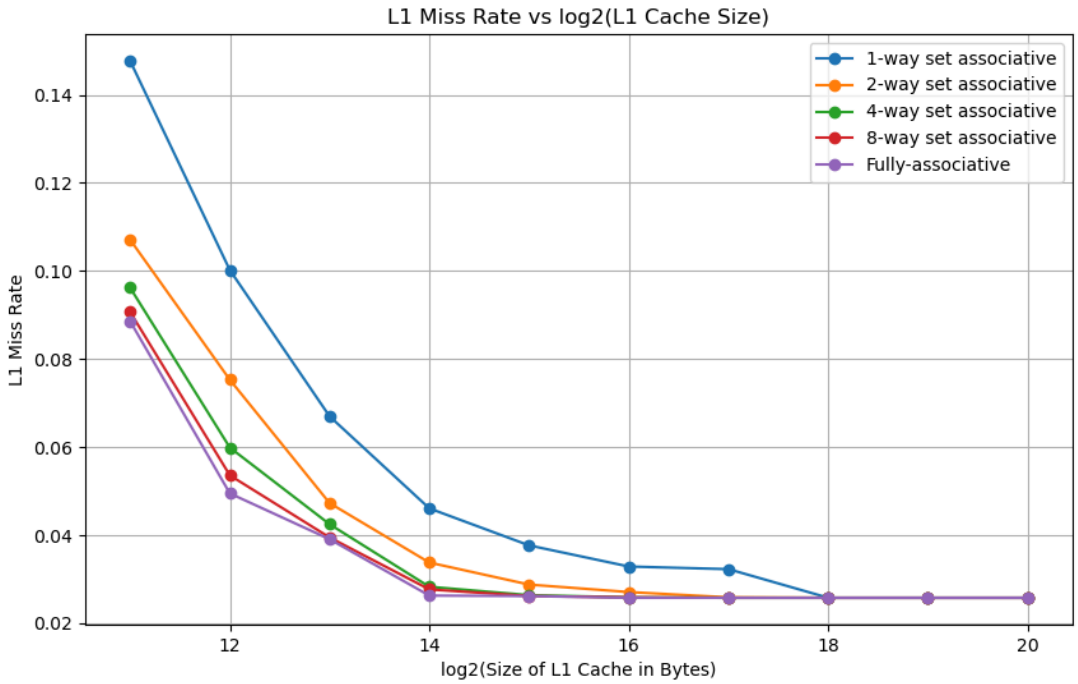
L1 Cache Investigation: SIZE and ASSOC

PLOT #1



DISCUSSION ABOUT THE TRENDS:

Direct-mapped cache exhibits the highest miss rate across all cache sizes, showing a steep decline as cache size increases, but it maintains a higher L1 miss rate even at larger cache sizes compared to the other configurations.

Associative caches (2-way, 4-way, 8-way, and fully associative) all exhibit lower miss rates than the direct-mapped configuration. The miss rate decreases with increasing associativity, as expected, since higher associativity reduces conflict misses.

Fully associative cache consistently has the lowest L1 miss rate.

Size effect: For all configurations, as the cache size increases (on the x-axis), the L1 miss rate decreases, showing diminishing returns as the cache size reaches about 64KB.

ESTIMATING THE THE COMPULSORY MISS RATE

Compulsory misses are not affected by cache associativity and occur when a block is accessed for the first time and has not yet been loaded into the cache. As cache size grows, these misses typically decrease. From the graph, we can see the miss rate reaching a minimum as the cache size increases, particularly for fully associative caches. After 128KB, the miss rate levels off across all associativity levels. Based on this, the compulsory miss rate seems to be

approximately 0.025 (or 2.5%), as this is the lowest observed miss rate across all cache sizes and associativity levels.

