The future of interconnection technology

by T. N. Theis

Continuing advances in interconnection technology are seen as essential to continued improvements in integrated circuit performance. The recent introduction of copper metallization, dual-damascene processing, and fully articulated hierarchical wiring structures, along with the imminent introduction of low-dielectric-constant insulating materials, indicates an accelerating pace of innovation. Nevertheless, some authors have argued that such innovations will sustain chip-level performance improvements for only another generation or two. In light of this pessimism, current trends and probable paths in the future evolution of interconnection technology are reviewed. A simple model is developed and used to estimate future wiring requirements and to examine the value of further innovations in materials and architecture. As long as current trends continue, with memory arrays filling an increasing fraction of the total area of high-performance microprocessor chips, wiring need not be a performance limiter for at least another decade. Alternative approaches, such as optical interconnections on chip, have little to offer while the incremental elaboration of the traditional wiring systems is still rapidly advancing.

Introduction

Advances in interconnection technology have played a key role in the continued improvements in integrated circuit density, performance, and cost per function. IBM has made sustained and major contributions, such as the planar multilevel metallization architecture [1] (introduced into IBM manufacturing in 1988), in which planarity was achieved by the extensive use of chemical–mechanical polishing (CMP) [2] and chemical vapor deposition (CVD) of conformal metals such as tungsten [1]. Planar wiring was a key innovation which improved structural integrity and facilitated continuous improvements in wiring pitch, number of metal levels, and design features such as stacked vias and local interconnections. These and other contributions to IBM interconnection technology were reviewed in the *IBM Journal of Research and Development* in 1995 by Ryan et al. [3]. A chronology of technology introductions was given, and the authors observed that the rate of introductions appeared to be increasing. This observation has been borne out by subsequent events.

IBM recently announced a six-level copper wiring process as a feature of its CMOS 7S logic technology [4] that is now in full-scale manufacturing. Along with the complete replacement of aluminum by copper for on-chip wiring, the technology represents a number of additional firsts in microelectronics. The industry-standard subtractive metal etching process has been replaced by a damascene (metal inlay) process—wire patterns (trenches and via holes) are etched in an insulator, then filled with copper, and the excess copper is then removed by chemical–mechanical planarization. For each wire level, both the via and trench structures are filled in a single step—the first use of a dual-damascene process for chip wiring. Furthermore, the copper is electrolytically deposited. With a minimum wiring pitch of 0.63 μm and aspect (height-to-width) ratios of the combined via and trench structures of the order of 3 to 4, CMOS 7S marks

0018-8646/00/$5.00 © 2000 IBM

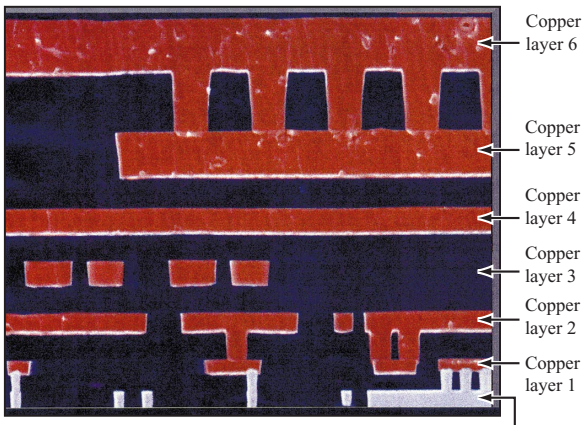

Figure 1

Cross-sectional scanning electron micrograph showing typical CMOS 7S interconnections with tungsten local interconnections and six levels of copper wiring. From [16], reproduced with permission of The Electrochemical Society, Inc.

the first industrial fabrication of deep-submicrometer-scale features by an electrochemical process.

Copper interconnections fabricated by a dual-damascene process offer advantages of performance, cost, and reliability over existing aluminum wiring processes [4]. Performance is gained because the resistivity of copper is approximately 40% lower than that of aluminum, so that copper wires exhibit approximately 40% lower RC delay than aluminum wires of the same cross section. Cost reduction comes from the elimination of some process steps and the simplification of other process steps in the dual-damascene process. Reliability is improved because the electrolytically deposited copper, when compared to aluminum, exhibits far less electromigration and far less stress migration. A detailed discussion of these advantages can be found in Reference [4].

At the same time, the semiconductor industry is moving toward the replacement of the traditional SiO_2 interlayer dielectric with one or more materials having a lower dielectric constant. Although the materials and related process integration issues are formidable, it is believed that a staged introduction of materials of lower and lower dielectric constant will be possible, leading to further improvements in wiring performance. Some companies have announced plans to introduce these insulators with the traditional aluminum wiring, while other companies are introducing copper wiring first, and then implementing new insulators. Paradoxically, in the midst of this rapid progress, a great deal of pessimism has been expressed regarding the prospects for further improvements. Much of this pessimism appears to be related to the fact that the industry is, for the first time, compelled to introduce truly hierarchical wiring systems.

Hierarchical wiring

A representative cross section of CMOS 7S wiring is shown in the scanning electron micrograph in **Figure 1**. The use of dual-damascene metal fill is evidenced by the absence of a seam between each copper wire layer and the via layer immediately below. The liner isolating the copper from the surrounding insulator is difficult to discern in the figure because it is very thin compared to the wire widths. Perhaps the most striking feature is the dramatic difference in size between the bottom and top levels of the wiring stack. Minimum contacted pitch is 0.63 μm at the high-density first copper level, 0.81 μm at succeeding levels, and the pitch and thickness of the fifth and sixth levels can be optionally scaled by a factor of 2×, as shown in the figure, for low RC delay.

CMOS 7S is thus an example of a hierarchical wiring system in which successive wire levels at increasing thickness and width enable long wire runs with low RC delay. Increasing the height and width of a wire and thickness of surrounding insulators, all by a factor of λ, leaves capacitance (C) per unit length unchanged, while resistance (R) per unit length is reduced by a factor of $1/\lambda^2$. In principle, RC delay can be reduced to arbitrarily low values by implementation of such "fat wires." This is sometimes referred to as reverse scaling.

A more complete model of RC delay in a logic circuit includes the effective internal resistance, R_t, of a driver transistor and the input capacitance, C_t, of the transistors that form the load, as well as the lumped resistance, R_w, and lumped capacitance, C_w, of the connecting wire. The total delay of the circuit is approximated by a weighted sum of delay terms of the form R_wC_w, R_wC_t, R_tC_w, and R_tC_t, with coefficients that depend on the circuit that is modeled [5]. Thus, reverse-scaling wire cross-sectional dimensions by λ while leaving wire length and transistor dimensions fixed causes R_wC_w and R_wC_t to decrease as $1/\lambda^2$ while the third wire-related delay term, R_tC_w, is unchanged. This last term, if significant, can be reduced by the use of a "wide" driver transistor which costs little in terms of additional chip area. Thus, all wire-related RC delays in this simple model circuit can easily be reduced. In general, the combination of reverse scaling and appropriate circuit design techniques prevents interconnection delays from overwhelming transistor delays in current microprocessor designs. Since a relatively small fraction of the wires represent long or critical paths, the cost of the additional metal area required to implement fat wires has so far been acceptable.

