Computer Organization & Architecture

3-5 Memory Hierarchy

Wang Guohua

School of Software Engineering

Contents of this lecture

- Why Have a Memory Hierarchy?
- An Example of Memory Hierarchy
- Fundamental Idea of the Memory Hierarchy
- How Does the Memory Hierarchy Work?
- Four Questions for the Memory Hierarchy Designers

Why Have a Memory Hierarchy? (1)

- Ideal Memory
  - Zero access time (latency)
  - Infinite capacity
  - Zero cost

Why Have a Memory Hierarchy? (2)

- The Problems
  - Ideal memory's requirements oppose each other
  - Bigger is slower
    - Bigger → Takes longer to determine the location
  - Faster is more expensive
    - Memory technology: SRAM vs. DRAM

Why Have a Memory Hierarchy? (3)

- The Problems (ctd.)
  - Memory Technology

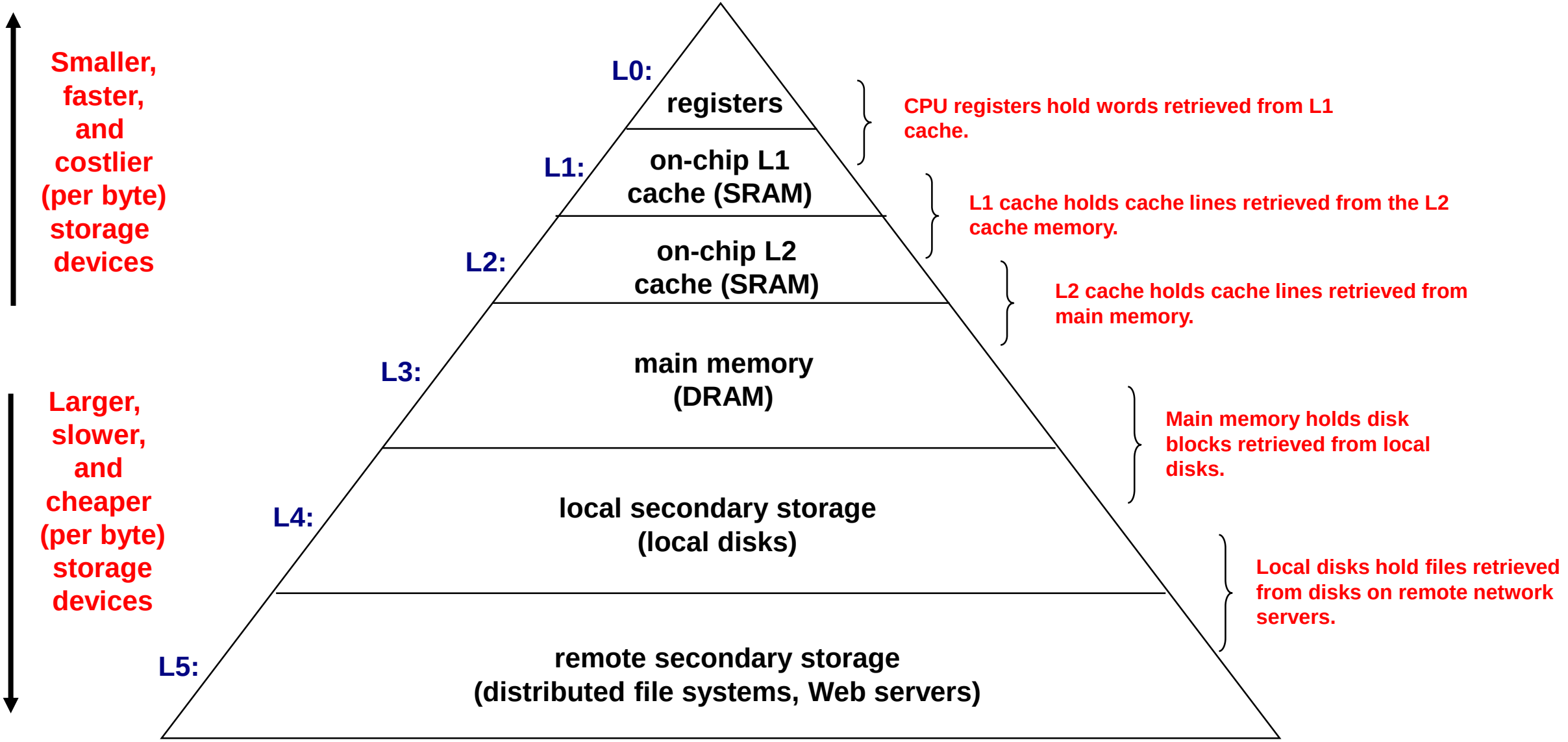
Memory Technology	Typical Access Time	$ per GB in 2012
SRAM	0.5 – 2.5 ns	$500 - $1000
DRAM	50 – 70 ns	$10 - $20
Flash Memory	5,000-50,000ns	$0.75-$1.00
Magnetic Disk	5,000,000 – 20,000,000 ns (5 – 20 ms)	$0.05 - $0.10

