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

Kaustav Banerjee, Member, IEEE, Shukri J. Souri, Pawan Kapur, and Krishna C. Saraswat, Fellow, IEEE

Center for Integrated Systems, Stanford University, Stanford, CA, 94305.
kaustav@ee.stanford.edu

Abstract

Performance of deep submicron VLSI is being increasingly dominated by the interconnects due to decreasing wire pitch and increasing die size. Additionally, heterogeneous integration of different technologies in one single chip is becoming increasingly desirable, for which planar (2-D) ICs may not be suitable. This paper 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 heterogeneous integration of technologies to realize a System-on-a-Chip (SoC) design. A comprehensive analytical treatment of these 3-D ICs has been presented and it has been shown that by simply dividing a planar chip into separate blocks, each occupying a separate physical level interconnected by short and vertical inter-layer interconnects (VILICs), significant improvement in performance and reduction in wire-limited chip area can be achieved, without the aid of any other circuit or design innovations. A scheme to optimize the interconnect distribution among different interconnect tiers is presented and the effect of transferring the repeaters to upper Si layers has been quantified in this analysis for a two-layer 3-D chip. Furthermore, one of the major concerns in 3-D ICs arising due to power dissipation problems has been analyzed and an analytical model has been presented to estimate the temperatures of the different active layers. It is demonstrated that advancement in heat sinking technology will be necessary in order to extract maximum performance from these chips. Implications of 3-D device architecture on several design issues have also been discussed with especial attention to SoC design strategies. Finally, some of the promising technologies for manufacturing 3-D ICs have been outlined.

Index Terms: 3-D ICs, heterogeneous integration, interconnect performance, optical I/Os, power dissipation, system interconnects, System-on-a-Chip design, VLSI design.
1 Motivation for 3-D ICs

The unprecedented growth of the computer and the information technology industry is demanding VLSI circuits with increasing functionality and performance at minimum cost and power dissipation. VLSI circuits are being aggressively scaled to meet this demand. This in turn has introduced some very serious problems for the semiconductor industry. Continuous scaling of VLSI circuits is reducing gate delays but rapidly increasing interconnect delays. The International Technology Roadmap for Semiconductors (ITRS) [1] predicts that, beyond the 130 nm technology node, performance improvement of advanced VLSI is likely to begin to saturate unless a paradigm shift from present IC architecture is introduced. Also, increasing interconnect loading affects the power consumption in high-performance chips. In fact, a significant fraction of the total chip power consumption can be due to the wiring network used for clock distribution, which is usually realized using long global wires. Additionally, interconnect scaling has significant implications for traditional computer-aided-design (CAD) methodologies and tools which are causing the design cycles to increase, thus increasing the time-to-market and the cost per chip function. Furthermore, increasing drive for the integration of disparate signals and technologies is introducing various system-on-a-chip (SoC) design concepts, for which existing planar (2-D) IC design may not be suitable.

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

1.1 Interconnect Limited VLSI Performance

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

Table 1. Optimal interconnect and inverter (FO4) delays at various technology nodes. Parameters necessary for the delay calculations are also shown.

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

Figure 2. 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.
1.1.1 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 Fig. 2(a). Fig. 2(b) shows an equivalent RC circuit for one segment of the system. \( V_o \) 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 \((V_o)\) is assumed to switch instantaneously when voltage at the input capacitor \((V_{st})\) reaches a fraction \( x \), \( 0 \leq x \leq 1 \) of the total swing. Hence the overall delay of one segment, \( \tau_0 \), is given by:

\[
\tau_0 = b(x)R_{tr}(C_L + C_P) + b(x)(eR_{tr} + rC_L)l + a(x)rc^2\quad (1)
\]

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 \( r_o \), \( c_o \) and \( c_p \) are the resistance, input and parasitic output capacitances of a minimum sized inverter respectively then \( R_{tr} \) can be written as \( r_o / s \) where \( s \) is the multiples of minimum sized inverters. Similarly \( C_p = s c_p \) and \( C_L = s c_o \). If the total interconnect length \( L \) is divided into \( n \) segments of length \( l = L/n \), then the overall delay, \( \tau_d \), is given by,

\[
\tau_d = n\tau = \frac{L}{l}b(x)r_0(c_0 + c_p) + b(x)cR_{tr} + rC_LL + a(x)rcLL\quad (2)
\]

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

\[
l_{opt} = \sqrt{\frac{b(x)r_0(c_0 + c_p)}{a(x)rc}}\quad (3)
\]

\[
s_{opt} = \sqrt{\frac{r_0c}{rc_0}}\quad (4)
\]

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. Therefor \( c_p = 3c_{NMOS} \), 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 \( c_o \). With these assumptions, The optimum values of \( l \) and \( s \) can be expressed as,

\[
l_{opt} = 3.24\sqrt{\frac{r_0c_{NMOS}}{rc}} \quad \text{and} \quad s_{opt} = 0.577\sqrt{\frac{r_0c}{rc_{NMOS}}}\]

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

\[
\tau_d = 3.24L\sqrt{0.4rct_{FO1}}\quad (5)
\]

where \( t_{FO1} = 6r_0c_{NMOS} \) 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 [13]. In terms of FO4, (5) can be approximately written as,

$$\tau_d = 2L \sqrt{0.4rc_{FO4}}$$  \hspace{1cm} (6)

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

$$t_{FO4} = 500L_{gate}$$  \hspace{1cm} (7)

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

1.1.2 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, $p_v$, is assumed to be equal to the product of the aspect ratio, $A.R.$, and $p_w$ and the wire height (thickness) is also assumed to be half the vertical pitch. $A$, and therefore, $r$ can then be expressed as:

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

$$r = 4 \frac{\rho}{A.R. p_w^2}$$  \hspace{1cm} (8)

1.1.3 Capacitance Calculations:

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

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

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

$$c = \left(\frac{4}{4 + A.R.}\right) \frac{\varepsilon_r}{A.R.}$$  \hspace{1cm} (9)

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

At 250 nm technology node, Copper (Cu) with low-k dielectric was introduced to alleviate the adverse effect of increasing interconnect delay [14-18]. However, as shown in Fig. 1, below 130 nm technology node, substantial interconnect delays will result inspite of introducing these new materials, which in turn will severely limit the chip performance. Further 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 the problem of long-lossy lines cannot be fixed by simply widening the metal lines and using thicker interlayer dielectric since this conventional solution will lead to a sharp increase in the number of metallization layers. Such an approach will increase the complexity, reliability, and cost, and will therefore be fundamentally incompatible with the industry trend of maximizing the number of chips per wafer, and 25% per year improvement in cost per chip function. Furthermore, with the aggressive scaling suggested by the ITRS [1], new physical and technological effects start dominating interconnect properties. It is imperative that these effects be accurately modeled, and incorporated in the wire performance and reliability analyses. The next three sub-sections provide quantitative analysis of the impact of these new effects, caused by scaling, on the resistivity of Cu interconnects.

