Interconnect opportunities for gigascale integration

by J. D. Meindl
J. A. Davis
P. Zarkesh-Ha
C. S. Patel
K. P. Martin
P. A. Kohl

Throughout the past four decades, semiconductor technology has advanced at exponential rates in both productivity and performance. In recent years, multilevel interconnect networks have become the primary limit on the productivity, performance, energy dissipation, and signal integrity of gigascale integration. Consequently, a broad spectrum of novel solutions to the multifaceted interconnect problem must be explored. Here we review recent salient results of this exploration. Based upon prediction of the complete stochastic signal interconnect length distribution of a megacell, optimal reverse scaling of each pair of wiring levels provides a prime opportunity to minimize cell area, clock period, power dissipation, or number of wiring levels. Using a heterogeneous version of Rent's rule, a design methodology for the global signal, clock, and power/ground distribution networks for a system-on-a-chip has been derived. Wiring area, bandwidth, and signal integrity are the prime constraints on the design of the networks. Three-dimensional integration offers the opportunity to reduce the length of the longest global interconnects in a distribution by as much as 75%. Wafer-level batch fabrication of chip input/output interconnects and chip scale packages provides new benefits such as I/O bandwidth enhancement, simultaneous switching-noise reduction, and lower cost of packaging and testing. Microphotonic interconnects have long-term potential to reduce latency, power dissipation, and crosstalk while increasing bandwidth.

1. Introduction

Semiconductor productivity and performance have increased at exponential rates in the last forty years. Three generic strategies have guided these advances: 1) scaling down minimum feature size, 2) increasing die size, and 3) enhancing packing efficiency (defined as the number of transistors or length of interconnect per minimum feature square of silicon area). Scaling of transistors reduces their cost, intrinsic switching delay, and energy dissipation per binary transition. Scaling of interconnects serves to reduce cost but *increases* latency (response time) in absolute value and energy dissipation relative to that of transistors. These increases result from relatively larger average interconnect lengths (measured in gate pitches) and larger die sizes for successive generations. Therefore, interconnects have become the primary limit on both the performance and the energy dissipation of gigascale integration (GSI).

Following this brief introduction, Section 2 quantifies the key facets of the interconnect problem. The principal generic opportunities to resolve it, including new materials

©**Copyright** 2002 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

0018-8646/02/$5.00 © 2002 IBM

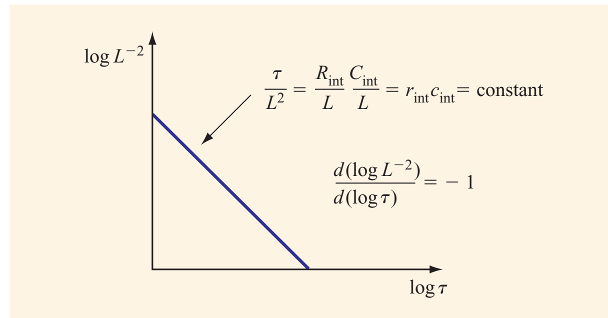


Figure 1

Interconnect reciprocal length squared $(1/L)^2$ vs. latency (τ) with log scales illustrating diagonal lines as loci of constant distributed resistance–capacitance product ($r_{int}c_{int}$). Reprinted with permission from [7]; © 1998 IEEE.

and processes, scaling, and novel architectures, are reviewed in Section 3 with emphasis on scaling. Reverse scaling of multilevel interconnect networks is based upon prediction of stochastic signal wiring distributions to achieve minimum area, power dissipation, clock period, or number of metal levels. A methodology to derive an integrated architecture for global signal, power, and clock distribution networks for a system-on-a-chip is reviewed in Section 4. Sections 5, 6, and 7 explore three unconventional approaches to alleviating the on-chip interconnect problem. These are respectively novel three-dimensional structures, high-density input/output interconnect enhancements, and compatible microphotonic interconnects. A brief conclusion is provided in Section 8.

2. The interconnect problem

What is the quintessential purpose of an interconnect? In a single word, it is *communication*. To give a more complete definition, it is communication between distant points with small latency. A lucid illustration that displays this key purpose is a graph whose vertical axis is reciprocal interconnect length squared and whose horizontal axis is interconnect latency [1]. Using logarithmic scales on both axes, a diagonal line is a locus of constant distributed resistance–capacitance product, the principal figure of merit of the large majority of interconnects used for GSI. As illustrated in **Figure 1**, reducing the distributed resistance–capacitance product moves the diagonal locus toward the lower left corner of the figure, providing smaller latency for a given interconnect length. However, during the past four decades interconnect scaling has *increased* the distributed resistance–capacitance product, moving toward the upper right corner of the figure and therefore demanding larger latency for a given interconnect length. In stark contrast, scaling of transistors reduces the power–delay product or switching energy of a binary transition, therefore moving toward the lower left corner of the power–delay plane to reduce simultaneously both average power transfer and delay.

To quantify the exploding disparity between the latency of interconnects and transistors, consider the comparisons illustrated in **Table 1**. For the 1-μm-generation technology of the late 1980s, the "*CV/I*," or intrinsic switching delay of a MOSFET [2] before it is loaded with parasitic or wiring capacitance, is approximately 20 ps. However, for the same generation, the total resistance–capacitance product or *RC* delay of a "benchmark" 1.0-mm-long interconnect is about 1.0 ps. In comparison, for the 100-nm generation projected for early production in 2005, the *CV/I* delay of a MOSFET *decreases* to 5 ps, while the *RC* latency of a 1.0-mm-long wire *increases* to 30 ps. The relevant observation is that as semiconductor technology is advancing from the 1.0-μm to the 100-nm generation, the *RC* delay or response time of a benchmark 1.0-mm-long interconnect is devolving from 20 times faster to six times slower than transistor intrinsic switching delay. Furthermore, the 1999 *International Technology Roadmap for Semiconductors* (ITRS) projection for 35-nm technology in 2014 suggests a 2.5-ps transistor delay and a 250-ps *RC* latency for a 1.0-mm-long interconnect [3]. For completeness, the time of flight (ToF) of a 1.0-mm-long interconnect is included in Table 1. As indicated, ToF delay is independent of scaling but does depend on the value of the relative permittivity of the interconnect dielectric.

To underscore the formidable challenge presented by interconnects to continued performance improvements for GSI, it is noteworthy that the numerical values for *RC* delay cited in Table 1 represent simple *best-case* calculations. For example, the results do not account for the adverse results of surface scattering, high-frequency skin effect, liner thickness for copper interconnects, or temperature rises in a multilevel wiring network.

Beyond latency, interconnects present an energy-dissipation problem illustrated in **Table 2** that also limits the performance of GSI as a consequence of practical constraints on the heat-removal capacity of the package of a gigascale chip or the energy-storage capacity of its portable power source. Again comparing technology generations, it is evident that the energy dissipation associated with a binary transition of a minimum-geometry MOSFET versus a 1.0-mm-long interconnect is respectively 33%, five times, and thirty times larger for the interconnect for the 1.0-μm, 100-nm, and 35-nm-technology generations. These gross imbalances clearly indicate that the power-dissipation problem of gigascale chips is essentially an interconnect problem.

Table 1 MOSFET and interconnect latency for 1.0-μm, 100-nm, and 35-nm-technology generations [3].

Technology generation	*MOSFET switching delay* ($t_d = CV/I$) (ps)	*RC response time* (L_{int} = 1 mm) (ps)	*Time of flight* (L_{int} = 1 mm) (ps)
1.0 μm (Al, SiO_2)	~20	~1	~6.6
100 nm (Cu, k = 2.0)	~5	~30	~4.6
35 nm (Cu, k = 2.0)	~2.5	~250	~4.6

Table 2 ITRS projections for switching delay, switching energy, clock frequency, total chip current drain, maximum number of wiring levels, maximum total wire length per chip, and chip pad count for 1.0-μm, 100-nm, and 35-nm-technology generations [3].

	Technology generation		
	1.0 μm	*100 nm*	*35 nm*
MOSFET switching delay (ps)	~20	~5	~2.5
Interconnect *RC* response time (ps) (L_{int} = 1 mm)	~1	~30	~250
MOSFET switching energy (fJ)	~300	~2	~0.1
Interconnect switching energy (fJ)	~400	~10	~3
Clock frequency	~30 MHz	~2–3.5 GHz	~3.6–13.5 GHz
Supply current (A) (V_{dd} = 5.0, 1.0, 0.5 V)	~2.5	~150	~360
Maximum number of wiring levels	3	8–9	10
Maximum total wire length per chip (m)	~100	~5000	—
Chip pad count	~200	~3000–4000	4000–4400

The preceding discussion of latency and energy-dissipation problems presented by interconnects is concerned with signal wiring. Historical records and ITRS projections [3] of clock frequencies for high-performance microprocessors summarized in Table 2 indicate 30 MHz, 3.0 GHz, and 13 GHz as the respective nominal clock frequencies for the 1.0-μm, 100-nm, and 35-nm-technology generations. These rapidly escalating requirements place enormous new demands on the interconnects that implement clock distribution networks of gigascale chips. Bandwidth, power dissipation, variation in the time of arrival of a clock pulse at different points on a chip (skew), and differences in clock pulse width (jitter) represent increasingly formidable issues.

Although gigascale signal and clock distribution network problems are daunting, power distribution may well match them in difficulty. As noted in Table 2, estimated maximum chip current drain is respectively 2.5 A, 150 A, and 360 A for the 1.0-μm, 100-nm, and 35-nm-technology generations. Concurrently, power-supply voltage scales down from 5.0 V to 1.0 V to 0.5 V for the corresponding generations. These aggressive expectations for high-current, low-voltage power distribution impose utterly unprecedented demands on interconnect networks.

Finally, the targets for number of wiring levels, maximum total interconnect length, and number of bonding pads or input/output interconnects per chip cited in Table 2 add significantly to expectations for future interconnect capabilities. In short, the highly demanding requirements that are projected for on-chip wiring compel comprehensive research over the most extensive and multidimensional solution space that can be conceived.

3. Reverse scaling

Approximate expressions for the latency (τ) of a single isolated interconnect that is *RC* limited with an ideal return path are given by

$$\tau_{90\%} \cong r_{int}c_{int}L^2 + 2.3R_{tr}c_{int}L + 2.3C_L(r_{int}L + R_{tr}), \quad (1a)$$

$$\tau_{90\%} \cong r_{int}c_{int}L^2 + 2.3R_{tr}c_{int}L \text{ for } C_L \ll c_{int}L, \quad (1b)$$

and

$$\tau_{90\%} \cong r_{int}c_{int}L^2 \text{ for } C_L \ll c_{int}L \text{ and } R_{tr} \ll r_{int}L, \quad (1c)$$

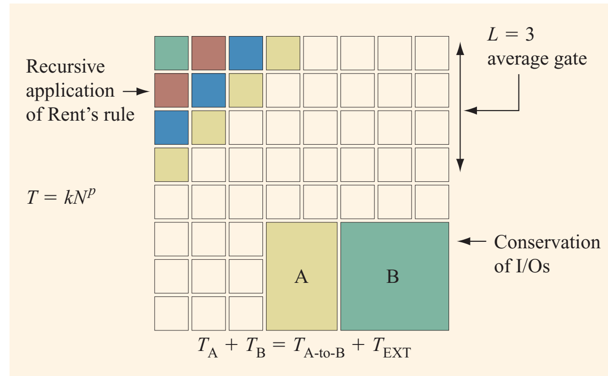


Figure 2

Diagram of a macrocell consisting of a random logic network of N microcells or individual logic gates configured in a square array. Rent's rule and the principle of conservation of interconnects are applied recursively to derive the complete stochastic interconnect length distribution of the random logic network.

where r_{int} and c_{int} are the interconnect resistance and capacitance per unit length, respectively, R_{tr} is the source resistance, C_L is the load capacitance, and L is the interconnect length. The latency of a low-resistance interconnect that is resistance-, inductance-, and capacitance- or RLC-limited is given by

$$\tau_{90\%} \cong \text{ToF} = L/[c_0/(\varepsilon_r)^{1/2}], \tag{2a}$$

where

$$\frac{R_{int}}{Z_0} \le 2 \ln \left(\frac{4Z_0}{R_{tr} + Z_0} \right), \qquad R_{tr} < 3Z_0, \qquad \text{and } C_L \ll c_{int} L \tag{2b}$$

are required for ToF response. In Equations (2), Z_0 is the characteristic impedance and $R_{int} = r_{int}L$ is the total resistance of the interconnect; c_0 is the velocity of light in free space, and ε_r is the relative permittivity of the interconnect insulator. Since RC-limited performance is far more common than ToF limitations, the RC case is considered in this section.

