Computer Organization and Architecture

Under Graduate Course
(B. Tech-Information Technology, 2nd Semester)
Jan 2020 - July 2020

By

Dr. Satish Kumar Singh

Associate Professor
Indian Institute of Information Technology, Allahabad
Email:sk.singh@iiita.ac.in

Memory System Design

CHARACTERISTICS OF THE MEMORY[2]

- 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
    - Cluster or sector on disks

CHARACTERISTICS CONT….

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

CHARACTERISTICS CONT....

- Performance
  - Access time
    - 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
  - Transfer Rate
    - Rate at which data can be moved
- Physical type
  - Semiconductor
  - Magnetic
  - Optical
  - Others

CHARACTERISTICS CONT....

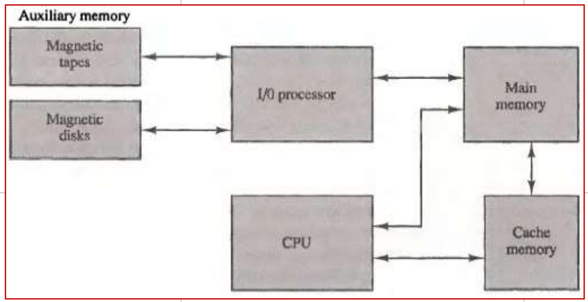
- Physical characteristics
  - Decay
  - Volatility
  - Erasable
  - Power consumption
- Organization
  - Physical arrangement of bits into words
  - Not always obvious
  - e.g. interleaved

MEMORY HIERARCHY

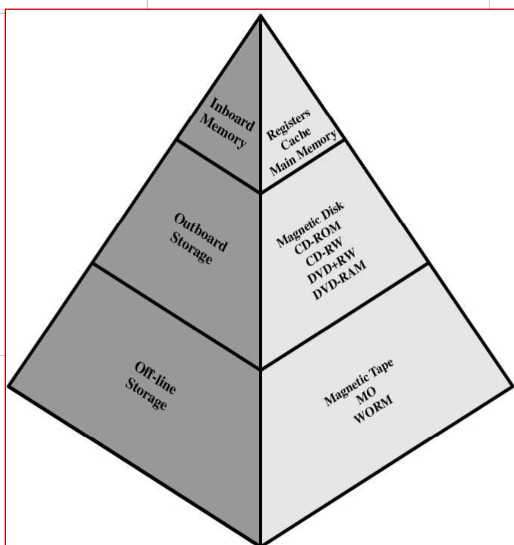
- Registers
  - In CPU
- Internal or Main memory
  - May include one or more levels of cache
  - "RAM"
- External memory
  - Backing store

MEMORY HIERARCHY CONT….

- Computational efficiency enhancement ← Using additional storage.
- General Memory used in computer systems,

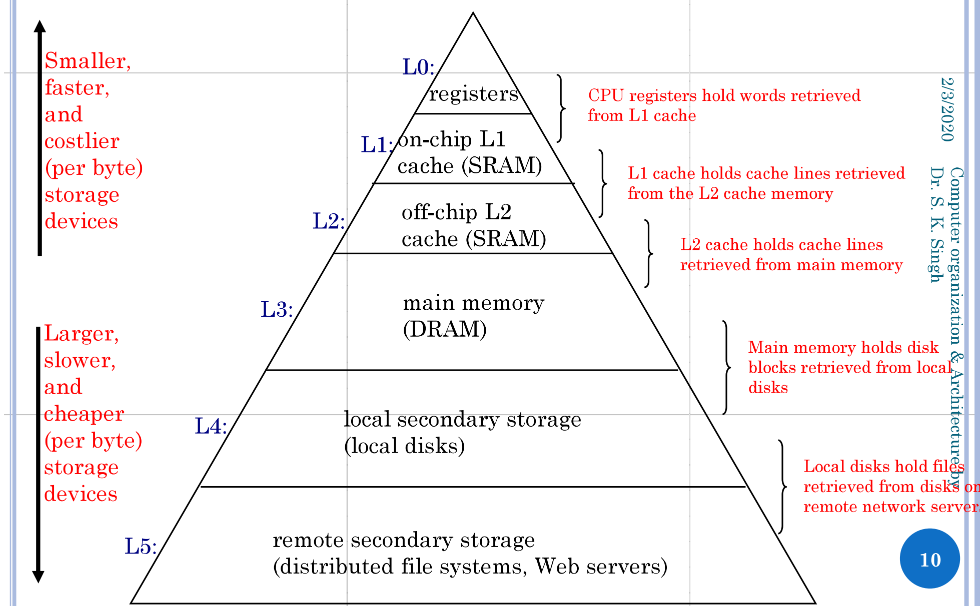


MEMORY HIERARCHY CONT….



Ref: COA, W. Stallings

AN EXAMPLE MEMORY HIERARCHY

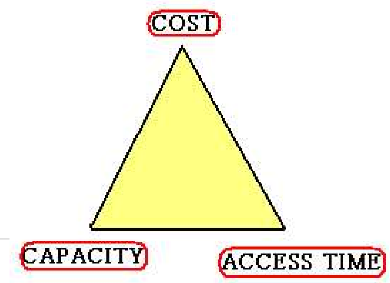


MEMORY HIERARCHY CONT….

