Impact of Small Process Geometries on Microarchitectures in Systems on a Chip

DENNIS SYLVESTER, MEMBER, IEEE AND KURT KEUTZER, FELLOW, IEEE

Invited Paper

Process effects in deep-submicrometer geometries are expected to change the physical organization, or microarchitecture, of integrated circuits. The factor that is expected to primarily impact integrated circuit microarchitectures is increasing delays in interconnect. We believe that, to properly microarchitect integrated circuits in small process geometries, it is necessary to get as detailed a picture as possible of the effects and then to draw conclusions about changes in microarchitecture. To this end, in this paper, we describe a comprehensive approach to accurately characterizing the device and interconnect characteristics of present and future process generations. This approach uses a detailed extrapolation of future process technologies to obtain a realistic view of the future of circuit design. We then proceed to quantify the precise impact of interconnect, including dynamic delay due to noise, on the performance of high-end integrated circuit designs. Having determined this, we then reconsider the impact of future processes on integrated-circuit design methodology. We determine that local interconnect effects can be managed through a deep-submicrometer design hierarchy that uses 50K–100K gate modules as primitive building blocks.

In light of this new system-on-a-chip microarchitecture, we then examine global interconnect issues. Our results indicate that, while global communication speeds will necessarily be lower than local clock speeds, International Technology Roadmap for Semiconductors expectations should be attainable to the 0.05-μm technology generation. Achieving these high clock speeds (10 GHz local clock) will be aided by the use of a newly proposed routing hierarchy that limits interconnect effects at each level of a design (local, isochronous, and global). In addition, key components of the interconnect architecture of the future include fat (or unscaled) global wires, intelligent repeater and shield wire insertion, and efficient packaging technologies.

Keywords**—***Integrated circuit interconnections, integrated circuit modeling, routing, ultralarge-scale integration.*

I. Introduction

Process effects in deep-submicrometer geometries are expected to change the physical organization, or microarchitecture of integrated circuits. The factor that is expected to primarily impact integrated circuit microarchitectures is increasing delays in interconnect. A number of independent sources have made forecasts that in deep-submicrometer (DSM) process geometries 80% or more of the delay of critical paths will be attributable to interconnect [1]–[3]. This forecast has been further supported by the broad industrial experience of significant problems in timing closure for current high-performance integrated circuit (IC) designs. Based on the present difficulties in IC design methodologies and predictions of increasing migration of delay to interconnect, there is a sense in both industry and academia that current synthesis and physical design methodologies require a significant overhaul.

Deep-submicrometer effects, particularly interconnect, have thus been billed as potential showstoppers to the continuation of Moore's law. Among the effects that are commonly mentioned are rising *RC* delay of on-chip wiring, noise considerations such as crosstalk and dynamic delay, reliability concerns due to rising current densities and oxide electric fields, and increasing power dissipation. Each of these issues has underlying physical explanations that shed insight into its potential impact as CMOS processes continue to scale. We address these issues and others in Section III.

We believe that to properly microarchitect integrate circuits in small process geometries, it is necessary to get as detailed a picture as possible of the effects and then to draw conclusions about changes in microarchitecture. Our goal in this paper is to propose a specific microarchitecture that limits the impact of interconnect on performance and allows device scaling to continue to drive IC performance. The design methodology we suggest is based on realistic projections of future technologies, simulation tools, analytical models, and actual design data. Furthermore, we describe several concepts that may help interconnect scale more readily at the

Manuscript received May 31, 2000; revised January 12, 2001.

D. Sylvester is with the University of Michigan, Ann Arbor, MI 48109-2122 USA.

K. Keutzer is with the University of California at Berkeley, Berkeley, CA 94720-1770 USA.

Publisher Item Identifier S 0018-9219(01)03199-1.

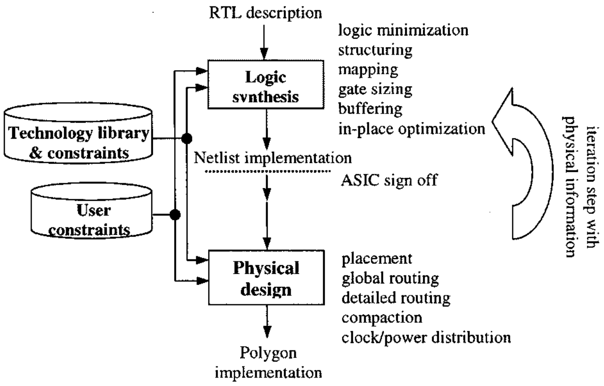


Fig. 1. Traditional ASIC design flow.

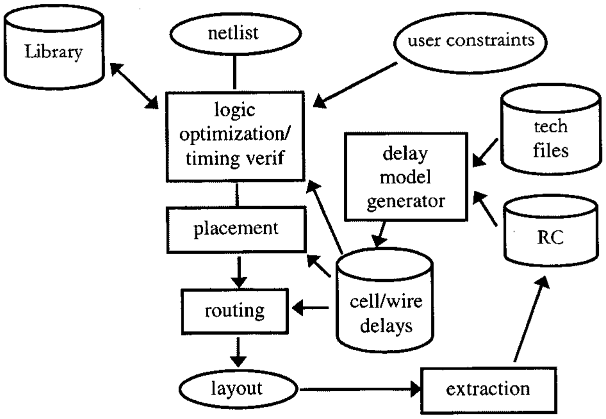


Fig. 2. Today's high-performance logical/physical flow.

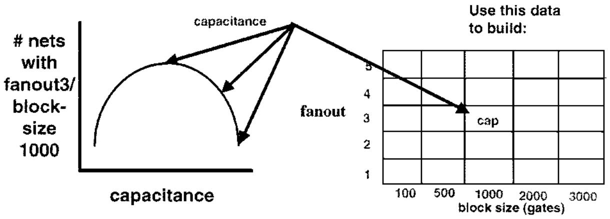


Fig. 3. General concept of wireload models.

global level in future high-speed systems. In this paper, we primarily focus on high-performance application-specific integrated circuits (ASIC) but use leading-edge microprocessors as a point of reference for global interconnect issues since the higher clock speeds of these designs make them susceptible to new interconnect phenomena (such as inductance and multiple clocking regions) sooner than ASICs.

Parts of this work form the basis for the Berkeley Advanced Chip Performance Calculator (BACPAC). This system-level performance model improves upon previous work [4], [5] by incorporating enhanced analytical models for delay, noise, power, and area. It also addresses the rising system-on-a-chip (SoC) hierarchy of future ASICs and microprocessors. The model and extensive documentation are available at http://www- device.eecs.berkeley.edu/~dennis/BACPAC. The models in BACPAC have been subsequently used in creating GTX [6], which adds significant flexibility to existing system performance models. GTX is part of the Gigascale Silicon Research Center and is available at http://vlsicad.cs.ucla.edu/GSRC/GTX.

II. Current Methodology

Integrated circuit microarchitectures can only be implemented in underlying process geometries by means of computer-aided design (CAD) tools, and microarchitectural possibilities are always limited by implementation feasibility. Therefore, we begin by briefly examining CAD tool flows.

A. Traditional ASIC Flow

Fig. 1 gives the implementation portion of a design flow used in traditional ASIC designs. In this flow, the design is described in a hardware-description language (HDL) such as VHDL or Verilog. The *technology library* is the database containing the data that models the predesigned cells in the underlying process technology for the logic synthesis and physical design tools. *User constraints* are the mechanism by which the user conveys constraints regarding the speed, area, and power of the design. Logic synthesis transforms the HDL description into a graph in which each vertex represents a cell in the technology library, and each edge represents a wired connection between the cells. This graph is called a *netlist*. Logic synthesis optimizes the circuit according to user constraints and ensures that design rules are met. A survey of logic synthesis can be found in [7].

Physical design is the process by which the synthesized *netlist* is transformed into a mask, which is then used to fabricate a three-dimensional integrated circuit. Information contained in the *technology library* and *user constraints* ensures that the output of physical design can be fabricated in the designated semiconductor process. User constraints restrict the location of pads or signals, the area resources available for implementation, and the timing behavior. An excellent reference for physical design operations is [8]. In the ASIC flow described in Fig. 1, the first half of the flow is the responsibility of the design center while the second half is the responsibility of the ASIC or semiconductor vendor.

B. Problems with Current Flow

The design flow in Fig. 1 contains a separation between the logic synthesis and physical design steps. For designs with aggressive performance goals, several iterations between synthesis and physical design are required to converge to a desired implementation. As a result, design teams have begun to perform more of the physical design steps themselves and the handoff to the semiconductor-vendor occurs only at the end. This approach is shown in Fig. 2 and is known as the *customer-owned tooling* (COT) approach. Note that the differences between the flow in Fig. 1 and that of Fig. 2 are more organizational than technical.

C. Problems with Logic Synthesis

During synthesis the capacitances and delays associated with final wiring are unknown. Models of interconnect, known as *wire-load models*, attempt to predict the amount of capacitance in a wire by reducing it to a function of fanout and block size. In this approach, shown in Fig. 3, a single capacitance value is used for *all* nets in a block with the same fanout. Clearly the capacitance values for nets in a block will vary and so the wire-load model is necessarily only an approximation of the actual future capacitance. When wire delay constitutes a small percentage of the total critical path delay, errors in capacitance estimation have little impact. However, if the capacitance values in wires increase, then wire-load models lead to increasing inaccuracy.

D. Problems with Physical Design

Placement and routing tools tend to use primitive models of circuit delay. Often the timing constraints inherent in the synthesized netlist are translated into a single static net prioritization. Place and route tools attempt to honor this prioritization scheme, but because they have only a primitive internal model of critical path delay, they cannot be sure that they have in fact obeyed the initial timing constraints. As delay migrates to interconnect, the possibility that excess capacitance in routing violates a timing constraint increases. If capacitance is increased due to cross-coupled capacitances in routing, this is only likely to be discovered after the final layout is extracted.

To summarize, the core problem with current design flows that is causing them to fail is the poor back-end prediction capabilities of logic synthesis (i.e., the wire-load models). To attack this problem at the microarchitectural level, we propose to use a block-based hierarchical design style that uses block sizes of 50K–100K gates. Within these blocks (or modules), we will demonstrate that interconnect effects can be effectively minimized via proper device sizing and intelligent scaling so that even relatively inaccurate wire-load models can be used at this level of hierarchy. We note now, and emphasize later, that this module size allows for a significant amount of flexibility and computing power.

III. Device and Interconnect in DSM

We propose that the size of a module that can still be reliably designed in the methodology of Figs. 1 and 2 without significant attention to DSM effects is approximately 50K–100K gates. Fig. 4 shows the approach we employ to arrive at this conclusion. To begin, we develop a strawman technology file that completely describes future process generations for both device and interconnect characteristics.

To facilitate our analysis, we develop models of future designs, at varying degrees of accuracy, that allow us to determine the performance of future ASICs. These models emphasize the important issues of delay, noise, and power. Before introducing our strawman technology, we will first give a high-level overview of device and interconnect issues in deep submicrometer.

A. MOSFETs

1) Voltage Scaling: While 5-V power supplies were prevalent at feature sizes >0.5 μm, DSM processes are seeing a continual drop in V_{DD}. One key reason for reducing voltages in scaled CMOS devices is that of reliability. For instance, gate oxides tend to break down if exposed to electric fields in excess of 5 to 6 MV/cm [9]. Since oxide thickness ($T_{\rm ox}$) is being scaled to increase current drive, the maximum voltage on the gate (V_{DD}) must also be scaled to keep electric fields within reason. Another form of device reliability that must be considered is the effect of hot electrons. When the electric field in the channel is too high near the drain, electrons may gain enough energy to inject themselves into the gate oxide. This accumulation of charge in the oxide results in shifts in V_t and changes in the device's current–voltage (I–V) characteristics. By reducing the voltage supply, both the channel and gate electric fields will be proportionately reduced for improved reliability.

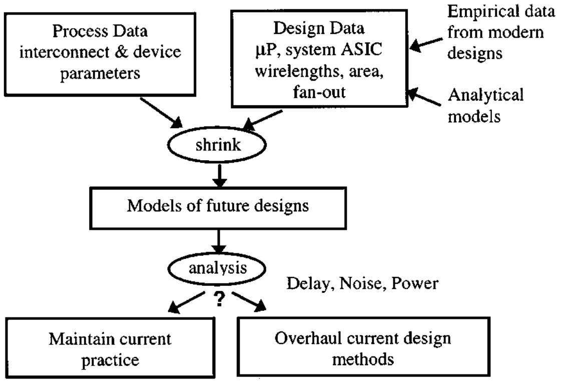


Fig. 4. Overview of our analysis methodology.

The impact of reduced supply voltages on power is another important reason why V_{DD} values are decreasing. Dynamic power consumption is defined by $P_{\rm dyn} = \alpha C V_{DD}^2 f$, where α is a switching activity ratio, C is the switched capacitance, and f is the operating frequency. The quadratic dependency on supply voltage makes the reduction of V_{DD} a primary goal in power minimization. This issue will be discussed more in Section VI-C.

2) Drive Current: Classical long-channel MOS theory states that device current in the saturation mode of operation is proportional to the square of gate drive ($V_{DD} - V_t$) and inversely proportional to channel length. At ultrashort channel lengths, most carriers travel at a maximum saturated velocity $v_{\rm sat}$ throughout the channel, which nearly eliminates the impact of channel length on current. A new and more accurate expression for this limiting case is [9] $I_{\rm dsat} = W v_{\rm sat} C_{\rm ox}(V_{DD} - V_t)$. From this expression, we can see the channel length independence as well as the move from quadratic to linear dependence on gate drive. Since $v_{\rm sat}$ is a material constant (approximately 8×10^6 cm/s for electrons, 6.5×10^6 cm/s for holes), we find that $I_{\rm dsat}$ (normalized to device width) will vary with $(V_{DD} - V_t)/T_{\rm ox}$. With V_t fixed at $V_{DD}/4$ (necessary to maintain sufficient gate drive), we obtain $I_{\rm dsat} \propto V_{DD}/T_{\rm ox}$. Since both of these parameters are decreasing, DSM MOSFETs are not expected to provide enhanced drive current per unit width. This marks a new regime for scaled CMOS devices, since up to the 0.35-μm process generation additional current drive was obtained with each process shrink. Data taken from published reports confirms this analysis of saturation current in DSM CMOS devices [10]–[12]. Our own survey of 16 reported technologies shows both n- and p-channel devices achieve only slight increases in drive current from 0.25 to 0.09 μm L_{drawn}.

B. Interconnect

Deep-submicrometer interconnect effects present a variety of problems to process engineers, circuit designers, and CAD tool developers. This section presents an overview of these effects.

1) RC Delay: The most commonly cited DSM interconnect problem is that of rising *RC* wire delays. For instance, the *RC* delay of a 1-mm metal 1 line in 0.5-μm technologies was 15 ps while in 0.1-μm technology it is 340 ps (without new materials). It can be clearly seen that wiring delay is capable of consuming the majority of the shrinking clock cycle time in DSM designs. We now look briefly at the reasons behind the rapid increase in *RC* delay and possible methods of slowing this trend.

