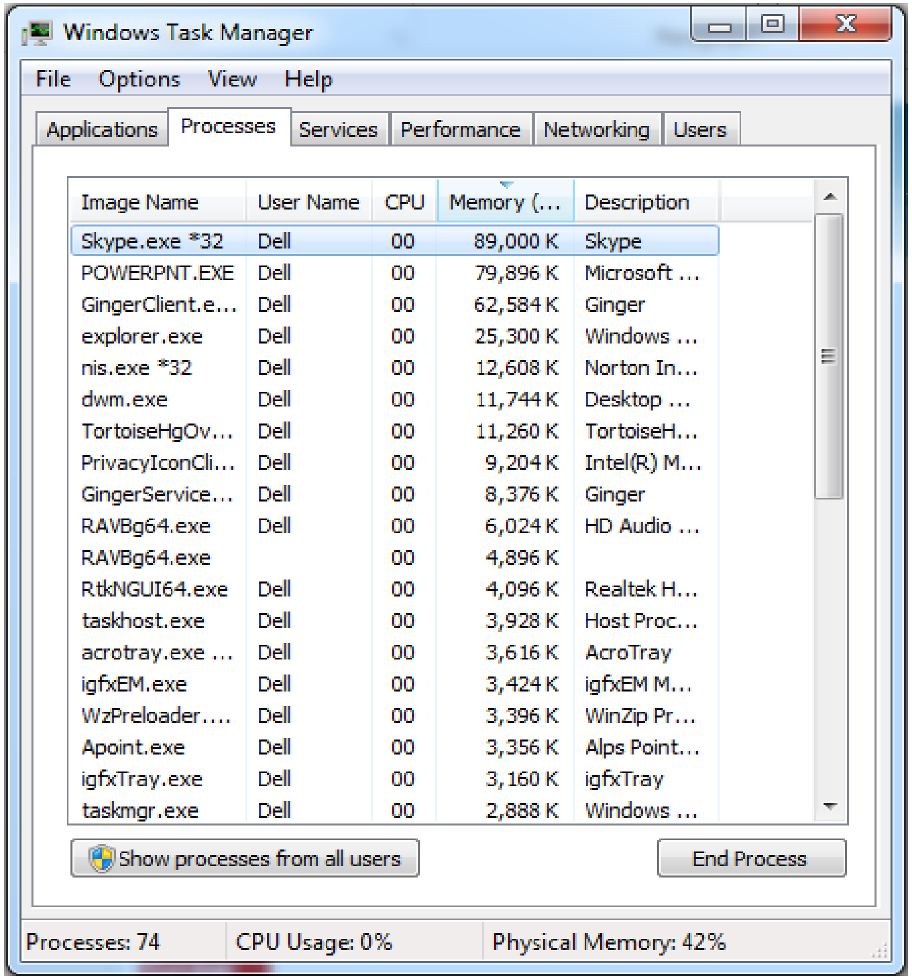
# Computer Memory Systems: Overview

**Dr. B. R. Bhowmik**
**Dept. of CSE**
**NIT Karnataka**

# Need for a Memory System

- Memory is one large block (array of bytes) that
  - Starts a 0, and ends at ($2^{32} – 1$)
  - Takes 1 cycle to access memory (read/write)
- All programs share the memory
  - We somehow magically avoid overlaps between programs running on the same processor.
  - All our programs require less than 4 GB of space.
  - Example: Windows Task Manager displays memory sharing by programs.




All the programs running on a Windows machine. The CPU of course runs one program at a time. Switches between programs periodically.

# Are All The Memory Cells Homogeneous?

| Cell Type | Area | Typical Latency |
|---|---|---|
| Master Slave D flip flop | 0.8 $\mu m^2$ | Fraction of a cycle |
| SRAM cell in an array | 0.08 $\mu m^2$ | 1-5 cycles |
| DRAM cell in an array | 0.005 $\mu m^2$ | 50-200 cycles |

**Typical Values**

# Should We Make Our Memory Using Only Flip-flops ?

- The area of a memory with SRAM cells becomes 10 times (10X) .
- The area of a memory with DRAM cells is 160X.
- Significantly consumes more power !!!

