I

FUNDAMENTAL LIMITATIONS IN MICROELECTRONICS

II

POWER SCHOTTKY DIODE DESIGN AND
COMPARISON WITH THE JUNCTION DIODE

III

PERMITTIVITY OF STRONTIUM TITANATE

Thesis by

Bruce Hoeneisen Frost

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1972

(Submitted April 6, 1972)

To Matico and Anahí

ACKNOWLEDGMENT

Grateful thanks are due to my advisor, Dr. Carver Mead, for his many contributions to this work, to Dr. Richard Neville, co-worker in the investigation on strontium titanate, to Mrs. Ruth Stratton, Miss Ingrid Vierheilig, and Mrs. Aldre Khoury, for their secretarial work, and especially to librarian Mrs. Paula Samazan for the many attentions and encouragement she has given me during these years at the Institute.

I gratefully acknowledge receiving financial support from Tektronix, Inc. and from the Institute. The work reported here was supported in part by the Office of Naval Research and by the General Electric Company.

ABSTRACT

Part I

The physical phenomena which will ultimately limit the packing density of planar bipolar and MOS integrated circuits are examined. The maximum packing density is obtained by minimizing the supply voltage and the size of the devices. The minimum size of a bipolar transistor is determined by junction breakdown, punch-through and doping fluctuations. The minimum size of a MOS transistor is determined by gate oxide breakdown and drain-source punch-through. The packing density of fully active bipolar or static non-complementary MOS circuits becomes limited by power dissipation. The packing density of circuits which are not fully active such as read-only memories, becomes limited by the area occupied by the devices, and the frequency is limited by the circuit time constants and by metal migration. The packing density of fully active dynamic or complementary MOS circuits is limited by the area occupied by the devices, and the frequency is limited by power dissipation and metal migration. It is concluded that read-only memories will reach approximately the same performance and packing density with MOS and bipolar technologies, while fully active circuits will reach the highest levels of integration with dynamic MOS or complementary MOS technologies.

Part II

Because the Schottky diode is a one-carrier device, it has both advantages and disadvantages with respect to the junction diode which is a two-carrier device. The advantage is that there are practically no excess minority carriers which must be swept out before the diode blocks

current in the reverse direction, i.e. a much faster recovery time. The disadvantage of the Schottky diode is that for a high voltage device it is not possible to use conductivity modulation as in the p i n diode; since charge carriers are of one sign, no charge cancellation can occur and current becomes space charge limited. The Schottky diode design is developed in Section 2 and the characteristics of an optimally designed silicon Schottky diode are summarized in Fig. 9. Design criteria and quantitative comparison of junction and Schottky diodes is given in Table 1 and Fig. 10. Although somewhat approximate, the treatment allows a systematic quantitative comparison of the devices for any given application.

Part III

We interpret measurements of permittivity of perovskite strontium titanate as a function of orientation, temperature, electric field and frequency performed by Dr. Richard Neville. The free energy of the crystal is calculated as a function of polarization. The Curie-Weiss law and the LST relation are verified. A generalized LST relation is used to calculate the permittivity of strontium titanate from zero to optic frequencies. Two active optic modes are important. The lower frequency mode is attributed mainly to motion of the strontium ions with respect to the rest of the lattice, while the higher frequency active mode is attributed to motion of the titanium ions with respect to the oxygen lattice. An anomalous resonance which multi-domain strontium titanate crystals exhibit below $65^{\circ}K$ is described and a plausible mechanism which explains the phenomenon is presented.

The contents of this thesis has been published under the following titles:

"Fundamental limitations in microelectronics. I. MOS technology", B. Hoeneisen and C. A. Mead, Solid-St. Electr., in press.

"Fundamental limitations in microelectronics. II. Bipolar technology:, B. Hoeneisen and C. A. Mead, Solid-St. Electr., in press.

"Current-voltage characteristics of small size MOS transistors", B. Hoeneisen and C. A. Mead, IEEE Trans. Electr. Devices, ED - 19, 382 (1972).

"Power Schottky diode design and comparison with the junction diode", B. Hoeneisen and C. A. Mead, Solid-St. Electr. 14, 1225 (1971).

"Permittivity of strontium titanate", R. C. Neville, B. Hoeneisen, and C. A. Mead, J. Appl. Phys., in press.

"Anomalous resonance of strontium titanate", R. Neville, B. Hoeneisen, and C. A. Mead, J. Appl. Phys., in press.

TABLE OF CONTENTS

PART I

FUNDAMENTAL LIMITATIONS IN MICROELECTRONICS

FUNDAMENTAL LIMITATIONS IN MICROELECTRONICS

Development of the planar technology in the late 1950's made integrated circuits possible. The number of devices per chip has doubled every year since the first planar transistors were manufactured in 1958, as shown in Fig. 1*. Although the chip area has increased by a factor of ≈20 in the last decade, the exponential growth in the number of devices per chip has largely been due to the steady decrease in size of individual devices. In spite of the increasing circuit complexity, the yields have remained approximately unchanged due to improvements in the technology. Although it is expected that this trend will continue in the near future, planar technology will soon reach rather fundamental limitations and the number of devices per unit area must level off.

The minimum size of present day devices is determined primarily by the tolerances in alignment of successive masks. Even with present day techniques, tolerances are improving steadily. As electron beam pattern generation becomes more generally available, mask alignment to a much higher precision may be envisioned. With these significant developments approaching, it is important to identify clearly the fundamental limitations which will ultimately limit circuit miniaturization.

The purpose of this study is to identify the physical phenomena which will ultimately limit the packing density of planar bipolar and

---

*G. E. Moore, private communication

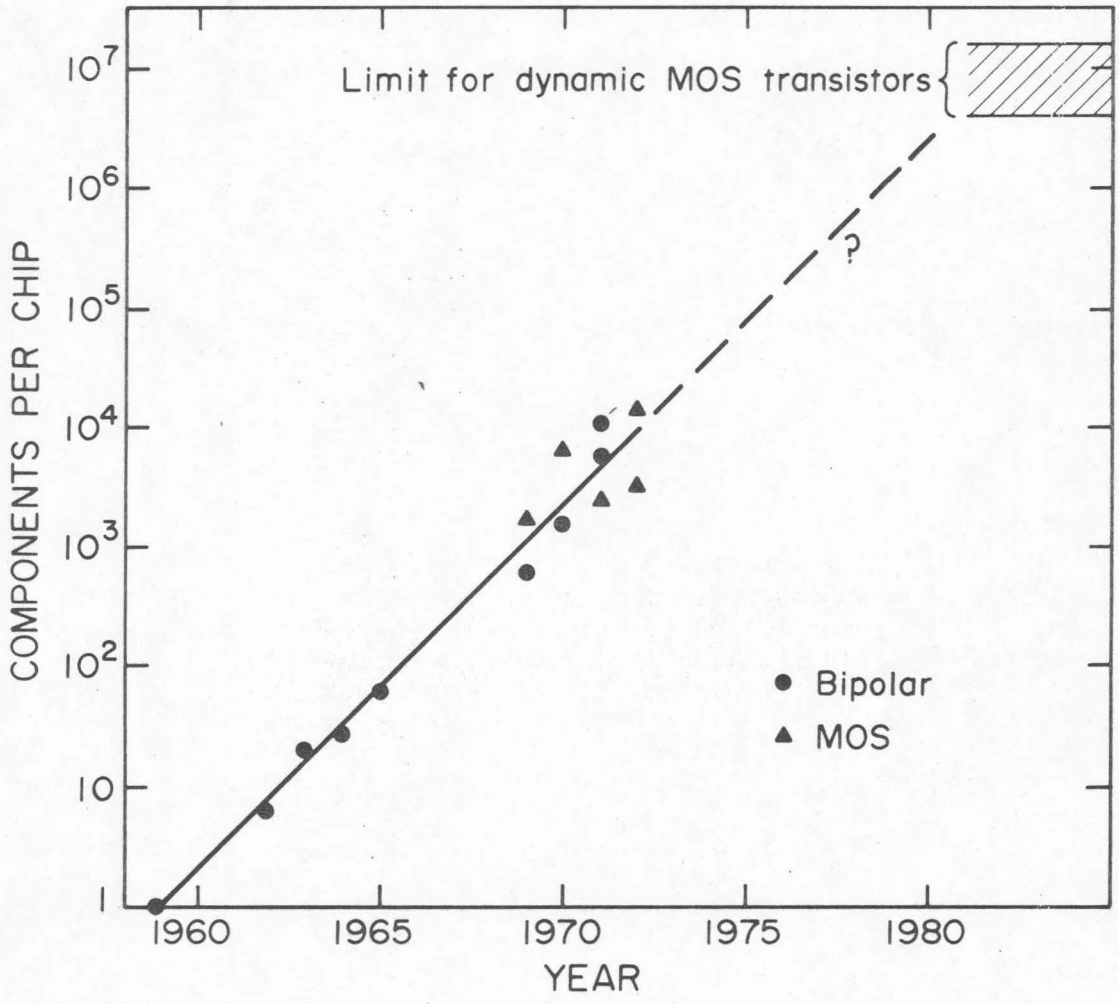

Fig. 1 History of integrated circuit complexity. Line corresponds to a two-fold increase in the number of components per chip per year. This figure is due to Gordon E. Moore.

MOS (metal-oxide-semiconductor) integrated circuits. It must be stressed that we consider only planar transistor circuits as we know them today, i.e., with silicon substrate and silicon dioxide dielectrics. The fundamental limitations we determine can be approached as tolerances and yields improve.

The limit we shall determine for fully dynamic MOS circuits is presented in Fig. 1. The uncertainty in chip size contributes to the uncertainty indicated in the figure. Notice that the maximum number of transistors per chip is approximately three orders of magnitude larger than present day circuits. At the current rate of growth such a limit would be reached within a decade.

For both MOS and bipolar integrated circuits the maximum number of devices per unit area is determined either by the power dissipation density or by the area occupied by transistors, interconnections, and passive devices. For a given frequency of operation, reduction of the supply voltage and/or the circuit capacitances permits a reduction of power dissipation and interconnection area per transistor. Reducing the size of the devices not only reduces the area occupied by these devices, but also reduces the circuit capacitances. In addition, lower voltage devices can be made smaller. Thus to maximize the circuit packing density it is necessary to minimize the supply voltage and the size of the individual devices.

1. Limitations of the Planar MOS Technology

1.1 Principal Limitations

To maximize the packing density of MOS integrated circuits it is necessary to minimize the supply voltage and the size of the individual devices.

The supply voltage has a lower bound which is determined by reproducibility of the gate turn on voltage, by the minimum oxide thickness which can be reliably manufactured and by noise margin considerations.

The area occupied by a present day MOS transistor can be reduced by decreasing its channel width and length. The channel length reduction has a limit, however, since when the drain and source depletion regions overlap, punch-through occurs. Further miniaturization is possible if the depletion widths are reduced by reducing the circuit supply voltage and increasing the substrate doping concentration. As the substrate doping concentration is increased the gate oxide electric field required to invert the substrate also increases. Thus the maximum allowable oxide field sets an upper limit to the substrate doping concentration. This concentration together with the junction built in voltage determines the minimum depletion region thickness of an operable device, which in turn determines the minimum device size.

Other size limitations are considered in detail although it is shown that they are not as stringent as the oxide field limitation mentioned above. These limitations include drain-substrate breakdown,

drain "corner" breakdown, and substrate doping fluctuations.

It will be shown that for static non-complementary circuits the maximum number of circuit functions per chip is determined by power dissipation, except for circuits such as read only memories in which a small fraction of the devices dissipate most of the power. The maximum packing density of fully dynamic or complementary MOS circuits is determined by the area occupied by transistors and interconnections.

Since a positive voltage is normally applied to the gate of an n-channel device, the silicon-silicon dioxide interface charge Q_{ss}, which is positive, does not have a tendency to increase with time[1]. As a result the flat band voltage of an n-channel MOS FET is inherently more stable than that of a p-channel device*. This is an important advantage in view of the high oxide fields and low threshold voltages of minimum size devices.

We will now consider the ultimate limitations of planar MOS transistors. More stringent limitations encountered in actual circuits are examined in the following section. The substrate doping concentration has an upper limit of $\simeq 2 \cdot 10^{19} cm^{-3}$ determined by field emission in the drain and source junctions. At higher doping concentrations the junction characteristic approaches that of a tunnel diode and isolation between the substrate and the drain and source regions is lost. Oxide "breakdown" limits the substrate doping concentration to $\simeq 1.3 \cdot 10^{19} cm^{-3}$. At higher concentrations the maximum electric field which can be applied to the gate oxide, ($\simeq 6 \cdot 10^{6} V/cm$) [2] does not

---

*It is assumed that normal processing precautions have been used to eliminate alkali ions in the oxide.

invert the substrate. The junction built in voltage produces a depletion thickness of 0.01μm into a substrate with $1.3 \cdot 10^{19}$ dopant atoms per cm^3. The channel length cannot be made smaller than approximately two depletion region thicknesses, or ≃0.02μm. Otherwise the two junctions would be in punch-through even with no applied bias.

The gate oxide thickness has a lower limit of ≃50Å determined by tunneling through the silicon dioxide energy gap. The isolation between gate and substrate is reduced for thinner oxides, since the oxide conductance per unit area increases exponentially with decreasing thickness[2].

Since high operating voltages preclude high packing density, it is important to determine how low an operating voltage may practically be achieved. Ultimately this voltage will depend upon the stability and reproducibility of the gate turn on voltage V_{GT} (given by Eq. (1A) of the appendix, page 42). For an n-channel silicon gate device the constant additive term $|V_{FB} + 2\phi|$ can be made as low as 0.1 to 0.3V depending on the silicon-silicon dioxide interface charge density Q_{ss}, the oxide thickness x_o, and the substrate doping concentration C_B. V_{FB} is the flat band voltage and 2ϕ is the substrate band bending at onset of strong inversion. Consider the source connected to the substrate, that is, $V_S = 0$. As long as the last term in Eq. (1A) is much larger than $|V_{FB} + 2\phi|$, the gate turn on voltage is proportional to $x_o\sqrt{C_B}$. Thus for a given relative manufacturing tolerance of x_o and C_B, the relative tolerance of V_{GT} is independent of V_{GT}, i.e., as V_{GT} is made smaller its

absolute controllability increases provided $V_{GT} \gtrsim |V_{FB}+2\phi|$. Therefore gate turn on voltages as low as $\simeq |V_{FB}+2\phi|$, i.e. a few tenths of a volt, can be achieved. For proper circuit operation the supply voltage should not be made much smaller than approximately $2V_{GT}$.

1.2 Minimum Size MOS Transistor

In this section we determine the approximate minimum size of MOS transistors as a function of the drain voltage V_{DD} . The results are approximate because they depend on a number of assumptions such as circuit configuration, gate turn on voltages, maximum gate oxide field, and flat band voltage, but should be within a factor of 2 of the actual limiting geometry. The circuit considered is an inverter as shown in the inset of Fig. 1A, page 44 . The source of the driver transistor 1 is connected to zero potential. The drain of the pull up transistor 2 is connected to V_{DD} , while its gate is connected to V_{GG} . All voltages are referred to the substrate. We arbitrarily chose $V_{GG} = 2V_{DD}$, the gate turn on voltage of transistor 1 to be $V_{GT1} = \frac{1}{2}V_{DD}$ and that of transistor 2 to be $V_{GT2} = \frac{3}{2}V_{DD}$ when $V_o = V_{DD}$. This situation is a particular case of the more general problem considered in Appendix 1 (see Fig. 1A). We shall assume that the gate flat band voltage V_{FB} is equal to -1V . This is approximately the flat band voltage of an n-channel MOS FET with an n^+ silicon gate, if the silicon-silicon dioxide interface charge Q_{ss} is made negligible ($\lesssim 10^{11}$ electronic charges per cm^2 for the thin gate oxides considered--an easily achievable value).

M. Lenzlinger and E. H. Snow[2] have studied the conduction mechanism of SiO_2 in detail. They conclude that conduction is contact

rather than bulk limited and is due to electrons tunneling from the metal or silicon contact, through part of the SiO_2 energy gap, into the SiO_2 conduction band. Thus the current density for a given electric field is independent of oxide thickness x_o provided that x_o is large enough. For an n-channel MOS transistor with an Al or n^+ silicon gate the oxide current density is[2] $\simeq 10^{-10}$ A/cm^2 for an oxide electric field of $\pm 6 \cdot 10^6$ V/cm, provided that $x_o \gtrsim 50$Å. Since the current density raises rapidly with electric field and destructive breakdown[3] of the gate oxide occurs at an electric field somewhat higher than $6 \cdot 10^6$ V/cm, it is clear that practical devices must operate with gate oxide fields substantially lower than this value. For the present work we shall arbitrarily choose the maximum allowable oxide electric field in a practical device to be $F_{ox} = 3 \cdot 10^6$ V/cm.

The minimum size of a MOS transistor, for a given drain voltage and substrate doping concentration, will now be determined. The device geometry considered is shown in the inset of Fig. 2. We shall take the minimum channel length, limited by drain-source punch-through, to be twice the drain depletion region thickness at the maximum drain voltage. Then punch-through occurs at a voltage somewhat higher than the maximum drain voltage. Neglecting junction curvature*, the drain depletion region thickness is

$$W = \sqrt{\frac{2\varepsilon(V_{DD} + \varphi)}{q\, C_B}} \qquad (1)$$

*This is a reasonable approximation, since for the geometry considered the depletion region thickness is never greater than the junction radius of the curvature.

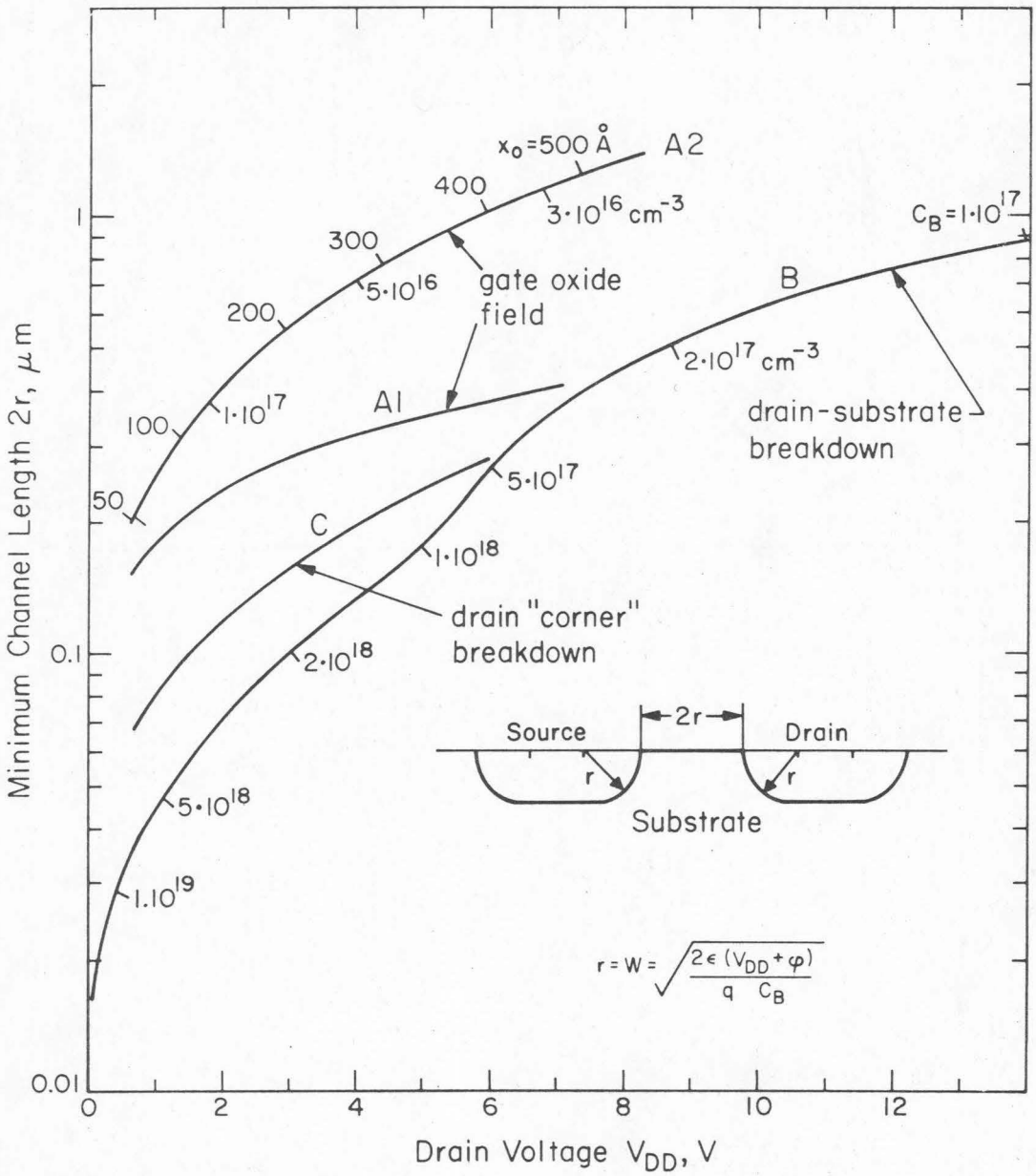

Fig. 2 Minimum channel length 2r of a MOS transistor, determined by oxide field (curves A1 and A2), drain-substrate breakdown (curve B) or drain "corner" breakdown (curve C), as a function of the drain voltage, V_{DD}. Curve A1 corresponds to the driver transistor and A2 to the pull up transistor of an inverter. The oxide thickness and substrate doping concentration of a minimum size pull up transistor are shown along curve A2.

where φ is the junction built in voltage. The minimum channel length $2r$, limited by drain-source punch-through, is obtained by setting $r \simeq W$.

Let us consider the gate oxide field limitation. The oxide field is a maximum near the edge of the source of the pull up transistor 2 when $V_o = 0V$ (see inset of Fig. 1A). The minimum gate oxide thickness of the pull up transistor 2 is obtained from Eq. (2A) of the appendix. The maximum substrate doping concentration is obtained from Eq. (3A) or from Fig. 1A. The minimum channel length is obtained from Eq. (1). The results are presented in Fig. 2, curve A2. It is assumed that gate oxide growth is a critical manufacturing step so that it is desirable to have both transistors 1 and 2 with the same oxide thickness x_o. (Conversely, the substrate doping concentration of transistor 1 could have been chosen equal to that of transistor 2. The oxide thicknesses of the transistors would then be different.) For a given oxide thickness, the substrate doping concentration of transistor 1 can be obtained from Eq. (1A) of the appendix and the required gate turn on voltage ($V_{GT1} = 0.5\ V_{DD}$). With this doping concentration the minimum channel length of transistor 1 is obtained from Eq. (1). The results are presented in Fig. 2, curve A1. Since both transistors have different substrate doping concentrations it is necessary to start with a wafer appropriate for the substrate of transistor 2, and then increase the doping concentration in the channel region of transistor 1 by ion implantation, for example.

For a given drain voltage, drain-substrate "breakdown" sets an upper limit to the substrate doping concentration, as shown in Fig. 2A,

page 45 . With this voltage and doping concentration, the minimum channel length $2r$ is calculated using Eq. (1). The results are presented in Fig. 2, curve B.

Drain "corner" breakdown can be estimated using an expression* by A. S. Grove et al[4]:

$$F_c \simeq \frac{2(V_D + \varphi)}{W} + \frac{(V_D + \varphi) - (V_G - V_{FB})}{\frac{\varepsilon}{\varepsilon_{ox}} x_o} \qquad (2)$$

Here F_c is the "corner" electric field and W the drain depletion region thickness in absence of the gate. V_D and V_G are the drain and gate voltages referred to the substrate. V_{FB} is the gate flat band voltage referred to the substrate, and ε_{ox} is the SiO_2 permittivity. Notice that the "corner" electric field is assumed to be simply the arithmetic sum of the drain junction electric field and the electric field induced in the silicon surface by the gate. When F_c reaches the critical value F_B shown in Fig. 4A, page 49 , drain "corner" breakdown occurs.

Let us again consider the inverter shown in the inset of Fig. 1A. The driver transistor 1 may have drain "corner" breakdown when its gate is low ($V_i = 0V$) and its drain is high ($V_o = V_{DD}$) . It is assumed that the gate oxide thickness x_o is chosen the same for both transistors. Then x_o is obtained, as before, by applying Eq. (2A) of the appendix to transistor 2. The minimum channel length $2r$

---

*To insure that the "corner" electric field is correct in the two limiting cases $W >> 3x_o$ and $W << 3x_o$, a factor of 2 has been added to the first term on the right hand side of Grove's[4] expression.

of transistor 1, limited by drain "corner" breakdown, is estimated by setting $r = W$, where W is obtained from Eq. 2 with $F_c = F_B \simeq 1.5 \cdot 10^6$ V/cm as shown in Fig. 4A. The results are presented in Fig. 2, curve C. The maximum substrate doping concentration limited by drain "corner" breakdown can be obtained from Eq. (1).

Notice that both the drain-substrate and drain "corner" breakdown limitations are less severe than the oxide field limitation. For this reason the junction radius of curvature can be made somewhat smaller than half the channel length as indicated in the inset of Fig. 2.

A minimum size transistor with $V_{DD} = 0.7V$ has a gate oxide thickness of 50Å as shown in Fig. 2. Since thinner oxides cannot be used due to tunneling from gate to substrate, $V_{DD} = 0.7V$ is a lower limit to the supply voltage of minimum size transistors. To reduce the supply voltage further it is necessary to reduce the substrate doping concentration, and therefore increase the device size.

1.3 Example

As a specific example we shall choose $V_{DD} = 2V$ and $V_{GG} = 4V$. The gate oxide thickness is calculated by applying Eq. (2A) of the appendix to transistor 2. The result is $x_o = 140$Å as indicated in Fig. 2, curve A2. The substrate doping concentration of transistor 2 is obtained from Eq. (1A) and the required gate turn on voltage ($V_{GT2} = 3V$ when $V_o = 2V$). The result is $C_{B2} = 9.2 \cdot 10^{16}$ cm^{-3} as indicated in Fig. 1A and in Fig. 2, curve A2. The substrate doping concentration of transistor 1, $C_{B1} = 2.7 \cdot 10^{17}$ cm^{-3}, is obtained from Eq. (1A), the oxide thickness $x_o = 140$Å, and the required gate turn

on voltage of transistor 1 (V_{GT1} = 1V). The maximum electric field in the gate oxide of transistor 1 is $1.5 \cdot 10^6$V/cm , which is smaller than F_{ox} .

For the voltages and doping concentrations considered in this example, drain-substrate breakdown and drain "corner" breakdown do not occur as shown in Fig. 2. From Eq. (1) the drain depletion region thickness is 0.12μm for transistor 1 and 0.205μm for transistor 2. The minimum channel length, limited by drain-source punch-through is approximately twice the drain depletion thickness or 0.24μm for transistor 1 and 0.41μm for transistor 2, as shown in Fig. 2, curves A1 and A2. A typical minimum size silicon gate MOS transistor is shown in Fig. 3. The drain-family and load line of the minimum size inverter we have just designed are presented in Fig. 4. These characteristics have been calculated using a MOS transistor model which includes velocity saturation of the charge carriers and is detailed in Appendix 3.

1.4 Doping Fluctuation Limitation

As the device size is reduced, the number of dopant atoms in a characteristic volume of the device becomes small enough so that its statistical fluctuations can no longer be neglected. The effect of substrate doping fluctuation is to alter the devices I-V characteristics, e.g., gate turn on voltage, and the devices breakdown characteristics, e.g., drain-source punch-through voltage. A chip with 10^6 devices will be considered. We shall require that, with an 80% certainty, the substrate doping fluctuations do not alter the gate turn on voltage or the punch-through voltage of any one of the 10^6 transistors by more than $\simeq$ 20% . This 20% variation corresponds to a substrate

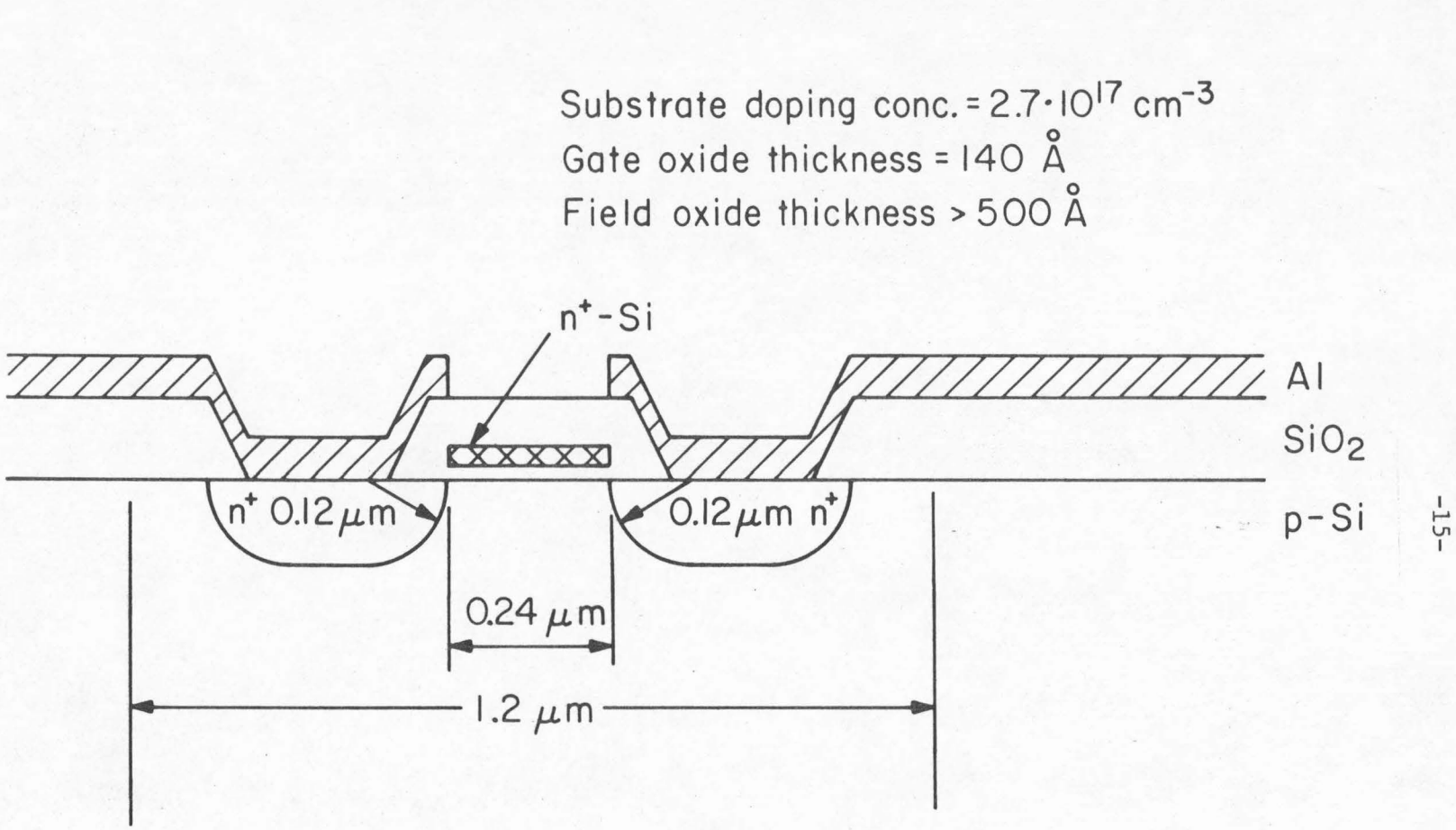

Fig. 3 Typical silicon gate MOS transistor of minimum size.

