3-D ICs: A Novel Chip Design for Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration

KAUSTAV BANERJEE, MEMBER, IEEE, SHUKRI J. SOURI, PAWAN KAPUR, AND
KRISHNA C. SARASWAT, FELLOW, IEEE

Invited Paper

Performance of deep-submicrometer very large scale integrated (VLSI) circuits is being increasingly dominated by the interconnects due to decreasing wire pitch and increasing die size. Additionally, heterogeneous integration of different technologies in one single chip is becoming increasingly desirable, for which planar (two-dimensional) ICs may not be suitable. This paper analyzes the limitations of the existing interconnect technologies and design methodologies and presents a novel three-dimensional (3-D) chip design strategy that exploits the vertical dimension to alleviate the interconnect related problems and to facilitate heterogeneous integration of technologies to realize a system-on-a-chip (SoC) design. A comprehensive analytical treatment of these 3-D ICs has been presented and it has been shown that by simply dividing a planar chip into separate blocks, each occupying a separate physical level interconnected by short and vertical interlayer interconnects (VILICs), significant improvement in performance and reduction in wire-limited chip area can be achieved, without the aid of any other circuit or design innovations. A scheme to optimize the interconnect distribution among different interconnect tiers is presented and the effect of transferring the repeaters to upper Si layers has been quantified in this analysis for a two-layer 3-D chip. Furthermore, one of the major concerns in 3-D ICs arising due to power dissipation problems has been analyzed and an analytical model has been presented to estimate the temperatures of the different active layers. It is demonstrated that advancement in heat sinking technology will be necessary in order to extract maximum performance from these chips. Implications of 3-D device architecture on several design issues have also been discussed with especial attention to SoC design strategies. Finally, some of the promising technologies for manufacturing 3-D ICs have been outlined.

Keywords**—***3-D ICs, heterogeneous integration, interconnect performance, optical I/Os, power dissipation, system interconnects, system-on-a-chip design, VLSI design.*

I. Motivation for 3-D ICs

The unprecedented growth of the computer and the information technology industry is demanding very large scale integrated (VLSI) circuits with increasing functionality and performance at minimum cost and power dissipation. VLSI circuits are being aggressively scaled to meet this demand. This, in turn, has introduced some very serious problems for the semiconductor industry. Continuous scaling of VLSI circuits is reducing gate delays but rapidly increasing interconnect delays. The International Technology Roadmap for Semiconductors (ITRS) [1] predicts that, beyond the 130-nm technology node, performance improvement of advanced VLSI is likely to begin to saturate unless a paradigm shift from present IC architecture is introduced. Also, increasing interconnect loading affects the power consumption in high-performance chips. In fact, a significant fraction of the total chip power consumption can be due to the wiring network used for clock distribution, which is usually realized using long global wires. Additionally, interconnect scaling has significant implications for traditional computer-aided-design (CAD) methodologies and tools which are causing the design cycles to increase, thus increasing the time-to-market and the cost per chip function. Furthermore, increasing drive for the integration of disparate signals (digital, analog, RF) and technologies (SOI, SiGe HBTs, GaAs, and so on) is introducing various system-on-a-chip (SoC) design concepts, for which existing planar (two-dimensional) IC design may not be suitable.

A. Interconnect Limited VLSI Performance

In single Si layer (2-D) ICs, chip size is continually increasing despite reductions in feature size made possible by advances in IC technology such as lithography and etching

Manuscript received August 30, 2000; revised February 28, 2001. This work was supported by the Defense Advanced Research Projects Agency AME Program and the MARCO Interconnect Technology Focus Center.

The authors are with the Center for Integrated Systems, Stanford University, Stanford, CA 94305 USA (e-mail: kaustav@ee.stanford.edu).

Publisher Item Identifier S 0018-9219(01)04184-6.

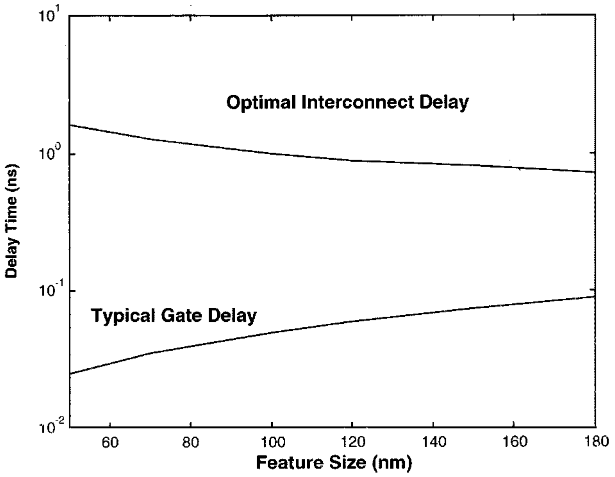


Fig. 1. Typical gate and interconnect delays as a function of technology nodes (minimum feature sizes). The interconnect delay assumes an optimally repeatered line and includes the delay due to the repeaters.

and reduction in defect density [1]. This is due to the ever-growing demand for functionality and higher performance, which causes increased complexity of chip design, requiring more and more transistors to be closely packed and connected [2]. Smaller feature sizes have dramatically improved device performance [3]–[5]. The impact of this miniaturization on the performance of interconnect wires, however, has been less positive [6]–[10]. Smaller wire cross sections, smaller wire pitch, and longer lines to traverse larger chips have increased the resistance and the capacitance of these lines resulting in a significant increase in signal propagation (RC) delay. As interconnect scaling continues, RC delay is increasingly becoming the dominant factor determining the performance of advanced ICs [1], [6]–[10]. Fig. 1 illustrates this problem, where the gate delay and the interconnect delay are shown as functions of various technology nodes based on Table 1 [1]. The interconnect delay has been calculated for an optimally buffered line, whose length equals the chip edge $\sqrt{A}$, where A is the chip area. The methodology used for the delay calculations is described as follows.

1) Interconnect and Gate Delay: Consider an interconnect of total length L. In order to minimize the delay associated with this interconnect, it can be optimally buffered by inserting repeaters between each interconnect segments of length l. The schematic representation is shown in Fig. 2(a). Fig. 2(b) shows an equivalent RC circuit for one segment of the system. V_{st} is the voltage at the input capacitance that controls the voltage source V_{tr}. R_{tr} is the driver transistor resistance, C_p is the output parasitic capacitance, and C_L is the load capacitance of the next stage, and r and c are the interconnect resistance and capacitance per unit length, respectively. The voltage source (V_{tr}) is assumed to switch instantaneously when voltage at the input capacitor (V_{st}) reaches a fraction x, $0 \leq x \leq 1$ of the total swing. Hence, the overall delay of one segment, τ_0, is given by

$$\tau_0 = b(x)R_{tr}(C_L + C_P) + b(x)(cR_{tr} + rC_L)l + a(x)rcl^2 \tag{1}$$

Table 1
Optimal Interconnect and Inverter (FO4) Delays at Various Technology Nodes. Parameters Necessary for the Delay Calculations are also Shown

Feature Size (nm)	180	150	120	100	70	50
Chip Area (cm^2)	4.5	4.5	5.76	6.22	7.13	8.17
Longest wire (cm)	2.12	2.12	2.4	2.49	2.67	2.86
ε_r (ILD, IMD)	3.5	3.5	2.7	2.5	2.5	2.5
ρ_{Cu} (μΩ-cm) @RT	1.673	1.673	1.673	1.673	1.673	1.673
p_{Global} (μm)	1.05	0.85	0.69	0.56	0.39	0.275
Global A.R.	2	2.2	2.4	2.5	2.8	2.9
c (pFcm^{-1})	2.633	2.867	2.393	2.301	2.557	2.643
r (Ωcm^{-1})	303.49	421.01	585.66	853.57	1571.33	3051.35
t_{FO4} (ps)	90	75	60	50	35	25
Interconnect Delay (ns)	0.72	0.81	0.88	0.99	1.27	1.62

where $a(x)$ and $b(x)$ only depend on the switching model, i.e., x. For instance, for $x = 0.5$, $a = 0.4$, and $b = 0.7$ [11], [12]. If r_0, c_0, and c_p are the resistance, input, and parasitic output capacitances of a minimum-sized inverter, respectively, then R_{tr} can be written as r_0/s where s is the multiples of minimum-sized inverters. Similarly, $C_P = sc_p$, and $C_L = sc_0$. If the total interconnect length L is divided into n segments of length $l = L/n$, then the overall delay, τ_d, is given by

$$\begin{aligned}\tau_d &= n\tau_0 \\ &= \frac{L}{l}\, b(x)r_0(c_0 + c_p) + b(x)\Big(c\frac{r_0}{s} + src_0\Big)L \\ &\quad + a(x)rclL.\end{aligned} \tag{2}$$

It should be noted in the above equation that s and l appear separately and therefore τ_d can be optimized separately for s and l. The optimum values of l and s are given as

$$l_{\text{opt}} = \sqrt{\frac{b(x)r_0(c_0 + c_p)}{a(x)rc}} \tag{3}$$

$$s_{\text{opt}} = \sqrt{\frac{r_0 c}{rc_0}}. \tag{4}$$

Note that s_{opt} is independent of the switching model, i.e., x.

Next we substitute (3) and (4) into (1), with $a(x) = 0.4$ and $b(x) = 0.7$. We also make two assumptions to simplify the delay calculations. 1) In the minimum-sized inverter, the PMOS is twice as large as the NMOS device. This is usually employed to match the transistor characteristics. Therefore, $c_p = 3c_{\text{NMOS}}$, where c_{NMOS} is the total source/drain junction capacitance of a minimum-sized NMOS. 2) The output parasitic capacitance c_p is equal to the load capacitance c_0. With these assumptions, the optimum values of l and s can be expressed as

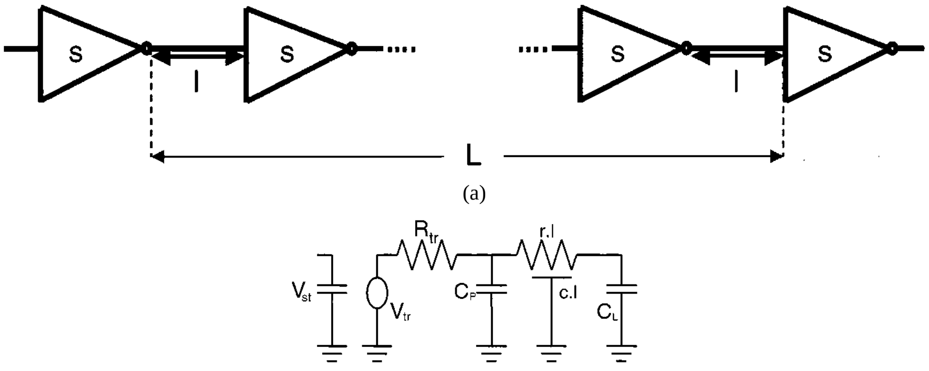
$$l_{\text{opt}} = 3.24\sqrt{\frac{r_0 c_{\text{NMOS}}}{rc}} \quad \text{and} \quad s_{\text{opt}} = 0.577\sqrt{\frac{r_0 c}{rc_{\text{NMOS}}}}$$

and the signal delay along an optimally buffered interconnect of length L can be expressed as

$$\tau_d = 3.24L\sqrt{0.4rct_{\text{FO1}}} \tag{5}$$

where $t_{\text{FO1}} = 6r_0c_{\text{NMOS}}$, and it represents the delay associated with an inverter that has a fan-out of one (FO1).

The delay in (5) can also be expressed in terms of the delay of a gate that has a fanout of four (FO4). The FO4 delay



(a)

(b)

Fig. 2. (a) Optimally repeatered interconnect of length L. Here, each repeater has a fanout of one (FO1). l is the optimal interconnect length between any two repeaters and s represents the optimal repeater size in multiples of the minimum-sized inverters for a given technology. (b) The equivalent RC circuit for one segment.

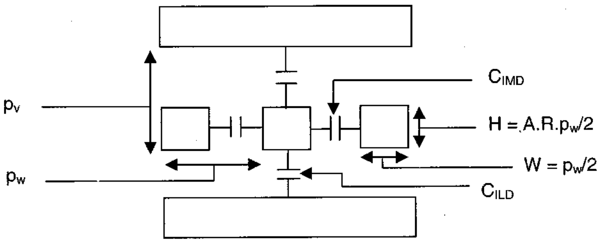


Fig. 3. Cross section of a multilevel interconnect structure showing interlevel (ILD) and intrametal (IMD) capacitances. The aspect ratio ($A.R.$) is defined as (H/W) and the horizontal pitch, p_w, is defined as the sum of line width and lateral spacing between adjacent lines. The vertical pitch, p_v, is defined as the sum of line thickness and vertical spacing between lines on adjacent levels.

is the delay through a buffer (inverter) that is driving four buffers which are identical to itself or a buffer that is simply four times as large. The FO4 delay is a useful metric since any combinational delay, composed of many different types of static and dynamic CMOS gates, can be divided by FO4, and this normalized delay holds constant over a wide range of process technologies, temperatures, and voltages [13]. In terms of FO4, (5) can be approximately written as

$$\tau_d = 2L\sqrt{0.4rct_{\text{FO4}}} \tag{6}$$

where $t_{\text{FO4}} = 15r_0c_{\text{NMOS}}$, which can be estimated from

$$t_{\text{FO4}} = 500L_{\text{gate}} \tag{7}$$

where L_{gate} is the transistor channel length in micrometers and t_{FO4} is in picosecond [13].

2) Resistance Calculations: The resistance per unit length, r, in (6) is generally given by

$$r = \frac{\rho}{A}$$

where A is the cross-sectional area of the interconnect. The width of the interconnect is assumed to be half the horizontal wire pitch, p_w. The vertical wire pitch, p_v, is assumed to be equal to the product of the aspect ratio, $A.R.$, and p_w, and the wire height (thickness) is also assumed to be half the vertical pitch. A and, therefore, r can then be expressed as

$$\begin{aligned} A &= A.R.\,\frac{p_w^2}{4} \\ r &= 4\,\frac{\rho}{A.R.\,p_w^2}. \end{aligned} \tag{8}$$

3) Capacitance Calculations: The cross section of the interconnect structure used for capacitance calculation is represented in Fig. 3. Accounting for the worst case switching, when adjacent wires switch opposite to the signal line, and ignoring any fringe capacitance, the total interconnect capacitance can be simply expressed as

$$C_{\text{total}} = 2(C_{\text{ILD}} + 2C_{\text{IMD}})$$

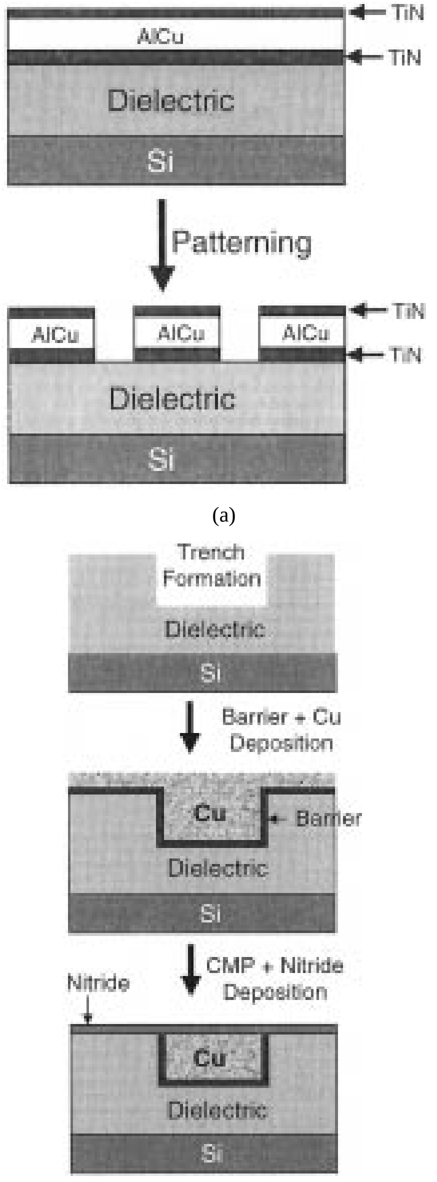
where $C_{\text{IMD}} = \varepsilon_{\text{IMD}} L\,A.R.$ and $C_{\text{ILD}} = \varepsilon_{\text{ILD}}(L/2\,A.R.)$. The factor of 2 in the denominator for C_{ILD} accounts for the overlap with the orthogonal wires on adjacent levels. The length of the overlap is taken to be half the length of the interconnect based on the assumption that wire width is half the pitch. Assuming $\varepsilon_{\text{IMD}} = \varepsilon_{\text{ILD}} = \varepsilon_r$, the capacitance per unit length, c in (6) can be expressed as

$$c = (1 + 4\,A.R.^2)\,\frac{\varepsilon_r}{A.R.}. \tag{9}$$

From Fig. 1, it can be observed that at the 50-nm technology node the interconnect delay is nearly two orders of magnitude higher than the gate delay. Therefore, as feature sizes are further reduced and more devices are integrated on a chip, the chip performance will degrade, reversing the trend that has been observed in the semiconductor industry thus far.

B. Physical Limitations of Cu Interconnects

At 250-nm technology node, copper (Cu) with low-k dielectric was introduced to alleviate the adverse effect of increasing interconnect delay [14]–[18]. However, as shown in Fig. 1, below 130-nm technology node, substantial interconnect delays will result in spite of introducing these new materials, which in turn will severely limit the chip performance. Further reduction in interconnect delay cannot be



(a)

(b)

Fig. 4. Illustration of (a) AlCu and (b) damascene Cu interconnect processes.

achieved by introducing any new materials. This problem is especially acute for global interconnects, which typically comprise about 10% of total wiring, for current architectures. Therefore, it is apparent that material limitations will ultimately limit the performance improvement as the technology scales. Also the problem of long-lossy lines cannot be fixed by simply widening the metal lines and using thicker interlayer dielectric since this conventional solution will lead to a sharp increase in the number of metallization layers. Such an approach will increase the complexity, reliability, and cost and will therefore be fundamentally incompatible with the industry trend of maximizing the number of chips per wafer and 25% per year improvement in cost per chip function. Furthermore, with the aggressive scaling suggested by the ITRS [1], new physical and technological effects start dominating interconnect properties. It is imperative that these effects be accurately modeled and incorporated in the wire performance and reliability analyses. The next three subsections provide quantitative analysis of the impact of these new effects, caused by scaling, on the resistivity of Cu interconnects.

Before proceeding with the analysis, it is important to understand the fundamental differences between the metallization processes for aluminum (Al) and Cu, as illustrated in Fig. 4. For Al-based interconnects [19], first a thin layer of barrier material, titanium (Ti) or titanium nitride (TiN), is uniformly deposited (blanket deposition) on top of a dielectric layer. The barrier layer is used to prevent any interaction between Al and the Si substrate, such as junction spiking. It is also used as an adhesion and texture promoter for the Al layer. The barrier layer is followed by Al[1] deposition and a very thin layer of TiN (capping layer), that is used as the antireflection coating for subsequent lithography processes. These (TiN) layers are also known to improve electromigration performance of Al interconnects. Thus, the metallization layer consists of Ti(TiN)–AlCu–TiN, which is then patterned using a dry-etching process.

In the case of Cu, pattern generation in blanket films by dry-etching processes is difficult because of the lack of volatile byproducts of Cu etching [20]. Hence, Cu films are deposited by the damascene process[2] [21] illustrated in Fig. 4(b). In this process, first a trench is patterned in the dielectric layer. This is followed by a barrier deposition, which coats the three surfaces of the trench. The barrier material is usually a refractory metal such as Ti or Ta or their nitrides [23]. As discussed later, there are different barrier deposition technologies. The barrier layer is necessary since Cu has poor adhesion to most dielectrics and can drift very quickly through them under electric bias to cause metal to metal shorts and to reach the underlying Si substrate where they can diffuse very rapidly through Si interstitial sites and form deep-level acceptors that can degrade device performance [24]. This is then followed by Cu deposition (usually by electroplating). Next, the unwanted Cu and barrier layers outside the trenches are removed using chemical–mechanical polishing (CMP) [25]. Finally, a layer of silicon nitride is deposited which passivates the top surface of the Cu metal in the trenches. Hence, due to the requirement of the barrier metal, effective cross section of the Cu interconnects will be less than the drawn dimensions.

It is commonly believed that material resistivity for Cu would not change significantly for future interconnects [1]. However, because of an increasing dominance of electron scattering from the interfaces and because of a greater fraction of interconnect area being consumed by metal barrier in the future (Fig. 5), the effective resistivity of Cu may rise significantly. In addition, the operational temperature of wires (∼373 K) is higher than room temperature (300 K) and can

[1] In reality, AlCu is employed where proportion of Cu is around 0.5% by weight. This is done to improve the electromigration lifetime of the interconnects.

[2] In practice, a dual-damascene processing scheme is employed where the via and the line are patterned sequentially and then filled with copper in one step [22]. However, since vias are present only at limited number of positions, Fig. 4(b) is an accurate representation of the cross section of Cu lines along most of the interconnect length.

increase further due to self-heating caused by the flow of current [12], [26]. The increase in temperature, in turn, would also increase the wire resistivity. Above effects are next quantified, and more realistic Cu resistivity trends are established.

1) Effect of Interconnect Dimensions on Cu Resistivity: As dimensions shrink, the electron scattering from the surface becomes comparable to the electron bulk scattering mechanisms such as phonon scattering. The dominance of the surface effect depends on the parameter, $k = d/\lambda_{mfp}$, where d is the smallest film dimension and λ_{mfp} is the bulk mean free path of electrons. Smaller k signifies a larger surface scattering effect. The surface scattering governed resistivity is given by [27]

$$\frac{\rho_s}{\rho_o} = \frac{1}{1 - \dfrac{3(1-P)\lambda_{mfp}}{2d} \displaystyle\int_1^{\infty} \left(\frac{1}{x^3} - \frac{1}{x^5}\right) \frac{1 - e^{-kx}}{1-Pe^{-kx}}\, dx}. \tag{10}$$

Here, ρ_s is the resistivity with surface scattering effect, ρ_0 is the bulk resistivity at a given temperature, k is as defined above, and x is the integration variable. The parameter P is a measure of extent of specular scattering at copper/barrier interface. Its value lies between 0 and 1. $P = 0$ signifies complete diffuse scattering causing maximum decrease in mobility; hence, a maximum increase in resistivity, whereas $P = 1$ indicates complete specular reflection leading to no change in resistivity. Values of P are influenced by technology-dependent factors and have been experimentally deduced before for various materials under various conditions [28], [29].

