A Stochastic Wire-Length Distribution for Gigascale Integration (GSI)—Part II: Applications to Clock Frequency, Power Dissipation, and Chip Size Estimation

Jeffrey A. Davis, Vivek K. De, and James D. Meindl, *Life Fellow, IEEE*

(Invited Paper)

***Abstract*— Based on Rent's Rule, a well-established empirical relationship, a complete wire-length distribution for on-chip random logic networks is used to enhance a critical path model; to derive a preliminary dynamic power dissipation model; and to describe optimal architectures for multilevel wiring networks that provide maximum interconnect density and minimum chip size.**

***Index Terms*— Average wire length, critical path, die area estimation, power dissipation model, Rent's Rule, wire-length distribution.**

I. Introduction

ONE of the main barriers to achieving gigascale integration is the limitation imposed by the wiring requirements of gigascale integration (GSI) systems [1], [2]. Multilevel wiring networks are extensively used in current VLSI systems to mitigate the impact of wiring on clock frequency, power consumption, and chip size [1]–[3]. Adequate modeling of a multilevel wiring networks is essential to elucidate the limitations and the opportunities for future GSI products.

The primary information used to model multilevel wiring networks is a complete wire-length distribution that is derived in a companion to this paper [4]. This new distribution is used to enhance a critical path model for clock frequency estimation; to estimate dynamic power dissipation from signal wires; and to determine the optimal scaling of the cross-sectional dimensions of interconnects in a multilevel wiring network and hence chip size.

The various applications of the wire-length distribution include the derivation of a critical path model is in Section II, a power dissipation model in Section III, and an optimal multilevel wiring architecture in Section IV.

II. Critical Path Model

Various critical path models have been proposed to estimate cycle time [2], [3], [5]. One essential component of a critical path model for CMOS circuits is the size of the wiring networks loading the critical path gates. The complete distribution enables a more accurate representation of the average wiring length and estimation of the longest interconnect in the system.

An established critical path model from [3] has all but one of its gates loaded by an average interconnect net. The remaining gate is loaded by a single global interconnect [3]. From [5], the limit on the clock period is given by

$$T_c \geq T_{\rm cs} + n_{\rm cp} t_d + T_{\rm LD} \tag{1}$$

where T_c is the clock period, $T_{\rm cs}$ is the clock skew, $n_{\rm cp}$ is the number of gates in the critical path, t_d is the average time delay of each gate loaded by an average wiring net, and $T_{\rm LD}$ is the time delay of the longest global interconnect in the system.

The average point-to-point interconnect length is determined directly for the interconnect density function from [4]

$$L_{\rm avg} = \frac{L_{\rm total}}{I_{\rm total}} = \frac{\displaystyle\int_{\ell=1}^{\ell=2\sqrt{N}} \ell i(\ell)\, d\ell}{\displaystyle\int_{\ell=1}^{\ell=2\sqrt{N}} i(\ell)\, d\ell} \tag{2}$$

where $L_{\rm total}$ is the total point-to-point interconnect length and $I_{\rm total}$ is the total number of interconnects. Evaluating (2) gives the final form of the average wire-length $L_{\rm avg}$ (see (3), shown at the bottom of the next page). This average interconnect length is dependent only on the number of gates in the system and Rent's exponent p. A graph of the average wire-length versus number of gates for various p values is seen in Fig. 1. A previous distribution gives higher average interconnect lengths due to overestimation of the number of longer interconnects as seen in Fig. 1 [6]–[8].

Each critical path gate is loaded by an average wiring net, and the length of an average wiring net is

$$L_{\rm avg\_net} = \frac{L_{\rm total\_net}}{I_{\rm nets}} = \frac{\chi L_{\rm total\_}p{\rm -to-}p}{\dfrac{I_{\rm total}}{\rm f.o.}} = \chi\ {\rm f.o.}\ L_{\rm avg} \tag{4}$$

where χ is a correction factor that converts the point-to-point interconnect length to wiring net length (using a linear net model $\chi = 4/({\rm f.o.} + 3)$ [4], $L_{\rm total\_net}$ is the total net length, f.o. is the average fanout, and $I_{\rm nets}$ is the total number of nets.

Manuscript received July 1, 1997; revised October 9, 1997. The review of this paper was arranged by Editor Y. Nishi. This work was supported by the Semiconductor Research Corporation (SJ-374.002) and the Advanced Research and Projects Agency (BAA9415-A-009).

The authors are with the Microelectronics Research Center, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0269 USA.

Publisher Item Identifier S 0018-9383(98)01663-3.

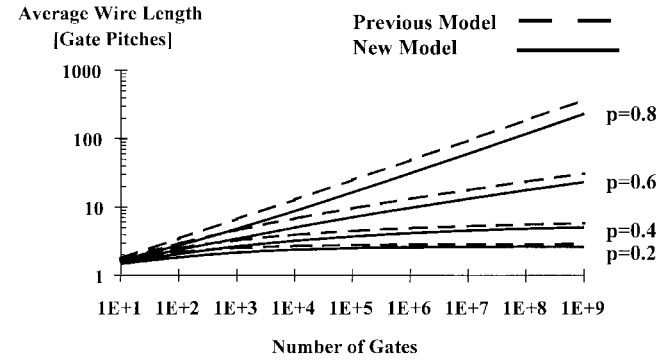


Fig. 1. Average interconnect length versus number of gates for $p = 0.2$, 0.4, 0.6, and 0.8.

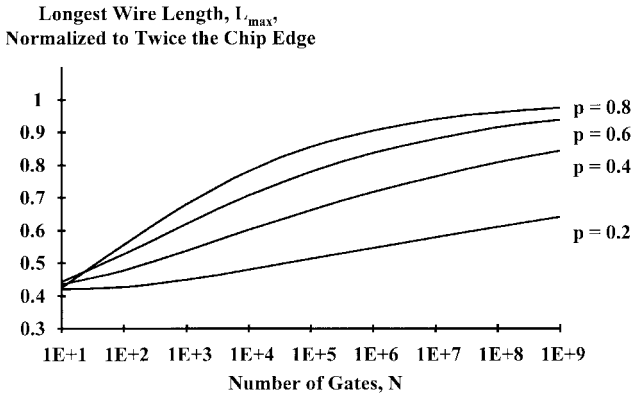


Fig. 2. Longest interconnect length normalized to twice the chip edge versus number of gates for $p = 0.2$, 0.4, 0.6, and 0.8.

