Computer Performance and Cache Memory

Muhammad Afzaal
m.afzaal@nu.edu.pk

Book Chapter

- “Computer Organization and Architecture”
- Author “William Stallings”
- 8th Edition
- Chapter 2
  - Section 2.2
- Chapter 4
  - Section 4.1
  - Section 4.2
  - Section 4.3

Designing for Performance

- Cost of systems continues to drop whereas performance and capacity continue to rise
- Today application programs are very complex and power hungry
  - Image processing
  - Speech recognition
  - Videoconferencing
  - Multimedia authoring
  - Simulation modeling
- We have to see how can we use the microprocessor in a way to utilize all it’s clock cycles

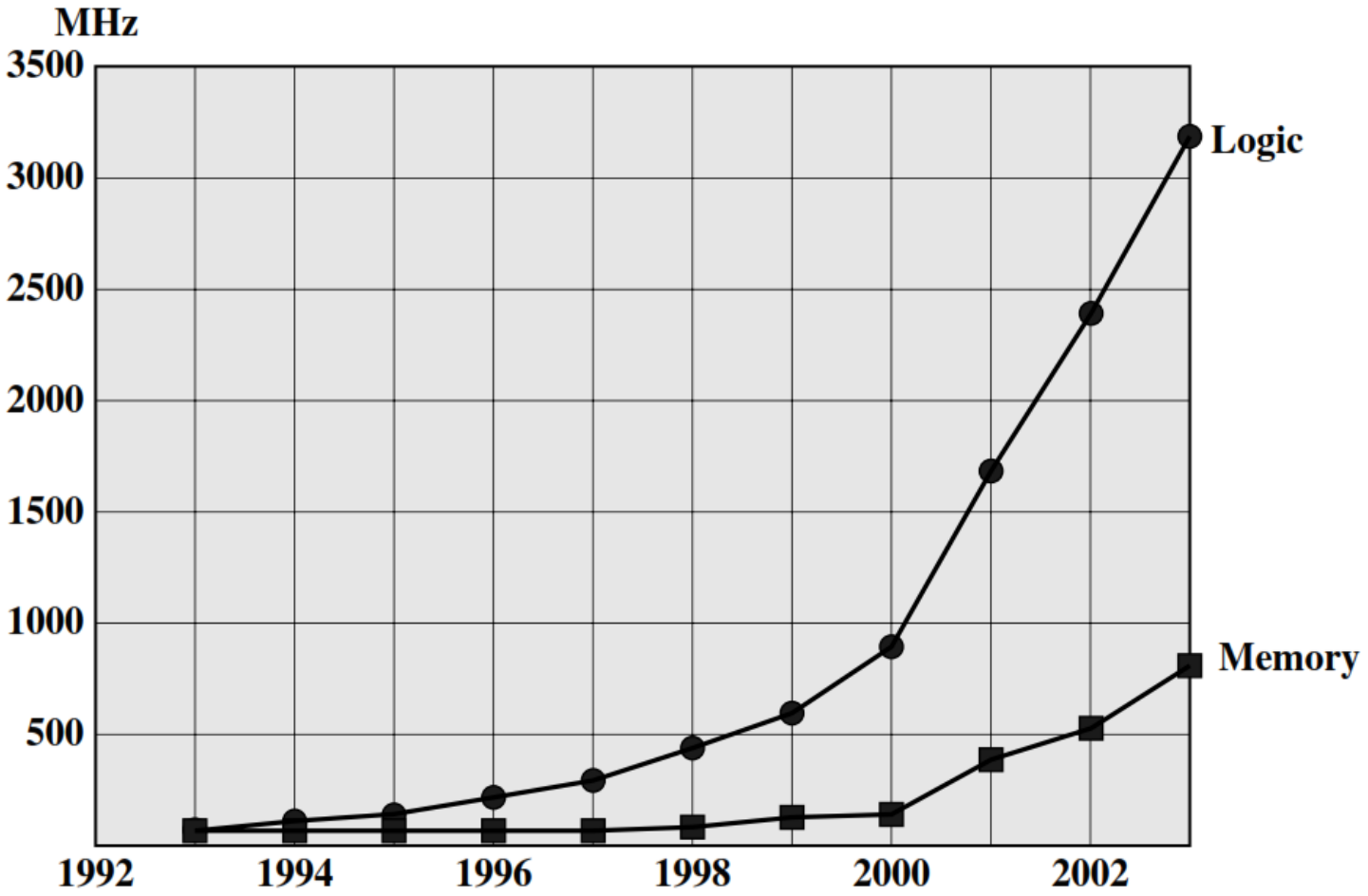
Microprocessor Speed

- Processor is not used at its full throttle always
- How to keep processor busy and get maximum out of it?
  - Branch Prediction
    - Processor predicts which branches or groups of instructions are likely to be executed next
  - Data Flow Analysis
    - Processor analyzes which instructions are dependent on each other
  - Speculative Execution
    - Processors execute instructions ahead of their actual appearance in the program execution

Performance Balance

- Processor power has increased rapidly
  - … what about the speed of other components of computer?
- Slow data transfer rate between main memory and processor
- If memory fails to keep pace with processor, the processor has to stall in a wait state

