THE BELL SYSTEM
TECHNICAL JOURNAL
DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 52 March 1973 Number 3

The Elimination of Tuning-Induced Burnout and Bias-Circuit Oscillations in IMPATT Oscillators

By C. A. BRACKETT

(Manuscript received August 18, 1972)

IMPATT diode microwave oscillators suffer from the effects of low-frequency instabilities, which include excessive up-conversion of bias-circuit noise, bias-circuit oscillations, and diode burnout induced by tuning at the microwave frequency. These instabilities are particularly troublesome in GaAs diodes, although also present in both Ge and Si to a lesser extent. Moreover, these instabilities are more prominent in higher efficiency, higher power diodes, presenting a severe systems problem in the practical utilization of GaAs diodes at their highest power and efficiency levels. In this paper, it is shown that these instabilities may be eliminated in a systematic and well controlled manner with little or no loss in microwave power or efficiency.

It is shown that the source of the unstable behavior is a low-frequency RF voltage-induced negative resistance which extends from dc to several tens, and perhaps hundreds, of megahertz, depending on the loaded Q of the microwave circuit. The negative resistance is an unavoidable fact of large-signal avalanche diode operation and is due to the rectification properties of the nonlinear microwave avalanche. Experiments are performed in which the negative resistance and its associated inductive reactance are

measured as functions of the dc bias current, baseband frequency, and microwave circuit loading and Q.

An analysis is performed and a simple small-signal equivalent circuit is derived for the diode terminal impedance which yields quantitative and qualitative understanding of the interaction of the IMPATT device with the microwave oscillator circuit.

Using the equivalent circuit and experimental characterization, a stability criterion is developed that is simply applied to any diode-circuit combination. Using this criterion, several examples of common configurations are investigated, including the maximum shunt capacity permissible in the bias circuit for stable operation, and the effects of line length in the bias circuit. Also discussed are two means of achieving stable operation, both of which have been tested experimentally and shown to work.

The scaling laws are derived in an approximate manner and used to show that diodes designed for higher frequencies will have less induced negative resistance. Thus, mm-wave silicon IMPATTs would have less trouble with low-frequency instabilities than those designed at 6 GHz.

I. INTRODUCTION

Low-frequency instabilities including noise, bias-circuit oscillations, and diode burnout at low bias currents are often encountered when tuning high-power, high-efficiency IMPATT diodes. Gallium arsenide IMPATTs, which produce high power at higher efficiency and lower noise than either silicon or germanium IMPATTs, are particularly prone to these instabilities. Microwave tuning operations in GaAs often burn out diodes at bias currents of one-half to one-quarter their thermally expected values. This problem has been so serious that it has cast doubt on the practicality of using GaAs IMPATTs at their fullest potential power and efficiency. These instabilities also make the testing and evaluation of diodes a very tedious procedure and cause the operating conditions of the oscillator to be very sensitive to microwave and bias-circuit load conditions.

The cause of these problems is an RF voltage-induced negative resistance in the IMPATT diode. This negative resistance has a low-pass frequency dependence, extending from essentially zero frequency up through several tens, and perhaps hundreds, of megahertz. The upper frequency limit is determined by the bandwidth of the microwave circuit. The magnitude and frequency dependence of the induced negative resistance are dependent on the microwave circuit tuning. It is possible to tune through very high values of negative resistance,

even at low dc bias currents. If this negative resistance is not stabilized, the results can be an excess amount of upconverted microwave noise, bias-circuit oscillations, or diode burnout due to the instability achieving an excessive amplitude. Stabilization removes the instability and eliminates these "tuning-induced" burnouts, as well as the excess noise and bias-circuit oscillations.

In this paper we (*i*) discuss the physical origins of this negative resistance, (*ii*) develop an equivalent circuit which agrees with experimental observations, (*iii*) present an experimental characterization of the effect, and (*iv*) discuss the principles and techniques of stabilization which have been used experimentally to remove the instabilities.

The written history of bias-circuit design is almost nonexistent, inasmuch as very little has been known about the low-frequency impedance of IMPATT diodes. The existence of a low-frequency negative resistance effect was observed by Clorfeine and Hughes,[1] and they correctly suggested that this was induced by large RF voltages through a rectification effect and postulated that it was the cause of bias-circuit oscillation problems. They did not offer any analytic description of the effect, nor did they present a systematic characterization of it.

The construction of bias circuits has been done rather empirically, with the understanding that higher-impedance bias circuits give less upconverted noise and bias-circuit oscillations. Olson[2] disclosed the use of a resistor in the center conductor of a coaxial cavity to lower noise and remove instabilities, and others have used transistor amplifiers to raise the bias-circuit impedance. Typical examples of bias-circuit instabilities are given in an Application Note[3] published by Hewlett-Packard, along with a discussion of typical bias-circuit designs that they have used in silicon IMPATT oscillator circuits.

The rectification effect itself was contained in Read's original proposal,[4] and it has appeared in several large-signal analyses since then. What has been missing is a simple picture of how the microwave circuit completes the feedback loop to provide negative resistance and a more definite link between this negative resistance, bias oscillations, and tuning-induced burnouts.

The equivalent circuit and experimental characterization of this work permit the elimination of tuning-induced burnout, bias oscillations, and excessive noise upconversion in a straightforward manner.

Outlining the paper, in Section II, the link between burnout, bias oscillation, and negative resistance is discussed. In Section III, the physical origins are discussed and a simple analysis is given, demonstrating reasonable agreement with experiment. In Section IV, the

experimental characteristics of the total terminal impedance of the diode are given as functions of the bias current, baseband frequency, and microwave loading conditions. A small-signal equivalent circuit of the diode's terminal impedance is derived in Section V which is simple, yet describes in detail the role played by the microwave circuit. Stabilization, the elimination of the instabilities, is discussed in Section VI, giving the principles, some examples, and possible techniques. Approximate scaling rules are derived in Section VII.

Section VIII presents a summary of the paper and its major conclusions in a form complete enough to give the more casual reader a good understanding of the negative resistance mechanism and its stabilization.

II. BURNOUT AND BIAS-CIRCUIT OSCILLATION

Before proceeding with consideration of the negative resistance itself, it may be valuable to establish a link between the existence of a negative resistance, bias-circuit oscillations, and tuning-induced burnout. As described later in this paper, a negative resistance has been measured and found to exist in the circuit used at frequencies from essentially dc up to several tens of MHz. Such a negative resistance may act as an amplifier of bias-circuit noise, resulting in a modulation of the oscillator amplitude and frequency and producing the upconversion of large amounts of noise to the microwave frequency. Under certain conditions, the bias circuit may break into sustained oscillations producing a combined AM and FM spectrum about the microwave signal. When either of the above events occurs (excessive noise upconversion or bias-circuit oscillation), there is grave danger of burning out the diode. Specifically, it is found that

(*i*) without stabilization of the negative resistance, low-current burnouts are frequent,
(*ii*) bias-circuit noise and/or oscillations accompany or precede the burnout,
(*iii*) stabilization of the negative resistance removes the noise and the bias-circuit oscillations, and stops the low-current burnouts,
(*iv*) the microwave tuning conditions likely to produce burnouts are the same as those that produce the noise and bias oscillations.

Thus, circumstantial evidence strongly links the negative resistance, the bias-circuit oscillations, and the low-current tuning-induced burnouts to each other. A specific sequence of events leading to burnout

is not postulated. Indeed, many such sequences can be visualized. Preliminary scanning electron microscope photographs indicate no difference between pure thermal (nonoscillating) burnouts and low-current tuning-induced burnouts. Both have the same characteristic gold fingers shorting out the junction that were first described by Evans, et al.[5]

III. PHYSICAL ORIGINS OF THE NEGATIVE RESISTANCE

The three essential ingredients which cooperate to produce the low-frequency negative resistance are

(*i*) a large-signal rectification effect wherein the dc voltage (at constant current) is lowered by an increase in the microwave voltage amplitude,

(*ii*) the dependence of the diode's microwave conductance, $g_d(V_a,I_d)$, upon the microwave voltage amplitude, V_a, and the dc bias current, I_d, and

(*iii*) the microwave circuit constraint

$$g_d(V_a,I_d) + G(\omega) = 0$$

where $G(\omega)$ is the microwave circuit conductance into which the diode oscillates. The first of these, the rectification effect, is summarized by Read[4] in the equation (for the Read diode)

$$V_d = \frac{W\tau}{2\epsilon A} I_d - \frac{m}{4WE_c} V_a^2 + \text{constant} \tag{1}$$

where

V_d = dc voltage,

W = depletion layer width,

τ = drift zone transit time,

ϵ = dielectric constant,

A = junction area,

$m = (E/\alpha)(d\alpha/dE)\,|_{E_c}$ = percentage change in the avalanche coefficient α for a percentage change in the electric field, and

E_c = critical field for avalanche breakdown.

Equation (1) is approximate, neglecting higher powers of V_a. Studies made with the nonlinear program introduced by Blue,[6] however, indicate that (1) is an excellent approximation.

The net dc terminal resistance from (1) is then

$$R_t = \frac{dV_{dc}}{dI_d} = R_{sc} - \frac{mV_a}{2WE_c}\frac{dV_a}{dI_d} \tag{2}$$

where $R_{sc} = W\tau/2\epsilon A$ is the space-charge resistance.

From the microwave circuit constraint, we see that the ac voltage amplitude V_a and the dc bias current I_d are [for fixed $G(\omega)$] not independent of each other. In fact,

$$\frac{dV_a}{dI_d} = -\frac{\left(\frac{\partial g_d}{\partial I_d}\right)_{V_a}}{\left(\frac{\partial g_d}{\partial V_a}\right)_{I_d}}. \tag{3}$$

In the usual situation $(\partial g_d/\partial I_d)_{V_a} < 0$ and $(\partial g_d/\partial V_a)_{I_d} > 0$ (remembering that g_d is itself negative), so $(dV_a/dI_d) > 0$ and the last term of (2) represents a negative resistance. It is this negative resistance which causes the instabilities in the bias circuit. It is associated, basically, with amplitude modulation of the microwave oscillation. The rectification property is due to the nonlinearity of the avalanche process whereby a sinusoidal field variation about the critical field, E_c, produces many more charges on the positive swing than it does on the negative. This means that either the dc current increases, or the dc voltage drops, as the RF voltage increases.

A simple picture then is as follows. A positive fluctuation of the diode current increases the microwave negative conductance (in magnitude). This causes the voltage amplitude, V_a, to increase in order to meet the circuit constraint, and the increase in V_a requires a drop in the dc voltage, thereby creating negative resistance.

Equation (2) can be written

$$R_t = R_{sc} - \frac{m}{2WE_c}\frac{1}{G}\frac{dP_{\text{out}}}{dI_d} \tag{4}$$

where G is the microwave circuit conductance at the oscillation frequency and P_{out} is the microwave output power. This form is convenient because G can be estimated and dP_{out}/dI_d can be easily measured in any particular circuit. Doing this, and using published data for the other constants, R_t has been calculated for typical silicon, germanium, and gallium arsenide IMPATT diodes designed to operate at 6 GHz. This data is shown in Table I and it indicates reasonable agreement with the experimental values also shown. Silicon has the

TABLE I—PHYSICAL DATA AND NEGATIVE RESISTANCE VALUES FOR SILICON, GERMANIUM, AND GALLIUM ARSENIDE IMPATTS

	Si	Ge	GaAs	Source
$C_{\text{Breakdown}}$ (pF)	~0.7	~0.7	2.5	Measured
$V_{\text{Breakdown}}$ (volts)	105	45	95	Measured
W (microns)	5	4	4	Ref. 7, p. 117
E_c (volts/cm)	3.5×10^5	2.1×10^5	6.5×10^5	Ref. 7, p. 117
$\frac{E_c}{\alpha}\frac{d\alpha}{dE}\Big\vert_{E_c}$	5.0	6.1	19.5	Calculated
$\frac{dP_{\text{out}}}{dI_d}$ (watts/amp)	13.6	10	24.4	Measured
G (mhos)	3×10^{-3}	3×10^{-3}	6×10^{-3}	Estimated[8]
R_- (ohms)	65	117	152	
R_{sc} (ohms)	25	59	25	Calculated
Theoretical R_t (ohms)	−40	−58	−127	Calculated
Experimental R_t (ohms)	−12	−50	−122	Measured

smallest negative resistance and the largest discrepancy between theory and experiment. Germanium has about (−)50 ohms of induced negative resistance, indicating that difficulties might be encountered if biased through a 50-ohm bias line. The gallium arsenide results indicate about (−)120 ohms induced negative resistance. Measured values have been everywhere between (−)85 ohms and (−)150 ohms. On the basis of this data alone, it would be expected that biasing GaAs diodes would be very difficult.

A complete experimental characterization of the low-frequency impedance is given in the next section.

IV. EXPERIMENTAL CHARACTERIZATION

The technique used to measure the low-frequency impedance of the diode was to induce bias-circuit oscillations and to barely quench them by adjusting the bias-circuit impedance. Then the diode was removed from the circuit and the impedance of the circuit as seen by the diode was measured at the frequency of the bias-circuit oscillation. The quenching impedance, Z_Q, measured in this way, is the negative of the small-signal impedance of the diode, Z_t, and could be measured as a function of the bias-circuit oscillation frequency, the microwave loading, and the bias current. The microwave frequency of oscillation was always tuned to that for maximum power output.

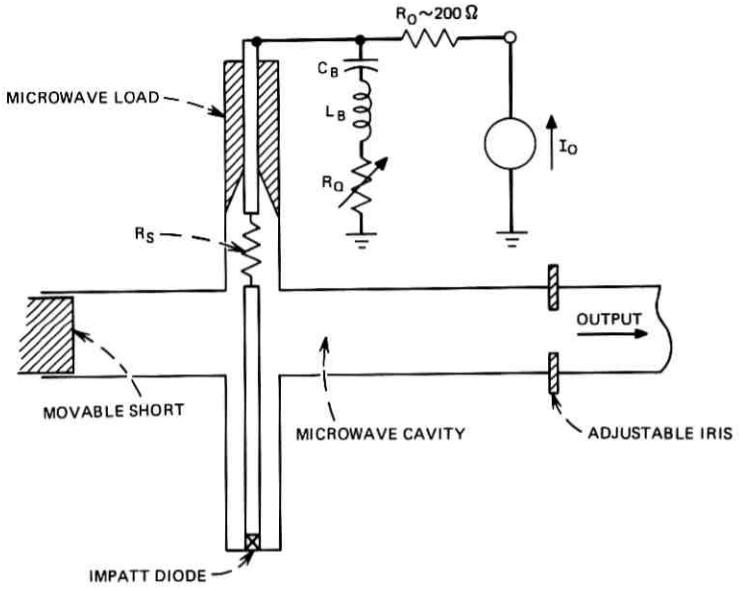


Fig. 1—Microwave cavity and bias configuration. The IMPATT diode is mounted at the end of a coax line which is coupled to the microwave cavity. The bias circuit includes a stabilizing resistor R_s, a resonant circuit C_B, L_B, R_Q used to induce bias-circuit oscillations, and a resistor R_o to isolate the bias circuit from the constant current supply I_o.

In order to perform this experiment, it was necessary to use an oscillator and bias circuit with the following two properties: (*i*) a bias circuit which could be made stable or unstable at all values of current and microwave loading, and (*ii*) microwave loading which could be varied continuously from above oscillation threshold down to zero external loading. A suitable circuit is one described by Magalhaes and Kurokawa[9] and illustrated in Fig. 1. It consists of a waveguide microwave cavity coupled to a coaxial transmission line. The waveguide cavity is formed between a movable short and an adjustable iris, which is used to couple the cavity to the output waveguide load. The adjustable iris is formed by rotating the rectangular waveguide cavity with respect to the output rectangular waveguide at a flanged joint.

The diode is situated at one end of the coaxial line (its distance from the waveguide was adjustable for tuning purposes) and the coaxial line is terminated on the opposite side of the cavity from the diode in a microwave load. The microwave load is an open circuit to dc and may be represented as a shunt capacitance of about 35 pF at bias-

circuit frequencies. The dc bias is fed in through the coaxial line from a constant-current supply. An isolation resistor of about 200 ohms was used to reduce the effects of the parasitic reactances of the power supply and its leads on the bias-circuit impedance. It was found that placing a series resistance R_s (2-watt carbon resistor) in the bias line would stabilize the induced negative resistance formed in the diode. A value of $R_s = 150$ ohms was sufficient to stabilize the diode at all bias currents up through the thermal limit of the resistor R_s and at all microwave loading levels.

The bias oscillation was induced by reducing R_s and adding a shunt resonant circuit with a quenching resistor R_Q. The resonant circuit

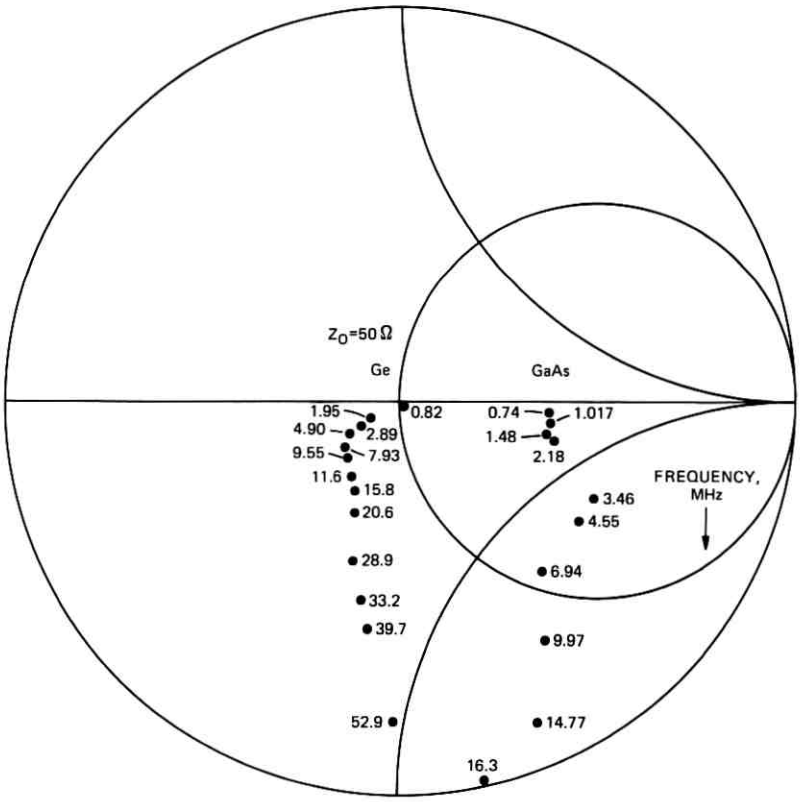


Fig. 2—Quenching impedance versus frequency of bias-circuit oscillations for Ge and GaAs measured at the diode's terminals. Bias currents were 100 mA and 120 mA for Ge and GaAs respectively. The quenching impedance is the negative of the diode's small-signal impedance, indicating a negative resistance and inductive reactance.

was used to tune the bias oscillation to the desired frequency and it was brought to the quenched, or small-signal, level by adjusting R_Q. Measuring the circuit impedance with the diode removed then gave the desired measurement of Z_Q.

The result of such a procedure at constant bias current and constant microwave loading conditions is shown in Fig. 2 for typical 6-GHz Ge and GaAs diodes. Since $Z_t = -Z_Q$, we see from Fig. 2 that the diode impedance has a general low-pass characteristic for the negative real part, becoming positive at frequencies higher than some frequency f_M. It is also seen that the reactive part of Z_t is inductive. This general character to Z_t has been observed in all diodes upon which these measurements have been made. The frequency f_M is the maximum frequency at which the bias circuit can be made to oscillate and is dependent upon the loaded Q of the microwave circuit, as will be explained in Section V. The scatter in the data from smooth contours is probably due to the difficulty of obtaining identical microwave tuning on successive measurements. The general range of low-frequency asymptotes for R_t for GaAs was from $(-)85$ ohms to $(-)150$ ohms. For Ge this range was from $(-)35$ ohms to $(-)75$ ohms, although less data is available for the Ge diodes. These experiments have been attempted on Si 6-GHz IMPATTs as well, with the results indicating

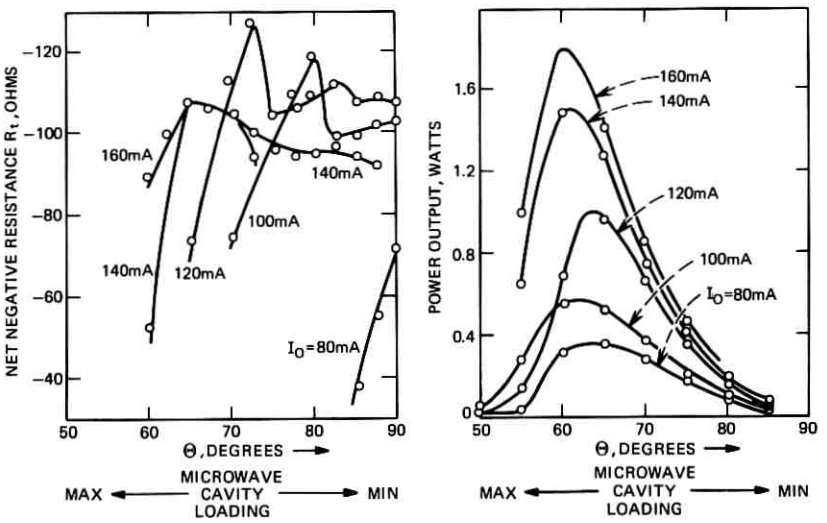


Fig. 3—Output power and induced net small-signal negative resistance as a function of bias current and microwave loading for GaAs. Microwave frequency was 4.7 GHz, bias-circuit frequency was ~5 MHz.

a maximum observed negative resistance of about (−)20 ohms. However, with the microwave circuit tuned to best overall operation, these 6-GHz Si diodes would not sustain bias-circuit oscillation, but only emit large amounts of baseband noise as the bias-circuit impedance was lowered. This is explained in Section V, where the equivalent circuit of this impedance is developed.

In Fig. 3, the small-signal-induced negative resistance, R_l, is shown as a function of bias current and microwave loading. This data was all taken at a bias-oscillation frequency of ~5 MHz. Also shown is the output power. The microwave loading is changed by rotating the output waveguide relative to the cavity waveguide, the angle between them denoted by Θ. For $\Theta \rightarrow 0$, the guides are aligned and the microwave cavity has its maximum loading, which implies zero (or at least minimum) microwave voltage amplitude. Increasing the rotation Θ, one passes through oscillation threshold to maximum power output and finally to maximum oscillation amplitude and zero output power in the minimally loaded condition. Thus, Θ may be considered to be an indication of microwave voltage amplitude or microwave load conductance. From Fig. 3 we make the following observations.

(*i*) At low bias currents, the negative resistance is small at normal maximum power loading conditions, but does become significant at the lightest loading ($\Theta \rightarrow 90$ degrees).

(*ii*) The negative resistance increases rapidly with bias current until it saturates at a level around (−)120 ohms (for this diode). Further increases in current tend only to translate the maximum negative resistance toward greater microwave loading conditions (smaller Θ), with the result that, at 160 mA, the loadings for maximum power and maximum induced negative resistance are nearly coincident.

(*iii*) There appears to be considerable structure to the negative resistance curves. This structure appears to be real because a decrease in either the current or the angle Θ from the condition $I_o = 140$ mA, $\Theta = 75$ degrees results in an increased amplitude of oscillation and sometimes burnout of the diode.

The reason the negative resistance increases with current at low currents and increases with Θ at the lower Θ is presumably associated with an increase of the voltage V_a. The apparent saturation of the negative resistance at higher currents and larger Θ is not understood. It may be associated with a saturation of the RF voltage V_a, or more

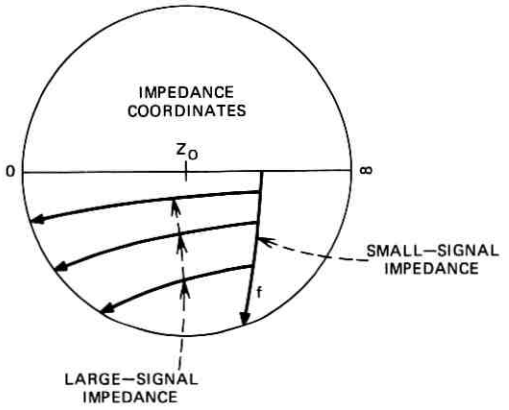


Fig. 4—Schematic illustration of the negative of the small-signal and large-signal diode impedance at bias-circuit frequencies. The large-signal impedances move in the direction of decreasing negative resistance with increasing bias oscillation amplitude.

likely with the detailed dependence of the diode's microwave admittance, y_d, on V_a and I_d at high V_a and high I_d.

One further experimental characteristic is that, if the bias oscillation is allowed to rise above the quenching level, the real part of Z_Q decreases as shown schematically in Fig. 4. This implies that the device impedance Z_l is open-circuit stable and is properly characterized as a negative resistance. It also explains the well-known fact that high-impedance termination of the bias circuit lowers the AM noise significantly.[10] Since the negative resistance acts as an amplifier of noise in the bias circuit, the lower the circuit impedance, the higher the gain and the larger the noise modulation of the microwave power become. In the following section, the small-signal equivalent circuit is worked out in detail and most of the observed features are explained on theoretical grounds.

The open-circuit stability and large-signal characteristics have a profound significance as to the elimination of both bias-circuit oscillations and diode burnout, as will be discussed at length in Section VI.

V. LOW-FREQUENCY EQUIVALENT CIRCUIT

It is evident by now that the microwave oscillator circuit plays an important role in the existence of the low-frequency negative resistance. The question arises, however, as to the role it plays in determining the bandwidth and magnitude of the negative resistance and why the reactance associated with R_l is always inductive. In this section

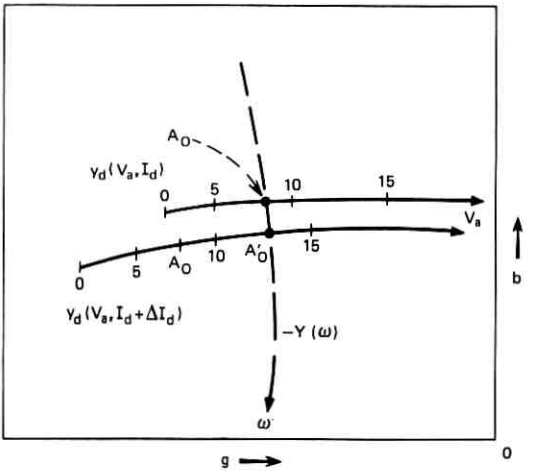


Fig. 5—Schematic admittance-plane description of the microwave large-signal device admittance, $y_d(V_a,I_d)$, and the negative of the microwave circuit admittance, $-Y(\omega)$. The intersection of y_d with $-Y$ determines the amplitude of the microwave oscillation A_o and the frequency ω. An increase in I_d requires an increase in V_a from A_o to A'_o as shown.

an equivalent circuit is developed, on an analytical basis, which fits the data of the previous section, and is used further in Section VI to consider the problems of stabilization.

To develop the equivalent circuit, a quasi-static model is used in which it is assumed that the diode's microwave admittance, $y_d(V_a,I_d)$, is an instantaneous function of V_a and I_d. For example, a sudden change in the dc bias current is assumed to shift the entire large-signal diode characteristic as shown in Fig. 5 from $y_d(V_a,I_d)$ to $y_d(V_a, I_d + \Delta I_d)$. Because of the large amount of energy stored in the microwave circuit, however, the RF voltage V_a cannot change instantaneously. Therefore, immediately after the change in bias current, the circuit constraint $y_d(V_a,I_d) + Y(\omega) = 0$ is not satisfied and can only be satisfied by the growth of V_a from $V_a = A_o$ to a new value A'_o as shown in Fig. 5. Such growth takes time and the amount of delay is related to the bandwidth of the microwave cavity. We can easily imagine that, if I_d were to fluctuate rapidly enough, faster than the circuit can respond, then V_a would maintain its average value and the induced negative resistance would disappear.

Using the quasi-static approach mentioned, the low-frequency (baseband) impedance, Z_t, of an IMPATT diode is derived in the

Appendix. It is found that

$$Z_t = R_{sc} - \frac{R_-}{1 + j\omega/\gamma} \tag{5}$$

with the definitions

$$R_- = \frac{m}{4WE_c} \frac{\sin \chi}{\sin \theta} \left| \frac{dV_a^2}{dI_d} \right| \tag{6}$$

and

$$\gamma = \frac{\sqrt{r^2 + s^2}}{\left| \frac{\partial Y}{\partial \omega} \right|} G \sin \theta . \tag{7}$$

Here, r and s are saturation parameters associated with the large-signal behavior of the diode admittance y_d as defined in eq. (17) of the Appendix, and $Y \equiv G + jB$ is the admittance of the microwave circuit. The angles χ and θ are defined in Fig. 6. In the admittance plane plot of this figure, $-Y(\omega)$ is the locus of the *negative* of the microwave circuit admittance in the neighborhood of the oscillation frequency ω_o; the arrow indicates the direction of increasing frequency. The large-signal admittance of the diode at frequency ω_o is assumed to intersect the $-Y(\omega)$ locus at the point P, which in fact determines the oscillation amplitude, V_a, and the frequency, ω_o. The line $y_d(V_a)$ indicates

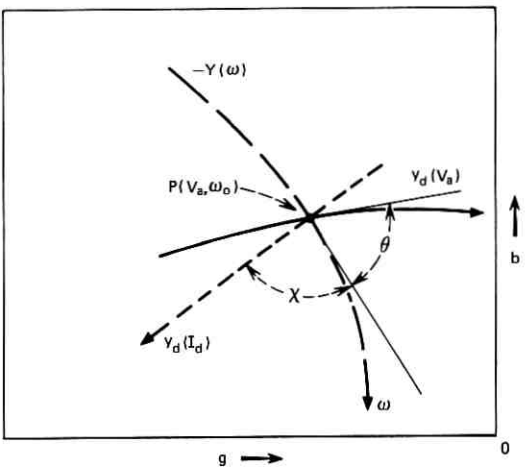


Fig. 6—Illustrating the definitions of the angles θ and χ at the intersection of the device admittance $y_d(V_a)$ and the circuit admittance $-Y(\omega)$. The intersection angle is θ for microwave voltage variations and is χ for bias-current variations. The angle $\theta + \chi$ is determined by the device alone.

the direction of the admittance plane in which the diode's admittance moves for increased voltage, V_a, at constant current. The line $y_d(I_d)$ is likewise the direction in which the diode's admittance moves for increased I_d at constant oscillation amplitude, V_a. The angles θ and χ give the intersection angles of $y_d(V_a)$ and $y_d(I_d)$ with the locus $-Y(\omega)$ as shown in the figure. It is well known[11] that one of the conditions for oscillation at the point P is that $0 \leqq \theta \leqq \pi$, with $\theta = \pi/2$ generally giving the best system performance (highest oscillator stability, lowest noise).

From (5) and (6) we see that at very low frequencies ($\omega/\gamma \ll 1$) and if $\sin \chi/\sin \theta = 1$ (a reasonable condition, as will be seen) then $Z_t = R_{sc} - R_-$ with R_- the same as found on a dc basis in eq. (2). We see that the space-charge resistance is a stabilizing factor, tending to lower the net negative resistance. At low enough frequencies, thermal effects become important and introduce further positive resistance. The latter is neglected here due to the relatively low cutoff frequency for thermal effects compared to γ.

Considering eq. (6) for R_-, we see that, if θ approaches either 0 or π, R_- can become very large. In fact, $\theta = \pi/2$ gives a minimum R_-, everything else being constant Intuitively, one can see why $\theta = 0, \pi$ are critical directions. For these conditions, $y_d(V_a)$ is nearly parallel to $-Y(\omega)$ and slight changes in the current I_d can cause large changes in V_a (and the frequency ω_o as well).

This dependence on θ is consistent with experimental observations. A certain set of Si diodes (6 GHz) were found not to have a large enough negative resistance to cause bias-circuit oscillations when tuned in what was considered to be the optimum manner. Detuning the cavity, however, and readjusting the power output to the same value gave very large bias-circuit oscillations with a very definite FM microwave spectrum. It is presumed that the FM is mostly due to θ being different from $\pi/2$.

Separately considered, the angle χ could be used to minimize R_- or even to make the net resistance positive. The difficulty with this is that the total angle $\theta + \chi$ is determined totally by the diode's characteristics and has nothing at all to do with the microwave circuit. The large-signal theories of Scharfetter and Gummel[12] and Blue[6] may be used to show that the angle $\theta + \chi$ is close to π over most of the useful range of diode parameters (V_a, I_d, ω_o). Comparing Figs. 5 and 6 of the paper by Gewartowski and Morris[8] confirms this conclusion experimentally, at least for the Si diode they investigated. If $\theta + \chi \cong \pi$ and $\theta \cong \pi/2$, then $\chi \cong \pi/2$ and $\sin \chi/\sin \theta \approx 1$ for best system require-

ments. This dependence of R_- on the angle χ indicates, however, that in laboratory testing (providing $\chi + \theta$ is not actually π) it should be possible to reduce the value of R_- and inhibit bias-circuit oscillations, while at the same time maintaining high power output. Experimentally, in coaxial circuits with many degrees of freedom in the tuning it is found that burnout and bias oscillations can usually be avoided by very carefully increasing the current by small increments and retuning slightly.

The parameter γ is shown in the Appendix to be one-half the 3-dB modulation bandwidth of the oscillator. γ achieves its maximum value at $\theta = \pi/2$. We therefore see from eq. (5) that the negative resistance has a low-pass characteristic and that $Re[Z_t - R_{sc}]$ has a cutoff frequency $\omega_c = \gamma$ equal to that of the microwave circuit.

The equivalent circuit of Z_t from eq. (5) is shown in Fig. 7. Z_t is inductive, but the frequency-independent representation involves a negative capacitance $-C_- = -(1/\gamma R_-)$. We note that, in accordance with the cutoff frequency mentioned above, the product $(-R_-)\cdot(-C_-) = +1/\gamma$ gives the time constant associated with the microwave oscillator bandwidth. A Smith chart plot of $(-)Z_t/R_{sc}$ for several values of R_-/R_{sc} and with $x = \omega/\gamma$ as a normalized frequency parameter is shown in Fig. 8. In this figure we see that the loci of $-Z_t/R_{sc}$ are straight vertical lines, and that the $x = \omega/\gamma$ values are obtained by drawing straight lines between the points $z = 0 - jx$ and $|z| = \infty$. For any value of R_-/R_{sc}, there exists a maximum frequency for which Z_t has a negative real part. We call this the maximum frequency of

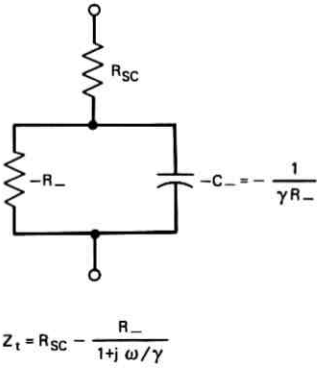


Fig. 7—Low-frequency equivalent circuit of an oscillating IMPATT diode neglecting thermal effects. R_{sc} is the space-charge resistance of the diode, γ and R_- are defined in the text.

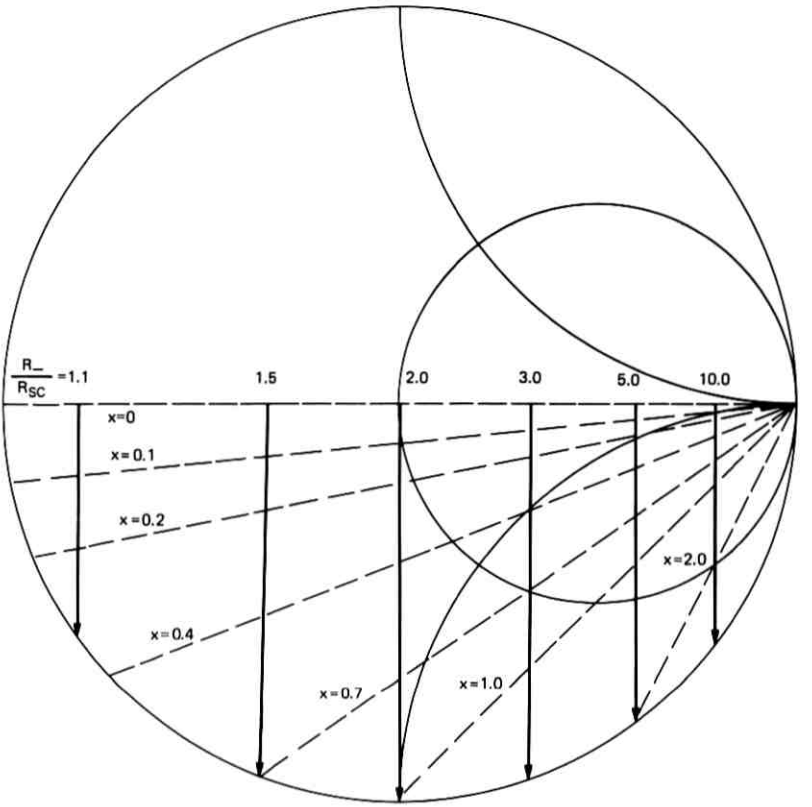


Fig. 8—The negative of the equivalent circuit impedance, Z_t/R_{sc}, normalized to the space-charge resistance, versus normalized frequency $x = \omega/\gamma$. The straight line characteristics of this plot make it useful in estimating the ratio of f_M from the values of R_-, R_{sc}, and γ.

oscillation, f_M, and it is given by

$$f_M = \frac{\gamma}{2\pi}\sqrt{\frac{R_-}{R_{sc}} - 1}. \tag{8}$$

Thus, the maximum frequency of oscillation may be greater than or less than $\gamma/2\pi$, depending on the ratio R_-/R_{sc}.

The normalization of Z_t to R_{sc} is convenient for simple construction of the loci and finding the ratio $2\pi f_M/\gamma$, but experimental results are more appropriately normalized to the bias line characteristic impedance as was done in plotting Z_Q in the previous section (Fig. 2). There are three independent parameters in the equivalent circuit for Z_t, namely,

R_-, R_{sc}, and γ. By measuring the impedance Z_t at three frequencies, one may calculate these parameters. Usually, R_{sc} can be measured in dc tests, so that only two measurements of $Z_t(\omega)$ are required. In Fig. 9, the data for the GaAs diode of Fig. 2 have been replotted along with the equivalent circuit curve obtained from the data $R_t(\omega = 0) = R_{sc} - R_- = -110\ \Omega$, $f_M \cong 17$ MHz, and $R_{sc} = 33.6\ \Omega$. The scatter in the data points is, as was mentioned before, due to the difficulty of disassembling the microwave circuit and reassembling it to the same RF conditions after every measurement. Under the circumstances, the agreement is considered to be good.

To illustrate the meaning of γ further, consider the simplest model

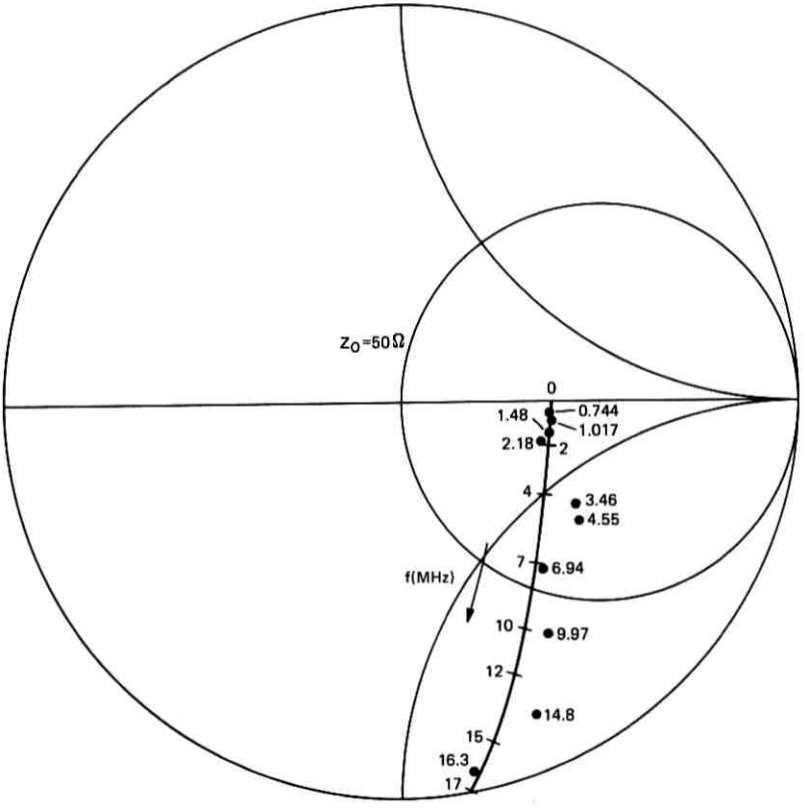


Fig. 9—Comparison of experimental (solid dots) and theoretical (solid curve) frequency dependence of the small-signal impedance Z_t. GaAs diode, $I_o = 120$ mA, $f_o = 4.7$ GHz, output power = 1 watt, $R_- = 143.6\ \Omega$, $R_{sc} = 33.6\ \Omega$, and $\gamma = 59 \times 10^6$ radians per second.

of oscillator and circuit possible in which only the diode's conductance is a function of V_a [this puts $r = 0$ in eq. (7)] and the circuit conductance does not depend upon frequency. Then eq. (7) becomes

$$\gamma = \frac{sG}{\left(\frac{\partial B}{\partial \omega}\right)} = \frac{s\omega_o}{2Q_L} \tag{9}$$

where Q_L is the loaded cavity Q. Since at maximum output power $s \cong 2$,[13] we can write under conditions for maximum power

$$\gamma \cong 2\pi \cdot \Delta f \tag{10}$$

where Δf = the full 3-dB bandwidth of the loaded cavity. Equation (8) then becomes

$$f_M = \Delta f \sqrt{\frac{R_-}{R_{sc}} - 1} = \frac{f_o}{Q_L} \sqrt{\frac{R_-}{R_{sc}} - 1}. \tag{11}$$

This dependence of f_M on Q_L was tested by making the experimental waveguide cavity length $\sim 3\lambda_g/2$ instead of $\sim \lambda_g/2$ and finding the maximum frequency of oscillation. This should raise Q_L by a factor of three and therefore decrease f_M by a factor of one-third. The measured result was a decrease in f_M by a factor of 0.41. This discrepancy may again have been due to not achieving the same RF conditions in the two cases.

In another test of eq. (11), the loaded cavity Q was measured after having made the bias-circuit impedance measurements necessary to determine f_M, R_{sc}, and R_- for the same microwave circuit conditions. The results were $\gamma = 113.3 \times 10^6$, $R_{sc} = 35\ \Omega$, $R_- = 119\ \Omega$ which, together with the assumption $s = 2$, give $Q_L \cong 261$. The microwave measurement yielded $Q_L \cong 268$, in good agreement.

We conclude that eq. (5) is a good representation of the terminal impedance at baseband frequencies in a large-signal IMPATT diode oscillator and that γ is well approximated by either (9) or (10).

VI. BIAS-CIRCUIT STABILIZATION

Up to this point, the major emphasis has been on characterizing the induced negative resistance and discussing its role as the cause of bias-circuit oscillations and tuning-induced burnout. We now turn our attention toward the principles and techniques of the stabilization of this negative resistance. By stabilizing we mean the achievement

of small-signal stability so that any fluctuation in bias-circuit current or voltage eventually decreases to zero. In addition to small-signal stability, we also require large-signal stability so that transients in power supply lines cannot cause instabilities to develop.

6.1 *Stability Criterion*

The stability criterion is derived by starting with the loop impedance equation

$$(Z_B + Z_t)i_B = e_n$$

where Z_B is the bias-circuit impedance, i_B the fluctuating component of current, and e_n is an assumed noise voltage. In the complex frequency plane ($s = \sigma + j\omega$), if $Z_B + Z_t$ has any zeros in the right-half plane ($\sigma > 0$), that component will grow in time, and the circuit plus diode is unstable. The Nyquist criterion is used to determine the number of zeros minus the number of poles of $Z_B + Z_t$ that are in the right-half plane. This is not so convenient here, since we are working with Z_t and Z_B as two separate sets of data and can only change the design of Z_B. We therefore consider the function $Z_Q - Z_B$, where $Z_Q = -Z_t$. This must have the same number of poles and zeros in the right-half plane. Now, however, we plot the loci of Z_Q and Z_B separately as ω goes from $-\infty$ to $+\infty$ and consider the vector pointing from $Z_B(\omega)$ to $Z_Q(\omega)$. We then see that the net number of rotations of the vector $Z_Q(\omega) - Z_B(\omega)$ indicates the number of zeros minus the number of poles in the right-half plane. In addition, since Z_B is a passive driving point impedance function, it cannot have any poles or zeros in the right-half plane. Therefore,

$$Z_Q(s) - Z_B(s) = \frac{R_-}{1 + s/\gamma} - R_{sc} - Z_B(s)$$

has no poles in the right-half plane and the number of rotations of the impedance vector $Z_Q(\omega) - Z_B(\omega)$ gives directly the number of right-half plane zeros.

To understand this in practical terms, we show schematically in Fig. 10 three conditions; stable, unstable, and conditionally stable for a typical bias circuit. The bias-circuit impedance, $Z_B(\omega)$, and the diode quenching impedance, $Z_Q(\omega)$, are separately plotted as a function of frequency over the complete frequency range for which $Z_t(\omega)$ has negative real part. The criterion then becomes: The bias circuit is stable if, at the point P where the two loci intersect, the frequency ω_{PQ} on the Z_Q locus is lower than the frequency ω_{PB} on the Z_B locus;

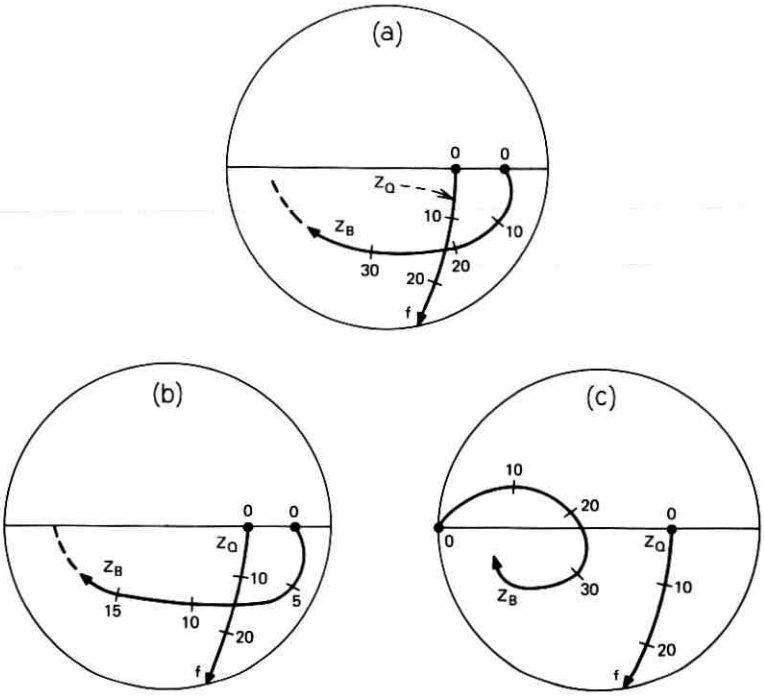


Fig. 10—Illustration of (a) stable, (b) unstable, and (c) conditionally stable bias-circuit impedance loci.

otherwise, an instability exists. In Fig. 10a, the criterion is satisfied, and the circuit is stable. In Fig. 10b, the criterion is not satisfied, and the circuit is not stable. In Fig. 10c, there is not even an intersection of Z_B with the small-signal locus Z_Q. However, $Z_B(\omega)$ does lie in the large-signal region of Z_Q (see Fig. 4), and such a condition is conditionally stable. At small-signal levels, no growing root appears, but if the excitation becomes large through some disturbance, then a large-signal instability can exist. We note that changes in the bias current or microwave loading can move Z_Q all the way down to zero impedance and therefore such a circuit configuration as shown in Fig. 10c would be almost certain to burn out the diode.

6.2 *Some Examples*

We now give some simple examples which illustrate the above criterion and shed light on some commonly used biasing schemes.

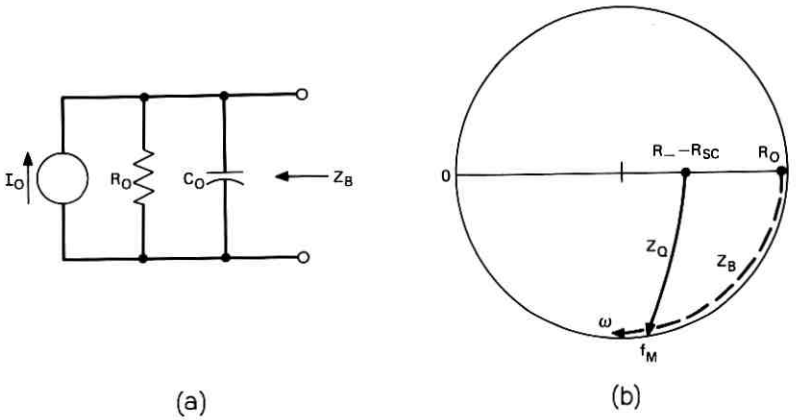


Fig. 11—Stability considerations for constant-current power supply having a lumped shunt capacitance C_o: Section 6.2.1.

6.2.1 *Constant-Current Supply—Maximum Equivalent Lumped Capacitance*

The circuit of Fig. 11a represents a constant-current supply with large shunt resistance and perhaps large shunt capacitance. Part (b) of the figure shows the impedance diagram assuming that $R_o \to \infty$. The maximum capacitance $C_{0\,\max}$ that can be tolerated before inducing instability is that which has a reactance equal to that of Z_Q at $f = f_M$, the maximum oscillation frequency. This is

$$C_{0\,\max} = \frac{1}{\gamma}\frac{1}{R_- - R_{sc}}.$$

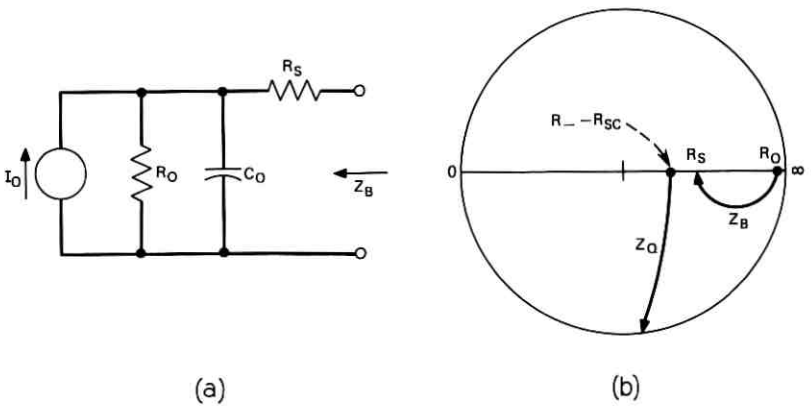


Fig. 12—Stability considerations for constant-current supply with added series resistance, R_s: Section 6.2.2.

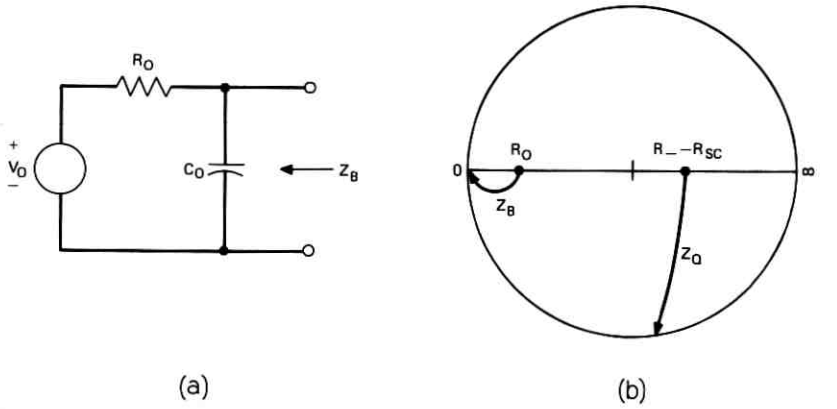


Fig. 13—Stability considerations for constant-voltage supply: Section 6.2.3.

