Fundamental Limitations in Microelectronics—I. MOS Technology[*]

B. Hoeneisen and C.A. Mead

Abstract—The physical phenomena which will ultimately limit MOS circuit miniaturization are considered. It is found that the minimum MOS transistor size is determined by gate-oxide breakdown and drain-source punch-through. Other factors which limit device size are drain-substrate breakdown, drain "corner" breakdown, and substrate doping fluctuations. However, these limitations are less severe than the oxide breakdown limitation mentioned above. Power dissipation and metal migration limit the frequency and/or packing density of fully dynamic and of complementary MOS circuits. In static non-complementary circuits, power dissipation is the principal limitation of the number of circuit functions per chip. The channel length of a minimum size MOS transistor is a factor of 10 smaller than that of the smallest present-day devices. The tolerances required to manufacture such a transistor are compatible with electron beam-masking techniques. It is thus possible to envision fully dynamic silicon chips with up to $10^7 - 10^8$ MOS transistors per cm^2.

1   Introduction

Development of the planar technology in the late 1950s 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[1].

Although the chip area has increased by a factor of $\approx$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 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.

The design rules for present-day MOS circuits involve limitations of several types. Spacing between the drain and source regions is typically limited by punch-through, a condition where the depletion regions of the two junctions overlap. Other spacings are set primarily by the tolerances in alignment of successive masks. Even with present-day techniques, tolerances are improving steadily. As electron beam pattern-generation techniques become more generally available, mask alignment to a much higher

---

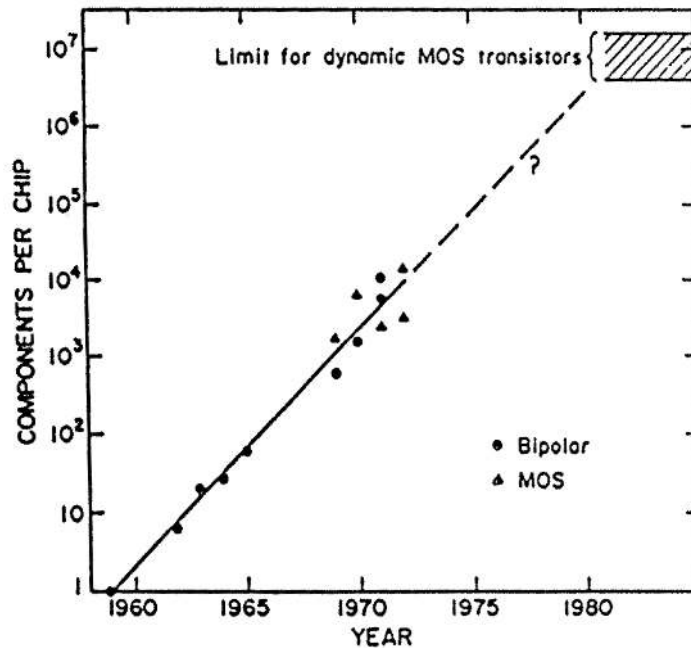   [1]G.E. Moore – private communication

Figure 1: History of integrated circuit complexity. Line corresponds to a twofold increase in the number of components per chip per year. This figure was provided by Gordon E. Moore.

precision may be envisioned. With these important developments approaching, it is important to identify clearly the fundamental limitations which will ultimately limit MOS circuit miniaturization.

It must be stressed that we do not determine the ultimate limits in microelectronics, but only the ultimate limits of MOS field-effect transistor circuits as we know them today. Only planar transistors with silicon substrate and silicon dioxide dielectric are considered. The limits we determine can be approached as tolerances and yields improve.

2 Principal Limitations of MOS-Integrated Circuits

The maximum number of circuit functions per unit area is determined either by power-dissipation density or by the area occupied by transistors, interconnections, and passive devices (if any). For given circuit capacitances and frequency of operation, a lower supply voltage implies lower currents, lower power dissipation, and lower interconnection area per transistor. Making the devices smaller not only reduces the area occupied by these devices, but also reduces the circuit capacitances. For a given frequency of operation and supply voltage, lower circuit capacitances imply lower currents, lower power dissipation, and lower interconnection area per transistor. In addition, lower voltage devices can be made smaller. Thus, we conclude that to maximize the packing density, it is necessary to minimize the supply voltages and the size of individual devices.

