Design and Fabrication of 3D Microprocessors

Patrick Morrow¹, Bryan Black², Mauro J. Kobrinsky¹, Sriram Muthukumar³, Don Nelson⁴, Chang-Min Park¹, and Clair Webb⁴

¹Components Research, Intel Corp., M/S RA3-252, 5200 N.E. Elam Young Parkway, Hillsboro, OR, 97124-6497
²Microprocessor Research Labs, Intel Corp., MS AN1, 1501 S Mopac, Suite 400, Austin, TX, 78746
³Assembly Technology Development, Intel Corp., M/S CH2-140, 5000 W. Chandler Blvd., Chandler, AZ, 85226
⁴Portland Technology Development, Intel Corp., M/S RA3-256, 5200 N.E. Elam Young Parkway, Hillsboro, OR, 97124-6497

ABSTRACT

Stacking multiple device strata can improve system performance of a microprocessor (µP) by reducing interconnect length. This enables latency improvement, power reduction, and improved memory bandwidth. In this paper we review some of our recent design analysis and process results which quantitatively show the benefits of stacking applied to µPs.

We report on two applications for stacking which take advantage of reduced wire length—“logic+logic” stacking and “logic+memory” stacking. In addition to optimizing minimum wire length, we considered carefully the thermal ramifications of the new designs. For the logic+memory application, we considered the case of reducing off-die wiring by stacking a DRAM cache (32 to 64MB) onto a high performance µP. Simulations showed 3x reduced off-die bandwidth, Cycles Per Memory Access (CPMA) reduction of 13%, and a 66% average bus power reduction. For logic+logic applications, we considered a high performance µP where the unit blocks were repartitioned into two strata. For this case, simulations showed that stacking can simultaneously reduce power by 15% while increasing performance by 15% with a minor 14º C increase in peak temperature compared to the planar design. Using voltage scaling, this translates to 34% power reduction and 8% performance improvement with no temperature increase. We found that these results can be further improved by a secondary splitting of the individual blocks. As an example, we split a 32KB first level data cache resulting in 25% power reduction, 10% latency reduction, and 20% area reduction.

We also discuss the fabrication of stacked structures with two complimentary process flows. In one case, we developed a 300mm wafer stacking process using Cu-Cu bonding, wafer thinning, and through-silicon vias (TSVs). This technology provides reliable bonding with non-detectable bonding-interface resistance and inter-strata via pitch below 8µm. We investigated the impact of this wafer stacking process to the transistor and interconnect layers built using a 65nm strained-Si/Cu-Low-K process technology and found no impact to either discrete N- and P-MOS devices or to thin 4Mb SRAMs. We verified fully functional SRAMs on thinned wafers with thicknesses down to 5µm. Although wafer stacking lends itself well to tight-pitch same-die-size stacking, die stacking enables integration of different size dies and includes opportunity to improve yield by stacking known good dies. We demonstrated a die stack process flow with 75µm thinned die, TSV, and inter-strata via pitch below 100µm. We also found negligible impact
to transistors using this process flow. Multiple stacks of up to seven 75µm thin dies with TSVs were fabricated and tested. Prospects for high volume integration of 3D into µPs are discussed.

**INTRODUCTION**

For several years now, there has been much discussion on the potential advantages of 3D strata stacking when used in semiconductor devices. In this paper, we summarize our recent findings evaluating 3D technology applied specifically to microprocessor systems. Although there are many potential advantages of 3D stacking, there are also many challenges, both in design, processing, and assembly. An overall assessment is necessary to make an informed decision on whether the technology provides sufficient benefits to warrant widespread implementation into future products. For microprocessor applications, stacking must be compatible with modern high volume manufacturing processes that include strain-enhanced Si devices and low-K dielectrics, both of which are sensitive to stresses. In addition, careful layout considering thermal ramifications is critical for the success of the system.