For the GaAs oscillators used for the data in Figs. 2 and 3, we had $R_- - R_{sc} \cong 120$ ohms and $\gamma \cong 2\pi \times 10^7$ rad/s, which gives $C_{0\,\max} = 140$ pF.

6.2.2 *Constant-Current Supply—Series Resistance*

The addition of series resistance between the shunt capacitance and the diode, as in Fig. 12a, gives the impedance diagram shown in Fig. 12b. Stability is predicted for $R_s > R_- - R_{sc}$. It should be noted here that this is the easiest possible way to stabilize the bias circuit, and were it not for the large dc power dissipation incurred in R_s, this would be ideal. The larger R_s is, the larger the difference between $Z_B(\omega)$ and $Z_Q(\omega)$, and the less the noise amplification is in the bias circuit, giving less upconverted noise in the microwave spectrum.

6.2.3 *Constant-Voltage Supply*

For a constant-voltage supply, the series resistance is usually very small, and C_0 is large. Figure 13 shows the circuit and impedance diagram and we see that this is a conditionally stable circuit. As the bias current is turned up from zero, the Z_Q locus moves through the low-impedance region, and this scheme is certain to burn out diodes. The addition of a large series resistance moves the whole Z_B locus to the high-impedance region and stability is indicated again.

6.2.4 *Constant-Current Supply—Series Inductance*

The addition of a large series inductance, as shown in Fig. 14, might be thought to hold the current more constant and therefore add stability. In fact, such an addition creates a series resonance which lowers

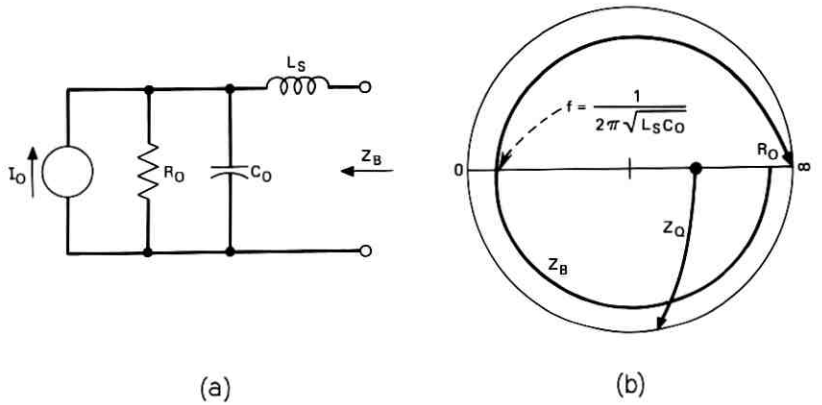


Fig. 14—Stability considerations for constant-current supply with added series inductance: Section 6.2.4.

the bias-circuit impedance at all frequencies below the resonance. Thus, the $Z_B(\omega)$ locus must intersect the $Z_Q(\omega)$ locus at a lower frequency (on Z_B), and this is in the direction of making the circuit unstable. Thus a series inductance worsens stability. Of course, some inductances have rather large series resistances, and this may, in itself, tend to improve stability.

6.2.5 *Constant-Current Supply—Resistive Choke Stabilizer*

The need to get rid of the dc loss in the stabilizing circuit forces one to consider parallel inductance-resistance networks as shown in Fig. 15. The constant-current supply is shown with parasitic shunt capacitance C_0 and series resistance R_s. R_s must be greater than $R_- - R_{sc}$ for this scheme to work, but it is used only to stabilize the lower frequency range, and may therefore be put further from the diode. Between R_s and the diode, there is additional shunt capacitance denoted by C_1. To stabilize the effects of C_1, a parallel inductance-resistance network can provide the locus Z_B shown (approximately) and therefore stabilize the oscillator. A simple design criterion may be formulated by assuming R_s to be very large. By choosing L/R large enough, the locus of Z_B may be made to resemble that of the pure series resistance, shunt capacitance case as closely as desired. From a graphical evaluation, the criterion $L \approx 3C_1R^2$ together with $R = 2$ to 5 times $R_- - R_{sc}$ gives an adequate margin of stability. For the GaAs oscillator circuit, an experimental model of this choke was made with $R = 470$ ohms, and $L = 75\,\mu$H. This would satisfy the above criterion for a capacitance

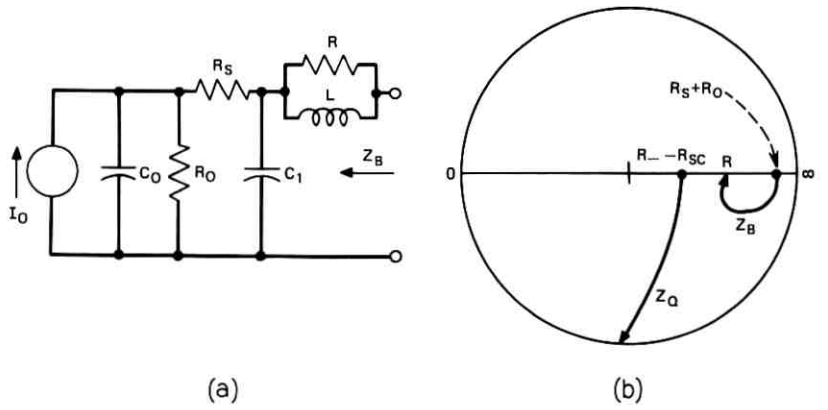


Fig. 15—Stability considerations for constant-current supply with resistive choke stability network: Section 6.2.5.

of 100 pF. In actual fact, it was possible to make $C_1 = 3000$ pF without causing instability. With $C_1 = 10{,}000$ pF, the diode burned out. The technique used to construct the choke is illustrated in Fig. 16. The body of the choke is made of a ferrite material similar to that used in low-frequency inductors. A hole is drilled through the center and a 470-ohm $\frac{1}{8}$-watt carbon resistor inserted. The winding around the ferrite is made of three layers of number 35 FORMVAR wire. Attached

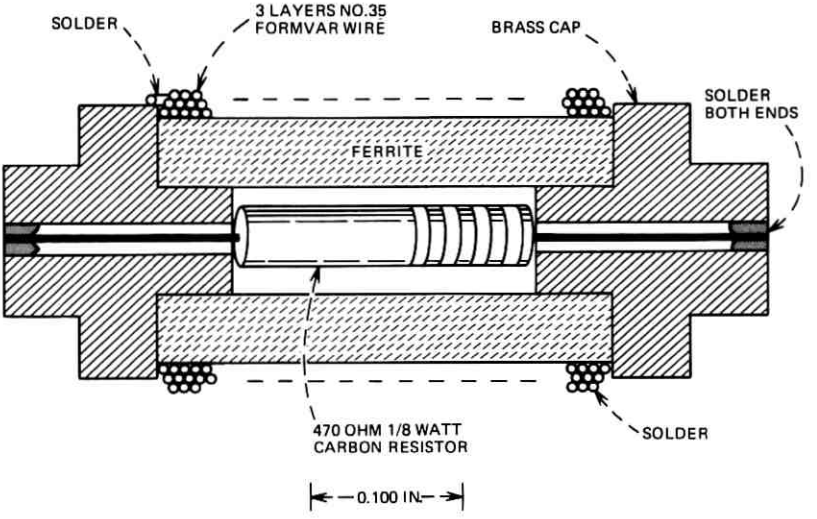


Fig. 16—Mechanical construction of the resistive choke stabilizing network.

to the ends are brass caps which thread into the center conductor of the coaxial bias line. This choke is placed at the same position indicated for R_s in Fig. 1. A broader bandwidth might be achieved by lowering the resistance and inductance values, with some sacrifice of stability margin.

In the particular model built, no attempt was made to insure low reflections at microwave frequencies from the bias circuit when the choke was inserted. It would be desirable to do this for a practical oscillator.

With this resistive choke stabilizing network, a GaAs IMPATT oscillator was made to give 3.4 watts of output power with 12 percent efficiency at 250 mA bias current. Without any stabilization, diodes of this same type (breakdown capacitance $\sim$2.5 pF, breakdown voltage $\cong$105 volts) consistently burned out at 80 to 100 mA. This network has vastly improved the stability of the bias circuit, but provides lossless dc power transmission from the constant-current supply to the diode.

Using a more sophisticated biasing network, designed upon the criteria developed here, but which provides a more controlled impedance locus than achievable with a simple RL choke, it has been possible to stabilize the bias circuit over all current ranges up through thermal burnout of the diode, without loss of RF or dc power.

A word of warning about the general use of chokes for stabilization: if the inductance is not large enough, such a choke actually worsens the instability and hastens burnout.

6.2.6 *Bias-Circuit Line Length*

In the previous examples it has been assumed that the terminations specified were all positioned exactly at the diode. In fact, this cannot be true and, therefore, a finite length of transmission line is necessary between the diode and the termination. This effect causes the termination impedance to be transformed to a lower impedance, and increases the likelihood of an instability. This places a limit on the length of line that can be tolerated between the diode and the stabilizing network. For example, for the GaAs data of Fig. 2, and a pure stabilizing resistance of 500 ohms, the maximum length of line is about $l_{max} = 0.09\,\lambda$ at a frequency of about 13 MHz. But $\lambda = c/(f\sqrt{\epsilon_r})$, where ϵ_r is the dielectric constant of the material in the bias line. This gives about 2 meters for air line with $\epsilon_r = 1$. In common microwave absorbent materials such as Eccosorb MF 124, $\epsilon_r = 27$ and l reduces to 40 cm. This is for a medium Q oscillator, Q_L being $\sim$250. For a low-Q

oscillator ($Q_L \sim 25$) the frequencies all will scale with $1/Q$ so the line length then reduces (in the Eccosorb material) to ~4 cm, which is beginning to be difficult to do.

If the termination is frequency dependent, the effect of line length can be more pronounced. An example, a length l of transmission line terminated in a pure capacitance, is illustrated in Fig. 17. Since at $f = f_M$, $X_Q/Z_o = -1.3$ for the GaAs oscillator (see Fig. 2), the frequency at which $X_B/Z_o = -1.3$ gives the maximum f_M tolerable without initiating bias instabilities for each value of C. This value of f_M is plotted versus C for $l = 10$, 20, and 30 centimeters, and may be used to derive either (*i*) the maximum value of shunt capacitance, (*ii*) the maximum length l for a given C and f_M, or (*iii*) the smallest value of microwave loaded Q [eq. (11)], which is tolerable before oscillation or instability occurs. For example, with a capacitance

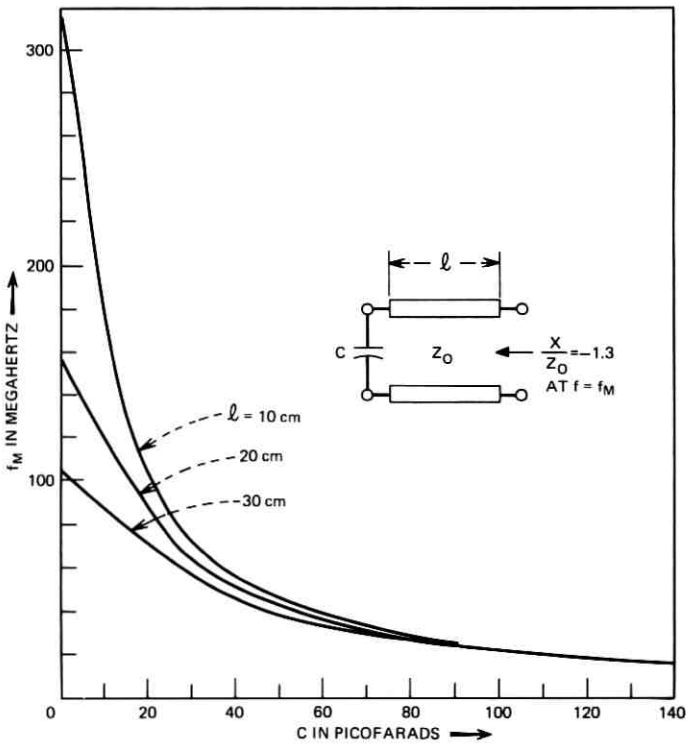


Fig. 17—Depicting the effect of a length of line, l, between the diode (having maximum bias-circuit oscillation frequency f_M) and a shunt capacity C, for the GaAs diode terminal impedance shown in Fig. 2.

$C = 50$ pF, $f_o = 6$ GHz, $l = 10$ cm, and $R_-/R_{sc} = 4.3$ we find $f_M \cong 47$ MHz and $Q_L = 232$. Thus the loaded Q would have to be greater than 232 in order that a 50-pF shunt capacitance at a distance of 10 cm from the diode would not cause instability in a 6-GHz GaAs oscillator similar to those used in these experiments.

6.2.7 *Higher-Impedance Bias Circuits*

If the bias-circuit characteristic impedance were made several times larger than the maximum induced value of negative resistance, then a matched termination at all frequencies would insure stable operation. Even small reflections could easily be tolerated, regardless of their distance from the diode. The difficulty with this is that it is very difficult to make good circuits with very-high-impedance lines. Traditionally, 50 ohms is used. Any increase above that value would materially assist the stabilization of the bias circuit. On the other hand, one could also decrease the impedance of the diode by making its area larger. In either case, difficulty will eventually be encountered because of the problem of matching the diode to the microwave circuit, but it is probable that some gain over the present situation can be had by these techniques.

VII. SCALING LAWS

The negative resistance of eq. (2) is obviously a function of the diode area and length, the maximum field strength, and the avalanche coefficient. It is therefore reasonable to inquire as to how R_- scales with different oscillator designs, material, and frequency. This can be worked out rather simply for the Read diode model assumed here, but for abrupt junction pn diodes this model fails to answer the question, how should the design of the diode be changed to decrease the value of R_-, while still achieving high-power, high-efficiency microwave operation?

For the Read diode, in the large-pulse approximation we may assume the negative conductance is approximately

$$g_d = -\frac{4}{\pi}\frac{I_d}{V_a}.$$

Then

$$\frac{dV_a}{dI_d} = -\frac{\left(\frac{\partial g_d}{\partial I_d}\right)_{V_a}}{\left(\frac{\partial g_d}{\partial V_a}\right)_{I_d}} = \frac{V_a}{I_d}.$$

Therefore,

$$\frac{I_d}{V_d}\frac{dV_a}{dI_d} = \frac{V_a}{V_d} \equiv \delta,$$

δ being the normalized ac voltage amplitude and V_d the breakdown voltage of the diode. The ratio δ is assumed here, as elsewhere,[14] to be invariant with frequency scaling, and with this assumption we find that R_- for a new, optimally designed diode, as compared with R_{-o} for a reference diode, is given by

$$\frac{R_-}{R_{-o}} = \left(\frac{m}{m_o}\right)\cdot\left(\frac{V_d}{V_{do}}\right)\cdot\left(\frac{I_{do}}{I_d}\right) = \left(\frac{m}{m_o}\right)\cdot\left(\frac{Z_d}{Z_{do}}\right). \tag{12}$$

Here Z_d is defined as the ratio of the breakdown voltage to the operating current and $W\cdot E_c$ has been set equal to V_d.

An interesting application of (12) is to the millimeter-wave Si IMPATTs designed to work at 100 GHz. For these diodes $V_d \cong 13$ volts, and a typical current is 0.1 ampere, giving $Z_d = 130$ ohms. For typical 6-GHz Si IMPATTs $V_d = 105$ volts, and a typical current is 0.20 ampere, giving $Z_{do} = 525$ ohms. Therefore,

$$\frac{R_-(\text{100-GHz diode})}{R_{-o}(\text{6-GHz diode})} \cong \left(\frac{m}{m_o}\right) \times 0.25.$$

It is found that R_{-o} is a few ohms and that $m/m_o < 1$, so it is predicted that the millimeter-wave IMPATT diodes should have very small induced negative resistance.

VIII. SUMMARY AND CONCLUSIONS

In this paper we have shown that a low-frequency negative resistance is induced in the IMPATT diode by the large-signal characteristics of the oscillator. This low-frequency negative resistance has been shown to be responsible for both bias-circuit oscillations and a class of low-current burnouts, normally called "tuning-induced" burnouts. The large upconversion of noise that often occurs is also caused by this same effect.

The origin of the negative resistance was shown to be the combination of the rectification property of the nonlinear ac avalanche, the coupling of fluctuations in the bias current to fluctuations in the microwave voltage amplitude by the microwave circuit oscillator constraint, and the fact that increases in the microwave voltage amplitude and dc bias current generally drive the large-signal microwave admittance of the diode in opposite directions on the admittance

plane. Calculations were made showing the correctness of this picture, and were applied to Si, Ge, and GaAs diodes designed for 6-GHz operation. The agreement with experimental results is good, showing a small value of negative resistance for Si (typically less than 10 ohms), somewhat larger for Ge (typically 35 to 75 ohms), and larger yet for GaAs (typically 85 to 150 ohms).

A reasonably complete experimental characterization of this negative resistance was given, showing its dependence on baseband frequency, dc bias current, and microwave loading conditions. It was found that, at low enough bias currents, the negative resistance was small for microwave loads near those which gave maximum power. At higher bias currents, however, the negative resistance approached its maximum value at or near optimum loading conditions. It was also found that the negative resistance had a low-pass frequency characteristic and an inductive reactance associated with it. The negative resistance was found to go to zero at a frequency called f_M, the theoretical maximum frequency of oscillation, and f_M was found experimentally to be roughly inversely proportional to the microwave circuit's loaded Q factor. It was found that, for large-signal bias-circuit oscillations, the negative resistance decreased, and the reactance remained approximately constant. This fact indicates open-circuit stability, and is very important in the stabilization of the bias circuit.

An analysis was performed and an equivalent circuit was derived, showing that the inductive reactance observed experimentally is best represented by a negative capacitance, $-1/\gamma R_-$, in parallel with the negative resistance, $-R_-$, all in series with the space-charge resistance, R_{sc}. This circuit predicts several aspects of the small-signal behavior of the induced negative resistance, and allows the stabilization techniques to be designed in a straightforward manner. The agreement of this equivalent circuit with the experimental results is good and it shows in detail how the interaction of the microwave circuit and diode admittances affect the low-frequency negative resistance. In particular, it is shown that the maximum frequency of oscillation, f_M, is given by

$$f_M = \frac{f_o}{Q_L}\sqrt{\frac{R_-}{R_{sc}} - 1}$$

where f_o is the microwave frequency of oscillation, Q_L is the loaded Q of the microwave cavity, R_{sc} is the space-charge resistance of the diode, and $R_- - R_{sc}$ is the low-frequency asymptote of the negative resistance. Smith chart plots of the negative of the diode's small-signal terminal impedance, $(-)Z_t$, are given for various values of

R_-/R_{sc} and normalized frequency. The shape of these loci is found to vary somewhat, though not drastically, with the choice of impedance normalization used. When normalized to 50 ohms (which is greater than R_{sc}), the contours always appear somewhat as shown in Fig. 2.

Stabilization techniques were discussed, beginning with the development of a simple means of judging the stability of a given oscillator and bias circuit. Several examples of stable and unstable conditions were given including:

(*i*) a calculation of the maximum equivalent lumped shunt capacitance that can still remain stable,

(*ii*) discussion of the effects of line length in the bias circuit which indicated the desirability of placing any stabilization network as close to the diode as possible, but at the same time indicating that it need not be exactly at the diode,

(*iii*) a stabilizing network consisting of a pure series resistance, R_s, in the bias circuit with $R_s \gg R_- - R_{sc}$ (it is required that there be no significant reactance or line length between the diode and R_s), and

(*iv*) an alternative stabilizing network consisting of a parallel inductance-resistance choke combination that does not dissipate dc power.

Using the principles of stabilization outlined here, it has been possible to stabilize the bias circuit of IMPATT oscillators over the full range of bias currents up to the diode's thermal burnout limit.

As a final topic, the frequency scaling of R_- was considered, and it was shown as an example that the 100-GHz Si IMPATTs should have very small induced negative resistances, and therefore the instabilities encountered at lower frequencies and in Ge and GaAs should not be a serious problem for the millimeter-wave Si diodes.

IX. ACKNOWLEDGMENTS

The author wishes to thank F. M. Magalhaes for early experimental contributions leading to the success of this investigation, and for providing the microwave oscillator circuits used in all of the experiments. The author also thanks K. Kurokawa for general guidance and especially for suggestions leading to the development of the equivalent circuit. The GaAs diodes were supplied by J. C. Irvin and D. J. Coleman, who first brought to the author's attention the severity of the problem in GaAs. The Si and Ge diodes were supplied by D. R. Decker and C. N. Dunn. Thanks also go to R. S. Riggs who performed a good many of the experiments reported here.

APPENDIX

It is the purpose of this Appendix to derive the equivalent circuit of the low-frequency diode impedance of an IMPATT diode oscillator. To do this, the Read diode model is assumed to describe the diode and a quasi-static technique is assumed to describe its interaction with the microwave circuit. Rewriting eqs. (8) and (9) of Kurokawa's work[11] to apply to admittance, and dropping the noise voltage terms, we may write

$$[G(\omega) - g]B'(\omega) - [B(\omega) + b]G'(\omega) + |Y'(\omega)|^2 \frac{1}{V_a}\frac{dV_a}{dt} = 0 \quad (13)$$

$$[G(\omega) - g]G'(\omega) + [B(\omega) + b]B'(\omega) + |Y'(\omega)|^2 \frac{d\varphi}{dt} = 0. \quad (14)$$

Here, $Y(\omega) = G(\omega) + jB(\omega)$ is the microwave circuit admittance, $y = -g + jb$ is the device admittance, V_a and φ are the microwave voltage amplitude and phase, and are assumed to be slowly varying functions of time, and the prime on G, B, and Y denotes differentiation with respect to frequency. Equations (13) and (14) describe the change in time of the amplitude and phase of a negative conductance oscillator. In steady-state oscillation, $dV_a/dt = 0$ and $d\varphi/dt = 0$, implying

$$G(\omega) - g(V_a,I) = 0, \quad (15)$$

and

$$B(\omega) + b(V_a,I) = 0, \quad (16)$$

which determine the steady-state oscillation amplitude A_o and frequency ω_o. We have assumed that g and b are independent of frequency, which only implies that the analysis is restricted to situations in which $Y(\omega)$ varies much more rapidly than y.

Let us now presume that the steady-state amplitude and bias current are perturbed by a small amount, that is

$$I_d = I_o + \Delta I$$

and

$$V_a = A_o + A_o\delta$$

where $\Delta I \ll I_o$ and $\delta \ll 1$. Then, linearizing the variation of g and b with both V_a and I_d gives

$$g = G - sG\delta + k_g\Delta I$$

and

$$b = -B + rG\delta - k_b\Delta I$$

where

$$s = -\frac{A_o}{G}\frac{\partial g}{\partial V_a}, \qquad r = \frac{A_o}{G}\frac{\partial b}{\partial V_a}$$
$$k_g = \frac{\partial g}{\partial I}, \qquad k_b = -\frac{\partial b}{\partial I} \tag{17}$$

are the RF voltage and dc current linearization parameters of the diode's admittance, evaluated at the steady-state large-signal oscillation point. We may then write

$$[sG\delta - k_g\Delta I]B' - [rG\delta - k_b\Delta I] + |Y'|^2\frac{d\delta}{dt} = 0.$$

Collecting terms,

$$\frac{d\delta}{dt} + \gamma\delta = \Delta f \tag{18}$$

where

$$\gamma = \frac{sGB' - rGG'}{G'^2 + B'^2} \tag{19}$$

and

$$\Delta f = \frac{B'k_g - G'k_b}{G'^2 + B'^2}\Delta I. \tag{20}$$

Introducing

$$G'/B' = \tan\theta_c,$$
$$r/s = \tan\theta_d,$$

and

$$\theta = \theta_c + \theta_d + \pi/2,$$

we obtain

$$\gamma = \frac{\sqrt{r^2 + s^2}}{\left|\frac{\partial Y}{\partial\omega}\right|} G\sin\theta, \tag{21}$$

which is eq. (7) of the text.

Equation (19) for Δf becomes

$$\Delta f = \frac{\sqrt{k_g^2 + k_b^2}}{\left|\frac{\partial Y}{\partial\omega}\right|}\sin(\chi)\Delta I. \tag{22}$$

The meaning of the angles θ_c, θ_d, and θ can be seen in Fig. 18 where it is seen that θ is the total angle of intersection between the negative of the circuit admittance curve, and the tangent to the large-signal

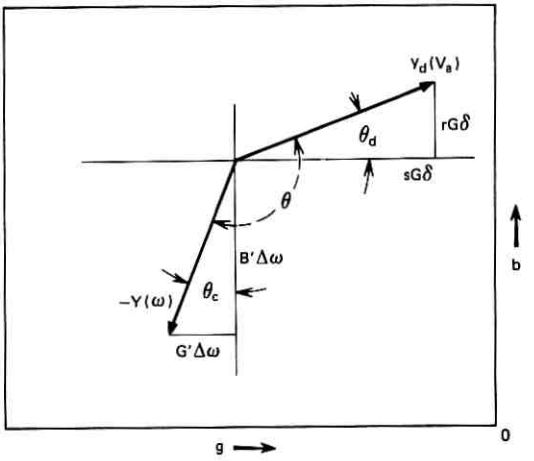


Fig. 18—Defining the angles θ_c, θ_d, and θ in the microwave large-signal admittance plane of the diode.

device variation with voltage amplitude, at the steady-state operating point. The angle χ is defined by

$$\chi = \frac{\pi}{2} - \theta_I - \theta_c$$

with θ_c defined as above and θ_I defined by

$$\tan \theta_I = \frac{k_b}{k_g}.$$

The graphical interpretation of χ is given by Fig. 19, and it is seen to be the angle between the negative of the circuit admittance and the direction of the variation of the large-signal device admittance with dc current, also at the steady-state operating point.

Assuming an $\exp(j\omega t)$ time dependence for δ and ΔI, that is $\delta = \text{Real}(\tilde{\delta} e^{j\omega t})$, and $\Delta I = \text{Real}(\tilde{\imath}_m e^{j\omega t})$, eq. (17) gives

$$\tilde{\delta} = \frac{\widetilde{\Delta f}}{\gamma(1 + j\omega/\gamma)}$$

or

$$\tilde{\delta} = \frac{\sqrt{k_g^2 + k_b^2}}{\sqrt{r^2 + s^2}} \frac{\sin \chi}{\sin \theta} \frac{1}{G} \frac{\tilde{\imath}_m}{(1 + j\omega/\gamma)}.$$

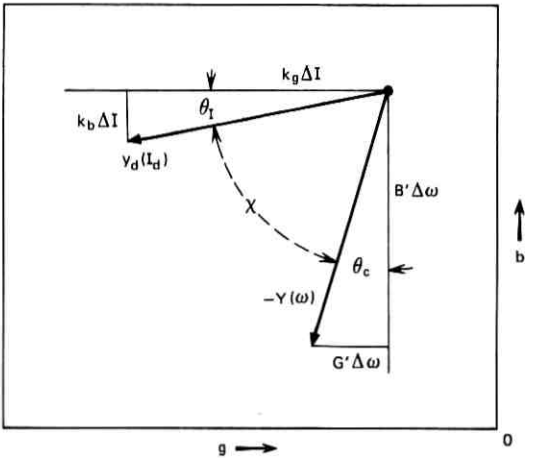


Fig. 19—Defining the angles θ_I, θ_c, and χ in the microwave large-signal admittance plane of the diode.

But

$$\frac{\sqrt{k_g^2 + k_b^2}}{\sqrt{r^2 + s^2}}\left(\frac{A_o}{G}\right) = \frac{\left|\frac{\partial y}{\partial I_d}\right|}{\left|\frac{\partial y}{\partial V_a}\right|} = \left|\frac{dV_a}{dI_d}\right|$$

so

$$\tilde{\delta} = \frac{1}{A_o}\frac{\sin\chi}{\sin\theta}\left|\frac{\partial V_a}{\partial I_d}\right|\frac{\tilde{\imath}_m}{(1 + j\omega/\gamma)}.$$

We may therefore interpret γ as the modulation frequency at which the microwave power in the sideband ($\sim|\delta|^2$) is one-half of its low-frequency response, that is, the common 3-dB modulation half-bandwidth of the oscillator.

The rectification equation, eq. (1) of the text, is now assumed to be true as V_a varies slowly with time. This gives, for the voltage at frequency ω (to first order in $\tilde{\delta}$),

$$\tilde{V}_m = R_{sc}\tilde{\imath}_m - \frac{m}{4W\epsilon_c}2V_a^2\tilde{\delta}$$

or

$$Z_t = \frac{\tilde{V}_m}{\tilde{\imath}_m} = R_{sc} - \frac{m}{4W\epsilon_c}\frac{\sin\chi}{\sin\theta}\left|\frac{dV_a^2}{dI_d}\right|\frac{1}{(1 + j\omega/\gamma)}. \qquad (23)$$

We may then write

$$Z_t = R_{sc} - \frac{R_-}{1 + j\omega/\gamma} \tag{24}$$

where R_- and γ are defined by eqs. (6) and (7) of the text. Z_t is the terminal impedance of the diode at baseband frequencies, and if $R_- > R_{sc}$, then a net negative resistance exists at low frequencies. The equivalent circuit for Z_t is shown in Fig. 7 and its properties are discussed in Section V of the text.

REFERENCES

1. Clorfeine, A. S., and Hughes, R. D., "Induced dc Negative Resistance in Avalanche Diodes," Proc. IEEE (Letters), *57*, No. 5 (May 1969), pp. 841–842.
2. Olson, H. M., Jr., "Negative Resistance Diode Coaxial Oscillator with Resistive Spurious Frequency Suppressor," U. S. Patent No. 3,621,463, November 16, 1971.
3. "Microwave Power Generation and Amplification using IMPATT Diodes," Hewlett-Packard Application Note 935, Hewlett-Packard Corp., Palo Alto, California.
4. Read, W. T., Jr., "A Proposed High-Frequency, Negative Resistance Diode," B.S.T.J., *37*, No. 3 (March 1958), pp. 401–446.
5. Evans, W. J., Scharfetter, D. L., Johnston, R. L., and Key, P. L., "Tuning Initiated Failure in Avalanche Diodes," J. Appl. Phys., *42*, No. 2 (February 1971), pp. 799–803.
6. Blue, J. L., "Approximate Large-Signal Analysis of IMPATT Oscillators," B.S.T.J., *48*, No. 2 (February 1969), pp. 383–396.
7. Sze, S. M., *Physics of Semiconductor Devices*, New York: John Wiley and Sons, 1969.
8. Gewartowski, J. W., and Morris, J. E., "Active IMPATT Diode Parameters Obtained by Computer Reduction of Experimental Data," IEEE Trans. Microwave Theory and Techniques, *MTT-18*, No. 3 (March 1970), pp. 157–161.
9. Magalhaes, F. M., and Kurokawa, K., "A Single-Tuned Oscillator for IMPATT Characterizations," Proc. IEEE (Letters), *58*, No. 5 (May 1970), pp. 831–832.
10. Gupta, M. S., "Noise in Avalanche Transit-Time Devices," Proc. IEEE, *59*, No. 12 (December 1971), pp. 1674–1687.
11. Kurokawa, K., "Some Basic Characteristics of Broadband Negative Resistance Oscillator Circuits," B.S.T.J., *48*, No. 6 (July–August 1969), pp. 1937–1955.
12. Scharfetter, D. L., and Gummel, H. K., "Large-Signal Analysis of a Silicon Read Diode Oscillator," IEEE Trans. Electron Devices, *ED-16*, No. 1 (January 1969), pp. 64–77.
13. Kurokawa, K., "Noise in Synchronized Oscillators," IEEE Trans. Microwave Theory and Techniques, *MTT-16*, No. 4 (April 1968), pp. 234–240.
14. Scharfetter, D. L., "Power-Impedance-Frequency Limitations of IMPATT Oscillators Calculated from a Scaling Approximation," IEEE Trans. Electron Devices, *ED-18*, No. 8 (August 1971), pp. 536–543.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 3, March, 1973
Printed in U.S.A.

Dynamic Data Reallocation in Bubble Memories

By P. I. BONYHARD and T. J. NELSON

(Manuscript received October 10, 1972)

Bubble technology offers several operations that have no equivalents in technologies based on magnetic recording. Examples of such operations are: transfer, reversal of the direction of propagation, and opening and closing of gaps in the data stream. This paper shows how such operations can be used to dynamically reallocate data in the bubble memory, causing it to become an integrated memory hierarchy. A considerable improvement in performance results. A model is presented which relates the bubble memory with dynamic reallocation to stack processing, a technique used in the evaluation of memory hierarchies. With the aid of this model it becomes possible to calculate the performance of the bubble memory using published data derived from the traces of selected typical programs. Memory design is optimized for the execution of such programs. Design parameters are proposed for a 2-Mb bubble memory with 128 detectors which, in the execution of the type of program for which data were available, requires an average of only 8.8 shifts for access and an average of 12.1 shifts per memory cycle. If bubbles are propagated at a rate of 1 MHz, the average access and cycle times for this memory become 8.8 μs and 12.1 μs, respectively. Such performance, in conjunction with the low cost per bit offered by bubble technology, is expected to have a major impact. The performance of this memory, when operated in conjunction with a faster buffer, is also calculated. The use of a 64-kb buffer is shown to reduce the average number of shifts for access to 1.05, and the average number of shifts per cycle to 1.9.*

I. INTRODUCTION

One major area of application of magnetic bubble technology is mass memory.[1] The potential advantages of bubbles over disk files and drums are: shorter access time, lower cost and power dissipation,

* The contents of this paper were presented at the 1972 Intermag Conference by P. I. Bonyhard.

reduced volume, and a higher degree of modularity.[2] Moreover, there are several operations which can be performed on the information in a bubble memory which cannot be performed by conventional magnetic recording. The purpose of this paper is to show that a magnetic bubble memory can function as a hierarchy, given certain operations which are not difficult to realize.

The operations relevant to this paper are: transfer, instantaneous reversal of the direction of propagation, and removal of bubbles from a propagate channel while closing the gap left by them or creation of a gap while inserting new bubble information into the channel. Transfer of bubbles from one propagate channel to another already plays an important role in the design of bubble mass memories.[1,2] By a process of (*i*) propagating forward n cycles, (*ii*) removing bits from the channel and closing the gap, (*iii*) propagating backward n cycles, and (*iv*) opening a gap and reinserting the bubbles, a permutation of the information in the channel is effected. It will be shown how a simple algorithm using one permutation element per propagate channel causes the memory to function as an integrated hierarchy. We also show how this dynamic reallocation of data can be combined with the major-minor loop organization[1] so as to optimize memory cost performance. It is concluded that the improvement is so substantial as to have a major impact in the computer industry.

II. DYNAMIC DATA REALLOCATION IN CLOSED LOOP SHIFT REGISTERS

Consider an assembly of simple, closed loop shift registers as shown in Fig. 1a. Information propagates in all registers synchronously under the influence of a common rotating drive field. One page of information is stored along a horizontal line, a particular page being formed by the black dots in Fig. 1a. It is inherent in the dynamic scheme that each page must carry its own reference address and some of the bits of the page are devoted to this purpose. Thus, extra loops, totaling $\log_2$ (number of pages), have to be provided.

Upon request for a specified page, information is shifted clockwise until the appropriate group of detectors, the position of which is marked D in Fig. 1b, detects the right address. This takes, say, x cycles of propagation. Now the page can be read or rewritten. The memory next is reset by shifting x cycles counterclockwise. Using the circuit arrangement of Fig. 1b, the addressed page will remain arrested at the detectors while all other pages move back. A T-bar realization of Fig. 1b is shown in Fig. 2. This design has been operated quasi-statically to demonstrate feasibility.

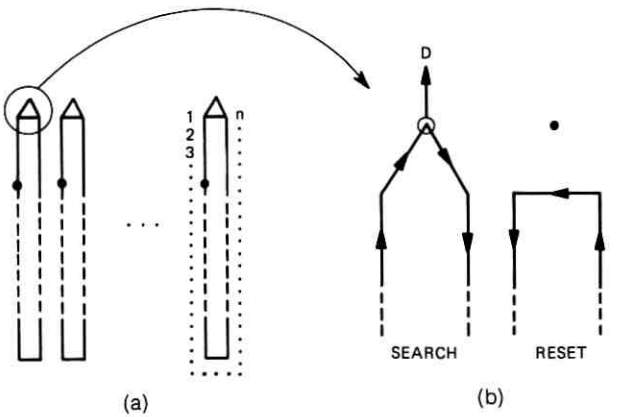


Fig. 1—Assembly of shift register loops for dynamic address reallocation.

Now let the physical page locations be numbered 1, 2, 3, $\cdots$, n, according to their distance from the detectors. A given page will reside in a location, the number of which is equal to the number of requests

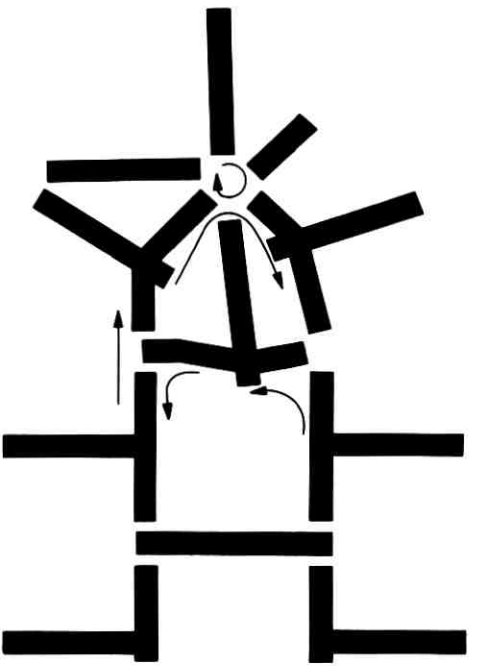


Fig. 2—Magnetic circuit used to realize the function of Fig. 1b.

that have been made, since the page was last requested, for pages originally in locations with higher numbers. Pages never requested will lie in the locations with the highest numbers. Thus recently used pages are near the "top," whereas less recently used pages are farther down. This replacement algorithm has been discussed before in the literature and has been named "the stack."[3]

It will now be shown how the average number of shifts necessary to reach an addressed page in the bubble stack can be calculated. Consider the stack to be arbitrarily divided into two parts, one part consisting of pages in locations 1, 2, $\cdots$, k_i, and the other part of pages in locations $k_i + 1$, $k_i + 2$, $\cdots$, n. It should be recognized that the first k_i page locations can be thought of as forming a "buffer" and

TABLE I—HIT RATIO DATA (REPRODUCED FROM REF. 4)

Buffer Size (Bits)	Classes	Page Size (Bits) 128	256	512	1k	2k	4k	8k	16k*	32k*
8k	1	0.894	0.915	0.928	0.924	0.884	0.792	0.495	—	—
	4	0.895	0.916	0.921	0.904	0.791	—	—	—	—
	16	0.891	0.903	0.860	—	—	—	—	—	—
	64	0.857	—	—	—	—	—	—	—	—
16k	1	0.931	0.949	0.957	0.958	0.950	0.912	0.824	0.56	—
	4	0.931	0.948	0.954	0.955	0.933	0.808	—	—	—
	16	0.930	0.943	0.943	0.909	—	—	—	—	—
	64	0.921	0.913	—	—	—	—	—	—	—
32k	1	0.951	0.969	0.973	0.978	0.977	0.966	0.939	0.86	0.64
	4	0.955	0.969	0.973	0.977	0.974	0.951	0.834	—	—
	16	0.955	0.968	0.972	0.970	0.933	—	—	—	—
	64	0.955	0.963	0.948	—	—	—	—	—	—
	256	0.934	—	—	—	—	—	—	—	—
64k	1	0.977	0.986	0.988	0.985	0.987	0.987	0.984	—	—
	4	0.981	0.986	0.988	0.986	0.987	0.985	0.965	—	—
	16	0.981	0.985	0.988	0.987	0.983	0.954	—	—	—
	64	0.979	0.984	0.985	0.974	—	—	—	—	—
	256	0.974	0.971	—	—	—	—	—	—	—
128k	1	0.985	0.993	0.994	0.996	0.993	0.992	0.994	—	—
	4	0.990	0.993	0.994	0.996	0.994	0.992	0.993	—	—
	16	0.990	0.994	0.995	0.997	0.995	0.991	0.957	—	—
	64	0.990	0.994	0.995	0.995	0.985	—	—	—	—
	256	0.989	0.992	0.986	—	—	—	—	—	—
256k	1	0.989	0.996	0.997	0.997	0.999	0.994	0.997	—	—
	16	0.994	0.996	0.997	0.998	0.998	0.996	0.997	—	—
	32	0.994	0.996	0.998	0.998	0.998	0.997	0.997	—	—
	64	0.994	0.996	0.998	0.998	0.998	0.988	—	—	—
	256	0.994	0.996	0.997	0.992	—	—	—	—	—

the last $n - k_i$ pages a "memory" in the terminology conventionally used for two-level memory hierarchies.[3,4] Whenever a page is brought from the memory to the buffer, the page currently in the k_ith position moves from the buffer to the memory. Clearly, this page is the least recently used page in the buffer, so that the replacement algorithm in this two-level hierarchy is "least recently used" (LRU). Hit ratios, that is, fractions of all requests that can be satisfied from the buffer without reference to the memory, can be found in literature. A particularly useful set of hit ratio data is given in Table II of Ref. 4 and is reproduced as Table I of this paper. The columns marked with asterisks contain entries obtained by graphical extrapolation from the original data.

In a two-level hierarchy, derived by cutting the bubble stack at the k_ith location, the number of bits per page is equal to the number of bubble loops, say, ℓ. The number of bits in the buffer is then ℓk_i. If the corresponding hit ratio is h_i, then the average number of shifts necessary per request might be estimated as

$$\bar{S} = \frac{k_i - 1}{2} h_i + \frac{n + k_i - 1}{2} (1 - h_i).$$

This is the average number of shifts in the buffer times the probability of a hit, plus the average number of shifts to the memory times the probability of a miss.

The bubble stack, however, can be cut at any location. A better estimate is obtained by dividing the stack into $m > 2$ levels

$$\bar{S} = \sum_{i=1}^{m} \frac{k_i + k_{i-1} - 1}{2} (h_i - h_{i-1}) \tag{1}$$

where $k_o = 0$, $k_m = n$, and $h_m = 1$. The best approximation to $\bar{S}$ is achieved when every value of k between 1 and n is taken into account:

$$\bar{S} \to \sum_{k=1}^{n} (k - 1)(h_k - h_{k-1}).$$

This result is, of course, self-evident, as $k - 1$ is the distance of the kth location and $h_k - h_{k-1}$ is the probability of hitting the kth location.

$\bar{S}$ has been calculated on the basis of eq. (1) using the entries that can be found in Table I. It is an overestimate because it divides the probability equally among the levels between entries. Actually, the

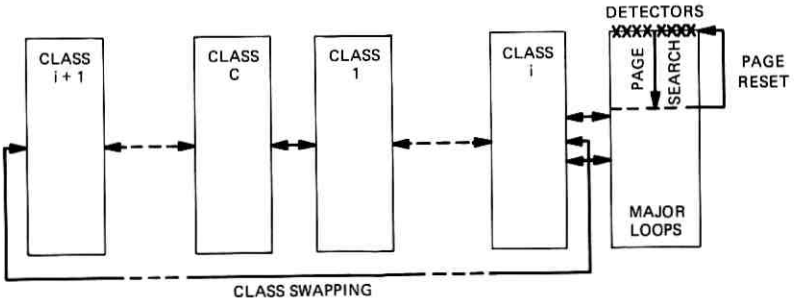


Fig. 3—Schematic representation of the operation of a major-minor loop organized memory with dynamic data reallocation in the major loops.

probability must be monotonically decreasing. The results are tabulated below:

Number of loops	128	256	512	1K	2K	4K	8K
Minimum number of bits per loop	2K	1K	512	256	128	64	32
Average number of shifts	57.8	25.9	12.13	5.67	3.08	1.44	1.05

The minimum number of bits per loop listed in the above tabulation is based on the information kindly provided by the author of Ref. 4 to the effect that the average program size used in deriving the hit ratio data was 256 kb. Thus closed loop shift registers can access data at least ten times faster with dynamic reallocation. A magnetic bubble memory with this feature is, in fact, an integrated hierarchy.

III. DYNAMIC DATA REALLOCATION IN MAJOR-MINOR LOOP ORGANIZED MEMORY PLANES

Dynamic data reallocation in closed loop shift registers does not appear to be a very attractive bubble mass memory design. Each loop must have its own detector, so that either the number of detectors or the average number of shifts is high. The average number of shifts for a given number of bits per detector does not compare too favorably with the major-minor loop organization.[1] Dynamic reallocation may, however, be combined with major-minor loop organization to form an extremely attractive design. Dynamic reallocation can be performed in the major loops only, in the minor loops only, or in both. The last of these three options is probably too complex to be of practical use. A design based on the first option is presented below. The second option

may be a valid alternative but, at least in the authors' opinion, it is less attractive.

Figure 3 illustrates schematically how the system operates. The information that may occupy the assembly of all the major loops concurrently forms a class. One of the classes is selected by shifting information in the minor loops until the required class is aligned for transfer, and then transferring it into the major loops. Two bits from each minor loop are transferred and the major loop is filled completely. Consequently, two bits of each class are stored in each minor loop in each memory plane. A plane consists of one major loop and all the minor loops associated with it, as well as one detector and one controlled bubble generator operating on the major loop. Once the right class resides in the major loops, the major loops operate exactly as the assembly of loops described in the previous section. The net result is that page locations are dynamically reallocated within each class, but pages are never permitted to cross class boundaries. Hit ratios that correspond to such a multiclass system are given in Ref. 4, and also appear in Table I.

It should be recognized that a page in this system consists of all those bits that may concurrently be detected. Thus there is one bit of each page in each plane. Also, the transfer mechanism considered here is of the "conductor" variety,[1] so that reversing the sense of propagation may be freely used to accomplish dynamic address reallocation in the major loops. As this consists of an equal number of forward and reverse shifts, the gaps left in the minor loops are correctly aligned at the time the class is to be transferred back.

Reaching a page in the memory is accomplished in two steps. First the right class is positioned and transferred into the major loops, and then the page is brought to the read/write port. The latter step has already been discussed in the previous section, but the former one, class swapping, needs some elaboration. Following the approach taken in Ref. 4, it is assumed that several programs are concurrently resident in the memory. More specifically, the memory size is chosen to be 2 Mb, whereas the average program size is 256 kb, so that an average of 8 programs may share the memory. It is assumed that each program resides in some number of contiguous classes which will be called active classes for the program currently being executed. The class occupying the major loops at any given time can be looked upon as a single-page buffer of a single-class two-level hierarchy, with all the other classes forming the rest of the pages in the memory. Of course,

for a single-page buffer the replacement algorithm is trivial. However, as can be seen in Table I, the hit ratios associated with such single-page buffers are still fairly high. Consequently, there is a good chance that the next request is addressed to the class already in the major loops, provided that the addressed class is not transferred back to the minor loops until it is known which class is addressed next. Therefore, it makes good sense to operate in this manner.

The average number of shifts necessary to bring out the new class, provided that a "class-page" failure has occurred, will now be calculated. It is assumed that the particular program executed at the time resides in m contiguous classes and the ith class was addressed last. In the absence of any further information, it appears to be reasonable to assume that each other class is equally likely to be addressed next. Consequently, the average number of shifts is

$$\bar{S}_m = \frac{2}{m}\left(\sum_{j=1}^{i-1} j + \sum_{j=1}^{m-i} j\right) = \frac{(i-1)i}{m} + \frac{(m-i+1)(m-i)}{m}.$$

It has to be remembered that there are two bits per class in each minor loop. The average $\bar{S}_m$ must be further averaged as i assumes all values between 1 and m, so that

$$\bar{\bar{S}}_m = \frac{1}{m}\sum_{i=1}^{m} \bar{S}_m = \frac{2}{m^2}\sum_{i=1}^{m} (i-1)i = \frac{2}{3}\frac{m^2-1}{m}.$$

To this two further shifts must be added to accomplish in and out transfer of two bits per minor loop. The total average number of class swapping shifts per page request is

$$S_{cs} = (1 - h_{cs})\left(\frac{2}{3}\frac{m^2-1}{m} + 2\right) \tag{2}$$

where h_{cs} is the class hit ratio discussed earlier in this section, and is the number of active classes.

To the class swapping shifts, given by eq. (2), one must add the average number of shifts necessary to reach a page within the selected class to get the total average number of access shifts per request to the memory. The average number of shifts within the selected class is given by eq. (1) with the hit ratios taken for the appropriate number of classes. The latter quantity must be doubled when calculating the

average number of shifts per memory cycle. The results are given below for three alternative designs labeled A, B, and C.

	A	B	C
Bits per memory	2M	2M	2M
Planes per memory	128	128	128
Bits per plane	16k	16k	16k
Classes	64	128	256
Bits per minor loop	128	256	512
Minor loops per plane	128	64	32
Bits per major loop	256	128	64
Active classes	8	16	32
Active bits per minor loop	16	32	64
Class hit ratio	≈0.64	≈0.56	0.495
Class swapping shifts	7.25	12.63	23.3
Class swapping shifts per request	2.61	5.56	11.8
Page search shifts	7.00	3.26	1.40
Total shifts per access	9.61	8.82	13.2
Total shifts per cycle	16.61	12.08	14.6

Details of the page search shift calculations are given in Table II. The memory size was chosen to be consistent with Ref. 4, where the

TABLE II—CALCULATION OF THE AVERAGE NUMBER OF PAGE SEARCH SHIFTS FOR THREE DESIGNS

Design	k_i	$\frac{k_i + k_{i-1} - 1}{2}$	h_i	$h_i - h_{i-1}$	$\frac{1}{2}(h_i - h_{i-1}) \times (k_i + k_{i-1} - 1)$
A	8	3.5	0.894	0.894	3.13
	16	11.5	0.931	0.037	0.43
	32	23.5	0.955	0.024	0.56
	64	47.5	0.981	0.026	1.24
	128	95.5	0.990	0.009	0.86
	256	191.5	0.994	0.004	0.77
					6.99
B	4	1.5	0.891	0.891	1.34
	8	5.5	0.930	0.039	0.21
	16	11.5	0.955	0.025	0.29
	32	23.5	0.981	0.026	0.61
	64	47.5	0.990	0.009	0.43
	128	95.5	0.994	0.004	0.38
					3.26
C	2	0.5	0.880	0.880	0.440
	4	2.5	0.934	0.054	0.135
	8	5.5	0.955	0.021	0.115
	16	11.5	0.980	0.025	0.287
	32	23.5	0.990	0.010	0.235
	64	47.5	0.994	0.004	0.190
					1.402

data came from, and the plane size was chosen as the one that would provide an economical design on the basis of current cost estimates.

Each of the three designs has an overhead associated with it, because of the extra planes needed to store the address tags of the reallocated pages. This overhead for designs A, B, and C is 6.3 percent, 5.5 percent, and 4.7 percent, respectively.

IV. THE USE OF A BUFFER TO IMPROVE PERFORMANCE

The best design outlined in the preceding section gave an average cycle time of about 12.1 shifts or, assuming 1-MHz operation, 12.1 μs. The question arises of how this figure may be further improved by the use of a smaller submicrosecond cycle time core or integrated circuit buffer. The answer can be readily calculated, provided that the buffer is divided into as many classes as there are active classes in the memory, and that the page size in the buffer is the same as in the memory. Now, if there are k_B page frames per class in the buffer, then the pages currently residing in location 1, 2, $\cdots$, k_B of all active classes will also appear in the buffer. This statement is not quite correct if the buffer stores only read (fetch) data, whereas write (store) data are sent from the central processor directly to the memory, but the difference is probably negligible.

It is, of course, also necessary to uniquely assign all nonactive memory classes to buffer classes, so that program swapping can take place. This can be done as follows. In terms of Design B, let the 7 most significant bits of the 14-bit page address be the class address. As each program occupies contiguous classes, the 4 least significant bits of the class address refer to classes in the same program for a typical program size. Thus the 1st, 2nd, 3rd, and 4th least-significant bits of the memory page address should be considered to be the buffer class address.