Figure 3. 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.
Before proceeding with the analysis, it is important to understand the fundamental differences between the metallization processes for aluminum (Al) and Cu, as illustrated in Figure 4. For Al based interconnects [19], first a thin layer of barrier material, Titanium (Ti) or Titanium Nitride (TiN), is uniformly deposited (blanket deposition) on top of a dielectric layer. The barrier layer is used to prevent any interaction between Al and the Si substrate, such as junction spiking. It is also used as an adhesion and texture promoter for the Al layer. The barrier layer is followed by Al 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 [20]. Hence Cu films are deposited by the damascene process\(^2\) [21] illustrated in Fig. 4(b). In this process, first a trench is patterned in the dielectric layer. This is followed by a barrier deposition, which coats the three surfaces of the trench. The barrier material is usually a refractory metal such as Ti or Ta or their nitrides [23]. As discussed later, there are different barrier deposition technologies. The barrier layer is necessary since Cu has poor adhesion to most dielectrics and can drift very quickly through them under electric bias to cause metal to metal shorts and to reach the underlying Si substrate where they can diffuse very rapidly through Si interstitial sites and form deep level acceptors that

---

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

2 In practice, a dual-damascene processing scheme is employed where the via and the line are patterned sequentially and then filled with Copper in one step [22]. However, since vias are present only at limited number of positions, Fig. 4(b) is an accurate representation of the cross-section of Cu lines along most of the interconnect length.
can degrade device performance [24]. This is then followed by Cu deposition (usually by electroplating). Next, the unwanted Cu and barrier layers outside the trenches are removed using chemical-mechanical-polishing (CMP) [25]. Finally, a layer of silicon nitride is deposited which passivates the top surface of the Cu metal in the trenches. Hence, due to the requirement of the barrier metal, effective cross section of the Cu interconnects will be less than the drawn dimensions.

![Diagram](image)

(a) (b)

Figure 5. Illustration of a) diffuse and specular surface scattering and b) effective cross-section reduction of Copper interconnects due to barrier.

It is commonly believed that material resistivity for Cu would not change significantly for future interconnects [1]. However, because of an increasing dominance of electron scattering from the interfaces and because of a greater fraction of interconnect area being consumed by metal barrier in the future, (Fig. 5), the effective resistivity of Cu may rise significantly. In addition, the operational temperature of wires (~373 K) is higher than the room temperature (300 K) and can increase further due to self-heating caused by the flow of current [12], [26]. The increase in temperature, in turn, would also increase the wire resistivity. Above effects are, next, quantified and more realistic Cu resistivity trends are established.

### 1.2.1 Effect of Interconnect Dimensions on Cu Resistivity

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

\[
\frac{\rho_s}{\rho_o} = \frac{1}{1 - \frac{3(1 - P)\lambda_{mfp}}{2d} \int_{0}^{\infty} \left( \frac{1}{x^3} - \frac{3}{x^2} \right) \frac{1 - e^{-kx}}{1 - Pe^{-kx}} \, dx}
\]

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

### 1.2.2 Effect of Barrier Thickness on Cu Resistivity

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

$$\rho_b = \rho_o \frac{l}{1 - \frac{A_b}{A \cdot R \cdot \left(\frac{p_w}{2}\right)^2}} \quad (11)$$

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

### 1.2.3 Simulation of Surface Scattering and Barrier Thickness Effects on Cu Resistivity

The resistivities for ITRS dictated future interconnects are evaluated in the light of above effects. The methodology for extracting future realistic resistivities using various barrier deposition technologies, operating temperatures and $P$ values is as follows. SPEEDIE (Stanford Profile Emulator for Etching and Deposition in IC Engineering) [30] was used to simulate the barrier profile for different deposition technologies, which was then used to extract the area consumed by the barrier. The simulations were performed on dimensions specified in the ITRS '99. The deposition time in the simulator was varied for each of the simulated geometries to obtain two conditions corresponding to a 5 nm and 10 nm minimum barrier thickness, respectively. The actual minimum barrier thickness in the future would be dictated by the quality of the barrier.

![Figure 6](image)

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

The effects of various barrier deposition technologies such as atomic layer deposition (ALD), Ionized physical vapor deposition (IPVD), collimated physical vapor deposition (CPVD), and simple physical vapor deposition (PVD) are quantified in Figure 6. The value of $P$ was taken to be 0.5 [28], temperature was 100 °C and minimum barrier thickness was chosen to be 10 nm in this figure. The resistivity of aluminum interconnects calculated with these physical effects is also shown in Figure 6 to demonstrate the diminishing advantage of copper over aluminum, for future scaled dimensions. This occurs
primarily because unlike copper, aluminum interconnects does not require barrier on all four surfaces and because its intrinsically higher bulk resistivity compared to copper lends the surface scattering effect less important at comparable dimensions. It can also be observed that more conformal deposition technologies such as ALD lead to a much slower rise in resistivity in future as the barrier deposited using these technologies leads to a lesser barrier cross sectional area consumption.

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

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

Figure 9. Typical VLSI design process flow.
Above discussion quantitatively illustrates that in the near future the material resistivity of copper will rise to prohibitively high values even with the best available deposition and barrier technologies. At some point, local and semiglobal tier effective resistivity of copper could become higher than corresponding resistivity for aluminum for same ITRS dictated dimensions. This will make the interconnect delay even higher than that depicted in Fig. 1 where bulk resistivity was assumed. This calls for a pressing need to develop Cu technologies with smooth surfaces along the wire perimeter to maximize elastic scattering of electrons such that the value of \( P \) in (10) may nearly equal one. There is also an urgent need for the development of barrierless Cu technology and for lowering the operating wire temperature by going with higher thermal conductivity packaging materials and/or with a radically new chip cooling mechanism.

1.3 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 [31]. As shown in Fig. 9, typically the design process starts at the behavioral level, which consists of a description of the system and what it is supposed to do (usually in C++ or Java programming languages). This description is then transformed to a Register Transfer Level (RTL) description using either the VHDL or Verilog languages. This is then transformed to a logic level structural representation (a netlist consisting of logic gates, flip-flops, latches etc.) by a process called logic synthesis. Finally, a physical mask-level 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 [32]. 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. 1, for deep submicron 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.

It is instructive to note that the magnitude of the interconnect problem for future deep submicron ICs with greater than \( 10^8 \) gates (269 million, at the 50 nm node [1]) cannot be fully comprehended by analyzing the impact of scaling on module-level designs (with around 50K gates) using standard wire-load models for average-length interconnects. This type of analysis, which has led some researchers to claim that interconnect delay is not a problem [33], is not quite adequate for deep submicron 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 [34]. In order to mitigate the interconnect scaling problems some researchers have proposed combined wire planning and constant-delay synthesis [11], [35]. This methodology is also based on a block-based design where the inter-block wires are planned or constructed and the remaining wires are handled through the constant-delay synthesis [36] within the blocks. The difficulty with this
method is that if the blocks are sufficiently large then the timing convergence problem persists. In contrast, if they are allowed to remain relatively small such that the constant-delay synthesis with wire-load models works, then the number of such blocks becomes so large that the majority of the wiring will be global and the physical placement of these point-like blocks becomes absolutely critical to the overall wire planning quality, which represents a daunting physical design problem. Another work proposed an interconnect fabric based on a ground-signal-ground wire grid to make wire loads more predictable [37]. However, this technique results in significant area penalty.

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

Thus it can be concluded that the interconnect problem in deep 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 [39]. We need to think beyond the current paradigm of design architecture.

1.4 System-on-a-Chip Designs

System-on-a-chip (SoC) is a broad concept that refers to the integration of nearly all aspects of a system design on a single chip [40], [41]. These chips are often mixed-signal and/or mixed-technology designs, including such diverse combinations as embedded DRAM, high-performance and low-power logic, analog, RF, programmable platforms (software, FPGAs, Flash etc.), as schematically illustrated in Fig 10. They can also involve more esoteric technologies like 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.
1.5 3-D Architecture

3-D integration (schematically illustrated in Fig. 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 [42]. Additionally, 3-D ICs can be very effective vehicles for large-scale on-chip integration of different systems.

