

CS6600 - Project 1: Cache and Memory Hierarchy Design

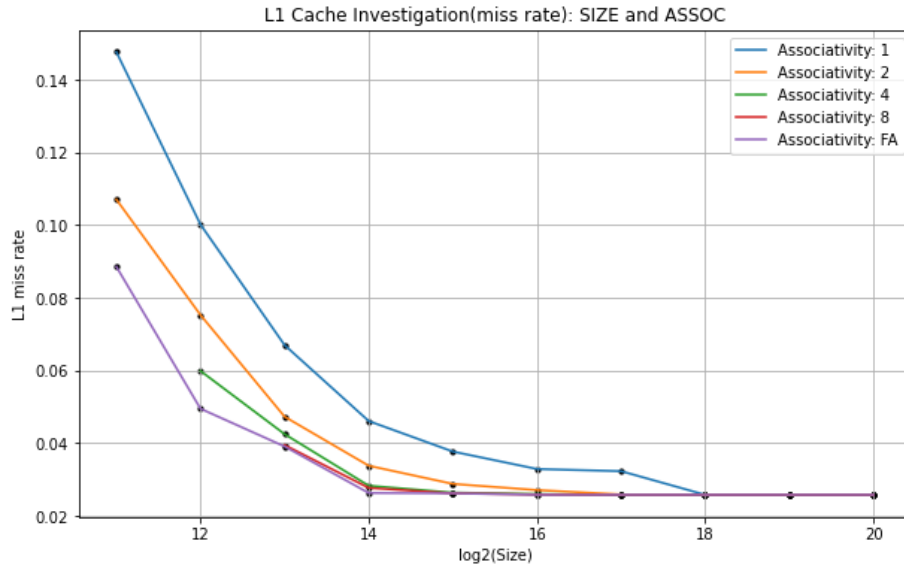
Gokulakrishnan R (CS21B029)

September 2024

L1 Cache Investigation: SIZE and ASSOC

Plot #1

We observe the L1 miss rate when size changes for various associativity values. For this experiment, we use BLOCKSIZE = 32, and assume that there are no Victim and L2 caches.



Discuss trends in the graph.

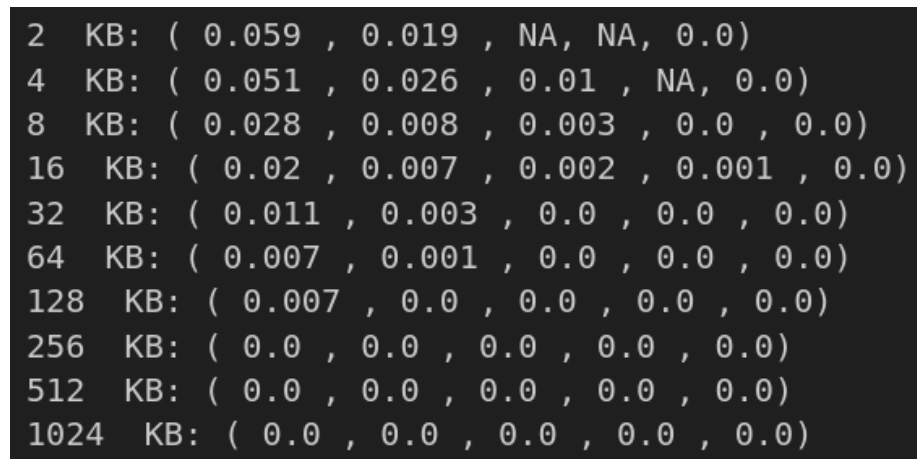
As we expect, we can see that the miss rate decreases as the cache size increases (for fixed associativity). It also becomes constant (0.0258) when the size crosses a threshold (different for each associativity). We can see that as associativity increases for a fixed L1 size, the miss rate decreases or remains the same when the cache size is very high.

Estimate the compulsory miss rate from the graph.

For all the associativities, we can see that the miss rate becomes constant when the size of the cache crosses a threshold. This means that the threshold is the minimum size of the cache (for that associativity) for which there are only compulsory misses. Thus, the compulsory miss rate of the given trace is 0.0258.

For each associativity, estimate the conflict miss rate from the graph.

For FA caches, the conflict miss rate is 0. We also know that for all other associativities, the conflict miss rate is the difference between the miss rate and the miss rate of a FA cache of the same size.

