Computer arithmetic and memory system

Memory Hierarchy

•Capacity, cost and speed of different types of memory play a vital role while designing a memory system for computers.

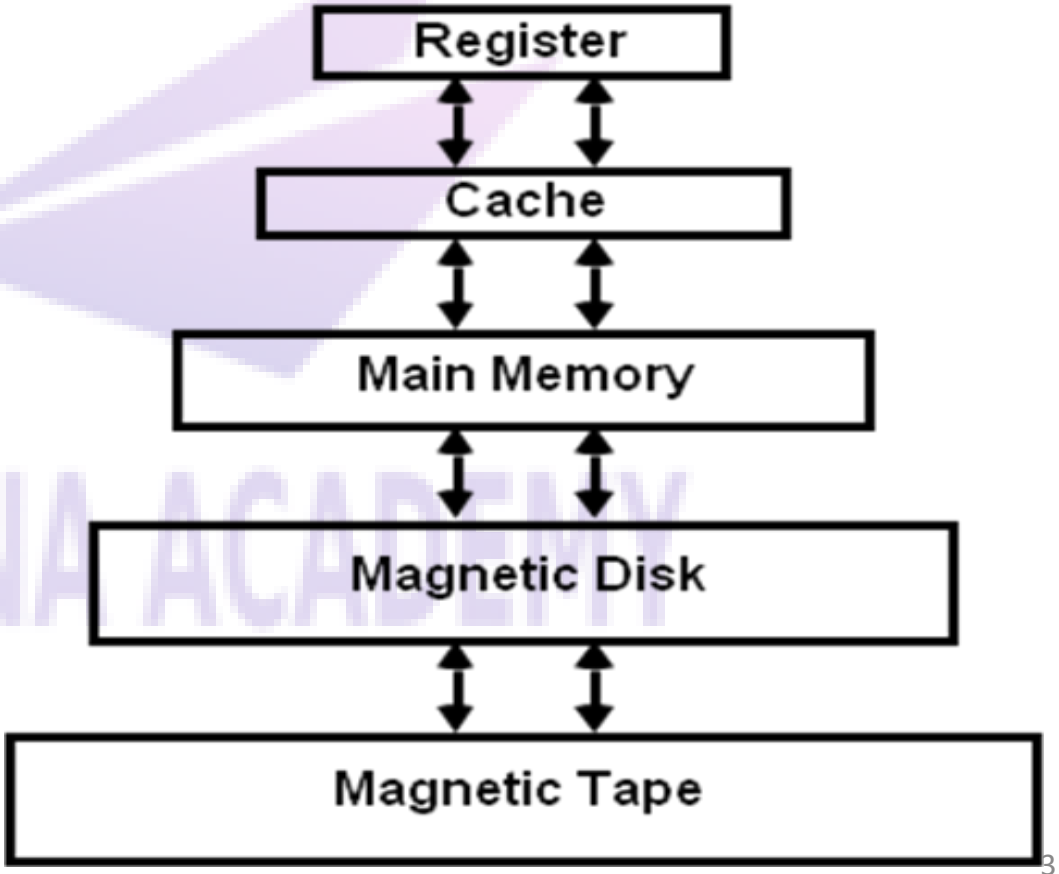
•There is a tradeoff between these three characteristics cost, capacity and access time.

•One cannot achieve all these quantities in same memory module because

- •If capacity increases, access time increases (slower) and due to which cost per bit decreases.
- •If access time decreases (faster), capacity decreases and due to which cost per bit increases.

•Memory Hierarchy is to obtain the highest possible access speed while minimizing the total cost of the memory system.

