Demystifying 3D ICs: The Pros and Cons of Going Vertical

W. Rhett Davis, John Wilson, Stephen Mick, Jian Xu, Hao Hua, Christopher Mineo, Ambarish M. Sule, Michael Steer, and Paul D. Franzon
North Carolina State University

Editor's note:
As 3D technologies become technologically viable, there is increasing interest in determining the achievable payoff. This article first presents an overview of 3D technologies and introduces the motivation for moving from 2D to 3D. It then presents a case study of a fast-Fourier-transform design to illustrate the advantages of going to the third dimension.
—Sachin Sapatnekar, University of Minnesota

An increasing number of integrated solutions involve the stacking of chips to reduce system size. You can find wire-bonded stacks of processors and memories in cell phones, PDAs, and flash cards. But is physical size of the system the only benefit of stacking chips? Does this miniaturization provide potential performance benefits? Until recently, practical interconnection of chip stacks was achievable only through 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 interconnection between chips in a stack, making true 3D ICs a reality. IC designers must know the benefits and drawbacks of these techniques so that they can decide whether or not their systems would work better as a 3D IC.

This article provides a practical introduction to the design trade-offs of the currently available 3D IC technology options. It begins with an overview of techniques, such as wire bonding, microbumps, through vias, and contactless interconnection, comparing them in terms of vertical density and practical limits to their use. We then present a high-level discussion of the pros and cons of 3D technologies, with an analysis relating the number of transistors on a chip to the vertical interconnect density using estimates based on Rent's rule. Next, we provide a more detailed design example of inductively coupled interconnects, with measured results of a system fabricated in a 0.35-μm technology and an analysis of misalignment and crosstalk tolerances. Lastly, we present a case study of a fast Fourier transform (FFT) placed and routed in a 0.18-μm through-via silicon-on-insulator (SOI) technology, comparing the 3D design to a traditional 2D approach in terms of wire length and critical-path delay.

Overview of vertical interconnect technologies

3D ICs offer an attractive alternative to 2D planar ICs: They provide increased system integration by either increasing functionality or combining different technologies. Currently, SoC solutions limit designers to one fabrication technology for both analog and digital circuits. The trend is to use inexpensive digital processes, which provide less than desirable performance for analog circuits, and to offload increased complexity to the analog designs. Using 3D ICs allows for integrating the best technology for a particular portion of an application into the chip cube.

Table 1 shows a summary of different 3D interconnect approaches, comparing them in terms of the method of assembly (die or wafer scale), maximum number of tiers (*tier* refers to the chips in a stack, as opposed to the layers in a chip), pitch of the vertical interconnect, and amount of routing resources consumed on the chip. Figure 1 illustrates each approach.

0740-7475/05/$20.00 © 2005 IEEE Copublished by the IEEE CS and the IEEE CASS

Table 1. Comparison of vertical interconnect technologies: wire bonded, microbump (3D package and face-to-face), contactless (capacitive and inductive), and through via (bulk and SOI).

	Wire	Microbump		Contactless		Through via	
Characteristic	**bonded**	**3D package**	**Face-to-face**	**Capacitive**	**Inductive**	**Bulk**	**SOI**
Assembly level	Die	Die	Die	Die	Die	Wafer	Wafer
Tier limit	Assembly process	Heat	Assembly process	Assembly process	Heat	Heat, yield	Heat, yield
Vertical pitch (mm)	35 to 100	25 to 50	10 to 100	50 to 200	50 to 150	50	5
Metal layers blocked by pad	All	Top 1 to 2	Top 1 to 2	Top	Top 1 to 2	All, top	All, top

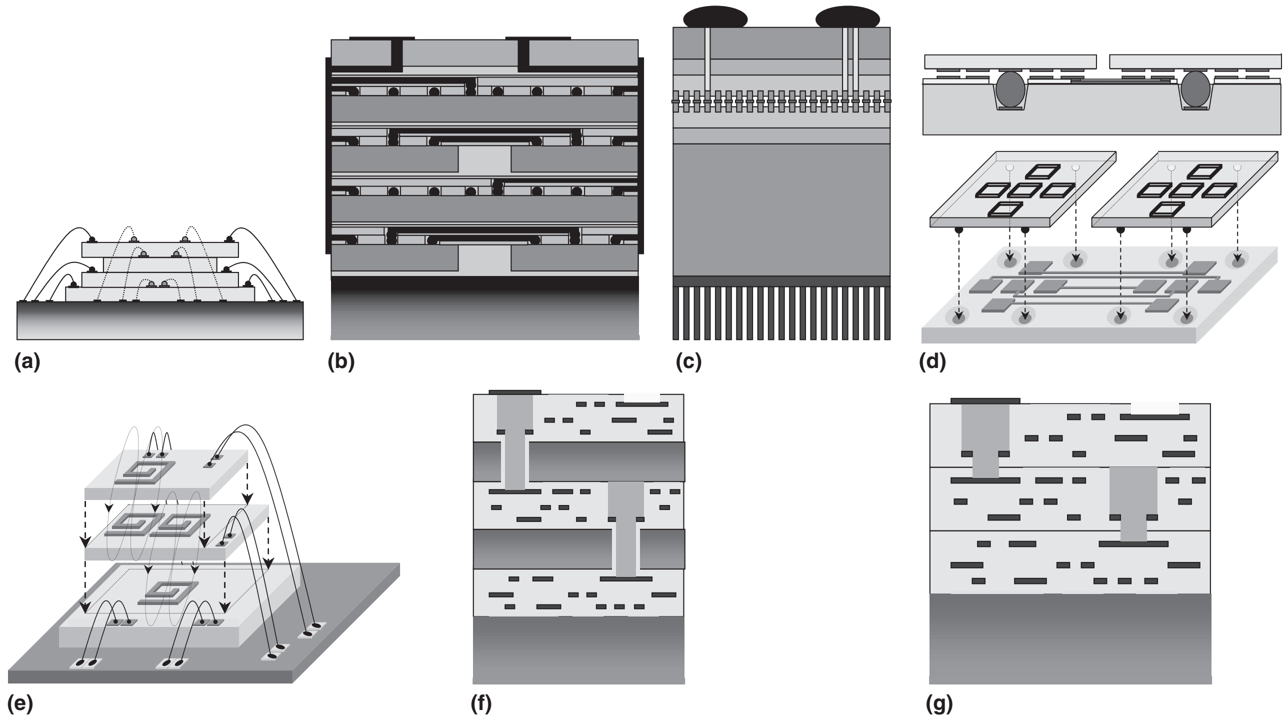


Figure 1. Illustration of vertical interconnect technologies: wire bonded (a); microbump—3D package (b) and face-to-face (c); contactless—capacitive with buried bumps (d) and inductive (e); through via—bulk (f) and silicon on insulator (g).

