RC24147 (W0701-009) January 2, 2007
Computer Science

IBM Research Report

Three Dimensional Integration Technology: A Microarchitectural Outlook

Eren Kursun
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Three-Dimensional Integration Technology: A Micro-Architectural Outlook

Eren Kursun
Reliability and Power-Aware Micro-architectures Department
IBM T.J. Watson Research Center
Yorktown Heights, NY

Abstract- Three-dimensional integrated circuits offer unique advantages compared to the planar counterparts: such as improved interconnect delay, enhanced performance and packaging density. Yet, 3D technology presents key challenges ranging from fabrication complexity to thermal issues that need to be addressed for performance and viability of vertically integrated microprocessor architectures. In this paper, we discuss the basics of 3D technology from microprocessor architecture design point of view. We present the main techniques in the 3D fabrication as well as key issues and challenges. One of our goals is to highlight the characteristics that require further attention from higher-level design stages. 3D technology research has rapidly expanded in the past few years and became a very active research area. We provide brief overviews of recent studies and tools from both academic and commercial domains. We also discuss the possible future directions for architecture-level 3D research.

1. Introduction and Motivation

Technology scaling has been one of the main factors that enabled faster, more sophisticated and lower-power electronic circuits. On the contrary, scaling trends have not had such a positive impact on the on-chip interconnects. Increasingly complicated interconnect networks of thinner and longer wires spanned more metal layers. As a result, interconnect scaling caused number of problems including higher signal propagation delay, routing complications and noise coupling. Current microprocessor architectures have complex interconnect networks with tens of kilometers of wiring per square centimeter, which dissipate a significant portion of the total active power (over 30% of the overall power dissipation according to [Black06]). On average, interconnect networks consist of 10^{17} coupling inductance/capacitances on more than 10 level metal stacks. As a result, over three quarters of the delay is attributed to interconnect for 65nm technology [Rickert04].

Performance of ULSI circuits such as microprocessors, is increasingly interconnect limited [Bohr95] [Ho01] [Meindl02]. Despite the new materials and techniques like Cu with low-k dielectric, wiring delay is expected to be limiting the performance for next generation as well. Figure 1 illustrates the negative impact of technology scaling on interconnects based on ITRS 2005 data [ITRS2005]. Even though the gate delay (in FO4) as well as local interconnect delay are reduced with smaller feature sizes, the global wiring delay shows exponential increase. Another main problem with on-chip interconnect is the difficulty of estimation: Since the on-chip interconnect is strongly related to the actual chip layout, it is very difficult to estimate the corresponding delay in the earlier stages of the design flow. Therefore the current synthesis based ULSI design flow is faced with timing closure problems.

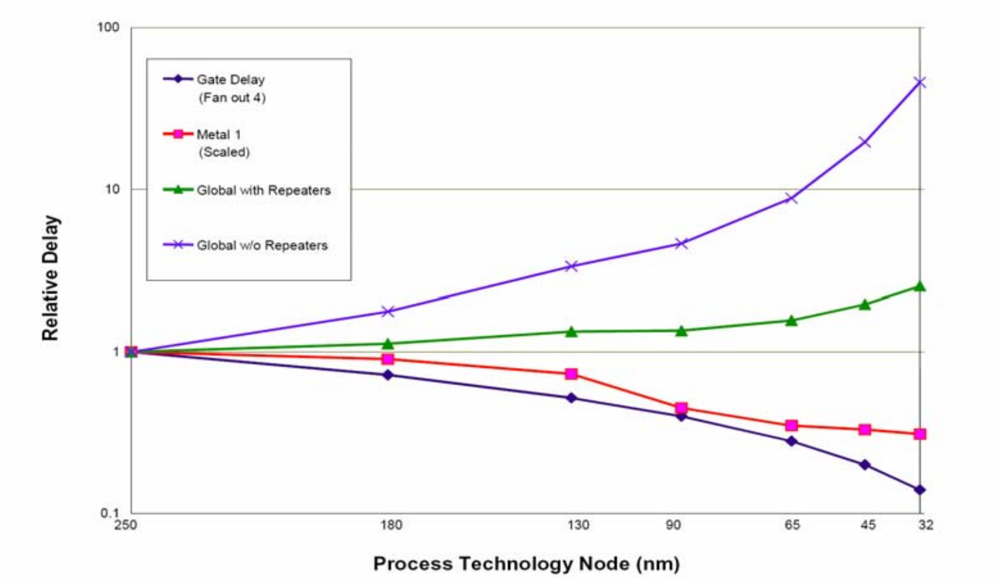


Figure 1. ITRS 2005 data on interconnect scaling

Advanced materials, optimizing interconnect dimensions and additional metal layers have been standard techniques used for handling interconnect scaling. A wide range of techniques has been proposed to address the exacerbated future challenges as well. These methods range from packet-based on-chip communication networks [Guerrier00] to non-Manhattan routing [Koh00], optical interconnects [Kobrinsky04] and 3D integrated circuits.

In this study we focus on one of these techniques, three-dimensional integration technology, as a means of alleviating interconnect driven problems. Vertical integration of silicon layers has gained substantial interest from industry and academia recently. As a result of the greater number of nearest neighbors, 3D technology reduces the interconnect length and latency considerably. According to [Davis01], increasing the number of active layers through vertical integration improves the interconnect performance by up to 145% for 50 nm node. Similarly, [Zhang01] reports higher benefits in area and interconnect delay from stacking device layers.

Three-dimensional integration can be used towards alleviating a number of problems faced by microprocessor design. One such issue is the infamous memory scaling problem: On average processor performance has been improving around 60% every year. However, the corresponding improvement in memory access time is less than 10%. This gap has been the main issue behind the limitations of logic-memory system. The interconnect latency and bandwidth improvement of 3D technology is quite promising to alleviate this gap. Three-dimensional IC technology enables integration of memory on the same chip and eliminating number of slow off-chip buses (<200 MHz frequency) by replacing them with on-chip interconnects (~2 GHz) [Loi06]. Blocks that are few millimeters apart in the planar 2D technology can be interconnected through vias that are few micrometers long in Face-to-Face 3D. As a general trend, the performance improvement through vertical integration is increasing roughly as the square root of the number of circuit layers in the stack according to [Topol06].

Memory scaling is not the only issue that the microprocessor architectures face: Future micro-architectural projections with tens of processor cores and multi-billion transistor counts are challenging the industry with packaging, interconnect, power and design complexity issues. Even though the complexity/functionality of microprocessors has been increasing according to the Moore's law, there are strong debates about the future challenges. Vertical integration technology presents promising characteristics for a number of scaling issues. These characteristics include reduced footprint area, along with increased packaging density, even for the same technology node. Footprint reduction is the main driving factor for adaptation of 3D in embedded systems. It is worth noting that interconnect is a more serious issue for microprocessors, yet the processor footprints are steadily increasing. (Intel®'s Montecito® with reportedly over 500 mm^2 footprint [Montecito05]) Projections indicate chip crossing latencies are likely to go up to tens of cycles in the near future [Bernstein03].

Yet another opportunity that 3D enables is the effective integration of disparate signals or technologies on the same chip - in a heterogeneous framework. The integration of disparate layers into mixed-signal or mixed-technology designs (such as analog, memory, RF, FPGA, CMOS, nano-tubes) creates exciting design opportunities. Additional device layers can be dedicated to auxiliary engines that perform management of processor resources, tracking software bugs and reliability checks. On the other hand, vertical integration has a number of challenges ranging from manufacturing complexity to thermal problems. Current research studies indicate most of these issues are manageable. For instance, with the recent advances in thermal vias and 3D manufacturing process, thermal problems are expected to be challenging, but not critical limitations [Black06].

Ultimately, the product value will be the factor that determines the future and wide acceptance of this technology. The product value consists of many variables that affect the entire design/manufacturing flow; including system manufacturing cost, die manufacturing cost, performance, power and other design parameters. 3D chips are already being shipped in other product domains such as embedded systems, sensor and memory applications. The preliminary indicators for 3D product value seem positive; yet the wide spread availability of three-dimensional microprocessor architectures is a matter of research in both processor design and 3D manufacturing technology.

Three-dimensional integration research has rapidly expanded in recent years and its span by far exceeded the initial manufacturing/packaging technology. There has been strong interest in 3D from both academia and industry. Our main goal in this study is to analyze the 3D technology from micro-architecture design point of view. We focus on the following in this study:

- We present basics of the 3D technology (fabrication stages, main advantage and disadvantages)
- We discuss the implications of the existing 3D technology on high-level design decisions
- We briefly explore state-of-the-art 3D tools and techniques, from physical design to architecture-level exploration
- We discuss future research topics and challenges from high-level design point of view

The rest of the paper is organized as follows: In Section 2 we discuss the basics of three-dimensional integration technology starting with the historical trends to variations. Section 3 presents the high-level overview of the manufacturing flow. In Section 4, we analyze the advantages and disadvantages of 3D. State-of-the-art 3D tools and techniques are briefly discussed in Section 5. Section 6 presents the architectural exploration of the 3D domain, including previous studies. Finally we conclude with a discussion of future 3D architectures in Section 7 and 8.

2. Preliminaries

2.1. History and Trends

Many consider the initial proposal of 3D ICs in late 1970s [Geis79]. Other process technology research followed during 80s [Mukai83], [Asaka86] and [Kunio89]. Nevertheless, 3D technology remained mostly a research technology until the late 1990s, since electronics design was largely limited by logic rather than interconnect.

At the initial stages of research, vertical integration was strongly limited by the number of vias: Up to 200-250 vertical interconnects between layers was reported by [Asaka86]. In recent years, novel integration techniques have provided progress in traditional 3D issues. Currently, the number of interlayer interconnect is reported to be over 100K/cm^2.

Until recently, practical interconnection of chip stacks was only achievable through wire bonding at the periphery, also called System-in-Package (SiP). This interest shifted towards wafer or chip level integration, as 3D solves many issues related to SiP (with lower cost per function and higher functional density). The interconnect lengths are also reduced from 10-20 mm of the SiPs to μm level in 3D.

Today, an increasing number of companies and academic groups are involved in 3D IC research. Vertically integrated chips have become available in embedded, wireless and memory sectors in recent years. These systems have strict area, volume and weight limitations. Thus, shrinking the footprint of the chip is highly desired. Microprocessor architectures are considered to be the next step in the adaptation of three-dimensional integration. (Companies interested in 3D range from IBM, Intel, Samsung, to start-ups such as Ziptronics, Xanoptics, Tezzaron, Sonics etc).

2.2. Types of Vertical Integration

In a broader sense, vertical integration of chips/wafers can be achieved in a variety of different ways: including through the periphery in SiP; through capacitive coupling in contactless integration; through micro-bumps and finally inter-layer vias [Davis05]. In this article we restrict ourselves to the approaches with increased inter-layer bandwidth with more applicability for microprocessor architectures. We focus on wafer-wafer integration through inter-layer interconnects for the most part.

Another variation of the three-dimensional integration concept is the 2.5D, whose definition is less clear. In general, 2.5D technology is considered to be a subset of 3D with minimum amount of modification to the manufacturing process flow which is achieved by integrating fully processed dies that are manufactured separately in their optimal technology. The integration can be achieved by through the silicon vias or terminals placed on both chips [Deng03] [Dong06].

3. Fabrication Technology

3D fabrication spans a wide range of techniques, each with significantly different characteristics. These vary in terms of their baseline interconnect length reduction, bandwidth; as a result, the architectural features they enable. In this section we discuss the basic 3D fabrication techniques from microprocessor design point of view. Hence we do not delve into many technical details in manufacturing process. Further details of the fabrication techniques can be found in: [Banerjee01], [Ieong03], [Reif02], [Burns00], [Burns06], [Kim2005], [Subramanian91], [Beyne04], and [Chan2001]). In this study we focus on wafer-wafer integration. It is important to note that there are other alternatives such as wafer-die, die-die and hybrid integration techniques as well.

There are 2 commonly used schemes for wafer integration: top-down and bottom-up approaches. Figure 2 shows example cases for these approaches: face-to-face bonding for top-down approach and multi-layer buried structures for bottom-up integration. As the names imply, top-down and bottom-up approaches indicate the general characteristics of the manufacturing flow.