The longest interconnect also is estimated using the interconnect density function, i.d.f., which has a range from 1 gate pitch to $2\sqrt{N}$ gate pitches [4]. The i.d.f. predicts that one interconnect exists in the interval between a given length $\ell_{\max}$ and $2\sqrt{N}$ from the following expression:

$$1 = \int_{\ell=\ell_{\max}}^{\ell=2\sqrt{N}} i(\ell)\, d\ell. \tag{5}$$

Because the interconnect density function is monotonically decreasing, then $\ell_{\max}$ is defined as the most probable value for the longest interconnect length. The expression for this length is also written in terms of the c.i.d.f., $I(\ell)$

$$I(2\sqrt{N}) - I(\ell_{\max}) = 1. \tag{6}$$

A graph of the longest interconnect, $\ell_{\max}$, normalized to twice the chip edge, $2\sqrt{N}$, for $p = 0.2,\ p = 0.4,\ p = 0.8$ values appears in Fig. 2 with $\alpha k = 3.0$.

III. Dynamic Power Dissipation Model

The dominate source of load capacitance for many VLSI CMOS circuits is the wiring capacitance [2], [9]. The interconnect density function provides a priori information concerning the distribution of capacitive loads present in a GSI system. Assuming a constant activity factor for each capacitive node, the average dynamic power dissipation of the signal interconnects P_{avg} is given by [2]

$$P_{\text{avg}} = a\tfrac{1}{2} C_{\text{total}} V_{\text{dd}}^2 f_c \tag{7}$$

where V_{dd} is the supply voltage, a is the average activity factor for each gate, C_{total} is the total capacitive load of the wiring network, and f_c is the clock frequency. The C_{total} term is estimated from the geometrical configurations of the interconnects in a multilevel wiring network and from the interconnect density function. A conventional multilevel network, for example, has two main interconnect types: 1) local interconnects and 2) global interconnects. The cross-sectional dimensions and the capacitive contributions for an arbitrary interconnect are illustrated in Fig. 3(a) and (b).

Assuming that the neighboring wiring planes in the multilevel network provide an effective ground plane, then total capacitance per unit length is given from Fig. 3(b)

$$c_{\text{total}} = 2c_{\text{ground}} + 2c_{\text{line\_to\_line}} \tag{8}$$

where c_{ground} is the line-to-ground capacitance, and $c_{\text{line-to-line}}$ is the line-to-line capacitance. The values of c_{ground} and $c_{\text{line-to-line}}$ that include fringing effects are given by [10] as

$$\begin{aligned}\frac{c_{\text{ground}}}{\varepsilon} &= \frac{W}{H_\varepsilon} + 1.086(1 + 0.685e^{-(H_\rho/1.343S)} \\ &\quad - 0.9964e^{-(S/1.421H_\varepsilon)}) \\ &\quad \cdot \left(\frac{S}{S+2H_\varepsilon}\right)^{0.0476} \left(\frac{H_\rho}{H_\varepsilon}\right)^{0.337} \\ \frac{c_{\text{line-to-line}}}{\varepsilon} &= \left(\frac{H_\rho}{S}\right)(1 - 1.897e^{(-H_\varepsilon/0.31S) - (-H_\rho/2.474S)} \\ &\quad + 1.302e^{-H_\varepsilon/0.082S} - 0.1292e^{-H_\rho/1.326S}) \\ &\quad + 1.722(1 - 0.6548e^{-W/0.3477H_\varepsilon})e^{-S/0.651H_\varepsilon}\end{aligned} \tag{9}$$

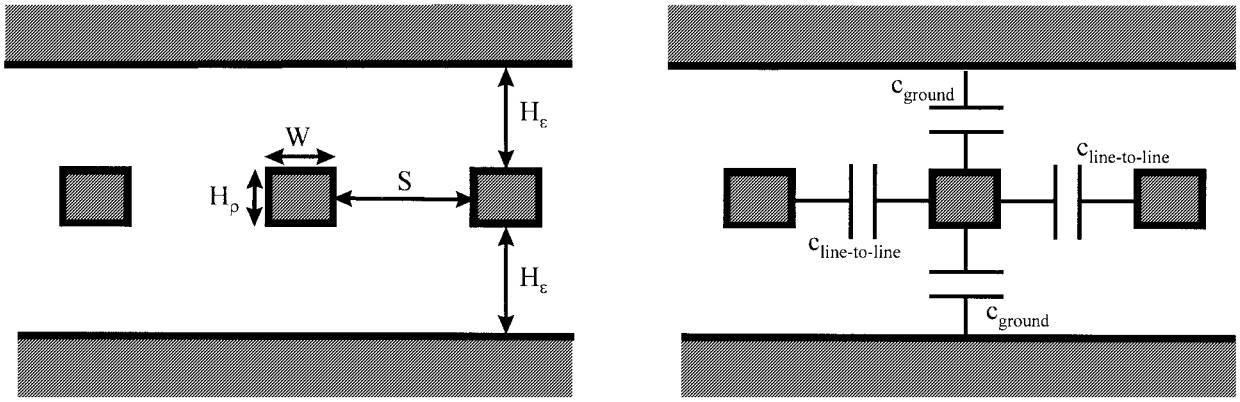
where the geometrical variables are defined in Fig. 3(a).

For the general case, the capacitances per unit length for the local and global interconnects are dissimilar, and the expressions for the total capacitance of the local levels $C_{\text{total,loc}}$ and the total capacitance of the global levels $C_{\text{total,glob}}$ are given by

$$C_{\text{total,loc}} = c_{\text{loc}} \sqrt{\frac{A_c}{N}} \chi \int_1^{L_{\text{loc}}} \ell i(\ell)\, d\ell \tag{10}$$

$$C_{\text{total,glob}} = c_{\text{glob}} \sqrt{\frac{A_c}{N}} \chi \int_{L_{\text{loc}}}^{2\sqrt{N}} \ell i(\ell)\, d\ell \tag{11}$$

$$L_{\text{avg}} = \frac{\left(\dfrac{p-0.5}{p} - \sqrt{N} - \dfrac{p-0.5}{6\sqrt{N}(p+0.5)} + N^P\left(\dfrac{-p-1+4^{p-0.5}}{2(p+0.5)p(p-1)}\right)\right)}{\left(N^{p-0.5}\dfrac{-2p-1+2^{2p-1}}{2p(p-1)(2p-3)} - \dfrac{(p-0.5)}{6p\sqrt{N}} + 1 - \dfrac{(p-0.5)\sqrt{N}}{(p-1)}\right)} \tag{3}$$



