RETROSPECTIVE: Victim Replication: Maximizing Capacity while Hiding Wire Delay in Tiled Chip Multiprocessors

Michael Zhang
Great Lake Advisors
Hillsborough
CA 94010

Krste Asanović
University of California Berkeley
Berkeley, CA 94720
krste@berkeley.edu

Following dramatic performance growth in the 1980s and 1990s, by the turn of the century microprocessors had evolved into large high-frequency, out-of-order superscalar uniprocessors that stretched the limits of microarchitecture, circuit design, and fabrication technology. The combination of worsening relative wire delay and the end of Dennard power-density scaling led to industry reluctantly embracing the reality that future general-purpose microprocessors would have to be multiprocessors, and that software developers would have to exploit parallelism to achieve future substantial performance improvements.

When we began our work on chip-scale multiprocessors in the early 2000s, there were a variety of proposals on how they should be structured. Many took the obvious (and sensible) approach of replicating well-known large-scale multiprocessor architectures and scaling them down to fit on a chip. Most designs assumed a "dance-hall" structure with processors on one side and cache banks on the other side of a central communication crossbar or fabric. However, we were interested in exploring the new opportunities and challenges as per-chip core count grew and technology scaling effects continued. In particular, we believed it would be more natural and scalable for future microprocessors to construct a tiled chip-scale multiprocessor, where each tile contained a processor, router, private caches, and a slice of the last-level cache and coherence manager. While we were inspired by the MIT RAW project [4] and other replicated architectures, these did not address hardware coherence, and we focused on how to build a large coherent chip-scale multiprocessor in this tiled style.

At the same time, others had been focused on the effects of wire delay on large caches, and had proposed various schemes to exploit non-uniform cache access (NUCA) delays by moving a cache line between different locations on a chip to reduce effective access latency. However, we found these schemes to be complex and unlikely to be practical. There were also unsolved challenges in where to place shared lines in multiprocessor variants. We realized that in the context of our tiled CMP structure, which was based on directory coherence, that we could allow there to be multiple copies of a cache line in different slices of the same cache [6]. We developed a simple cache policy to try to place lines evicted for capacity or conflict reasons into the local slice of the last-level cache, effectively repurposing that cache slice as a very large local victim cache. This scheme was both simple and effective, and adapted well across a variety of workloads, from single-threaded, to multi-threaded, to multi-programmed. While preparing the ISCA paper, we realized we could improve the scheme still further by removing the need to have a copy of the line at the home node, instead adopting a non-inclusive policy in the last-level cache, but unfortunately did not persist in trying to publish this work except as a technical report [5].

Looking back, although perhaps not readily apparent from the paper, most of the effort was developing new simulator infrastructure for this class of system and bringing up a representative set of benchmarks. In particular, we had to develop a new multiprocessor sampling methodology [1] to make our simulations tractable. At the time, we received some interest from various industry groups who were trying to figure out how to scale past a few cores per chip. One industry group built a large-scale emulation with a very large number of cores and confirmed our results, but were actually more excited by the energy savings from victim replication rather than the latency reduction. At the time, we were not able to perform a power analysis with our own simulation infrastructure or scale up to large core counts. The limitations of software-based multiprocessor simulation infrastructures later inspired our FPGA-based multiprocessor modeling work [3], and ultimately the FireSim project [2].

Today, around twenty years after we started this work, chip core counts are growing dramatically, particularly in the server space, and tiled CMPs are a common structure with various optimizations similar to victim replication and migration in use by some vendors.

References

[1] K. Barr, H. Pan, M. Zhang, and K. Asanović, "Accelerating multiprocessor simulation with a memory timestamp record," in *IEEE International Symposium on Performance Analysis of Systems and Software*, Austin, TX, March 2005.

[2] S. Karandikar, H. Mao, D. Kim, D. Biancolin, A. Amid, D. Lee, N. Pemberton, E. Amaro, C. Schmidt, A. Chopra, Q. Huang, K. Kovacs, B. Nikolić, R. Katz, J. Bachrach, and K. Asanović, "FireSim: FPGA-accelerated cycle-exact scale-out system simulation in the public cloud,"

in *International Symposium on Computer Architecture*, Los Angeles, CA, June 2018.

[3] Z. Tan, A. Waterman, H. Cook, S. Bird, K. Asanović, and D. Patterson, "A case for FAME: FPGA architecture model execution," in *International Symposium on Computer Architecture*, Saint-Malo, France, June 2010.

[4] M. B. Taylor, W. Lee, J. Miller, D. Wentzlaff, I. Bratt, B. Greenwald, H. Hoffmann, P. Johnson, J. Kim, J. Psota, A. Saraf, N. Shnidman, V. Strumpen, M. Frank, S. Amarasinghe, and A. Agarwal, "Evaluation of the Raw microprocessor: An exposed-wire-delay architecture for ILP and Streams," in *International Symposium on Computer Architecture*, Munich, Germany, 2004.

[5] M. Zhang and K. Asanović, "Victim Migration: Dynamically adapting between private and shared CMP caches," MIT Computer Science and Artificial Intelligence Laboratory, Tech. Rep. MIT-CSAIL-TR-2005-64, October 2005.

[6] M. Zhang and K. Asanović, "Victim Replication: Maximizing capacity while hiding wire delay in tiled chip multiprocessors," in *International Symposium on Computer Architecture*, Madison, WI, June 2005.