The simple relationship given by Equation (1c) serves as the basis for reviewing the principal generic opportunities for solving the key latency problem. The latency of an RC-limited interconnect can be expressed as the product of three factors, as indicated in Equation (3):

$$\tau = [\rho\varepsilon]\left[\frac{1}{HT}\right][L^2]. \tag{3}$$

The resistivity–permittivity factor $[\rho\varepsilon]$ identifies opportunities to reduce latency through new materials and processes such as the replacement of aluminum with copper [4]. The $[1/HT]$ factor, where H defines metal height and T defines insulator thickness, represents device- and circuit-level [1] opportunities to reduce latency through reverse scaling. Finally, L defines the length of an interconnect, and the $[L^2]$ factor represents system-level [5] opportunities to improve latency through the use of new microarchitectures that serve to "keep interconnects short." Solutions to the latency problem must be pursued at each of the levels represented in Equation (3): material and process, device, circuit, and system [1]. The scope of this section is confined to device-, circuit-, and system-level opportunities to reduce latency through reverse scaling. In comparison to alternatives such as new materials and processes as well as novel architectures, the compelling advantages of reverse scaling are 1) minimal time to implementation, 2) low cost of implementation, 3) low risk, and 4) high payoff.

The key to optimal reverse scaling is the capability to predict the complete stochastic interconnect density distribution for a projected next-generation product. Consider the case of a macrocell consisting of a random logic network of N microcells or logic gates. As illustrated in **Figure 2**, the macrocell can be modeled as a square array of logic gates. Rent's rule ($R = kN^p$) [6] and the principle of conservation of interconnects are applied recursively to the macrocell, as indicated in Figure 2. A closed-form expression for the complete stochastic signal wiring distribution resulting from this process is given [7] by the following:

Region 1: $1 \le L < \sqrt{N}$,

$$f(L) = \Gamma \frac{\alpha k}{2} \left(\frac{L^3}{3} - 2\sqrt{N} L^2 + 2NL \right) L^{2p-4}; \tag{4a}$$

Region 2: $\sqrt{N} \le L \le 2\sqrt{N}$,

$$f(L) = \Gamma \frac{\alpha k}{6} (2\sqrt{N} - L)^3 L^{2p-4}; \tag{4b}$$

$$\Gamma = \frac{2N(1 - N^{p-1})}{-N^p \dfrac{1 + 2p - 2^{2p-1}}{p(2p-1)(p-1)(2p-3)} - \dfrac{1}{6p} + \dfrac{2\sqrt{N}}{2p-1} - \dfrac{N}{p-1}}. \tag{4c}$$

Equation (4a) applies to shorter interconnects and Equation (4b) to longer interconnects in the distribution. These expressions reveal the dependence of interconnect density [$f(L)$ in units of number of interconnects of length L per gate pitch] versus interconnect length L in gate pitches. The dependencies on interconnect length L, number of gates in the network N, Rent's coefficient k, and Rent's exponent p are evident. As demonstrated in

Figure 3, this stochastic wiring distribution is found to be in close agreement with experimental data characterizing commercial products [7]. The key to obtaining close agreement between predicted and actual wiring distributions for a new product is to derive appropriate values of Rent's coefficient k and exponent p using data from previous generations of a product family. These two empirical parameters appear to have *genetic* characteristics.

An optimal architecture for a multilevel interconnect network that minimizes macrocell area, power dissipation, clock cycle time, or number of wiring levels can be derived using the stochastic interconnect distribution given by Equation (4). A derivation for minimum macrocell area begins with the formulation of a wiring area "supply and demand" equation (5a) [8]:

$$2e_{\mathrm{w}}A_{\mathrm{m}} = \chi p_n \sqrt{\frac{A_{\mathrm{m}}}{N}} \int_{L_{n-1}}^{L_n} Lf(L)\, dL; \tag{5a}$$

$$p_n = 2\sqrt{\frac{1.1\rho\varepsilon_{\mathrm{r}}\varepsilon_0 6.2 f_{\mathrm{c}}}{\beta}} \sqrt{\frac{A_{\mathrm{m}}}{N}}\, L_n; \tag{5b}$$

$$p_n = 2.5\,\frac{2f_{\mathrm{c}}}{\beta}\sqrt{6.2\rho\varepsilon_{\mathrm{r}}\varepsilon_0 R_0 C_0}\sqrt{\frac{A_{\mathrm{m}}}{N}}\, L_n. \tag{5c}$$

The *available* area for an orthogonal pair of wiring levels can be expressed as $2e_{\mathrm{w}}A_{\mathrm{m}}$, where e_{w} is a wiring efficiency factor that must be estimated from previous designs and A_{m} is the area of the macrocell. The *required* area is defined by the right-hand side of Equation (5a), where $\chi < 1$ converts point-to-point wire length to net length [7]. (Net length is the total length of wiring that connects the output terminal of a driver gate to the inputs of its load gates.) The factor p_n is wire pitch, the square-root factor is gate pitch (in cm), and the integral represents the total length of wire in gate pitches between its upper (L_n) and lower (L_{n-1}) length limits. On the basis of a distributed RC network model, Equation (5b) imposes a latency requirement on the longest interconnect (of length L_n) on a given pair of wiring levels. The required latency is expressed by β/f_{c}, where $\beta < 1$ and $1/f_{\mathrm{c}}$ is the clock period. In essence, the first and second equations are solved simultaneously for the minimum pitch p_n and maximum corresponding wire length L_n for each pair of wiring levels until L_n equals the maximum required wire length of the macrocell on its top pair of wiring levels. Equations (5a) and (5b) are solved simultaneously if repeaters are not used, while Equations (5a) and (5c) apply if optimal repeaters are used [8, 9]. The parameters R_0 and C_0 respectively represent the output resistance and input capacitance of a minimum-geometry MOSFET used as the basis for the repeater circuits [10].

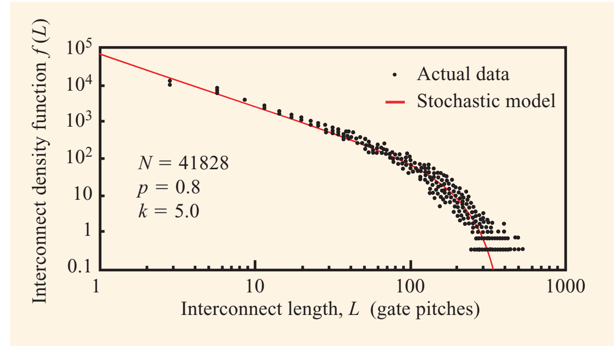


Figure 3

Interconnect density distribution. The vertical axis represents the density of interconnects of length L in units of number of interconnects per gate pitch, and the horizontal axis represents interconnect length in gate pitches. A gate pitch is the center-to-center spacing of the gates in the square array in Figure 2. Actual data is taken from a commercial microprocessor [7]. Reprinted with permission from [7]; © 1998 IEEE.

An example of minimization of macrocell area is illustrated in **Figure 4(a)**. A random logic network consisting of 12.4 million gates implemented with 100-nm-generation technology using eight levels ($n = 8$) of copper interconnects and operating at a clock frequency $f_{\mathrm{c}} = 578$ MHz is considered. Two alternative wiring network architectures are compared. The first architecture (shown on the left) is restricted to two and only two different cross-sectional dimensions (or two tiers) for eight levels of wiring. It requires two levels of 100-nm wiring and six levels of 540-nm wiring as well as a macrocell area $A_{\mathrm{m}} = 2.34\ \mathrm{cm}^2$ to interconnect the macrocell. The second architecture (shown on the right) is optimized to use three tiers of wiring in order to minimize cell area. It therefore requires four levels of 100-nm wiring, two levels of 150-nm wiring, and two levels of 300-nm wiring, as well as a macrocell area $A_{\mathrm{m}} = 0.70\ \mathrm{cm}^2$. The decisive macrocell-area advantage of the three-tier architecture is achieved using the methodology defined in Equations (5a), (5b), and (5c), whose central feature is demand prediction based upon a complete stochastic wiring distribution $f(L)$ [8, 9].

A second and currently more realistic example of an optimal multilevel network architecture is illustrated in **Figure 4(b)**. In this case the macrocell consists of an 11.3-million-gate random logic network implemented with 100-nm technology using eight levels of copper wiring ($n = 8$) and operating at a clock frequency of 1.56 GHz. If the pitch is chosen *a priori* to double for every pair of levels, the resulting architecture consists of two levels each of 100-, 200-, 400-, and 800-nm wiring, which require a 1.45-cm^2 area. In contrast, using the methodology prescribed by