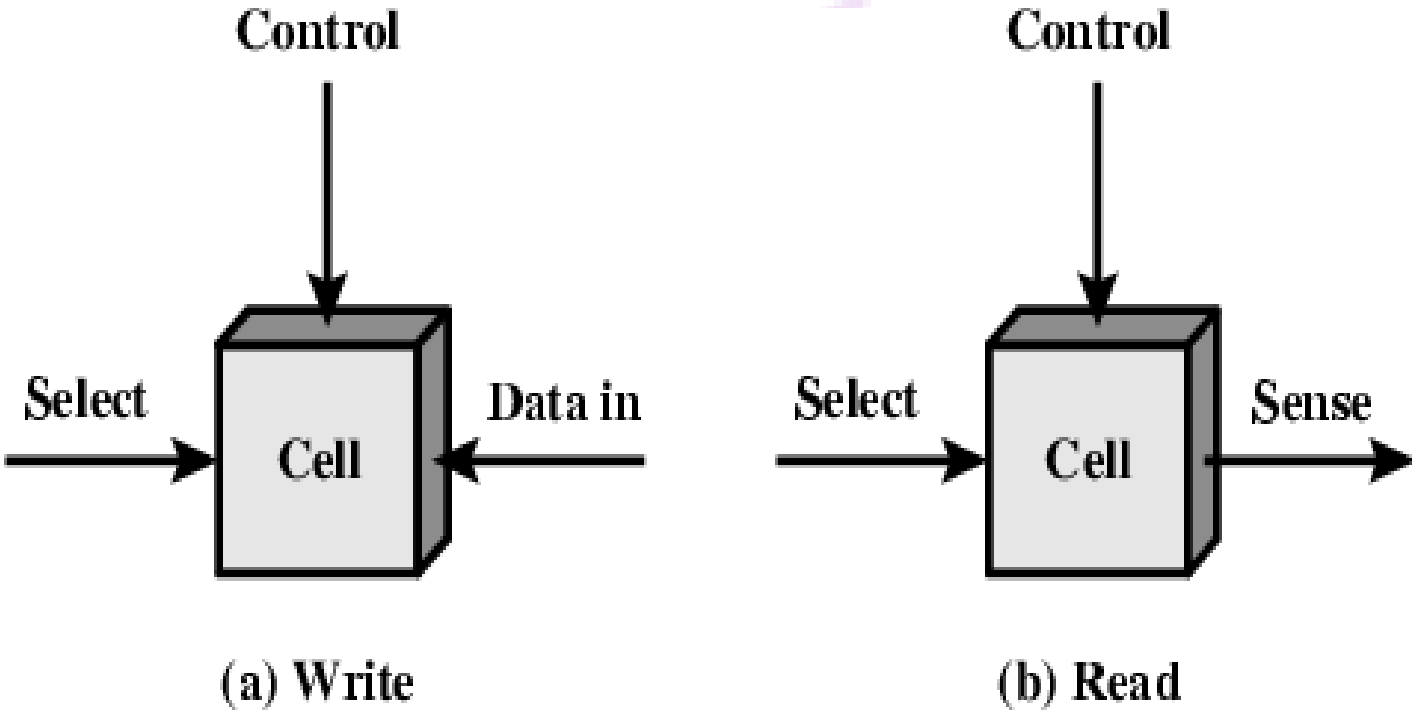
- As we go down in the hierarchy
  - Cost per bit decreases
  - Capacity of memory increases
  - Access time increases
  - Frequency of access of memory by processor also decreases
- **Hierarchy List**
  - Registers
  - L1 Cache
  - L2 Cache
  - Main memory
  - Disk cache
  - Disk
  - Optical
  - Tape



Internal or Main Memory

- The main memory is the central unit of the computer system.
- It is relatively large and fast memory to store programs and data during the computer operation.
- These memories employ semiconductor integrated circuits.
- The basic element of the semiconductor memory is the memory cell.
- The memory cell has three functional terminals which carries the electrical signal.
  - The select terminal: It selects the cell.
  - The data in terminal: It is used to input data as 0 or 1 and data out or sense terminal is used for the output of the cell's state.
  - The control terminal: It controls the function i.e. it indicates read and write.

- Most of the main memory in a general purpose computer is made up of RAM integrated circuits chips, but a portion of the memory may be constructed with ROM chips



RAM– Random Access memory

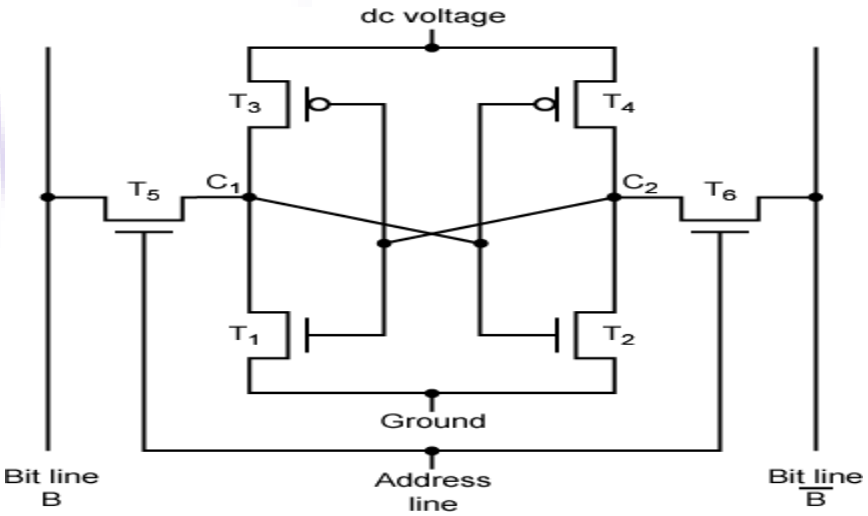
- Memory cells can be accessed for information transfer from any desired random location.
- The process of locating a word in memory is the same and requires of locating a word in memory is the same and requires an equal amount of time no matter where the cells are located physically in memory thus named 'Random access'.
- Integrated RAM are available in two possible operating modes, Static and Dynamic

Static RAM (SRAM)

- Consists of flip flop that stores binary information and this stored information remains valid as long as power is applied to the unit.