The average number of shifts necessary in the memory per request can now be calculated as follows. If the buffer hit ratio is h_B, then h_{cs} in eq. (2) should be replaced by h_B, whereas all contributions to the page search shifts, calculated in Table II, by values of $k_i \leqq k_B$ should be neglected. The results are tabulated below, still for Design B.

Buffer size (bits)	8k	16k	32k	64k
Buffer bit ratio	0.891	0.930	0.955	0.981
Pages per class in buffer	4	8	16	32
Class swapping shifts	1.37	0.88	0.56	0.24
Page search shifts	1.92	1.71	1.42	0.81
Total access shifts	3.29	2.59	1.98	1.05
Total shifts per cycle	5.21	4.30	3.40	1.86

V. DISCUSSION AND CONCLUSION

The availability of memories costing approximately the same or less per bit as a disk file or drum with cycle times of about 10 μs is likely to have a major impact on the computer industry. This appears to be the answer to the system designer's old dream of a memory at core speeds and disk costs. It looks like we can fulfill this dream with bubble memories, provided that we can shift bubbles at 1-MHz rates, which appears to be a reasonable objective. It may be added that the organizations discussed here should be also applicable to other forms of serial memory technologies, MOS registers, charge coupled registers, etc.

The major applications for bubble memories with dynamic address reallocation are likely to be in mini- and midi-computers and in other systems with relatively less powerful central processors. The larger systems and those with faster processors will probably find the buffered configuration more attractive. For instance, one may use a 64-kb, 0.5-μs access time, 1-μs cycle time buffer to give an average access time of about 1.5 μs, and an average cycle time of about 3.0 μs for the 2-Mb bubble memory. In many systems this would be vastly preferable to an all-integrated-circuit memory costing much more for the same capacity, even if the latter memory has a cycle time of 0.1 μs.

Further work should be directed towards the assessment of the applicability of these techniques to electronic telephone switching systems.

VI. ACKNOWLEDGMENTS

The authors have benefited from helpful discussions with many individuals: A. D. Friedman, J. E. Geusic, B. W. Kernighan, M. D. McIlroy, P. R. Menon, P. C. Michaelis, and H. E. D. Scovil in particular. Also, we thank R. L. Mattson who provided a preprint of the work which inspired this paper.

REFERENCES

1. Bonyhard, P. I., et al., "Applications of Bubble Devices," IEEE Trans. Magnetics, *MAG-6*, September 1970, pp. 447–451.
2. Bobeck, A. H., "Magnetic Domain Devices–A Tutorial," presented at the 1971 Intermag Conf. (also to be published in Scientific American).
3. Mattson, R. L., et al., "Evaluation Techniques for Storage Hierarchies," IBM Syst. J., *9*, No. 2, 1970, pp. 78–117.
4. Mattson, R. L., "Evaluation of Multilevel Memories," IEEE Trans. Magnetics, *MAG-7*, December 1971, pp. 814–819.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 3, March, 1973
Printed in U.S.A.

Restoring the Orthogonality of Two Polarizations in Radio Communication Systems, II

By T. S. CHU

(Manuscript received October 3, 1972)

The Poincaré sphere has been applied to the analysis of orthogonalizing two polarization ellipses by a differential phase shifter and a differential attenuator. The condition of minimum differential attenuation for removing a given amount of nonorthogonality is determined. A previously reported transformation via two nonorthogonal linear polarizations should be used for two slender ellipses.[1] *Another transformation via two oppositely rotating ellipses having parallel axes and equal axial ratios should be used for two fat ellipses. System applications of the transformations are discussed.*

I. INTRODUCTION

A method of recovering the orthogonality of two polarizations in a radio communication system was presented recently.[1] Two arbitrary polarization ellipses are first transformed simultaneously into two nonorthogonal linear polarizations by a differential phase shifter, and then the nonorthogonality is removed by a differential attenuator. It is of interest to ask whether that transformation is optimum for system applications.

Before proceeding further, we will define the optimum transformation. Clearly, it is desirable to minimize the differential attenuation that we must introduce to correct for nonorthogonality. However, in order to achieve maximum bandwidth, we also should minimize the differential phase shift even if we assume, as we do here, that the polarization characteristics of components in the system are not very frequency sensitive over the operating bandwidth. This assumption is, for example, expected to be valid for any depolarization in the main

beam of reflector-type antennas provided there is no polarization distortion in the feed radiation. If necessary, we can apply polarization correction separately to each of the subbands within a wide band.

In a practical dual-polarization radio system, the two orthogonal polarizations feeding the transmitting antenna are either linear or circular. Signal generators and detectors are always linearly polarized; however, conversion to circular polarization for radio transmission is sometimes made to avoid effects such as Faraday rotation. If the transmission medium and the radiating systems introduce only moderate polarization distortion, the dual polarization signals appear as two slender or two fat polarization ellipses at the receiving terminal, depending on whether linear or circular polarizations are being used. This classification into two types of elliptical shapes suggests an optimum transformation for each type. However, the validity of the transformations presented in the next section is independent of the shapes of the ellipses. A practical design of adjustable differential phase-shifters and attenuators has been suggested by E. A. Ohm.[2]

The following analysis will be presented with the aid of the Poincaré sphere. This geometrical representation of the polarization of a plane electromagnetic wave was introduced to radio engineers by G. A. Deschamps.[3] For convenience of the reader, a summary of the Poincaré spherical representation is given in the Appendix.

II. ANALYSIS

2.1 *Minimum Differential Attenuation*

Let us first find the condition for minimum differential attenuation required for removing a given amount of nonorthogonality between two polarizations. Each polarization is represented by a point on the Poincaré sphere. If the great circle arc connecting two points on the sphere is a semicircle, then the two polarizations are orthogonal to each other. The degree of nonorthogonality between two polarizations can be measured by the deviation of the great circle arc from a semicircle. Let two nonorthogonal elliptically polarized waves be represented by points M_1 and M_2 on Poincaré sphere as shown in Fig. 1a. The great circle arc connecting M_1 and M_2 intersects the equator at C. The longitudes of M_1 and M_2 with respect to C are twice the orientation angles of the two polarization ellipses with respect to the X-axis of a set of X-Y coordinates. This relationship is sketched in Fig. 1b. Since M_1 and M_2 are nonorthogonal to each other, $\overset{\frown}{M_1C} + \overset{\frown}{M_2C} < \pi$. Orthogonalization is accomplished by stretching the great circle arc $\overset{\frown}{M_1M_2}$ to the value of π.

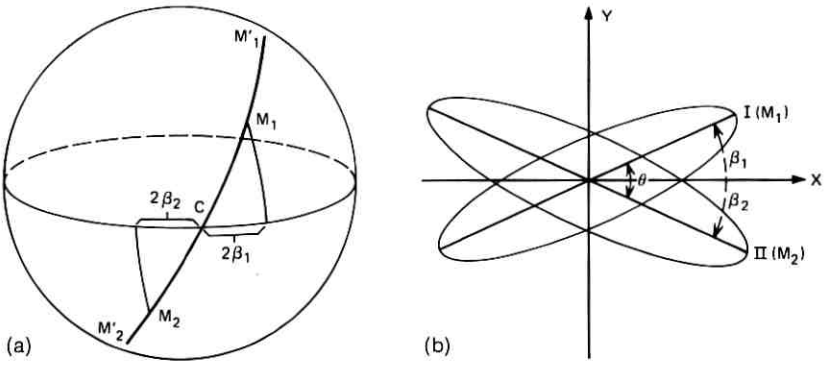


Fig. 1—Orthogonalization by differential attenuation.

If a differential attenuation of $\frac{\tan \frac{1}{2}\overset{\frown}{M_1C}}{\tan \frac{1}{2}\overset{\frown}{M_1'C}} = \frac{\tan \frac{1}{2}\overset{\frown}{M_2C}}{\tan \frac{1}{2}\overset{\frown}{M_2'C}}$ is imposed along the X-axis such that $\overset{\frown}{M_1'C} + \overset{\frown}{M_2'C} = \pi$, then the two polarizations M_1 and M_2 will be orthogonalized. Since $\tan \frac{1}{2}\overset{\frown}{M_1'C} \tan \frac{1}{2}\overset{\frown}{M_2'C} = 1$, the required differential attenuation is minimized by maximizing $\sqrt{\tan \frac{1}{2}\overset{\frown}{M_1C} \tan \frac{1}{2}\overset{\frown}{M_2C}}$ with a proper differential phase shift, which implies the constraint that $\overset{\frown}{M_1C} + \overset{\frown}{M_2C} = \overset{\frown}{M_1M_2}$. In this way, the minimum differential attenuation needed is found to be $\tan \frac{1}{4}\overset{\frown}{M_1M_2}$ when $\overset{\frown}{M_1C} = \overset{\frown}{M_2C}$. For the given degree of nonorthogonality, $\pi - \overset{\frown}{M_1M_2}$, the orthogonalization can be performed by the minimum differential attenuation $\tan \frac{1}{4}\overset{\frown}{M_1M_2}$ only if the two elliptical polarizations have the same axial ratio. Thus the two arbitrary elliptical polarizations must first be transformed by a differential phase shifter to allow a minimum differential attenuation.

Furthermore, one wishes to obtain two orthogonal linear or circular polarizations immediately following orthogonalization. This requirement determines the prerequisite condition that two nonorthogonal elliptically polarized waves should first be transformed into two nonorthogonal linear polarizations or two oppositely rotating elliptical polarizations having parallel major and minor axes and equal axial ratios, both of which are limiting cases of two elliptical polarizations with the same axial ratio.

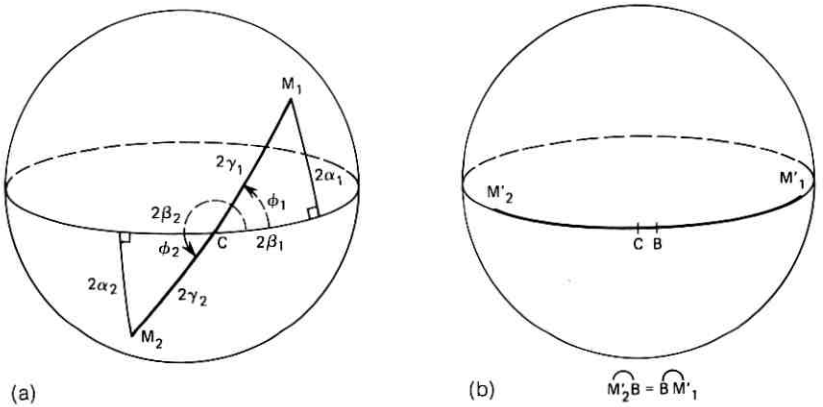


Fig. 2—Simultaneous transformation of two elliptical polarizations into two linear polarizations.

2.2 *Two Slender Ellipses*

If the transmitting antenna is fed by two orthogonal linear polarizations, moderate polarization distortion in a radio communication system will produce two slender polarization ellipses at the receiving terminal. At that point, the orthogonalization should begin by a simultaneous transformation of the two polarization ellipses into two linear polarizations. This transformation has been obtained previously;[1] however, it will be described here in terms of the Poincaré sphere.

Let two nonorthogonal elliptic polarizations be represented by two points, M_1 and M_2 in Fig. 2a. The intersection C of the great circle arc with the equator designates a set of coordinate axes X-Y which is shown in Fig. 1b. The polarization ratios of the two polarization ellipses in terms of these X-Y coordinates will be related by $\tan \phi_1 = \tan \phi_2$. Replacing each side of this equation by eq. (11) yields:

$$\tan 2\alpha_1 \csc 2\beta_1 = \tan 2\alpha_2 \csc 2\beta_2. \tag{1}$$

Substituting $\beta_2 = \beta_1 - \theta$ into eq. (1), one obtains the solution for the orientation of the X-Y axes

$$\beta_1 = \frac{1}{2} \cot^{-1} \left[\frac{\cos 2\theta - \dfrac{\tan 2\alpha_2}{\tan 2\alpha_1}}{\sin 2\theta} \right], \quad 0 < \beta_1 < \frac{\pi}{2}. \tag{2}$$

The arc $\overset{\frown}{M_1M_2}$ can be rotated onto the equator by applying the follow-

ing differential phase delay to the components in the Y direction

$$\phi = \tan^{-1} [\tan 2\alpha_1 \csc 2\beta_1]. \tag{3}$$

Now the angle ψ between the two linear polarizations represented by M_1' and M_2' in Fig. 2b is half of the sum of the arcs $\overset{\frown}{CM_1}$ and $\overset{\frown}{CM_2}$

$$\psi = \gamma_1 + \gamma_2 \tag{4}$$

where $\gamma_i = \tan^{-1} \sqrt{\dfrac{(1 + \tan^2 \alpha_i) - (1 - \tan^2 \alpha_i) \cos 2\beta_i}{(1 + \tan^2 \alpha_i) + (1 - \tan^2 \alpha_i) \cos 2\beta_i}}$ is obtained using eq. (10). This angle ψ may be changed to $\pi/2$ if a differential attenuation of tan $(\psi/2)$ is imposed on the components in the direction B bisecting the two linear polarizations. This direction will be oriented at an angle

$$x = \tfrac{1}{2}(\gamma_1 - \gamma_2) \tag{5}$$

with respect to the X direction of the coordinates in Fig. 1b.

The above equations can be easily identified with those in Ref. 1. One notes that the transformations are valid for the two polarizations located on the same side as well as on opposite sides of the equator of the sphere.

2.3 *Two Fat Ellipses*

If the transmitting antenna is fed by two orthogonal circular polarizations, two fat polarization ellipses will appear at the receiving terminal if the radio communication system is moderately contaminated by polarization distortion. Here the orthogonalization should be started by the simultaneous transformation of two given ellipses into two oppositely rotating ellipses having parallel major and minor axes and equal axial ratios. One looks for an X'-Y' cartesian coordinate system in terms of which the polarization ratios of the two ellipses become complex conjugates of each other after a proper differential phase delay is introduced along the axes. The X'-Y' coordinates correspond to the point C', and the proper differential delay is represented by Δ on the Poincaré sphere as shown in Fig. 3a. Let us write down the expression for the polarization ratio in terms of the X'-Y' coordinates[1]

$$P_i' = \sqrt{\frac{(1 + \tan^2 \alpha_i) - (1 - \tan^2 \alpha_i) \cos 2\beta_i'}{(1 + \tan^2 \alpha_i) + (1 - \tan^2 \alpha_i) \cos 2\beta_i'}}\, e^{j \tan^{-1}\left[\frac{2 \tan \alpha_i}{(1 - \tan^2 \alpha_i) \sin 2\beta_i'}\right]} \tag{6}$$

$$0 < |P_i' < \pi \text{ when } \tan \alpha_i > 0;$$

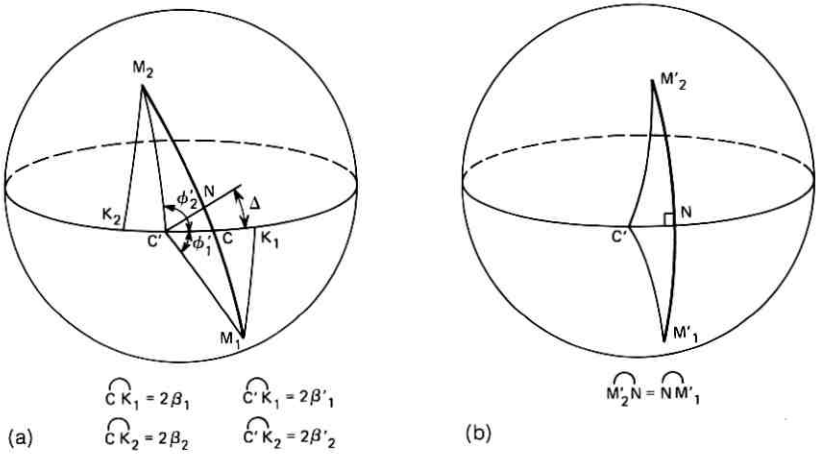


Fig. 3—Simultaneous transformation of two polarization ellipses into two ellipses with parallel axes and equal axial ratios.

where $i = 1, 2$;

$$-\frac{\pi}{2} < \underline{|P_i'|} < \frac{\pi}{2} \text{ when } \sin 2\beta_i' > 0 .$$

Combining the condition $|P_1'| = |P_2'|$ and the relation $\beta_2' = \beta_1' - \theta$, one obtains

$$\beta_1' = \frac{1}{2}\tan^{-1}\left[\frac{\dfrac{(1+\tan^2\alpha_2)(1-\tan^2\alpha_1)}{(1+\tan^2\alpha_1)(1-\tan^2\alpha_2)} - \cos 2\theta}{\sin 2\theta}\right], \quad 0 < \beta_1' < \frac{\pi}{2} \tag{7}$$

which fixes the X′-Y′ axes. The required differential phase delay Δ along the Y′ axis is determined by

$$\underline{|P_1'|} - \Delta = -(\underline{|P_2'|} - \Delta)$$

which gives

$$\Delta = \tfrac{1}{2}(\underline{|P_1'|} + \underline{|P_2'|}). \tag{8}$$

The above differential phase shift corresponds to the rotation of the great circle arc $\overset{\frown}{M_1M_2}$ about the point C′. The transformed ellipses are represented by M_1' and M_2' where $\overset{\frown}{M_1'N} = \overset{\frown}{M_2'N}$ and $\overset{\frown}{M_1'M_2'}$ is perpendicular to the equator as shown in Fig. 3b. The parallel major axes of the transformed ellipses are oriented with respect to the X′ axis at

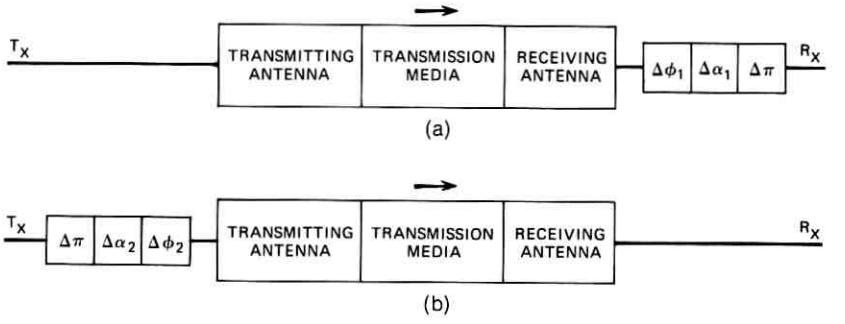


Fig. 4—Locations of the orthogonalization device.

an angle equal to $\frac{1}{2}\,\overset{\frown}{\mathrm{NC}}'$. Using eq. (13), we have

$$\tfrac{1}{2}\,\overset{\frown}{\mathrm{NC}}' = \tfrac{1}{2}\tan^{-1}\left\{\frac{2|P_1'|}{1-|P_1'|^2}\cos\tfrac{1}{2}\left[\underline{|P_1'|} - \underline{|P_2'|}\right]\right\}. \tag{9}$$

Now two oppositely rotating circular polarizations can be obtained by imposing a differential attenuation along the major axis of the transformed ellipses. The value of differential attenuation is also given by $\tan(\psi/2)$ where ψ is determined by eq. (4) of the preceding section.

In addition to the differential phase shift for orthogonalization as discussed above, circular polarization systems require $\Delta(\pi/2)$ phase shift* at both transmitting and receiving ends to convert to the final linearly polarized ports. The same amount of additional differential phase shift overall will also be needed in a system using orthogonal linear polarizations, if a $\Delta\pi$ phase shifter is used for polarization tracking.

III. DISCUSSION

The above analysis assumed that the device for orthogonalization would be located at the receiving terminal as shown in Fig. 4a. Since there always exist two elliptic polarizations which would become orthogonal after going through a linear transmission system with a certain polarization distortion, one can also put the differential ele-

* The notation, $\Delta(\pi/2)$, implies a differential phase shift of $\pi/2$.

ments at the transmitting terminal as shown in Fig. 4b. Another obvious corollary states that the dual-polarization radiation of an antenna always can be orthogonalized in any particular direction. The differential attenuator should be located as illustrated in Figs. 4a and b at the receiver or the transmitter in order to satisfy the condition for minimum differential attenuation. Sometimes it is desirable to use differential elements at both the transmitting and receiving terminals. For example, one may wish to eliminate the polarization distortions of the transmitting and receiving antennas separately.

It is often claimed that the use of circular polarization in a satellite communication system eliminates the need for polarization tracking. In order to realize this advantage, the depolarization of the satellite antenna radiation must be kept small over the entire coverage of ground stations. Furthermore, the matching requirement at each discontinuity of the waveguide feeding network and the radiating system is more stringent for circular polarization, because multiple reflections among the discontinuities often corrupt circular but not linear polarization.

IV. ACKNOWLEDGMENTS

The author wishes to thank A. B. Crawford and E. A. Ohm for helpful discussions.

APPENDIX

Poincaré Spherical Representation

A polarization ellipse shown in Fig. 5a is completely characterized by the axial ratio, A = minor axis/major axis, and the orientation of the ellipse. The sense of rotation can be taken into account by giving $+$ or $-$ sign to the axial ratio. Now we define $\tan^{-1} A$ as the ellipticity angle α with $-45° \leqq \alpha \leqq 45°$, and take the angle between OX and the major axis as the orientation angle β with $-90° < \beta < 90°$. Then a point M on a sphere with longitude 2β and latitude 2α as shown in Fig. 5b completely specifies a state of polarization.

The points on the equator represent linear polarizations. The arc between two points along the equator is twice the angle between two linear polarizations. The points on the upper and lower hemispheres correspond to clockwise and counterclockwise (wave approaching) elliptical polarizations respectively, while the poles designate circular polarizations. For each point which represents an elliptic polarization, the projection K onto the equator defines the orientation of its major

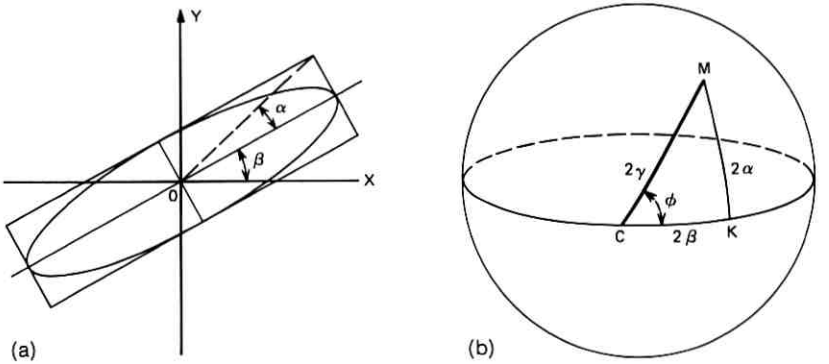


Fig. 5—Poincaré spherical representation of a polarization ellipse.

axis. Any set of X-Y coordinates can be specified by its X direction which in turn corresponds to a point C on the equator. Then the longitude of K can be measured with respect to C.

The polarization ratio P is a complex number defined as the ratio between the Y and X components of the electric vector. Using the expression for the polarization ratio in eq. (6) and the following formulas for the spherical triangle shown in Fig. 5b,

$$\cos 2\gamma = \cos 2\alpha \cos 2\beta \tag{10}$$

$$\tan \phi = \tan 2\alpha \csc 2\beta, \tag{11}$$

one can identify $P = \tan \gamma e^{j\phi}$. The orientation and the ellipticity can be expressed in terms of P as follows:

$$\sin 2\alpha = \sin 2\gamma \sin \phi \tag{12}$$

$$\tan 2\beta = \tan 2\gamma \cos \phi. \tag{13}$$

A differential phase delay of Δ of the Y component with respect to the X component will correspond to a clockwise rotation Δ of the arc $\overset{\frown}{CM}$ around the point "C" on the Poincaré sphere.

REFERENCES

1. Chu, T. S., "Restoring the Orthogonality of Two Polarizations in Radio Communication Systems, I," B.S.T.J., *50*, No. 9 (November 1971), pp. 3063–3069.
2. Ohm, E. A., private communication.
3. Deschamps, G. A., "Geometrical Representation of the Polarization of a Plane Electromagnetic Wave," Proc. IRE, *39*, May 1951, pp. 540–544.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 3, March, 1973
Printed in U.S.A.

Theory of Roundoff Noise in Cascade Realizations of Finite Impulse Response Digital Filters

By D. S. K. CHAN and L. R. RABINER

(Manuscript received September 14, 1972)

This paper presents a theoretical treatment of the roundoff noise problem for the special case of cascade realizations of Finite Impulse Response (FIR) digital filters.† *Explicit relations for evaluating roundoff noise with the usual assumption of uncorrelated samples are presented. Useful scaling methods are stated and classified as to conditions when these methods are optimum. Important differences between use of these scaling procedures for Infinite Impulse Response (IIR) filters and FIR filters are pointed out. Finally, useful properties of linear phase FIR filters are discussed.*

I. INTRODUCTION

In recent years, many techniques have been developed for the design of Finite Impulse Response (FIR) digital filters.[2–8] It is now possible to readily design filters with arbitrary frequency or time response characteristics using the windowing,[2,3] frequency sampling,[4] or optimal design[5–8] methods. While both the windowing and frequency sampling techniques yield suboptimal filters, they are useful because of their simplicity and ease of design. The optimal design technique is of special importance because the filters it generates can be proved to be optimum in a certain sense,[7] and because efficient algorithms exist for its implementation.[7,8]

As a result of these important developments, the FIR type of digital filter is becoming increasingly attractive as an alternative to the IIR (Infinite Impulse Response) type of filter for practical applications. A major advantage of FIR filters over IIR filters is that an FIR filter

† This paper is based on a thesis[1] submitted in partial fulfillment of the requirements for the degrees of Bachelor of Science and Master of Science in the Department of Electrical Engineering at the Massachusetts Institute of Technology in September 1972.

can have an exactly linear phase response while approximating an arbitrary magnitude frequency response. But even without considering this important advantage, current research[9] is revealing that in certain cases FIR filters are competitive with IIR filters in terms of speed and cost. Thus the implementation of FIR filters using finite-precision arithmetic is becoming an important area for research.

Up to the present, little is known as to how different types of FIR filter realizations behave with respect to quantization effects. Since hardware, specifically for the purpose of realizing FIR filters, has already been built for experimentation by various research groups,[10,11] it is important to obtain more knowledge to guide the implementation phase of FIR digital filter design. The purpose of this paper is to present a theoretical treatment of several problems associated with implementations of these filters.

II. PRELIMINARY REMARKS

The effects that quantization has on an IIR filter can be classified into three basic categories:

(*i*) Quantization of the values of samples derived from a continuous input waveform causes inaccuracies in the representation of the waveform (A-D noise).

(*ii*) Finite-precision representation of the infinite-precision filter coefficients alters the frequency response characteristics of the filter (coefficient accuracy problem).

(*iii*) Finite-precision arithmetic causes inaccuracies in the filter output (roundoff noise) which, together with the finite dynamic range of the filter, limit the signal-to-noise ratio attainable. Also, finite-precision arithmetic can lead to limit cycles where the output samples are generally highly correlated.

These same quantization effects also occur in finite wordlength FIR filters with the important exception that limit cycles cannot occur in nonrecursive realizations of FIR filters. In this paper only the third type of quantization effect, viz., roundoff noise, will be discussed. Furthermore, fixed-point arithmetic with rounding will be assumed.

Except for the first category above (A-D noise), all quantization effects depend in degree and character on the type of structure used to implement a filter. There are three well-known structures in which an FIR transfer function can be realized. They are the direct form, the cascade form, and the frequency-sampling structure.[12] Other less well-known structures based on polynomial interpolation formulas

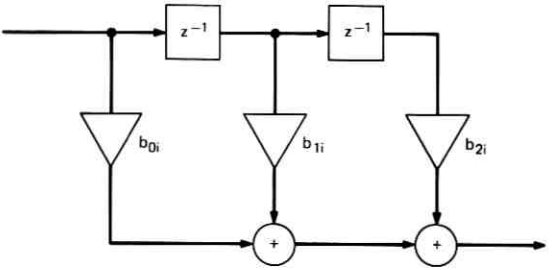


Fig. 1—Cascade-form filter section.

are also possible; these include the Lagrange, Newton, Hermite, and Taylor structures.[13] However, it is as yet unclear under what circumstances, if any, these other structures may be more advantageous than the well-known structures.

Only the cascade structure will be discussed in this paper. A second-order filter section, as shown in Fig. 1, will be used as the basic building block for the cascade structure. Although several minor variations to this configuration for the filter sections are possible,[1] the results presented here are sufficiently general so that they can be readily applied to other configurations as well.

Aside from section configuration, the prime issues that must be confronted in the realization of filters in cascade form are scaling and section ordering. Proper scaling must be performed on a filter in cascade form in order that full use of the dynamic range of each section can be made while avoiding error-producing overflows. By proper scaling, the signal-to-noise ratio of a filter can be maximized for a given quantization step size and section ordering.

Proper ordering must also be determined for a filter in cascade form if the filter is to be useful at all, since the noise output of a cascade filter can depend dramatically on the way it is ordered. For example, Schüssler[13] showed a 32nd-order FIR filter which, ordered two different ways, yields output noise variances that differ by a ratio on the order of 10^8. The problem of section ordering for cascade FIR filters has been investigated in depth,[1,14] and the results show that for higher-order filters, the variation of output noise variance across all orderings is much greater than 10^8.

Jackson[15] has formulated the roundoff noise problem for a general digital filter and has proposed an approach to the scaling of filters to satisfy dynamic range constraints. Most of his results can be specialized to the case of FIR filters by assuming a constant polynomial in the

denominator of the transfer function. However, the different perspective obtained by studying the FIR case separately affords many additional insights.

In this paper, formulas for the evaluation of roundoff noise variances in the FIR cascade structure are presented. Also, specific scaling methods for FIR filters are defined, two of which can be proved to be optimum for two classes of input signals. Finally, certain properties of the linear phase cascade structure useful in the study of section ordering are stated and can be proved. However, for reasons of space, no proofs are included in this paper. They can be found in Ref. 1.

III. DEFINITIONS

The general transfer function for an N-point FIR filter can be written in the form

$$H(z) = \sum_{k=0}^{N-1} h(k)z^{-k} \tag{1}$$

where the real-valued sequence $\{h(k), k = 0, \cdots, N-1\}$ is the impulse response of the filter. Alternatively, $H(z)$ can be expressed in the factored form

$$H(z) = \prod_{i=1}^{N_s} (b_{0i} + b_{1i}z^{-1} + b_{2i}z^{-2}) \tag{2}$$

where b_{ji}, $j = 0, 1, 2$, $i = 1, \cdots, N_s$ are real numbers and N_s, the number of factors, is defined as

$$N_s = \begin{cases} \dfrac{N-1}{2} & N \text{ odd} \\ \dfrac{N}{2} & N \text{ even} \end{cases}$$

and $b_{2N_s} = 0$ if N is even.

A linear phase filter is defined to be a filter, the transfer function $H(z)$ of which is expressible in the form

$$H(z)\,|_{z=e^{j\omega}} = H(e^{j\omega}) = \pm\,|H(e^{j\omega})|\,e^{-j\alpha\omega} \tag{3}$$

where α is a real positive constant with the physical significance of delay in number of samples. The factor $\pm$ is necessary since $H(e^{j\omega})$ actually is of the form

$$H(e^{j\omega}) = H^*(e^{j\omega})e^{-j\alpha\omega}$$

where $H^*(e^{j\omega})$ is a real function taking on both positive and negative values. It is useful to define a mirror-image polynomial (MIP) of degree N to be a polynomial of the form $\sum_{k=0}^{N} a_k z^k$, the coefficients of which satisfy the relation

$$a_k = a_{N-k} \qquad 0 \leqq k \leqq N.$$

Necessary and sufficient conditions on $H(z)$ such that a filter with transfer function $H(z)$ has an exactly linear phase response can then be stated as follows:

Theorem 1: $H(z)$ can be expressed in the form (3) *if and only if one of the following equivalent conditions hold:*

(*i*) $h(k) = h(N-1-k)$, $0 \leqq k \leqq N-1$.

(*ii*) *If z_i is a zero of $H(z)$, then z_i^{-1} is also a zero of $H(z)$. Also, if $z_i = +1$ is a zero of $H(z)$, then it occurs in even multiplicity.*

(*iii*) *Suppose z_i is a zero of the ith factor in* (2). *Let $S = \{i: z_i \text{ is real}\}$ and $Q = \{i: i \notin S\}$. Then $f(z) = \prod_{i \in S} (b_{0i} + b_{1i}z^{-1} + b_{2i}z^{-2})$ is a mirror-image polynomial in z^{-1}, and for all $i \in Q$, either $b_{0i} = b_{2i}$ or there exists $j \neq i$, $j \in Q$, such that*

$$\frac{b_{0i}}{b_{2j}} = \frac{b_{1i}}{b_{1j}} = \frac{b_{2i}}{b_{0j}}. \tag{4}$$

Furthermore, the following is a sufficient condition for $H(z)$ to be expressible in the form (3):

(*iv*) *In* (2), *for $1 \leqq i \leqq N_s$, either $b_{2i} = 0$ and $b_{0i} = b_{1i}$, or $b_{0i} = b_{2i}$, or there exists $j \neq i$, $1 \leqq j \leqq N_s$, such that*

$$\frac{b_{0i}}{b_{2j}} = \frac{b_{1i}}{b_{1j}} = \frac{b_{2i}}{b_{0j}}.$$

In all cases the value of α is $\alpha = (N-1)/2$.

It should be pointed out that a section with $b_{0i} = b_{2i}$ is necessarily one which synthesizes either two complex conjugate zeros on the unit circle, or two reciprocal zeros on the real axis, or two identical zeros at $+1$ or -1. Furthermore, two sections which satisfy (4) are precisely those sections which synthesize reciprocal zeros (i.e., if z_i is a zero of one section, then z_i^{-1} is a zero of the other section). Thus, taking (2) as the basis for the FIR cascade form, condition (*iv*) of Theorem 1 provides a way to assign zeros to individual sections of the cascade structure so that linear phase is guaranteed independent

of scaling or ordering. Hence, in this paper the following convention of zeros assignment for linear phase filters will be adopted: complex zeros are grouped by conjugate pairs, real zeros that are reciprocals of each other are paired together, while doubled or higher multiplicity zeros are grouped by pairs of the same kind. In this way the only zero that can occur by itself in a section is $z = -1$ (since by Theorem 1, $z = +1$ is not allowed as a zero of odd multiplicity).

The definition (3) of a linear phase filter requires the filter to have both constant group delay and constant phase delay. However, if only constant group delay is desired, a second type of "linear phase" filter can be defined in which the phase of $H(e^{j\omega})$ is a piecewise linear function of ω, i.e.,

$$H(e^{j\omega}) = \pm |H(e^{j\omega})| e^{j(\beta - \alpha\omega)}. \tag{5}$$

It can be shown[1] that with the constraint (1) on the form of $H(z)$, the only possible solutions for $\beta \in [-\pi, \pi]$ is $\beta = \pm (k\pi/2)$, $k = 0, 1, 2$. If $\beta = 0, \pm\pi$, (5) reduces to (3). Thus the only new cases added are when $\beta = \pm \pi/2$. These cases arise exactly when $z_i = +1$ occurs as a zero of $H(z)$ in odd multiplicity, or, equivalently, when $\{h(k)\}$ satisfies

$$h(k) = -h(N - 1 - k) \qquad 0 \leqq k \leqq N - 1.$$

Filters of this special type are useful in the design of wideband differentiators.[16] However, this type of filter will not be considered in this paper and the term "linear phase filter" will be restricted to refer to those filters satisfying (3).

IV. THEORY OF FIR CASCADE STRUCTURES

4.1 *Roundoff Noise in the Cascade Structure*

The analysis of roundoff noise in this paper is based on the usual model used for such analyses in digital filters.[15,17,18] In particular, each multiplier in a filter is modeled as an infinite-precision multiplier followed by a summation node where roundoff noise is added to the product so that the overall result equals some quantized level. Each noise sample is modeled as a random variable with uniform probability density on the interval $(-Q/2, Q/2)$ and zero density elsewhere, where Q is the quantization step size. Thus each sample is a zero-mean random variable with a variance of $Q^2/12$.

Furthermore, the following assumptions are made:

(*i*) Any two different samples from the same noise source are uncorrelated.

(*ii*) Any two different noise sources (i.e., associated with different multipliers), regarded as random processes, are uncorrelated.

(*iii*) Each noise source is uncorrelated with the input signal.

Thus each noise source is modeled as a discrete stationary white random process with a uniform power density spectrum of magnitude $Q^2/12$.

Applying this model to the filter section shown in Fig. 1, the addition of a noise source to the output of any multiplier is seen to be equivalent to adding a noise source to the output of the section. Therefore, to model a section of a cascade filter, k_i noise sources are added to the output of the section, where k_i is the number of multipliers with non-integer coefficients in the section. Or, equivalently, by assumption (*ii*) above, one noise source of variance $k_i(Q^2/12)$ can be added instead.

For the configuration shown in Fig. 1, k_i is in general equal to 3. However, when $b_{0i} = b_{2i}$, the signals of the two branches feeding the multipliers with coefficients b_{0i} and b_{2i} can first be summed before being multiplied by the common coefficient, thus reducing k_i to 2. Furthermore, by a sacrifice in speed (assuming serial arithmetic), it is possible, as demonstrated by practical hardware,[10] to reduce k_i to 1 for all i by summing all products in each section before performing rounding. It is of interest to point out that the same can be done in the direct form, resulting in effectively only one noise source of variance $Q^2/12$ feeding into the output of the filter.

Before proceeding further, some notations need to be developed. Let $H_i(z)$ denote the transfer function of the ith section of a filter $H(z)$, i.e.,

$$H(z) = \prod_{i=1}^{N_s} H_i(z) \tag{6}$$

where

$$H_i(z) = b_{0i} + b_{1i}z^{-1} + b_{2i}z^{-2}. \tag{7}$$

As a convention, filter sections will be numbered in increasing numbers according to increasing distance from the filter input (i.e., the section at the input is called the 1st section).

Furthermore, define

$$G_i(z) = \begin{cases} \prod_{j=i+1}^{N_s} H_j(z) & 0 \leqq i \leqq N_s - 1 \\ 1 & i = N_s \end{cases} \tag{8}$$

and let $\{g_i(k)\}$ be the impulse response of $G_i(z)$, i.e.,

$$G_i(z) = \sum_k g_i(k)z^{-k}. \tag{9}$$

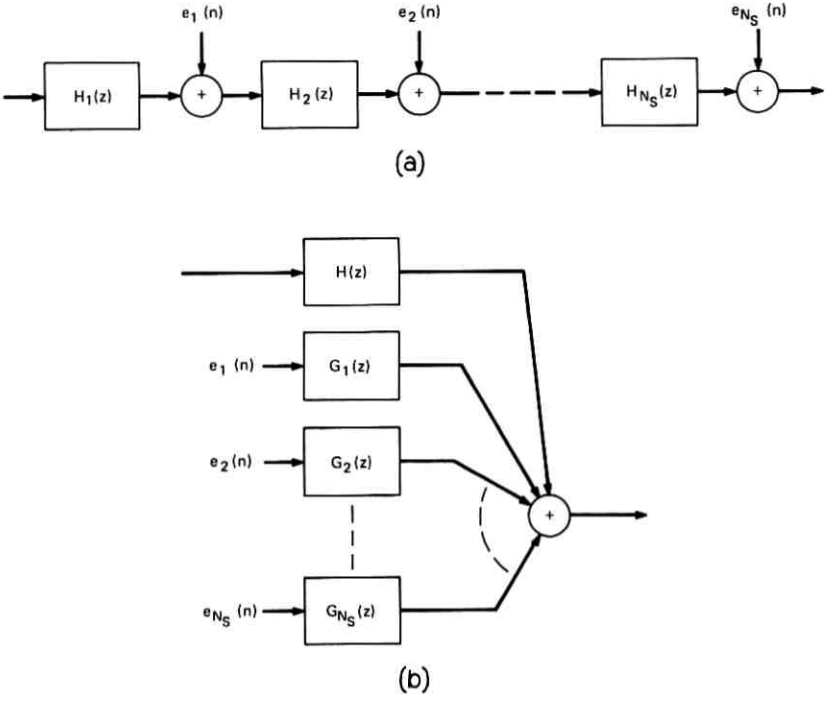


Fig. 2—Equivalent models for a filter in cascade form.

Then a cascade filter can be modeled as in Fig. 2a or equivalently as in Fig. 2b, where $\{e_i(n)\}$ is the noise source for the ith section. Letting $\{E_i(n)\}$ denote the noise sequence at the filter output due to the ith noise source alone gives

$$E_i(n) = \sum_k g_i(k)e_i(n - k). \tag{10}$$

By the stationarity of $\{e_i(n)\}$, the variance of $E_i(n)$ is independent of n; hence denoting this variance by σ_i^2, assumption (i) above leads to the relation

$$\begin{aligned}\sigma_i^2 &= \sum_k g_i^2(k)\overline{e_i(n - k)^2} \\ &= k_i \frac{Q^2}{12} \sum_k g_i^2(k).\end{aligned} \tag{11}$$

Now the total noise output is given by

$$E(n) = \sum_{i=1}^{N_s} E_i(n) = \sum_{i=1}^{N_s} \sum_k g_i(k)e_i(n - k). \tag{12}$$

Therefore by assumptions (*i*) and (*ii*)

$$\sigma^2 = \overline{E^2(n)} = \sum_{i=1}^{N_s} \sigma_i^2. \tag{13}$$

4.2 *Methods of Scaling to Meet Dynamic Range Constraints*

A practical digital filter, necessarily implemented as a physical device, must have a finite dynamic range. Especially when fixed-point arithmetic is employed, this dynamic range sets a practical limit to the maximum range of signal levels representable in a filter and acts to constrain the signal-to-noise ratio attainable.

In some filter structures such as the direct form, given the filter transfer function, the designer has no control over the relative signal levels at points within the filter. Only the gain of the overall filter can be varied. However, in a cascade realization with N_s sections there are $N_s - 1$ degrees of freedom available in addition to the overall filter gain and the ordering of sections.

To see this, a factorization for $H(z)$ is defined which is unique up to ordering of factors, in the form

$$H(z) = \beta \prod_{i=1}^{N_s} \hat{H}_i(z)$$

$$\hat{H}_i(z) = a_{0i} + a_{1i}z^{-1} + a_{2i}z^{-2} \tag{14}$$

where $\{a_{ij}\}$ satisfies

$$a_{0i} \geqq 0, \qquad \sum_{j=0}^{2} |a_{ji}| = 1 \qquad i = 1, \cdots, N_s. \tag{15}$$

Then the transfer function for the ith section in a cascade realization can be written as

$$H_i(z) = S_i \hat{H}_i(z) \tag{16}$$

where S_i is an arbitrary constant, subject only to the constraint that

$$\prod_{i=1}^{N_s} S_i = \beta. \tag{17}$$

Thus given β, $N_s - 1$ of the S_i's can be chosen at will.

Any rule for assigning values to $\{S_i\}$ will be referred to as a scaling method. Obviously, some scaling method must be employed in the design of a cascade filter whether or not one is concerned with dynamic range constraints, since numerical values must be assigned to the S_i's. When dynamic range is an issue, the constraints it imposes can be

met in some best manner by choosing the proper scaling method. In this paper, filters, designed so that no arithmetic overflow in them can cause distortion in the filter output, will be studied. Therefore, the investigation of scaling methods will be restricted to those methods which guarantee that for a given class of input signals no distortion-causing overflow occurs in the scaled filter.

It can be shown[19] that, in an addition operation, if two's complement arithmetic is used, as is usually the case, then as long as the final result is within the representable numerical range, individual partial sums can be allowed to overflow without causing inaccuracies in the result. In this paper, it is assumed that all additions in a filter are done using two's complement arithmetic. Then, to guarantee that no distortion caused by overflow occurs at a cascade filter's output, only the input and output of each filter section need be constrained not to overflow.

To simplify the discussion of scaling methods, the following definitions are used. Let

$$F_i(z) = \sum_{k=0}^{2i} f_i(k)z^{-k} = \prod_{j=1}^{i} H_j(z) \tag{18}$$

and

$$1 \leqq i \leqq N_s$$

$$\hat{F}_i(z) = \sum_{k=0}^{2i} \hat{f}_i(k)z^{-k} = \prod_{j=1}^{i} \hat{H}_j(z). \tag{19}$$

Also, let $\{v_i(n)\}$ be the output sequence of $F_i(z)$ or $H_i(z)$. Furthermore, assume that the maximum magnitude of numerical data representable in a filter is 1.0. Then the necessary overflow constraints on a cascade filter can be stated as

$$|v_i(n)| \leqq 1 \qquad 1 \leqq i \leqq N_s, \qquad \text{all } n. \tag{20}$$

Necessary and sufficient conditions for (20) to hold for two classes of input signals are given below. Theorem 2 deals with the class of input sequences $\{x(n)\}$ which satisfies $|x(n)| \leqq 1$ for all n. For simplicity, this class is refererd to as "class 1." Theorem 3 deals with the class of inputs with transform $X(e^{j\omega})$ which satisfies

$$\frac{1}{2\pi}\int_0^{2\pi} |X(e^{j\omega})|\, d\omega \leqq 1.$$

This class will be called "class 2." By virtue of the fact that

$$x(n) = \frac{1}{2\pi}\int_0^{2\pi} X(e^{j\omega})e^{j\omega n}d\omega \tag{21}$$

and hence

$$|x(n)| \leqq \frac{1}{2\pi} \int_0^{2\pi} |X(e^{j\omega})| \, d\omega, \tag{22}$$

class 2 is a subset of class 1.

Theorem 2: Suppose $|x(n)| \leqq 1$. *Then* $|v_i(n)| \leqq 1$, $1 \leqq i \leqq N_s$ *and all* n, *if and only if*

$$\sum_{k=0}^{2i} |f_i(k)| \leqq 1 \qquad i = 1, \cdots, N_s. \tag{23}$$

Theorem 3: Suppose $1/(2\pi) \int_0^{2\pi} |X(e^{j\omega})| \, d\omega \leqq 1$. *Then* $|v_i(n)| \leqq 1$, $1 \leqq i \leqq N_s$ *and all* n, *if and only if*

$$|F_i(e^{j\omega})| \leqq 1 \qquad i = 1, \cdots, N_s \qquad 0 \leqq \omega \leqq 2\pi. \tag{24}$$

Conditions (23) and (24) of Theorems 2 and 3 can be restated to give conditions on $\{S_i\}$. Recall that the $\hat{H}_i(z)$'s are unique once $H(z)$ is given, hence the $\hat{F}_i(z)$'s and $\{\hat{f}_i(k)\}$'s are also unique. Equations (16), (18), and (19) give

$$f_i(k) = \left(\prod_{j=1}^{i} S_j \right) \hat{f}_i(k) \tag{25}$$

and

$$F_i(z) = \left(\prod_{j=1}^{i} S_j \right) \hat{F}_i(z). \tag{26}$$

Therefore, conditions (23) and (24) can be restated respectively as

$$\left.\begin{aligned} \prod_{j=1}^{i} |S_j| &\leqq \left[\sum_{k=0}^{2i} |\hat{f}_i(k)| \right]^{-1} \qquad (27) \\ \text{and} \qquad \prod_{j=1}^{i} |S_j| &\leqq \left[\max_{0 \leqq \omega \leqq 2\pi} |\hat{F}_i(e^{j\omega})| \right]^{-1} \qquad (28) \end{aligned}\right\} i = 1, \cdots, N_s.$$

These then are conditions which, for the class of inputs concerned, a scaling method must satisfy. It will be shown next that in some sense optimum scaling methods are obtained when (27) and (28) are satisfied with equality. For ease of reference, two scaling methods will first be defined.

Define *sum scaling* to be the rule:

$$\prod_{j=1}^{i} S_j = \left[\sum_{k=0}^{2i} |\hat{f}_i(k)| \right]^{-1} \qquad i = 1, \cdots, N_s \tag{29}$$

or stated recursively,

$$S_i = \begin{cases} \left[\sum_{k=0}^{2} |\hat{f}_1(k)| \right]^{-1} & i = 1 \\ \left[\left(\prod_{j=1}^{i-1} S_j \right) \sum_{k=0}^{2i} |\hat{f}_i(k)| \right]^{-1} & i = 2, \cdots, N_s. \end{cases} \tag{30}$$

Also, define *peak scaling* to be the rule:

$$\prod_{j=1}^{i} S_j = \left[\max_{0 \leqq \omega \leqq 2\pi} |\hat{F}_i(e^{j\omega})| \right]^{-1} \qquad i = 1, \cdots, N_s \tag{31}$$

or

$$S_i = \begin{cases} \left[\max_{0 \leqq \omega \leqq 2\pi} |\hat{F}_1(e^{j\omega})| \right]^{-1} & i = 1 \\ \left[\left(\prod_{j=1}^{i-1} S_j \right) \max_{0 \leqq \omega \leqq 2\pi} |\hat{F}_i(e^{j\omega})| \right]^{-1} & i = 2, \cdots, N_s. \end{cases} \tag{32}$$

Theorem 4: Given an FIR transfer function to be realized in cascade form (as defined in Fig. 1) using fixed-point arithmetic of a given word-length, and given the ordering of filter sections, assume that:

(i) The number of noise sources in each section (i.e., k_i) is independent of the scaling method.
(ii) All filter coefficients can be represented to arbitrary precision.
(iii) No overflow is allowed to occur at the input and output of each section.
(iv) The overall gain of the filter is maximized subject to no overflow at the filter output.

Then each of the following scaling methods is optimum for the class of input signals stated, in the sense that it yields the minimum possible roundoff noise variance as defined in (13) among all scaling methods which satisfy conditions (iii) and (iv) above for the class of inputs considered.

(i) Sum scaling for class 1 signals.
(ii) Peak scaling for class 2 signals.

Thus optimal scaling methods are established for two classes of input signals. It is possible to define other classes of signals by considering the "*Lp* norm" of their transforms.[15,17] Specifically, the *Lp* norm of $X(e^{j\omega})$ is defined as

$$\|X(e^{j\omega})\|_p = \left[\frac{1}{2\pi} \int_0^{2\pi} |X(e^{j\omega})|^p d\omega \right]^{1/p} \qquad 1 \leqq p \leqq \infty \tag{33}$$

where for $p = \infty$ the limit as $p \to \infty$ of the right-hand side is meant. For each p, a class of signals can be defined consisting of those sequences with transforms which satisfy

$$\|X(e^{j\omega})\|_p \leqq 1. \tag{34}$$

Signals satisfying (34) will be referred to as Lp-norm constrained signals. Note that L_1-norm constrained signals are simply class 2 signals.

For proofs of the following useful theorem, refer to Refs. 1, 20, and 21.

Theorem 5: Let $X(e^{j\omega})$ and $Y(e^{j\omega})$ be transforms of sequences. Then

(*i*) $$\|X(e^{j\omega})\|_\infty = \max_{0 \leqq \omega \leqq 2\pi} |X(e^{j\omega})|$$

(*ii*) $$\|X(e^{j\omega})Y(e^{j\omega})\|_1 \leqq \|X(e^{j\omega})\|_p \|Y(e^{j\omega})\|_q$$

if $1/p + 1/q = 1, \quad 1 \leqq p,q \leqq \infty$

(*iii*) $$\|X(e^{j\omega})\|_r \leqq \|X(e^{j\omega})\|_s \quad \text{if} \quad 1 \leqq r \leqq s \leqq \infty .$$

Since with input $\{x(n)\}$,

$$v_i(n) = \frac{1}{2\pi} \int_0^{2\pi} F_i(e^{j\omega})X(e^{j\omega})e^{j\omega n}d\omega, \tag{35}$$

so that

$$|v_i(n)| \leqq \frac{1}{2\pi} \int_0^{2\pi} |F_i(e^{j\omega})X(e^{j\omega})|\,d\omega = \|F_i(e^{j\omega})X(e^{j\omega})\|_1, \tag{36}$$

by Theorem 5 (*ii*),

$$|v_i(n)| \leqq \|F_i(e^{j\omega})\|_p \|X(e^{j\omega})\|_q \qquad 1 \leqq i \leqq N_s. \tag{37}$$

Hence for L_q-norm constrained signals, i.e., if $\|X(e^{j\omega})\|_q \leqq 1$, the following scaling method (L_p-norm scaling) is obtained.

$$\|F_i(e^{j\omega})\|_p = 1 \qquad \begin{matrix} p = \dfrac{q}{q-1} \\ i = 1, \cdots, N_s, \end{matrix} \tag{38}$$

or stated in terms of $\{S_i\}$,

$$\prod_{j=1}^{i} S_j = [\|\hat{F}_i(e^{j\omega})\|_p]^{-1} \qquad i = 1, \cdots, N_s. \tag{39}$$

Notice that by virtue of part (*i*) of Theorem 5, L_∞-norm scaling is just peak scaling which has been shown to be optimum for class 2, or L_1-norm constrained, signals. Furthermore, by part (*iii*) of the theorem,

the following hierarchy of classes of signals is obtained:

class 1 $\supset$ class 2 $\supset$ L_p-norm constrained $\supset$ L_q-norm constrained
if $1 \leqq p \leqq q \leqq \infty$.

