Computer Organization & Architecture

Chapter 8 – Cache Performance

Zhang Yang 张杨

cszyang@scut.edu.cn

Autumn 2025

Contents of this lecture

- 8.7 Performance Consideration
  - Hit Rate and Miss Penalty
  - Mutilevel Cache
  - Caches on the Processor Chip

Hit Rate and Miss Penalty (1)

- Hit Rate
  - Hit: A successful access to data in a cache is called a hit.
  - Hit Rate: The number of hits stated as a fraction of all attempted accesses is called the hit rate.
  - High hit rates, well over 0.9, are essential for high-performance computers.

Hit Rate and Miss Penalty (2)

- Miss Penalty
  - Miss Rate: The number of misses stated as a fraction of attempted accesses.
  - Miss Penalty: The total access time seen by the processor when a miss occurs as the miss penalty.
  - The miss penalty consists almost entirely of the time to access a block of data in the main memory, in a system with one level of cache.

Hit Rate and Miss Penalty (3)

- Impact of the cache on the overall performance of the computer
  - The average access time experienced by the processor

    $t_{avg} = hC + (1 - h)M$

    - h: hit rate
    - C: the time to access information in the cache
    - M: miss penalty

Hit Rate and Miss Penalty (4)

- Example 8.1: Consider a computer that has the following parameters.
  - Access times to the cache and the main memory are τ and 10τ , respectively.
  - When a cache miss occurs, a block of 8 words is transferred from the main memory to the cache. It takes 10τ to transfer the first word of the block, and the remaining 7 words are transferred at the rate of one word every τ seconds. The miss penalty also includes a delay of τ for the initial access to the cache, which misses, and another delay of τ to transfer the word to the processor after the block is loaded into the cache (assuming no load-through). Thus, the miss penalty in this computer is given by: $M= \tau + 10\tau + 7\tau + \tau = 19\tau$

Hit Rate and Miss Penalty (5)

- Example 8.1 (ctd.)
  - Assume that 30 percent of the instructions in a typical program perform a Read or a Write operation, which means that there are 130 memory accesses for every 100 instructions executed. Assume that the hit rates in the cache are 0.95 for instructions and 0.9 for data. Assume further that the miss penalty is the same for both read and write accesses. Then, a rough estimate of the improvement in memory performance that results from using the cache can be obtained as follows:

$$\frac{\text{Time without cache}}{\text{Time with cache}} = \frac{130 \times 10\tau}{100(0.95\tau + 0.05 \times 19\tau) + 30(0.9\tau + 0.1 \times 19\tau)} = 4.7$$

Hit Rate and Miss Penalty (6)

- Example 8.1(ctd.)
  - An estimate of the increase in memory access time caused by misses in the cache is given by:

$$\frac{\text{Time for real cache}}{\text{Time for ideal cache}} = \frac{100(0.95\tau + 0.05 \times 19\tau) + 30(0.9\tau + 0.1 \times 19\tau)}{130\tau} = 2.1$$

Contents of this lecture

- 8.7 Performance Consideration
  - Hit Rate and Miss Penalty
  - Mutilevel Cache
  - Caches on the Processor Chip

Multi-level Cache (1)

- Most of today's systems employ multilevel cache hierarchies.
- The levels of cache form their own small memory hierarchy.
- Current day processor uses
  - Level-1 cache (8KB to 128KB) is situated on the processor.—Access time is typically about 4ns.
  - Level-2 cache (256KB to several MB) located internal or external to the processor. –Access time is usually around 15-20ns.

Multi-level Cache (2)

- Current day processor uses (ctd.)
  - Level-3 cache: As the latency difference between main memory and the fastest cache has become larger, some processors have begun to utilize as many as 3 levels of on-chip cache.
  - Level-4 cache: But by the 2010s some of the highest-performance designs returned to having large off-chip caches, which is often implemented in eDRAM and mounted on a multi-chip module, as a fourth cache level.

