

NC State University
Department of Electrical and Computer Engineering
ECE 463/563 (Prof. Rotenberg)
Project #1: Cache Design, Memory Hierarchy Design
REPORT (Version 2.0)

by

MAHARSHI MUKESHKUMAR OZA

NCSU Honor Pledge: "I have neither given nor received unauthorized aid on this project."

Student's electronic signature: Maharshi Mukeshkumar Oza
(sign by typing your name)

Course number: 563

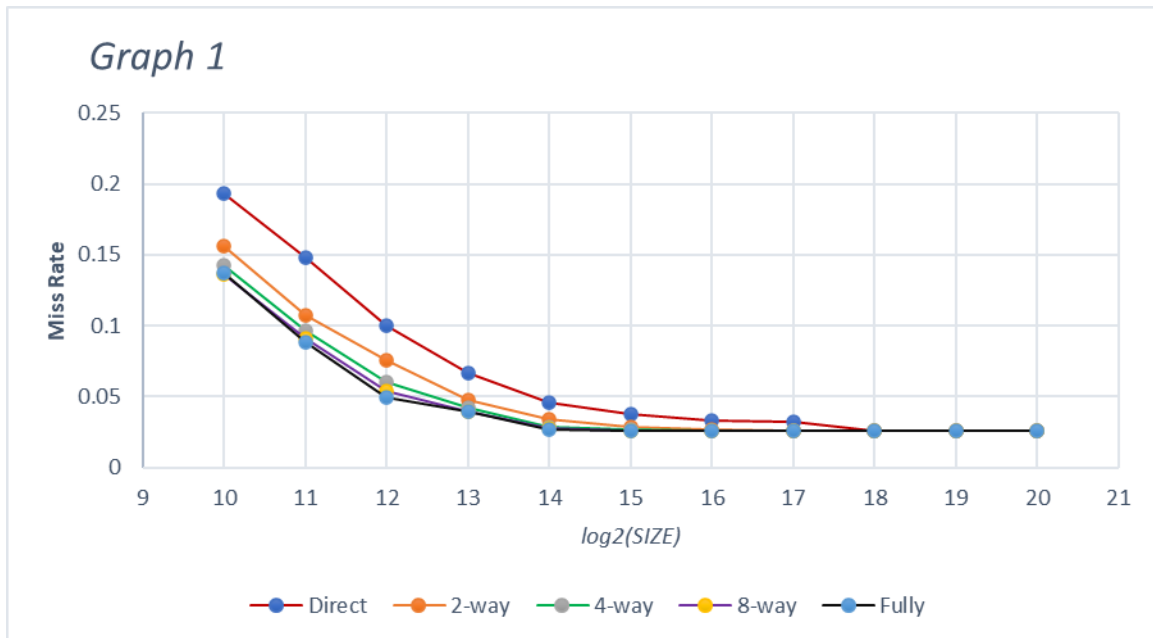
1. L1 cache exploration: SIZE and ASSOC

GRAPH #1 (total number of simulations: 55)

For this experiment:

- Benchmark trace: gcc_trace.txt
- L1 cache: SIZE is varied, ASSOC is varied, BLOCKSIZE = 32.
- L2 cache: None.
- Prefetching: None.

Plot L1 miss rate on the y-axis versus $\log_2(\text{SIZE})$ on the x-axis, for eleven different cache sizes: SIZE = 1KB, 2KB, ..., 1MB, in powers-of-two. (That is, $\log_2(\text{SIZE}) = 10, 11, \dots, 20$.) The graph should contain five separate curves (*i.e.*, lines connecting points), one for each of the following associativities: direct-mapped, 2-way set-associative, 4-way set-associative, 8-way set-associative, and fully-associative. All points for direct-mapped caches should be connected with a line, all points for 2-way set-associative caches should be connected with a line, *etc.*



Answer the following questions:

1. For a given associativity, how does increasing cache size affect miss rate?

Assuming associativity stays the same, increasing the cache size leads to a reduction in the miss rate, but only up to a certain point. Beyond that limit, further increasing the cache size no longer reduces the miss rate. As shown in the graph, this is observed as the cache size increases from 2KB to 1MB.

