THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XLVI JULY–AUGUST 1967 NUMBER 6

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published ten times a year by the American Telephone and Telegraph Company, B. S. Gilmer, President, C. E. Wampler, Vice President and Secretary, J. J. Scanlon, Vice President and Treasurer. Checks for subscriptions should be made payable to American Telephone and Telegraph Company and should be addressed to the Treasury Department, Room 2312C, 195 Broadway, New York, N. Y. 10007. Subscriptions $5.00 per year; single copies $1.25 each. Foreign postage $1.08 per year; 18 cents per copy. Printed in U.S.A.

THE BELL SYSTEM

TECHNICAL JOURNAL

VOLUME XLVI JULY-AUGUST 1967 NUMBER 6

The Si-SiO_2 Interface – Electrical Properties as Determined by the Metal-Insulator-Silicon Conductance Technique

By E. H. NICOLLIAN and A. GOETZBERGER

(Manuscript received December 28, 1966)

Measurements of the equivalent parallel conductance of metal-insulator-semiconductor (MIS) capacitors are shown to give more detailed and accurate information about interface states than capacitance measurements. Experimental techniques and methods of analysis are described. From the results of the conductance technique, a realistic characterization of the Si–SiO_2 interface is developed. Salient features are: A continuum of states is found across the band gap of the silicon. Capture cross sections for holes and electrons are independent of energy over large portions of the band gap. The surface potential is subject to statistical fluctuations arising from various sources. The dominant contribution in the samples measured arises from a random distribution of surface charge. The fluctuating surface potential causes a dispersion of interface state time constants in the depletion region. In the weak inversion region the dispersion is eliminated by interaction between interface states and the minority carrier band. A single time constant results. From the experimentally established facts, equivalent circuits accurately describing the measurements are constructed.

I. INTRODUCTION

The electrical properties of semiconductor surfaces have been studied for a long time. Until recently most investigators were concerned with either etched surfaces covered by a thin natural oxide

layer or with ultraclean surfaces that exist only in high vacuum. Development of improved oxidation techniques for silicon have made it now possible to study the electrical phenomena occurring at the interface between silicon and silicon dioxide. These studies resulted in the recognition that a semiconductor-insulator interface behaves differently in many ways from a "bare" surface. Of practical importance is the fact that an oxide-covered surface is more stable and can be made electrically more perfect than an unprotected surface.

The difference between oxidized and bare surfaces will be listed here briefly.

(*i*) There is no charge exchange with states at the air oxide interface. This type of state, the "slow surface state," has a long time constant and is encountered on etched surfaces. For thick oxide layers, only states at the oxide-semiconductor interface have to be considered. They are equivalent to "fast states" and will be called interface states in this paper. In addition to interface states, there may be traps within the oxide.[1, 2] Not much is known about these traps and they will not be a major topic of this paper.

(*ii*) All oxidized silicon surfaces contain a so-called surface charge. It consists of positive charges residing close to the interface. The surface charge originates from two sources, alkali ions[3] and built-in charges.[4, 5] Alkali ions can drift through the oxide at fairly low temperatures and thus be distinguished from built-in or residual charges which are fixed and have a total density which depends on oxidation rate and crystal orientation.[6] The present paper deals with interface states in alkali-free systems. Reference to residual surface charge will be necessary in connection with their influence on the uniformity of surface potential.

(*iii*) Thermally oxidized Si–SiO_2 interfaces are characterized by a much lower density of states than "bare" surfaces. The density of states over most of the silicon band gap in the oxidized systems is usually not much greater than 10^{12} cm^{-2}-eV^{-1} or smaller than 10^{10} cm^{-2}-eV^{-1} depending on the method of oxide preparation. For "bare" surfaces, density of states can be as high as 10^{14} cm^{-2}-eV^{-1}.

The electrical properties of interface states are characterized by their density, their position in the energy gap of the silicon, and their capture cross section. In addition, it is necessary to know whether they are of the donor or acceptor type. The most widely used tool for investigating these interface state properties is the metal-insulator-semiconductor (MIS) capacitor.

Dispersion of the capacitance can be used to obtain information

about the energy distribution and density of interface states as has been shown by Terman.[7] The capacitance technique, however, has severe limitations.[8] Essentially, the difficulty is that interface state capacitance must be extracted from measured capacitance which consists of oxide capacitance, depletion layer capacitance, and interface state capacitance. This difficulty does not apply to the equivalent parallel conductance because conductance arises solely from the steady-state loss due to the capture and emission of carriers by interface states and is thus, a more direct measure of these properties.[9] Conductance measurements yield more accurate and reliable results, particularly when the density of interface states is low as in the thermally oxidized system because only directly measured quantities are used with no matching to a theoretical curve required. Both the capacitance and equivalent parallel conductance as functions of voltage and frequency contain identical information about interface states. Greater inaccuracies arise in extracting this information from the capacitance. This is illustrated in Fig. 1 which shows capacitance

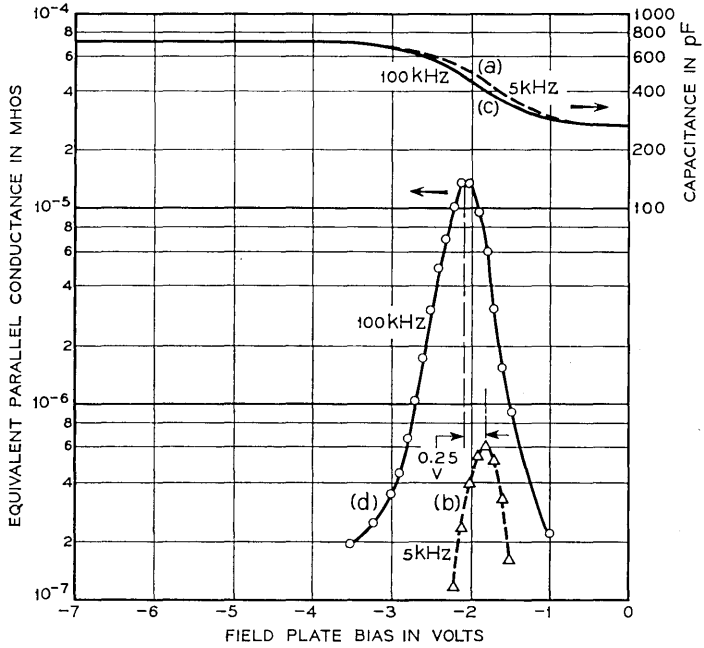


Fig. 1 — Capacitance and equivalent parallel conductance measured at 5 kHz and 100 kHz on a p-type sample having an acceptor density of 2.08×10^{16} cm^{-3} and interface state density in the $10^{11} cm^{-2}\text{-}eV^{-1}$ range. Interface state time constants and densities are given in Fig. 14 and Fig. 15 for this sample.

and equivalent parallel conductance measured at 5 kHz and 100 kHz. The largest capacitance spread is 14 percent while the magnitude of the conductance peak increases by over one order of magnitude in this frequency range.

In order to evaluate MIS measurements with regard to interface states, it is desirable to have one-dimensional current flow perpendicular to the interface. This condition does not apply when the silicon underneath the field plate and beyond is inverted. Then, the capacitance and conductance are dominated by the lateral ac current flow that has been explained elsewhere.[10, 11] Because of the positive polarity of surface charge, p-type silicon is normally inverted and is therefore most easily investigated in the depletion-accumulation region.

In the MIS capacitor, the density of states is normally low so that the largest silicon band bending possible is not limited by their influence. Maximum possible band bending in the MIS capacitor is slightly less than the band gap. In this paper, interface state properties are described for energies in the forbidden gap within the range $2u_B$ [see Fig. 5(a)] only at room temperature.

Lehovec and Slobodskoy[12] theoretically treat losses due to steady state capture and emission of both minority* and majority* carriers by a hypothetical single level interface state for the cases of zero and infinite recombination-generation rate through states in the silicon space-charge region. No direct experimental measurement and evaluation of these losses has been made until recently.[9] The theory and equivalent circuits developed in Refs. 12 and 13 constitute a very general description of all the losses and charge storage processes possible in an MIS structure. The gap between this general theory and what is actually measured in a sodium-free MIS structure is bridged in this work. It can be shown by conductance measurements that:

(*i*) loss in the oxide measured at room temperature is negligible,

(*ii*) in the samples studied, the dominant loss is by capture and emission of carriers by interface states rather than states in the silicon space charge region, and

(*iii*) the majority carrier time constant is always the dominant one.

These three results make it possible to study the capture and emission properties of the interface states for each type of carrier independently. It is this fact which makes the MIS capacitor a powerful tool. We shall divide the range of surface potential from accumulation to

* The term minority carrier denotes those carriers which are in a minority in the bulk of the semiconductor crystal. The term majority carrier denotes those in a majority in the bulk of the crystal.

inversion into four regions each of which is characterized by a different equivalent circuit. These regions are: depletion-accumulation, mid-gap, weak inversion, and strong inversion. The depletion-accumulation region includes those values of surface potential less than u_B where u_B is shown in Fig. 5(a). Mid-gap is the region a few kT/q volts wide about u_B. Weak inversion is between u_B and $2u_B$. Strong inversion includes those values of surface potential greater than $2u_B$. The following results have been obtained using the MIS conductance technique.

(*i*) A continuum of states across the bandgap of the silicon appears to be characteristic of the Si–SiO_2 interface.

(*ii*) Capture cross sections for holes and electrons are independent of energy over large portions of the bandgap.

(*iii*) Time constant dispersion at each bias in the depletion-accumulation region is dominated by statistical fluctuations of surface potential. In the samples measured, these fluctuations are caused primarily by a random distribution of built-in residual oxide charges and charged interface states over the plane of the interface.

(*iv*) A single time constant at each bias is observed in the region of weak inversion because the minority carrier capture resistance becomes negligibly small and the inversion layer provides a large lateral conductance across the silicon surface.

(*v*) Equivalent circuits are derived for the depletion-accumulation, mid-gap, and weak inversion regions. (See Fig. 31) The equivalent circuit which applies in strong inversion has been worked out previously.[11, 14] This circuit was proposed in Ref. 11 for the case where lateral ac current flow is absent and has been verified by experiment using n-type samples.[14]

II. SAMPLE PREPARATION

The sample preparation processing described in this section is intended to yield oxides which remain stable only during admittance-voltage measurements made at room temperature or lower. The stability requirements for device applications would be more stringent than this and would require somewhat different processing. No attempt is made to minimize built-in charge density. Both high and low interface state densities are purposely produced.

Measurements were made on both boron and arsenic doped silicon oriented in the [111] and [100] directions. Samples were prepared by first growing a 10-μ thick epitaxial layer having a nominal resistivity of 1 ohm-cm on a 0.005 ohm-cm substrate. The substrate always in-

corporates the same dopant as the epitaxial layer. The purpose of using an epitaxial layer is to minimize bulk series resistance to the point where it can be neglected—an important consideration when making conductance measurements.

The slices about 1 cm square were removed from the epitaxial grower and immediately placed in the oxidation apparatus described in Ref. 4. An oxide typically between 500 Å and 700 Å thick was then grown at 1000°C in steam at atmospheric pressure by the bias technique of Ref. 4.

Another group of samples was prepared by growing about the same thickness of oxide the same way except in dry oxygen at 1000°C and atmospheric pressure. Trace quantities of H_2 were removed from the O_2 stream by first passing it through a Deoxo* unit. The O_2 was then dried by passing it through a dry ice acetone trap. The oxide should be as thin as possible without being in the tunnelling range to make the measured admittance correspond more closely to the admittance of the interface. Oxidizing the sample immediately after the epitaxial layer is grown insures that the silicon surface has not had much opportunity to gather contamination from its surroundings. The oxide was grown by the bias technique to obtain an oxide in which ionic contamination is a minimum. The reason for taking pains to minimize ionic contamination in the oxide is that the presence of an ion near an interface state may perturb its energy and capture probability although such ions do not themselves act as interface states.[3]

Immediately after oxidation, the slice was placed in an oil diffusion pump evaporator and a number of chrome-gold field plates were evaporated through a metal mask held in intimate contact with the oxide surface. First, a few hundred Å of chromium was deposited to make a strong bond to the oxide. This was followed by 2000 Å of gold to provide good electrical contact.

The chromium and gold sources were heated to evaporation temperature in adjacent tungsten baskets. The baskets are 15 cm above the surface of the sample to reduce penumbra effects.[8] Since not only capacitance but also equivalent parallel conductance depends on the area of the field plate, a large field-plate diameter is used to get a large admittance thereby increasing detection sensitivity. However, making the field-plate diameter large increases the likelihood of encompassing defects and gross nonuniformities under it. A satisfactory diameter with these considerations in mind can be as large as

* This unit contains a catalyst which enables H_2 and O_2 to react at room temperature producing H_2O.

1500 μ. The oxide on the substrate side was then etched off in HF. Cr–Au in the same proportions as in the field plate was evaporated all over this side to make the back contact. The slice was placed in good thermal contact with a massive copper block during all the evaporation steps to keep it always at room temperature. This minimizes sodium migration into the oxide during evaporation. Minimizing sodium contamination in this step of the processing also reduces the likelihood of a patchy charge distribution underneath the field plate.

III. ORIGIN OF THE EQUIVALENT PARALLEL CONDUCTANCE

Two experiments are described which show that the dominant process causing the measured equivalent parallel conductance is capture and emission of carriers by interface states. Series resistance is not important because it can be made negligible by using epitaxial samples or it can be measured in strong accumulation and substracted from the total impedance. Normally, small-signal measurements are made so that harmonics of the signal frequency arising from the nonlinearity of the charge-voltage characteristic are unimportant (see Section 5.1).

3.1 *Loss in the Oxide*

The first experiment shows that equivalent parallel conductance measured at room temperature does not arise from loss in the oxide. Capacitance and equivalent parallel conductance for a p-type sample are plotted as functions of field plate bias in Fig. 2 before and after aging an unstable oxide in room air at 150°C. This aging produces an increased positive charge density in the oxide as seen by the shift of the capacitance curve to higher negative bias. This means that the electric field in the oxide resulting from the applied bias is greater for a given silicon surface potential after aging. However, the conductance curve has not changed appreciably in shape or magnitude. The conductance curve is shifted to higher negative bias by the same amount as the capacitance curve. This is seen by the fact that the peak of the conductance occurs at the same value of capacitance before and after aging. Therefore, the conductance is a function of silicon surface potential and not of field in the oxide. Therefore, it is not related to loss in the oxide.

3.2 *Loss in the Silicon Space-Charge Region*

The most important processes which can give rise to the measured conductance are capture and emission of carriers by interface states or

by states in the silicon space-charge region. It is interesting to note that the MIS capacitor can be used to measure the properties of either type of state provided one or the other dominates the loss processes.* Processing determines the densities and positions in the bandgap of levels of each type of state. In all samples made as described in Section II, interface state loss is dominant over the entire measurable range. In all these samples, no impurities other than arsenic or boron

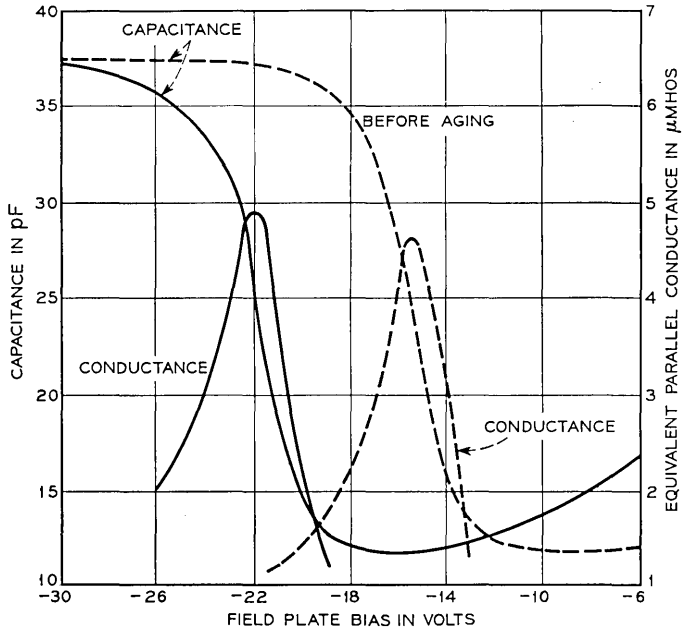


Fig. 2 — Measured capacitance and equivalent parallel conductance as functions of field plate bias before and after aging at 150°C in room air. Field plate diameter is 3.8×10^{-2}cm, silicon resistivity 50 Ω-cm, p-type, crystal orientation [111], oxide thickness 1200 Å, and measurement frequency 100 kHz.

are purposely introduced into the silicon. The densities and energy levels of other impurities or defects inadvertently introduced into the silicon during fabrication of the MIS capacitor are unknown. The same is true for interface states. It is best therefore, to establish by experiment which effect dominates. This is described next.

The magnitude of the equivalent parallel conductance measured is always much smaller when the oxide is grown in steam than when the

* Sah[15] has used the MIS capacitor to study states introduced into the silicon by heavy gold doping.

oxide is grown in dry oxygen, other parameters being the same.[4] In addition, the oxide can be reversibly cycled from the wet to the dry condition by heat treatment between 200°C and 400°C in various ambients. To enable a quantitative comparison, it is necessary that the resistivity of the silicon, oxide thickness, and field plate area be the same for both conditions. The easiest way to accomplish this is to cycle the same capacitor from the wet to a drier condition. Curves (a) and (b) in Fig. 3 are plots of capacitance and equivalent parallel

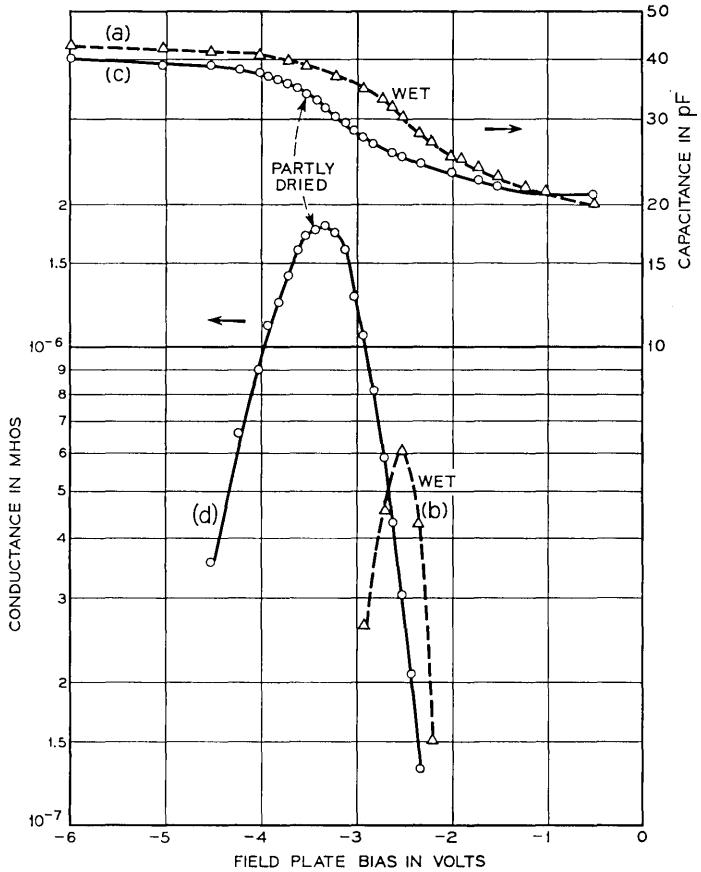


Fig. 3 — (a) capacitance and (b) equivalent parallel conductance vs field plate bias before drying; (c) capacitance and (d) equivalent parallel conductance vs field plate bias of the same capacitor after heating in dry nitrogen for 17 hours at 350°C. Field plate diameter is 3.8×10^{-2} cm, silicon resistivity 1 Ω-cm, p-type, crystal orientation [111], oxide thickness 910 Å grown in stream, and measurement frequency 100 kHz.

conductance for a steam grown oxide on p-type silicon. Curves (c) and (d) in this figure are measured after heating the capacitor in dry nitrogen (dried by passing through a liquid nitrogen trap) for 17 hours at 350°C. A similar set of curves for n-type is given in Fig. 4. All the curves in these two figures are measured at 100 kHz. Curves similar to these can be obtained at all frequencies in the range investigated (50Hz to 500 kHz). It is important to note only that the capacitance in Fig. 3 from curve (c) associated with the conductance peak of curve

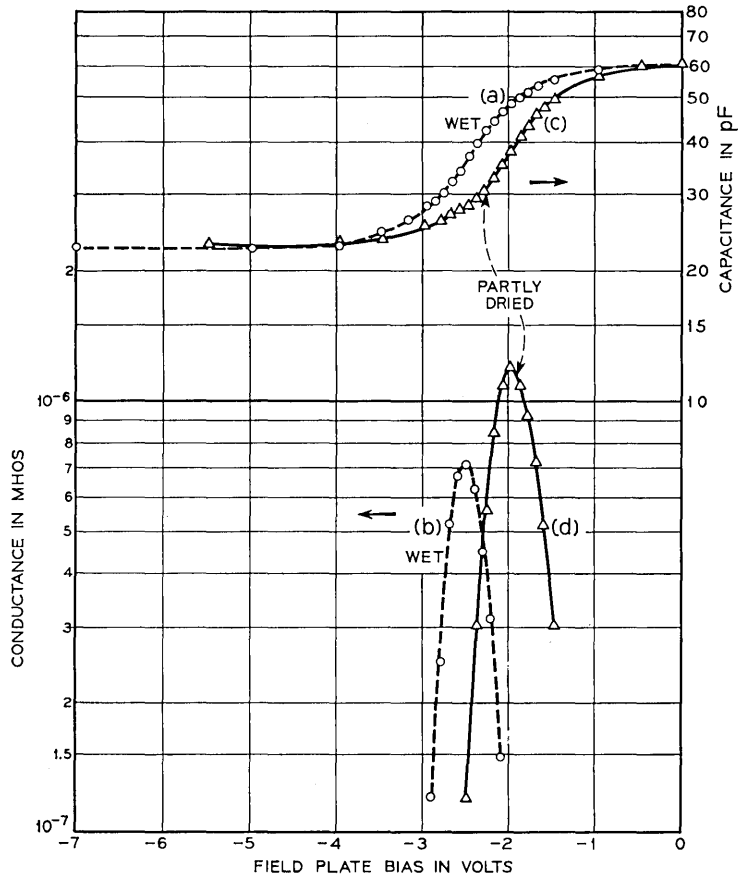


Fig. 4 — (a) capacitance and (b) equivalent parallel conductance measured vs field plate bias before drying; (c) capacitance and (d) equivalent parallel conductance vs field plate bias of the same capacitor after heating in dry nitrogen for 17 hours at 350°C. Field plate diameter is 3.8×10^{-2}cm, silicon resistivity 0.75 Ω-cm, n-type, crystal orientation [111], oxide thickness 600 Å grown in stream, and measurement frequency 100 kHz.

(d) is nearly the same as the capacitance from curve (a) associated with the conductance peak of curve (b). The same is true in Fig. 4. Therefore, all the relevant parameters before and after drying are the same except the magnitudes of the conductance peaks. The peak conductance of curve (b) in Fig. 3 is about one-third the peak conductance of curve (d). This result cannot be due to sodium contamination which might have been introduced into the oxide during the drying cycle, the initial wet oxide being sodium free (see Section II). The reasons for this are: (*i*) Sodium at the interface always remains charged in the range between accumulation and strong inversion.[3] Therefore, it does not become an interface state which could influence the conductance. (*ii*) A nonuniform charge distribution in the oxide caused by the sodium would spread the conductance curve over a larger range of bias which would decrease rather than increase the magnitude of the conductance peak if this effect were dominant. (*iii*) The p-type curves in Fig. 3 are shifted towards more negative bias while the n-type curves in Fig. 4 are shifted towards less negative bias with drying. Introduction of sodium during drying would shift both curves towards more negative bias. The increase of the conductance peak in both p and n-type is then a true drying effect.

The fact that the oxide can be reversibly cycled from the wet to the dry state at low temperatures rules out the possibility that a large density of structural defects or impurities which act as bulk states is introduced into the silicon during one method of oxide preparation and not the other. All of this strongly suggests that the measured equivalent parallel conductance is due mainly to losses in states which reside on the oxide side of the interface and exchange charge at 50 Hz to 500 kHz with the silicon because only the properties of such states could be so markedly altered by low temperature wet-dry cycling.

The most convincing evidence, however, is the dependence of interface state time constants on surface potential to be discussed later in this paper.

3.3 *Loss Due to Minority Carrier Transitions*

Minority carrier density can respond to the applied signal frequency by diffusion from the bulk to the interface and by recombination-generation processes through interface states and states in the silicon space charge region. These processes in the samples measured are too slow to allow the minority carrier density to follow 50 Hz to 500 kHz.[11, 14] It was indicated in the preceding part of this section that in this frequency range, interface state loss is solely responsible for the

measured equivalent parallel conductance. In the depletion region where minority carrier density is orders of magnitude less than majority carrier density, loss due to minority carrier transitions will be negligible compared to loss due to majority carrier transitions. Thus, in this region, the conductance measured arises simply from capture and emission of majority carriers by those interface states within a few kT/q of the Fermi level. In the region of weak inversion where minority carrier density is orders of magnitude greater than majority carrier density, minority carrier time constant becomes shorter than majority carrier time constant. Minority carriers will follow the signal frequency. However, current can flow and produce a loss only by transitions through the interface states from the minority carrier band to the majority carrier band. Therefore, this process will be controlled by the majority carrier time constant. This case will be treated in detail in Section IV. The important point is that the loss is determined entirely by the majority carrier time constant in both the depletion and weak inversion regions.

IV. THEORY

4.1 *Phenomenological Description of Loss*

A continuum of interface states over the bandgap of the silicon appears to be characteristic of the Si–SiO_2 interface. Fig. 5(a) is a schematic of an MIS capacitor showing the band bending in depletion for n-type silicon with an arbitrary distribution of interface states in the silicon bandgap. Initially, let the interface states be in equilibrium with the silicon. Then let a small ac signal between 50 Hz and 500 kHz be applied. Consider the first half of the cycle which moves the conduction band towards the Fermi level. This immediately increases the average energy of the electrons in the silicon. Because a conductance is observed, it is evident that the interface states do not respond immediately but lag behind. Therefore, electrons at a higher average energy in the silicon will be captured by interface states at a lower average energy. This results in an energy loss. On the other half of the cycle, the signal moves the valence band towards the Fermi level. Electrons in filled interface states now will be at a higher average energy than electrons in the silicon. As the electrons are emitted by the interface states into the silicon they will lose energy again. Thus, there will be an energy loss on both halves of the cycle which must be supplied by the signal source. The energy required for transitions between the band edge and an interface state is much higher than this and is

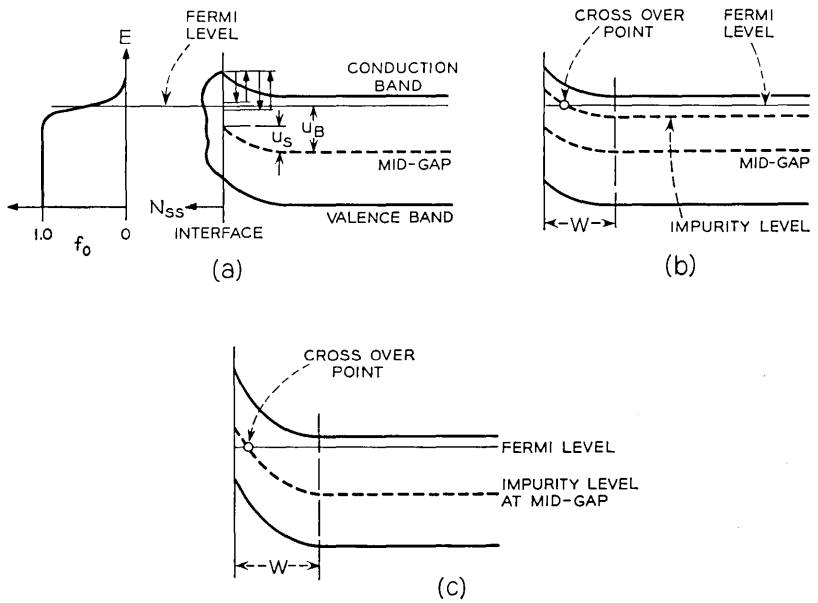


Fig. 5 — (a) Schematic showing band bending in depletion at silicon surface in an n-type MIS capacitor. An arbitrary distribution of interface states is shown for illustrative purposes. Majority carrier transitions to two levels within kT/q of the Fermi level are also shown. f_o is the Fermi function. (b) Band bending in depletion showing point at which Fermi level crosses a shallow impurity level. This impurity level is hypothetical. W is space charge width. (c) Band bending in weak inversion showing point at which Fermi level crosses a hypothetical impurity level located at mid-gap.

supplied by phonons. Only majority carrier capture and emission is considered here because this is the dominant loss mechanism for reasons given in Section III.

After a disturbance has been applied, the interface states will reach equilibrium with the silicon exponentially with a characteristic time, τ. The interface states are also capable of storing charge. Thus, there will be a capacitance, C_s, associated with them which is proportional to their density. The conductance related to the loss is $G_s = C_s/\tau$. It should be noted that G_s is independent of signal amplitude as long as C_s and τ are independent of it. This will be the case for small-signal amplitude. Interface state τ and C_s can then be extracted directly from a measurement of equivalent parallel conductance. Capture cross section is extracted from τ and density of states from C_s as will be shown in Section VII.

4.2 *Admittance of a Single Level State*

Our purpose is to derive expressions for the admittance of the MIS capacitor as a function of bias and frequency, and to obtain the corresponding equivalent circuits which fit in detail to experimental observations. The capacitance-voltage characteristics for the MIS capacitor without interface states has been described.[15, 16, 17] Capacitance dispersion caused by interface state loss was measured by Terman[7] to obtain interface state densities and time constants. Lehovec et al,[12] Bertz,[13] and Sah[15] developed equivalent circuits for a hypothetical single-level interface state from Shockley-Read statistics.[18] The equivalent circuits given in this previous work are either too general[12, 13, 15] or the measurement technique too inaccurate[7] to give a real and complete description of interface state properties. To describe the MIS conductance technique and interpret the measurements in the depletion and weak inversion regions, we shall rederive the theory for the admittance of the MIS capacitor. The starting point will be the theory for majority carrier capture and emission by a single-level interface state first calculated by Lehovec and Slobodskoy.[12]

Experimental evidence shows that only capture and emission of majority carriers are important when measuring in the depletion region. The quasi Fermi level and Fermi level are identical for majority carriers. Thus, application of an ac signal simply results in a time varying Fermi function. From Shockley and Read,[18] the capture rate of electrons, taken as majority carriers, by a single-level interface state is

$$R_n(t) = N_s c_n[1 - f(t)]n_s(t), \tag{1}$$

and the emission rate is

$$G_n(t) = N_s e_n f(t), \tag{2}$$

where N_s is the density of states cm^{-2}; c_n electron capture probability,* cm^3/sec, e_n electron emission constant, sec^{-1}, $f(t)$ the Fermi function at time t, and $n_s(t)$ electron density at the silicon surface at time t, cm^{-3}.

Net current density flowing is

$$i_s(t) = qN_s c_n[1 - f(t)]n_s(t) - qN_s e_n f(t), \tag{3}$$

where q is electronic charge in coulombs. Making the small-signal approximation, the admittance is, as is shown in Appendix A,

* This capture probability is the average over all states in the conduction band *near the silicon surface.*

$$Y_s = j\omega \frac{q^2}{kT} \frac{N_s f_o(1 - f_o)}{(1 + j\omega f_o/c_n n_{so})}, \tag{4}$$

where $j = \sqrt{-1}$, ω the angular frequency of the ac signal, sec^{-1}, k Boltzman's constant, eV $\times$ coul/°K, T the absolute temperature, °K, f_o the Fermi function established by the bias, $f_o = [1 + \exp(u - u_s)]^{-1}$, and n_{so} the electron density at the silicon surface established by the bias, cm^{-3}.

Equation (4) is the admittance of a series RC network with capacitance $C_s = q^2 N_s f_o(1 - f_o)/kT$ and time constant $\tau = f_o/c_n n_{so}$. Separating (4) into its real and imaginary parts, the equivalent parallel capacitance is

$$C = \frac{C_s}{1 + \omega^2 \tau^2}, \tag{5}$$

and the equivalent parallel conductance is

$$G_p = \frac{C_s \omega^2 \tau}{1 + \omega^2 \tau^2}. \tag{6}$$

4.3 *Equivalent Circuit of MIS Capacitor*

Let Q_T be the total charge density at a given bias. Then

$$Q_T = Q_{sc} + Q_s + Q_f, \tag{7}$$

where Q_{sc} is the silicon space-charge density, coul/cm^2, Q_s the interface state charge density, coul/cm^2, and Q_f the fixed-charge density in the oxide, coul/cm^2. The ac current density, $i_T(t)$, obtained by differentiating (7) with respect to time is

$$i_T(t) = i_{sc}(t) + i_s(t), \tag{8}$$

where $i_{sc}(t)$ and $i_s(t)$ are the ac current densities charging the silicon space-charge layer and the interface states, respectively, To obtain $i_{sc}(t)$ we have

$$i_{sc}(t) = \left(\frac{dQ_{sc}}{d\psi_s}\right)\left(\frac{d\psi_s}{dt}\right), \tag{9}$$

where $\psi_s(t) = \psi_{so} + \delta\psi_s$ is the silicon band bending in volts at time, t, ψ_{so} the silicon band bending established by the bias, and $\delta\psi_s = a \exp(j\omega t)$. From this, $d\psi_s/dt = j\omega\, \delta\psi_s$. Also, $dQ_{sc}/d\psi_s = C_D$ the depletion layer capacitance per cm^2. Substituting these into (9) we get

$$i_{sc} = j\omega C_D\, \delta\psi_s. \tag{10}$$

It is shown in Appendix A that $i_s(t)$ given by (3) can be written

$$i_s(t) = Y_s \, \delta\psi_s \, , \tag{11}$$

where Y_s is defined by (4). Substituting (10) and (11) into (8), we have

$$i_T(t) = (j\omega C_D + Y_s) \, \delta\psi_s \, . \tag{12}$$

From (12), it is seen that C_D appears in parallel with the series RC network of the interface states. The voltage, $v_a = v_o + \delta v_a$, applied to the capacitor divides between the silicon and the oxide film so that

$$v_a(t) = v_o + \delta v_a = \psi_s(t) + \frac{Q_T}{C_{ox}} \, , \tag{13}$$

where v_o is the dc bias, $\delta v_a = b \exp(j\omega t)$, and C_{ox} is the oxide capacitance per cm^2. Differentiating (13) with respect to t to get the ac terms only

$$j\omega \, \delta v_a = j\omega \, \delta\psi_s + \frac{i_T(t)}{C_{ox}} . \tag{14}$$

Substituting for $\delta\psi_s$ from (12)

$$\delta v_a = i_T(t)\left(Z_s + \frac{1}{j\omega C_{ox}}\right) , \tag{15}$$

where

$$Z_s = (j\omega C_D + Y_s)^{-1} .$$

The bracketed term in (15) is the impedance of a circuit in which C_{ox} is in series with Z_s. The equivalent circuit for the MIS capacitor from (4), (12), and (15) is given in Fig. 6.

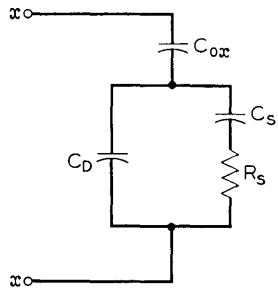


Fig. 6 — Equivalent circuit of MIS capacitor for capture and emission of majority carriers by a single level interface state. C_{ox} is the oxide capacitance per unit area, C_D the depletion layer capacitance per unit area, C_s the interface state capacitance per unit area, and R_s the majority carrier capture resistance. Bulk generation-recombination is zero.

The principle of the MIS conductance technique is easily illustrated by this simplified equivalent circuit. The admittance of the capacitor is measured by a bridge across the terminals x–x. The oxide capacitance is measured in the region of strong accumulation. The admittance of the network is then converted into an impedance. The reactance of the oxide capacitance is subtracted from this impedance and the resulting impedance converted back into an admittance. This leaves C_D in parallel with the series RC network of the interface states. The capacitance from (5) and (12) is

$$C_p = C_D + \frac{C_s}{1 + \omega^2\tau^2}, \tag{16}$$

and the equivalent parallel conductance divided by ω from (6) is

$$\frac{G_p}{\omega} = \frac{C_s\omega\tau}{1 + \omega^2\tau^2}. \tag{17}$$

G_p is divided by ω to make (17) symmetrical in $\omega\tau$. It should be noted that (16) and (17) are equivalent to the Debye equations,[19, 20] for a dielectric in which the polarization can relax exponentially with characteristic time, τ, after a change in applied electric field occurs.

Equation (16) describes the capacitance dispersion and is the basis of Terman's method. To extract C_s and τ from C_p using (16), C_D must be known. C_D can be calculated[16] using an estimated doping density. The doping density is not accurately known near the silicon surface because of pile up or depletion of the dopant during oxidation. Equation (17) on the other hand depends *only* on the interface state branch of the equivalent circuit. G_p/ω goes through a maximum when $\omega\tau = 1$ which gives τ directly. The value of G_p/ω at the maximum is $C_s/2$. Thus, *equivalent parallel conductance* corrected for C_{ox} gives C_s and τ *directly* from the *measured* conductance.

4.4 *Admittance of an Interface State Continuum*

The interface states are observed to be comprised of many levels so closely spaced in energy that they cannot be distinguished as separate levels. Thus, they appear as a continuum over the bandgap of the silicon. These observations are documented in Section VII and discussed in Section VIII and agreed with those of other investigators.[7, 21] A continuum of interface states appears to be characteristic of the Si–SiO_2 interface. The admittance given by (4) for a single-level state and the equivalent circuit in Fig. 6 must therefore be modified.

For a continuum of states at a finite absolute temperature, capture

and emission of majority carriers can occur by states located within a few kT/q on either side of the Fermi level. This results in a time constant dispersion. The admittance of the continuum first calculated by Lehovec[22] is obtained by integrating (4) over the bandgap

$$Y_{ss} = j\omega\left(\frac{q^2}{kT}\right)\int \frac{N_{ss}f_o(1 - f_o)\,d\psi}{(1 + j\omega f_o/c_n n_{so})}. \tag{18}$$

Here N_{ss} is the density of interface states, cm^{-2} eV^{-1}, and ψ is energy in ev. The integrand of (18) is sharply peaked about the Fermi level with a width of about kT/q. Thus, (18) can be easily integrated if both N_{ss} and c_n the capture probability* do not vary very much with ψ over a range kT/q. Experimentally, it is found (see Section VII) that neither vary appreciably over kT/q. Making the substitution $f_o(1-f_o) = (kT/q)\ (df_o/d\psi)$ transforms (18) into an integral over f_o. Integrating from zero to unity yields.

$$Y_{ss} = \frac{qN_{ss}}{2\tau_m}\ln(1 + \omega^2\tau_m^2) + jq\frac{N_{ss}}{\tau_m}\arctan(\omega\tau_m) \tag{19a}$$

where

$$\tau_m = 1/c_n n_{so}. \tag{19b}$$

Concentrating on the real part of (19a), we have for the continuum

$$\frac{G_p}{\omega} = \frac{qN_{ss}}{2\omega\tau_m}\ln(1 + \omega^2\tau_m^2). \tag{20}$$

At no value of bias does (20) even closely fit experimentally measured G_p/ω vs ω curves. As will be shown in the next section, (20) is just the first step in deriving an expression for G_p/ω which fits the experiment in depletion. To proceed further, it is necessary to consider separately the depletion region and the weak inversion region. In depletion, the time constant dispersion observed is far broader than given by (20). In weak inversion, the time constant dispersion disappears and the experimental G_p/ω vs ω curves are characterized by a single time constant at each bias.

4.5 *Depletion Region*

The model suggested here for the increased broadening of the time constant dispersion observed experimentally (see Section VII) in the depletion region assumes statistical fluctuations of surface potential in the plane of the interface. Majority carrier density at the silicon

* This capture probability is now the average over all the interface states in the range of the integral and all states in the conduction band near the silicon surface.

surface is related to surface potential by

$$\text{and} \quad \begin{aligned} n_{so} &= N_D \exp u_s = n_i \exp (u_s - u_B) \qquad \text{for n-type} \\ p_{so} &= N_A \exp (-u_s) = n_i \exp -(u_s - u_B) \quad \text{for p-type,} \end{aligned} \tag{21}$$

where n_i is intrinsic carrier density, cm^{-3}, $u_B = \ln n_i/N_D$ for n-type and $u_B = \ln N_A/n_i$ for p-type. u_B is the potential difference between mid-gap* and Fermi level in the quasi neutral region of the silicon in units of kT/q, N_D is the ionized donor and N_A the ionized acceptor density in the silicon, cm^{-3}, u_s is the silicon band bending or surface potential in units of kT/q. $u_s - u_B$ is the potential difference between mid-gap and Fermi level at the silicon surface.

The relation between time constant and surface potential is from (19b) and (21).

$$\text{and} \quad \begin{aligned} \tau_m &= \frac{1}{c_n n_i} \exp -(u_s - u_B) \quad \text{for n-type} \\ \tau_m &= \frac{1}{c_p n_i} \exp (u_s - u_B) \qquad \text{for p-type.} \end{aligned} \tag{22}$$

Equation (22) shows that small fluctuations in u_s will cause large fluctuations in τ_m and thus broadened time constant dispersion. If we assume that the built-in charges and charged interface states are randomly distributed in the plane of the interface, the electric field at the silicon surface will fluctuate over the plane of the interface. This is shown schematically in Fig. 7. Fluctuations in electric field will cause corresponding fluctuations in surface potential. To express this quantitatively,† we conceptually divide the plane of the interface into a number of squares of equal area. The area of each square is the largest area within which the surface potential is uniform. We shall call this particular area a characteristic area. Now, the admittance of the continuum given by (19) can be regarded as the result of integrating the admittance of a single level (4) over all the levels located within a characteristic area rather than within the entire area of the field plate. The total admittance is obtained by integrating the contribution from each characteristic area over all the characteristic areas within the area of the field plate. This must be done for each bias and frequency to generate the complete family of curves. The calculated

* Although the intrinsic Fermi level is slightly above mid-gap in silicon, this is neglected and intrinsic Fermi level is taken to be at mid-gap.

† This treatment is similar to the one used by W. Shockley in Ref. 23 for calculations of the influence of fluctuations of doping on avalanche breakdown.

G_p/ω curves will be compared with the measured G_p/ω curves to see whether the model agrees with observation. For simplicity, we shall derive in detail only an expression for the real part of the interface state admittance. The steps in deriving the imaginary part are similar.

Let $P(N)$ be the probability that there are N built-in charges and charged interface states in a characteristic area. The number of characteristic areas which contain between N and $N+dN$ surface charges is

$$dv = P(N)\, dN. \tag{23}$$

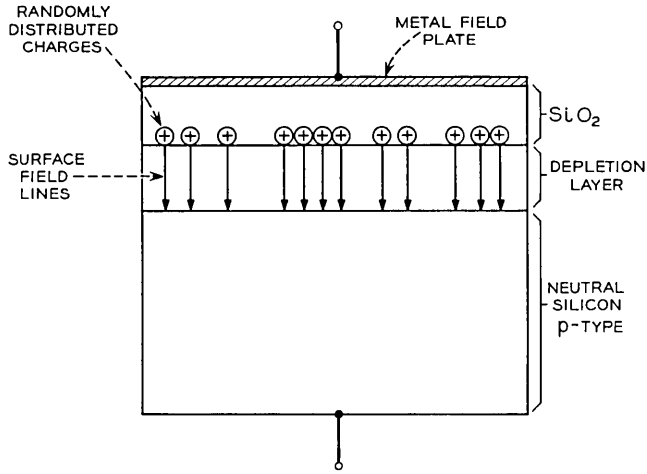


Fig. 7 — Cross section of MIS structure illustrating the random distribution of built-in charges and charged interface states over plane of the interface. Field lines are all terminated at edge of depletion layer rather than on ionized acceptor impurities in the layer for simplicity.

The total G_p/ω is G_p/ω for the continuum of levels located in a characteristic area (20) times the number of characteristic areas containing between N and $N+dN$ surface charges (23) integrated over all the characteristic areas. To integrate this product over all the characteristic areas, (20) is expressed as a function of u_s using (22) and $P(N)$ transformed to $P(u_s)$, the probability that the surface potential in a characteristic area is u_s. This transformation will be performed in two steps. The first transformation is

$$P(Q) = P(N)\, dN/dQ\ , \tag{24}*$$

* See Ref. 24 for a discussion of transforming the probability density as a function of one random variable to another.

where Q is the density of built-in charges and charged interface states, coul/cm^2. The second transformation is

$$P(u_s) = P(Q)\, dQ/du_s \,. \tag{25}$$

The number of characteristic areas in which the surface potential is between u_s and $u_s + du_s$ is $P(u_s)du_s$. Multiplying this by (20) and integrating, the total G_p/ω is

$$G_p/\omega = \frac{qN_{ss}}{2} \int_{-\infty}^{\infty} \frac{1}{\omega\tau_m} \ln (1 + \omega^2\tau_m^2) P(u_s)\, du_s \,. \tag{26}$$

It is assumed in (26) that both interface state density and capture probability are uniform over several kT of bandgap energy. That these are reasonable assumptions can be seen from Fig. 14 and 15.

Substituting (22) for p-type* into (26) to get (26) as a function of u_s

$$G_p/\omega = \tfrac{1}{2}qN_{ss} \int_{-\infty}^{\infty} \exp (u_B - u_s - u_0) \cdot \ln \{1 + \exp [2(u_0 + u_s - u_B)]\} P(u_s)\, du_s \,, \tag{27}$$

where $u_0 = \ln \omega/c_p n_i$.

The next task is to obtain $P(u_s)$ from $P(N)$. Let $\bar{N}$ be the mean number of built-in charges and charged interface states in a characteristic area. When $\bar{N}$ is large, $P(N)$ is given by the Gaussian approximation of a Poisson distribution

$$P(N) = (2\pi\bar{N})^{-\frac{1}{2}} \exp [-(N - \bar{N})^2/2\bar{N}]. \tag{28}$$

The characteristic area α is the ratio between the mean number of surface charges $\bar{N}$ in α and their mean density $\bar{n}$: $\alpha = \bar{N}/\bar{n}$. From this, the relation between N and Q required for the first transformation is

$$N = \alpha Q/q. \tag{29}$$

Combining (24), (28), and (29), we get

$$P(Q) = (2\pi\alpha\bar{Q}/q)^{-\frac{1}{2}}(\alpha/q) \exp [-\alpha(Q - \bar{Q})^2/2q\bar{Q}], \tag{30}$$

where $\bar{Q}$ is the mean of Q.

The transformation of $P(Q)$ to $P(u_s)$ consists of transforming an area in the space made up of all the points Q into an area in the space made up of all the points u_s. This requires a single-valued relation between

* This part of the derivation is carried out for p-type because the sign of the variables for n-type is negative making the analysis harder to follow.

Q and u_s. A relation between Q and u_s which is not single valued is obtained from (7) and (13). Retaining only the dc terms in (13) and expressing surface potential in units of $kT/q = \beta^{-1}$, we get

$$Q = Q_s + Q_f = C_{ox}(v_o + u_s/\beta) + Q_{sc}(u_s). \tag{31}*$$

$Q_{sc}(u_s)$ is the space-charge density which from Ref. 16 is

$$Q_{sc}(u_s) = (2q\epsilon_{si}N_A/\beta)^{\frac{1}{2}}[\exp(-u_s) + u_s - 1]^{\frac{1}{2}}, \tag{32}$$

where ϵ_{si} is the dielectric permittivity of silicon, farads/cm and N_A the ionized acceptor density, cm^{-3}. Equation (31) is not a single-valued relationship between Q and u_s because of (32). The difficulty in making the transformation from Q space to u_s space imposed by (31) and (32) is avoided by restricting the problem to the case where the fluctuations $Q - \bar{Q}$ are very small. For small fluctuations, (31) can be differentiated assuming N_A and oxide thickness to be uniform to get

$$dQ = \frac{C_{ox}}{\beta} du_s + dQ_{sc}. \tag{33}$$

dQ_{sc} is eliminated from (33) by using the relation $dQ_{sc} = (C_D/\beta)\, du_s = (\epsilon_{si}/W\beta)\, du_s$ where W is the space-charge width. Because $\bar{Q}$ is given by (31) when $u_s = \bar{u}_s$, dQ can be evaluated about $\bar{Q}$ at each bias by setting $W = W(\bar{u}_s)$. Doing this, (33) becomes

$$dQ = \frac{1}{W(\bar{u}_s)\beta}[W(\bar{u}_s)C_{ox} + \epsilon_{si}]\, du_s. \tag{34a}$$

Equation (34a) is the transformation equation we seek from an element of area in Q space to the corresponding element in u_s space. Replacing dQ and du_s in (34a) by the small fluctuations $Q - \bar{Q}$ and $u_s - \bar{u}_s$, respectively, we get

$$Q - \bar{Q} = \frac{1}{W(\bar{u}_s)\beta}[W(\bar{u}_s)C_{ox} + \epsilon_{si}](u_s - \bar{u}_s). \tag{34b}$$

Combining (25), (30), (34a), and (34b), we get

$$P(u_s) = (2\pi\sigma_s^2)^{-\frac{1}{2}} \exp[-(u_s - \bar{u}_s)^2/2\sigma_s^2], \tag{35}$$

where the standard deviation σ_s is

$$\sigma_s = \frac{W(\bar{u}_s)\beta}{[W(\bar{u}_s)C_{ox} + \epsilon_{si}]}\left(\frac{q\bar{Q}}{\alpha}\right)^{\frac{1}{2}}. \tag{36}$$

*In this paper, $Q_f + Q_s$ will be calculated from flat band voltage shift neglecting work function differences. Work function differences are neglected because flat band voltage shifts of several volts independent of the field plate metal are found in the samples used.

$\bar{Q}$ can be calculated from (31) and (32) at $u_s = \bar{u}_s$. In the samples measured, the total number of built-in charges was more than an order of magnitude greater than the total number of charged interface states. Therefore, the bias dependence and the variation during each cycle of the signal of the number of charged interface states is negligible.

Another cause of surface potential fluctuations which shall be taken into account is the random distribution of ionized acceptors in the space charge region. This is calculated in Appendix B using a model which has previously been used to determine the fluctuations of breakdown voltage.[23,25] Essentially, the procedure is similar to the one for the surface charges except a characteristic cube having a side equal to W is used instead of a characteristic area and Q rather than N_A is assumed to be uniformly distributed. The standard deviation from Appendix B, paragraph B.1 is

$$\sigma_B = \frac{q\beta[\bar{N}_A W(\bar{u}_s)]^{\frac{1}{2}}[1 - \exp(-\bar{u}_s)]}{2[W(\bar{u}_s)C_{ox} + \epsilon_{si}]}, \tag{37}$$

where $\bar{N}_A$ is the mean ionized acceptor density.

A theorem in statistics states[26] that if two or more independent variables are normally distributed their sum is also normally distributed with a mean and variance which is the sum of the mean and variance of each variable. Using this theorem, $P(u_s)$, which includes the influence of both the randomly distributed surface charges and ionized acceptors on the fluctuations of surface potential is

$$P(u_s) = [2\pi(\sigma_s^2 + \sigma_B^2)]^{-\frac{1}{2}} \exp[-(u_s - \bar{u}_s)^2/2(\sigma_s^2 + \sigma_B^2)]. \tag{38}$$

Substituting (38) into (27)

$$G_p/\omega = \tfrac{1}{2}qN_{ss}[2\pi(\sigma_s^2 + \sigma_B^2)]^{-\frac{1}{2}} \cdot \int_{-\infty}^{\infty} \exp[-(z + y)] \ln(1 + e^{2v})\, du_s, \tag{39}$$

where $y = \ln \omega\tau_m = u_0 + u_s - u_B$ and $z = (u_s - \bar{u}_s)^2/2(\sigma_s^2 + \sigma_B^2)$.

By similar arguments using (19), the equivalent parallel capacitance is

$$C_p = C_D(\bar{u}_s) + qN_{ss}[2\pi(\sigma_s^2 + \sigma_B^2)]^{-\frac{1}{2}} \cdot \int_{-\infty}^{\infty} \exp[-(z + y)] \arctan(e^v)\, du_s. \tag{40}$$

Depletion layer capacitance C_D now becomes a function of $\bar{u}_s$. The function $C_D(\bar{u}_s)$ has the same form as in Ref. 16 as long as the fluctuations $Q - \bar{Q}$ are small.

It will be shown in Section VIII that the influence of the randomly distributed ionized acceptors on the fluctuations of surface potential are small in the samples measured compared to the influence of built-in charges.

In summarizing the theory for the depletion region, it is shown how the interface state branch of the equivalent circuit is developed. The interface state branch of the equivalent circuit of Fig. 6 applies to a single level in the continuum. In each characteristic area where the surface potential is uniform, each level contributes a series RC network. Integrating the admittance of each level given by (4) over all the levels located in a characteristic area gives (19) and (20). Integrating again but this time the contribution to the total admittance from each characteristic area over all the characteristic areas gives (39) and (40). The interface state branch of the equivalent circuit therefore, can be represented by an infinite number of series RC networks connected in parallel as illustrated in Fig. 8(a).

4.6 *Weak Inversion*

For n-type, the silicon surface will be in weak inversion when the Fermi level is more than a few kT/q below mid-gap in the lower half of the bandgap. The minority carrier density at the silicon surface will now be orders of magnitude greater than the majority carrier density

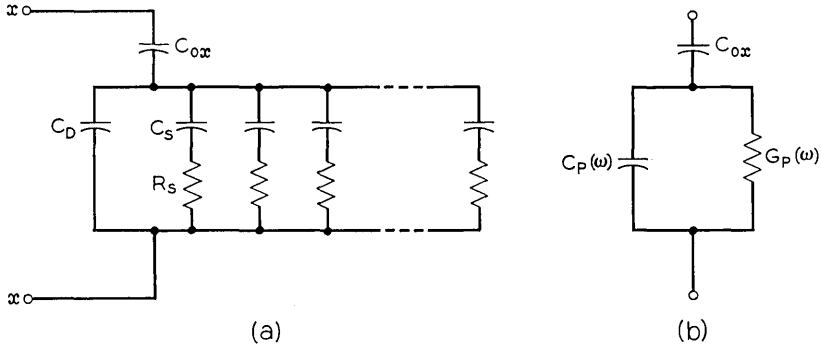


Fig. 8 — (a) Equivalent circuit for depletion region showing time constant dispersion caused primarily by statistical fluctuations of surface potential. C_{ox} is the oxide capacitance per unit area, and C_D the depletion layer capacitance per unit area. Each subnetwork consisting of C_s and R_s in series represents time constant of the continuum of interface states in a characteristic area. Bulk generation-recombination is zero. (b) Simplified version of (a) useful when extracting C_{ox} from the measured admittance. $C_p(\omega)$ is the capacitance per unit area at a given bias and frequency of the distributed network in parallel with C_D. $G_p(\omega)$ is the equivalent parallel conductance per unit area.

at the silicon surface. The minority carriers can no longer be ignored as in depletion and the equivalent circuit of Fig. 8(a) does not apply. To derive expressions for the interface state admittance we start again by considering a single level state. The capacitance associated with this state is $C_s = q^2 N_s f_o(1 - f_o)/kT$ and its time constant for majority carrier transitions (electrons) is $\tau_{n,s} = f_o/c_n n_{so}$ as derived previously in paragraph 4.2. The majority carrier capture resistance from this is

$$R_{n,s} = \tau_{n,s}/C_s = (kT/q^2)[c_n n_{so} N_s(1 - f_o)]^{-1}. \tag{41}$$

By similar reasoning, it can be shown that the minority carrier (holes) capture resistance is

$$R_{p,s} = (kT/q^2)[c_p p_{so} N_s f_0]^{-1}, \tag{42}$$

where c_p is the capture probability for holes cm^3/sec and p_{so} is the hole density cm^{-3} at the silicon surface when the Fermi level is at the interface state level.

The equivalent circuit for a single level is given in Fig. 9(a). From (41) and (42)

$$R_{n,s}/R_{p,s} = (c_p/c_n) \exp (u_F - 3u). \tag{43}$$

Equation (43) has been obtained by substituting $p_{so} = n_i \exp (-u)$, $n_{so} = n_i \exp (u)$, and $f_o = [1 + \exp (u - u_F)]^{-1}$ into (41) and (42). u is the potential difference between mid-gap and the interface state level in units of kT/q and u_F is the potential difference between mid-gap and the Fermi level at the silicon surface in units of kT/q. There will be no detectable loss due to transitions of electrons and holes to and from interface states unless $u \approx u_F$ because $f_o(1 - f_o)$ is so sharply peaked around the Fermi level. Equation (43) becomes

$$R_{n,s}/R_{p,s} = (c_p/c_n) \exp (-2u_F). \tag{44}$$

It is seen from (44) that $R_{n,s}$ will rapidly become orders of magnitude greater than $R_{p,s}$ when the Fermi level gets more than a few kT/q below mid-gap even if c_p is one or two orders of magnitude smaller than c_n. Time constant for majority carrier transitions is $\tau_{n,s} = C_s R_{n,s}$, and for minority carrier transitions it is $\tau_{p,s} = C_s R_{p,s}$. Thus, $\tau_{n,s}$ will be much greater than $\tau_{p,s}$ in weak inversion. If the period of the applied signal frequency is comparable to $\tau_{n,s}$ as it is in these experiments, then there will be virtually no loss associated with minority carrier transitions. Each interface state near the Fermi level is charged and discharged instantaneously by minority carriers so that $R_{p,s} \approx 0$.

Because $R_{p,s}$ is negligibly small, it can be replaced by a short circuit

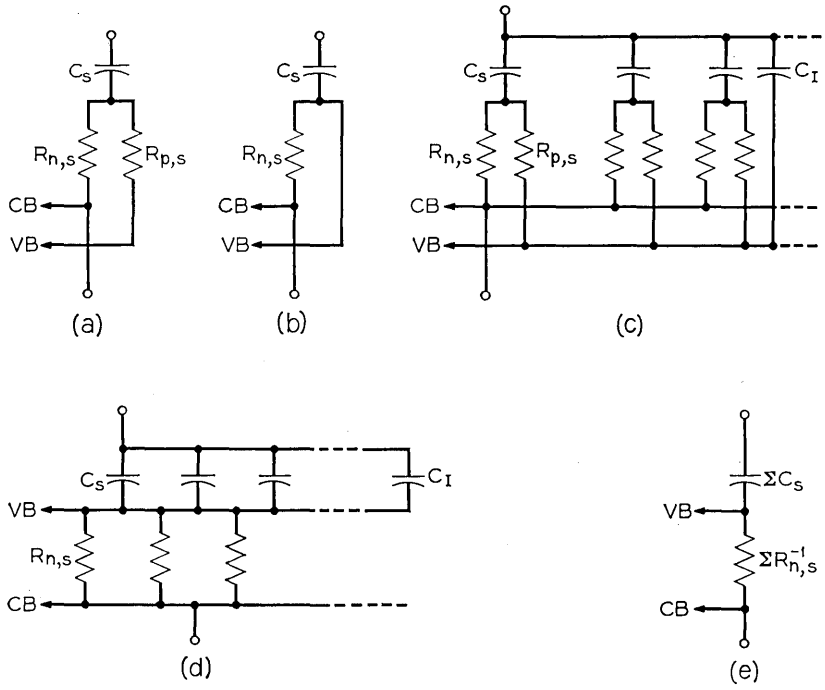


Fig. 9 — (a) Equivalent circuit for a single level interface state in region of weak inversion. C_s is the interface state capacitance per unit area, $R_{n,s}$ the capture resistance for electrons, and $R_{p,s}$ the capture resistance for holes. CB represents the conduction band and VB the valence band. (b) Equivalent circuit for a single level interface state for the case in which minority carriers are holes which respond instantaneously to the applied signal. For this representation, it is also necessary that the lateral conductance due to minority carriers along the surface is large compared to $R_{n,s}^{-1}$. (c) Circuit (a) generalized to include the fact that there is a continuum of states across the bandgap of silicon. C_I is the inversion layer capacitance per unit area. (d) Circuit (c) simplified by fast response of minority carriers. C_I is negligible in the region of weak inversion. (e) Simplification of circuit (d).

which connects the interface state capacitance directly to the minority carrier (valence) band as shown in Fig. 9(b). Those interface states in the continuum which are located in a given characteristic area each have a different time constant. These interface states must be represented by a distributed network such as Fig. 9(c). Only those levels in a narrow range of energy about a kT wide centered about the Fermi level participate in the capture and emission process. Therefore, the corresponding range of time constants is narrow. The large minority carrier density at the surface then means that the minority carrier

capture resistance $R_{p,s}$ of every one of these levels is negligibly small. Thus, the capacitance corresponding to each of these levels can be connected directly to the minority carrier band. The large inversion layer conductance shorts all the characteristic areas together. The net result is represented in Fig. 9(d) which can be further simplified to Fig. 9(e). G_p/ω vs ω now will be characterized by a single time constant at each bias. The expression for G_p/ω is (17) and for C_p it is (16). The value of C_s in (16) and (17) is the sum of all the individual interface state capacitances in parallel as shown in Fig. 9(c). The inversion layer capacitance C_I calculated in weak inversion using the relation given in Ref. 12 is negligible compared to ΣC_s obtained from measurement.

If interface state density N_{ss} is slowly varying with the position of the Fermi level, the statistical fluctuations of surface potential will have negligible effect on the total C_s. However, in analyzing the measurements on n-type samples in weak inversion using (17), it is found that C_s varies quite rapidly with position of the Fermi level (see Fig. 27). This implies that N_{ss} is also rapidly varying so that the statistical fluctuations of surface potential must be taken into account when calculating the relation between N_{ss} and total C_s. Total capture conductance is the sum of the capture conductances $R_{n,s}^{-1}$ of the individual interface states in parallel as shown in Fig. 9(e).

The statistical fluctuations of surface potential must be taken into account in calculating the total capture conductance because capture conductance depends on n_{so} as well as the rapidly varying N_{ss}. If desired, C_s can be calculated by integrating C_s for a single level over bandgap energy and then over surface potential using the statistical model developed earlier. Similarly, total capture conductance can be calculated by integrating capture conductance for a single level obtained from (41) over bandgap energy and then over surface potential. These integrals cannot be evaluated without knowing the functional dependence of N_{ss} on the position of the Fermi level. This dependence can be found by numerically fitting the calculations to the measurements. Although N_{ss} and majority carrier capture probability can be obtained in this way from measured C_s and τ_m, such an analysis will not be carried out here.

A comparison between the depletion and weak inversion regions will illustrate the processes occurring. In depletion except very near mid-gap, majority carrier density at the silicon surface is several orders of magnitude greater than minority carrier density at the silicon surface. Therefore, the time constant for minority carrier transitions

is much greater than the time constant for majority carrier transitions. The time constant for majority carrier transitions is comparable to the period of the applied signal frequency. Therefore, minority carrier density cannot follow the signal at all and ac current flows simply by transitions between interface states near the Fermi level and the majority carrier band. Thus, the interface state branch of the equivalent circuit in Fig. 8(a) can be obtained by open circuiting $R_{p,s}$ in Fig. 9(c).

In the weak inversion region except very near mid-gap, again the time constant for majority carrier transitions is comparable to the period of the signal frequency and several orders of magnitude greater than the time constant for minority carrier transitions. Now *both* majority and minority carrier densities respond to the signal so that ac current flows by generation and recombination through interface states near the Fermi level. This current flow is controlled by the majority carrier capture resistance as it is the largest. In this range, the interface states can easily communicate with the minority carrier band but the minority carrier band has no connection with the bulk. Thus, the main effect minority carriers have in the equivalent circuit is to tie all $R_{n,s}C_s$ together as shown in Fig. 9(d), thereby eliminating the time constant dispersion.

In the region around mid-gap where majority and minority carrier densities at the silicon surface are comparable, no simplification is possible and the equivalent circuit of Fig. 9(c) holds. This region is only a few kT/q wide and no analysis was attempted of measurements in this region.

V. MEASUREMENT APPARATUS, REPRODUCIBILITY AND SAMPLE DRIFT

5.1 *Measurement Apparatus*

The measurement of the equivalent parallel conductance of an MIS capacitor for the purpose of extracting interface state properties requires consideration of the following points which are much more important in conductance measurements than in capacitance measurements alone.

(*i*) Harmonics of the signal frequency due to the non-linearity of the charge-voltage characteristic of the MIS capacitor can give rise to a conductance. To insure that this is negligible and the conductance measured is due to interface state transitions only, the maximum swing of surface potential caused by the applied ac signal should be less than kT/q volts.

(*ii*) The conductance measurements reported here are in the range

between a nmho and about 10 μ mhos which corresponds to very small loss angles. Such small conductance values require that current leakage paths along the air-oxide interface be minimized.

(iii) To verify the theory developed in the previous section, capacitance, equivalent parallel conductance, and bias because of the exponential dependence in (22) must be measured very accurately.

The slices with the completed MIS capacitors were placed on a brass pedestal capped with platinum in a light-tight grounded box for measurement. A steady stream of dry nitrogen was supplied to the box. This keeps the oxide surface dry thereby preventing the formation of a current path for the ac signal along the oxide surface from the field plate to the back contact. It is especially important when making conductance measurements that such current paths be eliminated.

To avoid scratching the field plate and thereby reducing its area, a gold wire, 125 μ in diameter, mounted on a micromanipulator was used to make contact to the field plate.

For high sensitivity, it is necessary to measure accurately conductance when the loss angle is very small. The most suitable instrument for doing this over a wide range of frequencies is a capacitance bridge. The General Radio 1615-A capacitance bridge was found satisfactory over the frequency range from 50 Hz to 20 kHz and the Boonton 75-C capacitance bridge from 5 kHz to 500 kHz. Three terminal capacitance measurements were made with these bridges. Fig. 10 is a block diagram showing the arrangement.

There are three parameters which require special attention in these measurements. They are *(i)* frequency, *(ii)* signal amplitude, and *(iii)* dc bias.

5.1.1 *Frequency*

Frequency must be measured with a counter so that the precision in the calculated admittance of interface states and silicon is limited only by the precision of the bridges. The calibration marks on the oscillator dial in either the Boonton bridge or the hp signal generator are not accurate enough to obtain frequency (see Section VI).

5.1.2 *Signal Amplitude*

To keep harmonics of the signal frequency from giving rise to a spurious conductance, only signals of small amplitude can be applied. The small signal range is determined by experiment. Fig. 11 is a plot of measured capacitance and equivalent parallel conductance

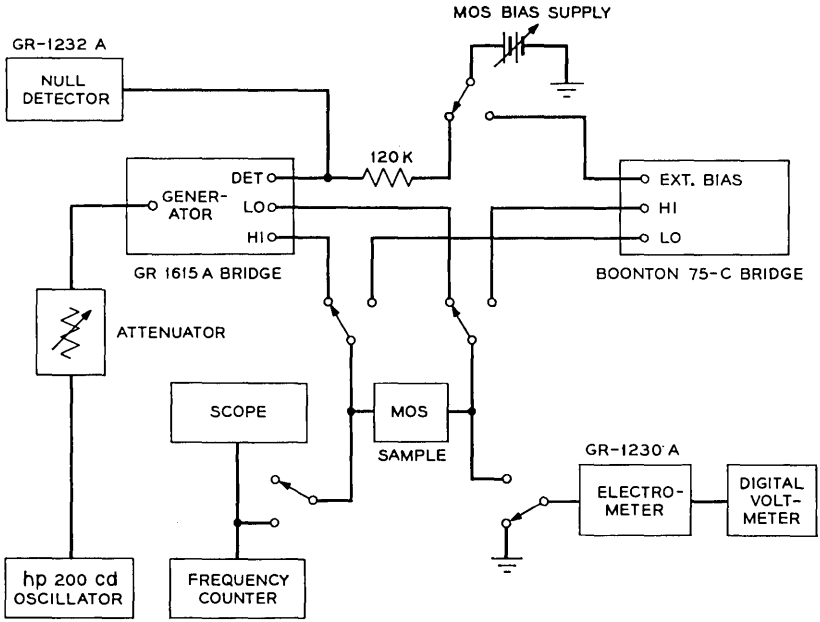


Fig. 10 — Block diagram of measuring apparatus.

normalized to their largest values as functions of peak-to-peak signal amplitude at a bias corresponding to peak conductance at 100 Hz, measured at a point where the slope of the MIS capacitance is large. There is a range of low signal amplitude at which both capacitance and equivalent parallel conductance are independent of signal amplitude. This is the experimentally determined small signal range. The maximum amplitude of the ac silicon surface potential produced by this signal is about two fifths the applied signal amplitude.

To see this, note that the applied signal divides between the oxide layer and the silicon. Expressing this mathematically, substitute (12) into (15):

$$\delta v_a = \delta\psi_s\left[1 + \frac{j\omega C_D + Y_s}{j\omega C_{ox}}\right]. \tag{45}$$

For a fixed signal amplitude, the amplitude of the silicon surface potential varies with bias because C_D and Y_s are functions of bias. Maximum silicon surface potential amplitude occurs when the Fermi level is near mid-gap because that is where C_D and $|\ Y_s\ |$ are smallest. The reason $|\ Y_s\ |$ is small is that the majority carrier density is small

making the interface state time constant long. Thus, to a good approximation $|\omega(C_{ox}+C_D)| > |Y_s|$ so that (45) becomes

$$\frac{\delta\psi_s}{\delta v_a} = \frac{C_{ox}}{C_{ox} + C_D}. \tag{46}$$

Typical values are: $C_{ox} = 5.73 \times 10^{-8}$ farads/cm^2 and $C_D = 8.60 \times 10^{-8}$ farads/cm^2 which gives $\delta\psi_s/\delta v_a = 2/5$.