(a) (b)

Fig. 3. The definition of (a) the interconnect dimensions and (b) the capacitive components.

where symbol χ is the correction factor that converts the total length of point-to-point interconnects to total net length as derived in [4], A_c is the chip area, N is the number of gates, $c_{\rm loc}$ is the distributed capacitance for the local interconnects, $c_{\rm glob}$ is the distributed capacitance for the global interconnects, and $L_{\rm loc}$ is the length in gate pitches of the longest local interconnect. Assuming that the longest local interconnect $L_{\rm loc}$ is less than the chip edge, the above expressions simplify to

$$C_{\rm total,loc} = c_{\rm loc}\chi\sqrt{\frac{A_c}{N}}\int_1^{L_{\rm loc}} \Gamma\frac{\alpha k}{2} \cdot\left(\frac{\ell^3}{3} - 2\sqrt{N}\ell^2 + 2N\ell\right)\ell^{2p-3}\,d\ell \tag{12}$$

$$C_{\rm total,glob} = c_{\rm glob}\chi\sqrt{\frac{A_c}{N}} \cdot\left[\int_{L_{\rm loc}}^{\sqrt{N}} \Gamma\frac{\alpha k}{2}\left(\frac{\ell^3}{3} - 2\sqrt{N}\ell^2 + 2N\ell\right)\ell^{2p-3}\,d\ell + \int_{\sqrt{N}}^{2\sqrt{N}} \Gamma\frac{\alpha k}{6}(2\sqrt{N}-\ell)^3\ell^{2p-3}\,d\ell\right]. \tag{13}$$

Evaluating these two expressions gives

$$C_{\rm total,loc} = c_{\rm loc}\sqrt{\frac{A_c}{N}}\,\frac{\Gamma\alpha k\chi}{2}\left(\frac{L_{\rm loc}^{2p+1}-1}{6p+3} - \frac{\sqrt{N}(L_{\rm loc}^{2p}-1)}{p} - \frac{2N(1-L_{\rm loc}^{2p-1})}{2p-1}\right) \tag{14}$$

$$C_{\rm total,glob} = c_{\rm glob}\sqrt{\frac{A_c}{N}}\,\frac{\Gamma\alpha k\chi}{2} \cdot\left(-\frac{L_{\rm loc}^{2p+1}}{6p+3} + \frac{\sqrt{N}}{p}L_{\rm loc}^{2p} - \frac{2N}{2p-1}L_{\rm loc}^{2p-1} - N^{p+(1/2)}\frac{2+2p-4^p}{(2p+1)p(2p-1)(p-1)}\right). \tag{15}$$

Adding (14) and (15) gives the total capacitive load posed by the wiring network

$$C_{\rm total} = C_{\rm total,loc} + C_{\rm total,glob}. \tag{16}$$

TABLE I
MICROPROCESSOR APPLICATION

PHYSICAL PARAMETER	VALUE
Number of Gates, N	8.0 million
Rent's Exponent, p	0.6
Rent's Coefficient, k	4.0
Minimum Feature Size, F	0.18 μm [11]
Supply Voltage, Desktop Vdd	1.5V [11]
Max number levels, n_{max}	6 [11]
Metal Resistivity, Copper	1.673e-6 Ω-cm
Dielectric Constant, Polymer	ε_r = 2.5
Wiring Efficiency Factor, e_w	0.4 [2]

The final expression for the average dynamic power dissipation for a multilevel wiring network that has local and global wiring levels is given by

$$P_{\rm avg} = a\frac{1}{2}V_{\rm dd}^2 f_c\sqrt{\frac{A_c}{N}}\,\frac{\Gamma\alpha k\chi}{2} \cdot\left[c_{\rm loc}\left(\frac{L_{\rm loc}^{2p+1}-1}{6p+3} - \frac{\sqrt{N}(L_{\rm loc}^{2p}-1)}{p} - \frac{2N(1-L_{\rm loc}^{2p-1})}{2p-1}\right) + c_{\rm glob}\left(-\frac{L_{\rm loc}^{2p+1}}{6p+3} + \frac{\sqrt{N}}{p}L_{\rm loc}^{2p} - \frac{2N}{2p-1}L_{\rm loc}^{2p-1} - N^{p+(1/2)}\frac{2+2p-4^p}{(2p+1)p(2p-1)(p-1)}\right)\right]. \tag{17}$$

Given the two-tier multilevel network with the physical characteristics as outlined in Tables I and II, assuming that $W = H_\rho = H_\varepsilon = S$ for the local and global tiers such that $c_{\rm loc} = c_{\rm glob} = 6.08\varepsilon$, and given the activity factor a is approximately 0.10, then the total power dissipation in the signal interconnects for this system is 10 W and the power density is 2.03 W/cm^2. The local interconnects dissipate 68% of the total power, and the global interconnects dissipate 32% of the total power. Even though the amount of area dedicated to the global tier and local tier is approximately the same, the global wires dissipate less power than the local wires because 1) the global and local distributed capacitance are identical

TABLE II
TWO-TIER CONVENTIONAL DESIGN AND CRITICAL PATH RESULTS

PARAMETERS	PITCH	# OF LEVELS
Local	0.36μm	2
Global	1.53μm	4
**********	********	********
Clock Frequency	410Mhz	
Wire-Limited Chip Area	4.92 cm^2	
Transistor Limited Chip Area	N *x* 200F^2=0.52cm^2	

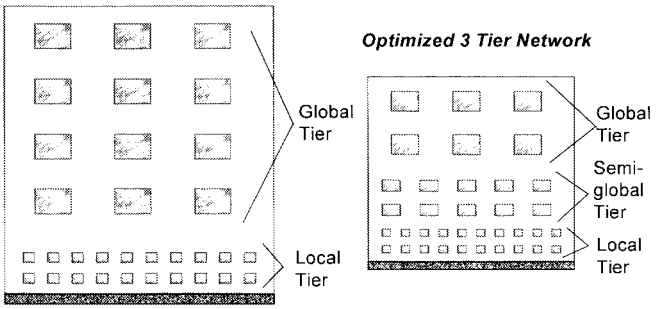


Fig. 4. Two-tier and three-tier multilevel interconnect architectures.

and 2) the total length of the global interconnects is less than the total length of the local interconnects.

