Measuring Data Access Latency in Large CPU Caches

Shaotong (Simon) Sun
B.S. CS, B.S. BIS, B.A. Phil, class of 2024

Abstract

This poster describes a new, multi-locality benchmark program for testing memory access latency and using it to study recent AMD machines equipped with 3D vertical cache (V-Cache) that can be over 1 GiB in total size on a single node. The latency study shows that these large caches differ from traditional LLCs in two aspects: the V-Cache is partitioned rather than shared, and the cache replacement policy is more similar to random than it is to LRU.

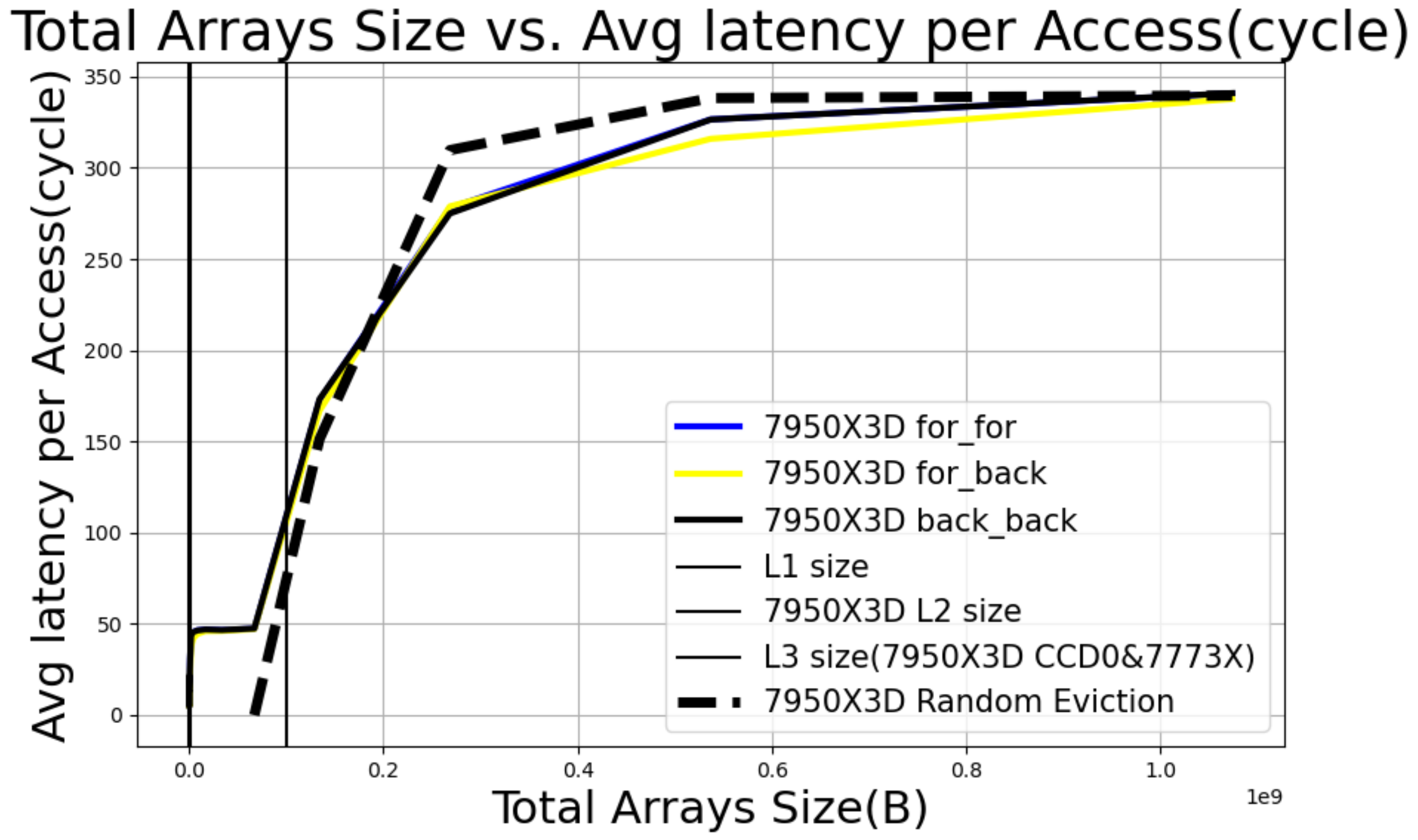
Our contributions are:
- Specifications of recent AMD machines with up to 768 MiB V-Cache on a single machine
- A benchmark program that tests access latency on data traversals that have different data reuse patterns:
  1. Cyclic: data is repeatedly traversed in the same order
  2. Sawtooth: the order is reversed each time the data is traversed
- results from V-Cache-based machines and a comparison with other systems

