THE BELL SYSTEM TECHNICAL JOURNAL

NOVEMBER 1983
VOL. 62, NO. 9, PART 1

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL (ISSN0005-8580) is published by the American Telephone and Telegraph Company, 195 Broadway, N. Y., N. Y. 10007; C. L. Brown, Chairman and Chief Executive Officer; W. M. Ellinghaus, President; V. A. Dwyer, Vice President and Treasurer; T. O. Davis, Secretary.

The Journal is published in three parts. Part 1, general subjects, is published ten times each year. Part 2, Computing Science and Systems, and Part 3, single-subject issues, are published with Part 1 as the papers become available.

The subscription price includes all three parts. Subscriptions: United States—1 year $35; 2 years $63; 3 years $84; foreign—1 year $45; 2 years $73; 3 years $94. Subscriptions to Part 2 only are $10 ($12 foreign). Single copies of the Journal are available at $5 ($6 foreign). 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 Bell System Technical Journal and sent to Bell Laboratories, Circulation Dept., Room 1E-335, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078.

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room 1J-319, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of the American Telephone and Telegraph Company. Comments selected for publication may be edited for brevity, subject to author approval.

Printed in U.S.A. Second-class postage paid at Short Hills, N. J. 07078 and additional mailing offices. Postmaster: Send address changes to The Bell System Technical Journal, Room 1E-335, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 62 November 1983 Number 9, Part 1

Calculation of Modes in an Optical Fiber Using the Finite Element Method and EISPACK

By T. A. LENAHAN*

(Manuscript received April 20, 1983)

This paper presents a method for computing the propagation modes of a circular optical fiber. Finite element analysis reduces Maxwell's equations to standard eigenvalue equations involving symmetric tridiagonal matrices. Routines from the Eigensystem Package (EISPACK) compute their eigenvalues and eigenvectors, and from these the waveforms, propagation constants, and delays (per unit length) of the modes are obtained. An extension allows loss of leaky modes to be calculated. Examples indicate that the method is reliable, economical, and comprehensive, applying to both single and multimode fibers.

I. INTRODUCTION

This paper presents a method for calculating the propagation modes of a circular optical fiber. The modal quantities, essential for telecommunications, include the waveforms (which describe the radial distribution of propagating power), the propagation constants (which determine cutoff conditions), and the delays per unit length (which

*Bell Laboratories.

determine pulse dispersion along the fiber). The technique combines the Finite Element Method (FEM) and routines from the Eigensystem Package (EISPACK)[1] to achieve an efficient calculation of these modal quantities for both single and multimode fibers.

The most popular approach to modal calculations for multimode fibers has been the Wentzel, Kramers, Brillouin (WKB) method, where Maxwell's equations are approximated by an easily integrated first-order differential equation.[2] WKB analysis provides a simple model for understanding optical transmission through a fiber[3] and has guided fiber design.[4]

For fibers with index of refraction profiles described by a power law, the WKB method is equivalent to geometric optics and to the model of a fiber with unlimited radial extension.[5] For such fibers the effect of the outer cladding is missed by the WKB and equivalent methods, and the bandwidth capability is often overestimated.[6]

Accuracy of the WKB method declines substantially with the number of propagating modes. For single-mode fibers the WKB method is not suitable, and instead, various analytic and numerical techniques have been used.

Analytic calculations have centered about the step-index profile and the infinitely extending parabolic profiles. Modal quantities have been expressed in terms of Bessel functions for the single step[7] and, also, the double step.[8] Parabolic profiles, analyzed by analogy with the harmonic oscillator,[9] have well-known expressions for their modal quantities. Coupled with perturbation analysis,[10] these results cover a broad range of profiles. But numerical approaches permit an even more comprehensive treatment.

Several numerical procedures have extended the analysis of the step-index profile. The solution in the core comes from a numerical integration; the solution in the cladding (where the index is assumed constant) is well known. Boundary conditions at the core-cladding interface link the two solutions and lead to a set of linear equations involving the propagation constant as a parameter. A search of the corresponding determinant for zeroes gives the propagation constants and, subsequently, the waveforms and delays.

In Ref. 11 this procedure is applied to Maxwell's first-order vector equations, and important results have been obtained for multimode[12,13] and single-mode[14,15] fibers. A similar scheme[16] deals with two coupled second-order differential equations equivalent to Maxwell's equations. The second-order scalar wave equation approximates Maxwell's equations by neglecting the relatively small gradient index terms. In Ref. 17 the basic procedure is applied to the scalar wave equation, and a variety of results have been obtained for single-mode fibers.[18,19]

In Ref. 20 the FEM is developed in terms of a variational principle

for the vector equations. Approximate solutions in the core and cladding are matched at the interface by performing a determinant search, as described before. The FEM has also been applied[21] to diverse single-mode waveguides by approximating the index profile by piecewise constant functions and then enforcing boundary conditions across the numerous interfaces. The result is a matrix eigenvalue equation in generalized form. Both of these FEM approaches seem limited to a small number of modes for economical operation.

The approach in this paper is characterized by a sequence of three steps. First, Maxwell's equations are transformed to ordinary differential equations in eigenvalue form. In one case, gradient index terms are neglected, and in the second, they are considered to first order so that their effect can be monitored. Next, finite element analysis, using the Galerkin technique,[22] reduces these differential equations to matrix equations in standard eigenvalue form. The matrices are symmetric and tridiagonal, and their positive eigenvalues correspond to the propagation modes. The routine BISECT in the EISPACK[1] library delivers the eigenvalues and TINVIT, also in EISPACK, delivers the corresponding eigenvectors. The eigenvalues give the propagation constants of the modes, the eigenvectors give the waveforms, and a combination of the two give the delays.

Like many other numerical techniques, this method can treat any uniform, circular fiber and meet usual standards of accuracy. Also, the effect of gradient index terms can be monitored. But by casting the equations in standard eigenvalue form, modern techniques of computational linear algebra (as used in EISPACK) can achieve substantial cost advantage over other numerical approaches. Typically, this method will process for a multimode fiber 25 modes per second on the *Cray-1** computer.

The calculation procedure is derived in the next section. Effects of material dispersion are incorporated into the analysis, and calculation of loss for leaky modes is also considered. Results are given in Section III for a variety of single and multimode examples. These results, as discussed in Section IV, illustrate the reliability, economy, and scope of the method.

II. ANALYSIS

This section derives from Maxwell's equations an algorithm that computes the propagation modes of an optical fiber. The algorithm is straightforward and can be carried out on any computer that has access to the EISPACK[1] or similar routines.

---

* Registered service mark of Cray Research, Inc.

2.1 Reduction of Maxwell's equations

The fiber is assumed perfectly straight and circular, and uniform along its length. The cylindrical coordinate system (r, θ, z) is defined so that the z-axis coincides with the fiber axis. The index of refraction can then be expressed by the function $n(r)$, the index profile. The profile can be any bounded function in the core ($r \leq R_1$), but it is constant (n_{cl}) in the cladding. The geometry is shown in Fig. 1.

The permittivity is

$$\epsilon = \epsilon_o n^2(r), \tag{1}$$

where ϵ_o denotes the free-space permittivity. The permeability μ is assumed throughout to be μ_o, the free-space value.

Maxwell's equations relate the electric field E and the magnetic field H by

$$\begin{aligned} \text{curl } E &= -i\omega\mu H \\ \text{curl } H &= i\omega\epsilon E, \end{aligned} \tag{2}$$

where ω denotes the frequency of excitation (in rad/s). Taking the curl of the second equation eliminates E to give

$$\nabla^2 H + \nabla g \times (\nabla \times H) + k^2 H = 0, \tag{3}$$

where

$$g = \ln(k^2), \tag{4}$$

and the wave number k has the forms

$$k = \omega(\mu\epsilon)^{1/2} = \omega n(\mu_o\epsilon_o)^{1/2} = \omega n/c = 2\pi n/\lambda, \tag{5}$$

with n the index of refraction, c the velocity of light in vacuum, and λ the free-space wavelength of the excitation.

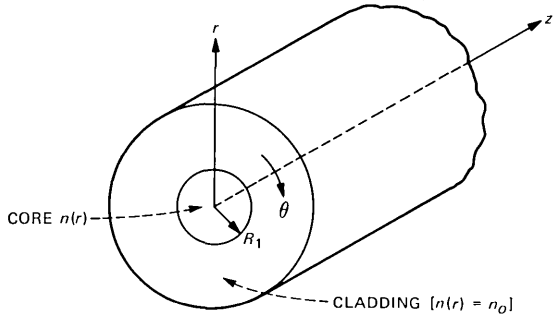


Fig. 1—Geometry of a uniform circular fiber with index profile $n(r)$.

Propagation modes are solutions of the form

$$H = H_o \exp(i\beta z), \tag{6}$$

where H_o is a vector field independent of z, and β is the propagation constant of the mode. Substituting this into the field equation gives, in cylindrical coordinates,

$$\begin{bmatrix} \left\{ \Delta^2 - \frac{1}{r^2} + k^2 - \frac{g_r}{r}\frac{\partial}{\partial r} r \left(\frac{2}{r^2} + \frac{g_r}{r}\right)\frac{\partial}{\partial\theta} \right\} & 0 \\ \left\{ -\frac{2}{r^2}\frac{\partial}{\partial\theta} \quad \nabla^2 - \frac{1}{r^2} + k^2 \right\} & 0 \\ 0 \quad i\beta g_r \quad \nabla^2 - g_r\frac{\partial}{\partial r} + k^2 \end{bmatrix} \begin{pmatrix} H_{o\theta} \\ H_{or} \\ H_{oz} \end{pmatrix} = \beta^2 \begin{pmatrix} H_{o\theta} \\ H_{or} \\ H_{oz} \end{pmatrix}, \tag{7}$$

where g_r means $\partial g/\partial r$. The transverse (i.e., r and θ) components of H are uncoupled from the longitudinal (i.e., z) component and satisfy an eigenvalue equation with eigenvalue β^2. The corresponding operators are indicated as a 2×2 submatrix in the full 3×3 matrix.

The angular dependence of the transverse field is given by

$$\begin{pmatrix} H_{o\theta} \\ H_{or} \end{pmatrix} = \begin{pmatrix} h_\theta \cos m'\theta \\ h_r \sin m'\theta \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} h_\theta \sin m'\theta \\ h_r \cos m'\theta \end{pmatrix}$$

$$\text{for} \quad m' = 0, 1, 2, \cdots, \tag{8}$$

where the functions h_θ and h_r depend only on r. The two forms correspond to different polarizations. Substituting these into eq. (7) gives

$$\begin{bmatrix} \left\{ \left(\frac{1}{r}\frac{d}{dr} r \frac{d}{dr} - \frac{g_r}{r}\frac{d}{dr} r - \frac{m'^2 + 1}{r^2} + k^2\right) \pm m' \left(\frac{2}{r^2} + \frac{g_r}{r}\right) \right\} \\ \left\{ \pm \frac{2m'}{r^2} \quad \left(\frac{1}{r}\frac{d}{dr} r \frac{d}{dr} - \frac{m'^2 + 1}{r^2} + k^2\right) \right\} \end{bmatrix} \cdot \begin{pmatrix} h_\theta \\ h_r \end{pmatrix} = \beta^2 \begin{pmatrix} h_\theta \\ h_r \end{pmatrix}, \tag{9}$$

where the $\pm$ sign depends on the polarization.

The case where $m' = 0$ reduces to two uncoupled equations:

$$\left(\frac{k^2}{r}\frac{d}{dr}\frac{r}{k^2}\frac{d}{dr} - \frac{1}{r^2} - \frac{g_r}{r} + k^2\right) h_\theta = \beta^2 h_\theta \tag{10}$$

for the Transverse Magnetic (TM) modes for which H_z is identically zero, and

$$\left(\frac{1}{r}\frac{d}{dr}r\frac{d}{dr}-\frac{1}{r^2}+k^2\right)h_r=\beta^2 h_r \tag{11}$$

for the Transverse Electric (TE) modes for which E_z is identically zero. Often, gradient index terms (those involving g_r) are neglected on the basis that profile variations are relatively small.[4] The difference or splitting in β^2 for the TM and TE modes measures the accuracy of this practice.

When $m' \neq 0$, eq. (9) is expressed as

$$\begin{pmatrix} A & B \\ B & A \end{pmatrix}\begin{pmatrix} h_\theta \\ h_r \end{pmatrix}+\frac{1}{2}\begin{pmatrix} a & b \\ -b & -a \end{pmatrix}\begin{pmatrix} h_\theta \\ h_r \end{pmatrix}=\beta^2\begin{pmatrix} h_\theta \\ h_r \end{pmatrix}, \tag{12}$$

where

$$A=\frac{1}{4}\frac{d}{dr}r\frac{d}{dr}-\frac{m'^2+1}{r^2}+k^2+\frac{a}{2}, \qquad a=-\frac{g_r}{r}\frac{d}{dr}r$$

$$B=\pm\frac{2m'}{r^2}+\frac{b}{2}, \qquad b=\pm\frac{m'g_r}{r}. \tag{13}$$

The initial term has eigenvectors of the form,

$$\begin{pmatrix} f_1 \\ f_1 \end{pmatrix}, \tag{14}$$

where $(A+B)f_1=\beta^2 f_1$, and

$$\begin{pmatrix} f_2 \\ -f_2 \end{pmatrix}, \tag{15}$$

where $(A-B)f_2=\beta^2 f_2$. The second term of eq. (12) is neglected because its first order perturbation contribution is zero, as in general

$$\begin{pmatrix} a & b \\ -b & -a \end{pmatrix}\begin{pmatrix} f \\ \pm f \end{pmatrix} \text{ is orthogonal to } \begin{pmatrix} f \\ \pm f \end{pmatrix}, \tag{16}$$

respectively. If two eigenvalues are equal or nearly equal, a degenerate-perturbation calculation may be required.

To first order, the modes when $m' \neq 0$ are either the $EH_{m',n}$ with transverse H fields of the form

$$H_{ot}=\begin{pmatrix} f\sin m'\theta \\ f\cos m'\theta \end{pmatrix} \text{ and } \begin{pmatrix} f\cos m'\theta \\ -f\sin m'\theta \end{pmatrix}, \tag{17}$$

where by eqs. (14) and (15) f satisfies

$$\left(\frac{k}{r}\frac{d}{dr}\frac{r}{k}\frac{d}{dr} - \frac{(m'+1)^2}{r^2} - \frac{(m'+1)}{2r}g_r + k^2\right)f = \beta^2 f, \tag{18}$$

or the $HE_{m',n}$ with

$$H_{ot} = \begin{pmatrix} f\cos m'\theta \\ f\sin m'\theta \end{pmatrix} \text{ and } \begin{pmatrix} f\sin m'\theta \\ -f\cos m'\theta \end{pmatrix}, \tag{19}$$

where f satisfies

$$\left(\frac{k}{r}\frac{d}{dr}\frac{r}{k}\frac{d}{dr} - \frac{(m'-1)^2}{r^2} + \frac{(m'-1)}{2r}g_r + k^2\right)f = \beta^2 f. \tag{20}$$

With $m = m' \pm 1$, eqs. (18) and (20) can be combined as

$$\left(\frac{k}{r}\frac{d}{dr}\frac{r}{k}\frac{d}{dr} - \frac{m^2}{r^2} \mp \frac{m}{2r}g_r + k^2\right)f = \beta^2 f. \tag{21}$$

The modes are indexed by the angular parameter m. For $m = 0$ there are two polarizations for each HE_{1n} mode. This group includes the fundamental mode HE_{11}, which propagates in single-mode fibers. For $m = 1$ there are two polarizations for the HE_{2n} modes, also the TM modes and the TE modes. For $m > 1$ there are two polarizations for the $HE_{m+1,n}$ modes and two for the $EH_{m-1,n}$ modes.

When gradient index terms are neglected, eq. (9) reduces to the scalar wave equation

$$\left(\frac{1}{r}\frac{d}{dr}r\frac{d}{dr} - \frac{m^2}{r^2} + k^2\right)f = \beta^2 f. \tag{22}$$

For $m = 0$ each solution represents two polarizations, i.e., two modes; for $m \neq 0$ each solution represents two polarizations and two angular harmonics, $m' = m \pm 1$, i.e., four modes.

The scalar wave equation and its counterparts [eqs. (10) and (21)] assume the form of a time-independent Schroedinger equation in two dimensions. In particular, the scalar wave equation can be expressed as

$$\left(\frac{1}{r}\frac{d}{dr}r\frac{d}{dr} - \frac{m^2}{r^2} + (k^2 - k_{\text{cl}}^2)\right)f = (\beta^2 - k_{\text{cl}}^2)f, \tag{23}$$

where

$$k_{\text{cl}} = 2\pi n_{\text{cl}}/\lambda. \tag{24}$$

The quantity $(k^2 - k_{\text{cl}}^2)$ plays the role of a potential that is 0 in the outer cladding and beyond; $(\beta^2 - k_{\text{cl}}^2)$ is an eigenvalue.

Results on Schroedinger's operators[23] imply that the propagation

modes correspond on a one-to-one basis to the positive eigenvalues and the number of such modes is finite (see Ref. 23, p. 366). This fact implies that increasingly accurate estimates of the propagation modes can be obtained by ever finer discretization of the equations. By contrast, finer discretizations introduce ever more negative eigenvalues corresponding to the continuum of radiation or unbound modes.

2.2 Finite element reduction

The Finite Element Method (FEM) can solve to desired accuracy the differential equations just derived. The present discussion specializes to the scalar wave equation [eq. (23)]; modifications needed for the equations containing gradient index terms are indicated in the appendix.

The solution function $f(r)$ is approximated by a piecewise linear function. This can be expressed in terms of interpolation functions (shown in Fig. 2) as

$$f(r) \cong \sum_{l=0}^{L+1} f(r_l) N_l(r), \tag{25}$$

where $r_l = l\delta$ for $l = 0, 1, \cdots, L + 1$ are evenly spaced sample points.

The first and last terms of the series are affected by end conditions on $f(r)$. At the center ($r = 0$) of the fiber $f(r)$ and its radial derivative must be bounded and well defined; so $f(0) = 0$ when $m > 0$ and df/dr $(0) = 0$ or equivalently $f(0) = f(\delta)$ to first order in δ when $m = 0$.

At the other end, $R = L\delta$ represents a truncation radius in the cladding, and $R + \delta$ represents the truncated part in a way that will now be explained. Assuming that the cladding extends indefinitely, the solution there is

$$f(r) = aK_m(\eta r), \tag{26}$$

where K_m denotes the mth order modified Bessel function of the second kind,[24] a is an unknown coefficient,

$$\eta = (\beta^2 - k_{\text{cl}}^2)^{1/2}, \tag{27}$$

and m denotes the azimuthal mode number used in the scalar wave

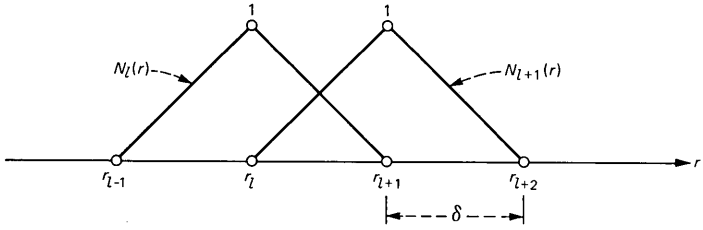


Fig. 2—Linear interpolation functions: hat functions.

equation [see eq. (23)]. It follows that the end condition for f at truncation is

$$f'/f = \eta K_m'(\eta R)/K_m(\eta R). \tag{28}$$

A Taylor's expansion of f to first order about $r = R$ yields

$$\begin{aligned} f(R + \delta) &= f(R) + \delta f'(R) \\ &= f(R)(1 + \delta \eta K_m'(\eta R)/K_m(\eta R)). \end{aligned} \tag{29}$$

Most modes of a multimode fiber are unaffected by the end condition in eq. (29) because their waveforms are negligibly small beyond the core. But the waveform of a single-mode fiber can extend beyond the core and then the end condition needs to be enforced.

Sample values of $f(r)$, the coefficients in eq. (25), are estimated by the Galerkin weighted residual method.[22] Although a piecewise linear approximation cannot satisfy the scalar wave equation, it can satisfy weighted averages of the equation. In the Galerkin technique the weightings are chosen as the basis functions $N_j(r)$ $j = 1, \cdots, L$. In terms of inner products, defined for any functions $p(r)$ and $q(r)$ as

$$(p, q) \equiv \int_o^{R+\delta} p(r)q(r)r\,dr, \tag{30}$$

Galerkin's technique yields

$$-\left(\frac{df}{dr}, \frac{dN_j}{dr}\right) - m^2\left(\frac{f}{r^2}, N_j\right) + \left((k^2 - k_{\text{cl}}^2)f, N_j\right) = (\beta^2 - k_{\text{cl}}^2)(f, N_j) \tag{31}$$

for $j = 1, \cdots, L$. The first term comes from an integration by parts.

Substituting the piecewise linear approximation of f [from eq. (25)] into the Galerkin equations gives L equations for $L + 2$ sample values, but the end conditions eliminate two of these values. The result is L equations in L unknowns, expressed as the matrix equation,

$$(A - m^2B + C)\mathbf{f} = \delta^2(\beta^2 - k_{\text{cl}}^2)D\mathbf{f}. \tag{32}$$

The column vector $\mathbf{f}$ denotes the sample values,

$$\mathbf{f} = [f(r_1), \cdots, f(r_L)]^t. \tag{33}$$

The $L \times L$ matrix A has jl element

$$a_{jl} = -\left(\frac{dN_l}{dr}, \frac{dN_j}{dr}\right)\delta^2 \tag{34}$$

for all j and l except, to accommodate the end conditions, a_{10} must be added to a_{11} when $m = 0$ and $a_{L,L+1}f_{L+1}/f_L$ must be added to a_{LL} for all

m. The $L \times L$ matrices B, C, and D have jl elements

$$b_{jl} = \left(\frac{1}{r^2} N_l, N_j\right) \delta^2,$$

$$c_{jl} = ((k - k_{\text{cl}}^2)N_l, N_j)\delta^2, \tag{35}$$

and

$$d_{jl} = (N_l, N_j).$$

The latter three matrices are evaluated by "lumping the masses," meaning that the integrals are evaluated numerically and the integration points are chosen as the sample points.[25] The trapezoidal rule then yields 0 for the off-diagonal elements and for the main diagonal gives

$$b_{ll} = \delta^2/l \; c_{ll} = l[k^2(r_l) - k_{\text{cl}}^2]\delta^4 \; d_{ll} = l\delta^2 \tag{36}$$

for $l = 1, \cdots, L$. The matrix A can be evaluated exactly. It is symmetric and tridiagonal (i.e., $a_{jl} = 0$ if $|l - j| > 1$) and has

$$a_{ll} = -2l\delta^2 \; a_{l,l+1} = (l + \tfrac{1}{2})\delta^2 \tag{37}$$

for $l = 1, \cdots, L$. When $m = 0$, a_{11} is $-1.5\delta^2$.

Eq. (32) converts to a standard symmetric eigenvalue problem by multiplying both sides by the diagonal matrix $D^{-1/2}$ and putting

$$\mathbf{g} = D^{1/2}\mathbf{f}. \tag{38}$$

The result is

$$T\mathbf{g} \equiv D^{-1/2}(A - m^2B + C)D^{-1/2}\mathbf{g} = \delta^2(\beta^2 - k_{\text{cl}}^2)\mathbf{g}, \tag{39}$$

standard form relative to the vector $\mathbf{g}$.

The key matrix, T, is symmetric and tridiagonal. For $m \neq 0$

$$t_{ll} = -2 - \frac{m^2}{l^2} + \delta^2[k^2(r_l) - k_{\text{cl}}^2]$$

$$t_{l+1,l} = t_{l,l+1} = \frac{1}{2}\left[\left(\frac{l+1}{l}\right)^{1/2} + \left(\frac{l}{l+1}\right)^{1/2}\right] \tag{40}$$

for $l = 1, \cdots, L - 1$; for $m = 0$

$$t_{11} = -\tfrac{3}{2} + \delta^2[k^2(r_l) - k_{\text{cl}}^2]; \tag{41}$$

for all m

$$t_{LL} = -2 - \frac{m^2}{L^2} + \delta^2[k^2(r_L) - k_{\text{cl}}^2] + t_{L,L+1}g_{L+1}/g_L. \tag{42}$$

The end condition in eq. (29) combined with eq. (38) gives

$$g_{L+1}/g_L = \frac{L+1}{L}\,[1 + \delta\eta K_m'(\eta R)/K_m(\eta R)]. \tag{43}$$

Equation (39) represents the FEM reduction of the scalar wave equation. The other differential equations, as indicated in the appendix, reduce to the same form, with T symmetric and tridiagonal. The sample distance δ is discussed in Section III.

2.3 Calculation of the propagation modes

The propagation modes are calculated by solving for the positive eigenvalues (μ) of T. The associated propagation constants are

$$\beta = (\mu/\delta^2 + k_{\text{cl}}^2)^{1/2} \tag{44}$$

with effective index of refraction

$$n_e = \beta/(2\pi/\lambda) = \beta(c/\omega). \tag{45}$$

When the end condition is enforced, the T matrix depends on β in its (L, L) entry, but a simple iteration procedure yields the proper β. The associated waveforms are given by $\mathbf{f}$ or $\mathbf{g}$.

The modal delays per unit length (τ_g) are the reciprocal of the group velocities,

$$\tau_g = \frac{d\beta}{d\omega}. \tag{46}$$

An efficient calculation of these uses the formula,

$$\frac{d\beta}{d\omega} = \left(\frac{\omega}{\beta}\right)\left(\frac{d\beta^2}{d\omega^2}\right) = \left(\frac{c}{n_e}\right)\left[\frac{dk_{\text{cl}}}{d\omega^2} + \frac{1}{\delta^2}\left(\frac{dT}{d\omega^2}\,\mathbf{g}\right)\cdot\mathbf{g}\right], \tag{47}$$

where $\mathbf{g}$ is assumed to be normalized (i.e., $\mathbf{g}\cdot\mathbf{g} = 1$). This follows standard procedure for taking the derivative of an eigenvalue with respect to a parameter.[26] Equations (40), (41), and (42) for T imply that $dT/d\omega^2$ is a diagonal matrix. Using the equivalence between $\omega^2 d/d\omega^2$ and $-\lambda 2d/d\lambda^2$, eq. (47) becomes

$$\tau_g = \sum_{l=1}^{L}\left[n^2(r_l) - \lambda^2\frac{dn^2}{d\lambda^2}(r_l)\right] g_l^2/cn_e. \tag{48}$$

To form the T matrix, the index profile $n(r)$ must be available for any excitation wavelength (λ). Sellmeier expansions[27] of the form

$$n^2 = 1 + \sum_{i=1}^{3} A_i/[1 - (l_i/\lambda)^2] \tag{49}$$

have been fitted to measured values of refractive index over the range of wavelengths 0.8 μm to 1.5 μm for bulk samples of pure SiO_2, 13.5-percent Ge doped SiO_2, and 1 percent F doped SiO_2 (denoted here as

a, b, and c, respectively). Also, results in Ref. 28 indicate that the index is approximately linear with concentration. Therefore, the index profile of an optical fiber having a graded Ge dopant is taken as

$$n(r, \lambda) = a(\lambda) + C_n(r)[b(\lambda) - a(\lambda)], \tag{50}$$

where $C_n(r)$ denotes the concentration profile for Ge, relative to 13.5 percent. For dual dopants, Ge and F, the index profile is

$$n(r, \lambda) = a(\lambda) + C_n(r)[b(\lambda) - a(\lambda)] + \tilde{C}_n(r)[c(\lambda) - a(\lambda)], \tag{51}$$

where $\tilde{C}_n(r)$ denotes the concentration profile for F, relative to 1 percent.

The concentration profiles in these equations may be specified or may be deduced from the index profile at a reference wavelength. Once the concentration profiles are available, the index profile and its λ^2 derivative can be determined from eq. (50) or eq. (51) for any wavelength.

The T matrix can be expressed in dimensionless form by replacing the sample spacing by

$$\delta = R_1/L_1, \tag{52}$$

where R_1 denotes the core radius and L_1 the number of samples in the core. This gives

$$\delta^2[k^2(r) - k_{\text{cl}}^2] = (2\pi R_1/\lambda)^2(n^2(r) - n_{\text{cl}}^2)/L_1^2, \tag{53}$$

and from eq. (50), (with $a = n_{\text{cl}}$),

$$n^2(r) - n_{\text{cl}}^2 = C_n(n_1^2 - n_{\text{cl}}^2) + (C_n^2 - C_n)(n_1 - n_{\text{cl}})^2, \tag{54}$$

where now $C_n(r)$ represents the Ge concentration normalized with respect to n_1, the index at the center. The latter term in eq. (54) can usually be neglected because $n_1 \sim n_{\text{cl}}$ and often, $C_n(r) \sim 1$ for most r. Then the T matrix can be expressed in terms of the usual V number,

$$V = (2\pi R_1/\lambda)(n_1^2 - n_{\text{cl}}^2)^{1/2} = R_1(k_1^2 - k_{\text{cl}}^2)^{1/2} \tag{55}$$

and the effective V number,

$$V_e = (2\pi R_1/\lambda)(n_e^2 - n_{\text{cl}}^2)^{1/2} = R_1(\beta^2 - k_{\text{cl}}^2)^{1/2} \tag{56}$$

for the (L, L) element. The FEM reduction becomes

$$T(V, V_e)\mathbf{g} = (V_e/L_1)^2\mathbf{g}, \tag{57}$$

where now an iteration on V_e is required. A similar expression, involving three V numbers, holds for dual dopants.

The effective index (n_e) is obtained from V_e in eq. (56). The delay

is

$$\tau_g = \frac{\omega}{\beta}\frac{d\beta^2}{d\omega^2} = \frac{1}{R_1^2}\frac{dV_e 2}{dV^2}\frac{dV^2}{d\omega^2} + \frac{dk_{\text{cl}}^2}{d\omega^2}$$

$$= \frac{1}{cn_e}\left\{n_{\text{cl}}^2 - \lambda^2\frac{dn_{\text{cl}}^2}{d\lambda^2} + \frac{V_e}{V}\frac{dV_e}{V}\right.$$

$$\left.\cdot\left[(n_1^2 - n_{\text{cl}}^2) - \lambda^2\frac{d}{d\lambda^2}(n_1^2 - n_{\text{cl}}^2)\right]\right\}. \tag{58}$$

Matrix T is always symmetric and tridiagonal. The routine BISECT in the EISPACK[1] library computes the positive eigenvalues for such matrices and TINVIT, also in EISPACK, computes the associated eigenvectors **g**. These routines operate efficiently and reliably.

2.4 Leaky modes

When the index profile drops below n_{cl} over part of its range, then leaky modes may exist. These modes have $n_e < n_{\text{cl}}$ and complex propagation constants. The corresponding attenuation accounts for radiation loss as the waveform spreads out radially. As is well known, leaky modes also can arise when $m \neq 0$ from the negative term, $-(m^2/r^2)$, in the propagation equation.

Leakage loss is calculated by truncating the waveform at the beginning of the outer cladding and enforcing the proper end condition. In terms of the complex propagation constant γ, eq. (27) for η changes to

$$\eta = (-\gamma^2 - k_{\text{cl}}^2)^{1/2}, \tag{59}$$

and the end condition [in eq. (43)] becomes complex. Now, the T matrix has a real (T_r) and an imaginary (T_I) part, but matrix T_I consists of all zeroes except the (L, L) diagonal element (call it x). First-order perturbation theory estimates an eigenvalue of T as

$$\mu = \mu_r + ixg_L^2 = \delta^2\eta^2, \tag{60}$$

where μ_r is an eigenvalue of T_r and g_L is the Lth component of the associated normalized eigenvector **g**. Again, an iteration is required because T depends on γ.

III. EXAMPLES OF MODAL CALCULATIONS

In this section modal calculations based on the results of Section II are illustrated in 10 examples. The next section summarizes and evaluates the results.

3.1 *Infinite parabolic profile*

The parabolic profile that extends without radial limit has index

$$n(r) = n_1[1 - 2\Delta(r/R)^2]^{1/2} \tag{61}$$

for all r, where R denotes the nominal radius of the fiber core. Under the scalar wave approximation, its modal waveforms are Laguerre Gaussian functions, its propagation constants are

$$\beta_{\mu m} = [k_{cl}^2 n_1^2 - 2Qk_{cl}n_1(2\Delta)^{1/2}/R]^{1/2}, \tag{62}$$

and its modal delays (neglecting material dispersion) are

$$\tau_{\mu m} = (\beta_{\mu m} + k_{cl}^2 n_1^2/\beta_{\mu m})/2\omega, \tag{63}$$

where the mode number is

$$Q = 2\mu + m + 1 \qquad m, \mu = 0, 1, \cdots, \tag{64}$$

with m the angular harmonic and μ the radial harmonic. Each Q represents a group of modes that have the same propagation constant and delay (see Ref. 9).

Modal delays were calculated assuming parameter values $\Delta = 0.013$, $R = 25$ μm, and $\lambda = 1.3$ μm. Convergence with respect to truncation radius (TR) was complete (within 0.1 ps/km) for each mode for $TR = 1.5R$. Convergence with respect to the number (L_1) of sample points in the core was within 1 ps/km in the rms delay for $L_1 = 400$, as indicated in Table I. An accuracy of 1 ps/km determines the bandwidth within 5 percent for a 10-GHz × km fiber and 0.5 percent for a 1-GHz × km fiber. The percent errors of the computed delays (with $L_1 = 500$) are shown in Fig. 3.

The most accurately computed mode within a mode group had radial number $\mu = 0$; their waveforms have no zero-crossings and are most easily approximated by piecewise linear functions. The least accurately computed mode (which was off by 5.5 ps/km) had the largest μ. As another measure of accuracy, the spread in the computed delays for each Q is also given in Table I.

To test the single-mode case, the parameters were changed to $\Delta =$

Table I—Convergence results for infinite parabolic profile in units of ps/km*

L_1	τ_{RMS}	Spread (1)	Spread (2)
100	144.0	136.1	96.5
200	129.4	33.9	23.7
300	127.1	15.1	10.5
400	126.3	8.5	5.9
500	125.9	5.5	3.8

* ($\Delta = 0.013$, $R = 25$ μm, $\lambda = 1.3$ μm)

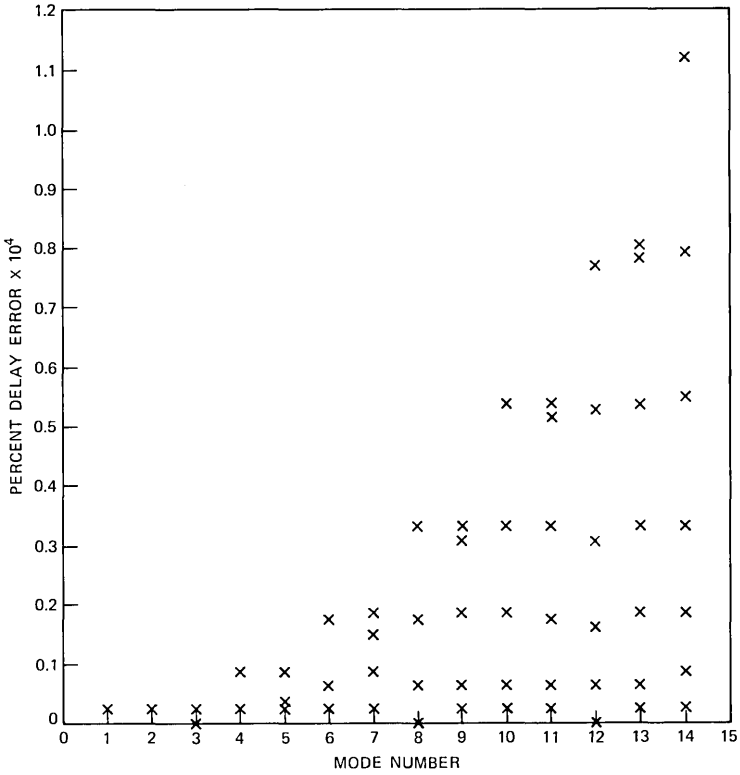


Fig. 3—Computation errors in modal delays for infinite parabolic profile.

0.005 and $R = 5$ μm, but again $\lambda = 1.3$ μm. The calculated effective index of the HE_{11} mode was accurate to seven decimal places and the delay to 0.1 ps/km when TR was $2.5R$ and L_1 was 300. This precision translates to nm accuracy in the zero-dispersion wavelength. The beam radius (BR) defined by the condition

$$f^2(BR) = f^2(0)/e \tag{65}$$

was accurate to four decimal places, as indicated in Table II. The number of samples in the core can be less in the single-mode case because the HE_{11} waveform does not oscillate; the truncation radius needs to be greater because the waveform extends farther into the cladding.

3.2 Parabolic fiber

Power-law fibers have the index profile

$$n(r) = \begin{cases} n_1(1 - 2\Delta(r/R)^\alpha)^{1/2} & r \le R \\ n_1(1 - 2\Delta)^{1/2} & r \ge R. \end{cases} \tag{66}$$

Assuming parameter values of $\Delta = 0.013$, $R = 25$ μm, $\lambda = 1.3$ μm, and $\alpha = 2$ (the parabolic case), effective index (n_e) and delay (τ_g) were computed for each propagation mode under the scalar wave approximation. The rms delay (τ_{RMS}) converged to 1.793 ns/km within 0.3 percent and the rms delay with the two highest-order-mode groups deleted (τ'_{RMS}) converged to 113 ps/km within 0.9 percent when $L_1 = 300$ and $TR = 0.7R$.

Figure 4 shows n_e and τ_g for each mode arranged by mode group.

Table II—Modal quantities for HE_{11} mode of parabolic profile*

L_1	TR	n_e	τ_g (μs/km)	BR (μm)
300	$3R/2$	1.4531412	4.8581500	2.6561930
300	$2R$	1.4531608	4.8577058	2.6644067
400	$2R$	1.4531608	4.8577059	2.6644135
300	$5R/2$	1.4531609	4.8577024	2.6644397
Actual Values		1.4531609	4.8577025	2.6643516

* ($\Delta = 0.013$, $R = 25$ μm, $\lambda = 1.3$ μm)

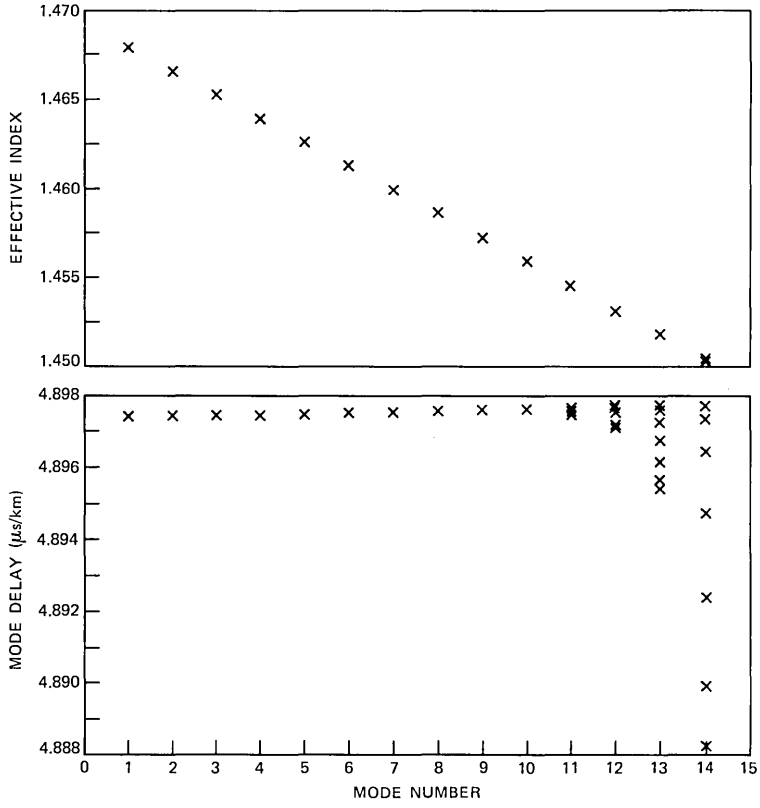


Fig. 4—Plots of n_e and τ_g versus mode number for parabolic fiber.

The delays spread almost 10 ns/km for $Q = 14$ and 3 ns/km for $Q = 13$, but less than 1 ns/km for the others. The spread in n_e is negligible, reaching a maximum of 0.01 percent.

Figure 5 shows n_e and τ_g plotted jointly on an expanded delay scale with outliers excised. To compare with WKB results, n_e and τ_g for the corresponding infinite parabola are indicated. The reference delays exceed all in their mode group: by as little as 0.2 ps/km for each of the first nine mode groups, but by more than 34 ps/km and 86 ps/km for $Q = 13$ and 14, respectively.

3.3 Contribution of gradient index terms

Gradient index terms cause modal delays to spread still further. Figure 6 shows the delay differences caused by these terms for each mode of the parabolic fiber. Of the 112 modal delays only 14 differed by more than 20 ps/km from the corresponding scalar wave values. The TM mode nearest cutoff had the largest difference (33 ps/km).

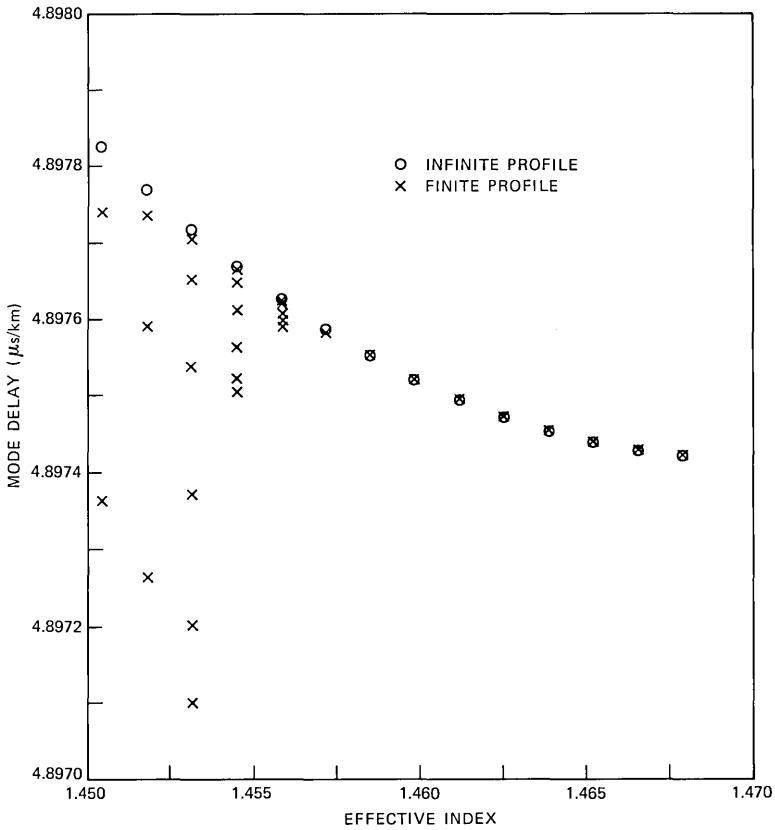


Fig. 5—Plots of τ_g versus n_e for cladded and infinite parabolic profiles.

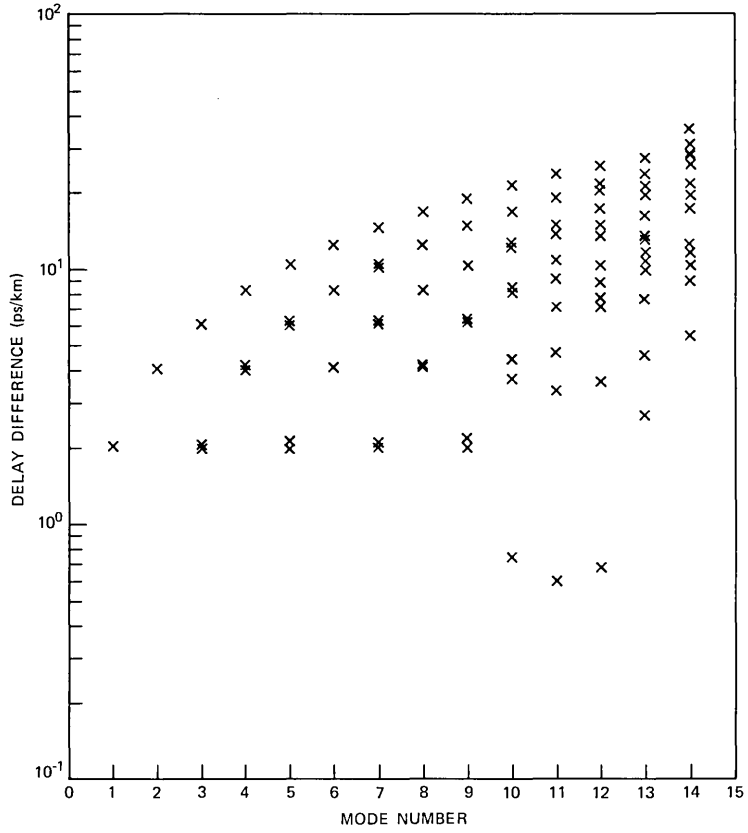


Fig. 6—Absolute differences in modal delays due to gradient index terms for parabolic fiber.

The change in τ_{RMS} was 5.9 ps/km and in τ'_{RMS} it was less than 1.5 ps/km. The error is less than 1.5 percent in either case. In the examples that follow gradient index terms are neglected.

3.4 Bandwidth of the parabolic fiber vs. λ

Material dispersion causes the bandwidth of power-law fibers to depend sharply on the excitation wavelength (λ). Here and in the following examples, the profile is assumed known at the reference wavelength of $\lambda = 0.8$ μm; the concentration profile and the index profile at other wavelengths are determined, as discussed in the previous section.

Bandwidth is defined as the half-power frequency of the transfer function over some distance. When intramodal dispersion is neglected, the transfer function is

$$G(\omega) = \sum_n a_n \exp(i\omega\tau_n), \tag{67}$$

where τ_n denotes the delay and a_n^2 the power of the nth mode. Defining the rms delay (τ_{RMS}) by

$$\tau_{\mathrm{RMS}} = \left[\sum_n a_n(\tau_n - \tau_{\mathrm{av}})^2 \Big/ \sum_n a_n\right]^{1/2} \tag{68}$$

and the average delay by

$$\tau_{\mathrm{av}} = \sum_n a_n\tau_n \Big/ \sum_n a_n, \tag{69}$$

then to second order in the $\omega(\tau_n - \tau_{\mathrm{av}})$, the bandwidth is

$$BW = 1/(2\pi\tau_{\mathrm{RMS}}), \tag{70}$$

in units of GHz × km for delays in ns/km.

Bandwidth was calculated on this basis over a range of wavelengths for the parabolic fiber of example 2, assuming the tapered modal-power distribution shown in Fig. 7. Lower-order modes are weighted more heavily than the higher and the highest are omitted as the higher modes are increasingly vulnerable to microbending and cladding absorption.

Figure 8 shows the spectral plot of bandwidth. The peak value occurs at $\lambda \simeq 0.963$ μm, compared to 0.983 μm reported in Ref. 13.

The spectral plot is also shown for the case where material dispersion is neglected. The figure shows that the bandwidth is lower and essen-

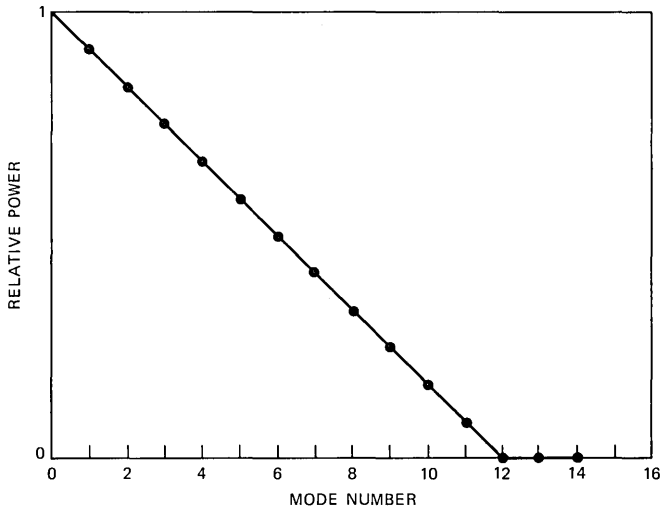


Fig. 7—A plot of the tapered modal-power distribution.

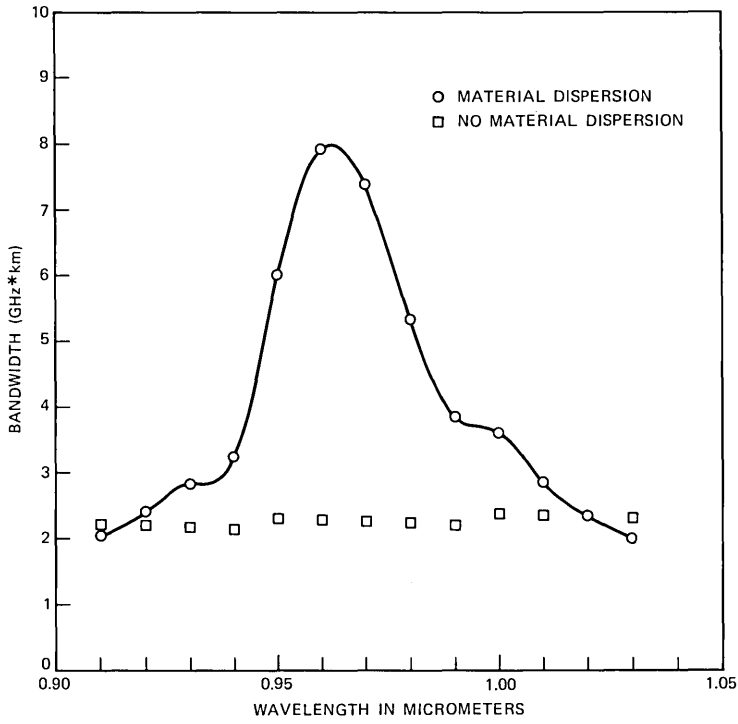


Fig. 8—Plots of bandwidth versus wavelength for parabolic fiber with and without material dispersion, using tapered modal-power distribution.

tially constant except for small discontinuities associated with new modes cutting in. Other calculation indicates that bandwidth increases to a peak at 9.5 GHz × km for $\alpha \simeq 1.98$ compared to an optimum $\alpha \simeq 1.97$ estimated by WKB analysis.[4]

3.5 Optimum α for power-law fibers

Material dispersion causes the optimum α of power-law fibers to depend on excitation wavelength. Figure 9 shows bandwidth versus α when $\Delta = 0.013$ and $R = 25$ μm for $\lambda = 0.82$ μm and $\lambda = 1.32$ μm. The optimum α's are 2.07 for the former and 1.88 for the latter, compared to values of 2.081 and 1.884, respectively, reported in Ref. 13. The peak bandwidths are 5.48 GHz × km and 6.02 GHz × km, respectively.

3.6 An alternate modal-power distribution

The fractional power that propagates in the cladding can be used to derive an alternate modal-power distribution. Modes with more than 0.1 percent of their power in the cladding will lose most of their power over 1 km for typical cladding losses of 1 dB/m. A realistic modal

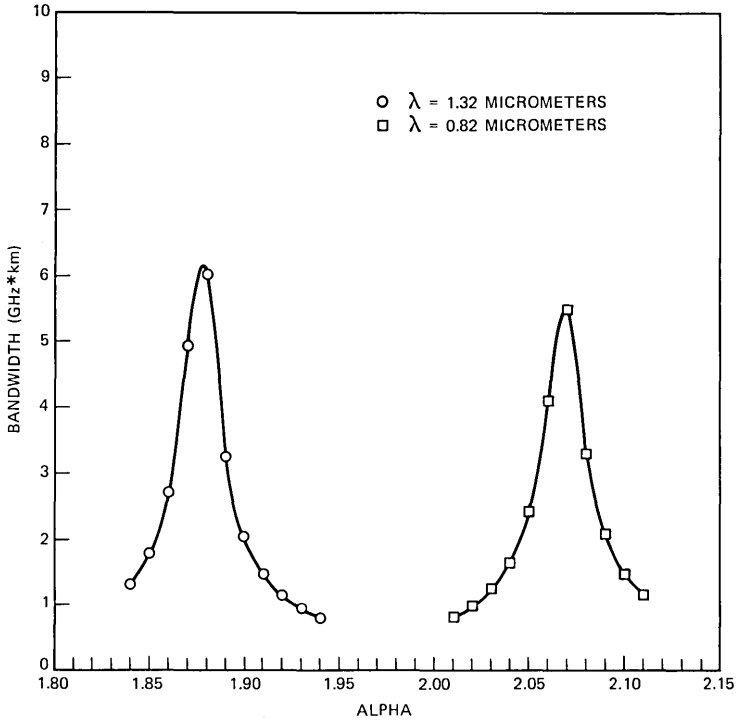


Fig. 9—Plots of bandwidth (in power-law fibers) versus α for $\lambda = .82$ μm and $\lambda = 1.32$ μm, using tapered modal-power distribution.

distribution would neglect these modes and for simplicity could weight the others equally.

Figure 10 shows bandwidth versus α for the two cases of example 5, assuming the new distribution. The optimum α is lower by 0.01 and the peak bandwidth is increased by 40 to 50 percent in either case. Table III indicates the modes neglected in the bandwidth calculation when $\lambda = 1.32$ μm and $\alpha = 1.88$.

3.7 Layer structure

Layer structure in actual fibers perturbs the ideal profile. In this example power-law profiles are approximated by steps that span equal areas and match the ideal profile at the mid-area points of the layers (see Fig. 11). Bandwidth at $\lambda = 1.32$ μm (using the modal distribution in example 6) is shown versus α and the number of steps (NS) in Fig. 12. The graphs indicate that the optimum α is essentially independent of NS, but the peak bandwidth and its sharpness decrease for smaller NS.

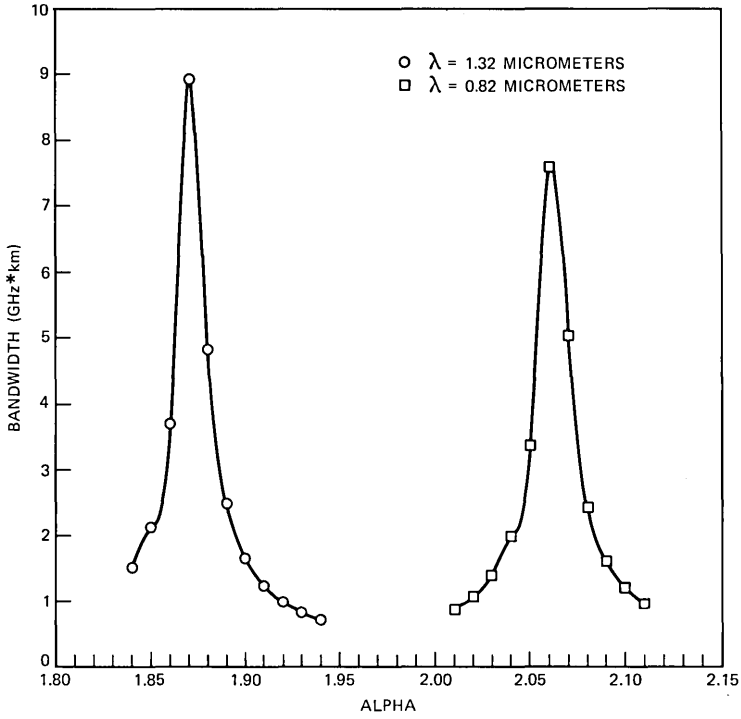


Fig. 10—Plots of bandwidth versus α for $\lambda = .82$ μm and $\lambda = 1.32$ μm assuming equal excitation of modes having less than 0.1 percent power in cladding.