IV. MULTILEVEL INTERCONNECT NETWORK ARCHITECTURE

A. Introduction

Through wiring information provided by the new interconnect distribution [4], a unique opportunity arises for optimizing multilevel interconnect network architectures. To illustrate this point, two types of multilevel wiring networks are examined: 1) a conventional two-tier multilevel network and 2) a three-tier multilevel network as illustrated in Fig. 4. A tier is defined as a collection of levels that have the same cross-sectional dimensions.

The local interconnect tier for both the two- and three-tier networks contains the shorter wire lengths in the multilevel network. Thus, the cross-sectional dimensions of the local levels are set to the minimum feature size to maximize the wire packing density. The longest local interconnect length $L_{\rm loc}$ is a critical length that is used with the interconnect distribution to determine the number of local interconnects as seen in Fig. 5.

The global interconnect tier for the two- and three-tier network contains the longer wires in the multilevel network. The cross-sectional dimensions for these global wires are determined by the clock frequency constraints on a global die-edge-length interconnect.

The semiglobal tier, which is only present in the three-tier network, has a pitch between those of the global tier and the local tier. The introduction of the semiglobal tier reduces the wiring area demand over the two-tier network because the shorter two-tier global interconnects have a smaller semiglobal pitch in the three-tier network as seen in Fig. 6. This semiglobal tier, therefore, reduces the overall wiring area demand for the multilevel network and thereby the wire-limited chip size.

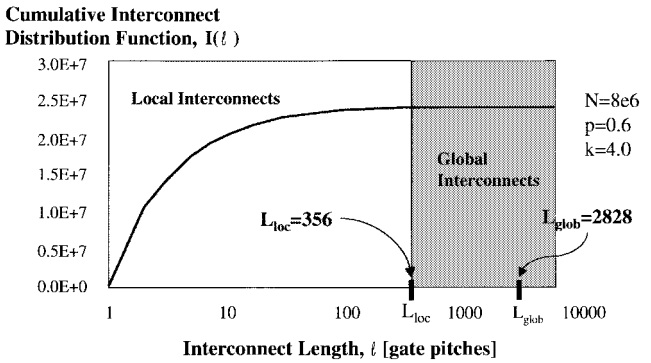


Fig. 5. The c.i.d.f. for a system with a two-tier multilevel architecture.

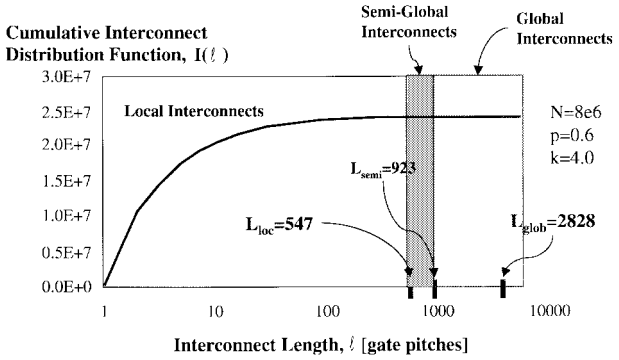


Fig. 6. The c.i.d.f. for a system with a three-tier multilevel architecture.

B. Two-Tier Wire-Limited Chip Size

The total required area for the two-tier wiring network is given by

$$A_{\rm required} = \sqrt{\frac{A}{N}}(p_{\rm loc}L_{\rm total,loc} + p_{\rm global}L_{\rm total,global}) \quad (18)$$

where A_c is the chip area, N is the number of gates, $p_{\rm loc}$ is the local pitch, $p_{\rm glob}$ is the global pitch, $L_{\rm total,loc}$ is the total length of the local interconnects (in gate pitches), and $L_{\rm total,global}$ is the total length of the global interconnects (in gate pitches). Using the interconnect density function [4], the total local length and the total global length are

$$L_{\rm total,loc} = \chi \int_1^{L_{\rm loc}} \ell i(\ell)\, d\ell \quad (19)$$

$$L_{\rm total,glob} = \chi \int_{L_{\rm loc}}^{2\sqrt{N}} \ell i(\ell)\, d\ell \quad (20)$$

where χ is the correction factor to model real wiring nets, where $\chi = 4/{\rm f.o.} + 3$ for a linear net model.

Assuming that the longest local length is less than a chip edge, then (19) and (20) become

$$L_{\rm total,loc} = \int_1^{L_{\rm loc}} \Gamma\frac{\alpha k}{2}\left(\frac{\ell^3}{3} - 2\sqrt{N}\ell^2 + 2N\ell\right)\ell^{2p-3}\, d\ell \quad (21)$$

$$L_{\rm total,glob} = \int_{L_{\rm loc}}^{\sqrt{N}} \Gamma\frac{\alpha k}{2}\left(\frac{\ell^3}{3} - 2\sqrt{N}\ell^2 + 2N\ell\right)\ell^{2p-3}\, d\ell + \int_{\sqrt{N}}^{2\sqrt{N}} \Gamma\frac{\alpha k}{6}(2\sqrt{N} - \ell)^3\ell^{2p-3}\, d\ell. \quad (22)$$

Evaluating (21) and (22) with substitution into (18) gives

$$A_{\text{required}} = \sqrt{\frac{A_c}{N}}\,\frac{\Gamma\alpha k\chi}{2}\left[p_{\text{loc}}\left(\frac{L_{\text{loc}}^{2p+1}-1}{6p+3} - \frac{\sqrt{N}(L_{\text{loc}}^{2p}-1)}{p} - \frac{2N(1-L_{\text{loc}}^{2p-1})}{2p-1}\right) + p_{\text{glob}}\left(-\frac{L_{\text{loc}}^{2p+1}}{6p+3} + \frac{\sqrt{N}}{p}L_{\text{loc}}^{2p} - \frac{2N}{2p-1}L_{\text{loc}}^{2p-1} - N^{p+(1/2)}\frac{2+2p-4^p}{(2p+1)p(2p-1)(p-1)}\right)\right]. \quad (23)$$