At the maximum sensitivity setting of the null amplifier, null is found in the small signal range at a null meter reading of $\approx 1/2$ μA on the Boonton bridge and $\approx 3/4$ μV on the General Radio bridge. These are the null readings determined essentially by the noise in the system. Therefore, harmonics of the signal frequency arising from the nonlinearity of the charge-voltage characteristics of the MIS capacitor are unimportant. However, as signal amplitude is increased beyond this range, nulls are obtained at successively higher null meter readings. This means that the amplitude of the harmonics has become large enough to get through the tuned null amplifier. In this range of signal amplitude, the conductance peak decreases in magnitude and the conductance curve is broadened.

Operating in the small signal range entails a reduction in bridge sensitivity. Therefore, optimum signal amplitude is at the high end of the small signal range. The small signal range is determined by the

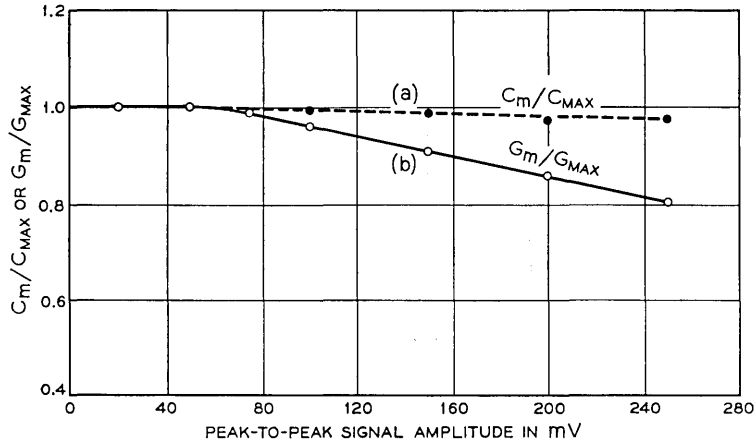


Fig. 11 — (a) Normalized capacitance and (b) normalized equivalent parallel conductance at fixed bias. The maximum value of capacitance is 367.8 pF and conductance 58.8 n mhos. Frequency is 100 Hz. Bias corresponds to peak conductance at this frequency.

oxide thickness and silicon resistivity. This range extends to higher signal amplitudes as the oxide is made thicker or silicon resistivity reduced. For oxide thicknesses between 500 Å and 700 Å and silicon resistivities around 1 ohm cm which is typical of the samples measured in this experiment, optimum peak-to-peak signal amplitude is found to be 50 mV. For other oxide thicknesses and silicon resistivities, optimum peak-to-peak signal amplitude can be found by repeating the measurements shown in Fig. 11. An external attenuator is needed on the hp signal generator to get an amplitude of 50 mV. Signal amplitude is measured with an oscillosocope and adjusted whenever frequency is changed. The maximum peak to peak silicon surface potential with this signal is $2/5 \times 50 = 20$ mV which is less than kT/q at 300°K and therefore small signal.

5.1.3 *DC Bias*

Because interface state time constants are an exponential function of silicon surface potential, dc bias must be measured to at least three decimal places. The Cimron 6900A digital voltmeter is more than adequate for this purpose.

DC bias is supplied to the MIS capacitor through a voltage divider in each bridge. Therefore, it must be measured across the sample rather than at the supply terminals. The voltmeter puts an additional capacitance and conductance across the sample which are read by the bridge. To avoid the correction this entails, the voltmeter is switched out of the circuit before the bridge is balanced. Because the input resistance of the voltmeter no longer shunts the MIS capacitor, the current through the voltage divider in the bridge decreases. Therefore, the voltage across the MIS capacitor is slightly larger than that measured. For the Cimron 6900A digital voltmeter which has an input impedance when balanced of 10^9 ohms, this is not a problem. However, for a digital voltmeter having an input impedance of a few megohms, this correction becomes apparent whenever the resistance values of the internal bridge voltage divider are changed. These changes occur when switching the conductance multiplier on the Boonton bridge, the conductance to loss factors switch on the General Radio bridge or in switching from one bridge to another. To eliminate this error, a General Radio 1230A electrometer which has an input resistance of 10^{15} ohms with the resistance selector set at ∞ can be used to measure bias voltage across the sample. The output of the electrometer is then read by the digital voltmeter.

To permit control of the bias voltage to within a fraction of a

millivolt, a voltage divider consisting of a bank of fixed resistors in series with a 10-turn helipot is used in conjunction with a heavy duty 45-volt battery in a shielded box.

5.2 *Reproducibility of Data and Sample Drift*

Sample preparation is designed to eliminate drift under negative bias at room temperature only. The admittance-voltage characteristics in the range of interest for both n- and p-type samples are located in the negative bias range because of the presence of a fixed residual positive charge in the oxide. Therefore, n-type samples are in strong accumulation and p-type in strong inversion at positive bias. No conductance measurements are made at positive bias because conductance due to interface states is too low to detect in strong accumulation or inversion.

Reproducibility of the measurements on a given MIS capacitor was assured as follows. Measurements are repeated after room temperature bias aging at high negative bias, after the lapse of several days, and by hysteresis measurements where the admittance-voltage characteristics are measured first with increasing bias and then with decreasing bias. Only those samples where the measurements were found to be reproducible within experimental error are used. Interface state properties for such capacitors on a few different slices are found to be similar.

VI. TECHNIQUE FOR ANALYZING THE DATA

6.1 *Properties of Equivalent Parallel Conductance*

The variation of G_p/ω with bias or frequency can be most easily seen from (17) and (22) for a single time constant. In (39) derived from the statistical model which fits the measurements exactly in the depletion-region, (see Fig. 21) the variation of G_p/ω with bias and frequency is obscured by the mathematical complexity. Equation (17) is adequate for illustrative purposes because of its simplicity and the fact that its variation with bias and frequency is qualitatively the same as (39) in the depletion region. Equation (17) fits the data exactly in the weak inversion region (see Fig. 25). G_p or G_p/ω can be obtained from measurements as a function of bias with frequency as parameter or as a function of frequency with bias as parameter. G_p or G_p/ω will go through a peak as a function of bias at each frequency as seen from (17) and (22). For a continuum of states, this peak occurs when the response time of the interface states which is being varied by the bias equals the period of the applied signal

frequency multiplied by a constant. When the frequency is increased, this peak shifts to a bias value corresponding to a surface potential nearer to flat bands. Fig. 1 shows capacitance and equivalent parallel conductance measured at 5 kHz and 100 kHz in a wet oxide. This figure illustrates how measured equivalent parallel conductance varies with bias and frequency. Fig. 1 also shows how inaccurate the capacitance method for extracting information about interface states is compared to the conductance method. It can be noted from Fig. 1 that

(*i*) Maximum capacitance dispersion in this frequency range is about 14 percent while equivalent parallel conductance changes more than an order of magnitude.

(*ii*) The conductance peak moves towards a bias voltage closer to flat band bias which in Fig. 1 is −2.6 volts as frequency is increased. Between 50 Hz and 500 kHz, conductance does not peak at flat band bias in the samples measured.* The peak shifts over the bias range from weak inversion to depletion in this frequency range.

(*iii*) Capacitance and equivalent parallel conductance are related to each other by the Kronig-Kramers relations. Rather than going through a mathematical proof of this, we note that (16), (17), (39), and (40) have the same form as the Debye equations.[19, 20] Because the Debye equations satisfy the Kronig-Kramers relations,[27] we infer that (16), (17), (39), and (40) do also.

G_p/ω as a function of frequency will go through a peak at each value of bias but in this case G_p will not. It can be seen from (17) that only G_p/ω is a symmetric in $\omega\tau$ when frequency is the variable while both G_p and G_p/ω are symmetric in $\omega\tau$ when bias is the variable.

G_p/ω vs *bias* (with frequency as parameter) will be spread over a bias range determined by the time constant dispersion and the density of interface states. The density of states spreads the G_p/ω vs bias curve in the same way that it spreads the high-frequency capacitance vs bias curve. However, G_p/ω vs ω (with bias as parameter) will be spread over a frequency range determined only by the time constanct dispersion. Because time constant dispersion determines the interface state admittance, G_p/ω vs ω curves are the more useful of the two for the purposes of this work.

6.2 *Extraction of* G_p/ω

It is necessary first to get G_p/ω and C_p from the measurements. C_p and G_p are shown in Fig. 8(b) to be the capacitance and equivalent

* To get the conductance to peak at flat band bias, frequencies higher than 500 kHz have to be applied.

parallel conductance of the portion of the equivalent circuit consisting of C_D in parallel with the distributed network representing the interface states. The admittance measured across the terminals x–x in Fig. 8(a) is $G_m + j\omega C_m$. Converting this to an impedance, subtracting away the reactance of C_{ox}, and converting back to an admittance, we get

$$\frac{G_p}{\omega} = \frac{\omega C_{ox}^2 G_m (G_m^2 + \omega^2 C_m^2)}{\omega^2 C_{ox}^2 G_m^2 + [\omega^2 C_m (C_{ox} - C_m) - G_m^2]^2} \tag{47}*$$

and:

$$C_p = \frac{C_{ox}(G_m^2 + \omega^2 C_m^2)[\omega^2 C_m (C_{ox} - C_m) - G_m^2]}{\omega^2 C_{ox}^2 G_m^2 + [\omega^2 C_m (C_{ox} - C_m) - G_m^2]^2}. \tag{48}*$$

The measured G_p/ω from (47) is related to N_{ss} and τ_m by (39) and C_p from (48) is related to C_D, N_{ss} and τ_m by (40) in the depletion region.

It should be noted that the circuit across the terminals x–x in Fig. 8(a) has a shorter time constant than that of just the interface state branch of the circuit because of C_{ox}. Therefore, G_p/ω from (47) will peak at a lower frequency than G_p/ω measured across terminals x–x when equivalent parallel conductance is measured as a function of frequency with bias as parameter. If equivalent parallel conductance is measured as a function of bias with frequency as parameter, G_p/ω from (47) will peak at a bias close to flat bands for the same reason as can be seen from (22).

To minimize errors introduced by correcting for C_{ox}, the oxide should be made as thin as practicable and frequency measured accurately with a counter as described in Section V.

6.3 *Bias vs Surface Potential*

From (22), there is an exponential relation between interface state time constant and surface potential. Therefore, bias is measured with a digital voltmeter as described in Section V and the relation between bias and surface potential is determined as accurately as possible. It is necessary to do this to prove that (22) fits the measurements and to extract c_n and c_p accurately. One of the most accurate methods for obtaining the relation between bias and surface potential has been developed by C. N. Berglund.[28] In this method, capacitance is measured as a function of bias at such a low frequency that the interface

* These equations were evaluated using the measured values of ω, C_{ox}, C_m, and G_m on an IBM 1620 computer.

states are in equilibrium with the silicon. This means that the loss associated with the capture and emission of carriers by the interface states is negligible.

Bias vs surface potential is obtained in two steps from the measured data. First, surface potential is found to within an additive constant at each bias. Then, the additive constant is found.

6.3.1 *Surface Potential within an Additive Constant*

Low-frequency capacitance C_{om} measured at bias v_o is

$$C_{om} = dQ_T/dv_o \, . \tag{49}$$

Differentiating (13) with respect to v_o

$$1 = d\psi_s/dv_o + (1/C_{ox})\, dQ_T/dv_o \, . \tag{50}$$

From (49) and (50)

$$d\psi_s/dv_o = 1 - C_{om}/C_{ox} \, . \tag{51}$$

The surface potential at any bias v_{o1} is found by integrating (51)

$$\psi_s(v_{o1}) = \int_{v_{o2}}^{v_{o1}} (1 - C_{om}/C_{ox})\, dv_o + \Delta, \tag{52}$$

Where Δ is an additive constant and v_{o2} is a bias in strong accumulation.

Thus, surface potential at each bias can be obtained to within an additive constant directly from measured data by an integration. The fact that an integration is carried out using directly measured quantities makes this method accurate.

It is found that in samples oxidized in steam as described in Section II, 50 Hz is a low enough frequency to measure C_{om} vs v_o and use (52) with negligible error. This is seen by the observation that the loss tangent is negligibly small in the bias range investigated here below about 80 Hz so that dispersion in the capacitance-voltage characteristic virtually disappears in this frequency range.

Another important consideration is that (52) will not give an accurate result if there are any *gross* nonuniformities. Gross nonuniformities occur when there are macroscopic areas under the field plate in which the charge density is significantly different from adjacent areas. Whether or not there are such nonuniformities present in a sample can be determined by evaluating (52) using C_{om} measured from strong accumulation to strong inversion. The resulting value of

ψ_s will be slightly less than the bandgap,* if there are no gross nonuniformities and considerably higher if there are.

The random distribution of built-in charges and charged interface states does not affect the accuracy of (52). Because of the statistical fluctuations of surface potential caused by this random charge distribution, bias is really found as a function of mean surface potential.

6.3.2 *Determination of Additive Constant*

The additive constant Δ in (52) and the doping density is calculated next. For this purpose, C_D must be found as a function of bias in the depletion region. Then C_D^{-2} is plotted as a function of $\psi_{se} - kT/q$ where

$$\psi_{se} = \int_{v_{o2}}^{v_{o1}} (1 - C_{om}/C_{ox})\, dv_a \,.$$

G_p/ω from (47) peaks at bias values within the depletion region at frequencies above a few hundred Hz. C_D is found from admittance measured at these frequencies using (48) to get first C_p. Then from (40)

$$C_p = C_D + C_s(\omega, \bar{u}_s), \tag{53}$$

where $C_s(\omega, \bar{u}_s)$ is the equivalent capacitance of the distributed network representing the interface states in Fig. 8(a) at the frequency and bias corresponding to $(G_p/\omega)_{\max}$ of the measured curve obtained using (47). It is shown in Appendix C that $C_s(\omega, \bar{u}_s) = 2(G_p/\omega)_{\max}$ where $(G_p/\omega)_{\max}$ is the measured peak. Then, using (53) we get

$$C_D = C_p - 2(G_p/\omega)_{\max} \,. \tag{54}$$

In the depletion region except near flat bands, C_D can be written in the following form using the approximate relation between C_D and $\bar{u}_s$ [16]

$$C_D^{-2} = 2(\bar{u}_s - 1)/q\epsilon_{si}N_A\beta, \tag{55}$$

where ϵ_{si} is the dielectric permittivity of silicon. Thus, a plot of C_D^{-2} measured far from flat bands in the depletion region vs $\bar{\psi}_{se} - kT/q$ measured in the same region will yield a straight line as shown in Fig. 12. The intercept of this line with the abcissa when it is extrapolated to $C_D^{-2} = 0$ gives Δ. From this, bias is found as a function of surface potential. Finally, the slope of the line in Fig. 12 gives the doping density from (55).

* This is particularly the case for samples having high doping density such as those used in this work.

6.4 N_{ss} and τ_m in Depletion

Equation (39) for G_p/ω from the statistical model is found to fit the measured G_p/ω vs frequency curves accurately at each bias (see Section VII). To prove this, (39) is numerically integrated on an IBM 7094 computer using the trapezoidal rule.* $\bar{u}_s$ corresponding to the particular bias at which G_p/ω vs frequency is measured is obtained from the relation between bias and surface potential just found. The characteristic area α is the only parameter which is varied to obtain the best fit to the measured curve. N_{ss} can be obtained independently.

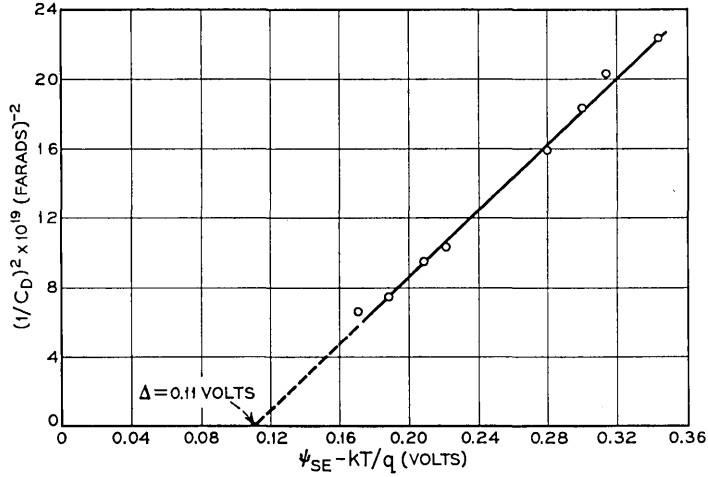


Fig. 12 — $1/C_D^2$ vs ψ_{se}—kT/q where C_D in farads is obtained from measurements in the depletion region of an n-type sample ($N_D = 1.15 \times 10^{15}$ cm^{-3}) and ψ_{se} from the integral in (52). Field plate diameter is 3.8×10^{-2} cm.

The parameter α is not arbitrary. It is found that $\alpha^{\frac{1}{2}}$ has about twice the value and the same bias dependence as the silicon depletion layer width. (see Fig. 22).

Fig. 13(a) shows G_p/ω found by integrating (39) plotted as a function of $\omega\tau_m$ with N_{ss} arbitrarily chosen as 1×10^{11} cm^{-2} eV^{-1} and α at a bias midway between flat bands and mid-gap. For comparison, Fig. 13(b) is a plot of (17) and Fig. 13(c) is a plot of (20) vs $\omega\tau_m$ for the same value of N_{ss}. From Fig. 13(a), it is seen that the maximum value of G_p/ω from (39) occurs when $\omega\tau_m = 2.5$. This was found to hold

* To do this, limits of integration in (39) must be finite. Because the fluctuations of surface charge are small, integrating from $u_s = -2u_B$ to $u_s = 2u_B$ will give a good approximation to the integral from $-\infty$ to ∞.

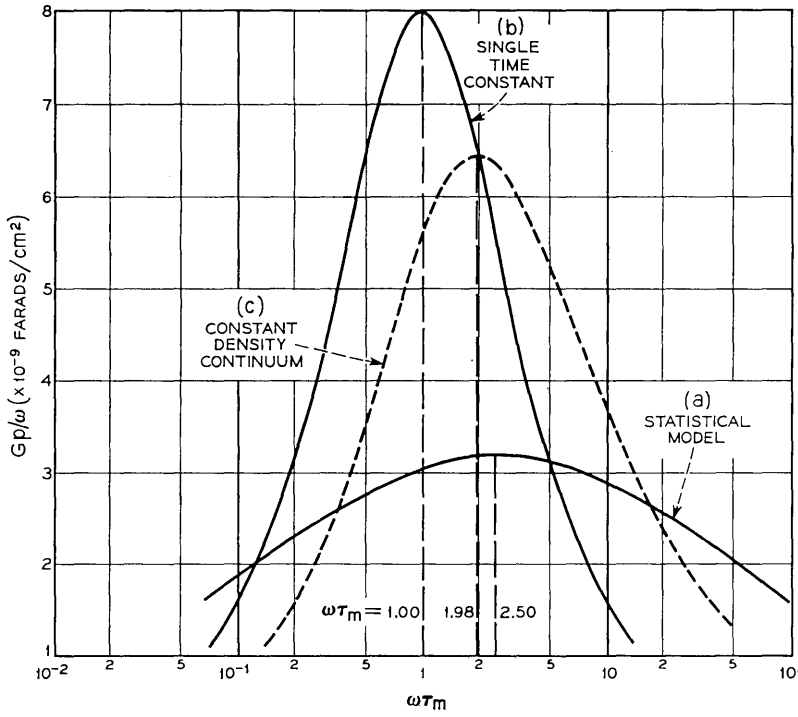


Fig. 13 — Calculated G_p/ω vs log $\omega\tau_m$. (a) Plot of (39) incorporating time constant broadening from statistical fluctuations of surface potential as well as continuum of states. $\alpha \doteq 2.5 \times 10^{-10}$ cm^2 and impurity density $= 2.1 \times 10^{16}$ cm^{-3} are used. (b) Plot of (20) calculated by Lehovec in Ref. (19) for a continuum of constant density and capture cross section. The same interface state density arbitrarily chosen as 1×10^{11} cm^{-2}-eV^{-1} is used in calculating all the curves.

at all values of bias measured. The condition $\omega\tau_m = 2.5$ is used to find τ_m from the frequency at which the measured G_p/ω vs frequency curve goes through a maximum at each bias. The τ_m found this way corresponds to the mean surface potential $\bar{u}_s$ at each bias. The condition $\omega\tau_m = 2.5$ can be used to get τ_m when G_p/ω is measured as a function of bias with frequency as parameter. In this case, ω is given by the fixed frequency but the τ_m obtained now corresponds to the bias at peak G_p/ω.

Interface state density vs bias is found by numerically integrating (39) with $\omega\tau_m = 2.5$, equating the result to the measured peak G_p/ω value, and solving for N_{ss} at each bias. The condition $\omega\tau_m = 2.5$ is equivalent to setting $u_0 = u_B - \bar{u}_s + \ln 2.5$ in (39). Doing this, (39)

becomes

$$\left(\frac{G_p}{\omega}\right)_{\max} = \tfrac{1}{2}qN_{ss}[2\pi(\sigma_s^2 + \sigma_B^2)]^{-\frac{1}{2}}I_{\max} \tag{56}$$

and

$$I_{\max} = \int_{-2u_B}^{2u_B} \exp\,[-(z + y_m)]\,\ln\,[1 + \exp\,(2y_m)]\,du_s\,, \tag{57}$$

where

$$y_m = u_s - \bar{u}_s + \ln 2.5 \quad \text{and} \quad z = (u_s - \bar{u}_s)^2/2(\sigma_s^2 + \sigma_B^2).$$

Equation (56) and (57) can be evaluated once α is known. The simplest way to find an approximate value for α is to use the experimentally established relation between $\alpha^{\frac{1}{2}}$ and W where W is the depletion layer width. W can be obtained from the relation

$$W = \frac{\epsilon_{si}}{C_D}\,, \tag{58}$$

where C_D is obtained from (54).

Solving (56) for N_{ss}

$$N_{ss} = 2\left(\frac{G_p}{\omega}\right)_{\max}[2\pi(\sigma_s^2 + \sigma_B^2)]^{\frac{1}{2}}(I_{\max})^{-1}q^{-1}. \tag{59}$$

Substituting (57) into (59) and using values of α from Fig. 22, (59) was evaluated to find N_{ss} as a function of the position of the Fermi level in the bandgap of the silicon.

6.5 *N_{ss} and τ_m in Weak Inversion*

This is a particularly simple case because G_p/ω vs frequency is characterized by a single time constant at each bias. G_p/ω vs frequency or bias curves are obtained using (47) as before. Fig. 6 and (17) apply so that the condition for maximum G_p/ω is $\omega\tau_m = 1$ and the value of $(G_p/\omega)_{\max}$ is $C_s/2$.

6.6 *Capture Probabilities and Cross Sections*

Log τ_m is plotted as a function of $u_B - \bar{u}_s$ from measurements in the depletion region to see if the data fit (22). It is found this way that the measurements fit (22) quite well for both holes and electrons (see Section VII). The fact that the experimental data fit (22) means that c_n and c_p are independent of bandgap energy. Therefore, c_n and c_p are not affected by statistical fluctuations of surface potential and are the true capture probabilities of the interface states. Capture

cross section is related to capture probability by the relation $c_n = \bar{v}\sigma_n$ for electrons where $\bar{v}$ is the average thermal velocity of electrons. c_n and c_p are calculated from experimental plots of (22) and σ_n and σ_p found using $v = 10^7$ cm/sec for both holes and electrons.

VII. EXPERIMENTAL RESULTS

7.1 *Depletion Region*

7.1.1 *Steam-Grown Oxides on [111] Orientation*

Fig. 14 shows log τ_m vs $u_B - \bar{u}_s$ calculated from the measurements as explained in Section VI. The values plotted correspond to peaks

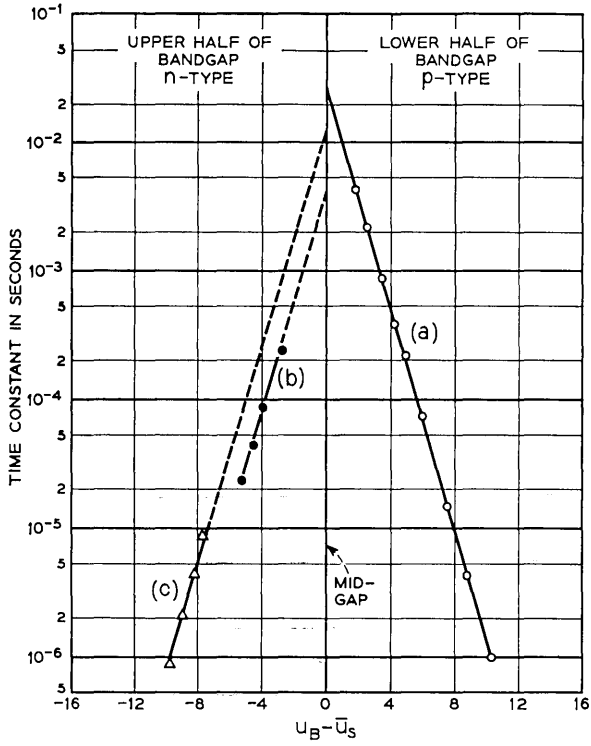


Fig. 14 — Log τ_m vs $u_B - \bar{u}_s$—showing fit of experimentally obtained points to (22) for both holes and electrons using steam grown oxide and [111] orientation. Points in lower half of the gap were measured on a p-type sample. Field plate diameter is 1.2×10^{-1} cm, acceptor density 2.08×10^{16} cm^{-3}, $u_B = 14.1$ and $C_{ox} = 5.74 \times 10^{-8}$ farads/cm^2. Points in upper half of the gap were measured on an n-type sample. Field plate diameter 3.8×10^{-2} cm, donor density 1.15×10^{16} cm^{-3}, $u_B = 13.5$, and $C_{ox} = 5.08 \times 10^{-8}$farads/cm^2.

of the G_p/ω curves obtained from measurements made at 300°K. The condition $\omega\tau_m = 2.5$ is used to get τ_m. Because only majority carrier transitions give rise to measured conductance, curves in the lower half of the silicon bandgap were obtained from measurements on p-type and in the upper half on n-type. It is seen that these curves fit (22) so that capture cross sections are obtained simply by extrapolating the curves to $u_B - \bar{u}_s = 0$ where (22) becomes $\tau_{mo} = (\bar{v}n_i\sigma)^{-1}$. This is then solved for σ to get capture cross section. Capture cross sections obtained this way from Fig. 14 are

σ_p (holes):	2.2×10^{-16} cm^2	from curve (a)
σ_n (electrons):	1.7×10^{-15} cm^2	from curve (b)
σ_n (electrons):	5.9×10^{-16} cm^2	from curve (c).

In all cases, $\bar{v} = 10^7$ cm/sec and $n_i = 1.55 \times 10^{10}$ cm^{-3} have been used.

Fig. 15 shows the corresponding values of N_{ss} as a function of the average position of the Fermi level E_F with respect to mid-gap potential at the silicon surface. Measurements in the lower half of the gap were made on p-type and in the upper half on n-type.

Fig. 16 illustrates the temperature dependence of τ. Curve (a) is calculated from measurements made at 204°K while curve (b) is just a replot of curve (a) from Fig. 14 for comparison.

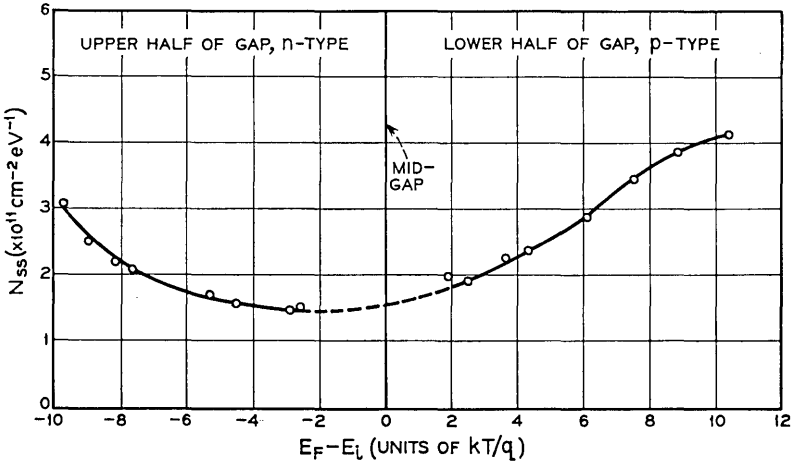


Fig. 15—N_{ss} from (59) vs $E_F - E_i$ measured in units of kT/q for same two samples in Fig. 14. Flat band is 14.1 in lower half of the gap and 13.5 in upper half. E_F is position of the Fermi level and E_i is mid-gap.

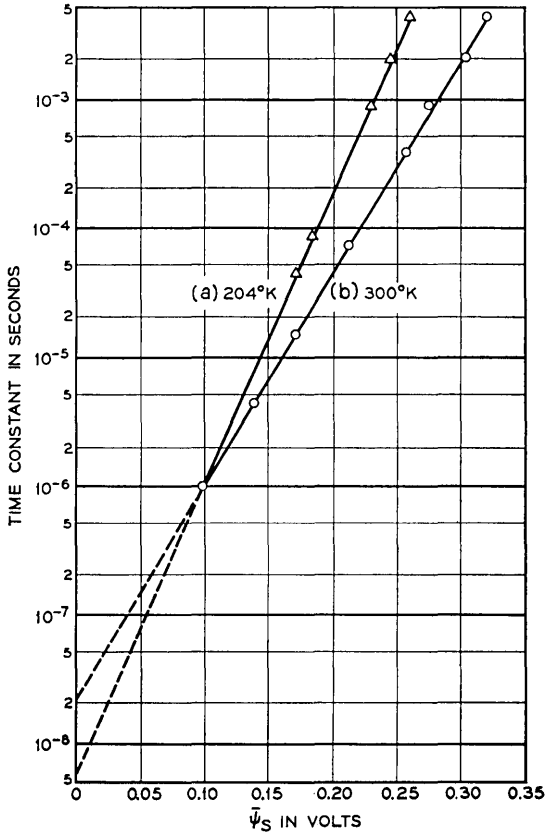


Fig. 16—Log τ_m vs $\bar{\psi}_s$ for p-type sample of Fig. 14: (a) measured at 204 °K and (b) at 300 °K. (b) Replot of (a) in Fig. 14. Curves show that experimentally obtained points satisfy temperature dependence of (22). (a) is the only one in this experiment measured at a temperature other than 300 °K.

7.1.2 *Steam-Grown Oxides on [100] Orientation*

Fig. 17 shows log τ_m vs $u_B - \bar{u}_s$ for the [100] orientation calculated from measurements made at 300°K. Capture cross sections are

σ_p (holes):	2.0×10^{-16} cm^2	from curve (a)
σ_p (holes):	4.0×10^{-16} cm^2	from curve (b)
σ_n (electrons):	1.2×10^{-15} cm^2	from curve (c).

Fig. 18 shows the corresponding values of N_{ss} in a plot similar to Fig. 15.

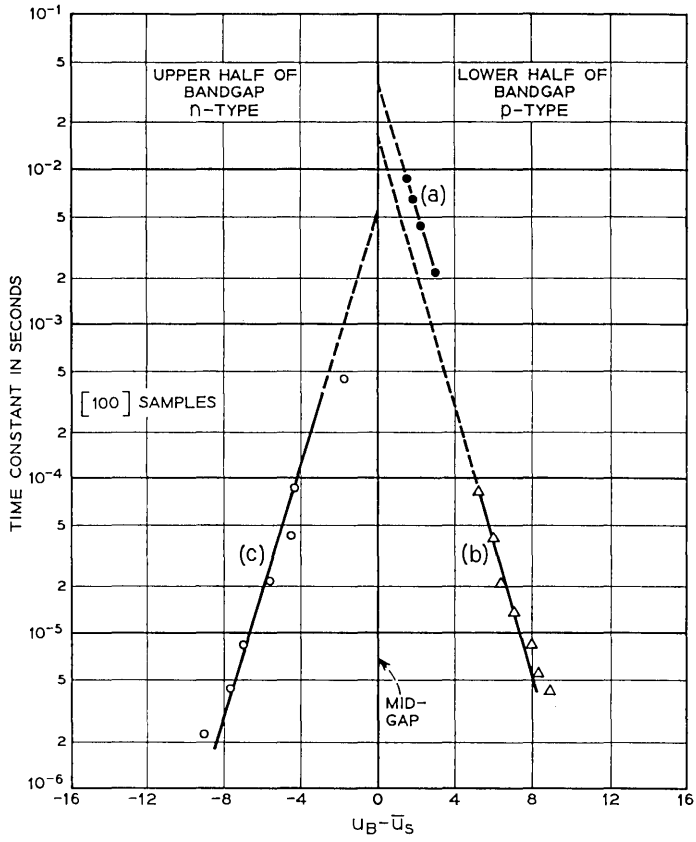


Fig. 17 — Log τ_m vs $u_B - \bar{u}_s$ showing fit of experimentally obtained points to (22) for both holes and electrons using steam-grown oxide and [100] orientation. Points in lower half of gap were measured on a p-type sample. Field plate diameter is 1.37×10^{-1} cm, acceptor density 1.19×10^{16} cm^{-3}, $u_B = 13.5$, and $C_{ox} = 4.94 \times 10^{-8}$farads/cm^2. Points in upper half of gap were measured on an n-type sample. Field plate diameter is 1.43×10^{-1} cm, donor density 1.44×10^{16} cm^{-3}, $u_B = 13.7$, and $C_{ox} = 4.24 \times 10^{-8}$farads/cm^2.

7.1.3 [111] *Orientation, Partly-Dried Oxides*

Fig. 19 shows log τ_m vs $u_B - \bar{u}_s$ for oxides initially grown in steam and partially dried by heating in dry N_2 as described in Section III. Only n-type was analyzed in detail, so that we have

$$\sigma_n \text{ (electrons)}: \quad 4.6 \times 10^{-16} \text{ cm}^2 \quad \text{from Fig. 19.}$$

Curve (a) in Fig. 20 shows the corresponding values of N_{ss}. Curve (b) in Fig. 20 is a replot from Fig. 15 for comparison.

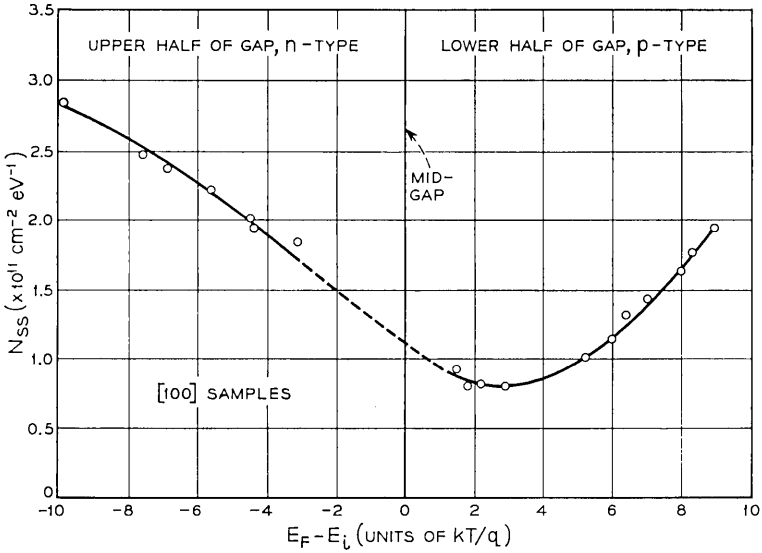


Fig. 18 — N_{ss} from (59) vs $E_F - E_i$ for same two samples in Fig. 17. Flat band is 13.5 in lower half of gap and 13.7 in upper half.

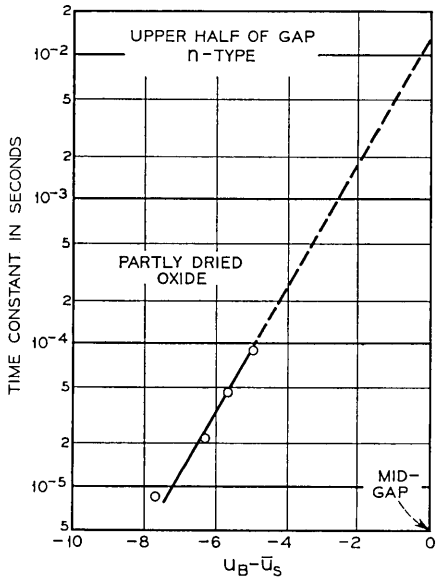


Fig. 19 — Log τ_m vs $u_B - \bar{u}_s$ after partial drying of n-type sample in Fig. 4. Measured points still fit (22). Donor density is 1.4×10^{16} cm^{-3}, $u_B = 13.7$, and $C_{ox} = 5.96 \times 10^{-8}$farads/cm^2.

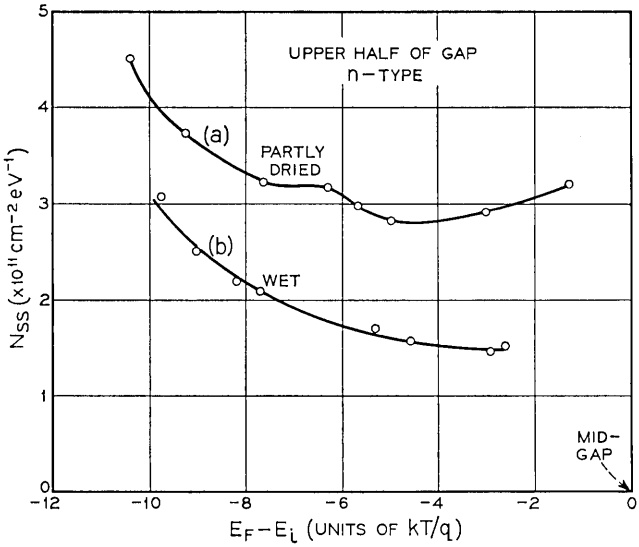


Fig. 20 — N_{ss} from (59) vs $E_F - E_i$. (a) n-type sample in Fig. 19 and (b) replot of n-type sample from Fig. 15 for comparison. For (a), flat band is 13.7 and for (b), flat band is 13.5.

7.1.4 *Range of* N_{ss}

As seen from Figs. 15 and 18, N_{ss} in steam-grown oxides is in the 10^{11} cm^{-2}-eV^{-1} range. Partly drying increases N_{ss} as seen from Fig. 20. Complete drying or growing the oxide in dry oxygen at 1000°C results in N_{ss} values in the 10^{12} cm^{-2}-eV^{-1} range or about an order of magnitude higher than for steam-grown oxides.

N_{ss} can be reduced below the values for steam-grown oxides by an order of magnitude by a method described by Balk.[29] First, an aluminum film serveral thousand Å thick is evaporated over the entire surface of a steam-grown oxide. The sample is then annealed in H_2 or N_2 at 350°C for 1/2 hour. After annealing, field plates are produced using photoresist techniques and chemically etching off the unwanted aluminum. This annealing process is found to reduce N_{ss} from the 10^{11} cm^{-2}-eV^{-1} range characteristic of steam-grown oxides to the 10^{10} cm^{-2}-eV^{-1} range. Thus, N_{ss} can be varied two orders of magnitude from the 10^{12} cm^{-2}-eV^{-1} range characteristic of dry oxides to the 10^{10} cm^{-2}-eV^{-1} range. Capture cross section appears to remain approximately the same throughout.

7.1.5 *Fit to Statistical Model*

The circles in Fig. 21 represent points calculated from measurements on the [111] p-type sample with a steam-grown oxide in Figs. 14 and 15. These points were measured with frequency as variable and bias fixed at -1.9 volts which corresponds to $u_B - \bar{u}_s = 6.04$. The best fit to these points calculated from (39) is the solid curve in Fig. 21. This fit was obtained by substituting (59) for N_{ss} in (39) and using $(G_p/\omega)_{max}$ from the data. The value of α required for this fit was found to be 2.5×10^{-10} cm^2.

Fig. 22 shows $\alpha^{\frac{1}{2}}$ vs space-charge width W. The values of α are obtained by finding the best fit of (39) to G_p/ω vs frequency curves taken at different bias settings on the same sample. Space-charge width is calculated from the depletion layer capacitance obtained as described in Section VI from the measured capacitance and equivalent parallel conductance.

7.2 *Weak Inversion*

The G_p/ω curves obtained after correcting for C_{ox} will peak in the weak inversion range at signal frequencies below about 1 kHz. Curves

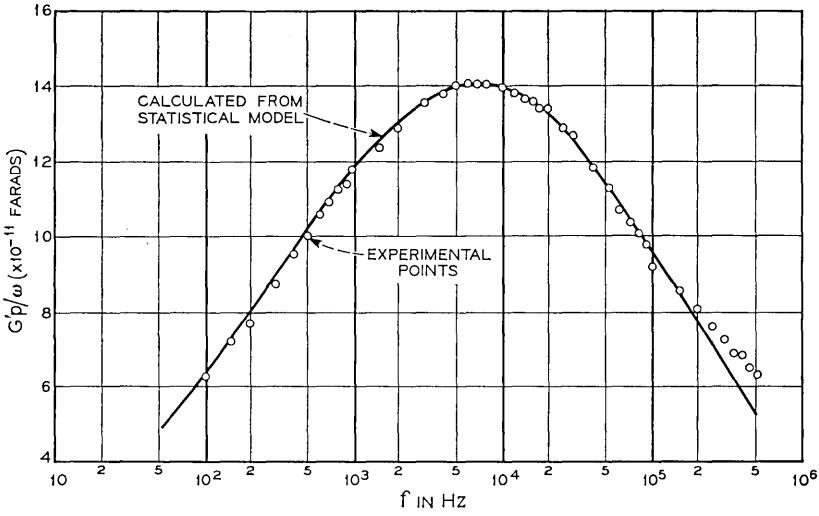


Fig. 21 — G_p'/ω vs frequency showing how experimental points fit (39). An α of 2.5×10^{-10} cm^2 is used in (39) to get this fit. G_p' is equivalent parallel conductance in mhos derived from the measurements. Experimental points are measured at a bias of -1.9 volts ($\bar{u}_s = 8.1$) in depletion region on the p-type sample of Fig. 14.

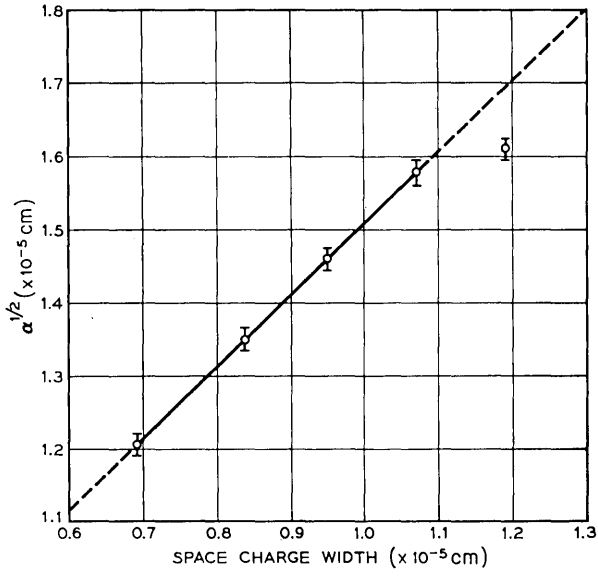


Fig. 22 — $\alpha^{\frac{1}{2}}$ vs space-charge width where α is obtained by finding the best fit of (39) to experimental points taken on p-type sample in Fig. 14. $\alpha^{\frac{1}{2}}$ can have any value within vertical bars through each point. Bars indicate precision with which (39) can be fitted to measured G_p/ω vs frequency points. No reason why the point at the largest W does not lie on the straight line is known.

of capacitance and equivalent parallel conductance vs field plate bias measured at 50 Hz on a [111], n-type sample with steam grown oxide are shown in Fig. 23. An n-type sample was used to avoid the masking effects of lateral ac current flow.[10] The capacitance peak below mid-gap persists down to 1.6 Hz, the lowest frequency used, and disappears at frequencies above 500 Hz. Above 500 Hz, the capacitance curve becomes the usual high-frequency curve.[5] The low-frequency capacitance dispersion is shown in Fig. 24.

G_p/ω vs frequency curves in this range can be fitted with a single time constant as shown in Fig. 25. The circles represent points calculated from measurements on an n-type sample with steam-grown oxide on the [111] orientation. These points were measured with frequency as variable and bias fixed at -3.3 volts which corresponds to $u_B - \bar{u}_s = -6.3$ in weak inversion. The best fit to these points calculated from (17) is the solid curve in Fig. 25. The values of τ and C_s used in (17) to obtain this fit are 1.7×10^{-3} seconds and 5.14×10^{-8} farads/cm^2, respectively.

Fig. 26 is a plot of log τ_m vs $u_B - \bar{u}_s$. These values are calculated

from measurements in the weak inversion region using $\omega\tau_m = 1$ to get τ_m. Fig. 27 is a plot of the corresponding values of C_s as a function of the position of the Fermi level with respect to mid-gap potential at the silicon surface. Values of C_s are obtained from the peak of the G_p/ω curve using $\omega\tau_m = 1$ and (17). Values of N_{ss} using $N_{ss} = C_s/q$ are given on the right-hand ordinate axis of Fig. 27. Although these values are not exact, they give the correct order of magnitude of N_{ss}.

Fig. 28 shows a low-temperature drying experiment similar to those in Figs. 3 and 4 except the curves in Fig. 28 were measured at 50 Hz. Curves (a) and (b) of Fig. 28 are plots of capacitance and equivalent parallel conductance vs bias for a steam-grown oxide on n-type silicon. Curves (c) and (d) are the capacitance and equivalent parallel conductance vs bias after heating in dry nitrogen at 350°C for 17 hours.

Fig. 29 is a plot of capacitance and equivalent parallel conductance vs bias measured at 50 Hz on a sample having N_{ss} in the 10^{10} cm^{-2}-eV^{-1} range. The sample is n-type oriented in the [100] direction and the low value of N_{ss} is produced by the annealing process described previously.

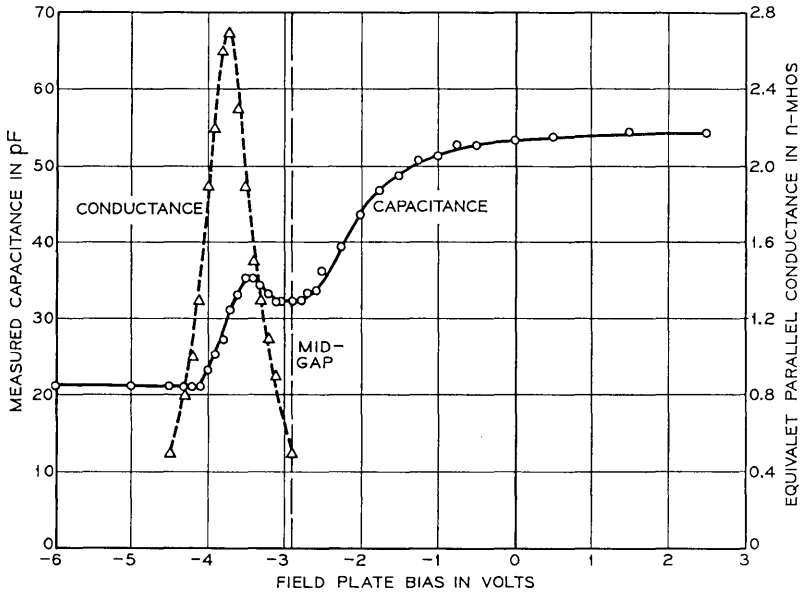


Fig. 23 — Capacitance and equivalent parallel conductance measured at 50 Hz vs field plate bias. Sample is the n-type one in Fig. 14. curves show capacitance and equivalent parallel conductance in the weak inversion region.

VIII. DISCUSSION

The discussion is divided into three parts. The first part deals with the depletion region and the second with the weak inversion region. The third part discusses the limitations of the conductance technique.

8.1 *Depletion Region*

8.1.1 *Steam-Grown Oxides on [111] Orientation*

It is seen from Fig. 14 that points calculated from the measurements fit (22) quite well for both n and p type samples. Time constant decreases by a factor of $1/e$ for every kT/q increase in $u_B - \bar{u}_s$. The exponential dependence of time constant on surface potential shown in Fig. 14 means that time constant is inversely proportional to majority carrier density at the silicon surface as seen from (21) and (22). This

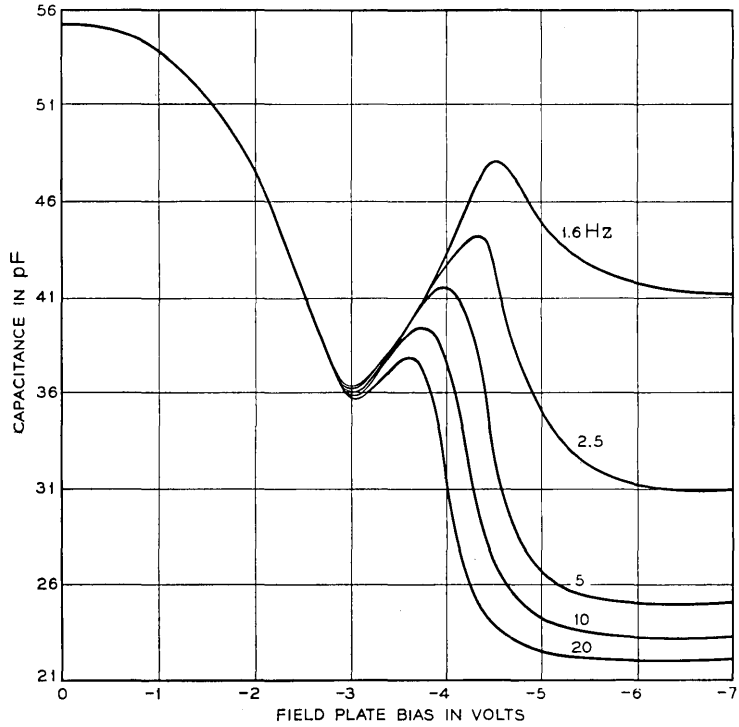


Fig. 24 — Capacitance vs field plate bias with frequency as parameter measured with apparatus described in Ref. 28, on n-type sample of Fig. 14. Curves show capacitance dispersion in weak inversion.

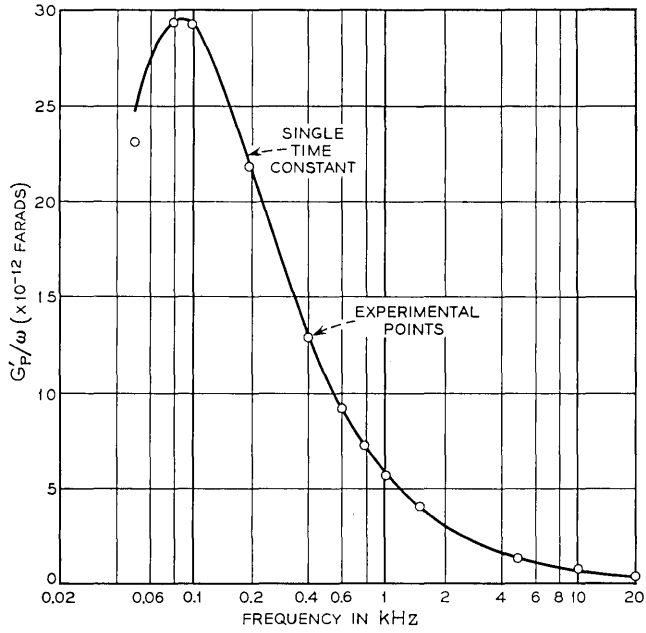


Fig. 25 — G_p'/ω vs frequency showing how experimental points fit (17). C_s = 5.14×10^{-8}farads/cm^2 and $\tau = 1.7 \times 10^{-3}$ seconds are used in (17) to get this fit. G_p' is equivalent parallel conductance in mhos derived from measurements. Experimental points are measured at a bias of −3.3 volts ($\bar{u}_s = 19.8$) in the weak inversion region on the n-type sample of Fig. 14.

is the experimental evidence that majority carrier transitions make the dominant contribution to measured equivalent parallel conductance in the depletion region.

The exponential dependence of time constant on surface potential is important evidence in addition to the drying experiments of Figs. 3 and 4, that the dominant loss is due to transitions to and from interface states. This follows from the fact that no other relaxation processes in the system have time constants which depend exponentially on surface potential. To see this, we shall consider generation-recombination though impurity levels in the silicon near mid-gap, majority carrier transitions to and from impurity levels in the silicon near the Fermi level and the transit time of majority carriers across the silicon space-charge region. Bulk generation-recombination is insignificant because the Fermi level does not cross impurity levels near mid-gap for band bending in depletion as seen in Fig. 5(b). Therefore, generation-recombination is taken to be zero in Figs. 6 and 8. The Fermi level can

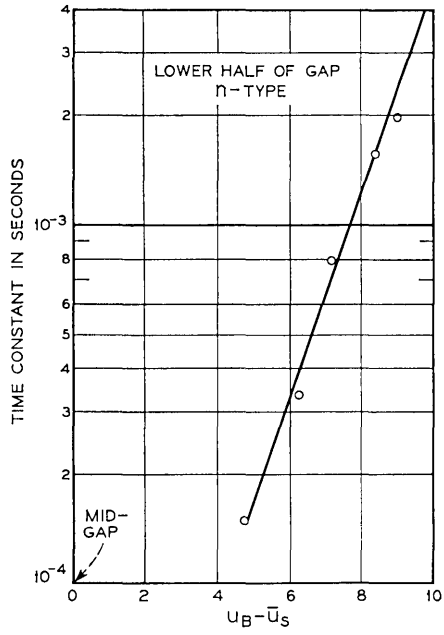


Fig. 26 — Log τ_m vs $u_B - \bar{u}_s$ for n-type sample in Fig. 14 in weak inversion region. Curve shows that equivalent parallel conductance in weak inversion is dependent on majority carrier density at the silicon surface.

still cross any impurity levels located near it [see Fig. 5(b)]. Majority carrier transitions to and from levels at this crossover point will contribute to the loss. The location of the crossover point from the interface will vary with bias. However, the crossover point is always at the same energy with respect to mid-gap so that majority carrier density there will be constant and independent of bias. Therefore, the time constant for this process will be independent of bias. As time constant is not observed to be bias independent in depletion in the samples measured, majority carrier transitions to and from levels in the silicon bulk must make a negligible contribution to the loss compared to majority carrier transitions to and from interface states.

Loss due to the transit of majority carriers across the space-charge layer is negligible.[11,12] It can be shown using the relations given in Refs. 11 and 12 that the value of space-charge resistance for majority carriers when the silicon surface is intrinsic is four orders of magnitude less than majority carrier capture resistance in the samples measured.

The fact that curves (a), (b), and (c) in Fig. 14 can be fitted so well by straight lines implies that capture cross section in these regions is

to a good approximation independent of energy in the bandgap. This independence is an experimental fact for which no satisfactory explanation exists at present. Because capture cross sections are independent of energy, they will be unaffected by statistical fluctuations of surface potential. The values of capture cross section obtained from the linear portions of Fig. 14 are therefore truly characteristic of the interface states present. Curves (b) and (c) in Fig. 14 measured on n-type are characterized by two types of interface states having different capture cross sections with an abrupt transition between them. The reason for this is not understood.

Fig. 14 shows that N_{ss} is slowly varying with energy. It is reasonable to conclude from this, and the observation, that the time constant varies monotonically with surface potential over the same portion of the band gap and that there is a continuum of interface states rather than a discrete level. It is also seen from Fig. 15 that N_{ss} is in the 10^{11} cm^{-2}-eV^{-1} range. This means that the capturing centers are spaced too far apart in the plane of the interface for the wavefunction of an

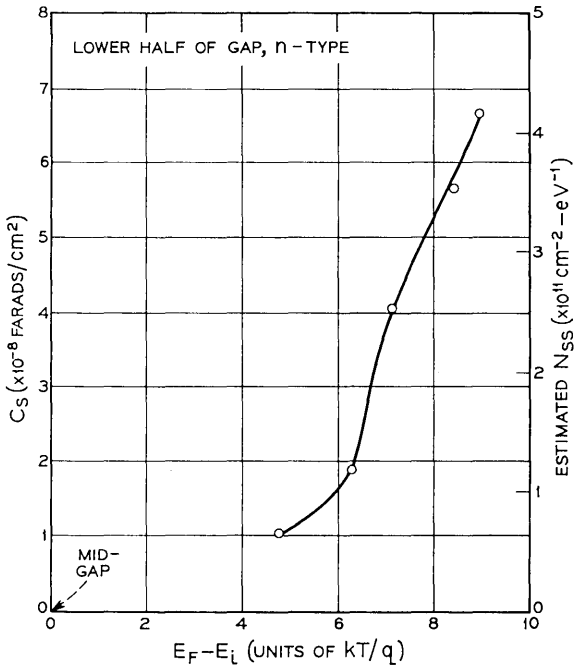


Fig. 27 — C_s and estimated N_{ss} vs $E_F - E_i$ for n-type sample in Fig. 14 in the weak inversion region. Estimated $N_{ss} = C_s/q$. C_s increases monotonically with $E_F - E_i$.

electron in one center to overlap a neighboring center. Transitions from one center to another are therefore, highly improbable so that it may be assumed that we do not have a band but just a continuum of levels closely spaced in energy. Transitions will only occur between the majority carrier band and levels within a few kT/q of the Fermi level. Cascading transitions from center to center which could occur in a band would probably have a different dependence of time constant on surface potential than shown in Fig. 14. Therefore, cascading transitions were not considered in the theory developed in Section IV.

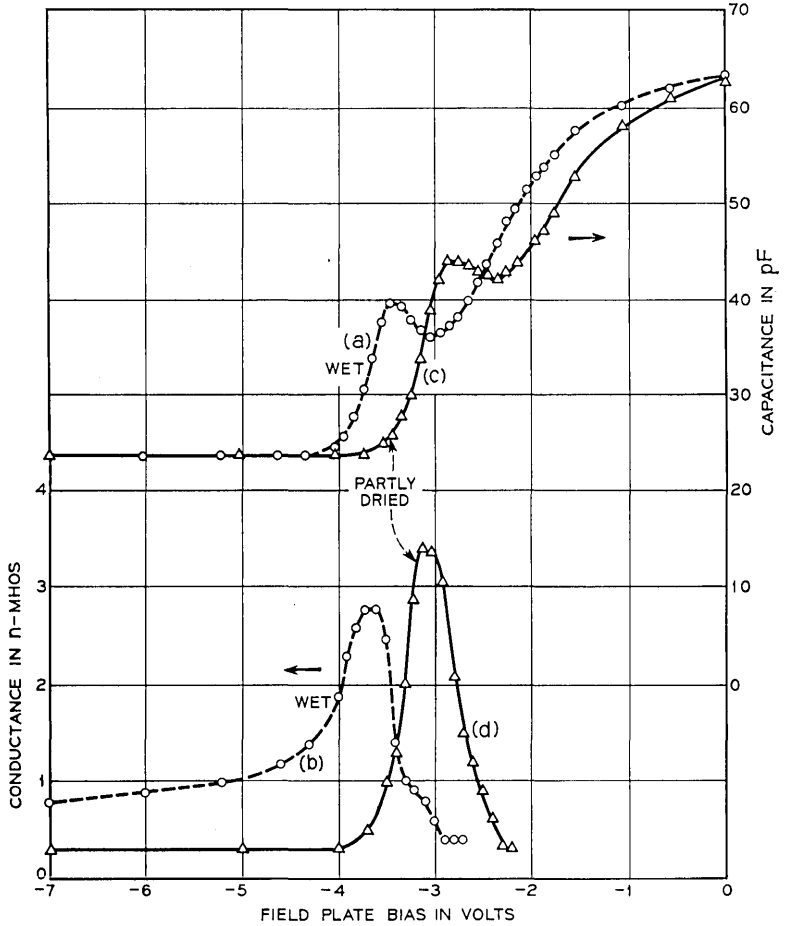


Fig. 28 — (a) capacitance and (b) equivalent parallel conductance measured vs field plate bias before drying; (c) capacitance and (d) equivalent parallel conductance vs field plate bias of same capacitor after heating in dry nitrogen for 17 hours at 350°C. This is the same sample as Fig. 4 only measured at 50 Hz.

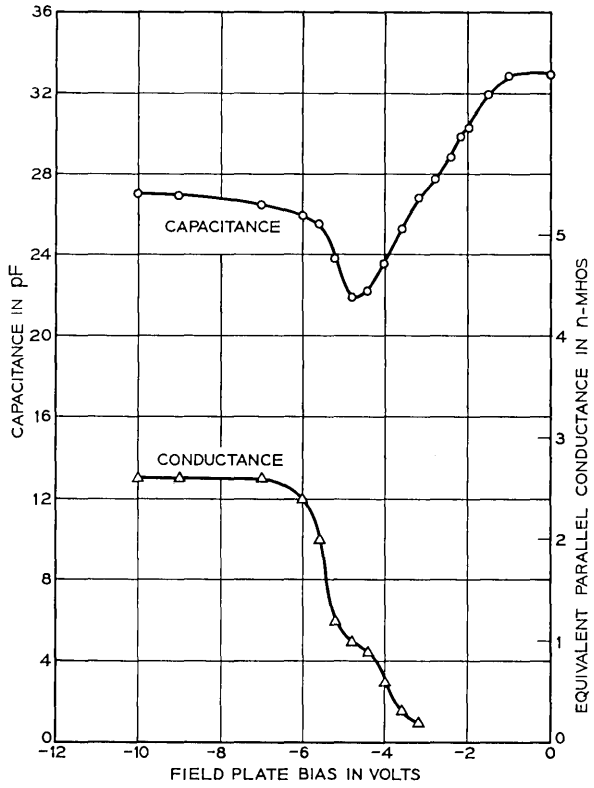


Fig. 29 — Capacitance and equivalent parallel conductance measured at 50 Hz vs field plate bias. Field plate diameter is 3.8×10^{-2}cm, silicon resistivity 1 ohm-cm, n-type, [100] oriented, and $C_{ox} = 2.8 \times 10^{-8}$farads/cm^2. Oxide is grown in steam then H_2 annealed as described in text to produce low N_{ss}.

The integration of (4) and (26) was performed for the case where N_{ss} and capture cross section do not vary appreciably over an energy interval of several kT/q. Figs. 14 and 15 show the experimental justification for this.

In Fig. 16, log τ_m is plotted as a function of $\bar{\psi}_s$ rather than $u_B - \bar{u}_s$ to illustrate more clearly the temperature dependence. An alternate and equivalent form of (22) which follows from (21) for p-type is

$$\tau_m = \frac{1}{\bar{v}\sigma_p N_A} \exp(q\psi_s/kT). \tag{60}$$

Curve (a) in Fig. 14 fits (60) quite well with a slope q/kT where $T = 204°$K. Curve (b) fits with the slope for 300°K. Thus, the temperature dependence expected from (60) is observed. This agrees with Shockley-

Read theory[18] which is used in Section IV. Capture cross section can be obtained from (60) by extrapolating the curves in Fig. 16 to $\bar{\psi}_s = 0$. Doing this, we get $\sigma_p = 8.8 \times 10^{-17}$ cm^2 from curve (a) and $\sigma_p = 2.1 \times 10^{-16}$ cm^2 from curve (b). The capture cross section from curve (b) is of course in agreement with σ_p from curve (a) in Fig. 14. However, it is 2.4 times larger than σ_p from curve (a) of Fig. 16. This is probably due to the error in finding the value of Δ from (52). Normally, Δ can be found only within 20 to 30 mV of its correct value by the technique of Section V. Part of this error is due to neglecting the difference between hole and electron effective mass. It is considered likely that in reality capture cross section is nearly independent of temperature in this range. The apparent variation is ascribed instead to the errors in finding the value of Δ.

8.1.2 *Steam-Grown Oxides on [100] Orientation*

The same general conclusions can be reached for the [100] samples from Figs. 17 and 18 as for the [111] samples just discussed but there are two differences. The first difference is that the time constant curves in Fig. 14 for the upper half of the gap in [111] p-type are similar to the curves for the lower half of the gap in [100] p-type in Fig. 17. Also, the curves in the lower half of the gap in Fig. 14 for [111] p-type are similar to the curves in the upper half of the gap in [100] n-type in Fig. 17. The second difference is that interface state density in Fig. 15 for the lower half of the gap in [111] p-type is about twice as great as in Fig. 18 for the lower half of the gap in [100] p-type. However, interface state density in Fig. 15 for the upper half of the gap in [111] n-type is about the same as in Fig. 18 for the upper half of the gap in [100] n-type. The reason for these differences is not understood.

Built-in charge density in the [111] samples varies from sample to sample from 5×10^{11} cm^{-2} to 9×10^{11} cm^{-2} in both n and p-type. This is about twice as great as in the [100] samples confirming the observations of Balk.[29] Also, the increase of N_{ss} towards the band edges as shown in Figs. 15 and 18 is consistent with the results of Gray and Brown.[30]

8.1.3 *[111] Orientation, Partly-Dried Oxides*

Fig. 19 shows that interface state time constants fit (22) after partial drying. Capture cross section for n-type has not changed appreciably although two types of states were not found. This may be due to the fact that too small a range of surface potential was covered in the measurements. Because capture cross section did not change, partial

drying does not appear to introduce new types of states—but only increases the number of states already present.

Comparison of curve (a) to curve (b) in Fig. 20 shows that interface state density for the partly-dried oxide is greater than for the wet oxide. This can be seen directly from Fig. 3 and 4 for p and n-type by the increase in the peak of the measured conductance with drying.

Fig. 3 for p-type shows by the shift of the curves to higher negative bias that a net positive charge has been induced in the oxide by partial drying. Fig. 4 for n-type shows by the shift of the curves towards less negative bias that a net negative charge has been induced in the oxide by partial drying. No conclusions about these results will be drawn because it is not certain that other changes independent of interface state density have not occurred in these samples.

The increase in interface state density with drying can explain the results observed in n^+p junctions in a previous paper.[31] In Ref. 31, it was found that drying an oxide heavily contaminated with alkali metal ions drastically reduced the current flowing in an n channel along the surface of the p-type material. The increased interface state density caused by drying results in more channel electrons being captured by interface states. This markedly reduces field-effect mobility which in turn results in an increase in the sheet resistance of the inversion layer. Although the inversion layer may be terminated by a defect at the edge of the wafer which shorts it to the p substrate, the current flow will be reduced to a much lower level than before.

8.1.4 *Range of* N_{ss}

The range of N_{ss} from 10^{12}cm^{-2}-eV^{-1} for very dry oxides through 10^{11} cm^{-2}-eV^{-1} for wet oxides to 10^{10} cm^{-2}-eV^{-1} for aluminum clad oxides annealed in H_2 or N_2 are empirical observations. It is not definitely known what these centers are. Therefore, an explanation can only be speculative. However, these results are qualitatively similar to previous work on "bare" surfaces which had at most a few monolayers of oxide.[32] Exposure to water vapor always reduced N_{ss} by a few orders of magnitude.[32] For the thick oxides studied in this work, water was most likely present in the form of OH groups which have a permanent dipole moment. An interesting model to explain the annihilation of fast states on "bare" surfaces has been proposed by Rzhanov.[33] Assuming that each interface state also has a permanent dipole moment, the lowest energy configuration will occur at the interface when the dipole moment of an H_2O or OH group is aligned antiparallel with the dipole moment of the nearest interface state. This could reduce the

capture probability of the interface state[34] to such an extent that there is no measurable loss due to transitions involving it. The only loss measured would be due to transitions to and from the remaining unpaired states. For thick oxides, this fits the observed decrease in N_{ss} with increasing wetness and the observation that measured capture cross section remains unchanged. This model is, however, only one of several capable of explaining the results.

8.1.5 *Statistical Model*

The time constant dispersion actually observed is much broader than expected for a continuum. This can be seen by comparing Fig. 21 to the curves in Fig. 13. Three mechanisms which could cause this broadening are considered. They are:

(*i*) Tunnelling of majority carriers to states randomly distributed into the oxide.[2]
(*ii*) Fluctuations of capture cross section over the plane of the interface.
(*iii*) Random fluctuations of surface potential over the plane of the interface.

The tunnelling mechanism is unlikely to be the dominant cause of the time constant dispersion for the following reasons. The range of frequencies used in the measurements is from 50 Hz to 500 kHz. Thus, only those interface states having time constants shorter than several milliseconds will be measured. The time constant distribution for tunnelling would be[2]

$$\tau_t = (n_{s_o}\bar{v}\sigma_o)^{-1} \exp(2K_o\xi), \tag{61}$$

where σ_0 is the true electron capture cross section of the states, $1/2K_o$ is the tunnelling decay constant, and ξ is the distance into the oxide measured from the interface. For any reasonable value of $1/2K_o$, such as 1 Å given in Ref. 2, and reasonable values of σ_o, such as 10^{-15} cm^2 to 10^{-16} cm^2, it can be seen from (61) that states located more than a few Å into the oxide would have time constants much too long to be measured in the frequency range between 50 Hz and 500 kHz. Hysteresis in the capacitance vs bias characteristic which can be observed by slowly varying bias from a low to a high value and then back is negligibly small. Hysteresis would be observed when there is an appreciable density of states deep in the oxide having time constants longer than the time it takes to vary the bias because the charging

and discharging of these states would lag behind the bias variation.[2] The fact that no hysteresis is observed means that the density of states in the oxide having time constants of the order of minutes is negligible in the samples measured. Finally, it is shown in Appendix D that for any reasonable value of K_o, the shape of the G_p/ω vs log ω curve calculated using (20) for the continuum of states and (61) for the tunnelling time constant distribution differs markedly from the observed G_p/ω vs log ω curve.

Although capture cross section is observed to be independent of energy over large portions of the band gap from Fig. 14, capture cross section could depend upon the location of the state in the plane of the interface. The assumption that capture cross section fluctuates over this plane does not give a broad enough time constant dispersion to fit the observed G_p/ω vs frequency curves.

Random fluctuations of surface potential over the plane of the interface is the model which fits best. One can see from (22) that small fluctuations of u_s will cause large fluctuations in τ_m. This broadens the time constant dispersion sufficiently to fit the experimental curves.

Random fluctuations of surface potential can be caused by:

(*i*) A random distribution of built-in charges and charged interface states over the plane of the interface.

(*ii*) A random distribution of ionized impurities in the silicon.

