September 1985 Vol. 64 No. 7

AT&T TECHNICAL JOURNAL

A JOURNAL OF THE AT&T COMPANIES

Optical Fibers

Laser Propagation

Microwave Communication

Representation Theory

Transmission Networks

Bandwidth-Error Exchange

Distribution Theory

Queueing

AT&T TECHNICAL JOURNAL (ISSN 8756-2324) is published ten times each year by AT&T, 550 Madison Avenue, New York, NY 10022; C. L. Brown, Chairman of the Board; L. L. Christensen, Secretary. The Computing Science and Systems section and the special issues are included as they become available. Subscriptions: United States—1 year $35; foreign—1 year $45.

Payment for foreign subscriptions or single copies must be made in United States funds, or by check drawn on a United States bank and made payable to the AT&T Technical Journal, and sent to AT&T Bell Laboratories, Circulation Dept., Room 1E335, 101 J. F. Kennedy Pky, Short Hills, NJ 07078.

Back issues of the special, single-subject supplements may be obtained by writing to the AT&T Customer Information Center, P.O. Box 19901, Indianapolis, Indiana 46219, or by calling (800) 432-6600. Back issues of the general, multisubject issues may be obtained from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

Printed in U.S.A. Second-class postage paid at Short Hills, NJ 07078 and additional mailing offices. Postmaster: Send address changes to the AT&T Technical Journal, Room 1E335, 101 J. F. Kennedy Pky, Short Hills, NJ 07078.

AT&T
TECHNICAL
JOURNAL

VOL. 64 SEPTEMBER 1985 NO. 7

AT&T Technical Journal
Vol. 64, No. 7, September 1985
Printed in U.S.A.

Thermal Buckling of Dual-Coated Fiber

By T. A. LENAHAN*

(Manuscript received August 8, 1984)

An analysis of buckling is presented for dual-coated fibers within their coating at low temperatures. Buckling causes microbending of the fiber axis, the prime source of added optical loss in the fiber. Buckling is caused by compressive stress exerted on the fiber by the coating, which arises because the thermal expansion coefficient of the coating is substantially larger than that of the fiber. Calculations show that buckling is more likely when the inner primary coating layer is soft and thick. Previous experimental results on fibers in cables indicate that softer, thicker primaries lead to more added loss at low temperatures, contrary to the simple model where lateral pressure imprints irregularities onto the path of the fiber. Such is the evidence that buckling occurs. The theory is applied to various coating designs. The calculated results rank the low-temperature performance reliably, but they also indicate that thermal strain falls short of the buckling strain. Other sources of strain are suggested, including a mechanism whereby irregularities induced by lateral pressure on the outside produce bending moments on the fiber.

I. INTRODUCTION

Optical fibers put in cables and/or subjected to low temperatures can have added transmission loss, attributed to microbending of the fiber axis.[1] Microbending can result from increased lateral pressure inside the cable, which imprints irregularities there onto the path of the fiber.[2] The imprinting is accentuated when the materials are stiffer, as at lower temperatures. Dual coatings (see Fig. 1) were introduced to reduce the effects of lateral pressure by buffering the fiber with a soft inner primary layer; the outer secondary layer is hard and robust

* AT&T Bell Laboratories.

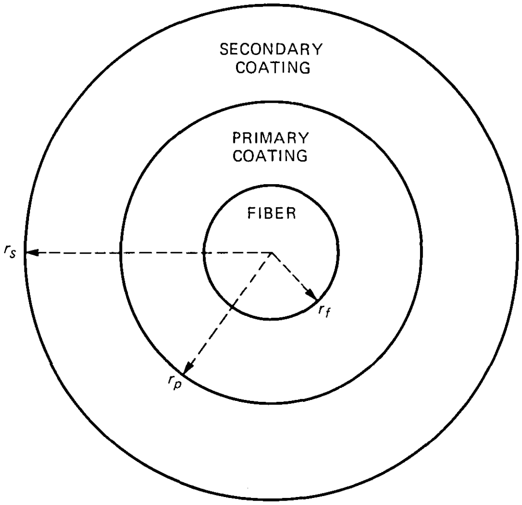


Fig. 1—A dual-coated fiber.

to allow handling. In practice, dual-coated fiber does indeed show substantially less added loss than single-coated fiber.[3]

Microbending also can arise from buckling of the fiber. Coating materials have thermal expansion coefficients two to three orders of magnitude larger than that of the fiber; hence, the coating exerts a compressive stress on the fiber at low temperatures. When the thermal strain exceeds a critical limit, the fiber buckles within the primary coating. Even below the buckling limit, thermal strain may amplify fiber bending already present. Buckling has been directly observed in (1) single-coated fiber within a soft elastomer under forced bending[4] and (2) dual-coated resistance wire under thermal stress.[5]

Indirect evidence exists for buckling of dual-coated fiber. Yabuta et al. report that thicker primary coatings lead to increased added loss at −50°C, irrespective of the outer secondary thickness, even though the primary material (silicone) remains soft at −50°C (see Fig. 7 of Ref. 6). This finding is not consistent with the simple lateral-pressure model, which says that more buffering gives less microbending. It is consistent with the buckling model because, as will be shown, thicker and softer primary layers provide less resistance to lateral displacement of the fiber and hence to fiber buckling. Results of the so-called S5 experiment,[7] involving various coating designs, likewise indicate that softer thicker primaries lead to more added loss at low temperatures.

In this paper thermal buckling of a dual-coated fiber within its coating is analyzed. Thermal strain on the fiber is calculated numerically and found to agree reasonably well with the simple rule of mixtures. Buckling strain, the minimum strain required for the fiber to buckle, depends mainly on the lateral rigidity of the fiber in its coating and is measured by a spring constant κ, which is also calculated numerically. The buckling analysis of Refs. 5 and 6 was flawed because these references incorrectly assumed that κ was the modulus of the primary.

The theory is applied to various designs including ones of Yabuta and ones used in the S5 experiment. Results indicate, as above, that the fiber is more likely to buckle when the primary coating is soft and thick. The calculated thermal strain by itself is less than the calculated buckling strain in all cases, but other sources of strain can combine with the thermal to reach the buckling strain. For instance, thermal stress combined with irregularities in the secondary caused by lateral pressure would put moments on the fiber.

Whatever the exact mechanism for microbending, coating designs should account for the possibility of buckling. Experimental results indicate that the primary should be sufficiently thin, especially if the primary material remains soft at low temperatures.

In the next section the buckling analysis is developed and the associated numerical calculations are indicated. In Section III the theory is applied to the resistance wire of Katsuyama et al.[5] and other silicone/nylon coatings of Yabuta et al.[6] and designs from the S5 experiment.[7] The paper is summarized in Section IV and certain aspects outside the model that favor buckling are discussed. One of the most important aspects, already mentioned, is the prospect that irregular lateral pressure in conjunction with thermal stress produces bending moments on the fiber.

II. ANALYSIS

The analysis of fiber buckling involves the force or strain needed to buckle the fiber and the force actually on the fiber. In this section, after the mechanical properties of the relevant materials are indicated, the thermal strain of the fiber is calculated, and then the buckling strain is calculated. The fiber and coating are assumed throughout to be perfectly circular and concentric, homogeneous, and uniform along their length.

2.1 Materials characterization

Two kinds of structures are considered: dual-coated optical fiber and the dual-coated wire of Katsuyama. The dual-coated optical fibers considered have outer secondaries that are either an ultraviolet (UV)-cured urethane acrylate material (Borden) or nylon. The Borden secondary has as its inner primary either UV-cured material supplied

Table I—Material parameters

		$T = -40°C$		$T = 0°C$		
Material	E_g (psi)	E (psi)	ν	E (psi)	ν	α (°C^{-1})
Glass	1.07E7	1.07E7	0.25	1.07E7	0.25	5.04$E - 7$
Wire	2.35E7	2.35E7	0.25	2.35E7	0.25	1.6$E - 5$
Desolite	4.60E5	9.15E3	0.495	2.47E2	0.49987	1.2$E - 4$
Hot Melt	—	3.34E2	0.49982	9.15E1	0.49995	1.4$E - 4$
Silicone	—	2.64E2	0.49986	1.47E2	0.49992	3.0$E - 4$
Borden	6.48E5	5.15E5	0.30	1.96E5	0.4244	0.6$E - 4$
Nylon	—	3.08E5	0.333	2.93E5	0.341	1.0$E - 4$

by De Soto (Desolite*) or a thermoplastic material (Hot Melt). The nylon secondary has silicone as its primary material; this combination is used for the wire of Katsuyama et al.[5] and for optical fiber studied by Yabuta et al.[6]

The modulus E, Poisson ratio ν, and thermal expansion coefficient α are the mechanical properties needed in the buckling analysis. The glass fiber and wire are elastic in the temperature range of interest; hence, a single value for the mechanical parameters characterizes these materials. The coatings are viscoelastic polymeric materials; their modulus values relax over time and depend on temperature. We simplify the modulus values here by considering only their values at 24 hours (a typical time span for temperature drops from room temperature to −40°C).

The values of E and α for the fiber are taken from Table III of Ref. 8 and for the wire from Ref. 5. Their Poisson ratio is taken as $\nu = 0.25$, the accepted value for a material in its glassy (stiff) state. These values appear in Table I.

The 24-hour modulus E of Desolite, Hot Melt, and Borden are shown as functions of temperature in Figs. 2 through 4, respectively. These moduli were synthesized from oscillatory data from a rheometric thermal mechanical spectrometer. Modulus values for 0 and −40°C appear in Table I. The thermal expansion coefficients (at low temperatures) of Desolite and Hot Melt and Borden have been taken from Refs. 9 and 8, respectively. Values of E at 0 and −40°C for silicone and nylon are taken from Ref. 5; the value of α for silicone comes from Ref. 10 and for nylon from Ref. 6. These all appear in Table I.

The Poisson ratios of the coating materials are estimated by assuming that the bulk modulus,

$$K = \frac{E}{3(1 - 2\nu)}, \tag{1}$$

* Registered trademark of De Soto, Inc.

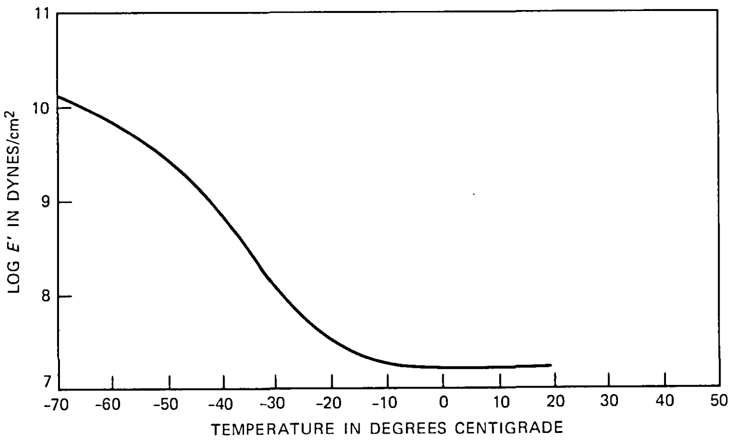


Fig. 2—Isochronal plot (24 hours) for the Young's modulus of Desolite versus temperature.

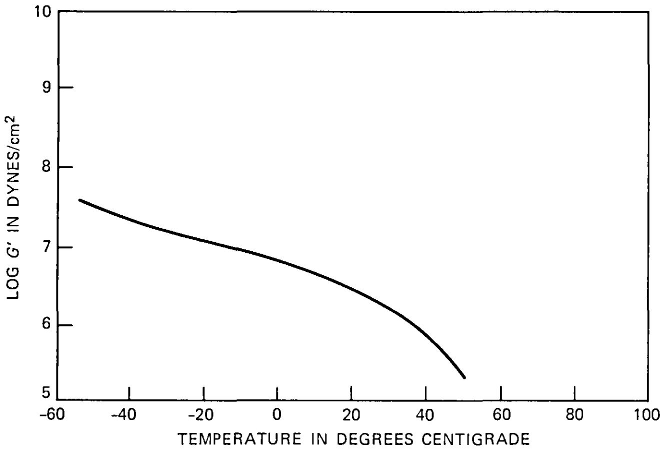


Fig. 3—Isochronal plot (24 hours) for the shear modulus of Hot Melt versus temperature.

is independent of temperature. Studies[11] have shown that K increases only by about a factor of 2 (though E increases several decades) as the material goes from a rubbery to a glassy state. Taking $\nu = 0.25$ at the low temperature glassy plateau (where $E = E_g$), ν is determined at any temperature by

$$\nu = 0.50 - 0.25E/E_g, \tag{2}$$

derived from the constancy of K.

The values of E_g are taken from the 1- or 2-second modulus at the lowest temperature measured (typically −60°C). For primary materials

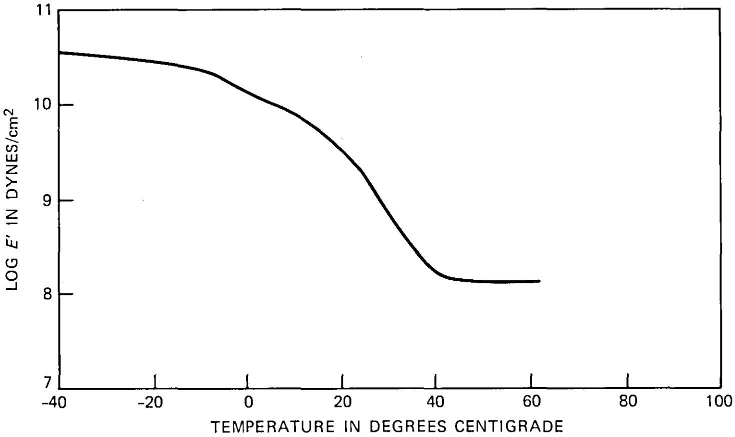


Fig. 4—Isochronal plot (24 hours) for the Young's modulus of Borden versus temperature.

where a plateau value was not attained or was unavailable, the E_g for Desolite was used to estimate ν. Available values of E_g and calculated values of ν appear in Table I.

2.2 Thermal strain

The different thermal expansion coefficients can be consolidated into an effective expansion coefficient α_{eff} for the coated fiber (or wire) as a whole. The rule of mixtures[8] approximates α_{eff} by weighting the various expansion coefficients by the cross-sectional area and the modulus of the corresponding materials. For N materials the formula is

$$\alpha_{\text{eff}} = \sum_{n=1}^{N} \alpha_n A_n E_n \Big/ \sum_{n=1}^{N} A_n E_n, \tag{3}$$

where α_n denotes the expansion coefficient, A_n the area, and E_n the modulus of the nth material.

The rule of mixtures is exact when the coupling of radial displacements is neglected, as if the fiber and coating layers were parallel springs joined at the ends. An analysis accounting for radial coupling is described in Appendix A. Deviations from the rule of mixtures are usually within a few percent, but examples have been found with deviations from −15 to +36 percent.

The thermal strain of the fiber for a temperature change T_0 to T_1 is

$$\epsilon_{\text{therm}} = \int_{T_0}^{T_1} (\alpha_{\text{eff}} - \alpha_{\text{fib}})dT, \tag{4}$$

where α_{fib} denotes the thermal expansion coefficient of the fiber by itself. In general, α_{eff} will depend on time and temperature through its dependence on E and α of the coating materials. With this information, the strain of the fiber can be determined as a function of time for any temperature cycling as in Ref. 8.

For simplicity, the α_{eff} will be taken here as the effective expansion coefficient averaged over the range of temperature change. As for α_{fib}, it is independent of temperature; so the thermal strain of the fiber for a temperature change T is simply

$$\epsilon_{\text{therm}} = (\alpha_{\text{eff}} - \alpha_{\text{fib}})T. \tag{5}$$

The value of α_{eff} will be estimated by using 24-hour modulus values at −40°C and, for comparison, 0°C.

2.3 Buckling analysis

The fiber (or wire) may buckle and follow a wavelike path because of compressive strain. The theory of elastic stability[12] is used to study the buckling of fibers within their coating.

The fiber is treated as a beam in an elastic medium. A force and moment balance yields the differential equation

$$E_f I \frac{d^4 y}{dz^4} + F \frac{d^2 y}{dz^2} + \kappa y = 0 \tag{6}$$

for the deflection y of the fiber as a function of the distance z along the fiber. Small deflections are assumed. The parameters E_f, I, and F denote the modulus of the fiber, the moment of inertia of the fiber ($\pi r_f^4/4$ for radius r_f), and the compressive force on the fiber, respectively. The parameter κ denotes the spring constant of the fiber, which is the ratio of the centering force exerted by the coating to the displacement of the fiber from center.

Early work incorrectly assumed that $\kappa = E_p$,[5,6] but Vangheluwe,[13] assuming a rigid secondary, determined that

$$\kappa = \frac{4\pi E_p(1 - \nu_p)(3 - 4\nu_p)}{(1 + \nu_p)\left[(3 - 4\nu_p)^2 \ln(r_p/r_f) - \dfrac{(r_p/r_f)^2 - 1}{(r_p/r_f)^2 + 1}\right]}, \tag{7}$$

where subscript p signifies the primary region and f the fiber (or wire). A numerical calculation of κ, which accounts for the elasticity of the secondary, is described in Appendix B. The two calculations give identical results when the secondary is assumed rigid. Results indicate that the elasticity of the secondary can reduce κ by more than 80 percent.

The buckling solution has the form

$$y = A \sin\left(\frac{2\pi z}{P}\right), \tag{8}$$

where A is an arbitrary amplitude and P is the pitch. (This form in conjunction with

$$x = A \cos\left(\frac{2\pi z}{P}\right) \tag{9}$$

for the orthogonal component covers the case of helical buckling.) Substituting y into eq. (6) gives

$$A\left[E_f I\left(\frac{2\pi}{P}\right)^4 - F\left(\frac{2\pi}{P}\right)^2 + \kappa\right]\sin\left(\frac{2\pi z}{P}\right) = 0. \tag{10}$$

Hence, the force required for the fiber to buckle with pitch P is

$$F = E_f I\left(\frac{2\pi}{P}\right)^2 + \kappa\left(\frac{P}{2\pi}\right)^2. \tag{11}$$

The minimum buckling force is

$$F_{\min} = 2\sqrt{E_f I \kappa} = r_f^2\sqrt{\pi E_f \kappa}, \tag{12}$$

corresponding to a pitch of

$$P_{\min} = 2\pi(E_f I/\kappa)^{1/4} = \pi r_f(4\pi E_f/\kappa)^{1/4}. \tag{13}$$

The corresponding strain of the fiber is

$$\epsilon_{\min} = F_{\min}/\pi r_f^2 E_f = \sqrt{\frac{\kappa}{\pi E_f}}. \tag{14}$$

This formula shows that fibers having smaller spring constants, which are associated with softer and/or thicker primaries, require less strain to buckle.

If $\epsilon_{\text{therm}} > \epsilon_{\min}$, then the fiber will buckle in its coating. If $\epsilon_{\text{therm}} < \epsilon_{\min}$, then buckling can still occur if $\epsilon_{\text{therm}} + \epsilon_{\text{res}} > \epsilon_{\min}$, where ϵ_{res} denotes residual strain caused by initial bending or other moments. Even if $\epsilon_{\text{therm}} + \epsilon_{\text{res}} < \epsilon_{\min}$, thermal bending can occur. In thermal bending, initial bending or other moments of the fiber are accentuated by the thermal stress. The size of the effect grows as $(\epsilon_{\min} - \epsilon_{\text{therm}} - \epsilon_{\text{res}})^{-1}$, as shown in Ref. 12.

III. APPLICATIONS

In this section the buckling analysis is applied to the dual-coated wire of Katsuyama, for which buckling was observed, and to various dual-coated fiber designs. Parameter studies are also presented.

3.1 Wire of Katsuyama and fibers of Yabuta

The wire of Katsuyama et al.[5] had a dual coating where the primary/secondary was composed of silicone/nylon. The material parameters for these are given in Table I. The wire radius was $r_w = 75\ \mu$m. The outer radii of the primary (r_p) and the secondary (r_s) were $r_p/r_s = 175\ \mu\text{m}/600\ \mu\text{m}$.

The calculated spring constant is $\kappa = 6.9E3$ psi, and effective thermal expansion coefficient is $\alpha_{\text{eff}} = 5.37E - 5$°C. If we assume a temperature drop of 100°C, the strain on the wire from eq. (5) is $\epsilon_{\text{therm}} = 0.38$ percent, and the minimum buckling strain from eq. (14) is $\epsilon_{\min} = 0.97$ percent with a pitch of $P_{\min} = 3.39$ mm. These values were obtained using the material parameters for −40°C. For 0°C, where the coating materials are somewhat softer, the calculations give $\kappa = 3.87E3$ psi and $\alpha_{\text{eff}} = 5.25E - 5°\text{C}^{-1}$; the same temperature change of 100°C gives $\epsilon_{\text{therm}} = 0.37$ percent and $\epsilon_{\min} = 0.72$ percent with a pitch of 3.92 mm. These calculated values are summarized in Table II.

As buckling did occur in the Katsuyama wire at −70°C, the difference $\epsilon_{\min} - \epsilon_{\text{therm}}$ estimates the residual strain ϵ_{res} in the wire at 30°C. Relative to the −40°C parameters, $\epsilon_{\text{res}} \sim 0.59$ percent; relative to the 0°C parameters, $\epsilon_{\text{res}} \sim 0.35$ percent.

Yabuta et al. (see Fig. 7 of Ref. 6) studied the silicone/nylon coating system on optical fibers. One design having a relatively thick primary had substantial added loss at −50°C; its dimensions were $r_p/r_s = 250\ \mu\text{m}/471\ \mu\text{m}$. Another design with a thinner primary had negligible added loss even though the secondary had somewhat more cross-sectional area; its dimensions were $100\ \mu\text{m}/448\ \mu\text{m}$. The fiber radius for both was $r_f = 62.5\ \mu$m.

Calculated values for κ, α_{eff}, ϵ_{therm}, $\epsilon_{\min}$, and $P_{\min}$ are given in Table II for both designs using both 0 and −40°C material parameters from Table I. For the lower-temperature parameters, the thick primary gives $\epsilon_{\text{therm}} = 0.61$ percent and $\epsilon_{\min} = 0.81$ percent with a pitch of 3.09 mm, and the thin primary gives $\epsilon_{\text{therm}} = 0.59$ percent and $\epsilon_{\min} = 3.10$ percent with a pitch of 1.57 mm. These values indicate that buckling

Table II—Buckling quantities for certain silicone coatings

Temp.	κ (psi)	α_{eff} (°C^{-1})	ϵ_{therm} (%)	$\epsilon_{\min}$ (%)	$P_{\min}$ (mm)
(a) Katsuyama et al.					
−40°C	6899	$5.367E - 5$	0.377	0.967	3.39
0°	3868	$5.253E - 5$	0.365	0.724	3.92
(b) Yabuta et al., poor					
−40°C	2184	$6.146E - 5$	0.610	0.806	3.09
0°	1219	$5.969E - 5$	0.592	0.602	3.57
(c) Yabuta et al., good					
−40°C	32393	$5.898E - 5$	0.585	3.104	1.57
0°	18571	$5.772E - 5$	0.572	2.350	1.80

in the first case is at least as probable as for the Katsuyama wire, but much less probable in the second.

3.2 S5 experiment

Four dual-coat designs were selected from the S5 experiment.[7] Their performance, based on added loss at low temperatures, ranged from good to bad. Figure 5 shows the added loss versus temperature for the best and worst cases.

Three of the four designs used UV-cured Desolite for the primary, the other used Hot Melt. All used Borden for the secondary. The coating dimensions varied in primary outer diameter/secondary outer diameter from 8/13 to 11/15, expressed in mils. The designs are

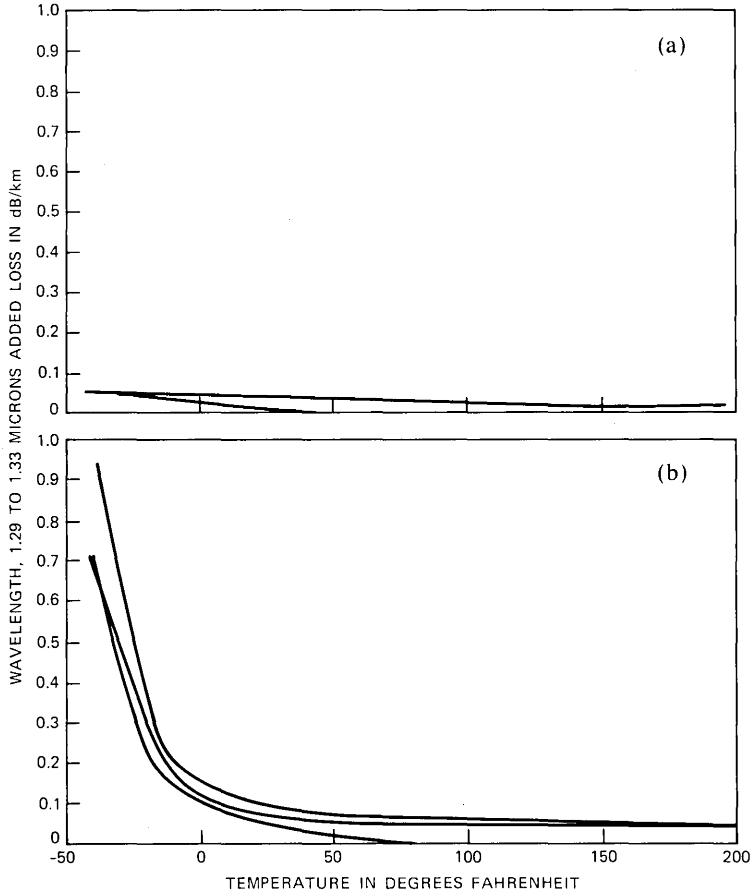


Fig. 5—Added loss at λ ~ 1.3 μm versus temperature for the (a) best coatings from the S5 experiment and (b) worst coatings from the S5 experiment.

Table III—S5 experiment

Size (mils)	Temp. (°C)	Coat (*P*)	Rank	κ (kpsi)	$\alpha_{\rm eff}$ (°C^{-1})	$\epsilon_{\rm therm}$ (%)	$\epsilon_{\rm min}$ (%)	$P_{\rm min}$ (mm)
8/13	−40	UV	Best	570.6	0.112*E* − 4	0.107	13.03	0.76
	0			28.6	0.487*E* − 5	0.044	2.92	1.61
10/13	−40	UV	Second	288.8	0.855*E* − 5	0.081	9.27	0.90
	0			10.3	0.349*E* − 5	0.030	1.75	2.07
11/15	−40	UV	Third	218.1	0.120*E* − 4	0.115	8.05	0.97
	0			7.47	0.494*E* − 5	0.044	1.49	2.18
11/15	−40	Hot Melt	Worst	10.1	0.138*E* − 4	0.133	1.73	2.08
	0			2.78	0.511*E* − 5	0.046	0.91	2.88

specified in Table III, together with calculated values of κ, $\alpha_{\rm eff}$, $\epsilon_{\rm therm}$ (for a 100°C temperature drop), and $\epsilon_{\rm min}$ and $P_{\rm min}$ using the material parameters for −40 and 0°C.

The calculated difference $\epsilon_{\rm min} - \epsilon_{\rm therm}$ tracks the performance rank of the design for both 0 and −40°C. The design with the stiffest and thinnest primary had the least added loss; the one with the softest, thickest primary had the most.

3.3 Parameter studies

The thermal and buckling strains depend on the geometry and materials of the coating. This dependence is now studied in general terms.

The thermal strain $\epsilon_{\rm therm}$ is proportional to the quantity, $\alpha_{\rm eff} - \alpha_{\rm fib}$. By the rule of mixtures,

$$\alpha_{\rm eff} - \alpha_{\rm fib} = \frac{A_pE_p(\alpha_p - \alpha_{\rm fib}) + A_sE_s(\alpha_s - \alpha_{\rm fib})}{A_fE_f + A_pE_p + A_sE_s}. \tag{15}$$

Usually, both the area A_p and modulus E_p of the primary are much smaller than A_s and E_s of the secondary. Neglecting terms with A_pE_p gives

$$\alpha_{\rm eff} - \alpha_{\rm fib} \simeq \frac{A_sE_s(\alpha_s - \alpha_{\rm fib})}{A_fE_f + A_sE_s} = \frac{\alpha_s - \alpha_{\rm fib}}{1 + (A_fE_f/A_sE_s)} \tag{16}$$

or

$$\alpha_{\rm eff} - \alpha_{\rm fib} \simeq (\alpha_s - \alpha_{\rm fib}) \Bigg/ \left[1 + \frac{E_f/E_s}{(r_s/r_f)^2 - (r_p/r_f)^2}\right]. \tag{17}$$

Figure 6 shows plots of this expression versus r_s/r_f, assuming $E_f/E_s = 20$ (as for Borden at −40°C) and $E_f/E_s = 50$ (as for Borden at 0°C) and also $r_p/r_f = 1.5$. The plots again show the well-known fact that larger, stiffer secondaries produce more thermal strain than smaller, softer ones.

The buckling strain $\epsilon_{\rm min}$ from eq. (14) is proportional to $\sqrt{\kappa}$ and inversely proportional to $\sqrt{E_f}$. The latter implies that wire filaments,

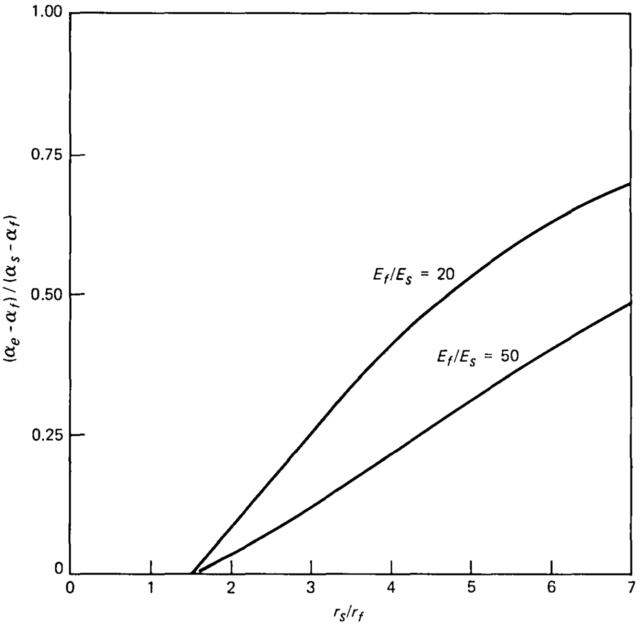


Fig. 6—Effective thermal expansion coefficient versus secondary radius for two secondary modulus values according to the rule of mixtures.

having a higher modulus, require less strain to buckle than glass fibers. The spring constant κ will be studied first for a rigid secondary to which the formula of Vangheluwe in eq. (7) applies. The elasticity of the secondary will be considered subsequently.

If $X \equiv r_p/r_f$ is close to 1, the approximation[14]

$$\ln X \simeq \frac{X^2 - 1}{X^2 + 1} + \frac{1}{3}\left(\frac{X^2 - 1}{X^2 + 1}\right)^3 \tag{18}$$

gives (for $\nu = \nu_p$)

$$\kappa \simeq \frac{4\pi E_p(1 - \nu)(3 - 4\nu)}{(1 + \nu)\left[8(1 - \nu)(1 - 2\nu)\dfrac{X^2 - 1}{X^2 + 1} + \dfrac{(3 - 4\nu)^2}{3}\left(\dfrac{X^2 - 1}{X^2 + 1}\right)^3\right]}. \tag{19}$$

When $\nu \simeq 1/2$, eq. (19) reduces to

$$\kappa \simeq \frac{4\pi E_p(X^2 + 1)^3}{(X^2 - 1)^3}; \tag{20}$$

when ν is not close to 1/2 and $X \simeq 1$, eq. (19) reduces to

$$\kappa \simeq \frac{\pi E_p}{2(1 - 2\nu)}\left(\frac{3 - 4\nu}{1 + \nu}\right)\frac{X^2 + 1}{X^2 - 1} = \frac{\pi K_p}{G}\left(\frac{3 - 4\nu}{1 + \nu}\right)\frac{X^2 + 1}{X^2 - 1}, \tag{21}$$

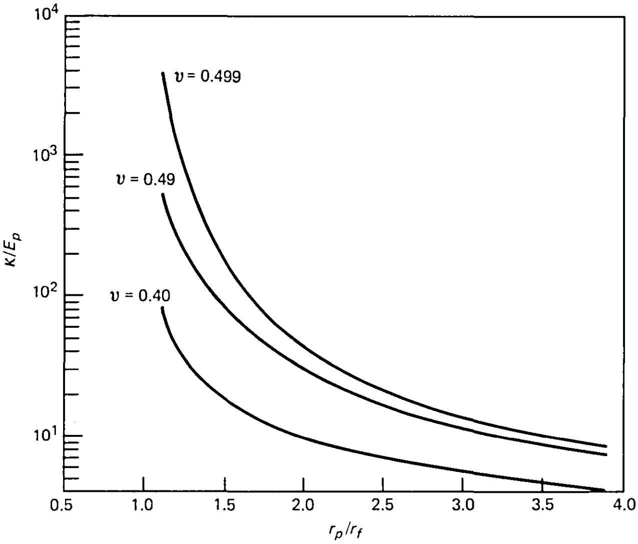


Fig. 7—Spring constant versus primary radius for three Poisson ratios assuming a rigid secondary.

where K_p denotes the bulk modulus of the primary. In all cases, κ increases without bound as X approaches 1 (i.e., as the primary becomes thinner). If X is large, the logarithmic term in the denominator dominates to give

$$\kappa = \frac{4\pi E_p(1 - \nu)}{(1 + \nu)(3 - 4\nu)\ln r_p/r_f}. \tag{22}$$

Thus, κ goes to 0 as r_p/r_f becomes large.

Figure 7 shows κ/E_p plotted versus r_p/r_f for $\nu = 0.40$, 0.49, and 0.499. The curves illustrate the asymptotic behavior of κ given above. They also indicate that thicker primaries have smaller κ with greatest sensitivity when ν is close to 1/2. The compliance of the secondary leads to lower κ.

Figure 8 shows κ versus r_s/r_f for E_s = 500, 200, and 100 kpsi with corresponding ν_s = 0.3, 0.42, and 0.46, respectively (chosen to keep the bulk modulus fixed). The primary was assumed to be Desolite at −40°C with $r_p/r_f = 1.5$. The curves show that κ is smaller for thicker, more compliant secondaries. The compliance of the secondary can cause κ to drop to 20 percent of the value from the formula of Vangheluwe.

IV. SUMMARY AND CONCLUSIONS

This paper has concerned buckling of dual-coated optical fiber caused by compressive stress on the fiber exerted by the coating at

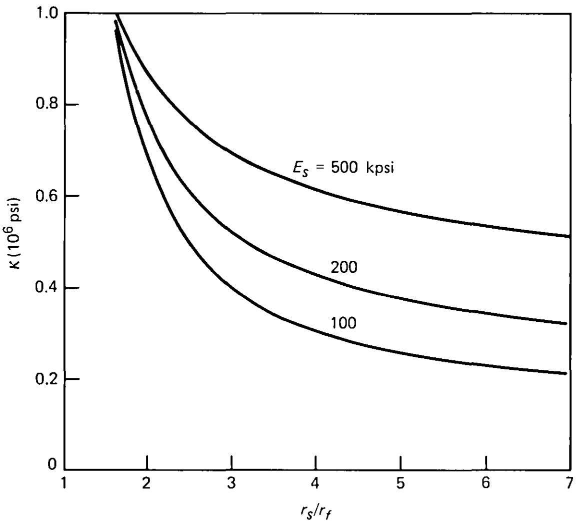


Fig. 8—Spring constant versus secondary radius for three Poisson ratios of the secondary assuming a Desolite primary at −40°C with $r_p/r_f = 1.5$.

low temperatures. The stress arises because the thermal expansion coefficient of the coating is substantially larger than that of the fiber.

Buckling is one of two mechanisms used to explain microbending of the fiber axis, the prime source of added optical loss in the fiber. In the other method, lateral pressure imprints irregularities around the fiber onto the path of the fiber. This occurs at room, as well as low, temperatures and in single-coated, as well as dual-coated, fiber. The antidote to lateral pressure is dual coating where the inside primary coating buffers the fiber and decouples it from the outside. However, too much buffering has been found to produce more added loss at low temperatures, presumably due to buckling.

Two basic ways exist for preventing buckling. The thermal strain on the fiber might be decreased, or the strain needed to buckle might be increased.

Thermal strain depends mostly on the secondary layer. Smaller secondaries put less strain on the fiber; secondaries of a more compliant material or ones with a smaller thermal expansion coefficient have the same effect. These changes are compatible with the dictates of the simple lateral-pressure model.

The buckling strain depends on the spring constant κ. Thinner primaries and, to a lesser extent, thinner secondaries have larger buckling strains. Stiffer primary materials and, to a lesser extent, stiffer secondary materials provide the same effect. The magnitude of these effects is indicated in Fig. 7 for the primary and in Fig. 8 for the secondary. In general, reduced buffering provides more resistance to buckling—just the opposite of what the simple lateral-pressure model dictates.

Experimental results indicate that buckling does occur. In the silicone/nylon coatings of Yabuta et al.,[6] thicker primaries were associated with substantially more added loss at low temperatures. Calculations in Section III showed that the coating with thin primary and low added loss at −40°C needs about four times as much strain to buckle as the one with thick primary and high added loss at −40°C. The primary material (silicone) remains relatively compliant (264 psi) at this temperature. The experimental results are consistent with the buckling mechanism, but not with the lateral pressure mechanism by itself. Similar results were found for fibers in the S5 experiment.[7] Thick primaries of the compliant Hot Melt material performed poorly compared with thinner primaries of De Soto, which stiffens at low temperatures. This outcome may be explained by the buckling calculations in Section III, which indicate that the spring constant for the latter coating is 50 times greater than for the former.

Nevertheless, the calculated thermal strain was less than the buckling strain in all cases, even for the wire of Katsuyama et al. where buckling was known to have occurred. The short fall may be explained by various factors outside the model, as follows:

1. Initial bending (e.g., stranding) involves a strain on the fiber, which adds to the thermal strain to bring the total closer to the buckling strain.

2. The thermal strain depends on the temperature drop, which was taken as 100°C. The precise temperature drop should be measured from a reference temperature where the residual stress on the fiber is 0. Because this temperature is uncertain, the temperature drop is also and might be more than 100°C.

3. Asymmetry of the coating or eccentricity of the fiber in its coating in conjunction with thermal stress produces bending moments along the fiber because, at equilibrium, moments on the coating must be balanced by moments on the fiber. These deformations can arise from imperfections in the coating process or lateral pressure in the cable. Thus, irregularities associated with lateral pressure can be transmitted to the fiber despite the buffering protection by the primary layer.

4. Thermal bending, a precursor to buckling, would produce a steadily increasing added optical loss as temperature drops for strains below the buckling strain.

Other omissions include details of the viscoelastic nature of the coating materials, the thermal dependence of the expansion coefficients, the thermal cycling, and thermal gradients. The buckling calculations must be regarded as estimates most valuable in making comparisons.

V. ACKNOWLEDGMENTS

The author thanks C. J. Aloisio for initially describing the buckling problem and continued discussion, C. R. Taylor for modulus measurements and continued discussion, C. H. Gartside III for discussion of the S5 experiment, L. L. Blyler, Jr., for measurements of thermal expansion coefficients, and D. P. Woodard and J. T. Loadholt for discussions.

REFERENCES

1. W. B. Gardner, "Miocrobending Loss in Optical Fibers," B.S.T.J., *54,* No. 2 (February 1975), pp. 457–65.
2. D. Gloge, "Optical-Fiber Packaging and Its Influence on Fiber Straightness and Loss," B.S.T.J., *54,* No. 2 (February 1975), pp. 245–62.
3. L. L. Blyler et al., "A New Dual-Coating System for Optical Fibers," Eighth European Conf. Opt. Commun., Cannes, France, 1982.
4. L. L. Blyler, Jr., et al., "Buckling of Optical Fibers Within Elastomers Used in an Embedded-Core Cable Structure," Int. Wire Cable Symp. Proc. 1983, pp. 144–50.
5. Y. Katsuyama et al., "Transmission Loss of Coated Single-Mode Fiber at Low Temperatures," Appl. Opt., *19,* No. 24 (December 15, 1980), pp. 4200–5.
6. T. Yabuta et al., "Excess Loss of Single-Mode Jacketed Optical Fiber at Low Temperature," Appl. Opt., *22,* No. 15 (August 1, 1983), pp. 2356–62.
7. C. H. Gartside, unpublished work.
8. G. S. Brockway and M. R. Santana, "Analysis of Thermally Induced Loss in Fiber-Optic Ribbons," B.S.T.J., *62,* No. 4, Part 1 (April 1983), pp. 993–1018.
9. L. L. Blyler, Jr., unpublished work.
10. C. R. Taylor, private communication.
11. J. D. Ferry, *Viscoelastic Properties of Polymers*, Second Edition, New York: Wiley, 1970.
12. S. P. Timoshenko, *Theory of Elastic Stability*, Second Edition, New York: McGraw-Hill, 1961.
13. D. C. L. Vangheluwe, "Exact Calculations of the Spring Constant in the Buckling of Optical Fibers," Appl. Opt., *23,* No. 13 (July 1, 1984), pp. 2045–6.
14. C. D. Hodgman, Editor, *C.R.C. Standard Mathematical Tables*, Twelfth Edition, Cleveland: Chemical Rubber Publishing Co., 1959, p. 373.
15. I. S. Sokolnikoff, *Mathematical Theory of Elasticity*, Second Edition, New York: McGraw-Hill, 1956.

APPENDIX A

Calculation of Thermal Stress

This appendix gives a method for calculating the effective expansion coefficient of a dual-coated fiber. The calculation goes beyond the rule of mixtures by treating the coupling of the radial displacements of the various layers.

Define cylindrical coordinates (r, θ, z) based at the center of the coated fiber. The displacement in each of the three regions can be represented by

$$U = \begin{pmatrix} U_r \\ U_\theta \\ U_z \end{pmatrix} = \begin{pmatrix} \dfrac{a_n}{r} + b_n r \\ 0 \\ Cz \end{pmatrix} n = 1, 2, 3. \tag{23}$$

The fiber has $n = 1$, the primary coating $n = 2$, and the secondary $n = 3$. This vector function satisfies the Cauchy Navier equation[15] for displacements and represents a solution without warping (because C is independent of n) or angular deformation.

The strain components are

$$\epsilon_{rr} = \frac{\partial U_r}{\partial r} = -\frac{a_n}{r^2} + b_n$$

$$\epsilon_{\theta\theta} = \frac{1}{r}\frac{\partial U_\theta}{\partial \theta} + \frac{U_r}{r} = \frac{a_n}{r^2} + b_n$$

$$\epsilon_{\theta\theta} = \frac{\partial U_z}{\partial z} = C$$

$$\epsilon_{\theta z} = \epsilon_{\theta r} = \epsilon_{rz} = 0. \tag{24}$$

The first invariant of the strain tensor is

$$e = \epsilon_{rr} + \epsilon_{\theta\theta} + \epsilon_{zz} = 2b_n + C. \tag{25}$$

Stress components are obtained from the generalized Hooke's law,

$$\sigma_{ij} = \lambda e \delta_{ij} + 2G\epsilon_{ij} - \beta T \delta_{ij}, \tag{26}$$

which gives

$$\sigma_{rr} = \lambda e + 2G\epsilon_{rr} - \beta T$$

$$\sigma_{zz} = \lambda e + 2G\epsilon_{zz} - \beta T, \tag{27}$$

where

$$\beta = \frac{E}{1 - 2\nu} \quad G = \frac{E}{2(1 + \nu)} \quad \lambda = \frac{2G\nu}{1 - 2\nu}, \tag{28}$$

α is the thermal expansion coefficient, and T is the temperature change. The material parameters are assumed to depend on the region or layer, but they are assumed constant within each layer.

The unknown coefficients used to describe the displacement U are determined from the various conditions. The displacement must be bounded, so $a_1 = 0$; and it must be continuous, so

$$a_2 = (b_1 - b_2)r_1^2 \qquad \text{and} \qquad a_3 = a_2 + (b_2 - b_3)r_2^2, \tag{29}$$

where r_1, r_2, and r_3 denote the radius of the fiber, the primary, and the secondary, respectively. The radial stress component σ_{rr} must be continuous, so

$$\begin{aligned} b_1(2\lambda_1 + 2G_1) + C\lambda_1 - \beta_1 T & \\ = b_2(2\lambda_2 + 2G_2) - a_2 2G_2/r_1^2 + C\lambda_2 - \beta_2 T & \\ b_2(2\lambda_2 + 2G_2) - a_2 2G_2/r_2^2 + C\lambda_2 - \beta_2 T & \\ = b_3(2\lambda_3 + 2G_2) - a_3 2G_3/r_2^2 + C\lambda_3 - \beta_3 T & \\ b_3(2\lambda_3 + 2G_2) - a_3 2G_3/r_3^2 + C\lambda_3 - \beta_3 T = 0. & \qquad (30) \end{aligned}$$

Finally, the total longitudinal force or integrated stress must be 0; hence,

$$\frac{1}{2}\,[r_1^2\sigma_{zz}^{(1)} + (r_2^2 - r_1^2)\sigma_{zz}^{(2)} + (r_3^2 - r_2^2)\sigma_{zz}^{(3)}] = 0 \qquad (31)$$

or

$$\begin{aligned} r_1^2[b_1 2\lambda_1 + C(\lambda_1 + 2G_1) - \beta_1 T] & \\ + (r_2^2 - r_1^2)[b_2 2\lambda_2 + C(\lambda_2 + 2G_2) - \beta_2 T] & \\ + (r_3^2 - r_2^2)[b_2 2\lambda_3 + C(\lambda_3 + 2G_3) - \beta_3 T] = 0. & \qquad (32) \end{aligned}$$

As for the moments, they are automatically 0 when the coatings are concentric.

These six linear equations in the six unknowns (a_2, a_3, b_1, b_2, b_3, C) can be solved by standard numerical methods. The object of the calculation is the coefficient C, which denotes the thermal strain of the structure. When T is set to 1, then C is α_e, the effective expansion coefficient.

APPENDIX B

Calculation of Spring Constant

This appendix gives a method for calculating the spring constant κ of a fiber within a dual coating. The calculation goes beyond the formula of Vangheluwe[13] by accounting for the elasticity of the outer secondary coating.

Define cylindrical coordinates (r, θ, z) based at the fiber center. The displacement function

$$U = \begin{pmatrix} U_r \\ U_\theta \end{pmatrix}$$

must satisfy the Cauchy Navier equation[15]

$$\nabla^2 U + C\nabla(\nabla \cdot U) = 0 \qquad C = \frac{1}{1 - 2\nu}. \qquad (33)$$

The outer surface of the secondary is assumed fixed, and when the rigid fiber is translated δ in the x direction,

$$\delta a_x = \delta(\cos\theta\; a_r - \sin\theta\; a_\theta), \tag{34}$$

the displacement at $r = r_1$ for small δ is

$$U = \begin{pmatrix} \delta \cos \theta \\ -\delta \sin \theta \end{pmatrix}. \tag{35}$$

This means that the angular dependence, in general, is

$$U = \begin{pmatrix} U_r(r) \cos \theta \\ U_\theta(r) \sin \theta \end{pmatrix}. \tag{36}$$

The four solutions of the Cauchy Navier eq. (33) having this angular dependence are

$$\begin{pmatrix} U_r(r) \\ U_\theta(r) \end{pmatrix} = \begin{pmatrix} \ln r/r_1 - \dfrac{C}{2 + C} \\ - \ln r/r_1 \end{pmatrix}, \begin{pmatrix} (2 - C)r^2 \\ (2 + 3C)r^2 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} r^{-2} \\ r^{-2} \end{pmatrix}, \tag{37}$$

as can be checked by direct substitution. The general solution is a linear combination of these four with four unknown coefficients for each of the two coating layers.

The total of eight unknown coefficients are determined by eight conditions. At r_1,

$$\begin{pmatrix} U_r \\ U_\theta \end{pmatrix} = \delta \begin{pmatrix} 1 \\ -1 \end{pmatrix};$$

at r_3,

$$\begin{pmatrix} U_r \\ U_\theta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix};$$

and at r_2,

$$\begin{pmatrix} U_r \\ U_\theta \end{pmatrix}$$

are continuous across the interface. These give six equations. The other two conditions involve the normal stress components σ_{rr} and $\tau_r\theta$, which must be continuous at r_2. In terms of U these are

$$\sigma_{rr} = \lambda \nabla \cdot U + 2\mu \frac{\partial U_r}{\partial r} = (\lambda + 2\mu) \frac{\partial U_r}{\partial r} + \lambda \frac{U_r + U_\theta}{r}$$

$$\tau_r\theta = \mu \frac{\partial U_\theta}{\partial r} - \mu \frac{U_r + U_\theta}{r}, \tag{38}$$

where $\lambda = 2\mu\nu C$ and the angular dependence of eq. (36) has been used. The eight linear equations in eight unknowns can be solved by standard numerical methods.

The force on the fiber for the deflection δ is

$$F = \int_{\text{Fiber}} \sigma \cdot a_x ds = \int_0^{2\pi} (\sigma_r \cos^2\theta - \tau_{r\theta} \sin^2\theta) r_1 d\theta$$

$$= \pi r_1 (\sigma_r(r_1) - \tau_{r\theta}(r_1)). \tag{39}$$

The stress quantity

$$\sigma_r - \tau_{r\theta} = (\lambda + 2\mu) \frac{\partial U_r}{\partial r} - \mu \frac{\partial U_\theta}{\partial r} + (\lambda + \mu) \frac{U_r + U_\theta}{r} \tag{40}$$

equals

$$\left[(\lambda + \mu) \frac{2}{2 + C} + 2\mu \right] 1/r$$

for the first solution in eq. (37) and is identically 0 for the other three. Therefore, when δ is set to 1,

$$\kappa = \pi \left[(\lambda_p + \mu_p) \frac{2}{2 + C_p} + 2\mu_p \right] a_1, \tag{41}$$

where p signifies the primary region and a_1 denotes the coefficient of the first solution,

$$\begin{pmatrix} \ln r/r_1 - \dfrac{C}{2 + C} \\ - \ln r/r_1 \end{pmatrix}$$

in eq. (37), in the primary region. Thus, of the eight unknown coefficients, only one is needed for getting the spring constant.

AUTHOR

Terrence A. Lenahan, S.B. and S.M. (Electrical Engineering), 1964, The Massachusetts Institute of Technology; Ph.D. (Applied Mathematics), 1970, University of Pennsylvania; AT&T Bell Laboratories, 1970—. Mr. Lenahan has done studies in various areas of mathematical physics, including electromagnetic propagation, elasticity, and fluid mechanics. He recently has been interested in mechanical and optical aspects of optical fibers. Member, AAS, AMS, SIAM, AAAS.

AT&T Technical Journal
Vol. 64, No. 7, September 1985
Printed in U.S.A.

Analysis of Laser Beam Propagation in a Turbulent Atmosphere

By R. H. CLARKE*

(Manuscript received January 30, 1985)

The beam propagation method, based on the parabolic approximation to the wave equation, is used in conjunction with Papoulis' redefinition for optical fields of Woodward's ambiguity function. A simple derivation is given of Tatarskii's formula for the lateral coherence function, and hence the mean intensity profile, of a laser beam propagating through a turbulent atmosphere. Statistics of the received signal and the effects of spatial nonstationarity of the turbulence can also be deduced using this technique, as can the effects of very large-scale variations in refractive index and receiver directivity.

I. INTRODUCTION

There has been a recent revival of interest in the propagation of laser beams through the atmosphere for communication purposes. King et al.[1] have conducted experiments that show that a laser is an effective standby substitute for a microwave link over a clear-air, line-of-sight path of several tens of kilometers. When the microwave link is subject to severe multipath fading, the laser signal is found to be much more stable. In these clear-air conditions the laser beam is mainly affected by the atmosphere's turbulence, which produces a spread of the propagating beam in excess of that expected due to diffraction. Over a 37-km path the lateral intensity profile of the laser beam is found to be random but to have an average Gaussian shape with a spread of about 6m between e^{-1} points.

More than two decades of research on the theory of optical propa-

* AT&T Bell Laboratories and Imperial College of Science and Technology, London.

gation through random media has been very competently reviewed by Strohbehn and others;[2] the chapter by Ishimaru[3] is particularly relevant here. The wide variety of approaches taken by different authors is apparent, ranging from the purely physical to the highly mathematical. The present paper seeks to produce a reasonably simple theoretical picture, which is accurate both physically and mathematically, and which will also be useful for engineering purposes.

The starting point, as so often elsewhere, is the parabolic approximation to the wave equation,[4] but it is used here in a manner that has become known as the "beam propagation method."[5] Other names for the method are the "split-step Fourier technique" of Tappert and Hardin[6] and the "multiple random phase-screen method."[7,8] The essential idea that makes a simple solution possible is that, because the fluctuations in refractive index are so weak and their scale size is so large compared to the wavelength, the phenomena of diffraction and scattering can be artificially separated. The propagation path is divided into many short sections, so that the propagating wave is barely disturbed by each section, but their cumulative effect can be considerable.

In each of these sections the irregularities are effectively removed, in the form of an accumulated random phase, to one or another of the boundary planes. Free-space diffraction is then allowed to occur within the now uniform section between the planes. The essential next step is that the Fresnel diffraction, which occurs between the planes in each of the sections, is described by what Papoulis[9] has called the "ambiguity function," after the name given by Woodward[10] to a similar function of fundamental importance in radar. For an optical field the ambiguity function was redefined by Papoulis as the Fourier transform of the lateral mutual coherence function (i.e., the lateral autocorrelation function of the field). The "field ambiguity function" so defined has the very useful property that it propagates in a uniform medium without changing its functional form. What does change is the argument, in a manner reminiscent of a wave traveling along a transmission line.

On encountering the artificially accumulated random phase at each boundary plane, the field ambiguity function is modified appropriately but then propagates through the next section again without change. This procedure can continue as long as the propagating beam is essentially forward scattered, which is true for a laser beam propagating through atmospheric turbulence at least out to 50 km, if not farther. Then over the final plane the lateral field autocorrelation function is obtained by taking the Fourier transform of the field ambiguity function. The mean intensity of the beam is contained in the field autocorrelation function as a special case.

It is gratifying that application of this simple procedure reproduces precisely what Ishimaru[3] has described as "Tatarskii's exact result"[4] for the lateral mutual coherence function of a laser beam propagating through atmospheric turbulence. In fact, because of its underlying physical clarity, the present procedure allows one to go a little farther than Tatarskii and describe the statistics of the propagating field in more detail, and also to deal with spatially nonstationary turbulence.

II. ELECTROMAGNETIC BASIS: THE BEAM PROPAGATION METHOD

In attempting to solve propagation problems in which the scale size of the refractive-index irregularities is large compared with the wavelength, and the magnitude of these fluctuations is very small, it is often helpful to factor out the term $\exp(-jkz)$, assuming propagation in the general direction of the z axis. This procedure is analogous to factoring out the time dependence $\exp(j\omega t)$. Thus, any phasor component $f(\boldsymbol{r}, t)$ of the quasi-monochromatic propagating field can be written as

$$f(\boldsymbol{r}, t) = u(\boldsymbol{r}, t)\exp(-jkz), \tag{1}$$

where $u(\boldsymbol{r}, t)$ is a slowly varying phasor function of position $\boldsymbol{r}$ and time t. The propagation constant k is some convenient mean value.

As a consequence of the assumed large scale size and tenuous nature of the refractive index irregularities, it can be shown[4,11] that the function $u(\boldsymbol{r}, t)$, is governed by the parabolic equation approximation to the wave equation:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - j2k\frac{\partial u}{\partial z} + 2k^2 n_1 u = 0, \tag{2}$$

where $n_1(\boldsymbol{r}, t)$ is the departure of the refractive index from its mean value, which will be assumed to be unity. Time fluctuations will be ignored in what follows, although they can be incorporated easily if required.

Consider the time-invariant solution of eq. (2) for a wave launched from the plane $z = 0$. A well-established approach[5,6,11] is to solve the equation in two iterative steps, and will be referred to here as the beam propagation method. First assume that there are no variations in refractive index, so that $n_1 = 0$. Then the solution for the field over any plane z is given by Fresnel's diffraction formula[12]

$$f(x, y, z) = j\,\frac{\exp(-jkz)}{\lambda z}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x', y', 0) \cdot\exp\left\{-\frac{jk}{2z}\left[(x - x')^2 + (y - y')^2\right]\right\} dx'dy' \tag{3}$$

in terms of the field $f(x, y, 0)$ over the aperture plane $z = 0$.

If diffraction now is ignored, by suppressing the first two terms in eq. (2), and if the irregularities in refractive index are restored, their first-order effect can be obtained from the solution of

$$\frac{\partial u}{\partial z} + jkn_1 u = 0, \tag{4}$$

which is, by straightforward integration,

$$u(x, y, z) = u(x, y, 0)\exp\{j\Phi(x, y)\}, \tag{5}$$

where

$$\Phi(x, y) = -k \int_0^z n_1(x, y, z)dz. \tag{6}$$

This is simply the phase induced by the refractive-index irregularities along straight-line paths, parallel to the z-axis, from 0 to z.

Thus the solution of eq. (2) is in the two artificially separated parts given by eqs. (3) and (5). The first part allows for diffraction but suppresses the effect of the irregularities, while the second part suppresses diffraction but allows for the effect of the irregularities. The two parts of the solution then must be combined in some suitable way, as illustrated in the following examples. The first is concerned with the nonrandom effect of the overall linear trend of refractive index with height, and the second considers the effect of the turbulence-induced, small-scale random irregularities in refractive index.

2.1 Effect of linear gradient in refractive index

If the refractive index varies linearly with height, with constant gradient g, then over a distance of Δz eq. (6) gives the phase variation with height x as linear also, namely as $\Phi(x) = -kgx\Delta z$. The mean linear trend in the atmosphere is usually negative, and so it can be seen that the effect of this negative gradient in refractive index will be to tilt the advancing wavefront forward through an angle $g\Delta z$. If this continous forward tilt is interpreted as bending the propagating beam, giving it a radius of curvature R, then the angle of tilt would be $\Delta z/R$. Hence $R = g^{-1}$, a well-known result.[13] But it should be emphasized that while the beam is being bent it is also experiencing diffraction, according to eq. (3), and so spreads as it bends as it propagates.

2.2 Effect of turbulent fine structure

Having seen that the beam-bending effect of the mean linear trend in refractive index can be treated separately, consider now a medium that is on average uniform but whose refractive index $n_1(\boldsymbol{r})$ is a zero-mean random process. The magnitude of these fluctuations in refractive index is typically 10^{-8}, for homogeneous turbulence conditions, with scale sizes of at least several millimeters, which is large compared to optical wavelengths. Hence the conditions for the beam propagation method to apply are fulfilled.

Consider a segment of the medium between the beam-launch plane $z = 0$ and the plane $z = \Delta z$. If the refractive-index irregularities are temporarily ignored, the field over the exit plane $z = \Delta z$ would be given by Fresnel's diffraction formula of eq. (3). Now restoring the irregularities but temporarily suppressing diffraction, their effect is accounted for in the accumulated random phase along parallel ray paths of length Δz:

$$\Phi(x, y) = -k \int_0^{\Delta z} n_1(x, y, z)dz. \tag{7}$$

The statistics of this random phase process will be important later and so are derived here.

Since the refractive-index fluctuation process $n_1(x, y, z)$ is zero mean,

$$\langle \Phi(x, y) \rangle = 0, \tag{8}$$

where the sharp brackets indicate taking the expectation.

If n_1 is wide-sense stationary, with autocovariance

$$B_{n_1}(\xi, \eta, \zeta) = \langle n_1(x, y, z)n_1(x + \xi, y + \eta, z + \zeta)\rangle \tag{9}$$

and variance

$$\sigma_{n_1}^2 = B_{n_1}(0, 0, 0), \tag{10}$$

then the autocovariance of the phase process

$$B_\Phi(\xi, \eta) = k^2 \int_0^{\Delta z} \int_0^{\Delta z} \langle n_1(x, y, z)n_1(x + \xi, y + \eta, z')\rangle dzdz'. \tag{11}$$

This is equivalent to[14]

$$B_\Phi(\xi, \eta) = k^2 \Delta z \int_{-\Delta z}^{\Delta z} B_{n_1}(\xi, \eta, \zeta)d\zeta. \tag{12}$$

Now it is convenient to assume that the width of the section $\Delta z \gg \zeta_0$, the scale size of the irregularities in the z direction, and so

$$B_\Phi(\xi, \eta) = k^2 \Delta z \int_{-\infty}^{\infty} B_{n_1}(\xi, \eta, \zeta)d\zeta. \tag{13}$$

It also follows from the condition $\Delta z \gg \zeta_0$ that the phase over the exit plane can always be taken to be Gaussian, whether the refractive index itself is Gaussian or not, as a consequence of the central limit theorem. Hence, the random phase process is completely described by the autocovariance of eq. (13), and its variance can be taken to be of the order

$$\sigma_\Phi^2 \sim k^2 \sigma_{n_1}^2 \zeta_0 \Delta z, \tag{14}$$

in which ζ_0 is sometimes referred to as the integral scale size of the refractive index in the direction of propagation.

Since the beam propagation method is, in essence, a perturbation technique, applied locally, it is important that over each section the phase variance $\sigma_\Phi^2 \ll 1$. In applications such as laser beam propagation through atmospheric turbulence, this condition is easily met, even with the constraint that $\Delta z \gg \zeta_0$.

III. STATISTICAL FIELDS: THE AMBIGUITY FUNCTION

The example of Section 2.2 will now be our main concern. So far we have the field distribution over the plane $z = \Delta z$ as $f_0(x, y, \Delta z) \exp\{j\Phi(x, y)\}$, where $f_0(x, y, \Delta z)$ is the free-space diffraction of the original aperture field and $\Phi(x, y)$ is the random phase process of eq. (7). If the field over the plane Δz is now allowed to diffract, Fresnel's diffraction formula would give the field over the next plane, assuming that the intervening region is free space. The irregularities could then be replaced and a different random phase process could account for them over this next plane. This new field then can be allowed to further diffract, and so on.

Over some plane z, well into the random medium, the phasor field component $f(x, y, z)$ will itself be random. The property of greatest utility would be its lateral mutual coherence function:

$$\Gamma(x, y; \xi, \eta; z) = \langle f^*(x - \xi/2, y - \eta/2, z) f(x + \xi/2, y + \eta/2, z) \rangle, \tag{15}$$

otherwise referred to as the lateral field autocorrelation function. (The asterisk denotes complex conjugate.) Hence the mean intensity

$$\langle I(x, y, z) \rangle = \langle |f(x, y, z)|^2 \rangle = \Gamma(x, y; 0, 0; z). \tag{16}$$

However, $\Gamma(\ )$ is not simple to calculate,[2] whereas its Fourier transform is. Papoulis[9] introduced the Fourier transform of $\Gamma(\ )$, calling it the ambiguity function of the optical field, and showed that it greatly simplified the calculation of Fresnel diffracted fields. In particular, a very useful property of the ambiguity function is that it propagates without changing its functional form in a uniform medium; what does change is the argument of the function.

The definition of ambiguity function that will be used here, for the field over the plane z, is

$$A(\mu, \nu; \xi, \eta; z)$$

$$= \frac{1}{\lambda^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Gamma(x, y; \xi, \eta; z) \exp\{jk(\mu x + \nu y)\} dx dy. \quad (17)$$

If the medium between this plane z and the plane $z + \Delta z$ is uniform, it can be shown that, under conditions when Fresnel diffraction occurs, the ambiguity function over the plane $z + \Delta z$ is simply[9]

$$A(\mu, \nu; \xi, \eta; z + \Delta z) = A(\mu, \nu; \xi - \mu \Delta z, \eta - \nu \Delta z; z). \quad (18)$$

The way in which the ambiguity function and this relation are used is described in the next two sections.

IV. BEAM WAVE PROPAGATION THROUGH TURBULENCE

Consider a beam of radiation launched into a turbulent medium in which the conditions for the application of the beam propagation method are satisfied, namely that the magnitude of the fluctuations in refractive index is very small and their scale size is very large in comparison with the wavelength of the propagating beam. If the field is launched from an aperture plane at $z = 0$, over which it is $f(x, y, 0)$, the ambiguity function over the $z = 0$ plane is given by eqs. (15) and (17) as

$$A(\mu, \nu; \xi, \eta; 0) = A_0(\mu, \nu; \xi, \eta) \quad (19)$$

$$= \frac{1}{\lambda^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f^*(x - \xi/2, y - \eta/2, 0)$$

$$\cdot f(x + \xi/2, y + \eta/2, 0) \cdot \exp\{jk(\mu x + \nu y)\} dx dy. \quad (20)$$

If the propagation path is divided into short sections of width Δz_i, $i = 1, 2, \cdots, N$, and if it is assumed that over each section the medium is uniform, with the effect of the irregularities swept forward onto the exit face, then the ambiguity function just to the left of the plane $z = \Delta z_1$ is

$$A(\mu, \nu; \xi, \eta; \Delta z_1^-) = A_0(\mu, \nu; \xi - \mu \Delta z_1, \eta - \nu \Delta z_1) \quad (21)$$

with eq. (20) substituted.

The accumulated phase can be inserted at this point by writing

$$f(x, y, \Delta z_1^+) = f(x, y, \Delta z_1^-) \exp\{j \Phi(x, y)\}, \quad (22)$$

where the phase $\Phi(x, y)$ is given by eq. (7) and has the statistical properties derived in Section 2.2. Now assuming that the $\Phi(x, y)$ process is stationary, over the lateral extent of the beam within the first section, it follows that the ambiguity function at the exit face of the section is

$$A(\mu, \nu; \xi, \eta; \Delta z_1^\dagger) = A_0(\mu, \nu; \xi - \mu\Delta z_1, \eta - \nu\Delta z_1)\exp\{B_{\Phi_1}(\xi, \eta) - B_{\Phi_1}(0, 0)\}, \quad (23)$$

where the phase autocovariance function $B_{\Phi_1}(\xi, \eta)$ is given by eq. (13), and the argument of the exponential is sometimes referred to as the phase structure function. So

$$B_{\Phi_1}(\xi, \eta) = \sigma_{\Phi_1}^2 \rho(\xi, \eta) \quad (24)$$

with the phase variance over this first section given by

$$\sigma_{\Phi_1}^2 = k^2 \sigma_{n_1}^2 \zeta_0 \Delta z_1, \quad (25)$$

where $\sigma_{n_1}^2$ is the variance of the refractive index fluctuations, ζ_0 is its scale size in the direction of propagation, and $\rho(\ )$ is the normalized phase autocovariance function. Thus eq. (23) is equivalently

$$A(\mu, \nu; \xi, \eta; \Delta z_1^\dagger) = A_0(\mu, \nu; \xi - \mu\Delta z_1, \eta - \nu\Delta z_1)\exp\{-\sigma_{\Phi_1}^2[1 - \rho(\xi, \eta)]\}. \quad (26)$$

It will be recalled from Section 2.2 that while $\Delta z_1 \gg \zeta_0$, it is to be kept small enough for $\sigma_{\Phi_1}^2 \ll 1$.