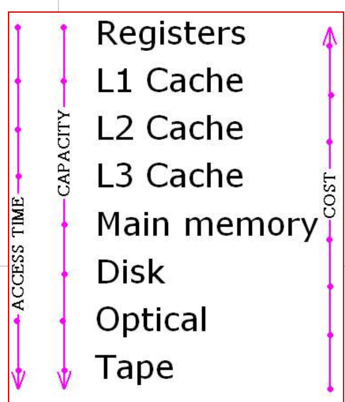
- Computer Memory
  - Main/Primary memory
    - Direct communication from CPU
    - Contains the program and data currently need by the processor
      - Processor registers (may be treated as)
      - RAM
      - ROM
      - CACHE (L1-L3)
  - Auxiliary/Secondary storage
    - Provides back-up storage for
    - System files, large data files and other backup information
      - Magnetic disks
      - Tapes
      - Optical
      - Others

MEMORY HIERARCHY CONT….

- Tradeoff triangle



Memory Hierarchy cont….



Main Memory-1[Random Access Memory]

- Key features
  - RAM is packaged as a chip
  - Basic storage unit is a cell (one bit per cell)
  - Multiple RAM chips form a memory
- Static RAM (SRAM)
  - Each cell stores bit with a six-transistor circuit
  - Retains value indefinitely, as long as it is kept powered
  - Relatively insensitive to disturbances such as electrical noise
  - ***Faster*** and ***More expensive*** than DRAM
  - Used to implement the ***Cache*** memory.
- Dynamic RAM (DRAM)
  - Each cell stores bit with a capacitor and transistor
  - Value must be refreshed every **10-100** ms
  - Sensitive to disturbances
  - ***Slower*** and ***Cheaper*** than SRAM
  - ***Reduced*** power consumption & larger ***Storage*** capacity
  - Used to implement the ***Main*** memory.

MAIN MEMORY-1[READ ONLY MEMORY]

- Key features
  - Portion of main memory is made of ROM chips.
  - ROM is packaged as a chip
  - Basic storage unit is a cell (one bit per cell)
  - Multiple ROM chips form a memory
  - ROM is also random access type.
  - ROM stores the permanent programs and tables of constants which do not required to be changed.
- Bootstrap Loader:
  - It is a initial program for starting the computer software operating after the power-switch is turned-on.
  - Due to noN-volatility the Bootstrap loader can be placed in ROM.

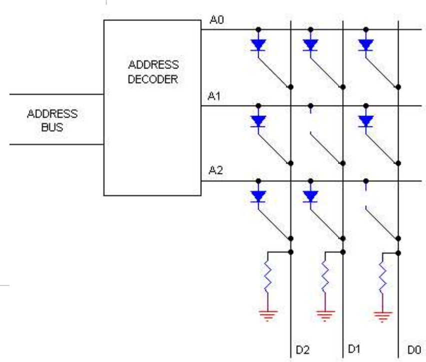
MAIN MEMORY-1[READ ONLY MEMORY]

- Programmable read only memory (PROM)
  - Re-programmed by using a special device called a PROM programmer.
  - Generally, a PROM can only be changed/updated once.
- Erasable Programmable read only memory (EPROM)
  - Erased by ultraviolet light
  - Reprogrammed by an EPROM programmer.
  - Erasing and programming many times however, the constant erasing and rewriting will eventually render the chip useless.
- Electrically Erasable Programmable read only memory (EEPROM)
  - Similar way to Flash memory
  - EEPROMs are used to store a computer system's BIOS, and can be updated without returning the unit to the factory.
  - BIOS updates can be carried out by computer users wishing a BIOS update.

What is BIOS?

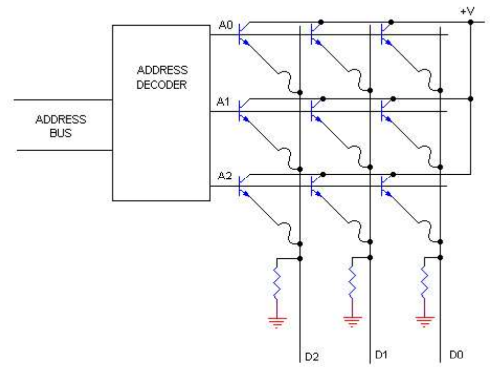
https://en.wikipedia.org/wiki/BIOS

ROM



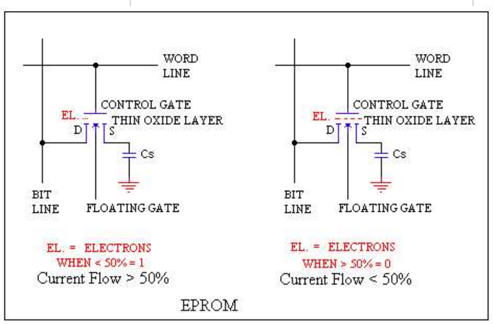
READ ONLY MEMORY (ROM)

PROM



PROGRAMMABLE READ ONLY MEMORY (PROM)

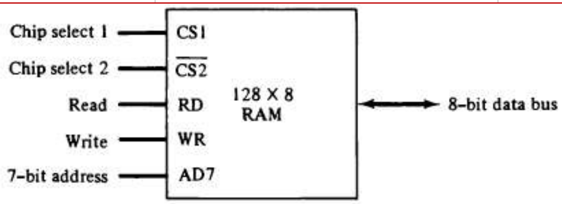
EPROM



NON-VOLATILE RAM (NVRAM)

- Key Feature: Keeps data when power lost
  - Several types
  - Most important is NAND flash
- NAND flash
  - Reading similar to DRAM (though somewhat slower)
  - Writing packed with restrictions:
    - Can't change existing data
    - Must erase in large blocks (e.g., 64K)
    - Block dies after about 100K erases
  - Writing slower than reading (mostly due to erase cost)
  - Chips often packaged with Flash Translation Layer (FTL)
    - Spreads out writes ("wear leveling")
    - Makes chip appear like disk drive