(*iii*) Random fluctuations of oxide thickness over the plane of the interface.*

Of these, the first two are the basis for deriving (39) in Section IV. The last was excluded because no dependence of time constant dispersion on oxide thickness was observed between 500 and 2000 Å in the samples measured. Fig. 21 shows the fit of (39) to experimental values of G_p/ω measured in the frequency range between 50 Hz and 500 kHz at a mean surface potential of 8.1. Fig. 21 shows that (39) accurately fits the experimental points. The fit of (39) to the experimental G_p/ω vs frequency points is equally accurate for all the other values of mean surface potential $\bar{u}_s$ measured in depletion. The only independent parameter in (39) varied to obtain these fits is the characteristic area α. By fitting (39) to the experimental G_p/ω vs frequency points at each value of $\bar{u}_s$ in this way, the values of α obtained are found to be related to space charge width W. This relation is shown in Fig. 22 where the length of the side of each square characteristic area $\alpha^{\frac{1}{2}}$ is plotted against

* This effect was pointed out to us by C. R. Crowell.[35]

W. Fig. 22 shows that $\alpha^{\frac{1}{2}}$ increases linearly with W and has a larger magnitude. To understand this relationship, consider a positive built-in charge or charged interface state located at a point in the plane of the interface. This charge is sandwiched between the plane of the field plate and the plane at the edge of the space-charge region a distance W away. An image charge is induced in each of these planes. The image charge in each plane induces another in the opposite plane and so on. Thus, the charge at the interface induces an infinite number of image charges in the field plate and space-charge edge planes. Let us now add a number of charges in the neighborhood of this charge in the plane of the interface and calculate the charge distribution induced by all the charges and their images in the plane at the edge of the space-charge region. The largest spacing between charges which results in a constant induced charge distribution in the plane at the edge of the space-charge region is $\alpha^{\frac{1}{2}}$. Thus, charges within a characteristic area are indistinguishable as separate charges which makes surface potential over a characteristic area uniform. The characteristic area is a measure of the granularity of the surface-charge distribution seen at the edge of the space-charge region. A calculation based on the image charge model shows that $\alpha^{\frac{1}{2}}$ has the values in Fig. 22 and a linear dependence on W. The value of α can be obtained within a factor of 2 from the image charge model so that α is not completely arbitrary.

We shall show next that the random distribution of built-in charges and charged interface states is the dominant cause of surface potential fluctuations in the samples measured. Rewriting (36) from Section IV for the standard deviation of surface potential caused by the random distribution of built-in charges and charged interface states gives

$$\sigma_s = \frac{W(\bar{u}_s)\beta(q\bar{Q}/\alpha)^{\frac{1}{2}}}{[W(\bar{u}_s)C_{ox} + \epsilon_{si}]}. \tag{62}$$

Curve (b) in Fig. 30 is the variance σ_s^2 calculated from (62) as a function of W. $\bar{Q}$ calculated from (31) and (32) with $u_s = \bar{u}_s$ is found to be independent of bias from -1.8 volts to -2.2 volts. This bias range corresponds to mean surface potential from 10.3 to 3.8. Assuming a constant value of interface state density of 3×10^{11} cm^{-2}-eV^{-1} in the lower half of the gap from Fig. 15, the change in total interface state density is $3 \times 10^{11} \times (10.3 - 3.9) \times 0.026 = 5.1 \times 10^{10}$ cm^{-2}. This is negligibly small compared to $\bar{Q}/q = 9.8 \times 10^{11}$ cm^{-2}. Curve (b) in Fig. 30 is calculated using $\bar{Q} = 1.57 \times 10^{-7}$ coul/cm^2, $C_{ox} = 5.7 \times 10^{-8}$ farads/cm^2 and the values of α from Fig. 22.

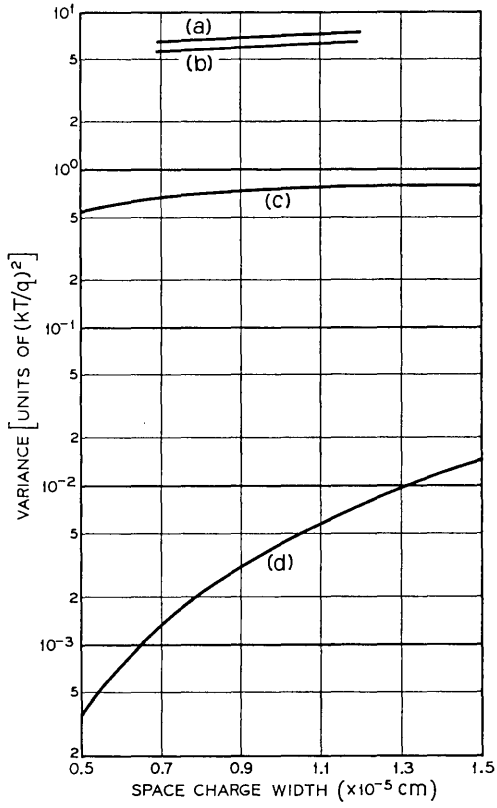


Fig. 30 — (a) total variance of surface potential ($\sigma_s^2 + \sigma_B^2$) caused by random distributions of built-in charges and charged interface states and ionized acceptors in the silicon space-charge region. σ_s and σ_B are calculated from (62) and (63), respectively. (b) variance of surface potential σ_s^2 caused by random distribution of built-in charges and charged interface states alone. σ_s is calculated from (62). (c) variance of surface potential σ_B^2 caused by random distribution of ionized acceptors in silicon space-charge region. σ_B is calculated from (63). (d) variance of surface potential σ_x^2 caused by fluctuations of oxide thickness. σ_x is calculated from (64).

For the standard deviation of surface potential caused by a random distribution of ionized acceptors in the silicon space-charge region, we rewrite (37) derived in paragraph B.1 of Appendix B

$$\sigma_B = \frac{q\beta[\bar{N}_A W(\bar{u}_s)]^{\frac{1}{2}}[1 - \exp(-\bar{u}_s)]}{2[W(\bar{u}_s)C_{ox} + \epsilon_{si}]}. \tag{63}$$

Curve (c) in Fig. 30 is the variance σ_B^2 calculated from (64) as a func-

tion of W. The values used are: $C_{ox} = 5.7 \times 10^{-8}$ farads/cm^2 and $\bar{N}_A = 2.1 \times 10^{16}$ cm^{-3}.

Finally, for the standard deviation of surface potential caused by oxide thickness fluctuation, we rewrite (101), derived in paragraph B.2 of Appendix B,

$$\sigma_x = \frac{qN_A\beta W^2(\bar{u}_s)[1 - \exp(-\bar{u}_s)]\, d\bar{x}}{[W(\bar{u}_s)\epsilon_{ox} + \bar{x}\epsilon_{si}]}. \tag{64}$$

Curve (d) in Fig. 30 is the variance σ_x^2 calculated from (64) as a function of W. The fluctuations of oxide thickness arise from the random nature of the oxidation process. It is reasonable to take $d\bar{x}$ as about one silicon lattice constant or 5Å. The values used in (64) to get curve (d) in Fig. 30 are: $\bar{x} = 595$ Å, $N_A = 2.1 \times 10^{16}$ cm^{-3}, and $d\bar{x} = 5$ Å.

Curve (a) in Fig. 30 is $\sigma_s^2 + \sigma_B^2$ vs W obtained by fitting (39) to the experimental G_p/ω vs frequency points. Fig. 30 clearly shows that the random distribution of built-in charges and charged interface states is the dominant cause of surface potential fluctuations in the samples measured. It can be easily seen that $\alpha^{\frac{1}{2}}$ should vary linearly with W as follows. Over the range of surface potential measured, the experimental G_p/ω vs frequency curves broaden very slightly with increasing surface potential. This is shown quantitatively in curve (a) of Fig. 30. Over the same range of surface potential, curve (c) in Fig. 30 is virtually independent of W. The dominant term σ_s^2 in curve (a) is therefore, nearly independent of W as shown in curve (b) of Fig. 30. For σ_s constant, an inspection of (62) will show that for $\bar{Q}$ constant and $WC_{ox} < \epsilon_{si}$ as is the case in the samples measured, $\alpha^{\frac{1}{2}}$ must vary linearly with W.

The density of built-in charges and charged interface states is in the high 10^{11} cm^{-2} range in the samples measured. Therefore, these charges are about 100 Å apart on the average which is too far for coulombic interaction between them to be important. This makes the assumption that they are randomly distributed a reasonable one. The mean number of built-in charges and charged interface states $\bar{N}$ in a characteristic area α can be calculated from the relation $\bar{N} = \alpha\bar{Q}/q$ (see Section IV) using values of α obtained from Fig. 22. $\bar{N}$ is found to vary from 140 charges at $W = 6.9 \times 10^{-5}$ cm to 246 charges at $W = 1.07 \times 10^{-5}$ cm. For these values of $\bar{N}$, there is negligible error in using the Gaussian approximation to the Poisson distribution (28) in deriving (39) in Section IV. The mean number of ionized acceptors calculated from the relation $\bar{M} = W^3\bar{N}_A$ varies from 26 charges at $W = 1.07 \times 10^{-5}$ cm to 7 charges at $W = 6.9 \times 10^{-5}$ cm. For all values of $\bar{M}$ except perhaps the last, there is negligible error in using the Gaussian approximation

to the Poisson distribution (81) in paragraph B.1 of Appendix B in deriving (39). Any error at $W = 6.9 \times 10^{-5}$ cm is not important to the final results because the contribution of σ_B^2 to the total variance is so small. The assumption of small fluctuations made in deriving (39) also seems to be reasonable because of the rather good fit obtained to the experimental points.

One can decide from (62), (63), and (64) which influence will be dominant for samples other than the ones used here. In spite of the fact that the samples investigated here were dominated by random fluctuations of built-in charges and charged interface states, it should be borne in mind that the other influences may become important in other cases. When built-in charge and interface state density is low and the oxide relatively thick, then bulk doping fluctuations will dominate. When the oxide is thin and the impurity density is high, random fluctuations of oxide thickness become important. The general conclusion is that there will always be random fluctuation of surface potential in simiconductor-insulator interfaces.

No data has been taken at flat bands or in the accumulation region. However, it is expected that the equivalent circuit in Fig. 8(a) will also be valid in this region with all the distributed time constants shifted to shorter values.

8.2 *Weak Inversion Region*

In this part, it will be shown that the loss in weak inversion can arise from generation-recombination either through bulk states or interface states and how the conductance technique can be used to distinguish between them. It will be shown also that the observed effects in this region can be explained with a continuum of states rather than a single level near a band edge.

When the silicon surface becomes inverted, the Fermi level will cross impurity levels near mid-gap as seen from Fig. 5(c). Generation recombination through bulk states as well as through interface states can contribute significantly to the measured loss. These two cases can be distinguished by the bias dependence of the measured equivalent parallel conductance. When generation-recombination through interface states dominates the loss, the equivalent parallel conductance goes through a peak as a function of bias in weak inversion because time constant varies inversely with majority carrier density at the surface (see paragraph 4.6). This case is illustrated in Fig. 23 where equivalent parallel conductance goes through a peak in weak inversion and drops to a very low value in strong inversion.

When generation-recombination through bulk states dominates the loss, equivalent parallel conductance does not go through a peak as a function of bias in weak inversion. The reason for this is that the loss due to this process is bias independent. This follows from the fact that hole and electron densities at the crossover point [see Fig. 5(c)], where the Fermi level crosses impurity levels located near mid-gap, are independent of bias. This case is illustrated in Fig. 29. The slight indication of a peak in the conductance curve means that generation recombination through interface states makes a slight contribution to the loss. However, the primary effect is that of a bias independent loss. The variation of measured conductance with bias is mainly due to variation of capacitance with bias.

Fig. 23 is measured on a sample having N_{ss} in the 10^{11} cm^{-2}-eV^{-1} range while Fig. 29 is measured on a sample having N_{ss} in the 10^{10} cm^{-2}-eV^{-1} range. Thus, samples can be made in which one or the other process dominates the loss. It should be noted that the conductance technique can be used to investigate generation-recombination through interface states or bulk states in the weak inversion region by choosing an appropriate sample. The investigation of generation-recombination through bulk states is beyond the scope of this work so that most of the samples investigated have interface state loss dominant in weak inversion. Therefore, bulk generation-recombination is made zero in the equivalent circuits of Fig. 9.

In general, samples having N_{ss} in the 10^{11} cm^{-2}-eV^{-1} range or higher will have interface state loss dominant in the weak inversion range. Additional evidence for this in such samples is shown in Fig. 28. The only difference between Fig. 28 and Fig. 4 is the low signal frequency which is used in Fig. 28 to reveal the behavior of equivalent parallel conductance in the weak inversion region. The conclusion is again drawn from Fig. 28 that interface state loss must be dominant as from Figs. 3 and 4 discussed previously in Section III.

The striking feature of the experimental G_p/ω vs frequency curves at each bias in weak inversion is that they can be accurately fitted by a single time constant using (17). An example of this is shown in Fig. 25. Three possible explanations of this will be considered.

(*i*) A single level state located at an energy in the silicon band gap corresponding to the bias at the capacitance peak in Fig. 23.

(*ii*) A single level near the valence band edge.

(*iii*) A continuum of interface states over the silicon band gap.

The first explanation can be ruled out by an examination of Figs. 24, 26, and 27. Fig. 24 shows that the capacitance peak shifts to greater

negative bias with decreasing signal frequency. The peak must always occur at the same bias for a single level. In Figs. 26 and 27, the range of $u_B - \bar{u}_s$ and $E_F - E_i$ corresponds to a range of bias which includes the capacitance peak in Fig. 23. Both τ_m in Fig. 26 and C_s in Fig. 27 increase monotonically through the corresponding capacitance peak in Fig. 23. For a single level state, C_s would go through a peak and τ_m first would be constant for E_F on the valence band side of the single level and then it would decrease exponentially with u_s for E_F on the conduction band side.

The range of $E_F - E_i$ and $u_B - u_s$ in Figs. 26 and 27 is limited by the capacitance dispersion shown in Fig. 24. This means that (52) cannot be used without significant error too far beyond the capacitance peak (at reasonable frequencies).

To rule out the possibility of a single level state near the valence band edge, G_p/ω vs frequency curves were obtained on a p-type sample This sample had steam-grown oxide on a 1 Ω-cm crystal oriented in the [100] direction. The [100] direction has a lower built-in charge density so that masking effects due to lateral ac current flow into the inverted layer beyond the field plate are minimized. A single time constant was again found to be characteristic of the G_p/ω vs frequency curves in the weak inversion region. The experiment was repeated on an n-type sample oriented in the [100] direction having a steam-grown oxide with the same result. The fact that the same result was obtained in weak inversion on both n and p-type samples and the lack of evidence for a single level state in the same energy range in accumulation in these samples makes such a state as an explanation of the effect very unlikely. The observations can be best explained in terms of a continuum of states as done in paragraph 4.6.

Evidence that this effect is dominated by the majority carrier density at the silicon interface is given by Fig. 26. Fig. 27 shows that C_s increases so rapidly with E_F that the approximation of a constant density of states over an interval of kT/q cannot be made. Thus, (18) cannot be simply integrated. The fact that C_s is so rapidly varying is the reason why the curve in Fig. 26 does not have the same magnitude of slope as the curves in Fig. 14. This rapid variation of C_s also explains the capacitance peak in Fig. 23. It can be seen that (16) will go through a peak if C_s has a strong dependence on bias.

Comparison of Fig. 27 with Fig. 15 shows that interface state density in the lower half of the silicon bandgap in n-type has a different magnitude and dependence on position of Fermi level than in the lower half of the gap in p-type. The reason for this is not understood.

8.3 *Limitations of the Conductance Technique*

Interface state densities in the 10^9 cm^{-2}-eV^{-1} range have been measured by this technique using the apparatus described in Section V. For a field plate diameter of 5×10^{-2} cm, this corresponds to about 2×10^6 states/eV. For a signal of 20 mV peak-to-peak, transitions involving only about 4×10^4 states can be detected.

It is important to use as thin an oxide layer as practicable to get large values of capacitance and equivalent parallel conductance for a given N_{ss} and to minimize the error in extracting C_{ox}.

The range of bandgap energy over which measurements were made in this work was limited by the frequency range 50 Hz to 500 kHz of the two bridges used.

IX. SUMMARY AND CONCLUSIONS

(*i*) Majority carrier transitions to and from interface states make the dominant contribution to the measured equivalent parallel conductance in the accumulation, depletion, and inversion regions. This fact makes it possible to use simplified equivalent circuits and extract interface state properties for each type of carrier independntly.

(*ii*) A continuum of Shockley-Read type interface state levels closely spaced in energy across the silicon band gap seems to be characteristic of the Si–SiO_2 interface.

(*iii*) Interface state densities in the 10^{11} cm^{-2}-eV^{-1} range are usual for "as-grown" steam oxides.

(*iv*) Interface state density can be increased to the 10^{12} cm^{-2}-eV^{-1} range by drying the oxide.

(*v*) Interface state density can be reduced to the 10^{10} cm^{-2}-eV^{-1} range by H_2 or N_2 annealing at low temperatures with an aluminum film deposited on a steam-grown oxide.[29]

(*vi*) Capture cross sections for electrons and holes are independent of energy over large portions of the silicon bandgap. They are also independent of temperature, the processes of drying or H_2 or N_2 annealing, and, except for subtle differences, crystal orientation.

(*vii*) In depletion, a time constant dispersion is observed which is much broader than expected for a continuum. The broadened time constant dispersion is determined mainly by statistical fluctuations of surface potential. In the samples studied, a random distribution of built-in charges and charged interface states was found to be the primary cause of the surface potential fluctuations. This model provides an accurate fit to the experimental points and yields a characteristic

area whose linear dimension has about twice the magnitude and has the same bias dependence as the silicon space-charge width. Charges within a characteristic area cannot be distinguished as separate charges. Random fluctuations of surface potential can also arise from a random distribution of ionized impurities in the bulk silicon and by random fluctuations of oxide thickness. In samples where the built-in charge and charged interface state density is low or the oxide layer thin, these can replace the random distribution of built-in charges and charged interface states as the dominant causes of surface potential fluctuations.

(*viii*) In weak inversion, the fast response time of minority carriers and the conductance of the inversion layer results in the observation of a single time constant at each bias.

(*ix*) The interface state branch of the equivalent circuit in the depletion-accumulation region where $\bar{u}_s < u_B$ is given in Fig. 31(a). The equivalent circuit in the mid-gap region where $\bar{u}_s = u_B \pm a$ few kT/q is shown in Fig. 31(b). Finally, the equivalent circuit in the weak inversion region where $u_B < \bar{u}_s < 2u_B$ is shown in Fig. 31(c).

(*x*) It is impossible to make an MIS structure in which surface potential is absolutely uniform because even if built-in charge and interface state density is made small, random distribution of ionized bulk impurities or random fluctuations of oxide thickness will still be present.

(*xi*) The accuracy required for extracting this detailed a picture of the Si–SiO_2 interface from MIS capacitor measurements could only be provided by the conductance technique (see Fig. 1).

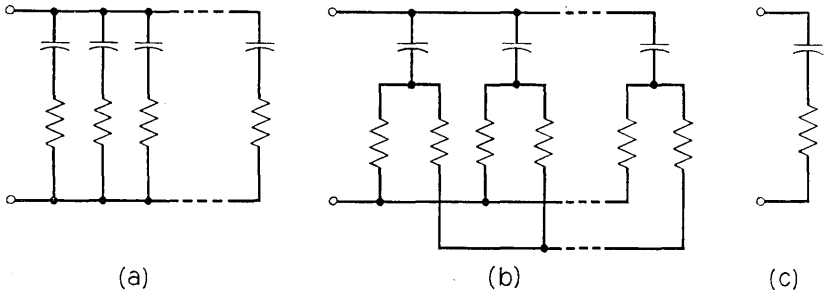


Fig. 31 — (a) Schematic of interface state branch of equivalent circuit in depletion-accumulation region where $\bar{u}_s < u_B$. (b) Schematic of interface state branch of equivalent circuit in mid-gap region where $\bar{u}_s = u_B \pm$ a few kT/q. (c) Schematic of interface state branch of equivalent circuit in weak inversion region where $u_B < \bar{u}_s < 2u_B$.

X. ACKNOWLEDGMENTS

The authors are indebted to Messrs. J. T. Nelson, C. N. Berglund, and C. Goldberg for their criticisms and many valuable suggestions during the course of this work. We also thank Mr. V. Heine for stimulating discussions, Mr. R. M. Ryder for a critical reading of the manuscript, Mr. A. D. Lopez for making the measurements, Mr. J. McGlasson for preparing the samples, Mr. A. Hartman for proofreading the manuscript, and Mrs. J. M. Schilling for help in the computer programming.

APPENDIX A

In this Appendix, interface state admittance is derived from (3):

$$i_s(t) = qN_s c_n[1 - f(t)]n_s(t) - qN_s e_n f(t). \tag{65}$$

Expressing $f(t)$ as the sum of a dc and an ac part:

$$f(t) = f_o + \delta f, \tag{66}$$

where f_o is the Fermi function established by the bias and δf the change caused by the ac signal. Similarly,

$$n_s(t) = n_{so} + \delta n_s , \tag{67}$$

where n_{so} is the electron density at the silicon surface when the Fermi level is at the trap level and δn_s the change caused by the ac signal. Substituting (66) and (67) into (65) and making the small signal approximation by neglecting second order terms, we get

$$i_s(t) = qN_s c_n[(1 - f_o)n_{so} + (1 - f_o)\,\delta n_s - n_{so}\,\delta f] - qe_n N_s(f_o + \delta f). \tag{68}$$

From detailed balance

$$qN_s c_n(1 - f_o)n_{so} = qe_n N_s f_o . \tag{69}$$

Solving (69) for e_n and substituting into (68) we get

$$i_s(t) = qN_s c_n\left[(1 - f_o)\,\delta n_s - n_{so}\frac{\delta f}{f_o}\right]. \tag{70}$$

The net current density can also be expressed as

$$i_s(t) = qN_s \frac{df}{dt}. \tag{71}$$

Equating (71) to (70)

$$\frac{df}{dt} = c_n(1 - f_o)\ \delta n_s - c_n n_{so} \frac{\delta f}{f_o}. \tag{72}$$

The small signal variation of the Fermi function is $\delta f = f_m \exp(j\omega t)$. Combining this with (66) results in

$$\frac{df}{dt} = j\omega\ \delta f. \tag{73}$$

Substituting (73) into (72) and solving for δf

$$\delta f = \frac{f_o(1 - f_o)\ \delta n_s}{n_{so}(1 + j\omega f_o/c_n n_{so})}. \tag{74}$$

Substituting (74) into (70) results in

$$i_s(t) = \frac{j\omega q N_s f_o(1 - f_o)}{(1 + j\omega f_o/c_n n_{so})} \frac{\delta n_s}{n_{so}}. \tag{75}$$

The ratio $\delta n_s/n_{so}$ is equal to the ac surface potential. This can be seen as follows:

$$\frac{\delta n_s}{n_{so}} = \frac{n_i \exp(u_s - u_B)\ \delta u_s}{n_i \exp(u_s - u_B)} = \delta u_s\ , \tag{76}$$

where n_i is the intrinsic carrier density and u_s is the silicon band bending or surface potential in units of kT/q and u_B the potential difference between mid-gap and Fermi level in the quasi neutral region of the silicon in units of kT/q. Expressing δu_s in volts is

$$\delta u_s = \frac{q}{kT}\ \delta \psi_s\ . \tag{77}$$

Equation (75) becomes

$$i_s(t) = j\omega \frac{q^2}{kT} \frac{N_s f_o(1 - f_o)\ \delta \psi_s}{(1 + j\omega f_o/c_n n_{so})}. \tag{78}$$

This can be written as

$$i_s(t) = Y_s\ \delta \psi_s\ , \tag{79}$$

where

$$Y_s = j\omega \frac{q^2}{kT} \frac{N_s f_o(1 - f_o)}{(1 + j\omega f_o/c_n n_{so})}. \tag{80}$$

APPENDIX B

B.1 *Random Fluctuations of Impurity Charges in the Silicon*

We shall use the model proposed by Shockley[23] for this calculation. Otherwise the procedure for finding the standard deviation of the surface potential is similar to that for the case of the surface charges.

Let $\bar{M}$ be the number of ionized acceptors in a characteristic volume in the silicon space-charge region. When $\bar{M}$ is large, the probability that there are M ionized acceptors in a characteristic volume is given by the Gaussian approximation of a Poisson distribution

$$P(M) = (2\pi\bar{M})^{-\frac{1}{2}} \exp [-(M - \bar{M})^2/2\bar{M}]. \tag{81}$$

The characteristic volume from Ref. 23 is a cube in which the length of a side is equal to the space-charge width W. The first transformation of the probability density function is from number of charges to number density

$$P(N_A) = P(M)\, dM/dN_A\ , \tag{82}$$

where N_A is the ionized acceptor density. The relation between M and N_A is

$$M = N_A W^3. \tag{83}$$

Combining (81), (82), and (83), the result of the first transformation is

$$P(N_A) = (2\pi\bar{N}_A W^3)^{-\frac{1}{2}} W^3 \exp [-W^3(N_A - \bar{N}_A)^2/2\bar{N}_A], \tag{84}$$

where $\bar{N}_A$ is the mean acceptor density.

The second transformation from number density to surface potential is

$$P(u_s) = P(N_A)\, dN_A/du_s\ . \tag{85}$$

Rewriting (31) and (32) results in

$$Q = C_{ox}(v_o + u_s/\beta) + (2q\epsilon_{si}N_A/\beta)^{\frac{1}{2}}[\exp(-u_s) + u_s - 1]^{\frac{1}{2}}. \tag{86}$$

Solving (86) for N_A is

$$N_A = \left(\frac{\beta}{2q\epsilon_{si}}\right) \frac{[Q - C_{ox}(v_o + u_s/\beta)]^2}{[\exp(-u_s) + u_s - 1]}. \tag{87}$$

Equation (87) is not a single-valued relation between N_A and u_s. We therefore restrict the problem as with the surface charges to the case where the fluctuations $N_A - \bar{N}_A$ are very small. For very small fluctuations, (87) can be differentiated assuming Q and oxide thickness to be uniform to get

$$dN_A = -\left\{\left(\frac{2N_A}{q\epsilon_{si}\beta}\right)^{\frac{1}{2}} \frac{C_{ox}}{[\exp(-u_s) + u_s - 1]^{\frac{1}{2}}} + \frac{N_A[1 - \exp(-u_s)]}{[\exp(-u_s) + u_s - 1]}\right\} du_s . \quad (88)$$

Equation (88) is now expressed as a function of u_s using (32)

$$C_D = \frac{\epsilon_{si}}{W} = \beta \frac{dQ_{sc}}{du_s} = \left[\frac{q\epsilon_{si}N_A\beta}{2[\exp(-u_s) + u_s - 1]}\right]^{\frac{1}{2}} [1 - \exp(-u_s)]. \quad (89)$$

Solving (89) for $\exp(-u_s) + u_s - 1$, we have

$$\exp(-u_s) + u_s - 1 = \frac{qN_A\beta}{2\epsilon_{si}} W^2[1 - \exp(-u_s)]^2. \quad (90)$$

Substituting (90) into (88) we get

$$dN_A = -\frac{2(WC_{ox} + \epsilon_{si})}{q\beta W^2[1 - \exp(-u_s)]} du_s . \quad (91)$$

Because $\bar{N}_A$ is given by (87) when $u_s = \bar{u}_s$, dN_A can be evaluated about $\bar{N}_A$ at each bias from

$$dN_A = -\frac{2[W(\bar{u}_s)C_{ox} + \epsilon_{si}]}{q\beta W^2(\bar{u}_s)[1 - \exp(-\bar{u}_s)]} du_s . \quad (92)$$

Equation (92) is the transformation equation sought. Replacing dN_A and du_s in (92) by the small fluctuations $N_A - \bar{N}_A$ and $u_s - \bar{u}_s$ respectively, we get

$$N_A - \bar{N}_A = -\frac{2[W(\bar{u}_s)C_{ox} + \epsilon_{si}]}{q\beta W^2(\bar{u}_s)[1 - \exp(-\bar{u}_s)]} (u_s - \bar{u}_s). \quad (93)$$

Combining (84), (85), (92), and (93), we get

$$P(u_s) = (2\pi\sigma_B^2)^{-\frac{1}{2}} \exp[-(u_s - \bar{u}_s)^2/2\sigma_B^2], \quad (94)$$

where the standard deviation σ_B is

$$\sigma_B = \frac{q\beta[\bar{N}_A W(\bar{u}_s)]^{\frac{1}{2}}[1 - \exp(-\bar{u}_s)]}{2[W(\bar{u}_s)C_{ox} + \epsilon_{si}]}. \quad (95)$$

B.2 *Random Fluctuations of Oxide Thickness*

Solving (86) for x, the oxide thickness,

$$\frac{\epsilon_{ox}}{C_{ox}} = x = \frac{\epsilon_{ox}(v_o + u_s/\beta)}{Q - \{(2q\epsilon_{si}N_A/\beta)[\exp(-u_s) + u_s - 1]\}^{\frac{1}{2}}}, \quad (96)$$

where ϵ_{ox} is the dielectric permittivity of the oxide layer. If Q is not zero, it will influence the fluctuations of surface potential caused by fluctuations in oxide thickness even if it is uniform. To get the influence of oxide thickness fluctuations independent of surface charge, we let $Q = 0$. Thus, (96) becomes

$$x = -\frac{\epsilon_{ox}(v_o + u_s/\beta)}{\{(2q\epsilon_{si}N_A/\beta)[\exp(-u_s) + u_s - 1]\}^{\frac{1}{2}}}. \tag{97}$$

Differentiating (97), assuming N_A to be uniform, results in

$$dx = -\frac{\epsilon_{ox}}{(2q\epsilon_{si}N_A/\beta)^{\frac{1}{2}}}\left\{\frac{1}{\beta[\exp(-u_s) + u_s - 1]^{\frac{1}{2}}} - \frac{(v_o + u_s/\beta)[1 - \exp(-u_s)]}{2[\exp(-u_s) + u_s - 1]^{\frac{3}{2}}}\right\} du_s . \tag{98}$$

Eliminating $\exp(-u_s) + u_s - 1$ from (98) by using (90) we get

$$dx = -\frac{W\epsilon_{ox} + x\epsilon_{si}}{q\beta N_A W^2[1 - \exp(-u_s)]}\, du_s . \tag{99}$$

Equation (97) gives $\bar{x}$ when $u_s = \bar{u}_s$ so that dx evaluated about $\bar{x}$ is

$$d\bar{x} = -\frac{W(\bar{u}_s)\epsilon_{ox} + \bar{x}\epsilon_{si}}{q\beta N_A W^2(\bar{u}_s)[1 - \exp(-\bar{u}_s)]}\, d\bar{u}_s . \tag{100}$$

$d\bar{x}$ is the standard deviation of x and $d\bar{u}_s$ the standard deviation of surface potential thus,

$$d\bar{u}_s = \sigma_x = \frac{qN_A\beta W^2(\bar{u}_s)[1 - \exp(-\bar{u}_s)]}{[W(\bar{u}_s)\epsilon_{ox} + \bar{x}\epsilon_{si}]}\, d\bar{x}. \tag{101}$$

APPENDIX C

There is a simple way to relate $C_s(\omega, \bar{u}_s)$ to measured $(G_p/\omega)_{\max}$. The distributed interface state network in Fig. 8(a) can be represented at one frequency and bias by an equivalent series R_sC_s network in which R_s and C_s are frequency and bias dependent. The equivalent circuit will then look like Fig. 6. If we let R_s and C_s be independent of frequency, G_p/ω vs frequency from (17) can be made to coincide with the measured G_p/ω vs frequency curve at each bias at $(G_p/\omega)_{\max}$ as shown in Fig. 32. Therefore, from (17), $C_s(\omega, \bar{u}_s) = 2(G_p/\omega)_{\max}$ at each bias and we obtain (54) by substituting this into (53) and solving for C_D.

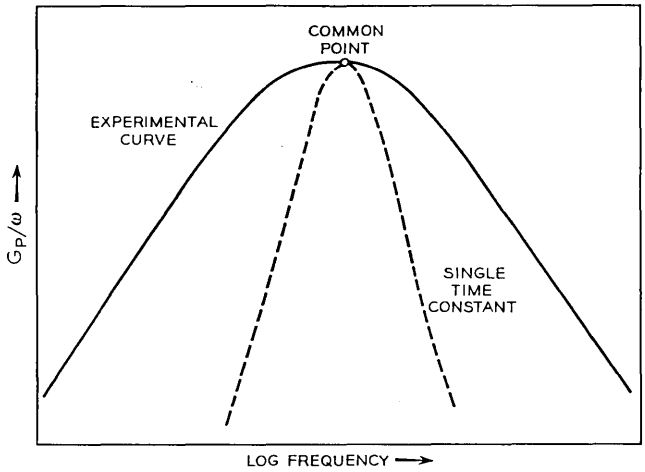


Fig. 32 — Schematic of G_p/ω vs log frequency showing how C_s is obtained from measured G_p/ω peak.

APPENDIX D

We shall calculate G_p/ω vs ω which would result from transitions between the conduction band near the surface and states distributed with uniform density into the oxide. The time constant distribution for tunnelling from (61) is

$$\tau_t = \tau_{ot} \exp (2K_o\xi), \tag{102}$$

where $\tau_{ot} = (n_{so}\bar{v}\sigma_o)^{-1}$.

The contribution to G_p/ω from states located between ξ and $\xi + d\xi$ in the oxide is

$$d(G_p/\omega) = (q/2\omega)(N_{ss}/x_L)\tau_{ot} \exp (-2K_o\xi) \cdot \ln [1 + \omega^2\tau_{ot}^2 \exp (4K_o\xi)]\, d\xi, \tag{103}$$

where (102) has been substituted for τ in (20) for the continuum of states, N_{ss}/x_L is the uniform density of states in the oxide, cm^{-3}-eV^{-1}, and x_L is the oxide thickness, cm.

Total G_p/ω from all the states in the oxide therefore, is

$$G_p/\omega = (q/2\omega)(N_{ss}/x_L)\tau_{ot} \int_0^{x_L} \exp (-2K_o\xi) \cdot \ln [1 + \omega^2\tau_{ot}^2 \exp (4K_o\xi)]\, d\xi. \tag{104}$$

To get G_p/ω vs ω, (104) is numerically integrated on an IBM 7094 computer using the trapezoidal rule. The result is shown in Fig. 33. Because we are interested only in the shape of the curve, the frequency scale is arbitrary.

G_p/ω is normalized to eliminate N_{ss}/x_L . A plot of normalized G_p/ω obtained from experiment is also shown in Fig. 33 for comparison.

The shape of the curve calculated from the tunnelling model is determined essentially by (102). The shortest time constant in this distribution is when $\xi = 0$ which is right at the interface. G_p/ω drops off rapidly as the period of the applied signal becomes comparable to and shorter than this time constant. When the period of the applied signal is long compared to the shortest time constant in (102), G_p/ω becomes a constant as we have assumed a constant density of states into the oxide. The main purpose of assuming a uniform density of states into the oxide is to illustrate the asymmetry of the G_p/ω vs log ω curve introduced by the tunnelling time constant distribution (102). The G_p/ω vs log ω curve will go through a maximum if the density of states decreases with distance into the oxide. As long as the density of states decreases more slowly with distance into the oxide than the time constants given by (102) increase, G_p/ω will decrease faster on the high frequency side of the peak than on the low-frequency side.

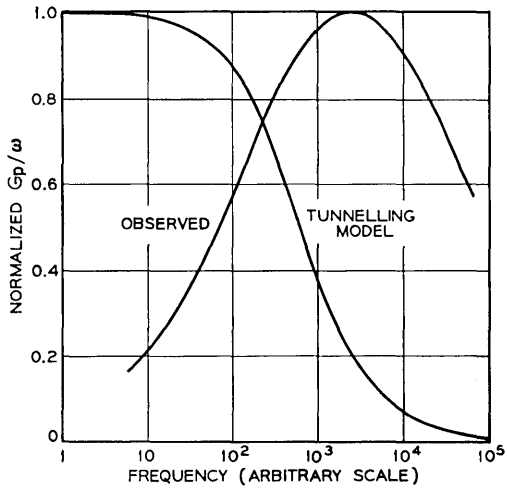


Fig. 33 — Normalized G_p/ω vs log frequency for tunnelling calculated from (104) and normalized G_p/ω vs log frequency obtained for measurement. Frequency scale is arbitrary to compare the shapes of two curves.

The G_p/ω vs log ω curve obtained from measurement is symmetric about its peak. A distribution of states into the oxide which would would give a G_p/ω vs log ω curve symmetric about its peak using the tunnelling model would be a peaked density distribution having its maximum some distance into the oxide. In addition, the width of this distribution would have to be considerably smaller than $1/2K_o$ to get G_p/ω vs log ω curves in the frequency range between 50 Hz and 500 kHz. Such a distribution would be rather arbitrary. Thus, it can be seen from the symmetry of the observed G_p/ω vs log ω curves that tunnelling to states distributed into the oxide is not likely to be the dominant cause of time constant dispersion in the samples measured.

LIST OF SYMBOLS

a	Maximum amplitude of the ac surface potential in volts.
b	Maximum amplitude of applied ac signal in volts.
C	Equivalent parallel capacitance of interface state branch of equivalent circuit having a single time constant in farads/cm^2.
CB	Representation of conduction band.
C_D	Depletion layer capacitance in farads/cm^2.
C_I	Inversion layer capacitance in farads/cm^2.
C_m	Measured equivalent parallel capacitance of the MIS capacitor in farads/cm^2.
c_n	Electron capture probability in cm^3/sec.
C_{om}	Low-frequency capacitance of MIS capacitor in farads/cm^2.
C_{ox}	Oxide layer capacitanec in farads/cm^2.
c_p	Hole capture probability in cm^3/sec.
C_p	Equivalent parallel capacitance of depletion layer and interface state branch of equivalent circuit in farads/cm^2.
C_s	Interface state capacitance associated with a single time constant in farad/cm^2.
$d\nu$	Number of characteristic areas in which the number of surface charges is between N and $N + dN$ or the surface potential is between u_s and $u_s + du_s$.
E_F	Fermi level in units of kT/q.
E_i	Mid-gap energy in units of kT/q.
e_n, e_p	Emission constants for electrons and holes in sec^{-1}.
f_m	Maximum value of the perturbation of the Fermi function by the ac signal.
f_o	Fermi function at the dc bias.

$f(t)$ Fermi function as a function of time.

G_m Measured equivalent parallel conductance of MIS capacitor in mhos/cm^2.

$G_n(t)$ Emission rate of electrons from a single level interface state in cm^{-2}-sec^{-1}.

G_p Equivalent parallel conductance of the interface state branch of the equivalent circuit in mhos/cm^2.

G_p' Equivalent parallel conductance corrected for C_{ox} in mhos.

$G_p(t)$ Emission rate of holes from a single level interface state in cm^{-2}-sec^{-1}.

G_s Majority carrier capture conductance of a single level interface state in mhos/cm^2.

i_s Current density charging the interface states in amp/cm^2.

i_{sc} Current density charging the space-charge layer in amp/cm^2.

i_T Current density through MIS capacitor in amp/cm^2.

k Boltzman's constant in eV-coul/°K.

M Number of ionized acceptors in a characteristic volume.

$\bar{M}$ The mean of M.

N Number of built-in charges and charged interface states in a characteristic area.

$\bar{N}$ The mean of N.

N_A Acceptor impurity density in cm^{-3}.

$\bar{N}_A$ Mean ionized acceptor density in cm^{-3}.

N_D Donor impurity density in cm^{-3}.

n_i Intrinsic carrier concentration in cm^{-3}.

N_s Density of a single level interface state in cm^{-2}.

n_{so} Electron density established by the dc bias in cm^{-3}.

N_{ss} Density of interface states, $dN_s/d\psi$ in cm^{-2}-eV^{-1}.

$n_s(t)$ Time dependent electron density at the silicon surface in cm^{-3}.

$P(M)$ Probability that there are M ionized acceptors in a characteristic volume.

$P(N)$ Probability of finding N charges in a characteristic volume.

$P(Q)$ Probability that the surface-charge density is Q in cm^2/coul.

p_{so} Hole density established by the dc bias in cm^{-3}.

$p_s(t)$ Time dependent hole density at the silicon surface in cm^{-3}.

$P(u_s)$ Probability that surface potential in a characteristic area is u_s.

q Electronic charge in coulomb.

Q Sum of built-in and interface state charge densities in coul/cm^2.

$\bar{Q}$ Mean surface-charge density in coul/cm^2.
Q_f Fixed charge density in the oxide in coul/cm^2.
Q_s Total interface state charge density in coul/cm^2.
Q_{sc} Silicon space-charge density in coul/cm^2.
Q_T Total charge density in coul/cm^2.
$R_{n,s}$; $R_{p,s}$ Electron and hole capture resistance of a single level in ohms-cm^2.
$R_n(t)$, $R_p(t)$ Capture rate of electrons and holes by a single level interface state in cm^{-2}-sec^{-1}.
R_s Majority carrier capture resistance of a single level interface state in ohm-cm^2.
t time in seconds.
T Absolute temperature in °K.
u Potential difference between mid-gap and an interface state level in units of kT/q.
u_B $\begin{cases} = \ln (n_i/N_D), \text{ for } n\text{-type silicon} \\ = \ln (N_A/n_i), \text{ for } p\text{-type silicon.} \end{cases}$
u_F Potential difference between mid-gap and Fermi level at silicon surface in units of kT/q
u_o $\ln \omega/c_n n_i$.
u_s Surface potential in units of kT/q.
$\bar{u}_s$ Mean surface potential in units of kT/q.
VB Representation of valence band.
$\bar{v}$ Average thermal velocity of holes and electrons in cm/sec.
$v_a(t)$ Time dependent voltage applied to MIS capacitor in volts.
v_o dc bias across MIS capacitor in volts.
v_{o1} Arbitrary dc bias in volts.
W Depletion layer thickness in cm.
$\bar{x}$ Mean oxide thickness measured from the interface in cm.
y $\ln \omega\tau_m = u_B + u_o - u_s$.
y_m $\bar{u}_s - u_s + \ln 2.5$.
Y_s Admittance of a single level interface state in mhos/cm^2.
Y_{ss} Admittance of an interface state continuum in mhos/cm^2.
Z_s MIS capacitor impedance less reactance of the oxide layer in ohms-cm^2.
α Characteristic area, cm^2
β $= 38.5\ q/kT$ in volt^{-1}.
Δ Additive constant in volts.
δf Time dependent part of Fermi function.
δn_s Incremental electron surface-charge density induced by δv_a in cm^{-3}.

δv_a	ac signal applied to MIS capacitor in volts.
$\delta\psi_s$	Silicon band bending by ac perturbation in volts.
ϵ_{si} , ϵ_{ox}	Dielectric permittivity of silicon and SiO_2 layer respectively in farad/cm.
ξ	Distance into oxide measured from interface in cm.
σ_B	Standard deviation of surface potential caused by fluctuations of ionized acceptors.
σ_n , σ_p	Electron and hole capture cross sections in cm^2.
σ_s	Standard deviation of surface potential caused by fluctuations of surface charges.
σ_x	Standard deviation of surface potential caused by oxide thickness fluctuations.
τ	Majority carrier time constant of a single level interface state, R_sC_s in seconds.
τ_m	Measured interface state time constant in seconds.
τ_{mo}	Time constant at $\bar{u}_s = u_B$.
$\tau_{n,s}$; $\tau_{p,s}$	Electron and hole time constant for a single level in seconds.
ψ	Energy in electron volts.
ψ_B	$(kT/q)\ u_B$ in volts.
ψ_s	Band bending or surface potential in volts.
$\bar{\psi}_s$	Mean surface potential in volts.
ψ_{se}	Surface potential plus an additive constant in volts.
ψ_{so}	Surface potential at the dc bias in volts.
ω	Angular frequency $= 2\pi f$ in sec^{-1}.

REFERENCES

1. Williams, R., Photoemission of Electrons from Silicon into Silicon Dioxide, Phys. Rev., *140*, 1965, A569.
2. Heiman, F. P. and Warfield, G., The Effects of Oxide Traps on the MOS Capacitance, IEEE Trans. Electron Devices, *ED-12*, 1965, p. 167.
3. Snow, E. H., Grove, A. S., Deal, B. E., and Sah, C. T., Ion Transport Phenomena in Insulating Films, J. Appl. Phys., *36*, 1965, p. 1664.
4. Goetzberger, A., Improved Properties of Silicon Dioxide Layers Grown Under Bias, J. Electrochem. Soc., *113*, 1966, p. 138.
5. Grove, A. S., Deal, B. E., Snow, E. H., and Sah, C. T., Investigation of Thermally Oxidized Silicon Surfaces Using Metal-Oxide-Semiconductor Structures, Solid State Electronics, *8*, 1964, p. 145.
6. Deal, B. E., Grove, A. S., and Snow, E. H., Characteristics of the Surface State Charge Q_{ss} of Thermally Oxidized Silicon, J. Electrochem. Soc., *114*, 1967, p. 266.
7. Terman, I. M., An Investigation of Surface States at a Silicon/Silicon Oxide Interface Employing Metal-Oxide-Silicon Diodes, Solid State Electronics, *5*, 1962, p. 285.
8. Zaininger, K. H. and Warfield, G., Limitations of the MOS Capacitance Method for the Determination of Semiconductor Surface Properties, IEEE Trans. Electron Devices, *ED-12*, 1965, p. 179.
9. Nicollian, E. H. and Goetzberger, A., MOS Conductance Technique for Measuring Surface State Parameters, Appl. Phys. Letters, *7*, 1965, p. 216.

10. Nicollian, E. H. and Goetzberger, A., Lateral AC Current Flow Model for Metal-Insulator-Semiconductor Capacitors, IEEE Trans. Electron Devices, *ED-12,* 1965, p. 108.
11. Hofstein, S. R. and Warfield, G., Physical Limitations on the Frequency Response of a Semiconductor Surface Inversion Layer, Solid-State Electronics, *8,* 1965, p. 321.
12. Lehovec, K. and Slobodskoy, A., Impedance of Semiconductor-Insulator-Metal Capacitors, Solid-State Electronics, *7,* 1964, p. 59.
13. Berz, F., Variation with Frequency of the Transverse Impedance of Semiconductor Surface Layers, J. Phys. Chem. Solids, *23,* 1962, p. 1795.
14. Goetzberger, A. and Nicollian, E. H., Temperature Dependence of Inversion Layer Frequency Response in Silicon, B.S.T.J., *46,* March, 1967, pp. 513-522.
15. Sah, C. T., Solid State Electronics Laboratory Technical Report No. 1, Section 16.1.4, Electrical Engineering Laboratories, University of Illinois, Urbana, Ill., 1964.
16. Lindner, R., Semiconductor Surface Varactor, B.S.T.J., *41,* May, 1962, pp. 803-831.
17. Moll, J. L., Variable Capacitance with Large Capacity Change, 1959 IRE Wescon Conv. Rec.—Part 3, Electron Devices, pp. 32-36.
18. Shockley, W. and Read, W. T., Statistics of the Recombination of Holes and Electrons, Phys. Rev., *87,* 1952, p. 835.
19. Frolich, H., *Theory of Dielectrics,* Oxford at the Clarendon Press, 1958, 2nd Ed., Chapter III.
20. Yager, W. A., The Distribution of Relaxation Times in Typical Dielectrics, Physics, *7,* 1936, p. 434.
21. Lehovec, K., Slobodskoy, A., and Sprague, J., Field Effect Capacitance Analysis of Surface States on Silicon, Phys. Status Solidi, *3,* 1964, p. 447.
22. Lehovec, K., Frequency Dependence of the Impedance of Distributed Surface States in MOS Structures, Appl. Phys. Letters, *8,* 1966, p. 48.
23. Shockley, W., Problems Related to p-n Junctions in Silicon, Solid-State Electronics, *2,* 1961, p. 35.
24. Davenport, W. B., Jr., and Root, W. L., *An Introduction to the Theory of Random Signals and Noise,* McGraw-Hill Book Co., Inc., New York, 1958, Chapter 3.
25. Goetzberger, A., McDonald, B., Haitz, R. H., and Scarlett, R. M., Avalanche Effects in Silicon p-n Junctions. II. Structurally Perfect Junctions, J. Appl. Phys., *34,* 1963, p. 1591.
26. Cramer, H., *The Elements of Probability Theory,* John Wiley & Sons, New York, 1955, Chapter 6.
27. von Hippel, A. R., *Dielectrics and Waves,* John Wiley & Sons, New York, 1954, Appendix I.
28. Berglund, C. N., Surface States at Si-SiO_2 Interfaces, IEEE Trans. Electron Devices, *ED-13,* 1966, p. 701.
29. Balk, P., Effects of Hydrogen Annealing on Silicon Surfaces, Extended Abstracts, Electronics Div. Electrochem. Soc., *14,* No. 1, 1965, p. 237.
30. Gray, P. V. and Brown, D. M., Density of SiO_2-Si Interface States, Appl. Phys. Letters, *8,* 1966, p. 31.
31. Kuper, A. B. and Nicollian, E. H., Effect of Oxide Hydration on Surface Potential of Oxidized P-Type Silicon, J. Electrochem. Soc., *112,* 1965, p. 528.
32. Many, A., Goldstein, Y., and Grover, N. B., *Semiconductor Surfaces,* John Wiley and Sons, New York, 1965, Ch. 9.
33. Rzhanov, A. V. and Neizvestnyi, I. G., Soviet Physics—Solid State, *3,* 1962, p. 2408.
34. Kurskii, Yu. A., Soviet Physics—Solid State, *4,* 1963, p. 1922.
35. Crowell, C. R., Interpretation of Tunnel and Capacitance Measurements in the Presence of Dielectric Film-Thickness Fluctuations, Appl. Phys. Letters, *8,* 1966, p. 328.

The Nonlinearity of the Reverse Current-Voltage Characteristics of a p-n Junction Near Avalanche Breakdown

By S. M. SZE and R. M. RYDER

(Manuscript received January 23, 1967)

For nonlinear applications such as high-speed switching, a device figure of merit is γ, the ratio of the second derivative to the first derivative of the current-voltage (I-V) characteristic, or $\gamma \equiv (d^2I/dV^2)/(dI/dV)$. At room temperature, the value of γ for an ideal forward-bias Schottky diode is about 40 V^{-1}. It is shown that although the ideal reverse breakdown characteristic could give a value of γ greater than 40 V^{-1}, because of the statistical distribution of impurities, the effect of space-charge resistance, and other complications, much lower values of γ are expected. Furthermore, the nonlinear characteristic is noisy, relatively slow, and causes some power consumption. It appears, therefore, that this nonlinearity is not likely to supersede Schottky barrier diodes in high-speed switching applications. It does not, however, rule out the possibility of microwave generation application.

For nonlinear applications such as high-speed switching, a device figure of merit is γ, the ratio of the second derivative to the first derivative of the current-voltage characteristic. The value of γ is a measure of the degree of nonlinearity, normalized to the operating admittance level. It is used here to compare the nonlinearity of a reverse-biased diode near breakdown to that of a forward-biased p-n junction or a Schottky barrier.

The current-voltage characteristic and γ for a forward-biased p-n junction or a Schottky barrier are given by

$$I = I_0[e^{qV/nkT} - 1] \tag{1}$$

$$\gamma \equiv (d^2I/dV^2)/(dI/dV) = \frac{q}{nkT}. \tag{2}$$

At room temperature, the value of γ for an ideal Schottky barriers ($n = 1$) is about 40 V^{-1} independent of bias.

The present note is undertaken to answer the following question: Is it possible at room temperature to have a value of γ greater than 40 V^{-1} for a p-n junction near its avalanche breakdown voltage? The answer is "yes" for an ideal isothermal breakdown characteristic and "no" for practical considerations.

Near avalanche breakdown the isothermal current-voltage characteristic without space-charge effect is given by

$$I = I_0 M = \frac{I_0}{1 - \int_0^W \alpha \exp\left[-\int_0^x (\alpha - \beta)\, dx'\right] dx}, \tag{3}$$

where M is the multiplication factor, W the depletion width depending on the applied voltage, and α and β are the ionization rates of electrons and holes, respectively.[1]

The value of γ computed from (3) is plotted in Fig. 1 [curve (a)] as a function of M for a silicon p^+n junction with n-type background doping of 5×10^{16} cm^{-3} and a breakdown voltage of 19.48 volts.[2] It is clear that the value of γ exceeds 40 V^{-1} at $M \cong 65$ corresponding to a voltage of 19.43 volts approximately 50 mV ($\equiv \Delta V$) smaller than the breakdown voltage. For a background doping of 1.5×10^{17} cm^{-3}, the breakdown voltage is 9.546 volts.[2] The value of γ exceeds 40 V^{-1} at $M \cong 50$ [curve (b)], corresponding to a voltage some 50 mV smaller than the breakdown voltage. Similar results are obtained for other dopings and different semiconductors.

Now, let us consider the space-charge effect. The space charge of holes and electrons produced by the avalanche generates a counter emf which reduces the field across the multiplying region.[3] Since the voltage reduction is proportional to the current, it is represented as a resistance; furthermore, it increases as the square of the length of the region in which the space charges accumulate.[4] For the above devices [curve (a)] with an n region approximately equal to the depletion width at breakdown, and a circular area of one mil in diameter, the equivalent space-charge resistance R_{sc} is about 50 ohms. The incorporation of the space-charge resistance will increase the applied voltage for any current level, thus, reducing the value of γ at any given M. The computed result is also shown in Fig. 1 [curve (c)]. One notices that γ reaches a maximum value of about 500 V^{-1} at an $M \cong 10^3$. For the above device area with a saturation current density of

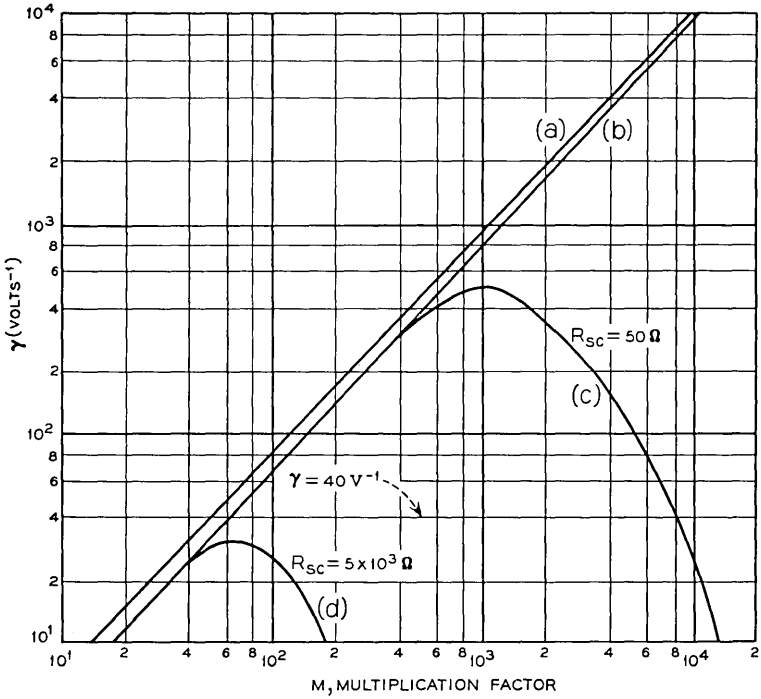


Fig. 1 — γ versus multiplication factor. (a) p^+n junction in silicon with background doping 5×10^{16} cm^{-3} and without space-charge effect. (b) p^+n junction in silicon with background doping 1.5×10^{17} cm^{-3} and without space-charge effect. (c) Same as (a) with 50-ohms space-charge resistance. (d) Same as (a) with 5000 ohms space-charge resistance.

about 10^{-9} amp/cm^2, the current at $M = 10^3$ is still very small, of the order of nanoamperes. For larger M, γ decreases rapidly to zero within 10 mV. If one increases the width of the n region, R_{sc} also increases; eventually for large enough R_{sc}, such as shown in Fig. 1 curve (d) the value of γ will be smaller than 40 V^{-1} for any M.

The effects of statistical spatial fluctuations of donor and acceptor ions on the breakdown voltages have been considered by Shockley.[5] It is found that this randomness leads to a characteristic fluctuation voltage of about 300 mV in silicon. The effect of these fluctuations in a p-n junction is to produce local regions with breakdown about a few hundred millivolts lower than the average in uncompensated material, and larger by $[(N_d + N_a)/(N_d - N_a)]^{1/2}$ in compensated material.

It is clear that this voltage fluctuation is an order of magnitude

larger than the value of ΔV which is the difference between the breakdown voltage ($M \to \infty$) and the voltage at which $M \cong 50$. Consequently, in a practical junction with this effect present, the curvature will be very much less than the ideal calculation of Fig. 1.

We conclude that the normalized curvature γ could be large for a junction near breakdown, but only if the junctions were ideal. It would be possible for γ to exceed 40/V provided space charge, heating, microplasma and statistical effects are small. This ideal high curvature would be limited to regions within a few millivolts of breakdown, where the multiplication ratio M is high, $\approx$50 or more.

Practical considerations severely limit the use of this nonlinearity:

(*i*) Because of statistical fluctuations in impurity distributions, there is a characteristic spatial fluctuation in breakdown voltage of a few hundred millivolts in silicon, as pointed out by Shockley.[5] This alone drastically reduces γ below the ideal value.

(*ii*) Series resistance, either ohmic or space charge, limits the region of high curvature to small currents, typically a few microamperes for a 1-mil diameter junction. Therefore, with usual shunt capacitances, the response time tends to be slower than for a good Schottky Barrier diode.

(*iii*) Microplasmas, if present, would cause erratic curvature and high noise.

(*iv*) The noise is high compared to a forward-biased or zero-biased Schottky Barrier. So is the power dissipation.

(*v*) Negative resistance complications are possible.[6]

In summary, although the ideal reverse breakdown characteristic could give a value of γ greater than 40 V^{-1}, because of the statistical distribution of impurities, the effect of space-charge resistance, and other complications above, much lower values of γ are expected. Furthermore, the nonlinear characteristic is noisy, relatively slow, and associates with some power consumption. It appears therefore, that this nonlinearity is not likely to supersede Schottky Barrier diodes in high-speed switching applications. It does not, however, rule out the possibility of microwave generation application. This latter application is not discussed here; it would utilize not only the large nonlinear resistance of the I-V characteristic near breakdown but also the nonlinear reactance and negative conductance due to the interaction between transit time and the avalanche process in the device.[6]

REFERENCES

1. Lee, C. A., Logan, R. A., Batdorf, R. L., Kleimack, J. J., and Wiegmann, W., Ionization Rates of Holes and Electrons in Silicon, Phys. Rev., *134,* 1964, p. 761.
2. Sze, S. M. and Gibbons, G., Avalanche Breakdown Voltages of Abrupt and Linearly Graded p-n Junctions in Ge, Si, GaAs, and GaP, Appl. Phys. Letters, *8,* 1966, p. 111.
3. Gummel, H. K. and Scharfetter, D. L., Avalanche Region in IMPATT Diodes, B.S.T.J., *45,* December, 1966, pp. 1797-1827.
4. Sze, S. M. and Shockley, W., Unit-Cube Expression for Space-Charge Resistance, B.S.T.J., *46,* May-June, 1967, pp. 837-842.
5. Shockley, W., Problems Related to p-n Junctions in Silicon, Solid-State Electronics, *2,* 1961, p. 35.
6. Misawa, T., Negative Resistance in p-n Junction Under Avalanche Breakdown Conditions, IEEE Trans. Electron Devices, *ED-13,* 1966, p. 143.

Subjective Evaluation of Transmission Delay in Telephone Conversations

By E. T. KLEMMER

(Manuscript received February 1, 1967)

An earlier experiment by Riesz and Klemmer on the effect of pure-transmission delay upon natural telephone conversations was extended in a test with more than double the time period and number of calls. The previous finding of little or no adverse reaction to round-trip pure delays of 600 and 1200 msec alone was confirmed. The previous finding of a large increase in dissatisfaction with both of these delays following exposure to 2400 msec was not obtained. Exposure to delays of 2400 msec led to no dissatisfaction with later calls at 600 msec, but some rejections at 1200 msec did occur. There is no contradiction of other results on normal telephone circuits with 2-wire terminations (and related echo sources, paths, and suppressors) wherein customer dissatisfaction is greater with 600 msec delay than with the much shorter delay of a normal long-distance circuit.

A previous paper by Riesz and Klemmer[1] in this journal described laboratory experiments on the effect of transmission delay upon the quality of telephone circuits for normal conversation. These experiments were of two types: (*i*) "pure delay" tests in which long transmission times were employed, but the side effects of echo and echo suppressors were avoided by using special 4-wire telephone circuits and (*ii*) "2-wire" tests which used long transmission times in normal 2-wire circuits (or circuits with 2-wire terminations) with echo sources and echo suppressors.

Since the publication of these experiments, several evaluations outside the laboratory have been done on circuits with long transmission times and naturally-occurring telephone calls (e.g., Helder,[2] Klemmer[3]). These studies have borne out the earlier laboratory finding of considerable dissatisfaction with 2-wire circuits for round-trip delays of 600 msec or more. The pure delay condition could not be evaluated in the field tests since it requires complete separation of the transmitting and receiving paths which is not available in the normal telephone

plant.* Even though the pure delay condition is not obtainable in the telephone network used by the public for domestic and overseas calls, it is of interest because it represents a limiting point for the effect of transmission delay on the quality of the circuits for conversation. It is not likely that any changes in echo suppressors or other circuit devices would produce better transmission quality than a 4-wire system with the same transmission time. Thus, if the degradation noted in the natural 2-wire circuits with long delay were present in similar amount on 4-wire circuits of the same delay, there would be little hope of improving transmission quality by improving echo control methods. If, on the other hand, the effects of pure delay alone are much less than that due to delays, echo, and echo suppression, then improvement in echo control methods is definitely indicated.

The results of the previous laboratory study with pure delay were very limited in number of people and calls, and also confounded by the introduction of extremely long delays (2400 msec) in the middle of the experiment. The data had shown little or no dissatisfaction with pure delays of 600 and 1200 msec prior to the introduction of the 2400-msec delay, but showed an increasing rejection of circuits with the lesser delays after exposure to the 2400-msec condition.

The chief purpose of the present experiment was to see if an increasing rejection rate would occur with continued exposure to pure delays of 600 and 1200 msec only. Therefore, the present experiment repeats the 12 weeks of the previous pure delay study but without the introduction of the 2400-msec delay.

After the 12 weeks of alternate days of 600- and 1200-msec delay, periods of 1800- and 2400-msec delay were inserted to re-evaluate the effect that these longer delays would have upon the users reaction to following days of 600- and 1200-msec delay.

I. SIMULATION APPARATUS: SIBYL

The simulator called Sibyl, which permits the insertion of experimental circuits into naturally-occurring telephone calls, was the same as that employed in the previous study[1] and is described by Irwin.[7] Elimination of echo and echo suppressors is attained by converting all telephone instruments to full 4-wire operation, separating the transmit and receive paths. Normal sidetone was provided within each telephone set.

*Other laboratory tests on the effect of pure delay upon conversational tasks have been reported by Bricker,[4] Krauss and Bricker,[5] and Vartabedian.[6] These did not involve naturally-occurring telephone calls, and thus, are not directly comparable to the tests described here.

II. SUBJECTS

Twenty administrative employees of Bell Telephone Laboratories were selected on the basis of questionnaires sent to several hundred people asking about frequencies of telephone calls to other extensions in Bell Laboratories. These people were selected to form a group with a high reported-rate of calling each other since the delay circuits could only be used when they called each other.

III. INSTRUCTIONS

The participants were told that some of their calls would go over special experimental circuits. They were not told which calls would be affected or anything about the nature of the experimental circuits. They were told that if they found any circuit "unsatisfactory for normal telephoning" they should dial the digit "4" without hanging up or breaking the connection, and the standard circuit would be restored. The instructions called for the originating party to reject the circuit, but actually either party could reject the experimental circuit and the few instances of rejection by the called party were also counted as rejected calls.

IV. EXPERIMENTAL DESIGN

The schedule of delays on the experimental circuits was as follows:

Weeks 1 through 12	600 and 1200 msec on alternate days
Weeks 13 through 14	1800 msec every day
Weeks 15 through 16	2400 msec every day
Weeks 17 through 26	600 and 1200 msec on alternate days.

The delay for the day was inserted on each call made by one subject to another subject unless the simulator was already in use. No calls involving other stations could be put over the experimental circuits (because of the 4-wire requirement), and therefore, only a small percentage of any subject's calls went over the experimental circuits. The subjects were not told of this limitation and none reported knowledge of it in the post-test interviews.

V. RESULTS

The percentage of calls rejected for each two-week period of the experiment is plotted in Fig. 1 for each delay separately for the first

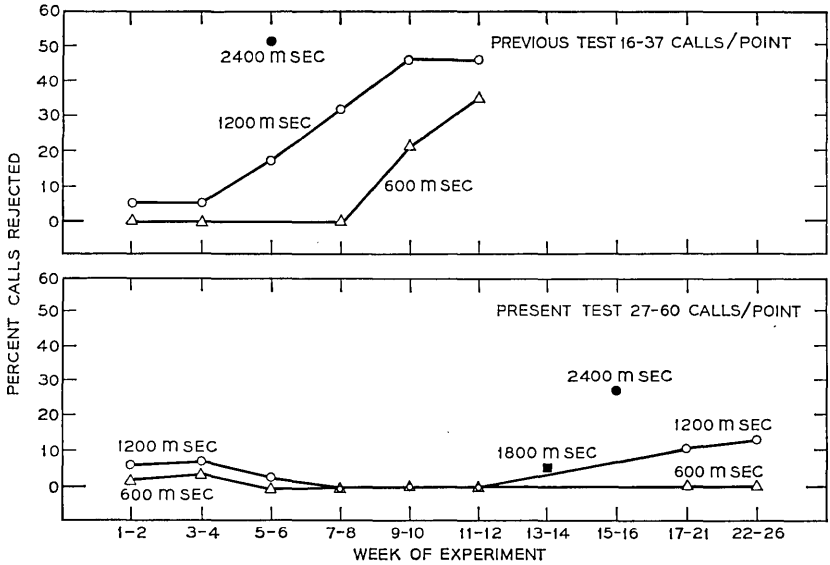


Fig. 1 — Rate of rejection of circuits as a function of weeks of the experiment from Riesz and Klemmer[6] and for the present experiment. Combined data from 18 and 20 subjects, respectively.

12 weeks of the test. The data from the previous study[1] is also shown. Clearly, the rising rejection rate of the previous study was not found, indeed, there were no rejections at either 600 or 1200 msec for weeks 7 through 12. In the first 12 weeks of the present study, a total of 527 calls were made over the delay circuits. This compares with 323 calls made during the entire earlier study. Thus, it is clear that increasing rejection of pure delays of 600 and 1200 msec is not to be expected from repeated exposures to these delays only.

The results during and after the longer delays in the experiment are also shown in Fig. 1. Two weeks at 1800 msec resulted in 5 percent rejections (3 calls of 60). Two weeks at 2400 msec resulted in 27 percent rejections (14 calls of 52). Six different people rejected calls at 2400 msec. Eighteen of the 20 subjects talked over the 2400-msec delay, and these 18 people made 97 percent of the calls over experimental circuits during the final 10 weeks when 600- and 1200-msec delays were again used on alternate days. Thus, the data following 2400-msec delay comes almost entirely from people who had been exposed to 2400 msec.

Because of vacations and illnesses there are not enough data during the final 10 weeks to plot biweekly points. Samples of comparable size to the first 12 weeks are obtainable by using two 5-week periods and these points are shown in Fig. 1. No additional rejections of the 600-msec circuit occurred during this period. Thus, there was no "sensitization" to that delay. The 1200-msec condition led to 12 percent rejection (7 of 57 calls) following exposure to long delays. (Note that in Fig. 1 the small difference in percentages between the two final periods at 1200 msec is not statistically meaningful.) The occurrence of some rejections at 1200 msec following exposure to the longer delays indicates some sensitization since 1200 msec had zero rejections from 6 weeks (94 calls) prior to the long delays.

Telephone interviews were conducted with the 16 subjects who were available at the end of the test. They were asked if they, or the people with whom they talked, had difficulty in talking or hearing on calls within the PBX during the time of the test, or if they noticed anything different about any calls. Eight (half) of the respondents said they had no difficulty and noticed nothing different about their calls. Eight reported difficulties due to: low volume (2); slow answer (2); fading (1); fuzziness (1); noise (1); and another party could not hear (1). One person, in addition, reported hearing echo (perhaps due to an air path feedback between receiver and transmitter). Of those who reported difficulty, one rated the condition not objectionable, two mildly objectionable, one moderately objectionable, and three seriously objectionable. One of those reporting seriously objectionable difficulty had never rejected a delay call, but several times dialed "4" on normal calls. He said the trouble was loss and dialing "4" did not help.

The interview data cannot be taken as a reliable measure of the circuit quality for three reasons: (*i*) The questions related to a large population of calls, only a few of which were actually over the delay circuits, (*ii*) The test lasted several months, and the subjects could hardly be expected to sort out accurately and remember all individual calls, and (*iii*) The subjects reported difficulties and attempted rejections on normal calls not involving the delay circuit at all. The interviews do, however, show that most of the time the participants were not aware that there was anything different about their connections when delays of 600 or 1200 msec were inserted. Indeed, eight people reported that they noticed nothing different about their circuits even though this group had actually talked on a total of 85 calls with 1800- or 2400-msec delay and many more calls at the smaller delays.

VI. DISCUSSION AND CONCLUSIONS

The first 12 weeks of the present study repeated the procedure of the previous pure delay tests except that the intermediate exposure to 2400 msec was omitted. Under these conditions of exposure to only 600 and 1200 msec, no increased dissatisfaction with delay circuits developed even though a much larger number of delay calls were made in the new study. This result alone obviously implies that exposure to the 2400-msec delay was responsible for the later rejections of the 600- and 1200-msec conditions in the previous study. The results of the later weeks of the present study provide limited support for this hypothesis, however. After exposure to 1800- and 2400-msec delay, *no* rejections of the 600-msec condition occurred but 12 percent of the 1200-msec calls were rejected. In the previous study, after exposure to 2400-msec delay, the 600-msec condition was rejected in 25 percent of the calls, and the 1200-msec condition was rejected in 43 percent of the calls.