The wire-limited chip area is calculated from the condition that the total required wiring area is equal to the total available area in a multilevel network [2]

$$A_{\text{available}} = A_c e_w n_{\text{levels}} = A_{\text{required}} \quad (24)$$

where A_c is the chip area, e_w is wiring efficiency factor, and n_{levels} is the number of levels available for the multilevel network. The wiring efficiency factor e_w accounts for router efficiency and additional space needed for power and clock lines [2]. Combining (23) and (24) gives the final expression for wire-limited chip area for a two-tier multilevel network

$$A_c = \sqrt{\frac{A_c}{N}}\,\frac{\Gamma\alpha k\chi}{2e_w n_{\text{levels}}}\left[p_{\text{loc}}\left(\frac{L_{\text{loc}}^{2p+1}-1}{6p+3} - \frac{\sqrt{N}(L_{\text{loc}}^{2p}-1)}{p} - \frac{2N(1-L_{\text{loc}}^{2p-1})}{2p-1}\right) + p_{\text{glob}}\left(-\frac{L_{\text{loc}}^{2p+1}}{6p+3} + \frac{\sqrt{N}}{p}L_{\text{loc}}^{2p} - \frac{2N}{2p-1}L_{\text{loc}}^{2p-1} - N^{p+(1/2)}\frac{2+2p-4^p}{(2p+1)p(2p-1)(p-1)}\right)\right]. \quad (25)$$

C. Three-Tier Wire-Limited Chip Size

The expression for the total required area of a three-tier multilevel wiring network is given by

$$A_{\text{required}} = \sqrt{\frac{A_c}{N}}(p_{\text{loc}}L_{\text{total,loc}} + p_{\text{semi}}L_{\text{total,semi}} + p_{\text{global}}L_{\text{total,gobal}}) \quad (26)$$

where A_c is the chip area, N is the number of gates, p_{loc} is the local pitch, p_{semi} is the semiglobal pitch, p_{glob} is the global pitch, $L_{\text{total,loc}}$ is the total length of the local interconnects, $L_{\text{total,semi}}$ is the total length of the semiglobal interconnects, and $L_{\text{total,global}}$ is the total length of the global interconnects. Using the wire-length distribution, the expressions for the total length of the local, semiglobal, and global interconnects are

$$L_{\text{total,loc}} = \chi\int_1^{L_{\text{loc}}} \ell i(\ell)\,d\ell \quad (27)$$

$$L_{\text{total,semi}} = \chi\int_{L_{\text{loc}}}^{L_{\text{semi}}} \ell i(\ell)\,d\ell \quad (28)$$

$$L_{\text{total,glob}} = \chi\int_{L_{\text{semi}}}^{2\sqrt{N}} \ell i(\ell)\,d\ell. \quad (29)$$

Assuming that the longest local interconnect is less than the chip edge, then the total length of the local interconnects is given by (21). Because the interconnect density function is piecewise defined, the total length of the semiglobal and global interconnects is dependent on whether the length of the longest semiglobal interconnect L_{semi} is less than or greater than the chip edge. The expressions for total interconnect length for both cases are

Case I: $L_{\text{semi}} < \sqrt{N}$

$$L_{\text{total,semi}} = \chi\int_{L_{\text{loc}}}^{L_{\text{semi}}} \Gamma\frac{\alpha k}{2}\left(\frac{\ell^3}{3} - 2\sqrt{N}\ell^2 + 2N\ell\right)\ell^{2p-3}\,d\ell$$

$$L_{\text{total,glob}} = \chi\int_{L_{\text{semi}}}^{\sqrt{N}} \Gamma\frac{\alpha k}{2}\left(\frac{\ell^3}{3} - 2\sqrt{N}\ell^2 + 2N\ell\right)\ell^{2p-3}\,d\ell + \chi\int_{\sqrt{N}}^{2\sqrt{N}} \Gamma\frac{\alpha k}{6}(2\sqrt{N}-\ell)^3\ell^{2p-3}\,d\ell. \quad (30)$$

Case II: $L_{\text{semi}} > \sqrt{N}$

$$L_{\text{total,semi}} = \chi\int_{L_{\text{loc}}}^{\sqrt{N}} \Gamma\frac{\alpha k}{2}\left(\frac{\ell^3}{3} - 2\sqrt{N}\ell^2 + 2N\ell\right)\ell^{2p-3}\,d\ell + \chi\int_{\sqrt{N}}^{L_{\text{semi}}} \Gamma\frac{\alpha k}{6}(2\sqrt{N}-\ell)^3\ell^{2p-3}\,d\ell$$

$$L_{\text{total,glob}} = \chi\int_{L_{\text{semi}}}^{2\sqrt{N}} \Gamma\frac{\alpha k}{6}(2\sqrt{N}-\ell)^3\ell^{2p-3}\,d\ell. \quad (31)$$

Evaluating (30) and (31) and substituting into (26) gives the expression for the total required wiring area for both cases

Case I: $L_{\text{semi}} < \sqrt{N}$

$$A_{\text{req}} = \sqrt{\frac{A_c}{N}}\,\frac{\Gamma\alpha k\chi}{2}\left[p_{\text{loc}}\left(\frac{L_{\text{loc}}^{2p+1}-1}{6p+3} - \frac{\sqrt{N}(L_{\text{loc}}^{2p}-1)}{p} - \frac{2N(1-L_{\text{loc}}^{2p-1})}{2p-1}\right) + p_{\text{semi}}\left(\frac{L_{\text{semi}}^{2p+1}-L_{\text{loc}}^{2p+1}}{6p+3} - \frac{\sqrt{N}(L_{\text{semi}}^{2p}-L_{\text{loc}}^{2p})}{p} - \frac{2N(L_{\text{semi}}^{2p-1}-L_{\text{loc}}^{2p-1})}{2p-1}\right) + p_{\text{glob}}\left(\frac{(2\sqrt{N})^{2p+1}+L_{\text{semi}}^{2p+1}}{6p+3} + \frac{\sqrt{N}(L_{\text{semi}}^{2p}+(2\sqrt{N})^{2p})}{p} - \frac{2N(L_{\text{semi}}^{2p-1}+2(2\sqrt{N})^{2p-1})}{2p-1} - \frac{4\sqrt[3]{N}}{3p-1}2(2\sqrt{N})^{2p-2}\right)\right]. \quad (32)$$

