Demystifying 3D ICs: The pros and cons of going vertical

by W. Rhett Davis, John Wilson, Jian Xu, Lei Luo, Hao Hua, Ambarish Sule,
Christopher A. Mineo, Michael B. Steer, Paul D. Franzon
to appear in IEEE Design & Test of Computers, Nov. 2005

An increasing number of integrated solutions involve the stacking of chips to reduce system size. Wire-bonded stacks of processors and memories can be founds in cell-phones, personal digital assistants, and flash cards. But is the physical size of the system the only benefit of stacking chips? Are there potential performance benefits to be had from this miniaturization? Until recently, practical interconnecting of chips stacks could be done only by wire-bonding at the periphery, offering little or no benefits in the way of interconnect density or reduction of parasitics. But several new technologies offer the means to cost-effectively achieve very high densities of interconnect between chips in a stack, making true 3D ICs a reality. Designers of integrated circuits need to be aware of the benefits and drawbacks to these techniques so that they can decide if their systems would work better as a 3D IC.

This paper provides a practical introduction to the design trade-offs of the currently available 3D IC technology options. It begins with a summary of techniques, including wire-bonding, micro-bumps, through-vias, and contactless interconnect, comparing them in terms of vertical density and practical limits to their use. A high-level discussion of the pros and cons of 3D technologies is presented, with an analysis relating the number of transistors on a chip to the vertical interconnect density using estimated based on Rent's rule. Next, the paper provides a more detailed design example of inductively-coupled interconnect, with measured results of a system fabricated in a 0.35 μm technology and an analysis of misalignment and crosstalk tolerances. Lastly, a case-study of an FFT placed and routed in a 0.18 μm through-via SOI technology is presented, comparing the 3D design to a traditional 2D approach in terms of wire-length and critical-path delay.

Summary of Vertical Interconnect Technologies

Three-dimensional (3D) ICs offer an attractive alternative to two-dimensional (2D) planar ICs due to increased system integration by either increasing functionality and/or combining different technologies. Currently, system-on-chip solutions limit designers to one fabrication technology for both analog and digital circuits. The trend is to use inexpensive digital processes, with less than desirable performance for analog circuits, and off-load increased complexity to the analog designs. By using 3D ICs, the best technology for a particular portion of an application can be integrated into the chip cube.

Table 1 and Figure 1 show a summary of different 3D interconnect approaches, comparing them in terms of the method of assembly (die-scale or wafer-scale), maximum number of tiers (where the term "tier" refers to the chips in a stack, in order to differentiate them from the "layers" in a chip), the pitch of the vertical interconnect, and the amount of routing resources consumed on the chip. The most common is the wire-bonded approach, in which individual die are stacked and wire-bonded. Connections between chips must go through the board or chip-carrier and back to other chips in the stack. This approach is limited by the resolution of wire-bonders, 35μm using 15μm wire

[Kulicke & Sofa - *Maxμmplus*], and becomes increasingly difficult as the number of I/Os in the chip stack increase. Unlike the other 3D approaches, the wire-bonds are limited to the periphery of the chip, which severely limits the interconnect density. In terms of chip routing resources, all metal layers are typically needed for the bonding pads, because the mechanical stresses require many metal layers to prevent tearing of the pad during bonding, and devices underneath the pad tend to be destroyed by the pressure.

Micro-bump technology involves the use of solder or gold bumps on the surface of the die to make connections. These bumps have a typical pitch of 50-500μm but have been demonstrated with smaller pitches. The mechanical stresses of assembly are much lower than with wire bonding, and so pads require only the top or sometimes top two metal layers, leaving lower layers free for routing or devices. 3D Package technology [4] involves embedding previously fabricated die into a set of carrier-wafers with a fixed size, allowing them to be assembled into a tight cube. Each die-carrier tier is micro-bump bonded to an epoxy routing tier that brings signals to the edges of the cube. The tiers are then laminated into a single stack and metallization is added on the sides to connect the routing tiers. The 3D package approach offers a much greater vertical interconnect density than the wire-bonded approach, although parasitic capacitances are not significantly reduced, because signals must still be routed to the periphery before coming back to the destination inside the cube. The number of tiers with the 3D package approach is not limited by the assembly process, but is more likely to be limited by the heat inside the cube. This method enables one or more chips, from the same or different fabrication technology, to be used in each layer of the stack. Face-to-face Micro-bump technology [3], offers the ability to shorten the wires between tiers and improve performance by reducing parasitics. It was determined that with proper placement of blocks in the 3D architecture, the use of high-power dynamic logic circuits, repeaters, pipelined stages, and long routing paths could be reduced. This decreased overall power consumption by 15% while simultaneously increasing performance by 15%. This approach is limited to two tiers; however, in order to get connections off of the chip stack, this technology must be used in conjunction with a wire-bonded or through-via approach.

Through-Via interconnect has the potential to offer the greatest interconnect density but also the greatest cost. Assembly is performed at the wafer-level, with a second wafer placed face-down down on the first wafer (face-to-face), with subsequent wafers being placed down (face-to-back) as the number of tiers grows. Holes are then etched through the upper wafer into the lower wafer, and then filled with tungsten to provide connectivity. Before the next chip is placed, the backside of the previously etched chip is thinned by polishing. The top tier has tungsten vias that protrude along with cuts for bond pads that provide power, ground, and I/O connectivity. Like the 3D package approach, through via approaches are not limited in the number of tiers that can be assembled. However the heat inside the stack can limit the number of tiers. Also, since assembly cannot be performed with known-good-die, the yield with this approach drops quickly as more tiers are added. Through-vias have been demonstrated in bulk technologies [2] by first coating the hole with an insulator, achieving pitches of 50μm. Silicon on Insulator (SOI) technologies [1] avoid the need for passivating the hole by polishing the substrate away completely, down to the buried oxide. The SOI technologies have achieved the smallest inter-tier pitches yet, on the order of 5μm. As

for routing resources, note in the figure that the through-via approaches consume all layers in the upper tier in addition to the top layer in the lower tier.