In view of the differences in results between the two studies regarding the influence of exposure to 2400-msec delay, it might be best to withhold judgment about the magnitude of the sensitization effect. There is no disagreement, however, on the more direct and important question about user's reaction to pure-transmission delays of 600- and 1200-msec round-trip delays when these are not confounded with the longer delays or speech-operated devices. Users are very seldom disturbed by these pure delays as is indicated by the fact that during the second 6 weeks of the present study the participants completed more than 200 calls without a single rejection.

This conclusion is supported by the field test results[2, 3] which show for round-trip delays of 600 msec (on 2-wire circuits with echo sources and echo suppressors) that less than 1 percent of the people interviewed immediately after a call over the delay circuit said anything which implies an awareness of the delay itself. This is true despite the fact that 25 percent or more of the respondents report some difficulty in talking or hearing on the 600-msec circuit and only half that many report difficulty on circuits with delays less than 100-msec delay (normal overseas cable circuits).

Although users are not aware of a transmission delay of 600 msec it is clear that the delay must affect the conversational patterns in such a way as to cause other types of difficulty in actual 2-wire circuits, difficulties such as speech mutilation by echo suppressors. In addition, there is evidence that for tasks other than naturally occur-

ring conversation, delays of 600 to 1800 msec may significantly lower performance (Krauss and Bricker,[5] Vartabedian[6]). Studies of speech dynamics under various transmission delays are underway to understand this effect better.

VII. ACKNOWLEDGMENTS

M. A. Pilla was responsible for the Sibyl simulator. A. Y. Kimura conducted the initial survey of potential participants. A. P. Winnicky supervised the data collection.

REFERENCES

1. Riesz, R. R. and Klemmer, E. T., Subjective Evaluation of Delay and Echo Suppressors in Telephone Communications, B.S.T.J., *42*, November, 1963, pp. 2919–2941.
2. Helder, G. K., Customer Evaluation of Telephone Circuits with Delay, B.S.T.J., *45*, September, 1966, pp. 1157–1191.
3. Klemmer, E. T., Transmission Quality for Conversation in Telephoning via Satellite, Human Factors, in press.
4. Bricker, P. D., *Satellite Communications Physics*, Chap. 5, Foster, R. M. (Ed.), Murray Hill, N. J., Bell Telephone Laboratories, 1963.
5. Krauss, R. M. and Bricker, P. D., The Effects of Transmission Delay on the Efficiency of Verbal Communication, J. Accous. Soc. Am., in press.
6. Vartabedian, A., The Effects of Time Delay in Four-Wire Teleconferencing, B.S.T.J., *45*, December, 1966, pp. 1673–1688.
7. Irvin, H. D., Studying Tomorrow's Communications . . . Today, Bell Laboratories Record, *36*, November, 1958, pp. 399–402.

The Effect of Intersymbol Interference on Error Rate in Binary Differentially-Coherent Phase-Shift-Keyed Systems

By W. M. HUBBARD

(Manuscript received March 7, 1967)

Two types of binary differentially-coherent phase-shift-keyed signals (designated AM-DCPSK and FM-DCPSK) which look attractive for high-speed digital communication systems are considered. The error rate as a function of signal-to-noise ratio is calculated for each type of signal. For the AM-DCPSK signal the effects of intersymbol interference from adjacent time slots, phase distortion in the pulses, nonideal delay lines in the differential phase detector and nonideal regeneration (in a sense described in the text) are considered. For the FM-DCPSK signal the effects of nonideal regeneration and of a degradation parameter δ are considered. The parameter δ can be readily associated with phase distortion and nonideal delay lines in the differential phase detector. By means of a straightforward but tedious calculation it can be related to intersymbol interference if the transfer characteristics of the channel are known. The results of the calculations are presented, in graphical form, for wide ranges of signal-to-noise ratio and of the parameters which describe the intersymbol interference and nonideal regenerator performance. Error rates from 10^{-10} to 10^{-4} are considered.

I. INTRODUCTION

This paper presents a summary of calculations which were performed before and during the construction of a 300-Mb/s repeater for a guided millimeter-wave communication system. Consequently, the problems which are considered are oriented toward problems which arise in connection with these high-speed systems, e.g., finite-width decision thresholds, imperfect phase shifts in the modulators, etc. Because of the nature of the channels envisioned for these systems, only

intersymbol interference from adjacent time slots is significant and the treatment of intersymbol interference will include only adjacent time-slot interference.

1.1 *Summary of Previous Results*

Several authors[1,2,3,4] have calculated the error rate as a function of signal-to-noise ratio, S/N, for an ideal differentially-coherent phase-shift-keyed (DCPSK) system, i.e., a system in which intersymbol interference can be ignored and in which regeneration is assumed to be ideal. The well-known result is

$$\Pi_0 = \tfrac{1}{2} \exp(-\mathrm{S/N}), \tag{1}$$

where Π_0 is the probability of error for the ideal case. The author[5] has considered the effects of nonideal regeneration. Sections II and III of this paper extend these error-rate calculations to include the effects of intersymbol interference for two types of DCPSK signals. These signals are discussed and compared in the next paragraph.

1.2 *Comparison of AM-DCPSK and FM-DCPSK Signals*

In a DCPSK system the information is carried in the phase of the signal at the sampling point in one time slot relative to the phase of the signal at a time, T, earlier where T is the reciprocal of the bit rate B. This phase change can be accomplished in several ways; the effect of intersymbol interference depends on how it is accomplished. In the following, two types of modulation which can be thought of as limiting cases (in a sense that should become clear in the following discussion) will be considered.

The first type to be considered consists of a sequence of amplitude modulated RF pulses occurring at the bit rate, B. The information is carried in whether the relative carrier phase between adjacent pulses is 0 or π. Since a phase shift of π radians is equivalent to a change in sign, the signal can be written in the form

$$S(t) = \sum_{n=0}^{\infty} a_n S_0(t - nT) \exp(j\omega_0 t), \tag{2}$$

where $a_n = +1$ or -1 according to whether the phase in the nth time slot is 0 or π, respectively. $S_0(t - nT)$ is a pulse-shaping term which reaches its maximum value at $t = nT$. Intersymbol interference arises from the fact that $S_0(t - nT)$ is not confined to a single time slot. Fig. 1(b) is an example of this type of signal.

Since the signal $S(t)$ in (2) is in fact a pure AM signal (even though

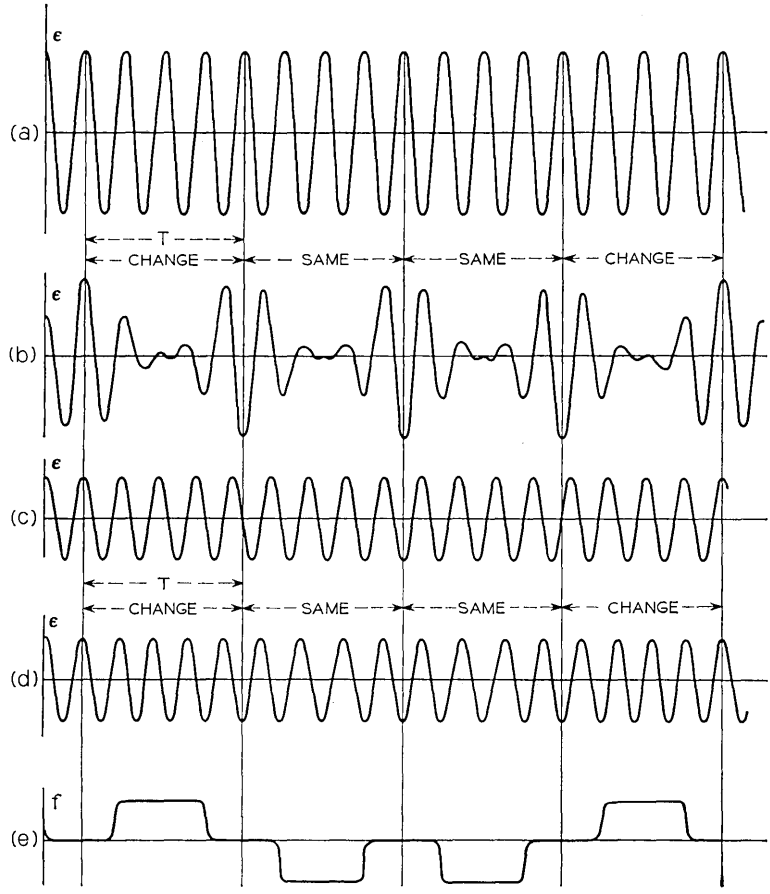


Fig. 1 — (a) Unmodulated IF-carrier for AM-DCPSK; (b) idealized AM-DCPSK signal; (c) unmodulated IF-carrier for FM-DCPSK; (d) idealized FM-DCPSK signal carrying the same information as in (b); (e) frequency vs time for the signal in (d).

the information is recovered by comparing phases) this type of modulation can be designated AM-DCPSK in order to distinguish it from a second type which will be described below. Error rate as a function of S/N for an AM-DCPSK signal with intersymbol interference and non-ideal regeneration is calculated in Section II.

The second type of modulation consists of a constant amplitude signal which shifts phase between sampling points by means of a frequency swing. This signal, which can be designated FM-DCPSK,

can be written in the form

$$S(t) = \exp j\left\{\omega_0 t + \int_0^t \omega(t')\,dt'\right\}, \quad \text{where} \quad \int_{(n-1)T}^{nT} \omega(t')\,dt' = \alpha_n \quad (3)$$

and α_n is a chance binary variable (which contains the information being transmitted) that can take on any two values which differ by π. In practice, the values $+\pi/2$ and $-\pi/2$ offer certain advantages so the following discussion will assume, for clarity, that $\alpha_n = \pm\pi/2$. This assumption is unnecessary for the calculation and does not influence the result in any way. An example of this type of signal is given in Fig. 1(d).

Error-rate vs S/N for an FM-DCPSK signal with intersymbol interference and nonideal regeneration is calculated in Section III.

1.3 *The Differential Phase Detector*

Both AM-DCPSK and FM-DCPSK signals can be detected using a product demodulator of the type shown in Fig. 2. For this device to function properly, the intermediate frequency f_0 and the bit rate, B, must be related according to

$$f_0 = \tfrac{1}{2}mB \qquad m = 1, 2, 3, \cdots \quad \text{for} \quad \text{AM-DCPSK} \quad (4)$$

and

$$f_0 = \tfrac{1}{2}(m + \tfrac{1}{2})B \qquad m = 1, 2, 3, \cdots \quad \text{for} \quad \text{FM-DCPSK.} \quad (5)$$

When these conditions are satisfied for the appropriate type of modulation, the output of the differential phase detector will be pro-

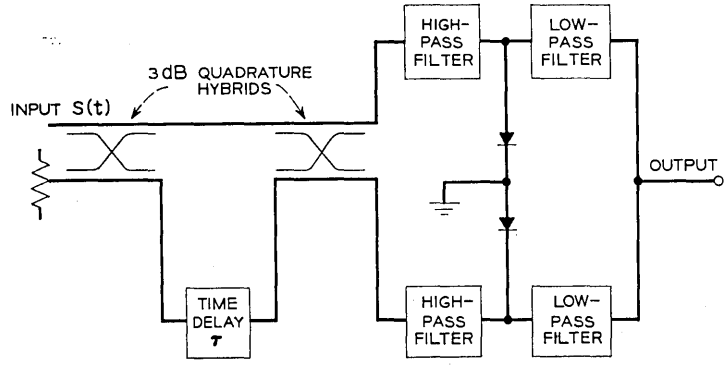


Fig. 2 — Differential phase detector.

portional to V, where

$$V = [(\mathbf{u} + \mathbf{v})^2 - (\mathbf{u} - \mathbf{v})^2] = uv \cos \psi. \tag{6}$$

Here $\mathbf{u}$ is a vector which represents the amplitude and phase of the signal at time t, $\mathbf{v}$ the amplitude and phase at time $t-T$, and ψ is the phase difference between times t and $t-T$.

1.4 *The Regenerator*

The output of the differential phase detector is fed into the regenerator where it is sampled at a particular point in each time slot. An ideal regenerator would regenerate a $+1$ if $\cos \psi > 0$ and a -1 if $\cos \psi < 0$ regardless of how small $|\cos \psi|$ might be. Since no realizable regenerator will accomplish this, we take as a model of a regenerator a device which regenerates $+1$'s and -1's according to the following inequalities:

$$\begin{aligned} &+1 \quad \text{if} \quad 1 \geqq \cos \psi \geqq \epsilon \\ &-1 \quad \text{if} \quad -\epsilon \geqq \cos \psi \geqq -1 \end{aligned} \tag{7}$$

$+1$ or -1 randomly and with equal probability if $|\cos \psi| < \epsilon$.

It is convenient to define a threshold T in terms of the parameter ϵ by means of the relation

$$\mathrm{S/T} = -20 \log \varepsilon \; dB.$$

The quantity S/T is the so-called signal-to-threshold ratio and represents the ratio of the expected value of signal power to the minimum value of signal power which will cause the regenerator to function reliably (in the absence of noise).

In high-speed systems the signal-to-threshold ratio is limited by practical considerations (at the present state of the art) to values of the order of 10 dB or less.

II. ERROR-RATE WITH AM-DCPSK MODULATION

2.1 *Intersymbol Interference*

The voltages at the two output ports of the second quadrature hybrid in Fig. 2 will at any instant consist of contributions from the following sources:

(*i*) The two pulses being compared.
(*ii*) Intersymbol interference from other pulses in the channel.

(*iii*) Interchannel interference from other channels propagating in the medium.
(*iv*) Noise.

The following assumptions are made:

(*i*) The noise is Gaussian and the noise on adjacent pulses is statistically independent.*
(*ii*) Interchannel interference is negligible.
(*iii*) Intersymbol interference comes only from adjacent pulses.
(*iv*) The sampling is accomplished instantaneously.

Since the phases of these output voltages are, in general, different we must represent these voltages by vectors (in a plane). Consider four adjacent pulses labeled LABR from left to right. The pulses labeled A and B are the two whose phases are to be compared. The ones labeled L and R are significant because they contribute to the intersymbol interference. Let **L**, **A**, **B**, and **R** be vectors of unit magnitude which lie along the $+X$ or $-X$ direction. Let ρ_a represent the ratio of $S_0(T)$ to $S_0(0)$ and ρ_r the ratio of $S_0(-T)$ to $S_0(0)$ where $S_0(t)$ is the quantity introduced in (2). Let **a** and **b** represent the noise on pulses A and B, respectively.

The outputs $\mathbf{r}_2$ and $\mathbf{r}_1$ of the two output ports of the second quadrature hybrid are then

$$\mathbf{r}_2 = \mathbf{S} + \mathbf{a} + \mathbf{b} = \mathbf{u} + \mathbf{v} \tag{8}$$

$$\mathbf{r}_1 = \mathbf{D} + \mathbf{a} - \mathbf{b} = \mathbf{u} - \mathbf{v}, \tag{9}$$

respectively, where

$$\mathbf{S} = \mathbf{A} + \mathbf{B} + \rho_a(\mathbf{B} + \mathbf{R}) + \rho_r(\mathbf{L} + \mathbf{A}) \tag{10}$$

$$\mathbf{D} = \mathbf{A} - \mathbf{B} + \rho_a(\mathbf{B} - \mathbf{R}) + \rho_r(\mathbf{L} - \mathbf{A}). \tag{11}$$

For a given pulse pattern **S** and **D** are determined. The quantities **a** and **b** represent four independent Gaussian variables with zero means and equal variances.

Consider first the means of the distributions of $\mathbf{r}_2$ and $\mathbf{r}_1$. Since **a** and **b** are Gaussian variables of zero mean these means are determined for a given pulse pattern by **S** and **D**, respectively.

There are $2^4 = 16$ possible patterns for the four pulses, L, A, B, and

* The latter assumption is never strictly true since the noise is band-limited. It has been shown,[6] however, that the effects of this correlation on error rate are negligible unless the noise bandwidth is smaller than about 1.4 times the bit-rate.

TABLE I

Case	L	A	B	R	$\frac{1}{2}S$	$\frac{1}{2}D$	$\frac{1}{4}(S^2 - D^2) = \frac{1}{4}PP_d$ (neglecting terms in ρ^2)
1	1	1	1	1	$1 + \rho_a + \rho_r$	0	$1 + 2\rho_a + 2\rho_r$
2	1	1	1	−1	$1 + \rho_r$	ρ_a	$1 + 2\rho_r$
3	−1	1	1	1	$1 + \rho_a$	$-\rho_r$	$1 + 2\rho_a$
4	−1	1	1	−1	1	$\rho_a - \rho_r$	1
5	1	1	−1	1	ρ_r	$1 - \rho_a$	$-1 + 2\rho_a$
6	1	1	−1	−1	$-\rho_a + \rho_r$	1	−1
7	−1	1	−1	1	0	$1 - \rho_a - \rho_r$	$-1 + 2\rho_a + 2\rho_r$
8	−1	1	−1	−1	$-\rho_a$	$1 - \rho_r$	$-1 + 2\rho_r$

R. Of these, eight can be obtained by reversing *all* of the phases in each of the other eight patterns. Since such a sign reversal has no effect on the error probabilities to be considered only eight patterns need be considered. These are numerated in Table I.

We now consider the output of the differential phase detector for an arbitrarily chosen pulse pattern. The criterion for making the decision as to whether the pulses are of the same phase or of opposite phase is that of determining whether

$$V = \{|\mathbf{r}_2|^2 - |\mathbf{r}_1|^2\} \tag{12}$$

is positive or negative. This criterion is equivalent to deciding whether the phase angle between the received pulses (including crosstalk and noise) is less than 90 degrees or greater than 90 degrees, respectively. It is worth mentioning that due to the correlation between the signals in the sum and difference arms 0 would in general not be the proper decision level if the phases of the pulse tails relative to the pulse peaks were fixed and known. Since this is probably not going to be the case in any reasonable system, the decision level will be taken at 0 in this calculation. (Zero is the optimum decision level for random phase in the tails.)

With the assumptions that have been made, the error probability including the effects of intersymbol interference can be determined by a straightforward extension of the calculation due to Bennett and Salz.[3] Substituting (8) and (9) into (12) gives

$$V = (P + x)(P_d + x_d) + yy_d\ , \tag{13}$$

where

$$P = S + D \qquad P_d = S - D$$
$$x = 2a_x \qquad x_d = 2b_x$$
$$y = 2a_y \qquad y_d = 2b_y\ .$$

Equation (13) is identical in form to (54) of Ref. 3. However, the symbols now include the effects of intersymbol interference. Following the method of Bennett and Salz[3] one obtains for the probability of error

$$\Pi = \frac{1}{\pi}\int_0^{\pi/2} \exp\left(-\frac{P^2P_d^2}{8\sigma^2(P_d^2\cos^2\theta + P^2\sin^2\theta)}\right) d\theta. \tag{14}$$

It must be recalled that P and P_d are pulse-pattern dependent. For a random message the error rate is obtained by averaging the expression in (14) over the eight possible values of P and P_d.

In order to determine the worst possible error rate for any arbitrary message, we can evaluate Π for the worst pulse pattern. From Table I it is apparent that this is Case 7 if there is no phase distortion in the pulse tails. In any event this case represents an upper limit on the error rate. Fortuitously, since $S = 0$ in this case, P^2 is equal to P_d^2 and the integral for Π becomes particularly simple.

$$\Pi = \frac{1}{\pi}\int_0^{\pi/2} \exp\left(-\frac{P^2}{8\sigma^2}\right) d\theta = \frac{1}{2}\exp\left[-\frac{(1-\rho_a-\rho_r)^2}{2\sigma^2}\right]. \tag{15}$$

Thus, the effects of intersymbol interference can be treated, for the worst pulse pattern at least, as a degradation of the signal-to-noise ratio by the amount

$$20\log(1-\rho_a-\rho_r)\ \text{dB}.$$

The extension of this calculation to the case where both intersymbol interference and a finite-width decision threshold[5] are present is straightforward. One replaces the integral in (90) of Ref. 3:

$$\int_0^\infty dz\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(-z \mid x, y)g(x, y)\,dx\,dy,$$

by the sum of two integrals:

$$\frac{1}{2}\int_{-4\epsilon}^{4\epsilon} dz\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(-z \mid x, y)q(x, y)\,dx\,dy + \int_{4\epsilon}^{\infty} dz\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(-z \mid x, y)g(x, y)\,dx\,dy.$$

These integrals are then evaluated by the method used in Appendix A of Ref. 5. Figs. 3, 4, and 5 show error rate as a function of signal-to-noise ratio in the case where ρ_a equals ρ_r in terms of signal-to-threshold ratio and crosstalk per tail. In these figures $\rho_T = \rho_a = \rho_r$ is expressed in dB.

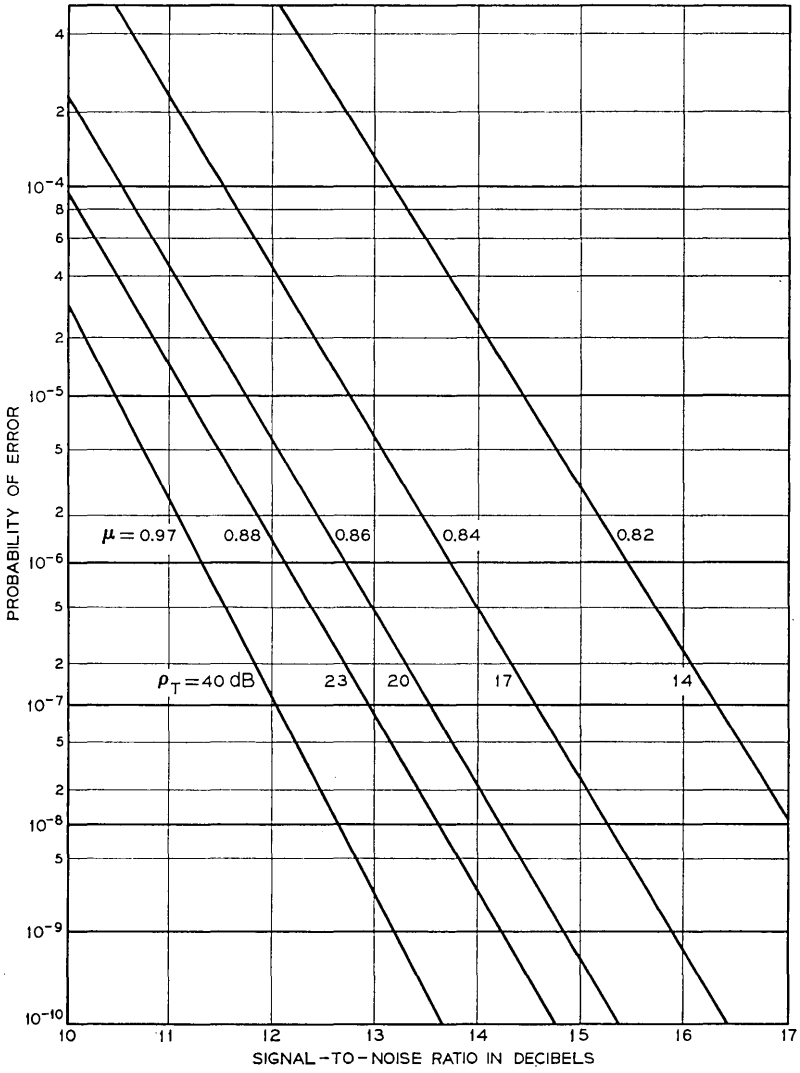


Fig. 3 — Probability of error vs signal-to-noise ratio for S/T = 15 dB AM-DCPSK.

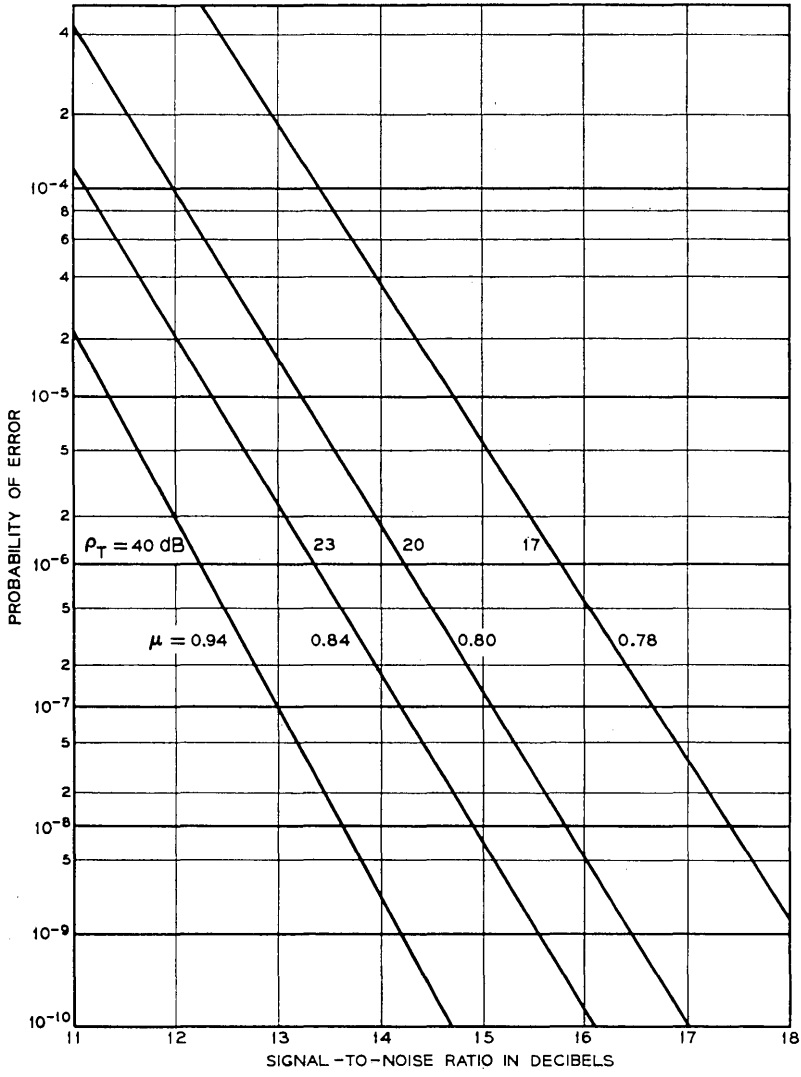


Fig. 4 — Probability of error vs signal-to-noise ratio for S/T = 9 dB AM-DCPSK.

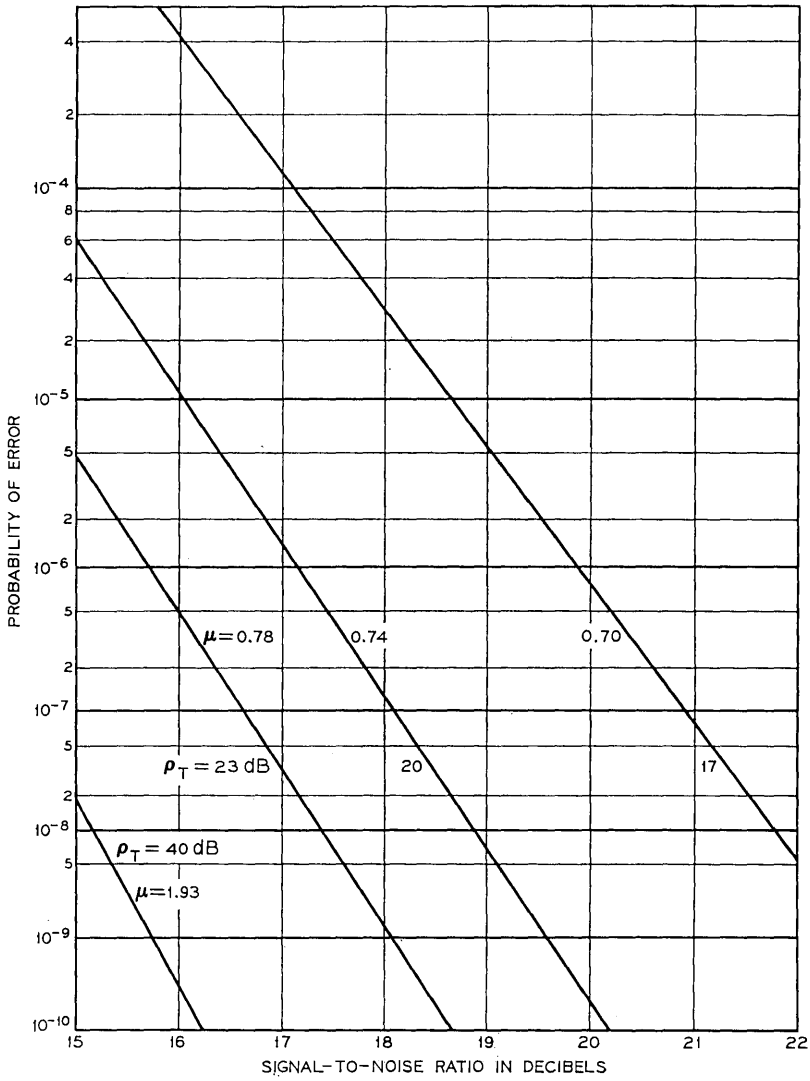


Fig. 5 — Probability of error vs signal-to-noise ratio for S/T = 6 dB AM-DCPSK.

The scale of the ordinate in Figs. 3, 4, and 5 is chosen such that a plot of Π_0 vs S/N from (1) gives a straight line. One finds from inspection of Figs. 3, 4, and 5 that even in the presence of intersymbol interference and nonideal regeneration, this linearity persists. The slope of the line does change, however. Fig. 6 shows the slope, μ, of these lines as a function of S/T for various values of ρ_T. Here the slope, μ, is defined such that $\mu = 1$ for the ideal case [see (1)].

2.2 *Phase Distortion*

In an AM-DCPSK system there are two important types of phase distortion which are readily treated by a modification of the foregoing calculation. One type is representable by a phase shift of the pulse in the nth time slot by β degrees relative to the phase which should have been transmitted in that time slot. In this case the vectors $\mathbf{B}$, and $\rho_a\mathbf{B}$, in (10) and (11) are rotated an amount β. This could arise, for example, from an improperly balanced pulse modulator. In the other type of phase distortion all of the delayed pulses in the differential phase detector are shifted an amount φ relative to their proper value. In this case the vectors $\mathbf{B}$, $\rho_r\mathbf{A}$, and $\rho_a\mathbf{R}$ are rotated an amount φ. This could arise, for example, from a delay line of improper length in the differential phase detector.

The analysis of these effects constitutes a straightforward extention of the calculation in paragraph 2.1 and only the results will be given here. Fig. 7 shows the degradation in error-rate performance which results from a phase shift β for $\rho_T = -\infty$, -26, and -20. This effect is virtually independent of S/T for S/T $>$ 6 dB. Fig. 8 shows the

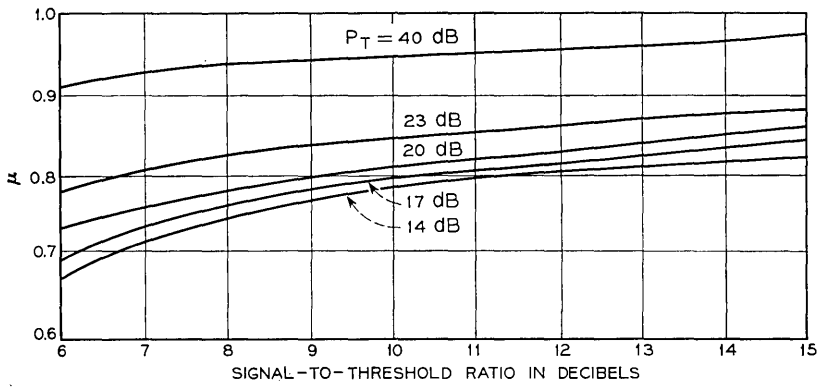


Fig. 6 — Slope, μ, of error probability vs S/N curve as a function of S/T.

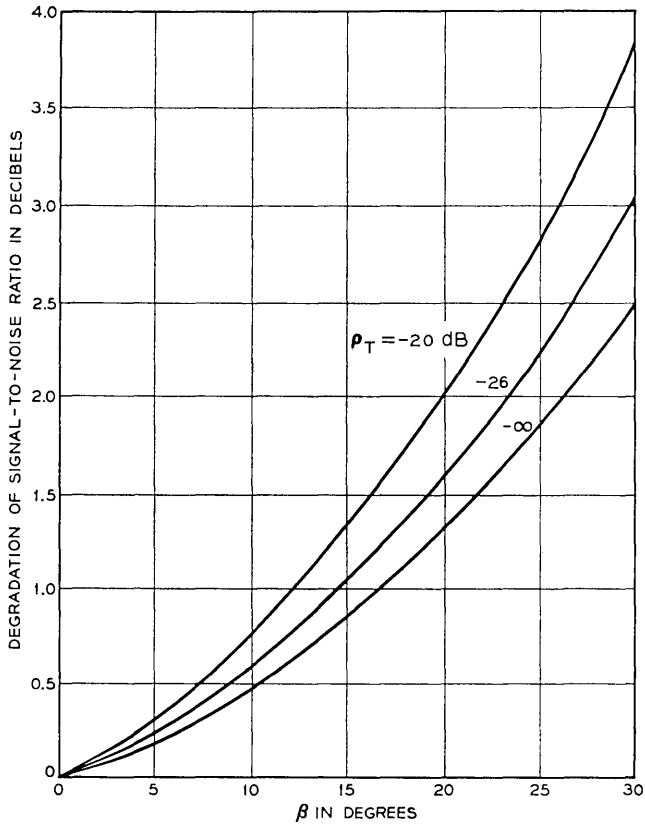


Fig. 7 — Degradation in signal-to-noise ratio due to a phase shift, β, in one of the received pulses.

degradation for phase shift φ. The result is virtually independent of S/T for S/T $>$ 6 dB and of ρ_T for $\rho_T < -20$ dB.

III. ERROR-RATE IN AN FM-DCPSK SYSTEM

In an FM-DCPSK system, it is useful to include a limiter in the receiver after the noise has been added. Therefore, the following calculation assumes that an ideal limiter is used. By ideal limiter is meant a device which receives at its input the signal

$$A(t) \exp [j\varphi(t)]$$

and at its output delivers the signal

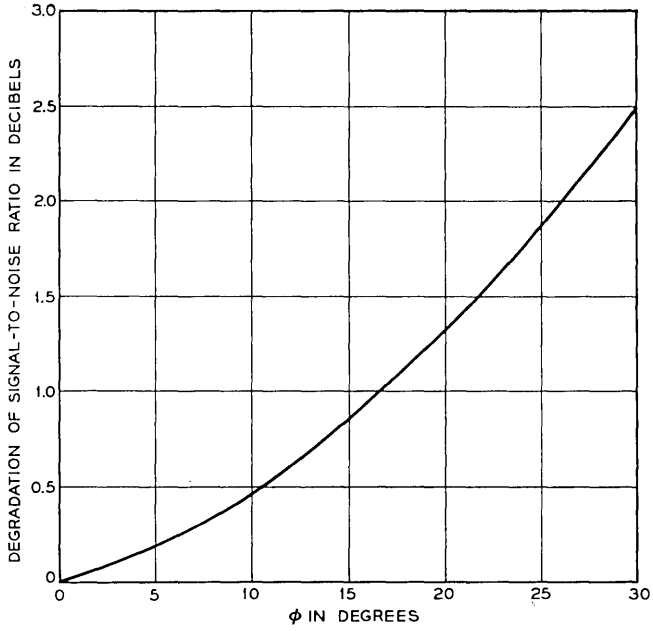


Fig. 8 — Degradation in signal-to-noise ratio due to improper delay line length.

$$A_0 \exp [j\varphi(t)].$$

Since the amplitudes u and v in (6) are constants after limiting, the output of the differential phase detector is simply

$$V = \cos \psi$$

and the error probability is directly related to the probability density of ψ.

The intersymbol interference is assumed to manifest itself in the form of a perturbation in the phase of the signal at the sampling time. More precisely, the intersymbol interference (in this model) introduces a (pattern dependent) phase shift δ_n so that the phase change in the nth time slot is $\pm(\pi/2 + \delta_n)$ instead of $\pm\pi/2$.

The value of δ_n (for each distinct pulse pattern) depends on the details of the signal waveform and the transfer function of the devices in the system. The determination of the δ_n's for particular systems is beyond the scope of this paper. For a discussion of this problem see, for example, Rice and Bedrosian.[7] This paper concerns itself with the

effect on error rate of a particular value of δ_n. In order to apply these results to the performance of a particular system, one needs to compute the values of δ_n for the various pulse patterns and then average these computed error rates over the possible pulse patterns.

The parameter δ can also be used to investigate the effects of phase distortion and nonideal delay lines in the differential phase detector just as β and φ were used in the AM-DCPSK case. One need only associate δ with the total (net) phase shift for these distortions (including that due to intersymbol interference).

Let α_0 and β_0 represent the phase shift due to intersymbol interference (and any other degradation in phase) on the two pulses being compared. This situation is represented by the phasor diagram in Fig. 9. Following the method described in Ref. 5, one obtains for the probability density function of ψ

$$p(\psi) = \frac{1}{2\pi}\exp(-1/\sigma^2) + \frac{1}{\pi}\exp(-1/2\sigma^2)$$
$$- \frac{1}{4\pi\sigma^2}\int_{-\pi/2}^{\pi/2}\cos(\alpha-\alpha_0)\cos(\alpha+\psi-\beta_0)$$
$$\cdot\exp\left[-\frac{\sin^2(\alpha-\alpha_0)+\sin^2(\alpha+\psi-\beta_0)}{2\sigma^2}\right]d\alpha$$
$$+ \frac{1}{4\pi\sigma^2}\int_{-\pi/2}^{\pi/2}\cos(\alpha-\alpha_0)\cos(\alpha+\psi-\beta_0)$$
$$\cdot\exp\left[-\frac{\sin^2(\alpha-\alpha_0)+\sin^2(\alpha+\psi-\beta_0)}{2\sigma^2}\right]$$
$$\cdot\operatorname{erf}\frac{\cos(\alpha-\alpha_0)}{\sqrt{2}\sigma}\operatorname{erf}\frac{\cos(\alpha+\psi-\beta_0)}{\sqrt{2}\sigma}\,d\alpha. \qquad (16)$$

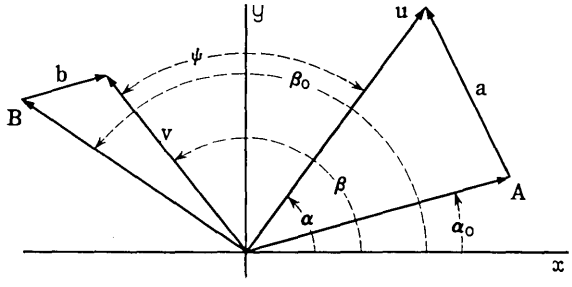


Fig. 9 — Phasor diagram for the differential detection process.

From the regenerator model (Paragraph 1.4) one observes that an error is made if $\cos \psi \geqq \epsilon$ and there is a 50-percent probability of error if $| \cos \psi | < \epsilon$. The error probability, Π, is therefore given for a transmitted signal which should result in $\psi = \pi$ by

$$\Pi = \int_{-\theta}^{\theta} p(\psi)\, d\psi + \frac{1}{2}\int_{\theta}^{\pi-\theta} p(\psi)\, d\psi + \frac{1}{2}\int_{\pi+\theta}^{-\theta} p(\psi)\, d\psi, \qquad (17)$$

where

$$\theta = \cos^{-1} \epsilon \qquad 0 \leqq \theta \leqq \pi/2.$$

Substituting (16) into (17) gives, after some simplification,

$$\Pi = \frac{1}{2} - \frac{1}{4\pi\sigma^2}\int_{-1}^{1}\int_{y_l(x)}^{y_u(x)} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) dy\, dx, \qquad (18)$$

where

$$y_u(x) = \sqrt{1-x^2}\cos(\varphi - \delta) + x\sin(\varphi - \delta)$$

$$y_l(x) = -\sqrt{1-x^2}\cos(\varphi + \delta) + x\sin(\varphi + \delta)$$

$$\delta = \alpha_0 - \beta_0\,, \qquad \varphi = \frac{\pi}{2} - \theta = \sin^{-1}\epsilon.$$

Now $y_u(x)$ and $y_l(x)$ are segments of (different) ellipses both of which have the following properties: They are centered at the origin, have their major axes along the line $x = y$, and are tangent to the lines $x = 1$, $x = -1$, $y = 1$, and $y = -1$. A typical pair of such ellipses is shown in Fig. 10. By symmetry the small area A in Fig. 10(a)

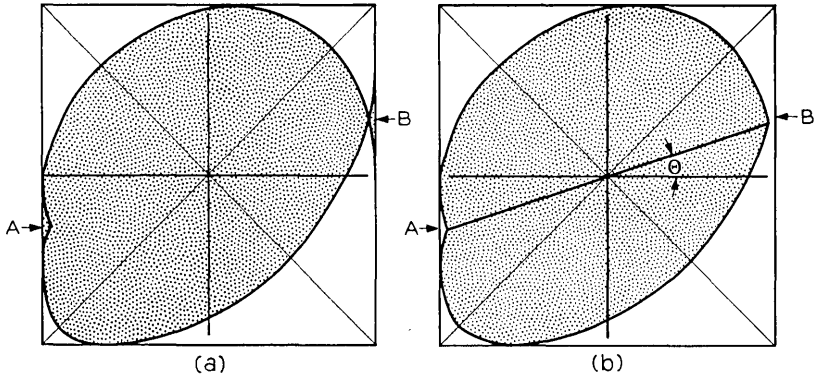


Fig. 10 — Region of integration in (16).

is equal to the small area B. Therefore, due to the spherical symmetry of the integrand, the integration indicated by the limits $-1 \leqq x \leqq 1$, $y_l(x) \leqq y \leqq y_u(x)$ [as shown by the shaded portion of Fig. 10(a)] can be replaced by the limits which correspond to integrating over the region bounded by the upper ellipse and the line $\theta = \Theta$ and the region bounded by the lower ellipse and this same line [see Fig. 10(b)]. Writing the integral in polar coordinates, one obtains a form which is easily integrated over r. When this is done, the following result is obtained:

$$\Pi = \frac{1}{4\pi}\left\{\int_{\Theta}^{\Theta+\pi} \exp\left\{-\frac{\cos^2(\varphi-\delta)}{2\sigma^2[1-\sin(\varphi-\delta)\sin 2\theta]}\right\} d\theta + \int_{\Theta+\pi}^{\Theta} \exp\left\{-\frac{\cos^2(\varphi+\delta)}{2\sigma^2[1-\sin(\varphi+\delta)\sin 2\theta]}\right\} d\theta\right\}. \tag{19}$$

But these integrals are periodic in period π, therefore, the Θ's can be deleted. The error rate can then be written

$$\Pi = \tfrac{1}{2}P_0(\varphi+\delta) + \tfrac{1}{2}P_0(\varphi-\delta), \tag{20}$$

where

$$P_0(\Phi) = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \exp\left(-\frac{\cos^2\Phi}{2\sigma^2[1-\sin\Phi\sin\theta]}\right) d\theta. \tag{21}$$

This integral is not soluble in closed form. The integral

$$P(\Phi) = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \exp\left(-\frac{\cos^2\Phi}{2\sigma^2}[1+\sin\Phi\sin\theta]\right) d\theta$$

is soluble and is a good approximation to $P_0(\Phi)$ over a wide range of values of σ and Φ. It can be written

$$P(\Phi) = \frac{1}{2}\exp\left(-\frac{\cos^2\Phi}{2\sigma^2}\right) I_0\left(\frac{\cos^2\Phi\sin\Phi}{2\sigma^2}\right)$$

where I_0 is the modified Bessel function of the first kind. A complete consideration of the accuracy of this approximation is quite tedious and will not be considered further because $P_0(\varphi)$ itself is so readily obtained by numerical integration of (21).

Fig. 11 shows $P(\varphi)$ for several values of S/N. Figs. 12, 13, 14, and 15 show Π(S/N) for S/T $= \infty$, 12, 9, 6 dB, respectively, for $\delta = 0$, 5, 10, and 15 degrees.

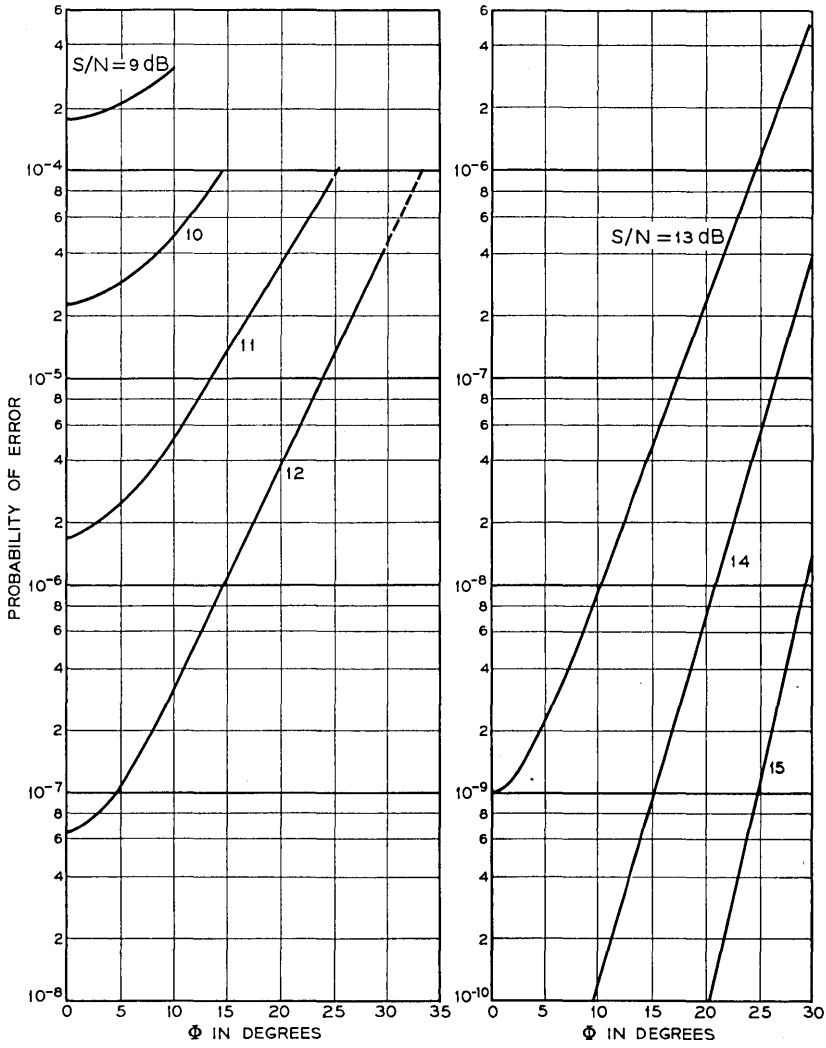


Fig. 11 — Numerical evaluation of the error-rate integral.

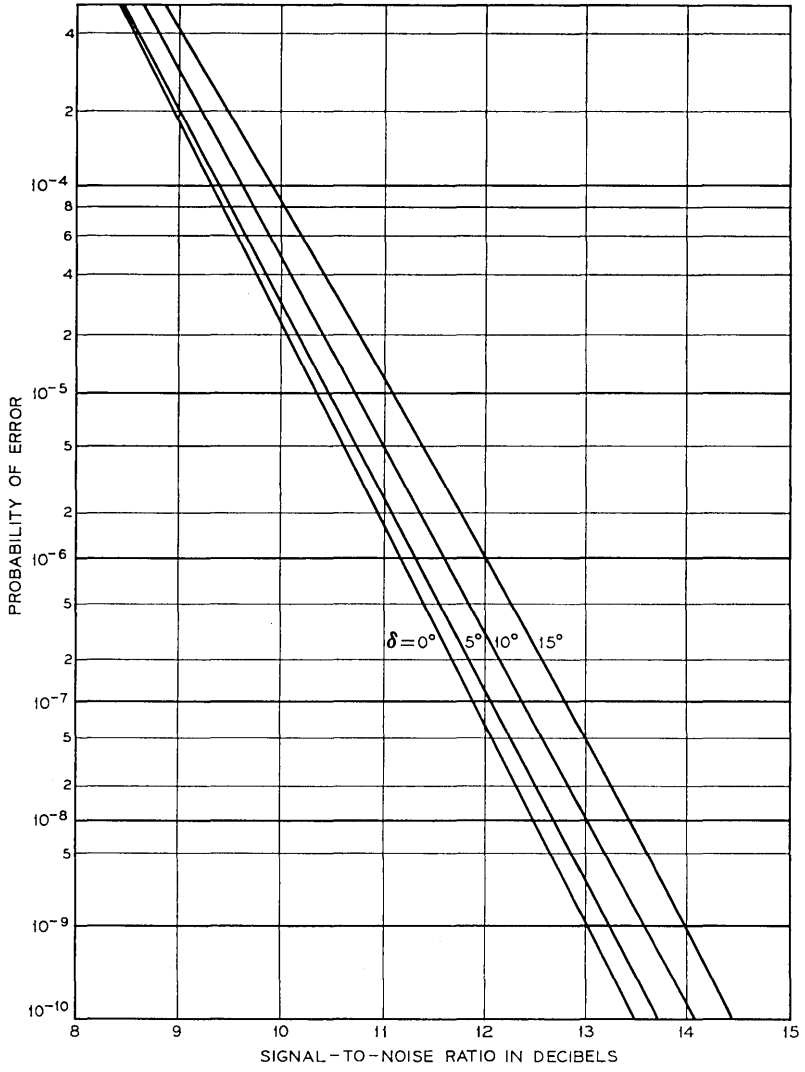


Fig. 12 — Probability of error vs signal-to-noise ratio for S/T = ∞ FM-DCPSK.

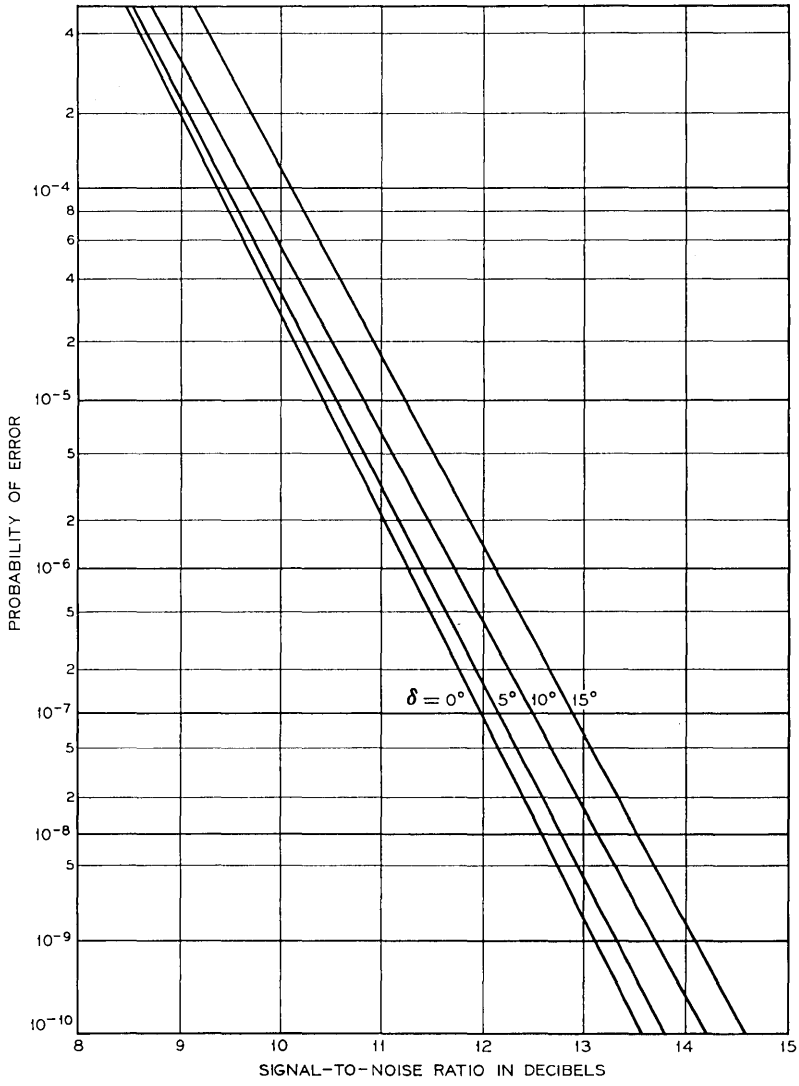


Fig. 13 — Probability of error vs signal-to-noise ratio for S/T = 12 dB FM-DCPSK.

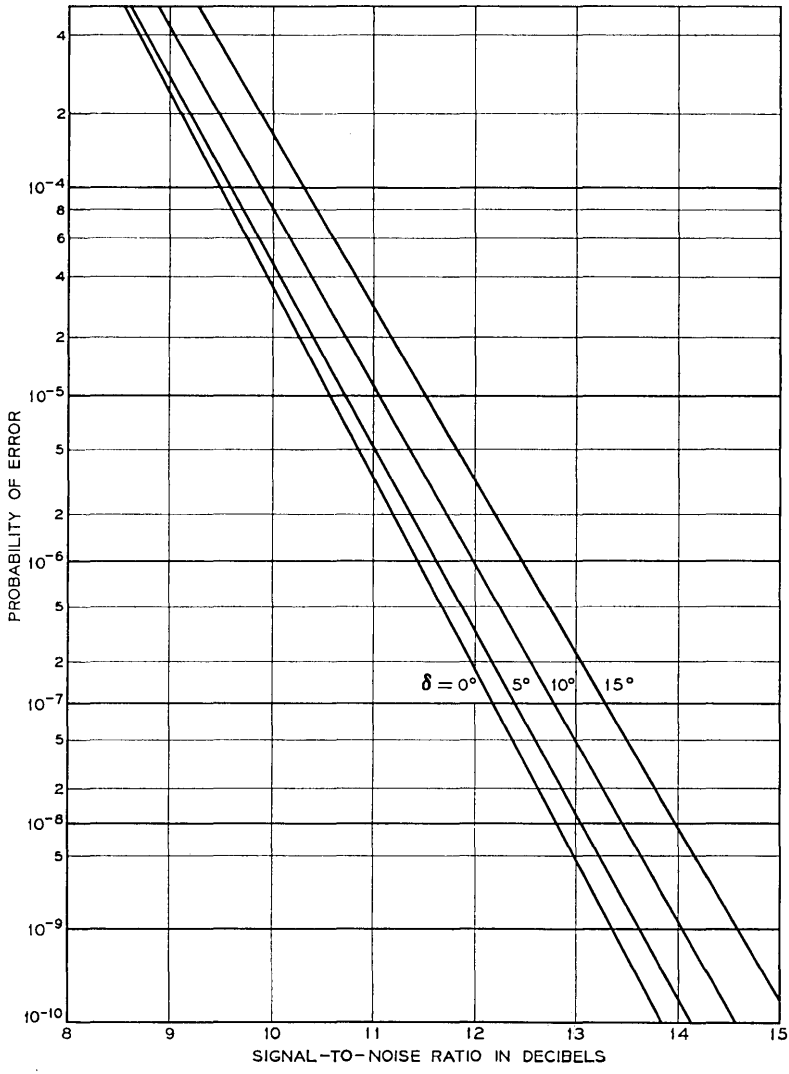


Fig. 14 — Probability of error vs signal-to-noise ratio for S/T = 9 dB FM-DCPSK.

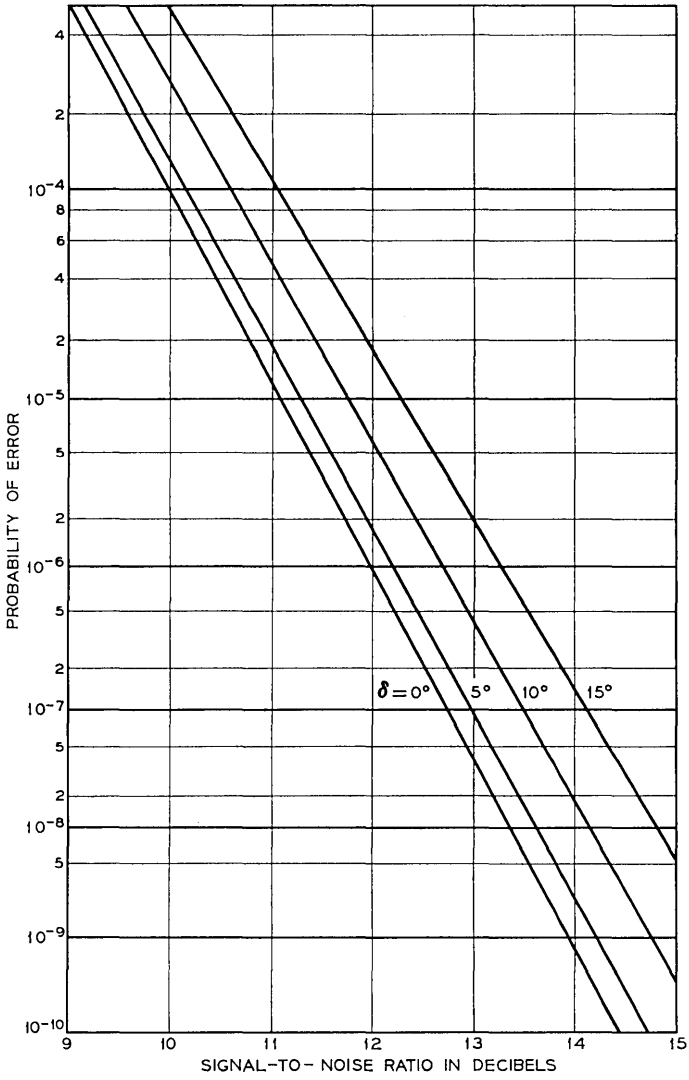


Fig. 15 — Probability of error vs signal-to-noise ratio for S/T = 6 dB FM-DCPSK.

V. ACKNOWLEDGMENTS

The author is grateful to Mrs. C. L. Beattie and Mrs. E. Kerschbaumer for programming the numerical calculations and to Mr. W. D. Warters for numerous enlightening conversations during the course of the calculations.

APPENDIX

Description of the Operation of the Differential Phase Detector for an FM-DCPSK Signal

The differential phase detector is shown in Fig. 2. For an FM-DCPSK signal with no noise or distortion the input is given by (3). One can readily show that the signals in the output ports of the second 3-dB quadrature hybrid are given by

$$A(t) = \tfrac{1}{2}\cos\left\{\omega_0 t + \int_{-\infty}^{t}\omega(t')\,dt'\right\} - \tfrac{1}{2}\cos\left\{\omega_0(t-\tau) + \int_{-\infty}^{t-\tau}\omega(t')\,dt'\right\}$$

$$B(t) = \tfrac{1}{2}\sin\left\{\omega_0 t + \int_{-\infty}^{t}\omega(t')\,dt'\right\} - \tfrac{1}{2}\sin\left\{\omega_0(t-\tau) + \int_{-\infty}^{t-\tau}\omega(t')\,dt'\right\}.$$

If the detectors, mounted as shown in Fig. 2, are regarded as having square-law behavior, the output is given by

$$\begin{aligned} V(t) &\propto B^2(t) - A^2(t) \\ &\propto \tfrac{1}{2}\cos\left\{\omega_0\tau + \int_{t-\tau}^{t}\omega(t')\,dt'\right\} + \text{terms in } 2\omega_0 t. \end{aligned}$$

The terms in $2\omega_0 t$ are removed by the low-pass filters. We can thus write the basic equation of the differential phase detector for a FM-DCPSK signal as

$$V(t) = \cos\left\{\omega_0\tau + \int_{t-\tau}^{t}\omega(t')\,dt'\right\}. \tag{22}$$

For use as a differential phase detector one chooses $\tau = T =$ the reciprocal of the bit rate, and an IF such that $\omega_0 T = (m+\frac{1}{2})\pi$ where m can be any integer. This equation then becomes

$$V(t) = \sin\left\{\int_{t-T}^{t}\omega(t')\,dt'\right\},$$

which by (3) must be

$$V(t) = \sin\alpha_n = \pm 1$$

at the sampling points $t = nT$.

REFERENCES

1. Lawton, J. G., Comparison of Binary Data Transmission, Proc. 1958 Conf. Mil. Electron.
2. Cahn, Charles, R., Performance of Digital Phase Modulation Communication Systems, IRE Trans. *CS,* May 1959, pp. 3–6.
3. Bennett, W. R. and Salz, J., Binary Data Transmission by FM Over a Real Channel, B.S.T.J., *42,* September, 1963, pp. 2387–2426.
4. Bussgang, J. J. and Leiter, M., Error-Rate Approximation for Differential Phase Shift Keying, IEEE Trans. *CS-12.* March, 1964. pp. 18-27.
5. Hubbard, W. M., The Effect of a Finite-Width Decision Threshold on Binary Differentially Coherent PSK Systems, B.S.T.J., *45,* February, 1966, pp. 306–320.
6. Hubbard, W. M., Effect of Noise Correlation on Binary Differentially Coherent PSK Communication Systems, B.S.T.J., *46,* January, 1967, pp. 277–280.
7. Bedrosian, E. and Rice, S. O., Distortion and Crosstalk of Linearly Filtered and Angle-Modulated Signals, unpublished work.

Experimental Verification of the Error-Rate Performance of Two Types of Regenerative Repeaters for Differentially Coherent Phase-Shift-Keyed Signals

By W. M. HUBBARD and G. D. MANDEVILLE

(Manuscript received February 9, 1967)

High-speed digital repeaters are being considered in the Bell System and elsewhere for both long- and short-haul communication systems. This paper describes two devices which were built to serve as prototypes of the IF portion of a millimeter guided-wave communication system but which might serve equally well as the IF portion of the repeaters for an optical communication system or for microwave radio systems. These two prototypes, designed for a differentially coherent phase-shift-keyed (DCPSK) signal, have been built and operated at a bit rate of 160 Mb/s using an 11.2-GHz IF signal. Both models operated with error rates very close to those predicted theoretically. One of the models seems to be particularly suitable for such systems; it consists of comparatively simple circuitry, and its operation is within 0.5 dB of the theoretical ideal behavior for a DCPSK system.

I. INTRODUCTION

High-speed long-haul communication by means of millimeter waves transmitted in the circular electric mode through a multimode circular waveguide was described by S. E. Miller[1] in 1954, and subsequently considered in some detail by Rowe and Warters.[2] This paper describes two experimental models of the IF portion of a repeater for such a system and compares their performance with theoretical predictions of error rate. These models could serve equally well as IF sections of repeaters for optical communication systems or microwave radio systems.

No attempt is made in these models to equalize the delay distortion of the medium, and no allowance is made for degradation from up converters, down converters, and millimeter-wave circuitry.

Both experimental model repeaters operated at a bit rate of 160 Mb/s and used an IF of 11.2 GHz. The bit rate and the IF were arbitrary choices made for convenience. In an actual system, a somewhat lower IF would be chosen to facilitate building solid-state amplifiers and a somewhat higher bit rate would probably be desirable. Since these model repeaters were to serve as prototypes for an even higher bit rate repeater, no components were used which did not seem capable of being modified to operate up to about twice this bit rate.

Section II discusses briefly the two particular choices of modulation scheme which were used, and describes some of the features common to both models. Section III describes the so-called AM-DCPSK repeater and Section IV the FM-DCPSK one.

These models demonstrated that a repeater which performs with error rates quite close to those predicted theoretically can be built.

II. DIFFERENCES AND SIMILARITIES OF THE TWO MODELS

The final version of a millimeter-wave repeater would almost certainly be an all solid-state system. This requirement imposes a limitation on the maximum power available, especially at millimeter-wave frequencies. Because of this power limitation, which generally manifests itself as a limitation on peak power, it is necessary to use a modulation scheme which affords good noise immunity. Optimum noise immunity (in a binary system) is obtained when the two signal states are anticorrelated.[3] This corresponds to coherent phase modulation where the two signals have identical envelopes (consistent with the power limitations) and differ in phase by π radians. Such a system operates by sending pulses with phase either 0 or π relative to some standard reference phase. With a system of this type, it is necessary to provide this standard phase at each repeater in order that regeneration may be performed. An alternative approach is differentially-coherent phase-shift-keying (DCPSK). In DCPSK, the phase of each pulse is used as the reference for determining the phase of the next following pulse. The information is, therefore, carried in the relative phase or, equivalently, it is carried in whether or not the phase changes between pulses. The penalty in noise immunity for using a DSPSK system instead of a coherent PSK system amounts to about 2 dB for error-rates of about 0.01 but decreases with increasing S/N. For signal-to-noise ratios which give error-rates of the order of 10^{-9} (the assumed acceptable error-rate for this experiment) the degradation in noise immunity is less than 0.5 dB. (See, for example,

Lawton.[4]) Since the system under consideration is to be operated with error rates of this order, the only form of modulation to be considered in this paper is DCPSK.

A block diagram of a typical repeater is shown in Fig. 1. The purpose of this figure is to show the overall layout of the repeater and to indicate where the IF portion of the repeater fits in. Figs. 2 and 3 are block diagrams of the two experimental model repeaters which this paper describes, along with the test equipment used in the experiments. The portions of these figures contained within the dotted lines are more detailed versions of the single block labeled "IF portion" in Fig. 1.

The two model repeaters used somewhat different modulation schemes to achieve the binary DCPSK. In binary DCPSK, the information is carried in the phase *change* of a signal between two sampling points. DCPSK signals can have several forms since the only requirement is that the phase must make either of two prescribed changes between each pair of adjacent sampling points. Two somewhat idealized classes of signals are considered below. These two classes represent limiting cases since any physically realizable signal would contain some of the properties of each. We designate these classes as AM-DCPSK and FM-DCPSK. The AM-DCPSK signals are created (at least conceptually) by generating a separate pulse for each time slot with one or the other of two phases. Each pulse is thought of as being generated independently of pulses in other time slots, but is not necessarily confined to a single time slot, i.e., intersymbol interference is allowed. As an example, such a signal might have the form

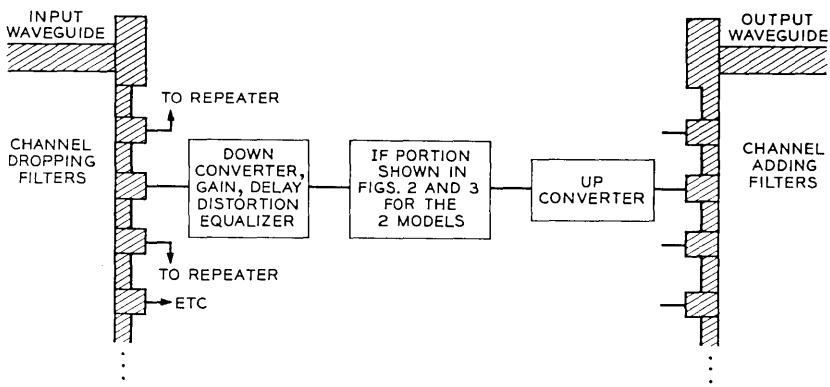


Fig. 1 — Block diagram of a complete repeater.

$$S(t) = \sum_{n=0}^{N} S_0(t - nT) \exp [j(\omega t + \alpha_n)], \tag{1}$$

where $S_0(t)$ is the pulse-shaping factor which is slowly varying compared with the exponential factor, and α_n is a chance variable which can take with equal probability either of two values which differ by π. The information is carried in whether $|\alpha_{n-1} - \alpha_n|$ equals 0 or π. In such a system, the amplitude is varied, but the carrier frequency of the individual pulses, $\omega/2\pi$, is a constant. If the pulses were nearly resolved, i.e., if

$$|S_0(t)| \ll S_0(0) \qquad \text{for} \qquad |t| \geqq T/2 \tag{2}$$

$S(t)$ would look like a pure AM signal as illustrated in Fig. 4(b).

The FM-DCPSK class is one in which the amplitude of the signal remains constant and the phase change (if any) between adjacent sampling times is effected by changing the carrier frequency. As an example, such a signal might have the form

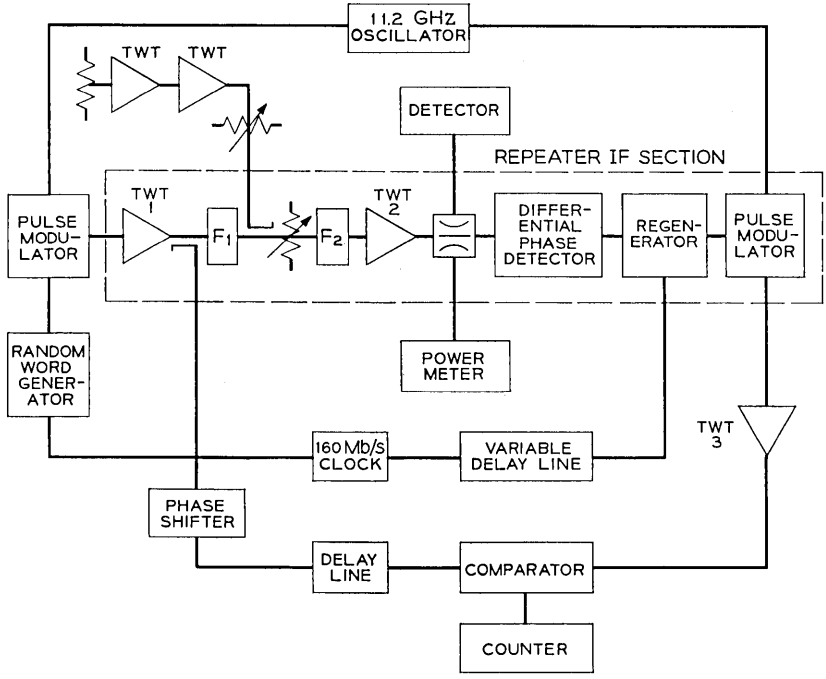


Fig. 2 — Block diagram of the AM-DCPSK model repeater.