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

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

![Schematic of a 3-D chip showing integrated heterogeneous technologies.](image_url)

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

2 **Scope of This Study**

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

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

The results obtained in Section 3 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. In Section 4 important concerns in 3-D ICs such as power dissipation, have been analyzed. It is demonstrated that advancement in IC cooling technology will be necessary for maximizing 3-D circuit performance.

Throughout this work no differences were assumed in the performance or the properties of the individual devices on any layer. Also the treatment is independent of the 3-D technology used. However, even if the properties of the devices on the upper Si layers are different, these layers can be used for memory devices or repeaters. Some of these applications are discussed in Section 5. Finally, in Section 6, various technology options for fabricating 3-D ICs have been outlined. For simplicity, technology effects on metal wire resistivity discussed earlier in Section 1.2 are ignored in the proceeding analysis (for both 2-D and 3-D ICs), where bulk resistivity is assumed.

3 Area and Performance Estimation of 3-D ICs

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

\[
T = k N^p
\]

(12)

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

To illustrate the application of this methodology, a logic system can be considered, the complexity of which necessitates that the final chip area is determined by the wiring requirement. Such ICs are considered wire-pitch limited, which is assumed throughout this work and considered valid for high-performance ICs. The wiring network is assumed to be a distribution of connecting wires ranging from the very short (to connect closest neighbor logic gates, or 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 problem of estimating the chip performance is then reduced to one of estimating this interconnect wiring distribution from which it is possible to determine a chip area and thus performance. To determine all the shortest wires in a logic system, the recursive property of Rent’s Rule is used, where the logic system is divided into logic gates and Rent’s Rule is applied to the interconnects between closest neighbor gates. This determines the number of interconnections between the closest logic gates. The longer wires are similarly determined by clustering the logic gates in growing numbers until the longest
interconnects are found. A summary of this methodology is given below and more details can be found in [47].

3.1 2-D And 3-D Wire-Length Distributions

The wire length distribution can be described by \( i(l) \), an Interconnect Density Function (i.d.f.), or by \( I(l) \), the Cumulative Interconnect Distribution Function (c.i.d.f.) which gives the total number of interconnects that have length less than or equal to \( l \) (measured in gate pitches), and is defined as,

\[
I(l) = \int_0^l i(x)dx
\]  \hspace{1cm} (13)

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

![Block diagram](image)

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

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

\[
T_A + T_B + T_C = T_{A\rightarrow C} + T_{A\rightarrow B} + T_{B\rightarrow C} + T_{ABC}
\]  \hspace{1cm} (14)

where \( T_A \), \( T_B \), and \( T_C \) are the number of I/Os for blocks A, B, and C respectively. \( T_{A\rightarrow C} \), \( T_{A\rightarrow B} \), and \( T_{B\rightarrow C} \) are the numbers of I/Os between blocks A and C, blocks A and B, and between blocks B and C respectively. \( T_{ABC} \) represents the number of I/Os for the entire system comprising 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,

\[ T_{A \rightarrow B} = T_A + T_B - T_{AB} \quad (15) \]

\[ T_{B \rightarrow C} = T_B + T_C - T_{BC} \quad (16) \]

Substituting (15) and (16) in (14) gives,

\[ T_{A \rightarrow C} = T_{AB} + T_{BC} - T_B - T_{ABC} \quad (17) \]

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

\[ T_B = k \left( N_B \right)^p \quad (18) \]

\[ T_{AB} = k \left( N_A + N_B \right)^p \quad (19) \]

\[ T_{BC} = k \left( N_B + N_C \right)^p \quad (20) \]

\[ T_{ABC} = k \left( N_A + N_B + N_C \right)^p \quad (21) \]

where \( N = N_A + N_B + N_C \). Substituting (18)-(21) in (17) gives,

\[ T_{A \rightarrow C} = k \left[ \left( N_A + N_B \right)^p - \left( N_B \right)^p + \left( N_B + N_C \right)^p - (N_A + N_B + N_C)^p \right] \quad (22) \]

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

\[ I_{A \rightarrow C} = \alpha k \left( T_{A \rightarrow C} \right) \quad (23) \]

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

\[ \alpha = \frac{f.o.}{I + f.o.} \quad (24) \]

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

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

\[ I(l) = I_{total} P(l) \quad (25) \]

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

\[ I_{total} = \alpha k N \left( -N^{p-1} \right) \quad (26) \]

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

\[ P(l) = \frac{1}{2N \left( -N^{p-1} \right)} \Gamma \left( \frac{l^2 \frac{p-1}{p} - 1}{6} + 2 \sqrt{N} - \frac{l^2 p - 1}{2p - 1} - N \frac{l^2 p - 2}{(p - 1)} \right) \quad (27) \]

for \( I \leq l \leq \sqrt{N} \), and
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, we first analyze the case where \( n \) Silicon layers are available. The application to two-layer (\( n=2 \)) case is straightforward. An \( N \) gate IC design is divided into \( N/n \) gate blocks. It is assumed that the routing algorithm and overall logic style is the same for both layers. This ensures that Rent’s constant, \( k \), and Rent’s exponent, \( p \), are the same for both layers. Applying Rent’s rule to all the layers, we have,

\[
T = k N^p = \left( \sum_{i=1}^{n} T_i \right) - T_{\text{int}} = n k \left( \frac{N}{n} \right)^p - T_{\text{int}}
\]

(32)

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

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

and
\[ T_{\text{ext},i} = T_i - \frac{T_{\text{int}}}{n} = k n^{p-1} \left( \frac{N}{n} \right)^p \] (33)

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

\[ k_{\text{eff,int}} = k \left( 1 - n^{p-1} \right) \]
\[ k_{\text{eff,ext}} = k n^{p-1} \] (34)

where \( K_{\text{eff,int}} \) is the effective number of I/Os per gate used for connecting other gates on the same layer and \( K_{\text{eff,ext}} \) is the effective number of I/Os per gate used to connect to gates on other active layers.

Figure 14. Schematic to illustrate a) conservation of total number of external I/O ports for maintaining constant functionality of chip, and b) two-layer 3-D chip with long horizontal interconnects replaced by short and vertical (VILICs) interconnects.
Extending this analysis to 2-layer \((n=2)\) 3-D IC’s (Fig. 14a), we have,

\[
T = k N^p = T_1 + T_2 - T_{int} = 2k \left( \frac{N}{2} \right)^p - T_{int}
\]  

(35)

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

\[
k_{\text{eff, ext}} = k 2^{p-1} \quad \text{and} \quad k_{\text{eff, int}} = k \left( l - 2^{p-1} \right)
\]

(36)

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

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

![Interconnect Length Distribution](image)

Figure 15. Wire-length distributions for the 2-D and 3-D ICs shown in Figure 14. 3-D significantly reduces requirement for longest wires. Metal tiers determined by \(L_{\text{Local}}\) and \(L_{\text{Semi-global}}\) boundaries as explained in the text.
3.2 Estimating 2-D and 3-D Chip Area

The analyses described in this work are performed on integrated circuits that are wire-pitch limited in size. The area required by the wiring network in such ICs is assumed to be greater than the area required by the logic gates. For the purposes of minimizing silicon real estate and signal propagation delays, the wiring network is segmented into separate tiers that are physically fabricated in multiple layers. An interconnect tier is categorized by factors such as metal line pitch and cross-section, maximum allowable signal delay and communication mode (such as 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 16 shows a schematic of a 3-tier interconnect structure.

