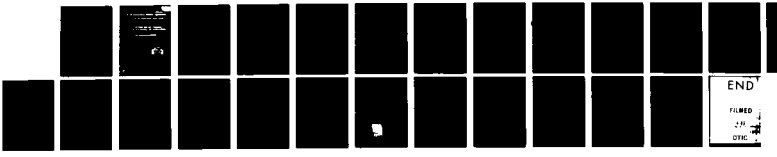
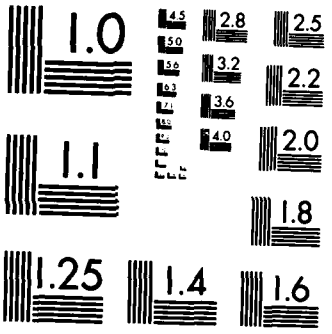
AD-A125 835 A RRITIQUE AND AN APPRAISAL OF VLSI MODELS OF COMPUTATION(U) ILLINOIS UNIV AT URBANA APPLIED COMPUTATION THEORY GROUP G BILARDI ET AL. AUG 81 1/1

UNCLASSIFIED ACT-28 N00014-79-C-0424 F/G 12/1 NL

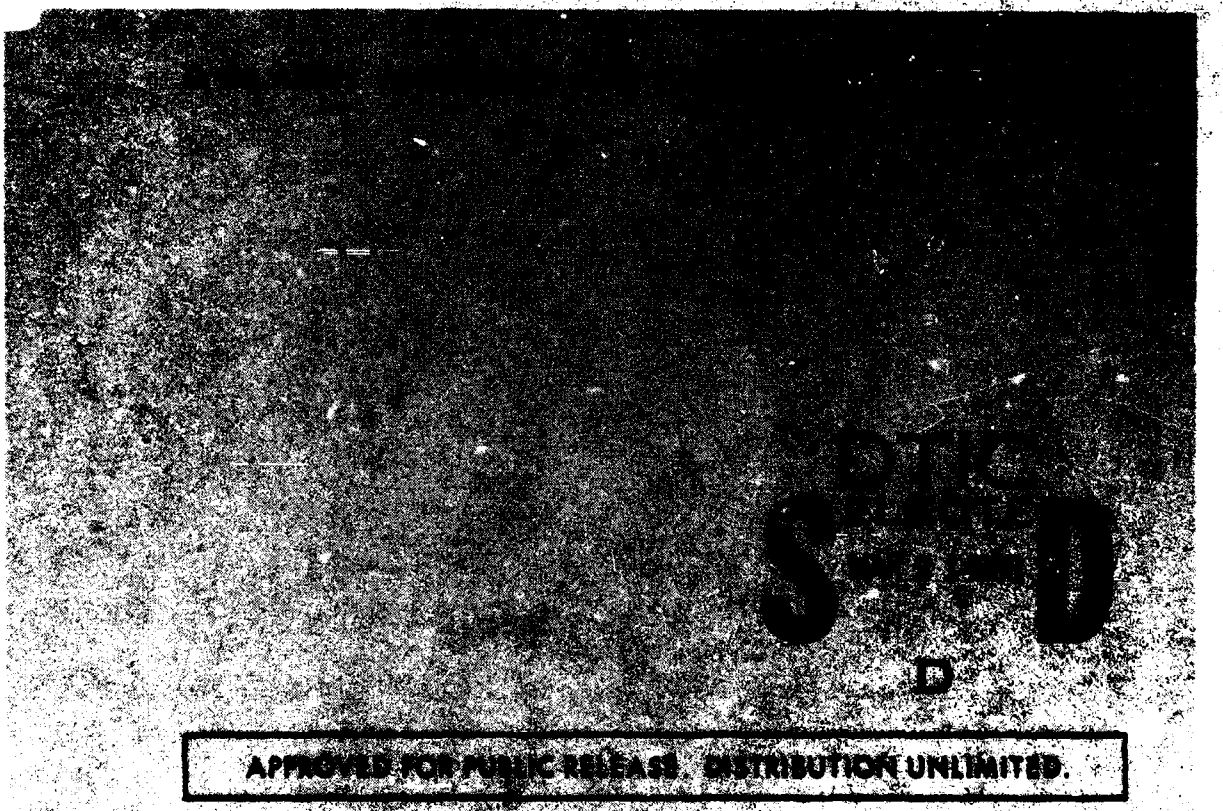




COORDINATED SCIENCE LABORATORY

APPLIED COMPUTATION THEORY GROUP

A CRITIQUE AND AN APPRAISAL OF VLSI MODELS OF COMPUTATION



REPORT R-914 UILU-ENG 81-2245

UNIVERSITY OF ILLINOIS - URBANA, ILLINOIS

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Critique and an Appraisal of VLSI Models of Computation		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER R-914; UILU-ENG 81-2245;ACT-28
7. AUTHOR(s) G. Bilardi, M. Pracchi and F.P. Preparata		8. CONTRACT OR GRANT NUMBER(s) MCS-81-05552 (NSF) N00014-79-C-0424 (JSEP)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Coordinated Science Lab, 1101 W. Springfield Ave. University of Illinois at Urbana-Champaign Urbana, Illinois 61801		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS National Science Foundation; Joint Services Electronics Program		12. REPORT DATE August 1981 13. NUMBER OF PAGES 21
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

VLSI models of computation, propagation delay, MOS technology, dispersive lines, RC-lines, diffusion equation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this paper we evaluate various proposed VLSI models of computation. While there is consensus on the appraisal of chip area, controversy remains with regard to computation time. Thus we have analyzed in detail the propagation of signals on dispersive lines. The results are expressed in terms of adimensional parameters characteristic of any given fabrication technology. The conclusion is that both current and projected silicon technologies fall within the realm of the capacitive model, where a dispersive line can be replaced by a capacitance proportional to its length. Diffusion phenomena appear therefore to exceed the present VLSI horizon.

