude. However, the dependence of the scintillation on geomagnetic latitude is not very clear, although scintillation appears to be more active near the geomagnetic equator.

Correlation between VHF and SHF scintillation

It has been attempted to extrapolate VHF scintillation data up to SHF by using λ^2 dependence. This leads to the prediction that there will be no scintillation at SHF, which contradicts the measurements. It should be pointed out that the irregularities that cause weak scattering at 4 and 6 GHz result in saturation of the signal at VHF. Moreover, they produce focusing effects upon the VHF. Therefore, from the VHF scintillation data, one can predict only the occurrence of scintillation at SHF, not its amplitude fluctuation.

Another possible explanation of the failure to predict SHF scintillation by using VHF data is that the spectrum of the irregularities follows a power law with an index near two instead of a Gaussian law, which is commonly used in the conventional diffraction model. The details of this theory are given in Reference 5 and will not be repeated here.

Conclusion

Based upon data acquired over a 15-month interval, probability distributions of the amplitude fluctuation of ionospheric scintillation have been presented. The scintillation was found mainly in the region with geomagnetic latitudes between 30°N and 30°S and was observed more often at the stations near the geomagnetic equator.

It has also been shown that the ionospheric scintillation at SHF has strong diurnal and seasonal variations. It is hypothesized that scintillation is caused by weak scattering from rather dense ionospheric irregularities. Because of the strong diurnal peak occurring about an hour after local sunset, it appears that these irregularities may be produced by the rapid decrease of the solar flux and the ensuing inhomogeneity of the ionosphere.

Acknowledgments

The author wishes to thank the engineering staff of the COMSAT Operations Division and the personnel at the earth stations who helped to collect the data. Dr. R. Barthle, J. Kaiser, J. L. Levatich, and Dr. E. S. Rittner of COMSAT and Dr. J. Aarons of AFCRL are acknowledged for many helpful discussions and criticism of the manuscript. The author is indebted to Dr. H. D. Craft, Jr. (now at Cornell University), who initiated the study.

References

[1] H. D. Craft, Jr., and L. H. Westerlund, "Scintillation at 4 and 6 GHz Caused by the Ionosphere," AIAA 10th Aerospace Sciences Meeting, San Diego, California, January 17–19, 1972.

[2] D. G. Singleton, "Saturation and Focusing Effects in Radio-Star and Satellite Scintillations," *Journal of Atmospheric and Terrestrial Physics*, Vol. 32, No. 2, February 1970, pp. 187–208.

[3] R. V. E. Lovelace, "Theory and Analysis of Interplanetary Scintillations," Ph.D. Thesis, Cornell University, Ithaca, New York, September 1970.

[4] P. Halley and B. Gatty, "Observation du F Diffus Equatorial a Djibouti et a Dakar," *Spread F and Its Effects Upon Radiowave Propagation and Communications*, edited by P. Newman, Maidenhead, England: Technivision, 1966, Chapter I-3.

[5] C. L. Rufenach, "A Power-Law Wavenumber Spectrum Deduced from Ionospheric Scintillation Observations," *Journal of Geo-Physical Research*, Vol. 77, No. 25, September 1, 1972, pp. 4761–4772.

Roger R. Taur was born in China and received his B.S. degree in electrical engineering from Cheng-Kung University, Tainan, Taiwan, and his M.S. and Ph.D. degrees in electrical engineering from Utah State University, Logan, Utah, where he was a Research Assistant in the Antenna and Propagation Laboratory of the Department of Electrical Engineering (1966–1970).

Dr. Taur worked in the area of antennas and propagation with Computer Sciences Corporation, Falls Church, Virginia, from 1970 to 1971 when he joined COMSAT Laboratories as a Member of the Technical Staff. He is currently engaged in research in the field of UHF-SHF wave propagation, ionospheric and tropospheric scintillation, and cross-polarization effects of propagation anomalies. Dr. Taur is a member of Sigma Xi and IEEE.

Index: communications satellites, earth terminals, multichannel communications, failure, monitors.

Monitoring interruptions at the satellite earth station

G. G. Szarvas and R. C. Trushel

Abstract

A monitoring system was designed and built to monitor and automatically record interruptions at the earth stations of a satellite communications system. This monitoring system uses a small computer with interfaces to the monitor points, the station clock, and a recorder. Hundreds of points can be monitored with periodic scanning. The resolution in the measurement of the interruption duration is of the order of 10 ms. The recorded data are postprocessed and listed after an off-line transmission to the computer center. In the field trial at the earth station at Andover, Maine, 50 monitor points, primarily carrier, supergroup, and group pilots, were connected to the system. According to the relationship of the pilots, the interruptions can be grouped as follows: satellite communications, domestic terrestrial, and distant terrestrial interruptions.

Introduction

The development of digital techniques and the use of telephone circuits for data transmission focus attention on such transmission characteristics as short interruptions and impulse noise. Although interruptions up to 200–500 ms are not objectionable in speech communications (call in progress), even millisecond interruptions can cause serious errors in data transmitted over voice circuits. Table 1 shows the effect of interruptions on various services [1].

The purpose of studies in this field is roughly twofold: to obtain an appropriate model of the channel [2] [4], and to investigate the causes of interruptions and impulse noise in transmission. The maintenance aspects have been investigated by Study Group IV of the C.C.I.T.T. [5].

Antenna tracking errors, switching of low-noise receivers and high-power amplifiers, and malfunctions in the earth station multiplexing and demultiplexing equipment are some of the factors which cause interruptions in satellite communications channels [6]. The interruption monitor system was primarily designed and built to measure interruptions at the earth station.

This presentation describes the data acquisition system, the processing of interruptions, and the system performance. Monitor point selection for the field trial at the earth station is also discussed.

System description

At this time, interruption recording at the earth stations is generally manual with a time resolution of roughly one second, based on station alarms. As a step forward, the following objectives for the interruption monitor system should be implemented:

a. The resolution time for interruption measurements should be of the order of 10 milliseconds.

b. The system should be able to monitor several hundred points.

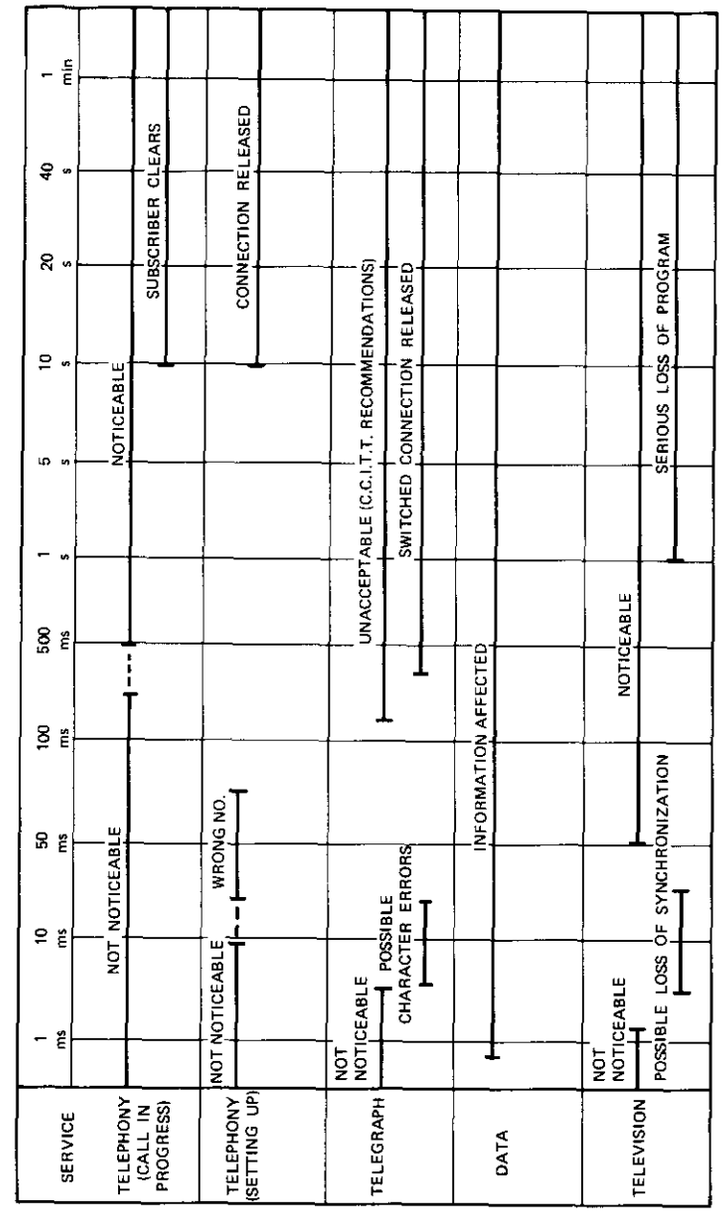
c. The system should be flexible with regard to configuration and time resolution.

d. The following data must be recorded when an interruption occurs: identification of the monitor point, duration (ΔT) of the interruption, and real-time marking of the interruption.

Figure 1 is a block diagram of the interruption monitor system, transmission, and postprocessing. Each monitor point in the earth station is connected to a threshold detector, which is generally a signal amplifier, with or without special filter characteristics, followed by an amplitude detector and a threshold comparator. The output of the threshold comparator is binary: a logic 1 corresponds to the interrupted state, and a logic 0 to the normal state.

The status of monitor points in groups of eight is periodically checked by the computer through the byte selector. The selection capacity of one byte selector unit is 32 bytes (256 monitor points). The system can be expanded without modification by installing a second unit.

TABLE 1. EFFECT OF INTERRUPTIONS ON VARIOUS SERVICES



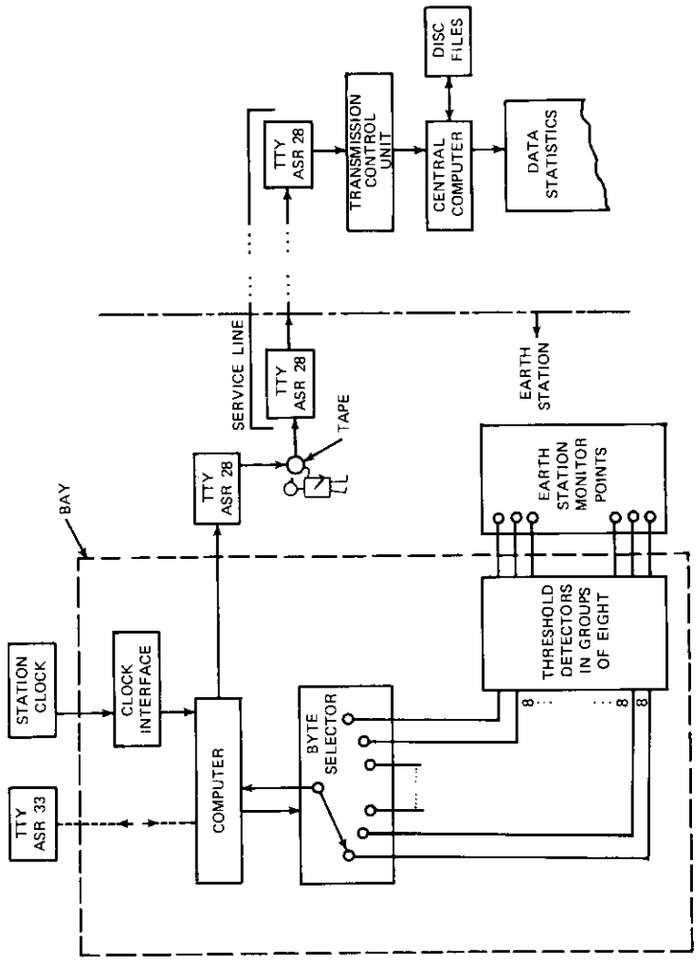


Figure 1. *Interruption Monitor System, Transmission, and Postprocessing*

The clock units for real-time counting and interruption duration measurements are transferred to the computer by the station clock through the clock interface. The real-time resolution is 0.1 s. In ΔT counting, the resolution can be 6.4, 12.8, or 25.6 ms. A second function of the 6.4-, 12.8-, or 25.6-ms pulse is to initiate a scanning order for the computer to check the status of the monitor points.

Interruption messages are recorded on the ASR 28 teletype, which acts only as a paper tape punch. The ASR 33 teletype is used for entering the program into the computer, for real-time initialization before an observation period, and for starting and stopping the observation period.

The computer is an 8-bit machine with 4,096 bytes of memory. The execution time for a computer instruction ranges from 1.5 to 12 μs. The computer has 11 interrupt lines. Two of these lines are used for station clock interfaces (scan interrupt and 0.1-s interrupt), and another is used for an ASR 28 teletype interface.

Processing interruptions

The interruption monitor program is driven by the three interrupts which were described previously. When none of these interrupts is active and all processing has been completed, the program is in a wait routine and the idle time is counted in seconds. Figure 2 is a block diagram describing interrupt processing.

Every tenth second, a pulse is sent to the computer to keep track of the real time. At every scan interrupt, the software ΔT counter is advanced and all of the monitor points are scanned. The present binary status of each monitor point is stored in a present status buffer, which is divided into two areas. If the status changes of the monitor points are completely processed before the next scan interrupt, the status of the monitor points on the next scan interrupt is stored in the same buffer area. If the status changes of the monitor points in one buffer area are not completely processed before the next scan interrupt, the second buffer area is used to store the status of the monitor points of all subsequent scan interrupts until the processing in the first buffer area is complete.