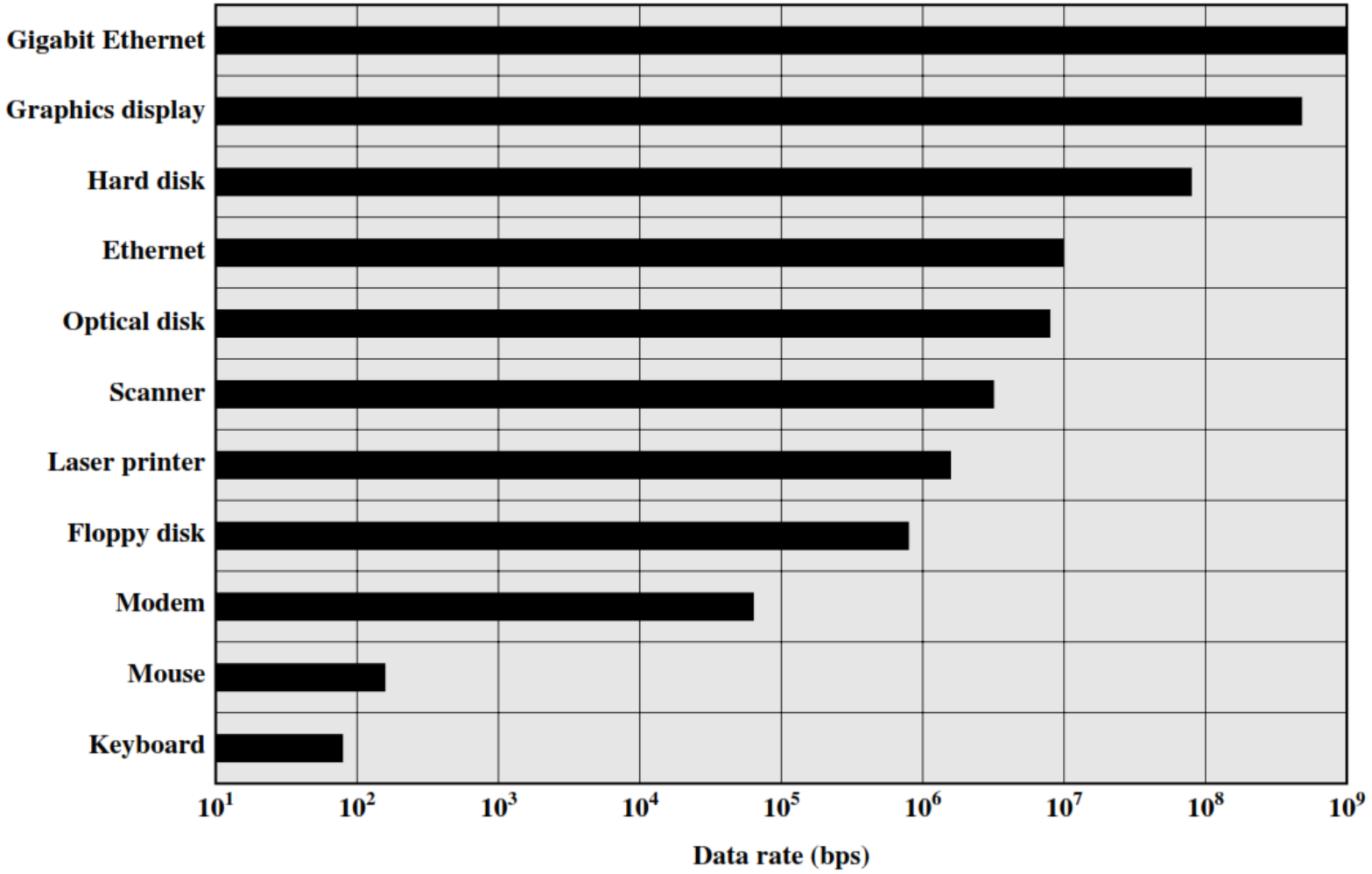
Processor and Memory Performance Gap



Some Tips to Increase Speed

- How to tackle this problem of speed between processor and main memory?
  - Increase data bus size
  - Include a cache in DRAM to make it more efficient
  - Reduce frequency of memory access
  - Increase the interconnect bandwidth between processor and memory

I/O Device Data Rate

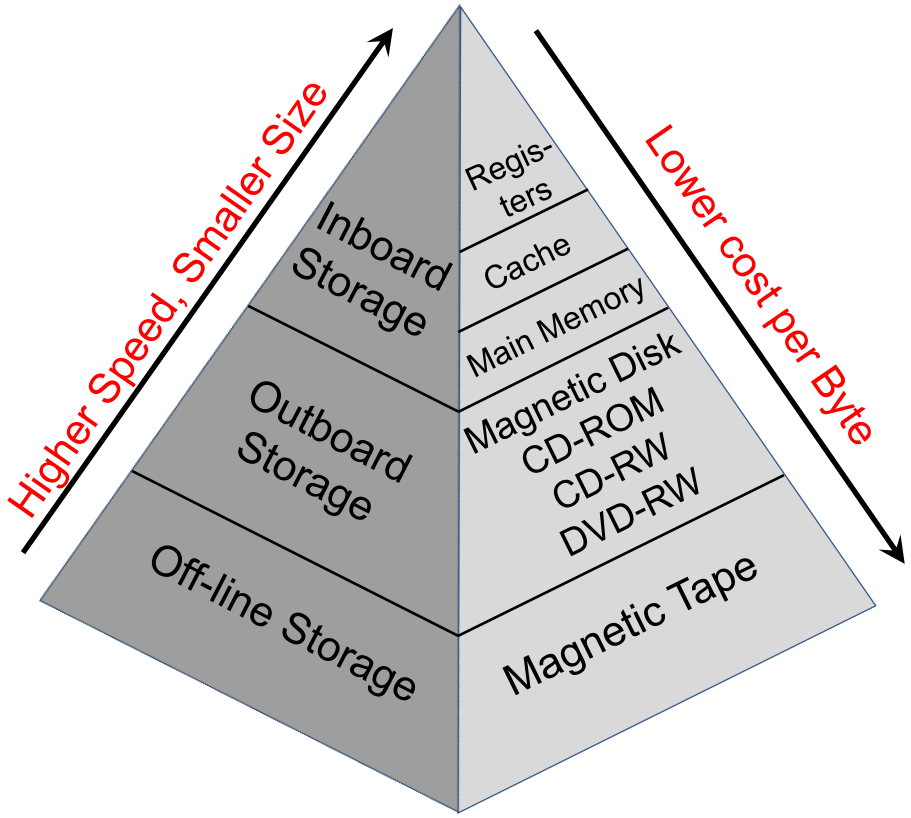


Improvement in Chip Organization and Architecture

- Increase hardware speed of processor
  - Shrinking the size of logic gates on processor reduces the propagation delay among them
- Increase size and speed of cache that is near to processor
  - A portion of processor can be dedicated to cache
- Changing the processor organization and architecture that can increase the instruction execution
  - Using parallelism