The supply voltage has a lower bound which is determined by reproducibility of the gate turn-on voltage, 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 power at any given time. 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 MOSFET is inherently more stable than that of a p-channel device.[2] 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 $\approx 2 \cdot 10^{19} \text{cm}^{-3}$ 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 $\approx 1.3 \cdot 10^{19} \text{cm}^{-3}$. At higher concentrations, the maximum electric field which can be applied to the gate oxide, ($\approx 6 \cdot 10^6 \text{ V/cm}^2$), does not invert the substrate. The junction built-in voltage produces a depletion thickness of $0.01 \, \mu\text{m}$ 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 regions thicknesses, or $\approx 0.02 \, \mu\text{m}$. Otherwise, the two junctions would be in punch-through, even with no applied bias.

The gate-oxide thickness has a lower limit of $\approx 50 \mathring{A}$ 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. 4 of the Appendix). 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_0, 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. 4 is much larger than $|V_{FB} + 2\phi|$, the gate turn-on voltage is proportional to $x_0\sqrt{C_B}$. Thus, for a given relative manufacturing tolerance of x_0 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 $V_{GT} \geq |V_{FB} + 2\phi|$. Therefore, gate turn-on voltages as low as $\approx |V_{FB} + 2\phi|$, 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 $2V_{GT}$.

---

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

3   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 Fig. 5. 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 $V_0 = V_{DD}$. This situation is a particular case of the more general problem considered in Fig. 5. 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 MOSFET with an n^+ silicon gate, if the silicon-silicon dioxide interface charge Q_{ss} is made negligible ($\leq 10^{11}$ electronic charges per cm^2 for the thin-gate oxides considered an easily achievable value).

M. Lentzlinger 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_0 provided that x_0 is large enough. For an n-channel MOSFET with an Al or n^+ silicon gate, the oxide current density is [2] $\approx 10^{-10}$A/cm^2 for an oxide electric field of $\pm 6 \cdot 10^6$ V/cm, provided that $x_0 \geq 50\mathring{A}$. 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 $F_{\text{ox}} = 3 \cdot 10^6$ 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 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,[3] the drain depletion region thickness is

$$W = \sqrt{\left(\frac{2\,\epsilon(V_{DD} + \varphi)}{qC_B}\right)} \tag{1}$$

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

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 $V_0 = \text{OV}$ (see Fig. 5). The minimum gate-oxide thickness of the pull-up transistor 2 is obtained from Eq. 5. The maximum substrate doping concentration is obtained from Eq. 6 or from Fig. 5. 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_0. (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. 4, and the required gate turn-on voltage ($V_{GT1} = 0.5\,V_{DD}$). With

---

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

Minimum Channel Length $2r$, μm (y-axis)

$x_0 = 500$ Å   A2
400
300
$3 \cdot 10^{16}$ cm^{-3}
$5 \cdot 10^{16}$ gate oxide field
200
$C_B = 1 \cdot 10^{17}$
B
100   $1 \cdot 10^{17}$ A1 $2 \cdot 10^{17}$ cm^{-3}
drain-substrate breakdown
50
C
$5 \cdot 10^{17}$
$1 \cdot 10^{18}$
drain "corner" breakdown
0.1 $2 \cdot 10^{18}$
$\leftarrow 2r \rightarrow$
Source   Drain
$5 \cdot 10^{18}$ r   r
Substrate
$1 \cdot 10^{19}$

