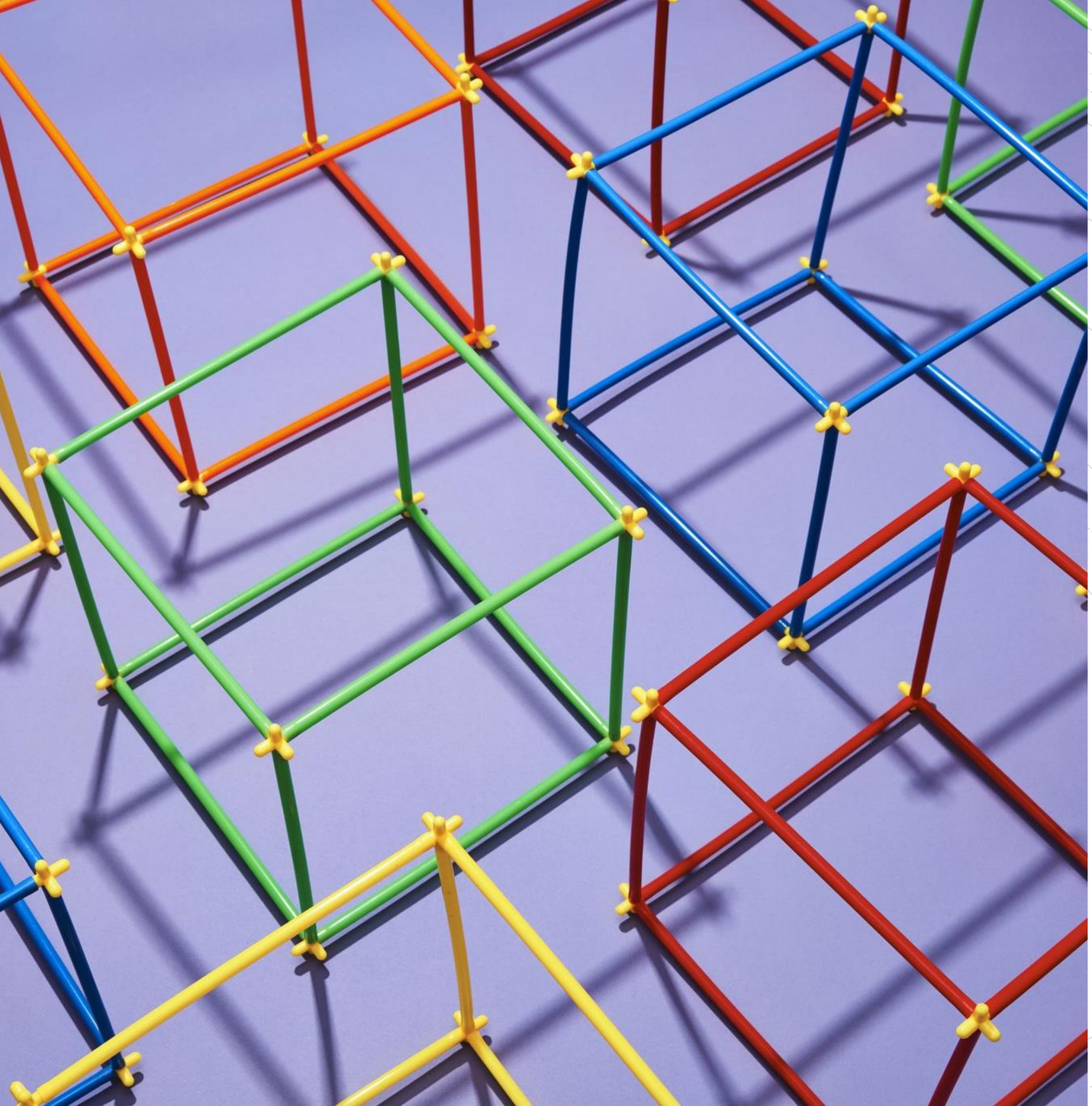


# Computer Architecture IN 2320

Chamalka Rajapaksha

# Computer Memory

- Why it needs a memory?
- Different types of memory

# Characteristics

- Location
- Capacity
- Unit of transfer
- Access method
- Performance
- Physical type
- Physical characteristics
- Organisation

# Location

- **CPU**
- **Internal**
- **External**

# Capacity

- Word size
  - The natural unit of organization
- Number of words
  - or Bytes

# Unit of transfer

## Internal

- Usually governed by data bus width

## External

- Usually a block which is much larger than a word

## Addressable unit

- Smallest location which can be uniquely addressed
- Word internally

# Access Methods

## Sequential

- Start at the beginning and read through in order
- Access time depends on location of data and previous location
- e.g. tape