Increasing line resistance is the main reason behind the increased wiring delay in DSM. Resistance is inversely proportional to the cross-sectional area of the wire. Due to the rising need for higher densities on-chip, wiring pitches are dropping rapidly at about the same rate as gate length. In an effort to keep resistance from increasing too quickly, many processes are scaling line thickness (or height) at a slower rate, which results in taller, thinner wires. For instance, [13] predicted an increase in wiring aspect ratio (AR $=$ height/width) from 1.8 at 0.25 μm to 3.0 at 0.05 μm.[1]

Besides the use of high AR lines, the other approach to reducing resistance is the use of better conductors for on-chip interconnect. Until recently, aluminum wires were used exclusively in back-end processes. Recent literature [10], [11] demonstrates the use of copper in sub-0.25-μm processes. The resistivity of copper interconnect is 30%–40% lower than that obtained using aluminum wiring (approximately 2.2 $\mu\Omega$·cm versus 3.3 $\mu\Omega$·cm). Note that the bulk resistivities of these materials are not achieved in production due to the use of cladding materials and reliability-enhancing impurities. Another advantage gained through the use of copper wiring is increased resistance to electromigration (EM) effects. Electromigration occurs in metals when a large current density is being driven through the line. Metal ions are physically moved down the wire by the large current, resulting in opens in the line or shorts to neighboring wires. Copper has a much lower susceptibility to ion transport from EM since it is a heavier metal. Results show copper to have an EM lifetime that is 100× longer than aluminum wiring at the same current density [10].

Wiring capacitance is also increasing, albeit slowly, in scaled processes due to the higher densities needed to route modern chips. For instance, line-to-line spacing and insulator thickness are both shrinking, resulting in an overall increase in line capacitance (this is true despite smaller linewidths). Since the reduction of packing density is not an option in DSM, the only way of reducing wiring capacitance is by using a low-k dielectric instead of SiO$_2$. SiO$_2$ has many natural advantages that make it the dielectric of choice in microelectronics. However, its relative dielectric constant of 3.9–4.1 leaves room for improvement by using advanced materials such as polyimides. Significant research is ongoing in the area of low-k process integration as preliminary work has suggested that gate delay and power can be reduced by 39% and 47%, respectively, in a 0.25-μm process with $\varepsilon = 3.1$ [14]. The ultimate goal in low-k dielectric process integration is the use of xerogels, which are highly porous materials with dielectric constants approaching that of air, $\varepsilon = 1$ [15].

2) Noise: As mentioned above, one of the methods to reduce resistance has been to slowly scale line thickness, resulting in taller, narrower wires. These high-aspect ratio lines have a detrimental side effect in that they result in a large amount of coupling capacitance. With AR > 1, lines tend to have more capacitance to neighboring wires than to upper and lower wiring layers, which effectively act as ground planes. In addition, spacing between wires is shrinking quickly in an attempt to maintain high packing densities, further increasing coupling capacitance. As evidence, line-to-line capacitance between wires on the same level makes up over 70% of the total wiring capacitance at lower levels even at 0.25-μm processes [16].

The impact of this rise in coupling capacitance can be seen in the form of noise. In this work, we will consider two distinct forms of noise in regard to DSM interconnect. The first is dynamic delay, which refers to the fact that total capacitance seen by a gate is no longer a constant value [17], [18]. Due to the rising contribution of coupling capacitance to total load capacitance, the Miller effect can have a large impact on actual delay times on a chip. The Miller effect states that, when both terminals of a capacitor are switched simultaneously, the effective capacitance between the terminals is modified. For instance, if wire A switches from 0 to V_{DD} while an adjacent wire B switches from V_{DD} to 0, the effective voltage swing between the two terminals is $2V_{DD}$. From $Q = CV$, the charge needed to switch wire A is now doubled with respect to the case where wire B is static. Alternatively, this is seen as a doubling of the "effective" capacitance. The increase in coupling capacitance is a potential timing hazard in that delay becomes a function of neighboring signal activity, making static timing analysis difficult.

The second form of noise we will discuss in this work is crosstalk [16]. In this scenario, a static wire (called the victim) is perturbed by switching activity on neighboring wires (aggressors). In the worst case, two aggressors switch in the same direction simultaneously, leading to an undesirable voltage spike on the victim line due to capacitive coupling. This sort of noise can cause false switching (especially in dynamic circuits), increased glitch power consumption, and voltage overshoot effects that may lead to enhanced device stress or forward-biasing of p-n junctions. Crosstalk is

[1] The most recent roadmap [1] has taken a more conservative stance regarding aspect ratios, more closely mimicking our projections.

Table 1
Projected MOSFET Characteristics in DSM

Process (µm)	T_{ox} [1] (Å)	V_{dd} (V)	V_t (V)
0.25	45-50 [40-50]	2.5	0.625
0.18	30-40 [19-25]	1.8	0.450
0.13	25-30 [15-19]	1.5	0.375
0.10	20-25 [10-15]	1.2	0.3
0.07	15-20 [8-12]	0.9	0.225
0.05	12-15 [6-8]	0.7	0.175

highly sensitive to the ratio of coupling capacitance to total capacitance, implying that it will become a larger issue as interconnect dimensions continue to scale.

IV. Future Process Technologies

Aiming to quantify the concerns raised in the previous section, we now present substantiated projections of DSM technology parameters at both the device and interconnect level. As conclusions are based on models and data, we take pains to detail our models and data. Care is taken to explain and justify our data and models, and comparisons are drawn to predictions in [1], [13]. Finally, where choices are made, we aim to be consistently conservative and err on the side of pessimism.

A. Device Roadmap

Table 1 presents the most important features of our strawman technology for DSM CMOS devices. Our gate oxide thickness projections are generally higher than those found in the 1999 ITRS [1]. Due to oxide reliability requirements, one should keep the electric field in the oxide below 5–6 MV/cm [7]. In addition, oxides thinner than 15 Å raise serious concerns about leakage due to direct tunneling through such an extremely thin layer [19].[2] These concerns are not likely to be resolved by the 0.13/0.1-μm generations as anticipated in [1]. Therefore, we predict oxides that are free from tunneling problems until the 0.07-μm technology node. Finally, $T_{\rm ox}$ values below 10 Å are extremely optimistic; this thickness represents only a few atomic layers of SiO_2. We predict oxides to "bottom out" around 10–12 Å due to leakage and fabrication issues.

Regarding voltage scaling, we select the high-performance V_{DD} values from ITRS 99 [1], rather than the low-power scenario. Also, V_t is set at $V_{DD}/4$ to provide sufficient current drive to keep performance climbing [7]. Beyond 0.1 μm, such small V_ts may yield high amounts of leakage current. In this case, several approaches may be taken. A dual-V_t process uses two different threshold voltages within the same design. Low V_t devices are used where speed is essential while the bulk of devices (e.g., 90%) have a higher V_t to keep overall leakage current small [20]. A second approach would be to use circuit techniques to raise the V_t in idle circuit blocks. This approach is similar in concept to standby or sleep modes in current microprocessors where the clock is disabled in low-activity regions in order to reduce dynamic power consumption [21]. However, sleep modes do not eliminate static power and further work will be needed in this area to minimize latency and ensure reliability.

Table 2
Interconnect Characteristics for Metals 1 and 2

Process (µm)	**0.25**	**0.18**	**0.13**	**0.1**	**0.07**	**0.05**
Thickness (µm)	0.5	0.46	0.34	0.26	0.2	0.14
Width / Space (µm)	0.3	0.23	0.17	0.13	0.1	0.07
Sheet Resistance (Ω/•)	0.044	0.048	0.065	0.085	0.11	0.16
T_{ins} (µm)	0.65	0.5	0.36	0.32	0.27	0.21
Dielectric Constant	3.3	2.7	2.3	2.0	1.8	1.5

Table 3
Top Metal Layer Parameters for All Generations

Thickness (µm)	Width/Space (µm)	Sheet Resistance (Ω/•)	T_{ins} (µm)
2.5	2.0	0.009	1.4

Combined with the characteristics described in Table 1, BSIM3 models have been developed to model DSM devices [22]. Default parameters are used except for $T_{\rm ox}$, V_t, mobility, and capacitances. Resulting I–V curves yield good fit (within 20% in linear region, 10% in saturation) with measured results from 0.15-μm devices ($T_{\rm ox} = 40$ Å). Simulated values of $I_{\rm dsat}$ for 0.25 to 0.1-μm processes show excellent correlation with published data.

B. Interconnect Roadmap

Tables 2 and 3 highlight key parameters from our interconnect roadmap. Table 2 presents dimensions for the lower two levels of metal for each process. Our interconnect hierarchy consists of several pairs of identical metal layers that fall under the categories of local, intermediate, and global wiring. The number of metal layers increases from 6 at 0.25 μm to 9 at 0.05 μm for enhanced connectivity. The first two levels of metal are used exclusively for routing local signals between gates within a larger block of gates (e.g., 50K). In these instances, the wirelength is typically short and the first concern is that of wiring density. Therefore, we predict a continuing drop in lower-level wiring pitch. This will not only provide additional local routing capability but will also allow for smaller standard cell sizes as [23] and [24] have shown cell size to be set by contacted wiring pitch.

[2] Stacked gate dielectrics using high-k materials on top of a very thin SiO_2 layer will help in reducing tunneling since a thicker high-k material can be used while maintaining the same electrical characteristics as a thinner SiO_2 region. Difficulties are in manufacturability and minimizing the amount of SiO_2 needed at the interface.

The second concern at the local level is noise. Due to the shrinking pitches, larger coupling capacitances lead to enhanced noise. In order to limit noise, we recommend the use of "flat" wiring where the aspect ratio is capped at about 2 [16]. Compared to predictions in [13], an AR of 2 will yield 30% smaller coupling capacitance at 0.07 μm than AR $= 2.7$. The use of thinner wires in the "flat" approach can be seen as a tradeoff between noise and resistance, where the lower resistivity of copper is taken advantage of in order to limit capacitances. This approach has been used in early copper designs to limit capacitance and is especially beneficial at lower levels where device resistance tends to be much larger than wire resistance (due to short wirelengths) [20]. An additional point in noise reduction is the scaling of insulator thickness, T_{ins}. In order to prevent coupling capacitance from becoming an even larger portion of total wiring capacitance, T_{ins} needs to be scaled appropriately. Also, by reducing T_{ins} vias with reasonable ARs can be used, allowing for easier fabrication and lower resistance.

Copper is used for all sheet resistance calculations with a resistance of 2.2 $\mu\Omega$·cm. The reduction of dielectric constant with scaling reflects the significant work being done on low-k materials to replace SiO_2 as the insulator of choice in ULSI [25]. The implementation of low-k dielectrics into processes represents a significant step in realizing very high performance designs. Smaller interconnect capacitances result in lower power dissipation as well as smaller delay times, making low-k dielectrics a more beneficial process advance than copper wiring. Our projections for low-k dielectrics are fairly conservative in comparison to [1], [13], especially beyond 0.13 μm.

Table 3 shows projections for top-level interconnect throughout scaling. In contrast to lower level metals, where average wirelengths scale down due to shrinking gate sizes, global wires actually become longer. For this reason, we select a large cross-section global wire that maintains a constant resistance of 44 Ω/cm. This low resistance value will allow for unattenuated distribution of power grids, clocks, global buses, and other important signals on the top layers of metal. The approach follows the concept of "fat" wires suggested in [5]. Transmission line characteristics will need to be well-modeled and controlled in this scheme [26].

Intermediate metal layers not included in Tables 2 and 3 provide inter-modular routing and offer pitches and thicknesses between those presented. For instance, metals 3 and 4 have a minimum pitch that is 50%–100% larger than that of metals 1 and 2.

V. Design Data

To supplement the analytical models and simulation tools to be used in the analysis section, we obtained detailed design data from modern ASICs. Empirical design data was compiled for thirteen 0.35-μm ASIC designs from Symbios Corporation. Average wirelengths in the designs vary somewhat but tend to be in the range of 200 to 300 μm. Average fan-out is consistently between 2.2 and 3. Designs regularly incorporated large macro blocks in addition to standard cells, effectively making them early SoC designs. The percentage of standard-cell logic area to total chip area ranged from 30% to 95% but logic area typically encompassed about 70% of the chip. Data was also obtained showing average wirelength for each fan-out in each design. This data is used in Section VI to create a critical path model for future ASICs.

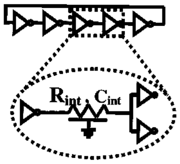


Fig. 5. Determination of gate/interconnect delay involves 2-input NAND ring oscillators and average wirelengths.

VI. Analysis

In this section, we develop models for various performance metrics (delay, noise, power) and apply them to DSM ASICs to confirm our position that 50K–100K gate modules can be designed using the flow in Figs. 1 and 2.

A. Delay

1) Defining Gate/Interconnect Delay: We begin our analysis by proposing a well-defined method of assessing gate delay versus interconnect delay. Typically for a given process, gate delay (t_{gate}) is determined using an unloaded (FO $= 1$, no significant wiring load) ring oscillator made up of inverters. This is an elegant way to determine t_{gate} as it is independent of device sizing since increasing device width contributes equally to larger drive current and load capacitance.

We propose a modified version of the ring oscillator concept to determine gate and interconnect delay. First, 2-input NAND gates replace inverters since they better represent on-chip logic gates. One of the inputs in these gates is tied to V_{DD}, resulting in worst case low-to-high delays. Next, the fan-out of the gates is varied from 1 to 4 (fan-outs greater than 4 are not of practical interest). At this point, gate delay is found for each fan-out individually. Finally, minimum-pitch interconnect of length L_{avg} is added between each stage. Average wirelength is a function of fan-out and is varied accordingly, using empirical design data as a reference. At this point, a stage delay (t_{stage}) is found for a given technology with fan-out ranging from 1 to 4. Interconnect delay, t_{wire}, is defined as the difference between the stage delay and intrinsic gate delay. The basic approach is shown schematically in Fig. 5.

Sizing of gates with interconnect loading becomes nontrivial as extremely large devices could be used to make t_{wire} negligible. Likewise, the use of minimum-sized gates would yield a misleading depiction of interconnect delay. Since this is not practical due to area and power considerations, we define an optimal driver size. The optimal size is defined by a W/L ratio such that an increase in W/L of 1 does not yield a 2% drop in stage delay. This criterion was chosen to most closely approximate the knee of the delay versus device sizing plot obtained when sweeping W/L.

2) Analytical Approach: In this section, we discuss the first of two different approaches to the delay analysis. A first-

Table 4
First-Order Analysis Results for Delay Scaling. 0.05-μm Value in ( ) is Adjusted to Match V_{DD} with [27]

Process (μm)	C_{gate} FO=2	C_{wire}	V_{DD}/I_{dsat}	Frequency	Frequency [27]
0.25	1	1	1	1	1
0.18	0.56	0.626	1.03	1.61	1.61
0.13	0.38	0.373	1.22	2.18	2.23
0.1	0.25	0.225	1.4	3.07	3
0.07	0.153	0.141	1.5	4.61	4.31
0.05	0.104	0.082	1.67	6.73 (5.77)	5.75

order analytical model is presented to observe general delay trends in future ASICs. A full-scale simulation approach that employs accurate device models and distributed interconnect effects is described in the following section.