SRAM Structure

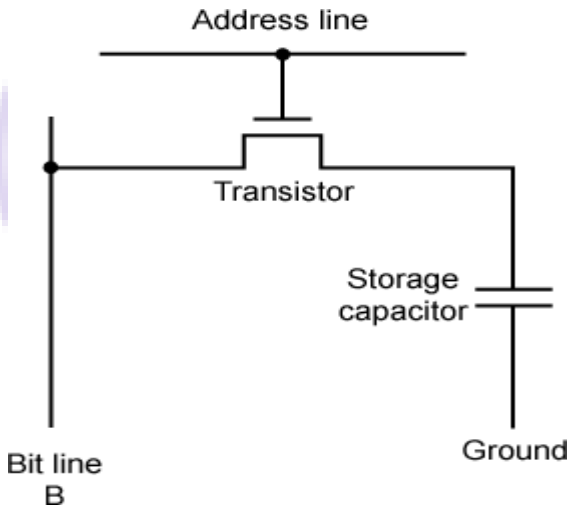
- Four transistors T1, T2, T3 and t4 are cross connected in an arrangement that produces a stable logical state.
- In logic state 1, point C1 is high and point C2 is low. In this state, T1 & T4 are off and T2 & T3 are on.
- In logic state 0, point C1 is low and C2 is high. In this state, T1 & T4 are on and T2 & T3 are off.
- The address line controls the two transistors T5 & T6. When a signal is applied to this line, the two transistors are switched on allowing for read and write operation.



- For a write operation, the desired bit value is applied to line B while it's complement is applied to line B complement. This forces the four transistors T1, T2, T3 & T4 into a proper state.
- For the read operation, the bit value is read from line B.

Dynamic RAM (DRAM)

- The dynamic RAM stores the binary information in the form of electrical charges and capacitor is used for this purpose.
- Since charge stored in capacitor discharges with time, capacitor must be periodically recharged and which is also called refreshing memory.



- The address line is activated when the bit value from this cell is to be read or written.
- The transistor acts as switch that is closed e. allowed current to flow, if voltage is applied to the address line; and opened i.e. no current to flow, if no voltage is present in the address line.

Dynamic RAM	Static RAM
Uses capacitor to store information	Uses flip flop to store information
More dense i.e. more cells can be accommodated per unit area	Needs more space
Slower, analog device	Faster, digital device
Less expensive, small in size	Expensive, big in size
Needs refreshing circuit	Don't require refreshing circuit
Used in main memory, larger memory units	Used in cache memory

ROM– Read Only memory

- Contains a permanent pattern of data that cannot be changed.
- Is non-volatile that is no power source is required to maintain the bit values in memory.
- While it is possible to read a ROM, it is not possible to write new data into it.
- The data or program is permanently presented in main memory and never be loaded from a secondary storage device with the advantage of ROM.
- A ROM is created like any other integrated circuit chip, with the data actually wired into the chip as part of the fabrication process.

- **Erasable Programmable ROM (EPROM)**
  - It is read and written electrically. However, before a write operation, all the storage cells must be erased to the same initial state by exposure of the packaged chip to ultraviolet radiation (UV ray).
  - Erasure is performed by shining an intense ultraviolet light through a window that is designed into the memory chip.
  - EPROM is optically managed and more expensive than PROM, but it has the advantage of the multiple update capability.

- **Electrically Erasable programmable ROM (EEPROM)**
  - This is a read mostly memory that can be written into at any time without erasing prior contents, only the byte or byte addresses are updated.
  - The write operation takes considerably longer than the read operation, on the order of several hundred microseconds per byte.
  - The EEPROM combines the advantage of non-volatility with the flexibility of being updatable in place, using ordinary bus control, addresses and data lines.
  - EEPROM is more expensive than EPROM and also is less dense, supporting fewer bits per chip.

- **Flash Memory**
  - Flash memory is also the semiconductor memory and because of the speed with which it can be reprogrammed, it is termed as flash.
  - It is interpreted between EPROM and EEPROM in both cost and functionality.
  - Like EEPROM, flash memory uses an electrical erasing technology.
  - An entire flash memory can be erased in one or a few seconds, which is much faster than EPROM.
  - In addition, it is possible to erase just blocks of memory rather than an entire chip.
  - However, flash memory doesn't provide byte level erasure, a section of memory cells are erased in an action or 'flash'.

External Memory

- The devices that provide backup storage are called external memory or auxiliary
- It includes serial access type such as magnetic tapes and random access type such as magnetic disks.

Magnetic Tape

- Is the strip of plastic coated with a magnetic recording Data can be recorded and read as a sequence of character through read / write head.
- It can be stopped, started to move forward or in reverse or can be rewound.
- Data on tapes are structured as number of parallel tracks running length wise.
- Earlier tape system typically used nine tracks.
- This made it possible to store data one byte at a time with additional parity bit as 9th track.
- The recording of data in this form is referred to as parallel recording.

Magnetic Disk

- Is a circular plate constructed with metal or plastic coated with magnetic material often both side of disk are used and several disk stacked on one spindle.
- Bits are stored in magnetize surface in spots along concentric circles called tracks.
- The tracks are commonly divided into sections called sectors.
- After the read/write head are positioned in specified track the system has to wait until the rotating disk reaches the specified sector under read/write head.
- Disk that are permanently attached to the unit assembly and cannot be used by occasional user are called hard disk drive with removal disk is called floppy disk.