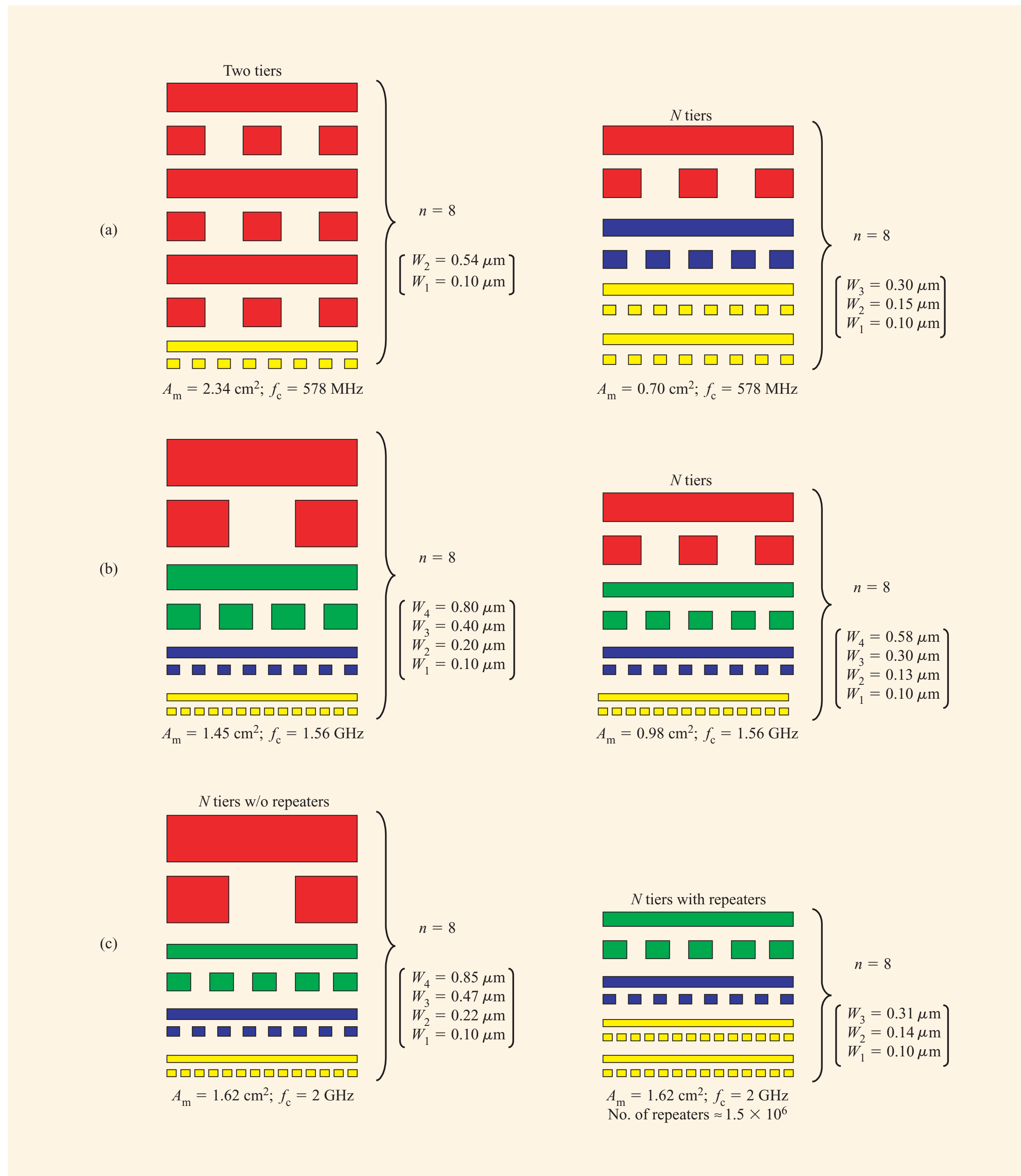


Figure 4

Optimization of macrocell area. (a) Minimum feature size $F = 0.1\,\mu$m; number of logic gates $N = 12.4$ million gates. Comparison of wiring-limited macrocell areas A_m required for a two-tier vs. an optimal three-tier multilevel interconnect network architecture for clock frequency $f_c = 578$ MHz. (b) $F = 0.1\,\mu$m; $N = 12.4$ million gates. Comparison of non-optimized and optimized four-tier architectures for $f_c = 1.56$ GHz. (c) $F = 0.1\,\mu$m; $N = 11.3$ million gates. Comparison of optimized four-tier architectures without and with repeaters for $f_c = 2$ GHz.

Equations (5a) and (5b), the optimal wire-level-pair dimensions are 100, 130, 300, and 580 nm, yielding a macrocell area of 0.98 cm^2 or a reduction of approximately 32%. If 1.5×10^6 optimal repeaters are used, the macrocell clock frequency can be increased to $f_c = 2.0$ GHz and the area reduced to 0.48 cm^2, as illustrated in **Figure 4(c)** [11].

As indicated by Equation (5a), determination of the area available for signal wiring on an orthogonal pair of levels requires estimation of the wiring efficiency factor e_w based on results of previous designs. As the number of wiring levels and the number of repeater circuits increase, via blockage tends to reduce wiring efficiency. The impact of via blockage can be estimated by calculation of a via blockage factor,

$$B_V = A_V / A_m , \tag{6a}$$

where A_V is the area blocked by vias on a given level of wiring and A_m is the macrocell or chip area. As illustrated in **Figure 5(a)**, *terminal* vias (which connect a particular interconnect net to a transistor) cause a "ripple effect" that reduces the number of wiring tracks available in a given area. In contrast, *turn* vias (which connect two wiring levels) do *not* cause via blockage. To elucidate with a simple example illustrated in Figure 5(a), $B_V = 0$ for the five uninterrupted wiring tracks on the left without terminal vias, and $B_V = 0.2$ for the four wiring tracks on the right, where 20% of the available wiring area is blocked by terminal vias [three of which are shown in Figure 5(a)] [12]. Assuming a uniform distribution of terminal vias as illustrated in **Figure 5(b)**, a general expression can be derived for B_V in terms of the geometry of the wiring layout [12]:

$$B_V = \sqrt{\frac{N_V (2W + s\lambda)^2}{A_m}} , \tag{6b}$$

where N_V is the total number of terminal vias for a particular metal level on a chip and W, s, and λ are defined in Figure 5(b). The number of terminal vias N_V for a given wiring level is determined by the total number of interconnects on wiring levels *above* the given wiring level using the methodology defined by Equations (4a), (4b), and (4c). From Equation (6), the via blockage factors for the eight wiring levels used in two macrocells (with $F = 100$ nm and $N = 12.4$ million gates) similar to those described in Figure 4 are illustrated in **Figure 6** [12]. Figure 6 reveals two striking features of via blockage due to signal interconnects as predicted by the new model. First, via blockage is more problematic for relatively small-area macrocells because of their greater interconnect density. More significantly, via blockage is severe on only the first level of wiring, where 15–30% of the total wiring area of a representative macrocell can be

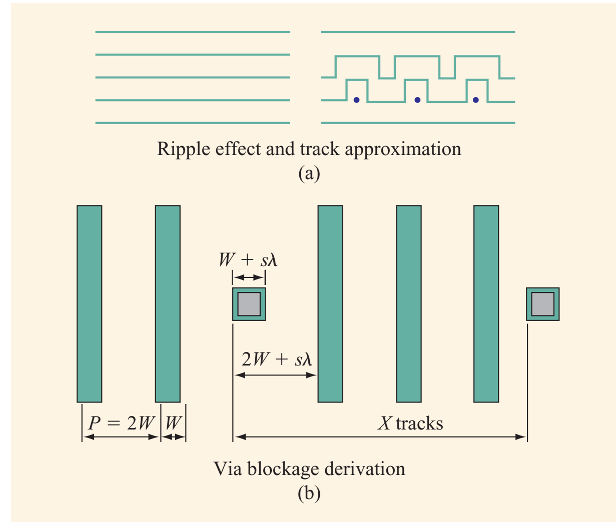


Figure 5

Sketch of wiring layouts used for derivation of via blockage factor $B_V = A_V/A_m$, where A_V is the area of a wiring level blocked by vias, A_m is the total macrocell or chip area, and N_V is the number of *terminal* vias piercing a wiring level as indicated by the ripple effect and track approximation. Reprinted with permission from [12]; © 2000 IEEE.

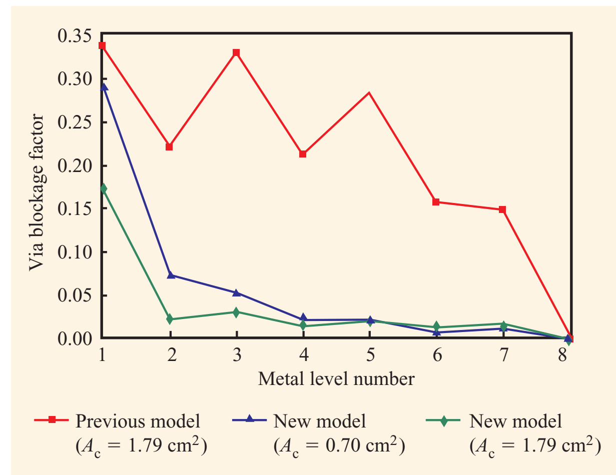


Figure 6

Via blockage factor vs. metal level number for two interconnect networks (with minimum feature size $F = 100$ nm and number of gates $N = 12.4$ million) similar to those described in Figure 4(a). Reprinted with permission from [12]; ©2000 IEEE. The previous model is described in [13].

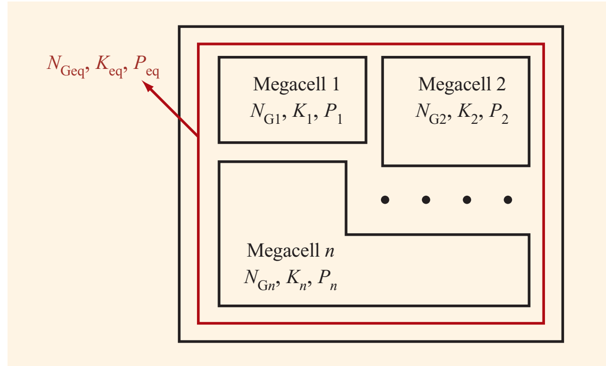


Figure 7

SoC layout used for definition of heterogeneous Rent's rule that applies to a heterogeneous set of megacells 1 through n comprising a system-on-a-chip. Reprinted with permission from [14]; ©2000 IEEE.

blocked. The via blockage estimate based on a previous model [13] is also illustrated in Figure 6.

4. System-on-a-chip (SoC)

The previous section deals with reverse scaling of signal wiring for a macrocell that may be modeled as a largely homogeneous block of microcells. A second commonly encountered situation is a system-on-a-chip consisting of a number of heterogeneous megacells such as control logic networks, cache memory arrays, arithmetic logic units, and register files. Each of these megacells can be characterized by a particular equivalent number of gates N_{Gi}, Rent's coefficient K_i, and Rent's exponent P_i [14]. The question to be addressed is the following: How can the global signal, power, and clock distribution networks for the heterogeneous SoC be designed to 1) fit all of the global wiring into the top two metal levels, 2) meet the required system clock frequency, and 3) limit the crosstalk noise to a specified maximum value? An initial response to this question follows.

The methodology begins by engaging a recently derived *heterogeneous version* of Rent's rule [14]. For the heterogeneous system-on-a-chip illustrated in **Figure 7**, this expanded version is defined by

$$T_{eq} = K_{eq} N^{P_{eq}}, \tag{7a}$$

where

$$K_{eq} = \left(\prod_{i=1}^{n} K_i^{N_{Gi}} \right)^{1/N_{Geq}}, \tag{7b}$$

$$P_{eq} = \frac{\sum_{i=1}^{n} P_i N_{Gi}}{N_{Geq}}, \tag{7c}$$

and

$$N_{Geq} = \sum_{i=1}^{n} N_{Gi}. \tag{7d}$$

In this power-law relationship (7a), Rent's coefficient K_{eq} is expressed as a weighted geometric average (7b) and Rent's exponent P_{eq} as a weighted arithmetic average (7c). Heterogeneous Rent's rule is used to derive three probability density distributions as summarized in **Figure 8** [14]. The first is a net fan-out (FO) distribution that defines the number of nets $N_{Net}(m)$ versus the number of net terminals $m = FO + 1$, where N_m is the total number of megacells in the SoC. The second is a net bounding area distribution that describes the number of nets versus the average net bounding area for nets with a specific number of terminals m. The average bounding area dimension of a square net $a(m)$ is shown in Figure 8, where η_p is an empirical placement efficiency factor that is estimated on the basis of previous designs [14]. The third distribution is an average net length distribution that describes the number of nets versus average net length for nets with a specific number of terminals m. An expression for the average value of net length $L_{av}(m)$ is given in Figure 8. These three distributions are combined to derive the total global signal wiring requirement L_{tot} as shown in Figure 8 [14].

Figure 9 summarizes this new methodology and compares model predictions with data from a commercial product. The graph in Figure 9 plots number of interconnect nets per mm or net density versus average interconnect net length in mm. The first dashed locus describes the density of nets with a fan-out of 1; the second describes nets with a fan-out of 2; the third a fan-out of 3, etc. The solid locus is the total interconnect net density distribution in number of nets per mm versus average interconnect length as calculated using the new model. The open circles represent data describing a commercial microprocessor consisting of 20 heterogeneous megacells [14].