To the first order, delay of a MOSFET is governed by a simple relationship between capacitance, voltage, and current: T_d = CV/2I. To study the scaling of delay through process generations, we employ this expression by approximating the trends of each of the variables involved. Several approximations are made and will be discussed in turn.

We begin by discussing a block-based design approach with 50K gate building blocks. We later discuss other block sizes. In a block of this size, wirelengths are typically small and line resistance is much less than the effective resistance of the MOSFET drivers. Therefore, we ignore the impact of wire resistance in these calculations. Consider a single gate with a fixed fan-out in two processes, with a scaling factor of S. By scaling both channel length and width by S, we get a quadratic reduction in module area assuming that wiring pitch is also reduced by S [23], [24]. We assume that average wirelength is a fixed fraction of the module side length so that L_{avg} scales by S. Since I_{dsat} is relatively constant in DSM processes, a reduction in channel width of S results in a proportional reduction in I_{dsat}.

Voltage swing, as discussed above, is also shrinking due to power and reliability concerns. From Table 1, we estimate the scaling factor for V_{DD} as 0.75. Finally, we assume that the load capacitance is made up of two components, interconnect and fan-out gate capacitance. Interconnect capacitance, C_{wire}, is found by multiplying the average wirelength by the capacitance per unit length. From 2-D simulations that incorporate low-k dielectrics (Table 2), we find the capacitance per unit length to be dropping by about 15% per generation, or a scaling factor of 0.85. Gate capacitance can be evaluated by estimating that T_{ox} is shrinking by 0.8× per generation (from Table 1). We set interconnect capacitance to be twice as large as the fan-out capacitance based on 0.25-μm calculations (FO = 2). Note that we are neglecting device junction and overlap capacitances in this analysis.

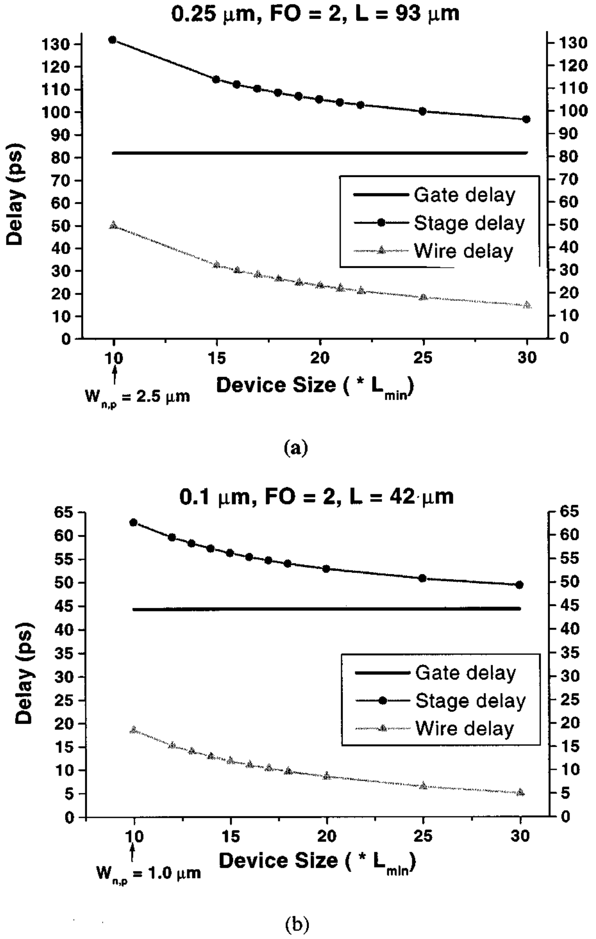
Normalizing CV/2I to 1 for the original process, the shrunken process yields a CV/2I value of 0.65. This represents a stage delay improvement of 35% from one generation to the next despite the smaller current drive. Translating to clock frequencies, these results predict a 54% increase for a generic process shrink. This analysis brings up several important points. First, it seems possible to maintain rising performance levels while scaling channel widths even in the era of velocity saturation. Second, the voltage and current components of CV/2I cancel each other to some extent, leaving the bulk of the delay improvement to reduction of load capacitance. This is important because it highlights the most vital issues in DSM. In order to keep on pace with performance projections, wirelengths need to be reduced, low-k dielectrics need to be introduced, and device capacitances need to drop. These are among the most important factors contributing to faster designs.

Table 4 shows detailed results from this first-order analysis using actual scaling numbers taken from our strawman technology. Gate capacitance is determined using $W/L = 20$ and a fan-out of 2. As mentioned in the previous discussion, the sharp decrease in load capacitance ($C_{\text{gate}} + C_{\text{wire}}$) compensates for a slow rise in V_{DD}/I_{dsat} to yield overall speed improvements. A comparison is made to gate delay predictions in [27], which uses 3-input NANDs and several empirical factors to account for device resistance and the use of dynamic logic. Our simple analysis gives very similar results regarding the scaling trends of gate delay in DSM.

3) Simulation Approach: The previous approach is useful in determining trends and important parameters in delay scaling. However, it makes several assumptions that we will now remove. For example, the analysis ignored the impact of global wires in clock cycle determination. Since designs are getting larger, global wires become longer, which means that they will necessarily take up an increasing portion of the clock cycle. In addition, wiring resistance was ignored due to the assumption that device resistance dominated. Also, only the gate oxide component of device capacitance was treated. In this section, we will look more closely at the factors involved in determining ASIC performance using simulation tools.

We begin by characterizing a small ASIC library for each process (0.25 through 0.1 μm) consisting of 2-input NANDs with varying fan-outs. As mentioned earlier, BSIM3 device models are used in simulations. According to the procedure outlined in Section VI-A1, we determine t_{gate}, optimal device sizing ($W_n = W_p$ in 2-input NANDs), and t_{wire}. After finding optimal device sizes at the 0.25-μm generation, we use this W/L ratio for all subsequent processes to allow for easier comparisons.

Fig. 6 shows plots resulting from device sizing optimization for 0.25 and 0.1 μm with a fan-out of 2. While the gate delay is constant throughout, the total stage delay decreases appreciably when increasing device sizes. In the limit, infinitely large devices will be unaffected by the presence of a relatively short wire, yielding a stage delay essentially equal to t_{gate}. We find that the optimal device size is around W/L = 16. At this size, interconnect represents 28% and 22% of the total delay at 0.25 and 0.1 μm, respectively. The fact that interconnect delay is actually decreasing is somewhat surprising. The primary reasons behind this conclusion are the presence of shorter average wires and new materials.



(a)

(b)

Fig. 6. (a) Device size determination for 0.25 μm, FO $= 2$. (b) Same plot for 0.1 μm illustrates drop in interconnect delay.

Previous studies reporting rises in interconnect delay have tended to focus on a fixed line length. Due to shrinking gate pitches, local wirelengths are expected to shrink with process scaling. As demonstrated in Table 4, we forecast a decrease in average wiring capacitance by a factor of 12 from 0.25 to 0.05 μm due to shorter lines and low-k dielectrics. This point underlies our conclusion that interconnect delay will not dominate within 50K gate modules of future designs, and that a block-based microarchitecture is well suited to DSM design methodologies.

However, the scaling down of wirelength does not hold at the global level. In contrast, these signals must necessarily get longer in order to ensure connectivity in larger chips. Due to the use of functional clustering (keeping modules that need to communicate often near one another), extremely long global wires ($L \geq$ chip edge length) can be minimized and excluded from critical paths. Recent study has suggested that typical global wires might be closer to half the chip edge length [28]. We use this as an approximate starting point for a global wire in 0.25 μm ($L = 1$ cm) and scale it up by 15% for each process shrink. Simulation results show that at a constant buffer width (fixed I_{dsat}), delay decreases from process to process mainly due to the drop in voltage swing. Low-k dielectrics cancel out the 15% expected increase in wirelength so that delay varies roughly with V_{DD} according to CV/2I. Of course, this requires the use of larger drivers in each process (in terms of W/L), which can lead to power penalties. The available power and area budget needs to be weighted against the buffer requirements to meet timing constraints.

A comparison between the approaches in Sections VI-A2 and VI-A3 is deferred until the impact of noise on delay can be considered. We will also extend our model from the delay of a single stage to the delay of an entire critical path.

B. Noise

The impact of noise on system performance was described qualitatively in Section III-B2. In this part of the analysis, we modify the results of Section VI-A3 to include dynamic delay effects and also discuss the impact of crosstalk noise on DSM designs. The general result of dynamic delay is increased stage delays due to higher effective capacitance and larger power dissipation due to bigger drivers. Also, the portion of total delay attributable to interconnect is increased. Signal integrity is viewed as a reliability problem. It places limitations on the routability of signals on lower level, minimum-pitch wiring. Increasing wiring pitch to accommodate crosstalk reduces layout densities, which is a problem in wire-limited designs.

1) Critical Path: We begin by creating a generic critical path model that will allow us to track ASIC clock frequencies through scaling. From empirical design data, we have determined that in 0.35-μm ASIC designs a critical path typically consists of about 14 stages. While we use a 2-input NAND as the gate type for each stage, we allow for different fan-out conditions, which are determined from fan-out distributions of the design data. Our model is broken down into blocks of 8, 3, 2, and 1 stages with fan-outs of 1, 2, 3, and 4, respectively. In addition, we include a global wire routed on the top metal layer, which is buffered to reduce delay. Dynamic delay effects are not considered on the global wire and a fixed buffer size is assumed throughout scaling.[3] Finally, very conservatively, 10% timing overhead is allotted for the critical path due to clock skew, process variation (both device and interconnect), and other unmodeled phenomenon. This modeled critical path closely reflects the characteristics of a typical path in a design.

2) Results: The noise analysis uses worst case neighboring wire switching activity to determine new delay values. We model the worst case by having two adjacent wires simultaneously switch in the opposite direction as the victim line. Since there is no wiring component to t_{gate}, the intrinsic gate delay will remain the same. However, as anticipated the larger effective coupling capacitance due to Miller effect results in larger t_{wire} values. Results are shown in Fig. 7, which illustrates the relationship between gate delay and stage delay for various processes. Clearly the presence of dynamic delay increases the portion of total delay attributable to interconnect. However, we still foresee a drop in the ratio of interconnect delay to total delay even considering noise effects. For example, considering noise effects t_{wire} comprises 35% of the total delay at 0.25 μm

[3] Scaling of the global buffer sizes does not change the results as significantly as one may think. This is due to the fact that the delay versus gate size plot becomes relatively flat for large drivers (as in Fig. 6).

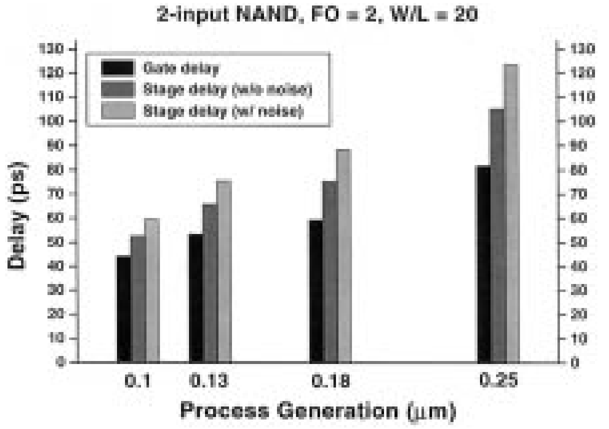


Fig. 7. Evolution of stage delay relative to a fixed gate delay with and without noise considerations.

Table 5
ASIC Performance Predictions Including Dynamic Delay Effects

Process (μm)	Logic Delay (ps)	Buffer Delay (ps)	Clock Skew/ Process Variation (ps)	Frequency (MHz)	ITRS Frequency (MHz)
0.25	1574	225	181	505	300
0.18	1274	150	128	713	500
0.13	952	120	110	846	700
0.10	752	110	89	1051	900

but only 27% at 0.1 μm. These values are far from the 80% forecasts commonly reported and reflect the impact of shrinking average wirelengths on a chip.

In general, worst case dynamic delay yields approximately an 80% increase in t_{wire}. This number corresponds very closely with the contribution of coupling capacitance to total line capacitance, which is about 75% throughout scaling. To compensate for the larger effective capacitance the optimal driver size is increased. These larger devices contribute to enhanced power dissipation and may also reduce layout density if drivers are made large enough. For a fixed optimal device size of $W/L = 20$, we incorporate the new delay numbers into the critical path model to determine expected ASIC clock frequencies in DSM. Global wire delay is taken from simulations with a two-stage buffering system and a fixed stage width of 100 μm. Results are presented in Table 5 and demonstrate similar trends as [1], [13] with a 100–150-MHz performance increase. Isolating the logic delay component of the critical path, we compare its evolution normalized to 0.25 μm with results from Table 4. We find that the inverse of logic delay scales as 1.5, 1.94, and 2.63 here compared to the analytical results of 1.61, 2.18, and 3.07. The discrepancies are due mainly to the inclusion of junction and overlap capacitances as well as dynamic delay effects.

3) Crosstalk: Crosstalk noise, or signal integrity, can be considered a reliability issue. Problems caused by crosstalk include functional errors due to false switching and enhanced device stress when bootstrapping occurs ($V_{ds} > V_{DD}$). Crosstalk is of the utmost concern when it can result in functional errors, such as in dynamic logic families such as domino [29] or in analog circuits. As a reliability problem, the most straightforward way to deal with signal integrity is the generation of accurate design rules. For instance, bounds may be set on the amount of tolerable crosstalk noise (e.g., 20% of V_{DD}). From this constraint, analytical models [30] can be used to define a critical line length for different metal layers and driver scenarios. This parameter, L_{noise}, is then compared to L_{delay}, which is defined as the maximum line length that can be used on a given metal layer before buffering becomes beneficial [31]. We have found that for typical driver conditions, crosstalk is a more severe restriction on routing in lower level metals than delay (i.e., $L_{\mathrm{noise}} < L_{\mathrm{delay}}$). In addition, L_{noise} is also typically smaller than the dimensions of a 50K gate module within which it is desirable to connect the majority of gates with metals 1 and 2. These findings indicate that noise can be a limiting factor in routing at the lower metal layers, which may lead to a loss in routing density due to possible increases in pitch, shielding wires, or the need to route in higher layers.

C. Power

1) Importance of Power: Low-power designs, especially microprocessors, have received a large amount of attention recently as portable and wireless applications gain marketshare. Also, even in the highest performance designs power has become an issue since the extremely high frequencies being attained (now exceeding 1 GHz) can easily lead to power dissipation in the many tens of watts. Dissipation of this amount of power requires heat sinks, resulting in higher costs and potential reliability problems. In this section, we discuss the reasons why power has become a significant issue and describe the three types of power consumption and how they can be expected to scale with CMOS processes.

In high-performance ASICs, there are three main reasons why power dissipation is rising. First, the presence of larger numbers of devices and wires integrated on a larger chip results in an overall increase in the total capacitance found on a design. Second, the drive for higher performance leads to increasing clock frequencies and dynamic power is directly proportional to the rate of charging capacitances (in other words, the clock frequency). Finally, the use of scaled voltages to improve reliability, decrease delay, and ironically, drop power consumption, leads to an increase in leakage current. While decreasing supply voltages does result in significant power savings overall, the static, or standby, current increases, which can be detrimental to low-activity designs that require very little standby power consumption. An excellent overview of power issues in large-scale CMOS circuits is given in [32] along with a discussion of power modeling techniques in [33].