RAM Chips

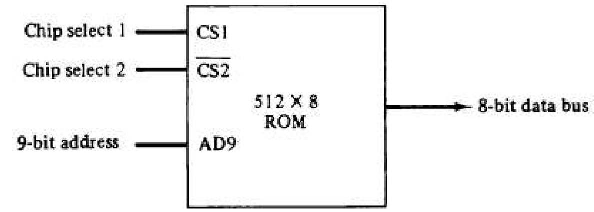


(a) Block diagram

CS1	$\overline{CS2}$	RD	WR	Memory function	State of data bus
0	0	×	×	Inhibit	High-impedance
0	1	×	×	Inhibit	High-impedance
1	0	0	0	Inhibit	High-impedance
1	0	0	1	Write	Input data to RAM
1	0	1	×	Read	Output data from RAM
1	1	×	×	Inhibit	High-impedance

(b) Function table

ROM Chips



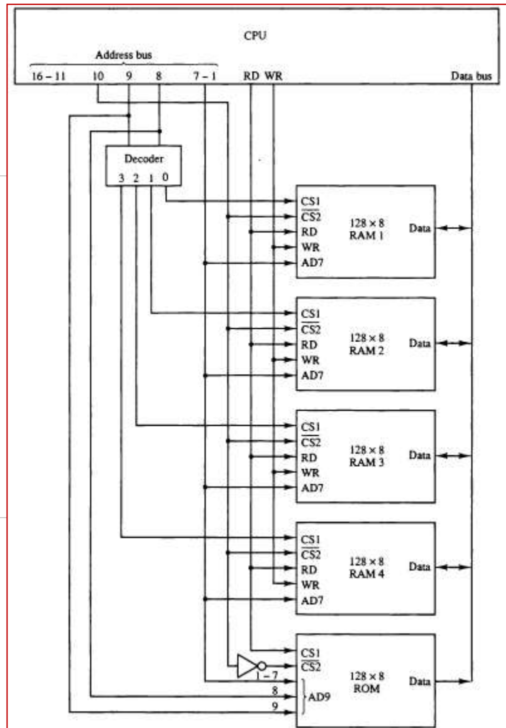
MEMORY ADDRESS MAP

- Calculation of amount of memory required.
- Assignment as RAM and ROM portions.
- Interconnection between processor and memory.
- So memory address map is the table/picTorial representation of assigned address space for each chip in the system.
- Example:
  - RAM of size 512x8 = (128x8)x4
  - ROM of size 512x8.
  - Generate the memory address map and show the connections of memory to processor.

MEMORY ADDRESS MAP FOR MICROCOMPUTER

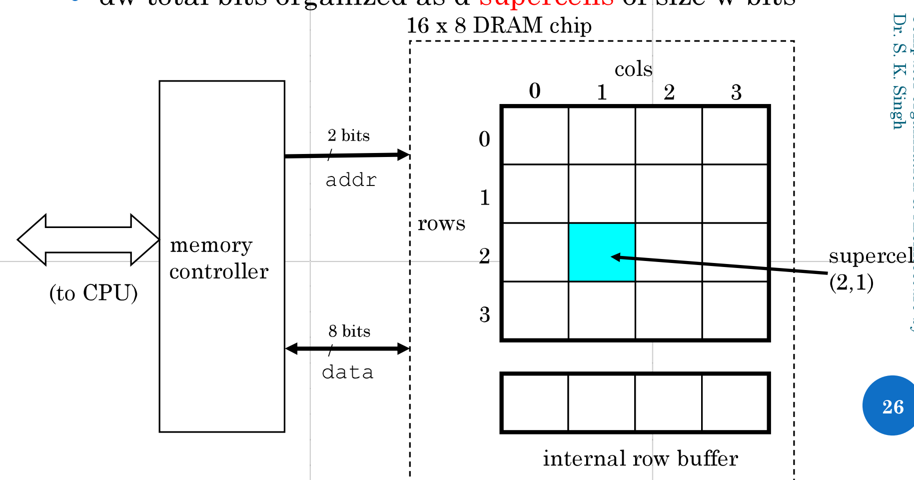
Component	Hexadecimal address	Address bus 10	9	8	7	6	5	4	3	2	1
RAM 1	0000–007F	0	0	0	x	x	x	x	x	x	x
RAM 2	0080–00FF	0	0	1	x	x	x	x	x	x	x
RAM 3	0100–017F	0	1	0	x	x	x	x	x	x	x
RAM 4	0180–01FF	0	1	1	x	x	x	x	x	x	x
ROM	0200–03FF	1	x	x	x	x	x	x	x	x	x

Memory connections to CPU



Conventional DRAM Organization

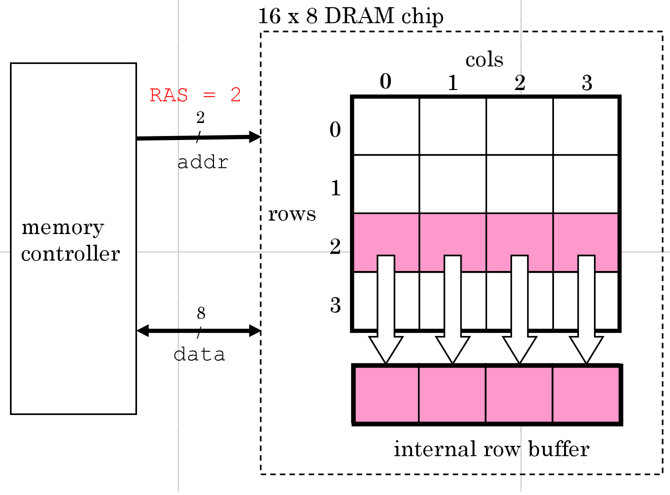
- d x w DRAM:
  - dw total bits organized as d supercells of size w bits



