A Two Level Neural Approach Combining Off-Chip Prediction with Adaptive Prefetch Filtering

Alexandre Valentin Jamet
alexandre.jamet@bsc.es
Barcelona Supercomputing Center (BSC)

***Abstract*—To alleviate the performance and energy overheads of contemporary applications with large data footprints, we propose the *Two Level Perceptron (TLP)* predictor, a neural mechanism that effectively combines predicting whether an access will be off-chip with adaptive prefetch filtering at the first-level data cache (L1D). TLP is composed of two connected microarchitectural perceptron predictors, named *First Level Predictor (FLP)* and *Second Level Predictor (SLP)*. FLP performs accurate off-chip prediction by using several program features based on virtual addresses and a novel selective delay component. The novelty of SLP relies on leveraging off-chip prediction to drive L1D prefetch filtering by using physical addresses and the FLP prediction as features. TLP constitutes the first hardware proposal targeting both off-chip prediction and prefetch filtering using a multi-level perceptron hardware approach. TLP only requires 7KB of storage.**

To demonstrate the benefits of TLP we compare its performance with state-of-the-art approaches using off-chip prediction and prefetch filtering on a wide range of single-core and multi-core workloads. Our experiments show that TLP reduces the average DRAM transactions by 30.7% and 17.7%, as compared to a baseline using state-of-the-art cache prefetchers but no off-chip prediction mechanism, across the single-core and multi-core workloads, respectively, while recent work significantly increases DRAM transactions. As a result, TLP achieves geometric mean performance speedups of 6.2% and 11.8% across single-core and multi-core workloads, respectively. In addition, our evaluation demonstrates that TLP is effective independently of the L1D prefetching logic.

I. Off-Chip Prediction and Prefetch Filtering

Emerging workloads spanning various domains [1], have a key property in common: massive working set sizes that do not fit in the existing cache hierarchies [4], making cache management a major performance bottleneck for processor design. Indeed, recent work [2] shows that these workloads spend up to 80% of their execution time waiting for DRAM.

To address the high-latency load requests of these emerging applications, prior work [5], [3] has introduced the concept of *off-chip prediction*. Hermes [3] is the state-of-the-art microarchitectural off-chip prediction scheme. At the core of Hermes, there is a perceptron predictor composed of several prediction tables, one per selected program feature, similar to prior work on perceptron-based microarchitectural prediction: from branch prediction [6] to cache replacement policies [7].

II. Experimental Setup

Our evaluation considers Champsim, a detailed trace-based simulator that models a 4-wide out-of-order CPU. The baseline system is similar to a Cascade lake micro-architecture. The micro-architecture simulated has either 1 or 4 cores, L1 instruction and data cache of 32KB each, an L2 cache of 1MB, and a banked L3 cache of 1.375MB per core. The system also includes a 16GB main memory based on DDR4 SDRAM with a data-rate of 3.2GB/s per core.

III. Experimental Results

In this work, we evaluate: i) the impact and limitations of the state-of-the-art off-chip prediction mechanism, Hermes [3]; and ii) the cost of inaccurate L1D prefetch requests.

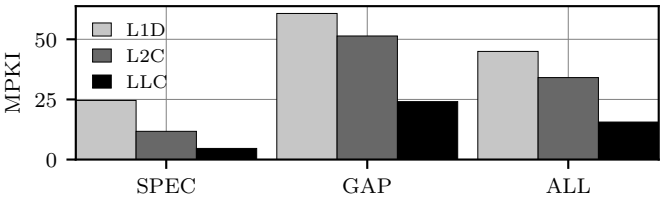

Fig. 1. MPKI of all caches (L1D, L2C, LLC) across the SPEC (SPEC CPU 2006 and SPEC CPU 2017) and GAP workloads.

Figure 1 shows the average Misses per Kilo Instruction (MPKI) rates of L1D, L2C and LLC. On average the MPKIs of L1D, L2C, and LLC are 45.0, 34.1, and 15.6, respectively. Therefore, 34.7% of L1D misses eventually require a DRAM access. Remarkably, workloads from domains such as graph processing put more pressure on the cache hierarchy, resulting in more frequent DRAM accesses. Indeed, Figure 1 reveals that, on average, the graph-processing (GAP) workloads trigger DRAM accesses for 39.7% of the L1D misses.

