Scalability of 3D-Integrated Arithmetic Units in High-Performance Microprocessors

Kiran Puttaswamy† and Gabriel H. Loh‡
School of Electrical and Computer Engineering†
College of Computing‡
Georgia Institute of Technology
Email: kiranp@ece.gatech.edu, loh@cc.gatech.edu

ABSTRACT

Three-Dimensional integration provides a simultaneous improvement in wire-related delay and power consumption of microprocessor circuits. Prior work has looked at the performance, power, and area benefits of 3D integration technology. In this paper, we investigate the scalability issues of 3D die-stacked arithmetic units. We explore the behavior of the 3D-integrated arithmetic circuits with increasing issue-width (parallel execution capability), transistor sizing, and temperature. We show that the 3D-integrated units have a lower latency degradation and lower rate of increase in energy consumption than planar circuits with increasing issue-widths and operating temperatures. We demonstrate that the 3D-integrated circuits have less sensitivity to transistor sizing than the planar circuits.

Categories and Subject Descriptors

B.2.2 [**Hardware**]: Performance Analysis and Design Aids

General Terms

Design, Die-stacked 3D integration, Arithmetic unit

Keywords

Issue-width, Scalability, Frequency, Temperature

1. INTRODUCTION

To exploit instruction level parallelism in applications, modern high-performance microprocessors execute multiple arithmetic operations simultaneously, usually within a clock cycle. For example, the Intel Core2 processor is reported to have about ten integer units and six floating point units [1]. Usually, the latencies of arithmetic units define a lower bound on the clock cycle of the microprocessor [2]. The number of arithmetic units that can simultaneously execute (i.e., issue-width) define an upper bound on the extraction of application parallelism by the microprocessor. Thus, designs of arithmetic units such as adders, shifters, and multipliers are critical in deciding the overall performance of microprocessors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
DAC 2007, June 4–8, 2007, San Diego, California, USA.
Copyright 2007 ACM ACM 978-1-59593-627-1/07/0006 ...$5.00.

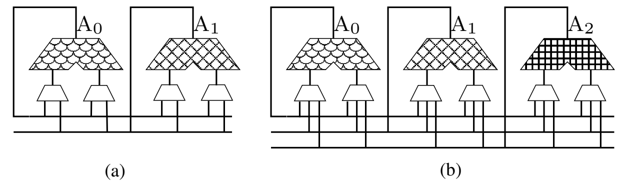

Figure 1: Bypass wiring complexity (a) $IW = 2$ (b) $IW = 3$

Traditionally, arithmetic unit designers identify one or more critical (longest) paths through the overall arithmetic computation and speed up those critical paths with innovative designs and/or transistor sizing on the critical path circuits. With increased transistor sizes, the current drive capability of the transistor increases which reduces the delay of the circuit. The undesirable effect of increasing transistor sizes is the increase in gate capacitances and leakage current, leading to increased power consumption and worsening power density issues. So, transistor sizing has to be balanced with power budget and thermal considerations.

Large issue widths enable microprocessors to execute more arithmetic operations simultaneously. Typically, an arithmetic operation (such as addition) requires at least two source operands. The source operands of a particular arithmetic operation are themselves the results generated by other arithmetic operations. If an instruction C has one of its source operands generated by instruction A and another source operand generated by instruction B, then C is said to be dependent on A and B. The results generated by A and B may be bypassed to the arithmetic unit executing C, to speed up the execution. As we increase the issue-width IW of the microprocessor, a given source operand for a given arithmetic operation can be generated by any of the IW arithmetic units. The circuitry required to bypass the results from each arithmetic unit to all the arithmetic units quickly dominates the delay. Figure 1 shows the increasing wire complexity when the issue width IW is increased from two (A_0 and A_1) to three (A_0, A_1, and A_2).

Given that technology scaling has created an ever widening gap between the relative delay of logic and wires [3, 4],

increasing the issue-width worsens the wire delay of the bypass network, thus reducing the benefits of multiple functional units on the overall performance of the processor. Three-Dimensional integrated circuit (3D IC) technology can reduce the wire delay by vertically stacking multiple die and connecting the stacked die with a high-density, low latency die-to-die (D2D) interconnect interface. There has recently been a great deal of interest in 3D ICs [5, 6, 7, 8]. Puttaswamy and Loh [5, 9, 10] have proposed designs of various SRAM and CAM-based components and arithmetic units of the microprocessor using a die-stacked 3D technology. Xie et al. [11] have explored the design space of 3D architectures and shown microarchitectural trade-offs for two processor case studies based on an Alpha processor and a Pentium processor. In this paper, we explore the scalability issues of 3D-integrated arithmetic circuits built in a die-stacked 3D technology [12, 13]. In particular, we show that the 3D-integrated arithmetic units have better scalability than the planar arithmetic units, with increasing issue-widths, transistor sizing, and operating temperatures.

