Computer Organization and Operating System

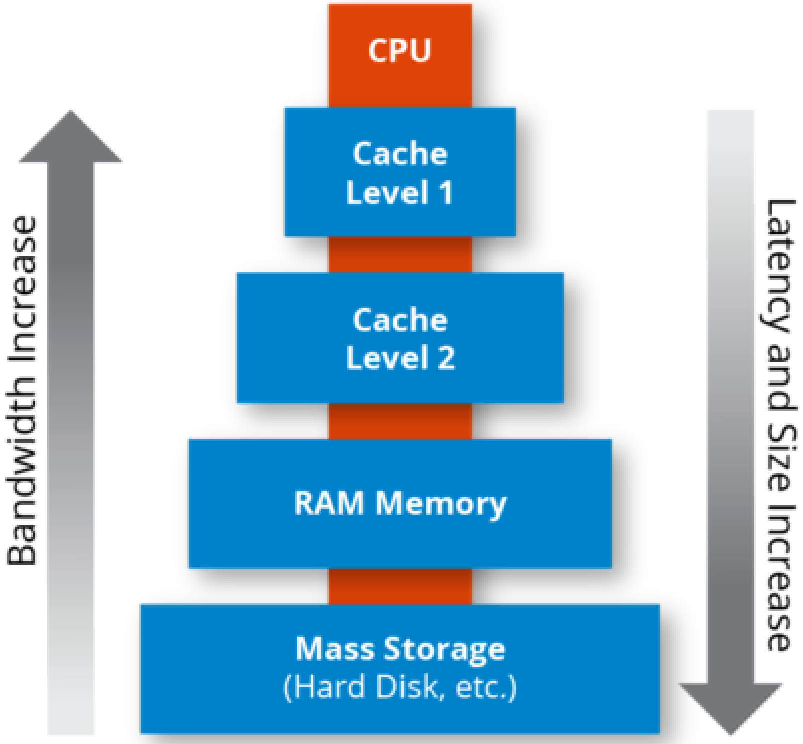
Memory and Cache

Akharin Khunkitti

KMITL

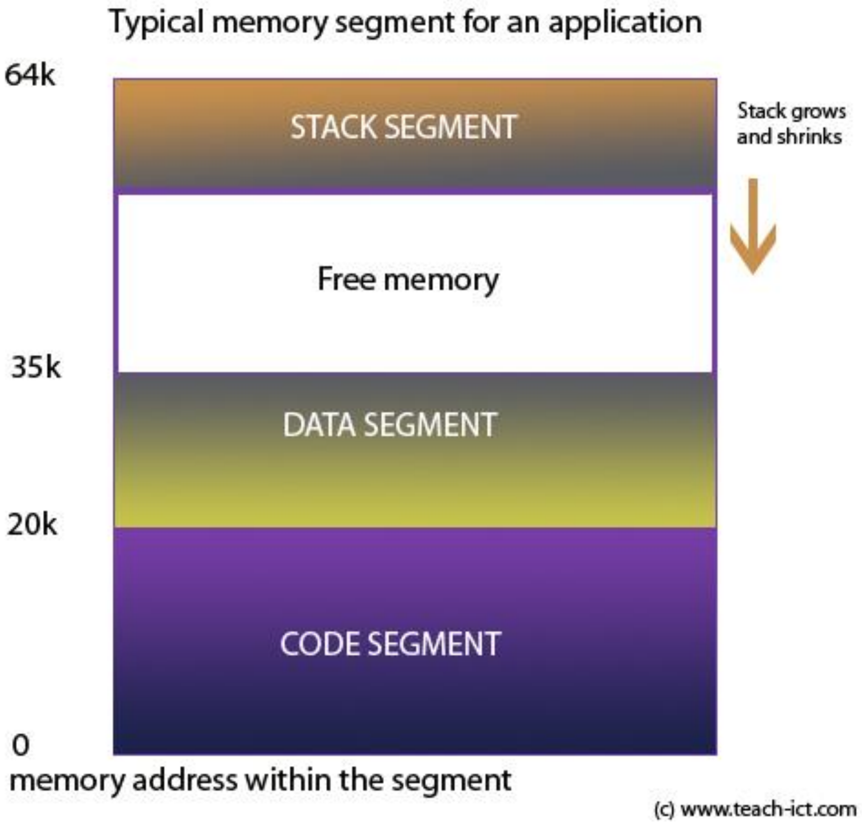
Topic

- Memory Functions
- Stack Operations
- Cache Memory
  - Overview, Types, Levels, Operation
- Storage Type
  - ROM, RAM, Disk, Hierarchical Storage
- Conclusion



Memory Functions

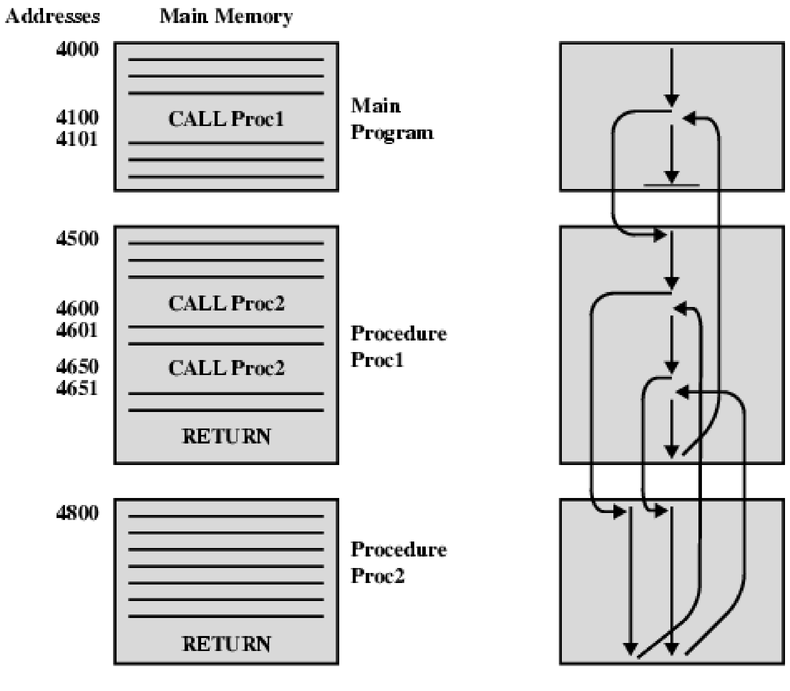
- Memory or Storage => Store
  - Programs
    - ➢Code Segment
  - Data
    - ➢Data Segment
  - States or Status
    - ➢Stack Segment
- Program / Procedure Calling
  - Program can call sub-programs (Functions)
  - Can be nested
  - Keep Status or Calling Address
    - Using **Stack** Area in Main Memory



(c) www.teach-ict.com

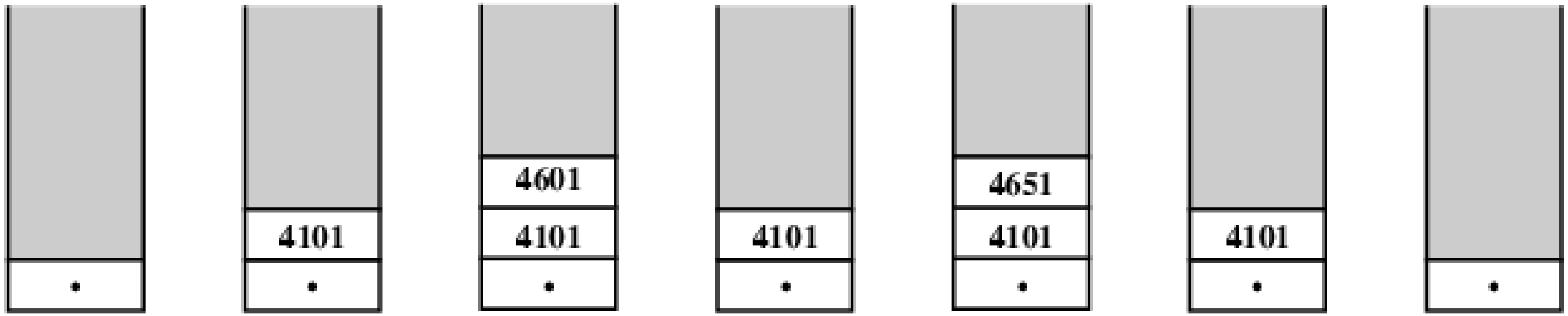
Stack Operation

- Stack in a defined area in Main Memory
- Keep Status or Addresses of calling programs
- Operation Style
  - Last-In First-Out
  - First-In Last-Out
- Controlled by "Stack Pointer"
- Can be used for Parameters Passing
  - Between Calling and Called programs



(a) Calls and returns

(b) Execution sequence



(a) Initial stack contents

(b) After CALL Proc1

(c) Initial CALL Proc2

(d) After RETURN

(e) After CALL Proc2

(f) After RETURN

(g) After RETURN

Parameters Passing by Stack

Structure of the stack in a function call

main:

// z = sumRange(A, x, y)

Pass (push) parameters on stack

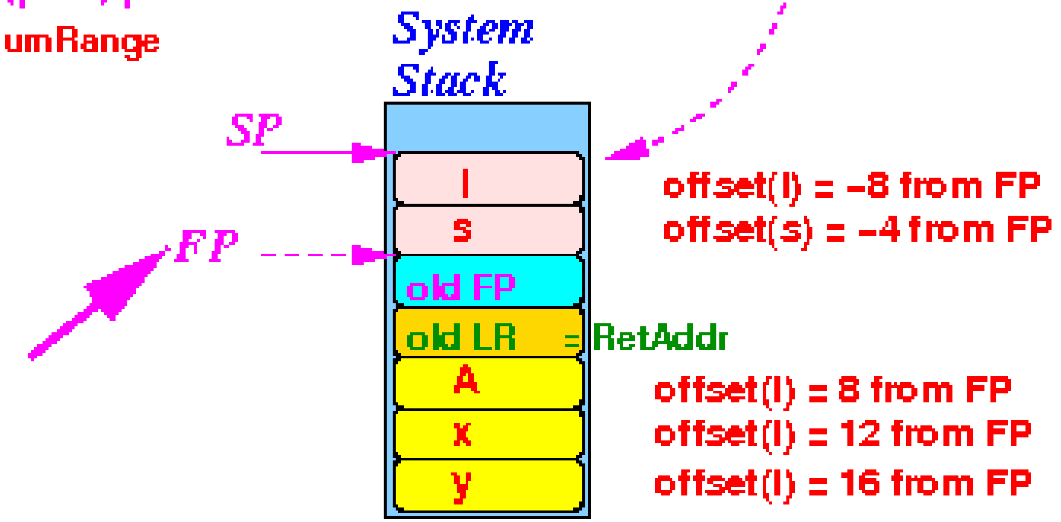
bl sumRange

RetAddr:

// sumRange subroutine

sumRange:

Allocate space for local variables



offset(l) = -8 from FP

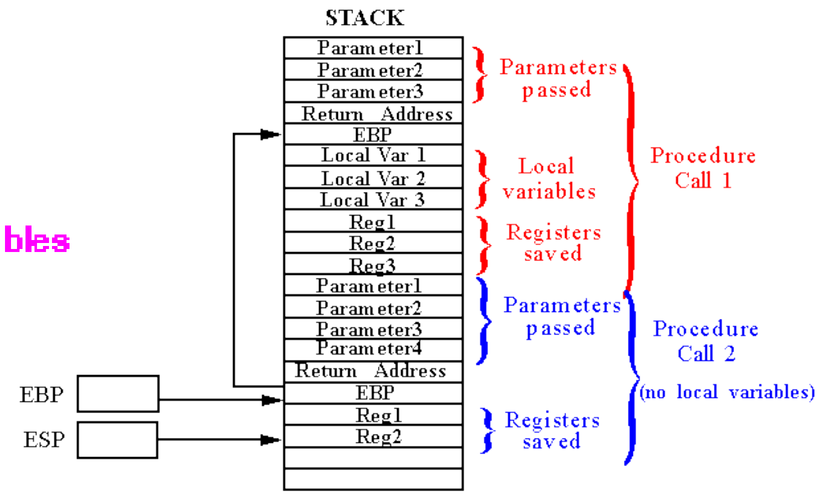
offset(s) = -4 from FP

offset(l) = 8 from FP

offset(l) = 12 from FP

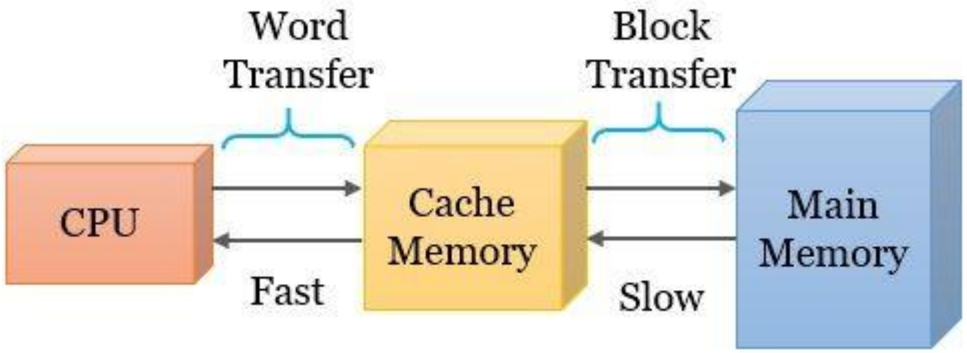
offset(l) = 16 from FP

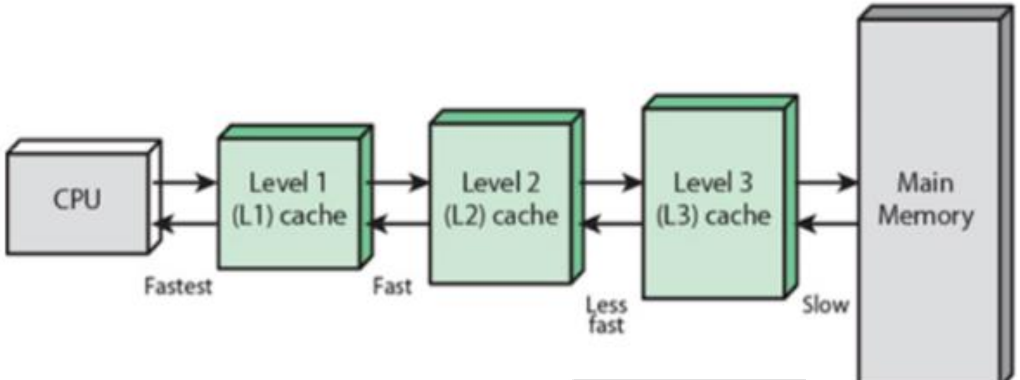
One call frame created per procedure call

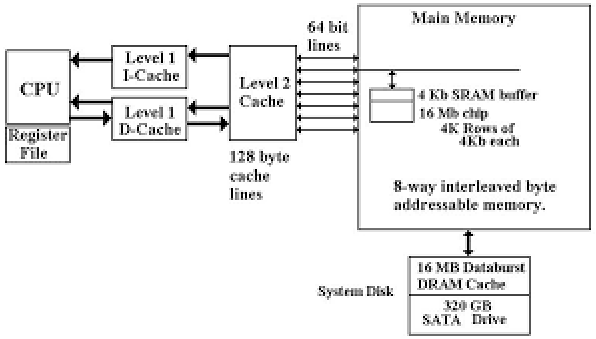


Cache Memory

- Cache Memory uses to improve computer performance
  - CPU is higher speed
  - Main Memory is much slower than CPU
  - CPU waits Memory => Slow
- Cache
  - Locate: Between CPU and Main Memory
  - Speed: Between CPU and Main Memory
- Cache => Fast => Small Amount
- May be located on CPU chip or module
- Can be multiple levels