In the next section, of width Δz_2, again artificially separating the phenomena of scattering and diffraction, by analogy with eq. (26)

$$A(\mu, \nu; \xi, \eta; \Delta z_1 + \Delta z_2^\dagger) = A(\mu, \nu; \xi - \mu\Delta z_2, \eta - \nu\Delta z_2; \Delta z_1^\dagger)\exp\{-\sigma_{\Phi_2}^2[1 - \rho(\xi, \eta)]\}, \quad (27)$$

in which it has been assumed that the turbulence is statistically uniform along the beam. (This condition can be relaxed, and clearly ought to be in some circumstances, but at the expense of greater complication.) Combining eqs. (26) and (27) gives

$$\begin{aligned} A(\mu, \nu; \xi, \eta; \Delta z_1 + \Delta z_2^\dagger) = A_0(\mu, \nu; \xi - \mu[\Delta z_1 + \Delta z_2],\\ \eta - \nu[\Delta z_1 + \Delta z_2])\exp\{-\sigma_{\Phi_2}^2[1 - \rho(\xi, \eta)]\}\\ \cdot\exp\{-\sigma_{\Phi_1}^2[1 - \rho(\xi - \mu\Delta z_2, \eta - \nu\Delta z_2)]\}. \quad (28) \end{aligned}$$

This argument can be continued out to a distance

$$z = \sum_{i=1}^{N} \Delta z_i \quad (29)$$

provided only that the propagation is essentially in the forward direction. This is ensured by the large scale size of the turbulence compared to the wavelength, and the small magnitude of the refractive index fluctuations. Then the ambiguity function at the plane z is

$$A(\mu, \nu; \xi, \eta; z) = A_0(\mu, \nu; \xi - \mu z, \eta - \nu z) \cdot \exp\left\{ - \sum_{n=1}^{N} \sigma_{\Phi_n}^2 \left[1 - \rho \left(\xi - \mu \left\{ z - \sum_{i=1}^{n} \Delta z_i \right\}, \eta - \nu \left\{ z - \sum_{i=1}^{n} \Delta z_i \right\} \right) \right] \right\}, \quad (30)$$

in which

$$\sigma_{\Phi_n}^2 = k^2 \sigma_{n_1}^2 \zeta_0 \Delta z_n \quad (31)$$

and with eq. (20) substituted.

The ambiguity function of eq. (30) can be written in integral form if the Δz_n can be taken to be sufficiently small. In the case of a He-Ne laser beam propagating through a turbulent atmosphere, Δz_n can be of the order of a meter, which is small enough when it is realized that both μ and ν are always small. The assumption of the statistical uniformity of the turbulence throughout the length of the path is still maintained, and so

$$A(\mu, \nu; \xi, \eta; z) = A_0(\mu, \nu; \xi - \mu z, \eta - \nu z) \cdot \exp\left\{ - k^2 \sigma_{n_1}^2 \zeta_0 \left[z - \int_0^z \rho(\xi - \mu z', \eta - \nu z') dz' \right] \right\}. \quad (32)$$

Finally, the lateral mutual coherence function is given by the inverse Fourier transform of eq. (17) as

$$\Gamma(x, y; \xi, \eta; z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(\mu, \nu; \xi, \eta; z) \exp\{- jk(\mu x + \nu y)\} d\mu d\nu \quad (33)$$

with eq. (32) substituted. This result is identical in form to Tatarskii's exact solution of the differential equation for the mutual coherence function, quoted by Ishimaru.[3] This agreement is encouraging, in that Tatarskii's approach[4] is basically more rigorous but is physically somewhat obscure, whereas the present method keeps the physical meaning of the mathematics well to the fore. As an example, the method given here avoids Tatarskii's assumption that the refractive index is delta-function correlated in the direction of propagation, which is a physical impossibility. In fact it seems that this so-called Markov approximation means physically that the induced phase processes in each section are statistically independent. Also, the summation form of the ambiguity function of eq. (30) could be advantageous in applications to paths along which the turbulence is nonstationary.

Incidentally, the mean field obtained from eq. (22), by taking its expectation, is

$$\langle f(x, y, \Delta z_1)\rangle = f_0(x, y, \Delta z_1)\exp\left\{-\frac{1}{2}\sigma_{\Phi_1}^2\right\}, \tag{34}$$

where $f_0(\ )$ is the field diffracted from the original aperture field in the absence of irregularities. Carrying out the same procedure over the N sections of the path gives

$$\langle f(x, y, z)\rangle = f_0(x, y, z)\exp\left\{-\frac{1}{2}k^2\sigma_{n_1}^2\zeta_0 z\right\}, \tag{35}$$

which is also referred to as the coherent part of the field. Note that when the total accumulated phase variance

$$\sigma_{\Phi_T}^2 = k^2\sigma_{n_1}^2\zeta_0 z \tag{36}$$

becomes much larger than unity, the coherent field becomes negligible.

Equations (35) and (33) give the first and second statistical moments, respectively, of the propagating field. The physical mechanism also strongly suggests, as a consequence of the central limit theorem, that the field is complex-Gaussian distributed. The statistical description of the random propagating field is therefore complete.

V. PROPAGATION OF A LASER BEAM THROUGH ATMOSPHERIC TURBULENCE

To find the mean intensity and other characteristics of a laser beam propagating through a turbulent atmosphere, it will be assumed that the beam is launched in its fundamental mode with a plane wave front, that is,

$$f(x, y, 0) = f_0 \exp\left\{-\frac{x^2 + y^2}{w_0^2}\right\}, \tag{37}$$

where w_0 is the beamwaist parameter and f_0 is the complex amplitude at the center of the beam. The ambiguity function for this field is, from eq. (20),

$$A_0(\mu, \nu; \xi, \eta) = \frac{\pi w_0^2|f_0|^2}{2\lambda^2}\exp\left\{-\frac{\pi^2 w_0^2}{2\lambda^2}(\mu^2 + \nu^2)\right\} \cdot\exp\left\{-\frac{\xi^2 + \eta^2}{2w_0^2}\right\}. \tag{38}$$

Equations (32) and (33) then yield the lateral mutual coherence function in this case as

$$\Gamma(x, y; \xi, \eta; z) = \frac{\pi w_0^2 |f_0|^2}{2\lambda^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left\{-\frac{\pi^2 w_0^2}{2\lambda^2}(\mu^2 + \nu^2)\right\}$$

$$\cdot \exp\left\{-\frac{(\xi - \mu z)^2 + (\eta - \nu z)^2}{2w_0^2}\right\}$$

$$\cdot \exp\left\{-k^2 \sigma_{n_1}^2 \zeta_0 \left[z - \int_0^z \rho(\xi - \mu z', \eta - \nu z')dz'\right]\right\}$$

$$\cdot \exp\{-jk(\mu x + \nu y)\} d\mu d\nu, \quad (39)$$

and the mean intensity is obtained from this by setting ξ and η equal to 0.

It is easily verified as a check, using a standard integral, that eq. (39) in the absence of any irregularities in refractive index (i.e., $\sigma_{n_1} = 0$) gives the correct intensity formula for free-space propagation of a laser beam,[15] namely,

$$I_0(x, y, z) = \frac{|f_0|^2 w_0^2}{w^2(z)} \exp\left\{-\frac{2(x^2 + y^2)}{w^2(z)}\right\}, \quad (40)$$

where the beamwaist parameter

$$w(z) = w_0 \sqrt{1 + \left(\frac{\lambda z}{\pi w_0^2}\right)^2} \quad (41)$$

depends on distance. The particular distance

$$z_F = \frac{\pi w_0^2}{\lambda} \quad (42)$$

usefully indicates the transition from the near field ($z \ll z_F$), when the beam is essentially collimated, to the far field ($z \gg z_F$), when the beam spreads out linearly with distance. In the experiment of King et al.,[1] for example, with the beamwaist at launch $w_0 = 8$ cm and the wavelength 0.63 μm, the distance $z_F = 32$ km.

To obtain some analytical results for the mean intensity, it will be necessary to resort to some approximation. One way of doing this is to replace the variable z' in the inner integral of eq. (39) by the constant z. Then the mean intensity, using eq. (16), is given approximately by

$$\langle I(x, y, z)\rangle = \frac{\pi w_0^2 |f_0|^2}{2\lambda^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left\{-\frac{\pi^2 w_0^2}{2\lambda^2}(\mu^2 + \nu^2)\right\}$$

$$\cdot \exp\left\{-z^2 \frac{\mu^2 + \nu^2}{2w_0^2}\right\} \exp\{-\sigma_{\Phi_T}^2[1 - \rho(-\mu z, -\nu z)]\}$$

$$\cdot \exp\{-jk(\mu x + \nu y)\} d\mu d\nu, \quad (43)$$

in which $\sigma_{\Phi_T}^2$ [see eq. (36)] is the total accumulated phase variance over the length of the path. Equation (43) can also be written equivalently, by substituting

$$\mu z = -\xi \quad \text{and} \quad \nu z = -\eta, \tag{44}$$

so that

$$\langle I(x, y, z)\rangle = \frac{\pi w_0^2 |f_0|^2}{2\lambda^2 z^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left\{-\left(\frac{\pi^2 w_0^2}{2\lambda^2 z^2} + \frac{1}{2w_0^2}\right)(\xi^2 + \eta^2)\right\}$$
$$\cdot \exp\{-\sigma_{\Phi_T}^2[1 - \rho(\xi, \eta)]\} \exp\left\{j\,\frac{k(\xi x + \eta y)}{z}\right\} d\xi d\eta. \tag{45}$$

If now the middle exponential in eq. (45) is written as the sum of two parts, as

$$\exp\{-\sigma_{\Phi_T}^2[1 - \rho(\xi, \eta)]\}$$
$$= \exp\{-\sigma_{\Phi_T}^2\} + \exp\{-\sigma_{\Phi_T}^2\}[\exp\{\sigma_{\Phi_T}^2 \rho(\xi, \eta)\} - 1], \tag{46}$$

then the mean intensity formula of eq. (45) conveniently splits into the sum of the coherent intensity and the incoherently scattered intensity, namely,

$$\langle I(x, y, z)\rangle = I_0(x, y, z)\exp\{-\sigma_{\Phi_T}^2\} + I_s(x, y, z), \tag{47}$$

in which $I_0(x, y, z)$ is given by eq. (40) and

$$I_s(x, y, z) = \frac{\pi w_0^2 |f_0|^2}{2\lambda^2 z^2} \exp\{-\sigma_{\Phi_T}^2\}$$
$$\cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left\{-\left(\frac{\pi^2 w_0^2}{2\lambda^2 z^2} + \frac{1}{2w_0^2}\right)(\xi^2 + \eta^2)\right\}$$
$$\cdot [\exp\{\sigma_{\Phi_T}^2 \rho(\xi, \eta)\} - 1] \exp\left\{j\,\frac{k(\xi x + \eta y)}{z}\right\} d\xi d\eta. \tag{48}$$

It is interesting to examine the two extreme cases of $\sigma_{\Phi_T}^2 \ll 1$ and $\sigma_{\Phi_T}^2 \gg 1$, which correspond respectively to short and long propagation paths.

5.1 Short paths

If, according to eq. (36), the path length z is short enough to make $\sigma_{\Phi_T}^2 \ll 1$, then eq. (47) shows that the coherent part will predominate. Appropriate approximations in eq. (48) give the incoherent scattered power in this case as

$$I_s(x, y, z) = \frac{\pi w_0^2 |f_0|^2}{2\lambda^2 z^2} \sigma_{\Phi_T}^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \rho(\xi, \eta)$$
$$\cdot \exp\left\{-\left(\frac{\pi^2 w_0^2}{2\lambda^2 z^2} + \frac{1}{2w_0^2}\right)(\xi^2 + \eta^2)\right\} \exp\left\{j\,\frac{k(\xi x + \eta y)}{z}\right\} d\xi d\eta. \tag{49}$$

If it is assumed that the initial laser beam width $w_0 \gg a$, where a is the typical scale size of the turbulence (of the order of 1 cm in the atmosphere), then the first exponential term in eq. (48) can be ignored and the mean intensity profile of the scattered field is the Fourier transform of $\rho(\xi, \eta)$.

Thus, over short paths the laser beam will be only slightly diminished in comparison to the free-space situation, and the energy lost will be scattered. When $w_0 \gg a$ there will be an intense central spot surrounded by a faint halo of scattered light. The form of the scattered intensity profile will be determined by the lateral correlation of the turbulent refractive-index fluctuations. On the other hand, if $w_0 \ll a$ the laser beam will snake its way through the turbulence, preserving its original profile but continually changing its direction in a random manner.

5.2 *Long paths*

If the path length is long enough to make $\sigma^2_{\Phi_T} \gg 1$, then according to eq. (47) the coherent part will be negligible, and so eq. (45) can be used directly to describe the now completely scattered field. Examining the middle exponential term in the integrand of eq. (45) reveals that for very large $\sigma^2_{\Phi_T}$ its behavior will be dominated by the behavior of $\rho(\xi, \eta)$ near the origin.[16] For turbulence

$$\rho(\xi, \eta) = 1 - \frac{\xi^2 + \eta^2}{a^2} + \cdots, \tag{50}$$

where a is now to be interpreted as the dissipation scale size of the assumed uniform and isotropic turbulence. (This assumption may not be true of course, but any naturally occurring refractive-index fluctuations will have an autocorrelation function that behaves parabolically in the neighborhood of the origin.[17]) Substituting eq. (50) into eq. (45) gives the mean intensity profile as

$$\langle I(x, y, z)\rangle = \frac{\pi w_0^2 |f_0|^2}{2\lambda^2 z^2} \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left\{-\left(\frac{\pi^2 w_0^2}{2\lambda^2 z^2} + \frac{1}{2w_0^2} + \frac{\sigma^2_{\Phi_T}}{a^2}\right)(\xi^2 + \eta^2)\right\} \cdot \exp\left\{j\,\frac{k(\xi x + \eta y)}{z}\right\} d\xi d\eta. \tag{51}$$

So, if $w_0 \gg a$, the $\sigma^2_{\Phi_T}/a^2$ term dominates, and a standard integral yields

$$\langle I(x, y, z)\rangle = \frac{|f_0|^2 w_0^2}{w^2(z)} \exp\left\{-\frac{2(x^2 + y^2)}{w^2(z)}\right\}, \tag{52}$$

where now

$$w(z) = \frac{\sqrt{2}\lambda\sigma_{\Phi_T} z}{\pi a}. \tag{53}$$

Equation (52) gives the mean intensity profile as being Gaussian in shape, as observed in the experiment of King et al.[1] The beamwaist parameter depends on $z^{3/2}$, since σ_{Φ_T} varies as $\sqrt{z}$. Some care must be taken with the longitudinal integral scale ζ_0 of the turbulence, which is needed to evaluate σ_{Φ_T} (see eq. 36). It is tempting to identify it with the dissipation scale size a, used in eq. (50), but that is probably an underestimate. On the other hand, Ishimaru's identification of it with the outer scale of turbulence[18] seems like an overestimate. The cautionary remark at the end of Section 5.3 should also be noted.

5.3 Lateral field autocorrelation

The autocorrelation of the propagating field over a plane is given by the lateral mutual coherence function of eq. (39). When the same approximation as for the intensity (eq. 43) is made, namely, replacing z' in the inner integral of eq. (39) by z, the lateral mutual coherence function becomes

$$\Gamma(x, y; \xi, \eta; z) = \frac{\pi w_0^2 |f_0|^2}{2\lambda^2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-\frac{\pi^2 w_0^2}{\lambda^2}(\mu^2 + \nu^2)\right\}$$
$$\cdot \exp\left\{-\frac{(\xi - \mu z)^2 + (\eta - \nu z)^2}{2w_0^2}\right\}$$
$$\cdot \exp\left\{-\sigma_{\Phi_T}^2[1 - \rho(\xi - \mu z, \eta - \nu z)]\right\}\exp\{-jk(\mu x + \nu y)\}d\mu d\nu. \tag{54}$$

Making the substitutions $p = \xi - \mu z$, $q = \eta - \nu z$ gives

$$\Gamma(x, y; \xi, \eta; z) = \frac{\pi w_0^2 |f_0|^2}{2\lambda^2 z^2}$$
$$\cdot \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-\frac{\pi^2 w_0^2}{2\lambda^2 z^2}[(\xi - p)^2 + (\eta - q)^2]\right\}$$
$$\cdot \exp\left\{-\frac{p^2 + q^2}{2w_0^2}\right\} \exp\left\{-\sigma_{\Phi_T}^2[1 - \rho(p, q)]\right\}$$
$$\cdot \exp\left\{-j\frac{k(\xi x + \eta y)}{z}\right\} \exp\left\{j\frac{k(px + qy)}{z}\right\} dpdq. \tag{55}$$

In the long-path limit, when $\sigma_{\Phi_T}^2 \gg 1$, and making use of eq. (50), the lateral mutual coherence function now becomes

$$\Gamma(x, y; \xi, \eta; z) = \frac{\pi w_0^2 |f_0|^2}{2\lambda^2 z^2} \exp\left\{-j\,\frac{k(\xi x + \eta y)}{z}\right\}$$

$$\cdot \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-\frac{\pi^2 w_0^2}{2\lambda^2 z^2}[(\xi - p)^2 + (\eta - q)^2]\right\}$$

$$\cdot \exp\left\{-\left(\frac{1}{2w_0^2} + \frac{\sigma_{\Phi_T}^2}{a^2}\right)(p^2 + q^2)\right\} \exp\left\{j\,\frac{k(px + qy)}{z}\right\} dpdq. \quad (56)$$

Performing the integration, and expressing the result in terms of the long-path mean intensity of eq. (51), gives

$$\Gamma(x, y; \xi, \eta; z) = \exp\left\{-j\,\frac{k(\xi x + \eta y)}{z}\right\}$$

$$\cdot \exp\left\{-\frac{\pi^2 w_0^2}{2\lambda^2 z^2}(\xi^2 + \eta^2)\right\} \langle I(x, y, x)\rangle. \quad (57)$$

It should be remembered, however, that this result is based on the approximation made in deriving eq. (54) from eq. (39). The physical significance of this approximation is now clear and is the following.

In effect, the accumulated random phase along the path has been incorporated in a single random phase-changing screen placed just in front of the radiating laser aperture. This observation follows from a very useful result given in Ratcliffe,[16] which is that the angular correlation function for the far field is the Fourier transform of the *magnitude* squared of the aperture field distribution. Using this result, one would expect the scattered field to be initially correlated over angles of the order of $\lambda/(\pi w_0)$. But that with increasing z, as the magnitude of the field becomes more finely divided, the lateral correlation scale size would be less than the $\lambda z/(\pi w_0)$ indicated by eq. (57). This speculation is borne out by some computer simulations.[19] A better approximation might be to replace the inner integral of eq. (39) by $z\rho(\xi - \mu z/2, \eta - \nu z/2)$. This would move the single equivalent random phase-changing screen to a point midway along the path. However, neither approximation is particularly satisfactory, and it is obviously safer to use the unapproximated eq. (39), although this would probably require numerical evaluation.

VI. CONCLUSIONS

The beam propagation method, based on the parabolic approximation to the wave equation, has been applied to the propagation of a laser beam through the clear but turbulent atmosphere. Papoulis' extension to optical fields of Woodward's ambiguity function was used. The resulting formula for the lateral mutual coherence function of the propagating laser beam, and hence its mean intensity profile, agrees with that of Tatarskii. The method has the advantage over alternative approaches of greater physical clarity. Incorporation of the effects of very-large-scale irregularities and of receiver directivity is then very simple, as also is the estimation of signal statistics and allowing for the consequences of spatial nonstationarity of the turbulence.

REFERENCES

1. B. G. King, P. J. Fitzgerald, and H. A. Stein, "An Experimental Study of Atmospheric Optical Transmission," B.S.T.J., *62,* No. 3 (March 1983), pp. 607–29.
2. J. W. Strohbehn, ed., *Laser Beam Propagation in the Atmosphere,* New York: Springer, 1978.
3. A. Ishimaru, "The Beam Wave Case and Remote Sensing," in *Laser Beam Propagation in the Atmosphere,* J. W. Strohbehn, ed., New York: Springer, 1978.
4. V. I. Tatarskii, "The Effects of the Turbulent Atmosphere on Wave Propagation," Israel Program for Scientific Translations, 1971.
5. J. Van Roey, J. van der Donk, and P. E. Lagasse, "Beam-Propagation Method: Analysis and Assessment," J. Opt. Soc. Amer., *71,* No. 7 (July 1981), pp. 803–10.
6. F. D. Tappert, "The Parabolic Approximation Method," in *Wave Propagation in Underwater Acoustics* (Lecture Notes in Physics, No. 70), J. B. Keller and J. Papadakis, eds., New York: Springer, 1977.
7. J. A. Fejer, "The Diffraction of Waves in Passing Through an Irregular Refracting Medium," Proc. Roy. Soc. A, *220* (December 1953), pp. 455–71.
8. R. W. Lee and J. C. Harp, "Weak Scattering in Random Media, With Applications to Remote Probing," Proc. IEEE, *57,* No. 4 (April 1969), pp. 375–406.
9. A Papoulis, "Ambiguity Function in Fourier Optics," J. Opt. Soc. Amer., *64,* No. 6 (June 1974), pp. 779–88.
10. P. M. Woodward, *Probability and Information Theory With Applications to Radar,* Elmsford, N.Y.: Pergamon, 1953.
11. R. H. Clarke, "Acoustic and Electromagnetic Waves Propagating in a Tenuous Random Medium," in *Aspects of Signal Processing With Emphasis on Underwater Acoustics,* G. Tacconi, ed., Boston, Mass.: Kluwer, 1977.
12. J. W. Goodman, *Introduction to Fourier Optics,* New York: McGraw-Hill, 1968, p. 60.
13. D. E. Kerr, *Propagation of Short Radio Waves,* New York: McGraw-Hill, 1951, p. 44.
14. A. Papoulis, *Probability, Random Variables and Stochastic Processes,* New York: McGraw-Hill, 1965, p. 325.
15. H. Kogelnik and T. Li, "Laser Beams and Resonators," Appl. Opt., *5,* No. 10 (October 1966), pp. 1550–67.
16. J. A. Ratcliffe, "Some Aspects of Diffraction Theory and Their Application to the Ionosphere," Reports on Progress in Physics, 1956, *19,* pp. 188–267.
17. L. A. Chernov, *Wave Propagation in a Random Medium,* New York: McGraw-Hill, 1960, p. 9.
18. A. Ishimaru, *Wave Propagation and Scattering in Random Media,* Vol. 2, New York: Academic, 1978, pp. 412 and 543.
19. M. R. Inggs and R. H. Clarke, "A Computer Simulation of Propagation Through a Tenuous Random Medium," Proc. IEEE-URSI Symp., Seattle, Washington, June 1979.

AUTHOR

Richard H. Clarke, B.Sc., 1956, Ph.D., 1960 (Electrical Engineering), University College, London; Assistant Professor, University of California, Berkeley, 1962–1964; Bell Laboratories, 1964–1968, and visiting member of technical staff, summers, 1981–1985. Mr. Clarke worked for the NATO ASW Research Centre, La Spezia, Italy, from 1969–1974 in their theoretical studies group. Since 1974 he has been teaching in electrical engineering at Imperial College, London. While at AT&T Bell Laboratories he worked on the theory of mobile radio and of propagation of random optical fields, and on the design of antireflection coatings for semiconductor laser diodes.

AT&T Technical Journal
Vol. 64, No. 7, September 1985
Printed in U.S.A.

Simple Analytical Representation of Antenna Spatial Radiation Patterns With Application to the Pyramidal Horn-Reflector Antenna

By J. SHAPIRA*

(Manuscript received January 14, 1985)

The spatial (three-dimensional) radiation characteristics of directive antennas are often needed in radio interference calculations and predictions. Presented is a method that allows a fast computation of the spatial radiation envelope characteristics of antennas from measured pattern information. This is achieved by fitting the measured data to simple functional forms that are based on salient physical properties of the antennas. An example is given in which radiation envelopes for a pyramidal horn-reflector antenna, widely used in AT&T service, are calculated from measured data. Superpositions of quadratic functions to fit main radiation lobes and logarithmic functions to represent the side-lobe envelopes are being used, and good agreement with the measured data is demonstrated.

I. INTRODUCTION

Ground scattering is a major source of interference in microwave communication links.[1] Its analysis involves repeated computations incorporating the antenna directivity in different directions throughout its three-dimensional (3D) coverage.

For a mathematical representation of the 3D directivity pattern of the antenna to be applicable for such purposes, it should be compatible with the data storage and computability constraints and commensurate with the accuracy requirements of the analysis package. Direct

* AT&T Bell Laboratories.

representation of measured data, for instance, requires storage of densely sampled data, about 20 points per beamwidth in every measurement cut, and a fraction of a beamwidth separation between cuts, resulting in the immense number of 400,000 points in the data storage for a 1°-beamwidth antenna. Interpolation methods, based on the bandlimitedness of the antenna spatial spectrum, can reduce the required database by at least two orders of magnitude,[2,3] but many applications require even further simplicity. Such a simplification can be offered by approximating the radiation pattern by its envelope only and disregarding the detailed side-lobe structure, which varies from one antenna unit to another of the same type and varies rapidly with frequency.

The envelope surface, representing a local angular average (or peak cover) of the radiation pattern, is much smoother and more repeatable. Its generation still requires all side-lobe peaks, and any general procedure of surface matching is not as straightforward and simple as desired. A major reduction in complexity may be achieved when use is made of salient features of the antenna. It is demonstrated, in what follows, that a complex surface may be well approximated to a high degree of accuracy with relatively simple analytic expressions by relying on basic antenna features.

It is worth mentioning here that simple models, using a single skirt function, have found use for specifications,[4,5] but their approximation is much too crude for other applications.

The approximation to the radiation pattern of the Pyramidal Horn-Reflector (PHR) antenna,[6] widely used by AT&T Communications, is worked out as an example encompassing only four coefficients in each region, out of a lookup table with 29 constants, while maintaining tight match over the main beam and no more than 5-dB deviation from the side-lobe peaks throughout. This work was briefly summarized in Ref. 7.

II. SURVEY OF PERTINENT ANTENNA FEATURES

2.1 Symmetries in the antenna pattern

The field distribution in a radiating aperture is transformed to the far field via the Fourier Transform (FT),

$$F(u, v) \propto \int_{\text{aperture}} dx \int dy f(x, y) e^{-jk(xu+yv)} \tag{1}$$

(see Fig. 1), where

$$u = \sin\theta \sin\phi$$
$$v = \sin\theta \cos\phi,$$

and k, being the wave number, equals $(2\pi)/\lambda$.

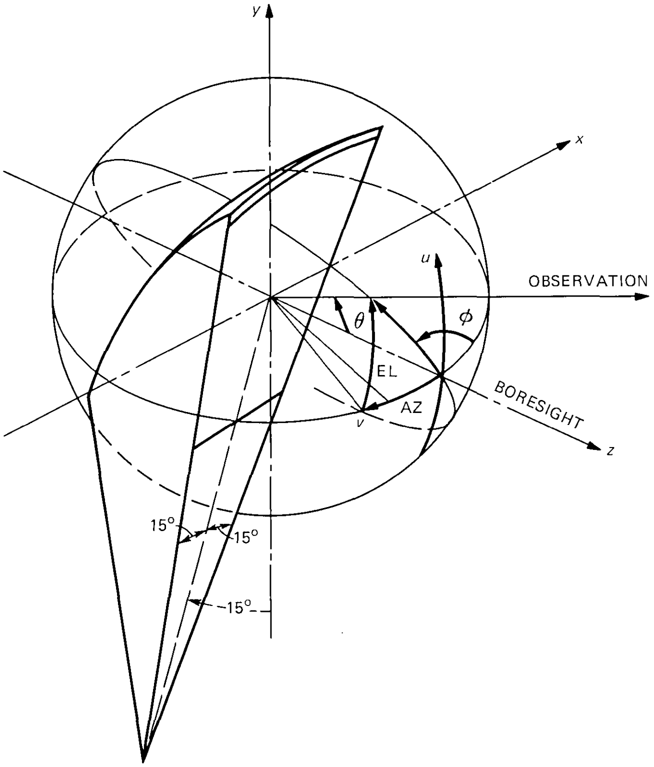


Fig. 1—Coordinate systems for the pyramidal horn-reflector antenna.

For any cut through the z axis (perpendicular to the aperture), one may rotate the coordinates to align the cut with the u and v axes and, thus, reduce eq. (1) to a one-dimensional FT

$$F(u, o) \propto \int_{\text{aperture}} dx e^{-jkxu} \int_{\text{aperture}} dy f(x, y), \tag{2}$$

where the symmetry rules of Table I apply.

The aperture of the PHR antenna, for example, is tilted and not perpendicular to its main beam boresight (see Fig. 1). In the vertical plane, the aperture distribution is, therefore, not real, and the resulting radiation pattern not symmetrical. A simple computation technique by which the field is projected onto a virtual vertical aperture is widely used (see Refs. 8 through 10 in connection with the PHR antenna) and produces a symmetrical pattern in the vertical plane, which is obviously in error. In the horizontal plane, however, the aperture is perpendicular to the pattern boresight and is symmetrical. Further,

Table I—Aperture symmetry rules

Rule	Aperture Distribution	Radiation Pattern
1	$f(x)$ real or $f(x) = \|f(x)\| e^{j\psi(x^2)}$	$\|F(u)\| = F\|(-u)\|$ $\arg F(u) = -\arg F(-iu)$
2	$f(x) = \pm f(-x)$	$F(u) = \pm F(-u)$
3	$f(x, y) = f_x(x) f_y(y)$	$F(u, \theta) = F_u(u) F_v(v)$ $f_x(x) < \longrightarrow F_u(u)$ $f_y < \longrightarrow F_v(v)$
4	$f(\rho, \phi) = f_\rho(\rho) e^{jn\phi}$ $\rho = \sqrt{x^2 + y^2}$ $\phi = tg^{-1}(y/x)$	$F(u, v) = \overline{F}_n(w) e^{-jn\Phi}$ $w = \sqrt{u^2 + v^2}$ $\Phi = tg^{-1}(u/v)$ (Hankel transform)

Table II—Asymptotic power density drop-off for rectangular distributions

Rectangular Aperture Distribution	Asymptotic Power Density Drop-Off
Uniform	u^{-2}
Cosine	u^{-4}
(Cosine)^2	u^{-6}
(Cosine)^3	u^{-8}
Taylor	u^{-2}
Dolph-Chebyshev	Constant

with the side walls of the aperture tilted, the field is not separable (rule 3 in Table I), nor is the far field.

2.2 Asymptotic drop-off of the radiation pattern envelope

The FT relationship between the aperture field and the radiation pattern may be evaluated asymptotically for large u. By integrating by parts one gets

$$F(u) = \int_a^b f(x) e^{-jkxu} dx$$
$$= \sum_{n=0}^{\infty} \left(\frac{1}{u}\right)^{n+1} f^{(n)}(x) e^{-j[kxu+(n-1)\pi/2]} \Big|_a^b, \qquad (3)$$

which represents the radiation pattern as an asymptotic series of diffraction terms by discontinuities at the aperture edges. The leading term in that inverse power series is determined by the order of the derivative of the aperture illumination that is discontinuous at the edges. A representative list of aperture distributions, along with the resulting power density drop-off, is listed in Table II (for an extensive list, refer to Ref. 11).

Phase-modulated aperture distributions (e.g., wide flare horns, shaped beam antennas) may have an additional, nondiffraction con-

Table III—Asymptotic power density drop-off for circular distributions

Circular Aperture Distribution	Asymptotic Power Density Drop-Off
Uniform	u^{-3}
$1 - (\rho/a)^2$	u^{-5}
$[1 - (\rho/a)^2]^2$	u^{-7}
$[1 - (\pi/a)^2]^3$	u^{-9}
Circular Taylor	u^{-3}

tribution resulting from saddle-point integration, which applies principally to the main beam (see, for example, Ref. 12).

The radiation pattern of a planar aperture with rectangular separable distributions (rule 3 in Table I) is a product of the principal plane patterns. Polar separable distributions generate patterns with drop-off rate $u^{-n-(1/2)}$ by integrating the Hankel transform[13] by parts and using asymptotic expressions for the Bessel functions. Representative circular distributions are listed in Table III.[14] Discrete element construction of the aperture distribution adds grating lobes to the antenna pattern when the elements are periodically displaced. The grating lobes are isolated, however, and their location can be predicted from the array structure.

A wedge diffraction pattern is a product of axial and cross patterns, with the latter culminating at the shadow boundaries of the incident and reflected illuminations and decaying from it as $(\sin \gamma/2)^{-1}$ for a thin wedge or as $(\sin 3\gamma/2)^{-1}$ for a right-angle wedge,[15] γ being the angle from the shadow boundary.

Diffraction by strips and cylinders is similarly a product of axial and cross patterns, where Snell reflection rules apply to the axial pattern. The diffraction pattern thus forms a cone around the axis, azimuthally and axially weighted by the respective pattern behavior.

2.3 Reconstruction of the antenna 3D radiation envelope approximation

The above survey shows simple pattern behavior for elemental radiators when represented in their natural coordinate systems. The generic form

$$F(u)(\mathrm{dB}) = a - b \log_{10} u \qquad (4)$$

may be used on each of the principal axes of the pattern of a separable distribution or edge diffraction and on representative radial cuts for a nonseparable distribution where azimuthal interpolation functions can close the gap. Paraboloids, circular or elliptical, are used to match the peak regions.

Contributions to the antenna radiation pattern come from the illumination of the aperture and its edges, diffraction by the feed and structural members, and the weather cover. All these can be classified by the categories surveyed in the previous section, and the characteristic pattern of each can be traced on the 3D antenna pattern in regions where it dominates. These traces are easiest to identify on the $\sin\theta$, ϕ polar plot of the antenna pattern, pivoted around the boresight, where they take elliptical shapes. For instance, an edge slanted by an angle α from boresight (z axis) in a plane slanted by β from the xz plane, diffracts the outgoing wave on a cone, the trace of which is an ellipse with axes

$$\begin{aligned} a &= \sin\alpha \\ b &= (1/2)\sin 2\alpha \end{aligned} \tag{5}$$

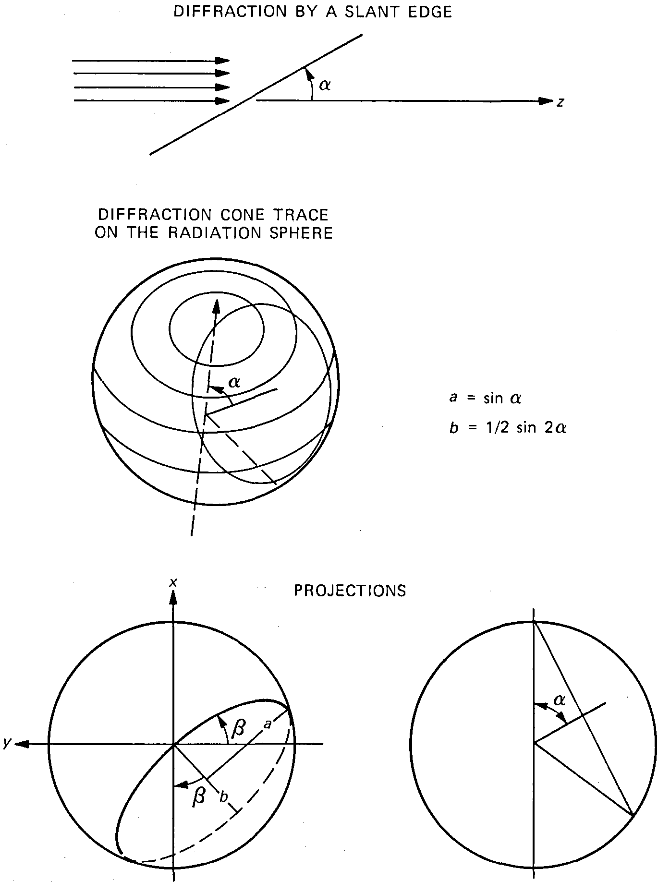


Fig. 2—Projections of the edge diffraction cone.

tangent to a line through the boresight forming an angle β with the y axis (see Fig. 2).

Once skeleton shape matching is obtained, parameters of eq. (4) are adjusted to match the envelope of each contributor in its natural coordinate system. The partial patterns are then retransformed to the antenna coordinate system where final patch up might be required in transition regions.

The desired application of the radiation envelope approximation should determine the coordinates of representation and the transformation formulas. In analyzing scattering interference in terrestrial transmission, for instance, the azimuth (AZ)-elevation (EL) coordinate system blends with the computations of the model much better than θ, ϕ coordinates, and the transformations and approximations are best done directly in that representation (see Fig. 1).

III. THE PYRAMIDAL HORN-REFLECTOR ANTENNA PATTERN

The PHR antenna is extensively used in terrestrial microwave links (see Fig. 3). The measured data of the frontal hemisphere of its 3D radiation pattern at 4 GHz consists of 91 ϕ cuts made every 1° with a sampling rate of 0.08° totaling about 200,000 measured points.[9] The sin θ, ϕ polar plot of its radiation distribution for horizontal polariza-

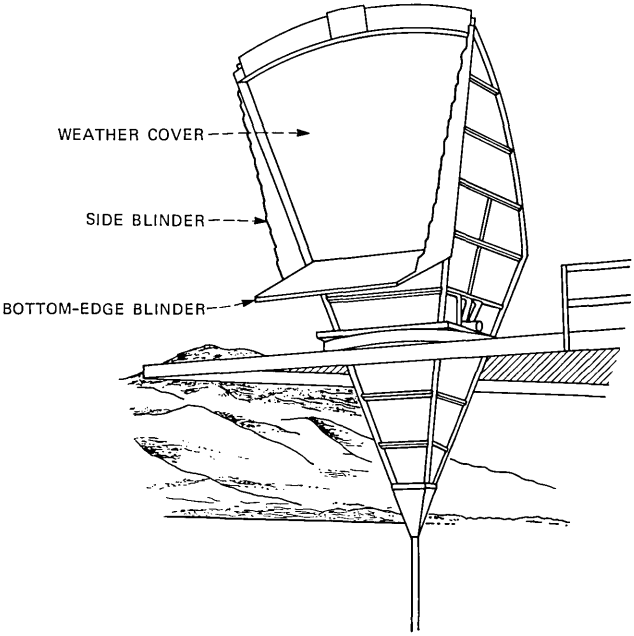


Fig. 3—The AT&T pyramidal horn-reflector antenna.

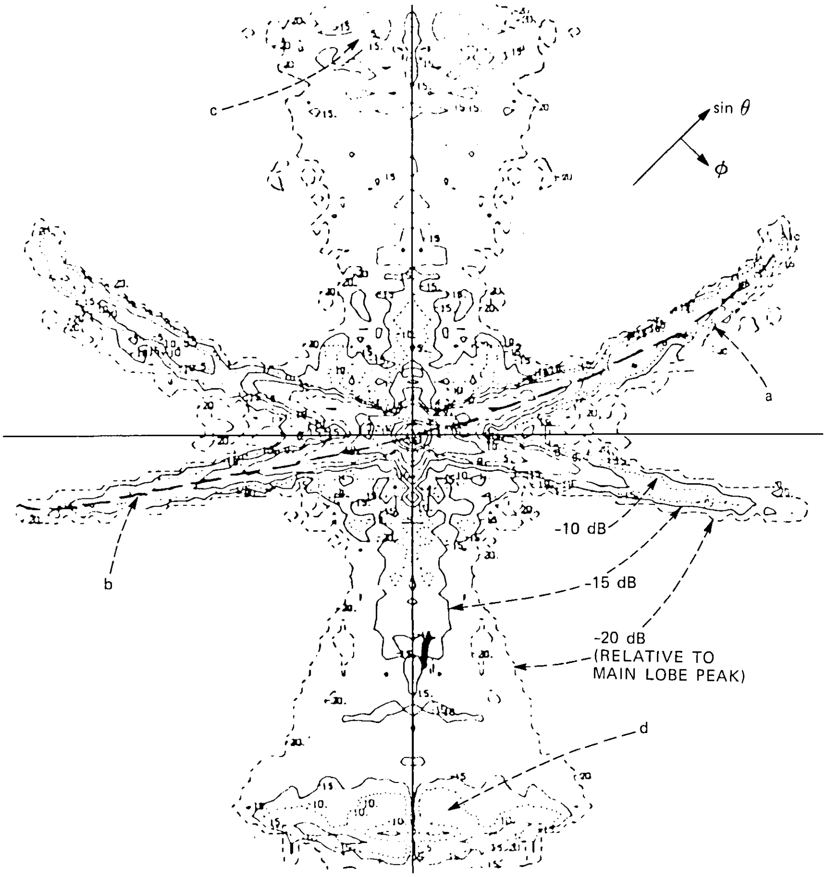


Fig. 4—Radiation pattern for the pyramidal horn-reflector antenna of Fig. 3 in sin θ, ϕ coordinates at 3900 MHz, horizontal polarization.

tion is shown in Fig. 4, truncated at 60 dB below the main beam peak, while the accompanying 3D pattern is shown in Fig. 5 in AZ-EL coordinates. The trace a in Fig. 4 could be matched to an ellipse with $\alpha = 14.5°$, $\beta = 14.5°$ and identified as side edge diffraction. The trace b, on the other hand, matches an ellipse with $\alpha = 3.6°$, $\beta = 14.5°$ corresponding to the side blinder (see Fig. 3).[16] A closer look at the side blinder attachment detail shows a step at the aperture edge, allowing for the aperture edge diffraction to dominate on one side and to be shadowed by the side blinder on the other. The large diffused lobe at d is due to reflection by the weather cover emerging down after a second bounce from the reflector (see Fig. 6), while the one at c is a spillover of the horn field illuminating the top edge of the reflector. The flare of the side-lobe ridge at the top and at the bottom is attributed to the curved top edge of the aperture.

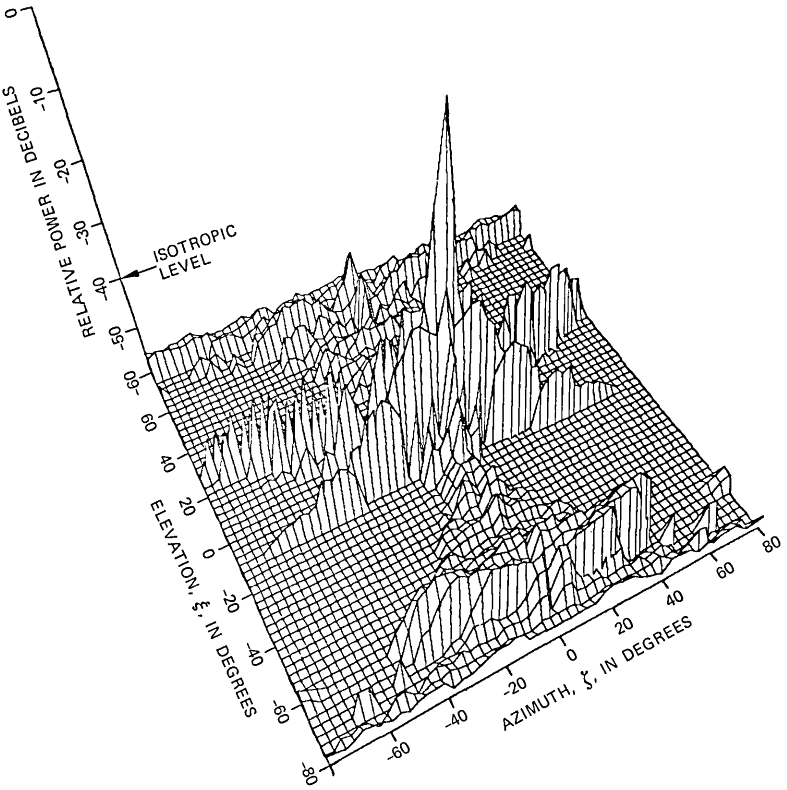


Fig. 5—Radiation pattern for the pyramidal horn-reflector antenna in AZ-EL coordinates at 3900 MHz, horizontal polarization.

The aperture field of the PHR antenna has the TE_{10}-mode distribution prevailing in the pyramidal horn, blown up by the reflector. The horizontally polarized field strongly illuminates the side walls but is much weaker at the upper and bottom edges. The case is reversed with vertical polarization as shown in Figs. 7 and 8; the side wall diffraction is highly suppressed, while that of the top and bottom edges is very strong at the center and decays to the sides. The spillover lobe, c, is much stronger, but the window lobe, d, is similar to that of the horizontal polarization.

IV. THE RADIATION ENVELOPE APPROXIMATION FOR THE PHR ANTENNA

The contributors to the radiation pattern, identified by examination of both the pattern and the antenna, are aperture illumination taper, top edge (curved), bottom edge, side edges (slanted $\alpha = 14.5°$, $\beta =$

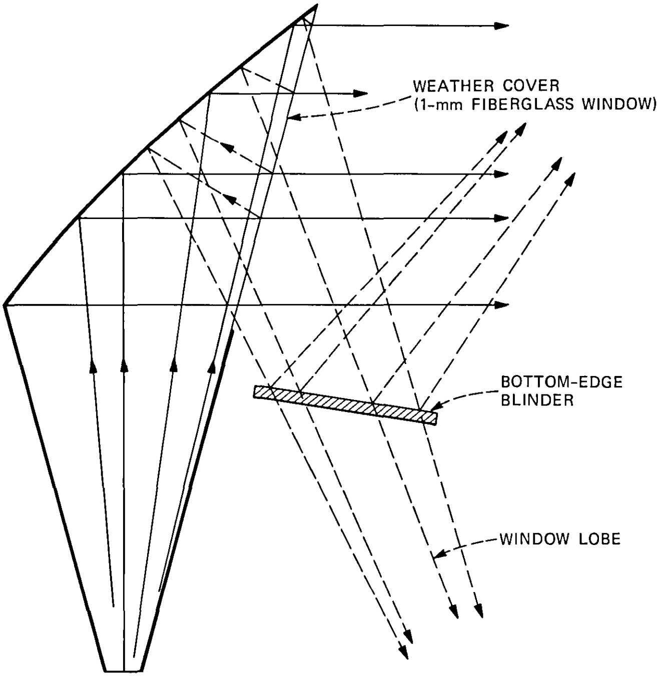


Fig. 6—Pyramidal horn-reflector antenna with bottom-edge blinder and window lobe.

14.5°), side blinders (slanted $\alpha = 3.6°$, $\beta = 14.5°$), weather cover (window lobe), top-edge spillover, and bottom-edge blinder (optional). Each of these contributions is approximated by the following generic function in the peak regions:

$$g = g_{\max} - K_u(u - u_0)^2 - K_v(v - v_0)^2, \tag{6}$$

where $g_{\max}$ is the peak level in decibels below the main beam peak, u and v are the local principal plane coordinates [see eq. (1)] for each contributor, and K_u, K_v are the parabolic coefficients to be matched. In the fall-off regions,

$$g = a_u + a_v - b_u \log_{10} u - b_v \log_{10} v \tag{7}$$

supplies the envelope, with a_u, a_v, b_u, b_v coefficients to be matched. The local coordinate systems are then transformed to the antenna sin AZ, sin EL coordinate system for preserving computational economy in the repeated transformations between the antenna and the scattering model[1] coordinates. Such a simplification is made possible by the fact that the antenna boresight is almost horizontal and its azimuthal plane aligns with that of the scattering model to enough accuracy. The

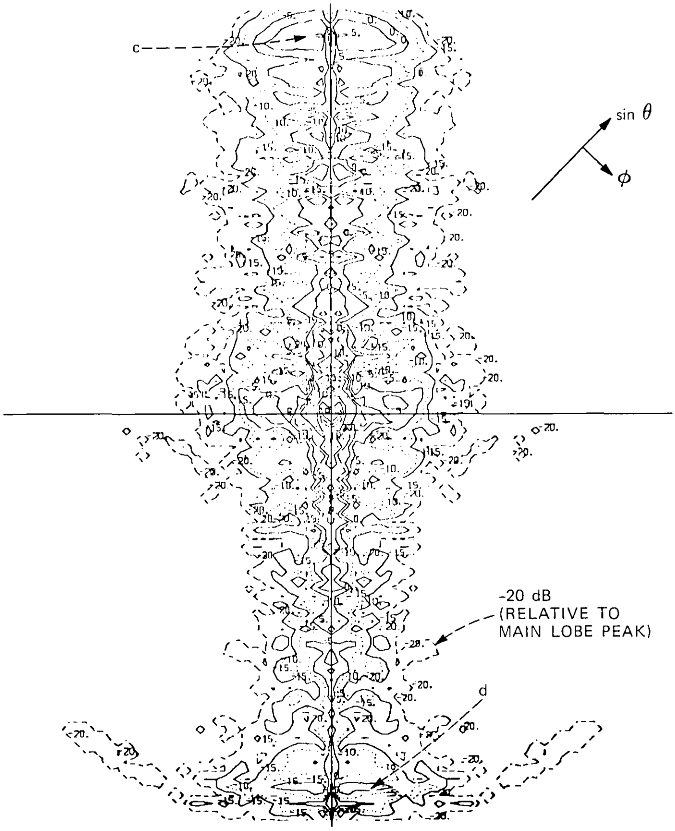


Fig. 7—Radiation pattern for the pyramidal horn-reflector antenna in sin θ, ϕ coordinates, vertical polarization.

sin AZ, sin EL coordinate system is presented in Fig. 1 and is related to the spherical coordinates via

$$\sin \mathrm{EL} = \sin \theta \sin \phi = u$$

$$\sin \mathrm{AZ} = \sin \theta \cos \phi / \sqrt{1 - \sin^2 \theta \sin^2 \phi}$$

and

$$\cos \theta = \cos \mathrm{AZ} \cos \mathrm{EL}$$

$$\sin \phi = \sin \mathrm{EL} / \sqrt{1 - \cos^2 \mathrm{AZ} \cos^2 \mathrm{EL}}.$$

The top edge contribution is flared to ±15° by scaling the AZ coordinate input to eq. (7).

The partial contributions to the radiation envelope thus obtained are drawn in Figs. 9 through 12 and 14 through 17 for horizontal and vertical polarizations, respectively. The overall radiation envelopes,

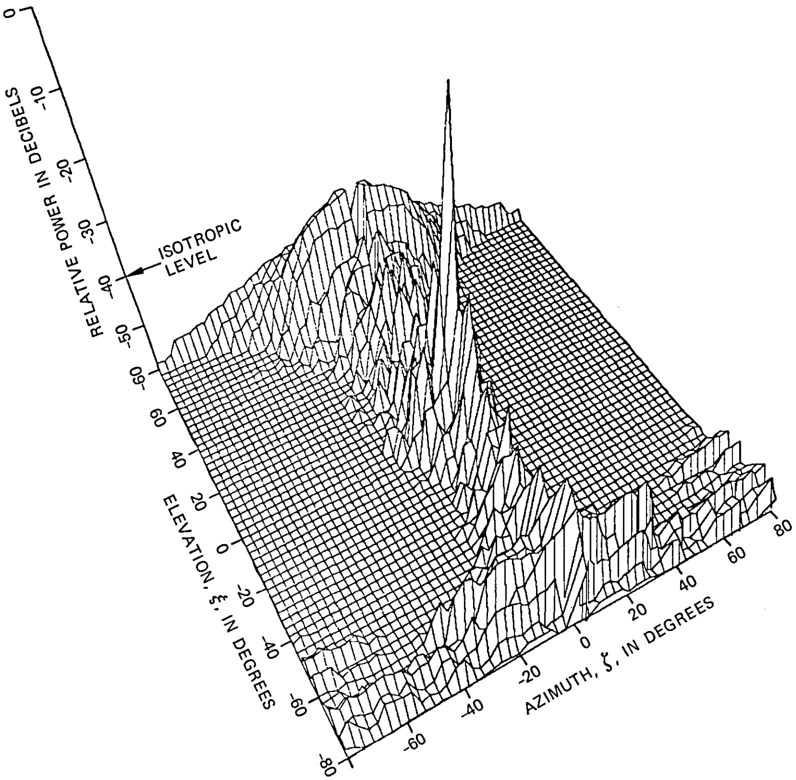


Fig. 8—Radiation pattern for the pyramidal horn-reflector antenna, vertical polarization.

drawn in Figs. 13 and 18 for these polarizations, respectively, are then represented by the largest partial contribution at every point, all the rest being ignored. Note that the three contributions aperture illumination taper, top edge, and bottom edge have been combined into one. The PHR antenna radiation envelope can, therefore, be well approximated by single functions of the type (6) or (7) in every region characterized by their respective coefficients. The total number of constants used by the program for each polarization is 36, only four or five of which are used in any individual function. Also, transformations using eq. (5) are required for the side edge and side blinder contributions.

V. SUMMARY

A procedure was described by which a simple and computationally economical mathematical model can be constructed to approximate

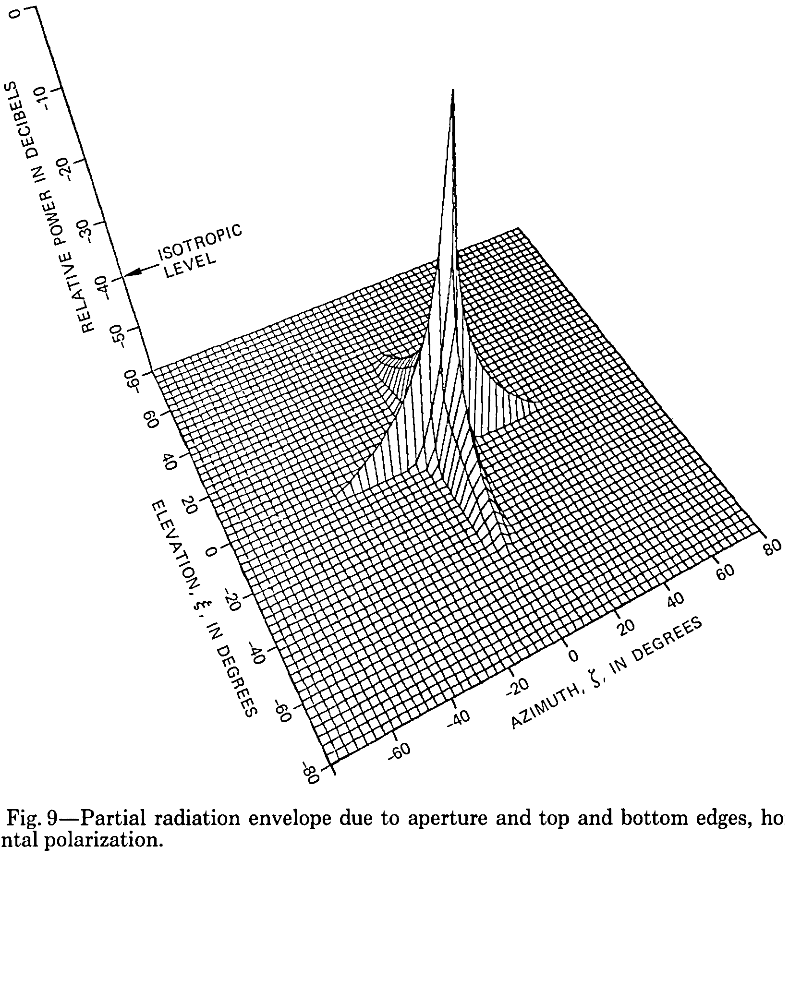


Fig. 9—Partial radiation envelope due to aperture and top and bottom edges, horizontal polarization.

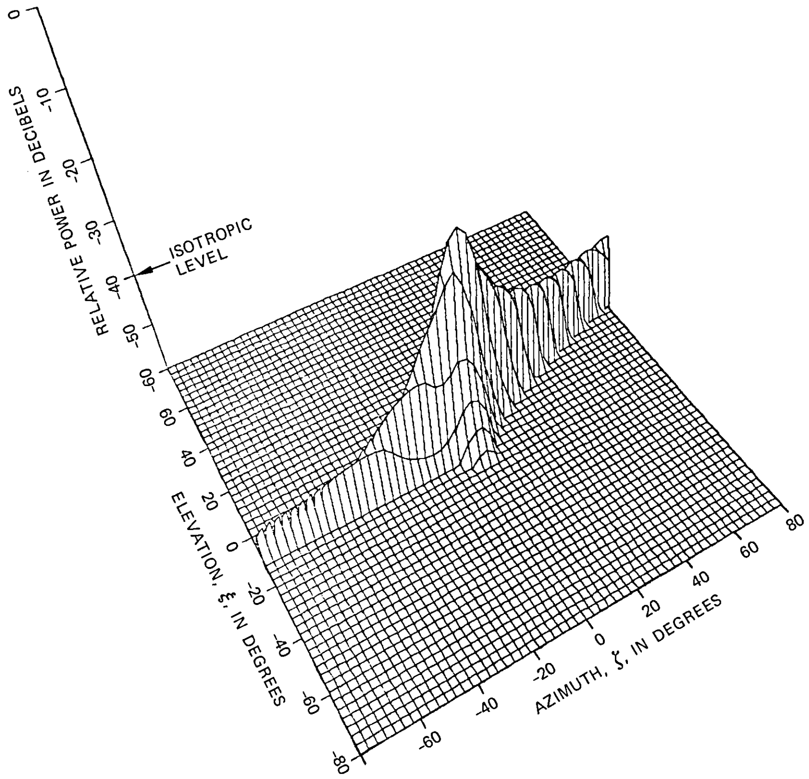


Fig. 10—Partial radiation envelope due to right edge, horizontal polarization.

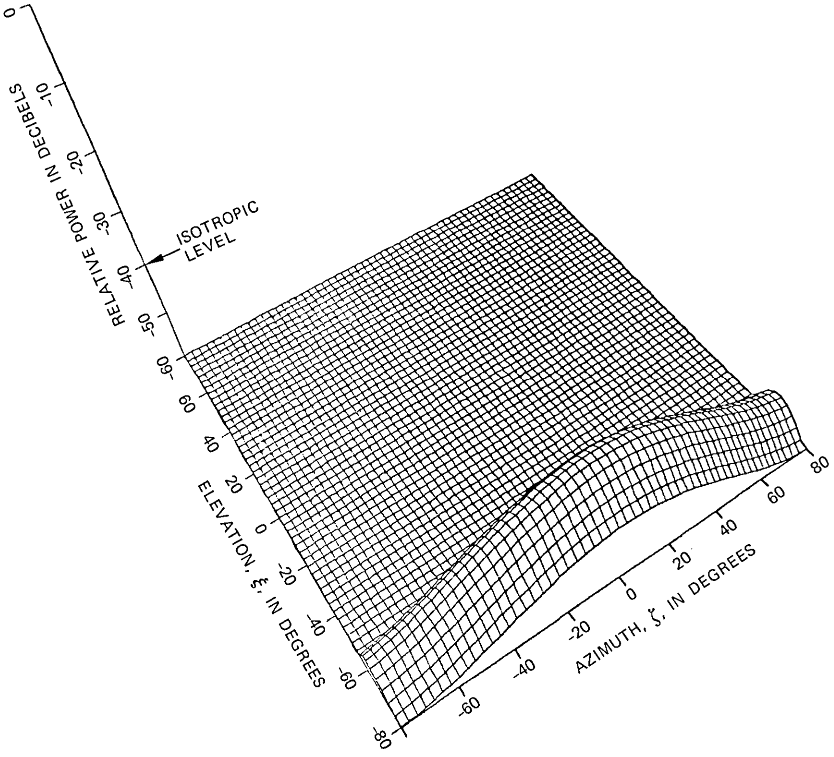


Fig. 11—Partial radiation envelope due to window lobe, horizontal polarization.

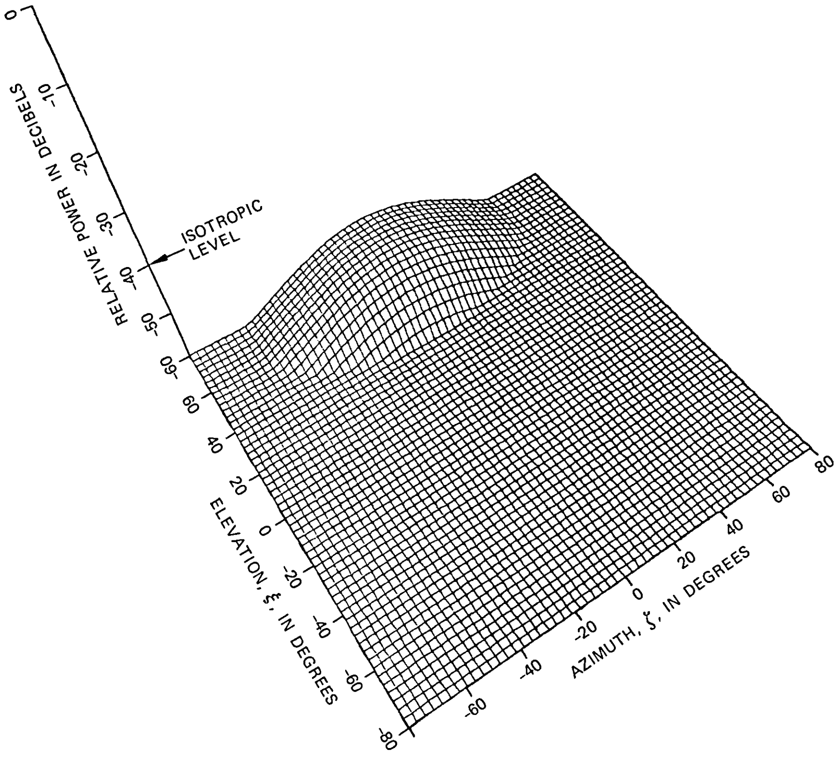


Fig. 12—Partial radiation envelope due to spillover, horizontal polarization.

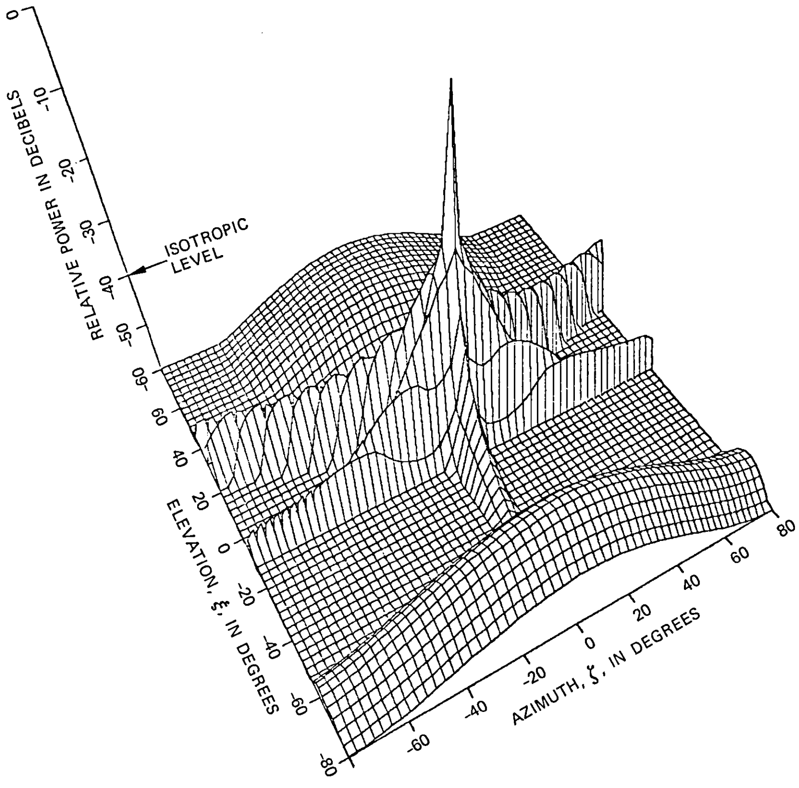


Fig. 13—Overall radiation envelope for the pyramidal horn-reflector antenna, horizontal polarization.

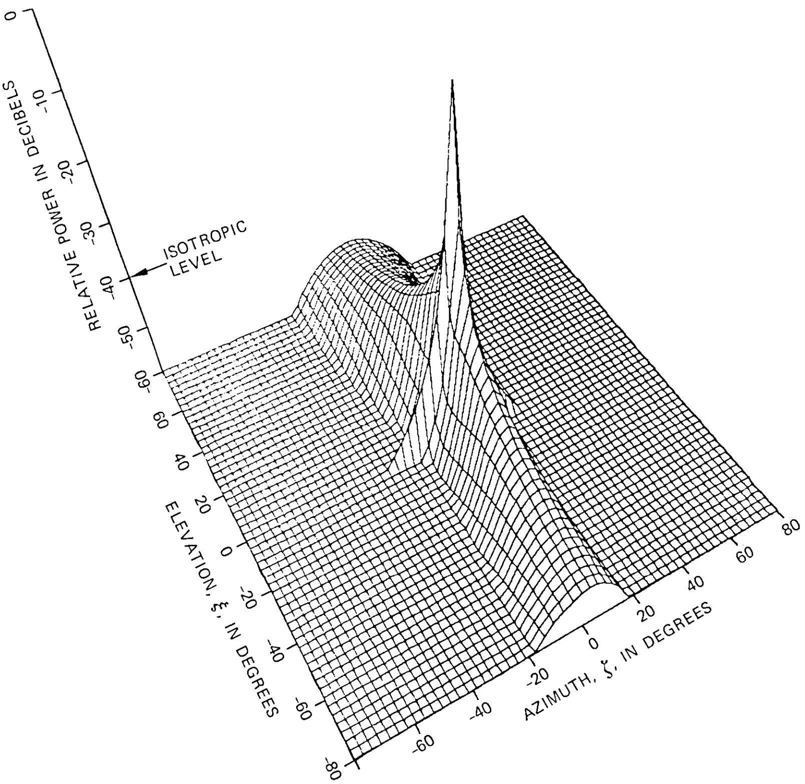


Fig. 14—Partial radiation envelope due to aperture and top and bottom edges, vertical polarization.

