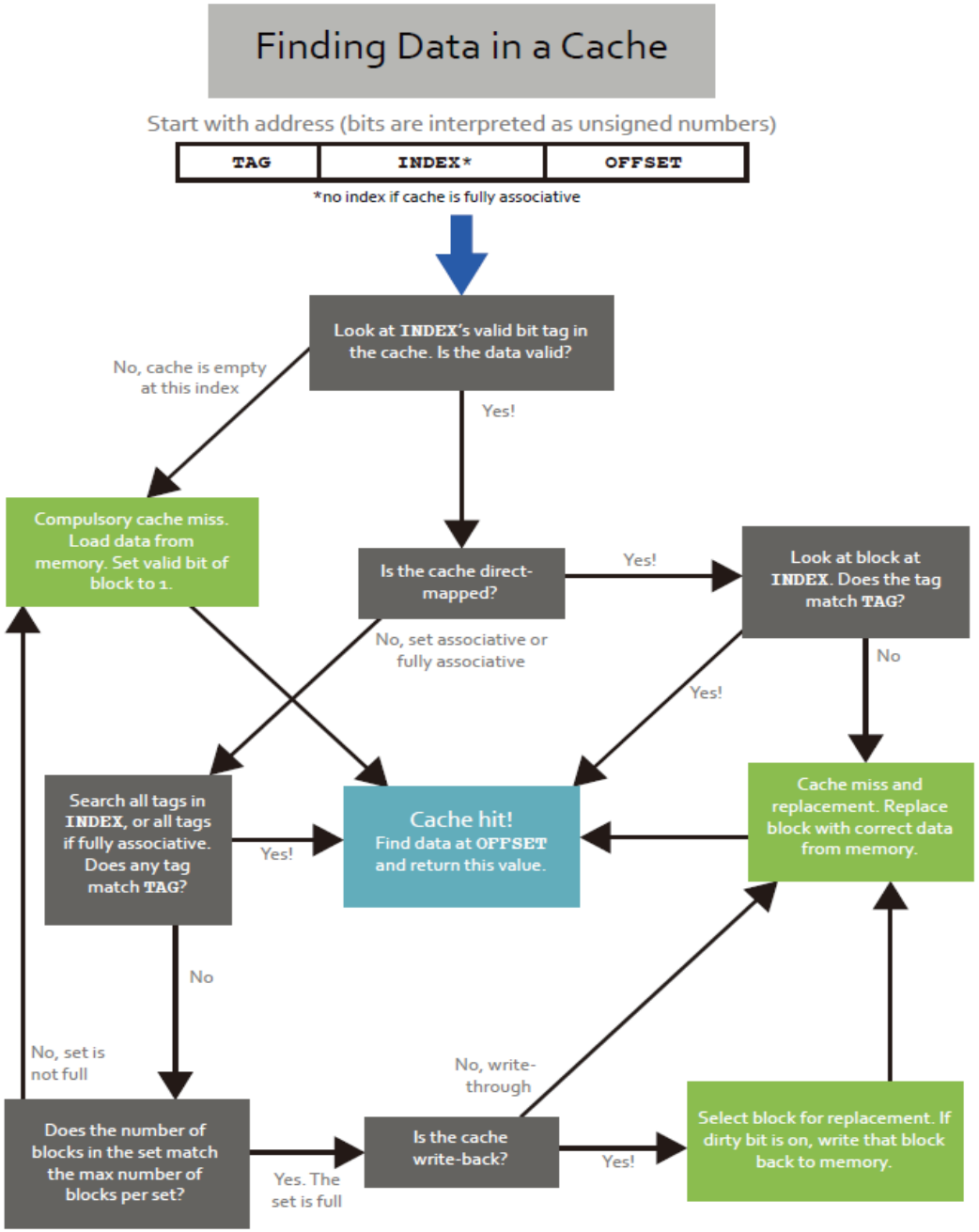
Multilevel Caches

HANG SU

Outline

- Multilevel Cache
- 3C’s
- AMAT

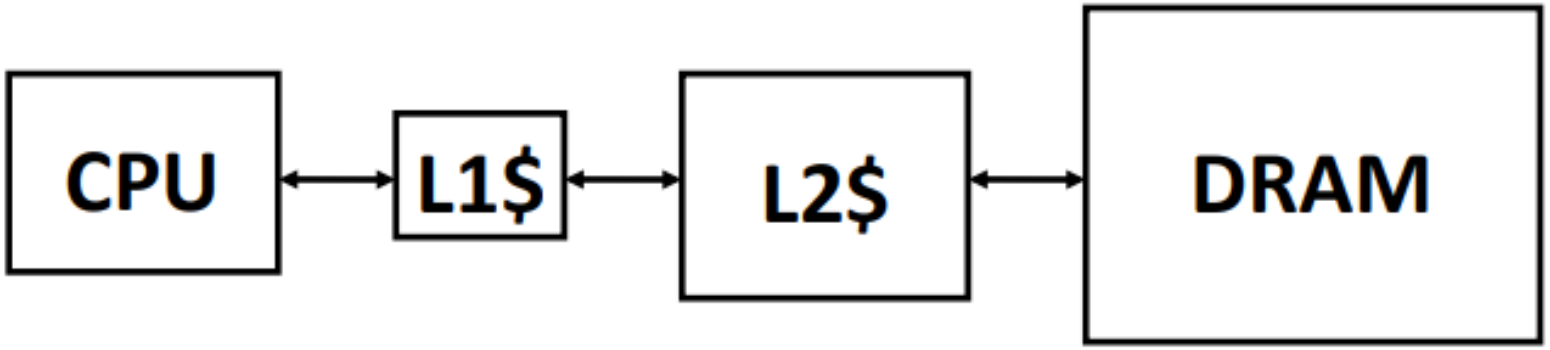
Cache Flowchart

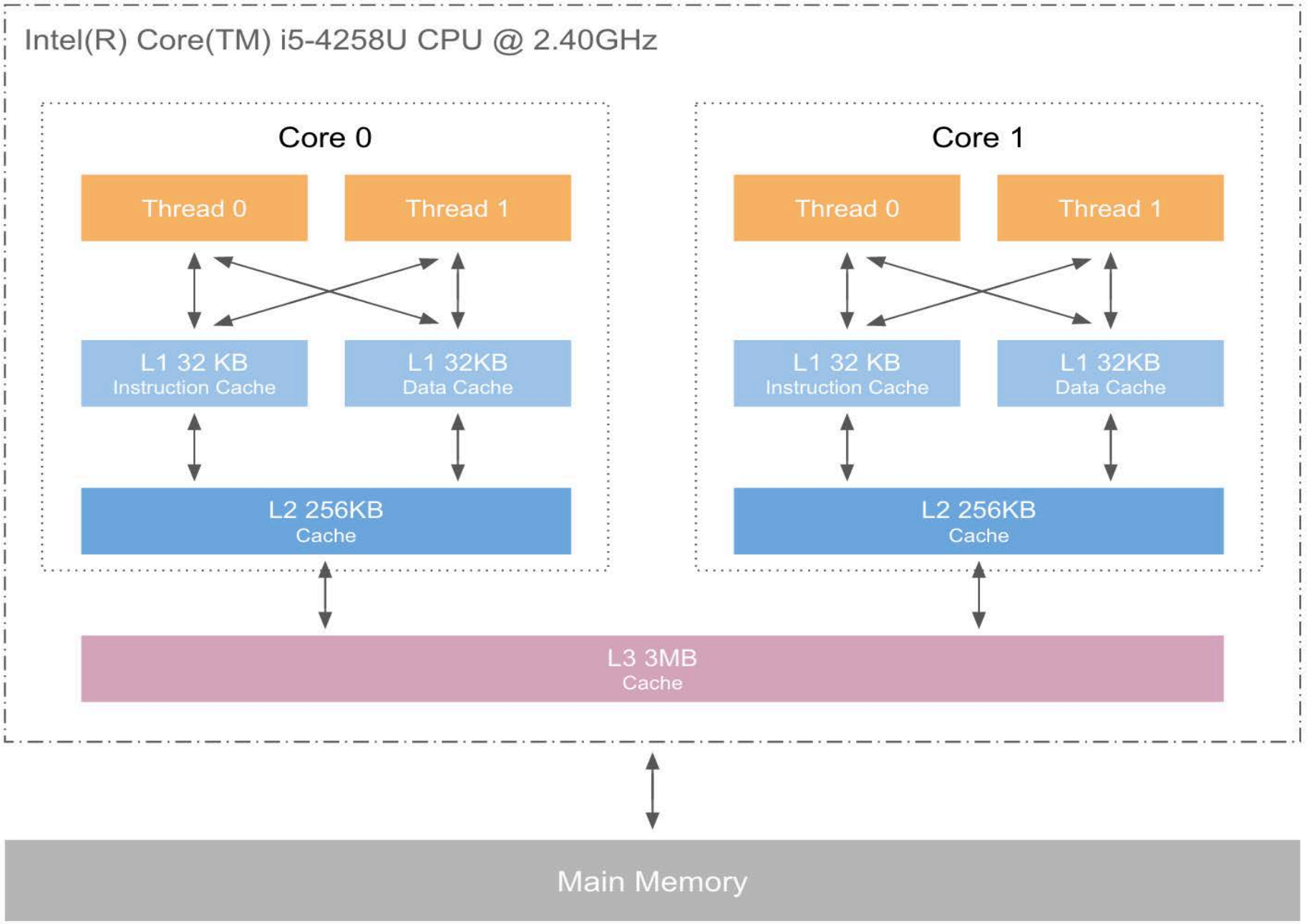


Multilevel Caches

Problem: Miss penalty is big

Solution: Use multiple cache levels




Intel(R) Core(TM) i5-4258U CPU @ 2.40GHz
Core 0
Thread 0
Thread 1
L1 32 KB
Instruction Cache
L1 32KB
Data Cache
L2 256KB
Cache
Core 1
Thread 0
Thread 1
L1 32 KB
Instruction Cache
L1 32KB
Data Cache
L2 256KB
Cache
L3 3MB
Cache
Main Memory

3C's

- Compulsory (cold start)
  - Insignificant if run a huge amount of instructions
- Capacity
  - Cannot contain all accessed data
  - Can be solved by an infinite cache
- Conflict