3.8 Delay vs. λ for a single-mode fiber

Delay (per km) of the HE_{11} mode was computed versus λ for an actual single-mode fiber based on its measured index profile. The profile measurement was performed on the associated preform and is shown in Fig. 13. The depression of the inner cladding is due to F doping.

As the calculation presumes radially symmetric profiles, the right and left sides were considered separately. About 750 sample points were used in either case.

The radial scale of the fiber profile was assumed to depend linearly on that of the preform. The fiber core radius (R) was estimated from the preform core radius (RP), fiber diameter (DF), and preform diameter (DP) as

$$R = 1.01(RP)(DF/DP), \tag{71}$$

where the 1-percent addition estimates the SiO_2 loss in the outer cladding during the draw process.

Table III—Modes with less than 0.1 percent of their power in cladding*

	<0.1 Percent		>0.1 Percent	
Q	m	μ	m	μ
14	1	6		
	3	5		
13	0	6	12	0
	2	5		
	4	4		
	6	3		
	8	2		
	10	1		
12	1	5	11	0
	3	4		
	5	3		
	7	2		
	9	1		
11	0	5	8	1
	2	4	10	0
	4	3		
	6	2		
10	1	4	3	3
			5	2
			7	1
			9	0

* ($\Delta = 0.013$, $R = 25$ μm, $\lambda = 1.32$ μm, $\alpha = 1.88$)

Figure 14 shows computed differential delay versus λ for the right and left profiles together with measurements. It was assumed that measurement matched calculation exactly at $\lambda = 1.32$ μm, because only relative delays were measured. The zero-dispersion wavelength (at the delay minimum) was computed as $\lambda_o = 1.3127$ μm for the left profile and $\lambda_o = 1.3113$ μm for the right, compared to measurement of $\lambda_o = 1.3114$ μm.

3.9 Leakage loss vs. λ for a single-mode fiber

Leakage loss (in dB/m) of the TE mode of the single-mode fiber in the preceding example was calculated for both sides of the profile as a function of λ and is shown in Fig. 15. The difference between the left and right profiles becomes evident in this figure. A loss of 4 dB/m has been identified[29] with effective cutoff, and corresponds to cutoff wavelengths, $\lambda_c = 1.17$ μm for the left profile and $\lambda_c = 1.23$ μm for the right, compared to a measurement of 1.28 μm. This discrepancy is discussed in the next section.

The equivalent step approximation (where the index of the core and the inner cladding are area-weighted averages of the profile) gives $\lambda_o = 1.308$ μm and $\lambda_c = 1.21$ μm. The λ_c value straddles the previous calculated values, but the λ_o value is somewhat less.

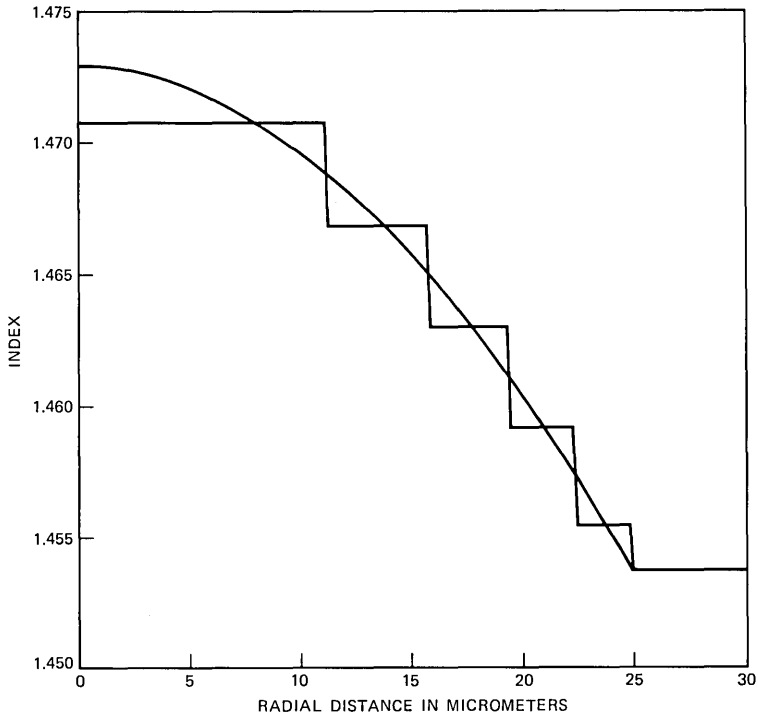


Fig. 11—A plot of a multilayered fit to an α profile (NS = 5, α = 1.90).

3.10 Design curves for single-mode fibers

The dimensionless formulation allows design curves to be generated efficiently. Figure 16 shows V_e versus V for the HE_{11} mode of three power-law profiles. The specific parameters (R, Δ, λ) determine V and, hence, V_e and dV_e/dV. Delay versus λ is computed according to eq. (58) and λ_o is estimated at the minimum of the delay curve. Figure 17 shows λ_o versus R for two values of Δ for triangular profiles ($\alpha = 1$).

IV. SUMMARY AND CONCLUSIONS

A method for computing modes of a circular optical fiber has been presented. The finite element method reduces Maxwell's equations to the standard eigenvalue problem, involving tridiagonal matrices. Routines from EISPACK exploit the tridiagonal form to compute the eigenvalues and eigenvectors efficiently. From these the modal quantities are obtained.

Using a piecewise linear approximation of the waveform is necessary to get tridiagonal form. Although piecewise quadratic and cubic approximations of the waveform can lead to smaller matrices, tridiagonal

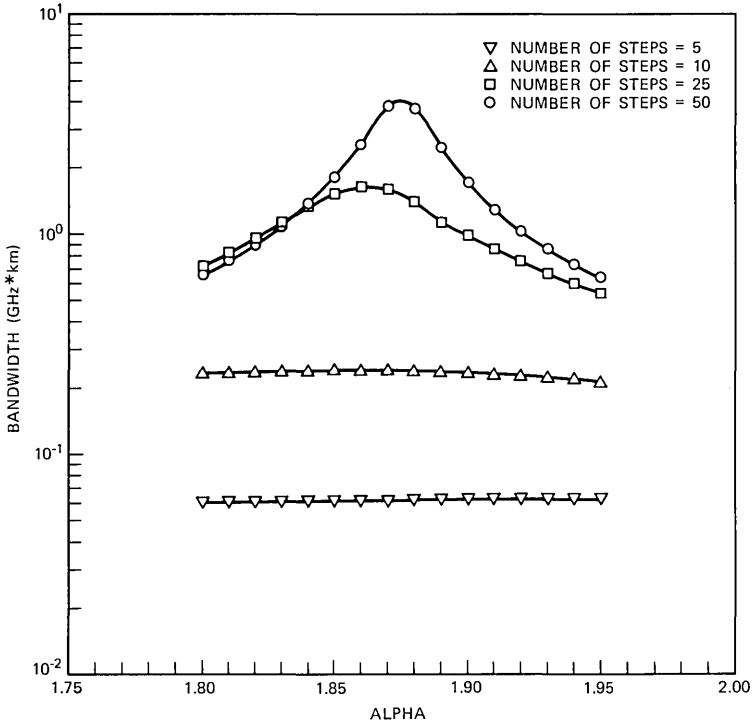


Fig. 12—Plots of bandwidth versus α at $\lambda = 1.32$ μm for four values of NS.

form is lost and the eigencalculations would be less efficient. Extension to elliptical and other nonradially symmetric fibers leads to similar difficulty in the eigen calculations.

The method applies to any circular fiber and can account for gradient index terms to first order. As illustrated in the 10 examples of Section III, the calculations can provide information on the radial distribution of propagating power, on pulse dispersion (arising from material, intermodal, and intramodal dispersion), and on leakage loss.

The accuracy of the method was tested for the infinite parabolic profile of the first example. After convergence was established, the modal quantities were compared with the actual values known through analysis. Agreement was excellent for both a multimode and a single-mode profile.

The WKB and scalar wave approximations were tested for the parabolic fiber in the next two examples. The cladding of the fiber altered the delays of the higher-order modes substantially, a facet missed by WKB analysis. The gradient index terms contributed less than 1.5 percent to the rms delay; and so, for most purposes the scalar wave approximation suffices for this fiber. The validity of these

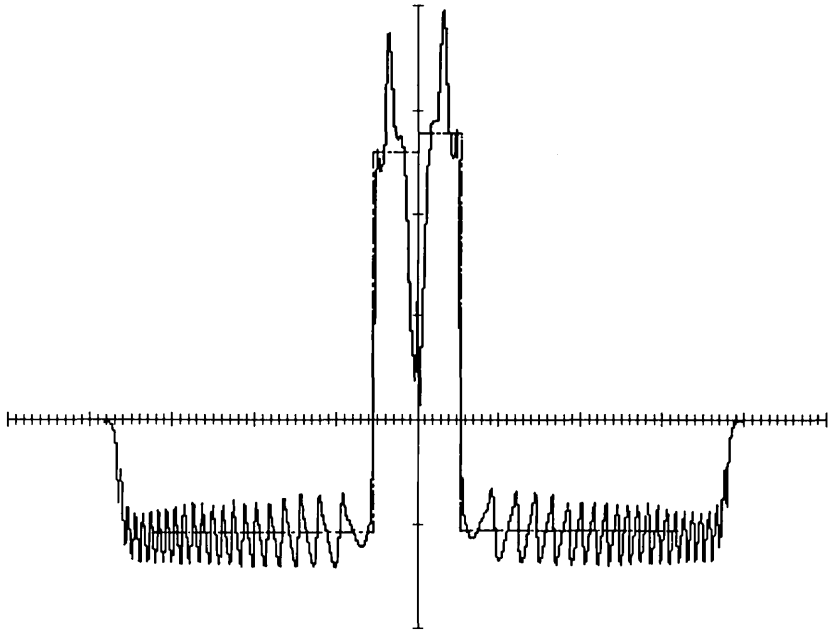


Fig. 13—The preform profile for a single-mode optical fiber.

approximations may change in other fibers; the value of the method is that it permits the test.

Material dispersion was incorporated into the calculation in the remaining examples. The variation of refractive index with wavelength and dopant concentration was modeled by Sellmeier expansions based on measurements of bulk samples with certain dopant concentration.[27] A linear dependence of index on concentration was assumed, but is not essential to the method. Irreversible thermal and stress effects[30] might also be incorporated.

Bandwidth was estimated in examples 4 and 5 for power-law fibers assuming a tapered modal-power distribution. With the higher-order mode groups deemphasized, results essentially agreed with WKB analysis. The usual sharp peak in bandwidth occurred in the spectral plot for the parabolic ($\alpha = 2$) fiber in example 4 and in the α-dependence in example 5. Calculation of optimum α agreed reasonably well in these cases with other numerical work and WKB calculations.

The bandwidth calculation was repeated in example 6 using a different modal-power distribution. Modes with more than 0.1 percent of their power in the cladding were neglected, as being too lossy to maintain power. The optimum α's were essentially the same, but peak

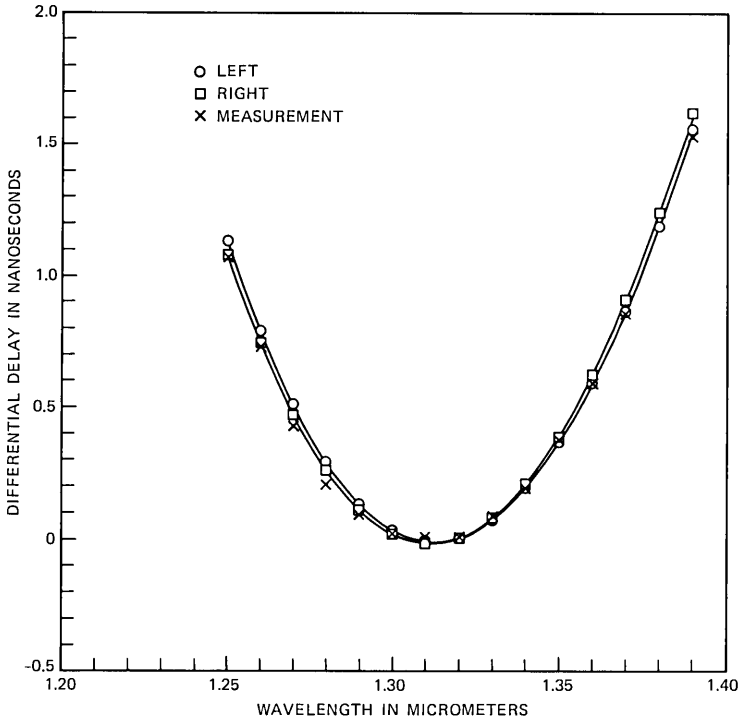


Fig. 14—Measurement and calculations for left and right profiles of differential delay versus wavelength.

bandwidths were somewhat higher. This example could be expanded to simulate differential mode attenuation, mode mixing, concatenation of dissimilar fibers, and other longitudinal variations.

Perturbations of α profiles usually lower bandwidth. Example 7 concerned the effect of layer structure on bandwidth and showed the expected decline with a smaller number of layers. Other perturbations such as ripple can also be treated. As a measure of its efficiency, the method used about two seconds per profile on the *Cray-1* for this example.

The last three examples concerned single-mode fibers. The first two of these involved an actual fiber whose profile was measured in the preform stage. Calculation of delay (per km) versus λ matched measurement extremely well, but calculated leakage loss exceeded measurement, giving a 7-percent average underestimate of cutoff wavelength (λ_c). Other calculations of leakage involving only slightly depressed inner claddings, where λ_c matches measurement to within 1 percent, suggest that the discrepancy may involve material changes in the F-doped glass caused by the draw process (consistent with obser-

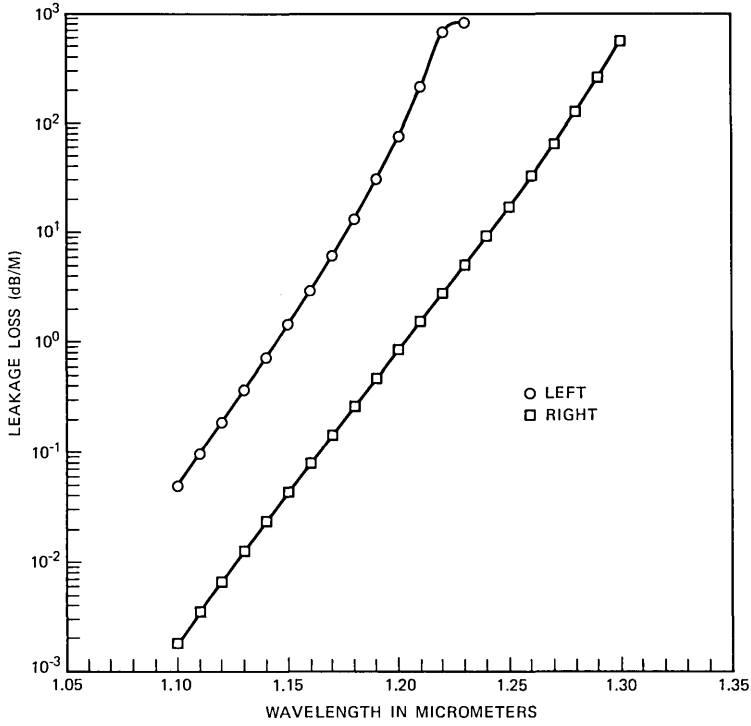


Fig. 15—Plots of computed leakage loss versus wavelength for left and right profiles.

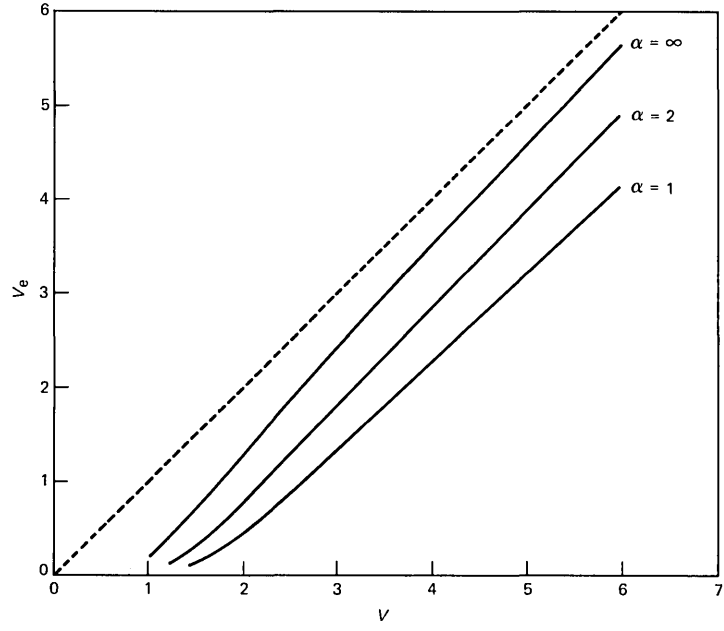


Fig. 16—Plots of V_e versus V for the HE_{11} mode of three power-law profiles.

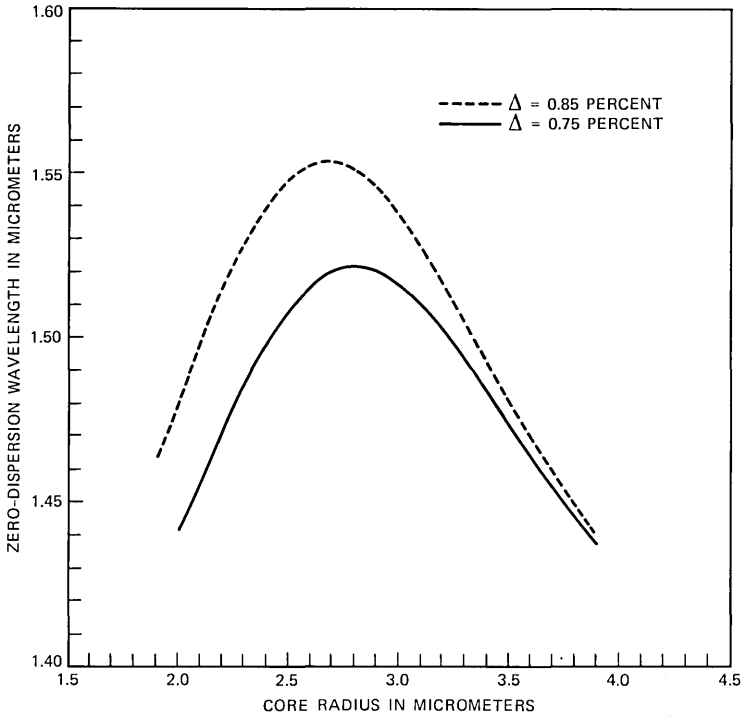


Fig. 17—Plots of computed zero-dispersion wavelength (λ_o) versus core radius (R) for two values of Δ, for triangular profiles.

vations in Ref. 31). Such changes need to be understood when predicting transmission performance from a preform profile. Similar values for λ_o and λ_c were obtained for the equivalent step profile (with depressed inner cladding), but the lower λ_o indicates the effect of profile structure.

The dimensionless formulation permits the greatest efficiency in getting design curves. Zero-dispersion wavelength (λ_o) for triangular profiles is calculated in example 10 showing the expected shift[32] to higher wavelengths. Spot size or cutoff wavelength can be obtained with similar efficiency.

In summary, the calculation method is reliable, and relatively inexpensive. In the context of circular fibers it is comprehensive, capable of simulating diverse effects.

V. ACKNOWLEDGMENTS

The author is grateful for useful discussions with I. A. White and A. J. Ritger on multimode fibers, W. T. Anderson and P. F. Glodis on

single-mode fibers, D. S. Burnett on the FEM, and L. Kaufman on EISPACK.

The author is also grateful to J. S. Nobles and D. L. Philen for fiber measurements and to H. W. Friedrichsen for programming support.

REFERENCES

1. B. T. Smith et al., *Matrix Eigensystem Routines—EISPACK Guide*, New York: Springer Verlag, 1976.
2. P. M. Morse and H. Feshback, *Methods of Theoretical Physics*, New York: McGraw-Hill, 1953, p. 1092.
3. J. E. Midwinter, *Optical Fibers for Transmission*, New York: Wiley, 1979, Chapter 6.
4. D. Gloge and E. A. J. Marcatili, "Multimode Theory of Graded-Core Fibers," B.S.T.J., *52,* No. 9 (November 1973), pp. 1563–78.
5. C. Pask, "Exact Expressions for Scalar Modal Eigenvalues and Group Delays in Power-Law Optical Fibers," J. Opt. Soc. Am., *69,* No. 11 (November 1979), pp. 1599–1603.
6. G. A. E. Crone and J. M. Arnold, "Anomalous Group Delay in Optical Fibers," Opt. Quant. Elect., *12,* No. 6 (November 1980), pp. 511–7.
7. E. Snitzer, "Cylindrical Dielectric Waveguide Modes," J. Opt. Soc. Am., *51,* No. 5 (May 1961), pp. 491–8.
8. S. Kawakami and S. Nishida, "Characteristics of a Doubly Clad Optical Fiber with a Low-Index Inner Cladding," IEEE Trans., *QE-11* (December 1974), pp. 879–87.
9. D. Marcuse, *Light Transmission Optics*, New York: Van Nostrand, 1972, pp. 267–72.
10. R. A. Sammut and A. K. Ghatak, "Perturbation Theory of Optical Fibers with Power-Law Core Profile," Opt. Quant. Elect., *10,* No. 6, (November 1978), pp. 475–82.
11. G. E. Peterson, A. Carnevale, U. C. Paek, and D. W. Berreman, "An Exact Numerical Solution to Maxwell's Equations for Lightguides," B.S.T.J., *59,* No. 7 (September 1980), pp. 1175–96.
12. G. E. Peterson, A. Carnevale, and U. C. Paek, "Comparison of Vector and Scalar Modes in a Lightguide with a Hyperbolic Secant Index Distribution," B.S.T.J., *59,* No. 9 (November 1980), pp. 1681–91.
13. G. E. Peterson, A. Carnevale, U. C. Paek, and J. W. Fleming, "Numerical Calculation of Optimum α for a Germania-Doped Silica Lightguide," B.S.T.J., *60,* No. 4 (April 1981), pp. 455–70.
14. U. C. Paek, G. E. Peterson, and A. Carnevale, "Dispersionless Single-Mode Lightguides with α Index Profiles," B.S.T.J., *60,* No. 5 (June 1981), pp. 583–98.
15. U. C. Paek, G. E. Peterson, and A. Carnevale, "Electromagnetic Fields, Field Confinement, and Energy Flow in Dispersionless Single-Mode Lightguides with Graded-Index Profiles," B.S.T.J., *60,* No. 8 (October 1981), pp. 1727–43.
16. Y. Kokubun and K. Iga, "Mode Analysis of Graded-Index Optical Fibers Using a Scalar Wave Equation Including Gradient-Index Terms and a Direct Numerical Integration," J. Opt. Soc. Am., *70,* No. 4 (April 1980), pp. 388–94.
17. W. L. Mammel and L. G. Cohen, "Numerical Prediction of Fiber Transmission Characteristics from Arbitrary Refractive-Index Profiles," Appl. Opt., *21,* No. 4 (February 1982), pp. 699–703.
18. S. J. Jang, L. G. Cohen, W. L. Mammel, and M. A. Saifi, "Experimental Verification of Ultra-Wide Bandwidth Spectra in Doubly-Clad Single-Mode Fiber," B.S.T.J., *61,* No. 3 (March 1982), pp. 385–90.
19. L. G. Cohen, W. L. Mammel, and S. Lumish, "Tailoring the Shapes of Dispersion Spectra to Control Bandwidths in Single-Mode Fibers," Optics Letters, *7,* No. 4 (April 1982), pp. 183–5.
20. K. Ikamoto and T. Okoshi, "Vectorial Wave Analysis of Inhomogeneous Optical Fibers Using Finite Element Method," IEEE Trans., *MTT-26* (February 1978), pp. 109–14.
21. C. Yeh, K. Ha, S. B. Dong, and W. P. Brown, "Single-Mode Optical Waveguides," Appl. Opt., *18,* No. 10 (May 1979), pp. 1490–504.
22. D. S. Burnett, *Finite Element Analysis, From Concept to Applications*, New York: Addison-Wesley, to be published.
23. M. Reed and B. Simon, *Methods of Modern Mathematical Physics IV: Analysis of Operators*, New York: Academic Press, 1978.

24. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Washington, D. C.: National Bureau of Standards, 1970, Chapter 9.
25. O. C. Zienkiewiez, *The Finite Element Method, 3rd Edition*, England: McGraw-Hill, 1977, p. 537.
26. T. Kato, *Perturbation Theory for Linear Operators*, New York: Springer-Verlag, 1966, p. 125 and p. 391.
27. J. W. Fleming, "Material Dispersion in Lightguide Glasses," Elect. Lett., *14*, No. 11 (May 1978), pp. 326–8.
28. S. E. Miller and A. G. Chynoweth, ed., *Optical Fiber Telecommunications*, New York: Academic Press, 1979, Chapter 7, p. 188.
29. W. T. Anderson, private communication.
30. P. L. Chu and T. Whitbread, "Measurement of Stresses in Optical Fiber and Preform," Appl. Opt., *21*, No. 23 (December 1, 1982), pp. 4241–5.
31. M. J. Saunders, "A Comparison of Single-Mode Refractive Index Profiles in Preforms and Fibers," 8th ECOC, paper C-17 (September 1982).
32. K. I. White, "Design Parameters for Dispersion-Shifted Triangular-Profile Single-Mode Fibers," Elect. Lett., *18*, No. 17 (August 19, 1982), pp. 725–7.

APPENDIX

Gradient Index Terms

Gradient index terms, involving the quantity $q_r = d/dr \ln k^2$, occur in eqs. (10) and (21). Although the index or its derivative may be discontinuous, they can be approximated by smooth functions, so k^2 and g_r can be taken as well defined, continuous functions of r. This appendix concerns the changes needed in the T matrix to accommodate these terms.

Equation (10) converts to symmetric form when $h_\theta = kf$ and both sides are divided by k. The result is

$$\left(\frac{k}{r}\frac{d}{dr}\frac{r}{k^2}\frac{d}{dr}k - \frac{1}{r^2} - \frac{g_r}{r} + k^2\right)f = \beta^2 f. \tag{72}$$

Applying the Galerkin technique gives for the first term,

$$\begin{aligned} A' \sim a'_{jl} &= -\left[\frac{1}{k^2}\frac{d}{dr}(kN_l), \frac{d}{dr}(kN_j)\right]\delta^2 \\ &= \left[\left(\frac{dN_l}{dr}, \frac{dN_j}{dr}\right) + \frac{1}{4}(g_r^2 N_l, N_j)\right. \\ &\quad \left. + \frac{1}{2}\left(g_r \frac{dN_l}{dr}, N_j\right) + \frac{1}{2}\left(g_r N_l, \frac{dN_j}{dr}\right)\right]\delta^2 \end{aligned} \tag{73}$$

in place of matrix A specified in eq. (34). The first quantity is a_{jl} as before, the second adds to the C matrix of eq. (35), and the last two when evaluated by the trapezoidal rule contribute only to the first side diagonals. The term g_r/r is handled in the same way as k^2. Combining these contributions gives a new T matrix that is symmetric and tridiagonal.

Likewise, eq. (21) converts to symmetric form when $k^{1/2}f$ replaces f.

The rest proceeds in the same way to give a symmetric, tridiagonal T matrix.

Material dispersion is incorporated by expressing the index in terms of concentration functions as before. Radial derivatives needed for the gradient index terms themselves involve derivatives of the concentration function. The ω^2 derivatives of these terms, needed for the modal delays, are found by differentiating the coefficients of the concentration functions.

AUTHOR

Terrence A. Lenahan, S.B. and S.M. (Electrical Engineering) 1964, Massachusetts Institute of Technology; Ph.D. (Applied Mathematics) 1970, University of Pennsylvania; Bell Laboratories, 1970—. Mr. Lenahan has done studies of various areas of mathematical physics including electromagnetic propagation, elasticity, and fluid mechanics. He has recently been interested in the propagation of light in optical fibers. Member, American Mathematical Society, SIAM, Sigma Xi.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 9, November 1983
Printed in U.S.A.

Measurements of 800-MHz Radio Transmission Into Buildings With Metallic Walls

By D. C. COX,* R. R. MURRAY,* and A. W. NORRIS*

(Manuscript received January 5, 1983)

In this paper we describe an experiment conducted to measure 800-MHz attenuation into buildings. This information is needed for refining the configuration and design of portable radiotelephone systems that will accommodate low-power portable sets. Signal levels have been measured in and around three small buildings and a house, using an instrumentation van with an erectable 27-foot-high antenna. These buildings and the house all have metallic materials in their walls and thus are expected to exhibit high attenuation. The van was parked at one location with respect to the three buildings, and at nine different locations, at distances ranging from 400 feet to 1600 feet, from the house. We found that small-scale signal envelope variations are approximately Rayleigh distributed. For the house, large-scale distributions of the small-scale signal medians are approximately log-normal. Median signal levels outside the house decrease as $d^{-4.5}$, where d is the distance from the van. Inside the house, levels decrease as $d^{-3.9}$ for first-floor locations and as $d^{-3.0}$ for second-floor locations. Average signal levels at 1000 feet, relative to free space, are −12.5 dB outside, −18.5 dB for the first floor, −16.5 dB for the second floor, and −28.9 dB for the basement. Other statistics of the signal levels and of attenuation into the house are also given in the paper. Cross-polarization couplings of −10 dB to 0 dB were measured in and around the three small buildings. The small-scale signal medians inside the three buildings range from 2 dB above to 24 dB below the averages of the signal medians outside buildings.

I. INTRODUCTION

The attenuation of radio signals propagating into buildings has a significant effect on the performance of portable radiotelephone sys-

* Bell Laboratories.

tems.[1] Frequencies in the 800-MHz to 900-MHz range are good candidates for such systems. Most earlier propagation measurements at these frequencies in the shadowing and multipath environment around buildings were made for mobile radio systems.[2] Ranges for the mobile radio data are generally greater than one mile. Battery power limitations on portable systems will likely restrict such systems to ranges on the order of 1000 feet. Earlier measurements made in buildings are limited in scope,[3-6] are oriented towards lower frequencies,[7-11] or are directed toward high-power portable sets with ranges greater than one mile.

One important portable radiotelephone environment comprises suburban residential areas characterized by discrete houses and other small buildings with densities ranging from less than one house per acre to a few houses per acre. Houses and small buildings with metallic materials in their walls are expected to exhibit high extreme values of attenuation.

In the residential environment, fixed radio terminals that communicate with portable sets could be placed at convenient locations outside buildings. These fixed terminals will be referred to as Portable Radiotelephone Terminals or PORTs.

We have implemented an experiment to provide 800-MHz attenuation information needed for refining the configuration and design of portable radiotelephone systems that will accommodate low-power portable sets. Measurements were made in and around three small buildings and a house. The buildings and the house all have metallic materials in their walls. An instrumentation van was parked at different locations to simulate different PORTs with distances ranging from 200 to 1600 feet. A 27-foot-high erectable antenna on the van simulated an unobtrusive PORT antenna. A portable signal source was moved in and around a building and signal levels were received and recorded in the van.

Section II of this paper describes the instruments used and how measurements were taken. Section III summarizes building attenuations and cross-polarization couplings measured in the three small buildings. Section IV contains statistical descriptions of the building attenuation for the house.

II. THE MEASUREMENTS

2.1 Instrumentation

2.1.1 Signal source

The portable signal source is a modified 815-MHz* handie-talkie.

* The actual frequency of the measurements is 815 MHz; however, the statistical results are not sensitive to small changes in frequency. Therefore, when frequency is referred to relative to the measurements, it will be rounded to 800 MHz.

The transmitting antenna is a half-wavelength coaxial sleeve dipole attached to the top of the hand-held unit. The dc power is provided by a self-contained nickel cadmium battery through a voltage regulator. The regulator minimizes output power drift due to normal battery discharge. The transmitter output is 0.8 watt. The output varies less than 0.3 dB and 700 Hz over continuous one-hour periods that include ambient temperature changes of 0°C to 25°C.

2.1.2 Instrumentation van

The instrumentation van (movable PORT) is a modified motor home, containing a 5-kw ac generator. An uninterruptable power supply isolates the instrumentation from generator voltage fluctuations that otherwise could affect measurement accuracy. The van is shown in Fig. 1.

Two 27-foot antenna masts are installed in pivoting mounts so they can be stored horizontally for transport and can be erected at the test site. They are 14.2 feet apart and are adjusted to vertical using built-in bubble levels. An 815-MHz collinear receiving antenna is mounted on one mast. A bracket on the other mast holds the 815-MHz signal source to provide a reference signal for calibrating the receiver. The centers of the two antennas are at the same height when erected.

2.1.3 Measuring receiver

The measuring receiver is an 815-MHz frequency-modulated (FM) communications receiver, modified to detect the received signal envelope. The receiving antenna is a collinear array (four dipole elements, 5.8-dB gain over dipole, 18-degree vertical beam width) mounted vertically at the top of the tilt-over mast on the instrumentation van. In the receiver modification, an 11.7-MHz intermediate frequency (IF) output is extracted before the limiter, down-converted to 13 kHz, bandpass-filtered (BW_{3dB} = 8 KHz), and linearly detected. The modified receiver has a −123 dBm sensitivity for 0 dB output s/n from the linear envelope detector and a 45-dB measuring range between levels 6 dB above the noise level and 3 dB below saturation. The measuring range is linear within ±1 dB over a 35-dB range at high signal levels. The entire receiver is enclosed in a sealed brass box to provide radio frequency (RF) isolation. A variable RF attenuator reduces the input signal in 1-dB steps to prevent overloading of the receiver.

2.1.4 Data acquisition

The analog receiver output drives a 12-bit-resolution digital storage oscilloscope and an integral 5-1/4-inch flexible disc drive for data storage. The oscilloscope is set to record 2048 samples in a 20-second measurement period. Up to 16 tracks of 2048 samples each can be

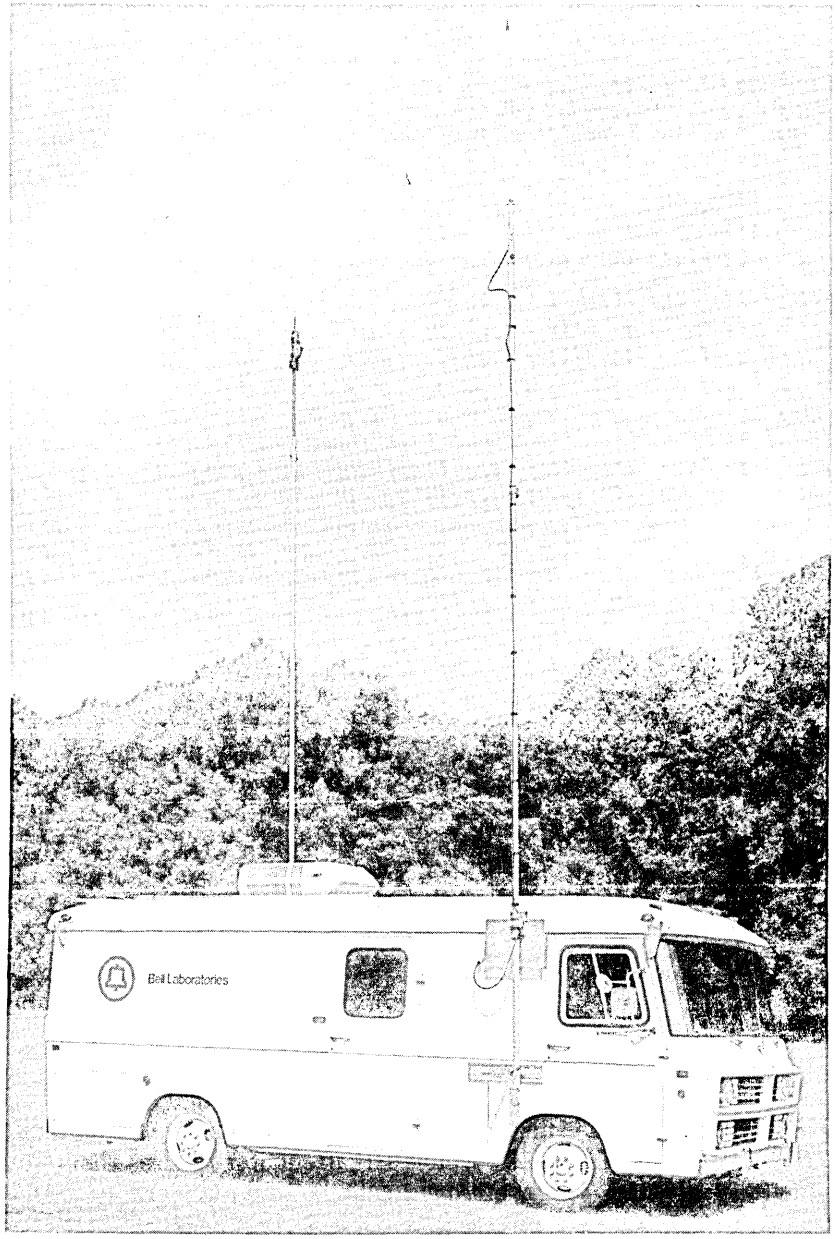


Fig. 1—The instrumentation van with the receiving antenna mast and the reference signal mast erected. The center of the four-element collinear receiving antenna is 27 feet above ground at the top of the mast mounted on the right side of the van near the front. The signal source is at the top of its mast mounted on the left side near the rear. The center of the signal-source dipole is also 27 feet above ground.

recorded on each disc. Data for each parked position of the van, i.e., one PORT location, are stored on a separate disc.

The recorded data are transferred to a desktop computer. At the time of transfer the following steps are performed:

1. Hand-recorded log information is appended to the signal strength data. The log information includes such items as the address of the house, the position of the van, the path azimuth, the path length, and the receiver input attenuator setting.
2. The data are scaled to convert the recorded signal voltages to dB relative to 0 dB at the reference location 14.2 feet from the receiving antenna. The scaling takes into account recorded dc offsets, recorded reference levels, and the receiver input attenuator settings.
3. Medians and cumulative distributions of signal level are calculated from the scaled data.
4. The scaled data are stored on flexible discs within the computer for further analysis.

2.2 Procedure

Van locations were selected using tax maps covering the immediate vicinity of the house. Points were chosen equally spaced in azimuth around the house for each of three radii at about 400, 800, and 1600 feet. Road layout and terrain irregularities influenced the final choices of vehicle placement. A single van location was used for the measurements in the other three small buildings.

Measurements were coordinated between the van and the measurement location over a 450-MHz voice link. The link comprises a 25-watt FM transceiver in the van and a 2-watt handie-talkie carried by the person making the measurements.

A typical procedure for measuring a building starts with the van parked at an appropriate position. The van location must be fairly level. If necessary, wooden ramps are put under wheels to aid in leveling. The portable transmitter is installed on its mast as a local reference source. The mast is erected and is plumbed to vertical. Similarly, the receiving antenna on the opposite side of the vehicle is erected and plumbed. Keeping both masts plumbed on the level van assures a fixed distance between the reference and receiving antennas for calibration purposes. The linear detector IF input is grounded, and the dc level is adjusted and recorded on a disc track. Next, the detector input is ungrounded and the receiver RF attenuator is adjusted so the RF reference level is within the receiver operating range. This level, which serves as a calibration reference at the fixed 14.2 foot distance, is then recorded.

The signal source is removed from the mast and taken to the building for signal level measurements. The unit is hand-held at arm's length,

for a scan height of 4.5 feet.[6] At a selected location within the building, a 20-second raster scan is made by moving the transmitter in a horizontal plane at 2.5 ft/s.[6] The 4-foot square scanned area consists of 12 parallel linear scans separated by 4-inch increments. During the scan period 2048 data points are taken at 100 samples/s.

During the scanning period, the oscilloscope is monitored to see that signal amplitudes are within the receiver operating range. If not, the RF input attenuation is adjusted and the scan is repeated. The remaining locations within the building and immediately outside are similarly scanned and recorded. After the measurements are made, the transmitter is reinstalled on the reference mast, erected and plumbed, and the dc level and RF signal reference level are again recorded. The closure error between beginning and ending reference-level recordings is usually <0.5 dB. If the closure is >1 dB, the measurements are repeated. Such high closure error occasionally occurs when the transmitter battery has discharged below the regulation limit of the voltage regulator.

2.3 Received signal characteristics and attenuation definitions

Motion of the signal source through the 4-foot-square areas inside and outside of buildings results in small-scale signal variations. The variations are caused by multipath propagation.[2,6] Inside houses and in areas shadowed from the van, where propagation is dominated by reflection and scattering, the variations in the received signal envelope are approximately Rayleigh distributed (see Ref. 6 and Section IV of this paper). Received signal minima are separated on the order of one half wavelength.[6] The medians (or means) of these small-scale variations are approximately stationary over the small areas but the medians for areas in different rooms in a building can be significantly different. Thus, the signal statistics can be modeled as a combination of a small-scale, quasi-stationary process (multipath) superimposed on a large-scale process (shadowing). This model is like the models used for mobile radio propagation.[2,12]

Only a single parameter is needed to describe the small-scale Rayleigh-distributed signal variation. The median can be determined if somewhat over half of the samples are above the measurement threshold. The mean, however, will be biased by the receiver noise level unless significantly more than half the samples are above the threshold. Therefore, medians of received signal variations for the 4-ft-square areas are used to characterize the small-scale signal variations at different measurement locations. The lowest median signal measured was 44 dB above the receiver noise level for the data presented in Sections III and IV. The lowest median cross-polarized signal measured was 38 dB above the noise level.

Building attenuation for a location inside a building can be defined as the difference between the signal level at the location and a "representative" signal level outside the building. Both levels are taken to be small-scale medians, measured in decibels, relative to a common reference level.

There are at least two possibilities that could be used for the representative level outside. The first choice is to use the average level of the signal in which the building is immersed. This level can be approximated by taking the average of several median levels, in decibels, measured at different locations surrounding the building. This decibel averaging is appropriate for large-scale variations of the median that are log-normally distributed. Large-scale variations are shown in Section IV to be approximately log-normally distributed. This definition of representative level seems appropriate for system considerations because service would have to be provided all around the exterior of a building.

The second possible choice for a representative level is to use the level of the signal incident on the building from the PORT (or received at the PORT from the incident region, since reciprocity holds). The incident level can be determined readily when the PORT is in line-of-sight of the building and there is little multipath from surrounding buildings. This was the case for the measurements in Ref. 6. However, when there are many intervening buildings between the PORT and the subject building and there is considerable multipath from surrounding buildings, the signal level incident on the building is not easily defined. This second choice of representative level appears least appropriate at the longest distances where the shadowing and multipath are most significant. These are also the distances of most concern in system considerations.

In Section III, comparisons are made of attenuations determined using both choices for representative levels. However, the more appropriate average outside level is used for the statistics in Section IV.

III. ATTENUATIONS AND CROSS-POLARIZATION COUPLINGS FOR THREE BUILDINGS

3.1 Building descriptions

The first building, shown in Fig. 2, is 20 feet by 38 feet, of corrugated steel construction and mounted on a concrete platform. There are metal screened doors inside the main metal doors, and all of the nine windows are metal screened, with the exception of three in the rear. The building inside is a single large area and is about one-third filled with equipment.

The second building is shown in Fig. 3. It is 21.5 feet by 28 feet, is covered with aluminum siding and has one solid door and nine win-

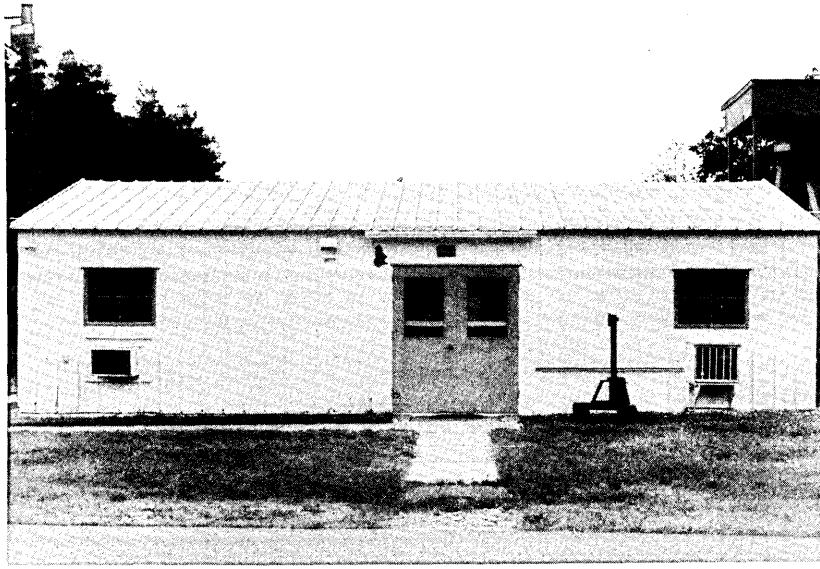


Fig. 2—Corrugated steel building on Crawford Hill as seen from the instrumentation van position.

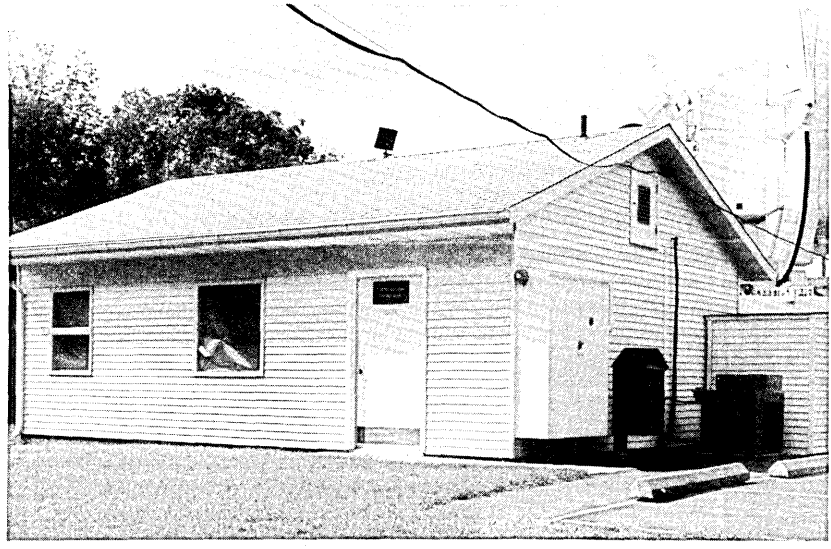


Fig. 3—Antenna Control building on Crawford Hill. The photograph was taken looking toward the building along the path from the van.

dows. All except the three largest windows are metal screened. The building interior is divided roughly in half on the long dimension and is about half filled with equipment.

The third building is of wooden construction, is 16 feet by 54 feet

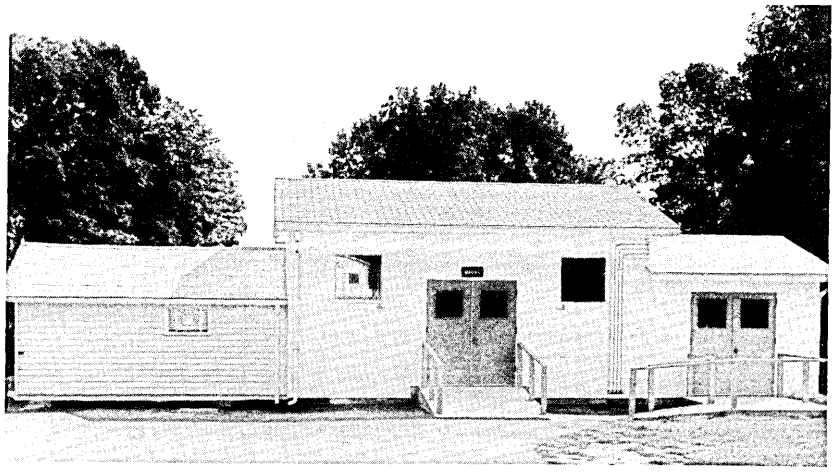


Fig. 4—Wooden building on Crawford Hill. The propagation path from the van was into the corner to the left in the photograph.

and contains vapor barrier foil insulation in the walls. This building, shown in Fig. 4, was empty. In front it has two doors with unscreened windows and three other unscreened windows. The room with double doors to the right in Fig. 4 is attached to an exterior wall. Measurements were not made in the attached room.

3.2 Attenuations for the three buildings

Signal levels were measured at four locations outside and two or three locations inside of the three buildings from a single van (PORT) location. These levels are summarized in Table I. The distances indicated in the table (i.e., 225 feet, 330 feet, and 845 feet) are the path length from the van antenna to the fronts of the buildings. The centers of the four outside locations for each building were about 5 to 10 feet away from the midpoints of the four sides of the building. The outside locations are numbered clockwise looking down on the building from above, starting with the side closest to the van. The small-scale median signal levels in decibels and the decibel average for the four outside locations are tabulated in the column labeled "Level/dB". These levels are relative to 0 dB at the signal level reference 14.2 feet from the van receiving antenna. The column labeled $1/r^2$/dB contains the signal levels that would occur in free space ($1/r^2$) at the distance of the measurement location. These levels are also relative to 0 dB at the 14.2 foot reference location. The $1/r^2$ level in the row labeled Avg. is the free space value at the building midpoint.

For the wooden building positioned at 330 feet from the van and for the building with metal siding positioned at 845 feet, the signal medians measured in front of the buildings are 1 to 2 dB greater than

Table I—Building attenuations for three buildings

Building Material	Location Outside	Level (dB)	$1/r^2$ (dB)	Location Inside	Relative to Average Attenuation (dB)	Relative to Incident Attenuation (dB)
All metal (225 ft)	Avg.	−29.0	−24.3			
	1	−24.9	−23.7	1	22.3	26.0
	2	−27.8	−24.1	2	18.7	22.3
	3	−36.2	−25.0	3	24.0	27.3
	4	−27.1	−24.1			
Metal siding (845 ft)	Avg.	−45.8	−35.7			
	1	−34.3	−35.6			
	2	−43.2	−35.5	1	9.0	20.4
	3	−49.5	−35.8	2	0.3	11.8
	4	−56.2	−35.8			
Wood (330 ft)	Avg.	−31.3	−27.5			
	1	−24.8	−26.8			
	2	−40.6	−27.7	1	−1.8	3.9
	3	−34.5	−27.8	2	2.6	9.0
	4	−25.4	−27.2			

the free-space values. The signal median in front of the all-metal building positioned at 225 feet is about 1 dB below the free-space value. These values are all consistent with the effects of a single reflection from the relatively flat dry ground between the van and the buildings.[13] The 27-foot van antenna height and a reflection coefficient phase angle of nearly 180 degrees, appropriate for small angles between the incident wave and smooth dry ground, yield signal maxima at distances above the ground of about 3 feet for 330 feet from the van and of about 8 feet for 845 feet from the van. For the same conditions, a minimum occurs about 4 feet above the ground at 225 feet from the van. Recall, the signal source is scanned about 4.5 feet above the ground. Therefore, at 330 feet, the scan is near a broad maximum that could be as much as 6 dB above free space if the ground were perfectly reflecting. The measured level at 330 feet is 2 dB above free space. At 845 feet, the scan should be down on the side of an 8-foot-high maximum. The measured level at 845 feet is 1.3 dB above free space. At 225 feet, the scan is near the first minimum above the ground. The measured level at 225 feet is 1.2 dB below free space.

The interior locations of the buildings are arbitrarily numbered in the column labeled "Location Inside". The column labeled "Relative to Average Attenuation in decibels" contains the differences between the median signal levels measured at the locations and the average of the four outside median levels for the building. Positive attenuation indicates the signal level inside is smaller than the average level outside.

Within a period of a few minutes, three separate measurements

were made at Location 1 inside the metal-sided building. The spread of the resulting three medians was only ±0.5 dB around the average value listed in the table. Two separate measurements were made at Location 2 outside the wooden building. The resulting two medians were within ±0.3 dB of the average value listed. Thus, the measurement repeatability is good and is consistent with the ±0.5 dB standard deviation of the medians expected for the statistical fluctuation resulting from the limited number (≈150) of independent samples in a 4-foot-square area. Measurements were made at the same locations in and around the metal building on three different days that had different ground moisture conditions, etc. The three medians for each location were averaged and the difference was taken between the location average and the individual medians. The standard deviation of the differences was 1 dB.

The column labeled "Relative to Incident Attenuation in decibels" contains the differences between the median levels at the inside locations and the median level for the outside location that is closest to the van. These outside closest locations have the largest signal levels measured for their corresponding buildings. The outside level is corrected for free space ($1/r^2$) for the distance between the outside location and the inside location being considered. This is the second definition of building attenuation described in Section 2.3. This second definition of attenuation is essentially the same definition of building attenuation that was used in Ref. 6 for the same all-metal building listed first in Table I (the values in Ref. 6 were not corrected for $1/r^2$). The second definition attenuation values are within 1 or 2 dB of the values in Ref. 6 for Locations 1 and 3 inside the metal building. At Location 2, however, the attenuation in Table I is 6 dB less than the earlier value. Items inside the building have been rearranged since the earlier measurements, but no reason for such a large change is evident. The second definition attenuation into the all-metal building is greater than attenuation from the front (van side) to the back of that building (10.0 dB front-to-back attenuation including $1/r^2$ correction). The second definition attenuation into the other buildings is less than the front-to-back attenuation (21.7 dB for the metal-sided bulding and 15.0 dB for the wooden one). This suggests that the dominant mechanism for signal propagation behind the metal building is scatter and/or reflection from objects behind and to the side of that building rather than passage through the building itself. The dominant mechanism for propagation behind the other two buildings is not evident from these simple measurements since either passage through multiple walls or scatter and/or reflection is consistent with the result.

For Location 1 in the wooden building, the first definition attenuation is negative. This indicates that the median level inside at that

location is greater than the average of the four outside medians. This is a reasonable situation because the signal levels at one side and at the back of that building are much lower than the signal level inside Location 1. This inside location has unscreened windows between it and the outside in the direction of the van. As mentioned in Section II, this first definition of attenuation seems more appropriate for use in system analysis since a system would have to serve the outside locations at all sides of a building. Of course, the distribution of outside levels relative to the outside average is also needed for a complete assessment of system performance.

The building attenuations by either definition are greatest for the all-metal building with metal screened windows. Since both of the other buildings have metal in their walls, attenuation into them probably depends strongly on coupling through the nonscreened windows.

3.3 Cross-polarization couplings for the three buildings

Cross-polarization coupling in multipath propagation is significant in reducing the effects of the random orientation of portable radiotelephones.[14] Cross-polarization coupling is defined as 20 $\log(E_x/E_t)$, where E_t is the average or median field magnitude of the polarization aligned with the transmitted polarization and E_x is the average (or median) field magnitude of the polarization orthogonal (crossed) to E_t. An indication of the cross-polarization coupling can be obtained by orienting the signal-source dipole antenna horizontally and scanning a measurement location with the dipole pointed towards the van (end-on orientation). The scan can be repeated with the dipole perpendicular to the direction of the van (broadside orientation).