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

A CRITIQUE AND AN APPRAISAL OF VLSI MODELS OF COMPUTATION

G. Bilardi, M. Pracchi, and F. P. Preparata, Fellow, IEEE
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign

Abstract

In this paper we evaluate various proposed VLSI models of computation. While there is consensus on the appraisal of chip area, controversy remains with regard to computation time. Thus we have analyzed in detail the propagation of signals on dispersive lines. The results are expressed in terms of adimensional parameters characteristic of any given fabrication technology. The conclusion is that both current and projected silicon technologies fall within the realm of the capacitive model, where a dispersive line can be replaced by a capacitance proportional to its length. Diffusion phenomena appear therefore to exceed the present VLSI horizon.

This work was supported in part by National Science Foundation Grant MCS-81-05552 and by the Joint Services Electronics Program Contract N00014-79-C-0424.

A CRITIQUE AND AN APPRAISAL OF VLSI MODELS OF COMPUTATION

1. Introduction

The central question in the design and analysis of algorithms is the definition of the model of computation to be adopted. Indeed, "performance" becomes meaningful only in relation to a given model. This model is normally the simplified abstraction of a class of real or imaginary machines; for example, the RAM or Random-Access-Machine, is the model of practically the totality of existing (Von Neumann type) processors. The model of computation is the simplest possible, compatibly with the requirement of being realistic. In other words, while a model aims at capturing the essential traits of a system or technology, its simplicity is what enables theoretical appraisals of performance.

Very-Large-Scale-Integration (VLSI), as a computing environment, is no exception. Indeed considerable attention has been paid [1][2][3][4] to the definition of a suitable model. The basic parameters of any VLSI computation model are chip area A and computation time T. VLSI systems display a trade-off between these two parameters, each of which represents a well-defined cost aspect: chip area is a measure of fabrication cost and computation time is a measure of operating cost.

A general feature of all proposed – and presumably of all future – VLSI models of computation is that a chip is viewed as a computation graph, whose vertices are called nodes and whose arcs are called wires. Nodes are, by and large, devices and are responsible for information processing (computations of boolean functions); wires are just electrical connections, and are responsible for both transfer of information and distribution of power

supply and timing waveforms.

A given computation graph is to be laid-out in conformity with the rules dictated by technology. These rules are geometric constraints on admissible layouts and typically concern widths of wires and transistor regions, clearances between wires, transistors, etc., number of metallic layers, permissible orientations, etc.. Once a layout — that is, a legal planar embedding of the computation graph — has been produced, the chip area A is normally the area of the smallest rectangle inscribing the layout, and is the sum of the areas of wires, transistors, and, possibly, of some wasted space. More formally we have:

Area Assumptions

A1. (Wire area) All wires have minimum width $\lambda > 0$ (which includes both the actual wire width and the clearance between wire and any other chip region) and at most $v \geq 2$ wires can overlap at any point (hypothesis of bounded number of layers). [All models.]

A2. (Transistor-port area) Transistors and I/O ports have minimum area $\geq \lambda^2$. [All models.]

A2.1 Transistors and I/O ports have fixed area $c_T\lambda^2$ and $c_P\lambda^2$, respectively, for constants c_T and c_P [Brent-Kung [2]; Chazelle-Monier [4]].

A2.2 The chip is subdivided into compact regions, called "self-timed"; within a self-timed region A2.1 holds, while drivers of inter-region wires have area proportional to the wire-length [Thompson [3]; Seitz [5]].

A3. (Chip area) The chip area A is at least the sum of the area of the wires, of the transistors, and of the I/O ports, and it is at most the area of the smallest rectangle (or convex region) enclosing a legal layout of the graph. [All models.]

These rules are quite simple and uncontroversial. Indeed no difficulty arises in appraising the area of a given computation graph.

Radically different — as to a consensus among researchers — is the situation regarding the computation time T. To acquire the necessary perspective, let us call "an elementary action" the change of output of a transistor and the transmission of this change on the wires connected to this output. Thus, given a computation graph — which supports a prescribed algorithm — the designer can describe the execution of the algorithm as a sequence of sets of elementary actions. In other words, execution is conveniently modeled by a single-source/single-destination (corresponding to begin and end, respectively) directed acyclic graph, whose arcs correspond to elementary actions. Each arc is weighted with the time taken by the action it represents. This knowledge, in principle, seems quite adequate for the evaluation of T, by simply taking the value of the most time-consuming source-destination path in the acyclic graph. The difficulty lies, however, in the assignment of values to the arc weights. Indeed, the proposed computational models basically differ in this weight assignment. More formally we have:

Time Assumptions

T1. (Propagation time along a wire).

T1.1 A bit requires a constant time τ to propagate along a wire, irrespectively of its length. (Brent-Kung). (We refer to this case as the synchronous model.)

T1.2 A bit requires a time $O(\log \ell)$ to propagate along a wire of length ℓ (Mead-Conway; Thompson). (We refer to this case as the capacitive model.)

T1.3 A bit requires a time $O(\ell^2)$ to propagate along a wire of length ℓ (Seitz; Chazelle-Monier). (We refer to this case as the diffusion model.)

T2. (Algorithm time) The computation time of an algorithm is the time of the longest sequence of wire propagation times between beginning and completion of the computation. [All models.]

The choices for T1 reflect the profound controversy on VLSI computation time. In a preliminary analysis, one is tempted to conclude that T1.3 is the most realistic choice. Indeed, a wire is characterized by a resistance and a capacitance which (in a given fabrication technique) both grow linearly with the wire length; therefore, the time constant of the transistor load grows proportionally to ℓ^2, whence the conclusion T1.3. Notice that the computational implications of T1.3 — as noted by Chazelle-Monier in [4] — are drastic. Indeed, chip wires of substantially different lengths are ruled out and connections must exist only between devices in very close proximity. As a consequence, the only permissible computation graphs are of the mesh type (or closely related), which rules out very fast parallel computation, such as performed by computing structures of the type of the shuffle-exchange [6], the cube-connected-cycles [7], or the tree-connected machine [8].

Asymptotically, the line of arguments sketched above is unimpeachable, and therefore — for the theoretician of algorithmics — valid, since asymptotic analysis is the cornerstone of concrete computational complexity. However, the asymptotics of VLSI have a much closer horizon than, for example, the asymptotics of the Turing machine. This horizon, in fact, is set by realistic bounds on the expectations — in the current technology — of minimum feature size and maximum chip size.

Within this horizon, the line parameters must be weighed against the nonnegligible output impedance of the driving transistor and the input impedance of the driven transistor. To appraise this interaction, it is therefore appropriate to take a critical look at the actual physical phenomena occurring during an "elementary action".

