3D Processing Technology and its Impact on iA32 Microprocessors

Bryan Black , Donald W. Nelson, Clair Webb, and Nick Samra
Intel Corporation
Email: Bryan.Black@intel.com

Abstract

This short paper explores an implementation of a new technology called 3D die stacking and describes research activity at Intel. 3D die stacking is the bonding of two die either face-to-face or face-to-back in order to construct the 3D structure. In this work a face-to-face bonding is utilized because it yields a higher density die-to-die inter-connect than is possible with face-to-back. With sufficiently dense die-to-die interconnect devices as complex as an iA32 microprocessor can be repartitioned or split between two die in order to simultaneously improve performance and power.

The 3D structure of this emerging technology is examined and applied in this paper to a real x86 deeply pipelined high performance microprocessor. In this initial study, it is shown that a 3D implementation can potentially improve the performance by 15% while improving power by 15%.

1. Introduction

3D die stacking is a new technology that is drawing the attention of the research community. Prior work has analyzed the implementation details of 3D structures [4][6][7], examined system-on-chip opportunities [1], explored cache implementations [4][7], adders [4], and projected wire benefits in full microprocessors [1][2][6]. At the 2004 Technology Venture forum the focus was on "3D Architectures for Semiconductor Integration and Packaging". It is clear that the embedded industry considers the emerging 3D technology as a very attractive method for integrating small systems.

This work examines one possible implementation of a 3D structure and applies it to a high performance iA32 microprocessor. An existing planar design database for a deeply pipelined x86 microprocessor is examined. The planar design is re-floor planned for 3D.

2. 3D Structure

The basic 3D structure is illustrated in figure 1. There are two die joined face-to-face with a dense die-to-die via interconnect. The die-to-die vias are placed on the top of the metal stack of each die and are heat bonded after alignment. In face-to-face bonding, backside vias are required to connect the C4 I/O to the active regions of the two die. Power is also delivered through these backside vias. We conducted a separate study to demonstrate that it is feasible to deliver power through these vias while minimizing droop and inductive effects.
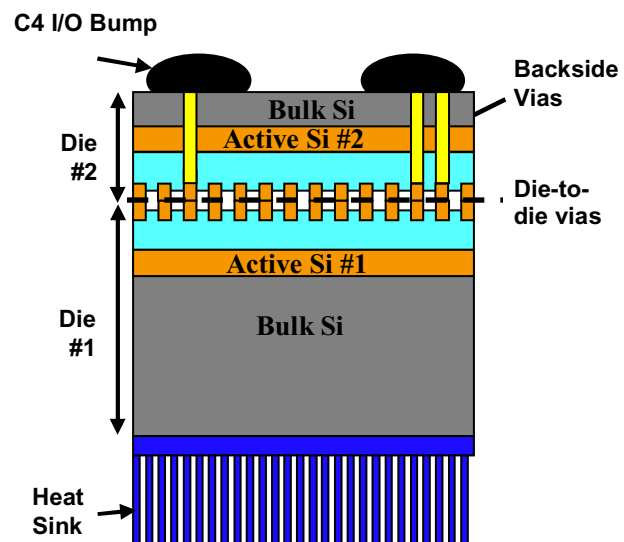

Figure 1. 3D structure[3]

2.1 Advantages

3D die stacking provides two unique advantages over a traditional planar process that can be exploited to improve the performance of a computing system.

1) Next generation transistor density (2x) in the current process generation
2) Enables disparate process technologies across strata

This work focuses on the first advantage. 2x transistor density yields 50% the original footprint which can dramatically reduce the size of the clock network, global wires, and even local wire as wells as the driver

strengths. Section 3 examines the performance and power impact of reduced wiring through both shorter inter-block and intra-block communication.

As described in prior work, the second advantage of 3D enables aggressive system-on-chip or system-on-stack (SOS) solutions. Additional die in a 3D stack can be used to integrate I/O devices, DRAM, solid state disk, etc. SOS is considered beyond the scope of this work and is left for future work.

