3D ICs Interconnect Performance Modeling and Analysis

Ph.D. Dissertation Draft

Shukri J Souri (ssouri@stanford.edu)

Advisor: Prof. K. Saraswat

Co-Advisor: Dr. J. McVittie

3rd Reading Committee member: Prof. F. Pease

Chapter 1

INTRODUCTION

The unprecedented growth of the computer and the information technology industry is demanding ULSI circuits with increasing functionality and performance at minimum cost and power dissipation. ULSI circuits are being aggressively scaled to meet this demand. This in turn has introduced some very serious problems for the semiconductor industry. Although continuous scaling of ULSI circuits is reducing feature sizes, gate delays and interconnect cross-sections, it is also rapidly increasing interconnect (RC) delays [1]. The ever increasing complexity of ICs demands a greater number of integrated transistors and gates which exponentially require more wiring for interconnectivity. Consequently, ICs are growing in size and, on average, interconnects are required to carry signals across longer distances [2]. Interconnect RC delays are thus increasing and not only due to increasing lengths but also due to the scaling of interconnect cross-sectional dimensions.

The International Technology Roadmap for Semiconductors (ITRS) projects performance improvement of advanced ULSI circuits likely to saturate beyond the 100 nm technology node, due to the rapidly dominating interconnect performance limitations, unless a paradigm shift from present IC architecture is introduced [3]. Considerable effort has already gone into pushing the interconnect performance limit out into the future. For instance, the interconnect system architecture has increased significantly in complexity, introducing a hierarchical structure to incorporate several metal layers where longer wires are routed to higher tiers and enjoy larger cross-sections to reduce interconnect resistance [4,5]. Also, Cu has become the interconnect metal of choice due to its lower resistivity as compared to Al, again

to reduce resistance. Much work is also being pursued in low dielectric constant (low-k) materials to replace deposited SiO_2 as an inter-layer (ILD) and inter-metal dielectric (IMD) to reduce interconnect capacitance [6-10]. Together with Cu and the multi-tiered interconnect architecture these solutions, although indeed lower RC delays, still have their limitations and are already considered in the ITRS projections [11,12].

Furthermore, interconnect delay is only part of the overall problem facing complex, high-performance ICs of the near future. Power consumption and dissipation, for instance, is rapidly becoming unmanageable as the clock frequencies continue to climb [13-17]. Any increase in interconnect loading significantly increases the power consumption in high-performance chips. In fact, around 40-70% 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 [91,92]. 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. Moreover, there exists an increasing drive for the integration of disparate signals and technologies, introducing various system-on-a-chip (SoC) design concepts, for which existing planar (2-D) IC design may not be suitable.

In addressing interconnect performance limits, this thesis analyzes the limitations of the existing interconnect technologies and design methodologies and presents a novel 3-dimensional (3-D) chip design strategy that exploits the vertical dimension to alleviate the interconnect related problems and to facilitate SoC applications. A detailed analysis of interconnect performance limitations in existing technologies is presented in Chapter 2 along with a review of previous work in the area of 3-D integration [18-30]. Historically, 3-D

integration studies have focused mainly on technology issues without much consideration on the performance improvements reaped as a result of migration towards 3-D. As ICs are evolving away from device-size limits towards wire-pitch limits in determining chip area, a need developed for an overall systems level performance analysis of 3-D integration. This analysis, the first of its kind in the literature, is presented in Chapter 3 where an interconnect-centric 3-D integration solution for ICs is proposed and a comprehensive analytical treatment of futuristic 3-D ICs is developed. The analysis shows that by simply dividing a planar chip into separate blocks, each occupying a separate physical level interconnected by short and vertical inter-layer interconnects (VILICs) significant improvement in performance and reduction in wire-limited chip area can be achieved, without using any other circuit or design innovations [31-34].

The resulting increase of power density and its effect on die temperature as ICs migrate to 3-D is addressed in Chapter 4. An analytical thermal model that incorporates the effect of heat conduction by vias is introduced and used to estimate the temperatures of the different active layers. It is demonstrated that advancements in heat sinking technology will be necessary in order to extract maximum performance from these chips. It is pointed out that thermal management solutions will be necessary not only for 3-D ICs but also for existing 2-D technology due to the generally increasing power dissipation.

Implications of 3-D IC architecture on several circuit designs and CAD methodologies and tools are discussed in Chapter 5 with special attention to SoC design strategies. Chapter 6 discusses challenges facing 3-D integration in general such as the thermal management issues, reliability and effect on yield. Some of the promising technologies for manufacturing 3-D ICs are discussed in Chapter 7 and conclusions are finally presented in Chapter 8.

Chapter 2

3-D INTEGRATION: MOTIVATION AND BACKGROUND

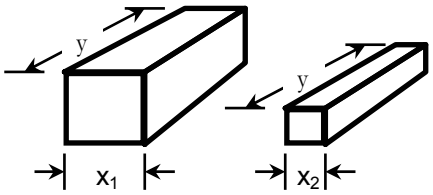
2.1 Interconnect Limited IC Performance

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, etching etc., and reduction in defect density [35]. 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 [35]. This trend can be clearly seen in Table I which shows the evolution of the total number of integrated transistors, chip area and number of metal layers for some commercial processors from Intel®. Smaller feature sizes have dramatically improved device performance [36-38]. The impact of this miniaturization on the performance of interconnect wires, however, has been less positive [1,39-41].

Year	1972	1982	1993	2000
Processor	8008	286	Pentium®	Pentium 4®
Technology Node	10 μm	1.5 μm	0.8 μm	0.18 μm
Frequency (MHz)	0.2	6	60	1400
Transistors	3500	120,000	3,100,000	42,000,000
Chip Area (mm^2)	15.2	68.7	217	224
Metal Layers	1	2	3	6

Table I: Evolution of characteristics for Intel® commercial processors

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 a simple illustration, consider a wire of length *y* whose cross-sectional dimensions are scaled down by a factor of 2 while the length is maintained constant as shown in Figure 1. The resistance, as a result, increases 4 fold while the capacitance remains constant. The total RC delay along such a wire, therefore, increases 4 times. This RC delay increase is exacerbated in fabricated ICs considering that the lengths are actually increasing.



$$x_2 = x_1/2$$
$$R_2 = 4R_1$$
$$C_2 = C_1$$
$$t_{w2} = R_2C_2y^2 = 4t_{w1}$$

Figure 1: RC delay along a length of wire increases quadratically to the scaling parameter.

As interconnect scaling continues, RC delay is increasingly becoming the dominant factor determining the performance of advanced ICs [1,3,39-41]. Figure 2 illustrates this problem by plotting a typical gate and the interconnect (RC) delays as functions of various technology nodes based on the International Technology Roadmap for Semiconductors, 1999, (ITRS '99) [3]. Throughout this analysis the interconnect RC delay is calculated for an optimally buffered line whose length equals the chip edge $\sqrt{A}$, where A is the chip area. This delay is considered a measure of IC performance and is used for comparison purposes. Chip size data is obtained from high-performance microprocessor projections from the ITRS which are summarized in Table II. The methodology used for the delay calculations is described below.

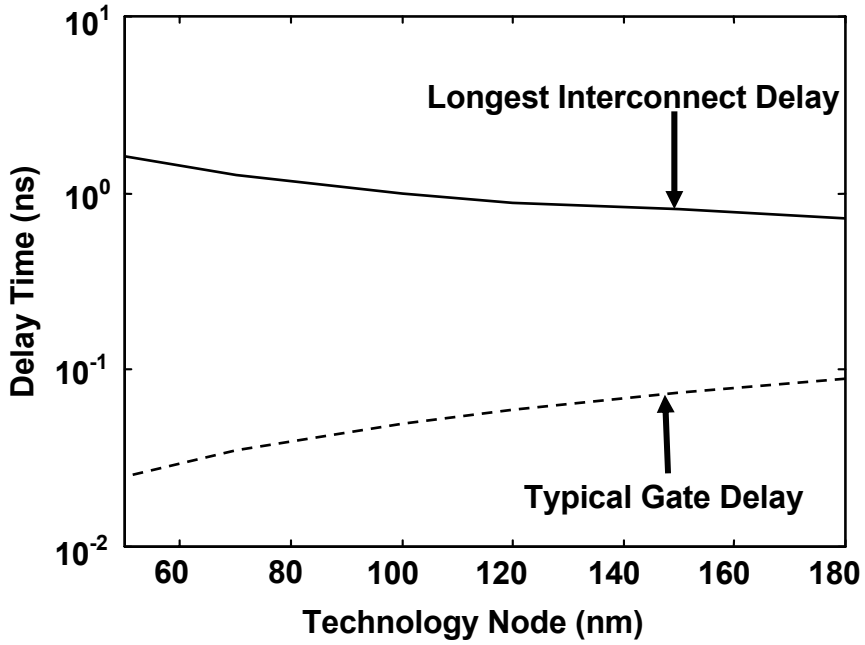


Figure 2: Typical gate delay and RC delay along longest interconnects for ITRS projections

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} ($\mu\Omega$-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_l ($pFcm^{-1}$)	2.633	2.867	2.393	2.301	2.557	2.643
R_l (Ωcm^{-1})	303.49	421.01	585.66	853.57	1571.33	3051.35
T_{FO4} (ps)	90	75	60	50	35	25
Delay (ns)	0.72	0.81	0.88	0.99	1.27	1.62

Table II: Optimal interconnect and inverter (FO4) delays at various technology nodes. Parameters necessary for delay calculations are also shown.

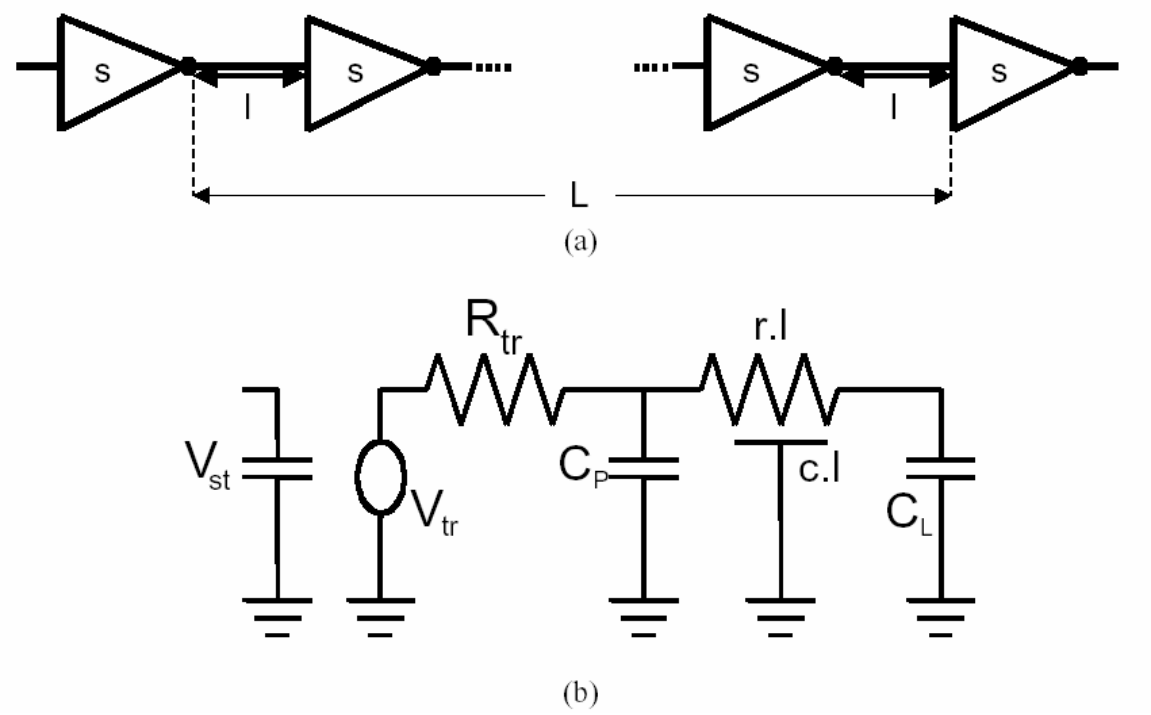


Figure 3: a) An optimally repeatered interconnect of length L. Here each repeater has a fanout of one (FO1). l is the optimal interconnect length between any two repeaters and s represents the optimal repeater size in multiples of the minimum sized inverters for a given technology b) the equivalent RC circuit.

2.1.1 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 Figure 3(a). Figure 3(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, *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, $0 = x = 1$ of the total swing. Hence the overall delay of one segment, τ_0, is given by:

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

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,

$$\tau_d = n\tau = \frac{L}{l}b(x)r_0(c_0 + c_p) + b(x)\left(c\frac{r_0}{s} + src_0\right)L + a(x)rclL \tag{2}$$

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:

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

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

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

Next we substitute (3) and (4) in (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 $c_p = 3c_{NMOS}$, where c_{NMOS} is the total source/drain junction capacitance

of a minimum sized NMOS, and 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,

$$l_{opt} = 3.24\sqrt{\frac{r_0 c_{NMOS}}{rc}} \qquad and \qquad s_{opt} = 0.577\sqrt{\frac{r_0 c}{r c_{NMOS}}}$$

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

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

where $t_{FO1} = 6r_0 c_{NMOS}$, and it represents the delay associated with an inverter that has a fanout 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 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 [42].

In terms of FO4, (5) can be approximately written as,

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

where $t_{FO4} = 15r_0 c_{NMOS}$, which can be estimated from:

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

where L_{gate} is the transistor channel length in microns and t_{FO4} is in picoseconds [42].

2.1.2 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_t, 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 r can then be expressed as:

$$A = A.R.\frac{p_w^2}{4}$$

$$r = 4\frac{\rho}{A.R.p_w^2} \tag{8}$$

2.1.3 Capacitance Calculations

The cross-section of the interconnect structure used for capacitance calculation is represented in Figure 4. 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:

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

where $C_{IMD} = \varepsilon_{IMD} LA.R.$ and $C_{ILD} = \varepsilon_{ILD}\frac{L}{2A.R.}$. 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 $\varepsilon_{IMD} = \varepsilon_{ILD} = \varepsilon_r$ the capacitance per unit length, c in (6) can be expressed as,

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

From Figure 2 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.

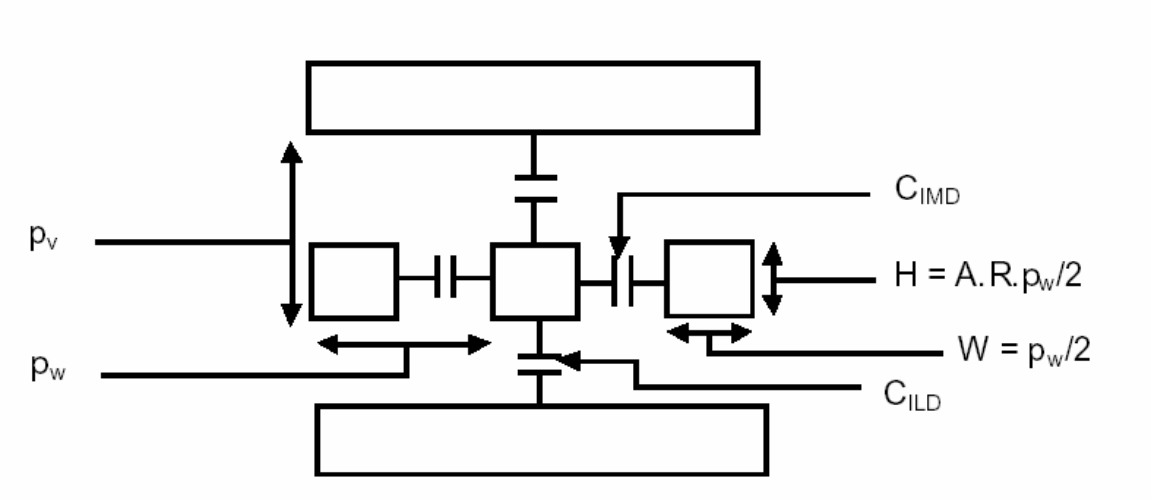


Figure 4: Cross-section of a multilevel interconnect structure showing inter-level (ILD) and intra-metal (IMD) capacitances. The aspect ratio (A.R.) is defined as (H/W) and the horizontal pitch, p_w, is defined as the sum of line width and lateral spacing between adjacent lines. The vertical pitch, p_v, is defined as the sum of line thickness and vertical spacing between lines on adjacent levels.

2.2 Hierarchical Wiring

Many solutions have been implemented and are being pursued to alleviate the adverse effects of increasing interconnect delay. One such solution is a hierarchical interconnect system architecture [5]. Here, metal lines are deposited in a multi-level vertical structure with vias connecting each metal layer to sub-levels. Reverse scaling is also applied such that the cross sectional dimensions of wires and their aspect ratios increase for higher level wires. A schematic depicting such architecture is shown in Figure 5. In this schematic the interconnect system is divided into 3 tiers: local, semi-global and global. The shortest of wires, responsible for nearest neighbor connectivity, are typically routed to the local tier where the smallest cross-sectional dimensions are afforded. Intermediate length wires, responsible for medium range inter-block communications are routed to the semi-global tier where they enjoy a larger wiring pitch to minimize RC delay. Finally, the longest wires, responsible for long distance, across-chip communications are routed to the global tier where the cross-sectional dimensions are the largest to facilitate minimum signal delay.

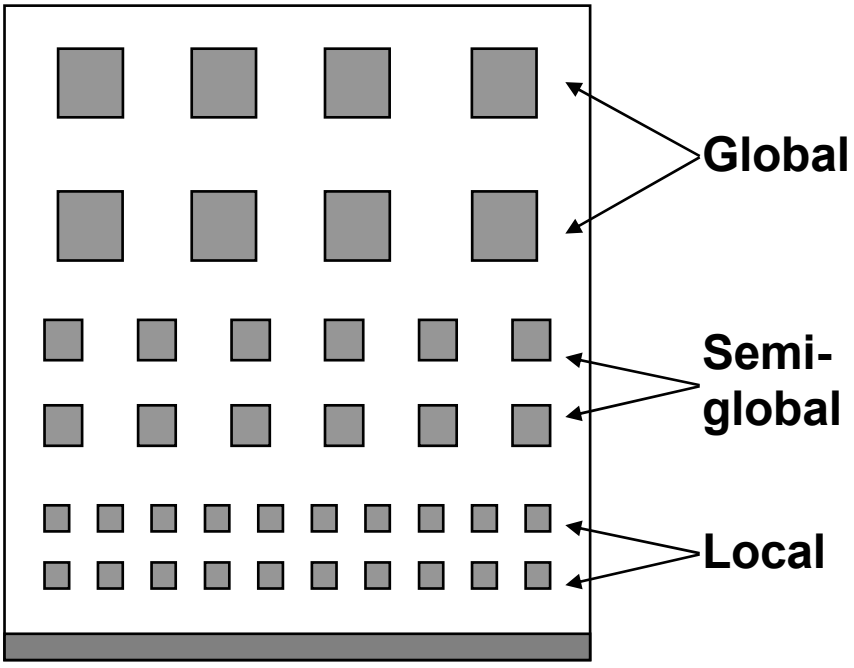


Figure 5: Schematic of a three-tier interconnection structure.

ITRS projections indicate to 10 or more layers of metal by the 50nm technology node. Increasing the number of metal layers plays a major role in preventing an explosive growth in chip area and, hence, interconnect delays. However, eventually the number of metal layers becomes highly unmanageable [5].

Reverse scaling also has its limitations. Reverse scaling refers to the increase in both the cross-sectional dimensions and in the aspect ratio of a wire. Increasing the lateral dimensions of the wires, while reduces wire resistance and delay, also increases the chip area and wire lengths which act to increase delay. Increasing aspect ratio, on the other hand, allows the manipulation of the vertical wire dimensions to reduce resistance. Conversely, however, any increase in the height of the wire, *H*, with respect to its width, *W*, also increases the wire-to-wire capacitance, C_{IMD}, as depicted in Figure 4. The effect of increasing the aspect ratio (AR) on the total RC delay of a particular wire is shown in Figure 6 where the normalized RC delay along a length of wire is plotted as a function of increasing AR. The calculation was performed based on the resistance and capacitance analysis presented above.

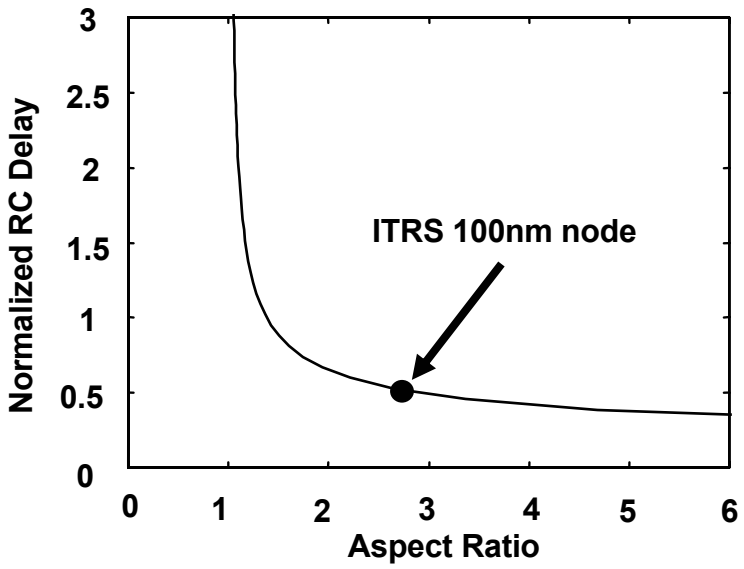


Figure 6: Effect of increasing AR on RC delay. Improvements saturate due to increasing C_{IMD}.

While hierarchical wiring has been instrumental in minimizing the limiting effect of interconnect delays it is clear that this solution has significant limitations. Other solutions

incorporate the benefits of lower resistivity metals, such as Cu, to replace Al and low-*k* dielectrics for ILD and IMD materials. Together, these solutions act to reduce the total interconnect RC delay. However, they are not without their limitations as described in the following section.

2.3 Limitations of Cu and low-*k* Technology

At 250 nm technology node, Cu with low-k dielectric was introduced to alleviate the adverse effect of increasing interconnect delay [6-10]. However, as shown in Fig. 2, below 100 nm technology node, substantial interconnect delays will result in spite of introducing these new materials, which in turn will severely limit the chip performance [3]. Further appreciable reduction in interconnect delay cannot be 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, as previously mentioned, 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 '99 [3], new physical and technological effects start dominating interconnect properties. It is imperative that these effects are accurately modeled, and incorporated in the wire performance and reliability analyses. Such modeling has been performed by P. Kapur at Stanford University [11] and the

following provides a summary of the impact of these new effects caused by scaling on the resistivity of Cu interconnects.

Before proceeding with the discussion, it is important to understand the fundamental differences between the metallization processes for Aluminum (Al) and Cu, as illustrated in Figure 7. For Al based interconnects [43], 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 deposition and a very thin layer of TiN (capping layer), that is used as the anti reflection 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 case of Cu, pattern generation in blanket films by dry-etching processes is difficult because of the lack of volatile byproducts of Cu etching [44]. Hence Cu films are deposited by the damascene process [45] illustrated in Fig. 7(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 [46]. 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 [47]. 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) [48]. Finally, a layer of silicon nitride is deposited to passivate 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.