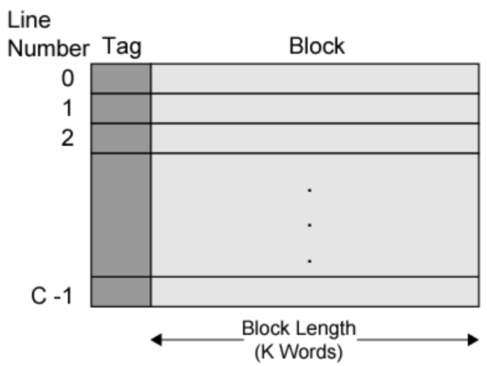




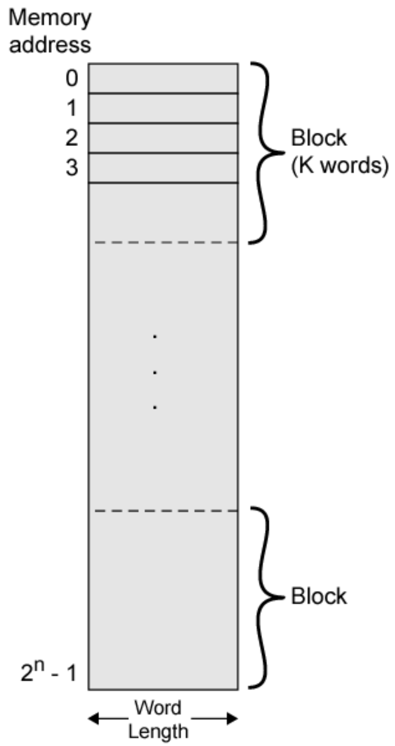


Cache/Main Memory Structure

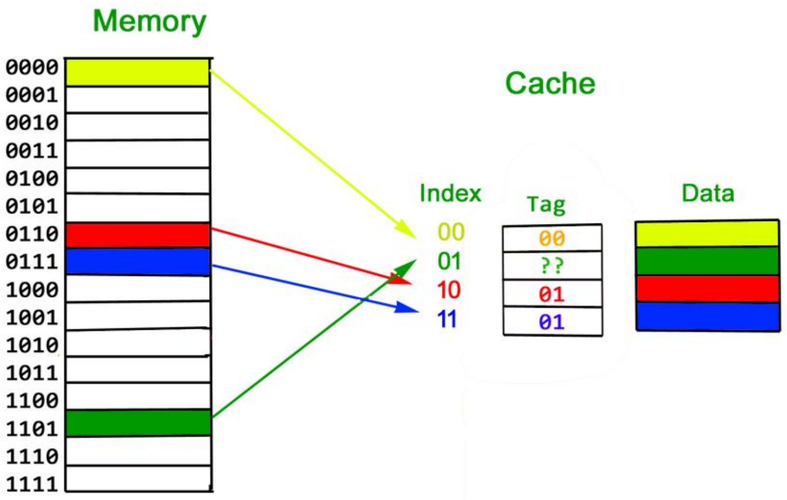
- Cache is less than Main Memory
  - Can not hold all memory
- Cache divided into Block
  - Called “Line”
  - Each Line has Tag
  - Tag uses to tell main memory address for the Line
    - By Mapping Functions

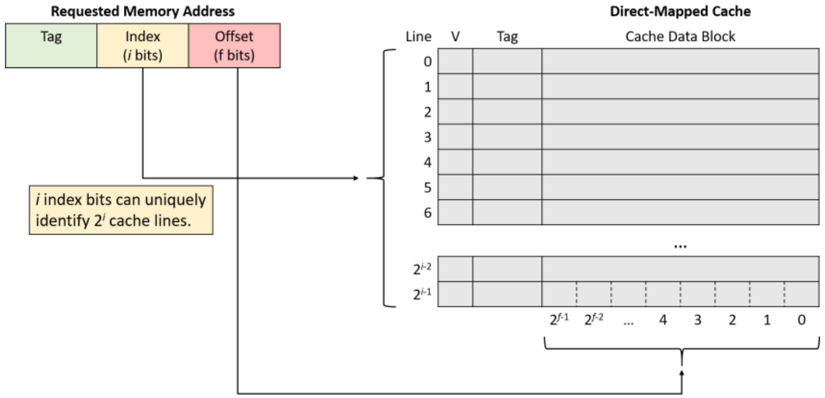


(a) Cache



(b) Main memory



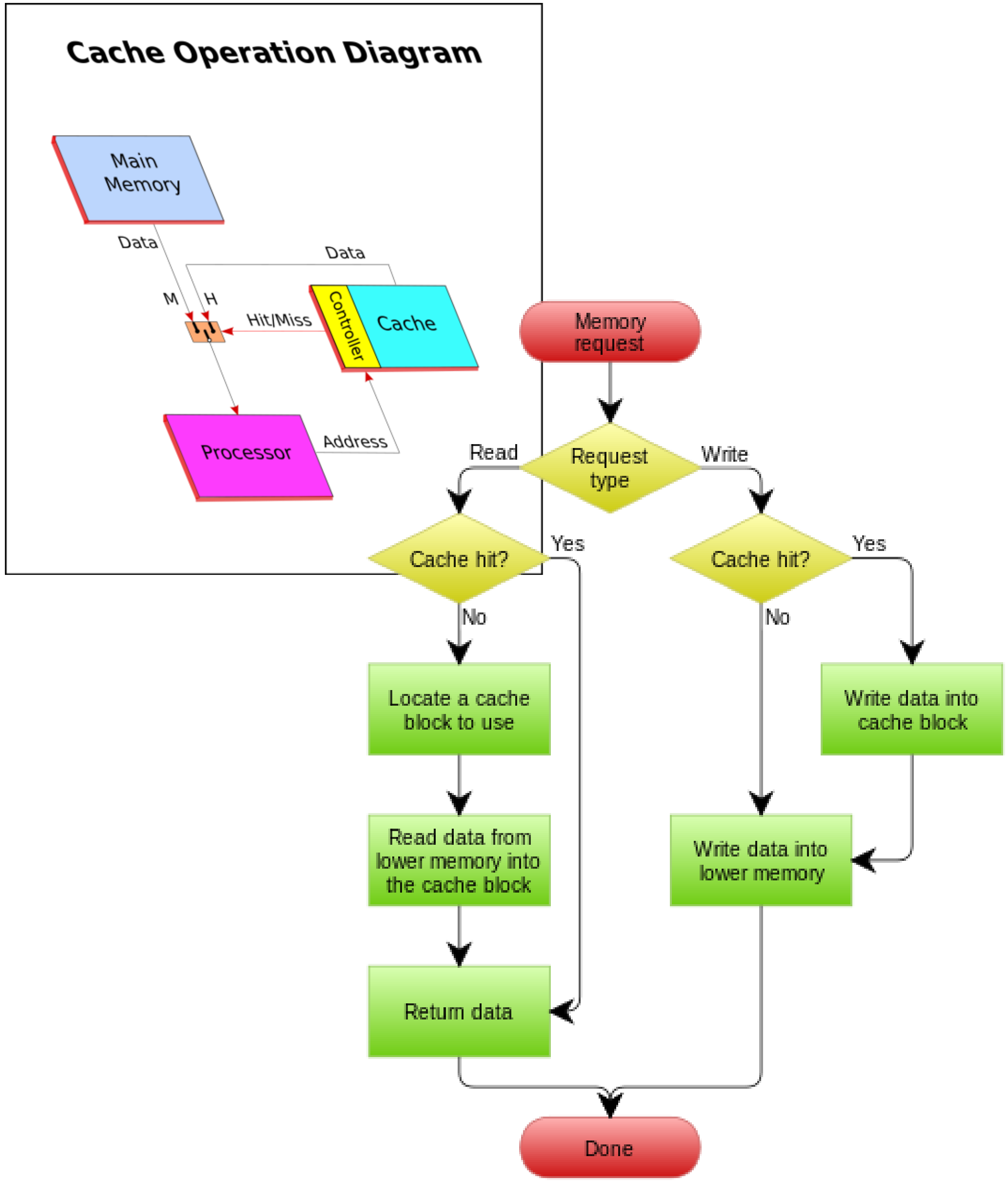


i index bits can uniquely identify 2^i cache lines.

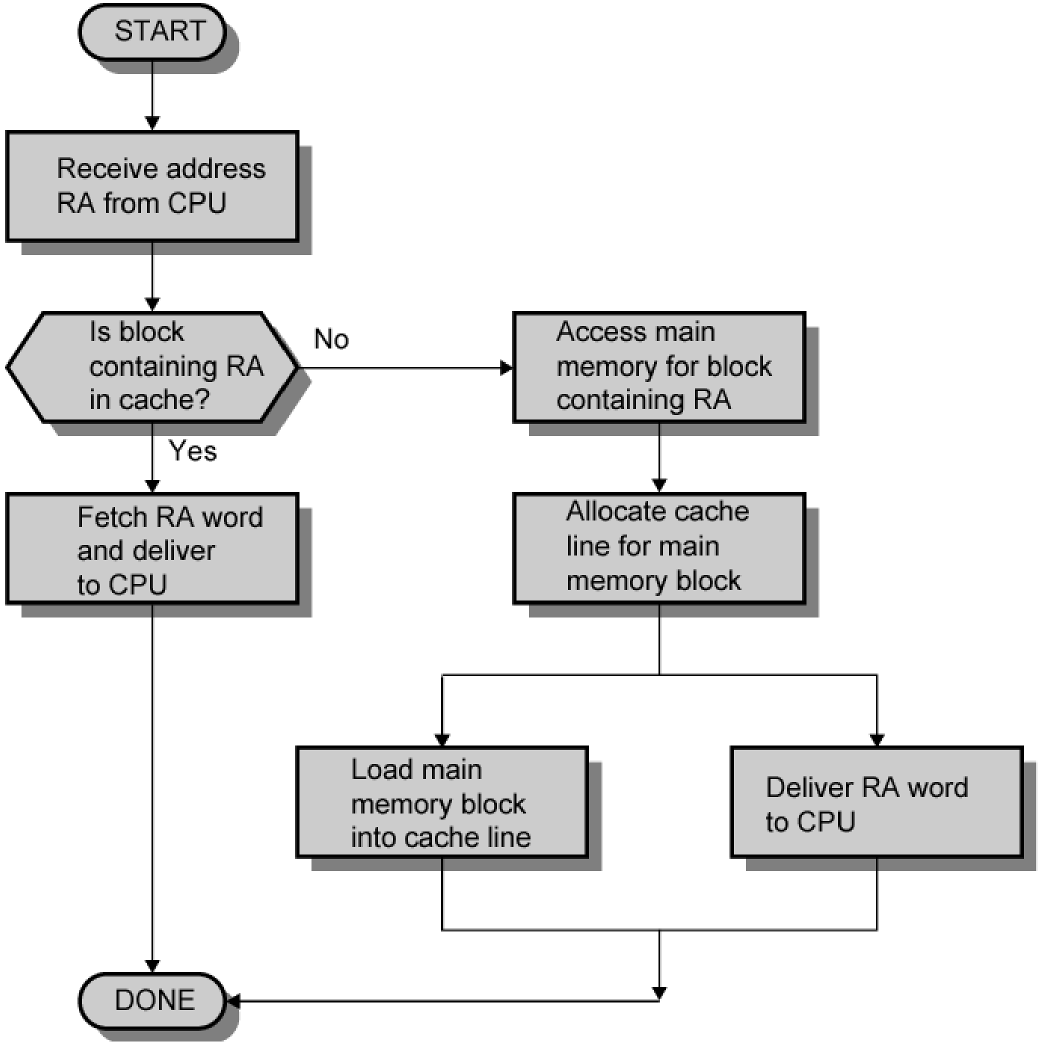
f offset bits can uniquely identify 2^f cache data block bytes.

Cache Operations

- CPU requests contents of memory location
- Check cache for this data
- If present, get from cache (fast)
    - Cache Hit
- If not present, read required block from main memory to cache
    - Cache Miss
- Then deliver from cache to CPU
- Cache includes tags to identify which block of main memory is in each cache slot



Cache Read Operations - Flowchart


START

Receive address RA from CPU

Is block containing RA in cache?

No

Access main memory for block containing RA

Yes

Fetch RA word and deliver to CPU

Allocate cache line for main memory block

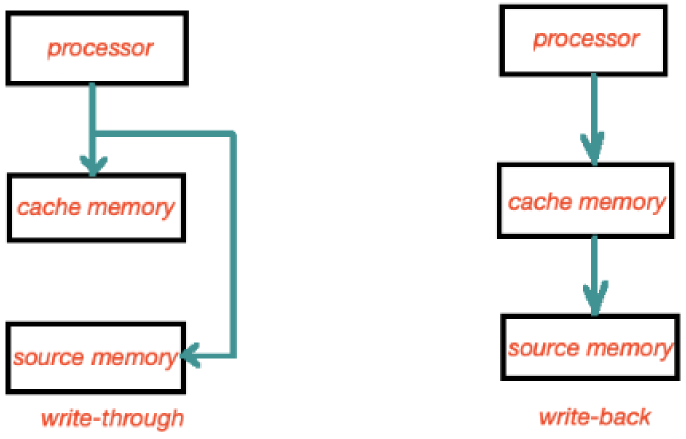
Load main memory block into cache line

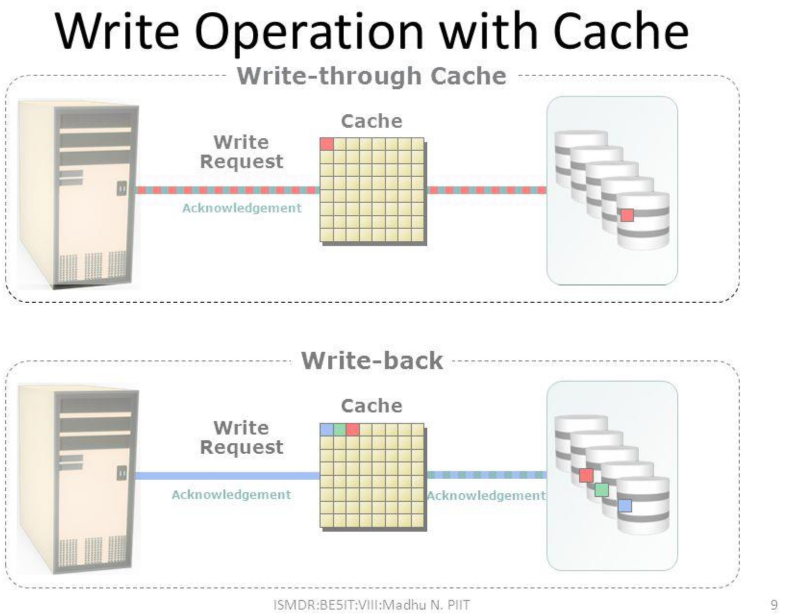
Deliver RA word to CPU

DONE

Cache Write Operations

- Cache Write Policy
  - Must not overwrite a cache block unless main memory is up to date
  - Multiple CPUs may have individual caches
  - I/O may address main memory directly
- Write Through
  - All writes go to main memory as well as cache
  - Multiple CPUs can monitor main memory traffic to keep local (to CPU) cache up to date
  - Lots of traffic
  - Slows down writes
  - Remember bogus write through caches!
- Write Back
  - Updates initially made in cache only
  - Update bit for cache slot is set when update occurs
  - If block is to be replaced, write to main memory only if update bit is set
  - Other caches get out of sync
  - I/O must access main memory through cache
  - N.B. 15% of memory references are writes