Will this situation continue? Can an existing hierarchical wiring system be scaled to future transistor densities, with acceptable wire-related RC delay at acceptable cost? First, the minimum wiring pitch must continue to scale with transistor dimensions. The scaling rate can be estimated from historical data, as shown in **Figure 2**. At the same time, an ever-increasing fraction of the total wires on chip will have to be implemented with larger pitches and thicknesses in order to meet RC constraints. To see that this is true, consider a circuit in which all transistor and wire dimensions are shrunk by a factor of $1/s$. Since the length of every wire in the circuit shrinks as $1/s$, R_w for any wire increases as s, while C_w decreases as $1/s$, and R_wC_w remains constant. However, in the simplest transistor scaling scenario, R_t is constant, C_t decreases as $1/s$, and thus R_tC_t decreases as $1/s$. Hence, some fraction of the wire runs which are acceptable in terms of RC delay before scaling will become unacceptable (compared to transistor switching delays) after scaling. To maintain parity with the improved transistor performance, such wires must be moved up the wiring hierarchy to the next larger pitch. This is awkward because it means that an existing wiring layout cannot be simply shrunk, but must instead be redesigned. Furthermore, the area required for the new wiring layout does not scale as $1/s^2$, so either the circuit area also fails to scale as $1/s^2$, or additional wiring levels must be added to contain those wires with pitches that do not scale. Since these fat wires require more area than traditional wires implemented at, or close to, minimum lithographic width, aggressive implementation of hierarchical wiring could drive the addition of wiring levels at a rate above the historical norm and the current industry projections, which are shown in **Figure 3**. The problem is compounded if, instead of shrinking an existing design (constant number of transistors), a new design is developed with more transistors and consequently additional long wires that must be implemented at large pitches.

Until recently the introduction of fat wires and truly hierarchical wiring structures has been avoided by the use of modified scaling approaches in which pitch is reduced while aspect ratio is increased [6]. However, with current aspect ratios as high as 2.2 [7], there is little incentive for further increases, at least from the point of view of RC reduction. This is because the well-known contribution of "sidewall" capacitance makes negligible the net reduction in RC [8], and in addition greatly exacerbates the problem of cross-coupling between adjacent wires. Introduction of low-dielectric-constant (low-ε) insulators can compensate for the increase in capacitance due to increasing aspect ratio. For instance, in the high-aspect-ratio limit, it can be readily shown that reducing the dielectric constant by $1/s$, while increasing the aspect ratio by s and reducing the wire pitch, wire length, and transistor dimensions by $1/s$, results in wire-related RC delay components (R_wC_w, R_wC_t, R_tC_w) that scale with transistor delay, R_tC_t. However, reducing the dielectric constant of the insulator does not reduce cross-coupling at all. This must be addressed by greater attention to circuit layout and improved design tools. Since the efficacy of future improvements in circuit

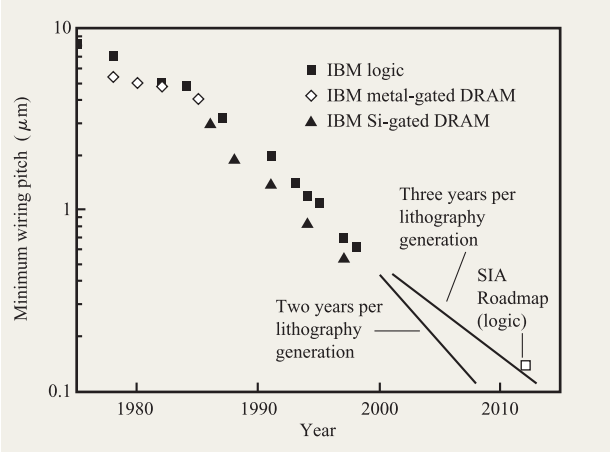

Figure 2

Minimum wire pitch used in IBM DRAM and CMOS logic technologies vs. year of introduction, and extrapolation of the current scaling trend into the future. Modified from [16], with permission of The Electrochemical Society, Inc.

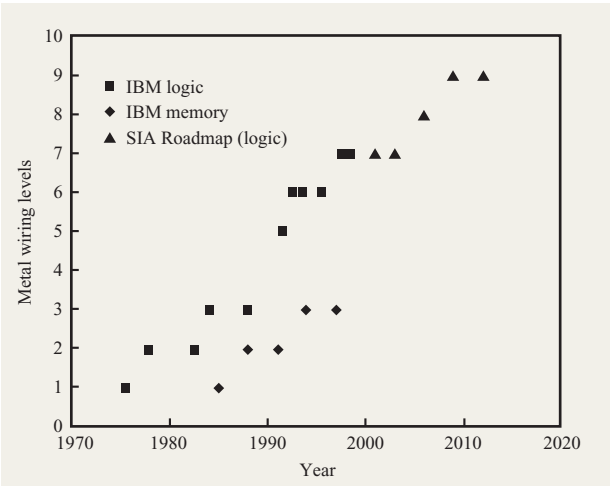

Figure 3

Number of wire levels used in IBM DRAM and CMOS logic technologies vs. year of introduction (includes tungsten local interconnections), and Semiconductor Industry Association (SIA) Roadmap values for future years. Modified from [16], with permission of The Electrochemical Society, Inc.