2. A mathematical model of wire switching

Perhaps the most characteristic feature of present-day VLSI technology is the fact that, irrespective of the choice of the devices (MOS-FET versus bipolar, for example) wires are realized as dispersive lines. This nature of wires is what determines the time behavior of networks (and must be reflected in the computation model) and the choice of devices, or of their operating regimes, has a nonessential effect on it. Therefore, with reference to dispersive line VLSI technology, any reasonable device selection is representative of the general problem.

In particular, we shall carry out our analysis with reference to the CMOS technology [9]. In figure 1a we have illustrated the circuit being considered. T_1 is an n-channel MOS transistor, initially cut-off. Its drain load — that is the wire AB and the gate capacitance of the driven transistor T_2 — is initially charged to voltage V_0. So, with reference to (I_{DS},V_{DS}) characteristic curves of figure 1b, P_1 is the initial operating point of T_1. At $t = 0$ a step voltage $v_g = V_0$ is applied at the gate of T_1; after a time τ_0 — negligible with respect to the other intervening times — the current I_0 corresponding to $v_g = V_0$ is established and the operating point moves to P_2. From this point on, the operating point moves on the $v_g = V_0$ curve towards the origin and the transistor load discharges through the channel. It is our objective to analyze this phenomenon.

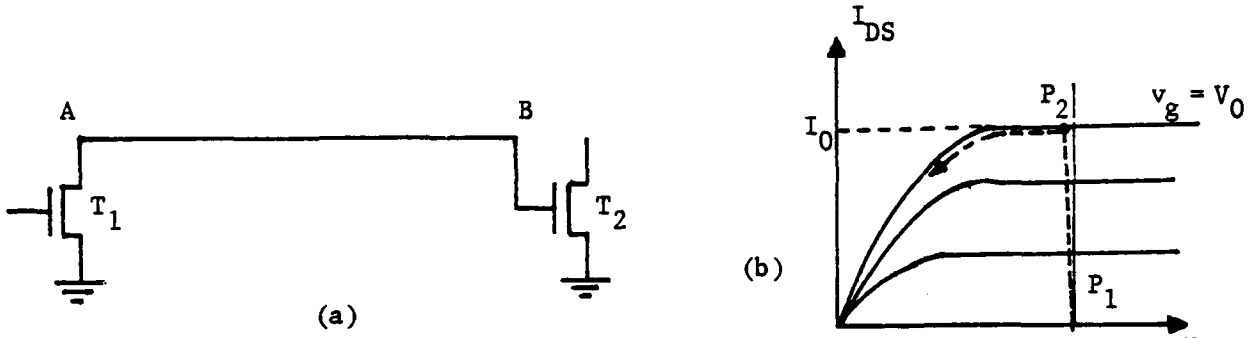


Figure 1. The CMOS configuration and the transistor characteristics.

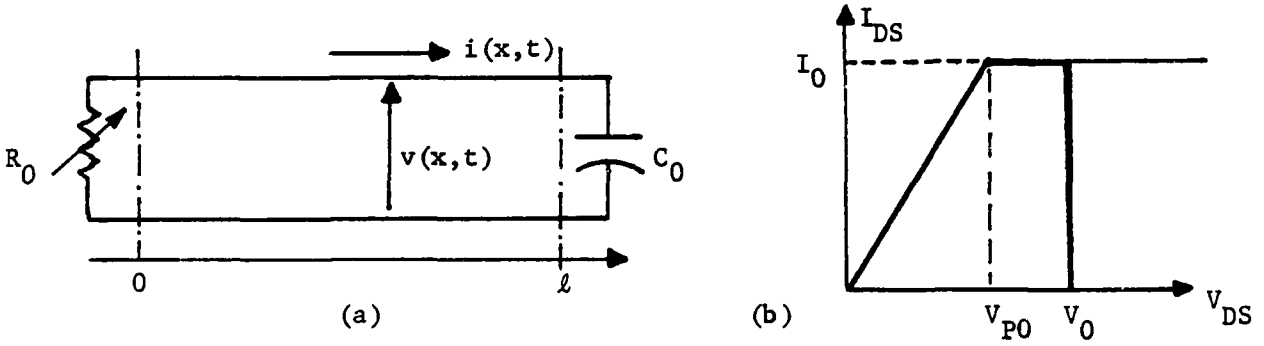


Figure 2. The model (a) and the idealized characteristics.

The circuit is modeled as in figure 2a, where C_0 is the gate capacitance of T_2 and the line, of length ℓ, has resistance r and capacitance c per unit of length. Transistor T_1 is modeled as a (variable) resistor R_0, to reflect the shape of the $v_g = V_0$ characteristic curve. In particular, we approximate the latter as in figure 2b with two straight line segments, meeting at the pinch-off voltage V_{PO}; the <u>saturated</u> regime is modeled by a horizontal segment starting at $(V_{DS}, I_{DS}) = (V_0, I_0)$, while the so-called <u>ohmic</u> regime is modeled by a segment passing through the origin. We shall now study the general discharge regime, and later specialize it to the two regimes defined above.

2.1 General solution.

Let $v(x,t)$ and $i(x,t)$ denote the values of the line voltage and current at abscissa x and time t, respectively. From Ohm's law and the definition of capacitance we obtain

$$\frac{\partial v}{\partial x} = -ri, \qquad \frac{\partial i}{\partial x} = -c\,\frac{\partial v}{\partial t}\,,$$

whence

$$\frac{\partial^2 v}{\partial x^2} = rc\,\frac{\partial v}{\partial t}, \quad \frac{\partial^2 i}{\partial x^2} = rc\,\frac{\partial i}{\partial t}\,. \tag{1}$$

These are instances of the classical diffusion equation (or heat equation), which has been assiduously studied over the past century. It seem natural to suspect that we are dealing with a standard textbook problem. However, our boundary conditions deserve special attention.