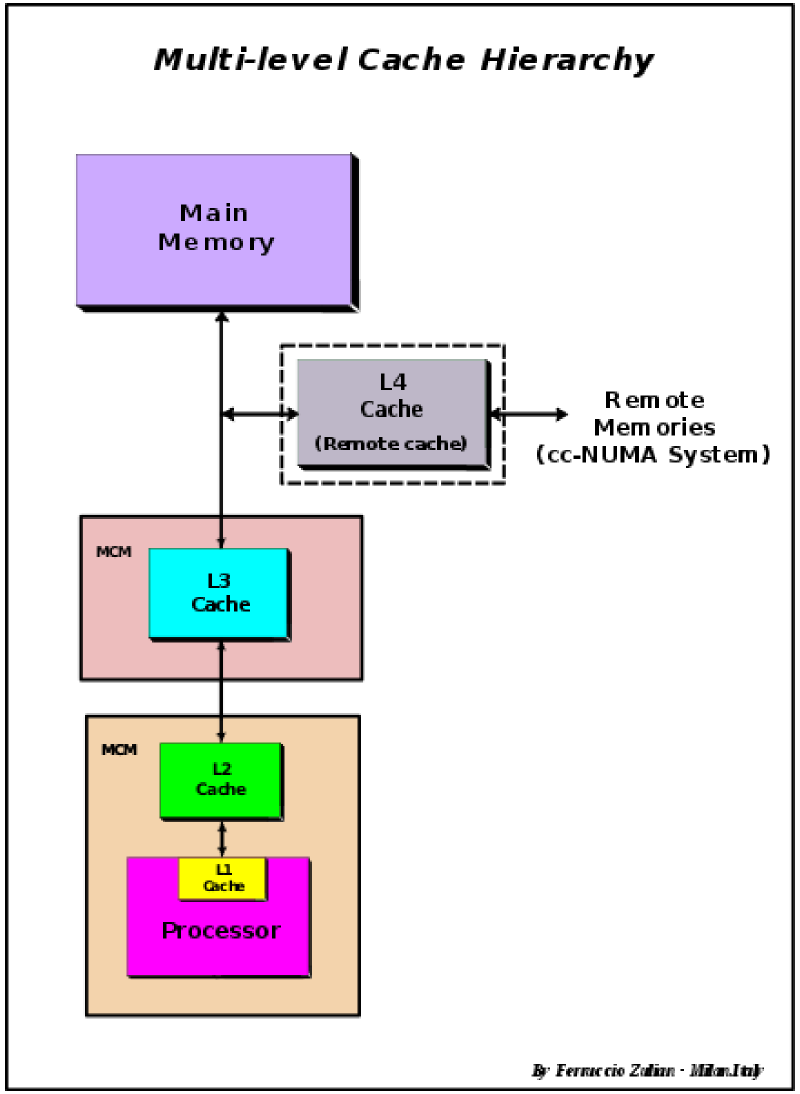




Write Operation with Cache

ISMDR:BESIT:VIII:Madhu N. PIIT 9

Multiple Level Cache - Hierarchy

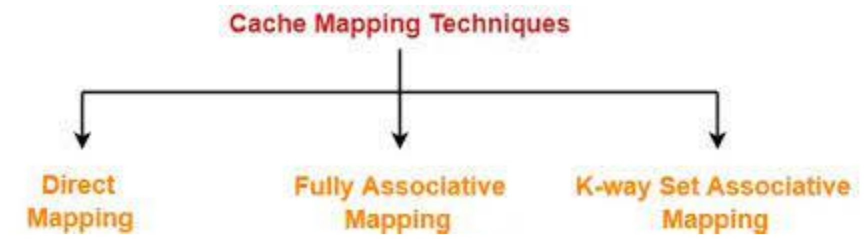


By Ferruccio Zulian - Milan.Italy

Cache Mapping Functions

- Map Cache Line to Main Memory Addresses
  - Compose of: Tag, Block, Word
  - Requested Address is in the cache or not?
    - Hit or Miss

- Direct Mapping
- Fully Associative Mapping
- Set Associative Mapping
  - K-Way Set Associative Mapping