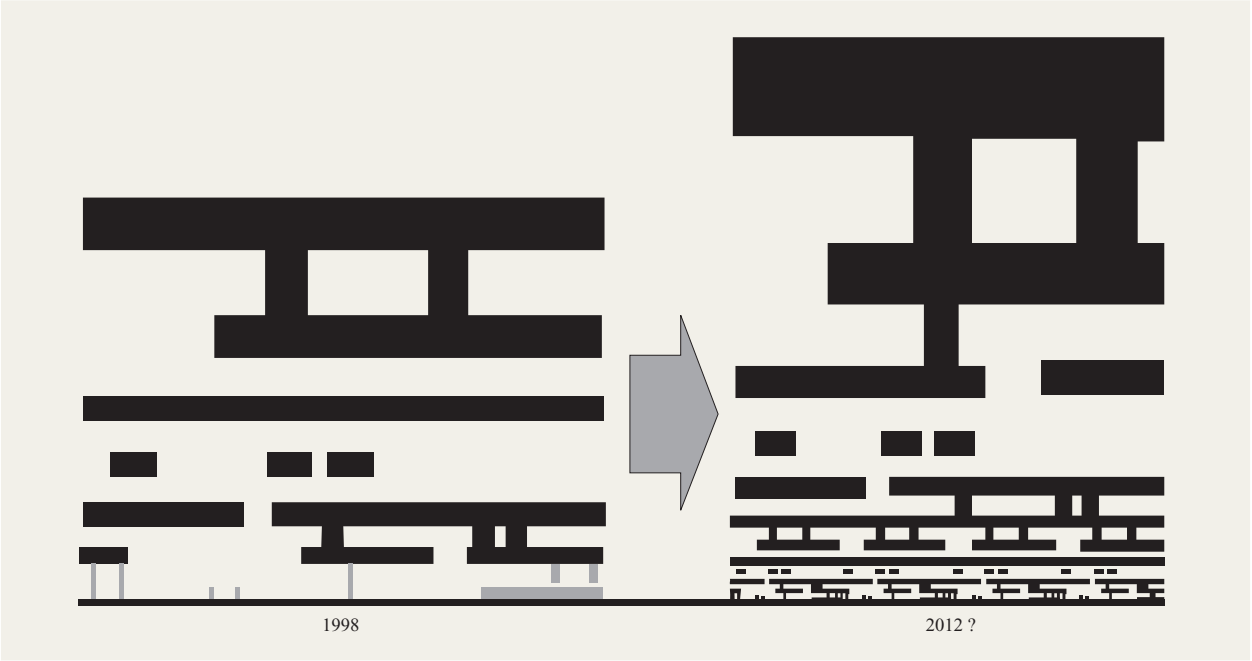

Figure 4

Schematic illustration of the likely evolution of interconnection architecture for high-performance CMOS logic. From [16], reproduced with permission of The Electrochemical Society, Inc.

layout methodology is unknown, it is difficult to predict the extent of future increases in wire aspect ratios. In the projection of future wiring needs which follows [Figure 6 (shown later) and associated discussion], a significant increase is assumed, roughly consistent with current Semiconductor Industry Association projections.

Of course, RC constraints are not the only factor driving the evolution of wiring systems. Wires for power distribution must be scaled to limit IR voltage drops, and must also avoid electromigration constraints. Long wires that operate as transmission lines must scale in width as the square root of the clock frequency for wires of constant length; the thickness of transmission lines need not increase at all with clock frequency once the thickness significantly exceeds the skin depth. Such considerations are additional reasons why, at this time, the metal thickness and minimum pitch of last or global wiring levels are no longer decreasing and will increase in the future. The combined effects of shrinking the minimum pitch, adding intermediate metal levels, and increasing the pitch of global wiring levels are schematically illustrated in **Figure 4**. Wiring systems will become increasingly hierarchical and increasingly three-dimensional, with an increasing disparity between minimum and maximum wire dimensions.

The wiring bottleneck

In principle, a hierarchical wiring system allows RC delay contributions of future wiring systems to be scaled to match improvements in device performance, but perhaps at the cost of introducing additional wire levels at a rate above the historical trend [3, 9] of roughly 0.75 levels per lithography generation. The introduction of copper wiring [4, 10] and, eventually, low-dielectric-constant insulators is seen by some as a partial solution, containing the proliferation of additional wiring levels for a generation or two before chip-level performance becomes limited by wiring [9]. Of course, wiring has long been identified as an eventual limiter of integrated circuit performance [11–14], but the National Technology Roadmap for Semiconductors (NTRS) now estimates that without radical material, design, or architectural innovations, this point will be reached at the 0.1-μm generation [15]. Here we show that any such estimate is extremely sensitive to some basic assumptions about chip design and architecture. In particular, wiring demands are very sensitive to the fraction of chip area that is devoted to random logic, as opposed to more regular arrays such as memory, registers, and so on. A brief version of this argument has been published elsewhere [16].

For purposes of this demonstration, the total area occupied by wires on a chip is approximated by

$$A_{\text{tot}} = \int w(L) L f(L, N)\, dL, \tag{1}$$

where $w(L)$ is the wire pitch as a function of wire length, L, and $f(L, N)$ is the wire length distribution function which must depend on N, the total number of transistors. The usual starting point in determining $f(L, N)$ is Rent's rule [17], an empirical relationship that specifies the number of wires which cross the boundary of a block of circuitry (input/output or I/O wires), K, in terms of the number of transistors or nodes within the block, N, and the number of wires, k, connecting each transistor to other transistors within the block,

$$K = kN^{p}. \tag{2}$$

The Rent exponent, p, is observed to vary from $0.55 < p < 0.85$, the lower values for highly regular circuits such as memory, and the higher values for random logic. Thus, memory chips require fewer interconnection levels than logic chips of roughly the same size, as is evident from Figure 3. Rent's rule is observed to hold for circuit blocks of widely varying size. If its validity is assumed at all length scales greater than the transistor-to-transistor spacing, a power-law distribution is obtained for the number of wires as a function of wire length. For wires on a chip, there is a natural cutoff in this distribution function at wire lengths longer than roughly the length of the chip. This cutoff has been treated with perhaps the greatest degree of rigor by Davis et al. [18], and we use the wire length distribution function derived by these authors. The minimum wire pitch implemented in the first level of wiring, $w_{1,\text{min}}$, is always determined by lithographic capabilities. Once $w_{1,\text{min}}$ and the aspect ratio are set, there will be maximum acceptable wire length at minimum pitch, $L_{1,\text{max}}$. The maximum run length, $L_{n,\text{max}}$, at minimum allowed pitch, $w_{n,\text{min}}$, for each higher-lying wire level, n, in the wiring hierarchy is chosen to satisfy a constraint on minimum performance of the wiring system. The particular constraint, or reverse-scaling relationship, considered here is

$$w_{n,\text{min}} \propto (L_{n,\text{max}})^{1/2}. \tag{3}$$

Assuming that aspect ratios and metal resistivity are unchanged between wiring levels, this constraint guarantees that maximum wire resistance is the same at each level in the wiring hierarchy (constant-resistance scaling). Thus, as successively longer lines are implemented at successively higher levels in the wiring hierarchy, the maximum $R_{\text{w}}C_{\text{w}}$ delay increases as L, and therefore increases no faster than the delay of a lossless transmission line. The maximum value for $R_{\text{t}}C_{\text{w}}$ also increases as L, and the maximum value of $R_{\text{w}}C_{\text{t}}$ is independent of L. These favorable scaling properties ensure that the application of well-known design approaches (repeaters, cascaded drivers) can reduce the delay of critical paths toward the physical limit set by transmission-line delays. Note, however, that Equation (3) is not a unique constraint. A less aggressive reverse-scaling scenario would require less wire area for a given circuit, and might still yield acceptable interconnection delay for a chip design.