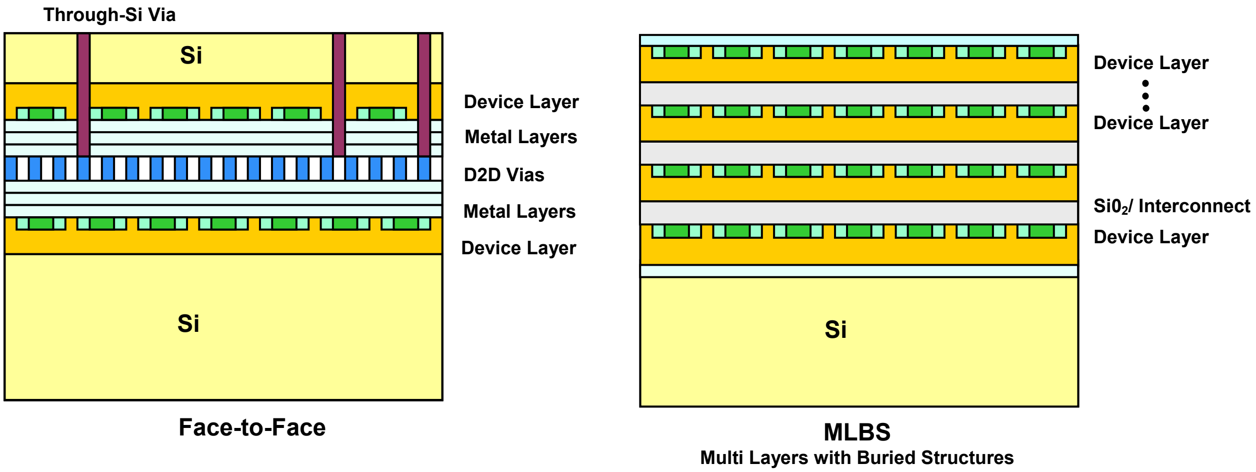


Figure 2. Top-down (F2F) and Bottom-up (MLBS) integration examples

3.1. Sequential Approach (i.e. Bottom-Up)

In the sequential approach the first-level device layer is created initially. A sequential layering process builds multiple device layers on top of it in the following stages (hence - bottom-up). The additional device layers can be integrated by various techniques such as:

- Laser beam re-crystallization [Kunio89], [Kawamura83]
- Lateral over-growth epitaxy [Neudeck00]
- Metal induced lateral crystallization [Chan00]

The back-end processing then builds interconnect patterns among devices. An example of bottom-up approach is MLBS (Multiple Buried Structures) [Xue01].

3.1.1. Multiple Layer Buried Structures

MLBS approach is considered to be a promising 3D integration technology. Figure 3 illustrates the high-level flow of this fabrication technology [Xue01], [Xue03]. As with many other techniques it can be implemented on Bulk Si or SOI with slight variations. In the first stage the front-end processing of devices on the initial wafer is completed. The inter-layer and in-plane interconnect patterns are then formed in a deposited dielectric (SiO_2). This step is repeated for second layer of interconnect or buried ground plane. The interconnect material is poly-Si or Tungsten in the current versions.

Following the initial device and interconnect formation, the second layer (donor wafer) is bonded to the first layer (host wafer) at room temperature. Chemical-mechanical polishing is used for thinning the donor layer and in order to achieve the required surface roughness for device fabrication. The front-end device processing and interconnect formation is then repeated for higher levels of 3D integration.

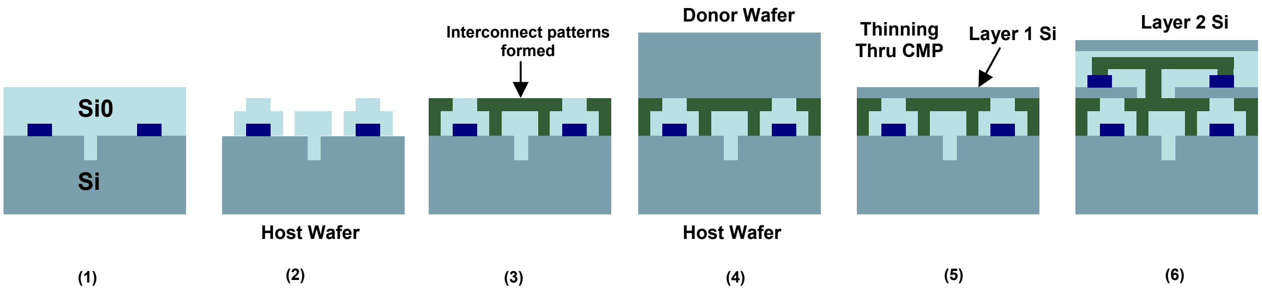


Figure 3. MLBS fabrication process (1) Device layer formation on host wafer and $Si0_2$ deposit (2) Interconnect patterns are etched (3) Interconnect formation (poly-Si or tungsten) (4) Donor wafer integrated (5) Thinning of the donor wafer through chemical-mechanical polishing (6) Additional interconnect and device layers integrated

Even though this approach requires modifications and extensions to the standard CMOS technology, it provides clear advantages in terms of interconnect bandwidth. The inter-layer via size can scale at a similar rate with the feature size in this technology. Hence it is capable of achieving via densities of higher than 10^7 via/cm^2.

3.2. Parallel Approach (i.e. Top-Down)

In the alternative scheme, multiple device layers are fabricated separately and later integrated to each other. The integration can occur in a number of ways such as Face-to-Face, or Face-to-Back (i.e. device layers facing each other or stacked on top of each other). Individual wafers are bonded following chemical-mechanical polishing and wet surface treatment stages. These stages are targeted towards thinning and polishing the surface for better bonding quality. Subsequently layers are bonded to each other, either using metal-to-metal or using polymeric or dielectric glue layers. Functional verification of each layer can be completed before the integration, which improves the overall yield. Yet the bonding of multiple wafers create complications as a result of wafer-wafer alignment problems, reliability issues regarding high temperature/pressure based bonding techniques. General flow for the parallel integration approach is illustrated in Figure 4.

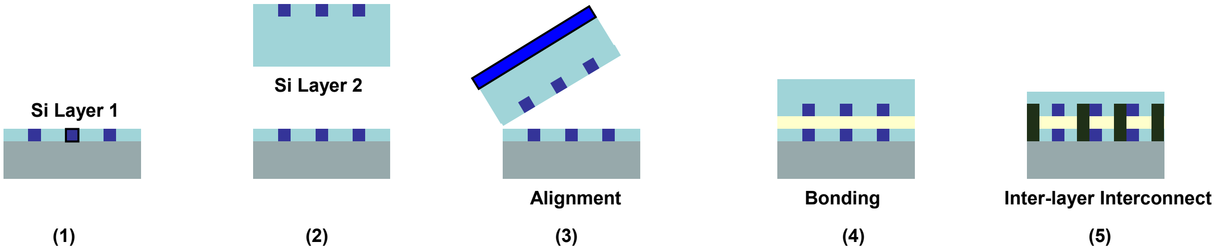


Figure 4. Top-down approach (1) Initial host silicon layer (2) Parallel fabrication of individual device layers (3) Alignment of the device layers (4) Bonding through the use of glue (or alternatively Cu-Cu thermo-compression) (5) Formation of inter-layer interconnect (Through-silicon-vias are used for Bulk F2B, regular metal layer interconnects for F2F)

Even though different wafer bonding techniques vary in their process flow they share common requirements as well: (i) Bonding (ii) Si substrate thinning (iii) Wafer alignment (iv)Vertical interconnection stages. We will elaborate on these common stages in section 3.3.

Since device layers are fabricated separately in parallel, the top-down technique is especially promising for the heterogeneous integration of disparate technologies (analog, digital, RF) on the same chip. It is very important to note that there are number of variations for each technique discussed in this study. In many cases, characteristics of different implementations may be significantly different from each other. The scope of this study is to highlight the basic ideas, characteristics, and challenges in 3D integration. Further details of individual techniques can be found in the listed references.

Among the aforementioned wafer bonding schemes, Face-to-Face bonding provides the higher interconnect density than the alternative bulk Si techniques. The reason is mainly the fact that Cu-Cu bonded vertical links do not go through silicon to connect the second device layer [Morrow04], [Morrow06]. In the case that more than 2 layers are integrated through face-to-face technique, Back-to-Back wafer bonding is required. In back-to-back integration deep vias are required to propagate the signals out to the package. Thus, B2B doesn't exhibit the same high inter-layer interconnect density as face-to-face. Hence it is interconnect limited.

Face-to-Back technique involves using through-the-silicon vias with unfavorable scaling properties as well. Via sizes in F2B are in the order of 1-10 μm according to [Bernstein06]. Face-to-Back bonding has bulk Si and SOI variations with quite different characteristics in terms of fabrication technology, interconnect bandwidth, etc. In the SOI-based approach the distance between the device layers is the smaller than both the bulk F2B and F2F. Furthermore, the interconnect bandwidth is the highest. The bulk F2B approach usually involves Cu-Cu integration of different layers, whereas the SOI version incorporates fusion or adhesive bonding.

3.3. 3D Manufacturing Stages and Techniques

In general the 3D manufacturing requires additional stages to the standard 2D silicon flow depending on the integration technique. These stages and the corresponding manufacturing approaches include:

- Silicon growth *(Bottom-up)*
- Bonding *(Top-down/bottom-up)*
- Alignment *(Top-down)*
- Wafer thinning *(Top-down/bottom-up)*
- Inter-layer interconnect *(Top-down/bottom-up)*

The following summarize the most common 3D techniques specifically targeting these manufacturing stages.

3.3.1. Beam Re-crystallization

Re-crystallization is a common bottom-up silicon growth technique for fabricating second silicon layer on top of an existing layer through depositing polysilicon and fabricating thin-film transistors. Laser or electron beams are used to induce re-crystallization that enhances the performance of the thin-film transistors. Major drawbacks of this technique are (i) high temperatures during polysilicon melting (ii) issues in controlling the grain size variation (iii) lower carrier mobility's in re-crystallized polysilicon. There were improvements in the temperature requirements for single-silicon thin film transistors that alleviate the latter problems [Kunio89], [Akasaka86].

3.3.2. Silicon Epitaxial Growth

An alternative technique for sequential silicon growth is silicon epitaxial growth. In this technique a hole is etched in the wafer and a-single crystal silicon is seeded from the ILD (inter layer dielectric). This silicon seed grows vertically in the hole then continues laterally and covers the ILD. Similar to the beam re-crystallization technique, silicon epitaxial growth suffers from high temperatures up to 1000^{o}C. Furthermore, this method cannot be used for metal layers [Neudeck99] [Lin97].

3.3.3. Solid Phase Crystallization

Solid phase crystallization involves deposition and crystallization of amorphous silicon [Nakata82] [Yamauchi94]. The resulting thin film transistor performance can be improved by removing grain boundaries. Since the deposition is low temperature, this technique provides alleviation to the increased temperature problems in silicon epitaxial growth. This technique provides promising results in terms of 3D fabrication.

3.3.4. Wafer Bonding

Wafer bonding is a crucial stage that determines the overall quality of the three-dimensional structure. The bonding quality can be assessed through:

- Interconnect pattern size and density
- Compatibility to the back-end process
- Alignment accuracy
- Thermal control (in order to limit the misalignment at higher temperatures)
- Electrical conductivity

- Mechanical support

There three commonly used techniques for wafer bonding, depending on the material: (i) Fusion bonding (oxide to oxide) (2) Adhesive bonding (polymer to polymer) (3) Metal bonding (Cu-Cu) [Reif02]. In the top-down scheme, wafer bonding starts with fully processed wafers, common to all of the above techniques [Rahman00]. Bonding through cu-cu interconnects provide both mechanical and electrical connections between device layers. It provides relatively large inter via dimensions and lower via density. On the other hand, it provides improved thermal conductivity between layers, compared to the other techniques

3.3.5. Alignment

A crucial aspect of top-down wafer bonding is the need for complete overlap of the device and interconnects patterns. Alignment of the silicon layers determines the quality of interconnect conductivity and density, as well as mechanical support between the layers. It is important to note that alignment accuracy changes between pre-binding and post-bonding stages. Increased bonding temperatures at the bonding stage degrade the alignment considerably. The resulting inter-layer interconnects are only limited to global interconnects due to the alignment limitations.

4. Main Advantages and Challenges

4.1. Advantages of 3D

4.1.1. Interconnect Reduction

One of the main benefits of vertical integration is the interconnect reduction. In general interconnect delay is proportional to wirelength^2/pitch^2. As a result of the wirelength reduction 3D architectures provide significant delay reduction especially in the semi-global and global wiring layers. 3D circuits are reported to improve the chip interconnect wire length around 25-45% for 2-5 layers of die stacking [Rahman00].