Wire bonded

The most common approach is wire bonded, in which wires connect the individual die in a stack. In general, connections between chips go through the board or chip carrier and back to other chips in the stack; however, it is possible to bond from chip to chip in the stack. This approach is limited by the resolution of wire bonders (for example, 35 μm for a 15-μm wire) and becomes increasingly difficult as the number of I/Os in the chip stack increases. Unlike other 3D approaches, wire bonds are possible only on the chip's periphery, which severely limits its 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 pressure tends to destroy devices underneath the pad.

Microbump

Microbump technology involves the use of solder or gold bumps on the surface of the die to make connections. These bumps typically have a pitch of 50 to 500 μm but sometimes have smaller pitches. The mechanical stresses of assembly are much lower than with wire bonding, so pads require only the top or sometimes top two metal layers, leaving lower layers free for routing or for devices.

3D package technology[1] involves embedding previously fabricated die into a set of carrier wafers with a fixed size, enabling engineers to assemble them into a tight cube. A layer of microbumps bond each die-carrier tier to an epoxy routing tier that brings signals to the edges of the cube. They then laminate the tiers into a single stack and add metallization to the sides to connect the routing tiers. The 3D package approach offers a much greater vertical interconnect density than the wire-bonded approach, but it does not significantly reduce parasitic capacitances because a microbump-bonded cube must still route signals to the periphery before sending them back to the destination inside the cube. With the 3D package approach, it is not the assembly process but rather the heat inside the cube that is likely to limit the number of tiers. The 3D package method enables the use of one or more chips, from the same or from different fabrication technologies, in each layer of the stack.

Face-to-face microbump technology[2] offers the ability to shorten the wires between tiers and improve performance by reducing parasitics. Black et al. determined that, with proper placement of blocks in the 3D architecture, they could reduce the use of high-power dynamic logic circuits, repeaters, pipelined stages, and long routing paths. This decreased overall power consumption by 15% while simultaneously increasing performance by 15%. This approach is limited to two tiers, however. Taking connections out of the chip stack requires the use of this technology in conjunction with a wire-bonded or through-via approach.

Through via

Through-via interconnection has the potential to offer the greatest interconnect density but also the greatest cost. Assembly occurs at the wafer level, placing a second wafer face down on the first wafer (face-to-face) and subsequent wafers face down (face-to-back) as the number of tiers grows. The manufacturing process then etches holes through the upper wafer into the lower wafer and fills the holes with tungsten to provide connectivity. Before placement of the next chip, 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. As in the 3D package approach, the assembly process in through-via approaches does not limit the number of possible tiers; rather, heat inside the stack is the limiting factor. Also, in this approach, the dies are not known to be good before assembly—so it's possible to attach a good die to a faulty one, making it necessary to reject the entire assembly. In such a situation, yield drops quickly with the addition of more tiers.

Bulk technologies[3] have demonstrated through-via interconnection by first coating the hole with an insulator, achieving pitches of 50 μm. Silicon-on-insulator (SOI) technologies[4] avoid the need for passivating the hole by polishing the substrate away completely, down to the buried oxide. SOI technologies have achieved the smallest inter-tier pitches yet, on the order of 5 μm. As for routing resources, the through-via approaches shown in Figures 1f and 1g consume all layers in the upper tier in addition to the top layer in the lower tier.

Contactless

Contactless or AC-coupled interconnection involves the use of capacitive or inductive coupling to communicate between tiers.[5] This approach eliminates the processing steps for creating inter-tier DC connectivity and eliminates the need to route signals to the periphery, allowing for reduced wire lengths. Also, because the contactless approach requires only a minimal amount of processing for chip thinning, the lack of specialized processing steps makes it much cheaper than microbump and through-via approaches.

Capacitive coupling[6,7] uses half capacitors formed from the top level of metal. The density of these interconnects depends on the distance between the tiers, the rise and fall times of the technology, and the dielectric constant of the gap. Kanda et al. and Drost et al. have demonstrated 50-μm pitches in a 0.35-μm CMOS technology; however, because of 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 Kanda et al. nor Drost et al. have suggested a method for supplying DC power to the top chip.

Typically, engineers use solder bumping to provide DC connectivity between chips or between a chip and a substrate. The difficulty with combining solder bump technology and AC-coupled interconnection (ACCI) 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) to allow sufficient coupling between the two half plates that form the interchip 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 recess the solder bump deep enough to bring the chip and substrate into close proximity.[5] Known as ACCI with buried bumps, this technique provides an interface that supplies AC and DC connectivity and lets chips attach to a substrate. Figure 1d shows cross-sectional and 3D views of a multichip module (MCM) using buried-bump technology. The solder bumps are used to create redundant power and ground bumps, and because all data is transferred across the AC coupling elements, this technique increases the assembled MCM's manufacturing yield. (The AC channels formed by the coupling elements are not susceptible to individual bump failure. So, unless the assembled module loses so many power and many ground bumps that the integrity of the power supply grid suffers, the module should yield.) Researchers at North Carolina State University have demonstrated this complete technique on a thin-film MCM across a 5.6-cm ACCI channel at 2.5 Gbps per channel using 0.35-μm CMOS chips. 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 interplate separation, and also provide stress relief between chip and substrate.[8]

Inductors can also be used to provide interchip communication.[5] Inductive coupling is more favorable for situations where the separation of the coupling elements, which is determined by the chip thickness, approaches the lateral dimensions of the coupling elements. This is the typical situation in a stack of three or more chips, which requires communications between chips throughout the stack. Figure 1e shows the basic concept for a three-tier stack. In this example, each tier is placed face-to-back, and wire bonds supply DC power and ground connections for each tier. The top and bottom tiers have either wire bonds or probe pads to supply clock and/or data for test and measurement. These 3D systems that use inductive coupling for tier-to-tier communications and wire bonding to provide DC power and external interfacing are inexpensive and relatively easy to construct. They provide a means to create high I/O connectivity in a multiple-tier 3D system. We present a demonstration system for inductive coupling later in this article.