In general, the semiconductor industry is driving toward lower power consumption, increased circuit performance, reduced form factor and increased heterogeneous integration. 3D is a technology that can enable improvements in all of these areas. For evolutionary process scaling, the signal delay time (RC) is expected to increase with technology node primarily due to the increasing resistance of the wires. In addition, the aggregated interconnect length increases since more wires are used in each layer and more metallization layers are being added. Hence, for microprocessor systems, we focused primarily on using 3D to reduce wiring [1-3]. Using today’s available manufacturing equipment alignment capability, 3D should be able to reduce global and semi-global wiring, which typically occupies the top one or two metallization layers.

Potential 3D applications can be divided into two generic regimes encompassing tight and loose inter-strata connection pitch. For 3D microprocessor systems, “logic+memory” stacking applications belong to the loose pitch category, while the “logic+logic” stacking applications belong to the tight pitch one. We studied the design and performance implications of stacking in both categories of applications. In addition, we investigated two complementary process flows that target each of these regimes. For the looser pitch (logic+memory) we developed a die stacking process and for the tighter pitch (logic+logic) we developed a wafer stacking process.

**DESIGN OF 3D MICROPROCESSOR SYSTEMS**

Since wire can consume more than 30% of the power within a microprocessor, our main focus for this study was wire reduction. With 3D integration, strata of different types can be stacked with a high bandwidth, low latency, and low power interface. Additionally, wire elimination using 3D provides new microarchitecture opportunities to trade off performance, power, and area. A generic 2-strata structure of a 3D die stack is illustrated in Figure 1.
We studied performance, power, and thermal effects of eliminating wire using stacking by:

1. Shortening wires dedicated to off-die interfaces to memory by stacking memory on logic (“Logic+Memory”). We quantify the effect of stacking a large SRAM or DRAM cache on the Intel® Core™ 2 Duo. The result is higher bandwidth (BW) and shorter latency access to large amounts of storage.

2. Decreasing the length of wires connecting blocks within an Intel® Pentium® 4 family product by splitting the microarchitecture across two die to construct a 3D floorplan. Such a “Logic+Logic” stacking, takes advantage of increased transistor density to eliminate wire between blocks of the microarchitecture. The result is shorter latencies between blocks yielding higher performance and lower power.

3. Decreasing the length of wires within a functional unit block. As an example, we split a data cache into two strata which also resulted in reduced latency and lower power.

**Memory+Logic Stacking**

Increased on die cache capacity, through stacking, improves performance by capturing larger working sets, reduces off-die bandwidth requirements by finding more data on die, and reduces system power by reducing main memory accesses and bus activity. Figure 2 illustrates the baseline microprocessor floorplan, power map, and thermal map. Figure 3 shows average results for 12 RMS benchmarks [2].

![Figure 2. Baseline microprocessor planar floorplan: (a) power map; (b) thermal map.](image)

![Figure 3. (a) Average performance for 2 thread RMS as cache capacity increases from 4MB to 64MB. (b) Peak temperature of 3D DRAM compared to 2D no DRAM](image)

The first bar in each group shows the baseline Cycles Per Memory Access (CPMA). The remaining bars show the CPMA with the three stacking options of 12MB SRAM, 32MB DRAM, and 64MB DRAM. The secondary Y-axis plots the off-die bandwidth for all four configurations.
The off-die bandwidth is reduced 3x with stacked cache while simultaneously decreasing the CPMA by 13%, yielding a 66% average power reduction in average bus power, due to reduced bus activity. We also demonstrated that the peak temperature only increases by 1.92ºC for memory+logic stacking of DRAM on a 92W microprocessor, because DRAMs consume very little power and the hotter die (processor) is next to the heat sink. An in-depth description of these results can be found in [2].

Logic+Logic Stacking

To explore the benefits of Logic+Logic stacking, we split the microarchitecture of the Intel® Pentium® 4 family processor, which is illustrated in Figure 4(a). Using Logic+Logic stacking, a new 3D floorplan can be developed that requires only 50% of the original footprint. The goal is to reduce inter-block interconnect by stacking and reducing intra-block, or within block interconnect through block splitting. The new 3D floorplan is shown in Figure 4(b).

