Die stacking using Through Silicon Vias (TSVs) is a promising path for short, dense, and low capacitance interconnects. Logic to memory and logic to logic stacking are specific examples of applications which directly benefit from TSV technology. Low capacitance TSVs offer power efficient path to reach > 1TB/s bandwidths. In the case of logic-to-logic stacking power can be reduced by up to 50% for a deeply pipelined machine (1). Despite clear technical advantages prolific adoption of TSV technologies has been limited in part by implementation costs. Materials and equipment improvements are improving the probability of finding TSV based stacking applications which justify the cost. Continued scaling will further reduce costs and create ongoing material and equipment opportunities and opportunities for new TSV applications. After reviewing specific logic to memory and logic to logic applications, future scaling directions will be discussed.

**Background**

There has been an exponential increase in discussions and advances in TSV process capability as evidenced by a Inspec® search on through silicon vias which reveals 250 publications in 2009 alone. Despite the volume of work image sensors remain the main commercial application to date. Yet the advantages of die stacking are so compelling that organizations continue to seek applications which can justify the cost of TSV processing and subsequent bonding, or like the approach considered in this paper, identify directions to reduce processing costs. The motivation to use TSVs to enable direct die stacking is clear. TSVs provide a means for dense, low capacitance and low latency interconnects between die. Short low capacitance connections between die offer decreased power and signal delay. Thinning die to enable efficient TSV fabrication also leads directly to reduced physical volume. Die stacking technology can obviously also be applied to integrated circuits derived from distinct substrate types or fabrication flows and enable heterogeneous integration of technologies like logic and DRAM.

Two fundamental elements required to enable die stacking are shown in figure 1 and are: 1) TSVs and 2) intra-strata connections. These two technology ingredients lead to significant cost adders for die stacking and the characteristics of these aspects must be compared to fully appreciate die stacking cost and capability space. TSV by definition they consume active Si area no matter how they are created. Intra-strata connections, or
the bonding layer, do not necessarily consume active silicon area and they do not need to be equal in number or tied to the same pitch as TSVs as shown in figure 1. In this paper we will not attempt to cover all the integration options for stacking die but will rather focus on the high level TSV and intra-strata scaling forces.

Figure 1. Cross section of stacked die showing TSVs and intra-strata connections, the two requisites of die stacking.

Applications

Three diverse stacking applications will be considered as a means to re-iterate the continued technical relevance of 3D stacking as well as the spectrum of requirements for TSV and intra-strata connections: Logic to Memory, and two flavors of Logic to Logic stacking.

Logic to Memory Stacking

As shown in figure 2 Intel microprocessor design has evolved from (a) deeply pipelined machines optimized for speed and featuring out of order execution towards (b) multi-core machines in which the number and design of cores is optimized to balance single and multi-threaded performance and improved performance per watt. Future processors could conceivably consist of (c) 10s to 100s of interconnected Intel Architecture cores with accelerators featuring a local and shared cache model (2). While the transition has relaxed inter-core latency issues and our early motivation for logic to logic to logic stacking it has had a huge impact on memory bandwidth requirements.

Figure 2. Evolution of Intel microprocessor design illustrating the transition from (a) single core deeply pipelined machines, towards (b) multi-core machines to balance single
and multithreaded performance, and continuing to (c) 10s to 100s of interconnected Intel Architecture cores and accelerators with shared cache (2).

While traditional memory bandwidth has been scaling at a predictable rate over the past decade (Figure 3) the transition towards 100s of interconnected cores with shared cache will create memory bandwidth demands in excess of 1 TB/s (2), well beyond the evolutionary trend. Although techniques consistent with multi-chip packages and other evolutionary designs could accommodate these rates, wide low capacitance TSV based interfaces offer improved energy efficiency.

![Effective memory bandwidth trend highlighting the revolutionary shift needed to efficiently support future multi-core architecture schemes.](image1)

Such a stack could be built as shown in figure 4. In this example the CPU was placed on top to facilitate cooling, and illustrates a face to face logic to memory interface though alternate embodiments are equally plausible. This application would require thousands of TSVs and the intra-strata connections would need to be about 4x greater. Since pitch at the logic to memory interface would only need to be on the order of 50-100 μm, extensions of traditional solder based flip chip assembly methods could be employed. To ensure high yield levels known good die techniques would be required at least for initial applications. Thus the overall stacking flow would need to be die based (die to die, or die to wafer). This will be an important consideration when we later consider TSV scaling forces.