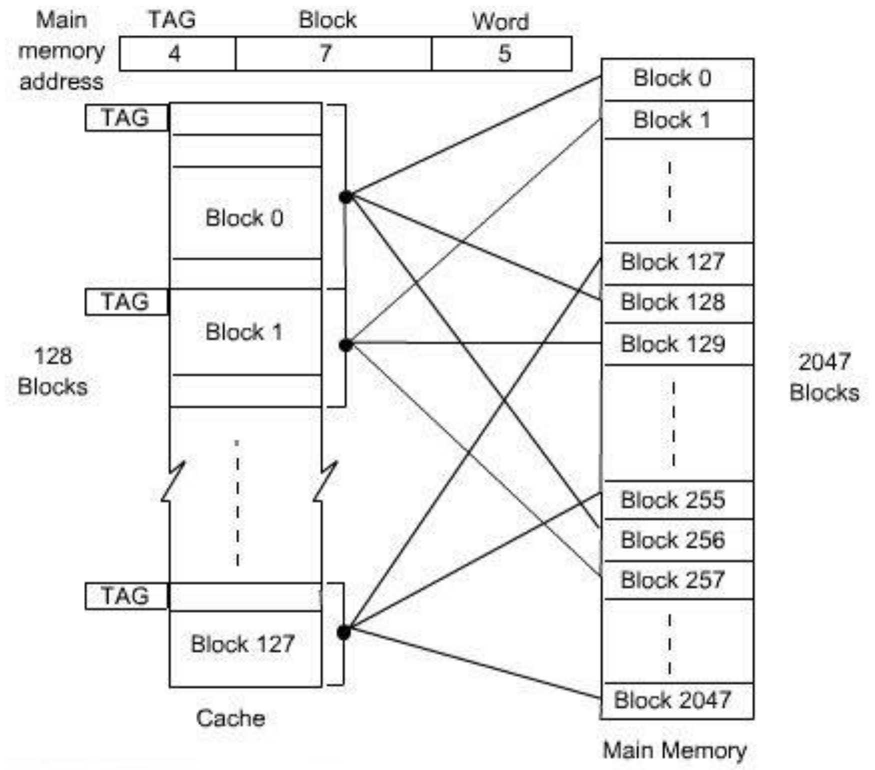


Cache Mapping Techniques

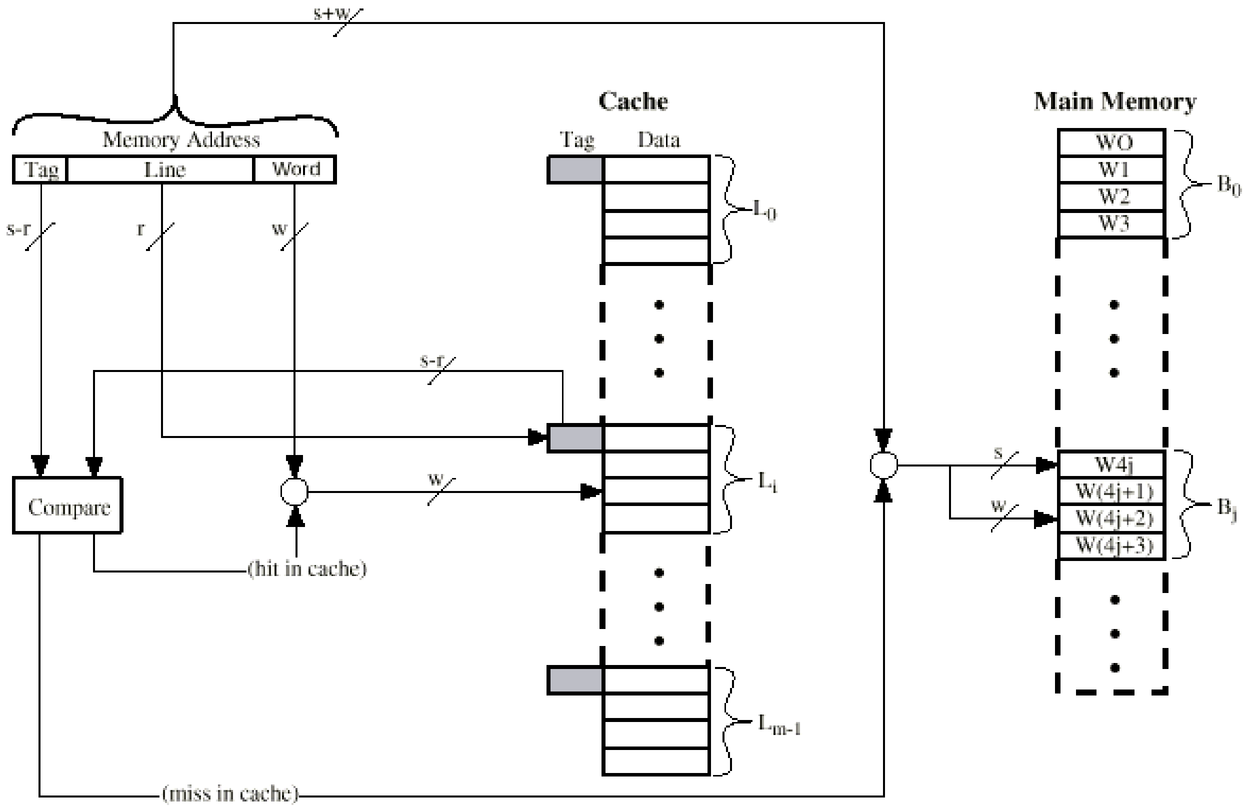
Direct Mapping

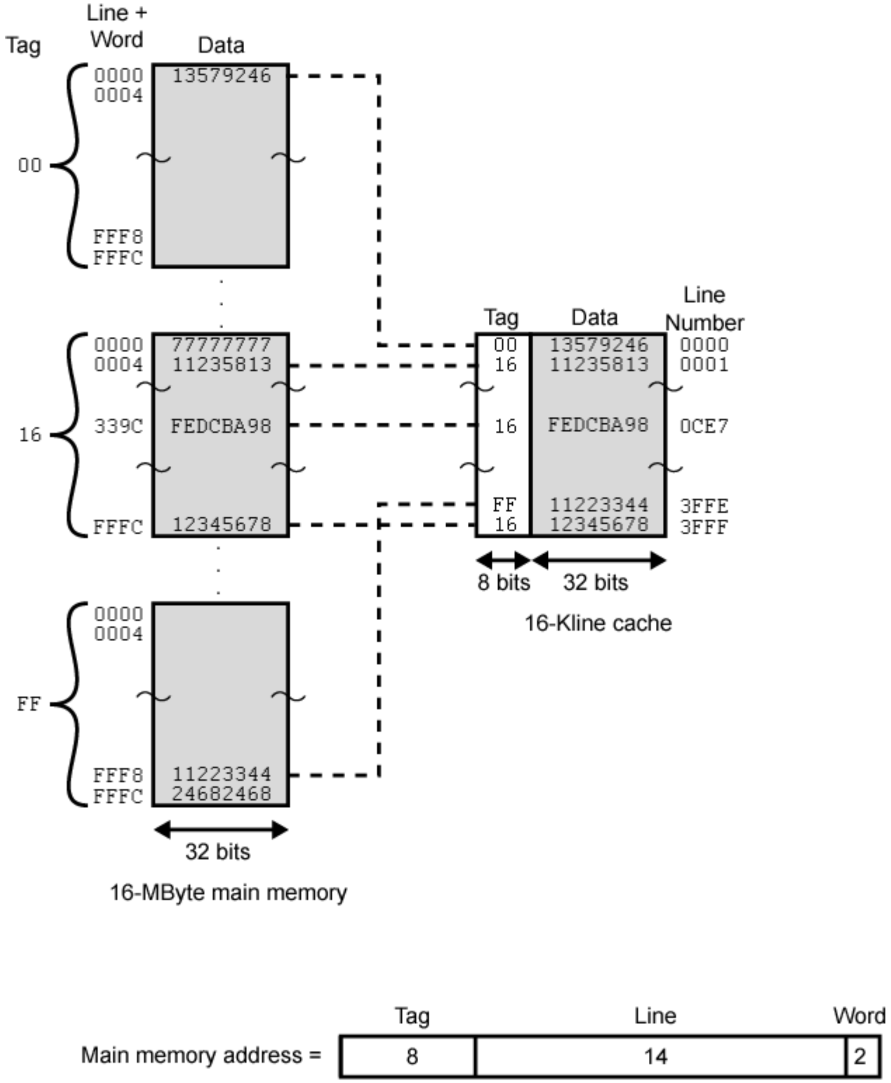
Fully Associative Mapping

K-way Set Associative Mapping

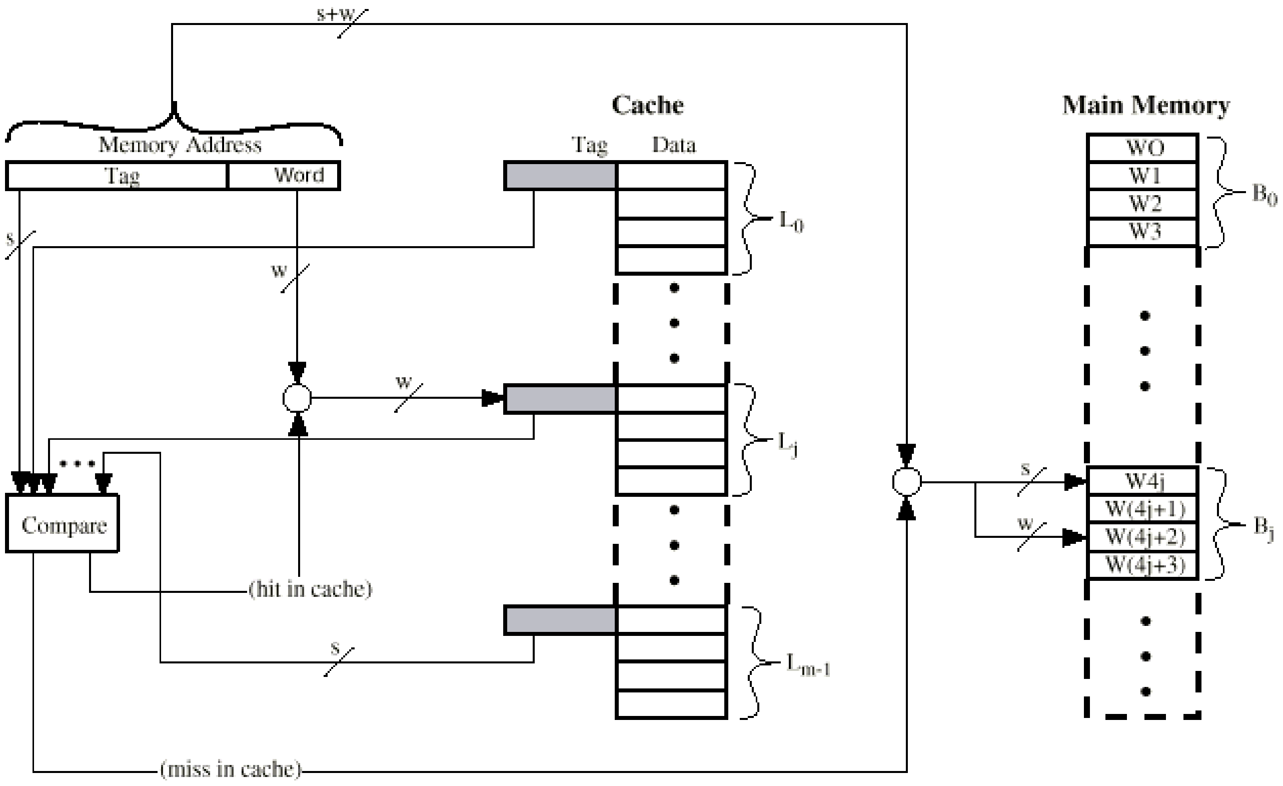


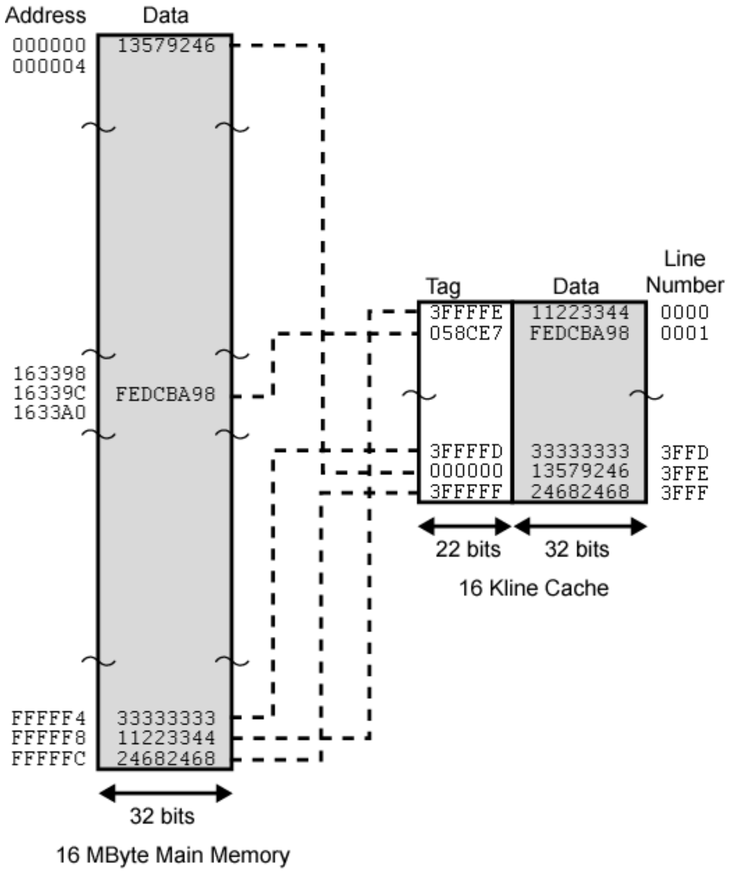
Direct Mapping





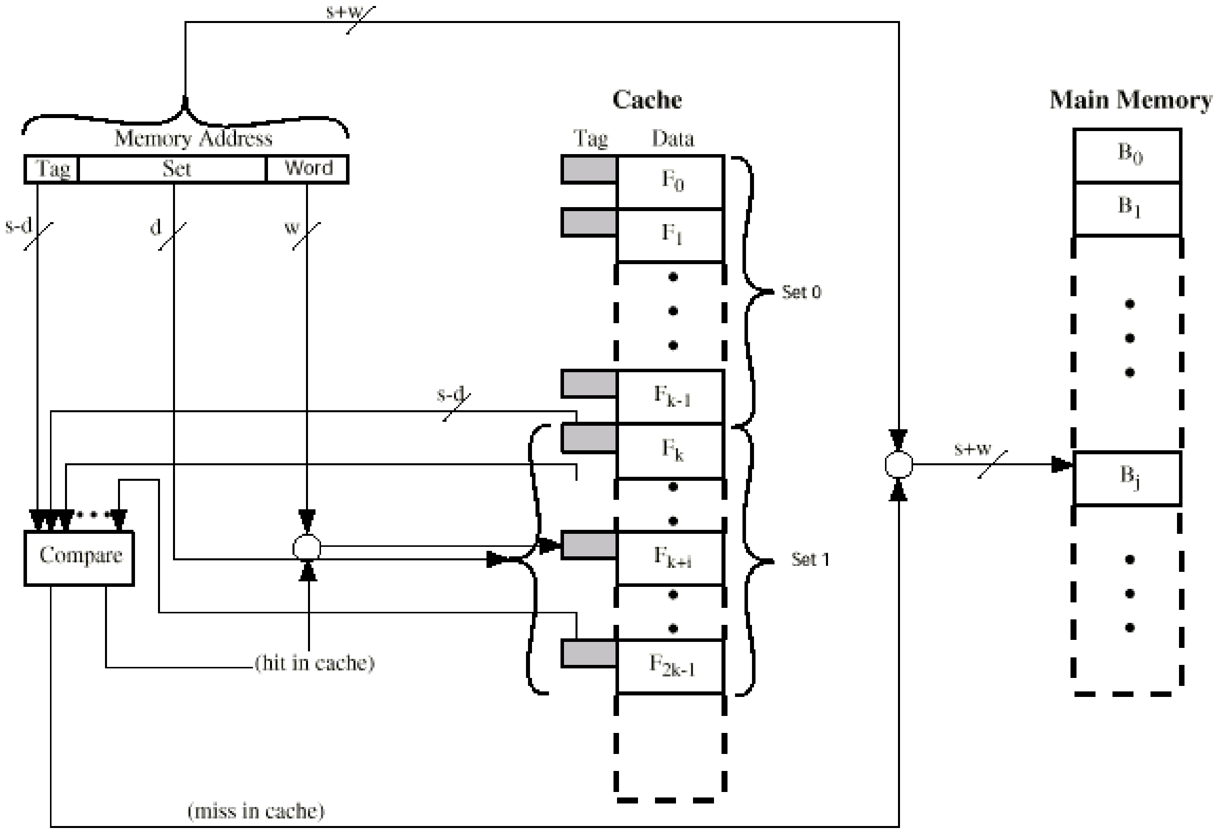
Fully Associative Mapping

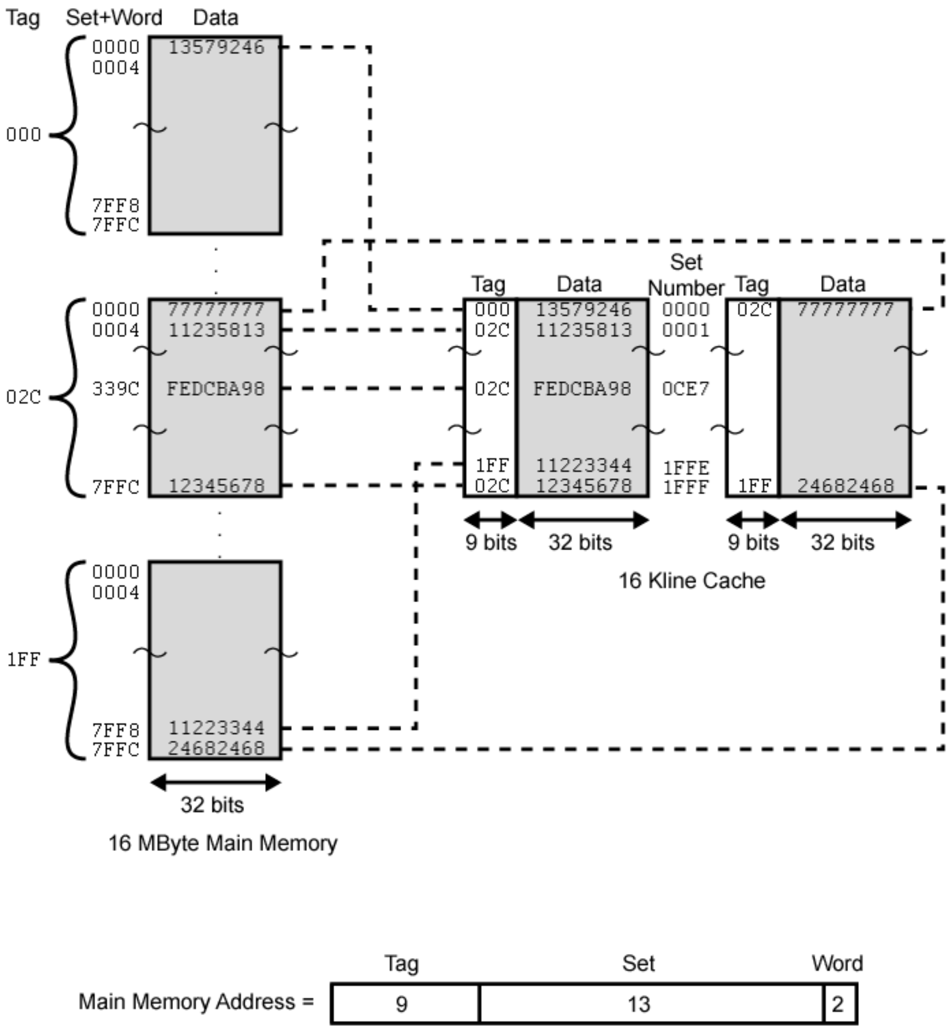




	Tag	Word
Main Memory Address =	22	2

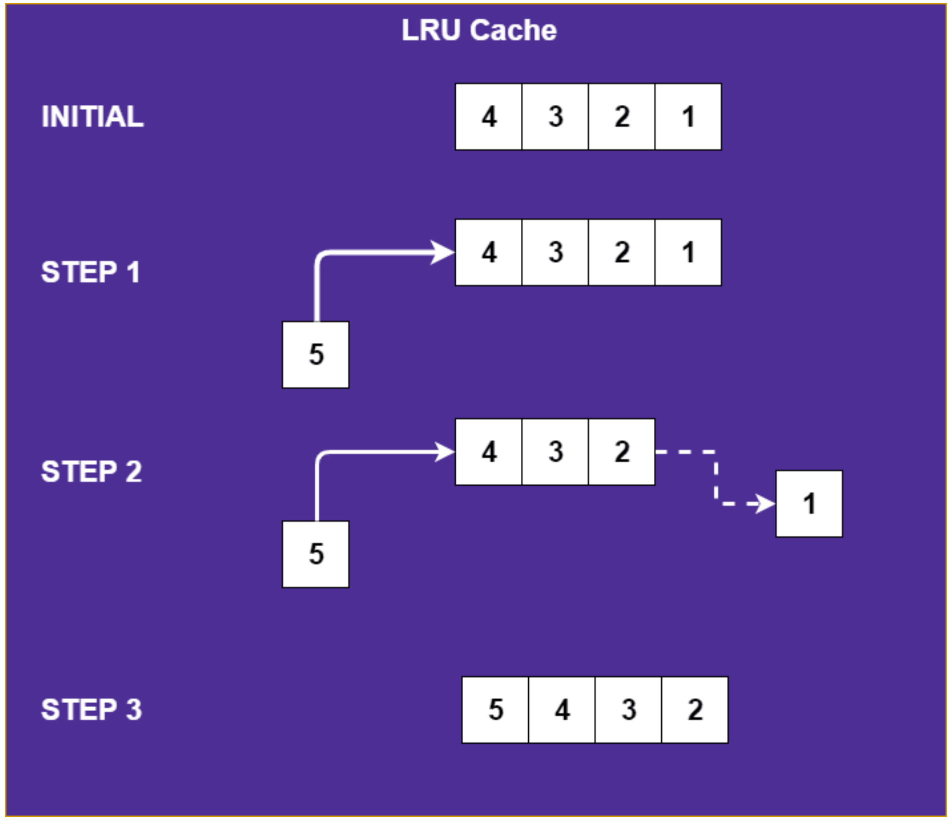
Two-Way Set Associative Mapping





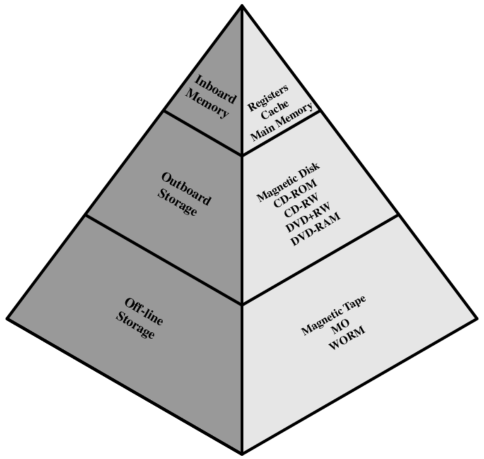
Cache Replacement Algorithms

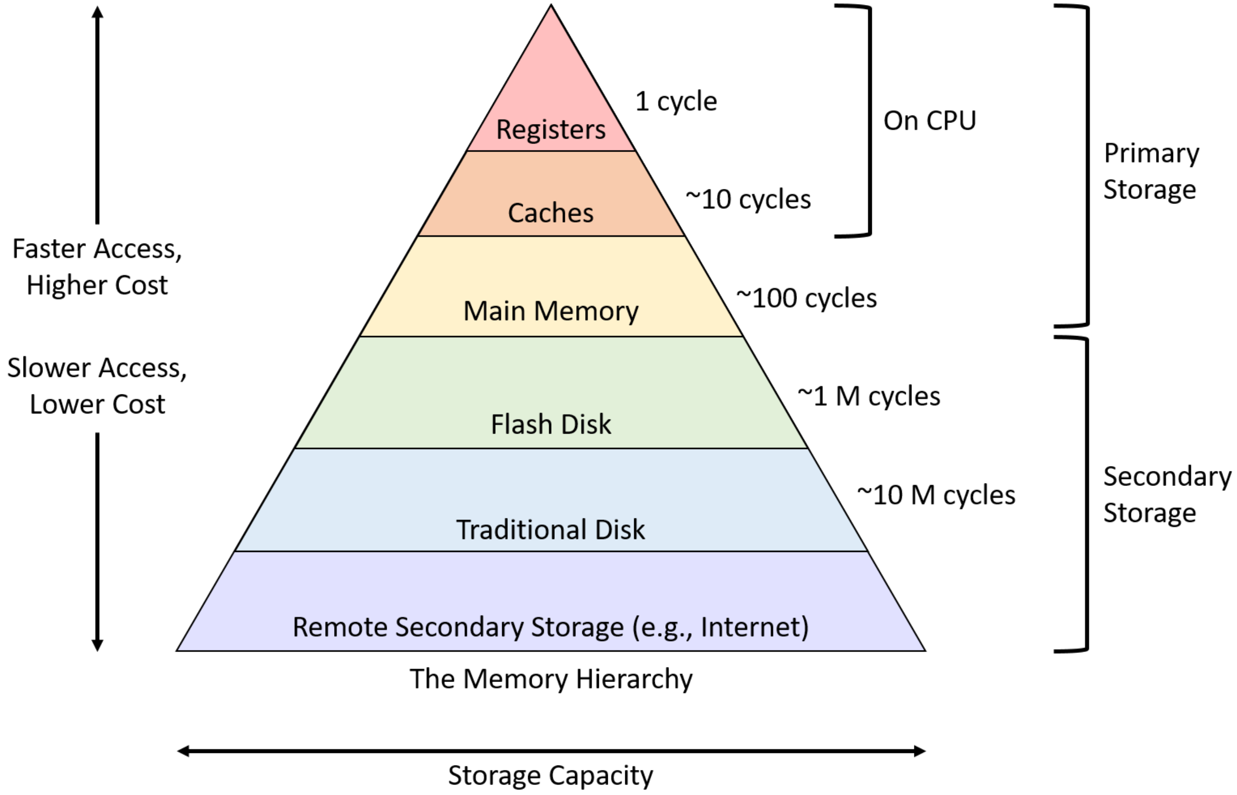
- Replacement Policy
- Decision Algorithms make cache free for putting main memory content into cache
- Hardware implemented algorithm (speed)
- Least Recently used (LRU)
- e.g. in 2 way set associative
  - Which of the 2 block is lru?
- First in first out (FIFO)
  - replace block that has been in cache longest
- Least frequently used
  - replace block which has had fewest hits
- Random



Storage Hierarchy

- Many types of Storage
- Speed => Cost => Capacity
- Forming into Hierarchical Storage
  - Registers
    - In CPU
  - Internal or Main memory
    - May include one or more levels of cache
    - “RAM”
  - External memory
    - Backing store
- List
  - Registers
  - L1 Cache
  - L2 Cache
  - L3 Cache
  - Main memory
  - Disk cache
  - Solid State Disk (SSD)
  - Hard Disk
  - Optical Disk
  - Tape