![Figure 4.](image)

(a) Baseline Intel® Pentium 4® Family Product 4 (b) 3D floorplan of (a)

Logic+Logic stacking allowed us to move blocks closer together reducing inter-block latency and power. Much of this effort concentrates on known performance sensitive pipelines. For example, load-to-use delay is critical to the overall performance of most benchmarks. The path between the first level data cache (D$) and the data input to the functional units (F) is drawn illustratively in Figure 4. The worst case path occurs when load data must travel from the far edge of the data cache, across the data cache to the farthest functional unit yielding at least one clock cycle of wire delay entirely due to planar floorplan limitations. Figure 5 shows that a 3D floorplan can overlap the D$ and functional units. In the 3D floorplan, the load data only travels to the center of the D$, at which point it is routed to the other die to the center of the functional units. As a result of stacking that same worst case path contains half as much routing distance, since the data is only traversing half of the data cache and half of the functional units, thus eliminating the one clock cycle of delay in the load-to-use delay. This stacking is also favorable for thermals because the D$ is relatively low power and the 3D power density of the D$ on top of the functional units is lower than the planar floorplan's hottest area over the instruction scheduler. A complete list of pipestages removed from the microarchitecture is shown in Table I. Approximately 25% of all pipestages are removed resulting in a 15% performance improvement from reduced instruction execution latency.

A risk of 3D stacking is the potential doubling of power density and the thermal consequences. This 3D floorplan increases power density by 1.3x while increasing temperature by 14ºC. Since there is a performance gain of 15% along with the 15% power reduction, it is possible to voltage and frequency scale the final results to reach a neutral peak temperature for
the 3D floorplan, shown in Table II. The 3D floorplan can attain neutral thermals by frequency and voltage scaling. The result is a 34% power reduction and 8% performance improvement. Scaling to neutral performance yields a 54% power reduction. These results show significant advantages for “Logic+Logic” 3D stacking of a microprocessor using only simple layout modifications. More in-depth details of this analysis can be found in [2]. We expect further improvements to be possible by further design optimizations for 3D.

<table>
<thead>
<tr>
<th>Functionality</th>
<th>% of Stages eliminated</th>
<th>Performance gain (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front-end pipeline</td>
<td>12.5%</td>
<td>~0.2%</td>
</tr>
<tr>
<td>Trace cache read</td>
<td>20%</td>
<td>~0.33%</td>
</tr>
<tr>
<td>Rename allocation</td>
<td>25%</td>
<td>~0.66%</td>
</tr>
<tr>
<td>FP inst latency</td>
<td>Variable</td>
<td>~4.0%</td>
</tr>
<tr>
<td>Int register file read</td>
<td>25%</td>
<td>~0.5%</td>
</tr>
<tr>
<td>Data cache read</td>
<td>25%</td>
<td>~1.5%</td>
</tr>
<tr>
<td>Instruction loop</td>
<td>17%</td>
<td>~1.0%</td>
</tr>
<tr>
<td>Retire to de-allocation</td>
<td>20%</td>
<td>~1.0%</td>
</tr>
<tr>
<td>FP load latency</td>
<td>35%</td>
<td>~2.0%</td>
</tr>
<tr>
<td>Store lifetime</td>
<td>30%</td>
<td>~3.0%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>~25%</td>
<td>~15%</td>
</tr>
</tbody>
</table>

Table II. Frequency and Voltage scaling the “Logic+Logic” stacked 3D floorplan; Conversion equations for Temperature, Power, Vcc, and Frequency are included.