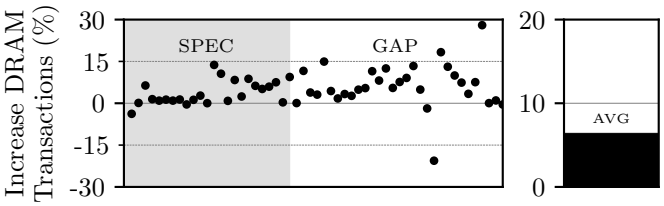

Fig. 2. Increase in DRAM transactions due to Hermes off-chip predictions relative to a baseline without off-chip prediction mechanism. Lower is better.

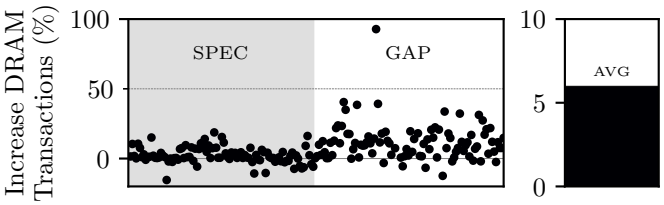

Fig. 3. Increase in DRAM transactions due to Hermes off-chip predictions relative to a baseline without off-chip prediction mechanism in the 4-core context. The x-axis ticks represent 200 different 4-core workload mixes of SPEC and GAP workloads. Lower is better.

Figures 2 and 3 indicate that Hermes places high pressure on DRAM, especially in the multi-core scenario, since it issues

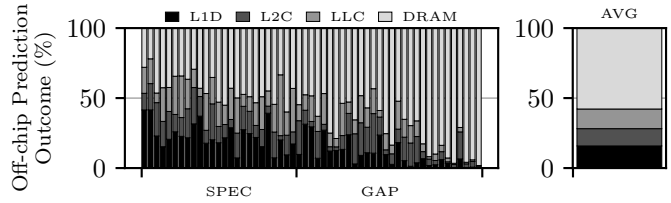

Fig. 4. Location of a block upon a Hermes off-chip prediction.

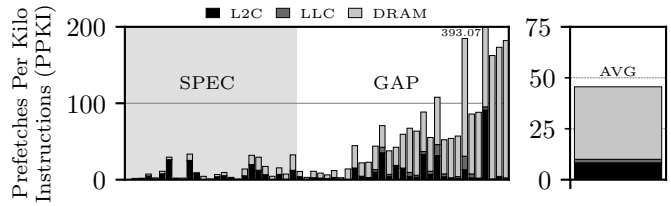

Fig. 5. Location where the inaccurate L1D prefetch requests are served across two state-of-the-art L1D prefetchers. Both SPEC and GAP workloads are separately sorted based on LLC MPKI, similar to Figure 2.

many speculative DRAM requests. Regarding the single-core evaluation, Hermes increases the number of DRAM transactions by 5.2%, 6.6%, and 6.4% over the baseline system that does not use any off-chip predictor for the SPEC, GAP, and all workloads combined, respectively. Figure 3, which presents the impact of Hermes on DRAM transactions in a multi-core context, shows that Hermes significantly increases DRAM transactions. Specifically, Hermes increases the average number of DRAM transactions by 2.2%, 9.6%, and 6.0% over the multi-core baseline for the SPEC mixes, GAP mixes, and all mixes, respectively. Notably, the increase in DRAM transactions for GAP workloads is higher than the increase for the SPEC workloads; this happens because the GAP suite is made of graph-processing applications that have larger data working sets than the general-purpose SPEC CPU workloads.