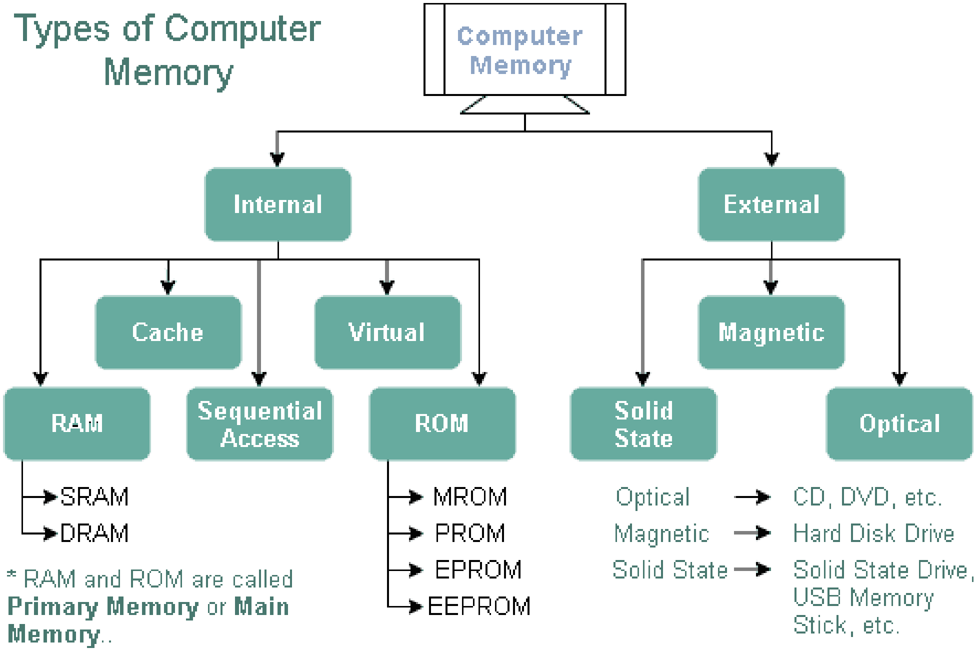




The Memory Hierarchy

Storage Types

- Internal Storage
  - Registers
  - Cache
  - Main Memory
- Storage Types
  - ROM – Read Only Memory
  - RAM – Random Access Memory
  - Static RAM – Registers, Cache
  - Dynamic RAM – Main Memory
  - Flash Memory – SSD
  - Hard Disk
  - Optical Disk
  - Tape

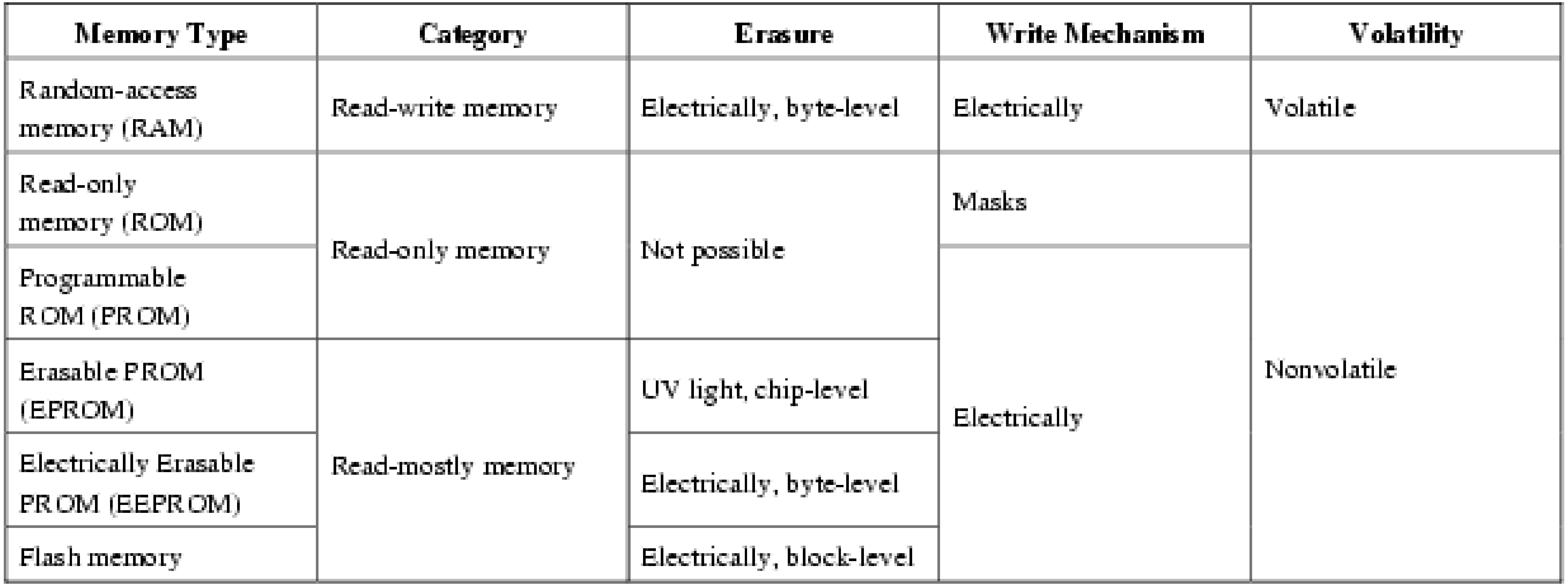


* RAM and ROM are called **Primary Memory** or **Main Memory**..

Physical Storage Types

- Semiconductor
  - RAM, ROM, etc.
- Magnetic
  - Disk & Tape
- Optical
  - CD & DVD
- Others
  - Bubble
  - Hologram

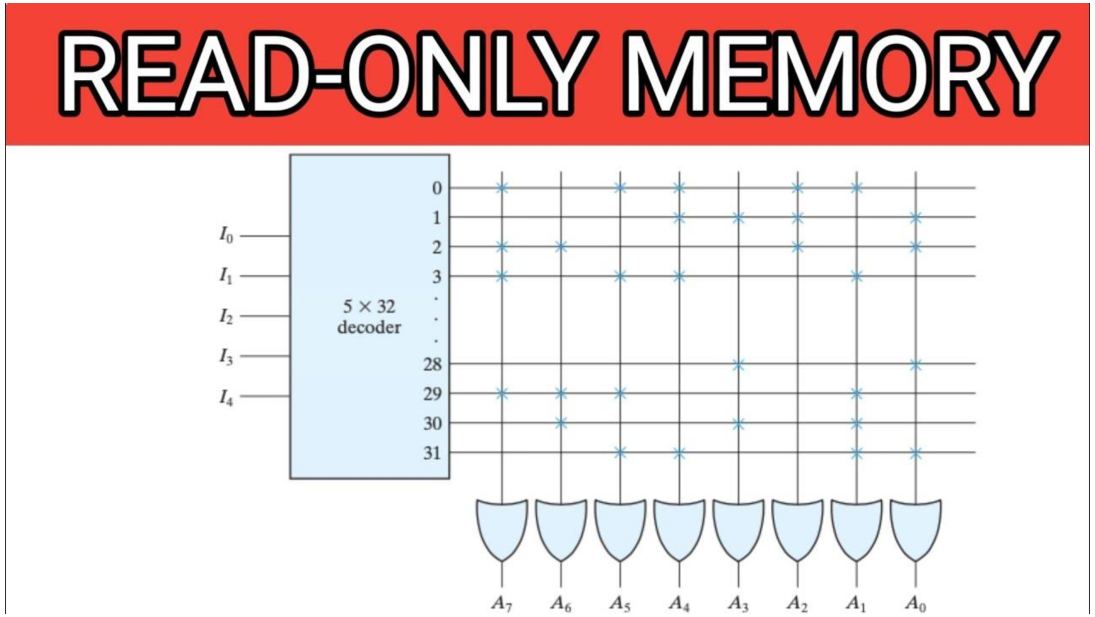
Semiconductor Memory Types



Memory Type	Category	Erasure	Write Mechanism	Volatility
Random-access memory (RAM)	Read-write memory	Electrically, byte-level	Electrically	Volatile
Read-only memory (ROM)	Read-only memory	Not possible	Masks	Nonvolatile
Programmable ROM (PROM)			Electrically	
Erasable PROM (EPROM)	Read-mostly memory	UV light, chip-level		
Electrically Erasable PROM (EEPROM)		Electrically, byte-level		
Flash memory		Electrically, block-level		

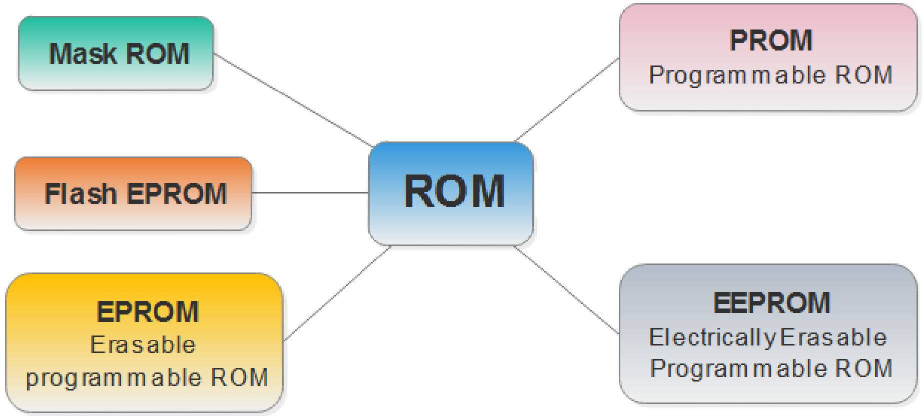
Read Only Memory (ROM)

- Permanent storage
  - Nonvolatile
- Microprogramming (see later)
- Library subroutines
- Systems programs (BIOS)
- Function tables



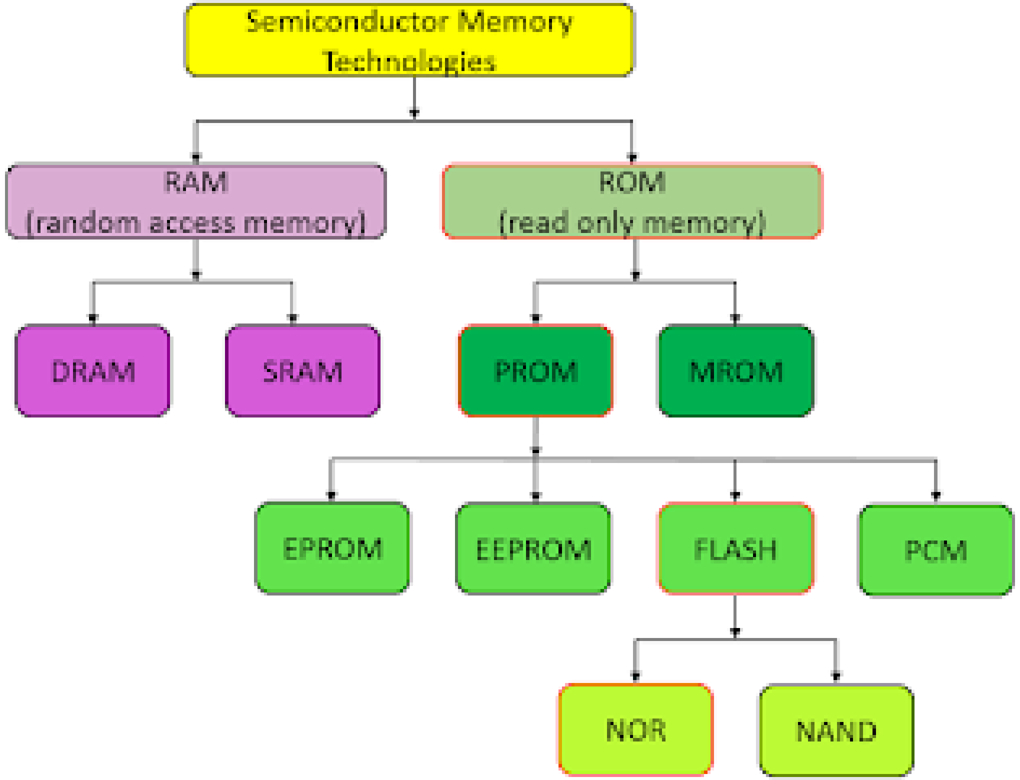
Types of ROM

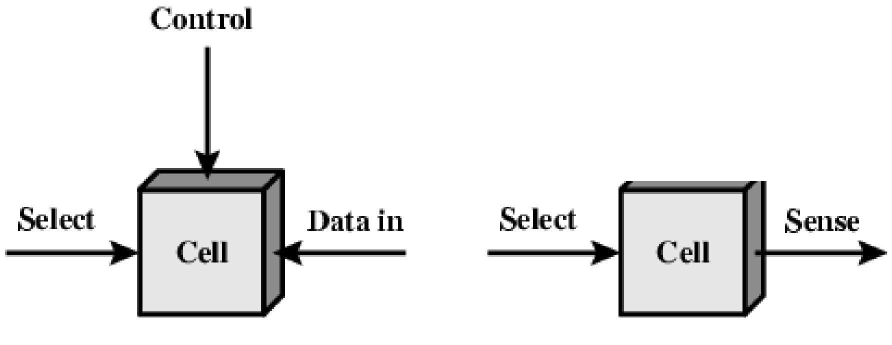
- ROM - Written during manufacture
  - Very expensive for small runs
  - Mask ROM
- Programmable (once)
  - PROM
  - Needs special equipment to program
- Read “mostly”
  - Erasable Programmable (EPROM)
    - Erased by UV
  - Electrically Erasable (EEPROM)
    - Takes much longer to write than read
 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


Mask ROM
Flash EPROM
EPROM
Erasable programmable ROM
ROM
PROM
Programmable ROM
EEPROM
Electrically Erasable Programmable ROM

Semiconductor Memory Types

- ROM – Read Only Memory
- RAM
  - Misnamed as all semiconductor memory is random access
  - Read/Write
  - Volatile
  - Temporary storage
  - Static or dynamic
- Static RAM - SRAM
- Dynamic RAM – DRAM

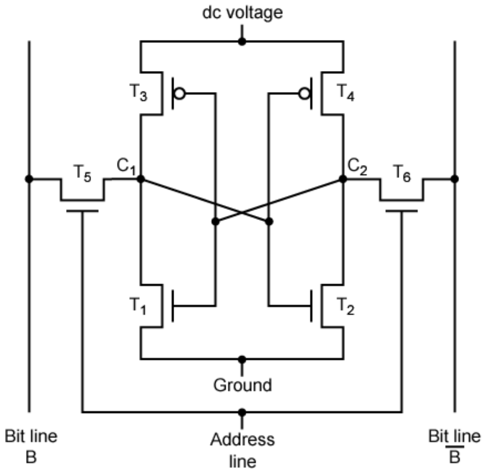


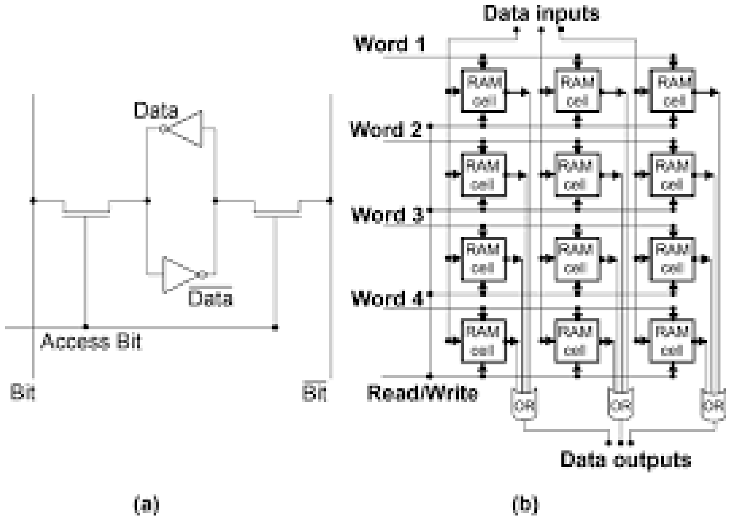


(a) Write

(b) Read

Static RAM

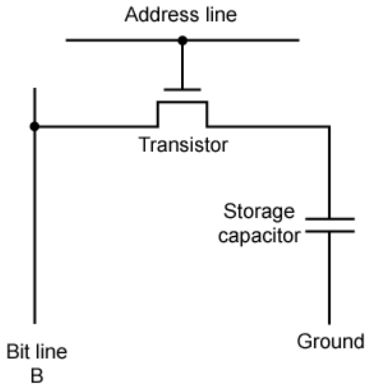




- Bits stored as on/off switches
- No charges to leak
- No refreshing needed when powered
- More complex construction
- Larger per bit
- More expensive
- Does not need refresh circuits
- Faster
- Cache
- Digital
  - Uses flip-flops