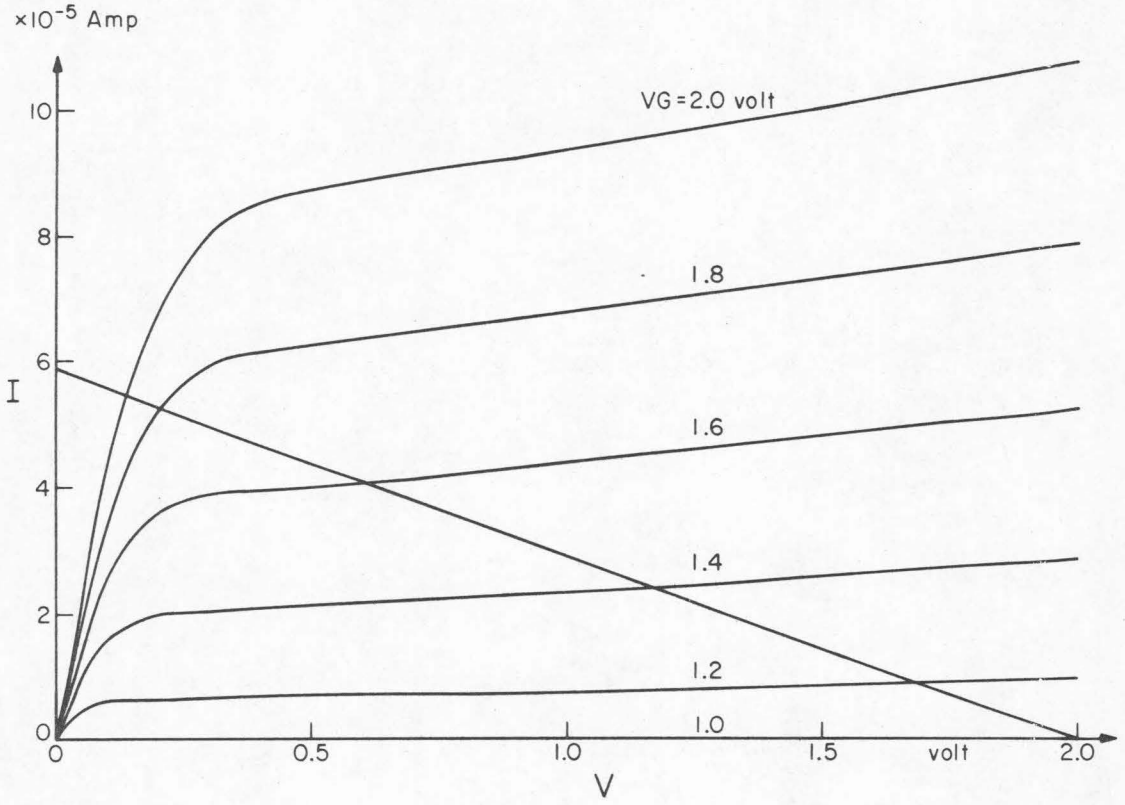

Fig. 4 Drain family of the driver transistor and load line of the pull-up transistor of a minimum size static inverter. $L_1 = Z_2 = 0.24$ µm, $L_2 = Z_1 = 0.96$ µm, $C_{B1} = 2.7 \cdot 10^{17}$ cm^{-3}, $C_{B2} = 9.2 \cdot 10^{16}$ cm^{-3}; $x_o = 140$ Å, $\mu = 250$ cm^2/V sec and $V_{FB} = -1.0$ V for both transistors. ($\alpha = 1$, see Appendix 3).

doping fluctuation of approximately 40% when measured in a volume W^3, W being a characteristic depletion thickness of the device. For a minimum size transistor with the geometry indicated in the inset of Fig. 2 we have $W \simeq r$. With an 80% certainty, the doping fluctuation does not exceed 40% in any one of the 10^6 cubes of volume r^3, if these cubes have on the average $\simeq 170$ ionized dopant atoms. The smallest size transistor shown in Fig. 2 corresponds to a driver transistor with a gate oxide thickness of 50Å, a substrate doping concentration of $4 \cdot 10^{17} cm^{-3}$, and a channel length $2r = 0.15\mu m$. Such a transistor has $\simeq 170$ dopant atoms in a volume r^3 of the substrate. Since this is an extreme case, we conclude that doping fluctuation is an important device limitation although less severe than oxide "breakdown".

1.5 Power Dissipation Limitation

In this section we shall show that for fully dynamic MOS circuits, the power dissipation density does not limit device size or packing density although it does set an upper limit to the frequency of operation. In static MOS circuits power dissipation is the most important limitation of the number of circuit functions per chip.

First we shall consider a fully dynamic or complementary inverter in which both transistors are never on simultaneously. Power dissipation occurs only when charging and discharging the load capacitance. It is assumed that each inverter output is connected to the input of the following inverter (fan out = 1), so that the load capacitance C is the sum of the gate and drain capacitance of transistor 1 (see inset of Fig. 1A). The power dissipation density of

densely packed dynamic inverters is

$$P = \frac{C V_{DD}^2}{S} f \qquad (3)$$

where f is the switching frequency, S the area occupied by an inverter, and $\frac{1}{2} CV_{DD}^2$ is the energy dissipated while charging or discharging the load capacitance C. It has been assumed that the clock driver is off the chip. The power dissipation required to gate the pull up transistor 2 on and off has not been taken into account, since it is dissipated off the chip. The power dissipation density at 10 MHz of several densely packed minimum size dynamic inverters is presented in Table 1.

In static inverters the gate voltage V_{GG} is constant so that the pull up transistor is always on. Thus, in addition to the power dissipation associated with charging and discharging the load capacitance, there is power dissipation due to current flowing through both transistors when they are simultaneously on. The drain characteristics of transistor 1 and the load line of a particular static inverter are shown in Fig. 4. From characteristics such as these the power dissipation density of several densely packed minimum size static inverters have been calculated assuming 50% duty cycle. The results are also presented in Table 1.

The power dissipation density of densely packed minimum size static inverters is seen to be very large. Thus power dissipation is the principal limitation of the number of circuit functions per chip, except in circuits such as read only memories in which only a small fraction of the devices dissipate most of the power.

TABLE 1. Power dissipation density P of an integrated silicon chip with densely packed, minimum size inverters. Assumptions: $V_{GG} = 2V_{DD}$, $V_{GT1} = 0.5\ V_{DD}$, $V_{GT2} = 1.5\ V_{DD}$. For dynamic inverter $L_1 = L_2 = Z_1 = Z_2 = 2r$. Inverter surface $S \simeq 90r^2$. Power calculated at 10 MHz. For static inverter $L_1 = Z_2 = 2r$, $L_2 = Z_1 = 8r$, $S \simeq 190r^2$. The load capacitance C is equal to the gate plus drain junction capacitances of transistor 1.

Type	V_{DD} volt	2r µm	x_o Å	C F	S cm^2	P W/cm^2
Dynamic	1	0.25	72	$4.6 \cdot 10^{-16}$	$1.4 \cdot 10^{-8}$	0.32
Dynamic	2	0.41	140	$6.7 \cdot 10^{-16}$	$3.8 \cdot 10^{-8}$	0.71
Dynamic	4	0.72	274	$11.0 \cdot 10^{-16}$	$12.0 \cdot 10^{-8}$	1.5
Static	1	0.18	72		$1.5 \cdot 10^{-8}$	960.0
Static	2	0.24	140		$2.7 \cdot 10^{-8}$	2000.0
Static	4	0.32	274		$4.9 \cdot 10^{-8}$	4100.0
Static	7	0.41	475		$8.0 \cdot 10^{-8}$	6900.0

The reason for this high power dissipation density is that for a minimum size static inverter the current through the pull up transistor is higher than necessary. If the current through the pull up transistor could be reduced until the charging time constant of the load capacitance were, say, 1/10 of one cycle, the current through the pull up transistor would be $10 \cdot CV_{DD} \cdot f$ when $V_o = 0$ volt, and the power dissipation density would be $P \simeq 6 \frac{CV_{DD}^2}{S} f$. In this case, as with a fully dynamic MOS FET circuit, power dissipation would only limit the operating frequency. The use of an MOS pull up transistor with the required current results in a channel length which is too long for the efficient use of area. This problem could be avoided if the pull up transistors are replaced by high Ω per square resistors. However, 10 MΩ resistors would typically be required. This problem will be discussed in detail for bipolar circuits.

1.6 Metal Migration Limitation

When a high current density flows through a metallic conductor, migration of the metallic atoms occurs[5]. This phenomenon is an important reliability consideration in both MOS and bipolar integrated circuit design. Divergence of the metallic migration current produces thinning of the conductor, which ultimately leads to catastrophic "strip burn out". Thus the instantaneous current density in aluminum conductors of integrated circuits should be kept substantially below $10^6 A/cm^2$.[5] Ohmic drop must also be considered. At $10^6 A/cm^2$ the ohmic drop in aluminum conductors is ≃3V/cm. We shall see that metal migration and power dissipation are two closely related limitations.

Let us consider the supply lines. For a given power dissipation P and supply voltage V_{CC}, the total average current that must be carried by the supply lines is $I = P/V_{CC}$. Therefore for a given power dissipation, supply voltage and conductor current density, the total supply line cross section area is independent of circuit complexity. Maintaining the present day power dissipation density and metalization thickness ($\simeq$ 1μm) we can expect the relative chip area occupied by supply lines to remain approximately independent of circuit complexity. For a given clock pulse rise and fall time constant τ the maximum instantaneous clock line current is $I \simeq CV_{CC}/\tau$, where C is the load capacitance. If τ is, say, 1/10th of a cycle, the total instantaneous current in the clock lines is $I \simeq 10P_d/V_{CC}$, where P_d is the dynamic power dissipation associated with C. The total current is again proportional to power dissipation, so we can expect that the relative chip area occupied by clock lines will also remain approximately independent of circuit complexity, provided that the metalization thickness is not reduced.

As power dissipation, metal migration does not limit the device size, but rather limits the packing density of static (noncomplementary) circuits and the frequency and/or packing density of dynamic (or complementary) circuits.

As an example, consider a chip with 10^6 fully dynamic minimum size inverters with $V_{DD} = 2V$ and V_{GG} switched between 0 and 4V. We shall assume that an aluminum line of width and thickness equal to $2r$ (i.e., 0.41μm) is connected to V_{GG} of 10^3 inverters. The gate capacitance of the 10^3 transistors is $\simeq$0.42 pF. With a maximum

allowable instantaneous current density in the metal line of $10^5 A/cm^2$ and a rise time equal to, say, 1/10 of a cycle, the maximum frequency of this particular circuit, limited by metal migration, would be 10 MHz.

2. Limitations of the Planar Bipolar Technology

2.1 Principal Limitations

To maximize the packing density of bipolar integrated circuits it is necessary to minimize the supply voltage and the size (or capacitance) of the individual devices.

The supply voltage cannot be lower than one diode drop ($\simeq$ 0.6V). Otherwise, the transistors could not be turned on. The minimum supply voltage for proper circuit operation is typically 2 to 3 diode drops depending on the circuit. For example, the minimum supply voltage of an RTL (resistor-transistor logic) gate is approximately two diode drops.

For a given supply voltage and doping concentration profile, the minimum transistor size is determined by punch-through, a condition where depletion regions overlap. To further reduce the device size it is necessary to reduce the depletion thicknesses by increasing doping concentrations and reducing the supply voltage. The maximum doping concentrations are determined by junction field emission "breakdown". (At high doping concentrations the principal reverse-conduction mechanism of p-n junctions is tunneling of carriers across the junction, as discussed in Appendix 2). Thus a minimum size transistor has its breakdown voltage equal to its punch-through voltage. Statistical fluctuations of the doping concentration can reduce the breakdown or punch-through voltages. It is therefore necessary to set the supply voltage somewhat lower than the breakdown or punch-through voltage of the device.

It will be shown that the maximum packing density of fully active bipolar circuits is determined by power dissipation. The packing density of integrated circuits which are not fully active (e.g., read-only memories in which only a small fraction of the devices dissipate most of the power) becomes limited by the area occupied by transistors, interconnections, and passive devices.

The analysis is necessarily approximate since it requires a number of assumptions such as the geometry of the devices, the circuit configuration and the maximum allowable power dissipation.

2.2 Breakdown and Punch-Through Limitation

An isoplanar bipolar transistor is shown in Fig. 5a. Since the base region is lightly doped compared to the emitter and collector, the depletion regions of the two junctions extend mainly into the base. For given voltages and base doping concentration, the minimum base region thickness is determined by punch-through, a condition where the depletion regions extending from the emitter and collector junctions overlap. The maximum base doping concentration C_B is determined by collector junction "breakdown" as shown in Fig. 2A of the appendix, page 45. The minimum base thickness r will be set equal to $r_C + r_E$, where $r_C = \sqrt{2\varepsilon(V_{CB}+\varphi)/qC_B}$ is the collector junction depletion thickness and $r_E = \sqrt{2\varepsilon\varphi/qC_B}$ is the emitter junction depletion thickness when the emitter is connected to the base. V_{CB} is the collector-base voltage, ε the silicon permittivity, φ the junction built-in voltage and q the electronic charge. The minimum base thickness r of an isoplanar transistor, determined by collector junction breakdown and base punch-through, is presented in Fig. 6 curve A. The

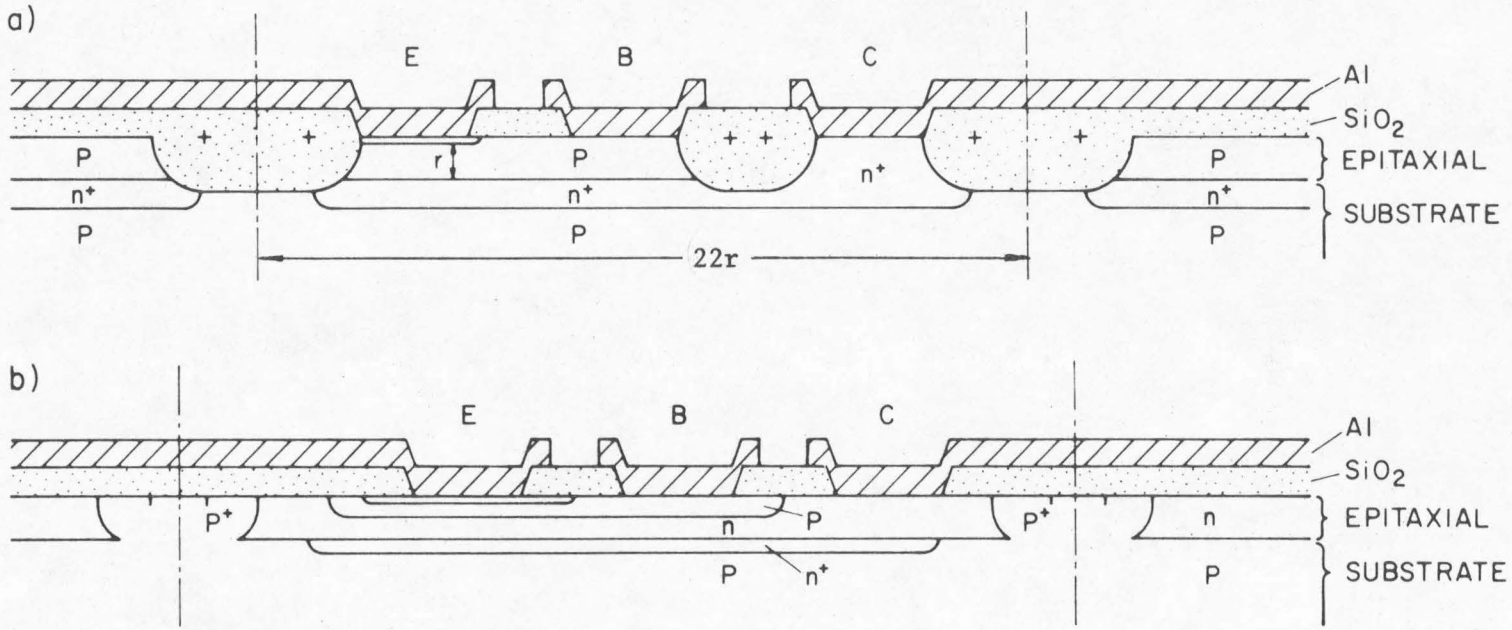

Fig. 5 Approximate cross section of minimum size bipolar transistors. An isoplanar transistor is shown in Fig. 5a and a transistor with diffused isolation is shown in Fig. 5b. The minimum size isoplanar transistor occupies an area of approximately 22r x 9r ≈ $200r^2$, where r is the base thickness.

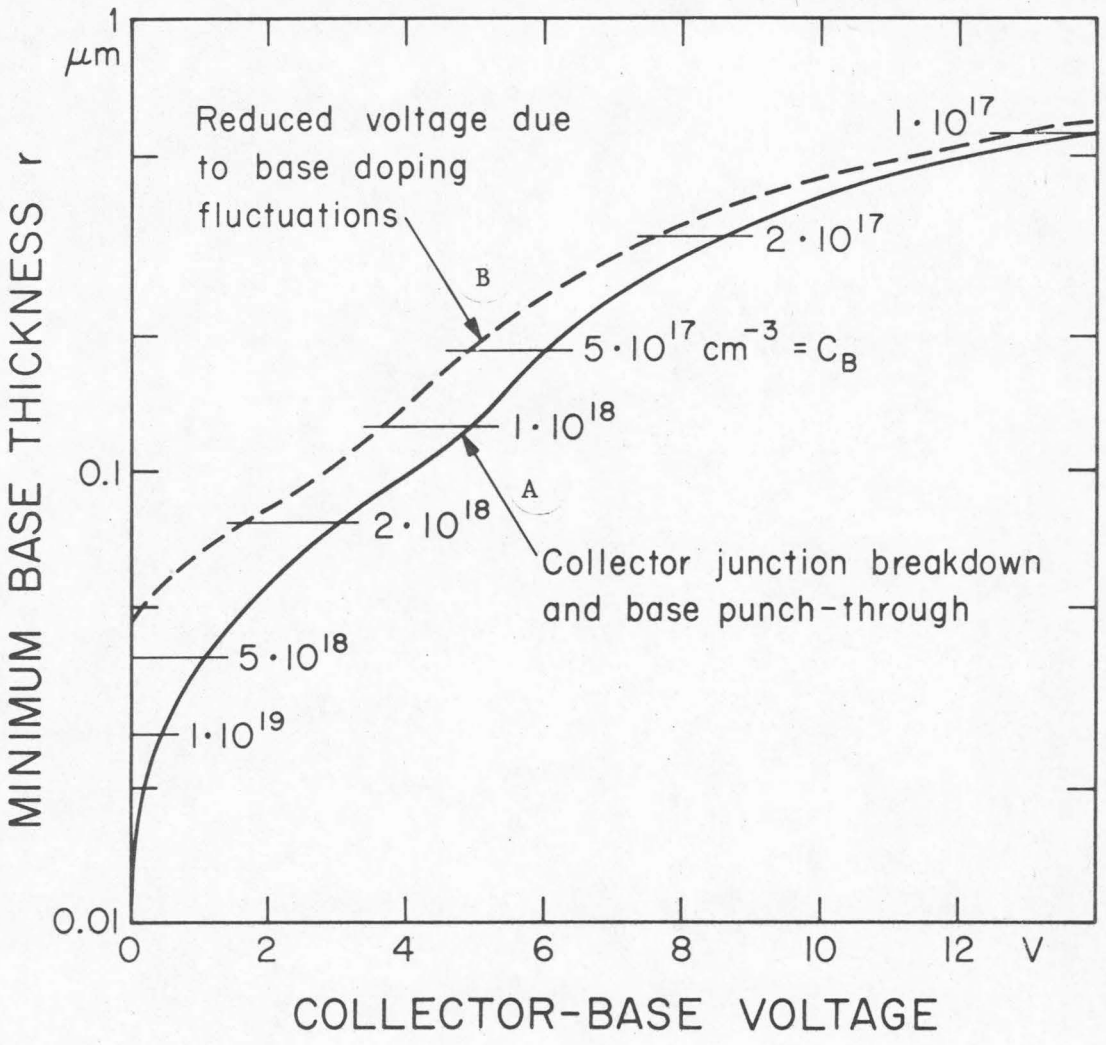

Fig. 6 Solid line indicates the minimum base thickness of an isoplanar transistor for a given collector-base voltage, determined by collector-junction breakdown and base punch-through. The collector-base voltage must be derated due to base doping fluctuations as indicated by the dashed line.

maximum substrate doping is equal to C_B , since both the substrate and base doping concentrations are limited by collector junction breakdown. If the substrate has a doping concentration C_B , the depletion region extending from the collector into the substrate has a thickness r_C . For a given base thickness, the minimum size of an isoplanar transistor can be determined approximately by geometrical considerations as shown in Fig. 5a. The isolation regions are broad enough to avoid punch-through between the collectors of adjacent transistors. The etch required to define the isolation regions is assumed isotropic and the silicon dioxide layer is grown thick enough to avoid inversion of the underlying silicon. The minimum size isoplanar transistor shown in Fig. 5a, with a substrate doping concentration equal to C_B and with square emitter, base and collector contacts, occupies an area of $\simeq 200\ r^2$ and has a collector capacitance of $\simeq 350 \cdot \varepsilon \cdot r$. If the circuit packing density is limited by power dissipation rather than by the area occupied by the devices, it is convenient to reduce the collector capacitance by reducing the substrate doping concentration. The isolation regions must then be made broader to avoid punch-through between the collector regions of adjacent transistors. A transistor with the geometry shown in Fig. 5a except for broader isolation regions, with square emitter, base and collector contacts and with a substrate doping concentration equal to $\frac{1}{10} C_B$, occupies an area $\simeq 280\ r^2$ and has a collector capacitance of $\simeq 180 \cdot \varepsilon \cdot r$.

A planar bipolar transistor with diffused isolation is shown in Fig. 5b. The area occupied by a minimum size transistor of this type can be estimated by making several approximations. The emitter region is degenerate and has a doping concentration of $\simeq 10^{21} cm^{-3}$. The doping concentration of the base, which is made lower than that of the emitter

to insure good emitter efficiency, can be as high as $\simeq 10^{19} cm^{-3}$. We make the simplifying assumptions that the doping concentration of the epitaxial collector region is considerably lower than that of the base, and that the collector junction depletion region does not reach the n^+ buried collector. For a given collector-base voltage, the maximum collector doping concentration C_C determined by collector junction breakdown can then be obtained directly from Fig. 2A of the Appendix. The collector junction depletion thickness is $r_C \simeq \sqrt{2\varepsilon(V_{CB}+\varphi)/qC_C}$. With this depletion thickness, the minimum size of a transistor with diffused isolation can be estimated using geometrical considerations as shown in Fig. 5b. The result is that a minimum size transistor with diffused isolation occupies $\simeq 2.1$ times more area and has a collector capacitance $\simeq 2.5$ times greater than a minimum size isoplanar transistor with the same voltage rating, contact area, and substrate doping concentration. The increased area and capacitance of the transistor with diffused isolation is mainly due to the separation between the base and isolation diffusion which is required to avoid punch-through between these regions. In what follows only the isoplanar transistor is considered.

2.3 Doping Fluctuation Limitation

As devices are made smaller the number of dopant atoms in a characteristic volume of the device decreases until its statistical fluctuations become important. Consider a minimum size isoplanar transistor designed to have a collector junction breakdown voltage equal to the base punch-through voltage V_{CB}, as in the previous section. The effect of base doping fluctuation is to decrease the breakdown voltage (if C_B increases) or decrease the punch-through voltage (if C_B decreases).

Thus due to base doping fluctuations the maximum allowable collector-base voltage must be derated. A chip with 10^6 isoplanar transistors will be considered. We shall arbitrarily choose the maximum allowable collector-base voltage V_{CBM} in such a way that, with an 80% certainty, none of the 10^6 transistors will have a collector breakdown or punch-through voltage lower than V_{CBM} .

The amount by which the collector-base voltage V_{CB} must be derated due to base doping fluctuations can be estimated as follows. The breakdown or punch-through voltages are altered significantly only if the base doping fluctuation causes a doping concentration variation in a volume $\simeq r_C^3$ or greater; r_C is the collector junction depletion thickness. The collector junction depletion region of a minimum size isoplanar transistor such as the one shown in Fig. 5a, can typically be divided into 30 cubes of volume r_C^3. The doping fluctuation ΔC_B which, with an 80% certainty, is not exceeded in any of the $30 \cdot 10^6$ cubes of volume r_C^3 is given by

$$\Delta C_B \cong 5.8 \cdot \frac{\sqrt{n}}{r_C^3} = 5.8 \cdot \sqrt{\frac{C_B}{r_C^3}} \cdot \qquad (4)$$

This calculation assumes that the number of dopant atoms in a volume r_C^3 has a Gaussian distribution with mean $n \equiv C_B\, r_C^3$ and standard deviation $\sqrt{n}$. The result does not depend strongly on the number of cubes assumed. The maximum allowable collector-base voltage V_{CBM} is determined either by breakdown (with a base doping concentration $C_B + \Delta C_B$) or by punch-through (with a doping concentration $C_B - \Delta C_B$), whichever is smaller, and is presented in Fig. 6 curve B. Notice that the maximum collector-base voltage determined by breakdown and punch-through must be derated by

about 1.4V due to base doping fluctuations. In a practical design the collector-base voltage must be derated further due to the manufacturing tolerances of the base doping concentration .

2.4 Power Dissipation Limitation

All bipolar transistor circuits have some sort of current limiting devices such as resistors and current sources. These current limiting devices can usually not be avoided even in complementary circuits due to the low impedance of the base-emitter junctions. The power dissipation of a circuit can be divided into static power dissipation (associated with the steady state currents) and dynamic power dissipation (associated with transient current which charge and discharge the circuit capacitances). The dynamic power dissipation is proportional to frequency. The static power dissipation of digital circuits is determined primarily by the current limiting devices in the circuit and not by the active transistors, since these are either on or off. The resistors can be chosen to obtain the desired static power dissipation density. These resistors and the circuit capacitances then determine the circuit time constants, which in turn set an upper limit to the frequency of operation. The maximum frequency of operation of bipolar integrated circuits is therefore determined either by dynamic power dissipation, by the circuit time constants which depend on the static power dissipation, or, in saturating circuits, by the lifetime of minority carriers in the base.

Let us examine the simple circuit shown in the inset of Fig. 7. The circuit consists of a series of resistor-transistor logic (RTL) inverters connected in cascade (fan out = 1), so that half of the inverters are in the high level state and half are in the low level state.

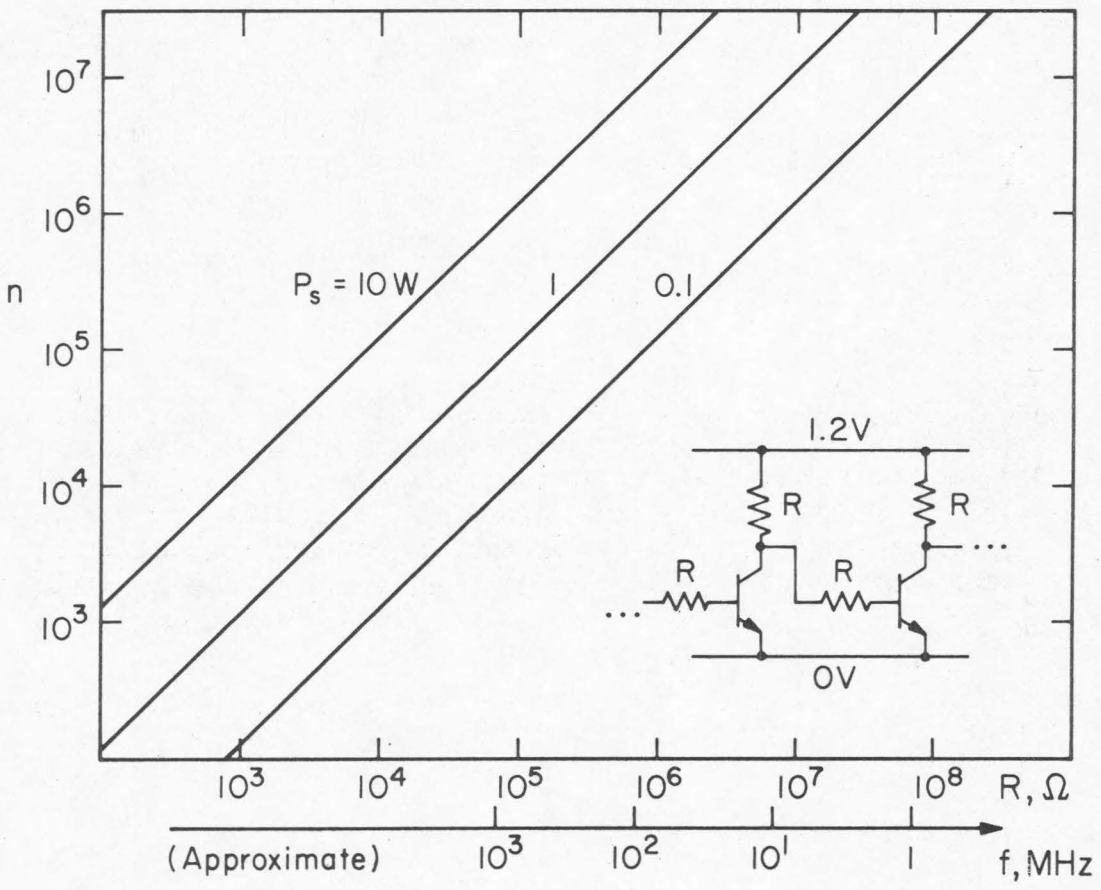

Fig. 7 For a given frequency of operation f and static power dissipation P_s, the maximum number of inverters n and resistor value R are determined for the particular circuit indicated. At the frequency f the dynamic power dissipation is $P_d \approx 0.36\ P_s$. It is assumed that the equivalent capacitance of the resistors is equal to the collector junction capacitance of the transistors (i.e. $180 \cdot \varepsilon \cdot r$ with $r = 0.07\mu m$).

The supply voltage is 1.2V so the minimum base thickness of the isoplanar transistors is $r = 0.07\mu m$ as shown in Fig. 6. If the substrate doping concentration is equal to 1/10th of the base doping concentration, the minimum size isoplanar transistor occupies an area of $\simeq 280\ r^2$ and has a collector capacitance of $\simeq 180 \cdot \varepsilon \cdot r$, as indicated in a previous section. We assume that each resistor (including inter-connections) occupies the same area as a transistor, so that each inverter occupies $\simeq 840\ r^2$. This area corresponds to a maximum density of $\simeq 2.4.10^7$ inverters per cm^2. At this packing density, the value R of the resistors required to have a static power dissipation density of 1 W/cm^2 is $\simeq 20\ M\Omega$. The rise-time of the collector of a transistor that is turned off is approximately $\tau = RC$, where C is the total capacitance associated with the collector node, i.e., collector junction capacitance plus capacitance of the two resistors connected to the collector. In our example $\tau \simeq 77$ nsec if we assume that the equivalent capacitance of each resistor is equal to the collector junction capacitance. Assuming that the recovery time of the transistor is shorter than τ, the maximum frequency of operation is determined by the collector risetime and is of the order of $1/5\tau \simeq 2.6MHz$. The dynamic power dissipation density at 2.6MHz is approximately $0.36W/cm^2$.