2) Effect of Barrier Thickness on Cu Resistivity: The second effect which contributes to the increase in the effective copper resistivity results from a finite cross-sectional area consumed by the higher resistivity metal barrier encapsulating copper. Barrier thickness, thus its area, depends on the deposition technology as well as the barrier material. Since barrier thickness can not scale as rapidly as the interconnect dimensions, it would occupy increasingly higher fraction of the interconnect cross section area while restricting the current flow only to the lower resistivity Cu. The effective resistivity just due to this effect is given by

$$\frac{\rho_b}{\rho_o} = \frac{1}{1 - \dfrac{A_b}{A.R. * \left(\dfrac{p_w}{2}\right)^2}}. \tag{11}$$

Here, ρ_b is the effective resistivity because of barrier, ρ_0 is the bulk resistivity at a given temperature, A_b is the area occupied by the barrier, $A.R.$ is the aspect ratio, and p_w is the horizontal pitch of the interconnect. From the above equation, it is obvious that as A_b increases, ρ_b increases.

3) Simulation of Surface Scattering and Barrier Thickness Effects on Cu Resistivity: The resistivities for ITRS dictated future interconnects are evaluated in the light of above effects. The methodology for extracting future realistic resistivities using various barrier deposition technologies, operating temperatures and P values is as follows. SPEEDIE

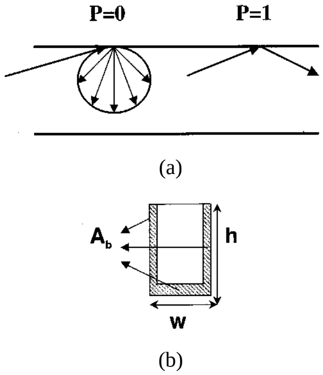


Fig. 5. Illustration of (a) diffuse and specular surface scattering and (b) effective cross section reduction of copper interconnects due to barrier.

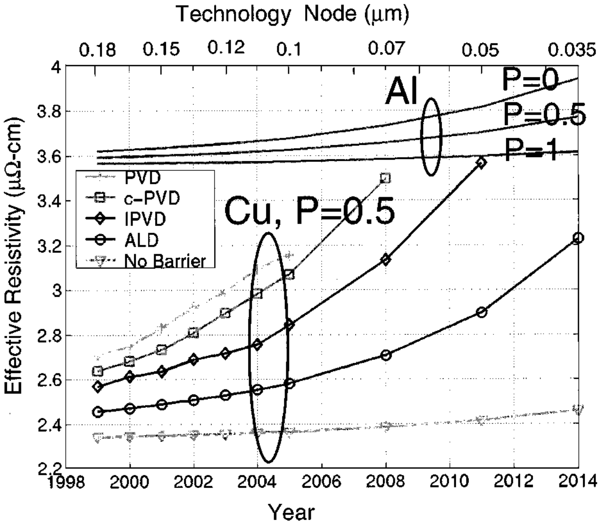


Fig. 6. Effective resistivity of Cu lines (calculated with both scattering and barrier effects at 100 °C and for $P = 0.5$) as a function of technology node (dimensions) based on ITRS, for various barrier deposition technologies. Resistivity of Al interconnects are also shown for different values of the scattering parameter, P.

(Stanford Profile Emulator for Etching and Deposition in IC Engineering) [30] was used to simulate the barrier profile for different deposition technologies, which was then used to extract the area consumed by the barrier. The simulations were performed on dimensions specified in the ITRS. The deposition time in the simulator was varied for each of the simulated geometries to obtain two conditions corresponding to a 5-nm and 10-nm minimum barrier thickness, respectively. The actual minimum barrier thickness in the future would be dictated by the quality of the barrier.

The effects of various barrier deposition technologies such as atomic layer deposition (ALD), ionized physical vapor deposition (IPVD), collimated physical vapor deposition (c-PVD), and simple physical vapor deposition (PVD) are quantified in Fig. 6. The value of P was taken to be 0.5 [28], the temperature was 100 °C, and the minimum barrier thickness was chosen to be 10 nm in this figure. The resistivity of aluminum interconnects calculated with these physical effects is also shown in Fig. 6 to demonstrate the diminishing advantage of copper over aluminum, for future

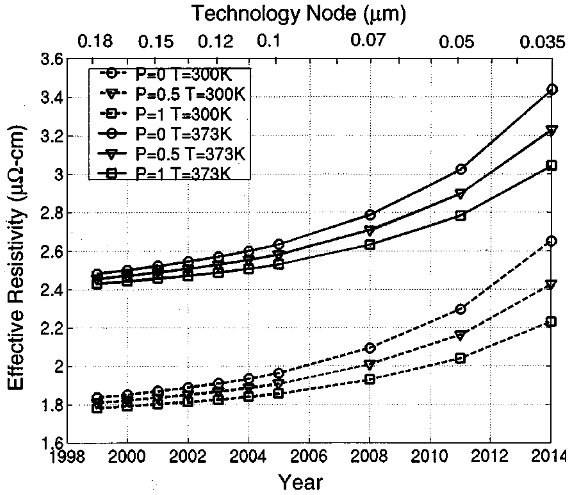


Fig. 7. Copper resistivity of global level interconnects versus year using most conformal technology (ALD), barrier thickness = 10 nm for various values of P and temperature.

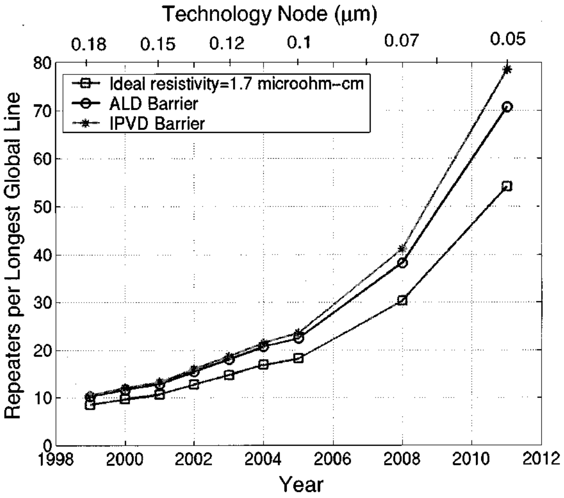


Fig. 8. Number of repeaters per longest global line as a function of technology nodes based on ITRS for different barrier technologies.

scaled dimensions. This occurs primarily because unlike copper, aluminum interconnects do not require barrier on all four surfaces, and because its intrinsically higher bulk resistivity compared to that of copper makes the surface scattering effect less important at comparable dimensions. It can also be observed that more conformal deposition technologies such as ALD lead to a much slower rise in resistivity in future as the barrier deposited using these technologies leads to a smaller barrier cross sectional area consumption.

Fig. 7 shows the effect of interface quality, characterized through the parameter P, on future global wire resistivity. Parameter P and temperature are varied. The minimum barrier thickness is 10 nm, and the deposition technology is assumed to be the best available, i.e., atomic layer deposition (ALD). From this figure, it is obvious that, under realistic wire temperature of 100 °C and P value of 0.5 [28], resistivities as high as 2.9 $\mu\Omega$·cm will be obtained in the year 2010. This gives about a 70% increase over the nominal bulk copper resistivity (1.7 $\mu\Omega$·cm) at room temperature. Under same conditions, simulations revealed resistivities of 3.45 $\mu\Omega$·cm and 3.95 $\mu\Omega$·cm, for the semiglobal and local interconnects, respectively. It was also found that using any other less conformal barrier deposition technology such as ionized physical vapor deposition (IPVD) or collimated PVD (c-PVD), the resistivity values for local and semiglobal interconnects become higher than aluminum technology for the same dimensions, in about a decade.

The incorporation of aforementioned technological constraints on copper resistivity leads to more realistic and higher line resistances per unit length, than that predicted using bulk Cu resistivity. As a result, the optimal interconnect length [$l_{\rm opt}$ in (3)] between repeater decreases, leading to an increase in the total number of repeaters per line. An example of this impact is shown in Fig. 8. This figure depicts the number of optimally spaced repeaters that minimize the line delay versus future years in a chip edge long global line. The P value was 0.5, the barrier thickness and temperature were 10 nm and 100 °C, respectively, for these calculations. As seen from this figure, the number of repeaters would be underestimated to be around 50 per line instead of, for example, about 80 using IPVD barrier, at the 0.05-μm technology node. Such an underestimation could lead to a significant underprediction of the area consumed by repeaters and the power dissipated by them.

The above discussion quantitatively illustrates that in the near future the material resistivity of copper will rise to prohibitively high values even with the best available deposition and barrier technologies. At some point, local and semiglobal tier effective resistivity of copper could become higher than the corresponding resistivity for aluminum for same ITRS dictated dimensions. This will make the interconnect delay even higher than that depicted in Fig. 1 where bulk resistivity was assumed. This calls for a pressing need to develop Cu technologies with smooth surfaces along the wire perimeter to maximize elastic scattering of electrons such that the value of P in (10) may nearly equal one. There is also an urgent need for the development of barrierless Cu technology and for lowering the operating wire temperature by going with higher thermal conductivity packaging materials and/or with a radically new chip cooling mechanism.

C. Deep-Submicrometer Interconnect Effects on VLSI Design

Interconnects in deep-submicrometer VLSI present many challenges to the existing CAD methodologies and tools [31]. As shown in Fig. 9, typically the design process starts at the *behavioral level*, which consists of a description of the system and what it is supposed to do (usually in C++ or Java programming languages). This description is then transformed to a *Register Transfer Level* (RTL) description using either the VHDL or Verilog languages. This is then transformed to a logic level structural representation (a netlist consisting of logic gates, flip-flops, latches, etc.) by a process called *logic synthesis*. Finally, a physical mask-level

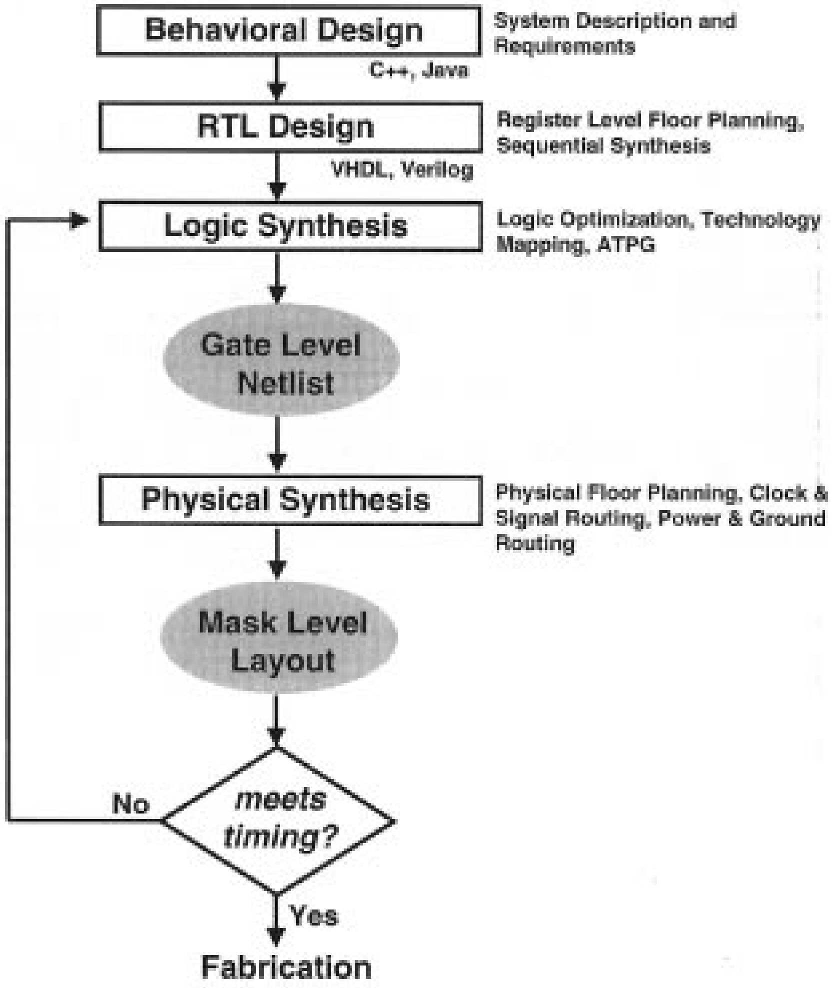

Behavioral Design
System Description and Requirements
C++, Java
RTL Design
Register Level Floor Planning, Sequential Synthesis
VHDL, Verilog
Logic Synthesis
Logic Optimization, Technology Mapping, ATPG
Gate Level Netlist
Physical Synthesis
Physical Floor Planning, Clock & Signal Routing, Power & Ground Routing
Mask Level Layout
meets timing?
No
Yes
Fabrication

Fig. 9. Typical VLSI design process flow.

layout file (such as GDSII) is generated using a process called *physical synthesis*, which generates the detailed floorplanning, placement, and routing.

For deep-submicrometer technologies, a significant manifestation of the interconnect effects arises in the form of *timing closure* problem, which is caused by the inability of logic synthesis (optimization) tools to account for logic gate interconnect loading with adequate precision prior to physical synthesis. This situation is illustrated in Fig. 9. Traditionally, logic optimization is performed using *wire-load models* that statistically predict the interconnect load capacitance as a function of the fanout based on technology data and design legacy information [32]. The wire-load model includes an average delay due to the interconnect connecting the output of a gate to the other gate inputs. This approach suffices if the interconnect delays (after physical synthesis) remain negligible. However, as shown in Fig. 1, for deep-submicrometer technologies, the interconnect delay associated with long global wires is a dominant fraction of the overall delay. As a result, the wire-load models become inaccurate for long and high fanout nets. This deficiency in the existing CAD flows causes a serious dilemma in deep-submicrometer designs. On one hand, the increasing circuit complexity (number of gate counts) requires the CAD methodologies to adopt higher levels of abstraction (*block-based* and *hierarchical design*) to simplify and accelerate the design process, while on the other hand, increasing interconnect delays and other interconnect related effects such as crosstalk, make it difficult for existing CAD tools to obtain timing convergence for the design blocks within a reasonable number of iterations.

It is instructive to note that the magnitude of the interconnect problem for future deep-submicrometer ICs with greater than 10^8 gates (269 million, at the 50-nm node [1]) cannot be fully comprehended by analyzing the impact of scaling on *module-level* designs (with around 50K gates) using standard wire-load models for *average-length* interconnects. This type of analysis, which has led some researchers to claim that interconnect delay is not a problem [33], is not quite adequate for deep-submicrometer VLSI. This is due to the fact that for deep-submicrometer designs, even if the average-length wires within small module-level blocks continue to produce wire delays such that the module-level designs can be individually handled by the traditional wire-load models, the number of such blocks required to realize the entire design would explode resulting in longer and more numerous interblock interconnects (*global wires*). Unfortunately, it is these long global wires that are mainly responsible for the increasing interconnect delays as pointed out in an earlier section. Furthermore, given the various technology and material effects arising due to interconnect scaling illustrated earlier, even some of the intramodule wire delays can become unexpectedly large contrary to usual assumptions as in [34]. In order to mitigate the interconnect scaling problems, some researchers have proposed combined *wire planning* and *constant-delay synthesis* [11], [35]. This methodology is also based on a block-based design where the interblock wires are planned or constructed and the remaining wires are handled through the constant-delay synthesis [36] within the blocks. The difficulty with this method is that if the blocks are sufficiently large then the timing convergence problem persists. In contrast, if they are allowed to remain relatively small such that the constant-delay synthesis with wire-load models works, then the number of such blocks becomes so large that the majority of the wiring will be global and the physical placement of these point-like blocks becomes absolutely critical to the overall wire planning quality, which represents a daunting physical design problem. Another work proposed an interconnect fabric based on a ground–signal–ground wire grid to make wire loads more predictable [37]. However, this technique results in significant area penalty.

Apart from the increasing signal transmission delays of global signals relative to the clock period and gate delay, there are signal integrity concerns arising from electromagnetic interference such as interconnect crosstalk, wire-substrate coupling and inductance effects, as well as voltage (IR) drop effects and signal attenuation induced intersymbol interference (ISI). Also, electromigration and thermal effects in interconnects impose severe restrictions on signal, bus, and power/ground line scaling [26], [38].

Thus, it can be concluded that the interconnect problem in deep-submicrometer VLSI design is not only going to get *bigger* due to ever increasing chip complexity, but will also get *worse* due to material and technology limitations discussed above. Hence, in the near future, existing design methodologies and CAD tools may not be adequate to deal with the wiring problem both at the modular and global levels.

Greater performance and greater complexity at lower cost are the drivers behind large-scale integration. In order to maintain these driving forces, it is necessary to find a way to keep increasing the number of devices on a chip, yet limit or even decrease the chip size to keep interconnect delay from affecting chip performance. A decrease in chip size will also assist in maximizing the number of chips per wafer; thus maintaining the trend of decreasing cost function. Therefore, innovative solutions beyond mere materials and technology changes are required to meet future IC performance goals [39]. We need to think beyond the current paradigm of design architecture.

D. System-on-a-Chip Designs

System-on-a-chip (SoC) is a broad concept that refers to the integration of nearly *all aspects* of a system design on a single chip [40], [41]. These chips are often mixed-signal and/or mixed-technology designs, including such diverse combinations as embedded DRAM, high-performance and low-power logic, analog, RF, programmable platforms (software, FPGAs, Flash, etc.), as schematically illustrated in Fig. 10. They can also involve more esoteric technologies like microelectromechanical systems (MEMS), bioelectronics, microfluidics, and optical input–output (I/O) devices. SoC designs are often driven by the ever-growing demand for increased system functionality and compactness at minimum cost, power consumption, and time to market. These designs form the basis for numerous novel electronic applications in the near future in areas such as wired and wireless multimedia communications including high-speed internet applications, medical applications including remote surgery, automated drug delivery, and noninvasive internal scanning and diagnosis, aircraft/automobile control and safety, fully automated industrial control systems, chemical and biological hazard detection, and home security and entertainment systems, to name a few.

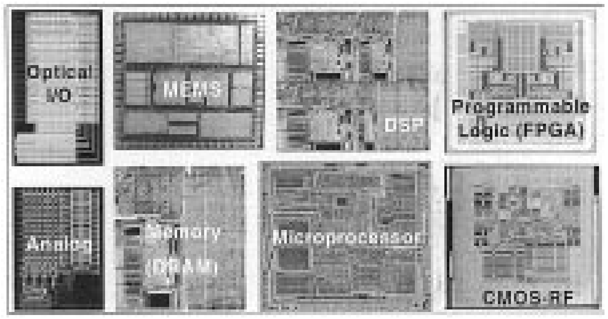


Fig. 10. Schematic of an SoC design using a planar (2-D) IC.

There are several challenges to effective SoC designs. Large-scale integration of functionalities and disparate technologies on a single chip dramatically increases the chip area, which necessitates the use of numerous long global wires. These wires can lead to unacceptable signal transmission delays and increase the power consumption by increasing the total capacitance that needs to be driven by the gates. Also, integration of disparate technologies such as embedded DRAM, logic, and passive components in SoC applications introduces significant complexity in materials and process integration. Furthermore, the noise generated by the interference between different embedded circuit blocks containing digital and analog circuits becomes a challenging problem. Additionally, although SoC designs typically reduce the number of I/O pins compared to a system assembled on a printed circuit board (PCB), several high-performance SoC designs involve very high I/O pin counts, which can increase the cost/chip. Finally, integration of mixed signals and mixed technologies on a single die requires novel design methodologies and tools, with design productivity being a key requirement.

E. 3-D Architecture

Three-dimensional integration (schematically illustrated in Fig. 11) to create multilayer Si ICs is a concept that can significantly improve deep-submicrometer interconnect performance, increase transistor packing density, and reduce chip area and power dissipation [42]. Additionally, 3-D

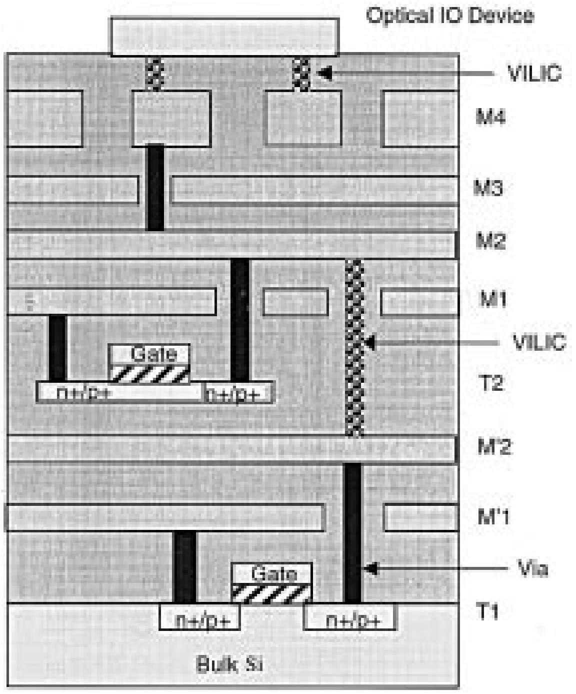


Fig. 11. Schematic representation of 3-D integration with multilevel wiring network and VILICs. T1: first active layer device, T2: second active layer device, Optical I/O device: third active layer I/O device. M′1 and M′2 are for T1, M1 and M2 are for T2. M3 and M4 are shared by T1, T2, and the I/O device.

ICs can be very effective vehicles for large-scale on-chip integration of different systems.

In the 3-D design architecture, an entire (2-D) chip is divided into a number of blocks, and each block is placed on a separate layer of Si that are stacked on top of each other. Each Si layer in the 3-D structure can have multiple layers of interconnect. These layers are connected together by vertical interlayer interconnects (VILICs) and common global interconnects as shown in Fig. 11. The 3-D architecture offers extra flexibility in system design, placement, and routing. For instance, logic gates on a critical path can be placed very close to each other using multiple active layers. This would result in a significant reduction in RC delay and can greatly enhance the performance of logic circuits. Also, the negative impact of deep-submicrometer interconnects on VLSI design discussed earlier can be reduced significantly by eliminating the long *global wires* that realize the interblock communications by vertical placement of logic blocks connected by short VILICs.

Furthermore, the 3-D chip design technology can be exploited to build SoCs by placing circuits with different voltage and performance requirements in different layers. The 3-D integration would significantly alleviate many of the problems outlined in the previous section for SoCs fabricated on a single Si layer. Three-dimensional integration can reduce the wiring, thereby reducing the capacitance, power dissipation, and chip area and therefore improve chip performance. Additionally, the digital and analog components in the mixed-signal systems can be placed on different Si layers thereby achieving better noise performance due to lower electromagnetic interference between such circuit blocks. From an integration point of view, mixed-technology assimilation could be made less complex and more cost-effective by fabricating such technologies on separate substrates followed by physical bonding. Also, synchronous clock distribution in high-performance SoCs can be achieved by employing optical interconnects and I/Os at the topmost Si layer (as illustrated in Fig. 11). Three-dimensional integration of optical and CMOS circuitry have been demonstrated in the past [43]. A schematic diagram of a 3-D chip is shown in Fig. 12 with logic, memory (DRAM), analog, RF, and optical I/O circuits on different active layers.

