RETROSPECTIVE: Cache Decay: Exploiting Generational Behavior to Reduce Cache Leakage Power

Stefanos Kaxiras, Zhigang Hu, Margaret Martonosi
Uppsala University, Shanghai ETiger Capital Partners LLC, Princeton University

I. Setting

The story behind the 2001 ISCA paper, "Cache Decay," takes us back a few years prior to 2001. During that time, there were clear indications that a significant portion (about 90% in some cases) of the data in a cache was not going to be accessed before being replaced (Doug Burger, Alain Kagi and Jim Goodman [1]). However, finding what to do with this was a challenge. Concurrently, the rise of leakage power consumption was becoming a growing concern within the computer architecture community. One of the early papers on leakage power, "$Gated V_{dd}$," introduced a mechanism that could be used by architects to address leakage at the architectural level [5]. The case study for $Gated V_{dd}$ was to downsize the cache by turning off parts of it, but that meant that performance degraded quickly. The concept of cache decay originated within this backdrop.

II. The Bell Labs-Princeton connection

Prior to the 2001 ISCA paper, Stefanos Kaxiras, who had graduated from Wisconsin-Madison in 1998, had joined Bell Labs (Murray Hill, NJ). Seeking to establish connections in the area and intrigued by Margaret Martonosi's work on power efficiency at the architectural level, Stefanos, along with other members of Bell Labs, visited Princeton during a Princeton Industry Affiliates meeting in 1999. This marked the beginning of the collaboration between Stefanos and Margaret. To kickstart their work on interesting projects, Zhigang Hu, a student of Margaret, who was seeking an industry internship, decided to spend the summer of 2000 at Bell Labs working with Stefanos.

III. Cache Decay

While attending ISCA 27 in June of 2000, Stefanos conceived the initial concept of Cache Decay, envisioning the cache lines to behave like DRAM and gradually leak away after they cease to be accessed. Upon returning to New Jersey, without fully appreciating the significance of the idea, Stefanos embarked on two projects with Zhigang. The first project, unrelated to Cache Decay, focused on utilizing simultaneous multithreading (SMT) for power efficiency, resulting in a 2001 paper published in CASES [4]. It was only by the middle of the summer that Stefanos and Zhigang shifted their attention to the second project, Cache Decay. Zhigang promptly established a cache model in SimpleScalar.

The idea was simple: Use $Gated V_{dd}$ to turn off individual cache lines that were no longer accessed and were simply waiting in the cache to be replaced. For this we had to somehow count the time since the last access to a cache line. If this time was sufficiently large, the cache line was deemed "dead" and could be turned off to save leakage power. The waiting time, until we call a cache line dead, was dubbed *decay interval*. However, initial results were disheartening as performance suffered more than we had anticipated. It turned out that we were exploring decay intervals that were too close to the inter-access interval of the cache lines (same order of magnitude). In fact, much larger decay intervals were needed. Despite the initial disappointment, Stefanos pushed on for a binary search for the optimal decay interval for each of the SPEC benchmarks we were using. We also used the notion of competitive algorithms to calibrate decay intervals by comparing the leakage energy of retaining data in a cache line to the alternative dynamic energy of needing to refetch it if needed. By the end of the summer we had excellent results: we were able to "turn off" about 80% of the cache with negligible impact on performance. Of course, at that point we did not even have a decent model of leakage power but that did not deter us from submitting to the "Power-Aware Computer Systems" workshop (PACS 2000) [3]. Prior to presenting the paper in the PACS workshop, Stefanos visited the Wisconsin Industry Affiliates meeting at Madison (early autumn of 2000) where he presented the idea. At that point, it became evident that Cache Decay held significant importance on various fronts. It was in the same meeting that the ISCA 2001 paper got part of its title. David Wood exclaimed after Stefanos' talk: "that's cache-line generational behavior!" His remark unmistakably alluded to his own 1991 paper that thoroughly explained this phenomenon [7].

IV. The ISCA 2001 paper

By the end of summer we were convinced that Cache Decay was a great idea and it was time to write the paper about it. To accomplish this, Stefanos made frequent commutes to Princeton, spending numerous afternoons and evenings there before driving back to NYC in the early hours of the following day. The summer time evaluation needed improvements, first and foremost needing a proper leakage model. Zhigang built a more rigorous evaluation framework, now reporting power numbers along with performance. Hierarchical decay counters were introduced to reduce the hardware cost (which translated to dynamic power!) of counting decay intervals. Adaptive Cache Decay was added to adjust the decay interval at runtime (based on application behavior) and a cost model was added by Margaret. Various problems with the design were fixed. For example, a problem first noted by Nevin Heintze (Bell Labs), of how to decay dirty cache lines (that need to be written back first), was addressed by cascading the decay counter signal to the cache sets, one at a time. The paper was submitted to ISCA in November 2000 and got easily accepted with top-

notch scores. It was presented by Zhigang in Gothenburg, Sweden. Unfortunately for Stefanos, at the last moment he had to turn back at the La Guardia Airport in NYC, and miss this international ISCA due to passport problems that prevented him from leaving the US and returning back to his family. Following the ISCA conference, Cache Decay gained increased visibility as it was shared through a series of invited talks delivered in both industrial and academic settings.