![Mechanical build up of a face to face logic to memory stack.](image2)
Logic to Logic Stacking – Functional Unit Block Stacking

Die stacking can also be advantageously applied as the geometrical solution to RC delay by adding an extra dimension to bring interacting functional unit blocks (FUBs) in close proximity (3). Bringing interacting functional blocks together will reduce delay and interconnect power. Circuit solutions such as repeaters can address delay concerns but only at the expense of die size and power. Figure 5 shows two specific examples: (a) a case in which a functional unit (FU) is located near a data cache (D$), and (b) a case in which a register file (RF) feeds both a single instruction multiple data (SIMD) block as well as a floating point (FP) unit. Starting with the D$ to FU path it is obvious that die stacking allows the designer to geometrically minimize path length. This particular Pentium was optimized for SIMD instructions over FP hence the FP unit was placed further from the RF. In the 3D case (figure 5b) designers have an additional degree of freedom to optimize for both SIMD and FP operations.

Figure 5. (a) 2D floorplan of an Intel® Pentium 4® product, and (b) the 2 strata for a 3D floorplan (1).

Since FUB stacking fundamentally changes latency and power requirements the benefits can be utilized in numerous combinations as shown in table I assuming Intel’s 65 nm technology (1).

<table>
<thead>
<tr>
<th></th>
<th>Pwr</th>
<th>Pwr %</th>
<th>Temp</th>
<th>Perf</th>
<th>Vcc</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>147</td>
<td>100%</td>
<td>99</td>
<td>100%</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Same Pwr</td>
<td>147</td>
<td>100%</td>
<td>127</td>
<td>129%</td>
<td>1</td>
<td>1.18</td>
</tr>
<tr>
<td>Same Freq</td>
<td>125</td>
<td>85%</td>
<td>113</td>
<td>115%</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Same Temp</td>
<td>97.28</td>
<td>66%</td>
<td>99</td>
<td>108%</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Same Perf</td>
<td>68.2</td>
<td>46%</td>
<td>77</td>
<td>100%</td>
<td>0.82</td>
<td>0.82</td>
</tr>
</tbody>
</table>

The mechanical build up is shown in figure 6 in which the 2 strata are bonded face to face and utilize TSVs to route power and IO into the system. Like the case of logic to memory the number of TSVs required will be in the thousands. Unlike the case of logic to memory however the number of intra-strata interconnects will likely be on the order of millions. The pitch requirements for such a high number of intra-strata connections are
considered too tight for solder based flip chip attach hence the focus on Cu to Cu bonding. To first order the intra-strata pitch should also scale with technology generation.

![Figure 6. Mechanical stack up for logic to logic stacking of FUBs. The application will require thousands of TSVs, which will equal the number of bumps at the die package interface (DPI), and millions of intra-strata connections.](image)

Initial motivation for logic to logic stacking was and remains a geometric solution for RC scaling. To date the 3D solution has not been needed due to tremendous opportunities afforded by 2D scaling. In a similar fashion the architectural evolution from large single core processors to multi-core processors has minimized motivation for logic stacking. However RC scaling is becoming more difficult and hence 3D stacking remains a viable consideration for future processors.

**Logic to Logic Stacking –Functional Unit Block Spliting**

The concept of FUB stacking can be taken a step further by considering splitting or repartitioning FUBs across strata. The easiest example to explain is that of a 32KB SRAM cache unit as shown in figure 7. Considering first the 2D plan in figure 7 (a), the timing from address generation to data result includes several long critical paths. As an example the long horizontal data bus and associated buffering is converted to a short vertical hop in the 3D implementation. Similarly for the case of the address bus the horizontal run is again converted to a vertical jump and the parallel vertical runs become one route shared between the two strata. When simulated using Intel 65nm technology it was shown that the read latency, size, and power were reduced by 10%, 20%, and 25% respectively (4).

Mechanical construction for the split FUB case is the same as the case for stacked FUBs as shown in figure 6 with the exception of intra-strata connections. In the case of split FUBs the number of intra-strata connections could be 10-100x higher than that of FUB stacking. An exact number will depend on specific FUB details and degree of splitting.
Figure 7. 32KB SRAM in (a) 2D and (b) 3D configurations. The 3D floor plan is shown as a front to back configuration for illustrative simplicity (4).

**TSV and Intra-strata Connection Scaling**

Considering that image sensors are the only commercially available TSV applications there is some humor considering TSV and intra-strata scaling for logic applications without a demonstrated commercial starting point. However TSV scaling improves the probability of finding logic applications which can justify the cost and provides a forward path for improved performance at reduced cost.

As highlighted by the applications considered the number of TSVs is expected to remain on the order of 1000s, not 100,000s with no fundamental driving force to scale in number over time. The number of TSVs is only dictated by the need to bring system level IO and power into and out of the active logic interface. In applications such as logic to logic stacking with requisite large numbers of intra-strata connects it is more economical to consider face to face bonding constructs without the penalty of lost Si die area.