Equations (1) and (3) imply a simple and idealized interconnection system. Equation (1) implies that $w(L)$ is a single-valued function of L, whereas in actual interconnection systems designers are generally free to run wires of a given length at various widths equal to or greater than the minimum specified width for each level. Such "wide" wires (as opposed to fat wires) are used to reduce wire resistance when this is a limiting factor (e.g., IR drops in power distribution), or to increase wire cross section where current density is constrained (e.g., by electromigration). Wide wires are also used to reduce $R_{\text{w}}C_{\text{w}}$ delays, but provide little benefit once the aspect ratio is reduced toward 1 or less. Equation (3) implies that all wires of length greater than $L_{n,\text{max}}$ are implemented in wire level $n + 1$ or above, whereas in actual wiring systems not all wires are critical in terms of delay, and the distribution of wire lengths in any particular wire level may not have a sharp cutoff. Thus, the wire area calculated using Equations (1) and (3) will differ from the wire area of a real wiring system, even if the wire distribution function, $f(L, N)$, accurately models the actual wire distribution function. However, we do not use Equations (1) and (3) to generate values for the wire area, but rather to calculate the *relative increase* in wire area as a hierarchical wiring system is implemented with wires of minimum width drawn at successively smaller lithographic dimensions while simultaneously adding wire levels and pitches to maintain the reverse-scaling scenario [Equation (3)]. Our results will be most accurate if wiring practices, such as the relative use of wide versus fat wires, do not change from lithography generation to lithography generation. Our underlying assumption is that if we underestimate the wiring area for the present generation, we underestimate it for future generations as well by the same factor, and the relative increase in wire area from generation to generation is accurately estimated. Actually, this underlying assumption is probably unduly pessimistic. With more levels available in future wiring systems, and with better circuit design tools able to optimize the use of the hierarchical levels, we may expect *less* use of wide wires, more area-efficient designs, and thus *less* relative increase in wiring area than is projected by this simple model.

It would now be straightforward to choose values of $L_{n,\max}$, $n > 1$, so as to distribute the wire area [Equation (1)] evenly among a discrete set of metal levels, and to repeat this procedure iteratively, varying the total number of metal levels so as to optimize the metal fill factor at each level. However, here we avoid this tedious iterative procedure by further idealizing our model wiring system—we assume that for wires longer than $L_{1,\max}$, wire pitch varies smoothly as a function of wire length,

$$w \propto L^{1/2}, \tag{4}$$

so that *each wire* just satisfies the reverse-scaling constraint on resistance as a function of length. Although such an ideal minimum-area wiring system cannot be realized in practice, a hierarchical wiring system becomes a better approximation of the ideal system as more wire levels are added. Future wiring will therefore better approximate the ideal system of our model. Since we compare future lithography generations to a current generation, the *relative* increase in wire area projected from our idealized model is expected to be pessimistically large.

To establish a baseline for our projections, we pick $L_{1,\max}$ roughly appropriate for an existing CMOS generation. The pitch of all runs of length $L \leq L_{1,\max}$ is $w_{1,\min}$, while the pitch of all runs of length $L \geq L_{1,\max}$ is

$$w = [w_{1,\min}(L/L_{1,\max})^{1/2}], \tag{5}$$

consistent with Equation (4). We now choose a lithographic scaling scenario for the wiring system. To maintain ideal scaling of transistor density, $w_{1,\min}$ must scale as $1/s$. We choose to scale $L_{1,\max}$ as $1/s^2$, which means that the maximum number of gate pitches spanned by a wire at minimum width scales as $1/s$, and the maximum number of transistors that can be connected by wires at minimum pitch scales as $1/s^2$. This choice guarantees that the maximum resistance of a wire of minimum width is constant over lithographic generations. (Note that this choice is independent of our choice of a constant-resistance reverse-scaling scenario, which ensures that the maximum wire resistance is the same for each wire level.) Scaling $L_{1,\max}$ as $1/s^2$ results in wires of minimum lithographic width having maximum wire-related delays of the forms R_wC_t, R_tC_w, and R_wC_w scaling as $1/s$, $1/s^2$, and $1/s^2$, respectively. That is, for wires of minimum lithographic width, all wire-related delays scale as fast as or faster than device delay.

The relative increase in metal area required to wire the system can now be calculated from Equation (1) as $w_{1,\min}$ is scaled with each lithographic generation. In the following discussion, each lithography generation corresponds to a full $1/\sqrt{2}$ reduction in all minimum lithographic dimensions, or a doubling of transistor density. Three scenarios are considered: 1) the number of transistors doubles with each lithographic generation ($N \propto s^2$) to fill a chip of constant size; 2) the number of transistors doubles every two lithographic generations ($N \propto s$) to fill a diminishing fraction, $1/s$, of a chip of constant size; and 3) the number of transistors is constant over lithography generations, filling a rapidly diminishing fraction, $1/s^2$, of a chip of constant size.

Using the reverse scaling relationship, Equation (5), and scaling $w_{1,\min}$ as $1/s$ and $L_{1,\max}$ as $1/s^2$, it is easy to verify that wires of constant length (the longest wires in scenario 1) are of constant width, and thus have a maximum delay of the form R_wC_t which scales with lithographic generation as $1/s$, and maximum delays of the forms R_tC_w and R_wC_w which are constant from generation to generation. Thus, for wires of fixed length, our model scales wire RC delay much more aggressively than the historic trend, in which delay has actually *increased* by a factor of about 1.26 per lithography generation [9]. However, the historic trend is not a good guide to the future, as wire-related delays have only recently become an issue. Since interconnection delay does not scale with device delay for interconnections which do not scale in length, our model is probably too conservative for cases in which the transistor count grows from generation to generation (scenarios 1 and 2) and the clock is to be distributed globally. For scenario 1, in which the longest wires are of fixed length, the use of optimized repeaters [19] allows delay to scale no faster than $1/s^{1/2}$. Thus, scenarios 1 and 2 are more appropriate for modeling the wire demands of a hierarchical clock system, in which a core runs at the highest clock frequency, with peripheral circuitry clocked at a lower rate. Furthermore, our model considers only RC delays. As time-of-flight delays become significant in future chips, the longer interconnections must approximate lossless transmission lines. In the limit that skin depth is small compared to metal thickness, the width of transmission lines of fixed length (scenario 1) must scale as the square root of clocking frequency in order to limit resistive loss. Since scaling of device delay accounts for only about half of the historical rate of improvement in clock frequency, scaling the width of wires of fixed length as s is roughly equivalent to scaling the width as the square root of clock frequency. In summary, our model probably underestimates the wire demands of blocks of random logic which grow in transistor count from generation to generation and which require global distribution of the clock frequency.