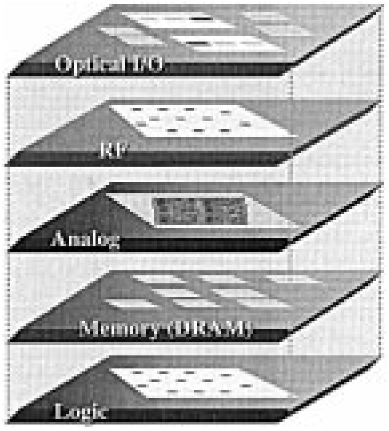


Fig. 12. Schematic of a 3-D chip showing integrated heterogeneous technologies.

II. Scope of This Study

A 3-D solution at first glance seems an obvious answer to the interconnect delay problem. Since chip size directly affects the interconnect delay, therefore by creating a second active layer, the total chip footprint can be reduced, thus shortening critical interconnects and reducing their delay. However, in today's microprocessors, the chip size is not just limited by the cell size, but also by how much metal is required to connect the cells. The transistors on the silicon surface are not actually packed to maximum density, but are spaced apart to allow metal lines above to connect one transistor or one cell to another. The metal required on a chip for interconnections is determined not only by the number of gates, but also by other factors such as architecture, average fan-out, number of I/O connections, routing complexity, etc. Therefore, it is not obvious that by using a 3-D structure the chip size will be reduced.

In this paper, the possible effects of 3-D integration of large logic circuits on key metrics such as chip area, power dissipation, and performance have been quantified by modeling the optimal distribution of the metal interconnect lines. To better understand how a 3-D design will affect the amount of metal wires required for interconnections, a stochastic wire-length distribution methodology derived for a 2-D IC in [44] has been modified for 3-D ICs to quantify effects on interconnect delay. Unlike previous work [45], wire-pitch limited chips are considered.

The results obtained in Section III indicate that, when critically long metal lines that occupy lateral space are replaced by short VILICs to connect logic blocks on different Si layers, a significant chip-area reduction can be achieved. VILICs are found to be ultimately responsible for this improvement. The assumption made here is that it is possible to divide the microprocessor into different blocks such that they can be placed on different levels of active silicon. In Section IV, important concerns in 3-D ICs such as power dissipation have been analyzed. It is demonstrated that advancement in IC cooling technology will be necessary for maximizing 3-D circuit performance.

Throughout this work, no differences were assumed in the performance or the properties of the individual devices on any layer. Also, the treatment is independent of the 3-D technology used. However, even if the properties of the devices on the upper Si layers are different, these layers can be used for memory devices or repeaters. Some of these applications are discussed in Section V. Finally, in Section VI, various technology options for fabricating 3-D ICs have been outlined. For simplicity, technology effects on metal wire resistivity discussed earlier in Section I-B are ignored in the proceeding analysis (for both 2-D and 3-D ICs), where bulk resistivity is assumed.

III. Area and Performance Estimation of 3-D ICs

We now present a methodology that can be used to provide an initial estimate of the area and performance of high-speed logic circuits fabricated using multiple silicon layer IC technology. The approach is primarily based on the empirical relationship known as Rent's Rule [46]. Rent's Rule correlates the number of signal input and output (I/O) pins T, to the number of gates N, in a random logic network and is given by the following expression:

$$T = kN^p. \tag{12}$$

Here, k and p denote the average number of fan-out per gate and the degree of wiring complexity (with $p = 1$ representing the most complex wiring network), respectively, and are empirically derived as constants for a given generation of ICs. The underlying assumption of this methodology is based upon the recursive application of Rent's Rule throughout an entire logic system.

To illustrate the application of this methodology, a logic system can be considered, the complexity of which necessitates that the final chip area is determined by the wiring requirement. Such ICs are considered wire-pitch limited, which is assumed throughout this work and considered valid for high-performance ICs. The wiring network is assumed to be a distribution of connecting wires ranging from the very short (to connect closest-neighbor logic gates, or intrablock connections), to the very long (for long-distance across-chip, or interblock communications). Furthermore, the performance of this logic system is assumed to be determined solely by this wiring network and specifically by the longest wires in the wiring network, as these represent the communications bottleneck due to their higher delay as compared to the shorter wires.

The problem of estimating the chip performance is then reduced to one of estimating this interconnect wiring distribution from which it is possible to determine a chip area and thus performance. To determine all the shortest wires in a logic system, the recursive property of Rent's Rule is used, where the logic system is divided into logic gates and Rent's Rule is applied to the interconnects between closest neighbor gates. This determines the number of interconnections between the closest logic gates. The longer wires are similarly determined by clustering the logic gates in growing numbers until the longest interconnects are found. A summary of this methodology is given as follows and more details can be found in [47].

A. 2-D and 3-D Wire-Length Distributions

The wire-length distribution can be described by $i(l)$, an interconnect density function (i.d.f.), or by $I(l)$, the cumulative interconnect distribution function (c.i.d.f.) which gives the total number of interconnects that have length less than or equal to l (measured in gate pitches) and is defined as

$$I(l) = \int_1^l i(x)\,dx \tag{13}$$

where x is a variable of integration representing length and l is the length of the interconnect in gate pitches. The derivation of the wire-length distribution in an IC is based on Rent's Rule. To derive the wire-length distribution $I(l)$ of an integrated circuit, the latter is divided up into N logic gates, where N is related to the total number of transistors N_t in an integrated circuit by $N = N_t/\phi$, where ϕ is a function of the average fan-in ($f.i.$) and fan-out ($f.o.$) in the system [48]. The gate pitch is defined as the average separation between the logic gates and is equal to $\sqrt{A_c/N}$where A_c is the area of the chip.

We first review the stochastic approach used for estimating the wire-length distribution of a 2-D chip and then modify it for 3-D chips. In order to derive the complete wire-length distribution for a chip, the stochastic wire-length distribution of a single gate must be calculated. The methodology is illustrated in Fig. 13. The number of connections from the single logic gate in Block A to all other gates that are located at a distance of l gate pitches is determined using Rent's Rule. The gates shown in Fig. 13 are grouped into three distinct but adjacent blocks (A, B, and C), such that a closed single path can encircle one, two, or three of these blocks. The number of connections between Block A and Block C is calculated by conserving all I/O terminals for blocks, A, B, and C, which states that terminals for blocks A, B, and C are either interblock connections or external system connections.

Hence, applying the principle of conservation of I/O pins to this system of three logic blocks, shown in Fig. 13, gives

$$\begin{aligned} &T_A + T_B + T_C \\ &\quad = T_{A\text{-to-}C} + T_{A\text{-to-}B} + T_{B\text{-to-}C} + T_{ABC} \end{aligned} \tag{14}$$

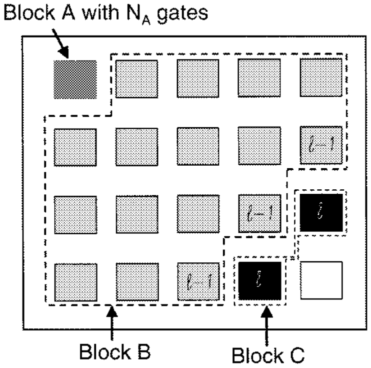


Fig. 13. Schematic view of logic blocks used for determining wire length distribution (adopted from [44]).

where T_A, T_B, and T_C are the number of I/Os for blocks A, B, and C, respectively. $T_{A\text{-to-}C}$, $T_{A\text{-to-}B}$, and $T_{B\text{-to-}C}$ are the numbers of I/Os between blocks A and C, blocks A and B, and between blocks B and C, respectively. T_{ABC} represents the number of I/Os for the entire system comprising all the three blocks. From conservation of I/Os, the number of I/Os between adjacent blocks A and B, and between adjacent blocks B and C can be expressed as

$$T_{A\text{-to-}B} = T_A + T_B - T_{AB} \tag{15}$$

$$T_{B\text{-to-}C} = T_B + T_C - T_{BC}. \tag{16}$$

Substituting (15) and (16) into (14) gives

$$T_{A\text{-to-}C} = T_{AB} + T_{BC} - T_B - T_{ABC}. \tag{17}$$

Now the number of I/O pins for any single block or a group of blocks can be calculated using Rent's Rule. If we assume that N_A, N_B, and N_C are the number of gates in blocks A, B, and C, respectively, then it follows that

$$T_B = k(N_B)^p \tag{18}$$

$$T_{AB} = k(N_A + N_B)^p \tag{19}$$

$$T_{BC} = k(N_B + N_C)^p \tag{20}$$

$$T_{ABC} = k(N_A + N_B + N_C)^p \tag{21}$$

where $N = N_A + N_B + N_C$. Substituting (18)–(21) into (17) gives

$$\begin{aligned} T_{A\text{-to-}C} = {} & k[(N_A + N_B)^p - (N_B)^p + (N_B + N_C)^p \\ & - (N_A + N_B + N_C)^p]. \end{aligned} \tag{22}$$

The number of interconnects between Block A and Block C ($I_{A\text{-to-}C}$) is determined using the relation

$$I_{A\text{-to-}C} = \alpha k(T_{A\text{-to-}C}). \tag{23}$$

Here, α is related to the average fan-out ($f.o.$) by

$$\alpha = \frac{f.o.}{1 + f.o.}. \tag{24}$$

Equation (23) can be used to calculate the number of interconnects for each length l in Fig. 13 in the range from one gate pitch to $2\sqrt{N}$ gate pitches, to generate the complete stochastic wire-length distribution for the logic gate in Block A. In the following step, Block A is removed from the system of gates for calculating the remaining wiring distribution in order to prevent multiplicity in interconnect counting. The same process is repeated for all gates in the system. Finally, the wire-length distributions for the individual gates are superimposed to generate the total wire-length distribution of the chip with N gates.

Davis *et al.* developed a closed-form analytical expression of the wire-length distribution for a 2-D IC [44], which can be expressed as

$$I(l) = I_{\text{total}} P(l) \tag{25}$$

where I_{total} is the total number of interconnects in a system derived from Rent's Rule as

$$I_{\text{total}} = \alpha k N \left(1 - N^{p-1}\right). \tag{26}$$

Here, $P(l)$ is the cumulative distribution function that describes the total probability that a given interconnect length is less than or equal to l and is given by the following expressions:

$$\begin{aligned} P(l) = {} & \frac{1}{2N(1-N^{p-1})} \\ & \cdot \Gamma\left(\frac{l^{2p}-1}{6p} + 2\sqrt{N}\,\frac{-l^{2p-1}+1}{(2p-1)} - N\,\frac{-l^{2p-2}+1}{(p-1)}\right) \end{aligned} \tag{27}$$

for $1 \leq l \leq \sqrt{N}$, and as shown in (28) at the bottom of the next page, for $\sqrt{N} \leq l \leq 2\sqrt{N}$. The factor Γ is defined by (29), shown at the bottom of the next page. Substituting (26)–(29) into (25) gives the closed-form expressions for the total wire-length distribution as follows:

$$\begin{aligned} & I(l) \\ & = \frac{\alpha k}{2}\Gamma\left(\frac{l^{2p}-1}{6p} + 2\sqrt{N}\,\frac{-l^{2p-1}+1}{(2p-1)} - N\,\frac{-l^{2p-2}+1}{(p-1)}\right) \end{aligned} \tag{30}$$

for $1 \leq l \leq \sqrt{N}$ and (31), shown at the bottom of the next page.

The simple use of Rent's Rule above applies to 2-D ICs and requires adaptation for a valid application to 3-D ICs. For the case of 3-D ICs, different blocks can be physically placed on different silicon layers and connected to each other using VILICs. The area saving by using VILICs can be computed by modifying Rent's Rule suitably. For generality, we first analyze the case where n silicon layers are available. The application to two-layer ($n = 2$) case is straightforward. An N gate IC design is divided into N/n gate blocks. It is assumed that the routing algorithm and overall logic style is the same for both layers. This ensures that Rent's Constant k and Rent's Exponent p are the same for both layers. Applying Rent's Rule to all the layers, we have

$$T = kN^p = \left(\sum_{i=1}^{n} T_i\right) - T_{\text{int}} = nk\left(\frac{N}{n}\right)^p - T_{\text{int}}. \tag{32}$$

Here, T is the number of I/Os for the entire design, T_i represents the number of I/Os for each layer, and T_{int} represents the total number of I/O ports connecting the n-layers. Hence, it follows that

$$T_{\text{int}} = n\left(1 - n^{p-1}\right) k \left(\frac{N}{n}\right)^p$$

and

$$T_{\text{ext},\, i} = T_i - \frac{T_{\text{int}}}{n} = kn^{p-1}\left(\frac{N}{n}\right)^p. \qquad (33)$$

Here, $T_{\text{ext},\, i}$ is the average number of external I/O ports per layer i. Comparing (33) with Rent's Equation, for each layer, i.e., $T = k(N/n)^p$, we find that for each layer

$$\begin{aligned} k_{\text{eff, int}} &= k\left(1 - n^{p-1}\right) \\ k_{\text{eff, ext}} &= kn^{p-1} \end{aligned} \qquad (34)$$

where $k_{\text{eff, int}}$ is the effective number of I/Os per gate used for interlayer connections and $k_{\text{eff, ext}}$ is the effective number of I/Os per gate used for external I/O connections.

Extending this analysis to two-layer ($n = 2$) 3-D ICs [Fig. 14(a)], we have

$$T = kN^p = T_1 + T_2 - T_{\text{int}} = 2k\left(\frac{N}{2}\right)^p - T_{\text{int}}. \qquad (35)$$

Since each layer will have ($T_{\text{int}}/2$) dedicated I/O ports for connection to the other layer, we have

$$k_{\text{eff, ext}} = k2^{p-1}$$

and

$$k_{\text{eff, int}} = k\left(1 - 2^{p-1}\right). \qquad (36)$$

Now the wire-length distribution analysis discussed above can be extended to 3-D ICs using the modified values of k for each layer. Fig. 15 shows the wire-length distributions for 2-D, and 3-D ICs with two active layers, using ITRS data for the high-performance 50-nm technology node. It can be observed that the wiring requirement is significantly reduced for the global wires in 3-D ICs. This is due to the fact that these long wires have been converted to short VILICs as schematically illustrated in Fig. 14(b).

For all the 2-D calculations presented in this paper, the values of k and p were chosen for each technology node such that the results fit the projected chip area provided in the ITRS. When applied to 3-D calculations the values of k and N were subjected to the 3-D transformations described above. Rent's Exponent p remained constant without transformation between 2-D and 3-D as discussed above.

B. Estimating 2-D and 3-D Chip Area

The analyses described in this work are performed on integrated circuits that are wire-pitch limited in size. The area required by the wiring network in such ICs is assumed to

$$P(l) = \frac{1}{2N\left(1 - N^{p-1}\right)}\,\Gamma \left(\begin{array}{l} \dfrac{N^{2p-1}}{6p} + 2\sqrt{N}\,\dfrac{-N^{2p-1}+1}{2p-1} - N\,\dfrac{-N^{2p-2}+1}{p-1} \\ +\dfrac{1}{3}\left(\begin{array}{l} -8N^{3/2}\,\dfrac{-l^{2p-3}+N^{p-(3/2)}}{2p-3} + 6N\,\dfrac{-l^{2p-2}+N^{p-1}}{p-1} \\ -6\sqrt{N}\,\dfrac{-l^{2p-1}+N^{p-(1/2)}}{2p-1} + \dfrac{-l^{2p}+N^{p}}{2p} \end{array} \right) \end{array} \right) \qquad (28)$$

$$\Gamma = \frac{2N\left(1 - N^{p-1}\right)}{\left(-N^p\,\dfrac{1+2p-2^{2p-1}}{p(2p-1)(p-1)(2p-3)} - \dfrac{1}{6p} + \dfrac{2\sqrt{N}}{2p-1} - \dfrac{N}{p-1}\right)} \qquad (29)$$

$$I(l) = \frac{\alpha k}{2}\,\Gamma \left(\begin{array}{l} \dfrac{N^{2p-1}}{6p} + 2\sqrt{N}\,\dfrac{-N^{2p-1}+1}{2p-1} - N\,\dfrac{-N^{2p-2}+1}{p-1} \\ +\dfrac{1}{3}\left(\begin{array}{l} -8N^{3/2}\,\dfrac{-l^{2p-3}+N^{p-(3/2)}}{2p-3} + 6N\,\dfrac{-l^{2p-2}+N^{p-1}}{p-1} \\ -6\sqrt{N}\,\dfrac{-l^{2p-1}+N^{p-(1/2)}}{2p-1} + \dfrac{-l^{2p}+N^{p}}{2p} \end{array} \right) \end{array} \right) \qquad (31)$$

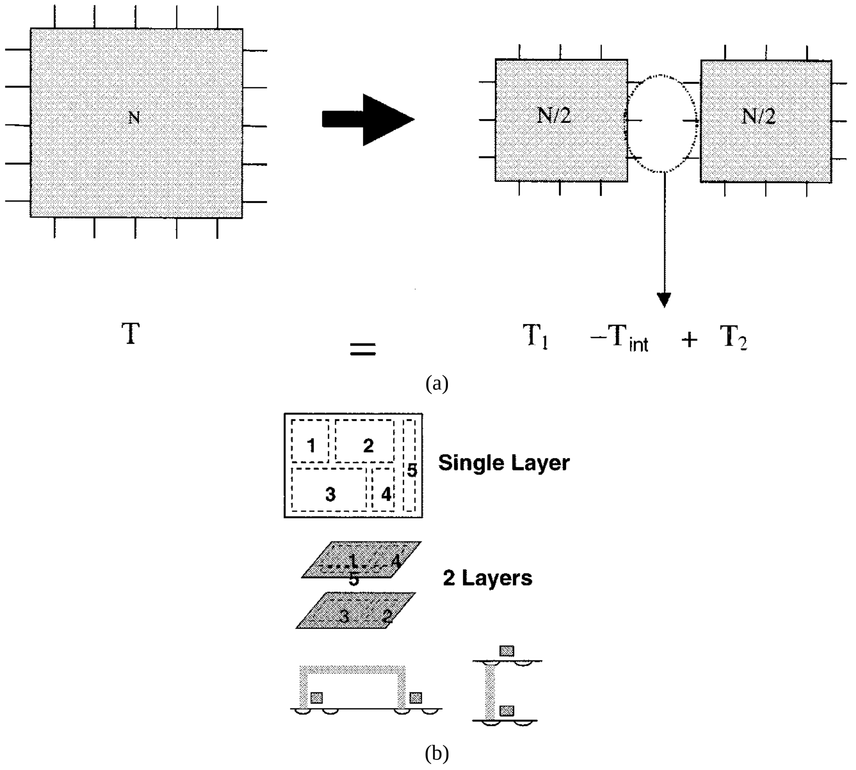


Fig. 14. Schematic to illustrate (a) conservation of total number of external I/O ports for maintaining constant functionality of chip, and (b) two-layer 3-D chip with long horizontal interconnects replaced by short and vertical (VILICs) interconnects.

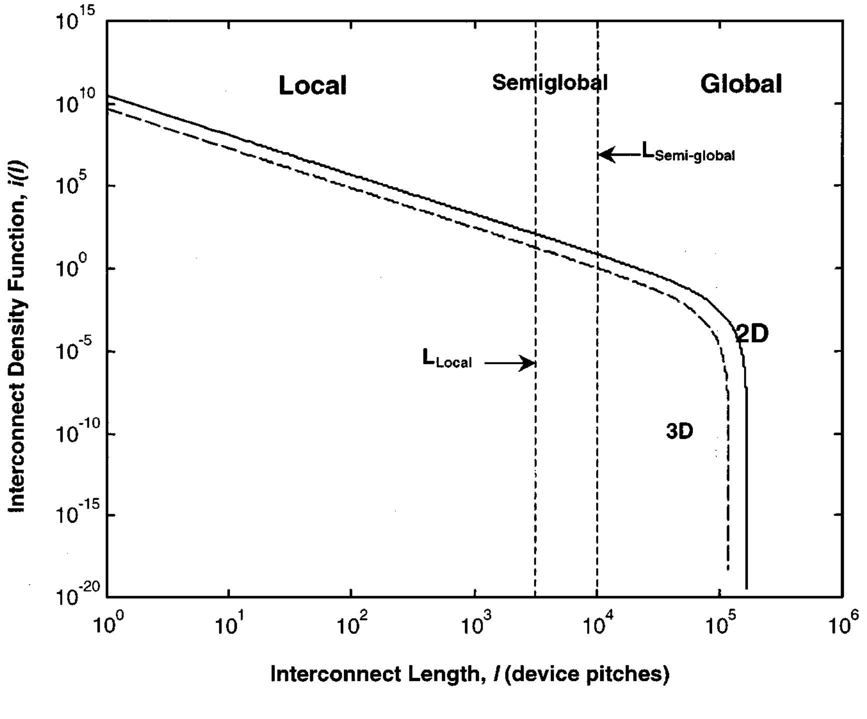


Fig. 15. Wire-length distributions for the 2-D and 3-D ICs shown in Fig. 14. 3-D significantly reduces requirement for longest wires. Metal tiers determined by $L_{\rm Local}$ and $L_{\rm Semi\text{-}global}$ boundaries as explained in the text.

be greater than the area required by the logic gates. For the purposes of minimizing silicon real estate and signal propagation delays, the wiring network is segmented into separate tiers that are physically fabricated in multiple layers. An interconnect tier is categorized by factors such as metal line pitch and cross-section, maximum allowable signal delay and communication mode (such as intrablock, interblock, power, or clocking). A tier can have more than one layer of metal interconnects if necessary, and each tier or layer is connected to the rest of the wiring network and the logic gates by vertical vias. The tier closest to the logic devices (referred to as the Local tier) is normally responsible for short-distance intra-block communications. Metal lines in this tier will normally be the shortest. They will also normally have the finest pitch. The tier furthest away from the device layer (referred to as the global tier) is responsible for long-distance across-chip interblock communications, clocking and power distribution. Since this tier is populated by the longest of wires, the metal pitch is the largest to minimize signal propagation delays. A typical modern IC interconnect architecture will define three wiring tiers: local, semiglobal, and global, spanning, for example, a total of 9–10 metallization layers as projected by ITRS for the 50-nm technology node. The semiglobal tier is normally responsible for interblock communications across intermediate distances. Fig. 16 shows a schematic of a three-tier interconnect structure.