If the multipath propagation were uniformly distributed in azimuth around the measurement location and were confined to a horizontal plane, horizontally polarized multipath would produce a median received signal level in a scanned horizontal dipole that was 3 dB less than the median that would be produced by the same multipath in a scanned loop oriented with its plane horizontal. The 3-dB decrease results from the nonuniform directivity pattern of the dipole in any plane containing the dipole. The median level received by the loop in the horizontally polarized multipath would be the same as the median level received by a scanned vertical dipole in vertically polarized multipath of the same average intensity. The multipath signal variations in all cases would be Rayleigh distributed.

For the measurement situation, equal median signal levels for end-on and broadside horizontal scans would be consistent with multipath having a uniform azimuthal distribution. Then, under the assumption that the propagation directions are confined to a horizontal plane, the

cross-polarization coupling would be 3 dB greater than the signal difference $\Delta = L_v - L_h$, where L_v is the median level of the signal from a scan with the source oriented vertically, and L_h is the median level of the signal from a scan with the source oriented horizontally. The medians have a statistical fluctuation with a standard deviation on the order of ±0.5 dB because of the limited number of independent samples. Therefore, for differences between the medians of the two horizontal scans of one or two decibels, they can be taken as equal and their average value can be used to determine Δ.

Table II summarizes the cross-polarization measurements made in and around the three buildings. The medians of the end-on and broadside scans are tabulated in columns labeled End-on and Broadside. The values are in decibels relative to the median of the signal scan at the same location with the source antenna oriented vertically. That is, columns End-on and Broadside indicate Δ for end-on and broadside scans. The column labeled cross-polarization is 3 dB greater than the average Δ for end-on and broadside scans.

The locations inside the metal building would yield a small positive value for cross-polarization coupling. This probably indicates a breakdown of the assumption that the multipath propagation is confined to a horizontal plane, so the coupling is taken as 0 dB. If the multipath propagation were uniformly distributed in all directions in three dimensions, the orientation of the antenna would be irrelevant. Since the data show only a small bias towards stronger median signal for the vertical antenna in the metal building, an alternative assumption for that building would seem to be uniformly distributed multipath propagation in all directions in three dimensions.

The cross-polarization values in Table II are all greater than −10 dB and most are greater than −6 dB. Another point worth noting is that the locations with the lowest signal levels (greatest attenuation) also have the largest cross-polarization coupling with values greater than −6 dB and usually greater than −3 dB. Since cross-polarization

Table II—Cross-polarization coupling for three buildings

Building Material	Location Outside	Location Inside	Broadside Scan (dB)	End-on Scan (dB)	Cross-polarization (dB)
All metal	3	—	−9.3	−6.8	−5.0
	—	1	−1.0	+0.4	0
	—	2	−4.4	−0.9	0
	—	3	−1.0	—	0
Metal siding	2	—	−13.6	−10.6	−9.1
	4	—	−5.0	−6.2	−2.6
	—	1	−4.0	−4.5	−1.3
Wood	2	—	−5.7	—	−2.7
	—	1	−8.4	−10.2	−6.6

coupling increases the effectiveness of diversity in mitigating the effects of random portable set orientation and multipath propagation,[14] this trend could be significant in determining system performance.

IV. ATTENUATION STATISTICS FOR A HOUSE

4.1 House description

The house is a two-story colonial located on a level, one acre lot. It is in a newly developed area with a density of one house per acre and with relatively few trees. The house has an area of 2400 square feet, consisting of living room, dining room, kitchen, and den on the first floor, and four bedrooms on the second floor. It also has a basement and a two-car attached garage. It is constructed with wood, with aluminum siding on three sides and nonmetallic siding on the front. All exterior walls contain insulation faced with a metal foil vapor barrier.

4.2 Small-scale statistics

The cumulative distributions of the envelope variations of multipath propagation for a scan of a 4-foot-square location are expected to be Rayleigh.[2,6] Figure 5 shows the distributions measured at four locations in and around the house. A Rayleigh distribution is a straight line with the particular slope indicated on the figure. The distributions are good approximations to the Rayleigh distribution for many locations inside the house and behind the house from the van. A few locations on the same side of the house as the van and only 400 feet from it experience essentially line-of-sight propagation. At these few locations the signal variation is small and the envelope distribution significantly departs from Rayleigh, as indicated by the example plotted as squares on the figure. The medians of the four distributions on the figure have been normalized to 0 dB. The two distributions that show the greatest departure from Rayleigh were selected as the extremes that have the greatest and least spread in attenuation of any of the distributions in the data set.

4.3. Large-scale statistics of small-scale medians for outside locations

The open data points in Fig. 6 are the medians of the measured signal envelopes for the small-scale locations outside the house. The medians are plotted versus the distance between the van antenna and the location. The median levels are in decibels relative to the signal level at the van reference. The outside locations are in front of the midpoints of the four outside walls of the house. The decibel averages of the four locations for each van position are indicated by the solid data points.

A strong dependence of signal level on distance is evident in Fig. 6.

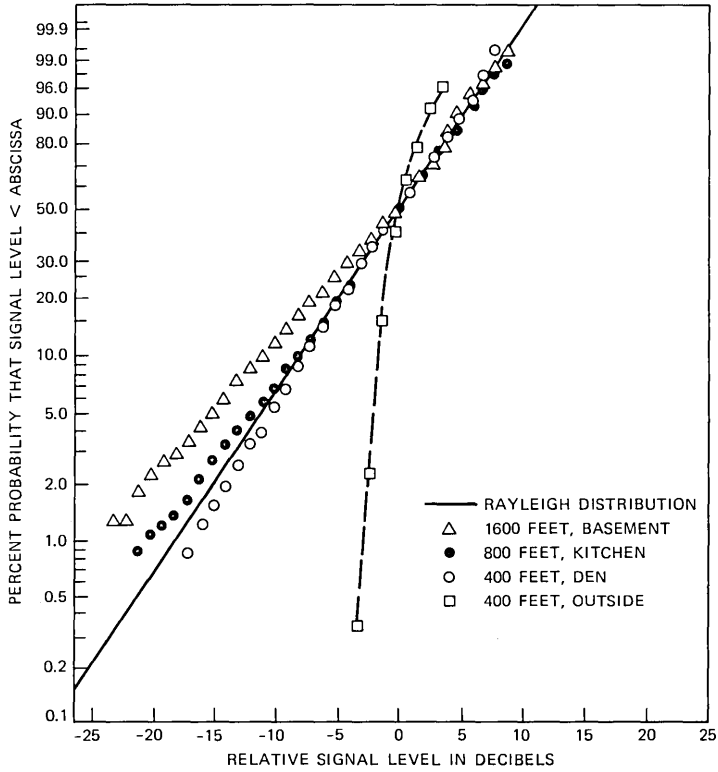


Fig. 5—Measured signal envelope distributions from four small (4-ft square) areas. On these coordinates, Rayleigh distributions are straight lines with the slope indicated. Distances between the van and the measurement location are indicated in the key.

The solid line is the linear-least-squares regression fit to the 36 medians from the four locations for each of the nine van positions. The least-squares regression fit to the averages of the four locations for each van position yields the same solid line. Signal level varies with distance as $d^{-4.5}$ for the solid line. The dotted line represents free-space propagation (d^{-2}) relative to 0 dB at the van reference. At 1000 feet from the van, the signal level represented by the solid line is 12.5 dB lower than the level would be in free space, i.e., the average excess attenuation over free space is 12.5 dB at 1000 feet. The distance dependence exponent is probably reliable only to several tenths because of the large spread in the data and the relatively small number of data points. The signal level outside was correlated to distance with a coefficient of 0.8.

The lower dashed line is the linear regression line for the data near 400 feet and 800 feet only. The upper dashed line is for the data near 800 feet and 1600 feet. The distance dependence for the 400-foot to

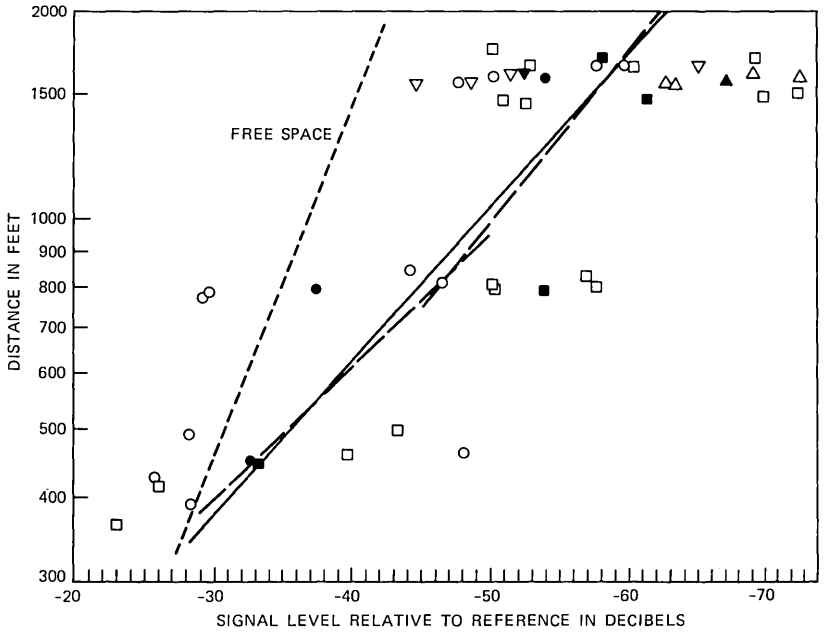


Fig. 6—Medians of the small-scale envelope variations for different outside measurement locations plotted versus distances between the locations and the van antenna. Open data points are medians for individual outside locations. Solid points are averages of the medians for the four measurement locations for each of the van locations. Points represented by the same symbols at nearly the same distance designate the four measurements associated with a particular van position. Signal levels are with respect to 0 dB at the signal reference located 14.2 feet from the van receiving antenna. The dotted line represents free-space propagation ($1/r^2$) relative to 0 dB at the signal reference. The solid line and the dashed lines are linear-least-squares regression lines, as discussed in the text.

800-foot distances is $d^{-5.1}$; the dependence for the 800-foot to 1600-foot distances is $d^{-4.1}$. Thus, the signal decreases faster with distance in the first several hundred feet than it does farther from the house. This is reasonable because, for the first few hundred feet, there are few obstructions along the path and propagation approaches line-of-sight. After several hundred feet, the number of intervening houses increases. Reflection from the ground also causes the signal to decrease more rapidly with distance than it does in free space.

The data points at 400 feet that are greater than free space values deserve comment. These points are from locations that are not shadowed from the van, i.e., they are within line-of-sight. A single ground reflection from a 27-foot-high antenna 400 feet away would produce a signal maximum about 4 feet above the ground. The signal maximum would be greater than the single-path free-space value. (It would be 6 dB greater for a reflection coefficient of unity.) The signal levels a few dB above free space are not unexpected 4.5 feet above the ground in

locations that have large reflecting surfaces (walls of houses) nearby in addition to the ground.

Figure 7 is the cumulative distribution of the data from Fig. 6 after the distance dependence, $d^{-4.5}$, is removed. The calculated standard deviation for the data points is 9.0 dB. A straight line on the coordinates in Fig. 7 represents a log-normal distribution of signal level. The measured distribution is a reasonable approximation to a log-normal distribution.

4.4 Large-scale statistics of small-scale medians for inside locations

The distance dependences for the linear-least-square fits to the medians for measurement locations inside the house were somewhat different for the different floors of the house. The variation with distance was $d^{-3.9}$ for the first floor, $d^{-3.0}$ for the second floor, and $d^{-3.2}$ for the basement. Since only one location was measured in the basement, this value for the basement may contain considerable statistical

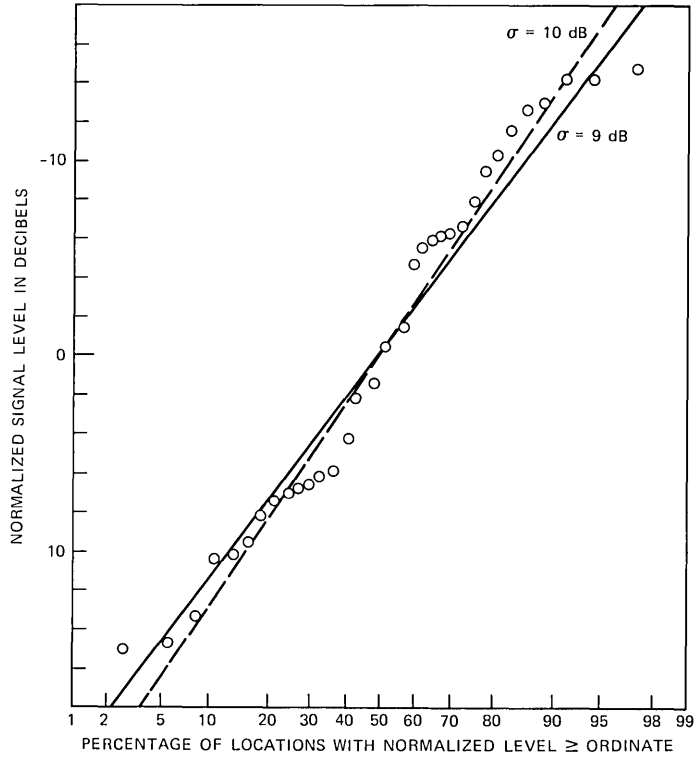


Fig. 7—Cumulative distribution of the medians of the small-scale envelope variations for outside locations after the values for the solid regression line in Fig. 6 are subtracted out. The solid line and the dashed line in this figure represent log-normal distributions with standard deviations of 9 dB and 10 dB, respectively.

error. The variation with distance for the composite set of data from the entire house (eight measurement locations inside the house and nine van positions) was $d^{-3.5}$. The less rapid decrease in signal level with distance for the second floor compared with the first floor is reasonable because the height of the second floor results in less attenuation from intervening houses. Since the decrease in signal level with distance is less rapid for the locations inside the house than outside, the apparent building attenuation will decrease somewhat with distance.

At 1000 feet, the excess attenuation over the free space values are 18.5 dB for the first floor, 16.5 dB for the second floor, 29.9 dB for the basement, and 19 dB for the composite data set for the entire house. The average building attenuation at 1000 feet, which is the difference between the linear regression values outside and inside, is then 6.0 dB for the first floor, 4.0 dB for the second floor, 16.4 dB for the basement, and 6.5 dB for the entire house.

4.5 Large-scale statistics of the building attenuation

The building attenuation considered in this section uses the first definition in Section 2.3, i.e., attenuation is with respect to the average of the four outside median levels. Building attenuations for the house are plotted versus distance in Fig. 8. The open data points represent data from the first floor; the solid data points and the points marked B represent data from the second floor and the basement, respectively. As noted in the previous section, the attenuation is weakly dependent on distance. Also, the attenuation into the basement is significantly greater than the attenuation into the other two floors. The linear regression lines are labeled on the figure for several combinations of the data. The variation of attenuation with distance is $d^{-0.6}$ for the first floor, $d^{-1.5}$ for the second floor, $d^{-1.4}$ for the basement, and $d^{-1.0}$ for the first and second floor together and for the entire house. The attenuation at 1000 feet taken from the regression lines is 6.0 dB for the first floor, 4.0 dB for the second floor, and 16.4 dB for the basement. These 1000-ft attenuation values are the same as those obtained in the previous section from the separate linear regression lines to the signal levels inside and outside.

Means (m) and standard deviations (σ) for several different groupings of the attenuation data are tabulated in Table III. It appears that most of the difference in the distance dependence for the first and second floors results from the different attenuation averages at 1600 feet.

The different distance dependences and different average attenuation for the different floors present a dilemma if they are real effects not confined to this data set. For radio system analysis, a single

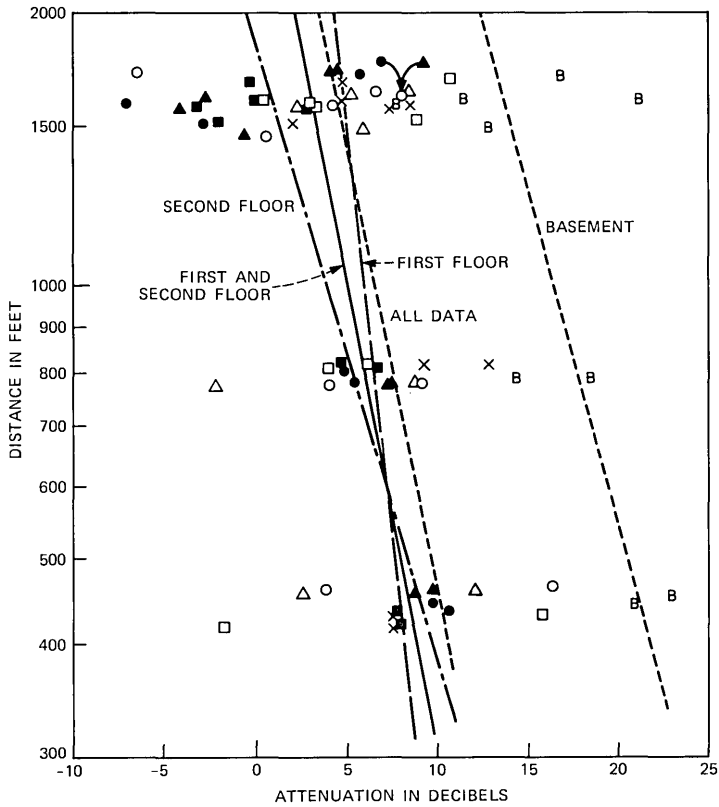


Fig. 8—Building attenuation by the first definition, i.e., relative to the average of the median levels for the four outside locations, plotted versus distance. Open data points and X points are from the first floor, solid data points are from the second floor, and points marked B are from the basement. All points marked with a particular symbol are from the same measurement location in the house but are for different van positions. The lines are linear-least-squares regression lines through different groupings of points.

Table III—Attenuation and standard deviations for the house (table values are in dB)

Floor	m σ	400 ft	800 ft	1600 ft
First	m	8.1	6.5	4.8
	σ	6.4	4.5	3.9
Second	m	9.1	6.1	0.7
	σ	1.1	1.3	4.6
Basement	m	22.1	16.4	13.8
	σ	1.7	2.8	5.1
First and second	m	8.5	6.3	3.1
	σ	4.8	3.4	4.7
First, second and basement	m	10.2	7.6	4.4
	σ	6.4	4.7	5.9

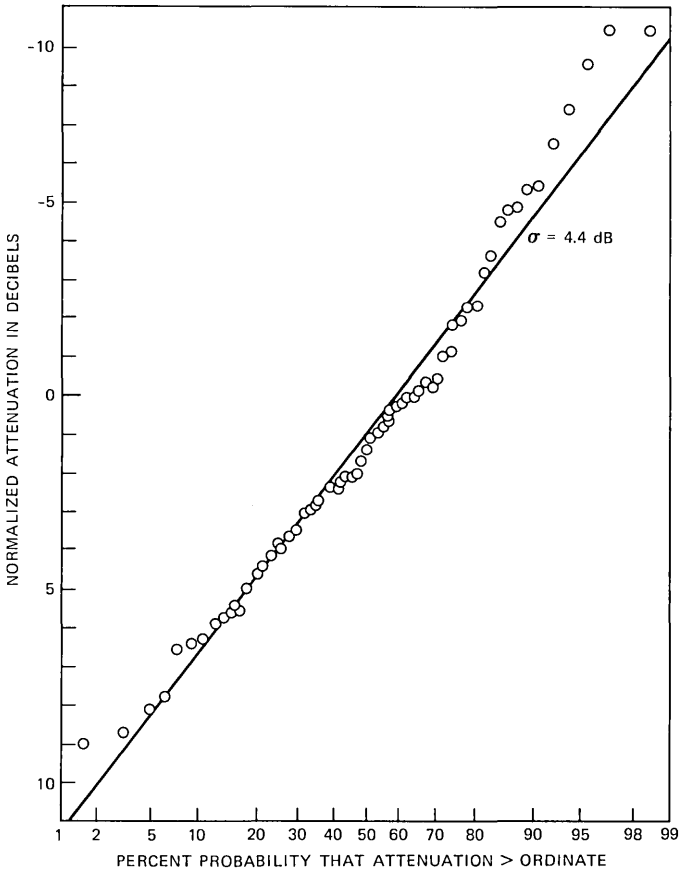


Fig. 9—Cumulative distribution of the building attenuations from Fig. 8 for the first and second floors after the values for the first- and second-floor regression line (solid line in Fig. 8) are subtracted out. The straight line represents a log-normal distribution with a standard deviation of 4.4 dB.

distance dependence and average attenuation are desirable; however, if only the composite regression line for the whole house were removed from the data, the variation remaining in the residual attenuation data would have a significantly larger standard deviation ($\sigma = 6.2$ dB) than if the individual regression lines for each floor were removed separately ($\sigma = 4.1$ dB). Also, the composite line for the entire house depends on the number of measurements made on each floor. For systems analysis, these numbers of measurements should be related to the probability of calls being made on those floors (or at those locations); but these probabilities are unknown.

The separation in decibels between the regression lines on Fig. 8 for the first and second floors is not extreme over the 4 to 1 distance

covered. Also, the attenuation averages for these floors for the three distances are not grossly different (all within ±2 dB). Therefore, it seems reasonable to simplify the description of the attenuation by combining the first and second floor data. The composite regression line for this combination is also shown on the figure. When this composite regression line is removed from the first and second floor data, the resulting cumulative distribution of attenuation is as shown in Fig. 9. The standard deviation of the attenuation data plotted in Fig. 9 is 4.4 dB, not significantly different from the 4.1-dB standard deviation for all the data with the regression line for each floor removed separately. The distribution in Fig. 9 is a reasonable fit to a log-normal distribution, especially considering the limited size of the data set. Combining the basement data with the data from the other floors is not reasonable because of the large difference in average attenuations. Since this difference can be expected to exist for all houses, the attenuation into basements will probably have to be described separately from the attenuation for the rooms on other floors.

V. CONCLUSIONS

Signal levels were measured in and around three small buildings from a single position of an instrumentation van. Measurements were also made in and around a house from nine different van positions ranging from 400 feet to 1600 feet away. The received signal envelope is approximately Rayleigh distributed over most small-scale areas about 4 feet square. For the three buildings, building attenuation of the small-scale signal medians ranged from −2 to 24 dB relative to the average signal level outside each building.

For the house, some of the parameters of the medians of the small-scale envelope variations and of the building attenuation defined in this paper are summarized in Table IV. After the distance dependence is removed from the median signal levels outside of the house, the distribution of the residual large-scale signal-level variations is ap-

Table IV—Summary of signal-level and attenuation parameters for the house

Parameter	Out-side	First Floor	Second Floor	Base-ment	First & Second Floors	Units
Signal-level distance dependence exponent	−4.5	−3.9	−3.0	−3.2	—	—
Signal-level relative to free space at 1000 ft.	−12.5	−18.5	−16.5	−28.9	—	dB
Building attenuation exponent	—	−0.6	−1.5	−1.4	−1.0	—
Average building attenuation at 1000 ft.	—	+6.0	+4.0	+16.4	+5.1	dB

proximately log-normal with a standard deviation of 9 dB. The standard deviation of the building attenuation for the first and second floors is 4.4 dB after distance dependence is removed. Because of the large average attenuation into the basement of the house, it does not appear reasonable to combine the basement attenuation statistics with the statistics for the rooms above ground level.

Cross-polarization coupling is strong inside and outside of the three buildings. For all locations measured, the coupling is greater than −10 dB and for most locations it is −6 dB or greater.

VI. ACKNOWLEDGMENTS

Initial design of the van masts and mounts was done by W. I. Tohlman and H. H. Hoffman. We wish to thank H. W. Arnold for permitting the measurements reported in this paper to be made in his house. The continued support of L. J. Greenstein and D. O. Reudink is greatly appreciated.

REFERENCES

1. D. C. Cox, "Co-Channel Interference Considerations in Frequency-Reuse Small Coverage-Area Radio Systems," IEEE Trans. Commun., *COM-30* (January 1982), pp. 135–42.
2. W. C. Jakes, *Microwave Mobile Communications*, New York: John Wiley and Sons, 1974.
3. J. M. Durante, "Building Penetration Loss at 900 MHz," Conference Proceedings, IEEE VTG Conference, 1973, p. 1–7.
4. P. I. Wells and P. V. Tryon, "The Attenuation of UHF Radio Signals by Houses," U. S. Dept. of Commerce Report, OT Report 76-98, August 1976; and IEEE Trans. Veh. Technol., *VT-26* (November 1977), pp. 358–62.
5. J. Shefer, "Propagation Statistics of 900 MHz and 450 MHz Signals Inside Buildings," Microwave Mobile Radio Symposium, March 7–9, 1973, Boulder, Colorado.
6. H. H. Hoffman and D. C. Cox, "Attenuation of 900 MHz Radio Waves Propagating Into a Metal Building," IEEE Trans. Ant. Propag., *AP-30* (July 1982), pp. 808–11.
7. L. P. Rice, "Radio Transmission into Buildings on 35 and 150 MHz," B.S.T.J., *38*, No. 1 (January 1959), pp. 197–210.
8. D. Mitchell and K. G. Van Wynen, "A 150 MC Personal Radio Signaling System," B.S.T.J., *40*, No. 5 (September 1961), pp. 1239–57.
9. G. V. Waldo, "Report on the Analysis of Measurements and Observations of New York City UHF-TV Project," IEEE Trans. Broadc., *9* (1963), pp. 7–36.
10. K. Tsujimura and M. Kuwabara, "Cordless Telephone System and its Propagation Characteristics," IEEE Trans. Veh. Technol., *VT-26* (November 1977), pp. 367–71.
11. M. Komura, T. Hogihira, and M. Ogasawara, "New Radio Paging System and Its Propagation Characteristics," IEEE Trans. Veh. Technol., *VT-26* (November 1977), pp. 362–6.
12. D. C. Cox, "Multipath Delay Spread and Path Loss Correlation for 910 MHz Urban Mobile Radio Propagation," IEEE Trans. Veh. Technol., *VT-26* (November 1977), pp. 340–4.
13. P. A. Matthews, *Radio Wave Propagation, VHF and Above*, London: Chapman and Hall, Ltd., 1965, Chapter 2.
14. D. C. Cox, Antenna Diversity Performance in Mitigating the Effects of Portable Radiotelephone Orientation and Multipath Propagation, IEEE Trans. Commun., *COM-31* (May 1983), pp. 620–8.

AUTHORS

Donald C. Cox, B.S., M.S., University of Nebraska, Lincoln, 1959 and 1960, respectively; Ph.D., Stanford University, 1968, all in Electrical Engineering;

Honorary Dr. of Science, University of Nebraska, Lincoln, 1983; Stanford University, 1963–1968; Bell Laboratories, 1968—. From 1960 to 1963, Mr. Cox worked on microwave communications system design at Wright-Patterson AFB, Ohio. From 1963 to 1968 he was at Stanford University doing tunnel diode amplifier design and research on microwave propagation in the troposphere. From 1968 to 1973 he was a Member of Technical Staff at Bell Laboratories, Holmdel, New Jersey, doing research in mobile radio propagation and on high-capacity mobile radio systems. He is now Supervisor of a group doing propagation and systems research for portable-radio telephony and for millimeter-wave satellite communications. Fellow, IEEE; member, Commissions B, C, and F of USNC/URSI, Sigma Xi, Sigma Tau, Eta Kappa Nu and Pi Mu Epsilon; Registered Professional Engineer, Ohio and Nebraska.

Roy R. Murray, B.S. (Electronic Engineering), 1975, Monmouth College; Bell Laboratories, 1965—. Mr. Murray has worked on high-speed multi-level digital modulators and, more recently, on UHF radio propagation into buildings. Currently, he is a member of the Telecommunication Systems Research Department. Member, Eta Kappa Nu.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 9, November 1983
Printed in U.S.A.

Penetration of Radio Signals Into Buildings in the Cellular Radio Environment

By E. H. WALKER*

(Manuscript received April 4, 1983)

Penetration of radio signals from Advanced Mobile Phone Service cell site transmitters has been measured in fourteen office and industrial buildings in the Chicago area. Signal levels on the first floor of buildings averaged 14 dB less than reference levels in the adjacent streets. This penetration loss was found to decrease with increasing height. Standard deviations of penetration loss ranging from 5 to 11 dB attest to the diversity of architecture and floor arrangements. Other relationships useful in planning for portable phone terminal applications are derived from the data and are presented in this paper. The data include over 4000 measurements, each being the average of 1024 samples of the local field taken over an eight-second period as the instrumentation traveled about 20 feet over the measurement path. The measurement path in each building was traversed for each of the several cell sites that transmitted signals of adequate strength into the buildings.

I. INTRODUCTION

The use of portable and hand-held terminals in 850-MHz cellular radio systems involves a radio field environment inside buildings that differs from the more familiar highway and street propagation environment of the mobile terminal. The laws that define propagation over open and urban terrain are complicated by a penetration loss in the transmission of the signal into the building's interior. The characterization of this loss, as encountered in large office and commercial

* Bell Laboratories.

buildings, was the purpose of an extensive series of measurements made in fourteen buildings served by the Advanced Mobile Phone Service (AMPS) Developmental Cellular System in Chicago. These measurements were then compared with the penetration loss in a single-family residence.

The measurements were made in January 1980 to serve as a basis for evolving system requirements related to the use of portable phone terminals in AMPS systems.

Indoor portable service in cellular systems offers a possible improvement over conventional systems. The multiplicity of serving cell sites should in many cases illuminate all sides of those buildings that enjoy a relatively exposed or unshadowed environment. It was expected that this effect would be somewhat less pronounced in buildings that are more sheltered in the urban core; the data support this conjecture. The data also support expectations that areas with windows would have lower penetration losses than areas without windows; the difference is about 6 dB. First-floor penetration losses averaged 14.2 dB. The loss decreased with height at 1.9 dB per floor, very close to the 2.0- and 2.5-dB rates reported in Refs. 1 and 2. The penetration loss measured in the aluminum-sided ranch house, as a matter of interest, is 7.3 dB, a value close to the findings of Cox et al. in Ref. 3.

Most of the measurements made in the Cellular Test Bed (Ref. 4) and reported in Ref. 5 were of the signal as received by a mobile transceiver. The measurements reported here include the entire distribution of setup channel signal levels throughout the measurement areas for each cell site transmitter that qualified as a server. As such, the data are perfectly applicable to conventional mobile portable and paging applications. The effect is that of taking measurements with the base station in several different locations.

II. ENVIRONMENT AND TEST METHOD

The Chicago AMPS Developmental Cellular System, as described in Ref. 4, uses ten separate setup channels for paging and mobile call origination in the Chicago service area. The channel signals are radiated continuously from omnidirectional antennas. The instrumentation used for the measurements reported in this paper can be tuned manually to any of these frequencies to measure one channel at a time. Figure 1 shows nominal coverage contours for the setup channels of the seven cell sites that served the selected buildings. The related voice-channel coverage is shown in Fig. 2 of the paper by D. L. Huff.[4] The center of each contour represents a cell site location. In general, the cell sites are spaced about 15 miles apart.

A group of buildings in the Chicago area was selected to represent a range of physical characteristics including location, architecture,

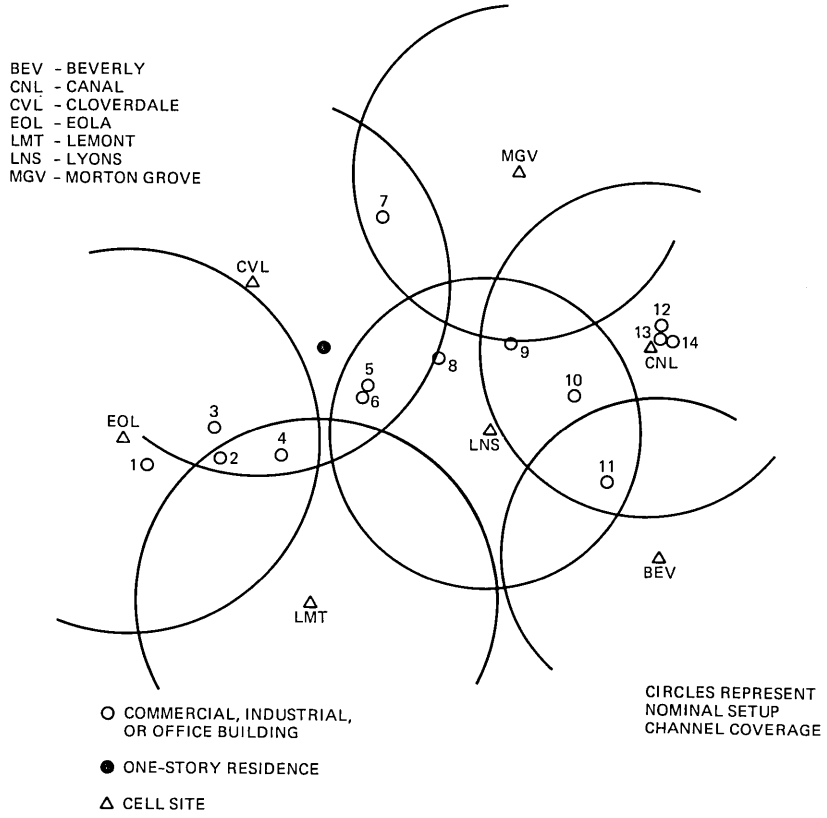


Fig. 1—AMPS Developmental Cellular System locations of 15 buildings and seven serving cell sites.

and local environment. Signal strength measurements were made along planned routes on selected floors of these buildings. A Propagation Measuring Set (PMS) was used to measure the strength of signals transmitted from cell sites.

Building penetration loss for a given floor area is defined to be the difference between the average of these measurements and an average of measurements made on the outside at street level.

Outside signal strength was measured at street level around the perimeter of the building, along the closest available path to the building's outside walls. These paths included driveways, streets, or parking lots, as required to achieve proximity to the building under test.

The PMS data output was transcribed for subsequent analysis by a computer program. The analysis was performed in the AMPS cellular test bed data processing facility described by DiPiazza, Plitkins and Zysman in Ref. 4.

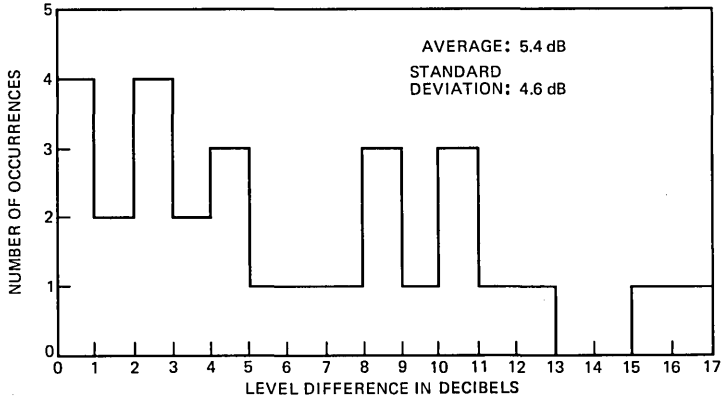


Fig. 2—Differences in level between the strongest and weakest serving channels on the same floor.

The purpose of the measurement program, as stated previously, was to determine the extent to which signals available from AMPS cell sites penetrate buildings. A first-order determination was desired; effects of 5 dB or more on the loss measurements were considered to be significant. A detailed investigation of the structure of interior fields in small buildings is reported by Cox in Ref. 3.

The measurements were subject to the effects of the following different conditions:

1. Different types of outside wall construction, such as steel-framed glass, brick, block masonry, etc.
2. Urban vs. suburban areas to identify difference between buildings in exposed locations and buildings sheltered in the dense urban core.
3. Changes in floor elevation.
4. Different building orientations with respect to serving cell sites.
5. Different percentages of window areas in the outside walls.
6. Different types of window treatment currently used to reflect sunlight and heat.

Time was an important factor in selecting the coverage objective for each building. Test runs were chosen to require about three hours of operations in any given building. The total time of recorded data is a small percentage of that.

Six significant categories of interior areas were identified. These were defined as open, enclosed, and hallways—each with and without windows. Typical of the areas that were considered to be "open" were cafeterias, lobbies, and large office spaces or conference rooms. Typical of "enclosed" areas were small, walled areas with space for only a few occupants, such as small office areas or workrooms.

Measurements in the first floors and second floors, together, were

thought to represent a worst case. Higher floors, thought to represent better penetration, were included with less regularity. Test routes were designed to distribute measurements over the floor areas in a uniform manner.

III. BUILDING SELECTION

The buildings chosen in the Chicago AMPS service area were those where multiple cell sites could provide outside building signal levels of at least −90 dBm.

In the downtown section of Chicago, tests were made in three buildings. These represent the urban group. The downtown buildings in which tests were made can be characterized as "single tenant buildings".

In the suburban area around Chicago, tests were made in eleven commercial buildings and one residential dwelling. Only the data from commercial buildings were included in subsequent analyses.

From the standpoint of architectural variety the program provided measurements in a warehouse, a manufacturing plant, open-space offices, cafeterias, lobbies, small offices, and hallways. Outside wall construction varied from unwindowed concrete slab to steel-framed windows occupying about 90 percent of the outside surface.

The locations of the buildings where tests were conducted are shown on the map of Fig. 1. The circles show the idealized boundaries of regions of cell site control channel coverage. Table I shows the distribution of floor levels in buildings where tests were conducted and the cell sites whose transmissions served the various floor levels. A more detailed description of each of the buildings is given in the appendix.

Several factors were considered when selecting floors and floor areas for measurement. From a test sampling standpoint the first and second floors were chosen when available since building penetration was expected to be poorest at the lower levels. Higher floor levels were included in the tests to provide data on the sensitivity of building penetration to height. On any particular floor, areas were chosen that represent the different types of interiors; test paths were chosen to be uniformly distributed over the floor. In two instances measurements were attempted in basement areas but signal strength levels were below the threshold of the PMS.

IV. MEASUREMENT TECHNIQUE

A portable Propagation Measurement Set (PMS) designed for field strength measurements was adapted for this test program by mounting it in a small "tea" cart equipped with a battery, battery charger, and a mobile antenna mounted on a 2-foot by 2-foot ground plane. The

Table I—Buildings, floors, and serving cell sites contributing to measurements

	Buildings	Floors	Cell Sites*						
			CVL	EOL	LMT	LNS	CNL	MGV	BEV
1	Nabisco	1		X					
	Naperville	2		X	X				
		4	X	X	X				
2	Metropolitan	1	X	X					
	Naperville	2	X	X	X				
		3	X	X	X				
3	Bell Labs	1	X	X					
	Naperville	3	X	X					
		5	X	X	X				
4	BSCTE	1	X	X	X				
	Lisle	3	X	X	X				
		5	X	X	X	X			
		9	X	X	X	X			
5	Illinois Bell	1	X	X	X	X			
	Oakbrook	2	X	X	X	X			
6	McDonalds	1	X		X	X			
	Oakbrook	2				X			
		8	X	X	X	X			
7	Chrysler	1	X					X	
8	Allied Van Lines	1				X			
		2	X		X	X			
	Hillside	4	X		X	X	X		
9	Illinois Bell	3				X	X	X	
	Oak Park	4				X	X	X	
10	Western Electric	1				X			
		2				X			
	Cicero	3				X	X	X	
11	Illinois Bell	1				X	X		X
	Kedzie	4				X	X		X
12	American Medical Assn	1				X	X	X	
		2				X	X	X	
13	Illinois Bell	1					X		
	Headquarters	3					X		
	Randolph	5					X	X	X
14	Illinois Bell	1					X		
	Adams	2					X		
		12					X		
		15					X		
One-story residence (data not included in analyses)		1	X	X	X	X			

* CVL–Cloverdale; EOL–Eola; LMT–Lemont; LNS–Lyons; CNL–Canal; MGV–Morton Grove; BEV–Beverly.

ground plane was positioned 4-1/2 feet above floor level, at approximately shoulder height.

The design of the PMS is based upon the instrumentation receiver of the mobile communications laboratory described by DiPiazza, Plitkins, and Zysman in Ref. 4.

The PMS has an internal calibration system that calibrates the receiver over an 80-dB range in 1-dB increments. In the measurement mode, the basic unit of measurement is a one-second power average derived from a log-amplifier output sampled at a 128-Hz rate. The data output of the set can be adjusted to provide one average value, also derived as a power average, for a group of 1 to 128 of these one-second averages, selectable in binary increments. The measurement range of the set is −40 dBm to approximately −122 dBm.

The PMS uses a paper tape printer for hard copy output. The output data include frequency, a sequential "major marker," the power average, and the one-second sample variance for the group of one-second samples contributing to each longer-term mean. In this application, a mean value was generated each eight seconds. The "major marker" is used to key measurements to the time of day and to logged operator observations.

Prior to a series of measurements, the PMS is calibrated and the calibration results are printed on paper tape. A received signal strength meter in the PMS provides a visual indication of the receiver output without operating the paper tape recorder. The measurement team uses it to verify test set operation and to choose the group of control channels for measurement. The set also contains start, stop, and pause controls to limit data output to areas of interest.

To increase the confidence level of outside measurements, two methods were used where possible to determine outside signal strength. One method used the Mobile Telephone Laboratory (MTL) vehicle, which has the capability of power averaging instantaneous signal strength samples. This facility is described by Huff in Ref. 4. The MTL receivers were sampled at a rate of 32 samples per second. The short-term (1/2-second) MTL power averages were grouped in segments representing building faces and then further power averaged to provide a single mean value for each face segment. The segment means were then decibel averaged over all faces to develop the outside reference.

The second method used the PMS at the time of the in-building measurements to spot check MTL data as a guard against system variations.

MTL measurements were not available for the buildings in the Chicago downtown area, the AMA building, the IBT Headquarters building at Randolph, and the Bell Training Center at West Adams. The PMS was used to collect both inside and outside measurements.

V. TEST PROCEDURES

The PMS was transported to the building locations in a step van. While the PMS was in the van, the van rooftop antenna was connected to it, and street-level data were recorded along the building perimeter.

The measurements collected along the entire perimeter were decibel averaged to provide the outside reference.

The largest possible variety of areas was included in the routing chosen for the PMS measurements in each particular building. In measuring areas within a building the data were marked so that the different interior types could be distinguished. The PMS has a provision for including a marker number in the data output. As the PMS measurements were made along each route, an accompanying commentary was made on a cassette recorder. This commentary provided information about the type of building interior area being measured and the associated major marker number. The major marker information and the interior type were used subsequently in coding the data for computer program input.

Also, this commentary was useful in flagging data that should be discarded and in coordinating the data with the floor area to which it applied.

As we noted earlier, the PMS has the capability to generate average values that can be selected to represent from 1 to 128 seconds of real time. The 8-second average was chosen for output data in these measurements. That value represents a compromise between the volume of data points and the number of feet of travel per data point. With the operational procedures used in these tests, the 8-second averages result in one data point for about each 20 feet of travel. A total data volume of approximately 4000 data points was accumulated for the entire series of in-building measurements.

During the initial planning phase, the expected setup channel signal strength in the vicinity of each building was predicted from propagation contours of the cell sites in the Chicago AMPS System (see Fig. 1). Before measuring each building floor, the operator of the PMS noted the received signal strength indicator (RSSI) value at several locations on the floor. If it was determined that a high percentage of the area was served at signal strengths above the PMS receiver threshold, the operator proceeded to collect data at each setup channel frequency over the entire prescribed route for the selected building floor.

In the measurements made in the first building in the measurements program (Bell Laboratories, Naperville), a high degree of repeatability was noted when the measurements from PMS output tapes were compared for repeated runs. Based on these visual observations it was decided that a single pass at each frequency of interest would provide sufficient accuracy for measurements in the remaining areas.

A comparison of repeated measurements over a common route is shown in Table II. The results indicate run averages are repeatable within approximately 1 dB.

Table II—Repeatability of measurements*

Floor	Channel	First Run Avg. Level (dBm)	First Run Std. Dev. (dB)	Second Run Avg. Level (dBm)	Second Run Std. Dev. (dB)	Level Diff. (dB)
3	780	−93.7	7.2	−94.6	7.4	0.9
3	798	−109.3	6.5	−108.6	8.0	−0.7
5	780	−77.4	10.4	−78.0	11.2	0.6
5	788	−94.3	9.0	−94.7	9.6	0.4
5	798	−85.5	9.5	−85.0	9.0	−0.5

* Measurements made at Bell Laboratories, Naperville, IL.

VI. DATA PROCESSING

As the first step in data processing, all measurements are coded to associate them with major marker numbers. The coding identifies the:

1. Building
2. Floor number (or outside street-level data)
3. Serving frequency
4. Time of day
5. Number of 1-second samples included in each average
6. Type of interior area (six possible types)
7. Flags indicating:
   (a) repeated measurements
   (b) data to be rejected because of operational trouble
   (c) omission of N data points (beginning or end) for reasons of ambiguity.

The time-of-day information permits verification of the major marker sequence and data group duration.

Data become ambiguous when the transition from one major marker to the next is accomplished without halting the measurements system. The ambiguous data point (8-second sample) contains portions of data from two major markers; it is deleted from the database.

The outside street-level data are used for developing a reference for building penetration loss calculations. After the data are encoded, the processing is performed in four steps (or passes) as described below:

1. Initial processing of the data (Pass I) develops statistics for all data for each separate category (building, floor, area type, and channel frequency). Invalid data points are eliminated. A summary listing is provided that includes the number of entries, the number of entries below PMS threshold, the average power level, and the standard deviation for each group of data from categories of areas. In addition to this summary, histograms for each group are developed.

2. The second pass introduces the outside street-level reference information and combines it with the Pass I data. This produces a display of the data for each frequency with its differential level relative to the outside reference (penetration loss).

3. The third pass combines the penetration loss for all the measurement frequencies on common floors, as generated in Pass II. This allows penetration loss to be used as a base for further combining to classify losses by area type, floor level, etc.

4. The fourth pass combines the penetration loss data from Pass III across different buildings. In the combining process similar data from different buildings are weighted to compensate for differences in the data volume collected in each specific building. The weighted average loss for a class of data from several buildings represents equal contribution from each building. Pass IV generates information about the effects of classes of building locations (suburban vs. urban) and penetration loss by floor and area type, for all buildings.

VII. SIGNIFICANT CONCLUSIONS

The conclusions made as a result of our extensive measurements are as follows:

1. Urban vs. suburban—First-floor penetration loss measurements were processed for three different groupings of buildings. The urban (downtown Chicago) building penetration for three buildings shows an average value of 18.0 dB. The penetration loss distribution for suburban buildings has an average value of 13.1 dB. These statistics are listed in Table III. The comparison indicates a loss for the urban group that is approximately 5 dB greater.

2. Penetration loss sensitivity to transmitter location—Different penetration losses are measured on common floors of common buildings as a function of the specific serving-site identity. These differences arise when a building is served by multiple cell sites. The distribution of the "max-min" decibel values for such multisite penetration loss differences is plotted in Fig. 2. The average max-min penetration loss difference for all the data is 5.4 dB with a sigma of 4.6 dB. This

Table III—Penetration loss

Buildings	Floor	Loss (dB)	St. Dev. (dB)	No. of Bldgs.
Urban	1	18.0	7.7	3
Suburban	1	13.1	9.5	10
All	1	14.2	9.3	13
	2	10.6	8.5	8
	3	6.8	9.8	6
	4	−0.8	10.7	4
	5	2.9	9.2	3
	8	−4.3	9.0	1
	9	−1.4	8.8	1
	12	15.3	4.9	1
	15	10.9	5.0	1

suggests that building penetration is a relatively strong function of the direction of illumination.

3. Penetration loss sensitivity to interior type—Table IV compares all interiors with and without windows for all buildings and all floors. The presence of windows is shown to reduce the average penetration loss by 6 dB, with a sigma of 5.2 dB. Not included in the data is the office area of the Chrysler building in Elk Grove where the copper-sputtered glass windows blocked out all measurable signal levels. For the windowed data category, Table IV lists penetration loss for three combinations of interior types. Open interior areas are shown to have 3 dB less penetration loss than hallways.

4. Floor height effect—A list of building penetration loss values ranked by floor height is shown in Table II. The high values of loss for the twelfth and fifteenth floors of a single building are the result of the shadowing effect of adjacent buildings. While these buildings effectively sandwiched the subject building, the adjacent street areas were relatively open. This may be a relatively common occurrence in the urban core and would appear to substantiate the observations of Durante.[1] Figure 3 is a scatter diagram with a straight line fit for penetration loss vs. floor level. The dots indicate data points that represent decibel averages of penetration losses. For each floor, there is a data point for each channel that served that floor. Only floor levels having three or more data points are included in the diagram and the building described above was omitted. The mean values for all data on each floor are indicated by "X's." The "least-squares" straight line fit to these means is also shown. The slope of the line is −1.9 dB per floor. The first floor intercept is 10.4 dB. While this loss rate agrees closely with the findings in Refs. 1 and 2, the first-floor penetration loss is about 10 dB lower. Subsequent measurements by Bell Laboratories in the Newark, N.J., Cellular Test Bed have confirmed the ranges of loss values and loss rates presented in this paper. It is assumed that differences in penetration loss with respect to values reported previously are probably due to differences in the methods of establishing the outside street-level reference.

Table IV—Penetration loss comparisons

Windowed/nonwindowed areas = 6.0 dB		(Standard Deviation = 5.2 dB)
For windowed areas:		
where enclosed areas are in conjunction with hallways		
	Enclosed/Hallways = 0 dB	(Standard Deviation = 6.0 dB)
where open areas are in conjunction with enclosed areas		
	Open/Enclosed = 1.0 dB	(Standard Deviation = 4.8 dB)
where open areas are in conjunction with hallways		
	Open/Hallways = 3.1 dB	(Standard Deviation = 5.2 dB)

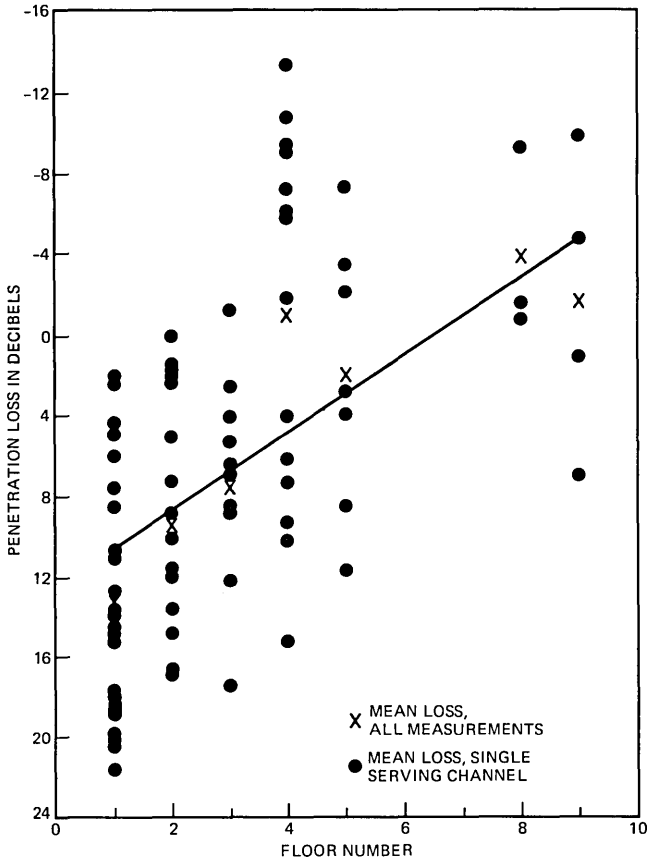


Fig. 3—Penetration loss measurements.

VIII. ACKNOWLEDGMENTS

The test team members for operations and data measurement were M. Patarca, L. F. Smith, and J. G. Rebele. Data keypunching and coding from PMS output tapes and corresponding annotation was performed by M. Patarca. The software and the processing of data to produce the results of the measurements program was completed by A. Some. K. K. Kelly collaborated in the selection and location of candidate buildings. Messrs. J. T. Kennedy, G. A. Lothian and J. R. Nevarez arranged basing the operations in Chicago and obtained the permissions necessary to access the buildings for measurements.

REFERENCES

1. J. M. Durante, "Building Penetration Loss at 900 MHz," Conference Proceedings, IEEE VTG Conf., 1973, pp. 1–7.

2. J. Dietz et. al., "Examination of the Feasibility of Conventional Land Mobile Operations at 950 MHz," FTC Office of the Chief Engineer, Research Division, Report R7102.
3. D. C. Cox, R. R. Murray, and A. W. Norris, "Measurements of 800-MHz Radio Transmissions into Buildings With Metallic Walls," B.S.T.J., this issue.
4. Special issue entitled "Advanced Mobile Phone Service," B.S.T.J., *58*, No. 1 (January 1979).
5. AT&T, Advanced Mobile Phone Service Developmental System Reports, Numbers 1–11, transmitted to the FCC by Illinois Bell, June 8, 1977–November 26, 1979.
6. L. P. Rice, "Radio Transmission into Buildings at 35 and 150 mc," B.S.T.J., *38*, No. 1 (January 1959), pp. 197–210.

APPENDIX

Building Descriptions

A.1 *Nabisco, Naperville*

Surroundings:	Open area, without nearby multistory buildings.
Construction:	Masonry with some corrugated siding. The interior has some windowed office space but a large portion is devoted to cereal processing equipment.
Cell site signals:	1st floor—Eola 2nd floor—Eola, Lemont 4th floor—Cloverdale, Eola, Lemont.

A.2 *Metropolitan Life Insurance Co., Naperville*

Surroundings:	Open area without nearby structures.
Construction:	Large glass sections alternated with large concrete slab sections.
Cell site signals:	1st floor—Cloverdale, Eola 2nd floor—Cloverdale, Eola, Lemont 3rd floor—Cloverdale, Eola, Lemont.

A.3 *Bell Laboratories, Naperville*

Surroundings:	Open area without nearby structures.
Construction:	Low light transmission glass in metal framework. Mostly windowed halls with interior office space.
Cell site signals:	1st floor—Cloverdale, Eola 3rd floor—Cloverdale, Eola 5th floor—Coverdale, Eola, Lemont.

A.4 *Bell System Center for Technical Education, Lisle*

Surroundings:	Open area without nearby structures. Located at the bottom of a hill to the East.
Construction:	Brick and glass with high-rise dormitory area. Windowed office space.
Cell site signals:	1st floor—Cloverdale, Eola, Lemont 3rd floor—Cloverdale, Eola, Lemont 5th floor—Cloverdale, Eola, Lemont, Lyons 9th floor—Cloverdale, Eola, Lemont, Lyons

A.5 *Illinois Bell, Oak Brook*

Surroundings: Stand-alone with other multistory buildings spaced several hundred yards distant.

Construction: Brick and glass windowed office space and work space.

Cell site signals: 1st floor—Cloverdale, Eola, Lemont, Lyons
2nd floor—Cloverdale, Eola, Lemont, Lyons.

A.6 *McDonalds, Oak Brook*

Surroundings: Stand-alone with other multistory buildings spaced several hundred yards distant.

Construction: Narrow windows with concrete in a ratio of approximately 50 percent. Interior office space uses half-height dividers. First floor is a lobby area.

Cell site signals: 1st floor—Cloverdale, Lemont, Lyons
2nd floor—Lyons
8th floor—Cloverdale, Eola, Lemont, Lyons.

A.7 *Chrysler, Elk Grove*

Surroundings: Open area without nearby multistory buildings.

