Thesis Background

Zhengyi Chen

October 11, 2023

The problem of cache replacement is general to computer systems of all scales and topologies: topologically massive systems, such as cellular stations[3] and CDNs[2, 1, 5], and data-path level implementations for processors[6, 4, 7] alike requires good solutions to maintain and maximize application performance to various levels of granularity. On the other hand, the set of feasible/performant solutions (i.e., cache replacement policies) to one system may or may not be inspiring to performance improvement on another system of different scale, objectives, tasks, constrained by a (mostly) different context of available inputs, metadata, etc.

We propose a framework for dynamic cache-replacement-strategy selection that balances computation cost, optimality, and working-set estimation for each strategy while incurring minimal performance penalties for a shared-kernel cooperative Distributed Shared Memory system. (We identify ...)

1 Existing Cache Replacement Strategies

1.1 LRU-derived Algorithms

1.2 FIFO-derived Algorithms

1.3 Cache Replacement in Processors

1.4 Machine Learning and Heuristics

2 The Cache Replacement Problem

3 Page Replacement in (SMP or?) Linux

References

[1] Sem Borst, Varun Gupta, and Anwar Walid. "Distributed caching algorithms for content distribution networks". In: *2010 Proceedings IEEE INFOCOM*. IEEE. 2010, pp. 1–9.

[2] Ohad Eytan et al. “It’s Time to Revisit LRU vs. FIFO”. In: *12th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 20)*. USENIX Association, July 2020. URL: https://www.usenix.org/conference/hotstorage20/presentation/eytan.

[3] Jingxiong Gu et al. “Distributed cache replacement for caching-enable base stations in cellular networks”. In: *2014 IEEE International Conference on Communications (ICC)*. 2014, pp. 2648–2653. DOI: 10.1109/ICC.2014.6883723.

[4] Aamer Jaleel et al. “High Performance Cache Replacement Using Re-Reference Interval Prediction (RRIP)”. In: *SIGARCH Comput. Archit. News* 38.3 (June 2010), pp. 60–71. ISSN: 0163-5964. DOI: 10.1145/1816038.1815971. URL: https://doi.org/10.1145/1816038.1815971.

[5] Madhukar R. Korupolu and Michael Dahlin. “Coordinated placement and replacement for large-scale distributed caches”. In: *IEEE Transactions on Knowledge and Data Engineering* 14.6 (2002), pp. 1317–1329.

[6] Moinuddin K. Qureshi et al. “Adaptive Insertion Policies for High Performance Caching”. In: *SIGARCH Comput. Archit. News* 35.2 (June 2007), pp. 381–391. ISSN: 0163-5964. DOI: 10.1145/1273440.1250709. URL: https://doi.org/10.1145/1273440.1250709.

[7] Subhash Sethumurugan, Jieming Yin, and John Sartori. “Designing a Cost-Effective Cache Replacement Policy using Machine Learning”. In: *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. 2021, pp. 291–303. DOI: 10.1109/HPCA51647.2021.00033.