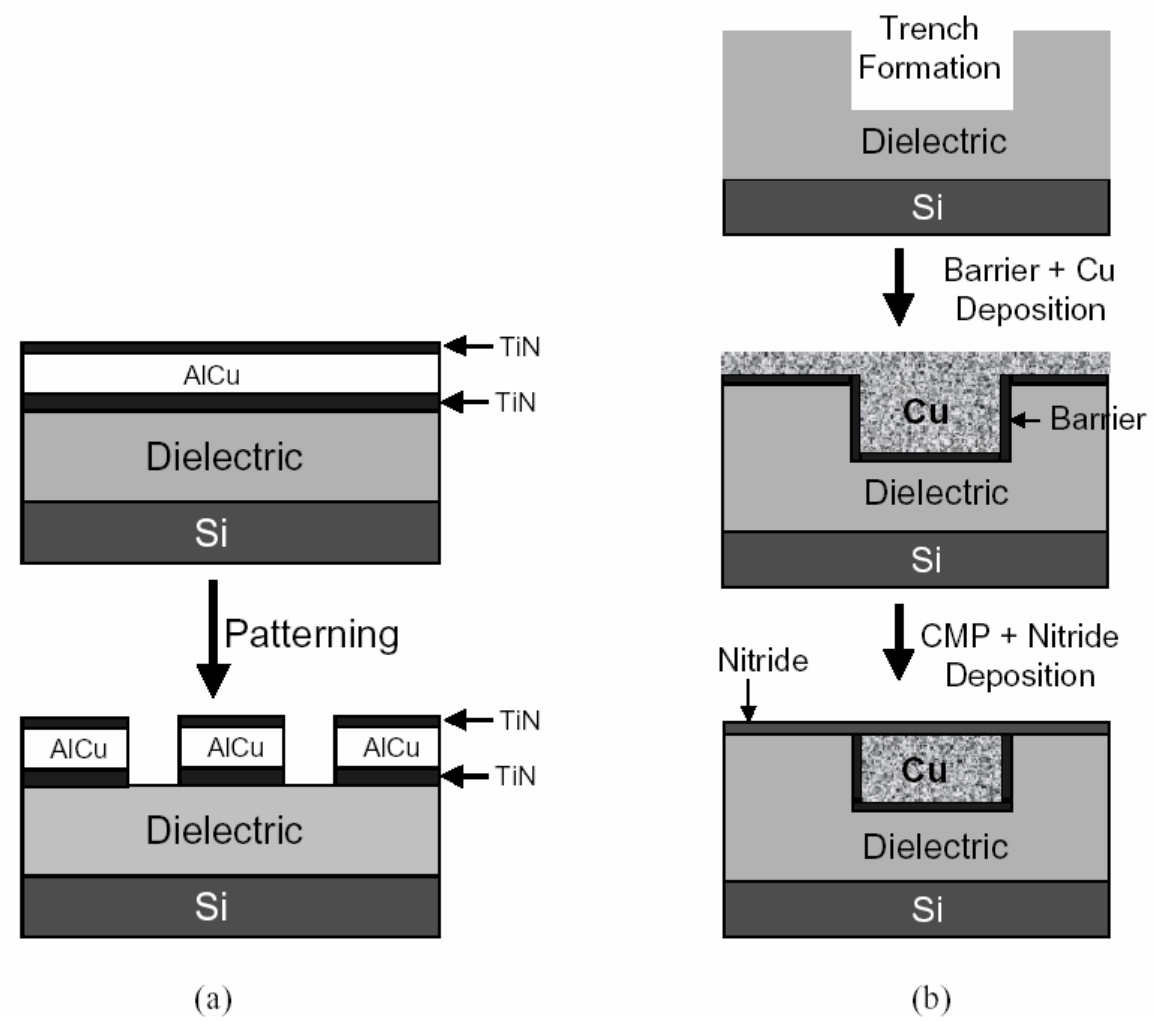


Figure 7: Illustration of a) AlCu and b) damascene Cu interconnect processes.

It is commonly believed that material resistivity for Cu would not change significantly for future interconnects [3]. However, as dimensions shrink, firstly, the electron scattering from the surface becomes comparable to electron bulk scattering mechanisms such as phonon

scattering. Secondly, since barrier thicknesses do not scale as rapidly as interconnect dimensions, a greater fraction of interconnect area are consumed by metal barrier in the future, (Fig. 8). These effects conspire to increase the effective resistivity of Cu significantly. In addition, the operational temperature of wires (~373K) is higher than the room temperature (300K) and can increase further due to self-heating caused by the flow of current [17,49]. The increase in temperature, in turn, would also increase the wire resistivity.

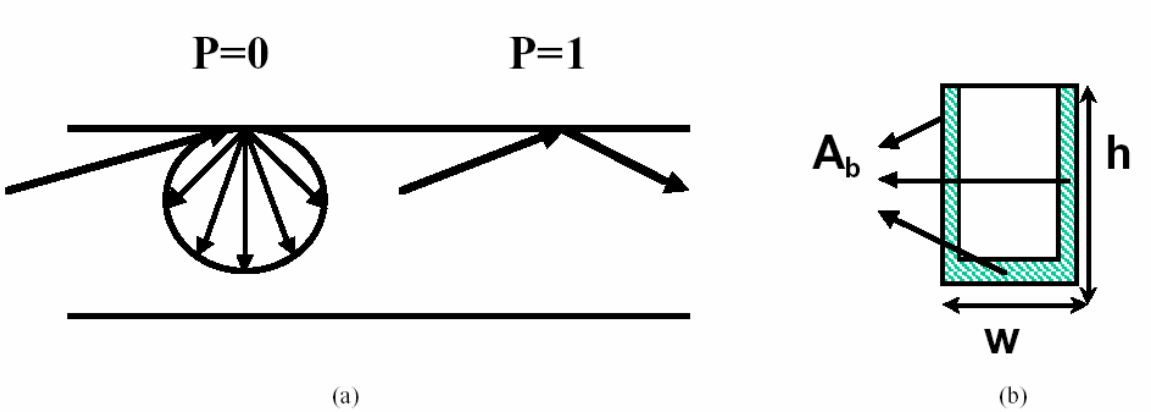


Figure 8: Illustration of a) diffuse and specular surface scattering and b) effective cross-section reduction of Cu interconnects due to barrier. *P*=0, signifies complete diffuse scattering causing maximum decrease in mobility, hence, a maximum increase in resistivity; whereas, *P*=1 indicates complete 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 [10,43].

In light of the above discussion, it becomes obvious that Cu interconnects and low-*k* dielectrics alone have limited impact in alleviating the interconnect delay problem. Other challenges facing VLSI design, such as CAD methodologies and SoC designs, are also discussed below.

2.4 Deep Submicron Interconnect Effects on VLSI Design

Interconnects in deep submicron VLSI present many challenges to the existing computer-aided-design (CAD) methodologies and tools [50]. 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 layout file (such as GDSII) is generated using a process called *physical synthesis*, which generates the detailed floorplanning, placement and routing.

For deep submicron 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 [51]. The wire-load model includes the intrinsic gate delay and an average delay due to the interconnect connecting the output of the gate to other gate inputs as well as the delay associated with the inputs of the following stage. This approach suffices if the interconnect delays (after physical synthesis) remain negligible. However, as shown in Fig. 2, for deep sub micron 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 submicron 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 coupling, make it difficult for existing CAD tools to obtain timing convergence for the design blocks within a reasonable number of iterations.

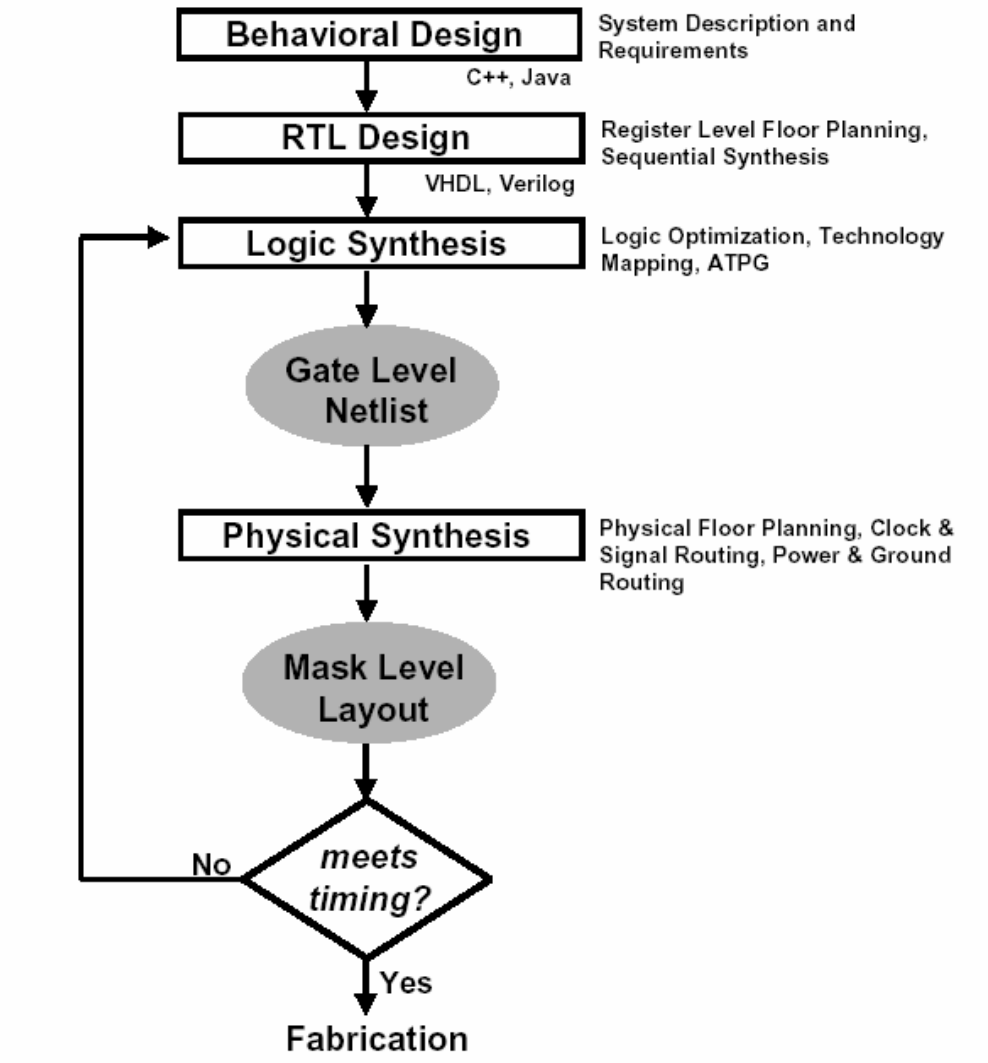


Figure 9: Typical VLSI design process flow.

It is instructive to note that the magnitude of the interconnect problem for future deep submicron ICs with greater than 10^8 gates 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 [52], is not quite adequate for deep sub micron VLSI. This is due to the fact that for deep submicron 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 inter-block 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 intra-module wire delays can become unexpectedly large contrary to usual assumptions as in [53]. In order to mitigate the interconnect scaling problems some researchers have proposed combined *wire planning* and *constant-delay synthesis* [54,55]. This methodology is also based on a block-based design where the inter-block wires are planned or constructed and the remaining wires are handled through the constant-delay synthesis [56] 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 [57]. 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 inter-symbol interference. Also, electromigration and thermal effects in interconnects impose severe restrictions on signal, bus, and power/ground line scaling [15,17].

Thus it can be concluded that the interconnect problem in deep submicron 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 [2]. We need to think beyond the current paradigm of design architecture.

2.5 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 [50,58]. 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 Figure 10. They can also involve more esoteric technologies like Micro-Electromechanical Systems (MEMS), bio-electronics, micro-fluidics, and optical input/output. 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 multi-media communications including high-speed internet applications, medical applications including remote surgery, automated drug delivery, and non-invasive 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.

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.

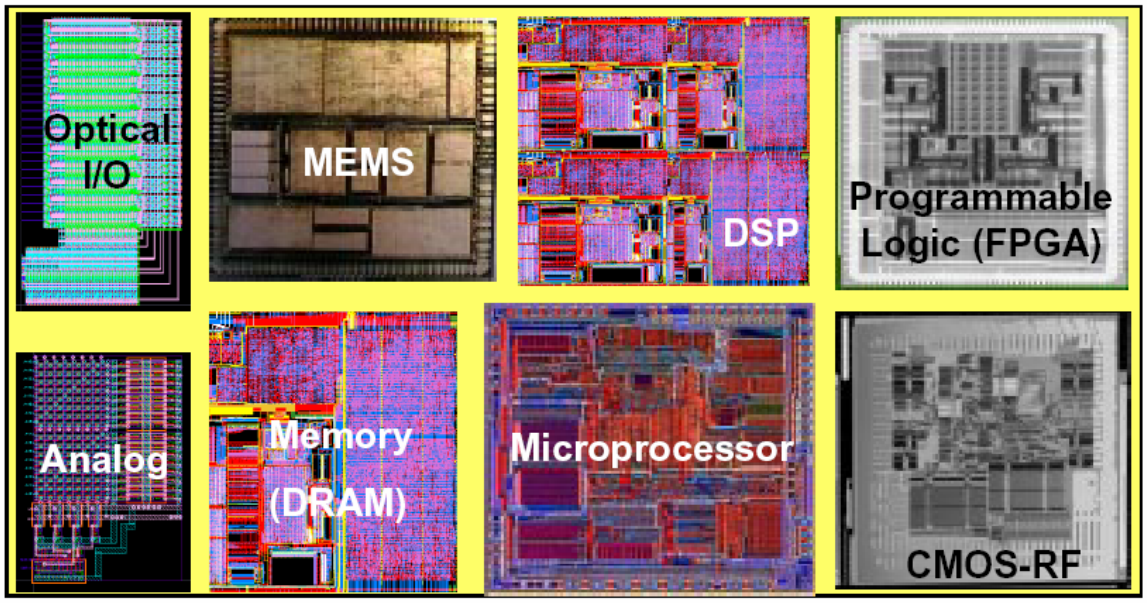


Figure 10: Schematic of a *System-on-a-Chip* design using a planar (2-D) IC.

2.6 3-D Integration

3-D integration (schematically illustrated in Figure 11) to create multilayer Si ICs is a concept that can significantly improve deep submicron interconnect performance, increase transistor packing density, and reduce chip area and power dissipation [51]. Additionally, 3-D ICs can be very effective vehicles for large-scale on-chip integration of heterogeneous systems.

3-D integration of ICs is not a new concept. Many researchers have investigated different technologies for 3-D IC fabrication [20-28]. There has also been some work on modeling interconnect performance and demonstrating the potential benefits for 3-D ICs [18,19]. The focus of previous work, however, has been mainly on device-size limited ICs. In other words, 3-D analysis was performed on circuits where either the number of integrated transistors is small or the complexity of wiring is minimal to warrant a simple interconnect system that does not play a dominant role in determining overall IC size or performance. Examples of such ICs may include low-performance microprocessors, ASICs or memory chips.

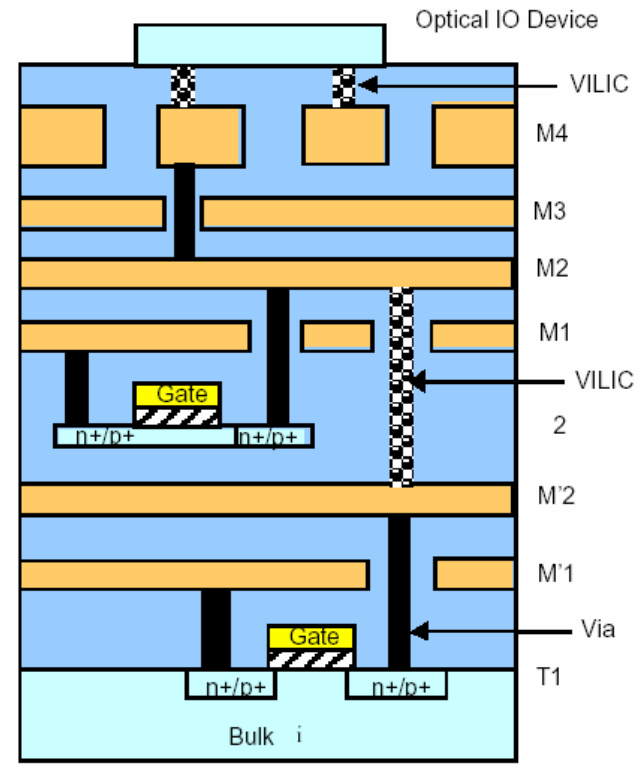


Figure 11: Schematic representation of 3-D integration with multilevel wiring network and VILICs. T1: first active layer device, T2: second active layer device, Optical I/O device: third active layer I/O device. M'1 and M'2 are for T1, M1 and M2 are for T2. M3 and M4 are shared by T1, T2, and the I/O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

For present and future high-end ICs, however, where the interconnect network is highly sophisticated occupying 6 layers of metal and projected to increase to 10 or more in a few years, such analysis cannot hold. There exists a need to develop a system's level understanding of the impact of 3-D integration on such wire-pitch limited ICs. Considering that the chip size is determined by the amount of wiring required, it becomes less than obvious that 3-D would help unless the wiring requirement that contributes to chip size is reduced.

With regards to device-size limited ICs, however, there has been much commercial activity in the area of 3-D integration. Matrix Semiconductor, Inc., located in Silicon Valley, CA, for instance, is successfully commercializing a 3-D memory chip (Figure 12) and has also successfully fabricated simple logic circuits in 3-D [59].

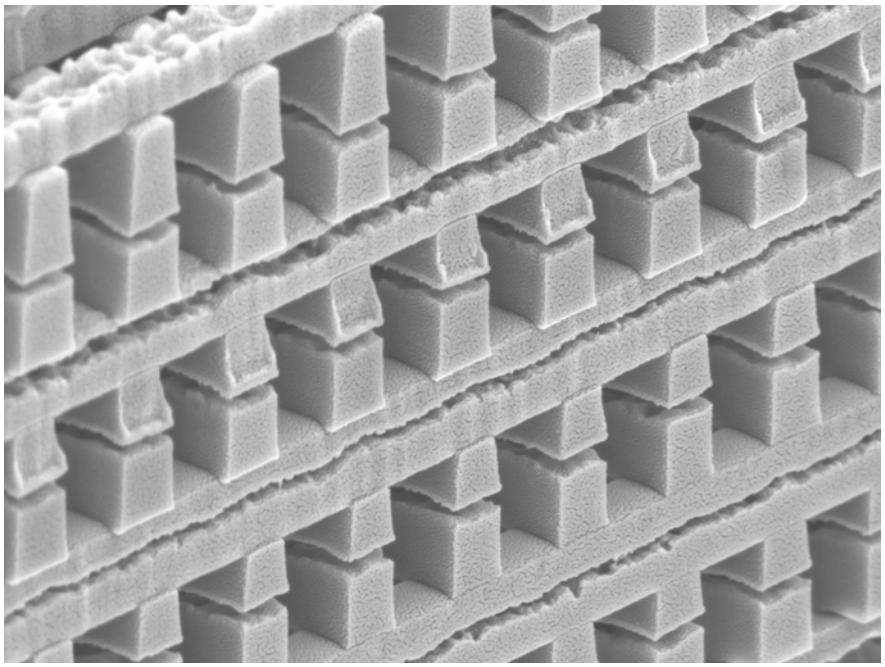


Figure 12: Vertical stack of memory cells can store eight bits of information in the area usually allotted to just one bit. *(Courtesy of Thomas H. Lee, Stanford University).*

Such feats of accomplishments promise to significantly increase device density and lower costs. While such examples indicate to the potential benefits of 3-D integration, much work remains necessary to achieve similar benefits for highly complex ICs such as high-performance microprocessors.

In response to the demand for a system's level interconnect network modeling for 3-D ICs, a 3-D design architecture for wire-pitch limited ICs is proposed and analyzed in this work [31-34]. Briefly the concept considers dividing an entire (2-D) chip into a number of logic and memory blocks which are arranged and allocated on separate layers of Si that are stacked on top of each other. Each Si layer in the 3-D structure may have its own dedicated or shared interconnect network. Each of these layers are connected together through vertical inter-layer interconnects (VILICs) and common global interconnects as shown in Fig. 11. The 3-D architecture offers extra flexibility in system design, block placement and routing. For instance, blocks on a critical path can be placed as nearest vertical neighbors 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 submicron interconnects on VLSI design discussed earlier can be reduced significantly by replacing certain long *global wires* that realize the inter-block communications with short VILICs due to vertical placement of logic blocks.

Furthermore, the 3-D chip design technology offers the capability to build SoCs by placing heterogeneous circuits, such as different voltage ICs 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. 3-D integration can reduce the wiring, thereby reducing the capacitance, power dissipation, and chip area and

improve chip performance. It would also lower the I/O pin count, and therefore be an economically attractive option for building high-performance SoCs. 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 Figure 11). 3-D integration of optical and CMOS circuitry have been demonstrated in the past [52]. A schematic diagram of a 3-D SoC is shown in Figure 13 with logic circuitry, distributed memory (SRAM) blocks (to reduce access time and enhance system performance), high-density DRAMs, analog/RF, and optical I/Os on different active layers.

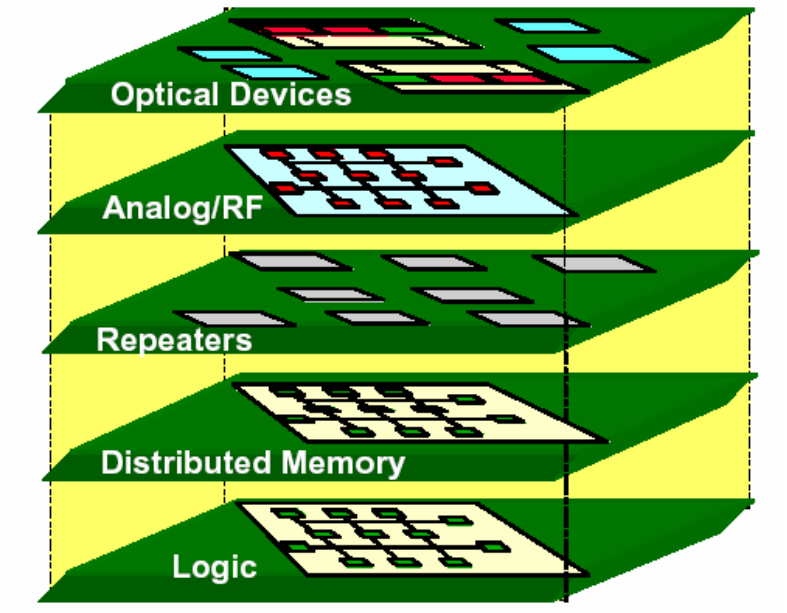


Figure 13: Schematic of a 3-D chip showing integrated heterogeneous technologies.

Chapter 3

PERFORMANCE MODELING AND ANALYSIS

As mentioned in the previous chapter, much work has been done in the area of 3-D integration. Yet most this effort has been focused either on the technology of fabricating 3-D ICs or on simple modeling of performance based on the assumption of device-size limited ICs [18-30]. In pursuing the effort of filling the gap, an interconnect systems' level quantitative analysis of wire-pitch limited 3-D ICs is presented in detail in this chapter [31-34].

3.1 Scope

A 3-D solution at first glance seems an obvious answer to the interconnect delay problem. Since chip size directly affects the interconnect delay, creating a second active layer can reduce the total chip footprint, 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 and the following chapters, the possible effects of 3-D integration of large logic circuits on key metrics such as chip area, power dissipation and performance is 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 [60,61] has been modified for 3-D ICs to quantify effects on interconnect delay. Unlike previous work [18,19], wire-pitch limited chips are considered.

The results obtained in the sections below indicate that when critically long metal lines that occupy lateral space are replaced with effective 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.

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 as discussed in later chapters. For simplicity, technology effects on metal wire resistivity as discussed in the previous chapter are ignored in the following analysis (for both 2-D and 3-D ICs), where bulk resistivity is assumed.

3.2 Concept

The basic concept for this 3-D analysis is illustrated in Figure 1. A general representation of a wire-pitch limited 2-D IC is considered as consisting of a number of logic blocks. By migrating to a 3-D structure it is assumed that logic blocks can be rearranged in

some fashion so as to occupy any number of active layers of Si. Such an arrangement makes the vertical dimension available for logic block interconnectivity. For instance, a global wire in the 2-D IC connecting 2 logic blocks across chip and contributing to the chip size can now be replaced with a Vertical Inter-Layer Interconnect (VILIC) connecting the same 2 blocks which can be arranged vertically stacked on top of each other. This VILIC is characterized by its much shorter length and smaller contribution to chip size as compared to the original global wire. By performing such replacements across the entire interconnect network, a significant fraction of lateral wires can thus be replaced with VILICs which ultimately reduces the horizontal wiring requirement and chip size and prevents the interconnect delay problem from dominating IC performance.

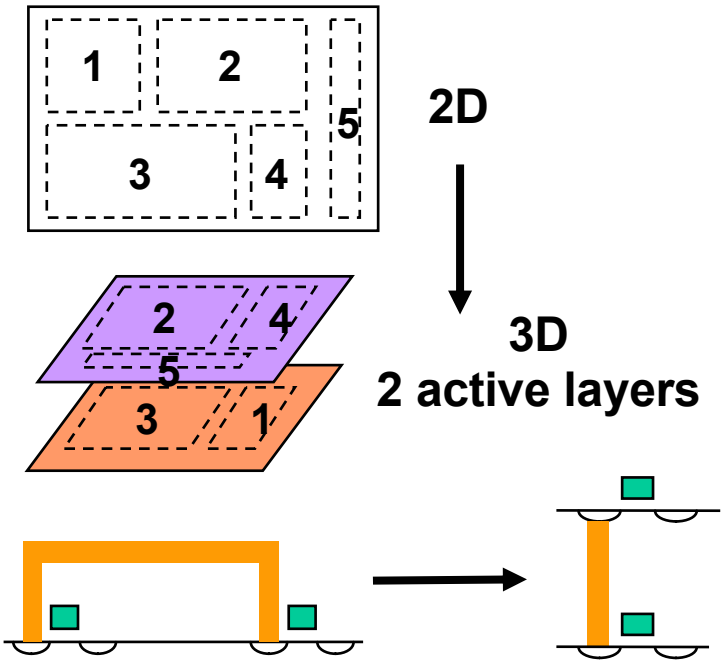


Figure 1: Horizontal interconnects are replaced with VILICs, reducing wiring requirement, chip area and interconnect delays.

This concept is reduced to a quantitative analysis in the following sections by considering a stochastic wire-length distribution model to estimate the reduction in the wiring requirement, chip area and improvements in interconnect delays.

