Cache Performance

Putting it all together

Cache capacity, associativity, set size, and block size are typically powers of 2.

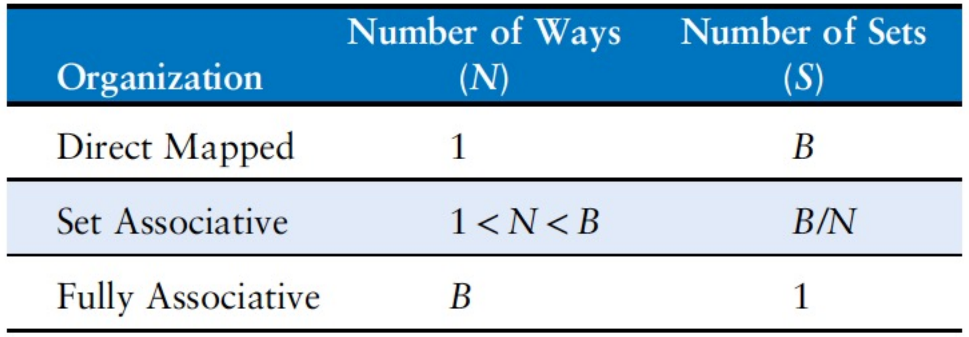
- Makes cache fields subsets of the address bits.

Increasing associativity N usually reduces miss rate caused by conflicts

- However, it requires more tag comparators.

Increasing block size takes advantage of spatial locality to reduce the miss rate.

- Decreases the number of sets and can lead to more conflicts.
- Increases the miss penalty.



Organization	Number of Ways (N)	Number of Sets (S)
Direct Mapped	1	B
Set Associative	$1 < N < B$	B/N
Fully Associative	B	1

Memory System Performance Analysis

$$Miss\ Rate = \frac{Number\ of\ misses}{Number\ of\ total\ memory\ accesses} = 1 - Hit\ Rate$$

$$Hit\ Rate = \frac{Number\ of\ hits}{Number\ of\ total\ memory\ accesses} = 1 - Miss\ Rate$$

$$AMAT = t_{\text{cache}} + MR_{\text{cache}}(t_{MM} + MR_{MM}t_{VM})$$

Performance Example

M

H

H

M

M

H

H

H

M

Cache access = 1ns
Mem access = 1000ns

Hit rate?

Miss rate?

Average mem. access time?

Improvement due to cache?

Accessing Main Memory is Slow

Reducing Miss Rate

Can reduce misses by changing capacity, block size, and/or associativity

What causes misses?

Compulsory

First request to a cache block will always be a miss

Cache Performance

Compulsory

First request to a cache block will always be a miss

Conflict

Several addresses map to the same set and evict still needed blocks

Altering parameters can affect one or more of these misses.

Increasing capacity reduce conflict misses but not compulsory.
Increasing block size could reduce compulsory misses due to spatial locality,
but increases conflict misses because more addresses map to the same set.

Cache Performance

Run benchmarks to determine the performance of different parameters.

- Increasing block size can lead to an increase in miss rates.

Depends on the size of the cache.

Large caches might see increased execution time anyways due to larger miss penalty.

Write Policy

Write-through cache

Data written to the cache is written simultaneously to memory.

Write-back Cache

A dirty bit is assigned to each block

1 when the block has been written and 0 otherwise

Dirty blocks are written back to main memory only when they are evicted from the cache.

Write through doesn't require the bit, but generally requires more main memory writes.

Modern caches are usually write-back because main memory access time is so large.

Summary

1. Cache memory used to reduce the need to go to RAM for reading data.

2. Cache strategy exploits spatial locality or temporal locality of data to keep most likely used information ready.

3. Different caching strategies have strengths and weaknesses.

   Cache performance must be verified empirically.