Contactless or AC-coupled interconnect involves the use of capacitive or inductive coupling to communicate between tiers [5]. This eliminates the need for the processing steps used to create inter-tier DC connectivity and signals need not be routed to the periphery, allowing reduced wire-lengths. Also, the lack of specialized processing steps in the contactless approach makes it much cheaper than micro-bump and through-via approaches, since only a minimal amount of processing is required for chip thinning. Capacitive coupling [6,7] uses half-capacitors formed using the top-level of metal. The density of these interconnects depends on the distance between the tiers, the rise/fall times of the technology, and the dielectric constant of the gap. Pitches of 50μm have been demonstrated in a 0.35μm CMOS technology; however, due to the proximity requirement between the plates of the capacitors, this approach requires the tiers to be face-to-face and is therefore limited to two tiers. The challenge for this configuration is supplying DC power to both chips. However, neither of these groups [6,7] has suggested a method for supplying DC power to the top chip. Specially designed probes, limited to test and measurement in a research laboratory, are used to supply power, ground, clock, and data to the chips.

Typically, solder bumping is used to provide DC connectivity between chips, or a chip and a substrate. The problem with solder bump technology and AC coupled between chips is the resulting gap between the two chips. For capacitive coupling to work the gap must be small enough, relative to the size of the plate, for there to be sufficient coupling between the two half plates that form the inter-chip capacitor. One solution is to use a high-k dielectric underfill to fill the gap [8]. Another approach is to form trenches in the substrate that allow for the solder bump to be recessed deep enough so that the chip and substrate are brought into close proximity [5]. This technique, known as ACCI with buried bumps, enables chips to be attached to a substrate, with an interface that supplies AC and DC connectivity. Shown in Figure 1, under “Contactless – Capacitive with buried bumps” is a cross sectional view and a 3D view of the MCM using the buried bump technology. The solder bumps are used to create redundant power and ground bumps; thereby, increasing the manufacturing yield of the assembled MCM, since all data is communicated across the AC coupling elements. The AC channels formed by the coupling elements are not susceptible to bump failure; unless the assembled module loses so many power and ground bumps that the integrity of the power supply grid suffers. Researchers at NCSU have demonstrated this complete technology using 0.35μm CMOS on a MCM-D with five metal layers; over interconnect length of 5.6cm at 2.5Gb/s/channel. The buried solder bump technique can also be combined with a high-k dielectric underfill to reduce the required area for coupling capacitors while relaxing the requirements on inter-plate separation, and also provide stress relief between chip and substrate [8].

Table 1. Comparison of Vertical Interconnect Technologies

		Assembly	Tier limit	Vertical Pitch	Chip Layer Resources
Wire-bonded		Die	~5	35-100µm	All
Micro-bump	3D Package	Die	heat	25-50µm	Top 1-2
	Face-to-face	Die	2	10-100µm	Top 1-2
Contactless	Capacitive	Die	2	50-200µm	Top
	Inductive	Die	heat	50-150µm	Top 1-2
Through-Via	Bulk	Wafer	heat, yield	50µm	All + Top
	SOI	Wafer	heat, yield	5µm	All + Top