In essence, the summation in Figure 8 defines the total length of global signal wiring required for a heterogeneous SoC. The next wiring resource requirement that must be defined concerns power distribution. **Figure 10** presents the results of modeling the required area for power distribution A_{power}, for the cases of peripheral bonding pads or an area array of bonding pads. For peripheral bonding pads, it is assumed that an equipotential ring

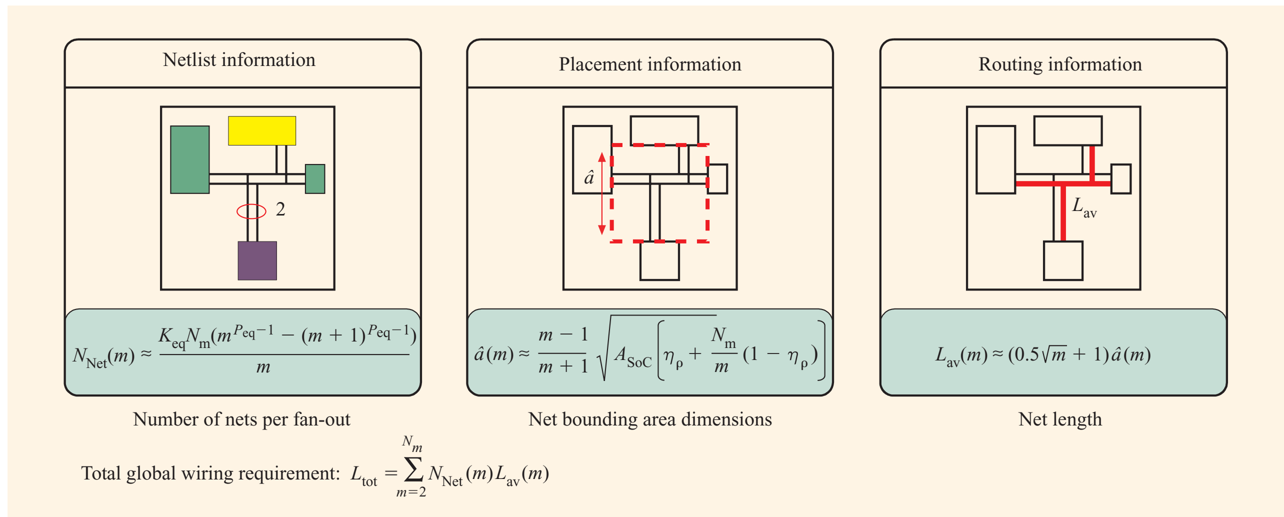


Figure 8

Signal wiring area models. Net length distribution [number of nets $N_{Net}(m)$ vs. number of net terminals (m)], average net bounding area $a(m)$, average net length L_{av}(m), and total global wire length requirement L_{tot}. Reprinted with permission from [14]; ©2000 IEEE.

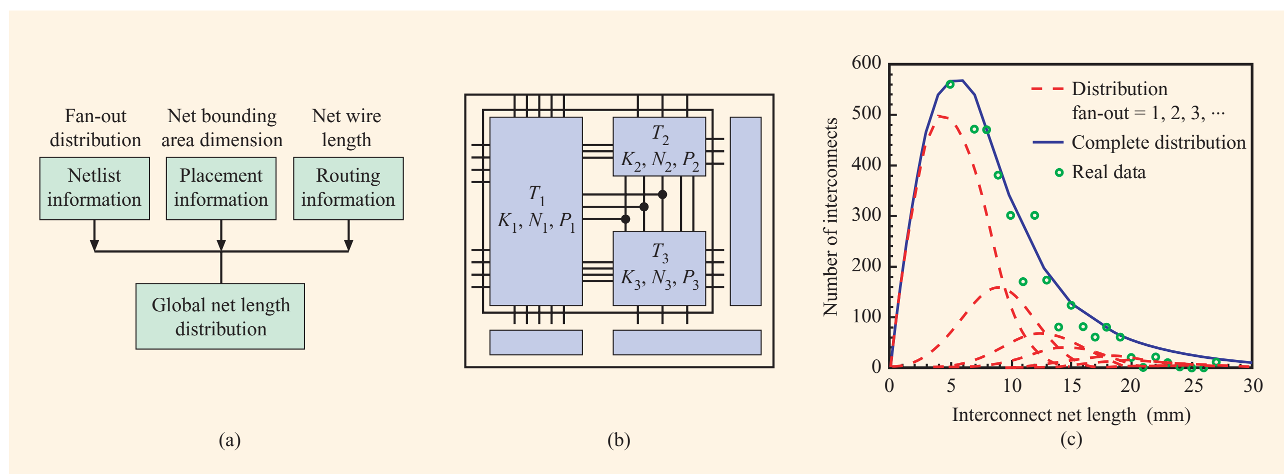


Figure 9

Summary of signal wiring (A_{signal}) derivation. (a) Summary diagram of derivation of interconnect net density distribution (number of interconnects per mm vs. average interconnect net length in mm for nets with a specific number of terminals m). (b) Block diagram of heterogeneous SoC. (c) Comparison of model predictions (solid curve) vs. actual data (open circles). Reprinted with permission from [14]; ©2000 IEEE.

surrounds the chip, as illustrated in Figure 10(a). For area array bonding pads, illustrated in Figure 10(b), it is assumed that V_{dd} is the potential of each bonding pad and that the current drain at each orthogonal intersection of the power grid lines is constant. A_{SoC} is the total SoC area [14]. In Figure 10, $\delta = \Delta V_{dd}/V_{dd}$ is the normalized voltage drop from a bonding pad to the most distant via at the intersection of an orthogonal pair of power grid lines, V_{dd} is supply voltage, H is metal height, P_{tot} is total chip power dissipation, and ρ_W is metal resistivity. Note that A_{Power} for area array bonding pads can be reduced effectively by increasing the number of bonding pads, n_{pad}.

The most critical clock distribution network requirement that must be met is imposed by the bandwidth necessary

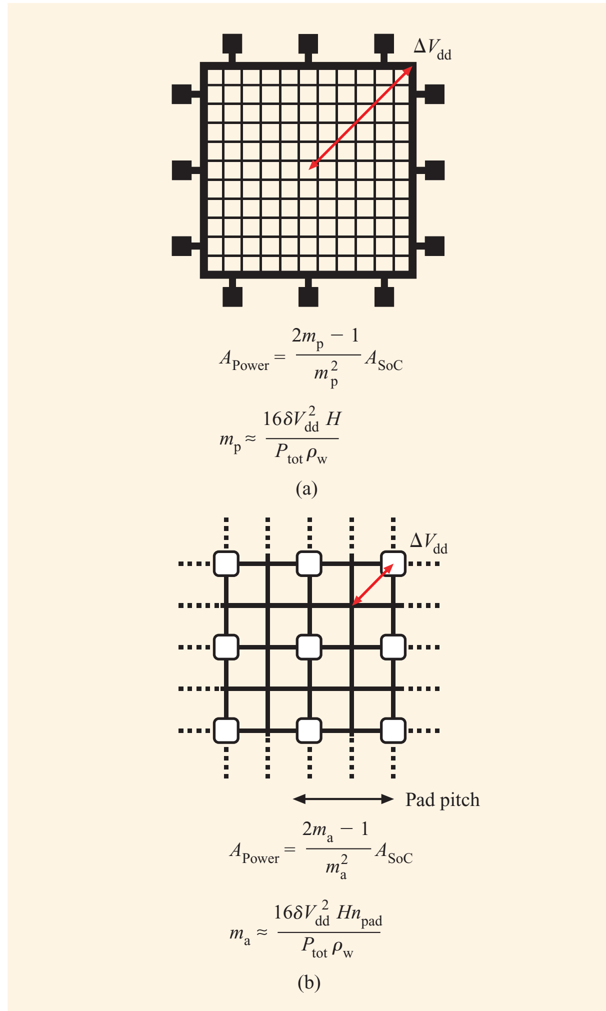


Figure 10

Summary of power wiring area (A_{power}) derivation for (a) peripheral bonding pads and (b) area array bonding pads as a function of chip or macrocell area A_{SoC}. $\delta = IR/V_{dd} = \Delta V_{dd}/V_{dd}$, where V_{dd} is supply voltage; H is metal height, P_{tot} is total chip power dissipation, ρ_w is wire resistivity, and n_{pad} is the number of power-supply pads.

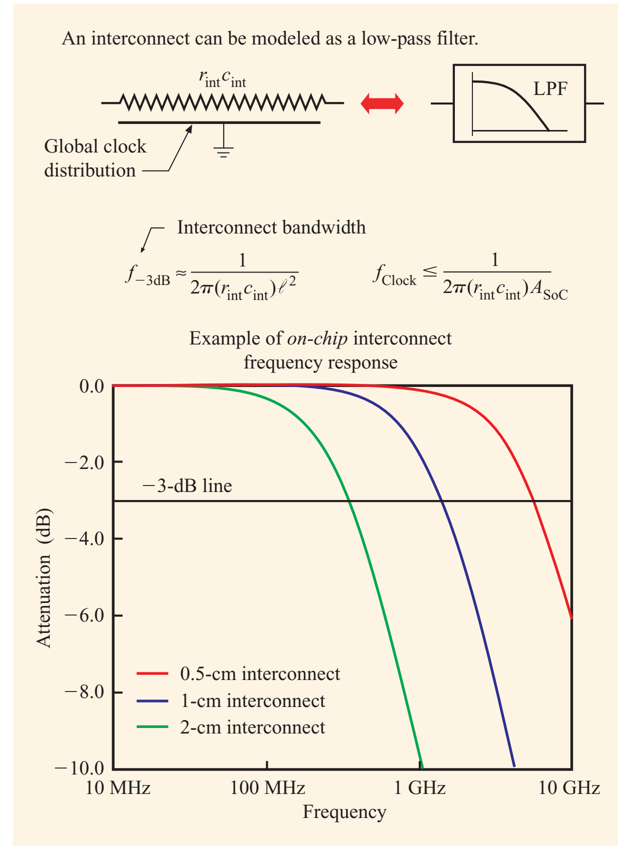


Figure 11

Summary of global H-array clock distribution network bandwidth requirements f_{-3dB} and limit on clock frequency f_{Clock} for a distributed *RC* network model of the H-array.

for rapid transitions of the clock waveform. It is assumed that global clock distribution is implemented with a balanced H-array. This array is modeled as a distributed *RC* network whose maximum length extends from the chip clock input pad to a terminal buffer/repeater of the global H-array. The approximate value of this maximum length is the dimension of the chip edge l. **Figure 11** defines the clock frequency limit f_{Clock} as a function of chip area $A_{SoC} = l^2$ [14].

The final performance requirement that is imposed on the global wiring network is a crosstalk noise limit. A model used for an approximate calculation of global crosstalk noise is illustrated in **Figure 12**. In this representation, a global signal line or "victim" is assumed to be surrounded by two near and two far "attackers." Simultaneous in-phase switching of the four attackers causes crosstalk noise on the victim due to coupling of both mutual capacitance and mutual inductance. A nearby high-quality return path is assumed to be available. Treating the five coupled lines as distributed *RLC* networks, a set of partial differential equations quantifies the problem [15, 16]. Some results of a solution to this set of equations are illustrated in **Figure 13**, which plots the ratio of crosstalk-to-binary signal voltage swing versus time [16]. Comparing the three- and five-line loci, it is evident that in the presence of a nearby high-quality return path, the near attackers shield the victim from the far attackers. Therefore, using the three-line model, simplified expressions for peak crosstalk voltage derived from the solutions of the set of partial differential equations are

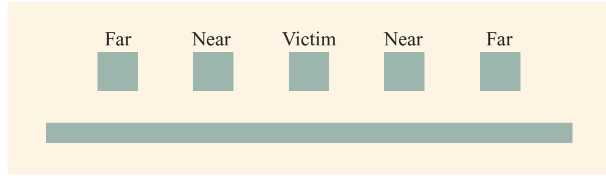


Figure 12

Five coupled distributed *RLC* lines above a ground plane. Model used for calculation of crosstalk noise induced in a global signal interconnect assuming a victim surrounded by near and far attackers in the presence of a nearby high-quality return path.