The rest of the paper is organized as follows: Section 2 describes the designs of planar and 3D-integrated arithmetic circuits. Section 3 presents the results. Section 4 makes some concluding remarks.

2. ARITHMETIC UNIT DESIGNS

We focus on three different arithmetic units, a Kogge-Stone adder, a barrel shifter, and a carry-save algorithm based array multiplier, due to their varying logic and wire requirements.

The Kogge-Stone adder belongs to a general class of adders called parallel-prefix adders. In parallel-prefix adders, addition is carried out in three steps, namely pre-processing, carry generation and post-processing. In the pre-processing step, special signals called generate and propagate signals are extracted from the input bits. In the carry generation step, the generate and the propagate signals are used to compute multiple carry bits simultaneously. In the post-processing step, the computed carry bits are used to compute the sum bits, providing the final sum of the addition.

The barrel shifter enables shifting the bit positions of the input N-bit number by any number of positions up to N-1. Besides a shift operation being a part of most instruction set architectures, the shifting is also used in floating-point arithmetic to align exponents and fractions. The barrel shifter is a wire-intensive structure since any input bit has to be capable of being routed to any of the other N-1 positions within one clock cycle.

Our array multiplier uses a well-known multiplier algorithm called the carry-save algorithm. Figure 2(a) shows a 4×4 carry-save array (CSA) algorithm implementation. In the carry-save technique, the carry information from adding the rows of partial products is not combined with the sum information until the very last step. Note that the sum arrays and carry arrays are being separately propagated until the end, where a carry-completing adder merges the final sum array and the carry array to produce the final product bits. Figure 2(b) shows the multiplier circuit and highlights two of the critical paths. The critical paths are dominated by propagation between the different rows and a short component of carry ripple in the last row.

For 3D-integrated circuits, our design objective is to reduce the wirelengths, especially those on the critical paths.

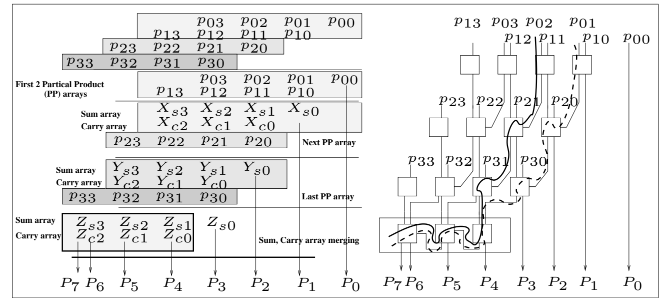

Figure 2: (a) Carry-save array (CSA) multiplier algorithm (b) CSA multiplier design (two critical paths highlighted)

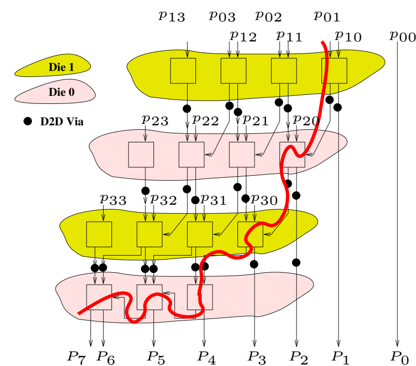

Figure 3: 3D Carry save array multiplier

In case of the adder and the shifter, we make use of odd/even bit-partitioned 3D designs as previously proposed in [9]. This partition of odd and even bits on the vertically stacked die reduces the wirelengths in successive logic levels by replacing some of the wires on the critical paths with short Die-to-die (D2D) vias. Figure 3 shows our proposed design of the 3D CSA multiplier. The 3D CSA multiplier replaces wires on the critical path with the D2D vias as shown in Figure 3, thus reducing the circuit latency and power.

3. RESULTS

For our studies, we run circuit simulations using HSPICE to obtain the latency and total energy for the planar and the 3D circuits. For HSPICE simulations, we use PTM transistor models [14] for a 65nm technology. We model the D2D vias to be 10μm length and 1μm pitch.

Table 1 shows the relative delay improvement of various 3D-integrated circuits over the planar circuits. The shifter, being the most wire dominated circuit derives the maximum benefit from 3D-integration. The adder and the multiplier circuits, being less wire-intensive, derive less benefit than the shifter circuits. Also, as the issue-widths increase to eight, the latency benefits increase up to 51% for the barrel shifter.

Next, we explore the frequency scalability of the 3D adders and planar adders with increasing issue-widths. Table 2 shows the operating frequency of the adders for increasing issue-widths. As issue-widths increase from two to eight, the

Table 1: Percent improvement in delays of 3D-integrated arithmetic units over planar units.