The method of processing the two buffer areas defines the waiting-line discipline for interruptions. When the processing time for the first buffer area is longer than two scan interrupt units and the status of the monitor points stored in the second buffer area changes, information errors occur. This type of error takes the form of a longer or shorter measured interruption, depending on whether the monitor point has returned to

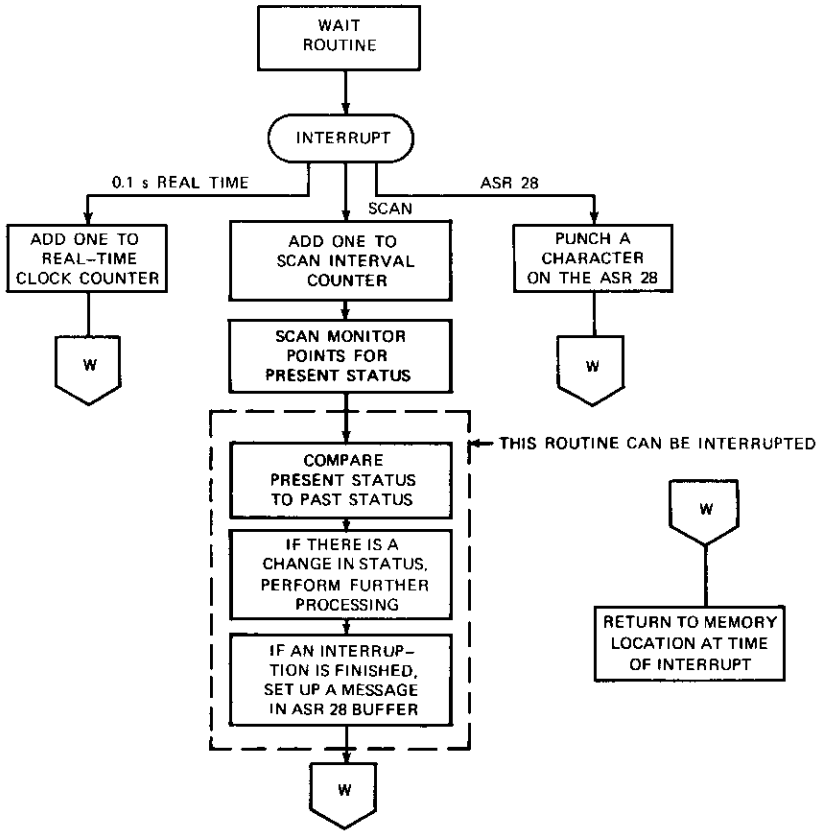


Figure 2. *Block Diagram Description of Interruption Processing*

service or gone out of service, respectively. When the duration of an interruption is less than the processing time for the first buffer area, the information is lost.

Interruptions are detected by comparing the present status of each monitor point with its past status (its status during the last scan). When a change in the status of a monitor point is detected, the program determines whether the monitor point has just gone out of service or has just come back into service. In the former case, the current value of the ΔT counter is stored in the interruption start time table. In the latter case, the stored ΔT counter value is subtracted from the current value of the ΔT counter to

calculate the duration of the interruption, and an interruption message is set up in the ASR 28 punch buffer.

When the punch buffer (with a storage capacity of 40 messages) is filled, another interruption message may not be added until 1.9 seconds, the time required to punch one entire interruption message, have elapsed. When this occurs, a software counter is advanced to record the number of lost messages.

At the end of an observation period, the program prints out any remaining interruption messages in the punch buffer and a trailer message, which contains the number of missed messages and the idle time of the computer. The punched tape is transmitted off-line to the COMSAT computer center for postprocessing.

The postprocessing programs translate the interruption monitor data into various readable and understandable formats. The first program is a chronological listing of all interruptions (Figure 3). The second is a chronological listing of interruptions per monitor point, and the third is a listing of the interruption duration distribution for each monitor point.

System performance

Scanning to measure the time duration of an interruption is an analog-to-digital conversion. In this type of uniform quantization, the quantizing width is the scanning time interval, T. The only difference between scanning as quantizing and normal quantizing is that, for the former, there is no fixed "zero level." That is, an interruption starts randomly in reference to the quantizing time moment. This increases the "quantizing noise." The time duration of an interruption measured to kT is in the interval $kT - \epsilon$ to $kT + \epsilon$ ($0 < \epsilon < T$, $k = 1, 2, 3, \ldots$). The error probability of a single measurement and the transformation of the continuous distribution function of interruptions into a discrete probability distribution are described in the appendices.

As a serving station, the system gives priority to the housekeeping activities (updating the counters for time keeping and scanning). If the 0.1-s interrupt is neglected, the arrival rate of housekeeping activities is identical to the selected scanning frequency, and the serving time is proportional to the number of monitor points. The interruption processing required by a change in the status of monitor points has a waiting-line discipline described in conjunction with the status buffer and is executed in the remaining time.

The portion of the total time allotted to housekeeping and interruption

TIME				MONITOR	DURATION		DESCRIPTION	
HR	min	s	s/10	POINT		FREQ	EQUIPMENT	PATH
14	43	50	3	026	5.5 min		RX SG PILOT INTRPT	
14	44	49	2	037	19.2 ms		RX GR PILOT INTRPT	
14	50	39	3	026	1.0 s		RX SG PILOT INTRPT	
14	50	40	9	026	1.6 s		RX SG PILOT INTRPT	
14	50	42	8	026	1.9 s		RX SG PILOT INTRPT	
14	51	36	9	026	54.2 s		RX SG PILOT INTRPT	
14	56	33	2	026	4.9 min		RX SG PILOT INTRPT	
14	57	22	5	054	6.4 ms		TX GR PILOT INTRPT	
14	58	4	7	049	6.4 ms		TX GR PILOT INTRPT	
14	58	4	8	049	6.4 ms		TX GR PILOT INTRPT	
14	59	52	6	046	6.4 ms		RX SG PILOT INTRPT	
14	59	52	7	046	57.6 ms		RX SG PILOT INTRPT	
14	59	52	8	046	32.0 ms		RX SG PILOT INTRPT	
14	59	52	9	041	6.4 ms		RX GR PILOT INTRPT	
14	59	52	9	043	6.4 ms		RX GR PILOT INTRPT	
14	59	52	9	044	6.4 ms		RX GR PILOT INTRPT	
14	59	52	9	045	6.4 ms		RX GR PILOT INTRPT	
15	0	4	7	080	6.4 ms		27.8-min INTRPT GEN	
15	2	34	4	026	55.3 s		RX SG PILOT INTRPT	
15	3	20	0	055	44.8 ms		TX GR PILOT INTRPT	
15	3	20	0	062	51.2 ms		TX GR PILOT INTRPT	
15	3	20	2	052	153.6 ms		TX GR PILOT INTRPT	
15	3	20	2	054	153.6 ms		TX GR PILOT INTRPT	
15	3	20	2	062	153.6 ms		TX GR PILOT INTRPT	
15	3	20	2	055	160.0 ms		TX GR PILOT INTRPT	
15	3	54	2	012	6.4 ms		TX 60-kHz CXR PILOT INTRPT	
15	3	54	9	026	1.3 min		RX SG PILOT INTRPT	
15	3	55	2	026	243.2 ms		RX SG PILOT INTRPT	

Figure 3. *Chronological Listing of Interruptions*

processing is shown as a function of the number of monitor points in Figure 4. In the same figure, the upper bound of performance is shown for working conditions A and B. For condition A, it is assumed that there is a recorder which can keep up with the processing rate and that the interruptions are uniformly distributed in time. The assumption of uniform distribution can be replaced with the assumption of an increased number of buffer areas to preserve the present status of the monitor points. The interruption processing time includes the transfer processing of messages to the recorder. Scale A shows the interruption processing capability per second.

For working condition B, the processing capability is shown as the time (in seconds) required to process 40 interruptions and to fill the output buffer of the ASR 28 teletype with interruption messages. The time required to record 40 interruption messages on the teletype is 76 seconds, which is a limit for the repetition rate of the bursts. The printing time of the teletype limits the long-term average of the system to 1890 interruptions per hour.

The following example may help to explain the use of Figure 4. For a load of 288 monitor points and a scan time interval of 6.4 ms, the system

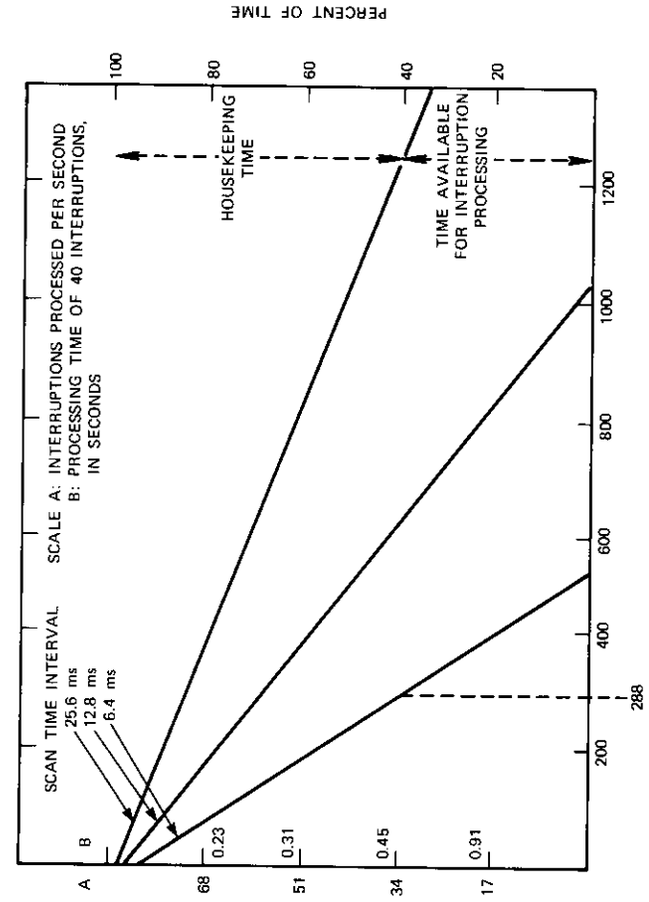


Figure 4. *Performance of the System*

processes 34 interruptions per second in working condition A and a burst-type load of 40 interruptions in 0.45 second in condition B.

Field trial

For a period of several weeks in the summer of 1971, the interruption monitor system was connected to 50 monitor points of the earth station at Andover, Maine. The objectives of the field trial were as follows:

a. to check system performance in field operation,
b. to compare the automatic recorded data with the manually recorded data of the earth station trouble register, and
c. to monitor interruptions down to the 10-ms range, primarily in the satellite communications network and then in the terrestrial network.

The majority of the signals used for monitoring interruptions were carrier, supergroup, and group pilots of the multiplexer/demultiplexer part of the earth station. Pilot access was simple and, consistent with current procedures, their alarms were also used for manual recording. Other signals connected to the system included 110 VAC, the technical power of the earth station, and an interruption generator signal with a time period of one-half hour for operational control of the system.

Andover is connected on the domestic side to the overseas switching centers of New York, White Plains, and Pittsburgh. The pilots of its through groups and supergroups are monitored in the multiplexer/demultiplexer unit. In the transmit side, group and supergroup pilot interruptions indicate domestic terrestrial interruptions. In the receive side, as the pilots are passed through the distant earth station, their interruptions indicate distant terrestrial network interruptions, excluding those caused by the carrier. The interruptions in the carrier pilots injected in the distant earth station indicate interruptions in the receive satellite communication. Figure 5 shows the topology of the pilots for evaluating the data.

The threshold level of the detectors was set at −6 dB below the normal signal level, a definition of interruption. (The alarm levels used for manual recording were closer to the normal level.) The bandwidth of the pilot filters gave an equivalent time constant of about 10 ms. The period of time for scanning was set at 6.4 ms, giving the same resolution in the time duration of interruptions.

The hierarchical relationship between the pilots in the demultiplexer part or in similar configurations can be utilized in a program development

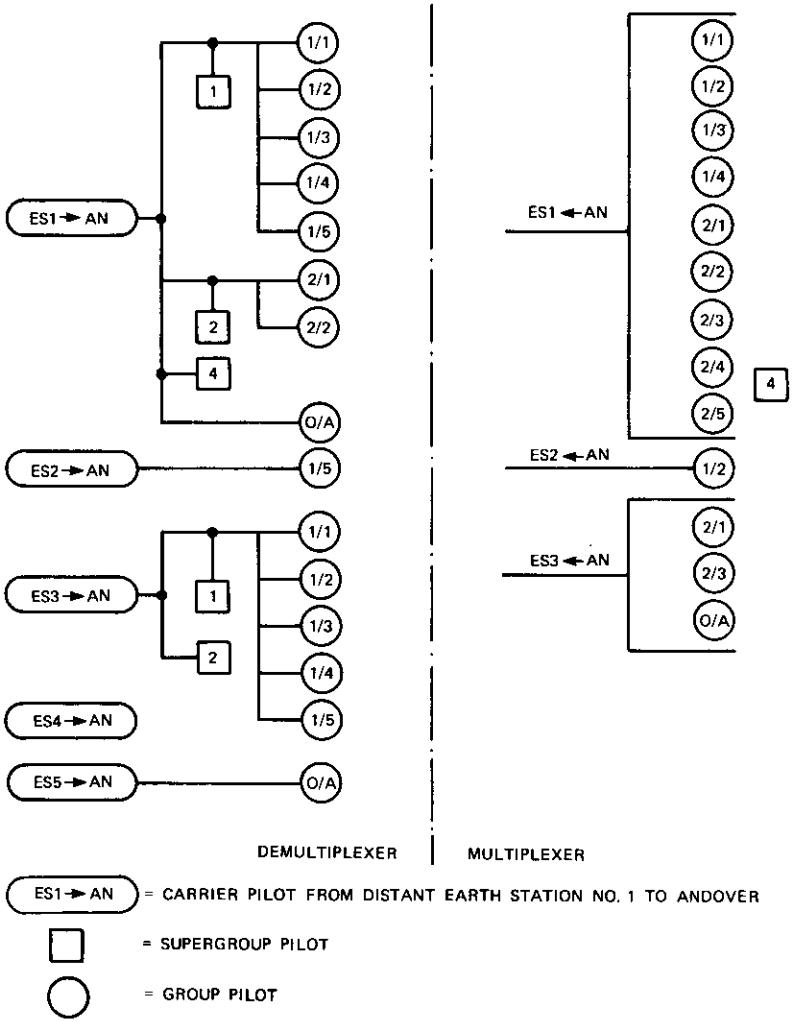


Figure 5. *Topology of the Pilots Used for Evaluation of Data*

either to diminish the processing work or for a redundant check. A receive carrier interruption causes an interruption, e.g., in all associated supergroups and groups. Therefore, processing of supergroup and group pilot interruptions can be omitted.