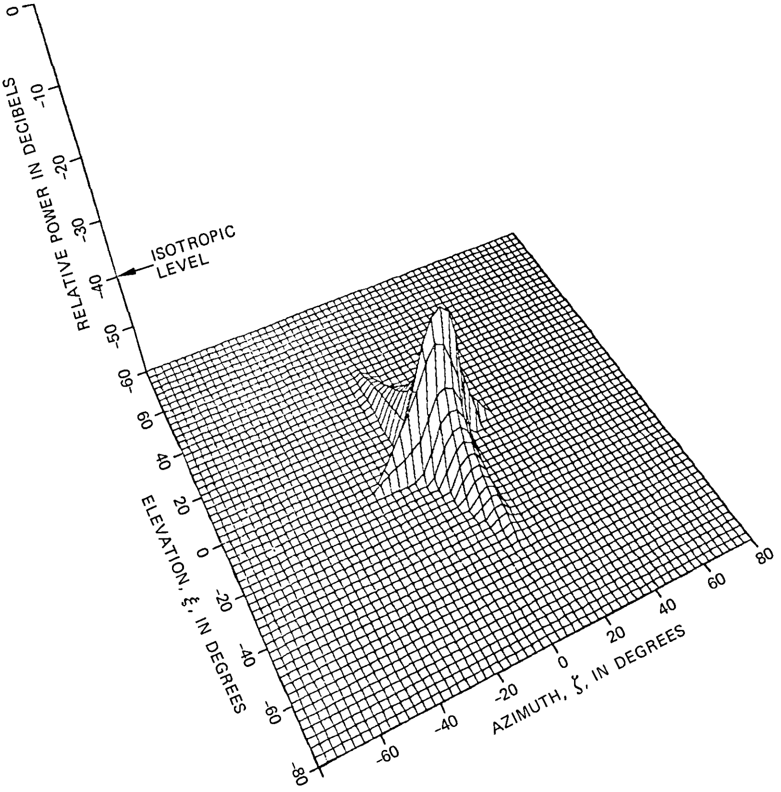


Fig. 15—Partial radiation envelope due to right edge, vertical polarization.

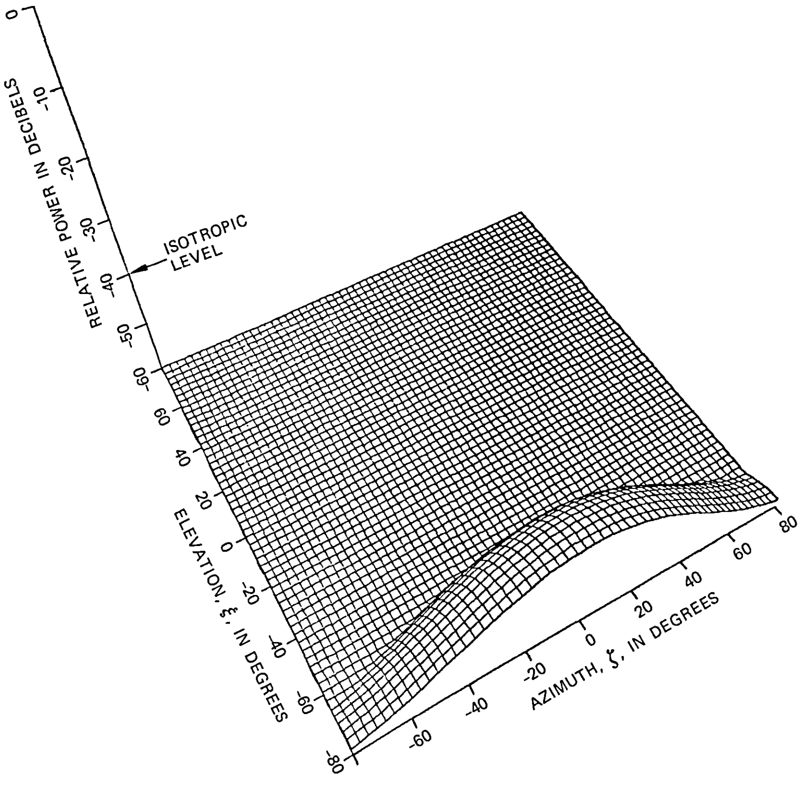


Fig. 16—Partial radiation envelope due to window lobe, vertical polarization.

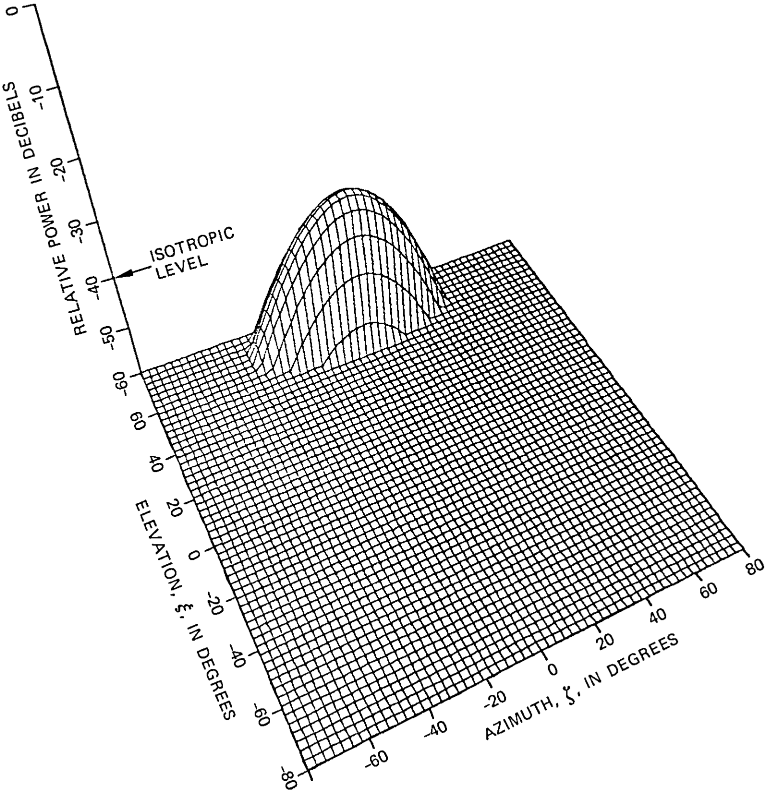


Fig. 17—Partial radiation envelope due to spillover, vertical polarization.

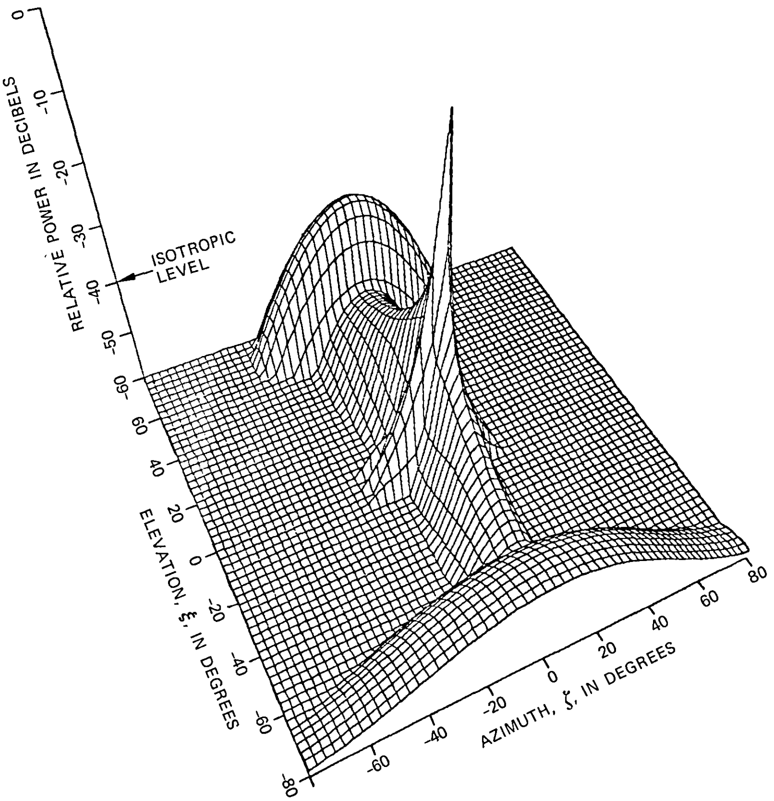


Fig. 18—Overall radiation envelope for the pyramidal horn-reflector antenna, vertical polarization.

the 3D radiation envelope of directive antennas by making proper use of the salient antenna features. Accuracy is a parameter in such a model. Test computations of carrier-to-interference ratios executed by using this model versus the measured 3D pattern in the terrain scattering interference model[1] agreed to better than 3 dB in all cases.

VI. ACKNOWLEDGMENTS

The author is grateful to Mr. W. A. Gianopoulos for his excellent programming work and to his colleagues Messrs. A. J. Giger, H. C. Wang, and P. E. Butzien for their comments and editing work.

REFERENCES

1. A. J. Giger and J. Shapira, “Interference Caused by Ground Scattering in Terrestrial Microwave Radio Systems,” ICC ’83 Conf. Rec. (June 1983), pp. 1254–61.
2. J. Shapira, “Sampling and Processing in Antenna Pattern Measurements,” Proc. 11th Convention of Electrical and Electronic Engineers in Israel (October 1979), IEEE Publication 79CH1566-9, p. B2-5.

3. O. M. Bucci et al., "Use of Sampling Functions for the Analysis of Radiating Structures," Final Report to ESTEC Contract N.4380/80/NL/DS, 1981.
4. International Radio Consultative Committee (CCIR), 13th Plenary Assembly, Radiation Diagrams of Antennas for Earth Stations in the Fixed Satellite Service for Use in Interference Studies, Rep. 391-4, Vol. IV, Int. Telecommun. Union, Geneva, Switzerland.
5. W. Kuebler, "A Procedure for Estimating the Power Pattern of a Narrow-Beam Antenna," IEEE Trans., *EMC-23,* No. 2 (May 1981), pp. 100–3.
6. H. T. Friis, "Microwave Repeater Research," B.S.T.J., *27,* No. 2 (April 1948), pp. 183–246.
7. J. Shapira, "Simple Representations of Antenna Spatial Radiation Patterns," Dig. IEEE, *AP-S,* 1983 Int. Symp., May 1983, pp. 112–5.
8. J. N. Hines, Tingye Li, and R. H. Turrin, "The Electrical Characteristics of the Conical Horn-Reflector Antenna," B.S.T.J., *42,* No. 4, Part 2 (July 1963), pp. 1187–211.
9. P. E. Butzien, "Three-Dimensional Radiation Characteristics of a Pyramidal Horn-Reflector Antenna," B.S.T.J., *60,* No. 6 (July–August 1981), pp. 913–21.
10. E. V. Jull, "Aperture Antennas and Diffraction Theory," Peter Peregrinus Ltd., 1981, pp. 68–73.
11. F. J. Harris, "On the Use of Windows for Harmonic Analysis With the Discrete Fourier Transform," Proc. IEEE, *66,* No. 1 (January 1978), pp. 51–83.
12. L. B. Felsen and N. Marcuvitz, *Radiation and Scattering of Waves,* Englewood Cliffs, N.J.: Prentice-Hall, 1973.
13. A. Papoulis, *Systems and Transforms With Applications in Optics,* New York: McGraw-Hill, 1968, p. 163.
14. W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design,* New York: Wiley, 1981.
15. R. G. Kouyoumjian, *The Geometric Theory of Diffraction and Its Applications in Numerical and Asymptotic Techniques in Electromagnetics,* New York: Springer-Verlag, 1975.
16. D. T. Thomas, "Design of Multiple-Edge Blinders for Large Horn-Reflector Antennas," IEEE Trans. Ant. Propag., *AP-21,* No. 2 (March 1973), pp. 53–158.

AUTHOR

Joseph Shapira, B.Sc. and M.Sc. (Electrical Engineering), Technion–Israel Institute of Technology, in 1961 and 1976, respectively; Ph.D. (Electrophysics), 1973, Polytechnic Institute of New York; Rafael–Armament Development Authority of Israel, 1963–1980, 1981–1982, 1983—; Electro-optics Industries of Israel, 1980–1981; Bell Laboratories, 1982–1983. During his first years with Rafael, Mr. Shapira founded and headed the Electromagnetics Department, encompassing research and development in antennas, arrays, radomes, propagation, target characteristics and electromagnetic compatibility. In 1979 he was appointed advisor on electromagnetic systems to the Director of Rafael. Later that year he was assigned Special Assistant to the Chief of Research and Development of the Israeli Ministry of Defense, where he directed reviews and planning of high technology major projects in radar, fire control, and communication systems. In 1980 he became the manager of the Optronics Systems Operation in the Electro-optics Industries of Israel. In 1981 he returned to Rafael as a Research Fellow. From 1982 to 1983 he was with Bell Laboratories, on a Sabbatical leave from Rafael, where he was engaged in terrain scattering research. He is now Deputy Director of the Guidance Division Rafael. Mr. Shapira is also affiliated with the Technion, Haifa, as a part-time Associate Professor in the Electrical Engineering Faculty. His areas of interest are electromagnetic engineering and electromagnetic compatibility. He won the IEEE antenna and Propagation Society Best Paper Award in 1974 as a coauthor of "Ray Analysis of Conformal Antenna Arrays." In 1980 he was awarded the A. D. Bergman Prize (presented by the President of Israel) for scientific and technical achievements in electromagnetic engineering, and his contributions to Rafael's technological capabilities. In 1981 Mr. Shapira was nominated the President of The Israel National Committee for Radio Science, and he was nominated to the delegation to the International Union of Radio Science. Senior member, IEEE.

AT&T Technical Journal
Vol. 64, No. 7, September 1985
Printed in U.S.A.

Convolutional Coding for High-Speed Microwave Radio Communications

By M. KAVEHRAD*

(Manuscript received January 1, 1985)

This paper describes the design, testing, and measured performance of a rate 11/12 self-orthogonal convolutional codec that meets the bit-error-rate objective of M-state quadrature amplitude modulation systems in terrestrial radio transmission. The objective is to reduce a bit error probability of 10^{-6} to 10^{-10} or smaller. In fact, the measured output error probability is well below 10^{-10} when the channel error probability is below 10^{-5}.

I. INTRODUCTION

To maintain the low bit error rates—as required for high-quality data transmission in the face of increased demand for bandwidth efficiency—through the use of M-ary quadrature amplitude modulation (M-QAM) signaling, application of error-correction coding in terrestrial digital radio transmission may be desirable. Accumulated "randomly scattered" errors from different error sources may make it difficult to meet an objective of very low average background error probability, for example, a 10^{-10} Bit Error Rate (BER) with a high-level modulation such as 64-QAM.

As will be explained in this paper, convolutional coding was considered as a potential candidate for this task. To answer some of the implementation questions, and because the code seems very attractive for high-speed transmission, we developed a rate 11/12 double error-correcting convolutional codec and measured its actual performance

* AT&T Bell Laboratories.

in terms oı codec input/output bit error probability. The code efficiency is 0.9166 and is in the range that is justifiable for spectrally limited terrestrial radio applications. The intention is to have the process of encoding and decoding applied to data rates in the high megabit range. As such, there is a need for considerable parallel processing in the encoder and decoder realizations. In this paper we describe a procedure for such processing for a rate 11/12 convolutional code. The parallel processing can be implemented using standard logic elements.

After a general description of convolutional coding, we describe the implemented codec in Section I and then present the test setup in Section II. Finally, in Section III we present the measured performance of the codec. Section IV provides our conclusions.

1.1 Code structure

In general, a convolutional encoder can be assumed to be a linear sequential network that maps a block of k_0 parallel bits entering the circuit into an $n_0 > k_0$ bit block over a certain period of time. If the bits leaving the encoder are the original data bits plus $(n_0 - k_0)$ parity-check bits, defined by the particular code polynomials, the code is called systematic. The ratio k_0/n_0 is defined as the code rate. In convolutional coding, the parity bits in a given block have not only been affected by the data bits in the present block but also the preceding blocks. (This is not the case for block codes.)

The constraint length of the code N is the number of bits n_0 in a coded block multiplied by the number of blocks m checked by the $n_0 - k_0$ parity checks. For block codes, $m = 1$.

In general, convolutional codes have the same limitations and, roughly speaking, the same inherent capabilities as block codes.

Like block codes, convolutional codes are capable of correcting random errors, burst errors, and combinations of random and burst errors. Since the parity bits in a convolutional code check information symbols in the blocks preceding the present block, the basic parity matrix is of the form

$$h = [P^T_{m-1} \quad OP^T_{m-2} \quad O \cdots P^T_0 I],$$

where $(n_0 - k_0)$ by k_0 matrices P^T_i are arbitrary, and O and I represent, respectively, the zero and identity matrices of order $n_0 - k_0$. The encoding process associates $(n_0 - k_0)$ parity checks, as specified by the matrix h, to every block of k_0 bits of the entering information, once every n_0 channel-bit times.

Decoding of a convolutional code is possible by algebraic or sequential methods. Although both techniques deal with convolutional codes, they do so in entirely different manners. As a result, most of the definitions applied in coding are of different forms, depending on the way that the code is decoded.

In algebraic methods, the decoder bases its decisions on the check digits within a constraint length, but in the sequential methods the decisions are made based on a section of the syndrome that is much longer than the code-constraint length. This substantially improves error-correcting performance.

Decoding simply associates an error pattern with the syndrome. After the decoder calculates the syndrome, it uses it to decode the oldest block in its registers. Like block codes, a convolutional code can be modeled in terms of parity-check and generator matrices.

An n-tuple b is a code word in the convolutional code family described by h if and only if $bH^T = 0$, where H^T is the transpose of the parity-check matrix

$$H = \begin{bmatrix} P_0^T I & & & \\ P_1^T & OP_0^T I & & \\ P_{m-1}^T & OP_{m-2}^T & O \cdots & P_0^T I \end{bmatrix},$$

where P_i^T, O, I were defined in the equation for h. The sequence b can be obtained by multiplying the information-bit sequence $d = d_{k-1} \cdots d_1 d_0$ by the generator matrix of the code, which is

$$G = \begin{bmatrix} IP_0 & OP_1 \cdots & OP_{m-1} \\ & IP_0 \cdots & OP_{m-2} \\ & & IP_0 \end{bmatrix},$$

that is, $b = dG$.

It can be seen that $GH^T = 0$. The syndrome of a received n-tuple r is defined as $rH^T = s$. Clearly, all n-tuples containing errors (not a code word) have nonzero syndromes.

Self-orthogonal codes are a class of convolutional codes that are rather simply implementable, and they can also be decoded with majority-logic decoding (a nonsequential decoding approach). The disadvantage of these codes is that for large values of n_0 and k_0 the random-error-correction ability of the code decreases, that is, their minimum distance gets small. For ease of implementation only the self-orthogonal codes with $n_0 - k_0 = 1$ are usually considered for construction. In self-orthogonal codes with minimum distance d, at least $d - 1$ rows of the parity matrix H are orthogonal (in the coding sense) on each of the k_0 bits of the zero block. For these codes, it can be shown that

$$N \leq n_0 \left[\frac{(n_0 - 1)(d - 1)(d - 2)}{2} + 1 \right],$$

where N is the constraint length and d the minimum distance of the code.

Suppose that a self-orthogonal code has the ability of correcting $t = (d - 1)/2$ random errors. If t or fewer errors occur within a constraint length of the $(m - 1)$th block, it can be shown that it is possible to arrange a set of $J = d - 1$ orthogonal check sums on each data bit in the $(m - 1)$th block of this code. It has been proved that the value assumed by the majority of these check sums will always be the correct value of the noise digit that was added to the considered data bit.[1] Therefore, by adding this bit to the data bit the correction can be made.

The threshold decoder corrects errors by reencoding the received bit stream, adding the regenerated parity bit to the corresponding received parity bit, and then forming the syndrome. Then the error pattern associated with the syndrome is added to the $(m - 1)$th block of the received sequence. For practical limitations, the m-bit section of the (semi-infinite) syndrome that is within a constraint length is stored in the syndrome register. As mentioned above, $J = d - 1$ of the check sums represented by the syndrome bits are orthogonal on each data bit in the $(m - 1)$th block.

Each of the k_0 majority gates functions as a voter for one of the orthogonal check-sum sets. Therefore, the syndrome register bits are weighted according to the coefficients of the generator polynomials to form the orthogonal check-sum sets, and fed to k_0 majority gates (threshold circuits). Hence, the number of inputs to each majority gate is the same as the number of terms in each generator polynomial, that is, $J = d - 1$.

The output of the ith threshold circuit, which is a zero when more than

$$\begin{Bmatrix} J/2, & J \textit{ even} \\ (J - 1)/2, & J \textit{ odd} \end{Bmatrix}$$

of the inputs are zeros, and one otherwise, is added to the ith data bit in the $(m - 1)$th block in the parity regenerator (reencoder) circuit. This correction bit is also used to invert each bit of the syndrome that is connected to the ith threshold circuit. After correcting all the errors, the bits are shifted out so that the next block can be processed.

With this introduction we can proceed with the design of the code for microwave radios.

1.2 Design model

The self-orthogonal convolutional code applied here, as stated earlier, is a rate 11/12 code. That is, the encoder appends one parity bit to the end of each block of 11 information bits to form a 12-bit coded block. The code constraint length, that is, the number of bits checked by every parity bit, is $N = 1716$. Therefore, the maximum term in the code generator polynomial is of the order of 142, which is the length of the encoder and syndrome shift registers. The code generator polynomial set is[1]

$$\begin{aligned}
G_1 &= 1 + D + D^3 + D^7 \\
G_2 &= 1 + D^8 + D^{24} + D^{47} \\
G_3 &= 1 + D^9 + D^{55} + D^{73} \\
G_4 &= 1 + D^{11} + D^{92} + D^{128} \\
G_5 &= 1 + D^{22} + D^{43} + D^{83} \\
G_6 &= 1 + D^{10} + D^{101} + D^{114} \\
G_7 &= 1 + D^{59} + D^{76} + D^{103} \\
G_8 &= 1 + D^{42} + D^{122} + D^{142} \\
G_9 &= 1 + D^5 + D^{57} + D^{95} \\
G_{10} &= 1 + D^{87} + D^{113} + D^{132} \\
G_{11} &= 1 + D^{66} + D^{99} + D^{133}.
\end{aligned} \tag{1}$$

We use these polynomials to form the encoder shift register shown in Fig. 1 and the syndrome register of the decoder in Fig. 2. The details of shift register tap connections for this type of code are shown in Fig. 3. The blocks in the block diagrams in Fig. 1 through 3 will be described in the following paragraphs.

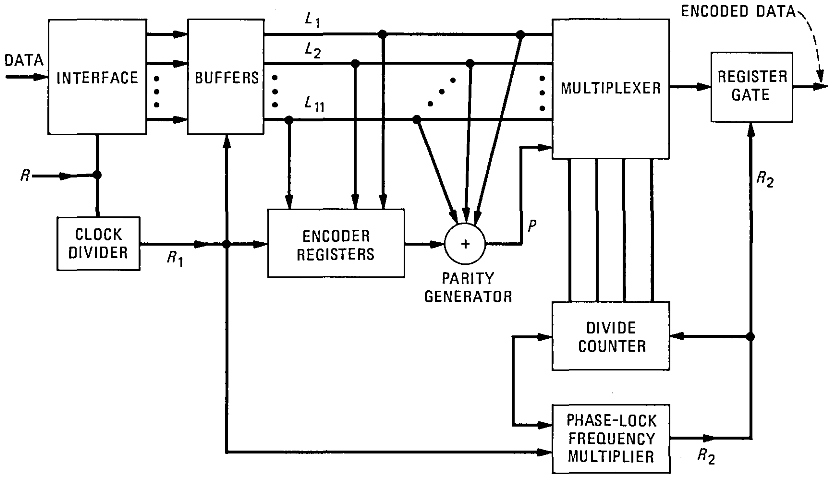


Fig. 1—Block diagram of encoder.

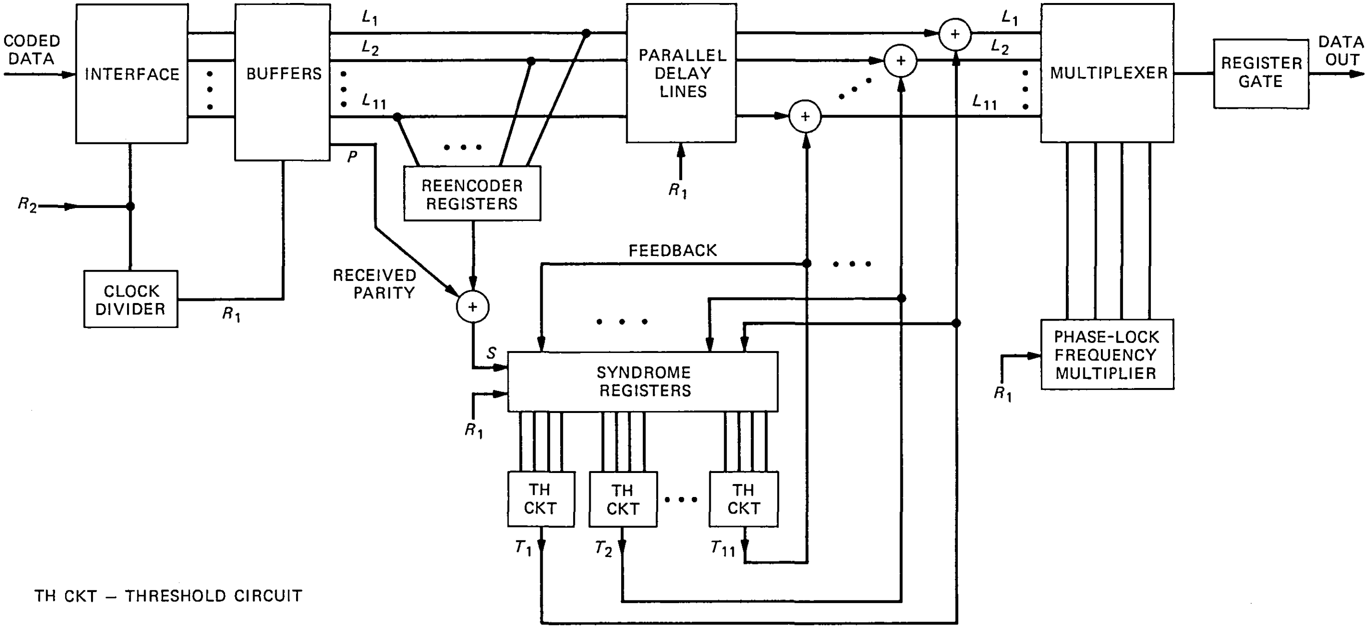


Fig. 2—Block diagram of decoder.

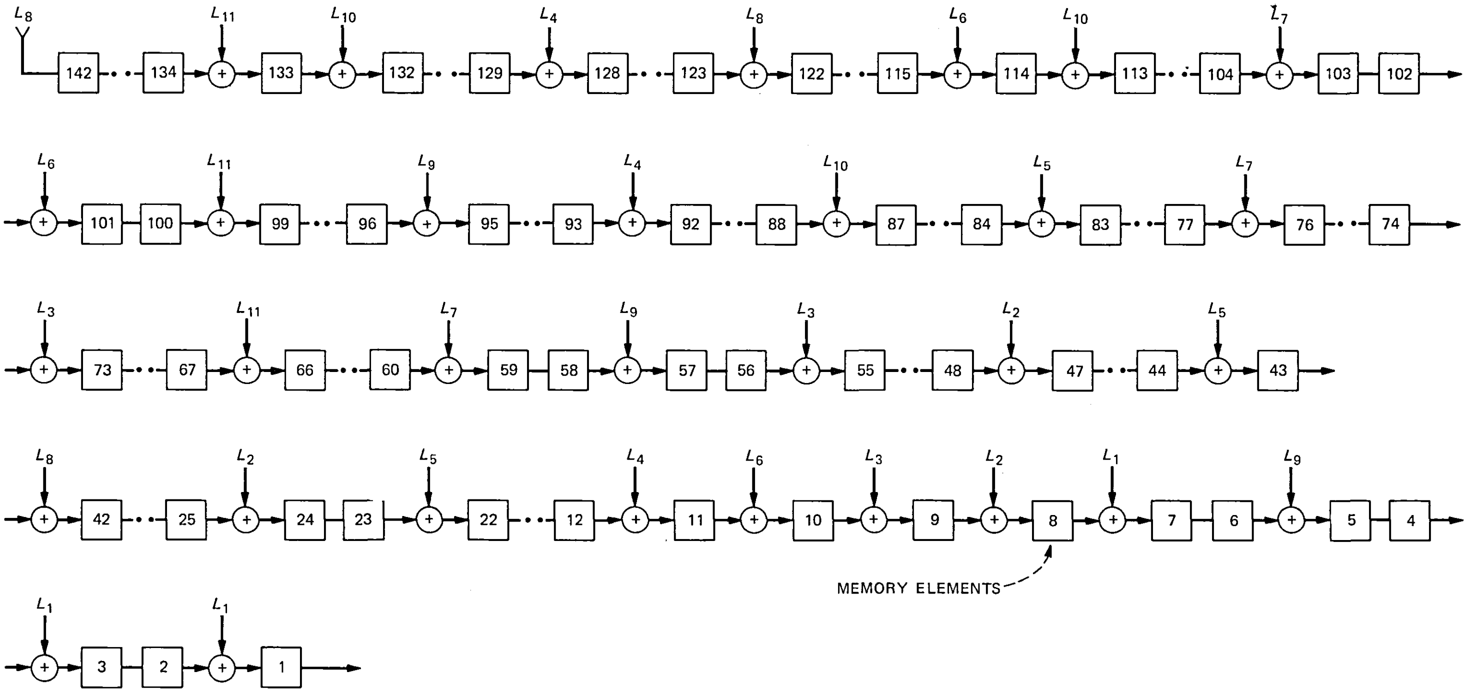


Fig. 3—Tap connection details.

The encoder circuit shown in Fig. 1 takes a serial data stream at a clock rate of R, and, via a Serial-to-Parallel (S/P) converter, the stream is converted to 11 parallel streams, each clocked at the rate $R_1 = R/11$. To realize the S/P converter circuit, an 11-bit serial-in/parallel-out shift register at the rate R was used. The clock conversion is performed by the divide-by-eleven circuit shown in Fig. 1. The timing alignment of the data streams is done through a set of flip-flops clocked at R_1. The encoder shift register consists of 142 delay elements, along with 32 exclusive OR gates placed at locations determined by the code generator polynomials. The delay elements were realized by using 8-bit Transistor–Transistor Logic (TTL) shift register Integrated Circuits (ICs). The generated parity and 11 data bits are then combined through a TTL multiplexer addressed by a 12-bit counter. The multiplexer has to be clocked at $R_2 = (12/11)R$, or more simply, $R_2 = 12R_1$. Hence, a clock multiplier was needed to generate the 12th harmonic of R_1. A digital phase-lock frequency multiplier was designed for this purpose. The circuit is shown in Fig. 4. It consists of a phase detector IC, a low-pass loop filter, a voltage-controlled multivibrator, and a divide-by-twelve counter. To generate the R_2 clock, a harmonic filtering method was tried first and discarded in favor of the phase-lock frequency multiplier. As stated earlier, the clock signal generated by the phase-lock frequency multiplier is used to address the multiplexer (mux) IC and the output of this is reclocked through a single flip-flop by the R_2 clock.

A block diagram of the decoder is shown in Fig. 2. The received coded data at rate $R_2 = (12/11)R$ is passed through an S/P converter to obtain 12 parallel bit streams. To alleviate the encoder/decoder synchronization problem in the experimental model, the R_2 clock generated on the encoder board was hard-wired to the decoder front-end circuit, where a divide-by-twelve counter was used to generate $R_1 = R_2/12$. This clock, in turn, is used to time align the 12 data streams through a set of flip-flops. The reencoder circuit is identical to the encoder shift register shown in Fig. 1. The generated parity and received parity bits are added through an exclusive OR gate to form

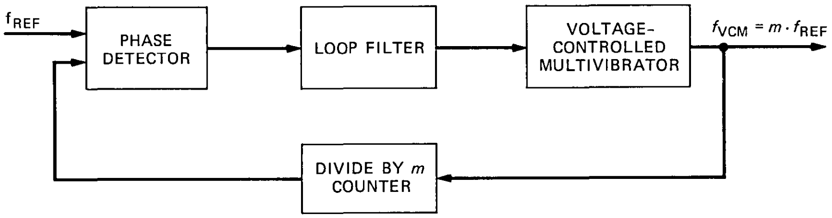


Fig. 4—Phase-lock frequency multiplier.

the syndrome bits. The syndrome register is set up the same way as the reencoder shift register. The syndrome register outputs are fed to a set of 11 threshold circuits. Each threshold circuit, as explained earlier, performs a majority logic vote; that is, if more than half the inputs to each circuit are binary ones, the output of that circuit will be a one, otherwise it will output a zero. Each binary one at the threshold circuit output indicates an error on a particular data line being checked by that threshold circuit. Because the code implemented here is a double error-correcting code, there are four inputs to each threshold circuit that operate on the orthogonal check bit set.

To have the error correction correctly performed, the 12 data lines have to be delayed by one syndrome register length. This can be done by using two RAMs in parallel; however, in this experimental model we used a set of shift register ICs, each containing a 128-bit delay and clocked at R_1 to acquire the delay needed. The error indicators are then modulo-2 added to the proper data bits, and the corrected data lines are multiplexed by a similar approach, as explained in the encoder circuit description. Again, a phase-lock frequency multiplier is used to provide the information clock rate at the decoder output. The multiplexer output is reclocked through a single flip-flop at the clock rate R. In addition, the error indicators are used to remove the effect of corrected errors from the check bits entering the syndrome register.

The fact that the main encoding/decoding operations here are done at a relatively low speed, because of the input serial-to-parallel conversions, makes this type of codec attractive for high-speed data transmission. Next we discuss the test procedure.

II. TESTING

To check the performance of the encoder and decoder, a test drawer was designed and built. The test drawer consisted of the encoder/decoder circuits, a thermal noise source, a summing amplifier, two attenuators, and a switch, as shown in Fig. 5. Functionally, this test drawer was to measure the error rate at the input/output of the codec.

The BER test set used here produces a random bit stream, representing the information bits, which is then encoded through the encoder. Noise is then added to the encoded signal before it enters the decoder. In order to measure both the input and output error probabilities with the same BER test set, we used the following method. The syndrome register flip-flops are set during the normal course of error correction and the decoder output stream closely resembles the encoder input data stream. However, if we reset the decoder syndrome register, the decoder error-correcting function is blocked. Consequently, the decoder outputs the unmodified noisy bit stream. There-

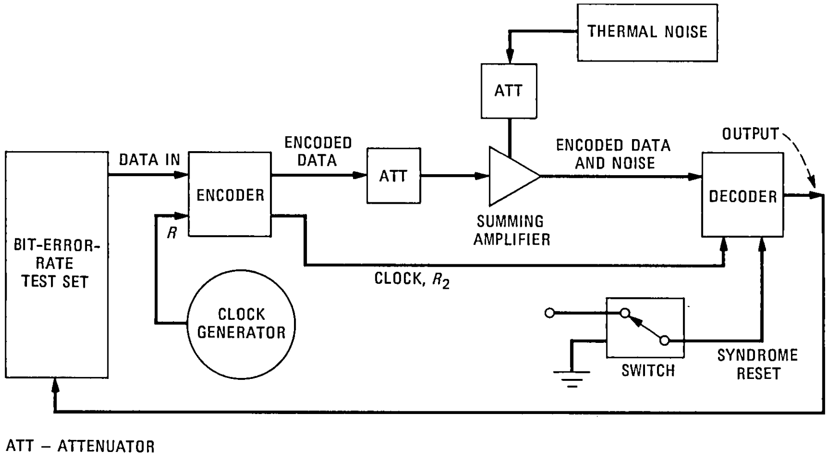


Fig. 5—Block diagram of test setup.

fore, the BER test set will measure the input error probability. The syndrome register resetting operation was done by the switch shown in Fig. 5. The noise generator was a standard thermal noise source. The step attenuators were to vary the signal and noise power in order to display different error rates at the input.

The test was performed at an input rate of 10 Mb/s. However, because of the serial-to-parallel operation at the encoder/decoder input, up to a 250-Mb/s data rate can be handled by this codec, using standard TTL integrated circuits.

III. CALCULATED AND MEASURED RESULTS

An approximate expression on the performance of the self-orthogonal convolutional codes for low channel error rates is presented in Ref. 2, and more details can be found in Ref. 3. The result is an asymptotic, upper bound on the bit error probability of the decoded bit. The bound is given by

$$P_b \lesssim \frac{1}{NR_c} \sum_{i=t+1}^{N} \binom{N}{i} p^i (1-p)^{N-i}, \qquad (2)$$

where

$\lesssim$ = asymptotically (in N_0)
N_0 = white noise spectral height
P_b = bit error probability after decoding
N = constraint length = 1716
R_c = code rate = 11/12
t = number of bit errors corrected per constraint length = 2
p = input bit error probability.

The expression for P_b in (2) is for any particular decoded bit in the first group in a constraint length, under the assumption either that (1) decoding is *direct* (without feedback syndrome correction) and the immediately preceding constraint span was free of decoder input errors, or (2) decoding is *with feedback* and the immediately preceding constraint span was free of decoder *output* errors. It happens to be valid, in general, only by virtue of the fact that the effects of prior history of the decoder are outweighed by the excess probability—included in eq. (2)—of those triple or higher weight input error patterns that do not cause output errors. This bound for the code implemented here is shown as one of the curves in Fig. 6.

As stated earlier, the set of error indicators can be used as a feedback to clean up the syndrome register. To investigate how much improve-

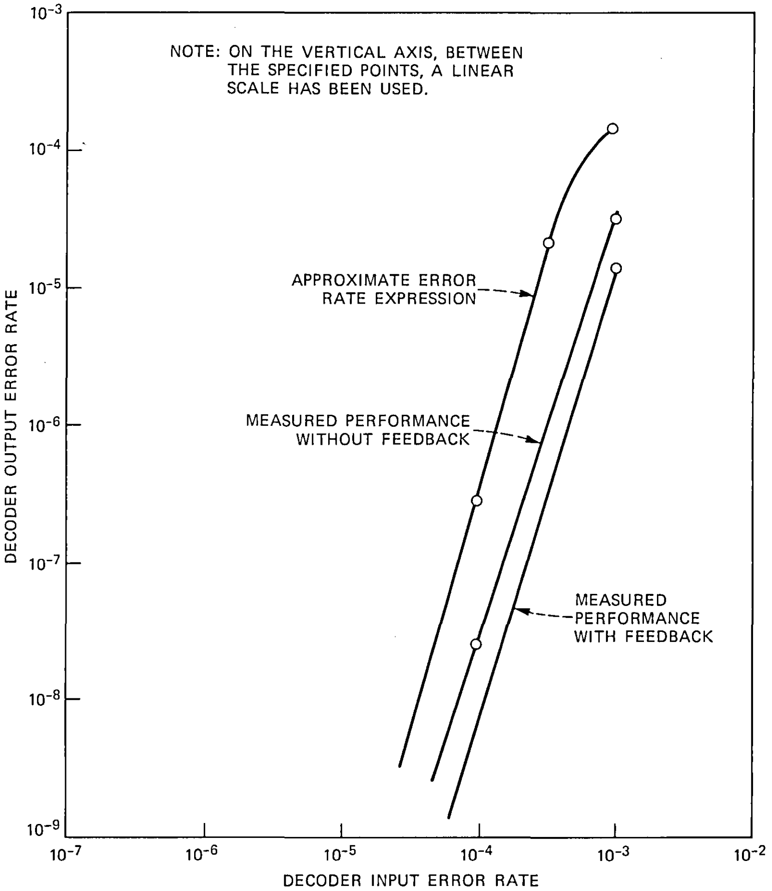


Fig. 6—Performance of a rate 11/12 codec.

ment is achieved by this operation, the output error probability measurements were taken for two cases: with and without syndrome error correction. These results are also shown in Fig. 6. As one can observe, having the set of feedback error indicators connected improves performance. In this case, the output probability of error is improved by almost one half of an order of magnitude at an input error rate of 10^{-4} by having the feedback links in Fig. 2 connected.

As stated earlier, the difference between the approximate bound and the measured performance of this double-error-correcting code can be due to the fact that a self-orthogonal, double-error-correcting, convolutional codec can correct many—triple, quadruple and longer—error patterns. However, the bound in eq. (2) only takes into account double-error correction.

Note that the BER test setup simulated only thermal noise effects. The possible effects of modem implementation and other nonthermal effects on a real channel need to be characterized. If these effects merely increase the decoder input error rate for a given bit energy to noise density (E_b/N_0), but maintain a pure Poisson distribution of those errors, then the output BER versus input BER results of the decoder test will still apply. Conversely, if the other effects cause significant departure from a Poisson arrival of decoder input errors, then the output BER performance versus input BER performance of the decoder will degrade from the test results previously described. In particular, if there is a tendency towards error clustering, a degradation could occur. For instance, clusters of three errors or more in a constraint span that are more frequent than that predicted by a Poisson model could be a source for performance degradation.

For example, in a gray coded 16-QAM modem, even a very small residual phase offset error in detection would significantly increase the probability of 2 bit errors in a 4-bit baud. Then both errors are prone to erroneous decoding if a third input error happens to occur nearby. When the objective is a 10^{-10} output BER, even a very slight effect of this sort can quickly result in an order-of-magnitude degradation in the codec performance.

IV. CONCLUSIONS

This paper has described the design, testing, and performance of a rate 11/12 self-orthogonal convolutional codec that meets the BER performance objective of M-QAM radio systems. The objective was to convert a bit error rate of 10^{-6} to an equivalent error rate of 10^{-10} or better. The measured error rate is well below 10^{-10} at an input error rate of 10^{-6}.

V. ACKNOWLEDGMENT

The author wishes to thank P. Dollard, P. J. McLane, and C.-E. Sundberg for the careful review of the manuscript and their useful suggestions.

REFERENCES

1. W. W. Wu, "New Convolutional Codes—Part I" IEEE Trans. Commun., *COM-23,* No. 9 (September 1975), pp. 942–56.
2. M. Kavehrad, "Implementation of a Self-Orthogonal Convolutional Code Used in Satellite Communications," IEEE, Trans. Elec. Circuits and Syst., *3,* No. 3 (May 1979), pp. 134–8.
3. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications,* New Jersey: Prentice Hall, 1983, p. 413.

AUTHOR

Mohsen Kavehrad, B.S. (Electrical Engineering), 1973, Tehran Polytechnic Institute; M.S. (Electrical Engineering), 1975, Worcester Polytechnic Institute; Ph.D. (Electrical Engineering), 1977, Polytechnic Institute of New York; Fairchild Industries, 1977–1978; GTE, 1978–1981; AT&T Bell Laboratories, 1981—. At AT&T Bell Laboratories Mr. Kavehrad is a member of the Communications Methods Research Department at Crawford Hill Laboratory. His research interests are digital communications and computer networks. Technical Editor, IEEE Communications Magazine; Chairman, IEEE Communications Chapter of New Hampshire (1984); Member, IEEE, Sigma Xi.

AT&T Technical Journal
Vol. 64, No. 7, September 1985
Printed in U.S.A.

Criteria for the Global Existence of Functional Expansions for Input/Output Maps*

By I. W. SANDBERG[†]

(Manuscript received March 8, 1985)

Much has been learned in recent years about the existence, determination, and properties of power-series-like expansions for expressing a nonlinear system's outputs in terms of its inputs. In particular, the existence and local convergence of expansions, and of certain "associated expansions," for important large classes of systems are now well established. While the focus of attention has been on questions such that the *size* of the inputs for which convergence is guaranteed is not the main issue, some related material has appeared that bears on the problem of determining the extent of the region of convergence. The result most closely related to this paper is a recent theorem that gives necessary and sufficient conditions under which f^{-1} has a generalized power-series expansion when f is an invertible locally-Lipshitz map between certain general subsets of two complex Banach spaces. In applications involving nonlinear models, ordinarily only *real* spaces of inputs and outputs are of direct interest. A "complexification" involving a certain solvability condition in complex spaces has to be able to be carried out to use the theorem referred to above. This paper reports on pertinent general results concerning invertible maps between subsets of real Banach spaces, with their complex extensions, and with generalized power-series expansions in both real and complex spaces. It focuses on questions concerning expansions for inverses of maps defined in real spaces. The results show that for a very large class of systems that have input/output maps, the ability to complexify is not just a useful sufficient

* This paper was presented at the Conference on Information Sciences and Systems, Johns Hopkins University, March 27–29, 1985.
[†] AT&T Bell Laboratories.

condition for expandability, but is in fact the *key* condition for an input/output map to be representable by a generalized power-series expansion.

I. INTRODUCTION

Convolution operator input/output representations for linear systems are well understood and are widely used. With regard to corresponding representations for nonlinear systems, much has been learned in recent years about the existence, determination, and properties of power-series-like expansions for expressing a system's outputs in terms of its inputs (see, for example, Refs. 1–7). In particular, the existence and local convergence of expansions, and of certain "associated expansions,"[3] are now well established for important large classes of systems.

While the focus of attention in Refs. 1 through 6 has been on questions such that the *size* of the inputs for which convergence is guaranteed is not the main issue, some related material has appeared that bears on the problem of determining the extent of the region of convergence. The result most closely related to this paper is a theorem in Ref. 7 which gives necessary and sufficient conditions under which f^{-1} has a generalized power-series expansion (in the sense of our Section 2.1) when f is an invertible locally-Lipshitz map between certain general subsets of two complex Banach spaces. Another theorem in Ref. 7 yields an algorithm for obtaining the expansion whenever it exists, and these two theorems are used therein to prove results concerning a certain system model considered in Ref. 2 and in earlier papers.

In applications involving nonlinear models, ordinarily only *real* spaces of inputs, outputs, and intermediate signals are of direct interest. A "complexification" involving the existence of a certain inverse map defined on a complex space has to be able to be carried out to use the theorems in Ref. 7. One of the main applications of the results in this paper is a proof that in an important general setting this complexification condition is always met when certain invertibility and expandability conditions are satisfied in the underlying real space. As a consequence, for a very large class of systems that have input/output maps, the ability to complexify emerges as the *key* condition for an input/output map to be representable by a generalized power-series expansion. (Under certain reasonable assumptions these expansions reduce to Volterra-like series.)[2,8]

To be more explicit, models of the kind mentioned above are characterized by five operators: a nonlinear operator N, and four linear operators a, b, c, and d. They have an input v and an output w, which belong to a space X of functions. Here X is taken to be a real Banach space; a, b, and c are assumed to be bounded maps of X into X, and we suppose that N is defined on all of X and takes X into X. One has $w = dv + bN(I - cN)^{-1}av$ (I the identity map on X) subject to some natural qualifications, from which it is clear that the study of such models* often involves the study of maps of the form $(I - cN)$. The X of particular interest to us is the space of real Lebesgue-measurable, n-vector-valued functions x defined on $[0, \infty)$, with the norm in X given by $\| x \| = \max_j \sup_{t \geqslant 0} | x_j(t) |$, where $x_j(t)$ is the jth component of $x(t)$.

Let A_0 and A be subsets of X such that $(I - cN)$ restricted to A_0 is an invertible map of A_0 onto A. Assume that both A_0 and A are open sets, and that A contains the zero element of X. Under these conditions, w is well defined for each v such that $av \in A$, the zero function is an allowed input, and the set of allowed inputs is open. The question that we ask is this: With $(I - cN)^{-1}$ the inverse of the restriction of $(I - cN)$ to A_0, assumed to be continuous, when is it true that $(I - cN)^{-1}u$ has a generalized power-series expansion that converges for $u \in A$? When it *is* true, the map from v to w has an expansion that converges whenever $av \in A$, assuming (and this is frequently very reasonable) that N is such that the existence of the expansion for $(I - cN)^{-1}u$ implies the existence of an expansion for $N(I - cN)^{-1}u$; see, for example, Corollary 1 in Appendix A or Theorems 1 and 7.

Theorem 2 in Section II provides an answer to the question, under the assumption that N has an extension into a complex space $\mathscr{B}$ associated with X, with this extension a certain type of globally convergent generalized power series. For the X of particular interest, this assumption is a reasonable one, and the corresponding $\mathscr{B}$ turns out to be just the natural complex associate of X. The answer given by Theorem 2 is that there must be two open subsets V_0 and V of $\mathscr{B}$ such that: $A \subset V$ (meaning that $u + i0 \in V$ for each $u \in A$; see Section 2.1), $A_0 \subset V_0$, V is a "star" in the sense that $zq \in V$ when $q \in V$ and z is a scalar such that $|z| \leqslant 1$, and the map $(I - cN)$ extended into $\mathscr{B}$ (see Section 2.2), and restricted to V_0, must be a homeomorphism of V_0 onto V, with the inverse of the restriction of the extended $(I - cN)$ locally Lipshitz on V. While this necessary and sufficient condition† may look complicated at first glance, its interpretation is straightforward: $(I - cN)^{-1}u$ has a power-series expansion that converges for $u \in A$ if and only if the equation $x - cNx = u$, when

* There is an error in the corresponding equation in Ref. 7, where B in (11) should be replaced with BN. This does not change the conclusion drawn there from Theorem 3; see, for example, our Theorem 7.

† The *sufficiency* of this type of condition is discussed in Ref. 7, Section 2.4.2. Also, with regard to the system model in Ref. 7 (p. 84), note that the existence of an expansion for w (in terms of v) implies the existence of an expansion for y and thus for x if, for example, B is the identity operator.

extended into the complex space $\mathscr{B}$, is, so-to-speak, uniquely locally-Lipschitz solvable in some open subset of $\mathscr{B}$ containing the points of A_0 for all right sides belonging to V, where V is any open star in $\mathscr{B}$ that contains the elements of A. (See Section 2.3.1. In this connection, notice that an open ball centered at the origin is an example of a star.)

As is suggested by the application described above, the results in this paper are concerned with invertible maps between subsets of real Banach spaces, with their complex extensions, and with generalized power-series expansions in both real and complex spaces, with the focus on questions concerning global expansions for inverses of maps defined in real spaces. Preliminaries are introduced in Section 2.1, and Sections 2.2 through 2.6 contain the paper's principal results.

There are several natural applications of the material in Section II other than the one already discussed. For example, consider again the five-operator model described above, and assume that the assumptions introduced are met. Assume in addition that an expansion representation for $(I - cN)^{-1}u$ does exist for $u \in A$. Suppose that this expansion also converges for $u \in B$, where B is some open subset of X for which $A \subset B$. Theorem 3 shows that then the map from v to w is in fact both well defined and has a generalized power-series expansion for $av \in B$.

It will become clear that the theorems in Section II are considerably more general than the applications discussed above are able to illustrate. For instance, they bear on cases in which the underlying function space is a set of functions of more than one independent variable. Also, in Section 2.5 corresponding results are given for certain implicitly defined maps. These latter results are useful in, for example, studies of globally convergent generalized power-series expansions for solutions of differential equations.

II. COMPLEX EXTENSIONS AND EXPANSION REPRESENTATIONS

2.1 Preliminaries

Throughout the paper X denotes a real Banach space. We associate with X (see Ref. 9, p. 312 and Ref. 10, p. 665) a complex Banach space $\mathscr{B}$ defined as follows: the elements of $\mathscr{B}$ are ordered pairs (x_1, x_2) of elements of X, addition and multiplication obey

$$(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2)$$

$$(\alpha + i\beta)(x_1, x_2) = (\alpha x_1 - \beta x_2, \alpha x_2 + \beta x_1),$$

and the norm of an element of $\mathscr{B}$ is given by

$$\| (x_1, x_2) \| = \sup_{\|\xi\|=1} [\xi^2(x_1) + \xi^2(x_2)]^{1/2},$$

where ξ denotes a general, real, bounded linear functional on X. We sometimes use $x_1 + ix_2$ to denote an element (x_1, x_2) of $\mathscr{B}$.

The map $x \longrightarrow x + i0$ of elements of X into elements of $\mathscr{B}$ isometrically* imbeds X into the complex space $\mathscr{B}$. In particular, in this sense, $\mathscr{B}$ is a complex extension of X. For example, if X is the space of bounded, Lebesgue-measurable, real n-vector-valued functions x defined on $[0, \infty)$, with $\| x \| = \max_j \sup_t | x_j(t) |$, then the elements v of $\mathscr{B}$ are bounded, Lebesgue-measurable, complex n-vector-valued functions defined on $[0, \infty)$, and (see Appendix B) one simply has $\| v \| = \max_j \sup_j | v_j(t) |$.

A *star* in $\mathscr{B}$ means a subset S of $\mathscr{B}$ such that $zv \in S$ for $v \in S$ and any complex scalar z with $|z| \leqslant 1$. A subset S of $\mathscr{B}$ is *c-convex* if for any bounded open set Δ of complex numbers, we have $(v + \Delta u) \subset S$ whenever $(v + \Gamma u) \subset S$, where Γ is the boundary of Δ.

In the paper, X_0 denotes a second real Banach space and $\mathscr{B}_0$ stands for its complex extension. We allow the possibility that $X_0 = X$.

Now let Y and W be any two Banach spaces, both real or both complex.

Given any positive integer m, by an m-linear map q from Y^m into W we mean that $q(y_1, \cdots, y_m)$ is linear (i.e., additive and homogeneous) separately in each y_j. Such a map is *symmetric* if $q(y_1, \cdots, y_m)$ is symmetric in the variables $y_1, \cdots, y_m$. A map h from Y into W is called a homogeneous polynomial of degree m if there exists an m-linear q from Y^m to W such that $h(y) = q(y, \cdots, y)$ for all y.† A homogeneous polynomial of degree zero is a constant map.

For S a subset of Y, let $\mathscr{P}(S, W)$ denote the set of all maps p from S into W such that there are homogeneous polynomials h_m of degree m $(m = 0, 1, \cdots)$ from Y to W, with the properties that $\sum_{m=0}^{\infty} h_m(s)$ converges in W for each $s \in S$, and

$$p(s) = \sum_{m=0}^{\infty} h_m(s), \quad s \in S. \tag{1}$$

The set $\mathscr{P}(S, W)$ is, of course, a set of maps p that admit a generalized power-series expansion in the sense indicated. If S contains an open ball in Y centered at the origin, then the expansion (1) for any $p \in \mathscr{P}(S, W)$ is unique in the sense that if

$$p(s) = \sum_{m=0}^{\infty} g_m(s), \quad s \in S,$$

with each g_m a homogeneous polynomial of degree m, then $g_m = h_m$ for all m (see Ref. 11, p. 174 and Ref. 1, Section 2.7).

Finally, we say that p belongs to $\mathscr{P}_F(S, W)$ if $p \in \mathscr{P}(S, W)$ and for

* By the Hahn-Banach theorem, $\| x \| = \sup\{\xi(x) \colon \| \xi \| = 1\}$.

† The same class of maps is obtained if "m-linear" is replaced with "symmetric m-linear."

each positive m there is a *continuous* symmetric m-linear q_m from Y^m into W such that $h_m(s) = q_m(s, \cdots, s)$ for all s. In particular, then each h_m is bounded in the sense that there is a constant $\rho_m > 0$ such that $\| h_m(s) \| \leqslant \rho_m \| s \|^m$ for all m and s, and every h_m is Fréchet differentiable on Y.

2.2 *Inverse maps and necessary conditions for the existence of series representations*

Throughout this section, and in Sections 2.3, 2.4, and 2.6, f is a map from X_0 into X, A and A_0 are open subsets of X and X_0, respectively, with $0 \in A$, f restricted to A_0 is a homeomorphism of A_0 onto A, and g: $A \rightarrow A_0$ is the inverse of the restriction of f. It is assumed that there is an $f^* \in \mathscr{P}_F(\mathscr{B}_0, \mathscr{B})$ such that $f(x) = f^*(x + i0)$ for $x \in X_0$. [Of course, by $f(x) = f^*(x + i0)$ we mean that $f(x) + i0 = f^*(x + i0)$.]

The following extension theorem is this paper's main result.

Theorem 1: If g has a power-series representation in the sense that $g \in \mathscr{P}(A, X_0)$, then there are open sets V and V_0 in $\mathscr{B}$ and $\mathscr{B}_0$, respectively, together with a map $g^*: V \rightarrow V_0$ such that V is a c-convex star, $A \subset V$, $A_0 \subset V_0$, and

1. *the restriction of f^* to V_0 is a homeomorphism of V_0 onto V with inverse g^**
2. $g^* \in \mathscr{P}_F(V, \mathscr{B}_0)$
3. $g(x) = g^*(x + i0), x \in A$.

2.2.1 *Proof of Theorem 1*

Two lemmas are used in the proof. The first of these follows.

Lemma 1: Let D be an open subset of X with $0 \in D$, let $h \in \mathscr{P}(D, X_0)$, and assume that h is continuous on D. Then there are an open c-convex star $Z \subset \mathscr{B}$ and a map h^* from Z to $\mathscr{B}_0$ such that $D \subset Z$, $h^* \in \mathscr{P}_F(Z, \mathscr{B}_0)$, and $h(x) = h^*(x + i0)$ for $x \in D$.

Proof of Lemma 1: We have

$$h(s) = \sum_{m=0}^{\infty} h_m(s), \quad s \in D, \tag{2}$$

where each h_m is a homogeneous polynomial of degree m. Let q_m be the unique symmetric m-linear map such that $h_m(s) = q_m(s, \cdots, s)$ for $s \in X$ and positive m (see Ref. 12, pp. 762–3). Let $h_0^* = h_0 + i0$, and define h_m^*: $\mathscr{B} \rightarrow \mathscr{B}_0$ for each $m \geqslant 1$ by

$$h_m^*(x_1 + ix_2) = \sum_{k=0}^{m} i^{(m-k)} \binom{m}{k} q_m(x_1, \cdots, x_1, x_2, \cdots, x_2), \tag{3}$$

with kx_1's and $(m - k)x_2$'s on the right side. It is not difficult to verify that the h_m^* are homogeneous polynomials of degree m (see Ref. 9, p.

313 and Ref. 13, p. 71). By Theorem 5.7 of Ref. 13 there is an open subset Z_1 of $\mathscr{B}$ such that $D \subset Z_1$ and the series

$$\sum_{m=0}^{\infty} h_m^*(v) \tag{4}$$

converges for $v \in Z_1$. Since D is open and h is continuous, it follows (see Ref. 13, Theorems 4.4 and 6.6) that h is analytic in D in the sense of Ref. 13, p. 75. Thus, using the hypothesis that $0 \in D$, h has a power-series expansion valid in a neighborhood of the origin, with the terms in the expansion continuous homogeneous polynomials. At this point the uniqueness result mentioned in Section 2.1 shows that the h_m in (2) are continuous.

By the continuity of the h_m, the q_m in (3) are continuous. Using (3) and the fact that $\| x_1 \| \leqslant 1$ and $\| x_2 \| \leqslant 1$ are implied by $\| x_1 + ix_2 \| \leqslant 1$ [see (7), below], we see that each h_m^* is bounded in the ball $\| x_1 + ix_2 \| \leqslant 1$. This shows (see Ref. 12, Theorem 26.2.4) that the h_m^* are continuous.

Let Z denote the interior of the region of convergence of the series (4), and let $h^*(v)$ be the sum (4) for any $v \in Z$. Obviously $Z_1 \subset Z$. Since the h_m^* are continuous, it follows (see Ref. 12, Theorem 26.6.1) that Z is a c-convex star. Since it is clear that $h_m^*(x + i0) = h_m(x)$ for $x \in X$, the proof of the lemma is complete.

Continuing with the proof of the theorem, by Lemma 1 there are an open c-convex star $V \subset \mathscr{B}$ and a map $g^*: V \rightarrow \mathscr{B}_0$ such that $A \subset V$, $g^* \in \mathscr{P}_F(V, \mathscr{B}_0)$, and $g(x) = g^*(x + i0)$ for $x \in A$.

We now turn to our second lemma.

Lemma 2: Let h be a Fréchet-differentiable map (*Ref.* 14, p. 149) *from an open connected subset D of* $\mathscr{B}$ *into* $\mathscr{B}_0$. *Assume that there is a point p in D and an open ball Q in X centered at the origin such that* $(p + Q) \triangleq \{s \in \mathscr{B}: s = p + (q + i0), q \in Q\} \subset D$ *and h maps* $(p + Q)$ *into the origin in* $\mathscr{B}_0$. *Then h vanishes everywhere in D.*

Proof of Lemma 2: Since h is Fréchet differentiable in a neighborhood of p, there are (see Ref. 12, Theorems 3.16.2 and 26.3.5) homogeneous polynomials H_m of degree m ($m = 1, 2, \cdots$) and a $\sigma > 0$ such that $s \in D$ and

$$h(s) = \sum_{m=1}^{\infty} H_m(s - p)$$

when $\| s - p \| < \sigma$. Thus, for some positive $\rho < \sigma$, we have

$$\sum_{m=1}^{\infty} H_m(q + i0) = \theta$$

for $q \in \{x \in X: \| x \| < \rho\}$, where θ is the zero element of $\mathscr{B}_0$. It easily follows (see Ref. 1, Section 2.7) that $H_m(x + i0) = \theta$ for each m and all $x \in X$.

Consider $H_m(x_1 + ix_2)$ with m, x_1, and x_2 arbitrary, and let P_m denote the polar form (see Ref. 12, pp. 762–3) associated with H_m. We see that $H_m(x_1 + ix_2)$ can be written as a finite sum of terms of the form $cP(y_1 + i0, \cdots, y_m + i0)$, in which $c \in \{\pm 1, \pm i\}$ and each y_j is either x_1 or x_2. On the other hand, $P_m(y_1 + i0, \cdots, y_m + i0)$ can be expressed (see Ref. 9, p. 306) as

$$(m!)^{-1} \sum_{\epsilon_1, \cdots, \epsilon_m = 0}^{1} (-1)^{m-(\epsilon_1 + \cdots + \epsilon_m)} H_m[\epsilon_1(y_1 + i0) + \cdots + \epsilon_m(y_m + i0)].$$

Therefore, using $H_m(x + i0) = \theta$ for $x \in X$, one has $H_m(x_1 + ix_2) = \theta$. This shows that $h(s) = \theta$ for s in an open ball in D. Since D is connected and h is G-differentiable in the sense of Ref. 12 (pp. 109–10), it follows (see Ref. 12, Theorem 3.16.4) that $h(s) = \theta$ throughout D, as claimed. We now return to the proof of the theorem.

Let $E(v)$ denote $f^*[g^*(v)] - v$ $(v \in V)$. Since $f^* \in \mathscr{P}_F(\mathscr{B}_0, \mathscr{B})$ and $g^* \in \mathscr{P}_F(V, \mathscr{B}_0)$, f^* and g^* are Fréchet differentiable on $\mathscr{B}_0$ and V, respectively (see Ref. 12, Theorems 26.6.4 and 3.17.1). Thus, using a version of the chain rule for differentiating a composite map, E is Fréchet differentiable on V. In addition, the set V is connected (because it is a star), and we have

$$E(x + i0) = f^*[g^*(x + i0)] - (x + i0) = 0, \quad x \in A.$$

Choose any point $p_1 \in A$, and let Q be an open ball in X centered at the origin such that $(p_1 + Q) \subset A$. Let $p = (p_1 + i0)$, and observe that $E[p + (q + i0)] = 0$ for $q \in Q$. By Lemma 2 (with $\mathscr{B} = \mathscr{B}_0$), $E(v) = \theta$ (the zero element of $\mathscr{B}$) for $v \in V$. This gives

$$f^*[g^*(v)] = v, \quad v \in V. \tag{5}$$

Since V is connected and g^* is continuous on V, $g^*(V)$ is connected. The continuity of f^* implies that $f^{*-1}(V)$ is open. From (5) it is clear that $g^*(V) \subset f^{*-1}(V)$. Let V_0 denote the maximal connected subset (i.e., the component) of $f^{*-1}(V)$ that contains $g^*(V)$. Since $f^{*-1}(V)$ is open, so is V_0. The map f^* obviously takes V_0 into V.

Now let $F(w)$ denote $g^*[f^*(w)] - w$ $(w \in V_0)$. It follows from Lemma 2 and $F(x + i0) = 0$ for $x \in A_0$ that

$$g^*[f^*(w)] = w, \quad w \in V_0. \tag{6}$$

Since (5) and (6) hold, f^* restricted to V_0 is a homeomorphism of V_0 onto V. The observation that $A_0 \subset V_0$ because $A_0 = g(A) \subset g^*(V) = V_0$ completes the proof of the theorem.

2.3 Complex-solvability criteria for global expandability

Here we use Theorem 1 to obtain necessary and sufficient conditions for the global expandability of the map g introduced at the beginning of Section 2.2. With regard to our result, Theorem 2 (below), we say that a map h from an open subset D of $\mathscr{B}$ into $\mathscr{B}_0$ is *locally Lipschitz* on D if for each $a \in D$ there are a positive number c_a and an open ball $\beta_a \subset D$ centered at a such that $\| h(v_1) - h(v_2) \| \leqslant c_a \| v_1 - v_2 \|$ for v_1 and v_2 in β_a.

Theorem 2: We have $g \in \mathscr{P}(A, X_0)$ if and only if (i) there are open sets V and V_0 in $\mathscr{B}$ and $\mathscr{B}_0$, respectively, with V a star, $A \subset V$, and $A_0 \subset V_0$ such that the restriction of f^* to V_0 is a homeomorphism of V_0 onto V, with the inverse of the restriction of f^* to V_0 locally Lipschitz on V, and (ii) the spaces $\mathscr{B}$ and $\mathscr{B}_0$ are homeomorphic, in the sense that there is a linear homeomorphism of $\mathscr{B}$ onto $\mathscr{B}_0$.

Proof: The necessity of (i) follows from Theorem 1 and the observation that g^* in Theorem 1, which belongs to $\mathscr{P}_F(V, \mathscr{B}_0)$, is Fréchet differentiable, hence continuously Fréchet differentiable (see Ref. 7, Lemma 2), and thus locally Lipschitz. Similarly, the necessity of (ii) is a consequence of Theorem 1 and the fact that the conclusion of Theorem 1 implies (see Ref. 15, p. 175, Problem 6) that the F-derivative (i.e., the Fréchet derivative) of f^* at any point in V_0 is an invertible map of $\mathscr{B}_0$ onto $\mathscr{B}$.

On the other hand, if (i) and (ii) are met, then, using Lemma 1 of Ref. 7, the inverse H of the restriction of f^* to V_0 is F-differentiable [and thus G-differentiable (see Ref. 12, pp. 109–10)] on V. It follows that $H \in \mathscr{P}(V, \mathscr{B}_0)$ (Ref. 12, Theorems 3.16.2 and 26.3.4). Let the series for H be given by

$$H(v) = \sum_{m=0}^{\infty} H_m(v), \quad v \in V.$$

By the conditions on f, for any $x \in A$ there is a $y \in A_0$ such that $f^*(y + i0) = (x + i0)$. Therefore, using $H[f^*(v)] = v$ $(v \in V_0)$ and $A_0 \subset V_0$, we see that H takes A into A_0. Thus, since $f^*[H(v)] = v$ $(v \in V)$, we have

$$g(x) = \sum_{m=0}^{\infty} H_m(x + i0), \quad x \in A.$$

Of course $H_0(0 + i0) \in X_0$. We claim that for each positive m there is an m-linear map Q_m from X^m into X_0 such that $Q_m(x, \cdots, x) = H_m(x + i0)$, $x \in A$. Since the H_m are homogeneous polynomials, we need only show that each H_m maps X into X_0, and we do that as follows.