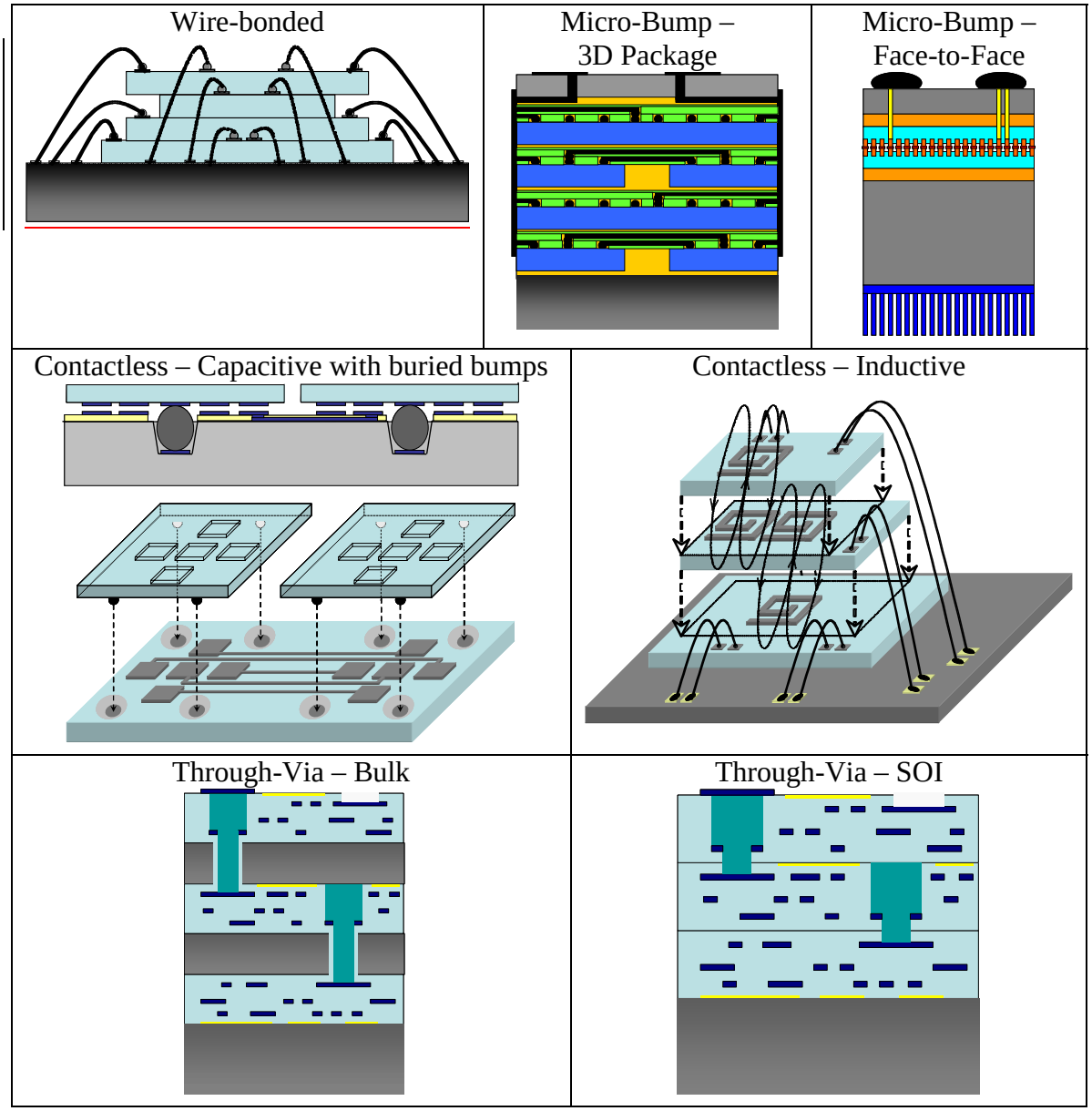


Figure 1. Illustration of Vertical Interconnect Technologies

Inductors may also be used to form inter-chip transformers between chips [5,9]. Inductive coupling is more suited for situations where the separation of the coupling elements, which is determined by the chip thickness, is approaching the lateral dimensions of the coupling elements. This is the typical situation if three or more chips are to be stacked and communications is required between chips throughout the stack. The basic concept for a three-tier stack is shown in Figure 1, under "Contactless – Inductive". In this example, each tier is placed face-to-back and DC power and ground connections for each tier are supplied via wirebonds. The top and bottom tier have either wirebonds or probe pads to supply clock and/or data, for test and measurement. 3D systems that use inductive coupling for tier-to-tier communications and wirebonding to providing DC power and external interfacing are inexpensive and relatively easy to construct. They provide a means to create high I/O connectively in a multi-tier 3D system. Presented later in this article is a demonstration system for inductive coupling.

3D Design: Why and Why Not?

After form-factor improvement, the main advantage of 3D IC technology arises from how it provides a major enhancement to interconnect resources. Used correctly, 3D IC technology should lead to improved bandwidth and throughput and reduced wire length. In the best-case, if we ignore the inter-tier vias, then we would expect the average wire-length to drop by a factor of $\sqrt{N_{tiers}}$. Both wire resistance and capacitance would drop proportionately, meaning a reduction in the power due to wires by a factor of $\sqrt{N_{tiers}}$ and a reduction in wire (RC) delay by a factor of N_{tiers}. Wires with repeaters would see a greater reduction in power and lesser reduction in delay, since repeaters are generally inserted so that delay increases linearly with wire length. Thus, for interconnect dominated architectures; we would expect a significant reduction in energy per operation.

Given high-density vertical interconnect; the question then becomes what are the architectures and applications that can take advantage of the order-of-magnitude improvement in routing resources. This question is just starting to be answered. Aside from imagers (e.g. [1]), the application space is just starting to be explored. Care has to be taken as any performance gain can easily be sacrificed if the increased heat density leads to degraded performance. For circuits operating in saturation, the degredation of mobility with temperature tends to be the dominant effect, and each 10^{o}C increase in operating temperature increases delay by almost 5% [10]. Doubling the heat density, without any improvement in cooling capacity, will lead to more than a 30% degradation in performance! Applications being explored include ones requiring large amounts of memory bandwidth (such as networking and scientific computing) and ones that are traditionally interconnect dominated (switches and FPGAs). All of these applications tend to be very power-hungry. In order for 3D IC technology to show a benefit, it must show that the reduction in interconnect delay outweighs the increase in temperature delay.

Inductive Coupling in 3D-ICs

A CMOS test chip was designed to investigate the use of inductive coupling for 3-D ICs. This demonstration system used a 0.35µm bulk CMOS process, not an advanced through-via-SOI 3D-IC process, and was used to illustrate the concept. For test and measurement, a two-chip stack test system was devised to position the top chip accurately above the bottom chip, and is shown in Figure 2(a). The top chip was used as the receiver and was thinned to the desired thickness, then stacked and aligned with the bottom chip. The top chip was glued onto a micromanipulator, which was used to provide precise positioning and to push the top and bottom chip together to close the gap between them. The corners of the test chip were used to simplify the alignment of the inductors and measurement of the system. Inductors may be placed in any location on the chip to create coupling channels with chip above or below.

Alignment marks were included in the layout of the chip and are visible for both top and bottom chips, and are visible in Figure 2. By referring to the alignment marks and adjusting the micromanipulator, it is possible to achieve perfect overlap of the coupled inductors. It also allows for arbitrary offsets of the inductors, making it possible to explore the tolerance of the transceiver system to misalignment in the 3-D assembly. Figure 2(b) shows a 3-D test structure under measurement.

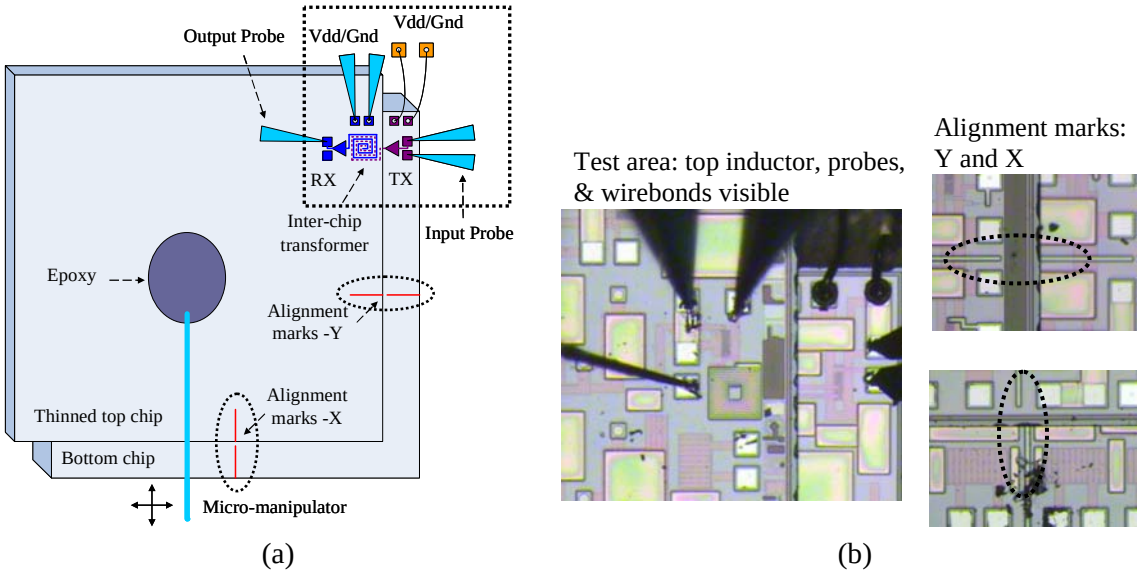
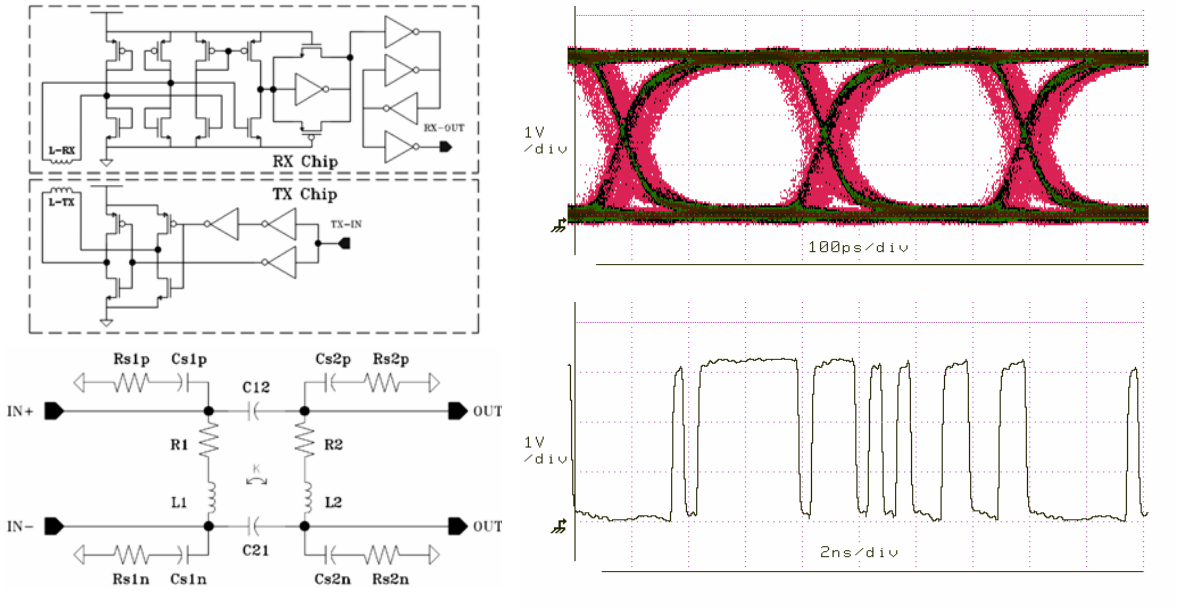