Multi-level Cache (3)

- Instruction and Data Caches
  - A unified or integrated cache is one where both instructions and data are cached.
    - Causing excessive cache misses.
  - Many modern systems employ separate caches for data and instructions. This is called a Harvard cache.
  - Advantages
    - Allows accesses to be less random and more clustered.
    - Less access time than unified cache (typically larger).

Contents of this lecture

- 8.7 Performance Consideration
  - Hit Rate and Miss Penalty
  - Mutilevel Cache
  - Caches on the Processor Chip

Caches on the Processor Chip (1)

- Please see 8.7.2 in textbook
- Most processor chips include at least one L1 cache. Often there are two separate L1 caches, one for instructions and another for data.
- In high-performance processors, two levels of caches are normally used, separate L1 caches for instructions and data and a larger L2 cache.

Caches on the Processor Chip (2)

- The average access time experienced by the processor in such a system is:

$$t_{avg} = h_1 C_1 + (1 - h_1)(h_2 C_2 + (1 - h_2)M)$$

where

h_1 is the hit rate in the L1 caches.

h_2 is the hit rate in the L2 cache.

C_1 is the time to access information in the L1 caches.

C_2 is the miss penalty to transfer information from the L2 cache to an L1 cache.

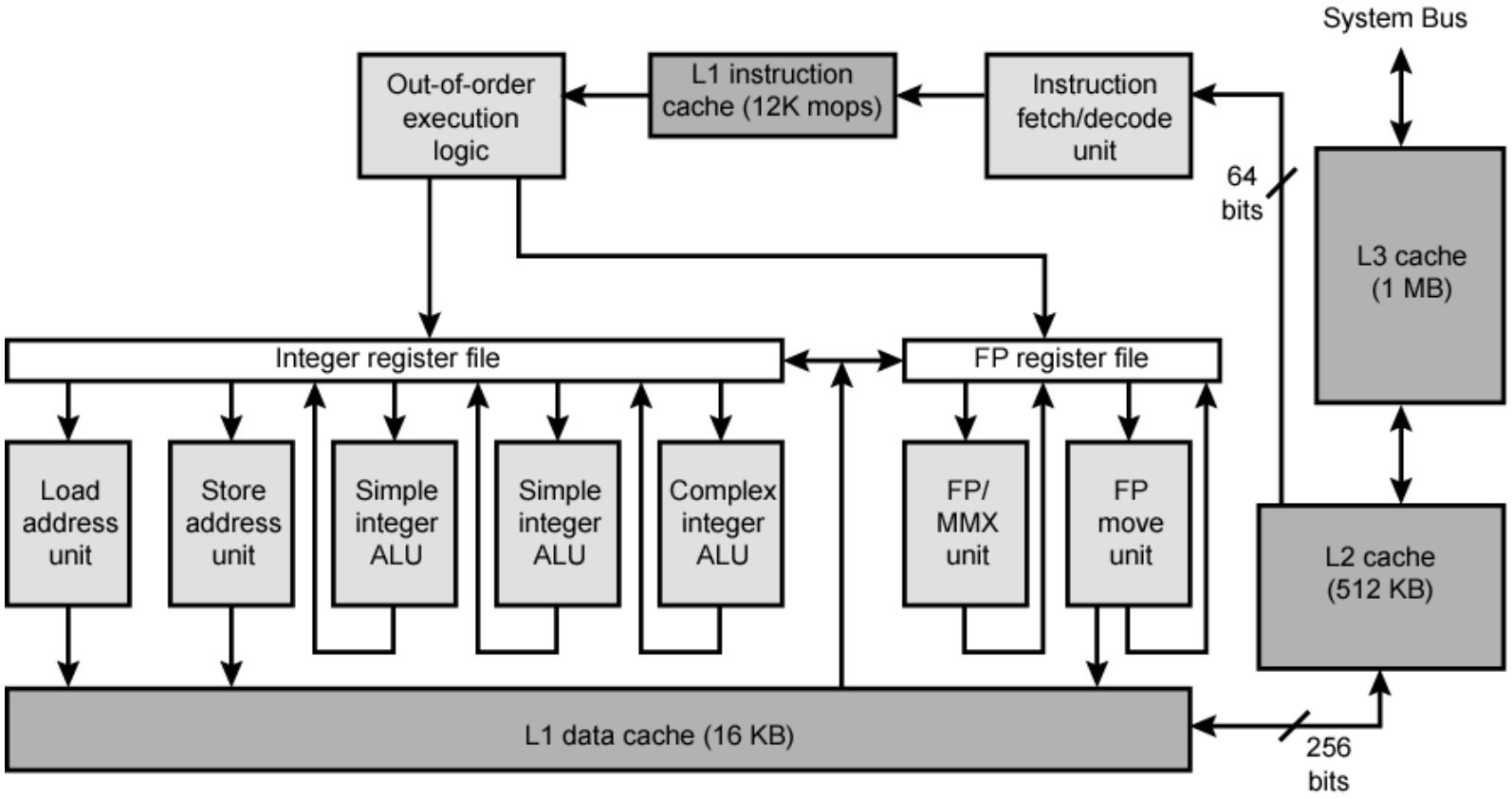
M is the miss penalty to transfer information from the main memory to the L2 cache.