Toward 3D design: Why and why not?

After form-factor improvement, 3D IC technology's main advantage is that it significantly enhances interconnect resources. Used correctly, 3D IC technology provides improved bandwidth and throughput, and reduced wire length. In the best-case scenario, if we ignored the inter-tier vias, we would expect the average wire length to drop by a factor of $(N_{\text{tiers}})^{1/2}$. Both wire resistance and capacitance would drop proportionately; that is, power would drop by a factor of $(N_{\text{tiers}})^{1/2}$ and wire (RC) delay would drop 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 interconnection, the question then becomes, What are the architectures and applications that can take advantage of the order-of-magnitude improvement in routing resources? Researchers are just starting to answer this question. Aside from imagers (such as the one Suntharalingam et al. describe[4]), exploration of the application space is just beginning. Important to note is the risk of losing performance gain if the increased heat density leads to degraded performance. For circuits operating in saturation, the degradation of mobility with temperature tends to be the dominant effect, and each 10° C increase in operating temperature increases delay by almost 5%. Doubling the heat density without any improvement in cooling capacity will lead to more than a 30% degradation in performance! Researchers are exploring applications, such as ones requiring large amounts of memory bandwidth (for example, networking and scientific computing) and ones that are traditionally interconnect dominated (switches and FPGAs). All of these applications tend to be very power hungry. To show benefit, 3D IC technology must demonstrate that the reduction in interconnect delay outweighs the increase in temperature delay.

Inductive coupling in 3D ICs

We designed a CMOS test chip to investigate the use of inductive coupling for 3D ICs. This demonstration system used a 0.35-μm bulk CMOS process, rather than an advanced through-via SOI 3D IC process.

Test and measurement system

For test and measurement, we devised a two-chip-stack test system to position the top chip accurately

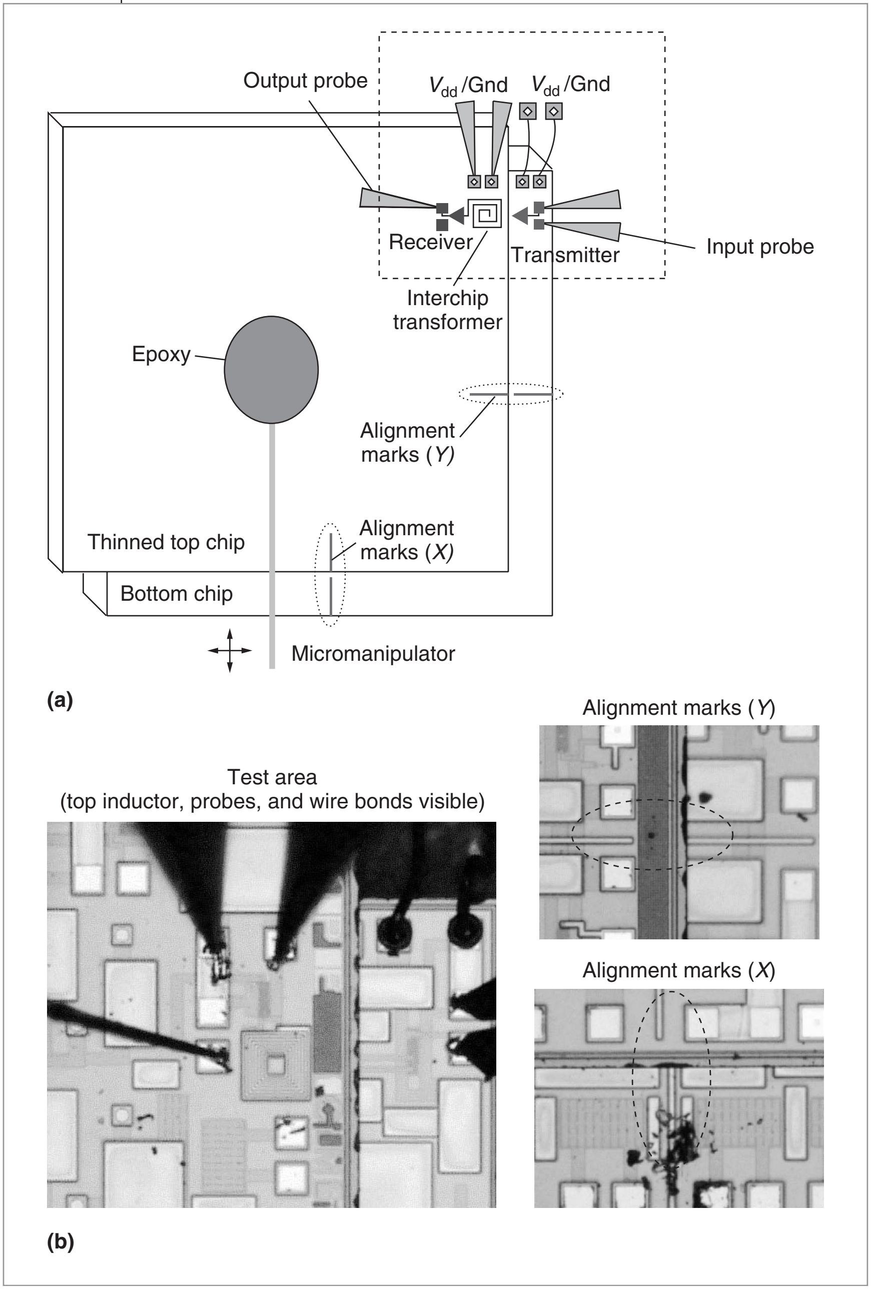


Figure 2. Testing 3D inductive coupling: test and measurement system (a), and microphotographs of a two-chip stack during measurement (b).

above the bottom chip. Figure 2a illustrates this test and measurement system. The system used the top chip as the receiver; we thinned it to the desired thickness, and then stacked and aligned it with the bottom chip. The top chip was glued onto a micromanipulator, which we used to provide precise positioning and to push the top and bottom chips together to close the gap between them. We used the corners of the test chip to simplify the alignment of the inductors and measurement of the system. Inductors can be placed in any location on the chip to create coupling channels with the chip above or below.

