Cache & Memory Profiling

October 01, 2025

Joyal Mathew

System Information

CPU	Operating System
AMD Ryzen 5 5625U	Ubuntu 22.04.5 LTS

Cache Sizes

L1 192 KiB
L2 3 MiB
L3 16 MiB

Experiments

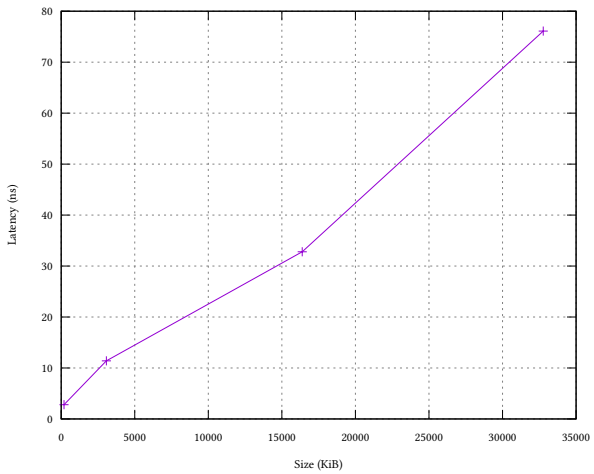
Intel Memory Latency Checker (MLC) was used to measure memory latency and bandwidth. Linux perf was used to count cache/TLB misses.

Zero-Queue Baseline

MLC's `idle_latency` command was used to measure zero-queue latency for different buffer sizes.

CPU frequency: 4.34 GHz.

Memory Regime	Buffer Size (KiB)	Latency (ns)	Cycles
L1 Cache	192	2.8	6.4
L2 Cache	3072	11.4	26.1
L3 Cache	16384	32.8	75.3
Main Memory	32768	76.1	174.7



The effects of cache are clearly visible. As we cross into different memory regimes, the idle latency increases greatly.

Bandwidth/Latency Sweep

Bandwidth and latency were measure for different access patterns and strides using MLC's `bandwidth_matrix` and `latency_matrix` commands.

Granularity/ Pattern	64 B	256 B	1024 B
Sequential	22798.7	20957.1	9880.9
Random	31310.8	20556.4	11311.8

Table 2: Bandwidth sweep (MB/s)

Granularity/ Pattern	64 B	256 B	1024 B
Sequential	91.6	107.4	100.7
Random	91.9	107.6	100.4

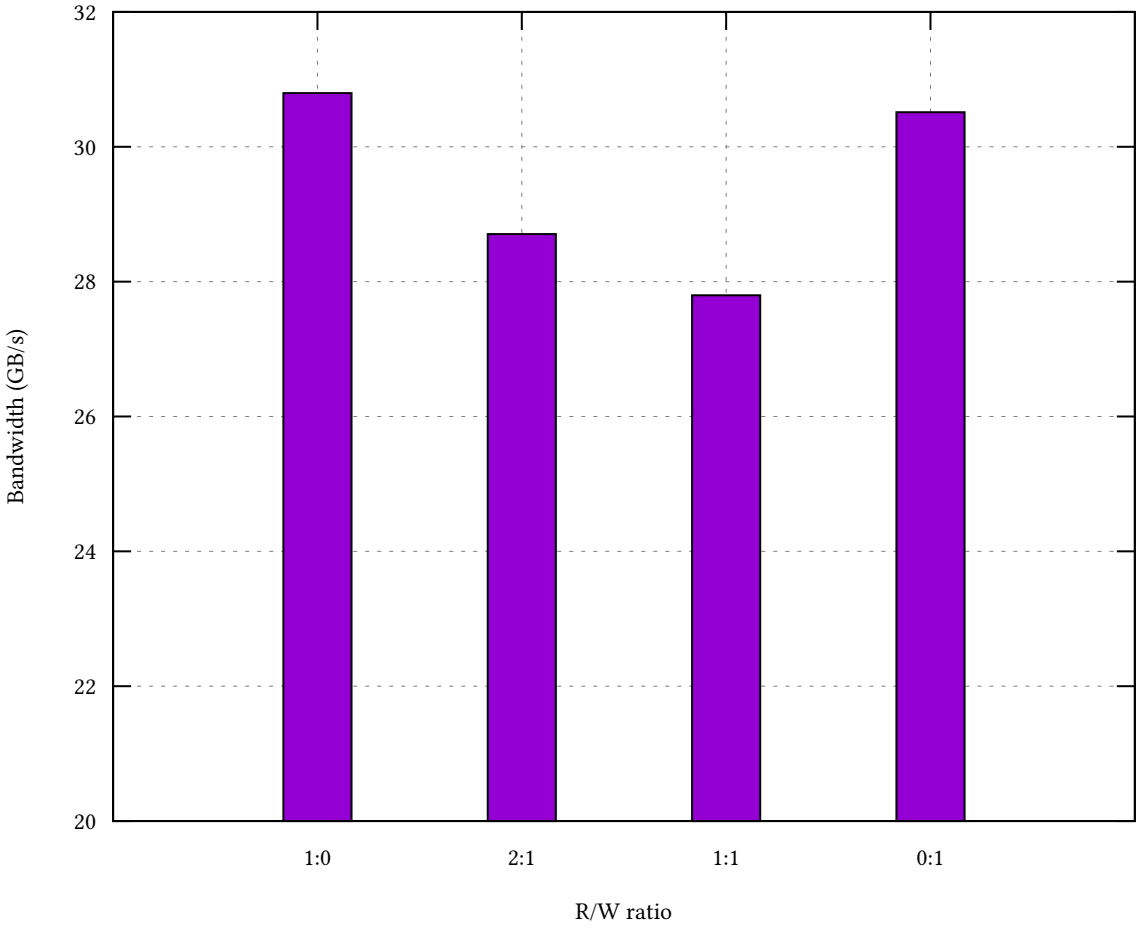
Table 3: Latency (ns)