2) Dynamic Power: Dynamic power consumption occurs as a result of charging capacitive loads at the output of gates. These capacitive loads are in the form of wiring capacitance, junction capacitance, and the input (gate) capacitance of fan-out gates. The expression for dynamic

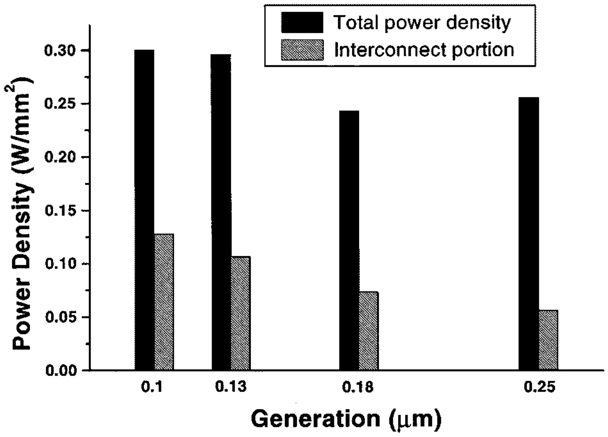


Fig. 8. Evolution of dynamic power density with scaling.

power was given previously and trends for several of the variables have already been discussed. Switching activity is a difficult parameter to estimate although it can frequently be approximated in the 0.1–0.2 range with reasonable accuracy.

In order to determine the impact of CMOS scaling on dynamic power consumption, we develop a simplified model of a 50K gate module that may exist in future SoCs. Given such information as packing density (devices/cm^2), wiring pitches, average device size (taken from Section VI-B2), and routing density (metal occupancy), we calculate the dynamic power density through process scaling [1], [13]. We do this by first estimating the size of a module, then calculating the interconnect and device components of the total capacitance. Fig. 8 summarizes the results and shows that dynamic power density is not increasing appreciably despite the rise in clock frequency. This analysis implies that dynamic power dissipation will increase approximately proportionally to the chip area. It should be noted that power dissipation in the clock network, off-chip drivers, and memory blocks are excluded from this analysis of a simple standard cell module. For larger block sizes (up to 200K gates), our analysis indicates that larger drivers are required, contributing to higher power densities. For instance, we expect up to a 50% rise in power density for a block-based design methodology using 200K gate modules for equivalent performance to 50K gates.

3) Static Power: The dominance of CMOS in modern circuit design is due in large part to its lack of static power consumption. This perceived benefit is becoming more suspect as voltages are scaled in order to limit dynamic power. Ideally, when the gate voltage of a MOSFET is below V_t there is negligible conduction. However, a small amount of leakage current flows at these conditions due to the inability of the gate to completely turn off the conducting channel. A good approximation of the amount of static current is given by ($T = 50\,°\mathrm{C}$) [9]:

$$I_{\text{static}} = 10\,\frac{\mu\mathrm{A}}{\mu\mathrm{m}} \cdot W \cdot 10^{(-V_t/95\,\mathrm{mV})}. \qquad (1)$$

It is seen that leakage current is an exponential function of threshold voltage. The need for scaling V_t to maintain current drive has been discussed and results in a marked rise in leakage current as we move into DSM. Leakage currents in the range of nA/μm become serious when considering the large integration levels in ULSI. Static power consumption is given by $P_{\text{static}} = I_{\text{static}} \cdot V_{DD}$. Table 6 calculates the leakage power density for a 50K gate module. There is a 2500× increase from 0.25 to 0.1 μm, demonstrating that leakage power is becoming a larger component of total power consumption that should not be ignored. The rapid rise in P_{static} calls for the use of multiple-V_t processes or novel circuit techniques to limit standby power consumption. Finally, the exponential relationship between static power and V_t is important since variation in V_t can be significant. Devices exhibiting V_ts at the lower tolerance limit of a process will exhibit considerably more leakage current than a nominal device. In the same manner, static current is strongly dependent on temperature, with high operating temperatures resulting in significantly worsened MOSFET subthreshold characteristics. Poor control of either V_t or operating temperature may lead to wild fluctuations in static power consumption.

Table 6
Scaling of Static Power Consumption Within a 50K Gate Module

Process	V_{DD}	V_t	W_{device}	I_{static}	Block area	P_{static}
(μm)	(V)	(V)	(μm)	(μA / block)	(mm^2)	(mW/mm^2)
0.25	2.5	0.625	5.75	3.03	2.5	0.003
0.18	1.8	0.450	4.1	151.8	1.43	0.191
0.13	1.5	0.375	3	677.3	0.83	1.22
0.1	1.2	0.3	2.3	3198	0.5	7.67

4) Short-Circuit Power: The final component of power dissipation is short-circuit power. Finite rise and fall times at the input of gates mean that both the pull-up and pull-down networks of a CMOS gate are conducting simultaneously for a short period of time. During this time, current is flowing between V_{DD} and ground, resulting in short-circuit power dissipation. Research in this area has demonstrated that well-designed circuits exhibit short-circuit power that is less than 20% of the dynamic component, with 5%–10% a more typical value [34]. Well-designed circuits strive to maintain reasonable input and output rise times so that short-circuit current cannot flow for an appreciable amount of time. In summary, short-circuit power dissipation is a manageable portion of the total power budget and can be approximated as 10% of the dynamic power consumption.

VII. Microarchitectural Implications

Logical hierarchy in an integrated circuit design is the hierarchy associated with the logical function of the circuit. Physical hierarchy is the hierarchy associated with the physical floorplan and layout of the circuit. In high-performance integrated circuit design, it is common for the logical hierarchy to be nearly identical with the physical hierarchy. This allows for large elements of a design to be designed independently both functionally and physically.

In response to our analysis above, we have proposed the use of a block-based hierarchical design style that limits physical block sizes to 50K–100K gates. The natural implication of this for IC microarchitecture is that the microarchitecture of the circuit be decomposed into blocks of 50K–100K gates. It is natural to ask: is this realistic? A review of the physical implementation of a recent high-end microprocessor (Alpha 21264) indicates that the floorplan naturally divides into approximately 25 blocks. Aside from memory elements, the modules, such as branch-prediction units, integer execution units, and reorder units, were each under 100K gates.

The largest ASICs manufactured today are entire systems-on-a-chip. These SoCs typically consist of a number of intellectual-property (IP) blocks connected by means of standard on-chip buses. The IP blocks can largely be broken into two categories: high-complexity blocks such as microprocessor cores or MPEG decompression units and lower complexity blocks such as peripheral units. We contend that most of even the high-complexity IP blocks will fit within 100K gates. For example, a 32-bit RISC microprocessor core contains approximately 50K gates (not including memory arrays), while a 24-bit DSP has 45K–50K gates and MPEG audio/video cores range from 60K–135K gates [35]. In other words, a design reuse methodology based on IP blocks is very suitable to our modular design style.

In summary, it appears that a hierarchical design approach in which designs are decomposed into 50K–100K gate blocks is suitable for the microarchitectures of both high-end microprocessors and high-end SoCs. In fact, it appears that such a methodology is already in use, at least for leading-edge designs. The next natural question is: how will these blocks be interconnected in the microarchitectures of future SoCs in small process geometries? To address this issue, we must look beyond the module boundaries and consider global interconnect issues.

VIII. Global Wiring Issues

To this point, we have presented an alternative analysis of DSM effects and analyzed how they are likely to impact future design methodologies. We proposed a new DSM design methodology based on the use of 50K–100K gate modules as primitive building blocks. Global wiring was included in the critical path delay analysis of Section VI by approximating a typical global wirelength (1 cm in 0.25 μm, scaled by 15%/generation) and optimally buffering it with respect to delay. However, a more detailed analysis is necessary since there are a host of *global* interconnect problems that were not sufficiently addressed there. To motivate this analysis, we begin by noting that semiconductor processing is advancing at such a rate as to enable the integration of hundreds and even thousands of 50K–100K gate blocks at sub-0.1-μm process geometries. Block-level placement and routing of thousands of such modules under global timing, power, noise, and area constraints will be the major design challenge of DSM (see Fig. 9).

The remainder of this work seeks to analyze and quantify the *global* impact of interconnect on future high-performance designs using the microarchitecture discussed above. Specifically, we shift our focus to microprocessors as these designs achieve the highest performance and will meet the limitations of global interconnect first. The main question to be answered is whether or not the new design methodology of this paper will create insurmountable problems in global routing. Other key questions concerning global interconnect to be examined include:

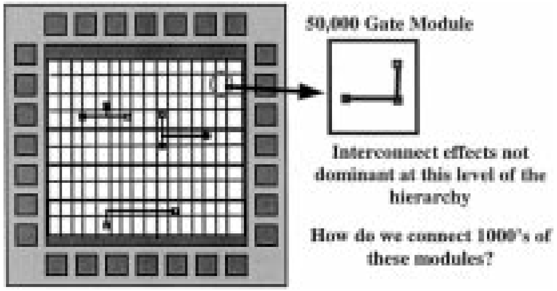

50,000 Gate Module
Interconnect effects not dominant at this level of the hierarchy
How do we connect 1000's of these modules?

Fig. 9. The design challenge in DSM is the assembly of thousands of 50K–100K gate modules considering chip-level interconnect effects.

- Over what sized region will we be able to propagate a high-speed signal ($>$1 GHz) across chip in a single clock cycle?
- How will this change in die sizes up to 750 mm^2?
- Can a high-speed clock be distributed reliably across a chip in light of increasing die sizes and process variation?
- Does noise at the global level pose a significant signal reliability concern?
- Will inductance result in severely degraded signal integrity?

Each of these topics will be discussed in detail, beginning with primary delay issues such as signal time-of-flight and scaling of global wires. Providing a comprehensive solution involves orchestration of a number of existing methodologies and technologies: unscaled global wires, flip-chip packaging, shield wires, etc. To demonstrate, a representative back-end process for 0.05-μm microprocessors is suggested.

IX. Primary Delay Issues

The foremost problem posed by long interconnect in DSM is that of the reverse scaling properties exhibited by wiring. This well-documented phenomenon implies that continual scaling (i.e., shrinking) of global interconnect, in conjunction with rising die sizes, will soon limit the attainable clock frequencies in a microprocessor. For instance, beginning with the 0.18-μm technology generation, [1] predicts a divergence of global and local clock frequencies due to the impact of global interconnect. In this section, we look at the concepts of signal time-of-flight and global conductor dimensions and the constraints they place on global communication.

A. Time-of-Flight

Larger die sizes and higher clock frequencies predicted in [1] imply that time-of-flight may become a limiting upper bound on speed. The time-of-flight (TOF) for a signal in a homogenous medium is given by

$$\mathbf{TOF = 33.33\sqrt{\varepsilon}\ ps/cm.} \qquad (2)$$

Here, ε is the dielectric constant of the medium ($\sim$4.0 for SiO_2). For example, a 750-mm^2 die, as predicted for the 0.05-μm technology generation [13], cannot support a global clock frequency greater than 5.48 GHz using Manhattan routing techniques.[4] This value is an upper bound since it uses air as a dielectric ($\varepsilon = 1$) and the entire clock cycle to traverse the longest potential path. A more realistic value would allot 80% of the clock cycle to this path and use a dielectric constant of $\sim$1.5, resulting in a maximum global clock speed of 3.58 GHz. This speed is near the projected global frequency of 3 GHz, but it is more interesting to look at the impact of TOF on locally clocked (isochronous) regions.

Due to TOF restrictions alone, an entire 750 mm^2 die is not reachable in a single clock period for frequencies in the range of 10 GHz (Fig. 10). Due to advances in processing, this clock speed will be realizable as approximately the delay through ten loaded stages in a 0.05-μm process. Assuming this divergence of clock speeds, TOF limitations on signal propagation do not play the major role in CMOS designs down to 0.05 μm; other effects such as *RC* delay, inductance, and reliable clock distribution will be more limiting factors. Fig. 10 reinforces this point by illustrating the relationship between TOF delays and die size. It is also interesting to note that the implementation of new low-k dielectric materials tends to offset the larger die size to provide a fairly constant maximum signal TOF for each technology generation. For instance, we expect a pathological corner-to-corner signal to have a time-of-flight of $\sim$220 ps regardless of technology node. Although this value does not increase due to the use of low-k materials, even a constant value will eventually render TOF issues important for global signaling. However, at 0.05 μm, this effect is still not dominant given a microarchitecture with multiple clock speeds.

B. Scaled Global Wiring

The current wiring paradigm calls for shrinking metal pitches at each generation in order to maintain sufficient routing densities. This is appropriate at the local level where wirelengths also decrease with scaling due to smaller gate sizes. In addition, this paradigm has worked at the global level before the DSM regime since *RC* delays of even global wires were insignificant compared to gate delays and clock cycle times. However, shrinking clock periods and transistor delays puts the current wiring scheme in danger. To examine more closely, a nominal global wiring pitch of 2 μm is used as a point of reference at 0.25 μm. For each successive technology shrink, the global wiring pitch is reduced by a factor of 0.75. This leads to a final wiring pitch at 0.05 μm of 0.48 μm. Aspect ratio is held at 1.5 for all technologies, which is in keeping with published reports by leading manufacturers [36].[5]

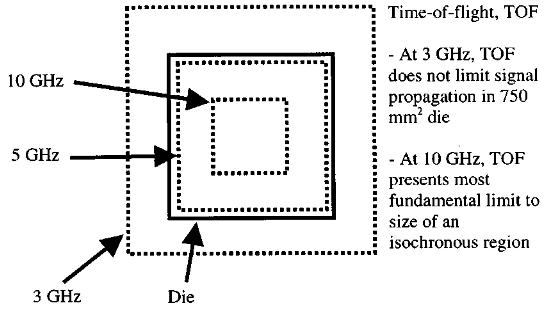

Time-of-flight, TOF
- At 3 GHz, TOF does not limit signal propagation in 750 mm² die
- At 10 GHz, TOF presents most fundamental limit to size of an isochronous region
10 GHz
5 GHz
3 GHz
Die

Fig. 10. Relationship between TOF delays and die size. The figure focuses on 50-nm microprocessors with $\varepsilon = 1.5$.

Table 7
Optimized Number of Repeaters for a Single Corner-to-Corner Wire in Each Process Technology

Technology Generation	# of repeaters (W_{min})	# of repeaters ($3 * W_{min}$)
0.25 μm	14	10
0.18 μm	22	14
0.13 μm	30	22
0.1 μm	50	36
0.07 μm	80	58
0.05 μm	138	98

With these wiring pitches, we examine critical line lengths and optimal buffer sizes. Critical line length is a concept based on the fact that there is a maximum wirelength for each metal level—wirelengths longer than this should be broken up using repeaters [31]. Since the pathological corner-to-corner wirelength will always be much longer than this critical line length (L_{crit}), we can use these two values to determine the number of repeaters needed to drive such a wire. Results are included in Table 7. Given this information, we now calculate the minimum delay for a pathological wire using repeaters and scaled global wire dimensions. Fig. 11 compares this minimum delay for two wire widths to the global cycle time supplied in [1]. We see that beyond a certain point the delay actually increases since the wire *RC* product is rising at the same time that the line length is increasing. Clearly, the scaling of global conductors is not compatible with the expected rise in global clock speeds. This trend has been identified by industry; IBM's 0.18-μm generation uses a larger global metal pitch than the previous generation. Fig. 12 explores the power ramifications of using scaled global wires: L_{crit} drops quickly due to rising line resistance, resulting in huge amounts of repeaters. At 0.05 μm, this wiring paradigm will consume $\sim$40% of the projected total power just for global interconnect distribution (repeaters + wires) while yielding severely degraded performance.

