

The Impact of 3-Dimensional Integration on the Design of Arithmetic Units

Kiran Puttaswamy[†] and Gabriel H. Loh[‡]

Georgia Institute of Technology

School of Electrical and Computer Engineering[†]

College of Computing[‡]

Email: kiranp@ece.gatech.edu, loh@cc.gatech.edu

Abstract—3-Dimensional integration technology stacks multiple die on top of each other with a dense die-to-die interface. This enables a circuit designer to replace long wires with short vertical interconnects, thus reducing wire-related delay and power consumption. In this research, we evaluate the impact of a 3D fabrication technology on the latency and power of arithmetic functional units. Specifically, we study integer adders and shifters as they have very different delay characteristics. An adder’s critical path latency is dominated by logic/gate delays, while a shifter’s latency is more greatly affected by wire delay. We demonstrate that the potential benefits of a 3D technology are the greatest when applied to wire-bound circuits. In particular, a barrel shifter implemented in 3D exhibits a 9% reduction in latency with a simultaneous 8% reduction in energy.

I. INTRODUCTION

In current and future technology generations, wires have an increasingly large effect on circuit latency and power consumption [1–3]. An emerging technology called three-dimensional integrated circuit (3D IC) technology stacks multiple die and connects the die with a dense, low-latency die-to-die via interface. This technology can potentially have a great impact on circuit and processor design. In a conventional planar (2D) technology, floorplanning and layout constraints may force two connected gates to be physically separated, thus requiring long, global wiring to connect the gates. In a 3D organization, these gates may be vertically stacked on top of each other on separate die, thus replacing the long global wire with a short die-to-die via.

The long term goal of our research is the design of a complete high-performance microprocessor in a 3D technology, and quantifying the performance and power benefits of such a design. While such a project is beyond the scope of this paper, we present the results of a fundamental step toward achieving that goal. In this work, we evaluate the impact of a 3D technology on arithmetic functional units, which are a critical component of any microprocessor. In particular, we study several common integer addition circuits and a barrel shifter. We chose to focus on these two types of functional units because the adder’s latency and power are dominated by logic, while the shifter is wire-bound. This allows us to compare and contrast the benefits of a 3D technology for two different circuit styles.

In this paper, Section II first presents a brief overview of 3D

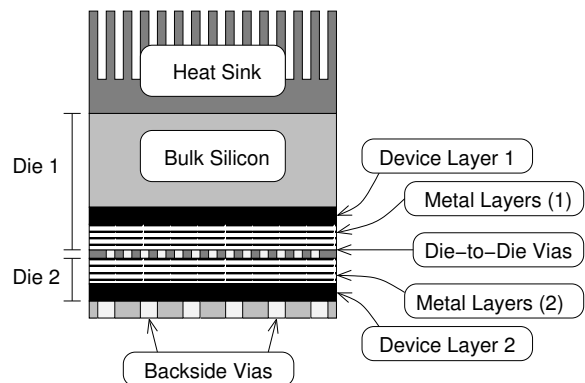


Fig. 1. A stack of two vertically (3D) integrated die.

integration technology. Section III describes the conventional 2D adder and shifter circuit implementations evaluated in this study. Section IV details how to implement the adders and shifter in a 3D technology to reduce area and critical wire delays. Section V provides the performance and power results of implementing functional units in 3D, and Section VI discusses future directions for research in 3D and makes some concluding remarks.

II. 3D INTEGRATION TECHNOLOGY

3D integration technology stacks two or more planar die and connects them with dense inter-die interconnects. Figure 1 shows a two-die stack where one die is inverted and aligned on top of the other such that the metal layers of the two die face each other. The two die are aligned and bonded with a short, highly dense die-to-die (D2D) interface. The distance between the top metal layers on the two die is very small, and the pitch of the D2D vias is of the same order as the top level metal [4]. One company, Tezzaron, has reported manufacturing a 3D D2D interface with a $2.4\mu\text{m}$ pitch with an expected second-generation pitch of $1.46\mu\text{m}$ [5], and we believe that the D2D via density will continue to scale for at least several generations. Likewise, Tezzaron reports that the feed-through capacitance and series resistance of the D2D via is very low. Our simulations show that the delay to transmit a signal between two die through the D2D via is less than a