Case II: $L_{\text{semi}} > \sqrt{N}$

$$A_{\text{req}} = \sqrt{\frac{A_c}{N}}\,\frac{\Gamma\alpha k\chi}{2}\left[p_{\text{loc}}\left(\frac{L_{\text{loc}}^{2p+1}-1}{6p+3} - \frac{\sqrt{N}(L_{\text{loc}}^{2p}-1)}{p} - \frac{2N(1-L_{\text{loc}}^{2p-1})}{2p-1}\right) + p_{\text{semi}}\left(-2N^{p+(1/2)}\frac{p+1}{(2p+1)p(2p-1)(p-1)} - \frac{L_{\text{semi}}^{2p+1}+L_{\text{loc}}^{2p+1}}{6p+3} - \frac{\sqrt{N}(L_{\text{semi}}^{2p}+L_{\text{loc}}^{2p})}{p} - \frac{2N(2L_{\text{semi}}^{2p-1}+L_{\text{loc}}^{2p-1})}{2p-1} + \frac{4\sqrt[3]{N}}{3p-1}L_{\text{semi}}^{2p-2}\right) + p_{\text{glob}}\left(\frac{L_{\text{semi}}^{2p+1}-(2\sqrt{N})^{2p+1}}{6p+3} + \frac{\sqrt{N}((2\sqrt{N})^{2p}-L_{\text{semi}}^{2p})}{p} + \frac{4N(L_{\text{semi}}^{2p-1}-(2\sqrt{N})^{2p-1})}{2p-1} + \frac{4\sqrt[3]{N}((2\sqrt{N})^{2p-2}-L_{\text{semi}}^{2p-2})}{3p-1}\right)\right]. \tag{33}$$

Substituting (32) and (33) into (24) gives the final expressions for the wire limited chip area

Case I: $L_{\text{semi}} < \sqrt{N}$

$$A_c = \sqrt{\frac{A_c}{N}}\,\frac{\Gamma\alpha k\chi}{2e_w n_{\text{levels}}}\left[p_{\text{loc}}\left(\frac{L_{\text{loc}}^{2p+1}-1}{6p+3} - \frac{\sqrt{N}(L_{\text{loc}}^{2p}-1)}{p} - \frac{2N(1-L_{\text{loc}}^{2p-1})}{2p-1}\right) + p_{\text{semi}}\left(\frac{L_{\text{semi}}^{2p+1}-L_{\text{loc}}^{2p+1}}{6p+3} - \frac{\sqrt{N}(L_{\text{semi}}^{2p}-L_{\text{loc}}^{2p})}{p} - \frac{2N(L_{\text{semi}}^{2p-1}-L_{\text{loc}}^{2p-1})}{2p-1}\right) + p_{\text{glob}}\left(\frac{(2\sqrt{N})^{2p+1}+L_{\text{semi}}^{2p+1}}{6p+3} + \frac{\sqrt{N}(L_{\text{semi}}^{2p}+(2\sqrt{N})^{2p})}{p} - \frac{2N(L_{\text{semi}}^{2p-1}+2(2\sqrt{N})^{2p-1})}{2p-1} - \frac{4\sqrt[3]{N}}{3p-1}2(2\sqrt{N})^{2p-2}\right)\right]. \tag{34}$$

Case II: $L_{\text{semi}} > \sqrt{N}$

$$A_c = \sqrt{\frac{A_c}{N}}\,\frac{\Gamma\alpha k\chi}{2e_w n_{\text{levels}}}\left[p_{\text{loc}}\left(\frac{L_{\text{loc}}^{2p+1}-1}{6p+3} - \frac{\sqrt{N}(L_{\text{loc}}^{2p}-1)}{p} - \frac{2N(1-L_{\text{loc}}^{2p-1})}{2p-1}\right) + p_{\text{semi}}\left(-2N^{p+(1/2)}\frac{p+1}{(2p+1)p(2p-1)(p-1)} - \frac{L_{\text{semi}}^{2p+1}+L_{\text{loc}}^{2p+1}}{6p+3} - \frac{\sqrt{N}(L_{\text{semi}}^{2p}+L_{\text{loc}}^{2p})}{p} - \frac{2N(2L_{\text{semi}}^{2p-1}+L_{\text{loc}}^{2p-1})}{2p-1} + \frac{4\sqrt[3]{N}}{3p-1}L_{\text{semi}}^{2p-2}\right) + p_{\text{glob}}\left(\frac{L_{\text{semi}}^{2p+1}-(2\sqrt{N})^{2p+1}}{6p+3} + \frac{\sqrt{N}((2\sqrt{N})^{2p}-L_{\text{semi}}^{2p})}{p} + \frac{4N(L_{\text{semi}}^{2p-1}-(2\sqrt{N})^{2p-1})}{2p-1} + \frac{4\sqrt[3]{N}((2\sqrt{N})^{2p-2}-L_{\text{semi}}^{2p-2})}{3p-1}\right)\right]. \tag{35}$$

D. A Case Study Comparison

1) Specifications: The physical parameters for this case study are taken from the National Technology Roadmap for Semiconductors (NTRS) for the year 2001 [11]. A system with eight million gates is examined, the physical characteristics of the system are outlined in Table I.

2) Assumptions and Models: A key assumption for the geometrical construction of each tier of the multilevel network is that all cross-sectional dimensions are equal (i.e., $W = S = H_\rho = H_\varepsilon$) and that the boundary between each tier has an oxide thickness of the tier with the larger interconnect dimensions.

The system clock frequency is determined from the enhanced critical path model used in Section II. For simplicity, the longest interconnects on the local and semiglobal tier have a time delay that is no greater than 25% of the clock period. The longer global interconnects have a full 90% of the clock period.

Assuming that the longest interconnect for each tier is driven by a single driver and the line capacitance is much greater that the transistor load capacitance, then the expression for the approximate 50% time delay of the circuit is [2], [12]

$$\tau = 0.4 r_{\text{line}} c_{\text{line}} x^2 + 0.7 R_{\text{tr}} c_{\text{line}} x + \frac{x\sqrt{\varepsilon_r}}{c_o} \tag{36}$$

where x is the length of the line in centimeters, r_{line} and c_{line} are the distributed resistance and capacitance of the line, R_{tr} is the equivalent transistor resistance, ε_r is the relative dielectric constant of the interlevel dielectric material, and c_o is the speed of light in free space. Making the assumption that the driver resistance matches the total line resistance (i.e., $xr_{\text{line}} = R_{\text{tr}}$), then (36) simplifies to

$$\tau = 1.1 r_{\text{line}} c_{\text{line}} x^2 + \frac{x\sqrt{\varepsilon_r}}{c_o}. \tag{37}$$

Assuming that the longest line in the critical path is given a maximum time delay that is some fraction, β, of the clock period and using the expression for line capacitance with the three-conductor two-ground plane model after [10] where all wire cross-sectional dimensions on each tier are equal, then

(37) becomes

$$\frac{\beta}{f_c} = 4\frac{1.1\rho\varepsilon_r\varepsilon_o 6.08}{p_w^2}x^2 + x\frac{\sqrt{\varepsilon_r}}{c_o} \tag{38}$$

where f_c is the clock frequency, p_w is the interconnect pitch of the tier, ρ is the resistivity of the metal, c_o is the speed of light in free space, and ε_r is the relative dielectric constant of the material.