Optical Disk

- The disk is form from resin such as Digitally recorded information is imprinted as series of microscopic pits on the surface of poly carbonate.
- This is done with the finely focused high intensity leaser.
- The pitted surface is then coated with reflecting surface usually aluminum or gold.
- The shiny surface is protected against dust and scratches by the top coat of acrylic.
- Information is retrieved from CD by low power laser.
- The intensity of reflected light of laser changes as it encounters a pit.
- The areas between pits are called lands.
- A land is a smooth surface which reflects back at higher intensity.
- The change between pits and land is detected by photo sensor and converted into digital signal.

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

DVD-Writable

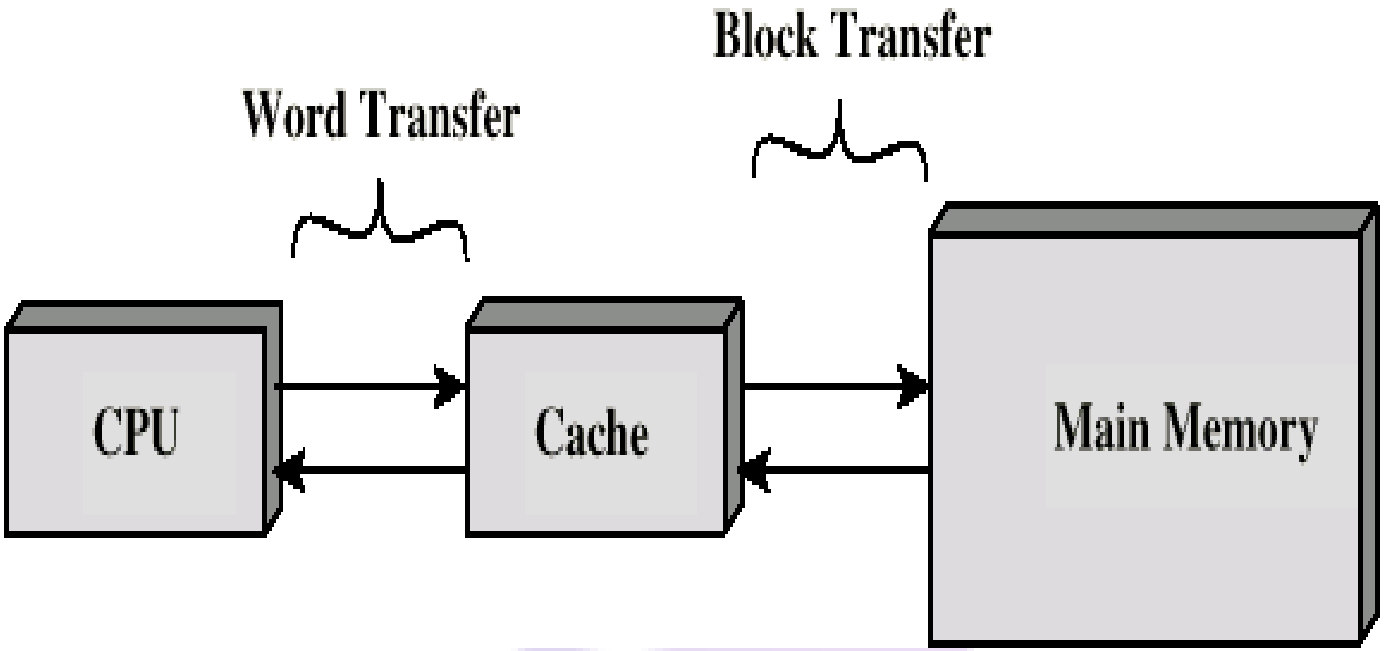
- Loads of trouble with standards
- First generation DVD drives may not read first generation DVD-W disks
- First generation DVD drives may not read CD-RW disks