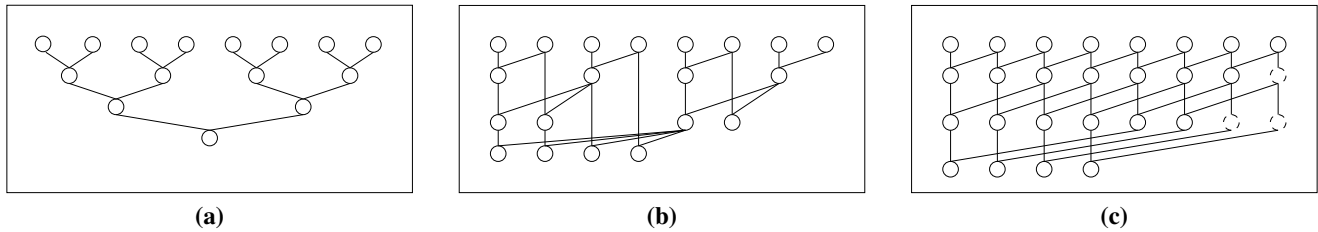


Fig. 2. 8-bit Planar Adders (a) Brent-Kung (b) Sklansky (c) Kogge-Stone. The nodes \circ represent the propagate-generate PG components of the parallel-prefix computation for the adder's carry logic, while the wires communicate the different partial-prefix computations between nodes.

single fan-out-of-four (FO4) delay.

The relative benefits of 3D technology will increase in future technology generations, making it a very attractive option for future designs. There has recently been a great deal of interest in 3D ICs. Researchers in academia and industry have studied the implementation of caches in a 3D technology [6–8]. Researchers in the general purpose micro-processor industry are evaluating the technology for feasibility and applicability [9, 10]. In the embedded market, companies are already shipping products that use 3D integration for 3D-stacked SRAM, DRAM and combined SRAM and micro-controller [11, 12].

Heat removal from localized hot-spots needs to be considered when two actively switching components are stacked on top of each other. The bottom die has a heat sink attached to it and has been reported to have a temperature profile similar to a planar die [13]. The top die generates heat that might get trapped locally due to the poor thermal conductivity of the dielectric layers used between metal layers and between the die. A 3D circuit that substantially reduces wire RCs will consume less power, thus reducing the thermal impact of the 3D structure. Researchers have also suggested using “dummy” thermal vias that do not carry signals, but simply provide low thermal resistance paths to help move heat away from potential hot-spots [14]. Between the overall power reduction and the effectiveness of thermal vias, we believe that thermal management in a 3D structure is a tractable problem.

3D technology will likely extend beyond stacking two die to continue scaling device density. After bonding two die in a face-to-face organization, coarser (less dense) backside vias are required. The differences between these D2D vias change the absolute values of the benefits reported in this paper, but the relative trends and overall conclusions remain the same. 3D integration also allows for heterogeneous technologies (CMOS, DRAM, analog, etc.) to be combined together, thus providing even greater functionality. In this study, we limit the scope of our explorations to a 2-die CMOS stack and leave the evaluation of more sophisticated 3D structures for future work.

III. PLANAR (2D) FUNCTIONAL UNIT IMPLEMENTATIONS

We evaluate two types of functional units: integer adders and a barrel shifter. In particular, we consider Brent-Kung (BK) adders [15], Sklansky (SK) adders [16] and Kogge-Stone (KS)

adders [17]. We also consider a classic barrel shifter. For both the adders and the shifter, we evaluate 64-bit implementations.

The critical paths through addition circuits are typically dominated by gate or logic delay rather than wire delay. For wide adders (64-bit), only a few of the upper levels of the propagate-generate logic make use of long wires. In contrast, a 64-bit barrel shifter contains many very long wires because input bits may need to be shifted 64 positions to the left or right. The comparison of circuits that are either logic-bound or wire-bound helps to illustrate the relative benefits of 3D for different situations and provides insight into what other components in a processor would benefit the most from a 3D implementation.

III-A. Addition Circuits

We chose three adders with varying logic and wire requirements. Figure 2 illustrates the carry generation logic for each of the adders. While the diagrams only show 8-bit versions of the circuits, all of our results are for full 64-bit adders. Figure 2(a) shows the carry generation tree of an 8-bit Brent-Kung adder. Brent-Kung (BK) adders share processing nodes and require propagation twice through the tree: first in a downward direction when the prefixes are synthesized, and then in an upward direction when the carries are generated. Figure 2(b) shows an 8-bit Sklansky (SK) adder. SK adders reduce the number of logic levels that need to be traversed at the cost of more logic and wiring. Each successive level of SK adders has exponentially increasing fanout and wire-length. Figure 2(c) shows an 8-bit Kogge-Stone (KS) adder. KS adders reduce the loading on the critical path by using even more logic to limit the fanout per node to only two. This in turn decreases the overall circuit latency while increasing the power due to the extra logic and wiring.