Using a three-tier interconnection structure, the semiglobal tier pitch that minimizes the wire limited chip area was determined. The maximum interconnect length on any given tier was determined by the interconnect delay criterion [47]. (It is assumed $t_{\text{delay\_max}} = 0.25T$ for semiglobal and local wires, with T as the clock period. The maximum length of a wire in the global tier is assumed to be equal to the chip edge dimension.) The cross-sectional dimensions of the global wires are determined by using the delay criteria at $t_{\text{delay}} = 0.9T$ [47].

The area of the chip is determined by the total wiring requirement. In terms of gate pitch, the total area required by the interconnect wiring can be expressed as

$$A_{\text{required}} = \sqrt{\frac{A_c}{N}}\,(p_{\text{loc}}L_{\text{total\_loc}} + p_{\text{semi}}L_{\text{total\_semi}} + p_{\text{glob}}L_{\text{total\_glob}}) \qquad (37)$$

where

A_c chip area;
N number of gates;
p_{loc} local pitch;
p_{semi} semiglobal pitch;
p_{global} global pitch;
$L_{\text{total\_loc}}$ total length of the local interconnects;
$L_{\text{total\_semi}}$ total length of the semiglobal interconnects;
$L_{\text{total\_glob}}$ total length of the global interconnects.

The total interconnect length for any tier can be found by integrating the wire-length distribution within the boundaries that define the tier (see Fig. 15, where broken vertical lines define the boundaries). Hence, it follows that

$$L_{\text{total\_loc}} = \chi \int_{1}^{L_{\text{loc}}} li(l)\,dl \qquad (38)$$

$$L_{\text{total\_semi}} = \chi \int_{L_{\text{loc}}}^{L_{\text{semi}}} li(l)\,dl \qquad (39)$$

$$L_{\text{total\_glob}} = \chi \int_{L_{\text{semi}}}^{2\sqrt{N}} li(l)\,dl \qquad (40)$$

where χ is a correction factor that converts the point-to-point interconnect length to wiring net length [using a linear net model, $\chi = 4/(f.o. + 3)$]. The boundaries shown in Fig. 15 represent the length of the longest wire for each tier, L_{loc} for the local, L_{semi} for the semiglobal, and L_{global} for the global tier.

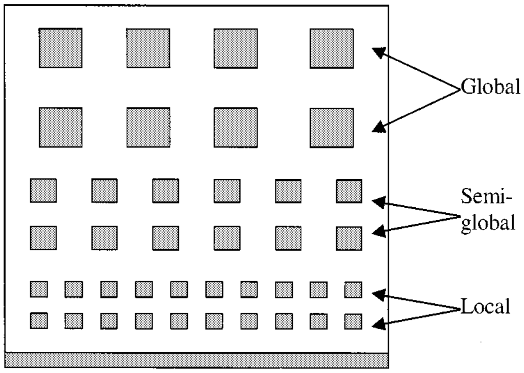


Fig. 16. Schematic of a three-tier interconnection structure.

We now present a modified analysis in terms of FO4 delay, discussed in Section I-A, to estimate optimal chip area. The main differences between this analysis and those in [42] and [46] arise from the fact that the delay used here is for an optimally buffered interconnect, given by (6), and has been expressed in terms of FO4 delay. By substituting (8) and (9) into (6), and using $\tau_d = \beta/f_c$, the length of the longest wire, L, and the pitch, p_w, for an arbitrary tier are related by the following expression:

$$\frac{\beta}{f_c} = \frac{\sqrt{0.4\rho\varepsilon_r\varepsilon_o(1 + 4\,A.R.^2)\,t_{\text{FO4}}}}{A.R.\,p_w}\sqrt{\frac{A_c}{N}}\,L \qquad (41)$$

where

β maximum delay fraction of clock period (25% for local and semiglobal, and 90% for global wires);
f_c clock frequency;
ρ resistivity of the interconnect metal;
ε_0 permittivity of free space;
ε_r relative permittivity of the dielectric material;
p_w horizontal wire pitch;
$A.R.$ wiring level aspect ratio;
t_{FO4} FO4 gate delay.

Equation (41) can be rearranged to solve for wire pitch or the length of the longest interconnect. The expressions for p_{global}, L_{semi} (which is a function of p_{semi}) and L_{loc} are given by

$$p_{\text{glob}} = \sqrt{\frac{A_c}{N}}\,\frac{L_{\text{glob}}}{A.R._{\text{glob}}}\,\frac{f_c}{\beta_{\text{glob}}} \cdot \sqrt{0.4\rho\varepsilon_r\varepsilon_o(1 + 4\,A.R._{\text{glob}}^2)t_{\text{FO4}}} \qquad (42)$$

$$L_{\text{semi}} = \frac{\beta_{\text{semi}}}{f_c}\,p_{\text{semi}}A.R._{\text{semi}}\sqrt{\frac{N}{A_c}} \cdot \frac{1}{\sqrt{0.4\rho\varepsilon_r\varepsilon_o(1 + 4\,A.R._{\text{semi}}^2)t_{\text{FO4}}}} \qquad (43)$$

$$L_{\text{local}} = \frac{\beta_{\text{local}}}{f_c}\,p_{\text{local}}A.R._{\text{local}}\sqrt{\frac{N}{A_c}} \cdot \frac{1}{\sqrt{0.4\rho\varepsilon_r\varepsilon_o(1 + 4\,A.R._{\text{local}}^2)t_{\text{FO4}}}}. \qquad (44)$$

Here, p_{loc} is assumed constant and equal to twice the minimum feature size. L_{global} is also assumed constant and equal to the chip die edge. Equation (43) for L_{semi} results in a nonunique set of possible solutions for A_c and p_{semi} which are determined numerically. The wire-pitch limited chip area (A_c) is calculated based on the condition that the total required wiring area (A_{required}) is equal to the total available area ($A_{\text{available}}$) in a multilevel network; hence, it follows that

$$A_{\text{available}} = A_c e_w n_{\text{levels}} = A_{\text{required}} \tag{45}$$

where e_w is the wiring efficiency factor that accounts for router efficiency and additional space needed for power and clock lines, and n_{levels} is the number of metal levels available for the multilevel network. For each possible solution of (43), new boundaries representing L_{loc} and L_{semi} are used with the wire-length distribution to find the new total area required by the interconnect wiring. From the total area required by the wiring, the chip area is estimated by dividing the interconnects among the required number of metal layers. The resulting chip areas are then plotted as a function of p_{semi} normalized to the constant local pitch. Three-dimensional chip areas are determined using the same analysis with the values of N and k transformed for 3-D accordingly.

C. Two Active Layer 3-D Circuit Performance

The above analysis is used to compare area and delay values for 2-D and 3-D ICs. The availability of additional silicon layers gives the designer extra flexibility in trading off area with delay. It is assumed that through technological advances, resistivity of Cu will be maintained at the bulk value. A number of different cases are discussed as follows.

1) Chip Area Minimization with Fixed Interconnect Delay: The model is applied to the microprocessor example shown in Table 2 for the 50-nm technology node [1] for the two cases where all gates are in a single layer (2-D) and where the gates are equally divided between two layers (3-D). In this calculation, VILICs are assumed to consume negligible area, interconnect line width is assumed to equal half the metal pitch at all times, and the total number of metal layers for 2-D and 3-D case was conserved. A key assumption for the geometrical construction of each tier of the multilevel interconnect network is that all cross-sectional dimensions are equal within that tier.

The possible solutions for A_c and p_{semi} resulting from the numerical solution of (43) are plotted for the high-performance IC (ITRS 50-nm technology node) in Fig. 17 which shows the possible chip areas with the normalized semiglobal tier pitch for a fixed operating frequency of 3 GHz. The solutions exhibit a minimum in A_c, which is taken to be the acceptable chip area. As p_{semi} increases from its value at the minimum A_c the semiglobal and global pitches increase resulting in a larger wiring requirement and thus a larger A_c. Furthermore, as p_{semi} increases, even longer wires can now satisfy the maximum delay requirement in the semiglobal tier. This results in global wires to be rerouted to the semiglobal tier, which in turn will require greater chip area. Under such circumstances, the semiglobal tier begins to dominate and determine the chip area. Conversely, as p_{semi} decreases from its value at the minimum A_c, the longer wires in the semiglobal tier no longer satisfy the maximum delay requirement of that tier and they need to be rerouted to the global tier where they can enjoy a larger pitch. The population of wires in the global tier increases and since these wires have larger cross sections they have a greater area requirement. Under such circumstances, the global tier begins to dominate and determine the chip area.

The curve for the 3-D case has a minimum similar to the one obtained for the 2-D case. It can be observed that the minimum chip area for the 3-D case is ~30% smaller than that of the 2-D case. Moreover, since the total wiring requirement is reduced (as shown in Fig. 15), the semiglobal tier pitch is reduced for the 3-D chip. This reduction in the semiglobal pitch increases the line resistance and the line-to-line capacitance per unit length. Hence, the same clock frequency, i.e., the same interconnect delay, is maintained by reducing the chip size. Ultimately, the significant reduction in chip area demonstrated by the 3-D results are a consequence of the fraction of wires that were converted from horizontal in 2-D

Table 2
Microprocessor Example (ITRS based 50-nm Technology Node)

PHYSICAL PARAMETER	VALUE
Number of Transistors, N_t	7053 million
Rent's Exponent, p	0.6
Rent's Coefficient, k	4.0
Minimum Feature Size, F	50 nm
Max number of wiring levels, n_{max}	9
Metal Resistivity, Copper	1.673 x 10^{-6} Ohm-cm
Dielectric Constant, Polymer	$\varepsilon_r = 1.5$
Wiring Efficiency Factor	0.4

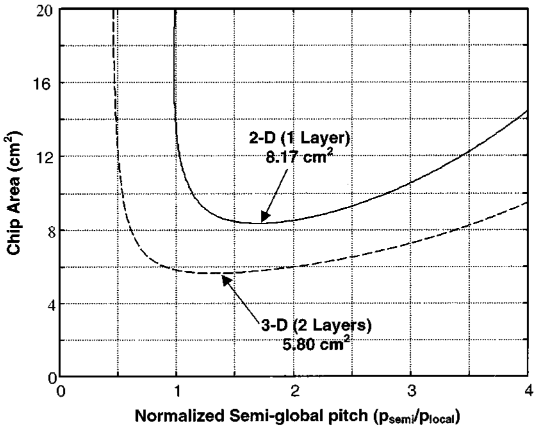


Fig. 17. Wire-limited chip area versus normalized semiglobal pitch (semiglobal pitch/local tier pitch) for 2-D and 3-D ICs at a fixed operating frequency of 3 GHz. As the normalized semiglobal pitch decreases, wires are rerouted to the global tiers, which have bigger pitch, and hence the chip area increases. Note that the estimated 2-D chip area of 8.17 cm^2 is also projected by ITRS for the 50-nm node. The number of metal layers for 2-D and 3-D ICs is nine (three per tier).

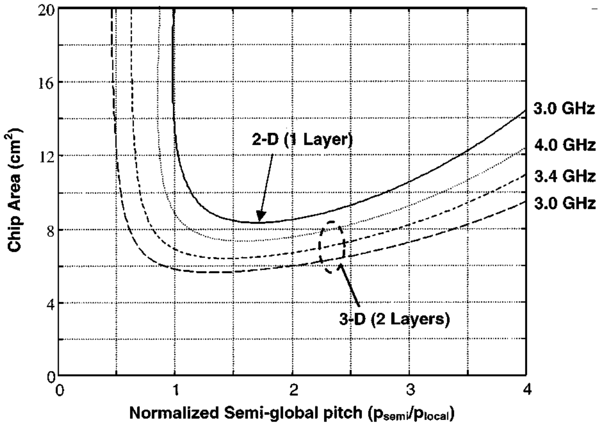


Fig. 18. 3-D chip operating frequency (performance) increases with increases in semiglobal wiring pitch. Chip area also increases but remains below the 2-D chip area. If 3-D chip area is made equal to 2-D chip area (=8.17 cm^2 at the 50-nm node), an operating frequency of 6 GHz can be obtained for the 3-D chip.

Table 3
Summary of Delay Performance Improvement for 3-D ICs. The Horizontal ILICs Differ from the Vertical ILICs in that they Consume Lateral Area

2-Layer Description of Inter-layer Interconnects (ILICs)	Delay Performance Improvement
Horizontal ILICs, equal global pitch	10%
Horizontal ILICs, equal chip area	17%
Vertical ILICs, equal global pitch	33%
Vertical ILICs, equal chip area	63%

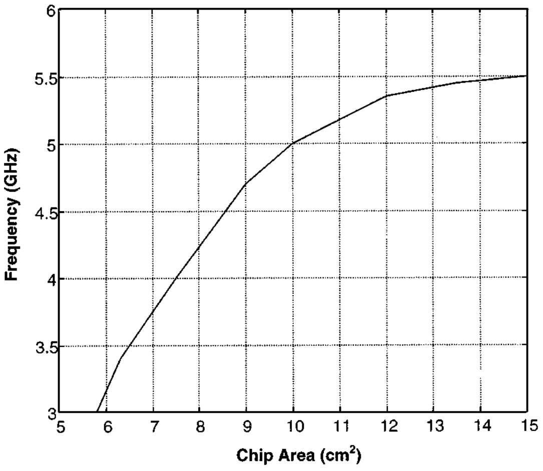


Fig. 19. Performance improvement with increasing chip area for a two-layer 3-D IC. Chip area is increased due to increasing wire pitch.

to vertical VILICs in 3-D. It is assumed that the area required by VILICs is negligible.

These results demonstrate, with the given assumptions, that a 3-D IC can operate at the same performance level, as measured by the longest wire delay, as its 2-D counterpart while using up about 30% less silicon real estate. However, it is possible for 3-D ICs to achieve greater performance than their 2-D counterparts by reducing the interconnect impedance at the price of increased chip area as discussed next.

2) Increasing Chip Area and Performance: 3-D IC performance can be enhanced to exceed the performance of 2-D ICs by improving interconnect delay. This is achieved by increasing the wiring pitch, which causes a reduction in resistance and line-to-line capacitance per unit length. The effect of increasing p_{semi} and p_{global} on the operating frequency and A_c is shown in Fig. 18. This illustrates how the optimum semiglobal pitch (i.e., p_{semi} associated with the minimum A_c) increases to obtain higher operating frequencies. Also, as the semiglobal tier pitch increases, chip area and, therefore, interconnect length also increases. However, it can be observed from Fig. 18 that the increase in chip area still remains well below the area required for the 2-D case. Fig. 18 also helps define a maximum-performance 3-D chip—a chip with the same (footprint) area (8.17 cm^2) as the corresponding 2-D chip, which can be obtained by increasing the semiglobal pitch beyond that for the 4-GHz case.

Two scenarios are considered: 1) global pitch is increased to match the global pitch for the 2-D case and 2) global pitch is increased to match the chip area (footprint) for the 2-D case. Table 3 shows that performance can be increased by 63% for case 2). Note that the delay requirement sets a maximum value of interconnect length on any given tier. Therefore, as interconnect lengths are increased, lines which exceed this maximum length criterion for that particular tier need to be rerouted on upper tiers.

Beyond the maximum performance point for the 3-D chip in Fig. 18 (normalized semiglobal pitch $\cong 1.75$), the performance gain becomes increasingly smaller in comparison to the decrease in performance resulting from the increase in chip area or interconnect delay. This eventually saturates the reduction in the overall interconnect delay, and therefore, as shown in Fig. 19, the clock frequency saturates. Furthermore, as the semiglobal pitch is increased beyond the maximum performance point, semiglobal wires need to be rerouted on the global tiers, which eventually leads to overcrowding of the global tier. Any further increases in the wiring density in the global tier forces a reduction in the global pitch as shown in Fig. 20.

The analysis presented so far was for a 50-nm two-Si-layer 3-D technology where the number of metal layers was preserved (in comparison to the 2-D case). In the next two sections, we extend this analysis to study the effect of more than two Si layers and also the effect of increasing the number of available metal layers.

D. Effect of Increasing Number of Silicon Layers

Three-dimensional technologies providing more than two active layers have also been considered. As the number of silicon layers increases beyond two, the assumption that all interlayer interconnects (ILICs) are vertical and consume negligible area becomes less tenable. For this particular example, it is assumed that 90% of all ILICs are horizontal (see Table 3). The area used up by these horizontal ILICs can be estimated from their total length and pitch. As shown in Fig. 21, the decrease in interconnect delay becomes progressively smaller as the number of active layers increases. This is due to the fact that area required by ILICs begins to

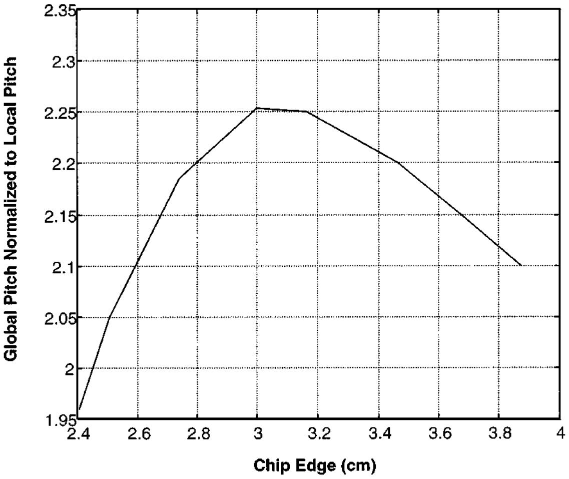


Fig. 20. As the chip size increases due to increasing wire pitch, interconnects are rerouted to higher tiers. The global tier becomes overcrowded for large chip areas and global pitch starts to decrease.

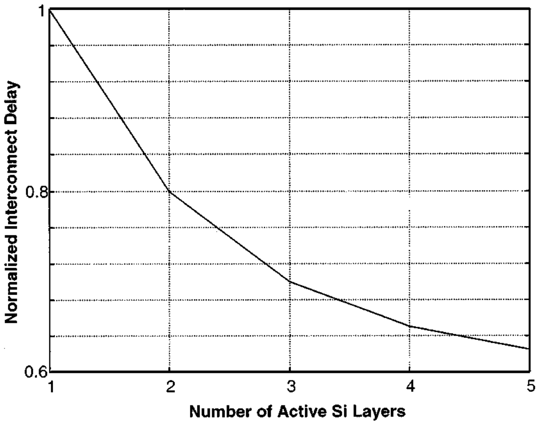


Fig. 21. Interconnect delay normalized to single layer delay as a function of the number of active Si layers shown for 50-nm node. The VILICs are assumed to consume lateral area.

offset any area saving due to increasing the number of active layers.

E. Effect of Increasing the Number of Metal Layers

In the above analysis, the total number of metal layers for 2-D and 3-D case was conserved. However, it is likely that there are local and semiglobal tiers associated with every active layer, and a common global tier is used. This would result in an increase in the total number of metal layers for the 3-D case. The effect of using 3-D ICs with constant metal layers discussed earlier and the effect of employing twice the number of metal layers as in 2-D are summarized in Fig. 22 for various technology nodes as per [1]. It can be observed that by using twice the number of metal layers the performance of the 3-D chip can be improved by an additional 35% (for the 50-nm node) as compared to the 3-D chip with same total number of metal layers as in 2-D. Fig. 22 also shows the impact of moving only the repeaters to the second Si layer. It can be observed that a performance gain of $\sim$9% is achieved for the 50-nm node. The gate delay and the interconnect delay (with repeaters) for the 2-D chip are identical to that shown in Fig. 1, and have been included in this figure for comparison. Finally, it can also be observed that for more aggressive technologies, the decrease in interconnect delay from 2-D to 3-D case is less impressive. This indicates that more than two active layers are possibly needed for those advanced nodes.

F. Optimization of Interconnect Distribution

In estimating chip area, the metal requirement is calculated from the obtained wire-length distribution. The total metallization requirement is appropriately divided among the available metal layers in the corresponding technology. Thus, in the example shown in Fig. 17, each tier, the local, the semiglobal and the global has three metal layers. The resulting area of the most densely packed tier, the local tier in this example, determines the chip area.

Consequently, higher tiers are routed within a larger than required area. An optimization for this scenario is possible by rerouting some of the local wires on the semiglobal tier and the latter on the global, without violating the maximum allowable length (or delay) per tier. This is achieved by reducing the maximum allowed interconnect length for the local and semiglobal tiers (L_{local} and $L_{\mathrm{Semi\text{-}global}}$ in Fig. 15) with varying fractions, w_1 and w_2, respectively. This is implicitly achieved by suitably reducing the parameter β in (43) and (44). Minimum chip area will be achieved when all the tiers are almost equally congested. The resulting calculations for chip area with optimized interconnect distribution for the 2-D IC analyzed in Fig. 17 are shown in Fig. 23. The 2-D chip area is seen to reduce by 9% as a result of this optimization. This wiring network optimization is also applied to 3-D ICs. The results are shown in Fig. 24 where the 3-D chip area is reduced by 11%.

IV. Challenges for 3-D Integration

A. Thermal Issues in 3-D ICs

An extremely important issue in 3-D ICs is heat dissipation [49], [50]. Thermal effects are already known to significantly impact interconnect/device reliability and performance in high-performance 2-D ICs [38], [51]. The problem is expected to be exacerbated by the reduction in chip size, assuming that same power generated in a 2-D chip will now be generated in a smaller 3-D chip, resulting in a sharp increase in the power density. Analysis of thermal problems in 3-D circuits is therefore necessary to comprehend the limitations of this technology and also to evaluate the thermal robustness of different 3-D technology and design options.

It is well known that most of the heat energy generated in integrated circuits arises due to transistor switching. This heat is typically conducted through the silicon substrate to the package and then to the ambient by a heat sink. With multilayer device designs, devices in the upper layers will also generate a significant fraction of the heat. Furthermore, all the active layers will be insulated from each other by layers of dielectrics (LTO, HSQ, polyimide, etc.) which typically have much lower thermal conductivity than Si [52], [53]. Hence, the heat dissipation issue can become even more acute for

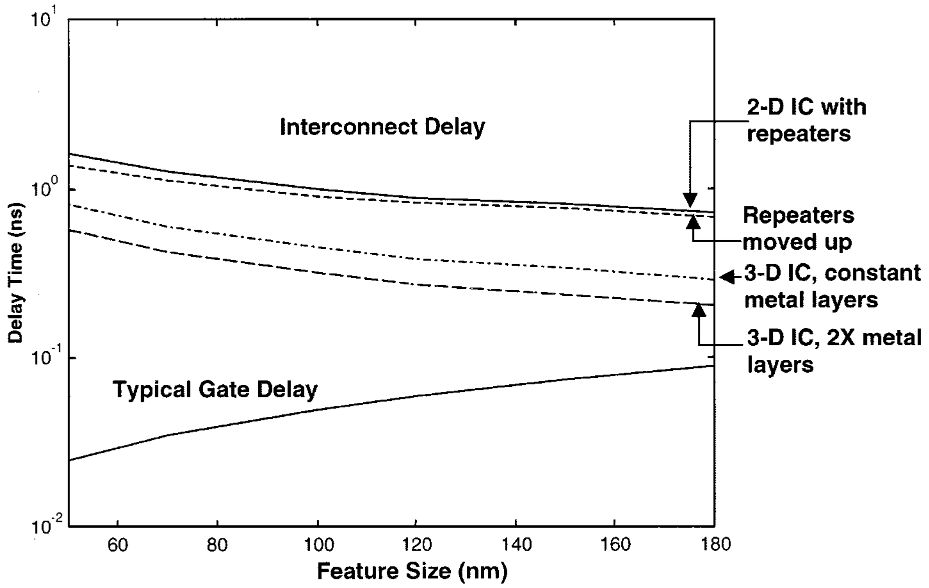


Fig. 22. Comparison of interconnect delay as a function of technology nodes (feature sizes) for 2-D and two-layer 3-D ICs. Moving repeaters to the upper active layer reduces interconnect delay by 9%. For the 50-nm node, 3-D IC (two active layers with same number of interconnects as the 2-D chip) shows significant delay reduction (63%). Increasing the number of metal levels in 3-D reduces interconnect delay by a further 35%. This figure is based on the assumption that 3-D chip (footprint) area equals 2-D chip area.