This example illustrates two important limitations of densely packed bipolar integrated circuits. First, the resistor values required to have a reasonable static power dissipation density are exceedingly large. Secondly, the maximum frequency of operation is limited to quite low values by both the circuit time constants and by dynamic power dissipation. Notice that at the maximum frequency $f = 1/5RC$, the dynamic

power dissipation of n inverters is $P_d = nf\,CV^2 = nV^2/5R$, which is approximately 0.36 P_s for the circuit considered above and is independent of C. P_s is the static power dissipation of n inverters.

For a given frequency of operation f and maximum power dissipation per chip P , the maximum number of inverters per chip n and the resistor value R can be determined from Fig. 7 for the particular circuit considered.

We conclude that the minimum power dissipation per circuit function is determined by the required frequency of operation and by the supply voltage and circuit capacitances. This minimum power dissipation sometimes can not be achieved because high enough ohm per square resistors are not available. Power dissipation limits the number of circuit functions per chip in fully active bipolar circuits. In circuits which are not fully active such as read-only memories in which only a small fraction of the devices dissipate most of the power, the area occupied by the devices and interconnections becomes the limiting factor.

2.5 Metal Migration Limitation

The considerations on metal migration made for MOS circuits are also valid for bipolar circuits. We limit our discussion here to a particular example. Consider a 1 cm^2 chip with 10^6 minimum size inverters, a supply voltage of 1.2V, a static power dissipation of 1W and a dynamic power dissipation of 0.36W at 100 MHz, as shown in Fig. 7. We assume that a supply line is connected to $2 \cdot 10^3$ inverters. The maximum mean current in this line is 2.3 mA. If a maximum current density of $10^5 A/cm^2$ is allowed, the minimum conductor cross section area required would be 2.3 $(\mu m)^2$.

For a metallization thickness of 1μm, all supply lines would occupy an area of approximately 0.23 cm^2.

2.6 High Valued Resistors

To improve the yield it is convenient to use the same processing steps to make the transistors and resistors in the integrated circuit. The highest sheet resistance is obtained by using the base region of the isoplanar transistors. Let us consider a minimum size 1.2V transistor with a base region thickness of $r = 0.07\mu m$ and a base doping concentration of $2.3 \cdot 10^{18} cm^{-3}$ as shown in Fig. 6. The I-V characteristic of a one-square pull up resistor is calculated using field effect transistor equations and is presented in Fig. 8. The current at $v = 0V$ is $\simeq 5 \cdot 10^{-5} A$ which corresponds to a linear resistor of $\simeq 24$ KΩ per square.

Alternative high ohm per square resistors are MOS resistors and cermet (ceramic-metal) thin film resistors. Sheet resistances of 7-25 KΩ per square are obtained with good controllability by using nonsaturated MOS transistors as resistors.(6) For example, the minimum size MOS pull up transistor considered in Fig. 4 has a sheet resistance of $\simeq 9K\Omega$ per square. Cermet resistors are manufactured with sheet resistances up to $\simeq 10K\Omega$ per square.

The maximum attainable sheet resistance is an important limitation of fully active bipolar circuits. Let us again consider the circuit shown in the inset of Fig. 7. The maximum packing density is obtained by increasing the length to width ratio of the resistors until the densly packed chip has an acceptable power dissipation density. The base region of a minimum size 1.2V isoplanar transistor has a sheet resistance of $\simeq 24K\Omega$ per square as shown in Fig. 8. The minimum width

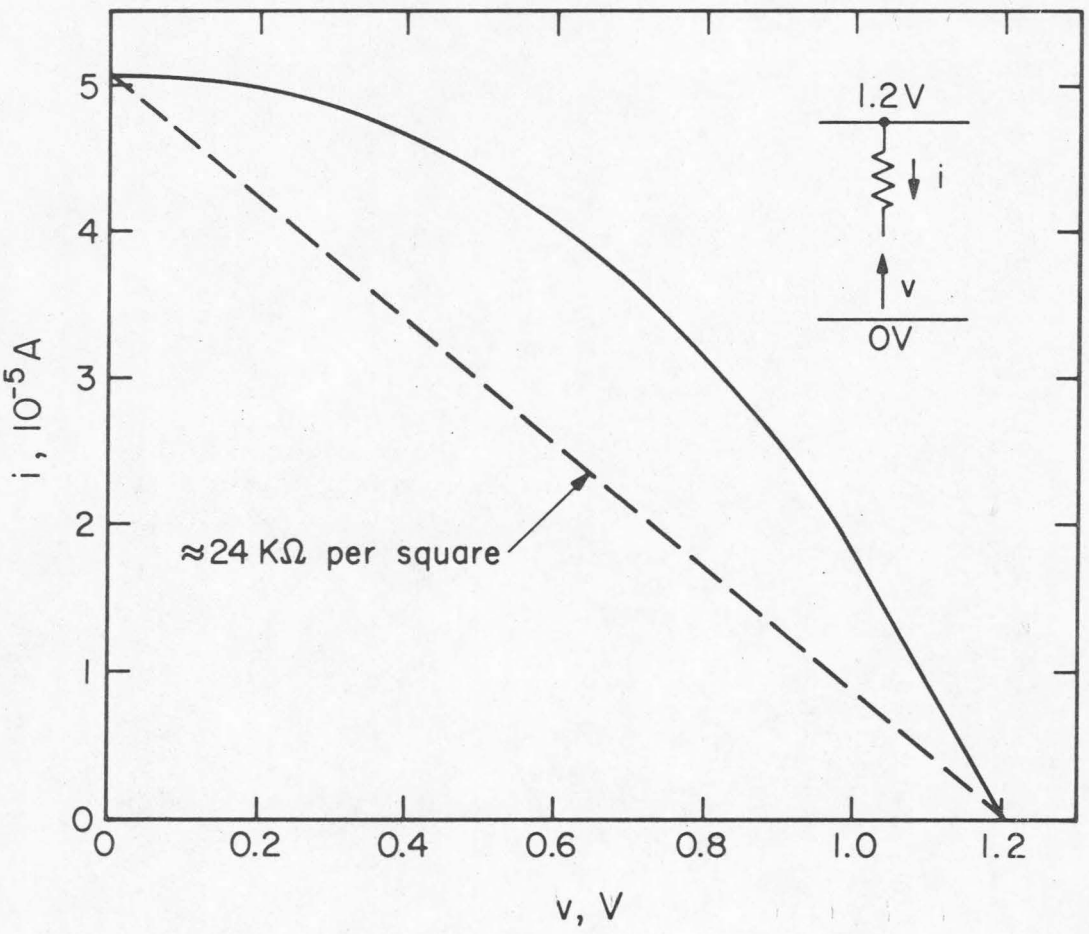

Fig. 8 I-V characteristic of the base region of a minimum size 1.2V isoplanar transistor used as a pull-up resistor. The emitter and collector regions are connected to 1.2V. The base width and length are equal.

resistor occupies $\simeq 6(\mu m)^2$ of chip area per MΩ . The static power dissipation density of a densly packed chip is $1 W/cm^2$ if the length to width ratio of the resistors is chosen to be $\simeq 110$. In this case most of the chip area is occupied by the resistors. The packing density is then $\simeq 6.5 \cdot 10^6$ resistors per cm^2 and the maximum frequency of operation is $\simeq 2$ MHz.

3. Conclusions

The maximum packing density of planar integrated circuits is obtained by minimizing the supply voltages and the area occupied by the devices. The principal physical limitations of MOS transistors which determine the minimum device size for given supply voltages are oxide breakdown, drain-substrate breakdown, drain "corner" breakdown and substrate doping fluctuations. These four limitations determine minimum device sizes of the same general order of magnitude, oxide breakdown being the most severe limitation as shown in Fig. 2. The minimum size of isoplanar bipolar transistors for a given supply voltage, is determined by collector junction breakdown, base punch-through and base doping fluctuations as shown in Fig. 6.

A minimum size 1.2V MOS driver transistor with a channel length to width ratio of 1 has a channel length of ≈ 0.2 μm, a gate oxide thickness of ≈ 83 Å, a substrate doping concentration of $3 \cdot 10^{17}$ cm^{-3}, an area of ≈ 0.6 $(\mu m)^2$ and a gate plus drain junction capacitance of $\approx 4 \cdot 10^{-4}$ pF. A minimum size 1.2V isoplanar bipolar transistor with a substrate doping concentration equal to 1/10th the base doping, has a base thickness of ≈ 700 Å, a base doping concentration of $\approx 2.3 \cdot 10^{18}$ cm^{-3}, an area of

$\approx 1.4\ (\mu m)^2$ and a collector capacitance of $\approx 1.3 \cdot 10^{-3}$ pF. For minimum size devices of equal voltage rating, the MOS transistor has a factor of 2-3 advantage in area and capacitance over the bipolar transistor.

In fully active bipolar or static non-complementary MOS circuits the number of devices per chip is limited by power dissipation. The required frequency of operation and the circuit capacitances determine an upper limit to the resistor values. The available resistors (diffused, MOS or cermet) often do not have a high enough sheet resistance for the efficient use of area. The resistors and the supply voltage determine the power dissipation per circuit function. If resistors and bipolar transistors are to be made with the same manufacturing steps, it is convenient to use the base region of the isoplanar transistors as depletion mode resistors. Then sheet resistances of 10 to 30 KΩ per square are obtained. As an example, let us consider a fully active circuit with a supply voltage of 1.2V as shown in Fig. 7. If base resistors with a length to width ratio of 10 are used, a static power dissipation of 1 W/cm^2 is obtained with $\simeq 3 \cdot 10^5$ inverters per cm^2 (i.e. $9 \cdot 10^5$ devices per cm^2). Taking into account the parasitic capacitances, the maximum frequency of operation of this circuit is $\approx$ 150 MHz as shown in Table 2. At 150 MHz the total power dissipation is ≈ 1.4 W/cm^2.

The power dissipation per transistor in read-only memories is low because there are many driver transistors per pull up resistor. The maximum packing density of read-only memories is therefore limited by the area occupied by the devices and interconnections. The maximum frequency is limited by metal migration and by the circuit time constants. Both MOS and bipolar read-only memories require two level metalization.

Table 2: Maximum packing density (devices per cm^2) and corresponding maximum frequency of operation (MHz) of several bipolar and MOS transistor integrated circuits

	Isoplanar Bipolar		Static non-complementary MOS		Dynamic MOS	
	Density	Frequency	Density	Frequency	Density	Frequency
	(1)	(2)	(1)	(2)		
1.2V inverters with 240 KΩ diffused or MOS resistors	$9 \cdot 10^5$	150	$9 \cdot 10^5$	200	-	-
					(3)	(4)
2V fully dynamic inverters	-	-	-	-	$3 \cdot 10^7$	30
	(3)	(6)	(3)	(6)		
1.2V, 1 cm^2 read only memory (5)	$1 \cdot 10^8$	0.5	$1 \cdot 10^8$	0.5	-	-

(1) Determined by static power dissipation $P_s = 1W/cm^2$
(2) Determined by the circuit time constants
(3) Determined by the area occupied by transistors and interconnections
(4) Determined by dynamic power dissipation $P_d = 2.1W/cm^2$ and/or metal migration
(5) Two level metallization assumed
(6) Determined by metal migration and/or the circuit time constants.

A 1 cm^2 static read-only memory can have up to $\approx 1.10^8$ transistors operating at a maximum frequency of $\approx$ 0.5 to 2 MHz, as shown in Table 2. The frequency of operation can be increased by decreasing the packing density or the area of the circuit. At these high packing densities, the maximum frequency is approximately inversely proportional to the number of devices in the read-only memory.

Fully dynamic or complementary MOS circuits have essentially no static power dissipation. The packing density of these circuits is therefore limited by the area occupied by the transistors and interconnections, and the frequency is limited by dynamic power dissipation as shown in Table 2 and Fig. 1.

We conclude that read-only memories will reach approximately the same performance and packing density with MOS and bipolar technologies, while fully active circuits will reach the highest levels of integration with dynamic MOS or complementary MOS technologies.

Two recent developments have been charge coupled devices and complementary bipolar circuits. Present day MOS charge coupled shift registers occupy approximately 1/4 the area of MOS transistor shift registers[7] due to the elimination of the supply lines and the source and drain diffusion regions. Charge coupled devices (CCD's) have gate oxide field and punch-through limitations similar to those of ordinary MOS transistors. We can therefore expect the maximum packing density of CCD shift registers to be of the order of 4 times greater than that of MOS transistor shift register, as with present masking techniques. In complementary bipolar circuits most of the static power dissipation is due to the base current of the transistors. The power dissipation per

transistor is therefore lower than in conventional circuits (e.g. Fig.7) but higher than in complementary MOS circuits. The maximum packing density of fully active complementary bipolar circuits lies between 10^6 and 10^7 devices per cm^2.

To manufacture minimum size devices mask alignment tolerance of 100 to 200 Å are required. Such tolerances are compatible with electron beam pattern generation techniques. For example, 1000 Å metal lines have already been fabricated[8]. We can therefore envision integrated circuits approaching the packing densities summarized in Table 2 in the near future.

REFERENCES - PART I

1. B. E. Deal et al., J. Electrochem. Soc. 114, 266 (1967).

2. M. Lenzlinger and E. H. Snow, J. Appl. Phys. 40, 278 (1969).

3. N. Klein, IEEE Trans. on Electron Devices, ED-13, 788 (1966).

4. A. S. Grove, O. Leistiko, and W. W. Hooper, IEEE Trans. on Electron Devices, ED-14, 157 (1967).

5. I. A. Blech and E. S. Meieran, Appl. Phys. Lett. 11, 263 (1967).

6. L. Vadasz, IEEE Trans. on Electron Devices, ED-13, 459 (1966).

7. L. Altman, Electronics 44, 13,50 (1971).

8. S. Magdo, M. Hatzakis, and C. H. Ting, IBM Jour. Research and Development 15, 446.

9. H. Weinerth, Solid-State Elect. 10, 1053 (1967).

10. A. G. Chynoweth et al., Phys. Rev. 118, 425 (1960).

11. S. L. Miller, Phys. Rev. 105, 1246 (1957).

12. J. Shields, J. Electronics and Control 6, 130 (1959).

13. J. C. Irvin, Bell System Tech. J. 41, 387 (1962).

14. R. A. Logan and A. G. Chynoweth, Phys. Rev. 131, 89 (1963).

15. S. M. Sze and G. Gibbons, Solid-State Elect. 9, 831 (1966).

16. D. Frohman-Bentchkowsky and L. Vadasz, IEEE Jour. Solid-State Circuits, SC-4, 57-64 (1969).

17. D. Frohman-Bentchkowsky and A. S. Grove, IEEE Trans. on Electron Devices, ED-16, 108-113 (1969).

APPENDIX 1: MAXIMUM SUBSTRATE DOPING CONCENTRATION

Circuit design considerations frequently require that the gate turn on voltage have a specified value V_{GT} at a specified source voltage V_S. This requirement and the maximum allowable gate oxide field F_{ox} set an upper limit to the substrate doping concentration.

The gate turn on voltage is

$$V_{GT} = V_{FB} + V_S + 2\phi + \frac{x_o}{\varepsilon_{ox}} \sqrt{2\varepsilon q\, C_B (V_S + 2\phi)} \tag{1A}$$

ϕ is the energy difference in eV between the Fermi level and the intrinsic Fermi level in the bulk of the substrate.

The minimum oxide thickness is

$$x_{o\ min} = \frac{(V_{G\ max} - V_{FB}) - (V_{S\ min} + 2\phi)}{F_{ox}} \tag{2A}$$

Here $V_{G\ max} - V_{S\ min}$ is the maximum gate-source voltage. The maximum substrate doping concentration is determined from Eqs. (1A) and (2A) by setting $x_o = x_{o\ min}$. The result is

$$C_{B\ max} = \left[\frac{V_{GT} - V_{FB} - V_S - 2\phi}{V_{G\ max} - V_{FB} - V_{S\ min} - 2\phi}\right]^2 \frac{\varepsilon_{ox}^2 F_{ox}^2}{2\varepsilon q (V_S + 2\phi)} \tag{3A}$$

The particular circuit shown in the inset of Fig. 1A will now be considered. To be specific we shall require that $V_{GT1} = \frac{1}{2} V_{DD}$ and $V_{GT2} = V_{GG} - \frac{1}{2} V_{DD}$ when $V_o = V_{DD}$. Here V_{GT1} and V_{GT2} are the gate turn on voltages of transistors 1 and 2 respectively (see

Fig. 1A). The maximum substrate doping concentration limited by oxide field is obtained by applying Eq. (3A) to each transistor. For transistor 1

$$C_{B1} \leq \left[\frac{\frac{1}{2} V_{DD} - V_{FB} - 2\phi}{V_{DD} - V_{FB} - 2\phi}\right]^2 \frac{\varepsilon_{ox}^2 F_{ox}^2}{2\varepsilon q \, 2\phi} \tag{4A}$$

For transistor 2

$$C_{B2} \leq \left[\frac{V_{GG} - \frac{3}{2} V_{DD} - V_{FB} - 2\phi}{V_{GG} - V_{FB} - 2\phi}\right]^2 \frac{\varepsilon_{ox}^2 F_{ox}^2}{2\varepsilon q (V_{DD} + 2\phi)} \tag{5A}$$

Equation (5A), which is a more severe limitation than Eq. (4A), is plotted in Fig. 1A for the case $V_{FB} = -1$ V and $F_{ox} = 3 \cdot 10^6$ V/cm.

APPENDIX 2: REVERSE BREAKDOWN OF LOW VOLTAGE SILICON JUNCTIONS

Several authors[9-12] have measured the reverse "breakdown" voltage of one-sided silicon step junctions. Their results are presented in Fig. 2A. The "breakdown" voltage V_B is defined as the applied voltage at a specified reverse current density. H. Wienerth[9] has shown that field emission is the main reverse conduction mechanism of low voltage diodes ($V_B \lesssim 3$V), whereas high voltage diodes ($V_B \gtrsim 8$V) are limited by avalanche breakdown. The reverse characteristics of diodes in the intermediate range ($3\text{V} \lesssim V_B \lesssim 8\text{V}$) can be explained[9] by avalanche multiplication of the field emission current.

A reverse biased n^+p junction is shown in Fig. 3A. Electrons can tunnel through the energy gap from the p to the n^+ side as shown in

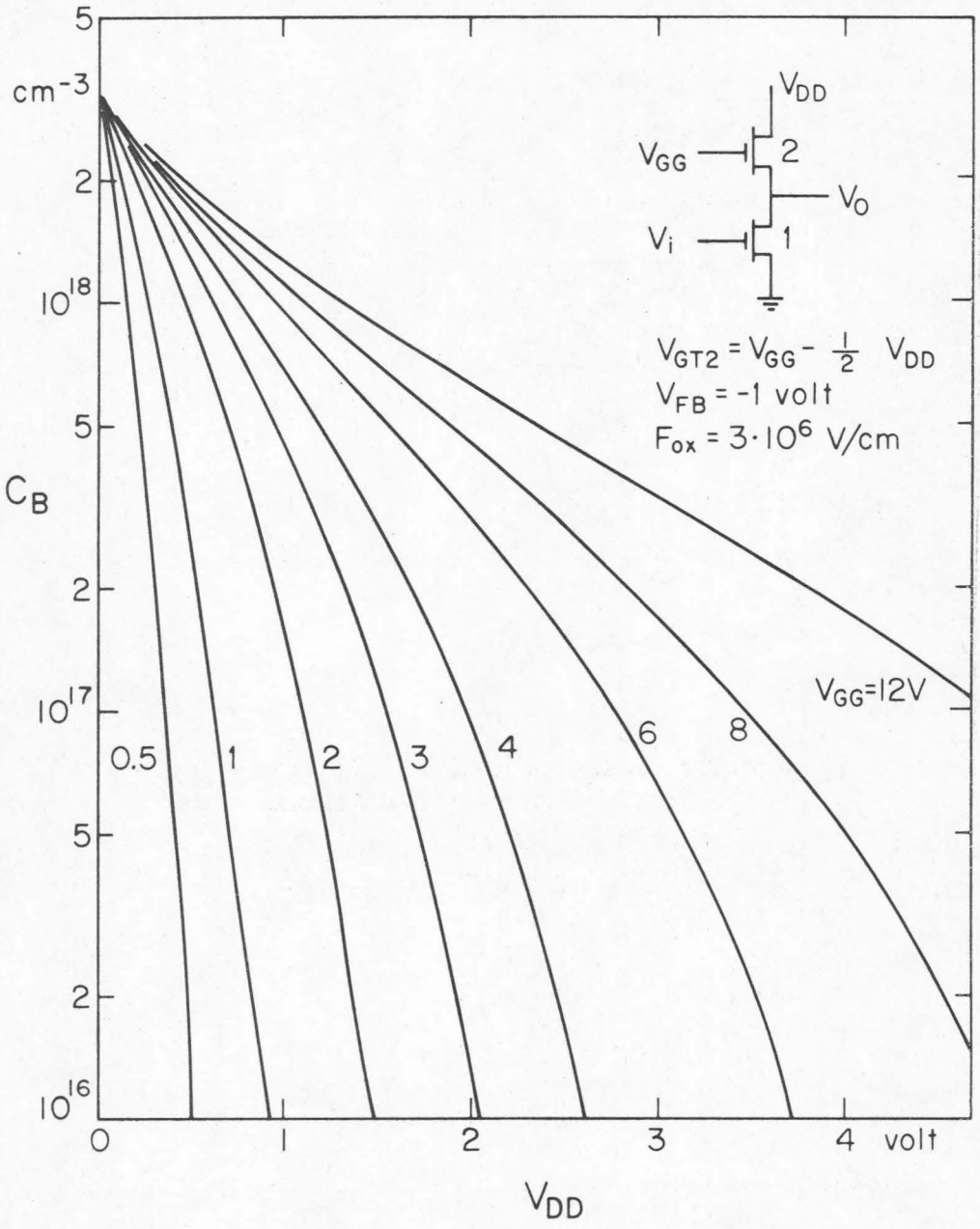

Fig. 1A Maximum substrate doping concentration C_B of the pull up transistor of an inverter, as a function of V_{DD} and V_{GG}, determined by the maximum allowable gate oxide electric field F_{ox}.

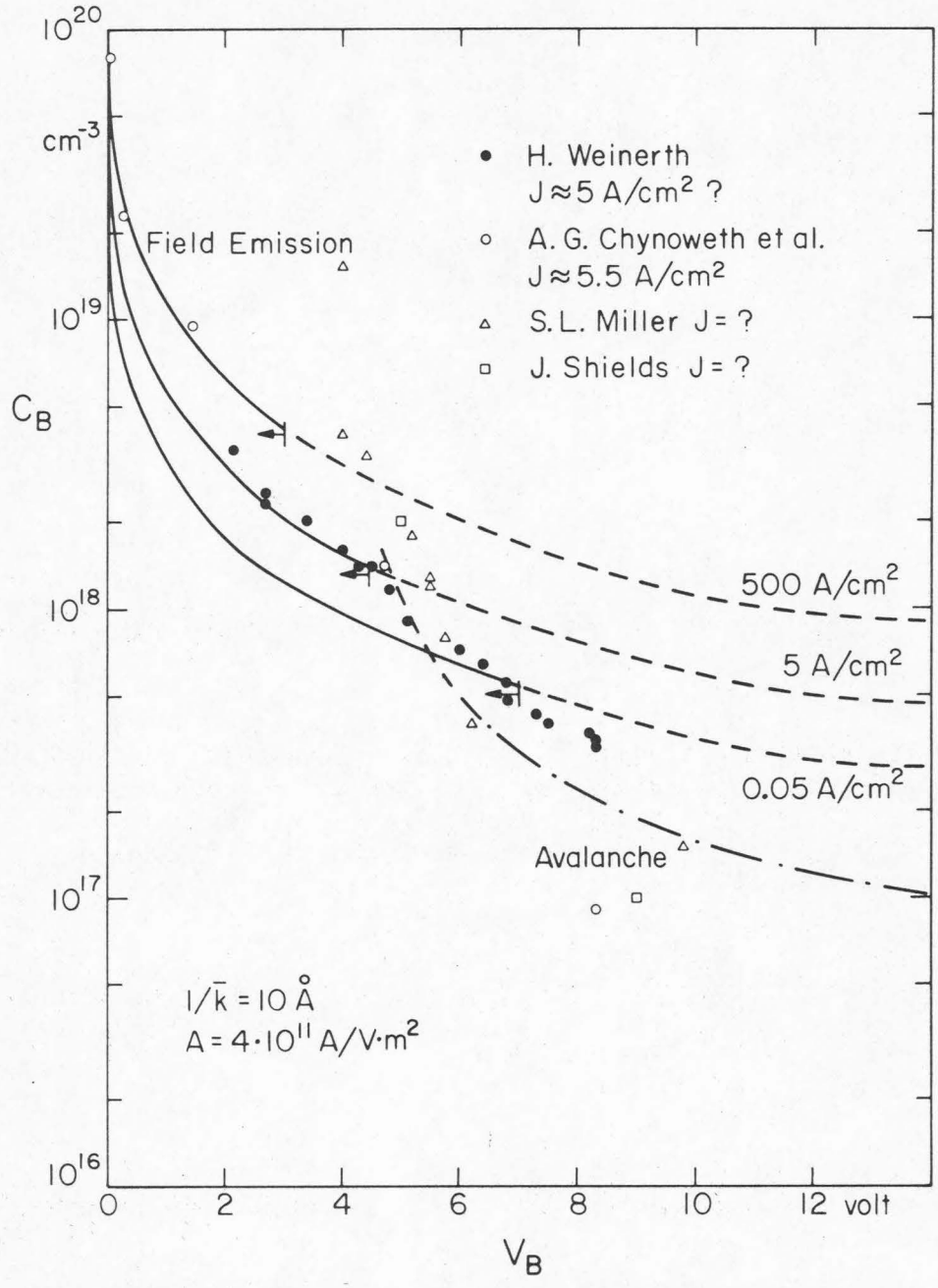

Fig. 2A Reverse "breakdown" voltage V_B of one sided silicon step junction diodes as a function of doping concentration C_B. "Breakdown" is defined to occur when the reverse current density reaches the indicated value. Experimental data by several authors are shown.[9-12] For the data of Weinerth[9] and Chynoweth et al., [10] doping concentration was obtained from the resistivity using a curve by J.C. Irvin.[13] The field emission curves are theoretical (see text). These curves can only be used to the left of the arrows, since at higher voltages avalanche multiplication is important. The experimental avalanche breakdown curve by S.L. Miller[11] is also shown.

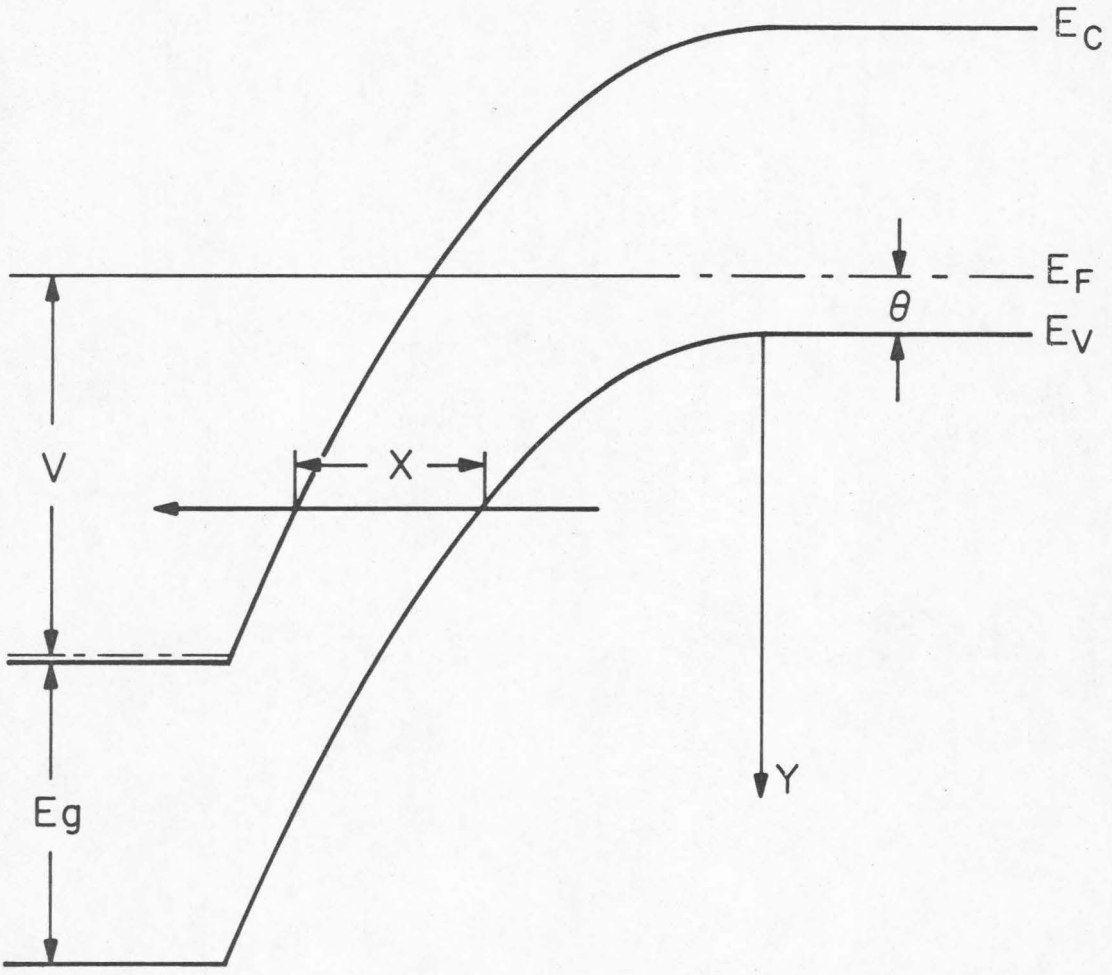

Fig. 3A Energy band diagram of a reverse biased n^+ p diode. The arrow shows the electron tunneling path.

the figure. This field emission current is equal to the product of the number of electrons per unit time attempting to cross the energy barrier, and their probability P of getting across. P is given approximately by the expression:

$$P = e^{-2\bar{k}x} \tag{6A}$$

where $\bar{k}$ is the average wave vector in the "forbidden" energy gap and

$$x = \sqrt{\frac{2\varepsilon}{qC_B}} \left[\sqrt{y + E_g} - \sqrt{y} \right] \tag{7A}$$

is the tunneling distance as shown in Fig. 3A. E_g and y are expressed in eV . q is the electronic charge, ε the permittivity of silicon and C_B the substrate doping concentration. The simplest reasonable approximation is to assume that the number of electrons attempting to cross the energy barrier per unit time is proportional to the energy range dy , so that the field emission current density is approximately

$$J = A \int_0^{V-\theta} e^{-2\bar{k}x} \, dy \tag{8A}$$

The average wave vector $\bar{k}$ was calculated from the tunnel diode data of R. A. Logan et al[14], and from data on the resistance of reverse biased zener diodes taken by H. Weinerth[9]. Both calculations give $\bar{k} \simeq 1/10 \, \text{Å}^{-1}$. The proportionality factor A was chosen to fit the experimental data by H. Weinerth (shown in Fig. 2A) at $V_B = 3$ volts.