Figure 4 shows that 42.2% of the total off-chip predictions are inaccurate since the corresponding blocks reside in the cache hierarchy (L1D, L2C, or LLC). Notably, a large fraction of the load requests corresponding to an inaccurate off-chip prediction are served by the L1D cache. Specifically, 17.7% of the total off-chip predictions are useless since their corresponding block resides in the L1D. In other words, delaying Hermes to issue an off-chip prediction after the L1D lookup completion would significantly reduce DRAM transactions. However, constantly delaying the off-chip predictions of Hermes until the L1D lookup is completed would result in suboptimal performance gains since more than 50% (57.8% on average in Figure 4) of the Hermes off-chip predictions are accurate. In these cases, issuing the DRAM access before the L1D access is resolved provides latency benefits. Thus, a mechanism to decide whether or not an off-chip prediction of Hermes should be issued before or after the L1D access completion has the potential to significantly reduce the number of useless DRAM accesses triggered by Hermes.

Figure 5 presents the breakdown of the inaccurate L1D prefetches issued by IPCP depending on where in the memory hierarchy (L2C, LLC, DRAM) the corresponding prefetch request is served. To do so, we use the Prefetches Per Kilo Instruction (PPKI) metric. Overall, 18.2%, 3.8%, and 78% of the total inaccurate prefetch requests are served by L2C, LLC, and DRAM, respectively. We observe that the majority of the inaccurate prefetch requests are the ones that were served from DRAM. This behavior is more prevalent for GAP workloads since these workloads have more complex patterns than SPEC.

IV. Conclusion

These findings demonstrate that the state-of-the-art approach for off-chip prediction incurs a significant overhead in terms of additional DRAM transactions, and that there are opportunities to eliminate this overhead and boost performance by unifying off-chip prediction and prefetch filtering. This research presents the Two Level Perceptron, a novel approach that unifies these two techniques in a single method.

Acknowledgment

This work was originally published in the Proceedings of the 2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA). Alexandre Valentin Jamet acknowledges his AI4S fellowship within the “Generación D” initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by NextGenerationEU through PRTR.

References

[1] Basak, A., Li, S., Hu, X., Oh, S.M., Xie, X., Zhao, L., Jiang, X., Xie, Y.: Analysis and optimization of the memory hierarchy for graph processing workloads. In: 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA). pp. 373–386. IEEE (2019). https://doi.org/10.1109/HPCA.2019.00051

[2] Beamer, S., Asanović, K., Patterson, D.: Reducing pagerank communication via propagation blocking. In: 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS). pp. 820–831. IEEE (2017)

[3] Bera, R., Kanellopoulos, K., Balachandran, S., Novo, D., Olgun, A., Sadrosadat, M., Mutlu, O.: Hermes: Accelerating long-latency load requests via perceptron-based off-chip load prediction. In: 2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO). pp. 1–18 (Oct 2022). https://doi.org/10.1109/MICRO56248.2022.00015

[4] Bienia, C., Kumar, S., Singh, J.P., Li, K.: The parsec benchmark suite: Characterization and architectural implications. In: Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques. p. 72–81. PACT ’08, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1454115.1454128, https://doi.org/10.1145/1454115.1454128

[5] Jalili, M., Erez, M.: Reducing load latency with cache level prediction. In: 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA). pp. 648–661 (April 2022). https://doi.org/10.1109/HPCA53966.2022.00054

[6] Jimenez, D.: Piecewise linear branch prediction. In: 32nd International Symposium on Computer Architecture (ISCA’05). pp. 382–393 (2005). https://doi.org/10.1109/ISCA.2005.40

[7] Jiménez, D.A., Teran, E.: Multiperspective reuse prediction. In: 2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO). pp. 436–448. MICRO-50 ’17, IEEE, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3123939.3123942, https://doi.org/10.1145/3123939.3123942

Alexandre Valentin Jamet studied two years of Higher School Preparatory Classes with a Physics and Engineering Sciences major at LGT Baimbridge, Guadeloupe. In the following years, he pursued his MSc degree in parallel with an Engineer Diploma from TELECOM Nancy with a major in Embedded Computing. He concluded his studies in Nancy in 2018. Since October, he has been working at the Barcelona Supercomputing Center (BSC) where he pursued doctoral studies. He received his PhD in Computer Architecture from the Universitat Politecnica de Catalunya in October 2024. He is now working as a Recognized Researcher in the Barcelona Supercomputing Center as part of the Artificial Intelligence For Science (AI4S) Fellowship.