Chapter 08: The Memory System

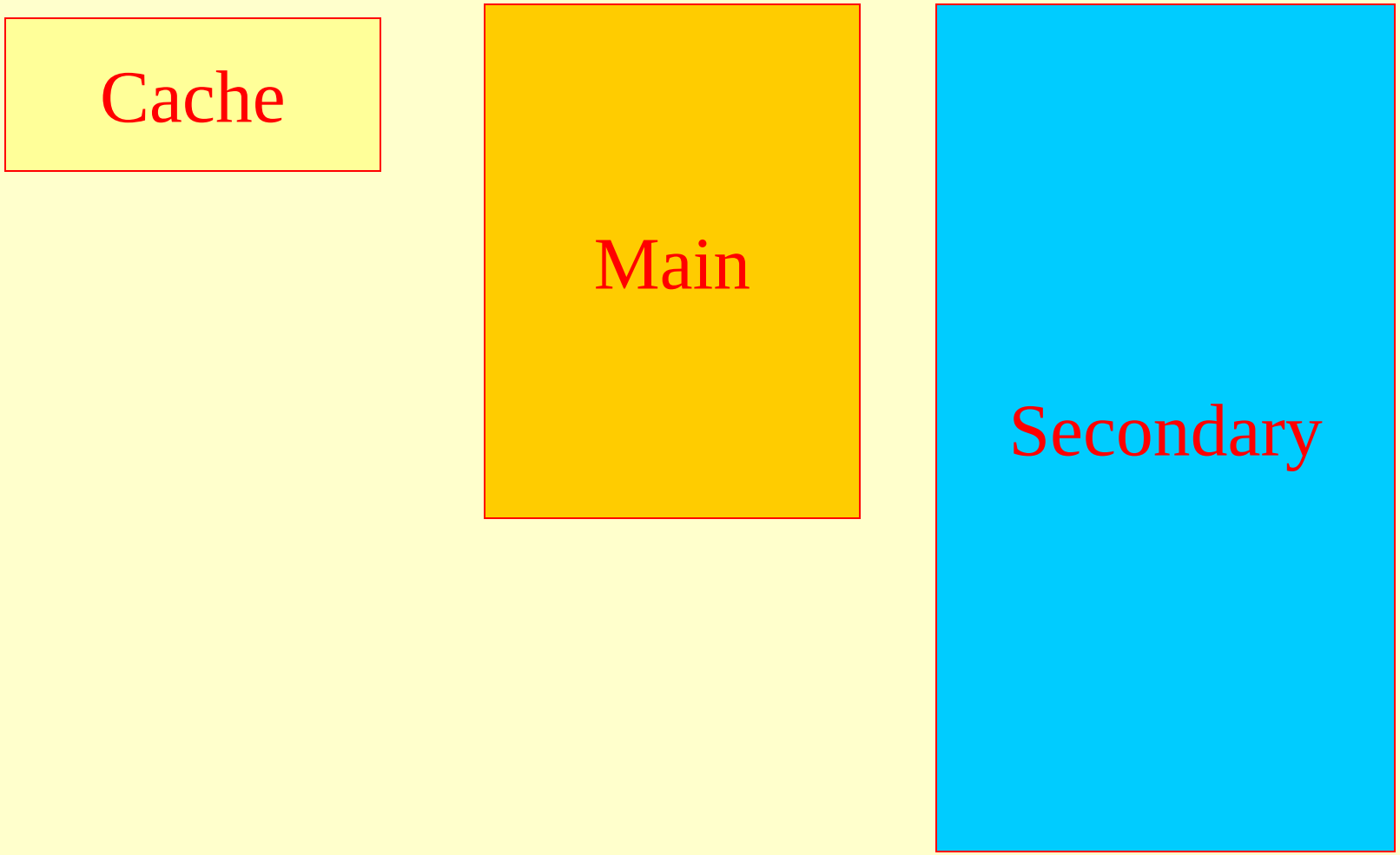
Lesson 02:
Memory Hierarchy

Objective

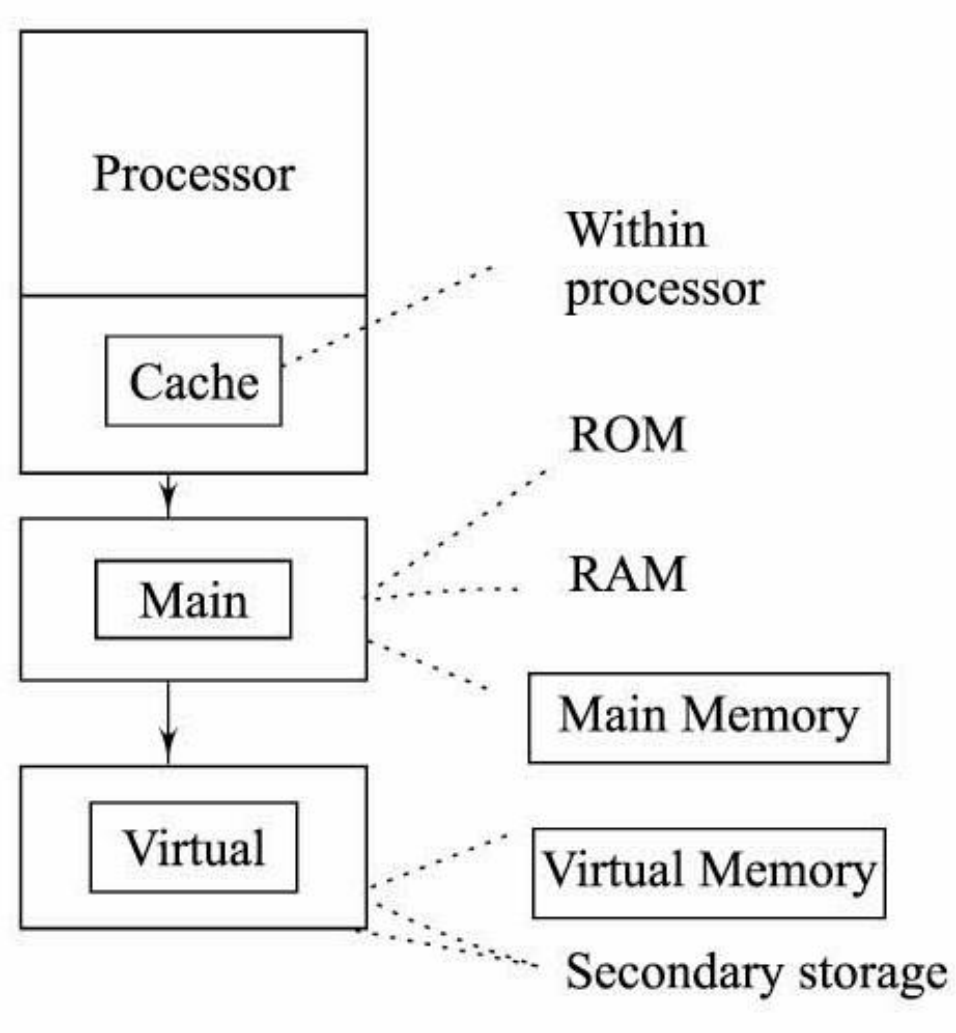
- Learn three levels of hierarchy of cache, RAM/ROM and secondary storage

Memory System Hierarchy

Three level Hierarchy



Memory System Three level Hierarchy



Memory System construction as a hierarchy

- Primary reason— the cost per bit of a memory technology is generally proportional to the speed of the technology

Memory hierarchy

- The levels closest to the processor, such as the cache, contain a relatively small amount of memory that is implemented in a fast memory technology to give a low access time

Access from Memory System Hierarchy

Demand-based system access

- In general, not possible to predict which memory locations will be accessed (referenced by processor instructions) most frequently
- Use of a demand-based system for determining which data to keep in the top levels of the hierarchy

Access Method

- When a memory request is sent to the hierarchy, the top level (cache) is checked to see if it contains the address
- If it does, then the request completes
- If not, the next lower level checked
- Process repeating until either the data is found or the bottom level of the hierarchy is reached, which is guaranteed to contain the data

Access Method

- If the top level in the hierarchy cannot handle a memory request, a block of sequential locations containing the referenced address is copied from the first level that contains the address into every level above that

Reasons for copying a block of sequential locations containing the address

1. First reason– that many storage technologies, such as page mode DRAMs and hard disks, allow multiple sequential words of data to be read or written in less time than an equal number of randomly located words
- It makes it faster to bring a multibyte block of data into the top levels of the hierarchy than to fetch each byte in the block from the lower levels of the hierarchy individually

Reasons for copying a block of sequential locations containing the address

2. Second– most programs display **locality of reference**