Bandwidth is much lower for random access due to reduced prefetcher effectiveness. When memory access is random, the CPU cannot predict and prefetch the next likley memory candidates. Additionally, when the stride is higher, less of the prefetched memory is utilized due to jumps. This further decreases bandwidth.

The effects are not as high for latency because it measures first-time access. That doesn't benefit as much from the memory prefetcher.

Read/Write Mix Sweep

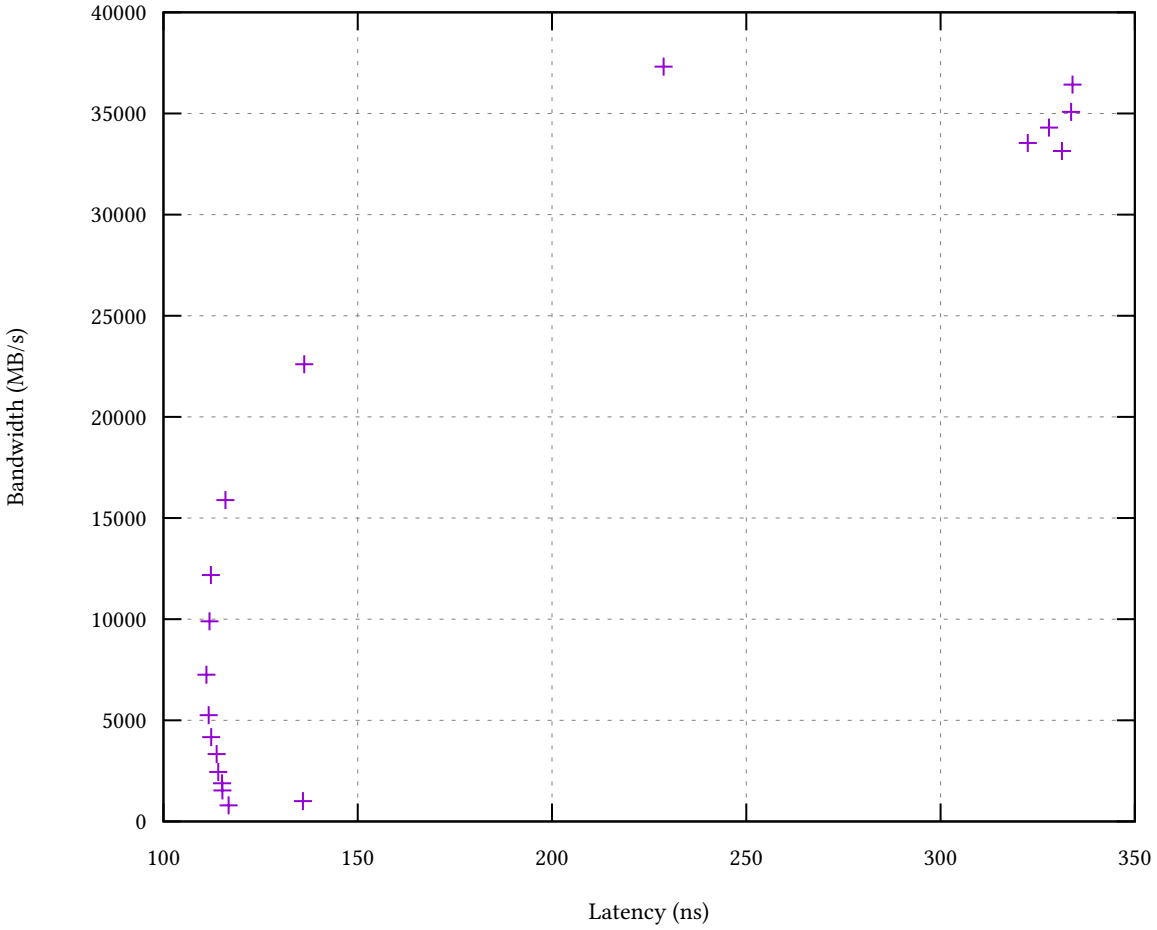
Bandwidth for a few R/W ratios were measured using the `bandwidth_matrix` command with the `-W/-R` options.