The use of threshold extension demodulators and/or automatic gain regulators causes the noise level to increase during the absence of signals.

When the signal is absent, noise peaks above the threshold level divide one continuous true interruption into many apparent short interruptions.

In the verification measurements for noise interaction, a looped-back contingency carrier was interrupted, and integrating networks with variable time constants were used in the amplitude detecting threshold circuits to filter the noise. In data postprocessing, the necessary minimum integrating time constant, 0.1 second for the carrier pilot and one second for the other pilots, was a criterion for joining the apparent distinct interruptions. This method makes it possible to avoid using more sophisticated signal detectors, e.g., correlation receivers, by accepting the possibility that the resolution of real interruptions occurring within the integration time will be lost. A software algorithm is being developed to incorporate both noise filtering and hierarchical processing in the interruption monitor program.

It is felt that a field trial of several weeks at a single earth station is not adequate for presenting numerical data. However, the following can be stated:

a. The interruption monitor system worked reliably in the field trial after some initial difficulties in transmitting and postprocessing data.

b. According to the hierarchical relationship between monitor points, postprocessing can separate the recorded data into satellite communications, domestic terrestrial, and distant terrestrial interruptions.

c. Interruptions down to the millisecond range were noted both in the satellite and in the terrestrial communications networks. These interruptions were recorded with a multipoint monitoring system.

d. The experimental cumulative distribution curve of the interruption duration can be approximated by a logarithmic normal distribution.

Conclusions

An interruption monitor of the type described in this paper can provide valuable information not otherwise available concerning interruption of the satellite communications system. For monitoring independent signals, the multiple-point monitoring system has the advantage of collecting significant statistical information in a shorter time. For monitoring interrelated signals, this same system permits fault isolation.

The expected flexibility of computer application for interruption measurements has been verified. Future earth stations in which a central computer will provide for monitor and control of multiple functions can be visualized.

Acknowledgments

The authors wish to thank Messrs. Helmo Raag, Thomas J. Celi, Ivor Knight, and Dr. Herbert H. Chu for their assistance, advice, and suggestions, and Mr. Donald Fifield and his staff for their assistance at the earth station. The system specification was formulated by a working group consisting of Messrs. A. C. Walle, Gomaa Abu-Taleb, and George G. Szarvas. The system hardware was implemented by Mr. Dennis Podgurski. The postprocessing programs were designed and implemented by Messrs. Sidney G. Embrey and Norman E. Schroeder and Mrs. Barbara M. Mosely.

References

[1] Ivor N. Knight, "The Effect of Short Interruptions on Satellite Services," *INTELSAT Earth Station Performance Seminar*, Washington, D.C., October 11–15, 1971, pp. 200–206.

[2] E. O. Elliott, "A Model of the Switched Telephone Network for Data Communications," *Bell System Technical Journal*, Vol. 44, No. 1, January 1965, pp. 89–109.

[3] W. K. Pehlert, Jr., "Performance of Error Control for High Speed Data Transmission on the Voice-Band Switched Telephone Network," *1969 IEEE International Conference on Communications*, Boulder, Colorado, June 1969, pp. 39.19–39.25.

[4] M. Muntner, "Error Statistics and Channel Performance," *Proceedings of the National Electronics Conference*, Vol. 25, Chicago, Illinois, December 1969, pp. 441–445.

[5] The International Telegraph and Telephone Consultative Committee (C.C.I.T.T.), "Short Breaks in Transmission," Question 2/IV, *White Book, IV Plenary Assembly, Mar del Plata*, September 23–October 25, 1968, pp. 9–29, Geneva; The International Telecommunications Union, 1969.

[6] "The Report of the Earth Station Performance Seminar," *INTELSAT Earth Station Performance Seminar*, Washington, D.C., October 11–15, 1971, pp. i–vi.

Appendix A. Error probability of a single measurement

An interruption with a time duration

$$\tau = kT - \epsilon, k = 1, 2, 3, \ldots, 0 < \epsilon < T$$

is shown in Figure A-1. The start of the interruption is observed at scanning time

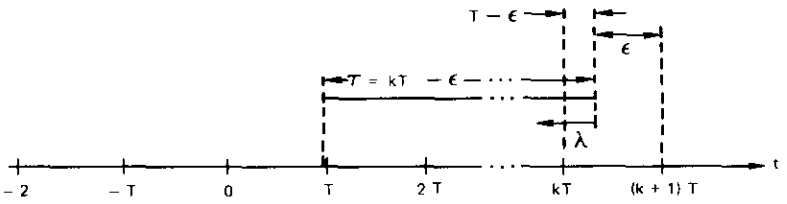


Figure A-1. *Random Position of* τ

T and its end is observed at $(k + 1)\ T$. Therefore, the duration of interruption is measured to $\Delta T = kT$.

Let λ be the length of a shift of τ. If τ is shifted to the left with $\lambda < (T - \epsilon)$, then ΔT is measured to kT. If τ is shifted further to the left with $(T - \epsilon) < \lambda < T$, then ΔT is measured to $(k - 1)\ T$. The shift corresponds to the random position of τ in relationship to the scanning time moments. In the same manner, an interruption of duration $\tau = kT + \epsilon$ can be measured to kT or $(k + 1)\ T$.

The random position of the start (or end) of the interruption τ over a scanning period time T corresponds to a uniform distribution with a constant probability density of $1/T$. Given the real interruption time, $\tau = kT \pm \epsilon$, the a priori probabilities of measuring an interruption to $(k \pm 1)\ T$ and kT are therefore

$$p_1 = p_{\Delta T|\tau}\left\{kT|(kT \pm \epsilon)\right\} = 1 - \frac{|\epsilon|}{T}$$

$$p_2 = p_{\Delta T|\tau}\left\{kT \pm 1|kT \pm \epsilon\right\} = \frac{|\epsilon|}{T} \qquad \text{(A1)}$$

respectively. These probabilities are shown in Figure A-2 as a function of ϵ.

The time duration of the interruption $\tau = kT \pm \epsilon$ is assumed to be uniformly distributed over the interval $(k - 1)\ T$ to $(k + 1)\ T$. Therefore,

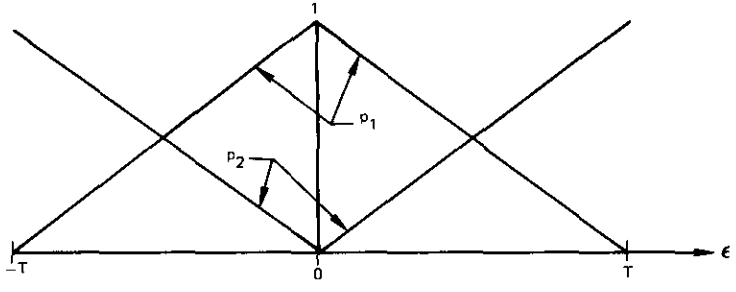


Figure A-2. *A Priori Probabilities*

$$p_\tau(kT \pm \epsilon) = \frac{1}{2T},\ (k - 1)\ T < \tau < (k + 1)\ T \qquad \text{(A2)}$$

and the probability

$$p_{\Delta T}(kT) = \int_{(k-1)\ T}^{(k+1)\ T} p_{\Delta T|\tau}\ p_\tau d\tau = \frac{1}{2} \quad . \qquad \text{(A3)}$$

According to the Bayes law and equations (A1), (A2), and (A3), the a posteriori probability can be written as follows:

$$p_{\tau|\Delta T}[(kT \pm \epsilon)|kT] = \frac{1}{T}\left(1 - \frac{|\epsilon|}{T}\right) \quad . \qquad \text{(A4)}$$

Equation (A4) is the probability density of ϵ, or the error in measurement. Therefore, the probability that the error of measurement is

$$|\epsilon| > \alpha, \quad 0 \leq \alpha \leq T$$

can be written as

$$p(|\epsilon| > \alpha) = 1 - 2\int_0^\alpha p_{\tau|\Delta T} d\epsilon = \left(1 - \frac{\alpha}{T}\right)^2 \quad . \qquad \text{(A5)}$$

This probability is plotted in Figure A-3, where, for example, the probability that the measurement error is larger than 0.6 T is 0.16.

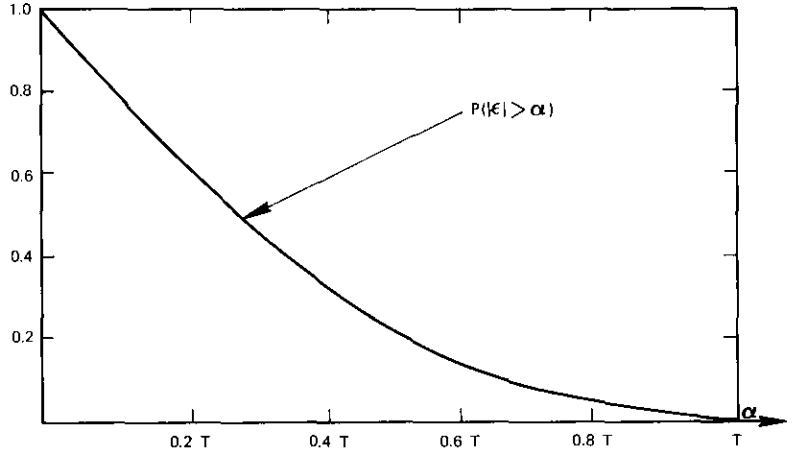


Figure A-3. *Error Probability of a Single Measurement*

Appendix B. Transformation of density function to discrete probabilities

In Figure B-1, $p_\tau(kT \pm \epsilon)$, $0 < \tau < \infty$, represents the density function of the time duration of an interruption. The measurement transforms this continuous function into discrete probabilities. Each elementary area $p_\tau d\tau$ contributes to two discrete probabilities. The discrete probability $p_{\Delta T}(kT)$ compresses the sampling area of p_τ from $(k - 1)\ T$ to $(k + 1)\ T$ to one line at kT. This area

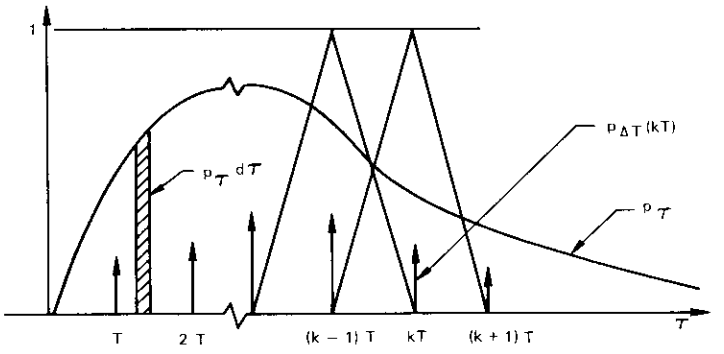


Figure B-1. *Transformation to Discrete Probabilities*

is weighted with an isosceles triangle with a height of 1 and a base length of $2T$. The probability $p_{\Delta T}$ can be expressed as

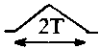
$$p_{\Delta T}(kT) = \int_{(k-1)\ T}^{kT} \left(\frac{\tau}{T} + 1 - k\right) p_\tau d\tau + \int_{kT}^{(k+1)\ T} \left(k + 1 - \frac{\tau}{T}\right) p_\tau d\tau \quad . \tag{B1}$$

Appendix C. Comparison of normal quantizing with scanning as quantizing

In normal quantizing [C 1], the given quantized signal amplitude is a deterministic rather than a statistical variable. This explains the differences in Table C–1, where scanning as quantizing is compared with normal quantizing, and T is the quantizing width.

To obtain more accurate results, the second- and higher-order moments of the discrete probabilities in normal quantizing are corrected by using Sheppard's corrections for grouping. These correction terms cannot generally be applied to scanning with the same accuracy because of its probabilistic type of grouping.

TABLE C–1. COMPARISON OF SCANNING AS QUANTIZING WITH NORMAL QUANTIZING

Property	Normal Quantizing	Scanning as Quantizing
Maximum Measurement Error	$\left\lvert\frac{T}{2}\right\rvert$	$\lvert T \rvert$
Weighting Function at Area Sampling		
Condition for Avoiding Aliasing	$\lvert\gamma\rvert < \frac{2}{T}$	$\lvert\gamma\rvert \sim \frac{2}{T}$
Accuracy in Recording Statistics		equal to or worse than that achieved with normal quantizing

*γ is the highest "frequency" in the frequency domain Fourier transform of the density function.