Memory Hierarchy

- Total memory capacity can be visualized as hierarchy of components
- As we go down the hierarchy, the following occurs
  - Decreasing cost per bit
  - Increasing capacity
  - Increasing access time
  - Decreasing frequency of access of the memory by the processor

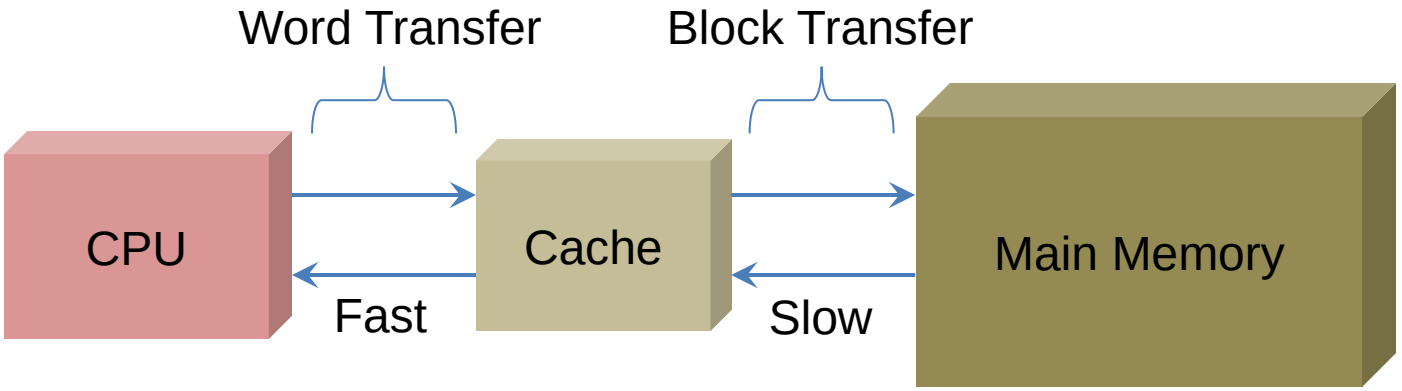


Memory Hierarchy

- Main Memory (Primary Storage)
  - Only programs and data currently need by the processor are stored here
- Auxiliary Memory (Secondary Storage)
  - Devices that provide backup storage
  - Magnetic disks and tapes
- Cache
  - Increases the speed of processing
  - Rapid rate and smaller size
  - Compensates speed differences between main memory and processor
  - Stores segments of programs and data currently in use