# What do we do ?

- We cannot create a memory of just flip flops
  - We will hardly be able to store anything.
- We cannot create a memory of just SRAM cells
  - We need more storage, and we will not have a 1 cycle latency.
- We cannot create a memory of DRAM cells
  - We cannot afford 50+ cycles per access.
- So, Tradeoffs are needed.

# Tradeoffs

- Tradeoffs are done among Area, Power, and Latency.
- Remember that
  - Increase Area → reduce latency, increase power.
  - Reduce latency → increase area, increase power.
  - Reduce power → reduce area, increase latency.
- So, we cannot have the best of all worlds.

# Characteristics of Memory Systems

- The complex subject of computer memory is made more manageable if we classify memory systems according to their following key characteristics.
  - Location
  - Capacity
  - Unit of transfer
  - Access method
  - Performance
  - Physical type
  - Physical characteristics
  - Organisation

# Memory Location

- Location refers to whether memory is internal (e.g., processor registers, main memory, cache) and external (e.g. optical disks, magnetic disks, tapes) to the computer. Thus, computer locations are two types.
  - Internal Memory often equated with main memory.
  - External Memory consists of peripheral storage devices accessible to the processor via I/O controllers.
- CPU
  - ✓ Assumed as another form of internal memory.
  - ✓ The processor requires its own local memory, in the form of registers.


Dr. B. R. Bhowmik, Dept. of CSE, NIT Karnataka

# Memory Capacity

- An obvious characteristic of memory is its **capacity**.
- *Internal memory capacity* is typically expressed in terms of bytes (1 byte = 8 bits) or words.
- *External memory capacity* is typically expressed in terms of bytes.
- Word: The natural unit of organisation of memory.
- Word Size or Length: The size of the word is typically equal to the number of bits used to represent an integer and to the instruction length.
  - Common word lengths are 8, 16, and 32 bits.
  - Unfortunately, there are many exceptions.
    - The CRAY C90 (an older model CRAY supercomputer) has a 64-bit word length but uses a 46-bit integer representation.
    - The Intel x86 architecture has a wide variety of instruction lengths, expressed as multiples of bytes, and a word size of 32 bits.

# Unit of Transfer

- For main memory, *Unit of transfer* is the number of bits read out of or written into memory at a time.
- The unit of transfer need not equal a word or an addressable unit.
- For Internal Memory: Usually governed by data bus width
- For External Memory: Usually a block which is much larger than a word.
- Addressable unit
  - Smallest location which can be uniquely addressed.
  - Addressing can be either at Word or at byte level by a system.
  - The relationship between the length in bits A of an address and the number N of addressable units is $2^A = N$.

# Access Methods

- The **method of accessing** units of data is four types:
  a) Sequential access,
  b) Direct access,
  c) Random access,
  d) Associative access.


Dr. B. R. Bhowmik, Dept. of CSE, NIT Karnataka

# Access Methods Contd…

- **Sequential Access**
  - Memory is organized into units of data, called records.
  - Access must be made in a specific linear sequence.
  - Start at the beginning and read through in order.
  - Access time depends on location of data and previous location. Thus, the time to access an arbitrary record is highly variable.
  - Tape units are sequential access.

- **Direct Access**
  - direct access involves a shared read–write mechanism.
  - Individual blocks or records have unique address based on physical location.
  - Access is by jumping to vicinity plus sequential search
  - Access time depends on location and previous location.
  - Disk units are direct access.



Dr. B. R. Bhowmik, Dept. of CSE, NIT Karnataka

# Access Methods Contd…

- **Random Access**
  - Individual addresses identify locations exactly.
  - Access time is independent of location or previous access.
  - Thus, any location can be selected at random and directly addressed and accessed.
  - Main memory and some cache systems are random access.

- **Associative Access**
  - This is a random access type of memory that enables one to make a comparison of desired bit locations within a word for a specified match, and to do this for all words simultaneously.
  - Data is located by a comparison with contents of a portion of the store.
  - Access time is independent of location or previous access.
  - Cache memories may employ associative access.

# Performance

- Like the capacity, another most important characteristic of memory is *performance*.
- Three performance parameters are used.
  a) Access time (latency)
  b) Memory Cycle time
  c) Transfer Rate

- **Access time**
  - Time between presenting the address and getting the valid data.
  - For random-access memory, this is the time it takes to perform a read or write operation.
  - For non-random-access memory, access time is the time it takes to position the read–write mechanism at the desired location.

