Characterizing the Impact of Graph-Processing Workloads on Modern CPU's Cache Hierarchy

Alexandre Valentin Jamet[1,2], Lluc Alvarez[1,2], and Marc Casas[1,2]

[1]Barcelona Supercomputing Center [2]Universitat Politècnica de Catalunya

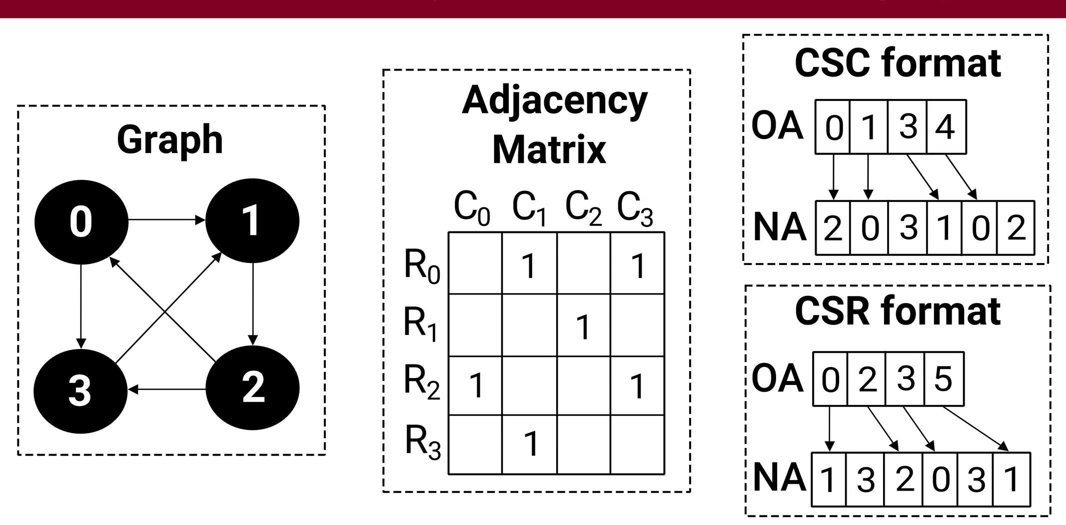
1. Graph-Processing Workloads

- Graph-processing is an emerging class of workloads that has applications in both industry and academia: social network analysis, web search engines, biomedical applications, etc.

Memory representation of graphs

- Graphs leverage data formats such as Compressed Sparse Row/Column (CSR/CSC) to manage a large amount of data.
- This format choice is due to provide a memory representation of the adjacency matrix of a given graph. Such a matrix is usually sparse.

Example: memory representation of a graph



2. Last-Level Cache Cache Replacement Policies

- Last-Level Cache (LLC) replacement policies represent a more complex challenge than their L1D and L2C counterparts as the LLC usually suffers from poor locality and highly filtered access patterns.

State-of-the-art LLC replacement policies

- State-of-the-art LLC replacement policies include:
  - **Reuse Distance Prediction** is based on the fundamental concept of Reuse Distance. These class of replacement policies primarily attempt to classify cache blocks into re-reference classes (*e.g.*, immediate reuse, distant reuse, etc.). The most representative policies of this class are SRRIP and DRRIP [5].
  - **Signature-based Hit Prediction (SHiP)** [2] builds on reuse distance prediction and learns the likelihood of a cache block experiencing hits. Decisions are taken to favor blocks that will provide hits.
  - **Multiperspective Reuse Prediction** [6] is a cache replacement policy that drives replacement, insertion, promotion, and bypass decisions in the LLC. It leverages a perceptron predictor to learn complex access patterns and extract knowledge from them.
  - **Optimal Replacement Approximation** are cache replacement policies that attempt to replicate the optimal replacement decision given by Belday's MIN. Hawkeye [3] and Glider [7] are representative.