## Direct

- Individual blocks have unique address
- Access is by jumping to vicinity plus sequential search
- Access time depends on location and previous location
- e.g. disk

## Random

- Individual addresses identify locations exactly
  - Each location physically wired-in addressing mechanism
- Access time is independent of location or previous access
  - e.g. RAM (Main memory and some caches)

## Associative

- Data is located by a comparison with contents of a portion of the store
  - E.g. within words
- Access time is independent of location or previous access
  - e.g. cache

# Memory Hierarchy

- Registers
  - In CPU
- Internal or Main memory
  - May include one or more levels of cache
  - “RAM”
- External memory
  - Backing store

# Performance



- Access time (latency)
  - Time between presenting the address and getting the valid data
- Memory Cycle time
  - Time may be required for the memory to ‘‘recover’’ before next access
  - Cycle time is access + recovery
- Primarily applied to random access memory
- Transfer Rate
  - Rate at which data can be moved
- 1/(cycle time), for random access memory
- TN=TA+N/R, For non-random access memory
  - TN-Average time to read or write N bits
  - TA-Average access time
  - N –Number of bits
  - R –Transfer rate, in bits per seconds

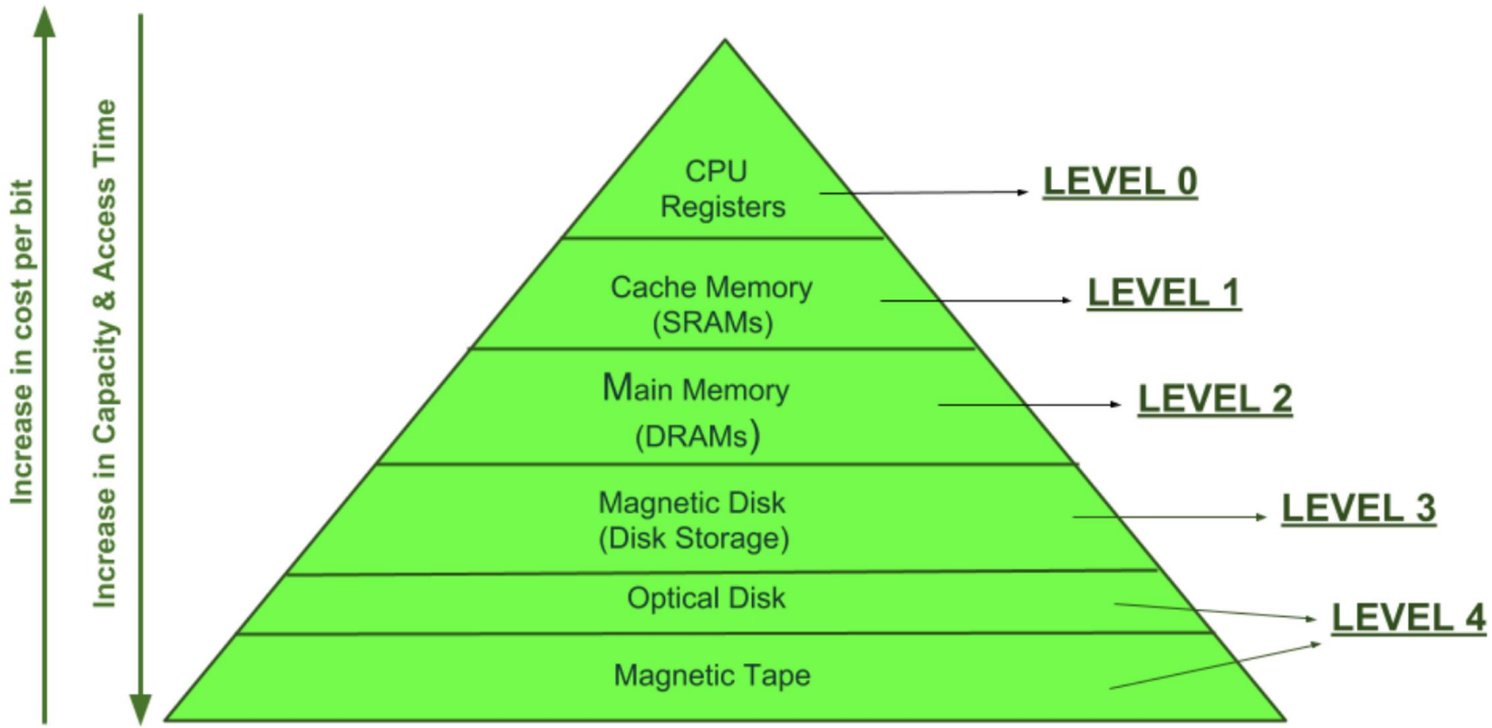
# Physical types



- Semiconductor
  - SRAM, caches
- Magnetic
  - Disk & Tape
- Optical
  - CD & DVD
  - Others
  - Bubble