When operations are homogeneous, bandwidth is much higher. Interspersed reads and writes degrate performance. This is due to the memory controller having to switch between issuing reads and writes.

Intensity Sweep

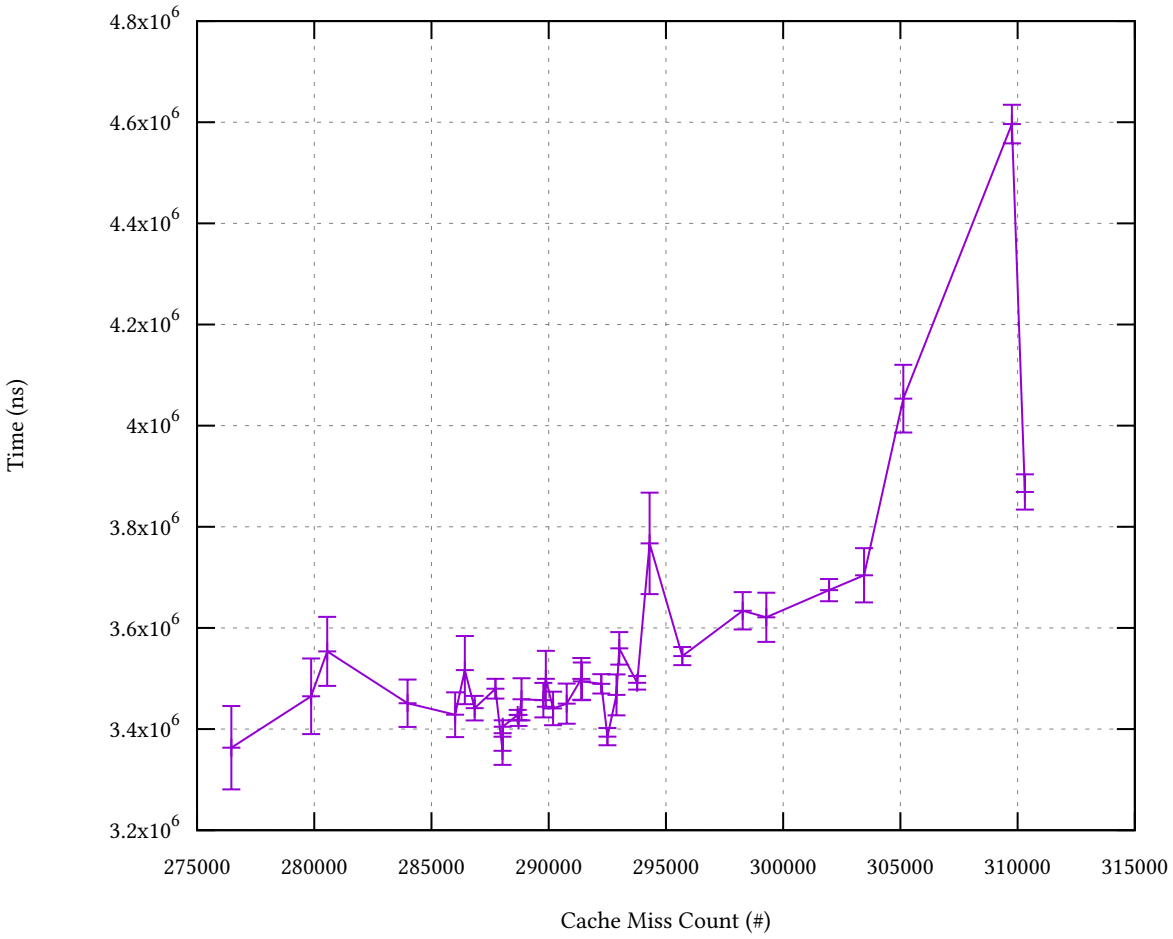
MLC's `loaded_latency` command was used to sweep intensity. We plot latency vs bandwidth here.