In general, class 1 and class 2 signals are the most useful to consider. L_2-norm constrained signals with L_2-norm scaling are useful when all inputs to a filter have finite energy bounded by a known value. For, by Parseval's Theorem, the energy of $\{x(n)\}$ is simply given by $(\|X(e^{j\omega})\|_2)^2$. Hence, if the input signals are first scaled so that their maximum energy is 1.0 (or the squared dynamic range of the filter), then L_2-norm scaling is sufficient to ensure no overflow.

L_2-norm scaling finds greater application for FIR filters than for IIR filters because, in the former case, it is applicable for a larger class of input signals. In particular, an Nth-order FIR filter has only N samples of memory; thus if the input signal to an Nth-order FIR filter consists of bursts of energy spaced more than N samples apart with zero energy in between, then the filter will effectively "see" only one burst at a time. Hence, the maximum energy of a burst can be used as the bound on the energy of the input as far as scaling is concerned. Thus an infinite-energy signal can have the effect of a finite-energy signal on an FIR filter.

Clearly, sum scaling and peak scaling can also be applied to IIR filters.[15] In fact, Theorems 2 and 3 are also valid for IIR filters. However, the input sequence needed in Theorem 2 to prove necessity in the case of IIR filters is an infinite-duration sequence extending to $-\infty$ with full dynamic range magnitudes, and signs that match those of $\{f_i(k)\}$ for some i. Since $\{f_i(k)\}$ for IIR filters is infinite in duration for all i, clearly such an input sequence is highly improbable. Hence, class 1 signals have been deemed too restrictive a description for ordinary inputs to an IIR filter, resulting in too stringent a scaling method.[15]

However, for FIR filters it is not difficult to find an input sequence within dynamic range which will require sum scaling to ensure no overflow, since only a small, finite portion of the sequence need match up with the $\{f_i(k)\}$'s. For example, if $F_1(z)$ has a zero with angle ω_0, $\pi/2 \leqq \omega_0 < \pi$, then all three samples of $\{f_1(k)\}$ have the same sign; hence an input sequence need only have three consecutive samples of value 1 before $|v_1(n)| = \sum_k |f_1(k)|$ for some n.

4.3 *Properties of the Linear Phase Cascade Structure*

Two theorems regarding certain properties of the linear phase cascade form are now given. These results are useful in the investigation of ordering of cascade filter sections.[14]

Theorem 6: Let $H_i(z)$ be the transfer function for the ith section of a linear phase FIR filter in cascade form, where

$$H_i(z) = b_{0i} + b_{1i}z^{-1} + b_{2i}z^{-2},$$

and let ω_i be the angle of one of its zeros, $-\pi \leqq \omega_i \leqq \pi$. Then for all i:

$$(i) \qquad \max_{\omega} |H_i(e^{j\omega})| = \begin{cases} |H_i(e^{j\pi})| & 0 \leqq |\omega_i| < \dfrac{\pi}{2} \\ |H_i(e^{j0})| & \dfrac{\pi}{2} \leqq |\omega_i| \leqq \pi \end{cases}.$$

$$(ii) \qquad \sum_{l=0}^{2} |b_{li}| = \max_{\omega} |H_i(e^{j\omega})| = \max\left(|H_i(e^{j0})|, |H_i(e^{j\pi})|\right).$$

The next theorem is concerned with the equivalence of certain orderings with regard to output noise variance. In particular, it states that with peak scaling each pair of sections in a filter which synthesize reciprocal zeros is completely interchangeable without affecting the output noise variance of the filter. With sum scaling, however, this is not necessarily true. Nevertheless, a weaker condition can be stated which says that, with sum scaling, if every pair of sections which synthesize reciprocal zeros of a filter is interchanged in position, then output noise variance is not changed. Figure 3 illustrates two such equivalent orderings.

Theorem 7: Let $\{H_i(z)\}$ and $\{H_i'(z)\}$ be the section transfer functions of two orderings for a linear phase filter $H(z)$, both scaled by the same method, thus

$$H(z) = \prod_{i=1}^{N_s} H_i(z) = \prod_{i=1}^{N_s} H_i'(z).$$

Then filters with section transfer functions $\{H_i(z)\}$ and $\{H_i'(z)\}$ produce identical output noise variances if either of the following conditions is true:

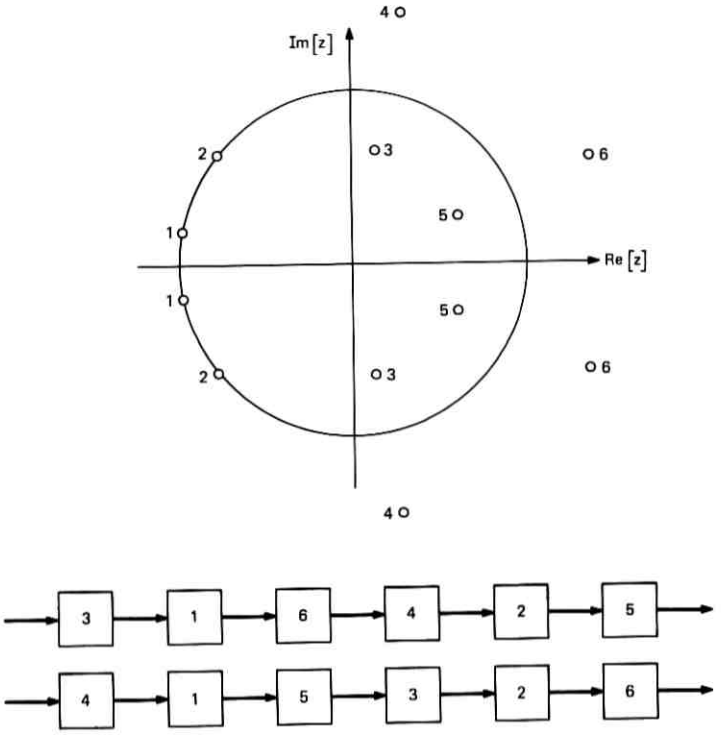


Fig. 3—Two orderings with equal output noise variances.

(*i*) *Peak scaling is used and for each* i, $H_i(z)$ *and* $H_i'(z)$ *have either the same zeros or reciprocal zeros* [*i.e.*, $H_i(z_i) = 0$ *if* $H_i'(z_i^{-1}) = 0$].

(*ii*) *Sum scaling is used and for all* i, z_i^{-1} *is a zero of* $H_i'(z)$ *whenever* z_i *is a zero of* $H_i(z)$.

REFERENCES

1. Chan, D. S. K., "Roundoff Noise in Cascade Realization of Finite Impulse Response Digital Filters," S. B. and S. M. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., September 1972.
2. Rabiner, L. R., "Techniques for Designing Finite-Duration Impulse-Response Digital Filters," IEEE Trans. Commun. Tech., *COM-19*, No. 2 (April 1971), pp. 188–195.
3. Kaiser, J. F., "Digital Filters," ch. 7 in *Systems Analysis by Digital Computer*, F. F. Kuo and J. F. Kaiser, eds., New York: Wiley, 1966.
4. Rabiner, L. R., Gold, B., and McGonegal, C. A., "An Approach to the Approximation Problem for Nonrecursive Digital Filters," IEEE Trans. Audio Electroacoustics, *AU-18*, No. 2 (June 1970), pp. 83–106.
5. Herrmann, O., "Design of Nonrecursive Digital Filters with Linear Phase," Elec. Ltrs., *6*, No. 11, 1970, pp. 328–329.

6. Rabiner, L. R., "The Design of Finite Impulse Response Digital Filters Using Linear Programming Techniques," B.S.T.J., *51*, No. 6 (July-August 1972), pp. 1177–1198.
7. Parks, T. W., and McClellan, J. H., "Chebyshev Approximation for Nonrecursive Digital Filters with Linear Phase," IEEE Trans. Circuit Theory, *CT-19* (March 1972), pp. 189–194.
8. Hofstetter, E., Oppenheim, A. V., and Siegel, J., "A New Technique for the Design of Nonrecursive Digital Filters," Proc. Fifth Annual Princeton Conf. Inform. Sci. and Syst., 1971, pp. 64–72.
9. Current research by L. R. Rabiner, J. F. Kaiser, and O. Herrmann.
10. Schüssler, W., personal communication.
11. Stitt, J. R., and Sonderegger, R. E., "Development and Application of a Programmable Real-Time 200 Coefficient Nonrecursive Digital Filter," Hardware Technical Note of Electronic Communications, Inc., St. Petersburg, Florida, 1969.
12. Rabiner, L. R., and Schafer, R. W., "Recursive and Nonrecursive Realizations of Digital Filters Designed by Frequency Sampling Techniques," IEEE Trans. Audio Electroacoustics, *AU-19*, No. 3 (September 1971), pp. 200–207.
13. Schüssler, W., "On Structures for Nonrecursive Digital Filters," Archiv Für Elektronik Und Übertragungstechnik, Band 26, 1972.
14. Chan, D. S. K., and Rabiner, L. R., "An Algorithm for Minimizing Roundoff Noise in Cascade Realizations of Finite Impulse Response Digital Filters," B.S.T.J., this issue, pp. 347–385.
15. Jackson, L. B., "On the Interaction of Roundoff Noise and Dynamic Range in Digital Filters," B.S.T.J., *49*, No. 2 (February 1970), pp. 159–184.
16. Rabiner, L. R., and Steiglitz, K., "The Design of Wide-Band Recursive and Nonrecursive Digital Differentiators," IEEE Trans. Audio Electroacoustics, *AU-18*, No. 2 (June 1970), pp. 204–290.
17. Jackson, L. B., "Roundoff-Noise Analysis for Fixed-Point Digital Filters Realized in Cascade or Parallel Form," IEEE Trans. Audio Electroacoustics, *AU-18*, No. 2 (June 1970), pp. 107–122.
18. Weinstein, C. J. "Quantization Effects in Digital Filters," Technical Report 468, Lincoln Laboratory, Lexington, Mass., November 1969.
19. Chu, Y., *Digital Computer Design Fundamentals*, New York: McGraw-Hill, 1962.
20. Fleming, W. H., *Functions of Several Variables*, Reading, Massachusetts: Addison-Wesley, 1965, pp. 200–204.
21. Rice, J. R., *The Approximation of Functions*, Reading, Massachusetts: Addison-Wesley, 1964, pp. 4–10.

The Bell System Technical Journal
Vol. 52, No. 3, March, 1973
Printed in U.S.A.

An Algorithm for Minimizing Roundoff Noise in Cascade Realizations of Finite Impulse Response Digital Filters

By D. S. K. CHAN and L. R. RABINER

(Manuscript received September 14, 1972)

Experimental results on roundoff noise in cascade realizations of Finite Impulse Response (FIR) digital filters are presented in this paper. The entire roundoff noise distribution (i.e., over all possible orderings) is given for several low-order filters using both sum and peak scaling. Based on observations about this distribution, as well as intuitive arguments about the effects of ordering on roundoff noise, an algorithm for minimizing roundoff noise is presented. Experimental verification of this algorithm for a wide range of filters is given.*

I. INTRODUCTION

As discussed in previous works,[1,2] the implementation of FIR filters using finite precision arithmetic has become an important issue in recent years. For cascade realizations of FIR filters, roundoff noise is a crucial problem. In Refs. 1 and 2, some of the theoretical bases for the analysis of roundoff noise in the FIR cascade form have been considered. This paper presents a large body of experimental results which depict the dependence of roundoff noise on several of the important parameters of a cascade FIR low-pass filter. Most importantly, these results point to an algorithm which can find efficiently, for a cascade filter, an ordering which has a noise variance very close to the minimum possible. Experimental verification of this algorithm for a wide range of filters is presented.

Low-pass, extraripple[3] filters are used throughout these investigations as being representative of FIR filters. It will be seen that most results

* This paper is based on a thesis[1] submitted in partial fulfillment of the requirements for the degrees of Bachelor of Science and Master of Science in the Department of Electrical Engineering at the Massachusetts Institute of Technology in September 1972.

should not depend on the type of filter used. Figure 1 shows the magnitude response of a typical extraripple filter (which by definition has linear phase) and the parameters which define it. Of the four parameters F_1, F_2, D_1, and D_2, only three can be independently specified. The parameters N (impulse response length), F_1, D_1, and D_2 will be chosen as independent variables in these investigations. The studied ranges of variation of these parameters are as follows: $7 \leqq N \leqq 129, 0.1 \geqq D_1 \geqq 0.001$, $0.1 \geqq D_2 \geqq 0.001$, $0 < F_1 < 0.5$, and $0 < F_2 < 0.5$ (sampling frequency = 1). These ranges comprise a large range of the significant values that these parameters take on. In the present state of the art in real-time digital filter hardware, 128th-order ($N = 129$) is the highest order that can be implemented in cascade form at a sampling rate of 10 kHz (e.g., typical for speech processing).[4,5] Furthermore, the stated ranges for D_1 and D_2 are those significant for many speech processing systems.[6]

While a great deal of experimental data has been collected, only representative examples will be presented here. For more examples see Ref. 1.

II. PRELIMINARY REMARKS

In Ref. 2 it is shown that given a transfer function $H(z)$ to be realized in cascade form and the order in which the factors of $H(z)$ are to be synthesized, there remain N_s degrees of freedom (including gain of filter) in the choice of filter coefficients, where N_s is the number of sections of the filter. Scaling methods are developed to fix these N_s degrees of freedom, and two particular methods, viz., sum scaling and peak scaling,* are shown to be optimum for the particular classes of input signals which they assume. These scaling methods will be applied in this paper.

The prime issues in the realization of filters in cascade form are threefold–scaling, ordering, and section configuration. Because of the simplicity of a 2nd-order FIR filter, there is little freedom in the choice of a structure for the sections of a cascade filter. In Ref. 2, the configuration shown in Fig. 2a is assumed because it turns out to be the most useful in a practical situation. Another configuration, Fig. 2b, is also mentioned in Ref. 2. Because, as seen from Fig. 2c, these two configurations can be readily accommodated in a more general sub-

* Sum (peak) scaling is defined to be a method of scaling where the scale factors for the cascade sections are chosen so that the sum of the impulse response magnitudes (peak of the magnitude of the frequency response) up to that section does not exceed one.

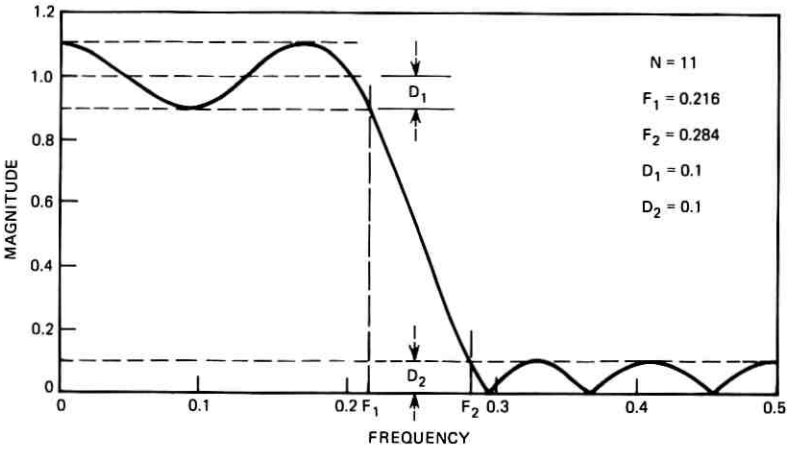


Fig. 1—Definition of filter parameters D_1, D_2, F_1, and F_2.

filter structure, it is here assumed that the configurations of Figs. 2a and b are both used in a cascade structure, depending on whether $b_{0i} \neq b_{2i}$ or $b_{0i} = b_{2i}$ respectively. The configuration of Fig. 2b has the advantage of having lower noise than the configuration of Fig. 2a. The option of summing all products in an increased length register before rounding is also possible for all configurations. However, the gain in signal-to-noise ratio does not seem to be worth the required sacrifice in speed (assuming serial arithmetic).[7] In any case, all resulting noise variances would simply be scaled down by a factor of from 2 to 3 if this strategy were used instead of rounding after each multiplication, as assumed here.

Other possible section configurations will be discussed later on. Since scaling is treated in depth in Ref. 2, the major concern here is the ordering of sections. Unlike the scaling problem, no workable optimal solution (in terms of feasibility) to the ordering problem has yet been found for cascade structures in general. The dependence of output roundoff noise variance on section ordering, given a scaling method, is so complex that no simple indicators are provided to assist in any systematic search for an ordering with lowest noise. Any attempt to find the noise variances for all possible orderings of a filter involves on the order of $N_s!$ evaluations, which clearly becomes prohibitive even for moderately large values of N_s. Thus there is little doubt that optimal ordering with time constraint is by far the most difficult issue to deal with in the design of filters in cascade form.

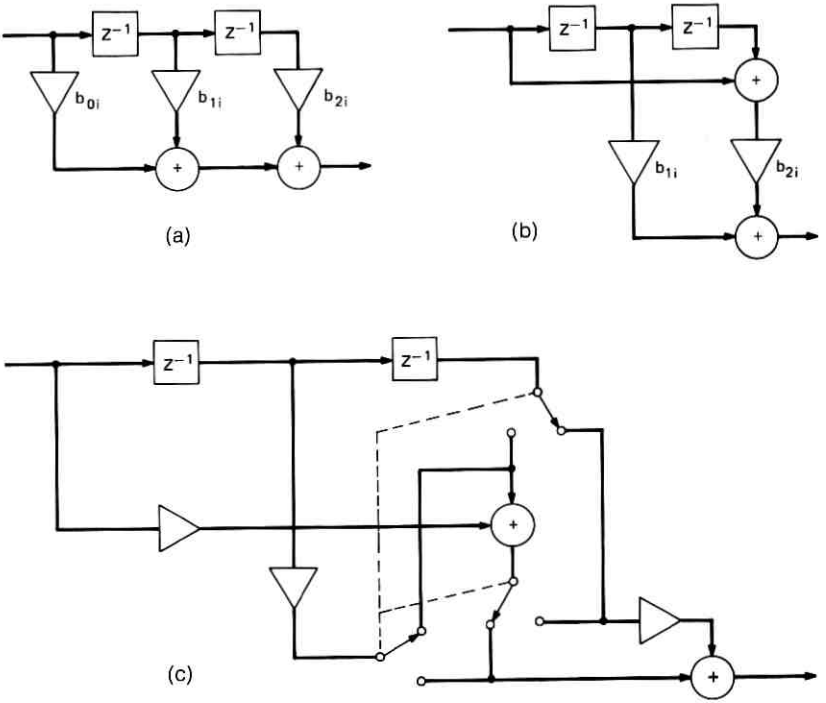


Fig. 2—(a) Cascade form filter section. (b) Alternate cascade filter section. (c) General cascade filter section.

Since finding an optimal solution to the ordering problem through exhaustive searching is very time consuming (if not impossible in any feasible amount of time) for all but very low-order filters, it is important to find out how closely a suboptimal solution can approach the optimum and how difficult it would be to find a satisfactory suboptimal solution. Even this concern, however, would be unfounded if the roundoff noise level produced by a filter were rather insensitive to ordering. Then the difference in performance between any two orderings may not be sufficient to cause any concern. However, Schüssler has demonstrated that quite the contrary is true.[7,8] He showed a 33-point FIR filter which, ordered one way, produces $\sigma^2 = 2.4\ Q^2$ (where Q is the quantization step size of the filter), while ordered another way yields $\sigma^2 = 1.5 \times 10^8\ Q^2$ (assuming all products in each section are summed before rounding). This represents a difference of 1.6 bits versus 14.6 bits of noise. Clearly, the difference is large enough so that the

problem of finding a proper ordering of sections in the design of a cascade filter cannot be evaded.

An important question to pursue in investigating suboptimal solutions is whether or not there exists some general pattern in which values of noise variances distribute themselves over different orderings. For example, for the 33-point filter mentioned above, are all noise values between the two extremes demonstrated equally likely to occur in terms of occurring in the same number of orderings? Or, perhaps, only a few pathological orderings have noise variance as high as that indicated. On the other hand, perhaps only very few orderings have noise variances close to the low value, in which case an optimum solution would be very valuable, while a satisfactory suboptimal solution may be just as difficult to obtain as the optimum.

In the next section, these questions will be answered by investigating filters of sufficiently low order so that calculating noise variances of $N_s!$ different orderings is not an unfeasible task. The implications of results obtained will then be generalized.

III. CALCULATION OF NOISE DISTRIBUTIONS

3.1 *Methods*

The definitions of sum scaling and peak scaling in Ref. 2 indicate that, for FIR filters, sum scaling is much simpler to perform than peak scaling. To achieve peak scaling, the maxima of the functions $\hat{F}_i(e^{j\omega})$* must be found for all i given an ordering. Even using the FFT, this represents considerably more calculations than finding $\sum_{k=0}^{2i} |\hat{f}_i(k)|$ for all i as is necessary for sum scaling. In the 33-point filter mentioned above, Schüssler used peak scaling on both the orderings. It will be shown in Section IV that, given a filter, peak and sum scaling yield noise variances that are not very different (within the same order of magnitude), and, in fact, experimental results indicate that they are essentially in a constant ratio to one another independent of ordering of sections. Hence the general characteristics of the distribution of roundoff noise with respect to orderings should be quite independent of the type of scaling performed. In order to save computation time, sum scaling will be used in these investigations.

Returning to the question of section configuration, for Infinite Impulse Response (IIR) filters Jackson[9] has introduced the concept of transpose configurations to obtain alternate structures for filter sec-

* $\hat{F}_i(e^{j\omega})$ and $\{\hat{f}_i(k)\}$ are defined in Ref. 2 and are the frequency response and impulse response of the cascade of sections 1 to i.

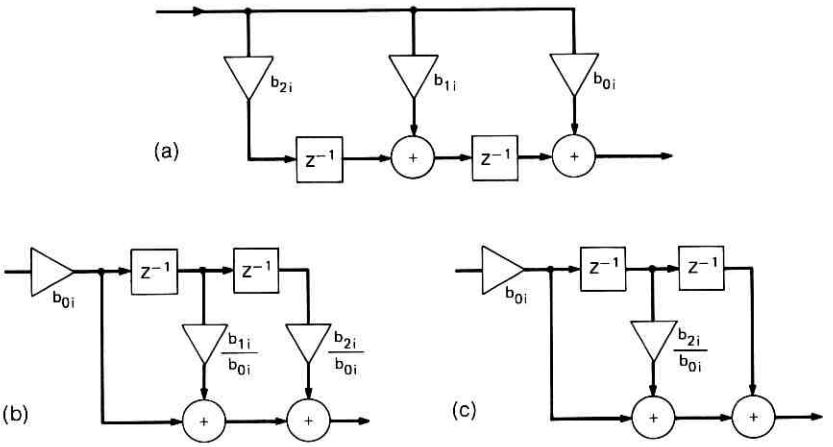


Fig. 3—(a) Transpose configuration of Fig. 2a. (b) Alternate configuration of Fig. 2a. (c) Alternate configuration of Fig. 2b.

tions. However, the application of this concept to Fig. 2a yields the structure shown in Fig. 3a, which is seen to have the same noise characteristics as the structure in Fig. 2a since, by the whiteness assumption on the noise sources, delays have no effect on them. Therefore, the structure of Fig. 3a has no advantages over other structures as far as roundoff noise is concerned. The only other significant alternate configuration for Fig. 2a is shown in Fig. 3b. The counterpart for Fig. 2b is Fig. 3c and is valid when $b_{0i} = b_{2i}$. Both of these new configurations have exactly the same number of multipliers as the original ones. However, one noise source is moved from the output to essentially the input of the section. Thus it is advantageous to use the structures in Figs. 3b and c for the ith section when

$$\frac{1}{b_{0i}^2} \sum_k g_{i-1}^2(k) < \sum_k g_i^2(k) \tag{1}$$

where $\{g_i(k)\}$ is, as defined in Ref. 2, the impulse response of the equivalent filter seen by the ith noise source. However, in order to have no error-causing internal overflow when the input and output of a section are properly constrained, the structures of Figs. 3b and c can be used only when $b_{0i} \leqq 1$. If $b_{0i} > 1$, either four multipliers are required, or Fig. 3b reduces to Fig. 2a.

In the investigations that follow, for each section of a filter, the configuration among Figs. 2a, 2b, 3b, and 3c which is applicable and

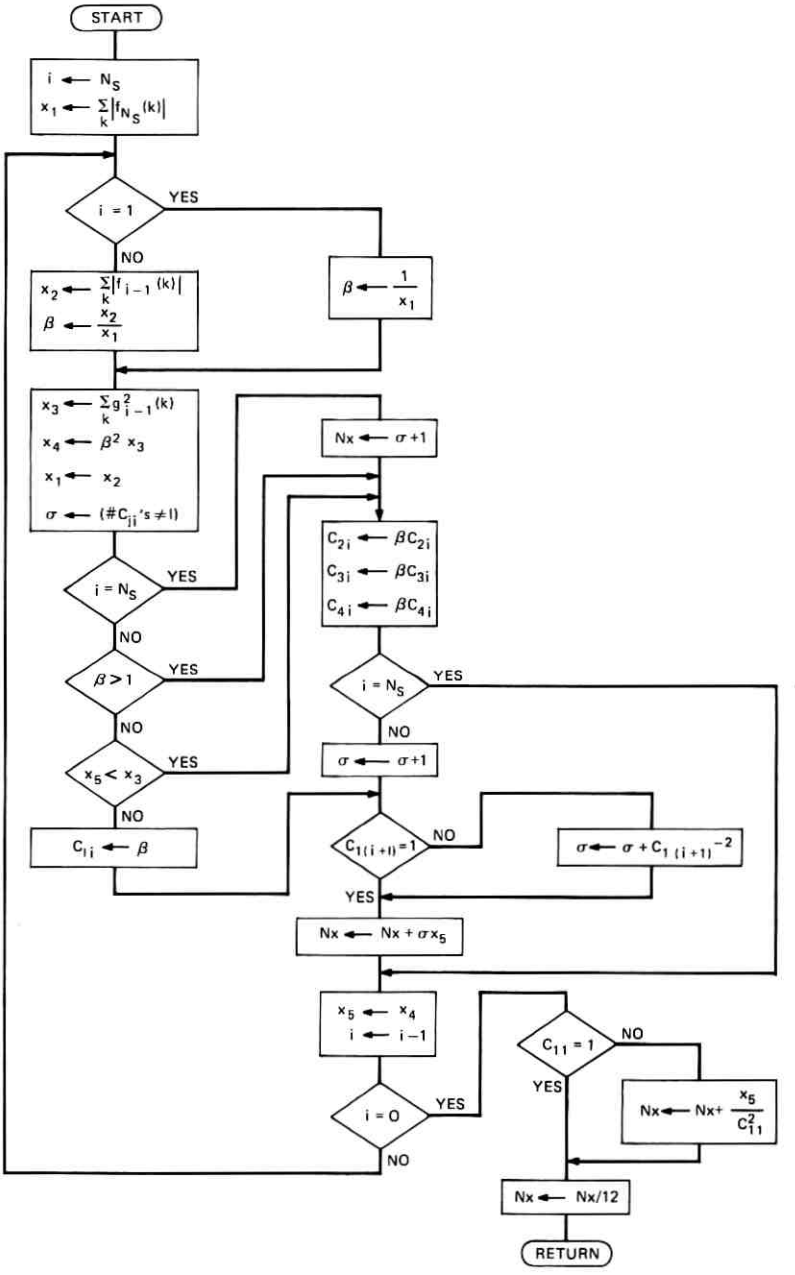


Fig. 4—Flow chart of scaling and noise calculation subroutine.

results in the least noise will be employed. It turns out that this flexibility in the choice of configuration has little effect on the noise distribution characteristics of a filter. For low-noise orderings, the configurations of Figs. 2a and b are almost always more advantageous than the other configurations. For high-noise orderings, the alternate configurations help to reduce the noise variance, but the difference is comparatively small. Thus in actual filter implementations the structures in Figs. 3b and c may be ignored.

Figure 4 is the flow diagram of a computer subroutine which is used to accomplish scaling, choice of configuration, and output noise variance calculation given a filter and its ordering. The input to the subroutine consists of N_s (the number of sections) and the sequence $\{C_{ji},\ 1 \leqq j \leqq 4,\ 1 \leqq i \leqq N_s\}$, the elements of which are unscaled coefficients of the filter, defined by

$$H_i(z) = C_{1i}(C_{2i} + C_{3i}z^{-1} + C_{4i}z^{-2}) \qquad 1 \leqq i \leqq N_s \tag{2}$$

where $H_i(z)$ is the ith section in the filter cascade. The sequences $\{f_i(k)\}$ and $\{g_i(k)\}$ in Fig. 4 are the impulse responses of the cascade of the first i sections and the last $(N_s - i)$ sections respectively. The coefficients $\{C_{ji}\}$ on input are assumed to be normalized so that, for all i, $C_{1i} = 1$ and at least one of C_{2i} and C_{4i} equals 1. On return $\{C_{ji}\}$ contains the scaled coefficients and Nx is the value of output noise variance computed in units of Q^2, where Q is the quantization step size of the filter.

Using this subroutine, the noise output of all possible orderings of several FIR filters ranging from $N_s = 3$ to $N_s = 7$ was investigated. By Theorem 7(*ii*) in Ref. 2, for any filter with at least one set of two complex conjugate pairs of reciprocal zeros, there are at most $N_s!/2$ orderings that differ in output noise variance. This is true since if all orderings are divided into two groups, according to the order in which the reciprocal zeros are synthesized in the cascade, then Theorem 7(*ii*) establishes a one-to-one correspondence between each ordering of one group and some ordering of the other group. Thus, in the investigation of all possible noise outputs of a filter, where possible, a pair of sections which synthesize reciprocal zeros is chosen and all orderings in which a particular one of these sections precedes the other are ignored.

3.2 *Discussion of Results*

Using the methods and procedures described, the noise distributions of 27 different linear phase, low-pass extraripple filters were investi-

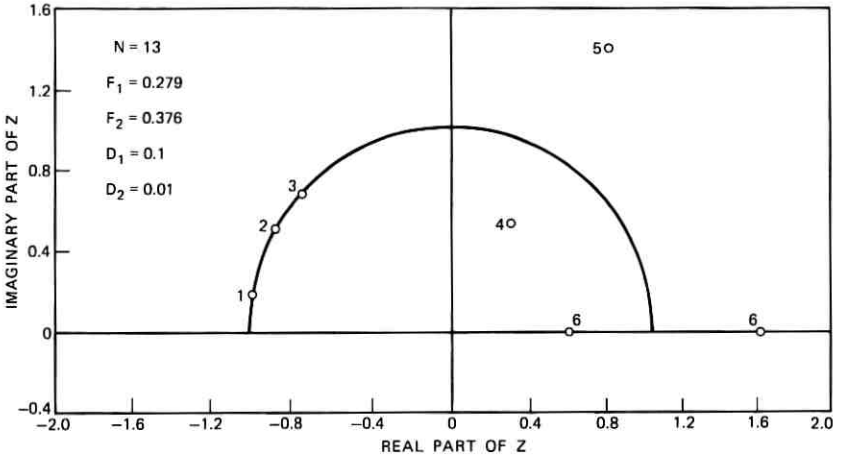


Fig. 5—Positions of the zeros of one filter.

gated. Twenty-two of these filters are 13-point filters, since $N = 13$ represents a reasonable filter length to work with. Thirteen-point filters have six sections each, corresponding to 6! or 720 possible orderings of sections. By reducing redundancy via Theorem 7 of Ref. 2, the number of orderings it is necessary to investigate reduces to 360 for all but 2 of the 22 filters.

The results of the investigations for all 27 filters will eventually be presented. Meanwhile, attention is focused on a typical 13-point filter. As an example, a filter with 4 ripples in the passband, 3 ripples in the stopband, and passband and stopband tolerances of 0.1 and 0.01, respectively, is used. By passband and stopband tolerances is meant the maximum height of ripples in the respective frequency bands. Figure 5 shows the positions of the zeros of the filter in the upper half of the z-plane. Each section of the filter is given a number for identification. The zeros that a section synthesizes are given the same number, and these are shown in Fig. 5. Table I shows a list, in order of increasing noise magnitude, of all 360 orderings investigated and their corresponding output noise variances in units of Q^2, computed according to Fig. 4. A histogram plot of the noise distribution is shown in Fig. 6a, and a cumulative distribution plot is shown in Fig. 6b.

Two characteristics of the histogram shown in Fig. 6a are of special importance because they are common to similar plots for all the filters investigated. First of all, most significant is the shape of the distribution. It is seen that most orderings have very low noise compared to

TABLE I—NOISE VARIANCE OF ALL 360 ORDERINGS OF A 13-POINT FILTER

Order	Noise	Order	Noise	Order	Noise
263451	1.0983	416253	1.5957	621453	2.4335
145263	1.1104	341625	1.6081	245613	2.4699
145362	1.1131	436152	1.6170	134652	2.4854
163452	1.1382	142635	1.6286	236415	2.5285
245163	1.1601	412653	1.6354	613425	2.5443
245361	1.1605	421653	1.6418	314652	2.5643
362451	1.1834	241635	1.6458	623415	2.5703
246351	1.2305	243615	1.6524	631425	2.6047
162453	1.2456	243561	1.6539	632415	2.6073
361452	1.2561	164352	1.6583	425631	2.6090
261453	1.2783	264153	1.6817	612435	2.6228
143652	1.2841	346215	1.6829	621435	2.6461
146352	1.3245	346125	1.6904	425613	2.6785
415263	1.3298	364251	1.7040	145632	2.6801
415362	1.3325	164253	1.7101	134562	2.6991
243651	1.3356	413562	1.7171	145623	2.6993
345261	1.3546	342615	1.7177	624351	2.7021
345162	1.3568	342561	1.7192	326415	2.7167
246153	1.3652	413625	1.7449	134625	2.7269
346251	1.3660	431562	1.7483	126453	2.7639
341652	1.3666	426315	1.7560	314562	2.7779
425163	1.3687	431625	1.7762	234651	2.7927
425361	1.3692	412563	1.7835	314625	2.8058
146253	1.3763	416325	1.7854	634251	2.8110
163425	1.3797	426135	1.7865	614352	2.8229
342651	1.4009	364152	1.7869	624153	2.8368
263415	1.4151	421563	1.7900	614253	2.8747
142653	1.4160	416235	1.8083	634152	2.8939
241653	1.4332	412635	1.8480	415632	2.8994
426351	1.4392	436215	1.8509	216453	2.9085
346152	1.4489	421635	1.8544	415623	2.9187
162435	1.4582	436125	1.8585	126435	2.9765
261435	1.4909	423615	1.8610	324651	2.9809
361425	1.4976	423561	1.8626	624315	3.0190
143562	1.4977	264315	1.8638	624135	3.0495
362415	1.5002	432615	1.8858	614325	3.0644
413652	1.5034	432561	1.8873	614235	3.0873
435261	1.5227	264135	1.8943	234615	3.1095
435162	1.5249	164325	1.8998	234561	3.1110
143625	1.5256	164235	1.9227	216435	3.1211
436251	1.5341	364215	2.0208	634215	3.1278
431652	1.5347	364125	2.0284	634125	3.1354
416352	1.5439	136452	2.0700	124653	3.2439
423651	1.5442	316452	2.1489	324615	3.2977
264351	1.5470	236451	2.2117	324561	3.2992
246315	1.5474	623451	2.2534	214653	3.3885
142563	1.5642	632451	2.2904	124563	3.3920
146325	1.5660	613452	2.3028	124635	3.4565
432651	1.5690	136425	2.3115	246531	3.4977
426153	1.5739	631452	2.3632	214563	3.5367
246135	1.5778	316425	2.3904	246513	3.5672
341562	1.5803	326451	2.3999	214635	3.6011
241563	1.5814	245631	2.4004	426531	3.7063
146235	1.5889	612453	2.4102	426513	3.7758

TABLE I—*Continued.*

Order	Noise	Order	Noise	Order	Noise
264531	3.8141	341526	5.8993	413265	16.5228
264513	3.8836	452613	5.9446	431265	16.5541
163245	4.0265	413526	6.0362	463521	16.5661
146532	4.0376	345216	6.0375	463512	16.6163
146523	4.0569	346521	6.0426	412365	16.6522
162345	4.0697	425316	6.0520	421365	16.6587
261345	4.1025	431526	6.0674	134265	17.5048
361245	4.1445	346512	6.0928	314265	17.5837
345621	4.2234	451632	6.1475	243165	17.7595
416532	4.2570	451623	6.1668	342165	17.8249
345612	4.2737	435216	6.2055	423165	17.9682
416523	4.2763	436521	6.2106	432165	17.9929
263145	4.2914	436512	6.2609	124365	18.2607
164532	4.3715	243516	6.3368	214365	18.4053
362145	4.3766	364521	6.3805	234165	19.2166
164523	4.3907	342516	6.4021	324165	19.4048
435621	4.3915	364512	6.4307	642351	21.7670
435612	4.4417	423516	6.5454	643251	21.8232
451263	4.5778	432516	6.5701	641352	21.8995
451362	4.5806	134526	7.0181	642153	21.9017
452163	4.6348	314526	7.0970	643152	21.9061
452361	4.6353	124536	7.2674	641253	21.9513
453261	4.7361	214536	7.4120	642315	22.0838
453162	4.7383	634521	7.4875	642135	22.1143
136245	4.9584	634512	7.5377	643215	22.1401
624531	4.9693	453621	7.6049	641325	22.1410
145236	4.9857	453612	7.6551	643125	22.1476
245136	5.0354	234516	7.7939	641235	22.1639
316245	5.0372	324516	7.9820	143256	23.0109
624513	5.0388	451236	8.4532	341256	23.0934
613245	5.1911	452136	8.5101	142356	23.1403
415236	5.2051	451326	8.8996	241356	23.1575
612345	5.2343	453126	9.0573	413256	23.2302
425136	5.2440	452316	9.3181	431256	23.2615
631245	5.2515	453216	9.4189	412356	23.3596
621345	5.2577	462351	11.8333	421356	23.3661
236145	5.4048	463251	11.8896	642531	24.0341
145326	5.4322	461352	11.9658	642513	24.1036
142536	5.4395	462153	11.9680	134256	24.2122
623145	5.4466	463152	11.9725	314256	24.2911
241536	5.4567	461253	12.0176	243156	24.4670
632145	5.4836	462315	12.1501	342156	24.5323
614532	5.5361	462135	12.1806	641532	24.6126
614523	5.5553	463215	12.2064	641523	24.6319
126345	5.5880	461325	12.2073	423156	24.6756
326145	5.5930	463125	12.2140	432156	24.7003
415326	5.6515	461235	12.2302	124356	24.9681
412536	5.6589	462531	14.1004	214356	25.1128
421536	5.6653	462513	14.1699	234156	25.9240
345126	5.6759	461532	14.6789	324156	26.1122
216345	5.7326	461523	14.6982	643521	26.4998
143526	5.8168	143265	16.3035	643512	26.5500
245316	5.8434	341265	16.3860	132645	81.4953
435126	5.8440	142365	16.4329	312645	81.5742
452631	5.8751	241365	16.4501	231645	85.2733

TABLE I—*Continued.*

Order	Noise	Order	Noise	Order	Noise
321645	85.4615	231465	116.6240	465213	180.9850
123645	87.0142	321465	116.8120	465132	181.3550
213645	87.1589	123465	118.3650	465123	181.3750
456231	99.7815	213465	118.5090	465321	182.8770
456213	99.8510	132456	119.5530	465312	182.9270
456132	100.2220	312456	119.6320	645231	190.8490
456123	100.2410	231456	123.3310	645213	190.9180
456321	101.7430	321456	123.5190	645132	191.2890
456312	101.7940	123456	125.0720	645123	191.3080
132465	112.8460	213456	125.2170	645321	192.8110
312465	112.9250	465231	180.9150	645312	192.8610

the maximum value possible. In fact, the lowest range of noise variance, in this case between zero and 2 Q^2, is the most probable range in terms of the number of orderings which produce noise variances in this range. The distribution is seen to be highly skewed, with an expected value very close to the low-noise end, in this case equal to 19.5 Q^2. In fact, from the cumulative distribution it is seen that approximately two-thirds of the orderings have noise variances less than 4 percent of the maximum, while nine-tenths of them have noise variances less than 14 percent of the maximum.

The second characteristic is that large gaps occur in the distribution so that noise values within the gaps are not produced by any orderings. While Fig. 6a shows this effect only for the higher noise values, a more detailed plot of the distribution in the range from zero to 28 Q^2, as in Fig. 6c, shows that gaps also occur for lower noise values. Thus noise values tend to occur in several levels of clusters. These observations provide the general picture of clusters of noise values, the separation of which increases rapidly as a function of the magnitude of the noise values, thus forming a highly skewed noise distribution.

The significance of these results is far-reaching. Given a specific filter, because of the abundance of orderings which yield almost the lowest noise variance possible, it is concluded that it should not be too difficult to devise a feasible algorithm which will yield an ordering, the noise variance of which is very close to the minimum. Thus, as far as designing practical cascade filters is concerned, it really is not crucial that the optimum ordering be found. In fact, it may be far more advantageous to use a suboptimal method, which can rapidly choose an ordering that is satisfactory, than to try to find the optimum. The reduction in roundoff noise gained by finding the optimum solu-

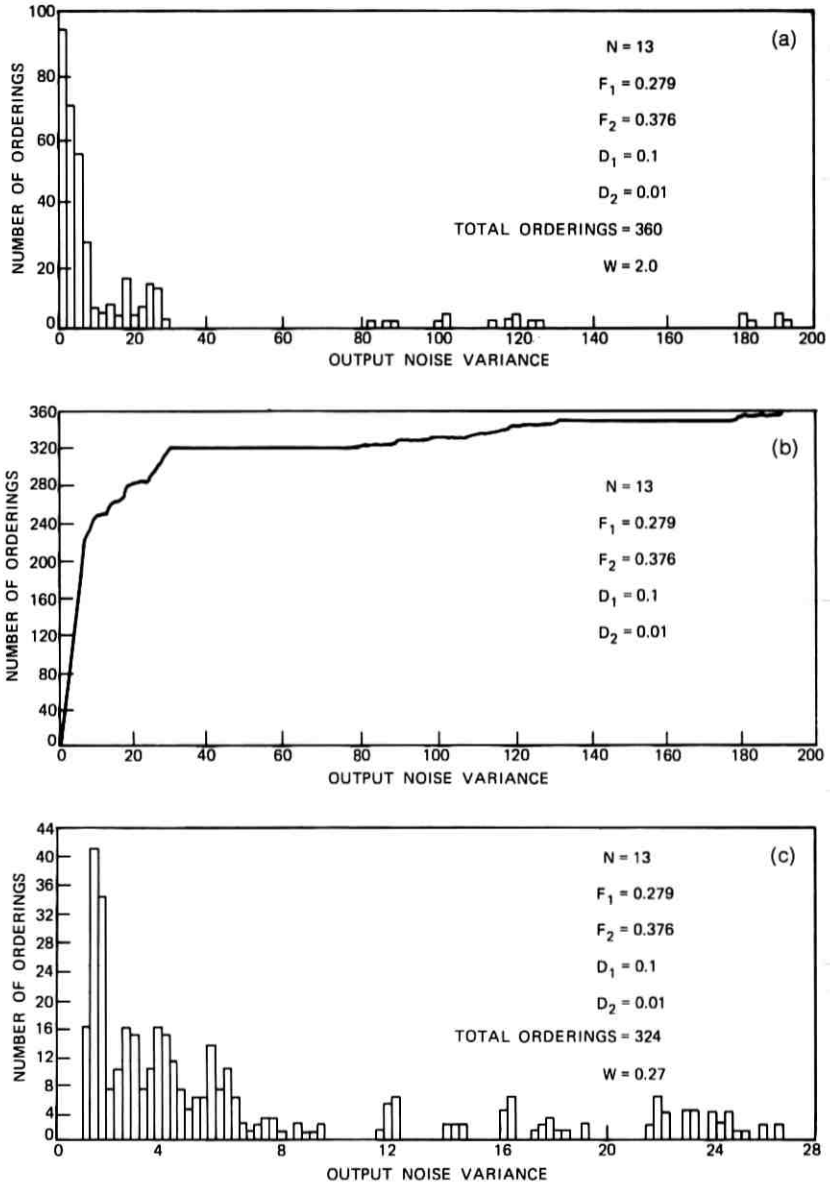


Fig. 6—(a) Noise distribution histogram of filter of Fig. 5. (b) Cumulative noise distribution of filter of Fig. 5. (c) Detailed noise distribution histogram of filter of Fig. 5.

tion is probably, at best, not worth the extra effort from the design standpoint. At least up to the present, no efficient method for finding an optimum ordering has been found.

In Section VI, a suboptimal method is presented which, given a filter, yields a low-noise ordering efficiently and has been successfully applied to a wide range of filters. Before presenting the algorithm, the behavior of roundoff noise with respect to scaling and other filter parameters will be further investigated. Also, the nature of high-noise and low-noise orderings will be discussed, so that they can be more easily recognized.

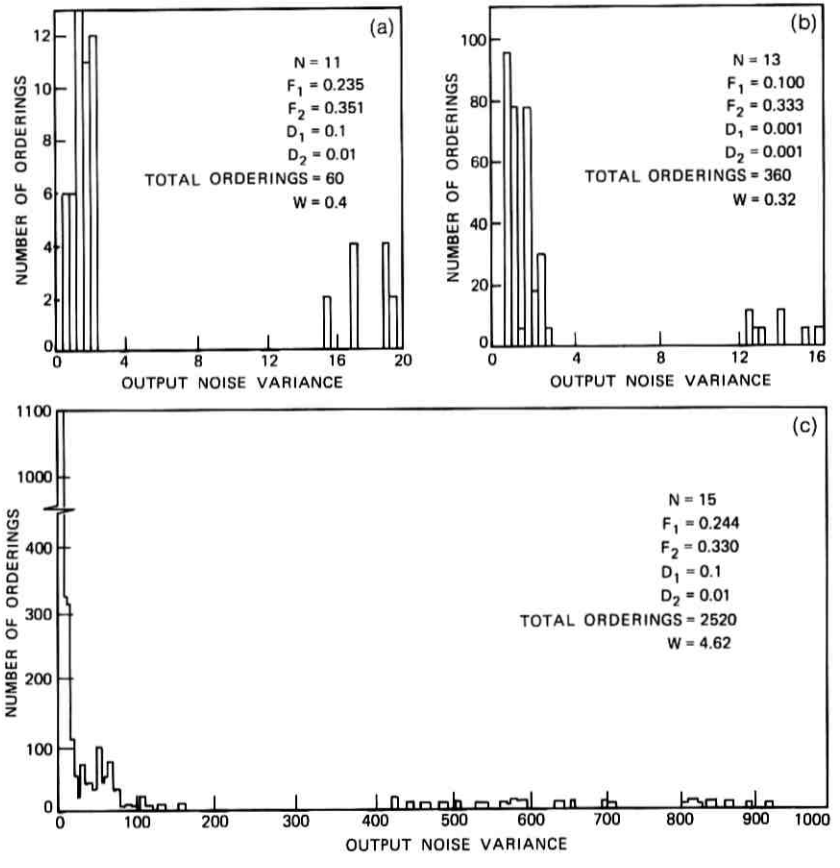


Fig. 7—(a) Noise distribution histogram of typical 11-point filter. (b) Noise distribution histogram of another 13-point filter. (c) Noise distribution histogram of a 15-point filter.

Before ending this section, the noise distribution histograms of an 11-point and one more 13-point filter are presented in Figs. 7a and b. These are seen to exhibit all the characteristics discussed above. The major difference between the noise distributions of the two 13-point filter examples presented lies in the magnitude of the maximum and average noise variances. This difference will be accounted for presently.

Also presented is the noise distribution for a 15-point filter, in Fig. 7c. The calculation of this distribution involves 2520 different orderings. This histogram shows even stronger emphasis on the distribution characteristics discussed and, together with Figs. 6a and 7a, suggests that the skewed shape and large-gap properties of the noise distribution of a filter become increasingly pronounced as the order of the filter increases. Thus it is expected that the results presented can be generalized for higher-order filters.

3.3 *Dependence of Distributions on Transfer Function Parameters*

From all the calculated noise distributions of the previous section, the interesting fact is observed that, though different filters may produce very different ranges of output noise variances when ordered in all possible ways, the noise variances for each filter always distribute themselves in essentially the same general pattern. The differences in noise variance ranges among different filters is accounted for by investigating the dependence of noise distributions on parameters which specify the transfer function of a filter.

The noise distributions of several low-pass extraripple filters with various values of the parameters, N, F_1, D_1, and D_2 were computed using the methods described. Since all these distributions have the same general shape, they can be compared by simply examining their maximum, average, and minimum values. A list of all the filters, the noise distributions of which have been computed, including those already discussed, is presented in Table II. These filters are specified by five parameters, namely the four already mentioned, plus N_p, the number of ripples in the passband. Since all the filters are extraripple filters, it is more natural to specify N_p than F_1. Of course, N_p and F_1 are not independent. The maximum, average, and minimum values of the noise distributions of each of these filters are listed in Table II. The last column in this table will be discussed in Section VI.

Filters numbered 1 to 5 in Table II are very similar except for their order in that they all have identical passband and stopband tolerances and approximately the same low-pass bandwidth. The maximum, average, and minimum values of their noise distributions are plotted

TABLE II—LIST OF FILTERS AND THEIR NOISE DISTRIBUTION STATISTICS

#	N	N_p	F_1	D_1	D_2	Noise Variance			
						Max	Avg	Min	Alg 1
1	7	2	0.212	0.1	0.01	1.24	0.84	0.37	0.37
2	9	3	0.281	0.1	0.01	6.26	2.54	0.73	0.73
3	11	3	0.235	0.1	0.01	19.41	4.79	0.68	0.68
4	13	4	0.279	0.1	0.01	192.86	19.55	1.10	1.10
5	15	4	0.244	0.1	0.01	923.63	54.45	1.02	1.16
6	13	3	0.100	0.001	0.001	15.84	3.01	0.65	0.69
7	13	4	0.261	0.05	0.004	119.48	12.91	0.96	1.02
8	13	1	0.012	0.01	0.01	9.91	1.61	0.32	0.35
9	13	2	0.067	0.01	0.01	16.30	2.94	0.44	0.47
10	13	3	0.138	0.01	0.01	42.63	5.94	0.71	0.73
11	13	4	0.213	0.01	0.01	69.76	8.52	0.82	0.91
12	13	5	0.288	0.01	0.01	76.43	11.01	1.44	1.52
13	13	6	0.364	0.01	0.01	52.54	10.33	1.92	2.43
14	13	3	0.201	0.1	0.01	96.25	12.09	0.81	
15	13	3	0.179	0.05	0.01	69.26	9.02	0.76	
16	13	3	0.154	0.02	0.01	50.63	6.87	0.72	
17	13	3	0.123	0.005	0.01	37.36	5.33	0.70	
18	13	3	0.106	0.002	0.01	32.83	4.80	0.69	
19	13	3	0.095	0.001	0.01	30.53	4.53	0.69	
20	13	3	0.124	0.01	0.1	132.57	17.56	1.02	
21	13	3	0.129	0.01	0.05	85.84	11.45	0.83	
22	13	3	0.135	0.01	0.02	54.94	7.47	0.75	
23	13	3	0.141	0.01	0.005	35.59	5.07	0.68	
24	13	3	0.144	0.01	0.002	26.44	4.37	0.68	
25	13	3	0.146	0.01	0.001	22.52	4.07	0.70	
26	15	4	0.185	0.01	0.01	417.08	27.38	1.00	
27	15	4	0.255	0.1	0.001	601.83	35.15	1.02	

on semilog coordinates in Fig. 8a. It is seen that all these statistics of the distributions have an essentially exponential dependence on filter length. The less regular behavior of the minimum values is believed to be caused by differences in bandwidth (F_1) among the filters.