Cache Memory Principles

Principles are

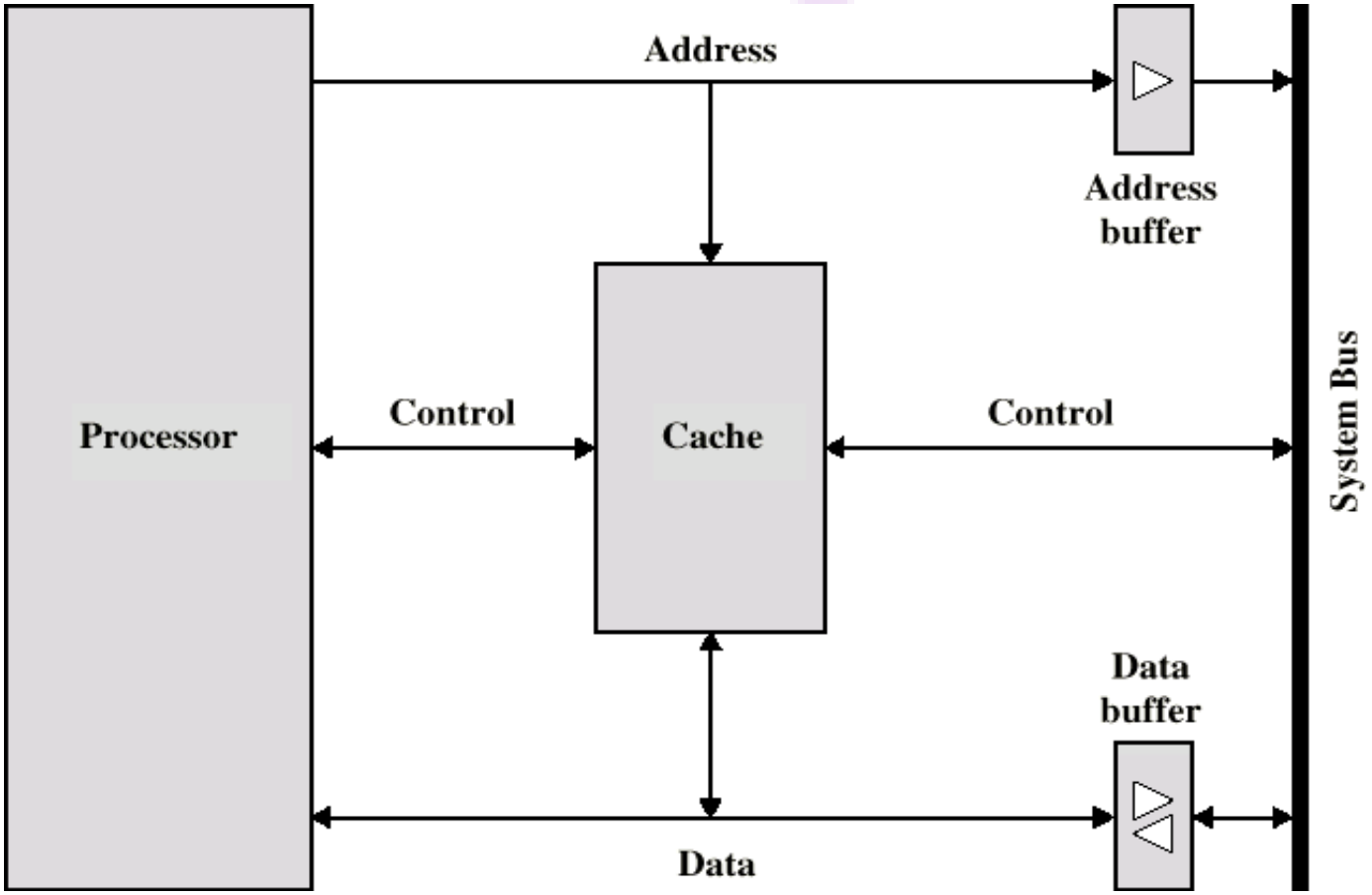
- Intended to give memory speed approaching that of fastest memories available but with large size, at close to price of slower memories
- Cache is checked first for all memory references.
- If not found, the entire block in which that reference resides in main memory is stored in a cache slot, called a line
- Each line includes a tag (usually a portion of the main memory address) which identifies which particular block is being stored
- Locality of reference implies that future references will likely come from this block of memory, so that cache line will probably be utilized repeatedly.

- **The proportion of memory references, which are found already stored in cache, is called the hit ratio**.
- When the processor attempts to read a word of memory, a check is made to determine if the word is in the cache.
- If so, the word is delivered to the processor.
- If not, a block of main memory, consisting of fixed number of words is read into the cache and then the word is delivered to the processor
- The locality of reference property states that over a short interval of time, address generated by a typical program refers to a few localized area of memory repeatedly.
- So if programs and data which are accessed frequently are placed in a fast memory, the average access time can be reduced.
- This type of small, fast memory is called **cache memory which is placed in between the CPU and the main memory**.



- When the CPU needs to access memory, cache is examined.

- If the word is found in cache, it is read from the cache and if the word is not found in cache, main memory is accessed to read A block of word containing the one just accessed is then transferred from main memory to cache memory.
- Typical cache organization

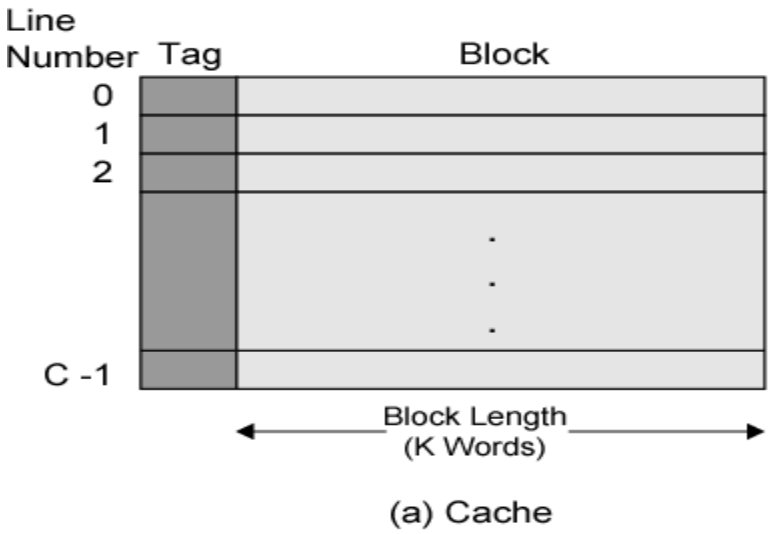


- Cache connects to the processor via data control and address line.
- The data and address lines also attached to data and address buffer which attached to a system bus from which main memory is reached.
- When a cache hit occurs, the data and address buffers are disabled and the communication is only between processor and cache with no system bus traffic.
- When a cache miss occurs, the desired word is first read into the cache and then transferred from cache to For later case, the cache is physically interposed between the processor and main memory for all data, address and control lines.