Figure 2: 3-D inductive coupling: (a) test and measurement system, (b) microphotographs of a two chip stack during measurement

A schematic of the current-mode transceiver circuits used for inductively coupled interconnects are shown in Figure 3(a). Also included is a simplified circuit model for the inter-chip transformer. The transmitter circuit is implemented using an H-bridge current steering structure, and is driven with NRZ signals. The receiver circuit can be divided into a sensing stage and a latching stage. The sensing stage detects current pulses from the secondary inductor and converts them into voltage pulses. The latching stage amplifies those voltage pulses and converts them into NRZ signals.

Both the transmit and receiver inductors were made by using double layer 150µm x 150µm spiral inductors with eight turns per layer, resulting in a measured self-inductance of 27nH. Measurements of this inductively coupled transceiver channel produced a maximum signaling rate of 2.8Gb/s for a 2^7-1 pseudorandom binary sequence (PRBS) when the top chip was thinned to 90µm. Bit-error-rate measurements at 2.5Gb/s showed no errors to over 2.5e13 bits, when measurements were stopped due to time constraints. The accumulated eye diagram at the receiver output is shown in Figure 3(b), along with a transient waveform at the RX output for a 2.0Gb/s arbitrary data pattern. The power dissipation for transmitter and receiver were 10.0mW and 37.6mW, respectively. The transceiver circuit does not require external support circuitry or a clock to recover the data and is able to maintain less than 100ps peak-to-peak jitter in the eye diagram at that receiver output. Previous implementations have required complex external support circuitry for clocking receiver, with delay and duty cycle control, and have only achieved data rates of 1.2Gb/s using 0.35µ CMOS technology [9]. They were able to save significant power in their receiver, but as mentioned, the design required significant supporting circuitry, which was not in their reported power consumption.



(a) (b)

Figure 3: (a) transceiver circuit & transformer model, (b) 2.8Gb/s measured eye diagram and 2^7 – 1 PRBS bit pattern measured at 2.0Gb/s

The coupling coefficient determinates the strength of receiving signal at the receiver input; it is not only sensitive to the vertical separation distance between two coupled inductors, but it is also sensitive to the horizontal offset. To investigate the tolerance of an inductively coupled transceiver system to the horizontal misalignment in a 3-D assembly process, measurements at arbitrary offsets between two coupled inductors in X and Y direction were performed. The transceiver system was tested at a data rate of 2.0Gb/s with the top chip thinned to 90µm, 105µm and 120µm and it was determined that the inter-chip transformer can tolerate 50µm, 20µm and 5µm misalignment, respectively.

The shmoo plot of Figure 4(a) shows the simultaneous values of chip thickness and horizontal offset for valid operation at 2.0Gb/s. To investigate tolerance to crosstalk between neighboring channels, measurements of 150µm diameter inductors in the same vertical plane, spaced on a 200µm pitch, and were found to have at least 40dB of isolation up to 5GHz. In other research, it has been shown that there is an optimal pitch that minimizes crosstalk between inductors in different planes [11].

If inductive coupling were used in a through-via-SOI 3D-IC process, there would have to be numerous considerations made. First, why use inductive coupling instead of the already available through-vias. The key reasons to consider inductive coupling are yield and lateral resources. Since the loss of a through-via used for data transmission could render an assembled 3D-IC useless, the combination of through-vias for redundant power and ground distribution with inductive coupling for inter-tier data transmission would increase the yield of the assembled 3D chip stack. Also, the use of inductively coupled I/Os does not eliminate all resources, both active devices and wiring, in its footprint for the tier(s) under consideration. Inductive I/Os would require one or two metal layers, depending on diameter. They could also be used to communicate between any of the tiers, not just adjacent tiers. As shown previously, the required inductor diameter is a function of the vertical separation. Results show that vertical separations 80% of the inductor diameter functioned with reasonable power levels. Reducing this ratio to 50% increases coupling, which would allow for reduced transceiver power.