Issue width	% Adder	% Shifter	% Multiplier
2	15.4	30.1	12.3
3	18.2	36.2	14.7
4	21.1	40.7	16.9
5	24.6	44.2	18.6
6	25.7	47.3	21.0
7	28.0	48.2	22.8
8	30.2	51.1	24.6

Table 2: Frequency scalability of planar and 3D adders with increasing issue-widths.

Issue width	Planar Frequency (GHz)	3D Frequency (GHz)	Percent improvement
2	3.95	4.67	18.24
3	3.69	4.51	22.19
4	3.43	4.35	26.69
5	3.20	4.24	32.68
6	3.05	4.10	34.55
7	2.87	3.99	38.95
8	2.73	3.91	43.18
Frequency loss	30.8%	16%	

frequency reduces by 30.8% for planar adders and only by 16% for 3D adders, demonstrating the wire relief provided by 3D technology. Also, the potential frequency improvement can be up to 43% for the 3D-integrated adders as compared to the planar adders.

Figure 4 shows the latency trends due to transistor sizing, for a four-issue processor configuration. The drive-strengths as well as the gate loads increase with increased transistor sizing. Until the drive-strengths reach the optimal capacity to drive the gate loads and the wire loads, the overall latency reduces for both the planar and the 3D units. Once the circuit has achieved sufficient drive-strength, further increase in the transistor sizes increases the gate load leading to latency degradation. The latency curve of the 3D circuits shows that the 3D implementations exhibit slower latency degradation than the corresponding planar circuits, due to reduced wires. This has important implications with continuous technology scaling. Prior research [15] has already identified that transistor geometries will become increasingly non-deterministic due to increased influence of process variations in current and future technology generations. Hence, 3D implementations are more likely to meet their delay target in the presence of process variations. Next, we consider the power consumption behavior of the 3D circuits in Figure 5 for increasing issue-widths. For smaller transistor sizes, we can see from Figure 5 that the issue-widths have less influence on the power consumption. As the transistor sizes increase, we see an increase in the power consumption for all issue widths, since both the switching power and the leakage power increase. For larger transistor sizes, the influence of issue-width becomes more noticeable. In particular, higher issue-widths have lower power consumption than lower issue-widths, for a given transistor

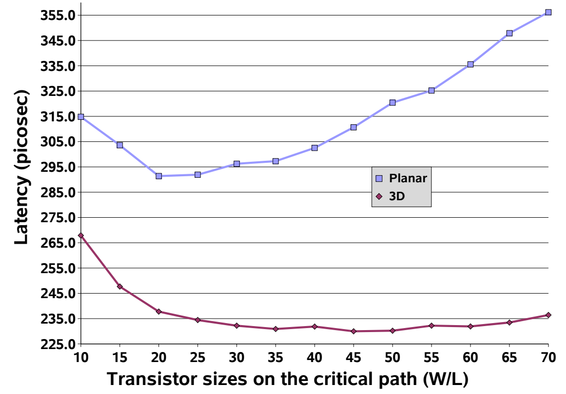

Figure 4: Latency vs. transistor sizing for a four-issue processor.

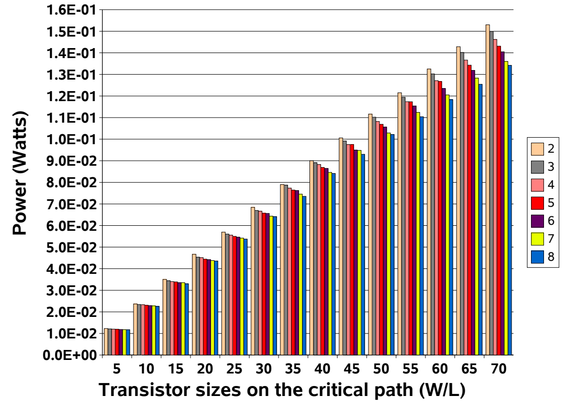

Figure 5: Effect of transistor sizing and issue-width on power.

sizing. This is due to the fact that higher issue-width circuits operate at lower frequencies than the lower issue width circuits (as shown in Table 2). Thus, increasing issue-widths have opposite effects on the delay and the power consumption. The higher the issue-widths, the larger the delay and lower the power consumption.

Figure 6 shows the comparison of energy consumption for the planar and the 3D circuits. Note that the planar circuits have a much higher increase in the total energy consumption as compared to 3D circuits, with increasing issue-widths. Planar circuits have a much higher rate of increase in delay than the rate of decrease in power, thus causing the energy to increase continuously. 3D circuits manage to balance the increase in delay by a corresponding decrease in power, thus keeping the overall increase to a small negligible amount. Thus, 3D implementations have better energy scalability with increasing issue-widths than the planar circuits.