Cache Operation

- CPU generates the receive address (RA) of a word to be moved (read).
- Check a block containing RA is in cache
- If present, get from cache (fast) and return.
- If not present, access and read required block from main memory to cache.
- Allocate cache line for this new found block.
- Load block for cache and deliver word to CPU
- Cache includes tags to identify which block of main memory is in each cache slot

Cache operation



Elements of cache design

Cache size

- Size of cache to be small enough so that the overall average cost per bit is close to that of main memory alone
- And large enough so that the overall average access time is close to that of cache alone
- **The larger the cache, the larger the number of gates involved in addressing the cache**
- Large caches tend to be slightly slower than small ones – even when built with the small integrated circuit technology and put in the same place on chip and circuit board
- The available chip and board also limits cache size

Mapping Function

- **The transformation of data from main memory to cache memory is referred to as memory mapping processes**.
- Because there are fewer cache lines than main memory blocks, an algorithm is needed for mapping main memory blocks into cache lines.
- There are three different types of **mapping functions in common use and are direct, associative and set associative**.
- All the three include following elements in each example.
  - The cache can hold 64 Kbytes
  - Data is transferred between main memory and the cache in blocks of 4 bytes each. This means that the cache is organized as 16Kbytes = 214 lines of 4 bytes each.

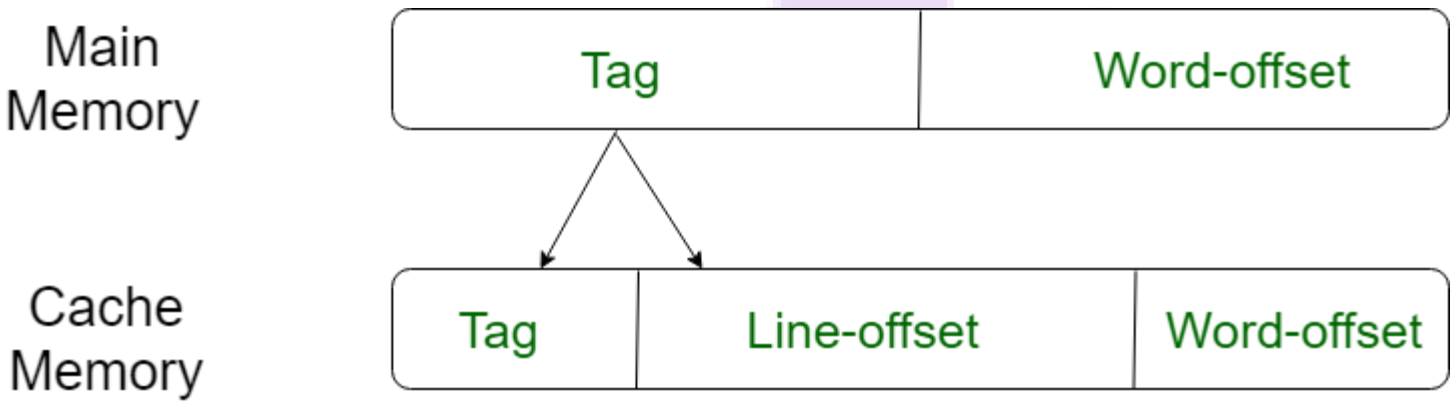
- The main memory consists of 16 Mbytes with each byte directly addressable by a 24 bit address (2^{24} = 16Mbytes). Thus, for mapping purposes, we can consider main memory to consist of 4Mbytes blocks of 4 bytes each.

Direct Mapping

- The simplest technique, known as direct mapping, maps each block of main memory into only one possible cache line.
- or **In Direct mapping, assign each memory block to a specific line in the cache**.
- **If a line is previously taken up by a memory block when a new block needs to be loaded, the old block is trashed.**
- An address space is split into two parts index field and a tag field.
- The cache is used to store the tag field whereas the rest is stored in the main memory.

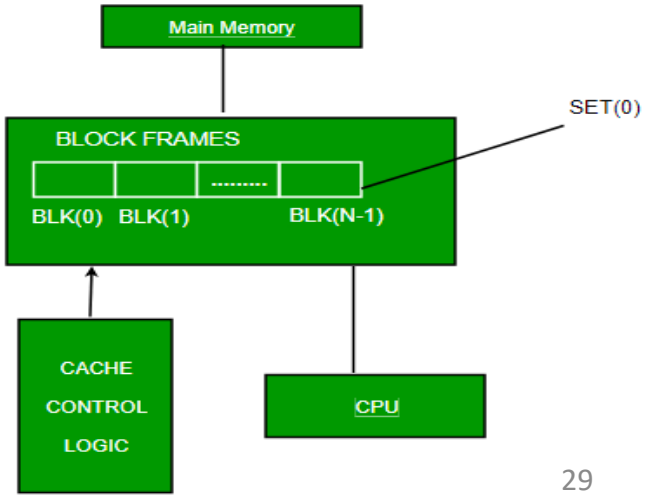
- **Direct mapping`s performance is directly proportional to the Hit ratio.**
  - **i = j modulo m**

    where i = cache line number, j = main memory block number, m = number of lines in the cache