The "breakdown" voltage given by Eq. (8A) is plotted in Fig. 2A for several current densities. Also is shown the experimental

avalanche breakdown curve by S. L. Miller[11]. The maximum electric field in the junction at "breakdown" was calculated from the data presented in Fig. 2A, using the standard expressions for one-sided step junctions. The results are plotted in Fig. 4A. The theoretical field emission curve fits the experiment quite well. H. Wienerth[9] calculated the field emission current of intermediate voltage diodes ($3V \lesssim V_B \lesssim 8V$), assuming that the reverse current is given by avalanche multiplication of the field emission current. These results (which are not shown) also fit the theoretical field emission curve quite well.

The "breakdown" voltage is reduced if the junction has curvature. The avalanche breakdown voltage as a function of curvature and substrate doping concentration has been calculated by S. M. Sze and G. Gibbons[15].

APPENDIX 3: CURRENT-VOLTAGE CHARACTERISTICS OF SMALL SIZE MOS TRANSISTORS

As the channel length of an MOS transistor is made smaller, two corrections to the standard theory[16] should be considered. First, the drain and source depletion regions are no longer negligible compared to the channel length. Second, charge carrier velocity saturation becomes important. One-dimensional analysis, including these two corrections, is used to find an upper and lower bound to the drain current.

One-dimensional analysis

The metal-oxide-semiconductor (MOS) transistor is essentially a two-dimensional device. It can, however, be analyzed approximately

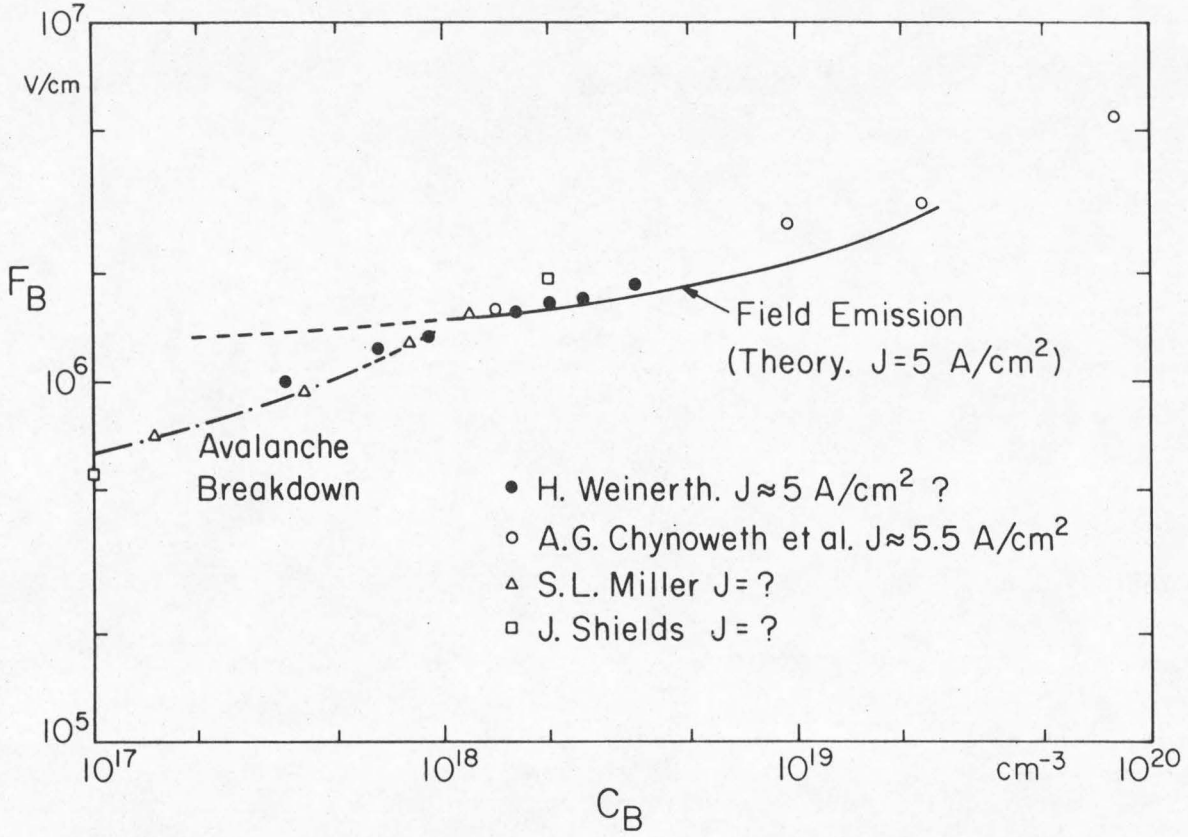

Fig. 4A Maximum electric field F_B in a one sided silicon step junction at "breakdown", as a function of doping concentration C_B. This electric field is calculated from the data of Fig. 2A.

as a one-dimensional device, provided that i) the channel plus substrate charge per unit area is determined exclusively by the gate voltage and not by the source or drain voltages (all voltages are referred to the substrate), and ii) the electric field component along the channel should be small compared to the normal component in the silicon surface. Both conditions are satisfied in the central region of the channel defined by $W_S < x < (L-W_D)$. L is the channel length and W_S and W_D are the source and drain depletion region thicknesses in absence of the gate, as shown in Fig. 5A. Thus the one-dimensional analysis can be applied only to the central region of the channel. The problem is then to determine the boundary conditions, i.e., the channel voltages at $x = W_S$ and $x = (L-W_D)$. This is a difficult problem because two-dimensional analysis cannot be avoided.

We shall consider an n-channel device. The charge per unit area induced in the silicon by the gate voltage is the sum of the channel charge per unit area $-Q$ and the depletion region space charge per unit area $-\sqrt{2\varepsilon q C_B V}$. V is the substrate band bending, C_B the substrate doping concentration, ε the silicon permittivity and q the electronic charge. The electric field in the gate oxide is $(V_G - V_{FB} - V) / x_o$, where V_{FB} is the flat band voltage and x_o the oxide layer thickness. Thus

$$Q + \sqrt{2\varepsilon q C_B V} = C_o(V_G - V_{FB} - V) \qquad (9A)$$

where $C_o \equiv \varepsilon_{ox}/x_o$ is the oxide capacitance per unit area. ε_{ox} is the oxide permittivity.

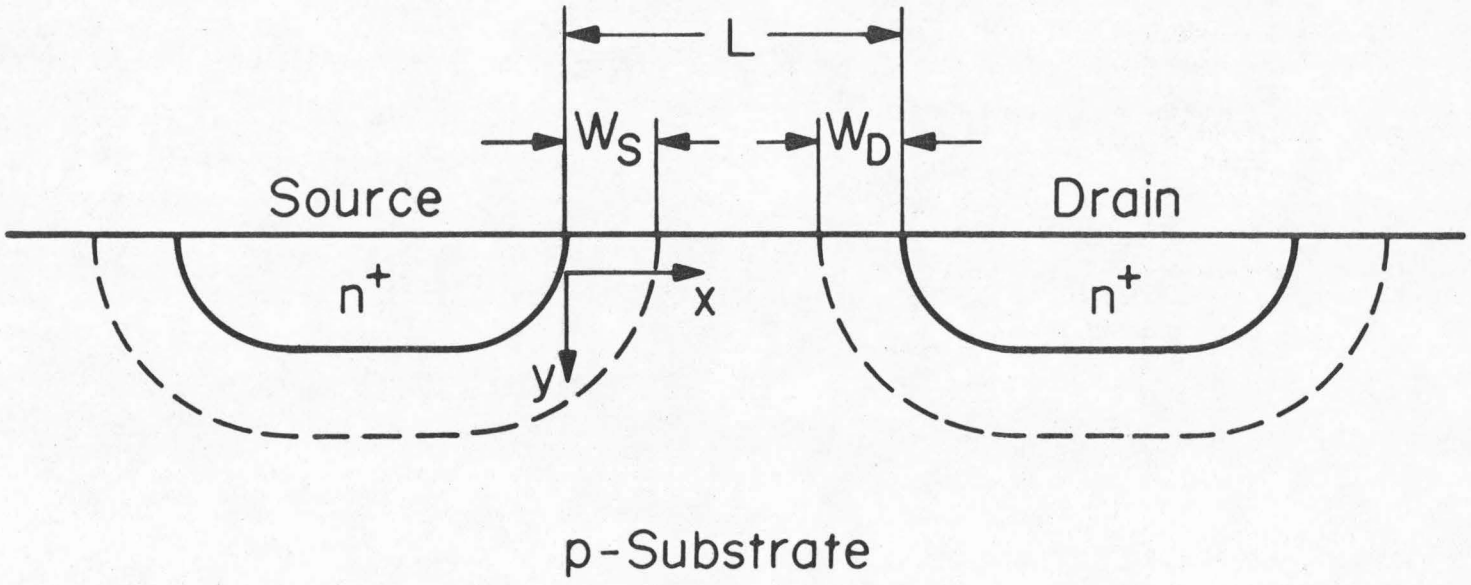

Fig. 5A Cross section of an n-channel MOS transistor. Drain and source depletion regions in absence of the gate extend to the dashed lines. The x,y coordinates used in the analysis are shown.

The drain current I_D can be obtained by solving the differential equation

$$I_D = Z \int_0^w qn(x,y)\ v(x,y)\ dy \equiv Z\,Q\,v_{eff} \tag{10A}$$

where n and v are the charge carrier concentration and velocity, Z the channel width, $Q \equiv \int_0^w q\,n\,dy$ is the channel charge per unit area given by Eq. (9A) and v_{eff} is the effective carrier velocity. We approximate the effective carrier velocity by

$$v_{eff} = v_o \frac{dV/dx}{\frac{dV}{dx} + \frac{v_o}{\mu_{eff}}} \tag{11A}$$

This expression has the correct behavior at the two asymptotes. For low electric fields in the x direction $v_{eff} \simeq \mu_{eff} \frac{dV}{dx}$, where μ_{eff} is the effective mobility. For high electric fields in the x direction $v_{eff} \simeq v_o$, where $v_o = 1 \cdot 10^7 \frac{cm}{sec}$ is the saturation velocity of electrons in silicon.

The solution of Eq. (10A) with the boundary conditions $V = V_1$ at $x = x_1$ and $V = V_2$ at $x = x_2$, is

$$I_D = \frac{ZC_o\mu_{eff}}{[L' + \frac{\mu_{eff}}{v_o}(V_2 - V_1)]} \left\{(V_G - V_{FB})(V_2 - V_1) - \frac{1}{2}[V_2^2 - V_1^2] - \frac{2}{3C_o}\sqrt{2\varepsilon q C_B}\,[V_2^{3/2} - V_1^{3/2}]\right\} \tag{12A}$$

where $L' \equiv x_2 - x_1$. Notice that the effect of charge carrier velocity saturation is to replace the length L' by $[L' + \frac{\mu_{eff}}{v_o}(V_2 - V_1)]$. Equation (12A) is only valid for $V_1 \leq V_2 < V_{2sat}$ where

V_{2sat} is the channel saturation voltage defined by

$$\left.\frac{\partial I_D}{\partial V_2}\right|_{V_2 = V_{2\,sat}} = 0 \qquad (13A)$$

It can be shown that when $V_2 = V_{2sat}$ the charge carriers at $x = x_2$ have the saturation velocity v_o, so that the channel charge per unit area at $x = x_2$ is

$$Q(V_{2sat}) = \frac{I_D}{Zv_o} \qquad (14A)$$

In addition, if $V_2 = V_{2sat}$ the electric field along the channel dV/dx is infinite at $x = x_2$. Since the electric field cannot be infinite we conclude that the channel voltage at $x = (L-W_D)$ is always smaller than V_{2sat}.*

The dependence of the effective mobility on the electric field component normal to the silicon surface was not taken into account when integrating Eq. (10A). This effect can be approximated by an empirical relation[16] between μ_{eff} and the "average" electric field component normal to the silicon surface, defined by

$$E_S = [V_G - V_{FB} - \frac{1}{2}(V_1 + V_2)]\,\frac{C_o}{\varepsilon}$$

Upper bound

At the source the substrate band bending V at onset of strong inversion is $V = V_S + 2\phi$. At the drain it is $V = V_D + 2\phi$. ϕ is

---

*This statement is also confirmed by a first order two-dimensional analysis of the fields near the drain[17].

the difference between the Fermi level and the intrinsic Fermi level in the bulk of the substrate. An upper bound to the drain current can be obtained from Eq. (12A) by setting L' equal to $(L - W_S - W_D)$, $V_1 = V_S + 2\phi$ and $V_2 = V_D + 2\phi$ or $V_2 = V_{2sat}$ whichever is smaller. This is equivalent to applying the one-dimensional analysis from $x_1 = W_S$ to $x_2 = (L - W_D)$, but instead of using the correct (but unknown) boundary conditions V_1 and V_2, we use a lower bound for V_1 and an upper bound for V_2. As a result the drain current obtained by this approximation is an upper bound to the actual current. Standard theory[16] and this upper bound are compared with experiment for a particular MOS transistor in Figs. 6A,a and b.

Lower bound

A lower bound to the drain current is obtained from Eq. (12A) by setting L' equal to the source-drain spacing L, $V_1 = V_S + 2\phi$ and $V_2 = V_D + 2\phi$ or $V_2 = V_{2sat}$ whichever is smaller. This is equivalent to assuming that the one-dimensional analysis is valid from source to drain. This approximation underestimates the channel charge per unit area near the drain and near the source, because even without the gate, the substrate is already depleted near the drain or the source. Thus we conclude that this approximation is a lower bound to the drain current. This lower bound is compared with experiment in Fig. 6A,c.

Summary and conclusions

The gate turn on voltage V_{GT} is obtained from Eq. (9A) by setting $V = V_S + 2\phi$ and $Q = 0$. The result is

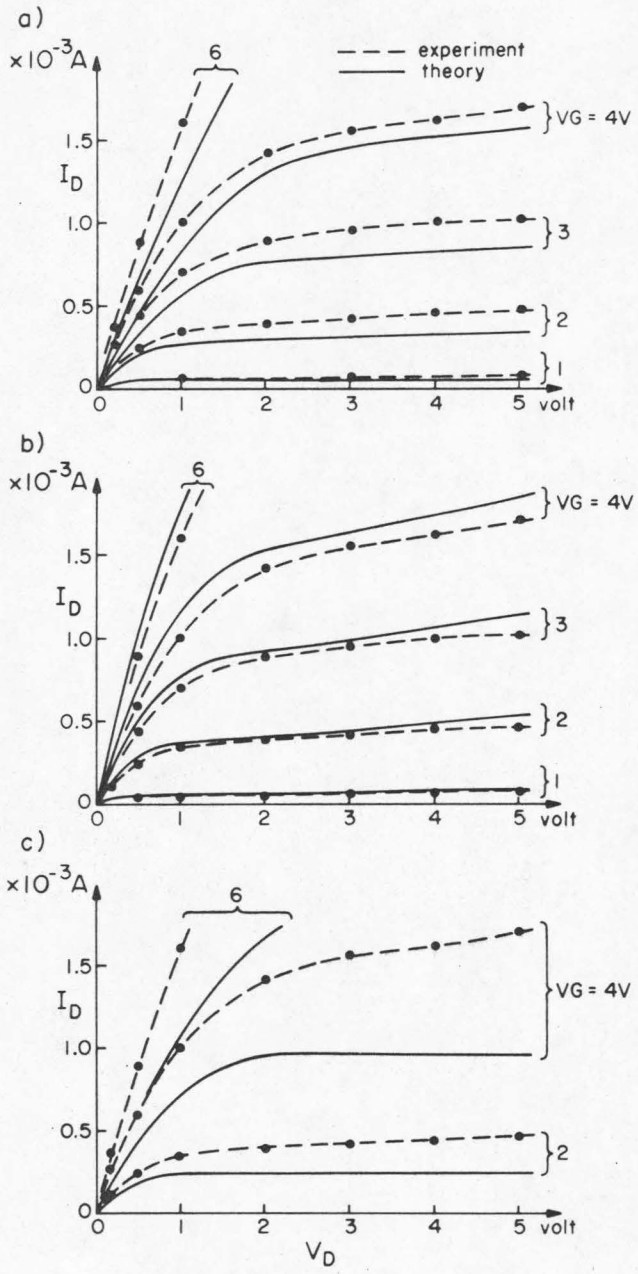

Fig. 6A (a) Standard theory, (b) upper bound and (c) lower bound are compared with the experimental drain characteristics of a particular MOS transistor. The transistor characteristics are L = 3.4μm, Z = 51μm, V_{FB} = -1.0V, x_0 = 1100Å, $C_B = 2.8 \cdot 10^{15} cm^{-3}$ and $\mu_{eff} = 0.0770 - 1.25 \cdot 10^{-9} \cdot E_s$ [m^2/Vsec]. The source voltage is V_s = 0V. All voltages are referred to the substrate.

$$V_{GT} = V_{FB} + V_S + 2\phi + \frac{1}{C_o}\sqrt{2\varepsilon q\, C_B (V_S + 2\phi)} \tag{15A}$$

If $V_G \leq V_{GT}$, $I_D = 0$. If $V_G > V_{GT}$

$$I_D = \frac{ZC_o\,\mu_{eff}}{[L - \alpha(W_S + W_D) + \frac{\mu_{eff}}{v_o}(V_D' - V_S)]} \{(V_G - V_{FB} - 2\phi - \frac{1}{2}V_D')V_D'$$

$$- (V_G - V_{FB} - 2\phi - \frac{1}{2}V_S)V_S - \frac{2}{3C_o}\sqrt{2\varepsilon q C_B}$$

$$\times [(V_D' + 2\phi)^{3/2} - (V_S + 2\phi)^{3/2}]\} \tag{16A}$$

where $V_D' = V_D$ or $V_D' = V_{Dsat}$, whichever is smaller.

The drain saturation voltage V_{Dsat} is defined by

$$\left.\frac{\partial I_D}{\partial V_D}\right|_{V_D = V_{Dsat}} = 0 \tag{17A}$$

(The drain depletion region W_D is kept constant during the differentiation). In Eq. (16A) it is understood that $V_D \geq V_S$ and that neither the source nor the drain junctions are in forward conduction. An upper bound to the drain current is obtained from Eq. (16A) by setting $\alpha = 1$. A lower bound is obtained with $\alpha = 0$. The correct value of the factor (or function) α can only be obtained from two-dimensional analysis or from experiment. The present analysis guarantees that $0 < \alpha < 1$.

Equation (16A) is the same as the standard expression except for two additional terms in the denominator: $-\alpha(W_S + W_D)$ which takes into account the depletion regions and $\frac{\mu_{eff}}{v_o}(V_D' - V_S)$ which is due to velocity saturation of the charge carriers.

The standard expression is quite good even for channel lengths as small as 4 μm because the two corrective terms are of similar magnitude and opposite sign. The upper and lower bounds do not differ by more than a factor of 2, even for the smallest MOS transistor determined by fundamental physical limitations.

PART II

POWER SCHOTTKY DIODE DESIGN AND COMPARISON WITH THE JUNCTION DIODE

List of Symbols

E_g	Semiconductor energy band gap, eV
ε	Absolute permittivity of the semiconductor
ξ	Electric field
ξ_B	Maximum electric field in the semiconductor at the reverse breakdown voltage V_R
ϕ	Schottky energy barrier measured from the Fermi level in the metal, eV
ϕ'	Built in potential in junction diode, eV
$\hbar$	Planck's constant divided by 2π
k	Boltzmann constant
ℓ	Semiconductor thickness
$[Le]_p$	Electron diffusion length in p region
m_o	Free electron rest mass
m	Effective mass of electrons in the metal
m*	Effective mass of conduction electrons in the semiconductor
n_o	Electron concentration in thermal equilibrium
n_i	Electron concentration of intrinsic semiconductor in thermal equilibrium
$[Po]_n$	Hole concentration in equilibrium in the n region
N_d	Donor concentration
N_t	Recombination-generation centers per unit volume
q	Electronic charge
T	Absolute temperature
τ_e	Electron lifetime
μ	Mobility
V_R	Reverse breakdown voltage of a plane diode, i.e., diode with no fringe field effects

v_{th}	Thermal velocity
w	Space charge region thickness
σ	Capture cross-section of generation-recombination center assumed equal for electrons and holes
$\eta = \frac{kT}{q} \log_e [\frac{N_c}{N_d}]$	Energy difference (in eV) between the conduction band edge and Fermi level in the bulk of the semiconductor. N_c is the conduction band effective density of states.

1. Schottky Diode Characteristics

Consider an n-type semiconductor with a rectifying Schottky barrier on one side and an injecting contact on the other, as shown in Fig. 1. We assume that both contacts are blocking for holes so that hole current can be neglected. For low reverse bias the dominant current flow mechanism is thermionic emission due to the electrons which acquire sufficient thermal energy to go over the energy barrier from the metal to the semiconductor as shown in Fig. 2. We refer to this thermionic current density as J_{TE} . In this figure is also shown schematically the energy distribution of electrons which contribute to the current. This distribution is peaked at some energy E_{xo} . For thermionic emission $E_{xo} \simeq E_F + q\phi$. As the reverse bias is increased the barrier becomes thinner and excited electrons with sufficient thermal energy tunnel through the energy barrier from the metal to the semiconductor. This current density (J_{TFE}) is referred to as thermionic field emission. In this case $E_F < E_{xo} < (E_F + q\phi)$. As the reverse bias is further increased E_{xo} decreases until it reaches the Fermi level, i.e., $E_{xo} \simeq E_F$. This is the field emission regime and is also shown in Fig. 2.

Apart from these reverse current mechanisms of the Schottky barrier, there is avalanche multiplication and breakdown in the semiconductor, as in a reverse-biased one-sided step junction diode. In Fig. 3 are plotted the current contributions of a reverse biased Schottky diode. It can be seen that the reverse voltage of a Schottky diode is limited either by avalanche breakdown

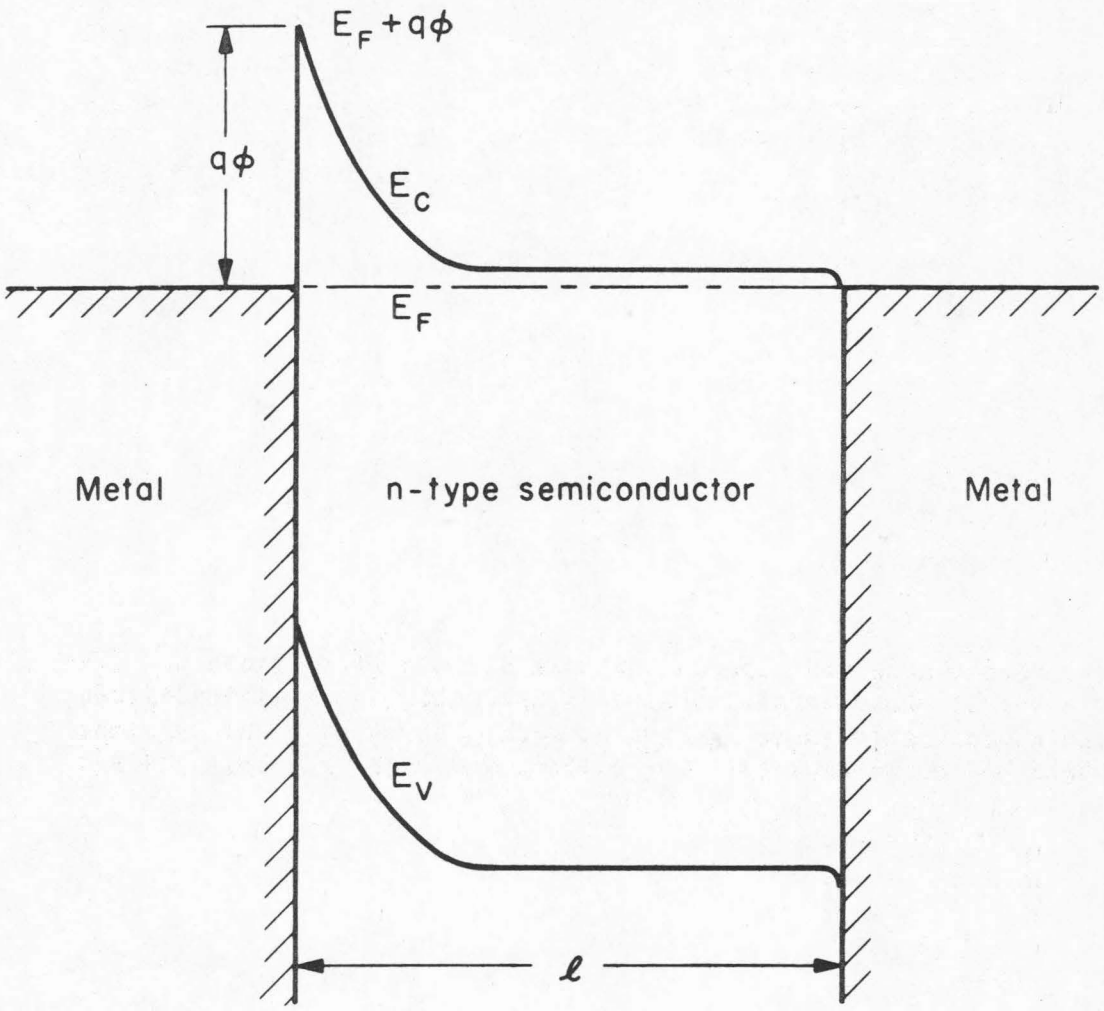

Fig. 1 Energy diagram of a Schottky diode. Electron energy is plotted vertically. The conduction and valence bands of the n-type semiconductor are shown. The rectifying Schottky barrier is on the left and the injecting contact for electrons is on the right.

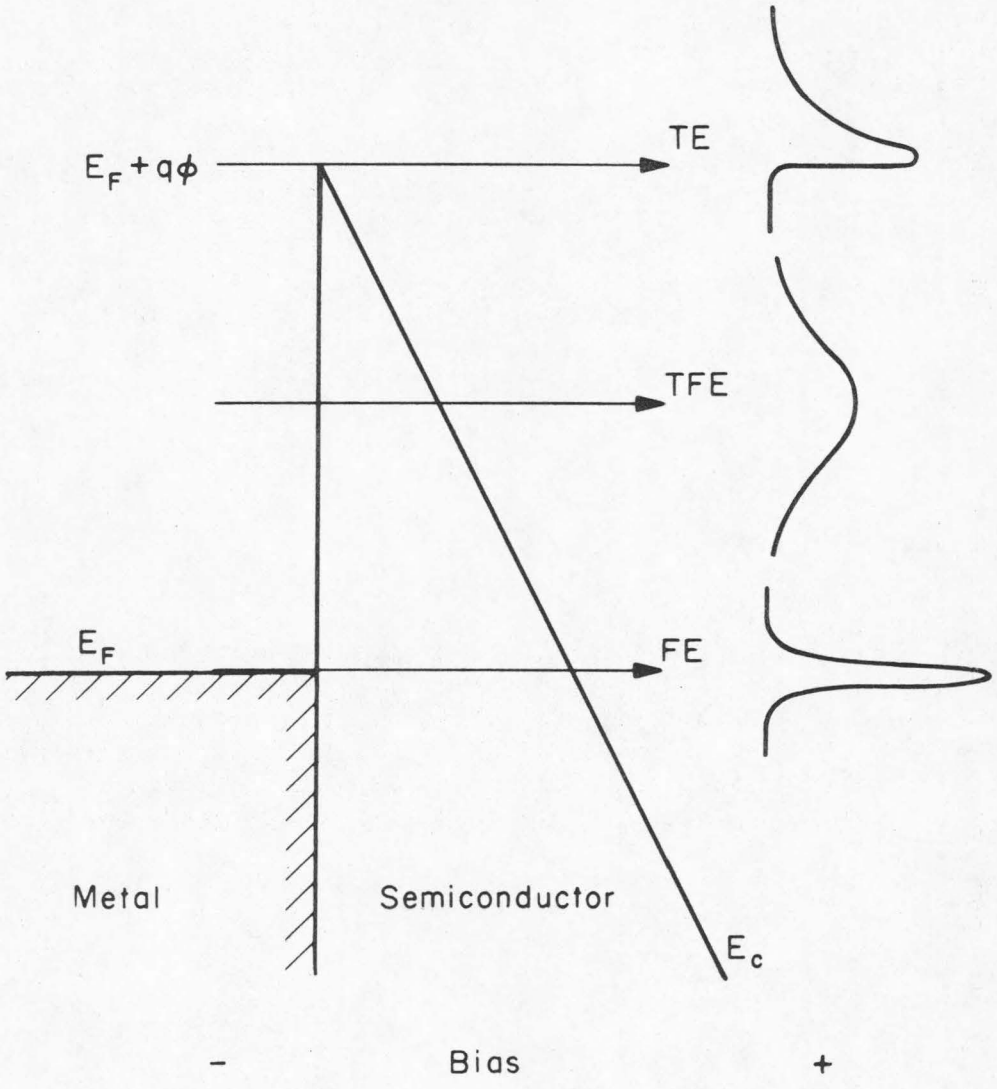

Fig. 2 Expanded view of the metal-semiconductor interface of the diode of Fig. 1 under high reverse bias. As the electric field is increased, the reverse current goes from thermionic emission (TE), to thermionic field emission (TFE), and finally to field emission (FE). The energy distributions of electrons that contribute to the current are shown schematically on the right for the three regimes.