Construction: Concrete slabs, only one floor. The interior is a large warehouse with metal storage bins for automobile parts. An attached office building had no measurable signal because copper-sputtered windows constituted 100 percent of outside surface.

Cell site signals: Cloverdale and Morton Grove.

A.8 *Allied Van Lines, Hillside*

Surroundings: Residential area without nearby multistory structures.

Construction: Concrete and glass with windowed office and work space.

Cell site signals: 1st floor—Lyons, Canal
2nd floor—Cloverdale, Lemont, Lyons
4th floor—Cloverdale, Lemont, Lyons, Canal, Morton Grove.

A.9 *Illinois Bell, Oak Park*

Surroundings: Corner buildings in area with other multistory buildings of about same height.

Construction: Brick and glass. Windowed office space and work space.

Cell site signals: 3rd floor—Lyons, Canal, Morton Grove
4th floor—Lyons, Canal, Morton Grove.

A.10 *Western Electric, Hawthorne*

Surroundings:	Corner building without nearby structures.
Construction:	Brick and smaller glass windows typical of factory construction. First floor is open area with heavy machinery; second floor is open area with assembly lines; third floor is office space with windowed space.
Cell site signals:	1st floor—Lyons 2nd floor—Lyons 3rd floor—Lyons, Canal, Morton Grove.

A.11 *Illinois Bell, Kedzie and 61st Street*

Surroundings:	Corner building in area with other multistory buildings of similar height.
Construction:	Brick and glass. Windowed office space and work space.
Cell site signals:	1st floor—Lyons, Canal, Beverly 4th floor—Lyons, Canal, Beverly.

A.12 *American Medical Association, State Street (Downtown)*

Surroundings:	Bounded on all sides by streets or parking area.
Construction:	Reflective glass windows and concrete. The interior office area is open with three-quarter partitions.
Cell site signals:	1st floor—Lyons, Canal, Morton Grove 2nd floor—Lyons, Canal, Morton Grove.

A.13 *Illinois Bell, Randolph (Downtown)*

Surroundings:	Bounded on four sides by city streets. Neighboring buildings are multistory.
Construction:	Concrete and glass with lobby on first floor and windowed work and office spaces on upper floors. Office partitions are three-quarter-high dividers.
Cell site signals:	1st floor—Canal 3rd floor—Canal 5th floor—Canal, Morton Grove, Beverly.

A.14 *Bell Training Center, West Adams (Downtown)*

Surroundings:	Neighboring multistory buildings in a business district. The building front is open to the street and both sides abut neighboring buildings. The rear of the building opens into a service courtyard.
Construction:	Concrete and glass. Interior upper floors have windowed office space.

Cell site signals: 1st floor—Canal
2nd floor—Canal
12th floor—Canal
15th floor—Canal.

A.15 *Private Home, Lombard*

Surroundings: Residential neighborhood.
Construction: Aluminum siding.
Cell site signals: Cloverdale, Eola, Lemont, Lyons.

The outside reference measurements for this building were incomplete because a complete perimeter path was not available. The measurements for this building are not combined with the results from other buildings.

The penetration loss for this building based on the available street-level reference measurements is 7.3 dB with a sigma of 6.7 dB.

AUTHOR

Edward H. Walker, B.S.E.E., 1956, Newark College of Engineering; Western Electric Company, 1940–1957; Bell Laboratories, 1958–1981. Mr. Walker has been involved in the design of radio circuits and the development of military radar systems. Since 1973 he has been responsible for the experimental evaluation of audio quality of the AMPS System and of mobile and portable terminals. His work on building penetration was completed prior to his retirement from Bell Laboratories in 1981. Member, IEEE.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 9, November 1983
Printed in U.S.A.

Transmission Errors and Forward Error Correction in Embedded Differential Pulse Code Modulation

By D. J. GOODMAN* and C.-E SUNDBERG†

(Manuscript received December 11, 1982)

We have derived formulas for the combined effects of quantization and transmission errors on the performance of embedded Differential Pulse Code Modulation (DPCM), a source code that can be used for variable-bit-rate speech transmission. Our analysis is more general and more precise than previous work on transmission errors in digital communication of analog signals. Special cases include conventional DPCM and Pulse Code Modulation (PCM). Our main result is a signal-to-noise ratio formula in which the effects of source characteristics (input signal, codec design parameters) and the effects of transmission characteristics (modulation, channel, forward error correction) are clearly distinguishable. We also present, in computationally convenient forms, specialized formulas that apply to uncoded transmission through a random-error channel, transmission through a slowly fading channel, and transmission with part or all of the DPCM signal protected by an error-correcting code. Numerical results show how channel coding can have different effects on conventional and embedded DPCM. They also show how the binary-number representation of quantizer outputs influences performance.

I. INTRODUCTION

1.1 Embedded Differential Pulse Code Modulation

Embedded coding can play a valuable role in variable-bit-rate speech transmission. With an embedded code the analog-to-digital (a/d) and

* Bell Laboratories. † University of Lund, Sweden.

digital-to-analog (d/a) converters operate at a constant, high bit rate, and the transmission system controls the instantaneous rate. Proposed applications for variable-bit-rate operation include a digital private branch exchange,[1] digital speech interpolation,[2] packet-switched voice transmission,[3] and mobile radio.[4]

Sophisticated versions of Differential Pulse Code Modulation (DPCM) are promising speech codes for these and other environments.[5–7] However, conventional DPCM is not suited to variable-bit-rate transmission because the decoder amplifies the effects of bit-rate adjustments. On the other hand, a slightly modified form of DPCM avoids this problem and produces an embedded code.[8]

Figure 1 shows the codec (coder, decoder) structure of embedded DPCM. Although up to E bits/sample can be transmitted, the signals presented to the two integrators have a resolution of only M bits/sample, the minimum bit rate of the channel. While Fig. 1 is a useful guide to practical implementations, Fig. 2, which is equivalent, is easier to analyze. It shows the quantizer at the encoder as a successive-approximation combination of two quantizers: a "minimal" quantizer with M bits/sample and a "supplemental" quantizer with E–M bits/sample, operating on the error signal of the minimal quantizer.*

In embedded DPCM, all of the bits from the minimal quantizer arrive at the decoder; the transmission system can delete some or all of the supplemental bits. With S bits/sample of the supplemental quantizer transmitted to the decoder, the rate is $D = M + S$ bits/sample, and the quantizing distortion is very close to that of a conventional codec with D bits/sample.

Errors in the two bit streams have different effects on the decoder output. Errors in the M, minimal bits, are enhanced by the decoder integrator, which has no effect on errors in the S, supplemental bits. This situation compares favorably with conventional DPCM, where all errors are integrated at the decoder. It also has implications for forward error correction in embedded DPCM. Figures 1 and 2 will be further explained in Section II.

1.2 The scope of this paper

Our principal contribution in this paper is an analysis of the combined effects of granular quantizing distortion and transmission errors on the mean-square error of embedded DPCM. The analysis is quite general: special cases include Pulse Code Modulation (PCM) ($M = 0$) and conventional DPCM ($S = 0$). The formulas for the noisy-

* Based on the structure of Fig. 2, Jayant has recently described an enhanced supplemental quantizer (called an explicit noise coder) that uses memory and delay to improve speech quality.[9]

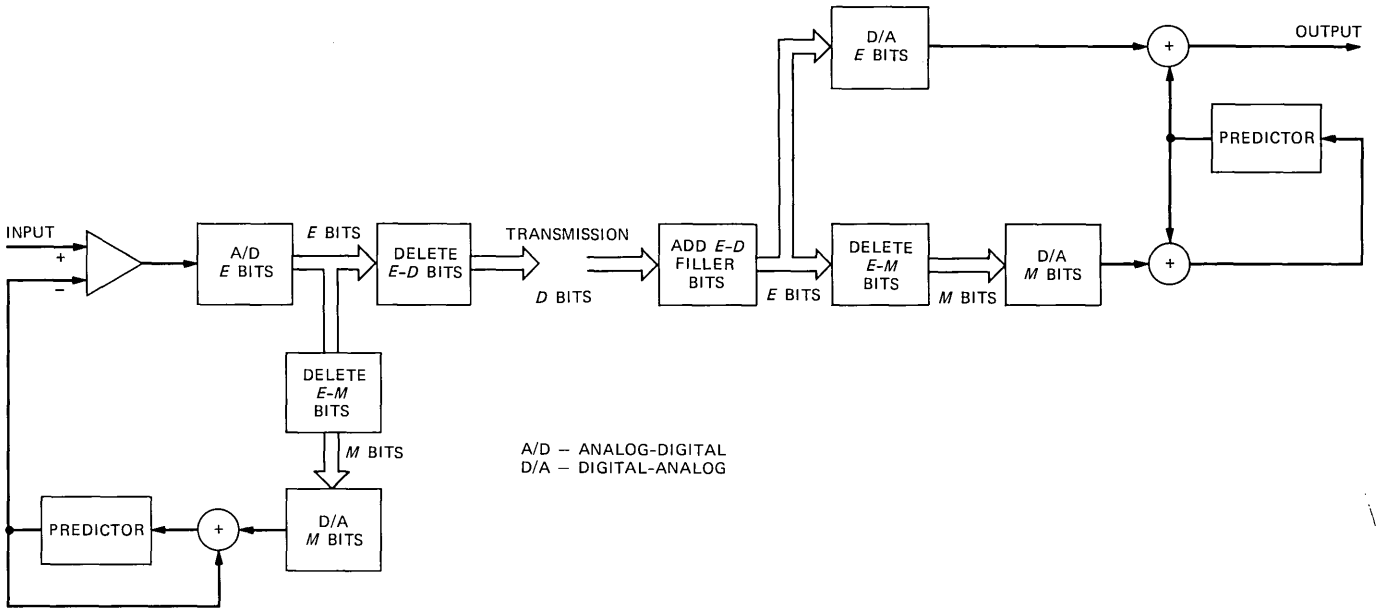


Fig. 1—Embedded DPCM encoder and decoder. Because both predictors operate on the same signal (with resolution M bits/sample) performance is unaffected by errors due to bit-rate adjustment. The channel can transmit between M bits/sample and E bits/sample.

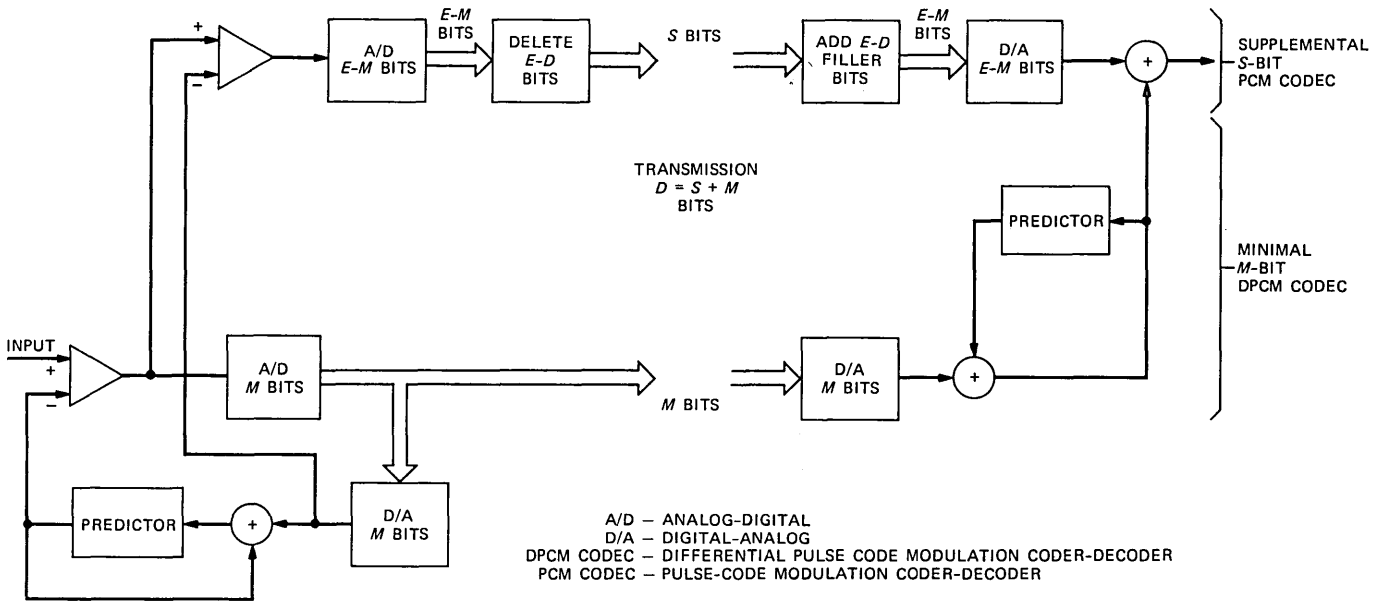


Fig. 2—Embedded DPCM encoder and decoder. *E*-bit analog-to-digital conversion is shown as a two-stage, successive-approximation process. All *M* minimal bits are transmitted. *S*, the number of supplemental bits transmitted, can range from 0 to *E*–*M*.

channel performance of conventional DPCM are new, and the specialization to PCM is more precise than previous work on the subject,[10-12] which includes several approximations that are accurate for multibit ($\geq$ 6-bit) quantizers, but are rather imprecise at lower bit rates. The method of analysis has the advantage of separating source effects from transmission effects. The source effects include the characteristics of the analog input signal and the codec design parameters. The transmission effects include modulation and demodulation, the channel, forward error correction, and diversity reception.

The main result is eq. (61), in which the transmission effects are contained in the discrete probability function $P(l)$, where l is an index of binary error patterns. The other symbols in (61) are source parameters and functions of source parameters. After deriving (61) we apply it to specific transmission environments and present, in Table VI, specialized formulas that are convenient for numerical computation.

In all, there are 78 formulas in Sections III through VI, most of them intermediate steps in derivations of a few key results. Anticipating that few readers will require all these details we provide here a summary of the analysis and we display a few numerical results. Sections I, II, and VII contain the main ideas of our work and sufficient information to allow readers to perform, on hand calculators, computations similar to the ones we present.

Section III introduces the notations for the signals and errors in the M-bit minimal DPCM codec and the S-bit supplementary PCM codec of Fig. 2. The analysis of Section III leads to (2), which expresses the sampled-data error sequence as a function of quantization errors, transmission errors, and integrator characteristics. Section IV begins the analysis of the mean-square value of (2) by deriving (35), the ratio of the mean-square codec input to the mean-square value of the encoder difference signal. Section V defines A factors, which are conditional mean squares of the errors due to specific binary error patterns, and derives (61), the general signal-to-noise ratio (s/n) formula. Section VI adapts (61) to specific transmission models and provides guides to numerical computation.

1.3 Examples of numerical results

Figure 3 shows the performance of embedded DPCM in four transmission environments, all of them employing Coherent Phase Shift Keying (CPSK) modulation at 32 kb/s in a white-Gaussian-noise channel. The encoder operates at 32 kb/s (8-kHz sampling, 4 bits/sample), and in format 1 all of this information is transmitted. Figure 3 indicates that when the channel s/n falls below 10 dB, the audio s/n deteriorates rapidly. In format 2, the least significant bit of each DPCM code word is deleted, and the remaining 3 bits/sample are

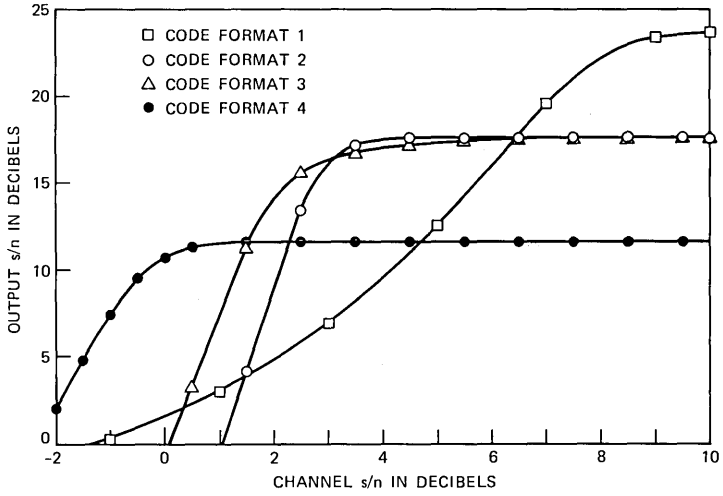


Fig. 3—Performance of embedded DPCM in four transmission environments.

protected by a rate 3/4 convolutional code. Although there is more quantizing noise than in format 1 (the s/n is 6 dB lower in the absence of transmission errors), the convolutional code permits accurate reception of the transmitted bit stream at channel s/n's down to 3 dB. Going one step further with this approach to channel coding, we have format 4, in which 16 kb/s of speech data are transmitted under the protection of a rate 1/2 code. The threshold of essentially error-free performance is now extended down to a channel s/n of about 0 dB.

In code format 3 the speech transmission rate is 24 kb/s, as in format 2, but now only 2 of the 3 bits/sample are protected by the convolutional code, which has rate 2/3. The threshold of curve 3 in Fig. 3 is about 1 dB lower than that of curve 2. On the other hand, format 3 is slightly worse than format 2 in intermediate channel conditions (s/n's between 3 and 5 dB). Over this range, format 2 is essentially error free, while format 3 is affected by errors in the unprotected third bit of each code word. The effect is small, however, because these errors are not amplified at the decoder.

With conventional, rather than embedded, DPCM, the corresponding picture, Fig. 4, is rather different, especially with respect to format 3. Here channel errors in the unprotected third bit are amplified by the integrator at the decoder. The result is a noticeably lower output s/n relative to format 2 (all three bits protected) when the channel s/n is between 3 and 6 dB. On the other hand, in clear channels the greater accuracy of prediction in the conventional encoder causes the output s/n of conventional DPCM at 24 kb/s (formats 2 and 3) and

32 kb/s (format 1) to be about 0.7 dB higher than that of embedded DPCM.

Figure 5, which applies to 24 kb/s speech transmission with a rate 2/3 code, summarizes the performance differences between conventional and embedded DPCM. Conventional DPCM has somewhat lower quantizing noise, which is reflected in the higher s/n in good channels. In intermediate conditions, when errors in the unprotected

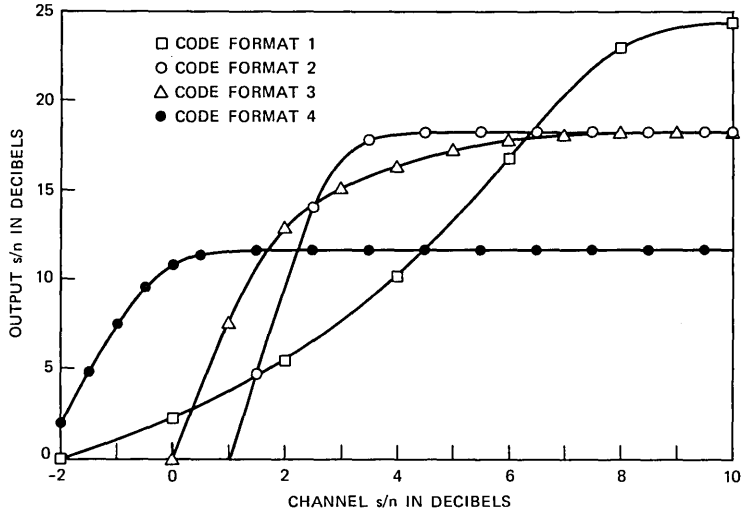


Fig. 4—Performance of conventional DPCM in four transmission environments.

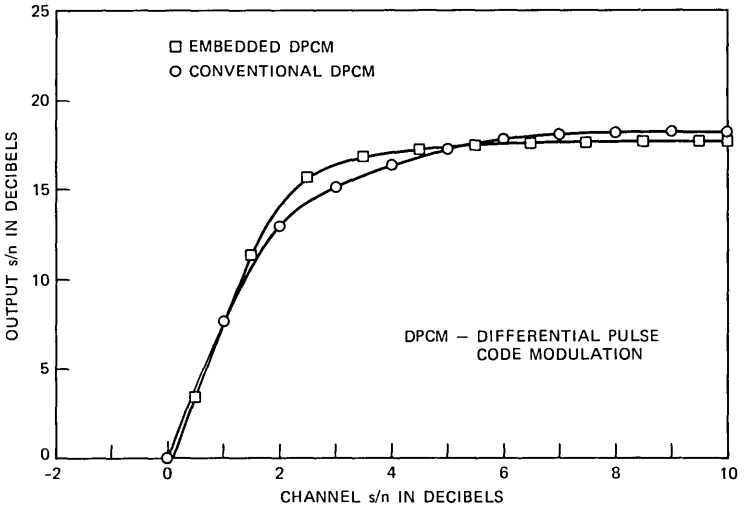


Fig. 5—A 24-kb/s speech transmission with a rate of 2/3 code (Format 3) used to summarize performance differences between embedded and conventional DPCM.

bit influence performance, embedded DPCM is better because the effects of these errors are not amplified at the decoder. In very difficult channels, errors in the two coded bits dominate performance and the s/n's of conventional and embedded DPCM are virtually equal.

II. EMBEDDED DPCM SIGNAL PROCESSING

Figure 1 shows the signal processing operations that take place in embedded DPCM encoding, transmission, and decoding. While the analog-to-digital converter at the encoder generates E bits/sample, the resolution of the signal presented to the integrator is limited to M bits/sample, where M is the minimum bit rate of the transmission system. The transmitted bit rate, D, can vary between M and E. At the receiver E–D filler bits are appended to the incoming signal. As in the encoder, E–M bits are deleted at the integrator input so that in the absence of transmission errors the encoder integrator and the decoder integrator produce the same approximation signal. When this signal is added to the full-resolution (D bits) quantizing error, the sum has nearly the quality of a conventional DPCM signal with D bits/sample.

While Fig. 1 demonstrates practical implementations, Fig. 2, which is equivalent, is easier to analyze. It represents the analog-to-digital conversion as a two-step, successive-approximation process. First the input to the converter is represented by M bits/sample. Then the error of this representation is processed by another analog-to-digital converter with E–M bits/sample. Taken together, the two digital signals comprise an E bits/sample representation of the DPCM difference signal. All of the M bits of the minimal analog-to-digital converter are transmitted. The other E–M bits are subject to deletion by the transmission system. At the receiver the minimal, M-bit signal is processed by a conventional DPCM decoder. The result is added to the supplemental, $S = D - M$ bit representation of the DPCM error signal to produce the system output.

III. SIGNAL ANALYSIS

3.1 Error sequence

To analyze Fig. 2, we introduce Fig. 6, which shows the signals that appear in the analysis and defines their notations. We are interested in the overall error signal

$$e(k) = x'(k) - x(k), \tag{1}$$

the difference between decoder output and encoder input. In particular we will derive the formula

$$e(k) = n_D(k) + e_D(k) + \sum_{i=1}^{\infty} b_i e_M(k - i), \tag{2}$$

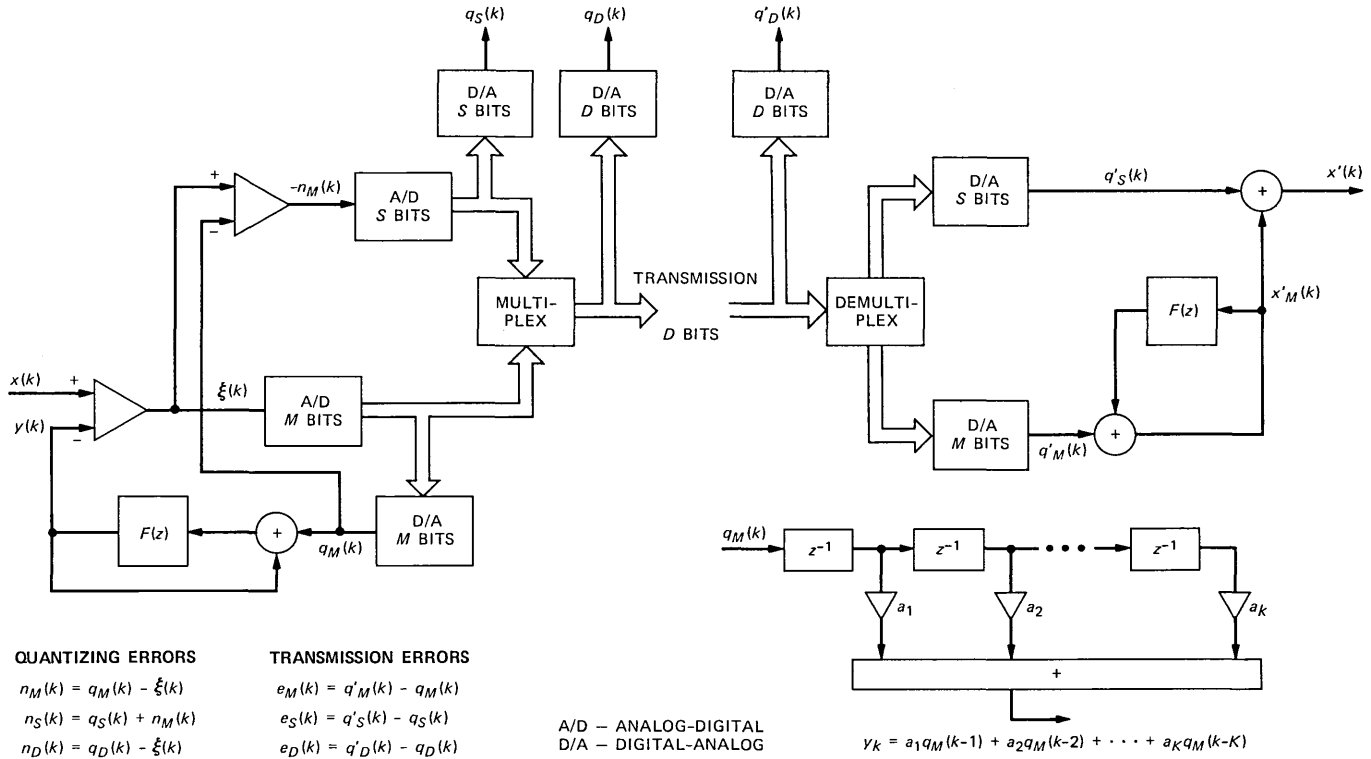


Fig. 6—The signal-processing operations of Fig. 2 and the notation used to analyze them. Three new digital-to-analog converters define signals that appear in the analysis.

where $n_D(k)$ is the quantization noise of the two-stage, D-bit analog-to-digital conversion, $e_D(k)$ is the effect of a transmission error on the entire D-bit transmitted code word, and $e_M(k)$ is the effect of a transmission error on the minimal, M-bit DPCM code word. The coefficients b_i are related to the predictor coefficients $a_1, a_2, \cdots a_K$ according to (12).

Formally,

$$e_M(k) = q'_M(k) - q_M(k), \tag{3}$$

the difference between the quantized inputs to the decoder and encoder integrators. To define $n_D(k)$ and $e_D(k)$, we view the combined code word with $M + S = D$ bits as a digital representation of $\xi(k) = x(k) - y(k)$. A D-bit digital-to-analog converter would produce the quantized signal $q_D(k)$, and so we have the definition of quantization error:

$$n_D(k) = q_D(k) - \xi(k). \tag{4}$$

At the receiver, where the D bits/sample are possibly corrupted by transmission errors, a digital-to-analog converter would produce $q'_D(k)$. The transmision error is

$$e_D(k) = q'_D(k) - q_D(k). \tag{5}$$

In the remainder of Section III we derive eq. (2); in Section IV and V we analyze its mean square.

3.2 Derivation of the error sequence

Here the signal analysis is facilitated by the transform notation of Table I. The reader may verify that the output of the minimal encoder, $Q_M(z)$, is related to input and quantizing noise[8] by

$$Q_M(z) = [X(z) + N_M(z)][1 - F(z)]. \tag{6}$$

Table I—Transform notation for codec signals

Encoder Signal	Description	Decoder Signal
$X(z)$	Encoder input, decoder output	$X'(z)$
	Minimal decoder output	$X'_M(z)$
$Y(z)$	Approximation signal	$Y'(z)$
$Q_M(z)$	M-bit representation of $X(z) - Y(z)$	$Q'_M(z)$
$N_M(z)$	$Q_M(z) - [X(z) - Y(z)]$ Quantizing error in $Q_M(z)$	
	$Q'_M(z) - Q_M(z)$ Transmission error in $Q'_M(z)$	$E_M(z)$
$Q_S(z)$	S-bit representation of $-N_M(z)$	$Q'_S(z)$
$N_D(z)$	$Q_S(z) + N_M(z)$ Quantizing error in $Q_S(z)$	
	$Q'_S(z) - Q_S(z)$ Transmission error in $Q'_S(z)$	$E_S(z)$
$F(z)$	Predictor $\sum_{i=1}^{K} a_i z^{-1}$	$F(z)$

In the minimal decoder

$$X'_M(z) = \frac{Q'_M(z)}{1 - F(z)}, \tag{7}$$

which leads to

$$X'_M(z) = X(z) + N_M(z) + \frac{E_M(z)}{1 - F(z)}. \tag{8}$$

In the supplemental encoder,

$$Q_S(z) = -N_M(z) + N_D(z), \tag{9}$$

and at the decoder

$$Q'_S(z) = -N_M(z) + N_D(z) + E_S(z). \tag{10}$$

Combining (8) and (10) we have the output of the entire decoder,

$$X'(z) = X'_M(z) + Q'_S(z) = X(z) + N_D(z) + E_S(z) + \frac{E_M(z)}{1 - F(z)}. \tag{11}$$

To transform (11) to time-domain notation, we defined b_i, $i = 0, 1, 2, \cdots$, to be the inverse z transform of $1/(1 - F)$, such that

$$\frac{1}{1 - F(z)} = \sum_{i=0}^{\infty} b_i z^{-i}. \tag{12}$$

Then we have

$$x'(k) = x(k) + n_D(k) + e_S(k) + \sum_{i=0}^{\infty} b_i e_M(k - i), \tag{13}$$

and the error is

$$\begin{aligned} e(k) &= x'(k) - x(k) \\ &= n_D(k) + e_S(k) + e_M(k) + \sum_{i=1}^{\infty} b_i e_M(k - i), \end{aligned} \tag{14}$$

where we have substituted $b_0 = 1$. Equation (14) is identical to (2) because

$$e_D(k) = e_M(k) + e_S(k). \tag{15}$$

IV. MEAN-SQUARE ERROR

The square of (2) is

$$\begin{aligned} e^2(k) = [n_D(k) + e_D(k)]^2 &+ \sum_{i=1}^{\infty} b_i^2 e_M^2(k - i) \\ &+ 2 \sum_{i=1}^{\infty} b_i [n_D(k) + e_D(k)] e_M(k - i) \\ &+ 2 \sum_{i=1}^{\infty} \sum_{j=i+1}^{\infty} b_i b_j e_M(k - i) e_M(k - j). \end{aligned} \tag{16}$$

To analyze the mean value of (16), we assume that the sequence $\{x(k)\}$ is drawn from a stationary ergodic random process. In our derivations we ignore all correlations in (16) between nonsimultaneous samples. That is, we assume

$$E\{[n_D(k) + e_D(k)]e_M(k - i)\} \approx 0; \qquad i \geq 1 \tag{17}$$

and

$$E\{e_M(k - i)e_M(k - j)\} \approx 0; \qquad i \neq j. \tag{18}$$

Equation (17) indicates that the overall error (quantizing plus channel distortion) in the kth sample is uncorrelated with errors in other samples of the minimal M-bit quantized samples. Equation (18) states that errors in different minimal samples are uncorrelated. These approximations are accurate because the sequence of samples at the input to a DPCM quantizer is decorrelated by the differential coding process and because transmission errors affecting different code words are independent or only weakly correlated.

The approximations, (17) and (18), remove the last two sums from the expected value of (16), leaving

$$E\{e^2(k)\} = E\{[n_D(k) + e_D(k)]^2\} + b_P E\{e_M^2(k)\}, \tag{19}$$

in which we summarize the influence of the predictor in

$$b_P = \sum_{i=1}^{\infty} b_i^2. \tag{20}$$

The expectations in (19) are related to the quantization and transmission of $\xi(k)$, the DPCM difference signal. In Section V, we present a complete theory of the errors due to these operations. While this theory relates these errors to $\sigma_\xi^2 = E\{\xi^2(k)\}$, we are ultimately interested in the s/n of the codec input, $x(k)$:

$$s/n = E\{x^2(k)\}/E\{e^2(k)\} = \sigma_x^2/\sigma_e^2. \tag{21}$$

To find this quantity, we will now derive σ_x^2/σ_ξ^2 and then combine it with the results of Section V. To begin the derivation, we refer to Fig. 6 and verify

$$Y(z) = F(z)[X(z) + N_M(z)], \tag{22}$$

which leads to

$$\xi(k) = x(k) - y(k) = x(k) - \sum_{i=1}^{K} a_i x(k - i) - \sum_{i=1}^{K} a_i n_M(k - i). \tag{23}$$

We write the mean-square value of (23) as

$$\sigma_\xi^2 = E\left\{\left[x(k) - \sum_{i=1}^{K} a_i x(k - i)\right]^2\right\} + E\left[\sum_{i=1}^{K} a_i n_M(k - i)\right]^2 - 2E\left\{\left[x(k) - \sum_{i=1}^{K} a_i x(k - i)\right]\left[\sum_{i=1}^{K} a_i n_M(k - i)\right]\right\}. \tag{24}$$

The first term in (24) depends on the spectral properties of $x(k)$ and on the predictor. The ratio of σ_x^2 to this quantity is called the prediction gain,

$$G = \sigma_x^2/E\left\{\left[x(k) - \sum_{i=1}^{K} a_i x(k-i)\right]^2\right\}. \tag{25}$$

It indicates that extent to which the predictor (in the absence of quantization) reduces the mean-square value of the signal to be quantized. Formally we have

$$G^{-1} = \sum_{j=0}^{K}\sum_{i=0}^{K} a_i' a_j' r_{i-j}, \tag{26}$$

in which $a_0' = 1$, $a_i' = -a_i$, $i = 1, 2, \cdots, K$, and r_n is a normalized covariance coefficient of the stationary input,

$$r_n = E\{x(k)x(k+n)\}/\sigma_x^2. \tag{27}$$

In evaluating the second and third terms of (24), we ignore correlation between different quantizing-noise samples and correlations between quantizing-noise samples and samples of the codec input. Thus we use (18) and the approximation

$$E\{n_M(k)x(j)\} \approx 0. \tag{28}$$

This allows us to write

$$\sigma_\xi^2 = G^{-1}\sigma_x^2 + a_P E\{n_M^2(k)\}, \tag{29}$$

where we define

$$a_P = \sum_{i=1}^{K} a_i^2. \tag{30}$$

The noise component of (29), which depends on the quantizer overload point and on the statistical properties of $\xi(k)$, is analyzed in Section V, where we restrict our attention to granular quantizing noise and derive $\sigma_q^2(B)$, the noise power of a B-bit quantizer with unity overload point. If the actual overload point is ξ_{max}, the noise power of the M-bit minimal quantizer is

$$E\{n_M^2(k)\} = \xi_{max}^2\sigma_q^2(M) \approx \Delta_M^2/12. \tag{31}$$

The quantizer step size is

$$\Delta_M = \xi_{max}2^{-(M-1)}, \tag{32}$$

and the approximation would be exact if $n_M(k)$ were uniformly distributed over the range $-\Delta_M/2$ to $\Delta_M/2$. Table II presents $\sigma_q^2(B)$ numerically and indicates the fractional error due to the above approximation. A parameter in Table II is the dimensionless load factor

Table II—Quantizing noise

Load Factor		Gaussian		Exponential	
		2 bits	3 bits	2 bits	3 bits
1.78	Noise power $\sigma_q^2(B)$	0.02066	0.005198	0.02189	0.005276
	Approximation error*	−0.01	0.00	0.05	0.01
3.16	Noise power $\sigma_q^2(B)$	0.02082	0.005207	0.02394	0.005419
	Approximation error*	0.00	0.00	0.13	0.04
5.62	Noise power $\sigma_q^2(B)$	0.02292	0.005209	0.02885	0.005836
	Approximation error*	0.09	0.00	0.28	0.11

* Relative error of the approximation $\sigma_q^2(B) \approx 2^{-2B}/3$.

$$L = \xi_{max}/\sigma_\xi. \tag{33}$$

Combining (29), (31), and (33), we arrive at

$$\sigma_\xi^2 = G^{-1}\sigma_x^2 + a_P L^2 \sigma_q^2(M)\sigma_\xi^2, \tag{34}$$

or the quantity we set out to derive:

$$\sigma_x^2/\sigma_\xi^2 = G[1 - a_P L^2 \sigma_q^2(M)]. \tag{35}$$

V. QUANTIZATION NOISE AND TRANSMISSION NOISE IN PCM

5.1 Granular and overload conditions

To analyze (19), we study, statistically, the quantization and transmission of the DPCM difference signal $\xi(k)$. In this type of study it is customary to separate the quantizing error into two components: overload distortion and granular noise. In speech communication this distinction is valuable for predicting subjective quality.[13,14] Moreover, in analyzing DPCM the distinction is essential because, except for a codec with an ideal integrator,[15] ($F(z) = z^{-1}$, which is pathologically vulnerable to transmission errors), there is no theory for computing the mean-square slope-overload distortion. Thus our analysis separates the transmission of clipped samples of $\xi(k)$ from samples subject to granular distortion. Our theory pertains only to the transmission of unclipped samples. For those samples we add two different distortions, quantizing noise and noise due to transmission errors. Unlike slope overload, both of these impairments are essentially uncorrelated with the signal. This gives us confidence that the mean-square sum is a reasonable quality measure.

Formally, we rewrite (19) as

$$\sigma_e^2 = p_{ov}E\{e^2(k) \,|\, |\xi(k)| > \xi_{max}\} + p_{gr}E\{e^2(k) \,|\, |\xi(k)| \leq \xi_{max}\}, \tag{36}$$

where $\xi(k)$ is the quantizer input and ξ_{max} is the overload point of the uniform DPCM quantizer. The probability of overload is p_{ov} and $p_{gr} = 1 - p_{ov}$ is the probability of granular quantization. By definition, the

quantizer is overloaded at time k if the quantization error exceeds half of a quantization step, i.e., if

$$|q_D(k) - \xi(k)| > \Delta_D/2. \tag{37}$$

The step size of the D-bit uniform quantizer is

$$\Delta_D = \xi_{\max} 2^{-(D-1)} = \Delta_M/2^{D-M}. \tag{38}$$

Our remaining analysis will be confined to the second expectation on the right side of (36) and in particular to the ratio,

$$s/n = \sigma_x^2/E\{e^2(k) \,|\, |\xi(k)| \leqslant \xi_{\max}\}. \tag{39}$$

To be concise in the remainder of this paper, we will omit the granular condition, $|\xi(k)| \leqslant \xi_{\max}$, from our notation of expected values.

5.2 Transmission model, normalized quantizer

To facilitate numerical evaluation of s/n's, we will present three tables of normalized error terms. The normalization relates these errors to a quantizer with a unity overload point and an input with probability density function $p_u(\cdot)$. If the quantizer of interest has an overload point of $\xi_{\max}$ and the input has the probability density $p_\xi(\cdot)$, the relevant errors are table entries scaled by $\xi_{\max}^2$. The two probability densities are related by

$$p_u(u) = \xi_{\max} p_\xi(\xi_{\max} u). \tag{40}$$

To confine our attention to the granular quantization condition, we perform our averages with respect to the conditional probability density

$$\begin{aligned} p_{\text{gr}}(u) &= \frac{p_u(u)}{\int_{-1}^{1} p_u(u)du}, \qquad |u| \leqslant 1 \\ &= 0 \qquad\qquad\qquad\quad |u| > 1. \end{aligned} \tag{41}$$

The model is illustrated in Fig. 7. The signal $u = \xi/\xi_{\max}$ is processed by a B-bit analog-to-digital converter with overload point 1 and step size

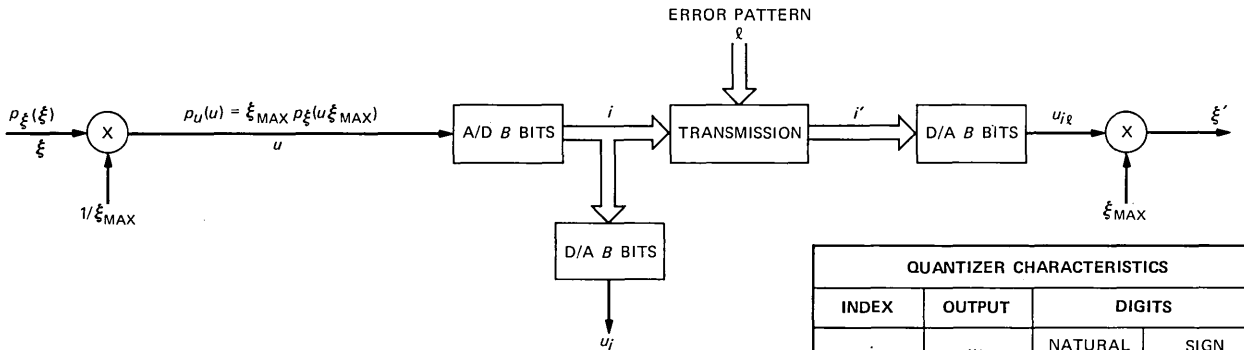
$$\Delta_B = 2^{-(B-1)}. \tag{42}$$

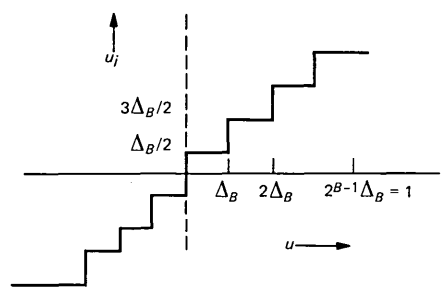
The digital output of the a/d is i, and the corresponding quantized signal is u_i, which is related to u by the graph in Fig. 7 and by

$$u_i = -1 + (i + 0.5)\Delta_B \quad \text{when}$$

$$-1 + i\Delta_B \leqslant u < -1 + (i + 1)\Delta_B, \qquad i = 0, 1, \cdots, 2^B - 1. \tag{43}$$

In Fig. 7, the B-bit code word i is transmitted, and i' is received,





$A_\ell(B) = E\{(u_{i\ell} - u_i)^2\}$

$\tilde{A}_\ell(B) = E\{(u_{i\ell} - u)^2\} - E\{(u_i - u)^2\}$

A/D – ANALOG-DIGITAL
D/A – DIGITAL-ANALOG

QUANTIZER CHARACTERISTICS			
INDEX	OUTPUT	DIGITS	
i	u_i	NATURAL BINARY	SIGN MAGNITUDE
2^B-1	$1-\Delta_B/2$	111	011
⋮	⋮	⋮	⋮
$2^{B-1}+1$	$\Delta_B/2$	100	000
2^{B-1}	$-\Delta_B/2$	011	100
⋮	⋮	⋮	⋮
1	$-1+3\Delta_B/2$	001	110
0	$-1+\Delta_B/2$	000	111

Fig. 7—Model for analyzing quantizing noise and the effects of transmission errors. The graph and table define the quantizer and two binary number representations.

with the transformation of i to i' characterized by a binary-error pattern with index l, l being an integer in the range 0, $2^B - 1$. To relate the error effect to l, we refer to the natural-binary representation of l and specify that a 1 in the bth *least*-significant position of l causes an inversion of the bth *most*-significant* bit in the binary representation of i. Thus $l = 0$ refers to error-free transmission ($i = i'$); $l = 1$ refers to an error only in the most significant bit; $l = 5$ refers to errors in the first and third most significant bits; etc.

With u the quantizer input and l the binary error pattern, we denote the received sample in Fig. 7, u_{il}. It is helpful to separate the complete error $u_{il} - u$ into quantization-noise and transmission-noise components as follows,

$$u_{il} - u = (u_i - u) + (u_{il} - u_i). \tag{44}$$

5.3 Conditional expectations of transmission-error effects

5.3.1 The general approach

Our goal is to evaluate the mean-square of (44) over the joint distribution of input statistics and binary-error patterns. The key to our analysis is the definition of A factors, which are conditional mean-square errors, each related to a specific binary-error pattern, l. By analyzing these conditional errors, we separate the effects of source characteristics from the effects of transmission characteristics. The source effects are embodied in the A factors; the transmission effects are embodied in probabilities of error patterns. These probabilities govern the weighted addition of the A factors to produce the final result.

This approach to analyzing transmission impairments was introduced by Rydbeck and Sundberg,[10–12] who were mainly concerned with quantizers with 6 to 8 bits/sample. This high resolution admitted various approximations that are inaccurate in the 2- to 4-bit quantizers of greatest interest for embedded DPCM transmission. Thus we proceed to a precise calculation of two types of A factors: conditional expectations related to the isolated effects of digital transmission errors and conditional expectations that include correlations between transmission errors and quantizing noise. In high-resolution quantizers this correlation is negligible, and the two types of A factors are essentially equal.

5.3.2 Analysis

To compute the mean-square value of (44), conditioned on error pattern l, we will identify three important quantities: $\sigma_q^2(B)$, the

* This reversal of the bit ordering of l relative to the binary representation of i will facilitate bookkeeping in subsequent computations.

granular-noise power of a B-bit quantizer; $A_l(B)$, the mean-square effect of error pattern l on the quantized signal u_i; and $\tilde{A}_l(B)$, the overall effect of error pattern l on the mean-square error of the analog output u_{il}.

To derive computationally convenient expressions for $\sigma_q^2(B)$, $A_l(B)$, and $\tilde{A}_l(B)$, we defined the integrals

$$p_i = \int_{v_i}^{v_{i+1}} p_{gr}(u)du \tag{45}$$

$$q_i = \int_{v_i}^{v_{i+1}} (u_i - u)p_{gr}(u)du \tag{46}$$

$$\sigma_{gr}^2 = \int_{-1}^{1} u^2 p_{gr}(u)du, \tag{47}$$

in which v_i is the lower boundary and v_{i+1} is the upper boundary of quantizing interval i:

$$v_i = -1 + i2^{-(B-1)}; \qquad i = 0, 1, \cdots, 2^B. \tag{48}$$

The first integral (45) is the probability of using interval i. The second integral (46) is the average quantization error in interval i. If B is large, Δ_B is small, and $q_i \approx 0$ because u_i is in the center of the quantization interval. The third integral (47) is the mean-square signal when the quantizer is in the granular condition.

Now we write the definitions followed by computational formulas for the quantizing noise and the effects of error pattern l:

$$\sigma_q^2(B) = E(u_i - u)^2 = \sigma_{gr}^2 + \sum_{i=0}^{2^B-1} (2q_i u_i - p_i u_i^2) \tag{49}$$

$$A_l(B) = E(u_{il} - u_i)^2 = \sum_{i=0}^{2^B-1} p_i(u_{il} - u_i)^2 \tag{50}$$

$$\tilde{A}_l(B) = E(u_{il} - u)^2 - E(u_i - u)^2 = A_l(B) + 2 \sum_{i=0}^{2^B-1} q_i(u_{il} - u_i). \tag{51}$$

$\tilde{A}_l(B)$, the difference between the total noise and the quantizing noise, includes the correlation between quantization effects and transmission-error effects. In multibit quantizers ($B > 4$) this correlation is small, and $\tilde{A}_l(B) \approx A_l(B)$, an assumption inherent in previous work on PCM. Because low-resolution quantizers are of interest in DPCM, we take account of this correlation in our present work.

Finally, we combine (49), (50), and (51) to write the mean-square value of (44)

$$\epsilon_l^2 = E(u_{il} - u)^2 = \sigma_q^2(B) + \tilde{A}_l(B). \tag{52}$$

5.3.3 Computations

Table III displays formulas for p_i, q_i, and σ_{gr}^2 that apply to inputs with Gaussian and exponential probability density functions. Because the input, u, of the normalized quantizer is related to the input, ξ, of the quantizer with overload point ξ_{max} by $u = \xi/\xi_{max}$, we have

$$\sigma_u^2 = \sigma_\xi^2/\xi_{max}^2 = 1/L^2, \tag{53}$$

where L is the dimensionless load factor defined in (33). Because L is a familiar quantizer design quantity, we have written the formulas in Table III as functions of L.

With the formulas in Table III, it is a simple matter to compute $\sigma_q^2(B)$ precisely. However, for $B \geq 4$ the approximation

$$\sigma_q^2(B) \approx \Delta_B^2/12 = 2^{-2B}/3 \tag{54}$$

is very accurate (within 3 percent of the exact value for $L \leq 5.6$). For $B = 2$ and 3, Table II shows the exact values of $\sigma_q^2(B)$ and the approximation errors

$$[\sigma_q^2(B) - \Delta_B^2/12]/\sigma_q^2(B) \tag{55}$$

for $L = 1.78, 3.16, 5.62$ ($L^2 = 10 \pm 5$ dB).

To compute $A_l(B)$, $\tilde{A}_l(B)$, it is necessary to know $u_{il} - u_i$, which depends on the binary number representation of u_i.

5.3.4 Binary number representations

We consider two representations: natural-binary and sign-magnitude, both defined in Fig. 7. Although in general the $A_l(B)$ and $\tilde{A}_l(B)$ depend on p_i and q_i, there are some special cases that are important and illuminating. For example, in the natural-binary code, the single-error A factors are independent of the signal statistics and of the quantizer. An error in the most significant bit causes $u_{il} - u_i = \pm 1$ provided it is the only error in the B-bit code word. Thus $A_1(B) = 1$. Likewise, any isolated error in the second most significant bit causes an output error of $\pm 1/2$, and in general a single binary error in position b causes the mean-square error

$$A_l(B) = (1/4)^{b-1}; \qquad l = 2^{(b-1)}. \tag{56}$$

In the sign-magnitude code, isolated binary errors in positions $b = 2, 3, \cdots, B$ have the same effects as corresponding errors in the natural-binary code. However, an isolated error in the most significant position transforms u_i to $u_{il} = -u_i$. The mean-square effect is

$$A_l(B) = 4E\{u_i^2\} \approx 4\sigma_u^2. \tag{57}$$

The approximation becomes more and more precise as B increases.

Table III—Formulas for computing A factors and quantizing noise B-bit transmission

General
$v_i = -1 + i2^{-(B-1)}; u_i = v_i + 2^{-B}$

	Gaussian	Exponential
Error integral, $Q(x)$	$\int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$	
Probability density, $p(u)$	$\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(uL)^2}{2}\right]$	$\frac{L}{\sqrt{2}} e^{-\lvert u \rvert L\sqrt{2}}$
Conditional (granular) density, $p_{gr}(u)$	$\frac{p(u)}{1-2Q(L)}; \lvert u \rvert \leq 1$ $0 \quad ; \lvert u \rvert > 1$	$\frac{p(u)}{1-e^{-L\sqrt{2}}}; \lvert u \rvert \leq 1$ $0 \quad ; \lvert u \rvert > 1$
Interval probability, p_i	$\frac{Q(Lv_i) - Q(Lv_{i+1})}{1-2Q(L)}$	$\frac{e^{-v_iL\sqrt{2}} - e^{-v_{i+1}L\sqrt{2}}}{2(1-e^{-L\sqrt{2}})}; 2^{B-1} \leq i \leq 2^B - 1$ $p_{2^{B-1}-i} \quad ; 0 \leq i \leq 2^{B-1} - 1$
Interval mean noise, q_i	$p_iu_i - \frac{e^{-(Lv_i)^2/2} - e^{-(Lv_{i+1})^2/2}}{L\sqrt{2}\pi[1-2Q(L)]}$	$p_iu_i - \frac{e^{-v_iL\sqrt{2}}\left(v_i + \frac{1}{L\sqrt{2}}\right) - e^{-v_{i+1}L\sqrt{2}}\left(v_{i+1} + \frac{1}{L\sqrt{2}}\right)}{2(1-e^{-L\sqrt{2}})}; 2^{B-1} \leq i \leq 2^B - 1$ $-q_{2^{B-1}-i} \quad ; 0 \leq i \leq 2^{B-1} - 1$
Conditional mean square, σ_{gr}^2	$\frac{1-2Q(L) - L\sqrt{\frac{2}{\pi}} e^{-L^2/2}}{L^2[1-2Q(L)]}$	$\frac{1 - e^{-L\sqrt{2}} - e^{-L\sqrt{2}}(L^2 + L\sqrt{2})}{L^2(1-e^{-L\sqrt{2}})}$

5.4 The s/n of the embedded codec

Referring to (52) and Fig. 7, we have the mean-square difference between ξ' and ξ, over all possible error patterns

$$E\{(\xi' - \xi)^2\} = \xi_{\max}^2 E\{\epsilon_l^2\} = \xi_{\max}^2 \left[\sigma_q^2(B) + \sum_{l=1}^{2^B-1} P(l)\tilde{A}_l(B)\right], \quad (58)$$

where $P(l)$ is the probability of error pattern l. The effect of the transmission errors on the quantized version of ξ is $\xi_{\max}^2 \sum_{l=1}^{2^B-1} P(l)A_l(B)$. Returning to (19), we have two expectations: the first is the combined (quantizing and transmission) noise of a D-bit signal; the second expectation is the noise due to transmission errors in the M-bit minimal signal. Thus, we can write (19) as

$$\sigma_e^2 = \xi_{\max}^2 \left[\sigma_q^2(D) + \sum_{l=1}^{2^D-1} P(l)\tilde{A}_l(D) + b_P \sum_{l=1}^{2^M-1} P(l)A_l(M)\right], \quad (59)$$

where $\xi_{\max}^2 = L^2\sigma_\xi^2$ is related to σ_x^2 by

$$\xi_{\max}^2 = \frac{L^2\sigma_x^2}{G[1 - a_P L^2 \sigma_q^2(M)]}. \quad (60)$$

The s/n, which is the principal subject of this paper, is, therefore,

$$s/n = \frac{G[1 - a_P L^2 \sigma_q^2(M)]}{L^2[\sigma_q^2(D) + \sum_{l=1}^{2^D-1} P(l)\tilde{A}_l(D) + b_P \sum_{l=1}^{2^M-1} P(l)A_l(M)]}. \quad (61)$$

With the exception of the two summations in the denominator, all of the quantities in (61) are properties of the input signal and the codec design parameters. These summations,

$$\sigma_{qt}^2 = \sum_{l=1}^{2^D-1} P(l)\tilde{A}_l(D) + b_P \sum_{l=1}^{2^M-1} P(l)A_l(M) \quad (62)$$

comprise the effects of transmission errors on the performance of embedded DPCM. We analyze them in Section VI.

VI. TRANSMISSION EFFECTS, BINARY-ERROR PROBABILITIES

The 2^D probabilities, $P(l)$, of binary-error patterns are properties of the digital transmission system, which includes a modulator, a channel, a demodulator, possibly a codec for forward error correction, and possibly a means for combining different versions (diversity branches) of the received signal. Depending on these components, the $P(l)$ exhibit properties that facilitate evaluation of the sums in (61). In the following subsections we consider three paradigms: (1) random errors with statistically independent transmission of all bits; (2) slow fading

with the bit-error probability constant over each code word, but independent from word to word; and (3) channel coding that makes all error patterns equally likely.