- Memory references that occur close together in time tend to have addresses that are close to each other, making it likely that the other addresses within a block will be referenced soon after the first reference to an address in the block

Average Access time and copied block sizes in Memory Hierarchy

Average Access Time

- As long as the probability that each address within the block will be referenced is sufficiently high, using multibyte blocks reduces the average access time
- Fetching the block takes less time than fetching each word within the block separately
- Different levels in the memory hierarchy will often have different block sizes, depending on the characteristics of the levels below them in the hierarchy

Block sizes in Different levels in the memory hierarchy

- Often have different block sizes
- Depend on the characteristics of the levels below them in the hierarchy

Example

- Caches tend to have block sizes of approximately 64 bytes copied from main memory
- Main memories generally have block sizes of around 4 kB copied from secondary memory

Example

- Time to fetch a large block of data from the virtual memory is only slightly longer than the time to fetch 1 byte
- Time to fetch a block of data into the cache from the main memory much closer to the time to fetch each byte individually

Top Level

- Caches are generally implemented using SRAM
- Most modern computers have at least two levels of cache memory in their memory hierarchy
- Caches have hardware to keep track of the addresses that are stored in them
- Tend to be relatively small, and have small block sizes, usually 32 to 128 bytes

Main Memory

- The main memory of a computer is generally constructed out of DRAM, relies on software to keep track of the addresses that are contained in it, and has a large block size, often several kilobytes

Virtual Memory

- Finally, virtual memory is usually implemented using disks and contains all of the data in the memory system
- Implementation by use of flash memory in case of mobile camera, PocketPC, phone, CD player

Summary

We learnt

- Three levels of hierarchy
- Cache top level
- RAM/ROM next level
- Secondary storage next lower level
- Different block sizes copied at higher level from lower level

End of Lesson 02 on

Memory Hierarchy