Large and Fast: Exploiting Memory Hierarchy

Chapter Five Of

Book of P. Hayes

Multilevel Caching

- ✔ A second level of cache is introduced on the same chip as microprocessor or on a separate chip.
- ✔ The purpose is to reduce the miss penalty.
- ✔ Primary cache is known as L1 cache.
- ✔ Secondary cache is known as L2 cache.
- ✔ If the desired data is in the secondary cache, then miss penalty= access time of secondary cache.
- ✔ If data is not available on the primary or secondary cache, then miss penalty= access time of main memory.

Performance of Multilevel Cache

- ✔ CPI = 1.0
- ✔ Clock rate = 500 GHz
- ✔ Clock period = 1/ 500 GHz = 0.2 ns
- ✔ Main memory access time = 100 ns.
- ✔ Miss rate at the primary cache = 2%

Miss penalty of main memory = 100 ns/0.2 ns/cycle = 500 clock cycles

CPI with one level caching = 1.0 + 2% × 500 =11.0

Performance of Multilevel Cache

- ✔ Secondary cache access time = 5 ns.
- ✔ Miss rate to main memory = 0.5%.

Miss penalty of secondary cache = 5 ns / 0.2 ns/clock cycles
=25 clock cycles

Total CPI = 1+ primary stalls per instruction + secondary stalls per instruction

$= 1 + 2\% \times 25 + 0.5\% \times 500 = 4.0$

The processor with secondary cache is faster by 11.0 / 4.0 = 2.8.

Multilevel Caching

- ✔ The primary cache focus on minimizing the hit time

  - Use smaller block size

- ✔ The secondary cache focus on minimizing the miss rate to reduce the penalty of long memory access time.

  -Use larger block size

- ✔ On chip L1 cache tend to have lower associativity than L2 cache.

Global Miss Rate:

The fraction of references that miss in all levels.

Local Miss Rate:

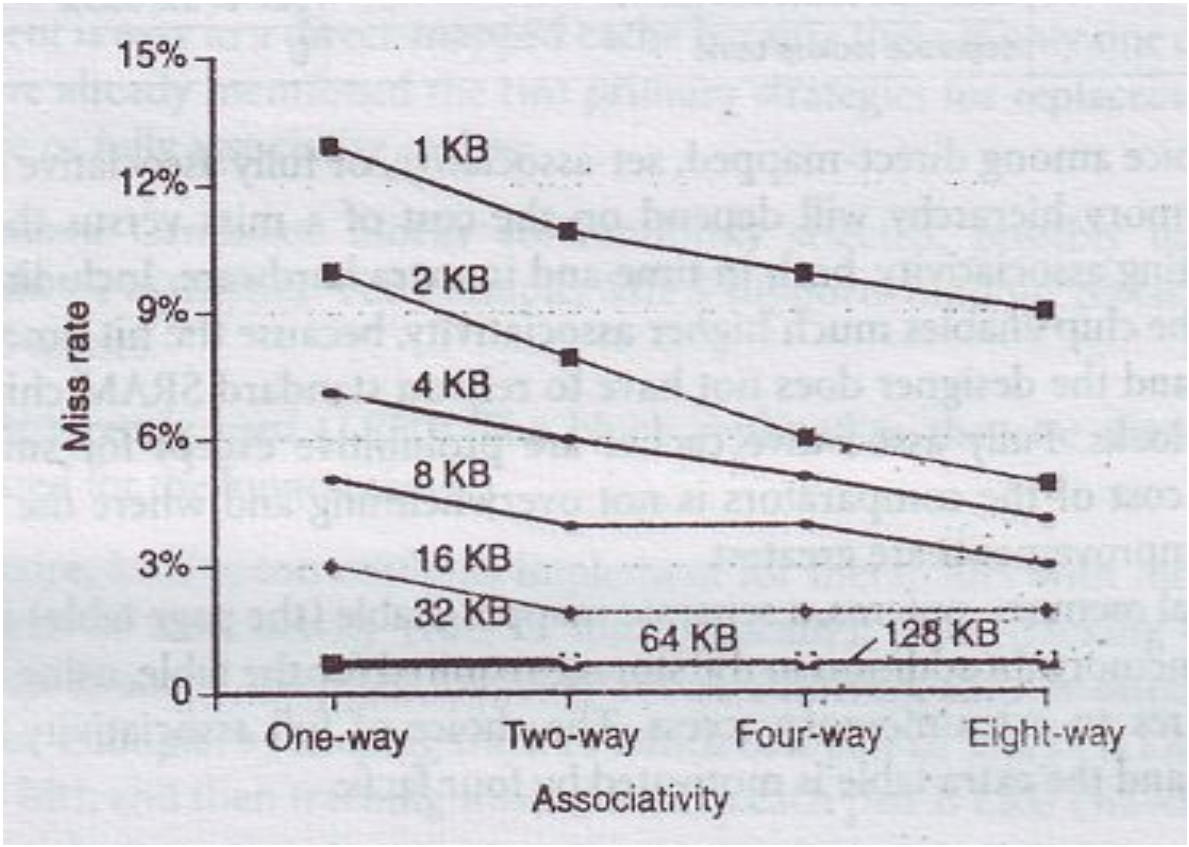
The fraction of references to one level of a cache that miss.

Block Placement

Scheme name	Number of sets	Blocks per set
Direct mapped	Number of blocks in cache	1
Set associative	$\frac{\text{Number of blocks in cache}}{\text{Associativity}}$	Associativity (typically 2–16)
Fully associative	1	Number of blocks in the cache

- ✔ Increasing the degree of associativity decreases the miss rate.
- ✔ Increasing associativity increases the access time and hardware cost.

Block Placement



Locating a Block

Associativity	Location method	Comparisons required
Direct mapped	index	1
Set associative	index the set, search among elements	degree of associativity
Full	search all cache entries	size of the cache
	separate lookup table	0

Block Replacement

1. Random Selection
2. Least Recently Used (LRU)

Three Sources of Misses

1. Compulsory Misses:

   Caused by the first access to a block
2. Capacity Misses:

   Caused by the limited capacity
3. Conflict Misses/ Collision Misses:

   When multiple blocks compete for the same location (direct mapping) or same set (set associative).

Solutions

Design change	Effect on miss rate	Possible negative performance effect
Increase cache size	decreases capacity misses	may increase access time
Increase associativity	decreases miss rate due to conflict misses	may increase access time
Increase block size	decreases miss rate for a wide range of block sizes due to spatial locality	increases miss penalty. Very large block could increase miss rate