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

Transceiver circuits

Figure 3a shows a schematic of the current-mode transceiver circuits used for inductively coupled interconnects (top) and a simplified circuit model for the interchip transformer (bottom). We implemented the transmitter circuit using an H-bridge current steering structure driven with non-return-to-zero (NRZ) signals. The receiver circuit has sensing and latching stages: 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.

We made both the transmit and receiver inductors using double-layer 150-μm × 150-μm spiral inductors with eight turns per layer, resulting in a measured self-inductance of 27 nH. Measurements of this inductively coupled transceiver channel produced a maximum signaling rate of 2.8 Gbps for a $2^7 - 1$ pseudorandom binary sequence when the top chip was thinned to 90 μm. Bit-error-rate measurements at 2.5 Gbps showed no errors for more than 2.5^{13} bits, at which point we stopped measuring because of time constraints. Figure 3b shows the accumulated eye diagram at the receiver output, along with a transient waveform at the RX output for a 2.0-Gbps arbitrary data pattern. The power dissipation for transmitter and receiver were 10.0 and 37.6 mW, respectively. The transceiver circuit does not require external support circuitry or a clock to recover the data and can maintain less than 100 ps of peak-to-peak jitter in the eye diagram at that receiver output. Previous implementations have required complex external support circuitry for the clocked sampling receiver, with delay and duty cycle control, and have only achieved data rates of 1.25 Gbps using 0.35-μm CMOS technology. This implementation saved significant power in its receiver, but, as mentioned, the design required significant supporting circuitry, which was not in the reported power consumption.[9]

The coupling coefficient determines the strength of the receiving signal at the receiver input; it is sensitive to both the vertical separation distance between the two coupled inductors and the horizontal offset. To investigate an inductively coupled transceiver system's tolerance to horizontal misalignment in a 3D assembly process, we performed measurements at arbitrary offsets between two coupled inductors in the *X* and *Y* directions. We tested the transceiver system at a data rate of 2.0 Gbps with the top chip thinned to 90, 105, and 120 μm, and we determined that the interchip transformer can tolerate 50-, 20-, and 5-μm misalignments, respectively. The shmoo

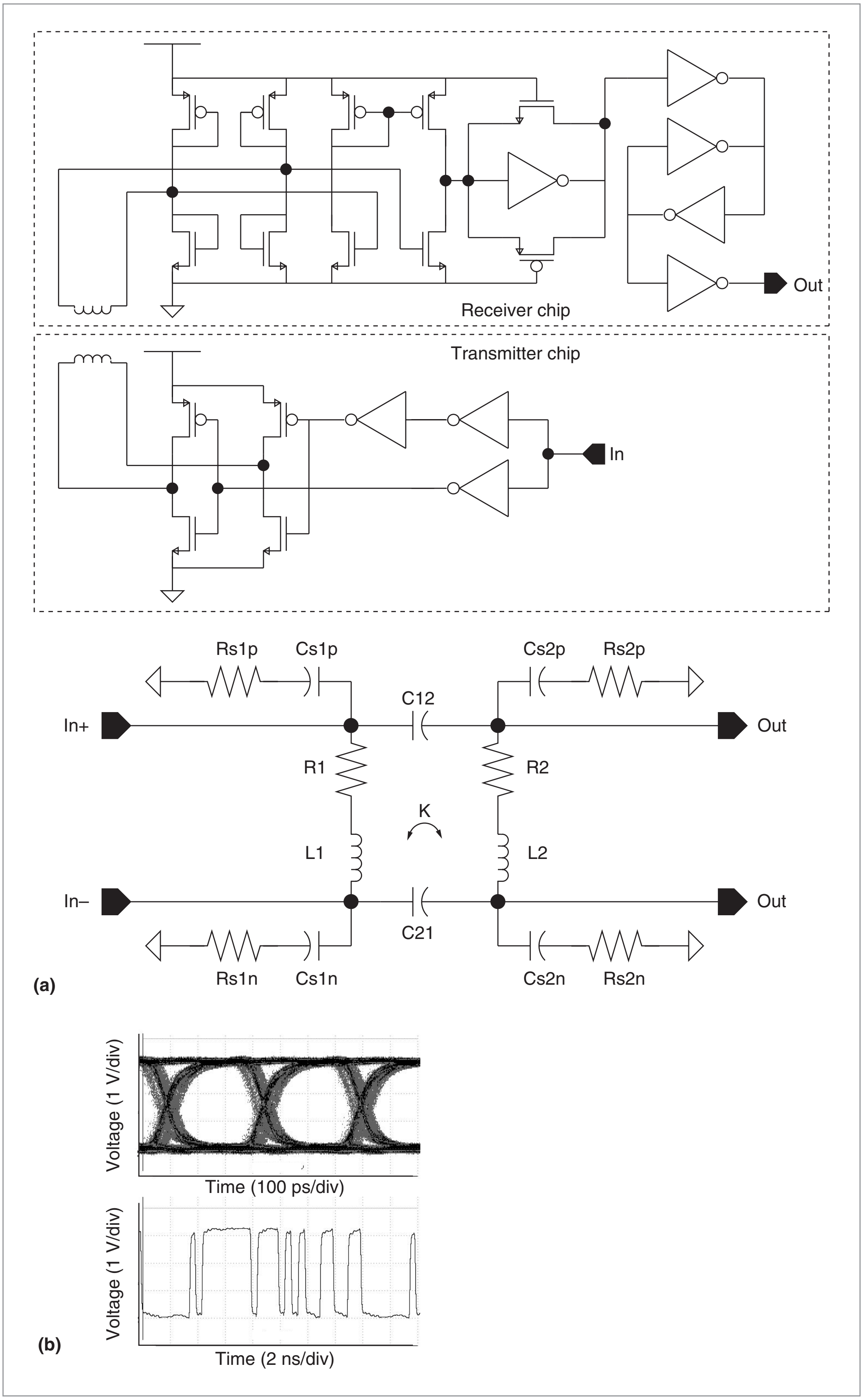


Figure 3. Transceiver circuit and transformer model (a); 2.8-Gbps measured eye diagram and $2^7 - 1$ pseudorandom binary sequence bit pattern measured at 2.0 Gbps (b).