![Figure 16. Schematic of a three-tier interconnection structure.](image)

Using a three tier interconnection structure, the semi-global tier pitch that minimizes the wire limited chip area was determined. The maximum interconnect length on any given tier was determined by the interconnect delay criterion [47] (It is assumed $t_{\text{delay,max}} = 0.25T$ for 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 $t_{\text{delay}} = 0.9T$ [47].

The area of the chip is determined by the total wiring requirement. In terms of gate pitch, the total area required by the interconnect wiring can be expressed as:
\[ A_{\text{required}} = \sqrt{\frac{A_c}{N}} \left( p_{\text{loc}} L_{\text{total, loc}} + p_{\text{semi}} L_{\text{total, semi}} + p_{\text{glob}} L_{\text{total, glob}} \right) \]  

(37)

where \( A_c \) is the chip area, \( N \) is the number of gates, \( p_{\text{loc}} \) is the local pitch, \( p_{\text{semi}} \) is the semiglobal pitch, \( p_{\text{glob}} \) is the global pitch, \( L_{\text{total, loc}} \) is the total length of the local interconnects, \( L_{\text{total, semi}} \) is the total length of the semiglobal interconnects and \( L_{\text{total, glob}} \) 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 [see Fig. 15 where broken vertical lines define the boundaries]. Hence it follows that,

\[ L_{\text{total, loc}} = \chi \int l l(t) dl \]  

(38)

\[ L_{\text{total, semi}} = \chi \int \frac{l l(t) dl}{L_{\text{loc}}} \]  

(39)

\[ L_{\text{total, glob}} = \chi \int \frac{l l(t) dl}{L_{\text{semi}}} \]  

(40)

where \( \chi \) is a correction factor that converts the point-to-point interconnect length to wiring net length (using a linear net model, \( \chi = \frac{1}{f_{0.1} + 3} \)). The boundaries shown in Fig. 15 represent the length of the longest wire for each tier, \( L_{\text{loc}} \) for the local, \( L_{\text{semi}} \) for the semiglobal and \( L_{\text{glob}} \) for the global tier.

We now present a modified analysis in terms of FO4 delay, discussed in Section 1.1, to estimate optimal chip area. The main differences between this analysis and those in [42], and [46] arise from the fact that the delay used here is for an optimally buffered interconnect, given by equation (6), and has been expressed in terms of FO4 delay. By substituting (8) and (9) in (6), and using \( \tau_d = \frac{\beta}{f_c} \), the length of the longest wire, \( L \), and the pitch, \( p_w \), for an arbitrary tier are related by the following expression:

\[ \frac{\beta}{f_c} = \frac{\sqrt{0.4 \rho \varepsilon_r \varepsilon_0 \left( l + 4 A.R. t_{\text{FO4}}^2 \right)_{\text{FO4}}}}{A.R. p_w} \sqrt{\frac{A_c}{N} L} \]  

(41)

Where \( \beta \) 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, \( \rho \) is the resistivity of the metal, \( \varepsilon_0 \) is the permittivity of free space, \( \varepsilon_r \) is the relative permittivity of the dielectric material, \( p_w \) is the wire pitch, \( A.R. \) is the wiring level aspect ratio and \( t_{\text{FO4}} \) is the FO4 gate delay. Equation (41) can be re-arranged to solve for wire pitch or the length of the longest interconnect. The expressions for \( p_{\text{glob}} \), \( L_{\text{semi}} \) (which is a function of \( p_{\text{semi}} \)) and \( L_{\text{loc}} \) are given by,

\[ p_{\text{glob}} = \sqrt{\frac{A_c}{N} A.R._{\text{glob}}} f_c \frac{1}{B_{\text{glob}}} \sqrt{0.4 \rho \varepsilon_r \varepsilon_0 \left( l + 4 A.R._{\text{glob}}^2 t_{\text{FO4}}^2 \right)_{\text{FO4}}} \]  

(42)

\[ L_{\text{semi}} = \frac{\beta_{\text{semi}}}{f_c} p_{\text{semi}} A.R._{\text{semi}} \sqrt{\frac{N}{A_c}} \frac{1}{\sqrt{0.4 \rho \varepsilon_r \varepsilon_0 \left( l + 4 A.R._{\text{semi}}^2 t_{\text{FO4}}^2 \right)_{\text{FO4}}}} \]  

(43)

\[ L_{\text{local}} = \frac{\beta_{\text{local}}}{f_c} p_{\text{local}} A.R._{\text{local}} \sqrt{\frac{N}{A_c}} \frac{1}{\sqrt{0.4 \rho \varepsilon_r \varepsilon_0 \left( l + 4 A.R._{\text{local}}^2 t_{\text{FO4}}^2 \right)_{\text{FO4}}}} \]  

(44)

Here \( p_{\text{loc}} \) is assumed constant and equal to twice the minimum feature size. \( L_{\text{global}} \) is also assumed constant and equal to the chip die edge. Equation (43) for \( L_{\text{semi}} \) results in a non-unique set of possible solutions for \( A_c \) and \( p_{\text{semi}} \) which are determined numerically. The wire-pitch limited chip area \( (A_{\text{required}}) \) is calculated based on the condition that the total required wiring area \( (A_{\text{required}}) \) is equal to the total available area \( (A_{\text{available}}) \) in a multilevel network, hence it follows that,
\[ A_{\text{available}} = A_c e_w n_{\text{levels}} = A_{\text{required}} \] (45)

where \( e_w \) is the wiring efficiency factor that accounts for router efficiency and additional space needed for power and clock lines, and \( n_{\text{levels}} \) is the number of metal levels available for the multilevel network. For each possible solution of (43), new boundaries representing \( L_{\text{loc}} \) and \( L_{\text{semi}} \) are used with the wire-length distribution to find the new total area required by the interconnect wiring. From the total area required by the wiring, the chip area is estimated by dividing the interconnects among the required number of metal layers. The resulting chip areas are then plotted as a function of \( p_{\text{semi}} \) 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.

### 3.3 Two Active Layer 3-D Circuit Performance

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

#### 3.3.1 Chip Area Minimization with Fixed Interconnect Delay

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

<table>
<thead>
<tr>
<th>PHYSICAL PARAMETER</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Transistors, ( N )</td>
<td>7053 million</td>
</tr>
<tr>
<td>Rent’s Exponent, ( p )</td>
<td>0.6</td>
</tr>
<tr>
<td>Rent’s Coefficient, ( k )</td>
<td>4.0</td>
</tr>
<tr>
<td>Minimum Feature Size, ( F )</td>
<td>50 nm</td>
</tr>
<tr>
<td>Max number of wiring levels, ( n_{\text{max}} )</td>
<td>9</td>
</tr>
<tr>
<td>Metal Resistivity, Copper</td>
<td>( 1.673 \times 10^{-6} \text{ Ohm-cm} )</td>
</tr>
<tr>
<td>Dielectric Constant, Polymer</td>
<td>( \epsilon_r = 1.5 )</td>
</tr>
<tr>
<td>Wiring Efficiency Factor</td>
<td>0.4</td>
</tr>
</tbody>
</table>

The possible solutions for \( A_c \) and \( p_{\text{semi}} \) resulting from the numerical solution of Equation (43) are plotted for the high-performance IC ITRS 50 nm technology node in Figure 17 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 \( p_{\text{semi}} \) 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 \( p_{\text{semi}} \) increases, even longer wires can now satisfy the maximum delay requirement in the semi-global tier. This results in global wires to be re-routed 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 \( p_{\text{semi}} \) 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 and since these wires have larger cross-sections they have a greater area requirement. Under such circumstances, the global tier begins to dominate and determine the chip area.