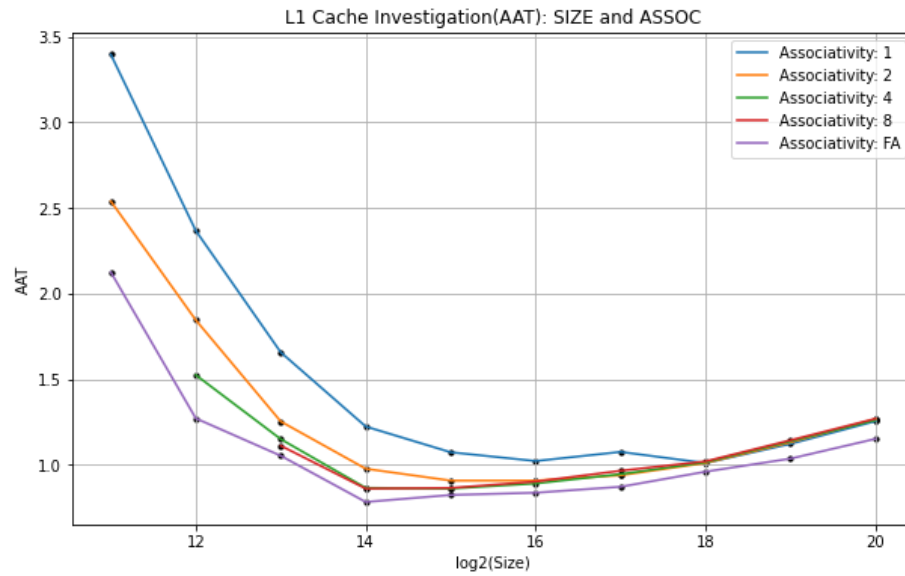


2	KB:	(0.059	,	0.019	,	NA	,	NA	,	0.0)
4	KB:	(0.051	,	0.026	,	0.01	,	NA	,	0.0)
8	KB:	(0.028	,	0.008	,	0.003	,	0.0	,	0.0)
16	KB:	(0.02	,	0.007	,	0.002	,	0.001	,	0.0)
32	KB:	(0.011	,	0.003	,	0.0	,	0.0	,	0.0)
64	KB:	(0.007	,	0.001	,	0.0	,	0.0	,	0.0)
128	KB:	(0.007	,	0.0	,	0.0	,	0.0	,	0.0)
256	KB:	(0.0	,	0.0	,	0.0	,	0.0	,	0.0)
512	KB:	(0.0	,	0.0	,	0.0	,	0.0	,	0.0)
1024	KB:	(0.0	,	0.0	,	0.0	,	0.0	,	0.0)

(The tuples in the image have conflict miss rate in the increasing order of associativity (1,2,4,8,FA))

Plot #2

We observe the AAT when size changes for various associativity values. For this experiment, we use BLOCKSIZE = 32, and assume that there are no Victim and L2 caches.

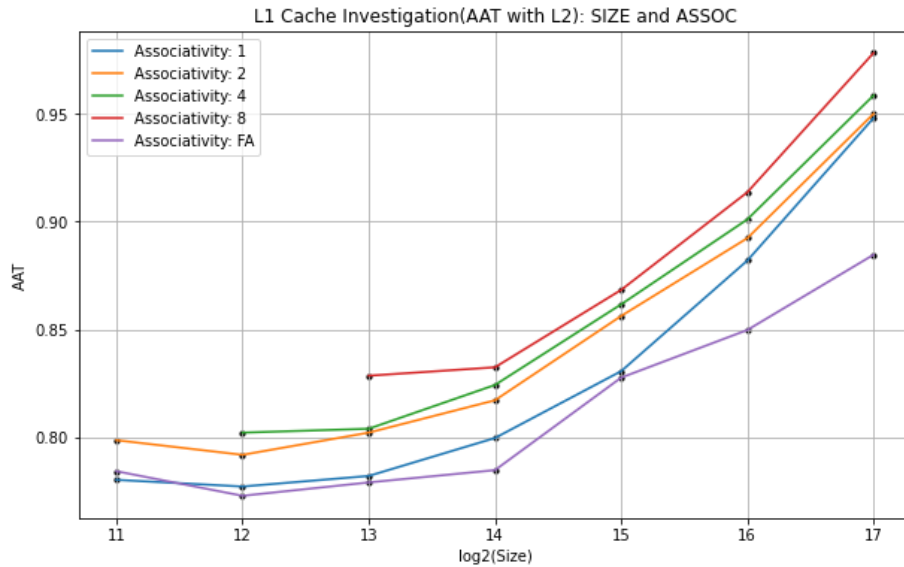


For a memory hierarchy with only an L1 cache and BLOCKSIZE = 32, which configuration yields the best AAT?