We assume that the initial conditions be provided by

$$v(x,0) = v_0(x), \quad x \in [0,\ell] \tag{2}$$

(or, alternatively, $i(x,0) = i_0(x)$) where $v_0(x)$ is an arbitrary function, while the boundary conditions at $x = 0$ and $x = \ell$ are supplied by the nature of the devices, that is,

$$\begin{cases} v(0,t) = -R_0 i(0,t), & t \geq 0\,, \qquad (3) \\ C_0\,\dfrac{\partial v}{\partial t}(\ell,t) = i(\ell,t), & t \geq 0\,. \qquad (4) \end{cases}$$

Here R_0 is a constant.

It is convenient to normalize time and distance, obtaining the normalized variables

$$\tau = \frac{t}{rc\ell^2}\,, \qquad \xi = \frac{x}{\ell}$$

After introducing the adimensional parameters $\rho = r\ell/R_0$ and $\gamma = c\ell/C_0$, the corresponding equations for voltage $V(\xi,\tau)$ and current $I(\xi,\tau)$ become

$$\frac{\partial^2 V}{\partial \xi^2} = \frac{\partial V}{\partial \tau} \qquad (1'a), \qquad \frac{\partial^2 I}{\partial \xi^2} = \frac{\partial I}{\partial \tau} \qquad (1'b)$$

$$V(\xi,0) = v_0(\ell\xi),\ \xi \in [0,1] \qquad (2'a), \qquad I(\xi,0) = i_0(\ell\xi),\ \xi \in [0,1] \qquad (2'b)$$

$$\frac{\partial V}{\partial \xi}(0,\tau) - \rho V(0,\tau) = 0 \qquad (3'a), \qquad \frac{\partial^2 I}{\partial \xi^2}(0,\tau) - \rho \frac{\partial I}{\partial \xi}(0,\tau) = 0 \qquad (\ )$$

$$\frac{\partial^2 V}{\partial \xi^2}(1,\tau) + \gamma \frac{\partial V}{\partial \xi}(1,\tau) = 0 \qquad (4'a), \qquad \frac{\partial I}{\partial \xi}(1,\tau) + \gamma I(1,\tau) = 0 \qquad (\ )$$

The diffusion equation is normally solved by separation of variables. Considering the current, we seek a general solution of the form $I(\xi,\tau) = g(\xi)h(\tau)$. Equation (1'b) is thus equivalent to the two equations

$$\frac{d^2 g}{d\xi^2} + \mu^2 g = 0, \qquad \frac{dh}{d\tau} + \mu^2 h = 0$$

for constant μ. Any function of the form $(A\cos\mu\xi + B\sin\mu\xi)e^{-\mu^2\tau}$ is a solution of (1'b); the constant μ is any of the eigenvalues of the problem, i.e., any choice which satisfies the boundary value problem (3'b),(4'b). Specifically, after some obvious algebra, from (3'b) and (4'b) we obtain the characteristic equation

$$\operatorname{tg}\mu = \frac{\gamma\rho}{\gamma+\rho} \cdot \frac{1}{\mu} - \frac{\mu}{\gamma+\rho} . \qquad (5)$$

The infinitely many solutions of (5) occur symmetrically with respect to 0. Therefore we restrict ourselves to $\mu > 0$. (A graphical display of the solution set is given in figure 3). The eigenvalues $\{\mu_i : i = 0,1,\ldots\}$ are indexed so that $\mu_0 < \mu_1 < \ldots$; note that $\mu_i > (2i-1)\pi/2$ for $i \geq 1$. As is well-known, to each μ_i there corresponds an eigenfunction $g_i(\xi)$ which

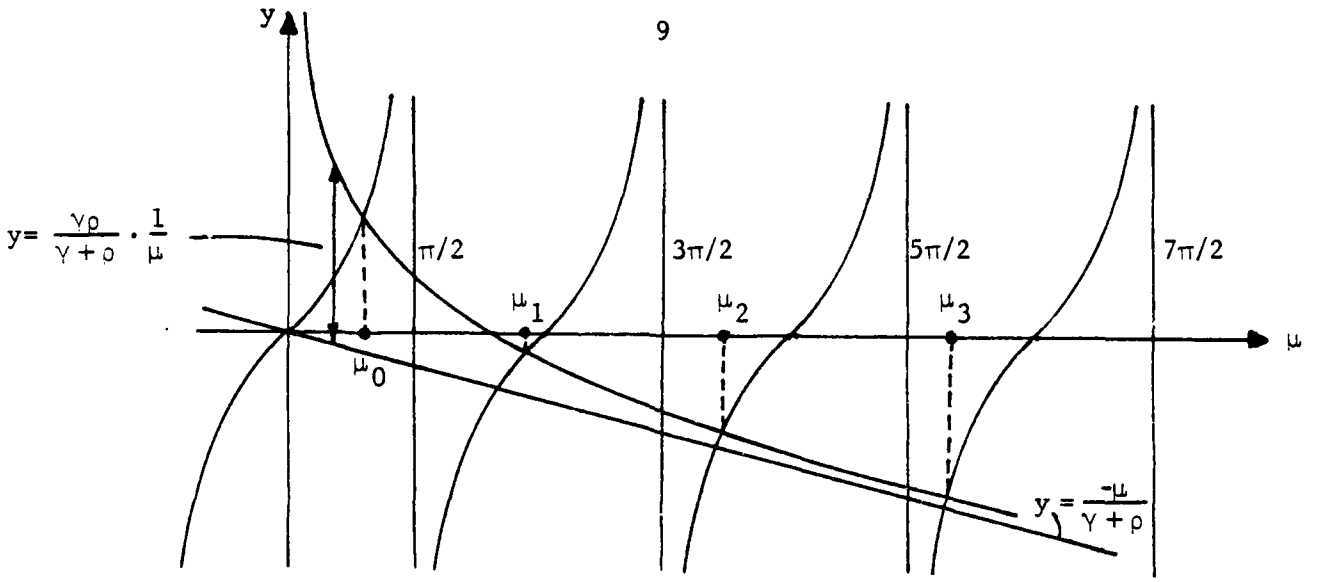


Figure 3. Illustration of the solutions of the characteristic equation.