Reading DRAM Supercell (2,1)

- Step 1(a): Row address strobe (RAS) selects row 2

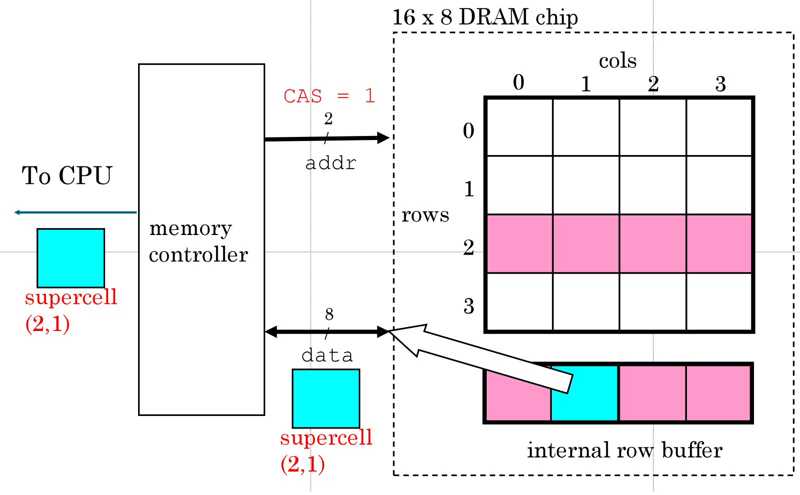
Step 1(b): Row 2 copied from DRAM array to row buffer



Reading DRAM Supercell (2,1)

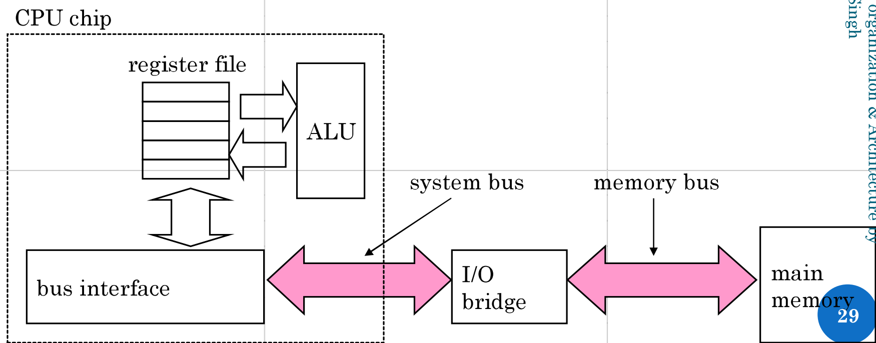
- Step 2(a): Column access strobe (CAS) selects column 1

Step 2(b): Supercell (2,1) copied from buffer to data lines, and eventually back to CPU



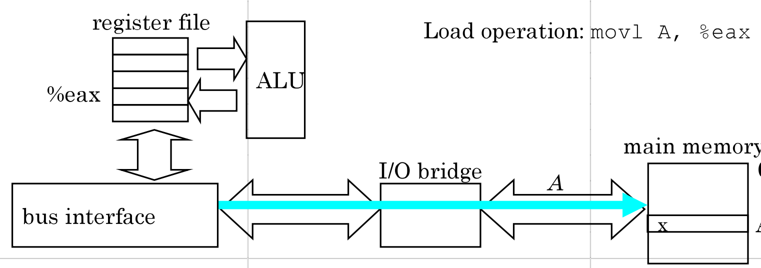
TYPICAL BUS STRUCTURE CONNECTING CPU AND MEMORY

- A bus is a collection of parallel wires that carry address, data, and control signals
- Buses are typically shared by multiple devices



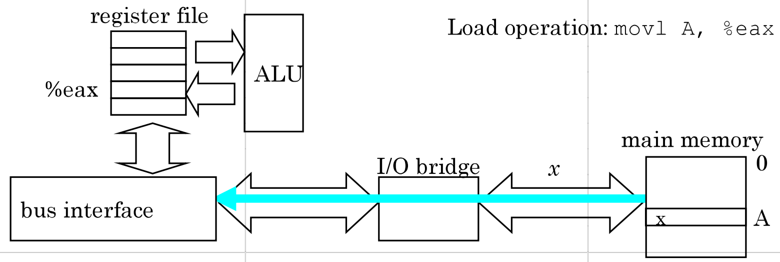
MEMORY READ TRANSACTION (1)

- CPU places address A on memory bus



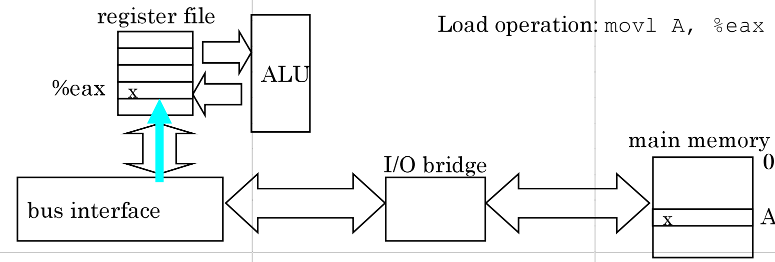
Memory Read Transaction (2)