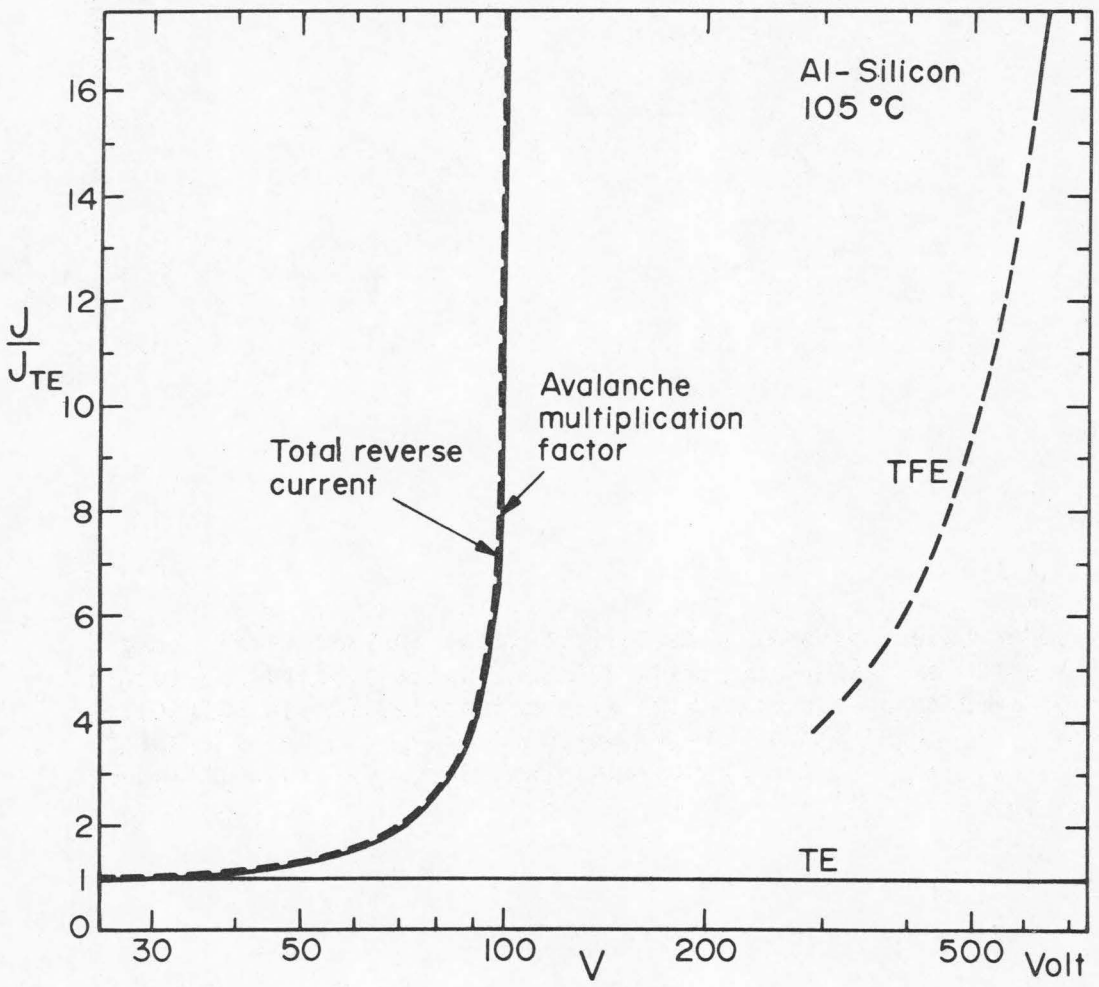

Fig. 3 Three current contributions of a reverse biased Schottky diode: thermionic emission (TE), thermionic field emission (TFE) and avalanche multiplication. In this case avalanche breakdown limits the reverse voltage of the diode. (This is an Al-Si diode. The silicon has a doping concentration $N_d = 1.6 \cdot 10^{15} cm^{-3}$. The TFE curve is calculated. The avalanche curve is only schematic).

or by thermionic field emission current*. It will be shown in connection with Fig. 9 that for silicon Schottky diodes with a reverse breakdown voltage higher than ≃ 5V avalanche is the limiting factor. For lower voltage diodes the limiting factor is thermionic field emission.

The thermionic emission reverse current density is

$$J_{TE} = \frac{qm(kT)^2}{2\pi^2 \hbar^3} \exp[-\frac{q\phi}{kT}] \qquad (1)$$

where ϕ is the Schottky barrier height and m is the effective mass of the electron in the metal. (Image force lowering has been neglected). Equation (1) is plotted in Fig. 4. Thermionic field emission current density is approximately

$$J_{TFE} = \frac{q^2\sqrt{kT}\, Am \quad \xi}{(2\pi)^{3/2}\, \hbar^2\sqrt{m^*}} \exp[-\frac{q\phi}{kT} + \frac{(\hbar q\ \xi)^2}{24 \cdot m^*(kT)^3}] \qquad (2)$$

where ξ is the electric field at the metal-semiconductor interface, m^* is the effective mass of the electron in the semiconductor, and A is of the order of unity. Equations (1) and (2) are derived in Appendix 1. The ratio of thermionic field emission to thermionic emission current J_{TFE}/J_{TE} is plotted in Fig. 5. It can be seen from this figure and from Fig. 3 that the thermionic field emission current resembles a "soft" reverse breakdown. Equation (2) uses the effective mass approximation. A more accurate determination of J_{TFE} and J_{FE} is outlined in Appendix 3.

---

*In exceptional cases pure field emission may be the limiting factor, see Section 6.

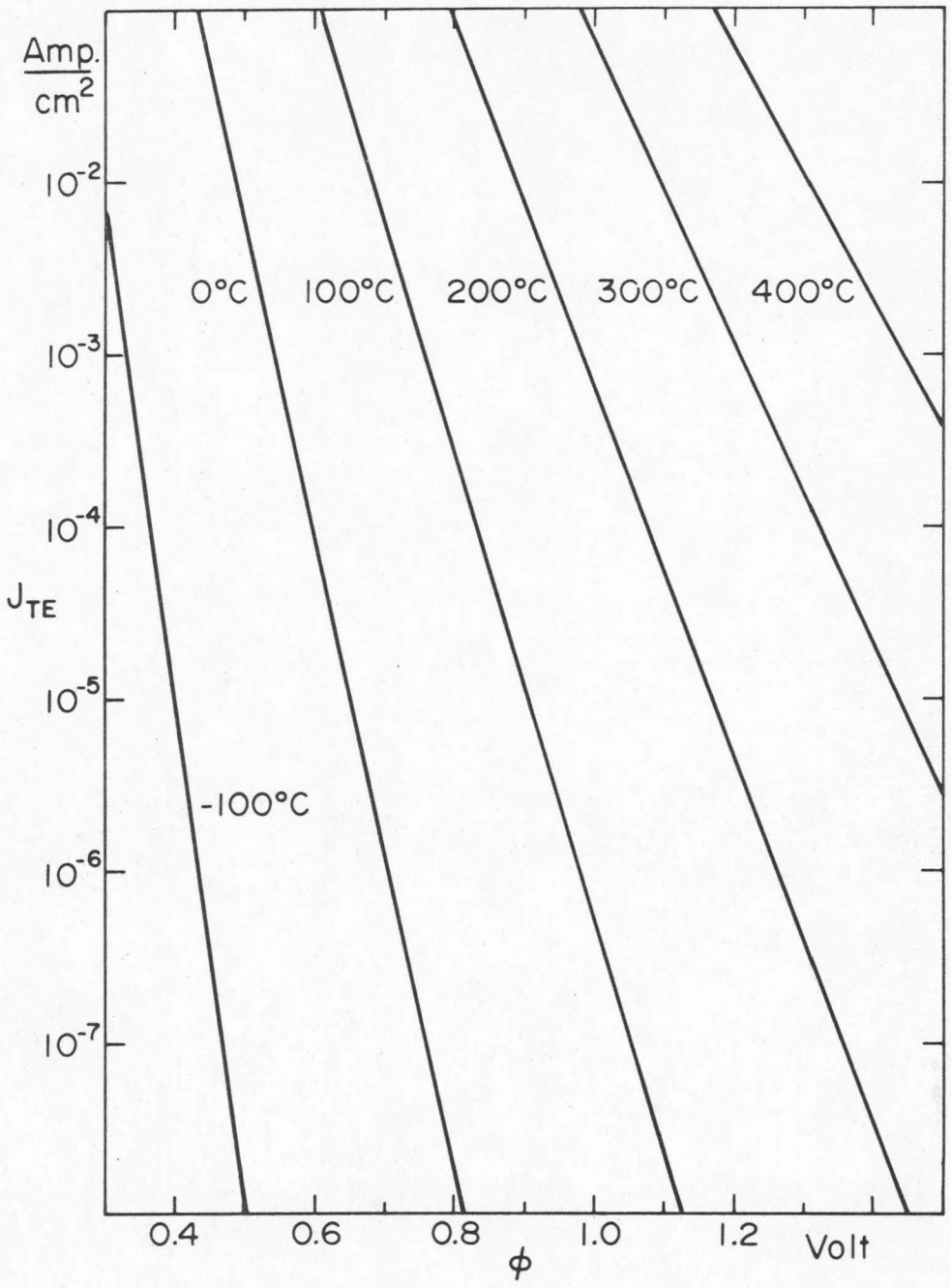

Fig. 4 Reverse thermionic emission current density J_{TE} of a Schottky diode as a function of barrier height ϕ and temperature. The effective mass m of the electrons in the metal has been set equal to the free electron rest mass m_o.

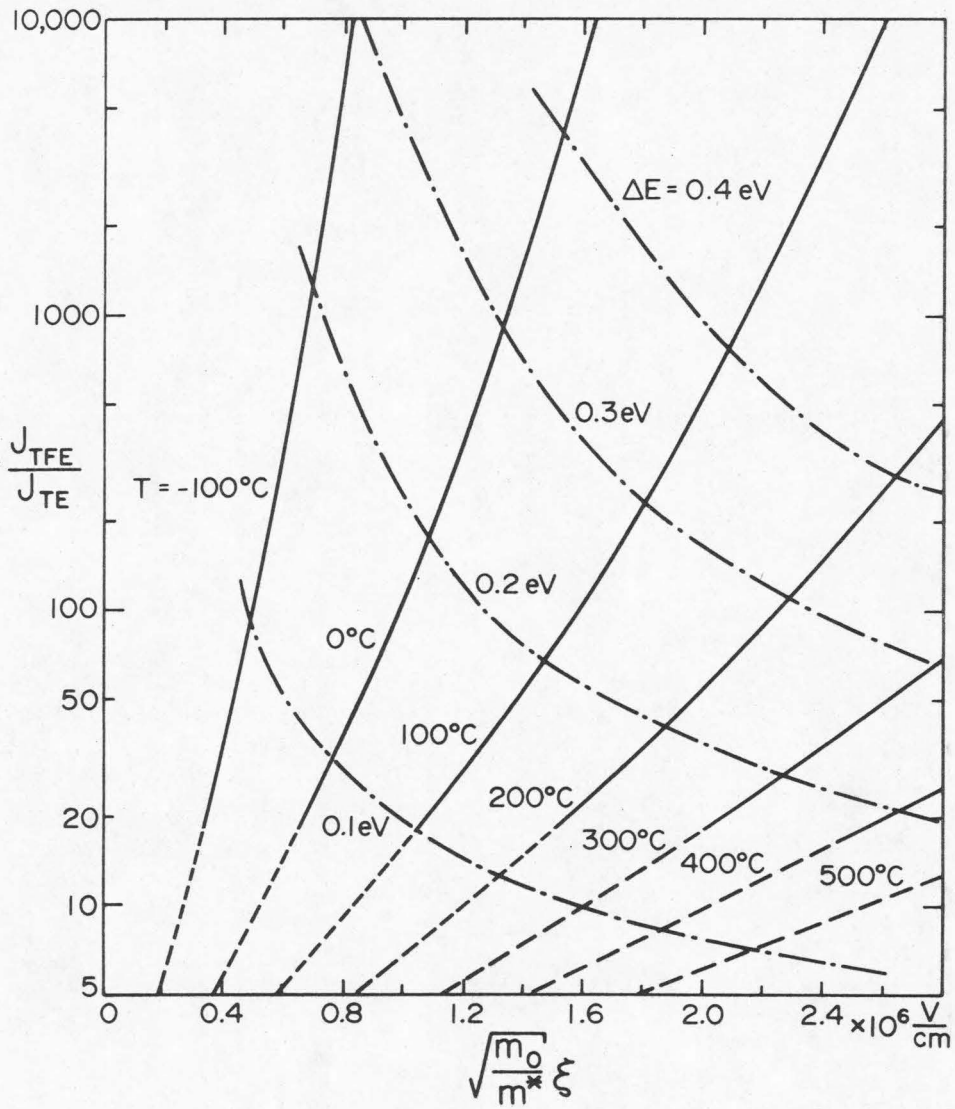

Fig. 5 Solid lines: Ratio of thermionic field emission to thermionic emission current, J_{TFE}/J_{TE}, as a function of temperature and the electric field ξ at the metal-semiconductor interface. For this plot A has been set equal to unity. The dot-dash lines are the contours at which the current distribution is peaked at the indicated energy below the barrier peak $\Delta E = E_F + q\phi - E_{x_0}$. Dashed lines: zone where WKB and integral approximations breakdown (see Appendix 1).

An idealized forward characteristic of the device is sketched in Fig. 6. From A to B we have a thermionic emission regime, i.e., most of the forward voltage drop occurs across the depletion region. From B to C the ohmic drop across the bulk of the semiconductor dominates. This is the ohmic regime. As we increase the forward bias the conduction electron concentration increases above its thermal equilibrium value due to injection from the back contact. At C the excess carrier concentration becomes comparable to the donor concentration. From C to D we have a space charge limited regime with a background of positively charged donors. At D the excess carrier concentration has raised the electron quasi-Fermi-level enough to begin de-ionizing the donors. At E all donors are de-ionized. From E to F we have simple space charge limited current. In all cases we have assumed that there are no charged traps in the semiconductor. In an actual device some of these regimes may not be present, depending on the relative position of the change-over voltages. Also there is considerable overlap of regimes which is not shown in the figure.

For currents below point C and for $V \gtrsim \frac{kT}{q}$ the forward voltage drop across the diode is

$$V = \frac{kT}{q} \log_e \left[\frac{mJ}{m^* J_{TE}}\right] + \frac{\ell}{qN_d\mu} J \qquad (3)$$

where J is the current density, ℓ the thickness of the semiconductor, and μ the electron mobility. The first term is the forward junction (thermionic emission) voltage drop and the second term is the ohmic drop across the semiconductor. It is assumed that the back contact is a good injecting contact so that its contribution to the

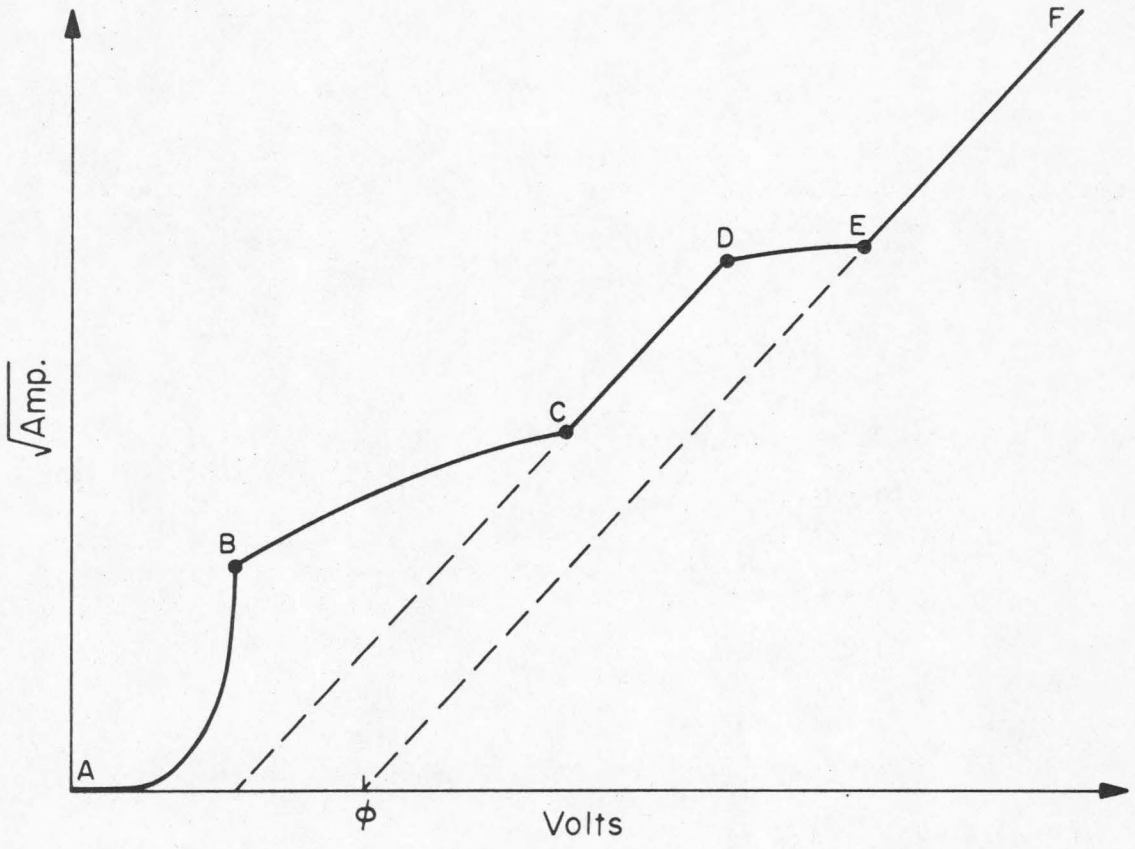

Fig. 6 Idealized forward I-V characteristic of an n-type Schottky diode free of charge traps. AB is determined by thermionic emission, BC by ohmic drop, CD by space charge limited current with a background of positively charged donors, and EF is simple space charge limited current. Overlap of regimes is not shown.

forward voltage drop may be neglected. The transition from the ohmic regime BC of Fig. 6 to the space charge limited regime CD occurs at

$$J_c \simeq \frac{8}{9} \frac{q^2 N_d^2 \mu \ell}{\varepsilon} \tag{4}$$

so Eq. (3) is valid as long as $J \ll J_c$. In normal applications where the semiconductor is not intrinsic, J_c is so large that Eq. (3) is valid for all practical current densities. In the space charge limited regime EF of Fig. 6 ($J \gg J_c$) the current density is related to the forward voltage drop through[1]

$$J = \frac{9}{8} \frac{\mu \varepsilon}{\ell^3} (V - \phi)^2 \tag{5}$$

assuming that no charged traps are present.

2. Schottky Diode Design

The design of a Schottky diode is an iterative process. First a specific metal-semiconductor Schottky barrier is chosen. We check whether an adequate injecting back contact can be made on the semiconductor and whether both contacts are metallurgically reliable. We also check whether the hole current can be neglected with respect to the electron current. For the n-type diode we are now considering this is the case if $Eg - \phi - \eta \gg kT/q$, where η is the energy difference (in eV) between the conduction band edge and the Fermi level in the bulk of the semiconductor. From an estimate of the maximum allowable reverse current density at low reverse voltage J_{TE} we determine the maximum allowable junction temperature T_{max} from Fig. 4. The maximum reverse voltage of the diode must now be specified.

From it the maximum reverse voltage V_R of a plane diode (i.e., a diode with no fringe field effects) can be estimated*. This voltage V_R uniquely determines the doping concentration N_d and the length ℓ of the semiconductor as shown below. In first approximation, breakdown occurs when the electric field at the metal-semiconductor interface reaches some critical value ξ_B. The maximum reverse voltage V_R is related to ξ_B through

$$V_R + \phi = \begin{cases} \xi_B \ell - \dfrac{qN_d}{2\varepsilon}\ell^2 & \ell \leq w = \dfrac{\varepsilon \xi_B}{qN_d} \\ \dfrac{\varepsilon \xi_B^2}{2q N_d} & \ell \geq w \end{cases} \tag{6}$$

Notice that the functional form of Eq. (6) depends on whether or not the thickness ℓ of the semiconductor is greater than the depletion width w at the maximum reverse voltage V_R. We wish to choose ℓ and N_d such that Eq. (6) is satisfied at the desired maximum reverse voltage V_R and that the conductance of the semiconductor per unit area $\sigma = qN_d\mu/\ell$ is maximized. These conditions are satisfied if

$$\ell = \frac{3}{2}\frac{(V_R + \phi)}{\xi_B} = \sqrt{\frac{\varepsilon(V_R + \phi)}{qN_d}} \tag{7}$$

and

$$N_d = \frac{4}{9}\frac{\varepsilon \xi_B^2}{q(V_R + \phi)} \tag{8}$$

Thus, for a given V_R we can calculate ℓ and N_d if ξ_B is known. This optimization is valid if the fringe field lowering of the reverse

---

*For a Schottky diode with guard ring a paper by Sze and Gibbons[2] is useful. Also, see Reference (3).

breakdown voltage is made relatively small. The reverse voltage of the diode is limited by thermionic field emission or by avalanche breakdown, whichever determines the smallest ξ_B. To determine the thermionic field emission limitation we must specify the maximum allowable reverse current J_{TFE} at V_R. From Fig. 5 the maximum electric field ξ_B at the metal-semiconductor interface limited by thermionic field emission may be determined for a given T_{max}. Although avalanche ionization rates can be estimated theoretically for an arbitrary semiconductor, it is more accurate to use experimental results. For some semiconductors, e.g., Ge, Si, GaAs and GaP[4,5,6], the avalanche breakdown voltages of one-sided step junctions have been measured as a function of the doping N_d. Assuming that the reverse voltage is limited by avalanche, we can determine N_d directly from these data using $V_{breakdown} = \frac{9}{8} V_R$. The 9/8 factor arises because the one-sided step junctions used in the experiments have $\ell > w$, whereas the Schottky diode has $\ell < w$ at the maximum reverse voltage. The argument leading to the 9/8 factor is shown in Fig. 7 and is valid if the ionization rates at $\xi = \frac{1}{3} \xi_B$ are negligible. Knowing N_d we can determine from Eq. (8) the maximum electric field ξ_B limited by avalanche breakdown. We use ξ_B determined by thermionic field emission or by avalanche breakdown (whichever is smaller) to calculate ℓ and N_d from Eqs. (7) and (8). The forward characteristics of the diode can now be obtained from Eq. (3). If this forward IV characteristic is not adequate for the specific application other metal-semiconductor Schottky barriers must be considered.

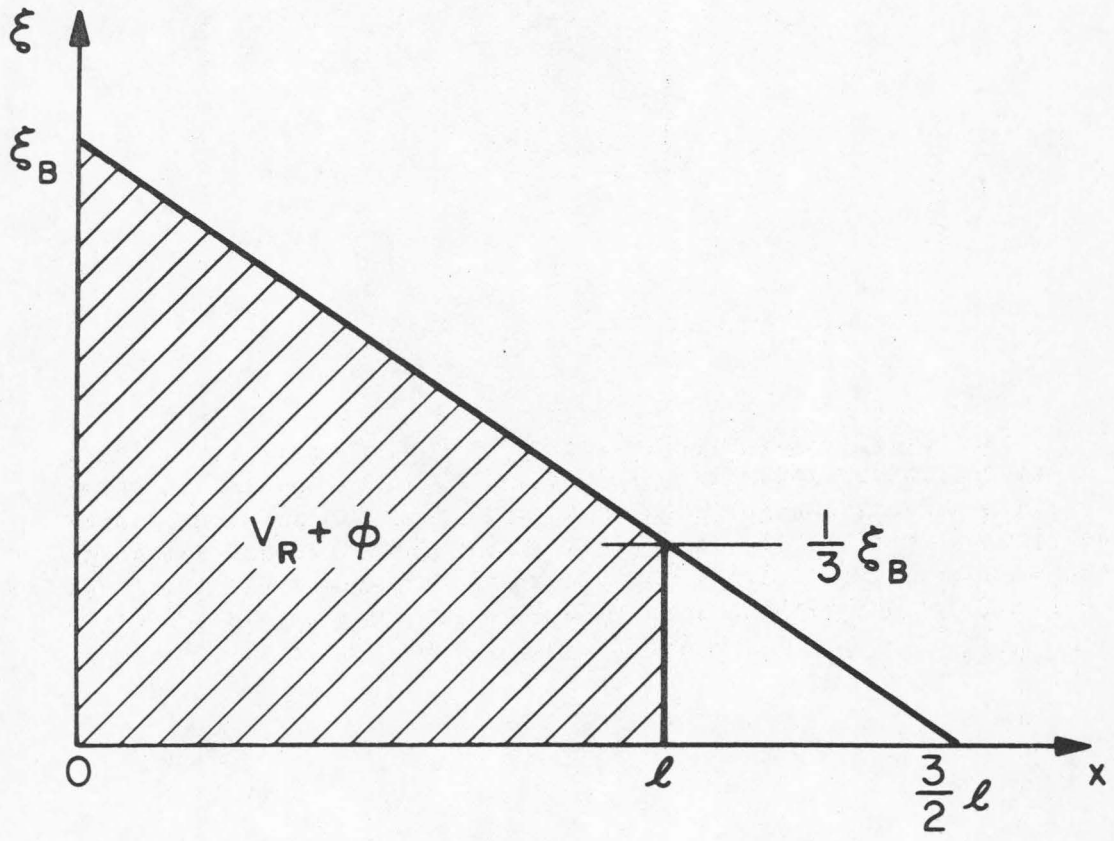

Fig. 7 Electric field ξ in an optimally designed Schottky diode at maximum reverse voltage V_R. The metal semiconductor interface is at $x = 0$ and the back injecting contact is at $x = \ell$. The cross-hatched area represents $V_R + \phi$. If the reverse voltage is limited by avalanche breakdown, the total triangle area is the avalanche breakdown voltage $V_R' + \phi'$ of a one-sided step junction diode (for which $\ell_n \geq \frac{3}{2}\ell$). Notice $(V_R' + \phi') = \frac{9}{8}(V_R + \phi)$.

3. Example

We wish to design an n-type silicon Schottky diode with a reverse breakdown voltage V_R of 100V. (This is the reverse breakdown voltage of a plane diode and can be approached by proper consideration of the fringe fields). First we consider a gold-silicon Schottky barrier which has $\phi \simeq 0.79V$ [7]. We shall set the maximum allowable reverse current density J_{TE} equal to 10 mA/cm^2 since this gives a reasonable upper limit to the power dissipation of the reverse biased diode. From Fig. 4 we find that the gold-silicon junction has $J_{TE} = 10\ mA/cm^2$ at $T = 155^oC$. Gold makes a poor metallurgical contact on silicon. Therefore, we shall consider aluminum which produces a barrier height of $\phi = 0.69V$ on silicon[8]. With this barrier height we obtain a reverse current density of $J_{TE} = 10\ mA/cm^2$ at 105^oC, so we shall limit the maximum temperature to this value. Assuming that the reverse voltage is limited by avalanche breakdown, we determine the doping concentration N_d directly from measured avalanche breakdown voltages in silicon one-sided step junctions. We enter $V = \frac{9}{8} \cdot 100V$ (instead of $V_R = 100V$, as discussed in the previous section) in the avalanche breakdown voltage versus doping concentration plot[4], to find $N_d = 1.6 \cdot 10^{15} cm^{-3}$. From Eq. (8) this gives $\xi_B = 2.4 \cdot 10^5 V/cm$. To determine the thermionic field emission limitation we specify the maximum allowable current, at the reverse voltage V_R, to be $J_{TFE} = 10 \cdot J_{TE}(105^oC)$. Using Fig. 5 with $T_{max} = 105^oC$ and $A \simeq 1$, we get $\sqrt{m_o/m^*}\ \xi_B \simeq 0.86 \cdot 10^6 V/cm$. For conduction electrons in silicon $m^* = 1.1\ m_o$, so $\xi_B \simeq 9 \cdot 10^5 V/cm$. Since the maximum electric field ξ_B determined by avalanche breakdown is

smaller, it is avalanche that limits the reverse voltage of the diode. Thus we set $N_d = 1.6 \cdot 10^{15} cm^{-3}$ and, from Eq. (7), $\ell = 6.4$ μm . Using the last term in Eq. (3) we obtain the resistance of the semiconductor times the diode area equal to $3.4 \cdot 10^{-3} \Omega\ cm^2$. We have used $\mu = 735\ cm^2/Vsec$ for electrons in silicon with a doping $N_d = 1.6 \cdot 10^{15} cm^{-3}$ and a temperature of 105°C. The transition from the ohmic regime to the space charge limited regime occurs at a current density given by Eq. (4): $J_c = 2.7 \cdot 10^4 A/cm^2$. The forward voltage drop given by Eq. (3) is valid for $J < J_c$, i.e., for all practical current densities. From Eq. (3) the forward voltage drop at $J = 100\ A/cm^2$ and 105°C is 0.64V. The finished device structure is shown in Fig. 8. The n^+ layer is to obtain a good ohmic contact to the n region and for mechanical support of the device. The three reverse current components of this diode have been plotted in Fig. 3. We have repeated the design for several values of reverse breakdown voltage; the results are plotted in Fig. 9. For this figure, the reverse voltage limited by thermionic field emission is defined to occur when $J_{TFE}/J_{TE} \simeq 10$ at 105°C.

4. Space Charge Limited Diode

In this section we shall compare the performance of two high voltage Schottky diodes; one with an n-type semiconductor and one with an intrinsic semiconductor. We assume that the forward voltage drop is substantially greater than the Schottky barrier energy so we neglect the voltage drop at the barrier. We also assume that the reverse breakdown voltage V_R occurs when the maximum electric

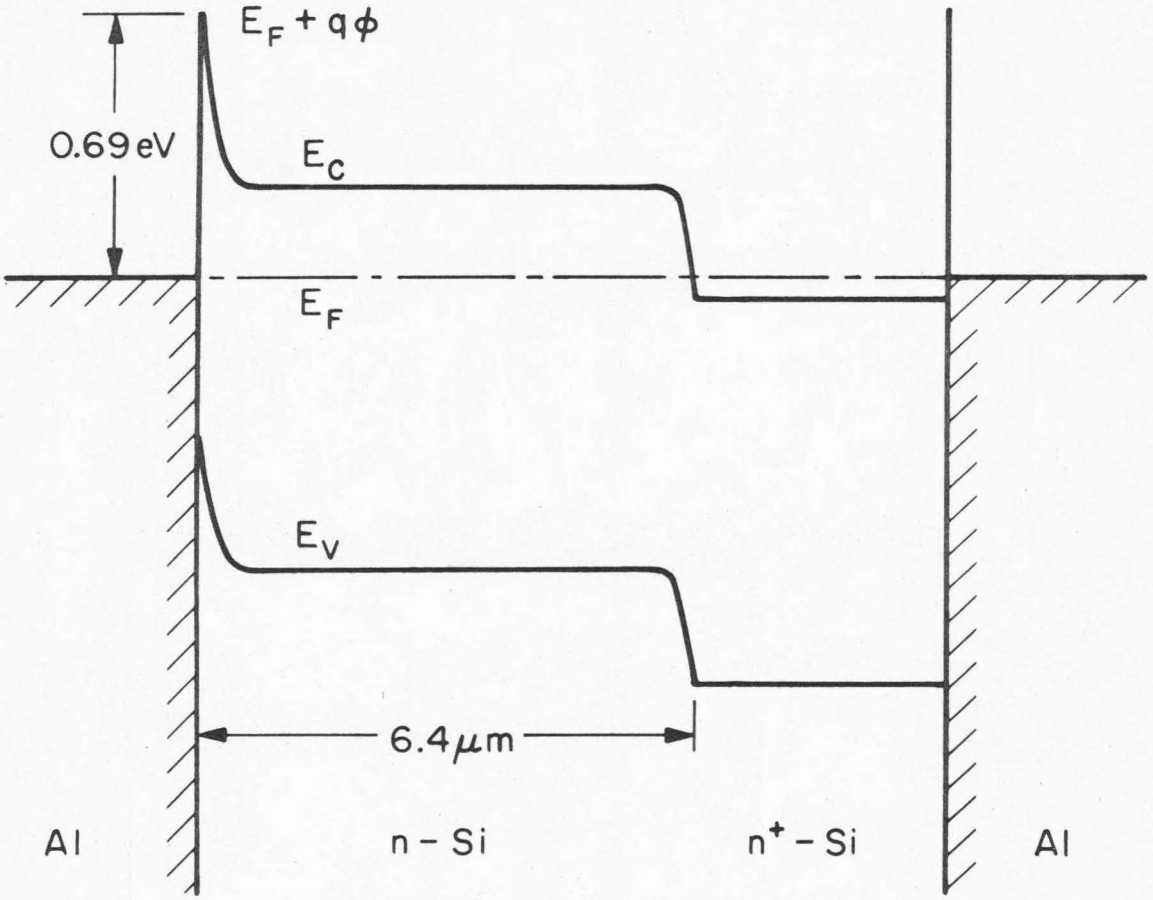

Fig. 8 Energy diagram of an optimally designed 100 V n-type aluminum-silicon Schottky diode.

