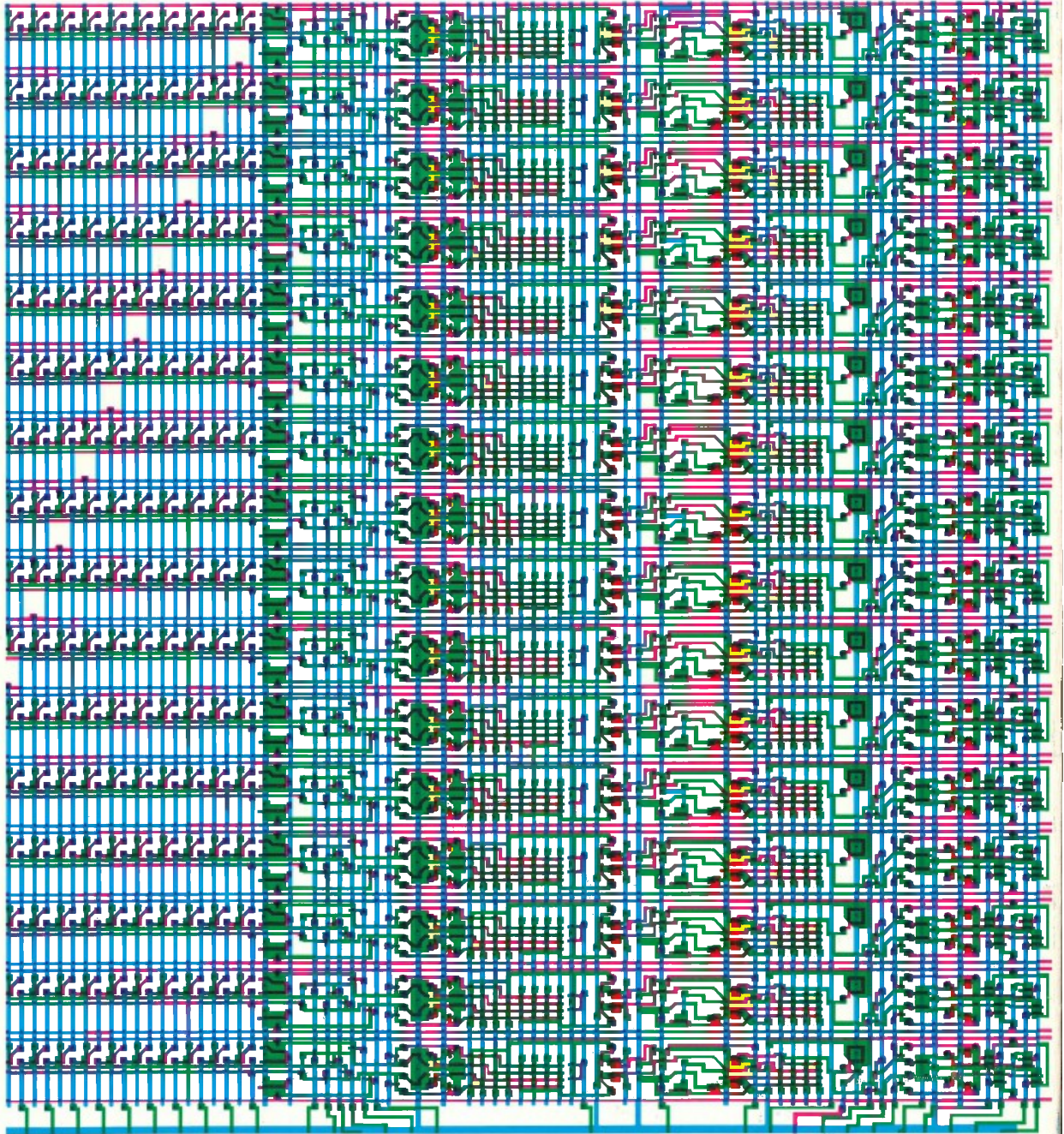


INTRODUCTION TO **VLSI** SYSTEMS

CARVER MEAD • LYNN CONWAY



This book is in the
Addison-Wesley Series in Computer Science

Consulting Editor
Michael A. Harrison

Library of Congress Cataloging in Publication Data

Mead, Carver A
Introduction to VLSI systems.

1. Integrated circuits—Large scale integration.
2. Microcomputers. 3. Digital electronics.
4. Computer architecture. I. Conway, Lynn A.,
joint author. II. Title.

