NASA CONTRACTOR REPORT

NASA CR-2620

EXPERIMENTAL INVESTIGATION OF A DOUBLE-DIFFUSED MOS STRUCTURE

H. C. Lin and J. L. Halsor

Prepared by
WESTINGHOUSE DEFENSE AND ELECTRONIC SYSTEMS CENTER
Baltimore, Md.
for Langley Research Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • MAY 1976

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-2620		

4. Title and Subtitle	5. Report Date
Experimental Investigation of a Double-Diffused MOS Structure	May 1976
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
H. C. Lin and J. L. Halsor	
	10. Work Unit No.
	506-18-21-03 (R4488)

9. Performing Organization Name and Address	11. Contract or Grant No.
Westinghouse Defense and Electronic Systems Center Systems Development Division Advanced Technology Laboratories Baltimore, Maryland	NAS1-12533
	13. Type of Report and Period Covered
	Contractor's Final Report

12. Sponsoring Agency Name and Address	14. Sponsoring Agency Code
National Aeronautics & Space Administration Washington, DC 20546	

15. Supplementary Notes

Langley Technical Monitor: Harry F. Benz

Final Report

16. Abstract

Double-diffused MOS (DMOS) transistors are capable of high-speed operation because of the reduced channel length that results when successive n- and p-type diffusions are done through the same source window. This effectively results in a channel length that is comparable to the base width of a double-diffused bipolar transistor (approx. 0.5 micrometer instead of several micrometers which is typical of standard MOSFET geometry). In addition, for optimum high frequency performance, it is desirable to minimize area and parasitic capacitance associated with standard metal gate electrodes. In this report, self-alined polysilicon gate technology is applied to the DMOS construction in a manner that retains processing simplicity and effectively eliminates parasitic overlap capacitance because of the self-alining feature.

In this report depletion mode load devices with the same dimensions as the DMOS transistors are integrated. The ratioless feature results in smaller dimension load devices, allowing for higher density integration with no increase in the processing complexity of standard MOS technology.

A number of inverters connected as ring oscillators were used as a vehicle to test the performance and to verify the anticipated benefits. The propagation time-power dissipation product and process related parameters were measured and evaluated.

This report includes (1) details of the process; (2) test data and design details for the DMOS transistor, the load device, the inverter, the ring oscillator, and a shift register with a novel tapered geometry for the output stages; and (3) an analytical treatment of the effect of the distributed silicon gate resistance and capacitance on the speed of DMOS transistors.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Metal Oxide Semiconductor Transistors Double-Diffused Field Shield	Unclassified - Unlimited Subject Category 33

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	78	$4.75

*For sale by the National Technical Information Service, Springfield, Virginia 22161

CONTENTS

	Page
INTRODUCTION	1
SILICON GATE DOUBLE-DIFFUSED MOS TRANSISTORS	2
INTEGRATED DOUBLE-DIFFUSED MOS STRUCTURE	3
ANALYSIS OF DMOS TRANSISTORS	5
DMOS INVERTER	7
Switching Transistor Design Considerations	11
Load Device Design Considerations	14
EFFECT OF SILICON GATE RESISTANCE ON THE FREQUENCY RESPONSE OF MOS TRANSISTORS	17
Silicon Gate as Distributed RC Network	18
Frequency Response of the Distributed RC Network	20
Time Constant of the Silicon Gate	21
Characteristics of the Silicon Gate Uniform Channel MOS Transistor	24
Characteristics of DMOS Transistors	26
DMOS RING OSCILLATOR	31
DMOS SHIFT REGISTER	34
AN OPTIMIZED OUTPUT STAGE FOR MOS INTEGRATED CIRCUITS	34
Output Stage Design Principles	36
Area Optimization	39
Optimum Stage Design	40
MASK DESIGN	44
FABRICATION	55
EXPERIMENTAL RESULTS	58
Typical Test Transistor Characteristics	58
Inverter Characteristics	60
Effect of Silicon Gate Resistance	60
Ring Oscillator Performance	66
SUMMARY AND CONCLUSIONS	69

Page

APPENDIX 72

Frequency Response of Distributed RC Lines 72

Transient Response 72

REFERENCES 74

FIGURES

		Page
1.	DMOS Structure	4
2.	Computed Drain Output DMOS Characteristics Velocity Saturating	8
3.	Computed Drain Output DMOS Characteristics Velocity not Saturating	9
4.	Computed Drain Output MOS Characteristics	10
5.	DMOS Inverter	12
6.	Threshold Voltage as Function of Substrate Concentration	16
7.	Silicon Gate MOS Transistor	19
8.	Distributed RC Network	19
9.	Equivalent Circuit of a Silicon Gate MOS Transistor	19
10.	Small-signal Response of Distributed RC Network Amplitude Response in dB vs. Normalized Frequency with Position as a Parameter	22
11.	Small-signal Response of Distributed RC Network Phase Response in Degrees vs. Normalized Frequency with Position as a Parameter	23
12.	Transient Response of Distributed RC Network. Normalized Gate Voltage vs. Normalized Time with Position as a Parameter. (Computed results in solid line; simulated results in dashed lines.)	25
13.	Computed Response of Silicon Gate MOS Transistor (Load Resistance $\approx$ 0)	27
14.	Computed Response of Silicon Gate MOS Transistor (Load Resistance $\approx$ 0)	28
15.	Computed Response of a Loaded Silicon Gate MOS Transistor (Load Resistance = 100 Ohms)	29
16.	Equivalent Circuit of Silicon Gate DMOS Transistor	30
17.	Computer Response of Silicon Gate DMOS Transistor (Load Resistance = 100 Ohms)	32
18.	Inverter Layout for Ring Oscillator	33
19.	Shift Register (a) Schematic Diagram (b) Clock Timing Diagram	35

Page

20. Cascade MOS Stages 38
21. Normalized Propagation Delay, Normalized Area, and Figure of Merit F, vs. Number of Stages for M = 100, K = 2 41
22. Normalized Propagation Delay and Area-Propagation Delay Square Product, F, vs. Number of Stages 43
23. Optimum Number of Enlarged Output Stages for Different Load to Node Capacitance Ratios 45
24. Minimum Propagation Delay Obtainable for a Given Area and a Fixed Load to Node Ratio of 100 46
25. Composite Layout of the Shift Register Cell 47
26. Overall Interconnection Mask 48
27. P^- Channel-Stop Diffusion Mask 49
28. N^+ Diffusion Mask 50
29. P Diffusion Mask 51
30. Silicon Etch Mask 52
31. Contact Mask 53
32. Interconnection Mask 54
33. Si Gate DMOS IC Structure 56
34. DMOS Integrated Inverter 59
35. Transistor V-I Characteristic of DMOS Inverter 61
36. Effect of Substrate Bias on Load Device Characteristics 62
37. DMOS Inverter Characteristics in 5V Operation 63
38. DMOS Inverter Characteristics in 3V Operation 64
39. Photomicrograph of DMOS Transistor Showing 2 Gate Contacts 65
40. Ring Oscillator 67
41. Ring Oscillation 68
42. Propagation Time and Power Dissipation of DMOS Inverters 70

EXPERIMENTAL INVESTIGATION OF A DOUBLE-DIFFUSED SHIELDED METAL-OXIDE SEMICONDUCTOR (DMOS) STRUCTURE

By H. C. Lin and J. L. Halsor
Westinghouse Electric Corporation
Advanced Technology Laboratories

INTRODUCTION

Double-diffused MOS (DMOS) transistors are capable of high-speed operation because of the reduced channel length that results when successive n and p-type diffusions are done through the same source window. This effectively results in a channel length that is comparable to the base width of a double-diffused bipolar transistor ($\sim$ 0.5 micrometer instead of several micrometers which is typical of standard MOSFET geometry). In addition, for optimum high-frequency performance, it is desirable to minimize area and parasitic capacitance associated with standard metal gate electrodes. In this work, we developed and applied self-aligned polysilicon gate technology to the DMOS construction in a manner that retains processing simplicity and effectively eliminates parasitic overlap capacitance because of the self-aligning feature.

In this work, we also designed and developed integrated depletion mode load devices with the same dimensions as the DMOS transistors. This ratioless feature results in smaller dimension load devices allowing for higher density integration with no increase in the processing complexity of standard MOS technology.

As a vehicle to test the performance and to verify the anticipated benefits, we designed and fabricated a number of inverters connected as ring oscillators. The propagation time-power dissipation product and process related parameters were measured and evaluated.

In the course of the investigation, a number of other technical advances were also made and published. This report is a description of these advances

and includes (1) details of the process; (2) test data and design details for the DMOS transistor, the load device, the inverter, the ring oscillator, and a shift register with a novel tapered geometry for the output stages; and (3) an analytical treatment of the effect of the distributed silicon gate resistance and capacitance on the speed of DMOS transistors.

SILICON GATE DOUBLE-DIFFUSED MOS TRANSISTORS

The progress of electron devices has been earmarked with ever increasing frequency performance. In the realm of semiconductor amplifying devices, the field has been dominated by bipolar transistors. The field effect transistors have generally been conceded as inferior to the bipolar transistor in high-frequency response.

Let us examine the reason for this inferiority: One figure of merit of a high-frequency electron device is the gain-bandwidth product GB ($=g_m/2\pi C_{in}$), where g_m is the transconductance and C_{in} is the input capacitance. For a bipolar transistor (where the junction capacitance can be made negligible by small geometry)

$$(GB)_{bipolar} = \frac{\mu kT}{\pi W_b^2 q} \tag{1}$$

where μ is the minority carrier mobility, W_b is the base width, k is Boltzmann's constant, T is the absolute temperature, and q is the electronic charge.

For a field effect transistor

$$(GB)_{FET} = \frac{\mu(V_p + V_G)}{2\pi L^2} \tag{2}$$

where L is the channel length, V_p is the pinchoff voltage (voltage where drain current saturates), and V_G is the gate voltage.

In ordinary operation, the magnitude of kT/q is only a fraction of a volt. Hence

$$|V_p + V_G| >> 2\left|\frac{kT}{q}\right| \tag{3}$$

The inequality should favor the field effect transistors for higher gain bandwidth product. However, in state-of-the-art FET's, the channel length L is much larger than the base width of bipolar transistors because the lateral L-dimension is determined by photolithographical techniques, while W_b is controlled in the fractional micrometer range by diffusion techniques. In practice, W_b can be well controlled in the fractional micrometer range while L can only be controlled to one order of magnitude higher. Since the gain bandwidth product has a reciprocal square-law dependence on these dimensions, the net result is that bipolar transistors are faster than field effect transistors by two orders of magnitude.

Another reason for a lower limit to the channel length is the voltage punch-through effect resulting from the widening of the depletion layer when the drain voltage becomes appreciable. This punch-through effect is particularly pronounced for FET's with low background impurity concentration. If means can be devised to reduce L and to grade the background channel substrate doping concentration along the channel to reduce the punch-through effect, the high frequency capabilities for FET's can be significantly extended.

A narrow channel length and a higher punch-through voltage can be achieved by double diffusion in much the same way as the conventional bipolar transistor. Such approaches have been pursued by Tarui[1], Cauge[2], and others. However, the structures described do not have a silicon gate. Nor do they contain an electrostatic shield to prevent field inversion in the areas outside the gates. This work is an attempt to fabricate an integrated circuit using silicon gate double-diffused transistors with a field shield to prevent surface inversion between devices.

INTEGRATED DOUBLE-DIFFUSED MOS STRUCTURE

A cross-section of a double-diffused N-channel MOS structure is shown in figure 1. The substrate can be either p or n-type. The window defining the source region is diffused sequentially with both an n-type dopant and a lower concentration p-type dopant. The window defining the drain region is diffused with only n-type dopant as in conventional transistors. With double diffusion, the channel is the p-type diffused region and the length can be made very short,

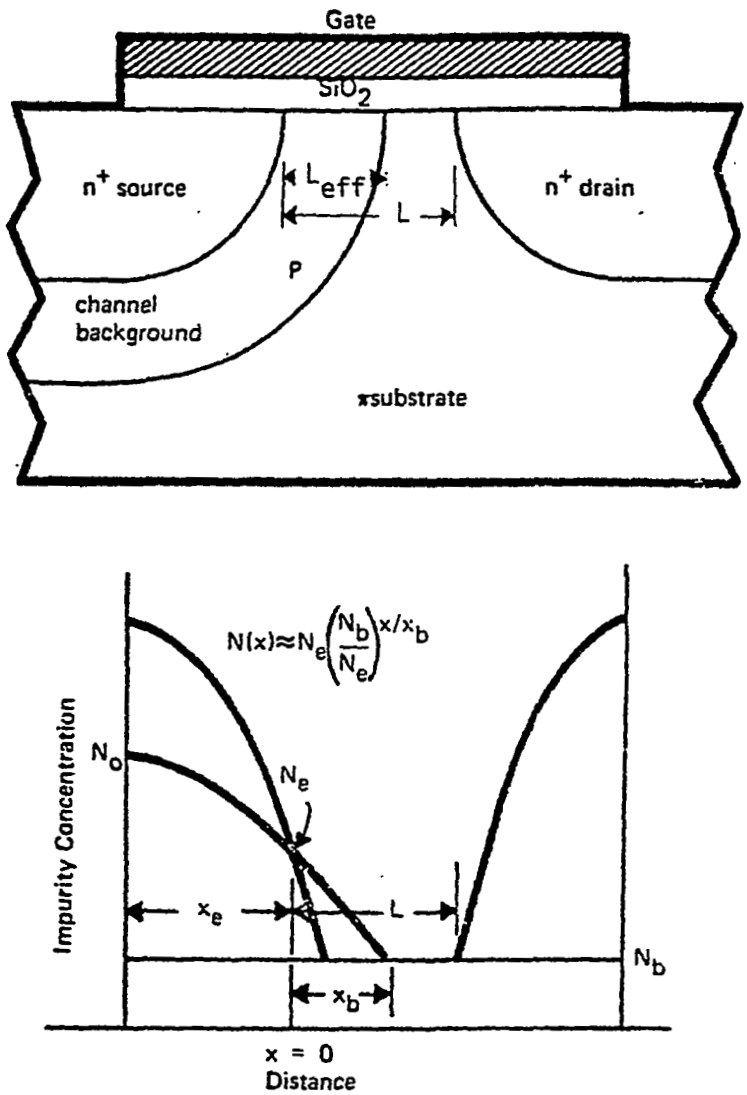


Figure I. DMOS Structure

say less than 1 μm. At the same time, because of the diffusion, the p-type doping is more concentrated toward the source and thus reduces the punch-through effect.

ANALYSIS OF DMOS TRANSISTORS

In our previous work, we analyzed the double-diffused MOS transistor as an idealized device[3,4] in that a zero resistivity metal gate was used and no velocity saturation effect was considered. When the effective channel is very short (in the micrometer or submicrometer range), the high electric field along the channel can saturate the carrier velocity. In our approach of using a silicon gate, the nonzero resistivity of the polycrystalline silicon gate can adversely affect the frequency response of the transistor. A more detailed analysis is made on this effect.

For the analysis of the DMOS transistor, the device can be divided into a number of incremental transistors of variable threshold voltage in much the same way an an electrical transmission line. The basic equation we used in our previous work is:

$$I_D = \frac{\mu C_G}{L} [V_G - V_T(x) - V(x)] \frac{dV}{dx} \tag{4}$$

where μ = mobility

C_G = gate capacitance

L = total channel length

V_G = gate voltage (relative to substrate)

$V_T(x)$ = threshold at a distance x from the edge of source

$V(x)$ = voltage of the channel at any point x

The results of our computation, which were reported in our previous report[4] as well as an IEEE publication[3], were arrived at by considering a constant mobility. If the mobilities are linearly dependent on electric field, the current is less than expected at high fields. For a conventional MOS transistor, the current tends to depart from the square-law relationship and become more linearized. Such an effect is taken into account in certain models[5] and

incorporated in some computer-aided design programs.*

The characteristic of a DMOS transistor can be computed by connecting a number of incremental transistors of different threshold voltages in series. The number of divisions depends on the accuracy desired. If too many divisions are used, it is uneconomical from the standpoint of computation time. We found that five divisions were sufficient to give a reasonable accuracy.

The computer-aided design program used is M-SINC.* The model of an MOS transistor used in this program should contain the following information

NAME	PARAMETER	DEFAULT CONDITION
VTO	Threshold Voltage with Source Grounded	0
UB	Low Field Mobility Value	450
ECRIT	Critical Field	6.E4
C1	Empirical Matching Coefficient for Mobility	0
COX	Oxide Capacitance per sq. cm.	3E-8
DNB	Doping Concentration in Substrate	1E15
XJD(XJS)	Overlap of Drain Length Under Gate (Source)	0
GOS	Choice of Depletion Equation	0
	0 - No Channel Length Depletion	
	1 - Square Root Dependence Thick Oxide Device	
	2 - Fringing Field Thin Oxide Device	
$\alpha 1$	First Fringing Field Coefficient	0.2
$\alpha 2$	Second Fringing Field Coefficient	0.6
CSS	Saturation Current of Junction Diode	1E-14
TEMP	Parameter Definition Temperature	300

The electric field is considered in the term

$$\mu_{eff} = \mu_{eff}(0) \quad \frac{E_s}{E_{so}}$$

where E_s is the electric field and E_{so} is the critical electric field above which the mobility is affected. If the electric field effect is neglected,

*M-SINC Computer-Aided Analysis Program developed by Stanford University.

the exponent C_1 is set at zero. For practical MOS transistors

$C_1 = 0.15$ for p-channel devices

$C_1 = 0.35$ for n-channel devices.

The results of the computation are shown in figure 2, and the results for no velocity saturation ($C_1 = 0$) are shown in figure 3. Generally, the effect of velocity saturation is to reduce the drain current. The reduction is greater at higher gate voltages. This effect also makes the transconductance constant as compared to conventional MOS transistors whose transconductance increases linearly with increasing gate voltage.

In our previous analysis,[3] it was pointed out that the DMOS transistor itself has a tendency toward constant transconductance even without considering carrier velocity saturation. This will be brought about by the following argument. The DMOS transistor can be considered to be an enhancement mode short channel transistor in series with a depletion mode transistor with a longer channel. The channel resistance is the series resistance of the two transistors and is then dominated by the higher of the two resistances. At low gate voltages, the enhancement mode transistor with high transconductance dominates the characteristics. At high gate voltages, the enhancement mode transistor has lower resistance, and the lower transconductance of the depletion mode transistor dominates the characteristics. The net result is to make the composite transconductance nearly constant. Thus, the DMOS transistor has two effects tending to lower the transconductance at high gate voltage and to make it more nearly constant.

The computer results of a transistor with uniform background is shown in figure 4. Note that at low gate voltages, the transconductance of a DMOS transistor is much greater than that of a conventional uniform background MOS transistor. However, at high gate voltages the superiority becomes less pronounced because of the linearization of the DMOS characteristics. For this reason, it is most advantageous to use the DMOS integrated circuit for low-voltage operation.

