Topic V35

Multi-Level Caches

Reading: (Section 5.4)

Multilevel Caches

Primary cache attached to CPU

- Small, but fast

Level-2 cache services references that miss primary cache

- Larger, slower, but still faster than main memory

Main memory services L2 cache misses

Multilevel Cache Considerations

Primary cache

- Focus on minimal hit time

L2 cache

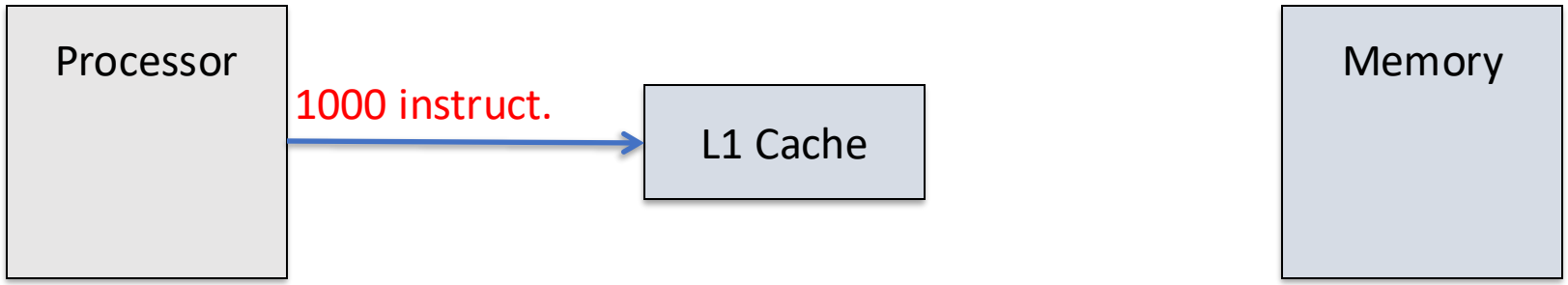
- Focus on low miss rate to avoid main memory access
- Hit time has less overall impact

Results

- L1 cache usually smaller than alternate design with a single cache

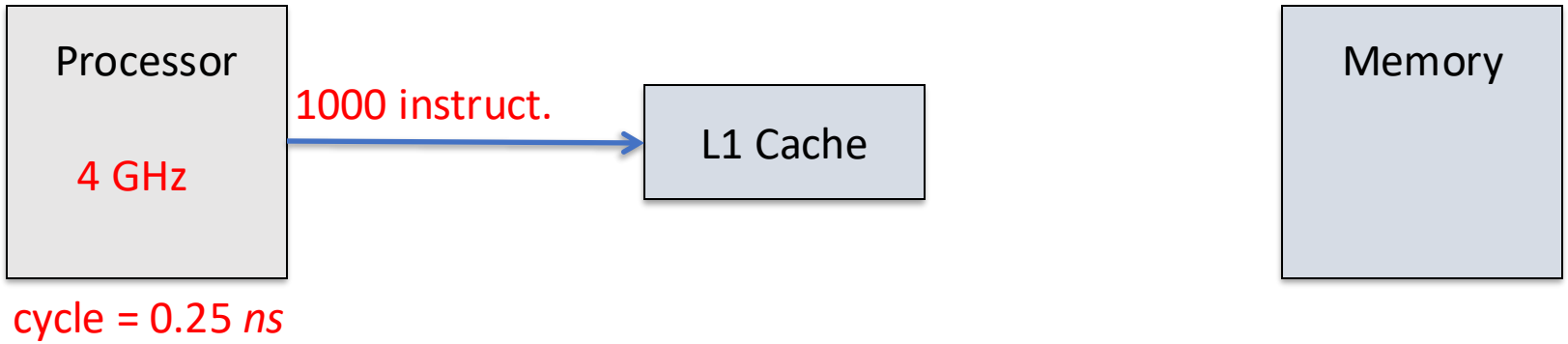
Multilevel Cache (example)

A processor with a clock rate of 4 GHz has a base CPI of 1.0 when all references hit in L1. The main-memory access time is 100 ns. The L1 miss rate is 2%.
How much faster will the processor be if we add an L2 cache that has a 5 ns access time for either a hit or a miss and is large enough to reduce the miss rate to main memory to 0.5% of the accesses to L1?
Assume that all accesses are instruction fetches.