simultaneously satisfies

$$g_i''(\xi) + \mu_i^2 g_i(\xi) = 0, \tag{1"}$$

$$g_i''(0) - \rho g_i'(0) = 0, \tag{3"}$$

$$g_i'(1) + \gamma g_i(1) = 0. \tag{4"}$$

Unfortunately, relation (3"), which is equivalent to $\rho g_i'(0) + \mu_i^2 g_i(0) = 0$, fails to realize the classical Sturm-Liouville condition [10], so that $\{g_i(\xi)\colon i = 1,2,\ldots\}$ is <u>not</u> a set of orthogonal functions. However by defining the "inner product" of functions on [0,1] in the following unconventional way

$$\langle\langle u,v\rangle\rangle \triangleq \int_0^1 u(\xi)v(\xi)d\xi + \frac{u(0)v(0)}{\rho} \tag{6}$$

It is easily shown that the eigenfunctions can be normalized so that

$$\langle\langle g_i, g_j\rangle\rangle = \delta_{ij}, \tag{7}$$

where δ_{ij} is the Kronecker symbol. Since (3") applied to the general expression $g(\xi) = A\cos\mu\xi + B\sin\mu\xi$ yields $A = -(\rho/\mu)B$, we have:

$$g_i(\xi) = G_i(\sin\mu_i\xi - \frac{\rho}{\mu_i}\cos\mu_i\xi) \tag{8}$$

where G_i is a constant. We can now project — in the sense of our inner product (6) — the initial condition $I(\xi,0)$ on the set $\{g_i(\xi)\}$, and obtain

$$I_i \triangleq \langle\langle I(\xi,0), g_i(\xi)\rangle\rangle ,$$

whence the general solution for the current is

$$I(\xi,\tau) = \sum_{i=0}^{\infty} I_i g_i(\xi) e^{-\mu_i^2\tau} . \tag{9}$$

2.2 Analysis of the saturated regime

As mentioned earlier, in the saturated regime starting at $t = 0$, capacitor C_0 and the line are at voltage V_0, and the transistor is modeled as a current generator with current value $-I_0$. The circuit is modeled as in figure 4. Therefore, boundary conditions (3) must be replaced by the

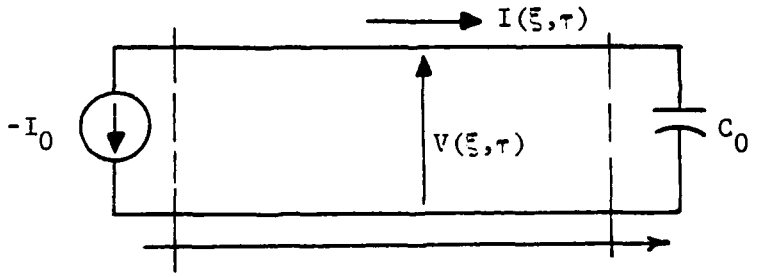


Figure 4. Model of the saturated regime.

new nonhomogeneous conditions

$$i(0,t) = -I_0 .$$

The current $I(\xi,\tau)$ can thus be expressed as

$$I(\xi,\tau) = I_0(\xi) + I_1(\xi,\tau)$$

where $I_1(\xi,\tau)$ satisfies homogeneous boundary conditions

$$\begin{cases} I_1(0,\tau) = 0 \\ \dfrac{\partial I_1}{\partial \xi}(1,\tau) + \gamma I_1(1,\tau) = 0 \end{cases} \tag{3'''}$$

with initial conditions $I_1(\xi,0) = -I_0(\xi)$, while the stationary term $I_0(\xi)$ satisfies the boundary conditions

$$\begin{cases} I_0(0) = -I_0 \\ \dfrac{\partial I_0}{\partial \xi}(1) + \gamma I_0(1) = 0. \end{cases}$$

The latter, and equation (1'b), immediately yield

$$I_0(\xi) = I_0\left(\frac{\gamma}{\gamma+1}\,\xi - 1\right).$$

Turning now to $I_1(\xi,\tau)$, note that condition (3''') implies $R_0 = \infty$, or equivalently, $\rho = 0$. As a consequence the boundary conditions for the eigenfunctions $\{g_i(\xi)\}$ become

$$\begin{cases} g_i(0) = 0 \\ g_i'(1) + \gamma g_i(1) = 0 \end{cases}$$

which are of the Sturm-Liouville type. Indeed, from (8) the eigenfunctions become

$$g_i(\xi) = \sqrt{\frac{2}{1 - \dfrac{\sin 2\mu_i}{2\mu_i}}} \cdot \sin \mu_i \xi, \quad (i = 1,2,\ldots)^{(1)}$$

---

(1) It should be noted that the eigenvalue $\mu_0 = 0$ does not yield a valid eigenfunction.

and they form an orthonormal set in the conventional sense. The coefficients I_i are therefore expressed as

$$I_i = -\langle I_0(\xi), g_i \rangle$$

(where $\langle\ \rangle$ denotes the conventional inner product) and

$$I_1(\xi,\tau) = \sum_{i=1}^{\infty} I_i g_i(\xi) e^{-\mu_i^2 \tau}$$

In conclusion

$$I(\xi,\tau) = I_0\left(\frac{\gamma}{\gamma+1}\xi - 1\right) + \sum_{i=1}^{\infty} I_i g_i(\xi) e^{-\mu_i^2 \tau}. \qquad (10)$$

We shall refer to the two terms in the right side of (10) as the stationary and transient terms, respectively.

The expression of $V(1,\tau)$, the voltage at the gate capacitor end of the line, is obtained from $I(1,\tau)$ and the capacitor equation, as

$$\boxed{V(1,\tau) = V_0 - \frac{rc\ell^2}{C_0}\int_0^{\tau} I(1,\vartheta)\,d\vartheta = V_0 - \frac{rc\ell^2 I_0}{C_0(1+\gamma)}\tau + \frac{rc\ell^2}{C_0}\sum_{i=1}^{\infty} I_i \frac{g_i(1)}{\mu_i^2}\left(1 - e^{-\mu_i^2\tau}\right).} \qquad (11)$$

(Note the corrective factor $rc\ell^2$ due to the normalization of time.) From this, by integrating $I(\xi,\tau)$ along the line, we obtain

$$V(0,\tau) = V(1,\tau) + r\ell \int_0^1 I(\eta,\tau)\,d\eta.$$

From this expression for $V(0,\tau)$ we can determine the time τ_{PO} at which $V(0,\tau) = V_{PO}$, i.e. the time at which the regime changes. Assuming that $V_{PO}/V_0 = 0.8$, by numerical evaluation we have ascertained that for $\gamma \leq 10^3$ at $\tau = \tau_{PO}$ the transient term of (10) is all but negligible ($< 10^{-8} \cdot I(\xi,\tau_{PO})$). Therefore in this range of γ, we may safely assume

$$I(\xi,\tau_{PO}) = I_0\left(\frac{\gamma}{\gamma+1}\xi - 1\right)$$

as the initial condition for the current in the ohmic regime.