To help integrate power considerations, a modification is made to Bakoglu's optimal buffer sizing expression [4] to provide an area-optimal buffer size. By multiplying a weighted area function ($W^{1/3}$, where W is the device

[4] A later version of the roadmap [1] projects even larger microprocessor die sizes (by about 10%). However, upcoming roadmap revisions point toward more compact dies ($\sim$500 mm^2 at 0.03–0.05 μm) that further limit the impact of time-of-flight effects.

[5] Global aspect ratio may be slowly increasing, which would yield better *RC* characteristics and somewhat worsened noise than our analysis suggests.

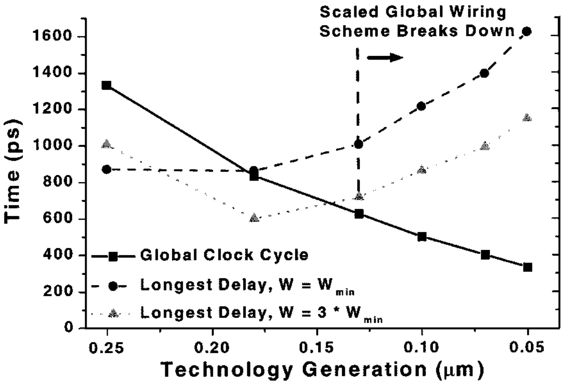


Fig. 11. Delays for corner-to-corner wires compared to global clock cycle from [2]. Scaled global wiring is used.

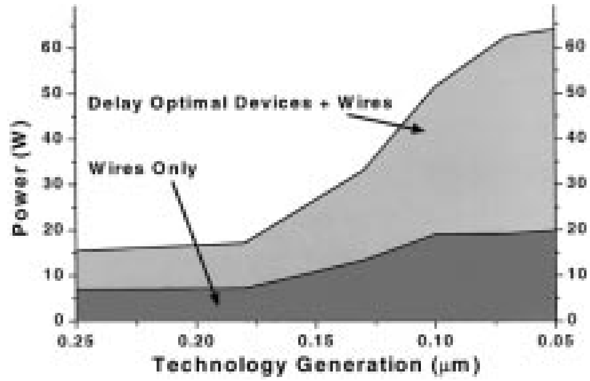


Fig. 12. Power for all repeaters and global interconnect (sized 3 × $W_{\min}$) where 50% of all devices are logic.

width) by the delay (T_d), we obtain a new objective function. Since optimization of delay alone usually results in overly large buffers with high power requirements, we have used the weighted area function to power and delay concerns. The product of T_d and $W^{1/3}$ is then differentiated and the minimum value is found at:

$$W_{\text{opt\_area}} = \frac{0.541}{R_w C_{\text{in}}} \left(-0.231 R_{\text{dev}} C_{\text{out}} - 0.126 R_w C_w + \begin{pmatrix} 0.053 R_{\text{dev}}^2 C_{\text{out}}^2 + 0.058 R_{\text{dev}} C_{\text{out}} R_w C_w \\ + 0.016 R_w^2 C_w^2 + 1.708 R_{\text{dev}} C_{\text{in}} R_w C_w \end{pmatrix}^{1/2} \right). \tag{3}$$

This expression gives a smaller value of W than the original formulation in [4]; the delay is consequently higher but the area and power savings are considerable. At line lengths that approach L_{crit}, this formula gives areas that are typically 50% to 65% smaller than [4]. The delay penalty remains under 20% until the line length is about 1/4 of L_{crit}, at which point the delay is not substantial (Fig. 13). At L_{crit}, the typical delay penalty is ~11%. As a simpler approximation to (3), we have found that at L_{crit} the $W_{\text{opt\_area}}$ is 50.8% of W_{opt} from [4] or simply 1/2.

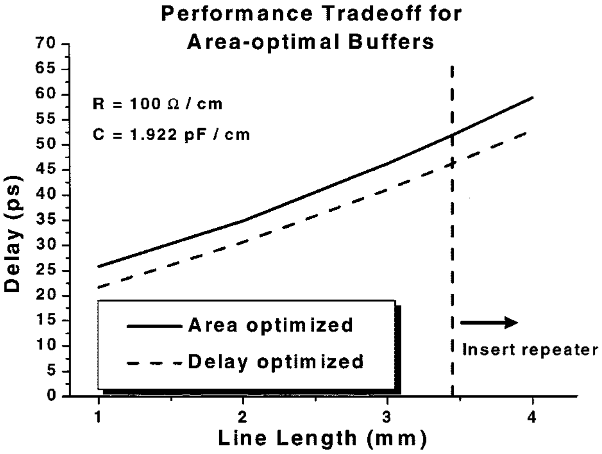


Fig. 13. Comparison between new area-optimal repeater size and delay-optimal repeater expression [4].

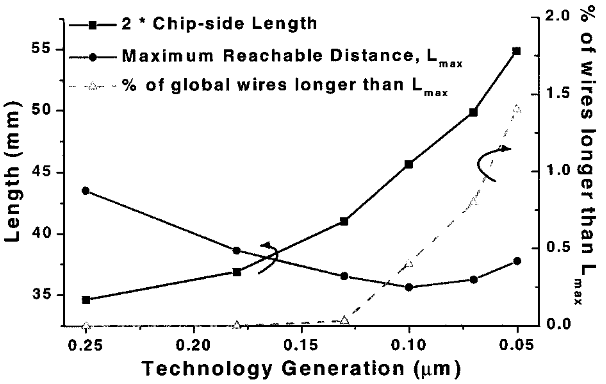


Fig. 14. $L_{\max}$ becomes less than twice the chip-side length past 130 nm for minimum-pitch fat wiring. Wiresizing allows $L_{\max}$ to exceed 2^*D_c at all technology nodes.

C. Fat Wiring

1) Performance Analysis: The use of fat, or unscaled, wires at the global metal levels was first suggested in [5]. Unscaled wiring has the benefit of a fixed low *RC* delay at the expense of fewer available routing tracks. In this work, we present a more comprehensive analysis of the performance and routing impact of using such fat wires. Let us first examine the performance aspect of fat wires. Fat wires in this work have a pitch of 2 μm and a thickness of 1.5 μm. The resistance of a cladded copper wire at these dimensions is 147 Ω/cm. Capacitance varies from approximately 2 pF/cm to ~0.75 pF/cm in 0.05-μm technology. Fig. 14 explores the maximum reachable distance, $L_{\max}$, for a minimum-pitch fat wire at each generation of interest. $L_{\max}$ is defined as the distance that can be traveled in 80% of the global clock cycle predicted in [1] and is found using analytical delay expressions [37]. This is compared to the pathological corner-to-corner wirelength, 2^*D_c. ITRS values for die size are used in the expectation that they will present an upper bound on chip area according to current design trends. We see that, even with fat wires, corner-to-corner wirelengths cannot be accommodated past 0.18 μm.

Table 8
Wiresizing Leads to Increased L_{crit} Values (0.05-μm Values Shown)

Linewidth (µm)	L_{crit} (S = S_{min}) (µm)	L_{crit} (S = W) (µm)
1	1661	1661
1.5	1917	1991
2	2098	2207
3	2344	2476
4	2504	2639
5	2617	2749

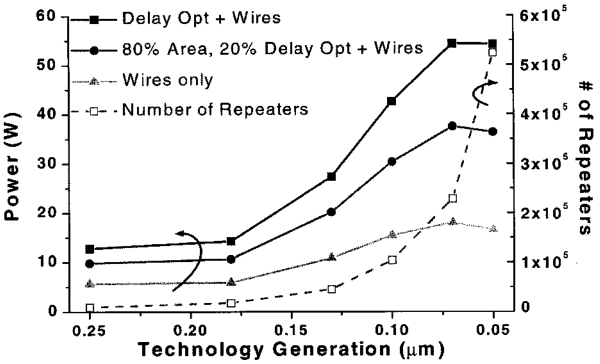


Fig. 15. Fat wiring yields lower power and better performance. Area-optimal drivers cut power by $<30\%$ and the number of repeaters is reduced from the scaled wire scenario.

However, the situation is not as bad as it appears. Based on empirical data and global wirelength models [28], [38] we project that the percentage of global wires with lengths $>L_{\max}$ is very small, under 2% throughout. Proper floorplanning may be able to further limit these wires. In addition, when using larger than minimum linewidths, $L_{\max}$ increases such that, even at 0.05 μm, it exceeds 2*D_c. Thus, for very long wires (a very small portion of the total global picture), wiresizing techniques can be used to maintain acceptable performance at the penalty of increased power. Table 8 demonstrates the effectiveness of wiresizing at 0.05 μm with L_{crit} calculated as in [31]. By doubling the linewidth from the minimum, L_{crit} rises by 26%, which translates directly to $L_{\max}$ since repeaters reduce delay dependency to linear. Two scenarios are shown: 1) spacing is kept at minimum and 2) spacing is equal to linewidth. The second case has severe routing resource penalties with little performance gains. Fig. 15 revisits the power issue using fat wires and repeaters with calculations based on [28], [39]. We find that power is actually slightly decreased from the alternative scenario using scaled wires. Furthermore, ITRS global clock frequencies can be met throughout the roadmap using fat wires. Also, the use of area-optimal repeaters rather than delay-optimal (80% area-optimal, corresponding to less critical paths) reduces global interconnect power consumption significantly, by ~30% at 0.05-μm technology.

2) Routing Resources: This section assesses the impact of fat wiring on routing resources by examining several factors that limit IC routing capacity including power distribution, via blockage, and clock routing. In the discussion, we will focus on the logic portion of a design as the wiring requirements for memory are generally much lower than that for logic.

The power distribution network uses significant routing resources, especially at the top layers. In this work, power grid dimensions are found by limiting the peak *IR* voltage drop to under 4% of V_{dd}. Analytical expressions are derived [40] to describe the *IR* drop of an arbitrary layer as a function of metal linewidth. When reliability constraints are met, the percentage of routing resources used is calculated for each metal layer.

It has been estimated empirically that for metal layers with equivalent pitches, an upper layer blocks 12%–15% of an underlying layer due to its need to connect to the substrate using vias [5].[6] However, when larger metal pitches are used on higher levels, the amount of blockage is reduced by using fixed-size (or nearly fixed-size) vias. The relationship between via blockage and metal pitch is linear in this case. With a multilevel interconnect system, vias connecting the top layer to the substrate necessarily block all underlying levels. Thus, metal one is blocked by all subsequent layers, resulting in a sizable loss in its routing capacity. Fortunately, upper layers have significantly larger pitches than bottom levels, reducing the via penalties associated with multilevel interconnect.

Clock distribution also serves to reduce available signal routing resources. Due to the regularity of H-tree structures, the total wiring required for such a network can be found fairly accurately. Given the number of clusters in the H-tree (see Section XII), the total wirelength is given by a simple analytical expression in terms of the chip-side length. Additional "within-cluster" routing is approximated using a heuristic that allocates wires from the central driver to the perimeter of the cluster in all directions [39].

Routing tools cannot fully utilize all the available routing resources for a given design. This is mainly due to the algorithms used within the routing tools. This effect is modeled by first calculating the available routing resources after clock routing, power/ground routing, and via blockages. At this point, the routing area is multiplied by a routing efficiency factor to give the estimated available routing resources. This routing efficiency factor is set at 0.5, which is based on discussions with industry CAD engineers.

Based on the above models, the wirability of a generic 0.05-μm microprocessor has been studied to determine the feasibility of the fat wiring scheme. Global wires are defined as any wires leaving a 50K gate module. We further classify these global wires into semiglobal and global. Wires that leave an isochronous (locally clocked) region are termed global wires as they will run at the slower global clock frequency and are routed on fat wiring levels. Wires that leave modules but not their isochronous regions are termed semiglobal and are routed on semiglobal wiring, which is

[6] Recent work has shown this approximation to be pessimistic [41]. Thus, we contend that our results on routing resource availability will also be slightly pessimistic.

not considered to be fat wires (i.e., they scale from process to process). This routing hierarchy is discussed further in the next section. Routing resources for both semiglobal and global wiring is considered.

There are found to be 100 isochronous regions, each of which contains 35 modules of 50K gates each. This design corresponds to a 50% logic (device count) microprocessor where 67% of the chip *area* is used for logic. Designs with extensive memory (75% or more) are more likely and will be more easily wirable.

We assume that semiglobal routing will use dimensions that are 3× the minimum allowable. For 0.05 μm, this corresponds to a contacted pitch of 0.45 μm. Aspect ratio is set at two to compromise between resistance and noise effects [16]. Using Rent's rule-based models [42], an average intra-isochronous wirelength of 1.15 mm is found for fan-out of two. While this value may seem small, it is actually greater than 50% of the isochronous region side-length. With an expected average wiring pitch of $1.5^{*}P_{\min}$ (conservatively allowing for wiresizing), the total required semiglobal routing resources for each isochronous region is 3.32 mm^2.

Available routing is now determined, starting with two levels of semiglobal wires. If the requirements exceed capacity, a third semiglobal level will be required. Via blockage depends on upper layers that have not yet been assigned—we use final results of fat wiring optimization to yield accurate results. Via blockage for the uppermost layer is 16% and for the underlying layer, 29%. Minimum-pitch wiring is found to be acceptable for power distribution with wiring usage of less than 2%. Clock distribution is performed on fat wiring layers so its impact on semiglobal resources is neglected. With a routing efficiency of 0.5 we find these two layers to provide 3.84 mm^2 of routing, which is greater than the requirements. Note, however, that significant wiresizing such that the average pitch exceeds $2^{*}P_{\min}$ would yield an unwirable design.

Let us now examine global routing. Using the same approach as before (where the global Rent's exponent is slightly smaller due to floorplanning emphasis and potential delay penalties associated with global routing), the average wirelength is calculated to be 7.84 mm (with fan-out of 2). There will be nets with much longer lengths than this, however many signals will only need to travel from one isochronous region to its nearest neighbor. The length of such a net will be in the range of 2–3 mm. Therefore, this wirelength is a conservative intermediate value between relatively rare pathological lines and more typical shorter global wires. Routing requirements are determined with an average wiring pitch of $2 \times P_{\min}$. We use a wider average pitch here than in the semiglobal case because we anticipate the use of shielding wires in order to deal with inductance. This will limit routing density and serve to increase the effective wiring pitch somewhat. Recall that very few ($<$2%) fat wires will require wiresizing to achieve delay targets.