TK7874.M37
ISBN 0-201-04358-0

621.3819'535

78-74688

Second printing, October 1980

Copyright © 1980 by Addison-Wesley Publishing Company, Inc. Philippines copyright 1980 by Addison-Wesley Publishing Company, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. Published simultaneously in Canada. Library of Congress Catalog Card No: 78-74688.

ISBN 0-201-04358-0
JKLMNOP-HA-898765

of MOS transistors. Figure 1.38(a) shows again a physical cross section through an MOS transistor, as a reference for the following figures. Part (b) reviews the case of equal source and drain potentials with the channel turned on fairly strongly, thus readily allowing charge to move between source and drain. Part (c) shows the situation when a small voltage difference, ΔV , has been applied between source and drain. Since the potential difference is maintained by external voltage sources, electrons will be forced to move from source to drain under the influence of the potential gradient, just as a liquid would flow from the higher to the lower level.

As the potential difference between source and drain is made larger, the variation in the "depth" of the fluid along the channel becomes significant (Fig. 1.38d). Continuity in the fluid requires that the charge move faster in the areas where the layer is thinner. This implies that the potential increases more rapidly closer to the drain region. With increasing drain potential the amount of charge flowing from source to drain per unit time increases, since the product of charge-layer depth and local gradient increases. However, there is a limit; once the drain potential exceeds the empty channel potential, the rate-of-charge flow will be limited by the drain-side edge of the barrier under the gate electrode. The MOS transistor has now reached saturation (Fig. 1.38e). The drain current density now is determined by the potential difference between the source and the empty channel and by the length of the channel (or the width of the barrier over which the charge has to flow); it is to first order independent of the drain voltage V_{db} .

Even in simple transistor circuits, the above fluid model helps one quickly develop a feeling for device and circuit operation. However, the real power of this intuitive model emerges when it is applied to complex structures where closed-form solutions describing charge motion can no longer be found. The empty potential under the various electrodes can first be plotted as in the above examples and the flow of charge then visualized using the analogy to the behavior of a fluid.

1.16 EFFECTS OF SCALING DOWN THE DIMENSIONS OF MOS CIRCUITS AND SYSTEMS

This section examines the effects on major system parameters resulting from scaling down all dimensions of an integrated system, including those vertical to the surface, by dividing them by a constant factor α . The voltage is likewise scaled down by dividing by the same constant factor α . Using this convention, all electric fields in the circuit will remain constant. Thus many nonlinear factors affecting performance will not change as they would if a more complex scaling were used.

Figure 1.39(a) shows a MOSFET of dimensions L, W, D , with $(V_{gs} - V_{th}) = V$. Figure 1.39(b) shows a MOSFET similar to that in part (a), but of dimensions $L' = L/\alpha, W' = W/\alpha, D' = D/\alpha$, and $V' = V/\alpha$. Refer now to Eqs. (1-1), (1-2), and (1-3). From these equations we will find that as the scale down factor α is increased, the transit time, the gate capacitance, and drain-to-source current of

every individual transistor in the system scale down proportionally, as follows:

$$\begin{aligned} \tau &\propto L^2/V, & \tau'/\tau &= [(L/\alpha)^2/(V/\alpha)]/[L^2/V], & \text{therefore, } \tau' &= \tau/\alpha; \\ C &\propto LW/D, & C'/C &= [(L/\alpha)(W/\alpha)/(D/\alpha)]/[LW/D], & \text{and } C' &= C/\alpha; \\ I &\propto WV^2/LD, & I'/I &= [(WV^2/\alpha^3)/(LD/\alpha^2)]/[WV^2/LD], & \text{and } I' &= I/\alpha. \end{aligned}$$

Switching power, P_{sw} , is the energy stored on the capacitance of a given device divided by the clock period (time between successive charging and discharging of the capacitance). A system's clock period is proportional to the τ of its smallest devices. As devices are made smaller and faster, the clock period is proportionally shortened. Also, the d.c. power, P_{dc} , dissipated by any static circuit equals I times V . Therefore, P_{sw} and P_{dc} scale as follows:

$$\begin{aligned} P_{sw} &\propto CV^2/\tau \propto WV^3/DL & \text{and} & & P'_{sw} &= P_{sw}/\alpha^2; \\ P_{dc} &= IV & \text{and} & & P'_{dc} &= P_{dc}/\alpha^2. \end{aligned}$$

Both the switching power and static power per device scale down as $1/\alpha^2$. The average d.c. power for most systems can be approximated by adding the total P_{sw} to one half of the d.c. power that would result if all level-restoring logic pull-downs were turned on. The contribution of pass-transistor logic to the average d.c. power drawn by the system is due to the switching power consumed by the driving circuits that charge and discharge the pass-transistor control gates.