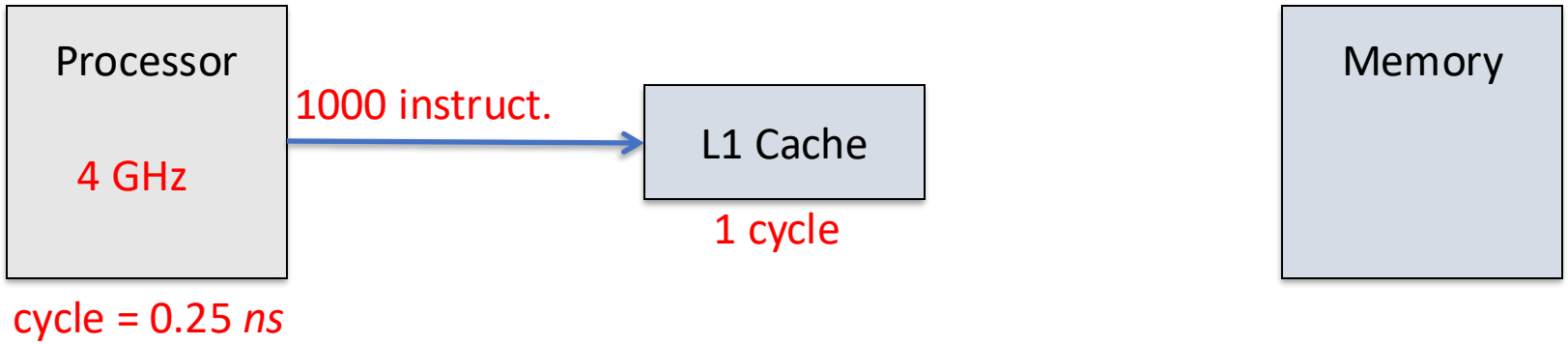
Multilevel Cache (example)

A processor with a clock rate of 4 GHz has a base CPI of 1.0 when all references hit in L1. The main-memory access time is 100 ns. The L1 miss rate is 2%.
How much faster will the processor be if we add an L2 cache that has a 5 ns access time for either a hit or a miss and is large enough to reduce the miss rate to main memory to 0.5% of the accesses to L1?
Assume that all accesses are instruction fetches.



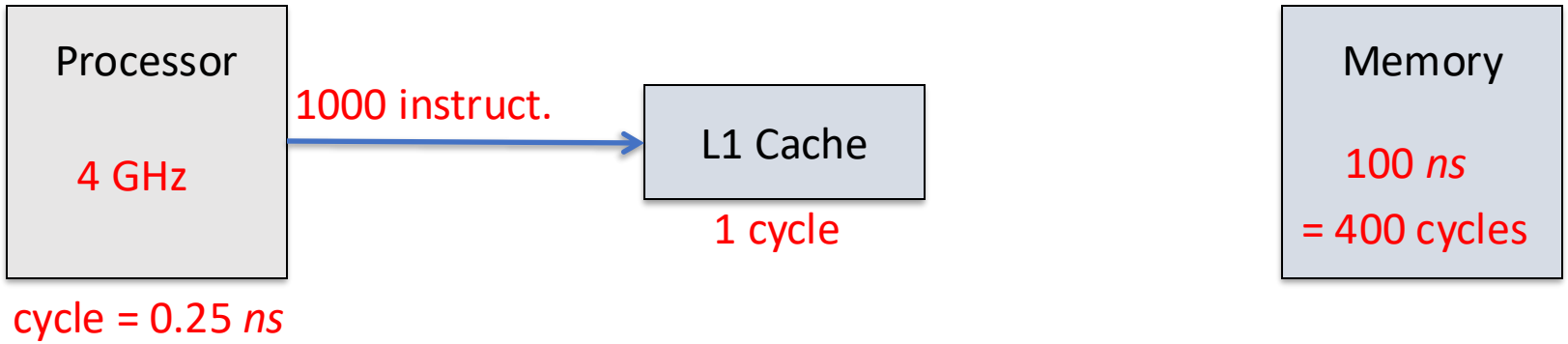
Multilevel Cache (example)

A processor with a clock rate of 4 GHz has a base CPI of 1.0 when all references hit in L1. The main-memory access time is 100 ns. The L1 miss rate is 2%.
How much faster will the processor be if we add an L2 cache that has a 5 ns access time for either a hit or a miss and is large enough to reduce the miss rate to main memory to 0.5% of the accesses to L1?
Assume that all accesses are instruction fetches.