From the graph, we can observe that the best AAT (0.7842 ns) is achieved when the cache is fully associative and the size of the cache is 16 KB.

Plot #3

We observe the AAT when size changes for various associativity values when an 8-associative 256KB L2 cache is present.



With the L2 cache added to the system, which L1 cache configuration yields the best AAT? What is the %improvement in this optimal AAT compared to the optimal AAT in plot #2?

From the graph, we can observe that the best AAT (0.7728 ns) is achieved when the cache is fully associative and the size of the cache is 4 KB.

$$\%improvement = 1.453\%$$

Compare the EDP and total area for the optimal-AAT configuration with L2 cache (plot #3) versus without L2 cache (plot #2).

With L2:

$$EDP = 147816822.3823 \text{ nJ.ns}$$

$$\text{total area} = 1.1784 \text{ mm}^2$$

Without L2:

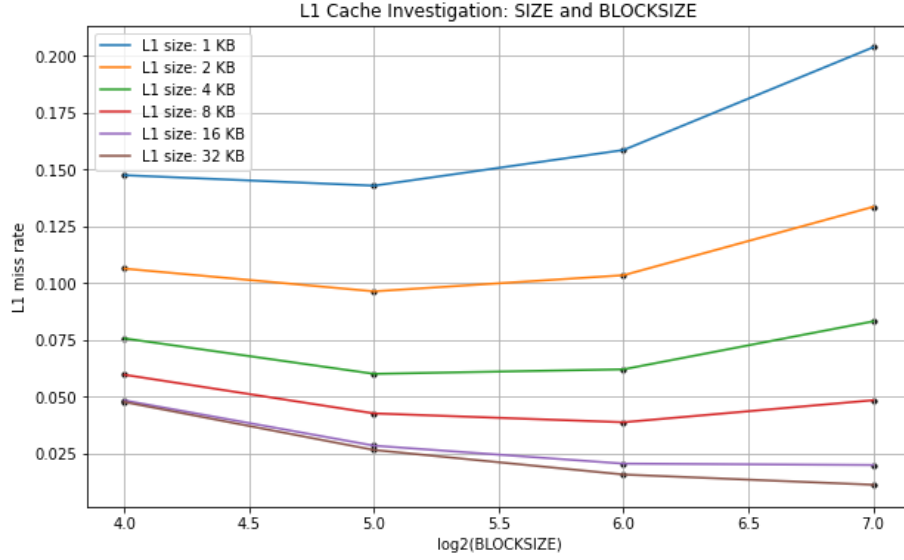
$$EDP = 460720976.6212 \text{ nJ.ns}$$

$$\text{total area} = 0.0634 \text{ mm}^2$$

L1 Cache Investigation: SIZE and BLOCKSIZE

Plot #4

We observe the L1 miss rate when the block size is varied for various sizes of the L1 cache. For this experiment, we use $\text{ASSOC} = 4$ and vary both the SIZE and the BLOCKSIZE. we assume that there is no L2 cache.



As block size is increased from 16 to 128, is the tradeoff between exploiting more spatial locality versus increasing cache pollution evident in the graph? Does the spatial locality-cache pollution trade-off depend on the cache size?

As block size increases, spatial locality exploitation increases, which tends to decrease the miss rate. But when done excessively, it results in cache pollution, which tends to increase the miss rate. For 1, 2, 4, and 8 KB L1 caches, we can clearly see the tradeoff in the graph. For these sizes, the miss rate first decreases and then increases as the block size increases. For 16 and 32 KB L1 caches, the miss rate seems to keep decreasing. This is because the block size is not increased enough for the cache pollution factor to overpower the spatial locality exploitation factor.

As we can infer from the graph, for bigger-sized caches, the optimal block size is higher compared to smaller-sized caches. For 16 and 32 KB L1 caches, the optimal block size has not been reached yet.

What is the optimal BLOCKSIZE for a 4-way set-associative 8KB L1 cache?