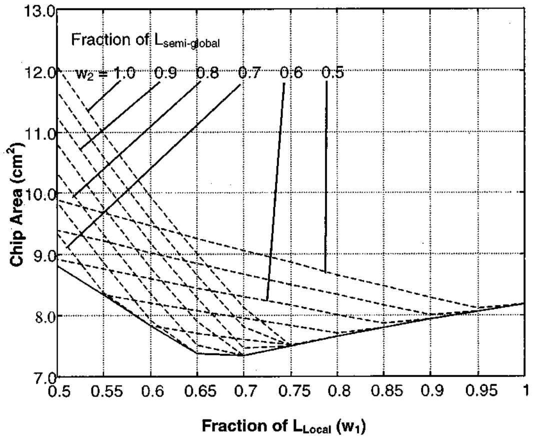


Fig. 23. Chip area for a 2-D IC with wiring network optimization. Solid line represents points of minimum area. (Based on ITRS data for 50-nm node.)

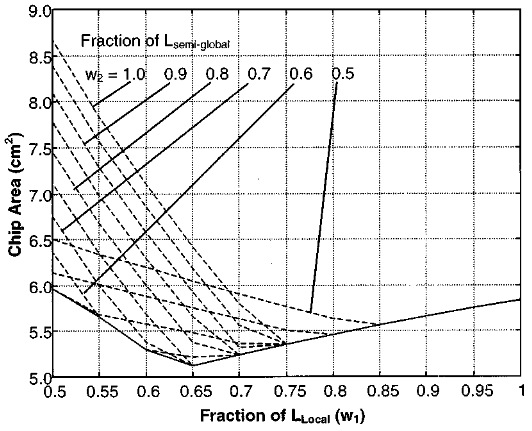


Fig. 24. Chip area for 3-D ICs with wiring network optimization. Solid line represents points of minimum area. (Applied to ITRS, 50-nm node.)

3-D ICs and can cause degradation in device performance, and reduction in chip reliability due to increased junction leakage, electromigration failures, and by accelerating other failure mechanisms [38].

In this section, a general methodology for estimating the temperatures of different active layers of a 3-D chip is presented and then applied to the specific example of a 3-D chip with two silicon layers. The analysis begins with die temperature estimation for 2-D circuits. In order to illustrate the thermal issues, a packaging technology-based package thermal resistance extracted at the present (180 nm) technology node for 2-D circuits has been used for both 2-D and 3-D chips.

1) Package Thermal Resistance Model for 2-D and 3-D ICs: Fig. 25 shows the total power dissipation (P) and chip area (A) for high-end microprocessors for various 2-D technology nodes based on [1]. It can be observed that, as technology scaling continues, chip area and power dissipation increases. The relationship between the die temperature rise (ΔT_{Die}) and P can be expressed as

$$\Delta T_{\text{Die}} = (T_{\text{Die}} - T_{\text{amb}}) = P \cdot R_\theta \tag{46}$$

where T_{amb} is the chip ambient temperature (≈25 °C), and R_θ is the effective thermal resistance from the Si devices to the heat sink and is mostly due to the package material between the Si and the heat sink. Neglecting interface resistances, R_θ can be expressed as

$$R_\theta = \left(\frac{t_{\text{Si}}}{K_{\text{Si}}} + \frac{t_{\text{Pkg}}}{K_{\text{Pkg}}}\right)\frac{1}{A} = \frac{R_n}{A}. \tag{47}$$

Here, t_{Si} and K_{Si} are the thickness and the thermal conductivity of the Si substrate, and t_{Pkg} and K_{Pkg} denote same parameters for the packaging material as shown in Fig. 26. A is the chip area through which heat flow takes place. R_n is the normalized package thermal resistance. Since the die size (length) is much larger than the thickness of Si, we assume one-dimensional heat flow. Hence, from (46) and (47), it follows that

$$\Delta T_{\mathrm{Die}} = R_n \frac{P}{A}. \tag{48}$$

Since the typical die temperature for present high-performance 2-D circuits (180-nm technology node) is known to be ~120 °C, the value of R_n can be calculated to be 4.75 °C/(W·cm^{-2}). Using this value of R_n, the die temperatures for other 2-D technology nodes based on [1] can be estimated from Fig. 25.

2) Analytical Die Temperature Model for 3-D ICs: A simple analytical model is proposed to estimate the temperature rise in each active layer of 3-D chips. The temperature rise (above the ambient temperature) of the jth active layer in an n-layer 3-D chip, schematically shown in Fig. 27(a), can be expressed as

$$\Delta T_j = \sum_{i=1}^{j} \left[R_i \left(\sum_{k=i}^{n} \frac{P_k}{A} \right) \right] \tag{49}$$

where

n total number of active layers;

R_i thermal resistance between the ith and the $(i-1)$th layers;

P_k power dissipation in the kth layer.

Note that this model does not take into account interconnect Joule heating. Assuming identical power dissipation (P) in each layer and identical thermal resistances (R) between layers, the temperature rise of the uppermost (nth) layer in an n-layer 3-D chip can be expressed as [50]

$$\Delta T_n = \left(\frac{P}{A}\right) \left[\frac{R}{2} n^2 + \left(R_1 - \frac{R}{2} \right) n \right] \tag{50}$$

where R_1 is mostly due to the package thermal resistance between the first layer and the heat sink (separated by the package layer of thickness t_{pkg}) and R is the thermal resistance between the ith and the $(i-1)$th layers for $i \neq 1$

$$R_1 = \frac{t_{\mathrm{Si\_1}}}{K_{\mathrm{Si}}} + \frac{t_{\mathrm{pkg}}}{K_{\mathrm{pkg}}}$$

and

$$R = \frac{t_{\mathrm{Si\_}i}}{K_{\mathrm{Si}}} + \frac{t_{\mathrm{glue},\,i-1}}{K_{\mathrm{glue},\,i-1}} + \frac{t_{\mathrm{ins},\,i-1}}{K_{\mathrm{ins},\,i-1}} \tag{51}$$

respectively.

Here, $t_{\mathrm{Si\_}i}$ is the thickness of the ith Si layer, and $t_{\mathrm{glue},\,i-1}$ and $t_{\mathrm{ins},\,i-1}$ are the thickness of the $(i-1)$th glue and insulator [Cu+ILD layer in Fig. 27(a)] layers, respectively. From (50), the temperature rise can be expected to increase linearly with power density and the square of the number of active layers n. However, for all practical 3-D ICs, $R_1 \gg R$, which gives rise to an approximately linear relationship between ΔT_n and n as shown in Fig. 27(b). Equation (50) also suggests that for most 3-D ICs with $n \leq 5$, R_1 will dominate the temperature rise of any layer.

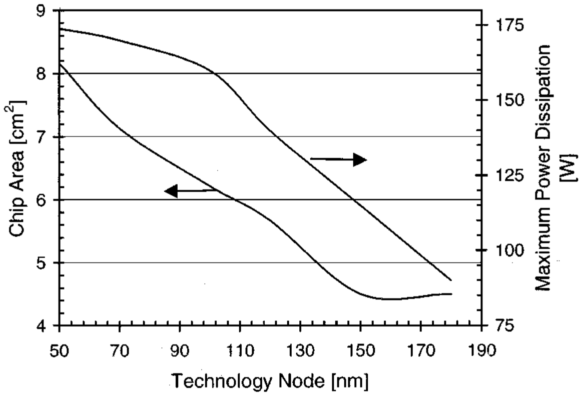


Fig. 25. Maximum power dissipation and chip area in 2-D circuits as a function of technology node based on ITRS.

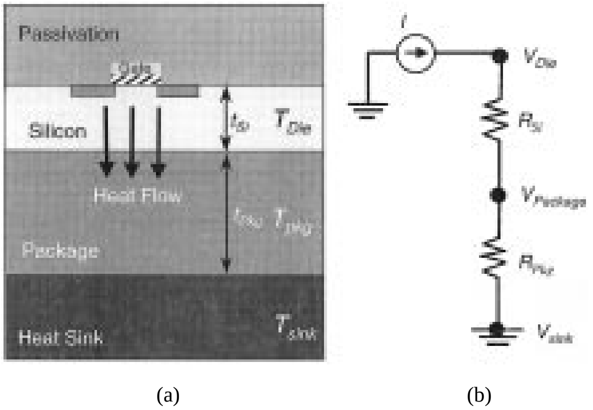


Fig. 26. Schematic view of (a) heat flow in 2-D circuits and (b) equivalent thermal circuit. T denotes temperature of different materials. R_{Si} and R_{Pkg} are the thermal resistances of the Si and the package material, respectively.

For the two-active-layer ($n = 2$) 3-D example used in our performance analysis earlier, the temperature of each of the layers ($j = 1$ and $j = 2$) can be expressed using (49) as

$$\Delta T_1 = \left[\frac{(P_1 + P_2)}{A} R_1 \right]$$

and

$$\Delta T_2 = \left[\frac{(P_1 + P_2)}{A} R_1 \right] + \left[\frac{P_2}{A} R \right] \tag{52}$$

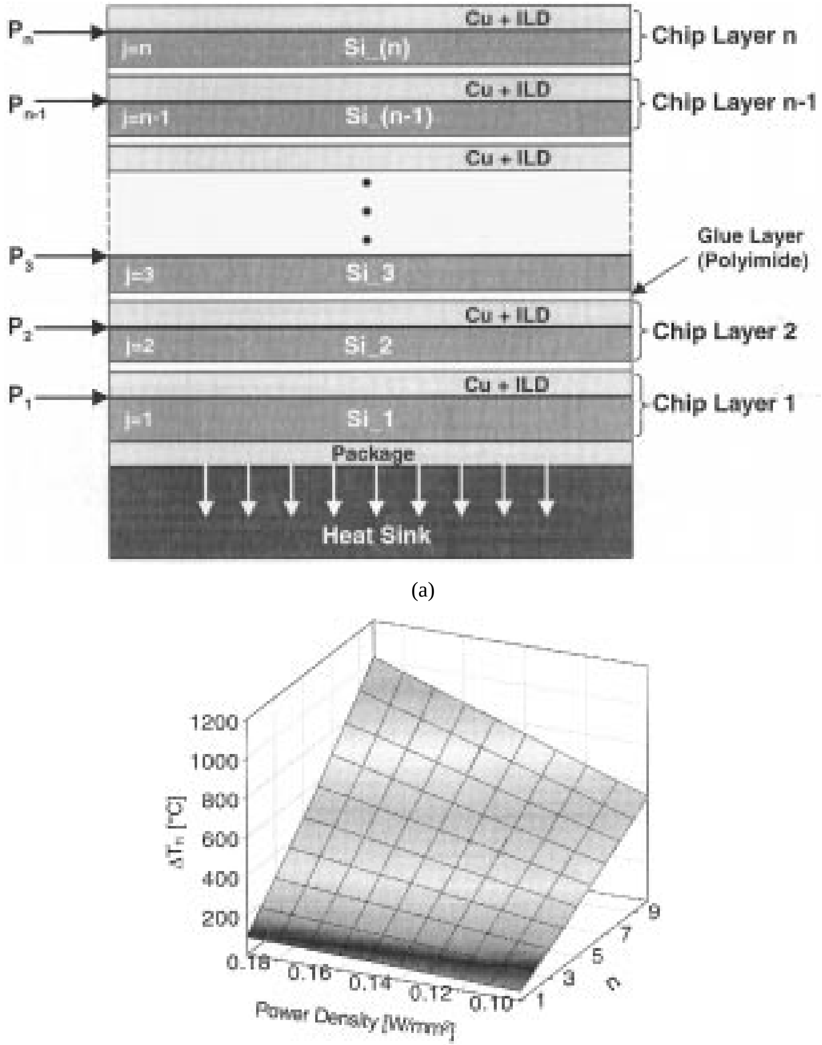
where

$$R_1 = \left(\frac{t_{\mathrm{Si\_1}}}{K_{\mathrm{Si\_1}}} + \frac{t_{\mathrm{pkg}}}{K_{\mathrm{pkg}}} \right)$$

which can be extracted from (48) assuming same packaging material for 2-D and 3-D chips. The temperature rise for the second active layer, ΔT_2, can therefore be expressed as

$$\Delta T_2 = \Delta T_1 + \frac{P_2}{A} R \tag{53}$$

where the second term on the right-hand side represents the effective temperature difference between active layer 2 and



(a)

(b)

Fig. 27. (a) Schematic of an n-layer 3-D chip with a heat sink at the bottom. P denotes the power dissipation in each layer. (b) Temperature increase as a function of n and the power density in each layer.

active layer 1. Since the 3-D chip area can be calculated from our model, and R_1 remains constant for both 2-D and 3-D circuits assuming same packaging material for all the technology nodes, the temperature of each of the active layers can be calculated.

3) Comparison Between 2-D and 3-D ICs: It has been recently shown that the power dissipation in 3-D circuits has a strong design dependence [42]. Three-dimensional design options where the chip area is the same as the corresponding 2-D chip gives the highest system performance (frequency) as discussed earlier (see Fig. 18). However, it also results in higher power dissipation giving rise to higher die temperatures. This is expected since same chip area between 2-D and 3-D is achieved by increasing the metal cross-sectional area for 3-D, which reduces the line resistance (R) and, hence, increases the operating frequency. However, since $P \propto Cf \propto 1/R$, the power dissipation increases, resulting in higher die temperatures.

We now present a comparison between the 2-D and 3-D ICs with respect to their performance, chip area, and power dissipation. Table 4 lists various parameters for a 2-D IC at the 50-nm technology node based on [1] and the performance analysis methodology presented in this paper. Corresponding parameters are also calculated for two limiting designs of a two-layer 3-D IC. In one case, the 3-D IC is designed to have the same chip area as that for the 2-D case and in the second design both the 2-D and 3-D ICs have the same operating frequency. As mentioned earlier, the 3-D design with the same chip (footprint) area (8.17 cm^2) as that for the 2-D case gives the maximum performance ($f_c = 6$ GHz). This design also gives the highest total interconnect network capacitance and the highest power density per layer, while the other 3-D design with the same operating frequency (3 GHz) as that for the 2-D IC gives the lowest total interconnect capacitance and the lowest power density per layer.

The total interconnect network capacitances shown for the 2-D and 3-D cases in Table 4 were calculated by summing the interconnect capacitances for each tier, local, semiglobal, and global, i.e.,

$$C_{\text{total}} = C_{\text{local}} + C_{\text{semi}} + C_{\text{global}}. \quad (54)$$

By using the maximum allowable delay per tier criteria, described in Section III-B, we calculate the longest wire on each tier. Also, as described in detail in Section III-B, the area under an interconnect density function plot (wire-length distribution) can be used to calculate the total length of wire on each tier. The capacitance for each tier is then calculated using (9), where $A.R.$ is the aspect ratio for that tier. The calculated capacitances for all the tiers are then summed to find the total interconnect capacitance.

Now, for the 2-D and 3-D chips, we can express the power dissipation as follows:

$$P = \tfrac{1}{2}\alpha C V_{dd}^{2} f_c. \qquad (55)$$

Here, we have only considered the dynamic power dissipation. For the 2-D case, P and f_c are given in [1] and C was estimated using (54). The product (αV_{dd}^2) was calculated using (55). For the corresponding 3-D ICs, the interconnect dominated capacitances, C_{total}, and the chip frequencies, f_c, are calculated using our model, and, in order to be consistent, the same value of the product (αV_{dd}^2) estimated for the 2-D IC was used for calculating the power dissipation.

For the 3-D IC design with the same chip area as that for the 2- D IC, it is obvious that the power density (power per unit area) is going to be higher since the operating frequency is twice as large. The die temperature for such a 3-D chip can be estimated using (52) (assuming same value of the package thermal resistance as that for the 2-D ICs) to be 211 °C and 294 °C for the first and the second active layers, respectively. Fig. 28 shows a plot of the required package thermal resistances to maintain the temperature of any layer at 120 °C as a function of the total chip power density for 2-D and 3-D ICs. For both the 2-D and the 3-D circuits, heat sink was assumed to be attached to the lower Si substrate only. It can be observed that maintaining the temperature of the upper silicon layer (3-D Si_2) in the 3-D chip at 120 °C requires lower package thermal resistance than that required for the first layer (3–D Si_1) of the 3-D chip or the 2-D chip. This is due to the extra thermal impedance between the two layers. Note that, in Fig. 28, lower values of package thermal resistances represent advanced packaging and cooling technologies. Also, the thermal problem can be significantly alleviated if heat sinks can be provided for both the active layers.

Note that in all calculations Joule heating of the interconnects has been ignored since most of the heat is dissipated by the transistors. However, interconnect Joule heating can increase the peak temperature in 3-D chips due to strong thermal coupling with the neighboring interconnects and the active layers giving rise to higher interconnect temperatures and, hence, higher interconnect resistance and also lower interconnect electromigration performance. In order to take these coupling effects into account, full chip thermal analysis using finite element simulations are needed as shown in [50].

From Fig. 28, it can be concluded that in order to operate the 3-D chips at their maximum performance limits, advancement in cooling and packaging technologies will be necessary to maintain acceptable chip temperatures. Lower operating temperatures for 3-D ICs can be achieved by employing a cooling design similar to the one illustrated in Fig. 29 [54] where coolant (water) pumped through microchannels etched at the back surface of a silicon substrate were used to achieve package thermal resistance of 0.09 °C/(W·cm^{-2}). Recent extensions of this approach are targeting even lower thermal resistances using closed-loop two-phase cooling systems with boiling convection in microchannels [55]. The geometry of the chip and the packaging layers for this cooling system are shown in Fig. 29.

Table 4
Comparison Between 2-D and 3-D ICs at the 50-nm Technology Node. Parameters for Two Limiting Cases of 3-D ICs have been Shown, One with the Same Chip Area as the 2-D IC and the Other with the Same Operating Frequency as the 2-D IC

50 nm Technology Node

	2-D	3-D	3-D
Active Layers	1	2	2
f_c (MHz)	3000	3000	6000
Feature Size (nm)	50	50	50
A_c (cm^2)	8.17	5.80	8.17
N_t (Millions) per Active Layer	7053	3526.5	3526.5
Gate Pitch (cm)	3.4E-5	4.06E-5	4.81E-5
p_{local} (μm) / A.R.	0.1 / 2.1	0.1 / 2.1	0.1 / 2.1
p_{semi} (μm) / A.R.	0.165 / 2.7	0.14 / 2.7	0.33 / 2.7
p_{global} (μm) / A.R.	0.275 / 2.9	0.23 / 2.9	0.55 / 2.9
L_{local} (gate pitches)	6190	5195	1313
L_{semi} (gate pitches)	10324	6826	4380
L_{global} (gate pitches)	83982	59384	59384
C_{total} (per active layer) (μF)	6.1285	2.370	5.6257
Total Power Dissipation (W)	174	135	639
Power Density per Layer (W/mm^2)	0.213	0.116	0.391

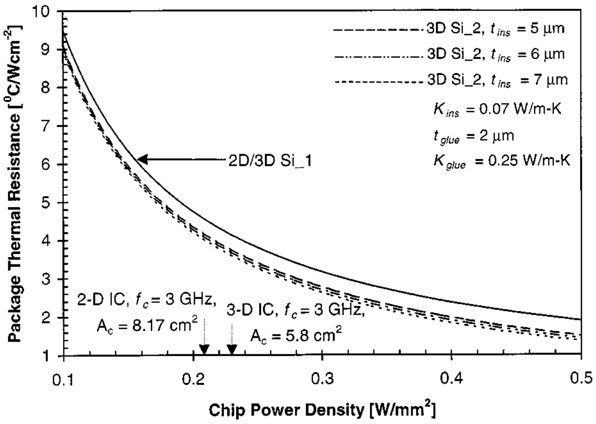


Fig. 28. Required package thermal resistance for 2-D and two-layer 3-D ICs to maintain the temperature of any layer at 120 °C as a function of chip power density. Heat sink is assumed at one end of the chip only. For the 3-D IC, as the dielectric thickness between the two active layers (t_{ins}) increases, lower values of the package thermal resistances are needed to maintain the temperature of the second active layer at 120 °C. The power densities corresponding to one of the 3-D designs discussed in the text, and the 2-D chip at 50-nm node, are also shown.