# Physical Characteristics

- Decay
- Volatility
- Erasable
- Power consumption




# Memory Hierarchy

# Memory Issues

- Capacity
- Latency
- Bandwidth

# Memory Access Latency

- What does memory access latency depend on?
  - Size of the memory
    - Larger is the size, slower it is
  - Number of ports
    - More are the ports, slower is the memory
  - Technology used
    - SRAM, DRAM, flip flops

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

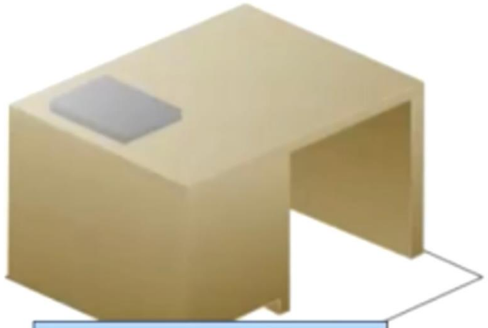
# Tradeoffs

- Area | Power | Latency
- Increase Area —> Reduce Latency, Increase Power
- Reduce Latency —> Increase Area, Increase Power
- Reduce Power —> Reduce Area, Increase Latency

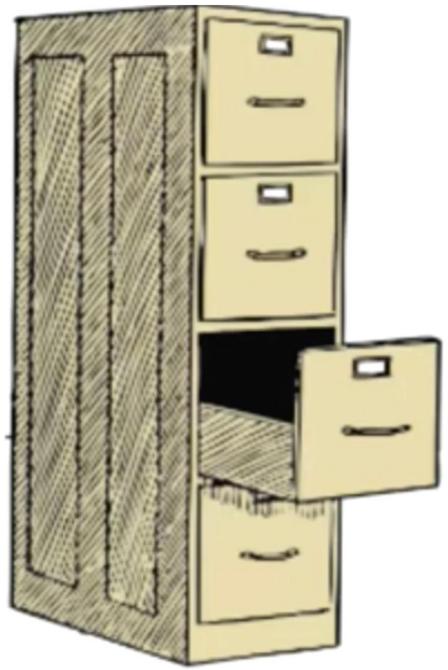
# What do we do?

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

shelf



desk



cabinet

# Example in real life - Protocol 1

- Idea - She tends to read the same set of books, over and over again, in the same window of time