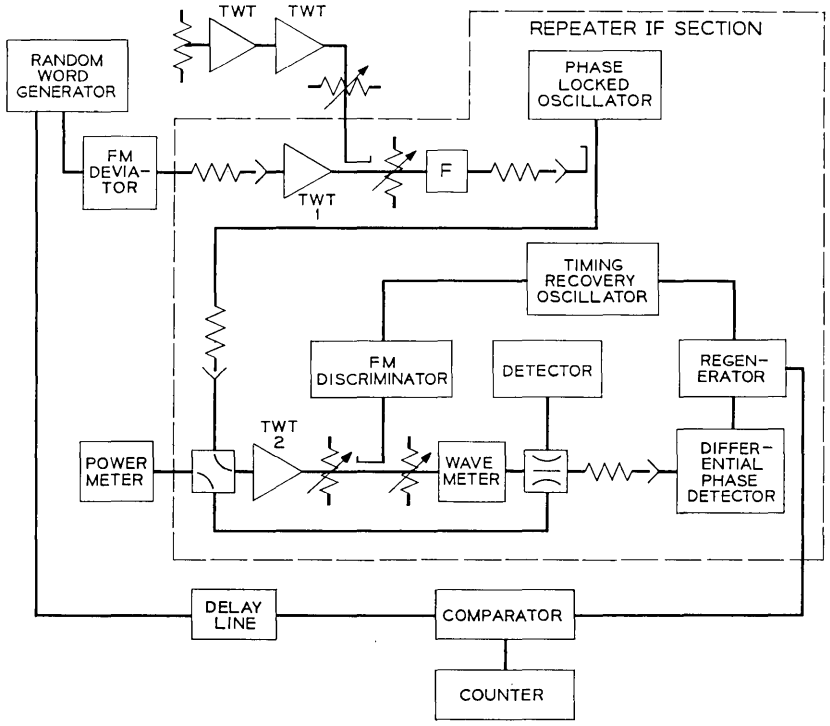


Fig. 3 — Block diagram of the FM-DCPSK model repeater.

$$S(t) = \exp j\left\{\omega_0 t + \int_0^t \omega(t')\,dt'\right\}, \quad \text{where} \quad \int_{(n-1)T}^{nT} \omega(t')\,dt' = \alpha_n \qquad (3)$$

and α_n is as defined for the AM-DCPSK system. An example of such a signal is shown in Fig. 4(d).

The signals used in the repeater to be described in Section III were essentially the AM-DCPSK class, although, in practice, some frequency modulation of the tails of the pulses is inevitable. Signals used in the repeater described in Section IV were essentially of the FM-DCPSK class, although some amplitude modulation is inevitable due to the finite bandwidth of the devices used in the experiment.

The AM-DCPSK signal has the advantage that a comparatively complete theoretical analysis of error rate in the presence of intersymbol interference is available. The FM-DCPSK signal has two advantages, namely, that a phase-locked oscillator can be used to pro-

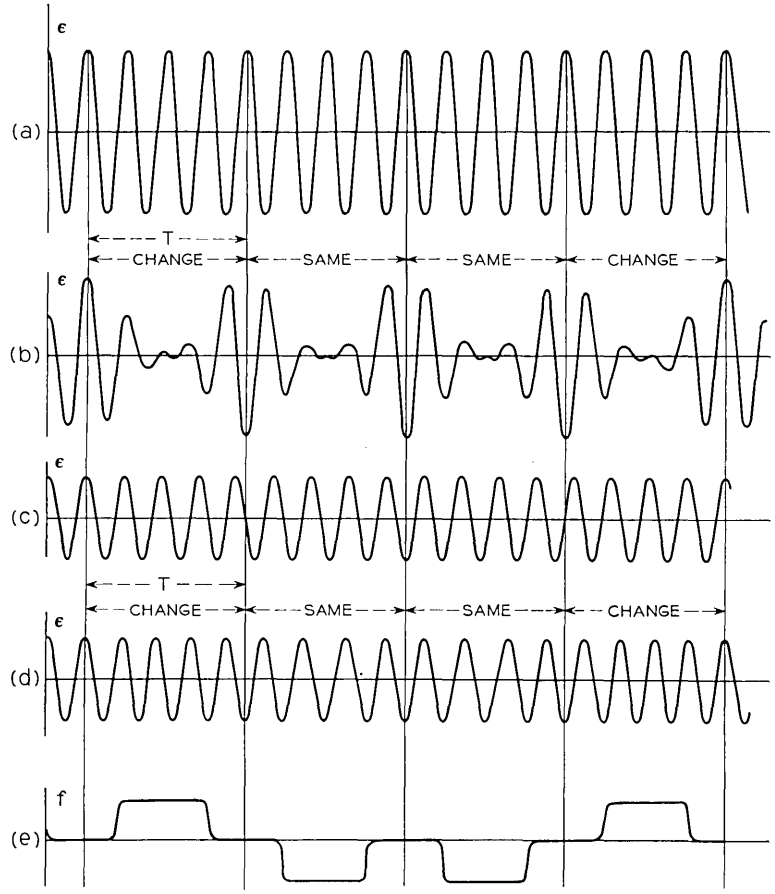


Fig. 4 — (a) unmodulated IF-carrier for AM-DCPSK; (b) idealized AM-DCPSK signal; (c) unmodulated IF-carrier for FM-DCPSK; (d) idealized FM-DCPSK signal carrying the same information as in (b); (e) frequency vs time for the signal in (d).

vide limiting and gain, and that it is comparatively easy to recover timing directly from the signal.

Certain features of the model repeaters shown in Figs. 2 and 3 are identical. Others are quite different. The components which are common to both models are described below. The others are described in Sections III and IV.

All of the traveling-wave tubes used are BTL experimental Model M1917.

2.1 *Random Word Generator*

The random word generator circuitry (for both experimental model repeaters) is identical to the baseband regenerator used with the AM-DCPSK model (to be described in Section III). Wideband random noise from three cascaded amplifiers is used for a random input signal to this "regenerator." The output is, therefore, a sequence of random positive and negative pulses occurring at the 160-megabit rate.

2.2 *Differential Phase Detector*

The differential phase detector ($D\varphi D$) is shown in Fig. 5. This device is used in both models as the $D\varphi D$ and, in addition, an adaptation of the device is used in the comparator for the AM-DCPSK model and another adaptation is used in the timing recovery circuit of the self-timed version of the FM-DCPSK model. The behavior of this device is discussed in the appendix of Ref. 5 for an FM-DCPSK signal. The behavior of the device with one time slot delay for an AM-DCPSK signal is quite straightforward and will be discussed briefly in Section III.

III. THE AM-DCPSK REPEATER

3.1 *General Description of the AM-DCPSK Repeater*

A block diagram of the AM-DCPSK Repeater is shown in Fig. 2. A random binary signal (at baseband) is provided by the random word generator at a bit rate of 160 Mb/s. This random "message" is then transferred onto the 11.2-GHz carrier by the first pulse modulator—

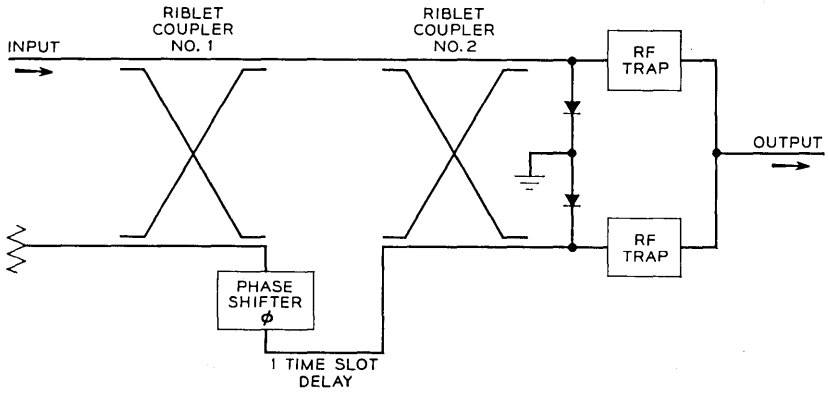


Fig. 5 — Differential phase detector ($D \varphi D$).

the two states now being the two phases 0 or π. The pulse signal is then amplified by TWT 1. Part of the signal (hereafter called the reference signal) is then tapped off in the directional coupler and stored in the delay line for later comparison with the regenerated signal. The remainder of the signal passes through filter F_1. Noise is added to the signal and the combined signal and noise are filtered by F_2 and amplified by TWT 2. The differential detector then processes this corrupted signal and the regenerator makes the decision as to whether a change or a same has occurred, regenerates a positive or negative pulse accordingly, translates back into differential binary and regenerates the translated signal into a form suitable for driving the second pulse modulator. This second pulse modulator (which derives its CW from the same source, a klystron, as the first pulse modulator) then provides a signal which carries the same message (except where errors are made) as the output of the first pulse modulator. This signal is amplified and combined with the reference signal in the comparator. The comparator then signals the counter when an error has been made.

3.1.1 *Differential Phase Detector for AM-DCPSK Pulses*

The AM-DCPSK signal is of the form given in (1). Straightforward analysis of Fig. 5 (assuming ideal square-law detectors) with the delay path equal to one bit interval shows the output of the $D\varphi D$ to be

$$\tfrac{1}{2}S_0^2(0)\cos(\alpha_n - \alpha_{n-1} - \varphi)$$

at the middle (the sampling time) of the nth time slot. For $\varphi = 0$, this signal is $+\tfrac{1}{2}S_0^2(t)$ when $\alpha_n = \alpha_{n-1}$ and is $-\tfrac{1}{2}S_0^2(t)$ when $\alpha_n = \alpha_{n-1} \pm \pi$. For $\varphi = \pi$ the opposite result obtains.

In the actual experiment, the signals are corrupted by both intersymbol interference and noise. The regenerator must make its decision on the basis of this corrupted signal and respond accordingly.

3.1.2 *Pulse Modulator*

The pulse modulators convert an 11.2-GHz CW signal into pulses having one of two possible phase states in accordance with the baseband signal pulse. Positive drive pulses produce an output of one phase and negative drive pulses produce an output shifted by 180 degrees.

The modulator consists of a Western Electric 2B hybrid junction with a matched pair of 1N78 diodes shunting the conjugate arms. One

arm contains a phase shifter in addition to the diodes. The device is shown schematically in Fig. 6.

The operation of the modulator depends upon the return loss of the diodes and the balance between them. They are dc biased to be matched so that very little power is reflected from them. Application of a baseband pulse to the diodes then produces reflected RF pulses in the conjugate arms. With the proper setting of the phase shifter they will add in the output circuit yielding an output pulse. When the drive pulse is of the opposite polarity, the reflected signals in the conjugate arms both differ by 180 degrees from the previous case. The resulting output pulse is then also shifted by 180 degrees.

In practice the phase shift from one state to the other is found to be 180 ± 5 degrees. The ±5 degrees represents the limit of accuracy of the measurement. The modulator loss, (the ratio of CW power in to pulse power out) increases as the CW input drive is increased or as the baseband drive is decreased. Under the low-drive conditions used in this experiment this loss is of the order of 20 dB.

3.1.3 *Baseband Regenerator*

The baseband regenerator (Fig. 7) samples the polarity of the differential detector output at the center of the time slots and translates this information into an output train of equal amplitude positive and negative pulses which are converted into 0 or π RF phases in the pulse modulator. In the absence of errors, this pulse train has a one-to-one correspondence with the original signal.

The regenerator consists of three direct-coupled stages of tunnel-

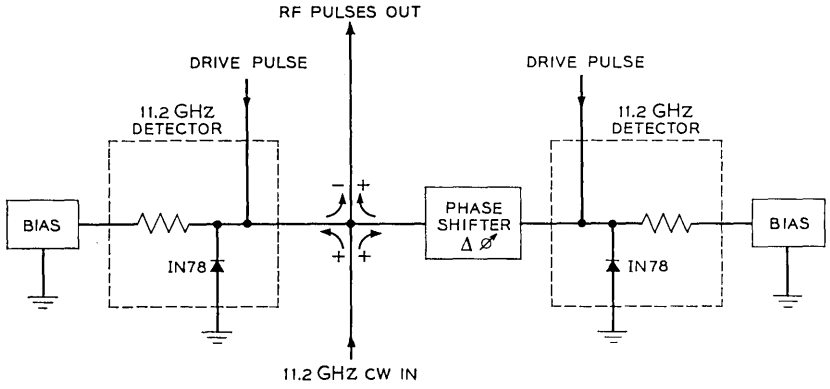


Fig. 6 — Pulse modulator schematic diagram.

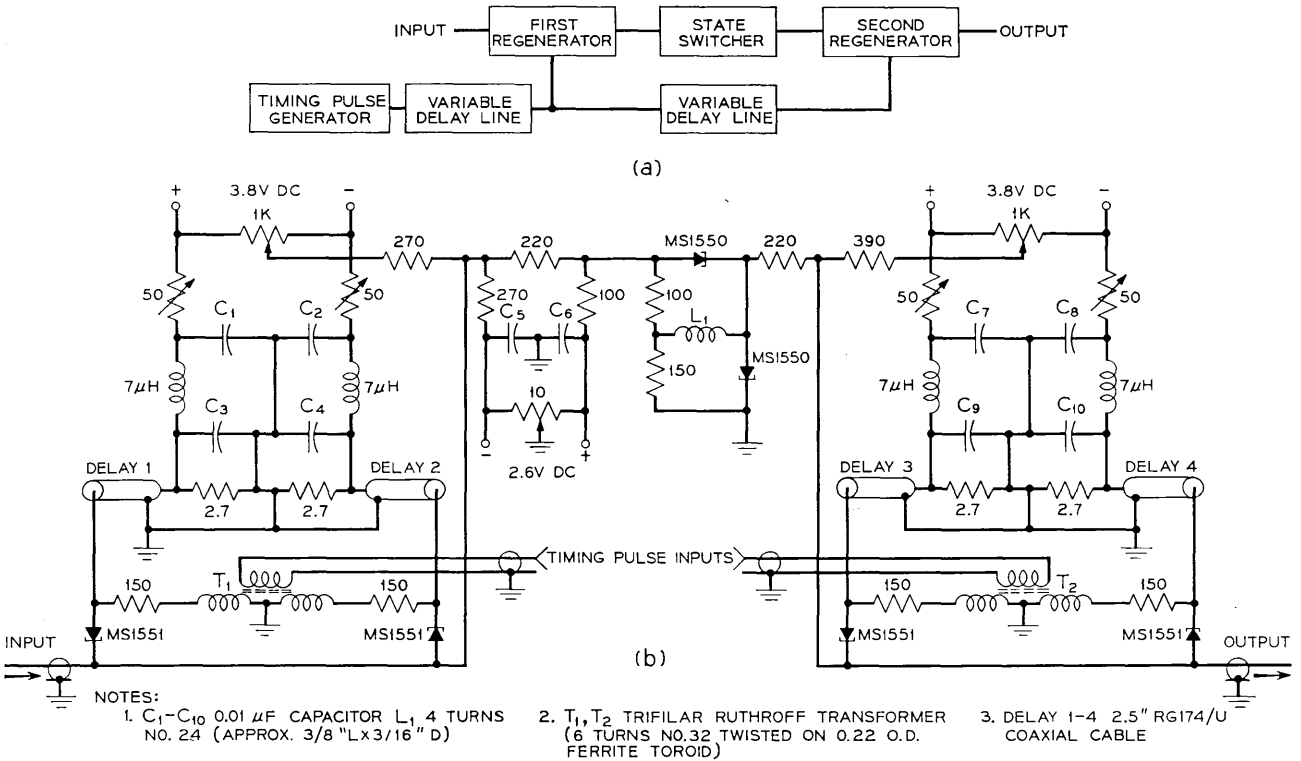


Fig. 7 — (a) Baseband regenerator; (b) baseband regenerator-schematic.

diode pairs. The first and third stages are a form of Goto-pair majority-logic circuit triggered at the 160-Mb/s rate by 1-nsec timing pulses. The middle stage is an adaptation of a well-known "flip-flop" circuit.

The input signal for the first stage is the output of the differential phase detector. (Some delay is provided in the coupling to minimize interference between the input signal and the pulses generated by the first stage.) The timing of the 160-Mb/s triggering pulses is adjusted so that they coincide with the centers of the time slots of the input signal. An output pulse is triggered in every time slot; its polarity is dependent upon the majority polarity of the input signal at the sampling time.

The second stage, the state-switcher, consists of two tunnel-diodes connected in series. Each time a positive pulse arrives the diodes exchange states. The diodes remain in their respective states so long as no positive pulse appears. The circuit time constant allows changes to be made as fast as the maximum requirement of the system—up to once every 6.25 nsec.

When the diodes exchange states the voltage at the junction of the two diodes changes between two discrete values relative to ground. This point is coupled to the junction of the third stage diode pair, and the timing of the third stage triggering pulse is adjusted so as to have decision times occur midway between any possible state changes.

The third stage, operating exactly as the first stage, then puts out pulses which are positive when the state switcher is in one state and negative when it is in the other state. These output pulses provide the drive for the RF pulse modulator described earlier.

3.1.4 *Comparator*

The function of the comparator is to compare the regenerated version of the corrupted signal with the uncorrupted reference signal and to indicate to the counter when an error has been made by the regenerator. Information in a DCPSK signal is carried not by the phase of an individual pulse but by the relative phase between adjacent pulses. The four possible situations are listed in Table I.
In the first two situations in Table I, an error was made; in the last two, no error was made.

As an example, the signal with phases

$$0\ 0\ \pi\ 0\ \pi\ \pi\ 0$$

is equally well represented, after regeneration, by itself or by its exact

TABLE I — POSSIBLE RELATIVE INFORMATION COMBINATIONS IN THE COMPARATOR

Relative phase between two reference pulses	Relative phase between two corresponding regenerated pulses
Change (π)	Same (0)
Same (0)	Change (π)
Same (0)	Same (0)
Change (π)	Change (π)

opposite, viz.,

$$\pi\ \pi\ 0\ \pi\ 0\ 0\ \pi.$$

We can refer to the case where the regenerator is reproducing the input exactly as the "G mode" and the case where it is reproducing the opposite of the input as the "U mode." (Such a distinction has meaning only where some absolute phase reference exists for the original and regenerated signals. This is the case in our experiment since both modulators are supplied from the same RF source; this is probably not the case in an actual system where the regenerated signal would not necessarily be transmitted at the same carrier frequency as the incoming signal.)

A little thought will convince the reader that the effect of an error by the regenerator is to cause the regenerated signal to change modes (U to G or G to U). In our example of the preceding paragraph, suppose that the change between the third and fourth symbols is (erroneously) interpreted by the regenerator (due to noise) to be a same. If the regenerator was originally functioning in the G mode, the result would be as shown

$$\text{Input}: \quad 0\ 0\ \pi\ 0\ \pi\ \pi\ 0$$

$$\text{Output}: \quad \underbrace{0\ 0\ \pi}\ \ \underbrace{\pi\ 0\ 0\ \pi}.$$

$$G \text{ Mode} \uparrow U \text{ Mode}$$

Error

Note that the regenerator responds to sames and changes, not to absolute phases.

Therefore, the comparator must be designed to detect transitions between modes of operation and to respond to these transitions.

The comparator is identical in design and construction to the dif-

ferential phase detector except that both inputs to the first Riblet coupler are used. It is shown in Fig. 8. Let $S_n(t)$ and $R_n(t)$ represent the regenerated and reference signals in the nth time slot, respectively. Then

$$S_n(t) = S_0(t) \exp [j(\omega t + \alpha_n)]$$

and

$$R_n(t) = S_0(t) \exp [j(\omega t + \beta_n + \theta)],$$

where α_n and β_n are the phases of the nth pulse in the regenerated and reference signals respectively and θ represents the phase shift in the long delay line (which is introduced into the reference signal path in order that corresponding time slots of the reference and regenerated signals be compared) and the phase shifter in the reference signal path. A straightforward analysis of Fig. 8 shows that the output of the comparator in the kth time slot is

$$V_k = \tfrac{1}{2}[\cos (\beta_k - \beta_{k-1} - \varphi) - \cos (\alpha_k - \alpha_{k-1} - \varphi) + \sin (\beta_k - \alpha_{k-1} + \theta - \varphi) + \sin (\alpha_k - \beta_{k-1} - \theta - \varphi)].$$

From this equation one can see the following results for the cases of interest. First $V(\varphi = 0, \theta) = -V(\varphi = \pi, \theta)$ for all values of θ.

Consider the case $\theta = 0$, $\varphi = 0$. There is a pulse of amplitude ± 1 whenever an error is made and no pulse when there is no error. The pulse is positive if the regenerated signal changes phase and the reference signal does not. It is negative if the reference signal changes phase and the regenerated signal does not.

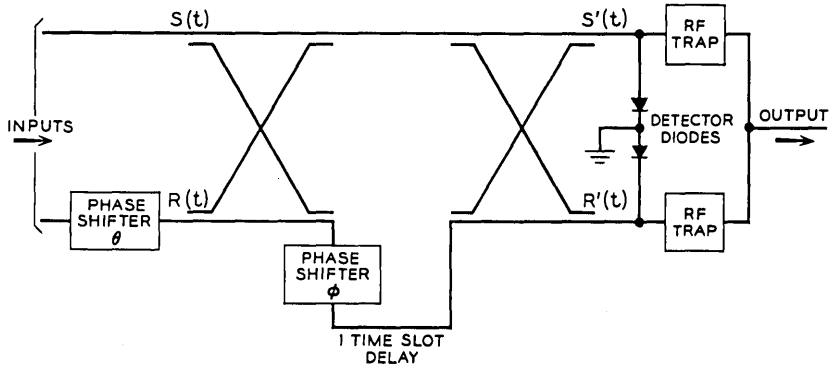


Fig. 8 — Comparator for AM-DCPSK repeater.

Consider the case $\theta = \pi/2$, $\varphi = 0$. There is a pulse of amplitude ± 2 whenever the regenerated signal changes from the G to the U mode but no pulse when the signal changes from U to G. ($\theta = -\pi/2$ gives a pulse on a U $\rightarrow$ G transition but not G $\rightarrow$ U.) The sign of the pulses is as described in the $\theta = 0$ case. This type of operation gives exactly a scale-of-two reduction in the counting rate and an increase of 6 dB in pulse height of the error signal. Since both of these effects are beneficial, this type of operation was generally used.

3.1.5 *Error Counting Circuit*

The detected output of the comparator contains cancelled pulses except where an error has occurred. Errors results in additions of the compared pulses which appear as either positive or negative output pulses. Since only pulses of one polarity can be counted at a time, the positive pulses are eliminated and only the negative pulses counted. Two settings of the comparator phase shifter, φ, 180 degrees apart, are therefore required to make a complete count of the errors.

The output signal is fed into a tunnel-diode unipolar amplitude discriminator (See Fig. 9). This is another Goto-pair of the type used in the regenerator but it is center-biased negatively to prevent positive input pulses from switching a diode. No timing is used. When a negative pulse of sufficient amplitude is received, the pre-biased diode switches to its second state and remains there for a maximum of 2 nsec, at which time the reflected pulse from a short-circuited delay line quickly returns it to its first state and it is ready to receive another pulse well before the next time slot.

The input level is adjusted so that only error pulses are passed. This is to prevent counting as errors the half-amplitude pulses which result from an occasional absence of pulses in time slots of the input signal, These are not properly to be considered as errors but only as trivial deductions from the 160-Mb/s total.

The output from the tunnel-diode amplitude discriminator is fed into a pulse amplifier followed by another amplitude discriminator and thence into a Hewlett-Packard 524-C Counter.

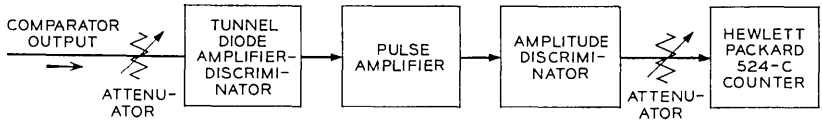


Fig. 9 — Error-counting circuit.

3.2 *Procedure*

3.2.1 *Adjustment of the Random Signal*

The statistics of the signal can be adjusted by changing the level and/or spectrum of the noise which drives the random word generator as well as by adjusting the bias on the various stages of the random word generator. Necessary conditions for a random signal, which were verified by monitoring both the envelope and the differentially detected "random" signals on the sampling scope are

(*i*) the number of pulses in the two phase states be equal,
(*ii*) the number of changes equal the number of sames, and
(*iii*) the four possible transitions (same → same, change → change, same → change, change → same) between time slots be equally populated.

The signal was made to satisfy these three requirements. They constitute our only check on randomness.

3.2.2 *Calibration of Noise and Signal Power*

The average noise power was read directly on a Hewlett-Packard Model 431B power meter. This calibration was checked frequently during the course of the experiment and was held to ± 0.1 dB.

The noise came primarily from a noise source consisting of two model M1917 travelling-wave tubes in tandem followed by a 280 MHz filter. (An additional -26.5 dBm of noise was introduced by TWT 1).

The peak signal power was determined by detecting the envelope of the signal with a calibrated detector and setting the level for a peak power of $+12.0$ dBm. This measurement depends on the calibration of the detector diode, ability to read the pulse height as displayed on the sampling scope accurately, and the choice of which pulse height to use, since intersymbol interference caused several pulse heights to be present. The experimental error here is estimated to be ± 0.4 dB.

Once the peak signal power, S, is set at $+12.0$ dBm, the average signal power corresponding to S, $\langle S \rangle$, can be read on the power meter. The value of $\langle S \rangle$ can be reproduced with an accuracy of ± 0.1 dB and other values of pulse power can be set by changing $\langle S \rangle$ the appropriate amount. Thus even though the absolute value of S is in question by ± 0.4 dB, the error in the *relative* power levels for the same pulse shape is only ± 0.2 dB.

3.2.3 *Determination of Threshold*

Consider a situation where the peak signal power, S, going into the differential phase detector is sufficient for proper operation of the regenerator and the noise power is negligible. Now if S is slowly decreased, a value S_1 will be reached where some errors occur. As S is further decreased, a second value S_2 (about 2.5 dB below S_1) is reached where the regenerator begins to completely disregard the input signal. We define the threshold, T, as

$$T = \tfrac{1}{2}(S_1 + S_2).$$

One finds experimentally that $T \approx 0$ dBm.

3.2.4 *Determination of Intersymbol Interference*

In the presence of intersymbol interference, the envelope-detected pulse will have several values at the decision point due to the four possible states of its nearest neighbors. The ratios of these pulse heights (in dB) can be measured directly on a sampling scope. From this the value of ρ_T, where

$$\rho_T = 20 \log \frac{S_0(0)}{S_0(T)}$$

with S_0 as defined in (1), can be calculated.

3.2.5 *Summary of Experimental Errors*

The experimental errors are summarized in Table II.

3.2.6 *Timing and Center Bias Adjustments*

The timing for the signal pulses and the regenerator is derived from a single 160-MHz clock. The decision time was set by adjusting the

TABLE II — SUMMARY OF EXPERIMENTAL ERRORS

Quantity	Experimental error (dB)
Peak signal power—absolute value	±0.4
Peak signal power—relative value	±0.2
Average noise power	±0.1
Threshold	±1.0
S/N (absolute)	±0.6
S/N (relative)	±0.3
S/T (absolute)	±1.2
S/T (relative)	±0.2
ρ_T	±1.5

timing of the regenerator pulse for minimum error count with a very small input signal.

The center bias setting of the first regenerator strongly influences the ratio of errors made in detecting changes, N_c, to errors made in detecting sames N_s. In the presence of time crosstalk, the optimum setting of the center bias is not zero but rather in the direction to favor changes. In the experiments performed with resolved pulses (no intersymbol interference) the center bias setting was determined by balancing N_c and N_s at a value of S/N such that $N_c + N_s \approx 80$ counts per second. In the experiments performed with unresolved pulses (significant intersymbol interference), the center bias setting was determined by finding the two values of center bias where, in the absence of signal and noise, the regenerator jumps from one state to the other and taking the midpoint between these values. This setting does *not* optimize (exactly) the error rate, but it does allow a more direct comparison with theory.

3.3 *Results and Comparison With Theory*

3.3.1 *Narrow-Pulse Experiment*

The first experiment to be discussed is the measurement of error-rate as a function of S/N for several values of S/T. For this experiment, the filter F_1 was removed entirely and filter F_2 was placed between the second noise tube TWT-N2 and the noise attenuator. The pulses were then completely resolved as seen in Fig. 10(a). The eye diagram (output of the differential phase detector) is shown in Fig. 10(b). The experimental results are plotted in Fig. 11 along with the corresponding theoretical curves.[6]

The theory for this comparison assumes that all pulses are of the same peak power S, whereas in fact the modulator produces pulses of two slightly different amplitudes. This is an important effect for small values of S/T and probably accounts for part of the discrepancy in the data for S/T = 4.5 dB and a great deal of the discrepancy for S/T = 3.0 dB. Also note that small errors in determination of threshold are extremely important when S/T $\approx$ 3.0 dB.

3.3.2 *Unresolved-Pulse Experiment*

Filters F_1 and F_2 (as shown in Fig. 2) were adjusted to have 6-dB bandwidths of 250 MHz and 262 MHz, respectively. The overall 6-dB bandwidth of the two filters in series as used in the experiment was then 173 MHz. This widened the pulse considerably. The pulse envelope

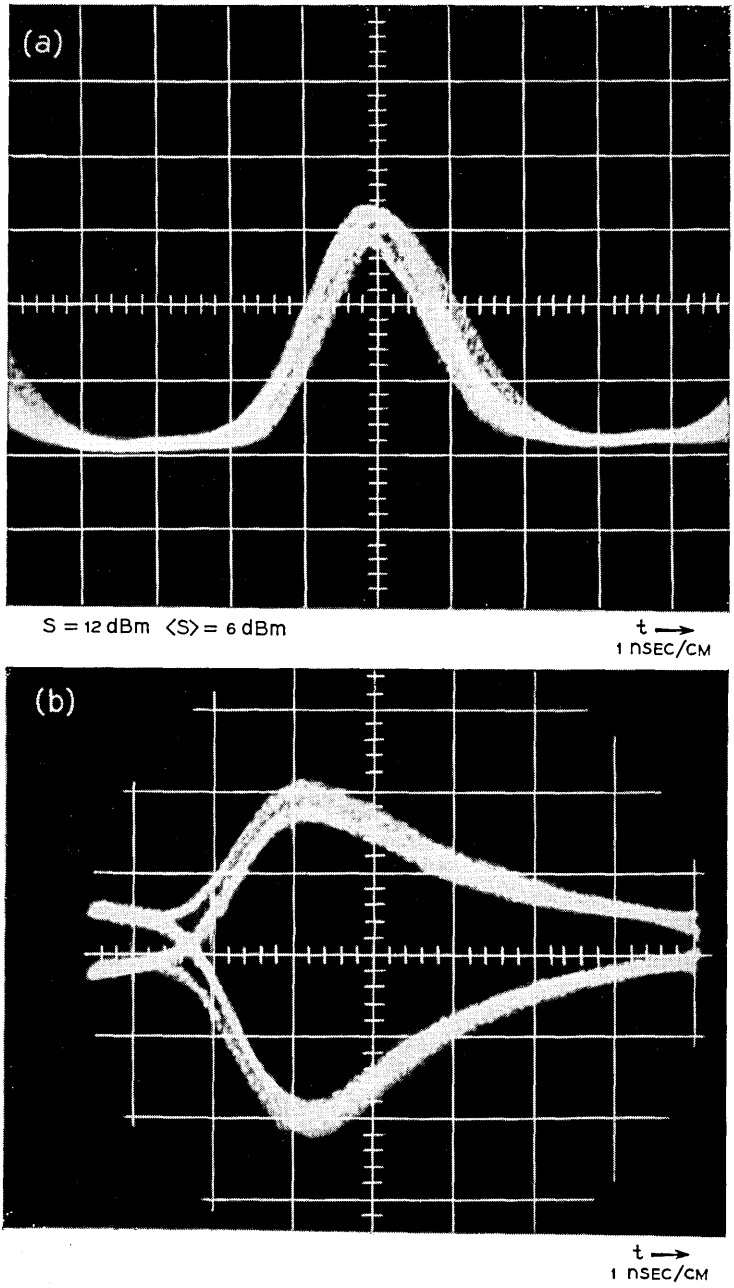


Fig. 10 — Narrow pulse experiment wave forms. (a) Envelope-detected RF pulse. (b) Eye-diagram (output to differential phase detector, input to regenerator).

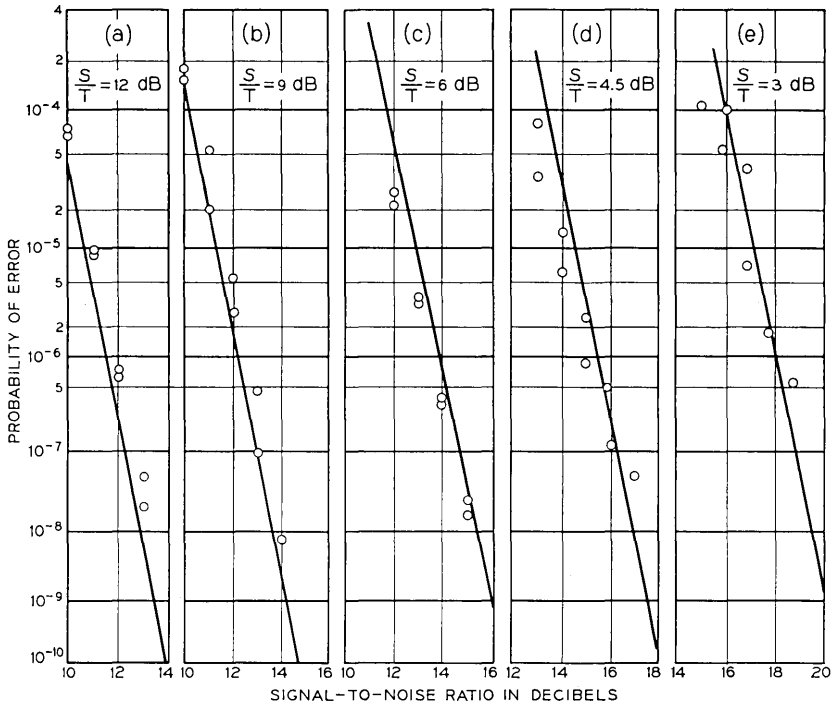


Fig. 11 — Error rate vs S/N for narrow pulses (negligible intersymbol interference). Points represent experimental values; curves are theoretical results from Ref. 6.

and the eye diagram for this signal are shown in Fig. 12. The intersymbol interference, ρ_T, was measured to be 22.5 ± 1.5 dB per tail. The experimental results are plotted in Fig. 13 along with the theoretical curves.

3.3.3 *Factor-of-Two Experiment*

Various arguments have been presented in the past which (erroneously) conclude that (at least under conditions of low error rate) an error in decision results in two errors (in adjacent time slots) in the regenerated signal. Salz and Saltzberg[7] have shown that these arguments are fallacious and that a single error in decision results, with about 90 percent probability, in a single error in the regenerated signal at the error rates used in these experiments.

Since these double errors, if they existed, would occur in adjacent time slots, our counter (video bandwidth is 10 MHz) would not be able

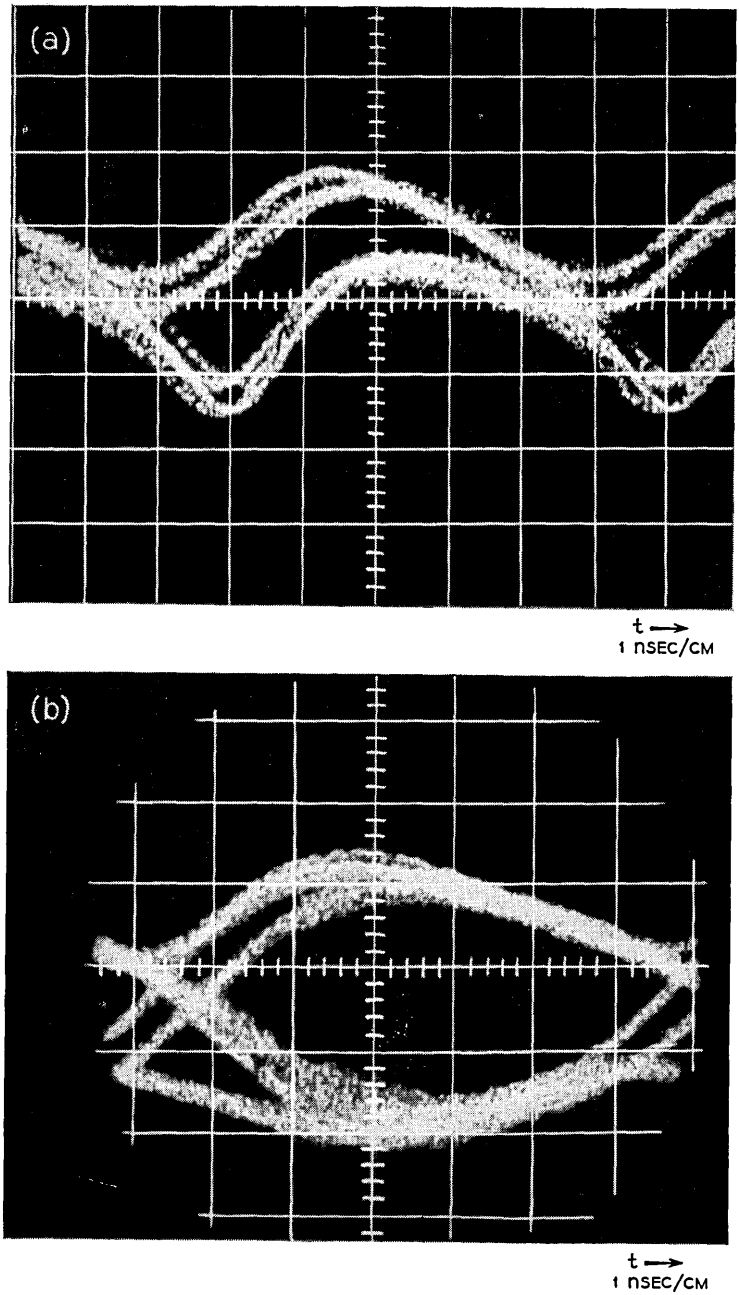


Fig. 12 — Widened pulse experiment wave forms. (a) Envelope-detected RF pulse. (b) Eye-diagram (output to differential phase detector, input to regenerator).

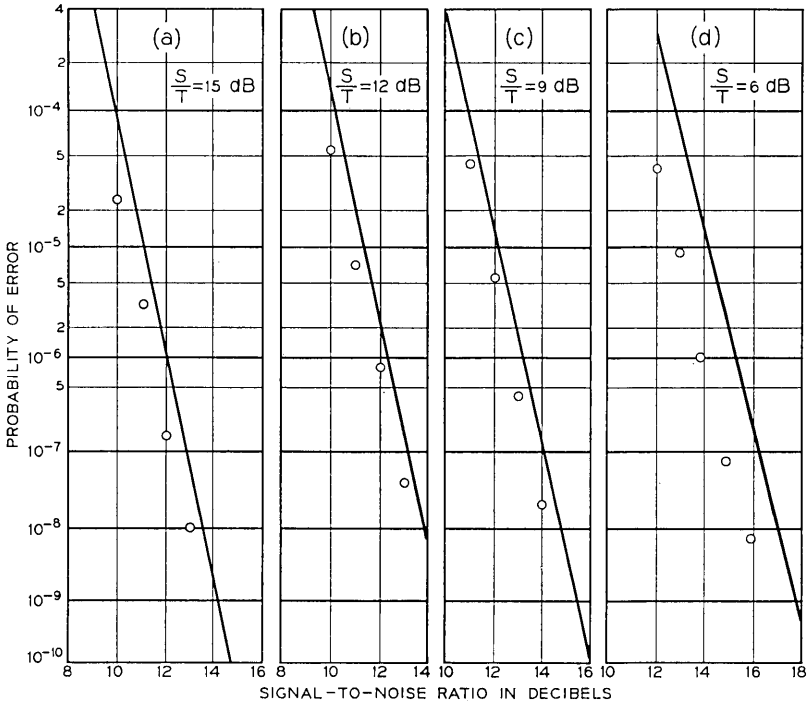


Fig. 13 — Error rate vs S/N for band-limited pulses. Points represent experimental values, curves are theoretical values from Ref. 5 evaluated $\rho_T = 22.5$ dB (the experimentally determined value for this waveform).

to resolve these pulses without benefit of the scale-of-two feature of the comparator. As explained in Paragraph 3.1.4 this scale-of-two can be removed by changing the phase of the reference signal 90 degrees. When this is done, the total count should double if the number of double errors is negligible. On the other hand, if double errors were created for every decision errors, the total count would increase only 50 percent since (for random signals) half of the "secondary" error-pulses from the comparator would be of the same polarity as their associated "primary" error pulses and hence could not be resolved in the counter.

Within experimental errors of a few percent, the total count is doubled when the scale-of-two is removed indicating that the errors in the regenerated signal do *not* occur in pairs.

3.3.4 *Test of Effect of Phase Error in the Differential Phase Detector on Error Rate*

The degradation in S/N produced by a small error, $\delta\varphi$, in the phase shift (or equivalently in the length) of the delay loop of the differential phase detector has been calculated.[5] This effect is of interest because $\delta\varphi$ is a function of temperature; hence, it cannot be made and kept arbitrarily small in a practical repeater. Fig. 14 shows the experimental points and the theoretical curve of equivalent degradation in S/N as a function of $|\ \delta\varphi\ |$. The points designated ○ were taken with $\delta\varphi < 0$, those designated □ with $\delta\varphi > 0$.

IV. THE FM-DCPSK REPEATER

4.1 *General Description of the FM-DCPSK Repeater*

Fig. 3 is a block diagram of the FM-DCPSK repeater. The overall operation of this repeater is quite similar to that of the AM-DCPSK repeater described in the preceding section. The random word generator, error counting circuit, and RF noise source are identical to those described in Paragraphs 2.1, 3.1.5, and 2.1, respectively.

4.1.1 *FM Deviator*

The pulse modulator used in the AM repeater is replaced with an FM deviator. This FM deviator consists of a frequency-modulated Esaki

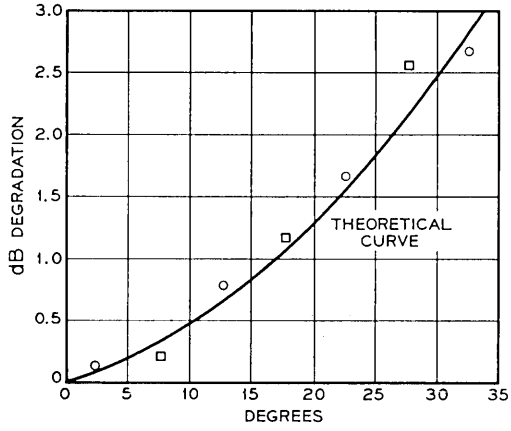


Fig. 14 — Degradation in effective S/N as a function of phase shift error in the differential phase detector.

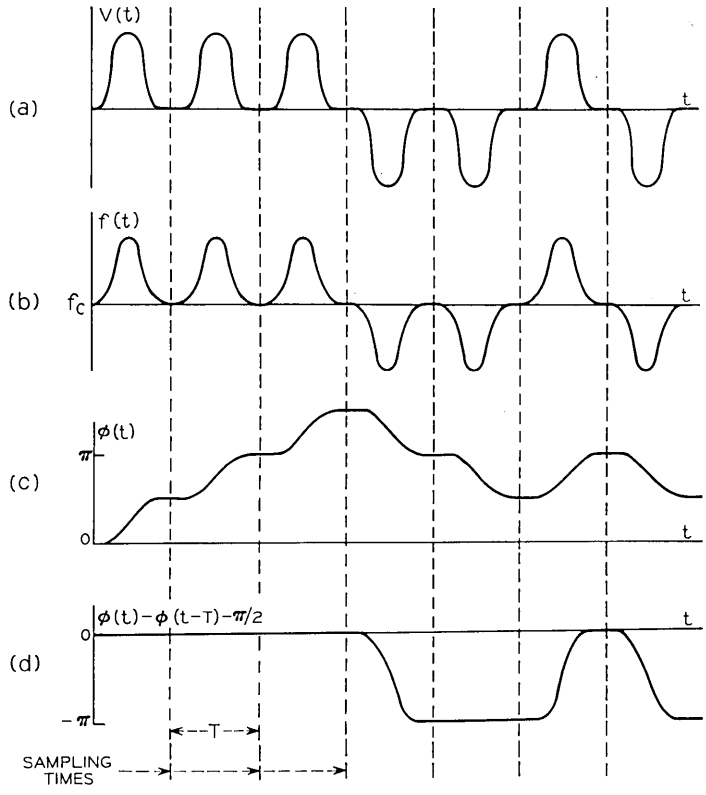


Fig. 15 — FM-DCPSK signal; (a) modulating voltage vs time; (b) frequency vs time; (c) phase vs time; (d) differential phase vs time.

diode oscillator. Modulation is produced by pulsing the bias voltage either positively or negatively in every time slot. The binary information is contained in the choice of polarity. This voltage pulse causes the oscillator frequency f to deviate from its nominal value f_c. This is illustrated in Fig. 15(a) and (b). This modulation causes a phase change $\delta\varphi_n$ given by

$$\delta\varphi_n = \int_{nT}^{(n+1)T} \{f(t) - f_c\}\, dt.$$

During the nth time slot sampling is performed after the phase shift has been completed (i.e, at the center of the intervals on which $f(t)$ equals f_c in Fig. 15(b); thus, $\delta\varphi_n$ represents the phase change between the nth and the $(n+1)$th samples.

Optimum noise immunity is obtained when the two possible signal states are anticorrelated, that is, when the two possible values of $\delta\varphi_n$ differ by π. This is achieved by adjusting the pulse amplitude until $\delta\varphi_n$ equals $\pm\pi/2$. In the following it will be assumed that this condition is satisfied.

The phase (relative to the phase of the carrier frequency) $\varphi(t)$ for the modulation illustrated in Fig. 15(a) and (b) is shown in 15(c). The differential phase detector used for this type of modulation is identical to the one described in the Paragraph 3.1.1. The phase shifter is adjusted to add (or subtract) a constant $\pi/2$ shift in the delayed signal so that it compares $\varphi(t)$ with $\varphi(t - T) + \pi/2$. Since $\varphi(t)$ and $\varphi(t - T)$ differ $\pm\pi/2$, $\varphi(t)$ and $\varphi(t - T) + \pi/2$ differ by 0 and π and the decision making circuitry operates just as it did for the system described in Section III. The quantity $\varphi(t) - [\varphi(t - T) + \pi/2]$ is illustrated in Fig. 15(d).

The requirements on the FM deviator differ from the usual requirements on an FM modulator in that only the area under the frequency-versus-time curve is important. There is no requirement that frequency be a linear or even a continuous function of the voltage.

Each of the FM deviators built for this experiment consists of a 0.5-mA GaAs Esaki diode mounted directly across 50-mil high X-band waveguide with one terminal grounded and the other brought out through an 11.2-GHz coaxial trap[8] as shown in Fig. 16. They have an output power of the order of -17 dBm. They are quite stable and have suffered no noticeable degradation over periods of up to 6 months.

The effect of the modulation on the amplitude of the output of the FM deviator turns out to be very small if the oscillator is properly tuned. Even after filtering the output to the bit-rate-bandwidth the amplitude variation is only about ± 1 dB. Since the $\pm\pi/2$ modulation gives complete symmetry about the carrier frequency f_c, the optimum value of the free-running frequency of the PLO driven by this signal is also f_c.

4.1.2 *Baseband Regenerator*

The baseband regenerator used in this repeater is just the first stage or "first regenerator" of the device used in the AM repeater. Instead of driving a state switcher this "first regenerator" drives the FM deviator directly. The function of the state switcher, namely, to translate from straight binary to differential binary is accomplished automatically by the manner in which the FM deviator functions.

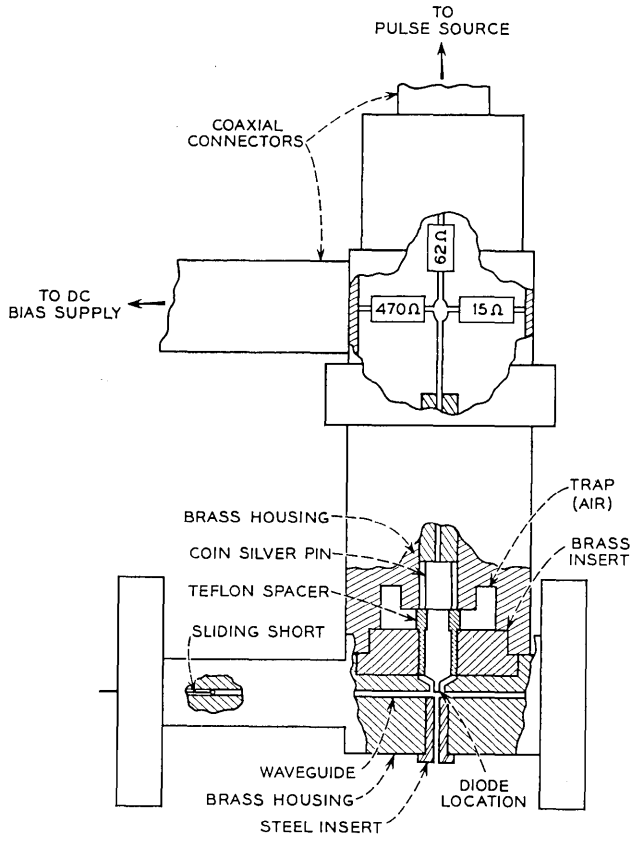


Fig. 16 — FM deviator.

4.1.3 *Timing Recovery*

One of the problems associated with any digital regenerative repeater is that of providing a clock to allow proper sampling of the received signal. The type of signal used in this experiment is particularly suitable for timing recovery because there is a frequency change in each time slot regardless of the message statistics. In this repeater, the timing was recovered by taking a portion of the incoming signal and putting it into a device which is like the differential phase detector except that the delay line is approximately 0.6 bit interval long. This device is described in the appendix.

4.1.4 *Limiter*

An Esaki diode limiter was incorporated into the circuit in order to take advantage of the improvements predicted in Ref. 6. This device consisted of an Esaki diode oscillator built directly in 50-mil high X-band waveguide. Input and output was accomplished by means of a Riblet coupler in the experimental repeater (only because a circulator was not at hand). In addition to limiting, the device gave about 20 dB of gain.

The repeater was operated both with and without the limiter; results of both types of operation are given in Paragraph 4.3.

4.1.5 *Comparator*

The comparison of input and output signals is made at baseband in this experiment (rather than at X-band as in the AM-DCPSK experiment). The output of the random word generator is split with a 6-dB matched splitter. One branch goes to the FM deviator, the other through a delay line to a baseband hybrid where it is compared with the output of the baseband regenerator. The polarities are such that if no error has been made the pulses cancel in the output arm of the hybrid whereas if an error has been made, the pulses add. This signal is then applied to the error counting circuit described in Paragraph 3.1.5.

Note that the baseband signal contains the information directly, not in differential form as does the X-band signal. Thus, some of the complicated logic of the X-band comparator is avoided.

4.2 *Procedure*

The procedure followed was similar to that described in Paragraph 3.2. The two significant differences are (*i*) the timing was derived from the timing recovery circuit, not from the 160-Mb/s clock which ran the random word generator, and (*ii*) the signal-to-noise ratio, when the limiter was used, was measured at the input to the limiter.

4.3 *Results and Comparison with Theory*

The results of experiments to measure error-rate versus S/N for the FM-DCPSK system are given in Fig. 17 for the case where a limiter was *not* used and in Fig. 18 for the case where a limiter was used. These results indicate that for an error rate of 10^{-9} the system operates with about 1.5-dB degradation from theoretical ideal without the limited and with about 0.5 to 0.8-dB degradation with the limiter. This

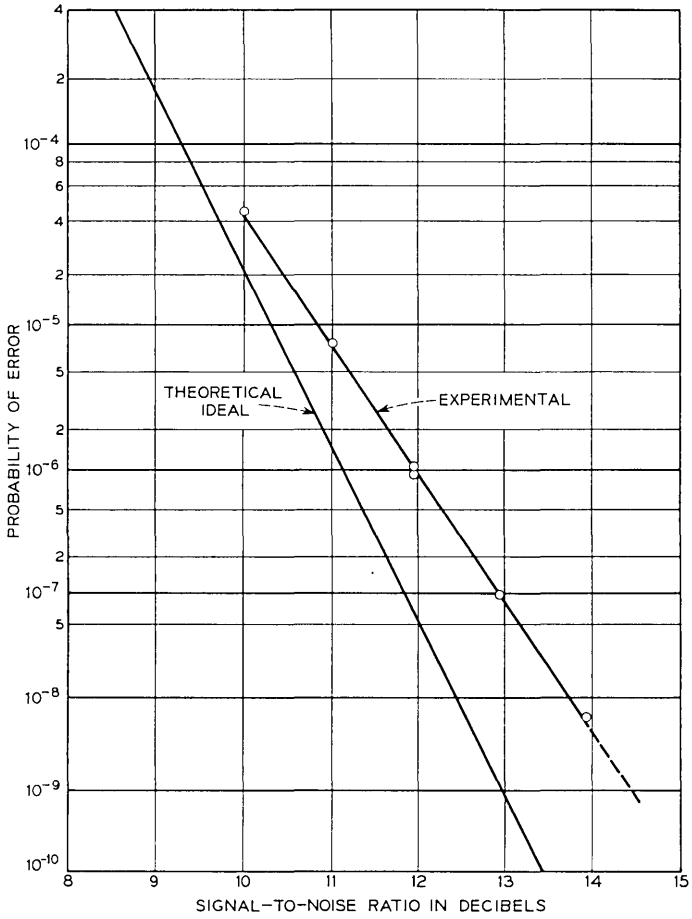


Fig. 17 — Error rate vs S/N for the unlimited FM-DCPSK signal.

degradation is due in part to the finite-width decision threshold discussed in Ref. 6 and in part to the intersymbol interference discussed in Ref. 5. The signal-to-threshold ratio is of the order of 12 dB for these experiments which, ignoring intersymbol interference, should account for about 0.4 dB of the degradation for the unlimited case. For the limited case the degradation due to a 12-dB signal-to-threshold ratio is about 0.1 dB. It is impossible to consider the intersymbol interference quantitatively because the exact form of the signal is unknown. Nevertheless, it is not unreasonable to assume that it

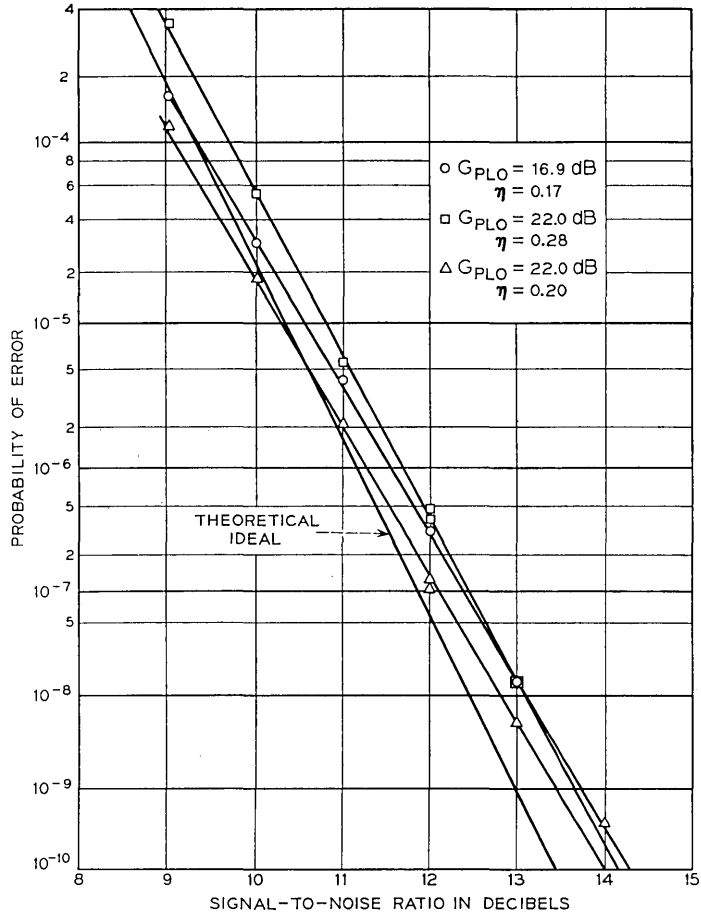


Fig. 18 — Error rate vs S/N for the limited FM-DCPSK signal.

accounts for a large part of the remaining discrepancy between the experimental and theoretical curves in Figs. 17 and 18.

V. CONCLUSIONS

Both AM-DCPSK and FM-DCPSK type systems are technically feasible. The FM-DCPSK type seems preferable in that timing information can be obtained by the method discussed above, whereas when the AM-DCPSK system is band limited to bandwidth of the order of the bit rate the pulses are no longer resolved and timing is quite

difficult to recover. In addition, a phase-locked oscillator serves as a suitable limiter for FM-DCPSK systems whereas a different—and perhaps more complicated—limiter would be necessary for the AM-DCPSK system. Finally, the baseband circuitry is much simpler for FM-DCPSK because the translation from straight binary to differential binary is accomplished automatically by inherent properties of the FM deviator.

VI. ACKNOWLEDGMENTS

The authors are particularly grateful to W. D. Warters for numerous suggestions throughout the course of these experiments, to H. M. James for assistance in construction of the X-band circuitry, and to J. H. Johnson for providing the X-band diodes for the limiters and the FM deviators.

APPENDIX

Timing Recovery from an FM-DCPSK Signal

The basic device described in the appendix of Ref. 5 and shown in Fig. 5 can be used as an FM discriminator for timing recovery by proper choice of the delay time τ. For this device one chooses τ such that $\omega_0\tau = m\pi$. Then (22) of Ref. 5 is, to first order, independent of the sign of $\omega(t')$ and thus independent of message statistics. One has

$$V(t) = \cos\left\{\int_{t-\tau}^{t} \omega(t')\, dt'\right\}. \tag{4}$$

The analysis cannot be carried further without assuming a particular form for $\omega(t')$. As an example, consider the signal given by (3) with $\omega(t')$ given by

$$\omega(t') = a_n \frac{\pi}{T} \cos^2 \frac{\pi t'}{T},$$

where $a_n = \pm 1$ depending on the message and changes sign only at the points $t' = (n + \frac{1}{2})T$. Substituting this into (4) gives

$$V_0(t) = \cos\left\{\frac{\pi\tau}{2T} + \tfrac{1}{2}\sin\left(\frac{\pi\tau}{T}\right)\cos\left(\frac{2\pi t}{T} - \frac{\pi\tau}{T}\right)\right\} \tag{5}$$

$$V_1(t) = \cos\left\{\frac{\pi}{T}\left[t - (n + \tfrac{1}{2})T - \frac{\tau}{2}\right] + \tfrac{1}{2}\cos\left(\frac{\pi\tau}{T}\right)\sin\left(\frac{2\pi t}{T} - \frac{\pi\tau}{T}\right)\right\}, \tag{6}$$

where (5) applies if a_n does *not* change sign on the interval $(t, t\text{-}\tau)$ and (6) applies if it does.

A study of (5) and 6) reveals that for $\tau \ll T$ the bit-rate-frequency component of V_0 and V_1 is quite small. As τ is increased, the bit-rate-frequency component of V_0 increases up to $\tau \approx 0.6T$ and then decreases to zero as τ is increased to T. The bit-rate-frequency component of V_1 increases as τ increases from zero to T. Thus, the optimum value of delay for the timing recovery discriminator is $\tau \approx 0.6T$.

The delay, τ, is related to T by another constant. Since we must have $\omega_0 T = (m + \frac{1}{2})\pi$ (see Appendix A of Ref. 5), and we also require $\omega_0 \tau = m'\pi$ for (4), τ and T are related by

$$\frac{\tau}{T} = \frac{m'}{m + \frac{1}{2}}.$$

Thus, one chooses m' such that $m'(m + \frac{1}{2})$ is as near 0.6 as possible.

REFERENCES

1. Miller, S. E., Waveguide as a Communication Medium, B.S.T.J., *33,* 1954, pp. 1209–1265.
2. Rowe, H. E. and Warters, W. D., Transmission in Multimode Waveguide with Random Imperfections, B.S.T.J., *41,* May, 1962, pp. 1031–70.
3. Kotel'nikow, V. A., *The Theory of Optimum Noise Immunity,* McGraw-Hill Book Co., New York 1959.
4. Lawton, J. G., Comparison of Binary Data Transmission Systems, Proc. Conf. Mil. Elec., 1958.
5. Hubbard, W. M., The Effect of Intersymbol Interference on Error Rate in Binary Differentially-Coherent Phase-Shift-Keyed Systems, B.S.T.J., this issue, pp. 1149–1172.
6. Hubbard, W. M., The Effect of a Finite-Width Decision Threshold on Binary Differentially Coherent PSK Systems, B.S.T.J., *45,* February, 1966, pp. 307–319.
7. Salz, J. and Saltzberg, B. R., IEEE Trans., *CS 12,* 1964, p. 202.
8. DeLoach, B. C., unpublished work.

The Suppression of Monocularly Perceivable Symmetry During Binocular Fusion

By BELA JULESZ

(Manuscript received February 10, 1967)

Symmetries that we can perceive with one eye can be made to disappear during binocular fusion—that is, a symmetrical pattern in one of a pair of stereoscopic images may not be seen when we view the pair stereoscopically. This phenomenon should not be confused with the classically-known binocular rivalry in which the left and right images cannot be fused and one of the images is alternately suppressed. The type of suppression phenomenon reported here is obtained for computer-generated random-dot patterns in which locally each picture element can be fused in a stable way. The binocularly suppressed symmetry can be one-, two-, and four-fold, and the experiments give some insight into the processes underlying the perception of symmetry. In addition to symmetries, it becomes possible to scramble text by exhibiting it stereoscopically.

I. BINOCULAR FUSION, RIVALRY, AND A THIRD POSSIBILITY

Recently, the author added a third possibility of perceptual response to the class of stereoscopic images traditionally consisting of binocular fusion and binocular rivalry.[1] In these computer-generated stereoscopic images the local and global properties are juxtaposed such that locally each picture element can be stereoscopically fused, causing the monocularly apparent global symmetry to disappear in the fused percept.

In the demonstration of Ref. 1, the left stereoscopic image consisted of randomly selected black and white dots with *bilateral* (one-fold) symmetry across the center *horizontal* axis. The right image was obtained from the left image by subdividing it into 20 horizontal stripes (of five picture element width) and shifting every even stripe to the left and every odd stripe to the right by two picture elements. Such a stereo image is shown in Fig. 1. When monocularly viewed, the hori-

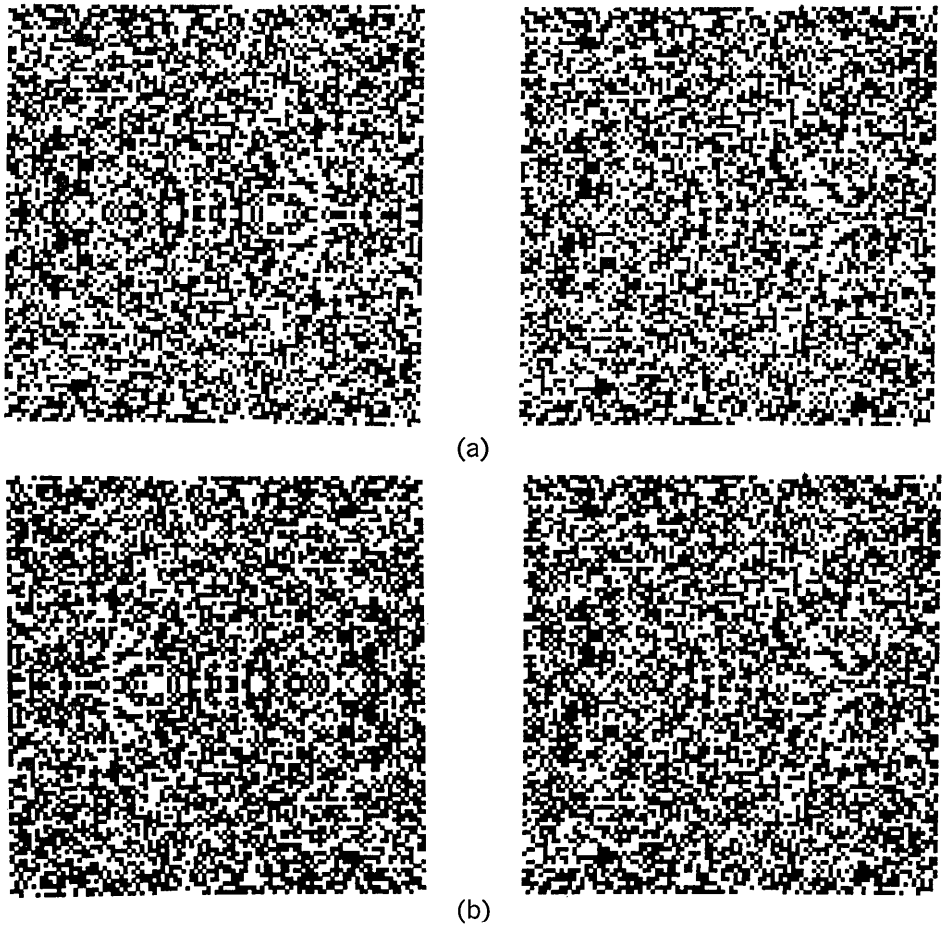


Fig. 1 — (a) Stereogram which, when monocularly viewed, contains an image of bilateral symmetry across the horizontal axis. When viewed stereoscopically, horizontal stripes are perceived in depth, and the symmetry is suppressed. (b) Stereogram identical to (a) except that the left image is mirror reflected to permit stereoscopic viewing with the supplied front-surface mirror as described in the appendix.

zontal bilateral symmetry is apparent in the left image, whereas the right image seems almost random. When binocularly viewed, the horizontal stripes are perceived alternately in depth, and if the symmetric pattern is viewed with the non-dominant eye the bilateral symmetry is suppressed in the fused percept.*

* These stereograms can be viewed stereoscopically by using a prism in front of one eye or other similar stereoscopic device. For those readers who do not

In Ref. 1 it was emphasized that the demonstration was only an example and others could be devised along these lines. Because of the theoretical importance of this stimulus class, the present article demonstrates several new examples. They are more powerful than that of Fig. 1 and help to clarify some factors in symmetry perception.

II. POSSIBLE IMPROVEMENTS OF THE ORIGINAL EXPERIMENT

Although the original demonstration in Fig. 1, served its purpose, it suffered from some inadequacies. The primary limitation of the stimulus is our difficulty in perceiving bilateral symmetry along a horizontal axis. Is it possible to create stereoscopic images in which more powerful monocular percepts of symmetry are binocularly suppressed? For instance, can the monocularly apparent bilateral symmetry across a *vertical* axis be made to disappear, or what is more, can we binocularly suppress the strongly perceivable two-fold or four-fold monocular symmetries?

Another criticism can be raised when inspecting Fig. 1. One might argue that the right image is not symmetrical, but because of a few recognizable clusters which are similar in the upper and lower half fields, the appearance of the right image deviates from complete randomness. Although we could argue that the bilateral symmetry in the left field is effortlessly perceived, whereas in the right field we have to scrutinize the stimulus to detect departure from randomness, nevertheless, attempts should be made to deal with this criticism.

A third objection might concern the loose way of specifying the amount of symmetry suppression in the binocular percept. Of course, it might be a simple task to let subjects rank order the amount of symmetry in the left and right monocular percepts and in the fused binocular percept. On the other hand, the phenomena shown here are universal, i.e., observers with adequate stereoscopic vision do deviate little in their judgments. The skeptic can easily check the validity of these findings himself. Therefore, instead of quantification, all experimental efforts have gone into the creation and display of increasingly sophisticated stimuli. It should be mentioned that after rank

possess such viewing aids, a front-surface mirror is included inside the back cover of this issue. Since all the stereograms except Figs. 1, 2 and 10 contain an image with vertical bilateral symmetry, they can be fused both ordinarily or with the aid of the mirror as described in the appendix. Figs. 1(b), 2(b), and 10(b) cannot be fused with the aid of a prism but only with a mirror—Figs. 1(a), 2(a), and 10(a) can only be fused with a prism. In order to obtain the described binocular percepts with mirror viewing and not the reversed depth percepts, the *left* images of the stereograms should be viewed with the aid of the mirror by the left eye and the right images directly by the right eye.

ordering the left and right monocular percepts with respect to symmetry (or lack of symmetry), the symmetry in the binocular percept does *not* have to be the mean of the monocular percepts, but can be even weaker than that of the monocular percept with the scrambled symmetry.

In the experiments reported here a systematic attempt was made to break up the clusters and strengthen the perception of monocular symmetries. Binocular suppression of symmetry was still observed.

III. NON-TOPOLOGICAL OPERATIONS IN BINOCULAR VISION

The existence of locally correlated but globally uncorrelated stereoscopic images is based on a fundamental difference between monocular and binocular vision. When viewed with one eye, the retinal projections of objects usually change size and shape in a continuous manner. Exceptions are a few hidden parts which may suddenly enter or depart the visual field. Therefore, monocular perception mostly operates on continuous (topological) transformations of the stimuli, and prolonged departure from spatial-temporal continuity results in strange phenomena and distortions. (For example, imagine a television picture out of horizontal synchronism.) On the other hand, binocular vision can easily combine non-topologically related stereoscopic images to yield stereopsis and fusion. Parts of one stereoscopic image can be broken up and shifted horizontally in the other image and, if these shifts are kept within the critical limit of disparity, the two images will be combined in a single three-dimensional percept. These shearings and displacements when skillfully applied to binocular viewing can destroy many monocular percepts, particularly that of symmetry.