Increasing number of metal layers is another issue that occurs due to the higher routing complexity in current microprocessor architectures. It has been shown that 3D integration enables reduction in the number of metal layers by 25% [Joyner01]. The corresponding wire length reduction yields 3.9x increase in the wire limited clock frequency as well as 84% decrease in the corresponding wire limited area.

Interconnect Power Reduction: The corresponding interconnect power is reduced by up to 25% for 2 layers and the energy delay product is enhanced 30-50% [Das04]. Interconnect power constitutes as much as 50% of the active power dissipation of the modern microprocessor architectures. Hence, 3D architectures have obvious power efficiency advantages.

4.1.2. Heterogeneous Integration

Vertical integration enables heterogeneous integration of disparate technologies such as RF, analog, memory. Utilizing disparate materials such as Si, SiGe, and GaAs also enables finely optimizing the processor design for various technologies.

In the simplest form, this idea can be used for integrating memory in multi-layer processor architecture. Number of research studies has shown significant benefit in integration of memory on logic layers [Liu05], [Zeng05], [Lee00]. Currently, integrating an entire system on a single piece of silicon (SoC) requires integrating digital, analog, flash, DRAM. This approach is advantageous in terms of reduced I/O, power dissipation, EMI and improved overall performance. Separate processing of individual technologies and integration the corresponding wafers provides significant relief to the existing problems. Cost of heterogeneous 3D architectures is expected to be lower than the existing SoC process which experience higher costs due to increased mask stages and incompatibilities of the individual processes. As a result 3D integration is considered to be a very promising alternative to SoC.

4.1.3. Packaging Density

3D integration also improves the packaging density of the circuit by integrating increased number of transistors while reducing the footprint. For next generation microprocessor architectures with multi-billion transistors, the processor footprint will likely become a design constraint. Furthermore, the manufacturing cost in high quantities is reported to be lowest for 3D, compared to SiP and SoC [Lu05].

The footprint reduction is not perfectly proportional to the number of layers, due to the area of head of the through-the-silicon vias for some technologies (top-down). On the other hand, inter-layer interconnect overhead is expected to be minimal for other techniques (such as MLBS).

The wire limited chip area is also reduced with the semi-global pitch. If the normalized semi-global pitch decreases, the wiring patterns need to be changed. The wires are rerouted to global metal tier with larger pitch. Hence the wire limited chip area increases. In total, the minimum wire limited chip areas for the 3D case is 30% lower than the planar counterpart [Banerjee01].

4.1.4. Noise Immunity

The reduction in the interconnect length and coupling capacitance provides improvement in the associated parasitic effects. It has been shown that 3D architectures have consistently better noise immunity compared to the planar counterparts. (such as reflection noise, crosstalk noise, simultaneous switching noise, and electromagnetic interference). The vertical separation of metal layers, with device layers and dielectrics, is one of the reasons for the improved noise immunity. In three-dimensional ICs, the line spacing 's' can be chosen much smaller without reaching the logic threshold voltage of CMOS gates. Even with the reduced absolute values of peak cross-talk amplitude, the noise immunity of vertically integrated circuits is improved. The main reason for the improvement is the reduced capacitances.

On the other hand, the electrical coupling between the top layer metal of the first active layer and the devices in the second layer is a concern for a number of vertical integration techniques. Especially the Face-to-Face integration is more prone to crosstalk problems; since it incorporates high bandwidth vertical interconnect patterns between close metal layers.

Similarly, thin-film SOI designs require special attention since the capacitive shielding between signal layers is not assured by silicon layer. Special conductive layers might be required for noise sensitive applications on 3D integration [Kuhn95]. Crosstalk problems are more of an importance for mixed signal integration of 3D ICs. Various solutions including ground plane structures have been proposed to alleviate the crosstalk problem [Kim05].

4.2. Issues and Challenges

Although the vertical integration technology is considered to be promising for a number of problems that we discussed in the previous section, it is not problem-free. The main issues in 3D packaging technology are: quality, density of vertical interconnects, electrical and thermal characteristics, availability of design tool kits, reliability, testability, packaging cost and fabrication cost. The integration and fabrication of vertically integrated ICs are more complicated, with increased number of steps and complications in alignment and bonding stages.

4.2.1. Vertical Interconnect

Vertical interconnect is interestingly both one of the main advantages and challenges of 3D ICs. Overall interconnect reduction is achieved through the efficient use of vertical signal transmission. Furthermore, total interconnect capacitance decreases through the use of 3D integration technology. However, there are number of challenges ranging from manufacturing complications to layer limitations that are associated with the inter-layer interconnects. For instance in the simplest case, vertical interconnects dominate and the corresponding wire capacitance increases above 6 layers [Hua06].

The interconnect bandwidth and length between device layers is the main determining factor for the feasibility of 3D integration technology. Vertical interconnect between device-layers is strongly tied to the alignment tolerances between wafers in parallel integration [Warner04]. There has been number of different approaches for the inter-layer interconnection.

One alternative is the use of through-the-silicon-vias (TSVs). Current TSV technologies are as small as 2-4 μm in diameter and average via structure is 15-20μm deep in the silicon. As a result of thinning of silicon wafers, through the silicon vias have become more feasible. TSV formation is an intricate stage as the current photolithographic process nears the resolution limit for 2 μm vias [Schaper05]. However, in general:

- Through-the-silicon vias scale slower than technology. Hence, the relative cost is expected to increase for the next generation 3D ICs.
- TSVs incur considerably higher area cost and bandwidth limitations compared to alternative schemes such as MLBSs
- They require additional pattern etching stage to the standard manufacturing process
- TSVs in B2B technology and I/O for F2F technology generate bandwidth limitations between layers

The scaling characteristics of inter-layer interconnect and their relative size to the logic determines how fine granularity 3D integration can be accomplished. Low-cost and high density inter-layer interconnect such as MLBS enable splitting processor blocks such as caches and register files to different device layers at a fine granularity. In such scenario, individual transistors of the memory cell can even be separated to multiple device layers.

Alternative Inter-Layer Communication Schemes: Other inter-layer interconnect techniques such as micro bumps and contactless interconnects are currently researched. In micro bump technology, solder or gold bumps on the surface of the die are used to make connections [Davis05]. The bumps typically have 50-500 μm pitch. The micro bump technology can be used for face-to-face bonding, limited to 2 tiers. It offers high-density I/O, however the existence of physical contacts requires a highly capacitive ESD structure. On the other hand, due to the delay/power/area increase, scaling of pads is quite difficult.

Another technique for vertical integration is contactless or AC coupled interconnects that use capacitive or inductive coupling to provide communication between device layers. One of the main advantages of this scheme is the simplicity of implementation. The resulting cost is less than micro bump or through-silicon-via techniques, as the additional manufacturing stages are few [Kanda03], [Drost04].

4.2.2. Temperature

One of the main challenges of three-dimensional integration is considered to be thermal challenges. In general, the increased packaging density of 3D ICs in terms of reduced 'Surface Area/Volume ratio' increases the power density. Hence it has a negative effect on the chip temperature profile. The advantages of 3D technology in terms of interconnect and packaging aspects become more prominent with increased number of layers, whereas the heating exacerbates.

In general the thermal problems in 3D integration can be summarized as:

(i) Increased power density due to reduced footprint and surface area
(ii) Increased distances from device layers to heat sink and spreader
(iii) Isolation of device layers through dielectrics (LTO, HSQ, polyimide etc) with low thermal conductivity (below 0.3W/mK) [Banerjee96] [Koo05]

Temperature considerations are critical even for the conventional microprocessors. In general, thermal challenges have serious reliability, timing, and data implications that threaten the proper functionality. Since the reliability of an electronic circuit is exponentially dependent on the junction temperature, even a 10-15^{o}C increase in the operating temperature results in 2x reduction in the lifetime of the device [Viswanath00].

Furthermore, the leakage power has exponential dependence on the on-chip temperatures. Increased average temperature is likely to elevate the leakage power due to the positive feedback loop between leakage and temperature. Static power dissipation exceeds the active power for technologies of 65 nm or smaller [Kim03] without special leakage management. 3D ICs are expected to incorporate increased number of transistors more effectively on the same chip. Therefore, they are very susceptible to elevated leakage power and on-chip temperatures. Techniques that target the management of on-chip temperatures and leakage are of critical importance for 3D technology.

There are a number of tools for specialized thermal analysis of 3D ICs ranging from: finite element based analysis, to HotSpot3D. The thermal evaluation of 3D architectures yields quite variable results depending on the baseline fabrication technology and the use of thermal vias. Yet increase in on-chip temperatures from current 2D designs is persistent in all cases.

Thermal analysis of earlier 3D techniques pointed to alarming on-chip temperatures [Banerjee01]. According to [Kleiner95] the self heating caused temperatures to exceed 200^{o}C as a result of the low thermal conductivities, which is much higher than the maximum temperature conventional silicon process usually permits [Hua06]. Most circuits are designed for the worst case temperature of 125^{o}C [Hua2006]. Thus, the numbers reported in these studies well exceeded Si temperature range. Similar dramatic increases are observed in studies such as [Im00], where even for the smallest power densities of 0.1 W/mm^2 2-3 layer 3D chips are above 200^{o}C.

On the other hand some recent studies that incorporate improved fabrication technologies along with thermal vias, present a much more optimistic outlook. Thermal characteristics of 2 and 4 layer 3D implementation of an Alpha 21364 processor were analyzed in [Puttaswamy06]. This study indicates mild temperature increases in maximum temperature up to 20-30 K for the 2 and 4 layer cases respectively. Similarly, [Black06] observed only an average 14^{o}C increase in the hotspot temperature for 2-layer face-to-face implementation of Intel® Pentium4® based microprocessor architecture. In the worst case temperature of the hotspot increased from 98^{o}C of the planar case to 124^{o}C in the 3D version. Although the average temperatures are down by 28% and maximum temperatures almost by half [Goplen05], one of the most important finding is that thermal vias are of limited effectiveness for challenging regions such as upper most silicon layers or other high temperature regions. They conclude that these regions have to be addressed separately by reducing the power dissipation that leads the thermal gradients.

In summary, the temperature differences between these two camps can be attributed to the special 3D technology used: SOI/Si, thermal vias, dielectrics etc. Thermal vias have been shown to be effective in improving the thermal conductivity between device layers in 3D. However, it is important to note that the correctness of either camp is yet to be determined through real measurements.

In general thermal modeling of three-dimensional integrated circuits is challenging. The major issues in the thermal analysis of 3D architectures include: (i) Lack of information about the baseline technology (3D process is still in research phase. There is no standard process similar to the current conventional Si technology.) (ii) Power modeling tools (such as the range of tools we have for planar architectures from high-level WATTCH analysis and lower level power modeling tools). Since interconnects constitute a significant portion of the driver capacitance of logic gates, the existing architectural power models and tools are not straightforward to use for 3D exploration. There seems to be a great need in accurate power modeling for different levels of abstraction for 3D ICs for better analysis and utilization of this technology.

Special Cooling Solutions

Even though the need for dynamic thermal management targeting is undeniable, some studies suggest that dynamic thermal management might not be sufficient for the smaller technology nodes around 45nm or smaller for 3D. Advanced heat sinks and cooling devices may be needed for thermally challenging 3D integrated circuits with more than 2 layers [Saraswat00]. As a result, special cooling solutions such as capillary loops with micro-scale evaporators, impingement cooling devices and micro-channels have been proposed. Closed loop cooling that has been shown to handle up to 500W/cm^2 [Upadhya03]. Among the cooling solutions micro-channels have been successfully adapted to 3D packages. Hotspot temperature of 300^{o}C can be reduced to 55^{o}C with micro-channel cooling [Koo05].

4.2.3. Yield

Yield of three-dimensional integration is currently a debated issue. Number of studies reports a decrease in the yield. They argue that the decrease in yield is partially caused by integration good wafers to the bad ones. Furthermore, the stages that employ high temperature/pressures in the wafer-level integration process are critical in effecting the yield. [Liu05] report yield decreases design cost by 17% and 29% for 2-3 layer 3D integrated circuits with 95% average wafer and bonding yields.