$r = W = \sqrt{\dfrac{2\epsilon}{q}\dfrac{(V_{DD}+\varphi)}{C_B}}$

0.01
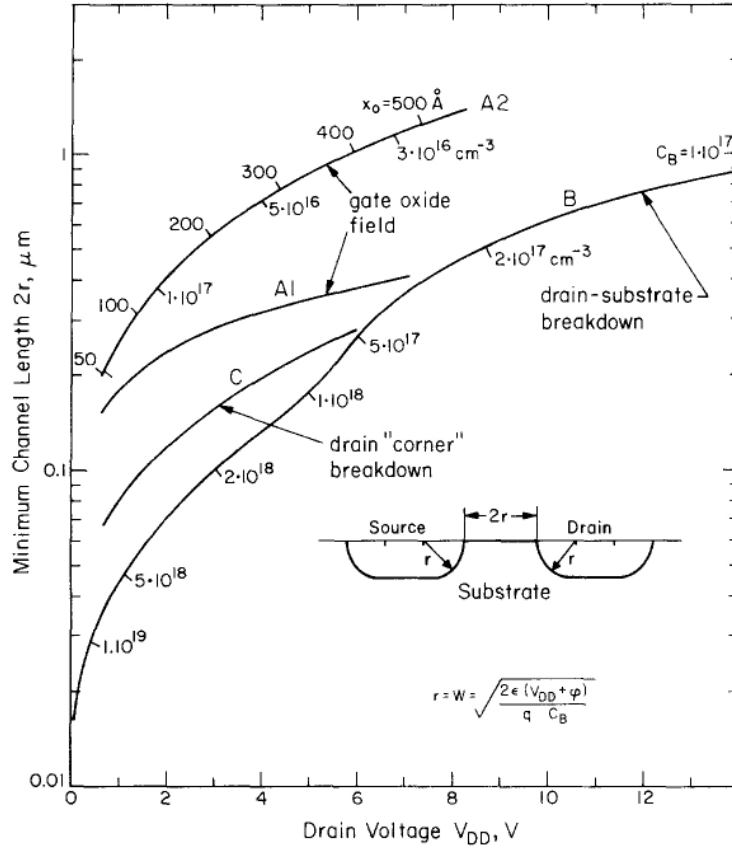0   2   4   6   8   10   12

Drain Voltage V_{DD}, V

Figure 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.

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. 6. 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[4] by A.S. Grove et al. [4]:

$$F_c \approx \frac{2(V_D + \varphi)}{W} + \frac{(V_D + \varphi) - (V_G - V_{FB})}{\frac{\epsilon}{\epsilon_{\mathrm{ox}}} x_0}. \tag{2}$$

Here, F_c is the "corner" electric field, and W is 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

---

[4]To insure that the "corner" electric field is correct in the two limiting cases $W \gg 3x_0$ and $W \ll 3x_0$, a factor of 2 has been added to the first term on the right-hand side of Grove's [4] expression.

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. 8, drain "corner" breakdown occurs.

Let us again consider the inverter showing Fig. 5. The driver transistor 1 may have drain "corner" breakdown when its gate is low ($V_i = $ OV) and its drain is high ($V_0 = V_{DD}$). It is assumed that the gate oxide thickness x_0 is chosen the same for both transistors. Then x_0 is obtained, as before, by applying Eq. 5 to transistor 2. The minimum channel length $2r$ of transistor 1, limited by drain "corner" breakdown, is estimated by setting $r = W$, where W is obtained from Eq. 2 with $F_c = F_B \approx 1.5 \cdot 10^6$ V/cm as shown in Fig. 8. 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 Fig. 2.

A minimum-size transistor with $V_{DD} = 0.7$V has a gate-oxide thickness of $50\mathring{A}$ as shown in Fig. 2. Since thinner oxides cannot be used due to tunneling from gate to substrate, $V_{DD} = 0.7$V 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.

4 Example

As a specific example, we shall choose $V_{DD} = 2$V and $V_{GG} = 4$V. The gate-oxide thickness is calculated by applying Eq. 5 of the Appendix to transistor 2. The result is $x_0 = 140\mathring{A}$ as indicated in Fig. 2, curve A2. The substrate doping concentration of transistor 2 is obtained from Eq. 4, and the required gate turn-on voltage ($V_{GT2} = 3$V when $V_0 = 2$V). The result is $C_{B2} = 9.2 \cdot 10^{16}$cm^{-3} as indicated in Fig. 5, 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. 4, and the required gate turn-on voltage of transistor 1 ($V_{GT1} = 1$V). 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 0.12 μm for transistor 1, and 0.205 μm for transistor 2. The minimum channel length, limited by drain-source punch-through is approximately twice the drain depletion thickness, or 0.24 μm for transistor 1, and 0.41 μm 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.
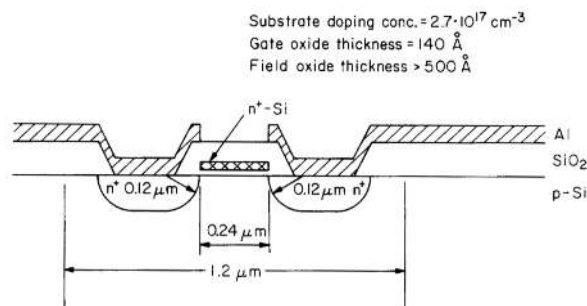