3.3 Rent's Rule

Before proceeding with the discussion, it is important to understand the major role that Rent's Rule plays in the analysis. Rent's Rule is an empirical relationship that was formulated in the early 1970s at IBM® that relates the total number of I/O pins, *T*, to the total number of gates, *N*, in a random logic arrangement [63], and takes the following form:

$$T = k\,N^{p} \tag{1}$$

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 and architecture of ICs. Furthermore, Rent's Rule exhibits a recursive property such that it holds for sub-systems within a logic network, a property that will be extensively used to estimate the total number of connections within an IC. Figure 2 is a plot from the original paper on Rent's Rule [63] showing the validity of the relationship.

3.4 Wire Length Distribution

The chip area and performance of an IC can be estimated from its wire length distribution. To quantify this distribution, a logic system is 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 analysis and considered valid for high-performance ICs.

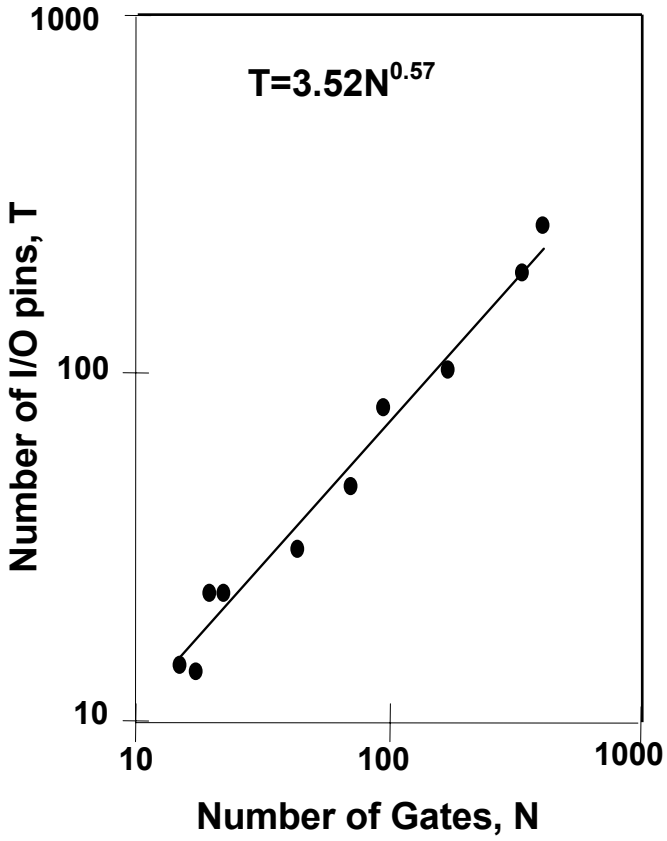


Figure 2: I/O pins vs. number of gates from the original paper [63] showing the validity of Rent's Rule.

The wiring network is assumed to be a distribution of connecting wires ranging from the very short (to connect closest neighbor logic gates, or intra-block connections), to the very long (for long distance across-chip, or inter-block 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 details of the following discussion can be found in [60,61].

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,

$$I(l) = \int_{1}^{l} i(x)\,dx \tag{2}$$

where x is a variable of integration representing length and l is the length of the interconnect in gate pitches. 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 [62]. 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 logic area of the chip.

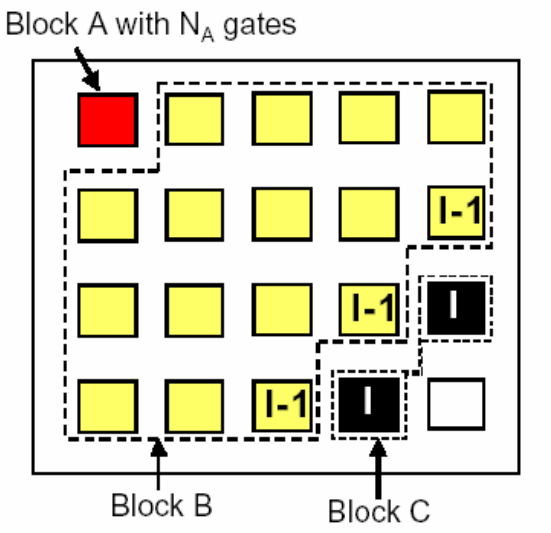


Figure 3: Schematic view of logic blocks used for determining wire length distribution (adopted from [60]).

The stochastic approach used for estimating the wire-length distribution of a 2-D chip is first reviewed and then modified 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. 3. 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. 3 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 inter-block 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. 3 gives,

$$T_A + T_B + T_C = T_{A-to-C} + T_{A-to-B} + T_{B-to-C} + T_{ABC} \tag{3}$$

where T_A, T_B, and T_C are the number of I/Os for blocks A, B, and C respectively. $T_{A\text{-}to\text{-}C}$, $T_{A\text{-}to\text{-}B}$, and $T_{B\text{-}to\text{-}C}$ are the numbers of I/Os from block A to C, block A to B, and from block B to C respectively. T_{ABC} represents the number of I/Os for the entire system comprising of 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,

$$T_{A-to-B} = T_A + T_B - T_{AB} \tag{4}$$

$$T_{B-to-C} = T_B + T_C - T_{BC} \tag{5}$$

Substituting (5) and (4) in (3) gives,

$$T_{A-to-C} = T_{AB} + T_{BC} - T_B - T_{ABC} \tag{6}$$

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,

$$T_B = k\left(N_B\right)^p \tag{7}$$

$$T_{AB} = k\left(N_A + N_B\right)^p \tag{8}$$

$$T_{BC} = k\left(N_B + N_C\right)^p \tag{9}$$

$$T_{ABC} = k\left(N_A + N_B + N_C\right)^p \tag{10}$$

where $N = N_A + N_B + N_C$. Substituting (7)-(10) in (6) gives,

$$T_{A-to-C} = k\left[\left(N_A + N_B\right)^p - \left(N_B\right)^p + \left(N_B + N_C\right)^p - \left(N_A + N_B + N_C\right)^p\right] \tag{11}$$

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

$$I_{A-to-C} = \alpha\, k\left(T_{A-to-C}\right) \tag{12}$$

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

$$\alpha = \frac{f.o}{1 + f.o} \tag{13}$$

Equation (12) can be used to calculate the number of interconnects for each length l in Fig. 3 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.

J. Davis *et al.* developed a closed form analytical expression of the wire-length distribution for a 2-D IC [60], which can be expressed as,

$$I(l) = I_{total}\, P(l) \tag{14}$$

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

$$I_{total} = \alpha k N \left(1 - N^{p-1}\right) \tag{15}$$

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,

$$P(l) = \frac{1}{2N\left(1 - N^{p-1}\right)} \Gamma \left(\frac{l^{2p} - 1}{6p} + 2\sqrt{N}\, \frac{-l^{2p-1} + 1}{(2p-1)} - N\, \frac{-l^{2p-2} + 1}{(p-1)} \right) \tag{16}$$

for $1 \le l \le \sqrt{N}$, and

$$P(l) = \frac{1}{2N\left(1 - N^{p-1}\right)} \Gamma \left(\begin{array}{l} \dfrac{N^{2p-1}}{6p} + 2\sqrt{N}\,\dfrac{-N^{2p-1}+1}{2p-1} - N\,\dfrac{-N^{2p-2}+1}{p-1} \\ + \dfrac{1}{3}\left(\begin{array}{l} -8N^{3/2}\,\dfrac{-l^{2p-3}+N^{\,p-(3/2)}}{2p-3} + 6N\,\dfrac{-l^{2p-2}+N^{p-1}}{p-1} \\ -6\sqrt{N}\,\dfrac{-l^{2p-1}+N^{p-(1/2)}}{2p-1} + \dfrac{-l^{2p}+N^{p}}{2p} \end{array} \right) \end{array} \right) \tag{17}$$

for $\sqrt{N} \le l \le 2\sqrt{N}$. The factor Γ is defined by,

$$\Gamma = \frac{2N\left(1 - N^{p-1}\right)}{\left(-N^{p}\,\dfrac{1 + 2p - 2^{2p-1}}{p(2p-1)(p-1)(2p-3)} - \dfrac{1}{6p} + \dfrac{2\sqrt{N}}{2p-1} - \dfrac{N}{p-1} \right)} \tag{18}$$

Substituting (15) – (18) in (14) gives the closed form expressions for the total wire length distribution as follows,

$$I(l) = \frac{\alpha k}{2}\,\Gamma \left(\frac{l^{2p}-1}{6p} + 2\sqrt{N}\,\frac{-l^{2p-1}+1}{(2p-1)} - N\,\frac{-l^{2p-2}+1}{(p-1)} \right) \tag{19}$$

for $1 \le l \le \sqrt{N}$, and

$$I(l) = \frac{\alpha k}{2}\,\Gamma \left(\begin{array}{l} \dfrac{N^{2p-1}}{6p} + 2\sqrt{N}\,\dfrac{-N^{2p-1}+1}{2p-1} - N\,\dfrac{-N^{2p-2}+1}{p-1} \\ + \dfrac{1}{3}\left(\begin{array}{l} -8N^{3/2}\,\dfrac{-l^{2p-3}+N^{\,p-(3/2)}}{2p-3} + 6N\,\dfrac{-l^{2p-2}+N^{p-1}}{p-1} \\ -6\sqrt{N}\,\dfrac{-l^{2p-1}+N^{p-(1/2)}}{2p-1} + \dfrac{-l^{2p}+N^{p}}{2p} \end{array} \right) \end{array} \right) \tag{20}$$

The simple use of Rent's Rule above applies to 2-D IC's and requires adaptation for a valid application to 3-D IC's. 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, an analysis is provided where n silicon layers are available. The application to the two-layer case (n=2) is straightforward. An N gate IC design is divided into N/n gate blocks. To maintain generality the blocks are assumed randomly re-distributed among the different layers. Other non-random boundary conditions governing logic redistribution would be IC design specific. Figure 4 illustrates the analysis for 2 layers. It is assumed that the routing algorithm and overall logic style is the same for all layers. This ensures that Rent's constant, k, and Rent's exponent, p, are the same for all layers. Applying Rent's rule to all layers, we have,

$$T = kN^p = \left(\sum_{i=1}^{n} T_i\right) - T_{\text{int}} = nk\left(\frac{N}{n}\right)^p - T_{\text{int}} \tag{21}$$

Here T is the number of I/Os entire design, T_i represents the number of I/Os for each layer and T_{int} represents the total number of I/O ports dedicated to interconnectivity of the n layers. p is Rent's exponent and k is the average number of I/Os per gate. Hence, it follows that,

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

$$T_{ext,i} = T_i - \frac{T_{\text{int}}}{n} = k\, n^{p-1}\left(\frac{N}{n}\right)^p \tag{22}$$

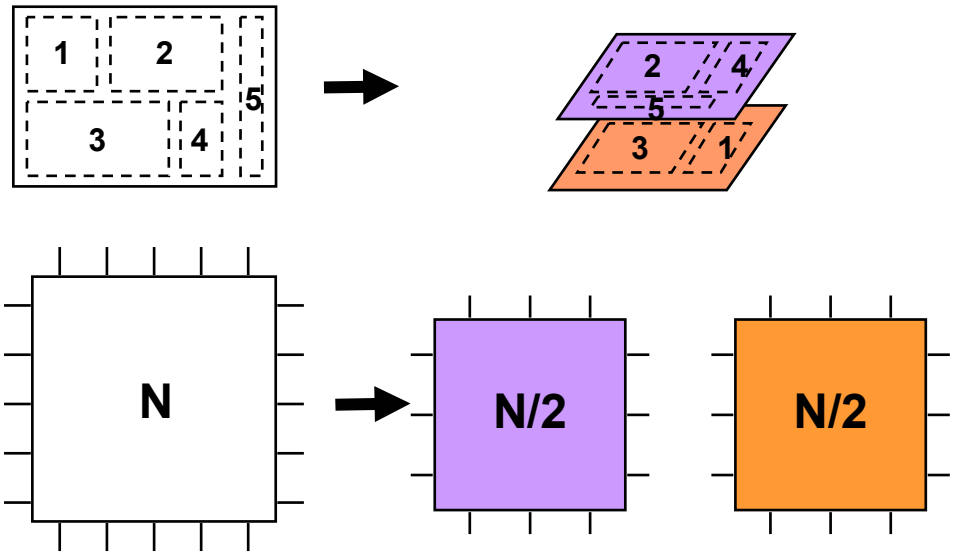


Figure 4: Schematic illustrating the migration process from 2-D to 3-D in Rent's Rule terms. Rent's Rule applies recursively to each 3-D layer as each is considered a sub-system. Conservation of I/Os is assumed.

Here $T_{ext,i}$ is the average number of *external* I/Os per layer, *i*. Comparing (22) this with Rent's equation for each layer, i.e., $T = k\left(\frac{N}{n}\right)^{p}$, then for each layer,

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

$$k_{eff,ext} = k\, n^{p-1} \tag{23}$$

where $k_{eff,int}$ is the effective number of I/Os per gate used for connecting other gates on the same layer and $k_{eff,ext}$ is the effective number of I/Os per gate used to connect to gates on other active layers. Extending this analysis to a 2-layer 3-D IC (n=2) as in Figure 4, then,

$$T = kN^p = T_1 + T_2 - T_{\text{int}} = 2k\left(\frac{N}{2}\right)^p - T_{\text{int}} \tag{24}$$

Since each layerhas (T_{int}/2) dedicated I/O ports for connection to the other layer, then,

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

$$k_{eff,\text{int}} = k\left(1 - 2^{p-1}\right) \tag{25}$$

Thus, Rent's Rule has been modified to apply to 3-D systems and these modifications can be incorporated into the stochastic model described above to quantify the wire length distribution of a 3-D IC. Specifically, $k_{eff,ext}$ can be used instead of k in the stochastic analysis to arrive at the horizontal wire length distribution that determines the wiring requirement and chip size. Meanwhile, $k_{eff,int}$ represents the fraction of wires that are to be replaced from the original 2-D wire length distribution with VILICs. Essentially the total wiring requirement has been resolved into 2 components in the 3-D configuration: a lateral component that determines the chip size and IC performance and a vertical component that represents the wires replaced. Since the vertical component of the wiring requirement is assumed not to contribute to the final chip size, a significant reduction in chip area, and hence improvement in performance, can be expected as will be discussed in later sections.

ITRS projections for high-performance microprocessors at the 50nm technology node are used to illustrate in detail the change in the wire length distribution as a representative 2-D configuration is migrated to 3-D. In later sections, the analysis is extended to all technology nodes projected by ITRS [3]. Table I summarizes some of these projections for the 50nm

technology node [3]. However, before proceeding with the comparison it is important to consider the effects of memory on this analysis.

Technology node	50nm
N_{Logic}	$769x10^6$
N_{Memory}	$6284x10^6$
Operating Frequency	3 GHz
Metal Levels	9
ρ_{Cu}	$1.67x10^{-6}$ Ω-cm
ε_r	1.5

Table I: Summary of ITRS 1999 projections for high-performance microprocessors at the 50nm technology node.

3.4.1 Incorporating On-Chip Memory

Present technology dedicates a physically separate portion of a microprocessor die for memory as illustrated in Figure 5. Such an arrangement is named *localized memory*. Considering the ITRS projections for the 50nm technology node, on-chip memory is projected to occupy approximately 50% of the chip area [11]. As such, it is imperative to take on-chip memory into account in this analysis. Futuristic trends in microprocessor design, however, project a distributed architecture whereby distributed memory modules are associated with individual logic blocks [2]. This arrangement is named *distributed memory*. In compliance with futuristic trends, the distributed memory arrangement is assumed during the application of the above analysis to the 50nm as well as across all technology nodes, where memory is assumed homogeneously distributed across the entire die.

Memory blocks, in general, require significantly less wiring complexity [53,62]. As such, the wiring requirement for on-chip memory in this wire length distribution analysis can be taken into account, alongside logic, by considering a lower Rent's exponent, *p*, dedicated to memory as compared to logic. A reasonable assumption for the value of *p* is in the range of 0.25 for memory [53,62].

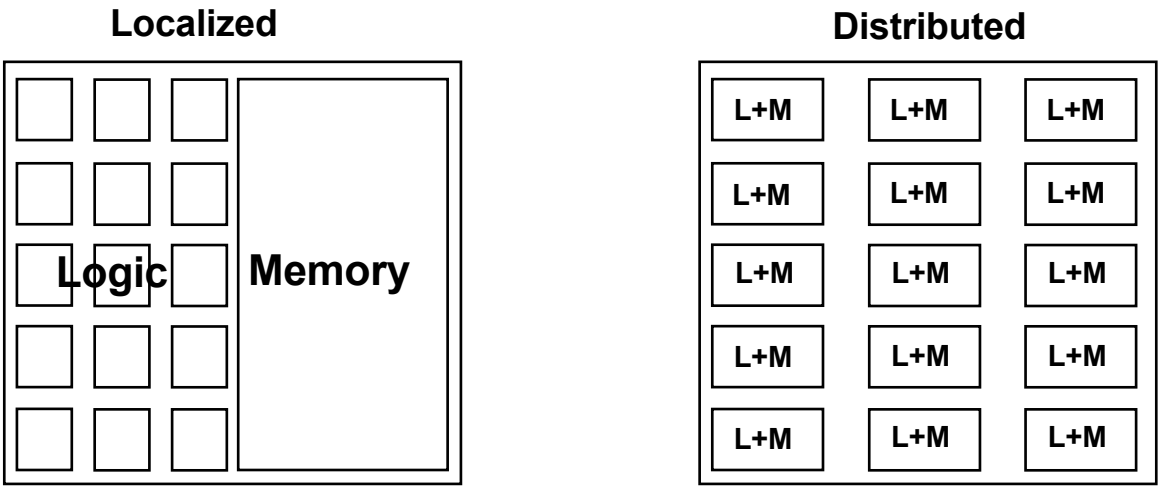


Figure 5: On-chip memory included in the analysis can be considered *localized* (current technology) or homogeneously *distributed* (future trends).

3.4.2 2-D and 3-D Wire Length Distributions

The results of the lateral wire length distribution modeling are presented below in Figure 6 for the ITRS 50nm technology node, comparing a representative 2-D IC to its 3-D counterpart with only 2 active layers of silicon. Although the analysis has been performed for all technology nodes and will be presented in later sections, the 50nm node serves as an example to illustrate this process. The modified Rent's Rule for 3-D is used in the wire length distribution analysis presented above and on-chip memory is taken into account. The values

for Rent's constant and exponent for logic, k and p, respectively, are derived from ITRS projections and vary for all technology nodes and are of the order of k=4 and p=0.65. As a technology generation is migrated to 3-D, k and p are assumed to remain invariant.

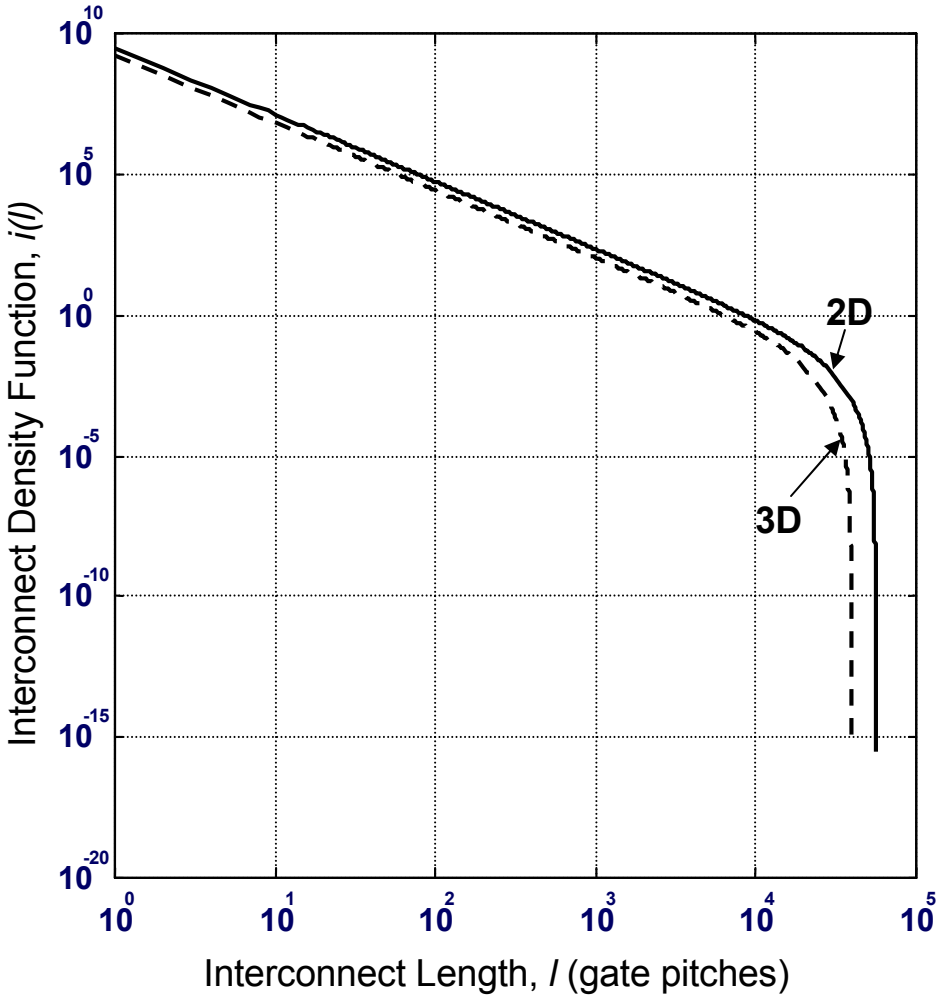


Figure 6: Lateral wire length distributions as a function of gate pitches for a 2-D and 3-D (2 active layer) arrangement for the ITRS 50nm technology node projection.

As is evident from Figure 6, migrating from 2-D to 3-D results in a significant reduction in the lateral wire length distribution. The difference between the 2 distributions

represents the population of wires that have been eliminated horizontally and replaced with VILICs which are assumed to have a negligible contribution to the chip size and performance. The distributions are plotted as a function of gate pitches which are yet to be determined. Of important note is the seeming uniform reduction in the 3-D distribution, reflecting the nature of the indiscriminate replacement of lateral wires with VILICs independently of their lengths. This is due to the previous assumption involving a random redistribution of logic and memory blocks from the 2-D to the 3-D configuration. Ideally, only the longest of wires would be replaced as they are mainly limiting interconnect delay, while the shortest of wires are left in place. However, this requires detailed knowledge of the specific IC design that would allow the intelligent redistribution of logic blocks. Figure 7 illustrates in schematic form the major difference between random and intelligent redistribution of blocks.

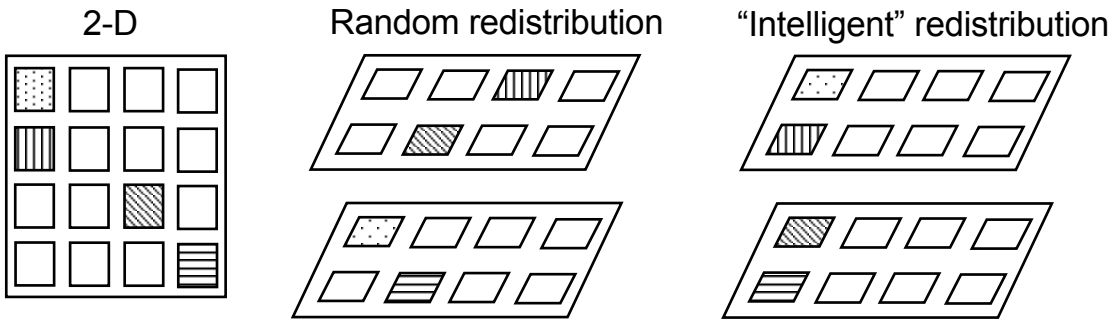


Figure 7: Random redistribution of logic and memory results in both long and short wires replaced with VILICs. A detailed knowledge of the IC design allows a more intelligent approach to redistribution such that only longer, performance limiting, wires are replaced.

To maintain generality, without limiting this analysis to any specific IC or design, the random redistribution approach has been assumed bearing in mind the implication that the

results obtained can be improved further for any specific design. As such, the results presented in this analysis are considered conservative in nature with room for improvement.

3.5 2-D and 3-D Chip Area Estimation

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 be greater than the area required by the logic gates. The chip size can be determined from the wire length distribution by proceeding with a process of interconnect tier allocation. 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 intra-block, inter-block, 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 inter-block 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 3 wiring tiers: local, semi-global and global, spanning, for example, a total of 9 to 10 metallization layers as projected by ITRS 1999 for the 50 nm technology node.

The semi-global tier is normally responsible for inter-block communications across intermediate distances. Figure 8 shows a schematic of a 3-tier interconnect structure.