6.1 Random binary errors

Errors in all bits are statistically independent of each other and occur with probability, P. The probability of error pattern l depends only on w, the Hamming weight of l, i.e., the number of ones in the B-bit binary representation of l. Thus,

$$P(l) = P^w(1 - P)^{B-w} = P^w \sum_{j=0}^{B-w} (-1)^j \binom{B-w}{j} P^j. \tag{63}$$

This expansion leads us to express the summations in (61) as polynomials in P. The coefficients of the polynomial involve the sums of all A factors with a fixed weight, w. Let us denote these sums $S_w(B)$ where, for example,

$$\begin{aligned} S_1(B) &= A_1(B) + A_2(B) + A_4(B) + \cdots + A_{2^{B-1}}(B); \\ S_2(B) &= A_3(B) + \cdots + A_{2^{B-1}+2^{B-2}}(B), \end{aligned} \tag{64}$$

and in general,

$$S_w(B) = \sum_{l_w} A_l(B); \qquad \tilde{S}_w(B) = \sum_{l_w} \tilde{A}_l(B), \tag{65}$$

where l_w is the set of all error patterns with Hamming weight, w. Combining (63) and (65), we can write

$$\sum_{l=1}^{2^B-1} P(l)A_l(B) = \sum_{w=1}^{B} \sum_{j=0}^{B-w} P^{j+w}(-1)^j \binom{B-w}{j} S_w(B). \tag{66}$$

The summations in (66) can be manipulated to form

$$\begin{aligned} \sum_{l=1}^{2^B-1} P(l)A_l(B) &= \sum_{w=1}^{B} P^w \sum_{j=1}^{w} \binom{B-j}{w-j} (-1)^{w-j} S_j(B) \\ &= \sum_{w=1}^{B} P^w T_w(B), \end{aligned} \tag{67}$$

where we define

$$\begin{aligned} T_w(B) &= \sum_{j=1}^{w} \binom{B-j}{w-j} (-1)^{w-j} S_j(B); \\ \tilde{T}_w(B) &= \sum_{j=1}^{w} \binom{B-j}{w-j} (-1)^{w-j} \tilde{S}_j(B). \end{aligned} \tag{68}$$

For the natural-binary and sign-magnitude representations we have

discovered and proved that for any input probability distribution, $T_w(B) = \tilde{T}_w(B) = 0$ for $w \geq 3$. Thus the tranmission term in (61) is

$$\sigma_{qt}^2 = \sum_{l=1}^{2^{D}-1} P(l)\tilde{A}_l(D) + b_P \sum_{l=1}^{2^{M}-1} P(l)A_l(M)$$
$$= \sum_{w=1}^{2} P^w[\tilde{T}_w(D) + b_P T_w(M)] \tag{69}$$

This formula is valid for channels with random binary errors, and Table IV presents values of $T_1(B)$, $T_2(B)$, $\tilde{T}_1(B)$, and $\tilde{T}_2(B)$ for three quantizer load factors, B = 2–6 bits and Gaussian and exponential inputs.

Table IV—Error constants for uncoded transmission

B	$T_1(B)$	$T_2(B)$	$\tilde{T}_1(B)$	$\tilde{T}_2(B)$	$T_1(B)$	$T_2(B)$	$\tilde{T}_1(B)$	$\tilde{T}_2(B)$
		Sign Magnitude				Natural Binary		
				Gaussian Inputs, Load Factor 1.78				
2	1.146	−0.146	1.085	−0.137	1.250	−0.354	1.194	−0.354
3	1.186	−0.186	1.170	−0.183	1.313	−0.439	1.298	−0.439
4	1.196	−0.196	1.192	−0.195	1.328	−0.461	1.325	−0.461
5	1.198	−0.198	1.197	−0.198	1.332	−0.466	1.331	−0.466
6	1.199	−0.199	1.198	−0.199	1.333	−0.467	1.333	−0.467
				Gaussian Inputs, Load Factor 3.16				
2	0.725	0.275	0.670	0.219	1.250	−0.775	1.167	−0.775
3	0.726	0.274	0.711	0.262	1.313	−0.899	1.292	−0.899
4	0.726	0.274	0.723	0.271	1.328	−0.930	1.323	−0.930
5	0.726	0.274	0.726	0.273	1.332	−0.938	1.331	−0.938
6	0.727	0.273	0.726	0.273	1.333	−0.940	1.333	−0.940
				Gaussian Inputs, Load Factor 5.62				
2	0.510	0.490	0.506	0.273	1.250	−0.990	1.137	−0.990
3	0.460	0.540	0.467	0.485	1.313	−1.165	1.292	−1.165
4	0.460	0.540	0.461	0.527	1.328	−1.196	1.323	−1.196
5	0.460	0.540	0.460	0.537	1.332	−1.204	1.331	−1.204
6	0.460	0.540	0.460	0.539	1.333	−1.206	1.333	−1.206
				Exponential Inputs, Load Factor 1.78				
2	0.943	0.057	0.898	0.000	1.250	−0.557	1.176	−0.557
3	0.958	0.042	0.950	0.024	1.313	−0.667	1.296	−0.667
4	0.964	0.036	0.962	0.031	1.328	−0.692	1.324	−0.692
5	0.966	0.034	0.965	0.033	1.332	−0.698	1.331	−0.698
6	0.966	0.034	0.966	0.033	1.333	−0.700	1.333	−0.700
				Exponential Inputs, Load Factor 3.16				
2	0.693	0.307	0.660	0.168	1.250	−0.807	1.147	−0.807
3	0.667	0.333	0.670	0.285	1.313	−0.958	1.291	−0.958
4	0.666	0.334	0.668	0.321	1.328	−0.990	1.323	−0.990
5	0.666	0.334	0.667	0.330	1.332	−0.998	1.331	−0.998
6	0.666	0.334	0.667	0.333	1.333	−1.000	1.333	−1.000
				Exponential Inputs, Load Factor 5.62				
2	0.537	0.463	0.527	0.205	1.250	−0.963	1.111	−0.963
3	0.465	0.535	0.492	0.430	1.313	−1.160	1.287	−1.160
4	0.459	0.541	0.468	0.512	1.328	−1.198	1.323	−1.198
5	0.458	0.542	0.461	0.534	1.332	−1.206	1.331	−1.206
6	0.458	0.542	0.459	0.540	1.333	−1.208	1.333	−1.208

6.2 Slow fading

Now the binary-error probability is a random variable that is constant over each code word but varies from word to word. In this case the effects of digital errors can be calculated as in (69) but with the average values $\overline{P^w}$ replacing P^w. These averages are computed over the distributions of channel s/n's that govern the random fluctuation of P from one code word to the next.

6.3 Error-correcting codes

To analyze the performance of embedded DPCM protected by an error-correcting channel code, we make three simplifying approximations. The first one, which pertains to the error-correcting code, states that when there is a decoding error, all error patterns are equally likely. Thus we assume that if the C most significant DPCM bits are protected by the code,

$$P(l) = \frac{1}{2^C - 1} P_W = \frac{1}{2^{C-1}} P_e, \qquad l = 1, 2, \cdots, 2^C - 1, \tag{70}$$

where P_W is the word-error probability and P_e is the binary-error probability of the channel code. They are related by

$$P_W = \frac{2^C - 1}{2^{C-1}} P_e. \tag{71}$$

The other two approximations apply when $C < D$, so that the C most significant DPCM bits are protected and the other D–C bits are uncoded. To simplify computations for this case, we (1) ignore simultaneous errors in the protected and unprotected parts of the D-bit word and (2) ignore multiple errors in the unprotected part. We consider separately three different relationships among C, the number of coded bits; D, the length of the entire DPCM code word; and M, the number of bits in the minimal quantizer.

6.3.1 Entire code word protected ($M \leq D = C$)

In this case $P(l)$ may be calculated according to (70) for all D-bit error patterns, $l = 1, 2, \cdots, 2^{D-1}$. This value of $P(l)$ is constant throughout the first sum in (62). In the second sum we have the probability of M-bit error patterns. For each M-bit error pattern there are 2^{D-M} D-bit patterns. Hence $P(l)$ in the second sum of (62) is higher by the factor 2^{D-M} than $P(l)$ in the first sum. Because each sum in (61) is a constant probability times a sum of A factors, we write

$$\sigma_{qt}^2 = \frac{1}{2^{C-1}} P_e[\tilde{A}_{sum}^{(D)}(D) + b_P 2^{D-M} A_{sum}^{(M)}(M)], \tag{72}$$

where we define the sum of the first $2^C - 1$ A factors

$$\tilde{A}_{\text{sum}}^{(C)}(D) = \sum_{l=1}^{2^C-1} \tilde{A}_l(D); \qquad A_{\text{sum}}^{(C)}(D) = \sum_{l=1}^{2^C-1} A_l(D). \tag{73}$$

Table V contains numerical values of $\tilde{A}_{\text{sum}}$ and A_{sum} for the sets of conditions of interest to us here.

6.3.2 *Entire minimal code word, parts of the supplemental code word protected ($M \leq C < D$)*

In this event we assume that all of the unprotected bits have the binary-error probability P and that as before the protected bits have binary-error probability P_e. Furthermore, we set to 0 the probability of simultaneous errors in the protected and unprotected parts of the code word. (These errors occur with probability related to P_eP.) We also set to 0 the probability of multiple errors in the unprotected part of the code word (which occur with probability less than P^2). Thus we break the first sum in (61) into two parts. The first part accounts for

Table V—Error constraints $A_{\text{sum}}^{(C)}(B)$ and $\tilde{A}_{\text{sum}}^{(C)}(B)$ *for coded transmission*

Load Factor		$B=2$		$B=3$			$B=4$			
		$C=1$	$C=2$	$C=1$	$C=2$	$C=3$	$C=1$	$C=2$	$C=3$	$C=4$
					Gaussian Inputs, Sign Magnitude					
1.78		0.90	2.15	0.87	2.11	4.37	0.87	2.10	4.35	8.78
3.16	A	0.47	1.72	0.41	1.56	3.45	0.40	1.52	3.37	6.91
5.62		0.26	1.51	0.15	1.11	2.92	0.13	1.05	2.77	5.84
1.78		0.84	2.03	0.86	2.08	4.31	0.86	2.09	4.34	8.75
3.16	$\tilde{A}$	0.39	1.56	0.39	1.52	3.37	0.39	1.51	3.35	6.86
5.62		0.15	1.28	0.13	1.07	2.84	0.13	1.04	2.75	5.80
					Gaussian Inputs, Natural Binary					
1.78		1.00	2.15	1.00	2.15	4.37	1.00	2.15	4.37	8.78
3.16	A	1.00	1.72	1.00	1.72	3.45	1.00	1.72	3.45	6.91
5.62		1.00	1.51	1.00	1.51	2.92	1.00	1.51	2.92	5.84
1.78		0.95	2.03	0.99	2.12	4.31	1.00	2.14	4.36	8.75
3.16	$\tilde{A}$	0.89	1.56	0.97	1.68	3.37	0.99	1.71	3.43	6.86
5.62		0.78	1.28	0.95	1.46	2.84	0.99	1.50	2.90	5.80
					Exponential Inputs, Sign Magnitude					
1.78		0.69	1.94	0.65	1.81	3.92	0.64	1.78	3.85	7.86
3.16	A	0.44	1.69	0.35	1.41	3.33	0.34	1.35	3.20	6.66
5.62		0.29	1.54	0.15	1.09	2.93	0.13	0.99	2.69	5.83
1.78		0.62	1.80	0.63	1.77	3.85	0.63	1.77	3.83	7.82
3.16	$\tilde{A}$	0.34	1.49	0.33	1.37	3.25	0.33	1.34	3.18	6.62
5.62		0.15	1.26	0.13	1.03	2.83	0.12	0.97	2.67	5.79
					Exponential Inputs, Natural Binary					
1.78		1.00	1.94	1.00	1.94	3.92	1.00	1.94	3.92	7.86
3.16	A	1.00	1.69	1.00	1.69	3.33	1.00	1.69	3.33	6.66
5.62		1.00	1.54	1.00	1.54	2.93	1.00	1.54	2.93	5.83
1.78		0.90	1.80	0.97	1.91	3.85	0.99	1.93	3.90	7.82
3.16	$\tilde{A}$	0.83	1.49	0.95	1.64	3.25	0.99	1.68	3.31	6.62
5.62		0.73	1.26	0.92	1.46	2.83	0.98	1.52	2.90	5.79

errors in the first C (protected) bits when the other D–C bits are error free, $l = 1, 2, \cdots, 2^C - 1$. The second part accounts for single errors in the remaining D–C bits when the first C bits are error free, $l = 2^C, 2^{C+1}, \cdots, 2^{D-1}$. The result is

$$\sigma_{qt}^2 = \frac{P_e}{2^{C-1}} \tilde{A}_{sum}^{(C)}(D) + P \sum_{i=C}^{D-1} \tilde{A}_{2^i}(D) + b_P 2^{C-M} \frac{P_e}{2^{C-1}} A_{sum}^{(M)}(M). \quad (74)$$

In the second term we use the approximation

$$P \approx P(1 - P)^{D-C+1}(1 - P_W) \quad (75)$$

for the probability of a single error in the unprotected part of the code word. A further approximation $\tilde{A}_{2^i}(D) \approx A_{2^i}(D)$ simplifies computation of the second term of (74) because, for natural-binary and sign-magnitude representations, (56) applies for $b > 1$. This allows us to derive

$$\sum_{i=C}^{D-1} A_{2^i}(D) = 4(4^{-C} - 4^{-D})/3 = A_{un}^{(C)}(D). \quad (76)$$

Thus for $M \leq C < D$ we have the formula

$$\sigma_{qt}^2 = \frac{1}{2^{C-1}} P_e[\tilde{A}_{sum}^{(C)}(D) + b_P 2^{C-M} A_{sum}^{(M)}(M)] + P A_{un}^{(C)}(D). \quad (77)$$

6.3.3 Part of the minimal code word protected ($C < M \leq D$)

Just as we decomposed the first sum in (61) into two parts in the previous case, we similarly decompose the second sum when some of the M minimal bits are unprotected. The result is

$$\sigma_{qt}^2 = \frac{P_e}{2^{C-1}} [\tilde{A}_{sum}^{(C)}(D) + b_P A_{sum}^{(C)}(M)] + P[A_{un}^{(C)}(D) + b_P A_{un}^{(C)}(M)]. \quad (78)$$

VII. NUMERICAL RESULTS

The useful computational formulas (61), (62), (69), (72), (77), and (78) are summarized in Table VI. In this section we apply these formulas to illustrate some of the properties of embedded DPCM and its relationship to conventional DPCM.

7.1 Source characteristics

All of our numerical results pertain to a Gauss-Markov input signal with adjacent-sample correlation $r_1 = 0.85$. The codec uses single integration with coefficient $a_1 = 0.85$ and the load factor, $L = \sqrt{10}$. For this configuration the coding gain is $G = 3.6$. If the embedded codec has a minimal quantizer with $M = 2$ bits, C_{source} in Table VI is 0.31. For conventional DPCM $C_{source} = 0.35$ with 3 bits/sample and

Table VI—Signal-to-noise ratio of embedded DPCM

$$\text{s/n} = \frac{C_{\text{source}}}{\sigma_q^2(D) + \sigma_{qt}^2};\ C_{\text{source}} = \frac{G[1 - a_P L^2 \sigma_q^2(M)]}{L^2}$$

(a) Notation

Symbol	Description	General Formula	Formula for Single Integration
G	Coding gain	(26)	$(1 - 2a_1 r_1 + a_1^2)^{-1}$
a_P	Predictor gain	(30)	a_1^2
L	Load factor	(33)	
M	Minimal codec bits		
D	Transmitted bits		
b_P	$1 + b_P$ is integrator gain	(20), (21)	$a_1^2/(1 - a_1^2)$
r_1	Adjacent-sample autocorrelation	(27)	
$\sigma_q^2(B)$	Quantizing noise	(49)	$2^{-2B}/3$ or Table II
σ_{qt}^2	Transmission-error effects	(62)	

(b) Error Formulas*

Transmission Format	Error Effect σ_{qt}^2
No channel code (Table IV)	$\sum_{w=1}^{2} P^w[\tilde{T}_w(D) + b_P T_w(M)]$
C bits coded (Table V)	
$M \leq D = C$	$\frac{P_e}{2^{C-1}}[\tilde{A}_{\text{sum}}^{(D)}(D) + b_P 2^{D-M} A_{\text{sum}}^{(M)}(M)]$
$M \leq C < D$	$\frac{P_e}{2^{C-1}}(\tilde{A}_{\text{sum}}^{(C)}(D) + b_P 2^{C-M} A_{\text{sum}}^{(M)}(M)] + P A_{\text{un}}^{(C)}(D)$
$C < M \leq D$	$\frac{P_e}{2^{C-1}}[\tilde{A}_{\text{sum}}^{(C)}(D) + b_P A_{\text{sum}}^{(C)}(M)] + P[A_{\text{un}}^{(C)}(D) + b_P A_{\text{un}}^{(C)}(M)]$

* P: binary-error rate, uncoded bits; P_e: binary-error rate, coded bits; $A_{\text{un}}^{(C)}(D) = 4(4^{-C} - 4^{-D})/3$.

0.36 with 4 bits/sample. Thus the quantizing-noise penalty of the embedded codec is 0.54 dB when 3 bits are transmitted, and 0.68 dB when 4 bits are transmitted. As indicated in Ref. 8 these penalties increase for higher values of L and a_1. They decrease rapidly as M increases.

7.2 Modulation, channel, error-correcting codes

In our numerical examples the modulation is coherent phase shift keying (CPSK) so that in a white-Gaussian-noise channel the binary error probability is

$$P = Q(\sqrt{2\rho}), \tag{79}$$

where ρ is the channel s/n and

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt. \tag{80}$$

Figures 3 and 4 depict performance with three different convolutional

codes. For all of them, we use the following truncated union bound to calculate binary-error probability:

$$P_e = \frac{1}{m} \sum_{i=d}^{d+4} w_i Q(\sqrt{2i\rho}). \tag{81}$$

where $m = 1, 2, 3$ for the rate 1/2, 2/3, and 3/4 codes, respectively, and the coefficients w_i and the free distance, d, characterize the convolutional coder and decoder. The codes considered here are punctured codes[16] with constraint length 5 (16 states in the decoder memory). Table VII contains their coefficients and free distances. The combination of (79) and (81) with the formulas in Table VI produces the curves in Figs. 3, 4, 5, and 8.

For transmission environments other than CPSK in a white-Gaussian-noise channel, there are formulas for P and P_e to be used in place of (79) and (81). There are many families of modulation schemes, channel conditions, error-correcting codes, and reception techniques that are of practical interest. This paper provides the tools for studying their effects on the performance of embedded and conventional DPCM. This is a subject worthy of further investigation.

7.3 Binary number representation

Without forward-error correction, the noise due to transmission errors is dominated by the effects of single errors in the most significant part of the transmitted code word. With the natural-binary representation, an error in the most significant bit always causes a noise impulse of half the peak-to-peak range of the quantizer (56). With the sign-magnitude representation, an error in the sign bit inverts the polarity of the quantized signal, thereby producing a noise impulse of approximately twice the magnitude of the quantizer input

Table VII—Source and channel code formats, convolutional code properties

	Format 1	Format 2	Format 3	Format 4
Source code				
bits/sample	4	3	3	2
bits/second	32K	24K	24K	16K
bits/sample protected	0	3	2	2
Channel code				
rate	No code	3/4	2/3	1/2
Free distance, d		4	5	7
Weight w_d		22	25	4
w_{d+1}	Error	0	112	12
w_{d+2}	Properties	1687	357	20
w_{d+3}		0	1858	72
w_{d+4}		66964	8406	225

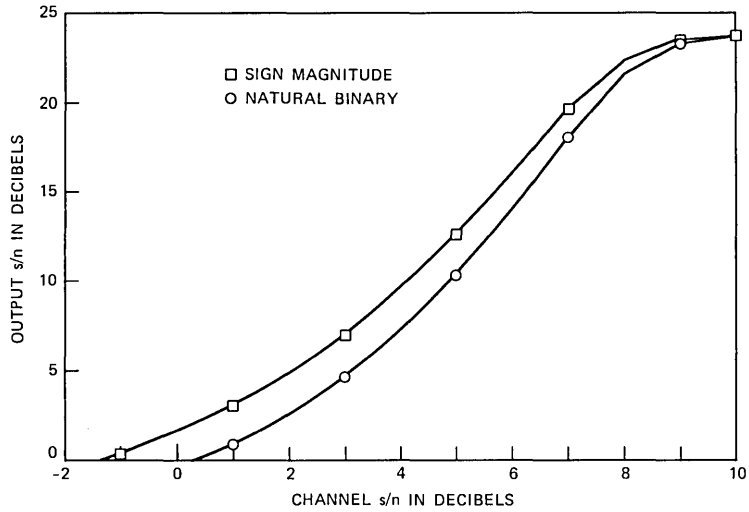


Fig. 8—Performance of embedded DPCM with sign-magnitude and natural-binary representations of quantizer outputs.

(57). Consequently, quantizers employing the sign-magnitude representation are somewhat less affected by transmission errors than quantizers with the natural-binary representation when the input probability distribution has its mode at zero. This is illustrated in Fig. 8, which pertains to uncoded 32 kb/s embedded DPCM transmission. When transmission errors are the dominant distortion, signals represented in the natural-binary format are about 2 dB noisier than signals represented by the sign-magnitude format.

With forward error correction, all error patterns are equally likely, and the two representations have essentially the same s/n.

VIII. ACKNOWLEDGMENT

We thank Lawrence Wong for his helpful comments on this manuscript.

REFERENCES

1. F. S. Boxall, "A Digital Carrier-Concentrator System with Elastic Traffic Capacity," IEEE Trans. Commun., *COM-22,* No. 10 (October 1974), pp. 1636–42.
2. J. A. Sciulli and S. J. Campanella, "A Speech Predictive Encoding Communication System for Multichannel Telephony," IEEE Trans. Commun., *COM-21* (July 1973), pp. 827–35.
3. T. Bially, B. Gold, and S. Seneff, "A Technique for Adaptive Voice Flow Control in Integrated Packet Networks," IEEE Trans. Commun., *COM-28,* No. 3 (March 1980), pp. 325–33.
4. D. J. Goodman and C.-E. Sundberg, "Combined Source and Channel Coding for Variable-Bit-Rate Speech Transmission," B.S.T.J., *62,* No. 7 (September 1983).
5. Proceedings of the 1982 IEEE International Conference on Acoustics Speech and

Signal Processing, Paris, May 1982. Session S8, pp. 954–83, contains four papers on adaptive DPCM at 32 kb/s.
6. J.-M. Raulin, et al., "A 60 Channel PCM-ADPCM Converter," IEEE Trans. Commun., *COM-30,* No. 4 (April 1982), pp. 567–73.
7. D. W. Petr, "32 Kbps ADPCM-DLQ Coding for Network Applications," Record of the IEEE Global Telecommunications Conference, Miami, Florida, November 1982, pp. 239–43.
8. D. J. Goodman, "Embedded DPCM for Variable Bit Rate Transmission," IEEE Trans. Commun., *COM-28,* No. 7 (July 1980), pp. 1040–6.
9. N. S. Jayant, "Variable Rate ADPCM Based on Explicit Noise Coding," B.S.T.J., *62,* No. 3 (March 1983), pp. 657–77.
10. N. Rydbeck and C.-E. Sundberg, "Analysis of Digital Errors in Non-Linear PCM Systems," IEEE Trans. Commun., *COM-24,* No. 1 (January 1976), pp. 59–65.
11. C.-E. Sundberg and N. Rydbeck, "Pulse Code Modulation with Error-Correcting Codes for TDMA Satellite Communication Systems," Ericsson Technics, *32,* No. 1 (1976), pp. 1–56.
12. N. Rydbeck and C.-E. Sundberg, "PCM/TDMA Satellite Communication Systems with Error-Correcting and Error-Detecting Codes," Ericsson Technics, *32,* No. 3 (1976), pp. 195–247.
13. D. J. Goodman, B. J. McDermott, and L. H. Nakatani, "Subjective Evaluation of PCM Coded Speech," B.S.T.J., *55,* No. 8 (October 1976), pp. 1087–109.
14. B. McDermott, C. Scagliola, and D. J. Goodman, "Perceptual and Objective Evaluation of Speech Processed by Adaptive Differential PCM," B.S.T.J., *57,* No. 5 (May–June 1978), pp. 1597–618.
15. L. J. Greenstein, "Slope Overload Noise in Linear Delta Modulators with Gaussian Inputs," B.S.T.J., *52,* No. 3 (March 1973), pp. 387–421.
16. G. C. Clark and J. B. Cain, *Error-Correction Coding for Digital Communications,* New York: Plenum Press, 1981, pp. 237–8, and p. 403.

AUTHORS

David J. Goodman, B.E.E., 1960, Rensselaer Polytechnic Institute; M.E.E., 1962, New York University; Ph.D. (Electrical Engineering), 1967, Imperial College, London; Bell Laboratories, 1967—. Mr. Goodman has studied various aspects of digital communications including analog-to-digital conversion, digital signal processing, subjective assessment of voiceband codecs, and spread spectrum modulation for mobile radio. He is Head, Communications Methods Research Department. In 1974 and 1975 he was a Senior Research Fellow, and in 1983 a Visiting Professor at Imperial College, London, England. Member, IEEE.

Carl-Erik W. Sundberg, M.S.E.E., 1966, and Dr. Techn., 1975, Lund Institute of Technology, University of Lund, Sweden; Bell Laboratories, 1981–1982. Mr. Sundberg is an Associate Professor in the Department of Telecommunication Theory, University of Lund, and a consultant in his field. He is Director of the consulting company SUNCOM, Lund. During 1976 he was with the European Space Research and Technology Centre (ESTEC), Noordwijk, The Netherlands, as an ESA Research Fellow. He has been a Consulting Scientist at LM Ericsson and SAAB-SCANIA, Sweden, and at Bell Laboratories. His research interests include source coding, channel coding (especially decoding techniques), digital modulation methods, fault-tolerant systems, digital mobile radio systems, spread spectrum systems, and digital satellite communication systems. He has published a large number of papers in these areas during the last few years. Senior Member, IEEE; member, SER, Sveriges Elektroingenjörers Riksförening.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 9, November 1983
Printed in U.S.A.

CCITT Compatible Coding of Multilevel Pictures

By H. GHARAVI* and A. N. NETRAVALI*

(Manuscript received February 25, 1983)

The Comite Consultatif International Telegraphique et Telephonique (CCITT) has recently recommended a code for two-level (black and white) graphics transmission. A large number of pictures in graphics communication contain areas that cannot be represented adequately by only two shades of gray. We describe techniques by which a composite picture, containing an arbitrary mixture of two- and multilevel areas, can be coded by schemes that are compatible with the CCITT code. First, the composite picture is segmented automatically into two types of areas: one requiring only two levels (text, drawings, etc.) and the other requiring multilevels (for example, photos). A Differential Pulse Code Modulation (DPCM) scheme is then used to code the multilevel areas. Code assignment for the outputs of the DPCM quantizer are based on the local conditional statistics, and the bit stream is processed to change the statistics of the run lengths so that the CCITT run-length code becomes efficient. Results of computer simulations are presented in terms of quality of processed pictures and the required bit rate. Simulations show that our CCITT compatible scheme is as efficient as an incompatible but optimum DPCM coding scheme.

I. INTRODUCTION

Simultaneous developments (algorithmic as well as systems) have taken place for many years in coding and transmission of two-level (black and white) document facsimile, and multilevel (many shades of gray) pictures.[1,2] The former type of pictures require very high spatial resolution to preserve the sharpness and have been coded by one-

* Bell Laboratories.

dimensional run-length coding and two-dimensional edge difference coding [Comite Consultatif International Telegraphique et Telephonique (CCITT) one- and two-dimensional codes].[3] On the other hand, multilevel pictures contain gradual luminance transitions, and therefore require lower spatial resolution. They have been coded by Differential Pulse Code Modulation (DPCM) and transform methods. Most pictures used in business facsimile systems and audiographics conferencing contain a mixture of two-level and multilevel segments or subpictures. Coding such pictures using two-level techniques would not be adequate from the point of view of the picture quality, and using multilevel techniques would generate enormous data rates. Thus, it is of interest to devise schemes that automatically divide a picture into segments, each segment with a specified amplitude (gray shades) and spatial resolution and code each segment as best suited for it. Another practical requirement is that of compatibility. A coding scheme that handles a mixture of two-level and multilevel segments should be upwardly compatible with the CCITT standard schemes for two-level pictures. System cost will be reduced if the scheme for two-level multilevel pictures uses hardware blocks that are also used by the two-level picture coder. We present such a scheme below. Principal characteristics of our scheme are:

1. Compatibility with the CCITT schemes for two-level pictures
2. Automatic segmentation of pictures into two-level and multilevel segments
3. High coding efficiency by preprocessing the multilevel segments to fit the CCITT codes
4. Lower spatial resolution for gray-level segments (if desired)
5. Nonlossy (information preserving) coding of two-level segments, and lossy coding of multilevel segments.

II. CODING ALGORITHM

The coding algorithm is explained in the following steps.

2.1 Segmentation

The function of the segmentor is to classify each picture element as one of the three:

1. Black } → Two-Level
2. White }

3. Gray → Multilevel

Figure 1 shows a neighborhood of the current picture element (pel) used for segmentation. We assume that each pel, obtained from the scanner, is specified by many shades of gray (e.g., 8 bits). The size of the neighborhood can be arbitrary. If it is too small, then many

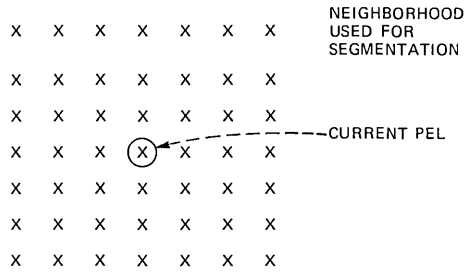


Fig. 1—Picture elements used for segmentation of the current pel. The size of the neighborhood is not necessarily 5 × 5 as suggested in the figure.

discontinuous segments of gray pixels will be generated. On the other hand, if the neighborhood is too large, then the ability to resolve small gray areas is lost. We slide this window of neighborhood over the pels along a scan line and classify each pel. We consider the boundary pels of the picture separately. Two thresholds, t_1 and t_2 ($t_1 < t_2$), are selected. It is hypothesized that most black pels will have intensity less than t_1, white pels will have intensity greater than t_2, and gray pel intensities may lie anywhere. Within the neighborhood let

n_1 = number of pels with intensity value $< t_1$

n_2 = number of pels with intensity value $> t_2$

n_3 = all the rest of the pels.

We define a state, S, consisting of three components: S_1, S_2, and S_3. A picture is segmented on the basis of the value of S. Let

$$S = S_1 + S_2 + S_3, \tag{1}$$

where

$S_1 = 1$, if $n_3 > n_1 + n_2$

$= 0$, otherwise

$S_2 = 1$, if previous pel is gray

$= 0$, otherwise

and

$S_3 = 1$, if $t_1 <$ intensity of present pel $< t_2$

$= 0$, otherwise.

The segmentation rule is then given by

$S \geq 2 = >$ current pel gray

$S < 2$ and intensity of current pel $> T = >$ current pel white

$S < 2$ and intensity of current pel $\leq T = >$ current pel black.

T is a threshold used to distinguish black elements from white ones, once they are known to be of the two-level type. If the range $(t_2 - t_1)$ is decreased by increasing t_1 and decreasing t_2, then more elements will be regarded as two-level and the quality of picture may suffer, but this will also decrease the bit rate.

We evaluated the performance of the segmentor, in particular its dependence on the block size and $(t_2 - t_1)$ by computer simulation. Since there are no standard mixtures of two-level and multilevel images, we created our own by taking a 512 × 512 gray-level image (shown in Fig. 2) and superimposing it on the CCITT documents four and five. Since this 512 × 512 original was scanned at low resolution (compared to 200 pels/inch used for CCITT documents), it contains significant sharp transitions that would not be present in a photograph scanned at 200 pels/inch. Also, because the original gray-level picture and the CCITT documents are rather "clean", segmentor works quite

Fig. 2—A 512 × 512 multilevel (8 bits/pel) picture used for simulation.

well. However, this may not be a typical situation if a nonideal scanner was used. We, therefore, added random noise to the entire composite picture. This noise had a variance of 425 (on an 8-bit scale, 0-255). Table I shows the performance of the segmentor with respect to block size for a composite picture made from CCITT document 4. Here $t_1 = 28$, $t_2 = 195$. As we view such a segmented picture we realize that a 5×5 block may be too small. A 9×9 block appears quite adequate even when the added noise variance reaches 758. Higher block sizes result in a larger number of contiguous gray pels, thereby decreasing the number of segments. Figure 3 shows a segmented picture. Due to equipment limitations we show only a 512×512 section of the

Table I—Performance of the segmentor

Block Size	No. of Gray Pels*		No. of Segments	
	Without Noise	With Noise†	Without Noise	With Noise
5×5	221,045	257,952	5394	27,910
9×9	222,818	226,691	3918	6662
15×15	224,595	224,598	3052	3062

* Total number of gray pels is 512×512.
† Variance of the noise = 758 (8-bit, 0-255 scale).

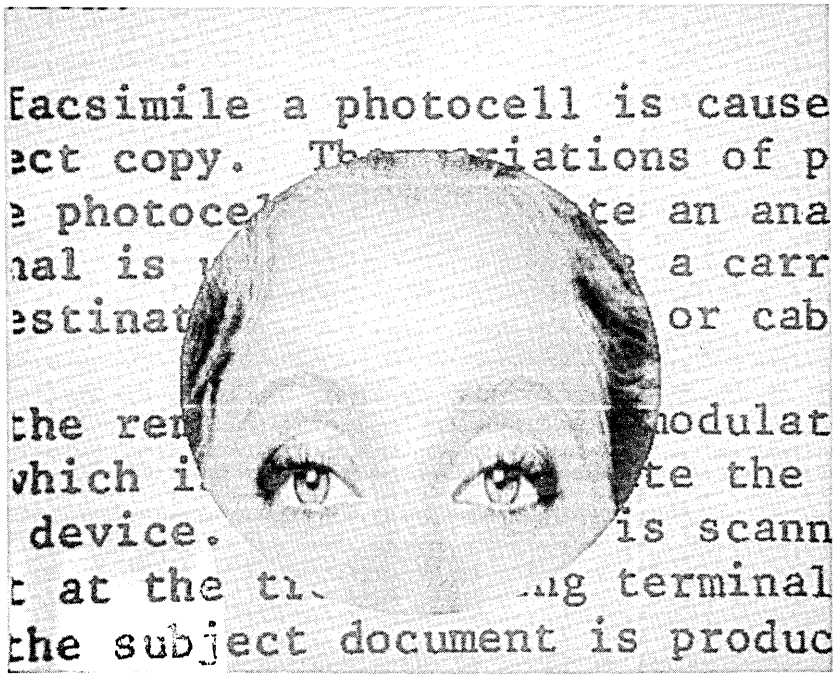


Fig. 3—A segmented picture. Pels classified black and white are reproduced with intensities 30 and 215, respectively. Gray-level pels are reproduced with 8-bit intensities.

composite picture. The segmentor has adequately separated the two-level areas from the multilevel areas.

2.2 Subsampling and interpolation

In most cases, areas of the picture that are segmented to be gray do not need as much spatial resolution as the two-level segments. If the two-level picture is at a very high spatial resolution (e.g., 200 pels/inch), then without any significant loss of quality, spatial resolution can be reduced in gray areas. Following is a scheme for subsampling and interpolation. A subsampling pattern is shown in Fig. 4. Interpolation is performed by averaging four surrounding pels, as in Fig. 4. Although we show only 2:1 subsampling, higher subsampling ratios may be used if the quality requirements are not very high. Also, two-dimensional subsampling may be performed, but this may increase the complexity.

2.3 Coding

After the pels (black, white, or gray) are classified and the gray areas to be transmitted are determined, a DPCM coder is used for gray areas. The resulting bit stream from the DPCM coder is preprocessed, multiplexed with bits from the two-level segments, and then coded by a CCITT one-dimensional or two-dimensional coder. Addresses for the segments of gray pels are coded separately and multiplexed with the coded data to transmit on the channel. A block diagram for the transmitter portion is shown in Fig. 5. Details of the algorithm are given below. Only a nonsubsampled case is illustrated; a subsampled case follows trivially.

2.3.1 Grey segment coding

The purpose of gray segment coding is to convert an 8-bit/pel signal representing gray areas into a coded 3-bit/pel signal, which can then be preprocessed and run-length coded. This procedure reduces the bit rate for gray pels to about 2 bits/pel.

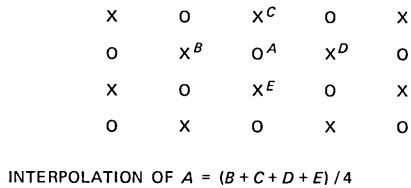


Fig. 4—Subsampling and interpolation pattern used in gray areas. Only one-dimensional subsampling is considered.

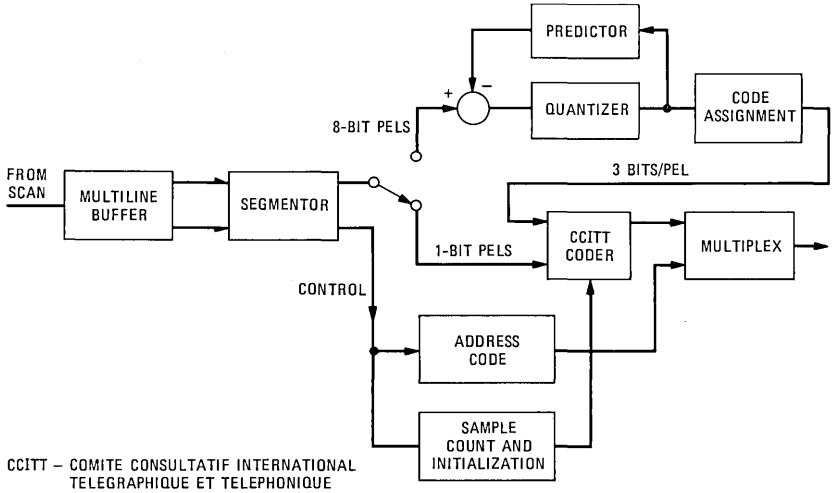


Fig. 5—Block diagram of the transmitter portion of the coder.

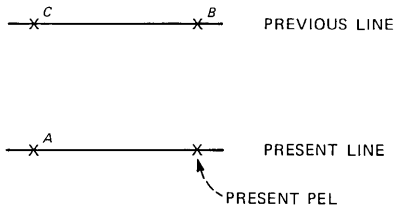


Fig. 6—Configuration of pels used for prediction. Only a nonsubsampled case is shown.

2.3.2 DPCM predictor

On Fig. 6 we see that the present pel is predicted by

$$\hat{X} = \text{prediction of the present pel}$$
$$= 0.5A + 0.25(B + C).$$

It is assumed in this figure that all elements A, B, C are gray elements. Appropriate modification is made if some of these are two-level elements.

2.3.3 DPCM quantizer

The prediction error is quantized by a symmetric seven-level quantizer with the transfer characteristics given in Fig. 7. For most pictures with a resolution of 100 pels/inch, this appears adequate, although in some cases dynamic range may not be sufficient. Subjective studies are needed to optimize the characteristics for a given set of pictures.

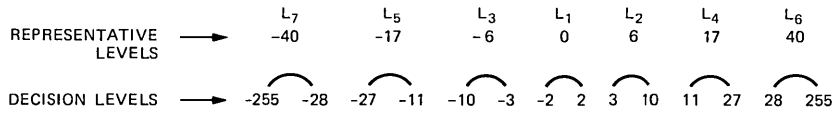


Fig. 7—Transfer characteristics of the quantizer.

Efficiency can be improved further by adapting the prediction and quantization.

2.3.4 Code assignment

To reduce statistical redundancy and create a bit stream that can be coded compatibly with the CCITT code, seven levels of the quantizer output are mapped into a three-bit code. First, a table of 49 states is constructed by looking at the seven outcomes of the quantized prediction values for both elements A and B (in Fig. 6). Given a state, the code words for the present pel are arranged in order of conditional frequency of occurrence. Such statistics are precomputed for a set of pictures. The code word that is most frequent (for a given state) is given the code [000], the next highest is given code [001], etc. In addition, to decrease the probability of occurrence of isolated '1', if the last bit of the code word for A is a '1', then the entire code word for the present pel is complemented (i.e., '0' $\rightarrow$ '1', and '1' $\rightarrow$ '0'). The table of 49 states and the corresponding code words are shown in Table II.

2.4 Preprocessing

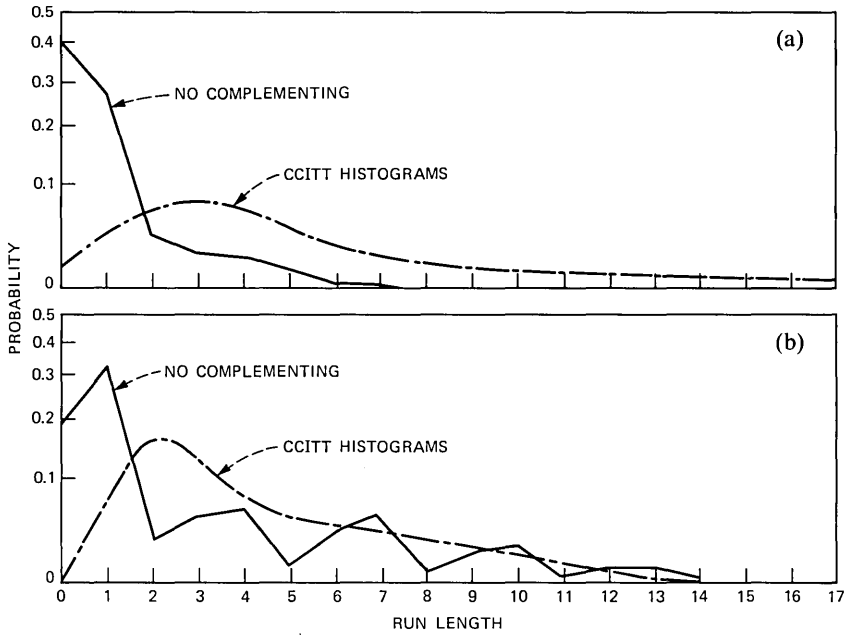
The code words for various states (e.g., runs) for the CCITT scheme are already defined based on the statistics. The statistics of the states for the gray-level segments are quite different. As an example, Fig. 8 shows histograms of the runs for black and white pels on which the one-dimensional CCITT code is based. The same figure also shows the histograms of the runs of the bits from gray-level picture (only 512 $\times$ 512) with the code assignment of the previous section but without any bit complementing. It is clear that the histograms are not similar in shape, and therefore using the CCITT code for runs of bits from gray segments would not be efficient. Since our experience shows that the two-dimensional CCITT code is not efficient for the gray segments, we give below a method of preprocessing that makes efficient use of the one-dimensional CCITT code. Let $n_c^b(i)$ and $n^b(i)$ be the histograms of the runs of black elements for the CCITT code and the gray-level segments, respectively. Also let $c(i)$ be the code assigned to the ith run by the CCITT coder. Let $j(i)$ be the sequence that is arranged in descending order of the histogram function $h_c^b(i)$, i.e.,

$$h_c^b(j(i)) \leq h_c^b[j(i-1)]\,.$$

Similarly, arrange the gray-level histogram $h^b(i)$ with the function $j^*(i)$. Then the code word for a length $j^*(i)$ of the gray segment is the same as $c(j(i))$. Thus, we arrange the two histograms in descending order and choose the code to be the same for entries of both the

Table II—Uncomplemented code words

No.	Quantizer Level for B	Quantizer Level for A	Decreasing Frequency of Occurrence –> 000	001	011	111	110	100	010
1	L_1	L_1	L_1	L_2	L_3	L_4	L_5	L_6	L_7
2	L_1	L_2	L_1	L_2	L_3	L_4	L_5	L_6	L_7
3	L_1	L_3	L_1	L_3	L_2	L_5	L_4	L_6	L_7
4	L_1	L_4	L_2	L_1	L_4	L_3	L_7	L_5	L_6
5	L_1	L_5	L_3	L_5	L_1	L_2	L_6	L_4	L_7
6	L_1	L_6	L_5	L_6	L_3	L_2	L_7	L_1	L_4
7	L_1	L_7	L_7	L_4	L_2	L_6	L_5	L_3	L_1
8	L_2	L_1	L_1	L_3	L_2	L_4	L_5	L_7	L_6
9	L_2	L_2	L_1	L_2	L_3	L_4	L_5	L_7	L_6
10	L_2	L_3	L_3	L_1	L_2	L_5	L_4	L_6	L_7
11	L_2	L_4	L_2	L_4	L_1	L_3	L_7	L_5	L_6
12	L_2	L_5	L_3	L_5	L_1	L_2	L_6	L_4	L_7
13	L_2	L_6	L_6	L_5	L_7	L_3	L_4	L_1	L_2
14	L_2	L_7	L_7	L_5	L_4	L_6	L_2	L_3	L_5
15	L_3	L_1	L_1	L_2	L_3	L_4	L_5	L_6	L_7
16	L_3	L_2	L_1	L_2	L_3	L_4	L_5	L_7	L_6
17	L_3	L_3	L_1	L_3	L_2	L_5	L_4	L_6	L_7
18	L_3	L_4	L_2	L_1	L_4	L_3	L_7	L_5	L_6
19	L_3	L_5	L_3	L_1	L_5	L_2	L_6	L_4	L_7
20	L_3	L_6	L_6	L_5	L_3	L_2	L_1	L_7	L_4
21	L_3	L_7	L_7	L_6	L_5	L_3	L_2	L_4	L_1
22	L_4	L_1	L_3	L_1	L_2	L_5	L_4	L_7	L_6
23	L_4	L_2	L_2	L_1	L_3	L_4	L_5	L_7	L_6
24	L_4	L_3	L_3	L_1	L_5	L_2	L_4	L_7	L_6
25	L_4	L_4	L_4	L_2	L_1	L_7	L_3	L_5	L_6
26	L_4	L_5	L_5	L_3	L_1	L_2	L_6	L_4	L_7
27	L_4	L_6	L_6	L_1	L_5	L_4	L_2	L_3	L_7
28	L_4	L_7	L_7	L_4	L_2	L_3	L_6	L_1	L_5
29	L_5	L_1	L_2	L_1	L_3	L_4	L_5	L_6	L_7
30	L_5	L_2	L_1	L_2	L_4	L_3	L_5	L_6	L_7
31	L_5	L_3	L_3	L_1	L_2	L_5	L_4	L_6	L_7
32	L_5	L_4	L_4	L_2	L_7	L_1	L_3	L_5	L_6
33	L_5	L_5	L_5	L_3	L_6	L_1	L_2	L_4	L_7
34	L_5	L_6	L_6	L_5	L_3	L_1	L_2	L_7	L_4
35	L_5	L_7	L_6	L_7	L_4	L_5	L_2	L_1	L_3
36	L_6	L_1	L_1	L_6	L_7	L_3	L_2	L_5	L_4
37	L_6	L_2	L_6	L_7	L_2	L_4	L_5	L_1	L_3
38	L_6	L_3	L_3	L_6	L_2	L_4	L_1	L_7	L_4
39	L_6	L_4	L_7	L_6	L_4	L_2	L_1	L_3	L_5
40	L_6	L_5	L_5	L_6	L_3	L_1	L_2	L_4	L_7
41	L_6	L_6	L_6	L_5	L_3	L_7	L_1	L_4	L_2
42	L_6	L_7	L_6	L_7	L_5	L_4	L_2	L_3	L_1
43	L_7	L_1	L_7	L_5	L_3	L_4	L_6	L_2	L_1
44	L_7	L_2	L_7	L_2	L_3	L_4	L_1	L_5	L_6
45	L_7	L_3	L_7	L_5	L_6	L_4	L_3	L_1	L_2
46	L_7	L_4	L_4	L_7	L_2	L_3	L_1	L_6	L_5
47	L_7	L_5	L_7	L_6	L_5	L_3	L_4	L_2	L_1
48	L_7	L_6	L_7	L_6	L_4	L_5	L_2	L_3	L_1
49	L_7	L_7	L_7	L_4	L_1	L_2	L_3	L_6	L_5



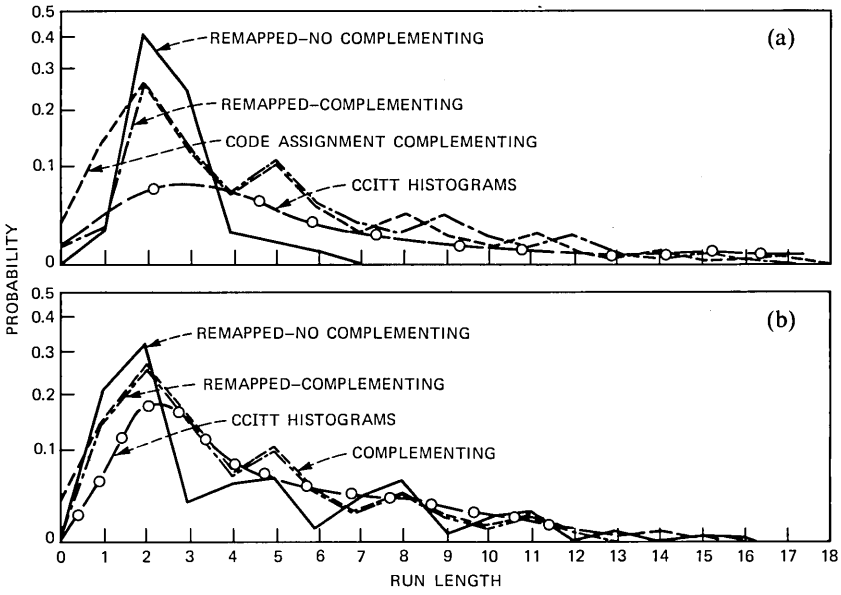
CCITT – COMITE CONSULTATIF INTERNATIONAL
TELEGRAPHIQUE ET TELEPHONIQUE

Fig. 8—Histograms of the (a) white and (b) black runs used in the 1D CCITT code and the processed gray-level pictures.

rearranged histograms. We found that in reality much of the gain in coding efficiency can be obtained by exchanging the code words for a few run lengths. This leads to simple preprocessing. Figure 9 shows the results of preprocessing on the histograms. It is clear that, if bit complementing is not used, after the preprocessing the code set is more attached to the histograms and therefore leads to more efficient code. However, if bit complementing is used, the histogram without any preprocessing is not too different from the CCITT histogram. Therefore, when bit complementing is used, the advantages of preprocessing are not large. Although this is the case for the picture we considered, more experiments are needed to evaluate statistics of typical pictures and usefulness of the preprocessing for such statistics.

2.5 Addressing

To encode positional information of the boundaries of a segmented picture, each composite line is considered as a sequence of alternating black and white runs corresponding to the lengths of two-level and gray-level segments, respectively. This is then coded by the two-



CCITT – COMITE CONSULTATIF INTERNATIONAL
TELEGRAPHIQUE ET TELEPHONIQUE

Fig. 9—Effect of preprocessing on the histograms of the (a) white and (b) black runs.

dimensional CCITT code and is transmitted at the beginning of each composite line.

2.6 Multiplexing and CCITT coding

The bits resulting from the above procedure for gray areas are multiplexed pel by pel with those of the two-level pels. Our experiments show that while it is advantageous to encode two-level areas by two-dimensional code, most of the two-dimensional correlation in gray segments is removed by two-dimensional prediction and code assignment. Therefore, gray areas are coded by one-dimensional code. Since the number of gray-level pels may vary from line to line, in order to maintain proper registration of two-level pels (for two-dimensional coding), a sample count is maintained and is used to initialize the two-dimensional coder once it comes out of the gray segment within a line.

III. SIMULATION RESULTS

Results of computer simulations are given in Tables III and IV. Table III shows results for 512 × 512 gray-level picture, and Table IV shows results for composite pictures with CCITT standard documents 4 and 5. It is clear from Table III that for the gray-level picture, without any preprocessing or bit complementing, coding efficiency is

Table III—Performance of coding algorithms for 512 × 512 gray picture

No.	Coding Algorithms	Coded bits (bits/pel)
1.	Entropy of the quantizer output (no subsampling)	1.84
2.	Entropy of the quantizer output (2:1 subsampling)	1.05
3.	Entropy with one-dimensional run-length coding (no complementing, no preprocessing, no subsampling)	2.76
4.	Entropy with one-dimensional run-length coding (no complementing, no preprocessing, 2:1 subsampling)	1.98
5.	One-dimensional run-length coding (no complementing, no preprocessing, CCITT code, no subsampling)	3.68
6.	One-dimensional run-length coding (no complementing, no preprocessing, CCITT code, 2:1 subsampling	2.86
7.	5+ preprocessing	3.14
8.	6+ preprocessing	2.24
9.	Entropy with one-dimensional run-length coding (complementing, no preprocessing)	1.88
10.	Entropy with one-dimensional run-length coding (complementing, no preprocessing, 2:1 subsampling)	1.17
11.	9+ CCITT code	2.05
12.	10+ CCITT code	1.31
13.	11+ preprocessing	1.98
14.	12+ preprocessing	1.19

Table IV—Performance of coding algorithms for composite pictures

No.	Coding algorithms	Coded bits	
		Document 4	Document 5
1.	One-dimensional CCITT code on noncomposite document	870803	547853
2.	Two-dimensional CCITT code on noncomposite document	577527	286911
3.	Two-dimensional code for two-level, one-dimensional code for gray level (no complementing)	1169270	893288
4.	Two-dimensional code for two-level, one-dimensional code for gray level (no complementing), 2:1 subsampling	961210	686777
5.	3+ preprocessing	1086589	811897
6.	4+ preprocessing	879796	601189
7.	(3) + complementing	909876	628626
8.	(4) + complementing	755133	470539
9.	(5) + complementing	894444	616956
10.	(6) + complementing	739874	453229
11.	Two-dimensional code for entire document	999125	714651
12.	Bits for addressing	18147	18153

rather low. This is a result of the mismatch of the run-length statistics. Considerable improvement is obtained by preprocessing the run lengths before applying the CCITT coder. Even higher improvement is obtained by the bit-complementing technique. Much of the mis-

match between the statistics is removed by the complementing technique, and therefore additional improvement obtained by preprocessing the complemented output is marginal. The use of complementing makes it possible to achieve bit rates that are close to the entropy of the coded output. Coding of composite pictures shows similar results. Another interesting conclusion from Table IV is that the two-dimensional CCITT code is not very efficient for gray-level segments of the composite picture. This is a result of lack of line-to-line correlation among bits that are outputs of the quantizer. Much of the line-to-line correlation is already removed by the two-dimensional prediction and the bit assignment based on conditional statistics.

IV. CONCLUSIONS

We have presented an algorithm that can automatically segment areas of a picture that require only two shades of gray from those that require many shades of gray. Gray areas are coded in a way that creates a bit stream that subsequently can be efficiently coded by a CCITT coder. We find that, for the gray areas, it is possible to achieve coding efficiencies close to the entropy of the DPCM quantizer output. Therefore, we conclude that it is possible to encode documents that contain an arbitrary mixture of two-level and multilevel areas using a CCITT coder that requires only a preprocessor at the transmitter and a postprocessor at the receiver.

REFERENCES

1. A. N. Netravali, ed., special issue on Digital Encoding of Graphics, Proc. IEEE, *68,* No. 7 (July 1980).
2. A. N. Netravali and J. O. Limb, "Picture Coding—A Review," Proc. IEEE, *68,* No. 3 (March 1980), pp. 366–406.
3. R. Hunter and A. H. Robinson, "International Digital Facsimile Coding Standards," Proc. IEEE, *68,* No. 7 (July 1980), pp. 830–46.