Figure 3: Typical silicon gate MOS transistor of minimum size.

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 MOSFET model which includes velocity saturation of the charge carriers [5].
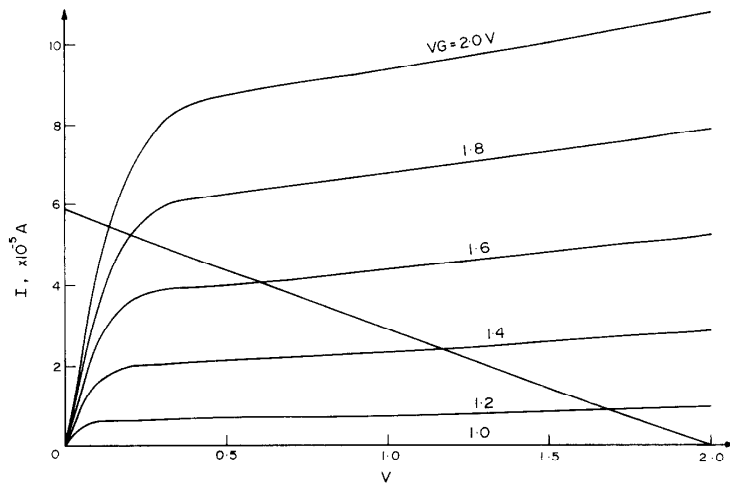

Figure 4: Drain family of the driver transistor and load line of the pull-up transistor of a minimum-size static inverter. $L_1 = Z_2 = 0.24$ μm, $L_2 = Z_1 = 0.96$ μm, $C_{B1} = 2.7 \cdot 10^{17}$cm^{-3}, $C_{B2} = 9.2 \cdot 10^{16}$cm^{-3}; $x_0 = 140\mathring{A}$, $\mu = 250$ cm^2/V sec and $V_{FB} = -1.0$ V for both transistors.

5   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 device's $I - V$ characteristics, e.g., gate turn-on voltage, and the device's 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-percent 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 $\approx$20 percent. This 20-percent variation corresponds to a substrate doping fluctuation of approximately 40 percent 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 Fig. 2, we have $W \approx r$. With an 80-percent certainty, the doping fluctuation does not exceed 40 percent in any one of the 10^6 cubes of volume r^3, if these cubes have in the average $\approx$170 ionized dopant atoms. The smallest size transistor shown in Fig. 2 corresponds to a driver transistor with a gate-oxide thickness of 50 $\mathring{A}$, a substrate doping concentration of $4 \cdot 10^{17}$cm^{-3}, and a channel length $2r = 0.15$ μm. Such a transistor has $\approx$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."