Reference

[C 1] A. K. Susskind, *Notes on Analog-Digital Conversion Techniques*, Cambridge, Massachusetts: M.I.T. Press, 1957.

George G. Szarvas was born in Tovaros, Hungary. He received the Dipl. Eng. degree with majors in electrical and mechanical engineering from the Technical University in Budapest (1950). From 1950 to 1956 he was a group leader working on the development of microwave receivers at the Telecommunications Research Institute in Budapest. From 1956 to 1963 he was employed by the Swedish companies L. M. Ericsson and Facit Electronics, where he worked with a time-division multiplex electronic exchange and with a computer for real-time application. From 1963 to 1969 he was manager of the basic technology for electronic circuits and instrumentation at the Swedish Teleadministration. He joined COMSAT Laboratories in 1969 and is a Member of the Technical Staff in the Digital Control Department.

Richard C. Trushel received his B.S. degree from St. Louis University. After receiving his M.S. degree in mathematics from the University of Iowa, he joined COMSAT Laboratories as a scientific computer applications programmer in 1969. For the past three years he has done real-time programming for mini-computer systems.

Index: telecommunications, modulation, phase-shift keying, random noise, electromagnetic interference, probable error.

Effects of cochannel interference and Gaussian noise in M-ary PSK systems

O. SHIMBO AND R. FANG

Abstract

The power series expansion technique can be used to develop an effective computational procedure for analyzing the combined effects of Gaussian noise and cochannel interference in M-ary coherent PSK systems. This procedure can be easily implemented on the computer. As numerical examples, the error probabilities of 4-, 8-, and 16-phase PSK systems as a result of impairments caused by Gaussian noise and one, two, three, and four equal-strength interferences are evaluated. Tradeoffs in system design between carrier-to-noise ratio and carrier-to-interference ratio can thus be made to achieve a given error probability performance for a given number of interferences.

Introduction

In some satellite communications environments, the effect of cochannel interference in addition to that of Gaussian noise can be quite significant.

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.

For instance, in the geostationary orbit, the effect of cochannel interference has been an important factor in efficient utilization of the available "parking space" in the "parking window." Also, for the case in which an earth station of a satellite system operates near some terrestrial communications facilities that share the same frequency spectrum, the resulting performance degradation in either facility must be accurately estimated. In this paper, the combined effects of Gaussian noise and cochannel interference on the error probability performance of M-ary coherent PSK systems are to be evaluated.

This problem has been studied previously by many investigators [1]–[7]. However, all of them either attack a less general problem or are unable to provide accurate numerical results for the general M-ary (e.g., 8- and 16-phase) PSK systems. Modification of the results in Reference 8 indicates that the combined effects of Gaussian noise and cochannel interference on the probability of error of M-ary PSK systems can be obtained rather easily.

First, the necessary modifications to the analyses in Reference 8 are adapted to solve the problem presented here. Then a modified computational procedure, which is implemented on COMSAT's computer, is described. Typical results are presented in figures which show the combined effects of Gaussian noise and one, two, three, and four equal-strength cochannel interference entries on the error probabilities of 4-, 8-, and 16-phase systems.*

Analysis

It is assumed that the receiver is an ideal PSK receiver as shown in Figure 1. It is further assumed that the filter in Figure 1 does not distort either the main signal or the cochannel interference. The case in which the filter produces intersymbol interference on the desired signal is outside the scope of the present paper and will be reported elsewhere.

The output signal of the filter in Figure 1 can be represented by

$$R(t) = A \sin(\omega_c t + \theta) + N_c(t) \cos \omega_c t + N_s(t) \sin \omega_c t + \sum_{l=1}^{H} B_l \sin(\omega_l t + \phi_l + \lambda_l) \quad (1)$$

* For convenience, the cochannel interferences are assumed to have equal strength in all illustrations, but this assumption is by no means necessary, as can be seen from the analysis in the next section.

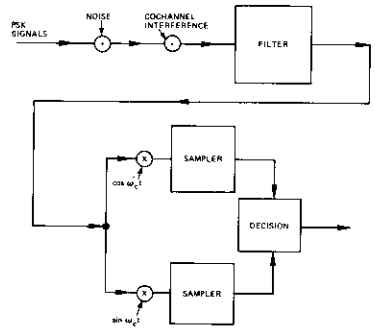


Figure 1. *Receiving Scheme of PSK Systems*

where A = amplitude of the desired signal
ω_c = angular carrier frequency of the desired signal
θ = modulating angle of the desired signal
N_c = in-phase component of noise
N_s = quadrature component of noise
B_l = amplitude of the lth carrier
ω_l = angular frequency of the lth carrier
ϕ_l = modulating angle of the lth carrier
λ_l = random phase of the lth carrier
H = number of cochannel interferences.

If it is assumed that θ is equally likely to be any of the M phases, the probability with which the received signal point P falls outside the decision cone D in Figure 2 represents the desired error probability. According to

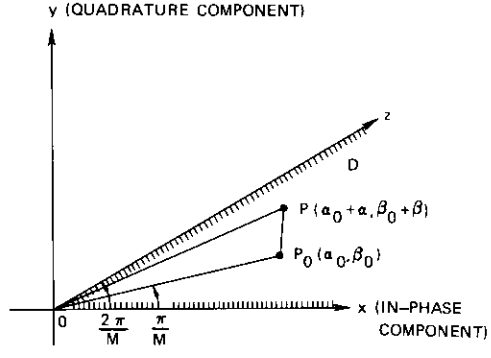


Figure 2. *Signal Space of PSK Systems*

the analysis given in Reference 8, the probability with which P falls inside D is

$$P_c = (2\pi\sigma_n^2)^{-1} \underset{\alpha,\beta}{E}\left[\int_0^\infty \int_0^{\Gamma x} \exp\left\{-\frac{1}{2\sigma_n^2}[(x-\alpha_0-\alpha)^2 + (y-\beta_0-\beta)^2]\right\} dx\, dy\right] \tag{2}$$

where σ_n^2 = baseband power of $N_c(t)$ or $N_s(t)$

$$\alpha_0 = A\cos\left(\frac{\pi}{M}\right)$$

$$\beta_0 = A\sin\left(\frac{\pi}{M}\right)$$

$$\Gamma = \tan\left(\frac{2\pi}{M}\right)$$

$$\alpha = \sum_{l=1}^{H} B_l \cos[(\omega_l - \omega_c)\, t + \phi_l + \lambda_l]$$

$$\beta = \sum_{l=1}^{H} B_l \sin[(\omega_l - \omega_c)\, t + \phi_l + \lambda_l]$$

$\underset{\alpha,\beta}{E}$ = averaging with respect to α and β.

Rewriting equation (2) in terms of the characteristic function Φ_0 of the cochannel interference yields

$$P_c = (2\pi)^{-2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^\infty\int_0^{\Gamma x} e^{-j(x-\alpha_0)u - j(y-\beta_0)v} \cdot e^{-1/2\,\sigma_n^2(u^2+v^2)} \Phi_0(u, v)\, du\, dv\, dx\, dy \tag{3}$$

where

$$\Phi_0(u, v) \equiv \underset{\alpha,\beta}{E}\left\{e^{ju\alpha + jv\beta}\right\} \tag{4}$$

If it is assumed that the random phases of the cochannel interference ϕ_l are mutually independent and uniformly distributed in $(0,2\pi)$, then $\Phi_0(u,v)$ can be expressed as

$$\Phi_0(u, v) = \prod_{l=1}^{H} J_0\left(B_l\sqrt{u^2+v^2}\right) \tag{5}$$

Note that the form of equation (3) is identical to that of equation (16) in Reference 8 except that $\Phi_0(u,v)$ is now given by equation (5) instead of by equations (17) and (18) in Reference 8.

To achieve faster convergence in the evaluation of error probabilities via the power series expansion method in Reference 8, define

$$\sigma^2 = \sigma_n^2 + \frac{\Delta}{2}\sum_{l=1}^{H} B_l^2 \tag{6}$$

Hence, σ^2 can be interpreted as the sum of the RF noise power and a part of the RF cochannel interference power (with the fraction Δ). With this definition, equation (3) can be modified as follows:

$$P_c = (2\pi)^{-2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^\infty\int_0^{\Gamma x} e^{-1/2\,(u^2+v^2)} \cdot e^{-j(x-\alpha_1)u - j(y-\beta_1)v} \Phi_1(u, v) \cdot e^{\lambda^2/2\,(u^2+v^2)}\, du\, dv\, dx\, dy \tag{7}$$

where

$$\alpha_1 = \frac{\alpha_0}{\sigma} = \frac{A}{\sigma}\cos\left(\frac{\pi}{M}\right) = \frac{A\cos(\pi/M)}{\sqrt{\sigma_n^2 + \frac{\Delta}{2}\sum_{l=1}^{H} B_l^2}} \tag{8a}$$

$$\beta_1 = \frac{\beta_0}{\sigma} = \frac{A}{\sigma}\sin\left(\frac{\pi}{M}\right) = \frac{A\sin(\pi/M)}{\sqrt{\sigma_n^2 + \frac{\Delta}{2}\sum_{l=1}^{H} B_l^2}} \tag{8b}$$

$$\lambda^2 = \frac{\frac{\Delta}{2}\sum_{l=1}^{H} B_l^2}{\sigma^2} = \frac{\frac{\Delta}{2}\sum_{l=1}^{H} B_l^2}{\sigma_n^2 + \frac{\Delta}{2}\sum_{l=1}^{H} B_l^2} \tag{8c}$$

$$\phi_1(u, v) = \prod_{l=1}^{H} J_0\left(\frac{B_l}{\sigma}\sqrt{u^2+v^2}\right) \tag{9}$$

Now expand

$$\Phi_1(u, v)\, e^{\lambda^2/2(u^2+v^2)}$$

into a power series:

$$\phi_1(u, v)\, e^{(\lambda^2/2)(u^2+v^2)} = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} b_{2m,2n}\, u^{2m}\, v^{2n} \tag{10}$$

Define the Hermite functions $\phi_n(x)$ as

$$\phi_n(x) = (2\pi)^{-1/2} H_n(x) \exp\left(\frac{-x^2}{2}\right) \tag{11}$$

for $n = 0, 1, 2, \ldots$, where $H_n(x)$ is the Hermite polynomial of degree n. These Hermite functions admit the following recurrence relationships:

$$\phi_n'(x) = -\phi_{n+1}(x) \tag{12a}$$

$$\phi_{n+1}(x) = x\phi_n(x) - n\phi_{n-1}(x) \tag{12b}$$

where the prime denotes the differentiation with respect to x and

$$\phi_{-1}(x) \equiv (2\pi)^{-1/2} \int_x^\infty \exp\left(\frac{-t^2}{2}\right) dt \equiv \frac{1}{2} \operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) \tag{12c}$$

Hence, substitution of equations (10)–(12) into equation (7) yields

$$\begin{aligned} P_e = &\int_0^\infty \int_0^{\Gamma x} \phi_0(x - \alpha_1)\ \phi_0(y - \beta_1)\ dy\ dx \\ &+ \sum_{m,n}{}' (-1)^{m+n}\ b_{2m,2n}[\phi_{2m-1}(-\alpha_1)\ \phi_{2n-1}(-\beta_1) \\ &- I_{2m,2n-1}(\alpha_1, \beta_1)] \end{aligned} \tag{13}$$

where the prime after the Σ denotes the exclusion of the term with $m = n = 0$, where $b_{0,0} = 1$, as can easily be seen from equations (9) and (10), and where

$$I_{m,n}(\alpha_1, \beta_1) \equiv \int_0^\infty \phi_m(x - \alpha_1)\ \phi_n(\Gamma x - \beta_1)\ dx \tag{14}$$

The quantity in equation (14) can be evaluated by using the recurrence method. The recurrence relationships for computing $I_{m,n}(\alpha_1,\beta_1)$ have been derived in Reference 8 and are as follows:

$$I_{m,n}(\alpha_1, \beta_1) = \phi_{m-1}(-\alpha_1)\ \phi_l(-\beta_1) - \Gamma\ I_{m-1,n+1}(\alpha_1, \beta_1) \tag{15}$$

for $M > 4$, and

$$\begin{aligned} I_{0,n}(\alpha_1, \beta_1) = &\sum_{l=0}^{n} \binom{n}{l} \phi_l\left(\alpha_1 \sin\frac{2\pi}{M} - \beta_1 \cos\frac{2\pi}{M}\right)\left(\cos\frac{2\pi}{M}\right)^{l+1} \\ &\cdot \left[\delta_{n,l} - \phi_{n-l-1}\left(\alpha_1 \cos\frac{2\pi}{M} + \beta_1 \sin\frac{2\pi}{M}\right)\right. \\ &\left.\cdot \left(-\sin\frac{2\pi}{M}\right)^{n-l}\right] \end{aligned} \tag{16}$$

in which

$$\begin{aligned} \delta_{n,l} &= 1 \text{ if } n = l \\ &= 0 \text{ if } n \neq l \end{aligned}$$

If $M = 4$, then the term $I_{2m,2n-1}(\alpha_1,\beta_1)$ in equation (13) should be set equal to zero and computation of any $I_{m,n}(\alpha_1,\beta_1)$ is unnecessary.