AUTHORS

Arun N. Netravali, B. Tech. (Honors), 1967, Indian Institute of Technology, Bombay, India; M.S., 1969, Ph.D. (Electrical Engineering), 1970, Rice University; Optimal Data Corporation, 1970–1972; Bell Laboratories, 1972—. Mr. Netravali has worked on problems related to filtering, guidance, and control for the space shuttle. At Bell Laboratories, he has worked on various aspects of digital processing and computing. He was a Visiting Professor in the Department of Electrical Engineering at Rutgers University. He is presently Director of the Computer Technology Research Laboratory. Mr. Netravali holds over 20 patents and has had more than 60 papers published. He was the recipient of the Donald Fink Prize Award for the best review paper published in the Proceedings of the IEEE. Editorial board, Proceedings of the IEEE; editor, Transactions on Communications; member, Tau Beta Pi, Sigma Xi; Senior Member, IEEE.

Hamid Gharavi, B.S.E.E., 1970, Tehran Polytechnic; M.S.C. (Digital Communication), Ph.D. (Electrical Engineering), 1979, University of Technology, Loughborough, England; Research fellow at University of Technology, Loughborough, England, 1979–1980; Lecturer at Auckland University, New Zealand, 1980–1981; Bell Laboratories, 1982—. Mr. Gharavi has worked on problems related to digital modulations for satellite communication. He also has worked on bandwidth compression of color television signals, source coding, graphics, and pattern recognition. Member, IEEE.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 9, November 1983
Printed in U.S.A.

The Queueing Network Analyzer

By W. WHITT*

(Manuscript received March 11, 1983)

This paper describes the Queueing Network Analyzer (QNA), a software package developed at Bell Laboratories to calculate approximate congestion measures for a network of queues. The first version of QNA analyzes open networks of multiserver nodes with the first-come, first-served discipline and no capacity constraints. An important feature is that the external arrival processes need not be Poisson and the service-time distributions need not be exponential. Treating other kinds of variability is important. For example, with packet-switched communication networks we need to describe the congestion resulting from bursty traffic and the nearly constant service times of packets. The general approach in QNA is to approximately characterize the arrival processes by two or three parameters and then analyze the individual nodes separately. The first version of QNA uses two parameters to characterize the arrival processes and service times, one to describe the rate and the other to describe the variability. The nodes are then analyzed as standard GI/G/m queues partially characterized by the first two moments of the interarrival-time and service-time distributions. Congestion measures for the network as a whole are obtained by assuming as an approximation that the nodes are stochastically independent given the approximate flow parameters.

I. INTRODUCTION AND SUMMARY

Networks of queues have proven to be useful models to analyze the performance of complex systems such as computers, switching machines, communications networks, and production job shops.[1-7] To facilitate the analysis of these models, several software packages have

* Bell Laboratories.

been developed in recent years, e.g., BEST/1,[8] CADS,[9] PANACEA,[10–12] and one based on Heffes.[13] These software packages contain algorithms for Markov models that can be solved exactly. For some applications, the model assumptions are at least approximately satisfied, so that the analysis can be very helpful. For many other applications, however, the model assumptions are not even approximately satisfied, so that the analysis can be misleading.

A natural alternative to an exact analysis of an approximate model is an approximate analysis of a more exact model. This paper describes a software package called the Queueing Network Analyzer (QNA), which was recently developed at Bell Laboratories to calculate approximate congestion measures for networks of queues. QNA goes beyond existing exact methods by treating non-Markov networks: The arrival processes need not be Poisson and the service-time distributions need not be exponential. QNA treats other kinds of variability by approximately characterizing each arrival process and each service-time distribution with a variability parameter. It is also possible to analyze large networks quickly with QNA because the required calculations are minimal, the most complicated part being the solution of a system of linear equations. The current version of QNA is written in FORTRAN.

Here is a rough description of the model: There is a network of nodes and directed arcs. The nodes represent service facilities and the arcs represent flows of customers, jobs, or packets. There is also one external node, which is not a service facility, representing the outside world. Customers enter the network on directed arcs from the external node to the internal nodes, move from node to node along the internal directed arcs, and eventually leave the system on one of the directed arcs from an internal node to the external node. The flows of customers on the arcs are assumed to be random so that they can be represented as stochastic processes.

If all servers are busy at a node when a customer arrives, then the customer joins a queue and waits until a server is free. When there is a free server, that customer begins service, which is carried out without interruption. Successive service times at each node are assumed to be random variables, which may depend on the type of customer but which otherwise are independent of the history of the network and are mutually independent and identically distributed. After the customer completes service, he goes along some directed arc from that node to another node. The customer receives service in this way from several internal nodes and then eventually leaves the network. A picture of a typical network (without the external node) is given in Fig. 1.

An important feature of the model is that there may be flows from node j to node i, as well as flows from node i to node j. This is of

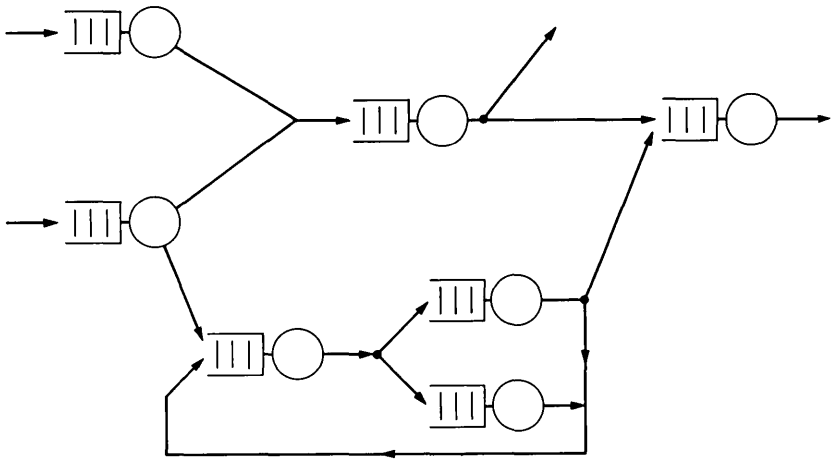


Fig. 1—An open network of queues.

course useful when customers can return to a node where they previously received service, but it is also useful when customers cannot return to a node where they previously received service. Then flows from node j to node i represent different customers than the customers that flow from node i to node j.

To be precise about the model, we give a list of basic assumptions. It is worth noting, however, that work is under way to extend QNA so that it can analyze systems in which each of the following assumptions is replaced by obvious alternatives. The general approximation technique is flexible, so that it is not difficult to modify and extend the algorithm.

Assumption 1. The network is *open* rather than closed. Customers come from outside, receive service at one or more nodes, and eventually leave the system.

Assumption 2. There are *no capacity constraints.* There is no limit on the number of customers that can be in the entire network and each service facility has unlimited waiting space.

Assumption 3. There can be *any number of servers* at each node. They are identical independent servers, each serving one customer at a time.

Assumption 4. Customers are selected for service at each facility according to the *first-come, first-served* discipline.

Assumption 5. There can be *any number of customer classes,* but customers cannot change classes. Moreover, much of the analysis in QNA is done for the aggregate or typical customer (see Sections 2.3 and VI).

Assumption 6. Customers can be created or combined at the nodes,

e.g., an arrival can cause more than one departure (see Section 2.2). (Think of messages.)

The general approach is to represent all the arrival processes and service-time distributions by a few parameters. The congestion at each facility is then described by approximate formulas that depend only on these parameters. The parameters for the internal flows are determined by applying an elementary calculus that transforms the parameters for each of the three basic network operations: superposition (merging), thinning (splitting), and flow through a queue (departure). These basic operations are depicted in Fig. 2. When the network is acyclic (e.g., queues in series), the basic transformations can be applied successively one at a time, but in general it is necessary to solve a system of equations or use an iterative method. To summarize, there are four key elements in this general approach:

1. *Parameters* characterizing the flows and nodes that will be readily available in applications and that have considerable descriptive power in approximations of the congestion at each node.
2. *Approximations for multiserver queues* based on the partial information provided by the parameters characterizing the arrival process and the service-time distribution at each node.
3. A *calculus for transforming the parameters* to represent the basic network operations: merging, splitting, and departure.
4. A *synthesis algorithm* to solve the system of equations resulting from the basic calculus applied to the network.

The current version of QNA uses two parameters to characterize the arrival processes and the service times, one to describe the rate and the other to describe the variability. (Three-parameter algorithms are being developed, however.) For the service times, the two param-

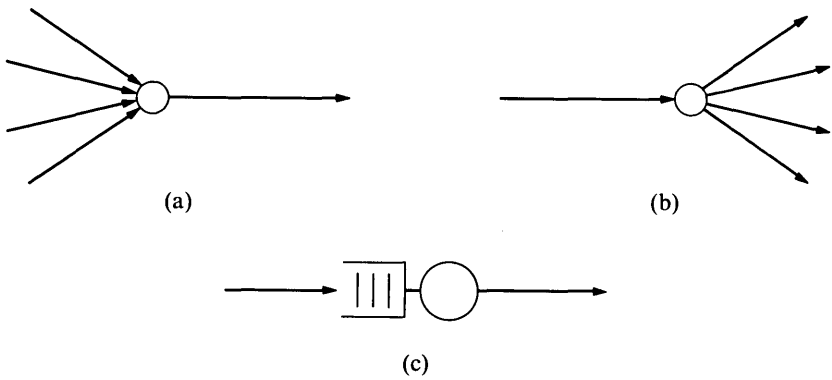


Fig. 2—Basic network operations: (a) Superposition or merging. (b) Decomposition or splitting. (c) Departure or flow through a queue.

eters are the first two moments. However, we actually work with the mean service time τ and the squared coefficient of variation c_s^2, which is the variance of the service time divided by the square of its mean. The user has the option of working with the service rate $\mu = \tau^{-1}$ instead of τ. For the arrival processes, the parameters are associated with renewal-process approximations. The first two parameters are equivalent to the first two moments of the renewal interval (interval between successive points) in the approximating renewal process. The equivalent parameters we use are the arrival rate λ, which is the reciprocal of the renewal-interval mean, and the squared coefficient of variation c_a^2, which is the variance of the renewal interval divided by the square of its mean.

We obtain the approximation of the flows by applying the general framework and the basic procedures for approximating point processes in Whitt,[14] incorporating refinements such as the hybrid procedures developed for merging by Albin.[15,16] Of course, the general idea of simple two-parameter approximations for stochastic point processes goes back at least to the equivalent random method for approximating overflow streams (see Wilkinson,[17] Cooper,[18] and references there). Renewal-process approximations for such point processes were introduced by Kuczura[19] (also see Rath and Sheng[20]). Two-parameter approximations for networks of queues similar to QNA have also been developed by others, apparently first by Reiser and Kobayashi[21] (also see Kuehn,[22] Sevcik et al.,[23] Chandy and Sauer,[24] Chapter 4 of Gelenbe and Mitrani,[4] and Shanthikumar and Buzacott[6]). These two-parameter approximations for networks of queues are also similar in spirit to two-parameter approximations for networks of blocking systems with alternate routing (see Katz[25]).

Some authors have referred to these two-parameter heuristic approximations for networks of queues as diffusion approximations,[4,21] but diffusion processes are not actually used. Diffusion approximations and associated heavy-traffic limit theorems have motivated some of the heuristic approximations in the literature and in QNA, and they are closely related to the asymptotic method for approximating point processes,[14] but the heuristic approximations in QNA are not the same as the more complicated diffusion approximations for networks of queues in Iglehart and Whitt,[26] Harrison and Reiman,[27] and Reiman.[28,29]

The approximation method in QNA is perhaps best described as a parametric-decomposition method,[22] because the nodes are analyzed separately after the parameters for the internal flows are determined. Moreover, when the congestion measures are calculated for the network as a whole, the nodes are treated (approximately) as being stochastically independent. This independence can be interpreted as

a generalization of the product-form solution that is valid for Markovian networks, i.e., in the Markov models the components of the vector representing the equilibrium number of customers at each node are stochastically independent, so that the probability mass functions for the vector is the product of the probability mass functions for the components. While QNA can be thought of as a decomposition method or an extended-product-form solution, an effort is made to capture the dependence among the nodes. The idea is to represent this dependence approximately through the internal flow parameters.

To see the motivation for QNA, consider the elementary open network containing a single node with a single server, an infinite waiting room, and the first-come, first-served discipline. Suppose there is a single customer class with each customer being served only once before departing. The standard Markov model of this elementary network, which is embodied in BEST/1, CADS, and PANACEA, is the classical M/M/1 queue,[1,3,18] which has a Poisson arrival process and an exponential service-time distribution. For the M/M/1 model, the expected waiting time EW (before the customer begins service) is

$$EW = \tau\rho/(1 - \rho), \tag{1}$$

where τ is the mean service time and ρ is the traffic intensity, which is assumed to satisfy $0 \leqslant \rho < 1$.

On the other hand, QNA uses an approximation for the GI/G/1 model to represent this one-node network. The GI/G/1 model has a renewal arrival process and both the interarrival-time distribution and the service-time distribution are general. In QNA, the arrival process is represented by a renewal process partially characterized by two parameters: the arrival rate λ and the variability parameter c_a^2. The service-time distribution is also partially characterized by two parameters: the mean service time τ and the variability parameter c_s^2. In contrast to (1), the formula for the expected waiting time in QNA is

$$EW = \tau\rho(c_a^2 + c_s^2)g/2(1 - \rho), \tag{2}$$

where $g \equiv g(\rho, c_a^2, c_s^2)$ is either one (when $c_a^2 \geqslant 1$) or less than one (when $c_a^2 < 1$); see (45). When $g(\rho, c_a^2, c_s^2) = 1$, (2) differs from (1) by the factor $(c_a^2 + c_s^2)/2$. When the arrival process is Poisson, $c_a^2 = 1$; when the service-time distribution is exponential, $c_s^2 = 1$. Hence, if the GI/G/1 model is actually an M/M/1 model, (2) reduces to (1). Of course, the user of QNA can set $c_a^2 = c_s^2 = 1$ and obtain (1). In fact, the values $c_a^2 = 1$ and $c_s^2 = 1$ are default values that the program uses if the user does not have variability parameters to provide. Each c^2 can assume any nonnegative value: $c^2 = 0$ for the degenerate deterministic distribution; $c^2 = k^{-1}$ for an erlang E_k, the sum of k i.i.d. exponential random variables; and $c^2 > 1$ for mixtures of exponential

distributions. Obviously, the difference between (2) and (1) can be large, so that (2) often significantly reduces the error.

To obtain (2), we studied the GI/G/1 queue partially characterized by the moments of the interarrival-time and service-time distributions. Building on previous work by Holtzman,[30] Rolski,[31] and Eckberg,[32] we investigated the set of possible values of EW given the partial information.[33–37] When $c_a^2 \geq 1$, formula (2) is always a possible value, i.e., there always is a GI/G/1 system with interarrival-time and service-time distributions having the specified moments in which (2) is correct. In general, (2) appears to be a reasonably typical value.

For the single-node example just considered, the arrival process was a renewal process. More generally, it is natural to think of all the non-Poisson arrival processes in the model as renewal processes, either because they are initially renewal processes or because the algorithm can be interpreted as approximating general arrival processes by renewal processes. Hence, with one customer class, it is natural to think of the model as a generalization of the open Jackson network M/M/m queues to an open Jackson network of GI/G/m queues. Each node is approximated by a GI/G/m queue having a renewal arrival process independent of service times that are independent and identically distributed with a general distribution. It is significant that QNA is consistent with the Jackson network theory: If there is a single class of customers, if all the arrival processes are Poisson, and if all the service-time distributions are exponential, then QNA is exact. However, for the general model few analytical results are available, so approximations are needed.

The software package QNA has a flexible input procedure: the model will accept more than one kind of input (see Section II). For the standard input, only limited information is required. Only two parameters are needed for each service-time distribution and each external arrival process. Also, a routing matrix is needed, which gives the proportion of those customers completing service at facility i that go next to facility j. (The algorithm is based on Markovian routing.) Hence, for n nodes, the input consists of $n^2 + 4n$ numbers.

There is also an alternate input by classes and routes. In this scheme there are different classes of customers and each class enters the network at a fixed node and passes through a specified sequence of nodes. For each class, there are two parameters characterizing its external arrival process and two parameters characterizing the service-time distribution at each node on its route. With this input by routes, different classes can have different service-time distributions at a given node and the same class can have different service-time distributions during different visits to the same node. For a class with n nodes on its route, the input consists of $3n + 2$ numbers. (This includes

the n nodes on the route.) QNA analyzes this route input by aggregation: All the classes are aggregated by QNA to convert the route input into the standard input. Afterwards, the special parameters of each class are used to describe its sojourn times.

QNA also provides a fairly rich output. Several different congestion measures are calculated for each node: the traffic intensity (utilization), the expected number of busy servers (offered load), and the mean and variance of the equilibrium delay and number of customers present. In fact, for single-server nodes the delay distribution itself is described. Congestion measures for the entire network are also calculated, under the approximation assumption that the nodes are stochastically independent given the approximate flow parameters. Means and variances of total service times, total delays, and total sojourn times (response times) are given. When the input is by routes, these characteristics are given for each customer class. Otherwise, these characteristics are given for any route requested by the user.

A desirable feature of QNA is the structure of the calculus to transform the parameters to characterize the internal flows. The calculus is linear for each network operation, so that the parameters for the internal flows are determined simply by solving systems of linear equations. For the rates, the system of linear equations is just the familiar traffic rate equations occurring in the Jackson network of M/M/m queues. After having obtained the rates, we obtain the variability parameters of the internal flows (the squared coefficients of variations) by solving another system of linear equations. As a by-product, the existence of a unique nonnegative solution for the flow parameters is trivially guaranteed. There is no guarantee that an iterative scheme will converge, and if it does, there is typically no guarantee that a solution is unique. The linearity also guarantees that the computation required is not great. Since there is only one linear equation per node in the network, QNA can be used to analyze large networks repeatedly at minimal cost.

The linear calculus for transforming the variability parameters incorporates results of recent studies to improve the accuracy of the approximations. The general framework for approximating point processes in Whitt[14] is used. Significant improvement over previous approximation methods of this kind has been obtained by paying particular attention to the difficult superposition operation. For superposition, we use a modification of the hybrid procedure developed at Bell Laboratories by Albin.[15,16,38,39]

We emphasize that QNA is approximate. In applications it is important to validate the QNA output by comparing it with simulations and/or measurements. QNA is designed so that it is easy to incorporate improvements and it is easy to tune QNA for particular applications.

QNA also provides a useful framework for developing new approximation procedures. Moreover, it is easy to use QNA in conjunction with other special algorithms available to analyze the nodes or the flows.

The rest of this paper describes QNA in more detail. The paper is organized—just as the output is—according to the main steps in the analysis. The input is described in Section II. Section III describes the preliminary analysis to eliminate immediate feedback. The procedures to determine the internal flow parameters are contained in Section IV, and the procedures to calculate approximate congestion measures for the nodes are contained in Section V. Section VI contains the procedures to calculate approximate congestion measures for the network as a whole.

In a sequel in this issue of the *Journal*,[40] we describe the performance of QNA by comparing it with simulation and other approximations of several networks of queues. The sequel illustrates how to apply QNA and demonstrates the importance of the variability parameters.

II. THE INPUT

In this section we describe the input options currently available for QNA. We anticipate more input options in the future. In Section 2.1 we describe the standard input, which is relatively compact. In Section 2.2 we describe a minor modification of the standard input, which allows for the creation or combination of customers at the nodes. For example, when a packet completes service at some node, it may cause several packets to be sent to other nodes. In Section 2.3 we describe an alternate input for different classes of customers having specified routes. We also describe the way QNA converts this input by classes and routes into the standard input of Section 2.1.

2.1 The standard input

With the standard input, there is a single customer class and no creation of customers at the nodes. Any number of networks can be processed during a single run, so the user first specifies the number of networks. Then, for each network, the user specifies the number of nodes, and for each node the number of servers. For each node in the network, there are two parameters for the service-time distribution and two parameters for the external arrival process. Finally, there is a routing matrix, indicating the proportion of customers that go to node j from node i. Here is a list of the input data for each network with the notation we use:

n = number of (internal) nodes in the network
m_j = number of servers at node j
λ_{0j} = external arrival rate to node j

c_{0j}^2 = variability parameter of the external arrival process to node j (squared coefficient of variation of the renewal interval in the approximating renewal process)

τ_j = mean service time at node j

c_{sj}^2 = squared coefficient of variation of the service-time distribution at node j

q_{ij} = proportion of those customers completing service at node i that go next to node j.

In matrix notation, $Q \equiv (q_{ij})$ is an $n \times n$ matrix and $\Lambda_0 \equiv (\lambda_{0j})$ is an $1 \times n$ vector. The user has the option of inputting τ_j or its reciprocal μ_j, the service rate at node j. (The same form must be used for all nodes.)

The user need not specify the variability parameters c_{0j}^2 and c_{sj}^2, in which case they are set equal to the default value one, corresponding to the M/M/1 model having a Poisson arrival proccess and an exponential service-time distribution. (Again, this option applies to all nodes.) Alternatively, the user can specify only the service-time variability parameters, c_{sj}^2, in which case either all the arrival-process variability parameters are automatically set equal to 1, yielding an M/G/1 approximation for each node, or the QNA algorithm is applied.

2.2 Creating and combining customers

QNA has an option to allow creating or combining customers at the nodes following the completion of service. For example, a message processed at some node might cause messages to be sent to several other nodes. Alternatively, messages might be divided into packets after service at one node and then later recombined into messages after service at another node. In a job shop, the focus might shift back and forth between units and lots, e.g., at different nodes we might consider bottles, six-packs, cases, and even truckloads.

With this option, the user must specify the multiplicative factor γ_j of customer creation or combination at node j for each j. There is customer creation (combination) at node j if $\gamma_j > 1$ ($\gamma_j < 1$). If customers are neither created nor combined, then $\gamma_j = 1$. If λ_j is the overall arrival rate to node j, then the departure rate, after this modification, is $\lambda_j \gamma_j$ and the rate of departure from the network j is

$$\lambda_j \gamma_j \left(1 - \sum_{k=1}^{n} q_{jk}\right).$$

When artificial nodes are used, the creation or combination can also be placed before service.

To obtain our approximation formulas, we work with the following models of customer creation and combination. These models require integer values, but the approximation formulas and the QNA input do

not. For customer creation, we replace each departure from node j with a batch of size γ_j. For customer combination, we replace γ_j^{-1} successive interdeparture intervals by a single one. From these models it is not difficult to calculate the impact of γ_j, e.g., the departure rate at node j is simply multiplied by γ_j.

2.3 Input by classes and routes

QNA provides the option of defining different customer classes. Each class has its own route or itinerary that specifies the sequence of nodes visited. Thus, for each class the routing is deterministic. Each class has an external arrival process that goes to the first node on the route. As usual, the external arrival process is characterized by rate and variability parameters. Also, each class may have its own service-time distribution at each node on its route. The service-time distributions can be different, not only for different classes, but also for different visits to the same node by the same class. These service-time distributions are also characterized by rate and variability parameters. (Alternatively, the user can elect to input the service-time parameters for each node. Then all classes have the same service-time distribution at each visit to a particular node.)

As with the standard input, the user must specify the number of nodes and the number of servers at each node. Now we need the number of routes too. The required data are:

n = number of nodes
m_j = number of servers at node j
r = number of routes.

Here is a list of the input data for the kth customer class of a network:

n_k = number of nodes on route k
$\hat{\lambda}_k$ = external arrival rate of class k
c_k^2 = variability parameter of the external arrival process for class k
n_{kj} = the jth node visited by customer class k
τ_{kj} = the mean service time of class k at the jth node of its route
c_{skj}^2 = the variability parameter of the service-time distribution of class k at the jth node of its route.

QNA converts this input by classes and routes into the standard input in Section 2.1. It then calculates the parameters of a typical or aggregate customer. Later, when computing sojourn times or response times of each customer class, QNA uses the service-time parameters of that customer class. The first version of QNA assumes as an approximation that each customer sees independent versions of the equilibrium distribution at each node. Hence, the waiting time before beginning service at each node is assumed to be the same for all classes and all visits.

We now indicate how QNA converts the input by classes and routes into the standard input of Section 2.1. For this purpose, let $1H$ be the indicator function of the set H, i.e., $1H(x) = 1$ if $x \in H$ and $1H(x) = 0$ otherwise.

First, we obtain the external arrival rates by

$$\lambda_{0j} = \sum_{k=1}^{r} \hat{\lambda}_k 1\{k{:}n_{k1} = j\}, \tag{3}$$

i.e., the external arrival rate at node j, λ_{0j}, is the sum of all route arrival rates $\hat{\lambda}_k$ for which the first node on the route is j. (Here the $1H$ notation is used for $H = \{k{:}n_{k1} = j\}$.) Similarly, the flow rate from i to j is

$$\lambda_{ij} = \sum_{k=1}^{r} \sum_{\ell=1}^{n_k-1} \hat{\lambda}_k 1\{(k, \ell){:}n_{k\ell} = i,\ n_{k,\ell+1} = j\} \tag{4}$$

and the flow from i out of the network is

$$\lambda_{i0} = \sum_{k=1}^{r} \hat{\lambda}_k 1\{k{:}n_{kn_k} = i\}. \tag{5}$$

From (4) and (5), we obtain the routing matrix Q. The proportion of customers that go to j from i is

$$q_{ij} = \lambda_{ij} \Big/ \left(\lambda_{i0} + \sum_{k=1}^{n} \lambda_{ik}\right). \tag{6}$$

If node i is an active part of the network, then the denominator will be strictly positive. Otherwise, QNA gives an error message.

Next, if the service-time parameters are given by routes, we obtain the service-time parameters for the nodes by averaging:

$$\tau_j = \frac{\sum_{k=1}^{r} \sum_{\ell=1}^{n_k} \hat{\lambda}_k \tau_{k\ell} 1\{(k, \ell){:}n_{k\ell} = j\}}{\sum_{k=1}^{r} \sum_{\ell=1}^{n_k} \hat{\lambda}_k 1\{(k, \ell){:}n_{k\ell} = j\}}. \tag{7}$$

The denominator in (7) will be strictly positive if node j is ever visited. Otherwise, as with (6), QNA supplies an error message.

We obtain the node variability parameters c_{sj}^2 using the property that the second moment of a mixture of distributions is the mixture of the second moments. Therefore, we have

$$\tau_j^2(c_{sj}^2 + 1) = \frac{\sum_{k=1}^{r} \sum_{\ell=1}^{n_k} \hat{\lambda}_k \tau_{k\ell}^2 (c_{sk\ell}^2 + 1) 1\{(k, \ell){:}n_{k\ell} = j\}}{\sum_{k=1}^{r} \sum_{\ell=1}^{n_k} \hat{\lambda}_k 1\{(k, \ell){:}n_{k\ell} = j\}}. \tag{8}$$

At this point, QNA has calculated enough information about the

standard input to compute the internal flow rates λ_j and the traffic intensities ρ_j as described in Section 4.1, i.e.,

$$\rho_j = \lambda_j \tau_j / m_j. \tag{9}$$

QNA uses this information to calculate the variability parameters c_{0j}^2 of the external arrival process. The hybrid approximation for superposition arrival processes in Section 4.2 is also used here because the external arrival process to node j is the superposition of the external arrival processes to node j from the different classes. If $\lambda_{0j} = 0$, then c_{0j}^2 does not matter and QNA sets $c_{0j}^2 = 1$. Otherwise,

$$c_{0j}^2 = (1 - \bar{w}_j) + \bar{w}_j \left[\sum_{k=1}^{r} c_k^2 \left(\hat{\lambda}_k 1\{k{:}n_{k1} = j\} \Big/ \sum_{\ell=1}^{r} \hat{\lambda}_\ell 1\{k{:}n_{k1} = j\} \right) \right], \tag{10}$$

where

$$\bar{w}_j \equiv \bar{w}_j(\rho_j, \bar{\nu}_j) = [1 + 4(1 - \rho_j)^2(\bar{\nu}_j - 1)]^{-1}, \tag{11}$$

ρ_j is the traffic intensity in (9), and

$$\bar{\nu}_j = \left[\sum_{k=1}^{r} \left(\hat{\lambda}_k 1\{k{:}n_{k1} = j\} \Big/ \sum_{\ell=1}^{r} \hat{\lambda}_\ell 1\{\ell{:}n_{\ell 1} = j\} \right)^2 \right]^{-1}. \tag{12}$$

Example 1: To help fix the ideas, we consider an elementary example with $n = 2$ nodes and $r = 3$ routes. Let the number of servers at the nodes be $m_1 = 40$ and $m_2 = 10$. Let the route input be described by vectors

$$(n_k, \hat{\lambda}_k, c_k^2; n_{k1}, \tau_{k1}, c_{sk1}^2; \cdots; n_{kn_k}, \tau_{kn_k}, c_{skn_k}^2). \tag{13}$$

Here suppose that the r vectors are:

$$\begin{aligned} &(2, 2, 1; 1, 1, 1; 1, 3, 3) \\ &(3, 3, 2; 1, 2, 0; 2, 1, 1; 1, 2, 1) \\ &(2, 2, 4; 2, 1, 1; 1, 2, 1). \end{aligned} \tag{14}$$

The first route corresponds to a Poisson arrival process at rate 2 to node 1, with all customers being fed back immediately for a second service before departing from the network. (Of course, the arrival process need not actually be Poisson; a Poisson process always has $c^2 = 1$ but other processes could have $c^2 = 1$ too.) The second class also enters at node 1, then goes to node 2 and back to node 1 before departing from the network, etc.

By (3), the external arrival rates are $\lambda_{01} = 5$ and $\lambda_{02} = 2$. By (4), the internal flow rates are $\lambda_{11} = 2$, $\lambda_{12} = 3$, $\lambda_{21} = 5$, and $\lambda_{22} = 0$. By (5), the flow rates out of the network are $\lambda_{10} = 7$ and $\lambda_{20} = 0$. By (6), the routing probabilities are: $q_{11} = 1/6$, $q_{12} = 1/4$, $q_{21} = 1$, and $q_{22} = 0$. By

(7), the mean service times are $\tau_1 = 2$ and $\tau_2 = 1$. By (8), $c_{s1}^2 = 1.67$ and $c_{s2}^2 = 1.00$. Note that both service times at node 2 in (14) have mean 1 and squared coefficient of variation 1, as with a common exponential distribution, so we should want $\tau_2 = c_{s2}^2 = 1$.

To obtain the internal arrival rates, we solve the traffic rate equations as in Section 4.1, i.e.,

$$\lambda_j = \lambda_{0j} + \sum_{i=1}^{n} \lambda_i q_{ij}, \tag{15}$$

to obtain $\lambda_1 = 12$, $\lambda_2 = 5$, $\rho_1 = 0.6$, and $\rho_2 = 0.5$.

Finally we obtain the variability parameters c_{0j}^2. First, from (12), $\bar{\nu}_1 = 25/13$ and $\bar{\nu}_2 = 1.0$. Then, from (11), $\bar{w}_1 = 0.629$ and $\bar{w}_2 = 1$, so that $c_{01}^2 = 1.38$ and $c_{02}^2 = 4$. Since there is only one external arrival process to node 2, we should have $\bar{\nu}_2 = \bar{w}_2 = 1$ and $c_{02}^2 = c_3^2 = 4$.

III. ELIMINATING IMMEDIATE FEEDBACK

In this section we describe a function that QNA can perform before calculating the internal flow parameters and analyzing the congestion. The user can elect to reconfigure the network to eliminate immediate feedback. This procedure, which was originally suggested by Kuehn,[22] usually improves the quality of approximations. Hence, it is recommended and is performed in the standard version of QNA.

Immediate feedback occurs whenever $q_{ii} > 0$. Since QNA assumes Markovian routing, each customer completing service at node i is immediately fed back to node i to be served again with probability q_{ii}. Each time the customer goes to the end of the line. With the decomposition method, QNA assumes the customer finds the equilibrium number of customers at the node each time, with each visit being an independent experiment.

QNA eliminates immediate feedback by giving each customer, upon arrival from another node, his or her total service time before going to a different node. This is equivalent to putting a customer immediately fed back at the head of the line instead of at the end of the line. Transitions from node i back to node i are eliminated and the new probability of a transition to node j becomes the old conditional probability given that the customer departs from node i. In other words, each visit to node i from elsewhere plus all subsequent times immediately fed back are interpreted as a single visit. The service time is increased to compensate.

The motivation for this procedure is easy to explain. For a multiserver node with Bernoulli (Markovian) feedback and iid service times that are independent of a general arrival process (not necessarily Poisson or renewal), the distribution of the queue length process (but not the waiting times) is the same after this transformation. Hence,

we calculate the approximate values of the mean and variance of the equilibrium queue length for the transformed node without feedback and use them to derive approximate waiting time characteristics. By Little's formula,[41,42] the expected waiting time is also exact, i.e., the only error is in the approximation of the arrival process by a renewal process and the approximations for the characteristics of the GI/G/m queue; there is no additional error due to the immediate feedback.

The first step of the reconfiguring procedure is quite simple: the new service time is regarded as a geometric mixture of the n-fold convolution of the old service-time distribution. The parameters τ_i, c_{si}^2, and q_{ij} are changed to $\hat{\tau}_i$, $\hat{c}_{si}^2$, and $\hat{q}_{ij}$ when $q_{ii} > 0$:

$$\begin{aligned}
\hat{\tau}_i &= \tau_i/(1 - q_{ii}) \\
\hat{c}_{si}^2 &= q_{ii} + (1 - q_{ii})c_{si}^2 \\
\hat{q}_{ii} &= 0 \\
\hat{q}_{ij} &= q_{ij}/(1 - q_{ii}), \qquad j \neq i.
\end{aligned} \tag{16}$$

Afterwards, when calculating congestion measures for node i, QNA makes further adjustments. When we eliminate immediate feedback according to (16), we no longer count the times a customer is fed back immediately as separate visits. Hence, we need to adjust the congestion measures that are expressed per visit. For example, since the expected number of visits to node i per visit from outside is $(1 - q_{ii})^{-1}$, to obtain the expected waiting time per original visit to node i, we multiply the values of the expected waiting time EW_i obtained from (16) by $(1 - q_{ii})$. Of course, the number of customers at each node is not affected by the feedback treatment.

Let $\bar{\lambda}_i$, $\bar{\tau}_i$, etc., represent the new adjusted values. In terms of the parameters λ_i, τ_i, etc., obtained using (16), the new adjusted values are:

$$\begin{aligned}
\bar{\lambda}_i &= \lambda_i/(1 - q_{ii}) \\
\bar{\tau}_i &= (1 - q_{ii})\tau_i \\
c_{si}^2 &= (c_{si}^2 - q_{ii})/(1 - q_{ii}) \\
E\bar{W}_i &= (1 - q_{ii})EW_i \\
\mathrm{Var}(\bar{W}_i) &= (1 - q_{ii})\mathrm{Var}(T_i') - \bar{c}_{si}^2\bar{\tau}_i^2 \\
\mathrm{Var}(T_i') &= c^2(T_i')(EW_i + \tau_i)^2 \\
c^2(T_i') &= c^2(\bar{T}_i')(1 + q_{ii}) + q_{ii} \\
c^2(\bar{T}_i') &= (\mathrm{Var}\ \bar{W}_i' + \bar{c}_{si}^2\bar{\tau}_i^2)(E\bar{W}_i + \bar{\tau}_i)^{-2} \\
\mathrm{Var}(\bar{W}_i') &= EN_i\bar{c}_{si}^2\bar{\tau}_i^2 + c^2(N_i)(EN_i)^2\bar{\tau}_i^2,
\end{aligned} \tag{17}$$

where N_i represents the equilibrium number of customers at node i and the T_i variables represent the sojourn time per visit at node i.

We obtain the variables $\bar{\tau}_i$ and $\bar{c}_{si}^2$ in (17) by inverting the operation in (16), so we receive the original data again. The last five formulas in (17) involving the second-moment characteristics of $\bar{W}_i$ are based on the results of N_i in the transformed system and heavy-traffic limit theorems for networks of queues by Reiman.[28,29] The main quantity desired is $\mathrm{Var}(\bar{W}_i)$; the variable $\bar{W}_i'$ is a preliminary approximation for $\bar{W}_i$.

In heavy traffic, the changes in the queue length at the nodes are negligible during a customer's sojourn in the network. Hence, if node i is visited X_i times by some customer, then the total sojourn time at node i, say T_i', is distributed approximately (in heavy traffic) as $X_i\bar{T}_i'$, where X_i is independent of $\bar{T}_i'$ and $\bar{T}_i'$ is the sojourn time per individual visit in (17). (We use T_i' and $\bar{T}_i'$ instead of T_i and $\bar{T}_i$ because we do not use the description of T_i obtained directly from (16) and $\bar{T}_i$ will differ from $\bar{T}_i'$.) By the independence, $ET_i'^2 = EX_i^2E\bar{T}_i'^2$. Since X_i is geometrically distributed with mean $(1 - q_{ii})^{-1}$, $c^2(X_i) = q_{ii}$, and we obtain the seventh formula in (17).

The sixth and eighth formulas in (17) just express the formula for c^2 in terms of the mean and variance and the fact that the sojourn time is the sum of a waiting time and a service time. The final formula for $\mathrm{Var}(W_i')$ is obtained by approximating W_i' by the sum of N_i iid service times, using standard formulas for the variance of a random sum (e.g., compute $EW_i'^2$ by first conditioning on N_i). Finally, we obtain the fifth formula for $\mathrm{Var}(\bar{W}_i)$ by splitting the variance of T_i' into waiting-time and service-time components and dividing by the expected number of visits to node i. As a consequence, $\mathrm{Var}(T_i')$ seems more reliable than $\mathrm{Var}(\bar{W}_i)$. This procedure makes $\mathrm{Var}(T_i)$, computed from $\mathrm{Var}(\bar{W}_i)$ by adding variance components as in Section VI, agree with the direct formula for $\mathrm{Var}(T_i')$ in (17).

The congestion measures based on (16) can be used to describe the total delays and total sojourn times of arbitrary customers in the network as in Section 6.2, but the congestion measures based on (17) are needed to describe the behavior of particular customers with specified routes as in Section 6.3. However, as stated above, $\mathrm{Var}(T_i')$ in (17) is an attractive alternative to $\mathrm{Var}(T_i)$ obtained via (16).

Experience indicates that eliminating immediate feedback often yields a better approximation (see Kuehn[22] and Sections V and VII of Whitt[40]). It is also often desirable to reconfigure the network to eliminate almost-immediate feedback, e.g., flows that return relatively quickly after passing through one or more other nodes (see Section V of Whitt[40]). Further study is needed to understand feedback phenomena and to develop improved approximations.

IV. THE INTERNAL FLOW PARAMETERS

In this section we indicate how QNA calculates the internal flow parameters. In Section 4.1 we focus on the flow rates, which are obtained via the traffic rate equations, just as with the Markov models. In Section 4.2 we display the corresponding system of linear equations yielding the variability parameters. The remaining subsections explain how the variability parameter equations were obtained. The basic operations of superposition, splitting, and departure are discussed in Section 4.3, Section 4.4, and Section 4.5, and their synthesis in Section 4.6.

4.1 Traffic-rate equations

In this step QNA calculates the total arrival rate to each node. Let λ_j be the total arrival rate to node j, let γ_j be the multiplicative factor of customer creation at node j as specified in Section 2.2, and let δ_j be the departure rate (to other nodes as well as out of the network) at node j. In general, $\delta_j = \lambda_j\gamma_j$. If there is no customer creation, then $\gamma_j = 1$ and the rate in equals the rate out.

The fundamental *traffic-rate equations* are just

$$\lambda_j = \lambda_{0j} + \sum_{i=1}^{n} \lambda_i\gamma_i q_{ij} \tag{18}$$

for $j = 1, 2, \cdots, n$, or in matrix notation

$$\Lambda = \Lambda_0(I - \Gamma Q)^{-1}, \tag{19}$$

where $\Lambda_0 \equiv (\lambda_{0j})$ is the external arrival-rate vector, $Q \equiv (q_{ij})$ is the routing matrix, and $\Gamma = (\gamma_{ij})$ is the diagonal matrix with $\gamma_{ii} = \gamma_i$ and $\gamma_{ij} = 0$ for $i \neq j$. When there is no customer creation, $\gamma_i = 1$ and $\Gamma = I$. Of course, (18) is just a system of linear equations. To solve them is equivalent to inverting the matrix $(I - \Gamma Q)$ in (19). When customers can be created at the nodes as in Section 2.2, special care should be taken to be sure that (18) has a solution. We need to have $sp(\Gamma Q) < 1$ where $sp(\Gamma Q)$ is the spectral radius of ΓQ.

Given the arrival rates, it is possible to solve for the *traffic intensities* or *utilizations* at each node, defined by

$$\rho_i = \lambda_i\tau_i/m_i, \qquad 1 \leqslant i \leqslant n. \tag{20}$$

If $\rho_i \geqslant 1$, then the ith node is *unstable.* If any node is unstable, the algorithm gives an error message, prints out the traffic intensities, and stops. The associated *offered load* at node i, which coincides with the expected number of busy servers [see p. 400 of Heyman and Sobel[41] or (4.2.3) of Franken et al.[42]] is

$$\alpha_i = \lambda_i\tau_i, \qquad 1 \leqslant i \leqslant n. \tag{21}$$

The parameters α_i and ρ_i coincide for a single server, with α_i tending to be more useful as the number of servers, m_i, increases (obviously when $m_i = \infty$).

After the arrival rates have been calculated for the nodes, QNA calculates related quantities for the arcs:

$$\begin{aligned} \lambda_{ij} &= \lambda_i \gamma_i q_{ij} && \text{—the arrival rate to node j from node i} \\ p_{ij} &= \lambda_{ij}/\lambda_j && \text{—the proportion of arrivals to j that came from } i,\ i \geqslant 0. \end{aligned} \tag{22}$$

Similarly, QNA calculates the following output rates:

$$\begin{aligned} d_i &= \lambda_i \gamma_i \left(1 - \sum_{j=1}^{n} q_{ij}\right) && \text{—the departure rate out of the network from node i} \\ d &= \sum_{i=1}^{n} d_i \text{ —} && \text{the total departure rate out of the network.} \end{aligned} \tag{23}$$

4.2 Traffic variability equations

The heart of the approximation is the system of equations yielding the variability parameters for the internal flows, i.e., the squared coefficients of variation for the arrival processes, c_{aj}^2. (These are derived in Sections 4.3 through 4.7.) The equations are linear, of the form

$$c_{aj}^2 = a_j + \sum_{i=1}^{n} c_{ai}^2 b_{ij}, \qquad 1 \leqslant j \leqslant n, \tag{24}$$

where a_j and b_{ij} are constants, depending on the input data:

$$a_j = 1 + w_j \left\{ (p_{0j} c_{0j}^2 - 1) + \sum_{i=1}^{n} p_{ij}[(1 - q_{ij}) + (1 - \nu_{ij})\gamma_i q_{ij} \rho_i^2 x_i] \right\} \tag{25}$$

and

$$b_{ij} = w_j p_{ij} q_{ij} \gamma_i [\nu_{ij} + (1 - \nu_{ij})(1 - \rho_i^2)], \tag{26}$$

where x_i, ν_{ij}, and w_j depend on the basic data determined previously, e.g., ρ_i, m_i and c_{si}^2, but not on the variability parameters c_{aj}^2 being calculated. The parameter γ_i is the multiplicative factor of customer creation or combination, introduced in Section 2.2. The variables x_i and ν_{ij} are used to specify the departure operation; the variable w_j is used to specify the superposition operation. The variables ν_{ij} and w_j are weights or probabilities that are used in convex combinations

arising in hybrid approximations for departure and superpositions, respectively. The variables x_j, ν_{ij}, and w_j are included to make modification of the algorithm based on (24) easy. The specific values in this version of QNA are:

$$x_i = 1 + m_i^{-0.5}(\max\{c_{si}^2, 0.2\} - 1), \tag{27}$$

$$\nu_{ij} = 0, \tag{28}$$

and

$$w_j = [1 + 4(1 - \rho_j)^2(\nu_j - 1)]^{-1} \tag{29}$$

with

$$\nu_j = \left[\sum_{i=0}^{n} p_{ij}^2\right]^{-1} \tag{30}$$

and p_{ij} in (22).

It is significant that it is easy to modify this system of equations. For example, other hybrid procedures for departures or superpositions can be introduced just by changing ν_{ij} and w_j, respectively. In this way, it is easy to calculate and compare the variability parameters for several different approximation procedures.

4.3 Superposition

The purpose of the following sections is to explain the key approximation equations (24) through (30), which yield the variability parameters for the internal flows. The approximations are all based on the basic methods in Whitt:[14] the asymptotic method and the stationary-interval method. We consider the basic operations—superposition, splitting, and departure—in turn, and then their synthesis.

For superposition, the stationary-interval method is nonlinear so it presents difficulties.[14–16,22] Moreover, there appears to be no natural modification that makes it linear. On the other hand, the asymptotic method is linear. By the asymptotic method, the superposition squared coefficient of variation c_A^2 as a function of component squared coefficients of variation c_i^2 and the rates λ_i is just the convex combination

$$c_A^2 = \sum_i \left(\lambda_i \Big/ \sum_k \lambda_k\right) c_i^2. \tag{31}$$

However, neither the asymptotic method nor the stationary-interval method alone works very well over a wide range of cases, e.g., see Section III of Whitt.[40] Albin[15,16] found that considerable improvement could be obtained by using a refined composite procedure, which is based on a convex combination of c_A^2 for the asymptotic method and c_{SI}^2 for the stationary-interval method. Her hybrid c_H^2 is of the form

$$c_H^2 = wc_A^2 + (1 - w)c_{SI}^2. \tag{32}$$

Unfortunately, since c_{SI}^2 is nonlinear, so is c_H^2. However, Albin found that a convex combination of c_A^2 and the exponential c^2 of 1 worked almost as well, having 4-percent average absolute error as opposed to 3 percent. Hence, we use such a hybrid procedure, namely,

$$\begin{aligned} c_H^2 &= wc_A^2 + (1 - w) \\ &= w \sum_i \left(\lambda_i \Big/ \sum_k \lambda_k \right) c_i^2 + 1 - w, \end{aligned} \tag{33}$$

where w is a function of ρ and the rates. Extensive simulation prompted Albin to suggest the weighting function

$$w = [1 + 2.1(1 - \rho)^{1.8}\nu]^{-1}, \tag{34}$$

where

$$\nu = \left[\sum_i \left(\lambda_i \Big/ \sum_k \lambda_k \right)^2 \right]^{-1}. \tag{35}$$

Note that if there are k component processes with equal rates then $\nu = k$. The parameter ν can be thought of as the number of component streams, with it being an equivalent number if the rates are unequal.

However, the weighting function (58) fails to satisfy an important consistency condition: We should have $w = 1$ when $\nu = 1$; if there is a single arrival process, the superposition operation should leave the c^2 parameter unchanged. Moreover, new theoretical results[39] indicate that the exponent of $(1 - \rho)$ in (34) should be 2. Hence, we use formula (33) based on the weight function w in (29).

4.4 Splitting

No approximation is needed for splitting because a renewal process that is split by independent probabilities (Markovian routing) is again a renewal process. However, approximation is of course indirectly associated with this step because the real process being split is typically not a renewal process and the splitting is often not according to Markovian routing.

Since a renewal process split according to Markovian routing is a renewal process, the asymptotic method and the stationary-interval method coincide. If a stream with a parameter c^2 is split into k streams, with each being selected independently according to probabilities p_i, $i = 1, 2, \cdots, k$, then the ith process obtained from the splitting has squared coefficient of variation c_i^2 given by

$$c_i^2 = p_i c^2 + 1 - p_i, \tag{36}$$

which is clearly linear. Formula (36) is easy to obtain because the

renewal-interval distribution in the split stream is a geometrically distributed random sum of the original renewal intervals.

4.5 Departures

For the stationary-interval method with single-server nodes, we apply Marshall's formula for the squared coefficient of variation of an interdeparture time, say c_d^2, in a GI/G/1 queue:[43,44]

$$c_d^2 = c_a^2 + 2\rho^2 c_s^2 - 2\rho(1-\rho)\mu EW, \tag{37}$$

where EW is the mean waiting time. Since EW appears in (37), the congestion at the node affects the variability of the departure process. A stationary-interval method approximation for c_d^2 is obtained by inserting an approximation for EW in a GI/G/1 queue. Our analysis[33–37] suggests that it suffices to use the linear approximation (2) with g set equal to one. When this is combined with (37), we obtain the simple formula

$$c_d^2 = \rho^2 c_s^2 + (1-\rho^2)c_a^2. \tag{38}$$

A simple extension of (38) for GI/G/m queues that is being used in the current version of QNA is

$$c_d^2 = 1 + (1-\rho^2)(c_a^2 - 1) + \frac{\rho^2}{\sqrt{m}}(c_s^2 - 1). \tag{39}$$

Note that (39) agrees with (38) when $m = 1$ and (39) yields $c_d^2 = 1$ as it should for M/M/m and M/G/∞ systems for which the stationary departure process is known to be Poisson. The third term in (39) approaches 0 as m increases, reflecting the way multiple servers tend to act as a superposition operation. A basis for further refinements of (39) is the asymptotic analysis of departure processes in Whitt.[45] This asymptotic analysis shows that in some cases the variability of the departure process depends on the arrival and service processes in a more complicated way.

As with superposition, the asymptotic method yields a more elementary approximation than the stationary-interval method. In fact, the asymptotic-method approximation for the departure process is just the arrival process itself, i.e., the asymptotic-method approximation for c_d^2 is just c_a^2.[44] The number of departures in a long interval of time is just the number of arrivals minus the number in queue, and the number in queue fluctuates around its steady-state distribution, whereas the number of arrivals goes to infinity.

It remains to combine the basic methods to form a refined hybrid procedure. However, limited experience indicates that this refinement is not as critical as for superposition. The stationary-interval method

alone seems to perform better for departure processes than for superposition processes.[44]

The most appropriate view for the departure process—the stationary-interval method or the asymptotic method—depends on the traffic intensities at the next nodes where the departures are arrivals. As the traffic intensity of the next node increases, the asymptotic-method approximation for the departure process becomes more relevant. For example, consider the case of two queues in series with parameters λ_{01}, c_{01}^2, μ_1, c_{s1}^2, μ_2, and c_{s2}^2. If $\mu_2 \to \lambda$ while μ_1 remains unchanged, then $\rho_2 \to 1$ and the second queue is in heavy traffic. Under such heavy traffic conditions, it has been shown[26] that the congestion measures at the second node are asymptotically the same as if the first facility were removed, i.e., as if the arrival process to the second node were just the arrival process to the first node. More generally, for any arrival process it has been proved that the asymptotic method is an asymptotically correct approximation for a queue in heavy traffic.[26]

Hence, it is natural to tune the departure approximation by using the traffic intensities in the following nodes. Since the departure process typically will be split and sent to different nodes with different traffic intensities, it is appropriate to do the tuning after splitting. Let c_{di}^2 be the departure c^2 at node i. Then

$$c_{ij}^2 = q_{ij}c_{di}^2 + 1 - q_{ij} \tag{40}$$

is the c^2 for the portion of the departures going to node j. We let c_{ij}^2 be a weighted combination of the approximations obtained by the asymptotic method and the stationary-interval method [using (39)]:

$$\begin{aligned} c_{ij}^2 = {} & \nu_{ij}(q_{ij}c_{ai}^2 + 1 - q_{ij}) \\ & + (1 - \nu_{ij})[q_{ij}[1 + (1 - \rho_i^2)(c_{ai}^2 - 1) \\ & + \rho_i^2 m_i^{-0.5}(c_{si}^2 - 1)] + 1 - q_{ij}], \end{aligned} \tag{41}$$

where ν_{ij} is chosen to satisfy $0 \leq \nu_{ij} \leq 1$ and be increasing in ρ_j with $\nu_{ij} \to 1$ as $\rho_j \to 1$. However, we have not yet found that positive ν_{ij} helps,[44] so the current version of the QNA uses (28).

From (38) it is clear that the departure process variability, as depicted by QNA, is an appropriate weighted average of the arrival-process variability and the service-time variability. Hence, when the service time is deterministic, so that $c_{sj}^2 = 0$, the departure process is less variable than the arrival process. However, the actual reduction of variability in a network caused by deterministic service times often is not as great as predicted by (38) or (39). Hence, we have replaced (39) by

$$c_d^2 = 1 + (1 - \rho^2)(c_a^2 - 1) + \frac{\rho^2}{\sqrt{m}}(\max\{c_s^2, 0.2\} - 1). \quad (42)$$

After making this change, we get (25) through (28).

4.6 Customer creation or combination

We treat customer creation or combination as a modification of the departure process. When there is customer creation at node i, we replace each departure by a batch of size γ_i. When there is combination at node i, we replace each interdeparture interval by the sum of γ_i^{-1} such intervals. These make more sense for integer values, but we do not require it. Hence, as described in Section 2.2, the departure rate from node i is $\gamma_i\lambda_i$ when the arrival rate is λ_i. We use the asymptotic method to obtain the variability parameter. Since the number of departures from node i in a large time interval is γ_i times the number of arrivals, the asymptotic-method approximation of the variability parameter for customer creation or combination is just to multiply c_{di}^2 by γ_i. (By the asymptotic method, $c^2 = \lim_{t\to\infty} \mathrm{Var}\, N(t)/EN(t)$; see Section 2 of Whitt.[14]) This is done before splitting.

4.7 Synthesis

We obtain the basic system of equations (24) through (30) by combining Sections 4.3 through 4.6 as follows:

$$\begin{aligned} c_{aj}^2 &= 1 - w_j + w_j \sum_{i=0}^{n} p_{ij} c_{ij}^2 \\ &= 1 - w_j + w_j \Bigg[p_{0j} c_{0j}^2 + \sum_{i=1}^{n} p_{ij} (\nu_{ij}[q_{ij}\gamma_i c_{ai}^2 + (1 - q_{ij})] \\ &\quad + (1 - \nu_{ij})\{\gamma_i q_{ij}[1 + (1 - \rho_i^2)(c_{ai}^2 - 1) \\ &\quad + \rho_i^2 m_i^{-0.5}(\max\{c_{si}^2, 0.2\} - 1)] + 1 - q_{ij}\}) \Bigg]. \end{aligned} \quad (43)$$

The first line is based on superposition, Section 4.3, and the second line is based on departure, splitting and customer creation, Sections 4.4 through 4.6.

V. CONGESTION AT THE NODES

Having calculated the rate and variability parameters associated with each internal arrival process, we are ready to calculate the approximate congestion measures for each node. At this point we have decomposed the network into separate service facilities that are analyzed in isolation. Each facility is a standard GI/G/m queue partially characterized by five parameters: the number of servers plus the first

two moments of the interarrival time and the first two moments of the service time. Instead of the moments we use the arrival rate λ, the mean service time τ, and the squared coefficients of variation c_a^2 and c_s^2. Since we are focusing on a single node, we omit the subscript indexing the node throughout this section.