6   Power Dissipation Density

In this section, we shall show that for fully dynamic MOSFET 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 MOSFET 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 Fig. 5). The
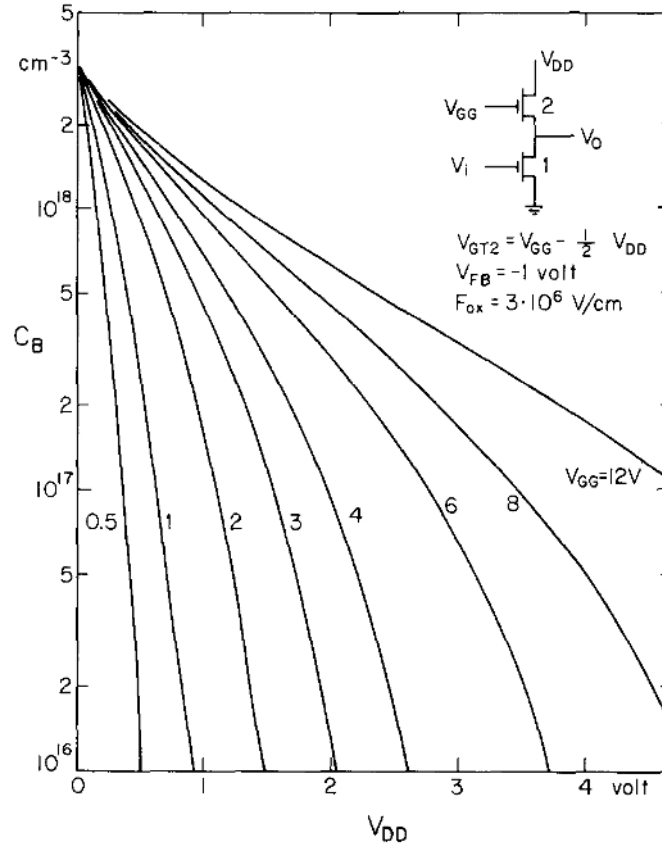

Figure 5: Maximum substrate doping concentration C_B of the pull-up transistor of an inverter, as a function of V_{DD} and V_{GG}, determined by the maximum allowable gate-oxide electric field F_{ox}.

power dissipation density of densely-packed dynamic inverters is

$$P = \frac{CV_{DD}^2}{S} f \tag{3}$$

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 10MHz 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

8

static inverters have been calculated assuming 50-percent duty cycle. The results are also presented in Table 1.

6.1 Table 1

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 $S \approx 90r^2$. Power calculated at 10 MHz. For static inverter $L_1 = Z_2 = 2r$, $L_2 = Z_1 = 8r$, $S \approx 190r^2$. The load capacitance C is equal to the gate plus drain junction capacitances of transistor 1

Type	V_{DD} (V)	$2r$ (μm)	x_0 (Å)	C (F)	S (cm²)	P (W/cm²)
Dynamic	1	0.25	72	4.6×10^{-16}	1.4×10^{-8}	0.32
Dynamic	2	0.41	140	6.7×10^{-16}	3.8×10^{-8}	0.71
Dynamic	4	0.72	274	11.0×10^{-16}	12.0×10^{-8}	1.5
Static	1	0.18	72		1.5×10^{-8}	960.0
Static	2	0.24	140		2.7×10^{-8}	2000.0
Static	4	0.32	274		4.9×10^{-8}	4100.0
Static	7	0.41	475		8.0×10^{-8}	6900.0

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 power at any given time.

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 $V_0 = 0$ V, and the power dissipation density would be $P \approx 6(\frac{CV_{DD}^2}{S})f$. In this case, as with a fully-dynamic MOSFET circuit, power dissipation would only limit the operating frequency. The use of a 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 ohm per square resistors. However, 10 MΩ resistors would typically be required.

7 Metal Migration Limitation

When a high-current density flows through a metallic conductor, migration of the metallic atoms occurs [6]. This phenomenon is an important reliability consideration in 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 lower than 10^6A/cm² [6]. This limitation is similar in nature to the power-dissipation limitation; it does not limit the minimum-device size, but rather limits the operating frequency and/or the number of circuit functions per chip.

In fully dynamic or in complementary MOS circuits, only capacitive currents flow, i.e., currents which either charge or discharge the circuit capacitances. Thus, for a given circuit configuration, the maximum allowable current density in the metallic conductors determines the maximum charging rate of the circuit capacitances and therefore the maximum operating frequency.

Consider a chip with 10^6 fully dynamic minimum-size inverters with $V_{DD} = 2$V and V_{GG} switched between 0 and 4V. We shall assume that an aluminum line of width and thickness equal to $2r$ (i.e., 0.41 μm) is connected to V_{GG} of 10^3 inverters. The gate capacitance of the 10^3 transistors is ≈ 0.42 pF. With a maximum allowable instantaneous current density in the metal line of 10^5A/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.

8   Conclusion

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. In static non-complementary MOS circuits, the number of devices per chip is limited by power dissipation, except for circuits such as read-only memories in which only a small fraction of the devices dissipate power at any given time. The maximum-packing density of fully dynamic or complementary MOS circuits is determined by the area occupied by the transistors and interconnections. Both power dissipation and metal migration limit the frequency of operation of fully dynamic or of complementary circuits.