IV. THE ROLE OF CLUSTERS IN SYMMETRY PERCEPTION

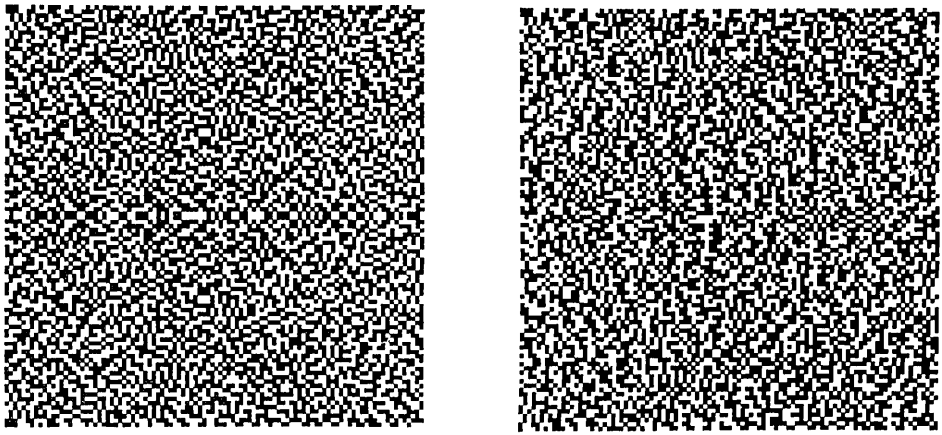
The right image of Fig. 1, although not symmetrical, is seen to deviate from randomness when carefully viewed. One of the main cues for bilateral symmetry has been removed, namely the large symmetrically shaped clusters in the immediate neighborhood (within $\pm$ 5 picture elements) of the horizontal symmetry axis. On the other hand, within the five picture-element wide horizontal stripes, some characteristic micropatterns are formed by chance; these can be recognized and matched in the upper- and lower-half fields. In the binocular percepts, because each of these similar micropattern pairs is seen at opposite depth levels, their matching becomes more difficult. Therefore,

the binocular percept seems even more random than the right monocular view.

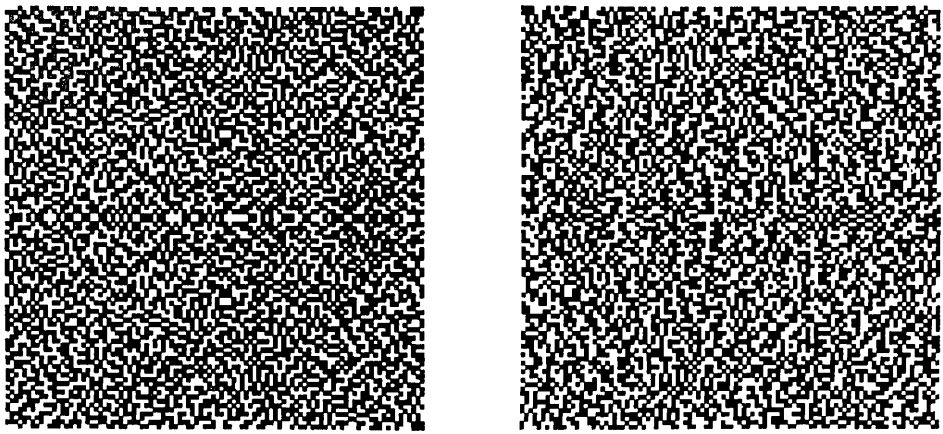
In addition to the microclusters formed by chance, the same random fluctuations can produce macroclusters which extend to large areas. The alternate horizontal shifts by a few picture elements, performed on the even and odd stripes are not adequate to break up these large dark or bright areas, and these are the last cues left in the binocular percept to reveal symmetry.

Is it possible to reduce the traces of symmetry by breaking up the micro- and macro-clusters in the random dot textures? Several attempts were made to break up the clusters. The first was a preprocessing of the left field of Fig. 1. Whenever a picture element was surrounded by more than six picture elements of the same brightness, its brightness value was complemented from black to white or vice versa; otherwise it was kept unchanged. This intervention prevented the formation of black and white clusters within a 3×3 array, but long horizontal and vertical lines were still present and were strongly perceivable both in the left and right images. A second cluster-breaking operation was therefore performed. If out of seven adjacent picture elements which lay on a horizontal or vertical line, six or more were white or six or more were black, the center one was complemented to opposite brightness value. The outcome of these two cascaded operations applied on a random dot pattern (similar to the left field of Fig. 1) yielded the left field of Fig. 2. This image looks very intricate, and the horizontal bilateral symmetry is mainly apparent because of the patterns close to the symmetry axis and because of some long diagonal line clusters that remained intact.

Another way of reducing recognizable micropatterns is to diminish the width of the horizontal stripes which are alternately shifted in the right image. In Fig. 2 the horizontal stripes in the right image are only two picture-elements wide (in contrast to the five picture-element stripe width in Fig. 1). Thus, no micropattern larger than two picture elements in the vertical extent stays unchanged. The right image of Fig. 2 gives a more random impression than the right image of Fig. 1. Unfortunately, with smaller stripe width, the amount of disparity has to be reduced in order to obtain stable stereopsis. Therefore, in Fig. 2 the even stripes are shifted to the left by two picture elements while the odd stripes are kept unchanged. The total disparity of Fig. 2 is only two picture elements, whereas in Fig. 1 it was twice this amount. Perhaps this precaution is unnecessary as seen in Figs. 7 and 8 which con-



(a)



(b)

Fig. 2 — (a) Stereogram which is similar to Fig. 1 except that the width and binocular disparity of the stripes is reduced and cluster formation of many proximate dots of equal brightness is prevented by preprocessing. When viewed stereoscopically, a transparent textured surface is perceived above the background (b) Stereogram identical to (a) except that the left image is mirror reflected to permit stereoscopic viewing with the supplied front-surface mirror as described in the appendix.

tain two picture-element wide stripes and a total disparity of four picture-elements, yet yield good stereopsis. When Fig. 2 is binocularly viewed, the thin stripes cannot individually be resolved but are perceived as a transparent textured plane in front of a solid textured background. In the fused image the symmetry is strongly suppressed, particularly when the non-dominant eye views the symmetric image. On the other hand, the reduced disparity makes it easier to find a few micropatterns in the two depth planes by scrutinizing the stimulus.

It should be possible to vary the stripe width in such a way that the upper half field contains combinations different from those of the lower half field, thus further reducing similar micropattern pairs. Nevertheless, instead of further attempts with horizontal bilateral symmetry, we turn our attention to vertical bilateral symmetry.

V. VERTICAL BILATERAL SYMMETRY

In these experiments the horizontal bilateral symmetry had two disadvantages over the vertical case. The first disadvantage has been already discussed; it was pointed out that horizontal symmetry is less perceivable than vertical symmetry. The second disadvantage will now be discussed. In Figs. 1 and 2 the left and right images contained the *same* picture elements except for a few picture elements that were uncorrelated in the small areas affected by the horizontal shifts. On the other hand, it is known from the random-dot stereoscopic image technique that areas presented to only one eye's view are perceived as continuations of the depth plane furthest behind.[2,3]

This perceptual response permits us to reduce further the similarity between the left and right image pairs. For instance, suppose we consider a left image with vertical bilateral symmetry. We now subdivide it into twenty *vertical* stripes of five picture-element width. The right image is identical to the left one, except each even vertical stripe is shifted to the left by two picture elements, as though it were a solid sheet. Because of the shift, bars two picture-elements wide along the left sides of the shifted stripes will be hidden by the dots which belong to the vertical stripes. The right sides of the shifted vertical strips will uncover new areas, which will be filled in with new random dots. Thus, every shifted even stripe will hide a bar (which belongs to the surround) of two picture-elements width (out of the five), and the odd stripes (which belong to the surround) of three picture-element width will be extended by a newly generated bar of two picture-elements width. Such a stereoscopic image is shown in Fig. 3, in which, in addi-

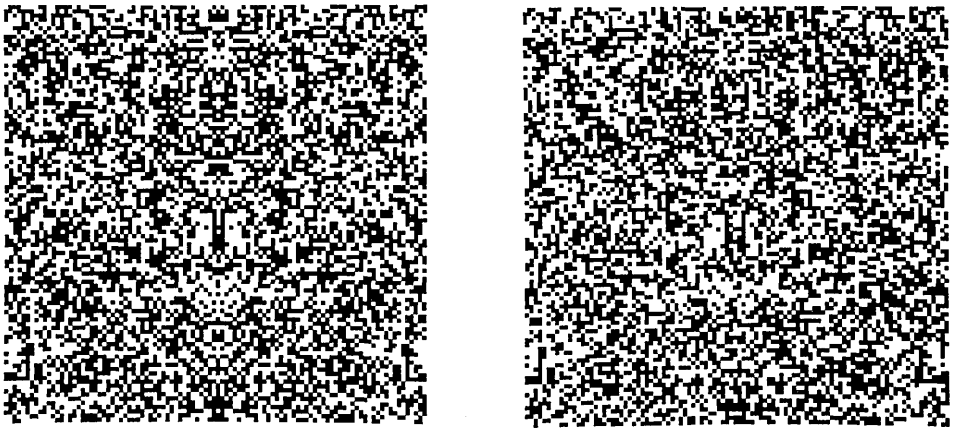


Fig. 3 — Stereogram which, when monocularly viewed, contains an image of bilateral symmetry across the vertical axis. When viewed stereoscopically, vertical stripes are perceived in depth, and the symmetry is suppressed.

tion to the shearing and shift, 2/5 = 40 percent of the dots are different in the right image as compared to the left. When the left image of Fig. 3 is monocularly viewed, the vertical bilateral symmetry is immediately apparent; when the images are binocularly viewed, vertical stripes are seen in front of a background, and symmetry is suppressed. The only cue for symmetry is the presence of some large black and white clusters which can be paired in the left and right half fields.

Fig. 4 is similar to Fig. 3, except that the same cluster breaking operation used in Fig. 2 was applied. Here the binocular percept is quite free from symmetry, except for a few diagonal checkerboard-like patterns which might be found under scrutiny. Anyway, there is no doubt that further preprocessing could greatly reduce these few remaining recognizable structures. It is also true that the bilateral symmetry is less apparent in the left image of Fig. 4 than before cluster breaking. It is a delicate operation to find the amount of cluster breaking which still gives a strong percept monocularly but which causes the binocular percept to disappear.

VI. TWO- AND FOUR-FOLD SYMMETRY

In a next experiment, a left image with two- and four-fold symmetry, respectively, was generated and a compatible right stereo image similar to Fig. 4 was devised. The only difference is that the vertical

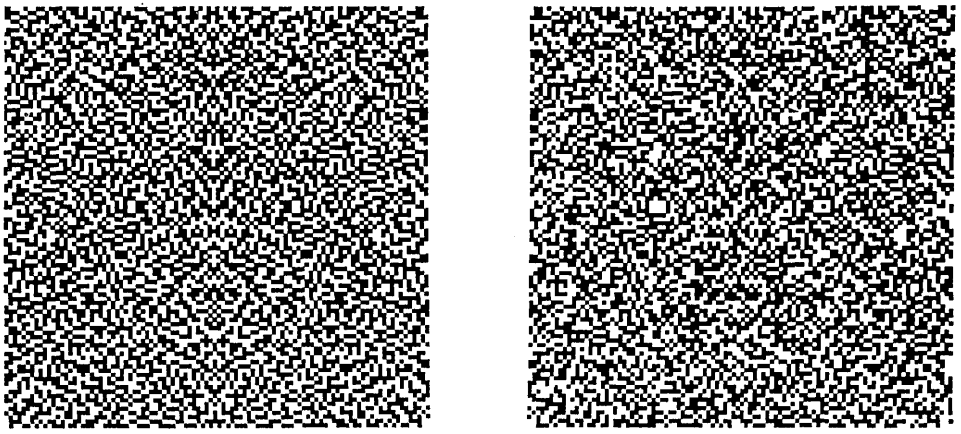


Fig. 4 — Stereogram which is similar to Fig. 3, except that clusters are broken up by preprocessing.

five picture-element wide alternate stripes are shifted in phase in the lower half field compared to the upper half field. Whereas, in the upper half field every even vertical stripe is perceived above the background, in the lower half field every odd vertical stripe is in front. Such stereo images are presented in Figs. 5 and 6 where the former contains a

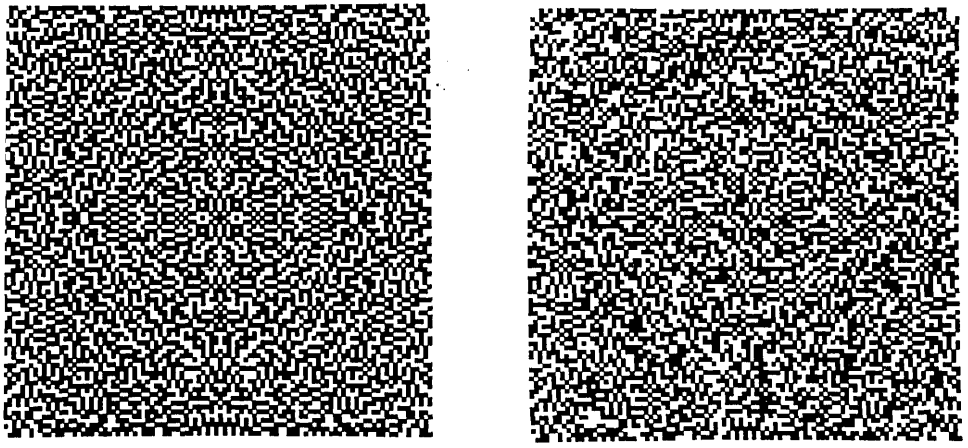


Fig. 5 — Stereogram which, when monocularly viewed, contains an image of two-fold symmetry. When viewed stereoscopically, vertical stripes are perceived in depth which are shifted in position in the lower half field with respect to the stripes in the upper half field. The symmetry is suppressed in the binocular view.

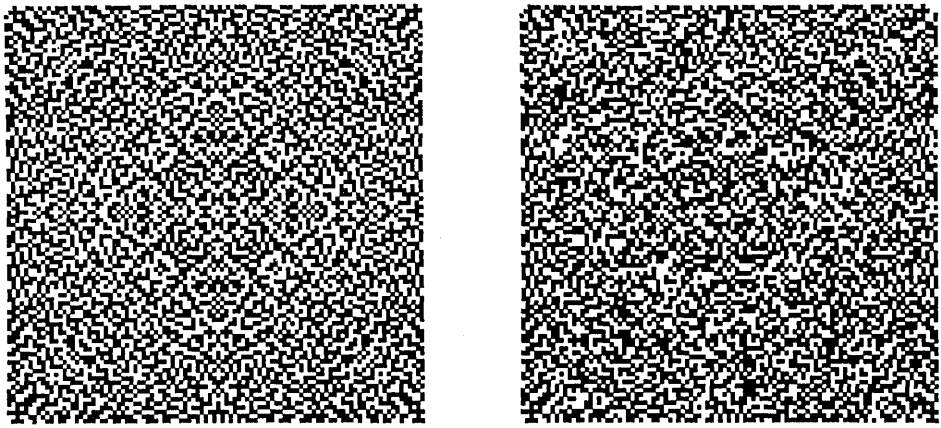


Fig. 6 — Stereogram which, when monocularly viewed, contains an image of octal symmetry. Otherwise similar to Fig. 5.

two-fold symmetry, whereas the latter a four-fold symmetry. The middle phase shift prevents the perception of the bilateral symmetry across the horizontal axis, while the alternate vertical stripes in depth suppress the symmetry across the vertical axis. The suppression of binocular symmetry is perhaps less striking for these cases than for the one-fold symmetry in Fig. 4; on the other hand, the two-fold and particularly the octal (four-fold) symmetry in the left field is so strong that the difference between the monocular and binocular perception of symmetry is very pronounced.

There are many other manipulations one could successfully apply in order to break up monocular global percepts. Instead of long stripes, rectangles can be selected at various depth levels and various widths. The horizontal and vertical stripes at various depth levels can be intermixed. The illustrations have served only to show the flexibility of the random-dot stereoscopic image technique.

VII. SOME CUES OF SYMMETRY PERCEPTION

It was already discussed how clusters affect symmetry perception and how much stronger bilateral symmetry is perceived across a vertical axis than across a horizontal one. Now, we are in a position to examine some of these problems in more detail. Since the images in these experiments are devoid of all familiarity cues, this preference to the vertical is characteristic of human perception and is much more

prominent than one would expect from psycho-physiological findings on the detection of horizontal-vertical lines.[4,5] It is also interesting that in two- and four-fold symmetry perception—which are more strongly perceived than bilateral vertical symmetry—the strength of perceived symmetry across the horizontal axis is greatly increased.

In the preceding experiments we restricted ourselves to mirror symmetries. For the horizontal bilateral symmetry $F(x, y) = F(x, -y)$, and for the vertical bilateral symmetry $F(x, y) = F(-x, y)$ has to be satisfied. For the four-fold symmetry in addition to both of these equations, $F(x, y) = F(-x, -y)$ has to be satisfied as well. Besides mirror symmetries one can study the perception of centric symmetries, such as shown in the right image of Fig. 7. For this case $F(x, y) = F(-x, -y)$; $F(x, y) \neq F(-x, y)$ and $F(x, y) \neq F(x, -y)$. This image was derived from the octal-symmetrical left image by alternately shifting every even horizontal stripe (of two picture-element width) to the right and every odd one to the left. The monocularly apparent octal symmetry is very strong in the left image, while the centric symmetry is hard to perceive in the right one. Nevertheless, some large clusters in the corners can be detected in both images, which are also recognizable in the stereoscopic view. It is interesting to note that these macroclusters are easily detected and matched in spite of their differences in fine details.

Fig. 8 is identical to Fig. 7 except for the clusters, which are broken

Fig. 7 — Stereogram which, in the left image contains octal (four-fold) symmetry and in the right image contains centric symmetry.

up by preprocessing. This experiment shows again that the larger the clusters are, the less similar they have to be in detail to be detected and matched at symmetrical locations. The centric symmetry in Fig. 8 is even harder to perceive than in Fig. 7. The symmetry in the binocular view is also greatly reduced, yet less so than in Fig. 6, since the left and right images of Fig. 8 are 100 percent correlated. In Figs. 2, 4, and 8, in which the macroclusters have been broken up in the left images, the microclusters have to be identical in their minutest details to be perceived as symmetrical. This is shown in Fig. 9, in which the left image was derived from a random-dot pattern with four-fold symmetry and then preprocessed by the cluster-cleaner. Obviously, the quadrants of the picture are not identical in their fine detail (since local changes by the cluster-cleaner affect the successive complementations) but are similar in their larger features. The right image of Fig. 9 was obtained by taking a quadrant of the left image and reflecting it across the horizontal and vertical axes sequentially. When the images are viewed at a short distance, they appear quite different; one looks symmetrical, the other not. When viewed from a distance, the two images appear identical and symmetrical, since only the similar macroclusters are perceivable.

This observation, that symmetry perception depends on cluster size, which in turn depends on texture density explains some of the difficulties with these demonstrations as they are presented in this article.

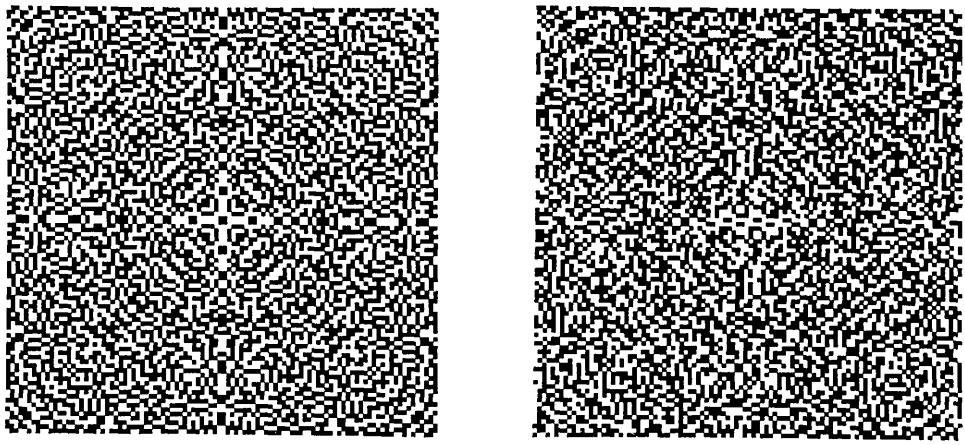


Fig. 8 — Stereogram which is similar to Fig. 7, except the macroclusters are removed by preprocessing.

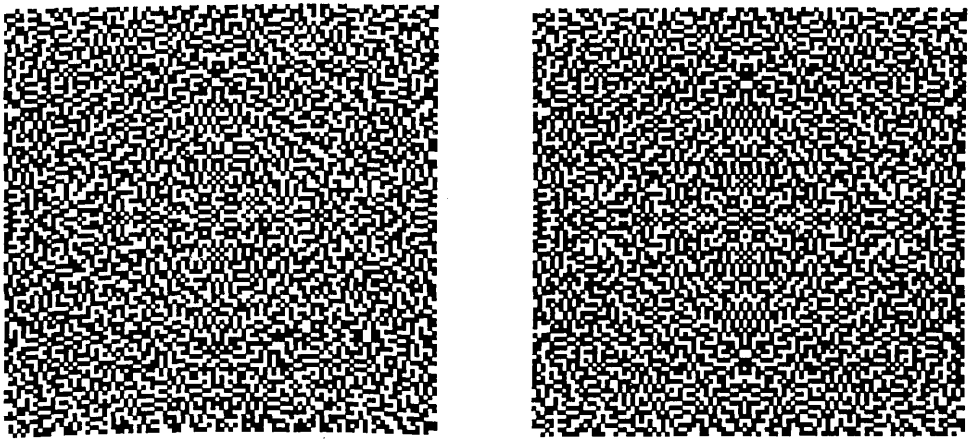


Fig. 9 — Two images which are similar in their macrostructure, but differ in their microstructure. When viewed at a short distance only the right image appears symmetrical, while from an increased viewing distance both appear symmetrical.

The figures as shown are already too small at the best viewing distance of 10 inches. If the images are made larger (in excess of 30 degree visual angle) the stereoscopically fusable local features will dominate stronger over the macroclusters.

Another observation, closely related to the previous ones, suggests a relationship between average cluster size and distance from the symmetry axis. Clusters on both sides of a symmetry axis can be detected as being symmetrical only within a certain distance from the axis, which is commensurate with the cluster size. Thus, small clusters must lie close to the symmetry axis, whereas larger ones may lie proportionally farther. This explains why the left image of Fig. 9 deviates from symmetry when viewed at a short distance. For large stimuli the large clusters cannot be perceived effortlessly, thus the nonsymmetrical microclusters dominate. When the stimulus size is reduced, the situation reverses; it is now the symmetrical macroclusters which can be effortlessly perceived, and differences in microclusters pass unnoticed.

Ernst Mach a century ago studied symmetry perception by using amorphic shapes.[6] His findings are similar to the results obtained by using random textures as long as large clusters of dark and light areas are contained in these textures. According to these findings vertical bilateral symmetry and centric symmetry can be spontaneously per-

ceived. Horizontal bilateral symmetry is less well perceived. When the macroclusters are broken up centric symmetry does not yield spontaneous perception, whereas the perception of bilateral symmetry is not impaired.

VIII. SUPPRESSION OF MONOCULAR PERCEPTS IN GENERAL

Until now the monocular percepts to be suppressed binocularly were restricted to global symmetries. Needless to say, many other monocular global percepts can be destroyed binocularly. Since symmetry is easy to generate, it was emphasized over other perceivable forms. For instance, the left random field could contain a written text of black or white letters that could be perceived and read with one eye. The right image would be totally correlated with the left one, except rows or columns would be shifted horizontally in some random way. In the resulting binocular percept, the letters would be scrambled both because of the horizontal shifts and because of these shifts causing various letter-segments to be perceived at different depth. This could render the binocular percept unreadable. While in symmetry perception there is always a possibility that various amounts of symmetry can be traced, for this example a dichotomy can be obtained. Either the perceived image contains a readable text, or not.

One might argue, that it is not surprising that a text in one channel disappears when some "noise" is added to it through the other channel. This argument seems convincing, but actually the opposite is true. This random pattern is not uncorrelated "noise" which masks the text in the other display, but a totally correlated pattern which gives rise to stereopsis. It is an interesting paradox that when the two images are uncorrelated, binocular rivalry occurs. One might expect for this case the largest masking of the text by the competing uncorrelated noise; the opposite is in fact the case, since during binocular rivalry the text is quite often visible as dominance alternates. For the type of stimuli demonstrated here, the local fusion is perfect and stable, therefore, if the text is masked in the binocular view, it stays so indefinitely.

Fig. 10 shows such a stereoscopic image. The word "YES" is inserted in the random texture of the left image by a computer program. The right image is derived from the left image by alternately shifting every even and odd horizontal stripe of two picture-element width to the left or right direction respectively. When this stereoscopic pair is presented binocularly to subjects—who have not seen the images monocularly yet—the impression is a transparent textured surface above a textured

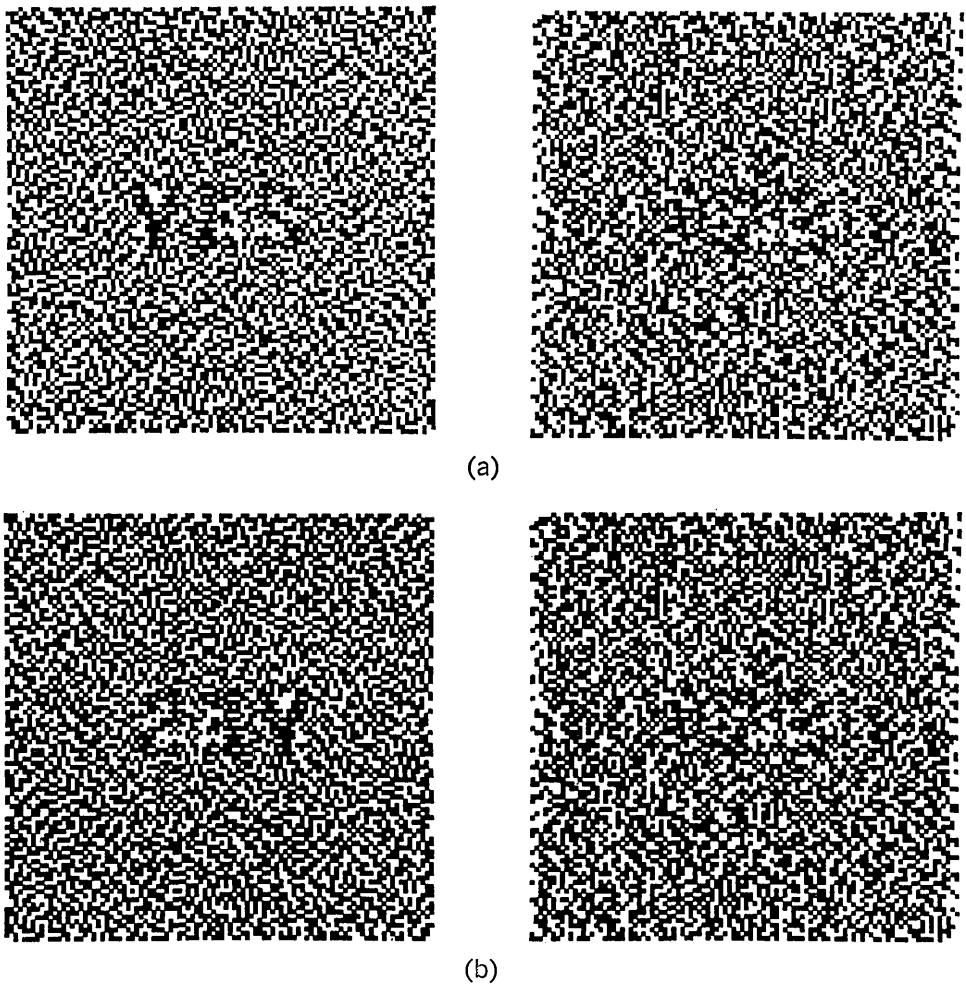


Fig. 10 — (a) Stereogram which, when monocularly viewed contains the word "YES." When viewed stereoscopically, the letters become fragmented, rendering the text unreadable. (b) Stereogram identical to (a) except that the left image is mirror reflected to permit stereoscopic viewing with the supplied front-surface mirror as described in the appendix.

background. All the eight subjects perceived both surfaces as randomly textured without being able to find and read the disguised word when instructed to search for it. After being instructed to close the right eye, each of the subjects recognized the word "YES" in the left image. The uniformly dark letters were intentionally sprinkled with random white

dots in order to aid local fusion. If the macroclusters out of which the letters are composed are not broken up, it is possible to detect them in the random texture of finer grain. This enables the subjects to shift the entire letters in the left field in registration with the unshifted segments of the same letters in the right image. In Fig. 10 the letter "Y" is more robust than the rest, and after it has been detected monocularly, some of the subjects could shift the entire letter in registration when binocularly viewed. Of course, by covering this letter with more random texture, it can be made to disappear in the binocular view as the rest.

These observations suggest a cluster processor prior to stereopsis. If clusters of a certain granularity exist in a finer or coarser textured surround, they might be extracted and separately fused from the rest. When the clusters in the stereoscopic images have the same average clustering, this preprocessing becomes inoperative and stereopsis occurs on a point-by-point basis.

There are many other examples in which monocular percepts are suppressed in the binocular view. In Refs. 2 and 7 several random-dot stereoscopic images have been presented in which one of the images was perturbed, yet this monocularly apparent perturbation was suppressed binocularly. For instance, when one image of the random-dot stereoscopic pair is blurred, binocular depth can be still experienced and the binocular percept appears as the sharp image.[2] This suppression of the contour-impoverished image by the contour-rich image seems to be a general property of binocular combination and was recently studied by Levelt.[8] Although, these phenomena corroborate the basic theme of this article, they differ from the examples given here, since they are a special case of binocular rivalry. According to Levelt, the greater the difference between contour content in a stereoscopic pair the more the contour-rich channel is weighted. On the other hand, in our experiments the stereoscopic pairs are weighted equally, and binocular rivalry never occurs—only perfect point-by-point fusion.

IX. CONCLUSIONS

In 1960 a perceptual phenomenon was reported in this Journal which demonstrated that binocular shapes can be perceived from monocularly shapeless and contourless, random-dot stereoscopic images.[2] The finding that correlated areas in the left and right images could give rise to binocular depth perception, regardless of the fact that these areas

were completely disguised when viewed by one eye, has several theoretical and practical implications. An obvious implication is that no monocular global percept (form recognition) is necessary for stereopsis. Recently, a reversal of this phenomenon was demonstrated by the author.[1] A stereoscopic image was devised in which the monocularly apparent shapes of bilateral symmetry across a horizontal axis disappeared when stereoscopically viewed. This phenomenon sharpens the implication of the earlier one. It suggests that whenever binocular combination occurs, this process precedes or dominates the recognition of horizontal bilateral symmetry.

The demonstration of Ref. 1 had a few inadequacies which were overcome in the experiments presented here. In order to improve on the weakly perceived bilateral symmetry across a horizontal axis, ways were found in which *vertical* bilateral symmetry could be presented monocularly without affecting symmetry suppression in the binocular view. Similar binocular suppression was obtained for monocularly apparent two- and four-fold symmetries. Micro- and macro-clusters, which are always present in random textures, were broken up in the stimulus and the results of this preprocessing on symmetry perception were studied.

An elaboration on these experiments substituted monocularly apparent symmetry with monocularly readable text to be incorporated in the stereoscopic images. In the binocular view the letters of the text were perceived as being scrambled and the letter fragments were perceived at various depths rendering the text unreadable.

X. ACKNOWLEDGMENTS

I want to thank Dr. E. C. Carterette for his interest and comments on Ref. 1, Dr. J. Krauskopf for his commenting on the manuscript, and Mr. R. A. Payne for his help in preparing the photographs.

APPENDIX

A Method of Viewing Stereograms

Figs. 1(b), 2(b), 3 through 9, and 10(b) may be viewed stereoscopically with the aid of the front-surface mirror supplied inside the back cover of this issue. Ordinary (back surface, glass) mirrors cannot be used. The mirror is made of polystyrene and is quite fragile—carefully remove the mirror, taking care not to scratch or mar the surface. For best results, the mirror should be flattened by fastening it to a piece of stiff, flat cardboard by means of transparent tape laid along the edges.

Fig. 11 — The author demonstrating a method of viewing stereograms.

Fig. 11 shows a method of viewing the stereograms. To obtain stereopsis, hold the mirror facing toward the left image so that the left eye views the image reflected in the mirror while the right eye gazes directly at the right image, as shown in Fig. 11. The mirror should be nearly perpendicular to the image plane. The reflected image will appear to float over the directly-viewed image. Adjust your position (or the mirror) until the reflected image and directly-viewed image are superimposed and the reflected image is the same size as the directly-viewed image. At this point the images fuse and give the stereoscopic effect.

REFERENCES

1. Julesz, B., Binocular Disappearance of Monocular Symmetry, Sci., *153,* 1966, pp. 657–658.

2. Julesz, B., Binocular Depth Perception of Computer-Generated Patterns, B.S.T.J., *39*, September, 1960, pp. 1125–1162.
3. Julesz, B., Binocular Depth Perception without Familiarity Cues, *Sci., 145,* 1964, pp. 356–362.
4. Hubel, D. H. and Wiesel, T. N., Shape and Arrangement of Columns in Cat's Striate Cortex, J. Physiol., *165,* 1963, pp. 559–568.
5. Campbell, F. W., Kulikowski, J. J., and Levinson, J., The Effect of Orientation on the Visual Resolution of Gratings, J. Physiol., *187,* 1966, pp. 427–436.
6. Mach, E., *The Analysis of Sensations* (1897), republished by Dover, New York, 1959.
7. Julesz, B., Stereopsis and Binocular Rivalry of Contours, J. Opt. Soc. Am., *53,* 1963, pp. 994–999.
8. Levelt, W. J. M., On Binocular Rivalry, Institute for Perception, Soesterberg, Netherlands, 1965.

Large-Signal Calculations for the Overdriven Varactor Upper-Sideband Upconverter Operating at Maximum Power Output

By J. W. GEWARTOWSKI and R. H. MINETTI

(Manuscript received March 6, 1967)

When a varactor frequency upconverter is used as the output device in a communications transmitter, it is often desirable to operate at maximum power output. For such operation, the design procedure includes a large-signal analysis under overdriven conditions, requiring the use of a computer for solution.

Equations for the instantaneous varactor charge, current, and voltage are derived assuming that only three currents are present: those corresponding to the signal, pump, and output frequencies. Numerical solutions corresponding to maximum power output are obtained for both graded-junction and abrupt-junction varactors. Values of power output, conversion efficiency, input impedances, load impedance, and bias voltage are presented for ranges of drive level (1 to 2) and varactor quality ($0.001 \leqq \omega_3/\omega_c \leqq 0.1$) sufficient to include most practical designs.

I. INTRODUCTION

Varactor upconverters are finding increased application in solid-state microwave transmitters. In this application a frequency modulated IF signal is mixed with an unmodulated microwave signal to obtain a frequency modulated output signal.[1] Because of the varactor's low loss and high power handling capacity, this output signal can be used as the transmitter output signal without further amplification. Microwave conversion efficiencies in a varactor upconverter are typically greater than 50 percent.[2]

A varactor upconverter can be either an upper-sideband or a lower-sideband upconverter. However, since the lower-sideband upconverter can present stability problems,[3] the upper-sideband upconverter is generally preferred for the above application.

This paper presents a general analysis of lossy varactor upper-sideband upconverters, using both graded- and abrupt-junction varactors. Results will be presented for operation at maximum power output for a prescribed drive level. The two input signals and the output signal are all three "large" signals in this mode of operation. Penfield and Rafuse have presented a theory for nonoverdriven abrupt-junction varactors;[3] however, their analysis gives a conservative value of output power, as explained by Nelson.[4] Nelson presents an improved upper bound on the varactor charge coefficients, which results in 25 to 54 percent greater output power than that given by Penfield and Rafuse. This paper will use Nelson's method of computing this upper bound, with a slight modification to allow for arbitrary output phase angles. Grayzel has presented a similar analysis for "punch through" varactors.[5] His analysis differs from ours in that he assumes a particular phase condition for the output current [equivalent to taking $\alpha = 0$ in our (5)]. Our results show that as much as 16 percent greater power output is obtained when α is optimized.

II. ANALYSIS

2.1 *Model and Assumptions*

The varactor model chosen for the analysis is the usual one consisting of a constant resistance R_s in series with a variable capacitance, as shown in Fig. 1. A polarity is assumed such that when the varactor is reverse biased, the voltage and charge stored are positive.

For voltages between the barrier potential and the breakdown voltage, the voltage and stored charge on the variable capacitance are related by

$$\frac{v_j - \Phi}{V_B - \Phi} = \left(\frac{q - q_\Phi}{Q_B - q_\Phi}\right)^{1/(1-\gamma)} \tag{1}$$

for

$$\Phi \leqq v_j \leqq V_B$$

$$q_\Phi \leqq q \leqq Q_B ,$$

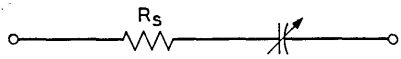


Fig. 1 — Varactor model consisting of a constant resistance in series with a time-varying capacitance.

where

v_j = voltage across the capacitance
Φ = barrier potential (negative)
V_B = breakdown voltage (positive)
q = charge on the capacitance
Q_B = charge at breakdown voltage (positive)
q_Φ = charge at the barrier potential (negative).

When the varactor is overdriven the voltage is assumed to clamp to the barrier potential

$$v_j = \Phi \quad \text{for} \quad q \leqq q_\Phi \,. \tag{2}$$

In this region the charge-storage effect is assumed to act like an infinite capacitance, so that any amount of charge can be stored without any additional voltage drop. This model is a good approximation for microwave varactors, where the minority carrier lifetime is considerably longer than an RF period. Experimental evidence has shown that this is still a reasonable approximation at frequencies as low as 40 MHz, a typical IF frequency.

It is necessary to have a parameter which measures the extent to which the varactor is overdriven. The *drive* is defined by[6]

$$drive = \frac{Q_B - q_{\min}}{Q_B - q_\Phi} \,, \tag{3}$$

where it is assumed that the varactor is always driven up to breakdown. Thus, $drive = 1$ corresponds to the "fully driven" case of Penfield and Rafuse.[3] Most practical high-power varactor devices operate overdriven, so that $drive > 1$.

Another useful parameter is the varactor cutoff frequency, given by[3]

$$f_c = \frac{\omega_c}{2\pi} = \frac{S_{\max}}{2\pi R_s} = \frac{V_B - \Phi}{2\pi R_s(1 - \gamma)(Q_B - q_\Phi)} \,, \tag{4}$$

where $S_{\max}$ is the elastance of the diode junction at the breakdown voltage. The last equivalence is obtained from (1) and the relationship $S = \partial v/\partial q$.

It will be assumed that only three currents are present in the varactor, those corresponding to the signal, pump, and output frequencies. When the pump and output frequencies are appreciably different, currents at other than the three frequencies mentioned are impeded by the selectivity characteristics of the circuits. However, if the pump and output frequencies are close together, then it is difficult to inhibit

currents at other sidebands. The effect of these sidebands on the results presented here has not been evaluated.

Circuit losses are not included explicitly. They may be accounted for by increasing the value of R_s or by calculating separate input and output circuit efficiencies, or by a combination of these two approaches.

2.2 *Equations*

The charge stored on the variable capacitance may be written as

$$q = Q_0 + 2Q_1 \sin \omega_1 t + 2Q_2 \sin \omega_2 t + 2Q_3 \sin (\omega_3 t + \alpha), \tag{5}$$

where ω_1 and ω_2 correspond to the two input frequencies, $\omega_3 = \omega_1 + \omega_2$ corresponds to the output frequency, and ω_2 is taken to be greater than ω_1. Since the two input signals at ω_1 and ω_2 are independent, their phases may be taken arbitrarily as shown. The output phase angle α, on the other hand, must be chosen to correspond to maximum power output at a prescribed value of *drive*.

The instantaneous current is obtained from (5) as

$$i = 2\omega_1 Q_1 \cos \omega_1 t + 2\omega_2 Q_2 \cos \omega_2 t + 2\omega_3 Q_3 \cos (\omega_3 t + \alpha). \tag{6}$$

The total voltage v on the varactor is given by

$$v = v_j + R_s i. \tag{7}$$

The following quantities may be obtained from (5), (6), and (7). The numerical integrations are performed using (1) and (2) for values of v_j.

Input resistance at ω_1 :

$$R_1 = R_s + \frac{1}{\omega_1 Q_1 T} \int_0^T v_j \cos \omega_1 t \, dt. \tag{8}$$

Input resistance at ω_2 :

$$R_2 = R_s + \frac{1}{\omega_2 Q_2 T} \int_0^T v_j \cos \omega_2 t \, dt. \tag{9}$$

Load resistance at ω_3 :

$$R_3 = -R_s - \frac{1}{\omega_3 Q_3 T} \int_0^T v_j \cos (\omega_3 t + \alpha) \, dt. \tag{10}$$

Input elastance at ω_1 :

$$S_1 = \frac{1}{Q_1 T} \int_0^T v_j \sin \omega_1 t \, dt. \tag{11}$$

Input elastance at ω_2 :

$$S_2 = \frac{1}{Q_2 T} \int_0^T v_i \sin \omega_2 t \, dt. \tag{12}$$

Output elastance at ω_3 :

$$S_3 = \frac{1}{Q_3 T} \int_0^T v_i \sin (\omega_3 t + \alpha) \, dt. \tag{13}$$

In general, the integration interval T must be large enough to obtain the desired degree of accuracy. To facilitate computation, ω_2 and ω_3 are selected to the nth and $(n + 1)$th integral multiples of ω_1 . The integration interval is then equal to $T = 2\pi/\omega_1$. Results for nonintegrally related frequencies can be obtained from these results by interpolation.

The resistances calculated above allow one to compute the powers at the three frequencies, using values of current from (6).

Input power at ω_1 :

$$P_1 = 2\omega_1^2 Q_1^2 R_1 \, . \tag{14}$$

Input power at ω_2 :

$$P_2 = 2\omega_2^2 Q_2^2 R_2 \, . \tag{15}$$

Output power at ω_3 :

$$P_3 = 2\omega_3^2 Q_3^2 R_3 \, . \tag{16}$$

In most applications ω_2 and ω_3 correspond to microwave frequencies, and ω_1 corresponds to a lower frequency. Thus, we define the microwave conversion efficiency by

$$\eta_{23} = \frac{P_3}{P_2} = \frac{\omega_3^2}{\omega_2^2} \frac{Q_3^2}{Q_2^2} \frac{R_3}{R_2}. \tag{17}$$

Another useful expression for this quantity is derived in the Appendix as

$$\eta_{23} = \frac{\omega_3}{\omega_2} \frac{R_2 - R_s}{R_2} \frac{R_3}{R_3 + R_s}. \tag{18}$$

The upconversion gain is defined by

$$G_{13} = \frac{P_3}{P_1} = \frac{\omega_3^2}{\omega_1^2} \frac{Q_3^2}{Q_1^2} \frac{R_3}{R_1}. \tag{19}$$

Another useful expression for this quantity is derived in the Appendix as

$$G_{13} = \frac{\omega_3}{\omega_1} \frac{R_1 - R_s}{R_1} \frac{R_3}{R_3 + R_s}. \tag{20}$$

Equation (18) and (20) are easily seen to correspond to the results given by the Manley-Rowe[7] relations as $R_s \to 0$.

One other important parameter is the bias voltage, computed numerically from the expression

$$V_0 = \frac{1}{T} \int_0^T v_i \, dt. \tag{21}$$

2.3 *Selection of the Charge Coefficients and Output Phase Angle*

The output power P_3 as given by (16) depends on Q_3 and R_3. R_3 in turn is a function of Q_0, Q_1, Q_2, Q_3, and other parameters. These charge coefficients are selected to give maximum output power at a prescribed value of *drive*. Obviously, if the drive level were not restricted, the output power could increase without limit, since overdrive of any magnitude is allowed by the theoretical model chosen for the varactor.

The instantaneous charge is given by (5). At the onset, five independent variables are unspecified, Q_0, Q_1, Q_2, Q_3, and α. Two of these variables may be eliminated by using relationships bounding the maximum and minimum instantaneous charge. The maximum charge is given by

$$q_{\max} = Q_B \tag{22}$$

and the minimum charge is obtained from the prescribed drive level using (3). Q_B and q_Φ are, of course, known values for a given varactor.

After applying these limits to eliminate two of the variables, the remaining three variables are varied to find numerically the values corresponding to maximum output power. In our numerical process Q_0 and Q_1 are eliminated, and Q_2, Q_3, and α remain as the independent variables. Note that once these quantities are selected, all other upconverter operating parameters can be computed using the equations in Section 2.2 together with a knowledge of the varactor characteristics.

The values of $q_{\max}$ and $q_{\min}$ discussed above must be selected with some care. One method of selection would be to take maximum and minimum values directly from (5). However, this method is inclined to give different results when the frequencies are harmonically related

than when they are not. To show this, let us compare two instantaneous charge waveforms of the type of (5), both having the same charge coefficients, but the first waveform having harmonically related frequencies and the second not. The first waveform will repeat its pattern with a period $2\pi/\omega_1$, whereas the second waveform will never repeat. Evidently the second waveform will generally have a larger peak-to-peak amplitude, since it is more "random," i.e., it has a larger assortment of peaks and valleys. This is especially true when ω_2 is not much larger than ω_1.

The values of $q_{\max}$ and $q_{\min}$ are computed using Nelson's approximate method.[4] For ω_2/ω_1 greater than 2 or 3, one may visualize the ω_2 and ω_3 terms of the charge waveform as being a high-frequency signal, amplitude modulated at ω_1. Thus, one may write (5) as

$$q = Q_0 + 2Q_1 \sin \omega_1 t + R \sin (\omega_2 t + \theta), \tag{23}$$

where

$$R = 2\sqrt{Q_2^2 + Q_3^2 + 2Q_2Q_3 \cos (\omega_1 t + \alpha)}$$

$$\theta = \tan^{-1}\left[\frac{Q_3 \sin (\omega_1 t + \alpha)}{Q_2 + Q_3 \cos (\omega_1 t + \alpha)}\right].$$

Let $\omega_1 t_{\max}$ and $\omega_1 t_{\min}$ be the instants corresponding to the maximum and the minimum values of the functions

$$Q_1 \sin \omega_1 t \pm \sqrt{Q_2^2 + Q_3^2 + 2Q_2Q_3 \cos (\omega_1 t + \alpha)}.$$

Then, the maximum and minimum values of instantaneous stored charge are given approximately by

$$q_{\max} = Q_0 + 2Q_1 \sin \omega_1 t_{\max} + 2\sqrt{Q_2^2 + Q_3^2 + 2Q_2Q_3 \cos (\omega_1 t_{\max} + \alpha)} \tag{24}$$

and

$$q_{\min} = Q_0 + 2Q_1 \sin \omega_1 t_{\min} - 2\sqrt{Q_2^2 + Q_3^2 + 2Q_2Q_3 \cos (\omega_1 t_{\min} + \alpha)}. \tag{25}$$

Equations (3), (22), (24), and (25) allow one to compute Q_0 and Q_1, given the values of Q_2, Q_3, and α for a given varactor and drive level.

2.4 *Calculation Procedure*

The parameters corresponding to maximum output power were obtained from a digital computer using a computer program based upon

the simplified flow chart of Fig. 2. Trial values of α were selected at 1 degree intervals, and trial values of Q_2 and Q_3 were selected at intervals of 0.0005 ($Q_B - q_\Phi$). These values were found to give results accurate to within plotting accuracy.

III. RESULTS

As indicated in Fig. 2, the results are a function of four variables, γ, *drive*, ω_2/ω_3, and ω_3/ω_c. γ determines the varactor type, graded junction ($\gamma = \frac{1}{3}$) or abrupt junction ($\gamma = \frac{1}{2}$).

The parameter *drive* is defined by (3). There is no theoretical upper limit for this quantity for the varactor model chosen; we shall arbitrarily take 2 as the upper limit for our calculation. In a practical varactor the drive level will be limited by forward bias current due to the finite minority carrier lifetime; however, values of 2 are usually attainable.

The parameter ω_3/ω_c expresses the effect of the loss R_s for a practical varactor.

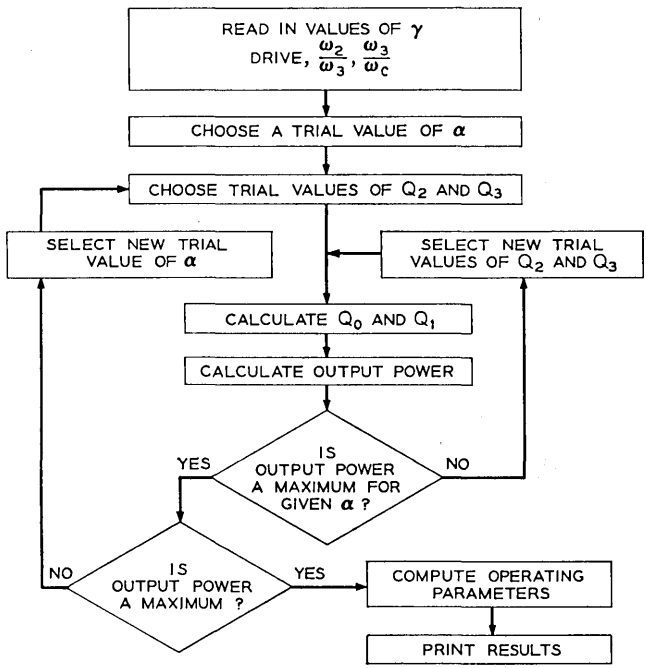


Fig. 2 — Simplified flowchart of the computer program.

The frequency ratio ω_2/ω_3 determines the ratios of the three frequencies, since $\omega_3 = \omega_1 + \omega_2$. Fortunately, it has been determined that the results can be normalized in such a way as to make them independent of this frequency ratio in most practical cases. These normalized quantities are as follows:

normalized resistances

$$\frac{R_1\omega_1}{S_{\max}}, \qquad \frac{R_2\omega_2}{S_{\max}}, \qquad \frac{R_3\omega_3}{S_{\max}};$$

normalized elastances

$$\frac{S_1}{S_{\max}}, \qquad \frac{S_2}{S_{\max}}, \qquad \frac{S_3}{S_{\max}};$$

normalized powers

$$\frac{P_1 S_{\max}}{(V_B - \Phi)^2\omega_1}, \qquad \frac{P_2 S_{\max}}{(V_B - \Phi)^2\omega_2}, \qquad \frac{P_3 S_{\max}}{(V_B - \Phi)^2\omega_3};$$

normalized microwave conversion efficiency and unconversion gain

$$\eta_{23}\frac{\omega_2}{\omega_3}, \qquad G_{13}\frac{\omega_1}{\omega_3}; \qquad \text{and}$$

normalized bias voltage

$$\frac{V_0 - \Phi}{V_B - \Phi}.$$

Results were computed for several frequency ratios. These ratios are designated by the harmonics present in the charge waveform, (5). Thus, for example, 1–7–8 denotes the case $\omega_2/\omega_1 = 7$ and $\omega_3/\omega_1 = 8$. Results were computed and compared for the cases 1–2–3, 1–3–4, 1–4–5, 1–5–6, 1–7–8, and 1–86–87. Fig. 3 shows the dependence of normalized efficiency $\eta_{23}\omega_2/\omega_3$ on the frequency ratio, for abrupt- and graded-junction varactors at the two extreme drive levels for $\omega_3/\omega_c = 0.001$. Fig. 4 shows the dependence of the normalized gain $G_{13}\omega_1/\omega_3$. Similar curves are obtained for the other operating parameters for $\omega_3/\omega_c = 0.001$. For $\omega_2/\omega_1 \geqq 5$ the percent variation in normalized resistance is less than 2 percent, and the percent variation in normalized elastance and bias voltage are less than 1 percent. As the loss factor ω_3/ω_c is increased, several of the normalized quantities show invariance as a function of the frequency ratio ω_2/ω_1 similar to that described above for $\omega_3/\omega_c = 0.001$. However, at high values of ω_3/ω_c six of these quanti-

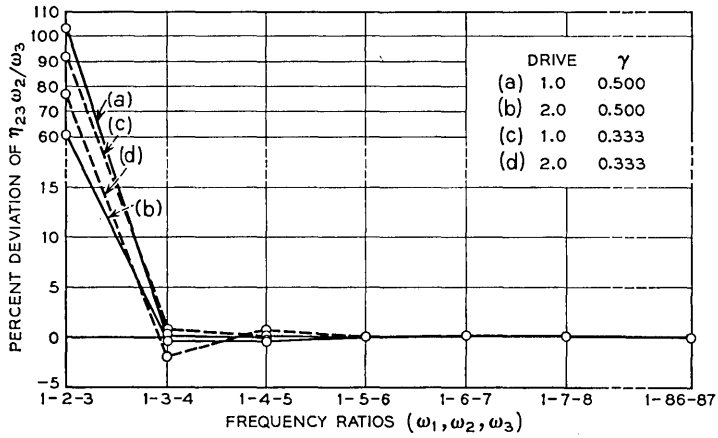


Fig. 3 — Variation of normalized microwave conversion efficiency as a function of the three upconverter frequencies. Percent deviations are calculated with respect to the 1–86–87 case. $\omega_3/\omega_c = 0.001$.

ties are no longer invariant with the ratio ω_2/ω_1. The six quantities which vary with ω_2/ω_1 are listed below (with maximum variation in percent of the 1–7–8 frequency ratio value as computed for $\omega_3/\omega_c = 0.1$):

(*i*) $R_1\omega_1/S_{\max}$ (26 percent)
(*ii*) $R_2\omega_2/S_{\max}$ (10 percent)
(*iii*) $P_1S_{\max}/(V_B - \Phi)^2\omega_1$ (26 percent)
(*iv*) $P_2S_{\max}/(V_B - \Phi)^2\omega_2$ (10 percent)
(*v*) $\eta_{23}\omega_2/\omega_3$ (9 percent)
(*vi*) $G_{13}\omega_1/\omega_3$ (35 percent).

Fortunately, simple correction functions can be applied to obtain the variation with the ratio ω_2/ω_1. Hence, all the results to be presented were computed using the 1–7–8 frequency ratios, and they may be considered to be applicable for $\omega_2/\omega_1 \geqq 5$, except for the six quantities mentioned under high-loss operation, for which simple corrections are provided.

The computed values of α for maximum power output are presented in Fig. 5. The computed values of normalized charge amplitudes for maximum power output are presented in Fig. 6 for the two extreme drive levels. As in the nonoverdriven case,[4] Q_1 and Q_2 are computed to be equal for maximum power output. Using these values and similar

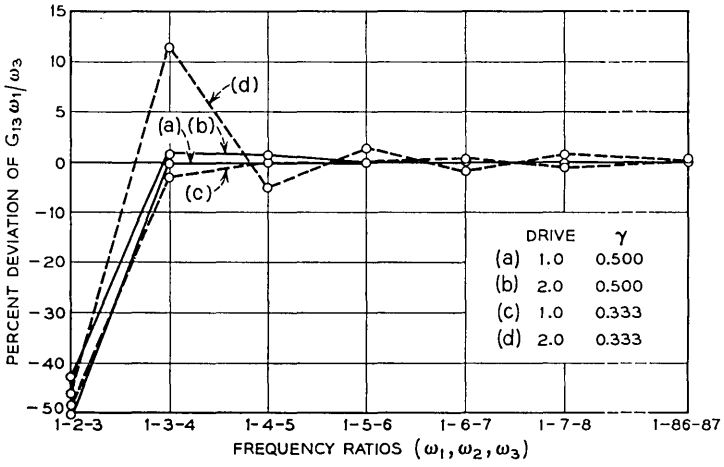


Fig. 4 — Variation of normalized upconversion gain as a function of the three upconverter frequencies. Percent deviations are calculated with respect to the 1-86-87 case. $\omega_3/\omega_c = 0.001$.

results for other drive levels, the operating characteristics of the upconverter are computed and presented below.

Figs. 7 and 8 show values of maximum output power for abrupt- and graded-junction varactors, respectively. It would appear that the abrupt-junction varactor has a considerable advantage in power output. However, if one considers the power-impedance product P_3R_3 (using data from Figs. 17 and 18), the difference is smaller; in fact, for high drive levels the graded-junction varactor has a larger P_3R_3 product.

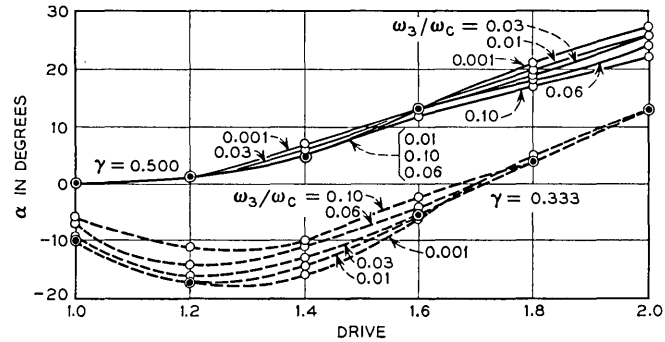


Fig. 5 — Computed values of the output current phase angle α for maximum power output.

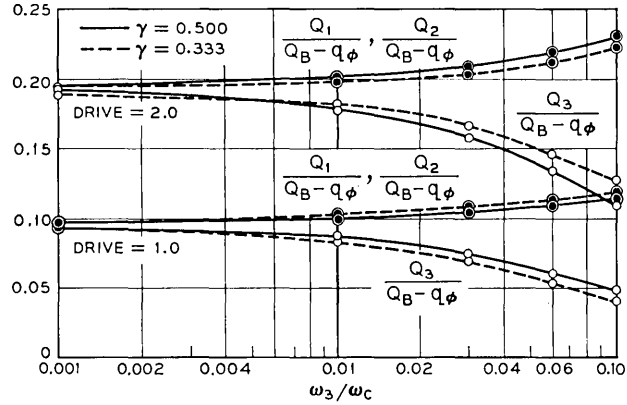


Fig. 6 — Normalized charge amplitude coefficients for maximum power output.

The results for maximum power output at unity drive agree with the results of Nelson,[4] despite the fact that Nelson assumed $\alpha = 0$ for all cases. The effect of a nonzero value of α is most pronounced for highly driven abrupt-junction varactors. For example, Fig. 7 gives a value of $P_3 S_{max}/(V_B - \Phi)^2 \omega_3$ equal to 0.0327 for $\omega_3/\omega_c = 0.001$ and *drive* = 2.0, which is 16 percent larger than the value obtained assuming α equal to zero.

The upconversion gain at maximum power output is presented in Fig. 9, and the microwave conversion efficiency is plotted in Fig. 10. The correction factors for high loss and $\omega_3 \neq 8\omega_1$ are described in the

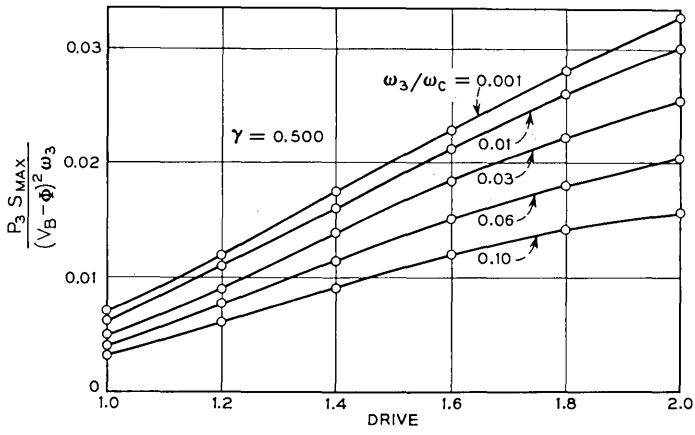


Fig. 7 — Maximum power output for abrupt-junction varactors.

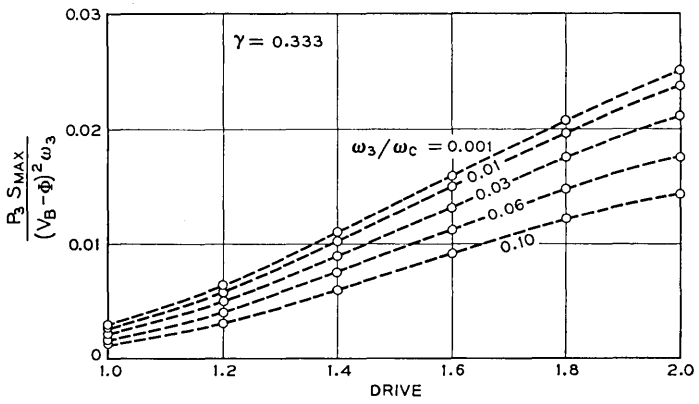


Fig. 8 — Maximum power output for graded junction varactors.

figure captions. For the graded-junction varactor, these curves show a definite advantage in operating at high drive levels. The curves are flatter for the abrupt-junction varactor and show maxima at intermediate drive levels.

Values of input elastance S_1 and S_2 are presented in Fig. 11. These

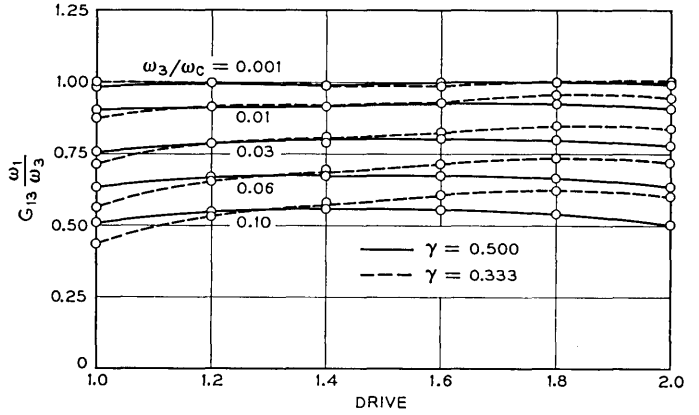


Fig. 9 — Upconversion gain for varactor upconverters operating at maximum power output. These results become inaccurate for high loss (high ω_3/ω_c) and $\omega_3 \neq 8\omega_1$. Accurate values of gain may be obtained from the values plotted by dividing by the correction factor

$$1 - \frac{\omega_3 - 8\omega_1}{8\omega_c} \frac{S_{\max}}{R_1\omega_1},$$

where $R_1\omega_1/S_{\max}$ is read directly from Fig. 13 or 14.

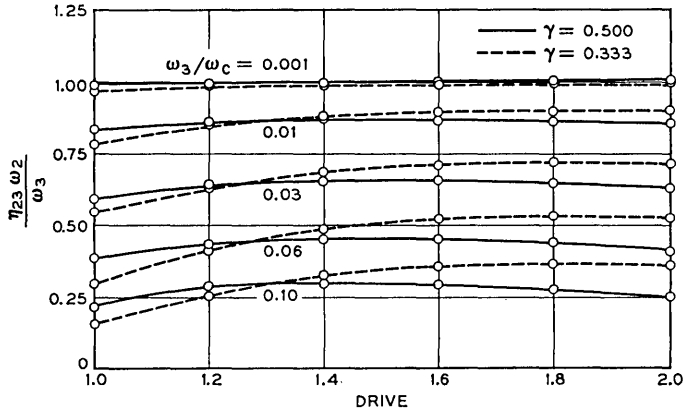


Fig. 10 — Microwave conversion efficiency for varactor upconverters operating at maximum power output. These results become inaccurate for high loss (high ω_3/ω_c) and $\omega_3 \neq 8\omega_1$. Accurate values of efficiency may be obtained from the values plotted by dividing by the correction factor

$$1 + \frac{\omega_3 - 8\omega_1}{8\omega_c}\frac{S_{\max}}{R_2\omega_2},$$

where $R_2\omega_2/S_{\max}$ is read directly from Fig. 15 or 16.

results are essentially independent of the loss parameter ω_3/ω_c. Values of the output elastance S_3 are presented in Fig. 12. Since S_3 was defined as the quadrature component of varactor voltage at ω_3, Kirchhoff's voltage law applied to the output loop requires that the load inductance present an equal but opposite reactance, i.e., $L_3 = S_3/\omega_3^2$.

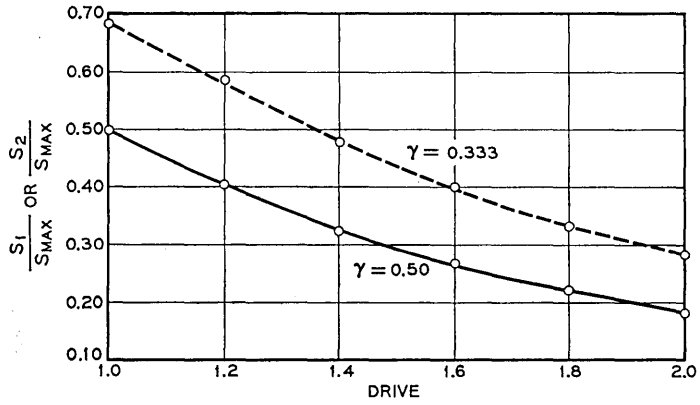


Fig. 11 — Input elastance at radian frequencies ω_1 and ω_2.

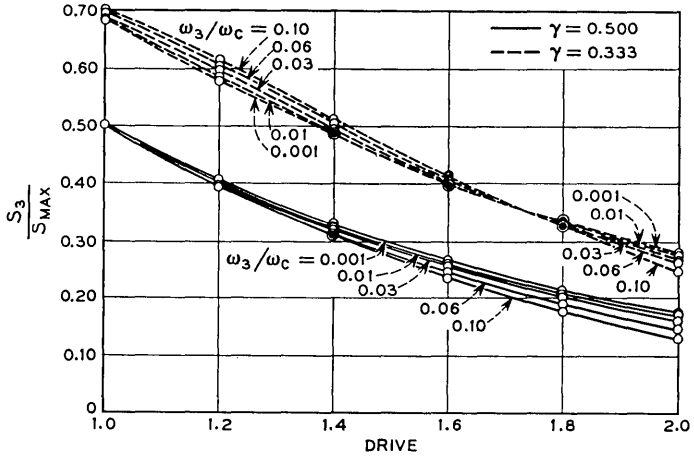


Fig. 12 — Output elastance at ω_3. The load inductance is equal to S_3/ω_3^2.

The input resistances R_1 and R_2 are presented in Figs. 13 to 16. The correction factors for high loss and $\omega_3 \neq 8\omega_1$ are described in the figure captions. The load resistance R_3 is plotted in Figs. 17 and 18. The abrupt-junction case has resistance maxima in the *drive* range of 1.4 to 1.6. This correlates with the range for maximum gain and efficiency as given by Figs. 9 and 10, as we would expect from (17) and (19). Similarly, the graded-junction case has maximum values of resistance, gain, and efficiency at a *drive* of approximately 1.8.

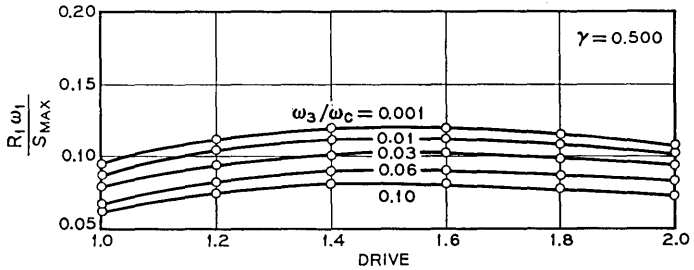


Fig. 13 — Input resistance at ω_1 for an abrupt-junction varactor. These results become inaccurate for high loss (high ω_3/ω_c) and $\omega_3 \neq 8\omega_1$. Accurate values of $R_1\omega_1/S_{max}$ may be obtained by subtracting the quantity

$$\frac{\omega_3 - 8\omega_1}{8\omega_c}$$

from the values plotted above.

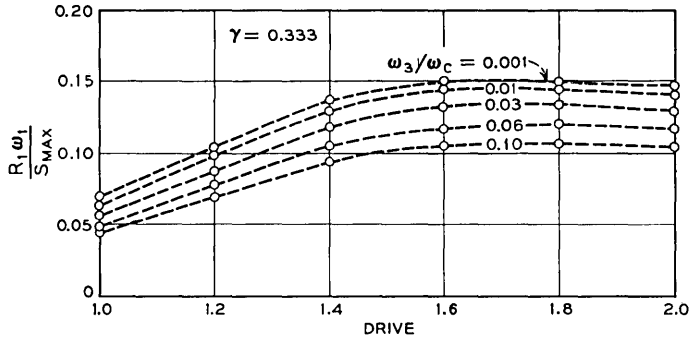


Fig. 14 — Input resistance at ω_1, for a graded-junction varactor. These results become inaccurate for high loss (high ω_3/ω_c) and $\omega_3 \neq 8\omega_1$. Accurate values of $R_1\omega_1/S_{\max}$ may be obtained by subtracting the quantity

$$\frac{\omega_3 - 8\omega_1}{8\omega_c}$$

from the values plotted.

If one considers the effect of the upconverter parameters upon the instantaneous bandwidth, he finds that large bandwidth requires large values of $R_i\omega_i/S_i$. Generally, the low-frequency ω_1 circuit is most crucial in this respect. Computing ratios of $R_1\omega_1/S_1$ from Figs. 11, 13, and 14 one finds that the abrupt-junction varactor has a higher ratio

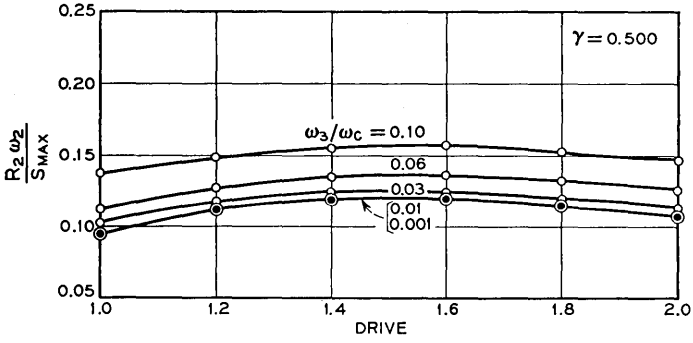


Fig. 15 — Input resistance at ω_2 for an abrupt-junction varactor. These results become inaccurate for high loss (high ω_3/ω_c) and $\omega_3 \neq 8\omega_1$. Accurate values of $R_2\omega_2/S_{\max}$ may be obtained by adding the quantity

$$\frac{\omega_3 - 8\omega_1}{8\omega_c}$$

to the values plotted above.