In a 3D-IC with three tiers, both the lateral and vertical crosstalk components would have to be considered when placing inter-tier I/Os. This would reduce the effective inter-tier I/O density when compared to using through-vias. Therefore, the required amount of inter-tier I/Os for a particular design would have to be established, before considering the use of inductive coupling in a through-via 3D-IC process. Shown in Figure 4(b) is an illustration of the concept of combining inductive coupling and through-vias for 3D-ICs. The illustration shows adjacent tier coupling elements with the appropriate regions vacant of inductors in other tiers above and below the inductors. Also shown are larger inductors that would be used for data transmission between non-adjacent tiers, and the corresponding regions vacant of inductors for tiers in between.

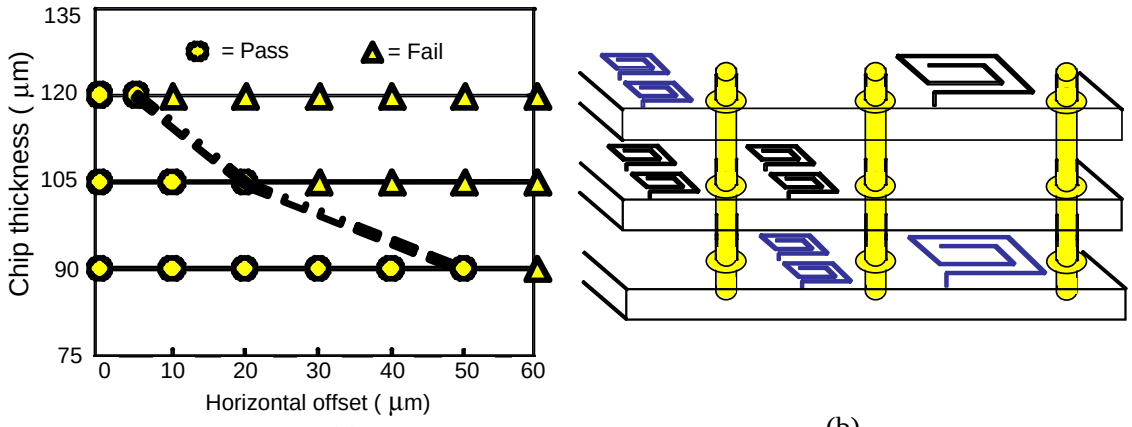


Figure 4: (a) valid operation at 2.0GB/s for various separations and offsets, (b) inductive coupling used in through-via-SOI 3D-IC process

Design Case Study

The through-wafer via 180nm SOI process developed at the Massachusetts Institute of Technology Lincoln Laboratories (MITLL) [1] offers three tiers and the highest-density vertical interconnect available, fitting an inter-tier via roughly in the area of a standard-cell. How much improvement can a typical designer expect to get from adapting their design to this technology? Ignoring the inter-tier vias, interconnect-dominated architectures should notice a reduction in the average wire-length by a factor of at most $\sqrt{3}$ or 42%. Other researchers have performed more thorough investigations of the potential wire-length improvement. Zhang *et al* [12] used stochastic estimates based on Rent's rule that show roughly a 40% reduction in the lengths of the longest wires but only a 30% reduction for the average wire. Das *et al* [13] developed a 3D placer and global router and applied them to the ISPD '98 benchmark circuits, showing reduction in average wire length of 11% when minimizing inter-tier cuts and 41% when minimizing wire-length. By all accounts, average wire-lengths should be significantly reduced. On the downside, the inter-tier via in the MITLL 3D technology creates a column that consumes all routing tracks for the tier, which can worsen routing congestion problems. Also, the parasitc capacitance of the inter-tier via degrades the benefit of reduced wire-length. We were curious to see how much delay and power would be reduced in a real design, once all of these factors were taken into account.

In collaboration with MITLL, we developed a design-kit for their technology for use with *Cadence Design Framework II* and the place-and-route tool *First Encounter*, also from *Cadence*. The kit provides design-rule and layout-vs.-schematic checking as well as a standard-cell library based on the IIT-SoC library from the Illinois Institute of Technology [14]. Figure 5 shows a 3D model of the technology, including metals 1-3 for each tier, with tiers labeled as A-C from bottom to top. One transistor is shown in each tier, with the drain nodes on the lower tiers connected to the gate nodes on the higher tiers through an inter-tier via. Note that the inter-tier vias consume all routing resources in the upper tier and cannot be stacked with the current technology. Note also that tiers B and C are flipped with respect to tier A. A practical place-and-route methodology must account for these factors in order to complete a design successfully.

Our approach is to use standard-cells to implement inter-tier vias. Each upper-tier cell has a corresponding lower-tier cell that must be placed underneath it to be design-rule correct. With this approach, placing and routing in three tiers is a simple matter of partitioning the design into three parts, placing the inter-tier vias, and then completing the placement and routing of each tier individually in *First Encounter*.

Our design methodology is as follows: first, we partition the design into three tiers with minimum cuts between the tiers, using the popular partitioning tool METIS [17]. We then floorplan each tier individually in *First Encounter* and do a preliminary placement with inter-tier vias. At this point, each lower-tier via-cell is in a different position from its corresponding upper-tier via-cell. We then take an average of the two positions, weighting the position by the number of connections in each tier. The via-cell positions are then fixed, and the design is re-placed and routed.

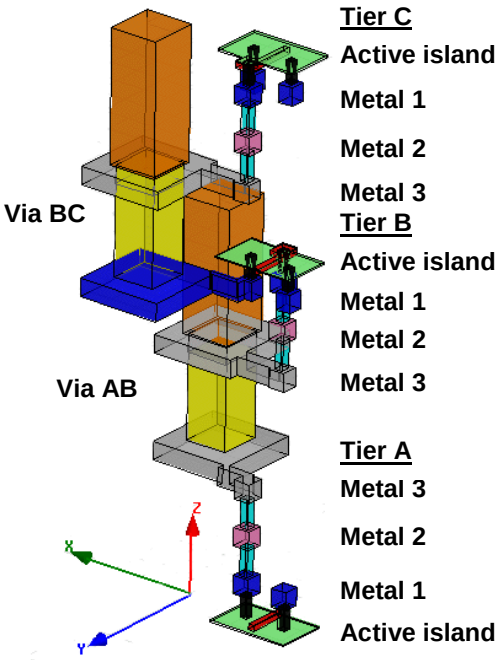


Figure 5: 3D Model of the MITLL process, showing two inter-tier vias and one transistor in each tier.