On the other hand, our model is certainly adequate for the case in which a block of random logic of fixed transistor count is shrunk through many lithographic generations. For scenario 3, the longest wires in the system scale in length and width as $1/s$ and have maximum delays of the forms R_wC_t, R_tC_w, and R_wC_w which all scale as $1/s$. Thus, our choice of constant-resistance reverse

scaling and constant maximum wire resistance over lithographic generations allows us to scale a circuit of fixed transistor count from generation to generation while ensuring that all wire-related delay components scale with device delay. This will be true regardless of the number of transistors in the initial design. For example, a processor designed at 0.25-μm lithographic ground rules and containing ten million transistors, operating at a clock frequency of 500 MHz, and with total clock cycle time partitioned evenly between device delay and interconnection delay, would, six lithographic generations later, occupy 1/64th the original area, operate at 4 GHz, and still have a total cycle time evenly partitioned between device delay and interconnection delay.

Note, finally, that none of these scenarios precludes additional improvements in clock frequency through advances in design and architecture. For many years, CMOS microprocessor clock frequencies have increased at a rate only partially explained by reduction in device delay due to scaling. Much of the additional improvement has come from architectural advances (primarily reduced switching operations per clock cycle) and design advances (for example, better timing of critical paths, leading to better "speed sorts" in manufacturing and testing). Such improvements do not require that wire RC delays scale any faster than device delay.

Figure 5 shows the relative increase in the number of wire levels required at each lithography generation for a hierarchical wiring system which is scaled according to the preceding discussion. In this and the other projections that follow, the value for the Rent exponent is chosen as $p = 0.85$, valid only for random logic, and a pessimistic estimate even in that case. At lithography generation 0, the maximum run at the minimum lithographically allowed wire width is chosen as $L_{1,\max} = 0.4N^{1/2}$ (that is, 0.4 chip lengths), a value roughly appropriate for the 0.25-μm lithography generation and aluminum wiring. Thus, the projection of wiring needs extends about six lithographic generations beyond 0.25 μm. (The general results and trends obtained here are insensitive to the precise choice of $L_{1,\max}$.)

As can be seen, filling a chip of constant size with an exponentially increasing number of transistors (scenario 1; $N \propto s^2$) soon leads to an explosion in the number of wire levels required to avoid an RC delay bottleneck. If the number of wiring levels at generation 0 is taken to be six, scenario 1 indicates that about nine times as many levels (54 levels!) would be required four lithography generations later. On the other hand, scenario 2 ($N \propto s$) and scenario 3 (constant transistor count) appear much more manageable. Scenario 2 may be closer to reality than scenario 1, since the largest chips for the most demanding logic applications are being increasingly filled with memory and other regular arrays, while random logic appears to be shrinking in absolute area.

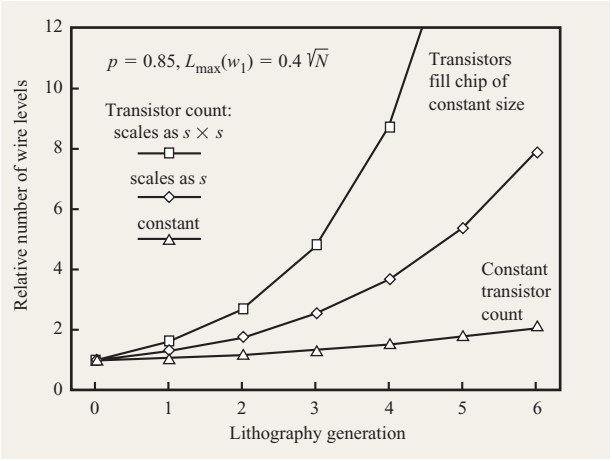

Figure 5

Relative number of wire levels required at each lithography generation for a hierarchical wiring system in which wire RC delay is scaled with transistor performance. From [16], reproduced with permission of The Electrochemical Society, Inc.

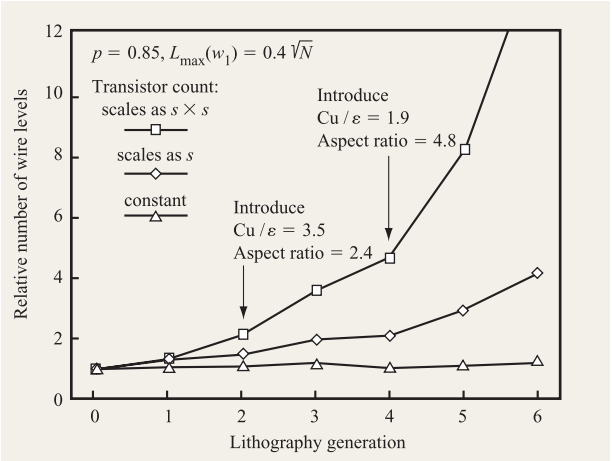

Figure 6

Relative number of wire levels required at each lithography generation as in Figure 5, but with the staged introduction of copper, insulators with successively lower dielectric constants, and increased aspect ratios. From [16], reproduced with permission of The Electrochemical Society, Inc.

What is the value of introducing new materials? **Figure 6** shows the same three scenarios as Figure 5, but we introduce new materials and higher-aspect-ratio wires in a

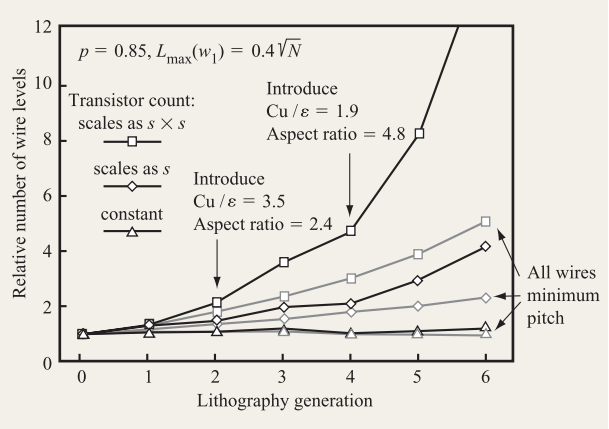

Figure 7

Relative number of wire levels required at each lithography generation as in Figure 6, compared with the relative number of wire levels required if all wires are implemented at a minimum lithographically defined width.