The relationship between the length in centimeters x and the length in units of average gate pitches L is $x = \sqrt{(A_c/N)}L$ (where A_c is the chip area and N is the number of gates). Then (38) becomes

$$\frac{\beta}{f_c} = 4\frac{1.1\rho\varepsilon_r\varepsilon_o 6.08}{p_w^2}\,\frac{A_c}{N}L^2 + L\sqrt{\frac{A_c}{N}}\,\frac{\sqrt{\varepsilon_r}}{c_o}. \tag{39}$$

Solving (39) for the wire pitch p_w, as a function of the interconnect length L [gate pitches] gives

$$p_w = 2L\sqrt{\frac{A_c}{N}}\sqrt{\frac{1.1\rho\varepsilon_r\varepsilon_o 6.08}{\left(\frac{\beta}{f_c} - L\sqrt{\frac{A_c}{N}}\,\frac{\sqrt{\varepsilon_r}}{c_o}\right)}}. \tag{40}$$

Solving (39) for the interconnect length L [gate pitches], as a function of the wire pitch p_w gives

$$L = \frac{1}{4}\,\frac{p_w^2}{2.2\rho\varepsilon_r\varepsilon_o 6.08}\sqrt{\frac{N}{A_c}} \cdot \left[-\frac{\sqrt{\varepsilon_r}}{c_o} + \sqrt{\frac{\varepsilon_r}{c_o^2} + 4\beta\frac{4.4\rho\varepsilon_r\varepsilon_o 6.08}{p_w^2 f_c}}\right]. \tag{41}$$

3) Two-Tier Results: The equation used to determine the wire-limited chip area for the two-tier network is given by (25). The local pitch p_{loc} is $2F$, where F is the minimum feature size and L_{glob} is a die-edge-length interconnect. The global pitch p_{glob} and the longest local interconnect L_{loc} are determined by the following expressions:

$$p_{\text{glob}} = \sqrt{\frac{A_c}{N}}2L_{\text{glob}}\sqrt{\frac{1.1\rho\varepsilon_r\varepsilon_o 6.08}{\left(\frac{\beta_{\text{glob}}}{f_c} - \sqrt{\frac{A_c}{N}}L_{\text{glob}}\frac{\sqrt{\varepsilon_r}}{c_o}\right)}}$$

$$L_{\text{loc}} = \frac{p_{\text{loc}}^2}{4(2.2)\rho\varepsilon_r\varepsilon_o 6.08}\sqrt{\frac{N}{A_c}} \cdot \left[-\frac{\sqrt{\varepsilon_r}}{c_o} + \sqrt{\frac{\varepsilon_r}{c_o^2} + 4\beta_{\text{loc}}\frac{4.4\rho\varepsilon_r\varepsilon_o 6.08}{p_{\text{loc}}^2 f_c}}\right] \tag{42}$$

where β_{glob} and β_{loc} are the fraction of the clock period for the longest global and local interconnect, respectively. This two-tier network was designed to have a clock frequency of 410 MHz within six metal levels which is consistent with the National Technology Roadmap for Semiconductors [11]. The results of the two-tier network are outlined in Table II.

This multilevel network is now redesigned as a three-tier network that either minimizes wire-limited chip size or maximizes the on-chip clock frequency.

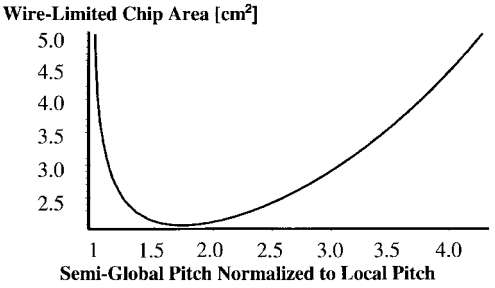


Fig. 7. Wire-limited chip area versus semiglobal pitch.

TABLE III
THREE-TIER THAT MINIMIZES CHIP AREA

PARAMETERS	PITCH	# OF LEVELS
Local	0.36μm	3
Semi-Global	0.61μm	1
Global	0.98μm	2
**********	********	********
Clock Frequency	410Mhz	
Wire-Limited Chip Area	2.08cm^2	
Transistor Limited Chip Area	N *x* 200F^2=0.52 cm^2	

4) Three-Tier Minimum Chip Size Results: The clock frequency for this three-tier network optimization is set to 410 MHz, and the semiglobal pitch is calculated to minimize the wire limited chip size. The expressions in (34) and (35) determine the wire-limited chip area for a given semiglobal pitch. The local pitch p_{loc} is $2F$, where F is the minimum feature size, and L_{glob} is assumed to be equal to the chip edge. When expressions (34) and (35) are used in conjunction with the following expressions for the global pitch p_{glob}, the longest semiglobal length L_{semi}, and the longest local length L_{loc}, then the wire-limited chip area can be estimated

$$p_{\text{glob}} = 2\sqrt{\frac{A_c}{N}}L_{\text{glob}}\sqrt{\frac{1.1\rho\varepsilon_r\varepsilon_o 6.08}{\left(\frac{\beta_{\text{glob}}}{f_c} - \sqrt{\frac{A_c}{N}}L_{\text{glob}}\frac{\sqrt{\varepsilon_r}}{c_o}\right)}}$$

$$L_{\text{semi}} = \frac{p_{\text{semi}}^2}{4(2.2)\rho\varepsilon_r\varepsilon_o 6.08}\sqrt{\frac{N}{A_c}} \cdot \left[-\frac{\sqrt{\varepsilon_r}}{c_o} + \sqrt{\frac{\varepsilon_r}{c_o^2} + 4\beta_{\text{semi}}\frac{4.4\rho\varepsilon_r\varepsilon_o 6.08}{p_{\text{semi}}^2 f_c}}\right]$$

$$L_{\text{loc}} = \frac{p_{\text{loc}}^2}{4(2.2)\rho\varepsilon_r\varepsilon_o 6.08}\sqrt{\frac{N}{A_c}} \cdot \left[-\frac{\sqrt{\varepsilon_r}}{c_o} + \sqrt{\frac{\varepsilon_r}{c_o^2} + 4\beta_{\text{loc}}\frac{4.4\rho\varepsilon_r\varepsilon_o 6.08}{p_{\text{loc}}^2 f_c}}\right]. \tag{43}$$

A graph of the implicit equation for wire-limited chip area in (34) as a function of the semiglobal pitch normalized to the minimum feature size pitch $2F$ appears in Fig. 7.