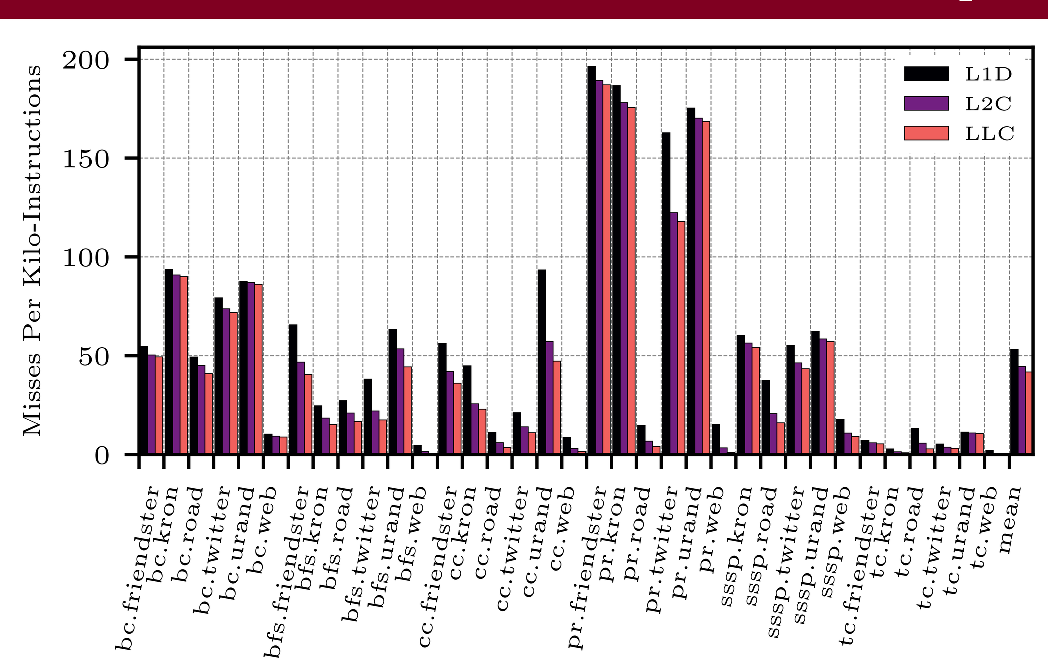
3. Methodology

- Champsim simulator modeling a Cascade Lake micro-architecture: L1D 32KB, L2C 1MB, LLC 1.375 MB.
- We model a 4GB DRAM with a data-rate of 3.2 GT/s.
- All workloads from GAP [1] Benchmark Suite, SPEC CPU 2006, SPEC 2017, XSBench.

4.A. Impact on Cache behavior

- Graph-processing workloads exhibit very high MPKI rates in all cache levels
- A very large portion (78.6%) of the accesses that trigger L1D misses also miss in the lower levels of the cache hierarchy and require a DRAM access.

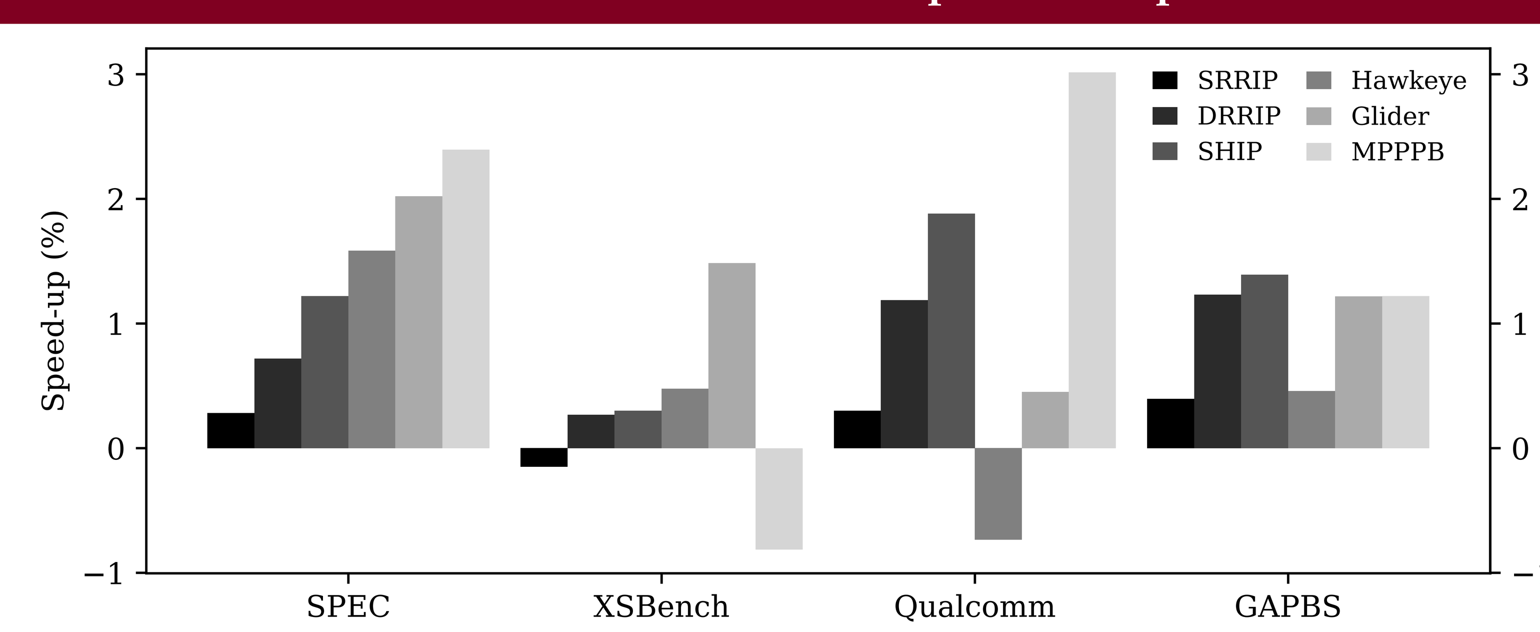
Misses-per-Kilo-Instructions in all Cache Levels of Graph Workloads