The norm in $\mathscr{B}_0$ has the property that $\| x_1 + ix_2 \| < \delta$ implies that $\| x_2 \| < \delta$, because, using the Hahn-Banach theorem,

$$\|x_2\| = \sup_{\|\xi\|=1} |\xi(x_2)| \leq \sup_{\|\xi\|=1} [\xi(x_1)^2 + \xi(x_2)^2]^{1/2} = \|x_1 + ix_2\|. \qquad (7)$$

Thus, the convergence of $\sum_{m=0}^{\infty} H_m(x + i0)(x \in A)$ implies that for each $x \in A$, the series $\sum_{m=0}^{\infty} KH_m(x + i0)$ converges in X_0, and that it converges to the zero element, where K is the map from $\mathscr{B}_0$ to X_0 defined by $x_2 = K(x_1 + ix_2)$. In particular, with β any open ball in A centered at the origin, one has $rx \in \beta$ and

$$0 = \sum_{m=0}^{\infty} KH_m(rx + i0) = \sum_{m=0}^{\infty} r^m KH_m(x + i0)$$

for $x \in \beta$ and $|r| < 1$. It follows (see Ref. 11, proof of Theorem 6) that $KH_m(x + i0) = 0$ for each m and any $x \in X$, showing that the H_m map X into X_0. This completes the proof.

2.3.1 Comments

Since the inverse image of an open set under a continuous map is open, we see that (i) is equivalent to the condition that there be an open subset S of $\mathscr{B}_0$ and an open star $V \subset \mathscr{B}$ with the following properties: $A_0 \subset S$, $A \subset V$, for each $v \in V$ there is a unique $w \in S$ that satisfies $f^*(w) = v$, and the map $v \rightarrow w$ is locally Lipschitz. This more sharply focuses attention on how machinery for proving existence, such as fixed-point techniques, might be used to establish expandability. A pertinent example can be found in Ref. 7, Appendix B. A simple, related additional example follows.

Let X be the space of real numbers, with the absolute value norm, and observe that the corresponding $\mathscr{B}$ is the usual space of complex numbers. Take $X_0 = X$, let $f(x) = x + x^3$ for all real x, and take A and A_0 to be $\{x: |x| < r\}$ and $f^{-1}(A)$, respectively, for some positive r. Notice that any $r > 0$ will do, and that our f^* is given by $f^*(z) = z + z^3$ for all complex z.

An easy contraction-mapping argument* shows that given $\rho \in (0, \sqrt{3}^{-1}]$ and any complex number a with $|a| < (\rho - \rho^3)$, there is a unique complex number z with $|z| < \rho$ such that $z + z^3 = a$, that z is real whenever a is, and that the map from a to z is locally Lipschitz. It follows from Theorem 2 that we have $g \in \mathscr{P}(A, X_0)$ for $r = r_0$, where $r_0 = 2(3\sqrt{3})^{-1}$.

Theorem 2 also can be used to show that g *does not* belong to $\mathscr{P}(A, X_0)$ if $r > r_0$: Suppose, for the purpose of obtaining a contradiction, that $g \in \mathscr{P}(A, X_0)$ for some $r > r_0$. Then for some V and V_0 as described in the theorem, f^* restricted to V_0 is a homeomorphism of V_0 onto V. By the proof of Theorem 2, the inverse g^* of f^* is differ-

* A good general source of information on the use of the contraction-mapping theorem is Ref. 16.

entiable. Using $g^*[f^*(z)] = z$ for $z \in V_0$, we have $g^{*\prime}[f^*(z)]f^{*\prime}(z) = 1$ $(z \in V_0)$, where "$\prime$" denotes the ordinary derivative. Since $f^{*\prime}(z) = 0$ at $z = z_0 \triangleq i(\sqrt{3})^{-1}$, V_0 cannot contain z_0. Using this fact and the continuity of g^*, it is not difficult to show that V cannot contain the point $f^*(z_0) = 2i(3\sqrt{3})^{-1}$. Since V is a star and $A \subset V$, $2(3\sqrt{3})^{-1} \notin A$, which is the contradiction sought. This finishes the discussion of the example.

Using Theorem 1, it follows at once that Theorem 2 remains true if the word "star" is replaced by "c-convex star."

The hypothesis concerning f^* at the beginning of Section 2.2 is equivalent to the condition that $f \in \mathscr{P}_F(X_0, X)$; see the proof of Lemma 1 and Ref. 13 (top of p. 75).

2.4 Convergence and the extent of invertibility

In this section we prove a result that shows, in particular, that if g is expandable on A, and if its expansion converges on a larger open set B, then there is a set B_0 that contains the points of A_0 such that f is in fact an invertible map of B_0 onto B.

Theorem 3: Let g belong to $\mathscr{P}(A, X_0)$, and let it have the generalized power-series representation

$$g(x) = \sum_{m=0}^{\infty} g_m(x) \tag{8}$$

for $x \in A$. Suppose that the right side of (8) converges in X_0 for $x \in B$, where B is an open subset of X such that $A \subset B$. Then there is an open subset B_0 of X_0 such that (i) $A_0 \subset B_0$, and f is a homeomorphism of B_0 onto B; and (ii) the inverse G of the restriction of f to B_0 has the representation

$$G(x) = \sum_{m=0}^{\infty} g_m(x), \quad x \in B.$$

Proof: Since $g \in \mathscr{P}(A, X_0)$ and g is continuous, $g \in \mathscr{P}_F(A, X_0)$ (see the proof of Lemma 1). Thus, by Ref. 13, Theorem 6.2, the function $h{:}B \rightarrow X_0$ defined by

$$h(x) = \sum_{m=0}^{\infty} g_m(x), \quad x \in B$$

is analytic in the sense of Ref. 13 and hence continuous. Using Lemma 1, there is an open connected set $W \subset \mathscr{B}$ and a Fréchet-differentiable map $h^*{:}\,W \rightarrow \mathscr{B}_0$ such that $B \subset W$ and $h(x) = h^*(x + i0)$ for $x \in B$.

By Lemma 2 and the hypothesis that $f[g(x)] = x$ $(x \in A)$, we find that

$$f^*[h^*(v)] = v, \quad v \in W. \tag{9}$$

Again using Lemma 2, and proceeding as in the proof of Theorem 1, one finds that

$$h^*[f^*(v)] = v, \quad v \in W_0, \tag{10}$$

where W_0 is the component of $f^{*-1}(W)$ that contains $h^*(W)$. Therefore, f^* is a homeomorphism of W_0 onto W.

We have, from (9), $f[h(x)] = x$ $(x \in B)$. Now let $B_0 = f^{-1}(B) \cap R_0$, where $R_0 = \{x \in X_0 : x + i0 = z, z \in W_0\}$. Since W_0 is open, R_0 is open in X_0. Thus, B_0 is an open subset of X_0, and from (10) one has $h[f(x)] = x$ for $x \in B_0$. This shows that f is a homeomorphism of B_0 onto B, with h the inverse of the restriction of f to B_0. Finally, using $A_0 \subset f^{-1}(A) \subset f^{-1}(B)$, and $A_0 = h(A) \subset h(B) \subset h^*(W) \subset W_0$ (which implies that $A_0 \subset R_0$), as well as the definition of B_0, it is clear that $A_0 \subset B_0$. This proves the theorem.

2.5 Results for implicit functions

Theorems along the lines of Theorems 2 and 3 are given here for maps that are defined implicitly in the sense of the implicit function theorem. In this section, X_1 stands for a third real Banach space, $\mathscr{B}_1$ denotes its complex extension in the sense of Secton 2.1, and $X_0 \times X$ and $\mathscr{B}_0 \times \mathscr{B}$ are product Banach spaces constructed from X_0 and X and $\mathscr{B}_0$ and $\mathscr{B}$, respectively.* We say that a map h defined on an open subset D of $\mathscr{B}$ into $\mathscr{B}_0$ is *Gâteaux differentiable* on D (see Ref. 12, pp. 109–10) if for each $v \in D$ and arbitrary $w \in \mathscr{B}$ the limit $\lim_{z \to 0} z^{-1}[h(v + zw) - h(v)]$ exists, in which z is a complex scalar.

As in Section 2.2, A is an open subset of X, with $0 \in A$. Here F is a map from $X_0 \times X$ into X_1 such that there is a continuous map $G: A \to X_0$ with the property that

$$F[G(x), x] = 0, \quad x \in A.$$

Assume that there is an $F^* \in \mathscr{P}_F(\mathscr{B}_0 \times \mathscr{B}, \mathscr{B}_1)$ such that $F(y, x) = F^*(y + i0, x + i0)$ for $(y, x) \in X_0 \times X$.

Theorem 4: $G \in \mathscr{P}(A, X_0)$ if and only if there is an open star $V \subset \mathscr{B}$, with $A \subset V$, and a continuous Gâteaux-differentiable map $G^*: V \to \mathscr{B}_0$ such that we have $G(x) = G^*(x + i0)$ $(x \in A)$ as well as

$$F^*[G^*(v), v] = 0, \quad v \in V. \tag{11}$$

Proof: First suppose that $G \in \mathscr{P}(A, X_0)$. By Lemma 1 there is an open star $V \subset \mathscr{B}$ and a map $G^* \in \mathscr{P}_F(V, \mathscr{B}_0)$ such that $A \subset V$ and $G(x) =$

* Except where indicated to the contrary, the choice of the norms in $X_0 \times X$ and $\mathscr{B}_0 \times \mathscr{B}$ is not important for our purposes. It would suffice to let the norm $\|\cdot\|$ in $X_0 \times X$ be given by $\|(x_0, x)\| = \max(\|x_0\|, \|x\|)$, and similarly for $\mathscr{B}_0 \times \mathscr{B}$.

$G^*(x + i0)$ for x in A. Since $G^* \in \mathscr{P}_F(V, \mathscr{B}_0)$, G^* is F-differentiable on V and thus continuous and G-differentiable in V.

It follows from a version of the chain rule (Ref. 15, pp. 171–2) that h defined by

$$h(v) = F^*[G^*(v), v], \quad v \in V$$

is F-differentiable on V. [Notice that $h(v) = (F^*Q)(v)$, where Q takes $v \in V$ into the point $(G^*(v), v)$ in $\mathscr{B}_0 \times \mathscr{B}$.] Since V is connected, and $F^*[G^*(\theta + x + i0), x + i0] = F[G(x), x] = 0$ (θ is the zero element of $\mathscr{B}$) for x in some open ball in X centered at the origin, by Lemma 2, one has (11).

Assuming, on the other hand, that we have a V and a G^* as indicated in the theorem, G^* is F-differentiable (see Ref. 12, Theorem 3.17.1) and an obvious modification of the part of the proof of Theorem 2 that concerns H shows that $G \in \mathscr{P}(A, X_0)$. This proves the theorem.

Theorem 5: Assume that $G \in \mathscr{P}(A, X_0)$, *and that G has the generalized power-series representation*

$$G(x) = \sum_{m=0}^{\infty} G_m(x) \tag{12}$$

for $x \in A$. *Suppose that the right side of* (12) *converges for* $x \in B$, *where B is an open subset of X such that* $A \subset B$. *Then* $F[G(x), x] = 0$ $(x \in B)$, *where G is defined for all* $x \in B$ *by* (12).

Proof: Paralleling the beginning of the proof of Theorem 3, there is an open connected set $W \subset \mathscr{B}$ and a Fréchet-differentiable $h^*: W \longrightarrow \mathscr{B}_0$ such that $B \subset W$ and $G(x) = h^*(x + i0)(x \in B)$. Using Lemma 2 and the observation in the proof of Theorem 5 concerning the applicability of a version of the chain rule, we have $F^*[h^*(v), v] = 0$ for $v \in W$, which implies that $F[G(x), x] = 0$ $(x \in B)$.

Remarks: Theorem 4 bears directly on problems concerning the existence of generalized power-series expansions for the solutions of nonlinear differential equations, because, as is well known, these equations can frequently be put in the form $F[G(x), x] = 0$, $x \in A$, where x takes into account inputs and/or initial conditions, and $G(x)$ is the corresponding solution. For related earlier work, see Ref. 8 and the references cited therein; the work includes, in particular, a description of the specific type of expansions that arise.

A result similar to Theorem 2 in Section 2.3 can be obtained for equations of the form $H(y, x) = w$, where y is a solution that depends on both x and w.* Specifically, suppose that H is a map from $X_0 \times X$ into X_1 such that $H(y, x) = H^*(y + i0, x + i0)$ for all x and y for some

* See Ref. 8, p. 75, for an example of how such an equation arises.

$H^* \in \mathscr{P}_F(\mathscr{B}_0 \times \mathscr{B}, \mathscr{B}_1)$. Assume now that the norm $\|\cdot\|$ in $X_0 \times X$ is defined by $\|(x_0, x)\| = \max(\|x_0\|, \|x\|)$, and similarly for $(X_1 \times X)$, $(\mathscr{B}_0 \times \mathscr{B})$, and $(\mathscr{B}_1 \times \mathscr{B})$. Assume also that $\mathscr{B}_0$ and $\mathscr{B}_1$ are homeomorphic in the sense of (ii) of Theorem 2.

Let A_{wx} and S, respectively, be open subsets of $(X_1 \times X)$ and $(X_0 \times X)$, with $(0, 0) \in A_{wx}$, such that for each $(w, x) \in A_{wx}$ there is a unique $y \in X_0$ for which $(y, x) \in S$ and $H(y, x) = w$. Assume that the map from (w, x) to y is continuous. Define $f:(X_0 \times X) \rightarrow (X_1 \times X)$ by $f(y, x) = [H(y, x), x]$ for all y and x, and let A_{yx} denote the open set $S \cap f^{-1}(A_{wx})$. Notice that f restricted to A_{yx} is a homeomorphism of A_{yx} onto A_{wx}. Thus, using Theorem 2 and the observation in the footnote in Appendix B, we see that the map from (w, x) to y described above belongs to $\mathscr{P}(A_{wx}, X_0)$ if and only if $(\mathscr{B}_1 \times \mathscr{B})$ and $(\mathscr{B}_0 \times \mathscr{B})$, respectively, contain open subsets V and V_0 with $A_{wx} \subset V$, $A_{yx} \subset V_0$, V a star, and $[H^*(y^*, x^*), x^*] \in V$ for $(y^*, x^*) \in V_0$, such that for each $(w^*, x^*) \in V$, there is a unique $y^* \in \mathscr{B}_0$ that satisfies $(y^*, x^*) \in V_0$ and $H^*(y^*, x^*) = w^*$, with the map from (w^*, x^*) to y^* locally Lipschitz.

2.6 Construction of the series for g of Section 2.2

Here we return to the setting introduced at the beginning of Section 2.2. Theorem 6 (in this section) provides an algorithm for determining the expansion of g whenever it exists. The theorem is a version of a result in Ref. 1 concerning complex spaces. We shall first prove a proposition that establishes the existence of certain derivatives that play a central role in the theorem.

Proposition: For each $m = 1, 2, \cdots$ the mth order Fréchet derivative (Ref. 14, *pp.* 179–81) $d^m f[g(0)]$ [*of f at the point* $g(0) \in X_0$] *exists, and one has*

$$f[g(0) + x] = f[g(0)] + \sum_{m=1}^{\infty} (m!)^{-1} d^m f[g(0)]x^m, \quad x \in X_0. \quad (13)$$

Proof: Using the hypothesis that $f^* \in \mathscr{P}_F(\mathscr{B}_0, \mathscr{B})$, the Fréchet derivative $df^*(w)$ and hence the derivatives $d^m f^*(w)$ $(m = 2, 3, \cdots)$ exist for $w \in \mathscr{B}_0$, and we have

$$f^*[g^*(0) + w] = f^*[g^*(0)] + \sum_{m=1}^{\infty} (m!)^{-1} d^m f^*[g^*(0)]w^m, \quad w \in \mathscr{B}_0 \quad (14)$$

(see Ref. 12, Theorems 26.6.3 and 3.16.2; Ref. 15, Lemma 3.6.1; Ref. 7, Lemma 2), where $g^*(0) = g(0) + i0$. Since df^* exists on $\mathscr{B}_0$, and the norm in $\mathscr{B}$ has the property that $\|x_1 + ix_2\| < \delta$ implies that $\|x_1\| < \delta$ and $\|x_2\| < \delta$ [see (7)], it is easy to see that df exists on X_0 and that one has $df(a)x = df^*(a + i0)(x + i0)$ for a and x in X_0. A simple inductive argument shows that $d^m f$ exists on X_0, with

$$d^m f(a)x_1 \cdots x_m = d^m f^*(a + i0)(x_1 + i0) \cdots (x_m + i0) \tag{15}$$

for $a \in X_0$, $(x_1, \cdots, x_m) \in X_0^m$, and each m. This proves the proposition, since it is clear that $f^*[g^*(0)] = f[g(0)]$. The relation (13) directs attention to an interpretation of the $d^m f[g(0)]$; it is not used otherwise.

Theorem 6: Let $g \in \mathscr{P}(A, X)$. Then $df[g(0)]$ is a homeomorphism of X_0 onto X, and

$$g(x) = g(0) + \sum_{m=1}^{\infty} g_m(x), \quad x \in A,$$

where the g_m are the homogeneous polynomials defined by

$$g_1(x) = df[g(0)]^{-1}x$$

and

$$g_m(x) = -df[g(0)]^{-1} \sum_{\ell=2}^{m} (\ell!)^{-1} \cdot \sum_{\substack{k_1+\cdots+k_\ell=m \\ k_j>0}} d^\ell f[g(0)]g_{k_1}(x) \cdots g_{k_\ell}(x), \quad m \geqslant 2.$$

2.6.1 Proof of Theorem 6

The inverse of $df[g(0)]$ exists because f is a homeomorphism of A_0 onto A with f and g, respectively, Fréchet differentiable on A_0 and A (Ref. 15, p. 175, Problem 6). Let

$$g(x) = g(0) + \sum_{m=1}^{\infty} g_m(x), \quad x \in A,$$

in which each g_m is a homogeneous polynomial on X of degree m.

With g^* and V associates of g and A, respectively, in accordance with Theorem 1, we have $g^* \in \mathscr{P}_F(V, \mathscr{B}_0)$ and $v = f^*[g^*(v)]$ for $v \in V$. Let $g_1^*, g_2^*, \cdots$ be continuous homogeneous polynomials such that

$$g^*(v) = g^*(0) + \sum_{m=1}^{\infty} g_m^*(v) \tag{16}$$

for $v \in V$. By the part of the proof of Theorem 2 concerning H, $g_m(x) = g_m^*(x + i0)$ for each m and each $x \in X$. Using (14) and $f^*[g^*(0)] = 0$,

$$\begin{aligned} v &= f^*[g^*(v)] \\ &= \sum_{\ell=1}^{\infty} (\ell!)^{-1} d^\ell f^*[g^*(0)] \left(\sum_{k_1=1}^{\infty} g_{k_1}^*(v) \cdots \sum_{k_\ell=1}^{\infty} g_{k_\ell}^*(v) \right), \quad v \in V. \end{aligned}$$

Since $g^* \in \mathscr{P}_F(V, \mathscr{B}_0)$, there is a $\sigma > 0$ such that the right side of (16) is *absolutely* convergent for $\| v \| < \sigma$ (Ref. 12, Theorem 26.6.6). Thus, using the boundedness of the $d^{\ell} f^{*}[g^{*}(0)]$ and an easily proved generalization of Theorem 5.5.3 of Ref. 14,

$$v = \sum_{\ell=1}^{\infty} (\ell!)^{-1} \sum_{k_1,\cdots,k_\ell=1}^{\infty} d^{\ell} f^{*}[g^{*}(0)](g^{*}_{k_1}(v) \cdots g^{*}_{k_\ell}(v)) \tag{17}$$

for $\| v \| < \sigma$, in which the sum over $(k_1, \cdots, k_\ell)$ is absolutely convergent.

At this point we use the proposition that there are positive constants M, β, K, and α such that

$$\| d^{\ell} f^{*}[g^{*}(0)] w_1 \cdots w_\ell \| \leqslant \ell^{\ell} M \beta^{\ell} \| w_1 \| \cdots \| w_\ell \|$$

and

$$\| g^{*}_{k}(v) \| \leqslant K(\alpha \| v \|)^{k}$$

for $\ell \geqslant 1$, $k \geqslant 1$, $v \in \mathscr{B}$, and $w_1, \cdots, w_\ell$ in $\mathscr{B}_0$ (see Ref. 7, Appendix A). It is clear that

$$\| d^{\ell} f^{*}[g^{*}(0)][g^{*}_{k_1}(v) \cdots g^{*}_{k_\ell}(v)] \| \leqslant \ell^{\ell} M(\beta K)^{\ell} (\alpha \| v \|)^{(k_1 + \cdots k_\ell)}.$$

Consider the sum

$$\sum_{\ell=1}^{\infty} \sum_{k_1,\cdots,k_\ell=1}^{\infty} (\ell!)^{-1} \ell^{\ell} M(\beta K)^{\ell} (\alpha \| v \|)^{(k_1 + \cdots + k_\ell)}. \tag{18}$$

Notice that

$$\sum_{k_1,\cdots,k_\ell=1}^{\infty} (\alpha \| v \|)^{(k_1 + \cdots + k_\ell)} = [\alpha \| v \| (1 - \alpha \| v \|)^{-1}]^{\ell} \tag{19}$$

for $\alpha \| v \| < 1$. Using (19) and Stirling's formula for $n!$, which gives $n! > (2\pi)^{1/2} n^{1/2} n^{n} e^{-n}$, it easily follows that the sum (18) converges for $\| v \|$ sufficiently small. Thus (see Ref. 14, Theorem 5.3.4) for such v the sum in (17) over $(\ell, k_1, \cdots, k_\ell)$, which equals

$$\sum_{\ell=1}^{\infty} \sum_{m=1}^{\infty} \sum_{\substack{k_1 + \cdots + k_\ell = m \\ k_j > 0}} (\ell!)^{-1} d^{\ell} f^{*}[g^{*}(0)](g^{*}_{k_1}(v) \cdots g^{*}_{k_\ell}(v)), \tag{20}$$

can be written as (see Ref. 14, Theorem 5.3.6)

$$\sum_{m=1}^{\infty} \sum_{\ell=1}^{m} \sum_{\substack{k_1 + \cdots + k_\ell = m \\ k_j > 0}} (\ell!)^{-1} d^{\ell} f^{*}[g^{*}(0)](g^{*}_{k_1}(v) \cdots g^{*}_{k_\ell}(v)). \tag{21}$$

By the uniqueness result for generalized power series mentioned in Section 2.1, and the fact that (21) equals v for v of sufficiently small norm,

$$df^{*}[g^{*}(0)] g^{*}_{1}(v) = v$$

and

$$df^*[g^*(0)]g_m^*(v)$$
$$= -\sum_{\ell=2}^{m} \sum_{\substack{k_1+\cdots+k_\ell=m \\ k_j>0}} (\ell!)^{-1} d^\ell f^*[g^*(0)][g_{k_1}^*(v) \cdots g_{k_\ell}^*(v)] \quad (m \geq 2)$$

for $\nu \in \mathscr{B}$. This, with $v = x + i0$ and (15), completes the proof.

2.6.2 Comments

For the case in which the expansion (14) for f^* has only a finite number of terms, the proof of Theorem 6 simplifies considerably, because then the equivalence of (20) and (21) is a consequence of just the absolute convergence of the sum over $(k_1, \cdots, k_\ell)$ in (17). A related result for this case is given in Ref. 17, p. 29.

For a different approach to the problem of determining the expansion of g, see Ref. 18.

The proof of Theorem 6 provides an alternative proof of Theorem 2 of Ref. 1, which is an analogous result concerning only complex spaces. It also yields a proof of the following "substitution theorem" (an earlier version proved in a different way appears in Ref. 19).*

Theorem 7: Take W_1, W_2, and W_3 to be three complex Banach spaces, and let S_1 and S_2 be nonempty open subsets of W_1 and W_2, respectively, with S_1 a star. Let $G \in \mathscr{P}_F(S_1, W_2)$ and let F be a Fréchet-differentiable map of S_2 into W_3. Assume that $G(S_1) \subset S_2$. Then $(FG)(\cdot) \in \mathscr{P}_F(S_1, W_3)$, the Fréchet derivatives $d^mG(0)$ and $d^mF[G(0)]$ exist for $m \geq 1$, and we have

$$(FG)(v) = F[G(0)] + \sum_{m=1}^{\infty} H_m(v), \quad v \in S_1, \tag{22}$$

where the H_m are the homogeneous polynomials given by

$$H_m(v) = \sum_{\ell=1}^{m} (\ell!)^{-1} \sum_{\substack{k_1+\cdots+k_\ell=m \\ k_j>0}} d^\ell F[G(0)](k_1!)^{-1}d^{k_1}G(0)v^{k_1} \cdots (k_\ell!)^{-1}d^{k_\ell}G(0)v^{k_\ell}.$$

Theorem 7 and Lemma 1 can be used to obtain results along the lines of Theorem 7 for cases in which the spaces of interest are real. This is discussed briefly in Appendix A.

REFERENCES

1. I. W. Sandberg, "Expansions for Nonlinear Systems," B.S.T.J., *61*, No. 2 (February 1982), pp. 159–99.

* Concerning the proof of Theorem 7, $(FG)(\cdot) \in \mathscr{P}_F(S_1, W_3)$ because G and hence $(FG)(\cdot)$ are Fréchet differentiable on the star S_1. Also, since F is Fréchet differentiable in a neighborhood of $G(0)$, F has a locally convergent expansion about that point.

2. I. W. Sandberg, "On Volterra Expansions for Time-Varying Nonlinear Systems," IEEE Trans. Circuits Syst., *CAS-30* (February 1983), pp. 61–7.
3. I. W. Sandberg, "The Mathematical Foundations of Associated Expansions for Mildly Nonlinear Systems," IEEE Trans. Circuits Syst., *CAS-30* (July 1983), pp. 441–55.
4. M. Fliess, M. Lamnabhi, and F. Lamnabhi-Lagarrique, "An Algebraic Approach to Nonlinear Functional Expansions," IEEE Trans. Circuits Syst., *CAS-30* (August 1983), pp. 554–70.
5. R. J. P. de Figueiredo, "A Generalized Fock Space Framework for Nonlinear System and Signal Analysis," IEEE Trans. Circuits Syst., *CAS-30* (September 1983), pp. 637–47.
6. S. Boyd, "Volterra Series: Engineering Fundamentals," Dissertation, University of California, Berkeley, March 1985.
7. I. W. Sandberg, "Nonlocal Input-Output Expansions," AT&T Tech. J., *64,* No. 1, Part 1 (January 1985), pp. 77–90.
8. I. W. Sandberg, "Volterra-like Expansions for Solutions of Nonlinear Integral Equations and Nonlinear Differential Equations," IEEE Trans. Circuits Syst., *30* (February 1983), pp. 68–77.
9. A. E. Taylor, "Additions to the Theory of Polynomials in Normed Linear Spaces," Tohôku Math. J., *44* (1938), pp. 302–18.
10. A. E. Taylor, "Analysis in Complex Banach Spaces," Bull. Amer. Math. Soc., *49* (1943), pp. 652–69.
11. L. M. Graves, "Riemann Integration and Taylor's Theorem in General Analysis," Trans. Amer. Math. Soc., *29* (January 1927), pp. 163–77.
12. E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups,* Providence: Amer. Math. Soc. Coll. Publ. XXXI, 1957.
13. A. Alexiewicz and W. Orlicz, "Analytic Operations in Real Banach Spaces," Studia Math., *14* (1954), pp. 57–78.
14. J. Dieudonné, *Foundations of Modern Analysis,* New York: Academic Press, 1969.
15. T. M. Flett, *Differential Analysis,* London: Cambridge University Press, 1980.
16. J. M. Holtzman, *Nonlinear System Theory,* Englewood Cliffs: Prentice-Hall, 1970.
17. A. Halme, "Polynomial Operators for Nonlinear Systems Analysis," Acta Polytech. Scand. Ma *24* (1972) (Dissertation), pp. 1–63.
18. I. W. Sandberg, "Iteration and Functional Expansions," Circuits Syst. Signal Process, *3,* No. 4 (1984), pp. 409–17.
19. I. W. Sandberg, "Series Expansions for Nonlinear Systems," Circuits Syst. Signal Process, *2,* No. 1 (1983), pp. 77–87.
20. A. D. Michal and M. Wyman, "Characterization of Complex Couple Spaces," Ann. Math., *42* (1941), pp. 247–50.
21. F. Riesz and B. Sz-Nagy, *Functional Analysis,* New York: Frederick Unger, 1955.

APPENDIX A

Substitution Results for Real Spaces

This appendix presents two useful corollaries of Theorem 7. Proofs are omitted because the corollaries can be proved using direct modifications of material already discussed.

Corollary 1: Assume that W_1, W_2, and W_3 are real Banach spaces, that S_1 and S_2 are open subsets of W_1 and W_2, respectively, and that $0 \in S_1$. Let G be a map of S_1 into W_2 such that the Fréchet derivative $d^mG(0)$ exists for $m = 1, 2, \cdots$, and G has the representation

$$Gx = G(0) + \sum_{m=1}^{\infty} (m!)^{-1} d^m G(0) x^m, \quad x \in S_1.$$

Suppose that $G(S_1) \subset S_2$. Let F map S_2 into W_3 such that $d^mF[G(0)]$ exists for each $m > 0$, and

$$F[y + G(0)] = F[G(0)] + \sum_{m=1}^{M} (m!)^{-1} d^m F[G(0)] y^m$$

for $y + G(0) \in S_2$, *where* M *is a positive integer (and thus* F *is assumed to be a polynomial). Then* $(FG)(\cdot) \in \mathscr{P}(S_1, W_3)$ *and* (22) *holds.*

Corollary 2: Suppose that W_1, W_2, *and* W_3 *are three real Banach spaces. Let* $G \in \mathscr{P}_F(S_1, W_2)$ *for some open subset* S_1 *of* W_1 *containing the point* 0, *and let* $F \in \mathscr{P}_F(S_2, W_3)$, *where* S_2 *is an open subset of* W_2 *containing* $G(0)$. *Then the Fréchet derivatives* $d^m G(0)$ *and* $d^m F[G(0)]$ *exist for* $m \geqslant 1$, *and there is an open subset* T_1 *of* S_1, *containing* 0, *such that* $G(T_1) \subset S_2$, $(FG)(\cdot) \in \mathscr{P}_F(T_1, W_3)$, *and one has* (22), *with* S_1 *replaced with* T_1.

APPENDIX B

A Comparison of Norms on Complex Spaces

Consider the Banach space $\mathscr{B}$ described in Section 2.1, and let $\mathscr{C}$ denote a Banach space consisting of the same set of points with a possibly different norm $\| \cdot \|_{\mathscr{C}}$.

Proposition: Let the norm $\| \cdot \|_{\mathscr{C}}$ *have the property that* $\| x_1 \| \leqslant \| x_1 + ix_2 \|_{\mathscr{C}}$ *for* $(x_1 + ix_2) \in \mathscr{C}$, *in which* $\| x \|$ *is the* X *norm of* x. *Then* $\| x_1 + ix_2 \| \leqslant \| x_1 + ix_2 \|_{\mathscr{C}}$ *for* $(x_1 + ix_2) \in \mathscr{B}$.

Proof: Assume, for the purpose of obtaining a contradiction, that $\| x_1 + ix_2 \| > \| x_1 + ix_2 \|_{\mathscr{C}}$ for some $(x_1 + ix_2)$. Then there is a ξ with $\| \xi \| = 1$ such that

$$\xi(x_1)^2 + \xi(x_2)^2 > \| x_1 + ix_2 \|_{\mathscr{C}}^2.$$

For this ξ, choose real α and β so that not both are zero and

$$\alpha\xi(x_2) + \beta\xi(x_1) = 0.$$

Using $(\alpha^2 + \beta^2)[\xi(x_1)^2 + \xi(x_2)^2] > (\alpha^2 + \beta^2)\| x_1 + ix_2 \|_{\mathscr{C}}^2$ and the observation that $[\xi(ax - by)]^2 + [\xi(bx + ay)]^2 = (a^2 + b^2)[\xi(x)^2 + \xi(y)^2]$ for real a and b, and x and y in X, one has

$$| \xi(\alpha x_1 - \beta x_2) | > \| (\alpha x_1 - \beta x_2) + i(\beta x_1 + \alpha x_2) \|_{\mathscr{C}}.$$

Since the left side is at most $\| \alpha x_1 - \beta x_2 \|$, we have a contradiction.

Comments: For X the space of bounded n-vector-valued functions described in Section 2.1, and $\mathscr{C}$ the corresponding complex Banach space with $\| v \|_{\mathscr{C}} = \max_j \sup_t | v_j(t) |$, the equality $\| \cdot \| = \| \cdot \|_{\mathscr{C}}$ holds, where $\| \cdot \|$ is the norm in $\mathscr{B}$. Indeed, $\| \cdot \| \leqslant \| \cdot \|_{\mathscr{C}}$ by the proposition, while with arbitrary $t \geqslant 0$ and $j \in \{1, \cdots, n\}$,

$$\xi(x_1)^2 + \xi(x_2)^2 = [x_{1j}(t)]^2 + [x_{2j}(t)]^2$$

for ξ the linear functional of unit norm on X defined by $\xi(x) = x_j(t)$, showing that $\|\cdot\| \geqslant \|\cdot\|_{\mathscr{C}}$.*

However, we have $\|\cdot\| \leqslant \|\cdot\|_{\mathscr{C}}$ but *not* $\|\cdot\| = \|\cdot\|_{\mathscr{C}}$ whenever X is a Hilbert space, $\|x_1 + ix_2\|_{\mathscr{C}} = (\|x_1\|^2 + \|x_2\|^2)^{1/2}$, and X is typical in the sense that it has a pair of nonzero elements x and y that are orthogonal. This follows from the inner-product representation of linear functionals in a Hilbert space, and the fact that for X, x, and y, as indicated above, one can show that

$$\sup_{\|v\|=1} \frac{(v, x)^2 + (v, y)^2}{\|x\|^2 + \|y\|^2} < 1,$$

where $(\cdot, \cdot)$ is the inner product in X.

Finally, we mention that the norm in $\mathscr{B}$ *cannot* be replaced with

$$\|(x_1, x_2)\| = (\|x_1\|^2 + \|x_2\|^2)^{1/2}, \tag{23}$$

because (23) does not define a norm in $\mathscr{B}$ unless X is a Hilbert space (see Ref. 20). It is not difficult to see that (23) does not suffice: If it did, we would have $\|ax - by\|^2 + \|bx + ay\|^2 = (a^2 + b^2)(\|x\|^2 + \|y\|^2)$ for any real numbers a and b, and arbitrary elements x and y of X. This would give $\|x - y\|^2 + \|x + y\|^2 = 2(\|x\|^2 + \|y\|^2)$, which is not valid unless X is a Hilbert space (see Ref. 21, p. 211).

AUTHOR

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; AT&T Bell Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of radar systems for military defense, synthesis and analysis of active and time-varying networks, with several fundamental studies of properties of nonlinear systems, and with some problems in communication theory and numerical analysis. His more recent interests have included compartmental models, the theory of digital filtering, global implicit-function theorems, and functional expansions for nonlinear systems. IEEE Centennial Medalist, Former Vice Chairman IEEE Group on Circuit Theory, and Former Guest Editor IEEE Transactions on Circuit Theory Special Issue on Active and Digital Networks. Fellow and member, IEEE; member, American Association for the Advancement of Science, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, National Academy of Engineering.

* This type of argument also shows that the norm $\|(a_0, a) + i(b_0, b)\|$ of an arbitrary element $(a_0, a) + i(b_0, b)$ of the complex extension of the Banach space $X_0 \times X$ with norm $\max(\|\cdot\|, \|\cdot\|)$ is $\max(\|a_0 + ib_0\|, \|a + ib\|)$, where $\|a_0 + ib_0\|$ and $\|a + ib\|$ are the $\mathscr{B}_0$ and $\mathscr{B}$ norms of $(a_0 + ib_0)$ and $(a + ib)$, respectively.

AT&T Technical Journal
Vol. 64, No. 7, September 1985
Printed in U.S.A.

Regular Mesh Topologies in Local and Metropolitan Area Networks

By N. F. MAXEMCHUK*

(Manuscript received January 14, 1985)

The throughput per user in loop and bus configured local area networks decreases linearly with the number of users. These networks cannot be extended to a metropolitan area with many users. A class of mesh networks is described that increases the throughput of conventional local area networks by decreasing the fraction of the network capacity needed to transmit information between a source and a destination. These networks have multiple paths that increase the reliability of the networks, and have point-to-point links that can cover a metropolitan area. In general, mesh networks require complex store-and-forward nodes that also route messages, control the flow of data entering the network, resequence packets at the destination, and recover packets with errors. However, there are characteristics of the local or metropolitan area that allow these functions to be simplified. As a result of these simplifications, loop access protocols are extended to mesh networks and the need to store and forward data is eliminated. A file transfer protocol that does not require packet resequencing is described. Three mesh networks are studied, and the desirable characteristics of networks are determined. One network, the Manhattan street network, has many of the desirable characteristics.

I. INTRODUCTION

Loop topologies[1] and random access strategies[2] were first applied to local data networks in the late 1960's. In that era,

- Low-bit-rate terminals were connected to large central computers,
- Computers and terminals were shared by a few computer experts,

* AT&T Bell Laboratories.

- Large-scale integrated circuits did not exist, and
- High-bit-rate transmission facilities were not readily available across public right of ways.

As a result, these networks trade reliability, total throughput, and the distance the network can span[3] for simple access and transmission strategies. Today, for comparison,

- Simple terminals are evolving into personal computers with bit mapped, rather than character, displays,
- Computer usage is becoming universal,
- Very-Large-Scale Integration (VLSI) is becoming commonplace, and
- The increased deployment of optical fibers and CATV systems makes it possible to obtain high-bit-rate communications over wider areas.

Personal computers use larger bandwidths than simple terminals to communicate with centralized support facilities and distribute processing. The increasing use of these devices and the increased distances that high-bit-rate networks can span increase the throughput required of the interconnecting network. In loop and bus systems the total throughput is constant. The average capacity available to each user decreases linearly as the number of users increases. Therefore, to support more users with greater individual requirements, alternative topologies must be considered. The complexity of the devices being connected to networks and advances in VLSI make more complex network interfaces feasible. This increases the class of networks and access strategies that can be considered.

A large number of users, dispersed over a large area, can be accommodated by interconnecting conventional local area networks with gateways. Schlatter and Massey have analyzed this type of network.[4] Their system consists of loops interconnected by switching elements, as proposed by Pierce.[5] Messages use a smaller fraction of the total network capacity than they would if the system were a single loop. Therefore, the maximum throughput of the system increases. Users who communicate the most often are placed on the same loop, which minimizes the interference between subgroups of users. The main disadvantage with this approach is that the gateways are different from the access units and are complex store-and-forward elements.

Yemini[6] and Saadawi and Schwartz[7] are investigating a tree topology. In this network, users are at the leaves of the tree and the nodes of the tree are switching points. Depending on the location of the destination, the switches direct messages toward the root of the tree or toward the leaves. In Yemini's system, the switches establish separate broadcast networks, and in Saadawi's, the switches store packets until the desired path is available. In these systems, messages only use a portion of the network capacity. Locating users who communicate frequently—who are near one another in the tree hierarchy—minimizes the interference between subgroups of users. The advantage of this approach over gateways is that there is only one type of element in the network. The disadvantage is that the network either stores and forwards packets or retains the distance constraints of broadcast networks.

To a certain extent these alternatives remove the throughput constraints of loop and bus systems. However, they still have single points of failure. To make networks more reliable, there must be multiple paths between each source and destination. By adding paths appropriately, the average and maximum distance between nodes decreases, messages use a smaller fraction of the network bandwidth, and the throughput increases. Multiple paths also make it possible to avoid heavily used segments of the network to equalize the load. Mesh networks, like loop networks, have point-to-point communication channels between nodes. This results in less expensive line drivers and receivers than multidrop broadcast systems and is compatible with current optical fiber transmission capabilities.

In general, mesh-configured networks and some of the local network alternatives require complex store-and-forward nodes. A queue of messages is maintained because packets arriving on several of the incoming links may be destined for the same outgoing link. In addition, store-and-forward networks must do routing, flow control, packet resequencing, and error control. Long-distance networks, such as the ARPA network,[8] perform these functions, but their interfaces are more complex than personal computers. Therefore, these networks are not a reasonable interconnection alternative for personal computers. There are, however, characteristics of the local or metropolitan area environment that make simpler mesh networks possible.

Local or metropolitan area networks differ from general long-distance networks in that the

- Physical location of the nodes does not dictate the topology of the network to as great an extent,
- Error rates are much lower, and
- Communications lines are less expensive.

In the local environment, it is not always necessary to connect the closest nodes together. Occasionally, connecting nodes that are further apart can make the topology of the network regular, and simplify tasks such as routing. The lower error rates make it more likely that a packet will traverse the entire network without error. Therefore, error control protocols can operate on an end-to-end, rather than on a link-by-link basis, and networks can have unidirectional links. When messages are not stored at the intermediate nodes for possible retransmission, the class of possible access and transmission strategies increases. In general, there is also a trade-off between the system complexity and communications efficiency. The cost of communications lines, and the additional access strategies and network options that are possible, result in a different network solution in local networks than in long-distance networks. The result is that local networks are significantly less complex.

There is a description in Section III of several regular networks with simple routing strategies. End-to-end error control protocols make it possible to extend the slotted system and register-insertion techniques developed for loop systems to mesh networks. This is shown in Section IV. With these techniques, flow control on mesh networks can be done by throttling the sources, as on loop networks. A trade-off exists between the buffering in these systems and the efficiency with which the communication lines are used. One attempt to take advantage of this trade-off is the Floodnet system.[9] There is a description in Section IV of how this trade-off is applied to mesh networks with slotted system and register-insertion interfaces. The result is that for certain topologies mesh networks without buffering are reasonable. Finally, there is a description in Section V of several file-transfer protocols that do not resequence packets.

II. EXAMPLE

Before discussing the implementation of mesh networks, we show that these networks provide a potential to increase the throughput of conventional local area networks. In this example, two-connected networks, with as few as 64 nodes, increase the throughput of bus configured networks by a factor of 20 to 30. This comparison assumes that the same rate communication lines are used in both the mesh and random access networks. A factor of two increase in throughput is obtained because there is twice as much capacity emanating from each node. However, the major portion of the increase occurs because messages in the mesh network use only a fraction of the total network capacity. Greater increases are obtained in larger networks.

Two traffic distributions are considered, a uniform distribution and a skewed distribution. In the uniform case, each node sends an equal amount of data to each of the other nodes. The skewed distribution corresponds to what might occur in a network of personal computers and file servers. The network is divided into communities of interest, each consisting of a file server and seven personal computers. A personal computer directs 80 percent of its traffic to its own file server and 20 percent to the other file servers. The computer receives an equal amount of traffic from the file servers.

For each traffic distribution, the throughput for six network topologies is investigated. The first two networks are the conventional broadcast bus and loop configured networks. The throughput of the bus network is calculated assuming that the link utilization can approach one, and is an upper bound on the achievable throughput. In the loop network, the packets only use the links between the source and destination. In this network, and in the remaining networks, the throughput is determined by increasing the traffic levels from the sources until the utilization on any link equals one. The remaining four networks are two-connected networks with two links arriving at, and two links emanating from, each node. The first of these networks is a conventional bidirectional loop. For the skewed distribution, the file server is in the middle of the seven personal computers it is servicing. The next two networks are regular arrays called the modified shuffle exchange and the Manhattan street network. These networks are described in Section III. The Manhattan street networks with 16, 32, 48, and 64 nodes are 4 × 4, 8 × 4, 8 × 6, and 8 × 8 arrays, respectively. For the skewed distribution, the seven personal computers and the file server in a community of interest are arranged in a 4 × 2 array on the network. This is shown for the 16-node network

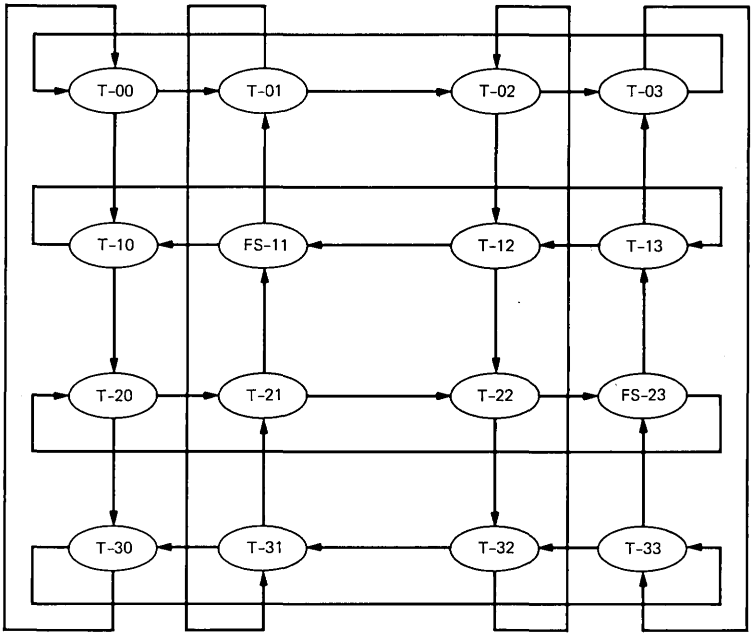


Fig. 1—A 16-node Manhattan street network with two communities of seven personal computers (T) and a file server (FS).

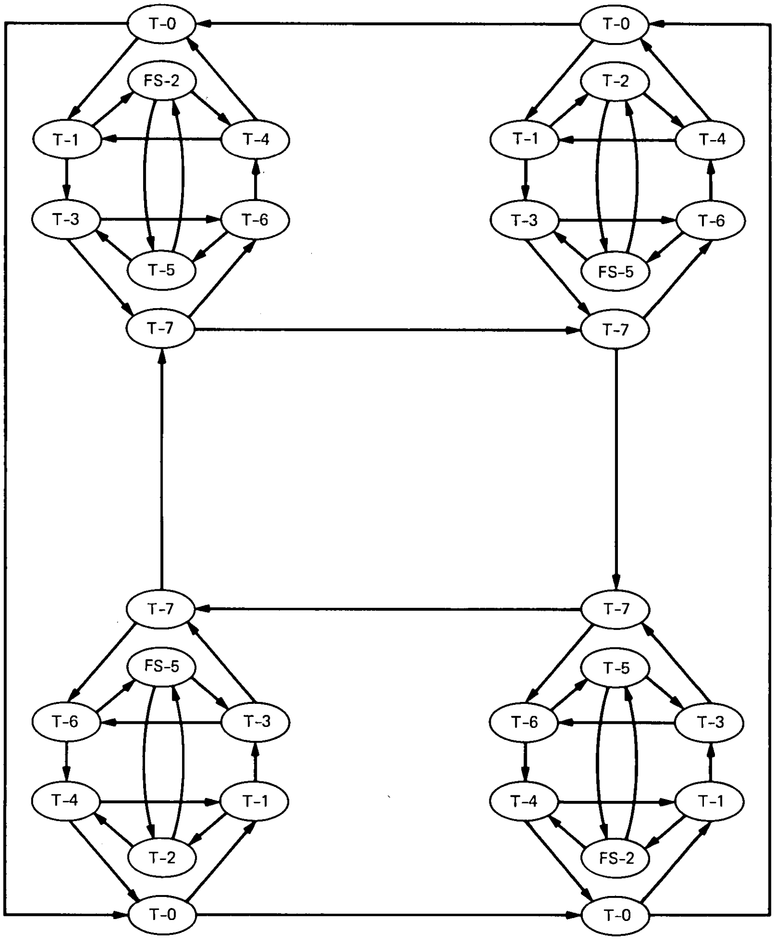


Fig. 2—A 32-node hierarchical network with four communities of seven personal computers (T) and a file server (FS) interconnected by shuffle-exchange networks. The four shuffle-exchange networks are connected by a bidirectional loop.

in Fig. 1. The final network is a hierarchical shuffle exchange, consisting of shuffle-exchange networks with the eight devices in a community of interest, interconnected by a bidirectional loop. Figure 2 shows a 32-node, hierarchical network consisting of four 8-node shuffle-exchange networks.

Traffic on the two-connected networks is placed on the shortest path between the source and destination. If there are several paths of equal length between a source and a destination, the path with the smallest flow is selected. Traffic with the shortest distance between a source and destination is assigned to the network first. Once a source destination requirement is assigned a path, the path is not changed if

Table I—The average megabits per user in networks with 10-Mb/s channels and the improvement over a broadcast network

	16 Nodes		32 Nodes		48 Nodes		64 Nodes	
Network	Mb/Usr	Imprv.	Mb/Usr	Imprv.	Mb/Usr	Imprv.	Mb/Usr	Imprv.
				Uniform Requirements				
BDCST	0.63	—	0.31	—	0.21	—	0.16	—
Loop	1.25	2.00	0.63	2.00	0.42	2.00	0.31	2.00
BDL	4.69	7.50	2.42	7.75	1.63	7.83	1.23	7.87
S-X	5.77	9.23	4.03	12.88			3.09	19.76
MSN	6.52	10.43	4.56	14.59	4.31	20.70	3.94	25.20
HS-X	4.69	7.50	1.96	6.28	1.41	6.75	1.13	7.20
				Skewed Requirements				
BDCST	0.36	—	0.18	—	0.12	—	0.09	—
Loop	0.71	2.00	0.36	2.00	0.24	2.00	0.18	2.00
BDL	2.63	7.37	2.08	11.67	1.80	15.11	1.59	17.82
S-X	1.61	4.52	1.01	5.68			0.81	9.10
MSN	2.63	7.37	2.17	12.17	2.12	17.80	2.12	23.76
HS-X	2.78	7.78	2.78	15.56	2.66	22.34	2.63	29.47

BDCST = Broadcast, BDL = Bidirectional, MSN = Manhattan street exchange, HS-X = Hierarchical street exchange, S-X = street exchange.

a link on the path becomes saturated, and the requirements are not split if two equally good paths exist. This procedure does not lead to the optimum throughput, but gives a reasonably good idea of what can be achieved.

The results of this investigation are presented in Table I. For each network, the average bit rate a user obtains in a network with 10-megabit-per-second transmission links, and the improvement this represents over a broadcast network, is presented. For the conventional broadcast and loop network, the fraction of the capacity a user obtains decreases linearly with the number of users, as expected. The loop system provides about twice as much throughput per user as the broadcast network because, on the average, a packet transmitted on this network uses only half of the network capacity. The two-connected networks obtain a factor of two increase in throughput because there is twice as much capacity emanating from each node, and an additional increase because the networks use a smaller fraction of the network capacity to transfer a packet between the sources and destinations.

The bidirectional loop, the Manhattan street network, and the hierarchical shuffle exchange respond well to the skewed requirements. These networks are capable of allowing complete connectivity while preventing users in different communities of interest from interfering with one another. This characteristic is extremely important in designing large networks. The shuffle-exchange and Manhattan street networks also respond well to a large group of users with uniform transmission requirements. This occurs because the average distance between users does not increase as rapidly in these networks as in the other networks. The average distance between users in the Manhattan street network is greater than that in the shuffle-exchange network; however, the throughput of the Manhattan street network is greater. The Manhattan street network can support a larger throughput because there are more equal-length shortest paths, and bottlenecks can be avoided.

III. TOPOLOGY

In this section, three two-connected networks are described, the bidirectional loop, the modified shuffle exchange, and the Manhattan street network. These networks have two independent paths between any node, and they can survive a single loop or node failure. While these networks are not optimal, they show what measures can be used to compare topologies, and what network characteristics are desirable. Lower bounds on two measures, the average and maximum shortest path between nodes, are derived and compared with these three topologies.

3.1 Bidirectional loops

In a bidirectional loop with N nodes, labeled 0 to $N - 1$, node i is connected to nodes $(i - 1) \bmod N$, and $(i + 1) \bmod N$. This is the only two-connected network with bidirectional paths between all of the nodes. If the transmission protocols require a response each time a packet of data is transferred between two intermediate nodes on a path, this is the only possible two-connected network. This network was initially considered as a mechanism to make loop networks more reliable.

This network has many of the topological advantages of loop systems. It

- Is defined for any number of nodes,
- Makes geographical sense,
- Has a simple rule for expanding the network by one node at a time, and
- Has a simple routing rule.

When a node is added to the network, the two existing nodes closest to this node are disconnected from one another and connected to the new node. Even if the network covers a large geographical area, there are not many long wires. Shortest-path routing in this network is straightforward. The nodes in the network are sequentially numbered from 0 to $N - 1$. The distance from a source node s to a destination node d is $(d - s) \bmod N$ on the incremental path and $(s - d) \bmod N$ on the decremental path. At the source, the shorter of these two paths is selected. Once a packet in this system starts on a path, it remains on that path. Therefore, a complete routing decision is made at the source, and the system is implemented as two separate loop systems.

A disadvantage of this network is that the node addresses and the value of N changes whenever a node is added to the system. Either an addressing and routing scheme that does not use this information must be found, or this information must be distributed each time the network is changed. Another disadvantage of this network is that the throughput is not as great as that in the shuffle-exchange and Manhattan street networks.

3.2 Modified shuffle exchange

The modified shuffle-exchange network is based on the shuffle-exchange multistage switch. The network is defined for N nodes, where N is constrained to be a power of two. Node i is connected to nodes $2^*i \bmod N$ and $(2^*i + 1) \bmod N$. This results in self-loops at nodes 0 and $N - 1$, which are not used to transmit packets. They also make the network less reliable in that a single link failure can disconnect a node from the network. In the modified network, the self-loops are removed and nodes 0 and $N - 1$ are connected to one another, as shown in Fig. 3. When the shuffle-exchange network is part of a hierarchical structure, as in the previous section, the self-loops are replaced by connections to the higher-level network.

Routing in this network is straightforward. Initially, ignore the two paths that were added to the modified network. Represent the address of node i by $M = \log_2 N$ bits, and label the paths to nodes $2^*i \bmod N$ and node $(2^*i + 1) \bmod N$ as 0 and 1, respectively. When a packet is transmitted from node i, the address of the new node has the low-order $M - 1$ bits of node i's address in the high-order $M - 1$ bits. The low-order bit of the new address is 0 or 1, depending on the path selected. To find the shortest path between a source and destination, match as many of the high-order bits of the destination address with the low-order bits of the source address as possible. To get to the destination, shift the low-order bits of the destination address that are not included in this match into the address and determine the path that must be selected.

For instance, assume that the source address is 11011 and the destination address is 11001. The first two bits of the destination address match the last two bits of the source address. The bits 001 must be shifted into the address to get to the destination, and the distance to the destination is 3. To get to the destination, first path 0 is taken to node 10110, then path 0 is taken to node 01100, and finally path 1 is taken to node 11001.

If none of the high-order bits of the destination match the low-order bits of the source, then the distance to the destination is $\log_2 N$ steps.

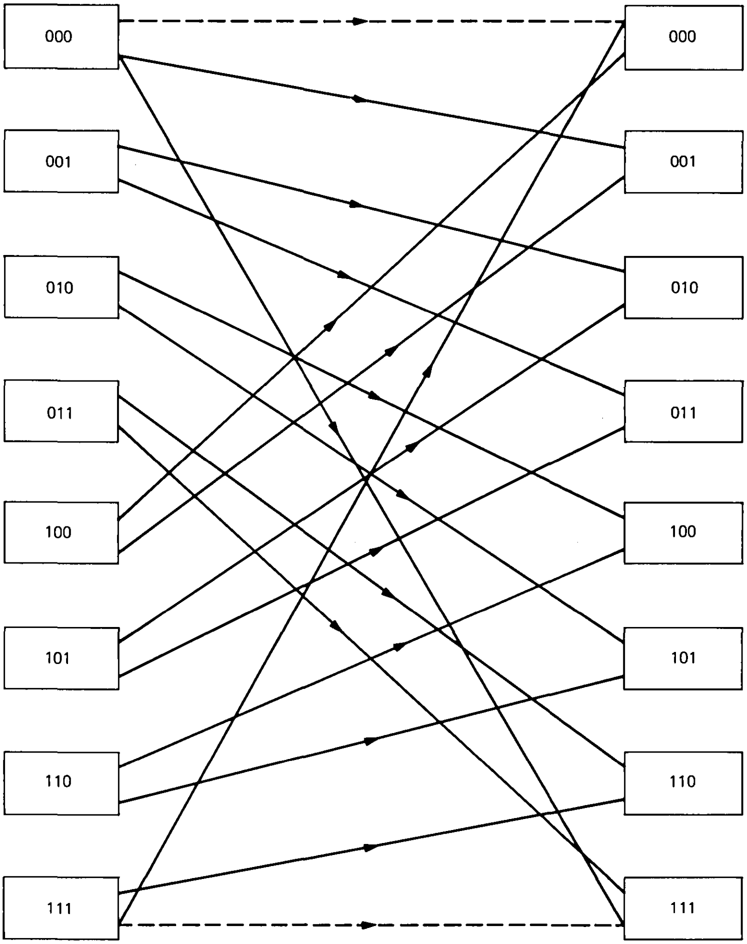


Fig. 3—An 8-node modified shuffle-exchange network.

This is the maximum distance between any source and any destination. In Section 3.4, this will be shown to be the minimum maximum distance between nodes for a two-connected network of this size.

The two paths that are added to the modified network make it possible to go from the all-one address to the all-zero address, and vice versa, in a single step. This shortens the average distance between nodes. These paths cannot change the maximum distance between nodes since this is already a minimum. The effect of these paths on the distance between nodes is shown in Table II. The network S-X is the shuffle exchange with two self loops and MS-X is the modified network. It is evident that the additional paths do not provide a great decrease in the average minimum distance between nodes. They should

Table II—The average and maximum shortest distance between nodes in the shuffle-exchange network and the modified shuffle-exchange network

	Average		
Nodes	S-X	MS-X	Maximum
4	1.50	1.33	2
8	2.11	1.96	3
16	2.83	2.73	4
32	3.65	3.58	5
64	4.53	4.49	6
128	5.46	5.44	7
256	6.42	6.40	8

be included to allow an alternate path when failures occur, but unless a simple routing rule is found to use them under normal operation, they should not be used.

There are several problems with this type of a network. The first problem is that the physical layout of this network does not make sense geographically. If half of the nodes in the network are in one area and half of the nodes are in another remote area, then half of the connections must be between the two remote areas. Therefore, the network can only be used in a small area where the length of the interconnections do not make a difference. Shuffle-exchange networks in physically disjoint areas must be interconnected by a hierarchical network that can make sense geographically, as in the example in Section II. The second problem is that the network is only defined if the number of nodes equals 2^i. At present, no way has been found to add one node at a time—changing a small number of connections—and move from a network with 2^i nodes to a network with 2^{i+1} nodes. The third problem is that the alternate paths between a source and a destination are not good. If the preferred path is blocked or inoperable, the alternate paths are much longer.

3.3 Manhattan street network

The Manhattan street network is based on a grid of alternatingly directed streets and avenues, as shown in Fig. 1. The nodes exist on the corner of a street and an avenue. The rationale for this type of a network is that routing from a particular street and avenue to a destination should be straightforward. As in a city with this layout, any destination street and avenue can be found without asking directions, even when some roads are blocked. In addition, it should be possible to lay out the network to make sense geographically.

The principal difference between a grid connecting corners with streets and a grid connecting nodes with wires is that the physical constraints associated with a two-dimensional surface can be violated more easily with wires. For instance, in the example in Section II, the file servers and terminals forming a community of interest are in the same neighborhood. Assume that the file servers in this system are in the same room and that the personal computers in the same community of interest are in the same physical area. By connecting the file server to the region of the network with the terminals, rather than basing the connections strictly on the physical location of devices, the file server appears to be in the same neighborhood as the terminals. This reduces the interference between terminals in different communities of interest.

The difference in physical constraints also allows the extremes of the grid to be connected. These connections form the grid on the surface of a torus instead of a flat surface. The advantage of this cyclic surface is that there are no corners. Therefore, the maximum distance from a source to a destination is not the distance between two corners of the grid, but the distance between the center and one of the corners. The graph can also be flipped so that the links leaving the center node are always pointed in the same direction. This allows the same routing decision function to be used at every node.

Consider a network with r rows and c columns. The current node has coordinates (i_s, j_s), and the destination node has coordinates (i_d, j_d). The current node is considered to be at location (0, 0), and the relative location of the destination (i, j), is expressed as

$$i = \left\{[1 - 2(j_s \bmod 2)](i_d - i_s) + \frac{r}{2} - 1\right\} \bmod r - \left(\frac{r}{2} - 1\right)$$

$$j = \left\{[1 - 2^*(i_s \bmod 2)](j_d - j_s) + \frac{c}{2} - 1\right\} \bmod c - \left(\frac{c}{2} - 1\right).$$

The current node is now in the center of the network. The value of i is between $-(r/2 - 1)$ and $r/2$, and j is between $-(c/2 - 1)$ and $c/2$. The factors $1 - 2^*(j_s \bmod 2)$ and $1 - 2^*(i_s \bmod 2)$ guarantee that the links leaving the current node point toward increasing i and j. The routing decision now depends only on the relative location of the destination and not on the current node.

The routing preference from the central node to outlying nodes for a 12 × 12, 12 × 14, and 14 × 14 Manhattan street network is shown in Fig. 4. In this network, the two links emanating from the central node are directed upwards and to the right. The routing preference is the shortest distance from the central node to the destination when the link to the right is taken, minus the shortest path to the destination

	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
6	0	4	0	4	0	4	0	0	0	0	0	0
5	4	4	4	4	4	4	0	0	0	0	0	0
4	0	4	4	4	4	4	4	0	0	0	0	0
3	4	4	4	4	4	4	0	0	0	0	0	0
2	0	4	4	4	4	4	4	0	0	0	0	0
1	4	4	4	4	4	4	0	-4	0	-4	0	0
0	0	0	4	0	4	0	-4	-4	-4	-4	-4	-4
-1	0	0	0	0	0	-4	-4	-4	-4	-4	-4	0
-2	0	0	0	0	0	0	-4	-4	-4	-4	-4	-4
-3	0	0	0	0	0	-4	-4	-4	-4	-4	-4	0
-4	0	0	0	0	0	0	-4	-4	-4	-4	-4	-4
-5	0	0	0	0	0	0	-4	0	-4	0	-4	0

	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
6	2	0	4	0	4	0	4	0	0	0	0	0	0	0
5	2	4	4	4	4	4	4	0	0	0	0	0	0	2
4	2	2	4	4	4	4	4	4	0	0	0	0	0	0
3	2	4	4	4	4	4	4	0	0	0	0	0	0	2
2	2	2	4	4	4	4	4	4	0	0	0	0	0	0
1	2	4	4	4	4	4	4	0	-4	0	-4	0	-2	2
0	-2	2	0	4	0	4	0	-4	-4	-4	-4	-4	-4	-2
-1	0	0	0	0	0	0	-4	-4	-4	-4	-4	-4	-2	-2
-2	-2	0	0	0	0	0	0	-4	-4	-4	-4	-4	-4	-2
-3	0	0	0	0	0	0	-4	-4	-4	-4	-4	-4	-2	-2
-4	-2	0	0	0	0	0	0	-4	-4	-4	-4	-4	-4	-2
-5	0	0	0	0	0	0	0	-4	0	-4	0	-4	0	-2

	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
7	2	2	2	2	2	2	2	-2	0	-2	0	-2	0	0
6	2	2	4	2	4	2	4	2	0	0	0	0	0	0
5	2	4	4	4	4	4	4	0	0	0	0	0	0	2
4	2	2	4	4	4	4	4	4	0	0	0	0	0	0
3	2	4	4	4	4	4	4	0	0	0	0	0	0	2
2	2	2	4	4	4	4	4	4	0	0	0	0	0	0
1	2	4	4	4	4	4	4	0	-4	0	-4	0	-2	2
0	-2	2	0	4	0	4	0	-4	-4	-4	-4	-4	-4	-2
-1	0	0	0	0	0	0	-4	-4	-4	-4	-4	-4	-2	-2
-2	-2	0	0	0	0	0	0	-4	-4	-4	-4	-4	-4	-2
-3	0	0	0	0	0	0	-4	-4	-4	-4	-4	-4	-2	-2
-4	-2	0	0	0	0	0	0	-4	-4	-4	-4	-4	-4	-2
-5	0	0	0	0	0	0	-2	-4	-2	-4	-2	-4	-2	-2
-6	0	0	2	0	2	0	2	-2	-2	-2	-2	-2	-2	-2

Fig. 4—Routing preference in a 12 × 12, 12 × 14, and 14 × 14 Manhattan street network.

when the upwards-directed link is taken. Therefore, a negative number implies that the right link leads to the shortest path to the destination, and a positive number implies that the upwards link yields the shortest path to the destination. The magnitude of the number shows how much longer the distance to the destination would be if a packet were forced to take a less desirable path. A zero implies that the distance to the destination is the same along either path. The figures show that to get to half of the nodes either path can be taken, to get to a quarter of the nodes the left path should be taken, and to get to the other quarter of the nodes the right path should be taken. The figures also

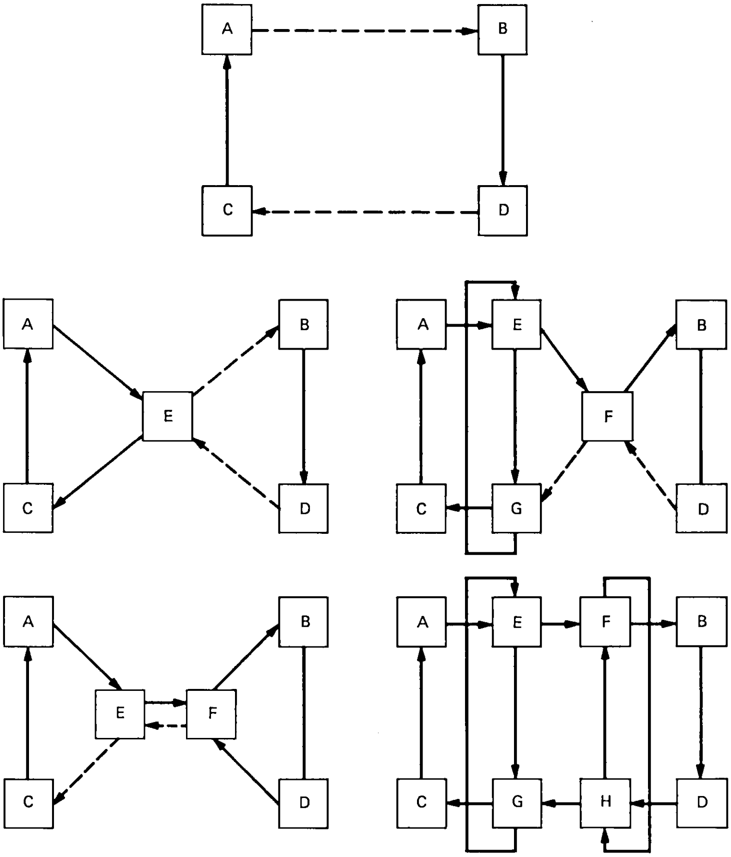


Fig. 5—Adding nodes E, F, G, and H one at a time to the basic rectangular structure consisting of nodes A, B, C, and D in a Manhattan street network.

show that if a packet is forced to take the wrong path, the increase in path length to the destination is never more than four.