Estimating Conflict Miss Rates for Each Associativity

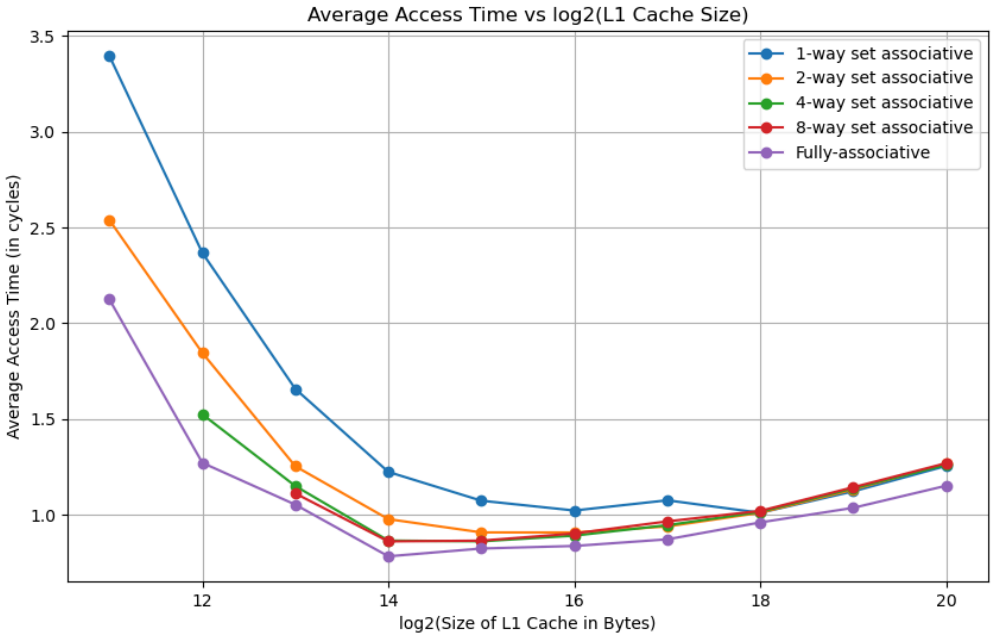
Conflict misses occur due to limited associativity and can be estimated by comparing the miss rates of direct-mapped and set-associative caches with fully associative caches, which do not experience conflict misses. The formula for conflict misses in a set-associative cache (SA) is:

conflict misses_{SA} = total misses_text_{SA} - total misses_{FA}

At around 16KB L1 cache size, the fully associative cache only has compulsory misses and no capacity misses, meaning that for all associativity levels, capacity misses will also be 0. Therefore, the difference in miss rates at 16KB, 32KB, 64KB, 128KB, etc., directly reflects the conflict miss rate.

- For the 2-way associative cache (green line), the conflict miss rate at 16KB is roughly 1%, decreasing to 0 by 128KB.
- For the 4-way associative cache (blue line), the conflict miss rate at 16KB is around 0.2%, gradually reducing to 0 by 128KB.
- For the 8-way associative cache (cyan line), the conflict miss rate at 16KB is approximately 0.1%, also diminishing to 0 by 128KB.

PLOT #2

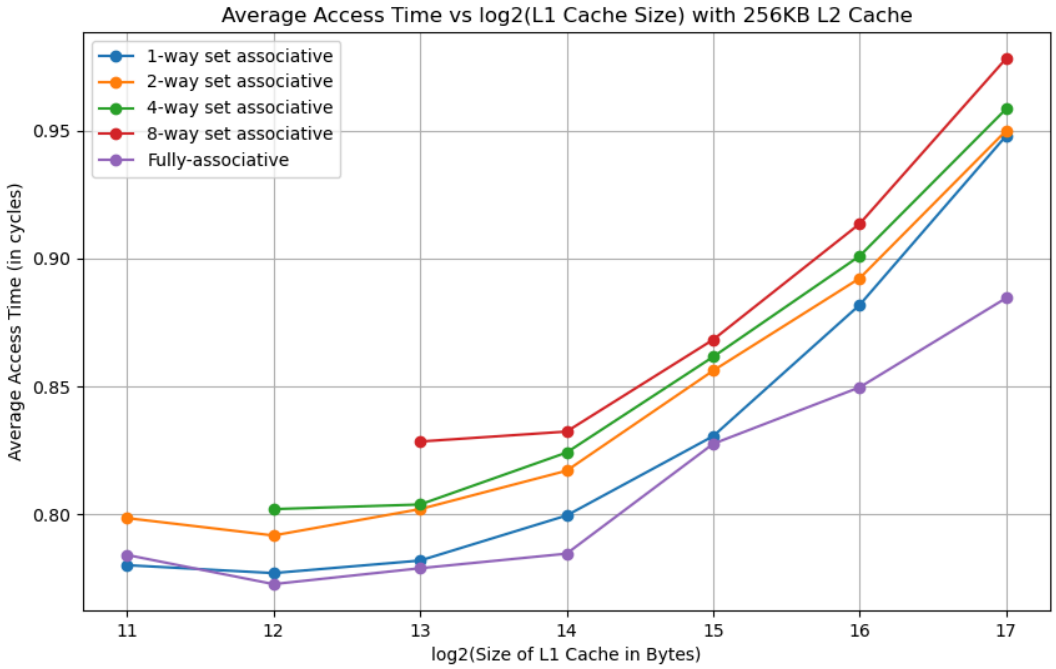


1. For a memory hierarchy with only an L1 cache and BLOCKSIZE = 32, which configuration yields the best AAT?
Average Access Time (AAT) is determined by considering both the cache miss rate and the access time for different configurations, factoring in the effects of associativity and cache size on performance.