Computation of the coefficients $b_{2m,2n}$ is slightly different from the computation in Reference 8. The new computational procedure will be provided in the next section so that the Σ' term in equation (13) can be evaluated. The first term in equation (13) can be computed by direct integration or by the following approximation derived in Reference 8:

$$\begin{aligned} &\int_0^\infty \int_0^{\Gamma x} \phi_0(x - \alpha_1)\ \phi_0(y - \beta_1)\ dy\ dx \\ &= \frac{1}{2}\operatorname{erfc}[2^{-1/2}\ C \cos\theta_2] + \frac{1}{2}\operatorname{erfc}[-2^{-1/2}\ C \cos\theta_1] \\ &+ \exp\left(-\frac{1}{2}C^2\right)\left\{(4\pi d)^{-1}\ C^2 \sin(\theta_1 - \theta_2)\cos(\theta_1 + \theta_2)\right. \\ &- (2Md)^{-1}\ C^2 + (4\pi d)^{-1}\ C\ [\cos\theta_2(2d + C^2\sin^2\theta_2)^{1/2} \\ &- \cos\theta_1(2d + C^2\sin^2\theta_1)^{1/2}] + (4\pi d)^{-1}\ (C^2 + 2d) \\ &\left.\cdot \left[\sin^{-1}\left\{\frac{C\cos\theta_2}{(C^2 + 2d)^{1/2}}\right\} - \sin^{-1}\left\{\frac{C\cos\theta_1}{(C^2 + 2d)^{1/2}}\right\}\right]\right\} \end{aligned} \tag{17}$$

where $2 \geq d \geq 4/\pi$, and

$$C \equiv (\alpha_1^2 + \beta_1^2)^{1/2} \qquad \theta_1 = \frac{\pi}{M} + \frac{2\pi}{M} - \theta_0$$

$$\theta_0 = \tan^{-1}\left(\frac{\beta_1}{\alpha_1}\right) \qquad \theta_2 = \frac{\pi}{2} - \theta_0 \quad .$$

Therefore, the error probability

$$P_e \equiv 1 - P_c \tag{18}$$

can be numerically computed from equations (18) and (13).

Recurrence relationship for $b_{2m,\,2n}$

Denote the total RF-interference-to-RF-noise power ratio as k, and define L as the ratio of the carrier power to the sum of the RF noise power and the total RF interference power. Namely,

$$k = \sum_{l=1}^{H} \frac{B_l^2}{2\sigma_n^2} \tag{19a}$$

$$L \equiv \frac{A^2}{2\left(\sigma_n^2 + \frac{1}{2}\sum_{l=1}^{H} B_l^2\right)} = \frac{C/N}{1+k} \quad . \tag{19b}$$

Then, equation (8) can be rewritten as

$$\alpha_1 = \left[\frac{2L(1+k)}{1+\Delta k}\right]^{1/2} \cos\left(\frac{\pi}{M}\right) \tag{20a}$$

$$\beta_1 = \left[\frac{2L(1+k)}{1+\Delta k}\right]^{1/2} \sin\left(\frac{\pi}{M}\right) \tag{20b}$$

$$\lambda^2 = \frac{\Delta k}{1+\Delta k} \quad . \tag{20c}$$

Assuming all interferences to be equal (otherwise only eqs. 27a and b would be modified),

$$B = B_1 = B_2 = \ldots = B_H \quad . \tag{21}$$

Thus, equations (9) and (10) indicate that

$$\Phi_1(u, v) \exp\left[\frac{1}{2}\lambda^2(u^2+v^2)\right]$$

$$= \left[J_0\left\{\left[\frac{2k}{H(1+\Delta k)}\right]^{1/2}[u^2+v^2]^{1/2}\right\}\right]^H \exp\left[\frac{1}{2}\lambda^2(u^2+v^2)\right]$$

$$= \sum_{m,n} b_{2m,2n}\, u^{2m}\, v^{2n} \quad . \tag{22}$$

Two methods can be employed to obtain $b_{2m,2n}$: the recurrence method and the convolution method.

Recurrence method

Define

$$\gamma = (u^2+v^2)^{1/2}$$

and

$$J_0\left\{\left[\frac{2k}{H(1+\Delta k)}\right]^{1/2}\gamma\right\} = \sum_{m=0}^{\infty} C_{2m}\,\gamma^{2m} \tag{23}$$

where $$C_{2m} = (-1)^m\,(2^m\,m!)^{-2}\left[\frac{2k}{H(1+\Delta k)}\right]^m \tag{24}$$

Then, equation (22) can be put into the following form:

$$\Phi_1(u, v) \exp\left[\frac{1}{2}\lambda^2(u^2+v^2)\right]$$

$$= \left[J_0\left\{\left[\frac{2k}{H(1+\Delta k)}\right]^{1/2}\gamma\right\}\right]^H \exp\left(\frac{1}{2}\lambda^2\gamma^2\right) \equiv \sum_{m=0}^{\infty} d_{2m}\,\gamma^{2m} \tag{25}$$

Since $\exp[(1/2)\,\lambda^2\gamma^2]$ can be expanded into the power series

$$\exp\left(\frac{1}{2}\lambda^2\gamma^2\right) = \sum_{l=0}^{\infty}\frac{1}{l!}\left(\frac{1}{2}\lambda^2\right)^l \gamma^{2l} \tag{26}$$

the coefficients d_{2m} can be obtained by comparing the coefficients of γ^{2m} on both sides of equation (25) after substituting equations (23) and (26) into equation (25). Hence, for $i \neq 0$,

$$d_{2i} = \frac{1}{2i}\sum_{p=0}^{i-1}[2H(i-p)\;C_{2i-2p} + \lambda^2 C_{2i-2p-2} - (2p)\;C_{2i-2p}]\;d_{2p} \tag{27a}$$

$$d_0 = 1 \quad . \tag{27b}$$

Now, equation (25) can be written as

$$\Phi_1(u, v) \exp\left[\frac{1}{2}\lambda^2(u^2 + v^2)\right] = \sum_{m=0}^{\infty} d_{2m}\,\gamma^{2m}$$

$$= \sum_{p=0}^{\infty} d_{2p}(u^2 + v^2)^p$$

$$= \sum_{m,n} b_{2m,2n}\,u^{2m}\,v^{2n} \qquad (28)$$

Using binominal expansion on $(u^2 + v^2)^p$ in equation (28) and comparing coefficients yields

$$b_{2m,2n} = d_{2(m+n)}\frac{(m+n)!}{m!n!} \qquad (29)$$

Convolution method

From equations (22), (23), and (25), it is obvious that d_{2m} can be computed by convolving C_{2m} H times and then convolving with the coefficients in equation (26). That is,

$$C_{2k}^{(1)} = \sum_{m=0}^{k} C_{2m}\,C_{2k-2m} \qquad (30a)$$

$$C_{2k}^{(2)} = \sum_{m=0}^{k} C_{2m}\,C_{2k-2m}^{(1)} \qquad (30b)$$

$$C_{2k}^{(H-1)} = \sum_{m=0}^{k} C_{2m}\,C_{2k-2m}^{(H-2)} \qquad (30c)$$

$$d_{2i} = \sum_{m=0}^{i} C_{2i-2m}^{(H-1)}\left(\frac{\lambda^2}{2}\right)^m \frac{1}{m!} \quad . \qquad (30d)$$

Computational procedure

The following is a modified computational procedure which is implemented on COMSAT's computer to determine the effects of Gaussian noise and up to four equal-strength cochannel interferences on the error probabilities of 4-, 8-, and 16-phase PSK systems:

a. Decide the values of M, H, k, L, and Δ. (To choose Δ, see Reference 8.)

b. Compute α_1, β_1, and λ^2.

c. Compute $1 - \int_0^{\infty}\int_0^{\Gamma x} \phi_0(x - \alpha_1)\,\phi_0(y - \beta_1)\,dy\,dx$ from equation (17).

d. Compute $b_{2m,2n}$ from equations (24), (27), and (29) or from equations (29) and (30).

e. Compute $\phi_n(-\alpha_1)$ and $\phi_n(-\beta_1)$ from the recurrence relationship in equation (12).

f. Compute $I_{m,n}(\alpha_1,\beta_1)$ from equation (15) if $M > 4$. (If $M = 4$, this computation is unnecessary.)

g. Compute $P_e = 1 - P_c$ from equation (13).

h. If the convergence in the computation of equation (13) is not fast enough, adjust the value of Δ to obtain a faster convergence (see Reference 8).

Numerical results and discussion

Figures 3–5 show the numerical results for one, two, three, and four cochannel interferences in 4-, 8-, and 16-phase PSK systems, respectively. In all figures, the vertical axes represent the error probabilities, the horizontal axes represent L in dB, and the parameters represent k in dB. Since k, as defined in equation (19a), is the total RF-interference-to-RF-noise power ratio, the two extrema with $k = +\infty$ dB and $-\infty$ dB correspond to noise-free and interference-free situations, respectively. Thus, the curves associated with $k = -\infty$ dB in all figures should be identical to those error probabilities resulting from Gaussian noise alone. Consequently, the curves with $k = -\infty$ dB in Figure 3 are identical, as are the curves with $k = -\infty$ in Figure 4 and Figure 5.

On the other hand, for the noise-free cases in which $k = +\infty$ dB, the error probabilities are always equal to zero for L above certain thresholds, since the amplitudes of these interferences are bounded. (When the carrier-to-total-interference power ratio exceeds a certain threshold, the received signal will always lie in the correct decision region and hence the error probability will be zero.) These thresholds are determined by the number of cochannel interferences for a given M-ary PSK system. For instance, it can be seen from Figure 3 that, for the 4-phase PSK system, the thresholds are 3, 6, 8, and 9 dB, respectively, as a result of one, two, three, and four cochannel interferences.

Also, from these sets of figures, it can be deduced that, for a fixed number of phases, M, a fixed interference-to-noise power ratio, k, and a fixed

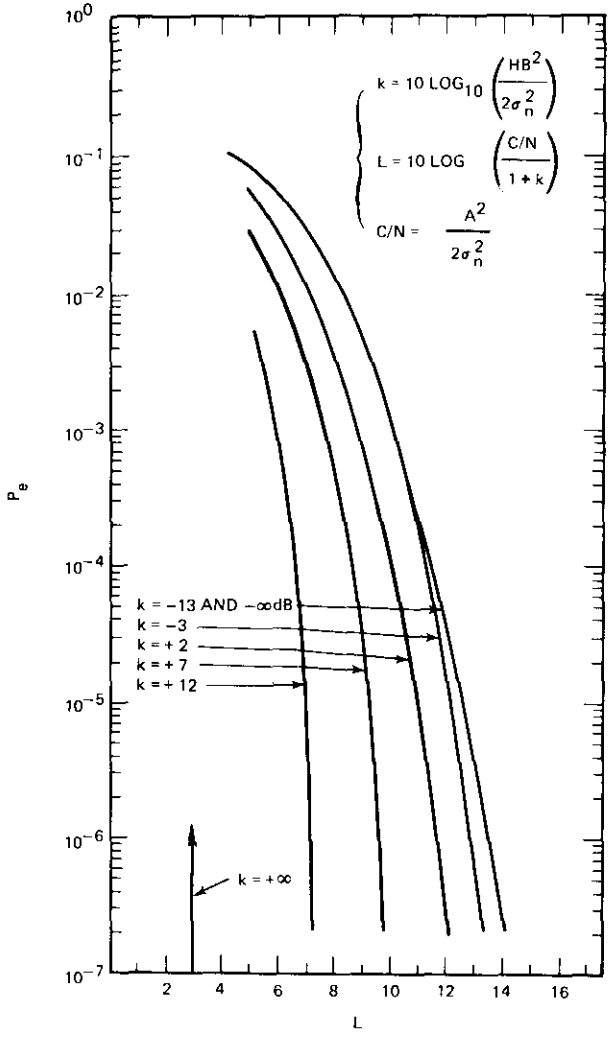


Figure 3a. *Error Probabilities of 4-Phase PSK in the Presence of Gaussian Noise and One Cochannel Interference ($H = 1, 4\phi$)*

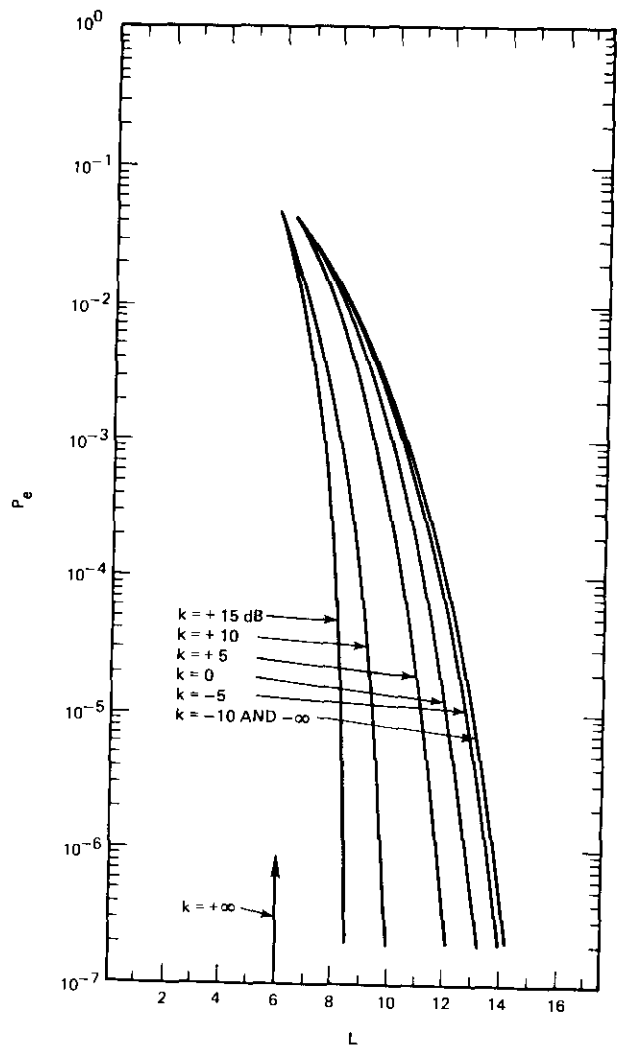


Figure 3b. *Error Probabilities of 4-Phase PSK in the Presence of Gaussian Noise and Two Cochannel Interferences ($H = 2, 4\phi$)*