  - Multiple locations map to one
  - Can be solved by ideal fully associative cache of the same size

3C's

- Larger cache size
  - + Reduces capacity and conflict misses
  - - Hit time will increase
- Higher associativity
  - + Reduce conflict misses
  - - May increase hit time (mux)
- Larger line size
  - + Reduces compulsory and capacity (reload) misses
  - - Increases conflict misses and miss penalty

More Misses

Coherence

- Keep the same shared memory for two different processors
- When one writes to memory, invalidates other processors' cache entry
- Coherence miss when working on the same data

Additional capacity miss

- Happens in multithreaded processor cores

Local Miss Rate

Relative

$$L_n\$ = \frac{L_n\$\ Miss}{L_n\$\ Access} = \frac{L_n\$Miss}{L_{n-1}Miss}$$

Global Miss Rate

Absolute

$$\frac{L_n\$\ Miss}{Total\ Access} =$$

$$\frac{L_nMiss}{L_{n-1}Miss} \times \cdots \times \frac{L_1Miss}{TotalAccess} =$$

$$LMR\ L_n \times \cdots \times LMR\ L_1$$

AMAT

AMAT = Hit time + Miss rate × Miss penalty

= L1 hit time + Local Miss Rate L1 × (

L2 hit time + Local Miss Rate L2 × (

…))

Example

	Access Time	Miss Rate
L1 cache	1 ns	10%
L2 cache	5 ns	1%
L3 cache	10 ns	0.2%
Main memory	50 ns	0%

No cache AMAT = 50 ns

L1 cache AMAT = $1 + 0.1 \times 50 = 6$ ns

L1-2 caches AMAT = $1 + 0.1 \times (5 + 0.01 \times 50) = 1.55$ ns

L1-3 caches AMAT = $1 + 0.1 \times (5 + 0.01 \times (10 + 0.002 \times 50)) = 1.5101$ns

Thanks