One problem with the shuffle-exchange network is the difficulty in changing the number of nodes in the network. Figures 5 and 6 show how nodes may be added to the Manhattan street network. Figure 5 shows how two columns are added to the basic square structure within the Manhattan street network. The dotted lines show the links that will be broken when the next node is added. Figure 6 shows how the procedure is continued to add nodes to partially full columns. Each time a new node is added, two links are broken and connected to the new node. This is no greater than the number of links that must be broken in the bidirectional loop. Eventually this procedure leads to a network with two additional rows or columns, and the pattern of alternatingly directed rows and columns is preserved.

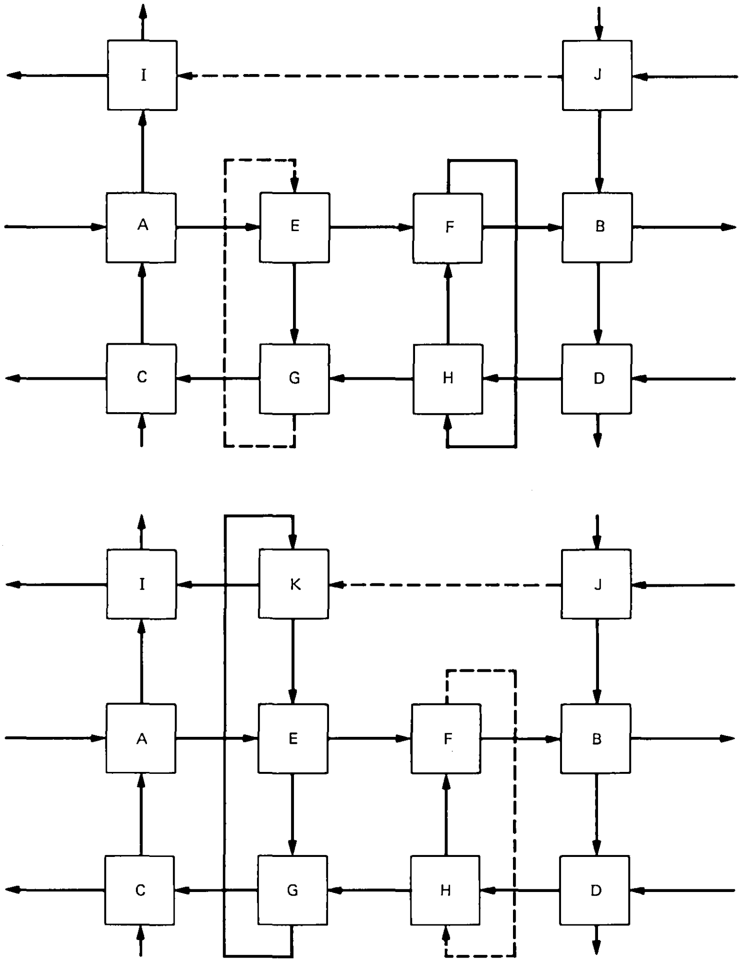


Fig. 6—Adding a node K to a partially full column in a Manhattan street network.

When adding a new node both the physical position of the node and the topology of the network must be considered. It is desirable to connect the node to the nearest existing nodes, but it is also desirable to start as few new rows or columns as possible, and to keep the number of rows and columns equal. When rows and columns are kept approximately equal, the average and maximum shortest paths between nodes increase as shown in Table III.

In the shuffle-exchange network it is occasionally better to establish a hierarchy of networks rather than make a single network larger. Hierarchical structures are also useful in Manhattan street networks. They are used to

Table III—Distances between nodes in 2*i* by 2*j* Manhattan street networks

			Shortest Paths Between Nodes	
Nodes	Rows	Columns	Average	Maximum
4	2	2	1.33	2
8	2	4	2.00	3
16	4	4	2.93	5
24	4	6	3.30	5
36	6	6	3.71	6
48	6	8	4.34	7
64	8	8	5.02	9
80	8	10	5.42	9
100	10	10	5.84	10
120	10	12	6.42	11
144	12	12	7.02	13
168	12	14	7.45	13
196	14	14	7.89	14
224	14	16	8.45	15
256	16	16	9.02	17

- Decrease the number of paths between physically distant sections of the network,
- Eliminate long paths between communities of interest, and
- Prevent traffic between communities of interest from affecting communications in other communities of interest.

The two-connected strategy can be maintained, as in Fig. 7. However, this will make routing more complex. An alternative is to connect one or more of the nodes in a local area to a higher-level network, as shown in Fig. 8. By using this approach, routing decisions in a local area are not affected by network changes in other areas, and addresses in different local areas are assigned independently. A hierarchical addressing and routing structure, similar to that used in the telephone system, can be used. For example, the address within the local area corresponds to a phone number, and the address of the local network on the higher-level network corresponds to the area code. When sending a packet within the local area an area code is not required.

3.4 An optimal two-connected network

Certain characteristics of the "best" networks are difficult to quantify. For instance, it should be possible to add nodes without making major reconfigurations, create geographically dispersed networks without adding excessive numbers of long links, and establish communities of interest. Other characteristics, such as the average and maximum number of links between nodes, can be compared and bounded.

Consider the class of two-connected networks. From a particular node, at most two nodes can be reached in one step, four additional

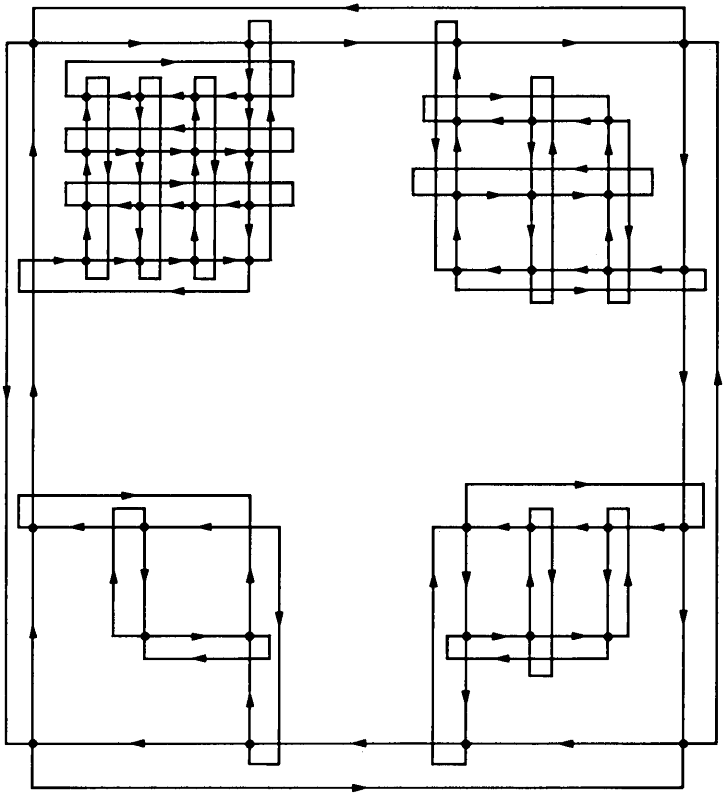


Fig. 7—A hierarchical Manhattan street network in which all of the nodes are two-connected.

nodes in two steps, and so on. The destination nodes form a binary tree. If, at any level in the tree, a destination node recurs, the number of new nodes that can be reached in future levels is reduced by the descendants of that node. Therefore, the maximum number of nodes that can be reached in m steps is

$$\sum_{i=1}^{i=m} 2^i = 2^{m+1} - 2.$$

If every node in a two-connected network can reach this number of new nodes in each m steps, then the network has the smallest average and maximum distance between nodes. In general, networks with these characteristics do not exist. However, this is a lower bound on these distance characteristics.

In the shuffle-exchange network with 2^j nodes, each node must reach $2^j - 1$ nodes, and the maximum minimum distance between

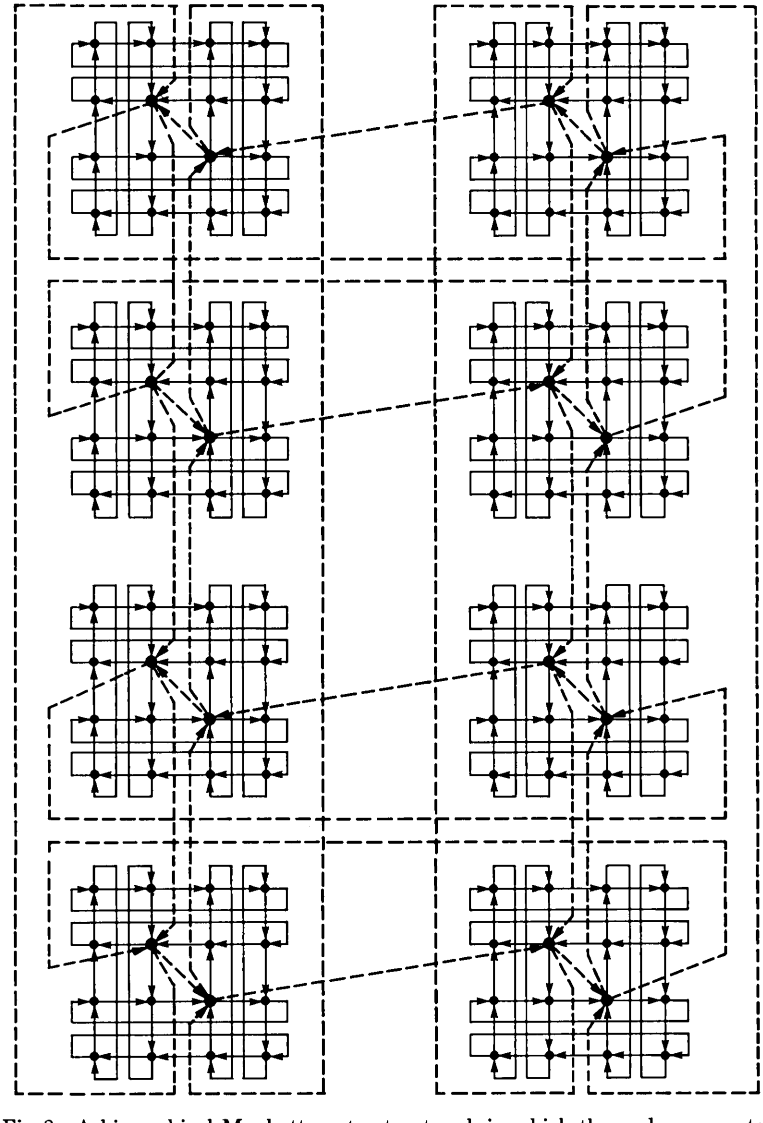


Fig. 8—A hierarchical Manhattan street network in which the nodes connected to the hierarchical network are four-connected.

nodes is j. The number of nodes that are reached in $j - 1$ steps in the optimum network is $2^j - 2$. Therefore, if $2^j - 1$ nodes must be reached on the optimum network, the minimum distance to the furthest node is j, and the largest minimum distance between nodes in the shuffle-exchange network is less than or equal to that in any network with 2^j nodes.

A comparison of the average and maximum distance between nodes

Table IV—A comparison of the distances between nodes in several networks

		Number of Nodes		
Net	Distance	16	64	256
Opt	Avg	2.53	4.19	6.06
	Max	4	6	8
S-X	Avg	2.73	4.49	6.40
	Max	4	6	8
MSN	Avg	2.93	5.02	9.02
	Max	5	9	17
BDL	Avg	4.26	16.25	64.25
	Max	8	32	128

for the optimum, shuffle-exchange, Manhattan street network, and the bidirectional loop is shown in Table IV. In both the optimum network and the shuffle-exchange network, the maximum distance between nodes varies as the log of the number of nodes. In the Manhattan street network, the maximum distance between nodes varies as the square root of the number of nodes, and, in the bidirectional loop, this distance varies linearly with the number of nodes. The same relationship is also noted between the number of nodes in these networks and the average distance between nodes. The average distance between nodes shows what fraction of the network resources is used to transfer a packet and provides an indication of the relative throughput of the networks. Although, as shown in Section II, there are other factors that also affect the throughput.

IV. IMPLEMENTATION

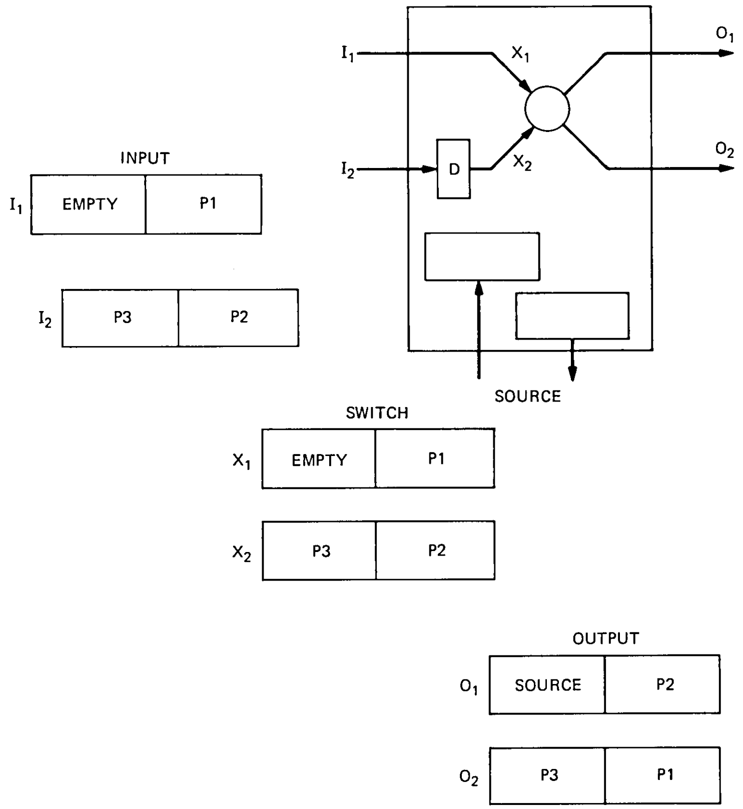
In a mesh network, as in a loop network, the communications lines are point-to-point links with a single transmitter and a single receiver. Transmission on these links is much simpler than in a broadcast network with a shared communication channel. The access protocols are simpler because there is only one source, and it is not necessary to multiplex users on the communication channel or resolve collisons. The receiver is simpler because the distance between the source and destination is constant, and the signal strength does not change by a large amount from packet to packet. Regenerating the signal to eliminate distance constraints is simpler because signals only propagate in one direction. Timing recovery is simpler because the source can transmit continuously and bit synchronization does not have to be reestablished at the beginning of each packet. In addition, the communication line does not have many taps and is compatible with the current generation of fiber-optic equipment.

In a two-connected network there are two links and a local source inputting data to a node, and two links and a local sink removing data from the node. Occasionally, multiple inputs try to transmit data to the same outgoing link. One way to resolve this problem is to queue packets waiting for a link. The network now assumes the complexity of a store-and-forward network. Not only must potentially large packet queues be maintained, but adaptive-routing, flow-control, deadlock-avoidance, and packet-resequencing issues must be addressed.

In this section, the slotted-system and register-insertion techniques, developed for loop communication systems, will be extended to mesh networks with equal in and out degrees. The general strategy guarantees that every packet arriving on an incoming link, and not destined for the node, will be transmitted on one of the outgoing links. Therefore, it is not necessary to maintain a packet queue for the links emanating from the node. The requirement that packets passing through the node take one of the outgoing paths results in longer paths when the shortest path to the destination is busy. However, it is possible to design networks to reduce the effects of incorrect paths. For instance, in the Manhattan street network the path to only half of the destinations is increased if a packet is forced to take one path rather than the other. In addition, if a packet is forced to take a less desirable path, the distance to the destination is increased by at most four.

The storage between the local source and the network is also limited. Packets from the local source are only transmitted when one of the outgoing links is not being used by an incoming link. It is assumed that either the local source can be throttled when the network is busy, or that the source provides data at a low rate relative to the network transmission rate. In the latter case, when a packet is lost, it must be recovered by a higher-level protocol. If the network delivers packets faster than the local sink can accept them, packets are either transmitted on one of the outgoing links or discarded. In the former case, the network is used for storage. Since new packets cannot enter the network when it is recirculating old packets, this transmission strategy acts as a flow-control mechanism. The assumptions on the local source and sink are implicit in all loop-configured systems without infinite storage.

The packets of data in a slotted system are fixed size. A node continuously transmits bits on each of the links emanating from the node, and periodically transmits a start-of-slot indication. The start-of-slot indication is followed by a packet of data or an empty slot. In the interval between the start-of-slot transmissions, at most one packet of data is received on each of the incoming links. The packets that are received between start-of-slot transmissions are forwarded after the start of slot is transmitted. These packets are switched to one of the outgoing links or the local sink before data from the local



INPUT

I_1	EMPTY	P1
I_2	P3	P2

SWITCH

X_1	EMPTY	P1
X_2	P3	P2

OUTPUT

O_1	SOURCE	P2
O_2	P3	P1

Fig. 9—Extension of slotted-loop systems to a mesh network.

source are given access to the slot. Since there are the same number of links arriving and leaving from each node, and the local source can be throttled, a queue of packets will not accumulate. The operation of a slotted system without a packet queue is shown in Fig. 9.

The interface for a register-insertion loop is shown in Fig. 10. Packets in this system are variable in size, but constrained to be less than the storage register W_l. The local source is only allowed to transmit when register W_l is empty. Since a packet from the local source is less than W_l, all data received from the loop while the local source is transmitting can be stored in W_l. When register W_l is not empty, bits from this register are transmitted on the loop. Therefore, the length of this register remains the same when bits are being received, and decreases when bits are not received. As long as this register is not empty, gaps between arriving packets are removed.

The register-insertion technique can be applied to a mesh network in which the in degree and out degree of the nodes are the same by

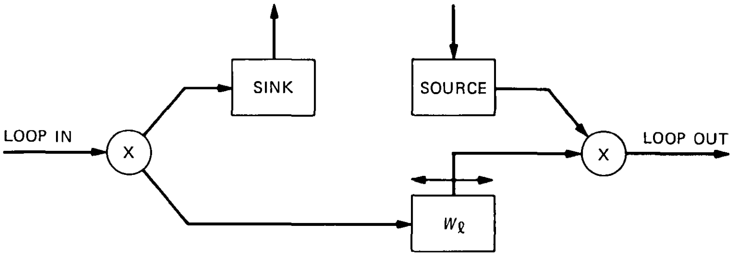


Fig. 10—A register-insertion access unit in a loop system.

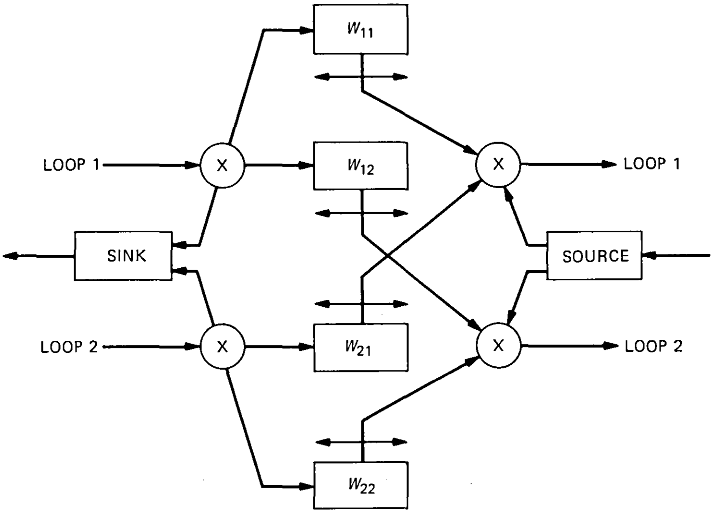


Fig. 11—Extension of register-insertion systems to a mesh network.

making the node appear as if several loops are passing through it. This is shown in Fig. 11. Registers W_{11} and W_{22} correspond to register W_l for loops 1 and 2, respectively. In addition to the local sink, register W_{12} appears to be a sink for loop 1. And, in addition to the local source, register W_{12} appears to be a source of data for loop 2. Therefore, register W_{12} allows messages on loop 1 to transfer to loop 2. As in a loop system, if buffer W_{12} is full the packet must continue around loop 1. Buffer W_{21} serves the same purpose for packets crossing from loop 2 to loop 1.

The register-insertion technique allows variable-length packets. When the system is busy, each node eliminates the null space between incoming packets to efficiently use the outgoing links. The slotted-system technique uses fixed-size packets. If there is less than a packet of data, the packet is partially empty. Therefore, the register-insertion technique uses the channel more efficiently. In a slotted system, the packets from the incoming links are aligned at the switching point. A packet only traverses a longer path if more than one incoming packet requires the same outgoing channel. In a register-insertion system, a packet can only transfer from one loop to the other if the crossover buffer is empty. It is possible that both packets on the through loops would rather be on the other loop, but cannot cross over because the buffers are full. Therefore, the register-insertion technique uses individual links more efficiently, but the slotted system takes a shorter path between the source and destination. The technique that provides the greater throughput depends on both the message-length distribution and the network topology.

A small amount of buffering can be included in either the slotted or register-insertion system to reduce the probability of a packet taking a longer path. In the slotted system, fixed-size packet buffers are inserted at the output channels. The probability that a packet must take a longer path is the probability that two arriving packets must take the same path and the buffer for that path is full. Without buffering, one packet must take the longer path whenever two arriving packets want to take the same path. In the register-insertion system, the additional storage is inserted at the crossover point. A packet cannot cross over if, when it is received, there are fewer bits available in the crossover buffer than there are in a maximum-size packet. In the original system, a packet cannot cross over if, when it arrives, the crossover buffer is not empty. Decreasing the probability that a packet takes a longer path decreases the fraction of the network resources that a packet uses, and increases the throughput of the system. The trade-off between buffering and system throughput remains to be investigated.

V. FILE TRANSFER

A file transfer consists of several packets being transmitted from a source to the same destination. In a system in which packets do not take the same path, it is possible that packets are not received in the same order that they are transmitted. Packets may be resequenced at the receiver, however, it is preferable to avoid this task.

One possible solution to this problem is to transmit one packet at a time and wait for an acknowledgment. This reduces the file-transfer rate. However, because of the small delays at each node, this solution is not as bad in mesh networks as it is in store-and-forward networks. For instance, in a slotted system the average delay per node is half a slot time, and the average round trip delay equals the average number of hops between nodes, $\bar{L}$, times the slot time. Therefore, there is an average of $\bar{L}$ slots between each packet in the file, and the file-transfer rate equals the channel rate divided by $\bar{L} + 1$. Higher file-transfer rates can be achieved by end-to-end protocols that take advantage of the delay characteristics of the system, or node protocols that take advantage of specific hardware structure.

Because of the small amount of delay at each node, it is unlikely that packets that take routes that have approximately the same length will arrive out of sequence. This probability can be reduced by allowing a small number of slots between packets in the same file transfer. A simple file-transfer protocol, which takes advantage of this characteristic, operates like the window protocols of store-and-forward networks and the go-back-N protocols of satellite systems. This protocol labels packets in a file transfer with a sequence number and a retry number. At the beginning of a file transfer, the transmitter and receiver start with the same sequence and retry number. The sequence number is the order of the packets. The receiver

- Increments its retry number when a packet with the expected retry number and a larger sequence number is received,
- Sends a positive acknowledgment if a packet has a sequence less than or equal to the expected number,
- Sends a negative acknowledgment with its retry number and expected sequence number if the packet has a larger sequence number than it expects, and
- Commits a packet if it has the expected sequence number.

The transmitter

- Stops saving a packet for retransmission when it receives a positive acknowledgment for a packet with a sequence number greater than or equal to the expected number,
- Adopts a new retry number and starts retransmitting from a negatively acknowledged sequence number if a negative acknowledgment with a larger retry number is received, and
- Periodically retransmits the last packet in a file transfer until it receives an acknowledgment.

The transmitter initially transmits packets in the file transfer in every available slot. However, when it receives negative acknowledgments, it increases the number of slots between subsequent packets.

Since this protocol only accepts packets in the correct order, packets do not have to be resequenced at the receiver. In a mesh network, several packets can be in transit between the source and destination. If the receiver misses a packet, it must send a negative acknowledgment for every packet with a larger sequence number than expected to be certain that the trasmitter receives the negative acknowledgment. The retry number is included so that the transmitter only backs up and starts retransmitting when it receives the first negative acknowledgment to an outstanding packet.

The transmitter adaptively changes the number of slots between packets according to network load and the rate of the receiver. When the network is lightly loaded, all packets follow the best path to the receiver, and arrive in sequence. If the receiver can accept packets at this rate, there are no negative acknowledgments, except for infrequent transmission errors, and the file-transfer rate equals the channel rate. When the network is heavily loaded, the packets follow different paths, are received out of sequence, and the file-transfer rate decreases. If the receiver cannot accept packets as fast as the transmitter can deliver them, the buffer in the interface unit is full when the packets arrive. In the systems described, these packets are directed to one of the output links at the node, and recirculate in the network until the buffer is available. These packets arrive out of sequence, negative acknowledgments are transmitted, and the transmitter slows down.

Additional improvements are possible by taking the structure of the nodes into account. In a slotted system, subsequent packets in a file transfer can be marked. At a node, packets in a file transfer, which follow immediately behind one another, can be directed along the same path. The end-to-end protocol can be implemented with empty slots only occurring when the source cannot deliver packets quickly enough, or when the channel at the source node is busy with traffic passing through the node. This will improve the file-transfer rate on moderately used channels. If this modification is used, the transmitter and receiver must negotiate the number of packets in a continuous sequence to be certain that the receiver can accept them at the channel rate.

In a register-insertion system, if the loop paths define a Hamiltonian circuit, packets in a file transfer can be constrained to follow these loops. Since a packet can never be denied access to this loop, all packets follow the same path and will be received in sequence. This allows file transfers to occur at the channel rate without resequencing, but requires file transfers to use a larger fraction of the network resources.

In both the register-insertion and slotted systems, it is possible to use a higher-level protocol to set up a limited number of virtual circuits along which file transfers can occur efficiently at the channel rate without resequencing. In the slotted system, the higher-level protocol is used to assign an input at a node to an output. When a file-transfer packet arrives at a node input, it is given first priority to the assigned output. Therefore, all packets in the file transfer follow the same path and do not have to be resequenced. The function of the higher-level protocol is to make the assigned paths for a file transfer use as few links as possible. The problem with this approach is that the preferred paths must be established at the beginning of each file transfer and disabled at the end of the transfer. In addition, a file transfer may be temporarily blocked by previously assigned paths, creating a need for a scheduler. In the register-insertion system, the preferred paths are established as the paths through the node. This has the same problems as the slotted system.

VI. CONCLUSION

Mesh networks increase the throughput of conventional local area networks by decreasing the fraction of the network capacity needed to transmit information between a source and a destination. These networks have multiple paths between each source and destination, thus increasing the reliability of local networks. The networks consist of point-to-point links, and can be extended to cover a metropolitan, rather than a local, area.

In general, mesh networks require complex store-and-forward nodes that also route messages, control the flow of data entering the network, resequence packets at the destination, and recover packets with errors. There are characteristics of the local or metropolitan area that allow these functions to be simplified. In the local environment, regular network topologies can be selected in which routing is straightforward. The lower error rates make it reasonable to recover errors on an end-to-end basis. This allows loop-access protocols to be extended to mesh networks, eliminating the need for buffering and additional flow-control protocols. Extensions for the slotted system and register-insertion techniques used in loop systems have been shown. Buffering can be included in these systems to improve the channel utilization; however, channels are less expensive in the local environment. The small node delays in these systems also make it reasonable to implement file-transfer protocols that do not require packet resequencing.

Three mesh networks have been studied, and the desirable characteristics of networks have been determined. Networks should have regular structures with simple routing rules, and should not have single points of failure. By minimizing the average and maximum distance between nodes, the fraction of the network resources used to transmit a packet decreases, and the throughput increases. This can be done by packing topologies with these characteristics noted, and by locating communities of terminals that communicate frequently in the same area of the network. Equal-length alternate paths between sources and destinations reduce the probability of bottlenecks and the need for buffering within a node. Networks will change, and it must be possible to add or delete nodes without changing a large number of connections. If the network covers a large area, it must be possible to limit the connections between nodes in different areas. Of the networks studied, the Manhattan street network has all of these characteristics.

REFERENCES

1. E. H. Steward, "A Loop Transmission System," Conf. Rec. Intern. Conf. Commun., San Francisco, June 1970, pp. 36-1-9.
2. N. Abramson, "The ALOHA-System—Another Alternative for Computer Communications," University of Hawaii Tech. Rep. B70-1, April 1970, AD707853.
3. R. M. Metcalf and D. R. Boggs, "Ethernet: Distributed Packet Switching for Local Computer Networks," Commun. ACM, *19* (July 1976), pp. 395–404.
4. M. Schlatter and J. L. Massey, "Capacity of Interconnection Ring Communication Systems With Unique Loop-Free Routing," IEEE Trans Inform. Theory, *IT-29,* No. 5 (September 1983), pp. 774–8.
5. J. R. Pierce, "How Far Can Loops Go," IEEE Trans. Commun., *COM-20,* No. 3 (June 1972), pp. 527–30.
6. Y. Yemini, "Tinkernet: Or Is There Life Between LANs and PBXs," Proc. ICC'83, *3,* Boston, June 1984, pp. 1501-5.
7. T. N. Saadawi and M. Schwartz, "Distributed Switching for Data Transmission Over Two-Way CATV," Proc. ICC'84, Amsterdam, May 1984, pp. 1409–13.
8. D. C. Walden, "Experiences in Building, Operating and Using the ARPA Network," Second USA-Japan Computer Conference, Tokyo, August 1975.
9. C. Petitpierre, "Meshed Local Computer Networks," IEEE Commun. Mag., *22,* No. 4 (August 1984), pp. 36–40.

AUTHOR

Nicholas F. Maxemchuk, B.S.E.E., 1968, City College of New York; M.S.E.E., 1970, Ph.D., 1975, University of Pennsylvania; RCA David Sarnoff Research Center 1968–1976; AT&T Bell Laboratories, 1976—. Mr. Maxemchuk is presently Head of the Distributed Systems Research Department. Since joining AT&T Bell Laboratories, he has done research on computer-communication networks, virtual and speech editing, and picture processing. From 1980 to the present, he has been on the adjunct faculty of the University of Pennsylvania, where he teaches a course on computer communications networks. From 1980 to 1985, he was the Associate Editor, then the Editor of Data Communications for the IEEE Transactions on Communications. Member, Eta Kappa Nu, Tau Beta Pi.

AT&T Technical Journal
Vol. 64, No. 7, September 1985
Printed in U.S.A.

Bandwidth-Error Exchange for a Simple Fading Channel Model

By B. F. LOGAN, JR.*

(Manuscript received January 30, 1985)

It is assumed that a data carrier signal is transmitted over a fading channel whose frequency response can be closely approximated over the transmission band by a first-order polynomial in frequency, the coefficients being slowly varying functions of time. An equivalent baseband model is obtained wherein the transmitted signal is of the form $s(t) = \sum_{-\infty}^{\infty} a_k f(t - kT)$, and the received signal is of the form $r(t) = s(t) + x(t)s'(t)$, where $x(t)$ is an unknown (e.g., random) function of time. The problem solved in this paper is that of finding the function f of prescribed bandwidth that minimizes the mean-square error, $E\{|r(nT) - a_n|^2\}$, under the assumption that the a_k are independent random variables of zero mean and unit variance, and $E\{|x(t)|^2\} = \alpha^2$. The results also apply to the sometimes more realistic hypothesis, $|x(t)| \leq \alpha$.

I. INTRODUCTION

A common method of transmitting data $\{a_k\}$ and $\{b_k\}$ is via a carrier signal of the form

$$s_c(t) = s_1(t) \cos \omega_c t + s_2(t) \sin \omega_c t, \tag{1}$$

where

$$s_1(t) = \sum_{-\infty}^{\infty} a_k f(t - kT), \qquad s_2(t) = \sum_{-\infty}^{\infty} b_k f(t - kT). \tag{2}$$

In the usual mathematical model, $f(t) = (\sin \Omega t)/\Omega t$, $\Omega = \pi/T$, so that

* AT&T Bell Laboratories.

$$s_1(nT) = a_n; \quad s_2(nT) = b_n.$$

Assuming then that $s_c(t)$ is transmitted over an ideal noiseless channel, $s_1(t)$ and $s_2(t)$ can be recovered by synchronous demodulation of the received signal, and then the data obtained by sampling s_1 and s_2 at the times nT.

A problem, communicated by G. Foschini, arises when $s_c(t)$ is transmitted over a channel having a relatively slow-varying fading characteristic. L. J. Greenstein and B. A. Czekaj have found that, in many cases, the fading channel response can be fairly well approximated over the transmission band by a first-order polynomial in frequency ω,

$$A_0(t) + A_1(t)\{\omega - \omega_c\} + B_1(t)i\{\omega - \omega_c\},$$

with slowly varying real coefficients.[1] Then, to a good approximation, synchronous demodulation of the received signal gives, instead of $s_1(t)$,

$$r_1(t) = A_0(t)s_1(t) + B_1(t)s_1'(t) + A_1(t)s_2'(t),$$

and a similar expression for the alteration of $s_2(t)$.

We assume that $A_0(t)$, the center-frequency channel gain, is positive and can be determined at the receiver (e.g., by measuring the average power in a narrow band about the center frequency), so that, by the use of automatic gain control we have available,

$$r_1^*(t) = s_1(t) + x_1(t)s_1'(t) + x_2(t)s_2'(t), \tag{3}$$

where $x_1(t)$ and $x_2(t)$ are unknown, for example, random, slowly varying functions of time. Then

$$r_1^*(nT) = a_n + x_1(nT)s_1'(nT) + x_2(nT)s_2'(nT). \tag{4}$$

It has been suggested,[2] as an alternative to using adaptive channel compensation, that error-free reception may be obtained by doubling the bandwidth of $f(t)$ in (2), i.e., by taking $f(t) = (\sin \Omega t)^2/(\Omega t)^2$, $\Omega = \pi/T$, so that $s_1'(nT) = s_2'(nT) = 0$. Here we want to determine the best attainable trade-off between mean-square error and bandwidth under certain assumptions on $\{a_k\}$, $\{b_k\}$, $x_1(t)$, and $x_2(t)$.

We have, from (2),

$$r_1^*(nT) - a_n = a_n\{f(0) - 1\} + \sum_{k\neq n} a_k f(nT - kT)$$
$$+ x_1(nT)\sum_{-\infty}^{\infty} a_k f'(nT - kT) + x_2(nT)\sum_{-\infty}^{\infty} b_k f'(nT - kT). \tag{5}$$

Now we assume that the a_k and b_k are independent random variables of zero mean and unit variance. Then the expected mean-square error over $\{a_k\}$ and $\{b_k\}$ is

$$E\{|r_1^*(nT) - a_n|^2\} = |1 - f(0)|^2 + \sum_{k\neq 0} |f(kT)|^2$$

$$+ \{x_1^2(nT) + x_2^2(nT)\} \sum_{-\infty}^{\infty} |f'(kT)|^2 + 2x_1(nT) \sum_{k\neq 0} f'(kT)f(kT)$$

$$+ 2x_1(nT)[f(0) - 1]f'(0). \quad (6)$$

[Note that the cross product $x_1(nT)x_2(nT)$ does not enter here because of the assumptions on $\{a_k\}$ and $\{b_k\}$.]

Now let us suppose that $x_1(t)$ and $x_2(t)$ are random (continuous) functions satisfying

$$E\{|x_1(t)|^2\} = \alpha_1^2, \qquad E\{x_1(t)\} = 0, \quad (7a)$$

$$E\{|x_2(t)|^2\} = \alpha_2^2. \quad (7b)$$

Then we have, for the expected, or mean-square error, ϵ^2,

$$\epsilon^2 = |1 - f(0)|^2 + \sum_{k\neq 0} \{f(kT)|^2 + \alpha^2 \sum_{-\infty}^{\infty} |f'(kT)|^2, \quad (8)$$

where $\alpha^2 = \alpha_1^2 + \alpha_2^2$.

As an alternative to the statistical assumptions, (7a) and (7b), let us suppose that

$$|x_1(t)| \leqslant \alpha_1, \quad (9a)$$

$$|x_2(t)| \leqslant \alpha_2. \quad (9b)$$

These assumptions may be more relevant in practice than the statistical assumptions. Note that if the coefficient of $x_1(nT)$ in (6) vanishes, which will be the case if $f(t)$ is even; then, under the assumptions (9a) and (9b), (8) will hold with the equality sign replaced by $\leqslant$. We wish to minimize the quantity on the right in (8) over bandlimited functions f of prescribed bandwidth. As we shall see, the minimum will be obtained for a real-valued even function, so that, indeed, the minimum will be an upper bound for ϵ^2 under the assumptions (9a) and (9b).

Note that, in the end, the quantity to be minimized depends only on how large $|x_1(t)|$ and $|x_2(t)|$ may be, slowly varying or not. The slowly varying hypothesis was used only to obtain (3) from the fading channel model. The same minimization problem is obtained, under the previous assumption on $\{a_k\}$, if

$$s(t) = \sum_{-\infty}^{\infty} a_k f(t - kT) \quad (10)$$

is transmitted over a channel such that the received signal is simply

$$r(t) = s(t) + x(t)s'(t), \quad (11)$$

where $x(t)$ is an unknown (continuous) function satisfying, if $x(t)$ is random,

$$E\{|x(t)|^2\} = \alpha^2, \qquad E\{x(t)\} = 0, \tag{11a}$$

or (say) if not, then

$$|x(t)| \leqslant \alpha. \tag{11b}$$

It is convenient, and sufficient, to consider the case $T = 1$ in (8), so that the quantity of interest is

$$\epsilon^2 = |1 - f(0)|^2 + \sum_{k\neq 0} f(k)|^2 + \alpha^2 \sum_{-\infty}^{\infty} |f'(k)|^2. \tag{12}$$

This is to be minimized over functions f in $B_2(\Omega)$; i.e., functions in L_2 of the form

$$f(t) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} F(\omega)e^{i\omega t}d\omega. \tag{13}$$

We need only consider the case $0 < \Omega \leqslant 2\pi$, since for $\Omega \geqslant 2\pi$ we can make $\epsilon^2 = 0$ by taking $f(t) = (\sin \pi t)^2/(\pi t)^2$. The transmission system, of course, is useless if $\epsilon^2 \geqslant 1$, but we can always make $\epsilon^2 < 1$ by appropriate choice (or scaling) of f, ($f \equiv 0$ giving $\epsilon^2 = 1$).

In general, we will not have $f(0) = 1$ for the optimal f. This is easily seen by defining the governing quantity in the problem, viz.,

$$\mu(\Omega; \alpha) = \inf_{\substack{f\in B_2(\Omega)\\ f(0)=1}} \left\{ \sum_{k\neq 0} |f(k)|^2 + \alpha^2 \sum_{-\infty}^{\infty} |f'(k)|^2 \right\}. \tag{14}$$

It follows from this definition that

$$\sum_{k\neq 0} |f(k)|^2 + \alpha^2 \sum_{-\infty}^{\infty} |f'(k)|^2 \geqslant \mu(\Omega; \alpha)|f(0)|^2, \qquad f \text{ in } B_2(\Omega). \tag{15}$$

Thus from (12) and (15) we have

$$\epsilon^2 \geqslant |1 - f(0)|^2 + |f(0)|^2\mu(\Omega; \alpha), \qquad f \text{ in } B_2(\Omega). \tag{16}$$

The quantity on the right is minimized by taking

$$f(0) = \gamma = \gamma(\Omega; \alpha) = \{1 + \mu(\Omega, \alpha)\}^{-1}, \tag{17}$$

giving

$$\epsilon^2 \geqslant \frac{\mu(\Omega; \alpha)}{1 + \mu(\Omega; \alpha)}. \tag{18}$$

Thus if the infimum in (14) is attained for $f = f(t; \Omega, \alpha)$, then equality will hold in (18) for the optimal function

$$f_0(t; \Omega, \alpha) = \gamma f(t; \Omega, \alpha). \tag{19}$$

In cases of practical interest, μ will be small, and hence $f_0(0)$ will be slightly less than 1.

II. RESULTS

The results are summarized in the following:

Theorem: Define for $\alpha > 0$, $\Omega > 0$,

$$\rho(\Omega;\ \alpha) = \frac{\pi}{\Omega} \cdot \frac{\alpha\Omega}{\arctan(\alpha\Omega)}, \tag{20}$$

where $\arctan(\cdot)$ *is between* 0 *and* $\pi/2$. *Then in (14) we have, for* $0 < \Omega \leqslant \pi$,

$$\mu(\Omega;\ \alpha) = \rho(\Omega;\ \alpha) - 1, \tag{21}$$

and for $\pi < \Omega \leqslant 2\pi$,

$$\mu(\Omega;\ \alpha) = \left\{(1 - \beta) + \beta \cdot \frac{\arctan(\alpha\beta\pi)}{\alpha\beta\pi}\right\}^{-1} - 1, \tag{22}$$

where

$$\beta = 2 - (\Omega/\pi). \tag{22a}$$

For $\Omega \geqslant 2\pi$,

$$\mu(\Omega;\ \alpha) = 0. \tag{23}$$

Furthermore, for $0 < \Omega \leqslant \pi$, *the infimum in (19) is attained only for*

$$f(t) = f(t;\ \Omega,\ \alpha) = \rho(\Omega;\ \alpha)\,\frac{1}{2\pi}\int_{-\Omega}^{\Omega} \frac{\cos \omega t}{1 + \alpha^2\omega^2}\,d\omega \tag{24}$$

and for $\pi < \Omega \leqslant 2\pi$, $\alpha > 0$, *only for*

$$\begin{aligned} f(t) &= (1 - \lambda)f(t;\ \beta\pi,\ \alpha) \\ &\qquad + \lambda\{\phi(t;\ \Omega)\cos \pi t - \frac{1}{\pi}\,\phi'(t;\ \Omega)\sin \pi t\}, \end{aligned} \tag{25}$$

where $f(t;\ \cdot,\ \alpha)$ *is defined in (24),* β *is defined in (22a), and*

$$\lambda = \frac{1 - \beta}{(1 - \beta) + \beta \cdot \dfrac{\arctan(\alpha\beta\pi)}{\alpha\beta\pi}}, \tag{25a}$$

$$\phi(t;\ \Omega) = \frac{\sin(\Omega - \pi)t}{(\Omega - \pi)t}. \tag{25b}$$

III. DISCUSSION

In the simple fading channel model we have fixed the sampling interval T to be unity, so that $\Omega = \pi$ corresponds to Nyquist-rate transmission, and solved the minimization problem for $0 < \Omega \leqslant 2\pi$. The case $0 < \Omega < \pi$ might be considered uninteresting, for in this case the bandwidth is too small for the data rate. However, the solution for this "uninteresting" case enters in the solution for the interesting case, $\pi < \Omega < 2\pi$.

From the results stated in the Theorem, the optimal function in $B_2(\Omega)$ for minimizing the mean-square error is found to be

$$f_0(t;\, \Omega,\, \alpha) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} \frac{\cos \omega t}{1 + \alpha^2\omega^2}\, d\omega, \qquad 0 < \Omega < \pi, \tag{26}$$

$$f_0(t;\, \Omega,\, \alpha) = f_0(t;\, 2\pi - \Omega,\, \alpha) + h_0(t;\, \Omega), \qquad \pi < \Omega \leqslant 2\pi, \tag{27}$$

where

$$h_0(t;\, \Omega) = \left\{\frac{\sin(\Omega - \pi)t}{\pi t}\right\}\cos \pi t - \frac{1}{\pi}\left\{\frac{d}{dt}\,\frac{\sin(\Omega - \pi)t}{\pi t}\right\}\sin \pi t$$
$$= \frac{1}{2\pi}\int_{\Omega_1 < |\omega| < \Omega} \left(1 - \frac{|\omega|}{2\pi}\right)\cos\omega t\, d\omega, \ (\Omega_1 = 2\pi - \Omega). \tag{28}$$

The resulting minimum mean-square error is

$$\epsilon_0^2(\Omega;\, \alpha) = 1 - \frac{\Omega}{\pi}\,\frac{\arctan(\alpha\Omega)}{\alpha\Omega}, \qquad 0 < \Omega \leqslant \pi, \tag{29}$$

$$\epsilon_0^2(\Omega;\, \alpha) = \left(2 - \frac{\Omega}{\pi}\right)\left\{1 - \frac{\arctan[\alpha(2\pi - \Omega)]}{\alpha(2\pi - \Omega)}\right\}, \qquad \pi < \Omega \leqslant 2\pi, \tag{30}$$

where $0 \leqslant \arctan(\cdot) < \pi/2$.

The main interest attaches to this quantity for $\pi \leqslant \Omega \leqslant 2\pi$ and α small, in which case

$$\epsilon_0^2(\Omega;\, \alpha) \doteq \frac{\alpha^2\pi^2}{3}\,\{2 - (\Omega/\pi)\}^3. \tag{31}$$

So there is an interesting trade-off between error and bandwidth in this case.

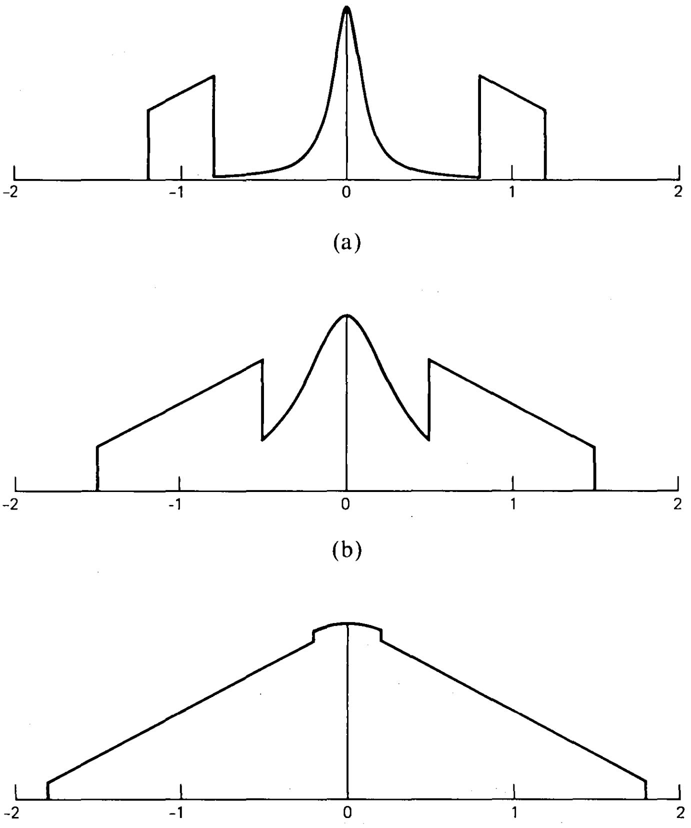
In the other direction we have

$$\lim_{\alpha\to\infty} \epsilon_0^2(\Omega;\, \alpha) = 2 - (\Omega/\pi), \qquad \pi \leqslant \Omega \leqslant 2\pi. \tag{32}$$

Thus, if α is very large, Ω must be very near 2π in order for the error to be small; i.e., one may as well take $\Omega = 2\pi$ for very large α. However, other practical considerations enter in the case of large α; e.g., the necessities of very accurate sampling and very close approximation of the extremal function. Also, additive noise, which has been neglected in this analysis, will be magnified by automatic gain control during periods of deep fading. So the results are deemed of no practical value in the case of large α.

Notice that the extremal function for the case $\pi < \Omega < 2\pi$ has two distinct components, the low-frequency component being the extremal function for the frequency $(2\pi - \Omega)$. The other component, $h_0(t; \Omega)$ is a bandpass function that does not depend on α, since as may be seen from the first line in (28), its derivative vanishes at the integers. From the second line in (28), it is seen that $h_0(t; \Omega)$ is a bandpass version (center frequency π, upper frequency Ω) of $(\sin \pi t)^2/(\pi t)^2$, the extremal function for $\Omega = 2\pi$. Graphs of $F_0(\omega; \Omega, \alpha)$, the Fourier transform of $f_0(t; \Omega, \alpha)$, are shown in Fig. 1 for several values of Ω and α^2.

It is doubtful that detailed statistics on fading channels, even if available, would be useful in practice, except to the extent that they could give a rough idea of what might be expected. Generally speaking, large errors cannot be tolerated over long intervals (negating, to some



(a)

(b)

(c)

Fig. 1—The Fourier transform $F_0(\omega/\pi; \Omega, \alpha)$ of the extremal function for various values of Ω and α. (a) $\Omega = 1.2\pi$, $\alpha^2 = 10$. (b) $\Omega = 1.5\pi$, $\alpha^2 = 1$. (c) $\Omega = 1.8\pi$, $\alpha^2 = 0.1$.

extent, the adoption of the mean-square error criterion), and one is faced with the dilemma of uncertainty in using such channels. In the model here, the "expected value of α^2" might be regarded from a pragmatical viewpoint as a mathematical non sequitur which might better be replaced by a design hypothesis, $|x(t)| \leq \alpha$. Then the use of $f_0(t;\ \Omega,\ \alpha)$ guarantees that the mean-square error will not exceed $\epsilon_0^2(\Omega;\ \alpha)$ if the hypothesis is true. The design problem is much like that of deciding how much insurance to buy.

On the other hand, one might adopt a hedging strategy, where extra bandwidth is used to *guard* against fading, while zero error is obtained in the absence of fading by using a truly interpolatory f. That is, in the problem of minimizing

$$\sum_{k\neq 0} |f(k)|^2 + \alpha^2 \sum_{-\infty}^{\infty} |f'(k)|^2 \tag{33}$$

over f in $B_2(\Omega)$, $\pi < \Omega \leq 2\pi$, $f(0) = 1$, one decides to make the first sum zero and minimize the second sum under the constraints. The extremal function for this problem, then, depends only on Ω, and is found to be[†]

$$f_1(t;\ \Omega) = \frac{\sin \pi t}{\pi t}\left\{\frac{\sin(\Omega - \pi)t}{\pi t} + \left(2 - \frac{\Omega}{\pi}\right)\cos(\Omega - \pi)t\right\},$$
$$(\pi < \Omega \leq 2\pi), \quad (34)$$

the minimum value of the second sum in (33) being, under the constraints,

$$\sum_{-\infty}^{\infty} |f_1'(k;\ \Omega)|^2 = \frac{\pi^2}{3}\left(2 - \frac{\Omega}{\pi}\right)^3, \qquad (\pi < \Omega \leq 2\pi). \tag{35}$$

The Fourier transform of f_1 is

$$\begin{aligned} F_1(\omega;\ \Omega) &= 1 \text{ for } |\omega| < 2\pi - \Omega, \\ &= 1 - \frac{|\omega|}{2\pi} \text{ for } 2\pi - \Omega < |\omega| < \Omega, \\ &= 0 \text{ for } |\omega| > \Omega. \end{aligned} \tag{36}$$

The graph of $F_1(\omega;\ \Omega)$ (see Fig. 2) suggests a log cabin; so it will be called the log-cabin characteristic, and $f_1(t;\ \Omega)$ will be called the log-cabin kernel.

One may choose an optimum scaling of the log-cabin kernel as a substitute for $f_0(t;\ \Omega,\ \alpha)$ in the original problem of minimizing, over f in $B_2(\Omega)$, $\pi < \Omega \leq 2\pi$,

[†] It is no surprise that $f_1(t;\ \Omega) = \lim_{\alpha\to 0} f_0(t;\ \Omega,\ \alpha)$.

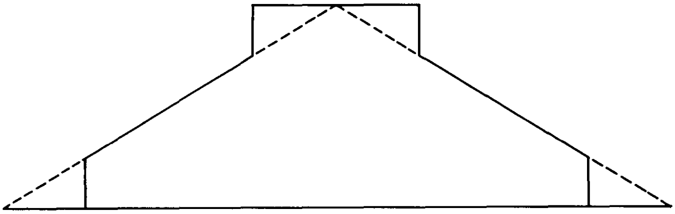


Fig. 2—The log-cabin characteristic.

$$\epsilon^2(f) = |1 - f(0)|^2 + \sum_{k\neq 0} |f(k)|^2 + \alpha^2 \sum_{-\infty}^{\infty} |f'(k)|^2. \tag{37}$$

The resulting function is

$$f_{10}(t;\, \Omega,\, \alpha) = \gamma f_1(t;\, \Omega), \tag{38}$$

where

$$\gamma = \left[1 + \frac{\alpha^2\pi^2}{3}\left(2 - \frac{\Omega}{\pi}\right)^3\right]^{-1}.$$

(In case α is small, the optimum scaling may be ignored, as it only amounts to a second-order correction.) The optimally scaled log-cabin kernel gives for the mean-square error in the original problem,

$$\epsilon_{10}^2(\Omega;\, \alpha) = \frac{\dfrac{\alpha^2\pi^2}{3}\left(2 - \dfrac{\Omega}{\pi}\right)^3}{1 + \dfrac{\alpha^2\pi^2}{3}\left(2 - \dfrac{\Omega}{\pi}\right)^3}, \qquad \pi < \Omega \leqslant 2\pi. \tag{39}$$

In Fig. 3, 10 $\log_{10}\{\epsilon_0^2(\Omega;\, \alpha)\}$ and 10 $\log_{10}\{\epsilon_{10}^2(\Omega;\, \alpha)\}$ are plotted, for various values of α^2, versus the normalized angular frequency, Ω/π, which corresponds to $2WT$ for a top frequency W and a sampling interval T. To evaluate the effectiveness of $f_0(t;\, \Omega,\, \alpha)$, some reference value of mean-square error must be adopted. It seems natural that $\epsilon_0^2(\Omega;\, \alpha)$ should be compared with $\epsilon_{10}^2(\pi;\, \alpha)$, the minimum mean-square error obtainable (at Nyquist rate) with an optimal scaling of $(\sin \pi t)/\pi t$. Thus, for $\alpha^2 = 1$, a 50-percent increase in bandwidth gives an improvement of approximately 6.3 dB. For small values of α^2, the improvement in decibels is approximately $-30 \log_{10}[2 - (\Omega/\pi)]$, which for $\Omega = 1.5\pi$ is approximately 9 dB; i.e., the mean-square error is reduced by a factor of 8 for a 50-percent increase in bandwidth. The difference between $\epsilon_0^2(\Omega;\, \alpha)$ and $\epsilon_{10}^2(\Omega;\, \alpha)$ is insignificant for moderate to small values of α^2. Of course, the extremely small errors given for Ω near 2π have to be discounted in practice as being purely theoretical values.

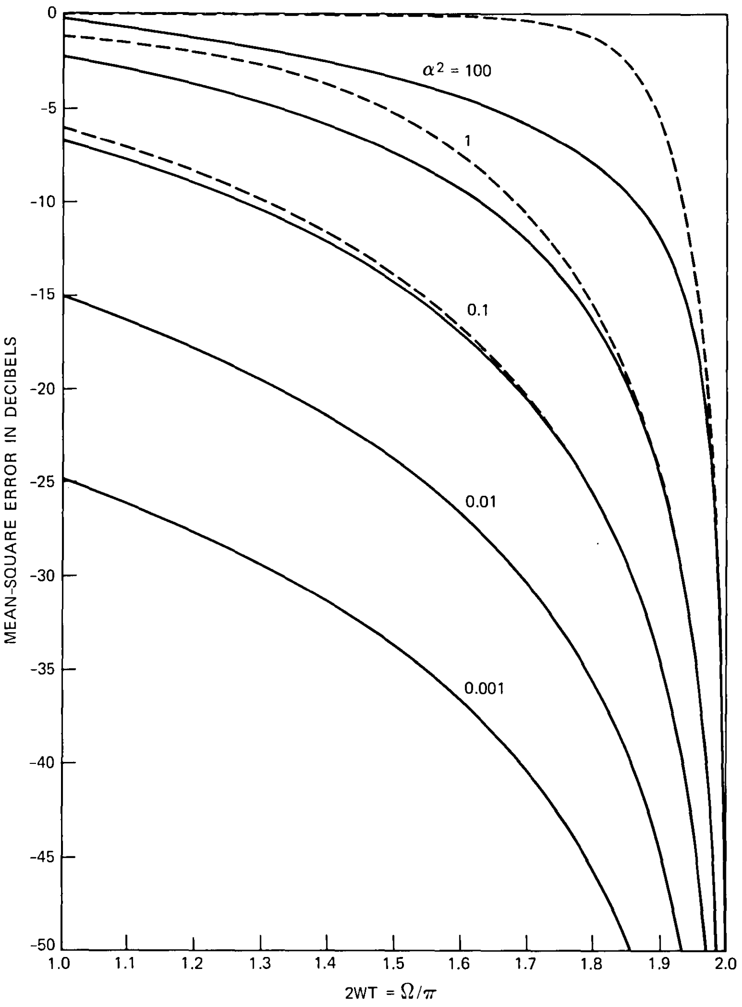


Fig. 3—Bandwidth-error exchange for various values of α^2. The solid line is for the optimal kernel; the dotted line is for the log-cabin kernel.

Whether or not the simple fading channel model is wholly acceptable, there is another reason for recommending or rationalizing the use of $f_0(t;\ \Omega,\ \alpha)$, or the log-cabin kernel; this reason being a reduced sensitivity to sampling jitter. We outline the justification of this reason.

In the problem of sampling jitter, it is supposed that

$$s(t) = \sum_{-\infty}^{\infty} a_k f(t - k), \tag{40}$$

where f is a real-valued function in $B_2(\Omega)$ and the a_k are independent random variables of zero mean and unit variance. In the absence of sampling jitter, f is taken, in case $\Omega = \pi$, to be $(\sin \pi t)/\pi t$, so that $s(t)$ is sampled at time(s) $t = n$ to obtain $a_n = s(n)$. In the presence of sampling jitter, $s(t)$ is inadvertently sampled at time(s) $t = n + \tau$, where τ is assumed to be a random variable with density $p(\tau)$, usually symmetric about $\tau = 0$. This jitter results in an error

$$\epsilon_n = a_n\{1 - f(\tau)\} - \sum_{k \neq n} a_k f(\tau + n - k). \tag{41}$$

The expected mean-square error over the $\{a_k\}$ is

$$\begin{aligned} \epsilon^2(\tau) &= \{1 - f(\tau)\}^2 + \sum_{k \neq 0} f^2(k + \tau) \\ &= 1 - 2f(\tau) + \sum_{-\infty}^{\infty} f^2(k + \tau), \end{aligned} \tag{42}$$

which, when averaged over τ, gives

$$\epsilon^2 = 1 - 2\int_{-\infty}^{\infty} f(\tau)p(\tau)d\tau + \int_{-\infty}^{\infty} \left\{ \sum_{-\infty}^{\infty} f^2(k + \tau) \right\} p(\tau)d\tau. \tag{43}$$

In case $0 < \Omega \leqslant \pi$, the sum in the last integral is $\int_{-\infty}^{\infty} f^2(t)dt$, independent of τ. In this case,

$$\epsilon^2 = 1 - 2\int_{-\infty}^{\infty} f(\tau)p(\tau)d\tau + \int_{-\infty}^{\infty} f^2(t)dt. \tag{44}$$

It is easy to show that ϵ^2 in (44), which is the same as ϵ^2 in (43) only for $0 < \Omega \leqslant \pi$, is minimized over f in $B_2(\Omega)$ by taking

$$f(t) = \int_{-\infty}^{\infty} p(\tau) \frac{\sin \Omega(t - \tau)}{\pi(t - \tau)} d\tau. \tag{45}$$

That is, $f(t)$ is obtained by bandlimiting $p(t)$ [projecting $p(t)$ on $B_2(\Omega)$].

Recall (26), where

$$f_0(t;\, \Omega,\, \alpha) = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} \frac{\cos \omega t}{1 + \alpha^2\omega^2} d\omega, \qquad 0 < \Omega \leqslant \pi.$$

From this, it is seen that

$$f_0(t;\, \Omega,\, \alpha) = \int_{-\infty}^{\infty} \frac{e^{-|\tau|/\alpha}}{2\alpha} \frac{\sin \Omega(t - \tau)}{\pi(t - \tau)} d\tau, \qquad 0 < \Omega \leqslant \pi. \tag{46}$$

Thus, for $0 < \Omega \leqslant \pi$, $f_0(t;\, \Omega,\, \alpha)$ may be interpreted as the function in $B_2(\Omega)$ which minimizes the mean-square error due to (reduced bandwidth *and*) a sampling jitter τ with density $(2\alpha)^{-1}e^{-|\tau|/\alpha}$, $-\infty < \tau < \infty$.

The fact that τ is unbounded requires the assumption of an ensemble of sampling mechanisms. However, this is not important in the case of small α, when all that really matters is the second moment of $p(\tau)$.

The mean-square error due to sampling jitter can be decreased at the expense of extra bandwidth. For $\Omega > \pi$, the sum $\sum_{-\infty}^{\infty} f^2(k + \tau)$ is no longer, because of aliasing, the same as $\int_{-\infty}^{\infty} f^2(t)dt$. The case $\pi < \Omega \leq 2\pi$ can be treated, as in the proof of the Theorem, by decomposing f as

$$f(t) = g(t) + h(t),$$

where g is the low-frequency component, bandlimited to $[-(2\pi - \Omega), (2\pi - \Omega)]$, and h is the bandpass component with center frequency π and upper frequency Ω. It turns out that the low-frequency component of the optimal f is obtained by bandlimiting the density $p(\tau)$. So in case of the two-sided exponential density, $g(t) = f_0(t;\, 2\pi - \Omega, \alpha)$ is optimal, as in $f_0(t;\, \Omega, \alpha)$. The optimal bandpass component $h(t)$ is not quite the same as $h_0(t;\, \Omega)$ in (28), but is very close to $h_0(t;\, \Omega)$ in case the density is symmetric with small support $[-a, a]$. In fact, as $a \to 0$, the optimal f in $B_2(\Omega)$, $\pi < \Omega \leq 2\pi$, for the (symmetric) sampling-jitter problem tends to $f_1(t;\, \Omega)$, the log-cabin kernel. So the log-cabin kernel is near optimal for reducing, at the expense of bandwidth, the mean-square error due to small symmetrically distributed sampling jitter.

An obvious generalization of the problem considered here is the minimization over f in $B_2(\Omega)$ of

$$\epsilon^2(f) = |1 - f(0)|^2 + \sum_{k \neq 0} |f(k)|^2 + \sum_{n=1}^{N} \alpha_n^2 \left\{ \sum_{k=-\infty}^{\infty} |f^{(n)}(k)|^2 \right\}. \qquad (47)$$

In response to a question raised by the Referee, a numerical approach to this problem is not recommended, at least for moderate N, because, first, the analytical solution is tractable, simpler, and exact; and second, certain anomalies can be anticipated.

For $0 < \Omega \leq \pi$, the solution is almost trivial, the extremal function being

$$f_0(t;\, \Omega, \bar{\alpha}_N) = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} \frac{\cos \omega t}{P(\omega^2;\, \bar{\alpha}_N)}\, d\omega, \qquad (0 < \Omega \leq \pi), \qquad (48)$$

where

$$P(\omega^2;\, \bar{\alpha}_N) = 1 + \sum_{n=1}^{N} \alpha_n^2 \omega^{2n}.$$

Always (for any Ω), the minimum mean-square error is

$$\epsilon_0^2(\Omega;\, \bar{\alpha}_N) = 1 - f_0(0;\, \Omega, \bar{\alpha}_N). \qquad (49)$$

Clearly, $\epsilon^2(f)$ can be made zero for $\Omega \geqslant (N+1)\pi$ by taking $f(t) = (\sin \pi t)^{N+1}/(\pi t)^{N+1}$. For $\pi < \Omega < (N+1)\pi$, aliasing complicates the problem. However, the aliasing can be handled in a systematic way, and the solution, if tedious, is straightforward, except in anomalous cases where only even-order derivatives appear in the last sum in (47). In the anomalous cases, the minimum, or properly speaking, the infimum of $\epsilon^2(f)$ over f in $B_2(\Omega)$ is not attainable for $\pi < \Omega < (N+1)\pi$. For example,

$$\epsilon^2(f) = |1 - f(0)|^2 + \sum_{k\neq 0} |f(k)|^2 + \alpha_2^2 \sum_{-\infty}^{\infty} |f''(k)|^2$$

can be made arbitrarily small for f in $B_2(2\pi)$, but only at the expense of making the norm (in L_2) of f very large.

There will be near-anomalous cases where the infimum *is* attained but for an f of large norm, i.e., cases where the solution is not satisfactory for reasons not taken into account. One consideration is the average power of the transmitted signal, which is directly proportional to $\int_{-\infty}^{\infty} f^2(t)dt$. Another consideration is the sensitivity to sampling jitter.

It is obvious, then, in the general problem, that $\epsilon^2(f)$ *should* be minimized with a constraint on the norm of f. This severely complicates the problem. The practical alternative is to solve the problem and then appraise the solution. What is needed is a better understanding of how the alphas affect the norm of the solution for $\Omega > \pi$. It turns out for the simple problem considered here that the norm of $f_0(t; \Omega, \alpha)$ decreases with Ω for $\pi < \Omega < 2\pi$. It would be nice to have, at least, sufficient conditions on the alphas for this decrease to obtain in the general problem. On the basis of the solution (48) to the general problem for $\Omega = \pi$, and the analogous sampling-jitter problem, it is conjectured that a sufficient condition on the alphas is that the reciprocal of the polynomial $P(\omega^2; \overline{\alpha}_N)$ in (48) be the Fourier transform of a positive function.

IV. PROOF OF THE THEOREM

The minimization problem is obviously equivalent to the problem of minimizing over f in $B_2(\Omega)$,

$$Q(f; \alpha) = \sum_{-\infty}^{\infty} |f(k)|^2 + \alpha^2 \sum_{-\infty}^{\infty} |f'(k)|^2, \tag{50}$$

subject to $f(0) = 1$.