2.3 Analysis of the ohmic regime

In this case, the phenomenon is governed by (1'b) (3'b) (4'b) with initial condition

$$I(\xi,0) = I_0\left(\frac{\gamma}{\gamma+1}\,\xi - 1\right). \qquad (2"b)$$

This expression is projected on the basis of the eigenfunctions (8) according to the unconventional rule (6), thereby obtaining

$$I_i = I_0\left(\frac{\gamma}{\gamma+1}\,H_i - K_i\right)$$

where H_i, and K_i $(i = 0,1,\ldots)$ are easily computable functions of the parameters μ, ρ, and γ. It follows that

$$I(\xi,\tau) = I_0 \sum_{i=0}^{\infty}\left(\frac{\gamma}{\gamma+1}\,H_i - K_i\right)g_i(\xi)e^{-\mu_i^2\tau} \qquad (12)$$

From $\left(\frac{\partial V}{\partial \xi} = -r\ell I\right)$ we readily obtain the expression of $V(\xi,\tau)$, as follows:

$$V(\xi,\tau) = V(0,\tau) - r\ell\int_0^{\xi} I(\eta,\tau)d\eta$$

$$= -R_0 I(0,\tau) - r\ell I_0 \sum_{i=0}^{\infty}\left(\frac{\gamma}{\gamma+1}\,H_i - K_i\right)e^{-\mu_i^2\tau}\int_0^{\xi} g_i(\eta)d\eta$$

that is,

$$\boxed{V(\xi,\tau) = R_0 I_0 \sum_{i=0}^{\infty}\left(\frac{\gamma}{\gamma+1}\,H_i - K_i\right)f_i(\xi)e^{-\mu_i^2\tau}} \qquad (13)$$

with

$$f_i(\xi) \triangleq g_i(0) + \rho\int_0^{\xi} g_i(\eta)d\eta = -\frac{\rho G_i}{\mu_i}\left(\cos\mu_i\xi + \frac{\rho}{\mu_i}\sin\mu_i\xi\right)^{(2)}$$

---

(2) It can be shown that the $\{f_i(\xi)\}$ are a set of eigenfunctions of the general solution of (1'a), (3'a), (4'a).

3. Discussion and conclusions.

Expressions (11) and (13), which respectively give the voltage $V(\xi,\tau)$ in the saturated and ohmic regimes, are the objective of our analysis. In any given technology the ratio $\gamma/\rho = cR_0/rC_0$ is a constant; therefore only one parameter describes the behavior. Several discharge curves have been plotted in figure 5, for the values of $\gamma = 10^{-2}$, 10^{-1}, 1, 10, 10^2, 10^3. Taking as propagation delay the instant t_{PR} for which $V(1,t_{PR}) = V_{TH} \simeq 0.2\ V_0$, we have plotted in figure 6 the relation between t_{PR} and γ. On a purely

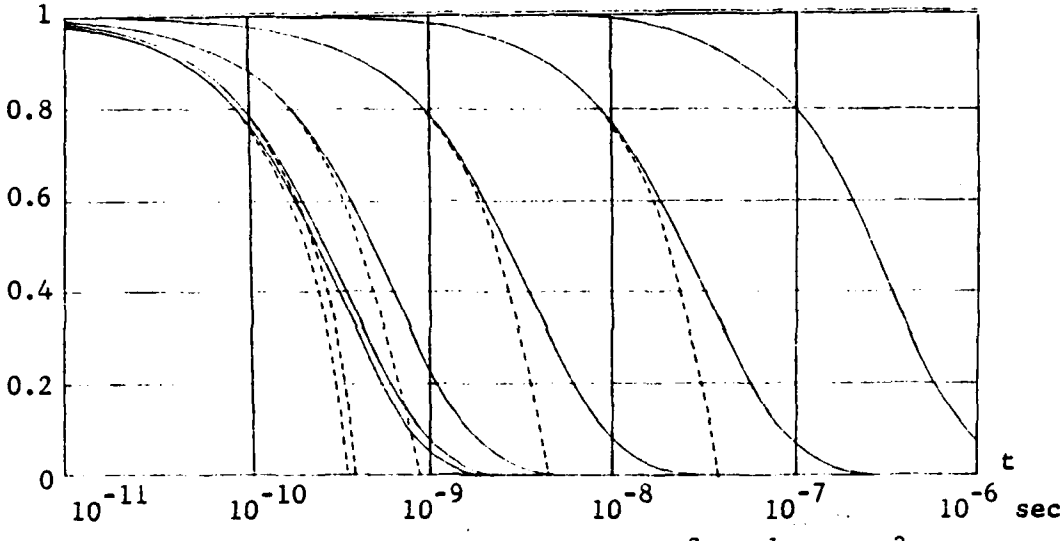


Figure 5. Discharge curves for $\gamma = 10^{-2}, 10^{-1}, \ldots, 10^3$. The broken lines describe the discharge at constant current.