From the graph, it is evident that the fully associative cache (magenta line) and the 8-way set associative cache (cyan line) provide the best AAT at smaller cache sizes (16KB and below). At 64KB, the AAT for all configurations (2-way, 4-way, and 8-way associative caches) converges closely, with the fully associative cache having a slight advantage. For larger cache sizes, the 2-way, 4-way, and 8-way associative caches perform similarly, with the 4-way cache showing a slight improvement among the three, while the fully associative cache continues to perform the best overall.

For a block size of 32 bytes and the optimal AAT, the fully associative cache offers the best performance across all cache sizes, with the best AAT achieved by the 16KB fully associative memory hierarchy at 0.785 ns, followed by the 16KB 8-way associative cache.

PLOT #3



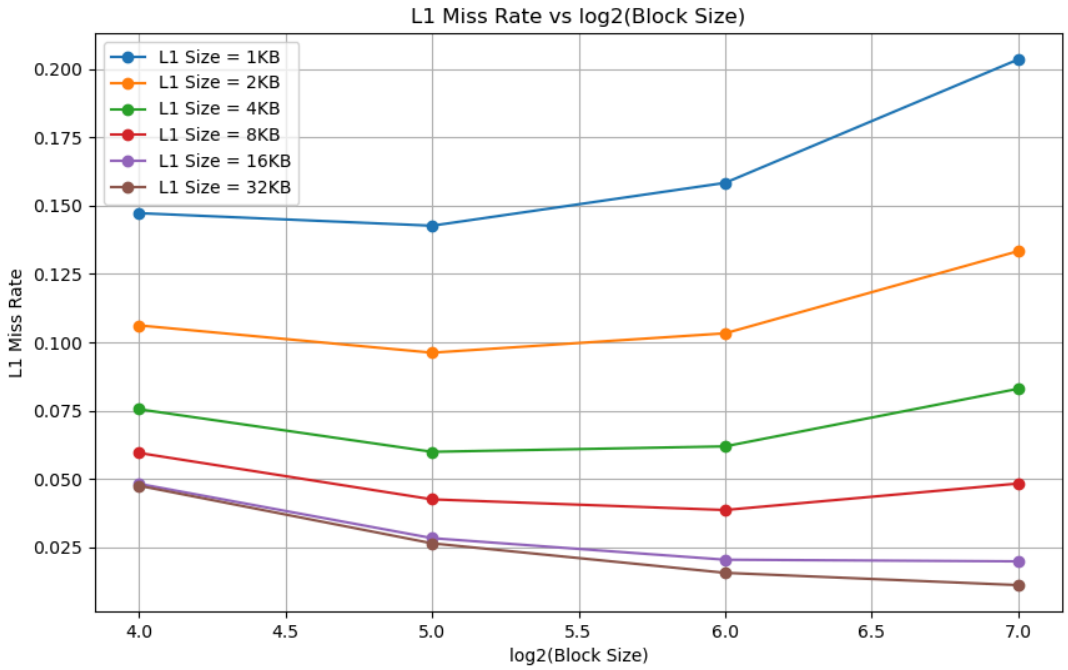
Optimal Memory Hierarchy
The 4KB fully associative cache appears to have the lowest AAT of about 0.77 ns , making it the best configuration when an L2 cache is added. Fully associative caches have the lowest AAT across almost all L1 sizes and have the best performance in terms of AAT. This is then

followed by directly mapped caches which have the next best performance in terms of AAT. This directly relates to a 1.91 % improvement in the AAT when a L2 is augmented to the L1 cache.

EDP and Area comparison with L1 cache

The Energy-Delay Product (EDP) for a system with only a 16KB fully associative (FA) cache is approximately 460,720,952.4227 nJ, while using a 256KB 8-way set associative L2 cache with a 4KB fully associative L1 cache reduces the EDP to around 147,816,815.9992 nJ, representing a 67.91% reduction in EDP. However, the area for the first configuration is 0.0634 units, and for the second configuration, it increases to 1.1784 units, which is a rise of 1.115 units.