Figure 8b shows a similar plot of the same distribution statistics for filters numbered 8 to 13 as a function of F_1. These filters have identical values of N, D_1, and D_2, and represent all six possible extraripple filters that have these parameter specifications. From Fig. 8b it is seen that with those parameters mentioned above held fixed, the noise output of a cascade filter tends to increase with increasing bandwidth.

Filters numbered 14 to 25 all have fixed values of N, N_p, and either D_1 or D_2. Plots of the distribution statistics of these filters as functions of D_1 and D_2 are shown respectively in Figs. 8c and 8d. These plots indicate that, as the transfer function approximation error for a filter decreases, so does its noise output. Though the plots are made

holding N_p rather than F_1 fixed, it is seen that, at least for the filters used in Fig. 8d, bandwidth increases with decreasing approximation error. Since the noise output of a filter is found to increase with band-

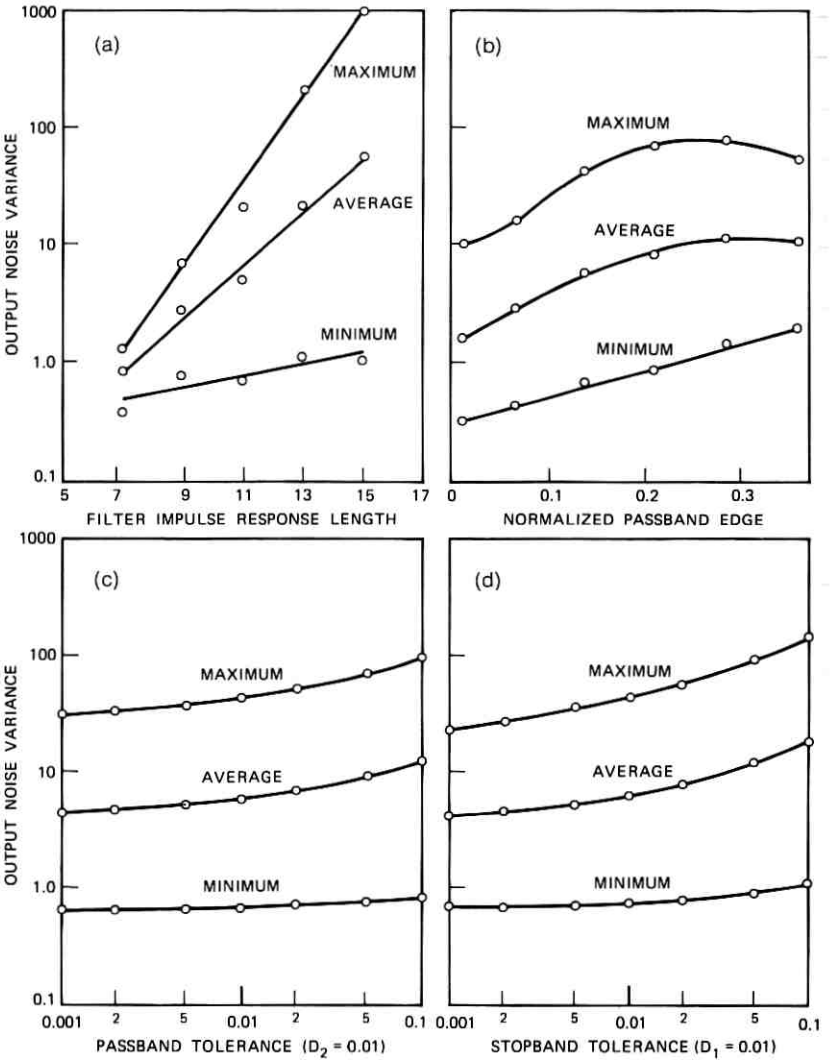


Fig. 8—(a) Output noise variance as a function of filter length. (b) Output noise variance as a function of bandwidth. (c) Output noise variance as a function of passband approximation error. (d) Output noise variance as a function of stopband approximation error.

width, it is expected that noise would still decrease with stopband tolerance D_2 if F_1 were fixed instead of N_p. In any event, the variation of F_1 among these filters is small.

Figures 8b to 8d are all plots of statistics for 13-point filters. Notice how the maximum, average, and minimum curves all tend to move together. In particular, the average curve almost always stays approximately halfway on the logarithmic scale between the maximum and minimum curves. This phenomenon is, of course, simply a manifestation of the empirical finding that noise distributions of different filters have essentially the same shape independent of differences in transfer characteristics.

To summarize, it has been found experimentally that with other parameters fixed, the roundoff noise output of a filter tends to increase with increases in all four independent parameters N, F_1, D_1, and D_2 which specify its transfer function. In particular, noise output tends to grow exponentially with N. It was not shown that the noise output of a filter with a fixed ordering and scaled a given way always varies in the way indicated when its transfer function parameters are perturbed. What has been shown is perhaps a more useful result from the design viewpoint. These findings imply that, other things being equal, a transfer function with, for instance, a higher value of D_2, is likely, when realized in a cascade form, to result in a higher noise output than a transfer function with a smaller value of D_2 realized by the same method. Though these results were found using only low-order filters, it is expected they could be generalized for higher-order filters as well. Section VI will present experimental evidence to confirm this expectation.

IV. COMPARISON OF SUM SCALING AND PEAK SCALING

The claim was made earlier that the results obtained on the noise distribution of filters ought to be quite independent of whether sum scaling or peak scaling is used. This claim will now be sustained heuristically and experimentally.

Let α denote the ratio of the maximum gain (over all frequencies) of a low-pass extraripple filter scaled by sum scaling to the maximum gain of the same filter peak scaled. Then it must be true that $\alpha \leqq 1$, since by definition the maximum gain for peak scaling is exactly one, while for sum scaling it must be no more than one if class 2 signals (which are a subset of class 1) are to be properly constrained (by Theorem 3, Ref. 2). Furthermore, it can be easily shown[1] that

$\alpha \geqq 1 - 2\epsilon$ where ϵ is defined as

$$\epsilon = \frac{\sum_k r(k)}{\sum_k |h(k)|} \tag{3}$$

with $\{h(k)\}$ being the impulse response of the filter and $\{r(k)\}$ being the magnitude of the negative portion of $\{h(k)\}$, i.e.,

$$r(k) = \begin{cases} -h(k) & h(k) < 0 \\ 0 & h(k) \geqq 0 \end{cases}. \tag{4}$$

For a low-pass filter, the envelope of $\{h(k)\}$ has the general shape of a truncated $(\sin x)/x$ curve, hence ϵ is expected to be a small number.

TABLE III—LIST OF FILTERS AND THE RESULTS OF ORDERING ALGORITHMS

$D1 = 0.01$ $D2 = 0.001$					Noise Variance		
					Peak Scaling		Sum Scaling
#	N	N_p	F_1	α	Alg 1	Alg 2	Alg 1
28	13	4	0.219	0.65	1.25	1.26	0.90
29	15	4	0.193	0.68	1.23	1.22	1.02
30	17	5	0.230	0.61	1.99	2.49	1.37
31	19	5	0.207	0.64	1.93	1.92	1.47
32	21	6	0.236	0.59	2.50	2.61	1.58
33	23	6	0.216	0.61	2.57	2.91	1.77
34	25	7	0.240	0.57	3.75	3.62	2.35
35	27	7	0.223	0.59	3.95	4.11	2.45
36	29	8	0.243	0.55	4.54	5.04	2.67
37	31	8	0.227	0.57	5.27	5.88	2.74
38	33	9	0.244	0.54	7.81	6.67	4.59
39	35	9	0.231	0.55	6.01	6.43	3.72
40	33	1	0.005	1.0	0.47	0.48	0.53
41	33	2	0.029	0.82	0.60	0.67	0.60
42	33	3	0.059	0.73	0.89	1.00	0.80
43	33	4	0.090	0.68	1.43	1.36	1.16
44	33	5	0.121	0.63	2.29	1.84	1.71
45	33	6	0.152	0.60	2.48	2.70	1.61
46	33	7	0.183	0.58	3.47	3.37	2.30
47	33	8	0.214	0.61	4.72	5.23	3.38
48	33	10	0.275	0.52	10.04	8.16	4.83
49	33	11	0.305	0.52	15.68	11.35	8.30
50	33	12	0.334	0.50	13.43	14.88	6.27
51	33	13	0.363	0.50	21.35	17.62	9.14
52	33	14	0.392	0.50	41.64	31.41	15.40
53	33	15	0.419	0.51	55.20	41.13	22.12
54	33	16	0.448	0.53	89.52	65.66	38.23

TABLE IV—LIST OF FILTERS AND THE RESULTS OF ORDERING ALGORITHMS

$N = 33$ $N_p = 8$					Noise Variance		
					Peak Scaling		Sum Scaling
#	F_1	D_1	D_2	α	Alg 1	Alg 2	Alg 1
55	0.211	0.01	0.002	0.59	5.63	5.33	3.34
56	0.208	0.01	0.005	0.57	5.13	5.69	3.18
57	0.205	0.01	0.01	0.56	5.05	5.27	3.34
58	0.202	0.01	0.02	0.55	7.63	8.31	4.01
59	0.197	0.01	0.05	0.53	11.34	12.53	6.92
60	0.193	0.01	0.1	0.51	46.33	22.99	16.88
61	0.238	0.1	0.01	0.58	9.90	9.01	5.61
62	0.227	0.05	0.01	0.58	8.91	7.35	5.52
63	0.214	0.02	0.01	0.56	8.87	5.75	4.32
64	0.196	0.005	0.01	0.56	5.47	4.69	3.68
65	0.185	0.002	0.01	0.56	5.95	4.08	3.41
66	0.178	0.001	0.01	0.57	4.10	4.11	2.85

In fact, it is easily shown[1] that, for a low-pass filter, if the passband tolerance D_1 is much less than the maximum passband gain, as is usually the case, then to an excellent approximation $\alpha = 1 - 2\epsilon$. It is now shown that if σ^2 is the output noise variance of a filter with sum scaling and σ'^2 is the output noise variance of the same filter except with peak scaling, then $\sigma^2 \geqq \alpha^2(\sigma'^2)$. To show this, the optimality properties for sum scaling and peak scaling, proved in Ref. 2, Theorem 4, are invoked. Since the gain of the sum-scaled filter is α times that of the peak-scaled filter, the ratio of their signal-to-noise ratios for class 2 inputs must equal α times the inverse ratio of their rms noise values (square root of variance), i.e., with S/N for sum scaling and S/N' for peak scaling,

$$\frac{S/N}{S/N'} = \alpha \cdot \frac{\sigma'}{\sigma}. \tag{5}$$

But since peak scaling is optimum for class 2 inputs and class 2 is a subset of class 1, then $S/N' \geqq S/N$. Thus $\sigma^2 \geqq \alpha^2\sigma'^2$. For an alternative derivation see Ref. 1.

In Tables III and IV, a list of filters and some results of Section VI are presented. Together with these results are listed measured values of α for each filter. Observe that, for these typical filters, α ranges from 0.5 to 1. Furthermore, for each filter, the last and third to last columns

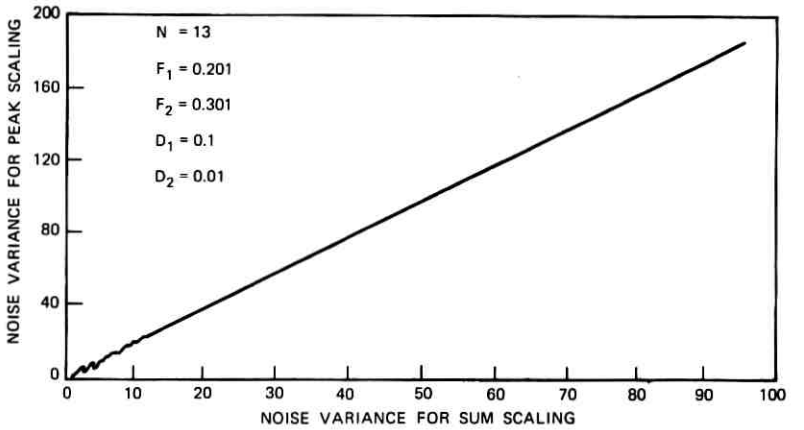


Fig. 9—Peak scaling versus sum scaling noise output comparison for a typical filter.

of Tables III and IV list the noise variances that result from the same ordering using sum scaling and peak scaling respectively. Comparing these, it is seen that, in almost every case, $\sigma^2 \leqq (\sigma'^2)$.[2] In particular, this is true if α is not too close to 1.0. The case where $\sigma^2 > (\sigma'^2)$ and $\alpha = 1$ is filter number 40 in Table III. However, except for the uninteresting cases of filters with all zeros on the unit circle, in general, $\alpha < 1$, and it is expected that $\sigma^2 \leqq (\sigma'^2)$. Thus for all practical purposes it can be assumed that

$$\alpha^2 \leqq \frac{\sigma^2}{\sigma'^2} \leqq 1. \tag{6}$$

From eq. (6), it is seen that the output noise variance for a filter with sum scaling is comparable, at least in order of magnitude, to that for the same filter ordered the same way with peak scaling applied. In fact, experimental results show that given a filter, the noise variances for sum scaling and peak scaling are in an approximately constant ratio for almost all orderings. An example of this result is shown in Fig. 9, where the noise variances for sum scaling and peak scaling of a typical filter are plotted against each other for each ordering. The resulting points are seen to form almost a straight line with slope approximately equal to 2, so that essentially $\sigma'^2 = 2\sigma^2$ for all orderings of this filter.

Thus the noise distributions of the previous section are essentially unchanged if peak scaling is used instead of sum scaling. To illustrate this, Fig. 10 shows the noise distribution for the filter of Fig. 6 with peak scaling used instead of sum scaling.

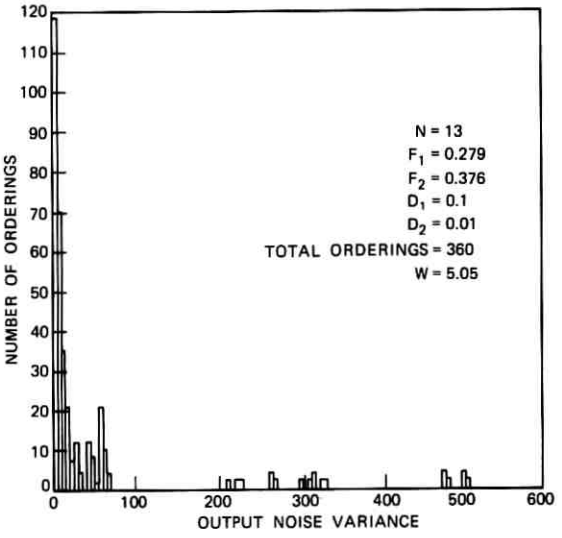


Fig. 10—Noise distribution histogram of filter of Fig. 5 using peak scaling.

The evaluation of noise variances with peak scaling is done in essentially the same way as that described in Fig. 4. Using a 128-point FFT to evaluate two transforms at a time (exploiting real and imaginary part symmetries) to give the maxima of the $F_i(e^{j\omega})$ for 360 orderings, the computations for peak scaling were found to require four times as much time as that for sum scaling.

V. AN INTUITIVE EXPLANATION OF ROUNDOFF NOISE DEPENDENCE ON ORDERING IN TERMS OF SPECTRAL PEAKING

That roundoff noise is distributed in the way shown with respect to orderings for a filter is an intriguing fact which is by no means obvious. The dependence of roundoff noise on ordering involves complicated matters like differing spectral shapes of different combinations of individual filter sections and the interactive scaling of signal levels within a filter necessitated by dynamic range limitations. As such, this dependence is much too complicated to visualize intuitively. It is proposed that the relative level of roundoff noise in a filter is adequately determined in order of magnitude by the amount of peaking in certain subfilter spectra. Thus it will be shown that, since the dependence on ordering of the amount of peaking of these spectra is not too difficult to visualize, by judging the relative amount of peaking of these spectra,

the relative merit of an ordering in terms of high-noise or low-noise output can be determined by inspection. These findings explain the general shape of noise distributions.

Given a linear phase filter with z-transform $H(z)$, define transfer functions $\bar{H}_i(z)$, $i = 1, \cdots, N_s$, to have the property that $\bar{H}_i(z)$ is proportional to the transfer function for the ith section of the filter, and each $|\bar{H}_i(e^{j\omega})|$ for all i has a maximum over ω equal to C, where C is chosen so that the overall filter frequency response $H(e^{j\omega})$ has a maximum in magnitude equal to one. Clearly $C \geqq 1$. Define transfer functions

$$\begin{aligned} \bar{A}_i(z) &= \prod_{j=1}^{i} \bar{H}_j(z) \\ \bar{B}_i(z) &= \prod_{j=i+1}^{N_s} \bar{H}_j(z) \end{aligned} \tag{7}$$

and define a number Pk to be the largest value of $\max_\omega |\bar{A}_i(e^{j\omega})|$ or $\max_\omega |\bar{B}_i(e^{j\omega})|$ for all i. It will be argued that, given an ordering, a large value of Pk indicates a high-noise output, while a low value of Pk indicates a low-noise output.

To see this, define $\bar{G}_i(e^{j\omega})$ to be $\bar{B}_i(e^{j\omega})$ with its maximum in magnitude over ω normalized to unity. Then it can be easily shown[1] that if

$$\begin{aligned} A_i &= \max_\omega |\bar{A}_i(e^{j\omega})| \\ B_i &= \max_\omega |\bar{B}_i(e^{j\omega})| \\ C_i &= k_i \frac{Q^2}{12} \cdot \frac{1}{2\pi} \int_0^{2\pi} |\bar{G}_i(e^{j\omega})|^2 d\omega \end{aligned} \tag{8}$$

(where k_i is the number of noise sources in the ith section), then the output noise variance due to the ith section is given by

$$\sigma_i^2 = A_i^2 B_i^2 C_i \qquad 1 \leqq i \leqq N_s - 1. \tag{9}$$

For the moment assume that C_i is a constant factor independent of ordering. Then σ_i^2 is proportional to $(A_i B_i)^2$. Note that for any i, A_i and B_i are the maxima of two functions the product of which is $H(e^{j\omega})$. Furthermore, for some i either $A_i = Pk$ or $B_i = Pk$. Now suppose $Pk \gg C$. Without loss of generality, it may be assumed $A_i = Pk$. Then argue that $A_i B_i \gg C$.

Clearly $A_i = |\bar{A}_i(e^{j\omega_0})|$ for some ω_0. Now $\bar{A}_i(z)\bar{B}_i(z) = H(z)$, and $H(z)$ is a function with zeros only in the z-plane other than the origin.

Also, at least in the case of well-designed band-select filters, the zeros of $H(z)$ are well spaced and spread out around the unit circle. The zeros of a typical filter are shown in Fig. 11a. Furthermore, $|H(e^{j\omega})| \leqq 1$. Thus in order for $|\bar{A}_i(e^{j\omega})|$ to have a large peak at ω_0, several zeros of $H(z)$ which occur in the vicinity of $z = e^{\pm j\omega_0}$ must be missing from $\bar{A}_i(z)$, while most of the remaining zeros must be in $\bar{A}_i(z)$. This means that $\bar{B}_i(z)$ has a concentration of zeros around $e^{\pm j\omega_0}$. By the result of Theorem 6(i) in Ref. 2, which says that the maximum of the magnitude frequency response of a filter section occurs at either $\omega = 0$ or $\omega = \pi$ depending on whether the zeros synthesized are in the left half or the right half of the z-plane, it is seen that most factors of $\bar{B}_i(e^{j\omega})$ must have maxima in magnitude which occur at exactly the same ω.

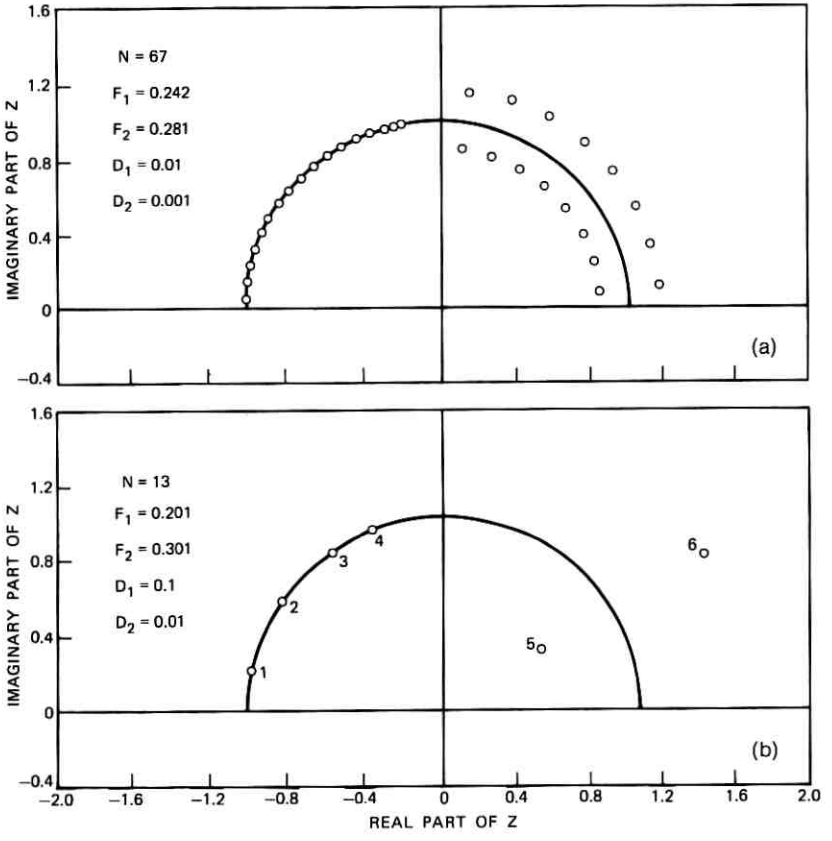


Fig. 11—(a) Zero positions of a typical 67-point filter. (b) Zero positions of filter number 14 of Table II.

Now C is found to be an increasing function of N_s^1, where typically $C > 2$. Hence $|\bar{B}_i(e^{j\omega})|$ is very likely to have a peak which is at least 1, or $B_i \geqq 1$. Thus $A_iB_i \gg C$. By the same token, if $B_i = Pk$ and $Pk \gg C$, then $A_iB_i \gg C$.

Hence if $Pk \gg C$, then for at least one i, $\sigma_i^2 = (A_iB_i)^2C_i$ where $A_iB_i \gg C$. Compared with a nominal value of say $A_iB_i = C$, the resulting difference in output noise variance can be great. When Pk takes on its lowest possible value, viz., $Pk = C$, the σ_i^2's are comparatively small for all i, hence it is expected that the resulting σ^2 is among the lowest values possible. Thus there exists a correlation between high values of Pk and high noise, and low values of Pk and low noise.

Concerning the assumption that C_i is constant independent of ordering, it is reasonable as long as only order-of-magnitude estimates are of interest. Since by definition $\max_\omega |\bar{G}_i(e^{j\omega})| = 1$ independent of ordering and i, it can be expected that variations in C_i with ordering are much less than variations in $(A_iB_i)^2$.

Based on these results, it can be concluded that an ordering which groups together, either at the beginning or at the end of a filter, several zeros all from either the left half or the right half of the z-plane is likely to yield very high noise. This observation is based on the fact that since zeros from the same half of the z-plane produce frequency spectra the maxima of which occur at exactly the same ω, several zeros from the same half of the z-plane can build up a large peak in the product of their spectra $\bar{H}_i(e^{j\omega})$. On the other hand, a scheme which orders sections so that the angle of the zeros synthesized by each section lies closest to the ω at which the maximum of the spectrum of the combination of the preceding sections occurs is likely to yield a low-noise filter.

The above observations are found to be true for all the filters the noise distributions of which were investigated. For example, from Table I, it is seen that those orderings which group together all three sections 4, 5, and 6 of the filter of Fig. 5 either at the beginning or at the end of the filter are precisely those which have the highest noise. Namely, they account for all noise variance values above 26.6 Q^2. Using the results on the noise distribution of a filter and the results of this section, it can thus be said that the comparatively few orderings of a filter which have unusually high noise can be avoided simply by judiciously choosing zeros for each section so that no large peaking in the spectrum, either as seen from the input to each section or from each section to the output, is allowed to occur. In particular, this can be done by ensuring that from the input to each section the zeros

synthesized well represent all values of ω, i.e., the variation in the density over ω of zeros chosen should be minimal.

These observations have the implication that low-noise orderings for a filter are those which choose zeros in such an order that they "jump around" about the unit circle and are well "interlaced," whereas high-noise orderings are those which allow clusters of adjacent zeros to be sequenced adjacently in the filter cascade. Since there are certainly far more ways to sequence zeros so that they satisfy the former property than the latter, it is clear why most orderings of a filter have low noise.

The values of Pk for all orderings of several filters were measured,[1] and they show good correlation with σ^2, thus supporting these arguments. For reasons of space, the results are not tabulated here. However, Figs. 12 and 13 show plots of the spectra from each section to the output of a high-noise and a low-noise ordering for a typical 13-point filter, the zeros of which are shown in Fig. 11b. Both orderings are peak scaled, so that the spectra from the input to each section of the filter have maxima equal to one. Thus, in reference to eq. (9), A_iB_i is equal to the maximum of the spectrum from section $(i + 1)$ to the output. The ordering of Fig. 12 has $\sigma^2 = 186\ Q^2$, while that of Fig. 13 has $\sigma^2 = 1.1\ Q^2$. It is seen that, as expected, A_iB_i has a large value for at least one i in the high-noise ordering, reaching a value of 60, while for the low-noise ordering $A_iB_i < 2.2$ for all i. Furthermore C_i, which is proportional to the integral of the square of the spectrum from section $(i + 1)$ to output with its maximum normalized to unity, does not vary too much between the two orderings. Finally, note that the spectrum of each section in the low-noise ordering does indeed tend to suppress the peak in the spectrum of the combination of previous sections. Thus the arguments of this section are supported.

VI. AN ALGORITHM FOR OBTAINING A LOW-NOISE ORDERING FOR THE CASCADE FORM

An extensive analysis of roundoff noise in cascade form FIR filters has been presented in this paper and in Ref. 2. However, an investigation of roundoff noise would not be complete without studying the practical question which in the first place had motivated all the analyses and experimentation. The question is, given an FIR transfer function desired to be realized in cascade form, how does one systematically choose an ordering for the filter sections so that roundoff noise can be kept to a minimum?

A partial answer to this question has already been given in the

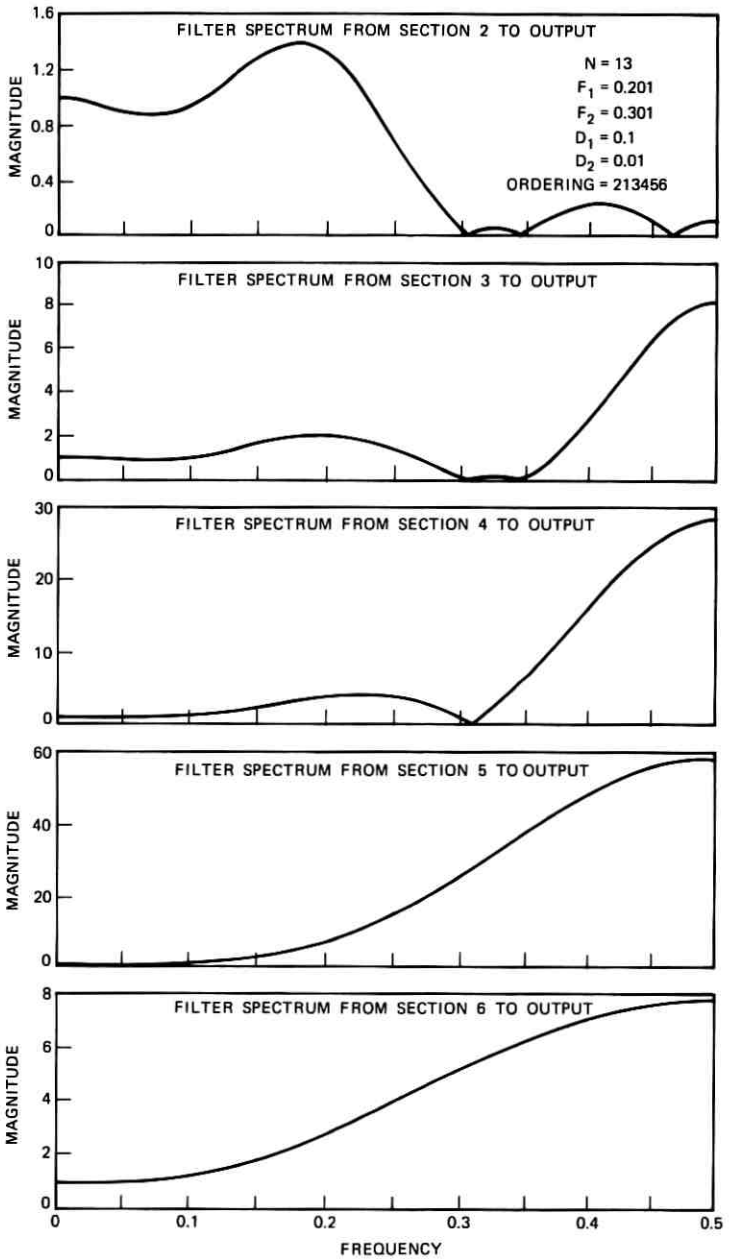


Fig. 12—Series of spectra from section i to the output where $i = 2, 3, 4, 5, 6$, for a high-noise ordering.

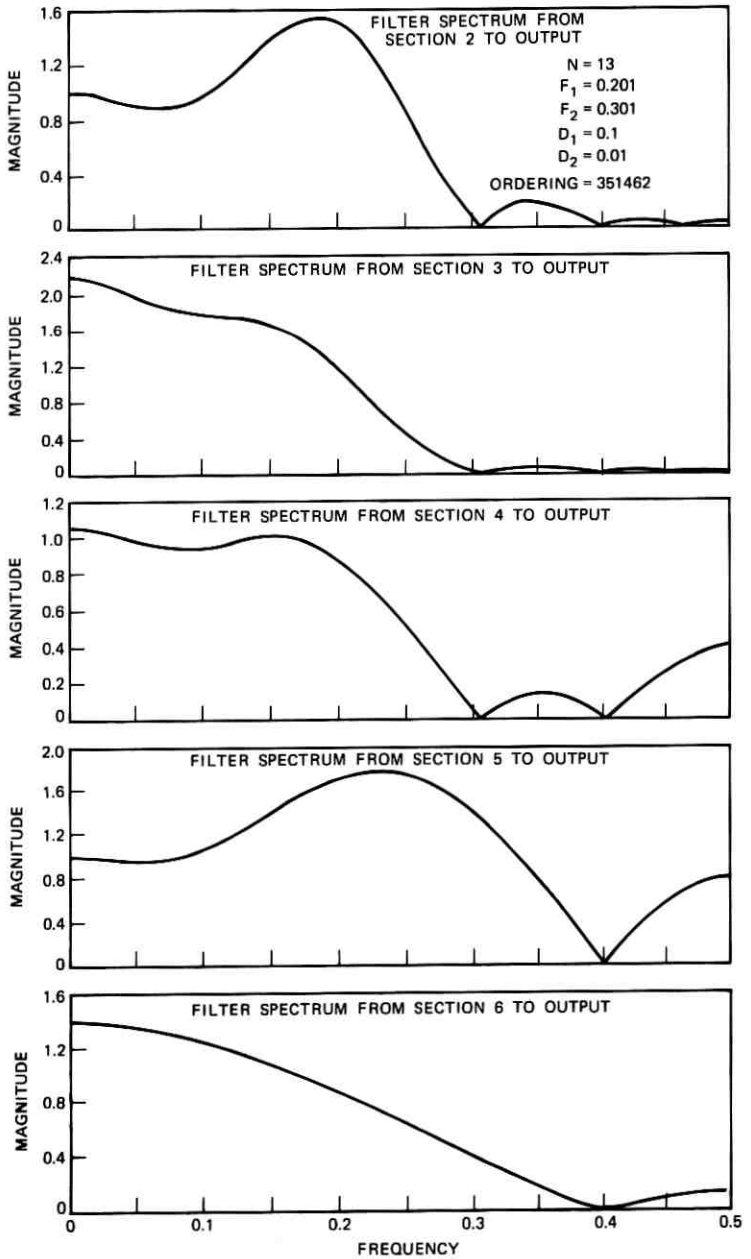


Fig. 13—Series of spectra from section i to the output where $i = 2, 3, 4, 5, 6$, for a low-noise ordering.

previous section. However, no completely systematic method has yet been devised for selecting an ordering for a filter guaranteed to have low noise. Ultimately, one wishes to find an algorithm which, when implemented on a computer, can automatically choose a proper ordering in a feasible length of time.

Avenhaus has studied an analogous problem for cascade IIR filters and has presented an algorithm for finding a "favorable" ordering of filter sections.[10] His algorithm consists of two major steps; a "preliminary determination" and a "final determination." In this section an algorithm is described for ordering FIR filters which is based upon the procedure used in the "preliminary determination" step of Avenhaus' algorithm. It has been found that a procedure appended to the proposed algorithm similar to Avenhaus' final determination step adds little that is really worth the extra computation time to the already very good solution obtainable by the first step. Hence such a procedure is not included in this algorithm.

No statement was made by Avenhaus as to what range of noise values can be expected of filters ordered by his algorithm, nor did he claim that his algorithm always yields a low-noise ordering (relatively speaking, of course). However, based on the results of Sections III through V, it will be argued heuristically that the proposed algorithm always yields filters which have output noise variances among the lowest possible. Together with extensive experimental confirmation, these arguments provide confidence that the proposed algorithm produces solutions that are very close to the optimum.

Application of Avenhaus' procedure to FIR filters allows the introduction of modifications which reduce significantly the amount of computation time required. Also, while IIR filters seldom require a higher order than the classic 22nd-order bandstop filter quoted by Avenhaus, practical FIR filters can easily require orders over 100. Though the same basic algorithm should still work for high orders, care must be exercised in performing details to avoid large roundoff errors in the computations. Through proper initialization, the proposed algorithm has been successfully tested for filters of order up to 128.

6.1 *Description and Discussion of Basic Algorithm*

The basic procedure or algorithm proposed by Avenhaus is simply the following. To order a filter of N_s sections, begin with $i = N_s$ and permanently build into position i in the cascade the filter section which, together with all the sections already built in, results in the smallest possible variance for the output noise component due to noise sources

in the ith section of the cascade. Because in the FIR cascade form noise is injected only into the output of each section, for FIR filters the procedure needs to be modified by considering the output noise due to the section in position $i - 1$ rather than i when choosing a section for position i. But the ith section is determined before the $(i - 1)$th section, hence the number of noise sources at the output of the $(i - 1)$th section is unknown at the time that a section for position i is to be chosen. This problem is overcome by assuming all sections to have the same number of noise sources. Then σ_i^2 is simply proportional to $\sum_k g_i^2(k)$ independent of what the ith section is, where $\{g_i(k)\}$ is the impulse response of the part of the filter from section $(i + 1)$ to the output.

Hence the revised basic algorithm for ordering FIR cascade filters is: *beginning with $i = N_s$, permanently build into position i the section which, together with the sections already built in, causes the smallest possible value for $\sum_k g_{i-1}^2(k)$.* Once this basic algorithm is determined, it is necessary only to decide on a scaling method and a computational algorithm for accomplishing the desired scaling and noise evaluation before an ordering algorithm is completed. Prior to discussing these issues, let us consider why the basic algorithm described above is always able to find a low-noise ordering.

The reason why the algorithm might not be able to find a low-noise ordering is that rather than minimizing $\sum \sigma_i^2$ directly, it minimizes each σ_i^2 individually where for σ_j^2, $1 \leqq j \leqq N_s - 1$, the search is essentially conducted over only $(j + 1)!$ out of the total of $N_s!$ possible orderings. Now this set of $(j + 1)!$ orderings depends on which sections were chosen for positions $j + 2$ to N_s in the cascade if $j < N_s - 1$. Hence in choosing a section for position j, previous choices might prevent attainment of a sufficiently small value for σ_{j-1}^2.

The basis for the following arguments is presented in Section V. Let $H(z)$ be an appropriately scaled filter. Given l, $1 \leqq l \leqq N_s - 1$, suppose σ_i^2 is small for all $i \geqq l$. Then the zeros of $\prod_{i=l+1}^{N_s} H_i(z)$ must be well spread around the unit circle in the z-plane since a clustering would cause large peaking in $\prod_{i=k+1}^{N_s} \bar{H}_i(e^{j\omega})$ for some $k \geqq l$, hence a large value of σ_k^2. But this means that the remaining zeros of $H(z)$, namely those in $\prod_{i=1}^{l} H_i(z)$, must also be well spread around the unit circle, since the zeros of $H(z)$ are distributed almost uniformly around the unit circle. Hence it ought certainly to be possible to find some pair of zeros in $\prod_{i=1}^{l} H_i(z)$ which, when assigned to position l, causes little peaking in $\prod_{i=1}^{l-1} \bar{H}_i(e^{j\omega})$ or $\prod_{i=l}^{N_s} \bar{H}_i(e^{j\omega})$, and thus results in a small value for σ_{l-1}^2. By induction, then, σ_i^2 can be chosen small for all i.

For small l it is true that there are very few zeros left as candidates for position l, but in these positions little peaking in the spectra can occur since the overall spectrum $\prod_{i=1}^{N_s} H_i(e^{j\omega})$ must be a well-behaved filter characteristic. Typically in a high-noise ordering, σ_l^2 reaches a peak for l somewhere in the middle between 1 and N_s, while σ_l^2 for small l has little contribution to σ^2, the total output noise variance. Hence the choice of sections for small l is not too crucial. Of course, the eligible candidates are still well-spaced zeros as for larger l, so that peaking should not be a problem.

Note that the reason the algorithm works so well is tied in with the result of Section III that most orderings of a filter have comparatively low noise. Because it is not difficult to find low-noise arrangements of zeros, $\sum \sigma_i^2$ can be minimized approximately by minimizing each σ_i^2 independently, searching over a much smaller domain. If the sum $\sum \sigma_i^2$ could not be segmented, searching for a minimum would be essentially an impossible task because of time limitations.

6.2 *Two Versions of the Complete Algorithm*

Having discussed why the basic algorithm works, the practical problem of implementing it is now discussed. First of all, there is the choice of scaling method to use in computing the $\sum_k g_i^2(k)$. As in the calculation of noise distributions in Section III, sum scaling is to be preferred since it can be carried out the fastest. Figure 14 shows a flow chart of the ordering algorithm in which sum scaling is employed. Calculation of σ^2 (Nx in the flow chart) is done exactly the same way as in the algorithm of Fig. 4.

Using this ordering algorithm, over 50 filters have been ordered and the noise variances in units of Q^2 (Q = quantization step size) of the resulting filters are shown in the last columns of Tables II through IV. Note that these noise variances are computed with sum scaling applied to the filters. The corresponding noise variance values for peak scaling have also been computed for the filters of Tables III and IV. These are shown in the third to the last column of the tables. The comparability of these noise values to those for sum scaling is a confirmation of the results of Section IV.

For an alternative implementation of the basic ordering algorithm, peak scaling can be used. To distinguish between the two different resulting algorithms, the former (sum scaling) will be referred to as alg. 1 and the latter as alg. 2. The only change needed in Fig. 14 to realize alg. 2 rather than alg. 1 is to replace $\sum_k |f_i(k)|$ by $\max_\omega |F_i(e^{j\omega})|$

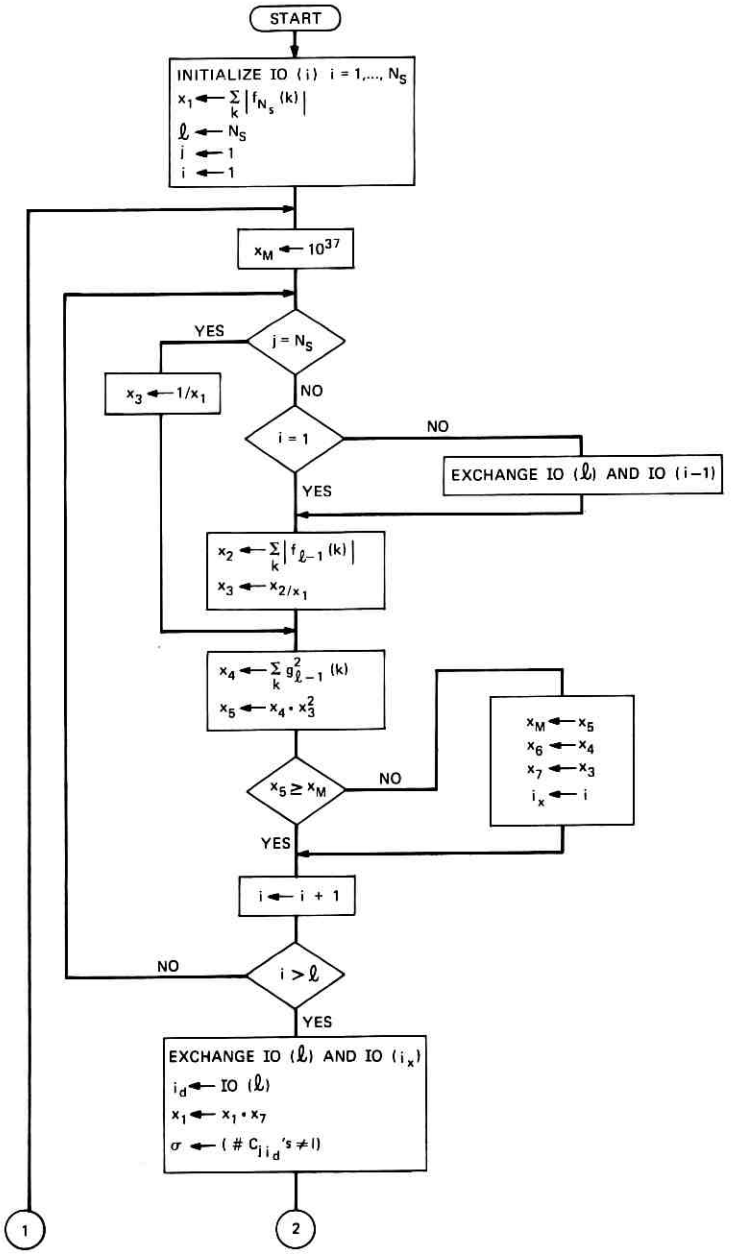


Fig. 14—Flow chart of ordering algorithm.

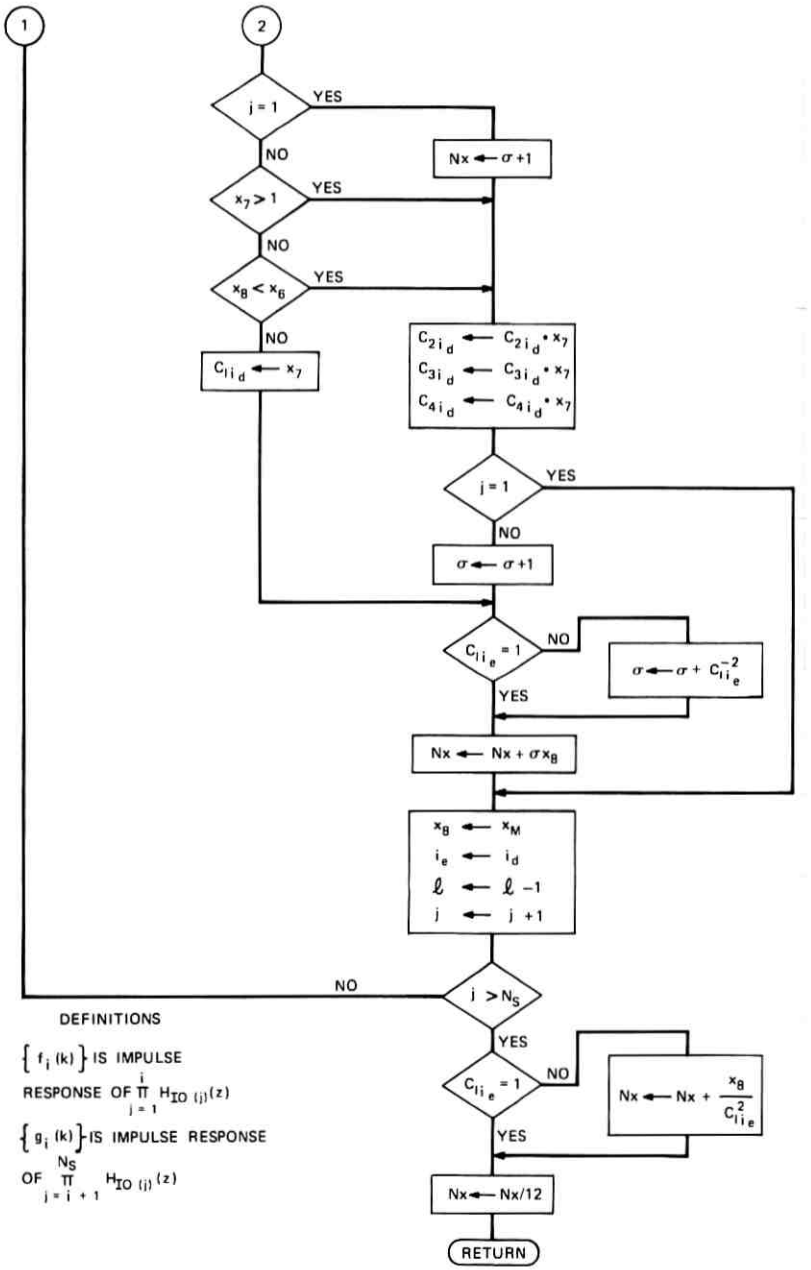


Fig. 14 (continued).

for given i whenever it appears. Results of using alg. 2 on the filters of Tables III and IV are shown in the next to last column of those tables. Observe that though the two algorithms in general yield different orderings for a given filter, the resulting noise variances are comparable. Thus, using both alg. 1 and alg. 2, two separate low-noise orderings for a given filter can be obtained. Further discussion of the results will be given shortly.

Even with a scaling method decided upon, the questions still remain of how $\sum_k g_i^2(k)$ and $\sum_k |f_i(k)|$ or $\max_\omega |F_i(e^{j\omega})|$ are to be computed and how the sequence $\{IO(i)\}$ which describes how the filter sections are ordered (see Fig. 14) is to be initialized. In obtaining the results of Tables III and IV, the following has been done: $\sum_k g_i^2(k)$ and $\sum_k |f_i(k)|$ were computed by evaluating $\{g_i(k)\}$ or $\{f_i(k)\}$ through simulation in the time domain (i.e., convolution) and $\max_\omega |F_i(e^{j\omega})|$ was determined by transforming $\{f_i(k)\}$ via an FFT and then finding the maximum value. Finally, $\{IO(i)\}$ was initialized to $IO(i) = i$, $i = 1, \cdots, N_s$. It will be seen that these procedures must be modified for higher-order filters. But meanwhile, the implications of these procedures in terms of dependence of computation time on filter length is considered.

Clearly, in algorithmically computing the impulse response of an N-point filter via convolution, the number of multiplies and adds

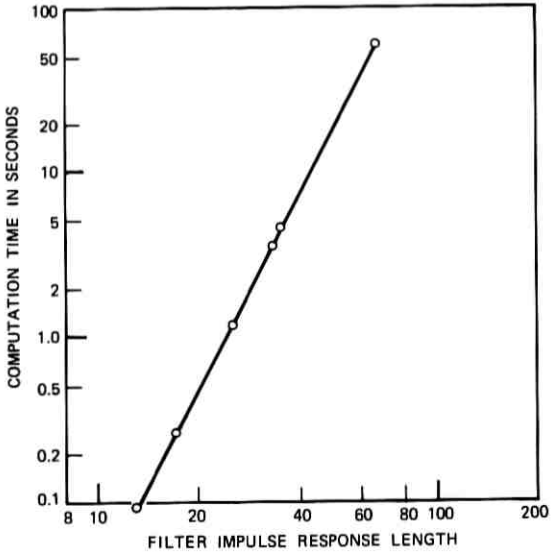


Fig. 15—Computation time versus filter length for ordering algorithm.

required to calculate each point varies as N, hence the time required to evaluate the entire impulse response must vary approximately as N^2. Now in the basic algorithm there are two nested loops, where the number of times the operations within the inner loop are performed is given by

$$\sum_{i=1}^{N_s} i = \frac{N_s(N_s + 1)}{2} \cong \frac{N^2}{8}.$$

Clearly for alg. 1, the evaluation of $\sum_k |f_{l-1}(k)|$ and $\sum_k g^2_{l-1}(k)$ dominates all operations within the inner loop in terms of time required. Since the total number of points that must be evaluated in order to compute $\{f_{l-1}(k)\}$ and $\{g_{l-1}(k)\}$ together turns out to be a constant independent of l, the combined operations must have approximately an N^2 time dependence. Hence it is predicted that the computation time required for alg. 1 must be approximately proportional to N^4. This prediction is verified in Fig. 15, where computation time for alg. 1 on the Honeywell 6070 computer is plotted against N on log-log coordinates for various values of N. As expected, these points lie on a straight line with a slope very nearly equal to 4.

For alg. 2, exactly the same procedures as in alg. 1 are carried out except that after each evaluation of $\{f_i(k)\}$ an FFT is performed. Thus for a given N, alg. 2 always requires more time than alg. 1, with the exact difference depending on the number of points employed in the FFT.

6.3 *Modification of Algorithm for Higher Filter Orders*

For filters of length greater than approximately 41, it is found that accuracy in the evaluation of impulse response samples by the methods described rapidly breaks down. This phenomenon is chiefly due to the fact that the initial ordering used is a very bad one. In particular, it is not difficult to see that this ordering (i.e., $\mathrm{IO}(i) = i$) has a noise variance which is among the highest possible and which increases at least exponentially with N. Thus all attempts at evaluating the impulse response of the filter by simulation in the time domain are marred by roundoff noise.

A natural possibility for resolving this problem is to perform calculations in the frequency domain. This has been tried as a modification to alg. 2. In particular, rather than computing $F_i(e^{j\omega})$ from

$\{f_i(k)\}$, it is evaluated as a product of $H_l(e^{j\omega})$, $l = 1, \cdots, i$, where $H_l(e^{j\omega})$ is computed from the coefficients of section l via an FFT. In this way the accuracy problem was solved, but computation time increased significantly. As an example, the 67-point filter listed in Table V was ordered using this method. The resulting noise variance was a reasonable 26.6 Q^2, but even with a 256-point FFT the computation time required amounted to 7.2 minutes, more than 7 times that required for alg. 1 to order the same filter.

A far better solution is as follows. The conclusion of Section III–that most orderings of a filter have relatively low noise–means that, if an ordering were chosen at random, it ought to have relatively low noise. The strategy is then to use a random ordering as an initial ordering for alg. 1. A given ordering of a sequence of numbers $\{\mathrm{IO}(i), i = 1, \cdots, N_s\}$ can be easily randomized using the following shuffling algorithm:[11]

Step 1: Set $j \leftarrow N_s$.
Step 2: Generate a random number U, uniformly distributed between zero and one.
Step 3: Set $k \leftarrow [jU] + 1$. (Now k is a random integer between 1 and j.) Exchange $\mathrm{IO}(k) \leftrightarrow \mathrm{IO}(j)$.
Step 4: Decrease j by 1. If $j > 1$, return to Step 2.