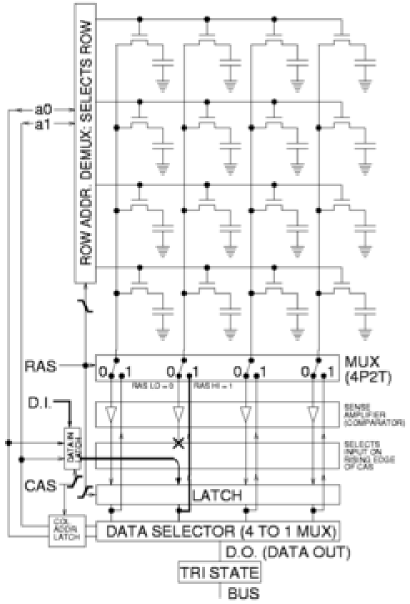
- SRAM Operation
- Transistor arrangement gives stable logic state
- State 1
  - C_1 high, C_2 low
  - T_1 T_4 off, T_2 T_3 on
- State 0
  - C_2 high, C_1 low
  - T_2 T_3 off, T_1 T_4 on
- Address line transistors T_5 T_6 is switch
- Write – apply value to B & compliment to B
- Read – value is on line B

Dynamic RAM



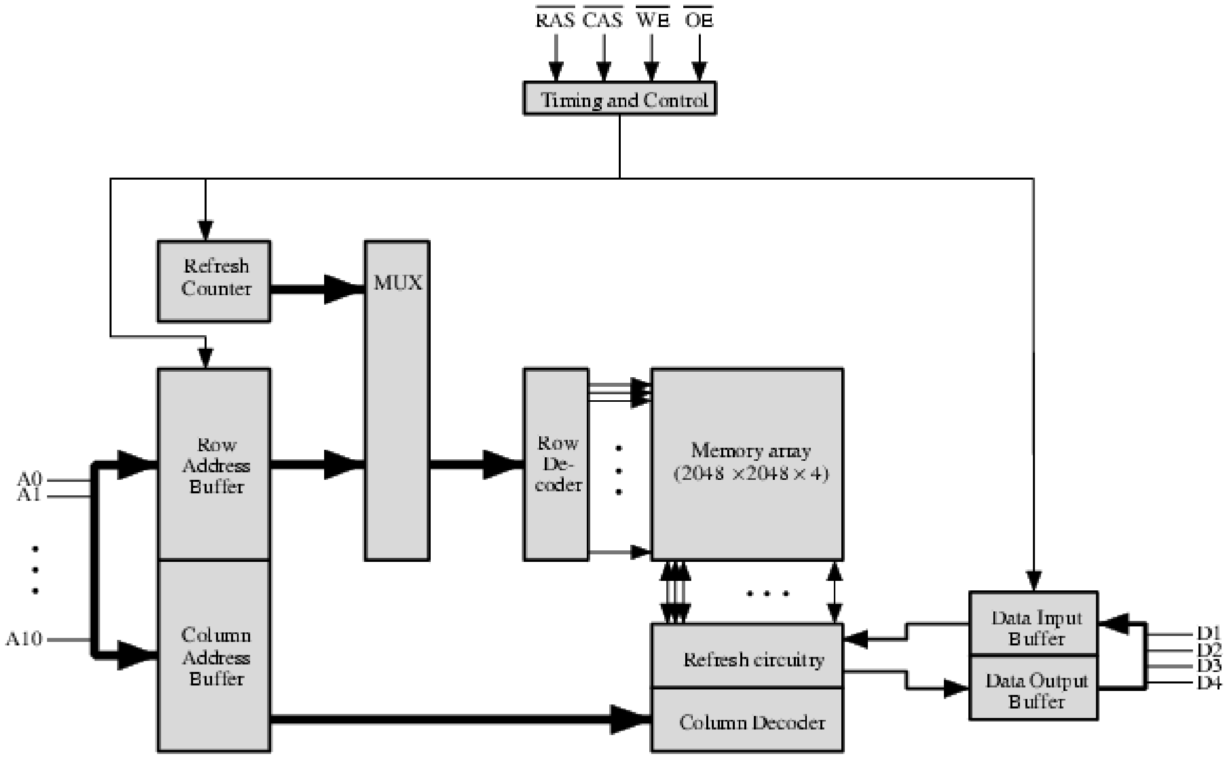
- DRAM
  - Bits stored as charge in capacitors
  - Charges leak
  - Need refreshing even when powered
  - Simpler construction
  - Smaller per bit
  - Less expensive
  - Need refresh circuits
  - Slower
  - Main memory
  - Essentially analogue
  - Level of charge determines value

- DRAM Operation
- Address line active when bit read or written
  - Transistor switch closed (current flows)
- Write
  - Voltage to bit line
    - High for 1 low for 0
  - Then signal address line
    - Transfers charge to capacitor
- Read
  - Address line selected
    - transistor turns on
  - Charge from capacitor fed via bit line to sense amplifier
    - Compares with reference value to determine 0 or 1
  - Capacitor charge must be restored


ROW ADDR. DEMUX: SELECTS ROW
a0
a1
RAS
0 1
RAS LO = 0
RAS HI = 1
MUX (4P2T)
SENSE AMPLIFIER (COMPARATOR)
D.I.
DATA IN LATCH
SELECTS INPUT ON RISING EDGE OF CAS
CAS
LATCH
COL ADDR LATCH
DATA SELECTOR (4 TO 1 MUX)
D.O. (DATA OUT)
TRI STATE
BUS

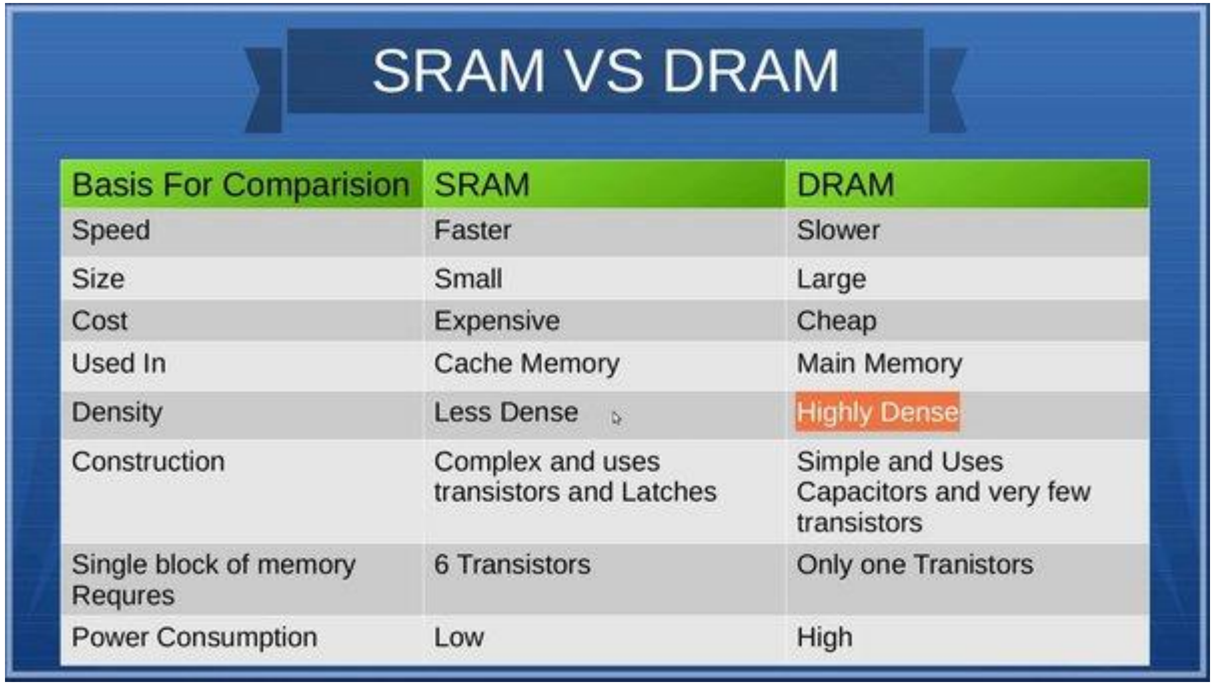
DRAM Refreshing

- Refresh circuit included on chip
- Disable chip
- Count through rows
- Read & Write back
- Takes time
- Slows down apparent performance



SRAM vs DRAM

- Both volatile
  - Power needed to preserve data
- Dynamic cell
  - Simpler to build, smaller
  - More dense
  - Less expensive
  - Needs refresh
  - Larger memory units
- Static
  - Faster
  - Cache

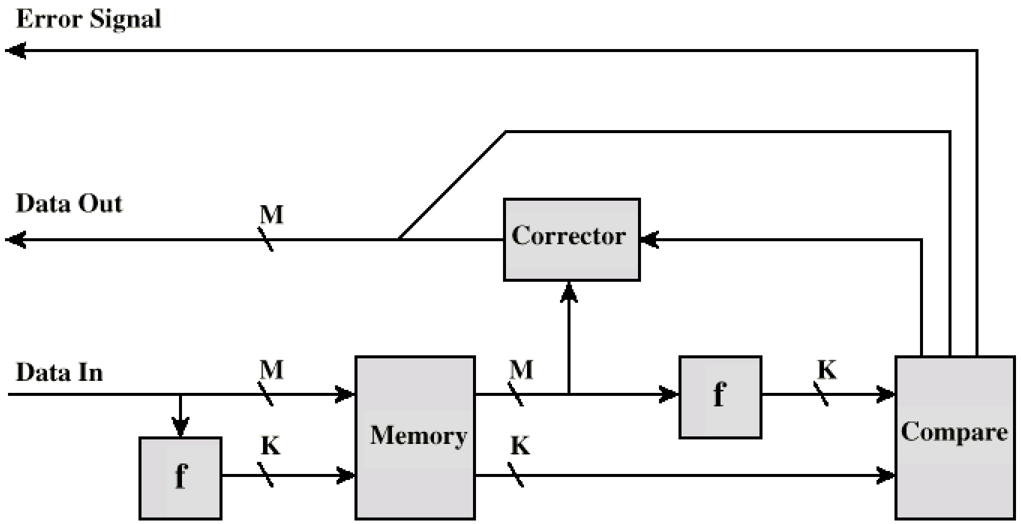


SRAM VS DRAM

Basis For Comparision	SRAM	DRAM
Speed	Faster	Slower
Size	Small	Large
Cost	Expensive	Cheap
Used In	Cache Memory	Main Memory
Density	Less Dense	Highly Dense
Construction	Complex and uses transistors and Latches	Simple and Uses Capacitors and very few transistors
Single block of memory Requres	6 Transistors	Only one Tranistors
Power Consumption	Low	High

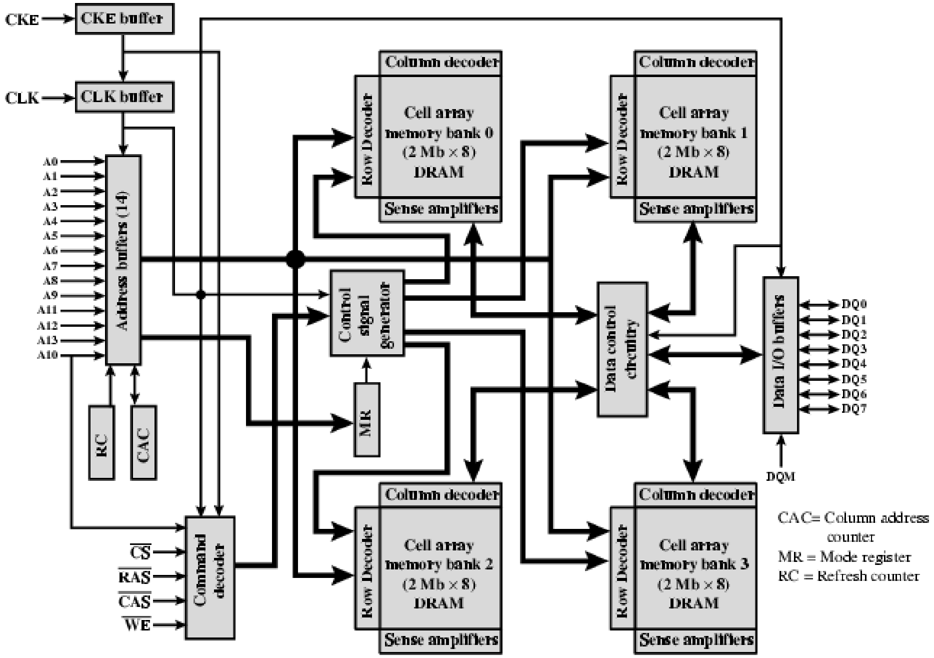
Memory Error Correction

- Hard Failure
  - Permanent defect
- Soft Error
  - Random, non-destructive
  - No permanent damage to memory
- Detected using Hamming error correcting code



Synchronous DRAM (SDRAM)

- 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)



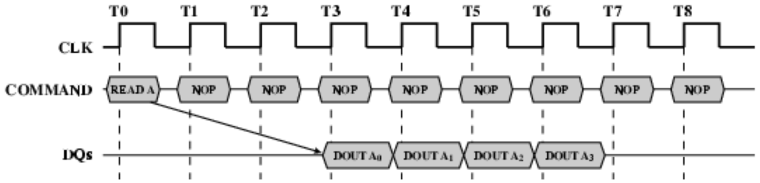
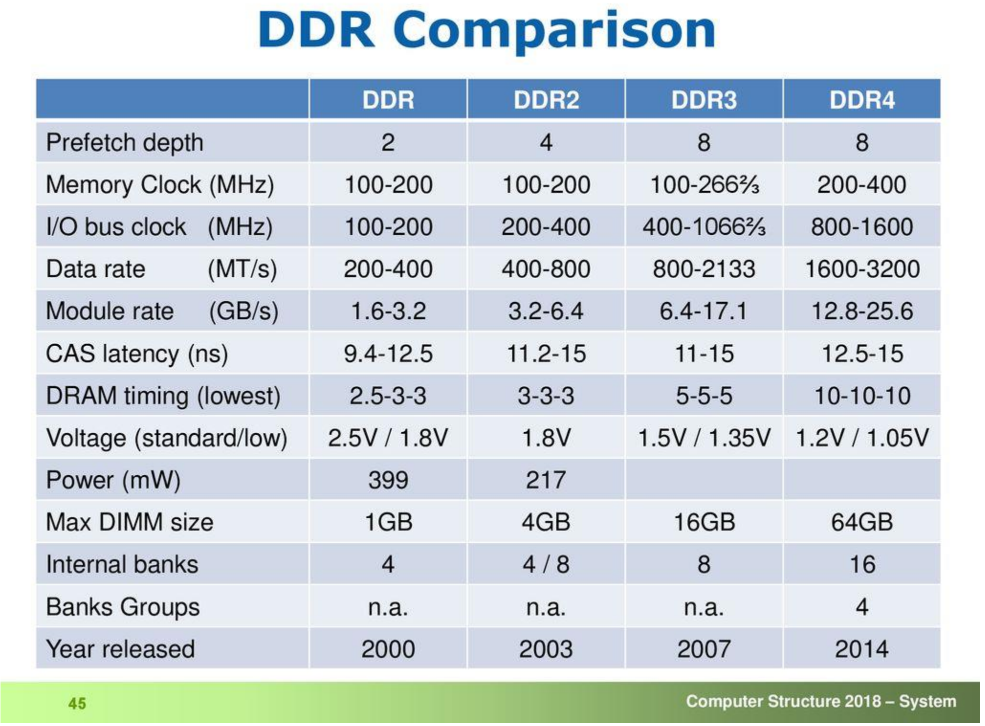