DMOS INVERTER

The DMOS transistor can be used in conjunction with a depletion-mode

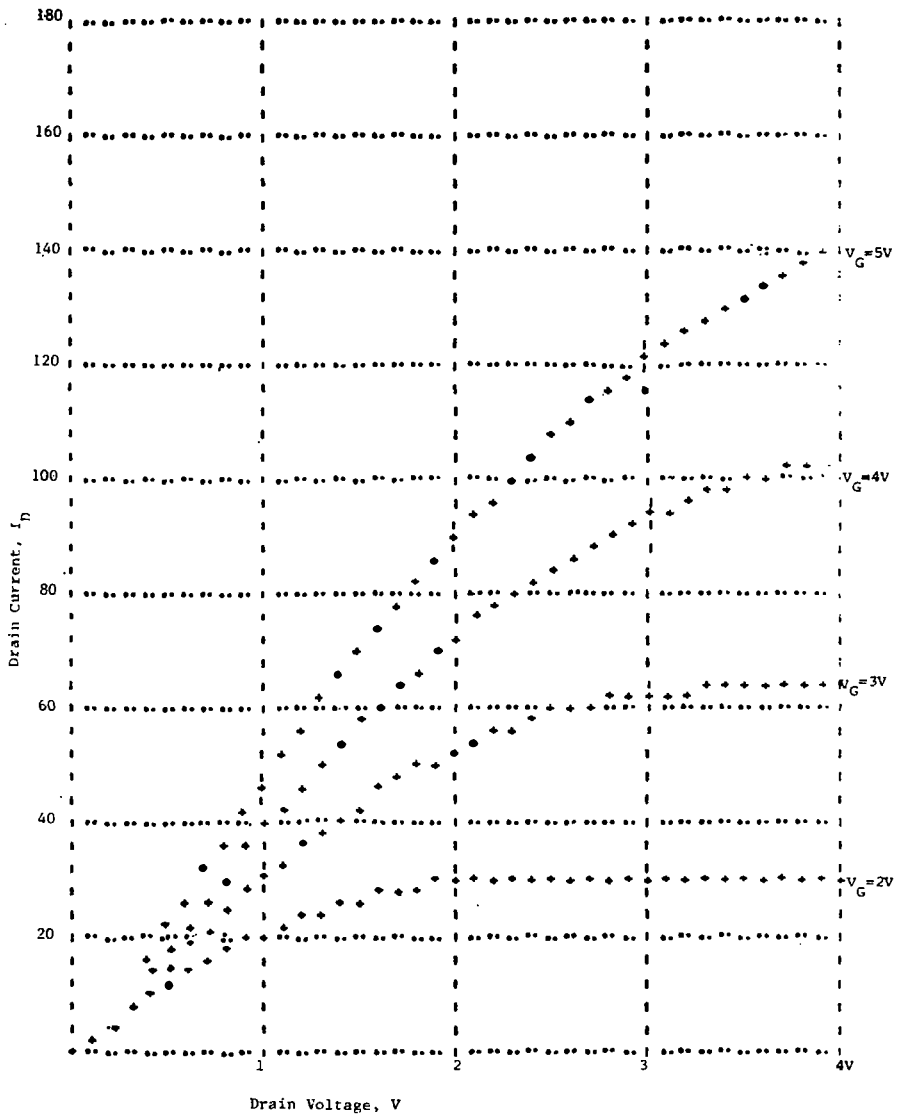


Figure 2. Computed Drain Output DMOS Characteristics Velocity Saturating

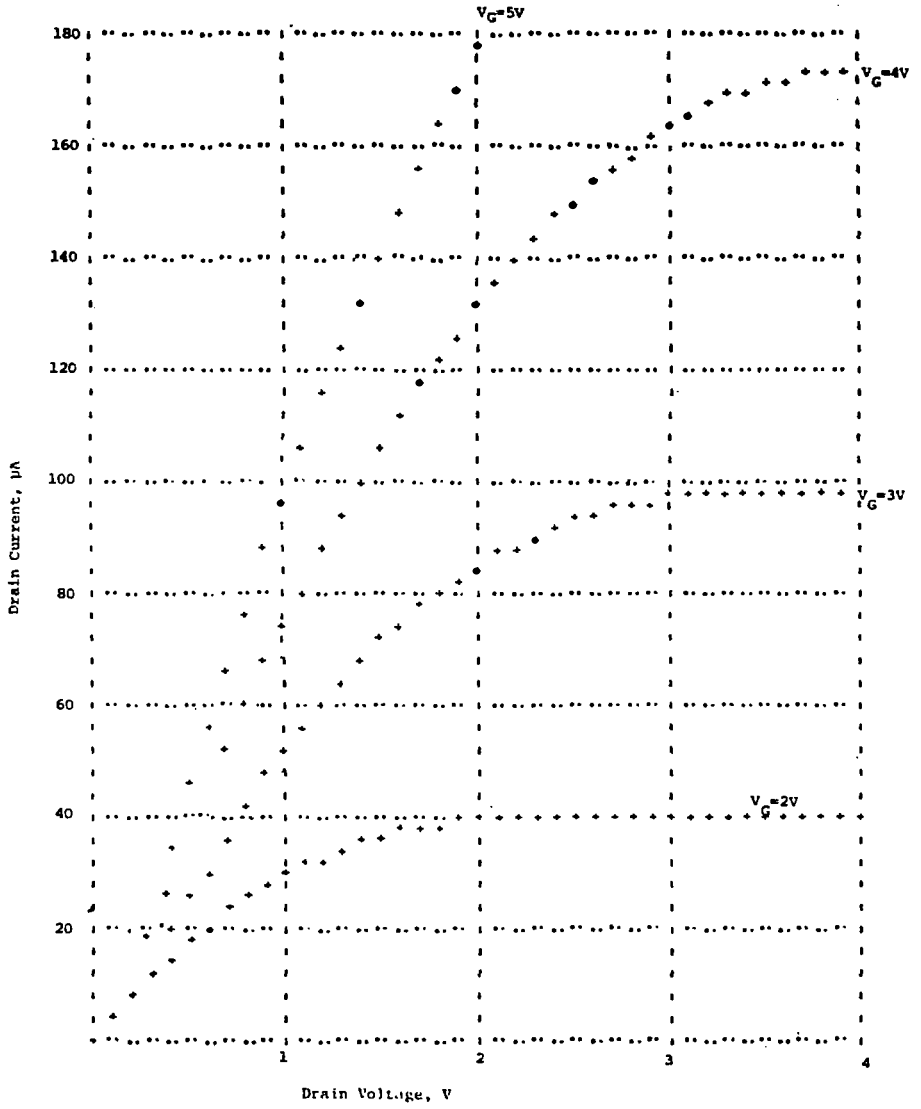


Figure 3. Computed Drain Output DMOS Characteristics
Velocity not Saturating

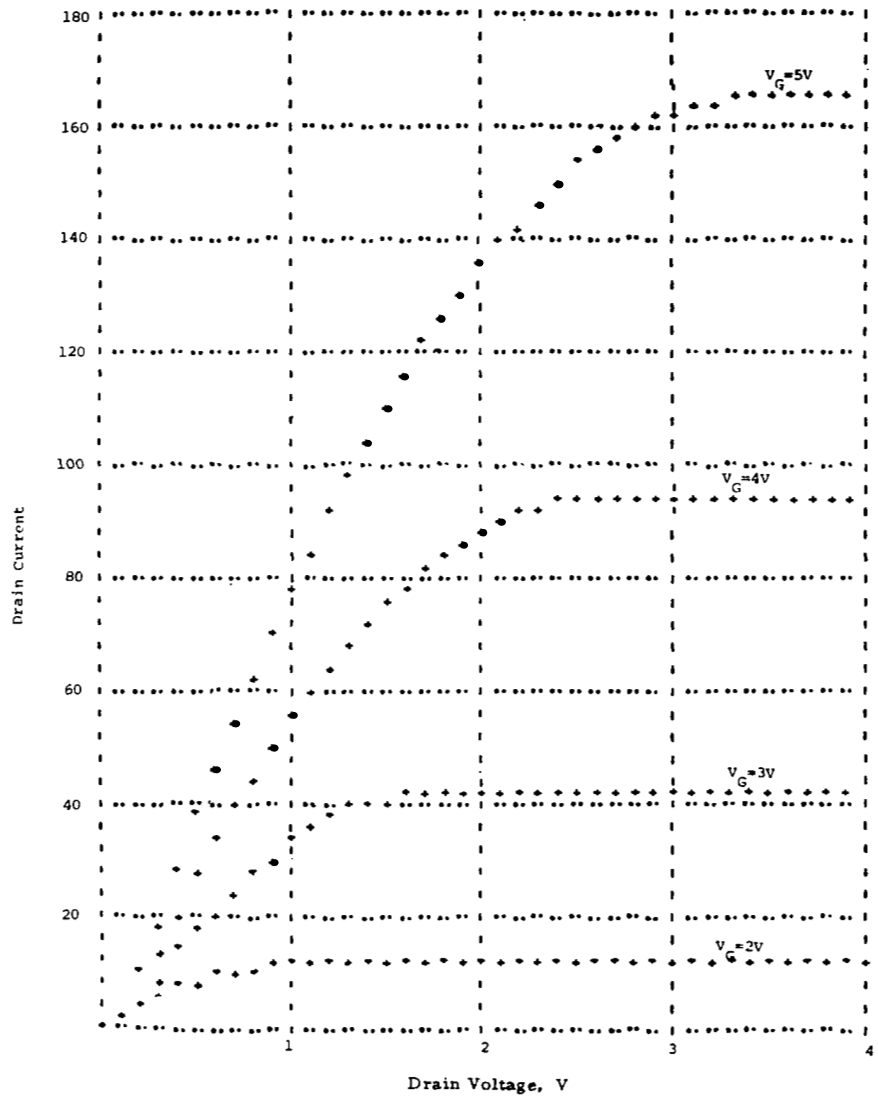


Figure 4. Computed Drain Output MOS Characteristics

transistor as an inverter as shown in figure 5. The depletion mode transistor is then used as a load device. The advantages of a depletion-mode load are well known.[6] It increases the speed and lowers the supply voltage. In combination with the speed of an active DMOS transistor, a very low-speed power product can be expected. In conventional integrated circuits, a depletion mode load requires an extra diffusion or ion implantation to obtain the proper background impurity concentration while the original substrate provides the background to obtain the proper threshold voltage. In our DMOS integrated circuit, no such extra processing step is required because the substrate is chosen to yield a depletion-mode load device and the threshold voltage for the enhancement-mode DMOS transistor is already provided by the double-diffusion. Besides this advantage, the load device can be made to have the same length-to-width aspect ratio as the active DMOS transistor. This ratioless geometry results in further reducing the area of an integrated circuit.

Switching Transistor Design Considerations

The choice of surface and background concentration is determined by the threshold voltage and the breakdown voltage. The threshold voltage is a function of surface state density and background concentration. For $\langle 1,0,0 \rangle$ oriented crystals, the surface state density is typically $2.5 \times 10^{11}/cm^2$. A threshold voltage of 2V requires a concentration of 2.5×10^{16} atoms/cm^2 as calculated from equation (2).

$$V_T(x) = \phi_{GS} - \frac{qN_{SS}}{C_{ox}} + 2\phi_F + \frac{\sqrt{2\epsilon_{si} qn(x) \left|2\phi_F + V(x)\right|}}{C_{ox}} \qquad (5)$$

where the following terminology is used.

ϕ_{GS} = Gate-background work function difference = $\frac{kT}{q} \ln \frac{4 \times 10^{20} N(x)}{n_i^2}$,
(for an n-doped Si gate with an assumed effective doping concentration of 4×10^{20} atoms/cm^3).

n_i = Intrinsic concentration = 1.5×10^{10} atoms/cm^3 at 27°C

N_{SS} = Surface-state density

ϕ_F = Fermi level associated with a given doping concentration

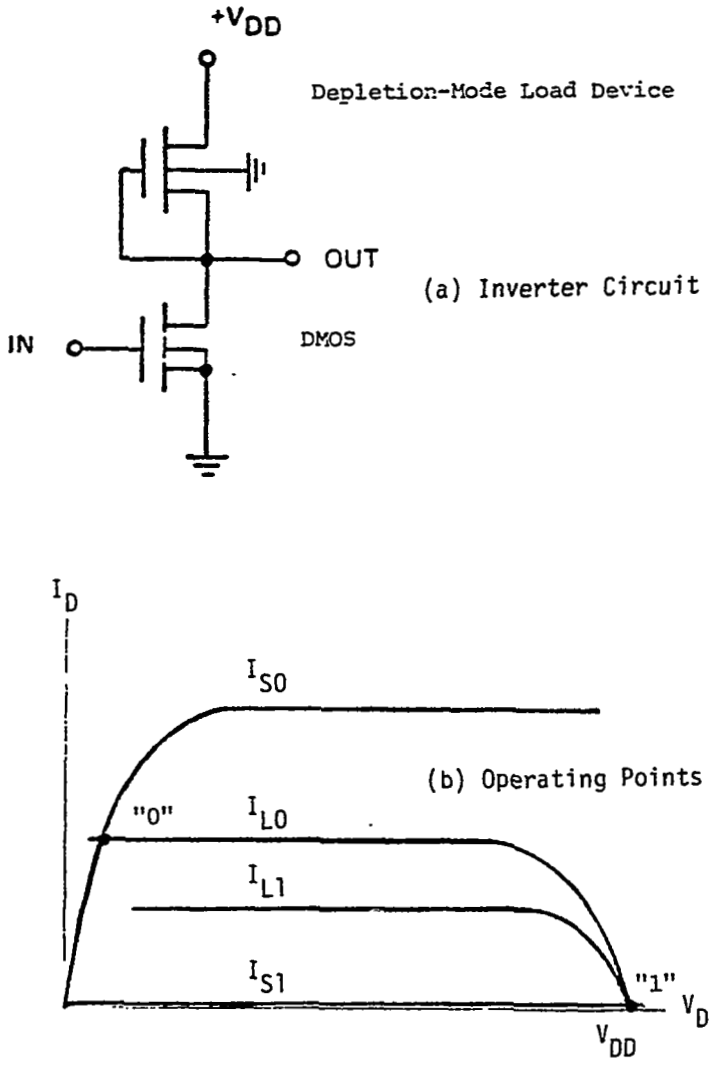


Figure 5. DMOS Inverter

$$= \frac{kT}{q} \ln \frac{N(x)}{n_i}.$$

C_{ox} = Gate capacitance per unit area ε_{ox}/t_{ox}.

ε_{ox} = Dielectric constant of oxide = $4 \times 8.85 \times 10^{-14}$ f/cm.

t_{ox} = Gate oxide thickness.

ε_{si} = Dielectric constant of silicon = ($11.7 \times 8.85 \times 10^{-14}$ f/cm).

The concentration n(x) for determining the threshold voltage is the value at the n + p junction formed by double diffusion and can be obtained by proper choice of surface concentration and junction depth as indicated in equation (5).

In our previous work,[3] it was shown that for the same threshold voltage, the ⟨1,1,1⟩ crystal orientation gives a higher gain-bandwidth product than the ⟨1,0,0⟩ crystal orientation. The reason is that ⟨1,1,1⟩ orientation has higher surface states and hence make the channel more conductive near the drain. This higher conductance accounts for the higher transconductance. The penalty for the higher gain-bandwidth product is the lower breakdown voltage at the isolation junction, as shown in table I, because the surface concentration of the p-type isolation layer must be higher (at least as high as the concentration at the drain channel junction) to avoid surface inversion. However, this is not a problem when we use low voltages.

Table I

TRANSISTOR DESIGN PARAMETERS

	⟨1,0,0⟩Si(N_{SS}=2.5x10^{11}/cm^2)			⟨1,1,1⟩(N_{SS} = 10^{12}/cm^2)		
p-type surface concentration	3.3x10^{17}	1.5x10^{17}	8.85x10^{16}	6.3x10^{17}	3.6x10^{17}	2.1x10^{17}
p-type background concentration	10^{15} atoms/cm^3			10^{15} atoms/cm^3		
p-type junction depth	3 μm	2.5 μm	2 μm	3 μm	2.5 μm	1 μm
n$^+$ type junction depth	2 μm	1.5 μm	1 μm	2 μm	1.5 μm	1 μm
Surface concentration at source junction	2.5x10^{16} atoms/cm^3			1x10^{17} atoms/cm^3		
Breakdown voltage	28V	26V	24V	13V	12V	11V
Threshold voltage	0.78V			0.78V		

Load Device Design Considerations

The load is an n-channel depletion mode device. The characteristic of the load device is dependent upon the substrate impurity concentration and the crystal orientation. For most economical design, it is desirable that the geometry of both the load device and the switch be minimum. The saturated drain current of the load device for the "high" output $(I_{DSAT})_{L1}$ and for the "low" output $(I_{DSAT})_{LO}$ should satisfy the conditions (see figure 5):

$$(I_{DSAT})_{L1} > (I_{DSAT})_{S1} \simeq 0 \tag{6}$$

$$(I_{DSAT})_{LO} < (I_{DSAT})_{SO} \tag{7}$$

where $(I_{DSAT})_{S1}$ and $(I_{DSAT})_{SO}$ are the saturated drain currents of the active switching device for the high and low output respectively. To satisfy equation (6), the load device must be in depletion mode; i.e., the threshold voltage must be negative. It should be kept in mind that the threshold voltage for the load device is different from the zero-substrate bias value because of the "body effect". The "body effect" is due to the reverse bias on the substrate with respect to the source. As shown in figure 4, the output terminal or the source of the load device is reverse biased with respect to the substrate when the output voltage is high. Under this condition, the threshold voltage or the negative of the pinchoff voltage V_P is given as

$$V_{TL} = \phi_{GS} - q\frac{N_{SS}}{C_{ox}} + 2\phi_F + \frac{\sqrt{2\varepsilon_s q N_s (2\phi_F + V_R)}}{C_{ox}} = -V_P \tag{8}$$

where V_R is the reverse bias of the substrate relative to the source (the output terminal).

The body effect is different for different output conditions. If the source of the switching transistor is grounded and the output is at the low state (or the switch is on), the source of the load device is near the ground and is not reverse biased with respect to the substrate. When the output is at the high state, the source of the load device is reverse biased with respect to the substrate, and the pinchoff voltage is decreased. Thus, in equation (6), the $(I_{DSAT})_{L1}$ should be calculated with substrate reverse

biased by approximately V_{DD} volts; $(I_{DSAT})_{LO}$ should be calculated with a substrate bias nearly equal to zero in equation (7). If the substrate is not the same potential as the source of the load device (i.e., not at "0 level") but is reverse biased by V_R, then the threshold voltage in equation (8) should be increased correspondingly. If we choose V_R = 5V, which is the supply voltage and

$$2\phi_F = \frac{kT}{q} \ln \frac{4 \times 10^{20} N_\pi}{n_i^2} \tag{9}$$

V_{TL} should be negative if the load device is to be in depletion mode. The saturated drain current of the DMOS can be represented by

$$I_{DS} = \frac{\mu C_{ox}}{2L_{eff}^2} (V_G - V_T)^2 \tag{10}$$

where L_{eff} is the effective channel length of a DMOS transistor which should be considerably shorter than the geometrical length. The saturated drain current of the load device is

$$I_{DL} = \frac{\mu C_{ox}}{2L^2} (V_P)^2 \tag{11}$$

The value of V_P is also given in equation (8). Since the effective channel length, L_{eff}, of the DMOS is always less than the geometrical length, L, the condition that $I_{DS} > I_{DL}$ can be satisfied if

$$V_G - V_T > V_P \geq 0 \tag{12}$$

Figure 6 is a plot of the threshold voltage (or the pinchoff voltage) as a function of the substrate concentration with different substrate bias V_P and surface state densities N_{ss} based on equation (8). The surface states range from $10^{11}/cm^2$ to $10^{12}/cm^2$, which are representative of practical oxidized silicon surfaces. The reverse biases were 0, 3, and 5V. The 0V bias curves

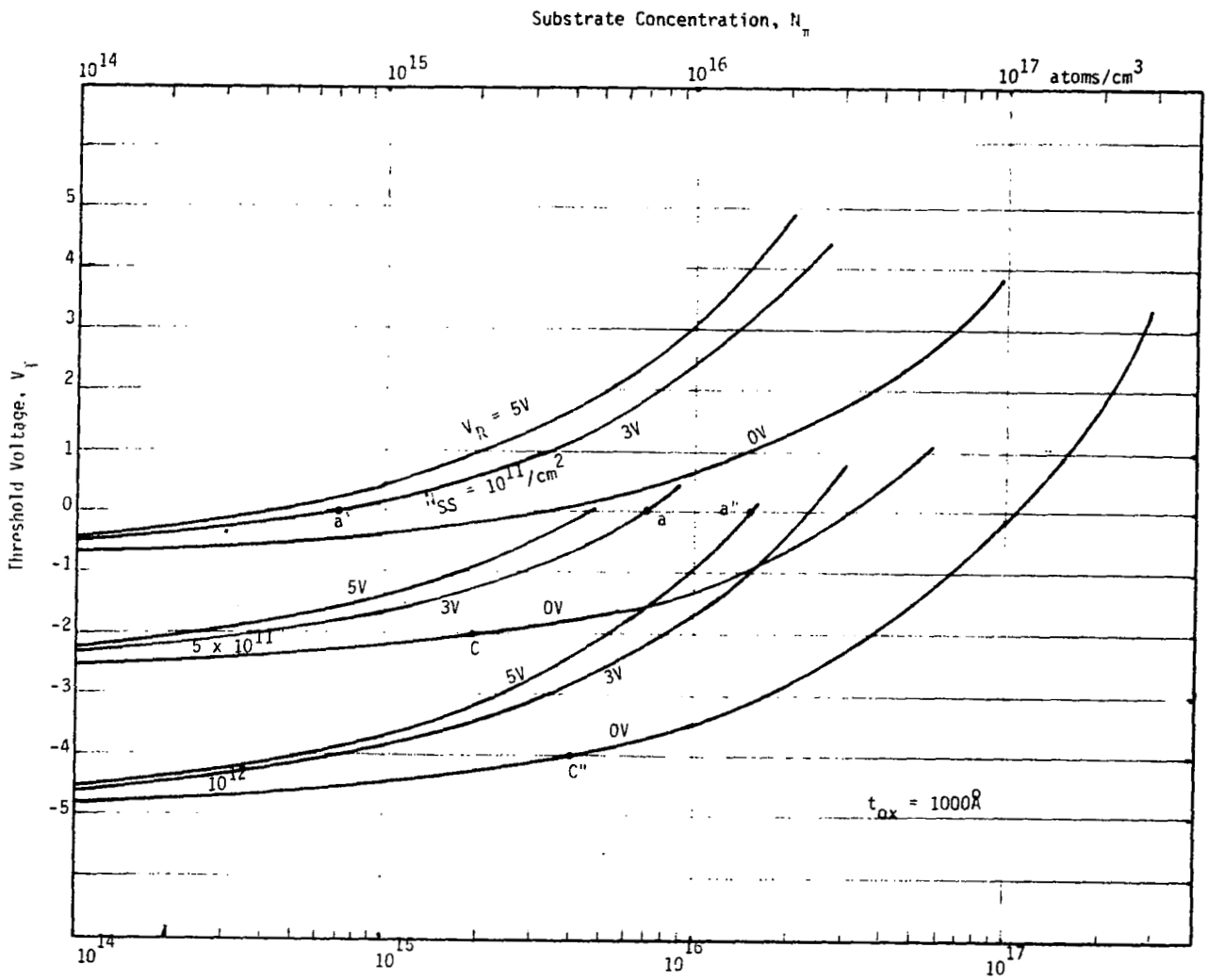


Figure 6. Threshold Voltage as Function of Substrate Concentration

are useful for designing the DMOS switch; the 3 and 5V curves are useful for the load device corresponding to two different supply voltages. For the load device to be in depletion mode, the threshold voltage must be negative with respect to a substrate bias approximately equal to the drain supply voltage. At the same time, the magnitude of the pinchoff voltage at zero bias must be less than $|V_G - V_T|$ to satisfy equation (12) keeping in mind that V_G can at most rise up to V_{DD}. For instance, from figure 6, if $N_{ss} = 5 \times 10^{11}/cm^2$ for $\langle 1,1,1 \rangle$ orientation, V_T of the switching device = 1V and $V_{DD} = 3V$, to satisfy equations (6) and (12) the substrate concentration N_π must be 2.2×10^{15} (point c) $< N_\pi < 7 \times 10^{15}$ (point a) atoms/cm^2. On the other hand, if we choose $N_{ss} = 10^{11}/cm^2$ for $\langle 100 \rangle$ orientation, then $N_\pi < 7 \times 10^{14}$ atoms/cm^3 (point a') which corresponds to a resistivity of 20 ohm-cm which is not easily controlled to a tight tolerance. This reasoning favors a moderate surface state density; i.e., a higher N_{ss} $\langle 1,1,1 \rangle$ substrate is more suitable than a lower N_{ss} $\langle 1,0,0 \rangle$ substrate.

At the other extreme, if one increases $N_{ss} = 10^{12}/cm^2$, then there is no N_π which can satisfy a 3V operation. Thus $V_{DD} = 5V$ must be increased to give 4×10^{15} (point c") $< N_\pi < 1.5 \times 10^{16}$ (point a") atoms/cm^3.

From these considerations, we choose a p type $\langle 1,1,1 \rangle$ substrate with

$$\rho_\pi = 2 \text{ to } 5 \text{ ohm-cm}$$
$$N_{ss} = 5 \times 10^{11}/cm^2$$
$$V_T = 1V$$

EFFECT OF SILICON GATE RESISTANCE ON THE FREQUENCY RESPONSE OF MOS TRANSISTORS[7]

Self-aligned silicon-gate MOS transistors have smaller dimensions and, therefore, are supposed to have a higher frequency response owing to the high g_m and the low gate-to-drain feedback capacitance. However, a silicon gate has much higher resistivity than a metallic gate. The nonzero resistivity can have an adverse effect on the frequency response. This effect will be explained in the following discussion.

Silicon Gate as Distributed RC Network

Figure 7 shows the structure of a polycrystalline silicon gate MOS transistor. For high transconductance and gain bandwidth product, a small gate length L (typical $< 10\mu m$) should be used. This dimension is often limited by the reproducible photoengraving capability of the process line used. The same photoengraving capability also limits the dimensions of the contact windows for the gate. If the gate contact were placed above the active area, it might short-circuit the gate to the source or the drain. This problem may be avoided in practice by not locating the gate contact directly above the active gate area but to the end where the polycrystalline silicon area may be conveniently enlarged as shown in figure 7.

Polysilicon gate material is not even approximately metallic and hence has nonzero resistivity. For a typical heavily doped, few thousand Angstrom thick silicon gate, the sheet resistivity may be in the order of tens of ohms per square and the series resistance to a far removed portion of a gate in the order of hundreds of ohms.

There are parasitic capacitances associated with the gate such as the gate-to-source capacitance, the gate-to-substrate capacitance, and the gate-to-drain capacitance. The silicon gate can then be considered as a uniformly distributed RC transmission line[8] as shown in figure 8. The distributed transistor can be considered as a number of lumped transistors with common drains and sources as shown in figure 9, but with their gates tapped from the distributed RC network. Because of their time delay along the distributed network, the effective gate voltages of each lumped transistor, V_g, are progressively delayed from the input signal, V_i. This effect is particularly noticeable toward the end of the network. The drain currents, since they are a monotonic function of gate voltages, are therefore progressively delayed. The total drain current is the summation of these time delayed from the input signal. For one lumped section of the RC distributed network, the small-signal voltage gain A_v is given as

$$A_v = \frac{g_m R_L}{1 + j\omega R_g C} \tag{13}$$

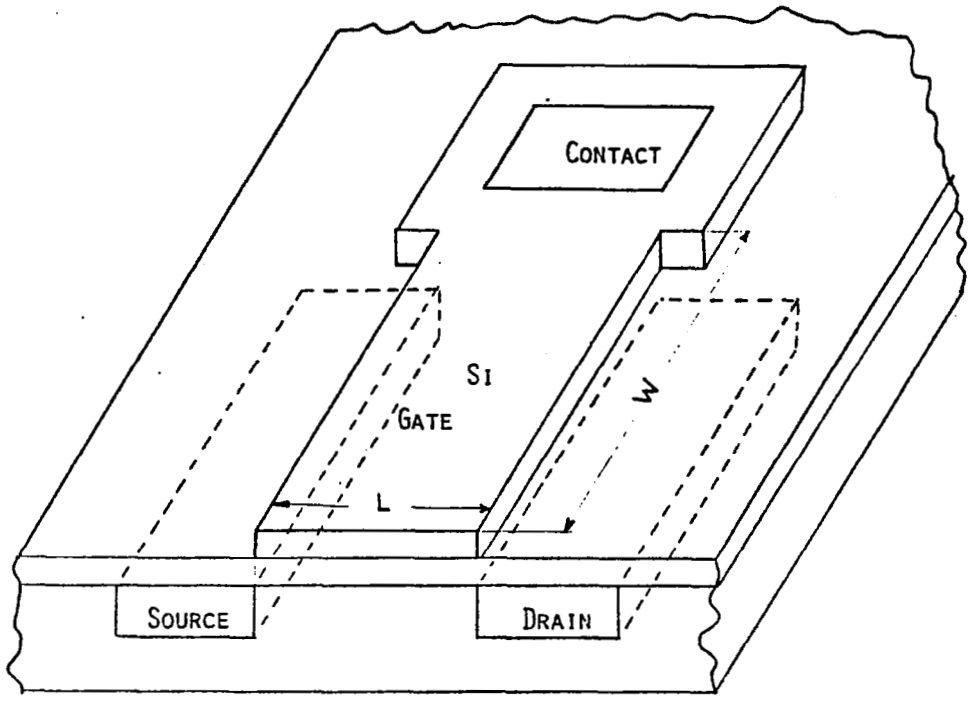


Figure 7. Silicon Gate MOS Transistor

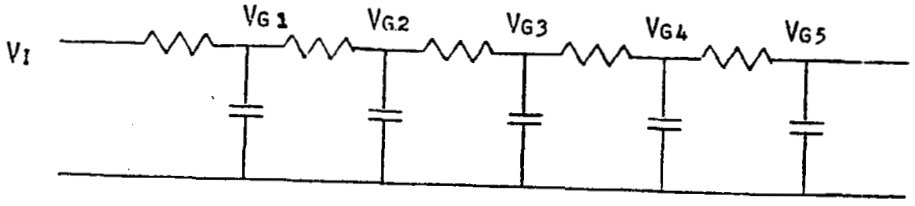


Figure 8. Distributed RC Network

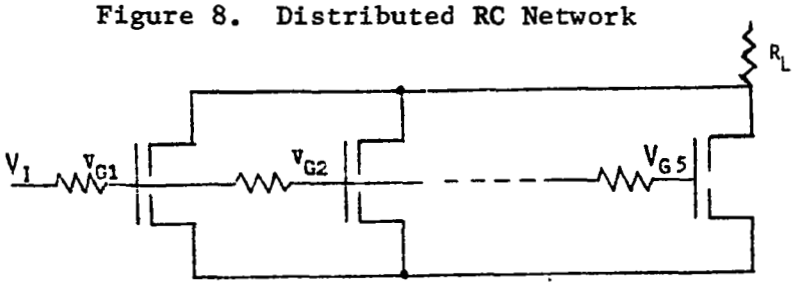


Figure 9. Equivalent Circuit of a Silicon Gate MOS Transistor

where g_m is the transconductance, R_g ($= \Sigma r_g$) is the series resistance, C ($= \Sigma c$) is the input capacitance, and R_L is the load resistance. Any increase in R_g or C will reduce the voltage gain.