The minimum channel length of a 2 V transistor is ≈ 0.4 μm. This length is a factor of 10 smaller than the channel of the smallest present-day devices. The mask alignment tolerances required to manufacture such a device are within the capabilities of electron beam pattern-generation techniques. Thus, we can envision fully dynamic or complementary integrated silicon chips with up to $\approx 3 \cdot 10^7$ MOS transistors per cm^2, operating in the 10 to 30 MHz range, as shown in Fig. 1.

The maximum packing density of read-only memories is determined by the area occupied by the devices and interconnections. For example, a read-only memory with a supply voltage of 1.2 V and with channel width-to-length ratios of 3/1 and 1/3 for the driver and pull-up transistors respectively, can have up to $\approx 1 \cdot 10^8$ transistors per cm^2 operating at a frequency of ≈ 0.5 MHz. Increasing the width-to-length ratios of the devices reduces the packing density and increases the maximum frequency by the same factor.

Present-day MOS charge-coupled shift registers occupy approximately 1/4 the area of MOS transistor shift registers (Fig. 7) due to the elimination of the supply lines and the source and drain diffusion regions. Charge-coupled devices (CCDs) 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.

References

[1] B.E. Deal et al., *J. Electrochem. Soc.,* Vol. 114, p. 266 (1967).

[2] M. Lenzlinger and E.H. Snow, *J. of Appl. Phys.,* Vol. 40, p. 278 (1969).

[3] N. Klein. *IEEE Transactions on Electron Devices,* ED-13, p. 788 (1966).

[4] A.S. Grove, O. Leistiko and W.W. Hooper, *IEEE Transactions on Electron Devices,* ED-14, p. 157 (1967).

[5] B. Hoeneisen and C.A. Mead, *IEEE Transactions on Electron Devices,* Vol. 19 (3), pp. 382-383 (1972).

[6] I.A. Blech and E.S. Meieran, *Applied Physics Letters,* Vol. 11, p. 263 (1967).

[7] L. Altman. *Electronics,* Vol. 44 (13), p. 50 (1971).

[8] H. Weinerth, *Solid-State Electronics,* Vol. 10, p. 1053 (1967).

[9] A.G. Chynoweth et al., *Phys Rev.,* Vol. 118, p. 425 (1960).

[10] S.L. Miller, *Phys. Rev.,* Vol. 105, p. 1246 (1957).

[11] J. Shields, *Journal of Electronics and Control,* Vol. 6, p. 130 (1959).

[12] R.A. Logan and A.G. Chynoweth, *Phys. Rev.,* Vol. 131, p. 89 (1963).

[13] J.C. Irvin, *Bell System. Tech. J.,* Vol. 41, p. 387 (1962).

[14] S.M. Sze and G. Gibbons, *Solid-State Electronics,* Vol. 9, p. 831 (1966).

9 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

$$V_{GT} = V_{FB} + V_S + 2\phi + \frac{X_0}{\epsilon_{\text{ox}}}\sqrt{[2\epsilon q C_B(V_S + 2\phi)]}. \tag{4}$$

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

$$X_{0\ \text{min}} = \frac{(V_{G\ \text{max}} - V_{FB}) - (V_{S\ \text{min}} + 2\phi)}{F_{\text{ox}}} \tag{5}$$

Here, $V_{G\ \text{max}} - V_{S\ \text{min}}$ is the maximum gate-source voltage. The maximum substrate doping concentration is determined from Eqs. 4 and 5 by setting $X_0 = X_{0\ \text{min}}$. The result is

$$C_{B\ \text{max}} = \left[\frac{V_{GT} - V_{FB} - V_S - 2\phi}{V_{G\ \text{max}} - V_{FB} - V_{S\ \text{min}} - 2\phi}\right]^2 \frac{\epsilon_{\text{ox}}^2\ F_{\text{ox}}^2}{2\epsilon q(V_S + 2\phi)}. \tag{6}$$