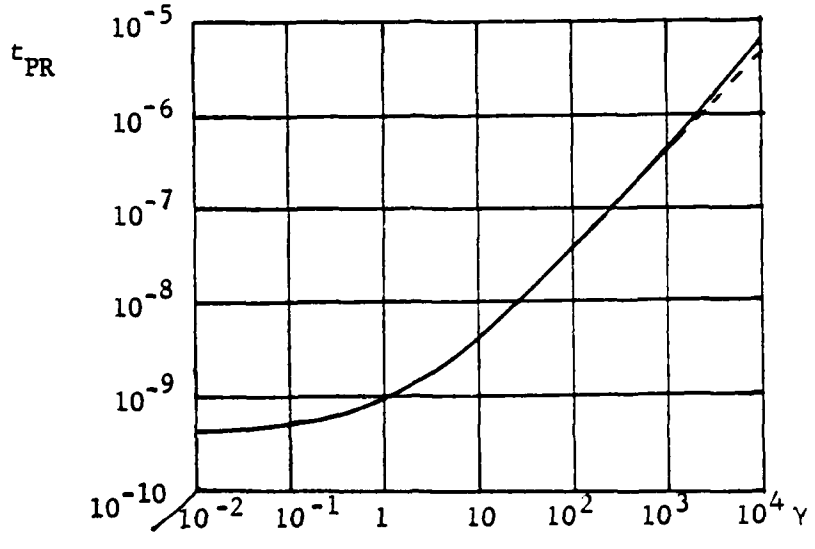


Figure 6. The relation between t_{PR} and γ.

qualitative basis, at this point it is interesting to observe the following facts (here we assume that $\rho \leq 10^{-4}\gamma$, as in current technology):

(i) For small values of γ (roughly, $< 10^{0}$), the propagation delay is practically constant and is determined by the characteristics of the devices. This is the unchallenged domain of the constant delay (or synchronous model).

(ii) For larger values of γ (roughly, $10^{0} < \gamma < 10^{3}$), the propagation delay is basically proportional to γ (i.e. to the wire length ℓ). This is the domain of the capacitive model.

(iii) For very large values of γ (roughly, $\gamma > 10^{3}$), the dependency of the propagation delay upon γ begins to deviate from linearity, i.e. the effects of the dispersive transmission line begin to be felt. This is the domain of the diffusion model.

On a less qualitative basis, we have examined expression (13) and evaluated $V(\xi,\tau)$ by summing at first a very large number of terms of the series in the right side, and next restricting the calculation to the first term (corresponding to $i = 0$). Since μ_0 is very close to 0 and $\mu_i > (2i-1)\pi/2$, as was to be expected the sum of all other terms is negligible with respect to the first term. Therefore we shall now consider the approximate – but basically valid – expression

$$V(\xi,\tau) = R_0 I_0\left(\frac{\gamma}{\gamma+1} H_0 - K_0\right) f_0(\xi) e^{-\mu_0^2 \tau}. \qquad (14)$$

The time constant of the discharge, in unnormalized time t, has the expression

$$t_D = \frac{rc\ell^2}{\mu_0^2},$$

where – we recall – μ_0 is the smallest positive solution of equation (5). If in (5) we expand $tg\mu_0$ in Taylor series we obtain

$$\mu_0 + \sum_{i=1}^{\infty} t_i \mu_0^{2i+1} = \frac{\gamma\rho}{\gamma+\rho} \frac{1}{\mu_0} - \frac{\mu_0}{\gamma+\rho}$$

whence

$$t_D = R_0 C_0 (1 + \gamma + \rho)(1 + \frac{\gamma+\rho}{1+\gamma+\rho} \sum_{i=1}^{\infty} t_i \mu_0^{2i}) \qquad (15)$$

Letting $\epsilon(\gamma,\rho) \triangleq (\gamma+\rho) \sum_{i=1}^{\infty} t_i \mu_0^{2i}/(1+\gamma+\rho)$, ϵ gives the relative deviation of t_D from $R_0 C_0(1 + \gamma + \rho)$, which is linear in ρ and γ and gives the delay in an idealized capacitive model. In this model the dispersive line is replaced by a single equivalent capacitance of value $c\ell(1 + \rho/\gamma)$, where ρ/γ is a constant in any given fabrication technology; indeed $C_0(1+\gamma+\rho) = C_0 + c\ell(1+\rho/\gamma)$. It is therefore of interest to obtain the behavior of ϵ as a function of ρ and γ. A set of contour lines of ϵ is plotted in a logarithmic diagram in figure 7.

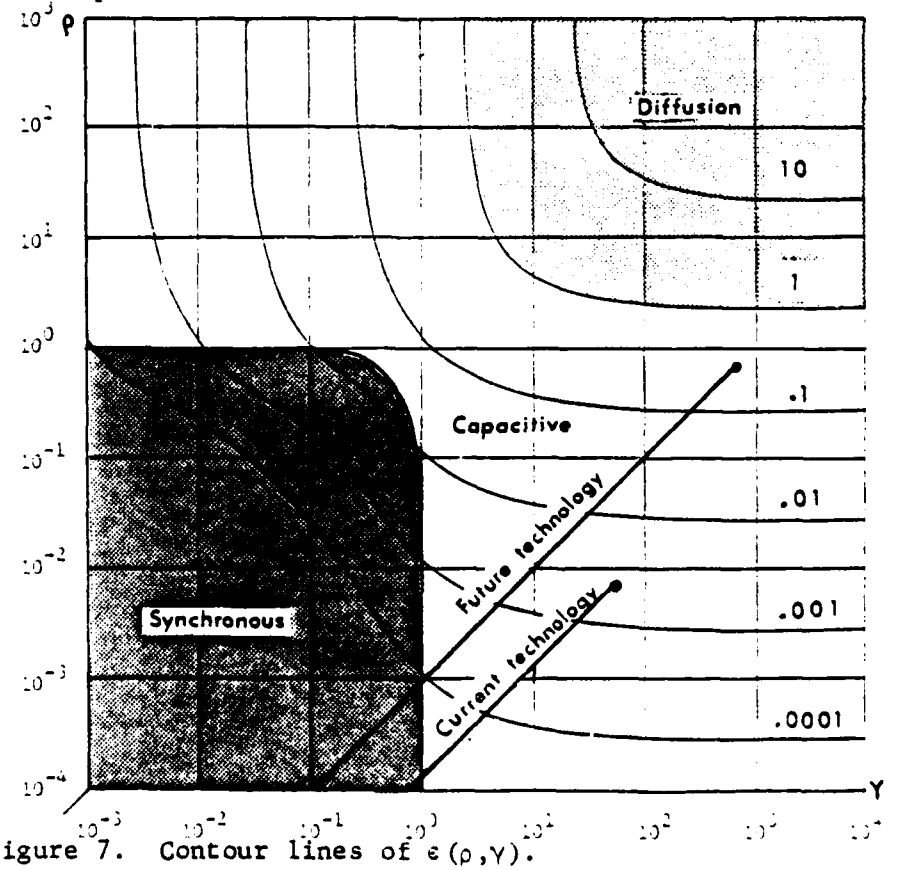


Figure 7. Contour lines of $\epsilon(\rho,\gamma)$.