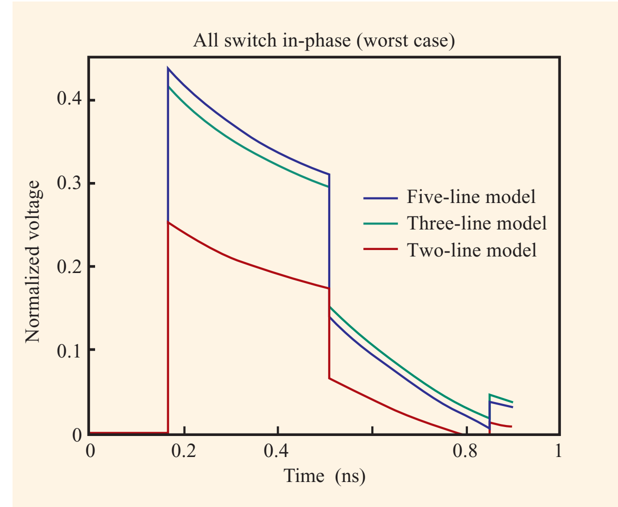


Figure 13

Effect of increasing number of aggressors. Normalized crosstalk voltage induced on a global victim interconnect by one, two, and four aggressor interconnects. Far lines have a negligible effect when a ground plane exists, which means that crosstalk is local. Reprinted with permission from [18]; © 2001 IEEE.

$$\frac{V_n}{V_{dd}} \cong \frac{1}{2} \frac{c_{mutual}}{c_{line} + c_{mutual}} \tag{8a}$$

for distributed *RC* models [17] and

$$\frac{V_n}{V_{dd}} \cong \frac{\pi}{2} \frac{1}{2} \frac{c_{mutual}}{c_{line} + c_{mutual}} \tag{8b}$$

for distributed *RLC* models [15, 16], where c_{mutual} is the line-to-line distributed capacitance and c_{line} is the line-to-return-path distributed capacitance.

A summary of the complete set of three compact models that define the primary global interconnect design requirements is given in **Table 3**. The models are expressed in terms of the physical parameters of the two global wiring levels illustrated in **Figure 14**. In the wiring resource requirement, the three terms respectively represent the signal and power wiring areas and unused area. The second and third models respectively describe

Table 3 Summary of complete set of requirements to be imposed on global signal, power, and clock distribution networks expressed in terms of the geometry of the two global wiring levels.

Equation description	*Simplified expressions for global wiring requirements in terms of w, s, H, and T_{ox}*
Wiring resource requirement	$(w+s)\sum_{m=2}^{N_m} N_{Net}(m)L_{av}(m) + \frac{2m_p - 1}{m_p^2} A_{SoC} + 0.5\left(1 - \frac{1}{m_p}\right)^2 A_{SoC} \le A_{SoC}$
	where $N_{Net}(m) \approx \frac{K_{eq}N_m[m^{P_{eq}-1} - (m+1)^{P_{eq}-1}]}{m+1}$
	$L_{av}(m) \approx (0.5\sqrt{m} + 1)\frac{m-1}{m+1}\sqrt{A_{SoC}\left[\eta_p + \frac{N_m}{m}(1-\eta_p)\right]}, \quad m_p = \frac{16\delta V_{dd}^2 H}{P_{tot}\rho_w}$
Wiring bandwidth requirement	$f_c \le \frac{1}{4\pi\rho_w\varepsilon_0\varepsilon_r(1/HT_{ox} + 1/ws)A_{SoC}}$
Wiring noise limit	$\frac{\pi}{4}\frac{1/ws}{(1/HT_{ox} + 1/ws)} \le$ % noise

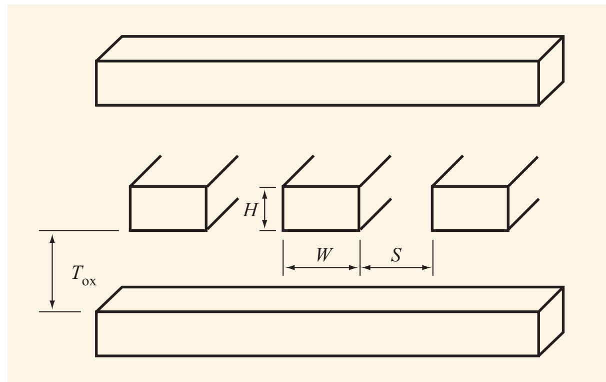


Figure 14

Physical representation of orthogonal interconnect system.

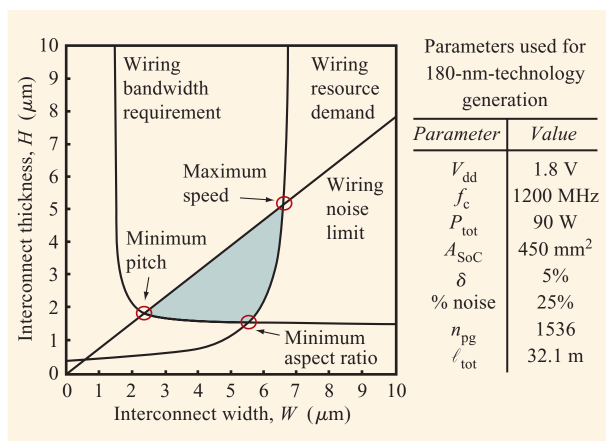


Figure 15

Global interconnect design plane plotting interconnect thickness H vs. width W for the interconnect requirements summarized in Table 3 as applied to a SoC consisting of 20 heterogeneous megacells including a total of approximately six million transistors. Reprinted with permission from [19]; © 2000 IEEE.

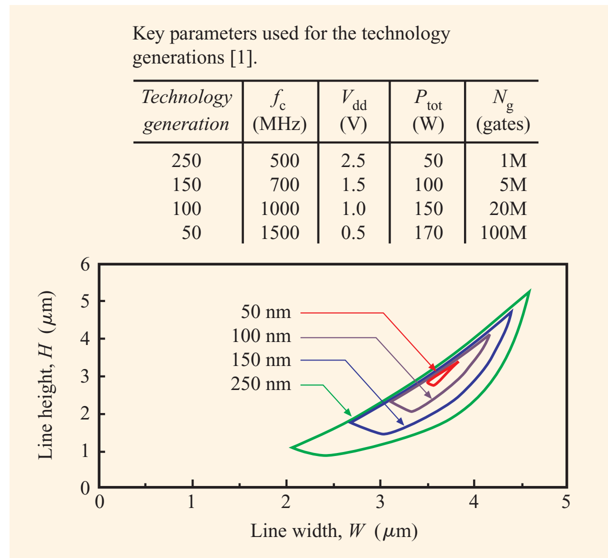
the clock wiring bandwidth and signal wiring crosstalk noise limit. In **Figure 15**, the three models are applied to a particular SoC consisting of 20 heterogeneous megacells containing a total of approximately six million transistors [14]. In the global interconnect design plane, the vertical axis represents interconnect thickness H and the horizontal axis, interconnect width W. The *allowable design region* that satisfies all primary global wiring requirements is the zone bounded by the resource, bandwidth, and noise limit loci. For example, an interconnect width $W \cong 2.4$ μm and height $H \cong 2.0$ μm satisfies the prime design constraints with a minimum pitch. Projections of the allowable design regions for several future generations of technology are illustrated in **Figure 16**. Here it is evident that the amount of compression of the allowable design region becomes unacceptable, and additional flexibility such as expansion of the number of global wiring levels appears to become necessary.

In summary, the methodology presented in this section enables early projections of key physical parameters of a global interconnect network that *simultaneously* satisfies the primary requirements of a SoC for signal, power, and clock distribution. The compact physical models that serve to implement the methodology offer a convenient opportunity to establish a quantitative guide to detailed design of a SoC. Therefore, the methodology may serve as a useful precursor to final design. Enhancements of this methodology that include, for example, the effects of clock skew, nonideal return paths, and simultaneous switching noise are needed.

5. Three-dimensional integration

Achieving three-dimensional (3D) integration in semiconductor technology requires the capability to stack multiple strata, each containing both transistors and multilevel interconnect networks, as discussed in preceding sections. This is a formidable challenge that is unlikely to be engaged seriously absent a convincing case for substantive benefits. Therefore, what are the primary benefits that can be projected for 3D integration? It appears that the singular generic advantage of 3D integration is a substantial reduction in length of the longest global interconnects used in a SoC.

Several rigorous derivations of stochastic interconnect distributions for 3D random logic networks [18–20] based upon the 2D distribution discussed in Section 3 [7, 8] have been reported. Using the analytic models derived in [20], the stochastic interconnect distributions for a 4.0-million-gate random logic network implemented with 1, 4, and 16 strata are illustrated in **Figure 17(a)**. Note that for simplicity these distributions assume that the interstratal pitch $r = 1$, which strictly imposes the condition that the interstratal pitch equals the intrastratal logic gate pitch. The loci of Figure 17(a) clearly indicate that multiple strata or 3D integration exerts very little impact on the density of local interconnects, but it has a profound effect on the length of the longest interconnects of the logic network. This observation is illustrated with greater clarity in **Figure 17(b)**. The right vertical axis indicates a length of approximately 4000 gate pitches for a corner-to-corner interconnect in a single-stratum implementation, 2000 gate pitches for a four-stratum implementation, and 1000 gate pitches for a 16-stratum implementation. For time-of-flight-limited global interconnects, this could result in a 4:1 reduction of latency and the possibility of an


Key parameters used for the technology generations [1].

Technology generation	f_c (MHz)	V_{dd} (V)	P_{tot} (W)	N_g (gates)
250	500	2.5	50	1M
150	700	1.5	100	5M
100	1000	1.0	150	20M
50	1500	0.5	170	100M

Figure 16

Global interconnect design plane projections illustrating compression of allowable design region and consequently the necessity for greater flexibility such as expansion of the number of wiring levels.

approximately fourfold increase in global clock frequency—for the expense of a 16-stratum implementation of the system.

A key simplifying assumption limiting the projections illustrated in Figures 17(a) and 17(b) is that the interstratal pitch equals the intrastratal gate pitch, or $r = 1$. Setting aside this assumption, a generic 3D wiring distribution for a 4.0-million-gate random logic network whose interstratal pitch is treated as an independent variable has been rigorously derived [20]. **Figure 17(c)** illustrates a key result of this new derivation for interstratal pitches $r = 1$ and $r = 50$. The two distributions are quite similar for short local and long global interconnect lengths. The only region in which the two loci deviate is the midrange of interconnect lengths, where interconnect length and stratal pitch are roughly equal. Consequently, it appears that interstratal separation distance is not a critical parameter in determining 3D wiring distributions.