By adding a step to randomize the initial ordering $\mathrm{IO}(i) = i$ in alg. 1, the inaccuracy problem was eliminated. The interesting question now arises that, since most orderings of a filter have relatively low noise, can we not obtain a good ordering simply by choosing one at random? The answer is yes in a relative sense, but, as will shortly be seen, a random ordering is far from being as good as one which can be obtained using the ordering algorithm.

The extra step of randomizing the initial ordering for alg. 1 requires negligible additional computation time, and a filter with impulse response length as high as 129 has been successfully ordered in this way. The time required to order this filter was approximately 13.5 minutes. Except for time limitations, there is no reason why even higher-order filters cannot be similarly ordered. Further results are described in the next section.

6.4 *Discussion of Results*

Note in Table II that alg. 1 can result in a noise variance which is very close to the minimum, if not the minimum. From this observation and the conclusions of Section III on the dependence of the minimum

TABLE V—LIST OF FILTERS AND THE RESULTS OF ALG 1′

$D_1 = 0.01$					Noise Variance			
					Ordering		Alg 1′	
#	N	N_p	F_1	D_2	Sequential	Random	Sum sc	Peak sc
38	33	9	0.244	0.001	1.0×10^{11}	6.2×10^{3}	4.59	7.81
67	47	12	0.237	0.001	4.3×10^{17}	2.2×10^{6}	6.47	12.07
68	67	17	0.242	0.001	3.3×10^{27}	1.5×10^{6}	16.77	30.03
69	101	25	0.241	0.001	$> 10^{38}$	1.4×10^{5}	41.93	73.55
70	129	20	0.153	0.0001	—	5.5×10^{11}	17.98	37.54

noise variance for a filter on different parameters, we can be quite confident that the noise variances shown in Tables III and IV are also very close to the minimum possible. The filters of Tables III and IV were chosen intentionally to show the dependence of the results of the ordering algorithms on various transfer function parameters. It is seen that the noise variances indeed behave in the way that would be expected from the results of Section III. In particular, σ^2 is seen to be essentially an increasing function of N, F_1, D_1, as well as D_2. The results of Tables III and IV are then a confirmation of the expectation that the conclusions of Section III on the general dependence of noise on transfer function parameters can be generalized to higher-order filters.

The results of using the modified alg. 1 (denoted alg. 1′) on a few filters are shown in Table V. Also shown in this table, for comparison, are the noise variances of these filters when they are in the sequential ordering $\mathrm{IO}(i) = i$ (where computable within the numerical range of the computer) as well as when they are in a random ordering (obtained by randomizing $\{\mathrm{IO}(i)\}$ where $\mathrm{IO}(i) = i$, as described above). Because of the potentially very large roundoff noise encounterable in these orderings, the noise variances were computed using frequency domain techniques. In particular, each $H_i(e^{j\omega})$ is evaluated via an FFT, peak scaling is then performed, and finally σ_i^2 is computed via $1/(2\pi) \int_0^{2\pi} |G_i(e^{j\omega})|^2 d\omega$ rather than $\sum_k g_i^2(k)$.

From Table V it is seen that though the noise variances of the random orderings are certainly a great deal lower than those of the corresponding sequential orderings, they are far from being as low as those obtained by alg. 1′. Thus it is certainly advantageous to use alg. 1

(or 1′) to find proper orderings for filters in cascade form. In all the examples given in Tables II through V, one can do little better in trying to find orderings with lower noise. With the possible exception of the uninteresting wideband filter, number 54 in Table III, all filters have less than 4 bits of noise after ordering by alg. 1, while the great majority have less than 3 bits. Thus it is not expected that these noise figures can be further reduced by much more than a bit or so.

Finally, in practice, cascade FIR filters of orders over approximately 50 are generally of little interest since there exist more efficient ways than the cascade form to implement filters of higher orders. For filters of, at most, 50th order, the computation time required for alg. 1 is less than 20 seconds on the Honeywell 6070 computer. Thus alg. 1 (or 1′) is a very efficient means for ordering cascade filters.

VII. SUMMARY

In this paper, experimental results have been presented which show that most orderings for an FIR filter in cascade form have very low noise relatively, and that the shape of the distribution of noise with respect to ordering is essentially independent of transfer function parameters as well as method of scaling (sum or peak). An explanation of these properties has been proposed, based on a characterization of high-noise and low-noise orderings. Furthermore, the dependence of noise on transfer function parameters and scaling has been investigated. These results point to an algorithm for ordering cascade FIR filters which has been implemented and tested for filters with a wide range of values of transfer function parameters. In every case, the algorithm gave results within expectation which were deduced to be very close to the optimum. Justification for the success of the algorithm has also been given.

REFERENCES

1. Chan, D. S. K., "Roundoff Noise in Cascade Realization of Finite Impulse Response Digital Filters," S. B. and S. M. Thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, September, 1972.
2. Chan, D. S. K., and Rabiner, L. R., "Theory of Roundoff Noise in Cascade Realizations of Finite Impulse Response Digital Filters," B.S.T.J., this issue, pp. 329–345.
3. Parks, T. W., and McClellan, J. H., "Chebyshev Approximation for Non-Recursive Digital Filters with Linear Phase," IEEE Trans. Circuit Theory, *CT-19* (March 1972), pp. 189–194.
4. Jackson, L. B., Kaiser, J. F., and McDonald, H. S., "An Approach to the Implementation of Digital Filters," IEEE Trans. Audio Electroacoustics, *AU-16*, No 3 (September 1968), pp. 413–421.

5. Schafer, R. W., "A Survey of Digital Speech Processing Techniques," IEEE Trans. Audio Electroacoustics, *AU-20*, No. 4 (March 1972), pp. 28–35.
6. Schafer, R. W., and Rabiner, L. R., "Design and Simulation of a Speech Analysis Synthesis System Based on Short-Time Fourier Analysis," IEEE Trans. Audio Electroacoustics, June 1973, in preparation.
7. Schüssler, W., private communication.
8. Schüssler, W., "On Structures for Nonrecursive Digital Filters," Archiv Für Electronik Und Übertragungstechnik, *26*, 1972.
9. Jackson, L. B., "On the Interaction of Roundoff Noise and Dynamic Range in Digital Filters," B.S.T.J., *49*, No. 2 (February 1970), pp. 159–184.
10. Avenhaus, E., "Realizations of Digital Filters with a Good Signal-to-Noise Ratio," Nachrichtentechnische Zeitscrift, May 1970.
11. Knuth, D. E., *Seminumerical Algorithms*, Vol. 2 of *The Art of Computer Programming*, Reading, Massachusetts: Addison-Wesley, 1969, p. 125.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 3, March, 1973
Printed in U.S.A.

Slope Overload Noise in Linear Delta Modulators With Gaussian Inputs

By L. J. GREENSTEIN

(Manuscript received September 22, 1972)

This paper derives a slope overload noise power formula for linear delta modulators having ideal integrators and Gaussian random inputs. Although the same problem has been treated by others, the present result is the only one applicable to all slope-following capacities and input spectra.

Despite its singleness of purpose, the paper divides logically into two parts. In Part 1, a common element in all previously published results is used to derive a new slope overload noise power formula. This derivation is analytically rigorous and provides some useful insights, but pertains to a particular kind of spectrum and so is incomplete.

The more universal result we seek is derived in Part 2. The approach here is far less rigorous and amounts to approximating the influences of other kinds of spectra by modifying the result of Part 1. The final expression contains four spectrum-related coefficients, for which simple formulas are given, and has an estimated accuracy of 1 dB for all cases of practical interest. Computed results are given for two important families of spectra and comparisons are made with previously published results.

Part 1

I. INTRODUCTION

1.1 *Objective*

This paper presents some new theoretical results on slope overload noise in linear delta modulators. In particular, we derive the mean power of the (unfiltered) slope overload noise at the demodulator output when the modulator input is a stationary Gaussian random process and the modulator feedback path contains an ideal integrator, Fig. 1. This same problem has already been treated by other investigators, notably Zetterberg,[1] Rice (with O'Neal),[2] Protonotarios,[3] and Abate,[4] but several factors have prompted a reexamination. One is

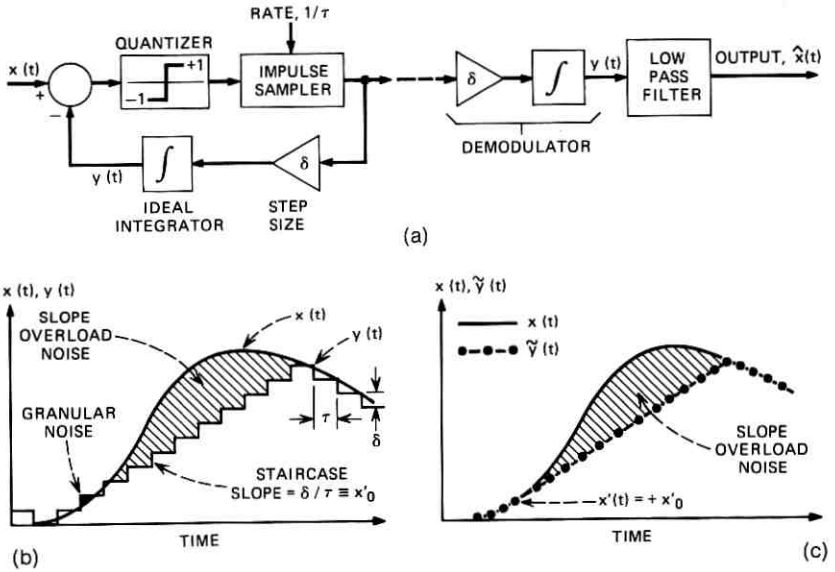


Fig. 1—Linear ΔM codec with perfect feedback integrator. (a) Equivalent block diagram. (b) Input and feedback signals. (c) Model used.

that some important disparities among the results of these separate studies remain unresolved; another is that the existing formulas do not, either individually or collectively, pertain to all slope-following capacities and input spectra; and finally, few explicit clues are available as to just how accurate each formula is and where it loses validity. In response to this situation, we have endeavored to find an expression for slope overload noise power that is accurate for all slope-following capacities and input spectra of possible interest. The result reported here satisfies that objective.

1.2 *Noise Descriptions and Definitions*

The idealized ΔM codec (coder/decoder) shown in Fig. 1 exemplifies the process we want to analyze: Every τ seconds the input signal ($x(t)$) is compared with a locally quantized version of itself ($y(t)$), and a unit impulse is generated with a polarity that is $\begin{pmatrix}\text{positive}\\ \text{negative}\end{pmatrix}$ if $\begin{pmatrix}x > y\\ x < y\end{pmatrix}$. The resulting binary impulse stream is applied through a feedback gain factor (δ) and ideal integrator to produce $y(t)$. At the decoder, $y(t)$ is reconstructed by using a replica

of the coder feedback network, and a final low-pass filter smooths the sharp edges of $y(t)$, yielding a closer approximation to $x(t)$.

Since $y(t)$ cannot change by more than δ units in τ seconds, δ/τ is the highest input signal rate-of-change that the ΔM codec can follow. We call δ/τ the ΔM *slope-following capacity* and denote it by x_o'. When $|dx/dt|$ exceeds this quantity, *slope overload* occurs and gives rise to the kind of error shown in Fig. 1b. In addition to this sporadic form of distortion, the "hunting" of $y(t)$ for $x(t)$ by means of quantum steps gives rise to a perpetual distortion called *granular noise*. Obviously, granular noise is reduced by decreasing δ, but at the expense of a reduced slope-following capacity and, hence, greater slope overload noise.

To delineate slope overload and granular noise for purposes of this analysis, let us suppose that $x(t)$ is passed first through a ΔM codec having infinitesimal δ and τ, but with the ratio between them the same as in the actual system (Fig. 1a). The decoded output (before the final filter) will then be the smooth function $\tilde{y}(t)$ shown in Fig. 1c. If $\tilde{y}(t)$ is then passed through the actual system, the decoded signal will be a very close approximation to $y(t)$, Fig. 1b. In agreement with previous conventions, we define slope overload noise to be the difference between $x(t)$ and $\tilde{y}(t)$; and the remaining distortion, $(\tilde{y}(t) - y(t))$, is defined to be granular noise. This essentially is the approach tacitly followed in most of the published literature on ΔM noise.[1-7]

We define a slope overload noise burst to be the nonzero difference between $x(t)$ and $\tilde{y}(t)$ over an interval $[t_a,t_b]$, which starts because $|dx/dt| > x_o'$ at $t = t_a$ and ends because $\tilde{y}(t)$ intersects $x(t)$ at $t = t_b$, (e.g., see Fig. 2, in which $t_a = 0$). The mean power of these bursts, averaged over all time, is the *slope overload noise power*.

Two qualitatively different kinds of noise bursts can be identified. One is the kind initiated when $|dx/dt|$ increases through x_o' while $\tilde{y}(t)$ is following $x(t)$, which we call *primary noise* (see Figs. 1c and 4b). The other arises when a prior burst terminates at a point where $|dx/dt|$ already exceeds x_o' (see Fig. 4c); the new burst that commences at this point we call *secondary noise*.

Finally, the *slope overload factor* is defined to be the ratio of the slope-following capacity to the rms input signal derivative, i.e.,

$$S \triangleq \frac{x_o'}{(dx/dt)_{\text{rms}}}. \tag{1}$$

It should be obvious that the noise power decreases monotonically

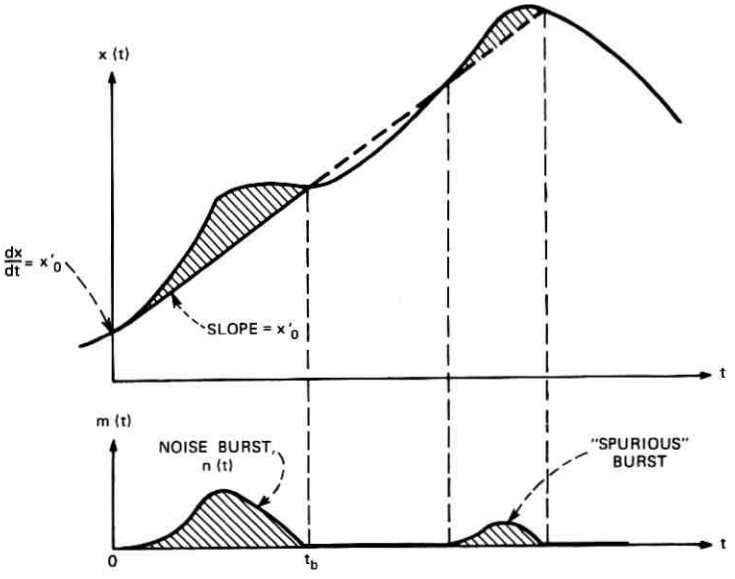


Fig. 2—Zetterberg's noise model.

with increasing S. It is also clear that when S is large (in which case $|dx/dt|$ rarely exceeds x_o'), secondary bursts are rare so that primary noise dominates the noise process; and that, by the same token, secondary noise dominates the noise process when S is small.

1.3 *Equivalent Input Process*

The generality of our analysis will be enhanced if we can assume that $\{x(t)\}$ is a bandlimited process. This assumption is made valid by regarding $\{x(t)\}$ as an equivalent process related to the true one as follows: Let the true input process be $\{x_o(t)\}$, having a power spectrum $X_o(f)$. Since the delta modulator really acts on discrete samples of the input separated by τ seconds, its response is the same as if the input were a process $\{x(t)\}$ having sample functions of the form

$$x(t) = \sum_{k=-\infty}^{\infty} x_o(k\tau) \frac{\sin(\pi(t - k\tau)/\tau)}{(\pi(t - k\tau)/\tau)} \tag{2}$$

and a power spectrum

$$X(f) = \sum_{n=-\infty}^{\infty} X_o\left(f + \frac{n}{\tau}\right); \qquad 0 \leqq f \leqq \frac{1}{2\tau} \text{ only.} \tag{3}$$

We will assume here that the sample function and slope overload noise depicted in Fig. 1c correspond to the bandlimited process $\{x(t)\}$, i.e., that the input spectrum is $X(f)$. If $\{x_o(t)\}$ is bandlimited by some frequency $W \leqq 1/(2\tau)$, the two processes, and consequently their spectra, are one and the same. (This condition is usually tacitly–though not always rightly–assumed in ΔM noise analyses.) If $\{x_o(t)\}$ is not so limited, the difference between the two processes is aliasing distortion, which can be analyzed separately.

1.4 *The Spectral Moments*

The spectral moments of the input process play a decisive role in establishing the past and present slope overload noise formulas. Following the convention of most authors, we define the nth moment to be

$$b_n = \int_0^\infty \omega^{2n} X(f) df. \tag{4}$$

The bandlimited nature of $X(f)$, as discussed above, guarantees the finiteness of all the moments. It is easy to show that the complete set of these finite moments, $(b_0, b_1, b_3, \cdots, b_n, \cdots)$, uniquely determines $X(f)$, just as the converse is true.

It should be obvious that b_n is the mean power of the nth derivative of $x(t)$, i.e., $b_n = \overline{(x^{(n)}(t))^2}$. Thus, for example, $\sqrt{b_1}$ is the rms derivative of the input and so (1) can be rewritten as

$$S = x_o'/\sqrt{b_1}. \tag{5}$$

Also, b_0 is nothing other than the input signal power, as distinct from the *ac* input power which we denote by σ^2. If the input contains no discrete dc component, then $b_0 = \sigma^2$; otherwise, b_0 exceeds σ^2 by the amount of the dc power.

Finally, we note from (4) that b_n is real and non-negative because $X(f)$ is. An additional relationship for b_n, derived by applying the Schwarz inequality to (4), is

$$b_n \geqq \frac{b_{n-1}^2}{b_{n-2}}; \qquad n \geqq 2, \tag{6}$$

that is, b_n for $n \geqq 2$ has a lower bound determined by the previous two moment values. We shall utilize this relationship later.

1.5 *Outline of the Work*

Our objective is to find a suitable approximation to the true slope overload noise power formula, the latter being denoted by $N(S)$. Section II describes the methods and results of previously published analyses. While the various approaches differ, they all suggest that this noise power is essentially determined by S and the first three spectral moments, b_0, b_1, and b_2. Section III treats this possibility as a premise and identifies a simple two-band process, the parameters of which can be chosen to yield any (b_0, b_1, b_2) and which can be analyzed precisely. The slope overload noise power for this process is derived and is denoted by $N_{tb}(S)$.

Unfortunately, $N_{tb}(S)$ is *not* precisely applicable to all spectra having the same b_0, b_1, and b_2. Instead, it is a lower-bound variation for all such spectra, which becomes increasingly unreliable in general as S decreases. To obtain a more universal estimate [denoted by $N_o(S)$] it is necessary to determine how much the slope overload noise power deviates from $N_{tb}(S)$ for spectra other than the two-band kind.

To this end, Section IV derives a noise power formula, $N_l(S)$, applicable to all spectra but confined to the large-S region ($S \geqq 3.5$), and Section V derives a noise power formula, $N_s(S)$, applicable to all spectra but confined to the small-S region ($S \approx 0$). Section VI then combines these results to obtain a two-region approximation, $N_o(S)$, which is accurate for all $X(f)$ and S. The result is given by (62) and (77), where α_1, α_2, a_1, a_2, a_3, and a_4 are related to the spectrum parameters b_0, b_1, b_2, b_3, b_4, b_5, $X(0)$, and $\int_o^\infty \omega^{-2}[X(f) - X(0)]df$. Further study shows that (77) alone can be used over all S with an accuracy of 1 dB or better up to $S = 6.5$, beyond which point $N(S)$ is at least 119 dB below the input signal power. For practical purposes, therefore, our final approximation to $N(S)$ is just (77), with a_1, a_2, a_3, and a_4 given by (72) through (75).

Section VII demonstrates the new result for two important families of spectra, and compares $N_o(S)$ with the noise power formula of Protonotarios. The latter is found to be highly accurate for most spectra of practical interest and $S \geqq 2.0$.

II. REVIEW OF PREVIOUS WORK

In the approaches of Zetterberg, Rice, and Protonotarios, $N(S)$ is approximated as the product of the mean energy ($\mathcal{E}$) per slope overload burst and the average rate of occurrence (R) of such bursts. A burst is assumed to commence whenever $|x'(t)|$ increases through

the value x'_o, which means that only primary noise is considered in these studies. The average rate of these events is known from the earlier work of Rice[8] to be

$$R = \frac{1}{\pi}\sqrt{\frac{b_2}{b_1}} \exp\{-(x'_o)^2/2b_1\}. \tag{7}$$

To find $\mathcal{E}$, Zetterberg uses the model shown in Fig. 2, where the solid line of slope x'_o represents the demodulator output for the duration of the burst, i.e., from $t = 0$ to $t = t_b$. Clearly, the noise is

$$n(t) = x(t) - [x(0) + x'_o t]; \qquad 0 \leqq t \leqq t_b \text{ only} \tag{8}$$

which is the first burst depicted in Fig. 2. To avoid deriving the random quantity t_b, Zetterberg regards the slope overload burst to be the entire excursion of $x(t)$ above the line $[x(0) + x'_o t]$, which is the variation $m(t)$ in Fig. 2 rather than just $n(t)$. This is obviously an approximation, since $m(t)$ can contain one or more spurious bursts that are not really part of the original noise burst, as seen. Such bursts, however, are low in both energy and probability of occurrence when $S \gg 1$.

Zetterberg proceeds by finding, at each t, the average of m^2 over the noise burst ensemble, and then integrating over all time and multiplying by R, (7). If done correctly, this leads to an estimate of $N(S)$ which is marred only by the inclusion of spurious burst contributions and the ignoring of secondary noise effects. In fact, however, Zetterberg errs in defining the initial conditions of the burst and in averaging over these conditions, leading to an incorrect solution.* In addition, Zetterberg makes a number of functional approximations and at least one algebraic error.* For all these reasons, his final result is incorrect and will not be repeated here.

The approach of Rice reverses that of Zetterberg, in that the energy per burst is found first and then averaged over the noise burst ensemble. The problem of spurious bursts is thus avoided, but at the expense of having to find t_b. Rice accomplishes this by expanding $x(t)$ into a power series about $t = 0$, and then ignoring fourth-order powers in t and higher, an approximation that gains in validity with increasing S. In addition, he approximates the third derivative of $x(t)$ at $t = 0$ by its conditional mean given that $x'(0) = x'_o$. This can be shown to be $-(b_2/b_0)x'_o$, and the relative variation of $x'''(0)$ about this value tends to be small when $S \gg 1$.

* See Protonotarios[3] for a discussion of these errors.

Using these two approximations, and correctly identifying and averaging over the initial conditions of the burst, Rice obtains the result

$$N_R(S) = \frac{243}{4\sqrt{2\pi}}\left(\frac{b_1^2}{b_2}\right)\frac{1}{S^5}\exp(-S^2/2). \tag{9}$$

Note that the ratio of noise power to signal power (b_0) depends solely on S and $\gamma \triangleq b_1^2/b_0b_2$. From (6) we see that $\gamma \leqq 1$ and observe that $\gamma = 1$ only for an infinitely narrowband spectrum.

It should also be noted that, as $S \to 0$, $N_R(S)$ increases without bound at a rate S^{-5}. This is clearly not a true representation since $N(S)$ should approach the ac signal power (σ^2) in the limit of zero slope-following capacity.*

Noting the disparity between Zetterberg's result and Rice's, even at large S where both should converge to exactness, Protonotarios has attempted to settle the issue by combining Zetterberg's more accurate model with Rice's correct averaging procedure. His analysis leads to a double integral solution which is exact except for the inclusion of spurious contributions and the ignoring of secondary noise effects. In reducing this formal result, Protonotarios makes some functional approximations and obtains the following:

$$N_P(S) = \frac{243}{4\sqrt{2\pi}}\left(\frac{b_1^2}{b_2}\right)\frac{1}{S^5}\exp(-S^2/2)A(\chi) \tag{10}$$

where

$$\chi = \frac{2^{5/8}}{3}S/(b_1^2/b_2b_0)^{1/8} \tag{11}$$

and $A(\cdot)$ is a function involving powers and exponentials of the argument, [eq. (66) of Ref. 3]. Once more, the ratio of noise power to signal power depends solely on S and γ. In the limit as $S \to \infty$, $A(\chi)$ approaches 1 and so $N_P(S)$ converges to Rice's result, $N_R(S)$. In the limit as $S \to 0$, $A(\chi)$ varies as S^4 so that $N_P(S)$ varies as S^{-1} rather than S^{-5}. This is still an unbounded increase, however, so that (10) is still not acceptable as a complete characteristic.

Abate's derivation of $N(S)$ follows an approach quite different from the others. Using simulation results reported by O'Neal[2] for three particular spectra, he has developed an empirical relationship between the ΔM sampling frequency and the slope overload factor for which

* If the input contains a discrete dc component, the delta modulator will follow it exactly so long as $S \neq 0$; hence the result $\lim_{S\to 0} N(S) = \sigma^2$.

granular-plus-slope overload noise power is minimized. Combining that relationship with a simplifying approximation to van de Weg's formula for granular noise power,[5] Abate obtains the following expression for slope overload noise power:

$$N_A(S) = \frac{8\pi^2}{27} \frac{b_1}{(2\pi W)^2} (1 + 3S) \exp(-3S) \tag{12}$$

where W is the signal truncation bandwidth.* We see that $N_A(S)$ goes to a finite value as $S \to 0$, and that its variation at large S is exponential; in both respects, it differs from $N_R(S)$ and $N_P(S)$. If we now define

$$K \triangleq \frac{8\pi^2}{27} \frac{b_2}{(2\pi W)^2 b_1}, \tag{13}$$

(12) reduces to the form

$$N_A(S) = K \left(\frac{b_1^2}{b_2}\right) (1 + 3S) \exp(-3S). \tag{14}$$

Evaluating K for the spectra studied by O'Neal, we find that it lies between 1.074 and 1.75 for the three cases, a range of just 2 dB. It is tempting to speculate, therefore, that (14) is a more correct form in general than (12), with K a universal constant on the order of unity, and that the apparent 2-dB spread in K over the three spectra is a result of experimental uncertainties. If we accept this notion, we again have the result that the ratio of noise power to signal power depends solely on S and γ.

III. NOISE POWER IN TERMS OF b_0, b_1, b_2

3.1 *The Two-Band Process*

The published results cited above suggest that $N(S)$ is essentially the same for all processes having the same values for the zeroth, first, and second moments. Assuming for the present that this supposition is correct, we call attention to the two-band process, the spectrum of which is shown in Fig. 3a. In the limit as the two bands become infinitely narrow, the zeroth, first, and second moments become precisely b_0, b_1, and b_2. By analyzing this simple process, therefore, we can derive an exact noise power formula $[N_{tb}(S)]$ in terms of b_0, b_1, and b_2 and then see how universal it really is.

* The three spectra treated by O'Neal are of the truncated Butterworth type, with corner frequency-to-bandwidth ratios of 0.068, 0.25, and ∞. The data used to derive (12) cover an S-range from about 1.4 to 4.2.

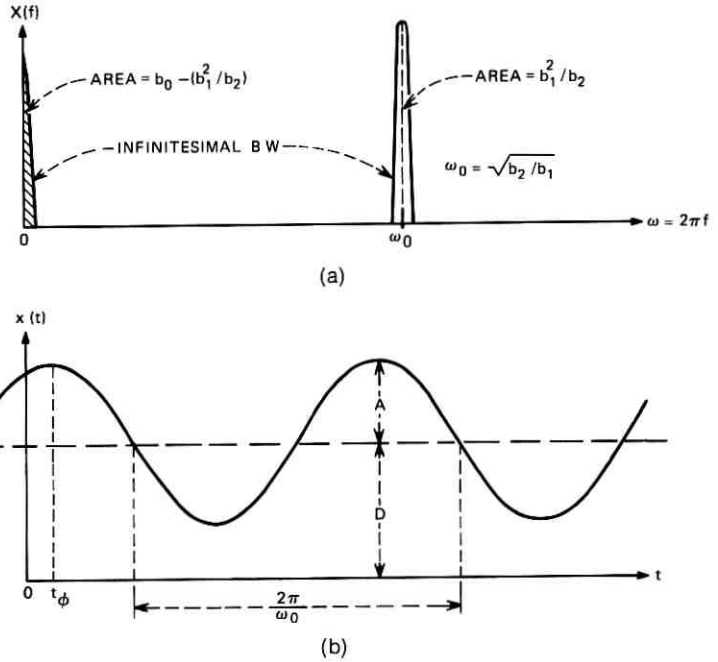


Fig. 3—The two-band random process. (a) Power spectrum. (b) Sample function.

To appreciate the simplicity of the two-band process for purposes of analysis, we should view it in the time domain. A sample function is shown in Fig. 3b and is seen to consist of a dc level, D, plus a sinusoid having radian frequency $\omega_o = \sqrt{b_2/b_1}$, phase $\phi = \omega_o t_\phi$, and amplitude A. D is a Gaussian variate, whose mean and variance across the sample function ensemble are 0 and $(b_0 - b_1^2/b_2)$, respectively; ϕ is a uniformly distributed variate on $[-\pi, +\pi]$; A is a Rayleigh variate of mean-square value $2b_1^2/b_2$; and D, ϕ, and A are mutually independent.

We can derive the slope overload noise power for this process by finding the mean noise power associated with a given sample function and averaging over the distributions on D, ϕ, and A. This approach is simplified by the fact that D and ϕ do not really influence the result. That is, the noise pattern for a given sample function converges ultimately to a variation about D that depends solely on the sinewave amplitude and frequency (Fig. 4).

We thus see that the noise power per sample function is the steady-state noise energy per half-cycle, denoted by $\mathcal{E}(A)$, times the rate of

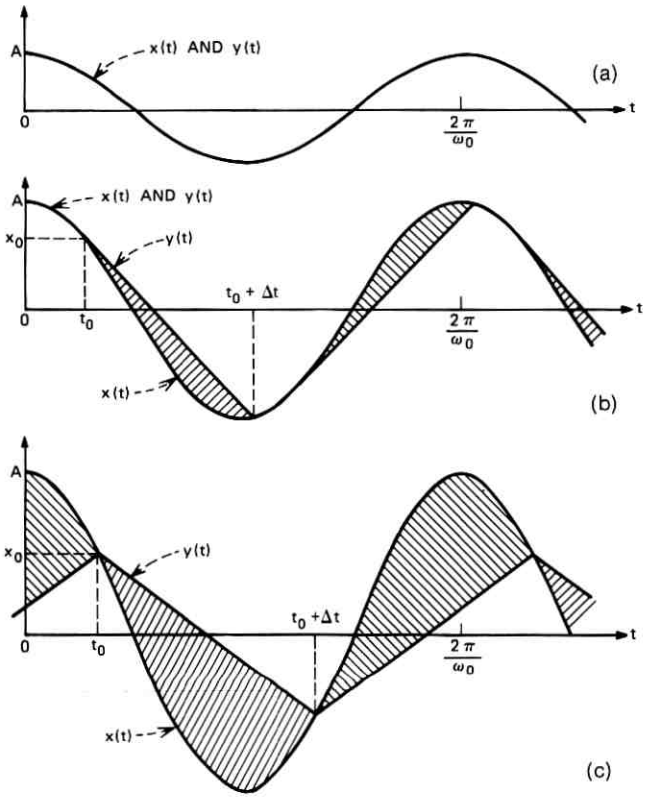


Fig. 4—The three regions of sinewave amplitude. (a) $0 \leqq A \leqq x_o'/\omega_o$ (no noise). (b) $(x_o'/\omega_o) < A < (x_o'/\omega_o)\sqrt{1 + \pi^2/4}$ (primary noise). (c) $A \geqq (x_o'/\omega_o)\sqrt{1 + \pi^2/4}$ (secondary noise).

occurrence of half-cycles, which is $\sqrt{b_2/b_1}/\pi$. The total noise power for the two-band process is then the average of this quantity over the distribution on A, i.e.,

$$N_{tb} = \frac{1}{\pi}\sqrt{\frac{b_2}{b_1}}\int_0^\infty \mathcal{E}(A)\underbrace{p(A)}_{\text{(pdf of } A)}dA\,. \tag{15}$$

Since A is known to be a Rayleigh variate of mean-square value $(2b_1^2/b_2)$, the only unknown is the energy function $\mathcal{E}(A)$.

3.2 *Derivation of* $\mathcal{E}(A)$

Three distinct regions of A can be identified, each giving rise to a distinct pattern of signal and noise. The region depicted in Fig. 4a

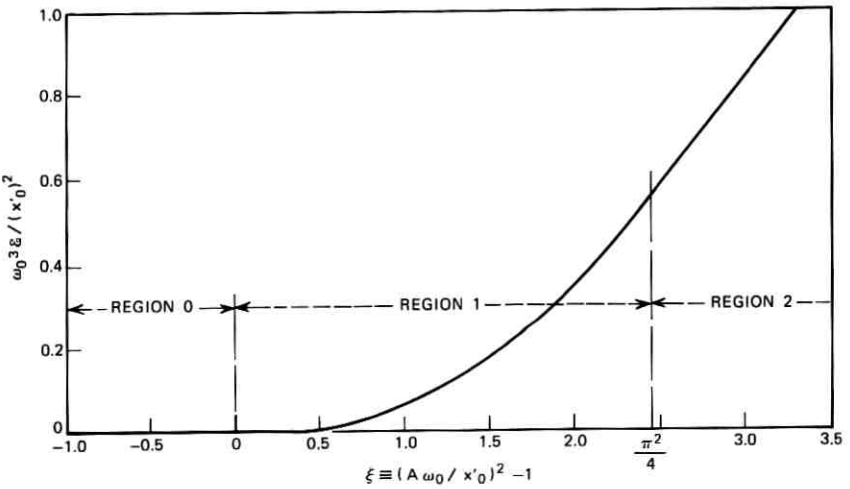


Fig. 5—Noise energy per half-cycle as a function of A.

(Region 0) is one in which no slope overload occurs because $|x'(t)| < x'_o$ at all t. Clearly, $\mathcal{E}(A) = 0$ in this case. The region depicted in Fig. 4b (Region 1) is one in which slope overload occurs in each half-cycle, but over less than the complete interval. We see that the noise in this case is primary noise, as defined earlier. Finally, the region depicted in Fig. 4c (Region 2) is one in which slope overload occurs for the entire duration of each half-cycle. The noise in this case is seen to be secondary noise.*

To find $\mathcal{E}$ in Region 1, we analyze the burst spanning $[t_o, t_o + \Delta t]$ in Fig. 4b. The noise energy in this burst is

$$\mathcal{E} = \int_{t_o}^{t_o+\Delta t} [A \cos \omega_o t - \{x_o - x'_o(t - t_o)\}]^2 dt. \tag{16}$$

The quantities x_o and t_o are found by means of trigonometric identities to be

$$x_o = \sqrt{A^2 - (x'_o/\omega_o)^2} \quad \text{and} \quad t_o = \frac{1}{\omega_o} \sin^{-1}\left(\frac{x'_o}{A\omega_o}\right). \tag{17}$$

Unfortunately, the quantity Δt cannot be solved for explicitly, but is

* For the special process under discussion, we see that a given sample function has either no noise at all, primary noise only, or secondary noise only. For more spectrally distributed processes, both kinds of noise occur in the same sample function and in a less regular pattern.

related to A by the transcendental equation

$$\left[\frac{\omega_o\Delta t - \sin(\omega_o\Delta t)}{1 - \cos(\omega_o\Delta t)}\right]^2 = \left(\frac{A\omega_o}{x_o'}\right)^2 - 1 \triangleq \xi. \tag{18}$$

We can relate $\mathcal{E}$ to A (or ξ) by combining (16), (17), and (18) and eliminating the common parameter, $\omega_o\Delta t$, between (16) and (18). The result is a unique correspondence between $\omega_o^3\mathcal{E}/(x_o')^2$ and ξ having the variation shown in Fig. 5 for Region 1. (Note that Region 1 corresponds to the ξ-interval $[0, \pi^2/4]$.) A highly accurate functional description of this variation has been found to be

$$\frac{\omega_o^3\mathcal{E}}{(x_o')^2} = \frac{81}{140\pi}\xi^{7/2}[0.43\exp(-0.5014\xi) + 0.57\exp(-2.10\xi)]; \quad 0 \leqq \xi < \pi^2/4. \tag{19}$$

Comparison with exact results shows this function to be accurate to within 0.8 percent.

To find $\mathcal{E}$ in Region 2, we analyze the burst spanning $[t_o, t_o + \Delta t]$ in Fig. 4c. In this case, x_o and t_o are given by (17), as before, but $\omega_o\Delta t$ is precisely π for all A. Applying these relationships to (16) and performing the indicated integration, we obtain

$$\frac{\omega_o^3\mathcal{E}}{(x_o')^2} = \frac{1}{2}\left[\xi - \left(3 - \frac{\pi^2}{6}\right)\right]; \qquad \xi \geqq \pi^2/4. \tag{20}$$

This variation is also shown in Fig. 5.

3.3 *Expression for $N_{tb}(S)$*

We now have expressions for $\mathcal{E}$ in the three regions of A and can average this complete result over the distribution on A (or ξ) to find N_{tb}, (15). By using (18), along with $\omega_o^2 = b_2/b_1$, $S = x_o'/\sqrt{b_1}$, and the fact that A is a Rayleigh variate of mean-square value $(2b_1^2/b_2)$, we obtain

$$p(\xi) = \frac{S^2}{2}\exp\left(-\frac{S^2}{2}\right)\exp\left(-\frac{S^2}{2}\xi\right); \qquad \xi \geqq -1$$
$$= 0; \qquad \text{elsewhere.} \tag{21}$$

Combining this with (15), (19), and (20), we obtain

$$N_{tb}(S) = \underbrace{\left(\frac{b_1}{b_2}\right)^2 F_1(S)}_{\substack{\text{Primary}\\ \text{Noise}}} + \underbrace{\left(\frac{b_1}{b_2}\right)^2 F_2(S)}_{\substack{\text{Secondary}\\ \text{Noise}}} = \left(\frac{b_1}{b_2}\right)^2 F(S) \tag{22}$$

where $F(S) \triangleq F_1(S) + F_2(S)$, and

$$F_1(S) = \frac{81}{560}\left(\frac{\pi}{2}\right)^8 S^4 \exp\left(-\frac{S^2}{2}\right)\left[0.43Q\left\{\frac{\pi^2}{4}\left(\frac{S^2}{2}+0.5014\right)\right\} + 0.57Q\left\{\frac{\pi^2}{4}\left(\frac{S^2}{2}+2.10\right)\right\}\right]; \quad (23)$$

$$F_2(S) = \left[1+\left(\frac{5\pi^2-36}{24}\right)S^2\right]\exp\left\{-\left(1+\frac{\pi^2}{4}\right)\frac{S^2}{2}\right\}; \quad (24)$$

$$Q\{u\} = \frac{1}{u^{9/2}}\left\{\frac{105\sqrt{\pi}}{16}\operatorname{erf}(\sqrt{u}) - \left[u^{7/2}+\frac{7}{2}u^{5/2}+\frac{35}{4}u^{3/2}+\frac{105}{8}u^{1/2}\right]\exp(-u)\right\}. \quad (25)$$

The variation $F(S)$ is shown in Fig. 6 along with its component parts, $F_1(S)$ and $F_2(S)$. These curves indicate, for the two-band process, which region of S is dominated by primary noise and which region by secondary noise.

The only inexactness in our result for $N_{tb}(S)$ arises from the functional fit to $\mathcal{E}$ in Region 1, (19). Since this fit is accurate to within ± 0.8 percent at all ξ, and both $\mathcal{E}(\xi)$ and $p(\xi)$ are non-negative func-

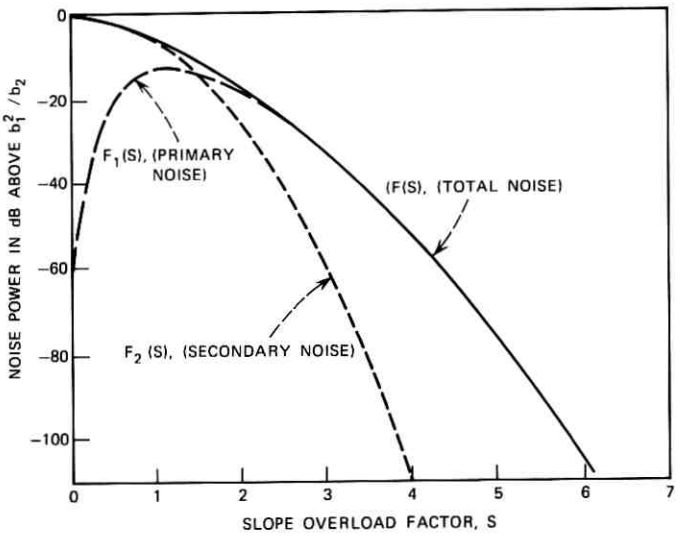


Fig. 6—Normalized noise powers for the two-band process.

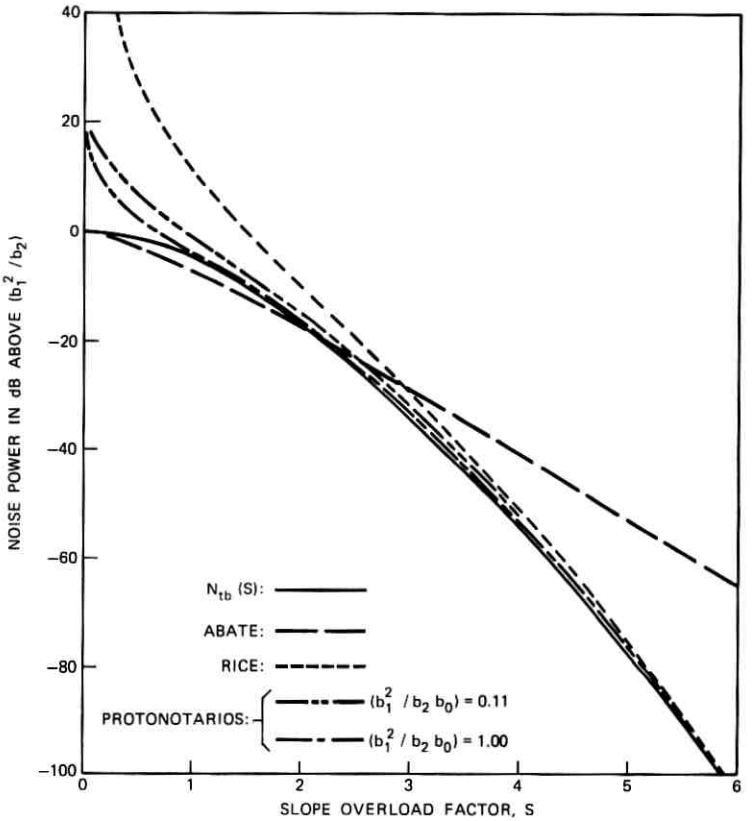


Fig. 7—Comparison of $N_{tb}(S)$ with previous results.

tions, the given expression for $N_{tb}(S)$ is also accurate to within ± 0.8 percent (or ± 0.035 dB).

3.4 *Comparisons and Interpretations*

The variation of N_{tb} with S is plotted in Fig. 7, along with the variations of N_R, (9); N_P, (10); and N_A, [(14) with $K = 1$]. Of particular interest is the fact that $N_{tb}(S)$ converges with the results of Rice and Protonotarios as $S \to \infty$. The approximations underlying the analyses of Rice and Protonotarios become increasingly valid as S increases, so the convergence of $N_{tb}(S)$ with their results is not surprising. The apparent dependence of noise power on the moments b_1 and b_2 alone at large S is explained by the fact that slope overload bursts are of

short duration in this region; hence, they are shaped primarily by the lower-order curvature of $x(t)$, which is reflected in b_1 and b_2. As S decreases towards small values, however, the higher-order curvature of $x(t)$, reflected in b_3, b_4, etc., also influences the noise bursts and, consequently, the noise power. We should therefore expect the true noise power $[N(S)]$ associated with a given spectrum to reflect other features of that spectrum at low S besides b_1 and b_2. If this is so, then $N_{tb}(S)$ cannot be assumed to be a universal formula.

To show that this expectation is correct, we note that, as $S \to 0$, N_{tb} converges to b_1^2/b_2, which is precisely the ac power (σ^2) in the two-band process. This is a quite general result, i.e., $N(S)$ always converges to σ^2 as $S \to 0$. However, some other process having the same b_0, b_1, and b_2 can have an ac power anywhere between b_1^2/b_2 and b_0, so that $N(S)$ will not converge to b_1^2/b_2 for all spectra having these moments in common.

We conclude, then, that $N(S)$ and $N_{tb}(S)$ converge at high S but become increasingly dissimilar as S decreases towards zero, the dissimilarity depending on the differences between the actual process spectrum and the two-band spectrum having the same b_0, b_1, and b_2. In Part 2 (following), we will derive a factor which relates $N_{tb}(S)$ and $N(S)$ when $X(f)$ is not a two-band spectrum.

Part 2

IV. NOISE POWER FOR $S \geqq 3.5$

We now derive a noise power formula applicable to all spectra, with S large (primary noise only). The derivation combines the power series method of Rice with the infinite-time averaging procedure used by Zetterberg and Protonotarios. The differences here are that many terms of the power series expansion are included, and steps are taken to minimize the contributions from spurious bursts. The complicated expression that results is reduced to a simple formula $[N_l(S)]$ that displays explicitly the influence of the higher-order spectral moments.

4.1 *Power Series Representation of the Noise Burst*

Let $t = 0$ be the time origin of a particular positive-going burst, so that $x'(0) = x_o'$ and $x''(0) > 0$, e.g., Fig. 2. Beginning at $t = 0$, $\bar{y}(t)$ follows the straight line $x(0) + x_o't$ until this ramp function intersects $x(t)$ again. We can therefore define the quantity

$$v(t) \triangleq x(t) - [x(0) + x'_o t]; \qquad t > 0 \tag{26}$$

and write the variation of the noise burst as

$$\begin{aligned} n(t) &= v(t) \quad \text{if } v(t') > 0 \quad \text{for all} \quad 0 < t' \leqq t \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{27}$$

Henceforth, we shall use the symbols n_t and v_t to denote $n(t)$ and $v(t)$, respectively.

Following Rice, we can express $x(t)$ by a power series expansion about $t = 0$. Noting that $x'(0) = x'_o$ and denoting $x''(0)$ by x''_o, v_t can be written as

$$v_t = \tfrac{1}{2}x''_o t^2 + \frac{1}{3!}x'''(0)t^3 + \cdots \frac{1}{m!}x^{(m)}(0)t^m + \cdots. \tag{28}$$

We now invoke the property of Gaussian random processes that the mth and $(m-2)$th derivatives at a given time instant are related by

$$x^{(m)} = -\frac{b_{m-1}}{b_{m-2}}x^{(m-2)} + d_m \tag{29}$$

where d_m is a zero-mean Gaussian variate of mean-square value

$$\overline{d_m^2} = b_m - \frac{b_{m-1}^2}{b_{m-2}}. \tag{30}$$

With this relationship we can write v_t as follows:

$$\begin{aligned} v &= \left\{x''_o\left[\frac{t^2}{2!} - \frac{b_3}{b_2}\frac{t^4}{4!} + \frac{b_3}{b_2}\cdot\frac{b_5}{b_4}\frac{t^6}{6!} - \cdots\right]\right. \\ &\quad \left. - x'_o\left[\frac{b_2}{b_1}\frac{t^3}{3!} - \frac{b_2}{b_1}\cdot\frac{b_4}{b_3}\frac{t^5}{5!} + \frac{b_2}{b_1}\cdot\frac{b_4}{b_3}\cdot\frac{b_6}{b_5}\frac{t^7}{7!} - \cdots\right]\right\} \\ &\quad + \left\{d_3\frac{t^3}{3!} + d_4\frac{t^4}{4!} - \left[\frac{b_4}{b_3}d_3 - d_5\right]\frac{t^5}{5!} - \left[\frac{b_5}{b_4}d_4 - d_6\right]\frac{t^6}{6!} + \cdots\right\} \\ &\triangleq A_t + \delta_t \end{aligned} \tag{31}$$

where A_t is the first bracketed quantity and δ_t is the second. We see that A_t is the conditional mean of v_t given that $x'(0) = x'_o$ and $x''(0) = x''_o$; and that δ_t is a linear sum of mutually independent zero-

mean Gaussian variates (d_3, d_4, etc.) whose mean-square value at time t is therefore

$$\overline{\delta_t^2} \equiv \sigma_t^2 = \left[\left(\frac{t^3}{3!} - \frac{b_4}{b_3}\frac{t^5}{5!} + \cdots\right)^2 \overline{d_3^2} + \left(\frac{t^4}{4!} - \frac{b_5}{b_4}\frac{t^6}{6!} + \cdots\right)^2 \overline{d_4^2} + \cdots\right]. \quad (32)$$

We thus have the result that the conditional pdf of v at time t is

$$p\{v_t | x_o', x_o''\} = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left\{-\frac{1}{2}\left(\frac{v_t - A_t}{\sigma_t}\right)^2\right\}. \quad (33)$$

4.2 *Derivation of Mean Noise Power*

Let $\overline{n_t^2}$ be the conditional mean square value of n at time t, given that $x'(0) = x_o'$ and $x''(0) \equiv x_o'' > 0$. To find the mean burst energy when $x'(0) = x_o'$, we must average $\overline{n_t^2}$ with respect to x_o'' and integrate the result over all time. We can then multiply this mean energy by the mean rate of primary noise burst occurrences to obtain the mean noise power. This approach is valid so long as secondary noise can be neglected, which it can in the region $S \geqq 3.5$ under consideration.

From Protonotarios,[3] we know that the correct density function for averaging with respect to the conditions $x'(0) = x_o'$ and $x_o'' > 0$ is

$$p(x_o'') = \frac{x_o''}{b_2} \exp\left\{-\frac{(x_o'')^2}{2b_2}\right\}; \qquad x_o'' \geqq 0 \quad (34)$$

while from Rice[8] we know that the mean rate of noise burst occurrences is $(\sqrt{b_2/b_1}/\pi) \exp(-S^2/2)$. The noise power in the region $S \geqq 3.5$ can thus be accurately given by

$$N(S) = \frac{1}{\pi}\sqrt{\frac{b_2}{b_1}} \exp\left(-\frac{S^2}{2}\right) \int_0^\infty \int_0^\infty \frac{x_o''}{b_2} \cdot \exp\left\{-\frac{(x_o'')^2}{2b_2}\right\} \overline{n_t^2(x_o', x_o'')} dx_o'' dt. \quad (35)$$

The remaining analytical task is to find $\overline{n_t^2(x_o', x_o'')}$.

The method of Protonotarios in this regard amounts to averaging v_t^2 over positive values using the conditional pdf of v_t, (33). To do this, however, is to incur the spurious burst problem depicted by Fig. 8a. The following line of reasoning leads to an improved procedure.

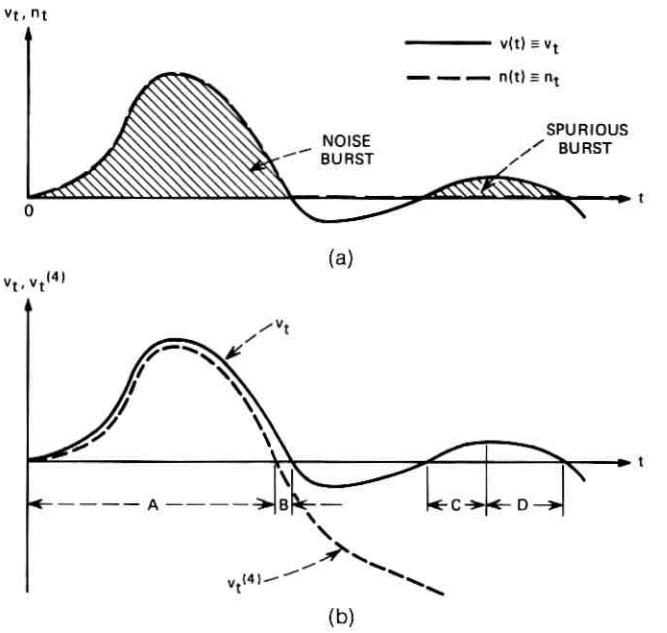


Fig. 8—An illustration of v_t and n_t. (a) v_t and n_t. (b) v_t and $v_t^{(4)}$.