The curve for the 3-D case has a minimum similar to the one obtained for the 2-D case. It can be observed that the minimum chip area for the 3-D case is \(~29\%\) smaller than that of the 2-D case. Moreover, since the total wiring requirement is reduced (as shown in Figure 15), the semi-global tier pitch is reduced for
the 3-D chip. 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 are 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.

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 29% 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.

Figure 17. Wire-limited chip area versus normalized semi-global pitch (semi-global pitch/local tier pitch) for 2-D and 3-D ICs at a fixed operating frequency of 3 GHz. As the normalized semi-global-pitch reduces, wires are rerouted to the global tiers, which have bigger pitch, and hence the chip area increases. Note that the estimated 2-D chip area of 8.17 cm² is also projected by ITRS for the 50 nm node. The number of metal layers for 2-D and 3-D ICs is nine (3 per tier).

### 3.3.2 Increasing Chip Area and Performance

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

Figure 18. 3-D chip operating frequency (performance) increases with increases in semi-global wiring pitch. Chip area also increases but remains below the 2-D chip area. If 3-D chip area is made equal to 2-

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

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

![Graph showing the relationship between chip area and frequency](image)

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

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

### 3.4 Effect of Increasing Number of Silicon Layers

3-D technologies providing more than two active layers have also been considered. As the number of Silicon layers increases beyond two, the assumption that all inter-layer interconnects (ILICs) are vertical and consume negligible area becomes less tenable. For this particular example it is assumed that 90% of all ILICs are horizontal (c.f. Table 3). The area used up by these horizontal ILICs can be estimated from their total length and pitch. As shown in Fig. 21, the decrease in interconnect delay becomes progressively smaller as the number of active layers increases. This is due to the fact that area required by ILICs begins to offset any area saving due to increasing the number of active layers.
Figure 20. As the chip size increases due to increasing wire pitch, interconnects are re-routed to higher tiers. The global tier becomes over-crowded for large chip areas and global pitch starts to decrease.

Figure 21. Interconnect delay normalized to single layer delay as a function of the number of active Si layers shown for 50 nm node. The VILICs are assumed to consume lateral area.
3.5 Effect of Increasing the Number of Metal Layers

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

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

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

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

![Fraction of $L_{semi-global}$ vs Chip Area](image)

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

4 Challenges for 3-D Integration

4.1 Thermal Issues in 3-D ICs

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

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

It is well known that most of the heat energy generated in integrated circuits arises due to transistor switching. This heat is typically conducted through the silicon substrate to the package and then to the ambient by a heat sink. With 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 than Si [52], [53]. 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 [38].

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

### 4.1.1 Package Thermal Resistance Model for 2-D and 3-D ICs

Figure 25 shows the total power dissipation \(P\) and chip area \(A\) for high-end microprocessors for various 2-D technology nodes based on [1]. It can be observed that as technology scaling continues, chip area and power dissipation is increasing. The relationship between the die temperature rise \(\Delta T_{\text{Die}}\) and \(P\) can be expressed as,

\[
\Delta T_{\text{Die}} = (T_{\text{Die}} - T_{\text{amb}}) = P \cdot R_\theta
\]  

(46)
Figure 25. Maximum power dissipaton and chip area in 2-D circuits as a function of technology node based on ITRS.

Figure 26. Schematic view of a) heat flow in 2-D circuits and b) equivalent thermal circuit. $T$ denotes temperature of different materials. $R_{Si}$ and $R_{pkg}$ are the thermal resistances of the Si and the package material respectively.
where $T_{amb}$ is the ambient temperature ($=25$ °C), and $R_\theta$ is the effective thermal resistance from the Si devices to the heat sink, and is mostly due to the package material between the Si and the heat sink. Neglecting interface resistances, $R_\theta$ can be expressed as,

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

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

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

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

### 4.1.2 Analytical Die Temperature Model for 3-D ICs

A simple analytical model is proposed to estimate the temperature rise in each active layer of 3-D chips. The temperature rise (above the ambient temperature) of the $j^{th}$ active layer in an $n$-layer 3-D chip, schematically shown in Fig. 27(a), can be expressed as,

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

where $n$ is the total number of active layers, $R_i$ represents the thermal resistance between the $i^{th}$ and the $(i-1)^{th}$ layers and $P_k$ is the power dissipation in the $k^{th}$ layer. Note that this model does not take into account interconnect Joule heating. Assuming identical power dissipation ($P$) in each layer and identical thermal resistances ($R$) between layers, the temperature rise of the uppermost ($n^{th}$) layer in an $n$-layer 3-D chip can be expressed as [50],

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

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

$$R_i = \frac{t_{Si,i-1}}{K_{Si}} + \frac{t_{pkg}}{K_{pkg}}$$

and

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

respectively.

Here, $t_{Si,i}$ is the thickness of the $i^{th}$ Si layer, and $t_{glue,i-1}$ and $t_{ins,i-1}$ are the thickness of the $(i-1)^{th}$ glue and insulator (Cu+ILD layer in Fig. 27 (a)) layers respectively. From (50) the temperature rise can be expected to increase linearly with power density and the square of the number of active layers, $n$. However, for all practical 3-D ICs, $R_i \gg R$, which gives rise to an approximately linear relationship between $\Delta T_n$ and $n$ as shown in Fig. 27(b). Equation (50) also suggests that for most 3-D ICs with $n \leq 5$, $R_i$ will dominate the temperature rise of any layer.
Figure 27. a) Schematic of an $n$-layer 3-D chip with a heat sink at the bottom. $P$ denotes the power dissipation in each layer. b) temperature increase as a function of $n$ and the power density in each layer.

For the two active layer ($n=2$) 3-D example used in our performance analysis earlier, the temperature of each of the layers ($j=1$ and $j=2$) can be expressed using (49) as,
\[
\Delta T_1 = \left( \frac{P_1 + P_2}{A} \right) R_J \text{ and } \\
\Delta T_2 = \left[ \frac{P_2}{A} R \right] + \left[ \frac{P_2}{A} R \right]
\]

where \( R_J = \left( \frac{t_{Si-I} + t_{pkg}}{K_{Si-I} + K_{pkg}} \right) \), which can be extracted from (48) assuming same packaging material for 2-D and 3-D chips. The temperature rise for the second active layer, \( \Delta T_2 \), can therefore be expressed as,

\[
\Delta T_2 = \Delta T_1 + \frac{P_2}{A} R
\]

where the second term on the right hand side represents the effective thermal resistance between active layer 2 and active layer 1. Since the 3-D chip area can be calculated from our model, and \( R_J \) remains constant for both 2-D and 3-D circuits assuming same packaging material for all the technology nodes, the temperature of each of the active layers can be calculated.