2. For a given cache size, how does increasing associativity affect miss rate?

Report for Project #1.

With a fixed cache size, a direct-mapped cache has the most cache misses, while a fully associative cache has the least. Different levels of cache associativity help reduce misses caused by conflicts or the cache running out of space. As associativity increases, the miss rate gets closer to the lowest possible level, which is determined by the data that must be loaded at least once (compulsory misses).

3. Estimate the *compulsory miss rate* from the graph and briefly explain how you arrived at this estimate.

compulsory miss rate = 0.0258

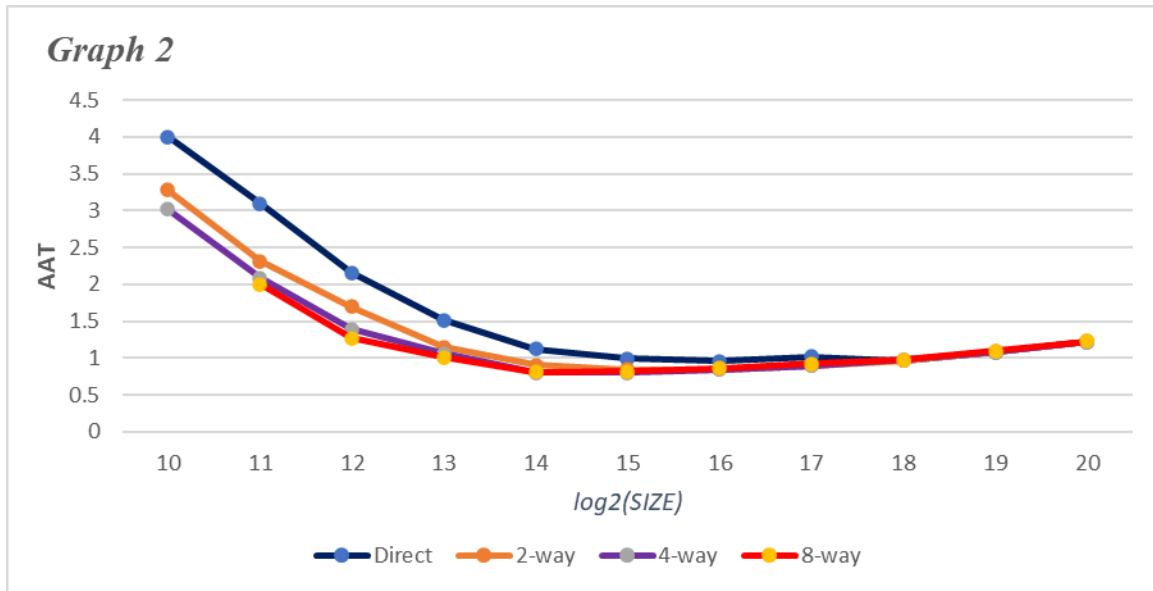
How I arrived at this estimate:

After a certain point, the improvement slows down because only the compulsory miss rate remains. All types of cache associativity help eliminate conflict and capacity misses until they reach the compulsory miss rate, which in this case is 0.0258. It can't be reduced any further beyond that.

GRAPH #2 (no additional simulations with respect to GRAPH #1)

Same as GRAPH #1, except make the following changes:

- The y-axis should be AAT instead of L1 miss rate.
- Do NOT include fully-associative cache configurations (power-hungry, impractical cost for implementing LRU). Thus, GRAPH #2 should contain four (not five) separate curves: direct-mapped, 2-way set-associative, 4-way set-associative, 8-way set-associative.



Answer the following question:

1. For a memory hierarchy with only an L1 cache and BLOCKSIZE = 32, which configuration yields the best (*i.e.*, lowest) AAT and what is that AAT?