III-B. Barrel Shifter

Figure 3 shows an 8-bit planar (2D) shifter. Each multiplexer level performs a conditional shift of a magnitude equal to the corresponding power of two. The shifter is designed to perform either right shifts or left shifts based on a control input. The worst-case wire length for each successive level doubles, as illustrated by the sequence of bolded lines in Figure 3. Furthermore, the height of each successive level also increases exponentially as more wiring tracks are required to route the signals across to the appropriate multiplexer inputs.

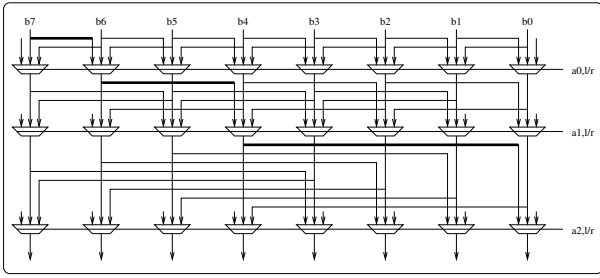


Fig. 3. 8-bit Planar Shifter

The increase in wire length in both dimensions leads to a significant wire-related latency and power cost at the deeper levels of the shifter.

IV. FUNCTIONAL UNIT IMPLEMENTATION IN 3D

We propose designs of 3D-integrated circuits to reduce the length of critical wires between processing nodes by stacking nodes in the vertical dimension. By stacking adjacent nodes on different die, we reduce the area footprint of the design to approximately half of its original size. We begin with a description of 3D adders.

IV-A. 3D Adder Circuits

For each of our 3D adder implementations, we performed a 3D bit-slicing where odd operand bits reside on one die, and even operand bits reside on the other die. Figure 4 shows an 8-bit BK adder in 3D. Every other node of the propagate-generate logic has been stacked on top of an adjacent node as indicated by the shading in Figure 4. This in turn reduces the total width of the circuit by one half, which reduces the length of each inter-node wire by one half as well. While all of the critical wires have been halved in length, some now contain D2D vias. The additional RC overhead of the D2D via is relatively small and is more than offset by the corresponding wire length reduction. The BK adder needs only $O(N)$ vias for an N -bit adder, which is easily satisfied by the dense D2D interface.

For the faster Sklansky and Kogge-Stone adders, we also stack adjacent processing nodes which in turn reduces overall footprint and critical wire lengths by one half. Figure 5 shows an 8-bit SK adder in 3D and Figure 6 shows an 8-bit KS adder in 3D. The SK adder requires D2D vias on half of the wires at each level, which totals to $O(N \lg N)$ D2D vias. The KS adder however only needs D2D vias for the first level of the propagate-generate logic, which adds up to only $O(N)$ D2D vias. For both cases, the compact 3D layout reduces critical wire lengths which in turn provides latency and power benefits.

IV-B. 3D Barrel Shifter

Similar to the adders, we 3D-partitioned the barrel shifter such that adjacent nodes are vertically stacked on different die, thus shrinking the circuit to half its original size. This partition of odd and even bit positions on the two die halves

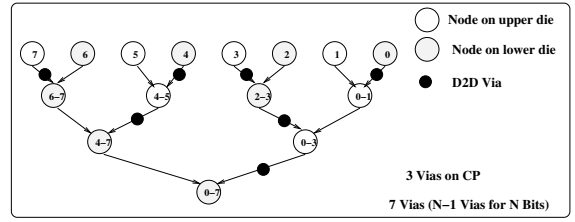


Fig. 4. 3D Brent-Kung Adder

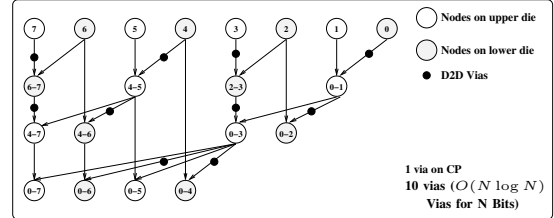


Fig. 5. 3D Sklansky Adder

the lengths of wires in every successive multiplexing level, which in turn provides greater latency and power savings with every additional layer of the circuit. The logic delays remain approximately the same as in the planar shifter design, but the wire delay component decreases quadratically with every halving of wire-length.