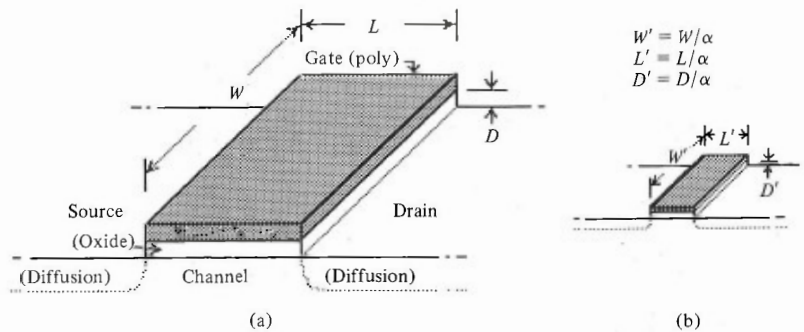


Fig. 1.39 (a) MOSFET, 1978. (b) MOSFET scaled down by Alpha, 19XX.

The switching energy per device, E_{sw} , is an important metric of device performance. It is equal to the power consumed by the device at maximum clock frequency multiplied by the device delay, and it scales down as follows:

$$E_{sw} \propto CV^2 \quad \text{and} \quad E'_{sw} = E_{sw}/\alpha^3.$$

Table 1.1 summarizes values of the important system parameters for current technology and for a future technology near the limits imposed by physical law:

	1978	19XX
Minimum feature size	6 μm	0.3 μm
τ	0.3 to 1 ns	≈0.02 ns
E_{sw}	≈10 ⁻¹² joule	≈2 × 10 ⁻¹⁶ joule
System clock period	≈30 to 50 ns	≈2 to 4 ns

A more detailed plot of the channel conductance of an MOS transistor near the threshold voltage is shown in Fig. 1.40. Below the nominal threshold, the conductance ($1/R$) is not in reality zero but depends on gate voltage and temperature as follows:

$$1/R \propto e^{(V_{gs} - V_{th})/(kT/q)},$$

where T is the absolute temperature, q is the charge on the electron, and k is Boltzmann's constant. At room temperature, $kT/q \approx 0.025$ volts. At present threshold voltages, as in the rightmost curve in Fig. 1.40, an off device is below threshold by perhaps $20 kT/q$, that is, by about 0.5 volts, and its conductance is decreased by a factor of the order of ten million. Said another way, if the device is used as a pass transistor, a quantity of charge that takes a time τ to pass through the on device, will take a time on the order of $10^7 \tau$ to "leak" through the off device.

The use of pass-transistor switches to isolate and "dynamically store" charge on circuit nodes is common in many memory applications using 1978 transistor dimensions. However, if the threshold voltage is scaled down by a factor of perhaps 5, as shown in the leftmost curve in Fig. 1.40, then an off transistor is only $4kT/q$ below threshold. Therefore, its conductance when off is only a factor of 100

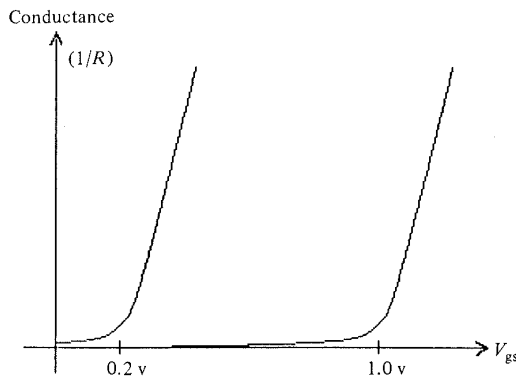


Fig. 1.40 Conductance as a function of threshold voltage.

or so less than when it is on. For such relatively large values of subthreshold conductance, charge stored dynamically on a circuit node by the transistor when on will safely remain on that node for only a few system clock periods. The charge will not remain on the node for a very large number of periods as it does in present memory devices using this technique. One way of possibly coping with this problem, as device dimensions and threshold voltages are scaled down, is to reduce the temperature of device operation.⁹

Suppose we scale down an entire integrated system by a scale-down factor of $\alpha = 10$. The resulting system will have one hundred times the number of circuits per unit area. The total power per unit area remains constant. All voltages in the system are reduced by the factor of 10. The current supplied per unit surface area is increased by a factor of 10. The time delay per stage is decreased by a factor of 10. Therefore, the power-delay product decreases by a factor of 1000.