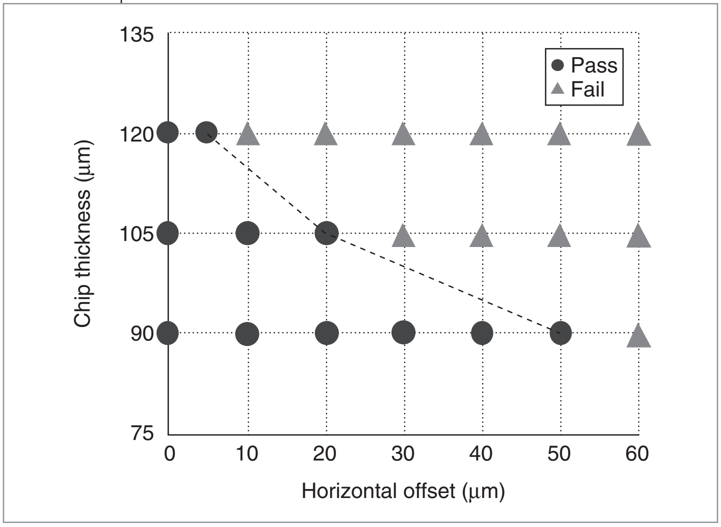


Figure 4. Valid operation at 2.0 Gbps for various separations and offsets.

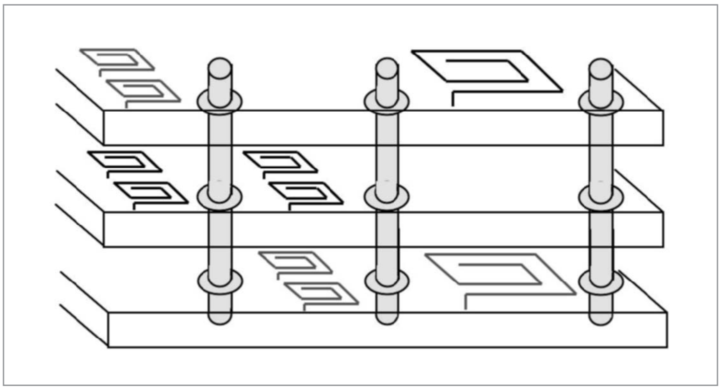


Figure 5. Inductive coupling used in through-via SOI 3D IC process.

plot in Figure 4 shows the simultaneous values of chip thickness and horizontal offset for valid operation at 2.0 Gbps. To investigate tolerance to crosstalk between neighboring channels, we measured 150-μm-diameter inductors in the same vertical plane, spaced on a 200-μm pitch, and found them to have isolation of at least 40 dB, up to 5 GHz. In other research, Mirua et al. have identified an optimal pitch that minimizes crosstalk between inductors in different planes.[9]

Key considerations

Using inductive coupling in a through-via SOI 3D IC process would require numerous considerations. First, why use inductive coupling instead of the already-available through vias? The key reasons to consider for inductive coupling are yield and lateral resources. Because the loss of a through via used for data transmission could render an assembled 3D IC useless, combining through vias for redundant power and ground distribution with inductive coupling for inter-tier data transmission would increase the assembled 3D chip stack's yield. 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 for communication between any of the tiers, not just adjacent tiers. As shown earlier, the required inductor diameter is a function of the vertical separation. Results show that vertical separations that were 80% of the inductor diameter functioned with reasonable power levels. Reducing this ratio to 50% increases coupling, which allows for reduced transceiver power.

In a 3D IC with three tiers, the placement of inter-tier I/Os must consider lateral and vertical crosstalk components. This would reduce the effective inter-tier I/O density when compared to using through vias. Therefore, designers would have to establish the required amount of inter-tier I/Os for a particular design before considering the use of inductive coupling in a through-via 3D IC process. Figure 5 illustrates 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 the tiers above and below the inductors. It also shows larger inductors for data transmission between nonadjacent tiers, and the corresponding regions vacant of inductors for the tiers in between.

Design case study

The through-wafer-via, 180-nm SOI process developed at the Massachusetts Institute of Technology Lincoln Laboratory (MITLL)[4] offers three tiers and the highest-density vertical interconnect available, fitting an inter-tier via in roughly the area of a standard cell. How much improvement can a typical designer expect from adapting his design to this technology? Ignoring the inter-tier vias, interconnect-dominated architectures should experience a reduction in the average wire length by a factor of at most $3^{1/2}$ or 42%. Other researchers have performed more thorough investigations of the potential wire-length improvement. Zhang et al. 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 average wires.[10] Das et al. developed a 3D placer and global router and applied them to the ISPD 98 benchmark circuits from the International Symposium on Physical Design. Their results showed an 11% reduction in average wire length when minimizing inter-tier cuts and a 41% reduction when minimizing wire length.[11]

So by all accounts, moving to 3D should significantly reduce average wire lengths. 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 increase routing congestion problems. Also, the inter-tier via's parasitic capacitance degrades the benefit of reduced wire length. We were curious to see how much delay and power we could reduce in a real design once we accounted for all of these factors.

Experimental setup

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-versus-schematic checking as well as a standard-cell library based on the IIT-SoC library from the Illinois Institute of Technology. Figure 6 shows a 3D model of the technology, including metals 1 to 3 for each tier, with tiers labeled A to C from bottom to top. We show one transistor 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. Also, tiers B and C are flipped with respect to tier A. A practical place-and-route methodology must account for these factors 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.[12] 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. Using this average, we fix the via-cell positions, and again place and route the design.