Figure 5.13 SDRAM Read Timing (Burst Length = 4, $\overline{CAS}$ latency = 2)

DDR SDRAM

- SDRAM can only send data once per clock
- Double-data-rate SDRAM can send data twice per clock cycle
  - Rising edge and falling edge



DDR Comparison

	DDR	DDR2	DDR3	DDR4
Prefetch depth	2	4	8	8
Memory Clock (MHz)	100-200	100-200	100-266⅔	200-400
I/O bus clock (MHz)	100-200	200-400	400-1066⅔	800-1600
Data rate (MT/s)	200-400	400-800	800-2133	1600-3200
Module rate (GB/s)	1.6-3.2	3.2-6.4	6.4-17.1	12.8-25.6
CAS latency (ns)	9.4-12.5	11.2-15	11-15	12.5-15
DRAM timing (lowest)	2.5-3-3	3-3-3	5-5-5	10-10-10
Voltage (standard/low)	2.5V / 1.8V	1.8V	1.5V / 1.35V	1.2V / 1.05V
Power (mW)	399	217		
Max DIMM size	1GB	4GB	16GB	64GB
Internal banks	4	4 / 8	8	16
Banks Groups	n.a.	n.a.	n.a.	4
Year released	2000	2003	2007	2014

45 Computer Structure 2018 – System

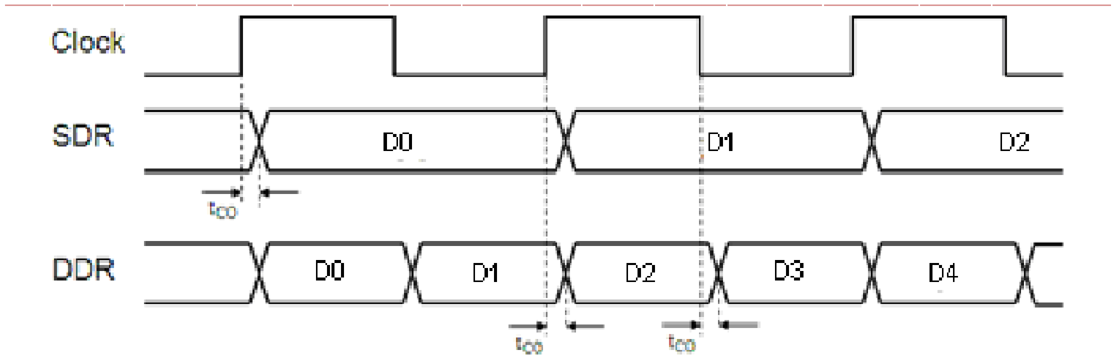
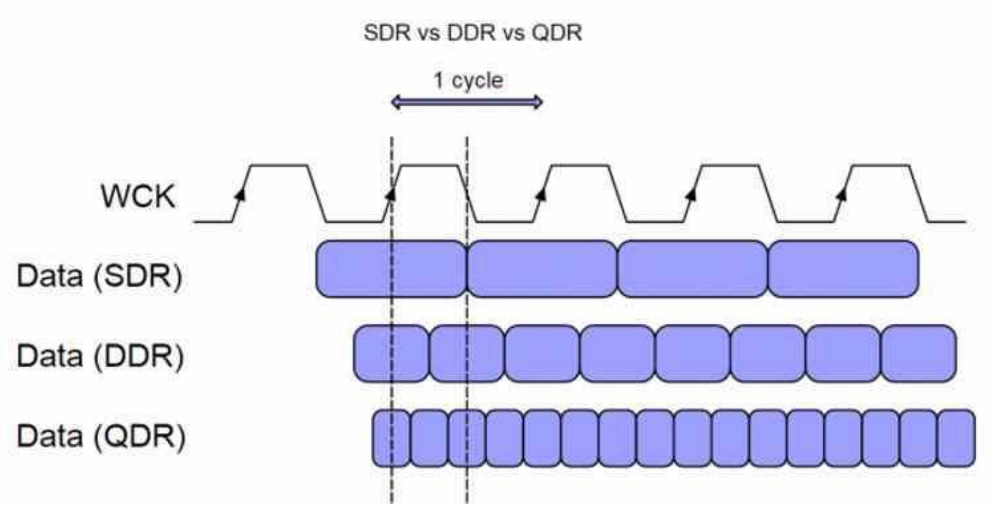


Figure 1: SDR and DDR timing diagram

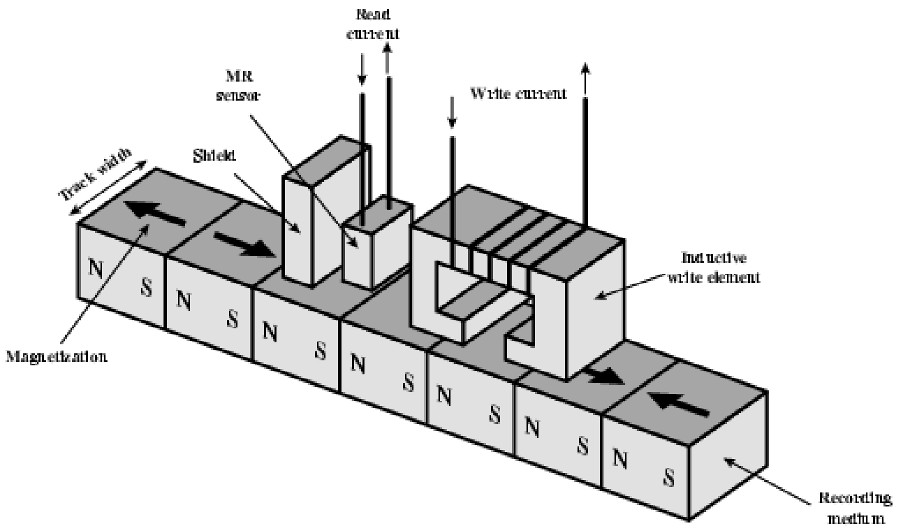


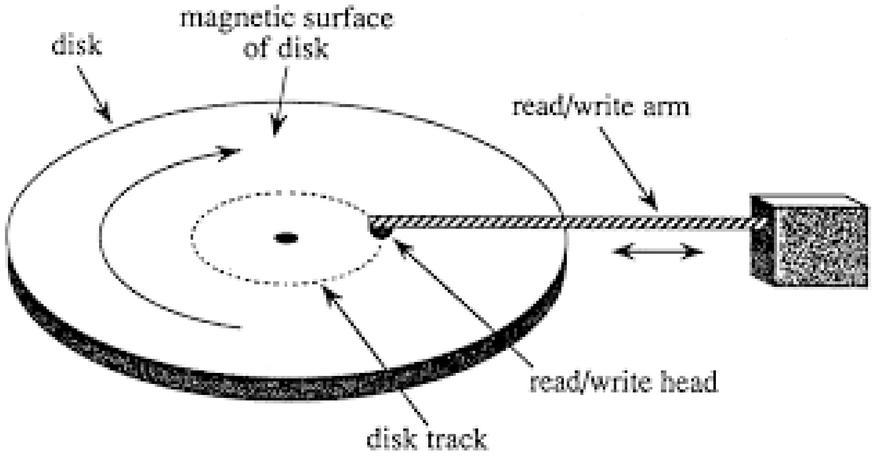
Types of External Memory

- Magnetic Disk
  - RAID
  - Removable
- Optical
  - CD-ROM
  - CD-Recordable (CD-R)
  - CD-R/W
  - DVD
- Magnetic Tape

Magnetic Disk

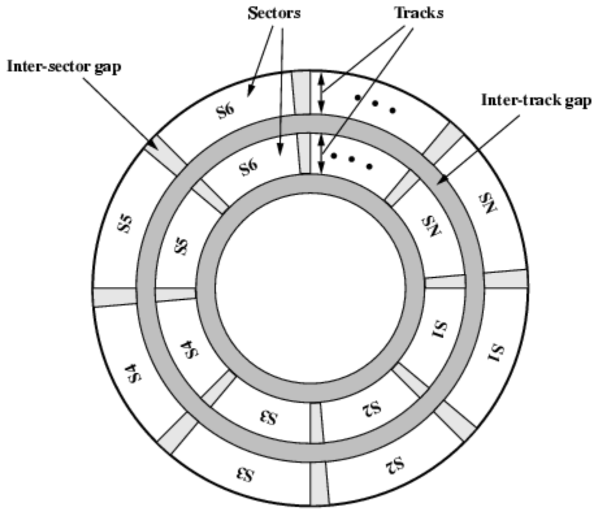
- Disk substrate coated with magnetizable material (iron oxide…rust)
- Substrate used to be aluminium
- Now glass
  - Improved surface uniformity
    - Increases reliability
  - Reduction in surface defects
    - Reduced read/write errors
  - Lower flight heights (See later)
  - Better stiffness
  - Better shock/damage resistance

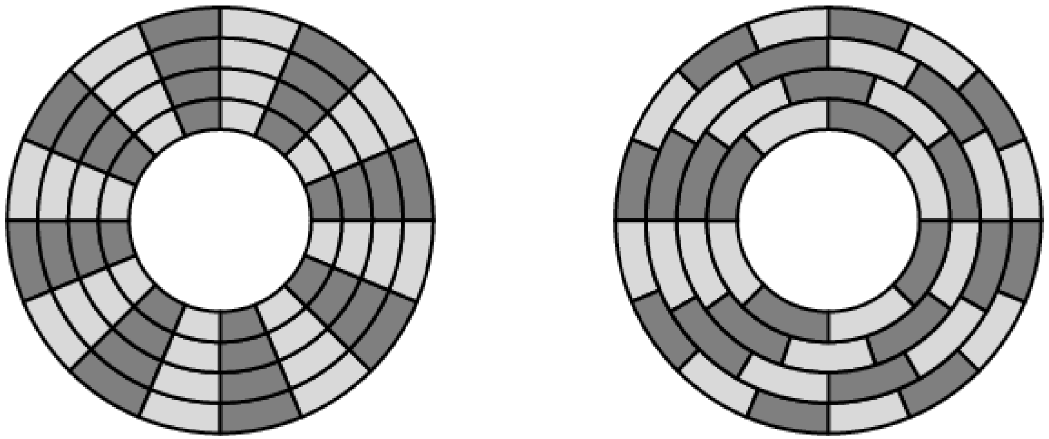




Data Organization and Formatting

- Concentric rings or tracks
  - Gaps between tracks
  - Reduce gap to increase capacity
  - Same number of bits per track (variable packing density)
  - Constant angular velocity
- Tracks divided into sectors
- Minimum block size is one sector
- May have more than one sector per block





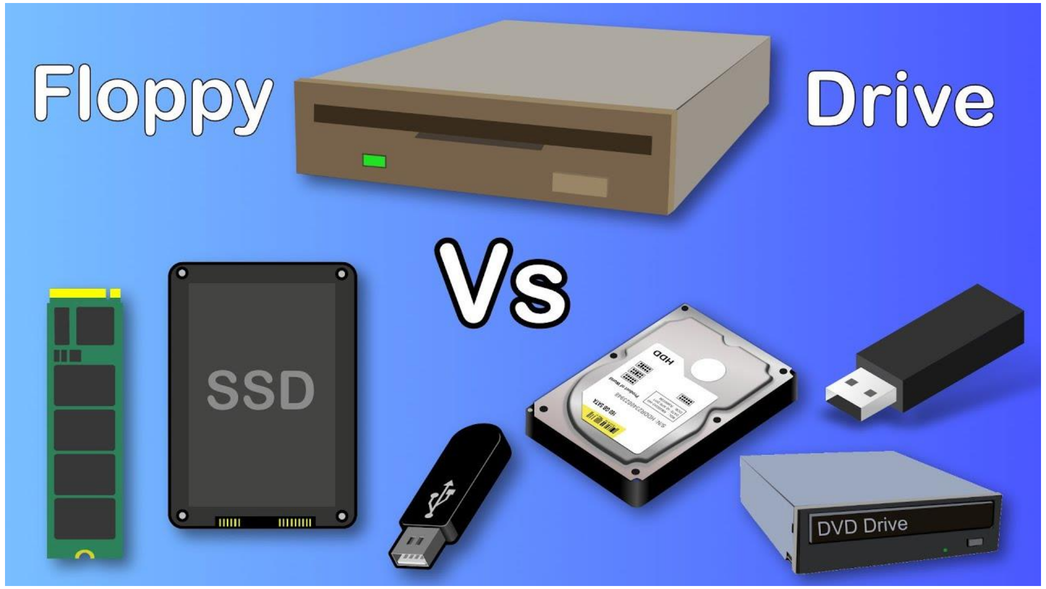
(a) Constant angular velocity

(b) Multiple zoned recording

Disk Types

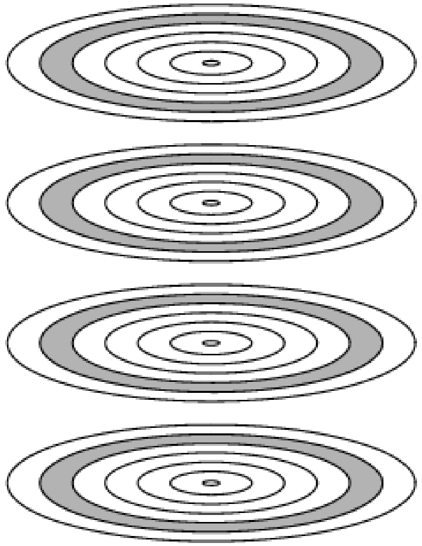
- Fixed head
  - One read write head per track
  - Heads mounted on fixed ridged arm
- Movable head
  - One read write head per side
  - Mounted on a movable arm
- Removable disk
  - Can be removed from drive and replaced with another disk
  - Provides unlimited storage capacity
  - Easy data transfer between systems
- Nonremovable disk
  - Permanently mounted in the drive

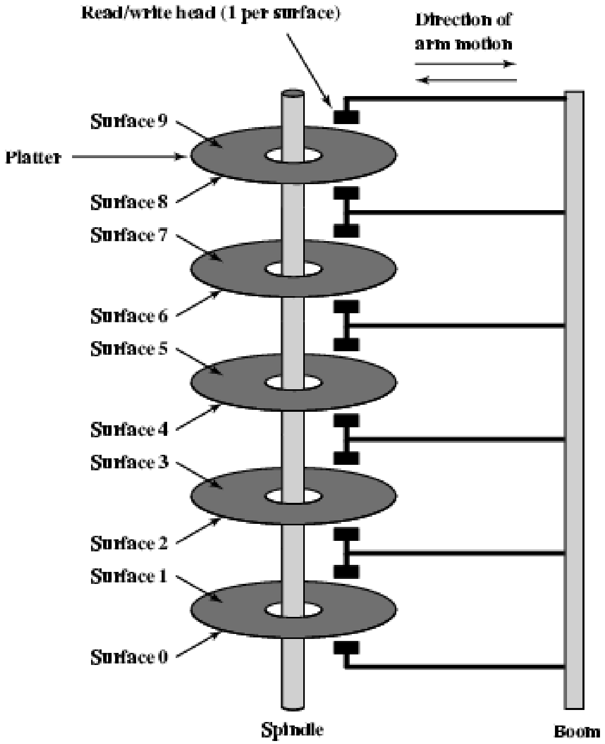
Floppy disks
• distribute software
• transfer files
• back-up data
8 inch
5¹/₄ inch
3¹/₂ inch