Frequency Response of the Distributed RC Network

The distributed RC network can be analyzed both for small-signal response and transient response. For small signal analysis, a one-dimensional distributed RC network is a good approximation because the enlarged metal contact area has very small sheet resistance and is sufficiently far from the source and drain diffusions compared to the oxide thickness such that the fringing capacitance is negligible. The mathematics is straightforward and is derived in the Appendix.

For small signal analysis, the voltage $V_g(x)$, from figure 8, along the width W of the distributed RC network is derived in the Appendix as equation (A10) as

$$V_g(x) = V_i \frac{\cosh \sqrt{j\omega rc}\, W(1 - x/W)}{\cosh \sqrt{j\omega rc}\, W} \tag{14}$$

where r and c are the resistance and capacitance respectively per unit length in the x direction. With manipulation, the amplitude and phase responses of this expression can be found to be

$$\left|\frac{V_g}{V_i}\right| = \left[\frac{\cosh^2 \sqrt{\omega rc/2}W(1 - x/W) + \sin^2 \sqrt{\omega rc/2}\, W(1 - x/W)}{\cosh^2 \sqrt{\omega rc/2}\, W + \sin^2 \sqrt{\omega rc/2}\, W}\right]^{1/2} \tag{15}$$

$$\angle V_g/V_i = \arctan\left[\tanh \sqrt{\omega rc/2}\, W \tan \sqrt{\omega rc/2} W\right] - \arctan\left[\tanh \sqrt{\omega rc/2}\, W \tan \sqrt{\omega rc/2}\, W\,(1 - x/W)\right] \tag{16}$$

These responses are plotted in figures 10 and 11. It can be seen that there is a considerable amount of attenuation and phase shift when the time constant, RC, of the total resistance $R = Wr$ and total capacitance $C = Wc$, is of the order of $1/\omega$.

The transient response of a distributed RC network, as derived in the

Appendix as equation (A16) is given as

$$\frac{V_g(x,t)}{V_i} = 1 + \sum_{n=1}^{\infty} \frac{4(-1)^n}{(2n-1)\pi} \cos\left[\pi\left(\frac{2n-1}{2}\right)\left(1 - \frac{x}{W}\right)\right] \exp\left[-\left(\frac{2n-1}{2}\right)\pi \frac{t}{W^2 rc}\right] \quad (17)$$

This equation is plotted as the solid curves in figure 12.

It can be seen that the signal is delayed along the line. At a normalized time of $t/W^2rc = 1$, the signal does not reach 90 percent of the final value at any length greater than 20 percent of the total length.

Time Constant of the Silicon Gate

The time constant of the silicon gate is a function of the sheet resistivity, ρ_s, and the gate oxide thickness. The resistance per unit length is given as

$$r = \rho_s / L \quad (18)$$

where ρ_s is the sheet resistivity and L is the source-to-drain distance of the silicon gate. The capacitance per unit length is given as

$$c = \frac{\varepsilon_{ox}}{t_{ox}} L \quad (19)$$

where ε_{ox} is the dielectric constant of the oxide and t_{ox} is the oxide thickness. The time constant $RC = W^2rc$ is then obtained from equations (18) and (19).

$$RC = \frac{W^2 \varepsilon_{ox} \rho_s}{t_{ox}} \quad (20)$$

For a typical MOS transistor,

$$\varepsilon_{ox} = 4 \times 8.85 \times 10^{-14} = 3.5 \times 10^{-13} \text{ F/cm}$$

$$\rho_s = 28 \ \Omega/\text{square}$$

$$t_{ox} = 1000\text{Å} = 10^{-5} \text{ cm}$$

$$rc = 10^{-6} \text{ sec/cm}^2$$

$$RC = W^2 rc = W^2 \times 10^{-6} \text{ sec.}$$

Thus, for a gate of 0.1 cm width, the time constant is then 10 ns. This order of magnitude of gate width is quite common for discrete MOS transistors and the output amplifier stages of an MOS integrated circuit. For low-power

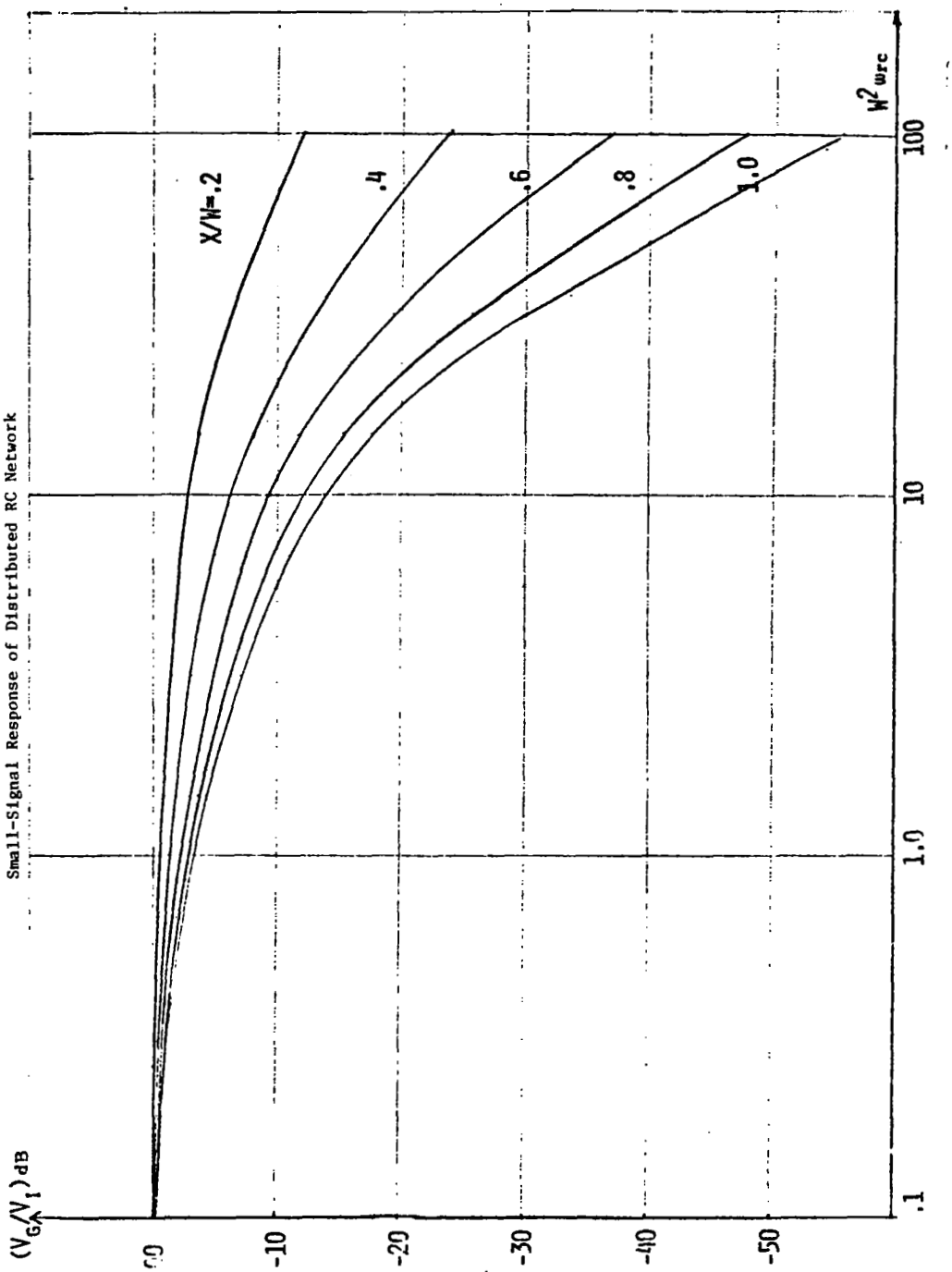


Figure 10. Small-signal Response of Distributed RC Network Amplitude Response in dB vs. Normalized Frequency with Position as a Parameter

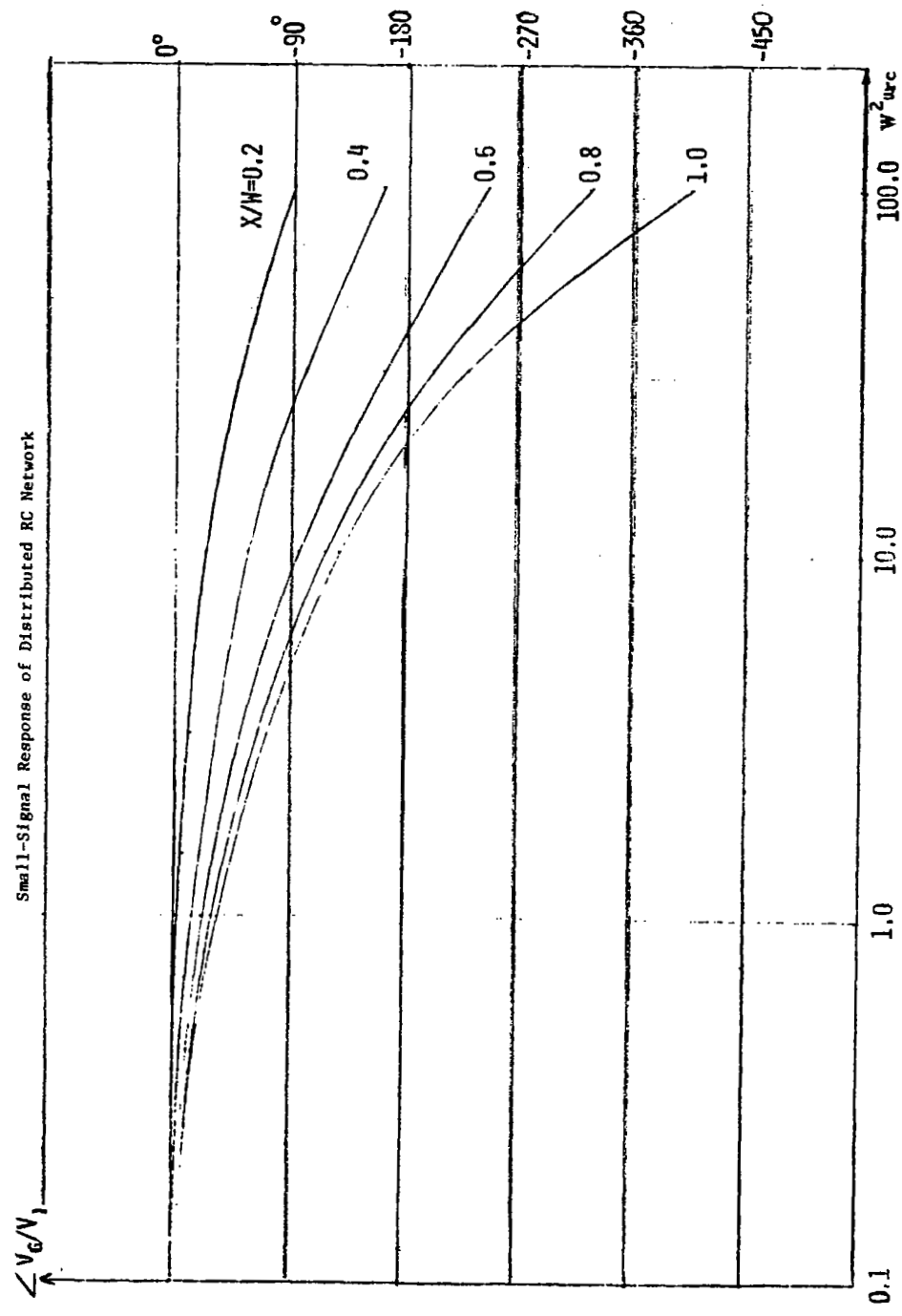


Figure 11. Small-signal Response of Distributed RC Network Phase Response in Degrees vs. Normalized Frequency with Position as a Parameter

internal stages on an MOS integrated circuit, the gate widths are typically one or two orders of magnitude narrower. The time constant, RC, which is proportional to W^2, is then in the subnanosecond or picosecond range.

Characteristics of the Silicon Gate Uniform Channel MOS Transistor

The characteristic of a MOS silicon gate distributed transistor can be computed from the equivalent lumped circuit with a number of transistors connected in parallel but with delayed signals applied to the gates as shown in figure 9. The characteristics of each individual transistor is given by the relationship

$$I_D = \frac{\mu C_G}{L^2}\left[(V_g - V_T)\, V_D - \frac{V_D^2}{2}\right] \text{ for } V_D < V_g - V_T \text{ (triode operation)} \tag{21}$$

$$I_D = \frac{\mu C_G}{2L^2}(V_g - V_T)^2 \text{ for } V_D > V_g - V_T \quad \text{(pentode operation)} \tag{22}$$

where μ is the carrier mobility, C_G is the total gate capacitance, V_T is the threshold voltage, and V_D is the drain voltage. In the present case, V_g is a function of time and distance from the contact as given by equations (11) and (17).

The computation can proceed by applying an input voltage at the gate contact in the equivalent circuit shown in figure 9 with given values of distributed gate resistance and capacitance. The resultant drain current can then be readily computed using a computer-aided-design program. The SPICE program* was used to simulate this relationship. For computational simplicity and reasonable accuracy, the transistor computed is divided into five increments. The parameter of each increment is assumed as follows:

$$L = 10\ \mu m \qquad W = 60\ \mu m$$

*Developed by the University of California, Berkeley.

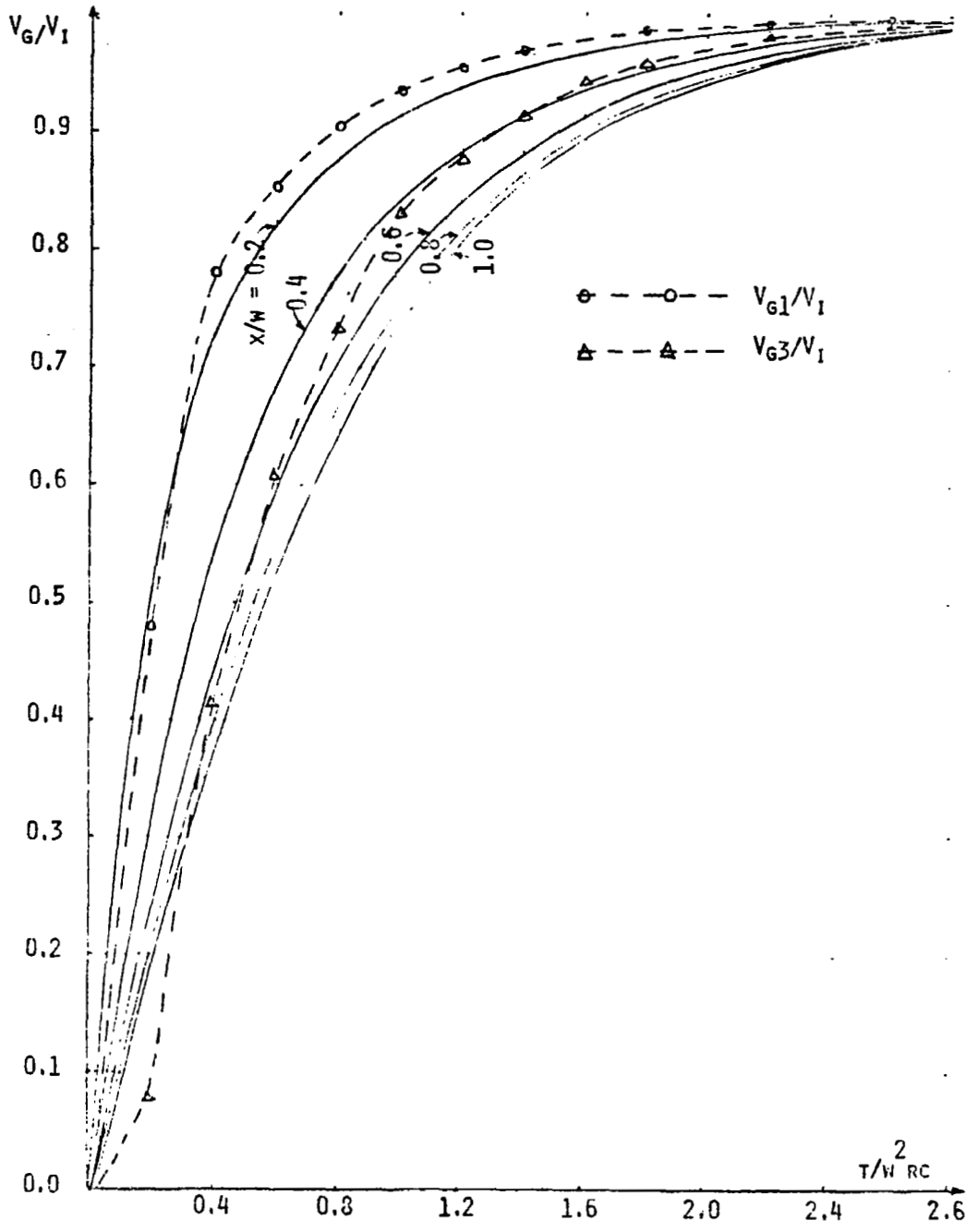


Figure 12. Transient Response of Distributed RC Network. Normalized Gate Voltage vs. Normalized Time with Position as a Parameter. (Computed results in solid line; simulated results in dashed lines.)

The silicon gate sheet resistivity = 25 ohms/square. A very small load resistance, 0.01 ohm, was used to sense the current. The transient response of the drain current with a 6-volt input pulse of nanosecond risetime and 1-nanosecond falltime, is shown in figure 13. For comparison, a zero sheet resistance aluminum gate transistor of the same geometry is also computed and the transient response is plotted in figure 14. The results in Table II were obtained.

Table II

Transient Response Simulation Results

	Si Gate	Al Gate
Sheet resistivity	25 ohms square	0
Risetime	4 nsec	0.5 nsec
Falltime	2 nsec	1 nsec

It can be seen that the nonzero sheet resistivity of the silicon gate very markedly slows down the transient response.

As a check, the gate voltages at the different incremental transistors were also computed. Two of these results are plotted in figure 12 as dashed lines to show general agreement.

The nonideal effect associated with the silicon gate is aggravated by the Miller effect. If there is gate-to-drain overlap capacitance, C_{gd}, present and the load resistance R_L is appreciable, the input capacitance is effectively increased to $C_{gd}(1 + g_m R_L)$. Thus, the time constant of the distributed RC gate is also increased.

To verify this prediction by simulation, the load resistance used in the computation (as shown in figure 9) was increased from 0.01 to 100 ohms. This computed transient response is then plotted in figure 15. Note that the risetime and falltime are lengthened as compared to that shown in figure 12 and 13, where no Miller effect is present.

Characteristics of DMOS Transistors

The double-diffused MOS (DMOS) transistor is designed for high-frequency operation by virtue of its much narrower effective channel length. The computation of its characteristics was described by Lin et al.[3] Because the background concentration between the source and the drain is graded, the threshold voltages vary along the channel. The DMOS distributed characteristics can

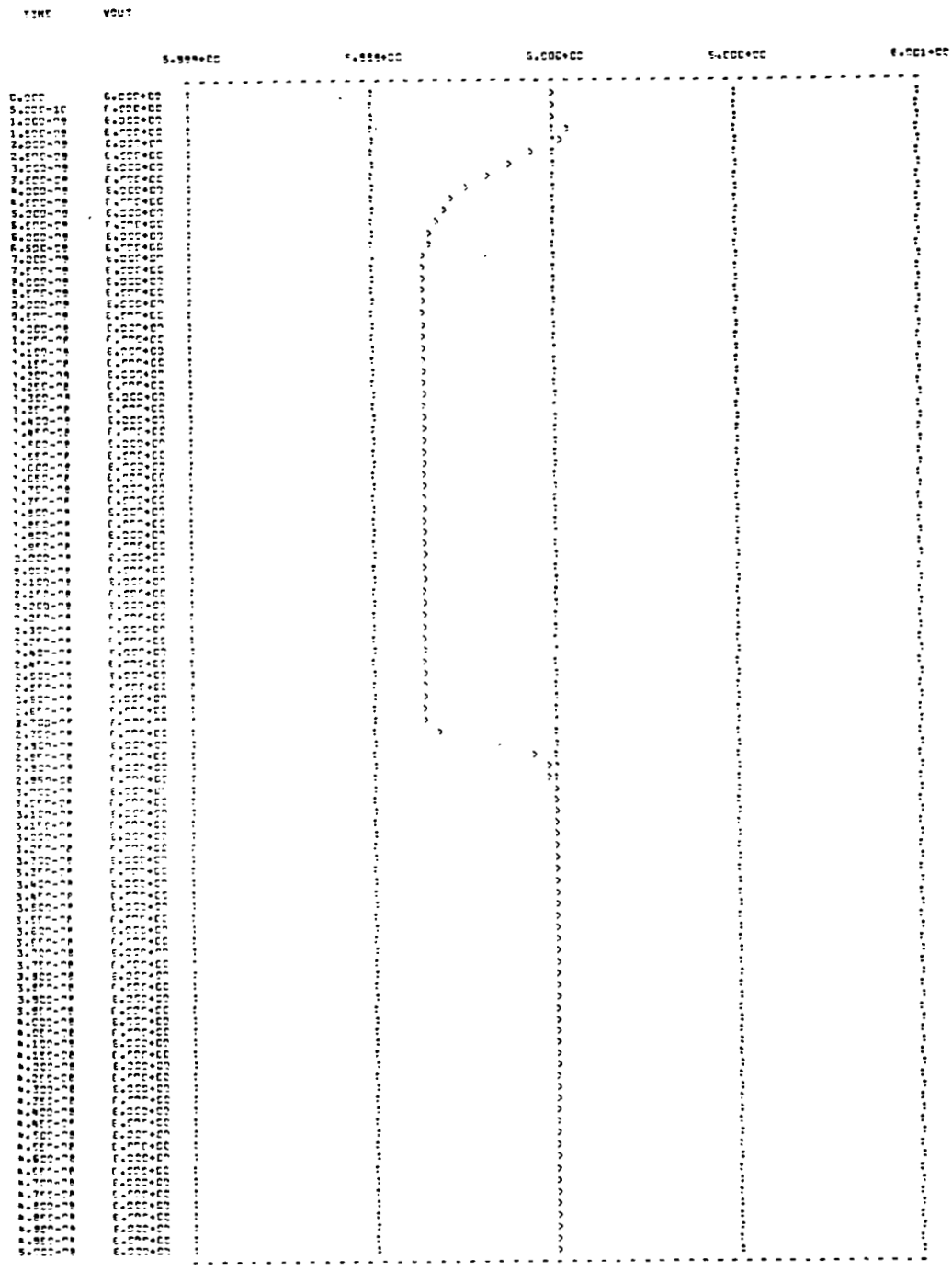


Figure 13. Computed Response of Silicon Gate MOS Transistor (Load Resistance $\approx$ 0)