The particular circuit shown in Fig. 5 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 $V_0 = V_{DD}$. Here, V_{GT1} and V_{GT2} are the gate turn-on voltages of transistors 1 and 2, respectively (see Fig. 5). The maximum substrate doping concentration limited by oxide field is obtained by applying Eq. 6 to each transistor. For transistor 1

$$C_{B1} \leq \left[\frac{\frac{1}{2}V_{DD} - V_{FB} - 2\phi}{V_{DD} - V_{FB} - 2\phi}\right]^2 \frac{\epsilon_{\text{ox}}^2 F_{\text{ox}}^2}{2\epsilon q 2\phi}. \tag{7}$$

For transistor 2

$$C_{B2} \leq \left[\frac{V_{GG} - \frac{3}{2}V_{DD} - V_{FB} - 2\phi}{V_{GG} - V_{FB} - 2\phi}\right]^2 \frac{\epsilon_{\text{ox}}^2 F_{\text{ox}}^2}{2\epsilon q(V_{DD} + 2\phi)}. \tag{8}$$

Eq. 8, which is a more severe limitation than Eq. 7, is plotted in Fig. 5 for the case $V_{FB} = -1$ V and $F_{\text{ox}} = 3 \times 10^6$ V/cm.

10 Appendix 2: Reverse Breakdown of Low-Voltage Silicon Junction Diodes

Several authors [8–11] have measured the reverse "breakdown" voltage of one-sided silicon step junctions. Their results are presented in Fig. 6. The "breakdown" voltage V_B is defined as the applied voltage at a specified reverse current density. H. Weinerth [8] has shown that field emission is the main reverse conduction mechanism of low-voltage diodes ($V_B \leq 3V$); whereas, high-voltage diodes ($V_B \geq 8V$) are limited by avalanche breakdown. The reverse characteristics of diodes in the intermediate range ($3 \text{ V} \leq V_B \leq 8 \text{ V}$) can be explained [8] by avalanche multiplication of the field-emission current.
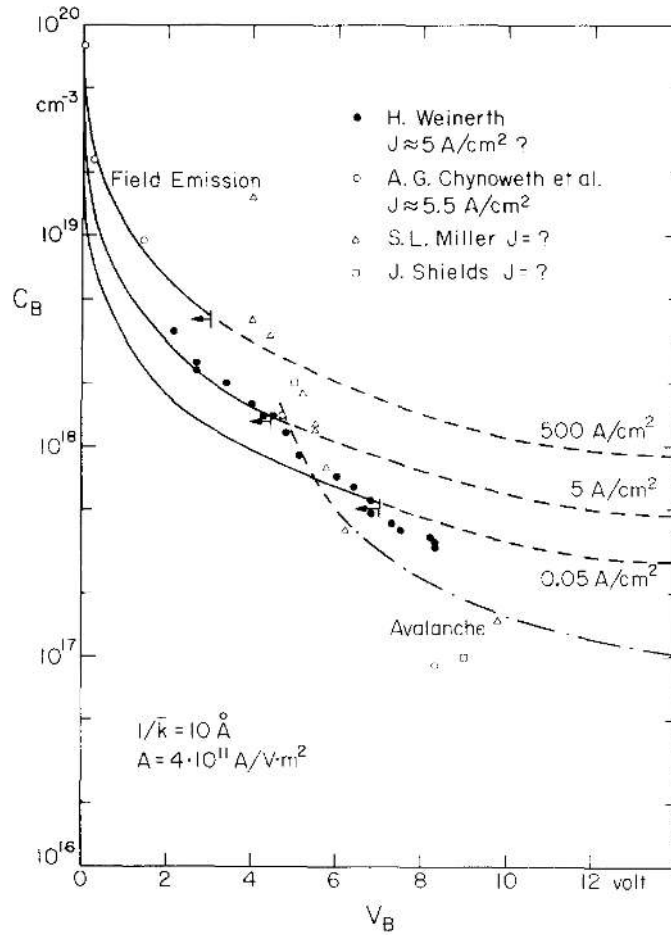

Figure 6: Reverse "breakdown" voltage V_B of one-sided silicon step-junction diodes as a function of doping concentration C_B. "Breakdown" is defined to occur when the reverse current density reaches the indicated value. Experimental data by several authors are shown [8–11]. For the date of Weinerth [8] and Chynoweth et al. [9], 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 [10] is also shown.