It seems reasonable to try to define in this diagram the regions of validity of each of the three models: synchronous, capacitive, diffusion. Specifically, referring to equation (15), we may (somewhat arbitrarily) define the region of the synchronous model as the one where $t_D \leq 2R_0C_0$ (that is, the time constant is at most twice that due to the devices alone); by an equally arbitrary criterion, we may define the region of the capacitive model as the one where $\epsilon \leq 1$ (a deviation which at most corresponds to doubling the propagation delay). This region is shown unshaded in figure 7. In the same diagram each technology is represented by a straight line of slope +1, since — as we noted — in any given technology $\rho = K_1\gamma$ (K_1, a constant). Current MOS technology is characterized by the following parameter values:

Feature width (λ) = 2.5 μm
Field oxide thickness = 1 μm
Gate oxide thickness = 600 Å
Aluminum thickness = 1 μm
Power supply voltage = 5 V

We assume that 2λ and 3λ be, respectively, the channel length of transistors and the width of aluminum wires; in addition the minimum channel width is chosen 4λ [1]. Recalling that the resistivity of aluminum is 0.28×10^{-7} Ωm that the dielectric constant of SiO_2 is 0.46×10^{-10} F/m, and that the electron mobility in Si is about 0.8×10^{-1} m^2/Vsec (we refer here to the n-channel portion of CMOS), we obtain the following values (see [1] [11]):

$$I_0 = 0.98 \text{ mA}, \quad R_0 = 4.05\times10^{3}\ \Omega, \quad C_0 = 4.12\times10^{-2} \text{ pF}$$
$$r = 3.78\times10^{3}\ \Omega/\text{m}, \quad c = 3.46\times10^{-10} \text{ F/m}$$

whence $\rho/\gamma = rC_0/cR_0 = 1.10\times10^{-4}$. The corresponding straight-line is shown in figure 7. In addition, assuming a maximum chip width of 10 mm, we have $\gamma \leq 84$. The corresponding point is also shown in figure 7.

In a scaled-down technology of the foreseeable future, not all parameters are likely to be changed according to a fixed ratio. Indeed it appears that "feature size", gate oxide thickness, and power supply will be scaled down, while there is a strong interest in maintaining the thicknesses of both aluminum and field oxide. Therefore a reasonable set of parameters of a future scaled-down technology will be

Feature width	= 0.5 μm
Field oxide thickness	= 1 μm
Gate oxide thickness	= 150 Å
Aluminum thickness	= 1 μm
Power supply voltage	= 3 V

Correspondingly we obtain: $I_0 = 1.4$ mA, $R_0 = 1.69\times10^{3}\ \Omega$, $C_0 = 6.5\times10^{-3}$ pF, $r = 1.89\times10^{4}\ \Omega/m$, $c = 6.93\times10^{-11}$ F/m, whence $\rho/\gamma = 0.992\times10^{-3}$. Moreover, assuming a maximum chip width of 50 mm, we obtain $\gamma_{max} \leq 5.65\times10^{2}$. The corresponding curve and point are also plotted in figure 7.

The conclusion we extract from the preceding analysis is that not only the current but also the projected MOS-FET VLSI technologies fall in the domains of either the synchronous or the capacitive models. In the latter propagation delay is proportional to the length of the wires. Note, however, that this propagation delay is computed in the hypothesis that both the driving and the driven transistors be standard (i.e., of minimum size). However, by raising the channel width of the driving transistor, the current I_0 increases and t_{PR} decreases. Indeed — as suggested by Carver-Mead [1] and Thompson [3] — if the channel width is proportional to the capacitive load for all transistors, one approaches constant propagation time and, presumably, current density becomes the limiting factor.

It must be noted, however, that projected future technology may reach the (conventional) boundary of the capacitive model region. Beyond this boundary, a possible design philosophy — as suggested by Chazelle-Monier [4] — is to introduce repeaters on long wires in order to achieve delay proportional to the wire length. Note, however, that in this case we can no longer avail ourselves of channel width control. Another alternative, entirely in the realm of speculation, could be the development of integrated nondispersive transmission lines, where speed of light considerations are the controlling phenomena.

Acknowledgement

The authors would like to thank K. Hess, M. Lightner, B. G. Streetman, G. W. Swenson and T. N. Trick for many invaluable discussions and for crucial comments on this manuscript.

REFERENCES

1. C. A. Mead and L. Conway, Introduction to VLSI Systems, Addison-Wesley, Reading, Mass. 1979.

2. R. P. Brent and H. T. Kung, "The chip complexity of binary arithmetic," Proceedings of the 12th Symposium on Theory of Computing, Los Angeles, pp. 190-200; April 1980.

3. C. D. Thompson, "A complexity theory for VLSI," Ph.D. Thesis, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Penn., August 1980.

4. B. Chazelle and L. Monier, "A model of computation for VLSI with related complexity results," Tech. Rep., Dept. of Comp. Sci., Carnegie Mellon University, February 1981.

5. C. L. Seitz, System Timing, Chap. 7, in C. Mead and L. Conway, Introduction to VLSI Systems, Addison-Wesley, Reading, Mass. 1979.

6. H. S. Stone, "Parallel processing with the perfect shuffle," IEEE Transactions on Computers, vol. C-20, pp. 153-161; 1971.

7. F. P. Preparata and J. Vuillemin, "The cube-connected-cycles: a versatile network for parallel computation," Comm. of the ACM, vol. 24, no. 5, pp. 300-309, May 1981.

8. J. L. Bentley and H. T. Kung, "A tree machine for searching problems," Tech. Rep., Dept. of Comp. Sci., Carnegie-Mellon University; August 1979.

9. Engineering Staff of AMI, MOS Integrated Circuits, Van Nostrand, New York, 1972.

10. R. L. Street, The analysis and solution of Partial Differential Equations, Brooks/Cole Publishing Company; Monterey, CA 1973.

11. B. G. Streetman, Solid State Electronic Devices, Second Edition, Prentice-Hall, Inc., Englewood Cliffs, NJ 1980.


4-8
DTI