On the other hand, an alternative argument proposes that the percentage yield is likely to remain the same for separating the larger die into number of 3D layers. These studies also indicate that the smaller dies have the advantage of increasing the candidate sites [Patti06]. As there are a number of different 3D approaches with significantly different stages, it is quite difficult to argue the correctness of either approach. Better understanding of the yield characteristics of 3D integration is required for vertically integrated circuits to be widely available. Moreover, the reliability of 3D ICs needs special attention for the ultimate viability of the 3D technology.

4.2.4. Electronic Design Automation Tools

Although design automation is not a problem directly related to the 3D manufacturing technology itself, the lack of design automation tools and techniques is a major limitation for wide range acceptance of 3D. There is a strong need for optimizing the existing EDA flow for 3D. The current electronic design automation tools target planar integration. There are a number of recent research studies on behavioral and physical synthesis stages of the design flow [Mukherjee04-05]. However, there is still a strong need for wide range of design automation tools for more efficient utilization of the underlying 3D technology.

4.2.5. Design Complexity and Time to Delivery

Increased number and complexity of manufacturing stages makes 3D integration flow costlier than the traditional silicon flow. Each of the 4-5 additional manufacturing stages for 3D technology has its individual challenges.

For instance, wafer alignment is a main determinant of the quality of design in the 3D process. Alignment tolerance effects the quality of interconnect between the layers, as well as the signal degradation in the vertical direction. Current alignment stages are capable of achieving 1μm alignment accuracy with through wafer alignment strategy (using a glass substrate to transfer the donor wafer on to the host). At the same time, wafer-bonding techniques have temperature complications. Surface cleanliness and smoothness is another issue that affects the overall bonding quality. Interconnection between device layers is restricted by the alignment tolerances. It requires high aspect ratio, low resistance and low parasitic links.

Time to delivery is expected to increase significantly as a result of the design complexity [Al-Sarawi98]. Yet the circumstances are rapidly changing as integration of more than billion transistors in the planar approach is becoming more problematic with each process generation. Relative cost and complexity for 3D integration may change in the next few years.

4.3. Implications of 3D Characteristics

From micro-architecture point of view, the number of device layers, interconnect delay and bandwidth between different device layers, thermal characteristics etc. are all of critical importance for the high-level design of the architecture. Furthermore, the granularity of 3D integration is very important as it ranges from transistor level device splitting of individual SRAM cells, to stacking of microprocessor cores to different layers. In the case of multi-core architectures, the topology, communication patterns between cores and the placement on multiple-device layers result in considerably large design spaces for architects to explore.

At the very basics, many high-level design decisions are likely to be influenced if not determined by the baseline 3D technology. For instance, the number of device layers is likely to be determined by the electrical resistance of inter-layer interconnects, thermal characteristics of >2 layer devices. Interconnect delay and bandwidth is mostly determined by the manufacturing process such as top-down or bottom-up approaches. The thermal characteristics are strongly tied to the thermal conductivity of dielectrics and glues used in the wafer-integration stage.

Interconnect density effects: Depending on the baseline technology the number of inter-layer vias can be the limiting factor for 3D integration (such as the case of B2B or F2B bonding in top-down approach). However, for bottom-up approach the via density can be well over 1000 K/cm^2, providing opportunity to split the architectural structures to different layers at very fine granularity. At fine granularity transistors of the SRAM cell can be split to more than one device layers. Furthermore, it is possible to vertically align the L2/L3 cache blocks and memory cells so that access times are significantly reduced.

Electrical characteristics: [Topol06] presents the electrical characteristics of the inter-layer interconnect along with testing schemes that for high via densities. High-aspect ratio Cu via technique is commonly used to achieve the desired via density. Although the high aspect ratio causes the electrical resistance to be couple of times larger than a regular back-end via, the resulting resistances were still only a few ohms [Ranganathan05]. Via resistances of less than 1Ω for silicon on insulator wafers were reported [Burns01]. Successful signal transmission through chains of up to 10,000 vias was reported by [Topol05] and the corresponding problems in high aspect ratio vias are studied in [Ranganathan05]. The electrical characteristics of vertical interconnects do not seem to be threatening the increased number of layers. However, the vertical wire capacitance increases and dominates above 6 layers [Hua06]. Thermal problems are likely to limit the integration of increased number of layers faster than interconnect.

Thermal characteristics: 3D integration caused temperature increase consistently during all the previous research studies. Even though the amount of increase varies depending on the existence of thermal vias or better glue materials, the general trend appears that 3D microprocessors will likely be thermally challenged. Special power and thermal management techniques appear to be strongly needed. Integration of more than 2 layers is increasingly challenging due to thermal reasons. Number of special packaging and cooling techniques has been proposed.

3D technology is rapidly changing and expanding. Hence it is very difficult to estimate the future implications for next generation microprocessors. However, the issues we discussed in this section, namely interconnect density, electrical and thermal characteristics do not appear to be major limitations for future 3D microprocessors.

5. Tools and Techniques

There is a strong need for electronic design automation tools and techniques that target 3D technology. Even though the existing tools can be modified and extended, the CAD flow need to be reformed to better utilize the opportunities vertical integration has to offer.

5.1. Physical Design Tools

In order to utilize the underlying infrastructure, architectural analysis tools need to be strongly integrated with the physical planning tools. Design stages such as placement, routing and via placement are of critical concern for effective implementation in 3D. Thus, the lower-level design stages heavily influence the architectural decisions. In this section we briefly describe the physical design and planning tools 3D. Since it dictates the bandwidth and latency of the inter-layer interconnect; number of vias and their placement has strong impact on the overall design quality in vertically integrated circuits. To address this problem various studies investigate the via placement problem for minimum interconnect delay [Pavlidis06]. Since our goal is to discuss 3D integration from architectural point of view, we will not go into technical details of the physical design studies. Further details can be found in the listed references.

Besides the electrical connectivity between layers, 3D designs commonly employ thermal vias that are solely targeting efficient heat transfer from thermally challenging parts to the cooler layers and heat sink. Thermal vias are around x100 larger than regular vias and they generally require special drilling stage during the manufacturing process. Therefore, each thermal via incurs additional area due to the larger size and the corresponding white space allocation for drilling. The trade off between the thermal benefits and placement/area requirements of thermal vias generates an optimization problem. Thermal via placement to reduce the thermal profile of the 3D chips is studied by [Goplen05]. According to the finite element based thermal analysis thermal vias should be intelligently placed based on the thermal profile to be effective. Thermal vias count for about 9% of the chip area is capable of reducing the on-chip temperatures by over 60%. However, their placement needs careful optimization due to:

- Area overhead of up to 10% even with optimal placement of thermal vias (under current technology parameters. The overhead is likely to increase relative to the technology scaling)
- The effectiveness of the thermal via is strictly limited by the temperature gradient between the points connected by it
- At higher levels of the 3D stack temperature gradients become less, hence the effectiveness of thermal vias has been reported to be much lower.

Integration of physical design tools with architectural techniques is strongly needed in 3D integration. To address this [Cong06] proposes an exploration tool for physical design and architectural evaluation. The physical design engine incorporates 3D floorplanning, routing, interconnect pipelining as well as thermal via insertion. Furthermore, the architectural evaluation engine focuses on trade offs on clock frequency and the number of pipeline stages.

Other studies on thermal via placement include [Cong05] [Wang06]. [Das03] propose standard cell placement tool, global routing tool, as well as layout editor targeted for 3D architectures. The initial analysis of circuits using this toolset provides up to 50% reduction in total wirelength.

5.2. Cache Analysis Tools

Caches are considered to be good candidates for implementation of three-dimensional integration as they are highly wirelength dominated regular structures. 3DCacti, was proposed recently for better analysis of the architectural design space of 3D cache memories [Tsai05]. The tool is built on Cacti 3.0 for 2D with modifications to capture 3D effects [Cacti3.0]. The cache partitioning mechanisms and experimental analysis reveal the variation in savings depending on parameters such as cache size, number of layers, and partitioning specifics. We discuss further details of this study later on in 6.2.2.

5.3. Temperature Modeling and Simulation for 3D

Accurate thermal modeling plays an important role for thermally challenged 3D architectures. As we discussed in Section 4.2.2.three-dimensional integration causes on-chip temperatures to increases consistently. The main reasons are increased packaging density, reduced surface area and limited thermal conductivity between device layers within the chip. In this section we briefly go over the existing thermal modeling tools for 3D temperature analysis. At steady state heat conduction in a three dimensional integrated circuit can be described in the following form: $K_x. \partial^2T/\partial x^2 + K_y. \partial^2T/\partial y^2 + K_z. \partial^2T/\partial z^2 + Q(x,y,z) = 0$, where T is the temperature, $K_{x,y,z}$ represent the thermal conductivities in x,y,z directions and Q is the heat generated per unit volume.

From this three-dimensional heat flow we can extract the temperature increase of each device layer of the 3D chip in the form for the kth layer of an n-layer 3D chip [Im00] (where thermal resistance and power dissipation of the ith layer is R_i and P_i respectively, excluding the interconnect heating).

$$\Delta T_k = \sum_{i=1}^{k} R \sum_{x=i}^{n} \frac{P_x}{A}$$

Temperature Estimation

While, these temperature expressions provide highly accurate temperature predictions, fast thermal simulation of points in a 3D stack requires effective approximation to the proposed equations. In general, temperature is heavily dependent on the power density. Hence the 3D circuit temperature can be estimated as follows: T= P.R= P*(t/k*A) = d*(t/k) where t is the thickness of the chip, k is the thermal conductivity of the material; R represents the thermal resistance and d is the power density [Hung06]. Next, we discuss the various temperature modeling techniques for three-dimensional integrated circuits:

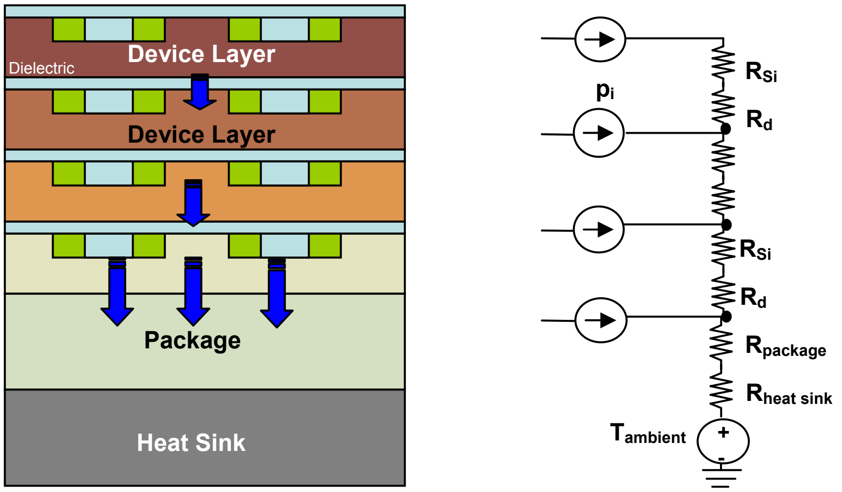


Figure 5. The heat flow in a 4-layer stacked IC and the corresponding analogous electrical network. P_i power dissipation of device layer i, thermal resistances for silicon, dielectric, package and heat sink.

5.2.1. Resistor-Capacitor Models

RC models are based on the analogy between electrical and thermal phenomena where: the heat flow is analogous to the electrical current passing through the thermal resistance. Thus the temperature difference between two nodes is equivalent to the voltage. The thermal capacitance is used to capture the delay before a change in power results in the thermal steady state. Figure 5 illustrates the analogous RC circuit for a three-dimensional integrated circuit. Each device layer is represented with the corresponding R_{si} and dielectric layers with R_d. It is important to note that the thermal resistance on the path to the heat sink increases considerably with each additional layer. Most dielectrics as well as silicon have relatively low heat conductance.

Hotspot-3D is an academic tool proposed by [Link06], which models the thermal behavior of the 3D ICs using the analogous RC circuits. It is a faster and improved version of the existing HotSpot 2.0 libraries including a multi-layer thermal analysis [HotSpot2.0]. The baseline resistor-capacitor modeling is largely unchanged while it extends to multiple layers of silicon in the HS3D version. Furthermore, thermal vias are modeled by varying the thermal resistance of materials. One of the important findings HotSpot3D enables is that, temperature profile remains relatively consistent regardless of the number of layers or the distribution power between layers. As a result of the minimal temperature variation between layers thermal-aware floorplanning provides minimal benefit. Another important point the study makes is on interconnect-driven folding in 3D designs. The experimental results provided in this study show that folding units onto multiple layers in order to minimize delay can cause higher peak temperatures for thermally aggressive blocks.

