Improving Microprocessor Performance through 3-D IC Technology

Paper and poster presentations at SRC TECHCON 2005
Presenter: Christianto C. Liu, Cornell University
Co-authors: Ilya Ganusov, Martin Burtscher, and Sandip Tiwari, Cornell University

Short Abstract

The growing gap between logic and memory performance has forced microprocessors to employ increasingly complex designs, including out-of-order execution and speculation, as well as memory hierarchies with more levels and larger caches to hide the main memory latency. In this paper, we examine how 3-D IC technology can be effective in improving the interactions between processors and their memory. We simulate SPEC2000 integer and floating-point programs on an extended version of the SimpleScalar v4.0 simulator. The results reveal that stacking an L2 cache on top of the CPU provides little benefit, whereas stacking main memory shows significant performance improvements. We also investigate the benefits of very large L2/L3 caches, which can be stacked on top of the CPU through 3-D technology. In addition, we show that stream prefetching is an effective technique to prefetch data into these large caches.

Extended Abstract

Microprocessors have been improving in performance at the rate of roughly 60% per year. On the other hand, memory access time has improved by less than 10% per year [1]. The resulting gap between logic and memory performance has forced microprocessor designs toward complex and power-hungry architectures such as out-of-order execution and speculation techniques. Moreover, processors have been designed with increasingly large cache hierarchies to hide the latency of main memory. In this paper, we examine how three-dimensional integrated circuit (3-D IC) technology can improve the interactions between the processor and memory.

In 3-D ICs (e.g., [2]), planar device layers are stacked, one on top of the other, in a 3-D structure where adjacent device planes can be connected by short, vertical wires (Fig. 1). As opposed to a conventional 2-D chip, where logic and memory units are situated on opposite ends of the chip, logic and memory in a 3-D chip can be stacked above each other to shorten the critical path. More importantly, by not having to go off-chip, a large performance improvement can be obtained [3]. On-chip DRAM memory macros can be designed without address multiplexing and without off-chip drivers and receivers. Furthermore, the bus turnaround time can be eliminated through dedicated read and write busses. The bandwidth can be dramatically improved since on-chip wiring is not pin-limited (as it is with off-chip accesses) and the CPU is able to fetch hundreds to thousands of bits from the DRAM at once, if necessary. Power consumption is also reduced because the large sources of off-chip capacitance are removed by keeping the memory on-chip.

We evaluate the performance gains due to 3-D technology using an extended version of the SimpleScalar v4.0 simulator [4]. The baseline 2-D processor core is representative of current technology (2 GHz CPU, 500 MHz memory, 64 KB L1I, 64 KB L1D, 1 MB L2). Details of the

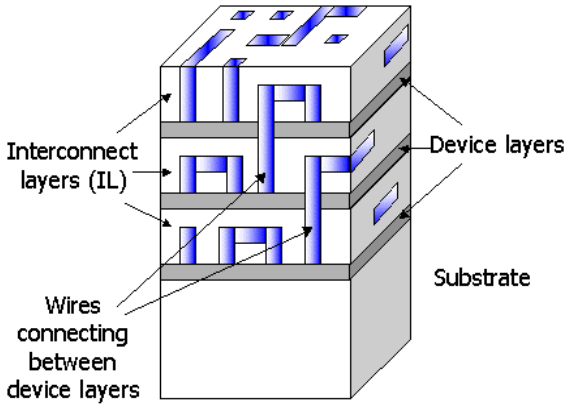

Fig. 1. Three-dimensional integrated circuits (3-D ICs) consist of planar device layers separated by silicon dioxide (or other insulating materials) and interconnected by short vertical wires.

simulated processor are not presented here due to space considerations. We use all 12 integer and 10 of the 14 floating-point programs from the SPECcpu2000 benchmark suite [5] for this study. The programs are run with the SPEC-provided reference inputs. We use the SimPoint toolset [6] to identify representative simulation points. Each program is simulated for 500 million instructions after fast-forwarding past the number of instructions determined by SimPoint.

We examine several different scenarios. First, we consider stacking the L2 cache on top of the CPU. However, the benefit of stacking a standard (a few MBs) L2 cache is small because proper floorplanning would have already placed the L2 strategically to minimize critical paths (e.g., to the L1 cache). Also, even if 3-D does reduce the latency of the L2 cache, the reduction is at most 1-2 cycles out of the roughly 10 cycle load-to-use latency that is typical for today's L2 caches. This reduction results in little performance improvement (Fig. 2). On the other hand, putting main memory on-chip provides a significant boost in performance. We also demonstrate the performance benefits of stacking very large L2/L3 caches, attainable through 3-D technology, as well as the effective use of stream prefetching [7] to preload data into these large caches. Due to lack of space, results of large L2/L3 caches and stream prefetching are not presented here, and will be addressed during the conference.

References:

[1] J. L. Hennessy and D. A. Patterson, *Computer Organization and Design*, 2nd ed., Morgan Kaufmann Publishers, San Francisco, 1997.

[2] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," *Proc. IEEE*, vol. 89, pp. 602-633, May 2001.