At lower latency levels, bandwidth increases sharply with latency. This falls off however in the "knee" between 150 and 200 ns of latency. From there, increases in latency provide diminishing returns in bandwidth. The peak bandwidth appears to be around 35 GB/s.

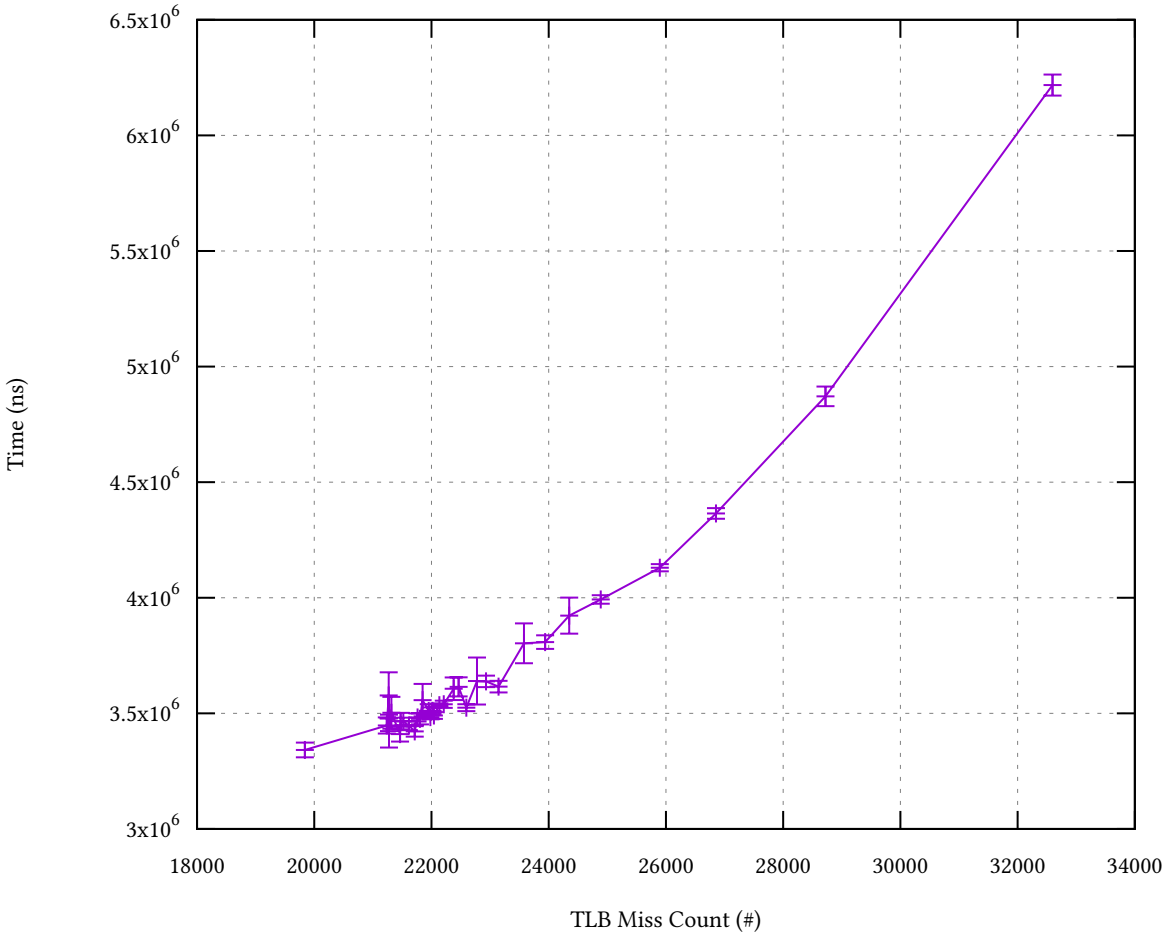
Cache-Miss Impact

A light multiplication kernel was used to measure the effects of cache misses. Random accesses (within page size) were interspersed to introduce cache misses.