As mentioned earlier active Si area is consumed in any TSV scheme, and can dominate the cost equation for TSV based die stacking applications. The presence of a CTE mismatched conductor material embedded in Si will impact local stress and neighboring device performance. This is exemplified in figure 8a showing the drain current change for both n and p channel devices in a 90 nm Si technology as a function of distance from the center of a 6 um diameter Cu TSV. For this specific case drain currents return to baseline values within 2 um of the TSV edge. To ensure all perform to baseline mobility expectations a keep out zone (KOZ) for active devices is employed around TSVs and must be included in the effective TSV area calculation of active Si lost to TSVs (figure 8b). Decreasing the TSV diameter has the dual benefit of reducing both the TSV and KOZ area (5) and will be a key driving force for logic stacking applications.
Figure 8. Active Si area cost for TSVs: (a) keep out zone (KOZ) to maintain background carrier mobility, and (b) total active Si area consumed by TSV is sum of TSV and KOZ.

Process costs also need to be considered. Figure 9 shows the relative costs of some of the more expensive steps for a 20 \( \mu \)m diameter TSV 100 \( \mu \)m deep. The interplay between module costs and integration schemes is complex. Etch costs are directly proportional to both depth and AR, while barrier seed deposition is dominated by AR assuming dimensions are at least on the order of microns. Polish cost is dominated by the amount of overburden to be removed which in turn is dictated by the plating process which can depend on via depth. To first order TSV process costs can be reduced by simply reducing AR by thinning the strata thickness but this strategy is bounded by thin die handling capabilities in some assembly schemes.

Figure 9. Relative cost of some of the more expensive TSV process costs assuming a 20 \( \mu \)m diameter TSV with an aspect ratio of 5.

Si area and module process costs can be summarized in figure 10. Since the primary technology motivation for TSVs is dense, low capacitance interconnects the Si area cost and technology benefit both improve with reduced Si area. To first order TSV process costs decrease with reducing strata thickness. Hence scaling will progress towards the lower left hand corner of this plot. However the ability to directly handle freestanding thin wafers or die is conceivably limited to somewhere around 50 \( \mu \)m. To reliably move beyond this limit we believe the most direct and efficient manufacturing process is to perform permanent wafer level bonding prior to wafer thinning. Such a path assumes same die size but viable device performance has been demonstrated down to 5 \( \mu \)m (6).
and thinner layers are conceivable. For applications such as logic to memory where the
constraint of equal die size is not practical or known good die is a critical consideration
efficient fabrication processes for high AR TSVs are critical. Industrial processes exist
today for AR of about 10, so future focus should really be on devising efficient
technologies for AR ≥ 20.

Figure 10. Si area and capacitance as a function of Si thickness.

Unlike TSVs the number of intra-strata connections and associated pitch will vary by
orders of magnitude depending on application. For logic to memory the number of intra-
strata connections needed is on the order of thousands and hence the minimum pitch
should be relatively loose at 50-100 μm. A wide range of assembly options including
extensions of solder based flip chip are applicable for these dimensions. However for
applications like stacked FUB where the number of intra-strata connections is on the
order of 10^6 the pitch requirements assuming the 10 nm node will be 1 μm, and
subsequent minimum alignment accuracy on the order of 250 nm. To date the only way
to practically achieve this has been with wafer level bonding. Throughput and high
alignment accuracy for wafer level bonding tools will be critical factors influencing the
adoption of wafer level bonding technology. Scientific approaches to reduce bonding
temperature will be key to enabling this.

Figure 11. Application requirements for # of intra-strata connections and pitch.
Conclusions

3D processing remains a promising new dimension for logic applications. Although this statement has been made throughout the last decade, modern process architectures and memory bandwidth hungry applications appear to be moving the cost-benefit pendulum more favorably towards TSV based stacking. For logic to memory applications the pendulum will continue to swing as cost efficient fabrication technologies for TSV ARs $\geq 20$ become available. For logic to logic stacking less expensive solutions involving processor design changes and continued RC scaling have prolonged the need for wafer level bonding. As more straightforward solutions such as RC scaling becomes more difficult the pendulum will continue to move in favor of logic to logic stacking. However the intercept for logic to logic stacking applications remains uncertain and later process node intercepts will require tighter wafer level bonding alignment accuracies. A fundamental understanding of how to reduce bonding temperatures will be critical to enable tight alignment as well as practical throughput for wafer level bonding.

Acknowledgements

No Intel authored 3D paper would be complete without acknowledging the spark and tenacity of Scott List for seeding much of the research at Intel. Intel has a long history of logic to memory stacking research with Johanna Swan and Sriram Muthukumar and their support teams as well. The KOZ data presented is an example of the careful work of Mauro Kobrinsky.

References