Intel CPU Cache (1)

- 80386 – no on chip cache
- 80486 – 8K using 16-byte lines and four-way set associative organization
- Pentium (all versions) – two on chip L1 caches
  - Data & instructions
- Pentium III – L3 cache added off chip
- Pentium 4
  - L1 caches: four way set associative
  - L2 cache
    - Feeding both L1 caches
    - 8-way set associative
  - L3 cache on chip

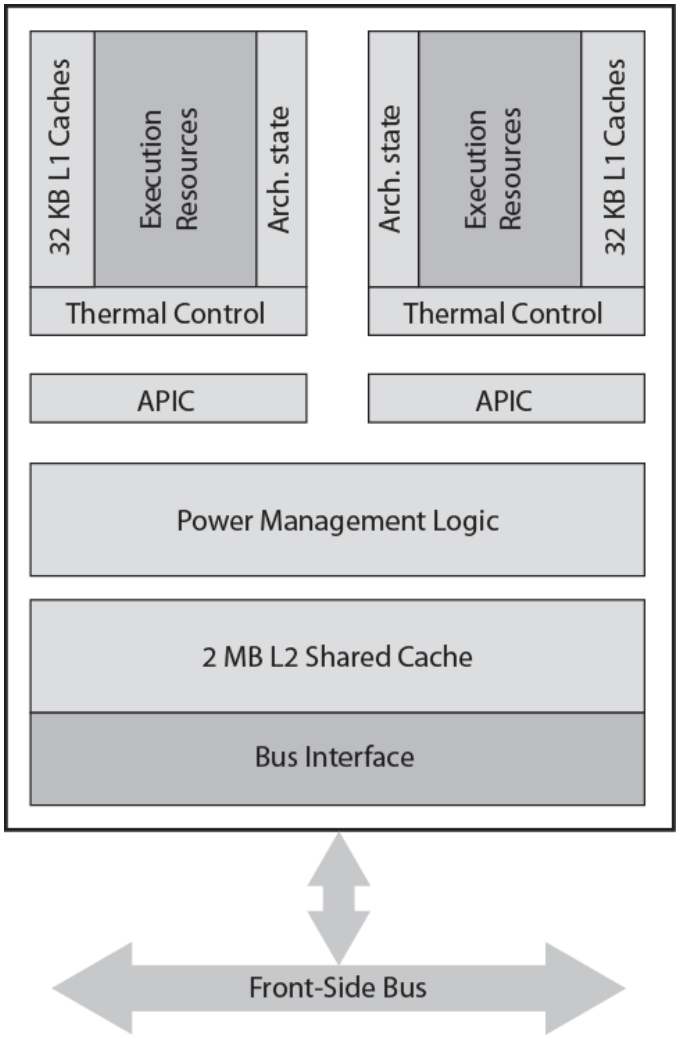
Intel CPU Cache (2)