Why Have a Memory Hierarchy? (4)

- Summary
  - We want fast, large, cheap memory.
  - But we cannot achieve all with a single level of memory
  - Idea
    - To create an illusion of “fast and large” main memory
    - Have multiple levels of storage (progressively bigger and slower as the levels are farther from the processor) and ensure most of the data the processor needs is kept in the fast(er) level(s)

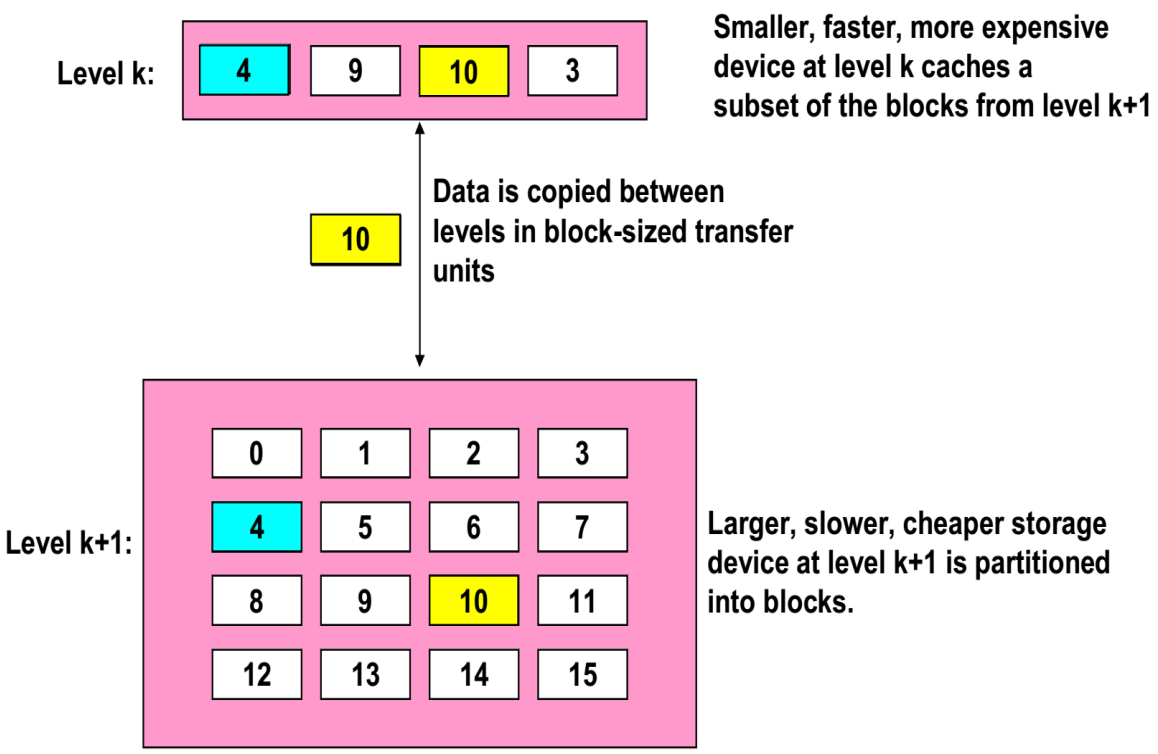
An Example of Memory Hierarchy

- Another Figure of The Memory Hierarchy