At this point, we forecast the need for two classes of fat wires. The extreme wires ($L \gg L_{\text{avg}}$) require very fat wiring tracks. However, the shorter global wires also discussed will only waste routing resources if they are routed at these fat levels. So, we allow for the classification of global wires between "shorter" global wires and "extreme" global wires. The former will have greater numbers but shorter average lengths whereas the latter will be very few in absolute numbers but their length and performance impact is significant. Given ITRS expectations for nine to ten metal layers at 0.05 μm, four to five layers remain for global usage (three for local and two for semiglobal). Using the more aggressive number, we further break this into two fat layers and two layers that will have ~1/2 the pitch and thickness of the fat layers. Using $P_{\text{fat}} = 2$ μm, we find the routing requirements to be 656.5 mm^2 by approximating that 2/3 of the wires will be routed on the lower global layers since their capacity is doubled.

Our analysis indicates that vias block 15% of metal 8, 14% of metal 7, and 27% of metal 6 routing tracks in this wiring scheme. Power and ground distribution accounts for ~5% on the top level (flip-chip is used with ~10-mV voltage drop) and 2% on subsequent layers. Clock distribution (both local and global) is found to consume 5% of routing area on the top two levels (we assume that, in order to limit process variation, a shielding plate is used underneath the top level distribution) and negligible amounts of the other global layers. Summing these contributions, we find a total available routing area of 806 mm^2. This system is therefore wirable, even in the presence of anticipated shield wires.

This analysis indicates that by adding global wiring layers during each generation (while remaining within the bounds of ITRS expected metal layers), large-scale microprocessors remain wirable at 0.05 μm ($>$1 billion transistors). For instance, six of the nine metallization levels at this process technology should be used solely for global routing, where global routing is defined as routing among 50K gate modules.

X. Routing Hierarchy

Our results indicate that due to global *RC* delays as well as time-of-flight considerations, the global clock will necessarily be slower than the achievable local clock frequency. Early examples of such a microarchitecture are beginning to appear in the most advanced microprocessors being developed today (1.5+ GHz); speed sensitive functions (e.g., integer execution units) may use a higher frequency clock than other modules. We expect this 2-clock architecture to fully evolve around the 0.13- or 0.1-μm generation, which is one to two generations later than predicted in [1]. The local clock speed will be set roughly by the delay time through ten loaded gates (approximately 10 GHz at 0.05 μm). This will continue to rise as long as faster devices can be made. However, the global clock speed will be determined by the propagation delay of the longest global interconnect. Alternatively, global signals can be divided into a number of clock periods by the insertion of latches along the repeater path. In any case, it is clearly in the best interests of the designer to keep the global clock as fast as possible (or as close to the local clock as possible) in order to reduce latency that will be associated with global accesses. The problem of increasing the global clock speed becomes equivalent to reducing the length

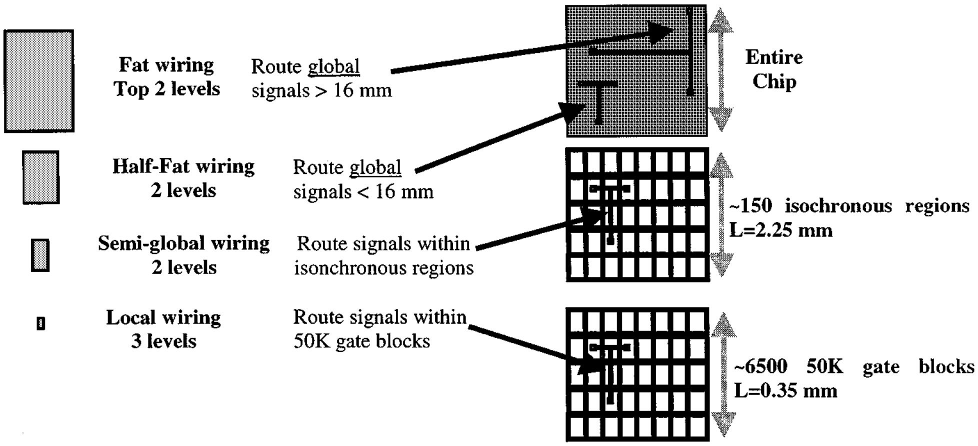


Fig. 16. Application of a new wiring hierarchy to a 50-nm microprocessor.

of the longest global interconnect. Thus, timing-driven floorplanning will be key in that reducing the pathological wirelength from 2*D_c to D_c will effectively double the global clock speed (since repeaters reduce the delay problem to a linear one).

Furthermore, with the use of fat wiring on global metallization layers, we demonstrate that cross-chip global wires will not require ten or 20 clock cycles for propagation at 0.1 μm as described elsewhere. In fact, we expect that global clock speeds will be kept within a factor of $\sim$4 of the local clocks to at least 0.05 μm. This implies that fewer than four local (fast) clock cycles will be needed to cross the chip using dedicated global wiring. As mentioned above, this relatively low latency can be achieved by dividing long global signal paths using latches or by using a separate global clock. The former may reduce some overhead incurred when introducing a separate clock but the power implications of each approach are not clear. In either case, fat or slowly scaled global wiring is the key to attaining low latency global signaling.

What is the size of a locally clocked, or isochronous region at 0.05 μm? This question can be answered by determining how far we can transmit a signal within a single local clock cycle. Using the top-level fat wiring ($P_{\min} = 2$ μm) we find that, in 80% of the 10-GHz local clock cycle, $L_{\max}$ is about 14 mm. This value corresponds to 16 isochronous regions with an area of 47 mm^2 each. However, a different approach could be used to determine isochronous region size. Using lower levels of metal (not fat wiring), we will obtain a smaller value of $L_{\max}$, leading to more isochronous regions. The advantage here is the exclusion of fat wiring from being used inside of these locally clocked zones, freeing up more routing tracks for longer global wiring.

This point forms the basis for a new wiring hierarchy that complements the envisioned DSM microarchitecture. The microarchitecture at 0.1 μm and beyond consists of three levels—the global level, the isochronous level, and the module level. The optimal wiring hierarchy used to interconnect these designs also consists of three levels—global routing (connecting elements at global clock frequency), semiglobal routing (connecting modules within isochronous regions), and local routing (connecting gates within modules). In this manner, each level of the wiring hierarchy has a dedicated purpose; to provide connectivity for its corresponding level of the microarchitecture. Fig. 16 and Table 9 further explain this new wiring hierarchy. Having outlined our general approach to managing global interconnect, we now consider a number of particular factors that might inhibit global wiring performance in deep submicrometer.

Table 9
Back-End Process Parameters for a 0.05-μm Microprocessor Using the Proposed Wiring Hierarchy

0.05 μm design	Fat	½ Fat	Semi-global	Local
(μm)	2 levels	2 levels	2 levels	3 levels
Pitch	2	1	0.45	0.15
Thickness	1.5	0.75	0.45	0.15
ILD Thickness	0.8	0.4	0.18	0.06

XI. Inductance

Inductive effects are expected to become more significant in future DSM designs since signal bandwidth is increasing as on-chip rise times decrease. A good definition for signal bandwidth is given by $0.35/T_{\text{rise}}$, which corresponds to a cutoff frequency at the -3-dB points [43]. For instance, a 1-GHz signal with 100-ps rise/fall times has a bandwidth of 3.5 GHz, which is significantly greater than the operating frequency.

If the product of inductance and angular frequency (2π*$0.35/T_{\text{rise}}$) is comparable to the line resistance, a first-order statement that inductive effects are important can be made. More accurately, expressions from [44] can be used to define a range of line lengths at which inductance should be considered.[7] We have examined each technology generation to determine the relevant wirelengths with regard to inductance. Fig. 17 examines the relationship between the intervals found and the critical line lengths for 0.05 μm. Fat wiring is used at $W_{\min}$. Wider wires have also been studied; their smaller resistance and larger capacitance make the range of significant wirelengths much greater. In Fig. 17, the

[7] These expressions assume that the line is being properly driven—overdriving RLC lines results in ringing effects and additional delay.

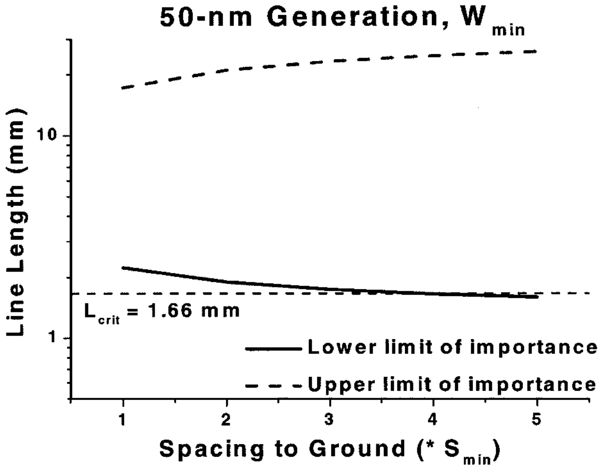


Fig. 17. Minimum pitch global wiring at 50 nm requires shield wires within a few pitches to limit inductive effects.

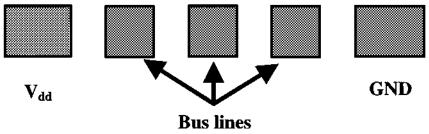


Fig. 18. Suggested global routing practice for sub-100-nm technologies. V_{dd}/GND lines supply close current return paths to reduce inductance.

spacing to ground is varied—inductance is a weak function of conductor geometry but a strong function of the distance to the current return path. Large inductive loops are created when the return path is far away whereas inductance can be limited by routing nearby ground lines to provide a stable current return path. This figure shows the importance of having a nearby current return path that can be accomplished by using shield wires as demonstrated in Fig. 18. It was shown in [45] that the use of interdigitated shield wires is normally a more effective approach to limiting inductance than the sandwiched ground plane approach taken in some commercial microprocessors [46]. Our results indicate that the range of inductive effects expands for scaled processes to include more and more of the useful wirelength spectrum for a design. Extensive use of shield wires will drastically reduce routing density, further emphasizing the need for added global metallization levels.

The most significant inductive effect on signal wiring may be in the form of mutual inductive coupling. The impact of shield wires on reducing inductive coupling noise is not clear. The nonlocality of inductive coupling makes analysis a much more difficult problem than capacitive coupling since the number of potential aggressors is substantially larger.

XII. Clock Distribution

Another major issue concerning chip-level interconnect is that of clock distribution. As the clock cycle shrinks, we see a corresponding drop in allowable clock skew. However, rising die sizes mean that a larger overall clock distribution network must be provided. These two points lead to a fine-grain clock network in which the growing network is made up of increasing numbers of shrinking components. In this section, we apply the well-established buffered H-tree to future designs to determine if it can continue to provide low-skew, high-speed clock distribution. Modified H-tree designs that take into account nonuniform clock sinks by modifying line lengths and driver strengths are directly extendable from this analysis.

Concentrating on local skew, we use BACPAC to find the required size of an H-tree for a 0.05-μm design. For a local clock cycle of 100 ps, we allot 5% of this to local skew and 5% to global skew. Modeling of global clock skew is complex since it requires an estimation of process variation at all intermediate levels of the H-tree. On the other hand, local skew is determined mainly by the size of a cluster (smallest component of an H-tree) and localized process variation at the last level of buffering.

Using this approach at 0.05 μm, an appropriate H-tree contains 4096 clusters and yields local skew of ~1 ps, or 10% of a loaded gate delay. We assume 10% variation in relevant performance parameters such as interconnect and gate capacitances and drive current. Each cluster has a size of 0.183 mm^2 and contains one or two 50K gate modules. This clock tree corresponds to the local clock distribution—the global clock must also be distributed over the entire die. While a clock tree containing 4096 clusters seems complex, the buffered H-tree structure has significant advantages that may allow it to continue as the clock network of choice in DSM SoCs. It does not use large amounts of wiring, has relatively low power consumption (compared to grid-based networks), and CAD tools exist to exploit the regularity of its structure in the design phase.

Global skew may be the most important component of skew due to the large die sizes expected in the future. If global clock skew becomes a bottleneck in chip design, new design styles must be looked at. Currently a potentially exciting new area of research lies in moving the clock distribution network from on-chip to the package level (see Fig. 19) [47]. Specifically, the use of flip-chip packaging allows for easy signal distribution since connection can be made anywhere on the die as opposed to only the periphery [48]. Also, flip-chip packaging has low parasitics (inductance, capacitance) so that the package-to-chip connection is relatively clean. In general, package-level *RC* parameters are three to four orders of magnitude smaller than those found in on-chip applications, allowing for easy transmission of digital waveforms over a large area (e.g., 750 mm^2) with little attentuation. Using this technique, global clock skew can be minimized while local clock skew constraints can be relaxed, allowing for larger clusters within a clock tree.

XIII. Noise

The issues of L^*di/dt noise and crosstalk noise are significant in that, given current design techniques, both issues are expected to become more problematic. Flip-chip packaging presents itself as a partial solution to L^*di/dt noise since flip-chip has very low inductive parasitics compared to conventional wirebonding.

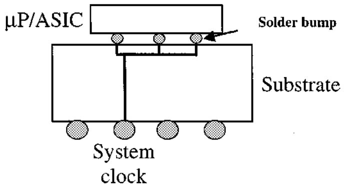


Fig. 19. Diagram showing package-level global clock distribution using flip-chip packaging [46].

Crosstalk noise and dynamic delay are also important due to the dominance of coupling capacitance over ground capacitance. Based on the interconnect scaling scenario presented here (aspect ratio of 1.5 for global lines, unscaled minimum pitch) and the analytical crosstalk model of [30], we have found that crosstalk at the global level will not be as significant as the local level due to the relatively large spacings and use of large repeaters—their capacitance will dampen the effects of coupling capacitance. For instance, the maximum line length of $L_{\rm crit}$ has a crosstalk voltage of 80 mV at 0.05 μm using area-optimal repeaters. This corresponds to 13% of the supply voltage, which is not overly high. Values over 20% are usually considered problematic [16]. However, the use of scaled global wires will lead to much larger values of crosstalk since $S_{\rm min}$ will be decreased.

Dynamic delay on critical timing paths will need to be limited by the use of shield wires, which are also helpful in reducing inductive effects. It is very unlikely, however, that a long net will have two neighbors for its entire run and these signals will switch simultaneously. Efficient screening tools and intelligent routers will be useful in assessing the noise susceptibility of global nets. Even in the worst case scenario, we expect only a 30% rise in delay for an optimally driven line of length $L_{\rm crit}$. Nonetheless, 30% delay variation is unacceptable on critical timing paths, hence the rising need for shield wires.

XIV. Power Distribution

Voltage drop in the power distribution networks of large-scale designs is a function of the instantaneous current being drawn from the supply as well as the distribution network resistance. Significant *IR* drops (e.g., $>$5% of V_{dd}) lead to delay variation and reduced noise margins. With rising power consumption yet dropping V_{dd} values, the supply current in future microprocessors will increase quickly. Larger wires are needed in order to limit *IR* drops, which negatively impacts global routing. In this section, we compare two packaging technologies to help determine if dc voltage loss problems can be managed in DSM.