Cache Memory

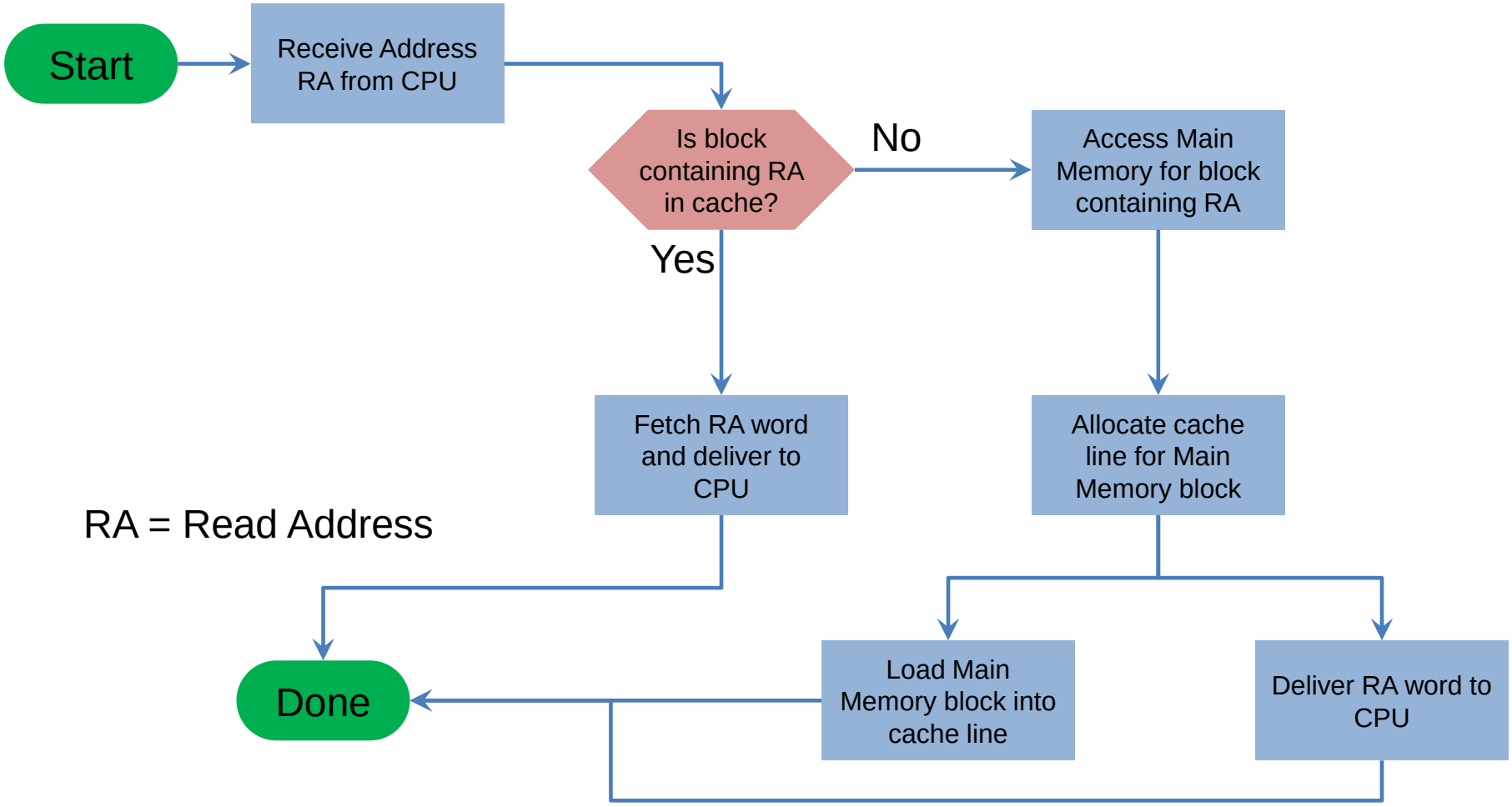
- Small amount of fast memory
- Sits between main memory and CPU
- May be located on CPU chip
- Faster than main memory by a factor of 5 to 10



Cache Operation

- CPU requests contents of memory location
- Check cache for this data
- If present, get from cache (fast operation)
- If not present, read required block from main memory to cache
- Then deliver from cache to CPU

Cache Memory Operation



Cache Performance

- When CPU refers to memory and finds the word in cache, it is said to produce a hit
- If not found, it is called miss

- Hit Ratio = Number of Hits/Total CPU references to memory

Elements of Cache Design (1/2)

- A few basic design elements to classify and differentiate cache architecture
- Cache Address
  - Logical
  - Physical
- Cache Size
- Mapping Function
  - Direct
  - Associative
  - Set Associative
- Line Size

Elements of Cache Design (2/2)

- Replacement Algorithm
  - Least Recently Used (LRU)
  - First in First out (FIFO)
  - Least Frequently Used (LFU)
  - Random
- Write Policy
  - Write through
  - Write back
  - Write once
- Number of Caches
  - Single or two level
  - Unified or Split