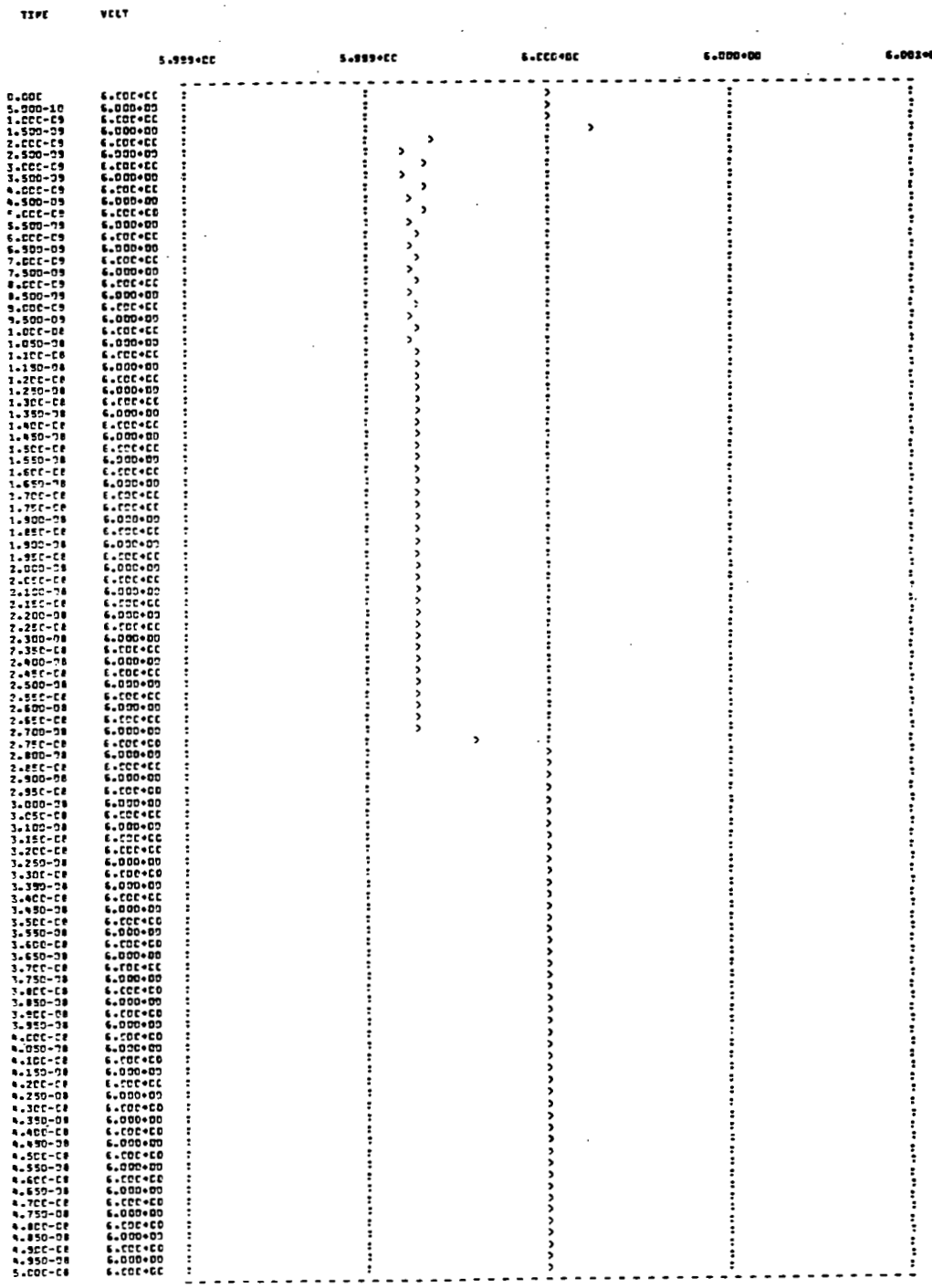


Figure 14. Computed Response of Silicon Gate MOS Transistor (Load Resistance $\approx$ 0)

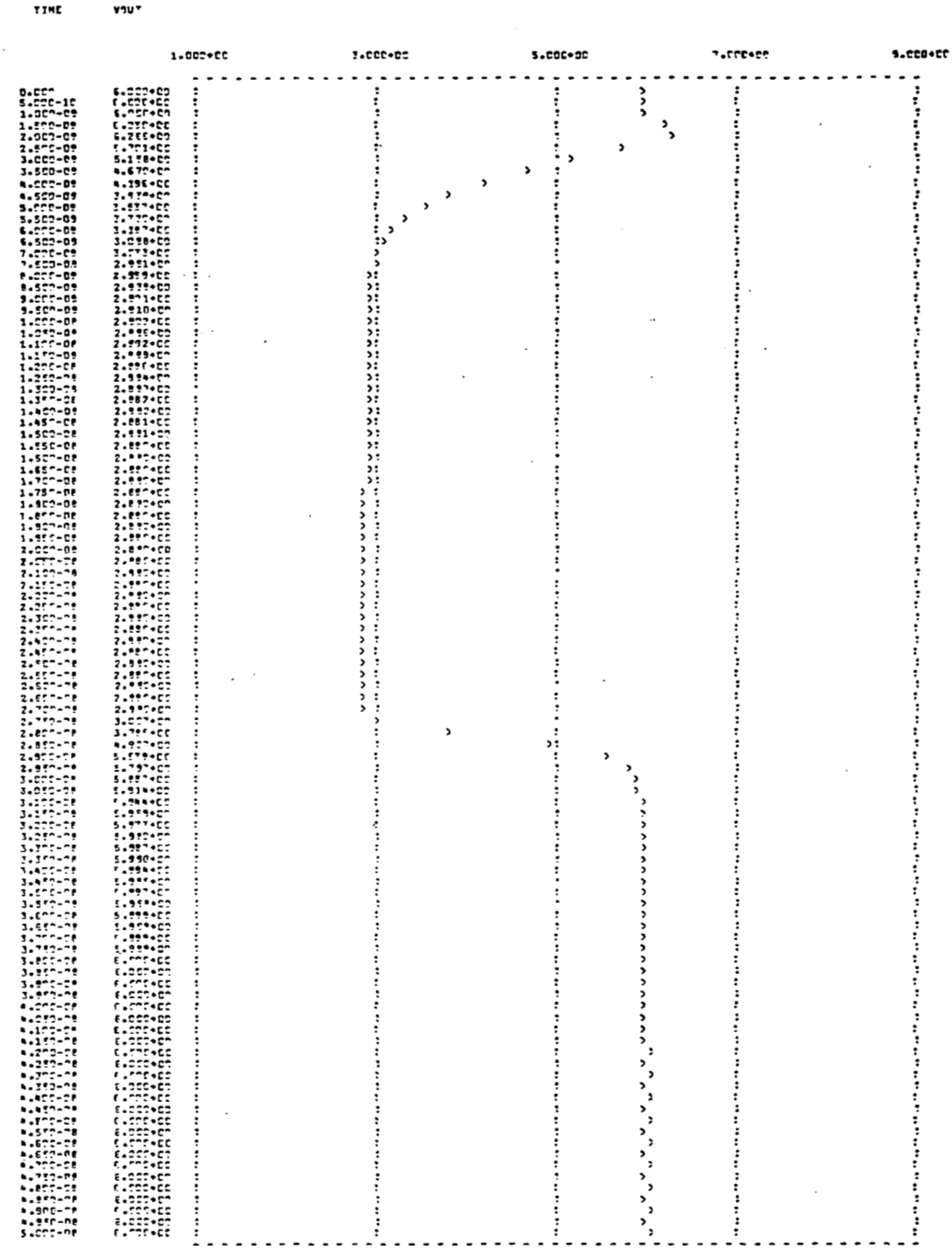


Figure 15. Computed Response of a Loaded Silicon Gate MOS Transistor (Load Resistance = 100 Ohms)

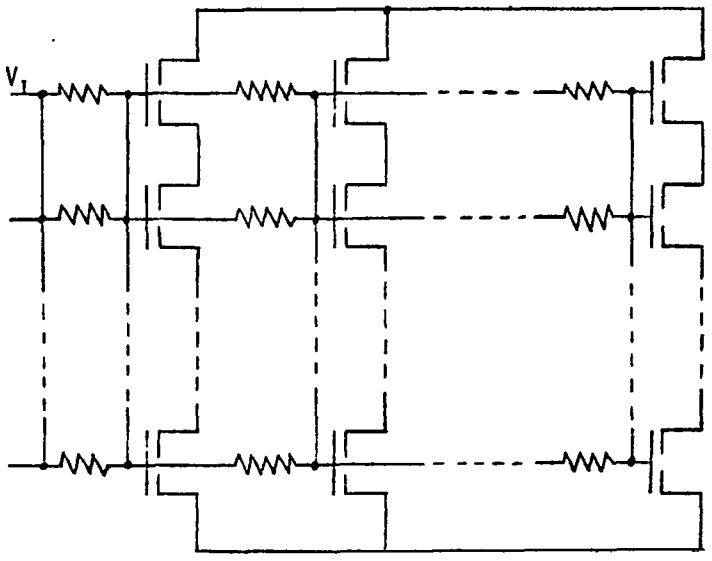


Figure 16. Equivalent Circuit of Silicon Gate DMOS Transistor

be computed by connecting a number of lumped incremental transistors in series each with a different equivalent threshold voltage. If a silicon gate is used for the DMOS transistor, a similar adverse effect on the speed can be expected due to the nonzero gate distributed series resistance.

The computation can be accomplished by using a series-parallel matrix equivalent circuit as shown in figure 16 for one transistor. Computational experience has shown that a DMOS can be approximated quite accurately by subdividing it into 5 series increments. For an n-channel DMOS transistor with background concentration = $10^{15}/cm^3$, the local equivalent threshold voltages, down an array of lumped sections written as a matrix column, were as follows: V_{T1} = 2.0V (also the overall threshold voltage), V_{T2} = 1.2V, V_{T3} = 0.4V, V_{T4} = -0.3V, V_{T5} = -0.8V. This computed result is shown in figure 17. As expected, in comparison with a conventional MOS transistor of the same gate length and same threshold voltage (shown in figure 13), the DMOS transistor shows a faster transient response. However, due to the time constant of the silicon gate, the risetime and falltime are not in the subnanosecond range.

DMOS RING OSCILLATOR

A number of inverters can be connected in cascade to form a ring oscillator. To form a ring oscillator, the number of inverters must be in odd number; then, the output level of the last inverter will be out of phase with the input level of the first inverter forcing the input to change level. The input signal must propagate through n number of inverters before the last inverter changes state. It takes time to propagate and takes two passes to complete a cycle. The propagation time t_p of a single inverter is

$$t_p = \frac{1}{2nf}$$

The design layout of each inverter in the ring oscillator is shown in figure 18. The last enhancement-mode DMOS transistor and the depletion-mode load device have the same aspect (width/length) ratio. The lower device is the DMOS transistor, which can be distinguished by the extra p-type diffusion mask outline, which is 1μm wider than the n+ source diffusion window. The drain-to-source spacing is conservatively drawn to 10μm. The channel width

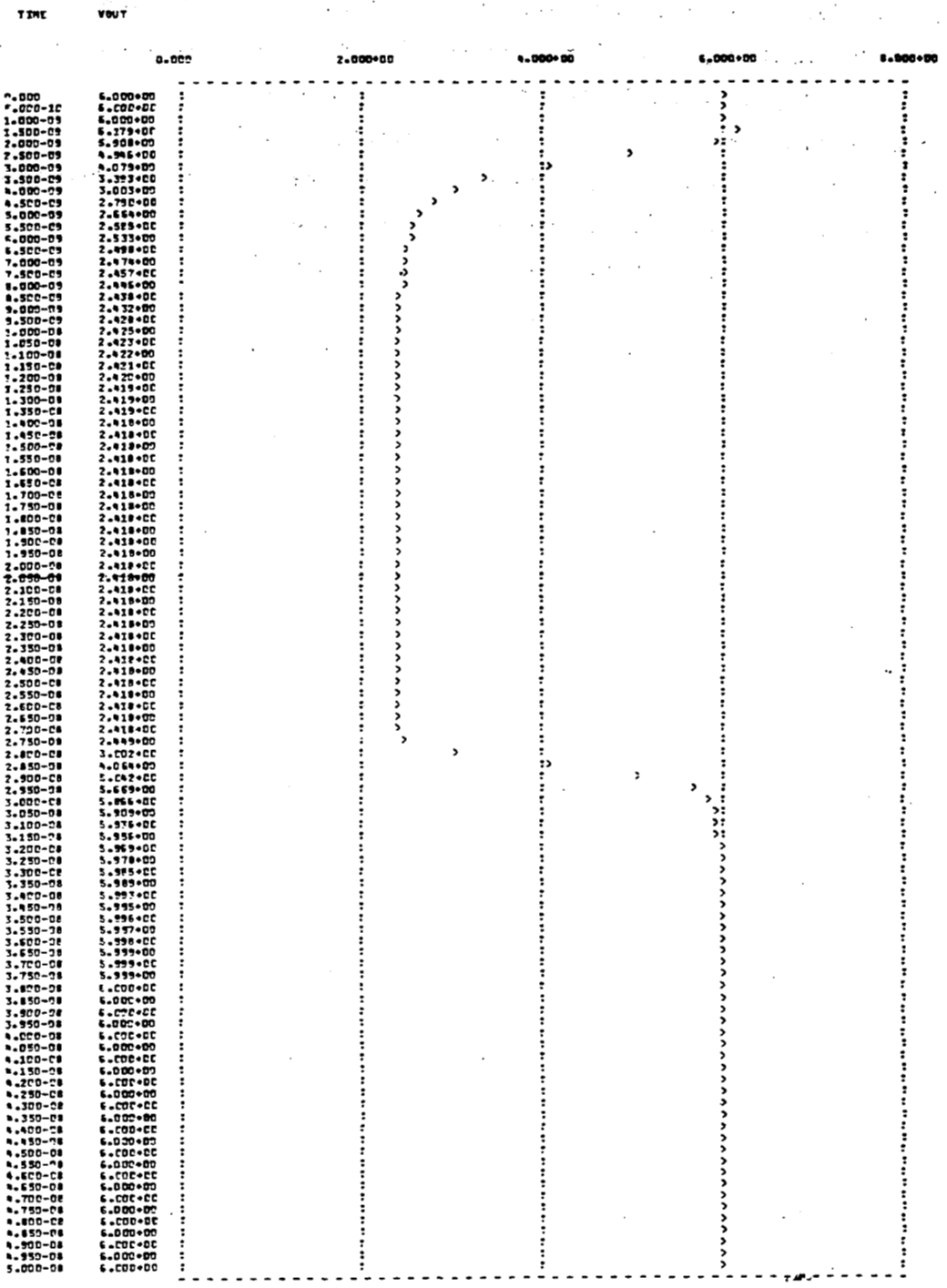


Figure 17. Computer Response of Silicon Gate DMOS Transistor (Load Resistance = 100 Ohms)

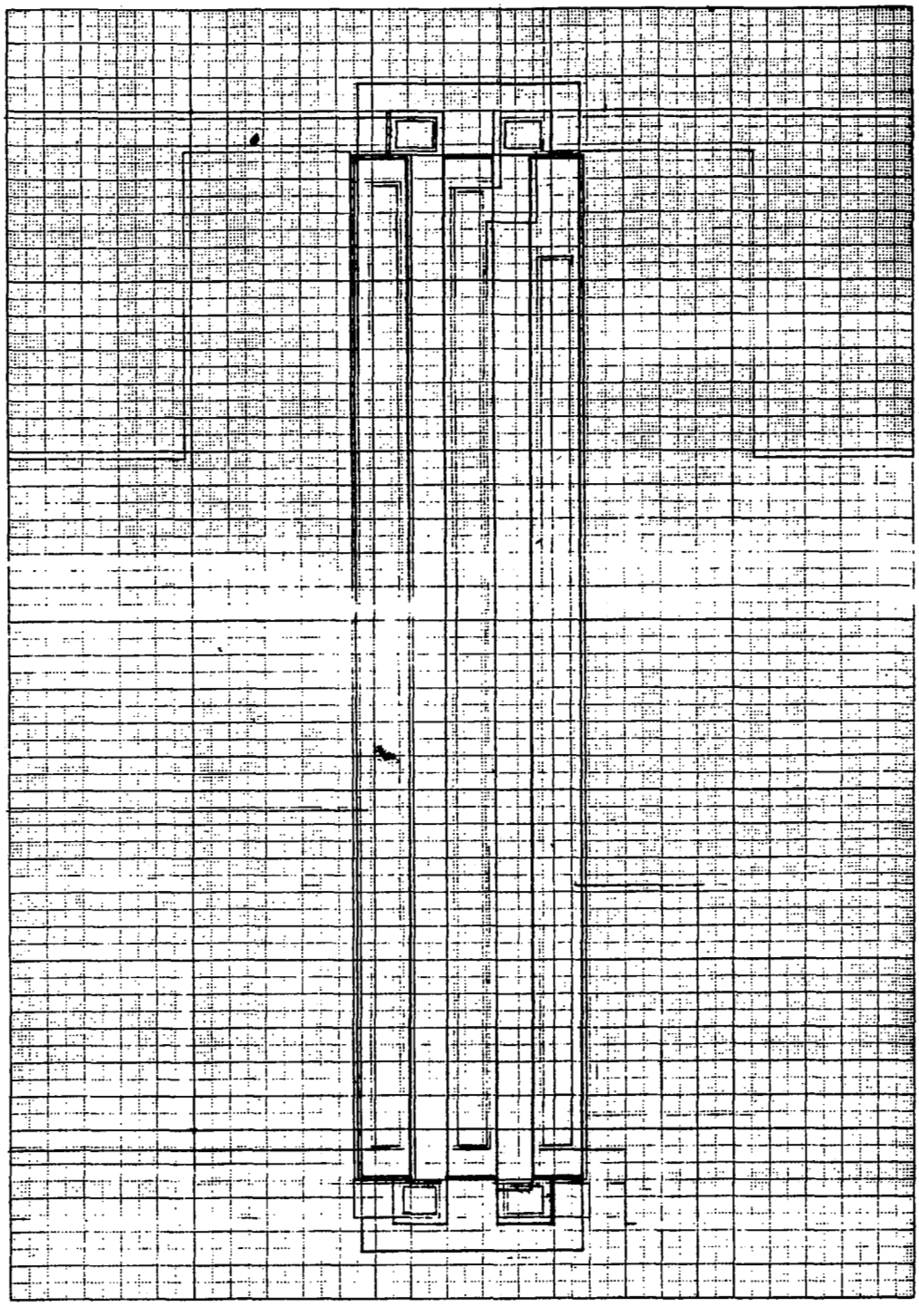


Figure 18. Inverter Layout for Ring Oscillator

is drawn to 300μm. The smaller dimension of the source and drain diffusions is 14μm. The smaller dimension of the contact window is 8μm. The contact to the silicon gate is made outside the active region.

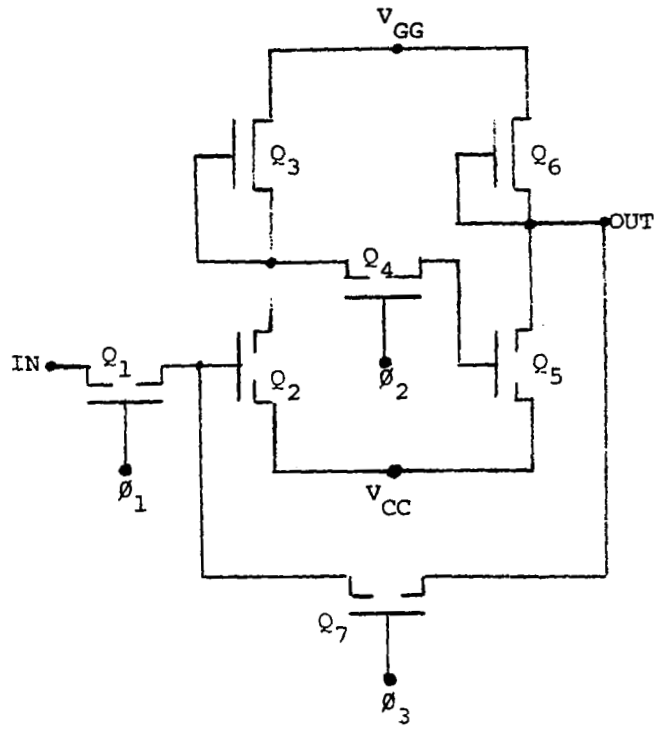
DMOS SHIFT REGISTER

A DMOS shift register cell design can be patterned after a conventional static shift register cell by use of two inverters which are cross-coupled through transmission gates to form a flip-flop as shown in figure 19. The transmission gates serve as active switches and should therefore be enhancement mode devices. When the transmission gate Q_1 is closed by the clock signal ϕ_1, information is coupled into the gate of the first inverter and appears as an inverter signal at the drain output of Q_2. During this time, the transmission gates Q_4 and Q_7 are opened so as not to load the input signal. After the clock pulse ϕ_1 is turned off, information is stored at the gate and the drain node capacitances of Q_2. Then the clock pulse ϕ_2 is applied to the transmission gate Q_4. The information stored at the drain node capacitance is transferred to the gate of Q_5 (or the input of the second inverter). Thus a doubly-inverted or in-phase signal appears at the drain output of Q_5. Thereupon, a third clock pulse Q_3, slightly delayed from ϕ_2, is applied to the transmission gate Q_7 to close it and latch the cross-coupled flip-flop. The timing relation of the different clocks is shown at the bottom of figure 19.

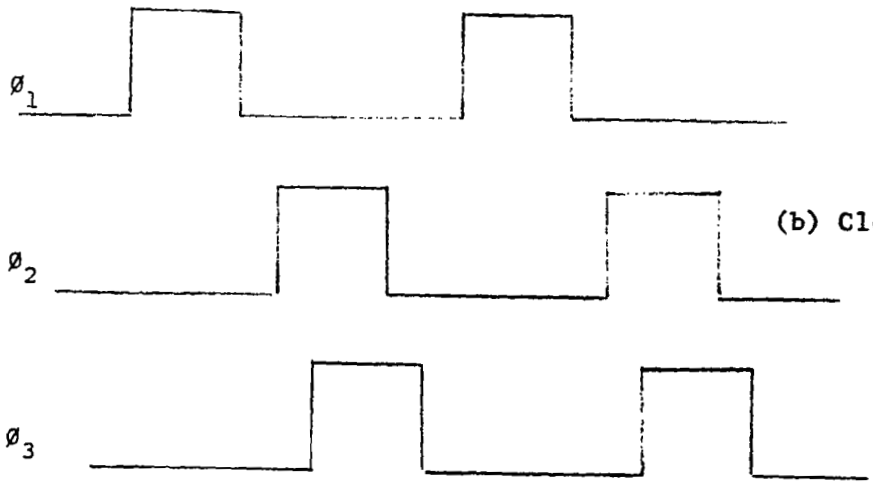
In the DMOS integrated circuit made by inverters and transmission gates already discussed, only two types of transistors are available, either the depletion-mode transistor, which is unsuitable for transmission gates or the enhancement-mode DMOS transmission gates. Thus, in the shift-register cell shown in figure 18, Q_1, Q_2, Q_4, Q_5, and Q_7 are DMOS, and the load devices Q_3 and Q_6 are of depletion-mode.

AN OPTIMIZED OUTPUT STAGE FOR MOS INTEGRATED CIRCUITS[9]

In MOS integrated circuits, the output transistors should be large enough to drive the required load capacitance. If there is a significant capacitive load of this last output transistor, it will capacitively load the previous



(a) Schematic Diagram



(b) Clock Timing Diagram

Figure 19. Shift Register

stage and slow it down. To improve the propagation time, the driver for the output stage should also be enlarged. Obviously, if every stage is enlarged, excessive area will be consumed. However, if the drivers are tapered in some fashion from a minimum area cell to maximum area output cell as shown in figure 20, both the propagation time and the area may be conserved. The following is an analysis of such a tapered output stage for optimizing propagation time and chip area. This technique is applicable in general to MOS integrated circuits where a high output drive capability is required.

Output Stage Design Principles

To drive typical load capacitances required in an integrated electronic system, the output transistors in an integrated circuit should be of suitable size to provide large charging and discharging currents. Unfortunately, large output transistors load the previous stage, which decreases its operating speed. To improve this situation, the driver for the output stage should also be enlarged. This analysis can obviously be carried backward on a stage-by-stage basis.