Power and ground on a chip are normally distributed using a grid of metal lines on each layer of metal with connections when equipotential lines cross each other. The power grid model we use is adapted from [49] and described in more detail in [40]. The maximum *IR* drop is calculated for each metal layer and then summed to obtain the total drop from the pads to the silicon level. Our hierarchy consists of a pair of lines (V_{dd} and ground) running parallel at minimum spacing. The distance between lines of the same potential on a given layer is called the grid pitch. For each grid pitch, there are two lines running the full chip length. We concentrate on the top layer of metal since the majority of the total voltage drop, $V_{\rm drop}$, occurs there.

Conventional wirebonding constrains power pads to be located at the chip periphery, creating very long lines from the supplies to the middle of the die. Thus, wirebonding results in very wide power distribution lines on the top metal layer hampering the routing of other signals such as clock or global buses. Imagine a gate in the center of the die. The resistive path from the pad to this gate will be at least $D_c/2$ in length.

Flip-chip technology allows for V_{dd}/GND to be distributed anywhere on the die using solder bumps. If V_{dd} and ground bumps alternate, then the effective distance between power connections to the grid becomes $2^*P_{\rm bump}$, where $P_{\rm bump}$ is the bump pad pitch. The effective length of the worst case resistive path from pad to underlying level has therefore been decreased from $D_c/2$ to $P_{\rm bump}$. This reduction corresponds to at least a 50× change, which can be directly translated to narrower wires and lower voltage drops.

XV. SoC Global Issues

To this point, we have focused on global issues in high-speed microprocessors as they have the largest die sizes and highest performance of all large-scale IC applications. However, the growing ASIC market is pushing for more advanced technologies and system-on-a-chip architectures. For this reason, it is interesting to look at the global wiring paradigm we have described in the context of SoCs. The major differences between SoCs and microprocessors in terms of global interconnect are that SoCs typically have smaller die sizes and lower performance. These points make the communication requirements much less stringent for SoCs. For instance, time-of-flight is not a concern in SoCs as traversing the die sizes of interest (on the order of several hundred mm^2) will not consume an appreciable portion of a relatively long ($<$2 GHz) clock cycle. Fat wiring is required in microprocessors due to very long global wires and quickly shrinking cycle times. However, with reduced die sizes in SoCs (and consequently reduced global wirelengths) global wiring may be scalable to the 0.1-μm generation or even beyond. As we have seen, the use of fat wiring also reduces the impact of noise since line spacing is unscaled. SoCs may have larger signal integrity problems if scaled global wiring is used in which case sophisticated wiresizing techniques or shield wires may be required. The inductance problem in DSM SoCs is significantly less than that in microprocessors for two reasons. First of all, the need for wide and fast buses is somewhat reduced in SoC designs and it is these structures that are most susceptible to inductive phenomenon. Secondly, if scaled global conductors are used due to relaxed timing requirements, the resistive component of the wires will serve to dampen the inductive effects.

Regarding packaging, SoC vendors may prefer to keep the packaging of their parts very simple, and as a result may not choose the flip-chip global clock distribution. However, the lower clock speeds will, in general, result in little or

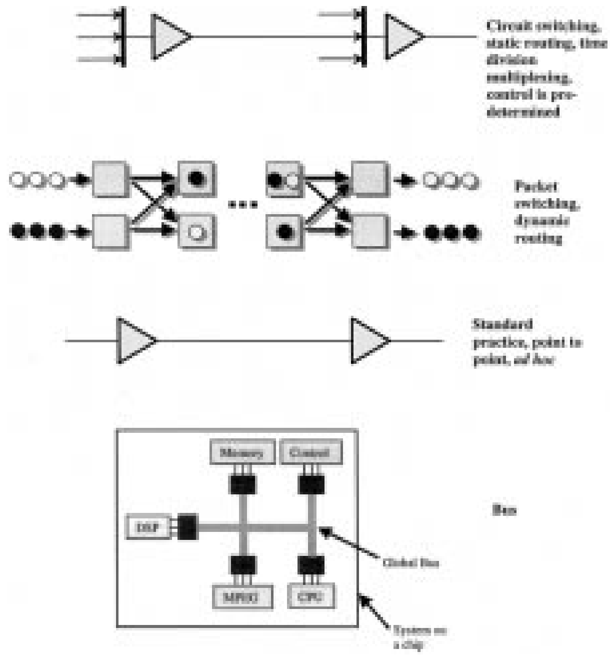

Circuit switching, static routing, time division multiplexing, control is pre-determined
Packet switching, dynamic routing
Standard practice, point to point, ad hoc
Memory
Control
DSP
MPEG
CPU
Bus
Global Bus
System on a chip

Fig. 20. Various global connection schemes in an SoC architecture.

no divergence of global and local clocks. In this case, an on-chip H-tree clock distribution network will be sufficient to distribute over the entire die at 1–2 GHz. The lower power constraints placed on SoCs by plastic packages also lead to smaller *IR* drop with a well-designed power supply grid. Eventually, however, we expect power consumption and related issues such as power distribution to become the limiting factor in DSM design.

XVI. Microarchitectural Implications

A. Choice of Channel Implementation

We begin by outlining a few alternatives for interconnecting multiple modules. The apparent choices are *ad hoc* or point-to-point connections, buses (e.g., AMBA from ARM, Core-Connect from IBM, or Silicon Backplane from Sonics), or switching networks. Among switching networks one can use either circuit-switched or packet-switched. These choices are illustrated in Fig. 20.

We do not attempt to consider the many high-level architectural issues, but restrict ourselves to looking at physical implications on microarchitectural choices. Based on our analysis of global interconnect resources, our conclusion is that there are sufficient routing resources for *any* of the interconnect choices named above, including point-to-point connections. As a result, we believe that choice of channel implementation will be principally determined by higher level architectural issues.

B. Signaling Approaches

The physical implementation of our proposed microarchitecture will have a large role in limiting delay/noise/power. For example, the current paradigm of using large repeater structures at regular intervals seems to work well for *RC* lines with moderate resistance levels. In future processes, global wires look more like RLC lines requiring attention to impedance matching in order to reduce ringing and overshoot effects. Also, with longer and more resistive wires, the sheer number of repeaters grows to staggering proportions as discussed above. This leads to substantial power dissipation as well as placing stringent requirements on power distribution networks and contributing to simultaneous switching noise problems.

A new global signaling convention would seek to minimize power dissipation (helping the power distribution and switching noise problems as well) while ensuring clean signal propagation (no ringing, noise, etc.). One possibility is the use of low-swing driver/receiver architectures. Depending on the actual architecture, dynamic power can be linearly to quadratically reduced with these methods typically at the expense of noise immunity and standby current. To attack the noise problem, differential signaling can be used that gives each signal line a complement, helping to reduce noise through common mode rejection. It can be shown that the use of differential low-swing signaling is actually more area-effective than using shield wires and also leads to better immunity against inductive coupling noise [50]. This is a promising area of future research.

C. Reliability

Another area where physical issues may influence communication approaches is reliability. To date, integrated circuits have been designed with the presumption that a signal launched at one latch in the integrated circuit will arrive reliably at its destination. Ensuring this quality is the problem of maintaining *signal integrity*. For some time boards and racks have been designed with attention to signal errors.

There are three levels of defense. The first is designing to maximize signal integrity. This is already in use in ICs. The second is designing to detect signal errors. The third is to allow for limited error correction. These latter two have not typically been required in IC designs, but there is some debate as to whether they will be required in the future. Our position is that this question reduces to one that we have already addressed: the ability to maintain signal strength and speed across long ($>$1 cm) distances on chip. On a single integrated circuit, the problem of signal reliability (in terms of bad data, the need for error correction hardware, etc.) is dramatically better than on a board since the distance between active devices is limited. For instance, although a global net in a very large die may approach several centimeters in total length, the longest practical distance between repeaters will not exceed $\sim$5 mm, which helps minimize the impact of noise sources such as capacitive and inductive coupling. Thus, the problem reduces to ensuring that there are adequate resources to provide proper wire sizes and sufficient repeaters. Our forecasts indicate that there will be such resources available.

D. Interconnect Libraries for Designing DSM ICs

In this section, we consider some constructive approaches to minimizing the impact of DSM effects. Recently, a new

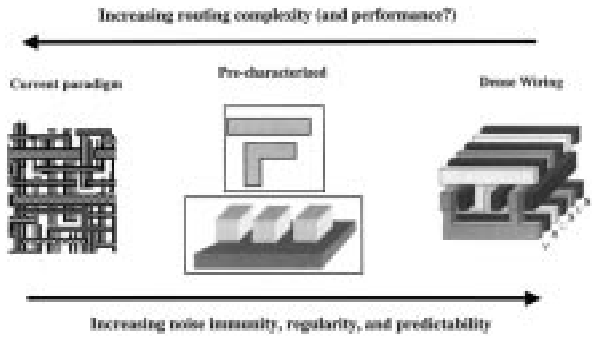

Increasing routing complexity (and performance?)
Current paradigm
Pre-characterized
Dense Wiring
Increasing noise immunity, regularity, and predictability

Fig. 21. Spectrum of routing complexity, bounded by current *ad hoc* methods and the dense wiring fabric.

layout fabric called the dense wiring fabric (DWF) was introduced to cope with the increasing unpredictability in integrated circuit routing [51]. This approach advocates placing a power and ground line next to *each* signal wire in a design, resulting in a fully occupied grid of wires on each metal layer. The advantages of DWF include predictable wire parasitics (capacitance and inductance) as well as simple power/ground grid routing (with low resistance). The main drawback of the DWF is the heavy area penalty associated with putting two shield wires next to each signal line. In [51], the authors saw an average area penalty of 65%, which translates directly to chip area and, hence, yield and cost issues.

Despite this substantial disadvantage, we believe that the DWF may be moving in the right direction, especially if issues such as dynamic delay, inductance, and systematic process variations prove difficult to be incorporated into the design flow (not just during post-routing analysis). However, we feel that a less extreme approach will yield a better tradeoff between the expected performance advantage of the current paradigm and the noise immunity and regularity of the DWF. Fig. 21 highlights a spectrum of interconnect design fabrics where the current *ad hoc* paradigm and the DWF represent the two end points. A good way to view this spectrum is in the allowable number of different interconnect geometries. In the DWF, there is only one possible geometry whereas in the present methodology there are an infinite number of routes possible.

Akin to a front-end standard cell based design approach, an *interconnect library* can be built with a finite number of possible interconnect geometries. Each of these structures can be precisely characterized in terms of parasitics, delays, and noise immunity *a priori*. This point means that the design process is greatly simplified and predictability is substantially improved over current *ad hoc* techniques. At the same time, the chip area will not be dramatically increased with this paradigm as is the case with the DWF. In fact, we anticipate the existence of a "sweet spot" for the library size where the actual performance penalty (in terms of area as well as timing) compared to the conventional technique is minimized while the regularity, predictability, and noise immunity of the interconnect library approach make it superior to either the DWF or *ad hoc* methods. Another advantage of the interconnect library approach is that design tools incorporating systematic pattern density and proximity effects inherent in the semiconductor manufacturing process will be more accurate and efficient than in the current paradigm.

XVII. Conclusion

The purpose of this paper is to examine the impact of DSM processing effects on the microarchitectures of future integrated circuits, principally future microprocessors and systems-on-a-chip style ASICs. To accomplish this, we have first modeled future process generations and the operation of critical paths in those process generations. In particular, we have focused on the problem of interconnect effects. Our results indicate that interconnect delay will be small ($<$25%) in blocks of 50K gates and will remain reasonable ($<$40%) even when pessimistic noise considerations are introduced. These results presume that *lines are adequately driven to compensate for capacitive loads and noise effects*. It appears that blocks of 100K gates will also be manageable although power dissipation penalties will begin to be significant at that point. Beyond 100K gates, size of these blocks is limited by several factors dealing with interconnect. For instance, by making modules too large wirelengths increase due to connectivity requirements within the block. Longer wires translate to larger devices to drive the wires, sacrificing area and power. Increased wirelength within a module also complicates IP generation, wire-layer assignment, and hurts performance by underutilizing local metallization levels.

Thus, we argue for a hierarchical block-based design methodology in which future integrated circuits will be implemented hierarchically with large macro-blocks of approximately 50K–100K gates. We further contend that such a methodology is not unrealistically restrictive. For microprocessor-oriented designs most blocks, such as an integer execution unit, fit within the 100K gate limit. Similarly, for system-on-a-chip ASICs in which the design consists of the assembly of predefined IP blocks, most significant modules, such as a 32-bit microprocessor, can also be implemented within 100K gates. In designs with 10^7 logic gates, this translates into a well-defined layout with 100 to 200 modules in the core area. Given this design hierarchy, the major design challenge in DSM becomes the assembly of thousands of 50K–100K gate modules considering chip-level interconnect effects such as time-of-flight, *RC* delay, inductance, and noise.

Having advocated a modular approach to design, we then consider implications on global interconnect. In our analysis of current global routing practices, we found that scaling of global wires is not sustainable beyond the 0.18-μm generation due to the rising *RC* delays of scaled-dimension conductors. Furthermore, we have reinforced the notion that there will be a divergence of local and global clock speeds around the 0.13–0.1-μm technology nodes due to the effects of time-of-flight and wiring *RC* delays. To combat the *RC* problem, we recommend the use of fat global wires as described in [5]. In this study, we used a comprehensive analysis of routing resources in a generic 0.05-μm microprocessor to determine whether the fat wiring scheme (with $P = 2$ μm) is feasible. We found that, by using additional

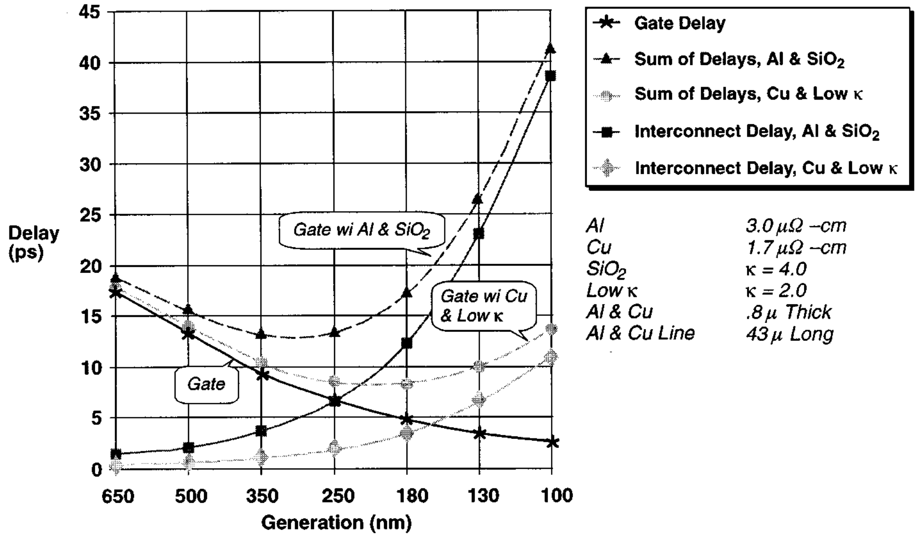


Fig. 22. Projected interconnect and gate delays according to [13].