2.2 Disadvantages

The obvious disadvantage of a 3D structure is thermal dissipation. A 2x increase in transistor density in the same floorplan footprint area without the shrink advantages of a new process generation can potentially lead to 2x power and heat density if hot spots are stacked on top of one another.

A less obvious challenge is die-to-die via density and scaling. The density of interconnect between the die in a 3D structure dictates the level of detail at which functionality can be decomposed, and re-engineered for 3D stacking. With sufficient density deeply coupled functionality can be folded, such as a microprocessor. If the die-to-die via interconnect is limited, decomposition must happen at a coarser granularity, and only disparate functionality can be stacked effectively. Independent of the initial die-to-die via density and the granularity of functional decomposition and folding, via density scaling is required to support a given implementation across multiple process generations. We conducted a separate study that analyzes in detail the die-to-die via density and scaling requirements for real iA32 microprocessors.

Although well beyond the scope of this paper 3D integration has significant impact on the backend development tools.
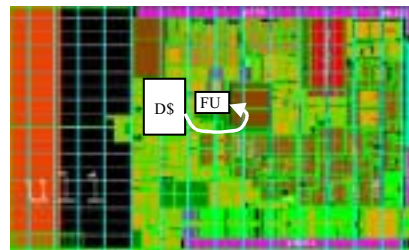
3. A 3D Microprocessor

3D technology is very exciting; however it must demonstrate improved product value. Value is complex and is comprised of many variables such as die manufacturing cost, system manufacturing cost, performance, power, size, and many other microprocessor or system design parameters. While the Intel[TM] Corporation is concerned about all aspects of system design, this work is focused entirely on single threaded high end microprocessor design.

This section focuses on exploiting the density benefit derived from die stacking. Figure 2 illustrates the planar floorplan of our test vehicle, which is a deeply pipelined high frequency iA32 microprocessor. Using, a 3D design structure a new floorplan can be developed that requires only 50% of the original footprint and reduces inter-block interconnect. It is also possible to fold individual blocks

to reduce intra-block interconnect. The new 3D floorplan is illustrated in Figure 3.

It was observed during the process of generating the 3D floorplan that the criteria for determining which blocks are stacked and which ones are folded is different depending on the physical region and the regional functionality of the microprocessor. 3D floorplanning is intrinsically more complicated than planar floorplanning. It is important to consider the temperature effects of stacking any block on a hot block. Large blocks prefer folding to stacking because internal wire and latencies can be eliminated. Folding large blocks also reduces the distance traveled by external signals that are crossing the block. It is possible to fold blocks to reduce latency and/or power.
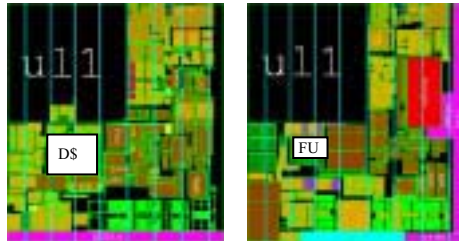

Figure 2. Planar floorplan of a deeply pipelined x86 microprocessor

In Figure 2 the path between the first level data cache (D$) and the data input to the functional units (FU) is drawn figuratively. The worst case path is when the load data must travel from the far edge of the data cache, across the data cache to the farthest functional unit. This is an example of inter-block interconnect that can be reduced with a 3D floorplan. In Figure 3 it can be seen that the D$ and FUs overlap in the new floorplan. In this configuration, 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 FUs. Now 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, effectively eliminating 1 clock cycle of delay in the load execution delay.

The large ul1 cache is an example of intra-block splitting. Although details are not provided the ul1 cache was creatively split to reduce wire delay throughout the overall structure as well as within the sub arrays. Overall the ul1 cache consumes 50% of the original footprint and reduces power and latency substantially.

The new floorplan in Figure 3 targeted egregious timing problems and piped RC delay. Many piped stages of RC delay were eliminated. For example the clock delay of store retirement was reduced by 30%, FP load latency was reduced by 35%, register file read was reduced 25%,

retirement and de-allocation were reduced by 20%, and other important latencies were improved. A total of 25% of all pipe stages were eliminated by the 3D floorplan.