4.1.3 Comparison Between 2-D and 3-D ICs

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

Table 4. Comparison between 2-D and 3-D ICs at the 50 nm technology node. Parameters for two limiting cases of 3-D ICs have been shown, one with the same chip area as the 2-D IC and the other with the same operating frequency as the 2-D IC.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>2-D</th>
<th>3-D</th>
<th>3-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Layers</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>( f_c ) (MHz)</td>
<td>3000</td>
<td>3000</td>
<td>6000</td>
</tr>
<tr>
<td>Feature Size (nm)</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>( A_c ) (cm(^2))</td>
<td>8.17</td>
<td>5.80</td>
<td>8.17</td>
</tr>
<tr>
<td>( N_t ) (Millions) per Active Layer</td>
<td>7053</td>
<td>3526.5</td>
<td>3526.5</td>
</tr>
<tr>
<td>Gate Pitch (cm)</td>
<td>3.4E-5</td>
<td>4.06E-5</td>
<td>4.81E-5</td>
</tr>
<tr>
<td>( p_{local} ) (( \mu )m) / A.R.</td>
<td>0.1 / 2.1</td>
<td>0.1 / 2.1</td>
<td>0.1 / 2.1</td>
</tr>
<tr>
<td>( p_{semi} ) (( \mu )m) / A.R.</td>
<td>0.165 / 2.7</td>
<td>0.14 / 2.7</td>
<td>0.33 / 2.7</td>
</tr>
<tr>
<td>( p_{global} ) (( \mu )m) / A.R.</td>
<td>0.275 / 2.9</td>
<td>0.23 / 2.9</td>
<td>0.55 / 2.9</td>
</tr>
<tr>
<td>( L_{local} ) (gate pitches)</td>
<td>6190</td>
<td>5195</td>
<td>1313</td>
</tr>
<tr>
<td>( L_{semi} ) (gate pitches)</td>
<td>10324</td>
<td>6826</td>
<td>4380</td>
</tr>
<tr>
<td>( L_{global} ) (gate pitches)</td>
<td>83982</td>
<td>59384</td>
<td>59384</td>
</tr>
<tr>
<td>( C_{total} ) (per active layer) (( \mu )F)</td>
<td>6.1285</td>
<td>2.370</td>
<td>5.6257</td>
</tr>
<tr>
<td>Total Power Dissipation (W)</td>
<td>174</td>
<td>135</td>
<td>639</td>
</tr>
<tr>
<td>Power Density per Layer (W/mm(^2))</td>
<td>0.213</td>
<td>0.116</td>
<td>0.391</td>
</tr>
</tbody>
</table>
We now present a comparison between the 2-D and 3-D ICs with respect to their performance, chip area and power dissipation. Table 4 lists various parameters for a 2-D IC at the 50 nm technology node based on [1] and the performance analysis methodology presented in this paper. Corresponding parameters are also calculated for two limiting designs of a two-layer 3-D IC. In one case the 3-D IC is designed to have the same chip area as that for the 2-D case and in the second design both the 2-D and 3-D ICs have the same operating frequency. As mentioned earlier, the 3-D design with the same chip (footprint) area (8.17 cm²) as that for the 2-D case gives the maximum performance ($f_c = 6$ GHz). This design also gives the highest total interconnect network capacitance and the highest power density per layer. While the other 3-D design with the same operating frequency (3 GHz) as that for the 2-D IC gives the lowest total interconnect capacitance and the lowest power density per layer.

![Figure 28](image_url)

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

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

$$C_{total} = C_{local} + C_{semi} + C_{global}$$

(54)

By using the maximum allowable delay per tier criteria, described in Section 3.2, we calculate the longest wire on each tier. Also, as described in detail in Section 3.2, the area under an interconnect density function
plot (wire length distribution) can be used to calculate the total length of wire on each tier. The capacitance for each tier is then calculated using (9), where A.R. is the aspect ratio for that tier. The calculated capacitances for all the tiers are then summed to find the total interconnect capacitance.

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

\[ P = \frac{1}{2} \alpha C V_{dd}^2 f_c \]  

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

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

![Figure 29. Schematic of a packaged Si chip with integrated microchannels etched in the substrate for pumping coolant to lower the package thermal resistance. BGA and OLGA denote ball grid array and organic layer ball grid array respectively. (Courtesy of Kenneth E. Goodson, Stanford University).](image)
Note that in all calculations Joule heating of the interconnects has been ignored since most of the heat is dissipated by the transistors. However, interconnect Joule heating can increase the peak temperature in 3-D chips due to strong thermal coupling with the neighboring interconnects and the active layers giving rise to higher interconnect temperatures and hence higher interconnect resistance and also lower interconnect electromigration performance. In order to take these coupling effects into account full chip thermal analysis using finite element simulations are needed as shown in [50].

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

Dummy thermal vias have been recently shown to be useful in reducing the temperature of interconnects in 2-D ICs [56]. A similar strategy can be used for the 3-D ICs, where inter-chip thermal vias that conduct heat but are electrically isolated can be employed to alleviate the heat dissipation problem in high-performance 3-D ICs. Furthermore, it is important to realize that thermal problems in 3-D ICs will be less severe for applications that do not require integration of high-performance logic. For example integration of memory, analog or RF blocks or any other circuits that have much lower power dissipation compared to high-performance logic may not require costly packaging and cooling solutions. However, any 3-D integration involving high-performance logic (even in the layer closest to the heat sink) would require careful thermal budgeting for the upper layer circuits, which would certainly be affected by the power dissipation of the logic layer according to (53). Additionally, non-uniform temperature distribution among the interconnects and devices in different active layers can lead to performance mismatch and degradation as recently demonstrated for 2-D high performance ICs [57], [58].

4.2 Electromagnetic Interactions (EMI) in 3-D ICs

4.2.1 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 [1]. Therefore line-to-line capacitance is the dominant portion of the overall capacitance. Hence, the presence of an additional Silicon layer on top of a global metal line may not have an appreciable effect on the line capacitance per unit length. For technologies with very small aspect ratio, the change in interconnect capacitance due to the presence of an additional Silicon layer could be significant, as reported in [59].

4.2.2 Interconnect Inductance Effects

For deep submicron interconnects on-chip inductive effects arising due to increasing clock speeds, decreasing rise times and increasing length of on-chip interconnects is a concern for signal integrity and overall interconnect performance [60]. Inductance can increase the interconnect delay per unit length and can cause ringing in the signal waveforms, which can adversely affect signal integrity [61], [62]. For long global wires (such as clock lines) inductance effects are more severe due to the lower resistance of these lines, which makes the reactive component of the wire impedance to become comparable to the resistive component, and also due to the presence of significant mutual inductive coupling between wires resulting from longer current return paths [63]. For Cu based technologies line resistances have decreased further and as a result, inductive effects are expected to become more significant. In 3-D ICs, the reduction of wire lengths will certainly help reduce inductance. Additionally, the presence of a second substrate close to the global wires might help lowering the inductance by providing shorter return paths, provided the substrate resistance is sufficiently low or if the wafers are bonded through metal pads as discussed in section 6.2.

4.3 Reliability Issues in 3-D ICs

3-D ICs will possibly 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 [50]. Additionally, heterogeneous integration of technologies using 3-D architecture will increase the need to understand mechanical and thermal behavior of new material interfaces, thin-film-material thermal and mechanical properties, and barrier/glue layer integrity. Additionally, from a manufacturing point of view, there might be yield issues arising due to the mismatch between the individual die-yield maps of different active layers, which may affect the net yield of 3-D chips. Such issues would demand a careful tradeoff between system performance, cost and the 3-D manufacturing technology.