Next, we look at the latency degradation due to increased operating temperature of the circuits. Figure 7 shows the comparison of latency degradation of the planar and the 3D circuits when the temperature of operation is varied from

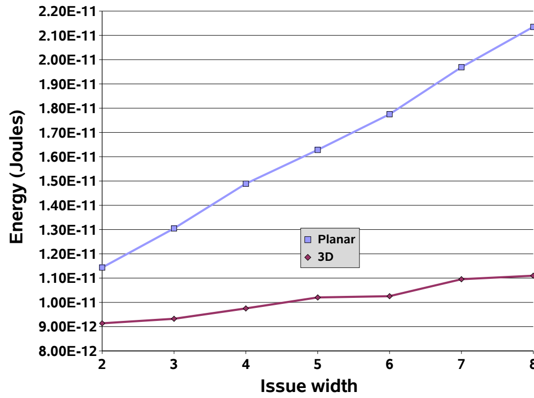

Figure 6: Energy comparison of planar and 3D circuits.

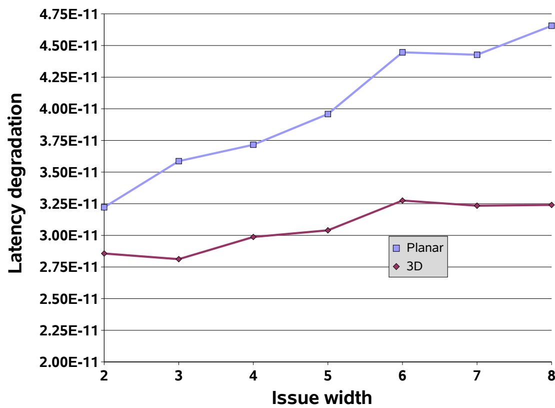

Figure 7: Latency degradation of planar and 3D circuits from 25C to 100C.

25 C to 100 C. When the circuits operate at higher temperatures, they experience latency degradation due to reduced mobility and increased wire resistances. Since the 3D-integrated circuits reduce the wires, they experience a lower rate of latency degradation than the planar circuits, for increasing issue-widths.

4. CONCLUSION

3D integration provides significant reductions in wire delay and power. In this paper, we demonstrated that the 3D-integrated arithmetic units have better scalability than the planar circuits, in the face of increasing issue-widths, process variations, and operating temperatures. The relative benefits of 3D technology will increase in future technology generations, making it a very attractive option for future designs. We believe that the better scalability of 3D circuits will play a crucial role in extending the silicon road-map 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).

5. REFERENCES

[1] Sandpile, "WWW Site," `http://www.sandpile.org`.

[2] S. Palacharla, "Complexity-Effective Superscalar Processors," Ph.D. dissertation, University of Wisconsin, 1998.

[3] ITRS, "International Technology Roadmap for Semiconductors," from `http://www.itrs.net`.

[4] S. Borkar, "Design Challenges of Technology Scaling," *IEEE Micro Magazine*, vol. 19, no. 4, pp. 23–29, July 1999.

[5] 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.

[6] 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.

[7] 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.

[8] Tezzaron Semiconductor, "WWW Site," `http://www.tezzaron.com`.

[9] K. Puttaswamy and G. H. Loh, "The Impact of 3-Dimensional Integration on the Design of Arithmetic Units," in *Proceedings of the International Symposium on Circuits and Systems*, Kos, Greece, May 2006, pp. 4951–4954.

[10] ——, "Dynamic Instruction Schedulers in a 3-Dimensional Integration Technology," in *Proceedings of the ACM Great Lakes Symposium on VLSI*, 2006.

[11] Y. Xie, G. H. Loh, B. Black, and K. Bernstein, "Design space exploration for 3d architectures," *J. Emerg. Technol. Comput. Syst.*, vol. 2, no. 2, pp. 65–103, 2006.

[12] P. Morrow, M. J. Kobrinsky, S. Ramanathan, C.-M. Park, M. Harmes, V. Ramachandrarao, H. mog Park, G. Kloster, S. List, and S. Kim, "Wafer-Level 3D Interconnects Via Cu Bonding," in *Proceedings of the 21st Advanced Metallization Conference*, San Diego, CA, USA, October 2004.

[13] R. Reif, A. Fan, K.-N. Chen, and S. Das, "Fabrication Technologies for Three-Dimensional Integrated Circuits," in *Proceedings of the 3rd International Symposium on Quality Electronic Design*, San Jose, CA, USA, March 2002, pp. 33–37.

[14] 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.

[15] K. Bowman, S. Duvall, and J. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 2, pp. 183 – 189, Feb 2002.