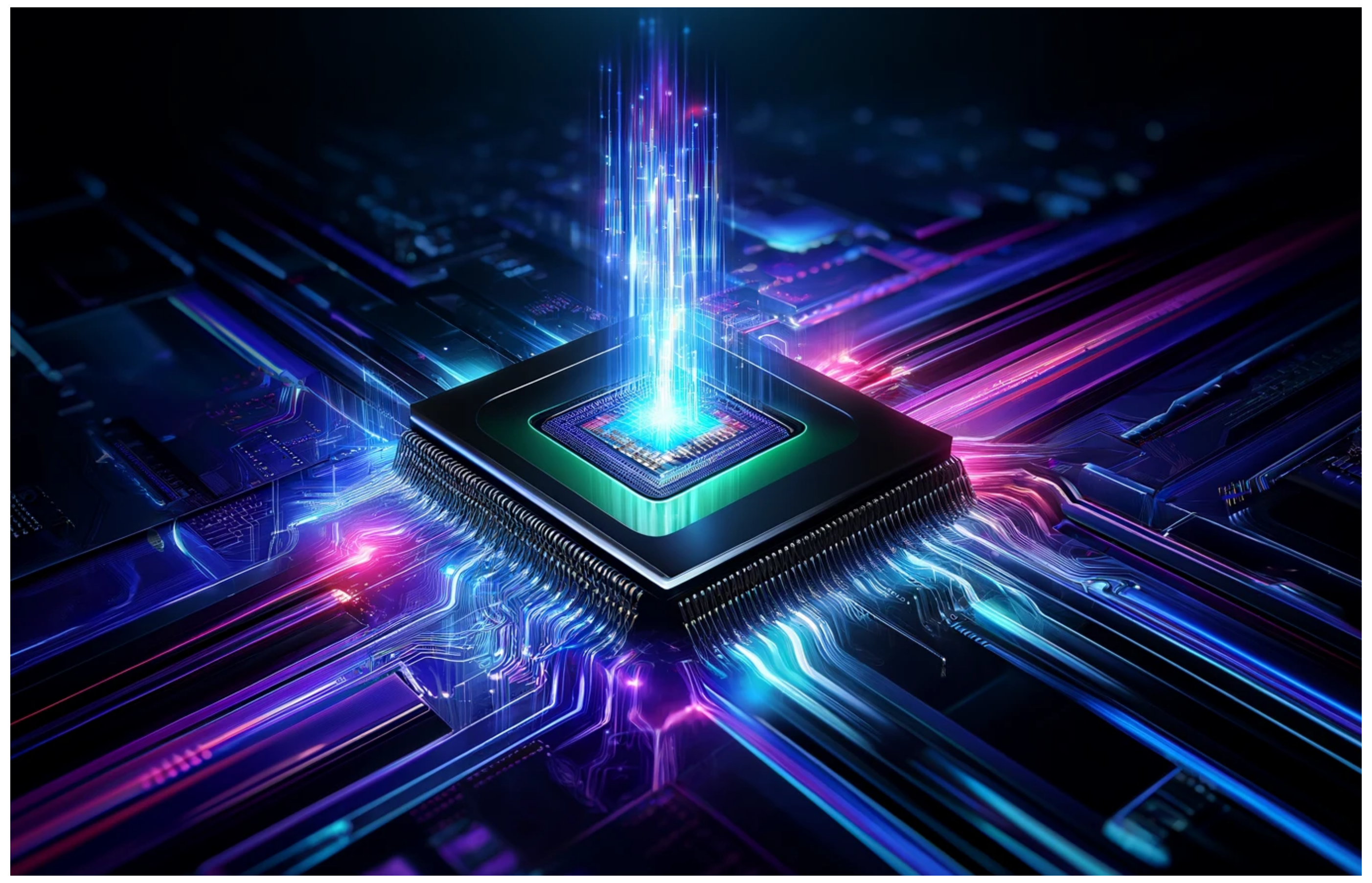
The results show that V-Cache has two unique design features:
1. The replacement policy is clearly different from the one used in the L2 cache on the same machine and the policy used on an Intel machine.
2. Only local V-Cache is used by a core, and having more V-Cache modules does not help if only a single core is used, i.e., partitioned LLC rather than shared LLC.