<table>
<thead>
<tr>
<th>3D provides 15% added perf; 15% pwr savings at same frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pwr</td>
</tr>
<tr>
<td>Baseline</td>
</tr>
<tr>
<td>Same Pwr</td>
</tr>
<tr>
<td>Same Freq</td>
</tr>
<tr>
<td>Same Temp</td>
</tr>
<tr>
<td>Same Perf</td>
</tr>
</tbody>
</table>

Perf vs. frequency 0.82% perf for 1% frequency
Freq vs. Vcc 1% for 1% in Vcc

Another way to take advantage of “Logic+Logic” stacking is to split a block between two die. Figure 5(a) illustrates a 32KB SRAM first level data cache of an IA32 microprocessor.

Figure 5. First level data cache (a) Baseline 2D (b) 3D implementation.

It is 2 way set and 16 bank data with parity. The cache includes the Address Generation Unit (AGU), the Data Translation Lookaside Buffer (DTLB), the Tag Address memory (TAG), 16 Data Banks with data bank muxes, way selects, alignment and parity logic and buffering for the
long metal data busses. The critical timing path from address generation to data result contains several long interconnect routes as shown by the dashed and dotted lines. This cache was rebuilt for 3D and characterized simulating in Intel’s 65nm CMOS technology. Using a 3D circuit implementation, a new floorplan can be developed that requires only 50% of the original footprint and reduces inter-block interconnects (see Figure 5(b)). By stacking the 2 halves, the critical address path is reduced significantly by sharing the address path between the top and bottom die (actual die are connected with metal layers facing each other). The horizontal data bus routing and its buffering are completely eliminated. The bank muxes and way select logic in the center of the unit can now share devices which reduces the total silicon area.

Redrawing the layout for the cache demonstrates a 20% area reduction, mostly from the sharing of sense and write circuits between the banks. Spice results show an expected and small increase in the sense time due to the increased sense amp sharing, however the reduction in horizontal metal routing over compensates and results in an overall 10% reduction in read latency. There is also a simultaneous power reduction of 25% due to shorter address, select, data wires, reduced clock loading, and sharing of clock distribution between the two die. A simultaneous area reduction, power reduction, and latency reduction has a powerful affect on the final product. An in-depth description of these 3D data cache results can be found in [3]. We expect to find similar results throughout a microprocessor design.

FABRICATION OF 3D MICROPROCESSOR SYSTEMS

Introduction

To date, the most viable fabrication techniques for 3D integration are accomplished using bonding. All of the bonding techniques can be organized generally into “dielectric bonding” and “metallic bonding” (see for example [4-8]). Although there are many dielectric bonding processes available ([5-7]), one major disadvantage of these techniques is that each inter-wafer connection requires a through-silicon-via (TSV) which consumes active device area, ultimately resulting in an expensive sacrifice of device density for inter-wafer bandwidth. In contrast, for the case of metal bonding, the bonded inter-wafer connections do not consume active device area, enabling a significantly higher signal bandwidth between device layers. However, in most integration schemes, TSVs are still needed for power, I/O, and signal routing through the device strata, as well as for inter-wafer connections for multi-strata stacking. Although these TSVs do consume some active device area, overall, metal bonding still provides a significant bandwidth advantage over dielectric bonding. This inter-wafer bandwidth is important for applications like 3D logic in microprocessors where the die-to-die via pitch requirements can be very tight [1-2].

Within these two generic bonding classifications, the actual bonding process can be completed either at a die to die level, a die to wafer level, or a wafer to wafer level. The choice depends on the application. Wafer stacking has potential throughput advantages especially where very tight alignment is required, while die stacking enables stacking of different size source die and allows pre-selecting only known good die before assembly to maximize yield. Due to thin die handling issues, wafer stacking also allows thinner strata compared to die to die stacking.