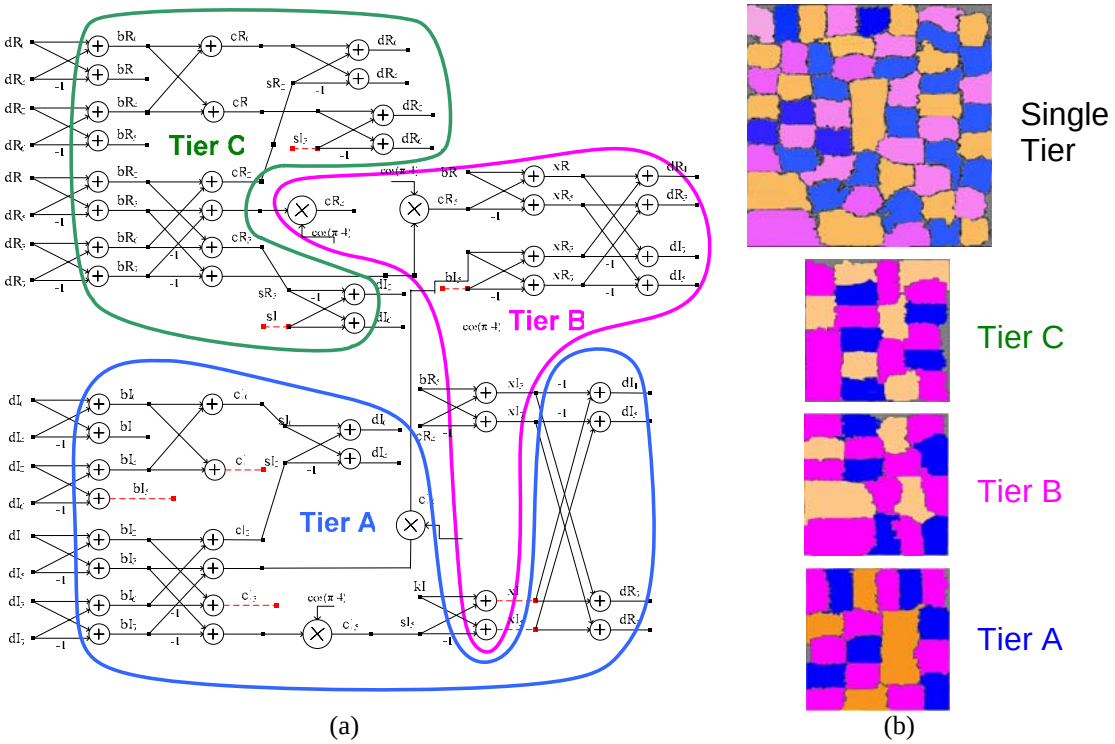


Figure 6: (a) Schematic of the 8-Point floating-point Winograd FFT test-case showing the partitioning between the tiers. (b) Comparison of the single-tier and three-tier floorplans, shown as First Encounter post-placement amoeba views. Each region represents one ADD or MULT operator.

We applied this approach to an 8-point Fast-Fourier Transform (FFT) design with floating-point arithmetic, using arithmetic units from [16]. The FFT was chosen, because the butterfly-structures tend to have long wires, and we wanted a design for which RC delay was a significant factor. We chose the Winograd algorithm [15] for this implementation, because it saves four multipliers over the traditional FFT array, even though it is slightly less regular. Figure 6(a) shows a high-level schematic and partition of the FFT, with dashed red-lines indicating the cuts between tiers. The final placement from *First Encounter* is shown in Figure 6(b) for both a single-tier and the multi-tier cases.

To accurately evaluate delay and power, we merge the extracted parasitic files (in the SPEF format) for each tier into a single file that can be imported into Synopsys *PrimeTime* and *PrimePower*. To complete the approach, we need an accurate estimate of the Inter-tier via resistance and capacitance. It is convenient to model an inter-tier via as a length of wire, but the thickness of the via (about 3 μm) is much less important than the the corresponding resistance and capacitance. Due to the large size of the via, it couples to many nearby wires. To better understand parasitics of 3D processing, we ran simulations using the 3D field-solver *Q3D* from *Ansoft* on each via and lengths of metal 2 wires in each tier. Table 2 shows the results, with intra-tier wire and via pitches shown with the inter-tier via pitch for comparison. Note that the parasitics vary widely, depending on whether the wires are isolated or shielded with surrounding wires, making it difficult to equate via capacitance with a length of wire. The capacitance of an inter-tier via can be approximated, however, as roughly equivalent to 8 to 20 μm of wire, depending on how much coupling is assumed to adjacent wires. Given that the average wire-length for this design is 5X-10X larger than the equivalent value, we can expect that the parasitics of inter-tier vias will have a small effect on the power and delay. The resistance is less significant, due to the large cross-sectional area of each via: about 0.1 Ω per via, which is equivalent to about 0.2 μm of a metal-2 wire.

Table 2. Interconnect Parameters for the MITLL 3D Process

Routing Pitches

Metal 1-3 wires	Via12 and Via23	ViaAB and ViaBC
0.6 μm	1.05 μm	5.6 μm

Simulated Capacitance Values

	Tier(s)	Isolated	Shielded
Inter-Tier Via	AB	0.82 fF	4.34 fF
	BC	0.89 fF	4.15 fF
Metal 2	A	0.051 fF/μm	0.222 fF/μm
	B	0.048 fF/μm	0.221 fF/μm
	C	0.045 fF/μm	0.221 fF/μm

Approximate Resistance Values

Metal 1-3 wires	Via12 and Via23	ViaAB	ViaBC
480 mΩ/μm	4 Ω	82 mΩ	87 mΩ

Results of the analysis are shown in Table 3. Area and wire-length estimates are taken from actual routed results. The total area from all three tiers was somewhat larger than the single-tier case due to the overhead of intertier vias. The average wire length dropped by 17% by using three-tiers, while the longest wire length was reduced by 41%. These finding support the prediction in [12] that the longest wires would benefit more than the average wire. Figure 7 shows a histogram of the wire-lengths that indicates that wires of all lengths were shortened. The big question is, however, why did we not get the 40% reduction in average wire-length that we were hoping for? We believe that the answer lies in the fact that our partitioning method seeks to minimize the number of cuts between tiers, rather than performing a true 3D optimization to minimize wire-length. We are currently working with the University of Minnesota to explore how their 3D placer [18] can improve the performance of this design.