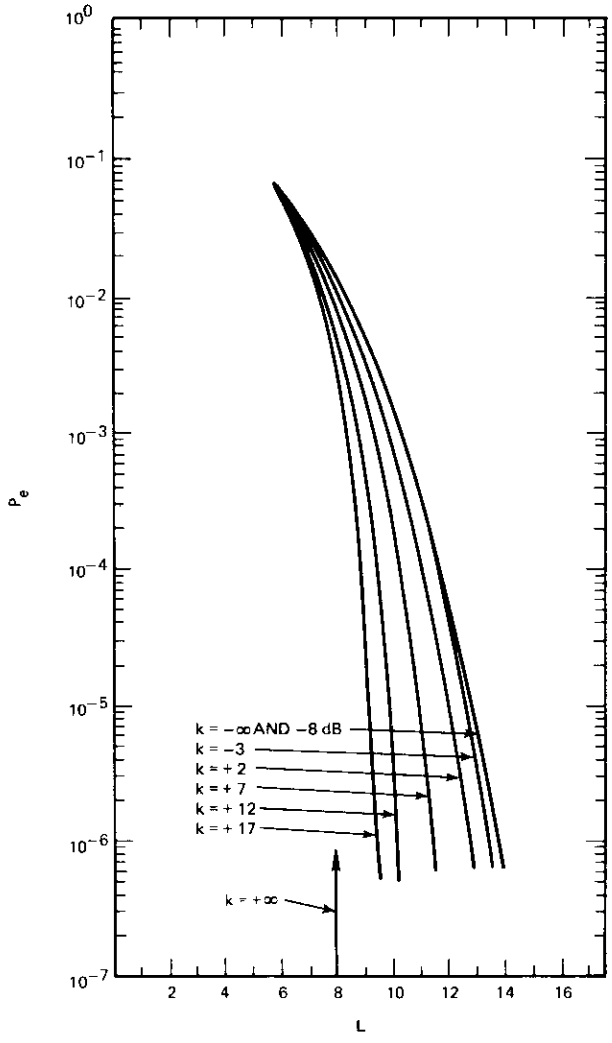


Figure 3c. *Error Probabilities of 4-Phase PSK in the Presence of Gaussian Noise and Three Cochannel Interferences ($H = 3, 4\phi$)*

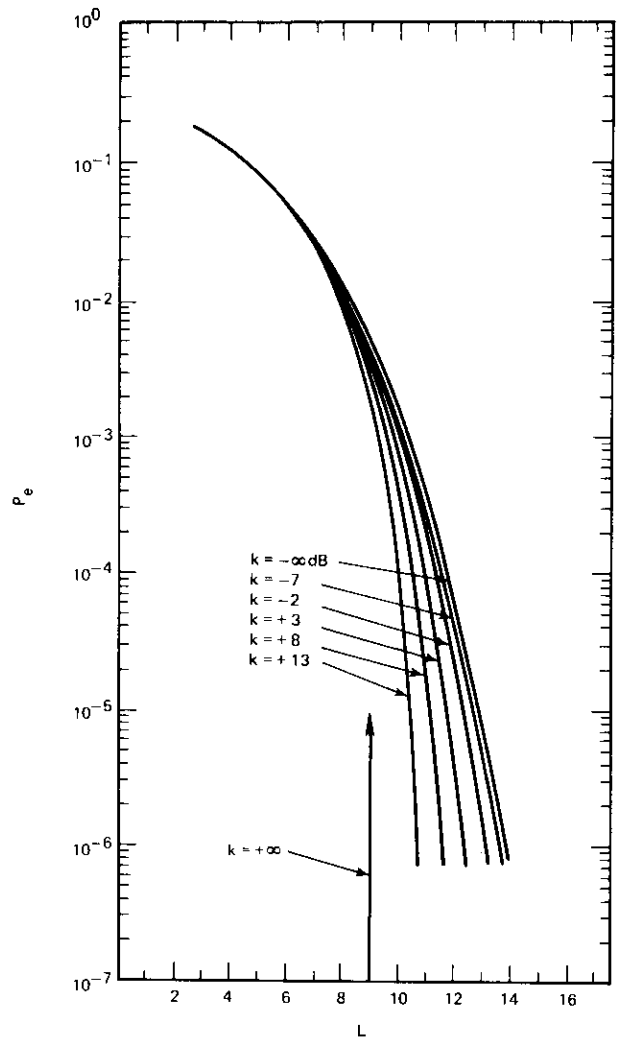


Figure 3d. *Error Probabilities of 4-Phase PSK in the Presence of Gaussian Noise and Four Cochannel Interferences ($H = 4, 4\phi$)*

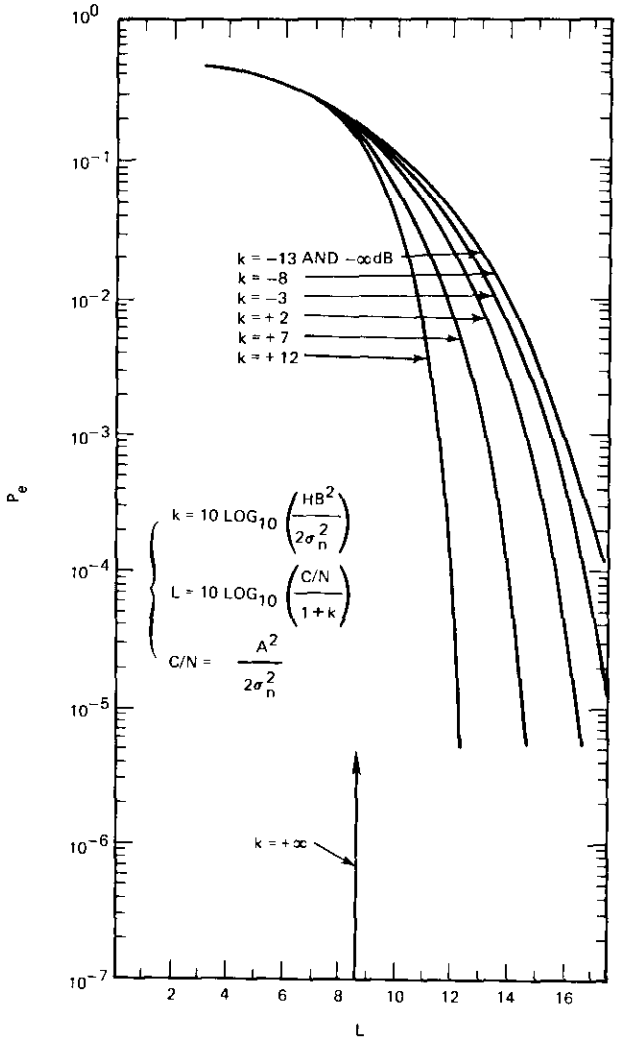


Figure 4a. *Error Probabilities of 8-Phase PSK in the Presence of Gaussian Noise and One Cochannel Interference ($H = 1$, 8ϕ)*

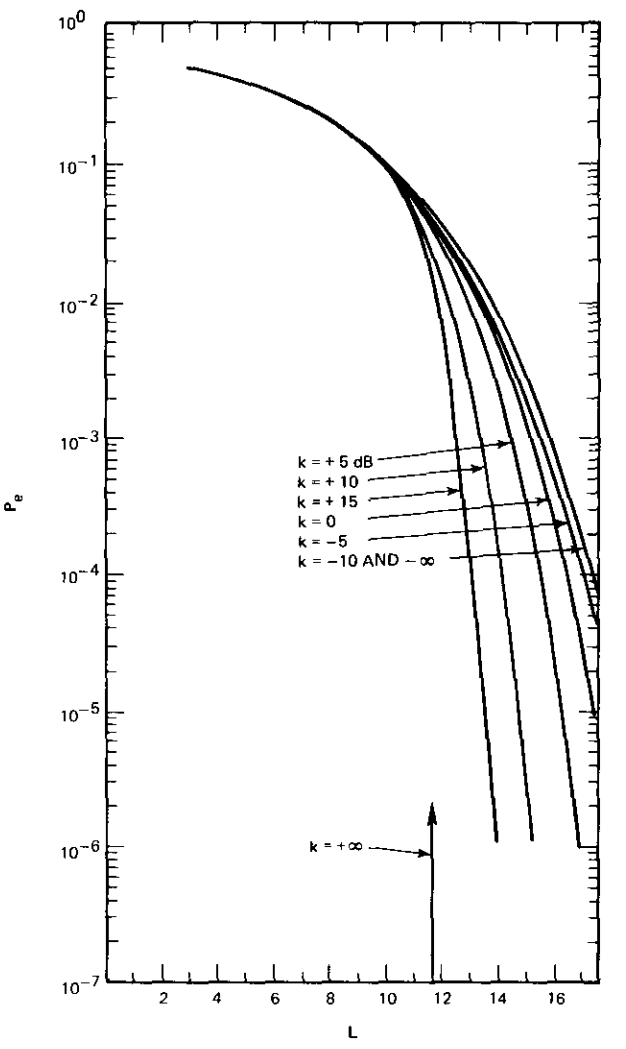


Figure 4b. *Error Probabilities of 8-Phase PSK in the Presence of Gaussian Noise and Two Cochannel Interferences ($H = 2$, 8ϕ)*

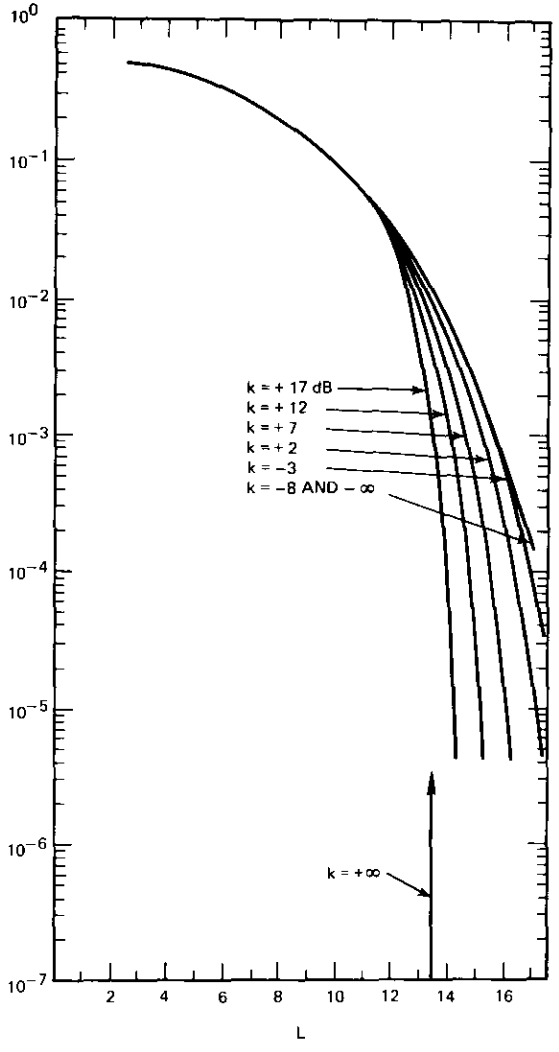


Figure 4c. *Error Probabilities of 8-Phase PSK in the Presence of Gaussian Noise and Three Cochannel Interferences ($H = 3, 8\phi$)*

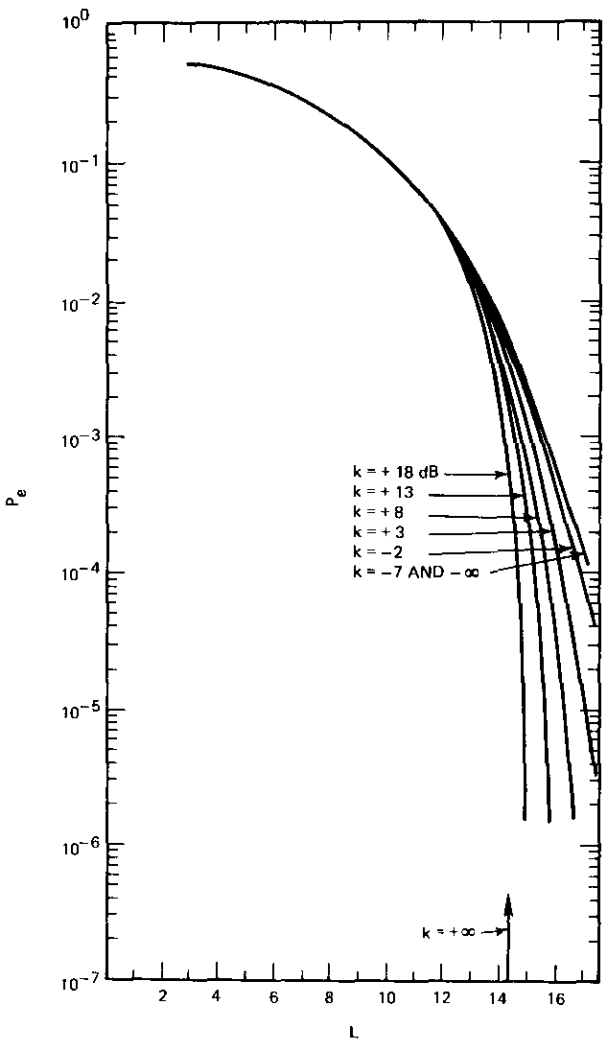


Figure 4d. *Error Probabilities of 8-Phase PSK in the Presence of Gaussian Noise and Four Cochannel Interferences ($H = 4, 8\phi$)*