Finally, one can choose either a via-first approach where the TSVs are made along with the fabrication of active circuitry and prior to thinning, dicing, and assembly, or a via last fabrication approach where the TSVs are made after the fabrication of active circuitry but prior to dicing and assembly. Via-first processing [9-11] minimizes the thin wafer handling and processing steps. However, limitations of via-first processing include device issues resulting from subsequent front and back end integration and additional constraints on device design rules. In contrast, the via-
last processing approach fabricates TSVs after completion of the active circuitry, avoiding issues with the standard process flow. Via-last processing also allows for more flexibility in defining via shape and multiple options for making electrical contact of the TSVs to the active circuitry. Since TSV fabrication is done after the active circuitry and wafer thinning, TSVs can be designed to “land on” (i.e. connect to) any of the available metal layers within the circuitry.

Based on these advantages, we focused on via-last integration and metal bonding, researching both die stacking and wafer stacking process flows. For “Logic+Memory” applications such as the stacked cache example discussed earlier, die stacking is the best choice for flexibility to integrate different memory sizes. For “Logic+Logic” designs which require tight pitch interconnections between strata, wafer stacking may be the more appropriate choice. In both of our approaches, the TSVs are fabricated at the wafer level to achieve maximum throughput. However, because of the different application requirements, the TSVs and final substrate thickness dimensions were quite different. For the die stacking process, the dies were on the order of 100 µm thick while for the wafer stacking process the top wafer thickness was on the order of 10 µm. Similarly, the TSV diameter was on the order of 50 µm for the die stacking process and on the order of 5 µm for the wafer stacking process.

**Wafer stacking via copper bonding**

Our 3D wafer stacking fabrication approach started with 300mm full thickness bulk-silicon wafers with 65 nm devices completed through a final standard copper metallization process layer, which will later provide the bonding structures. The 65nm technology wafers included strain-enhanced transistors and low-K dielectrics. A recess of the dielectric parts of the bonding layer is completed to ensure that each Cu structure will contact its counterpart. The wafers are then aligned face-to-face and bonded. For the active device and circuit test structures, the copper bonding pad size ranged from 5 µm by 5 µm to 6 µm by 40 µm. As we reported previously [4], the copper bonding process has a high yield and adds non-detectable resistance to the inter-wafer connections. After bonding, the top wafer is thinned down to a final pre-defined thickness in the range of 5-28 µm for efficient TSV processing. The thinning process includes both grinding and a sub-surface damage removal process resulting in mirror-like substrates with a total thickness variation better than 2 µm. The TSV are then fabricated, integrating copper as the conducting material and an electrical insulator on the sidewall to prevent shorting to the thinned silicon substrate. A backside metal layer was formed so that backside metal pads are electrically connected to the first metal interconnect layer of the thinned wafer through the TSV. The pads were used to make the connections for power and I/O. A cross sectional image of a fabricated stack and a schematic of the final structure are shown in Figure 6.

---

**Figure 6.** (a) XSEM of representative structure (wafer stacking). (b) Schematic.
Overall, we found no difference in electrical characteristics of thinned 65 nm strained-Si devices compared to non-bonded counterparts. We compared both N- and P-channel MOSFETs in the Cu-stacked configuration to those of non-bonded counterparts (i.e. wafers which were not subjected to the bonding, thinning, and TSV integration steps) [12]. Figure 7 shows some representative $I_{\text{off}}-I_{\text{on}}$ and $I_{\text{off}}-V_t$ data comparing stacked and thinned (14 $\mu$m, 19 $\mu$m) wafers with non-bonded baseline wafers.

The outliers were due to non-optimal backside patterning of some of the pads. Similarly, we found no degradation in ring oscillator circuit performance in the thin layer compared to non-stacked wafers [4]. Finally, we did not identify any differences in the functionality of thinned 65 nm 4 Mb SRAMs in the Cu-stacked configuration compared to that of their non-bonded counterparts (i.e. a non-bonded 65 nm 4 Mb SRAM fabricated on standard full thickness silicon wafer) as well as the 4Mb SRAMs directly below in the full thickness silicon [12]. The SRAM data was confirmed for wafers thinned down to a thickness of 5 $\mu$m. The SRAM layout was designed such that two SRAMs are stacked vertically, and both top (thinned) and bottom (full thickness) SRAMs are independently tested using the shared TSVs. The bottom SRAM pins were electrically connected to the TSV and backside metal pads through rerouting layers and the Cu bonding pads. This also showed the good electrical characteristics of the 3D bond pads. Finally, preliminary dicing tests of the wafer stack were successful and produced edge damage comparable with dicing of non-bonded wafers.