manner roughly following the National Technology Roadmap for Semiconductors [15]. At lithography generations 2 and 4, $L_{1,max}$ is held constant (rather than being scaled as $1/s^2$), and the material properties and aspect ratios are chosen so that all wire-related delays continue to scale with device delay for wires which scale in length as $1/s$. Specifically, at lithography generation 2 we introduce copper wires and a low-ε insulator ($\varepsilon \simeq 3.5$) and increase the aspect ratio by a factor of 4/3. At lithography generation 4 we introduce a lower-dielectric-constant insulator ($\varepsilon \simeq 1.9$) and further increase the wire aspect ratio by a factor of 2. The resulting very aggressive aspect ratio of 4.8 for lithography generations 4 through 6 is necessary to meet the constraint of our model that *all* wire-related RC components scale with device delay. This should not be taken as a recipe for design of a practical wiring system, since consideration of cross-coupling may limit aspect ratios to lower values. Reducing the dielectric constant to 1.3 and increasing the aspect ratio to a smaller value of 3.5 at lithography generation 4, consistent with Roadmap targets, would still reduce delays of the form R_tC_w and R_wC_w by $1/s$ or more, but would allow R_wC_t delays to increase. Since delay terms of the latter form are usually insignificant for longer wires, the lack of scaling for a single lithography generation would almost certainly be acceptable.

The number of wire levels still quickly becomes unmanageable in scenario 1, but remains tractable at least to generation 4 or 5 in scenario 2, and barely increases at all in scenario 3. Scenario 3 implies that any wire which spans a certain number of gate pitches at lithography generation 0 will still span that many gate pitches at lithography generation 4. The wire RC delay will scale approximately with the device delay; indeed, the rate of improvement in material properties was chosen to ensure this result. Such a scenario is attractive for several reasons. Existing circuit layouts can be relatively easily scaled to reduced lithographic dimensions, without the substantial redesign required if additional metal levels are added to the wiring hierarchy. Design tools need not evolve as rapidly as they must if increasingly hierarchical wiring systems are needed to wire circuits of fixed size. Such advantages of scalability of designs and tools, rather than the need to avoid a looming crisis in wirability, justify the improvements in materials and wire aspect ratio called for by the National Technology Roadmap for Semiconductors.

Scenario 3 of Figure 6 also implies that 64 present-day microprocessors could eventually be wired on a single chip, each core surrounded by memory and running at a multigigahertz clock rate, all with about six or seven levels of metal. Note that for many (perhaps most) applications, the processors can cooperate effectively without having to exchange information at the clock rate of the individual processors. Simple switching and routing protocols might therefore allow communication between processors at negligible cost in terms of additional wire area. The point is that while a wiring bottleneck must eventually limit progress in the implementation of ever-more-complex, ever-faster random logic circuits, architectural solutions may allow the limit to be avoided. Wiring issues need not halt the exciting progress in microelectronics technology for the foreseeable future.

And what if we dream about materials that do not yet exist? The capacitance per unit length is insensitive to the cross-sectional geometry, and the insulator dielectric constant cannot be reduced below 1. But cryogenic cooling dramatically reduces the resistivities of both aluminum and copper, and furthermore, wiring materials with a vanishingly small resistivity are *possible*. Surprisingly, the value of a zero-resistivity conductor is not that great, at least for the next few lithographic generations. **Figure 7** shows how the projected wiring needs of Figure 6 are reduced in the case of an ideal conductor which allows all wires, regardless of length, to be implemented at the minimum lithographically determined width. An ideal conductor would not greatly reduce the metal area of current wiring systems because the vast majority of all wires on a chip are short, and therefore *are* implemented at, or close to, minimum lithographic dimension. Furthermore, an ideal conductor only delays the inevitable explosion in wire levels that must occur for two-dimensional integration of ever more devices, so long as the Rent exponent is greater than 0.5.

Optical interconnections

With growing awareness of wiring as a potential performance bottleneck has come a growing interest in the possibility of replacing electrical connections with optical connections. Most proposals suggest a hybrid approach in which optical signals are generated in, routed through, and received in an optical "interposer" which is to be bonded to the silicon chip. This avoids the very difficult problem of integrating optically active elements, especially light emitters, into silicon technology. Recent progress in fabricating dense arrays of tiny vertical-cavity surface-emitting lasers (VCSELs), electro-optic modulators, and optical detectors gives some substance to these proposals [20].

Miller [20, 21] has listed and discussed potential technical advantages of optical interconnections, including elimination of resistive losses, reduced power consumption, avoidance of frequency-dependent cross-coupling, and improved electrical noise immunity. These advantages provide a convincing rationale for the eventual replacement of electrical connections at all length scales down to the chip scale, and perhaps for replacement of some wiring on chip. But when, at a given length scale, will optical interconnections be cheaper than wires that perform the same function? In the case of on-chip wiring, certainly not until the replacement of wires results in smaller-area and (therefore) cheaper chips.

Recent comprehensive tests of a six-layer copper/silicon dioxide wire structure showed that coupled wire pairs of approximately 1.2 μm thickness and 4.5 μm pitch will sustain clock frequencies in excess of 5 GHz over distances of about 1 cm [22]. At this frequency, these wires are already operating in the limit at which the electrical skin depth is of the order of the wire thickness. Thus, scaling to higher frequencies or longer runs requires only an increase in wire width, not thickness. To contain resistive losses, wire width should scale as the wire length and as the square root of the frequency. The minimum pitch of practical optical waveguides is of the order of 10 to 20 μm. So long as each optical waveguide replaces only one electrical wire, it will be decades before optical interconnections can compete on the basis of areal density. Waveguides are eliminated in proposals for free-space propagation of optical signals [20]. Waveguides based on the photonic bandgap principle [23] might reduce the width to a few micrometers, and should be investigated. However, the greatest hope for increased areal density of interconnections is to exploit the wavelength division multiplexing (WDM) capabilities of optical interconnections so that one optical connection replaces many wires.

Miller and Ozaktas [24] have shown that the maximum number of bits per second that can be carried by a simple electrical interconnection is

$$B \approx B_0 A/L^2, \tag{6}$$

in which A is the cross-sectional area of the conductor, L is the interconnection length, and B_0 is weakly dependent on geometry with a value of about 10^{16} bits/s for wires on a chip. The dimensionless ratio, $\sqrt{A}/L$, thus determines the bit rate which can be supported by a wire. Because optical waveguides are not subject to the resistive loss physics that gives rise to Equation (6), the bit rates supported by long optical interconnections of a given $\sqrt{A}/L$ can vastly exceed the bit rates supported by long wires with the same $\sqrt{A}/L$. Miller and Ozaktas estimate that this inherent superiority of optical interconnections, currently exploited at length scales of meters to thousands of kilometers, may eventually be exploited for on-chip interconnections as information bandwidths on chip approach 1 Tb/s.