Table 3 also shows the comparison of critical path delay and power between the single-tier and multi-tier versions. Note that the use of three tiers provided a speed-up of only 2.4%, which means that this design was still limited by gate-delay, rather than wire-delay. Total power was reduced by 23%, which is a combination of the effects of reduced wire capacitance, reduced clock-power, and reduced short-circuit power (since this design did not use repeaters). These findings imply that most easily achievable benefit from 3D integration may be a reduction in power, rather than an increase in speed.

Table 3. 2D and 3D FFT Place-and-Route Design Results

Technology	180 nm FD-SOI (1P, 3M)	
Cells	138,000	
Nets	143,000	
Power Supply	1.5 V	
	Single Tier	Three Tiers
Power @ 10 MHz	214 mW	164 mW
Area	3.6 mm× 3.6mm = 12.96 mm^2	2.1mm×2.1 mm×3 = 13.23 mm^2
Average Wire Length	98 μm	84 μm
Longest Wire Length	5.87 mm	3.47 mm
Crit. Path Delay	89.90ns	87.90ns
Cuts Between Tiers	321 (AB) 323 (BC)	193 (AC)

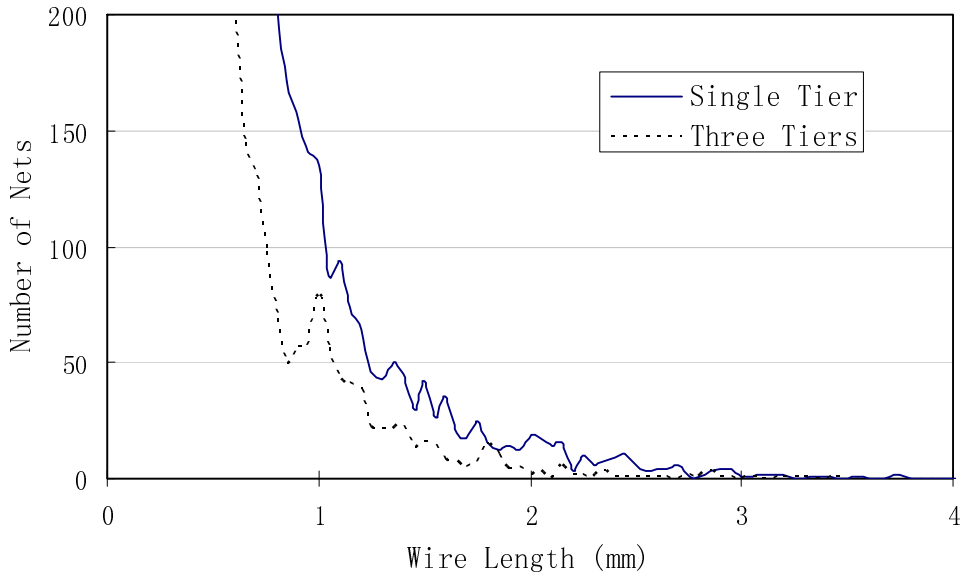


Figure 7. Histogram of wire-lengths from the 2D and 3D FFT Place-and-Route results (bin size = 50 μm).

The removal of heat is of great concern in an SOI process. Most of the heat in the system is generated in the transistor junctions, and since these junctions are floating in glass, there is nowhere for the heat to go. Following the approach used by Rahman and Reif [19], we can assume a 1-dimensional model in which all heat flows through silicon-dioxide to the substrate of tier A (called the "handle silicon"), where the heat-sink is presumably connected. We can assume that the thermal resistance from the thermal paste and heat-sink to ambient is 1.5 $^{\circ}K/W$ and calculate the thermal resistance between the active islands on tiers C and the thermal paste as 4.8 $^{\circ}K/W$. With this analysis, the worst case junction temperature for our FFT will be 1 degree above ambient, which is no cause for alarm. Higher power designs, however, may have cooling problems. For these designs, MITLL offers a special "back-metal", which is applied above tier C and connected to the BC-inter-tier vias. This approach would reduce the thermal resistance calculated above by 1.3 $^{\circ}K/W$ (27%). However, the existence of a heat-sink on the same surface as the pads makes the design of the package much more complicated.

The preceding analysis can be altered to estimate the performance of the FFT when using inductively-coupled interconnect. We perform this estimate by swapping each inter-tier via with an inductively coupled transceiver and re-computing the area, power, and delay. An inductor pitch of 55 μm allows for 1600 inter-tier I/O sites, and is small enough to meet the vertical interconnect requirement of a three-tier design. Even though 1600 inter-tier I/O sites are available, only 800 sites (50%) are useful due to vertical crosstalk limitations. In addition, the in-plane crosstalk constraints reduce the inductor diameter to be approximately 40 μm, allowing for at least 32 turns using minimum line width and spacing. This inductor diameter is significantly larger than that required to enable communications between the nearest inter-tier metal layers (~3μm). It is even larger than that needed for communicating at the distances between the nearest metal

layers between the upper and lower tiers (~10μm). Additional constraints from the required inter-tier vias for power and ground distribution will reduce the available number of I/O sites. In the through-via case study approximately 500 cut between tiers (inter-tier vias) were required.

The area of the inductors themselves is not included, since it is assumed that additional metal layers are added on top of each tier to implement them. The transceiver illustrated in Figure 3(a) has a total area of 1200 μm^2, and a delay of roughly 10 fan-out-4 inverters, which has been measured as 416 ps. The power of each transceiver is 48 mW at 2.8 GHz, but we predict that it can be re-designed to consume no static power and dynamic power of 170 μW at 10 MHz, and require significantly less area for low speed operation. The comparison in Table 4 shows that the increased transceiver area and delay were mildly significant, but the power increased by almost 60% relative to the single-tier implementation. These results indicate that inductively coupled interconnect is not an attractive choice for this kind of low-throughput design and should rather be constrained to designs with very high clock-rates.

Table 4. Comparison of a 2D FFT with a 3D inductively-coupled approach.

Design approach	Footprint Area	Power (@ 10 MHz)	Critical-Path Delay
Single-Tier	3.6mm×3.6 mm	214 mW	90 ns
Inductively-coupled	2.2mm×2.2 mm	339 mW	90 ns

Conclusions

A comparison of 3D through-via and inductively coupled design approaches was presented. Low-throughput, low-power systems such as the FFT case study show a clear power reduction from using multiple-tiers with through-vias, while high-throughput, high-power systems show a performance advantage with both approaches.