- Main memory reads A from memory bus, retrieves word x, and places it on bus



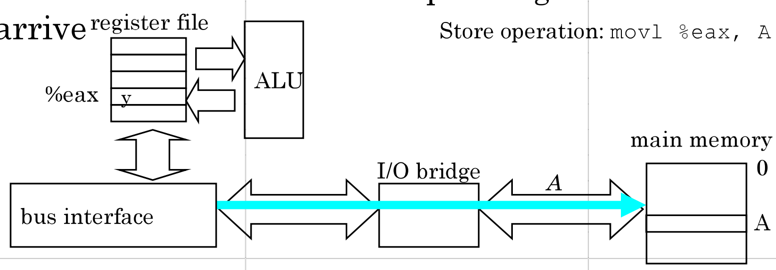
Memory Read Transaction (3)

- CPU reads word x from bus and copies it into register %eax



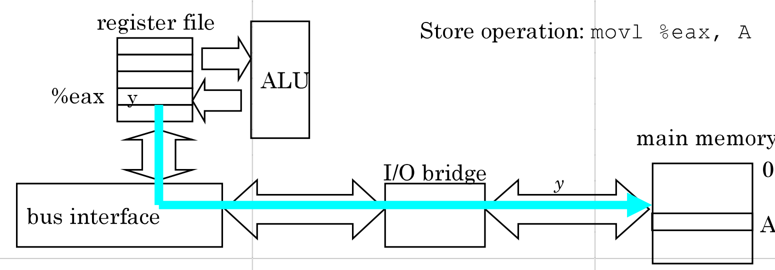
Memory Write Transaction (1)

- CPU places address A on bus; main memory reads it and waits for corresponding data word to arrive



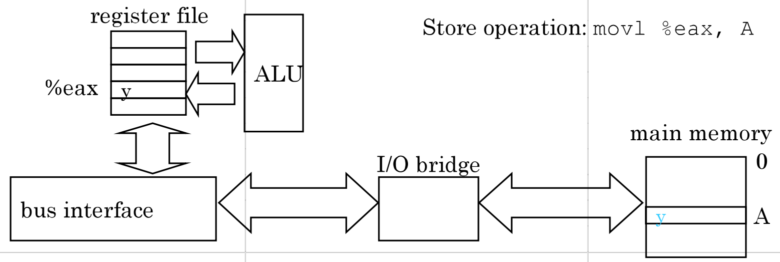
Memory Write Transaction (2)

- CPU places data word y on bus



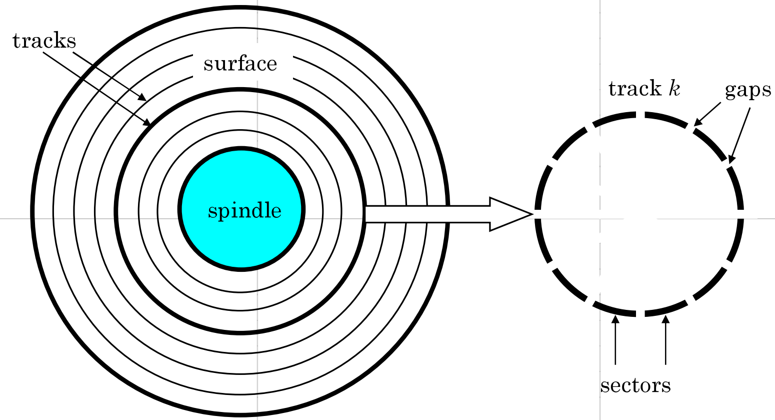
Memory Write Transaction (3)

- Main memory reads data word y from bus and stores it at address A



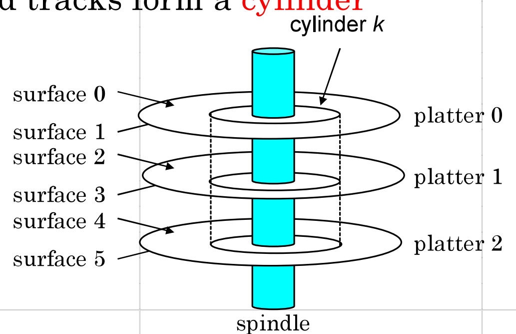
Disk Geometry

- Disks consist of platters, each with two surfaces
- Each surface consists of concentric rings called tracks
- Each track consists of sectors separated by gaps

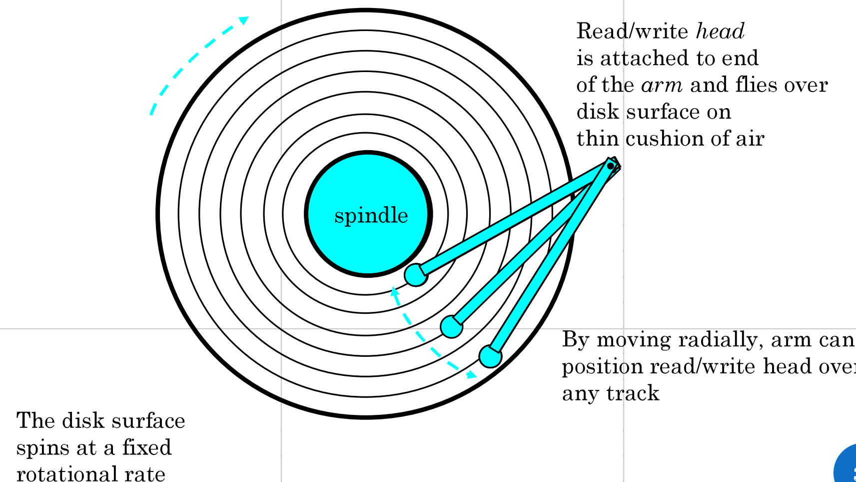


Disk Geometry (Muliple-Platter View)