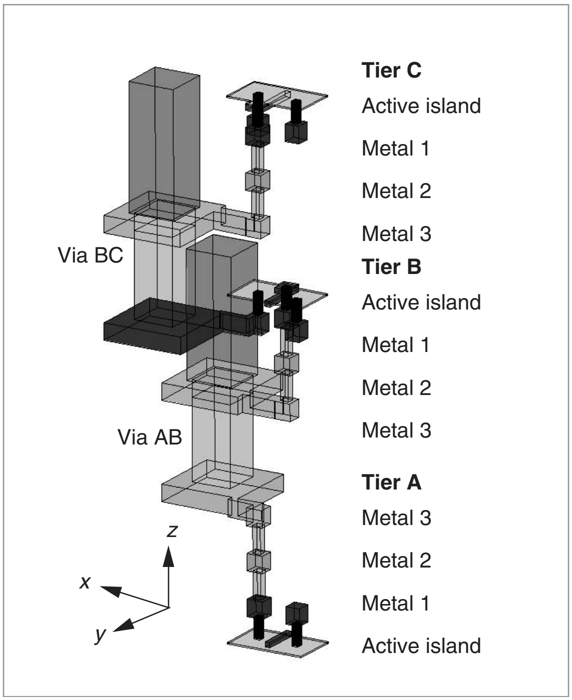


Figure 6. A 3D model of the MITLL process, showing two inter-tier vias and one transistor in each tier.

Our investigation applied this approach to an 8-point FFT design with floating-point arithmetic, using Phair's arithmetic units.[13] We chose the FFT because the butterfly structures tend to have long wires, and we wanted a design for which RC delay was a significant factor. In addition, we chose the Winograd algorithm for this implementation because it saves four multipliers over the traditional FFT array, even though it is slightly less regular. Figure 7a shows a high-level schematic and partition of the FFT, with dashed lines indicating the cuts between tiers. Figure 7b shows the final placement from First Encounter for both single-tier and multiple-tier cases.

Delay and power

To accurately evaluate delay and power, we merge the extracted parasitic files (in Standard Parasitic Exchange 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 corresponding resistance and capac-

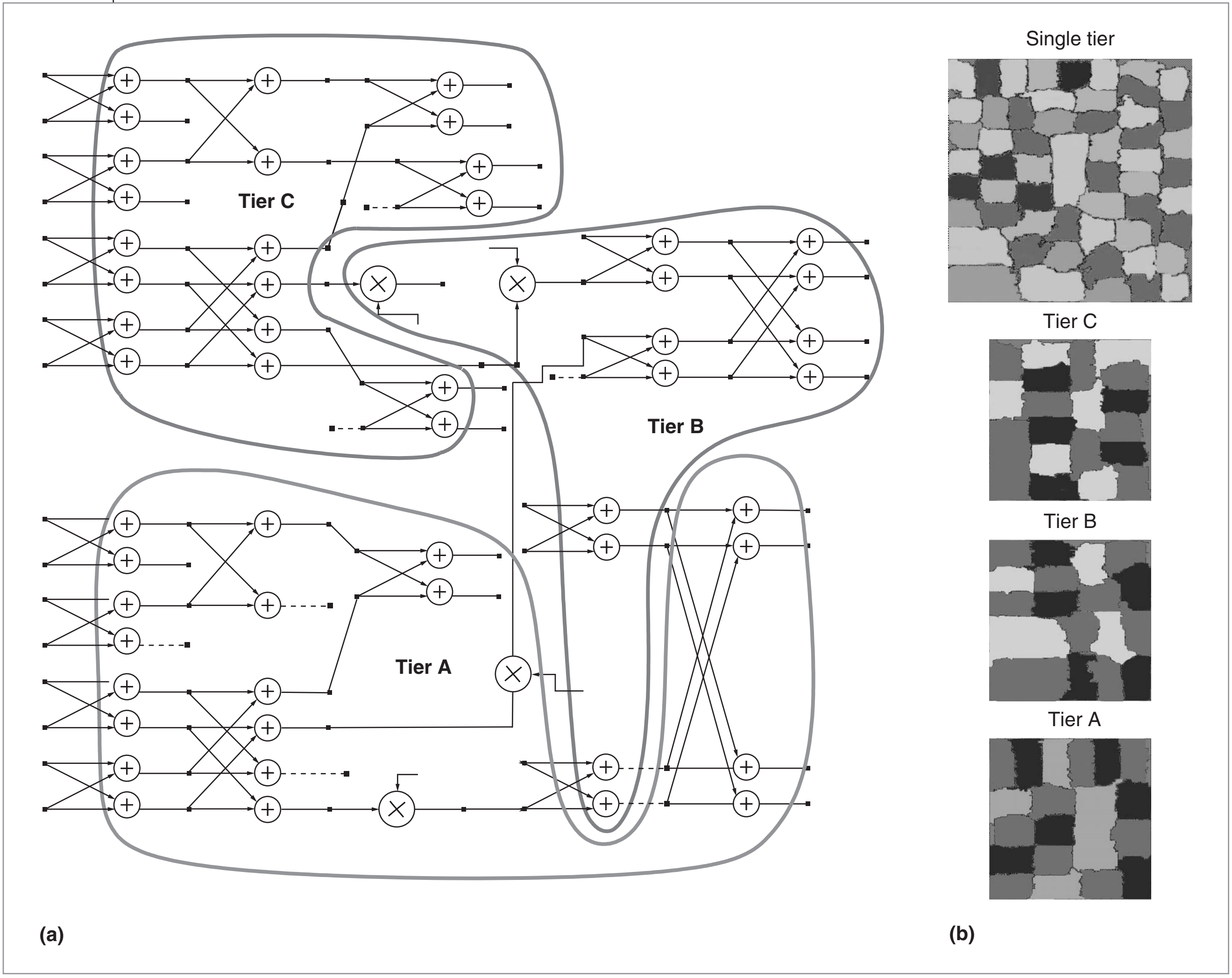


Figure 7. Schematic of the 8-point, floating-point Winograd FFT test case with partitioning between tiers (a), and comparison of the single-tier and three-tier floorplans, shown as First Encounter postplacement amoeba views (b). Each region represents one ADD or MULT operator.

itance. Because of its large size, the via couples to many nearby wires. To better understand parasitics of 3D processing, we ran simulations on each via and the lengths of metal-2 wires in each tier, using Ansoft's Q3D field solver. Table 2 shows the results, with the 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 wire length. We can, however, approximate the capacitance of an inter-tier via as roughly 8 to 20 μm of wire, depending on the amount of assumed coupling to adjacent wires. Given that the average wire length for this design is 5 to 10 times larger than the equivalent value, we can expect that the parasitics of inter-tier vias will have a small effect on power and delay. The resistance is less significant because of 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 3 shows the results of this analysis. We took area and wire length estimates from actual routed results, using a 180-nm fully depleted SOI process with 138,000 cells, 143,000 nets, and a 1.5-V power supply. The total area from all three tiers was somewhat larger than the single-tier case due to the overhead of inter-tier vias. Using three tiers resulted in a 17% drop in average wire length and a 41% drop in the longest wire length. These findings support Zhang et al.'s prediction that the longest wires would benefit more than the average wire.[10] Figure 8 shows a histogram of the wire lengths, which indicates

Table 2. Interconnect parameters for the MITLL 3D process.