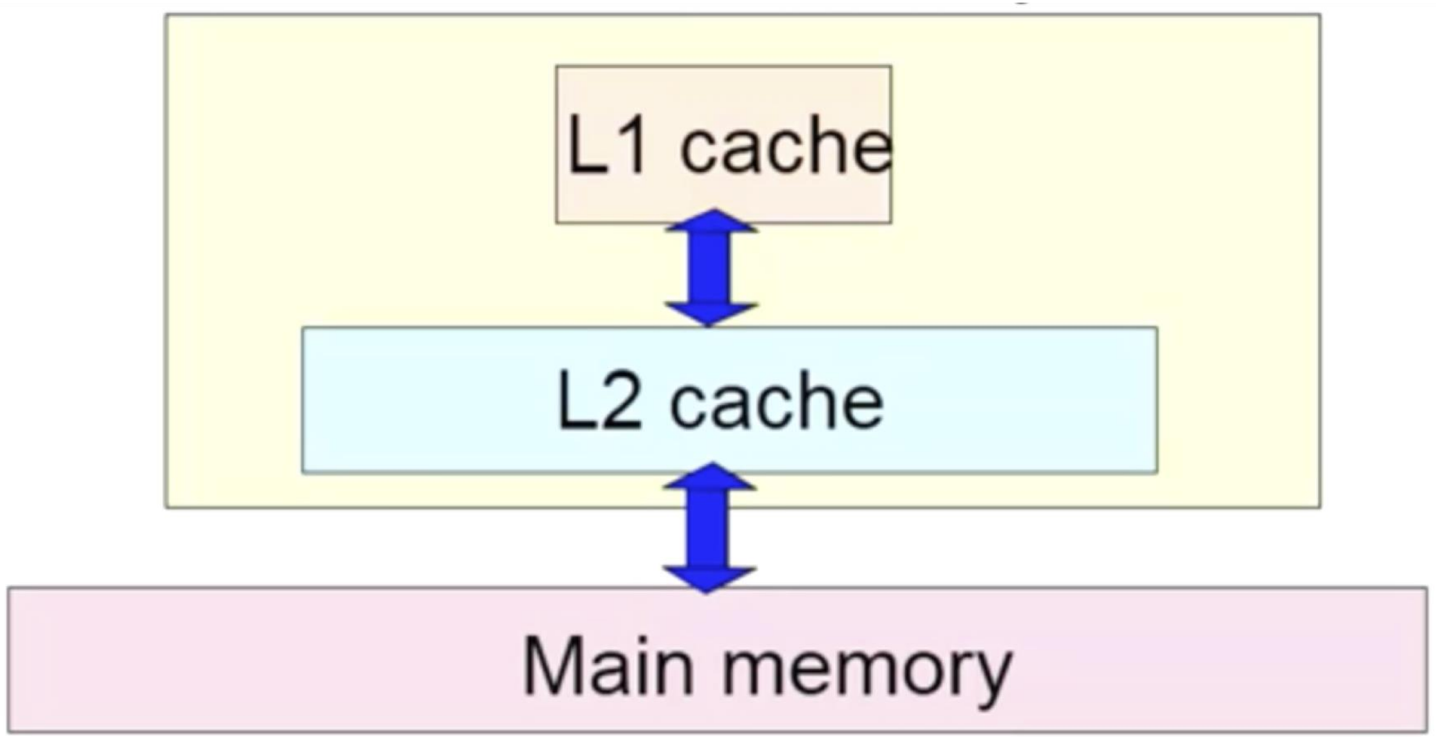
# Example in real life - Protocol 1

# Example in real life - Protocol 2

- If Amali takes a Computer Architecture module
  - She has  Computer Architecture books in her desk
- After the module is over
  - The architecture books go back to the shelf.
  - OOAD books come on to the desk
  - Idea - Bring all the OOAD books in one go. If she requires one, in high likelihood she may require similar books in near future.

# Locality of reference

- Temporal coherence - There is a higher probability of repeated access to any data item that has been accessed in the recent past
- Spatial coherence: There is a higher probability of access to any data item that is physically closer to any other data item that has been access in the recent past




# Memory Access Time

# Memory Access Time

- Protocol
  - First comes L1 cache. If the memory location is present, we have a cache hit
    - Perform the access (read/write)
  - Otherwise we have a cache miss
    - Fetch the values from the lower levels of the memory system, and populate the cache.
    - Follow this protocol recursively

# Memory Access Time

- Ave access time, Ts= H x T1 + (1-H) x (T1 + T2)
- H-fraction of all memory accesses that are found in the faster memory (Hit ratio)
- T1 –Access time to level 1
- T2 –Access time to level 2
- Access efficiency = T1/Ts

# Semiconductor Memory

- Basic element of a semiconductor memory is the memory cell
  - Cell is able be in one of two states
  - Read/write

- RAM
  - Misnamed as all semiconductor memory is random access
  - Read/Write
  - Volatile
  - Temporary storage
  - Static or dynamic

# DRAM

- Bits stored as charge in capacitors
- Charges leak
- Need refreshing even when powered
- cycle time traditionally longer than the access time
- Simpler construction
- Smaller per bit
- Less expensive
- Need refresh circuits
- Slower
- Main memory

# SRAM

- Bits stored as on/off switches
- No charges to leak
- No refreshing needed when powered
- More complex construction
- Larger per bit
- More expensive
- Does not need refresh circuits
- Faster
- Cache

# Types of ROM

- Written during manufacture
- Very expensive for small runs
- Programmable (once)
  - PROM
  - Needs special equipment to program
- Read “mostly”
  - Erasable Programmable (EPROM)
    - Erased by UV
  - Electrically Erasable (EEPROM)
    - Takes much longer to write than read
  - Flash memory
    - Erase whole memory electrically

# Synchronous DRAM

- Currently on DIMMs
- Access is synchronized with an external clock
- Address is presented to RAM
- RAM finds data (CPU waits in conventional DRAM)
- Since SDRAM moves data in time with system clock,
- CPU knows when data will be ready
- CPU does not have to wait, it can do something else
- Burst mode allows SDRAM to set up stream of data and fire it out in block
- DDR-SDRAM sends data twice per clock cycle (leading & trailing edge)
- SDRAM includes an on-chip burst counter that can be used to increment column addresses for very fast burst accesses.

# Questions?