4.B. Impact of Cache Replacement Policies

- SPEC CPU 2006 and SPEC 2017 suites no more represent a challenge for computer architects, as they are well studied and there a plenty of ingenious mechanisms that cope with their behavior.
- The current state-of-the-art LLC replacement policies do not generalize well to new benchmarks.
- Emerging workloads, and specifically graph-processing workloads do represent a challenge for computer architects, as they stress the cache hierarchy more than traditional workloads. Nonetheless, they reveal the need to take into account their behavior in the design of forthcoming CPUs.

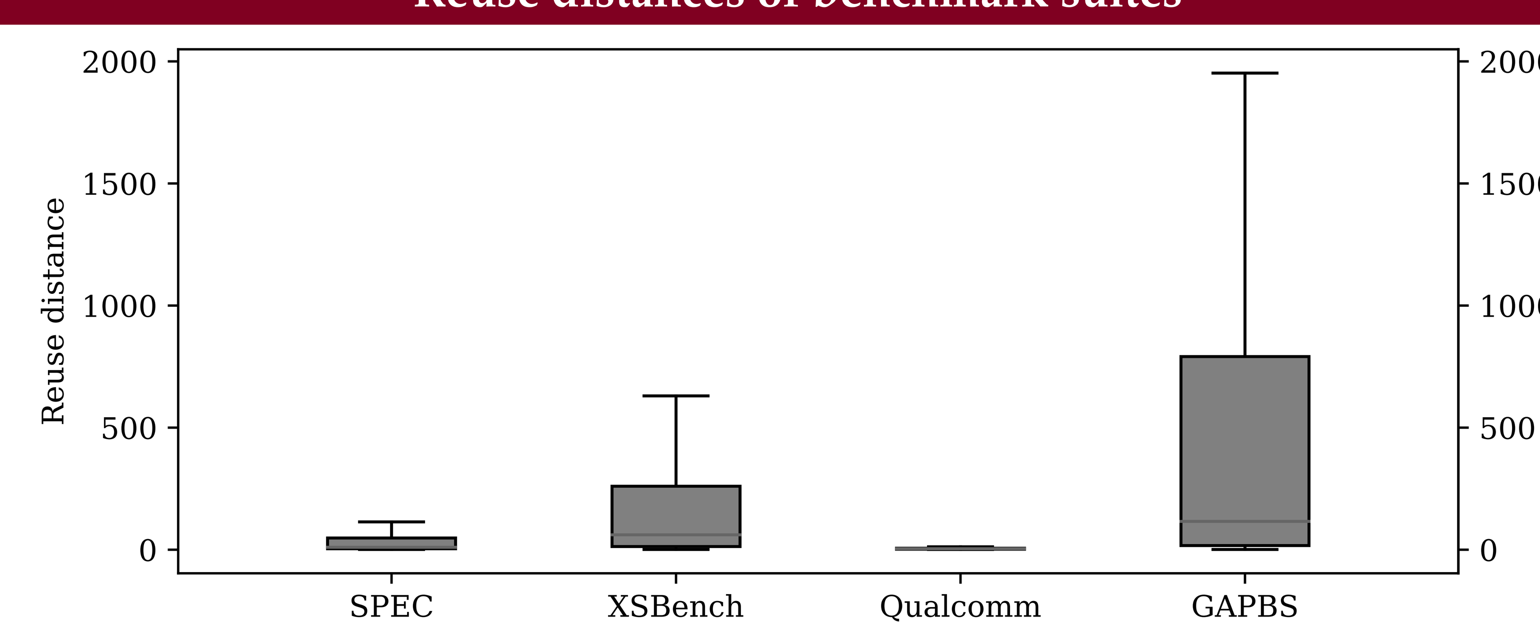
Performance benefits of LLC replacement policies