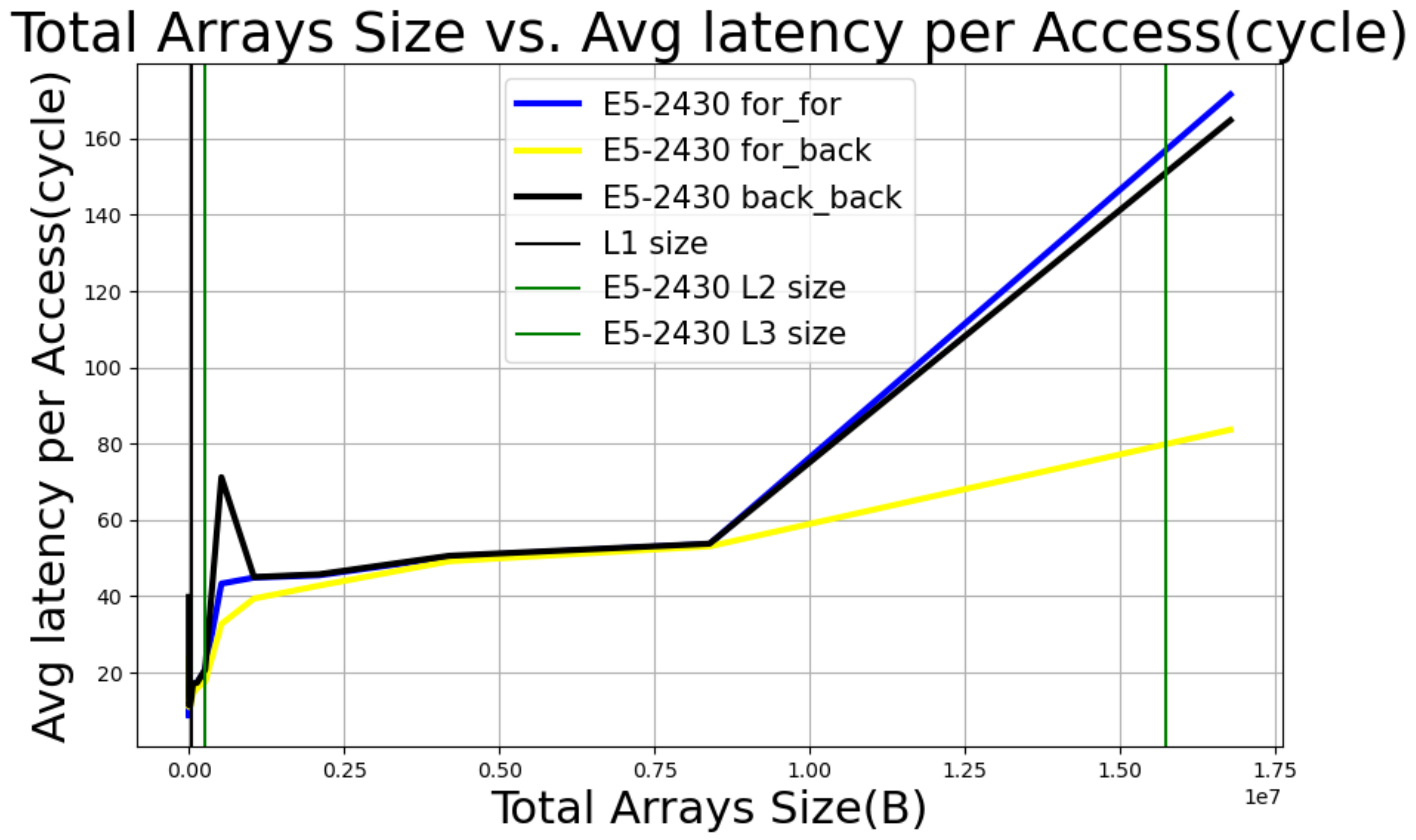
Dual Access Order Stream-in-Cache Benchmark (SiC)

On modern computers, the access latency is made obscure by non-data access costs, including instructions and auxiliary data, latency hiding through out-of-order execution and prefetching, contention on data bandwidth, and cache coherence.

In SiC, we enable techniques to overcome latency-hiding actions from the hardware.
- Triangular Access
- Dependent Loads
- Sawtooth and Cyclic Data Traversals
- Core Binding
- Cache Line Access
- Huge Pages







Machine Specifications

AMD R9 7950X3D[2]:
- Launched in February 2023
- Up to 5.7 GHz
- Zen4 Microarchitectures
- 2 CCDs, each contains:
  - 8 cores 16 threads
  - L1i: 256 KiB(8 × 32 KiB) L1d: 256 KiB(8 × 32 KiB) 8-way set associative (ECC) (write-back)
  - L2: 1 MiB (8 × 1024 KiB) 8-way set associative (ECC) (write-back)
  - L3: 128 MiB total size, 32 MiB + 64 MiB 3D V-Cache only on CCD0 16-way set associative (ECC) (L2 Victim Cache) (write-back)

Intel E5-2430[1]:
- Launched in 2011
- Up to 2.7 GHz
- Sandy Bridge-EN Microarchitectures
- 2 NUMA nodes, each contains:
  - 6 cores 12 threads
  - L1i: 192 KiB (6 × 32 KiB) L1d: 192 KiB (6 × 32 KiB)
  - L2: 1.5 MiB (6 × 256 KiB)
  - L3: 15 MiB

Acknowledgment:

I would like to express my sincere gratitude to Prof. Chen Ding, Mr. Yifan Zhu, and Ms. Mingzhe Du for their invaluable guidance and support throughout this project. The extensive knowledge and expertise they possessed have been instrumental in helping me navigate through the various challenges I faced during this research.

Reference:

1. https://www.intel.com/content/www/us/en/products/sku/64616/intel-xeon-processor-e52430-15m-cache-2-20-ghz-7-20-gts-intel-qpi/specifications.html
2. https://www.amd.com/en/product/12741