Consider the output circuit such as the shift register shown in figure 20. If the load capacitance, C_L, is equal to M times the interstage node capacitance, C_0, and the MOS devices are of minimum sizes, the propagation delay (which is proportional to load capacitance) through the last two stages is:

$$t_{pL} = (M + 1)\, t_{po} \tag{23}$$

where

t_{pL} = propagation delay through the last two stages
t_{po} = propagation delay at the low level interstage
$M = C_L/C_0$
C_L = load capacitance
C_0 = interstage node capacitance

The propagation time can be reduced by enlarging the sizes of the MOS devices of the last stage, I_1, by m times. The driver stage, I_2, is slowed down because the output transistors now see an m-fold increase in capacitive load. The propagation delay of the driver and output stages becomes:

$$t_{pL} = \frac{M}{m}(t_{po}) + m\,(t_{po}) \tag{24}$$

To find the minimum t_{pL}, we differentiate the above with respect to m and set the differentiated equation to zero:

$$\frac{d(t_{pL})}{dm} = \left(\frac{-M}{m^2} + 1\right) = 0 \tag{25}$$

or

$$m = M^{1/2}$$

and

$$t_p\,(min) = 2M^{1/2}t_{po} \tag{26}$$

For shorter propagation time, it is better to enlarge the sizes of the last few stages. For instance, one may increase stages I_1 by m_1 times, increase stage I_2 by m_2 times the previous stage, I_3 by m_3 times, etc. Then the propagation delay for the last four stages becomes:

$$t_{p4} = \left(\frac{M}{m_1} + \frac{m_1}{m_2} + \frac{m_2}{m_3} + m_3\right) t_{po} \tag{27}$$

By differentiating t_{p4}, we find:

$$m_1 = M^{3/4};\ m_2 = M^{2/4};\ m_3 = M^{1/4} \tag{28}$$

and

$$t_{p4} = 4M^{1/4}t_{po} \tag{29}$$

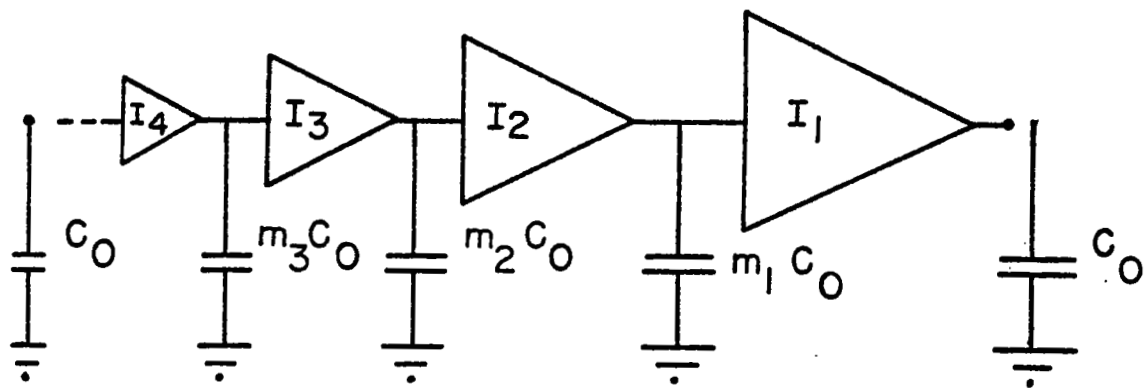


Figure 20. Cascade MOS Stages

The propagation delay of every stage is:

$$t_{pL} = M^{1/4} t_{po} \tag{30}$$

Note that the propagation time per stage resulting from enlarging the last three stages given in equation (30) is shorter than that of enlarging just the last stage. Also, the optimum sizes of the device decreases monotonically as $M^{3/4}$, $M^{2/4}$, $M^{1/4}$ times the minimum interstage sizes.

From the last two equations, one may generalize that for an output device of N stages, the propagation time per stage is

$$t_{pS} = M^{1/N} t_{po} \tag{31}$$

A computer analysis of a 10-stage output device utilizing a 2-phase ratioless logic interstage was made. The results of the analysis showed good agreement of the predicted propagation delay to the computed delay to within the accuracy of the program used.

Area Optimization

From the above derivation, one can see that the optimum device size should decrease on the order

$$m_1 = M^{(N-1)/N}, \quad m_2 = M^{(N-2)/N}$$

$$m_3 = M^{(N-3)/N}, \quad \ldots \quad M_{N-1} = M^{1/N}$$

$$m_N = 1$$

When a number of stages are connected in cascade, the area is the sum of all the stages. When the stages are tapered to optimize the speed, each preceding stage is reduced by a factor 1/m. If there are N-1 enlarged stages, the area of the last N stage is

$$A_N = A_o \left(1 + m + m^2 + m^3 + \ldots . m^{N-1}\right) \tag{32}$$

where A_0 is the area of the standard stage (not enlarged). This summation is equal to:

$$A_N = \frac{A_o \left(m^N - 1\right)}{m - 1} \tag{33}$$

But since

$$M = m^N$$

$$A_N = \frac{(M - 1)}{M^{1/N} - 1} A_o \tag{34}$$

This is the ratio of the enlarged area to the standard cell area. Equation (34) is plotted in figure 21 for M = 100. Also plotted is the propagation time of stage t_{pS}, which is normalized:

$$t_{pS}/t_{po} = M^{1/N} \tag{35}$$

From figure 21 we find that the increased area ratio is of the order of the load capacitance to node capacitance ratio, M; the area increases as the number of enlarged stages N increases; and the propagation time per stage decreases as the numbers of enlarged stages decreases.

Optimum Stage Design

The optimum design is a compromise between speed and area. The choice is based on a nebulous parameter, cost, as in any search method. We can arbitrarily choose a figure of merit, F, defined as

$$F = \left(\frac{A_N}{A_o}\right) \left(\frac{t_{pS}}{t_{po}}\right)^K \tag{36}$$

where K is a weighting exponent. If K is greater than unity, it means more

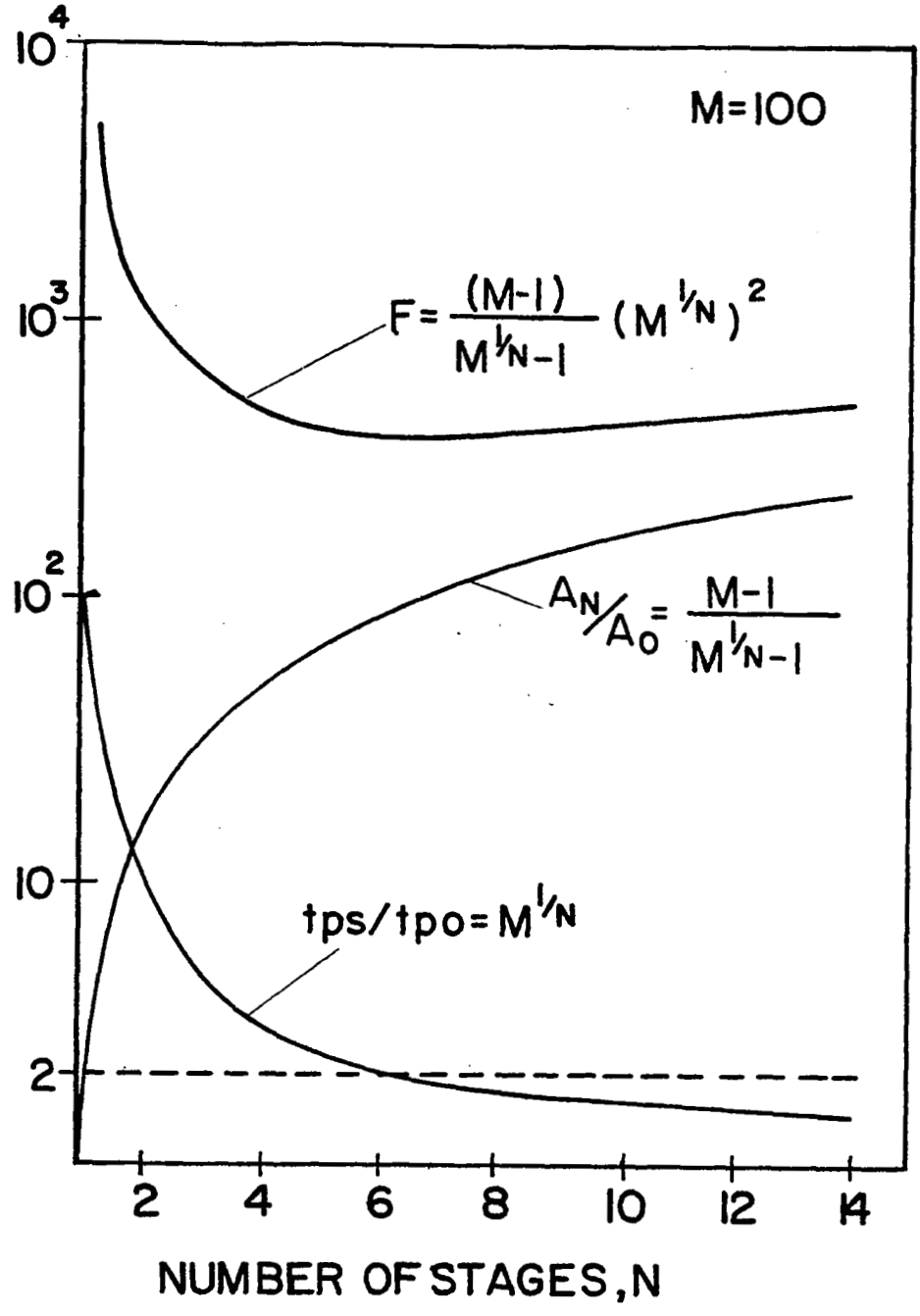


Figure 21. Normalized Propagation Delay, Normalized Area, and Figure of Merit F, vs. Number of Stages for M = 100, K = 2

weight is placed on speed than area. Substituting equations (34) and (35) into (36):

$$F = \frac{(M - 1)}{M^{1/N} - 1} \quad (M^{1/N})^K \tag{37}$$

The minimum F can be found by setting dF/dN to zero, yielding:

$$M^{1/N} = \frac{K}{K - 1} \tag{38}$$

For instance, if K = 2:

$$M^{1/N} = 2$$

This has several implications. From equation (35):

$$t_{pS}/t_{po} = M^{1/N}$$

The minimum area and propagation delay is achieved when:

$$t_{pS}/t_{po} = 2 \tag{39}$$

The optimum area is:

$$A_N \text{ (opt)} = (M - 1)\, A_o \tag{40}$$

Figure 22 is representative of how the figure of merit varies for several load to interstage capacitance values. Below the optimum value, t_{pS} decreases more slowly than the area increases; above this value, t_{pS} increases more quickly than the area decreases.

At some point, however, a large increase in propagation delay occurs. As seen in figure 21, the area-propagation delay square product increases drastically for small values of N. This is not the case for larger values of N. Within the design constraints of a particular circuit, the designer pays a smaller penalty in the area propagation delay square product by increasing the number

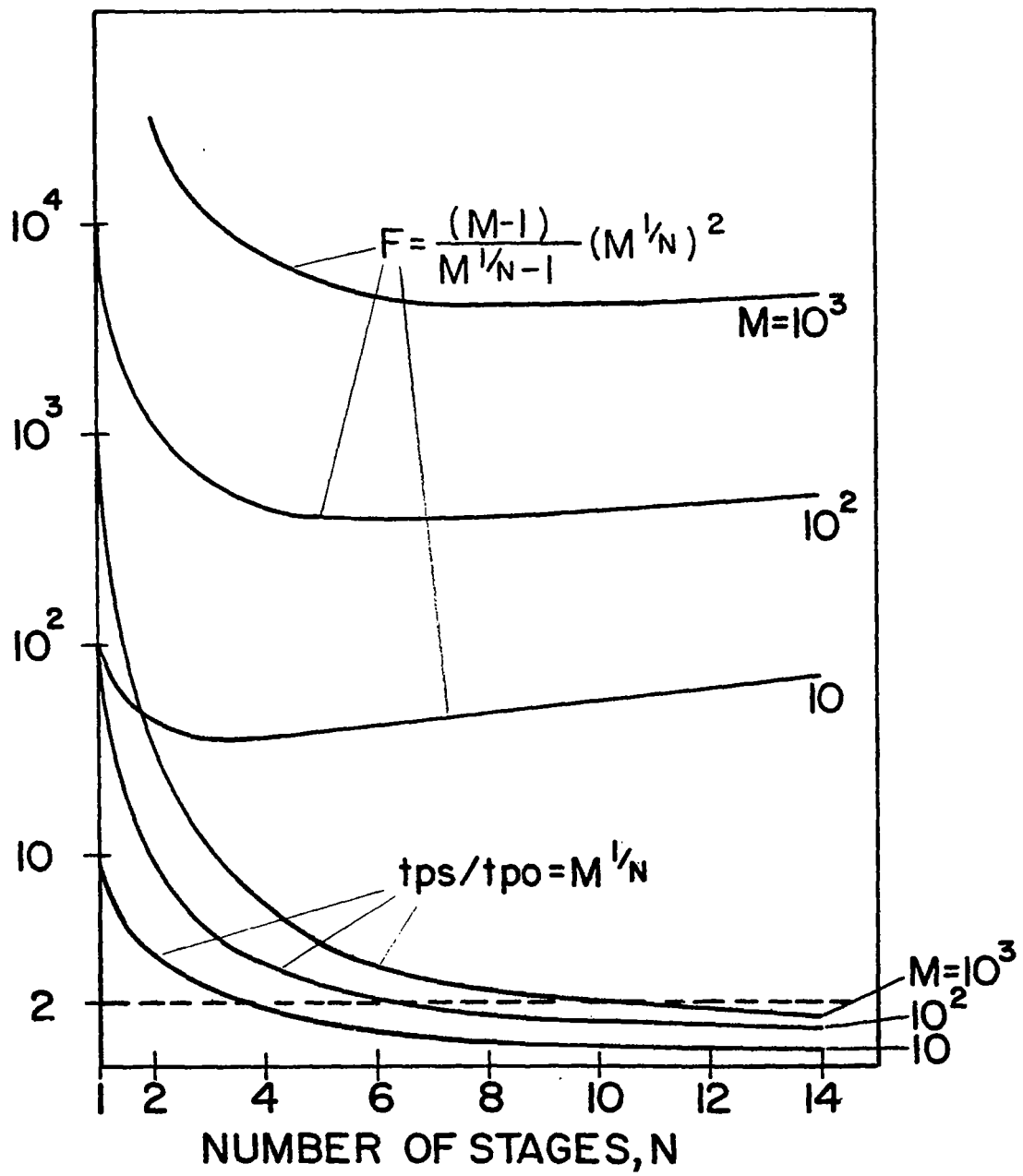


Figure 22. Normalized Propagation Delay and Area-Propagation Delay Square Product, F, vs. Number of Stages

of stages beyond the optimum point than decreasing the number of stages. Also since:

$$M^{1/N} = 2$$
$$M = m^{N} \tag{41}$$

An optimum condition exists when

$$m = 2 \tag{42}$$

or when the capacitance ratio per stage doubles. Simply stated, the best compromise between area increase and the square of propagation delay decrease exists when the output device is designed such that each required stage in the output circuit doubles in size or capacitance between the last interstage circuit and the load. Figure 23 shows the number of stages required for a given load to interstage capacitance ratio for optimum performance.

From a design standpoint, it is most economical to minimize the area. If the entire integrated circuit consists of B number of cells, of which N stages are enlarged, the total area is

$$A = (B - N)\, A_o + A_N \tag{43}$$

assuming each cell occupies A_o area. In most instances, B is large. The enlarged area A_N does not add any more bits but merely improves the loading and speed capability. Thus, a compromise exists between speed and area. If $(B - N) \gg A_N/A_o$, the addition of enlarged area is insignificant. If A_N becomes comparable or larger than $(B - N)\, A_o$, the enlarged buffer can add substantially to the area and cost of the integrated circuit. In this case, the design engineer may choose to commit some fraction of the total chip area to an output device. Figure 24 represents the minimum propagation delay obtainable for a given area and a fixed capacitive load.

MASK DESIGN

The layout of the basic shift-register cell is shown in figure 25. Since the transmission gates are bidirectional devices, it does not matter whether the source should be located at the input or the output side. From a topological standpoint, it is more convenient dand space saving for two transistors to

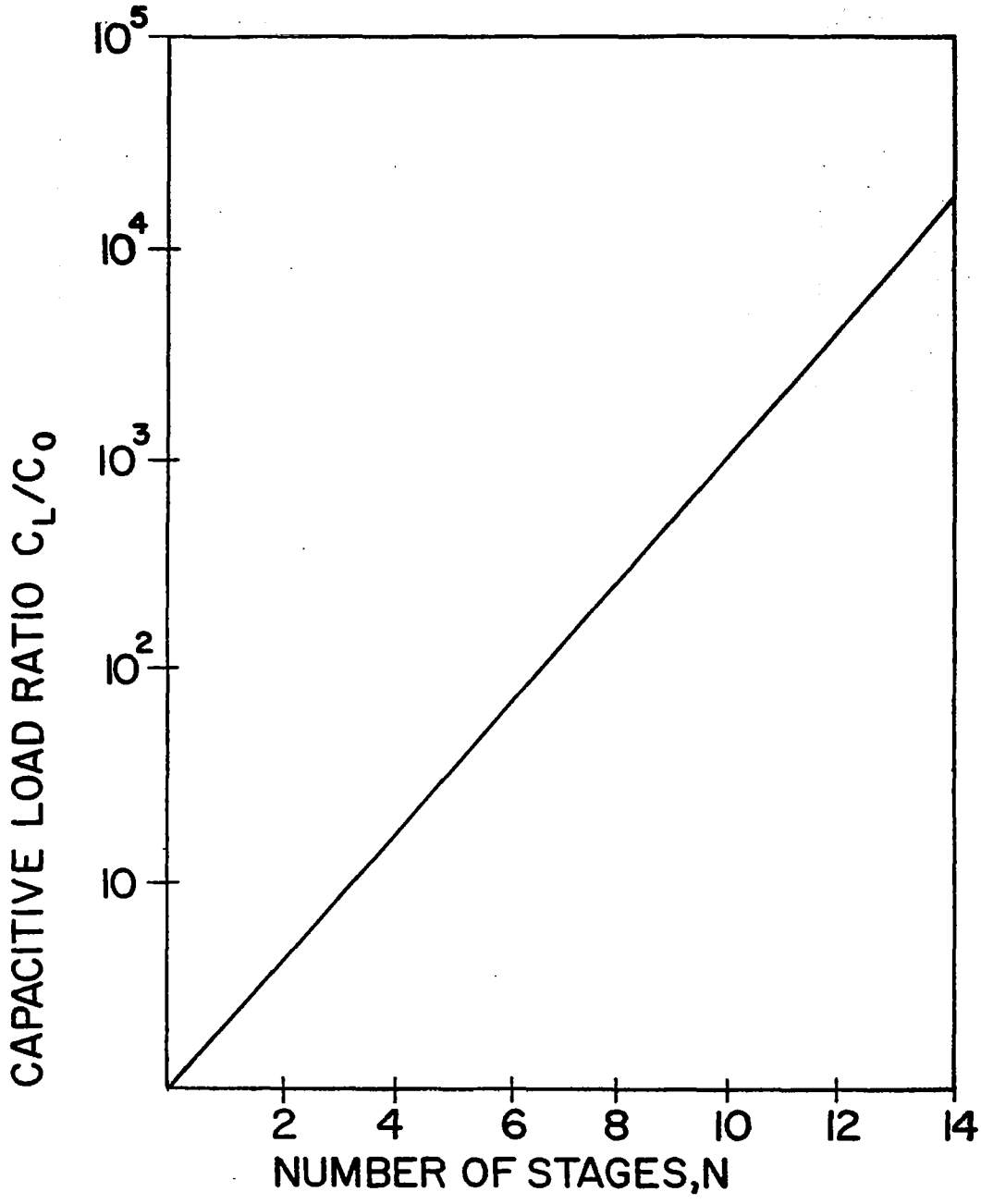


Figure 23 . Optimum Number of Enlarged Output Stages for Different Load to Node Capacitance Ratios

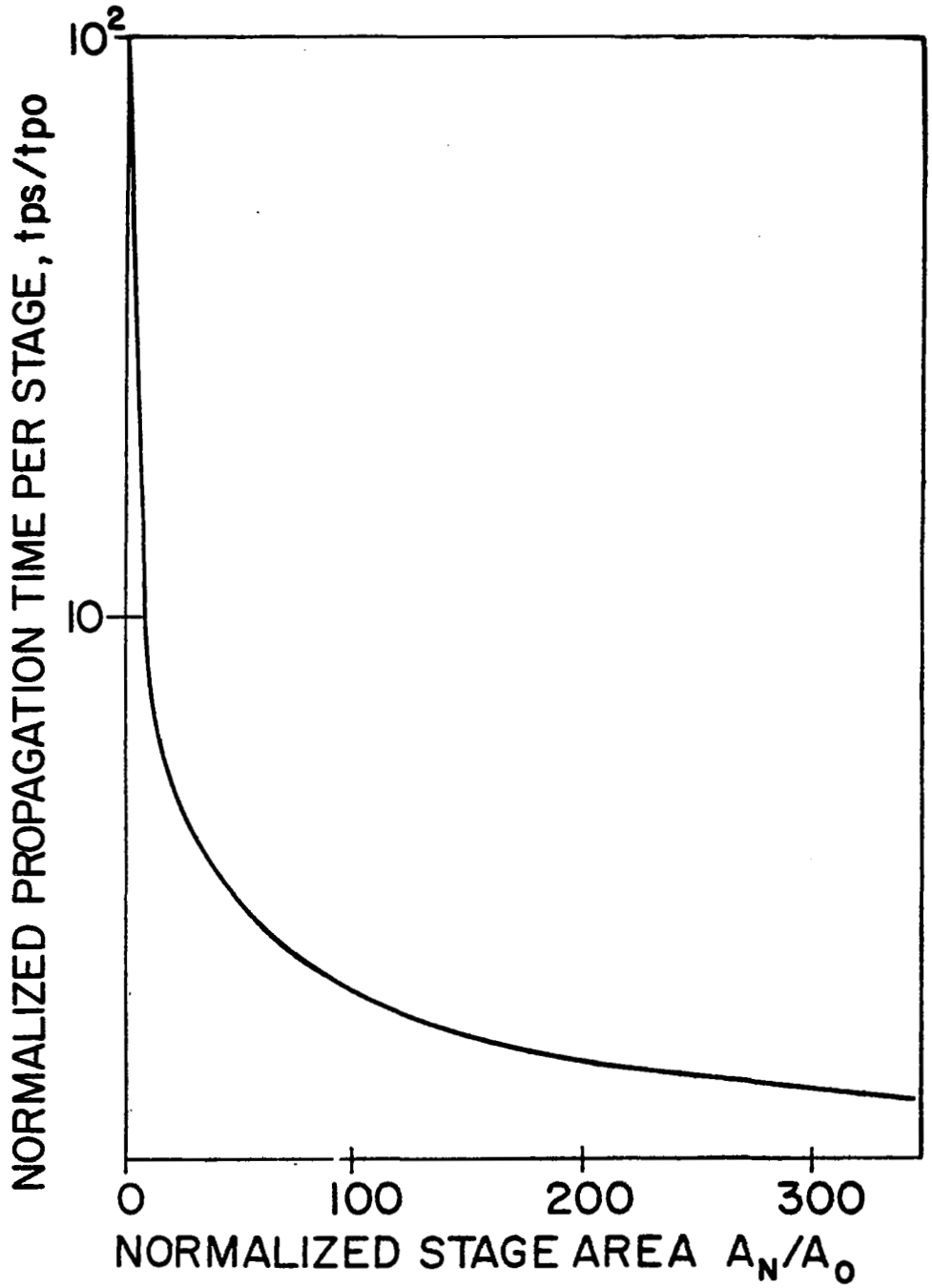


Figure 24. Minimum Propagation Delay Obtainable for a Given Area and a Fixed Load to Node Ratio of 100