Direct-mapped or set-associative cache configuration that yields the lowest AAT:

16 KB, 8-way set-assoc.

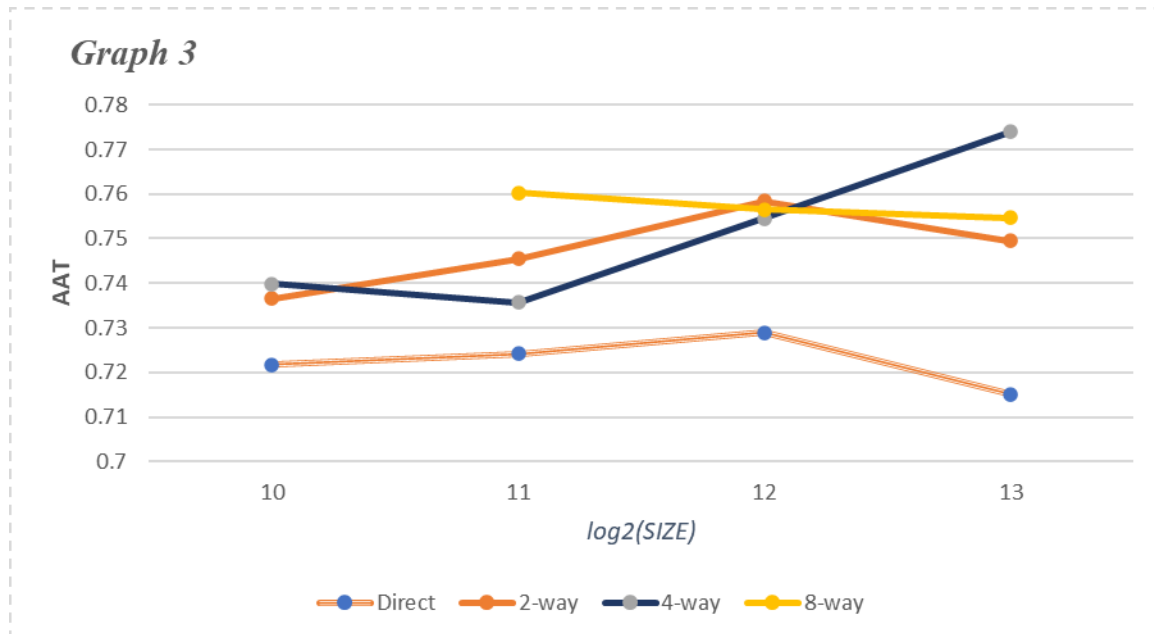
AAT for this configuration: 0.811124 nanoseconds.

$$\text{AAT} = \text{HT} + \text{MR} \times \text{MP}$$

GRAPH #3 (total number of simulations: 16)

Same as GRAPH #2, except make the following changes:

- Add the following L2 cache to the memory hierarchy: 16KB, 8-way set-associative, same block size as L1 cache.
- Vary the L1 cache size only between 1KB and 8KB (since L2 cache is 16KB). (And as with GRAPH #2, GRAPH #3 should contain four (not five) separate curves for L1 associativity: direct-mapped, 2-way set-associative, 4-way set-associative, 8-way set-associative.)



Answer the following questions:

1. With the L2 cache added to the system, which L1 cache configuration yields the best (*i.e.*, lowest) AAT and what is that AAT?

L1 configuration that yields the lowest AAT with 16KB 8-way L2 added:
8 KB, direct-mapped.

AAT for this configuration: 0.715107608

2. How does the lowest AAT with L2 cache (GRAPH #3) compare with the lowest AAT without L2 cache (GRAPH #2)?

The lowest AAT with L2 cache is less than the lowest AAT without L2 cache.