A demonstration system for inductively coupled interconnects in 3D ICs was built and tested. The inductive coupling behavior as a function of vertical separation and horizontal offset was also explored. For a 90μm thinned chip with a 150μm spiral inductor, the transceiver communicates PRBS NRZ signals at a data rate of 2.8Gb/s and tolerates up to 50μm misalignment. The transceiver circuit does not require any external support circuitry or a clock to recover the data and is able to maintain peak-to-peak jitter less than 100ps in the eye diagram at the receiver output, with BER measurements showing no errors to more than 2.5e13 bits at 2.5Gb/s.

In addition, a case-study of an FFT placed-and-routed in three tiers was presented, comparing it to an equivalent design in a single tier. A 26% reduction in average wire-length was demonstrated, along with a 33% reduction in the maximum wire-lengths. Because this particular design was not interconnect dominated, it does not make a compelling case for a move to 3D technologies, but the tools and models developed through this example will enable us and other researchers to perform further investigations more easily.

Ultimately, the move to 3D is likely to be limited by heat and yield. The designs that are most likely to benefit from the reduction in wire-lengths are the ones that already run the hottest. Unless methods can be found to effectively remove heat from the stack, any improvement from wire-length reduction can be lost in degraded transistor performance

due to heat. Designers will likely want to devote as few resources as possible to heat removal, which underscores the need for easy, accurate methods to estimate junction temperatures. Yield is another limiting factor. The through-via technologies offer the highest density but are assembled at the wafer-scale, rather than the die-scale. In wafer-scale assembly, if the yield of 3D vias is not extremely high, then the cost of the system will be prohibitively high.

Even though these difficulties seem great, however, they may be less daunting than the difficulties of designing at 65 nm and below. Ease of design may provide the final push that makes true 3D ICs a reality.

Acknowledgements

We would like to thank DARPA and SRC for supporting this work. We would also like to thank Cadence, Synopsys, and PTC, and Ansoft for generously providing the CAD tools. Thanks to MIT Lincoln Labs for providing access to their FD-SOI library and for their aid in developing our design kit. Lastly, thanks to James Stine at the Illinois Institute of Technology for generously providing access to the IIT-SoC standard-cell characterization scripts.

References

1. V. Suntharalingam *et al*, "Megapixel CMOS Image Sensor Fabricated in Three-Dimensional Integrated Circuit Technology", *ISSCC Digest of Technical Papers*, Feb. 2005, pp. 356-357.
2. V. N. Johnson, J. Jozwiak, and A. Moll, "Through Wafer Interconnects on Active pMOS devices," *IEEE Workshop on Microelectronics and Electron Devices*, 2004, pp. 82-84.
3. B. Black, D. W. Nelson, C. Webb, and N. Samra, "3D Processing Technology and its Impact on iA32 Microprocessors," *IEEE Intl. Conf. on Computer Design*, Oct. 2004, pp. 316-318.
4. R. M. Lea *et al*, "A 3-D Stacked Chip Packaging Solution for Miniaturized Massively Parallel Processing," *IEEE Trans. On Advanced Packaging*, vol. 22, no. 3, Aug. 1999, pp. 424-432.
5. S. Mick, J. Wilson, and P. Franzon, "4 Gbps High-Density AC Coupled Interconnection," *IEEE Custom Integrated Circuits Conf.*, May 2002, pp. 133-140.
6. K. Kanda, D. Dwi Antono, K. Ishida, H. Kawaguchi, T. Kuroda and T. Sakurai, "1.27Gb/s/pin 3mW/pin Wireless Superconnect (WSC) Interface Scheme", *ISSCC Digest of Technical Papers*, Feb. 2003, pp. 186-187.
7. R. J. Drost, R. D. Hopkins, R. Ho, and I. E. Sutherland, "Proximity Communication," *IEEE J. of Solid State Circuits*, vol. 39, no. 9, Sept. 2004, pp. 1529-1535.
8. T. Kim, J. Nath, J. Wilson, S. Mick, P. Franzon, M. Steer and A. Kingon, "A High K Nanocomposite for High Density Chip-to-Package Interconnections", *Materials Research Society Symposium Proceedings*, Vol. 833, 2005.

9. N. Miura, D. Mizoguchi, M. Inoue, H. Tsuji, T. Sakurai and T. Kuroda, “A 195Gb/s 1.2W 3D-Stacked Inductive Inter-Chip Wireless Superconnect with Transmit Power Control Scheme,” *ISSCC Digest of Technical Papers*, Feb. 2005, pp. 264-265.
10. J. M. Daga, E. Ottaviano, and D. Auvergne, “Temperature Effect on Delay for Low Voltage Applications”, *Proc. of Design Automation and Test in Europe*, Feb. 1998, pp. 680-685.
11. N. Mirua, D. Mizoguchi, T. Sakurai and T. Kuroda, “Cross Talk Countermeasures in Inductive Inter-chip Wireless Superconnect”, IEEE Custom Integrated Circuits Conf., Oct. 2004, pp. 99-102.
12. R. Zhang, K. Roy, C. Koh and D. B. Janes. “Power Trends and Performance Characterization of 3-Dimensional Integration for Future Technology Generations,” *Intl. Symp. On Quality Electronic Design*, Mar. 2001, pp. 217-222.
13. S. Das, A. Chandrakasan, and R. Reif. “Design Tools for 3-D Integrated Circutis,” *Proc. of the Asia and South Pacific Design Automation Conference*, Jan. 2003, pp. 53-56.
14. J. Stine, *The Illinois Institute of Technology System-on-Chip Design Flow*, available online at http://www.ece.iit.edu/~vlsi/scells.
15. S.Winograd, “On Computing the Discrete Fourier Transform”, *Mathematics of Computation*, vol. 32, no. 141, 1978, pp.175-199.
16. M. E. Phair, *Free Floating-Point Madness!* arithmetic units, available online at http://www.hmc.edu/chips/index.html.
17. G. Karypis and V. Kumar, *The METIS Serial Graph Partitioning Tool*, available online at http://www-users.cs.umn.edu/~karypis/metis.
18. B. Goplen and S. Sapatnekar, “Efficient Thermal Placement of Standard Cells in 3D ICs using a Force Directed Approach,” *Intl. Conf. on Computer-Aided Design*, Nov. 2003, pp. 86-89.
19. A. Rahman and R. Reif, “Thermal Analysis of Three-Dimensional (3-D) Integrated Circuits (ICs)”, Proc. of the IEEE International Interconnect Technology, June 2001, pp. 157 – 159.