It is interesting to note that dummy thermal vias have been recently shown to be useful in reducing the temperature of interconnects in 2-D ICs [56]. A similar strategy can be used for the 3-D ICs, where interchip thermal vias that conduct heat but are electrically isolated can be employed to alleviate the heat dissipation problem in high-performance 3-D ICs. Furthermore, it is important to realize that thermal problems in 3-D ICs will be less severe for applications that do

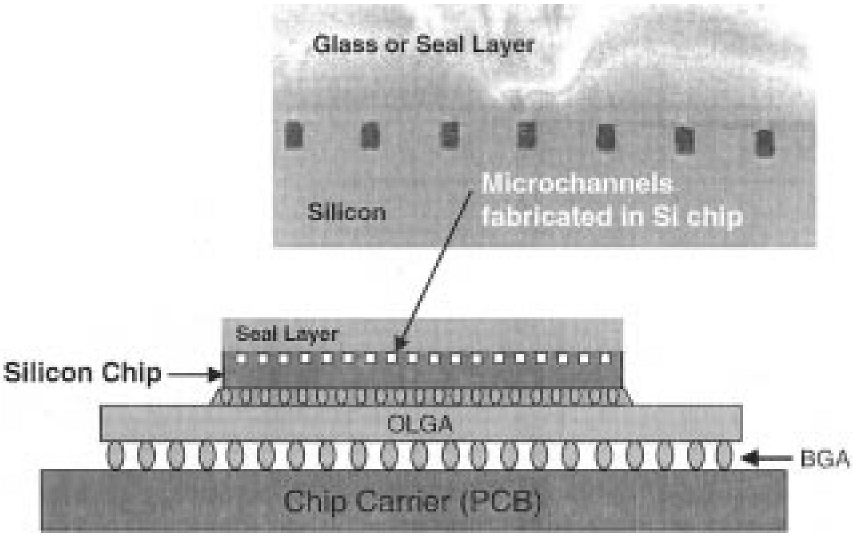


Fig. 29. Schematic of a packaged Si chip with integrated microchannels etched in the substrate for pumping coolant to lower the package thermal resistance. BGA and OLGA denote ball grid array and organic layer ball grid array, respectively. (Courtesy of Kenneth E. Goodson, Stanford University.)

not require integration of high-performance logic. For example, integration of memory, analog, or RF blocks or any other circuits that have much lower power dissipation compared to high-performance logic may not require costly packaging and cooling solutions. However, any 3-D integration involving high-performance logic (even in the layer closest to the heat sink) would require careful thermal budgeting for the upper layer circuits, which would certainly be affected by the power dissipation of the logic layer according to (53). Additionally, nonuniform temperature distribution among the interconnects and devices in different active layers can lead to performance mismatch and degradation as recently demonstrated for 2-D high-performance ICs with nonuniform substrate temperature distribution [57], [58].

B. *Electromagnetic Interactions (EMI) in 3-D ICs*

1) Interconnect Coupling Capacitance and Cross Talk: In 3-D ICs, an additional coupling between the top layer metal of the first active layer and the devices on the second active layer is expected to be present. This needs to be addressed at the circuit design stage. However, for deep-submicrometer technologies, the aspect ratio of global tier interconnects is ≥ 2.5 [1]. Therefore, line-to-line capacitance is the dominant portion of the overall capacitance. Hence, the presence of an additional silicon layer on top of a global metal line may not have an appreciable effect on the line capacitance per unit length. For technologies with very small aspect ratio, the change in interconnect capacitance due to the presence of an additional silicon layer could be significant, as reported in [59].

2) Interconnect Inductance Effects: For deep-submicrometer interconnects on-chip inductive effects arising due to increasing clock speeds, decreasing rise times, and increasing length of on-chip interconnects is a concern for signal integrity and overall interconnect performance [60]. Inductance can increase the interconnect delay per unit length and can cause ringing in the signal waveforms, which can adversely affect signal integrity [61], [62]. For long global wires (such as clock lines), inductance effects are more severe due to the lower resistance of these lines, which makes the reactive component of the wire impedance comparable to the resistive component, and also due to the presence of significant mutual inductive coupling between wires, resulting from longer current return paths [63]. For Cu-based technologies, line resistances have decreased further and, as a result, inductive effects are expected to become more significant. In 3-D ICs, the reduction of wire lengths will certainly help reduce inductance. Additionally, the presence of a second substrate close to the global wires might help lowering the inductance by providing shorter return paths, provided the substrate resistance is sufficiently low or if the wafers are bonded through metal pads as discussed in Section VI-B.

C. *Reliability Issues in 3-D ICs*

Three-dimensional ICs will possibly introduce some new reliability problems. These reliability issues may arise due to the electrothermal and thermomechanical effects between various active layers and at the interfaces (glue layers) between the active layers, which can also influence existing IC reliability hazards such as electromigration, and chip performance [50]. Additionally, heterogeneous integration of technologies using 3-D architecture will increase the need to understand mechanical and thermal behavior of new material interfaces, thin-film-material thermal and mechanical properties, and barrier/glue layer integrity. Additionally, from a manufacturing point of view, there might be yield issues arising due to the mismatch between the individual die-yield maps of different active layers, which may affect the net yield of 3-D chips. Such issues would demand a careful tradeoff between system performance, cost, and the 3-D manufacturing technology.

V. Implications for Circuit Design and System-on-a-Chip Applications

A. Repeater Insertion

For deep-submicrometer technologies, interconnect delay is the dominant component of the overall delay, especially for circuits with very long interconnects where the delay can become quadratic with line lengths. To overcome this problem, long interconnects are typically broken into shorter buffered segments. In [11], it was shown that, for point-to-point interconnects, there exists an optimum interconnect length and an optimum repeater size for which the overall delay is minimum. Repeater sizes for various metal layers for different technologies have been presented in [11] and [26]. For top-layer interconnect, the corresponding inverter sizes were approximately 450 times the minimum inverter size available in the relevant technology. These large repeaters present a problem since they take up a lot of active silicon and routing area. The vias that connect such a repeater from the top global interconnect layers block all the metal layers present underneath them, hence taking up substantial routing area. It has been predicted [64] that the number of such repeaters can reach 10 000 for high-performance designs in 100-nm technology. A methodology to estimate the chip area utilized by the repeaters is presented in the next section.

1) Chip Area Utilization by Repeater Insertion: The following is a description of the methodology used to estimate the fraction of chip area utilized by repeater insertion. Repeaters are assumed to be inserted along wires whose lengths exceed a certain critical length. This critical length is determined by the maximum allowable signal delay along the wire for each interconnect tier (as described in Section III-B). To illustrate, the local tier cannot have any nonrepeated lines that exceed a maximum allowable length, L_{opt} in (3). Any wires that are routed in the local tier whose length are required to be greater than L_{opt} must have repeaters inserted along their lengths in order to satisfy the maximum allowable signal delay for this tier. The maximum length of repeated interconnect wire in any given tier is not arbitrary. Repeated wires are assumed to have repeaters inserted optimally and the signal delay along such wires is given by (6). The maximum allowable length per interconnect tier is calculated based on (42)–(44). As an example, a schematic figure describing the critical lengths for the local tier is given in Fig. 30.

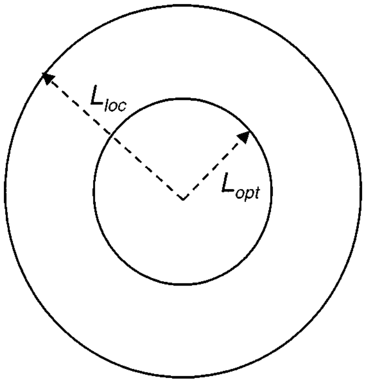


Fig. 30. Interconnect length boundaries for the local tier. L_{opt} is the maximum allowed length of an interconnect without repeater. L_{loc} describes the maximum length of any wire in the local tier. Interconnects with lengths $L_{\text{opt}} \leq l \leq L_{\text{loc}}$ require repeaters.

To estimate the fraction of chip area utilized by repeater insertion on all tiers, it is necessary to find the total number of repeaters, which is then multiplied by the size of a repeater. The size of a repeater is dependent on the wire that it is driving. For each tier, therefore, an optimum driver size can be calculated by multiplying the minimum repeater size, B_o, with a factor, $s_{\text{opt}} = \sqrt{r_o c/3rc_{\text{NMOS}}}$ (as described in Section I-A). To determine the total number of repeaters, it is necessary to determine the number of interconnects that require repeater insertion. For this we make use of Rent's Rule. As represented in Fig. 30, any given tier is divided into two regions. The central region of area πL_{opt}^2 is characterized by interconnects that are not repeated. Applying the recursive property of Rent's Rule, this central region can be considered as a logic block consisting of N_{central} logic gates. The number of I/Os connecting this central region to its surroundings is given by kN_{central}^p where k is Rent's Constant and p is Rent's Exponent. The probability, P_1, that the I/O of any gate within this area of πL_{opt}^2 reaches outside this area is given by

$$P_1 = \frac{kN_{\text{central}}^p}{kN_{\text{central}}} = N_{\text{central}}^{p-1}. \tag{56}$$

Assuming that the number of logic gates is related to the logic block area (A) by some constant of proportionality, i.e., $A = \pi L_{\text{opt}}^2 \propto N_{\text{central}}$, then P_1 for the local tier can be written as

$$P_1 = \kappa^{(p-1)} L_{\text{opt}}^{2(p-1)} \tag{57}$$

where κ is a constant of proportionality. Similarly, the probability, P_2, that the I/O of any gate within the local tier of area of πL_{loc}^2 reaches outside this area is given by

$$P_2 = \kappa^{(p-1)} L_{\text{loc}}^{2(p-1)}. \tag{58}$$

Hence, the probability that the I/O of any gate within the entire local tier to remain *inside* the tier is given by $(1 - P_2)$.

Therefore, the total probability, P_{loc}, that an interconnect will satisfy the length condition $L_{\text{opt}} \leq l \leq L_{\text{loc}}$ is given by

$$P_{\text{loc}} = P_1(1 - P_2). \tag{59}$$

Hence, the number of interconnects, I_R, that require repeater insertion for the local tier is simply the probability P_{loc} multiplied by the total number of I/Os of all the gates:

$$I_R = P_1(1 - P_2)k\kappa L_{\text{loc}}^2. \tag{60}$$

The optimum number of repeaters per unit length of wire $(1/l_{\text{opt}})$ is given by $\sqrt{0.4rc/4.2r_0c_{\text{NMOS}}}$ (see Section I-A). To estimate the total number of repeaters an average length of wire, l_{avg}, is considered, where

$$l_{\text{avg}} = \left(\frac{L_{\text{opt}} + L_{\text{loc}}}{2}\right). \tag{61}$$

Hence, the total number of repeaters can be expressed as

$$P_1(1 - P_2)k\kappa l_{\text{avg}} \sqrt{\frac{0.4rc}{4.2r_0c_{\text{NMOS}}}}\, L_{\text{loc}}^2. \tag{62}$$

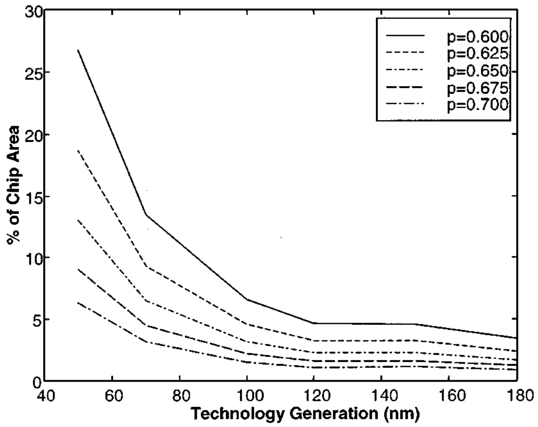


Fig. 31. Fraction of chip area used by repeaters for different technology nodes based on ITRS and different Rent's Exponents. As much as 27% of the chip area at 50-nm node is likely to be occupied by repeaters. The percentage of chip area occupied by repeaters decreases with increasing values of p. As p increases, more wires are allocated for chip I/O than for interblock connections, reducing the number of required repeaters for wires engaged in on-chip communications.

The total area used up by the repeaters in the local tier, $A_{R,\,\text{loc}}$, can therefore be expressed as

$$A_{R,\,\text{loc}} = P_1(1-P_2)k\kappa l_{\text{avg}}\sqrt{\frac{0.4rc}{4.2r_0c_{\text{NMOS}}}}\,L_{\text{loc}}^2 B_0 s_{\text{opt, loc}} \tag{63}$$

where B_o is the minimum repeater size ($\approx 60WL$) and $s_{\text{opt, loc}}$ is the optimum multiple of minimum repeater size for the local tier. All parameters in (63) can be calculated for a given technology node based on [1]. This procedure is repeated to account for all the interconnect tiers to estimate the total area, $A_{R,\,\text{total}}$, utilized by repeaters, i.e.,

$$A_{R,\,\text{total}} = A_{R,\,\text{loc}} + A_{R,\,\text{semi}} + A_{R,\,\text{glob}}. \tag{64}$$

Using the methodology presented above, the percentage of total chip area utilized by the repeaters were calculated at each technology node based on [1]. It can be observed from Fig. 31 that inserting these repeaters will cause significant area penalty, especially beyond the 70-nm node. However, this problem can be easily tackled using 3-D technology with just two silicon layers. The repeaters can be placed on the second silicon layer, thereby saving area on the first silicon layer and reducing the footprint area of the chip. Furthermore, if the second silicon layer is placed close to the common global metal layers, the vias connecting the global metal layers to the repeaters will not block the lower metal layers, thereby freeing up additional routing area.

Previously, Fig. 22 had also included delay simulation results for an otherwise single-active-layer IC except that the repeaters had now been moved to a second active layer. A conservative value of Rent's Exponent ($p = 0.65$) was used to estimate the reduction in chip area and therefore reduction in overall interconnect delay. At 50-nm node, an additional reduction of 9% in the overall interconnect delay results from the resulting area reduction.

B. Layout of Critical Paths

In typical high-performance ASIC and microprocessor designs, interconnect delay is a significant portion of the overall path delay [65]. Logic blocks on a critical path need to communicate to other logic blocks which, due to placement and other design constraints, may be placed far away from each other. The delay in the long interconnects between such blocks usually causes timing violations. With the availability of a second active layer, these logic blocks can be placed on different silicon layers and, hence, can be very close to each other, thereby minimizing interconnect delay. Even if highest quality devices are not made on the second active layer, the decrease in interconnect delay can be more than the increase in gate delay due to suboptimal transistor characteristics.

C. Microprocessor Design

In microprocessors and DSP processors, most of the critical paths involve on-chip caches [66]. The primary reason for this is that on-chip cache is (physically) located in one corner of the die whereas the logic and computational blocks, which access this memory, are distributed all over the die. By using a technology with two silicon layers, the caches can be placed on the second active layer and the logic and computational blocks on the first layer. This arrangement ensures that logic blocks are in closer proximity to on-chip caches.

Consider a microprocessor of dimensions $L \times L$. In typical current generation microprocessors, about half the physical area is taken up by on-chip caches. Hence, the worst case interconnect length in a critical path is $2L$ (typically the data transfer from cache takes more than one clock cycles but we assume single clock cycle transfers for simplicity). If on-chip caches are placed on the second active layer and the chip is resized accordingly to have dimensions $(L/\sqrt{2}) \times (L/\sqrt{2})$, then the worst case interconnect length is $\sqrt{2}L$ a reduction of about 30%. Even though this analysis is very simplistic compared to the more elaborate one presented in Section III and does not perform any optimization of the interconnect pitch, it demonstrates that going from a single silicon layer to two layers results in nontrivial improvement in performance. Recent studies [67] have shown that, by integrating level one and level two cache and the main memory on the same silicon using 3-D technology, access times for level 2 cache and main memory can be decreased. This, coupled with an increase in bandwidth between the memory, level 2 cache, and level 1 cache, reduces the level 2 cache/memory miss penalty and therefore reduces the average time per instruction and increases system performance.

D. Mixed Signal Integrated Circuits

With greater emphasis on increasing the functionality that can be implemented on a single die in the SoC paradigm, more and more analog, mixed-signal and RF components of the system are being integrated on the same piece of silicon (as illustrated in Fig. 10). However, this presents serious design issues since switching signals from the digital portions of the chip couple into the sensitive analog and RF circuit nodes through the substrate and degrade the fidelity (or equivalently, increase the noise) of the signals present in these blocks [68]. Furthermore, different fabrication technologies are required for the two applications. However, with the availability of multiple silicon layers, RF and mixed-signal portions of the system can be realized on a separate layer (using different technologies), thereby providing substrate isolation from the digital portion. A preliminary analysis shows a 30-dB improvement in isolation by moving the RF portions of the circuit to a separate substrate. Moreover, since the second Si layer may not be continuous, good isolation between different analog and RF components (such as the low-noise amplifier (LNA) and power amplifier) can also be achieved.

E. Optical Interconnects for Clocking and I/O Connections

For high-performance microprocessors with operating frequencies greater than a few gigahertz and large die sizes (on-chip frequency $= 3$ GHz, and die area $= 8.17$ cm^2 at the 50-nm technology node [1]), interconnects responsible for global communications, including the interconnect network used for the clock distribution, can contribute significantly to the key performance metrics (area, power dissipation, and delay) and to the overall cost of the chip. As the complexity (size) of the microprocessor increases, synchronization of various blocks in the chip becomes increasingly difficult [69]. This occurs mainly due to the variation in the placement of different blocks (or clock line lengths) and due to differences in their operating temperature that affects the clock skew and the net signal delay. Additionally, data input and output (I/O) requirements drive up the number of I/O pads and the corresponding size of the I/O circuitry (or chip area). Furthermore, in high-performance designs, around 40%–70% of the total power consumption could be due to the clock distribution network [70], [71], and as the total chip capacitance (dominated by interconnects) and the chip operating frequency increases with scaling, the power dissipation increases.

On-chip optical interconnects can eliminate most of the problems associated with clock distribution and I/O connections in large multigigahertz chips [72], [73]. They are attractive for high-density and high-bandwidth interconnections, and optical signal propagation loss is almost distance-independent. Also, the delays on optical clock and signal paths are not strongly dependent on temperature. Additionally, optical signals are immune to electromagnetic interactions discussed earlier with regards to metal interconnects. Hence, optical interconnects are very attractive for large-scale synchronization of systems within multigigahertz ICs. Furthermore, optical interconnects employing short optical (laser) pulses can reduce the optical power requirement [74]. They can also reduce the electrical power consumption since no photocurrent is generated during transition periods since optical power is incident on the transmitters and receivers only during valid output states [75]. The short duration of ultrafast laser pulses also results in large spectral bandwidth, which enables system concepts such as a single-source implementation of wavelength-division multiplexed optical interconnects [76], [77], a technique that allows multiple channels to be transmitted down a single waveguide.

Optical interconnect devices and networks integrated in a 3-D SoC IC (schematically illustrated in Fig. 11) can be employed to attain system synchronization and to enhance system performance. Integrated 3-D optical devices have been demonstrated directly on top of active silicon CMOS circuits [43], [78]–[80]. Also, polysilicon-based optical waveguides of submicrometer dimensions have been demonstrated for low-loss optical signal propagation and power distribution [81].

F. Implications on VLSI Design and Synthesis

VLSI design and synthesis (both logic and physical) for large digital circuits and high-performance SoC type applications based on 3-D ICs will necessitate some new design methodologies, design and layout tools, and test strategies. At an abstract level, physical design (placement and routing) can be viewed as a graph embedding problem. The circuit graph (synthesized and mapped circuit) is embedded on a target graph which is planar (which corresponds to the physical substrate of the conventional single silicon substrate technology). However, with more than one silicon layer available, the target graph is no longer planar, and therefore placement and routing algorithms need to be suitably modified. Moreover, since placement and routing information also affects synthesis algorithms, which in turn can affect the choice of architectures, this modification needs to be propagated all the way to synthesis and architectural level. Additionally, since 3-D ICs would likely involve *silicon-on-insulator* (SOI) type upper active layers, the design process will need to address issues specific to SOI technology to realize significant performance improvements [82], [83].

VI. Overview of 3-D IC Technology

A. Technology Options

Although the concept of 3-D integration was demonstrated as early as in 1979 [84], and was followed by a number of reports on its fabrication process and device characteristics [85]–[94], it largely remained a research technology since microprocessor performance was device-limited. However, with the growing menace of RC delay in recent times, this technology is being viewed as a potential alternative that can not only maintain chip performance well beyond the 130-nm node, but also inspire a new generation of circuit design concepts. Hence, there has been a renewed spur in research activities in 3-D technology [95]–[100] and their performance modeling [42], [67], [101]–[104].

Presently, there are several possible fabrication technologies that can be used to realize multiple layers of active-area (single crystal Si or recrystallized poly-Si) separated by interlayer dielectrics (ILDs) for 3-D circuit processing. A brief description of these alternatives is given as follows. The choice of a particular technology for fabricating 3-D circuits will depend on the requirements of the circuit system, since the circuit performance is strongly influenced

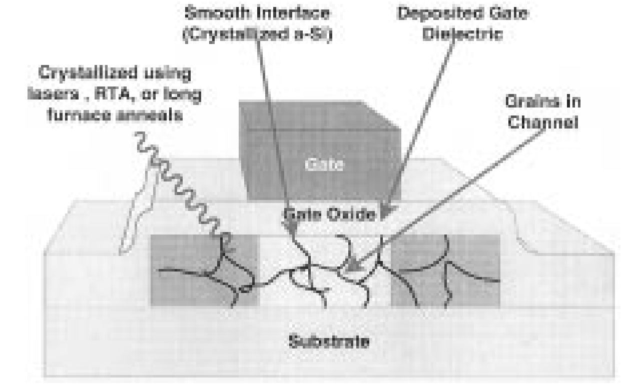


Fig. 32. Schematic of a thin-film transistor (TFT) fabricated on polysilicon depicting several grain boundaries in the active region.

by the electrical characteristics of the fabricated devices as well as on the manufacturability and process compatibility with the relevant 2-D technology.

1) Beam Recrystallization: A very popular method of fabricating a second active (Si) layer on top of an existing substrate (oxidized Si wafer) is to deposit polysilicon and fabricate thin-film transistors (TFT) (see Fig. 32). MOS transistors fabricated on polysilicon exhibit very low surface mobility values [of the order of 10 cm^2/(V·s)], and also have high threshold voltages (several volts) due to the high density of surface states (several 10^{12} cm^{-2}) present at the grain boundaries. To enhance the performance of such transistors, an intense laser or electron beam is used to induce recrystallization of the polysilicon film [84]–[94] to reduce or even eliminate most of the grain boundaries. This technique, however, may not be very practical for 3-D devices because of the high temperature involved during melting of the polysilicon and also due to difficulty in controlling the grain size variations [105], [106]. Beam recrystallized polysilicon films can also suffer from lower carrier mobility (compared to single-crystal Si) and unintentional impurity doping. However, high-performance TFTs fabricated using low-temperature processing [107], and even low-temperature single-crystal Si TFTs, have been demonstrated [108] that can be employed to fabricate advanced 3-D circuits.

2) Silicon Epitaxial Growth: Another technique for forming additional Si layers is to etch a hole in a passivated wafer and epitaxially grow a single-crystal Si seeded from open window in the ILD. The silicon crystal grows vertically and then laterally to cover the ILD (Fig. 33) [98]. In principle, the quality of devices fabricated on these epitaxial layers can be as good as those fabricated underneath on the seed wafer surface, since the grown layer is single crystal with few defects. However, the high temperatures ($\sim$1000 °C) involved in this process cause significant degradation in the quality of devices on lower layers. Also, this technique cannot be used over metallization layers. Low-temperature silicon epitaxy using ultra-high-vacuum chemical vapor deposition (UHV-CVD) has been recently developed [109]. However, this process is not yet manufacturable.