The generic benefit of substantial reductions in length of the longest global interconnects in a distribution resulting from 3D integration is an inherent advantage of 3D wiring. A concomitant inherent disadvantage of 3D structures is heat removal [21]. Beyond these general issues, the attraction of 3D integration for specific applications may be dominated by the peculiar features of the application itself. For example, two-dimensional sensor

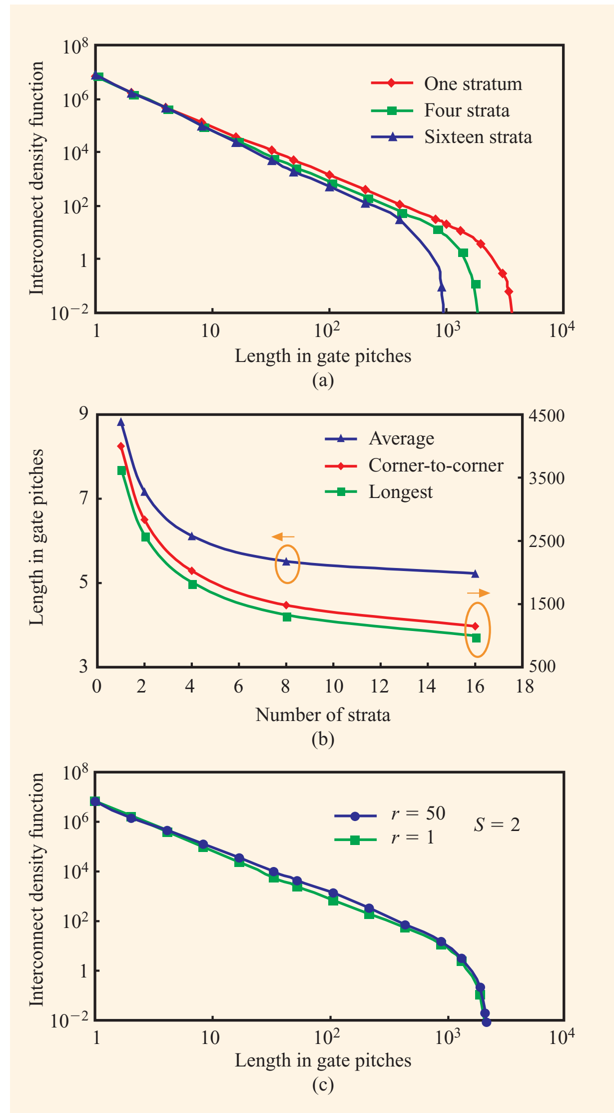


Figure 17

(a) Interconnect density per gate pitch vs. length in gate pitches for a random logic network of approximately four million gates implemented on one stratum, four strata, and sixteen strata. (b) Interconnect length vs. number of strata for average, corner-to-corner, and longest interconnects of the random logic network of part (a). (c) Interconnect density per gate pitch vs. length in gate pitches for the random logic network of part (a) implemented on two strata for interstratal pitches $r = 1$ and $r = 50$. Parts (a)-(c) reprinted with permission from [22]; © 2000 IEEE.

arrays that require direct access to each sensor cell for immediate signal preprocessing are interesting prospects for 3D integration [22]. More broadly, the capacity to explore opportunities for extraordinary performance enhancements through 3D integration would benefit from generic advances in capabilities to fabricate 3D structures.

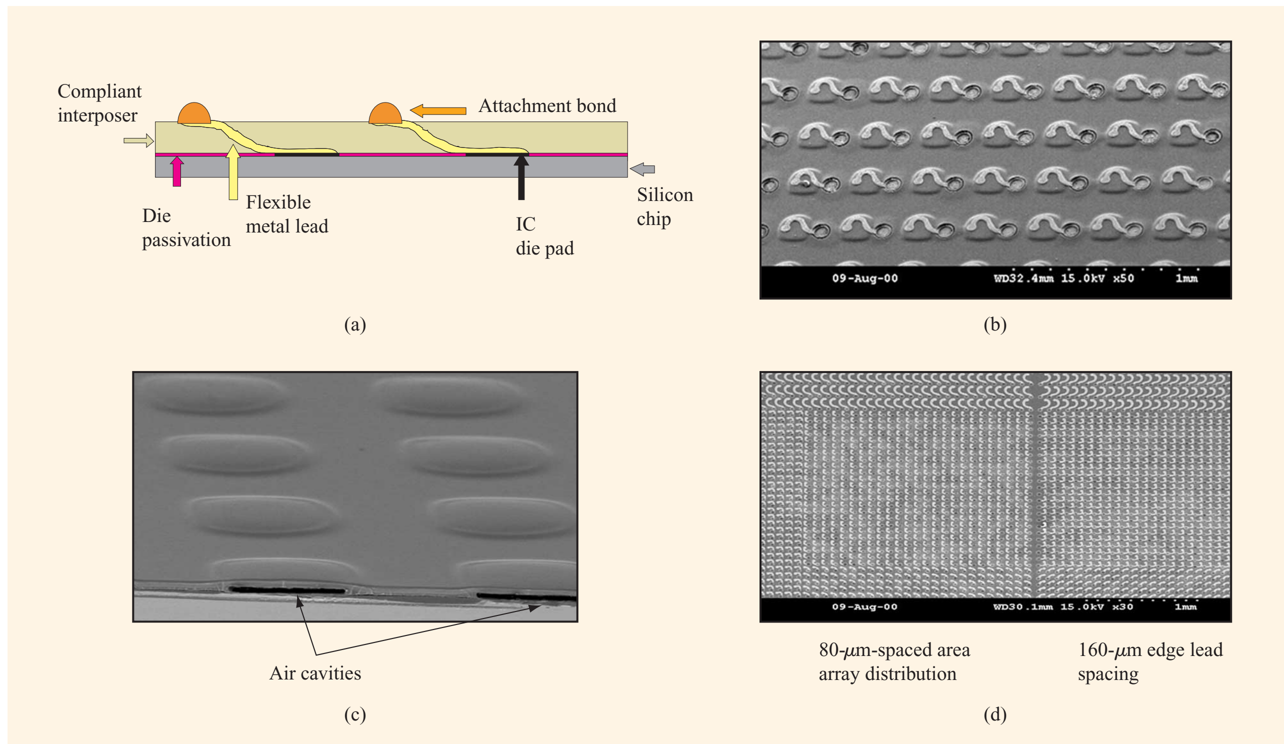


Figure 18

(a) Schematic representation of salient objective of Sea of Leads input/output interconnect technology. All silicon dice remain unseparated in a wafer until all compliant interposer packages, flexible metal input/output leads, and bonding material on the lead tips are batch-fabricated. Then all dice undergo full electrical testing and burn-in prior to dicing the wafer to yield known good packaged dice ready for shipment. (b) Photomicrograph of a Sea of Leads. (c) Cross-section photomicrograph of SoL air cavities. (d) Scanning electron micrograph of SoL with a density of 12 000 leads/cm^2. Parts (b) and (c) reprinted with permission from [31]; © 2001 IEEE.

6. Input/output interconnect enhancements

The intent of input/output interconnect enhancements is to improve the cost, size, reliability, and performance of a gigascale SoC. Historically, bonding wires have been the dominant approach to chip input/output (I/O) interconnects [23]. IBM pioneered the introduction of solder-bump I/O interconnects using flip-chip technology with a thin layer of glass passivation sealing the chip encapsulated in silicone gel, which prevented the formation of continuous water films [24, 25]. A particular novel technology that is currently under investigation for I/O enhancements is described as Sea of Leads (SoL) [26]. This technology proposes the use of wafer-level batch fabrication of compliant polymer packages, ultrahigh-density ($>10^4$/cm^2) x–y–z flexible metal leads, and solder-like bumps attached to the lead tips, as illustrated in **Figure 18(a)**. A short sequence of full-wafer SoL batch-fabrication processes constituting a "tail-end-of-the-line" (TEOL) are envisaged to follow conventional back-end-of-the-line (BEOL) wafer processing. The further intent of SoL technology is to complete all final electrical testing and burn-in operations prior to wafer dicing that yields *known good packaged die* ready for immediate shipment to customers. The flexible leads are designed to provide sufficient x–y–z compliance to accommodate typical differences in the thermal coefficients of expansion between a silicon chip and the substrate to which it is attached. The need for epoxy underfill is thereby precluded, and the possibility of convenient detachment of a chip from a substrate module is enabled.

Concurrent fabrication of packages and leads of all chips on a wafer extends the historically potent economies of wafer-level batch processing to the relatively costly die-by-die assembly, bonding, packaging, testing, and burn-in operations [27, 28]. Moreover, the size of the SoL package is the minimum for a chip-scale package (CSP). Significant reliability improvement may result from avoidance of epoxy underfill often needed to relieve stress on relatively rigid solder-ball connections between chip and substrate. **Figure 18(b)** is a photomicrograph of an SoL. The circular pattern is the via linking a die-bonding pad with the lead itself, which is the "question-mark-shaped" copper pattern.

This peculiar shape is designed to provide a high degree of *x–y* axis flexibility and thus accommodate chip–substrate thermal expansion differences. The somewhat rounded region beneath the copper lead defines the boundaries of a polymer interposer air cavity that is introduced to enhance *z*-axis compliance. This compliance is added in order to provide convenient and reliable temporary electrical contacts between an array of electrical test probes and the leads of the dice under test, especially when the probe tips are not in a precisely planar arrangement. A photomicrograph of the cross section of an air cavity is shown in **Figure 18(c)**. An SEM of a 1 × 1-cm die with an SoL density of 12 000 per cm^2 is shown in **Figure 18(d)**. The leads are oriented along the contours of expansion of the die to provide a higher degree of compliance proceeding radially outward from the center to the edge of the die.

Key performance enhancements that appear to be in the offing for SoL technology include the following [16, 26]:

1. Substantially increased input/output bandwidth for a chip resulting from the significantly larger (e.g., ~10×) number of signal leads that are available.
2. "Time-of-flight" global signal interconnect latency for a chip resulting from *exiting and then reentering* the die using external on-module wiring, or "exterconnects," to implement very-low-loss time-of-flight internal global wiring links for the chip.
3. Reduced global clock skew due to use of time-of-flight exterconnects to implement global clock trees.
4. Reduced global clock power dissipation through recycling the energy of reflected clock pulses distributed through low-loss exterconnects [29].
5. Suppression of far-attacker crosstalk noise on global signal interconnects through the use of exterconnects with nearby high-quality return paths provided by module power and ground planes.
6. Suppression of simultaneous switching noise (SSN) and reduced parasitic *IR* voltage drop in the power/ground distribution networks resulting from the significantly larger (e.g., ~10×) number of power and ground leads that are available.
7. Improved isolation and reduced interference in mixed-signal systems resulting from use of separate power/ground input/output leads for analog and digital signals.

Additional opportunities that are available through SoL include the capacity to satisfy the voracious appetite of 3D integration for I/O capacity and the potential for compatibility of electrical, rf wireless, and photonic I/O interconnects.

In short, SoL can be described as a "disruptive" technology, because the intent is to use batch-fabricated ultrahigh-density input/output leads to improve the cost, size, reliability, and performance of an SoC [16, 26].

7. Photonic interconnects

An exposition of interconnect opportunities for GSI would not be complete without consideration of photonic or optical interconnects [30–33]. In order to be competitive with electrical interconnects for GSI, photonics must provide small, low-power, high-speed, low-cost photon emitters, detectors, and conductors or waveguides that are compatible with CMOS technology. Consequently, this section focuses on *compatible photonics*, or photonic technologies with the long-range potential to satisfy the extremely stringent and particular demands of GSI.

The most challenging objective for CMOS-compatible photonic interconnects is an efficient room-temperature silicon light emitter. A novel silicon diode which exploits dislocation loops to introduce a local strain field that modifies the band structure to confine carriers near the junction and therefore enhance light emission was recently demonstrated [34].

Short of high-quality silicon photoemitters, a most interesting approach to compatible photonics is based upon heteroepitaxial deposition on Si of SiGe, followed by Ge, followed by GaAs, and finally by AlGaAs [31, 32]. The close lattice-constant match of Ge and GaAs provides a basis for growing high-quality single-crystal layers of GaAs. This heteroepitaxial approach to compatible photonics has the potential to provide III–V compound semiconductor lasers, Ge detectors, and polycrystalline or monocrystalline Si waveguides. **Figure 19** illustrates the current–voltage curves of heteroepitaxial SiGe and GaAs diodes on a Si substrate [32]. **Figure 20** displays photomicrographs of a right-angle bend and a junction in a polycrystalline Si waveguide [33]. Transmission loss is less than 0.5 dB in the bend and 1.0 dB in the junction. The waveguide width is 0.5 μm, which is comparable to dimensions of upper-level metal interconnects. These recent advances are encouraging demonstrations of the long-range promise of compatible microphotonic interconnects.

It has long been proposed that the most likely point of entry of photonic interconnects into silicon integrated electronics is in clock distribution [35, 36]. Recently, a polymer waveguide network with volume grating output couplers embedded in a printed wiring board (PWB) was proposed to transfer photons from a printed wiring board to one or more silicon photodetectors fabricated in a CMOS chip [37]. This approach to optical clock distribution does not utilize on-chip photon emitters and enables a planar package configuration.