V. Impact

Looking back at Cache Decay, 22 years after the paper, we believe that it played a pivotal role in shaping research in computer architecture for the 2000s and beyond.

First, it brought attention to the leakage power problem and demonstrated that it could be effectively addressed at the architectural level through power gating. Inspired by the success of Cache Decay, numerous techniques emerged for both memory and logic structures. While process technology solutions such as high-k dielectrics, FinFET transistors, multiple threshold voltages, and adaptive body biasing ultimately played a crucial role in addressing the leakage problem at the process technology level, Cache Decay and other architecture-level techniques have provided us with a deeper understanding of the performance implications resulting from this shift to power awareness.

Second, Cache Decay introduced a lightweight technique that capitalized on dead cache lines, highlighting that dead blocks significantly outnumber live blocks. This concept spurred further research in the community, leading to the repurposing of dead cache lines for various optimizations, such as compression, coherence, and reliability. In our own work presented at ISCA 2002, we built upon the notion of measuring time and generational behavior in the cache to exploit it for cache performance optimizations, rather than solely focusing on leakage mitigation [2]. Furthermore, Cache Decay, alongside other dead-block prediction schemes, paved the way for more advanced optimizations in the cache hierarchy, including a range of cache replacement algorithms that outperformed the traditional LRU approach, which began to emerge in the late 2000s. Even today, 22 years after its publication, Cache Decay continues to find new applications, such as for example in security, offering protection against cache side-channel attacks [6].

VI. What happened to the authors?

Stefanos Kaxiras left Bell Labs in 2003, transitioning to academia. He relocated to Europe and took an assistant professor position at the ECE department of the University of Patras, Greece. In 2010, he moved to a professor position at Uppsala University, Sweden, where he played a pivotal role in building the Uppsala group's reputation as one of Europe's leading architecture groups.

After completing his Ph.D. at Princeton, Zhigang Hu became a member of the IBM Thomas J Watson Research Center. In 2007, he made a transition to the investment services sector, humorously referring to it as "cash decay" during the economic downturn of 2008. Presently, Zhigang is the founder and CEO of Shanghai ETiger Capital Partners, LLC.

Margaret Martonosi is (still!) on the faculty at Princeton University, having started as an assistant professor of Electrical Engineering in 1994. She is now the H. T. Adams '35 Professor of Computer Science. With Computer Architecture well-represented in both the ECE and CS departments, Martonosi continues to maintain strong collaborations with both ECE and CS faculty and students.

References

[1] D. C. Burger, J. R. Goodman, and A. Kagi, "The declining effectiveness of dynamic caching for general-purpose microprocessors," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 1995.

[2] Z. Hu, S. Kaxiras, and M. Martonosi, "Timekeeping in the memory system: predicting and optimizing memory behavior," *ACM SIGARCH Computer Architecture News*, vol. 30, no. 2, pp. 209–220, 2002.

[3] S. Kaxiras, Z. Hu, G. Narlikar, and R. McLellan, "Cache-line decay: A mechanism to reduce cache leakage power," in *Power-Aware Computer Systems: First International Workshop, PACS 2000 Cambridge, MA, USA, November 12, 2000 Revised Papers 1*. Springer, 2001, pp. 82–96.

[4] S. Kaxiras, G. Narlikar, A. D. Berenbaum, and Z. Hu, "Comparing power consumption of an smt and a cmp dsp for mobile phone workloads," in *Proceedings of the 2001 international conference on Compilers, architecture, and synthesis for embedded systems*, 2001, pp. 211–220.

[5] M. Powell, S.-H. Yang, B. Falsafi, K. Roy, and T. Vijaykumar, "Gated-vdd: A circuit technique to reduce leakage in deep-submicron cache memories," in *Proceedings of the 2000 international symposium on Low power electronics and design*, 2000, pp. 90–95.

[6] J. P. Thoma, C. Niesler, D. Funke, G. Leander, P. Mayr, N. Pohl, L. Davi, and T. Guneysu, "Clepsydracache - preventing cache attacks with time-based evictions," in *Proc. of 32nd USENIX Security Symposium*, Anaheim, CA, aug 2023.

[7] D. A. Wood, M. D. Hill, and R. E. Kessler, "A model for estimating trace-sample miss ratios," in *Proceedings of the 1991 ACM SIGMETRICS conference on Measurement and modeling of computer systems*, 1991, pp. 79–89.