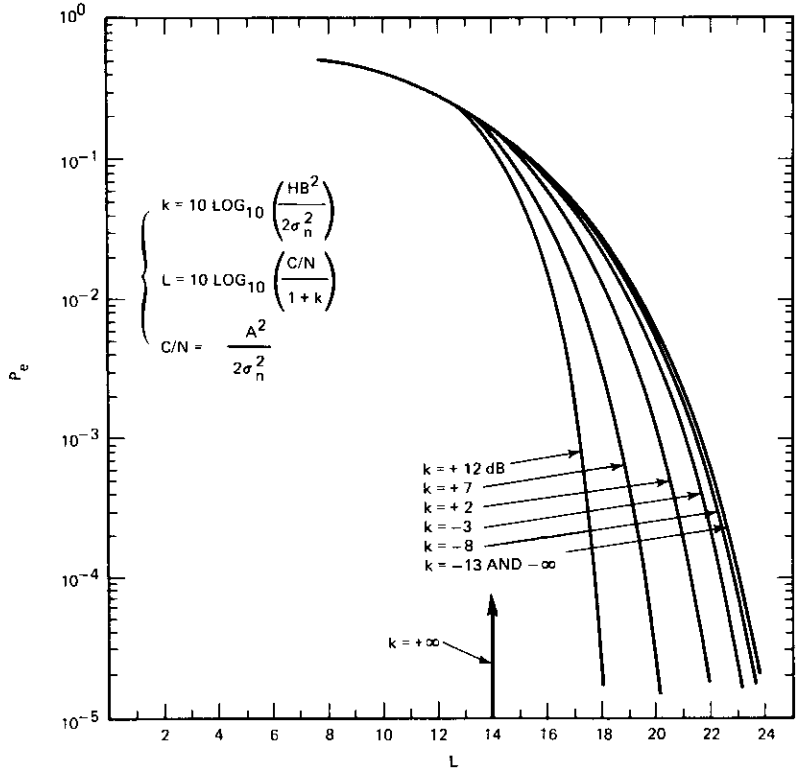


Figure 5a. *Error Probabilities of 16-Phase PSK in the Presence of Gaussian Noise and One Cochannel Interference ($H = 1$, 16ϕ)*

carrier-to-total-interference-and-noise power ratio, L, the error probability increases as the number of interferences, H, increases. This is quite reasonable. That is, as H increases to $+\infty$, the statistics of the sum of these H independent interferences become more close to Gaussian. Thus, the error probability should approach that caused by Gaussian noise with the same amount of power.

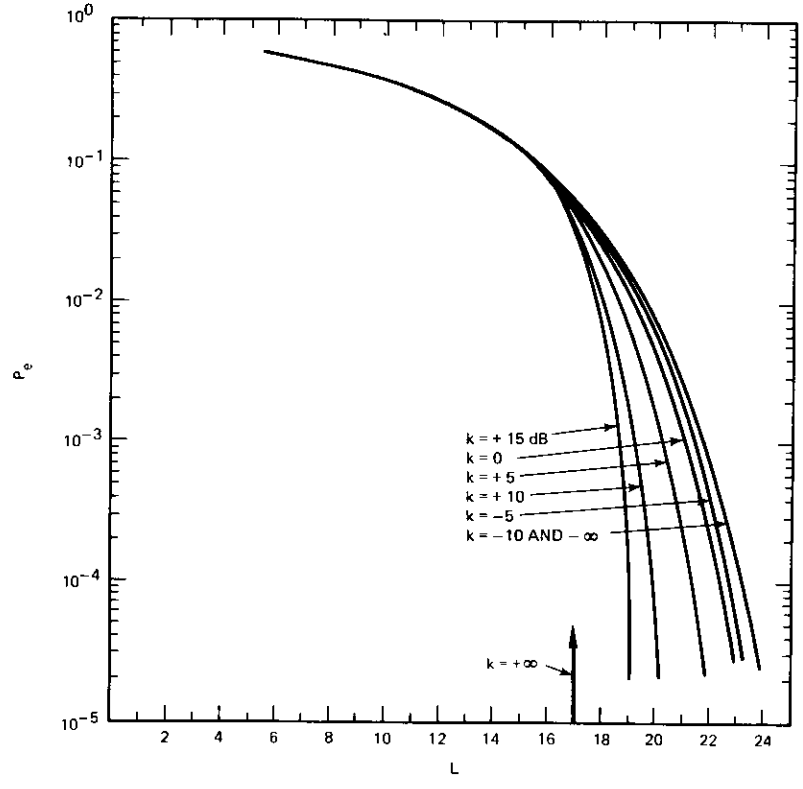


Figure 5b. *Error Probabilities of 16-Phase PSK in the Presence of Gaussian Noise and Two Cochannel Interferences ($H = 2$, 16ϕ)*

Conclusion

Adaptation of the results obtained in Reference 8 to analyze the effects of cochannel interference in M-ary PSK systems has been demonstrated. The necessary modifications are few. The computational procedure described in this paper has been programmed on COMSAT's computer to

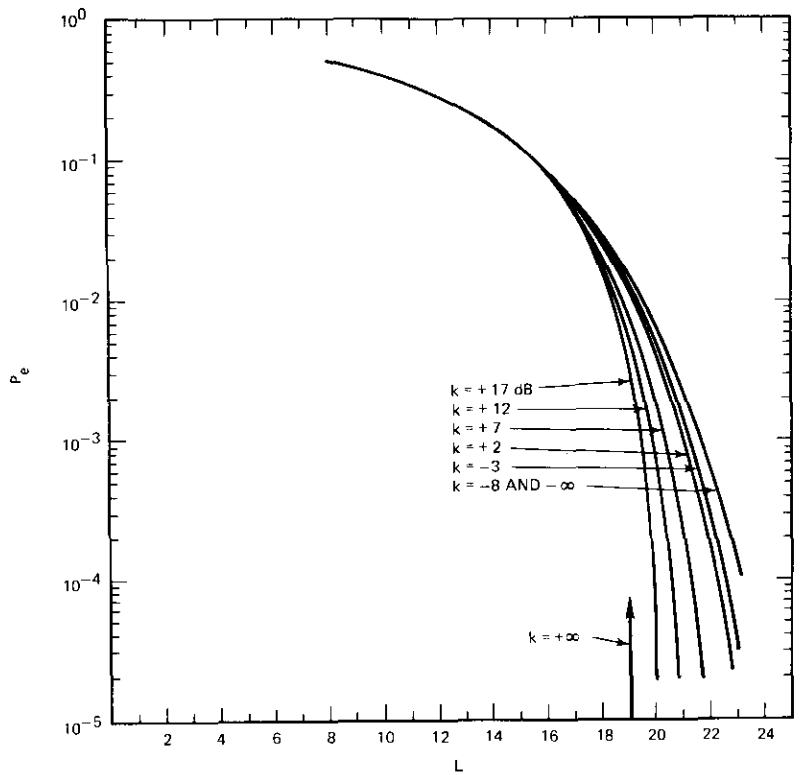


Figure 5c. *Error Probabilities of 16-Phase PSK in the Presence of Gaussian Noise and Three Cochannel Interferences ($H = 3$, 16ϕ)*

yield numerical results. These numerical results can also provide a tradeoff in system design between the required carrier-to-noise power ratio and carrier-to-interference power ratio to achieve a given error performance for a given number of interferences. For example, from Figure 3a for the case of 4-phase PSK and a single cochannel interference, to achieve a 10^{-4} error probability, L must equal 8.5 and 11.2 dB for $k = 7$ and -3 dB, respectively. Therefore, to increase the carrier-to-noise-and-interference power ratio, L, from 8.5 to 11.2 dB, the interference-to-noise power ratio can be reduced from 7 dB to -3 dB. Many other tradeoffs can be also made.

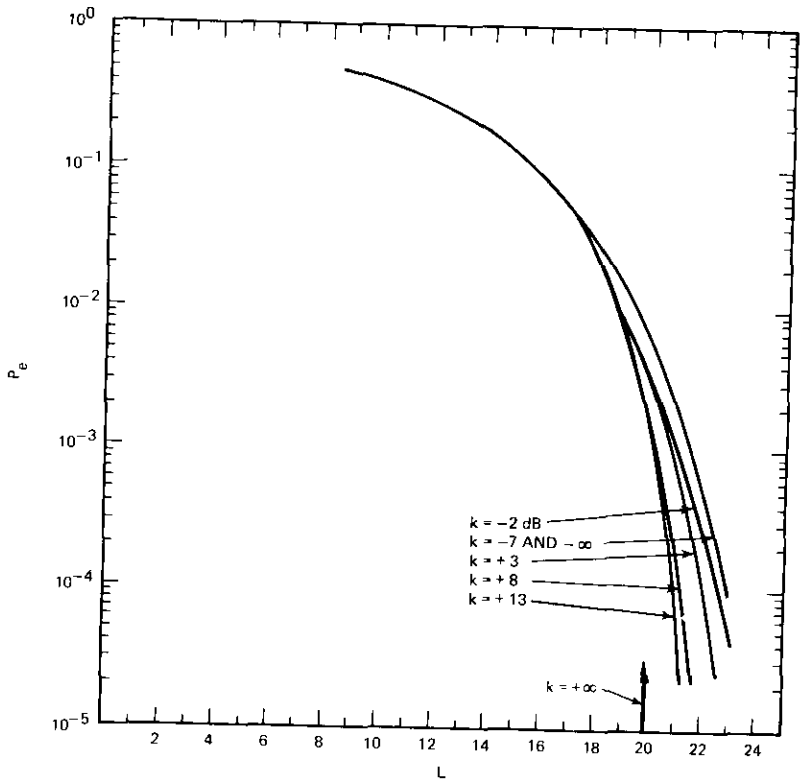


Figure 5d. *Error Probabilities of 16-Phase PSK in the Presence of Gaussian Noise and Four Cochannel Interferences ($H = 4$, 16ϕ)*

It should be noted here that intersymbol interference has been assumed to be negligible in this paper. The tradeoff between noise, cochannel interference, and intersymbol interference will be presented elsewhere.

Acknowledgment

The authors wish to thank Dr. K. Bhatnagar for his contributions to the computation of all curves given in this paper.

References

[1] A. S. Rosenbaum, "PSK Error Performance with Gaussian Noise and Interference," *Bell System Technical Journal*, Vol. 48, February 1969, pp. 413–442.

[2] A. S. Rosenbaum, "Binary PSK Error Probabilities with Multiple Co-channel Interferences," *IEEE Transactions on Communications Technology*, COM-18, June 1970, pp. 241–253.

[3] V. K. Prabhu, "Error Rate Consideration for Coherent Phase-Shift Keyed Systems with Co-channel Interference," *Bell System Technical Journal*, Vol. 48, March 1969, pp. 743–767.

[4] V. K. Prabhu, "Error-probability Upper Bound for Coherently Detected Signals with Co-channel Interference," *Electronics Letters*, Vol. 5, No. 16, August 7, 1969.

[5] M. C. Jeruchim and D. A. Kane, "Orbit/Spectrum Utilization Study," Vol. IV, Document No. 70SD4293, General Electric Company, December 31, 1970.

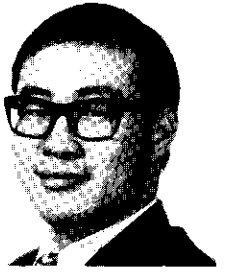
[6] J. Goldman, "Multiple Error Performance of PSK Systems with Co-channel Interference and Noise," *IEEE Transactions on Communications Technology*, COM-19, No. 4, August 1971, pp. 420–430.

[7] J. M. Aein, "On the Effects of Undesired Signal Interference to a Coherent Digital Carrier," Institute for Defense Analysis, Paper p-812, IDA Log No. HQ71-13729, February 1972.

[8] O. Shimbo, R. Fang, and M. Celebiler, "Performance of M-ary PSK Systems in Gaussian Noise and Intersymbol Interference," *IEEE Transactions on Information Theory*, IT-19, No. 1, January 1973, pp. 44–58, IT-19, No. 3, May, 1973, p. 365 (corrections).

Osamu Shimbo is Senior Scientist, Advanced Studies Laboratory, COMSAT Laboratories. Before joining COMSAT, he was a Senior Member of the Scientific Staff in the Research and Development Laboratories, Northern Electric Co., Ltd.; Senior Scientific Staff Member at Hirst Research Centre, General Electric Co., Ltd., of England; an exchange visitor in the Department of Electrical Engineering, Columbia University; and a Manager in the Transmission Laboratory of Oki Electric Industrial Co., Ltd., of Japan.

He holds a Bachelor of Engineering degree from Tohoku University, Japan (1956), and a Doctor of Engineering degree from Hokkaido University of Japan (1965) and has received awards for a paper on "Synchronization of PCM Systems" and for achievement in FM and PCM systems analysis.



Russell J. F. Fang was born in Chungking, China; received a B.S. degree in electrical engineering from the National Taiwan University, Taipei, China (1962), and M.S. and Ph.D. degrees from Stanford University in 1964 and 1968, respectively.