There are many procedures that could be applied at this point. We could fit complete distributions to the parameters,[14] and then apply any existing algorithm for solving a GI/G/m queue or a special case. Among the attractive options are procedures for analyzing the GI/G/1 queue,[46] the M/PH/m queue with phase-type service-time distributions,[47-52] the GI/H_k/m queue with hyperexponential service-time distributions[53,54] and the GI/E_k/m queue with Erlang service-time distributions.[55] Also available are approximations based on heavy-traffic and light-traffic limiting behavior.[56,57] The actual procedures used in this version of QNA, however, are quite elementary. Our study of the GI/G/1 queue[33-37] indicates that these elementary procedures are consistent with the limited information available. Since the arrival process is usually not a renewal process, and since only two moments are known for each distribution, there is little to be gained from more elaborate procedures. In fact, a user of QNA should be cautioned not to rely too heavily on detailed descriptions such as the tail of the waiting-time distribution. Such detailed descriptions may prove to be reasonably accurate, but they should certainly be checked by simulation.

We now describe the congestion measures provided by QNA. In Section 5.1 we treat the single-server node and in Section 5.2 we treat the multiserver node.

5.1 The GI/G/1 queue

We begin with the steady-state waiting time (before beginning service), here denoted by W. The main congestion measure is the mean EW, but we also generate an entire probability distribution for W. First, the approximation formula for the mean is as in (2):

$$EW = \tau\rho(c_a^2 + c_s^2)g/2(1 - \rho), \tag{44}$$

where $g \equiv g(\rho, c_a^2, c_s^2)$ is defined as

$$g(\rho, c_a^2, c_s^2) = \begin{cases} \exp\left[-\dfrac{2(1-\rho)}{3\rho}\dfrac{(1-c_a^2)^2}{c_a^2 + c_s^2}\right], & c_a^2 < 1 \\ 1, & c_a^2 \geq 1. \end{cases} \tag{45}$$

When $c_a^2 < 1$, (44) is the Kraemer and Langenbach-Belz approximation,[58] which is known to perform well.[33-37,59] When $c_a^2 > 1$, the original Kraemer and Langenbach-Belz refinement does not seem to help, so

it is not used. Note that (44) is exact for the M/G/1 queue having $c_a^2 = 1$.

Let the number of customers in the facility, including the one in service, be denoted by N. The probability that the server is busy at an arbitrary time, $P(N > 0)$, and the mean EN can be obtained from Little's formula (see Section 11.3 of Heyman and Sobel[41]):

$$P(N > 0) = \rho \tag{46}$$

and

$$EN = \rho + \lambda EW. \tag{47}$$

Formula (46) is exact even for stationary nonrenewal arrival processes and (47) is exact given EW.

For the probability of delay, $P(W > 0)$, denoted here by σ, we use the Kraemer and Langenbach-Belz approximation:[58]

$$\sigma \equiv P(W > 0) = \rho + (c_a^2 - 1)\rho(1 - \rho)h(\rho, c_a^2, c_s^2), \tag{48}$$

where

$$h(\rho, c_a^2, c_s^2) = \begin{cases} \dfrac{1 + c_a^2 + \rho c_s^2}{1 + \rho(c_s^2 - 1) + \rho^2(4c_a^2 + c_s^2)}, & c_a^2 \leqslant 1 \\ \dfrac{4\rho}{c_a^2 + \rho^2(4c_a^2 + c_s^2)}, & c_a^2 \geqslant 1. \end{cases} \tag{49}$$

Formula (48) also yields the correct value for M/G/1 systems, namely, ρ. Additional supporting evidence for (48) is contained in Whitt.[60]

We next focus on the conditional delay given that the server is busy, denoted by D. Obviously, $ED = EW/\sigma$. We first give an approximation formula for the squared coefficient of variation of D, c_D^2. This formula is the exact formula for the M/G/1 queue, with the service-time distribution being H_2^b when $c_s^2 \geqslant 1$ and E_k when $c_s^2 = k^{-1}$, where H_2^b is the hyperexponential distribution with balanced means and E_k is the Erlang distribution (see p. 256 of Cohen[61] and Section 3 of Whitt[14]). The idea underlying this approximation is that the conditional delay D in a GI/G/1 queue (rather than the total delay W) depends more on the service-time distribution than on the interarrival-time distribution. Hence, the M/G/1 formula for c_D^2 is used as an approximation for all GI/G/1 systems. The M/G/1 formula for c_D^2 is:

$$c_D^2 = 2\rho - 1 + 4(1 - \rho)d_s^3/3(c_s^2 + 1)^2, \tag{50}$$

where $d_s^3 = E(\nu^3)/(E\nu)^3$ with ν being a service-time random variable. Even $E(\nu^3)$ is available, it can be used in (50), but since $E(\nu^3)$ is not available with two parameters, we use approximations for d_s^3. The approximations are based on the H_2^b and E^k distributions.

Case 1: When $c_s^2 \geq 1$,

$$d_s^3 = 3c_s^2(1 + c_s^2), \tag{51}$$

which comes from the H_2^b formulas:

$$d_s^3 = \frac{3}{4}\left[\frac{1}{q^2} + \frac{1}{(1-q)^2}\right]$$

and

$$q = [1 + \sqrt{4(c_s^2 - 1)/(c_s^2 + 1)}]/2.$$

Case 2: When $c_s^2 < 1$,

$$d_s^3 = (2c_s^2 + 1)(c_s^2 + 1). \tag{52}$$

We obtain formula (52) by considering an Erlang E_k variable, which can be represented as the sum of k iid exponential random variables X_i with mean τ/k, where τ is the mean of the E_k variable. In this case

$$\begin{aligned} E(X_1 + \cdots + X_k)^3 &= kE(X_1^3) + 3k(k-1)E(X_1^2)E(X_1) \\ &\quad + k(k-1)(k-2)(EX_1)^3 \\ &= \left(\frac{\tau}{k}\right)^3 [6k + 6k(k-1) + k(k-1)(k-2)] \end{aligned}$$

so that

$$d_s^3 = \frac{(k+2)(k+1)k}{k^3} = \left(1 + \frac{2}{k}\right)\left(1 + \frac{1}{k}\right),$$

which reduces to (52) because $c_s^2 = k^{-1}$ for an E_k variable. Note that (51) and (52) agree at the boundary when $c_s^2 = 1$.

From (44), (48), and (50) through (52), we immediately obtain formulas for $\mathrm{Var}(D)$ and ED^2:

$$\begin{aligned} \mathrm{Var}(D) &= (ED)^2 c_D^2 = (EW)^2 c_D^2/\sigma^2 \\ E(D^2) &= \mathrm{Var}(D) + (ED)^2. \end{aligned} \tag{53}$$

From D we then obtain second-moment characteristics for W:

$$c_W^2 = \frac{E(W^2)}{(EW)^2} - 1 = \frac{\sigma E(D^2)}{(\sigma ED)^2} - 1 = \frac{c_D^2 + 1 - \sigma}{\sigma},$$

$$\mathrm{Var}(W) = (EW)^2 c_W^2 \quad \text{and} \quad E(W^2) = \mathrm{Var}(W) + (EW)^2. \tag{54}$$

We now indicate how QNA calculates an approximate probability distribution for W. The distribution has an atom at zero as given in (48) and a density above zero. The density is chosen so that W and D

have the first two moments already determined for them. (This is the general rule, but it is not quite followed in Cases 2 and 4 below.)

Case 1: $c_D^2 > 1.01$. Let D have the H_2^b density (hyperexponential with balanced means)

$$f_D(x) = p\gamma_1 e^{-\gamma_1 x} + (1-p)\gamma_2 e^{-\gamma_2 x}, \qquad x \geq 0, \tag{55}$$

where

$$p = [1 + \sqrt{(c_D^2 - 1)/(c_D^2 + 1)}]/2,$$

$$\gamma_1 = 2p/ED \quad \text{and} \quad \gamma_2 = 2(1-p)/ED. \tag{56}$$

Case 2: $0.99 \leq c_D^2 \leq 1.01$. Let D have the exponential density with mean ED.

Case 3: $0.501 \leq c_D^2 < 0.99$. Let the distribution of D be the convolution of two exponential distributions with parameters γ_1 and γ_2 ($\gamma_1 > \gamma_2$), i.e., let D have density

$$f_D(x) = \left(\frac{\gamma_1\gamma_2}{\gamma_1 - \gamma_2}\right)(e^{-\gamma_2 x} - e^{-\gamma_1 x}), \qquad x \geq 0, \tag{57}$$

where

$$\gamma_2^{-1} = \frac{ED + \sqrt{2\ \mathrm{Var}(D) - (ED)^2}}{2}$$

and

$$\gamma_1^{-1} = ED - \gamma_2^{-1}. \tag{58}$$

The associated tail probabilities are

$$P(D > x) = (\gamma_1 e^{-\gamma_2 x} - \gamma_2 e^{-\gamma_1 x})/(\gamma_1 - \gamma_2). \tag{59}$$

Case 4: $c_D^2 < 0.501$. Let D have an E_2 (Erlang) distribution with mean ED, which has $c^2 = 0.5$. Its density is

$$f_D(x) = \gamma^2 x e^{-\gamma x}, \qquad x \geq 0, \tag{60}$$

where $\gamma = 2/ED$. The associated tail probabilities are

$$P(D > x) = e^{-\gamma x}(1 + \gamma x), \qquad x \geq 0. \tag{61}$$

For deterministic service times, $d_s^3 = 1$, so that the smallest possible c_D^2 via (50) is $(1 + 2\rho)/3$. Hence, Case 4 above will not occur often.

Finally, we come to the second moment and variance of N, the number in system. For the M/G/1 queue, it is not difficult to compute $E(N^2)$. Since the steady-state number in the facility is equal to the number of arrivals during a customer's time in the facility, it is easy to compute the moments of N from the moments of W; for example,

$$E(N^2) = \lambda(EW + E\nu) + \lambda^2[E(W^2) + 2EWE\nu + E(\nu^2)]$$
$$= \lambda EW + \rho + \lambda^2 E(W^2) + 2\lambda\rho EW + \rho^2(c_s^2 + 1), \quad (62)$$

and

$$\mathrm{Var}(N) = \lambda EW + \rho + \rho^2 c_s^2 + \lambda^2 \mathrm{Var}(W). \quad (63)$$

We now modify the M/G/1 formulas (62) and (63) for the GI/G/1 queue. Let c_N^2 be defined by

$$c_N^2 = Y_1 Y_2 / Y_3, \quad (64)$$

where Y_1 is the M/G/1 value of $\mathrm{Var}(N)$ in (63) using (44) for EW and (54) and $\mathrm{Var}(W)$,

$$Y_2 = (1 - \rho + \sigma)/\max\{(1 - \sigma + \rho), 0.000001\}$$
$$Y_3 = \max\{(\rho + \lambda EW)^2, 0.000001\}, \quad (65)$$

and σ is the probability of delay in (48). The maximum is used in (65) to avoid dividing by zero. For the M/G/1 queue, $Y_2 = 1$; for GI/M/1 queues, Y_2 in (65) provides just the right correction, so that (64) is exact given the true value of σ, EW and $\mathrm{Var}(W)$. The correction Y_2 in (65) makes (64) too small for D/D/1 queues by a factor of $(1 + \rho)^{-1}$, but (64) is asymptotically correct in heavy traffic: $c_N^2 \to 1$ as $\rho \to 1$ if either $c_a^2 > 0$ or $c_s^2 > 0$.

From (47) and (64) we immediately obtain

$$\mathrm{Var}(N) = (EN)^2 c_N^2$$

and

$$E(N^2) = \mathrm{Var}(N) + (EN)^2. \quad (66)$$

When there is immediate feedback at the node and it is eliminated, adjustments are necessary in the formulas of this section, as indicated in Section III.

5.2 The GI/G/m queue

The first congestion measures for multiserver nodes provided by QNA are exact. Even for nonrenewal stationary-arrival processes, the expected number of busy servers is just the offered load [see p. 400 of Heyman and Sobel[41] and (4.2.3) of Franken et al.[42]]:

$$E \min\{N, m\} = \alpha = \lambda\tau \quad (67)$$

and the traffic intensity or utilization is

$$\rho = \alpha/m. \quad (68)$$

By Little's formula, as in (47),

$$EN = \alpha + \lambda EW. \quad (69)$$

QNA currently provides only a few simple approximate congestion measures for multiserver nodes. These are obtained by modifying the exact formulas for the M/M/m model.[18] Let characteristics such as $EW(c_a^2, c_s^2, m)$ represent the characteristic as a function of the parameters c_a^2, c_s^2, and m, and let characteristics such as $EW(\mathrm{M/M/m})$ be the exact value for M/M/m system. A simple approximation for EW based on heavy-traffic limit theorems[26,56,62,63] is:

$$EW(c_a^2, c_s^2, m) = \left(\frac{c_a^2 + c_s^2}{2}\right) EW(M/M/m). \tag{70}$$

Formula (70) has frequently been used for M/G/m queues and is known to perform quite well in that case.[64–67] By virtue of heavy-traffic limit theorems, we know that (70) is also asymptotically correct for GI/G/m systems as $\rho \to 1$ for fixed m. Limited additional study indicates that (70) is also reasonable for moderate values of ρ when $c_a^2 \geq 0.9$ and $c_s^2 \geq 0.9$, or when $c_a^2 \leq 1.1$ and $c_s^2 \leq 1.1$. The actual value may be significantly smaller (larger) when $c_a^2 < 0.9$ and $c_s^2 > 1.1$ ($c_a^2 > 1.1$ and $c_s^2 < 0.9$).

The simple approximation (70) is also supplemented by simple approximations for the second moments of W and N. They are obtained from:

$$c_W^2(c_a^2, c_s^2, m) = c_W^2(M/M/m)$$

and

$$c_N^2(c_a^2, c_s^2, m) = c_N^2(M/M/m). \tag{71}$$

Related second-moment characteristics are computed as in (54) and (66).

More detailed and sophisticated approximations for multiserver nodes are being studied. As we indicated before, a variety of methods and algorithms can be applied given the parameters of the arrival process.[47–57]

VI. TOTAL NETWORK PERFORMANCE MEASURES

In this section we describe the approximate congestion measures calculated by QNA for the network as a whole. In Section 6.1 we discuss congestion measures representing the system view, e.g., throughput and number of customers in the network; in Sections 6.2 and 6.3 we discuss congestion measures representing the customer view, e.g., number of nodes visited and response times. In fact, there are actually two different customer views. In Section 6.2 we discuss the view of an arbitrary, typical, or aggregate customer; in Section 6.3 we discuss the view of a particular customer with a specified route through the network.

6.1 System congestion measures

A basic total network performance measure is the *throughput*, which we define as the *total external arrival rate* λ_0,

$$\lambda_0 = \lambda_{01} + \cdots + \lambda_{0n}. \tag{72}$$

When no customers are created at the nodes, the total external arrival rate equals the total departure rate from the network, so that there is little ambiguity about what we mean by throughput. However, when customers are created or combined at the nodes, as in Section 2.2, there is more than one possible interpretation. We might be interested in the rate at which arrivals are processed, i.e., (72). For example, the customers created at the nodes might be regarded only as extra work that must be done to serve the arrivals. On the other hand, we might be interested in the rate at which customers leave the network or in the rate of service completions. The departure *rate from the network* is

$$d = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \lambda_i \gamma_i \left(1 - \sum_{j=1}^{n} q_{ij}\right) \tag{73}$$

and the total *rate of service completions* is

$$s = \sum_{i=1}^{n} s_i = \sum_{i=1}^{n} \lambda_i \gamma_i. \tag{74}$$

A description of the overall congestion is provided by the mean and variance of the number N of customers in the entire network. In general,

$$EN = EN_1 + \cdots + EN_n \tag{75}$$

and, as an approximation based on assuming that the nodes are independent, we have

$$\mathrm{Var}(N) = \mathrm{Var}(N_1) + \cdots + \mathrm{Var}(N_n). \tag{76}$$

Formula (76) is valid for the Markovian models as a consequence of the product-form solution, but is an approximation in general.

6.2 The experience of an aggregate customer

When we turn to the congestion experienced by individual customers, there are two very different approaches. The first approach keeps strict adherence to the model assumptions with the standard input in Section 2.1, and is based on interpreting the routing matrix as independent probabilities (Markovian routing). This means that each time any customer completes service at node i, that customer proceeds to node j with probability q_{ij}, independent of the current state

and history of the network. If the network is cyclic, this means that every customer has positive probability of visting some nodes more than once. This is the perspective of an aggregate customer. It might be that no individual customer actually ever visits the same node more than once.

If the aggregate view is desired, then the customer experience can be described by employing the basic theory of absorbing Markov chains as in Chapter III of Kemeny and Snell.[68] We can regard the external node as a single absorbing state to which all customers go when they leave the network or we can have more absorbing states, to distinguish between network departures from different nodes or different subsets of nodes. For this interpretation, the routing matrix Q is the transient subchain associated with the absorbing Markov chain and the inverse $(I - Q)^{-1}$ in (19) with $\Gamma = I$ is the *fundamental matrix* of the absorbing chain (see p. 45 of Kemeny and Snell[68]). Solving the traffic-rate equations is tantamount to solving for this fundamental matrix.

From the fundamental matrix it is easy to calculate the moments of the number n_{ij} of visits to any state j starting from any state i (on an external arrival process). For example, En_{ij} is just the (i, j)-th entry of $(I - Q)^{-1}$. It is also easy to calculate the probability of absorption into each of the absorbing states starting from any initial distribution. These various congestion measures are easily obtained working with $n \times n$ matrices.[68]

Suppose that we focus on an arbitrary, typical, or aggregate customer arriving on an external arrival process. Then that customer enters node i with probability λ_{0i}/λ_0, where λ_0 is defined in (72) and the expected number of visits to node i for each customer is

$$EV_i = \lambda_i/\lambda_0. \tag{77}$$

(We have used the fundamental matrix to get λ_i.) The mean of the time, T_i, that an arbitrary customer spends in node i during his or her time in the network is thus

$$ET_i = (EV_i)(\tau_i + EW_i) \tag{78}$$

and the expected total sojourn time (time spent in the network from first arrival to final departure) for an arbitrary customer is thus

$$ET = \sum_{i=1}^{n} ET_i = \sum_{i=1}^{n} EV_i(\tau_i + EW_i). \tag{79}$$

The variance of T_i is thus

$$\mathrm{Var}(T_i) = EV_i(\mathrm{Var}(W_i) + \tau_i^2 c_{si}^2) + \mathrm{Var}(V_i)(EW_i + \tau_i)^2. \tag{80}$$

The term $\mathrm{Var}(V_i)$ in (80) as well as EV_i is easily obtained from the fundamental matrix. In particular, $\mathrm{Var}(V_i) = EV_i^2 - (EV_i)^2$ and

$$EV_i^2 = \sum_{j=1}^{n} (\lambda_{0j}/\lambda_0)[F(2F_{dg} - 1)]_{ji}, \tag{81}$$

where F is the fundamental matrix $(I - Q)^{-1}$, F_{dg} is the $n \times n$ matrix with all off-diagonal entries 0 and diagonal entries the same as F.

To obtain an approximation for the variance of the total sojourn time in the network, we assume that the sojourn times at the different nodes are conditionally independent, given any particular routing. (This is not valid even for all acyclic networks of M/M/1 nodes,[69] but is often approximately true.[7,70]) In particular, for a customer entering some specified node and making V_j visits to node j, $1 \leqslant j \leqslant n$, before eventually leaving the network,

$$T = \left(\sum_{j=1}^{n} \sum_{k=1}^{V_j} T_{kj} \right), \tag{82}$$

where T_{kj} is the sojourn time for the kth visit to node j. Our approximation assumption is that the variables T_{kj} are mutually independent given the vector $(V_1, V_2, \cdots, V_n)$.

Hence,

$$\begin{aligned} E(T^2) &= \sum_{i=1}^{n} E\left(\sum_{k=1}^{V_i} T_{ki} \right)^2 \\ &\quad + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} E\left(\sum_{k=1}^{V_i} T_{ki} \sum_{\ell=1}^{V_j} T_{\ell j} \right) \\ &= \sum_{i=1}^{n} \{EV_i E(T_{1i}^2) + E[V_i(V_i - 1)]E(T_{1i})^2\} \\ &\quad + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} E(T_{1i})E(T_{1j})E(V_iV_j) \end{aligned} \tag{83}$$

so that

$$\mathrm{Var}(T) = \sum_{i=1}^{n} \mathrm{Var}(T_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} E(T_{1i})E(T_{1j})\mathrm{Cov}(V_i, V_j). \tag{84}$$

However, the current version of QNA ignores the covariance terms in (84) in the calculation of $\mathrm{Var}(T)$.

6.3 The experience of a particular customer

Another approach is to decouple the macroscopic and microscopic interpretations. This view is common in statistical mechanics. The total network may exhibit statistical regularity not evidenced in any single particle (customer). In this view, we think of the total system evolving as if customers were routed according to independent probabilities, even though individual customers may have very different

routing probabilities, perhaps nonrandom routing or acyclic routing. For example, we may consider the cyclic network entirely appropriate for the macroscopic view even though no individual customer ever visits any node more than once. In order for this view to be realistic, each individual customer should have a relatively negligible effect on the total network.

The procedure here is to solve for the equilibrium or macroscopic behavior of the network first and then afterwards consider particular customers. The particular customers will have their own routes through the network and perhaps their own service times at the nodes along the way. There are two cases, depending on whether the input is by classes and routes, as in Section 2.3 or the standard input as in Section 2.1.

6.3.1 Input by classes and routes

First, suppose that we are using the input by classes and routes in Section 2.3. Then the particular customers correspond to the customer classes specified in the input. Hence, each customer has a deterministic route through the network and possibly special service times at the nodes on the route. In this case, as described in Section 2.3, QNA first converts the input by classes and routes into the standard input in Section 2.1. Then QNA solves for the equilibrium behavior. Finally, congestion measures are calculated for the different classes under the assumption that they follow their originally specified special routes and that upon arrival at the nodes on the route they see independent versions of the equilibrium state of the network. Hence, in the notation of Section 2.3, for a customer in class k, the expected total service time is

$$\sum_{j=1}^{n_k} \tau_{kj}, \tag{85}$$

the expected total waiting time is

$$\sum_{j=1}^{n_k} E(W_{n_{kj}}), \tag{86}$$

and the expected total sojourn time or response time is the sum of (85) and (86). Similarly, for a customer in class k the variance of the total service time is

$$\sum_{j=1}^{n_k} \tau_{kj}^2 c_{skj}^2, \tag{87}$$

the variance of the total waiting time is

$$\sum_{j=1}^{n_k} \mathrm{Var}(W_{n_{kj}}), \tag{88}$$

and the variance of the total sojourn time is the sum of (87) and (88).

6.3.2 The standard input

With the standard input in Section 2.1, the user must specify the particular customers to be analyzed. In this case, the user specifies classes with routes and possibly service times (rate and variability parameters), but these data are not used in calculating the equilibrium behavior. The decoupling principle is used with greater force here; there need not be any consistency between the microscopic and macroscopic views: This additional input does not affect the equilibrium behavior of the total network.

In the current version of QNA the individual customer routes are deterministic, so that the additional input required is just as in Section 2.3 and the congestion measures are just as in (85) through (88) in Section 6.3.1. However, it is possible to modify QNA to allow random routes. Then the additional input would be just as in Section 2.1; for each class it would consist of a routing matrix plus parameters for the arrivals process and service times.

VII. ACKNOWLEDGMENTS

The software package QNA was written by Anne Seery. She used a subroutine for analyzing the M/M/m queue written by Shlomo Halfin. It has been a pleasure collaborating with Anne Seery on this venture. I also appreciate the help from many other colleagues and the continued support of my management: W. A. Cornell, C. S. Dawson, J. C. Lawson, C. J. McCallum, Jr., M. Segal, R. E. Thomas, and E. Wolman.

REFERENCES

1. L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*, New York: John Wiley and Sons, 1976.
2. M. Schwartz, *Computer-Communications Network Design and Analysis*, Englewood Cliffs: Pentice-Hall, 1977.
3. F. P. Kelly, *Reversibility and Stochastic Networks*, New York: John Wiley, 1979.
4. E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems*, New York: Academic Press, 1980.
5. C. H. Sauer and K. M. Chandy, *Computer Systems Performance Modeling*, Englewood Cliffs: Prentice-Hall, 1981.
6. J. G. Shanthikumar and J. A. Buzacott, "Open Queueing Network Models of Dynamic Job Shops," Int. J. Prod. Res., *19*, No. 3 (1981), pp. 255–66.
7. R. L. Disney, *Queueing Networks and Applications*, Baltimore: The John Hopkins University Lectures, 1982; to be published by the Johns Hopkins University Press.
8. "User's Guide for BEST/1," BGS Systems, Inc., Waltham, Massachusetts, 1980.
9. "User's Manual for CADS," Austin, TX: Information Research Associates, 1978.
10. J. McKenna, D. Mitra, and K. G. Ramakrishnan, "A Class of Closed Markovian Queueing Networks: Integral Representations, Asymptotic Expansions, and Generalizations," B.S.T.J., *60*, No. 5 (May–June 1981), pp. 599–641.

11. J. McKenna and D. Mitra, "Intergral Representations and Asymptotic Expansions for Closed Markovian Queueing Networks: Normal Usage," B.S.T.J., *61*, No. 5 (May–June 1982), pp. 661–83.
12. K. G. Ramakrishnan and D. Mitra, "An Overview of PANACEA, A Software Package for Analyzing Markovian Queueing Networks," B.S.T.J., *61*, No. 10, Part 1 (December 1982), pp. 2849–72.
13. H. Heffes, "Moment Formulae for a Class of Mixed Multi-Job-Type Queueing Networks," B.S.T.J., *61*, No. 5 (May–June 1982), pp. 709–45.
14. W. Whitt, "Approximating a Point Process by a Renewal Process, I: Two Basic Methods," Oper. Res., *30*, No. 1 (January–February 1982), pp. 125–47.
15. S. L. Albin, *Approximating Queues with Superposition Arrival Processes*, Ph.D dissertation, Department of Industrial Engineering and Operations Research, Columbia University, 1981.
16. S. L. Albin, "Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues," Department of Industrial Engineering, Rutgers University, 1982.
17. R. I. Wilkinson, "Theories for Toll Traffic Engineering in the U.S.A.," B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
18. R. B. Cooper, *Introduction to Queueing Theory*, Second Edition, New York: North Holland, 1981.
19. A. Kuczura, "The Interrupted Poisson Process as an Overflow Process," B.S.T.J., *52*, No. 3 (March 1973), pp. 437–48.
20. J. H. Rath and D. D. Sheng, "Approximations for Overflows from Queues with a Finite Waiting Room," Oper. Res., *27*, No. 6 (November–December 1979), pp. 1208–16.
21. M. Reiser and H. Kobayashi, "Accuracy of the Diffusion Approximation for Some Queueing Systems," IBM J. Res. Dev., *18* (March 1974), pp. 110–24.
22. P. J. Kuehn, "Approximate Analysis of General Queueing Networks by Decomposition," IEEE Trans. Commun., *COM-27*, No. 1 (January 1979), pp. 113–26.
23. K. C. Sevcik, A. I. Levy, S. K. Tripathi, and J. L. Zahorjan, "Improving Approximations of Aggregated Queueing Network Subsystems," in *Computer Performance*, K. M. Chandy and M. Reiser (eds.), Amsterdam: North Holland, 1977, pp. 1–22.
24. K. M. Chandy and C. H. Sauer, "Approximate Methods for Analyzing Queueing Network Models of Computing Systems," ACM Computing Surveys, *10*, No. 3 (September 1978), pp. 281–317.
25. S. Katz, "Statistical Performance Analysis of a Switched Communications Network," Fifth Int. Teletraffic Cong., Rockefeller University, New York, 1967, pp. 566–75.
26. D. L. Iglehart and W. Whitt, "Multiple Channel Queues in Heavy Traffic, II: Sequences, Networks and Batches," Adv. Appl. Prob., *2*, No. 2 (Autumn 1970), pp. 355–69.
27. J. M. Harrison and M. I. Reiman, "On the Distribution of Multidimensional Reflected Brownian Motion," SIAM J. Appl. Math., *41*, No. 2 (October 1981), pp. 345–61.
28. M. I. Reiman, "Open Queueing Networks in Heavy Traffic," unpublished work, 1981.
29. M. I. Reiman, "The Heavy Traffic Diffusion Approximation for Sojourn Times in Jackson Networks," *Applied Probability and Computer Science—The Interface*, Volume 2, R. L. Disney and T. J. Ott (eds.), Boston: Birkhauser, 1982, pp. 409–21.
30. J. M. Holtzman, "The Accuracy of the Equivalent Random Method with Renewal Inputs," B.S.T.J., *52*, No. 9 (November 1973), pp. 1673–9.
31. T. Rolski, "Some Inequalities for GI/M/n Queues," Zast. Mat., *13*, No. 1 (1972), pp. 43–7.
32. A. E. Eckberg, Jr., "Sharp Bounds on Laplace-Stieltjes Transforms, with Applications to Various Queueing Problems," Math. Oper. Res., *2*, No. 2 (May 1977), pp. 135–42.
33. W. Whitt, "On Approximations for Queues, I: Extremal Distributions," B.S.T.J., *63*, No. 1, Part 1 (January 1984), to be published.
34. J. G. Klincewicz and W. Whitt, "On Approximations for Queues, II: Shape Constraints," B.S.T.J., *63*, No. 1, Part 1 (January 1984).
35. W. Whitt, "On Approximations for Queues, III: Mixtures of Exponential Distributions," B.S.T.J., *63*, No. 1, Part 1 (January 1984).
36. W. Whitt, "The Marshall and Stoyan Bounds for IMRL/G/1 Queues are Tight," Oper. Res. Letters, *1*, No. 6 (December 1982), pp. 209–13.

37. W. Whitt, "Refining Diffusion Approximations for Queues," Oper. Res. Letters, *1*, No. 5 (November 1982), pp. 165–9.
38. S. L. Albin, "On Poisson Approximations for Superposition Arrival Processes in Queues," Management Sci., *28*, No. 2 (February 1982), 126–37.
39. W. Whitt, "Queues with Superposition Arrival Processes in Heavy Traffic," unpublished work, 1982.
40. W. Whitt, "Performance of the Queueing Network Analyzer," B.S.T.J., this issue.
41. D. P. Heyman and M. J. Sobel, *Stochastic Models in Operations Research*, Vol. I, New York: McGraw-Hill, 1982.
42. P. Franken, D. König, U. Arndt, and V. Schmidt, *Queues and Point Processes*, Berlin: Akademie-Verlag, 1981.
43. K. T. Marshall, "Some Inequalities in Queueing," Oper. Res., *16*, No. 3 (May–June 1968), pp. 651–65.
44. W. Whitt, "Approximations for Departure Processes and Queues in Series," Nav. Res. Log Qtr., to be published.
45. W. Whitt, "Departures from a Queue with Many Busy Servers," Math. Oper. Res., *9* (1984).
46. A. A. Fredericks, "A Class of Approximations for the Waiting Time Distribution in a GI/G/1 Queueing System," B.S.T.J., *61*, No. 3 (March 1982), pp. 295–325.
47. Y. Takahashi and Y. Takami, "A Numerical Method for the Steady-State Probabilities of a GI/G/c Queueing System in a General Class," J. Oper. Res. Soc. Japan, *19* (1976), pp. 147–57.
48. H. C. Tijms, M. H. van Hoorn, and A. Federgruen, "Approximations for the Steady-State Probabilities in the M/G/c Queue," Adv. Appl. Prob., *13*, No. 1 (March 1981), pp. 186–206.
49. H. Groenevelt, M. H. van Hoorn, and H. C. Tijms, "Tables for M/G/c Queueing Systems with Phase-Type Service," Report No. 85, Department of Actuarial Sciences and Econometrics, The Free University, Amsterdam, The Netherlands, 1982.
50. M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models—An Algorithmic Approach*, Baltimore: The Johns Hopkins University Press, 1981.
51. M. F. Neuts, "A Program for Analyzing the M/PH/c Queue," Department of Mathematics, University of Delaware, 1981.
52. P. Hokstad, "Some Numerical Results and Approximations for the Many Server Queue with Nonexponential Service Time," Department of Mathematics, University of Trondheim, Norway, 1981.
53. J. H. A. de Smit, "The Queue GI/M/s with Customers of Different Types or the Queues $GI/H_m/s$," Adv. Appl. Prob., *15*, No. 2 (June 1983), pp. 392–419.
54. J. H. A. de Smit, "A Program for Analyzing the $GI/H_k/s$ Queue," Department of Applied Mathematics, Twente University of Technology, Enschede, The Netherlands, 1982.
55. A. Ishikawa, "On the Equilibrium Solution for the Queueing System: $GI/E_k/m$," T.R.U. Mathematics, *15*, No. 1 (1979), pp. 47–66.
56. S. Halfin and W. Whitt, "Heavy-Traffic Limits for Queues with Many Exponential Servers," Oper. Res., *29*, No. 3 (May–June 1981), pp. 567–88.
57. D. Y. Burman and D. R. Smith," A Light-Traffic Theorem for Multiserver Queues," Math. Oper. Res., *8*, No. 1 (February 1983), pp. 15–25.
58. W. Kraemer and M. Langenbach-Belz, "Approximate Formulae for the Delay in the Queueing System GI/G/1," *Congressbook*, Eighth Int. Teletraffic Cong., Melbourne, 1976, pp. 235-1/8.
59. J. G. Shanthikumar and J. A. Buzacott, "On the Approximations to the Single Server Queue," Int. J. Prod. Res. *18*, No. 6 (1980), pp. 761–73.
60. W. Whitt, "Minimizing Delays in the GI/G/1 Queue," Oper. Res., *32* (1984), to be published.
61. J. W. Cohen, *The Single Server Queue*, Amsterdam: North-Holland, 1969.
62. J. Köllerström, "Heavy Traffic Theory for Queues with Several Servers. I," J. Appl. Prob., *11*, No. 3 (September 1974), pp. 544–52.
63. J. Köllerström, "Heavy Traffic Theory for Queues with Several Servers. II," J. Appl. Prob., *16*, No. 2 (June 1979), pp. 393–401.
64. A. M. Lee and P. A. Longton, "Queueing Processes Associated with Airline Passenger Check-In," Oper. Res. Quart., *10*, No. 1 (March 1959), pp. 56–71.
65. P. Hokstad, "Approximations for the M/G/n Queue," Oper. Res., *26*, No. 3 (May–June 1978), pp. 510–23.
66. S. A. Nozaki and S. M. Ross, "Approximations in Finite Capacity Multi-Server Queues with Poisson Arrivals," J. Appl. Prob., *15* (1978), pp. 826–34.

67. F. S. Hillier and O. S. Yu, *Queueing Tables and Graphs*, New York: North-Holland, 1981.
68. J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, Princeton: Van Nostrand, 1960.
69. B. Simon and R. D. Foley, "Some Results on Sojourn Times in Acyclic Jackson Networks," Management Sci., *25*, No. 10 (October 1979), pp. 1027–34.
70. P. C. Kiessler, "A Simulation Analysis of Sojourn Times in a Jackson Network," Report VTR 8016, Department of Industrial Engineering and Operations Research, Virginia Polytechnic Institute and State University, 1980.

AUTHOR

Ward Whitt, A.B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; Bell Laboratories, 1977—. At Yale University, from 1973–1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At Bell Laboratories he is in the Operations Research Department in the Network Analysis Center.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 9, November 1983
Printed in U.S.A.

Performance of the Queueing Network Analyzer

By W. WHITT*

(Manuscript received March 11, 1983)

This paper describes the performance of the Queueing Network Analyzer (QNA), a software package developed at Bell Laboratories to calculate approximate congestion measures for networks of queues. QNA is compared with simulations and other approximations of several open networks of single-server queues. This paper illustrates how to apply QNA and indicates the quality that can be expected from the approximations. The examples here demonstrate the importance of the variability parameters used in QNA to describe non-Poisson arrival processes and nonexponential service-time distributions. For these examples, QNA performs much better than the standard Markovian algorithm, which does not use variability parameters. The accuracy of the QNA results (e.g., the expected delays) in these examples is satisfactory for engineering purposes.

I. INTRODUCTION AND SUMMARY

This paper is a sequel to Whitt,[1] which described the software package called the Queueing Network Analyzer (QNA). QNA calculates approximate congestion measures for networks of queues. The first version of QNA treats open networks of multiserver queues with the first-come, first-served discipline and no capacity constraints. QNA is designed to treat non-Markovian models: The arrival processes need not be Poisson and the service-time distributions need not be exponential. QNA approximately characterizes other kinds of variability through variability parameters assigned to each arrival process and each service-time distribution. The first step in the algorithm is

* Bell Laboratories.

to solve for the flow rates and the variability parameters of the internal arrival processes. The second step is to compute approximate congestion measures for each queue separately by regarding it as a standard GI/G/m queue in which the renewal arrival process and the service-time distribution are each partially characterized by their first two moments or, equivalently, the rate and variability parameters. The third and final step is to calculate congestion measures for the network as a whole.

This paper describes the performance of QNA by comparing it with simulations and other approximations of networks of queues. Even though QNA can analyze multiserver queues, only single-server queues are considered here. Among the other approximations in each case are the M/M/1 and M/G/1 approximations, which can be obtained from QNA by using default options. The M/M/1 approximation, which is embodied in the Markovian algorithms, is obtained by setting all variability parameters equal to 1. With the M/M/1 approximation, the nodes are treated as independent M/M/1 queues with the correct rates. The M/M/1 approximation yields the exact equilibrium distribution of queue lengths for the Markov model with Poisson external arrival processes, exponential service-time distributions and one customer class. The M/G/1 approximation is obtained by setting the variability parameter of each arrival process equal to 1 and using the specified service-time variability parameter c_s^2; then the expected waiting time at each node is $(1 + c_s^2)/2$ times the M/M/1 value.

The congestion measures we consider in the examples here are the expected waiting time (before beginning service) and the expected sojourn time (waiting time plus service time) at a node or in the entire network. Of course, QNA produces other congestion measures, but we are comparing with previously published simulation results, which are mostly limited to expected waiting times and sojourn times.

We begin in Section II with a single GI/G/1 queue and discuss the implications of previous work on approximations for the GI/G/1 queue.[2-7] In Section III we consider a single queue with a superposition arrival process and compare QNA with simulations by Albin.[8-10] In Section IV we consider a network of eight queues in series analyzed by Fraker,[11] and in Sections V and VI we consider two networks analyzed by Kuehn:[12] Section V treats a tightly coupled two-node network and Section VI treats a nine-node network. In Section VII we treat a five-node network used to model a Bell Laboratories computer system. Finally, in Section VIII we consider a model from Gelenbe and Mitrani[13] for a packet-switched communication network. The examples in Sections VII and VIII have input by classes and routes as in Section 2.3 of Whitt.[1]

These examples indicate the approximation quality that can be

expected in applications of QNA. They also demonstrate the importance of the variability parameters when the external arrival processes are not nearly Poisson or the service-time distributions are not nearly exponential. These examples also illustrate how to apply QNA, e.g., to model superposition arrival processes (Section III), to eliminate almost immediate feedback (Section V), and to conduct sensitivity analyses for the variability (Section VII).

II. A SINGLE GI/G/1 QUEUE

We begin by considering the special network containing a single service facility, in particular, the GI/G/1 queue with service times and interarrival times each partially characterized by their first two moments or, equivalently, by the four parameters τ (the mean service time), c_s^2 (the squared coefficient of variation of the service time), λ (the arrival rate), and c_a^2 (the squared coefficient of variation of the interarrival time).[1] The subscript indexing the node is suppressed since there is only one node.

It is useful to consider this model because it has been extensively studied and is relatively well understood. In many cases we can analytically determine the quality of the approximations for the GI/G/1 queue. Hence, we can get an idea about the quality of the approximations for more general networks. Of course, approximations for a node in a general network might be worse because the internal arrival processes usually are not actually renewal processes. On the other hand, the network operations of superposition and splitting tend to make stochastic point processes more like Poisson processes, so that larger networks may actually be better behaved.

The model specification above determines the approximate congestion measures produced by QNA, but the model specification is not complete since there are many service-time and interarrival-time distributions with the given parameters. The formulas produced by the QNA are approximations for all these systems, so it is natural to ask how the approximate congestion measures compare to the set of all possible values that are consistent with the partial specification. Fortunately, it is often possible to identify the set of all possible values.[2–5] Moreover, it is often possible to locate the more likely values by identifying the set of all possible values under various natural constraints on the distribution. When this cannot be done exactly, it can often be done approximately using bounds.[6]

We now give a brief summary of results evaluating approximations for the expected waiting time or, equivalently (by Little's formula[14]), the expected queue length in the GI/G/1 queue based on the four parameters λ, c_a^2, τ, and c_s^2. First, recall that for the M/G/1 queue, with Poisson arrival process ($c_a^2 = 1$), the expected waiting time

actually depends on the service-time distribution only through the two parameters τ and c_s^2. Nonexponential interarrival-time distributions tend to be more difficult, however. For the GI/M/1 queue with an exponential service-time distribution ($c_s^2 = 1$), the expected waiting time depends on the interarrival-time distribution beyond the parameters λ and c_a^2. Given λ and c_a^2 in the GI/M/1 queue, the maximum relative error (upper bound minus lower bound divided by lower bound) in the mean queue length (number in system including anyone in service) is exactly c_a^2 (see Ref. 2). A similar, but somewhat less concise, result holds for the expected waiting time by virtue of Little's formula. The maximum relative error for the expected sojourn time (waiting time plus service time) is also c_a^2. This result suggests more generally that the reliability of the approximations might decrease when c_a^2 increases, which is consistent with numerical experience.

If we assume that the interarrival-time distribution is not too irregular, then the maximum relative error becomes much less. In the $H_k/M/1$ queue with a hyperexponential interarrival-time distribution (mixture of exponential distributions having $c_a^2 > 1$), the maximum relative error in the mean queue length is $(c_a^2 - 1)/2$ (see Ref. 4). It turns out that the extremal interarrival-time distributions for $H_k/M/1$ queues also are extremal for all interarrival-time distributions that have Increasing Mean Residual Life (IMRL) (also $c_a^2 > 1$) and all service-time distributions,[5] so that the maximum relative error in the mean queue length for IMRL/G/1 queues is $(c_a^2 - 1)/[2 + \rho(c_s^2 - 1)]$.

Other kinds of shape constraints for the GI/M/1 queue have been investigated by means of nonlinear programming.[3] In general, we conclude that if the distributions are not irregular, then the maximum relative error in the GI/G/1 queue might be about $0.05\ c_a^2$, e.g., about 10 percent when $c_a^2 = 2.0$.

From heavy-traffic limit theorems that describe the queue as $\rho \to 1$,[6] where $\rho = \lambda\tau$ is the traffic intensity, we know that asymptotically the queue length and waiting-time distributions depend on the interarrival-time and service-time distributions only through the four parameters λ, c_a^2, τ, and c_s^2. This suggests that more generally the quality of the approximations might improve as ρ increases. This is certainly consistent with experience for the GI/G/1 queue, but not necessarily for more complex networks, e.g., the tightly coupled two-node network here in Section V.

The heavy-traffic limit theorems are closely related to diffusion approximations because diffusion processes emerge as limits in the heavy-traffic limit theorems. We have recently compared various diffusion approximations for the expected waiting time in a GI/G/1 queue to known bounds.[6] We now show how QNA and other related approximations fit into this framework. Table I here compares four

Table I—Bounds and approximations for the expected waiting time, *EW*, in a GI/G/1 queue: Three cases

	Parameter Values		
	$c_a^2 = 0.5$ $c_s^2 = 4.0$ $\rho = 0.7$	$c_a^2 = 2.0$ $c_s^2 = 4.0$ $\rho = 0.7$	$c_a^2 = 0.8$ $c_s^2 = 4.0$ $\rho = 0.3$
Daley's general upper bound	5.75	9.00	1.82
Monotone failure rate upper bound[5]	5.83	7.50	1.07
Kraemer and Langenbach-Belz[7]	5.14	6.88	1.02
QNA[1]	5.14	7.00	1.02
M/G/1	5.83 H	5.83 L	1.07 H
M/M/1	2.33 L	2.33 L	0.43 L
MFR lower bound[5]	5.00	5.83	0.93

Notes: 1. In each case the mean service time is $\tau = 1$.
2. "H" indicates high (greater than or equal to) and "L" indicates low (less than or equal to) in comparison with the bounds.

approximations for the expected waiting time, *EW*, with various upper and lower bounds in the three cases in Table 1 of Ref. 6. The four approximations are the M/M/1, M/G/1, Kraemer and Langenbach-Belz,[7] and QNA.

The M/M/1 approximation is obtained by replacing both variability parameters c_a^2 and c_s^2 by 1. The M/M/1 approximation is produced by a direct application of the Markovian software packages. The M/G/1 approximation is obtained by replacing c_a^2 by 1 and using the specified value of c_s^2. The M/G/1 approximation *EW* is the exact value for the approximating M/G/1 system since *EW* depends on the service-time distribution only through its first two moments. Both the M/M/1 and M/G/1 approximations are produced by QNA using default options.

The QNA approximation is the Kraemer and Langenbach-Belz approximation when $c_a^2 \leq 1$ and is slightly greater when $c_a^2 > 1$.[1] The one case in which $c_a^2 > 1$ in Table I shows that the difference between the two approximations is small compared to the distance between the upper and lower Monotone Failure Rate (MFR) bounds.[5] The MFR bounds are for interarrival-time distributions with decreasing failure rate when $c_a^2 \geq 1$ and increasing failure rate when $c_a^2 \leq 1$. The MFR bounds are tight when $c_a^2 \geq 1$ but not when $c_a^2 < 1$ (see Ref. 5). Since the interarrival-time distribution need not have monotone failure rate, the range of all possible values is greater, but the MFR bounds indicate the more likely values.

For the three cases in Table I, the M/M/1 approximation performs very poorly, falling way outside the bounds. The M/G/1 approximation always coincides with one of the MFR bounds—the upper bound when $c_a^2 \leq 1$ and the lower bound when $c_a^2 \geq 1$—but it would be better to have an approximation somewhere in the middle between the bounds. When $c_a^2 \geq 1$, the QNA approximation is a convex combination of the

two MFR bounds.[5] Since the MFR bounds are tight when $c_a^2 \geq 1$, the QNA approximation always yields the exact value of EW for some GI/G/1 system with the given parameters.[5] When $c_a^2 \leq 1$, the QNA approximation is slightly less than the convex combination of the MFR bounds. The convex combination of the MFR bounds is known to be an upper bound for $E_k/G/1$ systems,[5] so that it is appropriate to use a smaller value. We conjecture that the QNA approximation always yields an exact value of EW for some GI/G/1 system with the given parameters when $c_a^2 \leq 1$ too.

Table I provides a sample of the comparisons possible using the previous studies.[2-6] Since QNA coincides with the Kraemer and Langenbach-Belz approximation when $c_a^2 \leq 1$ and the Sakasegawa-Yu approximation when $c_a^2 \geq 1$, previous comparisons such as Tables 13 and 14 in Klincewicz and Whitt[3] also apply to QNA.

III. A QUEUE WITH A SUPERPOSITION ARRIVAL PROCESS

In this section we consider one single-server queue with a superposition arrival process. Such a system with two component arrival processes is depicted in Fig. 1a. Since only one external arrival process at each node is allowed in QNA, this model cannot be analyzed directly, but it is easy to modify the model so that QNA does apply. We added dummy nodes with very low traffic intensity on each component arrival process, as shown in Fig. 1b. Since the new dummy nodes have low

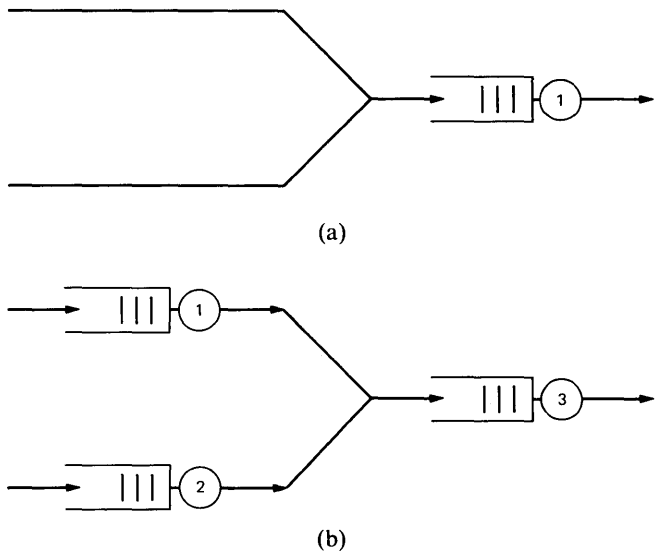


Fig. 1—(a) The original queue with a superposition arrival process. (b) The equivalent network with one external arrival process at each node.

traffic intensity, the rate and variability parameters of the departure processes from the dummy nodes will be almost identical to the corresponding parameters of the external arrival processes.

This model has recently been studied quite extensively by Albin[8-10] and is now relatively well understood. As in Section I, here we consider one illustrative example, which suggests the accuracy to expect more generally and shows the importance of the variability parameters.

The specific model we analyze is the $\Sigma GI_i/M/1$ system with an exponential service-time distribution and n iid stationary renewal processes as component arrival processes. The total arrival rate is 1, so the rate of each component process is n^{-1}. We consider six cases involving two values of n, $n = 2$ and 16, and three values of the traffic intensity ρ, $\rho = 0.3$, 0.7, and 0.9. In each case the component renewal-interval distribution is H_2^b, i.e., the mixture of two exponentials with balanced means: one with mean m_1 realized with probability p and the other with mean m_2 realized with probability $1 - p$, where $pm_1 = (1 - p)m_2$. In each case the squared coefficient of variation of the component renewal process interval is $c^2 = 6$. This is quite high variability, so that the component processes are not nearly Poisson. The three parameters of an H_2 distribution are determined by specifying the mean as n, $c^2 = 6$ and balanced means.

This model is taken from Chapter 3 and Appendix 6 of Albin.[8] Albin's approximations are based on the rate and variability parameters just as in QNA. In fact, the superposition approximation in QNA is a modification of Albin's procedure.[1] With Albin's procedure, the variability parameter of the superposition process is a convex combination of the variability parameters obtained from the stationary-interval method and asymptotic method described in Whitt.[15] The specific implementation of the stationary-interval method in Whitt[15] and Albin[8] is part of Kuehn's[12] algorithm for approximating networks of queues. Since Kuehn's implementation of the stationary-interval method is nonlinear, a different procedure is used in QNA. In QNA the variability parameter of the superposition process is a convex combination of the variability parameters obtained from the asymptotic method and a Poisson process.[1] Extensive experimentation has shown, however, that the approximation in QNA is very close to Albin's hybrid approximation, which performed very well in many experiments (about 3-percent average absolute relative percent error).

Several different approximations for the expected waiting time are compared with simulation in Table II. The simulation results were obtained in Albin.[8] The sample standard deviation is given in parentheses below the simulation estimate in Table II to indicate the statistical reliability of the estimate. The simulation program was written in FORTRAN using the "Super-Duper" subprogram in Mar-

Table II—A comparison of approximations and simulation of the expected waiting time, EW, in a $\Sigma GI/M/1$ queue with a superposition arrival process

No. of Renewal Processes, n	Traffic Intensity, ρ	Simulation Estimate	Approximation Method: M/M/1	Kuehn's Stationary-Interval Method	Asymptotic Method	Albin's Hybrid Procedure	QNA
2	0.3	0.205 (0.007)	0.128 (−37.6)	0.231 (+12.7)	0.281 (+37.1)	0.240 (+17.1)	0.236 (+15.1)
	0.7	4.57 (0.14)	1.63 (−64.3)	3.54 (−22.5)	5.32 (+16.4)	4.51 (−1.3)	4.64 (+1.5)
	0.9	26.3 (1.2)	8.1 (−69.2)	18.2 (−30.8)	28.2 (+7.2)	27.5 (+4.6)	27.5 (+4.6)
16	0.3	0.138 (0.004)	0.128 (−7.2)	0.147 (+6.5)	0.281 (+103.6)	0.153 (+10.9)	0.139 (+0.7)
	0.7	2.57 (0.07)	1.63 (−36.6)	1.88 (−26.8)	5.32 (+107.0)	2.36 (−8.2)	2.16 (−15.9)
	0.9	20.8 (0.94)	8.1 (−61.1)	9.4 (−54.8)	28.2 (+35.6)	21.1 (+1.4)	20.7 (−0.5)
Average Absolute Relative Percent Error			46.0	25.7	51.2	7.3	6.4

Notes: 1. The total arrival rate is 1 in each case.
2. The component renewal processes have H_2^b (hyperexponential) renewal-interval distributions with mean n, $c^2 = 6$, and balanced means.
3. The sample standard deviations appear below the simulation estimates in parentheses.
4. The relative percent error appears below the approximation values in parentheses.

saglia et al. to generate uniform random numbers.[16] A different random number seed was used for each simulation. The simulations began with an empty system, but the first 1000 customers were not counted to allow the system to approach steady-state. Each simulation consisted of 20 batches, with the number of customers per batch depending on the traffic intensity: 3,000 per batch for $\rho = 0.3$, 15,000 per batch for $\rho = 0.7$, and 50,000 per batch for $\rho = 0.9$. Even though much more simulation time was spent on the cases with higher traffic intensities, the statistical reliability was slightly less.

In Table II, in parentheses below the approximation values are the relative percent errors (RE), which are defined as

$$RE = 100(\text{Approx.} - \text{Simul.})/\text{Simul.} \tag{1}$$

At the bottom of Table II are the average absolute relative percent errors (ARE), which are defined as

$$ARE = \sum_{i=1}^{6} |RE_i|/6. \tag{2}$$

Dividing by the simulation value perhaps inflates the errors when $\rho =$

0.3 too much, but these summary measures provide a good overall comparison.

The stationary-interval method and the M/M/1 approximation are not bad for large n and small ρ because the superposition process converages to a Poisson process as $n \to \infty$ and the queue reflects this if ρ is not too big, in particular, if $n(1 - \rho)^2$ is sufficiently large.[10,17] However, for $\rho = 0.9$, these two methods perform poorly. On the other hand, the asymptotic method performs reasonably well for $\rho = 0.9$, but not well in other cases. In particular, the asymptotic method does not reflect the convergence to a Poisson process as $n \to \infty$; it gives the same answers for $n = 2$ and 16. So, for fixed ρ, the asymptotic method gets worse as n increases.

As Albin determined in extensive experiments,[8,9] her hybrid approximation is much better than either basic method alone. This example also illustrates how close QNA is to Albin's hybrid procedure. For queues with superposition arrival processes, we conclude that QNA usually gives reasonable results and strongly dominates the two basic methods.

In closing this section, we add a caveat. The component processes in the simulation for Table II, in Albin's hybrid procedure and in QNA, are all based on the case of balanced means. However, as discussed in Whitt,[4,5] given the first two moments, the one-parameter family of renewal processes with hyperexponential renewal-interval distributions range from a Poisson process to a batch Poisson process with geometrically distributed batch size. For a single renewal arrival process, the expected waiting time, EW, in an $H_2/G/1$ queue also ranges between these same extremes, i.e., the M/G/1 and the $M^B/G/1$ systems. Unfortunately, no related theory yet exists for superposition arrival processes. However, we can easily describe what happens in the two extremes. The superposition of independent Poisson processes is Poisson and the superposition of independent batch Poisson processes with geometrically distributed batches having a common mean is batch Poisson with geometrically distributed batches. Thus, we *conjecture* that the maximum and minimum values for EW with a superposition of iid H_2-renewal processes correspond to the $M^B/G/1$ and M/G/1 systems, respectively. If the conjecture is true, then we would have the same range of possible values for H_2-superposition arrival processes as for H_2-renewal processes. However, if the component processes are not too batchy, then the superposition process will become more Poisson as n increases. We should usually expect the superposition process to be more nearly Poisson as n increases. To summarize, this heuristic analysis suggests that the range of possible values for EW in the $\Sigma GI_i/G/1$ queue given the basic parameters λ, c_a^2, τ, c_s^2 may be about the same as for the GI/G/1 queue. In fact, the

superposition operation may actually make the arrival process better behaved.