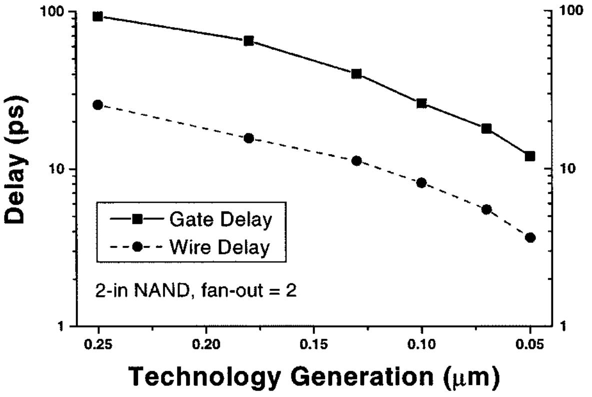


Fig. 23. New depiction of the gate versus wire delay debate.

metal layers for global routing, the fat wiring scheme is indeed scalable to the end of current semiconductor roadmaps. The role of clock distribution in future designs is also studied with the conclusion that the buffered H-tree clock network will enable low values of local skew as long as the number of clusters can be increased. Global skew, while not explicitly modeled, may require moving the global clock distribution network off-chip to the package level where wiring *RC* is much smaller. Noise issues are also discussed; $L^* di/dt$ noise and power supply *IR* drop become much less significant with the use of flip-chip packaging. A discussion of *IR* drop demonstrated that conventional wire-bonding packaging is nonscalable in terms of power supply reliability. Inductive issues were discussed by examining the importance of inductance for various line lengths in different technologies. We found that inductance is becoming a more significant problem, especially when using the fat-wire scheme with low resistance wiring. However, by using shield wires and trading off routing density for critical nets, self-inductance effects should be containable for the most part.

To concisely summarize these results, we revisit the well-known interconnect delay figure from [13] (reproduced in Fig. 22). Collected from the work we have done in this area, we propose an update to this often maligned figure that more accurately represents the scaling of local on-chip interconnections compared to gate delays. Based on a 2-input NAND with a fan-out of two, Fig. 23 realistically reflects gate and interconnect delays, as defined in this paper. On a log scale, it can be clearly seen that the two are scaling nearly identically. Our final point is to emphasize the importance of the device and interconnect interaction. Neither gates nor wires dominate circuit performance; it is the complex relationship between the two that determine how fast or slow a system will perform.

Constructively, we propose a wiring hierarchy that complements the modular design methodology presented. The modules are arranged together in isochronous (or locally clocked) regions that run at a higher clock speed than the global clock. These isochronous regions come together to form the entire design. The wiring hierarchy applies different levels of wiring to route each level of the design hierarchy. Local routing, where minimum pitch is set by lithography capabilities, is used within the modules. Semiglobal routing, whose dimensions are a fixed multiple of the local routing for each generation (i.e., semiglobal dimensions scale with processes), is used exclusively within isochronous regions. Finally, fat global routing connects these isochronous regions with very long wires. High-level architectures that minimize global communication requirements will work best with this design methodology to maximize performance. Using this wiring paradigm, we demonstrate a possible back-end structure for a 0.05-μm microprocessor. On the question of whether this physical wiring fabric will be used to support point-to-point connections, buses, or on-chip switching networks we are agnostics. We believe we have demonstrated that, with the stated assumptions, there will be adequate routing resources to support any of these methodologies and the precise communication structure used will be dictated by higher level architectural issues.

While the details of the microarchitectural implications of this paper may be debated, our primary conclusion is that the effects of small process geometries are less than was originally anticipated and we have pointed to a wide variety of techniques that can be used to successfully ameliorate them.

ACKNOWLEDGMENT

The authors gratefully acknowledge M. Shilman for information on packet-switched networks.

REFERENCES

[1] *International Technology Roadmap for Semiconductors*: Semiconductor Industry Association, 1999.

[2] D. Matzke, "Will physical scalability sabotage performance gains?," *IEEE Computer*, pp. 37–39, Sept. 1997.

[3] M. Bohr, "Interconnect scaling—The real limiter to high performance ULSI," in *Proc. IEDM*, 1995, pp. 241–244.

[4] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.

[5] G. A. Sai-Halasz, "Performance trends in high-performance processors," *Proc. IEEE*, pp. 20–36, Jan. 1995.

[6] A. E. Caldwell *et al.*, "GTX: The MARCO GSRC technology extrapolation system," in Proc. DAC, to be published.

[7] S. Devadas, A. Ghosh, and K. Keutzer, *Logic Synthesis*. New York: McGraw-Hill, 1994.

[8] N. Sherwani, *Algorithms for VLSI Physical Design Automation*. Boston, MA: Kluwer, 1995.

[9] C. Hu, "Device and technology impact on low power electronics," in *Low Power Design Methodologies*, J. Rabaey, Ed. Boston, MA: Kluwer, 1996, pp. 21–35.

[10] D. Edelstein *et al.*, "Full copper wiring in a sub-0.25 μm CMOS ULSI technology," in *Proc. IEDM*, 1997, pp. 773–776.

[11] S. Venkatesan *et al.*, "A high-performance 1.8V, 0.2-μm CMOS technology with copper metallization," in *Proc. IEDM*, 1997, pp. 769–772.

[12] L. Su *et al.*, "A high-performance 0.08 μm CMOS," in *Proc. VLSI Symp. Technology*, 1996, pp. 12–13.

[13] Semiconductor Industry Association, *National Technology Roadmap for Semiconductors*, 1997.

[14] M. Miyamoto, T. Takeda, and T. Furusawa, "High-speed and low-power interconnect technology for sub-quarter-micron ASIC's," *IEEE Trans. Electron Devices*, pp. 250–256, Feb. 1997.

[15] E. M. Zielinski *et al.*, "Damascene integration of copper and ultra-low-k xerogel for high performance interconnects," in *Proc. IEDM*, 1997, pp. 936–938.

[16] D. Sylvester, C. Hu, O. S. Nakagawa, and S.-Y. Oh, "Interconnect scaling: Signal integrity and performance in future high-speed CMOS designs," in *Proc. VLSI Symp. Technology*, 1998, pp. 42–43.

[17] F. Dartu and L. Pileggi, "Calculating worst-case gate delays due to dominant capacitance coupling," in *Proc. DAC*, 1997, pp. 46–51.

[18] G. Yee, R. Chandra, V. Ganesan, and C. Sechen, "Wire delay in the presence of crosstalk," in *Proc. TAU*, 1997, pp. 170–175.

[19] C. Hu, "Gate oxide scaling limits and projection," in *Proc. IEDM*, 1996, pp. 319–322.

[20] N. Rohrer *et al.*, "A 480 MHz RISC microprocessor in a 0.12 μm $L_{\rm eff}$ CMOS technology with copper interconnects," in *Proc. ISSCC*, 1998, pp. 240–241.

[21] J. Montanaro *et al.*, "A 160-MHz, 32-b, 0.5-W CMOS RISC microprocessor," *IEEE J. Solid-State Circuits*, pp. 1703–1714, Nov. 1996.

[22] *User's Manual*: BSIM3 version 3.1, UC-Berkeley, 1997.

[23] R. Payne, "Metal pitch effects in deep submicron IC design," *Electron. Eng.*, pp. 45–47, July 1996.

[24] T. R. Bednar, R. A. Piro, D.W. Stout, L. Wissel, and P. S. Zuchowski, "Technology-migratable ASIC library design," *IBM J. Res. Develop.*, pp. 377–385, July 1996.

[25] S.-P. Jeng *et al.*, "Implementation of low- dielectric constant materials for ULS circuit performance improvement," in *Proc. Symp. VLSI Technology, Systems, and Applications*, 1995, pp. 164–168.

[26] A. Deutsch *et al.*, "Modeling and characterization of long on-chip interconnections for high-performance microprocessors," *IBM J. Res. Develop.*, pp. 547–567, Sept. 1995.

[27] P. Fisher and R. Nesbitt, "The test of time: Clock cycle estimation and test challenges for future microprocessors," *IEEE Circuits Devices Mag.*, pp. 37–44, Mar. 1998.

[28] P. Zarkesh-Ha and J. D. Meindl, "Stochastic net length distributions for global interconnects in a heterogeneous system-on-a-chip," in *Proc. VLSI Symp. Technology*, 1998, pp. 44–45.

[29] D. A. Carlson, R. W. Castelino, and R. O. Mueller, "Multimedia extensions for a 550-MHz RISC microprocessor," *IEEE J. Solid-State Circuits*, pp. 1618–1624, Nov. 1997.

[30] O. S. Nakagawa, D. Sylvester, J. G. McBride, and S.-Y. Oh, "Closed-form modeling of on-chip crosstalk noise in deep-submicron ULSI interconnect," *Hewlett-Packard J.*, pp. 39–45, Aug. 1998.

[31] R. Otten, "Global wires: Harmful?," in *Proc. ISPD*, 1998, pp. 104–109.

[32] A. P. Chandrakasan, S. Sheng, and R. W. Broderson, "Low-power CMOS digital design," *Proc. IEEE*, pp. 473–484, Apr. 1992.

[33] D. Liu and C. Svensson, "Power consumption estimation in CMOS VLSI chips," *IEEE J. Solid-State Circuits*, pp. 663–670, June 1994.

[34] H. J. M. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE J. Solid-State Circuits*, pp. 468–473, Aug. 1984.

[35] [Online]. Available: http://www.mentor.com/inventra

[36] M. Bohr *et al.*, "A high performance 0.25 μm logic technology optimized for 1.8V operation," in *Proc. IEDM*, 1996, pp. 847–850.

[37] T. Sakurai, "Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSI's," *IEEE Trans. Electron Devices*, vol. 40, pp. 118–124, Jan. 1993.

[38] N. Vasseghi, K. Yeager, E. Sarto, and M. Seddighnezhad, "200-MHz superscalar RISC microprocessor," *IEEE J. Solid-State Circuits*, vol. 31, pp. 1675–1685, Nov. 1996.

[39] [Online]. Available: http://www-device.eecs.berkeley.edu/dennis/BACPAC

[40] D. Sylvester and K. Keutzer, "A global wiring paradigm for deep submicron design," *IEEE Trans. Computer-Aided Design*, vol. 19, pp. 242–252, Feb. 2000.

[41] A. B. Kahng, S. Mantik, and D. Stroobandt, "Requirements for models of achievable routing," in *Proc. ISPD*, 2000, pp. 4–11.

[42] W. E. Donath, "Placement and average interconnection lengths of computer logic," *IEEE Trans. Circuits Syst.*, vol. 26, pp. 272–277, Apr. 1979.

[43] S.-Y. Kim, N. Gopal, and L. Pillegi, "Time-domain macromodels for VLSI interconnect analysis," *IEEE Trans. Computer-Aided Design*, vol. 13, pp. 1257–1270, Oct. 1994.

[44] Y. I. Ismail, E. G. Friedman, and J. L. Neves, "Figures of merit to characterize the importance of on-chip inductance," in *Proc. DAC*, 1998, pp. 560–565.

[45] Y. Massoud, S. Majors, T. Bustami, and J. White, "Layout techniques for minimizing on-chip interconnect self-inductance," in *Proc. DAC*, 1998, pp. 566–571.

[46] D. W. Bailey and B. J. Benschneider, "Clocking design and analysis for a 600-MHz Alpha microprocessor," *IEEE J. Solid-State Circuits*, vol. 33, pp. 1627–1633, Nov. 1998.

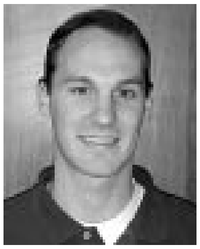
[47] Q. Zhu and S. Tam, "Package clock distribution design optimization for high-speed and low-power VLSI's," *IEEE Trans. Comp., Packag., Manufact. Technol.*, vol. 20, pp. 56–63, Feb. 1997.

[48] R. R. Tummala and E. Rymaszewski, *Microelectronics Packaging Handbook*. New York: Van Nostrand Reinhold, 1989.

[49] W. S. Song and L. A. Glasser, "Power distribution techniques for VLSI circuits," *IEEE J. Solid-State Circuits*, vol. 21, pp. 150–156, Feb. 1986.

[50] Y. Massoud, J. Kawa, D. MacMillen, and J. White, "Differential signaling in crosstalk avoidance strategies for physical synthesis," in *Int. Workshop Timing Issues*, to be published.

[51] S. P. Khatri, A. Mehrotra, R. K. Brayton, A. Sangiovanni-Vincentelli, and R. H. J. M. Otten, "A novel VLSI layout fabric for deep submicron applications," in *Proc. DAC*, 1999, pp. 491–496.



Dennis Sylvester (Member, IEEE) received the B.S. degree in electrical engineering *summa cum laude* from the University of Michigan, Ann Arbor, in 1995, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1997 and 1999, respectively.

He is an Assistant Professor in the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor. Previously, he worked as a Senior R&D Engineer in the Advanced Technology Group of Synopsys. He also worked at Hewlett-Packard Laboratories, Palo Alto, CA, from 1996 to 1998. During his graduate studies, he held a Semiconductor Research Corporation (SRC) Graduate Fellowship. He has published numerous papers in his field of research, which includes interconnect characterization and modeling, on-chip crosstalk, noise-aware timing analysis, and back-end statistical variation.

Dr. Sylvester was awarded the 2000 David J. Sakrison Memorial Prize for the best dissertation in the Berkeley EECS Department. He also received the 2000 Beatrice Winner Award at ISSCC, two outstanding research presentation awards from the SRC, and a Best Student Paper Award at the 1997 International Semiconductor Device Research Symposium. He has given tutorials and invited presentations at several conferences and workshops and serves on the technical program committee for the International Conference on Computer-Aided Design (ICCAD) as well as Technical Program Co-Chair for the 2001 International Workshop on System-Level Interconnect Prediction. He is a Member of Eta Kappa Nu.

Kurt Keutzer (Fellow, IEEE) received the B.S. degree in mathematics from Maharishi International University in 1978 and the M.S. and Ph.D. degrees in computer science from Indiana University in 1981 and 1984, respectively.

In 1984, he joined AT&T Bell Laboratories where he worked to apply various computer-science disciplines to practical problems in computer-aided design. In 1991, he joined Synopsys Inc., where he continued his research in a number of positions culminating in his position as Chief Technical Officer and Senior Vice-President of Research. He left Synopsys in January 1998 to become Professor of Electrical Engineering and Computer Science at the University of California, Berkeley where he serves as Associate Director of the Gigascale Silicon Research Center. He coauthored the book *Logic Synthesis*, (New York: McGraw-Hill, 1994). He has researched a wide number of areas related to synthesis and high-level design.

Dr. Keutzer has received three Design Automation Conference (DAC) Best Paper Awards, a Distinguished Paper Citation from the International Conference on Computer-Aided Design (ICCAD), and a Best Paper Award at the International Conference in Computer Design (ICCD). From 1989 to 1995 he served as an Associate Editor of IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS and he currently serves on the editorial boards of three journals: *Integration—The VLSI Journal*, *Design Automation of Embedded Systems*, and *Formal Methods in System Design*. He has served on the technical program committees of DAC, ICCAD, and ICCD as well as the technical and executive committees of numerous other conferences and workshops.