# Performance Contd…

- **Memory Cycle time**
  - Primarily applied to random-access memory.
  - Time may be required for the memory to "recover" before next access.
  - Consists of the access time plus any additional time required before a second access can commence. Thus, *Cycle time = access time + recovery time.*
  - Memory cycle time is concerned with the system bus, not the processor.

- **Transfer Rate**
  - Rate at which data can be moved into or out of a memory unit.
  - For random-access memory, it is equal to 1/(cycle time).
  - For non-random-access memory, the following relationship holds:

$$T_N = T_A + n/R$$

Where,

- $T_N$ = Average time to read or write N bits.
- $T_A$ = Average access time.
- n = Number of bits.
- R = Transfer rate, in bits per second (bps).



Dr. B. R. Bhowmik, Dept. of CSE, NIT Karnataka

# Physical Types

- **Physical type** means a variety of memory.
- The most common today's memory types are
  a) Semiconductor: RAM
  b) Magnetic: Disk & Tape
  c) Optical: CD & DVD
  d) Magneto-optical:  Bubble, Hologram

# Physical Characteristics

- Following physical characteristics of data storage are important.
  a) Decay
  b) Volatility
  c) Erasable
  d) Power consumption
- In a volatile memory, information decays naturally or is lost when electrical power is switched off.
- In a nonvolatile memory, information once recorded remains without deterioration until deliberately changed; no electrical power is needed to retain information.
- Magnetic-surface memories are nonvolatile.
- Semiconductor memory may be either volatile or nonvolatile.
- Nonerasable memory cannot be altered, except by destroying the storage unit. Semiconductor memory of this type is known as *read-only memory* (ROM).
- A practical nonerasable memory must also be nonvolatile.


Dr. B. R. Bhowmik, Dept. of CSE, NIT Karnataka

# Organisation

- Physical arrangement of bits into words is meant to an *organization*.
- For random-access memory, the **organization** is a key design issue.
- The obvious arrangement is not always used, e.g., interleaved.


Dr. B. R. Bhowmik, Dept. of CSE, NIT Karnataka

# Memory Hierarchy

- The design constraints on a computer's memory can be summed up by three questions:
  a) How much? → Capacity
     - ✓ The question of how much is somewhat open ended.
     - ✓ If the capacity is there, applications will likely be developed to use it.
  b) How fast? → Time is money
     - ✓ The question of how fast is, in a sense, easier to answer.
     - ✓ To achieve greatest performance, the memory must be able to keep up with the processor.
     - ✓ As the processor is executing instructions, we would not want it to have to pause waiting for instructions or operands.
  c) How expensive? → Cost
     - ✓ For a practical system, the cost of memory must be reasonable in relationship to other components.

# Memory Hierarchy Contd…

- *The Bottom Line:* there is a trade-off among the three key characteristics of memory.
  a) Capacity.
  b) Access time.
  c) Cost.
- A variety of technologies are used to implement memory systems, and across this spectrum of technologies, the following relationships hold:
  a) Faster access time, greater cost per bit
  b) Greater capacity, smaller cost per bit
  c) Greater capacity, slower access time
- The designer would like to use memory technologies that provide for large-capacity memory, both because the capacity is needed and because the cost per bit is low.
- However, to meet performance requirements, the designer needs to use expensive, relatively lower-capacity memories with short access times.

# Memory Hierarchy Contd…

- The way out of this dilemma is not to rely on a single memory component or technology, but to employ a **memory hierarchy**.
- A typical hierarchy is illustrated in Figure 1.
- As one goes down the hierarchy, the following occur:
  a) Decreasing cost per bit
  b) Increasing capacity
  c) Increasing access time
  d) Decreasing frequency of access of the memory by the processor