- Pentium 4 Block Diagram



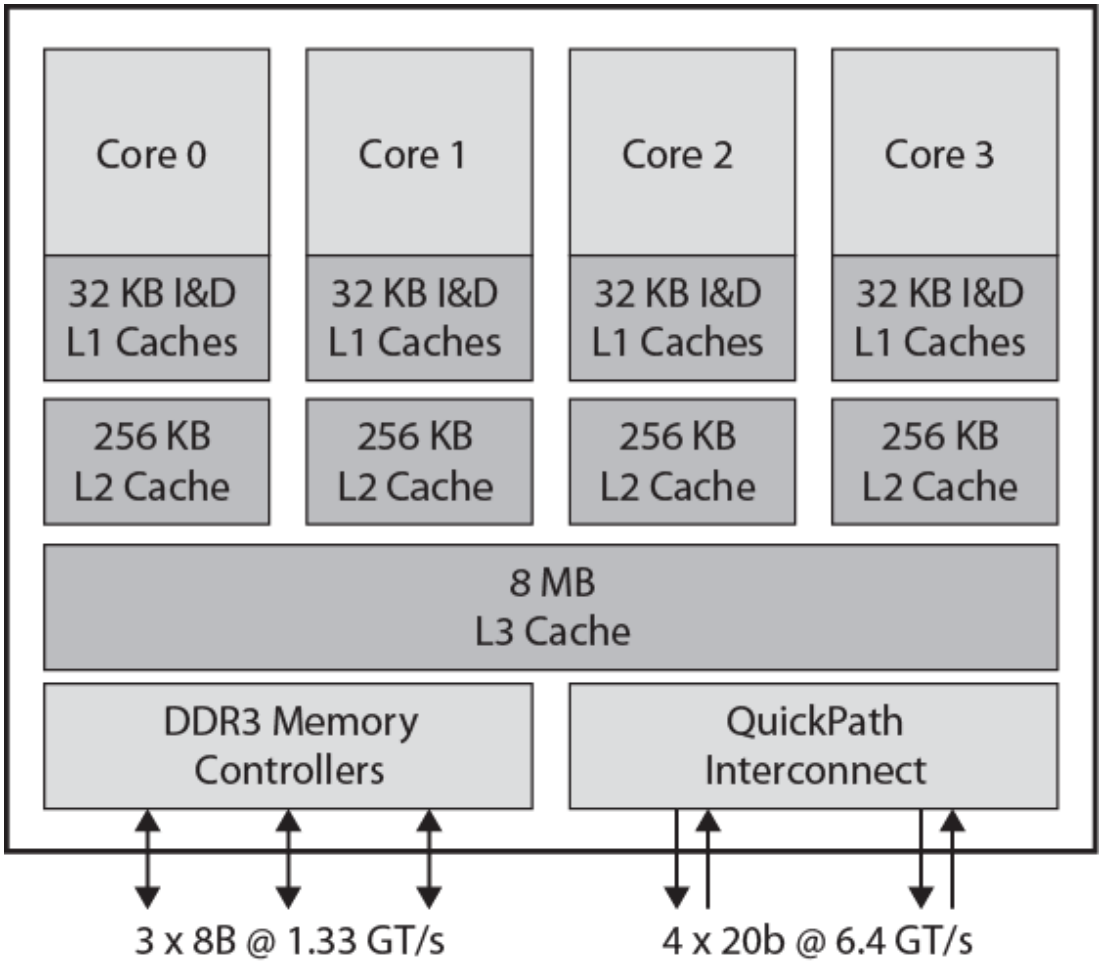
Intel CPU Cache (3)

- Intel Core Duo (2) Block Diagram
  - Dedicated L1 cache per core
    - 32KB instruction and 32KB data
  - 2MB shared L2 cache