Figure 3. 3D floorplan of a deeply pipelined x86 microprocessor

4. Results

In this section performance, power, and thermal results are presented along with the evaluation methodologies for each.

4.1 Performance

Only single thread performance is considered in this work. Performance is estimated based on partial results from an internal Intel developed cycle accurate full functional performance model that is used for microprocessor development. Over 650 benchmark traces are used that include SPECINT, SPECFP, hand written kernels, multimedia, internet, productivity, server, and workstation applications. Unfortunately, the complexity of this tool prevented a full simulation of all 3D changes. Consequently, only a part of the 3D changes could be simulated directly, while the rest were approximated based on past model behavior.

In Section 3 it was shown that from a quick draft of a 3D implementation of this iA32 microprocessor approximately 25% of all pipeline stages can be eliminated throughout the design. (Note: Our pipeline stage count includes all pipe stages in the machine including the cache hierarchy. It is not equivalent to the branch miss-predict penalty.) Pipe stage elimination improves performance by reducing average instruction execution latency. A 15% performance improvement was achieved by eliminating piped wire stages, reducing delay between blocks, and eliminating wire within blocks.

4.2 Power

Baseline power data for the planar design is gathered using performance model activities and detailed power roll ups from each block in the design. 3D power is estimated from the baseline by scaling according to the proposed design modifications. A 15% power reduction was the result of eliminating 50% of repeaters and repeater flops, along with a substantial reduction in global wire and a 50% reduction in the clock wire.

4.3 Thermals

Thermals are measured using an internally developed tool that accurately models all aspects of the 3D structure for thermal dissipation. During the exploration of this effort it was discovered that a naïve floorplan can increase heat by 10-15%. On further investigation it was discovered that the thermal problems can be addressed at the block level. In our microprocessor example the "hot" areas of the die are due to several blocks that utilize aggressive dynamic circuits to make timing. It was discovered that splitting these hot blocks across two strata reduces internal wire delay sufficiently to allow a relaxation in the implementation of the power inefficient circuits. The speculation is that fast hot blocks can be folded across two strata to reduce power consumption by as much as 50% while maintaining latency and power density. If this turns out to be possible in all the hot blocks, 3D thermal problems may be avoidable.

5. Conclusion

This work explores the recently emerging 3D technology and its application to a real iA32 high-end microprocessor. It is demonstrated that there are distinct advantages to a 3D structure that can be exploited to increase the performance and/or decrease the power of a heavily pipelined machine. The example in this work shows a 3D implementation of a real iA32 microprocessor compared to a planar implementation can improve performance by 15% while simultaneously decreasing power by 15%.

6. References

[1] Y. Deng and W. Maly, "Interconnect Characteristics of 2.5-D System Integration Scheme", ISPD 2001, pp. 171-175

[2] J. Joyner and J. Meindl, "Opportunities for Reduced Power Dissipation Using Three-Dimensional Integration", Proceedings of the IEEE 2002 International Interconnect Technology Conference, pp. 148-150

[7] P. Morrow, et al., "Wafer-level 3D interconnects via Cu bonding", to appear in the Proc. of the 2004 Advanced Metalization Conf.

[3] J. Mayega, O. Erdogan, P. Belemjian, K. Zhou, J. McDonald, and R. Kraft, "3D Direct Vertical Interconnect Microprocessors Test Vehicle", GLSVLSI 2003, pp. 141-146.

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

[5] A. Rahman and R. Reif, "System Level Performance Evaluation of Three-Dimensional Integrated Circuits", IEEE Transactions on VLSI Volume 8, Issue 6, Dec. 2000, pp. 671-678.

[6] A. Zeng, J. Lu, R. Gutmann, and K. Rose, "Wafer-level 3D manufacturing issues for streaming video processors", ASMC 2004, pp. 247-251.