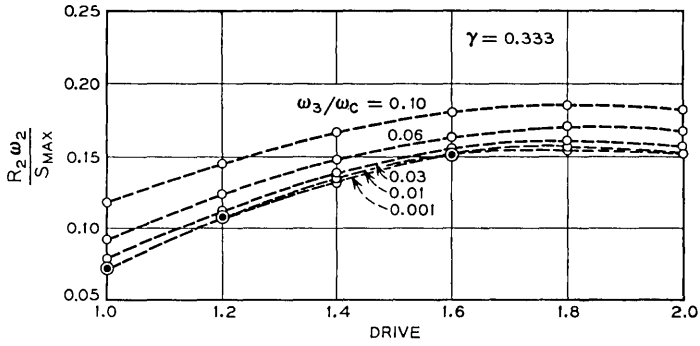


Fig. 16 — Input resistance at ω_2 for a graded-junction varactor. These results become inaccurate for high loss (high ω_3/ω_c) and $\omega_3 \neq 8\omega_1$. Accurate values of $R_{2\omega_2}/S_{max}$ may be obtained by adding the quantity

$$\frac{\omega_3 - 8\omega_1}{8\omega_c}$$

to the values plotted above.

at any prescribed drive level, albeit the difference becomes small (12 percent) at high drive levels. At *drive* = 1, the ratio $R_{1\omega_1}/S_1$ for the abrupt-junction varactor is nearly twice that of the graded-junction varactor.

Fig. 19 gives the dc bias voltage required for maximum output power.

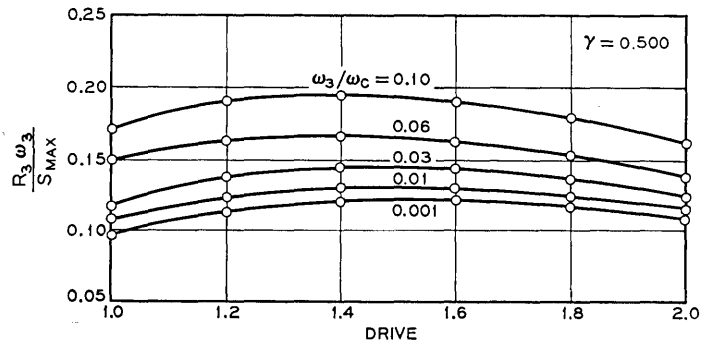


Fig. 17 — Load resistance for an abrupt-junction varactor.

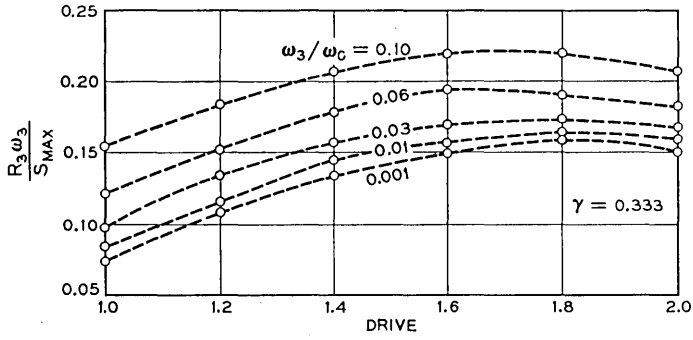


Fig. 18 — Load resistance for a graded-junction varactor.

IV. CONCLUSIONS

This analysis provides all the information required to design an upper-sideband upconverter for maximum power output, for abrupt-junction and graded-junction varactors. The results are sufficiently accurate for $\omega_2/\omega_1 \geqq 5$. The necessary load resistance and inductance are obtained from Figs. 12, 17, and 18. The input impedances at ω_1 and ω_2 are also presented, and the designer will ordinarily use this information to provide conjugate matching with the sources at these frequencies. Experimentally, the proper impedance matching conditions are facilitated by means of Swan's small-signal matching technique.[8]

It has been assumed throughout that currents are present in the varactor only at the three frequencies corresponding to the signal, pump, and upper-sideband.

At low drive levels, the abrupt-junction varactor provides both a

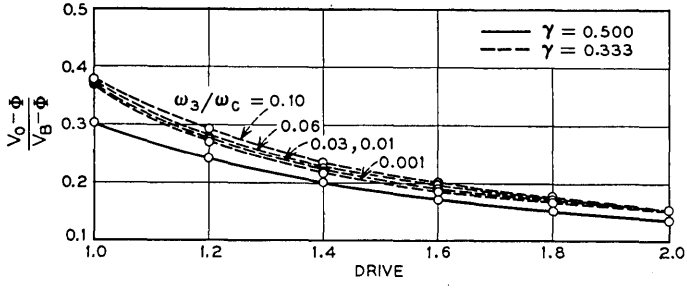


Fig. 19 — DC bias voltage required for maximum power output.

higher power-impedance product and also a greater bandwidth than that of the graded-junction varactor. However, at high drive levels the difference between the two varactor types is very small.

These results apply only to operation at maximum power output. In some applications, the designer is willing to sacrifice some power output in order to obtain a higher microwave conversion efficiency.[9] Such operation would require different operating parameters from those presented here.

APPENDIX

Efficiency and Gain Relations

The tuned upconverter may be represented by the equivalent circuit shown in Fig. 20. In this equivalent circuit the losses are separated from the frequency conversion device, and the power ratios for each are computed separately.

The power ratios for the lossless nonlinear capacitance are obtained from the Manley-Rowe relations:[7]

$$\sum_{m=0}^{\infty} \sum_{n=-\infty}^{\infty} \frac{mP'_{m+n}}{m\omega_1 + n\omega_2} = 0 \tag{26}$$

$$\sum_{m=-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{nP'_{m+n}}{m\omega_1 + n\omega_2} = 0, \tag{27}$$

where P'_{m+n} is the power into the lossless nonlinear capacitance at frequency $m\omega_1 + n\omega_2$. Restricting the exchange of power to the three radian frequencies of interest ω_1, ω_2, and $\omega_3 = \omega_1 + \omega_2$, (26) gives

$$\frac{P'_1}{\omega_1} + \frac{P'_3}{\omega_3} = 0 \tag{28}$$

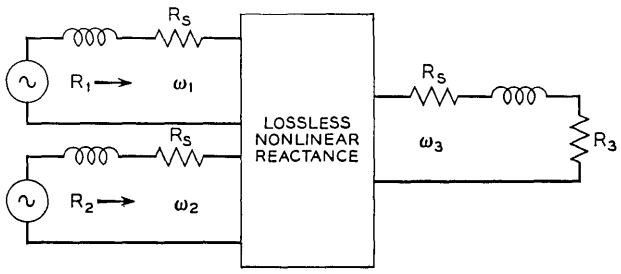


Fig. 20 — Basic equivalent circuit of the upconverter. Energy enters the nonlinear reactance at radian frequencies ω_1 and ω_2 and leaves at ω_3.

and (27) gives

$$\frac{P_2'}{\omega_2} + \frac{P_3'}{\omega_3} = 0. \tag{29}$$

In both cases P_3' is negative. If we redefine P_3' as the output power, the last two equations become

$$\frac{P_3'}{P_1'} = \frac{\omega_3}{\omega_1} \tag{30}$$

$$\frac{P_3'}{P_2'} = \frac{\omega_3}{\omega_2}. \tag{31}$$

In the input circuit at ω_2, a fraction $(R_2 - R_s)/R_2$ of the input power reaches the lossless nonlinear reactance. A fraction $R_3/(R_3+R_s)$ of the converted power reaches the load. Thus, the microwave conversion efficiency is given by

$$\begin{aligned} \eta_{23} &= \frac{R_2 - R_s}{R_2} \frac{P_3'}{P_2'} \frac{R_3}{R_3 + R_s} \\ &= \frac{\omega_3}{\omega_2} \frac{R_2 - R_s}{R_2} \frac{R_3}{R_3 + R_s}. \end{aligned} \tag{32}$$

Similarly, the upconversion gain is given by

$$G_{13} = \frac{\omega_3}{\omega_1} \frac{R_1 - R_s}{R_1} \frac{R_3}{R_3 + R_s}. \tag{33}$$

REFERENCES

1. Murphy, E. A., Posner, W., and Renkowitz, D., Solid-State 1-Watt FM Source at 6 Gc, ISSCC Digest Tech. Papers, *8*, February, 1965, pp. 104–105.
2. Hefni, I. E. and Spiwak, R. R., High-Efficiency Ultraflat Low-Noise Varactor Frequency Converter Using Low-Frequency Pumping, ISSCC Digest Tech. Papers, *9*, February, 1966, pp. 46–47.
3. Penfield, Jr., P. and Rafuse, R. P., *Varactor Applications,* MIT Press, Cambridge, Mass., 1962.
4. Nelson, C. E., A Note on the Large Signal Varactor Upper-Sideband Upconverter, Proc. IEEE, *54*, July, 1966, p. 1013.
5. Grayzel, A. I., The Design and Performance of 'Punch Through' Varactor Upper Sideband Up-Converters, NEREM Record, *8*, November, 1966, pp. 62–63.
6. Burckhardt, C. B., Analysis of Varactor Frequency Multipliers for Arbitrary Capacitance Variation and Drive Level, B.S.T.J., *44*, April, 1965, pp. 675–692.
7. Manley, J. M. and Rowe, H. E., Some General Properties of Nonlinear Elements—Part I. General Energy Relations, Proc. IRE, *44*, July, 1956, pp. 904–913.
8. Swan, C. B., Design and Evaluation of a Microwave Varactor Tripler, ISSCC Digest Tech. Papers, *8*, February, 1965, pp. 106–107.
9. Gravzel, A. I., A Note on the Abrupt Junction Large Signal Upconverter, Proc. IEEE, *54*, January, 1966, pp. 78–79.

Two Theorems on the Accuracy of Numerical Solutions of Systems of Ordinary Differential Equations

By I. W. SANDBERG

(Manuscript received March 13, 1967)

We consider the accuracy with which a numerical solution of the system of ordinary differential equations

$$\dot{x} + f(x, t) = 0, \qquad t \geqq 0$$

can be obtained by the use of a numerical integration formula of the well-known type

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k y'_{n-k} .$$

For the scalar case, under some natural assumptions, and assuming that α and β are real constants such that

$$\alpha \leqq \frac{\partial f(x, t)}{\partial x} \leqq \beta, \qquad t \geqq 0$$

at every point x, it is proved that if

$$F(z) \triangleq 1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + \tfrac{1}{2}(\alpha + \beta)h \sum_{k=-1}^{p} b_k z^{-(k+1)} \neq 0$$

for all $| z | \geqq 1$, then $\langle e \rangle$, the root-mean-squared error over a given interval, between the true samples of $x(t)$ and the y_n, satisfies

$$\langle e \rangle \geqq (1 + \rho)^{-1} \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle$$

in which ρ depends on α, β, the a_k, and the b_k, and $\langle \varphi \rangle$ takes into account the local roundoff and truncation errors as well as errors in the starting values for computing the y_n.

If the condition on $F(z)$ stated above holds, and if $\rho < 1$, then

$$\langle e \rangle \leqq (1 - \rho)^{-1} \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle .$$

The significance of the key assumptions is discussed and two examples are given.

I. INTRODUCTION

In this paper we present some results concerning the accuracy with which a numerical solution of the system of ordinary differential equations

$$\dot{x} + f(x, t) = 0, \qquad t \geqq 0 \quad [x(0) = x_0] \tag{1}$$

can be obtained by the use of a numerical integration formula of the well-known type[1]

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k y'_{n-k}, \qquad n \geqq p. \tag{2}$$

In (2) the y_n are approximations to the $x_n \triangleq x(nh)$, where h, a positive number, is the step size parameter; $y_0, y_1, \cdots, y_p$ are starting vectors, the last p of which are obtained by an independent method; and

$$y'_n \triangleq -f(y_n, nh).$$

If $b_{-1} \neq 0$, then y_{n+1} is defined implicitly, and (2) is said to be of closed type. It is assumed throughout that (2) can be solved* for y_{n+1} for all $n \geqq p$. Specializations of (2) include, for example, Euler's method:

$$y_{n+1} = y_n + hy'_n,$$

and the more useful formula

$$y_{n+1} = y_n + \tfrac{1}{2}h(y'_n + y'_{n+1}).$$

It is assumed throughout that for $t \geqq 0$, $f(x, t)$ is a well-defined real N-vector-valued function defined in the set of all real N-vectors x, that $f(x, t)$ satisfies (the usual weak) conditions which guarantee the existence and uniqueness of a solution of (1), and that the Jacobian matrix $\partial f(x, t)/\partial x$ exists for all x and all $t \geqq 0$.

Equation (2) ignores the roundoff error R_n introduced in calculating y_{n+1}, and, in order to take R_n into account, we shall consider instead

* It is well known that if f satisfies a uniform Lipshitz condition, and if h is sufficiently small, then (2) possesses a unique solution y_{n+1} which can be obtained by a simple iterative process.[1,2] However, this smallness condition is by no means always necessary. For example, if $b_{-1} > 0$ and, with α as defined in Section 2.3, if $\alpha > 0$, then for *any* $h > 0$ a unique solution y_{n+1} exists and can be computed by an iterative process which is only slightly more complicated than the usual one (see Section IV).

of (2):

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k y'_{n-k} + R_n , \qquad n \geqq p. \tag{3}$$

We may also assume that R_n takes into account the error in solving (2) for y_{n+1}, caused typically by truncating an iteration procedure after a finite number of steps. Predictor-corrector techniques can of course be viewed as giving rise to a degenerate (often one-step) iteration technique in which the "starting point" is generated by the predictor.

The truncation error T_n, a basic entity associated with the integration formula (2) and the differential equation (1), is defined for $n \geqq p$ by the relation

$$x_{n+1} = \sum_{k=0}^{p} a_k x_{n-k} + h \sum_{k=-1}^{p} b_k x'_{n-k} + T_n , \qquad n \geqq p$$

in which $x'_n = -f(x_n , nh)$. Clearly, T_n is a measure of how well the samples $x_{n-p} , x_{n-p+1} , \cdots , x_{n+1}$ of the solution of (1) satisfy the integration formula. The problem of estimating T_n is a classical one, and there are standard methods which lead to precise bounds.[1,2]

We now define a set of vectors $\{\varphi_n\}$ which plays a central role in the subsequent discussion:

$$\varphi_n = T_n - R_n , \qquad n \geqq p \tag{5}$$

$$\varphi_n = (x_{n+1} - y_{n+1}) - \sum_{k=0}^{p} a_k (x_{n-k} - y_{n-k}) + h \sum_{k=-1}^{p} b_k \{f[x_{n-k} , (n-k)h] - f[y_{n-k} , (n-k)h]\}, \qquad n = -1, 0, \cdots , (p-1) \tag{6}$$

(with the understanding that $x_n = y_n = f(x_n , nh) = f(y_n , nh) = 0$ for $n < 0$). Note that the φ_n for $n = -1, 0, \cdots , (p-1)$ are measures of the departures of the starting vectors from the exact values, and that $\varphi_n = 0$ for $n = -1, 0, \cdots , (p-1)$ if the starting vectors are exact.

We shall describe our results first for the scalar case (i.e., for $N = 1$).

II. RESULTS

Let e_n denote $(x_n - y_n)$, the difference between $x(nh)$ and its computed approximation. Suppose that $N = 1$, and that α and β are real

constants such that

$$\alpha \leqq \frac{\partial f(x, t)}{\partial x} \leqq \beta \tag{7}$$

for all $t \geqq 0$ (at every point x), and that

$$F(z) \triangleq 1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + \tfrac{1}{2}(\alpha + \beta)h \sum_{k=-1}^{p} b_k z^{-(k+1)} \neq 0 \tag{8}$$

for all $| z | \geqq 1$ (including "$z = \infty$"). We prove that then

$$\langle e \rangle \triangleq \left((M + 1)^{-1} \sum_{m=0}^{M} | e_m |^2 \right)^{\frac{1}{2}}, \tag{9}$$

the root-mean-squared value of the first $(M + 1)$ error terms [M is an *arbitrary* positive integer greater than or equal to $(p + 1)$] is bounded from *below* in terms of

$$\langle \varphi \rangle \triangleq \left((M + 1)^{-1} \sum_{m=0}^{M} | \varphi_{m-1} |^2 \right)^{\frac{1}{2}}, \tag{10}$$

the corresponding quantity for the φ's, in accordance with the inequality

$$\langle e \rangle \geqq (1 + \rho)^{-1} \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle. \tag{11}$$

[$F(z)$ is defined in (8)], in which

$$\rho \triangleq \tfrac{1}{2}(\beta - \alpha)h \max_{0 \leqq \omega \leqq 2\pi} \left| \frac{\sum_{k=-1}^{p} b_k e^{-i(k+1)\omega}}{F(e^{i\omega})} \right|. \tag{12}$$

We also prove that if in addition to the assumptions stated above, we have $\rho < 1$, then the sequence $\{e_n\}$ is bounded (i.e., there exists a positive constant c such that $| e_n | \leqq c$ for all $n \geqq 0$) whenever the sequence $\{\varphi_{n-1}\}$ is bounded, and

$$\langle e \rangle \leqq (1 - \rho)^{-1} \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle \tag{13}$$

(whether or not $\{\varphi_{n-1}\}$ is bounded).

Inequality (11) provides a limitation on obtainable accuracy under essentially only the weak assumption that the sequence of approximations $\{y_n\}$ defined by (2) approaches zero as $n \to \infty$ for all sets of starting values when $f(x, t) = \frac{1}{2}(\alpha + \beta)x$. Since, by assumption, $F(e^{i\omega}) \neq 0$ for $0 \leqq \omega \leqq 2\pi$, it is clear that $\rho < \infty$.

The condition that $\rho < 1$ is satisfied if and only if the locus of

$$\Theta(\omega) \triangleq \frac{\sum_{k=0}^{p} a_k e^{ik\omega} - e^{-i\omega}}{\sum_{k=-1}^{p} b_k e^{ik\omega}} \tag{14}$$

for $0 \leqq \omega \leqq 2\pi$ lies outside the "critical circle" C of radius $\frac{1}{2}(\beta - \alpha)h$ centered in the complex plane at $[\frac{1}{2}(\alpha+\beta)h, 0]$ (see Fig. 1).*

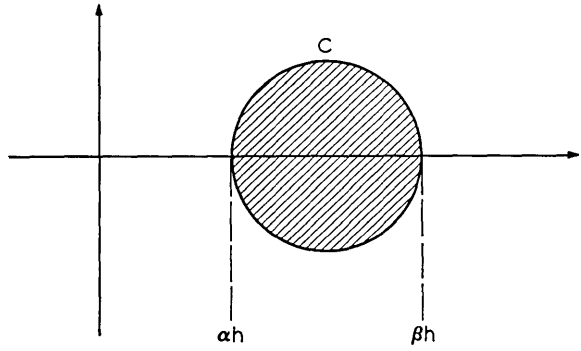


Fig. 1 — Location of the critical circle C (for $N = 1$).

Since

$$\rho = \tfrac{1}{2}(\beta - \alpha)h\{\min_{\omega} | \Theta(\omega) - \tfrac{1}{2}(\alpha + \beta)h |\}^{-1}, \tag{15}$$

we see that ρ is the ratio of the radius of C to the distance between c and θ, where c is the center of C and θ is a point nearest c on the locus of $\Theta(\omega)$.

2.1 *Discussion*

The quantity $\langle e \rangle$ of course of interest in problems in which we are concerned with a measure of the accuracy of a numerical solution

* If $\alpha > 0$, we can express both the conditions that $\rho < 1$ and $F(z) \neq 0$ for $|z| \geqq 1$ entirely in terms of a condition on the locus of $\Theta(\omega)^{-1}$ for $\omega \in [0, 2\pi]$. The requirements on $F(z)$ and ρ are met if the disk of radius $\frac{1}{2}[(\alpha h)^{-1} - (\beta h)^{-1}]$ centered at $\{\frac{1}{2}[(\alpha h)^{-1} + (\beta h)^{-1}], 0\}$ is not "encircled" or intersected by the locus of $\Theta(\omega)^{-1}$. There is a complication that arises as a result of the fact that $\Theta(\omega)^{-1}$ is typically not bounded. This complication often stems from a "consistency requirement" which implies that $1 - \sum^{p}_{k=0} a_k z^{-(k+1)}$ has at least one zero on the unit circle. However, due to also typical "convergence requirements," $1 - \sum^{p}_{k=0} a_k z^{-(k+1)}$ normally has only *simple* zeros on the unit circle, a fact that can be used to suitably define what is meant by the locus of $\Theta(\omega)^{-1}$ not encircling the disk. We leave the details of the necessary "principle of the argument" argument to the sufficiently interested reader.

over a large number of steps, as opposed to the accuracy of some final value obtained at the end of a large number of steps.

Although there is a vast and interesting literature concerned with various aspects of the problem of error estimation in digital computation (see, for example, Refs. 3, 4, and 5), the results presented above, and their proofs, appear to be most closely related to earlier results concerning the input-output stability of continuous-time nonlinear feedback systems.[6,*] Indeed, the writer is not aware of any lower-bound results in the numerical analysis literature of the type described above. However, some upper bounds concerning (2) of (for example) the form $|e_n| \leqq K$ with K independent of n (which imply $\langle e \rangle \leqq K$) have been obtained in certain cases. In this connection, our condition that guarantees the boundedness of $\{e_n\}$ is often weaker, and our upper bounds on $\langle e \rangle$ are often much stronger, because, for example, the φ_{m-1} can become very small as m becomes large.

Our approach can be applied to several other problems in numerical analysis. In particular, with reasonably direct modifications of our proofs, analogous theorems can be proved concerning the numerical integration of systems of second-order ordinary differential equations.

2.2 *Examples*

Euler's Method: $y_{n+1} = y_n + hy'_n$

Here $F(z) = 1 - [1 - \frac{1}{2}(\alpha + \beta)h]z^{-1}$, so that $F(z) \neq 0$ for $|z| \geqq 1$ if and only if $0 < \frac{1}{2}(\alpha + \beta)h < 2$, and $|F(e^{i\omega})| = |1 - [1 - \frac{1}{2}(\alpha + \beta)h]e^{-i\omega}|$. Thus (with $0 < \frac{1}{2}(\alpha + \beta)h < 2$),

$$\min_{\omega} |F(e^{i\omega})|^{-1} = [1 + |1 - \tfrac{1}{2}(\alpha + \beta)h|]^{-1}$$

$$\max_{\omega} |F(e^{i\omega})|^{-1} = [1 - |1 - \tfrac{1}{2}(\alpha + \beta)h|]^{-1}.$$

The locus of Θ is the circle shown in Fig. 2, since $\Theta(\omega) = 1 - e^{-i\omega}$. If $\alpha h > 0$ and $\beta h < 2$, then the critical disk (Fig. 2) is not intersected by the locus of Θ, the condition that $0 < \frac{1}{2}(\alpha + \beta)h < 2$ is satisfied, $\rho < 1$, and in accordance with the last paragraph of the section preceding Section 2.1:

$$\rho = \tfrac{1}{2}(\beta - \alpha)h \max([\tfrac{1}{2}(\beta + \alpha)h]^{-1}, [2 - \tfrac{1}{2}(\beta + \alpha)h]^{-1}).$$

---

* It is interesting to note that the possibility of exploiting feedback-theoretic ideas has been emphasized by Hamming.[2]

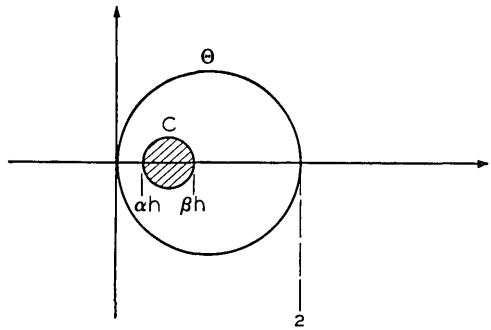


Fig. 2 — The locus of $\Theta(\omega)$ for Euler's method, and the critical circle C.

If $0 < (\alpha + \beta)h < 2$, then

$$\min_{\omega} | F(e^{i\omega}) |^{-1} = [2 - \tfrac{1}{2}(\alpha + \beta)h]^{-1},$$

$$\max_{\omega} | F(e^{i\omega}) |^{-1} = [\tfrac{1}{2}(\alpha + \beta)h]^{-1},$$

$$\rho = \frac{\beta - \alpha}{\beta + \alpha},$$

and

$$\langle e \rangle \leqq [1 + (\beta - \alpha)(\beta + \alpha)^{-1}]^{-1}[2 - \tfrac{1}{2}(\alpha + \beta)h]^{-1}\langle \varphi \rangle$$

$$\langle e \rangle \leqq [1 - (\beta - \alpha)(\beta + \alpha)^{-1}]^{-1}[\tfrac{1}{2}(\alpha + \beta)h]^{-1}\langle \varphi \rangle.$$

For estimates of T_n, see Ref. 1 or 2.

As a remark concerning the necessity of the condition $\rho < 1$, we note that if $\alpha h > 0$, but $\beta h > 2$, then for even the special case in which $f(x, t) = \beta x$, we have $e_0, e_1, e_2, \cdots$ unbounded, since $y_0, y_1, y_2, \cdots$ is unbounded (assuming merely that $y_0 \neq 0$).

The Formula $y_{n+1} = y_n + \tfrac{1}{2}h(y'_n + y'_{n+1})$:

In this important case

$$F(z) = 1 + \tfrac{1}{4}(\alpha + \beta)h - [1 - \tfrac{1}{4}(\alpha + \beta)h]z^{-1}, \quad \text{and}$$

$$\Theta(\omega) = \frac{1 - e^{-i\omega}}{\tfrac{1}{2}(1 + e^{-i\omega})} = 2i \tan\left(\frac{\omega}{2}\right).$$

We have $F(z) \neq 0$ for $|z| \geqq 1$ if and only if $(\alpha + \beta)h > 0$, and [assuming that $(\alpha + \beta)h > 0$]:

$$\min_{\omega} \mid F(e^{i\omega}) \mid^{-1} = [1 + \tfrac{1}{4}(\alpha + \beta)h + \mid 1 - \tfrac{1}{4}(\alpha + \beta)h \mid]^{-1}$$

$$\max_{\omega} \mid F(e^{i\omega}) \mid^{-1} = [1 + \tfrac{1}{4}(\alpha + \beta)h - \mid 1 - \tfrac{1}{4}(\alpha + \beta)h \mid]^{-1}.$$

The locus of Θ lies entirely on the imaginary axis of the complex plane,

$$\rho = \frac{\beta - \alpha}{\beta + \alpha},$$

and obviously $\rho < 1$ if $\alpha > 0$.

If $\alpha > 0$ and $(\alpha + \beta)h < 4$, then

$$\min_{\omega} \mid F(e^{i\omega}) \mid^{-1} = \tfrac{1}{2}$$

$$\max_{\omega} \mid F(e^{i\omega}) \mid^{-1} = [\tfrac{1}{2}(\alpha + \beta)h]^{-1}$$

and

$$\langle e \rangle \geqq [1 + (\beta - \alpha)(\beta + \alpha)^{-1}]^{-1}\tfrac{1}{2}\langle \varphi \rangle$$

$$\langle e \rangle \leqq [1 - (\beta - \alpha)(\beta + \alpha)^{-1}]^{-1}[\tfrac{1}{2}(\alpha + \beta)h]^{-1}\langle \varphi \rangle.$$

The last inequality can be written as simply

$$\langle e \rangle \leqq (\alpha h)^{-1}\langle \varphi \rangle.$$

For the integration formula under consideration,

$$T_n = \frac{h^3 x'''(\eta_n)}{12}$$

where η_n lies in the interval $[(n - p)h, (n + 1)h]$.

Here $p = 0$, and

$$\varphi_{-1} = (x_0 - y_0) + \tfrac{1}{2}h[f(x_0 , 0) - f(y_0 , 0)].$$

Thus, assuming for the purpose of illustration that roundoff errors can be neglected:

$$\langle e \rangle \leqq (\alpha h)^{-1}\left((M + 1)^{-1} \mid \varphi_{-1} \mid^2 + (M + 1)^{-1} \sum_{m=1}^{M} \left| \frac{h^3 x'''(\eta_m)}{12} \right|\right)^{\frac{1}{2}}$$

provided that $\alpha > 0$ and $(\alpha + \beta)h < 4$. If, for example, $(\beta - \alpha) \cdot (\beta + \alpha)^{-1} = \frac{1}{2}$ and $\frac{1}{2}(\alpha + \beta)h = \frac{1}{3}$, then the ratio of our upper bound on $\langle e \rangle$ to our lower bound is 18. If $(\beta - \alpha)(\beta + \alpha)^{-1} = \frac{3}{4}$ and $\frac{1}{2}(\alpha + \beta)h = \frac{1}{3}$, then the ratio is 42.

2.3 *Results for the Vector Case* ($N \geqq 1$)

We shall state our results for $N \geqq 1$ in a slightly more formal fashion.

Definitions:

(*i*) Let $\| q \|$ denote $(\sum_{k=1}^{N} q_k^2)^{\frac{1}{2}}$ for every real N-vector $q = (q_1, q_2, \cdots, q_N)$.

(*ii*) Let $\{\partial f(x, t)/\partial x\}_S$ and $\{\partial f(x, t)/\partial x\}_A$ denote, respectively, the symmetric and antisymmetric (i.e., skew symmetric) part of $\partial f(x, t)/\partial x$, the Jacobian matrix of $f(x, t)$.

(*iii*) Let $F(z) \triangleq 1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + \frac{1}{2}(\alpha + \beta)h \sum_{k=-1}^{p} b_k z^{-(k+1)}$

(*iv*) $e_n \triangleq x(nh) - y_n$, $n \geqq 0$ with the y_n for $n \geqq (p + 1)$ defined by (3).

(*v*) With M an arbitrary positive integer such that $M \geqq (p + 1)$, let

$$\langle e \rangle \triangleq \left((M + 1)^{-1} \sum_{m=0}^{M} \| e_m \|^2 \right)^{\frac{1}{2}}$$

$$\langle \varphi \rangle \triangleq \left((M + 1)^{-1} \sum_{m=0}^{M} \| \varphi_{m-1} \|^2 \right)^{\frac{1}{2}},$$

where the φ_{m-1} are defined in (5) and (6).

Assumptions:

Let the smallest eigenvalue of $\{\partial f(x, t)/\partial x\}_S$ be bounded from below by the real constant $\alpha (\alpha > -\infty)$ for all $t \geqq 0$, and let the largest eigenvalue of $\{\partial f(x, t)/\partial x\}_S$ be bounded from above by the real constant $\beta (\beta < \infty)$ for all $t \geqq 0$. Let the modulus of the largest eigenvalue of $\{\partial f(x, t)/\partial x\}_A$ be bounded from above by the real constant $\gamma (\gamma < \infty)$ for all $t \geqq 0$.

Definition:

Let

$$\rho \triangleq [\tfrac{1}{2}(\beta - \alpha)h + \gamma h] \cdot \max_{0 \leq \omega \leq 2\pi} \left| \frac{\sum_{k=-1}^{p} b_k e^{-i(k+1)\omega}}{1 - \sum_{k=0}^{p} a_k e^{-i(k+1)\omega} + \frac{1}{2}(\alpha + \beta)h \sum_{k=-1}^{p} b_k e^{-i(k+1)\omega}} \right| .$$

Theorem 1: If

(*i*) *the assumptions of Section I concerning* $f(x, t)$ *and* (2) *are satisfied,*

(*ii*) $1 + \frac{1}{2}(\alpha + \beta)hb_{-1} \neq 0$,

(*iii*) $F(z) \neq 0$ for all $| z | \geqq 1$,

then

$$\langle e \rangle \geqq (1 + \rho)^{-1} \min_{0 \leq \omega \leq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle.$$

Theorem 2: *If (i), (ii), and (iii) of Theorem 1 are satisfied, and if* $\rho < 1$, *then*

$$(i) \qquad \langle e \rangle \leqq (1 - \rho)^{-1} \max_{0 \leq \omega \leq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle,$$

and

$$(ii) \qquad \sup_{m \geqq 0} || \varphi_{m-1} || < \infty \quad \text{implies} \quad \sup_{m \geqq 0} || e_m || < \infty .$$

Corollary to Theorem 1: If (i) of Theorem 1 is satisfied and there exists at least one real constant k_1 *such that*

$$1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + k_1 h \sum_{k=-1}^{p} b_k z^{-(k+1)} \neq 0$$

for all $| z | \geqq 1$, *then there exists a positive constant* k_2, *which depends only on* $a_0, a_1, \cdots, a_p, b_{-1}, b_0, \cdots, b_p, \alpha, \beta$, *and* γ *such that*

$$\langle e \rangle \geqq k_2 \langle \varphi \rangle.$$

Theorems 1 and 2 are proved* in the following section. The proof of the corollary is very simple.

Since

$$1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + k_1 h \sum_{k=-1}^{p} b_k z^{-(k+1)} \neq 0$$

for all $| z | \geqq 1$, there exists a k_1' such that

$$1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + k_1' h \sum_{k=-1}^{p} b_k z^{-(k+1)} \neq 0$$

for all $| z | \geqq 1$, and

$$1 + k_1' h b_{-1} \neq 0.$$

Choose α' and β' such that $\alpha' \leqq \alpha$, $\beta' \geqq \beta$, and $\frac{1}{2}(\alpha' + \beta') = k_1'$. If we replace α and β with α' and β', respectively, we see that Theorem 1 applies.

---

* Our proofs actually yield sharper, but less explicit, bounds on $\langle e \rangle$ than those of Theorems 1 and 2. See (31) and (37).

III. PROOFS

Proof of Theorem 1:

We have

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k y'_{n-k} + R_n , \qquad n \geqq p$$

and

$$x_{n+1} = \sum_{k=0}^{p} a_k x_{n-k} + h \sum_{k=-1}^{p} b_k x'_{n-k} + T_n , \qquad n \geqq p.$$

Thus,

$$e_{n+1} = \sum_{k=0}^{p} a_k e_{n-k} - h \sum_{k=-1}^{p} b_k \{f[x_{n-k}(n-k)h] - f[y_{n-k} , (n-k)h]\} + \varphi_n , \quad n \geqq p$$

and, with φ_n defined for $n = -1, 0, \cdots, (p-1)$ by (6),

$$e_n = \sum_{k=0}^{p} a_k e_{n-1-k} - h \sum_{k=-1}^{p} b_k \{f[x_{n-1-k} , (n-1-k)h] - f[y_{n-1-k} , (n-1-k)h]\} + \varphi_{n-1} \qquad (16)$$

for $n \geqq 0$.

As a matter of convenience we define $a_{-1} \triangleq 0$.

Lemma 1: *There exist real sequences* $\{w_k\}_{k=0}^{\infty}$ *and* $\{v_k\}_{k=0}^{\infty}$ *both belonging to* l_1 (*i.e., with the property that* $\sum_{k=0}^{\infty} | w_k | < \infty$ *and* $\sum_{k=0}^{\infty} | v_k | < \infty$) *such that*

$$W(z) \triangleq \sum_{k=0}^{\infty} w_k z^{-k} = \frac{-h \sum_{k=-1}^{p} b_k z^{-(k+1)}}{1 - \sum_{k=-1}^{p} [a_k - \frac{1}{2}(\alpha + \beta)hb_k] z^{-(k+1)}} \qquad (17)$$

$$V(z) \triangleq \sum_{k=0}^{\infty} v_k z^{-k} = \frac{1}{1 - \sum_{k=-1}^{p} [a_k - \frac{1}{2}(\alpha + \beta)hb_k] z^{-(k+1)}} \qquad (18)$$

for all $| z | \geqq 1$.

The proof follows at once from the standard theory of z-transforms, in view of assumptions (*ii*) and (*iii*) of Theorem 1. The details are omitted.

Lemma 2: *Let* $\delta_n \triangleq f(x_n, nh) - f(y_n, nh) - \frac{1}{2}(\alpha + \beta)(x_n - y_n)$ *for all* $n \geqq 0$, *and let* $\{w_k\}$ *and* $\{v_k\}$ *be as described in Lemma 1. Then*

$$e_n = \sum_{k=0}^{n} w_{n-k}\,\delta_k + \sum_{k=0}^{n} v_{n-k}\varphi_{k-1} \tag{19}$$

for $n = 0, 1, \cdots, M$.

Proof of Lemma 2:

From (16) and our definition of δ_n:

$$\begin{aligned} e_n = \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k]e_{n-k-1} \\ - h\sum_{k=-1}^{p} b_k\,\delta_{n-k-1} + \varphi_{n-1}\,, \qquad n \geqq 0. \end{aligned} \tag{20}$$

We multiply both sides of (20) by $e^{-in\omega}$ and then sum from $n = 0$ to $n = M$ to obtain

$$\begin{aligned} \sum_{n=0}^{M} e^{-in\omega}e_n = \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k] \sum_{n=0}^{M} e^{-in\omega}e_{n-k-1} \\ - h\sum_{k=-1}^{p} b_k \sum_{n=0}^{M} e^{-in\omega}\,\delta_{n-k-1} + \sum_{n=0}^{M} e^{-in\omega}\varphi_{n-1} \end{aligned} \tag{21}$$

for all $\omega \in [0, 2\pi]$. Using $e_n = \delta_n = 0$ for $n < 0$, we have

$$\sum_{n=0}^{M} e^{-in\omega}e_{n-k-1} = e^{-i(1+k)\omega}\sum_{n=0}^{M} e^{-in\omega}e_n - e^{-i(1+k)\omega}\sum_{n=M-k}^{M} e^{-in\omega}e_n \tag{22}$$

and

$$\sum_{n=0}^{M} e^{-in\omega}\,\delta_{n-k-1} = e^{-i(1+k)\omega}\sum_{n=0}^{M} e^{-in\omega}\,\delta_n - e^{-i(1+k)\omega}\sum_{n=M-k}^{M} e^{-in\omega}\,\delta_n\,. \tag{23}$$

Thus, from (21), (22), and (23)

$$\begin{aligned} \sum_{n=0}^{M} e^{-in\omega}e_n = W(e^{i\omega})\sum_{n=0}^{M} e^{-in\omega}\,\delta_n + V(e^{i\omega})\sum_{n=0}^{M} e^{-in\omega}\varphi_{n-1} \\ + V(e^{i\omega})\Big\{h\sum_{k=-1}^{p} b_k e^{-i(1+k)\omega}\sum_{n=M-k}^{M} e^{-in\omega}\,\delta_n \\ - \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k]e^{-i(1+k)\omega}\sum_{n=M-k}^{M} e^{-in\omega}e_n\Big\} \end{aligned} \tag{24}$$

for all $\omega \in [0, 2\pi]$.

The expression within the braces in (24) can be written as

$$\sum_{n=0}^{\infty} s_n e^{-in\omega}$$

with $s_n = 0$ for $n = 0, 1, \cdots, M$ and for $n > (1 + M + p)$. Since $\{v_k\} \varepsilon l_1$, we have

$$V(e^{i\omega}) \sum_{n=0}^{\infty} s_n e^{-in\omega} = \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} v_{n-k} s_k .$$

Similarly,

$$V(e^{i\omega}) \sum_{n=0}^{M} e^{-in\omega} \varphi_{n-1} = \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} v_{n-k} (\varphi_{k-1})_M$$

in which

$$\begin{aligned} (\varphi_{k-1})_M &= \varphi_{k-1} , \qquad k \leqq M \\ &= 0 \qquad\qquad k > M \end{aligned}$$

and finally, since $\{w_k\} \epsilon l_1$,

$$W(e^{i\omega}) \sum_{n=0}^{M} e^{-in\omega} \delta_n = \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} w_{n-k} (\delta_k)_M ,$$

where

$$\begin{aligned} (\delta_k)_M &= \delta_k , \qquad k \leqq M \\ &= 0, \qquad k > M. \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{n=0}^{M} e^{-in\omega} e_n = & \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} w_{n-k} (\delta_k)_M \\ & + \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} v_{n-k} (\varphi_{k-1})_M \\ & + \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} v_{n-k} s_k \end{aligned}$$

for all $\omega \epsilon [0, 2\pi]$. Since

$$\sum_{k=0}^{n} v_{n-k} s_k = 0$$

for $n = 0, 1, \cdots, M$ we have*

$$e_n = \sum_{k=0}^{n} w_{n-k}\ \delta_k + \sum_{k=0}^{n} v_{n-k}\varphi_{k-1} \tag{25}$$

for $n = 0, 1, \cdots, M$. This completes the proof of Lemma 2.

Lemma 3: If (19) holds for $n = 0, 1, \cdots, M$, then

$$\left(\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} v_{n-k}\varphi_{k-1} \right|\right|^2\right)^{\frac{1}{2}} \leqq \left(\sum_{n=0}^{M} ||\ e_n\ ||^2\right)^{\frac{1}{2}} + \left(\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} w_{n-k}\ \delta_k \right|\right|^2\right)^{\frac{1}{2}} \tag{26}$$

and

$$\left(\sum_{n=0}^{M} ||\ e_n\ ||^2\right)^{\frac{1}{2}} \leqq \left(\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} w_{n-k}\ \delta_k \right|\right|^2\right)^{\frac{1}{2}} + \left(\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} v_{n-k}\varphi_{k-1} \right|\right|^2\right)^{\frac{1}{2}}. \tag{27}$$

Proof of Lemma 3:

Inequality (26) or (27) follows from (19) by two applications of Minkowski's inequality. We leave the details to the reader.

Inequality (27) is used only in the proof of Theorem 2.

Lemma 4:

$$\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} w_{n-k}\ \delta_k \right|\right|^2 \leqq \max_{0 \leqq \omega \leqq 2\pi} |\ W(e^{i\omega})\ |^2 \sum_{n=0}^{M} ||\ \delta_n\ ||^2.$$

Proof of Lemma 4:

By Parseval's identity,

$$\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} w_{n-k}\ \delta_k \right|\right|^2 = \frac{1}{2\pi} \int_0^{2\pi} \left|\left| \sum_{n=0}^{M} e^{-i\omega n} \sum_{k=0}^{n} w_{n-k}\ \delta_k \right|\right|^2 d\omega$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \left|\left| \sum_{n=0}^{M} e^{-i\omega n} \sum_{k=0}^{n} w_{n-k} (\delta_k)_M \right|\right|^2 d\omega$$

in which

$$(\delta_k)_M = \delta_k\ , \qquad k \leqq M$$

$$= 0, \qquad k > M.$$

Therefore,

$$\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} w_{n-k}\ \delta_k \right|\right|^2 \leqq \frac{1}{2\pi} \int_0^{2\pi} \left|\left| \sum_{n=0}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} w_{n-k} (\delta_k)_M \right|\right|^2 d\omega.$$

* We could have obtained (25) from (20) directly using standard z-transform theory, if we had introduced further assumptions which guarantee that the sequences $\{y_n\}$, $\{x(nh)\}$, and $\{\varphi_{k-1}\}$ are transformable. However, this would have complicated the statement of our results and would have weakened them in a non-trivial manner.

But since $\{w_k\} \in l_1$,

$$\sum_{n=0}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} w_{n-k}(\delta_k)_M = \sum_{n=0}^{\infty} e^{-i\omega n} w_n \sum_{k=0}^{M} e^{-i\omega k}\, \delta_k$$

$$= W(e^{i\omega}) \sum_{k=0}^{M} e^{-i\omega k}\, \delta_k \,.$$

Thus,

$$\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} w_{n-k}\, \delta_k \right|\right|^2 \leqq \frac{1}{2\pi} \int_0^{2\pi} \left|\left| W(e^{i\omega}) \sum_{k=0}^{M} e^{-i\omega k}\, \delta_k \right|\right|^2 d\omega$$

$$\leqq \max_{0 \leqq \omega \leqq 2\pi} |\, W(e^{i\omega})\, |^2 \frac{1}{2\pi} \int_0^{2\pi} \left|\left| \sum_{k=0}^{M} e^{-i\omega k}\, \delta_k \right|\right|^2 d\omega$$

$$\leqq \max_{0 \leqq \omega \leqq 2\pi} |\, W(e^{i\omega})\, |^2 \sum_{k=0}^{M} ||\, \delta_k\, ||^2$$

which proves Lemma 4.

Lemma 5:

$$\left(\sum_{n=0}^{M} ||\, \delta_n\, ||^2 \right)^{\frac{1}{2}} \leqq [\tfrac{1}{2}(\beta - \alpha) + \gamma] \left(\sum_{n=0}^{M} ||\, e_n\, ||^2 \right)^{\frac{1}{2}}.$$

Proof of Lemma 5:

We shall prove that

$$||\, \delta_n\, || \leqq [\tfrac{1}{2}(\beta - \alpha) + \gamma]\, ||\, e_n\, ||$$

for $n = 0, 1, \cdots, M$.

By definition

$$||\, \delta_n\, || = ||\, f(x_n\,, nh) - f(y_n\,, nh) - \tfrac{1}{2}(\alpha + \beta)(x_n - y_n)\, ||.$$

Let $q(a) = f[ax_n + (1 - a)y_n, nh]$ for $a \in [0, 1]$. Then

$$\frac{\partial q}{\partial a} = f'[ax_n + (1 - a)y_n\,, nh](x_n - y_n)$$

in which $f'[ax_n + (1 - a)y_n\,, nh]$ denotes the Jacobian matrix of $f(x, t)$, evaluated at $x = ax_n + (1 - a)y_n\,, t = nh$. Now, since $q(1) - q(0) = f(x_n\,, nh) - f(y_n\,, nh)$, we have

$$\int_0^1 \frac{\partial q}{\partial a}\, da = f(x_n\,, nh) - f(y_n\,, nh).$$

Therefore,

$$|| \delta_n || = \left|\left| \int_0^1 \{f'[ax_n + (1 - a)y_n , nh] - (\alpha + \beta)1_N\} \, da(x_n - y_n) \right|\right|$$

in which 1_N denotes the identity matrix of order N.

For H an arbitrary $N \times N$ matrix, let $|| H || \triangleq$ (largest eigenvalue of $H^*H)^{\frac{1}{2}}$, in which H^* denotes the complex-conjugate transpose of H (i.e., let $|| H ||$ denote the "spectral norm" of H). Then

$$\begin{aligned} || \delta_n || &\leqq \left|\left| \int_0^1 \{f'[ax_n + (1 - a)y_n , nh] - \tfrac{1}{2}(\alpha + \beta)1_N\} \, da \right|\right| \\ &\quad \cdot || x_n - y_n || \qquad (28) \\ &\leqq \int_0^1 || f'[ax_n + (1 - a)y_n , nh] - \tfrac{1}{2}(\alpha + \beta)1_N || \, da \, || e_n ||. \end{aligned}$$

With $\{f'\}_S$ and $\{f'\}_A$, respectively, the symmetric and antisymmetric parts of f', we have

$$|| f' - \tfrac{1}{2}(\alpha + \beta)1_N || \leqq || \{f'\}_S - \tfrac{1}{2}(\alpha + \beta)1_N || + || \{f'\}_A ||.$$

For each $a \varepsilon [0, 1]$, there exists[7] an orthogonal matrix T_1 such that $T_1\{f'\}_S T_1^{-1} \triangleq D = \text{diag}\,(\zeta_1 , \zeta_2 , \cdots \zeta_N)$ in which, since ζ_i is an eigenvalue of $\{f'\}_S$,

$$\alpha \leqq \zeta_i \leqq \beta$$

for $j = 1, 2, \cdots , N$. Using $|| T_1 || = || T_1^{-1} || = 1$,

$$\begin{aligned} || \{f'\}_S - \tfrac{1}{2}(\alpha + \beta)1_N || &= || T_1^{-1} DT_1 - \tfrac{1}{2}(\alpha + \beta)T_1^{-1}T_1 || \\ &\leqq || T_1^{-1} D - \tfrac{1}{2}(\alpha + \beta)T_1^{-1} || \cdot || T_1 || \\ &\leqq || T_1^{-1} || \cdot || D - \tfrac{1}{2}(\alpha + \beta)1_N || \cdot || T_1 || \\ &\leqq \max_i | \zeta_i - \tfrac{1}{2}(\alpha + \beta) \\ &\leqq \tfrac{1}{2}(\beta - \alpha). \qquad (29) \end{aligned}$$

Thus, $|| f' - \tfrac{1}{2}(\alpha + \beta)1_N || \leqq \tfrac{1}{2}(\beta - \alpha) + || \{f'\}_A ||$.

Consider $|| \{f'\}_A ||$. For each $a \varepsilon [0, 1]$ there exists[7] an orthogonal matrix T_2 such that $T_2\{f'\}_A T_2^{-1} \triangleq S$ is a direct sum of 2×2 block matrices of the form

$$B_i = \begin{bmatrix} 0 & b_i \\ -b_i & 0 \end{bmatrix}$$

and, if N is odd, and "1×1 matrix" containing the zero element. Clearly,

$$B_i^* B_i = \begin{bmatrix} b_i^2 & 0 \\ 0 & b_i^2 \end{bmatrix},$$

S^*S is a diagonal matrix, and its largest element is not greater than γ^2. That is,

$$|| \{f'\}_A || = || T_2^{-1} S T_2 || \leqq || S || \leqq \gamma,$$

and consequently

$$|| f' - \tfrac{1}{2}(\alpha + \beta) 1_N || \leqq \tfrac{1}{2}(\beta - \alpha) + \gamma \tag{30}$$

for all $a \varepsilon [0, 1]$. Finally, from (28) and (30)

$$|| \delta_n || \leqq [\tfrac{1}{2}(\beta - \alpha) + \gamma] \, || e_n ||.$$

At this point we have proved that with ρ as defined in Section 2.3

$$\left(\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} v_{n-k} \varphi_{k-1} \right|\right|^2 \right)^{\frac{1}{2}} \leqq (1 + \rho) \left(\sum_{n=0}^{M} || e_n ||^2 \right)^{\frac{1}{2}}. \tag{31}$$

We now need the following result.

Lemma 6:

$$\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} v_{n-k} \varphi_{k-1} \right|\right|^2 \geqq \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-2} \sum_{k=0}^{M} || \varphi_{k-1} ||^2.$$

Proof of Lemma 6:

Consider $S \triangleq \sum_{n=0}^{M} || \sum_{k=0}^{n} v_{n-k} c_k ||^2$, with the c_k's N-vectors. Choose $c_{M+1}, c_{M+2}, \cdots$ so that

$$\sum_{k=0}^{n} v_{n-k} c_k = 0, \qquad n \geqq (M + 1).$$

This is possible since $v_0 \neq 0$ [$v_0 = \lim_{z \to \infty} V(z)$]. Then

$$\begin{aligned} S &= \sum_{n=0}^{\infty} \left|\left| \sum_{k=0}^{n} v_{n-k} c_k \right|\right|^2 \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left|\left| \sum_{n=0}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} v_{n-k} c_k \right|\right|^2 d\omega. \end{aligned} \tag{32}$$

Under the *assumption* that

$$\sum_{k=0}^{\infty} || c_k ||^2 < \infty,$$

we can write

$$\sum_{n=0}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} v_{n-k} c_k = \sum_{n=0}^{\infty} e^{-i\omega n} v_n \sum_{k=0}^{\infty} e^{-i\omega k} c_k \tag{33}$$

in which the last sum over k is interpreted as the usual limit in the mean. From (32) and (33)

$$\begin{aligned} S &= \frac{1}{2\pi} \int_0^{2\pi} \left| \left| \sum_{n=0}^{\infty} e^{-i\omega n} v_n \sum_{k=0}^{\infty} e^{-i\omega k} c_k \right| \right|^2 d\omega \\ &\geqq \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-2} \frac{1}{2\pi} \int_0^{2\pi} \left| \left| \sum_{k=0}^{\infty} e^{-i\omega k} c_k \right| \right|^2 d\omega \\ &\geqq \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-2} \sum_{k=0}^{\infty} || c_k ||^2 \\ &\geqq \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-2} \sum_{k=0}^{M} || c_k ||^2 . \end{aligned} \tag{34}$$

We now *prove* that

$$\sum_{k=0}^{\infty} || c_k ||^2 < \infty .$$

Let

$$q_n \triangleq \sum_{k=0}^{n} v_{n-k} c_k , \qquad n \geqq 0.$$

Of course: $q_n = 0$ for $n \geqq (M + 1)$.

Let K be an arbitrary positive integer. Then

$$\sum_{n=0}^{K} e^{-in\omega} \sum_{k=0}^{n} v_{n-k} c_k = \sum_{k=0}^{K} e^{-in\omega} q_n .$$

With

$$\begin{aligned} (c_k)_K &= c_k , \qquad k \leqq K \\ &= 0, \qquad k > K \end{aligned}$$

we obtain

$$\sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} v_{n-k} (c_k)_K - \sum_{K+1}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} v_{n-k} (c_k)_K = \sum_{n=0}^{K} e^{-in\omega} q_n .$$

Therefore,

$$\begin{aligned} \sum_{k=0}^{K} e^{-ik\omega} c_k = &\left[1 - \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta) h b_k] e^{-i(k+1)\omega} \right] \\ &\cdot \sum_{K+1}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} v_{n-k} (c_k)_K + F(e^{i\omega}) \sum_{n=0}^{K} e^{-in\omega} q_n . \end{aligned} \tag{35}$$

The first term on the right side of (35) can be written as

$$\sum_{K+1}^{\infty} e^{-i\omega k} d_k .$$

Thus,

$$\sum_{k=0}^{K} || c_k ||^2 \leqq \sum_{k=0}^{K} || c_k ||^2 + \sum_{K+1}^{\infty} || d_k ||^2$$
$$= \frac{1}{2\pi} \int_0^{2\pi} \left|\left| F(e^{i\omega}) \sum_{n=0}^{K} e^{-in\omega} q_n \right|\right|^2 d\omega,$$

from which we obtain

$$\begin{aligned} \sum_{k=0}^{K} || c_k ||^2 &\leqq \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^2 \sum_{n=0}^{K} || q_n ||^2 \\ &\leqq \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^2 \sum_{n=0}^{M} || q_n ||^2. \end{aligned} \tag{36}$$

Since (36) holds for all $K > 0$, we have completed the proof of Lemma 6.*

Inequality (31) and Lemma 6 prove Theorem 1.

Proof of Theorem 2:

By Lemma 3 [inequality (27)], Lemma 4, and Lemma 5 of the preceding proof, we have

$$\left(\sum_{n=0}^{M} || e_n ||^2 \right)^{\frac{1}{2}} \leqq \rho \left(\sum_{n=0}^{M} || e_n ||^2 \right)^{\frac{1}{2}} + \left(\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} v_{n-k} \varphi_{k-1} \right|\right|^2 \right)^{\frac{1}{2}}. \tag{37}$$

Furthermore, by essentially the same argument used to prove Lemma 4, we find that

$$\left(\sum_{n=0}^{M} \left|\left| \sum_{k=0}^{n} v_{n-k} \varphi_{k-1} \right|\right|^2 \right)^{\frac{1}{2}} \leqq \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \left(\sum_{k=0}^{M} || \varphi_{k-1} ||^2 \right)^{\frac{1}{2}}.$$

Therefore, with $\rho < 1$,

$$\left(\sum_{n=0}^{M} || e_n ||^2 \right)^{\frac{1}{2}} \leqq (1 - \rho)^{-1} \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \left(\sum_{k=0}^{M} || \varphi_{k-1} ||^2 \right)^{\frac{1}{2}}.$$

We now prove that $\sup_{n \geqq 0} || \varphi_{n-1} || < \infty$ implies $\sup_{n \geqq 0} || e_n || < \infty$. Assume

* Alternatively, we can show using (35), that only a *finite* number of c_k are nonzero.

that $\sup_{n\geq 0} \| \varphi_{n-1} \| < \infty$. Let

$$u_n = \sum_{k=0}^{n} v_{n-k}\varphi_{k-1} .$$

Then we have

$$c_n = \sum_{k=0}^{n} w_{n-k}\, \delta_k + u_n , \qquad n = 0, 1, 2, \cdots , M \tag{38}$$

$$\delta_k = f(x_k , kh) - f(y_k , kh) - \tfrac{1}{2}(\alpha + \beta)(x_k - y_k) \tag{39}$$

in which, since $\{v_k\} \varepsilon l_1$, $\sup_{n\geq 0} \| u_n \| < \infty$.

There exist positive constants c_1 and c_2 such that $| w_n | \leqq c_1 e^{-c_2 n}$ for all $n \geqq 0$. By continuity, since $\rho < 1$, there exists $a\ \sigma \varepsilon (0, c_2)$ such that

$$\rho_\sigma \triangleq \max_{0\leq \omega \leq 2\pi} | W(e^{i\omega - \sigma}) | [\tfrac{1}{2}(\beta - \alpha) + \gamma] < 1$$

in which of course

$$W(e^{i\omega-\sigma}) = \sum_{n=0}^{\infty} w_n e^{-(i\omega-\sigma)n}.$$

From (38) and (39):

$$\tilde{e}_n = \sum_{k=0}^{n} \tilde{w}_{n-k}\, \tilde{\delta}_k + \tilde{u}_n , \qquad n = 0, 1, 2, \cdots , M$$

$$\tilde{\delta}_k = \tilde{f}(\tilde{x}_k , kh) - \tilde{f}(\tilde{y}_k , kh) - \tfrac{1}{2}(\alpha + \beta)(\tilde{x}_k - \tilde{y}_k)$$

where $\tilde{e}_n = e^{\sigma n} e_n$, $\tilde{w}_n = e^{\sigma n} w_n$, $\tilde{u}_n = e^{\sigma n} u_n$, $\tilde{\delta}_k = e^{\sigma k}\delta_k$, $\tilde{x}_k = e^{\sigma k} x_k$, $\tilde{y}_k = e^{\sigma k} y_k$, and $\tilde{f}(q, kh) = e^{\sigma k} f(e^{-\sigma k} q, kh)$ for all N-vectors q.

The Jacobian matrix $\tilde{f}'$ of $\tilde{f}$ is related to the Jacobian matrix of f by

$$\tilde{f}'(q, kh) = f'(e^{-\sigma k} q, kh)$$

from which we see that $\tilde{f}'$ satisfies the assumption concerning f' relative to the numbers α, β, and γ. Therefore, by the proof of Theorem 1,

$$\left(\sum_{n=0}^{M} \| \tilde{e}_n \|^2 \right)^{\frac{1}{2}} \leqq (1 - \rho_\sigma)^{-1} \left(\sum_{n=0}^{M} \| \tilde{u}_n \|^2 \right)^{\frac{1}{2}}$$

for all $M \geqq (p + 1)$.

Now,

$$\left(\sum_{n=0}^{M} \| \tilde{u}_n \|^2 \right)^{\frac{1}{2}} = \left(\sum_{n=0}^{M} \| e^{\sigma n} u_n \|^2 \right)^{\frac{1}{2}} \leqq \sup_{n\geq 0} \| u_n \| \left(\sum_{n=0}^{M} e^{2\sigma n} \right)^{\frac{1}{2}}$$

$$\leqq \sup_{n\geqq 0} || u_n || \left(\frac{e^{2\sigma(M+1)} - 1}{e^{2\sigma} - 1}\right)^{\frac{1}{2}}$$

$$\leqq \sup_{n\geqq 0} || u_n || \frac{e^{\sigma(M+1)}}{(e^{2\sigma} - 1)^{\frac{1}{2}}}$$

and so,

$$\left(\sum_{n=0}^{M} || \tilde{e}_n ||^2\right)^{\frac{1}{2}} \leqq (1 - \rho_\sigma)^{-1} \frac{e^{\sigma(M+1)}}{(e^{2\sigma} - 1)^{\frac{1}{2}}} \sup_{n\geqq 0} || u_n || \tag{40}$$

for all $M \geqq (p + 1)$.

From (38):

$$|| e_M || \leqq \left|\left| \sum_{k=0}^{M} w_{M-k}\, \delta_k \right|\right| + \sup_{n\geqq 0} || u_n ||. \tag{41}$$

We shall now use (40) to bound the first term on the right side of (41). Using the Schwarz inequality,

$$\begin{aligned} \left|\left| \sum_{n=0}^{M} w_{M-k}\, \delta_k \right|\right| &= e^{-\sigma M} \left|\left| \sum_{k=0}^{M} \tilde{w}_{M-k}\, \tilde{\delta}_k \right|\right| \\ &\leqq e^{-\sigma M}\left(\sum_{n=0}^{\infty} | \tilde{w}_n |^2\right)^{\frac{1}{2}}\left(\sum_{k=0}^{M} || \tilde{\delta}_k ||^2\right)^{\frac{1}{2}}. \end{aligned} \tag{42}$$

By the proof of Lemma 5,

$$|| \tilde{\delta}_k || \leqq [\tfrac{1}{2}(\beta - \alpha) + \gamma] || \tilde{e}_k ||,$$

which leads to

$$\left|\left| \sum_{k=0}^{M} w_{M-k}\, \delta_k \right|\right|$$

$$\leqq \frac{e^{\sigma}}{(e^{2\sigma} - 1)^{\frac{1}{2}}} [\tfrac{1}{2}(\beta - \alpha) + \gamma](1 - \rho_\sigma)^{-1} \sup_{n\geqq 0} || u_n || \left(\sum_{n=0}^{\infty} | \tilde{w}_n |^2\right)^{\frac{1}{2}}.$$

Since $\sigma \in (0, c_2)$ [see the paragraph below (39)], $|\tilde{w}_n|^2 \in l_1$, and therefore,

$$\sup_{M>(p+1)} \left|\left| \sum_{k=0}^{M} w_{M-k}\, \delta_k \right|\right| < \infty. \tag{43}$$

Finally, from (41) and (43), we have $\sup_{n\geqq 0} || e_n || < \infty$, which completes the proof of Theorem 2.

IV. A CONDITION FOR THE SOLVABILITY OF (2) FOR y_{n+1}

The problem of solving (2) for y_{n+1} is that of solving the equation

$$y + hb_{-1}f(y, t) = g \tag{44}$$

for y, with g a given N-vector. We write (44) as

$$Qy = g \tag{45}$$

in which the operator Q is defined by the condition that $Qv = v + hb_{-1}f(v, t)$ for all real N-vectors v.

We prove below that (with $\langle \cdot, \cdot \rangle$ denoting the usual inner product of real N-vectors):

$$\langle Qy_a - Qy_b , y_a - y_b \rangle \geqq k_1 \, || \, y_a - y_b \, ||^2 \tag{46}$$

$$|| \, Qy_a - Qy_b \, || \leqq k_2 \, || \, y_a - y_b \, || \tag{47}$$

for every pair of real N-vectors y_a and y_b, in which $k_1 = (1 + hb_{-1}\alpha)$ if $b_{-1} \geqq 0$, $k_1 = 1 + hb_{-1}\beta$ if $b_{-1} \leqq 0$: and $k_2 = \{1 + h \mid b_{-1} \mid \cdot [\max (|\, \alpha \,|, |\, \beta \,|) + \gamma]\}$. Since $k_2 < \infty$, according to a special case of Theorem I of Ref. 8, if $k_1 > 0$ (e.g., if $b_{-1} \geqq 0$ and $1 + hb_{-1}\alpha > 0$), then (45) possesses exactly one solution which can be determined by an iteration procedure that is only slightly more complicated than the usual procedure[1,2] (which is valid only under much stronger conditions on h).

To derive (46), let

$$q(\eta) = \eta y_a + (1 - \eta)y_b + hb_{-1}f[\eta y_a + (1 - \eta)y_b , t]$$

for $\eta \in [0, 1]$. Then

$$q'(\eta) = (y_a - y_b) + hb_{-1}f'[\eta y_a + (1 - \eta)y_b , t](y_a - y_b),$$

and so

$$\begin{aligned} Qy_a - Qy_b &= \int_0^1 q'(\eta) \, d\eta \\ &= \int_0^1 \{1_N + hb_{-1}f'[\eta y_a + (1 - \eta)y_b , t]\}(y_a - y_b) \, d\eta. \end{aligned} \tag{48}$$

Thus,

$$\langle Qy_a - Qy_b , y_a - y_b \rangle = \left\langle \int_0^1 \{1_N + hb_{-1}f'[\eta y_a + (1 - \eta)y_b , t]\} \, d\eta (y_a - y_b), (y_a - y_b) \right\rangle$$

$$= \int_0^1 \langle \{1_N + hb_{-1}f'[\eta y_a + (1 - \eta)y_b , t]\}(y_a - y_b), (y_a - y_b)\rangle \, d\eta$$

$$= || y_a - y_b ||^2 + hb_{-1} \int_0^1 \langle f'_S[\eta y_a + (1 - \eta)y_b , t](y_a - y_b), (y_a - y_b)\rangle \, d\eta,$$

in which f'_S denotes the symmetric part of f'. Thus, since the eigenvalues of f'_S are bounded from below by α, and from above by β:

$$\langle Qy_a - Qy_b , y_a - y_b\rangle \geqq (1 + hb_{-1}\alpha) || y_a - y_b ||^2, \qquad b_{-1} \geqq 0$$
$$\geqq (1 + hb_{-1}\beta) || y_a - y_b ||^2, \qquad b_{-1} \leqq 0.$$

Consider now the derivation of (47). By (48),

$$|| Qy_a - Qy_b ||$$
$$= \left|\left| \int_0^1 \{1_N + hb_{-1}f'[\eta y_a + (1 - \eta)y_b , t]\} \, d\eta (y_a - y_b) \right|\right|$$
$$\leqq \left|\left| \int_0^1 \{1_N + hb_{-1}f'[\eta y_a + (1 - \eta)y_b , t]\} \, d\eta \right|\right| \cdot || y_a - y_b ||$$
$$\leqq \int_0^1 || 1_N + hb_{-1}f'[\eta y_a + (1 - \eta)y_b , t] || \, d\eta \, || y_a - y_b ||.$$

But, with f'_A the antisymmetric part of f',

$$|| 1_N + hb_{-1}f'[\eta y_a + (1 - \eta)y_b , t] ||$$
$$\leqq 1 + h \mid b_{-1} \mid || f'[\eta y_a + (1 - \eta)y_b , t] ||$$
$$\leqq 1 + h \mid b_{-1} \mid \cdot || f'_S || + h \mid b_{-1} \mid \cdot || f'_A ||$$
$$\leqq 1 + h \mid b_{-1} \mid \max (\mid \alpha \mid, \mid \beta \mid) + h \mid b_{-1} \mid \gamma.$$

Therefore,

$$|| Qy_a - Qy_b || \leqq \{1 + h \mid b_{-1} \mid [\max (\mid \alpha \mid, \mid \beta \mid) + \gamma]\} || y_a - y_b ||$$

which is equivalent to (47).

REFERENCES

1. Ralston, A., *First Course in Numerical Analysis,* McGraw-Hill Book Co., New York, 1965.
2. Hamming, R. W., *Numerical Methods for Scientists and Engineers,* McGraw-Hill Book Co., New York, 1962.
3. Henrici, P., *Error Propagation for Difference Methods,* Wiley & Sons, Inc., New York, 1963.

4. Hildebrand, F. B., *Introduction to Numerical Analysis,* McGraw-Hill Book Co., New York, 1956.
5. Rall, L. B. (editor), *Error in Digital Computation,* Volumes 1 and 2, Wiley & Sons, Inc., New York, 1965.
6. Sandberg, I. W., On the Theory of Physical Systems Governed by Nonlinear Functional Equations, B.S.T.J., *44,* May-June, 1965, p. 871.
7. Macduffee, C. C., *The Theory of Matrices,* Chelsea Publishing Co., New York, 1956.
8. Sandberg, I. W., On the Properties of Some Systems that Distort Signals—I. B.S.T.J., *42,* September, 1963, p. 2033.

Design Considerations for a Semipermanent Optical Memory

By F. M. SMITS and L. E. GALLAHER

(Manuscript received April 7, 1967)

The potential of high-speed optical memories using electro-optic or acousto-optic light deflection for address selection is examined. It is shown that for such memories the total memory capacity decreases as the third power of the addressing rate and that capacities in excess of 10^8 bits are feasible with a random access rate of 10^6 addresses/sec.

A specific semipermanent memory design is then described which uses a laser light source, an acoustic xy light deflector and an array of 10^4 holograms as information storage elements. Each storage element contains 10^4 bits which appear as a pattern on a semiconductor read-out matrix when the storage element is illuminated through the xy deflector. Accordingly, the system has a total capacity of 10^8 bits with an access time of less than 10 μsec.

I. INTRODUCTION

The increased computational capabilities of modern day computers are accompanied by a sharply increased need for a semipermanent memory to store their base programs. The capacity requirements for a given base program memory generally increase as fast or perhaps faster than the computational complexity of the computer increases.

Along with the increased capacity, the next generations of memories must have increased speed capabilities since the speed of the program memory generally controls the overall computer processing speed. Ideally, a program memory should have a random access capability to each instruction or to each group of instructions if the instructions can be suitably grouped.

If individual instructions are accessed in one memory cycle, the speed of the program memory has to match the rate at which the desired instructions have to be made available to the logic or processor frame. If, however, a group of instructions can be obtained in one access

operation in the program memory, then the number of groups which must be accessed per unit time is reduced by the number of desired instructions contained within each group. This concept is usually referred to as paging; that is, each access operation within the memory selects a page, or set of instructions, rather than a single instruction. Operations within the memory output circuits then transfer the desired instruction from the accessed page, or pages, to the logic frame.

The twofold requirements of high speed and large capacity for a semipermanent memory suggests a review of the status and the capabilities of optical memories, particularly in view of the progress made in coherent optics within the past few years. The high speed requirements restrict the considerations to systems with access times well below 1 msec which make mechanical motion in the address selection impractical. The flying spot store was the first fully-developed optical memory with a short access time in the order of a few microseconds.[1] Recent publications in the field include studies of electro-optical systems,[2,3,4,5,6,7] holographic systems,[8,9] magneto-optical systems,[5,10,11,12] and ferro-electric optical systems.[13]

With a view towards high speed and high capacity a system using a laser light source and solid-state light deflectors appears particularly promising. This combination permits one to direct highly collimated and coherent light beams to a given address. Two suitable methods for deflecting light are available. They are the digital light deflector[14,15] and the acousto-optic deflector.[2] The high light intensity available from lasers, in conjunction with semiconductor detector arrays, makes it feasible to increase substantially the number of bits stored at each address as is required for page organization.