5 Implications for Circuit Design and System-on-a-Chip Applications

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

![Figure 30. Interconnect length boundaries for the local tier.](image)

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

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 inserter. Repeaters are assumed to be inserted along wires whose lengths exceed a certain critical length. This critical length is determined by the maximum allowable signal delay along the wire for each interconnect tier (as described in Section 3.2). To illustrate, the local tier cannot have any non-repeated lines that exceed a maximum allowable length, \( L_{\text{opt}} \) in (3). Any wires that are routed in the local tier whose length are required to be greater than \( L_{\text{opt}} \) must have repeaters inserted along their lengths in order to satisfy the maximum allowable signal delay for this tier. The maximum length of repeated interconnect wire in any given tier is not arbitrary. Repeated wires are assumed to have repeaters inserted optimally and the signal delay along such wires is given by (6). The maximum allowable length per interconnect tier is calculated.
based on (42), (43), and (44). As an example, a schematic figure describing the critical lengths for the local tier is given in Fig. 30.

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, $s_{opt} = \sqrt{\frac{r_o c}{3r c_{NMOS}}}$ (as described in Section 1.1).

To determine the total number of repeaters it is necessary to determine the number of interconnects that require repeater insertion. For this we make use of Rent’s Rule. As represented in Fig. 31, any given tier is divided into two regions. The central region of area $\pi L_{opt}^2$ 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, $P_I$, that the I/O of any gate within this area of $\pi L_{opt}^2$ reaches outside this area is given by,

$$P_I = \frac{k N_{central}^p}{k N_{central}^{p-l}} = N_{central}^{p-1}$$

(56)

Figure 31. Fraction of chip area used by repeaters for different technology nodes based on ITRS and different Rent’s exponents. As much as 27% of the chip area at 50 nm node is likely to be occupied by repeaters.
Assuming that the number of logic gates is related to the logic block area \((A)\) by some constant of proportionality, i.e., \(A = \pi L_{\text{opt}}^2 \propto N_{\text{central}}\), then \(P_1\) for the local tier can be written as:

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

where \(\kappa\) is a constant of proportionality. Similarly, the probability, \(P_2\), that the I/O of any gate within the local tier of area of \(\pi L_{\text{loc}}^2\) reaches outside this area is given by,

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

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

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

\[
P_{\text{loc}} = P_1 (1 - P_2)
\]

Hence, the number of interconnects, \(I_R\), that require repeater insertion for the local tier is simply the probability \(P_{\text{loc}}\) multiplied by the total number of I/O’s of all the gates:

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

The optimum number of repeaters per unit length of wire \((1/l_{\text{opt}})\) is given by \(0.4 \text{rc} / \sqrt{4.2 \tau_0 c_{\text{NMOS}}} \). To estimate the total number of repeaters an average length of wire, \(l_{\text{avg}}\), is considered, where:

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

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

\[
P_1 (1 - P_2) k \kappa l_{\text{avg}} \sqrt{\frac{0.4 \text{rc}}{4.2 \tau_0 c_{\text{NMOS}}} L_{\text{loc}}^2}
\]

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

\[
A_{R,\text{loc}} = P_1 (1 - P_2) k \kappa l_{\text{avg}} \sqrt{\frac{0.4 \text{rc}}{4.2 \tau_0 c_{\text{NMOS}}} L_{\text{loc}}^2} B_0 s_{\text{opt,loc}}
\]

Where \(B_0\) is the minimum repeater size \((\approx 60WL)\) and \(s_{\text{opt,loc}}\) is the optimum multiple of minimum repeater size for the local tier. All parameters in (63) can be calculated for a given technology node based on [1].

This procedure is repeated to account for all the interconnect tiers to estimate the total area, \(A_{R,\text{total}}\), utilized by repeaters, i.e,

\[
A_{R,\text{total}} = A_{R,\text{loc}} + A_{R,\text{semi}} + A_{R,\text{glob}}
\]

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

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

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

5.3 Microprocessor Design

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

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

5.4 Mixed Signal Integrated Circuits

With greater emphasis on increasing the functionality that can be implemented on a single die in the system-on-a-chip paradigm, more and more analog, mixed-signal and RF components of the system are being integrated on the same piece of Silicon (as illustrated in Fig. 10). However, this presents serious design issues since switching signals from the digital portions of the chip couple into the sensitive analog and RF circuit nodes from the substrate and degrade the fidelity (or equivalently, increase the noise) of the signals present in these blocks [68]. Furthermore, different fabrication technologies are required for the two applications. However, with the availability of multiple Silicon layers, RF and mixed signal portions of the system can be realized on a separate layer (using different technologies) thereby providing substrate isolation from the digital portion. A preliminary analysis shows a 30 dB improvement in isolation by moving the RF portions of the circuit to a separate substrate. Moreover, since the second Si layer 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 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 [1]), interconnects responsible for global communications, including the interconnect network used for the clock distribution, can contribute significantly to the key performance metrics (area, power dissipation, and delay) and to the overall cost of the chip. As the complexity (size) of the microprocessor increases, synchronization of various
blocks in the chip becomes increasingly difficult [69]. This occurs mainly due to the variation in the placement of different blocks (or clock line lengths) and due to differences in their operating temperature that affects the clock skew and the net signal delay. Additionally, data input and output (I/O) requirements drive up the number of I/O pads and the corresponding size of the I/O circuitry (or chip area). Furthermore, in high performance designs around 40-70% of the total power consumption could be due to the clock distribution network [70], [71], and as the total chip capacitance (dominated by interconnects) and the chip operating frequency increases with scaling, the power dissipation increases.

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

Optical interconnect devices and networks integrated in a 3-D system-on-a-chip IC (schematically illustrated in Fig. 11) can be employed to attain system synchronization and to enhance system performance. Integrated 3-D optical devices have been demonstrated directly on top of active silicon CMOS circuits [43], [78], [79], [80]. Also, polysilicon based optical waveguides of submicron dimensions have been demonstrated for low loss optical signal propagation and power distribution [81].

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 [82], [83].

6 Overview of 3-D IC Technology

6.1 Technology Options

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

Presently, there are several possible fabrication technologies that can be used to realize multiple layers of active-area (single crystal Si or recrystallized poly-Si) separated by 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.

6.1.1 Beam Recrystallization

A very popular method of fabricating a second active (Si) layer on top of an existing substrate (oxidized Si wafer) is to deposit polysilicon and fabricate thin film transistors (TFT) (see Fig. 32). MOS transistors fabricated on polysilicon exhibit very low surface mobility values (of the order of 10 cm²/V.s), and also have high threshold voltages (several volts) due to the high density of surface states (several 10¹² 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 [84-94], to reduce or even eliminate most of the grain boundaries. This technique however may not be very practical for 3-D devices because of the high temperature involved during melting of the polysilicon and also due to difficulty in controlling the grain size variations [105], [106]. Beam recrystallized polysilicon films can also suffer from lower carrier mobility (compared to single crystal Si) and unintentional impurity doping. However, high-performance TFTs fabricated using low temperature processing [107], and even low-temperature single-crystal Si TFTs have been demonstrated [108] that can be employed to fabricate advanced 3-D circuits.

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

6.1.2 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 (Fig. 33) [98]. In principle, the quality of devices fabricated on these epitaxial layers can be as good as those fabricated underneath on the seed wafer surface, since the grown layer is single crystal with few defects. However, the high temperatures (~1000 °C) involved in this process cause significant degradation in the quality of devices on lower layers. Also this technique cannot be used over metallization layers. Low temperature silicon epitaxy using ultra-high-vacuum chemical vapor
deposition (UHV-CVD) has been recently developed [109]. However, this process is not yet manufacturable.

### 6.1.3 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 (Fig. 34) [96], [110]. Interspace vias are etched to electrically connect both wafers after metallization and prior to the bonding process at ~400 °C (discussed in section 6.2 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 thermocompression of metal pads [110] offer low thermal-resistance interfaces between bonded wafers, which can help in heat dissipation.