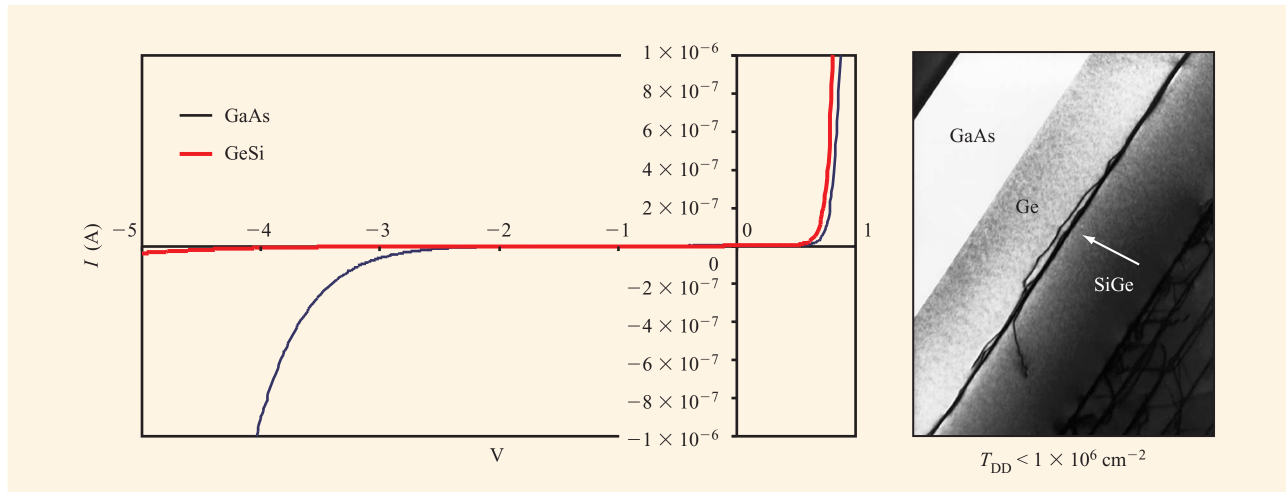


Figure 19

Monolithic Si/SiGe/Ge/GaAs; I–V curves for heteroepitaxial 180-μm × 180-μm SiGe and GaAs diodes on a Si substrate. (Courtesy of Prof. Gene Fitzgerald, Massachusetts Institute of Technology; reprinted with permission.)

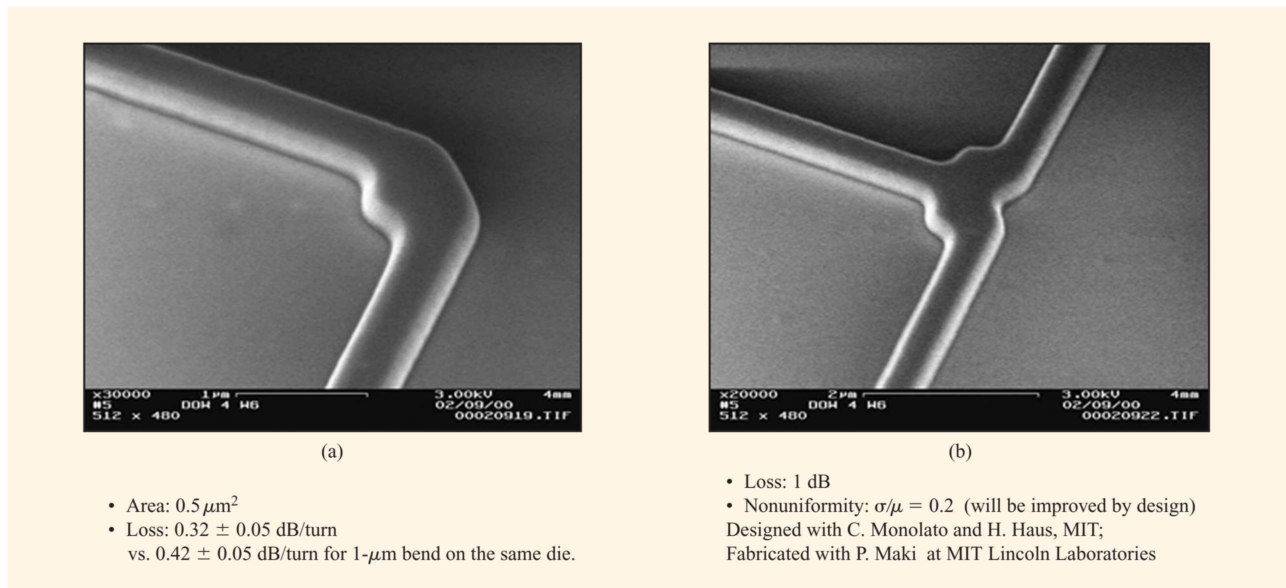


Figure 20

High-transmission-cavity (HTC) waveguide bends and splits. Photomicrograph of polycrystalline Si high-transmission-cavity waveguide bends (a) and junctions (b). (Courtesy of Prof. Lionel Kimerling, Massachusetts Institute of Technology; reprinted with permission.)

8. Conclusions

Interconnect latency is now the primary performance issue for GSI, and the problem promises to become more serious for future generations of technology. Opportunities to address the problem range, for example, from carbon nanotube conductors that may enable ultrahigh-speed ballistic transport [38] to new single-chip, distributed shared memory, cellular arrays of microprocessors [39, 40] that serve to keep interconnects short. The second interconnect problem that is not broadly recognized as such is energy dissipation. The keys to solving this problem are short interconnects, and transistors with

the smallest possible subthreshold swing and therefore the smallest possible binary signal swing. Crosstalk and simultaneous switching noise represent a third interconnect problem—signal integrity—which is difficult to describe using compact physical models.

For virtually any family of gigascale chips, the key to optimal reverse scaling of multilevel signal interconnect networks is prediction of the complete stochastic wiring distribution of a next-generation product. More general signal integrity models that can be incorporated into reverse scaling methodologies are needed.

The task of conjointly optimizing the architecture of the global signal, clock, and power/ground distribution networks of a system-on-a-chip consisting of a set of heterogeneous megacells is demanding. A first attempt to address this task comprehensively engages a new stochastic model for global signal wiring, a new model for global power/ground wiring area, a global clock bandwidth requirement, and a crosstalk noise requirement. Enhancements of current methodologies that include, for example, the effects of clock skew, nonideal return paths, and simultaneous switching noise are needed.

The generic benefit of substantial reductions (e.g., >50%) in the length of the longest global interconnects in a distribution is an inherent advantage of 3D integration. The capacity to explore novel opportunities for extraordinary performance enhancements through 3D integration would benefit from generic advances in capabilities to fabricate 3D structures.

In order to maintain historic rates of advance of monolithic semiconductor technology, more attention to ancillary features and particularly to input/output interconnects is unavoidable. Sea of Leads represents an early effort to more intimately couple the chip itself to its environment and then to exploit concomitant new opportunities. Key projected performance enhancements include substantially increased input/output bandwidth, reduced global signal interconnect latency, reduced global clock skew, reduced global clock power dissipation, greater suppression of simultaneous switching noise, and improved signal integrity in mixed-signal systems. More broadly, Sea of Leads represents an effort to extend the quintessential feature of semiconductor technology—wafer-level batch fabrication of several hundred chips—to the traditional die-by-die packaging and testing domains.

To be become widely used in GSI, photonics must provide small, low-power, high-speed, low-cost photon emitters, detectors, and conductors or waveguides that are compatible with CMOS technology [41]. Recent advances in heteroepitaxial deposition on Si of SiGe, followed by Ge, followed by GaAs to demonstrate light-emitting and -detecting diodes as well as Si waveguides, are promising.

Acknowledgments

The intellectual contributions to this paper by Azad Naeemi, Raguraman Venkatesan, Muhannad Bakir, Hiren Thacker, Qiang Chen, James Joyner, and Tony Mulé of the Georgia Institute of Technology Microelectronics Research Center are gratefully acknowledged. In addition, the authors wish to express their appreciation to DARPA, Contract No. F33615-97-C-1132, MARCO, Contract No. MDA 972-99-1-002, and the SRC, Contract No. 448:048, for their generous support.

References

1. J. D. Meindl, "Low Power Microelectronics: Retrospect and Prospect," *Proc. IEEE* **83,** 619–635 (April 1995).
2. M. Bohr, S. S. Ahmed, S. U. Ahmed, M. Bost, T. Ghani, J. Greason, R. Hainsey, C. Jan, P. Packan, S. Sivakumar, S. Thompson, J. Tsai, and S. Yang, "A High Performance 0.25 Micron Logic Technology Optimized for 1.8V Operation," *IEDM Tech. Digest*, pp. 847–850 (December 1996).
3. *International Technology Roadmap for Semiconductors* (ITRS), 1999 Edition, Semiconductor Industry Association, 4300 Stevens Suite Boulevard, Suite 271, San Jose, CA 95129.
4. T. N. Theis, "The Future of Interconnection Technology," *IBM J. Res. & Dev.* **44,** 379–390 (May 2000).
5. J. Hennessy, M. Heinrich, and A. Gupta, "Cache-Coherent Distributed Shared Memory: Perspectives in Its Development and Future Challenges," *Proc. IEEE* **87,** 418–429 (March 1999).
6. B. Landman and R. Russo, "On a Pin Versus Block Relationship for Partition of Logic Paths," *IEEE Trans. Computing* **C-20,** 1469–1479 (December 1971).
7. J. A. Davis, V. K. De, and J. D. Meindl, "A Stochastic Wire-Length Distribution for Gigascale Integration (GSI)—Part I: Derivation and Validation," *IEEE Trans. Electron Devices* **45,** 580–589 (March 1998).
8. J. A. Davis, V. K. De, and J. D. Meindl, "A Stochastic Wire-Length Distribution for Gigascale Integration (GSI)—Part II: Applications to Clock Frequency, Power Dissipation, and Chip Size Estimation," *IEEE Trans. Electron Devices* **45,** 590–597 (March 1998).
9. R. Venkatesan, J. A. Davis, K. A. Bowman, and J. D. Meindl, "Minimum Power and Area *N*-Tier Multilevel Interconnect Architectures Using Optimal Repeater Insertion," *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED)*, July 2000, pp. 167–172.
10. H. B. Bakoglu and J. D. Meindl, "Optimal Interconnect Circuits for VLSI," *IEEE Trans. Electron Devices* **ED-32,** 903–909 (May 1985).
11. H. B. Bakoglu and J. D. Meindl, "Optimal Interconnect Circuits for VLSI," *ISSCC Digest of Technical Papers*, February 1984, pp. 164–165.
12. Q. Chen, J. A. Davis, P. Zarkesh-Ha, and J. D. Meindl, "A Compact Physical Via Blockage Model," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.* **8,** 689–692 (December 2000).
13. G. A. Sai-Halasz, "Performance Trends in High-End Processors," *Proc. IEEE* **83,** 20–36 (January 1995).
14. P. Zarkesh-Ha, J. A. Davis, and J. D. Meindl, "Prediction of Net-Length Distribution for Global Interconnects in a Heterogeneous System-on-a-Chip," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.* **8,** 649–659 (December 2000).
15. (a) J. Davis and J. D. Meindl, "Compact Distributed RLC Interconnect Models Part I: Single Line Transient, Time