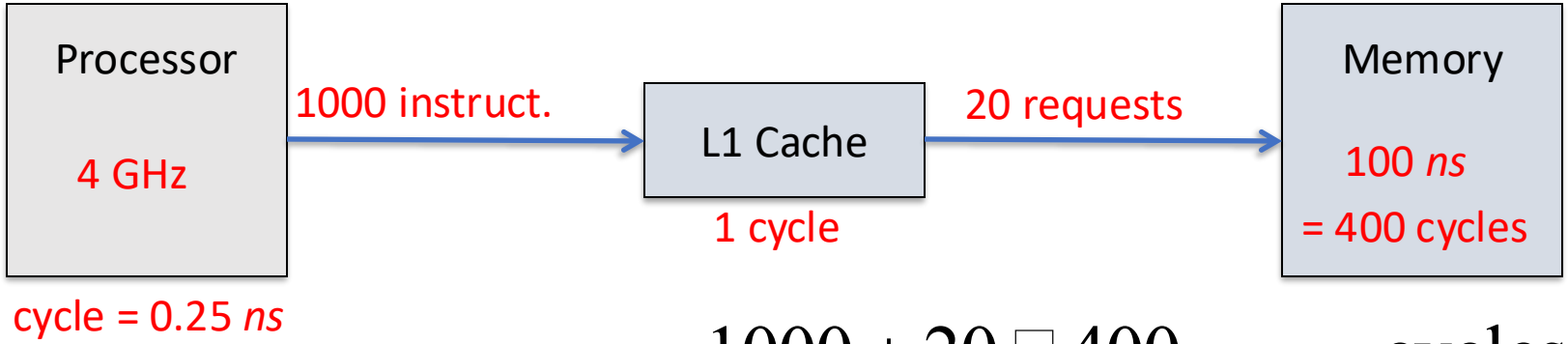
Multilevel Cache (example)

A processor with a clock rate of 4 GHz has a base CPI of 1.0 when all references hit in L1. The main-memory access time is 100 ns. The L1 miss rate is 2%.
How much faster will the processor be if we add an L2 cache that has a 5 ns access time for either a hit or a miss and is large enough to reduce the miss rate to main memory to 0.5% of the accesses to L1?
Assume that all accesses are instruction fetches.



Multilevel Cache (example)

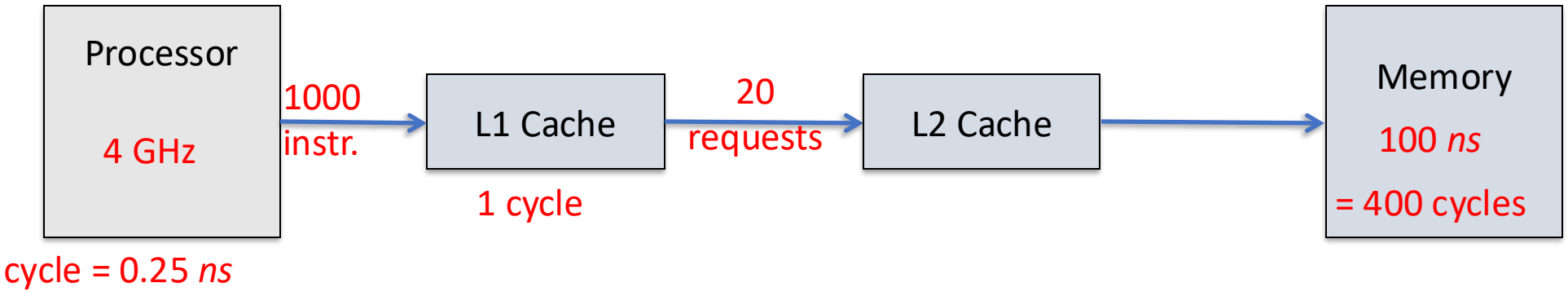
A processor with a clock rate of 4 GHz has a base CPI of 1.0 when all references hit in L1. The main-memory access time is 100 ns. The L1 miss rate is 2%.
How much faster will the processor be if we add an L2 cache that has a 5 ns access time for either a hit or a miss and is large enough to reduce the miss rate to main memory to 0.5% of the accesses to L1?
Assume that all accesses are instruction fetches.



$$\mathrm{CPI}_{\text{no L2}} = \frac{1000 + 20 \square 400}{1000} = 9 \frac{\text{cycles}}{\text{instruction}}$$