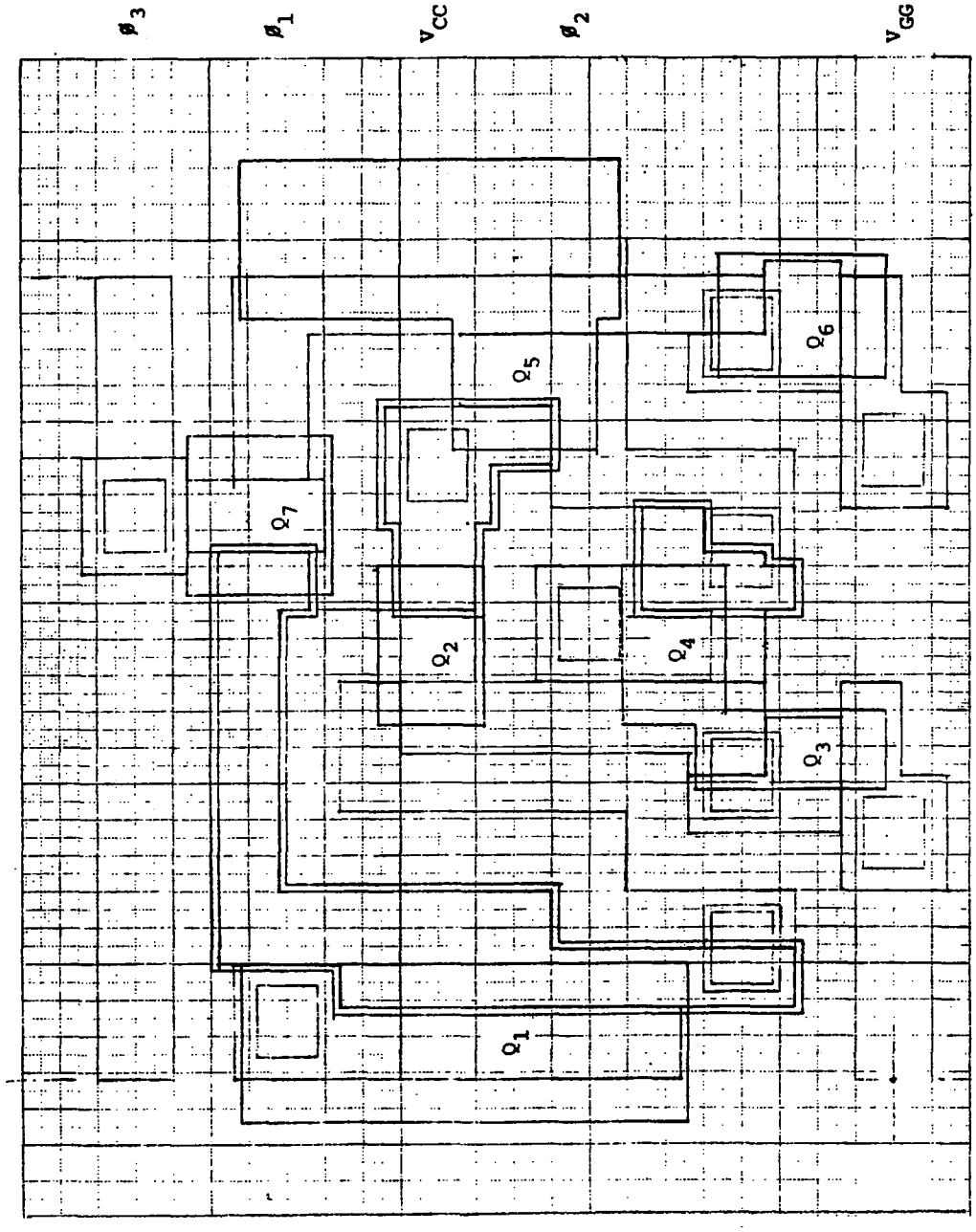


Figure 25. Composite Layout of the Shift Register Cell

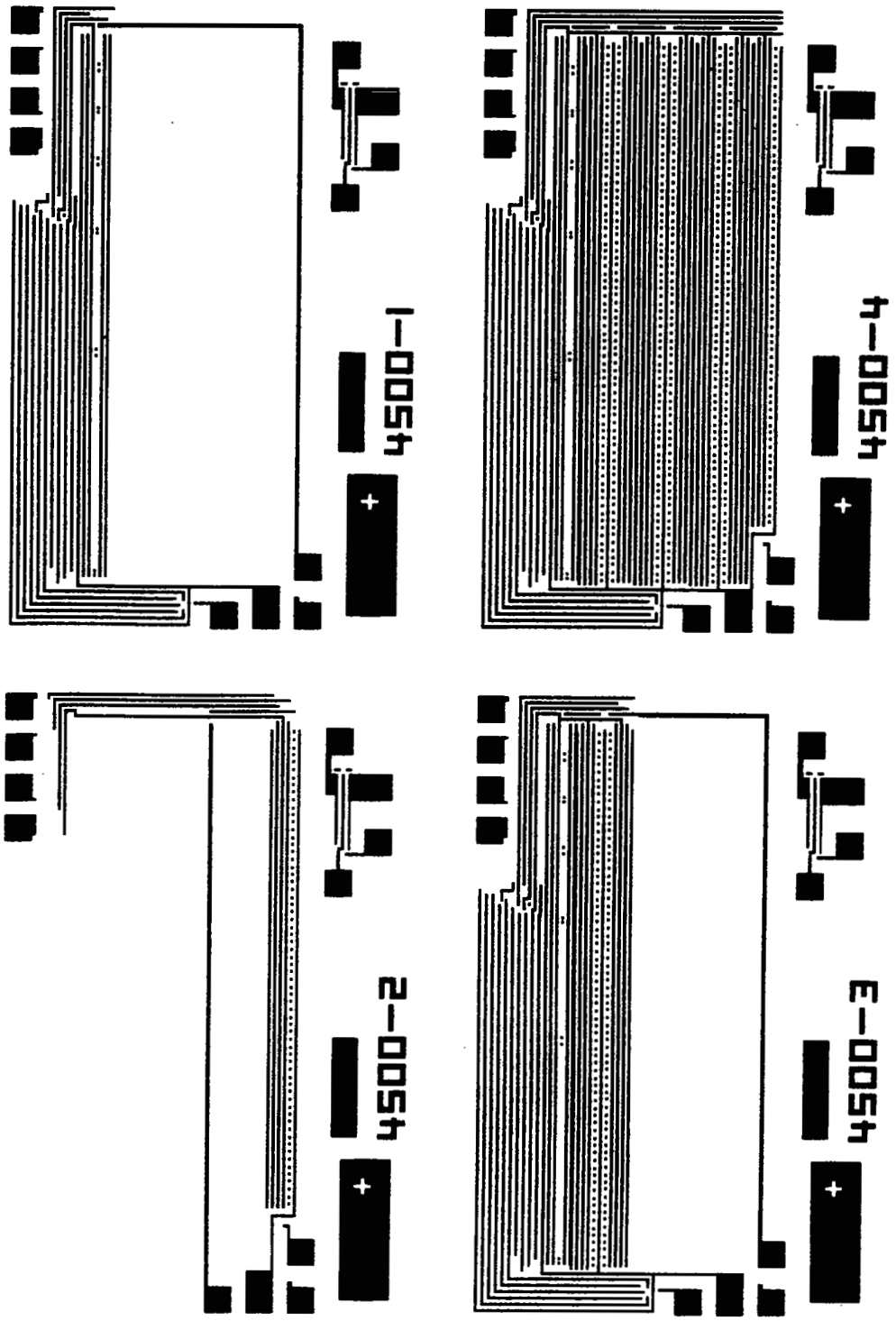


Figure 26. Overall Interconnection Mask

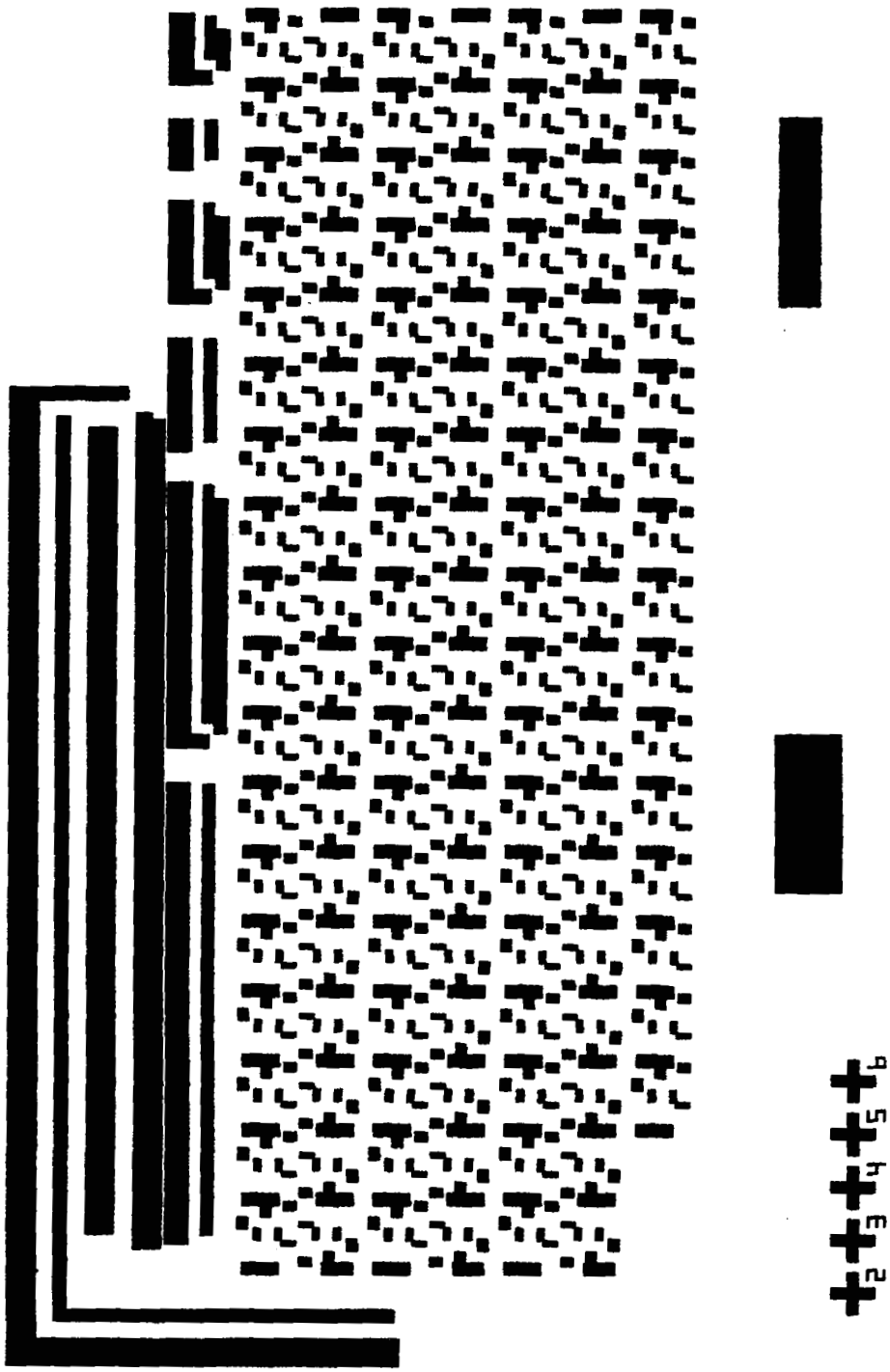


Figure 27. P⁻ Channel-Stop Diffusion Mask

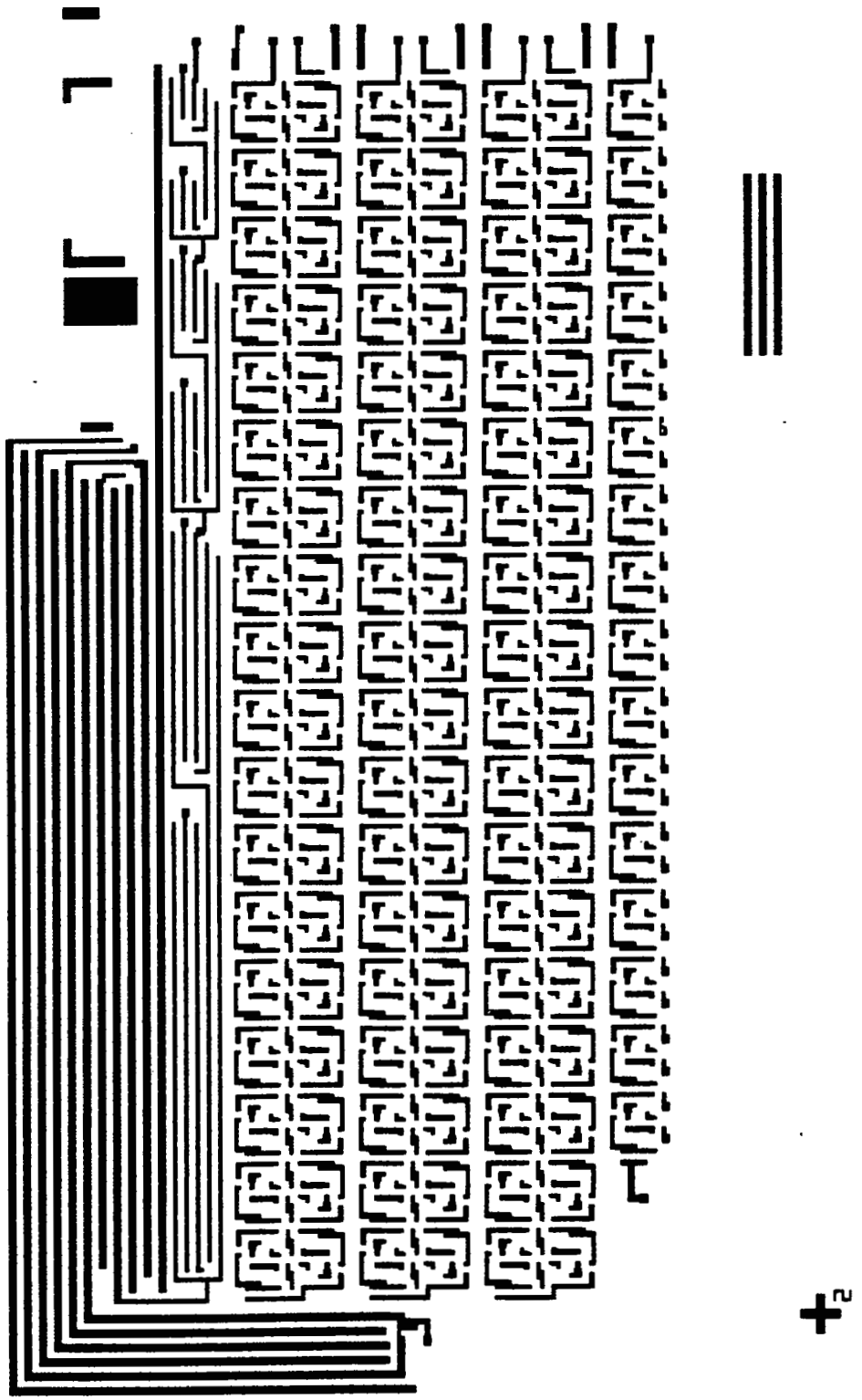


Figure 28. N^+ Diffusion Mask

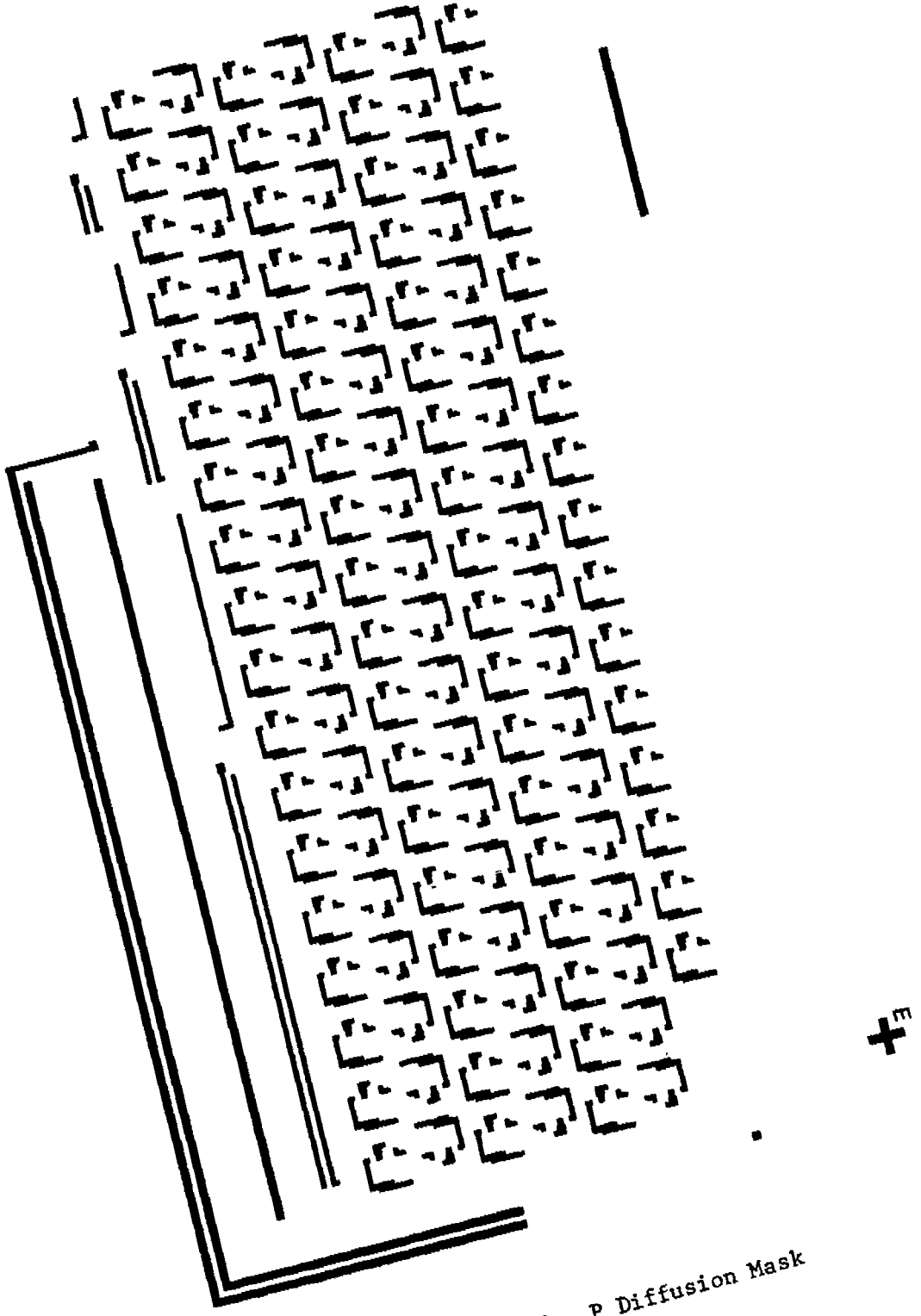


Figure 29. P Diffusion Mask

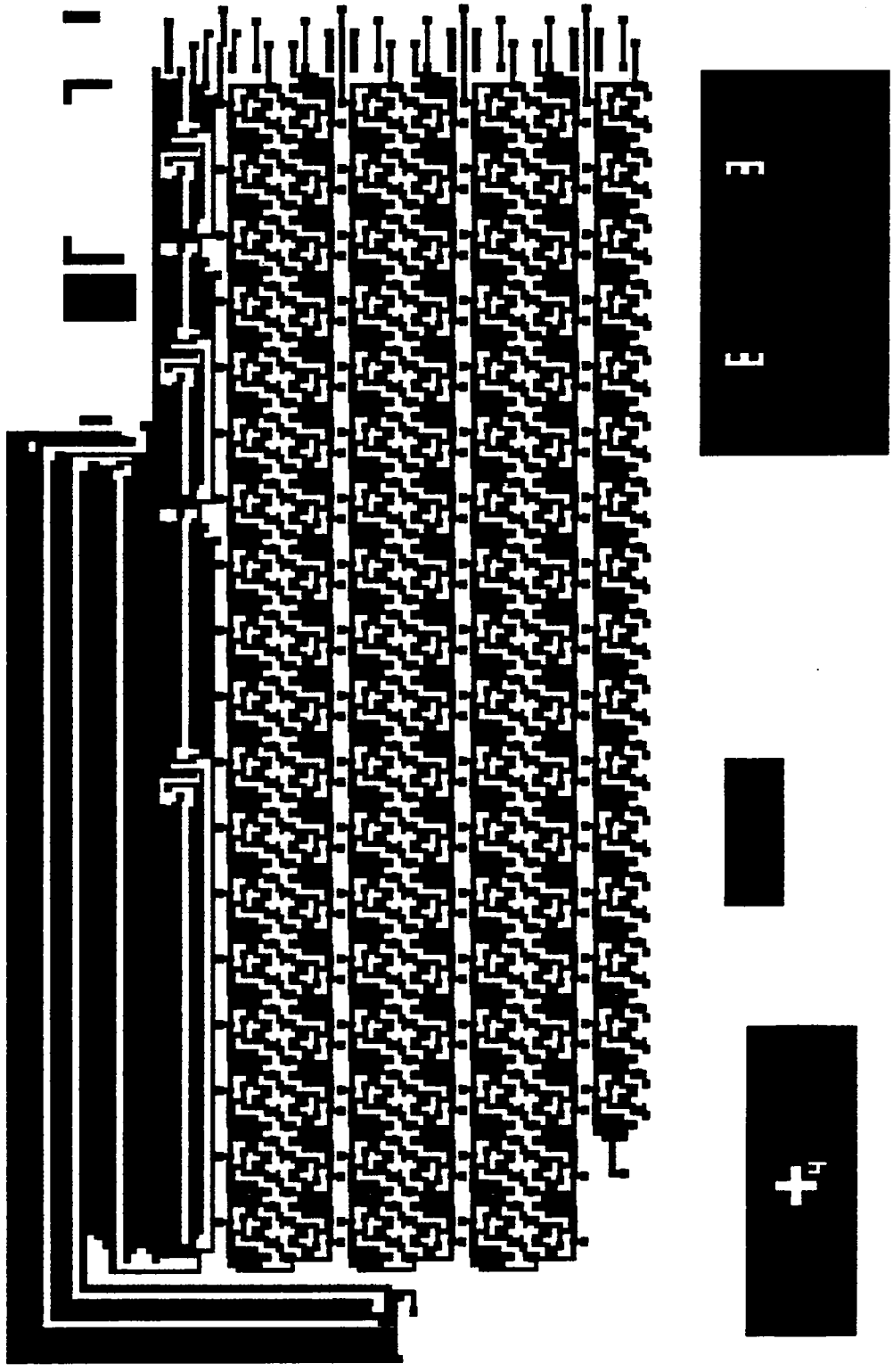


Figure 30. Silicon Etch Mask

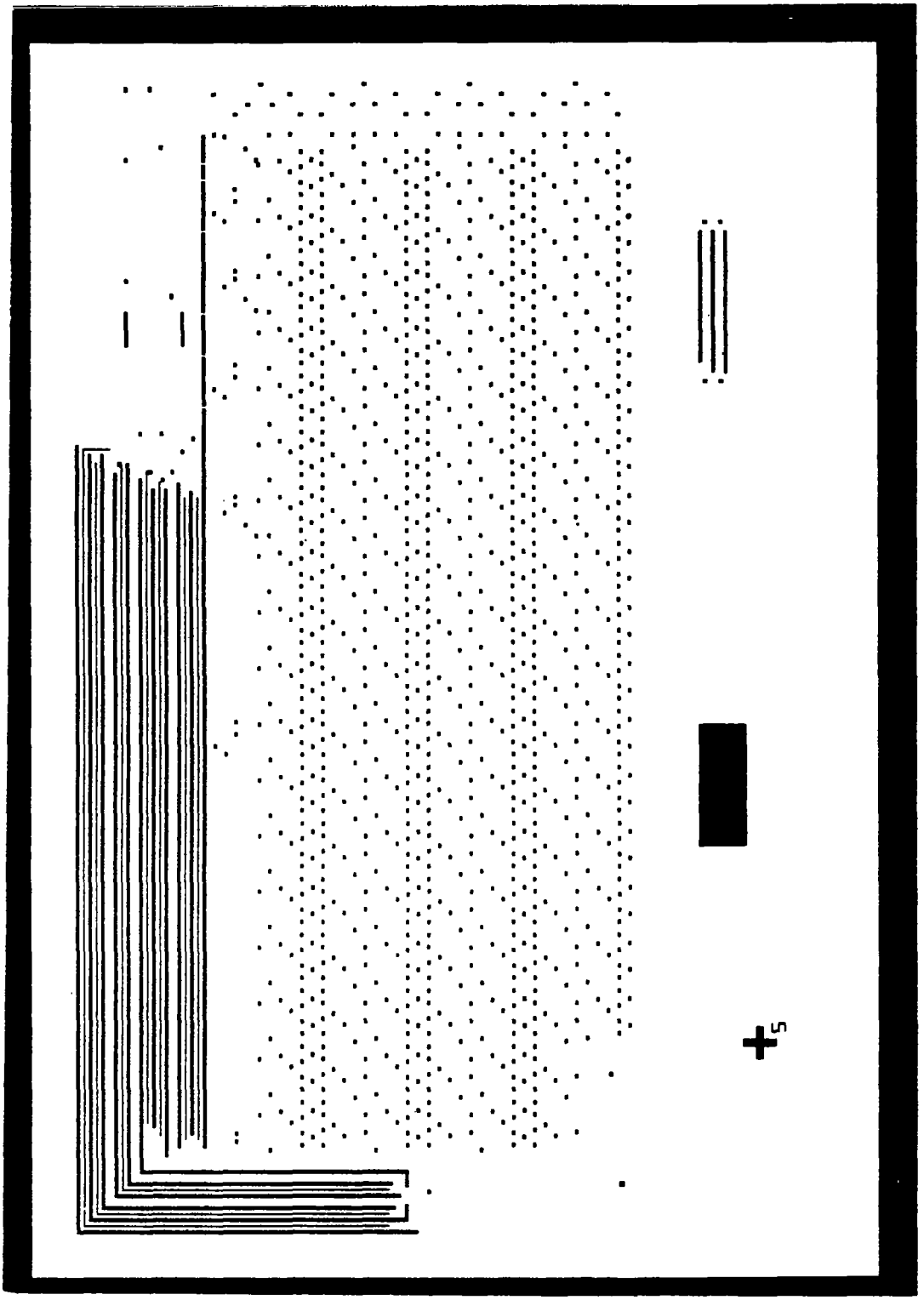


Figure 31. Contact Mask

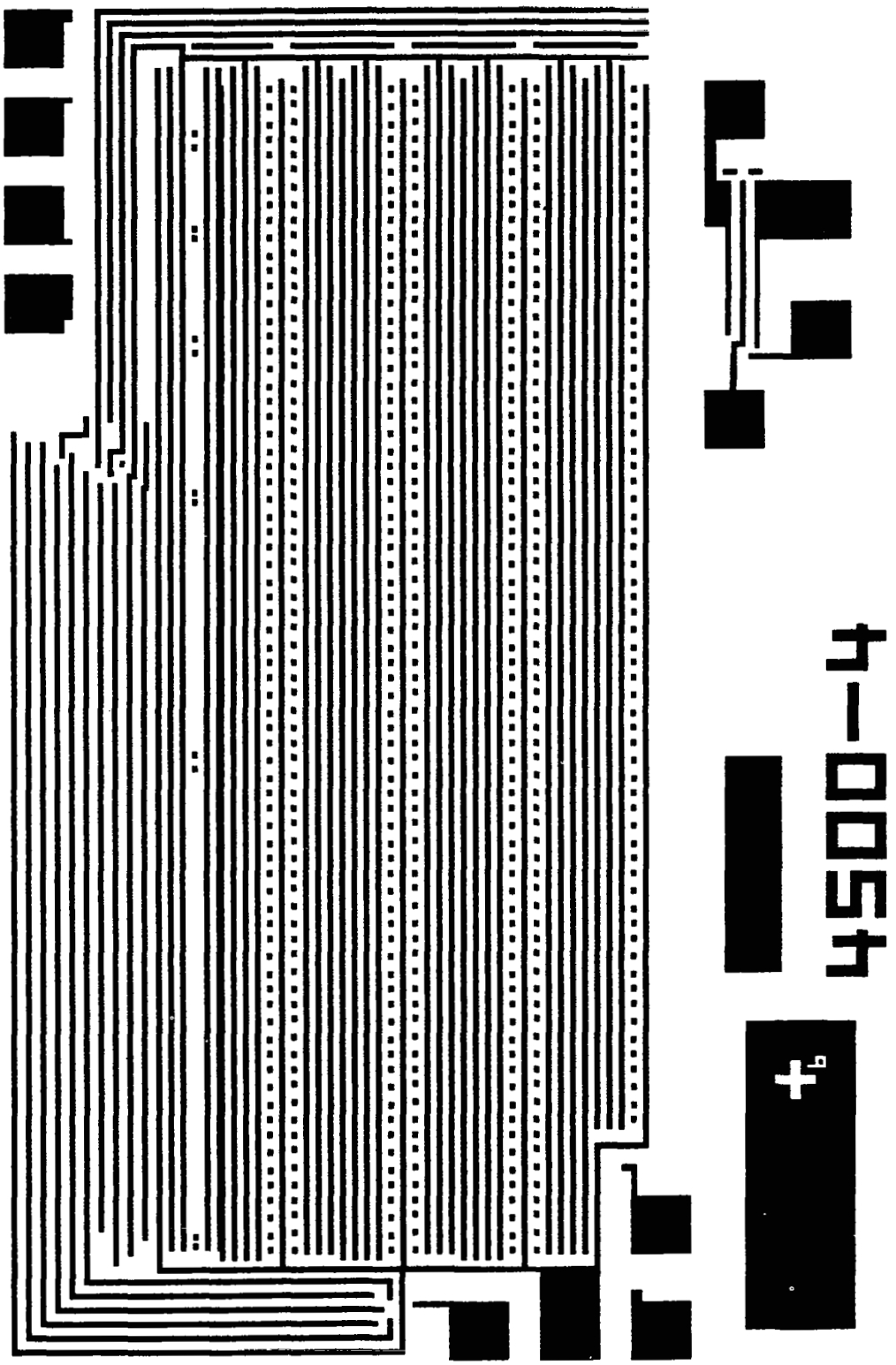


Figure 32. Interconnection Mask

have common sources or common drains.

The layout of the integrated circuits is shown in figures 26 through 32. There are four versions. Mask 4500-4 is the full 128 bit shift register. The first 124 bits are of low-level stages with minimum device geometry. The last 4 bits are progressively enlarged to optimize the speed-area product. The design has a zigzag path. The first two have 16 bits, the next six rows have 18 bits each, the eighth row has 3 progressively enlarged stages, and the last two rows comprise the last stage with the output transistors wrapping around the bend.

The second version (4500-2) is a 16-bit shift register using the first row of the low level stages. The change is accomplished by changing the interconnection pattern. The third version (4500-1) is a 4-bit shift register consisting of only the last four enlarged output stages. The fourth version (4500-3) is a 40-bit shift register consisting of two rows of low level stages with 36 bits and the last four output stages. In every version, there is a capacitor and a test inverter.

There are six masks (see Fabrication section). The masks were generated by the David Mann pattern generator.

FABRICATION

Double-diffused MOS (DMOS) transistors are capable of high-speed operation by effectively reducing the channel length using the difference of two successive diffusions through the same source window. For high transconductance, double diffusion should not be used for the drain. Thus, the drain window and the source window are usually not cut at the same time. As a result, self-alignment of the gate cannot be achieved. For best high-frequency performance, it is desirable to have self-aligned gates to reduce the area and the drain-to-gate capacitance. This work features a method for achieving self-alignment. In addition, a depletion mode load device with self-aligned gate is also incorporated in an integrated structure. The desired structure is achieved by using a polysilicon gate in combination with selective nitride masking.[10]

The desired inverter configuration is shown in figure 5 with a DMOS transistor driving a depletion load device. Figure 33 shows the processing

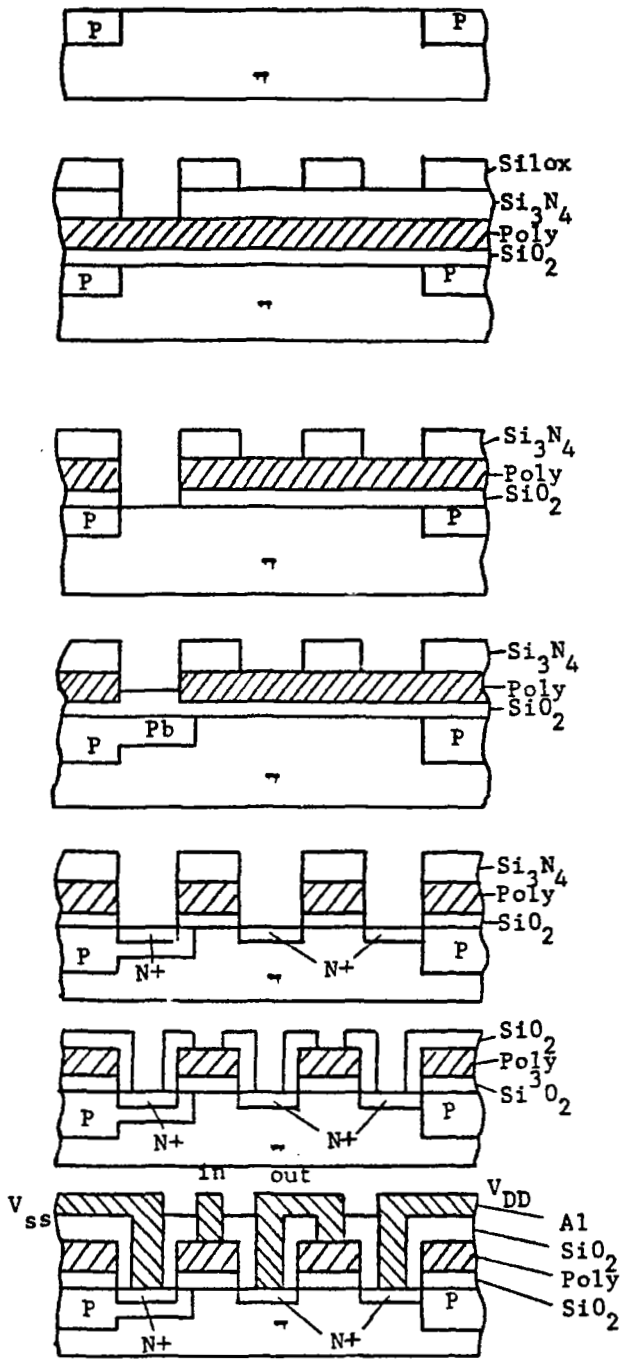


Figure 33. Si Gate DMOS IC Structure

sequence. A moderately doped p-type layer is first diffused into a π-type substrate for isolation. The gate dielectric polycrystalline silicon and silicon nitride layers are sequentially deposited on the substrate. Windows for N^+ source and drain diffusions are cut through the silicon nitride layer. Another photoengraving follows to open a slightly larger window in the photoresist above the source region of the DMOS transistor, permitting the nitride to serve as a mask for removing the polycrystalline silicon and gate oxide.

Next, a p-type channel diffusion penetrates into the source windows. Source windows are oxidized. Plasma etching then removes only the nitride and polycrystalline silicon layers from the drain windows. Now all the source and drain regions are covered only with oxide, which can be readily removed for the n-type diffusion. In this manner, self-alignment is achieved.

For the depletion-mode load device, the channel is masked against the second p-type diffusion. The low π-type background concentration, together with the positive surface staes, is sufficient to make the MOS transistor to be in depletion mode.

The processing parameters of the double diffused MOS are as follows:

a. Substrate: $\langle 1,1,1 \rangle$, p type 10 to 30Ω/cm.
b. Thermal Oxidation: 5000Å thick.
c. Photoengraving for p^+ guard ring.
d. p^+ boron diffusion: 100Ω/square, 5μm deep
e. Oxide removal.
f. Gate dielectric: 1000Å silicon dioxide plus 800Å silicon nitride.
g. Vapor growth of polycrystalline silicon by decomposition of silane: 3000Å.
h. Diffusion mask growth: 2000Å silicon nitride plus 5000Å silicon dioxide.