	Routing	Resistance	Simulated capacitance values	
Wire or via	**pitches (mm)**	**values**	**Isolated**	**Shielded**
Metal 2	0.6	480 mΩ/μm		
tier A			0.051 fF/μm	0.222 fF/μm
tier B			0.048 fF/μm	0.221 fF/μm
tier C			0.045 fF/μm	0.221 fF/μm
Vias 12 and 23	1.05	4 Ω	NA	NA
Via AB	5.6	82 mΩ	0.82 fF	4.34 fF
Via BC	5.6	87 mΩ	0.89 fF	4.15 fF

Table 3. Place-and-route design results for 2D and 3D FFT.

Parameter	Single tier	Three tiers
Power (mW) at 10 MHz	214	164
Area (mm^2)	3.6 mm × 3.6 mm = 12.96 mm^2	2.1 mm × 2.1 mm × 3.0 mm = 13.23 mm^3
Average wire length (mm)	98	84
Longest wire length (mm)	5.87	3.47
Critical-path delay (ns)	89.90	87.90
No. of cuts between tiers	NA	321 (AB), 323 (BC), 193 (AC)

that a 3D layout shortened all wire lengths. But the big question is, 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 can improve the performance of this design.

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

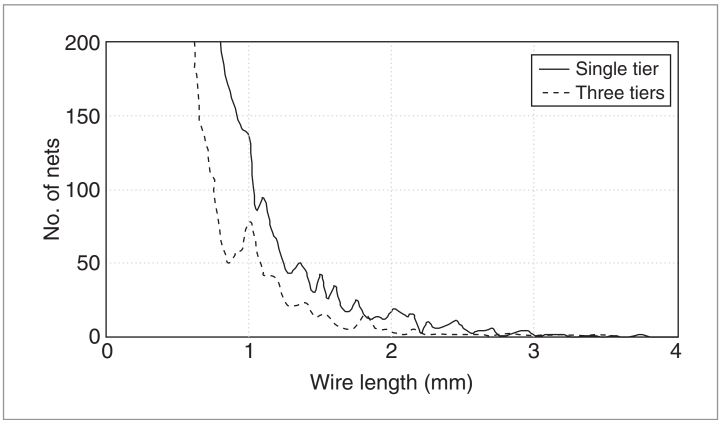


Figure 8. Histogram of wire lengths from the 2D and 3D FFT place-and-route results (bin size = 50 μm).

Heat

The removal of heat is of great concern in an SOI process. Most of the heat in the system comes from the transistor junctions, and because these junctions float in glass, there is nowhere for the heat to go. Following the approach that Rahman and Reif used,[14] we can assume a one-dimensional model in which all heat flows through silicon dioxide to the substrate of tier A (called the handle silicon), an area presumably connected to the heat sink. We can assume the thermal resistance—from the thermal paste and heat sink to ambient—is 1.5° K/W, and calculate the thermal resistance between the active islands on tier C and the ther-

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

Design approach	Footprint area	Power (mW) at 10 MHz	Critical-path delay (ns)
Single tier	3.6 mm × 3.6 mm	214	90
Inductively coupled	2.2 mm × 2.2 mm	339	90

mal paste as 4.8° 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, might have cooling problems. For these designs, MITLL offers a special *back metal*, which goes above tier C and connects to the BC inter-tier vias. This approach would reduce the thermal resistance we just calculated by 1.3° K/W (27%). However, the existence of a heat sink on the same surface as the pads complicates the package design.

We can alter this analysis to estimate the performance of the FFT when using inductively coupled interconnects. We perform this estimate by swapping each inter-tier via with an inductively coupled transceiver and recomputing the area, power, and delay. An inductor pitch of 55 μm allows for 1,600 inter-tier I/O sites and is small enough to meet the vertical interconnect requirement of a three-tier design. Even though 1,600 inter-tier I/O sites are available, only 800 sites (50%) are useful because of vertical crosstalk limitations. In addition, the in-plane crosstalk constraints reduce the inductor diameter to 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 (approximately 3 μm). It is even larger than the required size for communicating at the distances between the nearest metal layers between the upper and lower tiers (approximately 10 μm). Additional constraints from the required inter-tier vias for power and ground distribution will reduce the number of available I/O sites. The through-via case study required approximately 500 cuts between tiers (inter-tier vias).

We do not include the area of the inductors themselves, because we assume that additional metal layers are added on top of each tier to implement them. The transceiver in Figure 3a has a total area of 1,200 μm^2 and a delay of roughly 10 fan-out-of-4 inverters measured as 416 ps. The power of each transceiver is 48 mW at 2.8 GHz, but we predict that it can be redesigned to consume no static power and only 170 μW of dynamic power 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 interconnection is not an attractive choice for this type of low-throughput design and applies more successfully to designs with very high clock rates.

OUR CASE STUDY of an FFT placed and routed in three tiers demonstrated a 26% reduction in average wire length and a 33% reduction in maximum wire length. Because this particular design was not interconnect dominated, it does not make a compelling case for a move to 3D technologies. However, the tools and models we 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 researchers can discover methods to effectively remove heat from the stack, degraded transistor performance will negate any improvement from wire-length reduction. 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. Through-via technologies offer the highest density but are assembled at the wafer scale rather than the die scale. In wafer-scale assembly, if 3D vias do not produce extremely high yield, the cost of the system will be prohibitively high.