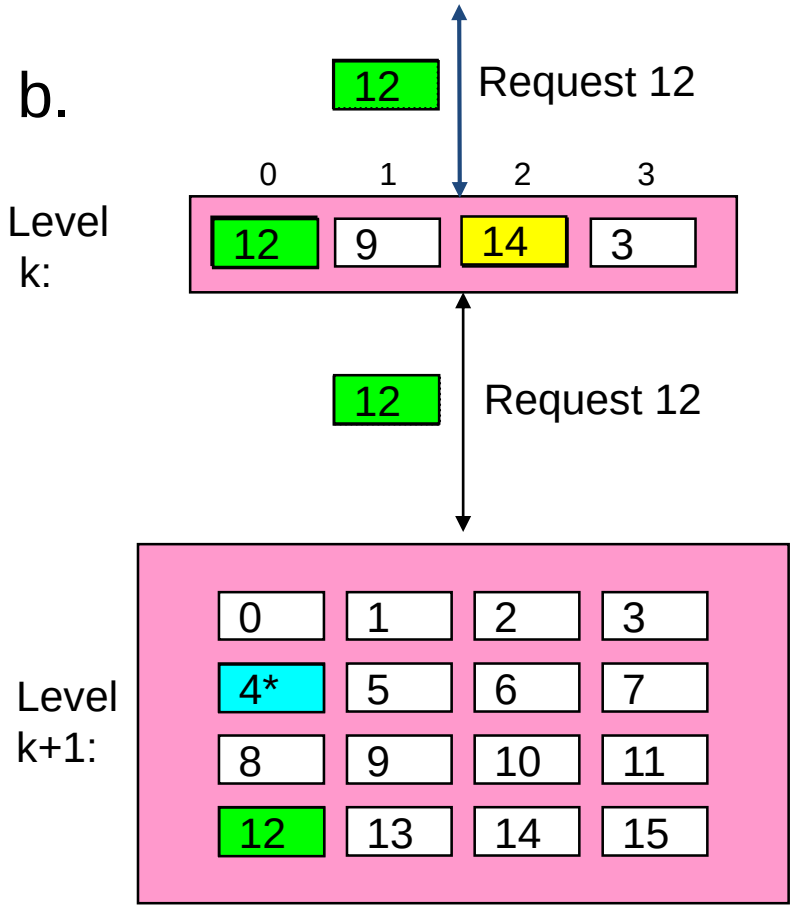
Fundamental Idea of The Memory Hierarchy (1)

- Buffer
  - A smaller, faster storage device that acts as a staging area for a subset of the data in a larger, slower device.
  - For each k, the faster, smaller device at level k serves as a buffer for the larger, slower device at level k+1.
  - All data is stored at the lowest level.
  - Data is copied between two adjacent levels.



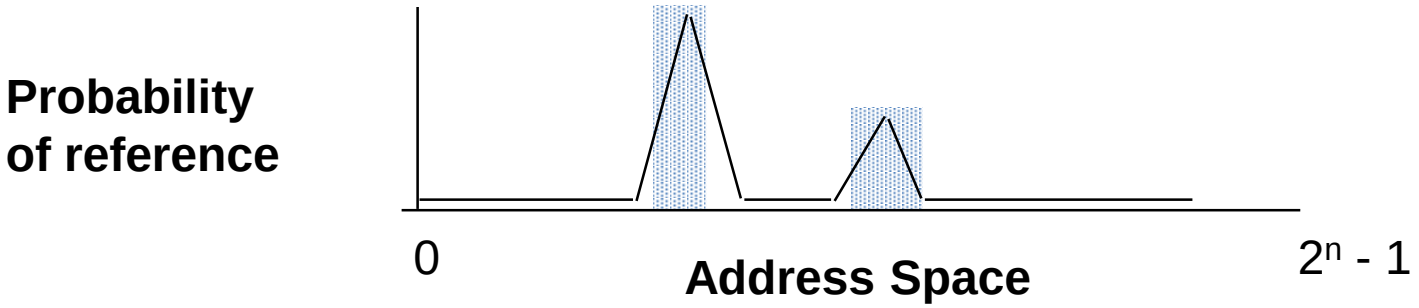
Fundamental Idea of The Memory Hierarchy (2)