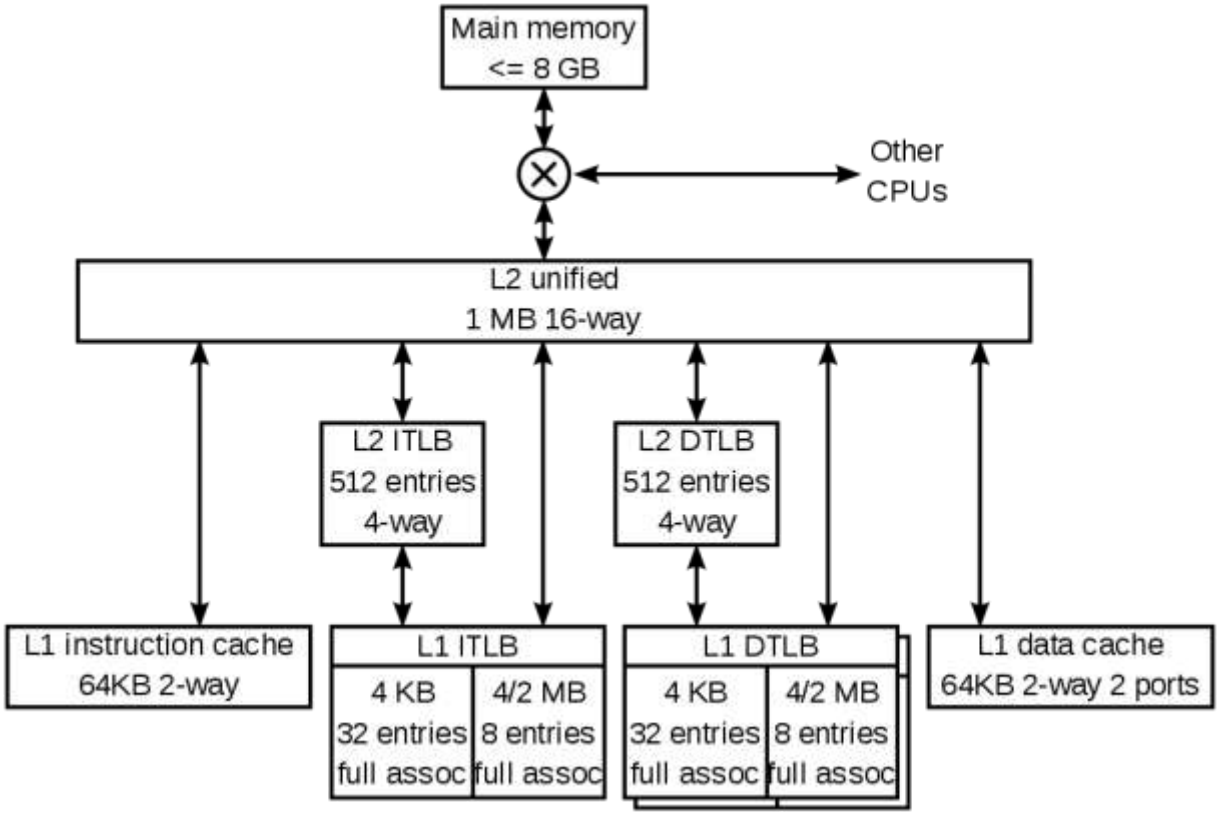
Intel CPU Cache (4)

- Intel Core i7 Block Diagram
  - Dedicated L2, shared L3 cache



AMD CPU Cache

- Cache hierarchy of the K8 core in the AMD Athlon 64 CPU



Summary

- 知识点：Multilevel Cache
  - Hit rate
  - Miss penalty
  - Average access time of single-level cache

$t_{avg}=hC+(1-h)M$

  - Average access time of two-level cache

$$t_{avg} = h_1C_1 + (1 - h_1)(h_2C_2 + (1 - h_2)M)$$