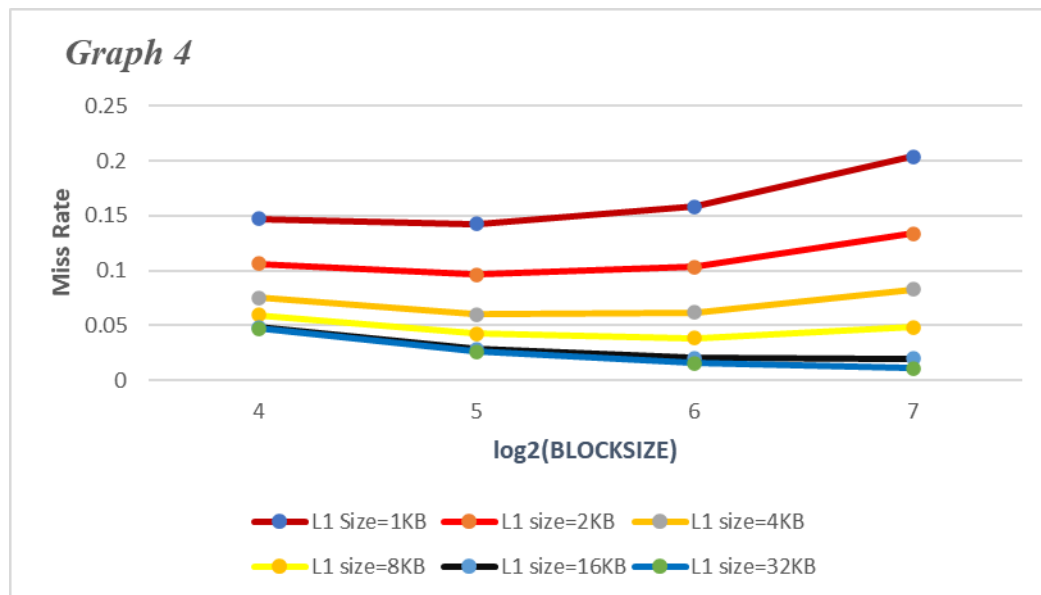
2. L1 cache exploration: SIZE and BLOCKSIZE

GRAPH #4 (total number of simulations: 24)

For this experiment:

- Benchmark trace: gcc_trace.txt
- L1 cache: SIZE is varied, BLOCKSIZE is varied, ASSOC = 4.
- L2 cache: None.
- Prefetching: None

Plot L1 miss rate on the y-axis versus $\log_2(\text{BLOCKSIZE})$ on the x-axis, for four different block sizes: $\text{BLOCKSIZE} = 16, 32, 64, \text{ and } 128$. (That is, $\log_2(\text{BLOCKSIZE}) = 4, 5, 6, \text{ and } 7$.) The graph should contain six separate curves (*i.e.*, lines connecting points), one for each of the following L1 cache sizes: $\text{SIZE} = 1\text{KB}, 2\text{KB}, \dots, 32\text{KB}$, in powers-of-two. All points for $\text{SIZE} = 1\text{KB}$ should be connected with a line, all points for $\text{SIZE} = 2\text{KB}$ should be connected with a line, *etc.*



Answer the following questions:

1. Do smaller caches prefer smaller or larger block sizes?

Smaller caches prefer smaller block sizes. For example, the smallest cache considered in Graph #4 (1KB) achieves its lowest miss rate with a block size of 16 B.

2. Do larger caches prefer smaller or larger block sizes?

Larger caches prefer larger block sizes. For example, the largest cache considered in Graph #4 (32KB) achieves its lowest miss rate with a block size of 128 B.

3. As block size is increased from 16 to 128, is the tension between *exploiting more spatial locality* and *cache pollution* evident in the graph? Explain.

Yes, the tension between *exploiting more spatial locality* and *cache pollution* is evident in the graph.

For example, consider the smallest (1KB) cache in Graph #4. Increasing block size from 16 B to 32 B is helpful (reduces miss rate) due to exploiting more spatial locality. But then increasing block size further, from 32 B to 128 B, is not helpful (increases miss rate) due to cache pollution having greater effect.