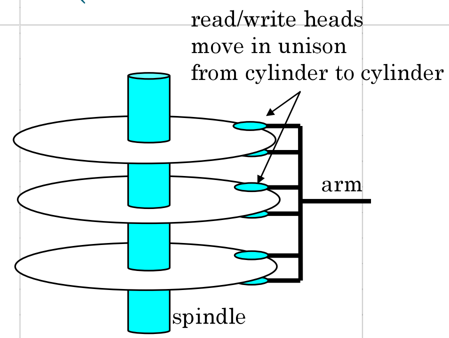
- Aligned tracks form a cylinder



Disk Operation (Single-Platter View)



Disk Operation (Multi-Platter View)



Disk Access Time

- Average time to access some target sector approximated by
  - Taccess = Tavg seek + Tavg rotation + Tavg transfer
- Seek time (Tavg seek)
  - Time to position heads over cylinder containing target sector
  - Typical Tavg seek = 9 ms
- Rotational latency (Tavg rotation)
  - Time waiting for first bit of target sector to pass under r/w head
  - Tavg rotation = 1/2 x 1/RPMs x 60 sec/1 min
- Transfer time (Tavg transfer)
  - Time to read the bits in the target sector.
  - Tavg transfer = 1/RPM x 1/(avg # sectors/track) x 60 secs/1 min

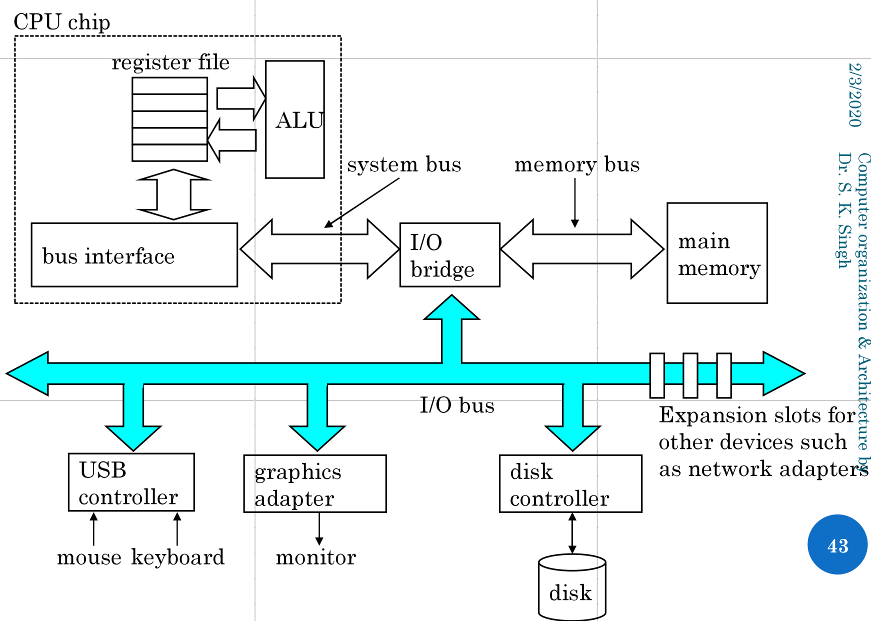
Disk Access Time Example

- Given:
  - Rotational rate = 7,200 RPM
  - Average seek time = 9 ms
  - Avg # sectors/track = 400
- Derived:
  - Tavg rotation = 1/2 x (60 secs/7200 RPM) x 1000 ms/sec = 4 ms
  - Tavg transfer = 60/7200 RPM x 1/400 secs/track x 1000 ms/sec = 0.02 ms
  - Taccess = 9 ms + 4 ms + 0.02 ms
- Important points:
  - Access time dominated by seek time and rotational latency
  - First bit in a sector is the most expensive, the rest are free
  - SRAM access time is about 4 ns/doubleword, DRAM about 60 ns
    - Disk is about 40,000 times slower than SRAM, and
    - 2,500 times slower then DRAM

Logical Disk Blocks

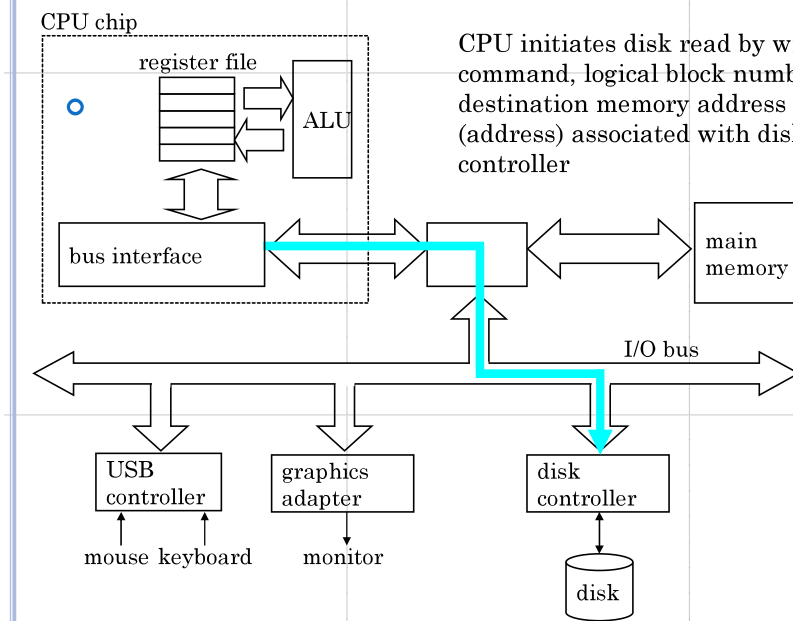
- Modern disks present a simpler abstract view of the complex sector geometry:
  - The set of available sectors is modeled as a sequence of b-sized logical blocks (0, 1, 2, ...)
- Mapping between logical blocks and actual (physical) sectors
  - Maintained by hardware/firmware device called *disk controller*
  - Converts requests for logical blocks into (surface,track,sector) triples
- Allows controller to set-aside spare cylinders for each zone
  - Accounts for the difference in "formatted capacity" and "maximum capacity"