As we can observe from the graph, the optimal BLOCKSIZE for a 4-way

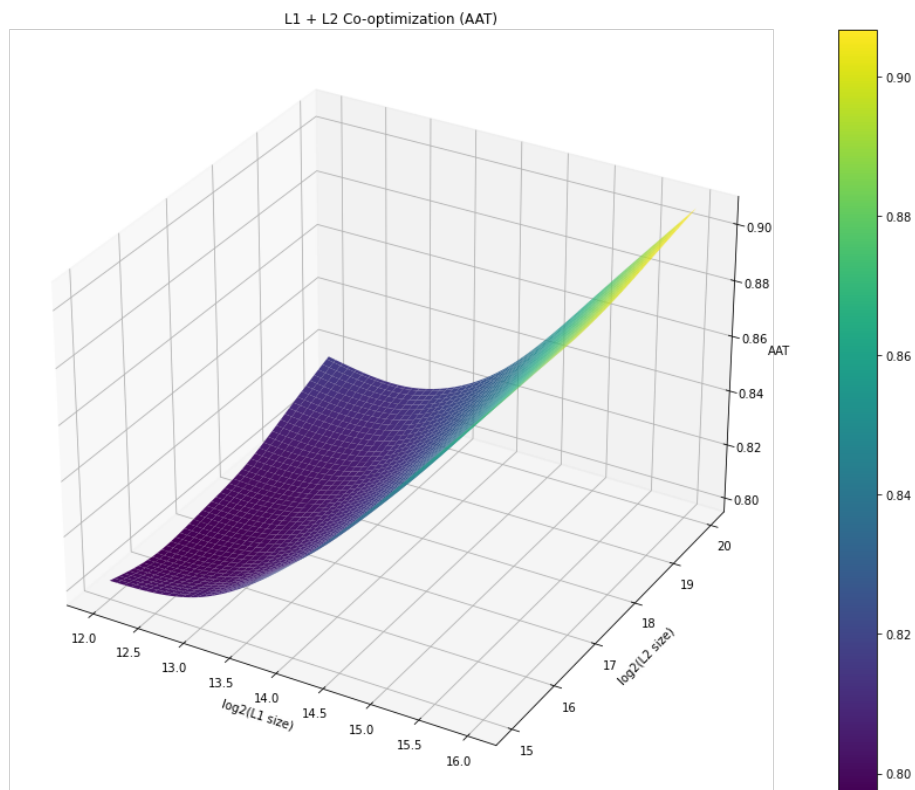
set-associative 8KB L1 cache is 64 bytes.

L1 + L2 Co-optimization

In this experiment, we use a 32-block, 4-way set associative L1 cache and a 32-block, 8-way set associative L2 cache.

Plot #5

We observe the AAT when both the L1 and L2 sizes are varied.



Which memory hierarchy configuration yields the best AAT?

The best AAT (0.7972 ns) is achieved when a 4 KB L1 cache and a 64 KB L2 cache are used.

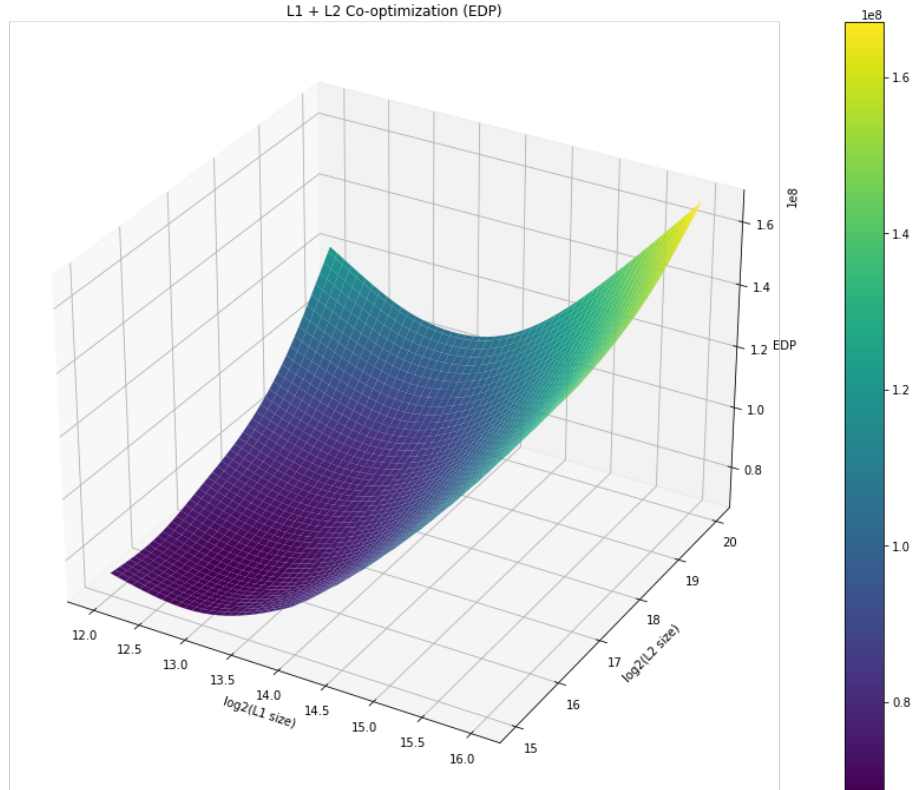
Can you propose a memory hierarchy configuration that has smaller total area, but provides an AAT within 5% of the best AAT?