The results of a recent theoretical investigation [25] suggest that it may be possible, utilizing photonic bandgap materials and techniques, to eventually miniaturize WDM components to the point at which they can be integrated on chip or in chip-scale interposers. This approach is of great interest and should be pursued. However, optical interconnections will never replace all wires on a chip. With minimum wire pitches of about 0.6 μm already in production, waveguides are simply too large to replace the vast majority of wires on current chips, not to mention the much smaller wires on future chips. Indeed, at sufficiently short distances, current wiring systems can, in principle, already support aggregate data rates greater than 1 Tb/s. Thus, for microprocessor architectures, the use of on-chip optical interconnections may be limited to replacement of only the widest and thickest signal wires.

Novel architectures

The speed of light will soon begin to limit interconnection performance, and this will be equally true of optical waveguides or conductive wires. For the last two decades, CMOS microprocessor clock speeds have increased at a rate of about 30 percent per year, so that the distance a signal can propagate in a clock cycle has decreased by a factor of about 0.75 each year. Currently, a signal propagates about 20 cm in a single 700-MHz clock cycle. If the trend in clock rate continues, this distance will shrink to 1 cm in about ten years. Reduction in insulator dielectric constant can increase signal propagation velocity by a factor of 2 at best, adding only two to three years to this scenario. One solution may be the implementation of a hierarchy of clock frequencies on a chip, with the highest frequency distributed only within a comparatively small logic core and its associated registers and fast cache. Such an architecture can be viewed as a simple mapping of current practice at the board and system level to future chips, and can obviously be extended to the implementation

of multiple processors on a chip, as suggested above in the discussion of Figure 6.

There also appears to be ample room for architectural improvements at the level of individual processors [26]. Such architectures reduce the need for global communications within a processor by clustering execution units with associated data and instruction storage. Within each of these clusters, communications are reduced by extending the storage hierarchy to the level of individual execution unit inputs. Finally, when global communications are required, these are implemented over a shared packet network by a variety of simple switching and routing schemes. Such architectures appear capable of greatly reducing bandwidth (hence wire area) requirements for global communications, although many issues such as compiler and run-time software optimization remain to be addressed [27].

An optimistic view

There are many reasons to take a guardedly optimistic view of future interconnection technology development. Some estimates of future wiring needs make the worst-case assumption that high-performance random logic completely fills chips of ever-increasing area. A more realistic assessment takes account of the fact that, in the most aggressive current microprocessor designs, random logic is actually shrinking while memory is growing as a fraction of total chip area. As discussed above, this significantly delays (but does not eliminate) the need for additional wiring levels.

Also, as wire delays become increasingly important, we can expect further innovation in design methodologies and architectures. Design tools must eventually account for signal delay at all levels of the wiring hierarchy and in all stages of the design process. Novel, currently experimental, architectures appear capable of significantly reducing wiring needs. We can foresee innovations such as stacked device structures and active devices in the interconnection levels (three-dimensional integration), and optical links on chip (no faster than electrical transmission lines, but operating at lower power and with the potential for much higher bandwidth through wavelength division multiplexing). Such advances are more likely as increased resources are devoted to the "interconnection problem." The Microelectronics Advanced Research Corporation (MARCO) [28], a recently formed subsidiary of the Semiconductor Research Corporation, has significantly increased university funding for long-time-horizon research in architecture and design, as well as materials, processes, and structures.

Finally, we are just beginning to implement fully hierarchical wiring systems and just learning to integrate copper with low-dielectric-constant insulators. Aspect ratios will be further increased as design tools and methodology are further improved to avoid the penalties imposed by increased cross-coupling. Copper wiring patterned by the dual-damascene method offers substantial process simplification, with an accompanying potential for cost reduction and improved manufacturing yields. Thus, the addition of many more wire levels, perhaps more than the NTRS estimate of nine in the year 2012, may be economically viable. Certainly, an important and inescapable challenge for future interconnection technology is to continue to articulate the wiring hierarchy, with an ever-increasing disparity between the minimum pitches of first and last metal levels, and a steadily increasing number of intermediate levels at intermediate pitches—an increasingly three-dimensional system. The extension of the technology should keep many talented scientists and engineers busy for the foreseeable future.

Acknowledgments

The author thanks James Ryan for providing much of the numerical data for Figure 2, Paul Solomon for an introduction, more than fifteen years ago, to the scaling properties of hierarchical wiring systems, George Sai-Halasz for many conversations over the years, Robert Dennard for discussions on the subject of cross-coupling in wiring systems, and Eric Kronstadt and Gregory Northrop for discussions on the rate of increase of microprocessor clock frequencies.

References

1. C. W. Kaanta, W. J. Cote, J. E. Cronin, K. L. Holland, P. I. Lee, and T. M. Wright, "Submicron Wiring Technology with Tungsten and Planarization," *Proc. IEEE IEDM*, pp. 209–212 (1987).
2. H. Landis, P. Burke, W. Cote, W. Hill, C. Hoffman, C. Kaanta, C. Koburger, W. Lange, M. Leach, and S. Luce, "Integration of Chemical-Mechanical Polishing into CMOS Integrated Circuit Manufacturing," presented at the 19th International Conference on Metallurgical Coatings and Thin Films, San Diego, CA, 1992.
3. J. G. Ryan, R. M. Geffken, N. R. Poulin, and J. R. Paraszczak, "The Evolution of Interconnection Technology at IBM," *IBM J. Res. Develop.* **39,** 371 (1995).
4. D. Edelstein, J. Heidenreich, R. Goldblatt, W. Cote, C. Uzoh, N. Lustig, P. Roper, T. McDevitt, W. Motsiff, A. Simon, J. Dukovic, R. Wachnik, H. Rathore, R. Schulz, L. Su, S. Luce, and J. Slattery, "Full Copper Wiring in a Sub-0.25 μm CMOS ULSI Technology," *Proc. IEEE IEDM*, pp. 773–776 (1997).
5. H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley Publishing Co., Inc., Reading, MA, 1990, pp. 202–204.
6. *Ibid*., p. 197.
7. S. Yang, S. Ahmed, B. Arcot, R. Arghavani, P. Bai, S. Chambers, P. Charvat, R. Cotner, R. Gasser, T. Ghani, M. Hussein, C. Jan, C. Kardas, J. Maiz, P. McGregor, B. McIntyre, P. Nguyen, P. Packan, I. Post, S. Sivakumar, J. Steigerwald, M. Taylor, B. Tufts, S. Tyagi, and M. Bohr, "A High Performance 180 nm Generation Logic Technology," *Proc. IEEE IEDM*, pp. 197–200 (1998).
8. Bakoglu, *op. cit*., p. 140.