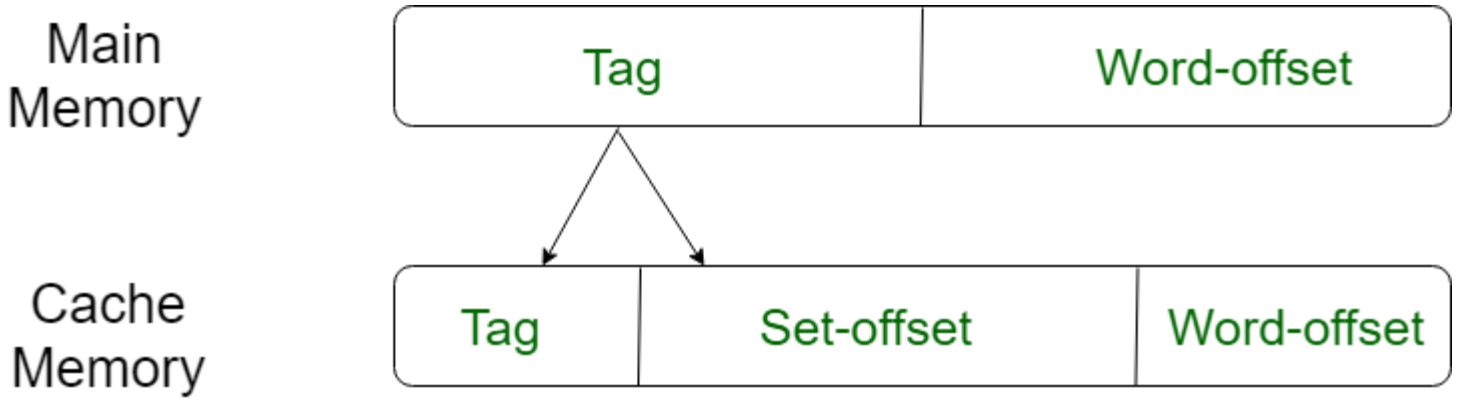
Associative Mapping

- **In this case, associative memory is used to store the content and addresses of the memory word**.
- Any block can go into any line of the cache.
- This means that the **word id bits are used to identify which word in the block is needed,** but the tag becomes all of the remaining bits.
- This enables the placement of any word at any place in the cache memory.
- **It is considered to be the fastest and most flexible mapping form**. In associative mapping, the index



Set-Associative Mapping

- This form of mapping is an enhanced form of direct mapping where the drawbacks of direct mapping are removed.
- Set associative addresses the problem of possible thrashing in the direct mapping method.
- It does this by saying that instead of having exactly one line that a block can map to in the cache, we will group a few lines together creating a ***set***.
- Then a block in memory can map to any one of the lines of a specific set.



- **Set-associative mapping allows each word that is present in the cache can have two or more words in the main memory for the same index address.**
- Set associative cache mapping combines the best of direct and associative cache mapping techniques.
- **In set associative mapping the index bits are given by the set offset bits.**
- In this case, the cache consists of a number of sets, each of which consists of a number of lines.

Relationships in the Set-Associative Mapping can be defined as:

- **m = v * k**
- **i= j mod v**

  where i = cache set number, j = main memory block number, v = number of sets, m = number of lines in the cache number of sets, k = number of lines in each set

Cache replacement policies

- There are multiple ways to replace data placed in cache
- As there is different mapping technique such as direct mapping, associative mapping and set-associative mapping
- Cache replacement algorithms are used to optimize the time taken by processor to process the information by storing the information needed by processor at that time
- And possibly in future so that if processor needs that information, it can be provided immediately.

List of algorithms are:

- Least Recently used (LRU)
  - **replace that block in the set which has been in cache longest with no reference to it**
  - Implementation: with 2-way set associative, have a USE bit for each line in a set. When a block is read into cache, use the line whose USE bit is set to 0, then set its USE bit to one and the other line's USE bit to 0.
  - Probably the most effective method
- First in first out (FIFO)
  - replace that block in the set which has been in the cache longest
  - Implementation: use a round-robin or circular buffer technique (keep up with which slot's "turn" is next

- Least-frequently-used (LFU)
  - replace that block in the set which has experienced the fewest references or hits
  - Implementation: associate a counter with each slot and increment when used
- Random
  - replace a random block in the set
  - Interesting because it is only slightly inferior to algorithms based on usage

Write Policy in Computer Organizations

- When a line is to be replaced, must update the original copy of the line in main memory if any addressable unit in the line has been changed
- If a block has been altered in cache, it is necessary to write it back out to main memory before replacing it with another block (writes are about 15% of memory references)
- Must not overwrite a cache block unless main memory is up to date
- I/O modules may be able to read/write directly to memory
- Multiple CPU's may be attached to the same bus, each with their own cache