Let $A_t^{(4)}$ denote the partial power series for A_t up to and including the fourth-order term, and let

$$v_t^{(4)} \equiv A_t^{(4)} + \delta_t. \tag{36}$$

We can show that $v_t^{(4)} \approx v_t$ when t is small and that $v_t > v_t^{(4)}$ in general. What we have found in particular is that, for S greater than about 2, the difference between v_t and $v_t^{(4)}$ does not widen appreciably over a typical burst duration (see Fig. 8b). Also, since $A_t^{(4)}$ is a strongly decreasing function of t beyond its peak, $v_t^{(4)}$ tends to be the same.

We therefore calculate $\overline{n_t^2}$ according to the following criterion: A given value of v_t is counted as part of the noise (i.e., n_t is assumed to be v_t) if either $\{v_t > 0, v_t^{(4)} > 0\}$ or $\{v_t > 0, v_t^{(4)} < 0, \dot{v}_t < 0\}$;* otherwise, v_t is not counted and n_t is assumed to be 0. For the sample function shown in Fig. 8b, this means that the energies in intervals A and B are both counted in the calculation of $\overline{n_t^2}$; the energy in interval C is not counted, since $v_t^{(4)} < 0$ and $\dot{v}_t > 0$; and the energy in interval D is counted, even though it is clearly spurious. We can reduce contribu-

* The symbol $\dot{v}_t$ stands for $dv(t)/dt$.

tions of the latter type, however, by using a suitably finite upper limit in the integration over time in (35). That is, we choose an upper limit sufficiently high that virtually all legitimate noise contributions are counted in the calculation but sufficiently low that some spurious contributions are omitted. By studying the time variations of A_t and δ_t under a wide variety of conditions, we have settled on an upper time limit of $3.5\sqrt{b_1/b_2}$.

Using the above criterion, we have been able to show that

$$\overline{n_t^2} = \frac{\sigma_t^2}{2}\left\{(1+2\alpha^2)[1+\operatorname{erf}(\alpha_4)] + \frac{2}{\sqrt{\pi}}(2\alpha-\alpha_4)\exp(-\alpha_4^2)\right\} + \frac{\sigma_t^2}{4}(\operatorname{erfc}(\dot{\alpha}))\left\{(1+2\alpha^2)[\operatorname{erf}(\alpha)-\operatorname{erf}(\alpha_4)] + \frac{2}{\sqrt{\pi}}[\alpha\exp(-\alpha^2)-(2\alpha-\alpha_4)\exp(-\alpha_4^2)]\right\} \quad (37)$$

where:

$$\alpha \triangleq A_t/\sqrt{2\sigma_t^2}; \quad (38)$$

$$\alpha_4 \triangleq A_t^{(4)}/\sqrt{2\sigma_t^2}; \quad (39)$$

and

$$\dot{\alpha} \triangleq \dot{A}_t/\sqrt{2\overline{(\dot{\delta}_t)^2}}. \quad (40)$$

To complete the derivation of $N(S)$, we insert this complicated expression into (35) and perform the integrations over x_o'' and t, remembering to use an upper limit of $3.5\sqrt{b_1/b_2}$ in the time integral.

4.3 *Simplified Formula*

The formal solution just described is believed to be more exact than that of Protonotarios [eq. (53) of Ref. 3], because it entails a substantial reduction in the spurious burst contributions. Furthermore, in reducing his formal solution to computation, Protonotarios makes a number of functional approximations which obscure the influence of the higher-order moments b_3, b_4, etc. In reducing the present solution to computation, we must also make approximations but of a different kind, i.e., those associated with performing numerical double integration and truncating the infinite power series in the integrand. The impact of the higher-order moments, however, is not approximated away in the process; in fact, the following discussion develops an empirical expression [denoted by $N_I(S)$] which exhibits these influences explicitly.

We define a dimensionless quantity, which we call the nth-moment excess, as

$$M_n \triangleq \frac{b_n - (b_{n-1}^2/b_{n-2})}{(b_2^{n-1}/b_1^{n-2})}; \qquad n \geqq 3. \tag{41}$$

From (6) we know that $M_n \geqq 0$ for all $n \geqq 3$ and, using Fig. 3, we can show that $M_n = 0$ for all $n \geqq 3$ if and only if $X(f)$ is the two-band spectrum. We can therefore say that $N(S)$ differs from $N_{tb}(S)$ only to the extent that M_3, M_4, etc., are not zero. A simple model that reflects this fact is one which describes the logarithmic difference between N and N_{tb} as a linear combination of the moment excesses,

$$\ln(N/N_{tb}) = AM_3 + BM_4 + CM_5 + \cdots \tag{42}$$

where the coefficients A, B, etc., are, in general, functions of S. To the extent that this representation is accurate, it is reasonable to assume further that the higher-order terms, involving M_6, M_7, etc., are negligible for $S \geqq 3.5$. The reason is that these moment excesses have little influence on slope overload bursts at such high slope-following capacities. The result of this line of reasoning is an approximation of the form

$$N(S \geqq 3.5) \doteq N_{tb}(S) \exp\{A(S)M_3 + B(S)M_4 + C(S)M_5\} \triangleq N_t(S). \tag{43}$$

We have tested this approximation by applying the double integral solution for $N(S)$, (35) and (37), to a number of spectra, using a ninth-order power series to represent v_t. With the computed noise powers for these spectra, we have then obtained results for $A(S)$, $B(S)$, and $C(S)$ over the range $3.5 \leqq S \leqq 10.0$ (Fig. 9) which appear to be quite accurate. In particular, the use of (43) with these results is found to predict $N(S \geqq 3.5)$, as computed from (35), to within 0.2 dB for all spectra of practical interest.

There exist, however, special conditions on $X(f)$ for which (43) yields too high estimates of the true noise power. This can occur when $X(f)$ contains an isolated high-frequency component which contributes materially to b_3, b_4, and/or b_5 but not to b_0, b_1, or b_2. Under such circumstances, $x(t)$ might change too rapidly during the noise bursts to be properly analyzed by the power series approach as used here. Calculations indicate that these cases fall outside the range of practical interest, so we need not consider them further.

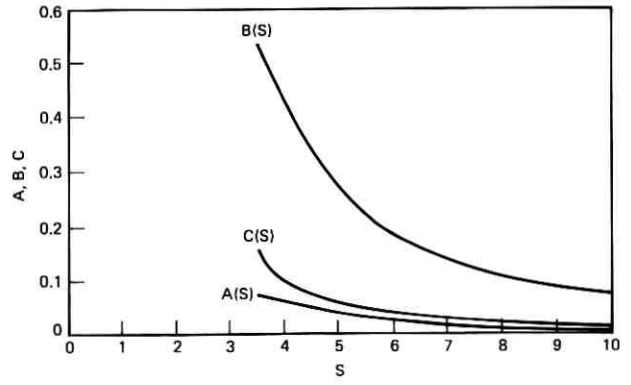


Fig. 9—Empirical results for A, B, and C; $S \geqq 3.5$.

V. NOISE POWER FOR $S \approx 0$

We now derive another noise power formula applicable to all spectra, but with S very small (secondary noise only). The representation we seek is a partial power series of the form

$$N(S \approx 0) \doteq \sigma^2[C_0 + C_1 S + C_2 S^2)] \triangleq N_s(S) \tag{44}$$

where σ^2 is the ac power of the input process.

5.1 *Formulation*

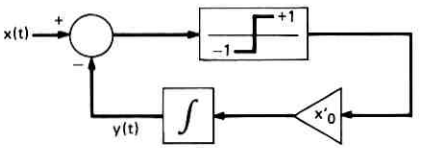
The analysis assumes $\{x(t)\}$ to be just the ac part of the input process, so that $\overline{x^2(t)} = \sigma^2$. Since dc components have no effect on noise power for $S \neq 0$, we are justified in ignoring them.

We consider x_o' to be so small compared to $\sqrt{b_1}$ (i.e., S so small compared to unity) that the delta modulator is always in slope overload. In this case, the decoded output [denoted here by $y(t)$ rather than $\bar{y}(t)$] is just a succession of alternating ramps of slope $\pm x_o'$, the slope polarity reversing whenever $y(t)$ intersects $x(t)$. This situation is illustrated in Fig. 10b. A valid model of the delta modulator for analyzing this case is given by Fig. 10a (cf. Ref. 9), from which we see that the noise signal is

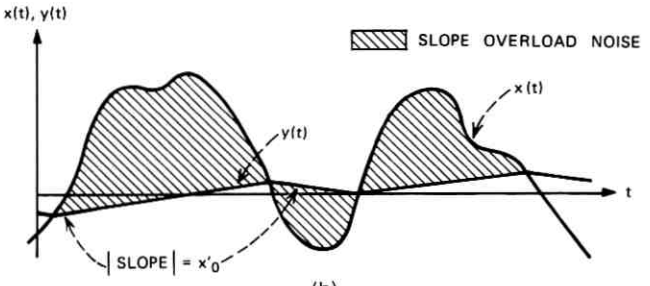
$$n(t) = x(t) - y(t) = x(t) - x_o' \int_{-\infty}^{t} n_q(u)du \tag{45}$$

where

$$n_q(\cdot) \triangleq \operatorname{sgn}\{n(\cdot)\}. \tag{46}$$



(a)



(b)

Fig. 10—Equivalent delta modulator and signals for $S \approx 0$. (a) ΔM model for $S \approx 0$. (b) Input and feedback signals.

The noise power in the region $S \approx 0$ can thus be accurately given by

$$N(S) = \overline{x^2(t)} - \underbrace{2x_o' \overline{\left[x(t) \int_{-\infty}^{t} n_q(u)du \right]}}_{-2\overline{x(t)y(t)}} + \underbrace{(x_o')^2 \overline{\int_{-\infty}^{t} \int_{-\infty}^{t} n_q(u)n_q(v)dudv}}_{\overline{y^2(t)}}. \quad (47)$$

If we envision each of these three terms as a power series in x_o', we see that $\overline{x^2(t)}$ has a zero-order term only (i.e., is independent of x_o'); $-2\overline{x(t)y(t)}$ may contain first-, second-, and higher-order terms; and $\overline{y^2(t)}$ may contain second-order terms and higher. We can thus rewrite (47) as

$$N(S) = \underbrace{K_0}_{\overline{x^2(t)}} + \underbrace{[K_1 x_o' + K_{2a}(x_o')^2 + \cdots]}_{\overline{-2x(t)y(t)}} + \underbrace{[K_{2b}(x_o')^2 + \cdots]}_{\overline{y^2(t)}}. \quad (48)$$

Retaining only terms up to second-order, substituting $x_o' = S\sqrt{b_1}$, and

comparing with (44), we see that

$$C_0 = \frac{K_0}{\sigma^2}; \qquad C_1 = \frac{K_1\sqrt{b_1}}{\sigma^2}; \qquad C_2 = \frac{(K_{2a} + K_{2b})b_1}{\sigma^2}. \tag{49}$$

Clearly, $K_0 = \overline{x^2(t)} = \sigma^2$, so that $C_0 = 1$ in all cases. The remainder of this section is devoted to finding C_1 and C_2.

5.2 *Derivation of* C_1

We will derive C_1 by evaluating the first-order term of $-2\overline{x(t)y(t)}$, (47), and then invoking (49). For x_o' very small, $n_q(t)$ does not differ appreciably from $x_q(t) \equiv \operatorname{sgn}\{x(t)\}$, which is the quantizer response to $x(t)$ applied alone. It is therefore fruitful to represent $n_q(t)$ as

$$n_q(t) = x_q(t) + \theta(t) \tag{50}$$

where $\theta(t)$ is a correction signal related to the finiteness of x_o' and is defined by

$$\theta = \begin{cases} +2 & \text{if } y < x < 0 \\ -2 & \text{if } y > x > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{51}$$

The relationships between $n_q(t)$, $x_q(t)$, and $\theta(t)$ are illustrated in Fig. 11. Although the peak magnitude of $\theta(t)$ is fixed, its "duty cycle" can be seen to depend directly on x_o'.

Inserting (50) into the cross term of (47) and interchanging the order of integration and averaging yields

$$-2\overline{x(t)y(t)} = -2x_o' \int_{-\infty}^{t} \overline{x(t)x_q(u)}du - 2x_o' \int_{-\infty}^{t} \overline{x(t)\theta(u)}du. \tag{52}$$

Since the duty cycle of $\theta(t)$ vanishes as $x_o' \to 0$, we can see that the first term must be identical to K_1x_o' in (48), while the second term contains $K_{2a}(x_o')^2$ plus higher-order terms in x_o'.

To evaluate the first term, we use a relationship applicable to $\{x(t)\}$ because it is a Gaussian process, namely,

$$\overline{x(t)x_q(u)} = \sqrt{\frac{2}{\pi}}\frac{\overline{x(t)x(u)}}{\sigma} = \sqrt{\frac{2}{\pi}}\frac{R_x(t-u)}{\sigma} \tag{53}$$

where $R_x(\tau)$ is the autocorrelation function of $\{x(t)\}$. We now apply this to the first term of (52) and use the fact that the area under $R_x(\tau)$ from $\tau = 0$ to $\tau = \infty$ is $\frac{1}{4}X(0)$. Equating the result to K_1x_o'

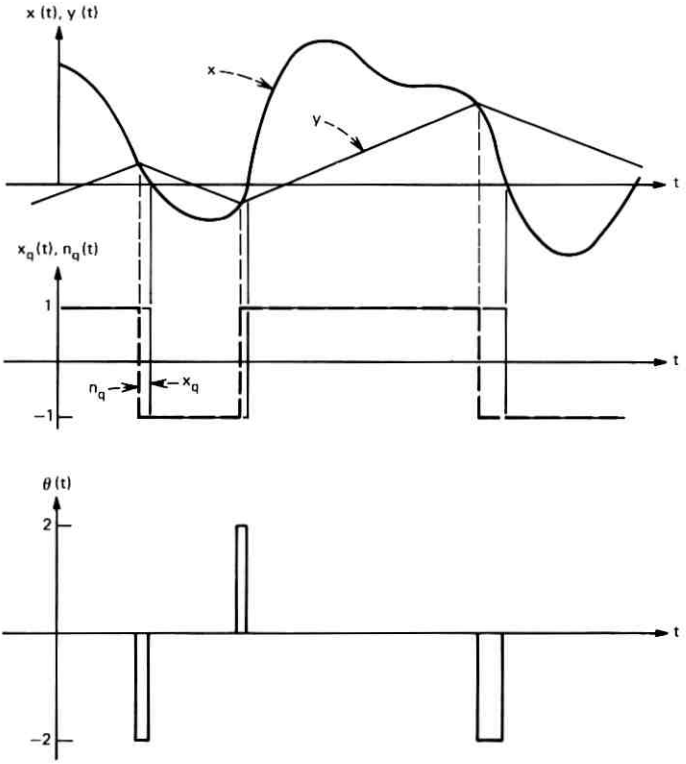


Fig. 11—The relationships between $x(t)$, $y(t)$, $x_q(t)$, $n_q(t)$, and $\theta(t)$.

we obtain

$$K_1 = -\frac{1}{\sqrt{2\pi}} \frac{X(0)}{\sigma} \tag{54}$$

which, from (49), yields

$$C_1 = -\frac{1}{\sqrt{2\pi}} \left[\frac{X(0)\sqrt{b_1}}{\sigma^3} \right]. \tag{55}$$

5.3 *Derivation of C_2*

The general approach used in the preceding analysis was also applied to finding C_2. Unfortunately, it leads to an approximation which is both highly complicated and not sufficiently precise for many spectra. For this reason, we have resorted to finding C_2 for a not-quite-Gaussian process having the same spectrum as the specified Gaussian one, and verifying that it is virtually exact under what appear to be "worst-

case" conditions. The details, which are documented but not published,[10] are quite involved and so we shall merely outline the approach.

Consider a wide sense stationary process $\{\tilde{x}(t)\}$ having sample functions of the form

$$\tilde{x}(t) = A \cos(\tilde{\omega}t + \phi) \tag{56}$$

where A is a Rayleigh-distributed amplitude of mean-square value $2\sigma^2$; ϕ is a random phase uniformly distributed over $[-\pi, +\pi]$; and $\tilde{\omega}$ is a random frequency whose pdf is

$$p(\tilde{\omega}) = \frac{1}{2\sigma^2} X(f = |\tilde{\omega}|/2\pi); \qquad \text{all } \tilde{\omega}. \tag{57}$$

It can be shown that the power spectrum for this process is $X(f)$, and that the ensemble pdf of $\tilde{x}$ at any t is Gaussian with zero mean and variance σ^2. Thus, $\{\tilde{x}(t)\}$ has certain properties in common with the Gaussian process $\{x(t)\}$ having the spectrum $X(f)$.

We can derive the mean slope overload noise power for the process $\{\tilde{x}(t)\}$ by finding the noise power for a single sample function, (56), and averaging over the distributions on A, ϕ, and $\tilde{\omega}$. Using the methods of Section III, we can show that the averaging over A and ϕ yields

$$\tilde{N}(x_o' \mid \tilde{\omega}) = \tfrac{1}{2}\overline{A^2} F\left(\frac{\sqrt{2}x_o'}{\sqrt{\overline{A^2}}\tilde{\omega}}\right) \tag{58}$$

where $F(\cdot)$ is defined by (23) through (25). Averaging (58) over the pdf of $\tilde{\omega}$, (57), and replacing $\overline{A^2}$ by $2\sigma^2$ and x_o' by $S\sqrt{b_1}$, we obtain

$$\tilde{N}(S) = \int_0^\infty X(f) F\left(\frac{\sqrt{b_1}S}{2\pi f\sigma}\right) df. \tag{59}$$

The derivatives of this function with respect to S can be found quite simply, although some caution is required when evaluating them at $S = 0$. The resulting value for $C_2 (= \frac{1}{2}\tilde{N}''(0))$ is found to be

$$C_2 = \lim_{S \to 0} \left\{ \frac{\sqrt{b_1}}{S\sigma^3} \int_0^\infty \frac{[X(f) - X(0)]}{2\pi f} \frac{d}{dS} F\left(\frac{\sqrt{b_1}S}{2\pi f\sigma}\right) df \right\}. \tag{60}$$

For most realistic processes, the continuous part of the power spectrum has zero slope at $f = 0$, i.e., $X'(0) = 0$. In that circumstance, (60) reduces to the much simpler form

$$C_2 = \left(\frac{\pi^2}{12} - 2\right) \frac{b_1}{\sigma^4} \int_0^\infty \frac{[X(f) - X(0)]}{(2\pi f)^2} df; \qquad X'(0) = 0. \tag{61}$$

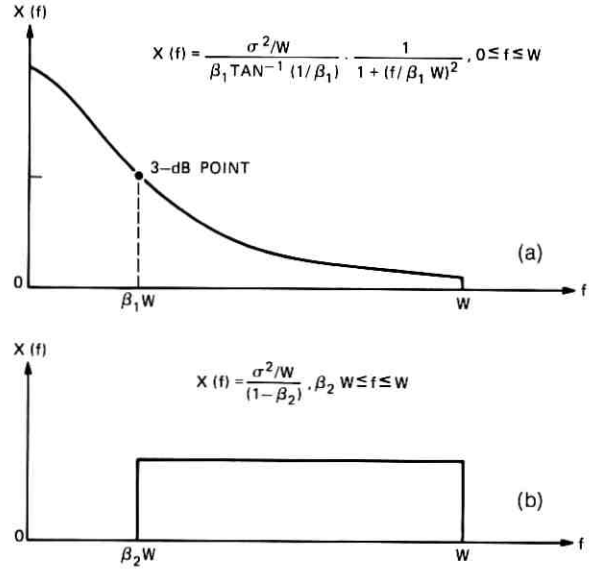


Fig. 12—Two families of spectra. (a) Truncated Butterworth spectrum. (b) Bandpass uniform spectrum.

In general, the above C_2 for the non-Gaussian process $\{\tilde{x}(t)\}$ cannot be compared with the C_2 for the Gaussian process $\{x(t)\}$ since we do not have a result for the latter case that is good for all $X(f)$. We do, however, have exact results under some special conditions on $X(f)$ for which C_2 is expected to be maximally dissimilar for the two processes. The very close agreement observed under these conditions persuades us that (60) is a reliable representation of C_2 for Gaussian processes having *any* $X(f)$.

5.4 *Final Result for* $N_s(S)$

Our final approximation to $N(S \approx 0)$ is (44) with $C_0 = 1$ and C_1 and C_2 given by (55) and (60). Note that when $X(0) = 0$, C_1 is zero and C_2 is negative, so that $N_s(S)$ is convex at the origin; and that when $X(0) \neq 0$, C_1 is negative and C_2 can be either positive or negative, depending on $X(f)$. For a spectrum like the one in Fig. 12b, this implies a functional discontinuity at $\beta_2 = 0$, which can be explained as follows: So long as $\beta_2 \neq 0$, the curvature of $N_s(S)$ is convex at the origin, but it becomes concave at some $S > 0$. As β_2 decreases, the curvature at the origin becomes sharper and the point of inflection occurs at successively lower values of S. Finally, in the limiting case

$\beta_2 = 0$, the point of inflection occurs at $S = 0$ and so $N_s(S)$ becomes concave at the origin. Thus, an abrupt increase in X at f close to 0, reflected in a large negative value of C_2, signifies a very small range of validity and calls for caution in using the new result for $N(S \approx 0)$.

For concreteness, suppose that the continuous part of $X(f)$ increases to a large peak density at some low frequency f_o, where $f_o \ll \sqrt{b_1}/2\pi\sigma$. To obtain a more useful approximation to $N(S \approx 0)$ under such circumstances, C_1 and C_2 should be recomputed as if the point of high density were at $f = 0$ instead of $f = f_o$, i.e., as if the spectrum were $X(f + f_o)$. The resulting $N_s(S)$, (44), then corresponds to a spectrum quite similar to $X(f)$ but without the abrupt change at low f. Consequently, it exhibits the correct "large-scale" curvature at low S while omitting the sharp "small-scale" curvature at S near 0 associated with the true C_1 and C_2. An empirically derived rule-of-thumb that leads to good results in all cases is the following: If $C_2 < -10$, shift the continuous part of $X(f)$ to the left until the nearest high-density peak is at $f = 0$, and recompute C_1 and C_2 for this modified spectrum. Then compute the quantity $C = (0.5\ C_2 - 10\ C_1)$ using these values. If C exceeds the original value of C_2, use the newly computed values of C_1 and C_2; otherwise use the original ones.

VI. GENERAL NOISE POWER FORMULA

We now seek an expression for slope overload noise power $[N_o(S)]$ which is accurate over all S for all spectra and is simple to use. Our approach is to derive separate functional descriptions for the regions $0 < S < 4.0$ and $S \geqq 4.0$. The descriptions are such that the resulting $N_o(S)$ and its first derivative are continuous at the boundary $S = 4.0$.

6.1 $N_o(S)$ *for* $S \geqq 4.0$

The result $N_l(S)$, (43), gives a very accurate approximation to $N(S)$ in the region $S \geqq 3.5$. However, our result for $N_{lb}(S)$, (22) through (25), is quite complicated, and the results for $A(S)$, $B(S)$, and $C(S)$ are available in graphical form only (Fig. 9). We now use these results to derive a simple but still accurate representation for this region.

Taking a suggestion from Rice's analysis,[2] we speculate that $N(S)$ for $S \geqq 4.0$ can be accurately approximated by

$$N(S \geqq 4.0) \doteq \frac{243}{4\sqrt{2\pi}} \frac{b_1^2}{b_2} \frac{\exp(-S^2/2)}{S^5} \cdot [\exp(-\alpha_1 \exp(-\alpha_2 S))] \triangleq N_o(S \geqq 4.0). \quad (62)$$

The quantity preceding the brackets is Rice's result $N_R(S)$, (9), while the bracketed term (with $\alpha_2 > 0$) is a monotonically decreasing function that converges to unity as $S \to \infty$. We choose α_1 and α_2 so that

$$N_o(4.0) = N_l(4.0) \text{ and } N_o'(4.0) = N_l'(4.0).$$

By using the data in Fig. 9 to obtain $A(S)$, $B(S)$, $C(S)$, and their first derivatives at $S = 4.0$, and the data in Fig. 6 to obtain $N_{lb}(S)$ and its first derivative at $S = 4.0$, we obtain

$$\alpha_2 = \frac{0.2141 - 0.024M_3 - 0.196M_4 - 0.067M_5}{0.6394 - 0.057M_3 - 0.426M_4 - 0.093M_5} \tag{63}$$

and

$$\alpha_1 = (0.6394 - 0.057M_3 - 0.426M_4 - 0.093M_5) \exp(4\alpha_2) \tag{64}$$

where M_3, M_4, M_5 are the moment excesses defined by (41). Note that when $X(f)$ is the two-band spectrum, ($M_n = 0$, $n \geqq 3$), α_1 and α_2 reduce to 2.45 and 0.336, respectively.

The above approximation to $N(S \geqq 4.0)$ can be tested by comparing it with the more precise results computed for various spectra using (35) and (37). The indications from such comparisons are that the approximation is accurate to within ± 0.3 dB over $S \geqq 4.0$ for all spectra.

6.2 $N_o(S)$ *for* $0 < S < 4.0$

The analyses of Sections IV and V inform us about $N(S)$ at the extremities of the region $0 < S < 4.0$, but not about the variation in between. To estimate this variation reliably, it is convenient to express $N(S)$ in the form

$$N(S) \equiv G(S)N_{lb}(S) \tag{65}$$

and then seek an approximation to $G(S)$. The latter is a spectrum-dependent function that differs from unity because–and to the extent that–$X(f)$ differs from a two-band spectrum having the moments b_0, b_1, and b_2 (Fig. 3). From physical reasoning given earlier, we expect that $G(S) \to 1$ as $S \to \infty$ and that, as S decreases from high values towards 0, $G(S)$ increases because–and to the extent that–b_3, b_4, etc., exceed the minimum values, (6), corresponding to the two-band spectrum. The maximum value of $G(S)$ should then be its value as $S \to 0$, which is $\sigma^2(b_2/b_1^2)$.

This reasoning is supported by careful scrutiny of the results for $N_l(S)$ and $N_s(S)$, Sections IV and V. From Section IV we can show

that $G(S)$ is a nonincreasing function* for $S \geqq 3.5$ and that it goes to unity as $S \to \infty$. From the results of Section V we can show that $G(S)$ is a nonincreasing function* at $S \approx 0$ and that it goes to $\sigma^2(b_2/b_1^2)$ as $S \to 0$. A logical consequence of these observations is that, for $0 < S < 3.5$, $G(S)$ is either a nonincreasing function, or has at least two extrema. The latter possibility has no physical basis in fact, and so we conclude that $G(S)$ is a nonincreasing function over all S, approaching a maximum value of $\sigma^2(b_2/b_1^2)$ as $S \to 0$ and a minimum value of 1 as $S \to \infty$.

Computations show that the decrease in $G(S)$ over $0 < S < 4.0$ is less than 10 dB for all spectra of possible practical interest. Thus if we can find a functional approximation to the quantity

$$g(S) \triangleq \ln G(S) = \ln \left[\frac{N(S)}{N_{tb}(S)}\right]; \qquad 0 < S < 4.0 \tag{66}$$

that is accurate to within ± 10 percent, the resulting approximation to $G(S)$ will be accurate to within ± 1 dB for all spectra of interest. This accuracy should be possible to achieve since $g(S)$ near $S = 0$ and $S = 4.0$ are known from the results of Sections III, IV, and V, and the above argument persuades us that $g(S)$ is nonincreasing over the region in between.

The function that we use to approximate $g(S)$ is

$$g_o(S) = \ln (\sigma^2 b_2/b_1^2) + a_1 S + a_2[\exp (a_3 S + a_4 S^2) - 1] \tag{67}$$

where $a_1, \cdots a_4$ are chosen to give $g_o'(0)$, $g_o''(0)$, $g_o(4)$, and $g_o'(4)$ the values predicted by the results of Sections III, IV, and V. These values are found by using (66), with $N_s(S)$ and $N_l(S)$ replacing $N(S)$ at $S = 0$ and $S = 4$, respectively. Thus, the following equations must be satisfied:

$$a_1 + a_2 a_3 = C_1; \tag{68}$$

$$a_2(a_3^2 + 2a_4) = 2C_2 - C_1^2 + \left(4 - \frac{\pi^2}{6}\right); \tag{69}$$

$$\ln (\sigma^2 b_2/b_1^2) + 4a_1 + a_2[\exp (4a_3 + 16a_4) - 1] = 0.057M_3 + 0.426M_4 + 0.093M_5; \tag{70}$$

and

$$a_1 + a_2(a_3 + 8a_4) \exp(4a_3 + 16a_4) = -0.024M_3 - 0.196M_4 - 0.067M_5 \tag{71}$$

* In fact, $G(S)$ is a *decreasing* function for all but the two-band spectrum, for which case it is precisely 1.

TABLE I—FORMULA COEFFICIENTS FOR THE BUTTERWORTH SPECTRUM

β_1	a_1	a_2	a_3	a_4
0.02	−0.177	2.26	−3.93	−8.76
0.068	−0.151	1.29	−3.70	−7.04
0.10	−0.137	1.06	−3.75	−6.99
0.25	−0.096	0.66	−3.90	−7.03
0.50	−0.066	0.50	−3.91	−6.67
1.0	−0.047	0.42	−3.86	−6.22
2.0	−0.040	0.38	−3.84	−6.00
∞	−0.036	0.37	−3.83	−5.90

where C_1, C_2 and $M_n (n \geqq 3)$ are given by (55), (60), and (41), respectively.

A set of solutions that is quite accurate in most cases is the following:

$$a_1 = -0.024M_3 - 0.196M_4 - 0.067M_5 \tag{72}$$

$$a_2 = \ln (\sigma^2 b_2 / b_1^2) - 0.155M_3 - 1.210M_4 - 0.361M_5 \tag{73}$$

$$a_3 = \begin{cases} (C_1 - a_1)/a_2; & a_2 > 0 \\ 0; & a_2 = 0 \\ |C_1 - a_1|/a_2; & a_2 < 0 \end{cases} \tag{74}$$

$$a_4 = \begin{cases} \dfrac{1}{2}\left[\dfrac{2C_2 - C_1^2 + \left(4 - \dfrac{\pi^2}{6}\right)}{a_2} - a_3^2\right]; & a_2 > 0. \\ 0; & a_2 \leqq 0 \end{cases} \tag{75}$$

The results for a_3 and a_4 when $a_2 \leqq 0$ are approximations only, but a_2 is nonpositive only when $X(f)$ is relatively narrowband, i.e., when $\sigma^2 b_2 / b_1^2 \approx 1$, in which case $G(S)$ is close to 1 over all S. Hence, these approximations lead to a quite accurate representation of $g(S)$. In the

TABLE II—FORMULA COEFFICIENTS FOR THE UNIFORM SPECTRUM

β_2	a_1	a_2	a_3	a_4
0	−0.036	0.37	−3.83	−5.90
0.05	−0.036	0.32	−4.44	−8.21
0.10	−0.036	0.27	0.14	−11.96
0.20	−0.035	0.17	0.21	−7.58
0.30	−0.033	0.08	0.41	−7.95
0.40	−0.029	0.02	1.44	−18.30
0.50	−0.024	−0.01	−1.76	0
0.80	−0.004	−0.01	−0.43	0
1.0	0	0	0	0

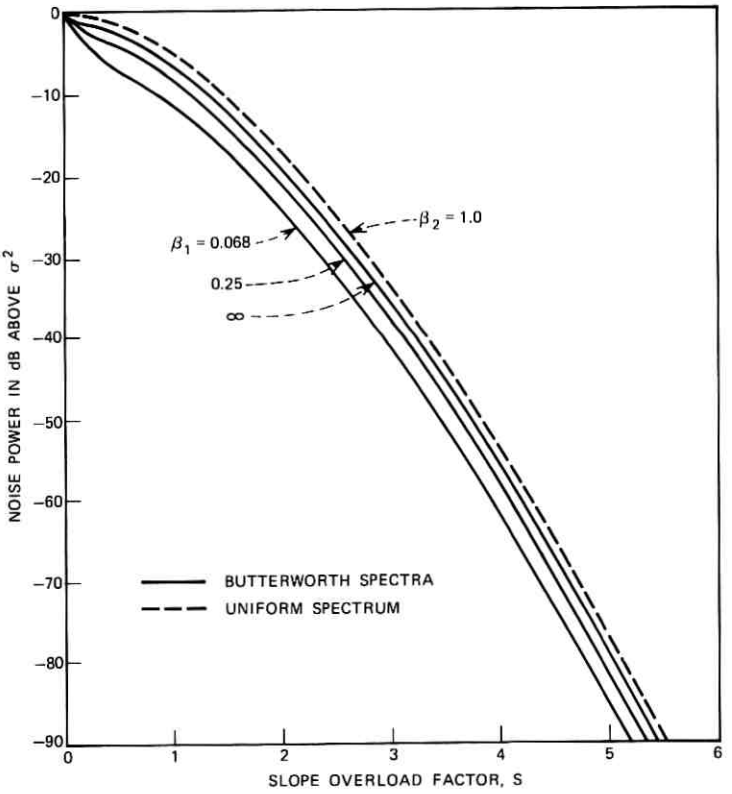


Fig. 13—Noise power results for several spectra.

more usual case where $a_2 > 0$, the above solutions are highly accurate so long as $4a_3 + 16a_4 \leqq -(4 + \ln a_2)$, a condition satisfied for most spectra. If this inequality is violated, solutions for $a_1, \cdots a_4$ can be found by graphical or computerized methods, but such situations appear to be rare.

With $g(S)$ approximated as above, $N_o(S)$ over $0 < S < 4.0$ can be expressed as the product $\exp(g_o(S))N_{tb}(S)$. To obtain a usable expression, however, $N_{tb}(S)$ should be given in a more simple form than eqs. (22) through (25). An approximation to $N_{tb}(S)$ which is accurate to within ± 0.3 dB is

$$N_{tb}(S) \doteq \frac{b_1^2}{b_2}[1 + 2.753S + 2.952S^2] \cdot \exp(-2.753S - 0.341S^2); \qquad 0 < S < 4.0. \quad (76)$$

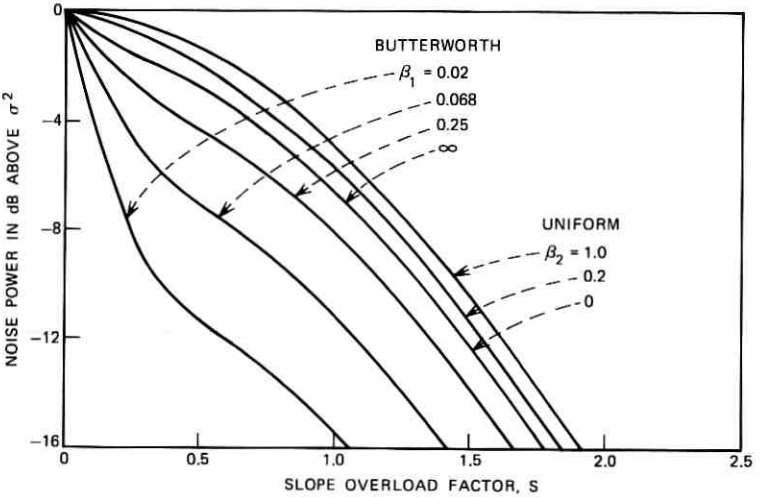


Fig. 14—Noise power results in greater detail.

Combining (76) with (67) leads to the following:

$$N_o(S) = \sigma^2[1 + 2.753S + 2.952S^2] \exp(-0.341S^2) \cdot \exp\{(a_1 - 2.753)S + a_2[\exp(a_3S + a_4S^2) - 1]\}; \quad 0 < S < 4.0. \quad (77)$$

6.3 *Final Expression*

A highly accurate two-region approximation to $N(S)$ is given by the combination of (62) and (77). We have determined, however, that (77) alone predicts $N(S)$ with an accuracy of 1 dB or better for S between 4.0 and 6.5 for all spectra. Moreover, we have determined that $N(S)$ is at least 119 dB below σ^2 for $S \geqq 6.5$ and for all spectra. Obviously, then, (77) can be used over all S for which $N(S)$ is not negligibly small. We therefore present (77) as our final expression for slope overload noise power, with $a_1, \cdots a_4$ defined by (68) through (71) and accurately approximated by (72) through (75) in virtually all cases.

VII. NUMERICAL RESULTS

The new result has been applied to the two families of spectra shown in Fig. 12. The Butterworth spectrum is characterized by an upper truncation frequency (W) and the ratio (β_1) of the 3-dB corner frequency to W. Note that, when $\beta_1 \rightarrow \infty$, the spectrum approaches that

TABLE III—N_P/N_o IN dB FOR THE BUTTERWORTH SPECTRUM

S \ γ	0.036 ($\beta_1 = 0.02$)	0.111 ($\beta_1 = 0.068$)	0.289 ($\beta_1 = 0.25$)	0.556 ($\beta_1 = \infty$)
1.0	+2.85	+1.76	+1.44	+1.70
2.0	+0.41	−0.20	−0.22	−0.10
3.0	+0.43	−0.01	−0.04	+0.22
4.0	+0.96	+0.74	+0.71	+0.81
5.0	+1.14	+1.12	+1.11	+1.13
6.0	+0.90	+0.85	+0.80	+0.70

of bandlimited white noise. The bandpass uniform spectrum is characterized by an upper truncation frequency (W) and the ratio (β_2) of the lower truncation frequency to W. Note that $\beta_2 = 0$ corresponds to bandlimited white noise in this case and that, as $\beta_2 \to 1$, the spectrum becomes infinitely narrowband.

For each of these two families, the formula coefficients, (a_1, a_2, a_3, a_4), have been computed as functions of the respective β-parameter. The results for the Butterworth spectrum are given by Table I and those for the uniform spectrum by Table II. The values shown are rounded to the decimal accuracy required.

Curves of N_o/σ^2 versus S for several Butterworth spectra of practical interest and for the narrowband spectrum are shown in Fig. 13. (The curve for the narrowband case represents an upper bound on $N(S)/\sigma^2$ for all spectra.) Also, an informative closeup that treats more spectra but over a reduced range of S is given by Fig. 14.

Since the new formula is highly accurate for all spectra, it can be used to estimate the accuracy of the Protonotarios formula, (10). Table III compares $N_P(S)$ with $N_o(S)$ for various Butterworth spectra spanning the range of $\gamma = b_1^2/b_2b_0$ of practical interest, while Table IV gives the comparison for various two-band spectra. The larger errors in Table IV must be regarded as untypical, since the two-band spectrum itself is rather artificial. For more typical spectra, we can

TABLE IV—N_P/N_o IN dB FOR THE TWO-BAND SPECTRUM

S \ γ	0.111	0.289	0.556	1
1.0	+5.04	+3.55	+2.49	+1.52
2.0	+2.45	+1.46	+0.75	+0.06
3.0	+1.97	+1.23	+0.70	+0.19
4.0	+2.05	+1.54	+1.12	+0.71
5.0	+1.79	+1.55	+1.30	+1.02
6.0	+0.84	+0.88	+0.74	+0.50

expect the errors in the Protonotarios formula for given S and γ to be more like the entries in Table III.

VIII. CONCLUSION

Aside from providing a new formula for slope overload noise power, this work has served to resolve the disparities between previously published formulas and can be used to identify the accuracies and limitations of those most widely used. Furthermore, the new result and the underlying methods of analysis can be extended to treat other important topics that have been mostly ignored before. Among these are the average duty cycle of slope overload, the slope overload noise spectrum, the effects of leaky feedback integrators on slope overload noise power, and the slope overload noise power for certain non-Gaussian processes.

IX. ACKNOWLEDGMENTS

The author is indebted to D. J. Goodman for his suggestions and encouragement throughout the execution of this work, and to D. L. Duttweiler for a number of helpful comments.

REFERENCES

1. Zetterberg, L. H., "A Comparison Between Delta and Pulse Code Modulation," Ericsson Technics, *2*, No. 1 (January 1955), pp. 95–154.
2. O'Neal, J. B., Jr., "Delta Modulation Quantizing Noise Analytical and Computer Simulation Results for Gaussian and Television Input Signals," B.S.T.J., *45*, No. 1 (January 1966), pp. 117–141.
3. Protonotarios, E. N., "Slope Overload Noise in Differential Pulse Code Modulation Systems," B.S.T.J., *46*, No. 9 (November 1967), pp. 2119–2162.
4. Abate, J. E., "Linear and Adaptive Delta Modulation," Proc. IEEE, *55* (March 1967), pp. 298–308.
5. van de Weg, H., "Quantizing Noise of a Single Integration Delta Modulation System With an N-Digit Code," Phillips Res. Rpt., *8* (1953), pp. 367–385.
6. Goodman, D. J., "Delta Modulation Granular Quantizing Noise," B.S.T.J., *48*, No. 5 (May-June 1969), pp. 1197–1218.
7. Iwersen, J. E., "Calculated Quantizing Noise of Single-Integration Delta-Modulation Coders," B.S.T.J., *48*, No. 7 (September 1969), pp. 2359–2389.
8. Rice, S. O., "Mathematical Analysis of Random Noise," B.S.T.J., *23*, No. 3 (July 1944), pp. 282–332, and *24*, No. 1 (January 1945), pp. 46–156.
9. Aaron, M. R., Fleischman, J. S., McDonald, R. A., and Protonotarios, E. N., "Response of Delta Modulation to Gaussian Signals," B.S.T.J., *48*, No. 5 (May-June 1969), pp. 1167–1195.
10. Greenstein, L. J., unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 3, March, 1973
Printed in U.S.A.

Tube Waveguide for Optical Transmission

By D. MARCUSE and W. L. MAMMEL

(Manuscript received September 25, 1972)

The dielectric optical waveguide described in this paper has an annular cross section, the refractive index of which is higher than the indices of the material inside and outside of the ring. The solution of the eigenvalue problem is an approximation which is valid for small refractive index differences of the three media. The resulting approximate eigenvalue equation is far simpler than its exact counterpart. The cutoff conditions of the first three modes and the eigenvalue of the lowest-order mode are presented graphically.

I. INTRODUCTION

Dielectric optical waveguides have become interesting since it has been demonstrated that the losses of dielectric materials can be quite low.[1–3] The conventional optical fiber waveguide consists of a solid dielectric core surrounded by a dielectric material with lower refractive index. This structure has the advantage of being particularly simple. In this paper we analyze a different type of dielectric optical waveguide. Our structure consists of a dielectric tube which is filled and surrounded by dielectric material with lower index of refraction. Such structures have been analyzed before.[4] A complete discussion of and references to the literature can be found in Ref. 5. However, the exact analytical treatment is extremely complicated so that the resulting theories are hard to apply to practical situations. Recently A. W. Snyder[6] has shown that the analysis of the conventional optical fiber waveguide becomes much simpler if one assumes that the difference of the refractive indices of core and cladding material is only slight. D. Gloge[7] has developed Snyder's theoretical work even further. On the basis of this simplified method of analysis, it is possible to treat rather complicated optical waveguide structures more simply and still obtain results that are applicable to most cases of practical interest. The simplification that results from the assumption of only slight index

differences is particularly appropriate since most practical dielectric optical waveguides satisfy this requirement.

An approximate eigenvalue equation for the modes of the tube waveguide and an equation describing the cutoff conditions for these modes are derived in this paper. The connection of the tube waveguide with the dielectric slab waveguide in the limit of large tube radii and with the solid-core optical fiber in the limit of small tube radii is discussed. The cutoff condition for the first three modes is represented graphically and the eigenvalue of the first mode is plotted.

For simplicity we assume throughout that the index of the material inside of the tube is smaller than or equal to the index of the material surrounding the tube.

II. APPROXIMATE MODE ANALYSIS OF THE TUBE WAVEGUIDE

The geometry of the tube waveguide is shown in Fig. 1. Following the ideas of Snyder and Gloge[6,7] we express the electric and magnetic field components in rectangular cartesian coordinates and write[8]

$$E_x = -\frac{i}{K_j^2}\left(\beta\frac{\partial E_z}{\partial x} + \omega\mu\frac{\partial H_z}{\partial y}\right) \tag{1}$$

$$E_y = -\frac{i}{K_j^2}\left(\beta\frac{\partial E_z}{\partial y} - \omega\mu\frac{\partial H_z}{\partial x}\right) \tag{2}$$

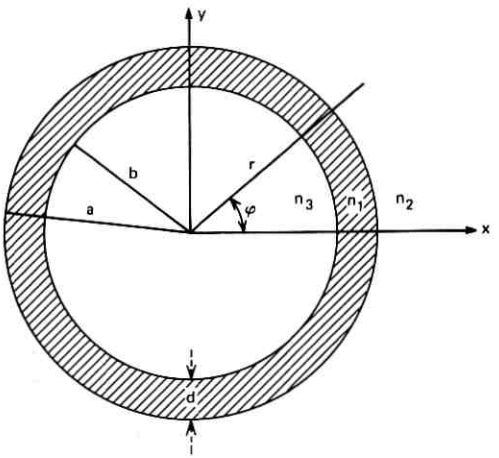


Fig. 1—Cross section of the tube waveguide.

$$H_x = -\frac{i}{K_j^2}\left(\beta\frac{\partial H_z}{\partial x} - \omega\epsilon\frac{\partial E_z}{\partial y}\right) \tag{3}$$

$$H_y = -\frac{i}{K_j^2}\left(\beta\frac{\partial H_z}{\partial y} + \omega\epsilon\frac{\partial E_z}{\partial x}\right). \tag{4}$$

It is assumed that the time and z dependence of the field are determined by the factor

$$e^{i(\omega t-\beta z)}. \tag{5}$$

β is the propagation constant in z direction, ω is the radian frequency. The parameter K_j is defined as

$$K_j^2 = n_j^2k^2 - \beta^2 \qquad j = 1, 2, \text{ or } 3 \tag{6}$$

with

$$k^2 = \omega^2\epsilon_o\mu_o. \tag{7}$$

Equations (1) through (4) are exact and are simply four of Maxwell's six equations written in a different form. If the refractive indices are all nearly the same, $n_1 \approx n_2 \approx n_3 \approx n$, we must have

$$\beta \approx nk, \tag{8}$$

so that

$$K_j \ll \beta. \tag{9}$$

It is thus apparent that the transverse field components are much larger than the longitudinal components if the refractive indices are all very nearly the same. The longitudinal field components are obtained as solutions of the wave equations.[8] We use the expressions

$$E_z = \frac{iA_jK_j}{2n_jk}\{Z_{\nu+1}(K_jr)\sin(\nu+1)\phi + Z_{\nu-1}(K_jr)\sin(\nu-1)\phi\} \tag{10}$$

$$H_z = -\sqrt{\frac{\epsilon_o}{\mu_o}}\frac{iA_jK_j}{2k}\{Z_{\nu+1}(K_jr)\cos(\nu+1)\phi - Z_{\nu-1}(K_jr)\cos(\nu-1)\phi\}. \tag{11}$$

The exponential factor (5) has been suppressed. $Z_\nu(K_jr)$ is a cylinder function and A_j is an arbitrary amplitude factor. Substitution of (10) and (11) into (1) through (4) results in

$$E_y = A_jZ_\nu(K_jr)\cos\nu\phi, \tag{12}$$

$$H_x = -n_jA_j\sqrt{\frac{\epsilon_o}{\mu_o}}Z_\nu(K_jr)\cos\nu\phi, \tag{13}$$

and

$$E_x = 0, \qquad H_y = 0. \tag{14}$$

These equations are approximations that hold provided that (8) is applicable.

In the three regions shown in Fig. 1, we use the following cylinder functions and parameters, assuming $n_1 > n_2 > n_3$,

$$\left.\begin{aligned} -iK_3 &= \theta = (\beta^2 - n_3^2k^2)^{\frac{1}{2}} \\ Z_\nu &= J_\nu(i\theta r) \end{aligned}\right\} \quad \text{for} \quad 0 \leqq r \leqq b \tag{15}$$

$$\left.\begin{aligned} K_1 &= \kappa = (n_1^2k^2 - \beta^2)^{\frac{1}{2}} \\ Z_\nu &= J_\nu(\kappa r) + BN_\nu(\kappa r) \end{aligned}\right\} \quad \text{for} \quad b \leqq r \leqq a \tag{16}$$

$$\left.\begin{aligned} -iK_2 &= \gamma = (\beta^2 - n_2^2k^2)^{\frac{1}{2}} \\ Z_\nu &= H_\nu^{(1)}(i\gamma r) \end{aligned}\right\} \quad \text{for} \quad a \leqq r \leqq \infty. \tag{17}$$

J_ν, N_ν are the Bessel and Neumann functions and $H_\nu^{(1)}$ is the Hankel function of the first kind. We use the functions with imaginary argu-

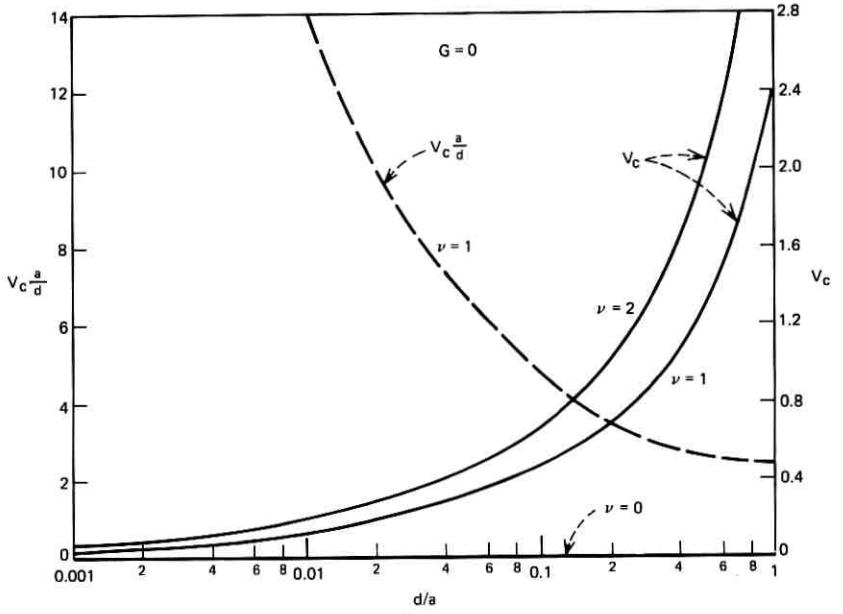


Fig. 2—The cutoff value of the normalized frequency V for $G = 0$ as a function of the ratio of wall thickness d to tube (outer) radius a. The cutoff value for the mode $\nu = 0$ is $V = 0$.

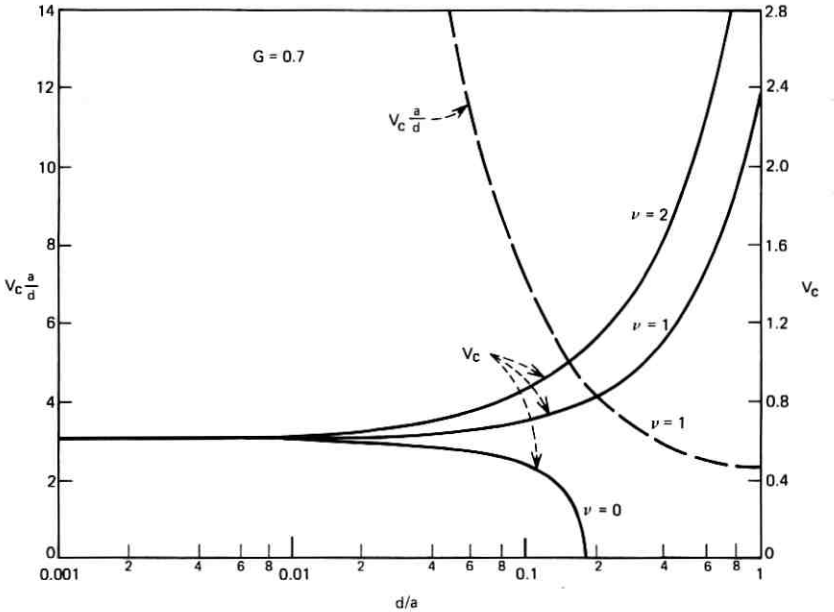


Fig. 3—The cutoff value of V as a function of d/a for $G = 0.7$.

ments instead of modified functions. This practice is in agreement with Jahnke and Emde[9] as well as with Gradshteyn and Ryzhik.[10] The amplitude coefficient A_j appearing in (10) through (13) also assumes different values in the three regions of space, $j = 1, 2, 3$.