5.2.2. Finite Element Analysis Models

In finite element based analysis the existing design space is decomposed into discrete points, then meshed into elements. Temperature calculation is performed for these discrete points. The temperatures within the element are then interpolated using weighted averages of the calculated points, as illustrated in Figure 6. (Any temperature within the cube can be approximated through the edge temperatures T_1-T_8)

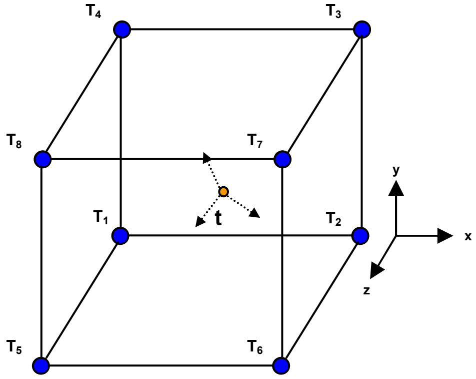


Figure 6. Finite element analysis model: each temperature t is represented as the weighted average of the calculated points {T_1, T_8}

Among the commercial FEA tools, Wilkerson et al. propose ACE+ that models the temperatures in a finite element based model solution [Wilkerson04]. While the HS3D puts special emphasis on the simulation speed, finite element based simulations of the CFD-ACE limits the number of thermal nodes to be analyzed as well as the thermal sampling interval due to its computational complexity. As a result of this accuracy-speed tradeoff, FEA based thermal analysis tools are mostly employed for static thermal analysis whereas HotSpot-based tools can be used for dynamic temperature variations.

Even though thermal modeling in both studies were validated using finite-element based analysis [Flotherm], these findings are quite different from [Cong06]s that uses CFD-ACE+ finite-element based tool in [Wilkerson04]. In the latter study experimental results indicate that the thermal profile changes significantly with the number of layers and block placement on various layers. The differences might be due to power modeling differences, as well as baseline process technology assumptions. In fact, HS3D uses Wattch [Wattch2000] for architecture-level power modeling. It is important to note that Wattch power models from 2D version is highly inaccurate for 3D, as the output capacitances are strongly interconnect dependent which change significantly in the multi-layer version.

5.2.3. Alternative Models

Other thermal modeling tools for multi-layer ICs were proposed in recent years. One such study is [Chiang01] SPICE-based tool with significantly less running time compared to finite-element based modeling. However, despite the shorter running time the thermal modeling accuracy is very high. This model is also capable of analyzing the interconnect temperature as well as modeling the effect of thermal via separation on the thermal conductivity of inter-layer dielectrics. The results were verified using finite element based model. Experimental analysis indicate that thermal characteristics of global interconnects require special attention due to the limitations on via placement (i.e. via distances are far greater for global interconnects). Another interesting finding indicates that thermal problems associated to low-k dielectrics are not as bad as they were estimated earlier.

Thermal characteristics of the processor deteriorate with the increased number of layers. Up to 26K increase in temperatures was observed by [Puttaswamy_GLSVSI_06_I]. However, this data does not include the thermal dependency of leakage power, as well as the lack of detailed modeling for static power dissipation. Thus the increase in the absolute temperature numbers is considerably limited in accuracy.

6. Architectural Exploration

Stacking versus Splitting?

Partitioning the processor resources to multiple silicon layers can be achieved at various granularities. At one end of the spectrum: partitioning happens at transistor-level, the pull-up PMOS transistors and pull-down NMOS transistors of the SRAM cells are separated to different layers. At the other extreme, the processor core is stacked with a memory layer on top, without partitioning the core resources to different layers. It is important to note that, the benefit of each technique depends on the underlying 3D fabrication technology (inter-layer interconnect bandwidth and latency, etc) as well as properties of the architectural block. An important determinant for partitioning efficiency is naturally the block characteristics:

- Interconnect limited larger structures (such as caches) benefit significantly from splitting the resources to multiple layers at finer granularity,
- Splitting does not provide significant benefit for non-interconnect limited structures (such as arithmetic logic units) where stacking may be more effective.

6.1. Processor Level Analysis

Even at processor level stacking three-dimensional integration provides promising improvement over planar counter parts. [Kleiner96] investigate the effects of moving the second-level cache as well as main memory on chip through the use of 3D technology. The analytical models on a RISC processor baseline indicate performance improvements in the order of 20-25% in average time per instruction over the conventional planar integration. In another study the second-level cache access time was reduced by 30% for the RISC processor/cache system [Kuhn96]. Another architectural implementation was conducted by Intel [Black04] on a deeply pipelined high performance x86 architecture. The results show a 15% improvement in performance as well as a simultaneous improvement in power dissipation by 15%.

6.2. Block Based Analysis

Multi-layer implementation can benefit the performance of individual blocks as well as the overall processor performance. The studies we discuss in this section target benefits of vertical integration on individual processor blocks such as caches, instruction schedulers, register file etc. As we discussed earlier, individual blocks vary significantly in terms of the benefit they see from 3D implementation.

In general, larger blocks that are strongly interconnect/wire dominated seem to be the best candidates for multi-layer implementation. Logic intensive blocks such as arithmetic logic units are shown to see minimal performance benefit 1-3% according to [Puttaswamy_ISCAS_06]. A similar study on dynamic instruction scheduler found that depending on the number of layers the latency improvement can be up to 12% with 22% reduction in the energy dissipation for 2-5 silicon layers [Puttaswamy_GLSVLSI_06_II].

6.2.1. Register File

Register files are critical structures with strong impact on microprocessor performance and cycle time. Furthermore, they are heavily ported and dominated by the associated interconnect. Hence, they benefit considerably from multi-layer stacking and splitting techniques. The benefit comes mostly from the reduction in the longer interconnect that runs through the registers and decoder/sense amplifiers.

Even though both caches and register files are SRAM structures, register files have larger footprint due to the higher port count. Hence finer granularity splitting techniques such as the ones that target port splitting are feasible for register files. 3D register files were studies by [Tremblay95]. Up to 6 times area reduction for an 8 window and 10 ported register file was shown. The access time is also reduced due to shorter bus lines and buffer sharing between cells. However, since power and temperature modeling was not available at the time, this study does not report any benefits on various folding schemes and their power/thermal advantages.

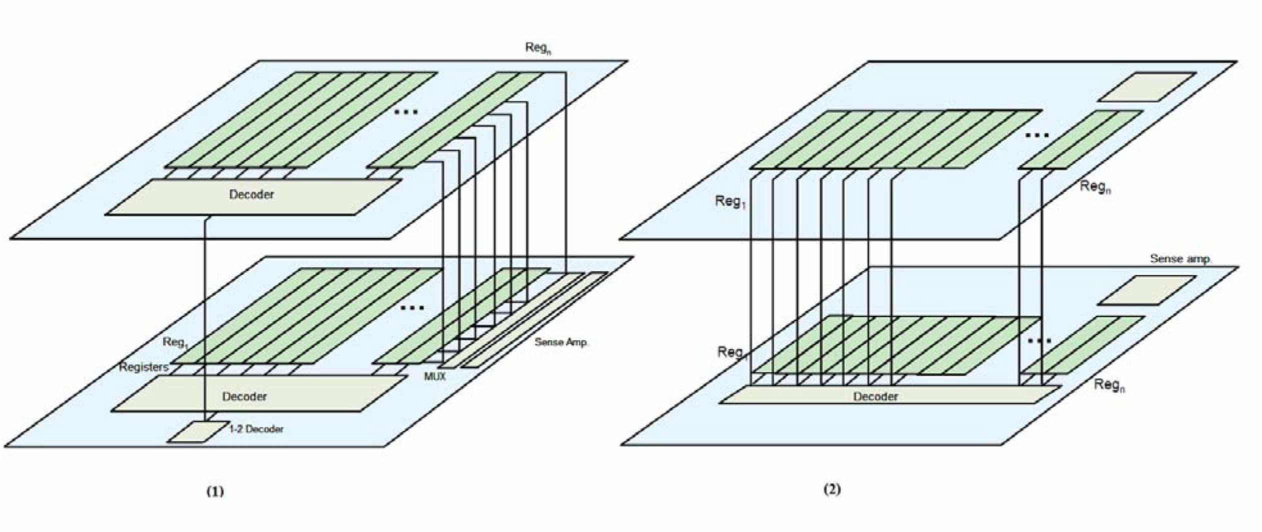


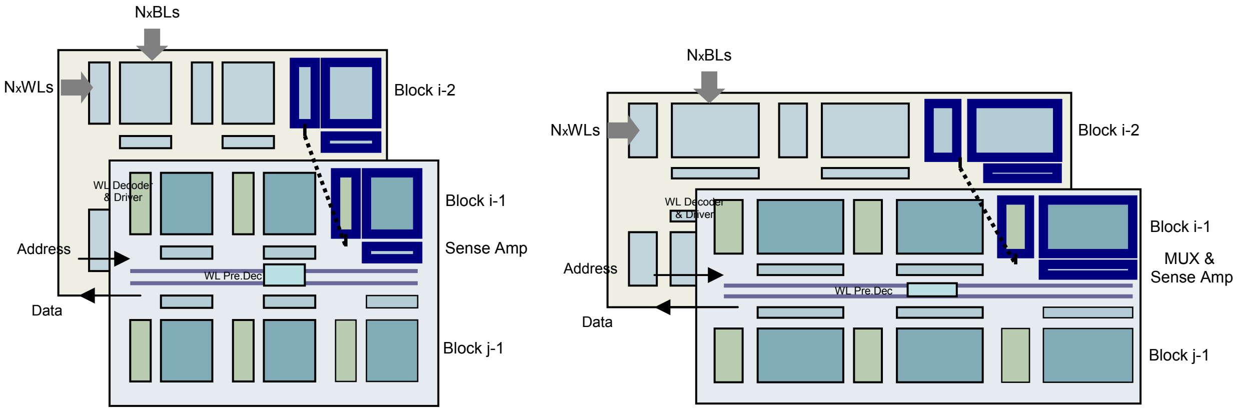
Figure.7. Register and bit partitioning in a two-layer architecture

[Puttaswamy_ISVLSI06] investigates a similar scenario with various folding techniques ranging from register partitioning to bit or port partitioning. Their analysis indicates that 24% latency improvement with above 50% energy reduction for a 256 entry register file.

6.2.2. Memory Hierarchy

The memory hierarchy displays the perfect range and capabilities of 3D integration technology. The partitioning can happen at various granularities ranging from individual transistors in an SRAM cell to incorporating a layer of memory on top the processor device layer. Alternatively the cache wordlines or bitlines can be assigned to different active device layers.

Figure 8 illustrates the wordline and bitline partitioning schemes for multi-layer caches. In wordline partitioning the local wordline decoder is used to feed the partitioned wordline drivers on multiple-layers. As a result, the delay of wordline access is reduced with the decrease in the number of transistors connected to the wordline driver in the specific layer. Furthermore, since the overall area is reduced, the distance from the address line to the wordline decoder is smaller by the number of device layers as well. In the alternative scheme of bitline folding, bitline length along with the pass transistors connected to a single bit is reduced. The sense amplifiers can be duplicated or shared between layers.



(1) Wordline **(2) Bitline**

Figure.8. 3D cache (1) wordline partitioning (2) bitline partitioning

Due to their regular structure, interconnect limitations and sheer sizes caches provide much higher benefit from 3D compared to other structures (such as scheduler, arithmetic logic unit). However, the savings in power and delay depend on the cache size, number of device layers, process technology and system requirements [Tsai05]. [Puttaswamy05] shows that 20% latency reduction along with 30% energy improvement is possible for a 512KB cache. Planar and 3D implementations of 16MB cache (in 130nm) is compared [Zeng05]. The resulting delay reduction in 3D implementation is around 40%. Vertical integration also yields almost 12% IPC gain even when the stacking is limited to caches in the microprocessor architecture. [Reed05] shows the benefits of multi-level cache implementation at data bank level. The resulting design has 15% area savings along with 25% reduction in the active power dissipation.