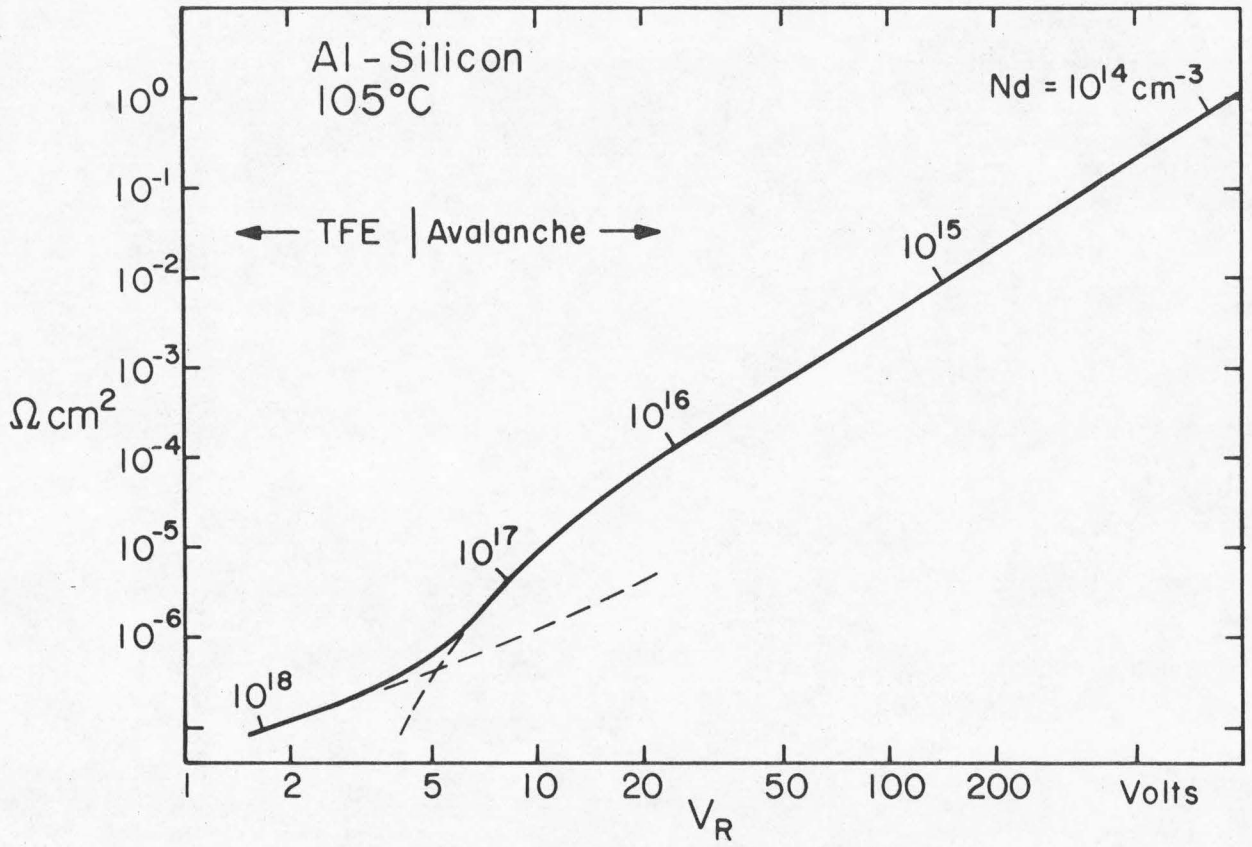

Fig. 9 Semiconductor resistance times diode area as a function of the reverse breakdown voltage V_R of an optimally designed n-type silicon Schottky diode. Notice that the reverse voltage is limited by avalanche breakdown for diodes with $V_R \gtrsim 5$ V and by thermionic field emission for lower voltage diodes. For this plot avalanche data by Shields[4] has been used.

field in the semiconductor reaches the critical value ξ_B . Removal of both these simplifying assumptions favors the n-type diode. The semiconductor of the i-type diode has a thickness $\ell_i = V_R/\xi_B$. The optimum thickness ℓ_n and doping concentration N_d of the semiconductor in the n-type diode are given by Eqs. (7) and (8). The forward voltage drop of the n-type diode is $\frac{\ell n}{qN_d \mu_n} J$ for current densities lower than that given by Eq. (4). At higher current densities the forward voltage drop is approximately $\sqrt{(8J\ell_n^3) / (9\mu_n \varepsilon)}$. The forward voltage drop of the space charge limited i-type diode is $\sqrt{(8J\ell_i^3)/(9\mu_i \varepsilon)}$. At low current densities the forward voltage drop across the n-type diode is smaller. At higher current densities the forward voltage drop across the i-type diode is smaller. We assume that $\mu_n = \mu_i$, which is reasonable since a high voltage Schottky diode has low doping concentration. The forward voltage drop V_F at which both diodes have the same current density can be calculated from the above equations. The result is $V_F \simeq 0.26 \cdot V_R$. In normal applications a diode is used at current densities for which $V_F << 0.26 \cdot V_R$. In this case the performance of an n-type diode is always better than that of a space charge limited diode, since for a given forward voltage drop the current density in the n-type diode is larger than that of the i-type diode. We conclude that for a Schottky diode it is not desirable to use an intrinsic semiconductor. On the contrary, for high voltage junction diodes a large performance advantage can be obtained by introducing a high resistivity layer between the heavily doped p and n regions, thus forming a $p^+ i\, n^+$ structure. Diodes of this type are discussed in the next section.

5. Optimum Design of Junction Diodes and Comparison with the Schottky Diode

Two junction diodes will be considered. First a $p^+ i n^+$ structure in which the high level diffusion length of electrons and holes is one third of the lightly doped i region thickness or larger. In this case electron and hole space charges nearly cancel in forward conduction, thus inhibiting space charge limiting of current. As a result the forward voltage drop across the i region is small. Because the conductivity of the i region increases with increasing forward current, the i region is said to be conductivity modulated. The second junction diode to be considered is a $p^+ n$ one-sided step junction in which conductivity modulation can be neglected, i.e., the hole diffusion length in the n region is much smaller than the n region thickness.

The optimum design and characteristics of Schottky, $p^+ n$ and $p^+ i n^+$ diodes are summarized in Table 1. Some characteristics of the $p^+ i n^+$ diodes are derived in Appendix 2. Notice that the optimum n region thickness ℓ_n and doping N_d of the $p^+ n$ diode are similar to those of the Schottky diode. The values of ℓ_n and N_d have been chosen to maximize the conductance of the n region per unit area for a given reverse breakdown voltage V_R. Notice also that the total forward voltage drop is the sum of the junction voltage drop V_j, plus a resistive voltage drop V_δ, which is the voltage drop across the semiconductor in the Schottky diode, the n region in the $p^+ n$ diode, or the i region in the $p^+ i n^+$ diode.

Table 1. See Refs. (12, 13) and Appendix 2

	Schottky	p^+n	p^+in^+
Reverse Breakdown Voltage V_R	V_R	V_R	V_R
Thickness ℓ	$\ell = \frac{3(V_R+\phi)}{2\xi_B}$	$\ell_n = \frac{3(V_R+\phi')}{2\xi_B}$	$\ell_i = \frac{(V_R+\phi')}{\xi_B}$
Doping Concentration N_d	$\frac{\varepsilon(V_R+\phi)}{q\ell^2}$	$\frac{\varepsilon(V_R+\phi')}{q\ell_n^2}$	$\ll \frac{2\varepsilon(V_R+\phi')}{q\ell_i^2}$
Reverse Current Density J_R	$J_R = AT^2\exp\left[-\frac{q\phi}{kT}\right]$ $A = \frac{qmk^2}{2\pi^2\hbar^3}$	$J_R = J_{gen} + J_{diff}$ $J_{gen} = \frac{qn_i}{2\tau} W$ $J_{diff} = q\left[\frac{P_o L_h}{\tau_h}\right]_n$	$J_R = J_{gen}$ $J_{gen} = \frac{qn_i}{2\tau} W$
Forward Current Density J	$J = \frac{m^*}{m} J_R \exp\left[\frac{qV_j}{kT}\right]$	$J = J_{gen}\exp\left[\frac{qV_j}{2kT}\right] + J_{diff}\exp\left[\frac{qV_j}{kT}\right]$ for $V_j \ll \frac{2kT}{q}\log_e\left[\frac{Nd}{n_i}\right]$	$J = \frac{\sqrt{2}qn_iL_i}{\tau_i}\tan h\left[\frac{\ell_i}{\sqrt{2}L_i}\right]\exp\left[\frac{qV_j}{2kT}\right]$
Resistive Voltage Drop V_δ	$\frac{27\cdot(V_R+\phi)^2 J}{8\mu\varepsilon\xi_B^3}$	$\frac{27\cdot(V_R+\phi')^2 J}{8\mu\varepsilon\xi_B^3}$	$\frac{\ell_i^2}{\mu\tau_i} = \frac{kT}{q}\left[\frac{\ell_i}{L_i}\right]^2$
Total Forward Voltage Drop V	$V_j + V_\delta$	$V_j + V_\delta$	$V_j + V_\delta$
Storage Charge per u. Area q_s	$\leq qN_d\,\ell\exp\left[\frac{q(Eg-\phi-\eta)}{kT}\right]$	$\approx \left[\tau_h\right]_n J$	$\approx \tau_i J$
Capacitive Charge per u. Area q_c	$\varepsilon\xi_B$	$\varepsilon\xi_B$	$\varepsilon\xi_B$
Total Switching Charge per unit Area q	$q_s + q_c$	$q_s + q_c$	$q_s + q_c$

When switching the diode from the forward conduction state to the reverse blocking state, the stored charge due to excess minority carriers must be either swept out or must recombine before the diode becomes blocking. If the switching is done fast enough, essentially the entire stored charge must flow in the reverse direction before the diode becomes blocking. In addition, to raise the reverse voltage from zero to the maximum reverse voltage V_R, a capacitive charge flows in the reverse direction creating the reverse bias depletion region. The stored charge per unit area q_s and the capacitive charge per unit area q_c are also shown in Table 1. The three diodes have the same capacitive charge if they are limited by avalanche breakdown, since then they have the same maximum electric field ξ_B at the breakdown voltage V_R. Once we have the optimum design and characteristics of the Schottky, p^+n and p^+in^+ diodes, we may compare their performance in any given application. To do this we assume that the three diodes are designed to have the same reverse breakdown voltage and that the barrier height of the Schottky diode is chosen so that the reverse current densities of the three diodes are of the same general order of magnitude at the maximum junction temperature. We may then compare directly the forward voltage drop and the stored charge of each device as shown in Table 1. It is interesting to note that for the Schottky and p^+n diodes the resistive voltage drop V_δ is only a function of $V_R^2 J$ and the semiconductor characteristics, provided that $V_R >> \phi'$. Thus, if the reverse breakdown voltage of a diode is doubled, its resistive voltage drop V_δ is the same at 1/4 the forward current density. As an example

we have plotted in Fig. 10 the forward voltage drop and the total switching charge per unit area of three 400V silicon diodes as a function of forward current density. It can be seen from this figure that at high current densities the $p^+ i n^+$ diode has the lowest forward voltage drop. The Schottky and p^+n diodes have the same resistive voltage drop, but the Schottky diode has a smaller junction voltage drop. The $p^+ i n^+$ diode has the highest stored charge, while the Schottky diode has the least.

6. Silicon Carbide Schottky Diode

We wish to design an n-type SiC Schottky diode with $V_R = 600V$ and $T_{max} = 200^oC$. We set the maximum allowable reverse current density (at 600V and 200^oC) equal to $10mA/cm^2$, which gives a reasonable upper limit to the power dissipation of the reverse biased diode. The following characteristics of SiC (15R polytype) at 200^oC are assumed:

$$m^* = 0.28\ m_o$$

$$\mu = 170\ cm^2/V\ sec$$

$$E_g = 2.80\ eV$$

$$\phi = \frac{1}{2} E_g = 1.40\ eV \quad \text{for Au, Ag or Al} \ ^{(9)}$$

$$\varepsilon = 10.2\ \varepsilon_o$$

Using thermionic field emission theory we find $\xi_B = 2.4 \cdot 10^6$ V/cm from Eq. (14A) of Appendix 1 with $J_{TFE} = 10\ mA/cm^2$ at 200^oC. The maximum electric field ξ_B limited by avalanche is approximately $6 \cdot 10^6$ V/cm [10] so avalanche does not limit the reverse voltage of this diode. Using Eq. (11A) of Appendix 1 we obtain

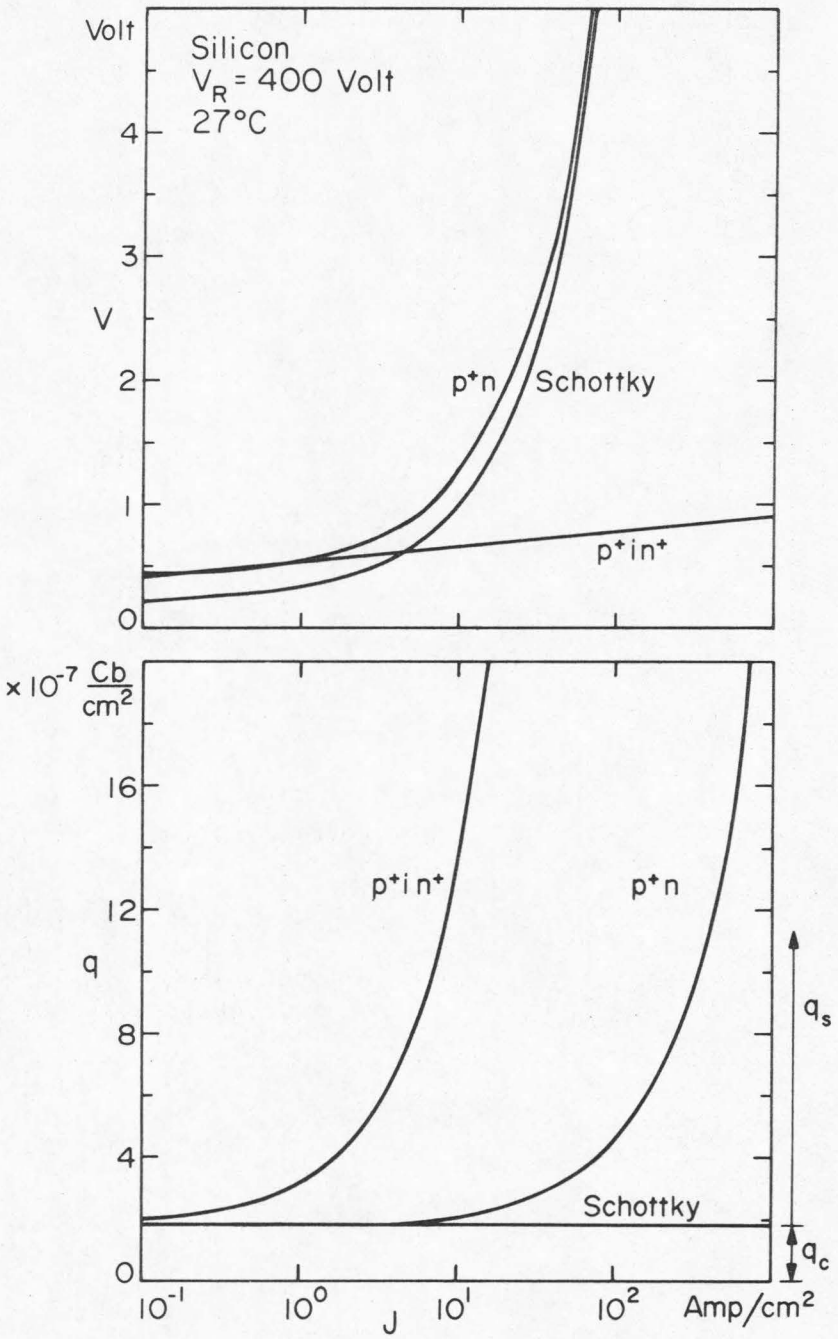

Fig. 10 Forward voltage drop V and total switching charge per unit area q as a function of current density for three 400 V silicon diodes. q is the sum of the capacitive charge q_c and the stored charge q_s. The high level lifetime of the charge carriers in the p^+ n diode is τ = 2.5 n sec. The Aℓ-Si Schottky diode has ϕ = 0.69 V and the p^+ i n^+ diode has $\frac{\ell_i}{L_i} = \sqrt{2}$ which corresponds to $\tau_i \approx$ 120 n sec.

$E_F + q\phi - E_{xo} = 1.18$ eV , indicating that the thermionic field emission theory is not applicable in this case for two reasons: i) tunneling occurs near the Fermi energy in the metal, so that the Boltzmann approximation to the Fermi distribution (see appendix) is not valid, i.e., we are in the transition from thermionic field emission to pure field emission; ii) tunneling occurs near the middle of the energy band gap of the semiconductor where the effective mass approximation is not valid. A more accurate calculation outlined in Appendix 3 gives $\xi_B = 2.09 \cdot 10^6$ V/cm and $E_F + q\phi - E_{xo} = 1.35$ eV . A diode with $V_R = 600$V and $T_{max} = 200^oC$ has $\ell = 4.3$ μm (from Eq. (7)) and $N_d = 1.8 \cdot 10^{16} cm^{-3}$ (from Eq. (8)). The forward voltage drop at $J = 300$ A/cm^2 and at 200^oC is 1.25V. A 600V SiC Schottky diode designed to operate at a maximum temperature of 300^oC would have a forward voltage drop at $J = 200$ A/cm^2 and at 300^oC of 1.3V. This is, however, a theoretical prediction which has yet to be proven experimentally.

7. Conclusion

Because the Schottky diode is a one-carrier device, it has both advantages and disadvantages with respect to the junction diode which is a two-carrier device. The advantage is that there are practically no excess minority carriers which must be swept out before the diode blocks current in the reverse direction, i.e., a much faster recovery time. The disadvantage of the Schottky diode is that for a high voltage device it is not possible to use conductivity modulation as in the p i n diode; since charge carriers are of one sign, no charge cancellation can occur and current becomes space charge limited. The

Schottky diode design is developed in Section 2 and the characteristics of an optimally designed silicon Schottky diode are summarized in Fig. 9. Design criteria and quantitative comparison of junction and Schottky diodes is given in Table 1 and Fig. 10. Although somewhat approximate, the treatment allows a systematic quantitative comparison of the devices for any given application.

The Schottky diode is superior to the junction diode if the latter has no conductivity modulation. If the stored charge is at all important the Schottky diode is also superior to the silicon junction diode with conductivity modulation if at the current densities of interest the Schottky device has a resistive voltage drop small compared to its junction voltage drop. This can be the case if the semiconductor of the Schottky diode has a high enough band gap. The relatively low energy band gap of silicon limits the maximum temperature of a silicon Schottky diode to 100-150^{o}C. Higher band gap materials such as SiC can be used with a higher Schottky barrier energy. Thus the maximum working temperature and/or the maximum electric field ξ_B limited by thermionic field emission or by pure field emission, is increased. Due to the higher band gap the maximum electric field ξ_B limited by avalanche breakdown is also increased. The semiconductor conductance per unit area (which is proportional to ξ_B^3) can therefore be increased. Thus the use of materials such as SiC may result in Schottky diodes with virtually no stored charge and with an I-V characteristic comparable to the best available p i n silicon diode.

REFERENCES - PART II

1. M. A. Lampert and P. Mark, Current Injection in Solids, Academic Press, New York, 1970.

2. S. M. Sze and G. Gibbons, Solid-State Electr. 9, 831 (1966).

3. A. Y. C. Yu and E. H. Snow, J. Appl. Phys. 39, 3008 (1968).

4. J. Shields, J. Electronics and Control 6, 130 (1959).

5. S. L. Miller, Phys. Rev. 105, 1246 (1957).

6. S. M. Sze and G. Gibbons, Appl. Phys. Lett. 8, 111 (1966).

7. D. Kahng, Solid Electr. 6, 281 (1963).

8. A. Y. C. Yu and C. A. Mead, Solid State Electr. 13, 97 (1970).

9. S. H. Hagen, J. Appl. Phys. 39, 1458 (1968).

10. C. Van Opdorp and J. Vrakking, J. Appl. Phys. 40, 2320 (1969).

11. C. R. Crowell, Solid State Electr. 12, 55 (1969).

12. J. Shields, Proc. Institution of Electrical Engineers (London) 106, Part B Supplement, 342 (1959).

13. A. S. Grove, Physics and Technology of Semiconductor Devices, John Wiley and Sons, Inc., New York, 1967.

14. For example, see G. H. Parker and C. A. Mead, Phys. Rev. 184, 780 (1969), and reference therein.

APPENDIX 1: THERMIONIC EMISSION AND THERMIONIC FIELD EMISSION IN A REVERSE BIASED SCHOTTKY DIODE

We shall consider only electron currents as shown in Fig. 2. The results can be modified for hole currents. The number of quantum mechanical states per unit volume, with momentum in the volume element $dp_x dp_y dp_z$ is

$$dN = \frac{dp_x dp_y dp_z}{(2\pi\hbar)^3} q' \tag{1A}$$

where q' is the spin degeneracy factor, which is 2 for electrons. The probability that a state of energy E is occupied by an electron is given by the Fermi probability distribution:

$$F(E) = \frac{1}{1 + \exp[\frac{E-E_F}{kT}]} \simeq \exp[-\frac{E-E_F}{kT}] \tag{2A}$$

for $(E - E_F)/kT > 1$. This equation is valid if we are not too far from thermodynamic equilibrium. The dispersion relation for an electron in the metal is

$$E = \frac{p^2}{2m} = \frac{p_x^2 + p_y^2 + p_z^2}{2m} \tag{3A}$$

and the x-component of its velocity is

$$v_x = \frac{p_x}{m} \tag{4A}$$

where m is the effective mass of the electron in the metal. (Equation (4A) is a relation between expectation values.) We choose the (y,z) plane to coincide with the metal-semiconductor interface. Then,

in the metal, the electrons with positive p_x are incident on the metal-semiconductor interface. The current contribution of electrons with positive p_x in the range from p_x to $p_x + dp_x$ is

$$dJ = \int qv_x F(E)dN = \left[\int_{-\infty}^{\infty} dp_y \int_{-\infty}^{\infty} dp_z \frac{qp_x}{m} \exp\left[-\frac{p_x^2 + p_y^2 + p_z^2}{2m\,kT}\right] \times \exp\left[\frac{E_F}{kT}\right] \frac{2}{(2\pi\hbar)^3}\right] dp_x$$

$$dJ = \frac{qmkT}{2\pi^2\hbar^3} \exp\left[-\frac{E_x - E_F}{kT}\right] dE_x \tag{5A}$$

where

$$E_x \equiv \frac{p_x^2}{2m} \tag{6A}$$

To get the thermionic emission current density we integrate (5A) from $E_x = E_F + q\phi$ to infinity. This gives

$$J_{TE} = \frac{qm(kT)^2}{2\pi^2\hbar^3} \exp\left[-\frac{q\phi}{kT}\right] \tag{7A}$$

To get the thermionic field emission current density we must multiply (5A) by the probability $P(E_x)$ that an incident particle will go through the barrier. This probability is a function of E_x and not of the total energy E. This becomes clear when the three-dimensional Schrödinger equation is reduced by separation of variables into three one-dimensional equations. Using the WKB approximation

$$P(E_x) = A \exp\left[-2 \int k dx\right] \tag{8A}$$

where A is of the order of unity and ik is the propagation vector in the semiconductor. The limits of the integration are the classical

turning points. If the tunneling occurs not too far from the band edge, the effective mass approximation is usable, so

$$k = \frac{\sqrt{2m^*(E_F + q\phi - q\xi x - E_x)}}{\hbar} \tag{9A}$$

We have assumed that the electric field ξ is constant in the vicinity of the junction. This is valid if the reverse voltage is high enough, i.e., $V_R >> \phi$. Image force lowering has also been neglected. Equation (5A) multiplied by (8A) then becomes

$$dJ = \frac{qmkTA}{2\pi^2\hbar^3} \exp\left[-\left\{\frac{E_x - E_F}{kT} + \frac{4\sqrt{2m^*}}{3\hbar q\,\xi}(E_F + q\phi - E_x)^{3/2}\right\}\right] dE_x \tag{10A}$$

The function in the exponent is plotted in Fig. 11. It has a maximum at

$$E_{xo} = E_F + q\phi - \frac{1}{2m^*}\left(\frac{\hbar q\xi}{2kT}\right)^2 \tag{11A}$$

The tunneling current is peaked at $E_x = E_{xo}$. Notice that as the electric field (or the reverse voltage) is increased from zero, the current goes gradually from thermionic emission (where $E_{xo} \simeq E_F + q\phi$) to thermionic field emission, and finally to field emission (where $E_{x_o} \simeq E_F$) . This last transition occurs at an electric field of the order of

$$\xi \simeq \frac{2kT}{\hbar}\sqrt{\frac{2m^*\phi}{q}} \tag{12A}$$

(which is equal to $\simeq 2.7 \cdot 10^6$ V/cm at $T = 300^oK$, $\phi = 1V$ and $m^* = m_o$), provided the effective mass approximation is still valid. To integrate Eq. (10A) approximately we expand the exponent about

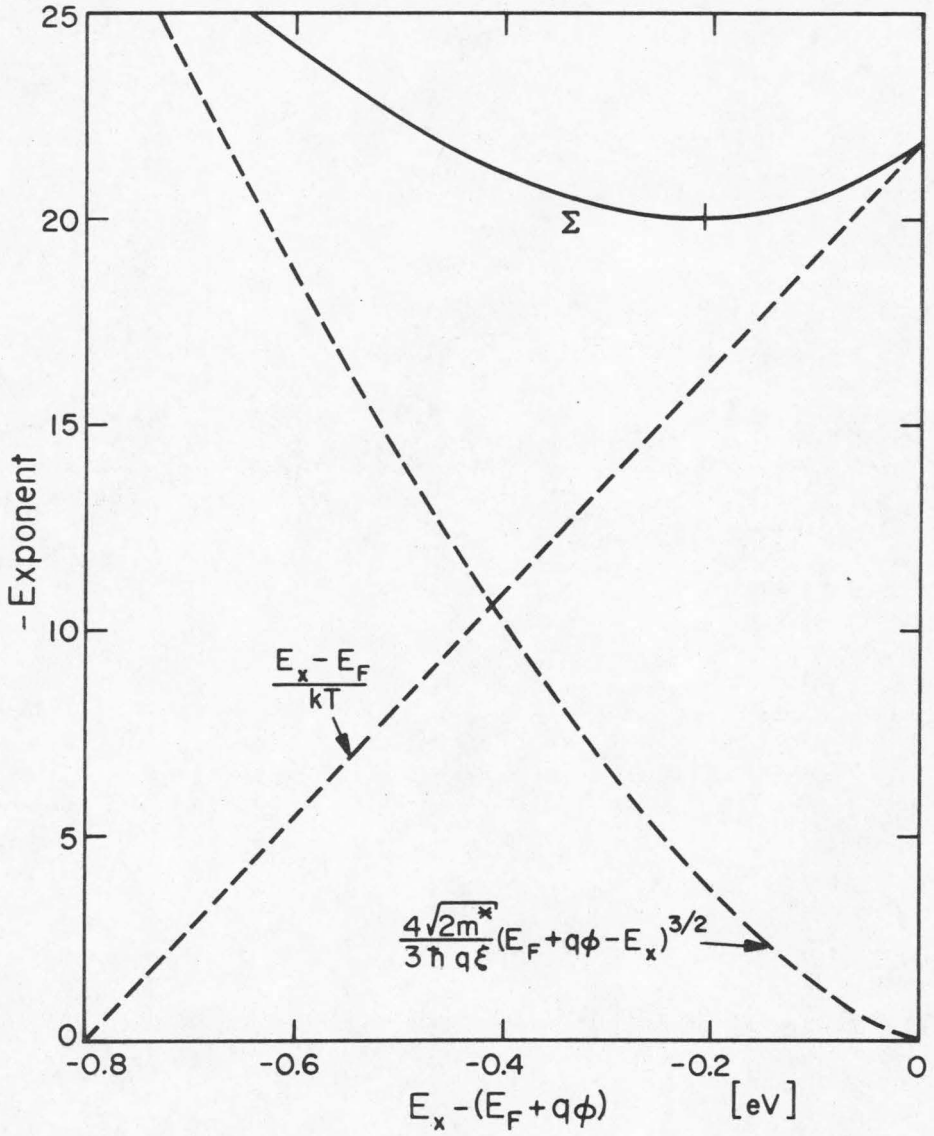

Fig. 11 Exponent in Eq. (10A) of Appendix 1 as a function of electron energy measured from the Schottky barrier peak.

$E_x = E_{x_o}$ keeping only up to the quadratic term, and extend the limits of integration from $-\infty$ to $+\infty$. This gives

$$J_{TFE} = \frac{q^2 \sqrt{kT}\, Am\xi}{(2\pi)^{3/2} \hbar^2 \sqrt{m^*}} \exp\left[-\frac{q\phi}{kT} + \frac{(\hbar q\xi)^2}{24 \cdot m^*(kT)^3}\right] \tag{13A}$$

From Eqs. (7A) and (13A) we get

$$\frac{J_{TFE}}{AJ_{TE}} = 4.32 \cdot 10^{-4} \cdot z \cdot 10^{2.15 \cdot 10^{-9} \cdot z^2} \tag{14A}$$

where $z \equiv \sqrt{m_o/m^*}\, \xi/T^{3/2}$ in MKS units. Equation (14A) is plotted in Fig. 5. The WKB approximation (Eq. (8A)) and the integral approximation are valid for $\int kdx \gtrsim 1$. A detailed solution of the Schrödinger equation for the potential shown in Fig. 2 gives the probability $P(E_x)$ of Eq. (8A) with A of the order of unity. There is some discussion[11] regarding the appropriate mass m in Eq. (7A). In our present approximation it is unimportant whether m is the effective mass of the electron in the metal or in the semiconductor.