In order to assess the potential usefulness of such a system, the achievable speed and memory capacity should be examined on as general a basis as possible. Such an assessment is presented in the following section of this paper. To illustrate the potential of optical semipermanent stores this is followed by a description of a specific system which is based entirely on present technology and which has a storage capacity of 10^8 bits, divided into pages of 10^4 bits each with a page access time of less than 10 μsec.

II. STORAGE CAPACITY AND SPEED POTENTIAL

The total number of bits that can be stored in an optical memory of the type discussed here has an upper limit that depends on the speed of the memory. This limit is set by physical limitations in the address-

ing rate as a function of the total number of addresses and by the number of bits at each address which for a given interrogation rate is limited by the light intensity.

For both methods of light deflection the ultimate limit in the addressing system can be expressed in terms of a number for the capacity speed product CSP $= N_a^{\frac{1}{2}} \nu_a$, where N_a is the total number of addresses and ν_a is the rate (addresses/sec) of random addressing.* At low addressing rates an upper limit on N_a is set by the diffraction limit.

Kurtz[16] discussed the capacity speed product for the electro-optic deflector in the form described by Tabor.[15] For a linear electro-optic material he obtains

$$\mathrm{CSP} = N^{\frac{1}{2}}\nu_a = \frac{6}{100\lambda\epsilon V_\pi^2\left[1 + \frac{(S_m - S)}{S_m}\right]} P_d ,$$

where the symbols have the following meaning:

- P_d reactive drive power
- ϵ dielectric constant of the electro-optic material
- V_π reduced half-wave voltage, that is the voltage which for an electrode spacing equal to the length produces half-wave phase retardation.
- λ the wave length of the light
- S_m The length of a module consisting of a polarization switch of length S and a Wollaston prism of length $S_m - S$.

In this formula, it is assumed that the smallest angular increment between addresses is given by $5\lambda/d$ with d the width of the aperture. This assumption allows for a $\lambda/4$ wavefront distortion, the aperturing due to the "walkoff",† and a guardband of $1.5\lambda/d$. It should correspond to about −20-dB crosstalk.

Obviously the driver power should be as large as possible. The only inherent limit is given by the heating of the electro-optic material which is a consequence of the finite Q of the crystal. For most linear electro-optic materials, heating does not present a serious problem.

The situation is more complicated for the quadratic electro-optic material KTN. Kurtz derived the capacity speed product for that case and particularly considered the limitations due to heating since

* The treatment of the capacity speed product draws heavily on an unpublished summary of the subject by K. D. Bowers.

† The light entering all but the first module has undergone at least some deflection and thus, "walked off" the optical axis.

the material has to be operated very close to its Curie temperature so that small changes in temperature will cause a large change in the electrical permittivity, hence in the polarization, hence in the electrically induced birefringence.

For the evaluation of specific cases the following assumptions are made:

driver power $P_d = 10\ W$
wave length of light $\lambda = 0.633\mu$
$S_m = 1.2$ S.

The resulting capacity speed products are listed in Table I. For KDP the listed values are for the longitudinal electro-optic effect since in that case it is larger than the transverse electro-optic effect. The values for all other materials are for the transverse electro-optic effect. The value listed for KTN is taken from Kurtz, but adjusted for the higher wavelength.

It can be seen that the highest capacity speed product achievable with presently known materials is on the order of 10^9 sec^{-1}.

The acousto-optic deflection of light has recently been reviewed by Gordon.[2] In this method, light is diffracted on the modulation of the refractive index produced by an acoustic wave. The deflection angle is changed by changing the sound frequency. The range of angular deflection is given by

$$\Delta\theta = (\lambda/v_s)\cdot\Delta f_s\ ,$$

where v_s is the sound velocity and Δf_s is the range of the acoustic frequencies available.

TABLE I

Material	$\kappa = \epsilon/\epsilon_0$	V_π (10^3 volt)	κV_π^2 (10^3 volt^2)	CSP (10^9 sec^{-1})
Linear electro-optic effect				
KDP[17]	20	8	13	0.07
CuCl[18]	8	7.2	4	0.23
$LiNbO_3$[19,20]	32	2.8	2.5	0.37
$LiTaO_3$[21]	43	2.4	2.5	0.37
ZnTe[22]	10	2.9	0.84	1.1
Quadratic electro-optic effect				
KTN[23]	1400	0.054	0.4	0.8

Since the incident light is apertured to the width d of the deflector, the emerging beam will have an angular spread due to diffraction. In order to have well-resolved output positions an angular separation between neighboring addresses of $4\lambda/d$ is assumed which allows for a $\lambda/4$ wave front distortion and for a guardband of $1.5\lambda/d$. This separation is somewhat less than was assumed in the electro-optic case where an additional allowance had to be made for the "walkoff." The number of resolvable addresses in one coordinate is thus obtained as

$$n = \Delta\theta/(4\lambda/d) = (\Delta f_s/4)\cdot(d/v_s).$$

The quantity d/v_s is the transit time for the acoustic energy across the optical aperture, that is, the time needed to change the frequency in the deflector, or in other words $v_s/d = \nu_a$. If the X deflector is followed by a similar Y deflector and the two are operated simultaneously, the total number of resolvable addresses is

$$N = n^2 = (\Delta f_s/4)^2(d/v_s)^2 = (\Delta f_s/4)^2/\nu_a^2 .$$

Therefore, the capacity speed product is

$$\text{CSP} = N^{\frac{1}{2}}\nu_a = \Delta f_s/4.$$

Considering the present state of the technology it is justified to assume $\Delta f_s = 400$ MHz which gives a capacity speed product of 10^8 sec^{-1}.

For both deflectors, the maximum number of addresses is limited by the optics of the system in conjunction with the practical limit on the size of the deflector element. This limit is estimated at 10^6 addresses. In the case of the electro-optic deflector this corresponds to an aperture of a few centimeters for the polarization switches, a size which will be difficult to exceed for electro-optic materials of adequate optical quality. On the other hand, in the acoustic case the generally required cylindrical optics makes it difficult to exceed the above limit.

The resulting limitations in the number of addresses versus addressing rate are depicted in Fig. 1 (a) and (b).

If one stores M bits at each address, the laser energy reaching this address has to be distributed over the M bits. In order to detect any one bit, a certain minimum amount of light energy (number of photons) is needed at the detector for any desired signal-to-noise ratio. For a given light power reaching a given address, this sets an upper limit to the number of bits that can be accommodated at each address.

If an ideal photon detector were used, the noise limit would correspond to 1 photon per bit reaching a detector in the time interval which

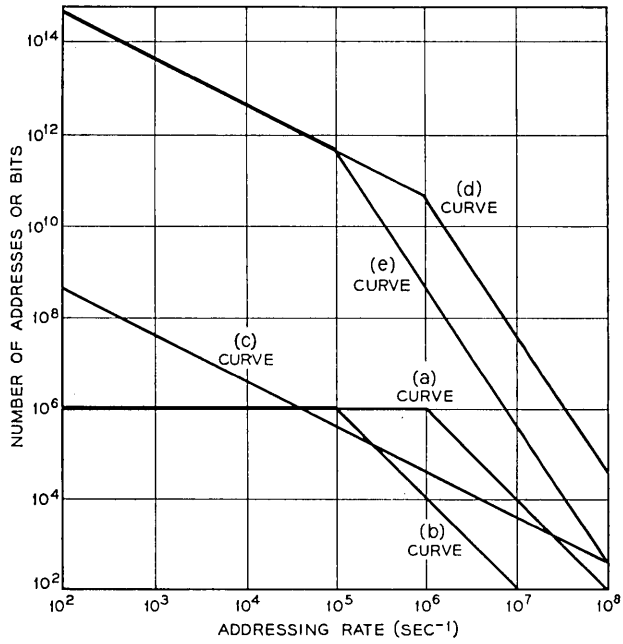


Fig. 1 — Capacity speed limitations in an optical memory. (a) Number of addresses vs addressing rate for the electro-optic deflector, (b) Number of addresses vs addressing rate for the acousto-optic deflector. (c) Number of bits per address vs addressing rate. (d) Number of total bits vs addressing rate for electro-optic deflector. (e) Number of total bits vs addressing rate for acousto-optic deflector.

is consistent with the desired memory cycle time. With a multiplying p-n junction, the noise limit is approximately 10^2 photons per bit and with a non-multiplying junction, the corresponding limit is 10^4 photons.[24] For an adequate signal-to-noise ratio, the number of photons reaching the detector should be 10^2 times the number corresponding to the noise limit.

For an array of photodetectors, non-multiplying junctions offer the greatest economy and flexibility. Thus, a minimum of 10^6 photons per bit should reach the detector in the time in which the read-out is to be performed.

At present, simple, reliable lasers operating in the visible range are available which emit in excess of 100 mW in continuous operation. In the light deflection and the projection onto the photodetectors, a fraction of the light intensity will be unavoidably lost. In a practical

optical system it will be very difficult to keep these losses below 10 dB. Accordingly, it can be assumed that 10 mW will be available at the detectors. If one further assumes that the rate at which bits are to be read out from any one detector should equal the addressing rate, an inverse relation between the number of bits that can be paralleled at one address and the bit rate per detector can be determined. Fig. 1(c) gives the resulting relation for a photon energy of 1.5 eV and 10^6 photons per bit.

The product between Fig. 1(c) and Fig.1(a) and (b), respectively, gives the overall limit of the total bit capacity. Fig. 1(d) gives this product for electro-optical deflection while Fig. 1(e) gives it for acousto-optical deflection. These curves indicate that for high addressing rates the resulting limit of the total capacity decreases with the third power of that rate.

The curves permit a general assessment of the potential usefulness of optical memories using present state of the art components. Designing a memory that operates near its physical limits increases the complexity and cost; a memory designed for an addressing rate and capacity which are well removed from these limits can be built much more readily.

For a memory capacity of 10^8 bits or more, the curves of Fig. 1 indicate that, for such a capacity, addressing rates of 10^6 addresses per second are fairly close to the physical limitations. However, if the concept of paging is used and the number of desired instructions per address is a number between 1 and 100, depending on the program design, then a range of addressing rates of 10^4 to 10^6 addresses per second gives a net rate approaching 10^6 instructions per second.

III. DESIGN CONSIDERATIONS FOR SYSTEM USING AN ACOUSTO-OPTIC DEFLECTOR

A system realizing the above concept utilizes a laser light source and an XY-deflector which directs the light to any one address in a matrix of storage elements as depicted in Fig. 2. Each storage element has associated with it an optical system through which the stored information is projected onto a common read-out matrix. Thus, upon illumination of a selected address a real image of an information matrix is generated in the read-out plane. Whether a matrix point receives light or not corresponds to a logical "1" or a logical "0". In the read-out plane a matrix of photodiodes or other light sensitive detectors will convert the light into electrical signals. Therefore, by illuminating

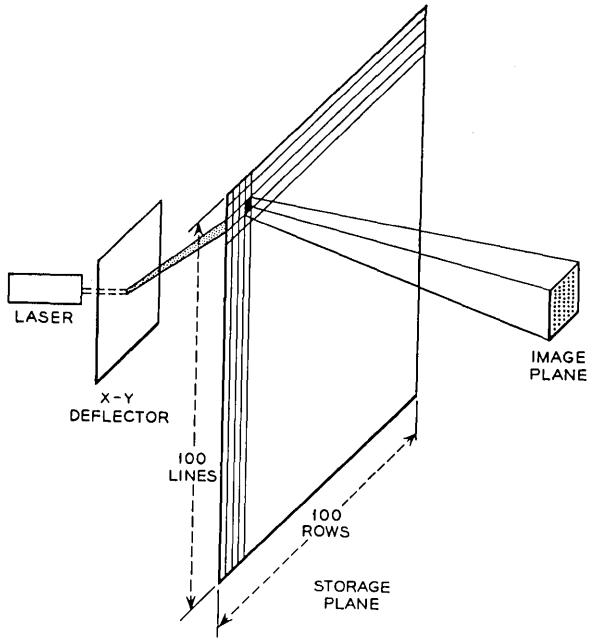


Fig. 2 — Schematic of optical memory system.

any one storage element with laser light, one transfers a large block of information from the storage location to the location of the read-out matrix.

At present, acousto-optic deflection is the most practical method of light deflection. Specifically, an addressing rate of 10^5 addresses per second and a total of 10^4 addresses is well within the range of available technology.[2] By storing 10^4 bits per address a total storage capacity of 10^8 bits is possible. Electro-optic light deflectors may ultimately replace the acousto-optic deflector since, theoretically, the former has a higher capacity speed product. However, due to materials problems the electro-optic light deflector is currently limited to speeds no higher than those achievable with the simpler acousto-optic deflector.

The following more specific considerations are based on the above configuration. It will become clear that the design is sufficiently far from all inherent limitations so that it will be possible to increase both capacity and speed through appropriate refinements within the established technology.

In the most straight forward approach a real image of the informa-

tion matrix is stored in the storage plane. The projection onto the read-out plane requires a separate lens system associated with each storage location, the resolution of which is dictated by the number of bits stored at each location. It might be possible that such a multi-lens system is realizable with molded plastic fly eye lenses.

To keep the overall size within reasonable limits, the individual real image will be quite small. This would make the system very sensitive to dust particles and alignment.

Since the use of coherent light has already been assumed, it is readily possible to use holograms as optical elements. In that case the lens and the storage function can be combined by preparing a hologram which upon illumination by the deflector generates a real image of an information matrix in the read-out plane. Within each hologram, the information about each bit is spread over the entire area,[25] so that the sensitivity to dust particles is significantly reduced. The presence of dust in that case gives an overall reduction of the signal-to-noise for all bits stored in one hologram. Uniformity of illumination is also not very critical. Furthermore, the exact positioning of the individual hologram is not very critical. The location of the image, however, will depend on the angular orientation of the hologram. With a 100×100 read-out matrix it will be adequate to control the angular positions of the individual holograms to one part in 1,000, a precision which is quite practical.

In the hologram case a positional requirement exists between the laser and the read-out matrix. It can be expected that the alignment of the laser beam with regard to the read-out matrix can be much more readily accomplished than the alignment of the picture in the case where the information is stored as a real image. In particular, with an acoustic deflector that portion of the light which emerges undeflected might be used in a positional servo-mechanism.

In the hologram case the complications lie primarily in the fabrication of the holograms. It appears feasible to store the required 10^8 bits of information on 10^4 individual holograms within an overall size of 2×2 inches.

The laser to be used in such a system should emit light of a wavelength of less than one micron so that silicon devices can be used as photodetectors. It is also desirable that the laser be operable in a pulsed mode.

The best operation would be achieved with a Q-switched solid state laser, since, in such lasers, power can be integrated for up to approximately 10^{-4} seconds and the energy so stored can be released in pulses

as short as 10^{-7} seconds.[26] Most solid-state detectors will operate with a minority-carrier lifetime of less than 10^{-6} seconds. In such devices, an optimum signal is obtained if the light pulse is short in comparison to the lifetime. A Q-switched solid-state laser will thus permit an optimum utilization of available laser intensity, since it is capable of integrating power up to the longest cycle times of interest and since it can discharge the energy in a time shorter than, or at worst comparable to, the lifetime in typical semiconductor devices.

In the simplest case the read-out matrix will consist of individual photodiodes located at each matrix point whereby all diodes can be read out simultaneously. For the specific case considered here, this requires 10^4 diodes and 10^4 read-out circuits. These circuits could be bi-stable elements such as flip-flops which are set by the photodiode.

As an alternative to a diode with a flip-flop, a pnpn diode in series with a resistor might also be used as a bi-stable element in the read-out matrix. The light could switch such a device from the low-current high-impedance state to a high-current low-impedance state.

The state of the individual bi-stable circuits could be interrogated through a multicoordinate address selection. Once the elements are set, the interrogation speed is only determined by semiconductor device considerations and it should be readily possible to interrogate "words" within the "page" with access and cycle times well below 100 nsec.

In this organization one is really dealing with a small (10^4 bit) semiconductor memory which is set by the optical system. This small semiconductor memory may be operated as a buffer store which can hold the information of one page for any desired period of time.

In other applications it may be necessary to have more than one page available in a random access high-speed buffer store. This requires that all information be rapidly transferred from the read-out matrix into a buffer store. In this case, the time for which the read-out matrix is to hold the information may become quite short and other methods of temporary information storage may be considered such as holding the information as a voltage on a capacitor. A particularly suitable method might be the use of photo transistors operated in the charge storage mode.[27] In that mode the light-generated charge is stored in a reverse biased junction for times much longer than a lifetime. The information is interrogated by applying a bias pulse to the transistor.

The charge storage mode has the additional advantage in that it can be used to integrate light over fairly long times making it possible to use CW lasers such as HeNe lasers with such detectors. This, how-

ever, leads to a somewhat less efficient operation than with a Q-switched solid-state laser since the CW laser will not integrate intensity during the time a new address is established.

IV. CONCLUSIONS

The preceding considerations lead to the conclusion that a semipermanent optical memory utilizing either acousto-optic or electro-optic light deflectors shows potential as a memory with 10^8 or more bits capacity and with page access as short as 1 μsec. Specifically, a memory with 10^8 bits capacity using acousto-optic deflectors can easily have page access times of 10 μsec. Such a memory using holograms as a storage medium and semiconductor devices as read-out elements appears well within the state of the art. Significantly shorter access times for similar capacities will probably require electro-optic deflectors using materials which are not yet available in adequate optical quality.

V. ACKNOWLEDGMENT

Valuable discussions with several of our colleagues, particularly with K. D. Bowers and L. K. Anderson, are gratefully acknowledged.

REFERENCES

1. Hoover, C. W., et al., Fundamental Concepts in the Design of the Flying Spot Store, B.S.T.J. *37,* September, 1958, p. 1161.
2. Gordon, E. I., A Review of Acousto-optical Deflection and Modulation Devices, Appl. Opt., *5,* 1966, p. 1629.
3. Kulcke, W., et al., Digital Light Deflectors, Appl. Opt. *5,* October, 1966, p. 1657.
4. Korpel, A., et al., A Television Display Using Acoustic Deflection and Modulation of Coherent Light, Appl. Opt., *5,* October, 1966, p. 1667.
5. Smith, D. O. and Harte, K. J., Content Addressed Memory Using Magneto- or Electro-optic Interrogation, IEEE Trans. Electron. Comp., *EC-15,* February, 1966, p. 123.
6. Poppelbaum, W. J., Computer Application of Electro-optics, Proc. Spring Joint Computer Conf., *28,* Spartan Books, Washington, D.C., 1966, p. 1.
7. Soref, R. A. and McMahon, D. H., Bright Hopes for Display Systems, Flat Panels, and Light Deflectors, Electron. *38,* November 29, 1965, p. 56.
8. Van Heerden, J., Theory of Optical Information Storage in Solids, Appl. Opt. *2,* April, 1963, p. 393.
9. Vitols, V. A., Hologram Memory for Storing Digital Data, IBM Tech. Disclosure Bull., *8,* April 1966, p. 1581.
10. Eide, J. E., *et al.,* Magneto-optical Display Panel with Memory, AD-629 586, General Electric Co., Electronics Lab., Syracuse, N.Y., under contract DA-28-043-AMC-01442(E), January, 1966.
11. Shafer, W., Suits, J. C., and Toxen, A. M., Optical Readout Using Europium Fluoride, IBM Tech. Disclosure Bull., *8,* December, 1965, p. 989.
12. Mee, C. D., Magneto-optical Readout Technique, IBM Tech. Disclosure Bull., *9,* February, 1967, p. 1155.

13. Smith, A. W., Ferroelectric Optical Switch, IBM Tech. Disclosure Bull., *9,* July, 1966, p. 180.
14. Nelson, T. J., Digital Light Deflection B.S.T.J., *43,* May, 1964, p. 821.
15. Tabor, W. J., A High-Capacity Digital Light Deflector Using Wollaston Prisms, B.S.T.J., *46,* May-June, 1967, p. 957.
16. Kurtz, S. K., Design of an Electro-Optic Polarization Switch for High-Capacity High-Speed Digital Light Deflection System, B.S.T.J., *45,* October, 1966, p. 1209.
17. Carpenter, R. O'B., The Electro-Optic Effect in Uniaxial Crystals of the Dihydrogen Phosphate Type. III Measurement of Coefficients, J. Opt. Soc. Am., *40,* 1950, p. 225.
18. West, C. D., Electro-Optic and Related Properties of Crystals with the Zink Blende Structure, J. Opt. Soc. Am., *43,* 1953, p. 335.
19. Lenzo, P. V., Spencer, E. G. and Nassau, K., Electro-Optic Coefficients in Single-Domain Ferroelectric Lithium Niobate, J. Opt. Soc. Am., *56,* 1966, p. 633.
20. Turner, E. H., High-Frequency Electro-Optic Coefficients of Lithium Niobate, Appl. Phys. Lett., *8,* 1966, p. 303.
21. Lenzo, P. V., Turner, E. H., Spencer, E. G., and Ballman, A. A., Electro-Optic Coefficients and Elastic-Wave Propagation in Single-Domain Ferroelectric Lithium Tantalate, Appl. Phys. Lett., *8,* 1966, p. 81.
22. Sliker, T. R. and Jost, J. M., Linear Electro-Optic Effect and Refractive Indices of Cubic ZnTe, J. Opt. Soc. Am., *56,* 1966, p. 130.
23. Chen, F. S., Geusic, J. E., Kurtz, S. K., Skinner, J. G., and Wemple, S. H., Light Modulation and Beam Deflection with Potassium Tantalate-Niobate Crystals, J. Appl. Phys., *37,* 1966, p. 388.
24. Anderson, L. K. and McMurtry, B. J., High-Speed Photo-detectors, Appl. Opt., *5,* October, 1966, p. 1573.
25. Leith, E. N. and Upatnicks, J., Wavefront Reconstruction with Diffused Illumination and Three-Dimensional Objects, J. Opt. Soc. Am., *54,* 1964, p. 1295.
26. Kiss, Z. J. and Pressley, R. J., Crystalline Solid Lasers, Proc. IEEE, *54,* 1966, p. 1236.
27. Weckler, G. P., Storage Mode Operation of a Phototransistor and its Adaptation to Integrated Arrays for Image Detection, Paper 4.6 IEEE Inter. Electron Devices Meeting, Washington, D. C., October 26–28, 1966.

Contributors to This Issue

LEE E. GALLAHER, B.S.E.E., 1951, and M.S.E.E., 1956, Case Institute of Technology; Bell Telephone Laboratories, 1955—. Mr. Gallaher first worked on the design of the flying-spot store for the Morris experimental central office. He later worked on the program stores for No. 1 ESS and is currently concerned with the design of memories and networks for future ESS systems. Member, Sigma Xi, Tau Beta Pi.

JAMES W. GEWARTOWSKI, B.S., 1952, Illinois Institute of Technology; S.M., 1953, Massachusetts Institute of Technology; Ph.D., 1958, Stanford University; Bell Telephone Laboratories, 1957—. Mr. Gewartowski was initially concerned with the development of high-power microwave tubes and electron guns. Since 1962 he has supervised a group concerned with varactor harmonic generators and upconverters and avalanche transit-time diode oscillators. He is co-author of the book, *Principles of Electron Tubes,* (Van Nostrand, 1965). Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

A. GOETZBERGER, Ph.D. in Science, 1955, University of Munich; Bell Telephone Laboratories, 1963—. Mr. Goetzberger is a supervisor in the metal insulator semiconductor group. Prior to 1963, he was with the Shockley Laboratory in Palo Alto, where he worked on junction imperfections and avalanche breakdown phenomena in silicon. He also participated in the development of a power transistor. Member, American Physical Society, IEEE, Electrochemical Society.

W. M. HUBBARD, B.S., 1957, Georgia Institute of Technology; M.S., 1958, University of Illinois; Ph.D., 1963, Georgia Institute of Technology; Bell Telephone Laboratories, 1963—. Mr. Hubbard's work has included analyses related to the design of millimeter-wave solid-state repeaters for use in a waveguide transmission system and the construction of prototype high-speed repeaters for this type of system. Member, Sigma Xi, Tau Beta Pi, Phi Kappa Phi, American Physical Society.

BELA JULESZ, Dipl. in Electrical Engineering, 1950, Budapest (Hungary) Technical University, Kandidat in Technical Sciences, 1956, Hungarian Academy of Sciences; Telecommunication Research Institute (Budapest) 1950–56; Bell Telephone Laboratories, 1956—. Mr. Julesz first taught and did research in communication systems and his thesis work reflected his later interest in analyzing and processing pictorial information. At the Laboratories he was first engaged in studies of systems for reducing television bandwidth. Since 1959 he has devoted full time to visual research, particularly in depth perception and pattern recognition, about which he has written extensively. Since 1964 Dr. Julesz is Head of the Sensory and Perceptual Processes Department, responsible for research in visual psychology and neurophysiology. Member, IEEE, AAAS, Psychonomic Society, Optical Society of America.

EDMUND T. KLEMMER, B.S., 1944, Webb Institute of Naval Architecture; M.A., 1949, Ph.D., 1952, Columbia University; Bell Telephone Laboratories, 1962—. Mr. Klemmer has studied customer dialing behavior, subjective evaluation methods, developed a preference scaling method, and measured the subjective quality of satellite circuits. Presently, he is responsible for studies of human performance in using the TOUCH-TONE® telephone for entering data into computer systems. Fellow, American Psychological Association; Member, Human Factors Society, Sigma Xi.

G. D. MANDEVILLE, 1933–34, Monmouth Junior College; 1935–36, Rutgers University; Western Electric Co., 1939–49; Bell Telephone Laboratories, 1949—. With Western Electric, Mr. Mandeville was concerned with radar development and shop test equipment. He headed the shop test equipment prove-in section for three years. With Bell Laboratories he has been associated with guided-wave research in the areas of waveguide and repeaters.

RICHARD H. MINETTI, B.S.E.E., 1966, Newark College of Engineering, evening division; Bell Telephone Laboratories, 1958—. Mr. Minetti has been engaged in development work on microwave tubes and semiconductor devices and is at present concerned with the characterization of avalanche transit time diodes.

E. H. NICOLLIAN, M.E., 1951, Stevens Institute of Technology; M.A. (Physics) 1956, Columbia University; Bell Telephone Laboratories, 1957—. Mr. Nicollian's work has been in semiconductor device physics. He is currently engaged in research on the electrical properties of

semiconductor-insulator interfaces. Member, American Physical Society, Electrochemical Society, RESA, AAAS.

R. M. RYDER, B.S., 1937, Ph.D. (Physics), 1940, Yale University; Bell Telephone Laboratories, 1940—. Mr. Ryder's work since 1948 has been in the area of transistor and other semiconductor device development, including varactor diodes for low-noise receivers, varactors for microwave power generation, high-speed switching diodes, microwave protectors, amplifiers, etc. He is now Department Head in charge of exploratory transistors and integrated circuits. Fellow, IEEE.

IRWIN W. SANDBERG, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of military systems, synthesis and analysis of active and time-varying networks, studies of properties of nonlinear systems, and with a few problems in communication theory. His current interests are in the area of numerical analysis. Member, IEEE, SIAM, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

FRIEDOLF M. SMITS, Dipl. Phys., 1950, Dr. rer. nat., 1950, University of Freiburg, Germany; research associate, Physikalisches Institut, University of Freiburg, 1950–54; Bell Telephone Laboratories, 1954–62; Sandia Corporation, 1962–65; Bell Telephone Laboratories, 1965—. Mr. Smits early work at Bell Telephone Laboratories includes studies of solid-state diffusion in germanium and silicon, and exploratory semiconductor device development. He supervised a group that conducted radiation damage studies on components, particularly solar cells, used in the Telstar satellite. At Sandia Corporation he was responsible for work on radiation effects particularly electron and neutron damage to semiconductors and semiconductor devices. His recent responsibility at Bell Telephone Laboratories is in the field of ultrasonic materials and ultrasonic devices including acousto-optic devices. Senior Member, IEEE; Member, American Physical Society, German Physical Society.

S. M. SZE, B.S., 1957, National Taiwan University, China; M.S., 1960, University of Washington; Ph.D., 1963, Stanford University; Bell Telephone Laboratories, 1963—. Mr. Sze has been concerned with semiconductor device physics and technology. At present, he is concerned with the studies of metal-insulator-semiconductor systems and metal-semiconductor barrier devices. Member, Sigma Xi, IEEE.

B. S.T. J. BRIEFS

Estimation of the Variance of a Stationary Gaussian Random Process by Periodic Sampling

By J. C. DALE

(Manuscript received February 14, 1967)

I. INTRODUCTION

This paper* applies previous work[1] on estimation of the mean of a stationary random process by periodic sampling to estimation of the variance with the added restriction that the process under consideration be Gaussian with known mean.

The samples are taken from a sample function of the random process, in a closed interval $(0, T)$ and are in general correlated. The estimator used is the average of equally-weighted squared samples. The variance of this estimator is derived and its behavior is predicted as a function of the number of samples and length of record.

II. THEORY

2.1 *General*

Let $x(t)$ be a sample function from a stationary, Gaussian random process $\{x(t)\}$ with known mean.†

An unbiased estimator of the variance is given by

$$\hat{\sigma}_x^2 = \frac{1}{N+1} \sum_{k=0}^{N} x^2\left(\frac{kT}{N}\right), \tag{1}$$

where T/N is the sampling period.

By invoking the Gaussian assumption, the variance of this estimator follows directly and is given by

$$\text{var } (\hat{\sigma}_x^2) = \frac{2}{N+1} \sum_{k=-N}^{N} \left(1 - \frac{|k|}{N+1}\right) R_x^2\left(\frac{kT}{N}\right),$$

* This work was supported by the U.S. Navy, Bureau of Ships under Contract N00 600-67-CO549.

† So long as the mean of $\{x(t)\}$ is assumed known, no generality of the derivation is lost by letting it be zero.

or in an equivalent form (2)

$$\text{var } (\hat{\sigma}_x^2) = \frac{2}{N+1} \int_{-\infty}^{\infty} q_{(N+1/N)T}(\tau) R_x^2(\tau) \sum_{k=-\infty}^{\infty} \delta\left(\tau - \frac{kT}{N}\right) d\tau.$$

$q_{(N+1/N)T}(\tau)$ is the triangular weighting function (See Ref. 1) and $R_x(\tau)$ is the autocorrelation function of $\{x(t)\}$.

By comparing the var $(\hat{\sigma}_x^2)$ to (3) in the previous paper[1] it is seen that (2) gives the variance of the sample mean of a stationary random process $\{x^2(t)\}$, whose autocovariance function is given by $2R_x^2(\tau)$. The spectrum of $\{x^2(t)\}$ is $2S_x(\omega) * S_x(\omega)$.

At this point the previous theory[1] applies directly. Using the same notation, the spectrum of the squared samples can be written as

$$G(\omega) = \frac{2}{N+1} \int_{-\infty}^{\infty} q_{(N+1/N)T}(\tau) R_x^2(\tau) e^{-i\omega\tau} \sum_{k=-\infty}^{\infty} \delta\left(\tau - \frac{kT}{N}\right) d\tau,$$

which is equivalent to (3)

$$G(\omega) = \frac{N}{T} \sum_{k=-\infty}^{\infty} F\left(\omega - k\frac{2\pi N}{T}\right).$$

In this case

$$F(\omega) = Q(\omega) * 2[S_x(\omega) * S_x(\omega)], \tag{4}$$

and $Q(\omega)$ is the transform of the weighting function.

$G(\omega)$ can be interpreted as $F(\omega)$, shifted by integral multiples of the sampling frequency, $2\pi N/T$. As before to obtain the variance of the estimate we need only be concerned with the value of $G(\omega)$ at $\omega = 0$. To minimize the variance of the estimate, the sampling frequency should be high enough to prevent overlapping of the sideband at $\omega = 0$. Satisfying this condition results in

$$\text{var } (\hat{\sigma}_x^2) = \frac{N}{T} F(0).\dagger \tag{5}$$

To answer the question of how many samples to take in time T to obtain minimum variance, consider (4). $Q(\omega)$ is approximately zero for

† Equation (5) is not quite true when both end points of the time record are included as samples. This is because $(N/T)\,F(0)$ is a function of N namely,

$$(N/T)\,F(0) = \frac{1}{2\pi} \int_{\infty}^{-\infty} 2[S_x(y) * S_x(y)] \left[\frac{\sin y(N + 1/2N)T}{y(N + 1/2N)T}\right]^2 dy.$$

Increasing N beyond the value given in (8) actually results in a higher variance on the estimate. This is apparent in the two examples, particularly for T small. This same effect was discussed in Ref. 1.

$|\omega| \geqq (2\pi/T)[N/(N+1)]$. If $S_x(\omega)$ is zero for $|\omega| \geqq 2\pi B$, then $S_x(\omega) * S_x(\omega)$ will be zero for $|\omega| \geqq 4\pi B$ and $F(\omega)$ will be approximately 0 for

$$|\omega| \geqq 2\pi\left[2B + \frac{1}{T}\left(\frac{N}{N+1}\right)\right]. \tag{6}$$

Therefore, choosing the sampling frequency so that

$$\frac{2\pi N}{T} \geqq 2\pi\left[2B + \frac{1}{T}\left(\frac{N}{N+1}\right)\right] \tag{7}$$

results in (5) being satisfied.

Solving (7) for N yields the required number of samples taken in time T to approximately minimize the variance of the estimate, namely

$$N = 2BT\left[\frac{1 + \sqrt{1 + 2/BT}}{2}\right]. \tag{8}$$

For $BT \gg 1$, N is approximately equal to $2BT$. Thus, twice the number of samples are required to obtain a minimum variance estimate of the variance than was previously shown to obtain a minimum variance estimate of the mean.

2.2 *Variance For Large T*

If T is allowed to become large $Q(\omega)$ will approach a delta function,

$$\lim_{T\to\infty} Q(\omega) = \frac{2\pi}{N+1}\,\delta(\omega). \tag{9}$$

This results in

$$F(\omega) = \frac{1}{\pi(N+1)}\int_{-\infty}^{\infty} S_x(y)S_x(\omega - y)\,dy. \tag{10}$$

If $S_x(\omega)$ is zero for $|\omega| \geqq 2\pi B$, and the sampling frequency satisfies (7), then the minimum value of variance of the estimate is given by

$$\text{var}\,(\hat{\sigma}_x^2)\,|_{\substack{\min \\ T\text{ large}}} = \frac{N}{T}F(0) = \frac{N}{T(N+1)\pi}\int_{-\infty}^{\infty} S_x^2(y)\,dy, \tag{11}$$

$$\approx \frac{1}{\pi T}\int_{-\infty}^{\infty} S_x^2(y)\,dy.$$

This is the same value obtained by continuous sampling.

III. EXAMPLES

The variance of the estimate of variance as a function of number of samples $(N+1)$ and length of record (T) has been computed for two examples.

The computation was done using an expression equivalent to (2).

3.1 *Rectangular Spectrum*

$$S_x(\omega) = \begin{cases} \frac{1}{2}; & -2\pi < \omega < 2\pi, \\ 0; & \text{elsewhere.} \end{cases} \tag{12}$$

Fig. 1 shows var $(\hat{\sigma}_x^2)$ plotted against number of samples. Each curve represents a different length of record as indicated by the values shown on the figure. It should be noted that the minimum value of var $(\hat{\sigma}_x^2)$ occurs at the number of samples predicted by (8). Also for small values of T the var $(\hat{\sigma}_x^2)$ reaches a minimum and then increases as more samples are taken. This is due to including both end points of the time record as samples.

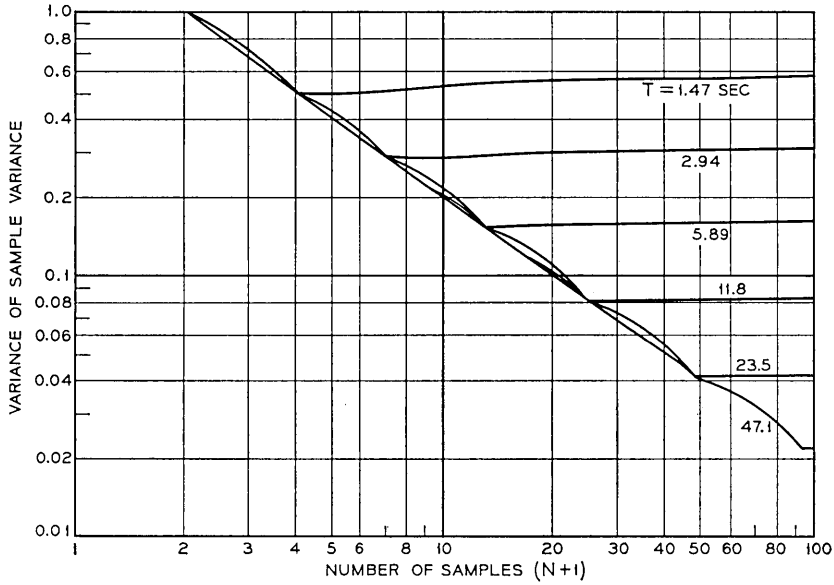


Fig. 1 — Variance of the sample variance as a function of the number of samples and length of record for a process with rectangular spectral density.

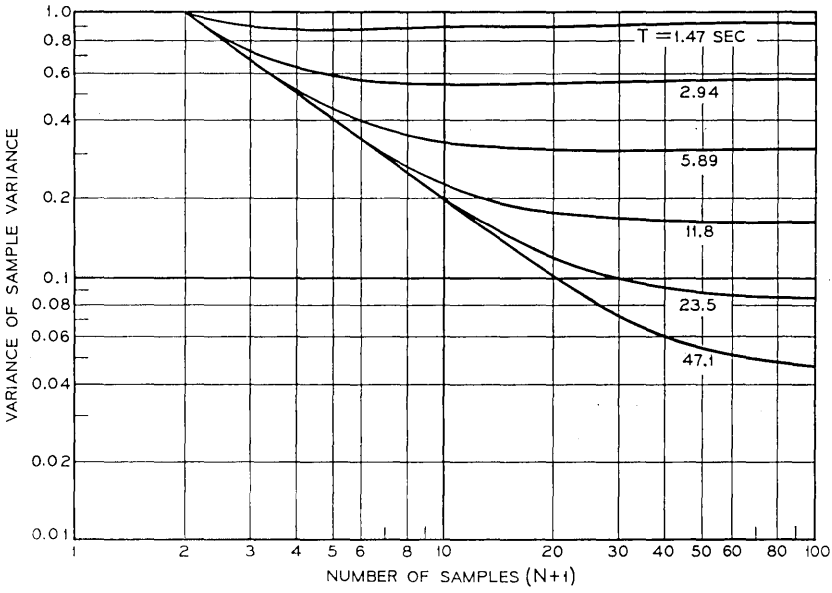


Fig. 2 — Variance of the sample variance as a function of the number of samples and length of record for a process with Markoff spectral density.

3.2 *Markoff Spectrum*

$$S_x(\omega) = \frac{2}{\omega^2 + 1}. \tag{13}$$

This sample shows a nonbandlimited spectrum. The results are shown in Fig. 2.

IV. CONCLUSION

By making the assumption that the random process $\{x(t)\}$ was Gaussian, it was possible to express var $(\hat{\sigma}_x^2)$ into an array of terms containing $R_x^2(kT/N)$, (2). In this form it is possible to apply the theory developed in the work on estimation of the mean.[1] The interesting result from this derivation was that when $BT \gg 1$, the variance of the sample variance is essentially minimized when $2BT$ samples are taken. This is in contrast to the BT samples required to minimize the variance of sample mean.

REFERENCES

1. Balch, H. T., Dale, J. C., Eddy, T. W., and Lauver, R. M., Estimation of the Mean of a Stationary Random Process by Periodic Sampling, B.S.T.J., *45*, May-June, 1966, pp. 733–741.

A Floating Gate and Its Application to Memory Devices

By D. KAHNG and S. M. SZE

(Manuscript received May 16, 1967)

A structure has been proposed and fabricated in which semi-permanent charge storage is possible. A floating gate is placed a small distance from an electron source. When an appropriately high field is applied through an outer gate, the floating gate charges up. The charges are stored even after the removal of the charging field due to much lower back transport probability. Stored-charge density of the order of $10^{12}/cm^2$ has been achieved and detected by a structure similar to an metal-insulator-semiconductor (MIS) field effect transistor. Such a device functions as a bistable memory with nondestructive read-out features. The memory holding time observed was longer than one hour. These preliminary results are in fair agreement with a simple analysis.

It has been recognized for some time that a field-effect device, such as that described by Shockley and Pearson,[1] can be made bistable utilizing switchable permanent displacement charges on ferroelectric material.[2] Subsequent studies of ferroelectric material have revealed,[3] however, that the inherent speed capability of a device incorporating a ferroelectric material is limited by domain motion, whose highest speed is limited by the acoustic velocity. In the absence of highly ordered, near-ideal thin film ferroelectric material, the speed capability of a bistable device, therefore, is in the microsecond range at best.[4] In addition, many ferroelectric materials suffer from irreversible mechanical disorder after many cycles of polarization switching,[2] rendering some uncertainty on the long term device reliability aspect.

An alternative to a ferroelectric gate is a floating gate chargeable by field emission, which hopefully circumvents the above mentioned difficulties. Consider a sandwich structure, metal $M(1)$, insulator $I(1)$, metal $M(2)$, insulator $I(2)$, and finally metal $M(3)$. (See Fig. 1). If the thickness of $I(1)$ is small enough so that a field-controlled electron transport mechanism such as tunneling or internal tunnel-hopping are possible, a positive bias on $M(3)$ with respect to $M(1)$ with $M(2)$ floating [$M(2)$ is called the floating gate henceforth], would cause electron accumulation in the floating gate, provided electron transport

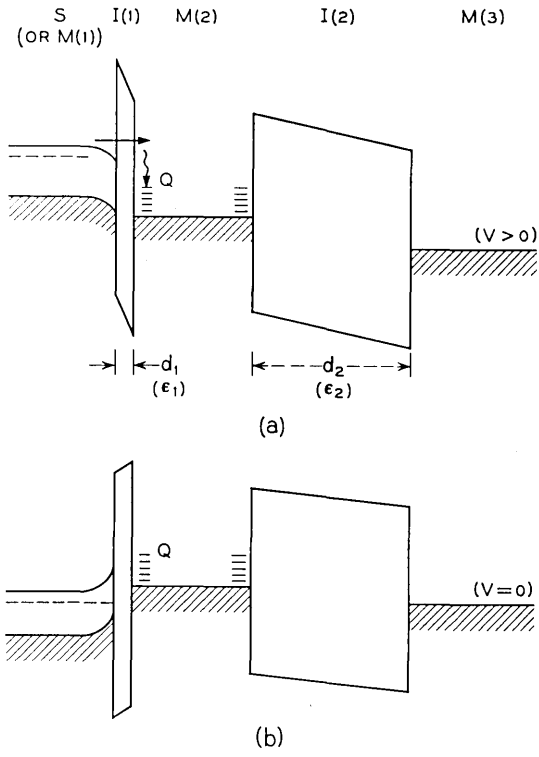


Fig. 1 — Energy band diagram of a floating gate structure with a semiconductor-insulator-metal-insulator-metal sandwich. For calculation of the stored charge, the semiconductor is replaced by a metal $M(1)$. (a) When a positive voltage step is applied to the outer gate. (b) When the voltage is removed. The stored charge Q causes an inversion of the semiconductor surface.

across $I(2)$ is small. These conditions can be met by choosing $I(1)$ and $I(2)$ such that the ratio of dielectric permittivity ϵ_1/ϵ_2 is small and/or the barrier height into $I(1)$ is smaller than that into $I(2)$. The sandwich structure is somewhat similar to the tunnel emitter metal-base transistor proposed by Mead[5] in its structure but with the following essential differences.

(*i*) $M(2)$ is much thicker than the hot electron range, so that emitted electrons are close to the Fermi-level of $M(2)$ before reaching $I(2)$.
(*ii*) No carrier transport is allowed across $I(2)$.
(*iii*) $M(2)$ is floating.

The stored charge Q, as a function of time when a step voltage function with amplitude V is applied across the sandwich, is given by

$$Q(t) = \int_0^t j\,dt' \qquad \text{coul/cm}^2. \tag{1}$$

When the emission is of Fowler-Nordheim tunneling type, then the current density, j, has the form

$$j = C_1 E^2 \exp(-E_0/E), \tag{2a}$$

where C_1 and E_0 are constants in terms of effective mass and the barrier height. (We have neglected the effects due to the image force lowering[6] of the barrier, etc., but the essential feature is expected to be retained even after detailed corrections are made). This type of current transport occurs in SiO_2 and Al_2O_3.

When the field emission is of the internal Schottky or Frankel-Poole type, as occurs in Si_3N_4,[7] then j follows the form

$$j = C_2 E \exp[-q(\Phi_1 - \sqrt{qE/\pi\epsilon_1})/kT], \tag{2b}$$

where c_2 is a constant in terms of trapping density in the insulator, Φ_1 the barrier height in volts, ϵ_1 the dynamic permittivity.

The electric field in $I(1)$ at all times is a function of the applied voltage V and $Q(t)$, and is obtainable from the displacement continuity requirement as

$$E = \frac{V}{d_1 + d_2(\epsilon_1/\epsilon_2)} - \frac{Q}{\epsilon_1 + \epsilon_2(d_1/d_2)}, \tag{3}$$

where d_1 and d_2 are the thickness of $I(1)$ and $I(2)$, respectively.

Fig. 2(a) shows the results of a theoretical computation using (1), (2a), and (3) with the following parameters: $d_1 = 50$ Å, $\epsilon_1 = 3.8\,\epsilon_0$ (for SiO_2), $d_2 = 1000$ Å, $\epsilon_2 = 30\,\epsilon_0$ (for ZrO_2), and V = 50 volts. One notes that the stored charge initially increases linearly with time and then saturates. The current is almost constant for a short time and then decreases rapidly. The field in $I(1)$ decreases slightly as the time increases. The above results can be explained as follows: When a voltage pulse is applied at $t = 0$, the initial charge Q is zero, and the initial electric field across $I(1)$ has its maximum value, $E_{max} = V/[d_1 + (\epsilon_1/\epsilon_2)d_2]$. As t increases, Q will first increase linearly with time. This is because of the fact that for small Q such that E remains essentially the same, the current will in turn remain the same, so $Q = j(E_{max}) \cdot t$. Eventually, when Q is large enough to reduce the

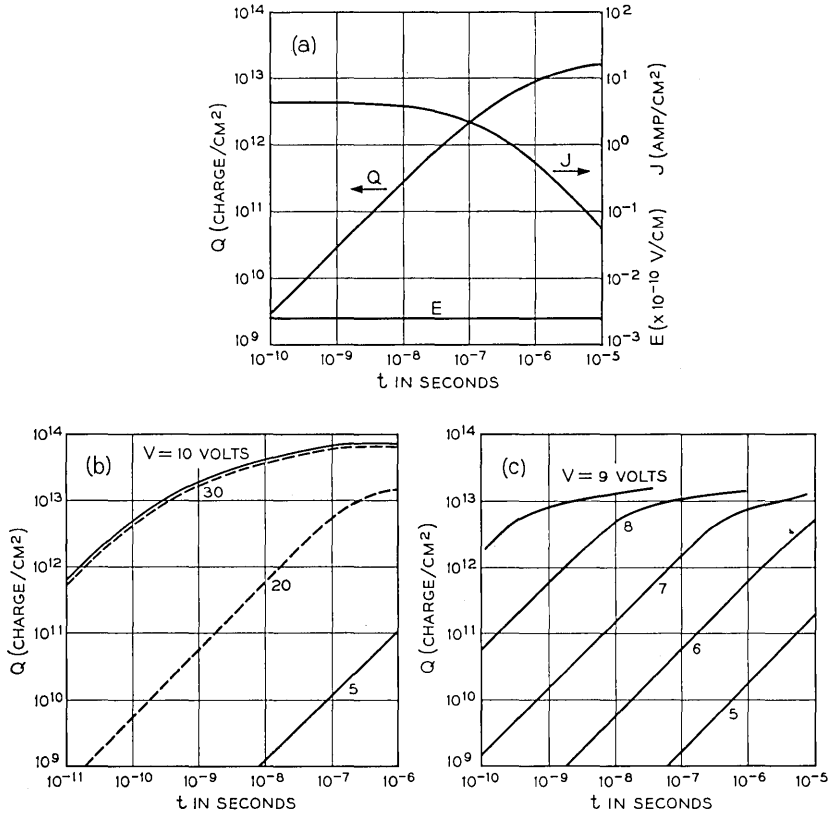


Fig. 2(a) — Theoretical results of the stored-charge density (Q), the current density (J), and the electric field across $I(1)$ as a function of time. V = 50 volts, $d_1 = 50$ Å, $\epsilon_1/\epsilon_0 = 3.8$ (for SiO_2), $d_2 = 1000$ Å, $\epsilon_2/\epsilon_0 = 30$ (for ZrO_2). (b) Theoretical results of the stored charge density as a function of time with the same ϵ_1 and ϵ_2 as in (a), and $d_1 = 10$ Å, $d_2 = 100$ Å (solid lines), $d_1 = 30$ Å, $d_2 = 300$ Å (dotted lines). (c) Theoretical results of the stored density as a function of time with $d_1 = 20$ Å, $\epsilon_1/\epsilon_0 = 60$ (for Si_3N_4), $d_2 = 200$ Å, $\epsilon_2/\epsilon_0 = 30$ (for ZrO_2) and various applied voltages.

value of E substantially, then the current will decrease rapidly with time and Q increases slowly.

Fig. 2(b) shows the stored charge as a function of time for the time ϵ_1 and ϵ_2 but different d_1, d_2, and V. It is clear that for a given structure, in order to store a given amount of charge, one can either increase the applied voltage or increase the charging time (pulse width) or both. Fig. 2(c) shows the calculated stored charge for the current transport described by (2b). Here $I(1)$ is a 20 Å thick Si_3N_4 film. There are

marked decreases in the gate voltages required for a given charge compared to Al_2O_3, SiO_2. This is largely due to the much lower barrier height (1.3 volts)[7] compared to SiO_2 ($\simeq$ 4.0 volts).[8]

It is noted that the field in $I(1)$ for appreciable charge storage is in the 10^7 V/cm range. When the outer gate voltage is removed, the field in $I(1)$ due to the stored charge on the inner gate is only 10^6 V/cm or so corresponding to 5×10^{12} charges/cm^2, a large enough charge to detect easily. Since the transport across $I(1)$ is highly sensitive to the field, (2a) and (2b), no charges flow back. The charge loss is actually controlled by the dielectric relaxation time of the sandwich structure,[9] which is very long. When it is desired to discharge the floating gate quickly, it is necessary to apply to the outer gate a voltage about equal in magnitude but opposite in polarity to the voltage which was used for charging. It is evident that net positive charges (loss of electrons) can also be stored in the floating gate if the discharging gate voltages are appropriately chosen in magnitude and duration.

It was mentioned that the stored-charge density of 5×10^{12}/cm^2 was sufficient for easy detection. One of the detection or read-out schemes is to use the surface field effect transistor (MOSFET or IGFET) first fabricated and described by Kahng and Atalla[10] in 1960. For inversion at a silicon surface, the charge required is only about 2×10^{11}/cm^2 for 1 ohm-cm n-type silicon. However, surface-state charges at the silicon-silicon dioxide interface may be as high as 10^{12}/cm^2, depending on the fabrication techniques used. For this reason we have chosen 5×10^{12}/cm^2 as the stored charge required for easy detection. When the Insulated Gate Field Effect Transistor (IGFET) principle is used for read-out, $M(1)$ is now replaced by silicon. This requires a slight correction in the calculation of charge flow through the insulator, but the major features of the results are not expected to be altered significantly. It is to be noted that about one half of the stored charge can be active in creation of the inversion layer since the other half resides near the $M(2)$-$I(2)$ interface due to Colombic repulsion.

To check the feasibility, a floating gate device has been fabricated using an IGFET as shown in Fig. 3(a). The substrate is an n-type silicon, 1 ohm-cm, and ⟨111⟩ oriented. $I(1)$ is a 50 Å SiO_2 thermally grown in a dry oxygen furnace. $M(2)$ and $I(2)$ are Zr (1000 Å) and ZrO_2 (1000 Å), respectively. $M(3)$ and the ohmic contact metals are aluminum deposited in a vacuum system. Fig. 3(b) is another version of the floating gate device using a thin film transistor (TFT) structure.[11]

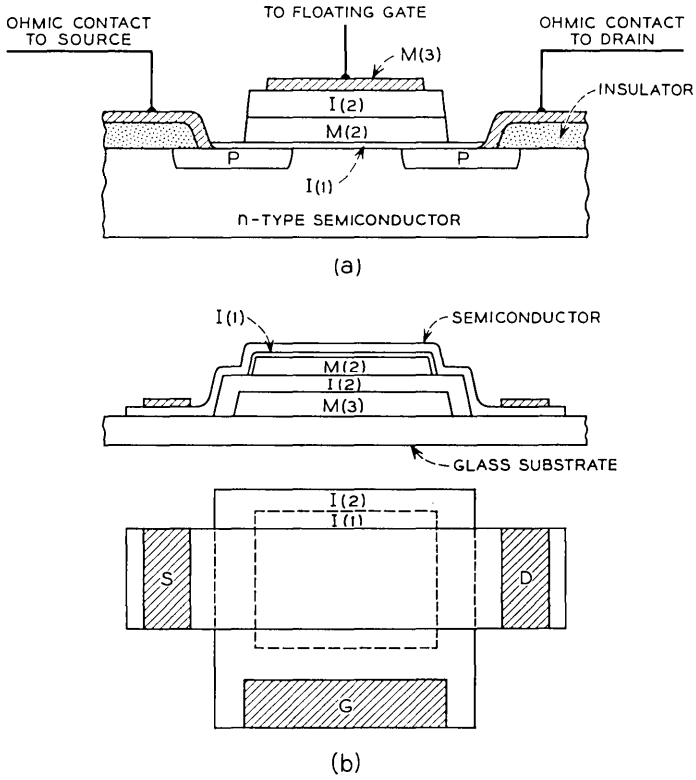
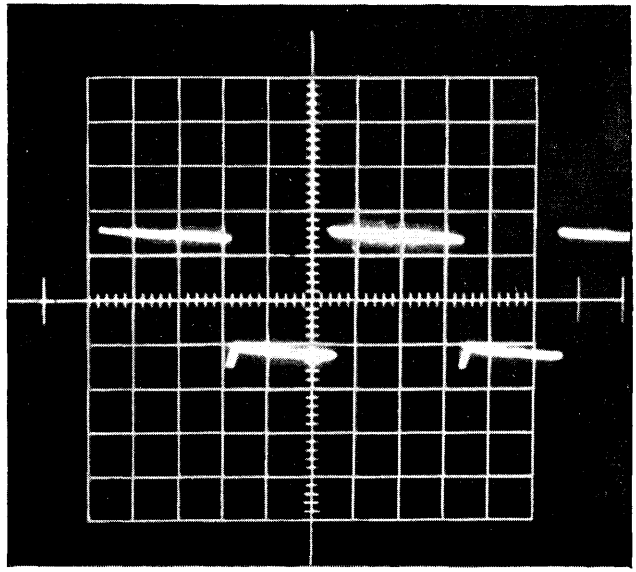
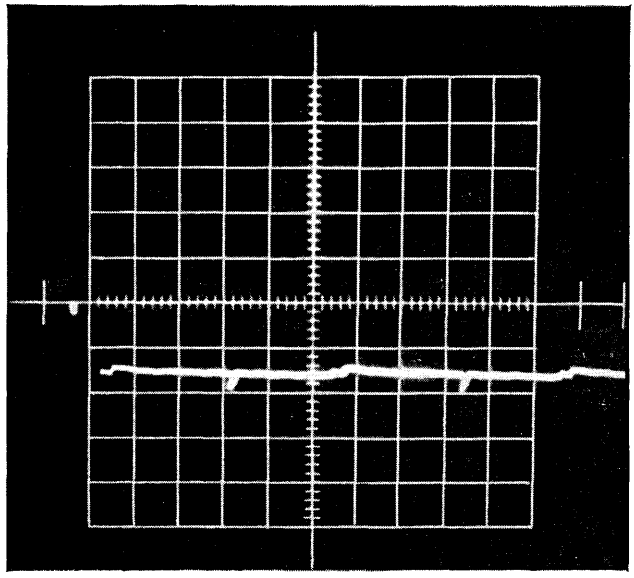


Fig. 3 — (a) Schematic diagram of a floating gate device using an IGFET. The numbers indicated correspond to those shown in Fig. 1. (b) Schematic diagram of a floating gate device using thin film transistor structure.

The IGFET-type floating gate devices have been tested in a pulsing circuit. Because of the relatively thick insulator layers, large voltages ($\approx$ 50 V) and long pulse width ($\approx$ 0.5 μs) have to be applied in order to store $\approx 5 \times 10^{12}$ charges/cm^2. Fig. 4 shows the experimental results. A positive pulse of 50 volts is first applied to the gate electrode, and 60 ms later a negative pulse of 50 volts is applied. Then the pulsing cycle repeats. In Fig. 4(a) the pulse widths are 0.5 μs. One notes that when the positive pulse is applied, a sufficient amount of charge is stored in the floating gate so that the silicon surface is inverted; a conducting channel is thus formed, and the channel current is "on." It can be seen that the channel current decreases only slightly at the end of 60 ms. When the negative pulse is applied, the stored charge is eliminated, and also the channel. The channel current reduces to its



(a)



(b)

Fig. 4 — Experimental results of the channel current of a IGFET-type floating gate device. A positive voltage pulse, V_1, with pulse width W_1, is first applied to the gate, and 60 ms later a negative pulse V_2 with pulse width W_2 is applied. Then the pulsing cycle repeats. Horizontal scale: 20 ms/div. Vertical scale: 0.1 ma/div. (a) $V_1 = V_2 = 50$ volts, $W_1 = W_2 = 0.5$ μs. (b) $V_1 = V_2 = 40$ volts, $W_1 = W_2 = 0.5$ μs.

"off" state. Fig. 4(b) shows results for pulses with the same widths but smaller amplitude (40 V). Since the stored charge is a strong function of the pulse amplitude, only a very small amount of charge is stored, too small to cause inversion. For non-leaky units, the memory holding time of longer than one hour has been observed.

It is clear that a modified IGFET such as a TFT can be used for read-out, as shown in Fig. 3(b). For an academic study of device operation, the floating gate can be partially exposed and a potential probe can be placed nearby.

In conclusion, it has been demonstrated that the controlled field emission to the buried "floating" gate may be capacitively induced by pulsing the outer gate electrode. This combination can therefore, be used as a memory device, with holding time as long as the dielectric relaxation time of the gate structure and with continuous nondestructive read-out capability. There seems to be no inherent reason why read-in read-out cannot be performed in a very short time, say in the nanosecond range or even shorter.

We wish to acknowledge helpful discussions and technical assistance rendered by M. P. Lepselter and P. A. Byrnes in connection with ZrO_2, and skillful technical assistance given to us by G. P. Carey and A. Loya, and J. F. Grandner and his group.

REFERENCES

1. Shockley, W., and Pearson, G. L., Phys. Rev., *74,* 1948, p. 232.
2. Ross, I. M., U. S. Patent 2,791,760, issued May 7, 1957.
3. Jona, F. and Shirane, G., *Ferroelectric Crystals,* MacMillan Co., 1962.
4. Heyman, P. M. and Heilmeier, G. H., Proc. IEEE, *54,* 1966, p. 842.
5. Mead, C. A., Proc. IRE, *48,* 1960, p. 359.
6. Simmons, J. G., J. Appl. Phys., *34,* 1963, p. 1793.
7. Sze, S. M., Current Transport and Maximum Dielectric Strength of Silicon Nitride Films, J. Appl. Phys., June, 1967.
8. Williams, R., Phys. Rev., *140,* 1965, A569.
9. Kahng, D., Semipermanent Memory Using Capacitor Charge Storage and IGFET Read-out, B.S.T.J., this issue, pp. 1296–1300.
10. Kahng, D. and Atalla, M. M., Silicon-Silicon Dioxide Field Induced Surface Devices, presented at the IRE-AIEE Device Res. Conf., Carnegie Inst. Tech., Pittsburgh. Pa., June 1960.
11. Weimer, P. K., An Evaporated Thin Film Triode, presented at the IRE-AIEE Device Res. Conf., Stanford University, Stanford, Calif., June 1961.

Semipermanent Memory Using Capacitor Charge Storage and IGFET Read-out

By D. KAHNG

(Manuscript received May 17, 1967)

One of the earliest computers used capacitors as its memory.[1] A mechanical means was used for both read-in and read-out operations. Electronic accessing was used in conjunction with vacuum tube or solid-state diodes in relatively modern computers such as the SEAC computer.[2] Capacitor storage is rarely used at present since magnetic memories meet the modern computer requirements much better. The inherent difficulty with capacitor storage was the limited holding time since a nonlinear resistor with small enough leakage currents to allow useful memory holding time was then not readily available. The old capacitor memory was charged through a diode with slow recovery time and with leakage current of 10^{-10} amp at best and required a large capacitor for any appreciable holding time. Furthermore, the read-out was usually destructive.

The capacitor storage merits re-examination in view of the advanced solid-state devices and technology now available. Coupled with an Insulated Gate Field Effect Transistor (IGFET),[3] the read-out can be nondestructive. Integrated circuit techniques may prove superior to the current magnetic memories for some applications where infrequent recycling is permissible. The inherent speed should be much faster than that of magnetic units.

Consider a capacitor C in series with a nonlinear element as shown in Fig. 1. The capacitor may represent the gate capacity of the IGFET plus any external capacitor in parallel with the gate capacitor. When a positive voltage pulse with amplitude V and duration τ is applied at the nonlinear element terminal, it can be shown that the stored charge $Q(\tau)$ and the decay time constant τ_c, defined as the time required to reach $1/e$ value of the initial stored charge Q_o, can be calculated for various nonlinear resistors.

I. POWER-LAW RESISTORS

The I-V characteristics are given by

$$I = KV_n^m, \tag{1}$$

then

$$Q(\tau) = C\left[V - \left(\frac{1}{V^{(m-1)}} + K(m-1)\frac{\tau}{C}\right)^{-1/(m-1)}\right] \tag{2}$$

and

$$\begin{aligned} \tau_e &= \frac{C^m}{(m-1)K}\frac{1}{Q_0^{(m-1)}}(e^{(m-1)} - 1) \\ &= \frac{(e^{(m-1)} - 1)}{(m-1)}\frac{Q_0}{I_0}, \end{aligned} \tag{3}$$

where I_o is the discharge current at the termination of charging. It is clear from (2) that m $\geqq$ 1 for physically meaningful $Q(\tau)$. For a long holding time, (3) tells us that the nonlinearity of the resistor should be large. These equations would describe the behavior of the storage unit comprising a space-charge-limited-current diode.[4] If traps are present, only a simple modification is needed in the analysis. Structures comprising photosensitive space-charge-limited-current diodes such as CdS diodes should allow optical read-in operations which might be advantageous for certain applications such as a vidicon.

II. TUNNEL SANDWICH DIODES

For this nonlinear element, the circuit in Fig. 1 should be modified to include the shunt capacitance of the tunnel sandwich. Thus, at the instance of pulse application, the voltage divides between the two capacitors. The I-V relationship for Fowler-Nordheim type tunneling is

$$I = KV_n^2 e^{-V_0/V_n}. \tag{4}$$

With the appropriate initial conditions, the stored charge $Q(\tau)$ is given by

$$Q(\tau) = C\left[V - \alpha V_0/\ln\left(e^{\alpha V_0/V} + \frac{KV_0}{C}\tau\right)\right], \tag{5}$$

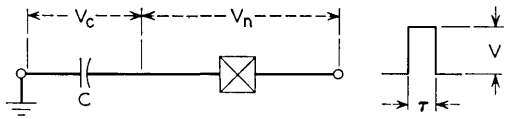


Fig. 1 — A capacitor being charged through a nonlinear resistor.

where α is defined as unity plus the ratio of the shunt capacitance of the tunnel sandwich, C_n to the charging capacitor, namely $\alpha = 1 + (C_n/C)$.

The discharge time constant τ_c can also be shown to be

$$\begin{aligned} \tau_c &= \frac{\frac{\alpha C}{KV_0}\left(e^{(\alpha CV_0/Q_0)(e-1)} - 1\right)}{e^{\alpha CV_0/Q_0}} \\ &= \tau_c e^{(\alpha CV_0/Q_0)(e-2)}. \end{aligned} \tag{6}$$

It is clear that charging time required for adequately large Q is very short, allowing fast read-in operation. The decay time can be seen to be large for tunneling across well-known insulators such as SiO_2 and Al_2O_3. Therefore, the decay is not controlled by (6) but rather by the dielectric relaxation time of the insulators used for the entire assembly including the IGFET gate material. The dielectric relaxation time of the best inorganic insulators is of the order of one day at room temperatures. However, certain organic insulators are known to have theoretical dielectric relaxation times of many years. Performance of a memory cell incorporating a tunnel sandwich diode is described in more detail elsewhere.[5]

III. SCHOTTKY BARRIER DIODES

For charging through a rectifier, Schottky barrier diodes are preferred over pn junction diodes since Schottky barriers are majority carrier devices and hence fast recovery is achievable.[6] I-V characteristics may be represented by

$$I = I_s(e^{\beta V} - 1). \tag{7}$$

For charging, we may neglect the unity in the bracket in (7), and the stored charge can be shown to be

$$Q(\tau) = C\left[V - \frac{1}{\beta}\ln\left(\frac{\tau}{\tau_c} + e^{-\beta V}\right)^{-1}\right], \tag{8}$$

where

$$\tau_c = \frac{C}{\beta I_s}.$$

For decay, it is easy to show

$$\tau_e = \frac{Q_0}{I_s}(1 - e^{-1}). \tag{9}$$

Fig. 2 shows the stored charge Q computed from (8) as a function of pulse duration for several pulse amplitudes. The characteristic time constant τ_c is not much less than 1 sec for a typical configuration ($C < 10^{10}$ Fd, $I_s > 10^{-12}$ amp). Therefore, it is seen that the stored charge is proportional to the pulse amplitude for sufficiently large V. This suggests that the device may be used as a multi-level storage unit.

The combination of Schottky barrier diodes and the IGFET is shown in Fig. 3. A similar structure has been fabricated and tested. The holding time was of the order of 10 sec when the charging was done by 15 volts pulse in agreement with (9) since the IGFET gate capacitance was about 10^{-12} Fd and I_s of the diode was 10^{-12} amp (measured at 10 volts reverse bias). The maximum pulse amplitude before the breakdown of the gate insulator was about 30 volts, allowing a longer holding time of about 30 seconds. The read-in time was less than 10^{-7} second and the reverse breakdown of the diode was used to turn the IGFET off, which also took less than 10^{-7} second.

A memory featuring nondestructive read-out, access times in the

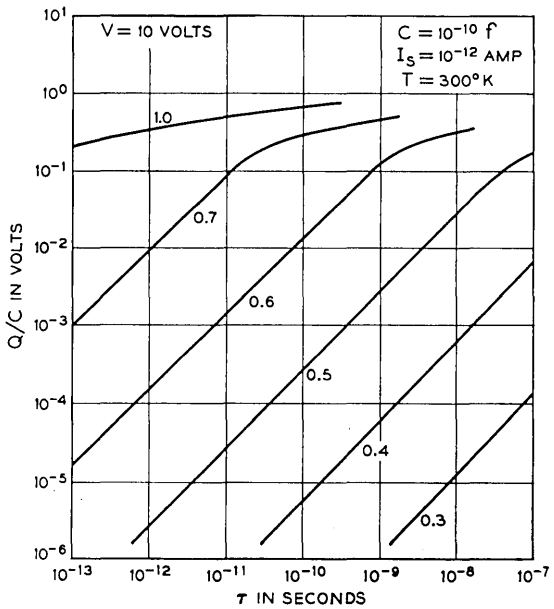


Fig. 2 — Theoretical stored charge/storing capacitance (the floating potential) as a function of time for charging through a Schottky barrier diode.

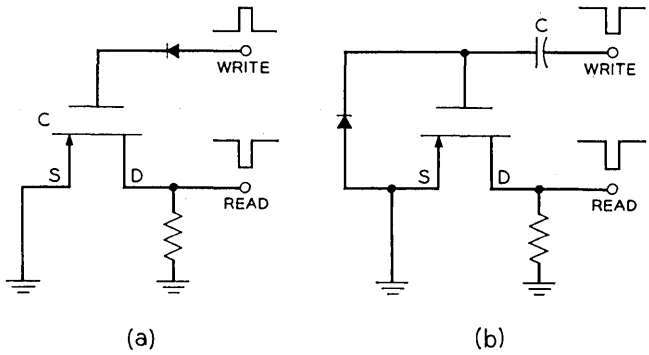


Fig. 3 — Combination of the capacitor storage unit with a p-channel IGFET for read-out. (a) Series connection, read-in should be positive. (b) Parallel connection, read-in should be negative.

submicrosecond range, and holding times of many seconds should find many applications. An integrated structure incorporating Schottky barrier diodes and IGFETs is readily obtainable with modern solid-state technology.

IV. ACKNOWLEDGMENT

The author gratefully acknowledges helpful discussions he had with H. K. Gummel, R. M. Ryder, J. E. Iwersen, and S. M. Sze.

REFERENCES

1. Richards, R. K., *Electronic Digital Systems,* John Wiley & Sons, Inc., New York, 1966, credits J. V. Atanasoff and his associates for the first capacitor memory bank constructed 1938–1942.
2. Holt, A. W., *Computer Handbook,* edited by H. D. Huskey and G. A. Koran, The Maple Press Co., York, Pa., 1962, Diode-Capacitor Memory System, Sect. 12.9.
3. Kahng, D. and Atalla, M. M., Silicon-Silicon Dioxide Field Induced Surface Devices, presented at the IRE-AIEE Device Research Conference, Carnegie Institute of Technology, Pittsburgh, Pa., June, 1960.
4. Rose, A., Phys. Rev., *97,* 1955, p. 1538.
5. Kahng, D. and Sze, S. M., A Floating Gate and its Application to Memory Devices, B.S.T.J., this issue, pp. 1288–1295.
6. Kahng, D. and D'Asaro, L. A., Gold-Epitaxial Silicon High-Frequency Diodes, B.S.T.J., *43,* 1964, pp. 225–232.