PLOT #4

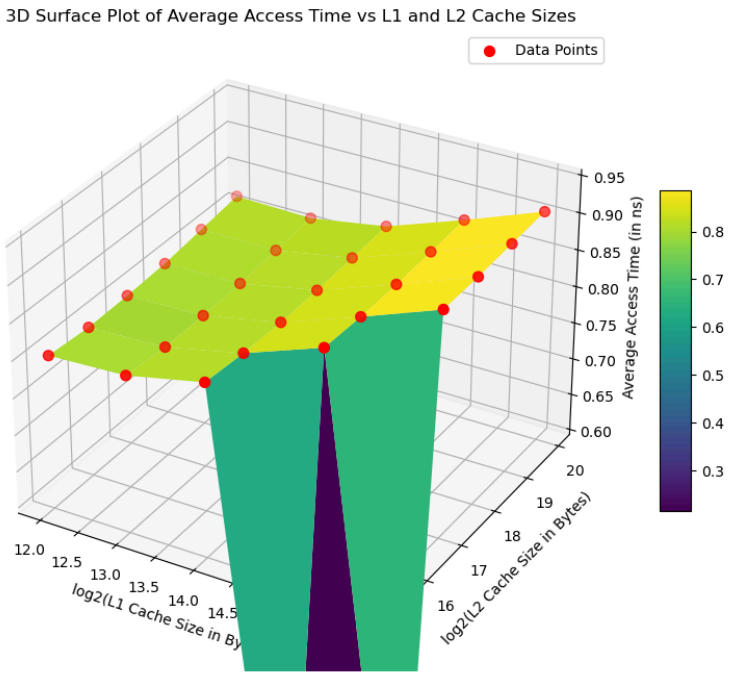


As the block size increases from 16 to 64 bytes, the miss rates generally decline, indicating that spatial locality is being utilized effectively.

- However, for most cache sizes, increasing the block size beyond 64 bytes results in rising miss rates.
- This suggests cache pollution, where fetching larger blocks loads the cache with data that isn't immediately useful, displacing other blocks.

This pattern shows that smaller caches are more vulnerable to cache pollution, while larger caches can benefit from bigger block sizes before experiencing pollution. For a 4-way set associative 8KB L1 cache, the optimal block size is 64 bytes, as it yields the lowest miss rate compared to other block sizes.

PLOT #5



The minimum Average Access time (AAT) observed in the data is 0.7972 nanoseconds.
This is achieved with the following memory hierarchy configuration:
The L1 cache size is 2^{12} = 4096 bytes.
• The L2 cache size is 2^{16} = 65536 bytes.
Thus, the optimal configuration that minimizes AAT is an L1 cache of **4096** bytes and an L2 cache of **65536** bytes.

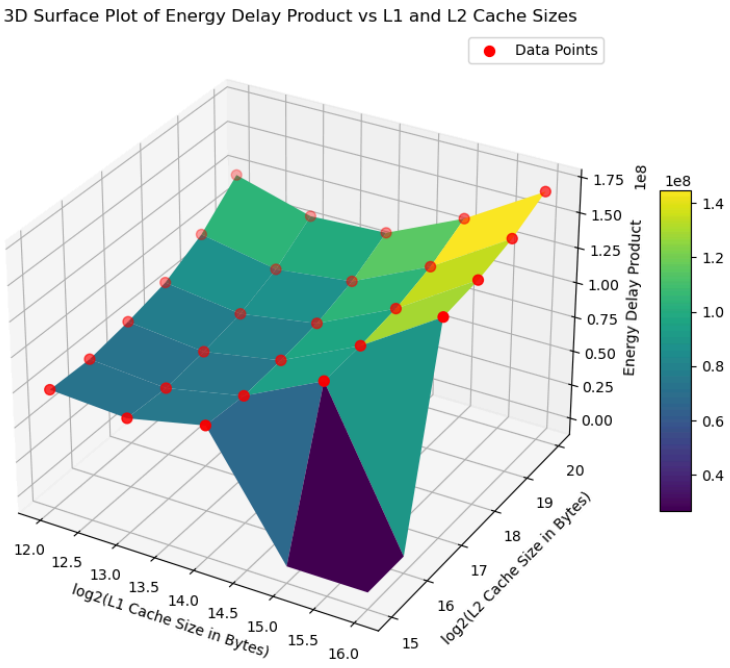
Can you propose a memory hierarchy configuration that has a smaller total area, but provides an AAT within 5% of the best AAT?