i. Photoengraving for n^+ source and drain windows through top oxide.
j. Photoengraving for p^+ source windows to substrate.
k. p^+ boron diffusion: 125Ω/square, 4μm deep.
l. Plasma etch of polycrystalline silicon from drain windows and then removal of oxide from both source and drain windows.
m. N^+ phosphorous diffusion: 10 Ω/square, 1.5μm deep.
n. Plasma etch of polycrystalline silicon to delineate discrete polysilicon areas.

o. Photoengraving of contact windows.

p. Metalization: 15,000Å, 70% Al plus 30% silicon.

q. Photoengraving of interconnect pattern.

r. Sintering.

EXPERIMENTAL RESULTS

An experimental self-aligned silicon gate DMOS integrated inverter is shown in figure 34. The DMOS transistor Q_1 has a length of 7 μm and a width of 420 μm. The depletion mode load device Q_2 has the same dimensions. In DMOS integrated circuits, it is possible to design both devices to have the same dimensions (ratioless) because of the higher conductance of the depletion-mode load device. This ratioless feature allows for higher density.

Typical Test Transistor Characteristics

The V-I characteristics of the DMOS transistor and the load device are shown in figure 35. Note that the transconductance of the DMOS transistor at 8 mA of drain current and 5V of drain voltage is approximately 3 mmho. From a theoretical standpoint,[5] the maximum achievable transconductance $g_m(max) = WC v_s$, where W = channel width = 300×10^{-4} cm.

C = gate capacitance per unit area.

= dielectric constant of oxide/gate oxide thickness

= 3.4×10^{-13} (F/cm)/0.15×10^{-4} (cm).

v_s = scattering limited velocity = 6×10^6 cm/s.

The theoretical g_m (max) using our parameters should be about 4 mmhos. Thus, our result is approaching the theoretical value.

In the several lots we fabricated, we had a spread of threshold voltages ranging 1 to 5V. However, the maximum transconductance achieved has been quite consistent. The threshold voltage of the DMOS transistor shown in figure 35 is approximately 3.5V.

The characteristics of a load device is shown in figure 35. It was processed in the same run as the DMOS shown in figure 35. Note that the saturated drain current averages about 0.5 mA. For proper inverter operation, the DMOS "on" current should be larger than the load device current.

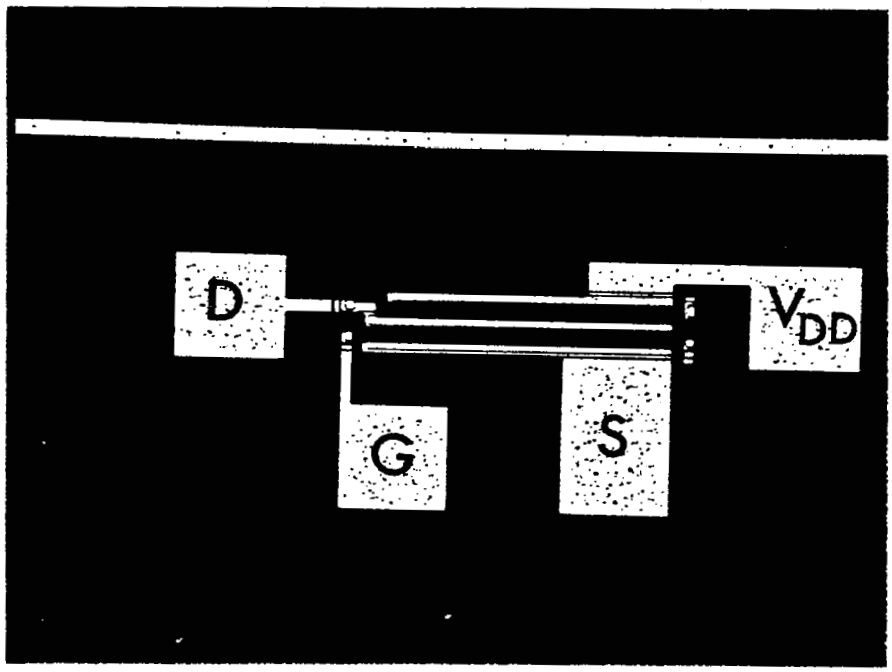


Figure 34. DMOS Integrated Inverter

Because the threshold voltage after fabrication was lower than expected before fabrication, provision was made to reverse bias the substrate to account for process variation, thereby increasing the threshold voltage. The effect of the substrate bias on the load device is to increase the threshold voltage, hence the I_{DSAT}, is shown in figure 36.

Inverter Characteristics

The characteristic of the test inverter is shown in figures 37 and 38. The transfer characteristics for 3 and 5V operation were tested. The effect of the substrate bias was also tested as shown in figure 37 and 38. These characteristics show that the inverter should work under all these conditions. However, the 3V operation with 1V substrate bias appears to give the best results in terms of noise margin.

Effect of Silicon Gate Resistance

The effect of the silicon gate resistance was also investigated. For this experiment, there are two gate contacts as shown in figure 39, one at each end of the gate. The gate length is 7 μm and the width-to-length ratio is 60.

The transistor was mounted in a microwave transistor package and tested with a Hewlett-Packard Network Analyzer. The transistor was tested first with one gate connection, then with both. The S-parameter readings are listed in Table III.

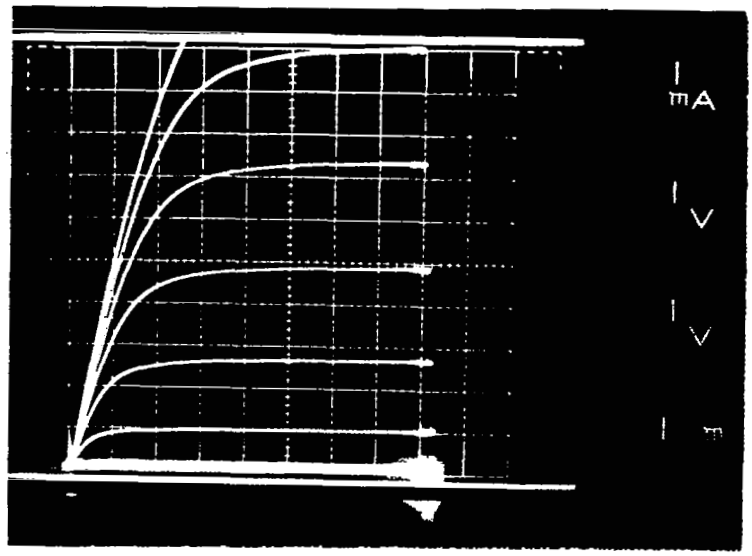
Table III

Experimentally Measured S-Parameters at 100 MHz

	H_{21}	S_{21}	G_1		G_{max}	U	K
1 gate connection	2.1	-21.5	19.1	5.8	3.4	0.28	2.44
2 gate connection	4.3	-14.3	13.5	6.3	5.5	0.36	1.58

From the above dimensional information and the previous processing constants, the RC time constant of the silicon gate is 2.7 nsec. The radian frequency (ω = 1/RC) is 3.5×10^8 radians/sec. According to figures 10, 11 and 12 the gate series resistance effect on amplitude response should become pronounced at 100 MHz, as indeed it is.

It can be seen from this data that the forward current gain, H_{21}, is doubled for the case of two gate contacts. The two gate contacts divide the



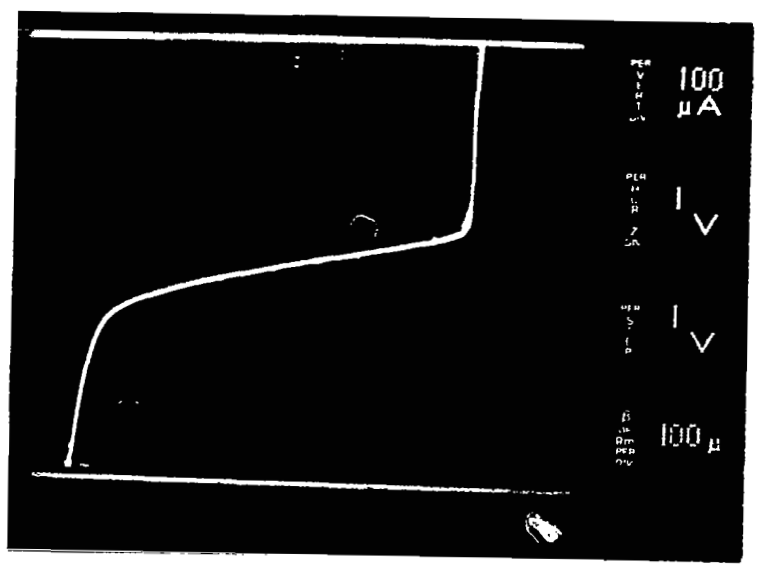


Figure 35. Transistor V-I Characteristic of DMOS Inverter

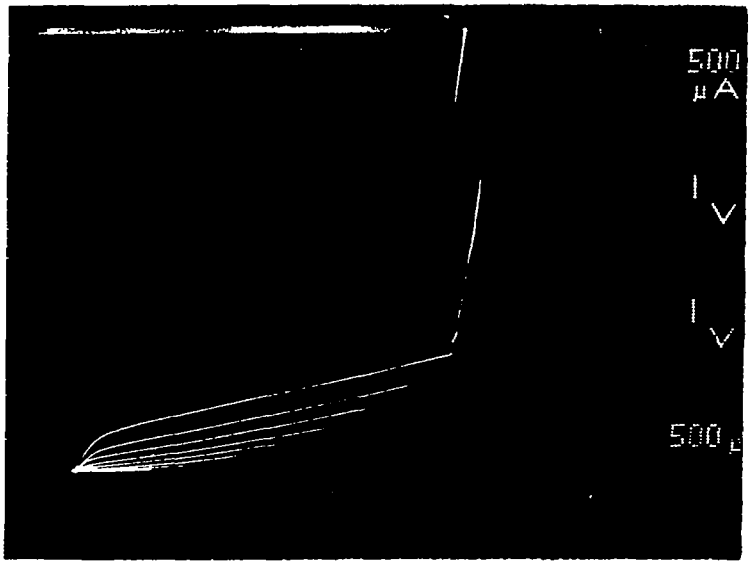
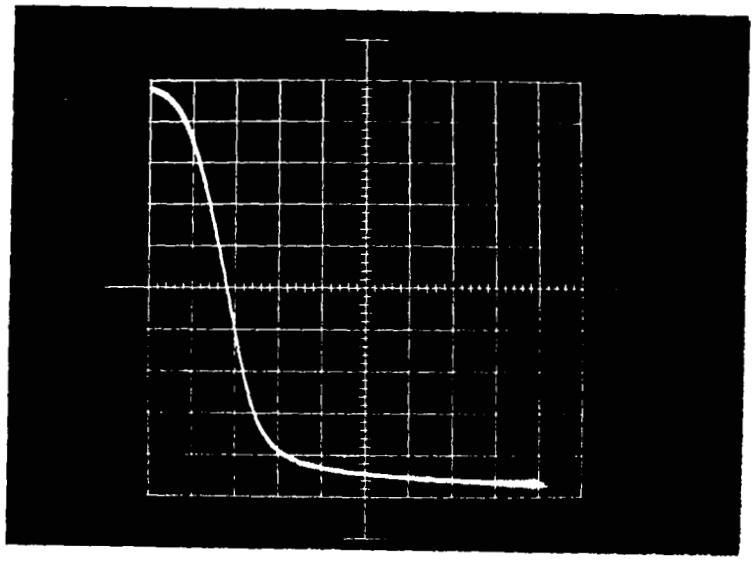