IV. EIGHT QUEUES IN SERIES

In this section we apply QNA to a network of eight single-server queues in series previously analyzed by Fraker.[11] The external arrival process is Poisson and all service-time distributions are Erlang. Fraker considered eight cases involving four traffic intensities (ρ = 0.3, 0.5, 0.7, and 0.9) and four Erlang service-time distributions ($M = E_1$, E_4, E_8, and $D = E_\infty$). Each of the traffic intensities and each of the service-time distributions are assigned randomly to two of the eight nodes. Fraker developed an approximation for these systems and compared it with simulations.

Tables III and IV describe Fraker's first two cases and the approximations for the expected waiting time at each node. The service-time squared coefficient of variation specifies the Erlang distribution since $c^2 = k^{-1}$ for E_k. Fraker made three simulation runs of 2500 customers, discarding the first 500 in each case to damp out the transient effects of starting the simulation. Statistics were collected for six blocks of 1000 customers each. Unfortunately, this is not enough to produce very good accuracy, especially for the nodes with higher traffic intensities. (Compare with the simulation length in Section III.) The statistical reliability can be seen from the results of the six runs displayed in Fraker.[11] (These also appear in Appendix 1 of Whitt.[18]) An idea of the variability can also be seen from node 1 because all the approximations except the M/M/1 approximation are exact for node 1. When ρ = 0.9 the length of a 95-percent confidence interval approximately equals the estimated value; when ρ = 0.7 the length of

Table III—A comparison of approximations and simulation of the expected waiting time at each node in Fraker's model of eight queues in series: Case 1

Node No.	Traffic Intensity, ρ_j	Squared Coefficient of Variation, c_{sj}^2	Simulated Value	Approximation Methods (Markovian Network) M/M/1	(Asymptotic Method) M/G/1	(Lag-1 Correlations) Fraker	QNA EW_j	QNA c_{aj}^2
1	0.7	1/8	0.98	1.63	0.92	0.92	0.92	1.00
2	0.5	1	0.30	0.50	0.50	0.38	0.38	0.61
3	0.5	0	0.19	0.50	0.25	0.13	0.16	0.71
4	0.7	1/4	0.73	1.63	1.02	0.62	0.63	0.58
5	0.3	0	0.01	0.13	0.07	0.01	0.01	0.42
6	0.9	1	7.50	8.10	8.10	6.03	5.56	0.40
7	0.9	1/8	3.91	8.10	4.55	4.16	4.09	0.89
8	0.3	1/4	0.00	0.13	0.08	0.01	0.01	0.33

Note: The arrival process is Poisson with rate 1 and the service-time distributions are Erlang.

Table IV—A comparison of approximations and simulation of the expected waiting time at each node in Fraker's model of eight queues in series: Case 2

Node No.	Traffic Intensity, ρ_j	Squared Coefficient of Variation, c_{sj}^2	Simulated Value	Approximation Methods (Markovian Network) M/M/1	(Asymptotic Method) M/G/1	(Lag-1 Correlations) Fraker	QNA EW_j	QNA c_{aj}^2
1	0.9	1	6.25	8.10	8.10	8.10	8.10	1.00
2	0.7	1/8	0.84	1.63	0.92	0.92	0.92	1.00
3	0.3	1/4	0.01	0.13	0.08	0.04	0.04	0.61
4	0.9	0	2.61	8.10	4.05	2.45	2.28	0.58
5	0.3	1/8	0.00	0.13	0.07	0.00	0.00	0.27
6	0.5	1/4	0.02	0.50	0.31	0.05	0.06	0.26
7	0.5	0	0.02	0.50	0.25	0.02	0.02	0.26
8	0.7	1	0.78	1.63	1.63	0.82	0.89	0.25

Note: The arrival process is Poisson with rate 1 and the service-time distributions are Erlang.

a 95-percent confidence interval is about 25 percent of the estimated value.

Table V compares the approximations with simulation for the nodes with traffic intensity $\rho = 0.7$ in all eight cases. Since the approximations are exact for the first node, the first node is not included for the cases in which $\rho_1 = 0.7$ (Cases 1, 5, and 6). For the approximations, the difference between the approximation value and the simulation value is displayed.

Tables III through V show that QNA performs about the same as Fraker's approximation, which is based on lag-1 correlations and is especially designed for queues with Erlang service times. Both these approximations performed significantly better than the M/G/1 approximation, which in turn performs significantly better than the M/M/1 approximation.

Additional analysis of Fraker's models plus other queues in series is contained in Whitt.[18] The performance of QNA in these other cases is consistent with the description here.

V. A TIGHTLY COUPLED NETWORK OF TWO NODES

In this section we consider a two-node network analyzed by Kuehn[12] and Gelenbe and Mitrani.[13] This network is depicted in Fig. 2. It has one external arrival process, which comes to node 1. Customers completing service at node 1 leave the system with probability 1/2; otherwise they go to node 2 and then back to node 1 to be served again. At node 2 customers are immediately fed back to node 2 for another service with probability q_{22}, but in most cases $q_{22} = 0$.

We first consider Kuehn's experiment. There are eight cases with

Table V—The expected waiting time at the nodes with $\rho_j = 0.7$ in Fraker's eight cases of eight single-server queues in series

Case No.	Node No.	Simulated Value of EW_j	Approximation Methods			
			M/M/1	M/G/1	Fraker	QNA
1	4	0.73	+0.90	+0.29	−0.11	−0.10
2	2	0.84	+0.79	+0.08	+0.08	+0.08
	8	0.78	+0.85	+0.85	+0.04	+0.11
3	4	1.08	+0.55	+0.55	−0.04	+0.04
	8	0.55	+1.08	+0.37	0.00	−0.02
4	3	1.52	+0.11	+0.11	−0.04	−0.04
	6	0.02	+1.61	+0.80	+0.09	+0.16
5	3	0.74	+0.89	+0.28	+0.10	+0.11
6	5	0.33	+1.30	+0.49	+0.05	−0.01
7	4	0.78	+0.85	+0.14	−0.06	−0.02
	7	0.17	+1.46	+0.65	+0.02	−0.01
8	5	0.50	+1.13	+0.52	+0.04	+0.02
Average		0.67	+0.96	+0.43	+0.01	+0.03
Average Absolute Difference			0.96	0.43	0.05	0.06

Note: The value for the approximations is the approximate value minus the simulated value.

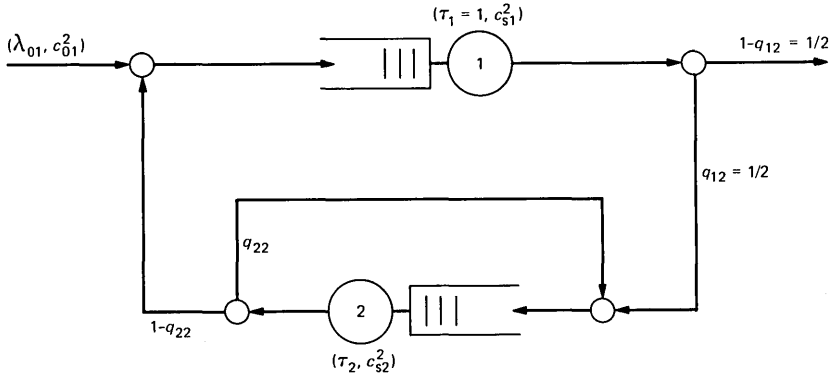


Fig. 2—Kuehn's first example: A network of two queues with one external arrival process.

three values of the external arrival rate for each case: $\lambda_{01} = 0.15$, 0.30, and 0.45. In each case the mean service time at node 1 is $\tau_1 = 1$ and the transition probability from node 1 to node 2 is $q_{12} = 1/2$, so that the traffic intensity at node 1 is $\rho_1 = 2\lambda_{01}$. For node 1 these three external arrival rates correspond roughly to light traffic ($\rho_1 = 0.3$), moderate traffic ($\rho_1 = 0.6$), and heavy traffic ($\rho_1 = 0.9$). In Cases 2 and 3 the traffic intensity at node 2 is the same as at node 1, i.e., $\rho_2 = 2\lambda_{01}$, but in all other cases it is $\rho_2 = \lambda_{01}$. In these other cases node 2 is always in relatively light traffic. The external arrival process is always a renewal process. Each interarrival-time distribution and service-time distribution is one of four distributions: deterministic (D with c^2

= 0), Erlang of order 4 (E_4 with $c^2 = 0.25$), exponential (M with $c^2 = 1$), or hyperexponential with balanced means (H_2^b with $c^2 = 2.25$). The eight cases are indicated in Table VI. The system type is described by a triple such as $M/H_2/E_4$, which means that the interarrival-time distribution is M and the service-time distributions at nodes 1 and 2 are H_2 and E_4, respectively.

The results are described in Tables VII and VIII. The simulation results and Kuehn's approximation are taken from Kuehn.[12] Two different approximation results are given for QNA in Table VII. The first column is the standard application of QNA with the network reconfigured to eliminate immediate feedback in the one case it occurs, at node 2 in Case 3. (See Section 3 of Ref. 1.) The final column of Table VII is an adjusted version of QNA to eliminate almost immediate feedback, which we discuss below.

For this network the quality of the standard QNA approximation is about the same as Kuehn's approximation. They both work well for low and moderate traffic intensities, e.g., about 10-percent average absolute relative percent error when $\rho_1 = 0.6$ (Table VIII), but not so well in heavy traffic. This two-node network presents an obvious difficulty for QNA. The network is tightly coupled so that many departures from node 1 rapidly return to node 1 for additional service. However, since these returning customers first pass through node 2, there is no immediate feedback, so that QNA does not reconfigure the network to eliminate the feedback. Nevertheless, it is evident that this almost immediate feedback for node 1 is very similar to immediate feedback and the potential exists for better results by reconfiguring the network to eliminate this feedback too.

In all cases except 2 and 3, almost immediate feedback is eliminated by applying the standard version of QNA with immediate feedback elimination twice. The first time we apply QNA to the full network and the second time we apply QNA with node 2 removed. When node

Table VI—The eight cases of Kuehn's two-node example in Fig. 2

System Number	System Type	Defining Parameters				
		c_{01}^2	τ_2	c_{s1}^2	c_{s2}^2	q_{22}
1	$M/H_2/E_4$	1.00	1.0	2.25	0.25	0.0
2	$M/H_2/E_4$	1.00	2.0	2.25	0.25	0.0
3	$M/H_2/E_4$	1.00	1.0	2.25	0.25	0.5
4	$M/H_2/H_2$	1.00	1.0	2.25	2.25	0.0
5	$M/E_4/E_4$	1.00	1.0	0.25	0.25	0.0
6	M/D/D	1.00	1.0	0.00	0.00	0.0
7	$H_2/H_2/E_4$	2.25	1.0	2.25	0.25	0.0
8	$E_4/H_2/E_4$	0.25	1.0	2.25	0.25	0.0

Notes: 1. In each case $\tau_1 = 1$ and $q_{12} = 1/2$.
2. In each case the arrival rate assumes one of three values: $\lambda_{01} = 0.15$, 0.30, and 0.45.

Table VII—A comparison of approximations and simulation of the expected total sojourn time (waiting time plus service time) in Kuehn's two-node network

System No.	External Arrival Rate, λ_{01}	Simulation (with 95-Percent Confidence Intervals)	Approximation Methods				
			M/M/1	M/G/1	Kuehn	QNA	QNA Adjusted
1	0.15	4.24 ± 0.08	4.04	4.50	4.51	4.51	4.24
	0.30	7.51 ± 0.29	6.43	8.13	8.22	8.26	7.25
	0.45	27.27 ± 5.63	21.82	32.67	33.47	34.10	27.35
2	0.15	5.95 ± 0.89	5.29	5.65	5.94	5.95	4.41
	0.30	10.91 ± 0.64	8.50	9.79	10.89	11.01	8.22
	0.45	49.49 ± 6.80	31.00	38.74	46.31	46.57	36.79
3	0.15	6.09 ± 0.19	5.29	5.65	6.11	6.12	4.48
	0.30	11.08 ± 0.63	8.50	9.79	11.60	11.66	8.25
	0.45	61.72 ± 17.99	31.00	38.74	51.27	51.43	35.07
4	0.15	4.50 ± 0.16	4.04	4.68	4.69	4.69	4.42
	0.30	7.96 ± 0.55	6.43	8.55	8.71	8.79	7.68
	0.45	29.91 ± 6.14	21.82	33.48	34.94	36.75	28.17
5	0.15	3.66 ± 0.05	4.04	3.63	3.63	3.64	3.81
	0.30	5.35 ± 0.16	6.43	5.13	4.93	5.00	5.67
	0.45	18.29 ± 3.27	21.82	14.67	12.80	12.80	18.12
6	0.15	3.43 ± 0.04	4.04	3.51	3.49	3.51	3.72
	0.30	4.83 ± 0.07	6.43	4.71	4.42	4.56	5.42
	0.45	13.59 ± 1.79	21.82	12.41	9.79	10.32	16.73
7	0.15	4.73 ± 0.12	4.04	4.50	4.60	4.62	4.78
	0.30	9.04 ± 0.56	6.43	8.13	8.59	8.92	9.14
	0.45	46.83 ± 10.64	21.82	32.67	35.88	39.74	39.50
8	0.15	3.67 ± 0.09	4.04	4.50	4.19	4.43	3.91
	0.30	5.78 ± 0.16	6.43	8.13	7.49	7.84	6.09
	0.45	17.46 ± 1.67	21.82	32.67	29.57	30.68	20.47

Notes: 1. In each case the traffic intensity at node 1 is $\rho_1 = 2\lambda_{01}$.
2. In Cases 2 and 3 the traffic intensity at node 2 is $\rho_2 = 2\lambda_{01}$; otherwise it is $\rho_2 = \lambda_{01}$.

2 is removed, the feedback to node 1 becomes immediate and the network is reconfigured by QNA to eliminate it. We use the second run with node 2 removed to determine the expected waiting time per visit at node 1. We use the first run to determine the expected number of visits to node 1 and the expected total sojourn time at node 2.

We do not treat nodes 1 and 2 symmetrically in Cases 1 and 4 through 8 because $\rho_1 = 2\rho_2$ so that ρ_2 is relatively small compared to ρ_1. Customers that return to node 1 via node 2 will not be delayed long at node 2 before coming back, but customers returning to node 2 via node 1 will be delayed relatively longer before coming back. If we had $\rho_1 < \rho_2$, we would remove node 1 in the second run of the QNA and focus instead on node 2.

In Cases 2 and 3 the traffic intensities at nodes 1 and 2 are equal, so the motivation for eliminating almost immediate feedback is less. What we have done for Table IV is first calculate the congestion

Table VIII—A comparison of approximation methods in Kuehn's two-node network with $\lambda_{01} = 0.3$ ($\rho_1 = 0.6$): The average absolute relative percent error in the expected total sojourn time compared with simulation

System Number	Approximation Methods					
	M/M/1	M/G/1	Kuehn	QNA Standard	QNA Adjusted	QNA Refined
1	−14.3	+8.3	+9.5	+10.0	−3.5	−3.5
2	−22.1	−10.3	−0.2	+0.9	−24.7	+0.9
3	−23.3	−11.6	+4.7	+5.2	−25.5	+5.2
4	−19.2	+7.4	+9.4	+10.4	−3.5	−3.5
5	+20.2	−4.1	−7.9	−6.5	+6.0	+6.0
6	+33.1	−2.5	−8.5	−5.6	+12.2	+12.2
7	−28.9	−10.1	−5.0	−1.3	+1.1	+1.1
8	+11.2	+40.7	+29.6	+35.6	+5.4	+5.4
Average Percent Error	21.5	11.9	9.4	9.4	10.2	4.7

measures for node 1 via the second run of QNA with node 2 removed as before. Then we use the results for node 1 to approximate the variability parameter of the arrival process to node 2. Finally, we analyze node 2 in isolation with the correct rates and this approximate arrival variability parameter. This works slightly better than the first procedure, but neither works well.

The results demonstrate that the standard version of QNA performs relatively well in Cases 2 and 3 when $\rho_1 = \rho_2$. The adjustment to eliminate almost immediate feedback yields a significant improvement when $\rho_1 > \rho_2$, but the results after adjustment to eliminate almost immediate feedback are much worse when $\rho_1 = \rho_2$.

As a refined procedure for this two-node network, we suggest eliminating almost immediate feedback at the node with higher traffic intensity when the traffic intensities differ significantly, and using the standard QNA algorithm otherwise. The refined procedure in Table VIII is standard QNA in Cases 2 and 3 and the adjusted QNA in all other cases.

Table VIII displays the relative percentage errors for all the approximations for the eight cases with $\lambda_{01} = 0.3$ ($\rho_1 = 0.6$). The refined procedure in the last column yields very good results. Table VIII also demonstrates that the standard version of QNA is significantly better than the M/M/1 approximation, but not uniformly better. In some cases, e.g., in Case 8, errors in opposite directions can cancel for the M/M/1 approximation.

The improvement from eliminating almost immediate feedback "by hand" suggests that it would be desirable to develop an automatic procedure for eliminating almost immediate feedback to incorporate

in QNA, and this is being investigated. It also indicates the potential for "tuning" QNA for particular applications.

We now consider Gelenbe and Mitrani's[13] experiment, which consists of five cases. The network is as depicted in Fig. 2 except the routing probability q_{12} is not exactly 1/2. The parameter values are given in Table IX and the results in Table X (pp. 137, 138 of Gelenbe and Mitrani[13]). We only include the best of three approximation schemes discussed by Gelenbe and Mitrani. Unfortunately, Gelenbe and Mitrani provided no information about the statistical reliability of the simulation estimates. Since ρ_1 and ρ_2 are nearly equal in each case, we did not try to eliminate almost immediate feedback.

The M/M/1 approximation values are obviously much too large because the M/M/1 approximation does not reflect the low variability of the service times. The M/G/1 approximation is much too low at node 2 because it does not benefit from the feedback elimination procedure. The Gelenbe-Pujolle procedure is better than the M/G/1 procedure, but not uniformly so. The QNA approximation is clearly

Table IX—The parameter values for Gelenbe and Mitrani's experiment with the two-node network in Fig. 2

Case No.	Parameter Values									
	λ_{01}	c_{01}^2	τ_1	c_{s1}^2	τ_2	c_{s2}^2	q_{11}	q_{12}	q_{21}	q_{22}
1	0.512	0.941	0.911	0.427	0.840	0	0	0.510	0.497	0.503
2	0.410	0.944	0.916	0.423	0.840	0	0	0.509	0.501	0.499
3	0.342	0.945	0.914	0.414	0.840	0	0	0.516	0.494	0.506
4	0.293	0.967	0.904	0.432	0.840	0	0	0.512	0.498	0.502
5	0.257	0.952	0.911	0.422	0.840	0	0	0.504	0.493	0.507

Table X—A comparison of approximations with simulations: The expected number of customers at each node in the two-node network in Gelenbe and Mitrani[13]

Case No.	Queue No.	Arrival Rate, λ_j	Traffic Intensity, ρ_j	Simulation Values	Approximation Methods			
					M/M/1	M/G/1	Gelenbe-Pujolle	QNA
1	1	1.04	0.952	13.82	19.83	14.14	11.96	11.98
	2	0.53	0.901	7.83	9.10	4.55	5.74	5.88
2	1	0.84	0.765	2.36	3.26	2.32	2.09	2.31
	2	0.43	0.713	1.87	2.48	1.24	1.69	1.84
3	1	0.71	0.646	1.60	1.82	1.29	1.20	1.41
	2	0.36	0.620	1.47	1.63	0.82	1.15	1.29
4	1	0.60	0.543	1.05	1.19	0.85	0.82	0.98
	2	0.31	0.519	1.01	1.08	0.54	0.78	0.90
5	1	0.52	0.472	0.76	0.89	0.63	0.62	0.76
	2	0.26	0.445	0.73	0.80	0.40	0.59	0.70

the best, but it may underestimate the congestion for high traffic intensities.

VI. KUEHN'S NINE-NODE NETWORK

We now consider a nine-node network analyzed by Kuehn,[12] which is depicted in Fig. 3. The mean service time at node j is $\tau_j = 1$ for each j. There are three external arrival processes with $\lambda_{0j} = 0.5$ for each j; these come to nodes 1, 2, and 3. Kuehn let the three external arrival processes be Poisson processes. As in Section V, all service-time distributions are D, E_k, M, or H_2^b. Kuehn considered two cases: homogeneous servers, in which all the service-time distributions are identical, and heterogeneous servers, in which nodes 1 through 3 have one service-time distribution and nodes 4 through 9 have another.

Kuehn compared his approximation with the M/M/1 and M/G/1 approximations and simulation for service-time variability parameters ranging from $c_{sj}^2 = 0$ to 4. He focused on the expected total sojourn time (waiting time plus service time) in the network and the expected sojourn time per visit in node 4. For this network Kuehn found, first, that the service-time variability parameters are significant (the sojourn-time measures increase significantly with c_{sj}^2); second, that his approximation tracks the simulation well; and, third, that the M/G/1 approximation also works well, but not quite as well as his approximation.

We obtained similar results applying QNA. The QNA approximation values are indistinguishable from the approximation values displayed graphically by Kuehn, which are consistently within the sim-

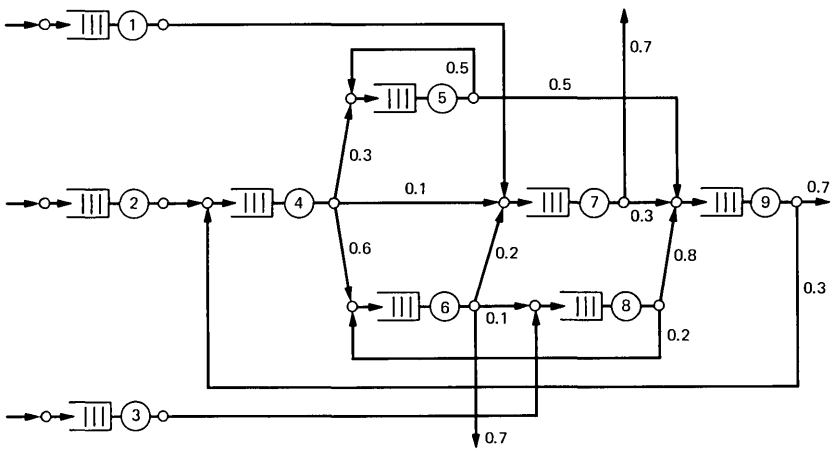


Fig. 3—Kuehn's second example: A network of nine queues with three Poisson external arrival processes.

ulation confidence intervals. In Table XI we display some of our results. Here we consider only the case of homogeneous servers in which $c_{sj}^2 = 0$, 1, or 4. However, we let the variability parameter of all external arrival processes by $c_{0j}^2 = 0.25$, 1.0, or 4.0. We thus obtain nine cases.

Kuehn did simulations only in the case $c_{0j}^2 = 1$. The different values obtained by QNA when $c_{0j}^2 \neq 1$ suggest that the congestion measures in this model are more sensitive to the variability of the service times than the variability of the interarrival times. Of course, we should expect that the M/G/1 approximation might perform well when the variability parameters of the external arrival processes are 1 or close to 1, but for relatively large and interconnected networks the M/G/1 approximation may perform well for other external arrival processes. It performs reasonably well here when $c_{0j}^2 = 0.25$ or 4.0.

VII. A COMPUTER SYSTEM MODEL: INPUT BY ROUTES

In this section we apply QNA to a network with input by customer classes and routes as in Section 2.3 of Ref. 1. We compare QNA to a simulation model used in the development of a computer system at Bell Laboratories. In this model there are five nodes and two customer classes.

The customer classes correspond to typical functions performed by the system. The route for each class represents a typical sequence of operations performed by the system to process one of these functions.

Table XI—Approximations of expected sojourn times in Kuehn's nine-node network in Fig. 3: The case of homogeneous servers, c_{sj}^2 identical for all j

Expected Total Sojourn Time in the Network		Service-Time Variability Parameters		
		$c_{sj}^2 = 0$	$c_{sj}^2 = 1$	$c_{sj}^2 = 4$
QNA	$c_{0j}^2 = 0.25$	6.1	11.5	28.3
	$c_{0j}^2 = 1.00$	7.5	12.9	29.6
	$c_{0j}^2 = 4.00$	12.3	17.7	34.4
M/G/1		8.2	12.9	27.1
M/M/1		12.9	12.9	12.9
Expected Sojourn Time per Visit in Node 4		**Service-Time Variability Parameters**		
		$c_{sj}^2 = 0$	$c_{sj}^2 = 1$	$c_{sj}^2 = 4$
QNA	$c_{0j}^2 = 0.25$	1.76	3.71	9.64
	$c_{0j}^2 = 1.00$	2.34	4.25	10.15
	$c_{0j}^2 = 4.00$	4.38	6.29	12.19
M/G/1		2.62	4.25	9.12
M/M/1		4.25	4.25	4.25

In the simulation model, both the routes and the service times at the nodes are deterministic for each class. Hence, the input for QNA is just as specified in Section 2.3 of Ref. 1 with the service-time variability parameters set equal to 0, i.e., $c^2_{skj} = 0$ for each class k and j on class k's route. The two routes we consider are given in Tables XII and

Table XII—The data for the first route in the model of Section VII

Number	Node	Mean Service Time	Number	Node	Mean Service Time
1	2	4	46	4	30
2	1	10	47	1	16
3	1	66	48	4	30
4	1	60	49	1	36
5	1	106	50	1	38
6	1	65	51	1	45
7	4	30	52	4	30
8	1	16	53	1	16
9	4	30	54	4	30
10	1	36	55	1	36
11	1	12	56	1	15.5
12	1	40	57	1	45
13	4	30	58	4	30
14	1	16	59	1	16
15	4	30	60	4	30
16	1	36	61	1	16
17	1	62	62	4	30
18	1	63	63	1	16
19	1	42	64	4	30
20	1	14.5	65	1	36
21	3	550	66	1	38
22	5	0.01	67	1	45
23	3	50	68	4	30
24	1	10	69	1	16
25	1	8	70	4	30
26	1	20.5	71	1	36
27	1	55.5	72	1	27
28	1	42	73	1	25
29	1	14.5	74	1	40
30	3	550	75	4	30
31	5	0.01	76	1	16
32	3	50	77	4	30
33	1	10	78	1	36
34	1	8	79	1	13
35	1	20.5	80	1	58
36	1	63.5	81	1	8
37	1	76	82	2	16
38	4	30	83	5	0.01
39	1	36	84	2	4
40	1	15.5	85	1	10
41	1	45	86	1	36
42	4	30			
43	1	16			
44	4	30			
45	1	16			

XIII. Note that node 1 frequently appears several times in succession, so that there is immediate feedback at node 1. Also note that the service times differ at different visits to the same node.

The customer classes arrive according to independent Poisson processes. In the case we consider the arrival rates of Classes 1 and 2 are 0.00015278 and 0.00030555, respectively.

The QNA, M/G/1, and M/M/1 approximations are compared with simulation in Table XIV. The simulation values are the average of three separate runs. The values from these separate runs are displayed to give an idea of the statistical reliability. The congestion measures compared are the expected waiting times at the nodes and the expected total waiting time (excluding service time) on three route segments. The first segment is the first 25 nodes of the second route; the second segment is the first 21 nodes on the first route; and the third segment is eight nodes from node 23 to node 30 on the first route. In Table XIV the waiting times at the nodes are measured in milliseconds while the waiting times on the route segments are measured in seconds.

From Table XIV, it is apparent that QNA with immediate-feedback elimination performs reasonably well, significantly better than the M/M/1 and M/G/1 approximations. Since the approximating variability parameters of the arrival processes are very close to 1, QNA without immediate-feedback elimination is very similar to the M/G/1 approximation. Hence, again we see that eliminating immediate feed-

Table XIII—The data for the second route in the model of Section VII

Number	Node	Mean Service Time	Number	Node	Mean Service Time	Number	Node	Mean Service Time
1	2	4	21	4	30	41	1	46
2	1	10	22	1	36	42	1	14.5
3	1	66	23	1	42	43	3	400
4	1	60	24	1	14.5	44	1	8
5	1	106	25	3	400	45	1	12
6	1	65	26	5	0.01	46	1	30
7	4	30	27	3	350	47	1	58
8	1	16	28	1	10	48	1	8
9	4	30	29	1	16	49	2	16
10	1	36	30	1	20.5	50	5	0.01
11	1	12	31	1	84	51	2	4
12	1	65	32	1	45	52	1	10
13	4	30	33	4	30	53	1	36
14	1	16	34	1	16			
15	4	30	35	4	30			
16	1	36	36	1	16			
17	1	62	37	4	30			
18	1	40	38	1	16			
19	4	30	39	4	30			
20	1	16	40	1	36			

Table XIV—A comparison of approximations and simulation of the expected waiting times for the model of Section VII

Method	Expected Waiting Time at the Nodes				Total Expected Waiting Time on a Route Segment		
	Node 1	Node 2	Node 3	Node 4	1 (25 nodes)	2 (21 nodes)	3 (8 nodes)
Simulation runs 1	51.9	0.058	224.1	2.11	1.25	1.12	0.66
2	58.5	0.058	245.6	2.27	1.37	1.27	0.77
3	57.6	0.055	236.2	2.17	1.35	1.23	0.74
Simulation average	56.0	0.057	235.3	2.18	1.32	1.21	0.72
M/M/1	59.2 (+5.7)	0.09	402.5 (+71.1)	6.53 (+200.0)	1.45 (+9.8)	1.32 (+9.1)	1.16 (+61.1)
M/G/1	43.5 (−22.3)	0.07	245.5 (+4.2)	3.26 (+49.5)	1.01 (−23.5)	0.91 (−24.8)	0.75 (+4.2)
QNA (eliminating feedback)	50.2 (−10.3)	0.07	244.1 (+3.7)	2.81 (+28.9)	1.11 (−15.9)	1.01 (−16.5)	0.79 (+9.7)

	Additional Information About the Network						
	1	2	3	4	1	2	3
Mean service time	33.1	8.0	350.0	30.0	1.28	1.31	0.75
Traffic intensity	0.64	0.11	0.54	0.18	---	---	---
c^2 of the service time from QNA	0.47	0.50	0.22	0.00	---	---	---
c^2 of the arrival process from QNA	0.92	1.00	0.99	0.90	---	---	---

Notes: 1. The relative percent errors appear below the approximation in parentheses.
2. The value $c_{s1}^2 = 0.47$ at node 1 is before adjustment for feedback; after adjustment it is 0.79.
3. The units of measurement are milliseconds for the nodes and seconds for the route segments.

back helps. The M/M/1 approximations at nodes 3 and 4 evidently are too large because the service times are nearly constant. However, at node 1 the M/M/1 approximation does pretty well, apparently because two different errors cancel. (We have not displayed the relative percentage errors at node 2 since it seems of little consequence because of the low traffic intensity.)

It is interesting to know how the system would perform if the external arrival processes are not Poisson and if the service times at the nodes on the routes are not deterministic. With QNA we can easily perform such sensitivity analyses. We can simply change the variability parameters of the external arrival processes and the service times on the routes. The results of such a study are given in Tables XV and XVI. Table XV gives the approximating variability parameter of the

arrival process at each node as a function of the external arrival process c^2 and the service time c^2. It is assumed that both external arrival processes have the same c^2 and that all service times on both routes have the same c^2. From Table XV, it is evident that the variability of the external arrival process hardly has any effect. Thus, QNA predicts that this model, and perhaps the system itself, will be robust to changes in the variability of the arriving traffic. The variability from outside is evidently dissipated on the long routes through the network.

Table XVI describes the impact of changing the service-time variability at the nodes on the routes. The service-time variability at the nodes would increase significantly and, thus, QNA predicts that the expected waiting times would also increase significantly. Simulations to test these predictions are planned.

Table XV—The approximate variability parameter of the arrival process at each node determined by the QNA, as a function of the given variability parameters: The model of Section VII

External Arrival Process c^2	Node	Service Time c^2 at Each Node on the Route				
		0.0	0.2	0.5	1.0	2.0
$c^2 = 1$	1	0.9155	0.9506	1.0034	1.0912	1.2669
	2	0.9995	0.9997	1.0000	1.0001	1.0016
	3	0.9937	0.9966	1.0010	1.0083	1.0230
	4	0.8950	0.9480	1.0274	1.1598	1.4245
$c^2 = 2$	1	0.9162	0.9513	1.0040	1.0918	1.2676
	2	1.0086	1.0088	1.0092	1.0097	1.0107
	3	0.9938	0.9967	1.0011	1.0084	1.0231
	4	0.8953	0.9483	1.0277	1.1601	1.4249
$c^2 = 4$	1	0.9175	0.9527	1.0053	1.0932	1.2689
	2	1.0268	1.0271	1.0274	1.0279	1.0290
	3	0.9939	0.9968	1.0012	1.0085	1.0231
	4	0.8959	0.9488	1.0283	1.1607	1.4254

Table XVI—The approximate service-time variability parameter c_{sj}^2 and mean delay EW_j at node j determined by the QNA, as a function of the variability of each service time on the route: The model of Section VII

Node Characteristic	Node	Service Time c^2 at Each Node on the Route				
		0.0	0.2	0.5	1.0	2.0
c_{sj}^2	1	0.787	0.905	1.082	1.377	1.968
	2	0.500	0.800	1.250	2.000	3.500
	3	0.220	0.464	0.830	1.441	2.661
	4	0.000	0.200	0.500	1.000	2.000
EW_j	1	50.2	54.9	61.6	72.6	94.6
	2	0.07	0.08	0.10	0.13	0.20
	3	244.1	293.3	367.7	491.4	739.0
	4	2.81	3.72	4.95	6.87	10.78

VIII. A PACKET-SWITCHED COMMUNICATION-NETWORK MODEL

In this section we consider a model of a packet-switched communication network analyzed in Section 4.3.1 of Gelenbe and Mitrani.[13] The basic model has 5 switching nodes and 12 one-way data links, as depicted in Fig. 4. However, in this model each data link is a server and the packets waiting for transmission on the link form the queue. Packets are assumed to arrive at the switching nodes according to independent Poisson processes. Each packet arriving from outside at node i has final destination j with probability d_{ij}. Each packet with destination j goes next to node r_{ij} from node i. Hence, there is a fixed route for each origin-destination pair.

We analyze this network using the input by classes and routes in Section 2.3 of Ref. 1. However, unlike Section VII, here the service-time parameters are associated with the nodes rather than the routes (which is an input option in QNA). The network of queues has 12 nodes with one server at each node and 20 routes. As specified by Gelenbe and Mitrani, the service rate at nodes 1, 2, 7, 8, 11, and 12 is 4.8 (in thousands of bits per second) and the service rate at the other nodes is 48. Since packet lengths are assumed constant, $c_{sj}^2 = 0$ for all j.

For this example the matrices $D \equiv (d_{ij})$ of destination probabilities and $R = (r_{ij})$ of next-node routes are:

$$D = \begin{bmatrix} 0.00 & 0.10 & 0.20 & 0.10 & 0.60 \\ 0.40 & 0.00 & 0.40 & 0.15 & 0.05 \\ 0.10 & 0.20 & 0.00 & 0.60 & 0.10 \\ 0.30 & 0.30 & 0.30 & 0.00 & 0.10 \\ 0.10 & 0.25 & 0.30 & 0.35 & 0.00 \end{bmatrix} \quad R = \begin{bmatrix} 0 & 3 & 3 & 3 & 2 \\ 4 & 0 & 5 & 5 & 4 \\ 6 & 6 & 0 & 9 & 8 \\ 10 & 10 & 10 & 0 & 12 \\ 1 & 1 & 7 & 11 & 0 \end{bmatrix}.$$

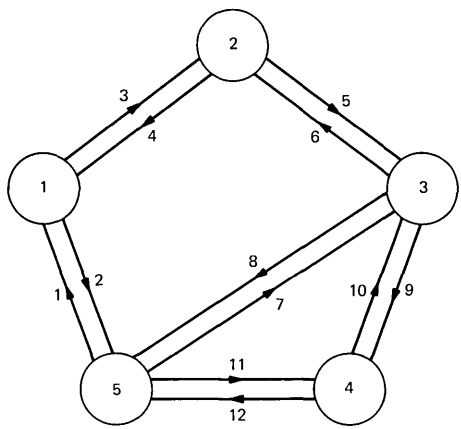


Fig. 4—Gelenbe and Mitrani's model of a packet-switched network: The 12 links are the nodes in the network of queues.

The input data for the routes are given in Table XVII. These data are obtained from the matrices D and R plus the external arrival rates of 6.00, 8.25, 7.50, 6.75, and 1.50 at the five switching nodes (p. 141 of Gelenbe and Mitrani[13]).

The results are compared with Gelenbe and Mitrani's approximation and simulation in Table XVIII. Since only 6000 packets reached their destination in the simulation, the statistical reliability of the simulation estimates cannot be very good, cf. Section III. Our analysis is revealing. First, Gelenbe and Mitrani describe their results as average buffer queue lengths, which might be thought to exclude the customer (packet) being served (transmitted). However, the results obviously include the customer in service. Second, the two M/M/1 approximations should agree, but they do not. Evidently, the arrival rates at switches 1 and 2 were not actually as cited in the text.[13] Hence, the numbers for the one heavily loaded link, link 2, cannot be meaningfully compared.

It is useful to consider the expected number waiting excluding the customer in service. This is easily obtained because the probability that the server is busy is exactly the traffic intensity, ρ (see Section 11.3 of Heyman and Sobel[14]). We thus obtain estimates of the expected number waiting by subtracting ρ from the numbers displayed in Table

Table XVII—The input data by routes for Gelenbe and Mitrani's model of a packet-switched communication network

Route Number	Origin-Destination Pair	External Arrival Process Parameters $\hat{\lambda}_k$	c_k^2	Number of Nodes on the Route	Node Sequence
1	1,2	0.60	1	1	3
2	1,3	1.20	1	2	3,5
3	1,4	0.60	1	3	3,5,9
4	1,5	3.60	1	1	2
5	2,1	3.30	1	1	4
6	2,3	3.30	1	1	5
7	2,4	1.24	1	2	5,9
8	2,5	0.41	1	2	4,2
9	3,1	0.75	1	2	6,4
10	3,2	1.50	1	1	6
11	3,4	4.50	1	1	9
12	3,5	0.75	1	1	8
13	4,1	2.03	1	3	10,6,4
14	4,2	2.03	1	2	10,6
15	4,3	2.03	1	1	10
16	4,5	0.68	1	1	12
17	5,1	0.15	1	1	1
18	5,2	0.38	1	2	1,3
19	5,3	0.45	1	1	7
20	5,4	0.53	1	1	11

XVIII. When this is done, some of the simulation estimates become negative, demonstrating that the input parameters are incorrect or the statistical reliability of the simulation is not very good. In this case, probably both problems exist.

We also observe that the traffic intensities at all link queues but the second are very small, so the numbers displayed in Table XVIII are mostly estimates of the traffic intensities themselves. Moreover, because the traffic intensities are small, the variability parameter of the departure process produced by QNA will be very close to the variability parameter of the arrival process (see Section 4.5 of Ref. 1). This would not be true for the second link with traffic intensity 0.835, but note that all departures from the second link leave the system. Since the external arrival processes are all Poisson, QNA should and does perform virtually the same as an M/G/1 approximation. In fact, since the service times are all constant ($c_{sj}^2 = 0$), the approximation reduces to the M/D/1 system. Moreover, we predict that a proper simulation of this model with the specified parameters will yield values very close to the M/D/1 approximation.

We also display in Table XIX the point-to-point (origin-destination) average total service times, delays, and sojourn times (service times plus delays) produced by QNA. No simulation values were available for comparison, however. The output is useful to indicate unacceptably high or low values. It is also useful to determine the separate contributions of service times and delays to sojourn times. Of course, in

Table XVIII—A comparison of approximations and simulation: The expected number waiting and being served on each link in the Gelenbe-and-Mitrani model of a packet-switched communication network depicted in Fig. 4

			Approximation Methods			
			M/M/1			
Link No.	Traffic Intensity, ρ_j	Simulation Value	Gelenbe and Mitrani	via QNA	Gelenbe-Pujolle	QNA
1	0.110	0.117	0.123	0.124	0.116	0.117
2	0.835	1.920	3.000	5.076	1.875	2.955
3	0.058	0.132	0.139	0.061	0.131	0.060
4	0.135	0.163	0.170	0.156	0.157	0.146
5	0.132	0.105	0.127	0.152	0.125	0.142
6	0.131	0.173	0.157	0.151	0.146	0.141
7	0.094	0.087	0.104	0.103	0.099	0.099
8	0.156	0.208	0.185	0.185	0.171	0.171
9	0.132	0.155	0.162	0.152	0.147	0.142
10	0.127	0.129	0.145	0.145	0.136	0.136
11	0.110	0.106	0.123	0.124	0.116	0.117
12	0.142	0.154	0.164	0.165	0.152	0.153

Table XIX—Average point-to-point service times, delays, and sojourn times for the Gelenbe-and-Mitrani model of a packet-switched communication network

Route No.	Mean Total Service Time on Route	Poisson Arrivals $c^2 = 1.0$		Bursty Arrivals $c^2 = 4.0$	
		Mean Total Delay on Route	Mean Total Sojourn Time on Route	Mean Total Delay on Route	Mean Total Sojourn Time on Route
1	0.021	0.001	0.021	0.001	0.022
2	0.042	0.002	0.044	0.003	0.044
3	0.063	0.004	0.066	0.005	0.068
4	0.208*	0.529*	0.737*	1.920*	2.128*
5	0.021	0.002	0.022	0.002	0.023
6	0.021	0.002	0.022	0.002	0.023
7	0.042	0.003	0.045	0.005	0.046
8	0.229*	0.530*	0.759*	1.922*	2.151*
9	0.042	0.003	0.045	0.004	0.046
10	0.021	0.002	0.022	0.002	0.046
11	0.021	0.002	0.022	0.003	0.024
12	0.208*	0.019	0.228*	0.077	0.285*
13	0.063	0.005	0.067	0.006	0.068
14	0.042	0.003	0.045	0.004	0.046
15	0.021	0.002	0.022	0.002	0.023
16	0.208*	0.017	0.226*	0.069	0.277*
17	0.208*	0.013	0.221*	0.025	0.233*
18	0.229*	0.014	0.243*	0.026	0.255*
19	0.208*	0.011	0.219*	0.043	0.251*
20	0.208*	0.013	0.221*	0.052	0.260*

Note: The larger values are marked with an asterisk.

Table XIX delays play a significant role only for routes using the second link.

We conclude by remarking that the assumption of Poisson arrivals for packets at each switch made by Gelenbe and Mitrani[13] often is not realistic. Often messages containing many packets arrive according to a Poisson process, but the packets arrive in a much more bursty manner. Hence, it is appropriate to use QNA with arrival-process variability parameters much larger than 1. The last two columns of Table XIX give the mean delays and sojourn times when the variability parameters of the external arrival processes are changed from $c^2 = 1.0$ to $c^2 = 4.0$. When this is done here, the large delays on routes 4 and 8 increase significantly. To a large extent, QNA was motivated by the need to be able to systematically study the effect of such variability.

IX. ACKNOWLEDGMENTS

I am grateful to Anne Seery for using QNA to generate much of the data here, as well as for writing the QNA program. I am grateful to E. B. Zucker for the simulation data in Section VII.

REFERENCES

1. W. Whitt, "The Queueing Network Analyzer," B.S.T.J., this issue.
2. W. Whitt, "On Approximations for Queues, I: Extremal Distributions," B.S.T.J., *63*, No. 1, Part 1 (January 1984).
3. J. G. Klincewicz and W. Whitt, "On Approximations for Queues, II: Shape Constraints," B.S.T.J., *63*, No. 1, Part 1 (January 1984).
4. W. Whitt, "On Approximations for Queues, III: Mixtures of Exponential Distributions," B.S.T.J., *63*, No. 1, Part 1 (January 1984).
5. W. Whitt, "The Marshall/Marshall and Stoyan Bounds for IMRL/G/1 Queues are Tight," Oper. Res. Letters, *1*, No. 6 (December 1982), pp. 209–13.
6. W. Whitt, "Refining Diffusion Approximations for Queues," Oper. Res. Letters, *1*, No. 5 (November 1982), pp. 165–9.
7. W. Kraemer and M. Langenbach-Belz, "Approximate Formulae for the Delay in the Queueing System GI/G/1," *Congressbook*, Eighth International Teletraffic Congress, Melbourne, Australia, 1976, pp. 235, 1–8.
8. S. L. Albin, *Approximating Queues with Superposition Arrival Processes*, Ph.D dissertation, Department of Industrial Engineering and Operations Research, Columbia University, 1981.
9. S. L. Albin, "Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues," Department of Industrial Engineering Rutgers University, 1982.
10. S. L. Albin, "On Poisson Approximations for Superposition Arrival Processes in Queues," Management Sci., *28*, No. 2 (February 1982), pp. 126–37.
11. J. R. Fraker, *Approximate Techniques for the Analysis of Tandem Queueing Systems*, Ph.D. dissertation, Department of Industrial Engineering, Clemson University, 1971.
12. P. J. Kuehn, "Approximate Analysis of General Queueing Networks by Decomposition," IEEE Trans. Commun., *COM-27*, No. 1 (January 1979), pp. 113–26.
13. E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems*, New York: Academic Press, 1980.
14. D. P. Heyman and M. J. Sobel, *Stochastic Models in Operations Research*, Volume I, New York: McGraw-Hill, 1982.
15. W. Whitt, "Approximating a Point Process by a Renewal Process, I: Two Basic Methods," Oper. Res., *30*, No. 1 (January–February 1982), pp. 125–47.
16. G. Marsaglia, K. Ananthanarayanan, and N. Paul, "Random Number Generator Package–Super Duper," School of Computer Science, McGill University, 1973.
17. W. Whitt, "Queues with Superposition Arrival Processes in Heavy Traffic," unpublished work, 1982.
18. W. Whitt, "Approximations for Departure Processes and Queues in Series," Navy Res. Log Qtrly., to be published.

AUTHOR

Ward Whitt, A. B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; Bell Laboratories, 1977—. At Yale University, from 1973–1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At Bell Laboratories he is in the Operations Research Department in the Network Analysis Center.

PAPERS BY BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Berreman D. W., **Numerical Modeling of Twisted Nematic Devices.** Phi T Roy A 309(1507):203–216, 1983.

Fishburn P. C., Pollak H. O., **Fixed-Route Cost Allocation.** Am Math Mo 90(6):366–378, 1983.

Stein A. H., An C. M., **A Mean-Value Theorem for Zeta-Functions Associated With Positive Definite Integral Forms.** Mich Math J 30(1):3–8, 1983.

ENGINEERING

Amitay N., Saleh A. A. M., **Broad-Band Wide-Angle Quasi-Optical Polarization Rotators.** IEEE Antenn 31(1):73–76, 1983.

Bartoli P. D., **CCITT Message Handling Facilities.** Aust Tele R 16(3):53–62, 1982.

Borsuk J. A., **Light-Intensity Profiles of Surface-Emitting InGaAsP LEDs—Impact on Coupling to Optical Fibers.** IEEE Device 30(4):296–303, 1983.

Campbell J. C., Qua G. J., Dentai A. G., **Optical Comparator—A New Application for Avalanche Phototransistors.** IEEE Device 30(4):408–411, 1983.

Capasso F., Tsang W. T., Williams G. F., **Staircase Solid-State Photomultipliers and Avalanche Photo-Diodes With Enhanced Ionization Rates Ratio.** IEEE Device 30(4):381–390, 1983.

Chin A. K., Zipfel C. L., Ermanis F., Marchut L., Camlibel I., Di Giuseppe M. A., Chin B. H., **The Migration of Gold From the P-Contact as a Source of Dark Spot Defects in InP/InGaAsP LEDs.** IEEE Device 30(4):304–310, 1983.

Dutta N. K., **Current Injection in Multiquantum Well Lasers (Letter).** IEEE J Q El 19(5):794–797, 1983.

Dutta N. K., Nelson R. J., Wright P. D., Besomi P., Wilson R. B., **Optical-Properties of a 1.3-μm InGaAsP Superluminescent Diode.** IEEE Device 30(4):360–363, 1983.

Heller A., **High-Efficiency of Photo-Electrochemical Cells—Reply (Letter).** Chem Eng N 61(22):4, 1983.

Johannes V. I., **Performance Parameters for Digital-Communications (Letter).** P IEEE 71(4):539, 1983.

Lee T. P., **Special Issue on Light-Emitting Diodes and Long-Wavelength Photodetectors—Foreword (Editorial).** IEEE Device 30(4):257–258, 1983.

Levy U., Gordon E. I., Logan R. A., **Laser Cathode-Ray Tube With a Semiconductor Double-Heterostructure Screen.** IEEE Elec D 4(5):155–156, 1983.

Ma K. T., Kusic G. L., **Tie Line Contingency Studies Based Upon Partial Information.** IEEE Power 102(6):1838–1842, 1983.

Saul R. H., **Recent Advances in the Performance and Reliability of InGaAsP LEDs for Lightwave Communication-Systems.** IEEE Device 30(4):285–295, 1983.

Zipfel C. L., Chin A. K., Di Giuseppe M. A., **Reliability of InGaAsP Light-Emitting Diodes at High-Current Density.** IEEE Device 30(4):310–316, 1983.

PHYSICAL SCIENCES

Acuna M. H. et al., **Physics of the Jovian and Saturnian Magnetospheres—Highlights of a Conference Held at the Applied Physics Laboratory, The Johns-Hopkins-University, October 22–24, 1981 (Review).** Space Sci R 35(3):269–292, 1983.

Allara D. L., Hebard A. F., Padden F. J., Nuzzo R. G., Falcone D. R., **Chemically-Induced Enhancement of Nucleation in Noble-Metal Deposition.** J Vac Sci A 1(2):376–382, 1983.

Bean J. C., **Recent Developments in Silicon Molecular-Beam Epitaxy.** J Vac Sci A 1(2):540–545, 1983.

Chabal Y. J., **Hydrogen Vibration on Si(111)7X7—Evidence for a Unique Chemisorption Site.** Phys Rev L 50(23):1850–1853, 1983.

Chang J. J., Julesz B., **Displacement Limits, Directional Anisotropy and Direction Versus Form Discrimination in Random-Dot Cinematograms.** Vision Res 23(6):639+, 1983.

Donnelly V. M., Flamm D. L., Ibbotson D. E., **Plasma-Etching of III-V-Compound Semiconductors.** J Vac Sci A 1(2):626–628, 1983.

Dubois L. H., Rowe J. E., **Surface Phonons on Clean-Covered and Adsorbate-Covered Nickel Disilicide ($NiSi_2$) Thin-Films.** J Vac Sci A 1(2):1232–1235, 1983.

Evans D., Celli V., Benedek G., Toennies J. P., Doak R. B., **Resonance-Enhanced Atom Scattering From Surface Phonons.** Phys Rev L 50(23):1854–1857, 1983.

Focht M. W., Schwartz B., **High-Resistivity in P-Type InP by Deuteron Bombardment.** Appl Phys L 42(11):970–972, 1983.

Hannay N. B., **The Information Society—Perkin Medal Address (Editorial).** Chem Ind L 1983(11):433–437, 1983.

Hart A. C., Krause J. T., **Coating Technique for High-Strength Lightguide Fusion Splices.** Appl Optics 22(11):1731–1733, 1983.

Jackel J. L., Hackwood S., Veselka J. J., Beni G., **Electrowetting Switch for Multimode Optical Fibers.** Appl Optics 22(11):1765–1770, 1983.

Lyon S. A., Chen Y. H., Lin J. F., Worlock J. M., **Optical-Transmission at 3.39 μm During Pulsed Laser Annealing of Silicon.** Appl Phys L 42(11):978–980, 1983.

Miller D. A. B., Chemla D. S., Eilenberger D. J., Smith P. W., Gossard A. C., Wiegmann W., **Degenerate 4-Wave Mixing in Room-Temperature GaAs/GaAlAs Multiple Quantum Well Structures.** Appl Phys L 42(11):925–927, 1983.

Moncton D. E., Brown G. S., **High-Resolution X-Ray-Scattering.** Nucl Instru 208(1–3):579–586, 1983.

Osinski J. S., Manzione L. T., **Characterization and Moldability Analysis of Epoxy Reaction Injection-Molding Resins (Review).** ACS Symp S 1983(221):263–282, 1983.

Ota Y., **Defect Evaluation of Si MBE Film.** J Cryst Gr 61(3):439–448, 1983.

Patel J. R., Golovchenko J. A., **X-Ray-Standing-Wave Atom Location in Heteropolar Crystals and the Problem of Extinction.** Phys Rev L 50(23):1858–1861, 1983.

Reichmanis E., Wilkins C. W., Price D. A., Chandross E. A., **The Effect of Substituents on the Photosensitivity of 2-Nitrobenzyl Ester Deep UV Resists.** J Elchem So 130(6):1433–1437, 1983.

Rousseau D. L., Ondrias M. R., **Resonance Raman-Scattering Studies of the Quaternary Structure Transition in Hemoglobin (Review).** Ann R Bioph 12:357–380, 1983.

Shapira Y., Brillson L. J., Heller A., **Investigation of InP Surface and Metal Interfaces by Surface Photo-Voltage and Auger-Electron Spectroscopies.** J Vac Sci A 1(2):766–770, 1983.

Sharma S. P., Sproles E. S., **Reaction of Palladium With Chlorine and Hydrogen-Chloride.** J Elchem So 130(6):1242–1247, 1983.

Temkin H., Logan R. A., Van Der Ziel J. P., **Integrated Arrays of 1.3-μm Buried-Crescent Lasers.** Appl Phys L 42(11):934–936, 1983.

Thiel F. A., **The Phase-Relations in the Cu-In-S-System and the Growth of $CuInS_2$ Crystals From the Melt—Reply (Discussion).** J Elchem So 130(6):1445, 1983.

Trapp K. D. C., Ermanis F., **Origin and Elimination of Crescent-Shaped Growth Defects in LPE Layers of InGaAs/InP Alloys.** J Elchem So 130(6):1381–1383, 1983.

Woodruff D. P., Johnson P. D., Smith N. V., **Inverse Photoemission.** J Vac Sci A 1(2):1104–1110, 1983.

Yao S. C., Chang Y., **Pool Boiling Heat-Transfer in a Confined Space.** Int J Heat 26(6):841–848, 1983.

SOCIAL AND LIFE SCIENCES

Anderson P. W., **Suggested Model for Prebiotic Evolution—The Use of Chaos.** P Nas Biol 80(11):3386–3390, 1983.

Weiss A., **A Sorting-cum-Learning Model of Education.** J Polit Ec 91(3):420–442, 1983.

CONTENTS, DECEMBER 1983

Part 1

Part 2

COMPUTING SCIENCE AND SYSTEMS

Part 3

THE AR6A SINGLE-SIDEBAND MICROWAVE RADIO SYSTEM:

THE BELL SYSTEM 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.


Bell System