9. M. T. Bohr, "Interconnect Scaling—The Real Limiter to High Performance ULSI," *Proc. IEEE IEDM*, pp. 241–244 (1995).
10. S. Venkatesan, A. V. Gelatos, V. Misra, B. Smith, R. Islam, J. Cope, B. Wilson, D. Tuttle, R. Cardwell, S. Anderson, M. Angyal, R. Bajaj, C. Capasso, P. Crabtree, S. Das, J. Farkas, S. Filipiak, B. Fiordalice, M. Freeman, P. V. Gilvert, M. Herrick, A. Jain, H. Kawasaki, C. King, J. Klein, T. Lii, K. Reid, T. Saaranen, C. Simpson, T. Sparks, P. Tsui, R. Venkatraman, D. Watts, E. J. Weitzman, R. Woodruff, I. Yang, N. Bhat, G. Hamilton, and Y. Yu, "A High Performance 1.8V, 0.29 μm CMOS Technology with Copper Metallization," *Proc. IEEE IEDM*, pp. 769–772 (1997).
11. K. C. Saraswat and F. Mohammadi, "Effect of Scaling of Interconnections on the Time Delay of VLSI Circuits," *IEEE J. Solid-State Circuits* **ED-32,** 275 (1982).
12. H. B. Bakoglu and J. D. Meindl, "Optimal Interconnection Circuits for VLSI," *IEEE Trans. Electron Devices* **ED-32,** 903 (1985).
13. P. M. Solomon, "The Need for Low Resistance Interconnects in Future High-Speed Systems," *Proc. SPIE* **947,** 104 (1988).
14. G. A. Sai-Halasz, "Directions in Future High-End Processors," *ICCD Digest*, p. 230 (1992).
15. *The National Technology Roadmap for Semiconductors: Technology Needs*, Semiconductor Industry Association, 1997, pp. 101–102.
16. T. N. Theis, "Challenges in the Extension of Hierarchical Wiring Systems," *Electrochemical Processing in ULSI Fabrication I and Interconnect and Contact Metallization: Materials, Processes, and Reliability*, Vol. **98-6,** P. C. Andricacos, J. O. Dukovic, G. S. Mathad, G. M. Oleszek, H. S. Rathore, and C. Reidsema Simpson, Eds., The Electrochemical Society, Inc., Pennington, NJ, 1999, pp. 1–11.
17. W. E. Donath, "Wire Length Distribution for Placements of Computer Logic," *IBM J. Res. Develop.* **25,** 152 (1981).
18. 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 (1998).
19. Bakoglu, *op. cit*., p. 214.
20. D. A. B. Miller, "Dense Two-Dimensional Integration of Optoelectronics and Electronics for Interconnections," *Heterogeneous Integration: Systems on a Chip*, A. Husain and M. Fallahi, Eds., *SPIE Critical Reviews of Optical Engineering*, Vol. **CR70,** SPIE, Bellingham, WA, 1998, pp. 80–109 and references therein.
21. D. A. B. Miller, "Physical Reasons for Optical Interconnection," *Int. J. Optoelectron.* **11,** 155 (1997).
22. A. Deutsch, H. Harrer, C. W. Surovic, G. Hellner, D. C. Edelstein, R. D. Goldblatt, G. A. Biery, N. A. Greco, D. M. Foster, E. Crabbe, L. T. Su, and P. W. Coteus, "Functional High-Speed Characterization and Modeling of a Six-Layer Copper Wiring Structure and Performance Comparison with Aluminum On-Chip Interconnections," *Proc. IEEE IEDM*, 1998, pp. 295–298.
23. J. D. Joannopoulos, R. D. Meade, and J. N. Winn, *Photonic Crystals*, Princeton University Press, New York, 1995; J. D. Joannopoulos, P. R. Villeneuve, and S. Fan, "Photonic Crystals: Putting a New Twist on Light," *Nature* **386,** 143–149 (1997).
24. D. A. B. Miller and H. M. Ozaktas, "Limit to the Bit-Rate Capacity of Electrical Interconnects from the Aspect Ratio of the System Architecture," *J. Parallel Distr. Computing* **41,** 42 (1997).
25. S. Fan, P. R. Villeneuve, J. D. Joannopoulos, and H. A. Haus, "Channel Drop Tunneling Through Localized States," *Phys. Rev. Lett.* **80,** 960 (1997).
26. S. Keckler and W. Dally, "Processor Coupling: Integrating Compile-Time and Run-Time Parallelism," *Proceedings of the 19th Annual International Symposium on Computer Architecture*, 1992, pp. 202–213.
27. W. J. Dally, "Interconnect-Limited VLSI Architecture," *Proceedings of the 1999 International Interconnect Technology Conference*, San Francisco, May 24–26, 1999, pp. 15–17.
28. *http://marco.fcrp.org*.

Received August 9, 1999; accepted for publication November 9, 1999

Thomas N. Theis *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (ttheis@us.ibm.com).* Dr. Theis received the B.S. degree in physics from Rensselaer Polytechnic Institute in 1972, and M.S. and Ph.D. degrees from Brown University in 1974 and 1978, respectively. A portion of his Ph.D. research was done at the Technical University of Munich, where he completed a postdoctoral year before joining IBM Research in 1979. Dr. Theis joined the Department of Semiconductor Science and Technology at the IBM Thomas J. Watson Research Center to study electronic properties of two-dimensional systems. He also collaborated in research on surface-enhanced Raman scattering, light emission from tunnel junctions, and conduction in silicon dioxide. The latter work helped to lay the basis for the present understanding of conduction in wide-bandgap materials. In 1982 he became manager of a group studying growth and properties of III–V semiconductors. Dr. Theis published extensively on the DX-center, a donor-related defect which limits the digital performance of some III–V transistors. In 1989 he was named Senior Manager, Semiconductor Physics and Devices. In 1993 he was named Senior Manager, Silicon Science and Technology, responsible for exploratory materials and process integration work bridging between IBM Research and IBM Microelectronics. While in this position, he was the principal author of IBM's successful contract proposal for the DARPA Low Power Electronics Program. This industry–university–SEMATECH joint program significantly advanced silicon-on-insulator materials, devices, and design techniques for low-power, high-performance microelectronics. While in this position, Dr. Theis also coordinated the transfer of copper interconnection technology from IBM Research to the IBM Microelectronics Division. The replacement of aluminum chip wiring by copper was an industry first and involved close collaboration among research, product development, and manufacturing organizations. He assumed his current position as Director, Physical Sciences, in February 1998. Dr. Theis is a member of the IEEE, the Materials Research Society, and the Project Management Institute; he is a Fellow of the American Physical Society. He is the author of more than sixty scientific and technical publications.