

1 Review

1.1 Give the formulas for splitting an address into tag, index, and offset (TIO).

1.2 Define the 3 C's of cache misses and how to improve each.

(a) Compulsory

(a) Capacity

(a) Conflict

2 Cache Misses (Sp'24 MT1, Q4)

- 2.1 Assume you have a 4 KiB 2-way set associative L1 data cache with 64-byte cache lines, and that you are using 32-bit addresses. How many bits are required for the tag, index, and offset?

Offset:

Index:

Tag:

- 2.2 Now consider the following program. Assume that each element is a long (8 bytes) and that the matrix elements are stored in row-major order. Assume that the first element in the matrix is aligned to the start of a cache line, that sum, i, and j are stored in registers, and the same cache as above.

```
int sum = 0;
int matrix[64][64];
for (int i = 0; i < 64; i++) {
    for (int j = 0; j < 64; j++) {
        sum += matrix[i][j];
    }
}
```

(a) What is the number of cache misses?

(a) How many hits?

(a) How many compulsory misses?

(a) How many conflict misses?

(a) How many capacity misses?

2.3 Bonus: How can we improve our hit rate?

3 Sectored Cache (Sub-blocks)

100	1		1		1		1	
300	1		1		0		0	
204	0		1		0		1	

3.1 What problem was this optimization trying to solve? Explain.

Exam Practice (Sp'22 MT1, Q3)

We have an 8 KiB, two-way set associative cache with 16-byte cache blocks. The replacement policy is LRU. Each cache block is broken up into four smaller sub-blocks. Each element in `buf` is 4 bytes. Assume `sum`, `i`, and `j` are stored in registers.

```
// Loop A
sum = 0;
for (int i = 0; i < 128; i++)
    for (int j = 0; j < 32; j++)
        sum += buf[i*32 + j];

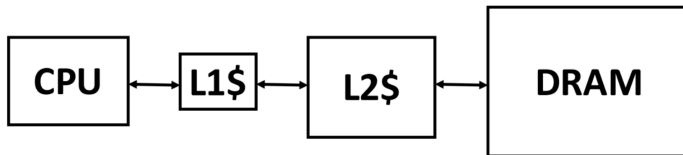
// Loop B
sum = 0;
for (int j = 0; j < 32; j++)
    for (int i = 0; i < 128; i++)
        sum += buf[i*32 + j];
```

3.2 What is the number of misses for Loop A and for Loop B with the sectored cache?

3.3 What is the number of misses for Loop A and for Loop B if the cache is not sectored (no sub-blocks)?

- 3.4 Qualitatively explain whether our sectoring scheme has improved the average memory access time of Loop A and Loop B. If so, why? If not, why not?

4 Hierarchical Caches



4.1 What problem was this optimization trying to solve? Explain.

Exam Practice

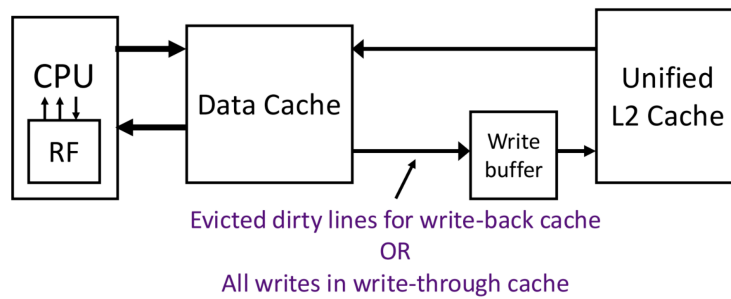
Assume we have a two-level cache hierarchy: $L1 \rightarrow L2 \rightarrow \text{Memory}$.

	Hit Time	Local Miss Rate
L1	2 cycles	8%
L2	15 cycles	25%
Main Memory	200 cycles	-

4.2 What is the AMAT of this system?

4.3 Say we double the size of L1. The L1 local miss rate falls from 8% to 5%, and the L2 local miss rate rises from 25% to 35%. Compute the new AMAT. Then explain in at most two sentences why the L2 local miss rate went up even though the hierarchical AMAT got better.

5 Write Buffers



5.1 What problem was this optimization trying to solve? Explain.

Exam Practice (Sp'22 MT1)

5.2 How do write buffers affect write miss penalties in a write-through cache? (A write miss occurs when the program attempts to STORE to an address that is not in the cache.) Explain your reasoning.

5.3 How do write buffers affect read miss penalties? (A read miss occurs when the program attempts to LOAD from an address that is not in the cache.) Explain your reasoning for both a write-back and a write-through cache.