Measuring Cache Performance

One level of cache

vs

Two levels of cache
Parameters needed

• One level:
  • Miss rate for instructions
  • Miss rate for data
  • Ratio of data access (load/store) instructions
  • CPI
  • Clock speed
  • Time penalty for miss
Example – single cache

• One level:
  • Miss rate for instructions  2%
  • Miss rate for data  3%
  • Ratio of data access (load/store) instructions  30%
  • CPI  1.5
  • Clock speed  2 GigaHertz
  • Time penalty for miss  100 nanoseconds
Compute effective CPI with cache misses
How much better would “perfect cache” be

• Data miss rate per instruction: 0.30 x 0.03 = 0.009
• Total miss rate per instruction: 0.02 + 0.009 = 0.029
• 2 GHz = 2 x 10^9 cycles/sec or 0.5 ns per cycle
• 100 ns penalty for miss = 200 cycles
• Miss penalty per instruction = 200 cycles x 0.029 = 5.8 cycles

• Total CPI = 1.5 + 5.8 = 7.3

• Perfect cache means no misses, so 1.5 CPI, better by 7.3/1.5 = 4.87
Add second level cache

• Parameters for 1st level the same.
  • Total miss rate per instruction: 0.02 + 0.009 = 0.029

• Parameters for second level:
  • Miss rate: 0.25% (2nd level cache much larger)
  • Response time = 5 ns = 10 cycles
  • Response time from main memory still 100 ns = 200 cycles
Analysis

• Miss penalty for 1\textsuperscript{st} to 2\textsuperscript{nd} level cache:
  miss rate per instruction x 10 cycles = 0.029 x 10 = 0.29 cycles

• Miss penalty for 2\textsuperscript{nd} level to memory:
  miss rate 2\textsuperscript{nd} level x 200 cycles = 0.0025 x 200 = 0.5 cycles

• Total CPI = 1.5 + 0.29 + 0.5 = 2.29

• Speedup due to 2\textsuperscript{nd} level cache: 7.3/2.29 = 3.19
Other factors

• Out-of-order execution can sometimes hide miss penalties, especially for 1\textsuperscript{st} level cache (usually not for second level)

• 1\textsuperscript{st} level cache should focus on fast hit time
  • Much smaller and often smaller block size
  • This can allow faster cycle time or fewer pipeline stages

• 2\textsuperscript{nd} level cache focus on minimizing miss rate
  • Much larger
  • Larger blocks
  • Higher level of associativity