# Memory Hierarchy - Diagram

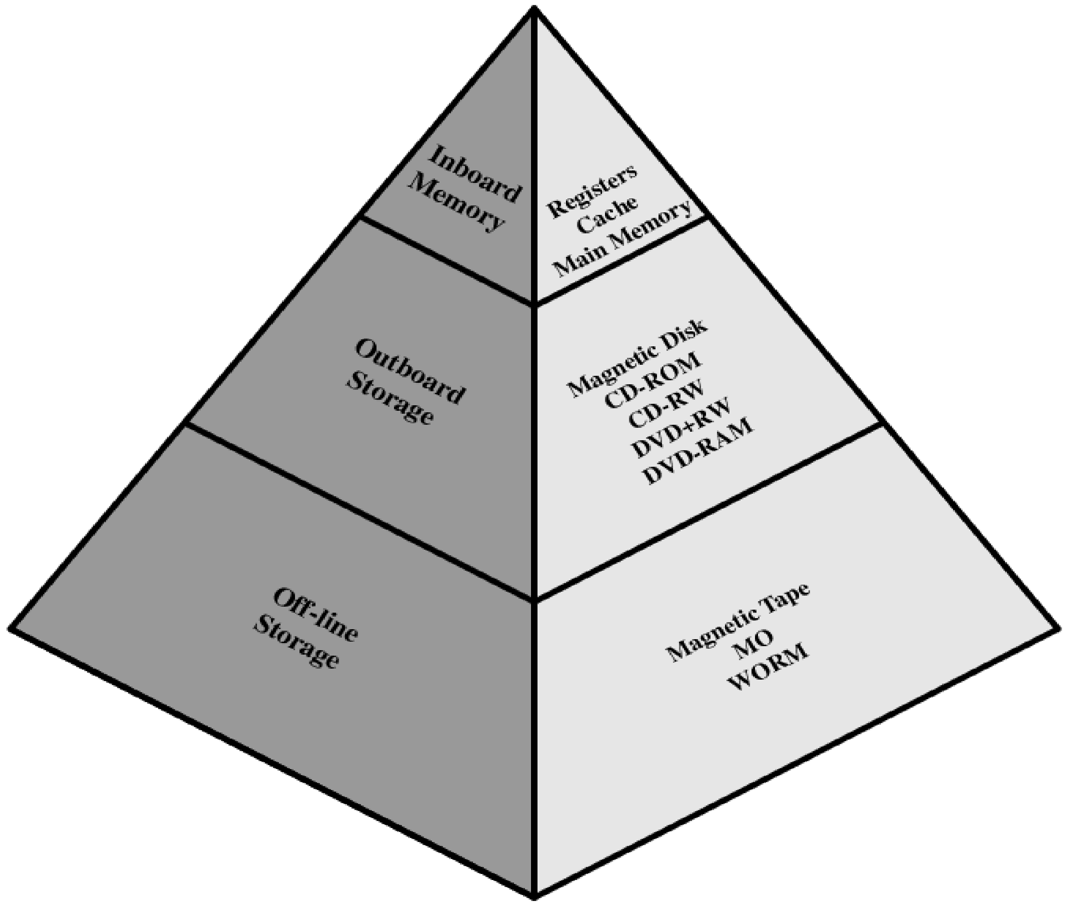



Figure 1. A typical hierarchy

# 3-Level Memory Classification

A. **Inbound storage / memory**
- Registers, Cache, Main memory.
- The class of Internal memory

B. **Outbound Storage**
- Magnetic Disk
- CD ROM
- CD RW
- DVD + RW
- DVD + RAM
- The class of External memory

C. **Off-line Storage**
- Magnetic Tape
- Hard Disk Drive
- The class of  Backing store

# Hierarchy List

- Registers
- L1 Cache
- L2 Cache
- Main memory
- Disk cache
- Disk
- Optical
- Tape

# So you want fast?

- It is possible to build a computer which uses only static RAM
- This would be very fast
- This would need no cache
- This would cost a very large amount

# Locality of Reference

- During the course of the execution of a program, memory references tend to cluster.
- The basis for the validity of condition (d) is a principle known as locality of reference. [see memory hierarchy slide]
- Programs typically contain a number of iterative loops and subroutines.
- Once a loop or subroutine is entered, there are repeated references to a small set of instructions.
- Similarly, operations on tables and arrays involve access to a clustered set of data words.
- Over a long period of time, the clusters in use change, but over a short period of time, the processor is primarily working with fixed clusters of memory references.

Thank You