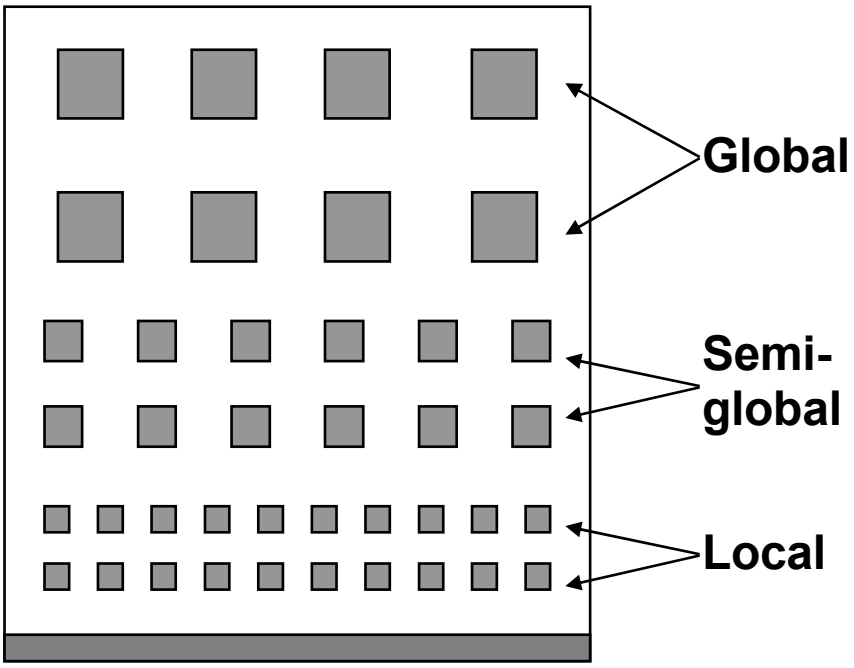


Figure 8: Schematic of a three-tier interconnection structure.

The process of allocating wires to a 3-tier interconnection structure involves determining the maximum interconnect length on any given tier by an interconnect delay criterion [61]. It is assumed $t_{delay\_max} = 0.25T$ for semi-global 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 $t_{delay} = 0.9T$ [61]. Once the wires are allocated, the semi-global tier pitch that minimizes the wire limited chip area is determined.

As the area of the chip is determined by the total wiring requirement, the total area required by the interconnect wiring, in terms of gate pitch, can be expressed as:

$$A_{required} = \sqrt{\frac{A_c}{N}} \left(p_{loc} L_{total\_loc} + p_{semi} L_{total\_semi} + p_{glob} L_{total\_glob} \right) \quad (26)$$

where A_c is the chip area, N is the number of gates, p_{loc} is the local pitch, p_{semi} is the semi-global pitch, p_{global} is the global pitch, L_{total_loc} is the total length of the local interconnects, L_{total_semi} is the total length of the semi-global interconnects and L_{total_glob} is the 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. Hence it follows that,

$$L_{total\_loc} = \chi \int_{1}^{L_{loc}} l\, i(l)\, dl \quad (27)$$

$$L_{total\_semi} = \chi \int_{L_{loc}}^{L_{semi}} l\, i(l)\, dl \quad (28)$$

$$L_{total\_glob} = \chi \int_{L_{semi}}^{2\sqrt{N}} l\, i(l)\, dl \quad (29)$$

where χ is a correction factor that converts the point-to-point interconnect length to wiring net length (using a linear net model, $\chi = 4/(f.o.+3)$). L_{loc} , L_{semi} and L_{global} represent the maximum length of wires in gate pitches for the local, semi-global and global tiers, respectively.

The maximum interconnect lengths L_{loc} and L_{semi} can now be calculated based on the delay of an optimally buffered interconnect, given by equation (2.6), and is expressed in terms of FO4 delay. By substituting (2.8) and (2.9) in (2.6), and using $\tau_d = \frac{\beta}{f_c}$, the length of the longest wire, L, and the pitch, p_n, for an arbitrary tier are related by the following expression:

$$\frac{\beta}{f_c} = L\sqrt{\frac{A_c}{N}}\frac{\sqrt{0.4\rho\varepsilon_r\varepsilon_o\left(1+4A.R.^2\right)t_{FO4}}}{A.R.p_w} \tag{30}$$

where β is the maximum delay fraction of clock period (25% for local and semi-global, and 90% for global wires), f_c is the clock frequency, ρ is the resistivity of the metal, ε_o is the permittivity of free space, ε_r is the relative permittivity of the dielectric material, p_w is the wire pitch, $A.R.$ is the wiring level aspect ratio and t_{FO4} is the FO4 gate delay. Equation (30) can be re-arranged to solve for wire pitch or the length of the longest interconnect. The expressions for p_{global}, L_{semi} (which are a function of p_{semi}) and L_{loc} are given by,

$$p_{glob} = \sqrt{\frac{A_c}{N}}\frac{L_{glob}}{A.R._{glob}}\frac{f_c}{\beta_{glob}}\sqrt{0.4\rho\varepsilon_r\varepsilon_o\left(1+4A.R._{glob}^2\right)t_{FO4}} \tag{31}$$

$$L_{semi} = \frac{\beta_{semi}}{f_c}p_{semi}A.R._{semi}\sqrt{\frac{N}{A_c}}\frac{1}{\sqrt{0.4\rho\varepsilon_r\varepsilon_o\left(1+4A.R._{semi}^2\right)t_{FO4}}} \tag{32}$$

$$L_{local} = \frac{\beta_{local}}{f_c}p_{local}A.R._{local}\sqrt{\frac{N}{A_c}}\frac{1}{\sqrt{0.4\rho\varepsilon_r\varepsilon_o\left(1+4A.R._{local}^2\right)t_{FO4}}} \tag{33}$$

Here p_{loc} is assumed constant and equal to twice the technology node. L_{global} is also assumed constant and equal to the chip die edge. Figure 9 schematically shows how wires in the wire length distribution are allocated into their respective interconnect tiers, with the dashed lines representing the tier boundaries that define the maximum length of wire per tier.

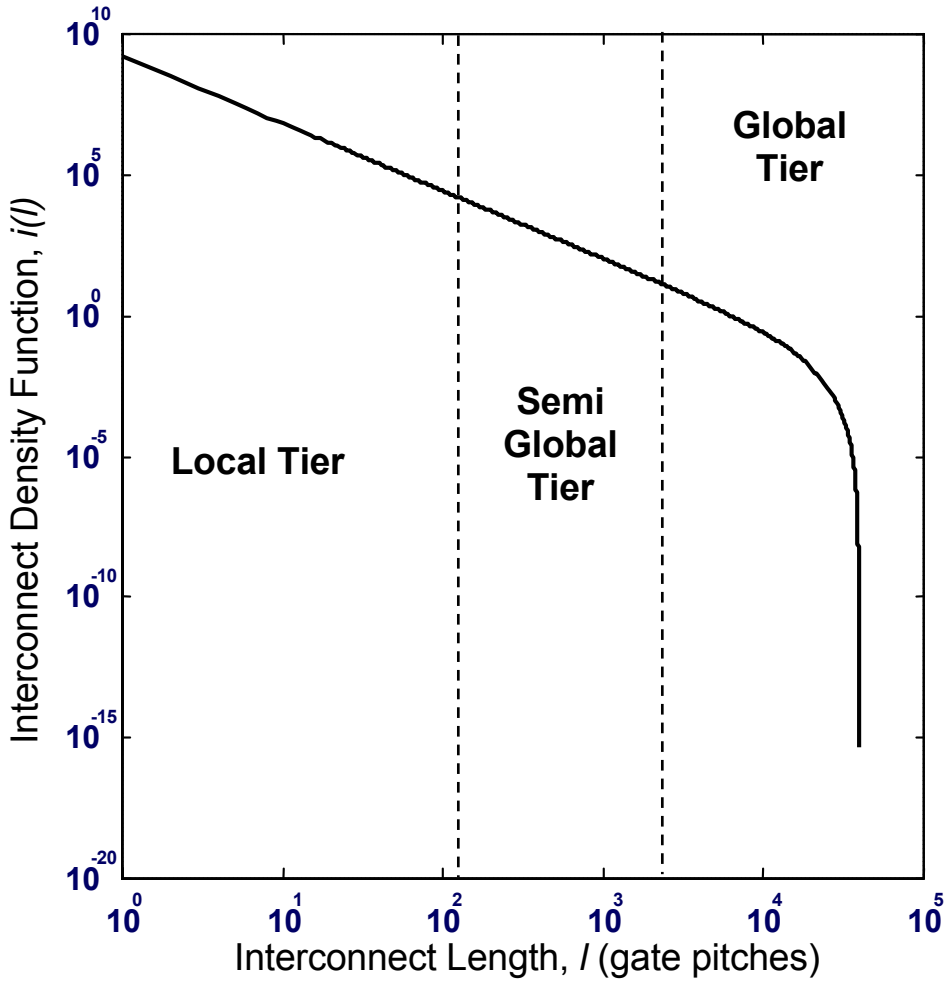


Figure 9: Wire length distribution for a 2 active layer 3-D configuration for the 50nm technology node schematically showing wire allocation to interconnect tiers with the dashed lines representing the tier boundaries L_{loc} and L_{semi}, respectively.

Using this interconnect tier allocation, the chip area, A_c, can be estimated using the above equations. However, equation (32) for L_{semi} results in a non-unique set of possible solutions for A_c. A further boundary condition is applied such that the wire limited chip area, A_c, is calculated considering that the total required wiring area, $A_{required}$, is equal to the total available area, $A_{available}$, in a multilevel network, hence it follows that,

$$A_{available} = A_c \, e_w \, n_{levels} = A_{required} \tag{34}$$

where e_w is the wiring efficiency factor that accounts for router efficiency and additional space needed for power and clock lines and is assumed to equal 0.4, and n_{levels} is the number of metal levels available for the multilevel network. An iterative procedure is employed to solve (32), such that for each possible solution of (32), 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 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. 3-D chip areas are determined using the same analysis with the values of N and k transformed to 3-D accordingly.

The model is applied to the microprocessor example shown in Table II for the 50 nm technology node [3] 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 per tier are equal.

The possible solutions for A_c and p_{semi} resulting from the numerical solution of Equation (32), for both 2-D and 3-D 2 active layers, are plotted for the high-performance IC ITRS 50 nm technology node in Figure 10 which shows the possible chip areas with the normalized semi-global 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

the minimum A_c the semi-global 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 semi-global tier. This causes re-routing of global wires to the semi-global tier, which in turn will require greater chip area. Under such circumstances, the semi-global tier begins to dominate and determine the chip area. Conversely, as p_{semi} decreases from the minimum A_c, the longer wires in the semi-global tier no longer satisfy the maximum delay requirement of that tier and they need to be re-routed to the global tier where they can enjoy a larger pitch. The population of wires in the global tier increases. 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.

PHYSICAL PARAMETER	VALUE
Logic Transistors	769×10^6
Memory Transistors	6284×10^6
Rent's Exponent, p	0.65
Rent's Coefficient, k	4
Operating Frequency	3 GHz
Technology node	50 nm
Number of wiring levels	9
ρ_{Cu}	1.673 x 10-6 Ω-cm
Dielectric Constant, Polymer	$\varepsilon_r = 1.5$
Wiring Efficiency Factor	0.4

Table II: ITRS projections for the 50nm technology node used to estimate chip areas for 2-D and 3-D 2 active layer configurations.

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 approximately 35%

smaller than that of the 2-D case. Moreover, since the total wiring requirement is reduced (as shown in Figure 6) the semi-global tier pitch is reduced for the 3-D chip at minimum A_c. This reduction in the semi-global 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 is a consequence of the fraction of wires that were converted from horizontal in 2-D to vertical VILICs in 3-D. It is assumed that the area required by VILICs is negligible.

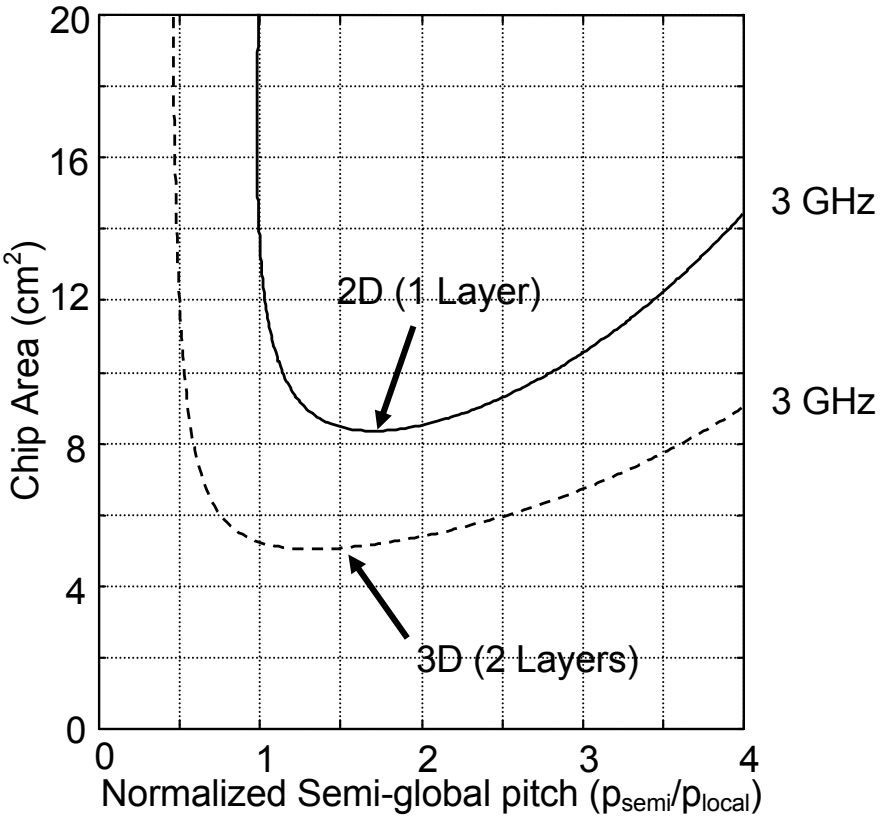


Figure 10: ITRS 50nm node wire-limited chip area vs. normalized semi-global pitch for 2-D and 3-D ICs at a fixed operating frequency of 3 GHz.

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 35% 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.

3.6 Improving Performance

So far, migration from 2-D to 3-D 2 active layers has demonstrated the ability of significantly reducing chip size while maintaining the performance invariant. However, as stated in previous chapters, the interest and motivation to migrate towards 3-D ICs is to prevent interconnect performance from limiting overall IC performance. While maintaining all other parameters constant, such as IC design, architecture, etc., the operating frequency is considered a measure of performance for the sake of comparison and the delay along the longest interconnect in an IC determines such an operating frequency. Therefore, it is of interest to study the possibility of reducing such interconnect delay using 3-D integration and hence improving performance. In the previous analysis, interconnects experienced a constant operating frequency boundary condition during migration to 3-D which caused the cross-sectional dimensions of interconnect wires to forcibly shrink and so reduce the chip area. In pursuing an improvement of performance such a boundary condition is removed and the question becomes: how much of a performance improvement is possible by migrating to 3-D?

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. For each performace

condition applied, the tier boundaries, L_{loc} and L_{semi}, are necessarily shifted in the wire length distribution (Figure 9) towards shorter wires such that the longest wire in each tier can satisfy the new delay condition. Consequently, wires that no longer satisfy the new delay condition are routed to higher tiers where they have larger cross sections and pitches. The effect of increasing p_{semi} and p_{global} on the operating frequency and A_c is shown in Figure 11. This illustrates how the optimal semi-global pitch (i.e. p_{semi} associated with the minimum A_c) increases to obtain higher operating frequencies. Also, as the semi-global tier pitch increases, chip area and subsequently interconnect length also increases, due to the routing of wires to higher tiers where they require larger area. However, it can be observed from Figure 11 that the increase in chip area still remains well below the area required for the 2-D case.

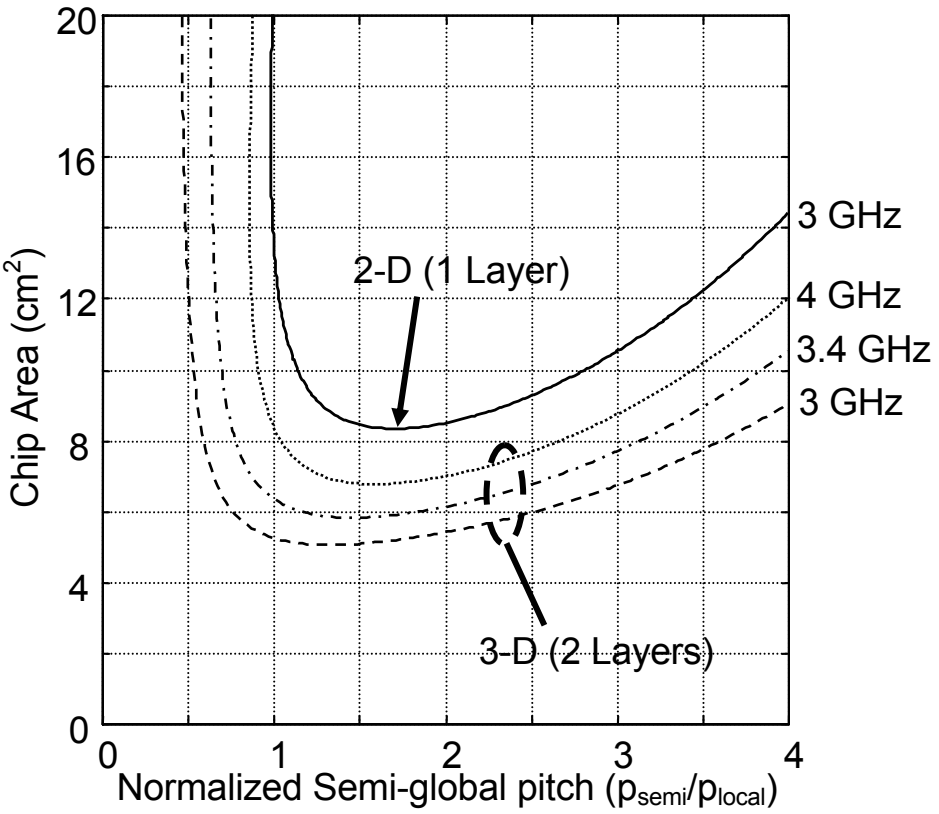


Figure 11: 3-D chip operating frequency (performance) and chip area increases with increasing semi-global wiring pitch.

While it is possible to improve performance by imposing smaller signal delay conditions, a competing effect of increasing chip size is also exhibited. Such increases in chip size are also non-linear considering the exponential nature of the wire length distribution as shown in Figure 9. To illustrate, any incremental reduction in the signal delay condition at the tier boundaries is accompanied by an exponentially increasing population of wires that are routed to higher tiers. Eventually, a saturation effect in the performance improvement can be expected as a function of chip area and is observed in Figure 12. Here, the minimum chip area is plotted for every time the operating frequency is increased.

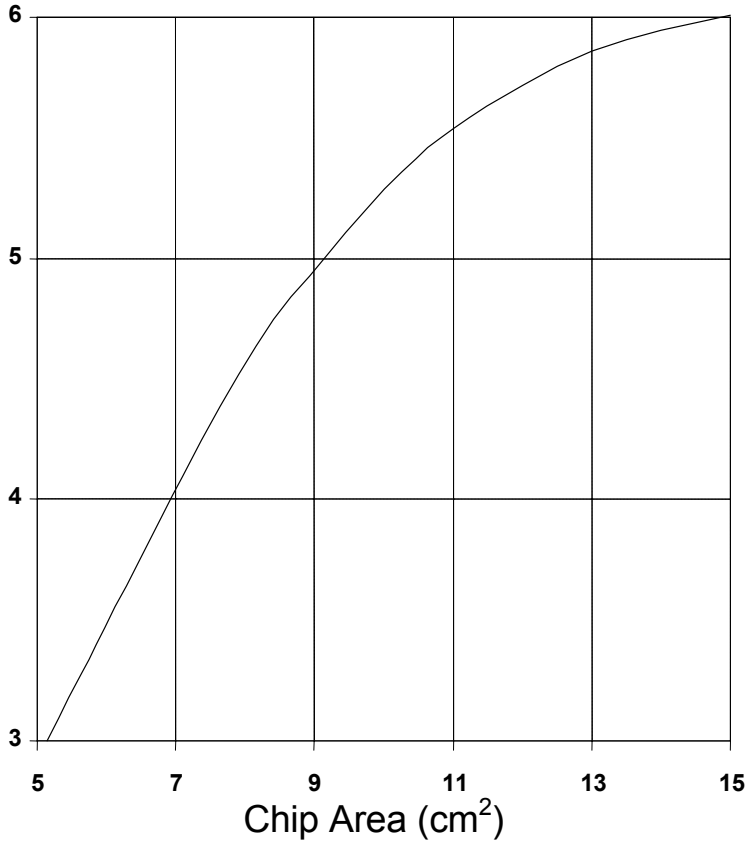


Figure 12: ITRS 50nm node wire-limited operating frequency vs. chip area for 3-D 2 active layer case showing a saturation effect in improvement due to exponential routing of wires to higher tiers.

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, this analysis is extended to study the effect of more than two Si layers at the 50nm technology node and also the effect of increasing the number of available metal layers over a number of ITRS projected nodes [3].

3.7 Effect of Increasing Number of Active Silicon Layers

A natural extension of the above analysis is to consider the effects of increasing the number of active silicon layers in the 3-D configuration beyond 2. The analysis itself is fairly straightforward as it only requires that logic and memory blocks are to be redistributed among a greater number of active layers, n, which is substituted in the above equations (22) and (23). The wire length distribution, chip area and performance improvement analyses are performed as before for the 2 active layer case.

Of importance is the choice of platform for a meaningful comparison of scenarios where the number of active layers is varied. Realistically, it is reasonable to compare the performances of all such scenarios while maintaining a constant die footprint or chip size. This allows for a valid comparison while maintaining a manufacturability perspective. From this point onwards, any comparison performed maintains the chip area across all comparable scenarios equal to the 2-D footprint at the respective technology node. For instance, if the comparison is performed at the 50nm node, then all configurations compared share an equal chip area of 8.17cm^2 [3]. The results for the number of active layer comparison are summarized in Figure 13 for the ITRS 50 nm technology node. The 2-D case is a special 3-D

case as it only has one active layer. The signal RC delay along the longest wire for each case is normalized and compared to the 2-D (single active layer case).

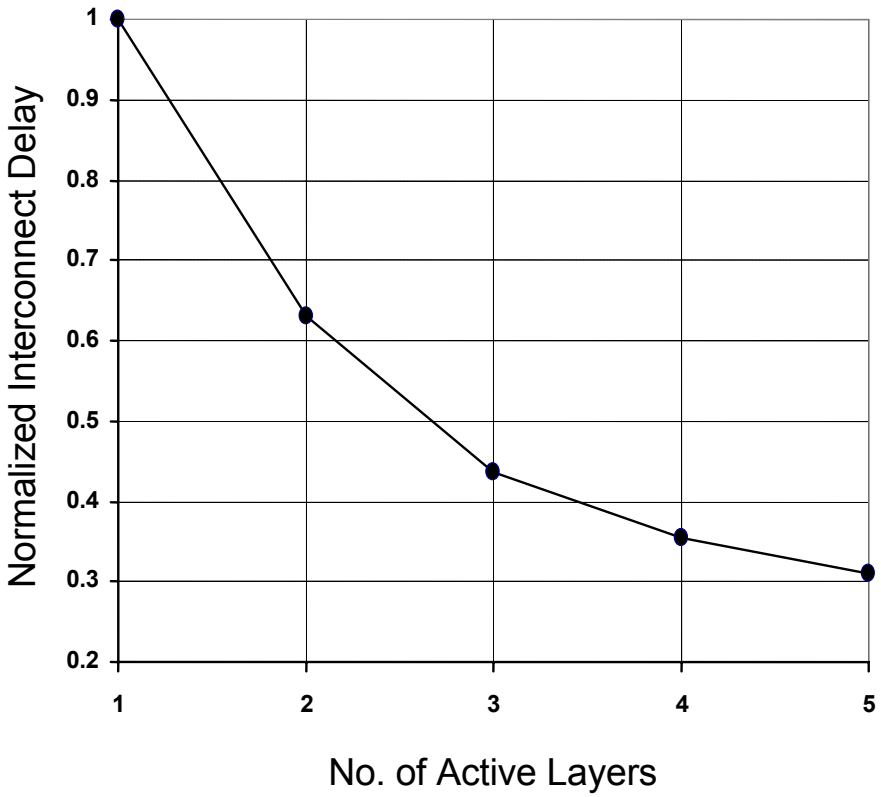


Figure 13: Comparison of the longest interconnect delay normalized to the 2-D case for a number of active layers for the ITRS 50nm node.

Figure 13 shows how further improvements in interconnect delay, which is translated to higher operating frequencies, can be obtained by introducing more active silicon layers in the 3-D configuration. However, a saturation trend is also observed here, whereby delay improvements are reduced with increasing number of active layers. This saturation occurs for the same reasons the same trend was observed in the previous section while improving performance. For all the cases considered in Figure 13, any incremental improvement in the

tier boundary delay condition results in an exponential number of wires routed to higher tiers where they contribute to increasing the chip area. By the time all chip areas are equal for comparison purposes, the improvements in delay suffer from this exponential increase in area.