In order to be able to match the boundary conditions, we transform the cartesian components to vector components in cylindrical polar coordinates

$$E_\phi = -E_x \sin\phi + E_y \cos\phi$$
$$= \tfrac{1}{2}A_j Z_\nu(K_j r)[\cos(\nu + 1)\phi + \cos(\nu - 1)\phi] \quad (18)$$

$$H_\phi = \tfrac{1}{2}n_j A_j \sqrt{\frac{\epsilon_o}{\mu_o}} Z_\nu(K_j r)[\sin(\nu + 1)\phi - \sin(\nu - 1)\phi]. \quad (19)$$

The boundary conditions require that E_z, E_ϕ, H_z, and H_ϕ are continuous at $r = b$ and at $r = a$. The boundary conditions thus provide us with eight equations. However, our approximate treatment of the modes yields only four arbitrary constants, A_1, A_2, A_3, and B. The situation is further aggravated if we observe that the field components contain the functions $\sin(\nu + 1)\phi$ and $\sin(\nu - 1)\phi$ and corresponding

expressions with the cosine function. Since the boundary conditions must hold for all values of ϕ, we must equate the coefficients of the functions with $\nu + 1$ independently of those with $\nu - 1$, thus doubling the number of equations. In spite of this apparently hopeless situation, approximate solutions are possible. The equations that result from the requirement that the tangential components of **E** are continuous at the boundaries differ from the corresponding equations resulting from the boundary conditions for **H** only by factors n_j. Since our approach is based on assuming that all n_j are nearly the same, we set these factors all equal to the same average value n and, after dividing by n, obtain equations that duplicate those obtained from the boundary conditions for **E**. To the approximation considered here, the **E** and **H** boundary conditions lead to the same set of equations, reducing their total number to one half of the original number. To satisfy the continuity of E_z and E_ϕ at $r = b$, we equate the coefficients of sin $(\nu + 1)\phi$ and obtain

$$A_3 = \frac{J_\nu(\kappa b) + BN_\nu(\kappa b)}{J_\nu(i\theta b)} A_1 \tag{20}$$

and the determinantal condition

$$i\theta \frac{J_{\nu+1}(i\theta b)}{J_\nu(i\theta b)} = \kappa \frac{J_{\nu+1}(\kappa b) + BN_{\nu+1}(\kappa b)}{J_\nu(\kappa b) + BN_\nu(\kappa b)}. \tag{21}$$

Equating the coefficients of sin $(\nu - 1)\phi$, we obtain a similar result with the only difference that $\nu + 1$ is replaced by $\nu - 1$ in (21). Equation (20) remains the same. With the help of the recursion relations[9,10] for cylinder functions, it is easy to prove that eq. (21) is not changed if $\nu + 1$ is replaced by $\nu - 1$. We thus find ourselves in the happy position of being able to satisfy–at least approximately–the eight boundary conditions at $r = b$ with the two available coefficients. In an analogous fashion, we obtain from the boundary conditions at $r = a$ the equation

$$A_2 = \frac{J_\nu(\kappa a) + BN_\nu(\kappa a)}{H_\nu^{(1)}(i\gamma a)} A_1 \tag{22}$$

and the determinantal condition

$$i\gamma \frac{H_{\nu+1}^{(1)}(i\gamma a)}{H_\nu^{(1)}(i\gamma a)} = \kappa \frac{J_{\nu+1}(\kappa a) + BN_{\nu+1}(\kappa a)}{J_\nu(\kappa a) + BN_\nu(\kappa a)}. \tag{23}$$

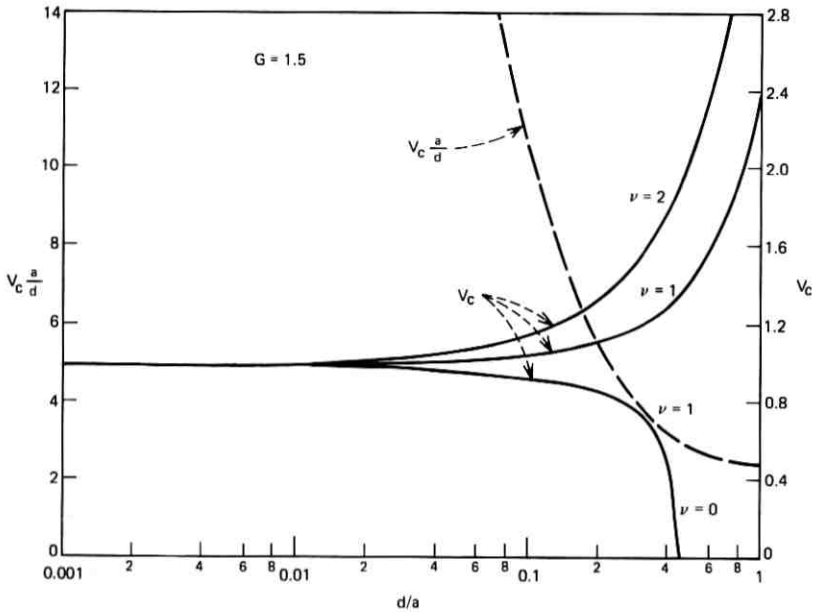


Fig. 4—Same as Fig. 3, $G = 1.5$.

The eigenvalue equation is obtained by eliminating B from (21) and (23). We find

$$\frac{\kappa J_\nu(i\theta b)J_{\nu+1}(\kappa b) - i\theta J_\nu(\kappa b)J_{\nu+1}(i\theta b)}{\kappa J_\nu(i\theta b)N_{\nu+1}(\kappa b) - i\theta N_\nu(\kappa b)J_{\nu+1}(i\theta b)} = \frac{\kappa H_\nu^{(1)}(i\gamma a)J_{\nu+1}(\kappa a) - i\gamma J_\nu(\kappa a)H_{\nu+1}^{(1)}(i\gamma a)}{\kappa H_\nu^{(1)}(i\gamma a)N_{\nu+1}(\kappa a) - i\gamma N_\nu(\kappa a)H_{\nu+1}^{(1)}(i\gamma a)}. \quad (24)$$

The constant B is given by

$$B = -\frac{\kappa J_\nu(i\theta b)J_{\nu+1}(\kappa b) - i\theta J_\nu(\kappa b)J_{\nu+1}(i\theta b)}{\kappa J_\nu(i\theta b)N_{\nu+1}(\kappa b) - i\theta N_\nu(\kappa b)J_{\nu+1}(i\theta b)}. \quad (25)$$

It can be shown that the eigenvalue equation (24) specializes to the eigenvalue equation ($d = a - b$)

$$\tan \kappa d = \frac{\kappa(\gamma + \theta)}{\kappa^2 - \gamma\theta} \quad (26)$$

of the asymmetric slab waveguide in the limit $a \to \infty$ and $b \to \infty$. The

distinction between TE and TM modes of the slab waveguide is lost in our approximation.

We are mainly interested in the cutoff conditions of the modes in order to find the separation between the lowest and the next higher mode which determines the range of single-mode operation. Cutoff is defined by the condition

$$\gamma = 0. \tag{27}$$

In order to determine the cutoff frequency, we use (27) in (24) and obtain the cutoff equation

$$\frac{J_\nu\left(iGV\frac{b}{d}\right)J_{\nu+1}\left(V\frac{b}{d}\right) - iGJ_\nu\left(V\frac{b}{d}\right)J_{\nu+1}\left(iGV\frac{b}{d}\right)}{J_\nu\left(iGV\frac{b}{d}\right)N_{\nu+1}\left(V\frac{b}{d}\right) - iGN_\nu\left(V\frac{b}{d}\right)J_{\nu+1}\left(iGV\frac{b}{d}\right)} = \frac{J_{\nu-1}\left(V\frac{a}{d}\right)}{N_{\nu-1}\left(V\frac{a}{d}\right)}. \tag{28}$$

The parameters appearing in this equation are defined as follows:

$$V = (n_1^2 - n_2^2)^{\frac{1}{2}}kd \tag{29}$$

and

$$G = \left(\frac{n_2^2 - n_3^2}{n_1^2 - n_2^2}\right)^{\frac{1}{2}}. \tag{30}$$

For future reference, we derive an interesting relation for the mode with $\nu = 0$. The cutoff condition (28) allows the solution $V = 0$ for $\nu = 0$. We find as the condition that $V = 0$, the relation

$$\frac{d}{a} = 1 - \left(\frac{n_1^2 - n_2^2}{n_1^2 - n_3^2}\right)^{\frac{1}{2}} = 1 - \frac{1}{(1 + G^2)^{\frac{1}{2}}}. \tag{31}$$

In the limit of infinite radii, $a \to \infty$ and $b \to \infty$, we obtain from (26) the cutoff condition for $n_1 > n_2 \geqq n_3$

$$V = \arctan G. \tag{32}$$

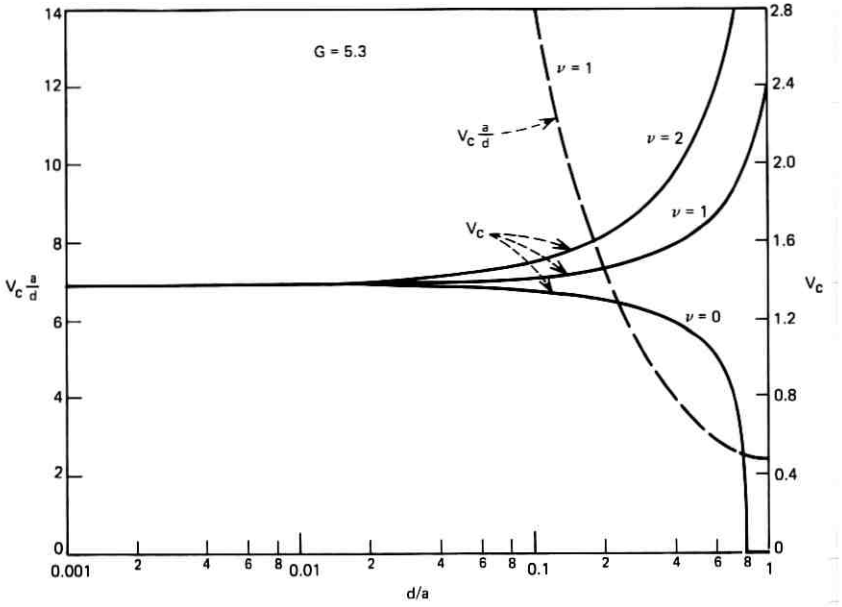


Fig. 5—Same as Fig. 3, $G = 5.3$.

III. DISCUSSION AND NUMERICAL RESULTS

We restrict our discussion to the case $n_2 \geqq n_3$. The cutoff values of V [see eq. (29)] as functions of d/a are plotted in Figs. 2 through 6 as solid lines for the first three modes, $\nu = 0$, $\nu = 1$, $\nu = 2$. The V parameter characterizes the tube waveguide in the limit of large radii where it can be regarded as an asymmetric slab waveguide that is bent into a circle. For large values of d/a, the waveguide approaches an optical fiber with solid cylindrical core. Such a fiber is characterized by a different V parameter which can be expressed in our notation as $(a/d)V$. In the limit $d/a = 1$, both parameters coincide. The parameter $(a/d)V$ is plotted in the figures as a dotted line only for the mode $\nu = 1$. Each curve, Figs. 2 through 6, is plotted for a different value of G [see eq. (30)]. The relation between the G values and the ratios of n_1/n_2 and n_1/n_3 is shown in Fig. 7.

Since the tube waveguide approaches the asymmetric slab in the limit $d/a = 0$ and the solid-core fiber in the limit $d/a = 1$, the cutoff values $V = V_c$ can be predicted at these limits. For $d/a = 1$, the cutoff value of the solid-core fiber is obtained from

$$J_{\nu-1}(V_c) = 0. \tag{33}$$

The values of V_c at $d/a = 1$ are, therefore, $V_c = 0$ for $\nu = 0$, $V_c = 2.405$ for $\nu = 1$, and $V_c = 3.83$ for $\nu = 2$.

At $d/a = 0$, we find the cutoff values for V from the asymmetric slab formula (32).

We see from Figs. 3 through 6 that the curve with $\nu = 0$ seems to end on the horizontal axis. Actually, the curve must be continued along the horizontal axis at $V = 0$ to the point $d/a = 1$. However, to the left of the point where the $\nu = 0$ curve reaches the horizontal axis, $V = 0$ is not a legitimate solution of the cutoff equation. The d/a value at which the $\nu = 0$ curve touches the horizontal axis is given by (31). This phenomenon finds the following simple explanation (private communication from E. A. J. Marcatili). As the mode approaches its cutoff value, we have $\gamma = 0$, and also $\theta = 0$, and $V = 0$. There is no longer any transverse field variation so that the field "senses" only the average value $\bar{n}^2$ of the dielectric constant of the tube,

$$\bar{n}^2 = \frac{\pi b^2 n_3^2 + \pi(a^2 - b^2)n_1^2}{\pi a^2}. \tag{34}$$

The mode is no longer guided if the average dielectric constant $\bar{n}^2$ is

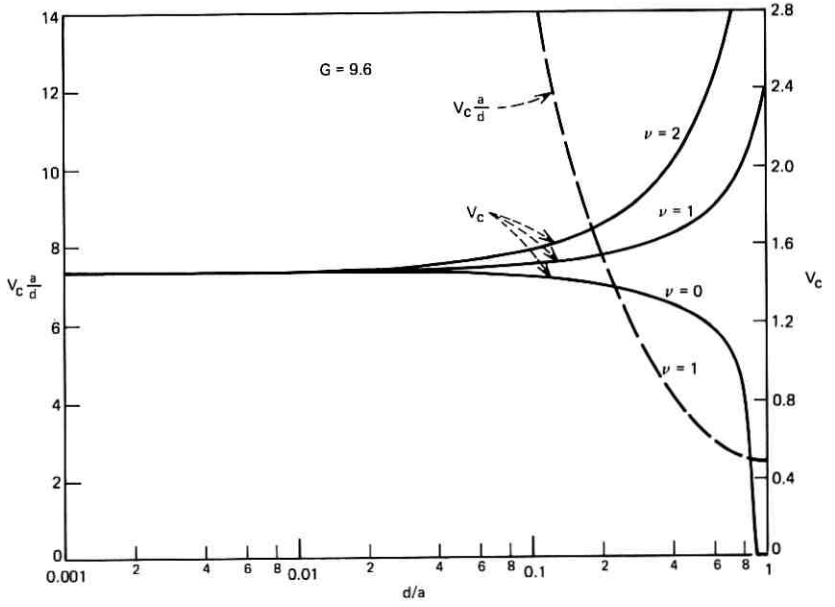


Fig. 6—Same as Fig. 3, $G = 9.6$.

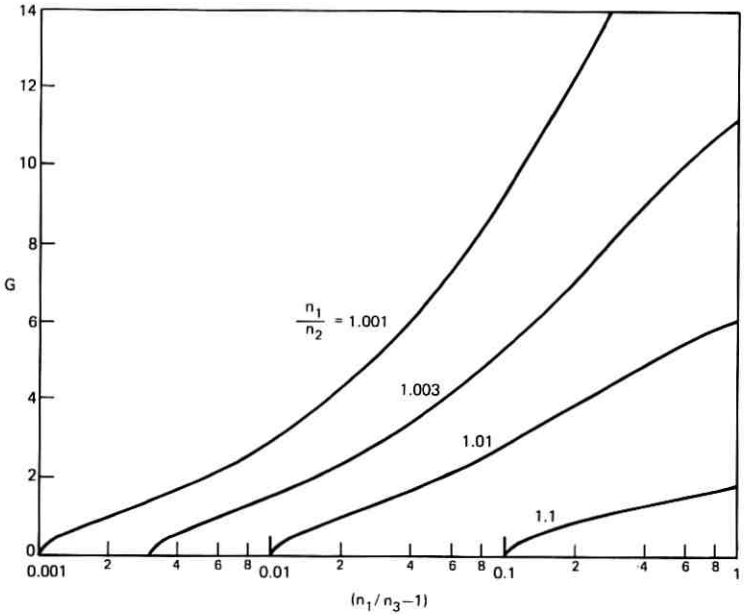


Fig. 7—The parameter G is shown as a function of $(n_1/n_3 - 1)$ for several values of n_1/n_2.

equal to the dielectric constants n_2^2 of the outer medium. The condition $\bar{n}^2 = n_2^2$ also leads to (31).

Figures 2 through 6 show that the cutoff points for all three modes ($\nu = 0$, 1, and 2) approach each other arbitrarily closely for small values of d/a. It is thus apparent that single-mode operation in the mode $\nu = 0$ becomes increasingly more difficult to achieve as the ratio d/a is decreased.

Finally, Fig. 8 shows a plot of the transverse decay parameter γa as a function of d/a. The curve is intended to provide information about single-mode operation of the tube waveguide. However, since it is desirable to operate with a tightly guided mode field, we chose as the operating point for V the cutoff value of the second mode, $V = V_{c1}$. It is a remarkable feature of the curves of Fig. 8 that they become independent of d/a for small values of this ratio. However, from the point of view of field confinement, we must remember that it is not γa but rather γd that characterizes the rate at which the field decays from the guiding structure–the tube–in radial direction. The crosses on the curves in Fig. 8 indicate the points at which we have $\gamma d = 1$, the circles indicate the points $\gamma d = 0.5$.

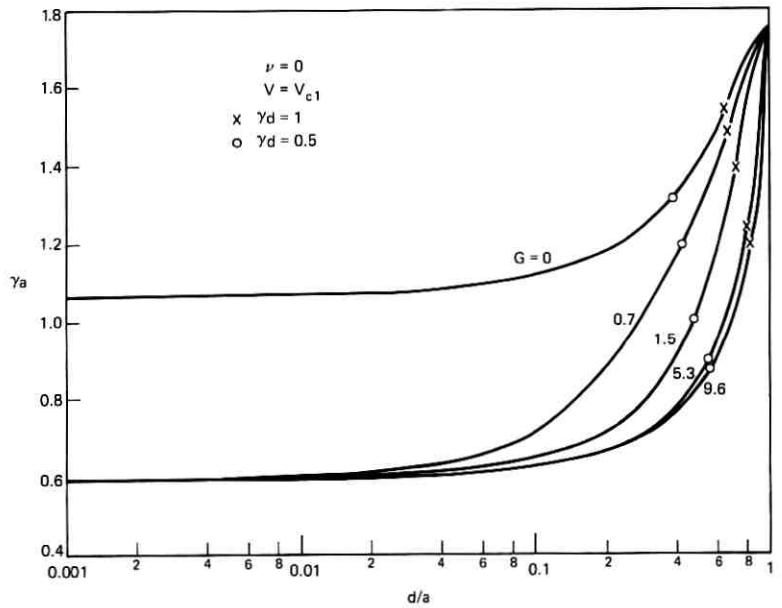


Fig. 8—This figure shows the parameter γa as a function of d/a for several values of G. V is taken equal to the cutoff value V_{c1} of the mode $\nu = 1$.

IV. CONCLUSIONS

The eigenvalue equation for the modes of the tube waveguide has been derived with the help of an approximate technique. The cutoff values of the normalized frequency parameter V [defined by (29)] are presented for the first three modes of the tube waveguide and the transverse decay parameter γa is plotted for the lowest-order mode under the assumption that this mode is operated at the point of cutoff of the next-higher-order mode.

One may wonder if it is possible to use the tube waveguide with single-mode operation to increase the mode radius compared to the mode radius of the lowest-order HE_{11} mode of the conventional fiber. This question has been studied under the assumption that both waveguides provide an equal amount of field confinement. The criterion for field confinement is the radial decay parameter γ. It was consequently assumed that the γ values of both waveguides are identical. Surprisingly, it was found that under the condition of single-mode operation for each guide and the requirement of equal field confinement, the mode radius of the tube waveguide is only slightly (approximately 40 percent) larger than the radius of the HE_{11} mode of the optical

fiber. A large mode radius appears desirable from the point of view of alignment tolerances of fiber splices. With a large mode radius, the alignment of the waveguide centers would be less critical. The tube waveguide does not appear to offer a significant advantage for this purpose.

REFERENCES

1. Kapron, F. P., Keck, D. B., and Maurer, R. D., "Radiation Losses in Glass Optical Waveguides," Appl. Phys. Ltrs., *17*, No. 10 (November 15, 1970), pp. 423–425.
2. Stone, J., "Optical Transmission Loss in Liquid-Core, Hollow Fibers," Conference on Integrated Optics-Guided Waves, Materials, and Devices, Optical Society of America, Digest of Technical Papers, pp. WA5-1–WA5-4.
3. Stone, J., "Optical Transmission in Liquid-Core Quartz Fibers," Appl. Phys. Ltrs., *20*, No. 7 (April 1972), pp. 239–240.
4. Beam, R. E., Astrahan, M. M., Jakes, W. C., Wachowski, H. M., and Firestone, W. L., "Dielectric Tube Waveguides," Final Report on Investigation of Multi-Mode Propagation in Waveguides and Microwave Optics, Microwave Laboratory, Northwestern University, Evanston, Illinois, May 1, 1949–November 30, 1950.
5. Kharadly, M. M. Z., and Lewis, J. E., "Properties of Dielectric-Tube Waveguides," Proc. IEEE, *116*, No. 2 (February 1969), pp. 214–224.
6. Snyder, A. W., "Asymptotic Expressions for Eigenfunctions and Eigenvalues of a Dielectric or Optical Waveguide," IEEE Trans. Microwave Theory Techniques, *MTT-17*, No. 12 (December 1969), pp. 1130–1138.
7. Gloge, D., "Weakly Guiding Fibers," Appl. Opt., *10*, No. 10 (October 1971), pp. 2252–2258.
8. Marcuse, D., *Light Transmission Optics*, New York: Van Nostrand Reinhold Company, 1972.
9. Jahnke, E., and Emde, F., *Tables of Functions with Formulas and Curves*, 4th Edition, New York: Dover Publications, 1945.
10. Grandshteyn, I. S., and Ryzhik, I. M., *Tables of Integrals, Series and Products*, 4th Edition, New York: Academic Press, 1965.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 3, March, 1973
Printed in U.S.A.

The Interrupted Poisson Process As An Overflow Process

By ANATOL KUCZURA

(Manuscript received September 18, 1972)

Traffic overflowing a first-choice trunk group can be approximated accurately by a simple renewal process called an interrupted Poisson process–a Poisson process which is alternately turned on for an exponentially distributed time and then turned off for another (independent) exponentially distributed time. The approximation is obtained by matching either the first two or three moments of an interrupted Poisson process to those of an overflow process. Numerical investigation of errors in the approximation and subsequent experience has shown that this method of generating overflow traffic is accurate and very useful in both simulations and analyses of traffic systems.

I. INTRODUCTION

A Poisson process which is alternately turned on for an exponentially distributed time and then turned off for another (independent) exponentially distributed time will be called an *interrupted* Poisson process–it can be viewed as a Poisson process modulated by a random switch. It was suggested by W. S. Hayward of Bell Laboratories that such a process be used to simulate overflow traffic. We will show that the interrupted Poisson process provides a simple and accurate method of simulating overflow traffic.

The objective is to reduce the cost of computer simulations of traffic systems by using the interrupted Poisson process to model the overflow traffic. Generation of actual overflow traffic by simulating the behavior of the trunk group from which it overflows is time consuming since a record must be kept of *all* calls which are offered to the subtending trunk group, whether they contribute to the overflow traffic or not. This is especially true when the traffic is overflowing a large trunk group. Moreover, the interrupted Poisson process provides a simple, approximate description of the overflow traffic and consequently facilitates analytical studies.

This method of generating overflow traffic has been used successfully in many studies of traffic systems. Examples of its application can be found in Refs. 1, 2, and 3 and also in numerous unpublished works. Since the method has been found to be very useful and requests for wider dissemination have been received by the author, this paper has been prepared.

In Section II we derive the distribution of the number of busy trunks in an infinite trunk group when the offered traffic is generated by an interrupted Poisson process. The corresponding distribution for an overflow input has been computed by Kosten.[4] Now, matching the first three moments of the two distributions, we obtain equations for the parameters of the interrupted Poisson process. We also give these equations for a two-moment match. The errors committed in approximating the distribution given by Kosten are also examined. In Section III we derive the interarrival time distribution for a traffic stream generated by an interrupted Poisson process.

II. MOMENT-MATCH EQUATIONS

Let the interrupted Poisson traffic be offered to an infinite trunk group. Let λ be the intensity of the Poisson process, $1/\gamma$ be the mean on-time of the random switch, $1/\omega$ be the mean off-time, and $1/\mu$ be the mean service time. Let the state of the system be described by (m, n) with state probabilities $p(m, n)$ where m is the number of servers busy, and n is the state of the switch taking on the value of 1 or 0 according to whether the process is on or off.

The equilibrium equations for the stationary state probabilities are

$$\begin{aligned}(m\mu + \omega)p(m, 0) &= \gamma p(m, 1) + (m + 1)\mu p(m + 1, 0), \qquad m \geqq 0,\\ (m\mu + \gamma + \lambda)p(m, 1) &= \omega p(m, 0) + (m + 1)\mu p(m + 1, 1)\\ &\qquad + \lambda p(m - 1, 1), \qquad m \geqq 1, \qquad (1)\\ (\gamma + \lambda)p(0, 1) &= \omega p(0, 0) + \mu p(1, 1).\end{aligned}$$

To solve this system of equations we introduce the probability generating function

$$G(z) = \sum p(m)z^m = G_1(z) + G_0(z),$$

where

$$G_1(z) = \sum p(m, 1)z^m, \qquad G_0(z) = \sum p(m, 0)z^m,$$

and

$$p(m) = p(m, 1) + p(m, 0).$$

Note that $G_1(1) = \sum p(m, 1)$ is simply the probability of the switch

being on and is given by $\omega/(\gamma + \omega)$. Similarly, $G_0(1) = \sum p(m, 0) = \gamma/(\gamma + \omega)$ is the probability of the switch being off. We will now derive the differential equations for G_1 and G_0 and obtain their solutions.

From (1) and the definition of $G(z)$ we have

$$\mu(z - 1)G_0'(z) + \omega G_0(z) - \gamma G_1(z) = 0,$$

$$\mu(z - 1)G_1'(z) + (\gamma + \lambda - \lambda z)G_1(z) - \omega G_0(z) = 0.$$

This system of equations is coupled. At the price of increasing the order, we can decouple the system by means of differentiation and simple substitution:

$$\mu(z - 1)G_0''(z) + [\mu + \gamma + \omega - \lambda(z - 1)]G_0'(z) - \frac{\lambda}{\mu}\omega G_0(z) = 0, \quad (2)$$

$$\mu(z - 1)G_1''(z) + [\mu + \gamma + \omega - \lambda(z - 1)]G_1'(z) - \frac{\lambda}{\mu}(\omega + \mu)G_1(z) = 0.$$

Since we are interested in the moments of the distribution of the number of busy servers, it will be convenient to obtain solutions of (2) valid about $z = 1$. Subsequently, the series can be rearranged at the origin to yield the state probabilities.

The change of variable

$$\xi = \frac{\lambda}{\mu}(z - 1)$$

transforms (2) into

$$\begin{aligned} \xi Q_0''(\xi) + (\epsilon - \xi)Q_0'(\xi) - \beta Q_0(\xi) &= 0, \\ \xi Q_1''(\xi) + (\epsilon - \xi)Q_1'(\xi) - (1 + \beta)Q_1(\xi) &= 0, \end{aligned} \quad (3)$$

where

$$\epsilon = 1 + \frac{\gamma + \omega}{\mu}, \qquad \beta = \frac{\omega}{\mu},$$

and

$$Q_1(\xi) = G_1(z), \qquad Q_2(\xi) = G_2(z).$$

Equations (3) have the solutions

$$\begin{aligned} Q_0(\xi) &= C_1 F_1(\beta; \epsilon; \xi), \\ Q_1(\xi) &= D_1 F_1(1 + \beta; \epsilon; \xi), \end{aligned}$$

where

$$ {}_1F_1(a; b; c) = \sum_{n=0}^{\infty} \frac{(a)_n c^n}{(b)_n n!} $$

with

$$ (a)_n = a(a+1)(a+2)\cdots(a+n-1), \qquad (a)_0 = 1, $$

is the confluent hypergeometric function and C and D are arbitrary constants. The conditions $G_1(1) = \omega/(\gamma+\omega)$ and $G_0(1) = \gamma/(\gamma+\omega)$ determine C and D and hence

$$ G(z) = \frac{\gamma}{\gamma+\omega} {}_1F_1\left[\beta; \epsilon; \frac{\lambda}{\mu}(z-1)\right] + \frac{\omega}{\gamma+\omega} {}_1F_1\left[1+\beta; \epsilon; \frac{\lambda}{\mu}(z-1)\right]. \quad (4) $$

The power series in $(z-1)$ on the right-hand side of (4) is convergent since ${}_1F_1$ is an entire function. The factorial moments of the distribution of the number of busy servers, being the coefficients of the expansion (4), are simply

$$ G^{(n)}(1) = \lambda^n \frac{(\omega)_n}{(\gamma+\omega)_n}, \qquad n = 0, 1, 2, \cdots, \quad (5) $$

where we have made the normalization $\mu = 1$. For computational purposes, the following recurrence relation is useful:

$$ G^{(n+1)}(1) = \lambda \frac{(\omega+n)}{(\gamma+\omega+n)} G^{(n)}(1). \quad (6) $$

To obtain the state probabilities $p(m) = p(m, 0) + p(m, 1)$ we rearrange the series (4) at the origin and identify the coefficients in this new expansion as the state probabilities

$$ p(m) = \frac{1}{m!} \sum_{k=0}^{\infty} G^{(m+k)}(1) \frac{(-1)^k}{k} = \frac{\lambda^m}{m!} \sum_{j=m}^{\infty} \frac{(-\lambda)^{j-m}}{(j-m)!} \frac{(\omega)_j}{(\gamma+\omega)_j}. \quad (7) $$

We now show that the interrupted Poisson process provides an accurate method of generating overflow traffic. Let a erlangs of Poisson traffic be offered to an Erlang B system of c trunks. Let the overflow traffic be routed to an infinite trunk group and let Y be the number of busy servers in the infinite trunk group under statistical equilibrium. The factorial moments $M_{(n)}$ of Y and the state probabilities

$f(m) = p[Y = m]$ have been computed by Kosten:[4]

$$M_{(n)} = a^n \frac{\sigma_0(c)}{\sigma_n(c)},$$
$$f(m) = \frac{a^{c+m}}{c!m!} \sum_{k=0}^{\infty} \frac{(-a)^k}{k!\sigma_{k+m}(c)}, \tag{8}$$

where

$$\sigma_0(c) = \frac{a^c}{c!}, \qquad \sigma_j(c) = \sum_{i=0}^{c} \binom{j+i-1}{i} \frac{a^{c-i}}{(c-i)}, \qquad j = 1, 2, \cdots.$$

For a more accessible reference which gives the derivation of the factorial moments, see the appendix prepared by J. Riordan in Ref. 5.

Now consider an interrupted Poisson traffic of original intensity λ offered to an infinite trunk group, and let X be the number of busy servers under statistical equilibrium. The factorial moments of X and the state probabilities are given by (5) and (7) respectively. In this system we have three parameters, λ, γ, and ω, which are to be chosen so that the interrupted Poisson process gives the best approximation to the overflow process. In the present analysis we choose the moment approximation and, in particular, take the factorial moments. Thus we require that

$$G^{(n)}(1) = M_{(n)}, \qquad n = 1, 2, 3, \tag{9}$$

and define the error

$$E_n(a, c) = f(n) - p(n), \qquad n = 0, 1, 2, \cdots. \tag{10}$$

With the aid of the recurrence relation (6), we can express (9) as

$$\lambda\left(\frac{\omega + n}{\omega + \gamma + n}\right) = a\delta_n, \qquad n = 0, 1, 2, \tag{11}$$

where

$$\delta_n = \frac{M_{(n+1)}}{aM_{(n)}} = \frac{\sigma_n(c)}{\sigma_{n+1}(c)}.$$

Equations (11) have the solution

$$\lambda = a\frac{\delta_2(\delta_1 - \delta_0) - \delta_0(\delta_2 - \delta_1)}{(\delta_1 - \delta_0) - (\delta_2 - \delta_1)},$$
$$\omega = \frac{\delta_0}{\lambda}\left(\frac{\lambda - a\delta_1}{\delta_1 - \delta_0}\right), \tag{12}$$
$$\gamma = \frac{\omega}{a}\left(\frac{\lambda - a\delta_0}{\delta_0}\right).$$

The parameters λ, ω, and γ can be computed from (12) whenever a and c are specified. Often, however, one is concerned with final-route traffic in which only the mean, α, and variance, v, (or peakedness ratio $z = v/\alpha$) are known. There are two ways to proceed in this case.

First, using Wilkinson's equivalent random method,[5] determine S, the number of trunks, and A, the equivalent random load corresponding to the overflow traffic of mean α and variance v. Now set $a = A$ and $c = S$ and use eqs. (12) to compute λ, ω, and γ for a three-moment match.

A second way to proceed is to determine the equivalent random load A as before, and set $\lambda = A$ in the last two equations of (12) to compute ω and γ for a two-moment match. A satisfactory value of the equivalent random load A is given by Rapp's approximation:[6]

$$A = \alpha z + 3z(z - 1), \tag{13}$$

where z is the peakedness ratio v/α. In terms of α, z, and A, these equations can be written as

$$\begin{aligned} \omega &= \frac{\alpha}{A}\left(\frac{A - \alpha}{z - 1} - 1\right), \\ \gamma &= \left(\frac{A}{\alpha} - 1\right)\omega. \end{aligned} \tag{14}$$

Note that for a two-moment match it is necessary to fix one of the parameters λ, ω, or γ. However, for a positive solution, they cannot be chosen arbitrarily–for positivity we must choose $\lambda > \alpha$, such as in (13).

To illustrate the procedure, let us take an example with overflow traffic of mean 0.61 and variance 0.95. By Wilkinson's equivalent random method, we obtain $S = 4$ trunks and $A = 3$ erlangs. The first three moments from (8) with $c = S = 4$ and $a = A = 3$ are

$$\begin{aligned} M_{(1)} &= 0.618 \\ M_{(2)} &= 0.708 \\ M_{(3)} &= 1.025. \end{aligned} \tag{15}$$

The three-moment match yields an interrupted Poisson process with parameters

$$\begin{aligned} \lambda &= 2.553 \\ \omega &= 0.646 \\ \gamma &= 2.022, \end{aligned}$$

and of course the same first three moments as in (15).

For the two-moment match, we set $\lambda = A = 3$ and from (14) obtain

$$\omega = 0.669$$

$$\gamma = 2.621.$$

The first two moments will be equal to $M_{(1)}$ and $M_{(2)}$ of (15) and the third moment is found to be

$$G^{(3)}(1) = 1.049$$

which is not significantly different from $M_{(3)}$ in (15). Computing the state probabilities, we obtain the following result in which the two-moment match of the negative binomial fit[5] was included for comparison:

State	Exact State Probabilities	Three-Moment Match	Two-Moment Match	Negative Binomial
0	0.6164	0.6161	0.6149	0.6086
1	0.2312	0.2318	0.2344	0.2464
2	0.0965	0.0962	0.0953	0.0924
3	0.0372	0.0372	0.0366	0.0337
4	0.0130	0.0130	0.0129	0.0122
5	0.0041	0.0041	0.0042	0.0043
6	0.0012	0.0012	0.0012	0.0015
7	0.0003	0.0003	0.0003	0.0005
8	0.0001	0.0001	0.0001	0.0002

We now examine the error $E_n(a, c)$. Since little additional computation is required to obtain the three-moment match, we feel it should be done to obtain a better fit. Consequently, we examined the error for a three-moment match only. Calculations of $f(n)$, the state probabilities as they are given exactly, and $p(n)$, the interrupted Poisson approximation, have been made. This was done for groups of 1, 2, 4, 8, 16, 32, 40, and 48 trunks in the primary group with offered occupancies of 0.75 and 1, and for a group of 64 with offered occupancy of 0.75. For larger trunk groups, significant loss of accuracy prevented successful computation. We note that for the case of one trunk the approximation is exact. Where comparison could be made with previously reported results for the negative binomial and the confluent hypergeometric approximations,[7] it was found that the three-moment match using the interrupted Poisson process gave uniformly better fit to the state probabilities.

A typical result of the computations made is displayed in Table I. It was found that for a fixed occupancy the errors would increase, reach a maximum, and then decrease as the number of trunks was increased. This behavior can be seen in Fig. 1 where

$$E = \max_n |E_n(a, c)| = \max_n |f(n) - p(n)|$$

TABLE I—THREE-MOMENT MATCH
($c = 16$, $a/c = 0.75$, $\lambda = 6.83$, $\omega = 0.366$, $\gamma = 3.09$)

State n	$f(n)$	$p(n)$	$E_n(12,16) = f(n) - p(n)$	$\frac{E_n(12,16)}{f(n)}$
0	0.64866359	0.64765765	0.00100593	0.00155078
1	0.17173954	0.17395867	−0.00221912	−0.01292144
2	0.08460030	0.08381668	0.00078362	0.00926265
3	0.04523738	0.04463748	0.00059990	0.01326127
4	0.02430285	0.02417444	0.00012841	0.00528357
5	0.01280780	0.01290307	−0.00009527	−0.00743819
6	0.00655960	0.00668314	−0.00012354	−0.01883330
7	0.00325188	0.00333032	−0.00007845	−0.02412386
8	0.00155792	0.00158885	−0.00003093	−0.01985106
9	0.00072097	0.00072378	−0.00000280	−0.00388859
10	0.00032234	0.00031441	0.00000793	0.02460252
11	0.00013930	0.00013020	0.00000910	0.06530564
12	0.00005822	0.00005142	0.00000681	0.11694969
13	0.00002356	0.00001937	0.00000418	0.17756000
14	0.00000923	0.00000697	0.00000226	0.24472624
15	0.00000351	0.00000240	0.00000111	0.31590026
16	0.00000129	0.00000079	0.00000050	0.38876238
17	0.00000046	0.00000025	0.00000021	0.46060149
18	0.00000016	0.00000008	0.00000009	0.53013187
19	0.00000005	0.00000002	0.00000003	0.59587179

is plotted against c for different values of the offered occupancy a/c and again in Fig. 2 where the relative error corresponding to the state n at which $|E_n(a, c)|$ attained its maximum is plotted.

III. INTERARRIVAL TIME DISTRIBUTION

In analytical studies of systems, a description of overflow traffic is sometimes needed in terms of the interarrival time distribution.[2] This distribution is complicated and may be difficult to compute whenever the size of the trunk group which the Poisson traffic is overflowing is large.[8] If the interrupted Poisson traffic is used to generate the overflow traffic, the resulting interarrival time distribution, say $A(t)$, is simple. Indeed, we show that this distribution is given simply by the mixture of two exponential distributions:

$$A(t) = k_1(1 - e^{-r_1 t}) + k_2(1 - e^{-r_2 t}), \tag{16}$$

where

$$r_1 = \tfrac{1}{2}\{\lambda + \omega + \gamma + \sqrt{(\lambda + \omega + \gamma)^2 - 4\lambda\omega}\},$$

$$r_2 = \tfrac{1}{2}\{\lambda + \omega + \gamma - \sqrt{(\lambda + \omega + \gamma)^2 - 4\lambda\omega}\},$$

$$k_1 = \frac{\lambda - r_2}{r_1 - r_2},$$

$$k_2 = 1 - k_1.$$

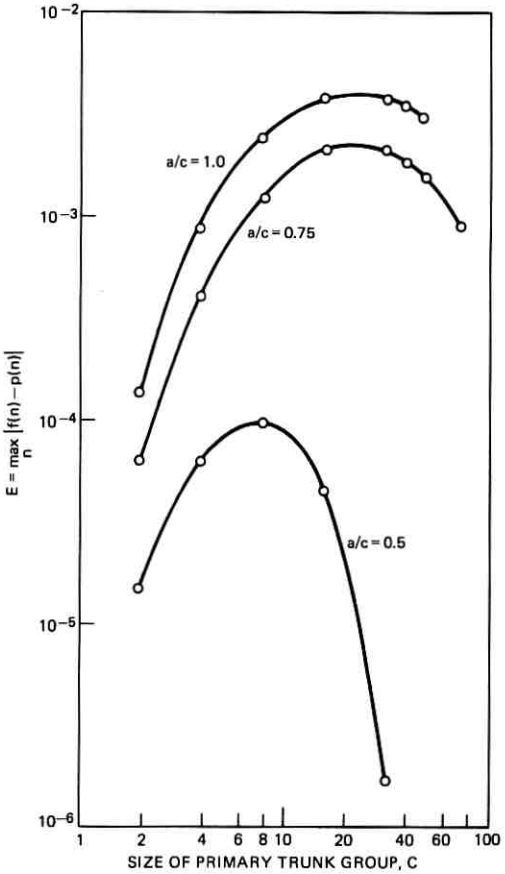


Fig. 1—Maximum absolute error E.

The proof goes as follows. Let W_n be the waiting time from $t = 0$ until the time of the nth arrival and $H_n(t)$ be the distribution of W_n. To obtain the interarrival time distribution, it is not necessary to find $H_n(t)$ for all n. The distribution $H_1(t)$ and proper choice of initial conditions at $t = 0$ is sufficient to find $A(t)$. We include the more general case here for completeness.

If $N(t)$ counts the number of arrivals in $(0, t)$ and

$$p_k(t) = P[N(t) = k],$$

then

$$H_n(t) = 1 - \sum_{k=0}^{n-1} p_k(t). \tag{17}$$

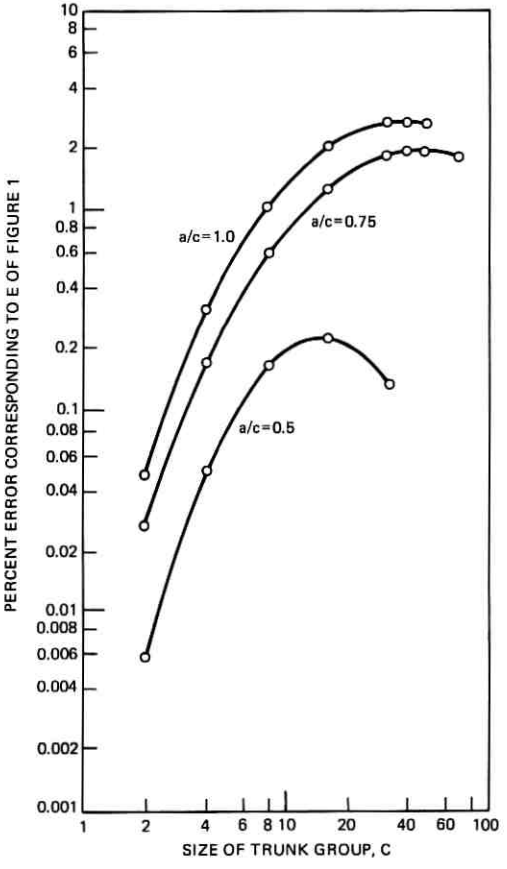


Fig. 2—Relative absolute error corresponding to Fig. 1.

This equation follows from the observation that $W_n > t$ if and only if $N(t) < n$.

Taking the Laplace-Stieltjes transform of both sides of (17) we obtain

$$\alpha_n(s) = 1 - s \sum_{k=0}^{n-1} \pi_k(s), \tag{18}$$

where

$$\alpha_n(s) = \int_0^\infty e^{-st} dH_n(t),$$

$$\pi_k(s) = \int_0^\infty e^{-st} p_k(t) dt.$$

The initial conditions used to obtain (18) are

$$p_k(0^+) = \begin{cases} 1, & k = 0 \\ 0, & k > 0. \end{cases} \tag{19}$$

We will compute $\pi_k(s)$ and hence determine $\alpha_n(s)$. From the expression for $\alpha_n(s)$ we then determine the interarrival distribution $A(t)$.

Let $p_{km}(t)$ be the probability that there were k arrivals in $(0, t)$, given that an arrival occurred at $t = 0$ and that at the instant t the switch is on if $m = 1$ and off if $m = 0$. These functions satisfy the system of differential equations

$$\begin{aligned} p'_{01}(t) &= \omega p_{00}(t) - (\lambda + \gamma)p_{01}(t), \\ p'_{k1}(t) &= \omega p_{k0}(t) - (\lambda + \gamma)p_{k1}(t) + \lambda p_{k-1,1}(t), \qquad k = 1, 2, \cdots, \\ p'_{k0}(t) &= -\omega p_{k0}(t) + \gamma p_{k1}(t), \qquad k = 0, 1, 2, \cdots, \end{aligned} \tag{20}$$

and the initial condition $p_{01}(0) = 1$.

Taking the Laplace transform of (20), we obtain

$$\begin{aligned} s\pi_{01}(s) &= \omega\pi_{00}(s) - (\lambda + \gamma)\pi_{01}(s) + 1, \\ s\pi_{k1}(s) &= \omega\pi_{k0}(s) - (\lambda + \gamma)\pi_{k1}(s) + \lambda\pi_{k-1,1}(s), \qquad k = 1, 2, \cdots, \\ s\pi_{k0}(s) &= -\omega\pi_{k0}(s) + \lambda\pi_{k1}(s), \qquad k = 0, 1, 2, \cdots, \end{aligned}$$

where

$$\pi_{kj}(s) = \int_0^\infty e^{-st} p_{kj}(t)dt.$$

This system of difference equations can be solved for $\pi_{k0}(s)$ and $\pi_{k1}(s)$ and hence $\pi_k(s)$, since $\pi_k(s)$ is the sum $\pi_{k0}(s) + \pi_{k1}(s)$. Omitting the details, we have

$$\pi_k(s) = \frac{s + \omega + \gamma}{\tau(s)} \left[\frac{\lambda(s + \omega)}{\tau(s)}\right]^k, \qquad k = 0, 1, 2, \cdots, \tag{21}$$

where

$$\tau(s) = s^2 + (\lambda + \gamma + \omega)s + \lambda\omega.$$

Substituting (21) into (18), we obtain

$$\alpha_n(s) = \left[\frac{\lambda(s + \omega)}{\tau(s)}\right]^n, \qquad n = 1, 2, \cdots. \tag{22}$$

The interarrival time distribution is given by the distribution of the

waiting time until the first arrival and hence its Laplace-Stieltjes transform is given by $\alpha_1(s)$. Inverting $\alpha_1(s)$, we obtain eq. (16).

REFERENCES

1. Neal, S. R., "The Equivalent Group Method for Estimating the Capacity of Partial-Access Service Systems Which Carry Overflow Traffic," B.S.T.J., *51*, No. 3 (March 1972), pp. 777–783.
2. Kuczura, A., "Queues With Mixed Renewal and Poisson Inputs," B.S.T.J., *51*, No. 6 (July-August 1972), pp. 1305–1326.
3. Kuczura, A., and Neal, S. R., "The Accuracy of Call-Congestion Measurements for Loss Systems with Renewal Input," B.S.T.J., *51*, No. 10 (December 1972), pp. 2197–2208.
4. Kosten, L., "Über Sperrungswahrscheinlichkeiten bei Staffelschaltungen," Electro Nachrichten Technik, *14* (January 1937), pp. 5–12.
5. Wilkinson, R. I., "Theories for Toll Traffic Engineering in the U.S.A.," B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
6. Rapp, Y., "Planning of Junction Network in a Multiexchange Area," Ericsson Technics, *20*, No. 1 (1964), pp. 77–130.
7. Descloux, A., "Overflow Traffic–On the Approximation of Overflow Distributions by Confluent Hypergeometric Distributions," unpublished work.
8. Riordan, J., *Stochastic Service Systems*, New York: John Wiley and Sons, 1962.

Contributors to This Issue

P. I. Bonyhard, B.Sc. (Physics), 1960, University of Leeds (England); Ph.D. (Numerical Automation), 1963, University of London (England); Bell Laboratories, 1965—. Mr. Bonyhard has worked on magnetic devices for digital systems, in particular plated wire, and, subsequently, bubbles. He is currently working on the design of bubble mass memories and on other possible applications for magnetic bubbles.

C. A. Brackett, B.S., 1962, M.S., 1963, and Ph.D., 1968, University of Michigan; Bell Laboratories, 1968—. At the University of Michigan, Mr. Brackett worked on traveling-wave tubes and beam-plasma amplifiers. At Bell Laboratories, he has been concerned with harmonic generation and multifrequency effects in IMPATT oscillators. He is presently engaged in work on GaAs injection lasers.

David S. K. Chan, S.B. (Electrical Engineering) and S.M. (Electrical Engineering), 1972, Massachusetts Institute of Technology; Bell Laboratories, 1970—. Mr. Chan has worked on digital switching and transmission systems, digital circuitry, and digital filtering. Member, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

Ta-Shing Chu, B.S., 1955, The National Taiwan University; M.S., 1957, and Ph.D., 1960, Ohio State University; Bell Laboratories, 1963—. Mr. Chu has been concerned with problems involving electromagnetic wave propagation and microwave antennas. Member, IEEE, Commission 2 of URSI, Sigma Xi, Pi Mu Epsilon.

Larry J. Greenstein, B.S.E.E., 1958, M.S.E.E., 1961, and Ph.D. (E.E.), 1967, Illinois Institute of Technology; Bell Laboratories, 1970—. Since joining Bell Laboratories, Mr. Greenstein has been engaged in studies of digital encoding, processing, and transmission. His current activities are in the area of radio communication. Member, AAAS, IEEE.

ANATOL KUCZURA, B. S. (Engineering Physics), 1961, University of Illinois; M.S. (Mathematics), 1963, University of Michigan; M.S.E.E., 1966, New York University; Ph.D. (Mathematics), 1971, Polytechnic Institute of Brooklyn; Bell Laboratories, 1963–1973. From 1963 to 1966, Mr. Kuczura worked in military systems engineering. Since 1966, he has been engaged in research on the application of probability theory and stochastic processes to the analysis of telephone traffic and queuing. Mr. Kuczura is now Director, Systems Analysis, at the North Electric Company's Paul H. Henson Research Center. Member, ORSA, SIAM, American Mathematical Society, Mathematical Association of America, AAAS, Chi Gamma Iota, Pi Mu Epsilon.

WANDA L. MAMMEL, A.B. (Mathematics), 1943, Winthrop College; M.Sc. (Applied Mathematics), 1945, Brown University; Bell Laboratories, 1956—. Ms. Mammel is engaged in finding mathematical methods for the numerical solution of a variety of problems. In particular, she has applied linear programming techniques to problems of crystal plasticity. At present she is working on problems in microwave propagation and optical waveguides.

DIETRICH MARCUSE, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Laboratories at the University of Utah. He is presently working on the transmission aspect of a light communications system. Mr. Marcuse is the author of two books. Fellow, IEEE; member, Optical Society of America.

T. J. NELSON, B.S.(E.E.), 1961, Iowa State University; M.E.E., 1963, New York University; Ph.D. (Physics), 1967, Iowa State University; Bell Laboratories, 1961–1963 and 1969—. Mr. Nelson has studied mode locking in lasers and magnetic bubble materials and devices. Member, Eta Kappa Nu, Sigma Xi, Phi Kappa Phi, American Physical Society.

LAWRENCE R. RABINER, S.B., S.M., 1964, Ph.D., 1967, Massachusetts Institute of Technology; Bell Laboratories, 1962—. Mr. Rabiner has worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech communications and digital signal processing techniques. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi; Fellow, Acoustical Society of America; Chairman of the IEEE G-AU Technical Committee on Digital Signal Processing; vice-president of the G-AU AdCom, associate editor of the G-AU Transactions; member of the technical committees on speech communication of both the IEEE and Acoustical Society.