From 1962 to 1963 he was with the Chinese Air Force Electronics Division, Pingtung, Taiwan; from 1964 to 1965 he was employed by the Stanford Research Institute, and from 1965 to 1968 he was with the Stanford Electronics Laboratories. Since 1968 he has been with COMSAT Laboratories. Among his current interests are interference and signal processing problems, feedback communications schemes, and high information-rate coding techniques.

Dr. Fang is a member of the IEEE and the Institute of Mathematical Statistics.

CTR Notes

Antarctic unattended earth station

D. W. LIPKE

Since mid-February 1972, when the first Pacific INTELSAT IV became operational, an unattended earth station located near McMurdo Station on Ross Island (166.6°E, 77.8°S), Antarctica, has transmitted data to the continental United States. The unmanned geophysical observatory (UGO) station, supported and owned by the National Science Foundation, is being used to assess the operation and performance of small terminals for relaying real-time scientific data from a remote collection point to a central processing center for reduction.

Two-way communications between Antarctica and Jamesburg, California, consisting of command and data transmissions to and from the UGO terminal, respectively, are provided by relaying the signals through an INTELSAT IV F-4 global-beam transponder. The communications link is extended from Jamesburg to a computer facility at Stanford University (National Science Foundation contractor) by commercial telephone lines.

Data can be transmitted from Antarctica in either of two modes:

a. a nonreturn-to-zero, biphase modulated carrier providing a data rate of 833 bps; or

b. a carrier which is frequency modulated by an analog data signal extending from 100 to 3000 Hz, and an FSK subcarrier (833 bps) at 5.5 kHz.

Selection of the mode of transmission and the modulation parameters (e.g., frequency deviation), and operational control of the unattended station can be performed at a remote location by means of a 100-bps command data link.

The primary components of the antarctic station are a fixed pointing antenna having a diameter of approximately eight feet, a 20-watt TWT output amplifier, electronics equipment for encoding scientific and operational data into the proper format for transmission, and a receiving system, which processes command signals and executes the desired opera-

---

D. W. Lipke is Manager, Special Services, Special Projects Division, COMSAT Laboratories.

tions. The electronics equipment is housed in an environmentally controlled capsule located atop a tripod support, approximately 20 feet above ground level. Although the station is now powered from the supply at McMurdo Station, the National Science Foundation is investigating the use of propane thermoelectric generators for remote applications.

The UGO station operates with a G/T of 7 dB/°K, an e.i.r.p. of 52 dBW, and an antenna elevation angle to the nominal satellite point (174°E) of 3.35° in a straight line, or approximately 3.6° when atmospheric bending is included. Since the antenna is not steerable, the transmission quality is a function of the antenna half-power beamwidth (1.5° at 6 GHz) and the satellite location. However, the orbital parameters of the satellite have caused it to remain within the 3-dB contour of the UGO antenna pattern, so that PSK transmissions from Antarctica have been received for 24 hours a day at Jamesburg with a carrier-to-noise density equal to or greater than 51 dB-Hz, the nominal design value, and an error rate of about one part in 10^7.

The successful operation of the antarctic terminal, the first unattended earth station to be used with an INTELSAT satellite, has demonstrated the utility of this type of station for special applications. The implementation of the program has been largely a result of the efforts of Stanford University personnel, in particular, Dr. Michael Sites, and personnel from the National Science Foundation.

Battery-powered electric propulsion for north-south stationkeeping

B. A. FREE AND J. D. DUNLOP

Introduction

Usually, synchronous communications satellites carry an onboard battery to power the satellite during eclipse, and a propulsion system to keep the satellite oriented and on station. Ni-Cd batteries and hydrazine monopropellant are the conventional agents for these tasks. The advantages of the Ni-H_2 battery over the Ni-Cd battery in terms of energy density, cycle lifetime, and trouble-free operation have been illustrated in a previous paper [1]. In the propulsion area, the advantages of solar-cell-powered electric propulsion have also been described [2]; the most prominent improvement is a substantial propulsion system weight reduction with respect to conventional hydrazine systems. This note describes the additional advantages which arise from an electric propulsion system powered by Ni-H_2 batteries.

Solar-cell-powered electric propulsion

Thrust programs may provide high thrust over a short thrusting time or low thrust over a long thrusting time as long as the total impulse requirement is met. When electric thrusters are directly powered by a dedicated portion of the solar array, the latter option is chosen to keep the power budget low. As a consequence, small, relatively inefficient ion thrusters are used, and the cumulative lifetime requirements are high.

The tradeoff between power and lifetime for solar-cell-powered electric propulsion is shown in Figure 1. The customary range of operation (5–10 mN) is shown as a shaded area. Confinement of the thrust level to this range for satellite sizes of the order of 1,000 kg leads to the type of thrust program illustrated in Figure 2. Note that two operational thrusters

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

Bernard A. Free is a Member of the Technical Staff in the Physics Laboratory, Applied Sciences Division, COMSAT Laboratories.

James Dunlop is Manager of Energy Storage, Physics Laboratory, Applied Sciences Division, COMSAT Laboratories.

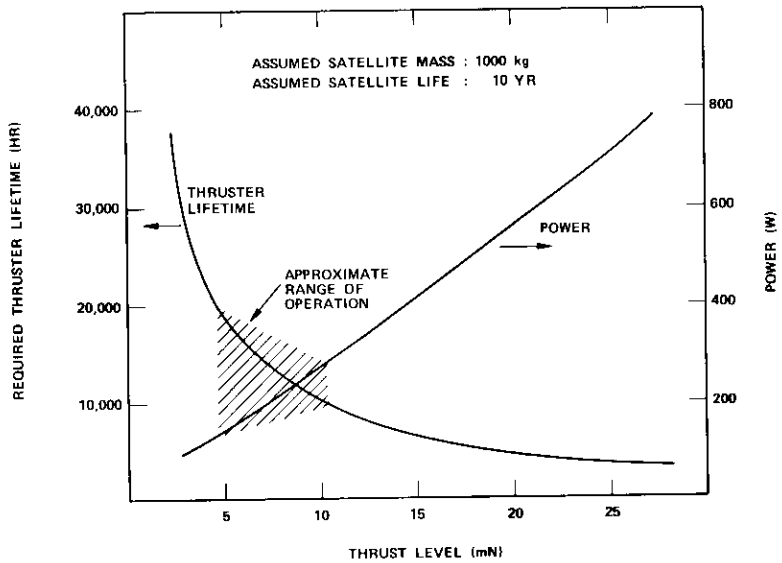


Figure 1. *Tradeoff Between Electric Thruster Lifetime and Required Power*

are required, one each for operation near the two daily nodes, so that thruster lifetime is half the propulsion system on-time.

New battery-powered electric propulsion concept

A new thrust program, shown in Figure 3, is now proposed. This program employs a high thrust for a short thrusting time. The higher power required is supplied by a battery. In an ideal situation, the propulsion system is designed to operate with the battery pack, which is already on board for eclipse operation. In this case, no additional batteries are required for electric thruster operation.

This concept is effective only with batteries which have a higher required cycle life expectancy without degradation of the useful energy density. Ni-Cd batteries cannot be used because the cycle life expectancy can be increased as required only by lowering the depth of discharge drastically, which results in a major decrease in the energy density and a major weight increase. On the contrary, Ni-H_2 batteries, currently under development, are expected to exhibit three to five times the cycle life expectancy of the Ni-Cd battery; hence they will be capable of satisfying all requirements.

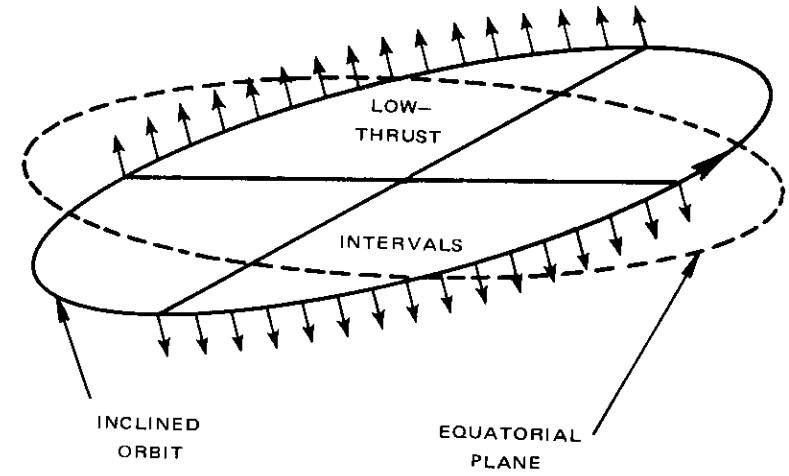


Figure 2. *Thrust Program for North-South Stationkeeping with a Small Electric Thruster*

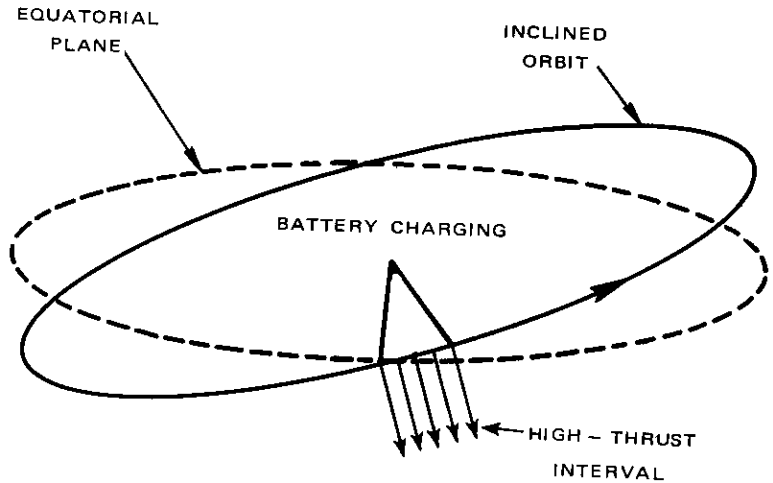


Figure 3. *Thrust Program for North-South Stationkeeping with Battery-Powered Electric Propulsion*

Consequences

The adoption of this energy-storage concept will make it possible to perform N-S stationkeeping with electric propulsion at higher power and thrust over a much shorter duty cycle, for example, six times the thrust

for one-sixth of the time. In addition to the much lighter weight of the battery pack and propulsion system, the following benefits will be realized:

a. More effective use will be made of batteries formerly used only during eclipse periods.

b. Larger thrusters will exhibit higher electrical efficiency, higher propellant utilization, more uniform beam profile, and longer wearout lifetimes.

c. Shorter duty cycles will reduce the required lifetime for each thruster from 20,000 hours to 6,000 hours, and greatly facilitate laboratory endurance testing.

d. Daily (rather than twice daily) correction will halve the number of propulsion system cycles.

e. Since either of two large thrusters is adequate, a 2-thruster system will be redundant. In the case of small thrusters, a 4-thruster system will be necessary for redundancy purposes.

f. The larger thruster will make it possible to execute other propulsion tasks, e.g., station changing and initial orbit trimming, more rapidly.

g. The new thrust level and duty cycle will bring all thruster requirements within the present state-of-the-art with respect to both efficiency and lifetime.

h. It will be possible to use a given thruster design for large or small satellites by suitably altering the duty cycle.

Thus, it appears that many potential problems associated with long-term electric-propulsion N-S stationkeeping can be eliminated or substantially reduced by utilizing the new energy-storage concept. In addition, flight hardware can be fabricated and tested much more rapidly as a result of the advanced state-of-the-art for large thrusters and the much shorter lifetime requirements.

References

[1] J. F. Stockel et al., "A Nickel-Hydrogen Secondary Cell for Synchronous Orbit Applications," *Proceedings of the 7th IEEE Intersociety Energy Conversion Engineering Conference—1972*, San Diego, California, September 1972, p. 87.

[2] B. A. Free, "Chemical and Electric Propulsion Tradeoffs for Communications Satellites," *COMSAT Technical Review*, Vol. 2, No. 1, Spring 1972, p. 123.

Translations of Abstracts

La cellule violette: Une cellule solaire au silicium améliorée

J. LINDMAYER ET J. ALLISON

Sommaire

Les cellules solaires au silicium actuelles ont un faible rendement quantique à des longueurs d'ondes courtes; en dessous de 0,5 μm la réponse typique tombe très nettement. A la suite de travaux intensifs on a pu étendre la réponse à des longueurs d'ondes aussi courtes que 0,3 μm, ce qui améliore sensiblement le courant de la cellule solaire. Un facteur de remplissage plus élevé a permis d'obtenir une efficacité de conversion encore plus grande. En combinant la réponse à une courte longueur d'onde et une courbe I-V plus marquée on est arrivé à une efficacité de conversion qui dépasse d'environ 30 pour cent celle des cellules actuelles utilisées pour l'espace. La cellule solaire améliorée s'appelle "la cellule violette".

Influence de l'ambiance de rayonnement dans l'espace sur la conception du satellite Intelsat IV

R. W. ROSTRON

Sommaire

Une étude approfondie a été effectuée afin de spécifier l'ambiance de radiations prévue pour INTELSAT IV et de prédire ses effets sur les cellules solaires au silicium. Les résultats de cette étude, provenant des données les plus récentes acquises par des satellites ou en laboratoire ont été présentées au constructeur du satellite sous forme d'un modèle de travail destiné à choisir les dimensions des panneaux solaires, et à déterminer la protection nécessaire pour les cellules solaires ainsi que d'autres composants électroniques sensibles aux radiations. On présente ici ce modèle, graphiquement et analytiquement, sous forme de l'intégrale dans le temps du flux d'électrons et de protons en fonction de l'énergie de ces particules. On présente aussi des courbes mon-