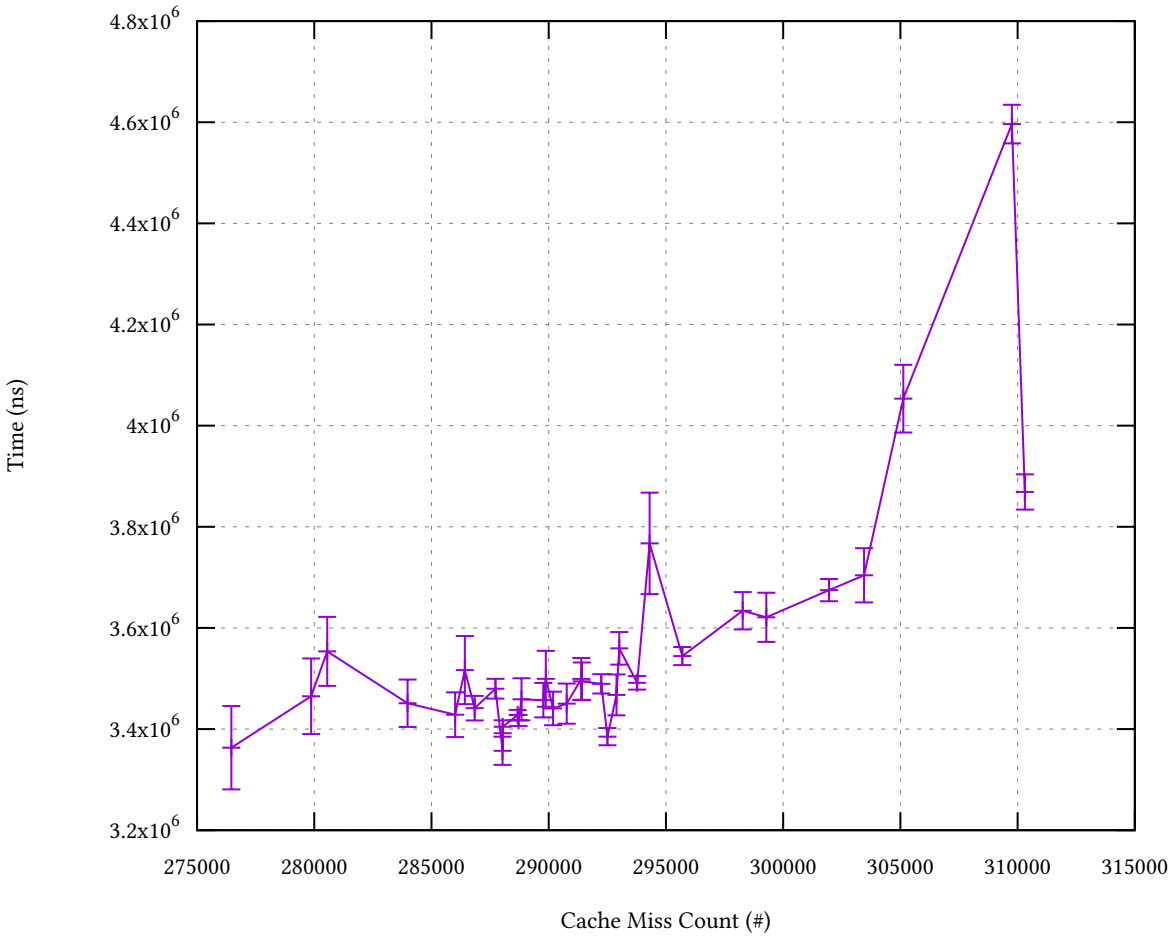


Execution time of the kernel appears to be directly related to the number of cache misses. This makes sense because when there are more cache misses, more memory accesses need to query higher up memory levels. These higher memory levels take longer to access and therefore increase the average memory access time (AMAT). This increases the overall run-time of the program.

TLB-Miss Impact



The same multiplication kernel was used to measure the effects of TLB misses. Random accesses (spanning many pages) were interspersed to introduce TLB misses.



Execution time increases directly with TLB miss count. This makes sense because for each TLB miss the CPU has to walk the page table to translate the required address. This is a very expensive operation.