APPENDIX 2: OHMIC DROP IN $p^+ i n^+$ DIODES

We consider a $p^+ i n^+$ diode in which the high level diffusion length of electrons and holes is one-third of the lightly doped i region thickness or larger. In this case in forward conduction the electron and hole concentrations are nearly equal and constant throughout the i region[12]. Otherwise large departures from space charge neutrality would occur. In forward conduction $n \simeq p$ in the i region, so the current density due to recombination in this region

is approximately (see Reference (13))

$$r_i = q\sigma\, v_{th}\, N_t \frac{pn}{p+n}\, \ell_i = q\sigma\, v_{th}\, N_t \frac{n}{2}\, \ell_i \tag{15A}$$

if recombination is controlled by recombination centers. The electron (or hole) concentration in the i region is approximately

$$n = p = n_i \exp[\frac{qV_j}{2kT}] \tag{16A}$$

since half of the applied junction voltage V_j is dropped across the n^+i junction and half across the ip^+ junction. The current density due to electrons injected into the p^+ region from the i region is approximately[9]

$$r_{p^+} = q\sigma\, v_{th}\, N_t \frac{p(0)n(0)}{p(0)+n(0)}\, [L_e]_{p^+} = q\sigma\, v_{th}\, N_t\, n(0)[L_e]_{p^+} \tag{17A}$$

because $n << p$ in this region. (We assume low level injection to the p^+ region). A similar expression holds for the current density due to holes injected into the n^+ region from the i region. $[L_e]_{p^+}$ is the electron diffusion length in the p^+ region. $n(0)$ is the electron concentration in the p^+ interface, and satisfies the inequality

$$n(0) \leq [n_o]_{p^+} \exp[\frac{qV_j}{kT}] = \frac{n_i^2}{p} \exp[\frac{qV_j}{kT}] \tag{18A}$$

since V_j is the total applied voltage across the n^+i and ip^+ junctions. The equals sign holds when there is no recombination in the i region. The ratio of current due to recombination in the i region to the current injected into the p^+ region is given by

$$\frac{r_i}{r_{p^+}} \geq \frac{p}{n_i} \frac{\ell_i}{2[L_e]_{p^+}} \exp[\frac{-qV_j}{2kT}] \simeq \frac{\ell_i}{2[L_e]_{p^+}} \exp[\frac{q(E_g - V_j)}{2kT}] \qquad (19A)$$

if the Fermi level in the p^+ region is near the valence band edge. A similar expression holds for the n^+ region. It has been assumed that the recombination center density N_t is the same in the n^+, i and p^+ regions. Since at practical current densities V_j is smaller than the bandgap E_g by several kT/q, we conclude that most of the recombination occurs in the i region, unless ℓ_i is very much smaller than the diffusion lengths L. Thus we shall consider a $p^+ i n^+$ diode in which all the recombination occurs in the i region. In forward conduction the electron and hole concentrations are nearly equal and constant throughout the i region. Otherwise large departures from space charge neutrality would occur. The forward current density is approximately

$$J = \frac{\sigma v_{th} N_t qn \ell_i}{2} = \frac{qn \ell_i}{\tau_i} \qquad (20A)$$

where $q_s = qn\ell_i$ is the total electron (or hole) charge per unit area in the i region and $\tau_i = 2/\sigma v_{th} N_t$ is the high level lifetime of electrons or holes. The voltage drop across the i region is approximately

$$V_\delta = \frac{\ell_i}{qn\mu} J = \frac{\ell_i^2}{\mu \tau_i} = \frac{kT}{q} [\frac{\ell_i}{L_i}]^2 \qquad (21A)$$

where, for our purposes, the high level diffusion length of electrons and holes in the i region is given by $L_i = \sqrt{\frac{kT}{q} \mu \tau_i}$. Here μ is the average of the electron and hole mobilities in the i region. This

expression for V_δ is within 20% of that given by Shields[12]*.

APPENDIX 3: FIELD EMISSION IN A REVERSE BIASED SCHOTTKY DIODE

To calculate the field emission current density we can no longer use the Boltzmann approximation to the Fermi probability distribution. Using the first half of Eq. (2A), Eq. (5A) becomes

$$dJ = \frac{qm\ kT}{2\pi^2\hbar^3}\left[\log_e\left\{1+\exp\left(\frac{E_x - E_F}{kT}\right)\right\} - \left(\frac{E_x - E_F}{kT}\right)\right] dE_x \qquad (22A)$$

dJ is the current density in the metal of the electrons which are incident on the metal-semiconductor interface (i.e., electrons with positive p_x) and which have E_x in the range dE_x at E_x. We must multiply Eq. (22A) by the probability $P(E_x)$ that an incident electron will cross the energy barrier. $P(E_x)$ is given by Eq. (8A):

$$P(E_x) = A\exp\left[-2\int k\,\frac{dx}{dE}\,dE\right] = A\exp\left[-\frac{2}{q\xi}\int_{E_x}^{E_c} k\,dE\right] \qquad (23A)$$

For field emission the energy distribution of electrons that contribute to current is peaked at $E_{xo} \approx E_F$, where E_F is the Fermi energy in the metal. Unless the Schottky energy barrier is a small fraction of the semiconductor band gap, the effective mass approximation is no longer valid. We use a two band model for the E vs. k dispersion relation[14].

$$|k|^2 = \frac{2m^*}{\hbar^2 E_g}\left[\frac{E_g^2}{4} - (E - E_i)^2\right] \qquad (24A)$$

---

*When comparing, note that the high level diffusion length used by Shields corresponds to $L_i / \sqrt{2}$

where $E_i \equiv (E_v + E_c) / 2$. The thermionic field emission or the field emission current density is obtained by multiplying Eqs. (22A) and (23A) and integrating. The integration is straightforward with a digital computer.

PART III

PERMITTIVITY OF STRONTIUM TITANATE

PERMITTIVITY OF STRONTIUM TITANATE

In the present investigation we interpret measurements of permittivity of perovskite strontium titanate performed by Dr. Richard Neville. For completeness the experimental procedures are presented briefly in the next section.

1. Experimental Evidence

Wafers of single crystal perovskite strontium titanate ($SrTiO_3$) were prepared as described in Reference (1). Capacitors formed by evaporating gold uniformly on one side of the wafers and circular gold contacts on the other side, were used to measure the permittivity of strontium titanate as a function of crystal orientation ([001], [011] and [111]), temperature (from 4.2 to 300°K), D.C. electric field (from 0 to ± 23000 V/cm) and frequency (from 1 KHz to 50 MHz). The wafer thickness was always much smaller than the contact diameter so that fringe fields could be neglected. The apparent relative permittivity is defined by

$$\varepsilon_r = \frac{\ell I}{jw\varepsilon_o AV} \qquad (1)$$

where ℓ is the wafer thickness, A is the contact area, I and V are the current and voltage phasors, w is the angular frequency and ε_o is the permittivity of free space. The apparent permittivity below 65°K is found to be dependent on how the samples are cooled. If the samples are cooled with no applied electric field, the apparent permittivity below 65°K is frequency dependent, exhibiting a "resonance" between 1 and 10 MHz as shown in Fig. 1. If the samples are cooled with an applied electric

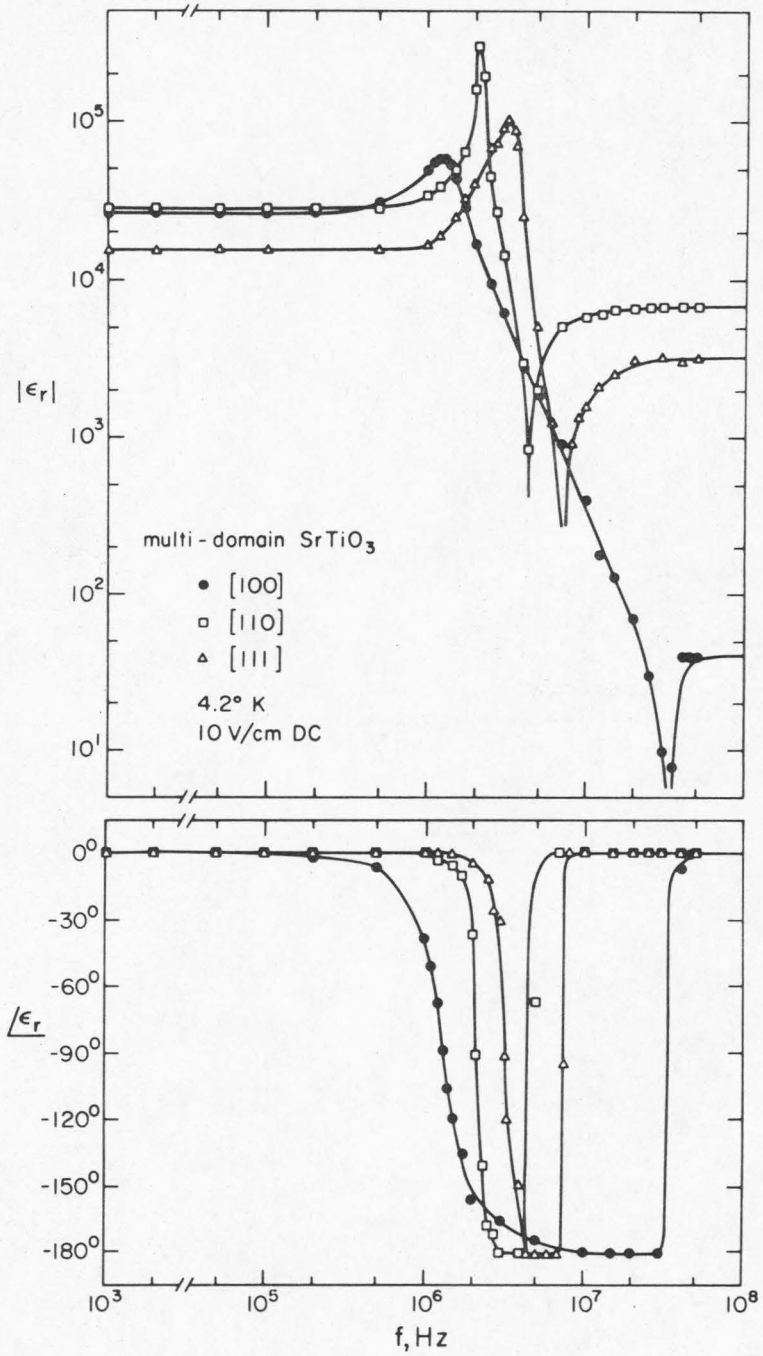

Fig. 1 Magnitude and angle of the apparent relative permittivity of perovskite strontium titanate in the [001], [011] and [111] directions, as a function of frequency. These measurements were made at 4.2°K and with an applied D.C. electric field of 10 V/cm. These crystals were cooled with no applied electric field. The sample orientation has a one-sigma error ± 1° and the permittivity has a one-sigma error of ± 3%. Data by R. Neville.

field in excess of $2 \cdot 10^4$ V/cm, the permittivity is found to be independent of frequency in the measured range from 1 KHz to 50 MHz, and is presented in Fig. 2 as a function of temperature, orientation and D.C. electric field.

We must now make a digression. The perovskite strontium titanate lattice is cubic above 110°K, tetragonal between 110°K and 65°K and orthorhombic below 65°K [2]. The unit cell axis are therefore orthogonal at all temperatures of interest. Above 110°K the three unit cell axes are equal. As the crystal is cooled below 110°K one of the three unit cell axes elongates slightly to form the tetragonal c axis. The two remaining unit cell axes differentiate when the crystal is cooled below 65°K, thus forming the orthorhombic lattice which has three unit cell axes of differing lengths. The distortion from a cubic lattice is small. For the orthorhombic lattice $\underline{a}:\underline{b}:\underline{c} = 0.9998:1:1.0002$ [2].

The frequency independent permittivity indicated in Fig. 2 is also obtained if the electric field (in excess of $2 \cdot 10^4$ V/cm) is applied to the crystal only while cooling through the 110°K and 65°K phase transition temperatures (i.e. from 113 to 107°K and from 68 to 63°K). Application of the electric field at other temperatures has no effect on the low temperature permittivity. If the electric field is applied only while cooling through the 65°K phase transition, the complete anomalous resonance is observed. If the electric field is applied only past the 110°K phase transition, about 10% of the anomalous resonance is observed. A single sample can exhibit successively the frequency dependent and frequency independent behavior, depending on how the crystal is cooled.

We interpret the foregoing observations as follows: When the

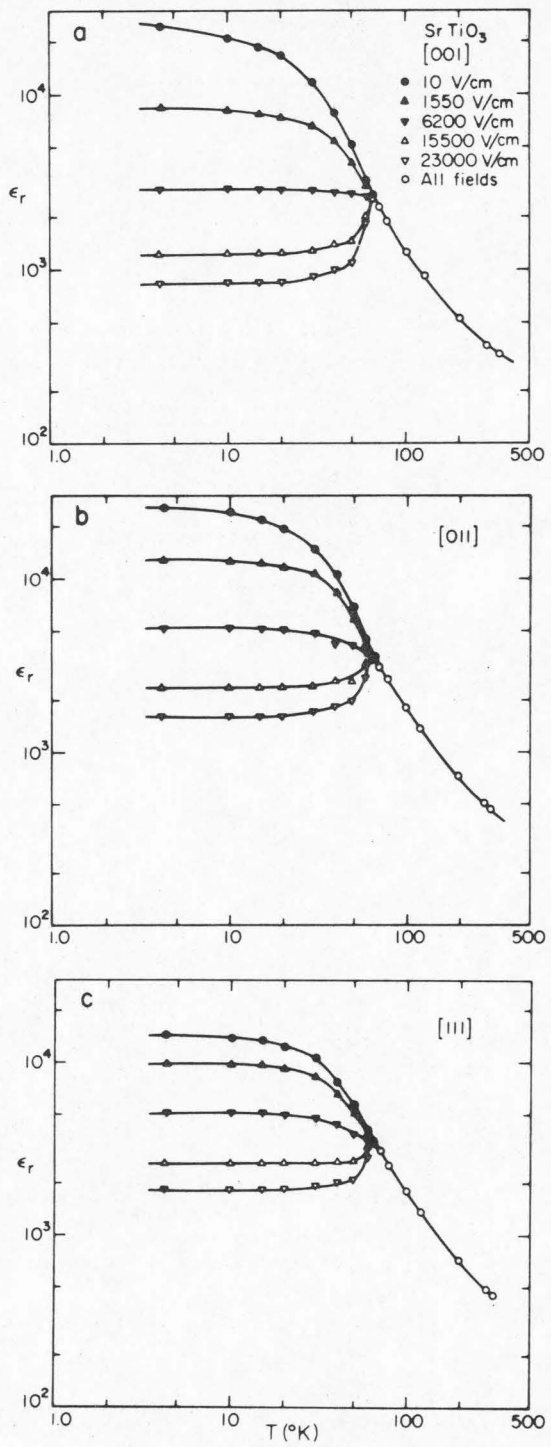

Fig. 2 Relative permittivity of perovskite strontium titanate as a function of temperature, D.C. electric field and orientation. These crystals were cooled <u>with</u> an applied electric field of 23000 V/cm. The permittivity is independent of frequency in the measured range from 1 KHz to 50 MHz. Data by R. Neville.

cubic crystal is cooled past 110°K with an applied electric field, one of the unit cell axis elongates slightly to form a single domain tetragonal crystal. While cooling the single domain crystal past 65°K with an applied electric field, the remaining two unit cell axis differentiate to form a single domain orthorhombic crystal. If the sample is cooled past 110°K with no applied electric field, tetragonal domains form with different c-axis orientation (as shown by Lytle[2] and Chang et al.[3]). While cooling the crystal past 65°K with no applied electric field, subdomains form with different a-axis orientation. We shall refer to the crystals cooled with (without) an applied electric field as single-domained (multi-domained).

In the measured directions ([001], [011] and [111]*), temperature range (from 4.2 to 300°K) and D.C. electric field range (from 0 to ± 23000 V/cm), the multi-domain samples have, within ±5%, the same low frequency ($< 3 \cdot 10^5$ Hz) permittivity as the single domain crystals.

The curves presented in Fig. 1 were found, within the experimental error, to be independent of the contact diameter, wafer thickness and wafer area in the measured ranges from 0.1 to 1.2 cm, 0.1 to 1 mm and 1 to 10 cm^2 respectively. In particular, the frequency of the anomalous resonance is not determined by the dimensions of the wafer or the contact. The apparent permittivity was also found to be independent of the amplitude of the A.C. electric field applied to the crystal, in the measured

---

*All crystal directions cited in this investigation correspond to the cubic lattice above 110°K.

range from 0.3 to 15 V/cm peak.* The phenomenon that causes the anomalous resonance is therefore linear, at least within the measurement range considered above. The frequency of the peak of the "resonance" does not change (within ±5%) with temperature as the crystal is heated above 4.2°K. The "resonance" gradually disappears as the temperature is increased and the apparent permittivity becomes frequency independent at 55 ± 10°K.

No thermal or electrical hysteresis or charge conduction was observed within the measurement ranges and measurement errors. Thus no evidence of a paraelectric to ferroelectric transition was found.

2. Single-Domain Strontium Titanate.

The high permittivity of strontium titanate and its dependence on temperature and frequency can be understood in terms of the electrically active optical phonons. From the data presented in Fig. 2, and without a detailed knowledge of the ion displacements, it is possible to (i) determine the crystal free energy vs. polarization curves at various temperatures, (ii) apply the Lyddane-Sachs-Teller relation[4] and (iii) verify the Curie-Weiss law[5]. A more detailed knowledge of the crystal structure and chemical bonds is required to determine (iv) the ion displacements in the electrically active optical phonon modes. Each of these topics will now be considered in turn.

---

*Except that the peak of the resonance is 10% lower with an applied A.C. electric field of 15 V/cm than with 0.3 V/cm peak.

2.1 Free Energy vs. Polarization

At constant temperature the increment of free energy per unit volume, F, of a dielectric is given by

$$dF = E\underline{e} \cdot d\underline{P} \tag{2}$$

where $\underline{e}$ is the unit vector in the direction of the macroscopic electric field E and $\underline{P}$ is the polarization density. The permittivity of the dielectric in the direction of the macroscopic electric field will be denoted by ε . Then by definition

$$\frac{\partial(\underline{e}\cdot\underline{P})}{\partial E} = \varepsilon-\varepsilon_o \tag{3}$$

where ε_o is the permittivity of free space. From equations (2) and (3)

$$F = \int_o^E (\varepsilon-\varepsilon_o)E\, dE \tag{4}$$

and

$$P \equiv \underline{e} \cdot \underline{P} = \int_o^E (\varepsilon-\varepsilon_o)dE \ . \tag{5}$$

The electric field dependence of the permittivity (see Fig. 2) can be expressed, within 1%, by $1/\varepsilon_r = A(T) + B(T)\ |E|$. The observed values of A(T) and B(T) are listed in Table 1. The free energy per unit volume F was calculated as a function of the polarization density P in the [001], [011] and [111] directions, using the data in Table 1. The results are presented in Fig. 3. At temperatures at and above 65^oK

Table 1 Relative permittivity of single domain strontium titanate in the [001], [011] and [111] directions as a function of temperature (from 4.2 to 300°K) and electric field (from -23000 to +23000 V/cm). The relative permittivity is given by $\epsilon_r = 1/(A+B\cdot|E|)$. B is expressed in m/V. The permittivity is independent of frequency in the measured range from 1 KHz to 50 MHz. The permittivity has a one-sigma error of 2% and the crystal orientation has a one-sigma error of 1°.

T	[001]		[011]		[111]	
°K	A	B	A	B	A	B
4.2	4.097×10^{-5}	4.907×10^{-10}	3.882×10^{-5}	2.466×10^{-10}	7.072×10^{-5}	2.005×10^{-10}
10	4.782×10^{-5}	4.877×10^{-10}	4.142×10^{-5}	2.455×10^{-10}	7.279×10^{-5}	1.996×10^{-10}
15	5.446×10^{-5}	4.848×10^{-10}	4.546×10^{-5}	2.414×10^{-10}	7.477×10^{-5}	1.971×10^{-10}
20	6.175×10^{-5}	4.817×10^{-10}	5.239×10^{-5}	2.391×10^{-10}	8.032×10^{-5}	1.946×10^{-10}
30	8.430×10^{-5}	4.438×10^{-10}	6.735×10^{-5}	2.220×10^{-10}	9.672×10^{-5}	1.809×10^{-10}
40	1.264×10^{-4}	3.777×10^{-10}	9.785×10^{-5}	1.881×10^{-10}	1.266×10^{-4}	1.530×10^{-10}
50	1.958×10^{-4}	3.156×10^{-10}	1.446×10^{-4}	1.578×10^{-10}	1.769×10^{-4}	1.282×10^{-10}
60	3.184×10^{-4}	9.852×10^{-11}	2.290×10^{-4}	5.278×10^{-11}	2.475×10^{-4}	3.879×10^{-11}
65	3.937×10^{-4}	0.0	2.833×10^{-4}	0.0	2.890×10^{-4}	0.0
70	4.484×10^{-4}	0.0	3.247×10^{-4}	0.0	3.300×10^{-4}	0.0
77	5.319×10^{-4}	0.0	3.788×10^{-4}	0.0	3.876×10^{-4}	0.0
100	7.813×10^{-4}	0.0	5.650×10^{-4}	0.0	5.780×10^{-4}	0.0
120	1.099×10^{-3}	0.0	7.299×10^{-4}	0.0	7.463×10^{-4}	0.0
200	1.923×10^{-3}	0.0	1.370×10^{-3}	0.0	1.399×10^{-3}	0.0
280	2.801×10^{-3}	0.0	2.020×10^{-3}	0.0	2.058×10^{-3}	0.0
300	3.030×10^{-3}	0.0	2.183×10^{-3}	0.0	2.232×10^{-3}	0.0

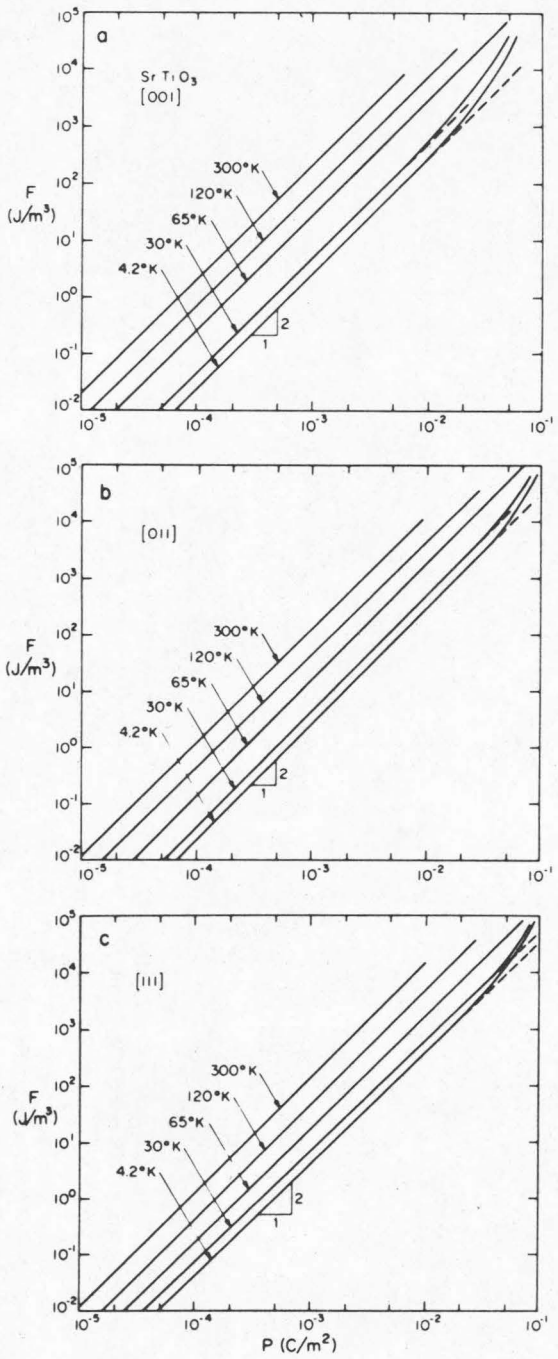

Fig. 3 Strontium titanate crystal free energy as a function of polarization, calculated for several temperatures and crystal orientations. The origins of free energy and polarization have been chosen so that F = 0 and P = 0 when the electric field E = 0.

the free energy is quadratic in the polarization. Below 65°K the free energy for large polarization departs from a P^2 dependence and increases more rapidly.

2.2 Lyddane - Sachs - Teller Relation

The LST relation [4] is given by:

$$\frac{\varepsilon(o)}{\varepsilon(\infty)} = \frac{w_L^2}{w_T^2} \tag{6}$$

where w_T is the frequency of a long wavelength transverse optic phonon, w_L is the frequency of the corresponding long wavelength longitudinal optic phonon, and $\varepsilon(o)$ and $\varepsilon(\infty)$ are permittivities at frequencies below and above w_T respectively. Equation (6) can be applied to all the optic phonon frequencies, yielding the generalized LST relation [6]:

$$\frac{\varepsilon(o)}{\varepsilon(\infty)} = \prod_i \frac{w_{Li}^2}{w_{Ti}^2} \tag{7}$$

Here $\varepsilon(o)$ and $\varepsilon(\infty)$ are the permittivities at low and at optic frequencies respectively. It is expected theoretically[5] and confirmed experimentally[7] that the only optical phonon frequency strongly temperature dependent is the transverse optic mode of lowest frequency, also called the "soft" phonon mode. Since the permittivity at optic frequencies is not very sensitive to temperature, we expect from Eq. (7) that the product $\varepsilon(o)\, w_{T\,Soft}^2$ is independent of temperature. Using the measured permittivities in the [001] direction and the "soft" phonon

frequencies measured by Cowley[7], Worlok and Fleury[8], Nilsen and Skinner[9], and Barker and Tinkham[10], it is found that

$$\varepsilon_r(o) w_{T\ Soft}^2 = 9.65 \pm 0.3 \cdot 10^{28} \text{sec}^{-2} \qquad (8)$$

independent of temperature and electric field in the measured ranges from 30^oK to 300^oK and 0 to 12000 V/cm*. $w_{T\ Soft}$ is expressed in radians per second. The [001] direction was chosen because Cowley measured the frequency of phonons propagating in this direction.

Using Eq. (6) it is possible to calculate the permittivity at frequencies above the "soft" phonon frequency. The lowest four transverse and longitudinal optic phonon frequencies are listed in Table 2. The longitudinal modes corresponding to each transverse mode were determined by requiring that $w_{Li} \geq w_{Ti}$[5]. Note that the modes 2 and 3 have $w_L = w_T$, which implies that for these modes the polarization due to ion displacements is zero[5]. This implication is confirmed by the displacements calculated by Cowley[7] for these two modes. The relative permittivity calculated from Eqs. (6) and (8) and the data presented in Table 2 is $\varepsilon_r = 12.1$ between the "soft" phonon frequency ($2.7 \cdot 10^{12}$ Hz at 296^oK) and $16.4 \cdot 10^{12}$ Hz, and $\varepsilon_r = 5.35$ above $16.4 \cdot 10^{12}$ Hz, in good agreement with the permittivity of strontium titanate at optical frequencies. Thus we have confirmed the generalized LST relation. The permittivity of single domain perovskite strontium titanate as a function of frequency is presented in Fig. 4.

*Non-systematic variations of $\varepsilon_r(o)\, w_{T\ Soft}^2$ as a function of electric field at 8^oK might be due to differences between our samples and those of Worlok and Fleury.

Table 2 Frequency (in units of 10^{12}Hz) of long wavelength transverse optical phonons and frequency of the corresponding long wavelength longitudinal optical phonons. Data from Cowley and others[7].

Mode	Transverse	Longitudinal
1	2.73 at 296°K	14.2
2	5.10	5.10
3	7.95	7.95
4	16.4	24.7

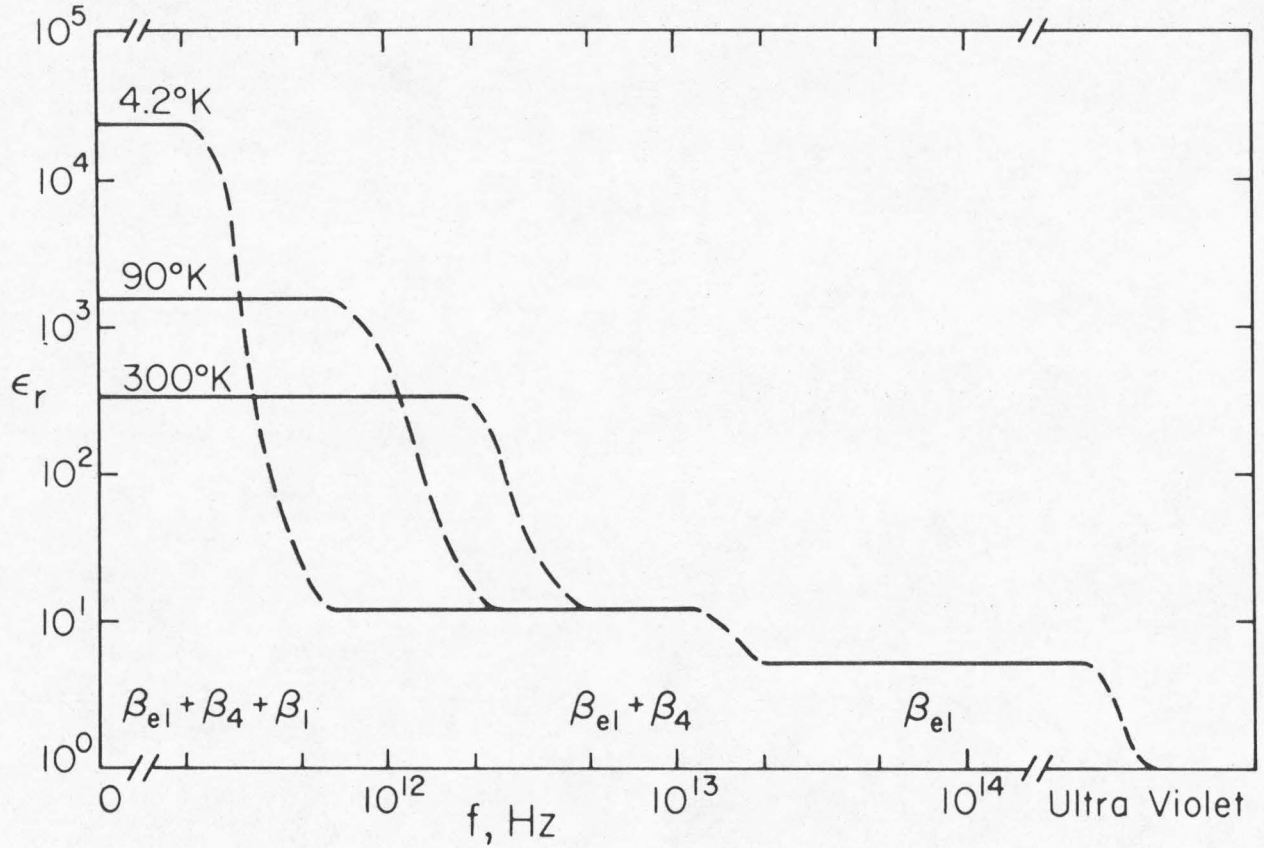

Fig. 4 Permittivity of strontium titanate in the [001] direction as a function of frequency.

2.3 Curie - Weiss Law

It can be seen in Fig. 5 that the permittivity of strontium titanate with essentially no applied D.C. electric field follows the Curie-Weiss relation[5] $1/\varepsilon_r \propto (T-T_c)$ from 40 to 300°K, with a Curie temperature of $T_c = 30 \pm 2°K$, which is the same for the three crystal orientations. Departures from the Curie-Weiss law occur at temperatures below 40°K with no applied D.C. electric field, and below 65°K with an applied D.C. electric field as shown in Fig. 2.

2.4 Ion Displacements

If a static electric field is applied to strontium titanate, the strontium and titanium ions are slightly displaced with respect to the oxygen ions. The displacement of each of the 5 ions in the unit cell are, in general, different and cannot be obtained from permittivity measurements alone[7]. There is some debate whether the "soft" transverse phonon is mainly due to motion of the strontium ions or the titanium ions. Cowley[7] favors titanium ion displacements while Bell and Rupprecht[11], Last[12], and Spitzer et al.[13] favor strontium ion displacements.