A reverse-biased n^+p junction is shown in Fig. 7. Electrons can tunnel through the energy gap from the p to the n^+ side as shown in 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:

$$P = e^{-2\mathbf{k}x} \tag{9}$$

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

$$x = \sqrt{\left(\frac{2\epsilon}{qC_B}\right)}\left[\sqrt{(y + E_g)} - \sqrt{(y)}\right] \tag{10}$$
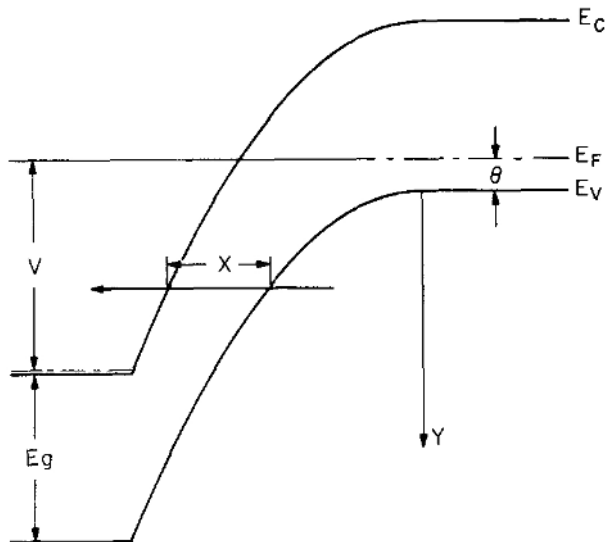
is the tunneling distance as shown in in Fig. 7.

Figure 7: Energy band diagram of a reverse-biased $n^+ - p$ diode. The arrow shows the electron tunneling path.

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

$$J = A \int_0^{V-\theta} e^{-2\mathbf{k}x} dy. \tag{11}$$

The average wave vector $\mathbf{k}$ was calculated from the tunnel-diode data of R.A. Logan et al. [12] and from data on the resistance of reverse-biased zener diodes taken by H. Weinerth [8]. Both calculations give $\mathbf{k} \approx \frac{1}{10}\mathring{A}^{-1}$. The proportionality factor A was chosen to fit the experimental data by H. Weinerth (shown in Fig. 6) at $V_B = 3\,\mathrm{V}$.

The "breakdown" voltage given by Eq. 11 is plotted in Fig. 6 for several current densities. Also is shown the experimental avalanche breakdown curve by S.L. Miller [10]. The maximum electric field in the junction at "breakdown" was calculated from the data presented in Fig. 6, using the standard expressions for one-sided step-junctions. The results are plotted in Fig. 8.

The theoretical field-emission curve fits the experiment quite well. H. Weinerth [8] calculated the field-emission current of intermediate-voltage diodes ($3\,\mathrm{V} \leq V_B \leq 8\,\mathrm{V}$), assuming that the reverse current is
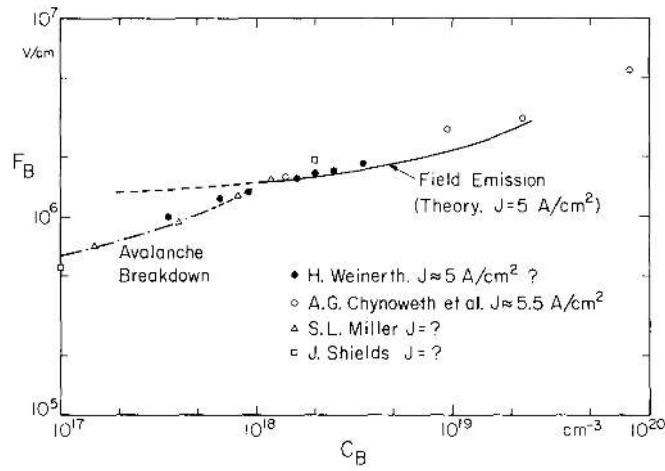
Figure 8: Maximum electric field F_B in a one-sided silicon step junction at "breakdown," as a function of doping concentration C_B. This electric field is calculated from the data of Fig. 6.

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 [14].