Based on the data we collected, we found that the bulk-Si wafer stacking based on face-to-face Cu-bonding and TSV is compatible with high performance strained-Si/Cu-low-k 65nm CMOS processes. Currently, the minimum pitch of the bond pads is limited only by the consistent cross-wafer (300 mm) alignment capability of the bonding equipment, which is expected to improve as tooling advances. This technology continues to look attractive for applications like logic+logic which requires modern high performance devices and tight pitch inter-strata interconnects.

**Die stacking via metal bonding**

The die stacking approach begins with wafer-level processing of the TSV. The first step is to attach a full thickness wafer with active devices to a wafer support system (WSS) and then backside thin. The WSS is required for handling of the thinned wafers (<100 $\mu$m) through the various process steps. Strippable adhesives were used for bonding the product wafer to the WSS and removed post TSV and interconnect fabrication and prior to assembly. The TSVs are formed in a similar fashion as for the wafer stacking approach described above, incorporating both a sidewall insulator and copper as the conductor. For this die stacking process designed for less tight via
pitch applications such as logic+memory, these TSV were significantly larger (>50\,\mu m) than those used in the wafer stacking process (<5\,\mu m) which relaxes some of the processing requirements for the TSV. After the TSV are completed, the thin wafer is de-bonded from the WSS and diced for assembly.

In our 3D die stacking prototype, we targeted the TSV to land on large arrays of W contacts as an alternative to making electrical connection to the backend copper metallization. This scheme allowed for minimal impact to signal and power delivery routing ability in the metal layers layout and/or device “keep out zones” for minimal die size growth. Vertical interconnect resistance of TSVs landing on W contact arrays will be dominated by the target geometry choice. Traditional W contact processing design rules does not allow for a solid metal contact landing pad, therefore electrical continuity was made through an array of contacts whose dimensions are defined by design rules of the process generation for that technology node.

Both silicon and glass were suitable as carrier wafers for our present WSS approach. Perforations in the WSS helped in ensuring a void-free adhesive bond layer. Adhesives used on the WSS need to have high thermal stability, high modulus, and an ability to be stripped at end-of-line. We evaluated both silicone and organic based adhesives looking at these criteria.

Post-silicon grind quality was found to be directly related to the modulus of the adhesive and the grinding processes. Pei et al. [13-14] have studied the subsurface damages induced in the silicon wafer during the grinding processes. For this die stacking study, we elected not to implement a post-grind subsurface damage reduction process. The total thickness variation resulting from this process was better than ±5\,\mu m, but typically not as good as the wafer stacking process since the wafer stacking process does not include a compliant adhesive layer.

Soft silicone based adhesives resulted in poorer grind quality compared to hard silicone based adhesives. Figure 8 shows the grind quality of the adhesives. Back-end layers are visible on the grind surface of the low modulus adhesive unlike the mirror finish grind quality observed for high modulus adhesives. Silicone based adhesives needed additional adhesion promoters added to the formulations to enable adhesion to polyimide passivation layers as modulus and thermal exposure increased. In addition, lower molecular weight silicones created cross-contamination issues with other processes in a clean room environment. Organic based adhesives did not have the issues observed with silicone based adhesives and showed better silicon grind quality; however they had additional difficulties stripping post TSV processing. Chemical, thermal, and optical stripping methods were evaluated but optimal strip solutions depended on the type of adhesive. The details related to WSS and adhesives are reported elsewhere [15].