From Fig. 7, a definite optimal semiglobal pitch clearly exists and has a value of $1.7p_{\text{loc}}$. The physical characteristics of this three-tier multilevel network that is optimized for minimum chip area are detailed in Table III. This optimal chip size is 42% of the chip size for the two-tier network.

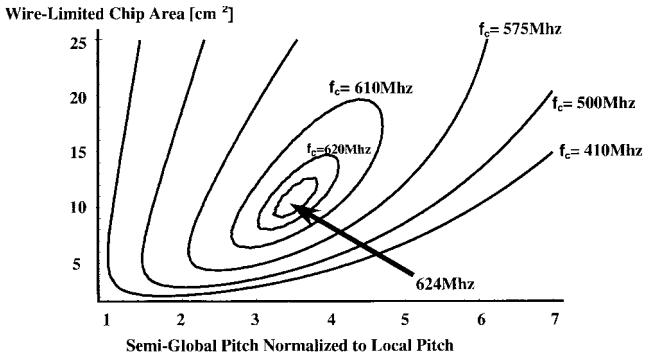


Fig. 8. Clock frequency versus minimum wire-limited chip area.

TABLE IV
THREE-TIER DESIGN THAT MAXIMIZES CLOCK FREQUENCY

PARAMETERS	PITCH	# OF LEVELS
Local	0.36μm	1
Semi-Global	1.2μm	2
Global	2.7μm	3
**********	********	********
Clock Frequency	624Ghz	
Wire-Limited Chip Area	9.64cm^2	
Transistor Limited Chip Area	N x 200F^2=0.52 cm^2	

5) Three-Tier Maximum Clock Frequency Results: This design maximizes the clock frequency of a multilevel wiring network given the constraints listed in Table I. Because the three-tier network increases the wiring density of the multilevel network, it can accommodate larger and therefore higher speed global interconnects within the six levels of metal. The chip area equation for the three-tier network in (34) and (35) is still used, but the clock frequency is varied.

The local pitch p_{loc} for this three-tier multilevel network is $2F$, where F is the minimum feature size. When the expressions in (34) and (35) are used in conjunction with the equations in (43), then the final implicit expression for the chip area is obtained. The graphical representation of the chip area for different clock frequencies appears in Fig. 8. The contours of different frequencies in Fig. 8 culminate to a peak frequency of 624 MHz.

The optimal clock frequency in Fig. 8 exists because increases in chip size and clock frequency reduce the length of the longest local interconnect, L_{loc}, in (43) and, therefore, increase the number of semiglobal and global interconnects. Incremental increases in chip area provide additional area to scale global cross-sectional dimensions and increase clock frequency; however, the incremental increase in chip area also increases the number of semiglobal and global interconnects which *lessens* the area available to scale global wiring dimensions. The optimal clock frequency occurs at the point where infinitesimal increases in chip area provides insufficient area available to scale global wires to increase clock frequency. Any attempt to increase the clock frequency beyond the optimal point in Fig. 8 causes the required wiring area in (26) to be greater than the available wiring area in (24), thereby making larger clock frequencies impossible to attain given constraints of the system in Table I.

The results of the three-tier network that maximizes clock frequency is outlined in Table IV. The maximum clock frequency achievable by the three-tier network is 1.5 times larger than the conventional two-tier network.

V. CONCLUSION

Based upon Rent's Rule, a new complete stochastic wiring distribution is utilized to enhance a critical path model; to develop a preliminary power dissipation model; and to determine optimal interconnect scaling in a three-tier multilevel network that minimizes chip size or maximizes clock frequency. For a given case study corresponding to 2001 technology [11], the optimized three-tier multilevel network provides a $0.42\times$ reduction in chip size or $1.5\times$ increase in clock frequency over the conventional two-tier design.

REFERENCES

[1] J. D. Meindl, "Low power microelectronics: Retrospective and prospect," *Proc. IEEE*, vol. 83, pp. 619–635, Apr. 1995.
[2] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.
[3] G. A. Sai-Halaz, "Performance trends in high-end processors," *Proc. IEEE*, vol. 83, p. 20–36, Jan. 1995.
[4] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI)—Part I: Derivation and validation," *Trans. Electron Devices*, this issue, pp. 580–589.
[5] J. D. Meindl and J. Davis, "Interconnect performance limits on gigascale integration (GSI)," *Mater. Chem. Phys.*, vol. 41, p. 161–166, 1995.
[6] W. E. Donath, "Wire-length distribution for placements of computer logic," *IBM J. Res. Develop.*, vol. 2, no. 3, p. 152–155, May 1981.
[7] J. R. Brews, "Electrical modeling of interconnects," in *Submicron Integrated Circuits*. New York: Wiley, 1989.
[8] W. E. Donath, "Placement and average interconnections lengths of computer logic," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 272–277, Apr. 1979.
[9] A. P. Chandrakasan and R. W. Broderson, *Low Power Digital CMOS Design*. Norwood, MA: Kluwer, 1995.
[10] J. H. Chern, J. Huang, L. Arledge, P. C. Li, and P. Yang, "Multilevel metal capacitance models for CAD design synthesis systems," *IEEE Electron Device Lett.*, vol. 13, pp. 32–33, Jan. 1992.
[11] Semiconductor Industry Assoc., *The National Technology Roadmap for Semiconductors*, 1994, pp. 12–16.
[12] T. Sakurai, "Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSI's," *IEEE Trans. Electron Devices*, vol. 40, pp. 118–124, Jan. 1993.

Jeffrey A. Davis, for a photograph and biography, see this issue, p. 589.

Vivek K. De, for a photograph and biography, see this issue, p. 589.

James D. Meindl (M'56–SM'66–F'68–LF'97), for a photograph and biography, see this issue, p. 589.