Now suppose $f = f_1 + if_2$, where f_1 and f_2 are real-valued functions in $B_2(\Omega)$ with $f_1(0) = 1$, $f_2(0) = 0$. Then, obviously,

$$Q(f; \alpha) = Q(f_1; \alpha) + Q(f_2, \alpha). \tag{51}$$

So we may restrict our attention to real-valued f in $B_2(\Omega)$. Next suppose that $f = f_1 + f_2$, where f_1 and f_2 are real-valued functions in $B_2(\Omega)$, f_1 being even, and f_2 odd. Once again we obtain (51), since f_1 and f_2 (f_1' and f_2') are orthogonal over the integers. So we may restrict our attention to real-valued even functions f in $B_2(\Omega)$, $f(0) = 1$, in seeking the minimum.

Now define

$$Q^*(f;\, \alpha) = \int_{-\infty}^{\infty} |\, f(t)\,|^2 dt + \alpha^2 \int_{-\infty}^{\infty} |\, f'(t)\,|^2 dt. \tag{52}$$

Then we have

Lemma 1. For f in $B_2(\Omega)$, $0 < \Omega \leqslant \pi$, we have

$$Q(f;\, \alpha) = Q^*(f;\, \alpha). \tag{53}$$

This follows from the well-known result,

$$\tau \sum_{-\infty}^{\infty} |\, f(k\tau + \theta)\,|^2 = \int_{-\infty}^{\infty} |\, f(t)\,|^2 dt,$$

$$f \text{ in } B_2(\Omega), \quad \theta \text{ real}, \quad 0 < \tau \leqslant \pi/\Omega, \tag{54}$$

and the fact that if f belongs to $B_2(\Omega)$ then so does f'. □

Next define

$$\rho(\Omega;\, \alpha) = \inf_{\substack{f \in B_2(\Omega) \\ f(0)=1}} \{Q^*(f;\, \alpha)\}, \qquad 0 < \Omega < \infty,$$

$$0 \leqslant \alpha < \infty. \tag{55}$$

Then it is a simple matter to establish:

Lemma 2. We have

$$\rho(\Omega;\, \alpha) = \frac{\pi}{\Omega} \cdot \frac{(\alpha\Omega)}{\arctan(\alpha\Omega)}, \tag{56}$$

and the infimum in (55) is attained for, and only for,

$$f(t) = f(t;\, \Omega,\, \alpha) = \rho(\Omega;\, \alpha) \cdot \frac{1}{2\pi} \int_{-\Omega}^{\Omega} \frac{\cos \omega t}{1 + \alpha^2\omega^2}\, d\omega. \tag{57}$$

It follows from Lemmas 1 and 2, and the definition (14) of $\mu(\Omega;\, \alpha)$, that

$$\mu(\Omega;\, \alpha) = \rho(\Omega;\, \alpha) - 1, \qquad 0 < \Omega \leqslant \pi. \tag{58}$$

So Lemmas 1 and 2 establish the first part of the theorem.

Proof of Lemma 2. From the previous argument [which applies as well to the integrals in (52)], the extremal f will be real and even; i.e.,

$$f(t) = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} F(\omega) \cos \omega t\, d\omega, \qquad f(0) = 1, \tag{59}$$

where $F(\omega)$ is real and even. Then

$$Q^*(f;\,\alpha) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} F^2(\omega)(1+\alpha^2\omega^2)d\omega. \tag{60}$$

Now suppose g is any real even function in $B_2(\Omega)$ satisfying $g(0) = 0$; i.e.,

$$g(t) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} G(\omega)\cos\,\omega t\;dt, \qquad \int_{-\Omega}^{\Omega} G(\omega)d\omega = 0. \tag{61}$$

Then

$$\begin{aligned} Q^*(f+g;\,\alpha) &= \frac{1}{2\pi}\int_{-\Omega}^{\Omega} [f(\omega)+G(\omega)]^2(1+\alpha^2\omega^2)d\omega \\ &= Q^*(f;\,\alpha) + Q^*(g;\,\alpha) \\ &\quad + \frac{1}{\pi}\int_{-\Omega}^{\Omega} F(\omega)G(\omega)(1+\alpha^2\omega^2)d\omega. \end{aligned} \tag{62}$$

The last integral vanishes, in accord with (61), if

$$F(\omega) = \frac{c}{1+\alpha^2\omega^2}. \tag{63}$$

Then setting $f + g = f_1$, we have f_1 in $B_2(\Omega)$, $f_1(0) = 1$, and

$$Q(f_1) = Q(f;\,\alpha) + Q(f_1 - f;\,\alpha) \geqslant Q(f;\,\alpha) \tag{64}$$

with equality throughout if, and only if, $f_1 - f \equiv 0$.

So the restriction to $[-\Omega, \Omega]$ of $F(\omega)$ given by (63) is the Fourier transform of the extremal function, provided

$$\frac{c}{\pi}\int_{0}^{\Omega} \frac{d\omega}{1+\alpha^2\omega^2} = 1, \tag{65}$$

requiring

$$c = \frac{\pi}{\Omega}\cdot\frac{(\alpha\Omega)}{\arctan(\alpha\Omega)}. \tag{66}$$

Then

$$\begin{aligned} \rho(\Omega;\,\alpha) &= \frac{1}{2\pi}\int_{-\Omega}^{\Omega} F^2(\omega)(1+\alpha^2\omega^2)d\omega \\ &= \frac{c^2}{2\pi}\int_{-\Omega}^{\Omega} \frac{d\omega}{1+\alpha^2\omega^2} = c. \qquad \square \end{aligned} \tag{67}$$

Now to find the extremal f in $B_2(\Omega)$, $\pi < \Omega \leqslant 2\pi$, we first introduce $B_2(a,\,b)$, the class of square-integrable bandlimited functions whose Fourier transforms vanish outside the intervals $[-b,\,-a]$, $[a,\,b]$, where

$0 \leqslant a < b < \infty$. (In case $a > 0$, the functions are called bandpass functions.) For f in $B_2(\Omega)$, $\pi < \Omega \leqslant 2\pi$, we may make the decomposition,

$$f(t) = g(t) + h(t), \qquad g \text{ in } B_2(\beta\pi), \qquad h \text{ in } B_2(\beta\pi, \Omega), \tag{68}$$

where $\beta = 2 - (\Omega/\pi)$, $0 \leqslant \beta < 1$. Next we use the general representation for h in $B_2(\beta\pi, \Omega)$,

$$h(t) = p(t)\cos \pi t + q(t) \sin \pi t, \qquad p, q \text{ in } B_2(\beta'\pi), \tag{69}$$

where $\beta' = 1 - \beta > 0$.

In $Q(f;\, \alpha)$, we need the values of $f(k)$ and $f'(k)$, which with the decomposition (68) become

$$f(k) = g(k) + (-1)^k p(k), \tag{70}$$

$$f'(k) = g'(k) + (-1)^k[p'(k) + \pi q(k)]. \tag{71}$$

Now it is convenient at this point to use the fact that the extremal f will be real, allowing us to write

$$\begin{aligned} Q(f;\, \alpha) = {} & \sum_{-\infty}^{\infty} [g(k) + (-1)^k p(k)]^2 \\ & + \alpha^2 \sum_{-\infty}^{\infty} \{g'(k) + (-1)^k[p'(k) + \pi q(k)]\}^2 \\ = {} & \sum_{-\infty}^{\infty} g^2(k) + \alpha^2 \sum_{-\infty}^{\infty} [g'(k)]^2 + \sum_{-\infty}^{\infty} p^2(k) \\ & + \alpha^2 \sum_{-\infty}^{\infty} [p'(k) + \pi q(k)]^2 + 2 \sum_{-\infty}^{\infty} (-1)^k g(k)p(k) \\ & + 2\alpha^2 \sum_{-\infty}^{\infty} (-1)^k g'(k)[p'(k) + \pi q(k)]. \end{aligned} \tag{72}$$

Now recall that g and g' belong to $B_2(\beta\pi)$; p and $[p' + \pi q]$ belong to $B_2(\beta'\pi)$, where $0 < \beta \leqslant 1$ and $\beta' = 1 - \beta$. Therefore, gp and $g'[p' + \pi q]$ belong to $B_1(\pi)$, the class of absolutely integrable (L_1) functions whose Fourier transforms vanish outside $[-\pi, \pi]$. Hence, their Fourier transforms vanish at the endpoints, $\pm\pi$, since the Fourier transform of a function of L_1 is continuous. If follows, for example, from the Poisson sum formula, that the last two sums in (72) vanish, and the other sums may be written as integrals, in accord with (54). Thus

$$\begin{aligned} Q(f;\, \alpha) = {} & \int_{-\infty}^{\infty} g^2(t)dt + \alpha^2 \int_{-\infty}^{\infty} [g'(t)]^2 dt \\ & + \int_{-\infty}^{\infty} p^2(t)dt + \alpha^2 \int_{-\infty}^{\infty} [p'(t) + \pi q(t)]^2 dt, \end{aligned} \tag{73}$$

g in $B_2(\beta\pi)$, p, q in $B_2(\beta'\pi)$.

We wish to minimize this quantity subject to

$$f(0) = p(0) + g(0) = 1.$$

First we make the last integral vanish by taking

$$q(t) = -\frac{1}{\pi} p'(t), \tag{74}$$

noting that this is not necessary in case $\alpha = 0$. Then

$$Q(f;\, \alpha) = Q^*(g;\, \alpha) + \int_{-\infty}^{\infty} p^2(t)dt, \quad g \text{ in } B_2(\beta\pi), \quad p \text{ in } B_2(\beta'\pi). \tag{75}$$

Now suppose

$$p(0) = \lambda \quad \text{and} \quad g(0) = 1 - \lambda. \tag{76}$$

Since

$$p(0) = \int_{-\infty}^{\infty} p(t) \frac{\sin \beta'\pi t}{\pi t}\, dt, \tag{77}$$

we have, from Schwarz's inequality,

$$\int_{-\infty}^{\infty} p^2(t)dt \geqslant p^2(0)/\beta', \tag{78}$$

with equality holding if, and only if,

$$p(t) = p(0)\frac{\sin \beta'\pi t}{\beta'\pi t}, \qquad (\beta'\pi = \Omega - \pi). \tag{79}$$

From Lemma 2, we have

$$Q^*(g;\, \alpha) \geqslant g^2(0)\cdot \rho(\beta\pi;\, \alpha), \tag{80}$$

with equality holding if, and only if,

$$g(t) = g(0)\cdot f(t;\, \beta\pi,\, \alpha). \tag{81}$$

Therefore

$$Q(f;\, \alpha) \geqslant \min_{\lambda} \{(1 - \lambda)^2 \rho(\beta\pi;\, \alpha) + \lambda^2/\beta'\} \tag{82}$$

where $\beta' = 1 - \beta > 0$.

The minimum occurs for

$$\lambda = \frac{1 - \beta}{(1 - \beta) + [\rho(\beta\pi;\, \alpha)]^{-1}}, \tag{83}$$

giving

$$Q(f;\ \alpha) \geqslant \{[\rho(\beta\pi;\ \alpha)]^{-1} + 1 - \beta\}^{-1} = m(\Omega,\ \alpha), \quad \pi < \Omega \leqslant 2\pi, \tag{84}$$

with equality holding (for $\alpha > 0$) if, and only if,

$$f(t) = (1 - \lambda)f(t;\ \beta\pi,\ \alpha) + \lambda\left\{\phi(t)\cos\pi t - \frac{1}{\pi}\phi'(t)\sin\pi t\right\}, \tag{85}$$

where $f(t;\ \Omega,\ \alpha)$ is defined in (57), λ is given by (83), and

$$\phi(t) = \frac{\sin(\Omega - \pi)t}{(\Omega - \pi)t}. \tag{86}$$

The uniqueness qualification, "for $\alpha > 0$", owes to the fact [cf. (74)] that we need not have $q(t) = -p'(t)/\pi$, in case $\alpha = 0$, in order to minimize $Q(f;\ \alpha)$ with $f(0) = 1$. Obviously, any function in $B_2(\Omega)$ satisfying $f(0) = 1$ and having $(\sin\pi t)/\pi t$ as a factor will minimize $Q(f;\ 0)$. Since equality may attain in (84) for f given by (85), we have

$$\mu(\Omega;\ \alpha) = m(\Omega;\ \alpha) - 1, \qquad \pi < \Omega \leqslant 2\pi, \tag{87}$$

and the theorem is proved. □

V. ACKNOWLEDGMENTS

I would like to thank the Referee and other reviewers of this paper for their comments and advice. Also, I am grateful to Mary Ann Gatto for furnishing the calculations and the figures.

REFERENCES

1. L. J. Greenstein and B. A. Czejak, "A Polynomial Model for Multipath Fading Channel Responses," B.S.T.J., *59*, No. 7 (September 1980), pp. 1197–225.
2. L. J. Greenstein, unpublished work.

AUTHOR

Benjamin F. Logan, Jr., B.S. (Electrical Engineering), 1946, Texas Technological College; M.S., 1951, Massachusetts Institute of Technology; Eng.D.Sc. (Electrical Engineering), 1965, Columbia University; AT&T Bell Laboratories, 1956—. While at MIT, Mr. Logan was a research assistant in the Research Laboratory of Electronics, investigating characteristics of high-power electrical discharge lamps. Also at MIT he engaged in analog computer development at the Dynamic Analysis and Control Laboratory. From 1955 to 1956 he worked for Hycon-Eastern, Inc., where he was concerned with the design of airborne power supplies. He joined AT&T Bell Laboratories as a member of the Visual and Acoustics Research Department, where he was concerned with the processing of speech signals. Currently, he is a member of the Mathematical Research Department. Member, Sigma Xi, Tau Beta Pi.

AT&T Technical Journal
Vol. 64, No. 7, September 1985
Printed in U.S.A.

A Large-Scale Distribution and Location Model

By J. G. KLINCEWICZ*

(Manuscript received February 25, 1985)

A large-scale, single-period, mathematical programming model of a multicommodity distribution system has been designed and implemented to analyze and to help reconfigure AT&T distribution facilities. Within this model, the number, size, and location of major stocking locations and subsidiary stocking locations are determined on the basis of various incurred costs. These costs include facility setup costs, facility closing costs, and shipping, inventorying, handling, and operating costs. The model incorporates various features that do not appear in standard facility location models, such as nonlinear economies of scale in operating cost, capacity constraints, special products that are handled by only a limited number of facilities, and establishment of subsidiary stocking locations where desirable. In this paper we describe the model, provide a mathematical programming formulation of the problem, and describe the algorithm that was developed to obtain good solutions in an efficient manner. The flexibility of the formulation and the efficiency of the solution technique make this model a unique and useful tool. It can provide insight when used to study an existing or proposed distribution system, and it has already been used in a variety of case studies.

I. INTRODUCTION

Large manufacturing and industrial concerns, such as AT&T, provide for the warehousing and distribution of finished goods. Decisions concerning the number, size, and locations of warehousing and distribution facilities greatly affect the cost of a large material logistics system.

*AT&T Bell Laboratories.

This paper describes a large-scale, single-period, mathematical programming model of a multicommodity distribution system. Various quantitative studies of distribution problems can be performed using this model. For example, an analyst can examine the trade-off between many small distribution facilities and a few large facilities, determine a consolidation strategy in areas of contracting demand, plan an expansion strategy in areas of increasing demand, and so forth.

This model currently is being used to analyze and help reconfigure AT&T distribution facilities in order to meet future material logistics requirements. The flexibility of the model has allowed it to be applied in a variety of in-depth case studies. These studies have included both national and regional studies, studies of different tiers within the distribution system, and studies involving different families of products. To provide this flexibility, components of the model are described in generic terms with a minimum of restrictive assumptions.

Within the model, the number, size, and location of major stocking facilities, called Distribution Centers (DCs), and subsidiary stocking facilities, called Local Distribution Centers (LDCs), are determined on the basis of various incurred costs, including facility setup costs, facility closing costs, and shipping, inventorying, handling, and operating costs. The model is quite complex and combines several features, or combinations of features, that do not appear in standard facility location models. Among these features are nonlinear economies of scale in operating costs, capacity constraints, special products that are handled by only a limited number of facilities, and establishment of subsidiary stocking locations where desirable. These features are discussed in Section II.

The generic distribution system that is considered is illustrated in Fig. 1. We describe this system briefly here; it is described in greater detail in Section II. Within the model, products are assumed to move from various vendors (and repair shops) to the DCs. These DCs, in turn, distribute these products to demand area locations. In certain instances, a group of demand areas can also be served by an LDC. Different products move according to different patterns among vendors, DCs, LDCs and demand areas; these different product "types" are described in particular in Section 2.2. The mathematical programming formulation of this problem is quite large; for example, a problem with 50 possible major distribution center locations, 20 subsidiary stocking locations, 100 demand areas, and 1 special product would involve 13,170 variables and 110,531 constraints. A sample problem of this size was solved in 115.6 CPU seconds on an Amdahl 470/V8 computer.

An extensive literature exists on operations research techniques for discrete facility location problems. A recent survey of facility location

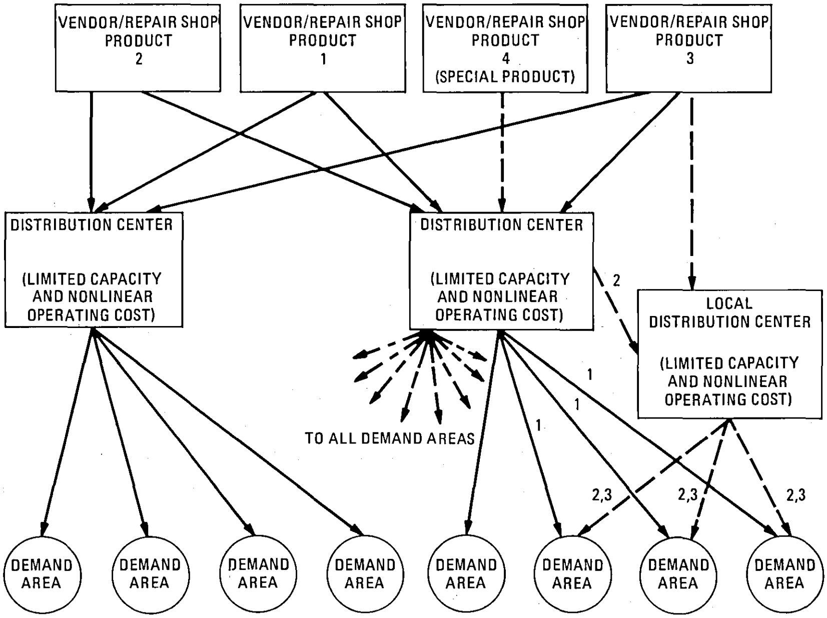

VENDOR/REPAIR SHOP PRODUCT 2
VENDOR/REPAIR SHOP PRODUCT 1
VENDOR/REPAIR SHOP PRODUCT 4 (SPECIAL PRODUCT)
VENDOR/REPAIR SHOP PRODUCT 3
DISTRIBUTION CENTER (LIMITED CAPACITY AND NONLINEAR OPERATING COST)
DISTRIBUTION CENTER (LIMITED CAPACITY AND NONLINEAR OPERATING COST)
2
LOCAL DISTRIBUTION CENTER (LIMITED CAPACITY AND NONLINEAR OPERATING COST)
1
1
1
TO ALL DEMAND AREAS
2,3
2,3
2,3
DEMAND AREA
DEMAND AREA
DEMAND AREA
DEMAND AREA
DEMAND AREA
DEMAND AREA
DEMAND AREA
DEMAND AREA

Fig. 1—The advanced distribution and location model.

work is given in Francis, McGinnis, and White.[1] We will here only mention briefly some representative problems and papers.

In the *p*-median problem, *p* facilities are chosen so as to minimize the sum of the distances from facilities to customers. A Lagrangian relaxation technique for a capacitated *p*-median problem is incorporated into the algorithm described in this paper (see Section 3.4). Surveys of work on *p*-median and related network location subjects can be found in Refs. 2 and 3.

A classic model is the Uncapacitated Facility Location Problem (UFLP), in which there exists a trade-off between setup costs of facilities and the costs of shipping products to demand areas.[4-7] Complete surveys of work on UFLP can be found in Refs. 8 and 9.

Some facility location problems can be modeled as generalized assignment problems (see Ref. 10). Here, customers with demands must be allocated among given facilities with limited capacity. Each customer must be served by only one facility. Lagrangian relaxation techniques for this problem are discussed in Fisher[11] and Ross and Soland.[10] A related procedure is incorporated into the algorithm described here in Section 3.2.

The Capacitated Facility Location Problem (CFLP)—in which a customer's demand can be split among several facilities—is considered, for example, in Refs. 12, 13, and 14. Some authors have considered the case in which there are concave operating costs associated with each facility. See, for example, Refs. 15, 16, and 17. In particular, the algorithm of Kelly and Khumawala,[15] which involves solving a sequence of linear transportation problems, is adapted here for use in the major optimization routine (see Section 3.3).

A Benders decomposition approach for a multicommodity problem was developed by Geoffrion and Graves.[18] Discussion of dynamic facility location can be found in Erlenkotter[19] and in a section of the capacity expansion survey by Luss.[20]

In Section II, we provide a complete description of the material logistics system being modeled and the mathematical programming formulation of the problem. Section III describes the solution technique that was implemented. The first subsection provides a general overview of the various stages of the algorithm. The subsequent subsections describe each stage in turn. Some brief discussion of implementation details and some concluding remarks are then given in Section IV.

II. DESCRIPTION OF THE PROBLEM

2.1 Types of locations and facilities

Five major types of facilities or physical locations can be identified in the material logistics system. Several products $k = 1, \cdots, K$ move among these facilities and locations.

Vendors are those locations, such as factories and manufacturing locations, from which products first enter the material logistics system. Each product k has its own set of fixed and known vendor locations $n = 1, \cdots, N_k$.

Repair shops are similar to vendors. Each product has its own set of fixed and known repair shop locations $m = 1, \cdots, M_k$. For each product, some fraction of demand ρ_k (possibly zero) is to be satisfied by repaired items.

Demand areas are geographical areas to which products are ultimately destined. These demand areas $j = 1, \cdots, J$, and the amount of demand D_{jk} for each product k at each demand area j, are assumed to be fixed and known.

Distribution centers are major intermediate stocking locations. The DCs must be chosen from among a set of locations $i = 1, \cdots, I$, which can be either "potential" or "existing." For potential locations, the model specifies a minimum capacity size B_i^1, a feasible capacity increment b_i^1, and a maximum capacity size $\bar{B}_i^1$. For existing facilities with fixed capacity, we assume $B_i^1 = \bar{B}_i^1$ and $b_i^1 = 0$. DC i can serve demand areas within a radius of δ_i^1 miles.

Local distribution centers are subsidiary stocking locations that handle certain types of products for several nearby demand areas. Depending upon the product type, LDCs receive products from either a DC or directly from vendors and repair shops; they then ship the products to demand areas. LDC locations must be chosen from among a set of existing and potential locations $\ell = 1, \cdots, L$. Potential facilities have minimum capacity B^2_ℓ, maximum capacity $\bar{B}^2_\ell$, and capacity increment b^2_ℓ. Existing facilities have fixed capacity $B^2_\ell = \bar{B}^2_\ell$ and $b^2_\ell = 0$. An LDC ℓ can serve demand areas within a radius of δ^2_ℓ miles.

Ordinarily, each DC deals only with certain demand areas that are "assigned" to it. The DC is (perhaps) associated with some LDCs. Each demand area that is assigned to a given DC obtains products only from that DC, and from at most one of the LDCs associated with it. Some products are exceptions to this rule; they are discussed in Section 2.2.

2.2 Types of products

Many different products move through this material logistics system. For modeling purposes, each "product" may represent an aggregation of several products. We distinguish four categories or types of products, according to how they are handled within the system. These types are described below and are pictured in Fig. 1. In Fig. 1, the three demand areas in the lower right corner are assigned to an LDC as well as a DC, whereas the other demand areas are assigned only to a DC.

Type 1 products can be handled only by a DC. These products are shipped to the DC from the vendor and repair shop. (We assume that the particular vendor and repair shop that supply a given DC are chosen so as to minimize shipping cost.) From the DC, they are shipped to all demand areas that are assigned to that DC.

Type 2 products can be delivered to demand areas either from a DC or else from an LDC. If a demand area is assigned to an LDC, the DC ships Type 2 products to the LDC for delivery to the demand area. This is indicated by the dashed lines in Fig. 1 that are labeled with the numeral 2.

Type 3 products can also be delivered to demand areas either from a DC or else from an LDC. The LDC receives shipment of Type 3 products directly from the vendor and repair shop. We assume that the LDC receives shipments from the same repair shops and vendors that serve the DC with which the LDC is associated. This is indicated by the dashed lines in Fig. 1 that are labeled with the numeral 3.

All Type 1, Type 2 and Type 3 products are handled with the arrangement whereby each demand area deals with only one DC and at most one LDC. (If a demand area is assigned to an LDC, it obtains all Type 2 and Type 3 products via that LDC.) However, Type 4 products ("special" products) are an exception to the rule. Only a small number p_k of DCs are chosen to handle a given Type 4 product k. These DCs then serve all demand areas. This is indicated by the alternate short-and-long dashed lines in Fig. 1.

Each product's type is given as input by the user. We let T_1, T_2, T_3, and T_4 denote the sets of indices of Type 1, Type 2, Type 3, and Type 4 products, respectively.

Each unit of demand that is handled by a DC or an LDC occupies some average amount of warehouse space (measured in units of capacity). The amount of space occupied depends upon, among other things, the size of the product, turnover time of inventory, and the type of facility (DC or LDC). We define the following parameters:

s_k^1 = warehouse space required per unit demand at a DC if a DC ships product k to a demand area,
s_k^2 = warehouse space required per unit demand at an LDC,
s_k^3 = warehouse space required per unit demand at a DC if a DC ships product k to an LDC (relevant for Type 2 products).

2.3 Costs

2.3.1 Facility setup costs

Facility setup costs are costs incurred when a facility (DC or LDC) is chosen to be open. The setup cost depends upon the size of the facility. We assume that, for DC i, opened with a capacity of x units, the setup cost is of the form $\alpha_i^1 + \beta_i^1 x$, where α_i^1 and β_i^1 are given constants that depend upon the particular facility. For LDC ℓ, the setup cost is of the form $\alpha_\ell^2 + \beta_\ell^2 x$, where α_ℓ^2 and β_ℓ^2 are given constants. Actual total setup costs realistically might be assumed to be amortized over several time periods. Since ours is a static, single-period model, the setup cost used in the model could be set equal to an amortized share of the total setup cost.

2.3.2 Facility closing costs

In the model, if a facility is not chosen to be open, a closing cost c_i^1 (for DC i) or c_ℓ^2 (for LDC ℓ) is incurred.

2.3.3 Shipping costs

Shipping cost parameters are given in cost per mile per unit demand. These cost parameters for product k follow:

t_k^1 = cost of shipping from vendor/repair shop to DC/LDC,
t_k^2 = cost of shipping from DC to demand area,
t_k^3 = cost of shipping from DC to LDC,
t_k^4 = the cost of shipping from LDC to demand area.

Obviously, for each type of product only certain of these parameters are applicable. (The model can also be easily modified so that shipping costs are measured per unit demand, independent of distance.)

2.3.4 Inventorying or storage cost

An inventorying or storage cost per unit demand is incurred at each DC and LDC. We define the following parameters:

η^1_{ki} = inventorying cost per unit demand at a DC i that ships product k to a demand area,
$\eta^2_{k\ell}$ = inventorying cost per unit demand at an LDC ℓ that ships product k to a demand area,
η^3_{ki} = inventorying cost per unit demand at a DC i that ships product k to an LDC (relevant for Type 2 products).

The value of η^3_{ki} can differ from η^1_{ki} because of differences in turnover time for inventory bound for an LDC and inventory bound directly for demand areas.

2.3.5 Handling cost

When a product is processed at a warehousing facility (either a DC or an LDC), there are some labor costs incurred. We define the following parameters:

h^1_{ki} = handling cost per unit demand at a DC i that ships product k to a demand area,
$h^2_{k\ell}$ = handling cost per unit demand at an LDC ℓ that ships product k to a demand area,
h^3_{ki} = handling cost per unit demand at a DC i that ships product k to an LDC ℓ (relevant for Type 2 products).

2.3.6 Operating cost

Other operating costs at each facility are represented as a continuous, nondecreasing, concave function of the space occupied in order to account for possible savings due to economies of scale. Let

σ^1_i = volume of space occupied at DC i, $i = 1, \cdots, I$, and
σ^2_ℓ = volume of space occupied at LDC ℓ, $\ell = 1, \cdots, L$.

Then, we define the cost functions:

$f^1_i(\sigma^1_i)$ = operating cost at DC i, $i = 1, \cdots, I$, and
$f^2_\ell(\sigma^2_\ell)$ = operating cost at LDC ℓ, $\ell = 1, \cdots, L$.

Within the software implementation, we assume these cost functions to be piecewise linear with a nonnegative intercept at the origin. In typical examples, each function used between three and five piecewise linear segments.

2.4 The mathematical programming model

2.4.1 Decision variables

To formulate the mathematical programming model, we first specify the necessary decision variables. All variables take on integer values. They are as follows:

$$y_i = \begin{cases} 1 \text{ if DC } i \text{ is open,} \\ & \text{for } i = 1, \cdots, I, \\ 0 \text{ otherwise,} \end{cases}$$

$$z_{i\ell} = \begin{cases} 1 \text{ if LDC } \ell \text{ is open and served by DC } i, \\ & \text{for } \ell = 1, \cdots, L \quad \text{and} \quad i = 1, \cdots, I, \\ 0 \text{ otherwise,} \end{cases}$$

$$x_{ij} = \begin{cases} 1 \text{ if demand area } j \text{ is assigned to DC } i, \\ & \text{for } i = 1, \cdots, I \quad \text{and} \quad j = 1, \cdots, J, \\ 0 \text{ otherwise,} \end{cases}$$

$$w_{\ell j} = \begin{cases} 1 \text{ if demand area } j \text{ is assigned to LDC } \ell, \\ & \text{for } \ell = 1, \cdots, L \quad \text{and} \quad j = 1, \cdots, J, \\ 0 \text{ otherwise,} \end{cases}$$

$$v_{ik} = \begin{cases} 1 \text{ if DC } i \text{ is used to serve special product } k, \\ & \text{for } i = 1, \cdots, I \quad \text{and} \quad k \in T_4, \\ 0 \text{ otherwise,} \end{cases}$$

$$u_{ijk} = \begin{cases} 1 \text{ if demand area } j \text{ receives special product } k \text{ from DC } i, \\ & \text{for } j = 1, \cdots, J \quad \text{and} \quad i = 1, \cdots, I \\ 0 \text{ otherwise,} & \text{and} \quad k \in T_4, \end{cases}$$

$$q_i^1 = \begin{array}{l} \text{number of size increments} \\ \text{above the minimum} \\ \text{opened at DC } i, \text{ for } i = 1, \cdots, I, \end{array}$$

$$q_\ell^2 = \begin{array}{l} \text{number of size increments} \\ \text{above the minimum} \\ \text{opened at LDC } \ell, \text{ for } \ell = 1, \cdots, L. \end{array}$$

2.4.2 Distance and other parameters

Shipping cost calculations require the distances between pairs of locations. For notational simplicity, we refer to all such pairwise distances by the notation d with two subscripts. The subscript i refers to a DC i, subscript ℓ to an LDC ℓ, and subscript j to a demand area j. Further, the subscript $n(i)$ refers to the vendor closest to DC i, and subscript $m(i)$ refers to the repair shop closest to DC i. Thus, $d_{n(i),i}$ is the distance to DC i from its nearest vendor, $d_{\ell j}$ is the distance from LDC ℓ to demand area j, and so forth.

A particular latitude/longitude point is associated with each facility location and each demand area and used to estimate road distances on an as-needed basis. The software allows users to provide distance data that would override the calculated distance.

To simplify the formulation, let the space (units of capacity) required for a DC to serve demand area j (excluding Type 4 products and provided it is not served by an LDC) be

$$S_j^1 = \sum_{k \in T_1 \cup T_2 \cup T_3} s_k^1 D_{jk}. \tag{1}$$

If demand area j is served by an LDC as well as a DC, then the amount of space required at the LDC is

$$S_j^2 = \sum_{k \in T_2 \cup T_3} s_k^2 D_{jk}, \tag{2}$$

and the amount of space required at the DC is

$$S_j^3 = \sum_{k \in T_1} s_k^1 D_{jk} + \sum_{k \in T_2} s_k^3 D_{jk}. \tag{3}$$

The space required at a DC to serve Type 4 product $k \in T_4$ for demand area j is

$$S_{jk}^4 = s_k^1 D_{jk}. \tag{4}$$

Likewise, it is convenient to aggregate the shipping costs, inventorying costs, and handling costs associated with assigning demand area j to DC i in an "assignment cost" A_{ij}^1, as follows:

$$\begin{aligned} A_{ij}^1 = \sum_{k \in T_1 \cup T_2 \cup T_3} & (t_k^1 d_{n(i)i}(1 - \rho_k) + t_k^1 d_{m(i)i} \rho_k \\ & + t_k^2 d_{ij} + h_{ki}^1 + \eta_{ki}^1) D_{jk}. \end{aligned} \tag{5}$$

If an LDC ℓ is involved, the assignment cost $A_{ij\ell}^2$ is expressed as

$$\begin{aligned} A_{ij\ell}^2 = & \sum_{k \in T_1} (t_k^1 d_{n(i)i}(1 - \rho_k) + t_k^1 d_{m(i)i} \rho_k + t_k^2 d_{ij} + h_{ki}^1 + \eta_{ki}^1) D_{jk} \\ & + \sum_{k \in T_2} (t_k^1 d_{n(i)i}(1 - \rho_k) + t_k^1 d_{m(i)i} \rho_k + t_k^3 d_{i\ell} + t_k^4 d_{\ell j} \\ & + h_{ki}^3 + h_{k\ell}^2 + \eta_{ki}^3 + \eta_{k\ell}^2) D_{jk} \\ & + \sum_{k \in T_3} (t_k^1 d_{n(i)\ell}(1 - \rho_k) + t_k^1 d_{m(i)\ell} \rho_k \\ & + t_k^4 d_{\ell j} + h_{k\ell}^2 + \eta_{k\ell}^2) D_{jk}. \end{aligned} \tag{6}$$

Finally, the cost A_{ijk}^4 of assigning Type 4 product k at demand area j to DC i is as follows:

$$A_{ijk}^4 = (t_k^1 d_{n(i)i}(1 - \rho_k) + t_k^1 d_{m(i)i} \rho_k + t_k^2 d_{ij} + h_{ki}^1 + \eta_{ki}^1) D_{jk}. \tag{7}$$

In the event that demand area j cannot be assigned to a DC i or an LDC ℓ because j lies outside the operating radius of the facility (i.e., $d_{ij} > \delta_i$ or $d_{\ell j} > \delta_\ell$), then the corresponding assignment costs can be set to an arbitrarily large number.

2.4.3 Formulating the model

Below, we provide the mathematical formulation for the problem. Then, the objective function and each of the constraints is explained in turn. The problem is formulated as follows:

$$\begin{aligned} \min \sum_{i=1}^{I} & (\alpha_i^1 + \beta_i^1(B_i^1 + b_i^1 q_i^1)) y_i \\ & + \sum_{\ell=1}^{L} (\alpha_\ell^2 + \beta_\ell^2(B_\ell^2 + b_\ell^2 q_\ell^2)) \sum_{i=1}^{I} z_{i\ell} \\ & + \sum_{i=1}^{I} c_i^1(1 - y_i) + \sum_{\ell=1}^{L} c_\ell^2 \left(1 - \sum_{i=1}^{I} z_{i\ell}\right) \\ & + \sum_{i=1}^{I} \sum_{j=1}^{J} A_{ij}^1 x_{ij} \left(1 - \sum_{\ell=1}^{L} w_{\ell j}\right) \\ & + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{\ell=1}^{L} A_{ij\ell}^2 x_{ij} w_{\ell j} \\ & + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k \in T_4} A_{ijk}^4 u_{ijk} \\ & + \sum_{i=1}^{I} f_i^1(\sigma_i^1) y_i \\ & + \sum_{\ell=1}^{L} f_\ell^2(\sigma_\ell^2) \sum_{i=1}^{I} z_{i\ell} \end{aligned} \tag{8a}$$

subject to

$$\sum_{i=1}^{I} x_{ij} = 1 \quad \text{for} \quad j = 1, \cdots, J, \tag{8b}$$

$$x_{ij} \leqslant y_i \quad \text{for} \quad i = 1, \cdots, I \quad \text{and} \quad j = 1, \cdots, J, \tag{8c}$$

$$\sum_{\ell=1}^{L} w_{\ell j} \leqslant 1 \quad \text{for} \quad j = 1, \cdots, J, \tag{8d}$$

$$\begin{aligned} w_{\ell j} \leqslant x_{ij} z_{i\ell} \quad \text{for} \quad & i = 1, \cdots, I \quad \text{and} \\ & \ell = 1, \cdots, L \quad \text{and} \quad j = 1, \cdots, J, \end{aligned} \tag{8e}$$

$$\sum_{i=1}^{I} z_{i\ell} \leqslant 1 \quad \text{for} \quad \ell = 1, \cdots, L, \tag{8f}$$

$$\sigma_i^1 = \sum_{j=1}^{J} S_j^1 x_{ij}\left(1 - \sum_{\ell=1}^{L} w_{\ell j}\right) + \sum_{\ell=1}^{L} \sum_{j=1}^{J} S_j^3 x_{ij} w_{\ell j} + \sum_{k \in T_4} \sum_{j=1}^{J} S_{jk}^4 u_{ijk} \quad \text{for} \quad i = 1, \cdots, I, \tag{8g}$$

$$\sigma_i^1 \leqslant B_i^1 + b_i^1 q_i^1 \quad \text{for} \quad i = 1, \cdots, I, \tag{8h}$$

$$\sigma_\ell^2 = \sum_{j=1}^{J} S_j^2 w_{\ell j} \quad \text{for} \quad \ell = 1, \cdots, L, \tag{8i}$$

$$\sigma_\ell^2 \leqslant B_\ell^2 + b_\ell^2 q_\ell^2 \quad \text{for} \quad \ell = 1, \cdots, L, \tag{8j}$$

$$B_i^1 + b_i^1 q_i^1 \leqslant \bar{B}_i^1 \quad \text{for} \quad i = 1, \cdots, I, \tag{8k}$$

$$B_\ell^2 + b_\ell^2 q_\ell^2 \leqslant \bar{B}_\ell^2 \quad \text{for} \quad \ell = 1, \cdots, L, \tag{8l}$$

$$\sum_i v_{ik} \leqslant \rho_k \quad \text{for} \quad k \in T_4, \tag{8m}$$

$$u_{ijk} \leqslant v_{ik} \quad \text{for} \quad j = 1, \cdots, J, \quad \text{and} \quad i = 1, \cdots, I \quad \text{and} \quad k \in T_4, \tag{8n}$$

$$\sum_i u_{ijk} = 1 \quad \text{for} \quad j = 1, \cdots, J \quad \text{and} \quad k \in T_4, \tag{8o}$$

$$y_i,\ z_{i\ell},\ x_{ij},\ w_{\ell j},\ v_{ik},\ u_{ijk} \in \{0, 1\} \text{ for } \begin{aligned}[t] & i = 1, \cdots, I, \\ & j = 1, \cdots, J, \\ & \ell = 1, \cdots, L, \text{ and} \\ & k \in T_4, \end{aligned} \tag{8p}$$

$$q_i^1,\ q_\ell^2 \in \{0, 1, \cdots, \infty\} \quad \text{for} \quad i = 1, \cdots, I \quad \text{and} \quad \ell = 1, \cdots, L. \tag{8q}$$

The first two summation terms in the objective function (8a) represent the setup cost incurred for open DCs and LDCs. The next two terms represent the closing costs that are incurred if a DC or an LDC is not open. In the fifth term, we include the assignment costs that are incurred if demand area j is assigned to DC i with no LDC involved. (Only in that case would $x_{ij}\ (1 - \sum_{\ell=1}^{L} w_{\ell j}) = 1$.) The sixth term, on the other hand, gives the assignment costs that are incurred if demand area j is assigned to DC i and LDC ℓ, and the seventh term considers the assignment costs for Type 4 products. The last two terms of the objective function represent the operating cost that is incurred at each open facility.

Constraints (8b) ensure that each demand area j is assigned to exactly one DC, and constraints (8c) ensure that demand areas are assigned only to DCs that are open. Constraints (8d) permit each demand area j to be assigned to at most one LDC. The condition (8e) guarantees that such an LDC assignment is made only if the LDC is open and is served by the same DC that serves the demand area j. Constraints (8f) ensure that each LDC is served by at most one DC. Constraints (8g) define σ_i^1, the space actually utilized at each DC i. This is obtained by adding the space required to serve demand areas that are assigned only to DC i, plus space required for demand areas assigned to both DC i and an LDC, plus space required to serve any Type 4 products that are assigned to DC i. In (8h), this space utilized is constrained to not exceed the capacity installed. Similarly, (8i) defines σ_ℓ^2, the space utilized at LDC ℓ, and (8j) constrains σ_ℓ^2 to not exceed the capacity installed at that LDC. Constraints (8k) and (8l) ensure that the capacity installed does not exceed the maximum permitted capacity for DCs and LDCs, respectively. In (8m), the number of DCs that serve each Type 4 product k does not exceed the permitted number p_k. Constraints (8n) guarantee that a demand area receives each Type 4 product k from a DC that handles that product. Constraints (8o) require that each demand area j be assigned to only one DC for a given Type 4 product. [In the event that no Type 4 products occur in the problem, constraints (8m) through (8o) do not appear]. Finally, conditions (8p) and (8q) enforce integer constraints on the variables.

The integer program (8) is quite large. For example, for 10 possible DC locations, 10 possible LDC locations, 50 demand areas, and 1 Type 4 product, there are 1640 variables and 6221 constraints. For a larger problem of 50 DCs, 20 LDCs, 100 demand areas, and 1 Type 4 product, there are 13,170 variables and 110,531 constraints. Nonlinearities appear in the objective function (8a) and in constraints (8e) and (8g), thus making it even more difficult to solve the program directly.

III. SOLUTION APPROACH

3.1 Overview

Because of the difficulty of the integer programming problem, the proposed algorithm contains some heuristic elements. In particular, we propose to first treat a simpler version of the problem and then adjust this solution with a series of heuristics to obtain a solution to the overall problem.

The elements of the problem that are judged to be most important are considered in the initial optimization. Other elements that are considered, by comparison, less important or elements that are expected to appear less often in actual case studies are treated by the secondary optimization. Specifically, the issues of Type 4 product distribution, LDC locations, and discrete facility sizing are set aside in the initial optimization. The initial optimization problem thereby becomes a type of capacitated facility location problem with concave costs. After obtaining a solution to this problem, the solution is modified in a step-by-step fashion to incorporate, in turn, Type 4 products, LDC locations and facility sizing. In the subsections that follow, we describe the various portions of the algorithm.

Within our software implementation, there is the option to specify that certain variables be fixed in advance, for example, that certain DCs or LDCs be fixed open or closed, that certain demand area assignments be forced or forbidden, or that certain DCs be prevented from handling Type 4 products. The necessary modifications to the algorithm are generally straightforward, and therefore not discussed explicitly here.

3.2 The preprocessor

In Section 2.4.2 we described various DC and LDC space requirement parameters and assignment cost parameters. These parameters are computed by the algorithm in a preprocessor routine. The importance of the space requirement parameters lies in the fact that, when not considering Type 4 products, the original *multicommodity* problem becomes a *single-commodity* problem. The "commodity" in this case is warehouse space; the DCs have supplies of space and the demand areas require space. Further, the assignment costs (5) and (7) provide a convenient aggregation of shipping, inventorying, and handling costs. (Because of storage requirements, however, the assignment cost for assignments that use LDCs [eq. (6)] is calculated as needed.)

3.3 The primary optimization

In the primary optimization routine, we determine a set of DCs to be opened. Initially, we assume that all demand is served by DCs alone and that the maximum possible capacity $\bar{B}_i$ is available at all open facilities. We associate a cost with each possible pairing of a DC i and a demand area j of the form

$$C_{ij} = A^1_{ij} + \beta^1_i S^1_j. \tag{9}$$

This represents the assignment cost for demand area j (as discussed in Section 2.4.2) plus a cost corresponding to the variable setup cost for the warehouse space required to serve j. We also compute a "net fixed cost"

$$F_i = \alpha^1_i - c^1_i, \tag{10}$$

which is equal to the difference between the fixed setup cost and the closing cost.

At first, we also set aside the requirement that each demand area be served by only one DC. After first obtaining a solution without this restriction, we then will adjust the solution to enforce the restriction.

The initial problem is then the following capacitated facility location problem with concave operating cost:

$$\min \sum_{i=1}^{I} F_i y_i + \sum_{i=1}^{I} f_i^1(\sigma_i^1) + \sum_{i=1}^{I} \sum_{j=1}^{J} C_{ij} x_{ij} \tag{11a}$$

subject to

$$\sum_{j=1}^{J} S_j^1 x_{ij} \leqslant \bar{B}_i^1 y_i \quad \text{for} \quad i = 1, \cdots, I, \tag{11b}$$

$$\sum_{i=1}^{I} x_{ij} = 1 \quad \text{for} \quad j = 1, \cdots, J, \tag{11c}$$

$$\sigma_i^1 = \sum_{j=1}^{J} S_j^1 x_{ij} \quad \text{for} \quad i = 1, \cdots, I, \tag{11d}$$

$$0 \leqslant x_{ij} \leqslant 1 \quad \text{for} \quad i = 1, \cdots, I \quad \text{and} \quad j = 1, \cdots, J, \tag{11e}$$

$$y_i \in \{0, 1\} \quad \text{for} \quad i = 1, \cdots, I. \tag{11f}$$

Several algorithms for problems of this form have been proposed.[15-17] We have implemented the iterative algorithm due to Kelly and Khumawala[15] that defines and solves a sequence of standard linear transportation problems. In these problems, the DCs are "sources" and demand areas are "sinks." Linear costs on arcs from sources to sinks are based on the values of C_{ij} and on the values of $f_i^1(\sigma_i^1)$ that are implied by the solution at the previous iteration. Key to the algorithm is the provision for a "dummy sink." Incoming arcs to this dummy sink have negative costs based on the setup costs F_i and the operating cost values $f_i^1(\sigma_i^1)$ at the previous iteration. Intuitively, this dummy sink offers to "buy back" the capacity at a DC. Costs on arcs into this sink are chosen from iteration to iteration in such a way as to discourage opening very underutilized DCs and to encourage taking advantage of the economies-of-scale in the concave operating cost. Complete details are given in Ref. 15. Within our implementation, the linear transportation problems are solved using primal network flow-convex,[21] which is a state-of-the-art simplex network flow code.

Upon solving problem (11), we obtain a set G of open DCs, each of which serves some group of demand areas. Some set U of demand areas will be served by more than one DC. Typically, in our computational experience the numbers of such demand areas that have their demand "split" in this way among DCs is small. We next seek to resolve these splits and obtain a solution in which each demand area is served by only one DC.

Within this phase, we momentarily leave fixed those demand areas that are assigned to only one DC in the solution obtained to (11). The demand areas $j \in U$ whose assignments have been split are considered to be unassigned. The fixed assignments, on the other hand, take up warehouse space at the open DCs. The remaining available warehouse space at each DC i is denoted $\tilde{B}_i$.

To take the unassigned demand areas and assign them among the open DCs, we must, if possible, (approximately) solve this problem:

$$\min \sum_{i \in G} \sum_{j \in U} C_{ij} x_{ij} \tag{12a}$$

$$\text{subject to } \sum_{i \in G} x_{ij} = 1 \quad \text{for} \quad j \in U, \tag{12b}$$

$$\sum_{j \in U} S_j^1 x_{ij} \leq \tilde{B}_i \quad \text{for} \quad i \in G, \tag{12c}$$

$$x_{ij} \in \{0, 1\} \quad \text{for} \quad j \in U \quad \text{and} \quad i \in G. \tag{12d}$$

This is of the form of a generalized assignment problem.[10,11] As described below, we first attempt to solve this using Lagrangian relaxation (see Ref. 11) and branch-and-bound techniques. Later, we add additional DCs to the set G in the event that (12) is infeasible.

If we associate nonnegative Lagrange multipliers λ_i, $i \in G$ with constraints (12c) and incorporate these constraints into the objective (12a), we obtain the relaxed problem:

$$\min \sum_{i \in G} \sum_{j \in U} C_{ij} x_{ij} + \sum_{i \in G} \lambda_i \left(\sum_{j \in U} S_j^1 x_{ij} - \tilde{B}_i \right) \tag{13a}$$

subject to

$$\sum_{i \in G} x_{ij} = 1 \quad \text{for} \quad j \in U, \tag{13b}$$

$$x_{ij} \in \{0, 1\} \quad \text{for} \quad i \in G \quad \text{and} \quad j \in U. \tag{13c}$$

The relaxed problem can be solved by inspection. For each j, let index e be chosen so that

$$C_{ej} + \lambda_e S_j^1 = \min_{i \in G} \{C_{ij} + \lambda_i S_j^1\}. \tag{14}$$

(Ties can be broken arbitrarily.) Then, for each demand area j,

$$x_{ij} = \begin{cases} 1 \text{ if } i = e \\ 0 \text{ otherwise} \end{cases} \quad \text{for } i \in G. \tag{15}$$

This is incorporated into an iterative scheme where the λ_i are increased from iteration to iteration if the solution obtained violates constraints (12c). This update is of the form

$$\lambda_i \leftarrow \lambda_i + \tau \left(\sum_{j \in U} S_j^1 x_{ij} - \tilde{B}_i \right), \tag{16}$$

where "←" indicates "is replaced by" and τ is a positive scalar. A formula for τ is

$$\tau = \frac{\gamma Z}{\left\{ \sum_{i=1}^{I} \left\{ \sum_{j=U} S_j^1 x_{ij} - \tilde{B}_i \right\}^2 \right\}^{1/2}}, \tag{17}$$

where γ is a positive scalar (in our implementation, $\gamma = 0.25$ has been used successfully) and Z is an estimate of the difference between the optimal objective function value (12a) in the original problem and the current value (13a) in the relaxed problem. Within our implementation, we obtain an estimate of the magnitude of (12a) based on arbitrarily assigning each $j \in U$ to one of the DCs to which it has a split assignment. The value of Z is then chosen to be 10 percent of this estimate of (12a).

Similar updating rules appear in Held, Wolfe, and Crowder.[22] Similar relaxation techniques for generalized assignment problems are discussed in Fisher.[11]

If the Lagrangian relaxation problem (13) does not yield a feasible solution to (12) within a reasonable number of iterations (40 iterations have been used successfully in our implementation), then a branch-and-bound technique for (12) is initiated. A description of branch-and-bound algorithms for integer programming problems can be found, for example, in Ref. 23. We outline below the major features of the algorithm that we have implemented.

At each node of the branch-and-bound tree, assignments for some of the demand areas $j \in U$ are assumed fixed. At each node, a Lagrangian relaxation problem, similar to (13), is solved for the unfixed demand areas. A lower bound on the optimal objective function value (12a), given the fixed assignments at the node, is obtained by evaluating the relaxed objective function (13a) at solution (15). Likewise, another lower bound at the node can also be calculated by assuming that each unfixed demand area is assigned to the least costly DC. The branch-and-bound routine uses the maximum of the two lower bounds. If a feasible solution is found by Lagrangian relaxation, it is an upper bound. A very simple branching rule is used in which variables are chosen for branching in order of increasing j. Nodes are chosen for branching by the least lower bound value among most recently created unfathomed nodes. If the lower bound at a given node is within, say, 10 percent of a current upper bound, then no further branching is done from that node.

If problem (12) is not feasible, an additional facility is chosen to be in the open set and then problem (13) is resolved with the DCs in the larger set G fixed open and all other DCs fixed closed. This additional new facility is chosen according to a rule that estimates the change in assignment cost that would result from its opening. That is, for each DC $i \notin G$, compute

$$\Omega_i = \sum_{j \notin U} \min_{k \in G}\{\max(C_{kj} - C_{ij}, O)\}. \tag{18}$$

Each term in this summation gives, for a demand area j, the minimum savings in assignment cost that would result if DC i were available. The DC with maximum Ω_i value is chosen to be open and is thus added to G. (This rule is similar to the "largest omega" rule proposed by Khumawala[6] as a branching rule in a branch-and-bound algorithm for plant location problems.) The Kelly and Khumawala algorithm for problem (11) is then repeated with facilities $i \in G$ open and all others forced to be closed.

In Fig. 2, we provide a flowchart of the primary optimization routine. This figure summarizes the procedures described in this subsection.

3.4 Assigning Type 4 products

As explained in Section 2.2, for each Type 4 product k, only a limited number p_k of DC facilities are chosen to handle it. These facilities can be different for each k. We assume that Type 4 products occupy, at most, only a small percentage (perhaps 10 percent or less) of the warehouse space required.

Given the demand area assignments determined by the primary optimization algorithm, each open DC $i \in G$ only has some amount $\tilde{B}_i$ of warehouse space still available. If demand area j is assigned to DC i for Type 4 product k, we associate cost

$$C^4_{ijk} = A^4_{ijk} + \beta^1_i S^4_{jk}, \tag{19}$$

which represents the assignment cost plus a share of the variable setup cost.

Given the set G, the problem of choosing p_k locations to serve product k can be formulated as the following capacitated p-median problem:

$$\min \sum_{i \in G} \sum_{j=1}^{J} C^4_{ijk} u_{ijk} \tag{20a}$$

subject to

$$\sum_{i \in G} v_{ik} \leq p_k, \tag{20b}$$

$$\sum_{i \in G} u_{ijk} = 1 \quad \text{for} \quad j = 1, \cdots, J, \tag{20c}$$

$$u_{ijk} \leq v_{ik} \quad \text{for} \quad j = 1, \cdots, J \quad \text{and} \quad i \in G, \tag{20d}$$

$$\sum_{j=1}^{J} S^4_{jk} u_{ijk} \leq \tilde{B}_i \quad \text{for} \quad i \in G, \tag{20e}$$

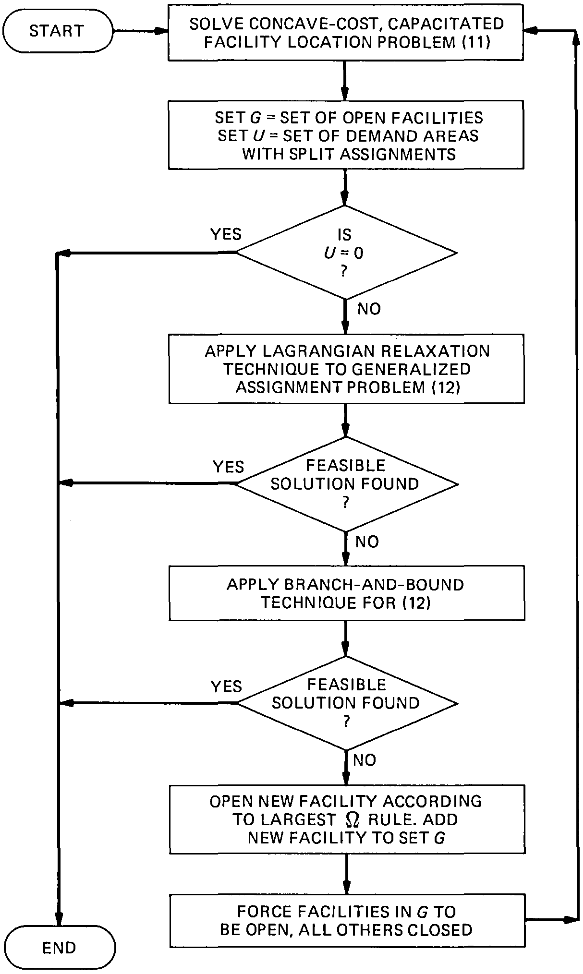

START
SOLVE CONCAVE-COST, CAPACITATED FACILITY LOCATION PROBLEM (11)
SET G = SET OF OPEN FACILITIES
SET U = SET OF DEMAND AREAS WITH SPLIT ASSIGNMENTS
IS U = 0 ?
YES
NO
APPLY LAGRANGIAN RELAXATION TECHNIQUE TO GENERALIZED ASSIGNMENT PROBLEM (12)
FEASIBLE SOLUTION FOUND ?
YES
NO
APPLY BRANCH-AND-BOUND TECHNIQUE FOR (12)
FEASIBLE SOLUTION FOUND ?
YES
NO
OPEN NEW FACILITY ACCORDING TO LARGEST Ω RULE. ADD NEW FACILITY TO SET G
FORCE FACILITIES IN G TO BE OPEN, ALL OTHERS CLOSED
END

Fig. 2—The primary optimization routine.

$$u_{ijk},\, v_{ik} \in \{0, 1\} \quad \text{for} \quad j = 1, \cdots, J \quad \text{and} \quad i \in G. \qquad \text{(20f)}$$

Without constraints (20e), this is the well-known p-median problem.[3,24,25] The additional constraints (20e) are capacity constraints that take into account the amount of warehouse space that is actually available to serve product k. Problem (20) is solved for each Type 4 product in turn. If there is more than one Type 4 product, the available capacities $\tilde{B}_i$ are updated each time (20) is solved.

To solve problem (20), we apply a Lagrangian relaxation technique. Nonnegative Lagrange multipliers μ_i are associated with constraints (20e) and unrestricted multipliers λ_j with constraints (20c). These constraints are incorporated into the objective function to produce the relaxed problem:

$$\min \sum_{i\in G} \sum_{j=1}^{J} C_{ijk}^{4} u_{ijk} + \sum_{j=1}^{J} \lambda_j \left(\sum_{i\in G} u_{ijk} - 1 \right) + \sum_{i\in G} \mu_i \left(\sum_{j=1}^{J} S_{jk}^{4} u_{ijk} - \tilde{B}_i \right) \tag{21a}$$

subject to

$$\sum_{i\in G} v_{ik} \leqslant p_k, \tag{21b}$$

$$u_{ijk} \leqslant v_{ik} \quad \text{for} \quad j = 1, \cdots, J \quad \text{and} \quad i \in G, \tag{21c}$$

$$u_{ijk},\ v_{ik} \in \{0, 1\} \quad \text{for} \quad j = 1, \cdots, J \quad \text{and} \quad i \in G. \tag{21d}$$

The relaxed problem can be solved by inspection. For each DC i, compute

$$R_i = \sum_{j=1}^{J} \min(C_{ijk}^{4} + \lambda_j + \mu_i S_{jk}^{4},\ O). \tag{22}$$

This represents the contribution to the objective function (21a) that is possible if $v_{ik} = 1$. For given λ_j and μ_i, it is optimal in (21) to choose those DCs corresponding to the p_k smallest values of R. (If $v_{ik} = 1$, then it is optimal to choose $u_{ijk} = 1$ only if $C_{ijk}^{4} + \lambda_j + \mu_i S_j^{4} \leqslant O$.)

If these optimal values of u_{ijk} satisfy constraints (20e) and (20c), then we obtain a solution to the original problem (20). If not, then the values of λ_j and μ_i are modified for the next iteration. If $\sum_{i\in G} u_{ijk} > 1$, then λ_j is increased, whereas if $\sum_{i\in G} u_{ijk} < 1$, λ_j is decreased. Likewise, if $\sum_{j=1}^{J} S_{jk}^{4} u_{ijk} > \tilde{B}_i$, then μ_i is increased. To avoid oscillations, we do not allow values of μ_i to decrease. Thus, if $\sum_{j=1}^{J} S_{jk}^{4} u_{ijk} \leqslant \tilde{B}_i$, the constraint is satisfied and μ_i is held fixed. Procedures for updating these multipliers are analogous to those described in Section 3.3.

If a feasible solution is not found within a reasonable number of iterations (again, 40 has been used successfully), then a solution is generated based on the last set G_k of R_i values. The DCs corresponding to the p_k smallest R_i values are chosen to serve product k (i.e., for these DCs set $v_{ik} = 1$). For each demand area j, we find index e such that

$$C_{ejk}^{4} + \mu_e S_{jk}^{4} = \min_{i\in G_k} \{C_{ijk}^{4} + \mu_i S_{jk}^{4}\}, \tag{23}$$

where the μ_i are also taken from the last iteration. Then, for each j, set

$$u_{ijk} = \begin{cases} 1 \text{ if } i = e \\ 0 \text{ otherwise} \end{cases} \quad \text{for } i \in G. \tag{24}$$

No immediate modification of this solution is made, even in the event that this solution violates capacity constraints (20e). The solution will be adjusted to feasibility later in the facility sizing routine (see Section 3.5). In the meantime, the amount of violation is small, since Type 4 products are assumed to constitute only a small portion of the demand.

The description of the algorithm above can be modified in a straightforward way so that the special DCs are restricted to be chosen from among a predetermined list of DCs for each special product. If none of the facilities from the predetermined list were to appear in the solution obtained in the primary optimization routine, then the model would serve demand areas directly from the Type 4 product vendors. This possibility could be completely avoided by fixing one or more of the facilities on the list to be open a priori.

3.5 LDC locations

After a set of open DCs and demand area assignments are chosen, both for Type 4 products (Section 3.3) and other products (Section 3.2), we attempt to introduce LDCs into the solution. First, we determine a tentative set of open LDC locations and accompanying demand area assignments. With LDCs in place, utilized capacity in some DCs may now be decreased, thus allowing additional demand areas to be assigned. We then check if any "small" DCs can now be closed advantageously and their demand areas reassigned.

3.5.1 Determining tentative LDC assignments

In this routine, we first order the DCs in terms of decreasing throughput (i.e., amount of warehouse space occupied). For each DC in order, we perform an "LDC assignment procedure" to determine the set of LDCs that should be associated with the DC. We consider the larger DCs first, since they are more likely to serve a larger geographic region and, hence, more likely to benefit from the presence of subsidiary LDC locations. The following is the LDC assignment procedure for DC i:

Step 1. Let Q_i denote the set of demand areas j that are assigned to DC i. Let E denote the set of possible LDC locations that have not already been assigned to another DC. For each $j \in Q_i$, determine the LDC $e \in E$ that satisfies

$$A^2_{ije} = \min_{\ell \in E} A^2_{ij\ell}. \tag{25}$$

For each such $j \in Q_i$, in turn, make a tentative assignment of j to e if

$A^2_{ije} < A^1_{ij}$, unless the assignment would result in the violation of capacity bound $\bar{B}^2_e$. If assigning j to a particular LDC e would violate the capacity bound, check other demand areas that are already assigned to e; if a feasible solution with lower total cost can be obtained by removing another demand area from LDC e and replacing it with demand area j, then do so. Let P_ℓ denote the set of demand areas j tentatively assigned to LDC ℓ at the end of step 1.

Step 2. Consider those LDC locations to which demand areas $j \in Q_i$ have been tentatively assigned at the end of step 1. Sort these LDCs in order of increasing throughput. Thus, underutilized LDCs, which are less likely to be cost-effective, are considered first.

Step 3. Consider each LDC ℓ on the list in order. If the throughput due to tentative assignments is greater than some threshold amount (say, some fraction of the minimum capacity B^2_ℓ), leave its assignments unchanged. Otherwise, remove the LDC from the list, cancel the tentative demand area assignments to LDC ℓ, and attempt to reassign the demand areas $j \in P_\ell$, if possible, to other LDCs. (That is, find another LDC $\bar{e}$ such that

$$A^2_{ij\bar{e}} = \min_{\substack{\ell \in E \\ \ell \neq e}} A^2_{ij\ell}. \tag{26}$$

If $A^2_{ij\bar{e}} < A^1_{ij}$, tentatively assign j to $\bar{e}$.)

Step 4. Resort LDCs remaining on the list in order of increasing throughput.

Step 5. For each LDC ℓ remaining on the list, estimate the total cost for serving demand areas $j \in P_\ell$ using DC i and LDC ℓ. This estimate includes the assignment costs $\sum_{j \in P_\ell} A^2_{ij\ell}$ plus the fixed setup cost α^2_ℓ, plus variable setup costs $\beta^2_\ell \sum_{j \in P_\ell} S^2_j$, plus the concave operating cost $f^2_\ell \left(\sum_{j \in P_\ell} S^2_j\right)$, and minus the closing cost c^2_ℓ. Compare this with an estimate of the cost for serving demand areas $j \in P_\ell$ using DC i alone. This estimate includes the assignment costs $\sum_{j \in P_\ell} A^1_{ij}$, plus variable setup costs $\beta^1_i \sum_{j \in P_\ell} S^1_j$, plus the difference in the concave operating cost function due to the additional throughput. If the cost using the LDC is less, then open LDC ℓ and make the tentative assignments permanent. If not, then attempt to tentatively reassign the demand areas $j \in P_\ell$ to other LDCs still on the list. □

Obviously, more sophisticated procedures can be designed to decide assignments for LDCs. However, since it is expected that demand areas will only be assigned to LDCs that are relatively proximate, the potential number of economically attractive assignments is limited. Thus, more sophisticated procedures have not been found necessary.

After completing this procedure for each DC i that is open, we have a tentative set of open LDCs and LDC assignments.

3.5.2 Closing small DCs

We allow DCs with relatively small throughput to be closed and their demand areas reassigned. We first sort open DCs in a list in order of increasing throughput. In this way, the smaller DCs that are more likely to close will be considered first. For each DC i on the list, we then execute a "DC closing procedure," which follows.

Step 1. If the throughput of DC i exceeds some minimum threshold (say, half the minimum capacity B_i^1), then leave it as is; go on to the next DC. If not, continue with step 2.

Step 2. For each $j \in Q_i$, attempt to tentatively reassign the demand area to another DC. This reassignment should be to a DC that is feasible; that is, the DC should have sufficient spare capacity to handle the demand area, and the demand area should be within the radius of operation for the DC. If there is more than one such feasible DC, choose the one that minimizes the assignment cost for j. If, for some $j \in Q_i$, no feasible reassignment is possible, then cancel all tentative reassignments, keep DC i open, and go on to the next DC. Otherwise, continue with step 3.

Step 3. Compare the additional assignment cost, operating cost and facility closing cost brought on by reassigning demand areas $j \in Q_i$. Compare this with the savings in setup cost for DC i. If it is advantageous to close DC i, make the tentative reassignments permanent. Otherwise, cancel all tentative reassignments and go on to the next DC. □

At this point, the set of open DCs and LDCs is determined. There remains only the question of sizing these open facilities, which we address in the next subsection.