We are looking for a memory hierarchy configuration with AAT between 0.7573 and 0.8371 ns. The total area of the memory hierarchy configuration with best

AAT is 0.3858 mm^2 . When a 4 KB L1 cache and a 32 KB L2 cache are used, the AAT is 0.8017 ns, and the total area is 0.2853 mm^2 .

Plot #6

We observe the EDP when both the L1 and L2 sizes are varied.



Which memory hierarchy configuration yields the best EDP?

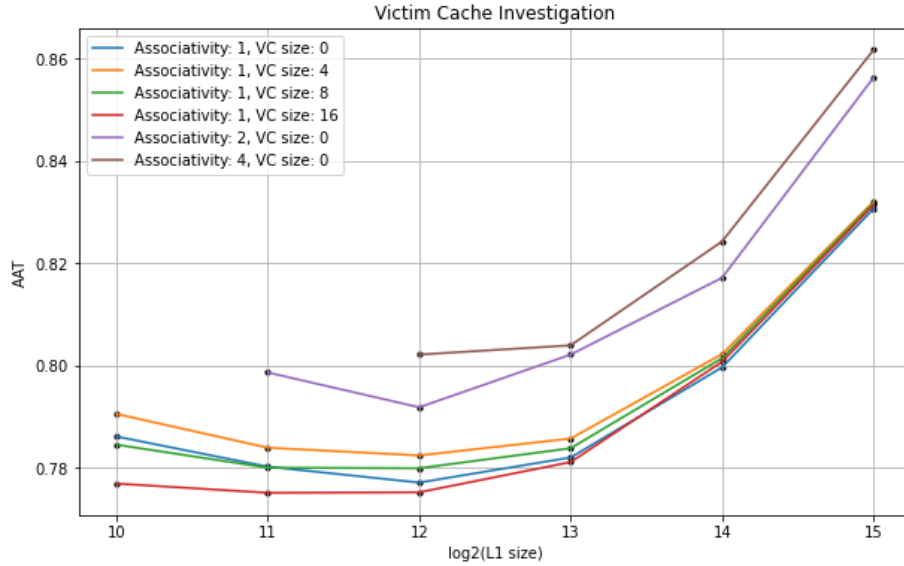
The best EDP (68614171.431 nJ.ns) is achieved when an 8 KB L1 cache and a 64 KB L2 cache are used.

Victim Cache Investigation

In this experiment, we use a 32-block set associative L1 cache and a 256 KB, 32-block, 8-way set associative L2 cache.

Plot #7

We observe the AAT when both the size of the L1 cache is varied for various configurations of the L1 cache and various sizes of the victim cache.



Does adding a Victim Cache to a direct-mapped L1 cache yield performance comparable to a 2-way set-associative L1 cache of the same size?

From the graph, we can observe that, when size < 4 KB, AATs of a 2-way set associative L1 cache with 0 blocks of victim cache are very close to AATs of direct-mapped L1 cache with 8 blocks of victim cache, when size ≥ 4 KB, AATs of a 2-way set associative L1 cache with 0 blocks of victim cache are very close to AATs of direct-mapped L1 cache with 16 blocks of victim cache.

As expected, adding victim caches to direct mapped caches gives them an illusion of associativity.

Which memory hierarchy configuration yields the best AAT?

A 2 KB direct-mapped cache with a victim cache that can hold 16 blocks has the best AAT (0.7751 ns).

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

We are looking for a memory hierarchy configuration with AAT between 0.7363 and 0.8139 ns. The total area of the memory hierarchy configuration with the best AAT is 1.1785 mm^2 . When a 2 KB direct-mapped L1 cache with a victim

cache that can hold 64 blocks and a 32 KB L2 cache are used, the AAT is 0.7641 ns, and the total area is 0.2690 mm^2 .