In the previous scenario, the total area was calculated to be 0.3858. In the new configuration, with an L1 cache of 4096 bytes and an L2 cache of 32768 bytes, the total area is 0.2853, and the AAT is 0.80 ns, staying within the 5% threshold.
To calculate the percentage decrease in the total area, we use the formula:

Percentage Decrease = (0.3858 - 0.2853)/(0.3858) = 26.04%

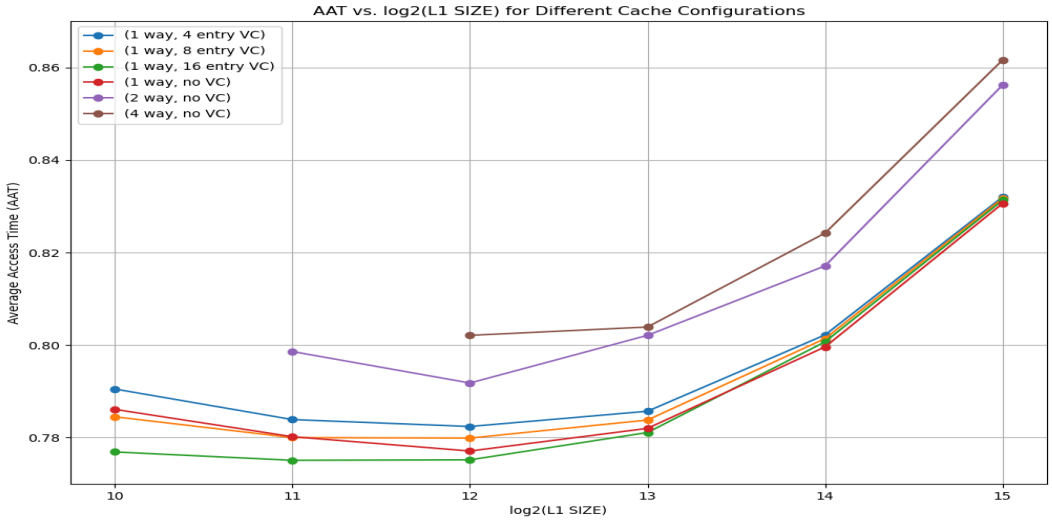
Thus, the new configuration results in a 26.04% decrease in the total area compared to the previous configuration while being within the 5% threshold of AAT.

PLOT #6



The minimum Energy-Delay Product (EDP) observed is 68614172.0543 for which the L1 has a size of 2^{13} = 8 kB and L2 has a size of 2^{16} = 64 kB.Thus the optimal configuration is an 8 kB 4-way L1 Cache and a 64 kb 8-way L2 cache, both with a 32 byte blocksize

PLOT #7



Comparison of direct mapped L1 with VC to 2-way L1

From the plot, we can compare the performance of a 2-way L1 cache without a victim cache to the 1-way L1 caches that utilize a victim cache. It is evident that even a victim cache with just 4 entries significantly reduces the Average Access Time (AAT), leading to much better performance compared to the 2-way L1 without a victim cache.

- For a 1-way L1 with a 4-entry victim cache, the average AAT is 2.7% lower.
- For a 1-way L1 with an 8-entry victim cache, the average AAT is reduced by 2.87%.
- For a 1-way L1 with a 16-entry victim cache, the average AAT is 3.14% lower.

Optimal configuration for least AAT

It is evident that the least AAT is achieved by a 1-way L1 Cache of 2 kB, with a 16-entry VC

Alternate memory hierarchy

The specified constraints are as follows:
- Average Access Time (AAT): The minimum AAT is 0.7751 ns, indicating that the alternate configuration can have an AAT of up to 0.8138 ns.
- Area: The optimal configuration has an area of 1.1785 sq. mm.

From the graph, it is evident that a 1-way L1 cache with a size of 1 kB and a 4-entry victim cache (VC) has an AAT of 0.7905 ns. The area of this configuration is 1.1717 sq. mm, which is below the specified area constraint.

Therefore, the alternate memory configuration consists of a 1-way L1 cache of 1 kB with a 4-entry VC. This configuration has an area that is 0.577% smaller and an AAT that is 1.89% greater than the optimal configuration.