I/O Bus

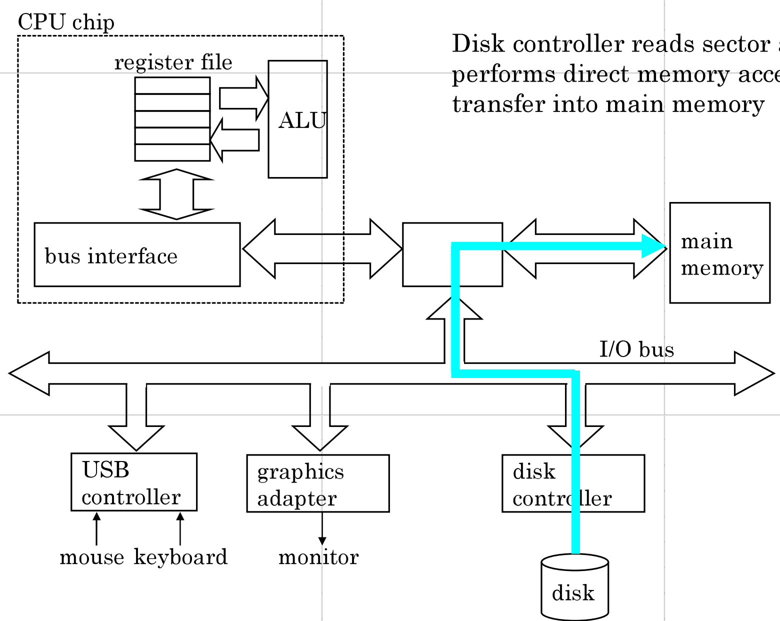


Reading a Disk Sector (1)

CPU initiates disk read by writing command, logical block number, and destination memory address to a port (address) associated with disk controller

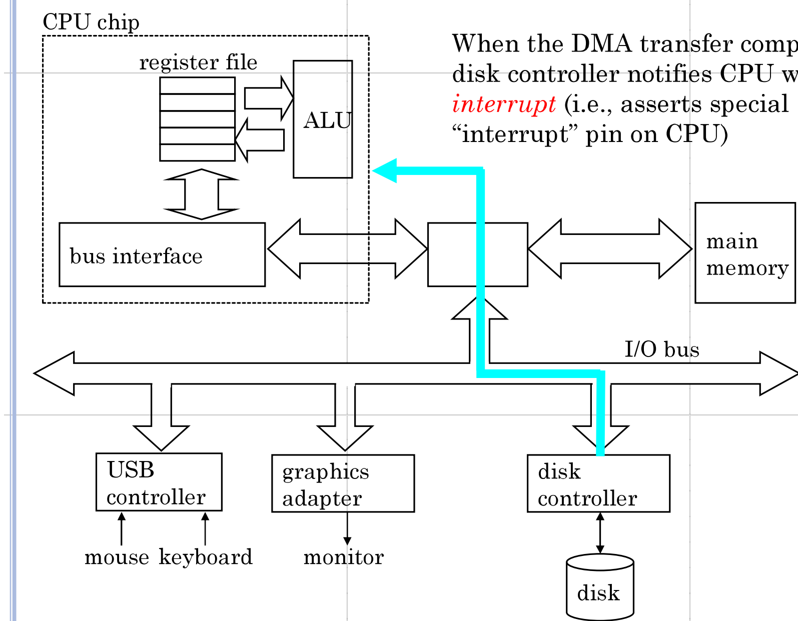


Reading a Disk Sector (2)



Disk controller reads sector and performs direct memory access (DMA) transfer into main memory

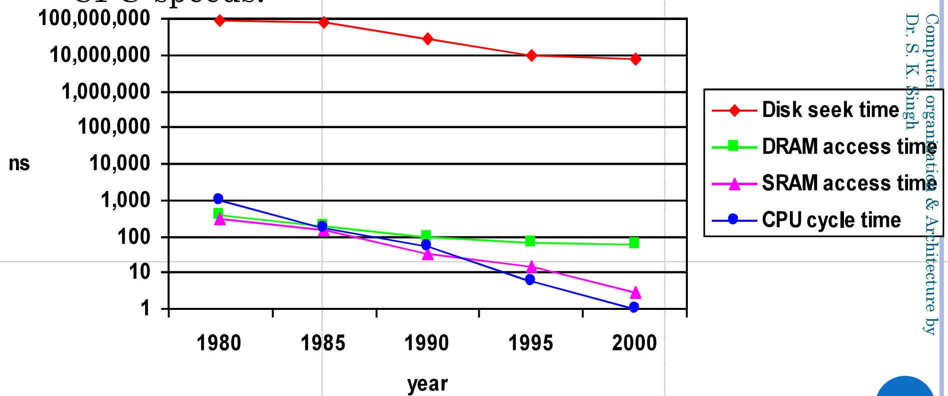
Reading a Disk Sector (3)



When the DMA transfer completes, disk controller notifies CPU with *interrupt* (i.e., asserts special "interrupt" pin on CPU)

THE CPU-MEMORY GAP

- The increasing gap between DRAM, disk, and CPU speeds.