3.6 Facility sizing

The model must determine the number q_i^1 of increments installed such that $B_i^1 + b_i^1 q_i^1 \leqslant \bar{B}_i^1$. At this point, there is a tentative set of demand area assignments that require amount σ_i^1 [see (8g)] of warehouse space at each DC i. The simplest sizing routine would be to choose q_i^1 to be the smallest integer such that $\sigma_i^1 \leqslant B_i^1 + b_i^1 q_i^1$. There are two reasons why we might want to modify this approach:

1. DC i is overcapacitated by an amount W_i (perhaps because of assignments that were made for Type 4 products).

2. By moving some demand areas to other DCs, we can perhaps install one less increment of space, thereby saving variable set-up cost $\beta_i^1 b_i^1$. Suppose that if we reduced the requirements for warehouse space at DC i by W_i units, we could install one less increment of space; call this value W_i the "excess" space requirement.

The facility sizing routine attempts to adjust demand area assignments in order to allow certain facilities to be installed at a smaller capacity. It begins by ordering the DCs in a list. First, any overcapacitated DCs are entered in the list. Next, other DCs are entered in the list in order of increasing W_i. For each DC on this list, in order, we then perform the following "DC sizing routine."

Step 1. For each demand area $j \in Q_i$ find its next best "feasible" DC assignment and compute the associated assignment cost differential. (By feasible we mean that assigning the demand area to the DC would not cause certain capacity limits to be exceeded. If DC i, the facility being sized, is overcapacitated, we take this limit to be the maximum capacity of the other DC; otherwise, we take it to be the capacity size required to serve the current assignments to the other DC. Assign an arbitrarily large cost differential if no other feasible assignment is possible.) At this point, consider only reassignments to a DC alone. The use of LDCs will be considered in step 5 below.

Step 2. Sort demand areas $j \in Q_i$ in order of increasing cost differential.

Step 3. Tentatively reassign a sufficient number of demand areas from the top of the list so as to decrease the throughput at DC i by an amount greater than or equal to the excess W_i.

Step 4. If DC i is overcapacitated, make permanent the reassignments found in step 3. If not, check the cost differentials for the reassignments. If the sum of the cost differentials for the reassigned demand areas is greater than the savings $\beta_i^1 b_i^1$ in variable setup cost, then cancel the reassignments. Otherwise, make the reassignments permanent.

Step 5. If demand area reassignments were made permanent in step 4, then consider the possible use of LDCs for each such demand area. In particular, if demand area j were permanently reassigned to a DC i', examine those LDCs ℓ that are now associated with DC i' (i.e., $z_{i'\ell} = 1$). Determine if total costs can be reduced by assigning demand area j to one of these LDCs. If so, assign j to the LDC that results in the minimum total cost. □

After this procedure is completed, the values of q_i^1 (number of space increments for DCs) and q_ℓ^2 (number of space increments for LDCs) are chosen to be the smallest integers that provide sufficient warehousing space to handle the assigned demand areas.

IV. IMPLEMENTATION DETAILS AND CONCLUDING REMARKS

To provide an effective tool for decision makers, our model was designed to be flexible and efficient. Flexibility in the implementation allows the user to analyze many different real situations using the same "generic" model. Further flexibility comes from an implementation that permits some variables to be fixed a priori, thus allowing the user to impose various "nonquantifiable" conditions on the model. For example, various DCs or LDCs can be fixed open or closed. Certain DCs can be forbidden from handling Type 4 products; certain demand areas can be assigned a priori to a given DC or LDC. These conditions can be imposed to reflect some physical constraint or corporate policy. Such conditions can also be imposed after studying a previous solution obtained from the model. In this way, the model is "forced" to consider alternate solutions of interest to the user. The user may also wish to perform a sensitivity analysis in which the model is run several times with variations in one or more cost parameters.

Efficiency of the algorithm is essential so that multiple runs as described above can be accomplished in a short time frame without excessive computation. The current implementation allows for a maximum of 10 different products, 50 possible DC locations, 40 possible LDC locations, and 200 demand areas. Some average run times for a variety of problems that have been encountered in practice are given in Table I. (Problem I, in particular, was the basis for a nationwide distribution planning study.) These times were obtained on an Amdahl 470/V8 operating under MVS; the code was compiled using the FORTRAN 77 compiler. All times are within an acceptable range for performing multiple runs in an economic study. Note that, as is typical in combinatorial optimization problems, run times can vary among different problems of the same size. For example, problem H is smaller than problem E, but took over twice the CPU time (262.2 seconds versus 115.6 seconds). (This variation is due primarily to the difference in the number of iterations required by the Kelly and Khumawala algorithm, which is used within the primary optimization procedure described in Section 3.3.)

The many cost components considered (including nonlinear operating costs), the ability to incorporate such features as different

Table I—Average execution times

Number of Problem	Number of Demand Areas	Number of DCs	Number of LDCs	Number of Special Products	CPU Seconds
A	30	8	10	1	4.6
B	54	8	25	0	11.1
C	149	8	25	1	16.7
D	54	48	25	0	61.2
E	100	50	20	1	115.6
F	100	48	25	0	126.6
G	149	48	25	0	170.1
H	100	50	0	0	262.2
I	158	50	0	0	603.7

product types and capacity bounds and subsidiary warehouses, the flexibility offered by the ability to fix variables a priori, and the efficiency in run times make this model a unique and useful tool. It should provide genuine insight when used to study existing or proposed material logistics systems.

V. ACKNOWLEDGMENTS

Hanan Luss provided very valuable advice and support throughout the development of this model. His suggestions greatly improved the solution algorithm and its presentation here. Patty Donohue developed the user interface routines and provided essential help and support during the implementation and testing of the program. Dave Aff, Marv Berkowitz, and Steve Waters contributed to many helpful discussions concerning the model formulation and data requirements.

REFERENCES

1. R. L. Francis, L. F. McGinnis, and J. A. White, "Locational Analysis," Eur. J. Oper. Res., *12* (March 1983), pp. 220–52.
2. G. Y. Handler and P. B. Mirchandani, *Location on Networks*, Cambridge: M.I.T. Press, 1979.
3. B. C. Tansel, R. L. Francis, and T. J. Lowe, "Location on Networks: A Survey, Part I: The p-Center and p-Median Problems," Manage. Sci., *29* (April 1983), pp. 482–97.
4. M. A. Efroymson and T. L. Ray, "A Branch-Bound Algorithm for Plant Location," Oper. Res., *14* (May–June 1966), pp. 361–8.
5. D. Erlenkotter, "A Dual-Based Procedure for Uncapacitated Facility Location," Oper. Res., *14* (November–December 1978), pp. 992–1009.
6. B. M. Khumawala, "An Efficient Branch-and-Bound Algorithm for the Warehouse Location Problem," Manage. Sci., *18* (August 1972), pp. 718–31.
7. A. A. Kuehn and M. J. Hamburger, "A Heuristic Program for Locating Warehouses," Manage. Sci., *9* (July 1963), pp. 643–66.
8. G. Cornuejols, G. L. Nemhauser, and L. A. Wolsey, "The Uncapacitated Facility Location Problem," Manage. Sci. Res. Rep. No. MSRR 493, Carnegie-Mellon University, August 1983.
9. J. Krarup and P. Pruzan, "Simple Plant Location Problem: Survey and Synthesis," Eur. J. Oper. Res., *12* (January 1983), pp. 36–81.
10. G. T. Ross and R. M. Soland, "Modeling Facility Location Problems as Generalized Assignment Problems," Manage. Sci., *24* (November 1977), pp. 345–57.
11. M. L. Fisher, "The Lagrangean Relaxation Method for Solving Integer Programming Problems," Manage. Sci., *27* (January 1981), pp. 1–8.
12. P. S. Davis and T. L. Ray, "A Branch-Bound Algorithm for the Capacitated Facilities Location Problem," Nav. Res. Logis. Quart., *16,* (September 1969), pp. 331–4.
13. A. M. Geoffrion and R. D. McBride, "Lagrangean Relaxation Applied to Capacitated Facility Location Problems," AIIE Trans., *10* (March 1978), pp. 40–7.
14. R. M. Nauss, "An Improved Algorithm for the Capacitated Facility Location Problem," J. Oper. Res. Soc., *29* (December 1978), pp. 1195–201.
15. D. L. Kelly and B. M. Khumawala, "Capacitated Warehouse Location With Concave Costs," J. Oper. Res. Soc., *33* (September 1982), pp. 817-26.
16. B. M. Khumawala and D. L. Kelly, "Warehouse Location With Concave Costs," INFOR, *12* (February 1974), pp. 55–65.
17. R. M. Soland, "Optimal Plant Location With Concave Costs," Oper. Res., *22* (March–April 1974), pp. 373–85.
18. A. M. Geoffrion and G. W. Graves, "Multicommodity Distribution System Design by Benders Decomposition," Manage. Sci., *20* (January 1974), pp. 822–44.
19. D. Erlenkotter, "A Comparative Study of Approaches to Dynamic Location Problems," Eur. J. Oper. Res., *6* (February 1981), pp. 133–43.

20. H. Luss, "Operations Research and Capacity Expansion Problems: A Survey," Oper. Res., *30* (September–October 1982), pp. 907–47.
21. C. L. Monma and M. Segal, "A Primal Algorithm for Finding Minimum-Cost Flows in Capacitated Networks With Applications," B.S.T.J., *61* (July–August 1982), pp. 949–68.
22. M. Held, P. Wolfe, and H. Crowder, "Validation of Subgradient Optimization," Math. Program., *6* (January 1974), pp. 62–88.
23. H. M. Salkin, *Integer Programming,* Reading, MA: Addison-Wesley, 1975.
24. S. C. Narula, U. I. Ogbu, and H. M. Samuelsson, "An Algorithm for the p-Median Problem," Oper. Res., *25* (July–August 1977), pp. 709–12.
25. S. L. Hakimi, "Optimal Locations of Switching Centers and the Absolute Centers and Medians of a Graph," Oper. Res., *12* (May–June 1964), pp. 450–9.

AUTHOR

John G. Klincewicz, S.B. (Mathematics), 1975, The Massachusetts Institute of Technology; M.A., 1977, and Ph.D. (Operations Research), 1979, Yale University; AT&T Bell Laboratories, 1979—. At AT&T Bell Laboratories, Mr. Klincewicz is a member of the Operations Research Department. His research interests include applications of mathematical programming and the development of algorithms for network flow problems and facility location problems. Member, Operations Research Society, Mathematical Programming Society.

AT&T Technical Journal
Vol. 64, No. 7, September 1985
Printed in U.S.A.

A Priority-Based Admission Scheme for a Multiclass Queueing System

By K. M. REGE and B. SENGUPTA*

(Manuscript received November 21, 1984)

We consider a queueing problem involving multiple priority classes where the station is divided into waiting and service areas. The service area has a finite number of positions where a customer of a particular class has access to only a subset of these positions. The admission into the service area is controlled by a mechanism that allows customers within a priority class to enter the service area on a first-come first-served basis. The customers of different classes are assumed to be indistinguishable once they have entered the service area. We consider service under three different disciplines: last-come first-served preemptive resume, multiple server, and processor sharing. We show that the waiting time of a customer is related to that of a customer in an equivalent M/G/1 queue. We characterize the Laplace-Stieltjes transform of the time spent in the service area. We also discuss three potential applications in the area of computer and communication systems.

I. INTRODUCTION

This paper is concerned with investigating a queueing system in which customers from n different job classes representing various priority levels receive service in a service area with a finite capacity of m. The capacity, m, of the service area refers to the maximum number of customers that can be present in the service area at any time. We describe an admission scheme that allows preferential access to the service area by the higher-priority customers. This scheme may give rise to a smaller waiting time before entry into the service area for the

* Authors are employees of AT&T Bell Laboratories.

higher-priority customers. The customer classes are assumed to be indistinguishable after they have gained access to the service area. The performance measures that we characterize are the distributions of waiting time and the mean time spent in the service area. This work has potential applications in the design of multiprogramming levels for computer systems and the design of window levels for communication systems.

More formally, let n independent classes of customers arrive at a queueing station, each according to a Poisson process. Let the mean arrival rate of class i be λ_i. The classes are arranged according to decreasing order of priority, that is, class 1 has the highest priority and class n has the lowest priority. The queueing station is divided into n waiting areas (one for each class) and a service area. The service area can hold at most m customers simultaneously. Service is provided at a state-dependent rate of μ_i whenever there are i customers present in the service area. We consider three service disciplines within the service area: Last-Come First-Served Preemptive Resume (LCFS-PR), m Server (MS), and Processor Sharing (PS). The admission into the service area is controlled by means of a gate that allows customers from the waiting areas to enter the service area on the basis of the contents of the service area. The admission policy gives preferential treatment to higher-priority customers in gaining access to the service area. In particular, the admission policy is governed by two rules:

1. When a customer of class i is admitted to the service area, the waiting areas of classes $1, \cdots, i-1$ must be empty.

2. When a class i customer enters the service area, the number of customers in the service area (excluding itself) must be less than k_i. The sequence $\{k_i,\ i = 0, \cdots, n+1\}$ is a set of strictly decreasing, nonnegative integers with $k_0 = \infty$, $k_1 = m$, and $k_{n+1} = 0$.

This admission scheme reserves some slots in the service area for the exclusive use of the high-priority customers. In the case of the MS model, this means that at times the low-priority customers will not be allowed to enter the service area although some servers are idle. Obviously this is not the most efficient way of utilizing the servers' capacity. However, when the designer's overriding concern is to reduce the delays suffered by the high-priority customers, it is useful to reserve certain slots exclusively for the high-priority customers. The queueing station is shown schematically in Fig. 1.

In this paper, we show that this problem has an interesting structure, which can be exploited to characterize (1) the waiting-time distribution, by class, and (2) the mean time spent in the service area, by class. For the special case of the PS discipline, we show how to obtain this mean when n equals two. Since writing this paper, it has been brought to our attention that Schaack and Larson[1] have independently

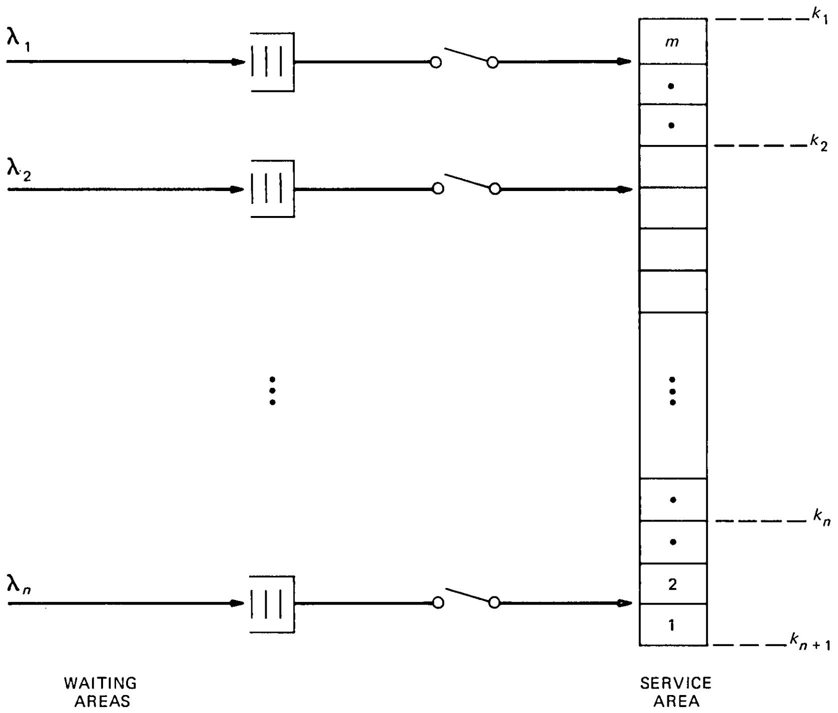


Fig. 1—The queueing station.

studied the special case of the MS discipline and reported the same results as we do here.

This paper is organized into four sections. In Section I, we discuss some potential applications for this model. In Section II, we derive the waiting-time distribution, by class. We characterize the mean time spent in the service area, by class, in Section III. Section IV summarizes our conclusions.

II. POTENTIAL APPLICATIONS

We discuss three potential applications in this section. In the first, we propose this scheme for sharing of multiprogramming threads by several job classes in a computer system. Our second and third examples propose this admission scheme for sharing a window size by several job classes at the link level and the application level, respectively, in a communication system.

Avi-Itzhak and Heyman[2] had first proposed that a state-dependent server be used to approximate the CPU and disk subsystem of a computer. We depict a multiple CPU and disk subsystem in Fig. 2,

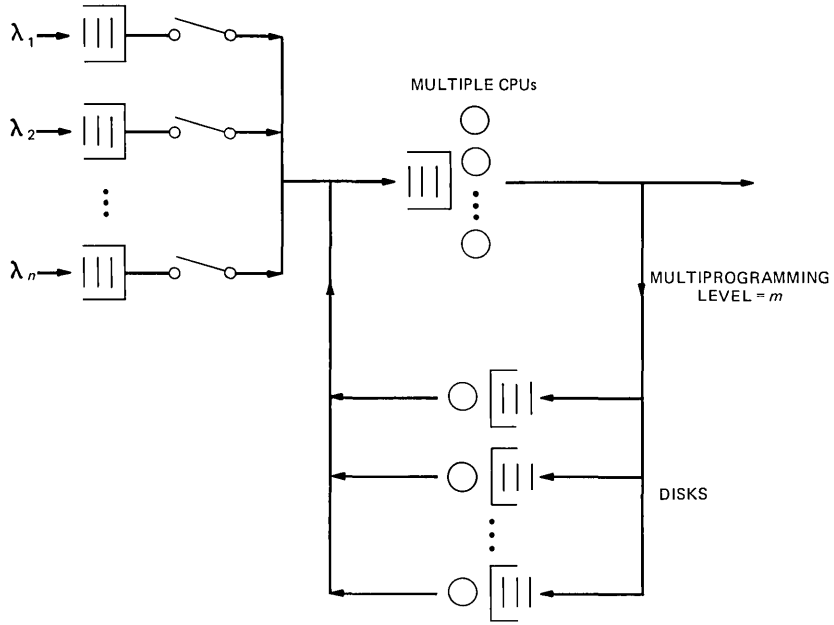


Fig. 2—A computer with shared multiprogramming threads.

which is approximated by a state-dependent server in our model. The service rate μ_i of our model is obtained by solving for the throughput in the closed queueing network of Fig. 2 with a population size of i. It is assumed in our model that the service requirements in terms of CPU and I/O times are approximately the same for all classes of customers. Also, the service discipline for the state-dependent server that is most appropriate for this application is PS. The closed queueing network of Fig. 2 can be solved by mean-value analysis described by Reiser and Lavenberg.[3] The circumstances under which a single state-dependent server is a good approximation of the CPU and I/O subsystem was investigated by Fredericks.[4] This approximation is usually good when each customer makes many trips to the I/O devices and when the CPU and I/O times required by a customer are not too unbalanced. In our model, m is the multiprogramming level of the computer, usually determined by considerations such as available memory and the extent to which the jobs require concurrent access of the same databases. Given m, our model can be used to determine a way to allocate available multiprogramming threads to the various job classes so that some requirements on mean response time can be met.

A second application would be in the modeling of a link layer protocol such as high-level data link control. Assume that a provider of packet-switching service offers n grades of service, each with its own response-time requirement. The different grades of service may be provided by appropriately sharing a link-level window size of m among the packets of n service grades. The admission scheme proposed in the Introduction is one means of offering different levels of service to customers. In this application, the queueing station represents a node in the network where the customers in the service area correspond to the jobs ready for transmission. Service provided to a customer constitutes its transmission to an adjacent node and the return of the acknowledgment. Since data links are often characterized by relatively low utilizations, the value of μ_i may be approximately proportional to i, at least for small values of i. The constant of proportionality may be taken as the mean round-trip time to receive an acknowledgment on a link that has no traffic. This linearity would imply that there is hardly any wait for transmission to commence once a packet has entered the service area. For larger values of i, some saturation of μ_i will take place as the presence of a large number of packets in the service area starts to choke the capacity of the link. The limiting value of μ_i may be chosen as the rate at which acknowledgments can be returned in a fully utilized link. We show the closed queueing network used for calculating μ_i in Fig. 3. This is an approximation along the lines of one proposed by Schwartz.[5]

The third potential application is from the point of view of a user of a data network. This potential application is similar in spirit to the previous one except that we are concerned with the high-level protocol of host-to-host traffic using a data network. The network itself is approximated by a state-dependent server. The closed queueing network used for calculating μ_i is shown in Fig. 4. This approximation was proposed by Reiser.[6,7] The user of the network must allocate a window size of m to n different types of traffic. The admission scheme described earlier may enable the user to determine a way to share the

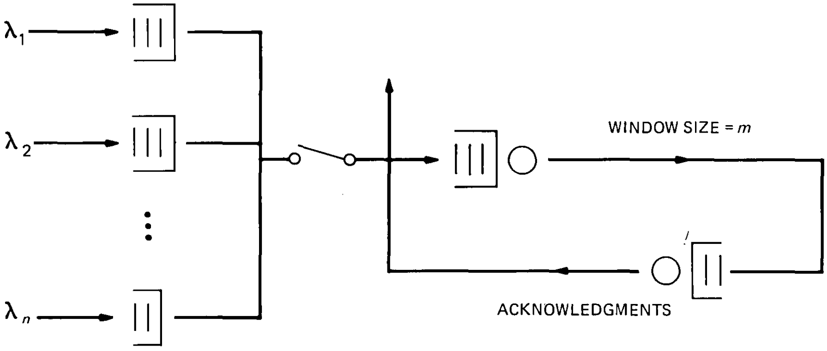


Fig. 3—Link layer protocol.

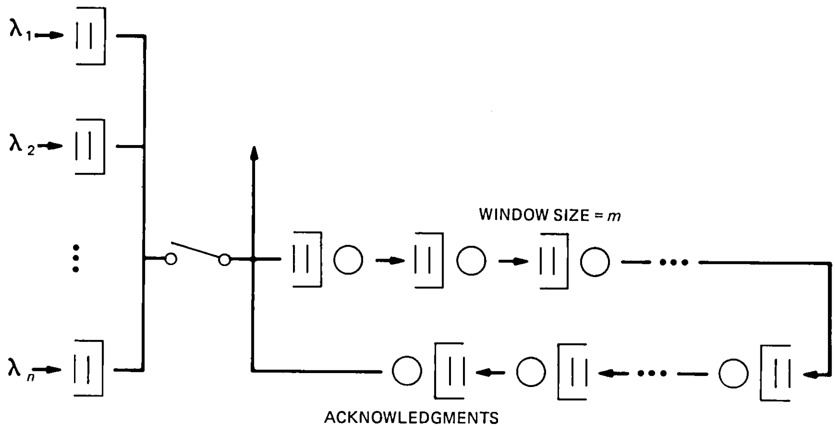


Fig. 4—High-level protocol.

window size among the types of traffic to meet certain response-time criteria.

III. THE WAITING-TIME DISTRIBUTION

In this section, we will characterize the waiting-time distributions without assuming anything about the service discipline. The key result of this section is that, given that a customer of class l is required to wait, its waiting-time distribution is related to that of a suitable M/G/1 queue.

The state of the system is completely specified by an n-tuple $(J_1, J_2, \cdots, J_i, \cdots, J_n)$, where $J_i (i = 2, \cdots, n)$ represents the number of customers of class i waiting and J_1 represents the number of customers of class 1 waiting plus the number of customers present in the service area. With this description of the state space, one can, in principle, write down a system of equations for the steady-state probability vector $P(\mathbf{j}) = P(J_1 = j_1, J_2 = j_2, \cdots, J_n = j_n)$. In particular, for $l = 0, \cdots, n-1$

$$(\Lambda + \mu_{j_1})P(\mathbf{j}) = \sum_{k=l+1}^{n} \lambda_k P(\mathbf{j} - \mathbf{e}_k)\delta(j_k) + \sum_{k=1}^{l} \lambda_k P(\mathbf{j} - \mathbf{e}_1)\delta(j_1)$$

$$+ \mu_{j_1+1} P(\mathbf{j} + \mathbf{e}_1) + \mu_{j_1} P(\mathbf{j} + \mathbf{e}_{l+1})(1 - \delta(j_1 - k_{l+1}))\delta(l), \quad (1)$$

where

$$k_{l+1} \leq j_1 < k_l,$$

$$\Lambda = \sum_{k=1}^{n} \lambda_k,$$

$\delta(x) = 1$ if $x > 0$ and 0 otherwise, and $\mathbf{e}_k$ is an n-tuple with a 1 in the kth position and 0 elsewhere. For $j_1 \geq m$, μ_{j_1} is to equal to $\mu_m (= \mu)$, since for $j_1 \geq m$, the number of customers in the service area is m. The solution of this equation is not easy; however, it is possible to solve it for the case where $n = 2$. Since the solution of this equation does not concern us at present, we show how to calculate this for $n = 2$ in Appendix A.

We will now concentrate our attention on the stochastic process defined by the random variable J_1. Let $\{u_i\}$ be the steady-state marginal probability distribution of J_1. Since arrivals are Poisson, a customer of class l is required to wait outside with probability $\sum_{j=k_l}^{\infty} u_j$. Clearly, the waiting time is 0 whenever an arrival finds that $J_1 < k_l$. Let us now start observing the system when J_1 changes its value from $k_l - 1$ to k_l. Let t_0 denote this instant. At t_0, a customer of type j, with $j \leq l$, arrives to find exactly $k_l - 1$ customers in the service area and is immediately admitted for service. Let t_f denote the first instant after t_0 when J_1 moves from k_l to $k_l - 1$ with no type l customers waiting outside. During the open interval (t_0, t_f), several type l customers may get admitted to the service area. If $n(n \geq 0)$ type l customers are admitted to the service area during the open interval (t_0, t_f), let $t_1, t_2, \cdots, t_n$ denote instants when these admissions took place (refer to Fig. 5). Note that at the instants $t_1, t_2, \cdots, t_n$, a departure occurs from the state $J_1 = k_l$ and there is at least one type l customer waiting outside. Also, at these instants there cannot be any higher-priority customers in the waiting area. Now let us focus our attention on the intervals (t_0, t_1), (t_1, t_2), $\cdots$, (t_{n-1}, t_n) and (t_n, t_f). [In case n equals 0, we need consider just one interval (t_0, t_f).] The lengths of these intervals are governed by the customers inside the service area, which by assumption are indistinguishable, and by arrivals of

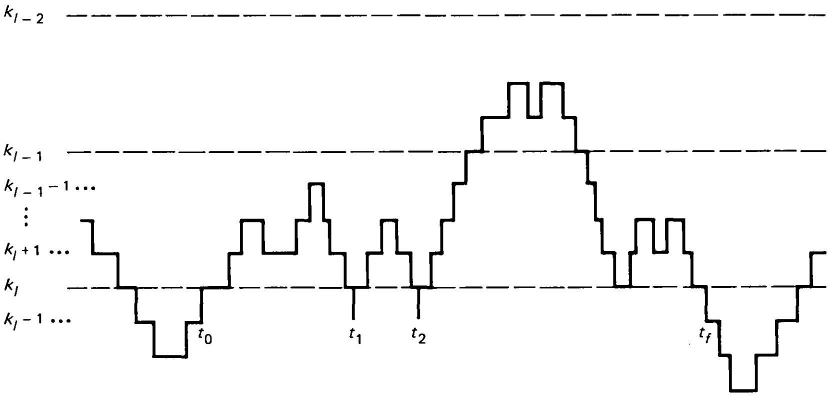


Fig. 5—A typical sample function of the process J_1. (The notches at t_1 and t_2 represent the event that a departure occurred from the state $J_1 = k_l$ and a waiting type l customer was immediately admitted to the service area.)

customers of priority higher than that of type l customers (i.e., the types of these customers are less than l) that are Poisson. Also, since the service requirements are memoryless and since the end points of these intervals are marked by identical states so far as customers of priority higher than that of type l customers are concerned, the lengths of these intervals are independent and identically distributed (i.i.d.) random variables. Let H_l denote the generic random variable corresponding to these lengths and let $H_l(s)$ denote the Laplace-Stieltjes Transform (LST) of its distribution function.

Now a type l customer is required to wait if it arrives during an open interval similar to the interval (t_0, t_f) described in the previous paragraph. Since the interadmission times for type l customers during such an interval are i.i.d. random variables with LST of distribution function $H_l(s)$, it follows that the waiting-time distribution of a type l customer, given that it is required to wait, is identical to that of a customer in an M/G/1 queue [with arrival rate λ_l and LST of service-time distribution $H_l(s)$], which arrives to find the server busy. Let $W_l(s)$ denote the LST of the waiting-time distribution of a type l customer given that it is required to wait. Then, from page 223 of Ref. 8, we have

$$W_l(s) = \frac{(1 - H_l(s))(1 - \rho_l)}{(s - \lambda_l + \lambda_l H_l(s))E(H_l)}, \tag{2}$$

where

$$\rho_l = \lambda_l E(H_l)$$

and

$$l = 2, \cdots, n.$$

The LST of the unconditional waiting-time distribution for a customer of type l is given by

$$\sum_{j=0}^{k_l-1} u_j + W_l(s) \sum_{j=k_l}^{\infty} u_j. \tag{3}$$

The waiting-time distribution of a class 1 customer is easier to characterize. Given that a class 1 customer has to wait, its waiting time is the same as the sojourn time in an M/M/1 queue where the server is working at a rate of $\mu_m (=\mu$, say). So,

$$W_1(s) = \frac{\mu - \lambda_1}{\mu - \lambda_1 + s}, \tag{4}$$

and the unconditional waiting-time distribution is given by

$$\sum_{j=0}^{k_1-1} u_j + W_1(s) \sum_{j=k_1}^{\infty} u_j. \tag{5}$$

In the remainder of this section, we will show how to calculate u_j and $H_l(s)$. Let us define a random variable B_l to be the elapsed time from the instant J_1 changes from $k_l - 1$ to k_l until the next instant when the value of J_1 drops from k_l to $k_l - 1$ and $J_l = 0$. In other words, B_l is the length of an interval similar to (t_0, t_f) discussed earlier. From the preceding discussion it should be clear that B_l constitutes a busy period of an M/G/1 queue with arrival rate λ_l and the LST of service-time distribution $H_l(s)$. Let $B_l(s)$ be the LST of the distribution of B_l. Then $B_l(s)$ and $H_l(s)$ are related by (see page 212 of Ref. 8)

$$B_l(s) = H_l[s + \lambda_l - \lambda_l B_l(s)]. \tag{6}$$

Since $B_1(s)$ represents the busy period of an M/M/1 queue,

$$B_1(s) = \frac{\mu + \lambda_1 + s - [(\mu + \lambda_1 + s)^2 - 4\mu\lambda_1]^{1/2}}{2\lambda_1} \tag{7}$$

from page 215 of Ref. 8.

Next, we define the random variable C_j to denote the first passage time from the state $J_1 = j$ to $J_1 = j - 1$, where $k_{l-1} > j > k_l$. Let $C_j(s)$ be the LST of the distribution of C_j. It should be clear that the waiting customers of class $l, \cdots, n$ play no role in determining this first passage time. Further, $J_1 = j$ implies that $J_2 = J_3 = \cdots = J_{l-1} = 0$. From these observations, now it is possible to write down the following equations for $C_j(s)$:

$$C_j(s) = \left(\frac{1}{\Lambda_l + \mu_j + s}\right)(\Lambda_l C_{j+1}(s) C_j(s) + \mu_j)$$

$$\text{for} \quad j = k_l + 1, \cdots, k_{l-1} - 2;$$

$$C_{k_{l-1}-1}(s) = \left(\frac{1}{\Lambda_l + \mu_{k_{l-1}-1} + s}\right)(\Lambda_l B_{l-1}(s) C_{k_{l-1}-1}(s) + \mu_{k_{l-1}-1});$$

$$H_l(s) = \left(\frac{1}{\Lambda_l + \mu_{k_l} + s}\right)(\Lambda_l C_{k_l+1}(s) H_l(s) + \mu_{k_l});$$

and

$$\Lambda_l = \sum_{k=1}^{l-1} \lambda_k \quad \text{for} \quad l = 2, \cdots, n. \tag{8}$$

Thus, eq. (8) defines a recursive technique for obtaining $H_l(s)$ from $B_{l-1}(s)$ via the functions $C_j(s)$ for $k_{l-1} > j > k_l$. By using eq. (6), one can obtain $B_l(s)$ from $H_l(s)$; and eq. (7) provides the value of $B_1(s)$, the boundary for eq. (8) when $l = 2$.

Finally, we show how to characterize u_j for $j = 0, \cdots$, which is the steady-state marginal distribution of J_1. To do this, we define n Semi-

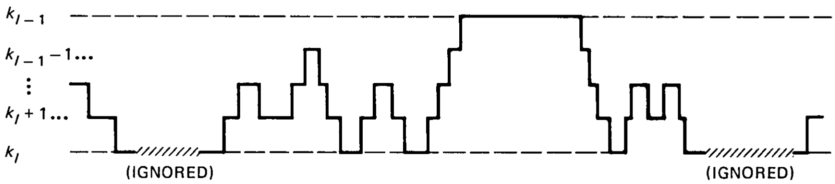


Fig. 6—Sample function of the lth SMP derived from the sample function of J_1 shown in Fig. 5.

Markov Processes (SMP), where the lth SMP has states $k_l, \cdots, k_{l-1}$, where $l = 2, \cdots, n+1$. The state of the lth SMP is the realization of the random variable J_1 with two differences. When $J_1 > k_{l-1}$, we will assume that the state of the SMP is k_{l-1}. Further, we will simply ignore the times when $J_1 < k_l$. A typical sample function of the lth SMP shown in Fig. 6 may help illustrate the structure of these SMPs. The transition probability for the SMP from state j to $j+1$ is $\Lambda_l/(\Lambda_l + \mu_j)$ and j to $j-1$ is $\mu_j/(\Lambda_l + \mu_j)$, where $k_l < j < k_{l-1}$. The holding time in state $j\,(k_l < j < k_{l-1})$ is exponential with rate $\Lambda_l + \mu_j$. For state k_{l-1}, the holding time is B_l and this state makes a transition into state $k_{l-1} - 1$ with probability 1. State k_l makes a transition into $k_l + 1$ with probability 1, and the holding time is exponential with rate Λ_l. In the description of the SMP, only one statement needs clarification, that is, the holding time in state k_l. Since we are ignoring all times when $J_1 < k_l$, this is equivalent to ignoring the transition of J_1 from k_l to $k_l - 1$. The result follows from observing that the transition from k_l to $k_l + 1$ occurs at an exponential rate of Λ_l. It is relatively easy to solve these n SMPs using methods described in Ross.[9] Let π_{jl} be the steady-state probability of state j in the lth SMP ($l = 2, \cdots, n+1$; $j = k_l, \cdots, k_{l-1}$). Then it should be clear that

$$P(J_1 = j \mid J_1 \geq k_l) = u_j \Big/ \sum_{i=k_l}^{\infty} u_i = \pi_{jl}$$

$$\text{for} \quad j = k_l, \cdots, k_{l-1} - 1 \quad (9)$$

and

$$P(J_1 \geq k_{l-1} \mid J_1 \geq k_l) = \sum_{i=k_{l-1}}^{\infty} u_i \Big/ \sum_{i=k_l}^{\infty} u_i = \pi_{k_{l-1}l}.$$

It is easy to use (9) to calculate $u_j\,(j = 0, \cdots, m-1)$ recursively, starting with the solution of the $(n+1)$st SMP and working backwards through to the second SMP. For $j \geq m$, the server always works at a rate of $\mu_m (=\mu)$, and the random variable J_1 behaves like the number in the system for an M/M/1 queue. Thus, we have

$$u_j = (1 - \lambda_1/\mu)(\lambda_1/\mu)^{j-m}\left(1 - \sum_{j=0}^{m-1} u_j\right)$$

for $j \geq m$. We note at this point that the probability that the system is idle is given by u_0.

IV. THE TIME SPENT IN THE SERVICE AREA

In this section, we describe methods of obtaining the mean time spent in the service area by class for the LCFS-PR and MS disciplines. For the PS discipline, we describe a method for characterizing the means when $n = 2$.

4.1 The LCFS-PR discipline

In this discipline, we will assume that on entry into the service area, a customer occupies the lowest-numbered service position that is empty. Further, the server renders service to the customer in the highest-numbered service position that is nonempty. Thus, a customer occupies the same service position from entry until departure. Let $T_j(s)$ be the LST of the distribution of time spent in the service area by a customer who occupies position j on entry into the service area. Then,

$$T_{k_1}(s) = \mu/(\mu + s)$$

and

$$T_j(s) = \begin{cases} C_j(s) & \text{if} \quad k_{l-1} > j > k_l \\ H_l(s) & \text{if} \quad j = k_l \end{cases} \quad \text{and} \quad l = 2, \cdots, n. \qquad (10)$$

It is now easy to use the results of Section II to obtain this LST or any other characterization of this distribution.

4.2 The multiple-server discipline

In this discipline, let $\mu_i = i\sigma$. Then the time spent in the service area is simply exponential with parameter σ.

4.3 The processor sharing discipline

The exact solution for the mean time spent in the service area can be obtained by first noting that customer classes are indistinguishable on entry into the service area. We denote the state of the system by an n-tuple $(J_1, J_2, \cdots, J_n)$ as seen by a customer of class l after entry into the service area and let $Q_l(\mathbf{j})$ denote the probability that the state of the system is $\mathbf{j}$ when an arbitrary class l customer is admitted to the service area, where j_1 includes the newly admitted customer. Further, let $x(\mathbf{j})$ be the mean time spent in the service area for a customer who sees state $\mathbf{j}$ immediately on being admitted to the service area. If we let t_l be the mean time spent in the service area by a class l customer, then

$$t_l = \sum_{j \in A_l} x(\mathbf{j}) Q_l(\mathbf{j}), \tag{11}$$

where

$$A_l = \{(j_1, j_2, \cdots, j_n) \mid j_k = 0 \text{ for } 2 \le k < l\} \quad \text{for} \quad l = 1, \cdots, n.$$

It is possible to obtain $Q_l(\mathbf{j})$ from the following observations:

1. For $j_1 < k_l$,

$$Q_l(\mathbf{j}) = P(\mathbf{j} - \mathbf{e}_1).$$

For a customer of class l to see j_1 including itself, there must be $j_1 - 1$ ahead of it in the service area.

2. Whenever a customer of class l enters the service area, $J_2 = J_3 = \cdots = J_{l-1} = 0$ and $J_1 \le k_l$.

3. For $j_1 = k_l$, one of two disjoint events must occur. Either the class l customer arrived to see $k_l - 1$ customers ahead of it in the service area or it must have waited in the waiting area prior to admission. The former case is identical to the first observation above. In the latter case, we have to characterize the distribution of $(J_l, \cdots, J_n)$ at the time of entry into the service area. The random variable J_l behaves like the number waiting as seen by a customer about to enter service given that it had to wait in an M/G/1 queue with an arrival rate of λ_l and a service time of H_l. The distribution of $J_k (k = l + 1, \cdots, n)$ is simply the convolution of what was seen on arrival and the number of new arrivals of type k during the wait of the customer of class l.

In principle, it is possible to write down $Q_l(\mathbf{j})$ in terms of $P(\mathbf{j})$ from the observations made above. The notation is cumbersome, so we will not go into the details here. The exact derivation when $n = 2$ is given in Appendix B.

Further, for $k_{l+1} \le j_1 < k_l$, the $x(\mathbf{j})$ satisfy

$$\begin{aligned}(\Lambda + \mu_{j_1}) x(\mathbf{j}) = 1 &+ \sum_{k=l+1}^{n} \lambda_k x(\mathbf{j} + \mathbf{e}_k) + \sum_{k=1}^{l} \lambda_k x(\mathbf{j} + \mathbf{e}_1) \\ &+ \left(\frac{j_1 - 1}{j_1}\right) \delta(j_1) \mu_{j_1} \{x(\mathbf{j} - \mathbf{e}_1) \delta(j_1 - k_{l+1}) + [x(\mathbf{j} - \mathbf{e}_1)(1 - \delta(j_{l+1})) \\ &+ x(\mathbf{j} - \mathbf{e}_{l+1}) \delta(j_{l+1})](1 - \delta(j_1 - k_{l+1}))\}. \end{aligned} \tag{12}$$

The solution of this equation is not easy; however, it is possible to solve it for the case where $n = 2$. We present this solution in Appendix C.

V. CONCLUDING REMARKS

In the earlier sections, we have shown how to characterize distributions of the waiting time and the time spent in the service area. Of interest in many applications would also be the sojourn time (i.e., the elapsed time between arrival into and departure from the system) of customers. In principle, it is possible to characterize the sojourn time distribution for the two-class PS problem by the methods used in Ref. 10.

We note that the time spent in the service area for the first-come first-served discipline is somewhat difficult to characterize, whereas the results for the waiting time is the same as that in Section II. The reason for this difficulty can be seen by first assuming that we are about to characterize a two-class problem. Then the mean time spent in the service area has to be found conditioned on J_1, J_2 and the position of the tagged customer in the service area. The difference equations for this mean time thus will be in three variables and are hard to solve.

REFERENCES

1. C. Schaack and R. C. Larson, "An N Server Cutoff Multi-Priority Queue," Working Paper No. OR135-85, MIT, Cambridge (February 1985).
2. B. Avi-Itzhak and D. P. Heyman, "Approximate Queueing Models for Multiprogramming Computer Systems," Oper. Res., *21*, No. 6 (November–December 1973), pp. 1212–30.
3. M. Reiser and S. S. Lavenberg, "Mean Value Analysis of Closed Multichain Queueing Networks," IBM Research Report RC7023, 1978.
4. A. A. Fredericks, "Approximations for Customer Viewed Delays in Multiprogrammed Transaction Oriented Computer Systems," B.S.T.J., *59*, No. 9 (November 1980), pp. 1559–74.
5. M. Schwartz, "Performance Analysis of the SNA Virtual Route Pacing Control," IEEE Trans. Commun., *COM-30*, No. 1 (January 1982), pp. 172–84.
6. M. Reiser, "Admission Delays on Virtual Routes With Window Flow Control," Performance of Data Communications Systems, G. Pujolle, Ed., New York: North-Holland, 1981.
7. M. Reiser, "A Queueing Network Analysis of Computer Communication Networks With Window Flow Control," IEEE Trans. Commun., *COM-27*, No. 8 (August 1979), pp. 1199–209.
8. L. Kleinrock, *Queueing Systems, Vol. 1: Theory*, New York: Wiley, 1975.
9. S. M. Ross, "Applied Probability Models With Optimization Applications," San Francisco: Holden-Day, 1969.
10. K. M. Rege and B. Sengupta, "Sojourn Time Distribution in a Multiprogrammed Computer System," AT&T Tech. J., *64*, No. 5 (May–June 1985), pp. 1077–90.

APPENDIX A

The Steady-State Probabilities for a Two-Class Problem

In this appendix, we show how to obtain the solution to eq. (1) when $n = 2$. For the sake of notational ease, we will refer to the random variables J_1 and J_2 as I and J in this and the other appendices. Further, i and j will be the realizations of the random variables I and J, respectively. Let P_{ij} be the steady-state probability that $I = i$ and $J = j$. Equation (1) reduces to the following equations when n is 2:

$$(\lambda_1 + \lambda_2 + \mu_i)P_{ij} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + \mu_{i+1}P_{i+1,j} \qquad \text{if} \quad i > k_2, \quad j > 0, \tag{13}$$

$$(\lambda_1 + \lambda_2 + \mu_i)P_{i0} = \lambda_1 P_{i-1,0} + \lambda_2 P_{i-1,0} + \mu_{i+1}P_{i+1,0} \qquad \text{if} \quad 0 < i < k_2, \tag{14}$$

$$(\lambda_1 + \lambda_2 + \mu_{k_2})P_{k_2,0} = (\lambda_1 + \lambda_2)P_{k_2-1,0} + \mu_{k_2}P_{k_2,1} + \mu_{k_2+1}P_{k_2+1,0} \tag{15}$$

$$(\lambda_1 + \lambda_2 + \mu_i)P_{i0} = \lambda_1 P_{i-1,0} + \mu_{i+1}P_{i+1,0} \quad \text{if} \quad i > k_2, \tag{16}$$

$$(\lambda_1 + \lambda_2 + \mu_{k_2})P_{k_2,j} = \lambda_2 P_{k_2,j-1} + \mu_{k_2}P_{k_2,j+1} + \mu_{k_2+1},\, P_{k_2+1,j} \qquad \text{if} \quad j > 0 \tag{17}$$

and

$$(\lambda_1 + \lambda_2)P_{00} = \mu_1 P_{10}. \tag{18}$$

It should be clear that u_0 (of Section III) is the same as P_{00}. So, one can obtain P_{i0} for $0 \le i \le k_2$ by using (18) and (14). Even though the coefficients of P_{ij} in eqs. (13) and (16) depend on i, it is easy to see that these are constant coefficient partial difference equations when $i \ge m$. Further, these equations are very similar to the ones solved by Rege and Sengupta.[10] Using these methods, it is easy to show that

$$P_{i0} = B_0\sigma_2^i \tag{19}$$

and

$$P_{ij} = \frac{\rho_2}{(\sigma_1 - \sigma_2)}\left(\sum_{\nu=i+1}^{\infty} \sigma_1^{i-n}P_{\nu,j-1} + \sum_{\nu=m}^{i} \sigma_2^{i-n}P_{\nu,j-1} + B_j\sigma_2^i\right) \qquad \text{for} \quad i \ge m \text{ and } j > 0, \tag{20}$$

where

$$\sigma_1 = \left(1 + \rho_1 + \rho_2 + \sqrt{(1 + \rho_1 + \rho_2)^2 - 4\rho_1}\right)/2,$$

$$\sigma_2 = \left(1 + \rho_1 + \rho_2 - \sqrt{(1 + \rho_1 + \rho_2)^2 - 4\rho_1}\right)/2,$$

$$\rho_1 = \lambda_1/\mu, \quad \rho_2 = \lambda_2/\mu,$$

and $\{B_j, j = 0, 1, \cdots\}$ constitute a sequence of unknown constants to be determined from the boundary conditions (15) and (17).

Now we will show how to determine the unknown constants in two steps. First, we will show this for B_0 and then for B_j. It is possible to determine B_0 by assuming two trial values and using (19) and (16) to recursively calculate two sets of P_{i0} for $i = m + 1, m, \cdots, k_2$. Since each of these P_{i0} is a linear function of the unknown constant B_0, now it is easy to use linear interpolation to obtain the correct value of B_0 that agrees with $P_{k_2,0}$ already obtained from (14). One can now use

(15) to obtain $P_{k_2,1}$. To calculate $B_j (j > 0)$, let us assume the $P_{k_2,j}$ has been obtained from (15) for $j = 1$ or from (17) for $j > 1$. Further assume that $P_{ik} (0 \leq k \leq j - 1$ and all $i)$ are known. As before, we start with two trial values of B_j and recursively calculate two sets of P_{ij} for $i = m + 1, \cdots, k_2$ by using (20) and (13). Since each of the P_{ij} is a linear function of B_j, we can use linear interpolation to obtain the correct value of B_j that agrees with $P_{k_2,j}$ already obtained from (15) for $j = 1$ or from (17) for $j > 1$. Finally, one can use (17) to obtain $P_{k_2,j+1}$.

APPENDIX B

State Probability As Seen by Customers on Admission to the Service Area

Here we describe the procedures for computing $Q_1(i,j)$ and $Q_2(i,j)$ that denote the state probabilities as seen upon admission to the service area by type 1 and type 2 customers, respectively. The procedure for computation of $Q_2(i,j)$ is described first.

It is clear that if, on arrival, a type 2 customer finds less than k_2 customers in the system, it does not have to wait before entering the service area. So,

$$Q_2(i,j) = P(i - 1, j) \quad \text{for} \quad i < k_2. \tag{21}$$

It is also obvious that $Q_2(i,j) = 0$ for $i > k_2$, since a type 2 customer cannot enter the service area if the number of customers in the service area other than itself is greater than or equal to k_2. A type 2 customer will see $I = k_2$ upon its admission to the service area in one of two ways: (1) if there are $k_2 - 1$ customers in the system just before its arrival, or (2) if $I \geq k_2$ at the time of its arrival and it has to wait until all type 2 customers ahead of it in the waiting area are admitted to the system and a departure occurs from the state $I = k_2$.

From the analysis of Section II, $W_2(s)$ is the LST of the distribution of the waiting time of a type 2 customer given that it has to wait. The generating function of the number of arrivals of type 2 during this wait is $W_2(\lambda_2(1 - z))$. Further, the probability that a type 2 customer has to wait before entering the service area is $\sum_{i=k_2}^{\infty} \mu_i$.

Thus,

$$Q_2(k_2, j) = \hat{\delta}_j P_{k_2-1,0} + \frac{(d^j/dz^j) W_2(\lambda_2(1 - z))|_{z=0}}{j!} \sum_{j=k_2}^{\infty} \mu_i, \tag{22}$$

where $\hat{\delta}_j = 1$ if $j = 0$ and 0 otherwise.

To compute $Q_1(i,j)$, we note that type 1 customers do not have to wait if there are less than k_1 customers in the service area at the time of their arrival. Thus,

$$Q_1(i,j) = P(i - 1, j) \qquad \text{for} \quad i < k_1. \tag{23}$$

For $i > k_1$,

$$Q_1(i,j) = \sum_{i'=k_1}^{\infty} \sum_{j'=0}^{j} P(i',j')$$

$$\cdot \int_0^{\infty} \frac{\mu(\mu t)^{i'-k_1}}{(i'-k_1)!} e^{-\mu t} \frac{(\lambda_1 t)^{i-k_1}}{(i-k_1)!} e^{-\lambda_1 t} \frac{(\lambda_2 t)^{j-j'}}{(j-j')!} e^{-\lambda_2 t} dt. \quad (24)$$

For $i = k_1$, there are two possibilities: (1) if the type 1 customer finds $k_1 - 1$ customers in the service area upon arrival and does not have to wait, (2) if it has to wait but no type 1 arrivals occur during its wait. Thus,

$$Q_1(k_1, j) = P(k_1 - 1, j) + \sum_{i'=k_1}^{\infty} \sum_{j'=0}^{j} P(i',j')$$

$$\cdot \int_0^{\infty} \frac{\mu(\mu t)^{i'-k_1}}{(i'-k_1)!} e^{-\mu t} e^{-\lambda_1 t} \frac{(\lambda_2 t)^{j-j'}}{(j-j')!} e^{-\lambda_2 t} dt. \quad (25)$$

In deriving (24) and (25) above, we have used the facts that the waiting time of a type 1 customer (given that it has to wait) has a gamma distribution and that the type 1 and type 2 arrivals that occur during this wait are independent and Poisson.

APPENDIX C

Characterization of the Time Spent in the Service Area by a Tagged Customer in a Two-Class Processor Sharing System

Let $X(s)$ denote the LST of the distribution of the time spent in the service area by an arbitrary "tagged customer." Similarly, let $X_{i,j}(s)$ denote the LST of the conditional distribution of this random variable given that the state $\mathbf{J}$ of the system at the time the tagged customer was admitted to the service area was (i, j). (Here i is assumed to include the tagged customer.) Let $\mathbf{X}_i(s)$ denote the row vector with entries

$$\{\mathbf{X}_i(s)\}_j = X_{i,j}(s) \quad \text{for} \quad i \geq k_2, \quad j \geq 0. \quad (26)$$

When $i < k_2$, there can be no type 2 customers waiting outside, that is, $j = 0$. So we let $X_i(s)$ denote the quantity $\mathbf{X}_i(s)$. We are interested in determining the quantities $\mathbf{X}_i(s)$ for $i \geq k_2$ and $X_i(s)$ for $1 \leq i < k_2$.

Assume that the tagged customer was admitted to the service area at time 0 and let $\tilde{T}$ denote the time at which the tagged customer finishes service and quits the system. Then,

$$\{\mathbf{X}_i(s)\}_j = E[e^{-s\tilde{T}} \,|\, \mathbf{J}_{0^+} = (i, j)] \quad \text{for} \quad i \geq k_2, \quad j = 0, 1, \cdots,$$

and

$$X_i(s) = E[e^{-s\tilde{T}} \,|\, \mathbf{J}_{0^+} = (i, 0)] \quad \text{for} \quad i \leq i \leq k_2 - 1. \quad (27)$$

Let T_i denote the first passage time (after time 0) into the state $(i, .)$, that is,

$$T_i = \mathrm{Min}\{t{:}t \geq 0, \mathbf{J}_t = (i, .)\} \quad \text{for} \quad i \geq 1. \tag{28}$$

Also, for $i \geq k_2$ and $j, k = 0, 1, 2, \cdots$, let ${}_iR_{j,k}(s)$ denote the quantity

$$E[e^{-sT_{i-1}}1\{\tilde{T} > T_{i-1}\}1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \mid \mathbf{J}_{0^+} = (i, k)],$$

where $1\{A\}$ denotes the indicator function of the event A. Observe that for $i \geq m$ the server continues to work at the maximum multiprogramming level until the first passage time into the state $(i-1, \cdot)$ so that, sample path by sample path, the above expectation is independent of i. Thus,

$${}_iR_{j,k}(s) = R_{j,k}(s) \quad \text{for} \quad i \geq m. \tag{29}$$

Also, during this time no type 2 jobs are admitted to the service area so that j cannot be less than k, that is,

$$R_{j,k}(s) = 0 \quad \text{for} \quad j < k. \tag{30}$$

Let $R(s)$ denote the matrix with entries

$$\{R(s)\}_{j,k} = R_{j,k}(s) \quad \text{for} \quad j, k = 0, 1, 2, \cdots. \tag{31}$$

Then $R(s)$, clearly, is lower triangular. Moreover, by a sample path argument it can be shown that $R_{j,k}(s)$ depends upon the difference, $j-k$, which represents the number of type 2 arrivals during $(0, T_{i-1})$, and not on k, which refers to the number of type 2 customers waiting outside at time 0^+. Thus $R(s)$ has the form

$$R(s) = \begin{bmatrix} r_0(s) & & & \\ r_1(s) & r_0(s) & & 0 \\ r_2(s) & r_1(s) & r_0(s) & \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{32}$$

C.1 A matrix equation for R(s)

By conditioning on the first event to occur, we write $R_{j,k}(s)$ as

$$\begin{aligned} R_{j,k}(s) = {} & \frac{\mu(m-1)/m}{\Lambda + \mu + s}\,\delta_{j,k} + \frac{\Lambda_1}{\Lambda + \mu + s}\,E[e^{-sT_{i-1}}1\{\tilde{T} > T_{i-1}\} \\ & \cdot 1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \mid \mathbf{J}_{0^+} = (i+1, k)] + \frac{\lambda_2}{\Lambda + \mu + s}\,E[e^{-sT_{i-1}} \\ & \cdot 1\{\tilde{T} > T_{i-1}\}1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \mid \mathbf{J}_{0^+} = (i, k+1)], \end{aligned} \tag{33}$$

where the first term corresponds to the event "departure of a nontagged customer," the second to the event "arrival of a type 1 customer," and the third to the event "arrival of a type 2 customer." Now the expectation in the second term can be written as

$$E[e^{-sT_{i-1}}1\{\tilde{T} > T_{i-1}\}1\{\mathbf{J}_{T_{i-1}^{+}} = (i-1, j)\} \,|\, \mathbf{J}_{0^+} = (i+1, k)]$$

$$= \sum_{j'=0}^{\infty} E[e^{-s(T'+T'')}1\{\tilde{T} > T' + T''\}1\{\mathbf{J}_{T'+T''} = (i-1, j)\}$$

$$\cdot 1\{\mathbf{J}_{T'} = (i, j')\} \,|\, \mathbf{J}_{0^+} = (i+1, k)], \quad (34)$$

where T' is the first passage time into the state $(i, \cdot)$ and T'' is the time elapsed since the first passage into $(i, \cdot)$ until the first passage into the state $(i-1, \cdot)$. Clearly, $T_{i-1} = T' + T''$ since the Markov process $\mathbf{J}$ is skip-free. From the Markov property of $\mathbf{J}_{T'}$ it follows that the right-hand size of (34) equals

$$\sum_{j'=0}^{\infty} E[e^{-sT''}1\{\tilde{T} - T' > T''\}1\{\mathbf{J}_{(T'+T'')^+} = (i-1, j)\} \,|\, \mathbf{J}_{T'^+}$$

$$= (i, j'), \tilde{T} > T'] \times E[e^{-sT'}1\{\tilde{T} > T'\}1\{\mathbf{J}_{T'^+} = (i, j')\} \,|\, \mathbf{J}_{0^+}$$

$$= (i+1, k)],$$

which is nothing but

$$\sum_{j'=0}^{\infty} R_{j,j'}(s)R_{j',k}(s) = \{R^2(s)\}_{j,k}.$$

Thus, (34) can be written as

$$R_{j,k}(s) = \frac{\mu(m-1)/m}{\Lambda + \mu + s}\,\delta_{j,k} + \frac{\lambda_1}{\Lambda + \mu + s}\,\{R^2(s)\}_{j,k}$$

$$+ \frac{\lambda_2}{\Lambda + \mu + s}\,R_{j,k+1}(s) \quad (35)$$

or in a matrix form

$$R(s) = \frac{\mu((m-1)/m)}{\Lambda + \mu + s}\,I + \frac{\lambda_1}{\Lambda + \mu + s}\,R^2(s) + \frac{\lambda_2}{\Lambda + \mu + s}\,R(s)\Delta. \quad (36)$$

In (35) and (36), $\delta_{j,k}$ is the Kronecker δ, that is, $\delta_{j,k}$ equals 1 if $j = k$ and 0 otherwise; and the matrix Δ is given by

$$\Delta = \begin{bmatrix} 0 & & & \\ 1 & 0 & & 0 \\ & 1 & 0 & \\ 0 & & 1 & 0 \\ & & \cdot & \cdot & \cdot \\ & & \cdot & & \cdot \end{bmatrix}. \quad (37)$$

The structure of $R(s)$ as given in (32) makes it possible to compute the entries $r_i(s)$ recursively. $r_0(s)$ satisfies the quadratic equation

$$r_0(s) = \frac{\mu(m-1)/m}{\Lambda+\mu+s} + \frac{\lambda_1}{\Lambda+\mu+s}\, r_0^2(s) \tag{38}$$

so that $r_0(s)$ is given by

$$r_0(s) = \frac{(\Lambda+\mu+s) \pm \sqrt{(\Lambda+\mu+s)^2 - 4\lambda_1\mu(m-1)/m}}{2\lambda_1}. \tag{39}$$

Since, for $s \geq 0$, $|r_0(s)|$ must be less than or equal to 1, the larger of the two roots is unacceptable. Thus,

$$r_0(s) = \frac{(\Lambda+\mu+s) - \sqrt{(\Lambda+\mu+s)^2 - 4\lambda_1\mu(m-1)/m}}{2\lambda_1}. \tag{40}$$

Once $r_0(s)$ is known, $r_i(s)$, for $i \geq 1$, can be computed recursively since $r_i(s)$ is expressible in terms of $r_0(s), \cdots, r_{i-1}(s)$. We note here that $r_0(s)$ has a form similar to that of $\sigma_2(s)$ in Ref. 10. In fact, if $\lambda_2 = 0$, then $r_0(s)$ and $\sigma_2(s)$ are identical and represent the same quantity.

C.2 The structure of $\mathbf{X}_i(s)$

Now that we have an explicit expression for $R(s)$, we shall attempt to express the quantities $\mathbf{X}_i(s)$ in terms of $R(s)$.

The kth entry of $\mathbf{X}_i(s)$ can be written as

$$\{\mathbf{X}_i(s)\}_k = E[e^{-s\tilde{T}}1\{\tilde{T} > T_{i-1}\} \,|\, \mathbf{J}_{0^+} = (i, k)] \\ + E[e^{-s\tilde{T}}1\{\tilde{T} \leq T_{i-1}\} \,|\, \mathbf{J}_{0^+} = (i, k)]. \tag{41}$$

The first term on the right-hand side of (41) can be expanded as

$$E[e^{-s\tilde{T}}1\{\tilde{T} > T_{i-1}\} \,|\, \mathbf{J}_{0^+} = (i, k)] = \sum_{j=0}^{\infty} E[e^{-s[(\tilde{T}-T_{i-1})+T_{i-1}]} \\ \cdot 1\{\tilde{T} > T_{i-1}\}1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \,|\, \mathbf{J}_{0^+} = (i, k)], \tag{42}$$

which, because of the Markov property of $\mathbf{J}_t$, reduces to

$$\sum_{j=0}^{\infty} E[e^{-s(\tilde{T}-T_{i-1})} \,|\, \mathbf{J}_{T_{i-1}^+} = (i-1, j),\ \tilde{T} > T_{i-1}]E[e^{-sT_{i-1}} \\ \cdot 1\{\tilde{T} > T_{i-1}\}1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \,|\, \mathbf{J}_{0^+} = (i, k)].$$

It follows from the memoryless property of the tagged customer's service-time distribution that $E[e^{-s(\tilde{T}-T_{i-1})} \,|\, \mathbf{J}_{T_{i-1}^+} = (i-1, j),\ \tilde{T} > T_{i-1}]$ equals $\{\mathbf{X}_{i-1}(s)\}_j$; and, for $i \geq m$, $E[e^{-sT_{i-1}}1\{\tilde{T} > T_{i-1}\}1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \,|\, \mathbf{J}_{0^+} = (i, k)]$ equals $\{R(s)\}_{j,k}$. Thus (42) reduces to

$$E[e^{-s\tilde{T}}1\{\tilde{T} > T_{i-1}\} \,|\, \mathbf{J}_{0^+} = (i, k)] = \{\mathbf{X}_{i-1}(s)R(s)\}_k \quad \text{for} \quad i \geq m. \tag{43}$$

To derive an expression for the second term in (41), we note that, for $i \geq m$, no sample path that figures in the expectation

$E[e^{-s\tilde{T}}1\{\tilde{T} \le T_{i-1}\} \mid \mathbf{J}_{0^+} = (i, k)]$ allows the multiprogramming level to fall below m before the departure of the tagged customer. Thus the server continues to operate at rate μ until the tagged customer's departure. Also, for any two initial states (i_1, k_1) and (i_2, k_2), as long as $i_1, i_2 \ge m$, there is a one-to-one correspondence between sample paths describing the trajectory of $\mathbf{J}_t$ between 0 and $\tilde{T}$, which make equal contributions to the expectation. Thus,

$$E[e^{-s\tilde{T}}1\{\tilde{T} \le T_{i-1}\} \mid \mathbf{J}_{0^+} = (i, k)] = a(s) \quad \text{for} \quad i \ge m, \tag{44}$$

where $a(s)$ is independent of i and k.

To derive an expression for $a(s)$, it will be convenient to introduce a quantity $\sigma(s)$ defined by

$$\sigma(s) = E[e^{-sT_{i-1}}1\{\tilde{T} > T_{i-1}\} \mid \mathbf{J}_{0^+} = (i, k)] \qquad \text{for} \quad i \ge m, \quad k = 0, 1, \cdots. \tag{45}$$

Note that although the definition of $\sigma(s)$ involves the initial state (i, k), $\sigma(s)$ is independent of the latter as long as $i \ge m$. Also note that

$$\sigma(s) = \sum_{k=0}^{\infty} r_k(s). \tag{46}$$

By conditioning on the first relevant event to occur, $\sigma(s)$ can be written as

$$\sigma(s) = \frac{\mu(m-1)/m}{\lambda_1 + \mu + s} + \frac{\lambda_1}{\lambda_1 + \mu + s} \cdot E[e^{-s(T_{i-1}-T_{i+1})}1\{\tilde{T} > T_{i-1}\} \mid \mathbf{J}_{T_{i+1}} = (i+1, \cdot), T_{i-1} > T_{i+1}]. \tag{47}$$

In (47) it can be seen that arrivals of type 2 customers are completely ignored since they do not affect the mechanics involved. Since the first passage time from the state $(i+1, \cdot)$ to $(i-1, \cdot)$ involves the sum of the first passage times from $(i+1, \cdot)$ to $(i, \cdot)$ and from $(i, \cdot)$ to $(i-1, \cdot)$, which are i.i.d., (47) reduces to

$$\sigma(s) = \frac{\mu(m-1)/m}{\lambda_1 + \mu + s} + \frac{\lambda_1}{\lambda_1 + \mu + s}\,\sigma^2(s). \tag{48}$$

Equation (48) is identical to the one describing $\sigma_2(s)$ in Ref. 10, with λ replaced by λ_1. The desired root of (48) is the smaller of the two roots so that

$$\sigma(s) = \frac{(\lambda_1 + \mu + s) - \sqrt{(\lambda_1 + \mu + s)^2 - 4\mu\lambda_1(m-1)/m}}{2\lambda_1}. \tag{49}$$

Now $a(s)$ can be obtained in a straightforward manner from $\sigma(s)$. By conditioning on the first relevant event to occur, we write

$$a(s) = \frac{\mu/m}{\mu + \lambda_1 + s} + \frac{\lambda_1}{\mu + \lambda_1 + s} E[e^{-s(\tilde{T}-T_{i+1})} 1\{\tilde{T} \le T_{i-1}\} \,|\, \mathbf{J}_{T_{i+1}^+}$$

$$= (i + 1, \cdot), \tilde{T} > T_{i+1}] = \frac{\mu/m}{\mu + \lambda_1 + s} + \frac{\lambda_1}{\mu + \lambda_1 + s}$$

$$\cdot [E[e^{-s(\tilde{T}-T_{i+1})} 1\{\tilde{T} \le T_{i-1}\} 1\{\tilde{T} \le T'\} \,|\, \mathbf{J}_{T_{i+1}^+} = (i + 1, \cdot), \tilde{T} > T_{i+1}]$$

$$+ E[e^{-s(\tilde{T}-T_{i+1})} 1\{\tilde{T} \le T_{i-1}\} 1\{\tilde{T} > T'\} \,|\, \mathbf{J}_{T_{i+1}^+} = (i + 1, \cdot), \tilde{T} > T_{i+1}]], \quad (50)$$

where T' denotes the first passage time into the state $(i, \cdot)$ after the state has reached $(i + 1, \cdot)$ at time T_{i+1}. Following arguments similar to the ones used earlier, it can be shown that (50) can be written as

$$a(s) = \frac{\mu/m}{\mu + \lambda_1 + s} + \frac{\lambda_1}{\mu + \lambda_1 + s} [a(s) + \sigma(s) a(s)],$$

that is,

$$a(s) = \mu/m[\mu + s - \lambda_1 \sigma(s)]^{-1}. \quad (51)$$

The desired vector $\mathbf{X}_i(s)$ can now be written as

$$\mathbf{X}_i(s) = \mathbf{a}(s) + \mathbf{X}_{i-1}(s) R(s) \quad \text{for} \quad i \ge m, \quad (52)$$

where $\mathbf{a}(s)$ is the row vector $(a(s), a(s), a(s), \cdots)$.