V. METHODOLOGY AND RESULTS

For our studies, we use circuit-level simulations using HSPICE to obtain the latency and total energy for the worst-case switching of the 2D and 3D circuits. We manually sized transistors and computed wire lengths based on a custom floorplan. We only considered CMOS gates (as opposed to domino or dual-rail). We use BSIM transistor models [18] for a 70nm technology and wire parameters extrapolated to 70nm from a TSMC 180nm technology. We model the D2D via to be $1 \mu\text{m}$ in dimension. We use distributed RC models for the inter-gate wiring, and we include parasitic capacitances such as transistor drain capacitances.

Table I shows the latencies and energy consumptions for our 2D and 3D functional units. Overall, the 64-bit adder circuits only exhibit a modest latency improvement when implemented in 3D. As the designs use more and more wire (the KS adder is the most wire-intensive design evaluated and shows the greatest latency reduction), the relative benefit does increase, thus illustrating how 3D can effectively target wire delay. Similarly, the overall energy reduction is not as great for the SK and KS adders because their large amount of logic duplication reduces the relative power contribution of the wires.

The results for the 3D barrel shifter demonstrates how 3D integration can provide greater benefits when the wire-delay component dominates the circuit. The 3D shifter exhibits almost a 9% improvement in latency while *simultaneously* reducing energy consumption by nearly 8%. For power-

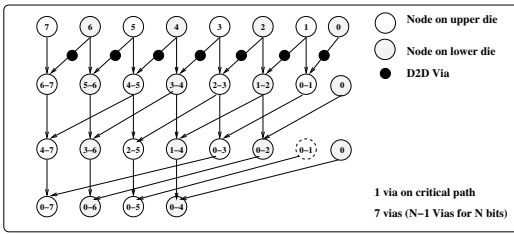


Fig. 6. 3D Kogge-Stone Adder

TABLE I

LATENCY AND ENERGY COMPARISON OF 64-BIT PLANAR AND 3D CIRCUITS. THE COLUMNS MARKED ‘%’ SHOW THE RELATIVE REDUCTION IN LATENCY AND ENERGY.

	Latency			Energy		
	2D (ps)	3D (ps)	%	2D (pJ)	3D (pJ)	%
BK adder	683.8	673.9	1.5%	11.8	11.4	3.4%
SK adder	673.3	661.1	1.8%	15.4	15.3	0.6%
KS adder	396.1	381.4	3.7%	22.7	22.1	2.6%
Shifter	559.7	510.7	8.8%	16.9	15.6	7.7%

conscious designs, the latency reduction can potentially be traded to further reduce power. For example, narrower and/or longer transistors may be used to reduce switching and leakage power while still maintaining a latency equivalent to the original 2D implementation.

VI. CONCLUSIONS

Three-dimensional integration provides significant reductions in wire lengths and overall circuit footprint. This paper has demonstrated the impact of 3D on two different types of functional units. For circuits such as barrel shifters which are dominated by wire-delay, 3D can potentially provide simultaneous benefits for performance and power. On the other hand, logic-bound circuits like adders do not gain as much from a 3D implementation, and 3D may be better utilized to reduce the wire-lengths between such circuits rather than within the circuits.

The impact of 3D fabrication on the design of high-performance circuits, and especially on a complete microprocessor, is still a largely unexplored research topic. Our results suggest that researchers should focus on wire-bound processor structures such as large SRAMs including caches, branch predictors, and register files; large wire-bound CAM structures such as the dynamic instruction scheduling logic of modern superscalar processors [19]; and long wire-dominated critical paths such as the branch misprediction notification path [20] and the result bypass network. We believe that the ability of 3D to provide simultaneous power and performance benefits will play a crucial role in extending the silicon roadmap for a few more generations.

ACKNOWLEDGMENTS

Funding and equipment for this project have been provided by Intel Corporation and a grant from the Microelectronics Advanced Research Corporation (MARCO).