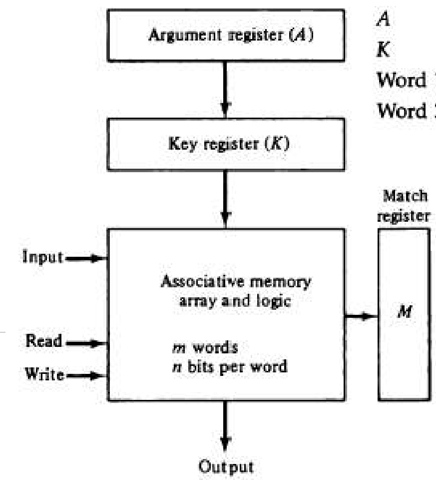
ASSOCIATIVE MEMORY

- Memory access time is a crucial parameter with the performance point of view.
- The memory access time can be reduced considerably by,
  - Taking very fast memory devices
  - Decreasing the number of access to memory
    - Depends on location of content
    - Search algorithm
- Searching time required for an item can be reduced is data is accessed on the basis of content rather than the address of data itself.
- Memory unit addressable by its content is know as the content addressable memory (CAM)
- Entire memory is addressed simultaneously and in parallel format while on the basis of word contents.

ASSOCIATIVE MEMORY CONT….

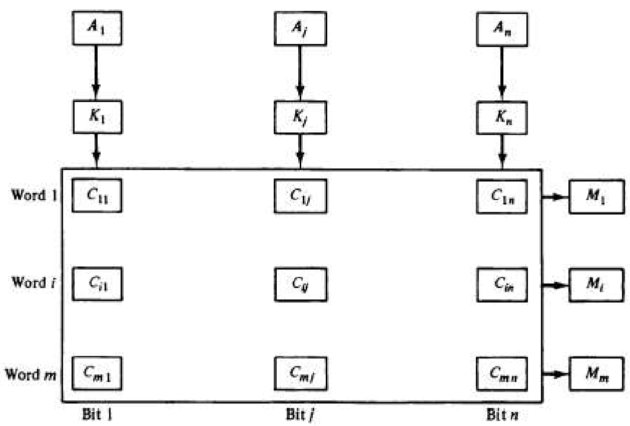
- No address is given while writing in CAM.
- Memory itself is capable of finding the unused / empty location to store a word.
- While reading the word from CAM a part of content of the word is specified,
- Memory locates all such matching words and mark them to be read.
- It is very costly because,
  - Each cell must have storage as well as matching logic capability.
- CAM are used only when the search time is very critical.

HARDWARE ORGANIZATION FOR ASSOCIATIVE MEMORY



A	101 111100	
K	111 000000	
Word 1	100 111100	no match
Word 2	101 000001	match

Associative Memory with m-words and n-cells per word



Associative memory cont….

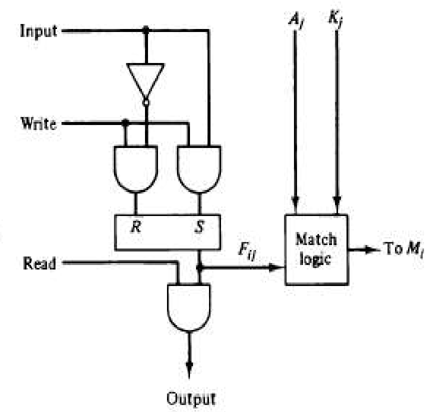
$$x_j = A_j F_{ij} + A_j' F_{ij}'$$

$$M_i = x_1 x_2 x_3 \cdots x_n$$

$$x_j + K_j' = \begin{cases} x_j & \text{if } K_j = 1 \\ 1 & \text{if } K_j = 0 \end{cases}$$

$$M_i = (x_1 + K_1')(x_2 + K_2')(x_3 + K_3') \cdots (x_n + K_n')$$

$$M_i = \prod_{j=1}^{n} (A_j F_{ij} + A_j' F_{ij}' + K_j')$$



READ AND WRITE OPERATIONS OF ASSOCIATIVE MEMORY

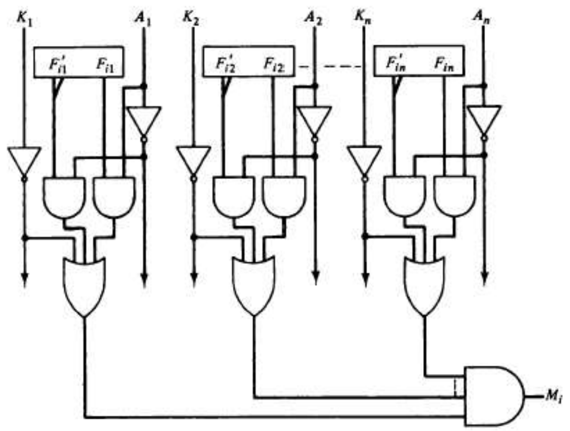


Figure 12-9 Match logic for one word of associative memory.

Thank you