3) Processed Wafer Bonding: An attractive alternative is to bond two fully processed wafers on which devices are fabricated on the surface, including some interconnects, such that the wafers completely overlap (Fig. 34) [96], [110]. Interchip vias are etched to electrically connect both wafers after metallization and prior to the bonding process at $\sim$400 °C (discussed in Section VI-B). This technique is very suitable for further processing or the bonding of more pairs in this vertical fashion. Other advantages of this technology lie in the similar electrical properties of devices on all active levels and the independence of processing temperature since all chips can be fabricated separately and later bonded. One limitation of this technique is its lack of precision (best-case alignment is ± 2 μm), which restricts the interchip communication to global metal lines. However, for applications where each chip is required to perform independent processing before communicating with its neighbor, this technology can prove attractive. Also, this limitation can be eliminated to a large extent by using bonding pads which can be sized for alignment [see Fig. 37(b)], provided that the footprint area is sufficient. Additionally, bonding techniques based on the thermocompression of metal pads [110] offer low thermal-resistance interfaces between bonded wafers, which can help in heat dissipation.

4) Solid Phase Crystallization (SPC): As an alternative to high-temperature epitaxial growth discussed above, low-temperature deposition and crystallization of amorphous silicon (a-Si), on top of the lower active layer devices, can be employed. The amorphous film can be randomly crystallized to form a polysilicon film [111]–[114]. Device performance can be enhanced by eliminating the grain boundaries in the polysilicon film. For this purpose, local crystallization can be induced using low-temperature processes ($<$600 °C) such as using patterned seeding of Germanium (Fig. 35) [97], [115]. In this method, Ge seeds implanted in narrow patterns made on a-Si can be used to induce lateral crystallization and inhibit additional nucleation. This results in the formation of small islands, which are nearly single-crystal. CMOS transistors can then be fabricated within these islands to give SOI-like performance. Another approach based on the seeding technique employs metal (Ni) seeding to induce simultaneous lateral recystallization and dopant activation after the fabrication of the entire transistor on an a-Si layer. This technique known as the metal induced lateral crystallization (MILC) (see Fig. 36) [116], [117] offers even lower thermal budget ($<$500 °C) and can be employed to fabricate high-performance devices (MOSFETS or optical devices) on upper active layers even with metallization layers below.

The SPC technique offers the flexibility of creating multiple active layers and is compatible with current CMOS processing environments. Recent results using the MILC technique prove the feasibility of building high-performance devices at low processing temperatures, which can be compatible with lower level metallization [118]. It is found that the electrical characteristics of these devices are still inferior to single-crystal devices [119]. However, technological advances to overcome the thermal budget problem have been made to allow fabrication of high-performance devices using SPC [120]–[122].

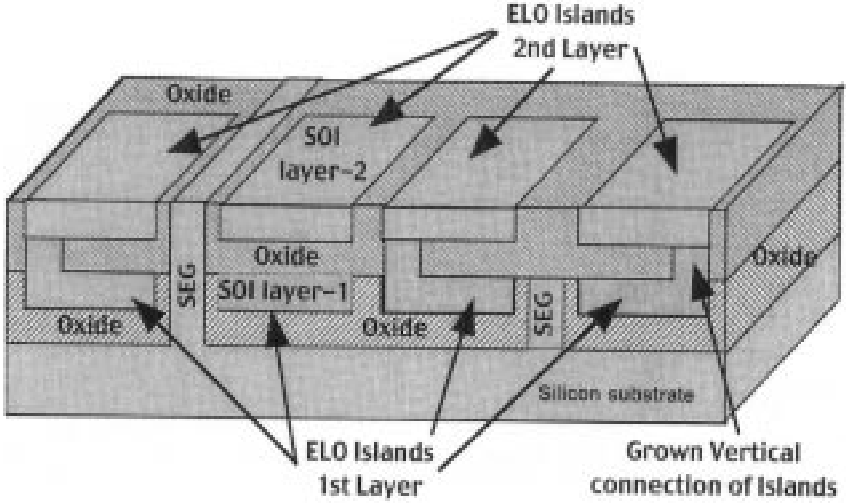


Fig. 33. Schematic of an epitaxially grown second active layer. ELO denotes epitaxial layer overgrowth. (Courtesy of Gerold W. Neudeck, Purdue University, West Lafayette, IN.)

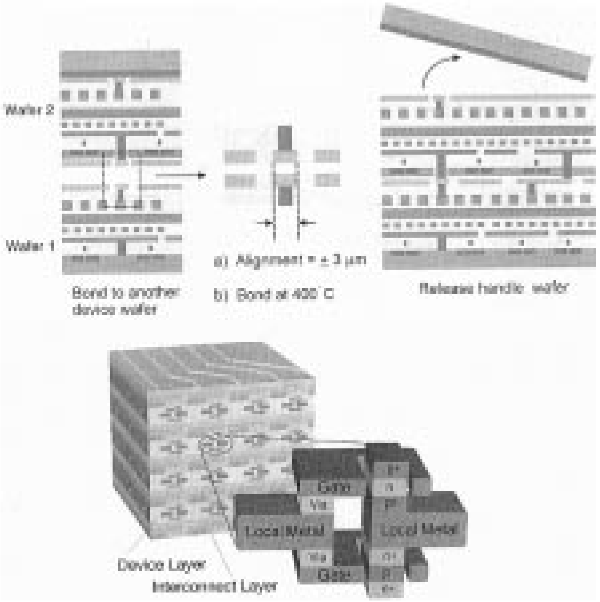


Fig. 34. Schematic of final steps used in one of the wafer bonding technologies based on metal thermocompression (top) and a finished 3-D chip (bottom). (Courtesy of Rafael Reif and Dimitri Antoniadis, Massachusetts Institute of Technology, Cambridge, MA.)

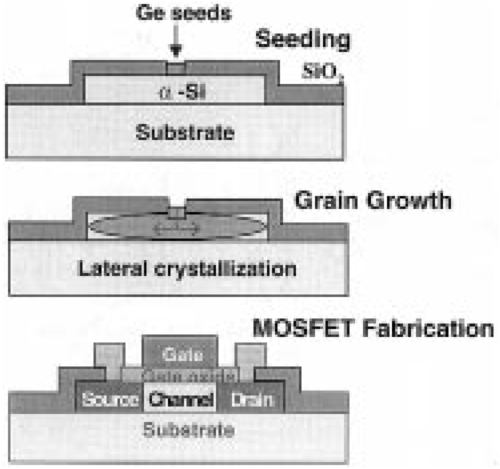


Fig. 35. Schematic of the Ge seeded solid phase crystallization (SPC) process flow.

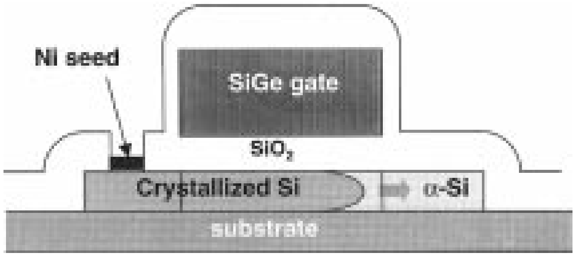
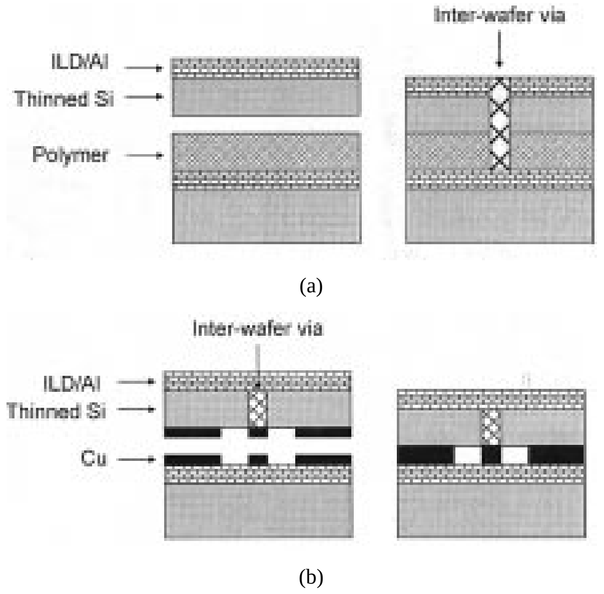


Fig. 36. Schematic of the MILC process flow using Ni seeding.

It is possible to conceive of several 3-D circuits for which SPC will be a suitable technology, such as in upper-level nonvolatile memory, or by simply sizing up the upper level transistors to match their single-crystal CMOS counterparts. For example, deep-submicrometer polysilicon TFTs [123], stacked SRAM cells [124], [125], and EEPROM cells [126] have already been demonstrated. With technological improvements, the MILC (Ni seeding) process can be used to fabricate islands of single-grain devices to maximize circuit performance.

B. *Vertical Interlayer Interconnect Technology Options*

The performance modeling presented in this study directly relates improved chip performance with increased utility of VILICs. It is therefore important to understand how to connect different active layers with a reliable and compatible process. Upper-layer processing needs to be compatible with metal lines underneath connecting lower layer devices and metal layers. With Cu technologies, this limits the processing temperatures to <450 °C for upper layers. Otherwise, Cu diffusion through barrier layers, and the reliability and thermal stability of material interfaces can degrade significantly. Tungsten is a refractory metal that can be used to withstand higher processing temperatures, but it



(a)

(b)

Fig. 37. Schematic of the wafer bonding techniques (a) with adhesive layer of polymer in between, and (b) through thermocompression of copper metal. (Courtesy of Rafael Reif, Massachusetts Institute of Technology, Cambridge, MA.)

has higher resistivity. Current via technology can also be employed to achieve VILIC functionality. The underlying assumption here requires that intralayer gates are interconnected using regular horizontal metal wires and vias, while interlayer interconnects can be VILICs connecting the wiring network for each layer, as schematically illustrated in Fig. 11.

Recently, interlayer (VILIC) metallization schemes for 3-D ICs have been demonstrated using direct wafer bonding. These techniques are based on the bonding of two wafers with their active layers connected through high aspect ratio vias, which serve as VILICs. One method is based on the optically adjusted bonding of a thinned ($\sim$10 μm) top wafer to a bottom wafer with an organic adhesive layer of polyimide ($\sim$2 μm) in between [127]. Interchip vias are etched through the ILD (inter level dielectric), the thinned top Si wafer and through the cured adhesive layer, with an approximate depth of 20 μm prior to the bonding process [see Fig. 37(a)]. The interchip via made of chemical vapor deposited (CVD) TiN liner and CVD-W plug provides a vertical interconnect (VILIC) between the uppermost metallization levels of both layers. The bonding between the two wafers (misalignment $\leq$1 μm) is done using a flip-chip bonder with split beam optics at a temperature of 400 °C.

A second technique relies on the thermocompression bonding between metal pads in each wafer [110]. In this method, Cu–Ta pads on both wafers [illustrated in Fig. 37(b)] serve as electrical contacts between the interchip via on the top thinned Si wafer and the uppermost interconnects on the bottom Si wafer. The Cu–Ta pads can also function as small bond pads for wafer bonding. Additionally, dummy metal patterns can be made to increase the surface area for wafer bonding. The Cu–Ta bilayer pads with a combined thickness of 700 nm are fused together by applying a compressive force at 400 °C. This technique offers the advantage of a metal–metal interface that will lower the interface thermal resistance between the two wafers (and, hence, provide better heat conduction) and can be beneficial as a partial ground plane for lowering the electromagnetic effects discussed in Section IV-B.

VII. Summary

In this paper, we have motivated the need for 3-D IC technologies with multiple active layers, as a promising alternative to the present single Si layer IC technologies, to alleviate the interconnect delay problems in near-future high-performance logic circuits, and to realize large scale integration of heterogeneous technologies in one single die.

In Section I, the interconnect delay problem associated with Cu/low-k technologies was discussed using estimated delay values based on the data from the ITRS. The implications of material effects arising at deep-submicrometer dimensions such as increasing metal resistivity of copper due to increased electron surface scattering and the effect of a finite barrier layer thickness on line resistance were quantified. The increasing impact of interconnect delays on VLSI design was also discussed and the limitations of various proposed solutions to overcome the interconnect problem were highlighted, especially in light of ITRS-based interconnect trends and their associated effects. It was concluded that Cu/low-k interconnects alone will not be able to overcome the deep-submicrometer interconnect problems, and that the existing design-based solutions are also not adequate to deal with the wiring problem. Additionally, various limitations of the existing planar (2-D) ICs with regards to their utility for heterogeneous integration of technologies were also discussed.

In Section III, a detailed performance analysis methodology was presented for the 3-D ICs to predict area, delay, and power dissipation, and provide examples of some of these tradeoffs which result in area and/or delay reduction over the 2-D case. A scheme to optimize the interconnect distribution among different interconnect tiers was also presented and the effect of transferring the repeaters to upper Si layers was quantified in this analysis for a two-layer 3-D chip. Our analysis predicts significant performance improvements over the 2-D case. The primary target technology for this analysis has been the ITRS-based 50-nm node with two active layers of silicon. Other technology nodes with two active layers were also considered. It was shown that the availability of additional silicon layers gives extra flexibility to designers which can be exploited to minimize area, improve performance and power dissipation, or any combinations of these.

Additionally, in Section IV, we addressed some of the concerns associated with 3-D circuits including that of heat dissipation. An analytical thermal model for estimating the temperature rise of individual active layers in 3-D ICs was presented. It was demonstrated that, for circuits with two silicon layers running at maximum performance, maintenance of acceptable die temperatures might require advanced packaging and heat-sinking technologies. Implications of reliability and electromagnetic interactions (such as capacitance and inductance effects) arising in 3-D ICs were also briefly discussed.

In Section V, we highlighted some scenarios in current and future VLSI and SoC type applications involving mixed signals and technologies, where the use of 3-D circuits will have an immediate and beneficial impact on performance. We also briefly discussed the implications of using this technology on the design process, as conventional VLSI design methodologies and tools and gate level and architecture level synthesis algorithms need to be suitably adapted. Finally, in Section VI, an overview of some of the manufacturing technologies under investigation, which can be used to fabricate these circuits, was provided.

VIII. Conclusion

Deep-submicrometer VLSI interconnect scaling trends and the growing need for heterogeneous integration of technologies in one single die have created the necessity to seek alternatives to the existing (2-D) single-active-layer ICs. In this paper, we have shown that 3-D ICs are an attractive chip architecture that can alleviate the interconnect related problems such as delay and power dissipation and can also facilitate integration of heterogeneous technologies in one single chip. In fact, several applications of 3-D ICs have been recently demonstrated [128]–[131], which show the potential of this technology for effective implementations of *SoC* designs that are expected to form the backbone of most future electronic systems. While many technological challenges need to be overcome for the successful realization of *completely monolithic* 3-D ICs, advanced 3-D *packaging techniques* to realize heterogeneous ICs [132] can be precursors to the future monolithic 3-D ICs.

Acknowledgment

The authors would like to acknowledge J. Hutchby of SRC for providing initial seed funding and A. Mehrotra, University of Illinois at Urbana–Champaign, for several technical discussions during the initial phase of this project. They would also like to thank G. Chandra, T.-Y. Chiang, C.-O. Chui, S. Im, A. Joshi, and R. Shenoy, all from Stanford University, for several interesting discussions and for providing feedback.

References

[1] *The International Technology Roadmap for Semiconductors (ITRS)*, 1999.

[2] C. R. Barrett, "Microprocessor evolution and technology impact," in *Symp. VLSI Tech. Dig.*, 1993, pp. 7–10.

[3] C. Hu, "MOSFET scaling in the next decade and beyond," *Semicond. Int.*, pp. 105–114, 1994.

[4] B. Davari, R. H. Dennard, and G. G. Shahidi, "CMOS scaling for high performance and low power—The next ten years," *Proc. IEEE*, vol. 83, pp. 595–606, Apr. 1995.

[5] G. A. Sai-Halasz, "Performance trends in high-end processors," *Proc. IEEE*, vol. 83, pp. 20–36, Jan. 1995.

[6] K. C. Saraswat and F. Mohammadi, "Effect of interconnection scaling on time delay of VLSI circuits," *IEEE Trans. Electron Devices*, vol. ED-29, pp. 645–650, 1982.

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

[8] J. D. Meindl, "Low power microelectronics: Retrospect and prospect," *Proc. IEEE*, vol. 83, pp. 619–635, Apr. 1995.

[9] S.-Y. Oh and K.-J. Chang, "2001 needs for multi-level interconnect technology," *Circuits Dev.*, pp. 16–21, 1995.

[10] M. T. Bohr and Y. A. El-Mansy, "Technology for advanced high-performance microprocessors," *IEEE Trans. Electron Devices*, vol. 45, pp. 620–625, Mar. 1998.

[11] R. H. J. M. Otten and R. K. Brayton, "Planning for performance," in *Proc. 35th Annual Design Automation Conf.*, 1998, pp. 122–127.

[12] K. Banerjee, A. Mehrotra, W. Hunter, K. C. Saraswat, K. E. Goodson, and S. S. Wong, "Quantitative projections of reliability and performance for low-k/Cu interconnect systems," in *38th IEEE Ann. Int. Reliability Physics Symp. Proc.*, 2000, pp. 354–358.

[13] K. Yang, S. Sidiropoulos, and M. Horowitz, "The limits of electrical signalling," in *Symp. High Performance Interconnects, Hot Interconnects V*, Aug. 1997.

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

[15] S. Venkatesan *et al.*, "A high performance 1.8V, 0.20 μm CMOS technology with copper metallization," in *IEDM Tech. Dig.*, 1997, pp. 769–772.

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

[17] N. Rohrer *et al.*, "A 480MHz RISC microprocessor in a 0.12 μm $\mathrm{L_{eff}}$ CMOS technology with copper interconnects," in *Int. Solid-State Circuits Conf., Tech. Dig.*, 1998, pp. 240–241.

[18] B. Zhao *et al.*, "A Cu/low-k dual damascene interconnect for high performance and low cost integrated circuits," in *Symp. VLSI Technology, Tech. Dig.*, 1998, pp. 28–29.

[19] T. J. Licata, E. G. Colgan, J. M. E. Harper, and S. E. Luce, "Interconnect fabrication processes and the development of low-cost wiring for CMOS products," *IBM J. Res. Develop.*, vol. 39, pp. 419–435, July 1995.

[20] M. J. Hampden-Smith and T. T. Kodas, "Copper etching: New chemical approaches," *MRS Bull.*, vol. 18, p. 39, June 1993.

[21] S. P. Murarka, J. Steigerwald, and R. J. Gutmann, "Inlaid copper multilevel interconnections using planarization by chemical–mechanical polishing," *MRS Bull.*, vol. 18, pp. 45–51, June 1993.

[22] J. G. Ryan, R. M. Geffken, N. R. Poulin, and J. R. Paraszczak, "The evolution of interconnect technology at IBM," *IBM J. Res. Develop.*, vol. 39, pp. 371–381, July 1995.

[23] S.-Q. Wang, "Barriers against copper diffusion into silicon and drift through silicon dioxide," *MRS Bull.*, vol. 19, pp. 30–40, Aug. 1994.

[24] R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd ed. New York: Wiley, 1986, pp. 1–56.

[25] K. A. Perry, "Chemical mechanical polishing: The impact of a new technology on an industry," in *Symp. VLSI Technol., Tech. Dig.*, 1998, pp. 2–5.

[26] K. Banerjee, A. Mehrotra, A. Sangiovanni-Vincentelli, and C. Hu, "On thermal effects in deep sub-micron VLSI interconnects," in *Proc. 36th ACM Design Automation Conf.*, 1999, pp. 885–891.

[27] J. C. Anderson, Ed., *The Use of Thin Films in Physical Investigation*. London, U.K.: Academic, 1966.

[28] F. Chen and D. Gardner, "Influence of line dimensions on the resistance of Cu interconnections," *IEEE Electron Device Lett.*, vol. 19, pp. 508–510, Dec. 1998.

[29] *Handbook of Thin Film Technology*, L. I. Maissel and R. Glang, Eds., McGraw-Hill, New York, 1970.

[30] J. P. McVittie *et al.*, *SPEEDIE 3.5 Manual*. Stanford, CA: Stanford Univ. Press, 1998.

[31] L. P. P. P. van Ginneken, Magma Design Automation Inc., Cupertino, CA, private communication.

[32] Synopsys Design Compiler—White Paper (1997). [Online]. Available: http://www.synopsys.com/products/logic/dc_wp97.html#6

[33] D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron," in *Proc. Int. Conf. Computer Aided Design*, 1998, pp. 203–211.

[34] ——, "A global wiring paradigm for deep submicron design," *IEEE Trans. Computer-Aided Design*, vol. 19, pp. 242–252, 2000.

[35] W. Gosti, A. Narayan, R. K. Brayton, and A. L. Sangiovanni-Vincentelli, "Wireplanning in logic synthesis," in *Proc. Int. Conf. Computer Aided Design*, 1998, pp. 26–33.

[36] J. Grodstein, E. Lehman, H. Harkness, B. Grundmann, and Y. Watanabe, "A delay model for logic synthesis of continuously sized networks," in *Proc. Int. Conf. Computer Aided Design*, 1995.

[37] S. P. Khatri, A. Mehrotra, R. K. Brayton, A. Sangiovanni-Vincentelli, and R. H. J. M. Otten, "A novel VLSI layout fabric for deep sub-micron applications," in *Proc. 36th ACM Design Automation Conf.*, 1999, pp. 491–496.

[38] K. Banerjee, Ph.D. dissertation, Univ. California, Berkeley, 1999.

[39] W. J. Dally, "Interconnect-limited VLSI architecture," in *Int. Interconnect Technology Conf. Proc.*, 1999, pp. 15–17.

[40] M. J. M. Pelgrom, "System-on-chip concepts," in *ULSI Devices*, C. Y. Chang and S. M. Sze, Eds. New York: Wiley Inter-Science, 2000.

[41] H. De Man, "System design challenges in the post PC era," presented at the 37th ACM Design Automation Conf., 2000.

[42] S. J. Souri, K. Banerjee, A. Mehrotra, and K. C. Saraswat, "Multiple Si layer ICs: Motivation, performance analysis, and design implications," in *Proc. 37th ACM Design Automation Conf.*, 2000, pp. 873–880.

[43] A. V. Krishnamoorthy *et al.*, "3-D integration of MQW modulators over active submicron CMOS circuits: 375 Mb/s transimpedance receiver-transmitter circuit," *IEEE Photon. Technol. Lett.*, vol. 7, pp. 1288–1290, Nov. 1995.

[44] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI)—Part I: Derivation and validation," *IEEE Trans. Electron Devices*, vol. 45, Mar. 1998.

[45] L. Robinson, L. A. Glasser, and D. A. Antoniadis, "A simple interconnect delay model for multilayer integrated circuits," in *IEEE VMIC Conf.*, 1986.

[46] B. S. Landman and R. L. Russo, "On a pin versus block relationship for partitions of logic graphs," *IEEE Trans. Comput.*, vol. C-20, Dec. 1971.