Memory hierarchy is consuming an increasing amount of space on a microprocessor with each generation. By stacking the main memory on top of the microprocessor layer [Kleiner96] observes performance improvements in the order of 20-25% in average time per instruction over the 2D implementation. The first tier of the chip is the microprocessor, second and third layers consist of second-level cache and the DRAM main memory. In general thermal management of the upper layers in vertically integrated chips is challenging. However, since the power density of the second-level cache and memory are quite low (0.75 W/cm^2 and 0.15 W/cm^2 respectively, compared to the 15W/cm^2 of the processor layer) the resulting design had favorable thermal characteristics.

A similar study [Liu05] looked at incorporating memory on chip along with the caches. Since 3D integration enables integration of more functionality than planar designs, it is possible to incorporate a larger L2 cache or increase the cache hierarchy with a larger L3 cache. [Liu05] observed significant performance improvement on most benchmarks with an on-chip L3 cache. Especially benchmarks with frequent L2 cache misses such as ‘mcf’ in SPEC2000 benchmark suite display the highest performance boost.

Recently [Black06] studied thermal and performance of 3D face-to-face technology, through SRAM/DRAM stacking on a microprocessor as well as implementing multi-layer version of a traditional micro architecture. The results on multi-layer implementation of a planar architecture show 15% performance improvement along with 15% reduction in power dissipation. Maximum on-chip temperature was only increased by 14^{o}C. Furthermore, through stacking a 32MB DRAM, cycles per memory access was reduced by 13% on average. Stacking an additional DRAM layer did not cause any noticeable temperature degradation.

Current multi-core architectures share a common bus to the memory. As the number of processing elements connected to the same memory increase the bus bottleneck is becoming more problematic. [Lee00] proposes a multi-port memory which acts as a real shared memory between processing units. The memory unit can be placed in the top layer of the 3D architecture. As all the aforementioned studies illustrate, three-dimensional integration provides unique advantages for implementation of caches and the rest of the memory hierarchy. Even though the benefit may vary with the cache organization, size, granularity of partitioning etc., each cache structure saw improvement over the planar counterpart.

7. Future Three-Dimensional Micro-Architectures

7.1. Truly 3D Architectures

One of the key challenges in three-dimensional microprocessor architectures is to utilize the benefits of underlying process technology (in terms of interconnect delay, bandwidth, power and thermal profile, layout limitations). The current exploratory 3D design philosophy is mostly restricted to mapping the existing 2D processor designs on a multi-layer silicon technology. However, the resulting architectures are not well tailored towards the characteristics of the baseline. Hence the final implementation is sub-optimal in terms of effectively incorporating the benefits of 3D. Truly 3D microprocessors are likely to be shaped by a number of factors:

- Adaptation to the underlying structure at every design granularity (Such as dividing the functionality among device layers possibly even at the finest granularities)
- Increased packaging density and connectivity is likely to reduce the on-chip critical paths (possible reduction in pipeline stages. Current designs with architectural blocks of limited connectivity may evolve into tightly connected blocks, shorter stages etc)
- Addition processor blocks (such as on-chip controllers, snap-on auxiliary engines to help with performance, security tasks, software bugs etc)
- Memory hierarchy and organization is likely to be effected by 3D (such as additional cache levels, increased memory bandwidth, changes in communication between cores/caches/memory, organization of memory)

The underlying 3D architecture imposes a number of critical design constraints on higher-level design. These constraints range from the structure of the cache hierarchy, to partitioning decisions for individual blocks and inter-block communication constraints. [Ozturk06] provides an interesting analysis of architectural decisions towards truly 3D architectures. This study investigates optimal placement of processor cores and storage blocks for minimum access time under temperature constraints. The experimental results indicate the extent that 3D specific architecture design enables further optimization. There is a strong need for understanding and exploration of the baseline technology so that it can be exploited at higher level decisions such as architecture level. This requires fast and efficient architectural and design automation tools as well as better, intuitive understanding of 3D.

7.2. Multi-Core and Multi-Threading

Multi-core architectures have gained increasing popularity in the recent years, thanks to their reduced design complexity, promising scaling ability, and their natural fit to the multi-programmed and throughput driven applications. As a result, architectures such as IBM Cell, Sun Niagara have been introduced in the past years. Current predictions indicate that future microprocessor architectures will likely have increased number of cores.

Three-dimensional integration provides number of benefits for the implementation of multi-core architectures

- Increased packaging density - enables incorporating higher number of cores on the chip (and/or more complicated cores may be possible as well with this technology)
- Vertical integration provides reduced latency and increases bandwidth for the memory hierarchy
- 3D enables further improvement in the topology and organization of the multi-core (multiple cores may share a set of performance enhancers and caches, thanks to the increased connectivity)

It is crucial to analyze the effects of 3D on CMP architectures with higher number of processor cores. The critical design decisions such as cache hierarchy and inter-core communication will greatly be effected by 3D process. The only study on chip multi processor implementation in 3D we are aware of at this point is [Li06], which focuses on the second level caches in CMP architectures in 3D. The proposed router and design topology utilizes the network architecture of L2 cache memory. Furthermore, the experimental analysis of multi-dimensional implementation of the L2 cache and CMP architecture provide promising results. There is very limited amount of research on potential benefits of 3D technology on multi-threaded/multi-core architectures. Future microprocessors are expected to have increased number of cores and threads; hence there is a strong need for research studies that provide better understanding of 3D in these scenarios.

7.3. Power/Thermal Modeling and Management

Power and thermal issues have already become key constraints for processor design. As a result of the higher transistor density 3D architectures are challenged by thermal problems even though the interconnect power dissipation is reduced. It has been shown by various studies ([Puttaswamy06]-[Hua06]), that thermal problems will unlikely be deal-breakers for 3D technology. Yet there is no doubt that 3D architectures will be thermally challenged and on-chip temperatures need to be managed more aggressively. Dynamic thermal management techniques targeting 3D architectures (most likely in multi-core CMP versions) are needed to effectively manage the on-chip temperatures. Both local and global management of on-chip resources on multiple device layers (in terms of power dissipation, performance, on-chip temperatures) is critical.

In order to achieve these goals fast and accurate power and temperature modeling tools are needed especially at higher-level design stages. Current planar tools need to be extended for 3D as well as verified with real on-chip experiments. Without the existence of such tools the thermal analysis and research opportunities are limited.

7.4. Heterogeneous Integration

Heterogeneous integration of different materials (such as Si, SiGe, GaAs) and disparate technologies (such as analog, RF, CMOS, nano-tubes) provide exciting opportunities for microprocessor design. Three-dimensional integrated circuits may provide less expensive alternative to the costly and complicated SoC process.

Figure 9 illustrates a scenario with DRAM memory integrated along with the processor and layers of analog and optical I/O devices. In this specific example the processor layer can be consisting of larger PPE cores as in the IBM Cell processor. The accelerator layer can incorporate number of SPE units that can effectively enhance the performance of the bottom layer. The DRAM memory sits on top of the two processing layers, providing reduced interconnect latency and improved bandwidth. The top layer is dedicated to specialized I/O.

Alternatively, specialized recovery and management units can be placed at the top layer of the device stack. As proposed by [Mysore06] this top layer can be snapped-on through micro-bumps, and can assist the processor with security issues/bugs as well as performing run-time optimizations on the hardware. Incorporating field programmable gate arrays in the structure for dynamic tuning of resources is yet another possibility, where the top layer can configure itself to enhance the performance by adapting to the changes in the application.

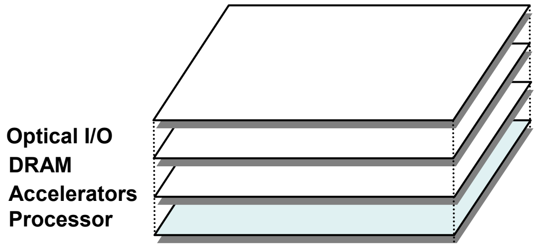


Figure 9. 3D processor chip stacked with memory, accelerator units and optical I/O layers

8. Conclusion

3D integration is considered to be a promising technology which can potentially help alleviate a number of problems modern ULSI circuits face. The advantages include and not limited to:

- Interconnect length and complexity reduction
- Increased packaging density
- Improved connectivity within the chip (in terms of number of devices connected and delay)

- Enhanced noise tolerance
- Heterogeneous integration of disparate technologies and materials
- Interconnect power reduction (dynamic power reduction as a result)

On the other hand 3D technology also causes chip temperatures to increase, along with additional manufacturing stages with possible complications (such as substrate alignment) [Islam02]. Recent studies such as [Black06] observe that temperature increase in 3D architectures is consistent, yet manageable. Their experimental results indicate only 14^{o}C increase for a 2-layer F2F implementation of an existing architecture, with maximum hotspot temperatures below 125^{o}C threshold. Manufacturing cost and time-to-market are likely to be the other determining factors for 3D ICs wide acceptance in microprocessor architectures.

3D integration has already expanded to many research fields, and is far from being considered as a solely packaging/manufacturing issue. However, most of the research efforts in 3D integration have been on fabrication technology and design automation tools/techniques. Design automation flow targeting planar technology has to be updated for successful adaptation of 3D.

The recent efforts in architectural exploration have been quite restricted due to the limitations of the lower level design tools and techniques. Furthermore, these techniques mostly consider the case that take the existing microprocessor architectures at block or core level and simply mapping onto multiple device layers. The resulting architecture is not designed to utilize the characteristics of the baseline 3D technology. Hence it is highly sub-optimal. Although the current mappings of existing microprocessor architectures/blocks to multiple layers show performance, packaging density and even power dissipation benefit, there is much room for further optimization. We believe that architectures need to be designed targeting specifically the characteristics of 3D for better utilization of the underlying technology. The need for architectural tools such as 3D dynamic and leakage power optimization is strong. The baseline characteristics of 3D technology will guide architecture-level decisions.

Even though the main benefit of 3D is probably alleviating existing issues in interconnect limitations, it has potential for shaping the next generation multi-core multi-threaded microprocessors. Novel architectures with multiple layers of processor, memory, accelerators of various materials (such as Si, SiGe etc) provide new and exciting opportunities for the evolution of microprocessor architectures in the next generations.

References

[Asaka86] Y. Asaka, "Three-Dimensional IC Trends", Proceedings of the IEEE, Vol.74, Issue.12, pp.1703-1714, 1986

[Akasaka86] Y. Akasaksa, T. Nishimura, "Concept and Basic Technologies for 3-D Integrated Circuit Structure", IEEE International Electron Device Meeting Technical Digest, pp.488-491, 1986

[Al-Sarawi98] S.F. Al-Sarawi, D. Abbot, P. D. Franzon, “A Review of 3-D Packaging Technology”, IEEE Transactions on Components, Packaging, and Manufacturing Technology: Advanced Packaging, Vol.21, Issue 1, pp. 2-14, 2998

[Banerjee96] K. Banerjee, A. Amerasekera, G. Dixit, C.Hu, “The Effect of Interconnect Scaling and Low-K Dielectric on the Thermal Characteristics of the Integrated Circuit Metal”, IEEE International Electron Device Meeting Technical Digest, pp.65-68, 1996

[Banerjee01] K. Banerjee, S. Souri, P. Kapur, K. Saraswat,“3-D ICs: A Novel Chip Design For Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration, Proceedings of IEEE Vol.89, No.5, 2001

[Bernstein03] K. Bernstein, C. T. Chuang, R. Joshi, R. Puri, “Design and CAD Challenges in Sub-90nm CMOS Technologies”, International Conference on Computer Aided Design, pp.129-136, 2003

[Bernstein06] K. Bernstein, “Introduction to 3D Integration”, International Solid State Circuits Conference Tutorial, 2006

[Beyne04] E. Beyne, “3D Interconnection and Packaging: Impending Reality of Still a Dream?”, IEEE International Solid State Conference, 2004

[Black04] B. Black, D.W. Nelson, C. Webb, N. Samra, “3D Processing Technology and Its Impact on iA32 Microprocessors”, International Conference on Computer Design, pp.316-318, 2004

[Black06] B. Black, M. M. Annavaram, E. Brekelbaum, J. DeVale, L. Jiang, G. H. Loh, D. McCauley, P. Morrow, D. W. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. P. Shen, C. Webb, “Die Stacking (3D) Microarchitecture”, International Symposium on Microarchitecture, 2006

[Bohr95] M. T. Bohr, “Interconnect Scaling – The Real Limiter to High Performance ULSI”, IEEE Electron Devices Meeting, pp.241-244, 1995

[Burns00] J. Burns, L. McIlrath, J. Hopwood, C. Keast, D. P. Vu, K. Warner, P. Wyatt, “An SOI-Based Three-Dimensional Integrated Circuit Technology”, IEEE International Silicon on Insulator Conference, pp.20-21, 2000

[Burns01] J. Burns, L. Mcllrath, C. Keast, C. Lewis, A. Loomis, K. Warner, P. Wyatt, “Three Dimensional Integration for Low Power, High-Bandwidth Systems on a Chip”, IEEE International Solid State Circuits Conference, 2001

[Burns06] J.A. Burns, C. K. Chen, J. M. Knect, P. W. Wyatt, “A Wafer-Scale 3-D Circuit Integration Technology”, IEEE Transactions on Electron Devices, Vol.53, No.10, pp. 2507-2516

[Cacti3.0] P. Shivakumar, N. P. Jouppi, “Cacti 3.0: An Integrated Cache Timing, Power, and Area Model”, Western Research Lab, Research Report, 2001/2.

