

Locality of Reference:-

- Spatial: fetch nearby locations also when getting some data
- Temporal: set of instr. likely to be referenced soon.

Cache Mapping:-

1) Direct Mapped:

Tag	Line	Word
-----	------	------

2) Set Associative:

Tag	Set	Word
-----	-----	------

3) Fully Associative:

Tag	Word
-----	------

Word = block size \times word length

Line = no. of lines in cache
 $= \frac{\text{cache size}}{\text{line/block size}}$

Set = $\frac{\text{line}}{\text{set size}}$ (no. of sets in cache)

Tag = Main mem bits -
 word + set/line bits

Valid bit:-

- for each Block
- 0: initial value / data changed main mem.
- 1: when data is loaded.

Write Hit:-

- 1) Write Through:
 - Both updated
 - consistency ✓
 - Latency ↑

2) Write Back:

- Change Cache only
- Dirty bit (modified bit) = 1 (in cache)
- while "replacing" check dirty bit & update in main mem. at that time.

Adv:

- Latency ↓
- write everything at once

Cache Miss:-

- data not in cache
- miss penalty: /
- Latency: time for 1st word retrieval
- Bandwidth: retrieval of rest of the blocks.

Read Miss:-

- 1) Load through/early restart:
 - DON'T wait to load everything
 - continue once word is found & retrieval runs in //.

Write Miss:-

- 1) Write no allocate:
 - write directly to mem.
 - valid bit = 0
- 2) Write allocate:
 - load data written to cache
 - easy access if same data needed again

Types of Misses:-

- 1) Compulsory miss:
 - initial miss.
- 2) Capacity miss:
 - not enough space to store
- 3) Conflict Miss: *
 - discarded block needs to be retrieved again

Replacement Algos:-

- 1) Random
- 2) LRU
- 3) FIFO
- 4) LFU (least freq. used.)

- * 4) Coherence miss:
 - due to flushing to maintain coherence.

Optimisations:-

- 1) ↑ block size
 - 2) ↑ cache size
 - 3) ↑ associativity
 - 4) Multilevel cache
 - 5) Read priority > write
 - 6) Avoid address translation when indexing cache.
- 1, 2, 3: ↓ miss rate
 4, 5: ↓ miss penalty
 6: ↓ Hit time

vishwa

Size ↑ speed ↓ capacity ↑
Cost ↓ Access time ↑

DATE

Performance Analysis :-

$$\text{Avg. Mem. Access Time (AMAT)} = \text{Hit} + \text{Miss rate} \times \text{miss penalty}$$

access + receive + request

$$\left[\text{request time} + \underbrace{\frac{\text{Block Size}}{\text{bus size}}}_{\text{receive}} + \text{access time} \right]$$

CPU Execution Time:

$$= \text{IC} \times \text{Cycle time} \times \text{Overall CPI}$$

$$\text{Overall CPI} = \text{Base CPI} + \text{CPU stall} + \text{Mem stall}$$

↳ taken 0

$$\text{Mem stall} = \% \text{ mem access} \times \% \text{ miss rate} \times \text{miss penalty}$$

for unified cache :

$$\text{mem stall} = (1 + \% \text{ mem access}) \times \% \text{ miss rate} \times \text{miss penalty}$$

for not unified cache :

$$\text{mem stall} = \% \text{ access} \times \% \text{ d-cache miss} \times \text{penalty} + \% \text{ i-cache miss} \times \text{penalty}$$

for sep. mem access % :-

$$\text{mem stall} = \% \text{ access} \times \% \text{ load} \times \% \text{ d-cache miss} \times \text{read penalty} + \% \text{ access} \times \% \text{ store} \times \% \text{ d-cache miss} \times \text{write penalty} + \% \text{ i-cache miss} \times \text{read penalty}$$