This is a rather attractive scaling in all ways except for the current density. The delivery of the required average d.c. current presents an important obstacle to scaling. This current must be carried to the various circuits in the system on metal conductors, in order that the voltage drop from the off-chip source to the on-chip subsystems will not be excessive. Metal paths have an upper current density limit imposed by a phenomenon called metal migration (discussed further in Chapter 2). Many metal paths in today's integrated circuits are already operated near their current density limit. As the above type of scaling is applied to a system, the conductors get narrower but still deliver the same current on the average to the circuits supplied by them.

Therefore, it will be necessary to find ways of decreasing system current requirements to approximately a constant current per unit area relative to present current densities. In n -channel silicon gate technology, this objective can be partially achieved by using pass-transistor logic in as many places as possible and avoiding restoring logic except where it is absolutely necessary. Numerous examples of this sort of design are given later in this text. This design approach also has the advantages of tending to minimize delay per unit function and to maximize logic functions per unit area. However, when scaled down to submicron size, the pass transistors will suffer from the subthreshold current problem. It is possible that when the fabrication technologies have been developed to enable scaling down to submicron devices, a technology such as complementary MOS, which does not draw any d.c. current, may be preferable to the n MOS technology used to illustrate this text. However, even if this occurs, the methodology developed in the text can still be applied in the design of integrated systems in that technology.

The limit to the kind of scaling described above occurs when the devices created are no longer able to perform the switching function. To perform the switching function, the ratio of transistor on-to-off conductance must be $\gg 1$, and therefore the voltage operating the circuit must be many times kT/q . For this reason, even those circuits optimized for operation at the lowest possible supply voltages still require a VDD of ≈ 0.5 volts. Devices in 1978 operate with a VDD of

approximately five volts and minimum channel lengths of approximately six microns. Therefore, the kind of scaling we have envisioned here will take us to devices with approximately one-half micron channel lengths and current densities approximately ten times what they are today. Power per unit area will remain constant over that range. Smaller devices might be built but must be used without lowering the voltage any further. Consequently the power per unit area will increase. Finally, there appears to be a fundamental limit¹⁰ of approximately one-quarter micron channel length, where certain physical effects such as the tunneling through the gate oxide and fluctuations in the positions of impurities in the depletion layers begin to make the devices of smaller dimension unworkable.

REFERENCES

1. T. K. Young and R. W. Dutton, "MINI-MSINC: A Minicomputer Simulator for MOS Circuits with Modular Built-in Model," Stanford Electronics Laboratories, Technical Report No. 5013-1, March 1976.
2. L. Nagel and D. Pederson, "Simulation Program with Integrated Circuit Emphasis (SPICE)," 16th Midwest Symposium on Circuit Theory, Waterloo, Ontario, April 12, 1973.
3. W. M. Penney and L. Lau, eds., *MOS Integrated Circuits*, Princeton, N.J.: Van Nostrand, 1972, pp. 60-85.
4. R. C. Jaeger, "Comments on 'An Optimized Output Stage for MOS Integrated Circuits'," *IEEE J. Solid-State Circuits*, June 1975, pp. 185-186.
5. T. Kilburn; D. B. G. Edwards; and D. Aspinall, "A Parallel Arithmetic Unit Using a Saturated Transistor Fast-Carry Circuit," *Proc. IEE*, Pt. B, vol. 107, November 1960, pp. 573-584.
6. Staff of the Computation Lab, "Description of a Relay Calculator," *Annals of the Harvard Computation Lab*, vol. 24, Harvard University Press, 1949.
7. T. J. Chaney and C. E. Molnar, "Anomalous Behavior of Synchronizer and Arbiter Circuits," *IEEE Transactions on Computers*, April 1973, pp. 421-422.
8. C. H. Séquin and M. F. Tompsett, *Charge Transfer Devices*, New York: Academic Press, 1975.
9. F. H. Gaensslen; V. L. Rideout; E. J. Walker; and J. J. Walker, "Very Small MOSFETs for Low-Temperature Operation," *IEEE Transactions on Electron Devices*, March 1977.
10. B. Hoeneisen, and C. A. Mead, "Fundamental Limitations in Micro-electronics-I. MOS Technology," *Solid-State Electronics*, vol. 15, 1972, pp. 819-829.