[3] R. E. Matick and S. E. Schuster, "Logic-based eDRAM: origins and rationale for use," *IBM J. Res. & Dev.*, vol. 49, pp.145-165, Jan. 2005.

[4] E. Larson, S. Chatterjee, and T. Austin. "Mase: a novel infrastructure for detailed microarchitectural modeling," in Proc. of the Second International Symposium on Performance Analysis of Systems and Software, 2001, pp. 1-9.

[5] *SPEC CPU2000 V1.2*, Standard Performance Evaluation Corporation, [Online]. Available: http://www.spec.org/osg/cpu2000, 2001.

[6] T. Sherwood, E. Perelman, G. Hamerly, and B. Calder, "Automatically characterizing large scale program behavior," in Proc. of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems, 2002, pp. 45-57.

[7] S. Palacharla and R. E. Kessler, "Evaluating stream buffers as a secondary cache replacement," in Proc. of the 21st Annual International Symposium on Computer Architecture, 1994, pp. 24-33.

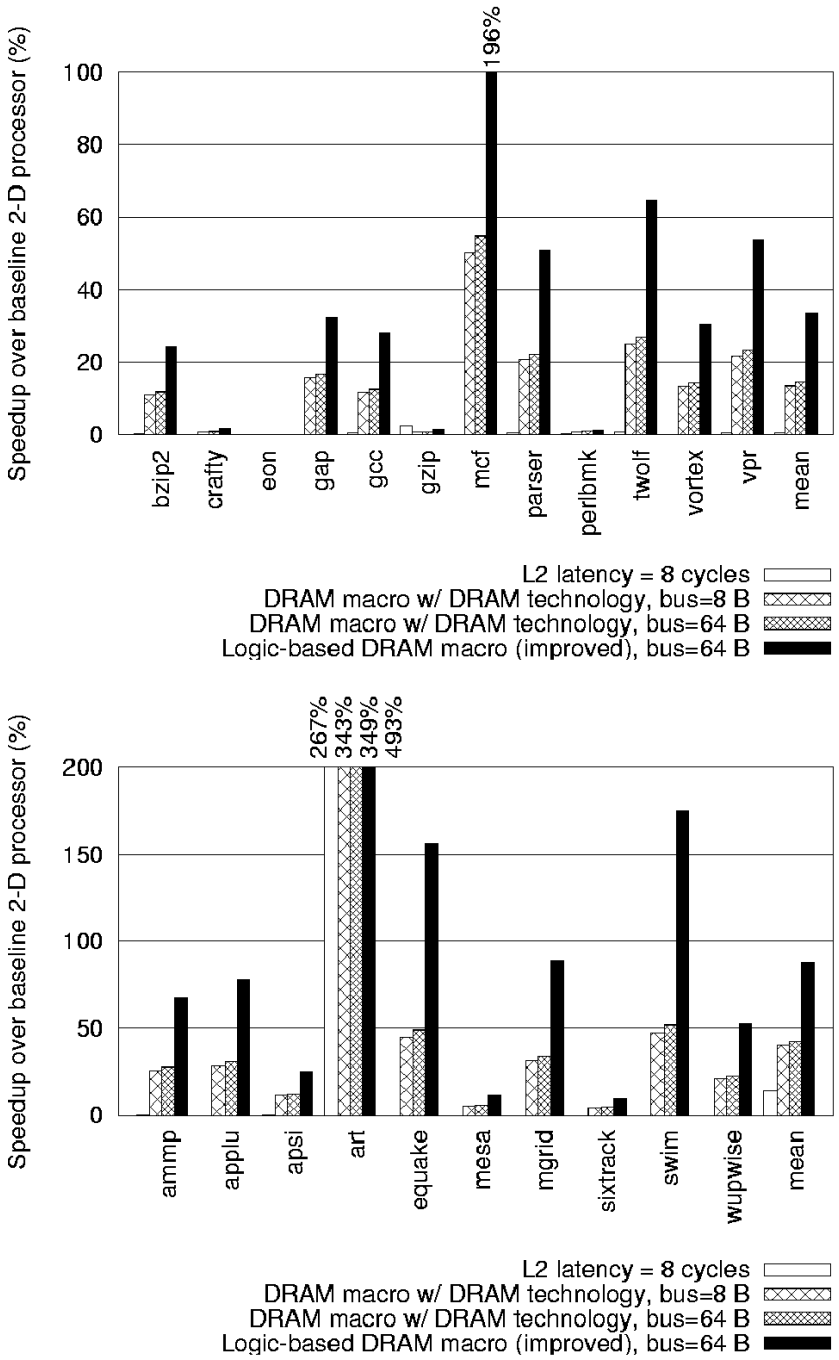

Fig. 2. Integer (top) and floating-point (bottom) benchmark performance of 3-D processors with (1) hypothetical 2-cycle reduction in L2 hit latency, and (2) three possible scenarios of on-chip, stacked main memory with different memory bus widths. The last case represents logic-based DRAM (less dense but much faster than standard DRAM) with improved memory controller capabilities.