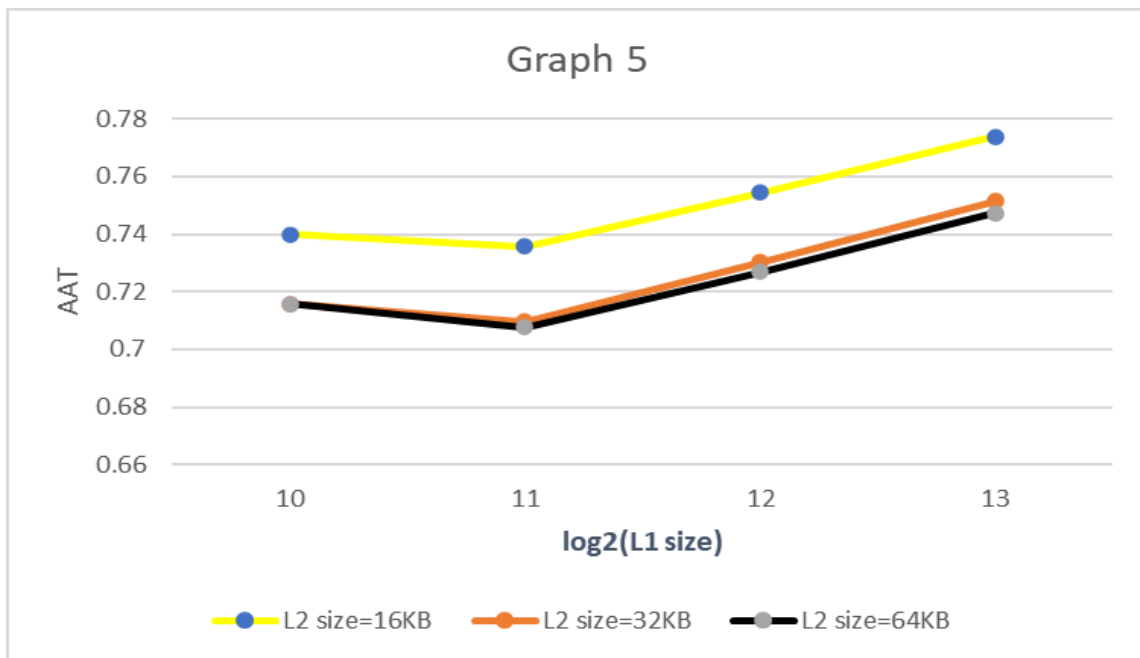
3. L1 + L2 co-exploration

GRAPH #5 (total number of simulations: 12)

For this experiment:

- Benchmark trace: gcc_trace.txt
- L1 cache: SIZE is varied, BLOCKSIZE = 32, ASSOC = 4.
- L2 cache: SIZE is varied, BLOCKSIZE = 32, ASSOC = 8.
- Prefetching: None.

Plot AAT on the y-axis versus $\log_2(\text{L1 SIZE})$ on the x-axis, for four different L1 cache sizes: L1 SIZE = 1KB, 2KB, 4KB, 8KB. (That is, $\log_2(\text{L1 SIZE}) = 10, 11, 12, 13$.) The graph should contain three separate curves (*i.e.*, lines connecting points), one for each of the following L2 cache sizes: 16KB, 32KB, 64KB. All points for the 16KB L2 cache should be connected with a line, all points for the 32KB L2 cache should be connected with a line, *etc.*



Answer the following question:

1. Which memory hierarchy configuration in Graph #5 yields the best (*i.e.*, lowest) AAT and what is that AAT?

Configuration that yields the lowest AAT: L1 Size = 2KB and L2 Size = 64KB.
Lowest AAT: 0.707657833

4. Stream buffers study (ECE 563 students only)

TABLE #1 (*total number of simulations: 5*)

For this experiment:

- Microbenchmark: stream_trace.txt
- L1 cache: SIZE = 1KB, ASSOC = 1, BLOCKSIZE = 16.
- L2 cache: None.
- PREF_N (number of stream buffers): 0 (pref. disabled), 1, 2, 3, 4
- PREF_M (number of blocks in each stream buffer): 4