[47] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI)—Part II: Applications to clock frequency, power dissipation, and chip size estimation," *IEEE Trans. Electron Devices*, vol. 45, Mar. 1998.

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

[49] S. A. Kuhn, M. B. Kleiner, P. Ramm, and W. Weber, "Thermal analysis of vertically integrated circuits," in *IEDM Tech. Dig.*, 1995, pp. 487–490.

[50] S. Im and K. Banerjee, "Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs," in *IEDM Tech. Dig.*, 2000, pp. 727–730.

[51] K. Banerjee, "Thermal effects in deep submicron VLSI interconnects," in *IEEE Int. Symp. Quality Electronic Design*, 2000.

[52] K. Banerjee, A. Amerasekera, G. Dixit, and C. Hu, "The effect of interconnect scaling and low-k dielectric on the thermal characteristics of the IC metal," in *IEDM Tech. Dig.*, 1996, pp. 65–68.

[53] K. E. Goodson and Y. S. Ju, "Heat conduction in novel electronic films," *Annu. Rev. Mater. Sci.*, vol. 29, pp. 261–293, 1999.

[54] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for VLSI," *IEEE Electron Device Lett.*, vol. EDL-2, pp. 126–129, May 1981.

[55] K. E. Goodson, Stanford University, private communication.

[56] T.-Y. Chiang, K. Banerjee, and K. C. Saraswat, "Effect of via separation and low-k dielectric materials on the thermal characteristics of Cu interconnects," in *Tech. Dig. IEEE Int. Electron Devices Meeting*, 2000, pp. 261–264.

[57] A. H. Ajami, M. Pedram, and K. Banerjee, "Effects of non-uniform substrate temperature on the clock signal integrity in high performance designs," in *Custom Integrated Circuits Conf.*, 2001, pp. 233–236.

[58] A. H. Ajami, K. Banerjee, and M. Pedram, "Non-uniform chip-temperature dependent signal integrity," in IEEE Symp. VLSI Technology, 2001, to be published.

[59] S. A. Kuhn, M. B. Kleiner, P. Ramm, and W. Weber, "Interconnect capacitances, crosstalk, and signal delay in vertically integrated circuits," in *IEDM Tech. Dig.*, 1995, pp. 487–490.

[60] C.-K. Cheng, J. Lillis, S. Lin, and N. Chang, *Interconnect Analysis and Synthesis*. New York: Wiley, 1999.

[61] K. Banerjee and A. Mehrotra, "Accurate analysis of on-chip inductance effects and implications for optimal repeater insertion and technology scaling," in IEEE Symp. VLSI Circuits, 2001, to be published.

[62] ——, "Analysis of on-chip inductance effects using a novel performance optimization methodology for distributed RLC interconnects," in Proc. 38th ACM Design Automation Conf., 2001, to be published.

[63] Y.-C. Lu, K. Banerjee, M. Celik, and R. W. Dutton, "A fast analytical technique for estimating the bounds of on-chip clock wire inductance," in *Proc. IEEE Custom Integrated Circuits Conf.*, 2001, pp. 241–244.

[64] J. Cong and L. He, "An efficient technique for device and interconnect optimization in deep submicron designs," in *Int. Symp. Physical Design*, 1998, pp. 45–51.

[65] P. D. Fisher, "Clock cycle estimation for future microprocessor generations," Tech. Rep., SEMATECH, 1997.

[66] D. Greenhill *et al.*, "A 330 MHz 4-way superscalar microprocessor," in *ISSCC Dig. Tech. Papers*, 1997, pp. 166–167.

[67] M. B. Kleiner, S. A. Kuhn, P. Ramm, and W. Weber, "Performance improvement of the memory hierarchy of RISC-systems by application of 3-D technology," *IEEE Trans. Comp., Packag., Manufact. Technol. B*, vol. 19, pp. 709–718, 1996.

[68] B. Razavi, "Challenges and trends in RF design," in *Proc. 9th Annu. IEEE Int. ASIC Conf. and Exhibit*, 1996, pp. 81–86.

[69] J. M. Rabaey, *Digital Integrated Circuits: A Design Perspective*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[70] H. Kawaguchi and T. Sakurai, "A reduced clock-swing flip-flop (RCSFF) for 63% power reduction," *IEEE J. Solid-State Circuits*, vol. 33, pp. 807–811, May 1998.

[71] T. Sakurai, "Design challenges for 0.1 μm and beyond," in *Proc. ASP DAC*, 2000, pp. 553–558.

[72] J. W. Goodwin, F. J. Leonberger, S. C. Kung, and R. A. Athale, "Optical interconnections for VLSI systems," *Proc. IEEE*, vol. 72, pp. 850–866, 1984.

[73] D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," Proc. IEEE, to be published.

[74] A. L. Lentine, L. M. F. Chirovsky, and T. K. Woodward, "Optical energy considerations for diode-clamped smart pixel optical receivers," *IEEE J. Quantum Electron.*, vol. 30, pp. 1167–1171, 1994.

[75] G. A. Keeler, B. E. Nelson, D. Agarwal, and D. A. B. Miller, "Skew and jitter removal using optical pulses for optical interconnection," *IEEE Photon. Technol. Lett.*, vol. 12, pp. 714–716, 2000.

[76] E. A. De Souza, M. C. Nuss, W. H. Knox, and D. A. B. Miller, "Wavelength-division multiplexing with femtosecond pulses," *Opt. Lett.*, vol. 20, pp. 1166–1168, 1995.

[77] D. Agarwal, G. A. Keeler, B. E. Nelson, and D. A. B. Miller, "Wavelength division multiplexed optical interconnects using femtosecond optical pulses," in *Proc. IEEE LEOS Annu. Meeting*, 1999, pp. 828–829.

[78] K. W. Goossen *et al.*, "GaAs MQW modulators integrated with silicon CMOS," *IEEE Photon. Technol. Lett.*, vol. 7, pp. 360–362, Apr. 1995.

[79] A. V. Krishnamoorthy and K. W. Goossen, "Optoelectronic-VLSI: Photonics integrated with VLSI circuits," *IEEE J. Select. Topics Quantum Electron.*, vol. 4, pp. 899–912, June 1998.

[80] T. K. Woodward and A. V. Krishnamoorthy, "1-Gb/s integrated optical detectors and receivers in commercial CMOS technologies," *IEEE J. Select. Topics Quantum Electron.*, vol. 5, pp. 146–156, 1999.

[81] L. C. Kimerling, "Photons to the rescue: Microelectronics becomes microphotonics," *The Electrochemical Society Interface*, pp. 28–31, 2000.

[82] C. T. Chuang, P. F. Lu, and C. J. Anderson, "SOI for digital CMOS VLSI: Design considerations and advances," *Proc. IEEE*, vol. 86, pp. 689–720, Apr. 1998.

[83] D. Allen, D. Behrends, and B. Stanisic, "Converting a 64b PowerPC processor from CMOS bulk to SOI technology," in *Proc. 36th ACM Design Automation Conf.*, 1999, pp. 892–897.

[84] M. W. Geis, D. C. Flanders, D. A. Antoniadis, and H. I. Smith, "Crystalline silicon on insulators by graphoepitaxy," in *IEDM Tech. Dig.*, 1979, pp. 210–212.

[85] J. P. Colinge and E. Demoulin, "ST-CMOS (stacked transistor CMOS): A double-poly-NMOS-compatible CMOS technology," in *IEDM Tech. Dig.*, 1981, pp. 557–560.

[86] G. T. Goeloe, E. W. Maby, D. J. Silversmith, R. W. Mountain, and D. A. Antoniadis, "Vertical single-gate CMOS inverters on laser-processed multilayer substrates," in *IEDM Tech. Dig.*, 1981, pp. 554–556.

[87] S. Kawamura, N. Sasaki, T. Iwai, M. Nakano, and M. Takagi, "Three-dimensional CMOS IC's fabricated by using beam recrystallization," *IEEE Electron Device Lett.*, vol. EDL-4, pp. 366–368, Oct. 1983.

[88] S. Akiyama, S. Ogawa, M. Yoneda, N. Yoshii, and Y. Terui, "Multilayer CMOS device fabricated on laser recrystallized silicon islands," in *IEDM Tech. Dig.*, 1983, pp. 352–355.

[89] M. Nakano, "3-D SOI/CMOS," in *IEDM Tech. Dig.*, 1984, pp. 792–795.

[90] K. Sugahara, T. Nishimura, S. Kusunoki, Y. Akasaka, and H. Nakata, "SOI/SOI/bulk-Si triple level structure for three-dimensional devices," *IEEE Electron Device Lett.*, vol. EDL-7, pp. 193–195, Mar. 1986.

[91] Y. Akasaka and T. Nishimura, "Concept and basic technologies for 3-D IC structure," in *IEDM Tech. Dig.*, 1986, pp. 488–491.

[92] S. Tatsuno, "Japan's push into creative semiconductor research: 3-dimension IC's," *Solid State Technol.*, pp. 29–30, Mar. 30, 1987.

[93] T. Nishimura, Y. Inoue, K. Sugahara, S. Kusunoki, T. Kumamoto, S. Nakagawa, M. Nakaya, Y. Horiba, and Y. Akasaka, "Three dimensional IC for high performance image signal processor," in *IEDM Tech. Dig.*, 1987, pp. 111–114.

[94] T. Kunio, K. Oyama, Y. Hayashi, and M. Morimoto, "Three dimensional ICs, having four stacked active device layers," in *IEDM Tech. Dig.*, 1989, pp. 837–840.

[95] S. Strickland *et al.*, "VLSI design in the 3rd dimension," in *Integration*. New York: Elsevier Science, 1998, pp. 1–16.

[96] D. Antoniadis, Massachusetts Institute of Technology, Cambridge, MA. private communication.

[97] V. Subramanian and K. C. Saraswat, "High-performance germanium-seeded laterally crystallized TFT's for vertical device integration," *IEEE Trans. Electron Devices*, vol. 45, pp. 1934–1939, Sept. 1998.

[98] G. W. Neudeck, S. Pae, J. P. Denton, and T. Su, "Multiple layers of silicon-on-insulator for nanostructure devices," *J. Vac. Sci. Technol. B*, vol. 17, no. 3, pp. 994–998, 1999.

[99] K. C. Saraswat, S. J. Souri, V. Subramanian, A. R. Joshi, and A. W. Wang, "Novel 3-D structures," in *IEEE Int. SOI Conf.*, 1999, pp. 54–55.

[100] K. C. Saraswat, K. Banerjee, A. Joshi, P. Kalavade, S. J. Souri, and V. Subramanian, "3-D ICs with multiple Si layers: Performance analysis, and technology," in *197th Meeting Electrochemical Soc.*, Toronto, Canada, May 14–18, 2000.

[101] S. A. Kuhn, M. B. Kleiner, P. Ramm, and W. Weber, "Performance modeling of the interconnect structure of a three-dimensional integrated RISC processor/cache system," *IEEE Trans. Comp., Packag., Manufact. Technol. B*, vol. 19, no. 4, pp. 719–727, 1996.

[102] S. J. Souri and K. C. Saraswat, "Interconnect performance modeling for 3D integrated circuits with multiple Si layers," in *Int. Interconnect Technology Conf. Proc.*, 1999, pp. 24–26.

[103] A. Rahman, A. Fan, J. Chung, and R. Reif, "Wire-length distribution of three-dimensional integrated circuits," in *Int. Interconnect Technology Conf. Proc.*, 1999, pp. 233–235.

[104] R. Zhang, K. Roy, and D. B. Jones, "Architecture and performance of 3-dimensional SOI circuits," *IEEE Int. SOI Conf.*, pp. 44–45, 1999.

[105] A. W. Wang and K. C. Saraswat, "A strategy for modeling of variations due to grain size in polycrystalline thin film transistors," *IEEE Trans. Electron Devices*, vol. 47, pp. 1035–1043, 2000.

[106] V. Subramanian, M. Toita, N. R. Ibrahim, S. J. Souri, and K. C. Saraswat, "Low-leakage Germanium-seeded laterally-crystallized single-grain 100 nm TFTs for vertical integration applications," *IEEE Electron Device Lett.*, vol. 20, pp. 341–343, July 1999.

[107] A. Kohno, T. Sameshima, N. Sano, M. Sekiya, and M. Hara, "High performance poly-Si TFTs fabricated using pulsed laser annealing and remote plasma CVD with low temperature processing," *IEEE Trans. Electron Devices*, vol. 42, no. 2, pp. 251–257, 1995.

[108] M. A. Crowder, P. G. Carey, P. M. Smith, R. S. Sposili, H. S. Cho, and J. S. Im, "Low-temperature single crystal Si TFT's fabricated on Si-films processed via sequential lateral solidification," *IEEE Electron Device Lett.*, vol. 19, no. 8, pp. 306–308, 1986.

[109] H.-Y. Lin, C.-Y. Chang, T. F. Lei, J.-Y. Cheng, H.-C. Tseng, and L.-P. Chen, "Characterization of polycrystalline silicon thin film transistors fabricated by ultrahigh-vacuum chemical vapor deposition and chemical mechanical polishing," *Jpn. J. Appl. Phys.*, pt. 1, vol. 36, pp. 4278–4282, July 1997.

[110] A. Fan, A. Rahman, and R. Reif, "Copper wafer bonding," *Electrochem. Solid State Lett.*, vol. 2, pp. 534–536, 1999.

[111] T. Noguchi, "Appearance of single-crystalline properties in fine-patterned Si thin film transistors (TFTs) by solid phase crystallization (SPC)," *Jpn. J. Appl. Phys.*, pt. 2, vol. 32, pp. 1584–1587, Nov. 1993.

[112] T. W. Little, H. Koike, K. Takahara, T. Nakazawa, and H. Oshima, "A 9.5-in. 1.3-Mpixel low-temperature poly-Si TFT-LCD fabricated by solid-phase crystallization of very thin films and an ECR-CVD gate insulator," *J. Soc. Inform. Display*, vol. 1/2, pp. 203–209, 1993.

[113] N. Yamauchi, "Polycrystalline silicon thin films processed with silicon ion implantation and subsequent solid-phase crystallization: Theory, experiments, and thin-film transistor applications," *J. Appl. Phys.*, vol. 75, no. 7, pp. 3235–3257, 1994.

[114] D. N. Kouvatsos, A. T. Voutsas, and M. K. Hatalis, "Polycrystalline silicon thin film transistors fabricated in various solid phase crystallized films deposited on glass substrates," *J. Electron. Mat.*, vol. 28, no. 1, pp. 19–25, 1999.

[115] J. A. Tsai, A. J. Tang, T. Noguchi, and R. Reif, "Effects of Ge on material and electrical properties of polycrystalline $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ for thin film transistors," *J. Electrochem. Soc.*, vol. 142, no. 9, pp. 3220–3225, 1995.

[116] S.-W. Lee and S.-K. Joo, "Low temperature poly-Si thin film transistor fabrication by metal-induced lateral crystallization," *IEEE Electron Device Lett.*, vol. 17, no. 4, pp. 160–162, 1983.

[117] S. Y. Yoon, S. K. Kim, J. Y. Oh, Y. J. Choi, W. S. Shon, C. O. Kim, and J. Jang, "A high-performance polycrystalline silicon thin-film transistor using metal-induced crystallization with Ni solution," *Jpn. J. Appl. Phys.*, pt. 1, pp. 7193–7197, Dec. 1998.

[118] A. R. Joshi and K. C. Saraswat, "Sub-micron thin film transistors with metal induced lateral crystallization," in *Proc. 196th Meeting Electrochemical Soc.*, Honolulu, HI, 1999.

[119] K. C. Saraswat, K. Banerjee, A. R. Joshi, P. Kalavade, P. Kapur, and S. J. Souri, "3-D ICs: Motivation, performance analysis, and technology," in *Proc. 26th Eur. Solid-State Circuits Conf. (ESSCIRC)*, Stockholm, Sweden, Sept. 19–21, 2000.

[120] J. Nakata and K. Kajiyama, "Novel low-temperature recrystallization of amorphous silicon by high energy beam," *Appl. Phys. Lett.*, pp. 686–688, 1982.

[121] Y. W. Choi, J. N. Lee, T. W. Jang, and B. T. Ahn, "Thin-film transistors fabricated with poly-Si films crystallized at low temperature by microwave annealing," *IEEE Electron Device Lett.*, vol. 20, pp. 2–4, Jan. 1999.

[122] A. Heya, A. Masuda, and H. Matsumura, "Low-temperature crystallization of amorphous silicon using atomic hydrogen generated by catalytic reaction on heated tungsten," *Appl. Phys. Lett.*, vol. 74, no. 15, pp. 2143–2145, 1999.

[123] R. K. Watts and J. T. C. Lee, "Tenth-micron polysilicon thin-film transistors," *IEEE Electron Device Lett.*, vol. 14, pp. 515–517, Nov. 1993.

[124] M. Rodder and S. Aur, "Utilization of plasma hydrogenation in stacked SRAMs with poly-Si PMOSFETs and bulk Si NMOSFETs," *IEEE Electron Device Lett.*, vol. 12, pp. 233–235, May 1991.

[125] T. Yamanaka *et al.*, "Advanced TFT SRAM cell technology using a phase-shift lithography," *IEEE Trans. Electron Devices*, vol. 42, pp. 1305–1312, July 1995.

[126] M. Cao, T. Zhao, K. C. Saraswat, and J. D. Plummer, "A simple EEPROM cell using twin polysilicon thin film transistor," *IEEE Electron Device Lett.*, vol. 15, pp. 304–306, Aug. 1994.

[127] P. Ramm *et al.*, "Three dimensional metallization for vertically integrated circuits," *Microelectron. Eng.*, vol. 37/38, pp. 39–47, 1997.

[128] H. Kurino *et al.*, "Intelligent image sensor chip with three dimensional structure," in *IEDM Tech. Dig.*, 1999, pp. 879–882.

[129] K.-W. Lee *et al.*, "Three-dimensional shared memory fabricated using wafer stacking technology," in *IEDM Tech. Dig.*, 2000, pp. 165–168.

[130] J. Burns *et al.*, "Three-dimensional integrated circuits for low-power, high-bandwidth systems on a chip," in *ISSCC Dig. Tech. Papers*, 2001, pp. 268–269.

[131] M. Koyanagi *et al.*, "Neuromorphic vision chip fabricated using three-dimensional integration technology," in *ISSCC Dig. Tech. Papers*, 2001, pp. 270–271.

[132] K. Ohsawa *et al.*, "3-D assembly interposer technology for next-generation integrated systems," in *ISSCC Dig. Tech. Papers*, 2001, pp. 272–273.

Kaustav Banerjee (Member, IEEE) received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1999.

Since March 1999, he has been with Stanford University as a Research Associate at the Center for Integrated Systems. He also works as a Technical Consultant in the EDA industry. His research interests include signal integrity, reliability and performance optimization issues in high-performance VLSI, high-speed mixed-signal and RF applications. He is also interested in various aspects of integrated heterogeneous circuits and systems including system-on-chip designs. At Stanford, he leads an interdisciplinary research team of ten doctoral students. As part of the MARCO Interconnect Focus Center at Stanford, he is actively involved in the research of 3-D ICs. He is also involved in several collaborative research initiatives with other leading Universities. He coadvises doctoral students in the Electrical Engineering Departments of the University of Southern California, Los Angeles, and the Swiss Federal Institute of Technology, Lausanne, Switzerland. In the past, he has held several summer research positions at the Semiconductor Process and Device Center of Texas Instruments Inc., Dallas, during 1993–1997. He has authored or coauthored over 50 research publications in archival journals and refereed international conferences and has presented numerous invited talks and tutorials. He is currently the Technical Program Chair of the International Symposium on Quality Electronic Design (ISQED) and also serves on the technical program committees of the International Symposium on Physical Design (ISPD), the EOS/ESD Symposium and the International Reliability Physics Symposium (IRPS).

Dr. Banerjee is the recipient of a Best Paper Award at the 2001 Design Automation Conference (DAC).

Shukri J. Souri received the B.A. degree in engineering science from Oxford University, Oxford, U.K., and the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1992 and 1994, respectively. He is currently pursuing the Ph.D. degree in the Electrical Engineering Department at Stanford.

From 1994 to 1997, he was a Member of Technical Staff with the Corporate Research Division, Raychem Corporation, Menlo Park, CA, where he worked on electroceramic semiconducting and ferroelectric materials and devices for circuit protection applications. His research interests include 3-D IC performance modeling and interconnect network architecture. He has also worked on 3-D integration using advanced seeding and crystallization techniques. He is a co-inventor on a number of U.S. patents and has several publications in his areas of interest.

Pawan Kapur was born and raised in IIT Kanpur, India. He received the B.S. degree in physics and mathematics *summa cum laude* from Moravian College, Bethlehem, PA, in 1995 and the M.S. degree in electrical engineering from Stanford University, CA, in 1998. He is currently pursuing a Ph.D. in electrical engineering, also from Stanford University.

The focus of his current research is on interconnect modeling for integrated circuits which includes exploring limitations of both electrical interconnects and assessing the advantages and limitations of possible optical interconnect integration in ICs.

Krishna C. Saraswat (Fellow, IEEE) received the B.E. degree in electronics and telecommunications in 1968 from Birla Institute of Technology and Science, Pilani, India, and the M.S. and Ph.D. degrees in electrical engineering in 1969 and 1974, respectively, from Stanford University, Stanford, CA.

During 1969–1970, he worked on microwave transistors at Texas Instruments, Dallas, TX and since 1971, he has been with Stanford University, California, where presently he is a Professor of Electrical Engineering and Associate Director of the NSF/SRC Engineering Research Center for Environmentally Benign Semiconductor Manufacturing. During 1996–1997 he was the Director of the Integrated Circuits Laboratory at Stanford. He is working on a variety of problems related to new and innovative materials, device structures, and process technology of silicon devices and integrated circuits. Special areas of his interest are thin film MOS transistors (TFTs) on insulator for 3-D multilayer ICs; thin film technology for VLSI interconnections and contacts; process and equipment modeling; ultrathin MOS gate dielectrics; rapid thermal processing; and development of tools and methodology for simulation and control of a manufacturing technology. His group has developed several simulators for process, equipment and factory performance simulations, such as, SPEEDIE for etch and deposition simulation, SCOPE for IC factory performance simulations and a thermal simulator for RTP equipment design. Currently, he is also involved in the development of an interconnect process simulator. He has authored or coauthored more than 360 technical papers.

Prof. Saraswat received the Thomas D. Callinan Award by the Electrochemical Society in May 2000 for his contributions to the dielectric science and technology. He was Coeditor of the IEEE TRANSACTIONS ON ELECTRON DEVICES during 1988–1990. He is a member of The Electrochemical Society and The Materials Research Society.