3.8 Effect of Increasing Number of Metal Layers

Up to this section, the analysis considered has conserved the total number of metal layers between 2-D and 3-D configurations. Specifically, for the ITRS 50nm technology node projection, this has been 9 layers of metal. However, this is unlikely to be the case. For instance, depending on the interconnect system and IC architecture, it is possible that each active layer may have specifically associated with it a certain number of wiring levels, such as a local and semi-global tiers, with an overall global tier shared by all active layers. Such a scenario would significantly increase the number of available metal layers.

Furthermore, the total number of metal layers may depend on the actual 3-D integration technology used (this will be discussed in more detail in a later chapter). For example, die bonding is a process whereby 2 independently fabricated ICs, each with their own interconnect structure, can be bonded together at some interconnect level, essentially forming a 3-D IC with 2 active layers. This particular case would result in a total number of metal layers that is the total sum for each IC.

With this in mind, the previous analyses are performed again with the number of metal layers doubled in the 3-D configurations. The analyses were performed over all technology nodes and the results are presented in the following section.

3.9 Summary of 3-D Integration Performance Analysis Results

The above discussion mainly focused on the performance analysis of 3-D integration for the ITRS 50nm technology node. In this section, the analysis is extended to a number of ITRS node projections which are listed in Table III.

	1999	2001	2003	2005	2008	2011
Technology Node	180nm	150nm	120nm	100nm	70nm	50nm
Ac (cm^2)	4.5	4.5	5.76	6.22	7.13	8.17
N_{Logic} (M)	12	24	48	96	272	769
N_{Memory} (M)	98	196	393	786	2222	6284

Table III: A summary of ITRS technology nodes and projections used in the analysis.

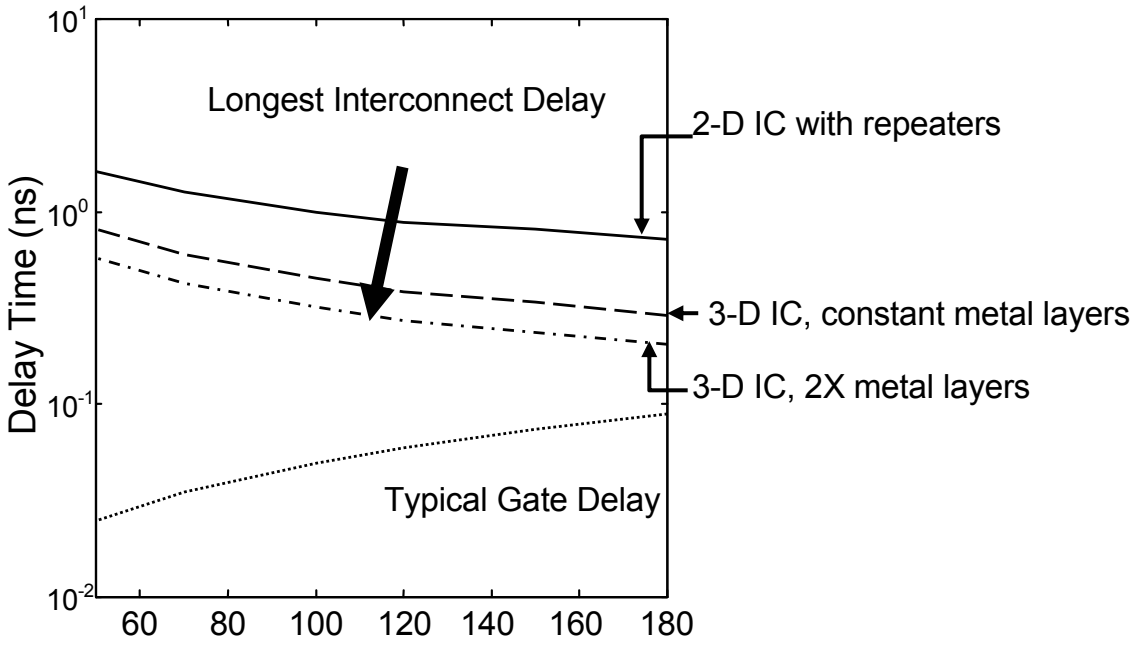


Figure 14: A summary of signal delay results for several cases covering a wide range of technology nodes as projected by ITRS 1999 [3].

As in Section 3.7, for comparison purposes, all cases shown share the same footprint as in the 2-D case for each technology node. The results are plotted and shown in Figure 14. The signal delay is plotted as a function of technology node for the following cases: 2-D longest interconnect delay and gate delay (as in Figure 2.2); 3-D longest interconnect delay with 2 active layers and conserved number of metal layers; and 3-D longest interconnect delay with 2 active layers with twice the number of metal layers. Figure 14 shows that by using 3-D integration, it is possible to reduce the interconnect delay by approximately 37% as compared to 2-D. This reduction in delay can be improved by a further 30% by doubling the number of metal layers.

This performance analysis has shown that by migrating 2-D ICs to a 3-D configuration it is possible to push interconnect delays to much lower values and prevent interconnect delays from limiting the performance of future advanced ICs. Furthermore, the analysis shown is considered conservative due to the random redistribution of logic and memory blocks assumption. Further improvements are possible by taking the specific IC design into consideration and applying some intelligence in the logic/memory block redistribution such that only long and performance limiting interconnects are replaced with VILICs.

Chapter 4

THERMAL ANALYSIS OF 3-D ICS

An extremely important issue in 3-D ICs is that of heat dissipation [13,14]. Thermal effects are already known to significantly impact interconnect/device reliability and performance in high-performance 2-D ICs [15,16,64]. The problem is expected to be exacerbated by the reduction in chip size, assuming that the 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.

Thermal issues, however, are not expected to plague 3-D systems alone. Figure 1, for instance, shows a disturbing trend in power density for Intel® commercial processors. While 3-D can potentially complicate the power consumption, dissipation and die temperature issues of ICs, nonetheless, superior thermal management solutions will soon be required from which both 2-D and 3-D systems can benefit. With such a cautionary note, the following discussion will provide an analytical thermal treatment of ICs which can be applied to both 2-D and 3-D ICs.

The majority of the thermal 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 multi-layer 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 as compared to Si [17, 65]. Hence, the heat dissipation issue can become even more acute for 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 [15].

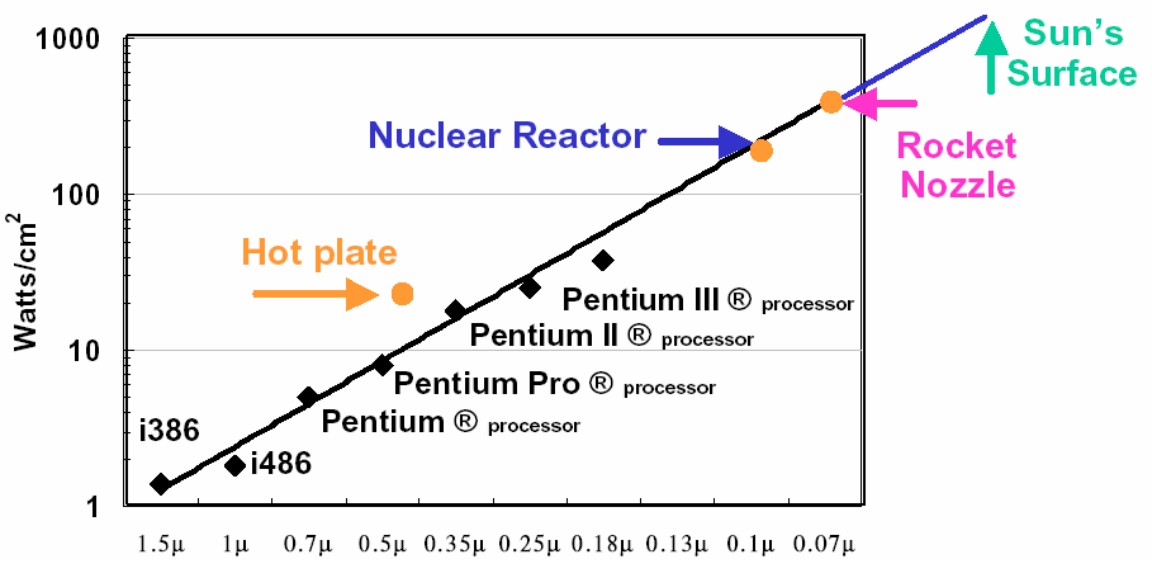


Figure 1: Evolution of power density for Intel® processors [64]. *(Courtesy of Intel®).*

In this chapter, a detailed thermal analysis of high performance three dimensional (3-D) ICs is presented under various integration schemes. The analysis presented here is the culmination of collaborative work with T. Y. Chiang at Stanford University. A complete thermal model including power consumption due to both transistors and interconnect joule heating from multiple strata is presented. With the effect of vias, as efficient heat dissipation paths, taken into account, this model provides more realistic temperature rise estimation for 3-D ICs. These vertical links and vias have much higher thermal conductivity and hence can effectively reduce the thermal resistance caused by the ILD layers. Ignoring the effect of these

structures can result in overly pessimistic estimations predicting unacceptably high 3-D chip temperatures. Recently, a model has been developed to quantify the via thermal effect in 2-D structures [66,67]. Here, this compact analytical model is applied to evaluate temperature rise in 3-D structures, incorporating via effect and power consumption due to both devices in active layers and interconnect joule heating [68]. The results show excellent agreement with the 3-D finite element simulations using ANSYS [70]. With the effect of vias, as efficient heat dissipation paths, taken into account, this model provides more realistic temperature rise estimation for 3-D ICs as compared to previous work [34]. Furthermore, tradeoffs between power, performance, chip area and thermal impact are evaluated.

4.1 Thermal Modeling Incorporating Via Effect

According to ITRS [3], although the average power density for high performance microprocessor will remain relatively constant throughout the technology nodes, current density in the wires will rise significantly (Fig. 2).

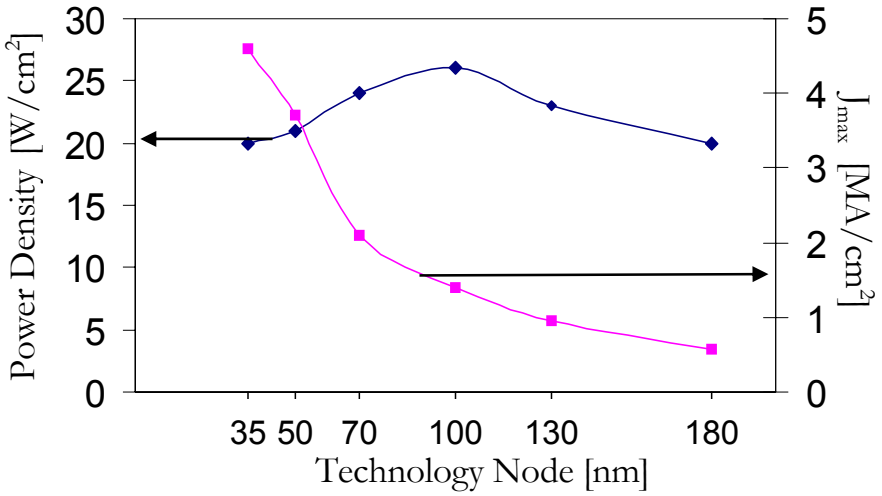


Figure 2: Trends of chip power density and interconnect J_{max} along technology nodes suggested by ITRS [3]. Chip power density is the total power of the chip divided by chip size.

Furthermore, Cu resistivity will increase due to barriers, surface scattering and skin effect. Thus interconnect joule heating will become significant. In addition, low-k dielectrics with poor thermal conductivity (Fig. 3), will not only lead to higher interconnect temperature in 2-D ICs but also impact the device temperature in various active layers in 3-D ICs (Fig. 4).

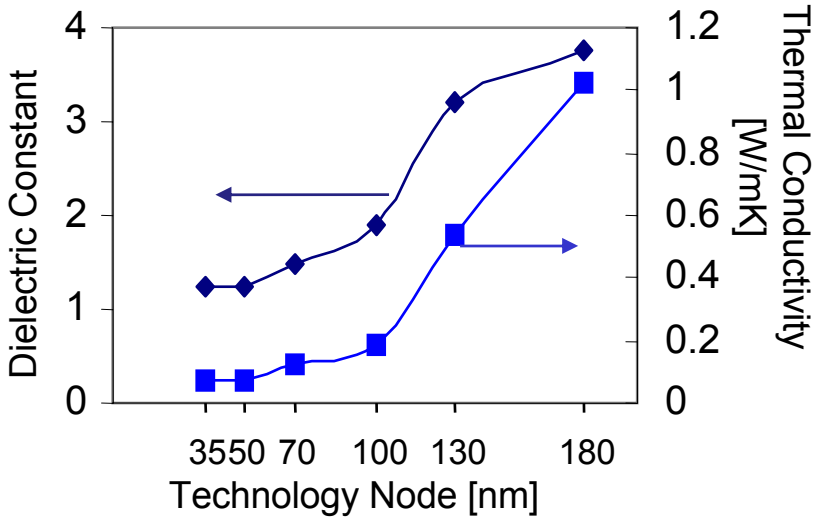


Figure 3: Both dielectric constant and thermal conductivity of ILD materials decrease with advanced technology nodes.

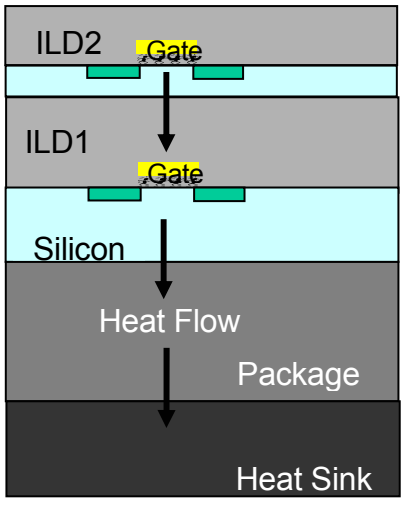


Figure 4: Schematic of multi-level 3-D IC with a heat sink attached to Si substrate.

As seen in Figure 5, the ratio of thermal resistance caused by ILD layers (R_{ILD}) to required package (including glue layers, heat sink) thermal resistance (R_{pkg}) increases rapidly for future technology nodes. The required R_{pkg} is the maximum allowed value which gives the maximum junction temperature specified in ITRS. With multiple active layers, R_{ILD} will become the dominant factor to determine temperature rise in 3-D ICs.

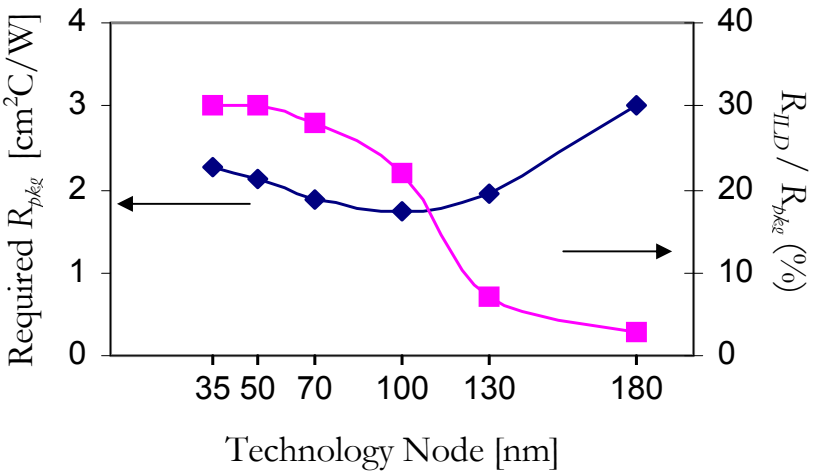


Figure 5: The required package thermal resistance, R_{pkg}, to achieve the maximum junction temperature specified in ITRS and the ratio of R_{ILD} and R_{pkg} vs. technology nodes.

Absent from previous work [34] is the effect of vias on the thermal behavior of 3-D ICs. Vias are vertical metal connections that provide efficient heat conduction paths between heat generating layers that are separated by thermal insulators such as ILDs. As a result, all previous analyses provide unrealistic and pessimistic projections for 3-D IC die temperature rises. Figure 7 shows the effect of considering vias as heat conduction paths, and their separation, on the effective thermal conductivity of different ILD materials [69]. As the via density increases, the effective thermal conductivity of the ILD materials approaches that of the vias themselves.

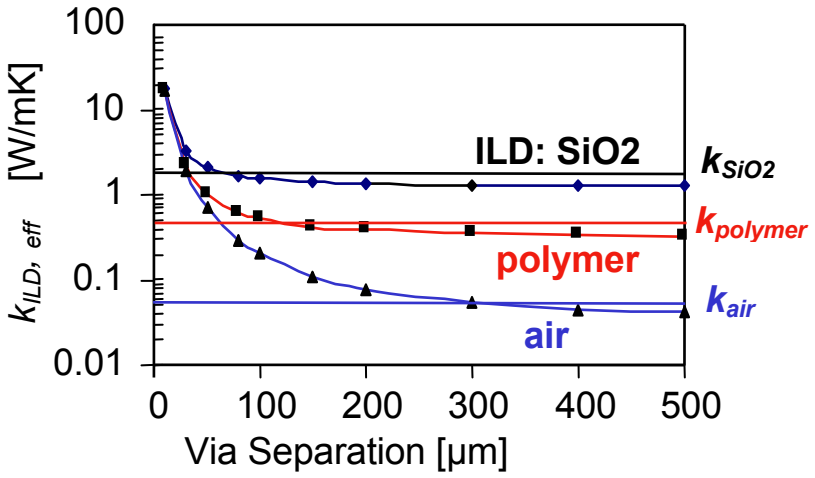


Figure 6: The effect of including vias and their separation on the effective thermal conductivity of different ILD materials.

The analytical expression, derived based from first principles, to evaluate temperature rise in 3-D structure is given below [66,68]:

$$\overline{T_{Si\_N}} = T_{amb} + \sum_{m=1}^{N} \{ [\underbrace{\sum_{n=1}^{N_m} \frac{t_{ILD,mn}}{k_{ILD,mn}\, s_{mn}} \eta_{mn} \left(\sum_{i=n}^{N_m} j_{rms,mn}^2 \, \rho H_{mn} + \sum_{j=m+1}^{M} 2\Phi_j \right)}_{\text{Temperature rise caused by ILDs}}] + \underbrace{R_m \left(\sum_{k=m}^{M} Q_k \right)}_{\substack{\text{Temp. rise caused by PKG,} \\ \text{glue layer, Si sub.}}} \}$$

where:

- T_{amb}: ambient temperature.
- M : number of strata.
- N_m: number of metal levels in the mth stratum.
- mn : the nth interconnect level in the mth stratum.
- t_{ILD}: thickness of ILD.
- k_{ILD}: thermal conductivity of ILD materials.
- s : heat spreading factor [66].
- η : via correction factor, $0 \leq \eta \leq 1$ [66].

- j_{rms} : root-mean-square value of current density flowing in the wires.
- ρ : electrical resistivity of metal wires.
- H : thickness of metal wires.
- $\varnothing$: power density of active device layer.
- Q : total power consumption of m^{th} stratum, including power consumed by active layer and interconnect joule heating.
- R : thermal resistance of glue layer and Si layer for each of the stratum, with R_1 represents the total thermal resistance of package, heat sink and Si substrate.

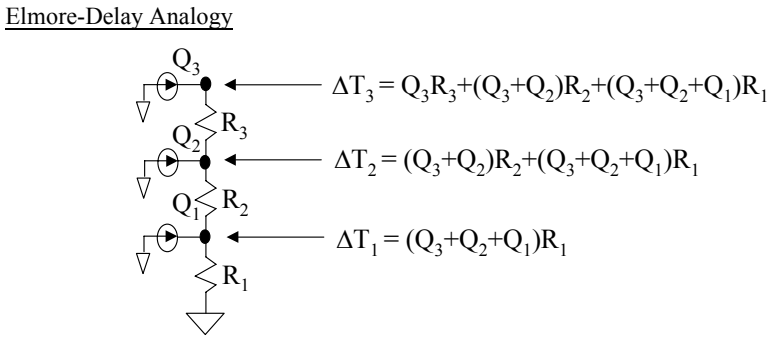


Figure 7: Elmore-Delay model – electrical analogy for the thermal mode employed [66].

Via effect is incorporated into the expression by the via correction factor η ($0 \leq \eta \leq 1$), with $k_{ILD,eff} = k_{ILD}/\eta$, where $k_{ILD,eff}$ is the effective thermal conductivity of ILD with the help of the via effect and k_{ILD} is the nominal thermal conductivity if the via effect is ignored [66]. Power consumption due to both active (device) layers and interconnect joule heating are included. This expression can be better understood by comparing it with the Elmore-delay model, shown in Figure 7, following an electrical-thermal analogy. The model is validated by comparing with full chip thermal simulation done using ANSYS in [70]. The result from analytical expression shows excellent agreement with ANSYS [68]. However, the analytical model takes much less computation time and provides better insight.

4.2 Power Analysis of 3-D ICs

In general, performing any comparison between 2-D and 3-D configurations of the same IC is a task wrought with difficulties. The source of the dilemma is the lack of common ground on which to perform the comparison without bias. For instance, the operating frequency can be maintained constant, yet the chip areas and hence power densities must change. Conversely, if the chip areas are assumed invariant, as was the basis for the comparison in Chapter 3, then the operating frequencies are necessarily different. Such an assumption serves well for a performance comparison. Comparing power dissipation and temperature rises as a result of migration from 2-D to 3-D, on the other hand, is a different matter. To address this thermal comparative issue, the analysis in this section is again focused on the ITRS projection at the 50nm technology node. However, several 3-D integration cases and scenarios are explored to cover an adequate space for the comparison.

A summary of the wire-pitch limited 2-D and of the different 3-D integration cases that are used in the comparison is listed in Table I. All the data in this table are calculated based on the 50nm technology node and the thermal resistance of the package is assumed to be 2.15cm^2 °C/W from ITRS projections for 2-D ICs at the 50 nm node. The data in the 2-D column represents the standard 2-D IC. 3-D, Case 1, is a special 3-D integration case in that memory and logic from the 2-D are each dedicated to separate active layers without any modifications to the wiring. The resulting chip area, A_o, is determined by the larger logic area and power dissipation is unchanged relative to 2-D. The remaining 4 3-D cases are obtained, and compared to 2-D, by modifying the chip wiring. Their characteristics are summarized below:

- 3-D, Case 2: Equal f_c and decreased A_c;
- 3-D, Case 3: Equal f_c and A_c;
- 3-D, Case 4: $2f_c$ and equal A_c;
- 3-D, Case 5: Equal A_c with f_c determined by maintaining 2-D P_{Total}.

where A_c is the chip area, f_c is the operating frequency and P_{total} is the total power dissipation. The different characteristics of these 3-D integration cases give rise to different power dissipations as summarized in Table I. The dynamic power dissipation components considered are due to logic, interconnect (local, semi-global and global), clock distribution and repeaters and are calculated using $P_{Dynamic}=1/2\alpha CV_{dd}^2 f_c$ where α is the activity factor (assumed to be 0.1), V_{dd} is the supply voltage obtained from ITRS, f_c is the operating frequency and C is capacitance. Other power dissipating components include memory, I/O pads and static components such as leakage and short-circuit currents, are all combined under P_{Other}.

The capacitance, C, is calculated for each component to determine the associated power dissipated. For P_{Logic}, the device capacitance is calculated by considering gate oxide capacitance, overlap capacitance, and junction capacitance all of which are calculated from ITRS data [3]. Interconnect capacitances for the local, semi-global and global tiers are found from the wire-length distribution and the dimensions of the wire pitches for each tier. Clock distribution capacitances are calculated using the BACPAC model proposed in [71] by considering a buffered H-Tree model. Power dissipated by repeaters is calculated based on the driver capacitances and the number of repeaters which is modeled in the next chapter. P_{Other} is

determined in the 2-D case to be the sum of remaining components to achieve the ITRS projected total power dissipation for this generation. Since this component is assumed dominated by dynamic dissipation it is considered linearly dependent on the operating frequency for all 3-D cases.

	2-D	3-D, Case 1	3-D, Case 2	3-D, Case 3	3-D, Case 4	3-D, Case 5
Active Layers	1	2	2	2	2	2
f_c (MHz)	3000	3000	3000	3000	6000	3559
Feature Size (nm)	50	50	50	50	50	50
Chip Area (cm^2)	8.17	4.25	4.51	8.17	8.17	8.17
Memory Area (cm^2)	3.92	3.92	3.92	3.92	3.92	3.92
Logic Area (cm^2)	4.25	4.25	5.1	12.42	12.42	12.42
P_{Logic} (W)	34.8	34.8	34.8	34.8	69.6	41.28
P_{Local} (W)	17.4	17.4	17.44	20.66	10.44	6.19
$P_{Semi\text{-}Global}$ (W)	14.63	14.63	6.89	8.16	30.68	18.2
P_{Global} (W)	6.96	6.96	4.18	5.63	11.78	6.99
P_{Clock} (W)	34.8	34.8	22.97	27.21	56.93	33.76
$P_{Repeaters}$ (W)	45.24	45.24	29.7	35.19	73.6	43.65
P_{Other} (W)	20.17	20.17	20.17	20.17	40.34	23.93
P_{Total} (W)	174	174	136.15	151.82	293.37	174
Power Density Per Active Layer (Wcm^{-2})	21.30	20.47	15.09	9.29	17.95	10.65

Table I: Comparison of power dissipation due to logic, interconnect, clock distribution and repeaters for 2-D and 3-D ICs with 2 active layers for ITRS 1999 50nm technology node. 3-D IC cases are presented for comparison by varying the chip area, A_c, and operating frequency, f_c, and represent the same 2-D IC (conserving feature size, number of transistors and functionality) converted to 3-D with 2 active layers.