- General Buffering Concepts
  - Program needs object d, which is stored in some block b.
  - Hit
    - Program finds b in the buffer at level k.
      - E.g., block 14.
  - Miss
    - b is not at level k, so level k buffer must fetch it from level k+1.
      - E.g., block 12.



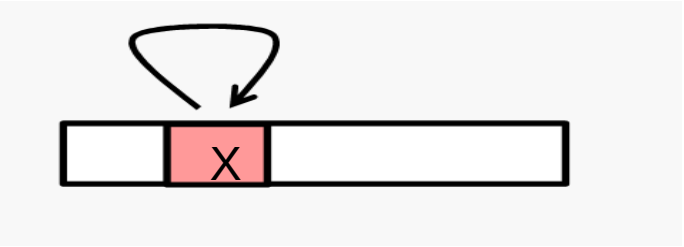
How Does The Memory Hierarchy Work? (1)

- By taking advantage of the Principle of Locality
  - Well-written programs tend to exhibit good locality.
  - Programs access a relatively small portion of the address space at any instant of time.

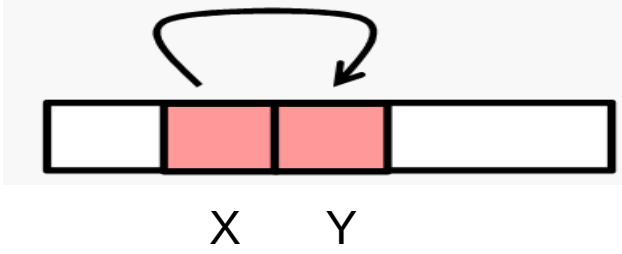


How Does Memory Hierarchy Work? (2)

- The Principle of Locality
  - Temporal Locality – locality in time
    - If item was referenced recently, it will likely be referenced again soon.



  - Spatial locality – locality in space
    - If an item is referenced, items near it might be referenced soon



How Does Memory Hierarchy Work? (3)

- In the memory hierarchy
  - Temporal Locality (Locality in Time):

    => Keep most recently accessed data items closer to the processor
  - Spatial Locality (Locality in Space):

    => Move blocks consists of contiguous words to the upper levels

How Does Memory Hierarchy Work? (4)

- Summary
  - Programs tend to access the data at level k more often than they access the data at level k+1.
  - Thus, the storage at level k+1 can be slower, and thus larger and cheaper per bit.
  - Net effect: A large pool of memory that costs as much as the cheap storage near the bottom, but that serves data to programs at the rate of the fast storage near the top.

Four Questions for Memory Hierarchy Designers

- Q1: Where can a block be placed in the upper level? *(Block Placement)*
- Q2: How is a block found if it is in the upper level? *(Block Identification)*
- Q3: Which block should be replaced on a miss? *(Block Replacement)*
- Q4: What happens on a write? *(Write Policy)*

Quiz

1. A memory hierarchy _____.

   A. limits programs' size but allows them to execute more quickly

   存储器层次结构不会限制程序大小

   B. is a way of structuring memory allocation decisions

   存储器层次结构不是一种存储器分配方式

   C. takes advantage of the speed of SRAM and the capacity of disk

   存储器层次结构利用了SRAM的速度和磁盘的容量，给用户造成存储器速度快、容量大的假象

   D. makes programs execute more slowly but allows them to be bigger

   存储器层次结构使得程序执行得更快