[Chan00] V. W. C. Chan, P. C. H. Chan, M. Chan, “Three Dimensional CMOS Integrated Circuits on Larger Grain Polysilicon Films”, IEEE Electron Device Meeting, pp.169-172, 2000

[Chan2001] M. Chan, “The Potential and Realization of Multi-Layers in Three Dimensional Integrated Circuits”, IEEE International Conference on Solid State and Integrated Circuit Technology, Vol.1, pp. 40-45, 2001

[Chiang01] T.-Y. Chiang, K. Banerjee, K. C. Saraswat, “A New Analytical Thermal Model for Multilevel VLSI Interconnects Incorporating Via Effects”, IEEE International Interconnect Technology Conference, pp.92-94, 2001

[Cong05] J. Cong, Y. Zhang, “Thermal via Planning for 3-D ICs”, Asia South Pacific Design Automation Conference, 2005.

[Cong06] J. Cong, A. Jagannathan, Y. Ma, G. Reinman, J. Wei, Y. Zhang, ”An Automated Design Flow for 3D Microarchitectural Evaluation”, Design Automation Conference Asia and South Pacific, pp.384-398, 2006.

[Das03] S. Das, A. Chandrakasan, R. Reif, “Design Tools for 3-D Integrated Circuits”, Proceedings of the 2003 Design Automation Conference Asia and South Pacific, pp.53-56, 2003

[Das04] S. Das, A. Chandrakasan, R. Reif, “Timing, Energy and Thermal Performance of Three-Dimensional Integrated Circuits”, Proceedings of Great Lakes Symposium on VLSI, pp.338-343, 2004

[Davis01]J. A. Davis, R. Venkatesan, A. Kaloyeros, M. Beylansky, S. J. Souri, K. Banerjee, K. C. Saraswat, A. Rahman, R. Reif, and J. D. Meindl, “Interconnect Limits on Gigascale Integration (GSI) in the 21st Century,” Proc. IEEE 89, 305 (2001)

[Davis05]W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, P. D. Franzon, ”Demystifying 3D ICs: The Pros and Cons of Going Vertical”, IEEE Design and Test of Computers, pp.498-510, 2005

[Deng03] Y. Deng, W. Maly, “Physical Design of the 2.5D Stacked System”, IEEE International Conference on Computer Design, 2003

[Deng04] Y.Deng, W.Maly, “2.5D System Integration: A Design Driven System Implementation Schema”, Asia and South Pacific Design Automation Conference, 2004

[Dong06] S.Dong, S.Zheng, X.Hong, “Floorplanning for 2.5D System Integration Using Multi-Layer-BSG Structure”, IEEE International Symposium on Circuits and Systems, 2006

[Drost04] R.J.Drost, R.D.Hopkins, R.Ho, I.E.Sutherland, “Proximity Communication”, IEEE Journal of Solid-State Circuits, Vol.39, No.9, pp.1529-1535, 2004

[Flotherm] Flometrics Corporation, Flotherm Modeling Software

[Forthun92]J. Forthun and C. Belady, “3-D Memory for Improved System Performance”, Proceedings of the International Electron Packaging Conference, 1992, p. 667.

[Geis79] M. W. Geis, D. C. Flanders, D. A. Antoniadis, H. I. Smith, “Crystalline Silicon on Insulators by Graphoepitaxy”, IEEE International Electron Device Meeting, 1979

[Goplen05] B. Goplen, S. Sapatnekar, “Thermal Via Placement in 3D ICs”, International Symposium on Physical Design, pp.167-174, 2005

[Guerrier00] P. Guerrier, A. Greiner, “A Generic Architecture for On-Chip Packet-Switched Interconnections”, Proceedings of Design Automation and Test in Europe, pp.250-256, 2000

[Ho01] R. Ho, K. W. Mai, M.A. Horowitz, “The Future of Wires”, Proceedings of the IEEE, Vol. 89, No.4, pp.490-504, 2001

[HotSpot2.0] K. Skadron, M. R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, D. Tarjan, “Temperature-Aware Micro-architecture: Modeling and Implementation”, ACM Transactions on Architecture and Code Optimization, Vol.1, No.1, pp. 94-124, 2004

[Hua06] H. Hua, C. Mineo, K. Schoenfliess, A. Sule, S. Melamed, “Exploring Compromises among Timing, Power and Temperature in Three-Dimensional Integrated Circuits”, ACM Design Automation Conference, pp. 997-1002, 2006

[Hua2006] H. Hua, C. Mineo, K. Schoenfliess, A. Sule, S. Melamed, W. R. Davis, "Performance Trends in Three-Dimensional Integrated Circuits", IEEE International Interconnect Technology Conference, pp.45-47, 2006

[Hung06] W.L. Hung, G. M. Link, Y. Xie, N. Vijaykrishnan, M. J. Irwin, "Interconnect and Thermal-Aware Floorplanning for 3D Microprocessors", IEEE International Symposium on Quality Electronic Design, 2006

[Ieong03] M. Ieong, K. Guarini, V. Chan, K. Bernstein, R. Joshi, J. Kedzierski, W. Haensch, IEEE Custom Integrated Circuits Conference, pp.10.2.1-7, 2003

[Im00] S. Im, K. Banerjee, "Full Chip Thermal Analysis of Planar (2-D) and Vertically Integrated (3-D) High Performance ICs", IEEE Electron Devices Meeting, Technical Digest, pp.727-730, 2000

[Islam02] R. Islam, C. Brubaker, P. Lindner, C. Schaefer, "Wafer Level Packaging and 3D Interconnect for IC technology", Proceedings of Advanced Semiconductor Manufacturing, pp.212-217, 2002

[ITRS2005] International Technology Roadmap for Semiconductors, 2005 Edition, Interconnect Report, http://www.itrs.net/Links/2005ITRS/Interconnect2005.pdf

[Joyner01] J. W. Joyner, R. Venkatesan, P. Zarkesh-Ha, J. A. Davis, J. D. Meindl, "Impact of Three-Dimensional Architectures on Interconnects in Gigascale Integration", IEEE Transactions on Very Large Scale Integration Systems, Vol.9, No.6, pp. 922-928, 2001

[Jung04] S-M. Jung et.al. "The Revolutionary and Truly 3-Dimensional 25F2 SRAM Technology with the Smallest S3 Cell and SSTFT for Ultra High Density SRAM", VLSI Technology Digest of Technical Papers, 2004

[Kanda03] K. Kanda, D. D. Antono, K. Ishida, H. Kawaguchi, T. Kuroda, T. Sakurai, "1.27Gb/s/pin 3mW/pin Wireless Superconnect (WSC) Interface Scheme", IEEE International Solid-State Circuits Conference, Digest of Technical Papers, Vol.1, pp.186-187, 2003

[Kawamura83] S. Kawamura, N. Sasaki, T. Iwari, M. Nakano, M. Takagi, "Three-Dimensional CMOS ICs Fabricated by using Beam Recrystallization", IEEE Electron Device Letters, pp.366-369,1983

[Kim03] N. S. Kim, T. Austin, D. Blaauw, T. Mudge, K. Flautner, J. S. Hue, M. J. Irwin, M. Kandemir, V. Narayanan, "Leakage Current: Moore's Law Meets Static Power", IEEE Computer, Vol.36, Issue 12, 2003

[Kim05] S. K. Kim, C.C. Liu, L. Xue, S. Tiwari, "Crosstalk Attenuation with Ground Plane Structures in Three-Dimensionally Integrated Mixed Signal Systems", IEEE Microwave Symposium Digest, pp.4, 2005

[Kim2005] S. K. K. Kim, S. Tiwari, "Low Temperature Wafer-Scale 3D ICs: Technology and Characteristics", IEEE International Conference on Integrated Circuit Design and Technology, pp.183-186, 2005

[Kleiner96] M. B. Kleiner, S.A. Kuhn, P. Ramm, W. Weber, "Performance Improvement of the Memory Hierarchy of RISC-Systems by Application of 3D Technology", IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part B, Vol.19, No.4, pp.709-717, Nov, 1996

[Kleiner95] M. B. Kleiner, S. A. Kuhn, P. Ramm, W. Weber, "Thermal Analysis of Vertically Integrated Circuits", IEEE Electron Devices Meeting, pp.487-490, 1995

[Kobrinsky04] M. J. Kobrinsky, B. A. Block, J. F. Zheng, B. C. Barnett, E. Mohammed, M. Reshotko, F. Robertson, S. List, I. Young, K. Cadien, "On-Chip Optical Interconnects", Intel Technology Journal, Vol.8, Issue 02, 2004

[Koh00] C.K.Koh, P.H.Madden, "Manhattan or non-Manhattan? A Study of Alternative VLSI Routing Architectures", Proceedings of Great Lakes Symposium on VLSI, pp.47-52, 2000

[Koo05] J. M. Koo, S. Im, L. Jiang, K. E. Goodson, "Integrated Micro-channel Cooling for Three-Dimensional Electronic Circuit Architects", American Society of Mechanical Engineers, Journal of Heat Transfer, Vol.127, Issue 1, pp.49-58, 2005

[Kuhn95] S.A. Kuhn, M. B. Kleiner, P. Ramm, W. Weber, "Interconnect Capacitances, Crosstalk and Signal Delay in Vertically Integrated Circuits", IEEE International Electron Device Meeting Technical Digest, pp.487-490, 1995

[Kuhn96] S. A. Kuhn, M. B. Kleiner, P. Ramm, W. Weber, "Performance Modeling of the Interconnect Structure of a Three-Dimensional Integrated RISC Processor/Cache System", IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part B, Vol.19, No.4, Nov, pp.719-727, 1996

[Kunio89] T. Kunio, K. Oyama, Y. Hayashi, M. Morimoto, "Three Dimensional ICs, Having Four Stacked Active Device Layers", IEEE International Electron Devices Meeting, 34.6.1-4, 1989

[Lee00] K. W. Lee, T. Nakamura, T. Ono, Y. Yamada, T. Mizukusa, H. Hashimoto, K.T. Park, H. Kurino, M. Koyanagi, "Three-Dimensional Shared Memory Fabricated Using Wafer Stacking Technology", IEEE Electron Devices Meeting, pp.165-168, 2000

[Li06] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, V. Narayanan, M. Kandemir, "Design and Management of 3D Chip Multiprocessors Using Network-in-Memory", Proceedings of International Symposium on Computer Architecture, 2006

[Lin97] H. Y. Lin, C. Y. Chang, T. F. Lei, J. Y. Cheng, L. P. Chen, "Characterization of Polycrystalline Silicon Thin Film Transistors Fabricated by Ultra-High Vacuum Chemical Deposition and Chemical Mechanical Polishing", Japanese Journal of Applied Physics, Bol.36, pp.4278-4282, 1997

[Link06] G. M. Link, N. Vijaykrishnan, "Thermal Trends in Emerging Technologies", Proceedings of International Symposium on Quality Electronic Design, pp.8, 2006

[Liu05] C. C. Liu, J. H. Chen, R. Manohar, S. Tiwari, "Mapping System-on-Chip Designs from 2D to 3D ICs", IEEE International Symposium on Circuits and Systems, Vol.3, pp.2939-2942, 2005

[Loi06] G. L. Loi, B. Agrawal, N. Srivastava, S. C. Lin, T. Sherwood, K. Banerjee, "A Thermally-Aware Performance Analysis of Vertically Integrated Processor Memory Hierarchy", Proceedings of IEEE/ACM Design Automation Conference, pp.991-996, 2006

[Lu03] J. Q. Lu, A. Jindal, P. D. Persans, T. S. Cale, R. J. Gutmann, "Wafer-Level Assembly of Heterogeneous Technologies", International Conference on Compound Semiconductor Manufacturing, 2003

[Lu05] J. Lu, "Wafer-Level 3D Hyper-Integration Platform", Spring 2005.