### 6.1.4 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 [111-114]. Device performance can be enhanced by eliminating the grain boundaries in the polysilicon film. For this purpose, local crystallization can be induced using low temperature processes (< 600 °C) such as using patterned seeding of Germanium (Fig. 35) [97], [115]. In this method Ge seeds implanted in narrow patterns made on a-Si can be used to induce lateral crystallization and inhibit additional nucleation. This results in the formation of small islands, which are nearly single crystal. CMOS transistors can then be fabricated within these islands to give SOI like performance. Another approach based on the seeding technique employs metal (Ni) seeding to induce simultaneous lateral recrystallization and dopant activation after the fabrication of the entire transistor on an a-Si layer. This technique known as the Metal Induced Lateral Crystallization (MILC) (see Fig. 36) [116], [117] offers even lower thermal budget (< 500 °C) and can be employed to fabricate high-performance devices (MOSFETS or optical devices) on upper active layers even with metallization layers below.

The SPC technique offers the flexibility of creating multiple active layers and is compatible with current CMOS processing environments. Recent results using the MILC technique prove the feasibility of building high performance devices at low processing temperatures, which can be compatible with lower level metallization [118]. It is found that the electrical characteristics of these devices are still inferior to single crystal devices [119]. However, technological advances to overcome the thermal budget problem have been made to allow fabrication of high-performance devices using SPC [120], [121], [122].

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 [123], stacked SRAM cells [124], [125], and EEPROM cells [126] have already been demonstrated. With technological improvements, the MILC (Ni seeding) process can be used to fabricate islands of single-grain-devices to maximize circuit performance.

### 6.2 Vertical Inter-Layer Interconnect Technology Options

The performance modeling presented in this study directly relates improved chip performance with increased utility of VILICs. It is therefore important to understand how to connect different active layers with a reliable and compatible process. Upper-layer processing needs to be compatible with metal lines underneath connecting lower layer devices and metal layers. With Cu technologies, this limits the processing temperatures to < 450 °C for upper layers. Otherwise, Cu diffusion through barrier layers, and the reliability and thermal stability of material interfaces can degrade significantly. Tungsten is a refractory metal that can be used to withstand higher processing temperatures, but it 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, as schematically illustrated in Fig. 11.

Figure 33. Schematic of final steps used in one of the wafer bonding technologies based on metal thermocompression (top) and a finished 3-D chip (bottom). (Courtesy of Rafael Reif and Dimitri Antoniadis, Massachusetts Institute of Technology, Cambridge, MA).
Figure 34. Schematic of an epitaxially grown second active layer. ELO denotes epitaxial layer overgrowth. (Courtesy of Gerold W. Neudeck, Purdue University, West Lafayette, IN).

Figure 35. Schematic of the Ge seeded Solid Phase Crystallization (SPC) process flow.
Figure 36. Schematic of the MILC process flow using Ni seeding.

Figure 37. Schematic of the wafer bonding techniques a) with adhesive layer of polymer in between, and b) through thermocompression of Copper metal. (Courtesy of Rafael Reif, Massachusetts Institute of Technology, Cambridge, MA).
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 [127]. Interchip vias are etched through the ILD (inter level dielectric), the thinned top Si wafer and through the cured adhesive layer, with an approximate depth of 20 µm prior to the bonding process (see Fig. 37a). The interchip via made of chemical vapor deposited (CVD) TiN liner and CVD W plug provides a vertical interconnect (VILIC) between the uppermost metallization levels of both layers. The bonding between the two wafers (misalignment ≤ 1 µm) is done using a flip-chip bonder with split beam optics at a temperature of 400 °C.

A second technique relies on the thermocompression bonding between metal pads in each wafer [110]. In this method Cu/Ta pads on both wafers (illustrated in Fig. 37b) serve as electrical contacts between the interchip via on the top thinned Si wafer and the uppermost interconnects on the bottom Si wafer. The Cu/Ta pads can also function as small bond pads for wafer bonding. Additionally, dummy metal patterns can be made to increase the surface area for wafer bonding. The Cu/Ta bilayer pads with a combined thickness of 700 nm are fused together by applying a compressive force at 400 °C. This technique offers the advantage of a metal-metal interface that will lower the interface thermal resistance between the two wafers (hence provide better heat conduction) and can be beneficial as a partial ground plane for lowering the electromagnetic effects discussed in Section 4.2.

**SUMMARY**

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

In Section 1, the interconnect delay problem associated with Cu/low-k technologies was discussed using estimated delay values based on the data from the ITRS. The implications of material effects arising at deep submicron dimensions such as increasing metal resistivity of Copper due to increased electron scattering and the effect of a finite barrier layer thickness on line resistance were quantified. The increasing impact of interconnect delays on VLSI design was also discussed and the limitations of various proposed solutions to overcome the interconnect problem were highlighted, especially in light of ITRS based interconnect trends and their associated effects. It was concluded that Cu/low-k interconnects alone will not be able to solve the deep submicron interconnect problem, and that existing design based solutions are also not adequate to deal with the wiring problem. Additionally, various limitations of the existing planar (2-D) ICs with regards to their utility for heterogeneous integration of technologies were also discussed.

In Section 3, a detailed performance analysis methodology was presented for the 3-D ICs to accurately predict 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. Our analysis predicts significant performance improvements over the 2-D case. The primary target technology for this analysis has been the ITRS based 50 nm node with two active layers of Silicon. Other technology nodes with two active layers were also considered. It was shown that the availability of additional Silicon layers gives extra flexibility to designers which can be exploited to minimize area, improve performance and power dissipation or any combinations of these.

Additionally, in Section 4, we addressed some of the concerns associated with 3-D circuits including that of heat dissipation. An analytical thermal model for estimating the temperature rise of individual active layers in 3-D ICs was presented. It was demonstrated that for circuits with two Silicon layers running at maximum performance, maintenance of acceptable die temperatures might require advanced packaging and heat-sinking technologies. Implications on reliability and electromagnetic interactions (such as capacitance and inductance effects) arising in 3-D ICs were also briefly discussed.
In Section 6, we highlighted some scenarios in current and future VLSI and systems-on-chip type applications involving mixed signals and technologies, where the use of 3-D circuits will have an immediate and beneficial impact on performance. We also briefly discussed the implications of using this technology on the design process, as conventional VLSI design methodologies and tools, gate level and architecture level synthesis algorithms need to be suitably adapted. Finally, in Section 6, an overview of some of the manufacturing technologies under investigation, which can be used to fabricate these circuits, was provided.

CONCLUSIONS

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 paper we have shown that 3-D ICs are an attractive chip architecture that can alleviate the interconnect related problems such as delay and power dissipation, and can also facilitate integration of heterogeneous technologies in one single chip. In fact, several applications of 3-D ICs have been recently demonstrated [128], [129], [130], [131], 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 [132] can be precursors to the future monolithic 3-D ICs.

ACKNOWLEDGMENTS

This work was supported by the DARPA AME Program and the MARCO Interconnect Technology Focus Center. The authors would like to acknowledge Amit Mehrotra, University of Illinois at Urbana-Champaign, for several technical discussions during the initial phase of this project. They would also like to thank Gaurav Chandra, Chi On Chui, Sungjun Im, Amol Joshi and Rohit Shenoy, all from Stanford University, for several interesting discussions and for providing feedback.
REFERENCES


[55] K. E. Goodson, Stanford University, Private Communication.