Multiple Platter

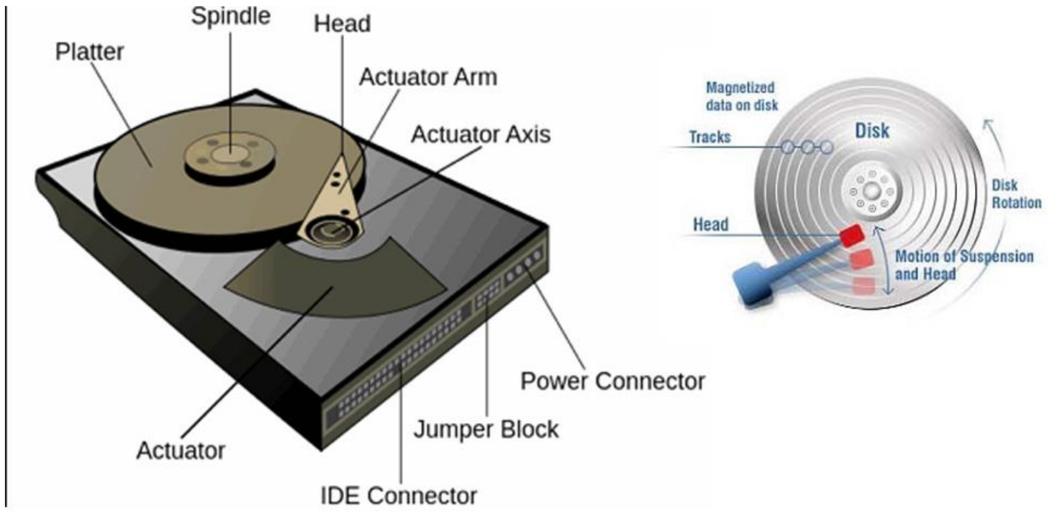
- One head per side
- Heads are joined and aligned
- Aligned tracks on each platter form cylinders
- Data is striped by cylinder
  - reduces head movement
  - Increases speed (transfer rate)

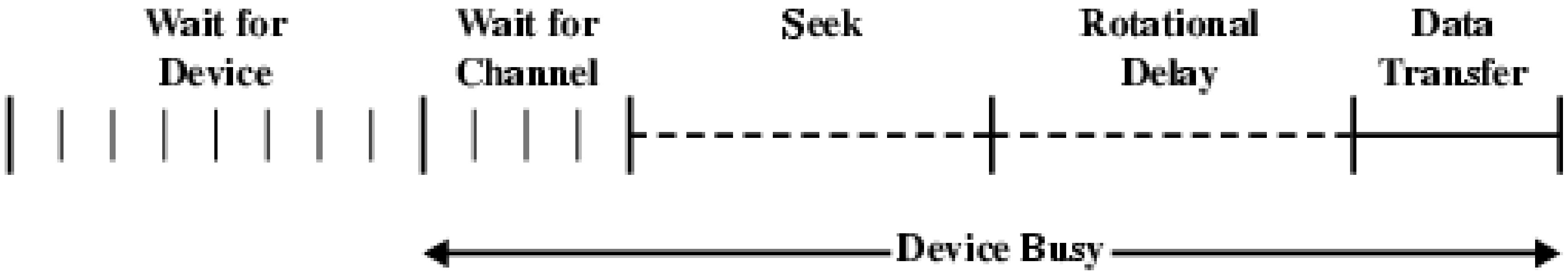




Disk Speed

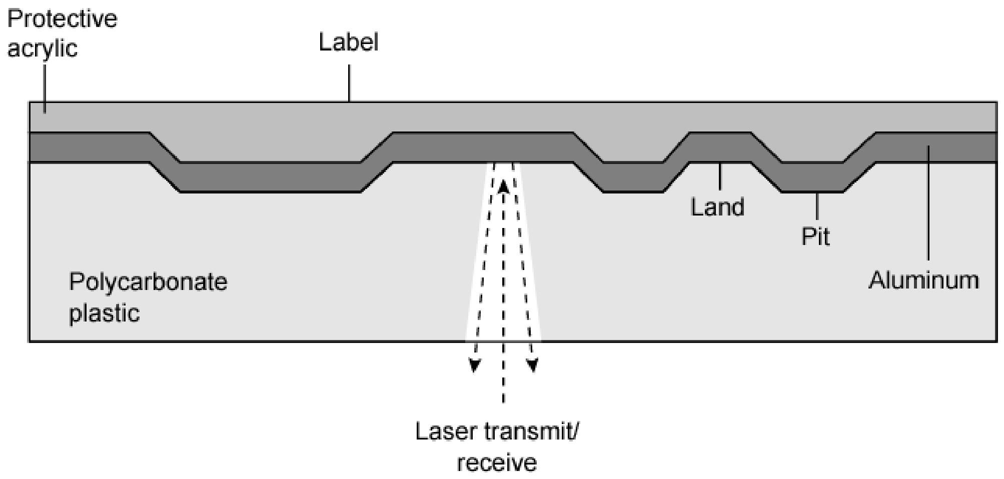
- Seek time
  - Moving head to correct track
- (Rotational) latency
  - Waiting for data to rotate under head
- Access time = Seek + Latency
- Transfer rate





Optical Storage CD-ROM

- Originally for audio
- 650Mbytes giving over 70 minutes audio
- Polycarbonate coated with highly reflective coat, usually aluminium
- Data stored as pits
- Read by reflecting laser
- Constant packing density
- Constant linear velocity

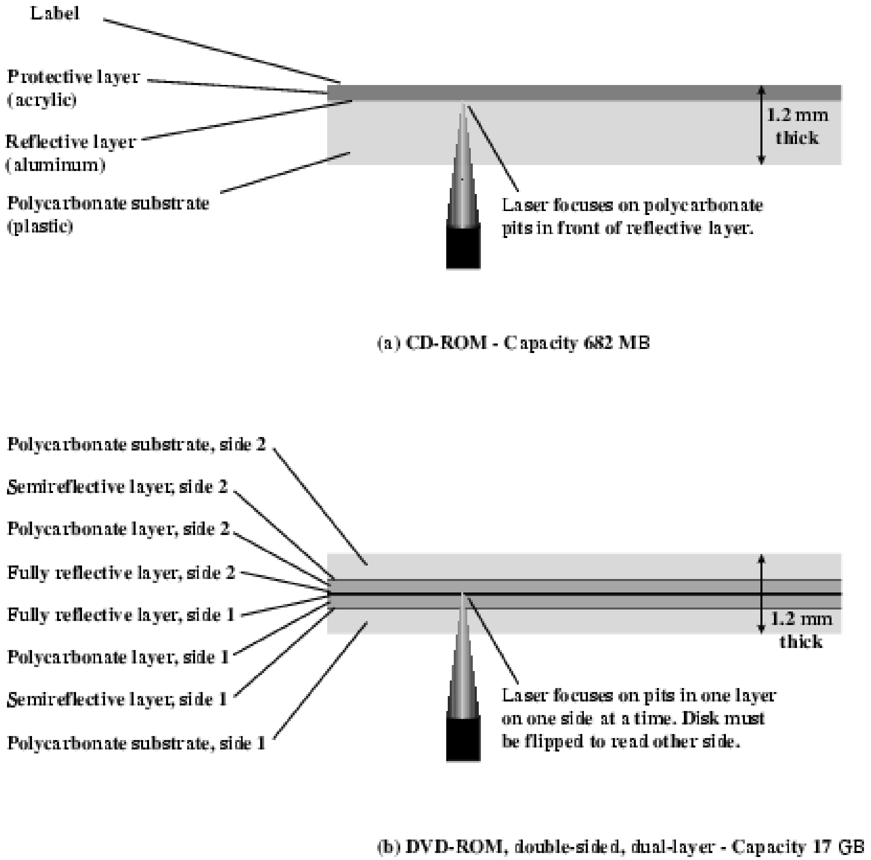


Other Optical Storage

- CD-Recordable (CD-R)
  - WORM
  - Now affordable
  - Compatible with CD-ROM drives
- CD-RW
  - Erasable
  - Getting cheaper
  - Mostly CD-ROM drive compatible
  - Phase change
    - Material has two different reflectivities in different phase states

DVD - technology

- Multi-layer
- Very high capacity (4.7G per layer)
- Full length movie on single disk
  - Using MPEG compression
- Finally standardized (honest!)
- Movies carry regional coding
- Players only play correct region films
- Can be "fixed"

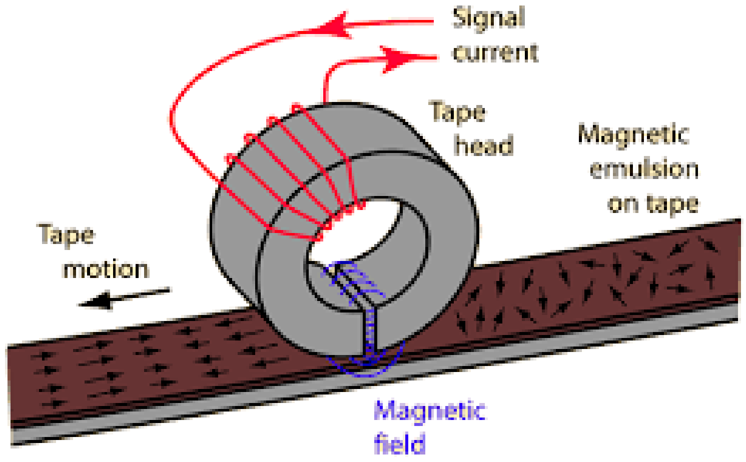

Label
Protective layer (acrylic)
Reflective layer (aluminum)
Polycarbonate substrate (plastic)
1.2 mm thick
Laser focuses on polycarbonate pits in front of reflective layer.
Polycarbonate substrate, side 2
Semireflective layer, side 2
Polycarbonate layer, side 2
Fully reflective layer, side 2
Fully reflective layer, side 1
Polycarbonate layer, side 1
Semireflective layer, side 1
Polycarbonate substrate, side 1
1.2 mm thick
Laser focuses on pits in one layer on one side at a time. Disk must be flipped to read other side.

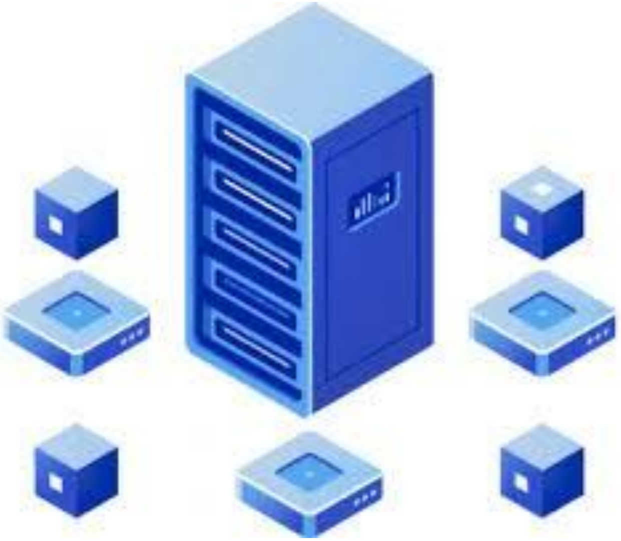
(a) CD-ROM - Capacity 682 MB

(b) DVD-ROM, double-sided, dual-layer - Capacity 17 GB

Magnetic Tape

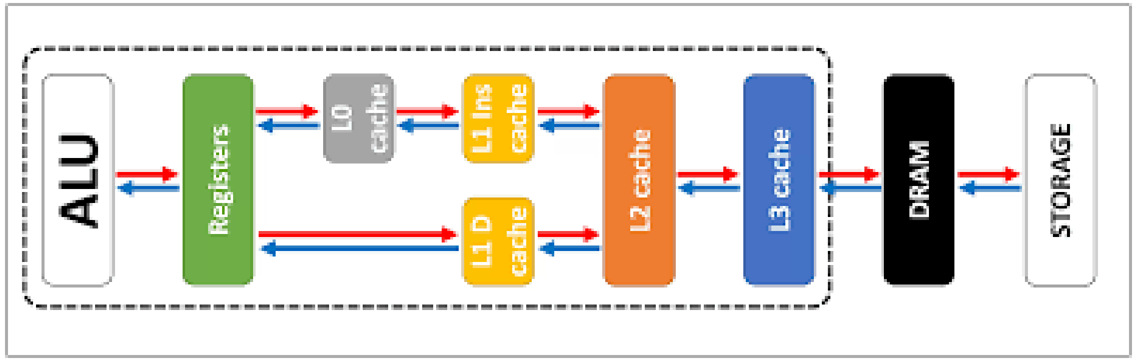
- Serial access
- Slow
- Very cheap
- Backup and archive

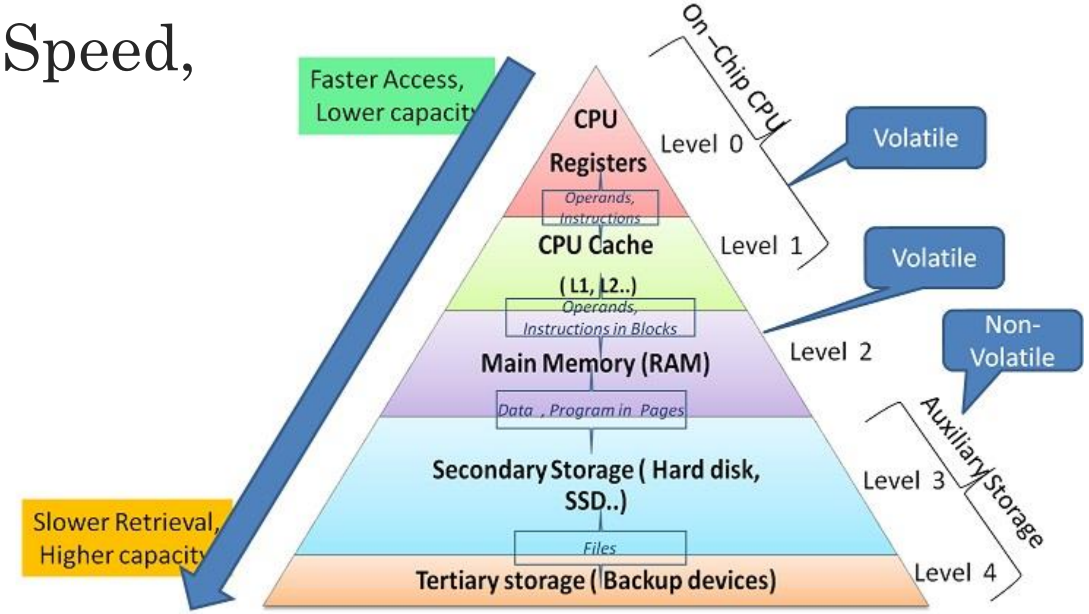




Conclusion

- Memory Functions
  - Store Programs, Data, States
- Stack Operations
  - Use Main Memory, LIFO
- Cache Memory
  - Improve Main Memory Access Speed, But Write Concerned
- Storage Types
  - ROM, SRAM, DRAM
  - Magnetic Storage Disk
  - Optical Disk
  - Hierarchical Storages
    - Speed vs Cost vs Capacity





END

Questions?