[Meindl02] J. D. Meindl, J.A. Davis, P. Zarjesh-Ha, C. S. Patel, K. P. Martin, P. A. Kohl, "Interconnect Opportunities for Gigascale Integration", IBM Journal of Research and Development, Vol. 46, No.2/3, 2002

[Montecito05] B. Doyle, P.Mahoney, E. Fetzer, S.Naffziger, "Clock Distribution on a Dual-Core, Multi-Threaded Itanium Family Processor", IEEE International Conference on Integrated Circuit and Technology, 2005

[Morrow04] P. Morrow, M. J. Kobrinsky, S. Ramanathan, C. M. Partk, M. Harmes, V. Ramachandraro, H. M. Park, G. Kloster, S. List, S. Kim, "Wafer-level 3D Interconnects via Cu Bonding", Proceedings of the Advanced Metallization Conference, 2004

[Morrow06] P. R. Morrow, C. M. Park, S. Ramanathan, M. J. Kobrinsky, M. Harmes, "Three-Dimensional Wafer Stacking Via Cu-Cu Bonding Integrated With 65-nm Strained-Si/Low-k CMOS Technology", Proceedings of IEEE Electron Device Letters, Vol. 27, No. 5, 2006

[Mukai83] R. Mukai, N. Sasaki, T. Iwai, S. Kawamura, M. Nakano, ”Indirect Laser Annealing of Poly-silicon for Three-Dimensional IC's”, IEEE International Electron Devices Meeting, 14.4, 1983

[Mukherjee04] M. Mukherjee, R. Vemuri, “Simultaneous Scheduling, Binding and Layer Assignment for Synthesis of Vertically Integrated 3D Systems”, IEEE International conference on Computer Design, 2004

[Mukherjee05]M. Mukherjee, R. Vemuri, “On Physical-Aware Synthesis of Vertically Integrated 3D Systems”, IEEE International Conference on VLSI Design, 2005

[Mysore06] S. Mysore, B. Agrawal, N. Srivastava, S. C. Lin, K. Banerjee, T. Sherwood, “Introspective 3D Chips”, International Conference on Architecture Support for Programming Languages and Operating Systems, 2006

[Nakata82] J. Nakata, K. Kajiyama ,“Novel Low-Temperature Re-crystallization of Amorphous Silicon by High Energy Beam”, Applied Physics Letters, pp.686-688, 1982

[Neudeck99] G. W. Neudeck, S. Pae, J. P. Denton, T. Su, “Multiple Layers of Silicon-on-Insulator for Nanostructure Devices”, Journal of Vacuum Science and Technology-B, Vol.17, No.3, pp.994-998, 1999

[Neudeck00] G. W. Neudeck, T. Su, J. P. Denton, ”Novel Silicon Epitaxy for Advanced MOSFET Devices”, IEEE Electron Device Meeting, pp.169-172, 2000

[Pavlidis06] V. Pavlidis, E. Friedman, ”Via Placement for Minimum Interconnect Delay in Three-Dimensional Circuits”, IEEE International Symposium on Circuits and Systems, pp.4587-4590, 2006.

[Ozturk06] O. Ozturk, F. Weng, M. Kandemir, Y. Xie, “Optimal Topology Exploration for Application Specific 3D Architectures”, IEEE Asia South Pacific Design Automation Conference, pp.390-395, 2006

[Patti06] R. S. Patti, “Three-Dimensional Integrated Circuits and the Future of System-on-Chip Designs”, Proceedings of IEEE, Vol.94, Issue.6, 2006

[Puttaswamy05] K. Puttaswamy, G. H. Loh, “Implementing caches in a 3D Technology for High Performance Processors”, International Conference on Computer Design, ICCD, pp.525-532, 2005

[Puttaswamy_GLSVLSI_06_II] K. Puttaswamy, G. H. Loh, ”Dynamic Instruction Schedulers in a 3-Dimensional Integration Technology”, ACM/IEEE Great Lakes Symposium on VLSI, pp.153-158, 2006

[Puttaswamy_GLSVLSI_06_I] K. Puttaswamy, G. H. Loh, ”Thermal Analysis of a 3D Die-Stacked High Performance Microprocessor”, ACM/IEEE Great Lakes Symposium on VLSI, pp.19-24, 2006

[Puttaswamy_ISCAS_06] K. Puttaswamy, G. H. Loh, “The Impact of 3-Dimensional Integration on the Design of Arithmetic Units”, IEEE International Symposium on Circuits and Systems, ISCAS, pp.4951-4954, 2006

[Puttaswamy_ISVLSI06] K. Puttaswamy, G. Loh, “Implementing Register Files for High Performance Microprocessors in a Die-Stacked (3D) Technology”, Proceedings of Emerging VLSI Technologies and Architectures, 2006

[Rahman00] A. Rahman, A. Fan, R. Reif, “Comparison of Key Performance Metrics in Two and Three Dimensional Integrated Circuits”, Proceedings of IEEE International Interconnect Technology Conference, pp.18-20, 2000

[Ranganathan05] N. Ranganathan, K. Prasad, N. Balasubramanian, Z. Qiaoer, S. C. Hwee, “High Aspect Ratio Through-Wafer Interconnect for Three-Dimensional Integrated Circuits”, IEEE Electronic Components and Technology Conference, pp.343-348, 2005

[Reed05] P. Reed, G. Yeung, B. Black, “Design Aspects of Microprocessor Data Cache Using 3D Interconnect Technology”, IEEE International Conference on Integrated Circuit Technology, 2005

[Reif02] R. Reif, A. Fan, K-N Chen, S. Das, "Fabrication Technologies for Three-Dimensional Integrated Circuits", International Symposium on Quality Electronic Design, 2002

[Rickert04] P. Rickert, "Problems or Opportunities?: Beyond the 90nm Frontier", Keynote Speech, International Conference on Computer Aided Design, 2004.

[Saraswat82] K. C. Saraswat, F. Mohammadi, "Effect of Interconnection Scaling on Time Delay of VLSI Circuits", IEEE Transactions on Electron Devices, Vol. ED-29, pp. 645-650, 1982

[Saraswat00] K. C. Saraswat, K. Banerjee, A. R. Joshi, P. Kalavade, P. Kapur, S. J. Souri, "3-D ICs: Motivation, Performance Analysis and Technology", Proceedings of European Solid State Circuits Conference, pp.406-414, 2000

[Schaper05] L. W. Schaper, S. L. Burkett, S. Sipesshoefer, G. V. Vangara, Z. Rahman, S. Polamreddy, "Architectural Implications and Process Development of 3-D VLSI Z-Axis Interconnects Using Through Silicon Vias", IEEE Transactions on Advanced Packaging, Vol 28, No.3, pp.356-366, 2005

[Souri00] S. Souri, K. Banerjee, A. Mehrotra, K. C. Saraswat, "Multiple Si Layer ICs: Motivation, Performance Analysis, and Design Implications", Proceedings of Design Automation Conference, 2000

[Subramanian91] C. K. Subramanian, G. W. Neudeck, "A Full-Wafer SOI Process For 3-Dimensional Integration", IEEE Microelectronics Symposium, pp.195-198, 1991

[Topol05] A. W. Topol, D. C. La Tulipe, L. Shi, S. M. Alam, D. J. Frank, S. E. Steen, J. Vichiconti, D. Posillico, M. Cobb, S. Medd, J. Patel, S. Goma, D. DiMilia, M. T. Robson, E. Duch, M. Farinelli, C. Wang, R. A. Conti, D. M. Canaperi, L. Deliginanni, A. Kumar, K. T. Kwietniak, C. D' Emic, J. Ott, A. M. Young, K. W. Guarini, M. Ieong, "Enabling SOI-Based Assembly Technology for Three-Dimensional (3D) Integrated Circuits (ICs)", International Electron Devices Meeting Technical Digest, pp.363, 2005

[Topol05_II]A. W. Topol, D. C. La Tulipe, L. Shi, S. M. Alam, A. M. Young, D. J. Frank, S. E. Steen, J. Vichiconti, D. Posillico, D. M. Canaperi, S. Medd, R. A. Conti, S. Goma, D. Dimilia, C. Wang, L. Deligianni, M. A. Cobb, K. Jenkins, A. Kumar, K. T. Kwietniak, M. Robson, G. W. Gibson, C. D'Emic, E. Nowak, R. Joshi, K. W. Guarini, and M. Ieong, "Assembly Technology for Three Dimensional Integrated Circuits," Proceedings of the VLSI/ULSI Multilevel Interconnection Conference, 2005, p. III-D.

[Topol06] A. W. Topol, D. C. La Tulipe Jr., L. Shi, D.J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini, M. Ieong, "Three-Dimensional Integrated Circuits", IBM Journal of Research and Development, Vol.50, No. 4/5, pp.491, 506, 2006

[Tremblay95] M. Tremblay, B. Joy, K. Shin, "A Three Dimensional Register File for Superscalar Processors", Proceedings of 28th Hawaii International Conference on System Sciences, HICSS 1995

[Tsai05]Y.- F. Tsai, Y. Xie, N. Vijaykrishnan, M.J. Irwin, "Three-Dimensional Cache Design Exploration Using 3DCacti", IEEE International Conference on Computer Design, October 2005

[Upadhya03] G. Upadhya, P. Zhou, K. Goodson, M. Munch, T. Kenny, "Closed-Loop Cooling Technologies for Microprocessors", IEEE International Device Meeting, 2003

[Viswanath00] R. Viswanath, W. Vijay, A. Watwe, V. Lebonheur, "Thermal Performance Challenges from Silicon to Systems", Intel Technology Journal Q3, 2000

[Warner04] K. Warner, C. Chen, R. D'Onofrio, C. Keast, S. Poesse, "An Investigation of Wafer-to-Wafer Alignment Tolerances for Three-Dimensional Integrated Circuit Fabrication", IEEE International Silicon on Insulator Conference, pp.71-72, 2004

[Wattch2000] D. Brooks, V. Tiwari, M. Martonosi, “WATTCH: A Framework for Architectural-Level Power Analysis and Optimizations”, Proceedings of International Symposium on Computer Architecture, pp.83-94, 2000

[Wilkerson04] P. Wilkerson, A. Raman, M. Turowski, ”Fast, Automated Thermal Simulation of Three-Dimensional Integrated Circuits”, Thermal and Thermomechanical Phenomena in Electronic Systems, ITHERM, pp. 706-713, Vol.1, June, 2004

[Wong06] E. Wong, J. Minz, S. K. Lim, ”Effective Thermal Via and Decoupling Capacitor Insertion for 3D System-on-Package”, IEEE Electronic Components and Technology Conference, 2006.

[Xie06] Y. Xie, G. H. Loh, B. Black, K. Bernstein, “Design Space Exploration for 3D Architectures”, ACM Journal of Emerging Technologies in Computing Systems, JETCS, Vol.2, pp. 65-103, 2006

[Xue01] L. Xue, C. Liu, S. Tiwari, “Multi-Layers with Buried Structures (MLBS): An approach to Three-Dimensional Integration”, IEEE International Conference on Silicon on Insulator, pp.117-118, 2001

[Xue03] L. Xue, C. C. Liu, H.S. Kim, S. Kim, S. Tiwari, ”Three-Dimensional Integration: Technology, Use and Issues for Mixed-Signal Applications”, IEEE Transactions on Electron Devices, Vol.50, No.3, pp. 601-609, 2003

[Yamauchi94] N.Yamauchi, “Polycrystalline Silicon thin Films Processed with Silicon Ion Implantation and Subsequent Solid-Phase Crystallization: Theory, Experiments and Thin-Film Transistor Applications”, Journal of Applied Physics, Vol.75, No.7, pp.3235-3257, 1994

[Zeng05] A. Zeng, J. Lu, K. Rose, R. J. Gutmann, “First-Order Performance Prediction of Cache Memory with Wafer Level 3D Integration”, IEEE Design and Test of Computers, pp.548-555, 2005

[Zhang01] R. Zhang, K. Roy, C-K. Koh, D. B. Janes, “Power Trends and Performance Characterization of 3-Dimensional Integration”, Proceedings of IEEE International Symposium on Circuits and Systems, Vol.4, pp.414-417, 2001