In 3-D case 2, the total power dissipation is seen to decrease primarily due to the reduction in the wiring requirement thus reducing the interconnect power dissipation and the number of required repeaters, and minimizing the clock distribution network. 3-D case 3 is

associated with a larger chip area which requires longer interconnect lines, a larger number of repeaters and clock-distribution network all of which increase the power dissipation as compared to 3-D case 2. 3-D case 4 shows a dramatic increase in the power dissipated primarily due to the significant increase in operating frequency. 3-D case 5 illustrates the increase in the operating frequency if the chip area and the power dissipation requirements are maintained constant to 2-D.

Although the total power consumption can be reduced, in some cases, by going from 2-D to 3-D ICs, as shown in Table I, due to the reduction in the interconnect and the clock network related capacitance, the heat removal capability can deteriorate as the upper active layers experience longer heat dissipation paths to the heat sink. The analytical thermal model, including the via effect, discussed in the previous section is therefore employed here to project the temperature rise for each active layer for all the integration schemes outlined in Table I.

Figure 8 compares the die temperatures for the 2-D and the different 3-D integration cases. The 2-D die temperature represents the projected, and therefore acceptable, temperature from ITRS [3] of 104C. The die temperatures for all the 3-D integration cases occupy a wide range from 99C to 149C. The lowest temperature achieved is that for 3-D case 3. Here, although the wiring has been reduced by migrating to 3-D, the operating frequency and the chip area have been maintained constant as compared to 2-D. The power density, then, is significantly lower than in the 2-D case giving rise to a lower die temperature. 3-D case 4 exhibits the highest die temperature of the 3-D cases where the wiring has been modified. In this case, although the die footprint is equal to 2-D, the operating frequency is significantly increased, increasing the power density and giving rise to a high die temperature of 137C. In

general, 3-D ICs have the advantage of either reduced chip area (cases 1 & 2) or increased operating frequency (cases 4 & 5) or reduced die temperature (case 3). Although in 3-D case 4, the temperature is significantly increased, it is still much lower than that estimated with via effect ignored [34,70].

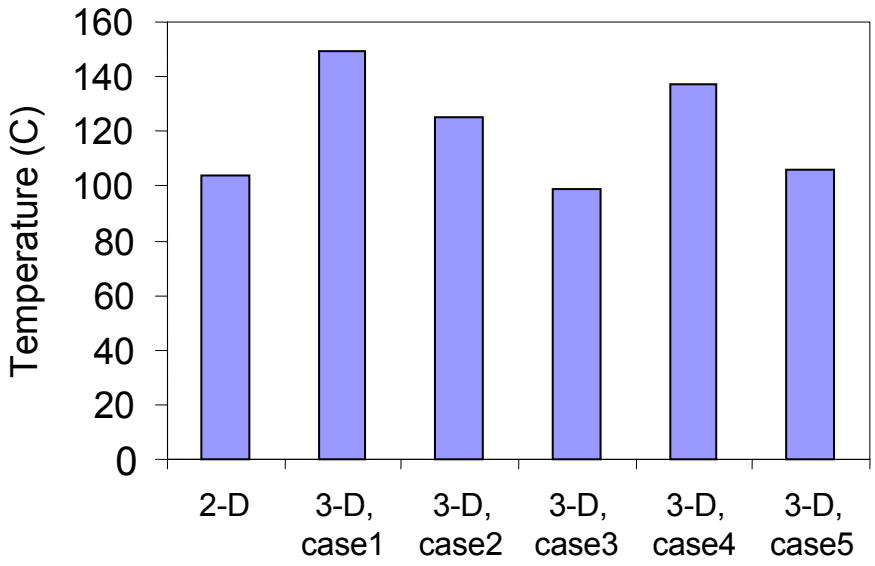


Figure 8: Comparison of temperature performance among 2-D ICs and four different two-active-layer 3-D ICs scenarios.

The comparison can be developed further. It is desirable to put memory in close proximity to logic circuitry to reduce latency in high performance microprocessors. 3-D ICs provides an excellent opportunity to stack memory and logic. The power consumption in on-chip memory is generally less than 10% of total power consumption and the area occupied by memory and logic are comparable for the 50nm ITRS technology node [3]. With these assumptions, several schemes are developed, and applied in the following analysis to 3-D case 4. These schemes are illustrated in Figure 9 where four 3-D stack schemes are shown, each with a different configuration of memory and logic. By applying the analytical thermal mode

discussed above, the active layer temperature performance for each scheme is shown in Figure 10.

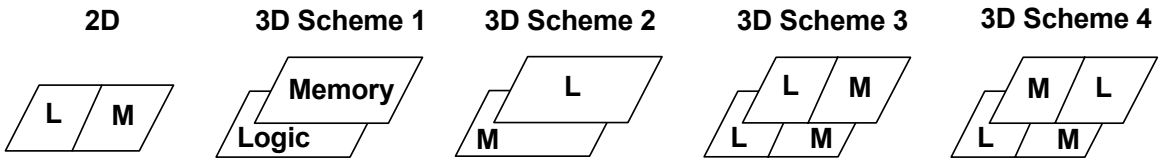


Figure 9: 3-D integration schemes.

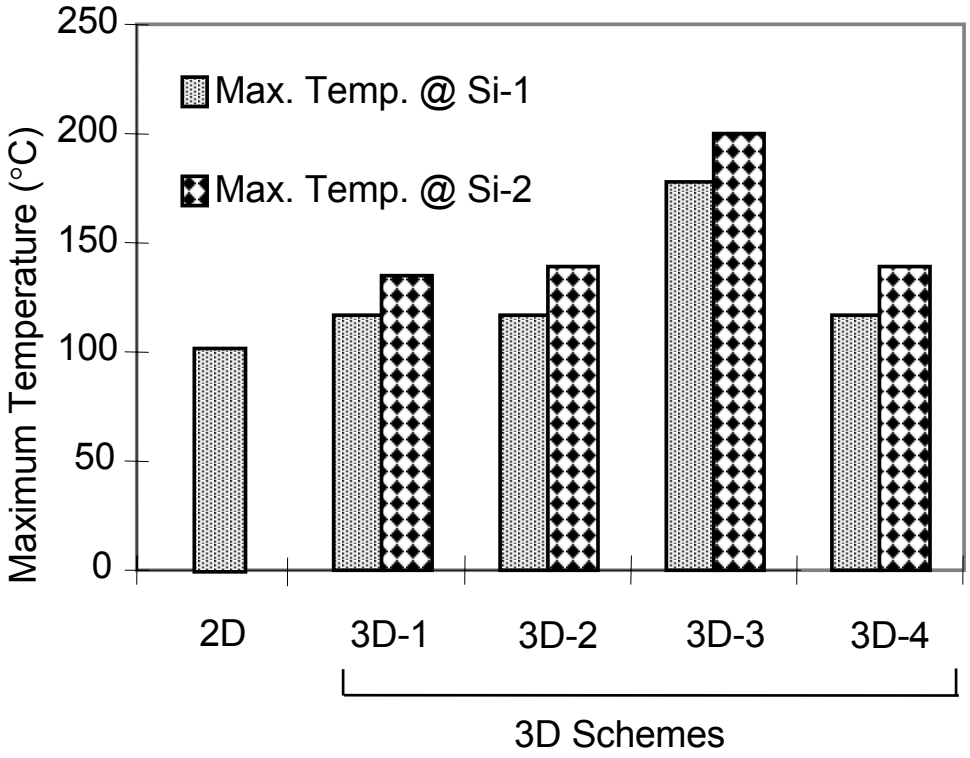


Figure 10: Comparison of temperature rises for 2-D and active layer temperatures for several 3-D schemes.

In general, the 3-D schemes all exhibit higher active layer temperatures as compared to 2-D due to the higher operating frequency assumed for 3-D. Furthermore, temperatures at the

second active layers are consistently higher than the first, which is closest to the heat sink. Although this temperature discrepancy between active layers is due to the presence of a thermally insulating ILD layer with vias acting as heat conduction paths, the temperature rise is minimal as compared to previous work [34,70] due to the incorporation of the via effect. The lowest rise in temperature is associated with schemes where logic and memory alternate in the vertical dimension. The highest temperature reached is for scheme 3, where logic is placed on top of logic. In this scheme, since logic is responsible for the majority of the power dissipated, the local power density is much higher as compared to other schemes, giving rise to higher active layer temperatures.

By using a compact analytical thermal model incorporating the effect of vias as efficient heat conduction paths, a comparison has been presented to demonstrate the possibility of a high performance 3-D integration scenario through thermally responsible design. It is shown that with careful thermal designs, 3-D ICs can have similar thermal capability as that of 2-D ICs. For the case where higher temperature is not avoidable (high performance), better circuit design and advanced packaging or thermal management solutions will be necessary, from which not only 3-D ICs can benefit, but also 2-D.

Chapter 5

IMPLICATIONS FOR CIRCUIT DESIGN AND SOC APPLICATIONS

3-D integration offers a new dimensional degree of freedom to circuit design and IC architecture. In the preceding chapters, the analysis has been intentionally generalized free of any constraints as to the particular design or architecture. It has generally been assumed to apply to wire-pitch limited, high-performance ICs which consist of integrated logic and memory elements. Furthermore, in terms of functionality, each active layer in a given 3-D configuration has been assumed identical. In reality, benefits of migration to 3-D need not be limited to such cases.

In the sections below, a number of issues are discussed where each active layer is designed with specific and distinct functionality. For instance, in Chapter 4, the possibility of separating memory from logic components, where each is allocated to a separate active layer, was discussed. Other functionalities may be allocated to distinct active layers as well. Repeaters, for example are studied in the following section. 3-D also opens up the possibility of integrating heterogeneous technologies on-chip. Analog ICs can be designed and fabricated on active layers on top of digital. Or optical I/O layers can be integrated on the top-most active layer to facilitate optical off-chip communications or on-chip clock distribution. As such, 3-D integration provides a path towards System-On-Chip designs.

5.1 Repeater Insertion

For deep submicron 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 [54] 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 [17, 54]. 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 [72] 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 repeaters is presented in the following discussion.

5.1.1 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 [34,73]. Repeaters are assumed 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 Chapter 3). To illustrate, the local tier cannot have any non-repeated lines that exceed a maximum allowable length, L_{opt} in Equation (2.3). 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 give tier is not arbitrary. Repeated wires are assumed to have repeaters

inserted optimally and the signal delay along such wires can be described by Equation (2.6), also shown below:

$$\tau_d = 2L\sqrt{0.4rct_{FO4}} \qquad [1]$$

where L is the wire length, $t_{FO4} = 15r_o c_{NMOS}$ is the fanout-of-four delay of a minimum repeater size, and r and c are the line resistance and capacitance per unit length, respectively. The maximum allowable length per interconnect tier is calculated based on Equations (3.31), (3.32) and (3.33). As an example, a schematic figure describing the critical lengths for the local tier is given in Figure 1.

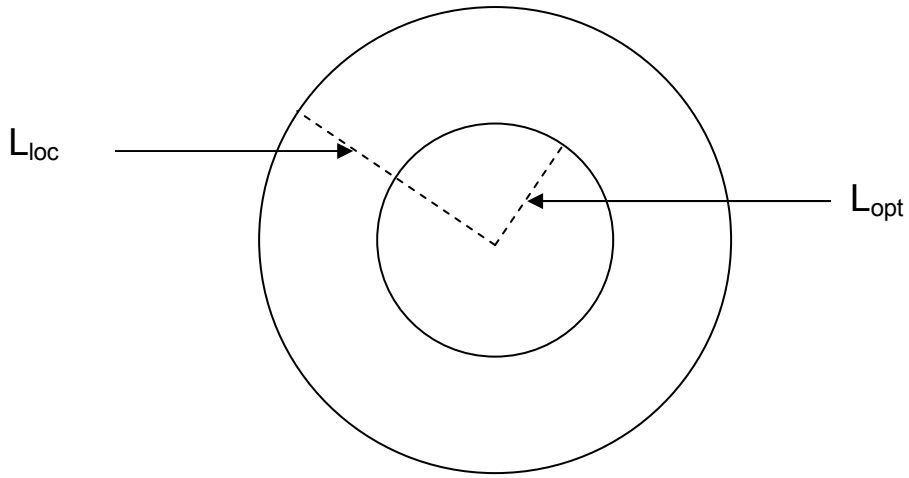


Figure 1: A schematic showing the interconnect length boundaries for the local tier. L_{opt} is the maximum allowed length of an interconnect without repeatering. L_{loc} describes the maximum length of any wire in the local tier. Interconnects with lengths $L_{loc} \leq l \leq L_{opt}$, require repeaters. Appropriate conditions are applied to the remaing tiers.

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_{opt}, given by (see Equation 2.4):

$$s_{opt} = \sqrt{\frac{r_o c}{3 r c_{NMOS}}} \qquad [2]$$

where r_o and c_{NMOS} are the minimum repeater resistance and NMOS capacitance, respectively, and r and c are the line resistance and capacitance per unit length.

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 schematically in Figure 1, 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/O's connecting this central region to its surroundings is given by $k N_{central}^{\ p}$ where k is Rent's constant and p is Rent's Exponent. The probability, P_1, of an I/O of any gate within this area of πL_{opt}^2 to reach outside this area can be represented as:

$$P_1 = \frac{k N_{central}^{p}}{k N_{central}} = N_{central}^{p-1} \qquad [3]$$

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

$$P_1 = \kappa^{p-1} L_{opt}^{2(p-1)} \qquad [4]$$

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 πL_{loc}^2, reaches outside this area is given by:

$$P_2 = \kappa^{p-1} L_{loc}^{2(p-1)} \quad [5]$$

Hence, the probability that the I/O of any gate within the entire local tier remains *inside* the tier is given by *(1-P_2)*. Therefore, the total probability, P_{loc}, that an interconnect will satisfy the length condition $L_{opt} \le l \le L_{loc}$ is given by:

$$P_{loc} = P_1(1 - P_2) \quad [6]$$

Then, 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/O's of all the gates:

$$I_R = P_1(1 - P_2) k \kappa L_{loc}^2 \quad [7]$$

The optimum number of repeaters per unit length of wire *(l/l_{opt})* is given by $\sqrt{\frac{0.4rc}{4.2r_0 c_{NMOS}}}$. To estimate the total number of repeaters an average length of wire, l_{avg}, is considered where:

$$l_{avg} = \left(\frac{L_{opt} + L_{loc}}{2} \right) \quad [8]$$

Hence, the total number of repeaters can be expressed as:

$$P_1(1 - P_2) k \kappa l_{avg} \sqrt{\frac{0.4rc}{4.2r_0 c_{NMOS}}} L_{loc}^2 \quad [9]$$

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

$$A_{R,loc} = P_1(1 - P_2) k \kappa l_{avg} \sqrt{\frac{0.4rc}{4.2r_0 c_{NMOS}}} L_{loc}^2 B_0 s_{opt,loc} \quad [10]$$

where B_o is the minimum repeater size ($\approx 60WL$) and $s_{opt,loc}$ is the optimum multiple of minimum repeater size for the local tier. All parameters in Equation (10) can be calculated for a given technology node based on [3]. This procedure is repeated to account for all the interconnect tiers to estimate the total area, $A_{R,total}$, utilized by repeaters, i.e.,

$$A_{R,total} = A_{R,loc} + A_{R,semi} + A_{R,glob} \quad [11]$$

Using the methodology presented above, the percentage of logic area utilized by repeater insertion is calculated at each technology node based on [3] for a range of Rent's Exponents. It can be observed from Figure 2 that inserting these repeaters will result in significant area penalty, especially beyond the 70nm technology node.

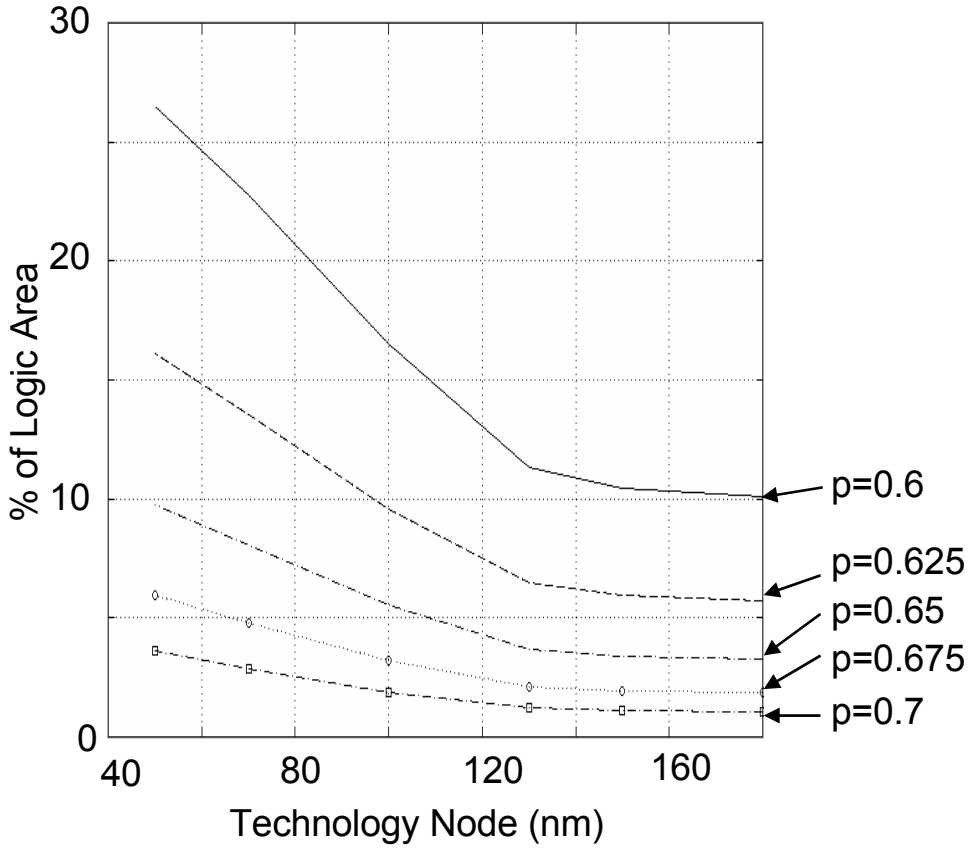


Figure 2: Fraction of logic area used by repeaters for different technology nodes based on ITRS projections [3] and different Rent's Exponents. As much as 27% of the logic area at the 50nm node is likely to be occupied by repeaters.

The area penalty due to repeater insertion is highly dependent on Rent's Exponent, p, as shown in Figure 2. The exponent, as described in Chapter 3, Section3, is an empirical description of an ICs wiring complexity and architecture. From Rent's Rule ($T=kN^{p}$), it can be inferred that when p=1, the all of the gate I/Os are dedicated to off-chip communications. For this extreme scenario, none of the gates or logic blocks is involved with inter-block communications. As such, no repeaters, *per se*, are necessary for insertion as no communication signals are available for repeatering. Hence, the number of repeaters required when p is relatively (~1) high is small. Of course, I/O drivers are still necessary to propagate the I/O signals all the way to the blocks. Conversely, considering the other extreme when p = 0 implies that the vast majority of interconnects are involved in inter-block communications, carrying signals across chip and will therefore require a large number of repeaters.

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.

5.1.2 Effect of second active layer for repeaters on interconnect delay

The effect of 3-D integration on interconnect delay across a number of technology nodes as projected by ITRS is summarized in Figure 3.14 of Chapter 3, Section 3.9. The effect of placing repeaters on a separate active layer of silicon on the interconnect delay can also be estimated and updated into the figure. By moving the repeaters from a 2-D configuration to a

second active layer in a 3-D configuration while maintaining all else constant results in a logic area reduction of approximately 10% for the 50nm technology node at p=0.65 as shown in Figure 2. Since logic occupies about half the footprint at the 50nm node, this results in a 5% chip area reduction in 3-D for the same operating frequency. For valid comparison with the data in Figure 3.14, and in accordance with the comparative procedure discussed in Chapter 3, the performance of this 3-D case (with repeaters on a second active layer) is increased such that the resulting footprint equals that in 2-D. Figure 3.14 is reproduced below in Figure 3 where the improvements in interconnect delay, reflecting the performance increase due to moving repeaters to a second active layer, are included. Figure 3 shows that the improvements in interconnect delay as a result of repeater displacement are approximately 9% at the 50 nm technology node.

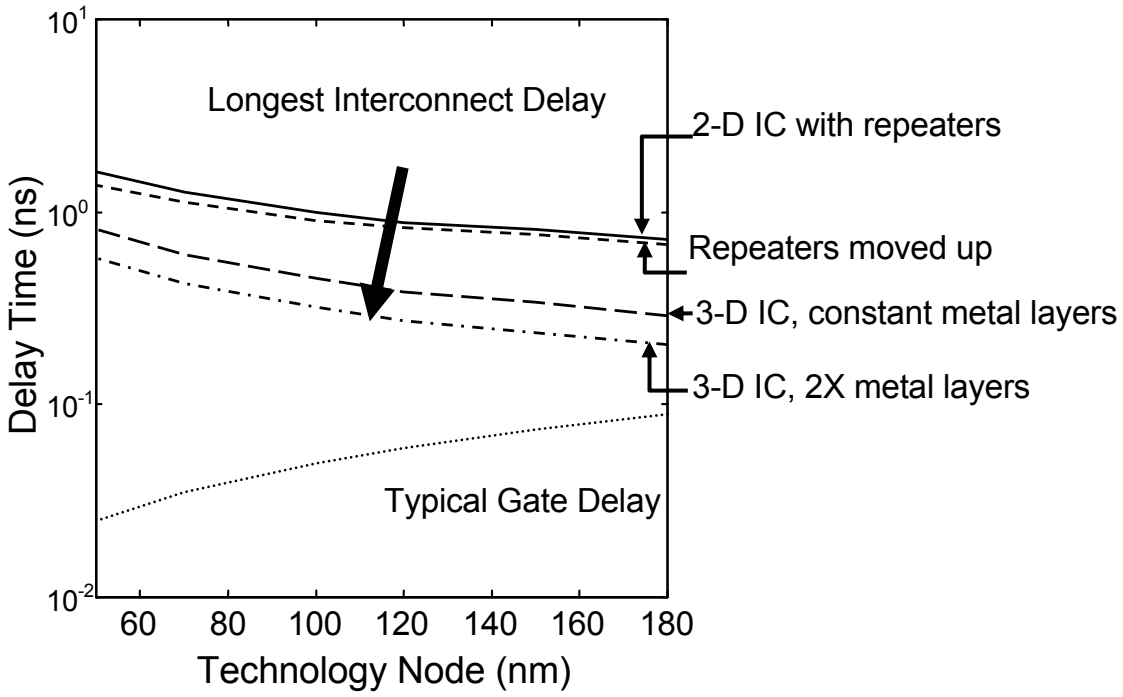


Figure 3: An updated summary of signal delay results from Fig. 3.14 for several cases covering a wide range of technology nodes as projected by ITRS 1999 [3].

5.2 Layout of Critical Paths

In typical high performance ASIC and microprocessor designs, interconnect delay is a significant portion of the overall path delay [86]. Logic blocks on a critical path communicating with other logic blocks, 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. Depending on the 3-D technology used (discussed in Chapter 7), the quality of higher layer devices may be considerably worse as compared to single crystal, bulk devices, which can render higher active layers incompatible with high performance logic. However, even if highest quality devices are not made on the second active layer, the decrease in interconnect delay can be more that the increase in gate delay due to sub-optimal transistor characteristics.

5.3 Microprocessor Design

In microprocessors and DSP processors, most of the critical paths involve on-chip caches [87]. The primary reason for this is that, in present technology, 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, as described in Chapter 4. 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 technology microprocessors, about half the physical area is projected to be consumed by on-chip caches at the 50nm technology node [3]. 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 and ignore design implications such as pipelining for simplicity). If on-chip caches are placed on the second active layer and the chip is resized accordingly to have dimensions $\frac{L}{\sqrt{2}} \times \frac{L}{\sqrt{2}}$, 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 Chapter 3, and does not perform any optimization of the interconnect pitch, it demonstrates that going from single silicon layer to two layers can result in nontrivial improvement in performance. Recent studies [88] 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 average time per instruction and increases system performance.