Figure 36. Effect of Substrate Bias on Load Device Characteristics

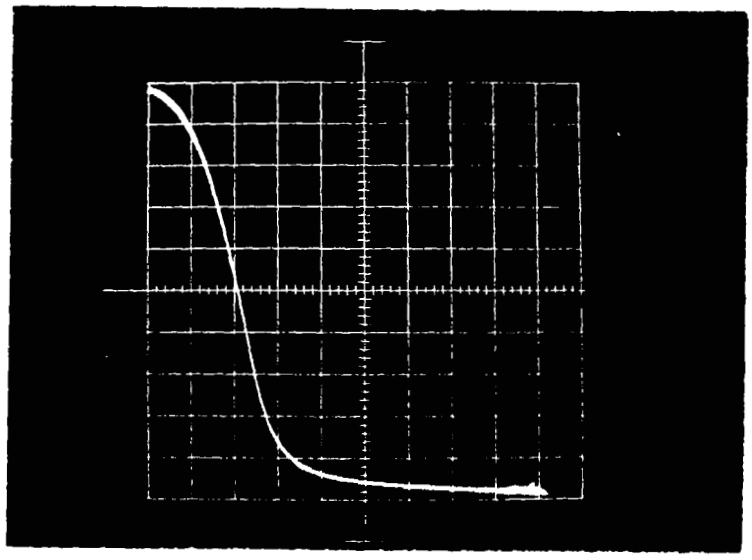


V_{out}

(a) Zero Bias

V_{in}

0.5V/Division

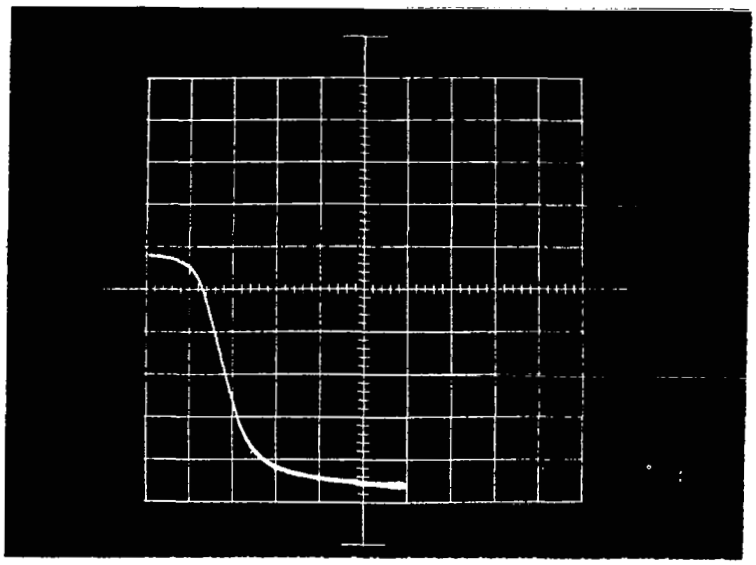


V_{out}

(b) 1V Bias

V_{in}

Figure 37. DMOS Inverter Characteristics in 5V Operation

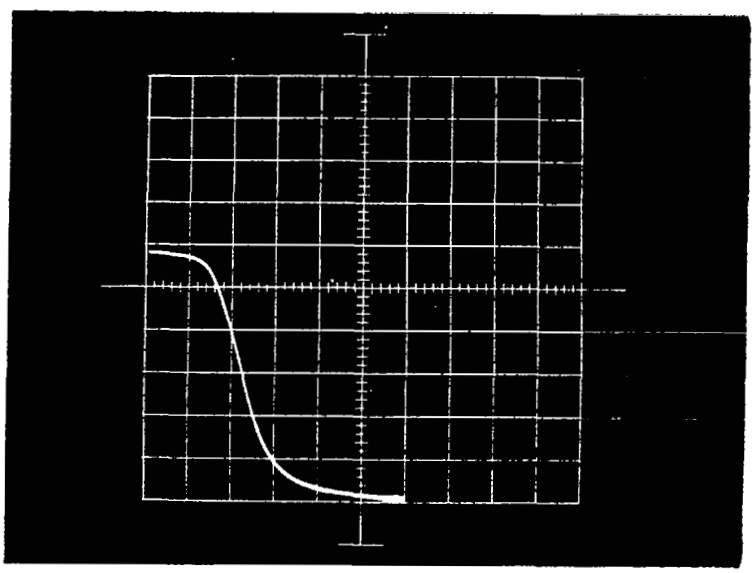


V_{out}

(a) Zero Bias

V_{in}

0.5V/Division



V_{out}

(b) 1V Bias

V_{in}

Figure 38. DMOS Inverter Characteristics in 3V Operation

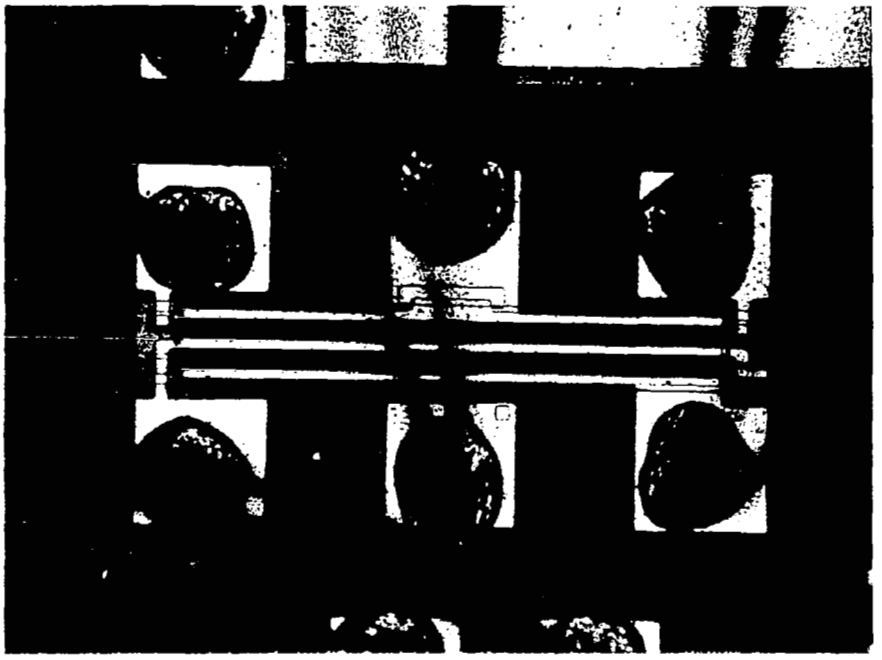


Figure 39. Photomicrograph of DMOS Transistor Showing 2 Gate Contacts

current into two halves and the voltage drop along each part of the distributed gate is then reduced. This effectively higher gate voltage over more of the gate length increases the output drain current, hence H_{21}.

The maximum available power gain, G_{max}, is directly proportional to the transconductance and inversely proportional to the input resistance. With two gate contacts, the input resistances, of which the silicon gate resistance contributes substantially, are effectively halved. Meanwhile, the transconductance is effectively increased because of reduced voltage drop along the silicon gate. Hence, the maximum available gain is increased for two gate contacts. These measurements are therefore indicative of the degrading effect on frequency response by the silicon gate.

Ring Oscillator Performance

A number of DMOS inverters can be connected in cascade as a ring oscillator to test the propagation delay and the speed-power product. The one we tested was a seven-inverter ring oscillator and a buffer as shown in figure 40. The ring oscillator works satisfactorily from 1.3 to 8.5V. The test results are tabulated in Table IV. The oscillation waveforms are shown in figure 41 for supply voltages of 1.3, 3, and 8.5V.

Table IV

Ring Oscillator Performance

	V_B = 0V			V_B = -1V			V_B = -2V			V_B = -3V		
V_{DD}	t_{pd} ns	P_o mW	$t_{pd}P_o$ pJ	t_{pd} ns	P_o mW	$t_{pd}P_o$ pJ	t_{pd} ns	P_o mW	$t_{pd}P_o$ pJ	t_{pd} ns	P_o mW	$t_{pd}P_o$ pJ
1.3V	7.57	.32	2.42									
1.5	7.57	.35	2.68	11.43	.37	4.22						
2				9.29	.51	4.76						
3	7.57	1.09	8.23	8.64	1.01	8.75	11.14	.889	9.90			
4	7.29	1.7	12.38	8.0	1.6	12.8	9.14	1.44	13.16	16.25	1.38	22.43
5	6.93	2.38	16.45	7.5	2.3	17.25	8.36	2.12	17.72	9.93	1.94	19.26
6	6.64	3.34	22.17	7.14	3.3	23.57	7.71	2.93	22.59	8.64	2.74	23.67

The propagation delay, t_{pd} is calculated as follows:

$$t_{pd} = \frac{1}{2nf}$$

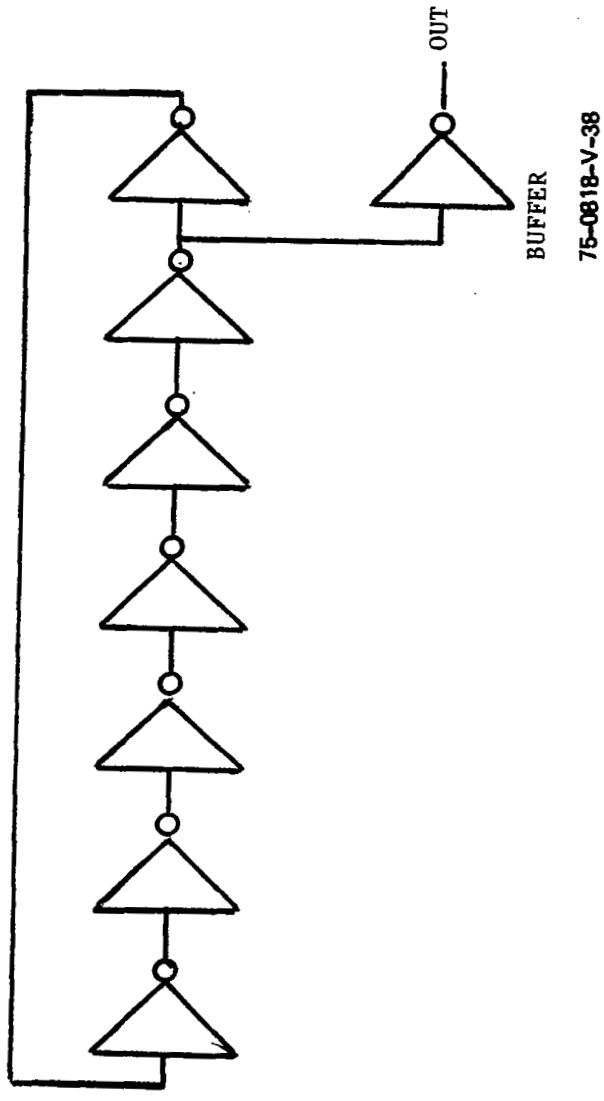
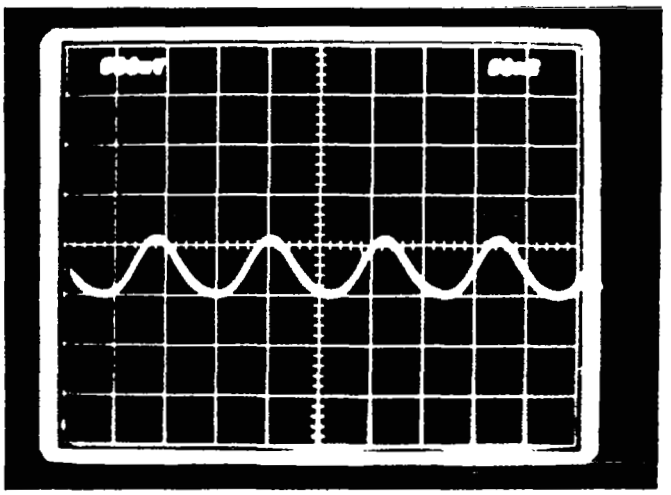
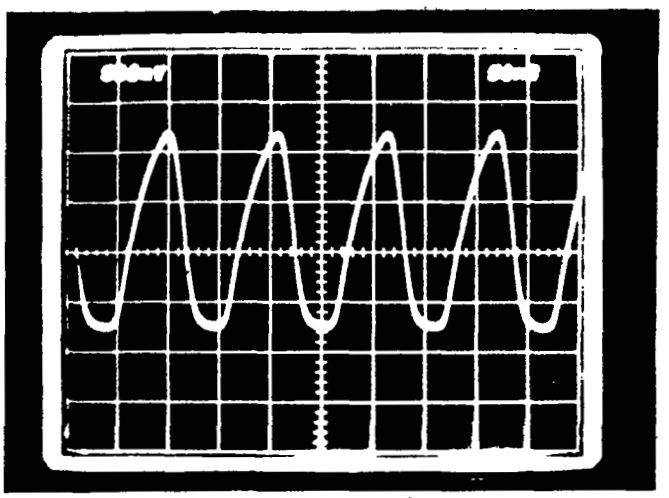


Figure 40. Ring Oscillator



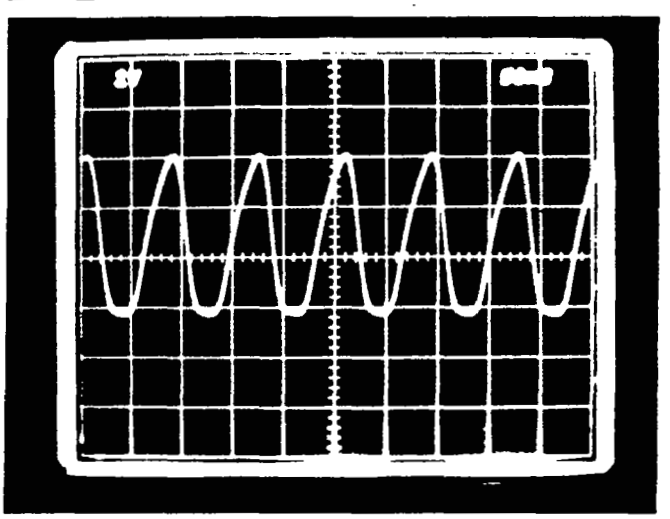
41a

V_{DD} = 1.3V



41b

V_{DD} = 3



41c

V_{DD} = 8.5V

Figure 41. Ring Oscillation

where f is the frequency of oscillation and n is the number of inverters in the ring (n = 7 in our case). The power dissipation per stage P_o is equal to the total power dissipation divided by the number of inverters used in the ring oscillator and the buffer (=8 in our case). The propagation delay-power dissipation product, $t_{pd}P_o$, is a figure of merit of a logic gate. Certain units have t_{pd} as low as 4 nsec.

It should be noted that the propagation delay of 7 nsec is considerably better than the conventional TTL gates. The propagation delay-power dissipation product at a supply of 1.3V is only 2 picojoules which is an order of magnitude lower than for TTL gates. The fact that the DMOS gates can operate as low as 1.3V is another advantage over the TTL gates. It appears that the DMOS gates are better than the TTL gates in most important aspects. Figure 42 is a plot of propagation delay and power dissipation as a function of supply voltage.

SUMMARY AND CONCLUSIONS

The purpose of this work was to study the feasibility of using self-aligned silicon gate technology for DMOS integrated circuits. In the course of the experimental investigation, we have accomplished a number of technological advances:

First, we developed a process for fabricating DMOS integrated circuits using silicon gate technology. Up to now, the DMOS transistor technology has been limited to aluminum gates, which are not self-aligned and therefore require overlap areas with resulting additional feedback capacitance. Our development makes it possible to combine the low parasitic capacitance and facilitated processing advantages of self-aligned gates with the short effective channel length of DMOS to produce a structure with much improved high-frequency performance.

Secondly, we developed a technique to optimize the design of the depletion-mode load device without the need for additional diffusion steps as is normally done in standard MOS processing. This was done by proper choice of the substrate orientation and resistivity. Moreover, our design approach resulted in a load device having the same length-to-width ratio as the DMOS transistor.

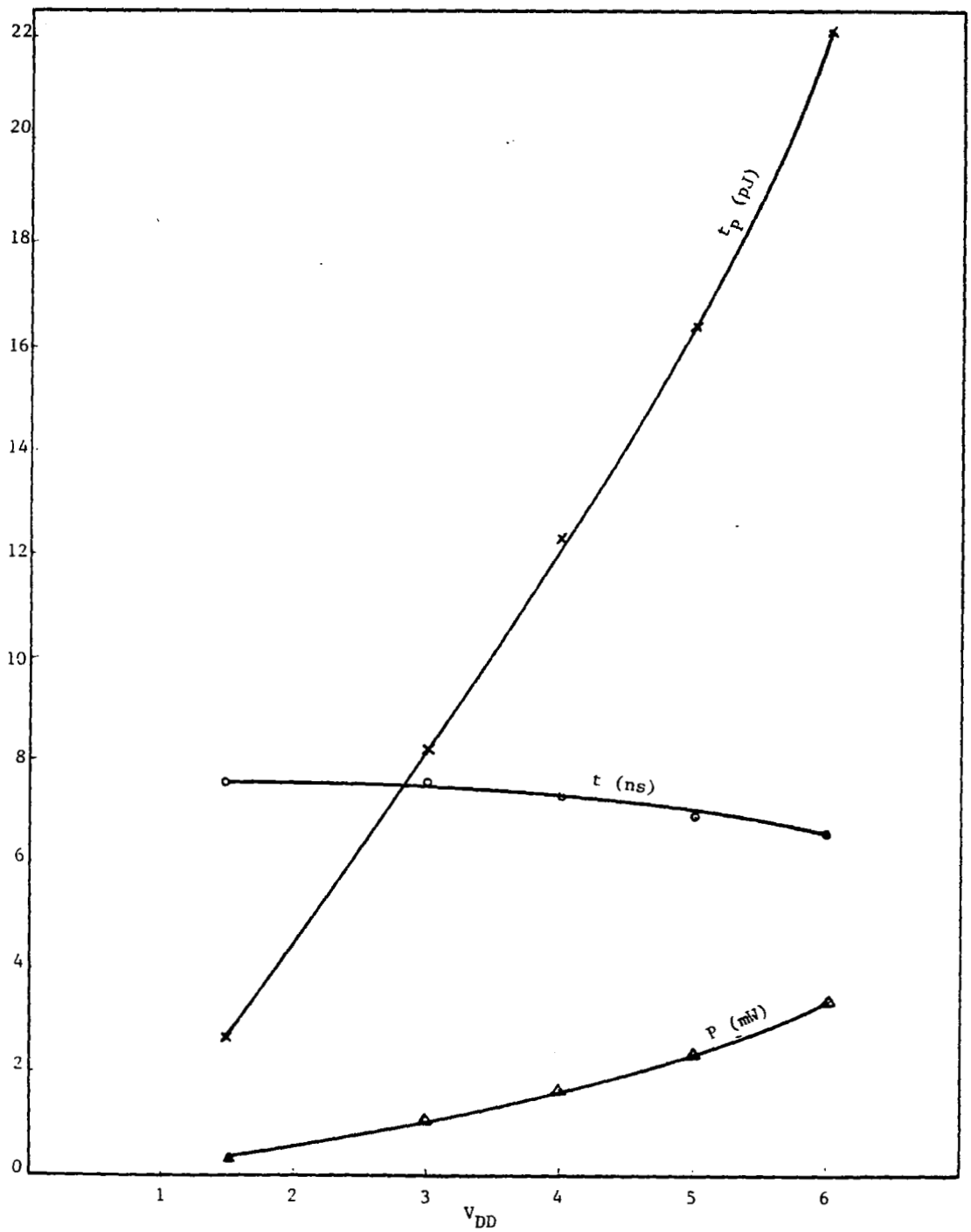


Figure 42. Propagation Time and Power Dissipation of DMOS Inverters

This ratioless geometry greatly reduces the size of an integrated circuit as compared to other approaches where the size of the load device is relatively larger than the MOS transistor.

Thirdly, we developed a novel method of tapering the geometry of the output stages for optimizing the propagation time and the chip area. This method has general application to most clocked MOS LSI integrated circuits. This design approach was incorporated in the design of the DMOS shift register under investigation.

Fourthly, we recognized the effect of the finite silicon gate resistance on the frequency response of an MOS transistor. The resistance creates a distributed RC network along the gate and tends to slow down the transient response. The effect is more pronounced for wider silicon gates (i.e., larger devices). Care should be exercised to minimize this adverse effect on large output devices.

Finally, we successfully fabricated an integrated DMOS ring oscillator with a propagation time of 4 to 7 nanoseconds capable of operating as low as 1.3V and a speed-propagation-power dissipation product of 2 pJ. Such a performance is better than conventional TTL gates. The results are particularly significant in that the ring oscillator transistors were test pattern devices and not optimized for high-speed performance. Further improvement in performance is to be expected by elimination of bonding pad capacitances and by optimizing the transistor geometry.

The results of our investigation have been published in four different articles. The results of the DMOS process development and the tapered output stage concept were presented in separate papers at the 1974 Government Microelectronic Applications Conference (GOMAC) in Boulder, Colorado. An expanded paper on the latter was also published in the IEEE Journal of Solid-State Circuits (April 1975). The effect of silicon gate resistance on the frequency response of MOS transistors appeared in the IEEE Transactions on Electron Devices (May 1975). Publication of the experimental results and design considerations for the integrated DMOS ring oscillator is anticipated for later this year.

APPENDIX

Frequency Response of Distributed RC Line

Small Signal: Consider an incremental RC section as shown in figure 8. The differential equations for such a section are:

$$c \frac{dv}{dt} = \frac{di}{dx} \tag{A1}$$

$$ri = \frac{dv}{dx} \tag{A2}$$

For steady-state sinusoidal voltages and currents,

$$\frac{d^2v}{dx^2} = j\omega\, rcv \tag{A3}$$

$$\frac{di}{dx} = j\omega\, cv \tag{A4}$$

The solutions for these equations are

$$v = Ae^{\gamma x/W} + Be^{-\gamma x/W} \tag{A5}$$

$$i = 1 - \frac{1}{z_o} (Ae^{\gamma e/W} - Be^{-\gamma x/W}) \tag{A6}$$

where A and B are arbitrary constants and

$$\gamma = (j\omega rcW^2)^{1/2} \tag{A7}$$

$$z_o = (r/j\omega c)^{1/2} \tag{A8}$$

For an open ended line as shown in figure 1, the boundary conditions are

$$(i)_{x=W} = 0; \quad (v)_{x=0} = V_i \tag{A9}$$

Substituting into equations (A5) and (A6)

$$\frac{V(x)}{V_i} = \frac{\cos h\gamma\, (1-x/W)}{\cosh\, \gamma} \tag{A10}$$

Transient Response

The Laplace transform of a step input voltage is V_i/s. From equation (A10) the transform of any voltage x is,

$$\phi(x,s) = \frac{V_i \cosh (1-s/W)\, W \sqrt{rcs}}{s \cosh W\sqrt{rcs}} \tag{A11}$$

The inverse transform of this equation is

$$v(x,t) = \frac{1}{2\pi j} \oint \frac{V_i e^{st} \cosh (1-x/W)\, W\sqrt{rcs}}{s \cosh W\sqrt{rcs}}\, ds \tag{A12}$$

The poles for cosh W rcs are given by the roots of the equation

$$W\sqrt{rcs} = j \left(\frac{2n-1}{2}\right) \pi \tag{A13}$$

where n is an integer.

From Cauchy's integral theorem at any pole

$$\oint \frac{P(s)}{D(s)}\, ds = 2\pi j R \tag{A14}$$

where the residue $R = \frac{P(s)}{D'(s)}$ (A15)

Thus the integral (A12) can be evaluated.

$$V(x,t) = V_i \left\{1 + \sum_{n=1}^{\infty} \frac{4(-1)^n}{(2n-1)\pi} \cos \left[\pi\left(\frac{2n-1}{2}\right)\left(1 - \frac{x}{W}\right)\right] \exp \left[-\left(\frac{2n-1}{2}\right) \frac{\pi t}{W^2 rc}\right]\right\} \tag{A16}$$

REFERENCES

1. Tarui, Y.; Hayashi, Y; and Sekigawa, T: "Diffusion Self-Aligned MOST - A New Approach for High Speed Devices," in Proc. 1st Conf. Solid State Devices, Tokyo, 1969.

2. Cauge, T. P.; Kocsis, J.; Sigg, H. J.; and Vendelin, G. D.: "A Double Diffused MOS Transistor Achieves Microwave Gain," Electronics, Vol. 44, Feb. 1971, p. 99.

3. Lin, H. C.; and Jones, W. N.: "Computer Analysis of the Double-Diffused MOS Transistor for Integrated Circuits," IEEE Trans. Electron Devices, ED-20, Mar. 1973, pp. 275-283.

4. Lin, H. C.; and Halsor, J. L.: "Experimental Investigation of a Shielded Complementary Metal-Oxide Semiconductor (MOS) Structure," NASA CR-132456, 1974.

5. Frohman-Bentchkowsky, D; and Vadasz, L.: "Computer-Aided Design and Characterization of Digital MOS Integrated Circuits," IEEE Journal of Solid State Circuits, SC-4, April 1969, pp. 57-64.

6. Lin, H. C.; and Varker, C. J.: "Normally-On Load Device for IGFET Switching Circuits", 1969 NEREM Rec., Vol. 11, Nov. 1969, p. 125.

7. Lin, H. C.; Arzoumanian, F.; Halsor, J. L.; Giuliano, M. N.; and Benz, H.F.: "Effect of Silicon Gate Resistance on the Frequency Response of MOS Transistors," IEEE Trans. on Electron Devices, ED-22, May 1975, pp. 255-265.

8. Engeler, W. E.; and Brown, D. M.: "Performance of Refractory Metal Multilevel Interconnection System", IEEE Transactions on Electron Devices, ED-19, Jan. 1972, pp. 54-61.

9. Lin, H. C.; and Linholm, L.: "Optimized Output Stage for MOS Integrated Circuits", IEEE Journal of Solid-State Circuits, SC-10, April 1975, pp. 106-109.

10. NASA Tech Brief, "Polycrystalline Silicon Delineation with Application to a Double-Diffused MOS Transistor", (LARS 11536 and 11598).

11. Ohta, K.; et al,: "A High-Speed Logic LSI Using Diffusion Self-Alignment Depletion MOST", Digest of Solid-State Circuits Conference, XVIII, Feb. 1975, pp. 124-125.