**Figure 8**: Silicon grind quality using low modulus and high modulus adhesives.
Electrical performance of TSVs fabricated on a representative silicon test chip and targeting landing on an array of W contacts was characterized to understand whether the TSVs meet the requirements of vertical interconnects in a 3D-stacked multi-chip package and to understand the impact of TSVs on fundamental transistor device performance.

Individual TSV resistance was measured using a Kelvin structure for TSVs landing on W contacts. Figure 9 shows a typical wafer level Kelvin resistance distribution.

![Figure 9: Single Kelvin TSV resistance landing on W contacts](image)

Calculated theoretical resistances were estimated based on geometry and configuration for this design such as number of W contacts effectively connected to the TSV, W contact dimensions, TSV barrier layer thickness and dimensions and resistivity of the material sets chosen. The measured low resistance distribution agrees with theoretical calculations and the high resistance portion resulted from non-uniform wafer processing issues associated with the current non-optimized process. Capacitance from a line load of 10 TSVs was measured to be less than 1pF and was limited by the resolution of the instrumentation.

For the large TSV dimensions used here, it is critical to understand the impact to transistor performance due to TSV processing and to establish device “keep out zones”. TSVs were fabricated on a representative silicon test chip at increasing distances to paired planar NMOS discrete transistors. Control measurement data for saturation currents ($I_{DSAT}$) and threshold voltage ($V_T$) was collected prior to TSV processing and compared to post processing data for long and short channel NMOS devices. Figure 10 is a normalized wafer-level plot of $I_{DSAT}$ and $V_T$ shifts for long and short channel NMOS devices in close and far proximity from the TSV.

![Figure 10: $I_{DSAT}$ and $V_T$ shifts for short and long channel n-MOS relative to proximity of TSVs](image)
Overall, measurements indicated only small transistor performance changes for both long and short channel devices in relative proximity of devices to TSVs. Both $I_{DSAT}$ and $V_T$ are independent of TSV proximity for long channel devices. Long channel devices show a median $I_{DSAT}$ change of $<1\%$ and $V_T$ shows a $6mV$ median shift. The tail distributions for the measurements are due to process uniformity issues as discussed in the process development section. Short channel devices show a slightly stronger influence with the presence of a TSV relative to long channel devices. $V_T$ measurements show a median shift of $13mV$ for short channel devices far from the TSV and $17mV$ for short channel devices near to the TSV. $I_{DSAT}$ measurements show a $2\%$ median degradation for short channel devices located at near proximity to the TSV and $2.5\%$ degradation for devices located at far proximity from the TSV. Although it is possible that damage to the silicon during grinding processes can induce stress within the channel region degrading carrier mobility, even without a subsurface damage removal step for this study, the overall shift in device parameters was negligible in all cases. We were clearly able to define device keep out zone design rules that do not have a major impact on area.

Preliminary assembly of TSV enabled test vehicles was demonstrated using this approach. A single metal layer modular daisy chain was used to develop a stack assembly process with TSVs. Figure 11 shows a representative cross-section view of a 7-die stack assembled using TSVs in a front to back die alignment. In this example, each die was about $75 \mu m$ thick. Traditional copper-solder joints and copper-tin-copper thermo-compression interconnects were successfully integrated with this stacking process. Details of the die stacking process can be found here [16].

![Figure 11: Cross-section of 7 die assembly stacking](image)

**CONCLUSIONS AND FUTURE PERSPECTIVES**

Based on our studies, 3D technology is a viable option to augment conventional scaling. We have not identified any major showstoppers to implementation and there is potential for substantial performance improvement and power saving. Overall, even with our simple design modifications, we were able to realize benefits from 3D, and further architecture exploration is likely to expand that further. Moving forward, there are other emerging technologies that need to be compared to 3D solutions. For example, embedded DRAM [17] can also be considered as an alternative to our example of logic+cache memory stacking. For logic+logic applications, items such 3D design tools and high volume manufacturing equipment issues related to very fine alignment capability still need improvement. It should be stressed that the better the alignment, the tighter the pitch and the more architecture opportunities become available in the future.
REFERENCES