The trace “stream_trace.txt” was generated from the loads and stores in the loop of interest of the following microbenchmark:

```
#define SIZE 1000

uint32_t a[SIZE];
uint32_t b[SIZE];
uint32_t c[SIZE];

int main(int argc, char *argv[]) {
...
    // LOOP OF INTEREST
    for (int i = 0; i < SIZE; i++)
        c[i] = a[i] + b[i];    // per iteration: 2 loads (a[i], b[i]) and 1 store (c[i] = ...)
...
}
```

Fill in the following table and answer the following questions:

PREF_N, PREF_M	L1 miss rate
0,0 (pref. disabled)	0.2500
1,4	0.2500
2,4	0.2500
3,4	0.0010
4,4	0.0010

1. For this streaming microbenchmark, with prefetching disabled, do L1 cache size and/or associativity affect the L1 miss rate (feel free to simulate L1 configurations besides the one used for the table)? Why or why not?

With prefetching disabled, L1 cache size and/or associativity do not affect L1 miss rate (for this streaming microbenchmark).

The reason: Once the L1 cache size reaches a certain point (1KB in this case), all the conflict and capacity misses are minimized, so the miss rate levels out and only compulsory misses remain. Compulsory misses happen when data is accessed for the first time, and increasing the cache size won't help with that. Even if the cache grows larger, new data blocks will continue to replace old ones, leading to compulsory misses. In this case, since a 16B block only holds 4 data blocks, you'll miss on the 5th block because the cache size doesn't guarantee all blocks can stay in memory.

Why increasing L1 cache associativity doesn't help:

Associativity helps when there are conflict misses, which happen when multiple data blocks try to occupy the same spot in the cache. But in this scenario, the data is being accessed in a sequential order (e.g., `a[0]`, `a[1]`, `a[2]`, etc.), so there's no overlap or competition for cache space. Each block is used only once and then replaced by the next one. Therefore, increasing associativity won't help because the misses here are not due to conflicts—they're compulsory misses, since new blocks are constantly being brought into the cache with each access.

2. For this streaming microbenchmark, what is the L1 miss rate with prefetching disabled? Why is it that value, *i.e.*, what is causing it to be that value? Hint: each element of arrays `a`, `b`, and `c`, is 4 bytes (`uint32_t`).

The L1 miss rate with prefetching disabled is 0.2500, because the miss rate is 25% because the cache can hold 4 elements in a block, but every 5th element causes a miss.

3. For this streaming microbenchmark, with prefetching disabled, what would the L1 miss rate be if you doubled the block size from 16B to 32B? (hypothesize what it will be and then check your hypothesis with a simulation)

The L1 miss rate with prefetching disabled and a block size of 32B is 0.126, because if we increase the block size by 2 times, the miss rate will generally decrease by x times as well. This means we can store $4x$ elements in our cache block, and the miss rate will occur after every $4x$ -th element. Therefore, the ratio is $1/4x$. Assuming $x = 2$ in this case, the miss rate is 12.5%.

4. With prefetching enabled, what is the minimum number of stream buffers required to have any effect on L1 miss rate? What is the effect on L1 miss rate when this many stream buffers are used: specifically, is it a modest effect or huge effect? Why are this many stream buffers required? Why is using fewer stream buffers futile? Why is using more stream buffers wasteful?

Minimum number of stream buffers needed to have any effect on L1 miss rate: 3

With this many stream buffers, the effect on L1 miss rate is huge. Specifically, the L1 miss rate is nearly 0.001. We only miss on the first elements of `a`, `b` & `c` (hence a total of 3 misses).

This many stream buffers are required because each buffer has to prefetch data for all the 1 array, and in total there are 3 arrays which are `a`, `b` and `c`.

Using fewer stream buffers is futile because there is going to be at-least 1 of 3 that is not going to be prefetched in advance.

Using more stream buffers is wasteful because there are only 3 arrays `a`, `b` and `c` whose data needs to be prefetched so adding more stream buffers is not going to be used.