Delay, and Overshoot Expressions," *IEEE Trans. Electron Devices* **45,** 580–589 (November 2000). (b) J. Davis and J. D. Meindl, "Compact Distributed RLC Interconnect Models Part II: Coupled Line Transient Expressions and Peak Crosstalk in Multilevel Networks," *IEEE Trans. Electron Devices* **45,** 590–597 (November 2000).
16. A. Naeemi, C. Patel, M. Bakir, P. Zarkesh-Ha, K. Martin, and J. Meindl, "Sea of Leads: A Disruptive Paradigm for a System-on-a-Chip (SoC)," *ISSCC Digest of Technical Papers*, February 2001, pp. 280–281.
17. T. Sakurai, "Closed-Form Expressions for Interconnect Delay, Coupling, and Crosstalk in VLSIs," *IEEE Trans. Electron Devices* **40,** 118–124 (January 1993).
18. A. Naeemi, J. A. Davis, and J. D. Meindl, "Analytical Models for Coupled Distributed RLC Lines with Ideal and Non-Ideal Return Paths," *IEDM Tech. Digest*, pp. 689–692 (December 2001).
19. P. Zarkesh-Ha and J. D. Meindl, "An Integrated Architecture for Global Interconnects in a Gigascale System-on-a-Chip (GsoC)," *IEEE Symposium on VLSI Technology*, *Digest of Technical Papers*, June 2000, pp. 194–195.
20. A. Rahman, A. Fan, J. Chung, and R. Reif, "Wire-Length Distribution of Three-Dimensional Integrated Circuits," *Proceedings of the IEEE International Interconnect Technology Conference*, June 1999, pp. 233–235.
21. S. J. Souri and K. C. Saraswat, "Interconnect Performance Modeling for 3D Integrated Circuits with Multiple Silicon Layers," *Proceedings of the IEEE International Interconnect Technology Conference*, June 1999, pp. 24–26.
22. J. Joyner, P. Zarkesh-Ha, J. Davis, and J. Meindl, "A Three-Dimensional Stochastic Wire Length Distribution for Variable Separation of Strata," *Proceedings of the IEEE International Interconnect Technology Conference*, June 2000, pp. 132–134.
23. H. B. Bakoglu, *Circuits, Interconnects, and Packaging for VLSI*, Addison-Wesley Publishing Co., Inc., Reading, MA, 1990, Ch. 3.
24. J. Burns, L. McIlrath, C. Keast, C. Lewis, A. Loomis, K. Warner, and P. Wyatt, "Three-Dimensional Integrated Circuits for Low-Power, High Bandwidth Systems on a Chip," *ISSCC Digest of Technical Papers*, February 2001, pp. 268–269.
25. B. L. Gehman, "Bonding Wire Microelectronic Interconnections," *IEEE Trans. Components, Hybrids, Manuf. Technol.* **CHMT-3,** 375 (September 1980).
26. E. M. Davis, W. E. Harding, R. S. Schwartz, and J. J. Corning, "Solid Logic Technology: Versatile, High-Performance Microelectronics," *IBM J. Res. & Dev.* **8,** 102–114 (April 1964).
27. P. A. Totta and R. P. Sopher, "SLT Device Metallurgy and Its Monolithic Extension," *IBM J. Res. & Dev.* **13,** 226–238 (May 1969).
28. A. Naeemi, P. Zarkesh-Ha, C. Patel, and J. D. Meindl, "Performance Improvements Using On-Board Wires for On-Chip Interconnects," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, 2000, pp. 325–328.
29. C. S. Patel, C. Power, M. Realff, P. A. Kohl, K. P. Martin, and J. D. Meindl, "Low Cost High Density Compliant Wafer Level Package," *Proceedings of the International Conference on High-Density Interconnect and Systems Packaging*, Denver, April 26–28, 2000, pp. 262–268.
30. C. S. Patel, M. Realff, S. Merriweather, C. Power, K. Martin, and J. D. Meindl, "Cost Analysis of Compliant Wafer Level Packages," *Proceedings of the 50th Electronic Components and Technology Conference* (ECTC), Las Vegas, May 22–24, 2000, pp. 1634–1639.
31. H. Reed, M. Bakir, C. Patel, K. Martin, J. Meindl, and P. Kohl, "Compliant Wafer Level Package (CWLP) with Air-Gaps for Sea of Leads (SoL) Interconnections," *Proceedings of the International Interconnect Technology Conference*, San Francisco, June 4–6, 2001, pp. 151–153.
32. P. Zarkesh-Ha and J. D. Meindl, "Stochastic Net Length Distribution for Global Interconnects in a Heterogeneous System-on-a-Chip," *Symposium on VLSI Technology*, *Digest of Technical Papers*, June 1998, pp. 44–45.
33. A. V. Krishnamoorthy and D. A. B. Miller, "Scaling Optoelectronic-VLSI Circuits into the 21st Century: A Technology Roadmap," *IEEE J. Quantum Electron.* **2,** 55–76 (April 1996).
34. L. M. Giovane, J. Foresi, M. Morse, L. Liao, A. Agarwal, X. Duan, L. Kimerling, J. Michel, A. Thilderkvist, and E. Fitzgerald, "Materials for Monolithic Silicon Microphotonics," *Proceedings of the Symposium on Materials Devices for Silicon-Based Optoelectronics*, Warrendale, PA, 1998, pp. 45–56.
35. E. A. Fitzgerald and L. C. Kimerling, "Silicon-Based Microphotonics and Integrated Optoelectronics," *Mater. Res. Soc. Bull.* **23,** 4 (April 1998).
36. L. C. Kimerling, "Silicon Microphotonics," *Appl. Surf. Sci.* **159–160,** 8–13 (June 2000).
37. W. L. Ng, U. Lourenco, R. Gwilliam, S. Dedain, G. Shao, and K. Homewood, "An Efficient Room-Temperature Silicon-Based Light-Emitting Diode," *Nature* **410,** 192–194 (March 8, 2001).
38. J. W. Goodman, S. Kung, and R. Ahtale, "Optical Interconnections for VLSI Systems," *Proc. IEEE* **72,** 850–866 (July 1984).
39. S. K. Tewsbury and L. A. Hornak, "Optical Clock Distribution in Electronic Systems," *J. VLSI Signal Proc.* **16,** 225–246 (June–July 1997).
40. A. V. Mulé, S. Schultz, E. Glytsis, T. Gaylord, and J. Meindl, "Input Coupling and Guided Wave Distribution Schemes for Broad-Band Intrachip Guided Wave Optical Clock Distribution Networks Using Volume Grating Coupler Technology," *Proceedings of the IEEE International Interconnect Technology Conference*, June 4–6, 2001, pp. 128–130.
41. C. Zhou, J. Kong, and H. Dai, "Electrical Measurements of Individual Semiconducting Single-Walled Carbon Nanotubes of Various Diameters," *Appl. Phys. Lett.* **76,** 1597–1599 (March 2000).
42. V. Mulutinovic and P. Stenstrom, "Special Issue on Distributed Shared Memory Systems," *Proc. IEEE* **87,** 399–403 (March 1999).
43. W. J. Dally and J. W. Poulton, *Digital Systems Engineering*, Cambridge University Press, New York, 1998.
44. D. A. B. Miller, "Rationale and Challenges for Optical Interconnects to Electronic Chips," *Proc. IEEE* **88,** 728–748 (June 2000).

Received May 22, 2001; accepted for publication January 7, 2002

James D. Meindl *Microelectronics Research Center, Georgia Institute of Technology, 791 Atlantic Avenue, NW, Atlanta, Georgia 30332 (james.meindl@mirc.gatech.edu).* Dr. Meindl is the Director of the Joseph M. Pettit Microelectronics Research Center and the Joseph M. Pettit Chair Professor of Microelectronics at the Georgia Institute of Technology. He is also Director of the Interconnect Focus Center, a multi-university research effort managed jointly by the Microelectronics Advanced Research Corporation and the Defense Advanced Research Projects Agency for DoD. His current research interests focus on physical limits on gigascale integration. Dr. Meindl is a Life-fellow of IEEE and the American Association for the Advancement of Science, and a member of the American Academy of Arts and Sciences and the National Academy of Engineering and its Academic Advisory Board. He received his bachelor's, master's and doctor's degrees in electrical engineering from Carnegie Institute of Technology (Carnegie Mellon University).

Jeffrey A. Davis *Microelectronics Research Center, Georgia Institute of Technology, 791 Atlantic Avenue, NW, Atlanta, Georgia 30332 (jeff@ece.gatech.edu).* Dr. Davis received the B.E.E., M.S.E.E., and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology in 1993, 1997, and 1999, respectively. He joined the faculty at Georgia Tech as an Assistant Professor in 1999. In June 2000, Dr. Davis received the best student paper award for the 1999 International Interconnect Technology Conference (IITC) for interconnect modeling and design exploration of gigascale integrated (GSI) systems. In January 2001 he received the National Science Foundation CAREER Award to explore novel alternatives to global interconnect design for future GSI systems. Dr. Davis is currently the general chair of the 2002 System Level Interconnect Prediction (SLIP) workshop (*www.sliponline.org*).

Payman Zarkesh-Ha *LSI Logic Corporation, 1551 McCarthy Boulevard, Milpitas, California 95035 (payman@lsil.com).* Dr. Zarkesh-Ha is a Research Staff Member in the Interconnect Modeling Group in the Device Technology Division of LSI Logic Corporation. He received a B.S. degree in electrical engineering from the University of Science and Technology, Tehran, Iran, in 1992, an M.S. degree in electrical engineering from Sharif University, Tehran, Iran, in 1994, and a Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, in 2001. He subsequently joined LSI Logic Corporation, where he has worked on interconnect architecture design for the next ASIC generations. He is an author or coauthor of two patents and 25 technical papers. Dr. Zarkesh-Ha is a member of the Institute of Electrical and Electronics Engineers.

Chirag S. Patel *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (pchirag@us.ibm.com).* Dr. Patel is a Research Staff Member in the Science and Technology Department at the Thomas J. Watson Research Center. He received B.S., M.S., and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology in 1995, 1996 and 2001, respectively. In 2000 Dr. Patel received an outstanding paper award at the High Density Interconnect Conference for his work on the compliant wafer-level package. In 2001, he joined IBM at the Thomas J. Watson Research Center, where he has worked on advanced and exploratory packaging technologies. He is an author or coauthor of more than 25 technical papers. Dr. Patel is a member of the Institute of Electrical and Electronics Engineers.

Kevin P. Martin *Microelectronics Research Center, Georgia Institute of Technology, 791 Atlantic Avenue, NW, Atlanta, Georgia 30332 (kevin.martin@mirc.gatech.edu).* Dr. Martin is a Senior Research Scientist in the Microelectronics Research Center at the Georgia Institute of Technology. He received his B.S. degree in 1976, his M.S. degree in 1979, and his Ph.D. degree in 1982, all in physics from Ohio State University. He has since held research positions at Boston University, the Francis Bitter National Magnet Laboratory at MIT, the University of Oregon, and Georgia Tech (where he has worked since 1987). His research activities include the integer and fractional quantum Hall effect, resonant tunneling in quantum-well nanostructures, physics of compound semiconductors, plasma processing of semiconductors, properties of semiconductor nanostructures, nanofabrication, wafer-level packaging, and high-performance ultrahigh-density input/output interconnects. Dr. Martin is an author and coauthor of more than 50 peer-reviewed journal papers and coinventor of six allowed patents. He has been a member of the American Physical Society, the American Vacuum Society, and the Institute of Electrical and Electronics Engineers.

Paul A. Kohl *Microelectronics Research Center, Georgia Institute of Technology, 791 Atlantic Avenue, NW, Atlanta, Georgia 30332 (paul.kohl@che.gatech.edu).* Dr. Kohl received a Ph.D. degree from the University of Texas in 1978. He worked at AT&T Bell Laboratories from 1978 to 1989 in the area of materials and processes for semiconductor devices. In 1989, he joined the School of Chemical Engineering at the Georgia Institute of Technology, where he is currently Regents' Professor. His research interests include new materials and chemical processes for semiconductor and electrochemical devices.