Let us consider the crystal structure of perovskite strontium titanate. The unit cell is a cube with a Ti^{4+} ion at the body center, surrounded by O^{2-} ions at the face centers and Sr^{2+} ions at the cube corners. The charge indicated for each ion is correct only if the crystal bonds are purely ionic. The charges are smaller if the bonds are partially covalent. The atomic weights and approximate ionic radii are

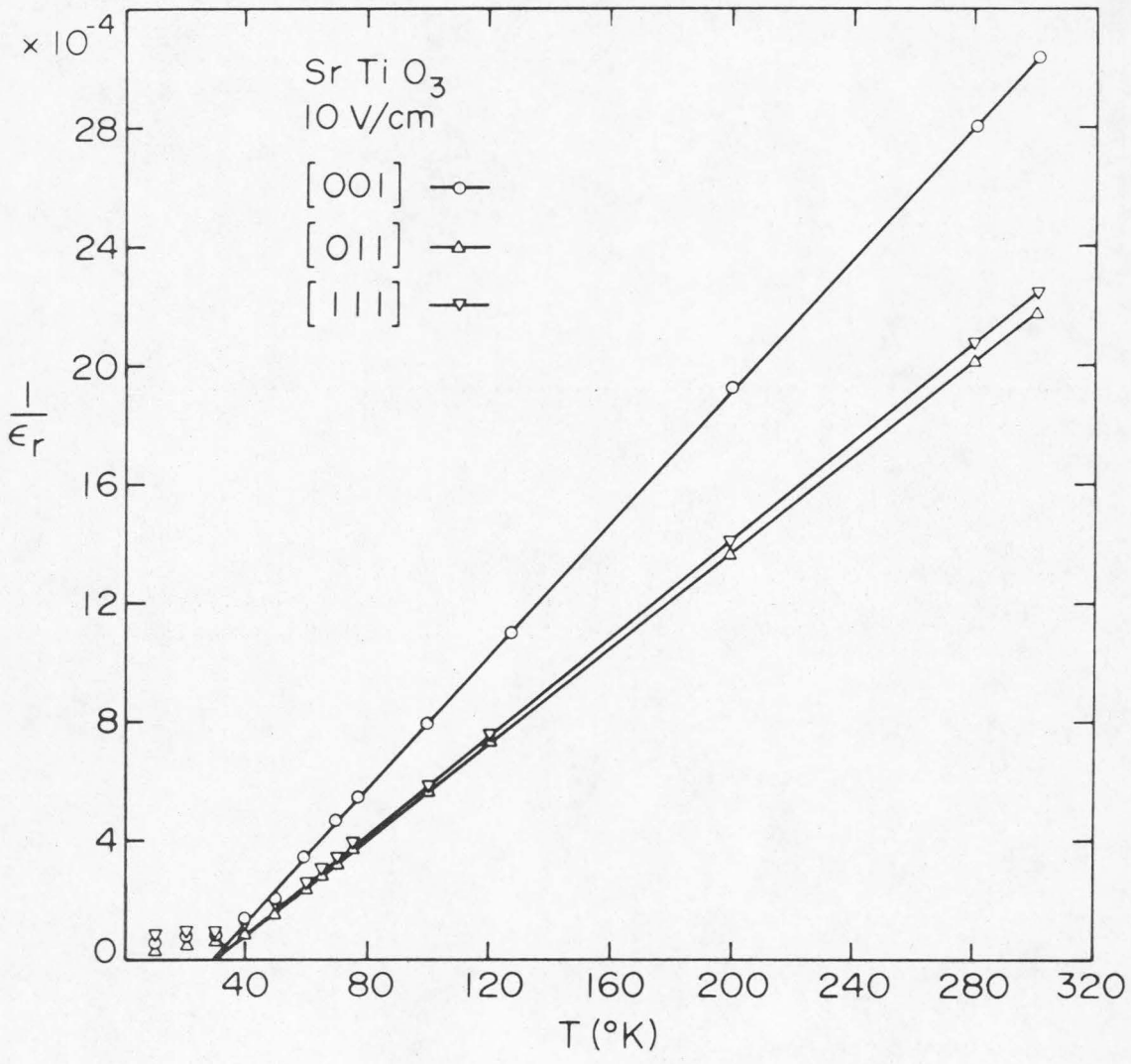

Fig. 5 Inverse relative permittivity of strontium titanate as a function of temperature with crystal orientation as a parameter. The applied D.C. bias field is 10 V/cm. Data by R. Neville.

summarized in Table 3. If the ions are assumed to be solid spheres, the lattice constant would be 4.00Å if determined by the O-Ti distance, 3.45Å if determined by the O-Sr distance and 3.73Å if determined by the O-O distance. The lattice constant is thus determined by the O-Ti bond. The actual unit cell cube edge is 3.9Å[2]. If the ions were solid spheres, the Sr^{2+} ion would be free to move at least 0.3Å from its equilibrium position. This freedom of motion of the Sr ion might give rise to the "soft" phonon mode. We shall see, however, that the ion displacements are quite small, so that Ti ion motion might also be important.

The greatest polarization density indicated in Fig. 3 is $9.87 \cdot 10^{-2} C/m^2$, which is obtained by applying an electric field of 23000 V/cm in the [011] direction at $4.2^{o}K$. Taking into account the electronic polarizability[5], this polarization density corresponds to a 0.075Å displacement of the Sr^{2+} ions, or to a 0.037Å displacement of the Ti^{4+} ions with respect to the rest of the lattice.

Let us consider the free energy vs. polarization curves of Fig.3. At low polarization densities $F \propto P^2$ which is characteristic of linear restoring forces. At higher polarization densities there is departure from $F \propto P^2$, indicating a "hardening" of the restoring forces. The nearest neighbor Ti-O distance decreases fastest if the Ti ion is displaced in the [001] direction, and least if the Ti ion is displaced in the [111] direction. Thus if Ti ion displacements are significant, we expect the greatest departure from $F \propto P^2$ for electric fields applied in the [001] direction and the smallest departure for fields applied in the [111] direction. This is in fact observed in Fig. 3. If only the

Table 3 Approximate ionic radii and atomic weights of Ti^{4+}, O^{2-} and Sr^{2+}, obtained from the 51st edition of the Handbook of Chemistry and Physics published by The Chemical Rubber Company.

	Ti^{4+}	O^{2-}	Sr^{2+}
Radii, Å	0.68	1.32	1.12
Atomic Weight	47.9	16.0	87.6

Sr ion were displaced with respect to the rest of the lattice, we would expect the greatest departure from $F \propto P^2$ in the [011] direction and the smallest departure in the [001] direction. The "hardening" of the restoring forces is therefore indicative that the Ti ion motion is significant.

Verification of the theory of ionic dielectrics[5] is complicated in strontium titanate by the fact that there are two important active optic modes (modes 1 and 4 of Table 2). The theory for cubic lattices with two ions per unit cell can be generalized in a simple manner by making the important approximation that the local electric field is the same at all ion sites. To test the theory we shall make the following additional simplifying approximations: (a) the lattice formed by the oxygen ions is rigid, (b) Mode 1 of Table 2 is assigned to oscillation of the strontium ions with respect to the rest of the lattice as indicated by Last[12], and (c) Mode 4 is assigned to oscillation of the titanium ions with respect to the oxygen lattice as indicated also by Last.

With these approximations the permittivity of the crystal is given by

$$\varepsilon = \varepsilon_o \frac{1+2\beta}{1-\beta} \tag{9}$$

where ε_o is the permittivity of free space, $3\, a^3\, \varepsilon_o \beta$ is the polarizability of the unit cell[5] and a^3 is the unit cell volume. At frequencies below the "soft" transverse optic mode β has three contributions: β_{el} due to electron displacements,

$$\beta_4 \equiv \frac{q_4^2}{3\, C_4\, a^3\, \varepsilon_o} \qquad (10)$$

due to titanium displacements, and

$$\beta_1 \equiv \frac{q_1^2}{3\, C_1\, a^3\, \varepsilon_o} \qquad (11)$$

due to strontium displacements. C_1 is the restoring force constant acting on each Sr ion defined by $C_1 x_1 = q_1 E_{loc}$, where q_1 is the charge of the Sr ion, x_1 is its displacement with respect to the oxygen lattice and E_{loc} is the local electric field. C_4 is the restoring force constant acting on the Ti ion, and q_4 is its charge. At frequencies between the "soft" mode and $16.4 \cdot 10^{12}$Hz, (see Table 2), $\beta = \beta_{el} + \beta_4$ and at optic frequencies $\beta = \beta_{el}$, as indicated in Fig.4. Applying Eq. (9) to the three frequency ranges we obtain $\beta_{el} = 0.592$, $\beta_4 = 0.195$ and $\beta_1 = 0.204$ at $300^{\circ}K$*. Since $\beta_4 \simeq \beta_1$ the two active optic modes give approximately the same contribution to the low frequency permittivity of strontium titanate, i.e. if a static electric field is applied to the crystal, then $x_1 q_1 \simeq x_2 q_2$.

The frequency of the long wavelength longitudinal mode 4, W_{L4}, is given by

$$\mu_4 W_{L4}^2 = C_4\left(1 + \frac{2\beta_4}{1+2\beta_{el}}\right) \qquad (12)$$

---

*The near equivalence of β_1 and β_4 gives strength to the approximation that E_{loc} is the same at all ion sites.

and for mode 1

$$\mu_1 W_{L1}^2 = C_1 \left[1 + \frac{2\beta_1}{1+2(\beta_{e1}+\beta_4)} \right] \qquad (13)$$

The reduced mass of the Sr ion oscillating against the rest of the lattice is $\mu_1 = 45.8 \cdot m_o$ and that of the Ti ion oscillating against the oxygen lattice is $\mu_4 = 24.0 \cdot m_o$, where m_o is the proton rest mass (see Table 3). From Eqs. (12) and (13) and the values of w_{L1} and w_{L4} in Table 2, we obtain $C_1 = 525$ N/m and $C_4 = 820$ N/m .

The theory can now be checked by calculating the charges of the Sr and Ti ions using Eqs. (10) and (11). The results are $q_1 = 2.56$ electronic charges for the strontium ions and $q_4 = 3.14e$ for the titanium ions, leaving -1.90e for the oxygen ions. Considering the approximations made in this analysis the calculated charges are in reasonable agreement with the expected ion charges in the strontium titanate crystal. If we attribute mode 1 of Table 2 to Ti motion and mode 4 to Sr motion (instead of vice-versa), we get 3.56e for the strontium ions and 2.25e for the Ti ions, in disagreement with the expected charges of the ions.

3. Multi-Domain Strontium Titanate

The curves of Fig. 1 correspond, within experimental error, to the equivalent circuit presented in Fig. 6. The equivalent circuits of three typical samples are summarized in Table 4. The low frequency permittivity of strontium titanate calculated using the Lyddane-Sachs-Teller

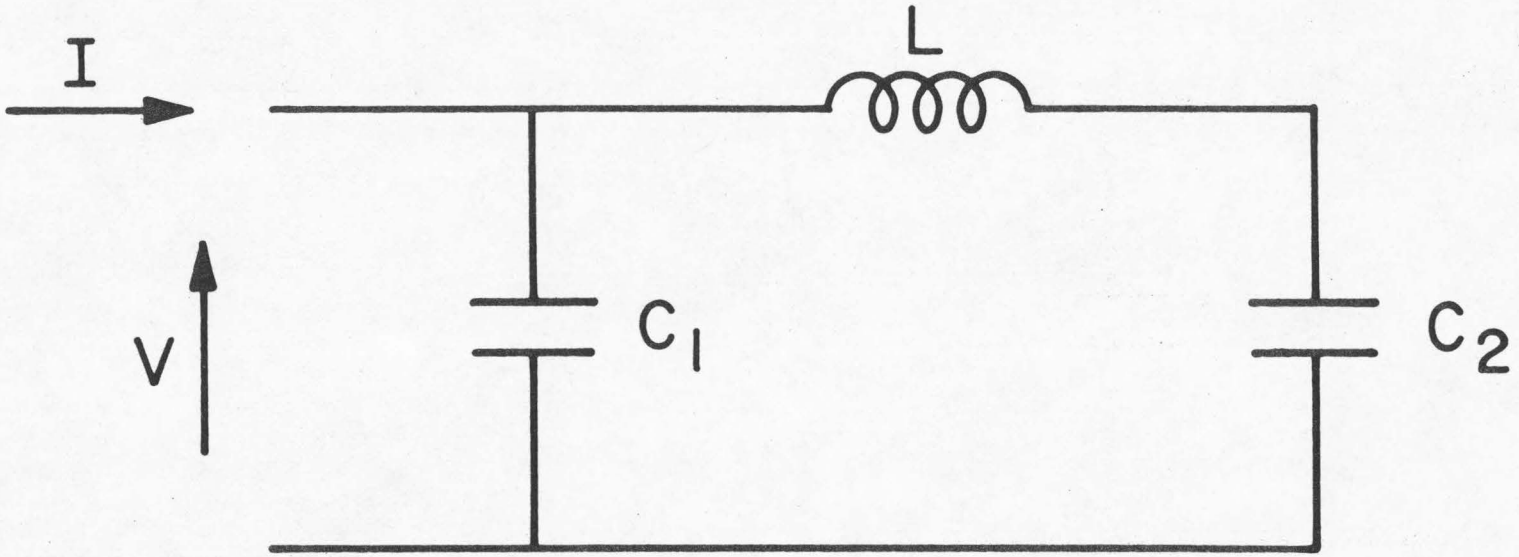

Fig. 6 Equivalent circuit of capacitors with multidomain strontium titanate dielectric below 65°K.

Table 4 Contact area A and wafer thickness ℓ of typical multi-domain samples oriented in the [001], [011] and [111] directions. The components C_1, C_2 and L, and the quality factor Q of the equivalent circuit of each sample are specified at 4.2°K and a D.C. electric field of 10 V/cm

Symbols	A	ℓ	C_1	C_2	L	Q
Units	mm^2	mm	pF	pF	μH	-
[001]	4.28	.125	12.2	8010	1.87	≃2.5
[011]	4.28	.125	2080	6470	0.89	≃ 10
[111]	4.28	.125	970	3730	0.663	≃ 4

relation, the measured optic phonon frequencies and the measured optical permittivity, is in quantitative agreement with the measurements on the single-domain crystals and on the multi-domain crystals below $3 \cdot 10^5$ Hz, as shown in Section 2.2. The anomalous resonance indicated in Fig. 1 cannot be attributed to a mechanical oscillation of the crystal as a whole since the resonant frequency is independent of the sample dimensions. The anomalous resonance is not due to an optical phonon since the lowest optical phonon has a natural frequency (given by Eq. 8) of $3.2 \cdot 10^{11}$ Hz at 4.2°K. Since the anomalous resonance is related to the presence of domains, it is probably due either to domain boundary motion or to a mechanical oscillation with a frequency determined by the domain boundary spacing. If the resonance were due to domain boundary motion, motion of the domain boundaries would be expected below the resonance frequency and not above it. The permittivity of single domain and multi-domain samples would therefore be equal above the resonant frequency, contrary to experiment. We have therefore assumed that the resonance is due to a mechanical oscillation with a frequency determined by the domain boundary spacing. The resonant frequency f is then of the order of $1/\sqrt{s\rho L^2}$, where ρ is the density, L is the domain boundary spacing and s is the "effective" elastic compliance constant of the particular mode of oscillation. Using a typical domain boundary spacing of 10 μm [3,2] and a resonant frequency of 2 MHz (see Fig. 1) we get $1/s \approx 2 \cdot 10^6$ N/m^2 at 4.2°K, which is smaller than the elastic stiffness constants c_{11}, c_{12} or c_{44} (measured above 110°K [11]) by a factor of $\approx 10^5$. A cubic crystal subject to a tensile stress in the [001] direction

has a Young modulus $1/s = (c_{11}^2 + c_{11}\,c_{12} - 2c_{12}^2)/(c_{11} + c_{12})$ and a Poisson ratio $c_{12}/(c_{11} + c_{12})$. Notice that if $c_{11} = c_{12}$, $1/s = 0$, the lattice becomes unstable and the crystal is free to undergo the cubic to tetragonal transition. In this transition the volume of the unit cell is conserved to first order in the strain. Measurements indicate that c_{11} approaches c_{12} as the temperature approaches 110°K from above [11]. Since from 110°K to 4.2°K the crystal undergoes a second (and perhaps a third[2]) phase transition, it is likely that the crystal lattice is marginally stable in this temperature range. The low value predicted for $1/s$ at 4.2°K is therefore possible.

The lattice of a multidomain crystal is presented schematically in Fig. 7. Notice that in the cubic to tetragonal transition two adjacent domains must "rotate" with respect to each other by $2\alpha \approx (c-a)/c$ in order that the two domains coincide at the domain boundary. The anomalous resonance is attributed to the "soft" oscillation in which the c/a ratio, and therefore α , vary with time. The kinetic energy per unit volume of such an oscillation is

$$T = \frac{1}{2}\rho\frac{L^2}{12}\dot{\alpha}^2 = \frac{1}{2}\rho\frac{L^2}{12}\dot{e}^2 \quad , \tag{14}$$

where e is the elastic strain. This oscillation can couple to the electric field if the crystal is ferroelectric and/or piezoelectric. If the crystal were ferroelectric, the low frequency permittivity of multi-domain crystals would be greater than that of single-domain crystals. This difference in permittivity is because the lattice in the multi-domain crystal can "rotate" with respect to the electric field, whereas

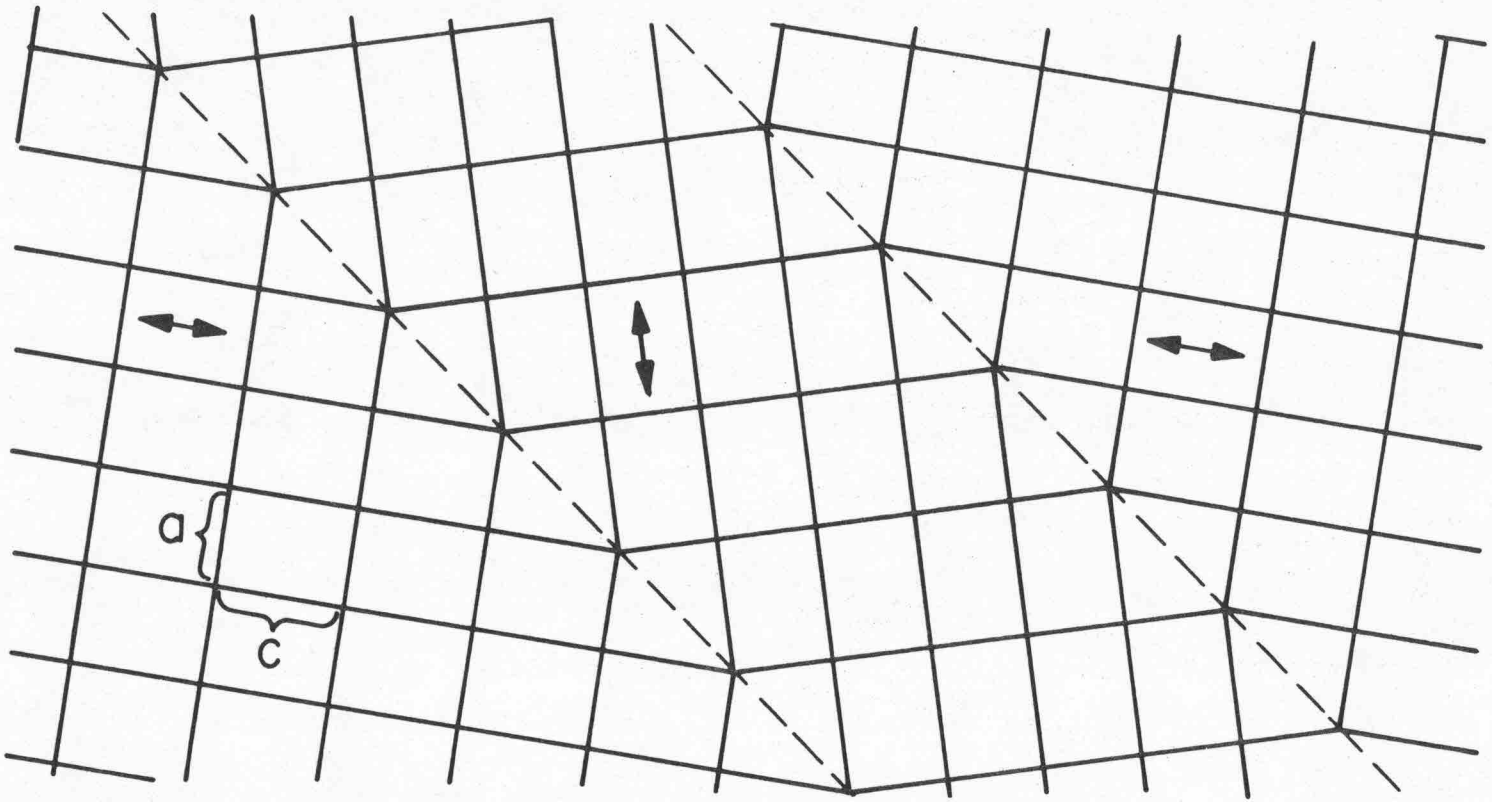

Fig. 7 Lattice of multidomain strontium titanate (schematic). Two domain boundaries are shown. The arrows indicate the [001] direction.

it cannot in the single domain sample. Since the low frequency permittivities are found to be equal, we conclude that the anomalous resonance is not due to ferroelectricity. In addition we have found no evidence that the crystal is ferroelectric. We therefore assume that the crystals are piezoelectric. In schematic one-dimensional notation a piezoelectric crystal is described by the equations [5]

$$P = Zd + E\chi \quad , \qquad e = Zs + Ed \quad , \tag{15}$$

where

P = polarization density, Cb/m^2 ,

Z = stress, N/m^2 ,

d = piezoelectric strain constant, m/V ,

E = macroscopic electric field, V/m ,

χ = dielectric susceptibility, F/m ,

e = elastic strain, - ,

s = elastic compliance constant, m^2/N .

Neglecting losses, the increment in kinetic energy per unit volume δT is equal to the work per unit volume done by the stress $-Z\delta e$. This equality determines the differential equation

$$\rho\frac{L^2}{12}\ddot{e} + \frac{1}{s}e = E\frac{d}{s} \tag{16}$$

which has the solution (in the frequency domain)

$$e = \frac{Ed}{1-w^2\rho\frac{L^2}{12}s} \tag{17}$$

The current density is given by

$$J = \frac{\delta}{\delta t} (P + \varepsilon_o E) \tag{18}$$

which (in the frequency domain) is

$$J = j w \varepsilon E - j w \frac{d^2}{s} E + \frac{j w \frac{d^2}{s} E}{1 - w^2 \rho \frac{L^2}{12} s} , \tag{19}$$

where $\varepsilon \equiv \varepsilon_o + \chi$ is the permittivity of the single domain crystal. Notice that Eq. 19 can be represented by the equivalent circuit of Fig.6. The apparent permittivity derived from (Eq.19) is plotted in Fig. 8. We should note that in a complete analysis the s and d in (Eq. 19) are actually functions of the tensor components s_{ijkl} and d_{ijk} respectively [14] and that the s in $\sqrt{\frac{12}{s\rho L^2}}$ and in $\frac{d^2}{s}$ (see Fig. 8) are different in general.

4. Conclusions

Dr. Richard Neville has measured[1] the permittivity of single crystal, single-domain and multidomain perovskite strontium titanate in the [001], [011] and [111] directions as a function of temperature (from 4.2 to 300°K), electric field (from 0 to ± 23000 V/cm) and frequency (from 1 KHz to 50 MHz). Strontium titanate crystals below 110°K are normally multi-domained. We have shown however that single domain crystals can be formed by cooling the samples past the 110 and 65°K phase transition temperatures with an applied electric field in excess of $2 \cdot 10^4$ V/cm. The samples become single domained possibly because the tetragonal crystal is piezoelectric. Other workers have formed single domain crystals by cool-

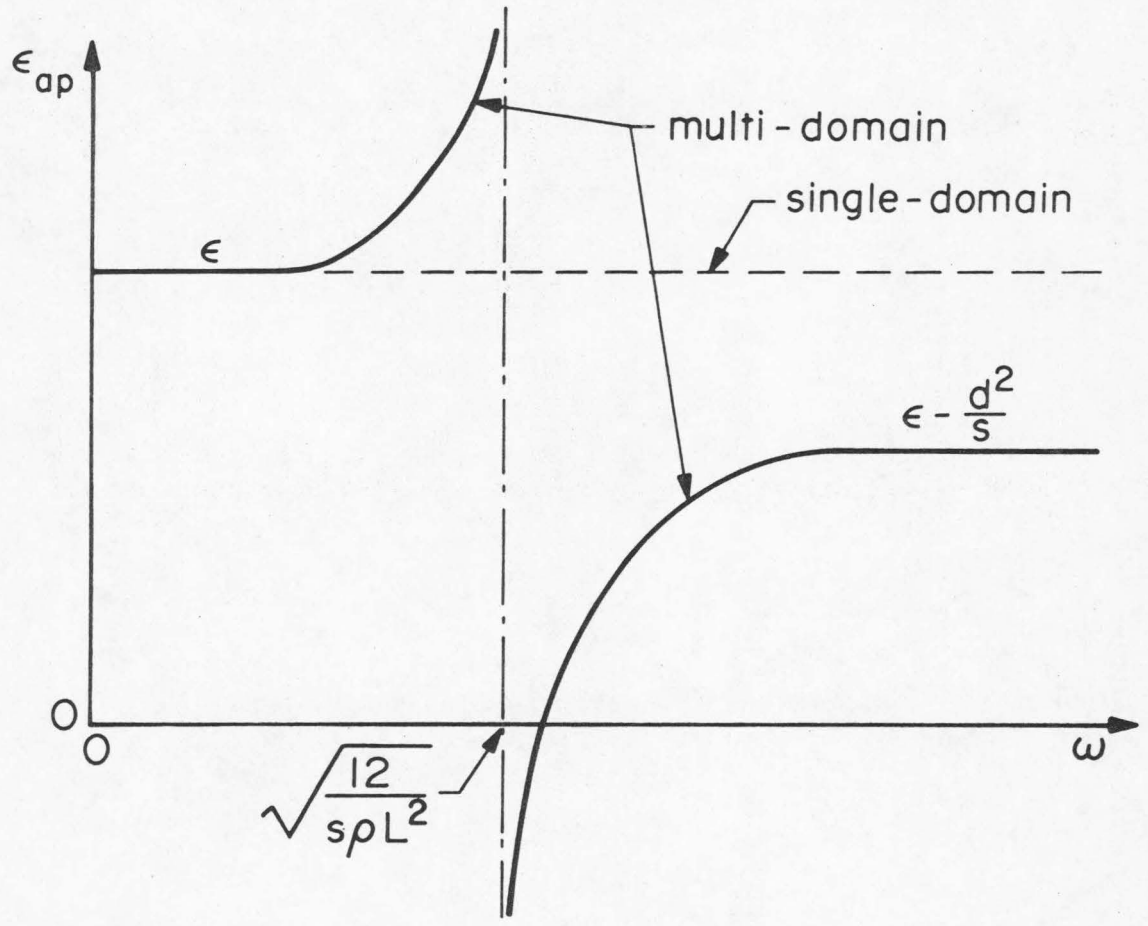

Fig. 8 Apparent permittivity of single domain and multidomain strontium titanate as a function of angular frequency, derived from Eq. 19.

ing the samples past the transition temperatures with an applied uniaxial stress[3].

The crystal free energy was calculated as a function of polarization. The Curie-Weiss law is satisfied in the temperature range from 40°K to 300°K, giving a Curie temperature of 30 ± 2°K for the three crystal orientations. The Lyddane-Sachs-Teller relation is satisfied for temperatures between 30°K and 300°K and electric fields between 0 and 12000 V/cm. A generalized LST relation was used to calculate the permittivity of strontium titanate from zero to optical frequencies. The calculated optical frequency permittivity is in excellent agreement with the literature. It was concluded that only two of the long wavelength optic modes propagating in the [001] direction generate dipole moments.

The perovskite strontium titanate crystal is formed of a tightly bound oxygen-titanium lattice which leaves ample spaces for the strontium ions. The O-Ti lattice determines the lattice constant and, most likely, the mechanical characteristics of the crystal. Each small Ti ion is tightly surrounded by six half oxygen ions. Thus the lattice formed by the oxygen ions screens the Ti ions from the Sr ions. We are led to believe that the two electrically active modes are mainly due to strontium motion and to titanium motion respectively. The low frequency active optical mode is attributed to motion of the heavy, loosely bound Sr ions with respect to the rest of the lattice and the high frequency active mode is attributed to motion of the lighter and more tightly bound Ti ions with respect to the oxygen lattice as indicated by Last. These assumptions

are seen to be consistent with the theory of the permittivity of ionic crystals. The two active optical modes give approximately equal contributions to the low frequency permittivity of strontium titanate. It was seen that departures of the free energy from $F \propto P^2$ indicate Ti ion motion. The restoring force that acts on a titanium ion begins to "harden" when this ion is displaced by approximately 0.002Å from its equilibrium position.

An anomalous resonance which multi-domain perovskite strontium titanate crystals exhibit below 65°K is described. The resonance is attributed to the presence of domains. The elastic stiffness constants c_{11} and c_{12} approach each other as the crystal is cooled to 110°K. When $c_{11} = c_{12}$, the lattice becomes unstable and the cubic to tetragonal transition occurs. If $c_{11} \approx c_{12}$, the Young's modules in the [001] direction is small and little energy is required to change the c/a ratio of the unit cell (keeping the volume constant to first order in the strain). An oscillation in which the c/a ratio varies with time can account for the observed anomalous resonance if the crystal is piezoelectric. The frequency of the resonance would be determined by the domain boundary spacing. Further experimental evidence is required to have a more reliable and detailed understanding of the anomalous phenomenon.

REFERENCES - PART III

1. R. Neville, B. Hoeneisen and C. A. Mead, J. Appl. Phys., in press.
2. F. W. Lytle, J. Appl. Phys., 35, 2212 (1964).
3. T. S. Chang et al., Appl. Phys. Lett., 17, 254.
4. R. H. Lyddane, H. Sachs and E. Teller, Phys. Rev. 59, 673 (1941).
5. C. Kittel, "Introduction to Solid State Physics", John Wiley & Sons, Inc., New York (1968).
6. A. S. Barker, Jr., Phys. Rev., 136, A1290 (1964).
7. R. A. Cowley, Phys. Rev., 134, A981 (1964).
8. J. M. Worlock and P. A. Fleury, Phys. Rev. Lett., 19, 1176 (1967).
9. W. G. Nilsen and J. G. Skinner, Journal of Chem. Physics, 48, 2240 (1968).
10. A. S. Barker, Jr. and M. Tinkham, Phys. Rev., 125, 1527 (1962).
11. R. O. Bell and G. Rupprecht, Phys. Rev., 129, 90 (1963).
12. J. T. Last, Phys. Rev., 105, 1740 (1957).
13. W. G. Spitzer, R. C. Miller, D. A. Kleinman and L. E. Horvath, Phys. Rev., 126, 1710 (1962).
14. J. F. Nye, "Physical Properties of Crystals", Oxford University Press, (1957).