5.4 Mixed Signal Integrated Circuits and SoC

With greater emphasis on increasing the functionality that can be implemented on a single die in the system-on-a-chip (SoC) paradigm, more and more analog, mixed-signal and RF components of the system are being integrated on the same piece of silicon [77]. 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 from the substrate and degrade

the fidelity (or equivalently, increase the noise) of the signals present in these blocks [89]. 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 [31]. Moreover, since the second Si layer is not continuous, good isolation between different analog and RF components (such as the low-noise amplifier (LNA) and power amplifier) can also be achieved.

5.5 Optical Interconnects for System Clocking and I/O Connections

For high performance microprocessors with operating frequencies greater than a few GHz and large die sizes (on-chip frequency = 3 GHz, and die area = 8.17 cm^2 at the 50 nm technology node [3]), 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 microprocessors increases, synchronization of various blocks in the chip becomes increasingly difficult [90]. 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 [91, 91],

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 multi-GHz chips [93, 94]. 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, as discussed in the following chapter, with regards to metal interconnects. Hence optical interconnects are very attractive for large-scale synchronization of systems within multi-GHz ICs. Furthermore, optical interconnects employing short optical (laser) pulses, can reduce its optical power requirement [95]. 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 [96]. 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 [97, 98], a technique that allows multiple channels to be transmitted down a single waveguide.

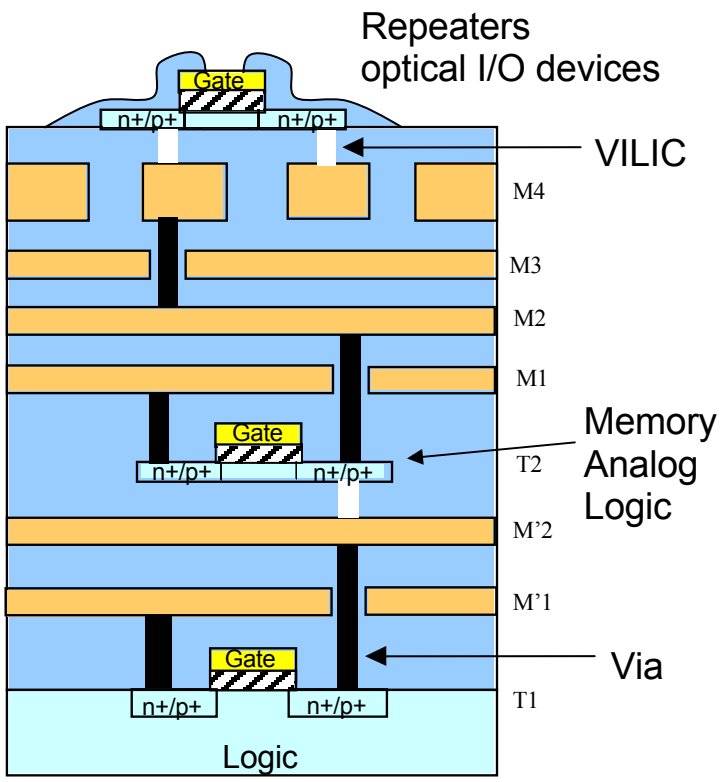


Figure 4: A schematic representation of 3-D integration of heterogeneous technologies, incorporating optical I/O, analog and logic layers.

Optical interconnect devices and networks integrated in a 3-D system-on-a-chip IC (schematically illustrated in Figure 4) can be employed to attain system synchronization and to enhance system performance. Furthermore, use of optical interconnects for clock distribution can significantly alleviate the power dissipation problem in 3-D ICs [11], and hence reduce the cost per chip. Integrated 3-D optical devices have been demonstrated directly on top of active silicon CMOS circuits [79], [99-101]. Also, polysilicon based optical waveguides of submicron dimensions have been demonstrated for low loss optical signal propagation and power distribution [102].

5.6 Implications on VLSI Design and Synthesis

VLSI design and synthesis (both logic and physical) for large digital circuits and high-performance system-on-a-chip 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 SOI (silicon-on-insulator) type upper active layers, the design process will need to address issues specific to SOI technology to realize significant performance improvements [103, 104].

Chapter 6

CHALLENGES FOR 3-D INTEGRATION

The list of challenges facing 3-D integration is as inexhaustible as the list of benefits. These can include, but are not limited to, thermal management, device quality, electromagnetic interactions, technology, reliability and yield. This chapter discusses some of those issues and offers some possible solutions.

6.1 Thermal Management

A detailed discussion and analysis of the modeled thermal behavior of 3-D ICs was presented in Chapter 4. Although taking into account vias as efficient heat conduction paths in the analysis projects more realistic die temperatures, as compared to previous wok [34,70], the temperature rises can nonetheless be detrimental to device performance, for high performance configurations, unless suitable thermal management solutions become available.

Lower operating temperatures for 3-D ICs can be achieved by employing a cooling design similar to the one illustrated in Figure 1 [82] where a coolant (e.g. water) is pumped through microchannels etched at the back surface of a silicon substrate were used to achieve very low package thermal resistances of $0.009^{o}C/(Wcm^{-2})$. This represents a highly significant reduction from ITRS projection of package thermal resistance at the 50 nm technology node of $2.15°C/(Wcm^{-2})$. Recent extensions of this approach are targeting even lower thermal resistances using closed-loop two-phase systems with boiling convection in microchannels [83]. The geometry of the chip and the packaging layers for this cooling system are shown in Figure 1.

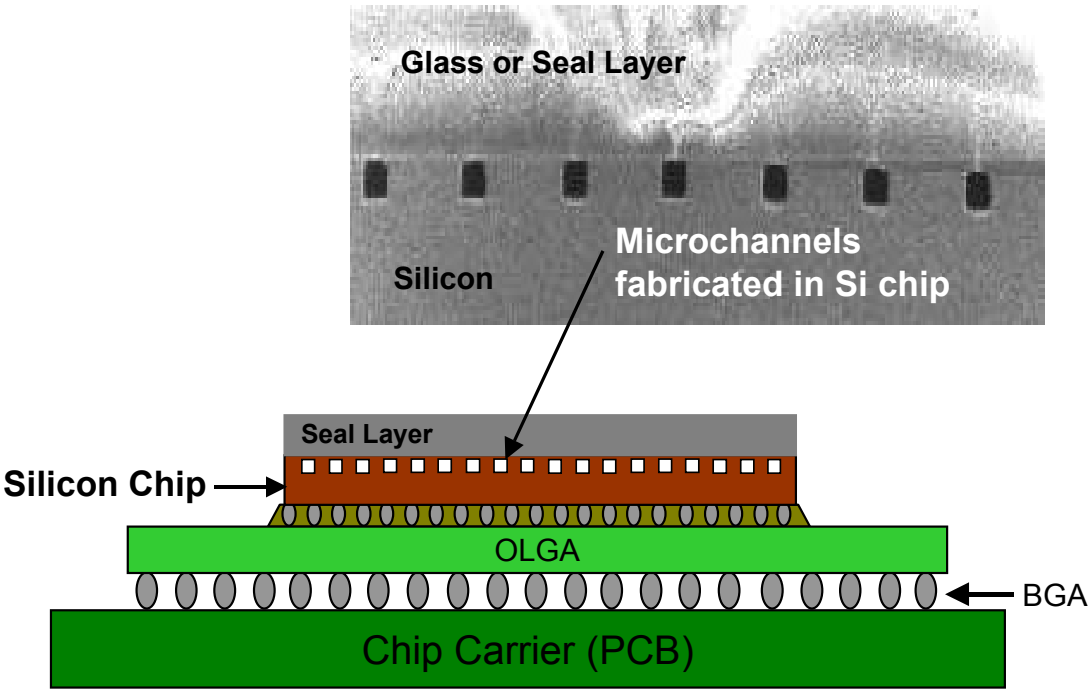


Figure 1: Schematic of a packaged Si chip with integrated microchannels etched in the substrate for pumping coolant to lower the package thermal resistance. BGA and OLGA denote ball grid array and organic layer ball grid array respectively. (*Courtesy of Kenneth E. Goodson, Stanford University*).

Using vias as efficient heat conduction paths, dummy thermal vias have been shown to be useful in reducing the temperature of interconnects in 2-D ICs [103]. A similar strategy can be employed for 3-D ICs, where inter-chip thermal vias that conduct heat but are electrically isolated can be distributed 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 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 of high-performance and high-power circuitry (even in the layer closest to the heatsink) would require careful thermal budgeting. Additionally, non-

uniform temperature distribution among the interconnects and devices in different actie layers can lead to performance mismatch and degradation as demonstrated for 2-D high-performance ICs [104,105].

6.2 Electromagnetic Interactions (EMI) in 3-D ICs

6.2.1 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 submicron technologies, the aspect ratio of global tier interconnects is ≥ 2.5 [3]. 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 [18].

6.2.2 Interconnect Inductance Effects

For deep submicron interconnects on-chip inductive effects arising due to increasing clock speeds and decreasing rise times are a concern for signal integrity and overall interconnect performance [84]. Inductance causes ringing in the signal waveforms, which can adversely affect signal integrity. For global wires inductance effects are more severe due to transmission line effects and also due to the lower resistance of these lines, which makes the wire impedance due to inductance comparable to that due to the resistance, and also due to the presence of significant mutual inductive coupling between wires resulting from longer current

return paths [85]. In 3-D ICs, the presence of a second substrate close to the global wires might help lower 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 the following chapter.

6.3 Reliability Issues in 3-D ICs

3-D ICs will most likely introduce some new reliability problems. These reliability issues may arise due to the electro-thermal and thermo-mechanical 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 [14]. There will be an increasing need to understand mechanical and thermal behavior of new material interfaces, thin-film material thermal and mechanical properties, and barrier 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 will affect the net yield of 3-D chips.

Chapter 7

3-D IC TECHNOLOGY

7.1 Technology Options

Although the concept of 3-D integration was demonstrated as early as in 1979 [108], and was followed by a number of reports on its fabrication process and device characteristics [20-26, 109-114], 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 100 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 [27-32] and their performance modeling [32-34], [116-117].

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 inter-layer dielectrics (ILDs) for 3-D circuit processing. A brief description of these alternatives is given below. 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 by the electrical characteristics of the fabricated devices as well as on the manufacturability and process compatibility with the relevant 2-D technology.

7.1.1 Beam Recrystallization

A 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)

as illustrated in Figure 1. MOS transistors fabricated on polysilicon exhibit very low surface mobility values (of the order of 10 cm^2/Vs), 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 re-crystallization of the polysilicon film [108-114], 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 temperatures involved during melting of the polysilicon and also due to difficulty in controlling the grain size variations [118, 119]. 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 [120], and even low-temperature single-crystal Si TFTs have been demonstrated [121] that can be employed to fabricate advanced 3-D circuits.

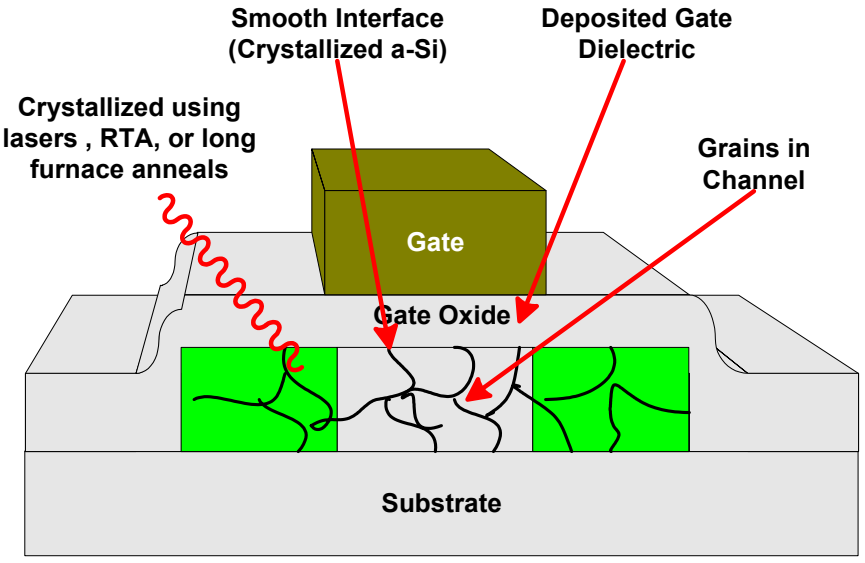


Figure 1: Schematic of a thin film transistor (TFT) fabricated on polysilicon depicting several grain boundaries in the active region.

7.1.2 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, as shown in Figure 2 [30]. 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 (~1000 ^{0}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 [122]. However, this process is not yet manufacturable.

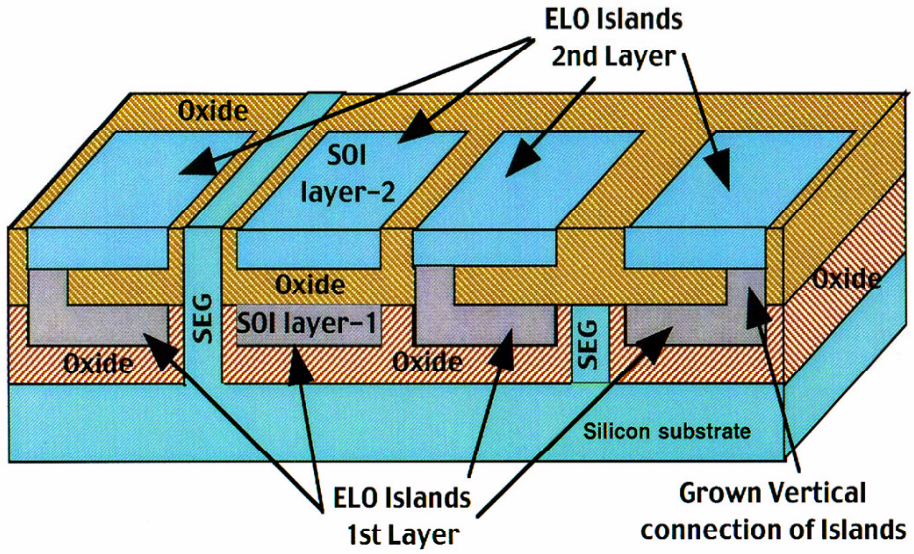


Figure 2: Schematic of an epitaxially grown second active layer. ELO denotes epitaxial layer overgrowth. (*Courtesy of G. W. Neudeck, Purdue University*).

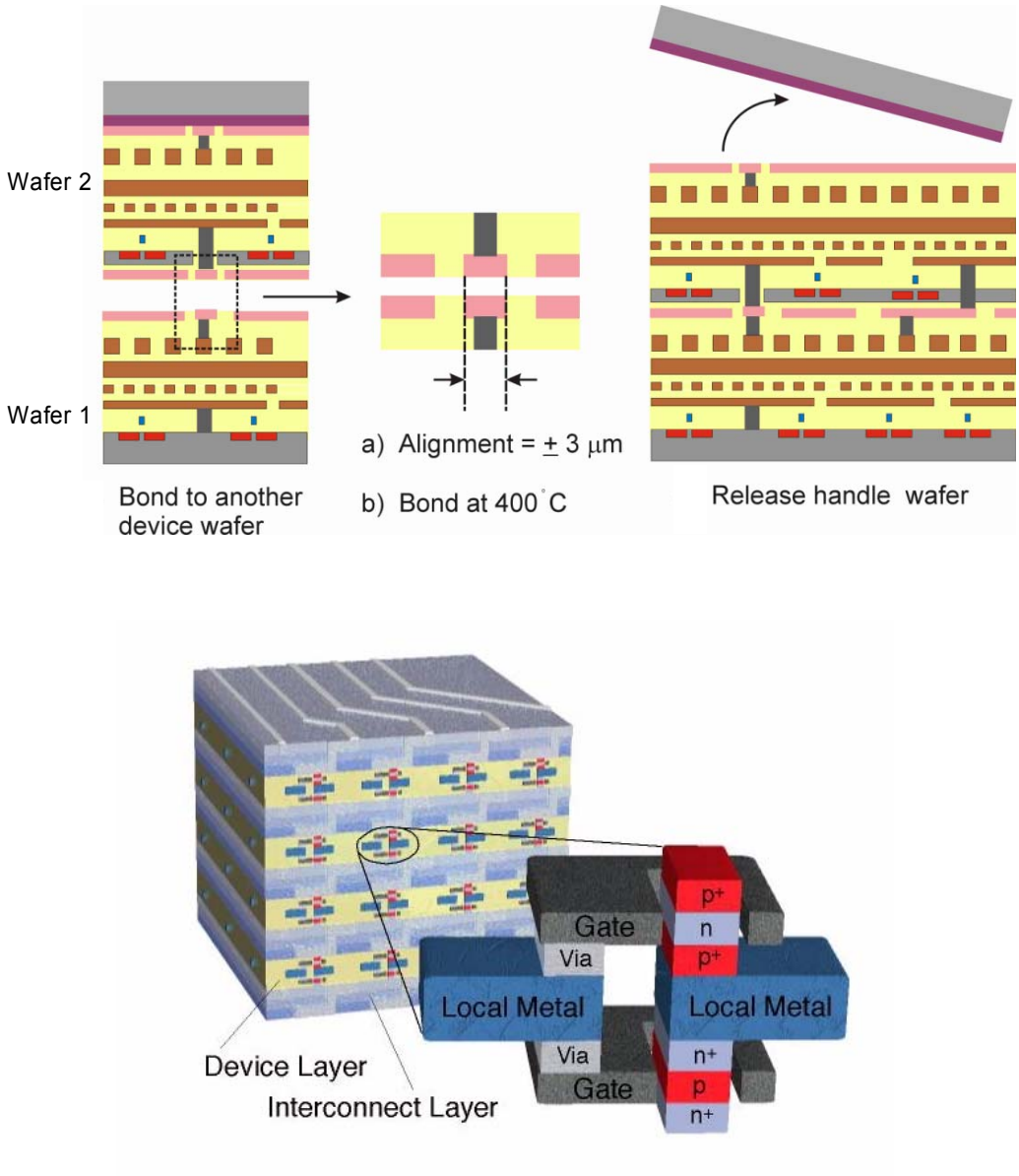


Figure 3: Schematic of steps used in one of the wafer bonding technologies based on metal thermo-compression (top) and a finished 3-D chip (bottom). *(Courtesy of Rafael Reif and Dimitri Antoniadis, Massachusetts Institute of Technology).*

7.1.3 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 (Figure 3) [28, 123]. Inter-chip vias are etched to electrically connect both wafers after metallization and prior to the bonding process at ~400 ^{0}C (discussed in Section 7.2 below). 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 +/- 2 μm) which restricts the inter-chip 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. Additionally, bonding techniques based on the thermo-compression of metal pads [123] offer low thermal-resistance interfaces between bonded wafers, which can help in heat dissipation.

7.1.4 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 [124-126]. 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 ^{0}C) such as using patterned seeding of Germanium (Fig. 4) [29]. 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. 5) [129-131], offers even lower thermal budget (< 500 ^{0}C) and can be employed to fabricate high-performance devices (MOSFETS or optical devices) on upper active layers even with metallization layers below.

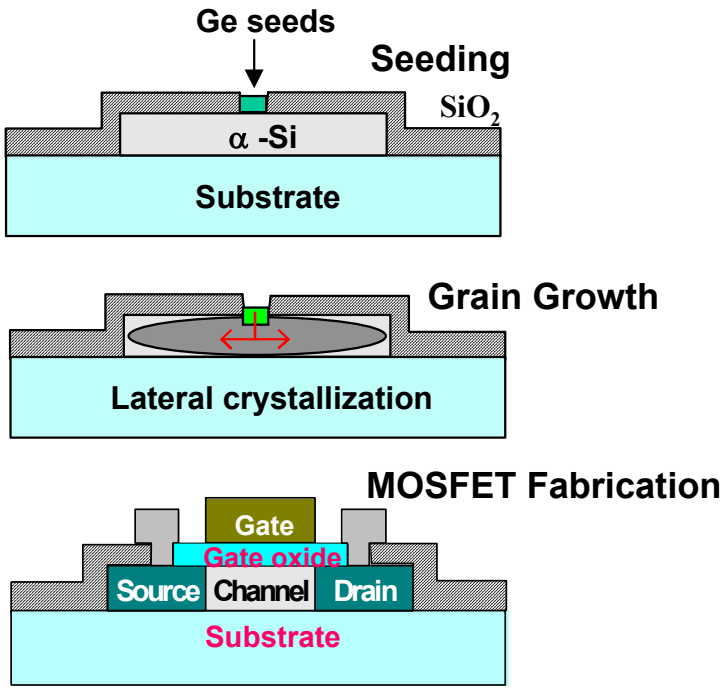


Figure 4: Schematic of the Ge seeded SPC fabrication steps.

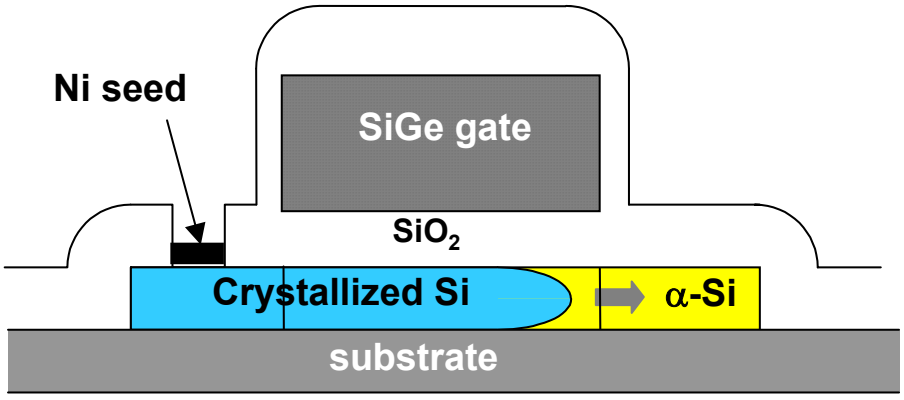


Figure 5: Schematic of the MILC process using Ni seeding.

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 [131]. It is found that the electrical characteristics of these devices (although superior among their peers) are still inferior to single crystal devices. However, technological advances to overcome the thermal budget problem have been made to allow fabrication of high-performance devices using SPC [132-134].

It is possible to conceive of several 3-D circuits for which SPC will be a suitable technology, such as in upper-level non-volatile memory, or by simply sizing up the upper level transistors to match their single crystal CMOS counterparts. For example, deep sub-micron polysilicon TFTs [135], stacked SRAM cells [136, 137], and EEPROM cells [138] 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.

7.2 Vertical Inter-Layer Interconnect Technology Options

The performance modeling presented in this thesis 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 ^{0}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 has higher resistivity. Current via technology can also be employed to achieve VILIC functionality. The underlying assumption here requires that intra-layer gates are interconnected using regular horizontal metal wires and inter-layer interconnects can be vias connecting the wiring network for each layer.

Recently, inter-layer (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 (~ 10 μm) top wafer to a bottom wafer with an organic adhesive layer of polyimide (~ 2 μm) in between [139]. Inter-chip 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, as illustrated in Figure 6a. The inter-chip 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 ≤ 1 μm) is done using a flip-chip bonder with split beam optics at a temperature of 400 ^{0}C.