REFERENCES

- [1] V. Agarwal, M. S. Hrishikesh, S. W. Keckler, and D. Burger, “Clock Rate Versus IPC: The End of the Road for Conventional Microarchitectures,” in *Proceedings of the 27th International Symposium on Computer Architecture*, Vancouver, Canada, June 2000, pp. 248–259.
- [2] R. Ho, K. W. Mai, and M. A. Horowitz, “The Future of Wires,” *Proceedings of the IEEE*, vol. 89, no. 4, pp. 490–504, April 2001.
- [3] R. Ronen, A. Mendelson, K. Lai, S.-L. Lu, F. Pollack, and J. P. Shen, “Coming Challenges in Microarchitecture and Architecture,” *Proceedings of the IEEE*, vol. 89, no. 3, pp. 325–340, March 2001.
- [4] Y. Deng and W. Maly, “2.5D System Integration: A Design Driven System Implementation Schema,” in *Proceedings of the Asia South Pacific Design Automation Conference*, Yokohama, Japan, January 2004, pp. 450–455.
- [5] S. Gupta, M. Hilbert, S. Hong, and R. Patti, “Techniques for Producing 3D ICs with High-Density Interconnect,” in *Proceedings of the 21st International VLSI Multilevel Interconnection Conference*, Waikoloa Beach, HI, USA, September 2004.
- [6] K. Puttaswamy and G. H. Loh, “Implementing Caches in a 3D Technology for High Performance Processors,” in *Proceedings of the International Conference on Computer Design*, San Jose, CA, USA, October 2005.
- [7] Y.-F. Tsai, Y. Xie, N. Vijaykrishnan, and M. J. Irwin, “Three-Dimensional Cache Design Using 3DCacti,” in *Proceedings of the International Conference on Computer Design*, San Jose, CA, USA, October 2005.
- [8] P. Reed, G. Yeung, and B. Black, “Design Aspects of a Microprocessor Data Cache using 3D Die Interconnect Technology,” in *Proceedings of the International Conference on Integrated Circuit Design and Technology*, Austin, TX, USA, May 2005, pp. 15–18.
- [9] B. Black, D. Nelson, C. Webb, and N. Samra, “3D Processing Technology and its Impact on IA32 Microprocessors,” in *Proceedings of the 22nd International Conference on Computer Design*, San Jose, CA, USA, October 2004, pp. 316–318.
- [10] K. W. Guarini, A. W. Topol, M. Jeong, R. Yu, L. Shi, M. R. Newport, D. J. Frank, D. V. Singh, G. M. Cohen, S. V. Nitta, D. C. Boyd, P. A. O’Neil, S. L. Tempest, H. B. Pogge, S. Purushothaman, and W. E. Haensch, “Electrical Integrity of State-of-the-Art 0.13 μ m SOI CMOS Devices and Circuits Transferred for Three-Dimensional (3D) Integrated Circuit (IC) Fabrication,” in *Proceedings of the International Electron Devices Meeting*, December 2002, pp. 943–945.
- [11] Tezzaron Semiconductor, “WWW Site,” <http://www.tezzaron.com>.
- [12] Samsung Electronics Corporation, “Samsung Electronics Develops World’s First Eight-die Multi Chip Package for Multimedia Cell Phones,” January 10 2005, press Release from <http://www.samsung.com>.
- [13] P. Wilkerson, A. Raman, and M. Turowski, “Fast, Automated Thermal Simulation of Three-Dimensional Integrated Circuits,” in *Proceedings of the 9th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, Las Vegas, NV, USA, June 2004, pp. 706–713.
- [14] T.-Y. Chiang, K. Banerjee, and K. C. Saraswat, “Compact Modeling and SPICE-Based Simulation for Electrothermal Analysis of Multilevel ULSI Interconnects,” in *Proceedings of the International Conference on Computer-Aided Design*, 2001.
- [15] R. P. Brent and H. T. Kung, “A Regular Layout for Parallel Adders,” pp. 260–264, March 1982.
- [16] J. Sklansky, “Conditional Sum Addition Logic,” *IRE Transactions on Electronic Computers*, vol. 9, no. 2, pp. 226–231, June 1960.
- [17] P. M. Kogge and H. S. Stone, “A Parallel Algorithm for the Efficient Solution of a General Class of Recurrence Equations,” pp. 786–793, August 1973.
- [18] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, “New Paradigm of Predictive MOSFET and Interconnect Modeling for Early Circuit Design,” in *Proceedings of the 2000 Custom Integrated Circuits Conference*, Orlando, FL, USA, May 2000, pp. 201–204.
- [19] S. Palacharla, “Complexity-Effective Superscalar Processors,” Ph.D. dissertation, University of Wisconsin, 1998.
- [20] G. Hinton, D. Sager, M. Upton, D. Boggs, D. Carmean, A. Kyler, and P. Roussel, “The Microarchitecture of the Pentium 4 Processor,” *Intel Technology Journal*, Q1 2001.