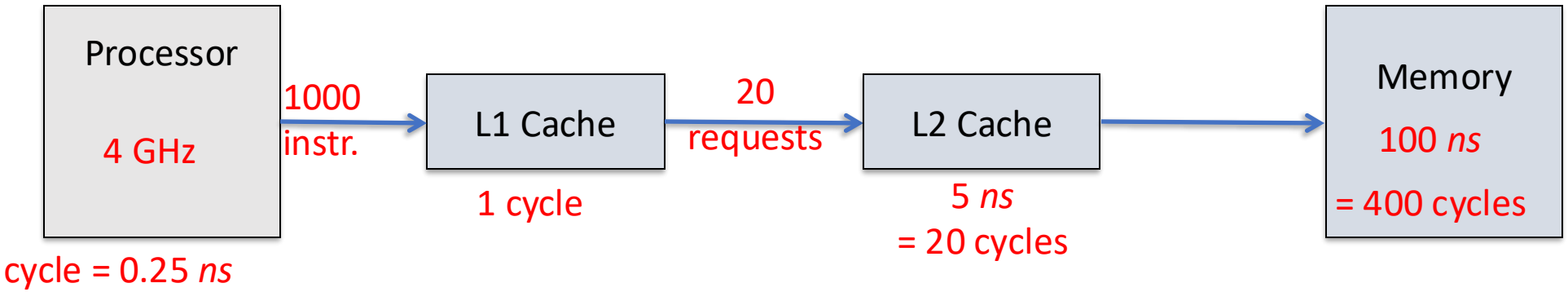
Multilevel Cache (example)

A processor with a clock rate of 4 GHz has a base CPI of 1.0 when all references hit in L1. The main-memory access time is 100 ns. The L1 miss rate is 2%.
How much faster will the processor be if we add an L2 cache that has a 5 ns access time for either a hit or a miss and is large enough to reduce the miss rate to main memory to 0.5% of the accesses to L1?
Assume that all accesses are instruction fetches.



Multilevel Cache (example)

A processor with a clock rate of 4 GHz has a base CPI of 1.0 when all references hit in L1. The main-memory access time is 100 ns. The L1 miss rate is 2%.
How much faster will the processor be if we add an L2 cache that has a 5 ns access time for either a hit or a miss and is large enough to reduce the miss rate to main memory to 0.5% of the accesses to L1?
Assume that all accesses are instruction fetches.



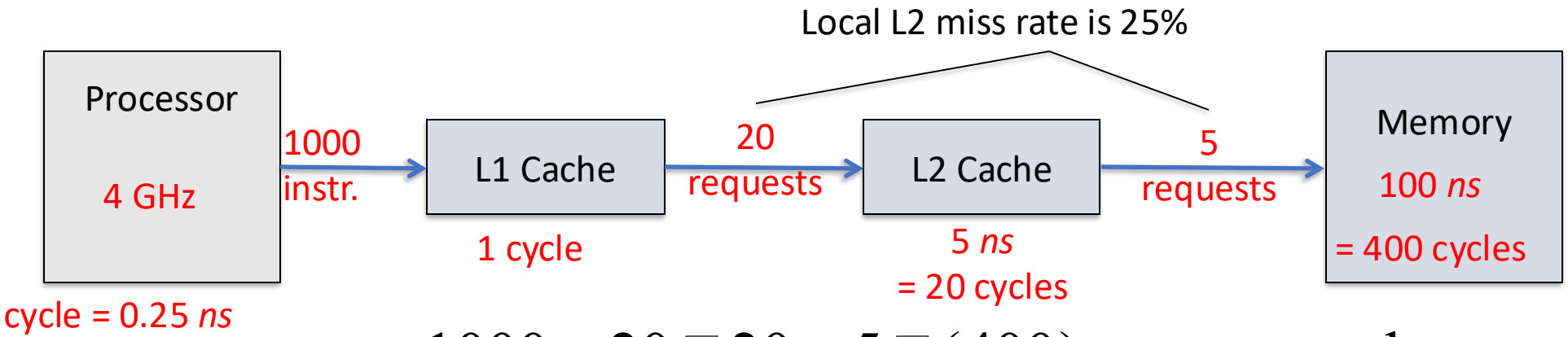
Multilevel Cache (example)

A processor with a [illegible] references hit in L1. The main-mem[illegible] How much faster wi[illegible] a 5 ns access time for either a hit [illegible] s rate to main memory to 0.5% of [illegible] Assume that all acce[illegible]

$$CPI_{\text{no L2}} = \frac{1000 + 20 \times 400}{1000} = 9 \frac{\text{cycles}}{\text{instruction}}$$

Processor with L2 is $\frac{9}{3.4} = 2.6$ times faster

Global miss rate of L2



$$CPI_{L2} = \frac{1000 + 20 \times 20 + 5 \times (400)}{1000} = 3.4 \frac{\text{cycles}}{\text{instruction}}$$

Interactions With Advanced CPUs

- Out-of-order CPUs can execute instructions during cache miss
  - Pending store stays in load/store unit
  - Dependent instructions wait in reservation stations
    - Independent instructions continue
- Effect of miss depends on program data flow
  - Much harder to analyze

Interactions With Software

Misses depend on memory access patterns

- Algorithm behavior
- Compiler optimization for memory access