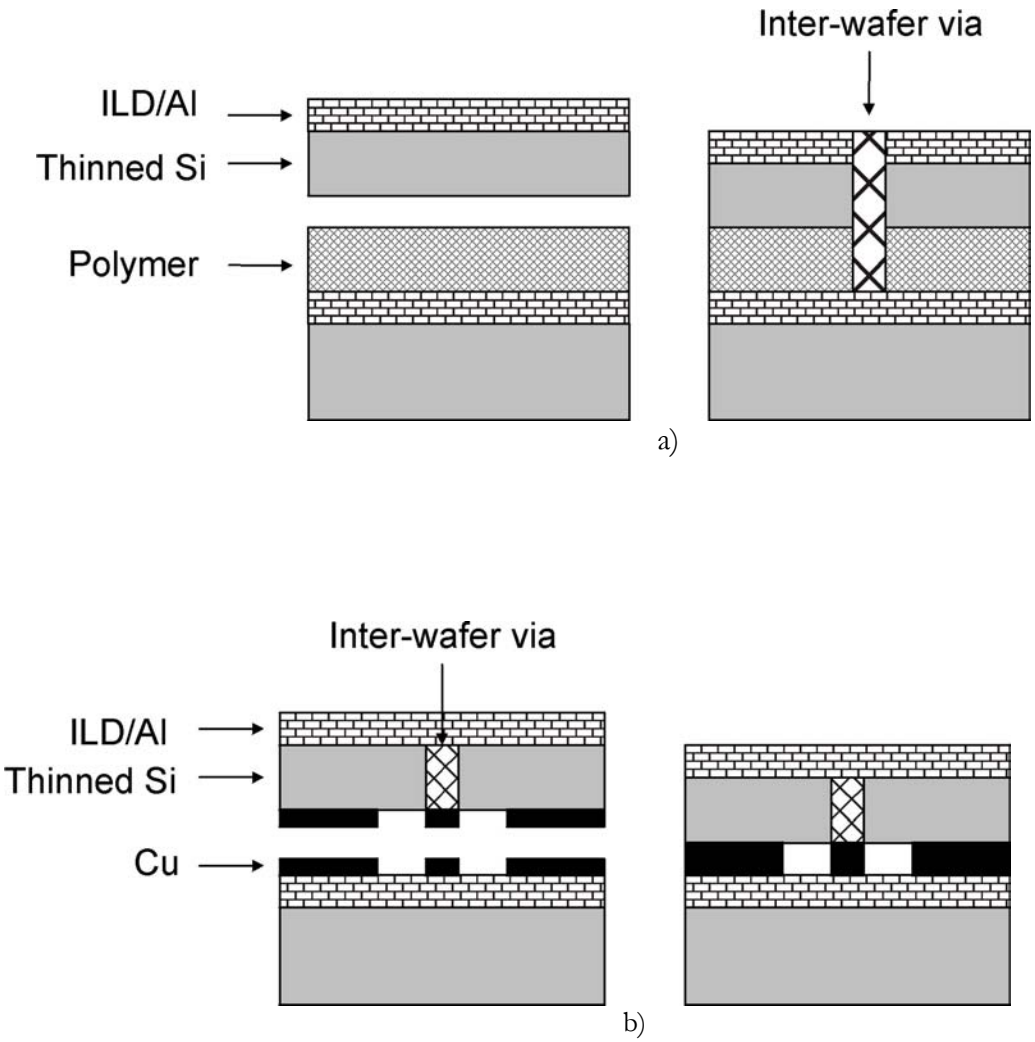


Figure 6: Schematic of the wafer bonding techniques a) with adhesive layer of polymer in between, and b) through thermocompression of Cu metal. *(Courtesy of Rafael Reif, Massachusetts Institute of Technology).*

A second technique relies on the thermocompression bonding between metal pads in each wafer [123]. In this method Cu/Ta pads on both wafers (illustrated in Figure 6b) serve as electrical contacts between the inter-chip 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 bi-layer pads with a combined thickness of 700 nm are fused together by applying a compressive force at 400 ^{0}C. This technique offers the advantage of a metal-metal interface that will lower the interface thermal resistance between the two wafers (hence provide better heat conduction) and can be beneficial as a partial ground plane for lowering the electromagnetic effects discussed in the previous chapter.

Chapter 8

CONCLUSIONS

Interconnect systems of present architecture and design are rapidly dominating chip size, IC performance and power dissipation. Although a number of remedies are incorporated to alleviate the interconnect delay problem, such as hierarchical wiring, introduction of Cu interconnects and low-k dielectrics for ILD materials, all have their limitations. ITRS projects that high-performance ICs will become interconnect performance limited at and beyond the 100nm technology node unless technological breakthroughs are introduced to alleviate the problem [3].

3-D integration offers the potential for such a breakthrough. Although the concept of 3-D is not new, the major contribution of this work is the development of quantitative systems level analysis of wire-pitch limited ICs in a 3-D design. Among other contributions of this work is the adaptation and application, for the first time, of Rent's Rule to a 3-D network of logic; incorporation of memory into the performance analysis; investigating functional separation of active layers by considering separate active layers for repeaters; and, through a joint contribution with T. Y. Chiang, the development of a realistic thermal analysis of 3-D ICs incorporating the effect of vias as efficient heat conduction paths on the thermal behavior.

3-D integration offers a number of advantages: it reduces the chip area due to the replacement of lateral wires with VILICs; it improves the performance of ICs by allowing larger wire pitches due to a reduced wiring requirement; and it offers the potential for the integration of heterogeneous technologies paving the way towards true SoC designs. At the

same time, 3-D integration poses several challenges that include, but not limited to, potentially unacceptable thermal behavior due to higher performance, technological challenges of fabricating such systems and impact on yield and reliability.

All these issues have been discussed in detail in the preceding chapters. The interconnect delay problem associated with Cu/low-k technologies was discussed using estimated RC values based on the data from ITRS 1999 in Chapter 2. The implications of material effects arising at deep submicron dimensions such as increasing metal resistivity due to increased electron scattering and finite metal barrier layer thickness were described. 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 solve the interconnect problem and that the design based solutions have not been adequate to deal with the wiring problem.

A methodology was presented in Chapter 3 to obtain the wire length distributions for 3-D ICs which can be used to estimate chip area, delay, and power dissipation, and provide examples of some of these trade-offs 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. The analysis predicts significant performance improvements of up to 100% over the 2-D case. The primary target technology for this analysis has been the ITRS '99 based 50 nm node with two active layers of silicon. Other

technology nodes with two active layers have also been 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, some of the thermal behavior concerns associated with 3-D circuits were addressed in Chapter 4. A compact analytical thermal model for temperature rise in 3-D ICs was introduced, which was used to estimate the temperature of individual active layers. It was demonstrated that for circuits with two silicon layers running at maximum performance, maintenance of acceptable die temperatures may require advanced heat-sinking technologies.

Furthermore, some scenarios in current and future VLSI and systems-on-chip type applications were discussed in Chapter 5 involving mixed signals and technologies, where the use of 3-D circuits will have an immediate and beneficial impact on performance. Briefly discussed were the implications of using this technology on the design process, as conventional VLSI design methodologies and tools, gate level and architecture level synthesis algorithms need to be suitably adapted. Also the effect of dedicating a separate active layer to repeaters on IC performance was quantified and presented.

Chapter 6 briefly discusses some of the challenges facing 3-D integration. Novel thermal management solutions are introduced and implications on reliability and electromagnetic interactions (such as capacitance and inductance effects) arising in 3-D ICs were also briefly discussed. Finally, an overview of some of the manufacturing technologies under investigation, which can be used to fabricate these circuits, was provided in Chapter 7.

Deep submicron 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 thesis, 3-D ICs are shown as 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. Several applications of 3-D ICs have been already been demonstrated in [59, 140-143], which show the potential of this technology for effective implementations of *System-on-a-Chip* 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 [144] can be precursors to the future monolithic 3-D ICs.

BIBLIOGRAPHY

1. M. T. Bohr, "Interconnect scaling-the real limiter to high performance ULSI," *IEDM Tech. Dig.*, 1995, pp. 241-244.

2. W. J. Dally, "Interconnect-limited VLSI architecture," *Int. Interconnect Technology Conf. Proceedings*, 1999, pp. 15-17.

3. *The International Technology Roadmap for Semiconductors* (ITRS), 1999.

4. M. T. Bohr and Y. A. El-Mansy, "Technology for advanced high-performance microprocessors," *IEEE Trans. Electron Devices,* vol 45, no. 3, pp. 620-625, 1998.

5. T. N. Theis, "The Future of Interconnection technology", *IBM Journal of R&D,* Vol. 44, 2000.

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

7. S. Venkatesan et al., "A high performance 1.8V, 0.20 μm CMOS technology with copper metallization," *IEDM Tech. Dig.*, 1997, pp. 769-772.

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

9. N. Rohrer et al., "A 480MHz RISC microprocessor in a 0.12 μm L_{eff} CMOS technology with copper interconnects," *Int. Solid-State Circuits Conf., Tech. Digest*, 1998, pp. 240-241.

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

11. P. Kapur, Ph.D. Thesis, Stanford University, 2002.

12. F. Chen and D. Gardner, "Influence of line dimensions on the resistance of Cu interconnections," *IEEE Electron Device Letters*, vol. 19, no.12, pp. 508-510, 1998.

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

14. S. Im and K. Banerjee, "Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs," *IEDM Tech. Dig.*, 2000.

15. K. Banerjee, Ph.D. Thesis, University of California at Berkeley, 1999.

16. K. Banerjee, "Thermal effects in deep submicron VLSI interconnects," Tutorial Notes, *IEEE International Symposium on Quality Electronic Design*, March 20-22, 2000.

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

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

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

20. Akasaki, "Three-Dimensional IC Trends", *Proceedings of the IEEE*, Dec. 1986

21. Takahashi *et al.* "Three-dimensional memory module", *IEEE Trans. on Advanced Packaging*, Feb. 1998

22. Gann, "High density packaging of flash memory", *1998 Proceedings. Seventh Biennial IEEE*, 1998

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

24. S. Tatsuno, "Japan's push into creative semiconductor research: 3-dimension IC's," *Solid State Technology*, March 30, pp. 29-30, 1987.

25. 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," *IEDM Tech. Dig.*, 1987, pp. 111-114.

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

27. S. Strickland, et al., "VLSI design in the 3rd dimension," *INTEGRATION*, Elsevier Science, pp. 1-16, 1998.

28. D. Antoniadis, "3-dimensional 25 nm – scale CMOS technology," *Advanced Microelectronics Program Review Proceedings Book*, Sept. 1-2, Lexington, MA, 1998.

29. V. Subramanian and K. C. Saraswat, "High-performance germanium-seeded laterally crystallized TFT's for vertical device integration," *IEEE Trans. Electron Devices*, vol. 45, no. 9, pp. 1934-1939, 1998.

30. 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* 17(3), pp. 994-998, 1999.

31. S. J. Souri, K. Banerjee, A. Mehrotra, and K. C. Saraswat, "Multiple Si layer ICs: motivation, performance analysis, and design implications," *Proc. 37th ACM Design Automation Conf.*, 2000, pp. 873-880.

32. K. C. Saraswat, S. J. Souri, V. Subramanian, A. R. Joshi, and A. W. Wang, "Novel 3-D Structures," *IEEE Int. SOI Conf.*, 1999, pp. 54-55.

33. S. J. Souri and K. C. Saraswat, "Interconnect performance modeling for 3D integrated circuits with multiple Si layers," *Int. Interconnect Technology Conf. Proceedings*, 1999, pp. 24-26.

34. K. Banerjee, S. Souri, P. Kapur, K. Saraswat, "3-D ICs: A Novel Chip Design for Improving Deep Submicron Interconnect Performance and Systems-on-Chip Integration". *Proceedings of the IEEE*, May 2001.

35. C. R. Barrett, "Microprocessor evolution and technology impact," *Symp. VLSI Technol., Digest*, 1993, pp. 7-10.

36. C. Hu, "MOSFET scaling in the next decade and beyond," *Semiconductor International*, pp. 105-114, 1994.

37. B. Davari, R. H. Dennard, and G. G. Shahidi, "CMOS scaling for high performance and low power-The next ten years," *Proc. of the IEEE*, vol. 83, no. 4, pp. 595-606, 1995.

38. G. A. Sai-Halasz, "Performance trends in high-end processors," *Proc. of the IEEE*, vol. 83, no. 1, pp. 20-36, 1995.

39. 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.

40. J. D. Meindl, "Low power microelectronics: retrospect and prospect," *Proc. of the IEEE*, vol. 83, no. 4, pp. 619-635, 1995.

41. S-Y Oh and K-J Chang, "2001 needs for multi-level interconnect technology," *Circuits and Devices*, pp. 16-21, 1995.

42. K. Yang, S. Sidiropoulos, M. Horowitz, "The Limits of Electrical Signalling," *Symposium of High Performance Interconnects*, *Hot Interconnects V*, IEEE Computer Society, August 1997.

43. 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 Journal of Research and Development*, vol. 39, no. 4, pp. 419-435, July 1995.

44. M. J. Hampden-Smith, and T. T. Kodas, "Copper etching: new chemical approaches," *MRS Bulletin*, vol. 18, no. 6, p. 39, June 1993.

45. S. P. Murarka, J. Steigerwald, and R. J. Gutmann, "Inlaid copper multilevel interconnections using planarization by chemical-mechanical polishing," *MRS Bulletin*, vol. 18, no. 6, pp. 45-51, June 1993.

46. S.-Q. Wang, "Barriers against copper diffusion into silicon and drift through silicon dioxide," *MRS Bulletin*, vol. 19, no. 8, pp. 30-40, August 1994.

47. R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd edition, John Wiley and Sons, New York, NY, pp. 1-56, 1986.

48. K. A. Perry, "Chemical mechanical polishing: the impact of a new technology on an industry," *Symp. On VLSI Technol*ogy, Tech. Dig., 1998, pp. 2-5.

49. 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," *38th IEEE Annual International Reliability Physics Symposium Proceedings*, 2000, pp. 354-358.

50. *Private Communications*, Lukas P.P.P. Van Ginneken, Magma Design Automation, Cupertino, CA.

51. S. Devadas, A. Ghosh, and K. Keutzer, *Logic Synthesis*, McGraw Hill, 1994.

52. D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron," *Proc. Int. Conf. on Computer Aided Design*, 1998, pp. 203-211.

53. D. Sylvester and K. Keutzer, "A global wiring paradigm for deep submicron design," *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, vol. 19, no. 2, pp. 242-252, 2000.

54. R. H.J.M. Otten and R. K. Brayton, "Planning for performance," *Proc. 35th Annual Design Automation Conference*, 1998, pp. 122-127.

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

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

57. 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," *Proc. 36th ACM Design Automation Conference,* 1999, pp. 491-496.

58. J.P. McVittie *et al.*, SPEEDIE 3.5 Manual, Stanford University, 1998.

59. T. H. Lee, "A Vertical Leap for Microchips", *Scientific American*, Jan. 2002.

60. 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, no. 3, March 1998.

61. 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, no. 3, March 1998.

62. Bakoglu, H.B., *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley Company: Reading, Mass., 1990.

63. B. S. Landman, and R. L. Russo, "On a pin versus block relationship for partitions of logic graphs," *IEEE Trans. Computers,* vol. C-20, no. 12, Dec. 1971.

64. F. Pollack, "New Challenges in Microarchitecture and Compiler Design", *PACT 2000*, Oct 2000.

65. K. E. Goodson and Y. S. Ju, "Heat conduction in novel electronic films," *Annu. Rev. Mater. Sci.*, 29: pp. 261-293, 1999.

66. T-Y, Chiang *et al.*, "A new analytical thermal model for multilevel ULSI interconnects incorporating via effect", *IITC*, 2001.

67. T-Y, Chiang *et al.*, "Impact of vias on the thermal effect of deep sub-micron Cu/low-k interconnects", *VLSI Tech. Symp.*, 2001. pp. 141-142

68. T-Y, Chiang *et al.*, "Thermal analysis of heterogeneous 3D ICs with various integration scenarios", *Electron Devices Meeting, 2001. IEDM Technical Digest.* 2001

69. T-Y, Chiang *et al.*, "Analytical thermal model for multilevel VLSI interconnects incorporating via effect", *IEEE Electron Device Letters* , Volume: 23 Issue: 1 , Jan. 2002

70. S. Im and K. Banerjee, "Full Chip Thermal Analysis of Planar (2-D) and Vertically Integrated (3-D) High Performance ICs," *Technical Digest IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, December 11-13, 2000, pp. 727-730.

71. D. Sylvester *et al.*, "Analytical modeling and characterization of deep-submicrometer interconnect", *Proc. IEEE,* pp. 634-664, May 2001.

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

73. *Private Communications,* A. Mehrotra,

74. *The Use of Thin Films in Physical Investigation*, Ed. J.C Anderson, Academic Press, London, 1966.

75. *Handbook of Thin Film Technology*, Edited by L.I. Maissel and R. Glang, Chapter 13, McGraw Hill Book Company, 1970.

76. R. Ho, K. Mai, H. Kapadia, and M Horowitz, "Interconnect scaling implications for CAD," *Proc. Int. Conf. on Computer Aided Design*, 1999.

77. M. J. M. Pelgrom, "*System-On-Chip Concepts*," Chapter 11, *ULSI Devices*, Eds. C. Y. Chang and S. M. Sze, Wiley Inter-Science, 2000.

78. H. De Man, "System Design Challenges in the Post PC Era," Keynote Address Presentation Slides, *37th ACM Design Automation Conf.*, 2000.

79. A. V. Krishnamoorthy, et al., "3-D integration of MQW modulators over active submicron CMOS circuits: 375 Mb/s transimpedance receiver-transmitter circuit," *IEEE Photonics Technology Letters*, vol. 7, no. 11, pp. 1288-1290, 1995.

80. H. B. Bakoglu and J. D. Meindl, "Optimal interconnection circuits for VLSI," *IEEE Trans. Elec. Dev.* vol. 32, no. 5, pp. 903- 909, 1985.

81. 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," *IEDM Tech. Dig.*, 1996, pp. 65-68.

82. D. B. Tuckerman, R. F. W. Pease, "High-performance heat sinking for VLSI," *IEEE Electron Device Lett.*, vol. EDL-2, no.5, pp.126-129, 1981.

83. *Private Communication*, K. E. Goodson, Stanford University.

84. E. E. Davidson, B. D. McCredie, and W. V. Vilkelis, "Long lossy lines (L^3) and their impact upon chip performance," *IEEE Trans. Components, Packaging, and Manufacturing Technology-Part B*, vol. 20, no. 4, pp. 361-375, 1997.

85. A. Deutsch et al., "Frequency-dependent crosstalk simulation for on-chip interconnections," *IEEE Trans. Advanced Packaging*, vol. 22, no. 3, pp. 292-308, 1999.

86. P. D. Fisher, "Clock cycle estimation for future microprocessor generations," *Technical Report*, SEMATECH 1997.

87. D. Greenhill et al, "A 330 MHz 4-way superscalar microprocessor," *ISSCC, Digest of Tech. Papers*, 1997, pp. 166-167.

88. 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. Components, Packaging, and Manufacturing Technology-Part B,* vol. 19, no. 4, pp. 709-718, 1996.

89. B. Razavi, "Challenges and trends in RF design," *Proc. 9th Annual IEEE Int. ASIC Conf. and Exhibit*, 1996, pp. 81-86.

90. J. M. Rabaey, "*Digital Integrated Circuits: A Design Perspective*," Prentice Hall Inc., 1996.

91. H. Kawaguchi and T. Sakurai, "A reduced clock-swing flip-flop (RCSFF) for 63% power reduction," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 5, pp. 807-811, 1998.

92. T. Sakurai, "Design challenges for 0.1 um and beyond," *Proceedings of the ASP DAC*, 2000, pp. 553-558.

93. J. W. Goodwin, F. J. Leonberger, S. C. Kung, R. A. Athale, "Optical interconnections for VLSI systems," *Proceedings of the IEEE*, vol. 72, pp. 850-866, 1984.

94. D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," *To appear in Special Issue of Proceedings of the IEEE.*

95. 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.*, 30, pp. 1167-1171, 1994.

96. G. A. Keeler, B. E. Nelson, D. Agarwal, and D. A. B. Miller, "Skew and jitter removal using optical pulses for optical interconnection," *IEEE Photonics Letters*, 12, pp. 714-716, 2000.

97. E. A. De Souza, M. C. Nuss, W. H. Knox, and D. A. B. Miller, "Wavelength-division multiplexing with femtosecond pulses," *Optics Letters*, 20, pp. 1166-1168, 1995.

98. D. Agarwal, G. A. Keeler, B. E. Nelson, and D. A. B. Miller, "Wavelength division multiplexed optical interconnects using femtosecond optical pulses," *Proc. IEEE LEOS Annual Meeting*, 2, pp. 828-829, 1999.

99. K. W. Goossen, et al., "GaAs MQW modulators integrated with silicon CMOS," *IEEE Photonics Letter*, vol. 7, no. 4, pp. 360-362, 1995.

100. A. V. Krishnamoorthy and Keith W. Goossen, "Optoelectronic-VLSI: photonics integrated with VLSI circuits," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 4, no. 6, pp. 899-912, 1998.

101. T. K. Woodward and A. V. Krishnamoorthy, "1-Gb/s integrated optical detectors and receivers in commercial CMOS technologies," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 5, no. 2, pp. 146-156, 1999.

102. L. C. Kimerling, "Photons to the rescue: microelectronics becomes microphotonics," *The Electrochemical Society Interface*, pp. 28-31, Summer 2000.

103. 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," *Tech. Dig. IEEE International Electron Devices Meeting*, 2000, pp. 261-264.

104. A. H. Ajami *et al.*, "Effects of non-uniform substrate temperature on the clock signal integrity in high performance designs," *Custom Integrated Circuits Conference*, 2001.

105. A. H. Ajami *et al.*, "Non-uniform chip-temperature dependent signal integrity," *IEEE Symposium on VLSI Technology*, 2001.

106. C. T. Chuang, P. F. Lu, and C. J. Anderson, "SOI for digital CMOS VLSI: design considerations and advances," *Proc. IEEE*, vol. 86, no. 4, pp. 689-720, 1998.

107. D. Allen, D. Behrends, and B. Stanisic, "Converting a 64b PowerPC processor from CMOS bulk to SOI technology," *Proc. 36th ACM Design Automation Conference,* 1999, pp. 892-897.

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

109. J. P. Colinge and E. Demoulin, "ST-CMOS (Stacked Transistor CMOS): a double-poly-NMOS-compatible CMOS technology," *IEDM Tech. Dig.*, 1981, pp. 557-560.

110. 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," *IEDM Tech. Dig.*, 1981, pp. 554-556.

111. 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, no. 10, pp. 366-368, 1983.

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

113. M. Nakano, "3-D SOI/CMOS," *IEDM Tech. Dig.*, 1984, pp. 792-795.

114. 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, no. 3, pp. 193-195, 1986.

115. 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. Components, Packaging, and Manufacturing Technology-Part B,* vol. 19, no. 4, pp. 719-727, 1996.

116. A. Rahman, A. Fan, J. Chung, and R. Reif, "Wire-length distribution of three-dimensional integrated circuits," *Int. Interconnect Technology Conf. Proceedings*, 1999, pp. 233-235.

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

118. 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 Dev.* vol. 47, pp. 1035-1043, 2000.

119. V. Subramanian, M. Toita, N. R. Ibrahim, S. J. Souri and K. C. Saraswat," Low-leakage Germanium-seeded Laterally-crystallized Single-grain 100nm TFTs for Vertical Integration Applications," *IEEE Electron Dev. Lett.,* vol. 20, no. 7, pp. 341-343, 1999.

120. 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.

121. 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.

122. 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., Part 1,* vol.36, (no.7A), pp. 4278-4282, July 1997.

123. A. Fan, A. Rahman, and R. Reif, "Copper wafer bonding," *Electrochemical and Solid State Letters*, vol. 2(10), pp. 534-536, 1999.

124. T. Noguchi, "Appearance of single-crystalline properties in fine-patterned Si thin film transistors (TFTs) by solid phase crystallization (SPC)," *Jpn. J. Appl. Phys., Part 2,* no.11A, vol.32, pp. 1584-1587, Nov. 1993.

125. 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. Society for Information Display*, 1/2, pp. 203-209, 1993.

126. 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.*, 75(7), pp. 3235-3257, 1994.

127. 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. Electronic Materials*, vol. 28, no. 1, pp. 19-25, 1999.

128. J. A. Tsai, A. J. Tang, T. Noguchi, and R. Reif, "Effects of Ge on material and electrical properties of polycrystalline $Si_{1-x}Ge_x$ for thin film transistors," *J. Electrochem. Soc.*, vol. 142, no. 9, pp. 3220-3225, 1995.

129. 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.

130. 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., Part 1,* pp. 7193-7197, Dec. 1998.

131. A. R. Joshi and K. C. Saraswat, "Sub-micron thin film transistors with metal induced lateral crystallization," Abstract no. 1358, *Proc. 196th Meeting of the Electrochemical Society*, Honolulu, HI, 1999.

132. J. Nakata and K. Kajiyama, "Novel low-temperature recrystalization of amorphous silicon by high energy beam," *Appl. Phys. Lett.*, pp. 686-688, 1982.

133. 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, no. 1, pp. 2-4, 1999.

134. 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.

135. R K. Watts and J. T. C. Lee, "Tenth-micron polysilicon thin-film transistors," *IEEE Electron Device Lett.,* vol. 14, no. 11, pp. 515-517, 1993.

136. 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, no. 5, pp. 233-235, 1991.

137. T. Yamanaka et al., "Advanced TFT SRAM cell technology using a phase-shift lithography," *IEEE Trans. Electron Devices,* vol. 42, no. 7, pp. 1305-1312, 1995.

138. 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, no. 8, pp. 304-306, 1994.

139. P. Ramm *et al.*, "Three dimensional metallization for vertically integrated circuits," *Microelectronic Engineering*, 37/38 pp. 39-47, 1997.

140. H. Kurino *et al.*, "Intelligent image sensor chip with three dimensional structure," *IEDM Technical Digest*, 1999, pp. 879-882.

141. K-W. Lee *et al.*, "Three-dimensional shared memory fabricated using wafer stacking technology," *IEDM Technical Digest*, 2000, pp. 165-168.

142. J. Burns *et al.*, "Three-dimensional integrated circuits for low-power, high-bandwidth systems on a chip," *ISSCC Digest of Technical Papers*, 2001, pp. 268-269.

143. M. Koyanagi *et al.*, "Neuromorphic vision chip fabricated using three-dimensional integration technology," *ISSCC Digest of Technical Papers*, 2001, pp. 270-271.

144. K. Ohsawa *et al.*, "3-D assembly interposer technology for next-generation integrated systems," *ISSCC Digest of Technical Papers*, 2001, pp. 272-273.