If we introduce two more vectors $\alpha(s)$ and $\beta(s)$, where

$$\beta(s) = \frac{\mu/m}{\mu/m + s} [1, 1, 1, \cdots], \quad (53)$$

and

$$\alpha(s) = \mathbf{a}(s) - \beta(s) + \mathbf{X}_{m-1}(s) R(s), \quad (54)$$

the vectors $\mathbf{X}_i(s)$ for $i \ge m$ can be expressed as

$$\mathbf{X}_i(s) = \beta(s) + \alpha(s)[R(s)]^{i-m}. \quad (55)$$

(The above equation can be proved by mathematical induction.)

It can be seen that as $i \rightarrow \infty$, the term $\alpha(s)[R(s)]^{i-m}$ vanishes, so that $\mathbf{X}_i(s)$ approaches its limiting value $\beta(s)$. In other words, when i is large, the tagged customer receives its entire service at the rate μ/m as expected.

C.3 Boundary conditions

For $i \ge m$, eq. (55) characterizes the vectors $\mathbf{X}_i(s)$ in terms of known quantities $\beta(s)$, $a(s)$, $R(s)$, and the unknown $\mathbf{X}_{m-1}(s)$. In other words, if $\mathbf{X}_{m-1}(s)$ is known, $\mathbf{X}_i(s)$ can be obtained, for $i \ge m$, directly from (55). To completely characterize the time spent in the service area by a tagged customer, it remains to derive the boundary conditions, that is, a system of equations from which the quantities $X_1(s)$, $X_2(s)$, $\cdots$, $X_{k_1-1}(s)$, $\mathbf{X}_{k_1}(s)$, $\cdots$, $\mathbf{X}_{m-1}(s)$ can be obtained.

Without going into the details of derivation—it involves arguments similar to the ones used in the earlier analysis—we state the boundary conditions, which are as follows:

$$\mathbf{X}_i(s) = \frac{\mu_i/i}{\Lambda + \mu_i + s}\,\mathbf{1} + \frac{(i-1)\mu_i/i}{\Lambda + \mu_i + s}\,\mathbf{X}_{i-1}(s) + \frac{\lambda_1}{\Lambda + \mu_i + s}\,\mathbf{X}_{i+1}(s)$$

$$+ \frac{\lambda_2}{\Lambda + \mu_i + s}\,\mathbf{X}_i(s)\Delta \quad \text{for} \quad k_2 + 1 \le i \le m - 1, \tag{56}$$

$$\mathbf{X}_{k_2}(s) = \frac{\mu_{k_2}/k_2}{\Lambda + \mu_{k_2} + s}\,\mathbf{1} + \frac{(k_2 - 1)\mu_{k_2}/k_2}{\Lambda + \mu_{k_2} + s}\,\{\mathbf{X}_{k_2}(s)\Delta^T + X_{k_2-1}(s)\mathbf{e}_0\}$$

$$+ \frac{\lambda_1}{\Lambda + \mu_{k_2} + s}\,\mathbf{X}_{k_2+1}(s) + \frac{\lambda_2}{\Lambda + \mu_{k_2} + s}\,\mathbf{X}_{k_2}(s)\Delta, \tag{57}$$

$$X_{k_2-1}(s) = \frac{\mu_{k_2-1}/(k_2 - 1)}{\Lambda + \mu_{k_2-1} + s} + \frac{(k_2 - 2)\mu_{k_2-1}/(k_2 - 1)}{\Lambda + \mu_{k_2-1} + s}\,X_{k_2-2}(s)$$

$$+ \frac{\Lambda}{\Lambda + \mu_{k_2-1}}\,\{\mathbf{X}_{k_2}(s)\}_0, \tag{58}$$

$$X_i(s) = \frac{\mu_i/i}{\Lambda + \mu_i + s} + \frac{(i-1)\mu_i/i}{\Lambda + \mu_i + s}\,X_{i-1}(s) + \frac{\Lambda}{\Lambda + \mu_i + s}\,X_{i+1}(s)$$

$$\text{for} \quad 2 \le i < k_2 - 1, \tag{59}$$

and

$$X_1(s) = \frac{\mu_1}{\Lambda + \mu_1 + s} + \frac{\Lambda}{\Lambda + \mu_1 + s}\,X_2(s), \tag{60}$$

where $\mathbf{e}_0 = [1\ 0\ 0\ 0\ \cdots]$ and $\mathbf{1} = [1\ 1\ 1\ \cdots]$.

Equations (55) through (60) give a characterization of the LST of the distribution of the time spent in the service area by the tagged customer given the state of the system at the time it was admitted to the service area. To derive the mean time spent in the service area, we need to differentiate these quantities at $s = 0$. Noting that $\boldsymbol{\alpha}(0) = \mathbf{0}$, we have

$$\mathbf{X}_i'(0) = \boldsymbol{\beta}'(0) + \boldsymbol{\alpha}'(0)[R(0)]^{i-m} \quad \text{for} \quad i \ge m. \tag{61}$$

The boundary conditions also are obtained by differentiating the corresponding equations at $s = 0$:

$$\mathbf{X}_i'(0) = -\frac{1}{(\Lambda + \mu_i)}\,\mathbf{1} + \frac{(i-1)\mu_i/i}{(\Lambda + \mu_i)}\,\mathbf{X}_{i-1}'(0) + \frac{\lambda_1}{\Lambda + \mu_i}\,\mathbf{X}_{i+1}'(0)$$

$$+ \frac{\lambda_2}{\Lambda + \mu_i}\,\mathbf{X}_i'(0)\Delta \quad \text{for} \quad k_2 + 1 \le i < m - 1 \qquad (62)$$

$$\mathbf{X}_{k_2}'(0) = -\frac{1}{\Lambda + \mu_{k_2}} + \frac{(k_2 - 1)\mu_{k_2}/k_2}{\Lambda + \mu_{k_2}}\,[\mathbf{X}_{k_2}'(0)\Delta^T + X_{k_2-1}'(0)\mathbf{e}_0]$$

$$+ \frac{\lambda_1}{\Lambda + \mu_{k_2}}\,\mathbf{X}_{k_2+1}'(0) + \frac{\lambda_2}{\Lambda + \mu_{k_2}}\,\mathbf{X}_{k_2}'(0)\Delta, \qquad (63)$$

$$X_{k_2-1}'(0) = -\frac{1}{\Lambda + \mu_{k_2-1}} + \frac{(k_2 - 2)\mu_{k_2-1}/(k_2 - 1)}{\Lambda + \mu_{k_2-1}}\,X_{k_2-2}'(0)$$

$$+ \frac{\Lambda}{\Lambda + \mu_{k_2-1}}\,\{\mathbf{X}_{k_2}'(0)\}_0, \qquad (64)$$

$$X_i'(0) = -\frac{1}{\Lambda + \mu_i} + \frac{(i-1)\mu_i/i}{\Lambda + \mu_i}\,X_{i-1}'(0) + \frac{\Lambda}{\Lambda + \mu_i}\,X_{i+1}'(0)$$

$$\text{for} \quad 2 \le i < k_2 - 1, \qquad (65)$$

and

$$X_1'(0) = -\frac{1}{\Lambda + \mu_1} + \frac{\Lambda}{\Lambda + \mu_1}\,X_2'(0). \qquad (66)$$

AUTHORS

Kiran M. Rege, B. Tech. (Electrical Engineering), 1977, I.I.T., Bombay; Ph.D. (Electrical Engineering), 1981, University of Hawaii; AT&T Bell Laboratories, 1982—. Mr. Rege spent 1984 on leave of absence, teaching at I.I.T. Bombay in the Department of Electrical Engineering. His technical work at AT&T Bell Laboratories includes modeling and analysis of switching, computer, and communication systems. His research interests include communication theory, queueing theory, and performance analysis of computer and communication systems.

Bhaskar Sengupta, B. Tech. (Electrical Engineering), 1965, I.I.T. Kharagpur, Eng. Sc.D. (Operations Research), 1976, Columbia University; AT&T Bell Laboratories, 1981—. Mr. Sengupta has worked in IBM and Service Bureau Company and was an Assistant Professor at the State University of New York at Stony Brook. He was also a consultant to Turner Construction Company in New York. At AT&T Bell Laboratories he works on performance problems for communication, computer, and manufacturing systems.

AT&T Technical Journal
Vol. 64, No. 7, September 1985
Printed in U.S.A.

Laplace Transform Inequalities With Application to Queueing

By D. L. JAGERMAN*

(Manuscript received April 11, 1985)

Inequalities satisfied by the Laplace transforms of convex and log-convex functions are obtained. Applications are made to the M/G/1 queue waiting time and to an important teletraffic congestion problem, arising in parcel blocking studies.

I. INTRODUCTION

The purpose of this paper is to establish certain inequalities satisfied by the Laplace transform of convex functions and to illustrate their use. The notion of α-convexity on which these results are based is fully discussed.[1] This notion had its origin in the author's investigations concerning the inversion of the Laplace transform;[2] subsequently, it has been applied to obtaining the results reported on here. The property of α-convexity forms a natural bridge between ordinary convexity ($\alpha = 0$) and the stronger property of log-convexity (all α). This enables the formulation of a criterion in terms of the transform of a function for ascertaining the log-convexity of the function, vd. Theorem 2 and (11).

An infinite set of inequalities satisfied by the Laplace transform of a log-convex function is obtained, vd. Theorem 3; these inequalities are illustrated by two applications. The first provides necessary conditions for the log-convexity of the complementary waiting time distribution in the First-In First-Out (FIFO) M/G/1 queue. The condi-

* AT&T Bell Laboratories.

tions are expressed in terms of the transform of the complementary service time distribution.

The second application concerns the important teletraffic problem of ascertaining the time congestion of a call that overflows a primary group and is offered to a secondary group. Simple upper and lower bounds are obtained for a function $[\delta_j(x, a)]$ arising in Brockmeyer's analysis of the problem and in terms of which the time congestion is obtained.[3] These results form an important part of "parcel blocking" analyses.[4]

II. α-CONVEXITY

A function $f(x)$ is said to be α-convex on an interval I if $e^{\alpha x}f(x)$ is convex on I. Clearly, ordinary convexity corresponds to $\alpha = 0$. A sufficient condition for convexity of $f(x)$ is[5]

$$f''(x) \geq 0, \quad x \in I. \tag{1}$$

Introducing the function $h(x)$ by

$$h(x) = e^{-\alpha x} \frac{d^2}{dx^2} [e^{\alpha x}f(x)], \tag{2}$$

$$= f''(x) + 2\alpha f'(x) + \alpha^2 f(x). \tag{3}$$

Then, in view of (1), the condition for α-convexity is

$$h(x) \geq 0, \quad x \in I. \tag{4}$$

It should be observed that α-convexity does not imply convexity ($\alpha = 0$). Thus consider, for example, $f(x) = x^3$, which is α-convex for $x \geq 0$, $\alpha \geq 0$. For $\alpha = 1$, however, x^3 is α-convex for $-3 - \sqrt{3} \leq x \leq -3 + \sqrt{3}$.

The α-convexity of a function may permit stronger bounds to be obtained on integrals of the function than ordinary convexity. For example, let $p(x) \geq 0$, and let $f(x)$ be convex on I, then Jensen's inequality states[5]

$$\int_I f(x)p(x)dx \geq f(\mu) \int_I p(x)dx,$$

$$\mu = \int_I xp(x)dx \Big/ \int_I p(x)dx. \tag{5}$$

If $f(x)$ is α-convex on I, then, since

$$\int_I f(x)p(x)dx = \int_I e^{\alpha x}f(x)e^{-\alpha x}p(x)dx, \tag{6}$$

one has

$$\int_I f(x)p(x)dx \geq e^{\alpha\mu}f(\mu) \int_I e^{-\alpha x}p(x)dx,$$

$$\mu = \int_I xe^{-\alpha x}p(x)dx \bigg/ \int_I e^{-\alpha x}p(x)dx. \quad (7)$$

This result can be stronger than (5).

Let the Laplace transform, $\tilde{f}(s)$, of $f(x)$ be defined by

$$\tilde{f}(s) = \int_0^\infty e^{-sx}f(x)dx, \quad s > -\gamma, \quad (8)$$

and let $f(x)$ be α-convex for $x \geq 0$; then the following theorem may be stated.

Theorem 1:

$$\tilde{f}(s) \geq \frac{1}{s+\alpha}\, e^{\frac{\alpha}{s+\alpha}} f\left(\frac{1}{s+\alpha}\right), \quad \alpha > -s;$$

or, equivalently,

$$f(x) \leq \frac{1}{x}\, e^{-\alpha x}\tilde{f}\left(\frac{1}{x} - \alpha\right), \quad \alpha < \frac{1}{x} + \gamma.$$

Proof: Use of (7) with $p(x) = e^{-sx}$.

A function $f(x) > 0$ is said to be log-convex on I if $\ln f(x)$ is convex on I. Thus, the condition for log-convexity is

$$f''(x)f(x) - f'(x)^2 \geq 0, \quad x \in I. \quad (9)$$

In particular, log-convexity implies convexity, hence $e^{\alpha x}f(x)$ is convex; thus, a log-convex function is α-convex for all α. The following theorem asserts also the converse.

Theorem 2: A function $f(x) > 0$ is log-convex on an interval I if and only if it is α-convex on I for all α.

Proof: It is necessary to prove only that $f(x)$ is log-convex on I if it is α-convex on I for all α. This follows from (3) on observing that the discriminant of the quadratic in α is $f'(x)^2 - f''(x)f(x)$; hence, α-convexity for all α implies (9) and the consequent log-convexity of $f(x)$.

Corollary: The sum of log-convex functions is log-convex.

Proof: The sum of α-convex functions corresponding to the same α is clearly again α-convex for the same α; hence, the statement follows on applying the theorem.

A function $f(x)$ is said to be completely monotone on I if

$$(-1)^r f^{(r)}(x) \geq 0, \quad x \in I, \quad r = 0, 1, 2, \cdots . \tag{10}$$

The Bernstein theorem,[6] which states that $f(x) \geq 0$ if and only if $\tilde{f}(s)$ is completely monotone for s real and in the domain of convergence of (8), may be used to translate condition (4) in terms of $\tilde{f}(s)$. Accordingly, let $\tilde{f}(s)$ converge for $s > 0$ and let

$$\tilde{h}(s) = (s + \alpha)^2 \tilde{f}(s) - (s + 2\alpha) f(0+) - f'(0+), \quad s > 0. \tag{11}$$

Then $f(x)$ is α-convex if and only if $\tilde{h}(s)$ is completely monotone for $s > 0$. Thus, also, $f(x)$ is log-convex if and only if $\tilde{h}(s)$ is completely monotone for $s > 0$ and all α.

III. INEQUALITIES FOR $\tilde{f}(s)$ FROM LOG-CONVEXITY

It will now be assumed that $f(x)$ is log-convex for $x > 0$ and that $\tilde{f}(s)$ converges for $s > 0$. Thus, one has

$$\min_{\alpha} \, (-1)^n \tilde{h}^{(n)}(s) \geq 0, \quad s > 0, \quad n = 0, 1, 2, \cdots . \tag{12}$$

The following theorem will now be proved.

Theorem 3: If $f(x)$ is log-convex for $x > 0$, then for all $s > 0$, one has

$$\tilde{f}(s)^{-1} \leq \frac{sf(0+) - f'(0+)}{f(0+)^2}, \quad f(0+) \neq 0,$$

$$\frac{d}{ds}\, \tilde{f}(s)^{-1} \geq \frac{1}{f(0+)}, \quad f(0+) \neq 0,$$

$$(n - 1)\tilde{f}^{(m)}(s)\tilde{f}^{(m-2)}(s) \geq n\tilde{f}^{(m-1)}(s)^2, \quad n \geq 2.$$

The equality signs are achieved for $f(x) = e^{-\gamma x}$.

Proof: One has for

$$\alpha = \frac{f(0)}{\tilde{f}(s)} - s, \tag{13}$$

$$\min_{\alpha} \tilde{h}(s) = sf(0) - f'(0) - \frac{f(0)^2}{\tilde{f}(s)} \geq 0. \tag{14}$$

This establishes the first inequality. For

$$\alpha = \frac{\tilde{f}(s)}{\tilde{f}'(s)} - s, \tag{15}$$

one has

$$\min_{\alpha}[-\tilde{h}'(s)] = \frac{\tilde{f}(s)^2}{\tilde{f}'(s)} + f(0) \geq 0, \tag{16}$$

which, after a little manipulation, yields the second inequality. For the choice

$$\alpha = -n \frac{\tilde{f}^{(n-1)}(R)}{\tilde{f}^{(n)}(R)} - R, \tag{17}$$

direct calculation shows that

$$\min_{\alpha}[(-1)^n \tilde{h}^{(n)}(s)] = (-1)^n n \left[(n-1)\tilde{f}^{(n-2)}(s) - n \frac{\tilde{f}^{(n-1)}(s)^2}{\tilde{f}^{(n)}(s)} \right] \geq 0. \tag{18}$$

In view of the complete monotonicity of $\tilde{f}(s)$, the remaining inequalities are established.

IV. CONVEXITY IN M/G/1

An application of Theorem 3 will now be made to waiting time in the FIFO M/G/1 queue.[1] In the following, λ is the arrival rate, μ is the service rate, $\rho = \lambda/\mu < 1$, μ_2 is the second moment about the origin of service time, and $\tilde{\beta}(s)$, $\tilde{F}(s)$ are the Laplace transforms of the complementary service and waiting time distributions, respectively. Theorem 4 may now be stated.

Theorem 4: Necessary conditions for the complementary waiting time distribution, $F(x)$, to be log-convex are

$$\frac{2}{\mu^2 \mu_2} \frac{1}{s + \dfrac{2}{\mu\mu_2}} \leq \tilde{\beta}(s) \leq \frac{1}{s+\mu}, \qquad \mu_2 \geq \frac{2}{\mu^2}.$$

Proof: The Pollaczek-Khintchine formula[1] may be written in the form

$$\tilde{F}(s) = \frac{1}{s} \frac{\rho - \lambda\tilde{\beta}(s)}{1 - \lambda\tilde{\beta}(s)}. \tag{19}$$

Clearly,

$$\tilde{\beta}(s) \sim \frac{1}{s}, \qquad s \rightarrow \infty; \tag{20}$$

hence,

$$\tilde{F}(s) \sim \frac{\rho}{s} - \frac{\lambda(1-\rho)}{s^2}, \qquad s \rightarrow \infty. \tag{21}$$

Thus,

$$F(0+) = \rho, \qquad F'(0+) = -\lambda(1-\rho). \tag{22}$$

Application of the first inequality of Theorem 3 yields

$$s\,\frac{1-\lambda\tilde{\beta}(s)}{\rho-\lambda\tilde{\beta}(s)} \le \frac{s}{\rho} + \mu\,\frac{1-\rho}{\rho}. \tag{23}$$

Observing that $\tilde{\beta}(s)$ is monotone decreasing, and that $\lambda\tilde{\beta}(0) = \rho < 1$, one has $(s > 0)$

$$1 - \lambda\tilde{\beta}(s) > 0, \qquad \rho - \lambda\tilde{\beta}(s) > 0. \tag{24}$$

Hence, multiplying (23) through by $\rho - \lambda\tilde{\beta}(s)$, and solving for $\tilde{\beta}(s)$, one obtains

$$\tilde{\beta}(s) \le \frac{1}{s+\mu}. \tag{25}$$

Using the second inequality of Theorem 3 in the form

$$\tilde{F}(s)^{-1} \ge \tilde{F}(0)^{-1} + \frac{s}{F(0+)} \tag{26}$$

with the evaluation

$$\tilde{F}(0) = \frac{1}{2}\,\frac{\lambda\mu_2}{1-\rho} \tag{27}$$

yields the required lower bound for $\tilde{\beta}(s)$. Finally, multiplying the upper and lower bounds by s and evaluating the limit $s \to \infty$ provides the last inequality of the theorem.

V. AN OVERFLOW MODEL

Consider the traffic model of Fig. 1, in which a is the offered load, assumed Poisson, to the primary trunk group of x trunks whose overflow of m erlangs and peakedness z is offered to the secondary trunk group of c trunks. Clearly,

$$m = aB(x, a), \tag{28}$$

in which $B(x, a)$ is the Erlang loss function.[7,8] The blocking (call congestion) experienced by a call overflowing the primary is given by the formula (equivalent random method)[8]

$$B_c = \frac{B(x+c, a)}{B(x, a)}. \tag{29}$$

An alternative expression for B_c may be obtained from the Brockmeyer analysis.[3] Let $P(k)$ be the probability k trunks are busy on the secondary group and let $m(k)$ be the corresponding load offered to the secondary; then

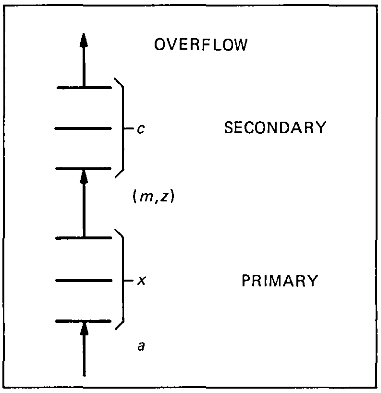


Fig. 1—Overflow system.

$$m = \sum_{k=0}^{c} m(k)P(k), \tag{30}$$

$$B_c = \frac{m(c)P(c)}{m}. \tag{31}$$

One has also that the time congestion, B_T, is given by

$$B_T = P(c). \tag{32}$$

In order to obtain $P(c)$, the function $\delta_j(x, a)$ is introduced by

$$\delta_j(x, a) = \frac{G_x(-j-1, a)}{G_x(-j, a)}, \tag{33}$$

in which $G_j(x, a)$ are Poisson-Charlier polynomials.[7] Then

$$B_T = \delta_c(x, a)B(x + c, a). \tag{34}$$

It has been found[4] that $\delta_j(x, a)$ satisfies the recursion

$$\delta_0(x, a) = B(x, a)^{-1},$$

$$\delta_j(x, a) = 1 + \frac{x-a}{j} + \frac{a}{j}\,\delta_{j-1}(x, a)^{-1}, \qquad j \geq 1. \tag{35}$$

Another recursion for $\delta_j(x, a)$ may be obtained by considering the integral

$$I_j(x, a) = \int_0^\infty e^{-ay}(1 + y)^x y^j dy. \tag{36}$$

Since

$$G_x(-j, a) = (-1)^x \frac{a^j}{\Gamma(j)} \int_0^\infty e^{-ay}(1 + y)^x y^{j-1} dy, \tag{37}$$

one has from (33)

$$\delta_j(0, a) = 1, \qquad j \geq 0,$$

$$\delta_j(x, a) = \frac{a}{j}\frac{I_j(x, a)}{I_{j-1}(x, a)}, \qquad j \geq 1. \tag{38}$$

One easily establishes

$$I_j(x + 1, a) = I_j(x, a) + I_{j+1}(x, a), \tag{39}$$

$$I_j(x, a) = \frac{x}{a} I_j(x - 1, a) + \frac{j}{a} I_{j-1}(x, a); \tag{40}$$

hence, from (39) and (40)

$$\delta_j(x, a) = \frac{x}{j}\frac{I_j(x - 1, a)}{I_{j-1}(x, a)} + 1. \tag{41}$$

Use of (39) finally yields

$$\delta_j(x, a) = \frac{x}{a + j\delta_j(x - 1, a)}\,\delta_j(x - 1, a) + 1, \quad x \geq 1. \tag{42}$$

The recursion of (42) with initial value of (38) provides a convenient and stable method for the exact computation of $\delta_j(x, a)$; however, in many practical investigations, it is useful to have upper and lower bounds showing simple and explicit dependence on the arguments. For this purpose, let $P_x(t)$ be the recovery function for a group of x trunks, that is, the probability that all trunks are busy at time t given they were all busy at time zero; then[9]

$$\tilde{P}_x(s) = \frac{G_x(-R, a)}{RG_x(-R - 1, a)}. \tag{43}$$

Comparison of (43) with (33) shows that

$$\tilde{P}_x(j) = \frac{1}{j\delta_j(x, a)}. \tag{44}$$

It is also known that $P_x(t)$ is log-convex in t for $t \geq 0$.[9] The following theorem may now be proved.

Theorem 5: For $j \geq 1$, $x \geq 0$, $a \geq 0$, *the following inequalities hold*

$$\frac{1}{2}\left[1 + \frac{x - a + 1}{j} + \sqrt{\left(1 + \frac{x - a + 1}{j}\right)^2 + 4\,\frac{j(a - 1) - x}{j^2}}\right]$$

$$\leq \delta_j(x, a) \leq \frac{1}{2}\left[1 + \frac{x - a}{j} + \sqrt{\left(1 + \frac{x - a}{j}\right)^2 + \frac{4a}{j}}\right].$$

Proof: The upper bound is due to A. A. Fredericks,[4] who shows that

$$\delta_j(x, a) \leq \delta_{j-1}(x, a), \qquad j \geq 1; \tag{45}$$

hence, from (35) one gets

$$\delta_j(x, a) \leq 1 + \frac{x - a}{j} + \frac{a}{j}\,\delta_j(x, a)^{-1}. \tag{46}$$

Solution of the quadratic provides the upper bound of the theorem.

Since $P_x(0) = 1$, the second inequality of Theorem 3 applied to $\tilde{P}_x(s)$ in the form

$$\tilde{P}_x(s + h)^{-1} - \tilde{P}_x(s)^{-1} \geq h, \qquad h \geq 0, \tag{47}$$

yields, from (44),

$$(j + h)\delta_{j+h}(x, a) - j\delta_j(x, a) \geq h, \qquad h \geq 0. \tag{48}$$

Setting $h = 1$ and writing (48) in the form

$$\delta_{j-1}(x, a)^{-1} \geq \frac{j - 1}{j\delta_j(x, a) - 1} \tag{49}$$

yields, after substitution into (35),

$$j\delta_j(x, a) \geq j + x - a + a\,\frac{j - 1}{j\delta_j(x, a) - 1}. \tag{50}$$

From (48) it follows that $j\delta_j(x, a) \geq 1$; hence, in (50), multiplying through by $j\delta_j(x, a) - 1$ yields the inequality

$$\delta_j(x, a)^2 - \left(1 + \frac{x - a + 1}{j}\right)\delta_j(x, a) - \frac{j(a - 1) - x}{j^2} \geq 0. \tag{51}$$

Solution of this quadratic finally yields the lower bound of the theorem.

REFERENCES

1. D. L. Jagerman, "Waiting Time Convexity in the M/G/1 Queue," AT&T Tech. J., *64,* No. 1, Part 1 (January 1985), pp. 33–41.
2. D. L. Jagerman, "An Inversion Technique for the Laplace Transform With Application to Approximation," B.S.T.J., *57,* No. 3 (March 1978), pp. 669–710.
3. E. Brockmeyer, "The Simple Overflow Problem in the Theory of Telephone Traffic," Teleteknik, *5* (1954), pp. 361–74.
4. A. A. Fredericks, "Approximating Parcel Blocking via State Dependent Birth Rates," Proc. 10th ITC, Montreal, Canada, 1983. (Paper #2, Session #5.3).
5. G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*, Cambridge: Cambridge University Press, 1959.
6. J. A. Shohat and J. D. Tamarkin, "The Problem of Moments," Math. Surveys No. 1, Providence, R.I.: American Mathematical Society, 1943.
7. D. L. Jagerman, "Some Properties of the Erlang Loss Function," B.S.T.J., *53,* No. 3 (March 1974), pp. 525–51.
8. D. L. Jagerman, "Methods in Traffic Calculations," AT&T Bell Lab. Tech. J., *63,* No. 7 (September 1984), pp. 1283–310.
9. D. J. Jagerman, "Nonstationary Blocking in Telephone Traffic," B.S.T.J., *54,* No. 3 (March 1975), pp. 625–61.

AUTHOR

David L. Jagerman, B.E.E., 1949, Cooper Union; M.S., and Ph.D. (Mathematics), 1954 and 1962, respectively, New York University; AT&T Bell Laboratories, 1963—. Mr. Jagerman has been engaged in mathematical research on quadrature, interpolation, and approximation theory, especially related to the theory of widths and metrical entropy, with application to the storage and transmission of information. For the past several years, he has worked on the theory of difference equations and queueing, especially with reference to traffic theory and computers. He is currently preparing a text on difference equations with application to stochastic models. He is also an adjunct professor at Stevens Institute, where he teaches selected topics on mathematics applied to computer science.

PAPERS BY AT&T BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Baker B. S., **A New Proof for the 1st-Fit Decreasing Bin-Packing Algorithm.** J Algorithm 6(1):49–70, Mar 1985.

Barron E. N., **Viscosity Solutions for the Monotone Control Problem.** SIAM J Con 23(2):161–171, Mar 1985.

Bentley J. L., McGeoch C. C., **Amortized Analysis of Self-Organizing Sequential Search Heuristics.** Comm ACM 28(4):404–411, Apr 1985.

Cargill T. A., **Implementation of the Blit Debugger.** Software 15(2):153–168, Feb 1985.

Coffman E. G., Langston M. A., **A Performance Guarantee for the Greedy Set-Partitioning Algorithm.** ACT Inform 21(4):409–415, 1984.

Crawford S. G., McIntosh A. A., Pregibon D., **An Analysis of Static Metrics and Faults in C Software.** J Syst Soft 5(1):37–48, Feb 1985.

Dembo R. S., Klincewicz J. G., **Dealing With Degeneracy in Reduced Gradient Algorithms.** Math Progr 31(3):357–363, Mar 1985.

Fishburn P. C., **SSB Utility Theory and Decision Making Under Uncertainty.** Math Soc Sci 8(3): 253–285, Dec 1984.

Fishburn P. C., **SSB Utility Theory—An Economic Perspective.** Math Soc Sc 8(1):63–94, Aug 1984.

Johnson D. S., **The NP-Completeness Column—An Ongoing Guide.** J Algorithm 6(1):145–159, Mar 1985.

Kesler T. E. et al., **The Effect of Indentation on Program Comprehension.** Int J Man M 21(5):415–428, Nov 1984.

Morrison J. A., Mitra D., **Heavy-Usage Asymptotic Expansions for the Waiting Time in Closed Processor-Sharing Systems With Multiple Classes.** Adv Appl P 17(1):163–185, Mar 1985.

Nozari A., **Control of Entry to a Nonstationary Queuing System.** Nav Res Log 32(2):275–286, May 1985.

Pike R., Locanthi B., Reiser J., **Hardware Software Trade-Offs for Bitmap Graphics on the Blit.** Software 15(2):131–151, Feb 1985.

Prabhu N. U., Reeser P. K., **A Random Family of Queuing Systems With a Dynamic Priority Discipline.** Math Oper R 10(1):24–32, Mar 1985.

Tarjan R. E., **Amortized Computational Complexity.** SIAM J Alg 6(2):306–318, Apr 1985.

Vardi Y., **Empirical Distributions in Selection Bias Models.** Ann Statist 13(1):178–203, Mar 1985.

Vardi Y., Shepp L. A., Kaufman L., **A Statistical Model for Positron Emission Tomography.** J Am Stat A 80(389):8–20, Mar 1985.

Whitt W., **The Best Order for Queues in Series.** Manag Sci 31(4): 475–487, Apr 1985.

Whitt W., **The Renewal-Process Stationary-Excess Operator.** J Appl Prob 22(1):156–167, Mar 1985.

Willard D. E., **Reduced Memory Space for Multidimensional Search Trees.** Lect N Comp 182:363–374, 1985.

ENGINEERING

Agrawal G. P., **Coupled-Cavity Semiconductor Lasers Under Current Modulation—Small-Signal Analysis.** IEEE J Q El 21(3):255–263, Mar 1985.

Auborn J. J., Barberio Y. L., **An Ambient-Temperature Secondary Aluminum Electrode—Its Cycling Rates and Its Cycling Efficiencies.** J Elchem So 132(3):598–601, Mar 1985.

Auston D. H., Cheung K. P., **Coherent Time-Domain Far-Infrared Spectroscopy.** J Opt Soc B 2(4):606–612, Apr 1985.

Benes V. E., **New Exact Nonlinear Filters With Large Lie Algebras.** Syst Contr 5(4):217–221, Feb 1985.

Benvenuto N., **Moments of Error-Frequency Response Due to Coefficient Inaccuracy for Sampled Data Filters (Letter).** IEEE Acoust 33(2):436–437, Apr 1985.

Bowers J. E., Hemenway B. R., Wilt D. P., **Etching of Deep Grooves for the Precise Positioning of Cleaves in Semiconductor Lasers.** Appl Phys L 46(5):453–455, Mar 1 1985.

Bowers J. E., Koch T. L., Hemenway B. R., Wilt D. P., Bridges T. J., Burkhardt E. G., **High-Frequency Modulation of 1.52-μm Vapor-Phase-Transported InGaAsP Lasers.** Electr Lett 21(7):297–299, Mar 28 1985.

Broer M. M., Golding B., **Low-Temperature Optical Dephasing of Rare-Earth Ions by Tunneling Systems in Glass.** J Luminesc 31(Dec):733–737, Dec 1984.

Brus L. E., **On the Development of Bulk Optical Properties in Small Semiconductor Crystallites.** J Luminesc 31(Dec):381–384, Dec 1984.

Burrus C. A., Bowers J. E., Tucker R. S., **Improved Very-High-Speed Packaged InGaAs Pin Punch-Through Photodiode.** Electr Lett 21(7):262–263, Mar 28 1985.

Calderbank A. R., Mazo J. E., Wei V. K., **Asymptotic Upper Bounds on the Minimum Distance of Trellis Codes.** IEEE Commun 33(4):305–309, Apr 1985.

Capasso F., Levine B. F., **New Transport Phenomena in Variable Gap Semiconductors and Their Device Applications.** J Luminesc 30(1–4):144–153, Feb 1985.

Chemla D. S., **Two-Dimensional Semiconductors—Recent Development.** J Luminesc 30(1–4):502–519, Feb 1985.

Chen C. Y., Garbinski P. A., Kasper B. L., **Bit Rate Dependence of Receiver Sensitivities in $Ga_{0.47}In_{0.53}As$ Photoconductive Detectors.** Electr Lett 21(7):273–274, Mar 28 1985.

Chen C. Y., Olsson N. A., Tu C. W., Garbinski P. A., **Monolithic Integrated Receiver Front End Consisting of a Photoconductive Detector and a GaAs Selectively Doped Heterostructure Transistor.** Appl Phys L 46(7):681–683, Apr 1 1985.

Chi G. C., Mogab C. J., **Rie Planarization Process for Magnetic-Bubble Devices.** IEEE Magnet 21(2):1170–1173, Mar 1985.

Coffman E. G., Kadota T. T., Shepp L. A., **On the Asymptotic Optimality of 1st-Fit Storage Allocation (Letter).** IEEE Soft E 11(2):235–239, Feb 1985.

Dautartas M. F., Suh S. Y., Forrest S. R., Kaplan M. L., Lovinger A. J., Schmidt P. H., **Optical Recording Using Hydrogen Phthalocyanine Thin Films.** Appl Phys A 36(2):71–79, Feb 1985.

Donegan J. F., Bergin F. J., Imbusch G. F., Remeika J. P., **Luminescence From $LiGa_5O_8$-Co.** J Luminesc 31(Dec):278–280, Dec 1984.

Downer M. C., Fork R. L., Shank C. V., **Femtosecond Imaging of Melting and Evaporation at a Photoexcited Silicon Surface.** J Opt Soc B 2(4):595–599, Apr 1985.

Dudderar T. D., Hall P. M., Gilbert J. A., **Holo-Interferometric Measurement of the Thermal Deformation Response to Power Dissipation in Multilayer Printed Wiring Boards.** Exp Mech 25(1):95–104, Mar 1985.

Eisenstein J. P., **High-Precision Torsional Magnetometer—Application to Two-Dimensional Electron Systems.** Appl Phys L 46(7):695–696, Apr 1 1985.

Eisentein G., Tucker R. S., Korothy S. K., Koren U., Veselka J. J., Stulz L. W., Jopson R. M., Hall K. L., **Active Mode Locking of an InGaAsP 1.55-μm Laser in a Fiber Resonator With an Integrated Single-Mode-Fibre Output Port.** Electr Lett 21(5):173–175, Feb 28 1985.

Eisinger J., **Fluorometry of Absorbent and Turbid Samples and the Lateral Mobility in Membranes of Intact Erythrocytes.** J Luminesc 31(Dec):875–880, Dec 1984.

Glass A. M., Klein M. B., Valley G. C., **Photorefractive Determination of the Sign of Photocarriers in InP and GaAs.** Electr Lett 21(6):220–221, Mar 14 1985.

Glassgold A. E., Huggins P. J., Langer W. D., **Shielding of Co From Dissociating Radiation in Interstellar Clouds.** Astrophys J 290(2):615–626, Mar 1985.

Gottscho R. A., **Automated Pressure Scanning of Tunable Dye Lasers.** Rev Sci Ins 56(4):529–531, Apr 1985.

Greene B. I., Wolfe R., **Femtosecond Relaxation Dynamics in Magnetic Garnets.** J Opt Soc B 2(4):600–605, Apr 1985.

Hegarty J., Olsson N. A., Goldner L., **CW Pumped Raman Preamplifier in a 45-km-Long Fiber Transmission-System Operating at 1.5 μm and 1 Gbit/s.** Electr Lett 21(7):290–292, Mar 28 1985.

Hegarty J., Sturge M. D., **Exciton Holeburning in GaAs/GaAlAs Multiquantum Wells.** J Luminesc 31(Dec):494–496, Dec 1984.

Hong M., Maher D. M., Ellington M. B., Hellman F., Geballe T. H., Ekin J. W., Holthuis J. T., **Further Investigations of the Solid-Liquid Reaction and High-Field Critical Current Density in Liquid-Infiltrated Nb-Sn Superconductors.** IEEE Magnet 21(2):771–774, Mar 1985.

Jackel J. L., Veselka J. J., Lyman S. P., **Thermally Tuned Glass Mach-Zehnder Interferometer Used as a Polarization Insensitive Attenuator (Letter).** Appl Optics 24(5):612–614, Mar 1 1985.

Kash K., Shah J., **Hot-Electron Relaxation in $In_{0.53}Ga_{0.47}As$.** J Luminesc 30(1–4):333–339, Feb 1985.

Katehakis M. N., Johri P. K., **Optimal Repair of a 2-Component Series System With Partially Repairable Components.** IEEE Reliab 33(5): 427–430, Dec 1984.

Kevan S. D., **Electronic Coherence Length Following Pulsed-Laser Annealing of Cu(001).** Phys Rev B 31(6):3343–3347, Mar 15 1985.

Korotky S. K., Eisenstein G., Alferness R. C., Veselka J. J., Buhl L. L., Harvey G. T., Read P. H., **Fully Connectorized High-Speed Ti-$LiNbO_3$ Switch Modulator for Time-Division Multiplexing and Data Encoding.** J Light W T 3(1):1–6, Feb 1985.

Law H. H., Wilson W. L., Gabriel N. E., **Separation of Gold Cyanide Ion From Anion-Exchange Resins.** Ind Eng PDD 24(2):236–238, Apr 1985.

Levine B. F., Bethea C. G., Campbell J. C., **1.52-μm Room-Temperature Photon-Counting Optical Time Domain Reflectometer.** Electr Lett 21(5):194–196, Feb 28 1985.

Lifshitz N., **Dependence of the Work-Function Difference Between the Polysilicon Gate and Silicon Substrate on the Doping Level in Polysilicon.** IEEE Device 32(3):617–621, Mar 1985.

Lipson J., Young C. A., Yeates P. D., Masland J. C., Wartonick S. A., Harvey G. T., Read P. H., **A Four-Channel Lightwave Subsystem Using Wavelength Division Multiplexing.** J Light W T 3(1):16–20, Feb 1985.

Luryi S., **An Induced Base Hot-Electron Transistor.** IEEE Elec D 6(4):178–180, Apr 1985.

Marcuse D., Stone J., **Experimental Comparison of the Bandwidths of Standard and Dispersion-Shifted Fibers Near Their Zero-Dispersion Wavelengths.** Optics Lett 10(3):163–165, Mar 1985.

McCaughan L., **Long Wavelength Titanium-Doped Lithium-Niobate Directional Coupler Optical Switches and Switch Arrays.** Opt Eng 24(2):241–243, Mar–Apr 1985.

Miller R. C., Kleinman D. A., **Excitons in GaAs Quantum Wells.** J Luminesc 30(1–4):520–540, Feb 1985.

Mitchell J. W., Wittman P. K., **Metastable Transfer Emission Spectroscopy—Recent Advances and Applications.** J Luminesc 31(Dec):592–594, Dec 1984.

Mollenauer L. F., **Femtosecond Measurement of Configurational Relaxation With the Soliton Laser.** J Luminesc 31(Dec):9–14, Dec 1984.

Mucha J. A., **The Gases of Plasma Etching—Silicon-Based Technology.** Sol St Tech 28(3):123–127, Mar 1985.

Ng K. K., Bayruns R. J., Fang S. C., **The Spreading Resistance of MOSFETS.** IEEE Elec D 6(4):195–198, Apr 1985.

Olsson N. A., Temkin H., Logan R. A., Johnson L. F., Dolan G. J., Vanderziel J. P., Campbell J. C., **Chirp-Free Transmission Over 82.5 km of Single-Mode Fibers at 2 Gbit/s With Injection-Locked DFB Semiconductor Lasers.** J Light W T 3(1):63–67, Feb 1985.

Panish M. B., Temkin H., Sumski S., **Gas Source MBE of InP and $Ga_xIn_{1-x}P_y$-As_{1-y}—Materials Properties and Heterostructure Lasers.** J Vac Sci B 3(2):657–665, Mar–Apr 1985.

Personick S. D., **Switches Take to Optics.** Electronwk 58(11):55–58, Mar 18 1985.

Petroff P. M., **Defects in III-V Compound Semiconductors.** SEM Semimet 22(PA):379–403, 1985.

Prabhu K. A., **A Predictor Switching Scheme for DPCM Coding of Video Signals.** IEEE Commun 33(4):373–379, Apr 1985.

Sandberg I. W., **Multilinear Maps and Uniform Boundedness.** IEEE Circ S 32(4):332–336, Apr 1985.

Sermage B. et al., **Subnanosecond Carriers Lifetime Measurement in 1.3μm InGaAsP.** J Luminesc 31(Dec):500–502, Dec 1984.

Seth S. C., Agrawal V. D., **Cutting Chip-Testing Costs.** IEEE Spectr 22(4):38–45, Apr 1985.

Shah N. J., Pei S. S., Tu C. W., Hendel R. H., Tiberio R. C., **11 PS Ring Oscillators With Submicrometer Selectively Doped Heterostructure Transistors.** Electr Lett 21(4):151–152, Feb 14 1985.

Shang H. T., Lenahan T. A., Glodis P. F., Kalish D., **Design and Fabrication of Dispersion-Shifted Depressed-Clad Triangular-Profile (DDT) Single-Mode Fiber.** Electr Lett 21(5):201–203, Feb 28 1985.

Shank C. V., **Progress in Femtosecond Measurement Techniques.** J Luminesc 30(1–4):243–247, Feb 1985.

Snell R. L., Bally J., Strom S. E., Strom K. M., **Radio and Optical Observations of the Jets From L1551 IRS-5.** Astrophys J 290(2):587–595, Mar 15 1985.

Stavola M., **Two-Center Optical Transitions in Condensed Matter.** J Luminesc 31(Dec):45–49, Dec 1984.

Stormer H. L., **The Fractional Quantum Hall Effect.** Festkorperp 24:25–44, 1984.

Suh S. Y., Snyder D. A., Anderson D. L., **Writing Process in Ablative Optical Recording.** Appl Optics 24(6):868–874, Mar 15 1985.

Taylor C. R., Aloisio C. J., Matsuoka S., **Mechanical Relaxation of Flame-Retardant Polycarbonate Using the Cole-Cole Method.** Polym Eng S 25(2):105–112, Feb 1985.

Temkin H. et al., **Index-Guided Arrays of Schottky-Barrier Confined Lasers.** Appl Phys L 46(5):465–467, Mar 1 1985.

Tsang W. T., **Molecular-Beam Epitaxy for III-V Compound Semiconductors.** SEM Semimet 22(PA):95–207, 1985.

Tucker R. S., Korotky S. K., Eisenstein G., Koren U., Stulz L. W., Veselka J. J., **20-GHz Active Mode-Locking of a 1.55-μm InGaAsP Laser.** Electr Lett 21(6):239–240, Mar 14 1985.

Valdmanis J. A., Fork R. L., Gordon J. P., **Generation of Optical Pulses as Short as 27 Femtoseconds Directly From a Laser Balancing Self-Phase Modulation, Group-Velocity Dispersion, Saturable Absorption, and Saturable Gain.** Optics Lett 10(3):131–133, Mar 1985.

Vandenberg J. M., Gurvitch M., Hamm R. A., Hong M., Rowell J. M., **New Phase Formation and Superconductivity in Reactively Diffused Nb_3Sn Multilayer Films.** IEEE Magnet 21(2):819–822, Mar 1985.

Varaiya P. P., Walrand J. C., Buyukkoc C., **Extensions of the Multiarmed Bandit Problem—The Discounted Case.** IEEE Auto C 30(5):426–439, May 1985.

Vieira N. D., Mollenauer L. F., **Single-Frequency, Single-Knob Tuning of a CW Color Center Laser.** IEEE J Q El 21(3):195–201, Mar 1985.

Whalen M. S., Wood T. H., **Effectively Nonreciprocal Evanescent-Wave Optical-Fibre Directional Coupler.** Electr Lett 21(5):175–176, Feb 28 1985.

Zucker J. E., **Raman-Scattering Resonant with Two-Dimensional Excitons in Semiconductor Heterostructures.** J Luminesc 31(Dec):375–380, Dec 1984.

PHYSICAL SCIENCES

Allara D. L., Nuzzo R. G., **Spontaneously Organized Molecular Assemblies. 1. Formation, Dynamics, and Physical Properties of Normal-Alkanoic Acids Adsorbed From Solution on an Oxidized Aluminum Surface.** Langmuir 1(1):45–52, Jan–Feb 1985.

Allara D. L., Nuzzo R. G., **Spontaneously Organized Molecular Assemblies. 2. Quantitative Infrared Spectroscopic Determination of Equilibrium Structures of Solution-Adsorbed Normal-Alkanoic Acids on an Oxidized Aluminum Surface.** Langmuir 1(1): 52–66, Jan–Feb 1985.

Ashkin A., Dziedzic J. M., **Observation of Radiation-Pressure Trapping of Particles by Alternating Light Beams.** Phys Rev L 54(12):1245–1248, Mar 25 1985.

Bair H. E., **Curing Behavior of an Epoxy-Resin Above and Below TG.** Polym Prepr 26(1):10–11, Apr 1985.

Bates F. S., **Measurement of the Correlation Hole in Homogeneous Block Copolymer Melts.** Macromolec 18(3):525–528, Mar 1985.

Batlogg B., Boppart H., **High-Resolution Pressure Volume Measurements by a Strain-Gauge Technique.** Rev Sci Ins 56(3):459–461, Mar 1985.

Bohr J. et al., **Model-Independent Structure Determination of the INSB(111) 2 × 2 Surface With Use of Synchrotron X-Ray Diffraction.** Phys Rev L 54(12):1275–1278, Mar 25 1985.

Brawer S. A., **A Theory of Dense Liquids Based on Monte-Carlo Simulations of Very Small Clusters.** J Chem Phys 82(4):2092–2105, Feb 15 1985.

Campisano S. U., Gibson J. M., Poate J. M., **Interface and Precipitation Effects in Solid-Phase Epitaxy of Sb Implanted Amorphous Si.** Appl Phys L 46(6):580–581, Mar 15 1985.

Capasso F., Cho A. Y., Mohammed K., Foy P. W., **Doping Interface Dipoles—Tunable Heterojunction Barrier Heights and Band-Edge Discontinuities by Molecular-Beam Epitaxy.** Appl Phys L 46(7):664–666, Apr 1 1985.

Cardillo M. J., **Concepts in Gas Surface Dynamics.** Langmuir 1(1):4–10, Jan–Feb 1985.

Chabal Y. J., Raghavachari K., **New Ordered Structure for the H-Saturated Si(100) Surface—The (3 × 1) Phase.** Phys Rev L 54(10):1055–1058, Mar 11 1985.

Chitanvis S. M., Lax M., **Scattering of Waves From Rough Surfaces—Specular and Incoherent Fields.** Supp Pr T P (80):40–46, 1984.

Cholli A. L., Yamane T., Jelinski L. W., **Combining Solid-State and Solution-State P-31 NMR to Study In Vivo Phosphorus Metabolism.** P Nas US 82(2):391–395, Jan 1985.

Chrien R. E., Kopecky J., Liou H. I., Wasson O. A., Garg J. B., Dritsa M., **Distribution of Radiative Strength From Neutron Capture by Pu-239.** Nucl Phys A 436(2):205–220, Apr 1 1985.

Colvard C., Fischer R., Gant T. A., Klein M. V., Merlin R., Morkoc H., Gossard A. C., **Phonon Freedom and Confinement in GaAs-Al$_x$Ga$_{1-x}$As Superlattices.** Superlatt M 1(1):81–86, 1985.

Cook R., Helfand E., **Time-Correlation Functions for a One-Dimensional Polymer Model.** J Chem Phys 82(3):1599–1605, Feb 1 1985.

Coppersmith S. N., Littlewood P. B., **Inductive Response From Nonlinear Mixing in CDWs.** Lect N Phys 217:236–239, 1985.

Coppersmith S. N., Varma C. M., **Shift in the Longitudinal Sound Velocity Due to Sliding Charge-Density Waves.** Lect N Phys 217:206–210, 1985.

Couchman P. R., **The Composition-Dependent Glass Transition in Theory and Practice.** Polym Prepr 26(1):13–14, Apr 1985.

Disalvo F. J., Waszczak J. V., **Absence of a Charge-Density Wave in the Structurally One-Dimensional Phosphide VP4.** Mater Res B 20(3):351–354, Mar 1985.

Dutta N. K., Wessel T., Olsson N. A., Logan R. A., Koszi L. A., Yen R., **Fabrication and Performance Characteristics of 1.55-μm InGaAsP Multiquantum Well Ridge Guide Lasers.** Appl Phys L 46(6):525–527, Mar 15 1985.

Eaton J. A., Johnson H. R., O'Brien G. T., Baumert J. H., **Ultraviolet Spectra and Chromospheres of R-Stars.** Astrophys J 290(1):276–283, Mar 1 1985.

Eisenstein J. P., Stormer H. L., Narayanamurti V., Gossard A. C., **High-Precision Dehaas-Vanalphen Measurements on a Two-Dimensional Electron Gas.** Superlatt M 1(1):11–14, 1985.

Elliman R. G., Gibson J. M., Jacobson D. C. Poate J. M., Williams J. S., **Diffusion and Precipitation in Amorphous Si.** Appl Phys L 46(5):478–480, Mar 1 1985.

Endo M. et al., **Structure of Ion-Implanted Graphite Fibers.** J Chim Phys 81(11–1):803–808, Nov–Dec 1984.

Farrell H. H., Levinson M., **Scanning Tunneling Microscope as a Structure-Modifying Tool.** Phys Rev B 31(6):3593–3598, Mar 15 1985.

Fisher D. S. et al., **Scaling in Spin Glasses.** Phys Rev L 54(10):1063–1066, Mar 11 1985.

Franey J. P., Kammlott G. W., Graedel T. E., **The Corrosion of Silver by Atmospheric Sulfurous Gases.** Corros Sci 25(2):133–143, 1985.

Gallagher P. K., Obryan H. M., **Characterization of $LiNbO_3$ by Dilatometry and DTA.** J Am Ceram 68(3):147–150, Mar 1985.

Gibson J. M., Hull R., Bean J. C., Treacy M. M. J., **Elastic Relaxation in Transmission Electron-Microscopy of Strained-Layer Superlattices.** Appl Phys L 46(7):649–651, Apr 1 1985.

Goodby J. W., Patel J. S., Leslie T. M., **Ferroelectric Switching in the Tilted Smectic Phases of R-(−)-4-N-Hexyloxybenzylidene4′-Amino-(2-Chloropropyl)Cinnamate (HOBACPC).** Ferroelectr 58(1–4):441–456, 1984.

Gornik E., Lassnig R., Strasser G., Stormer H. L., Gossard A. C., Wiegmann W., **Specific-Heat of Two-Dimensional Electrons in GaAs-GaAlAs Multilayers.** Phys Rev L 54(16):1820–1823, Apr 22 1985.

Graedel T. E., Plewes J. T., Franey J. P., Kammlott G. W., Stoffers R. C., **Sulfidation Under Atmospheric Conditions of Cu-Ni, Cu-Sn, and Cu-Zn Binary and Cu-Ni-Sn and Cu-Ni-Zn Ternary Systems.** Metall T-A 16(2):275+, Feb 1985.

Greywall D. S., **He_3 Melting-Curve Thermometry at Millikelvin Temperatures.** Phys Rev B 31(5):2675–2683, Mar 1 1985.

Haight R., Bokor J., Stark J., Storz R. H., Freeman R. R., Bucksbaum P. H., **Picosecond Time-Resolved Photoemission Study of the InP(110) Surface.** Phys Rev L 54(12):1302–1305, Mar 25, 1985.

Harbison J. P., Derkits G. E., **Tungsten Patterning as a Technique for Selective Area III-V MBE Growth.** J Vac Sci B 3(2):743–745, Mar–Apr 1985.

Hayes J. R., Levi A. F. J., Wiegmann W., **Hot-Electron Spectroscopy of GaAs.** Phys Rev L 54(14):1570–1572, Apr 8 1985.

Hensel J. C., Tung R. T., Poate J. M., Unterwald F. C., **Specular Boundary Scattering and Electrical Transport in Single-Crystal Thin Films of $CoSi_2$.** Phys Rev L 54(16):1840–1843, Apr 22 1985.

Higashi G. S., Rothberg L. J. Fleming C. G., **Vibrational Spectroscopy of Growth Surfaces During Photochemical Deposition of Aluminum from Trimethylaluminum Vapor.** Chem P Lett 115(2):167–172, Mar 29 1985.

Hockberger P., Connor J. A., **Alteration of Calcium Conductances and Outward Current by Cyclic Adenosine-Monophosphate (CAMP) in Neurons of Limax-Maximus.** Cell Mol N 4(4):319–338, Dec 1984.

Hohenberg P. C., **Nonequilibrium Steady States With Spatial Patterns.** Phys Scr T9:93–94, 1985.

Ibbotson D. E., Mucha J. A., Flamm D. L. Cook J. M., **Selective Interhalogen Etching of Tantalum Compounds and Other Semiconductor Materials.** Appl Phys L 46(8):794–796, Apr 15 1985.

Iye Y., **Non-Ohmic Transport in the Magnetic-Field-Induced Charge-Density-Wave Phase of Graphite.** Phys Rev L 54(11):1182–1184, Mar 18 1985.

Jayaraman A., **The Diamond Anvil Cell and High-Pressure Research.** J Physique 45(NC8):355–363, Nov 1984.

Jayaraman A., Kaplan M. L., Schmidt P. H., **Effect of Pressure on the Raman and Electronic Absorption Spectra of Naphthalenetetracarboxylic and Perylenetetracarboxylic Dianhydrides.** J Chem Phys 82(4):1682–1687, Feb 15 1985.

Kevan S. D., Stoffel N. G., Smith N. V., **Surface States on Low-Miller-Index Copper Surfaces.** Phys Rev B 31(6):3348–3355, Mar 15 1985.

Knox W. H., Fork R. L., Downer M. C., Miller D. A. B., Chemla D. S., Shank C. V., Gossard A. C., Wiegmann W., **Femtosecond Dynamics of Resonantly Excited Excitons in Room-Temperature GaAs Quantum Wells.** Phys Rev L 54(12):1306–1309, Mar 25, 1985.

Krigas T. M. et al., **Model Copolymers of Ethylene With Butene-1 Made by Hydrogenation of Polybutadiene-Chemical-Composition and Selected Physical Properties.** J Pol SC PP 23(3):509–520, Mar 1985.

Lanzerotti L. J. et al., **Hydromagnetic Field Line Resonances and Modulation of Particle Precipitation.** Planet Spac 33(3):253–262, Mar 1985.

Larson R. G., **Derivation of Strain Measures From Strand Convection Models for Polymer Melts.** J Non-Newt 17(1):91–110, Jan 1985.

Leslie T. M., **The Ferroelectric Phases Derived From the 4-N-Alkoxycinnamic Acids.** Ferroelectr 58(1–4):9–20, 1984.

Leung S. Y., **Thermal Considerations in Multiwafer Liquid-Phase Epitaxy (LPE) Boat Design.** J Cryst Gr 69(2–3):291–300, Nov 1984.

Licini J. C., Dolan G. J., Bishop D. J., **Weakly Localized Behavior in Quasi-One-Dimensional Li Films.** Phys Rev L 54(14):1585–1588, Apr 8 1985.

Littlewood P. B., **Pinning, Metastability and Sliding of Charge-Density Waves.** Lect N Phys 217:369–376, 1985.

Lynn K. G., Mills A. P., West R. N., Berko S., Canter K. F., Roellig L. O., **Positron or Positronium-Like Surface State on Al(100).** Phys Rev L 54(15):1702–1705, Apr 15 1985.

Lyons A. M., **Photodefinable Carbon Films—Electrical Properties.** J Non-Cryst 70(1):99–109, Feb 1985.

Malik R. J., Capasso F., Stall R. A., Kiehl R. A., Ryan R. W., Wunder R., Bethea C. G., **High-Gain, High-Frequency AlGaAs/GaAs Graded Band-Gap Base Bipolar Transistors With a Be Diffusion Setback Layer in the Base.** Appl Phys L 46(6):600–602, Mar 15 1985.

Malyj M., Espinosa G. P., Griffiths J. E., **Structure and Delocalized Vibrational Modes in Vitreous $Si_x(Se_{1-y}Te_y)_{1-x}$.** Phys Rev B 31(6):3672–3679, Mar 15 1985.

McNevin S. C., Becker G. E., **Investigation of Kinetic Mechanism for the Ion-Assisted Etching of Si in Cl_2.** J Vac Sci B 3(2):485–491, Mar–Apr 1985.

Mills A. P., Crane W. S., **Low-Energy Positron-Diffraction Study of NAF and LIF.** Phys Rev B 31(6):3988–3992, Mar 15 1985.

Mitchell J. W., **Chemical Analysis of Electronic Gases and Volatile Reagents for Device Processing.** Sol St Tech 28(3):131–137, Mar 1985.

Nakahara S., Okinaka Y., **On the Effect of Hydrogen on Properties of Copper.** Scrip Metal 19(4):517–519, Apr 1985.

Nishikawa K. I., Okuda H., Hasegawa A., **Heating of Heavy Ions on Auroral Field Lines in the Presence of a Large-Amplitude Hydrogen Cyclotron Wave.** J Geo R-S P 90(NA1):419–428, Jan 1 1985.

Ogielski A. T., Morgenstern I., **Critical Behavior of Three-Dimensional Ising Spin-Glass Model.** Phys Rev L 54(9):928–931, Mar 4 1985.

Osheroff D. D. et al., **Novel Magnetic-Field Dependence of the Coupling of Excitations Between Two Fermion Fluids.** Phys Rev L 54(11):1178–1181, Mar 18 1985.

Paalanen M. A., Ruckenstein A. E., Thomas G. A., **Spins in Si-P Close to the Metal-Insulator Transition.** Phys Rev L 54(12):1295–1298, Mar 25 1985.

Pai C. S., Cabreros E., Lau S. S., Seidel T. E., Suni I., **Rapid Thermal Annealing of Al-Si Contacts.** Appl Phys L 46(7):652–654, Apr 1 1985.

Panek M. G. et al., **Thermolysis Rates and Products of the Putative Ketochloroallyl Groups in Poly(vinyl-Chloride), as Inferred From the Behavior of Analogous Model Compounds.** Polym Prepr 26(1):120–121, Apr 1985.

Papadopoulos S., Barr D., Brown W. L., Wagner A., **The Energy Spread of Ions From Gold Liquid-Metal Ion Sources as a Function of Source Parameters.** J Physique 45(NC9):217–222, Dec 1984.

Patel D. J., Kozlowski S. A., Weiss M., Bhatt R., **Conformation and Dynamics of the Pribnow Box Region of the Self-Complementary D(C-G-A-T-T-A-T-A-A-T-C-G) Duplex in Solution.** Biochem 24(4):936–944, Feb 12 1985.

Patel D. J., Kozlowski, S. A., Hare D. R., Reid B., Ikuta S., Lander N., Itakura K., **Conformation, Dynamics, and Structural Transitions of the TATA Box Region of Self-Complementary $d[(C\text{-}G)_N\text{-}T\text{-}A\text{-}T\text{-}A\text{-}(C\text{-}G)_N{}^*]$ Duplexes in Solution.** Biochem 24(4):926–935, Feb 12 1985.

Patel J. S., Leslie T. M., Goodby J. W., **A Reliable Method of Alignment for Smectic Liquid Crystals.** Ferroelectr 58(1–4):457–464, 1984.

Patterson G. D., Carroll P. J., **Light-Scattering Spectroscopy of Pure Fluids.** J Phys Chem 89(8):1344–1354, Apr 11 1985.

Pfeiffer L., Paine S., Gilmer G. H., Van Saarloos W., West K. W., **Pattern Formation Resulting From Faceted Growth in Zone-Melted Thin Films.** Phys Rev L 54(17):1944–1947, Apr 29 1985.

Rubinstein M., Helfand E., **Statistics of the Entanglement of Polymers—Concentration Effects.** J Chem Phys 82(5):2477–2483, Mar 1 1985.

Schillinger F. C. et al., **C-13 Nuclear Magnetic-Resonance Characterization of Random Ethylene Vinyl-Chloride Copolymers.** Macromolec 18(3):356–360, Mar 1985.

Scott T. W., Braun C. L., **Picosecond Measurements of Geminate Charge Pair Recombination in Photoionized Liquids.** Can J Chem 63(1):228–231, Jan 1985.

Sette F., Stohr J., Kollin E. B., Dwyer D. J., Gland J. L., Robbins J. L., Johnson A. L., **Na-Induced Bonding and Bond-Length Changes for Co on Pt(111)—A Near-Edge X-Ray-Absorption Fine-Structure Study.** Phys Rev L 54(9):935–938, Mar 4, 1985.

Silberberg Y., Smith P. W., Miller D. A. B., Tell B., Gossard A. C., Wiegmann W., **Fast Nonlinear Optical Response From Proton-Bombarded Multiple Quantum Well Structures.** Appl Phys L 46(8):701–703, Apr 15 1985.

Silfvast W. T., Wood O. R., Lundberg H., Macklin J. J., **Stimulated Emission in the Ultraviolet by Optical Pumping From Photoionization-Produced Inner-Shell States in Cd^+.** Optics Lett 10(3):122–124, Mar 1985.

Sinclair J. D., Psota-Kelty L. A., Weschler C. J., **Indoor Outdoor Concentrations and Indoor Surface Accumulations of Ionic Substances.** Atmos Envir 19(2):315–323, 1985.

Stall R. A. et al., **Morphology of GaAs and $Al_xGa_{1-x}As$ Grown by Molecular-Beam Epitaxy.** J Vac Sci B 3(2):524–527, Mar–Apr 1985.

Stavola M., Parsey J. M., Forrest S. R., Kaplan M. L., Schmidt P. H., Young M. S. S., **Transient Capacitance Analysis of III-V Semiconductors With Organic-On-Inorganic Semiconductor Contact Barrier Diodes.** Appl Phys L 46(5):506–508, Mar 1 1985.

Takase Y. et al., **Observation of Parametric Instabilities in the Lower-Hybrid Range of Frequencies in the High-Density Tokamak.** Phys Fluids 28(3):983–994, Mar 1985.

Tamargo M. C., Hull R., Greene L. H., Hayes J. R., Cho A. Y., **Growth of a Novel InAs-GaAs Strained Layer Superlattice on InP.** Appl Phys L 46(6):569–571, Mar 15 1985.

Tsang W. T., **Growth of InP, GaAs, and $In_{0.53}Ga_{0.47}As$ by Chemical Beam Epitaxy.** J Vac Sci B 3(2):666–670, Mar–Apr 1985.

Tsang W. T., **Selective Area Growth of GaAs and $In_{0.53}Ga_{0.47}As$ Epilayer Structures by Chemical Beam Epitaxy Using Silicon Shadow Masks—A Demonstration of the Beam Nature.** Appl Phys L 46(8):742–744, Apr 15 1985.

Tsang W. T., Chiu T. H., Chu S. N. G., Ditzenberger J. A., **$GaSb_{0.5}As_{0.5}/Al_{0.35}Ga_{0.65}Sb_{0.48}As_{0.52}$ Superlattice Lattice Matched to InP Prepared by Molecular-Beam Epitaxy.** Appl Phys L 46(7):659–661, Apr 1 1985.

Varma C. M., **Aspects of Strong Electron-Phonon Coupling Related to the CDW Transition at Temperatures Above It.** Lect N Phys 217:99–105, 1985.

Wang T. T., Von Seggern H., West J. E., Keith H. D., **High-Field Poling of Poly(Vinylidene Fluoride) Films Using a Current Limiting Circuit.** Ferroelectr 61(4):249–256, 1984.

Weiner J. S., Chemla D. S., Miller D. A. B., Wood T. H., Sivco D., Cho A. Y., **Room-Temperature Excitons in 1.6-μm Band-Gap GaLnAs/AlLnAs Quantum Wells.** Appl Phys L 46(7):619–621, Apr 1 1985.

Rousseau D. L., Ondrias M. R., **Resonance Raman-Spectra of Photodissociated Hemoglobins—Implications on Cooperative Mechanisms.** Biophys J 47(4):537–545, Apr 1985.

SOCIAL AND LIFE SCIENCES

Eisinger J., Flores J., Tyson J. A., Shohet S. B., **Fluorescent Cytoplasm and Heinz Bodies of Hemoglobin Koln Erythrocytes—Evidence for Intracellular Heme Catabolism.** Blood 65(4):886–893, Apr 1985.

Fishburn P. C., Brams S. J., **Manipulability of Voting by Sincere Truncation of Preferences.** Publ Choice 44(3):397–410, 1984.

Starr S. J., Shute S. J., Thompson C. R., **Relating Posture to Discomfort in VDT Use.** J Occup Med 27(4):269–271, Apr 1985.

CONTENTS, OCTOBER 1985

AT&T TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Current Contents/Engineering, Technology & Applied Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Science Abstracts (Series A, Physics Abstracts; Series B, Electrical and Electronic Abstracts; and Series C, Computer & Control Abstracts), Science Citation Index, Sociological Abstracts, Social Welfare, Social Planning and Social Development*, and *Solid State Abstracts Journal*. Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

AT&T