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

Acknowledgments

DARPA and the Semiconductor Research Corp. supported this work. We thank Cadence, Synopsys, PTC, and Ansoft for generously providing the CAD tools, and Steve Lipa of North Carolina State University for developing the Cadence design kit. We thank the MIT Lincoln Laboratory for providing access to its FD-SOI library and for its aid in developing our design kit. We also thank Jay Diepenbrock of IBM's Integrated Supply Chain for the use of IBM's test equipment and

James Stine of the Illinois Institute of Technology for generously providing access to the IIT-SoC standard-cell characterization scripts. Finally, we thank Evan Erickson and Eun Chu Oh of North Carolina State University for their assistance.

References

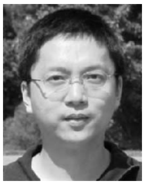
1. R.M. Lea et al., “A 3-D Stacked Chip Packaging Solution for Miniaturized Massively Parallel Processing,” *IEEE Trans. Advanced Packaging*, vol. 22, no. 3, Aug. 1999, pp. 424-432.
2. B. Black et al., “3D Processing Technology and Its Impact on iA32 Microprocessors,” *Proc. IEEE Int’l Conf. Computer Design* (ICCD 04), IEEE CS Press, 2004, pp. 316-318.
3. V.N. Johnson, J. Jozwiak, and A. Moll, “Through Wafer Interconnects on Active pMOS Devices,” *Proc. IEEE Workshop on Microelectronics and Electron Devices*, IEEE Press, 2004, pp. 82-84.
4. V. Suntharalingam et al., “Megapixel CMOS Image Sensor Fabricated in Three-Dimensional Integrated Circuit Technology,” *Int’l Solid-State Circuits Conf. Digest of Technical Papers* (ISSCC 05), IEEE Press, 2005, pp. 356-357.
5. S. Mick, J. Wilson, and P. Franzon, “4 Gbps High-Density AC Coupled Interconnection,” *Proc. IEEE Custom Integrated Circuits Conf.* (CICC 02), IEEE Press, 2002, pp. 133-140.
6. K. Kanda et al., “1.27Gb/s/pin 3mW/pin Wireless Superconnect (WSC) Interface Scheme,” *Int’l Solid-State Circuits Conf. Digest of Technical Papers* (ISSCC 03), IEEE Press, 2003, pp. 186-187.
7. R.J. Drost et al., “Proximity Communication,” *IEEE J. Solid State Circuits*, vol. 39, no. 9, Sept. 2004, pp. 1529-1535.
8. T. Kim et al., “A High-K Nanocomposite for High Density Chip-to-Package Interconnections,” *Proc. Symp. Materials, Integration and Packaging Issues for High-Frequency Devices II*, vol. 833, Materials Research Soc., 2004; http://www.mrs.org/publications/epubs/proceedings/fall2004/g/.
9. N. Mirua et al., “Cross Talk Countermeasures in Inductive Inter-chip Wireless Superconnect,” *Proc. IEEE Custom Integrated Circuits Conf.* (CICC 04), IEEE Press, 2004, pp. 99-102.
10. R. Zhang et al., “Power Trends and Performance Characterization of 3-Dimensional Integration for Future Technology Generations,” *Proc. Int’l Symp. Quality Electronic Design* (ISQED 01), IEEE CS Press, 2001, pp. 217-222.
11. S. Das, A. Chandrakasan, and R. Reif, “Design Tools for 3-D Integrated Circuits,” *Proc. Asia and South Pacific Design Automation Conf.* (ASP-DAC 03), IEEE Press, 2003, pp. 53-56.
12. G. Karypis and V. Kumar, “METIS Serial Graph Partitioning,” Nov. 1998; http://www-users.cs.umn.edu/~karypis/metis/metis/index.html.
13. M.E. Phair, “Free Floating-Point Madness!” May 2002; http://www.hmc.edu/chips/index.html.
14. A. Rahman and R. Reif, “Thermal Analysis of Three-Dimensional (3-D) Integrated Circuits (ICs),” *Proc. IEEE Int’l Interconnect Technology Conf.*, IEEE Press, 2001, pp. 157-159.

W. Rhett Davis is an assistant professor of electrical and computer engineering at North Carolina State University (NCSU). His research interests include methodologies and CAD tools for SoC design, and low-power circuit design for telecommunications and embedded systems. Davis has a PhD in electrical engineering from the University of California, Berkeley.

John Wilson is a research assistant professor at NCSU. His research interests include AC-coupled interconnection, 3D ICs, RF microelectromechanical systems, and analog circuit design. Wilson has a PhD in electrical engineering from NCSU.

Stephen Mick is a postdoctoral researcher at NCSU. His research interests include developing I/O techniques and interfaces for high-speed chip-to-chip communications, and both micro- and nanofabrication technologies. Mick has a PhD in electrical engineering from NCSU.

Jian Xu is pursuing a PhD in the Department of Electrical and Computer Engineering at NCSU. His research interests include high-speed chip-to-chip communications, vertical signaling in 3D ICs, and chip-package codesign. Xu has a BS in electrical engineering from Huazhong University of Science and Technology, China.



Hao Hua is pursuing a PhD in electrical engineering at NCSU. His research interests include design methodology for 3D ICs and crosstalk avoidance methodology in SoC design. Hua has a BS in electrical engineering from Fudan University, China.

Christopher Mineo is pursuing a PhD at NCSU. His research interests include timing verification and delay variability. Mineo has an MS in computer engineering from NCSU.

Ambarish M. Sule is pursuing a PhD at NCSU. His research interests include logic design, and verification and energy-delay optimization techniques. Sule has an MS in electrical and computer engineering from NCSU.

Michael Steer is a professor of electrical and computer engineering at NCSU. His research interests include RF and microwave system design, circuit modeling (including transient harmonic balance analysis), and characterization of interconnects in high-speed digital systems. Steer has a PhD in electrical engineering from the University of Queensland, Australia.

Paul D. Franzon is a Distinguished Alumni Professor of Electrical and Computer Engineering at NCSU. His research interests include the technology and design of complex systems incorporating VLSI, microelectromechanical systems, advanced packaging, and molecular electronics. Franzon has a PhD from the University of Adelaide, Australia.

■ Direct questions and comments about this article to John Wilson, 7620 Elliott Drive, Raleigh, NC 27613; jmwilson@eos.ncsu.edu.