4.C. Impact on Reuse Distances

- We observe that GAP workloads experience much longer reuse distances than any other workloads [4].
- Such a behavior hinders the efficacy of LLC replacement policies.

Reuse distances of benchmark suites



6. Conclusions

- This work highlights the poor efficacy of state-of-the-art LLC replacement policies when applied to Graph workloads as compared to other workloads.
- We show the necessity of leveraging other micro-architectural approaches to accelerate these workloads.

5. References

References

- [1] S. Beamer, K. Asanović, and D. Patterson, "The GAP Benchmark Suite." arXiv, May 16, 2017. Accessed: Mar. 08, 2023. [Online]. Available: http://arxiv.org/abs/1508.03619
- [2] C.-J. Wu, A. Jaleel, W. Hasenplaugh, M. Martonosi, S. C. Steely, and J. Emer, "SHiP: signature-based hit predictor for high performance caching," in Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture - MICRO-44 '11, Porto Alegre, Brazil: ACM Press, 2011, p. 430. doi: 10.1145/2155620.2155671.
- [3] A. Jain and C. Lin, "Back to the Future: Leveraging Belady's Algorithm for Improved Cache Replacement," in Proceedings of the 43rd International Symposium on Computer Architecture, in ISCA '16. Seoul, Republic of Korea: IEEE, Jun. 2016, pp. 78–89. doi: https://doi.org/10.1109/ISCA.2016.17.
- [4] A. V. Jamet, L. Alvarez, D. A. Jiménez, and M. Casas, "Characterizing the impact of last-level cache replacement policies on big-data workloads," in 2020 IEEE International Symposium on Workload Characterization (IISWC), Oct. 2020, pp. 134–144. doi: 10.1109/IISWC50251.2020.00022.
- [5] A. Jaleel, K. B. Theobald, S. C. S. Jr, and J. Emer, "High Performance Cache Replacement Using Re-Reference Interval Prediction (RRIP)," in Proceedings of the 37th annual international symposium on Computer architecture, Saint-Malo, France: IEEE, Jun. 2010, pp. 60–71. doi: https://doi.org/10.1145/1815961.1815971.
- [6] D. A. Jiménez and E. Teran, "Multiperspective reuse prediction," in Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture - MICRO-50 '17, in MICRO '17. Cambridge, Massachusetts: IEEE, 2017, pp. 436–448. doi: 10.1145/3123939.3123942.
- [7] Z. Shi, X. Huang, A. Jain, and C. Lin, "Applying Deep Learning to the Cache Replacement Problem," in Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, Columbus OH USA: ACM, Oct. 2019, pp. 413–425. doi: 10.1145/3352460.3358319.

Acknowledgements

- Marc Casas is supported by Grant RYC-2017-23269 funded by MCIN/AEI/ 10.13039/501100011033 and by "ESF Investing in your future".

Biography

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 2018, he has been a Ph.D. candidate at the Computer Architecture departments of Barcelona Supercomputing Center (BSC) and Universitat Politècnica de Catalunya (UPC), Spain.