Write Through

- All write operations are made to main memory as well as to cache, so main memory is always valid
- Other CPU's monitor traffic to main memory to update their caches when needed
- This generates substantial memory traffic and may create a bottleneck
- Anytime a word in cache is changed, it is also changed in main memory
- Both copies always agree
- Generates lots of memory writes to main memory
- **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

- **When an update occurs, an UPDATE bit associated with that slot is set, so when the block is replaced it is written back first**
- **During a write, only change the contents of the cache**
- **Update main memory only when the cache line is to be replaced**
- Causes “cache coherency” problems -- different values for the contents of an address are in the cache and the main memory
- Complex circuitry to avoid this problem
- Accesses by I/O modules must occur through the cache

- Multiple caches still can become invalidated, unless some cache coherency system is used. Such systems include:
  - Bus Watching with Write Through - other caches monitor memory writes by other caches (using write through) and invalidates their own cache line if a match
  - Hardware Transparency - additional hardware links multiple caches so that writes to one cache are made to the others
  - Non-cacheable Memory - only a portion of main memory is shared by more than one processor, and it is non-cacheable

Number of Caches

L1 and L2 Cache

On-chip cache (L1 Cache)

- It is the cache memory on the same chip as the processor, the on-chip cache.
- It reduces the processor's external bus activity and therefore speeds up execution times and increases overall system performance.
- Requires no bus operation for cache hits
- Short data paths and same speed as other CPU transactions

Off-chip cache (L2 Cache)

- It is the external cache which is beyond the processor.
- **If there is no L2 cache and processor makes an access request for memory location not in the L1 cache, then processor must access DRAM or ROM memory across the bus.**
- Due to this typically slow bus speed and slow memory access time, this results in poor performance.

- On the other hand, if an L2 SRAM cache is used, then frequently the missing information can be quickly retrieved.
- It can be much larger
- It can be used with a local bus to buffer the CPU cache-misses from the system bus

Unified and Split Cache

- **Unified Cache**
  - Single cache contains both instructions and data. Cache is flexible and can balance "allocation" of space to instructions or data to best fit the execution of the program.
  - Has a higher hit rate than split cache, because it automatically balances load between data and instructions (if an execution pattern involves more instruction fetches than data fetches, the cache will fill up with more instructions than data)
  - Only one cache need be designed and implemented

Split Cache

- Cache splits into two parts first for instruction and second for data. Can outperform unified cache in systems that support parallel execution and pipelining (reduces cache contention)
- Trend is toward split cache because of superscalar CPU's
- Better for pipelining, pre-fetching, and other parallel instruction execution designs
- Eliminates cache contention between instruction processor and the execution unit (which uses data)

Memory Write Ability and Storage Permanence

- ***Memory Basics***
  - mxn memory stores m words of n bits each.
  - m=2k where k = no of address input signals
  - n indicates no of data signals
  - r/w selects read or write
  - enable : read or write only when asserted
- ***Write Ability***
  - Write ability is the manner and speed that a particular memory can be written . The ranges of write ability are:
    - High End (Processor writes to memory quickly – RAM)
    - Middle Range (Processor writes to memory slowly – FLASH , EEPROM)
    - Lower Range (Special equipment used to write to memory – EEPROM , OTP Rom)

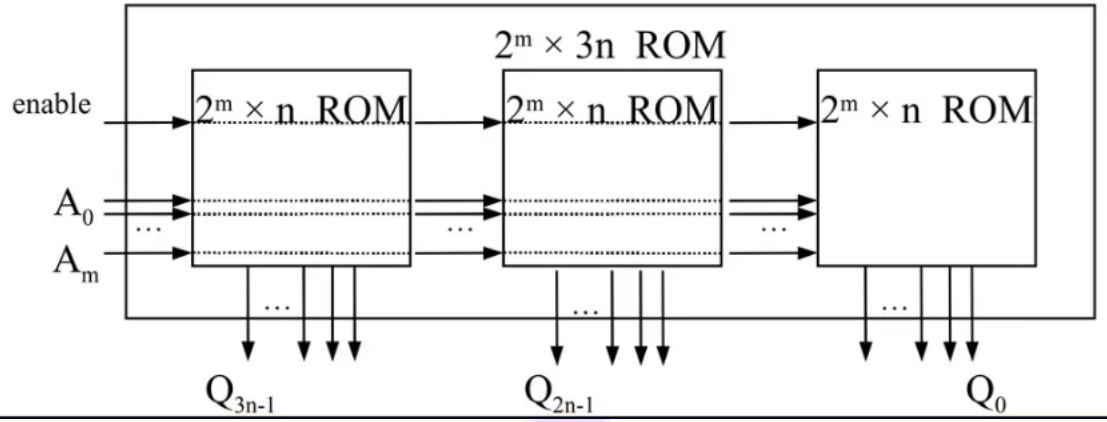
- Low End (bits stored only during fabrication – Masked ROM)

- ***Storage Permanence***
  - Storage permanence is the ability of memory to hold its stored bits after they have been written. The ranges are :-
    - High end (Never loss bits – Masked ROM)
    - Middle Range (holds bits for days or years after power cutoff – NVRAM)
    - Lower Range (holds bits as long as power is supplied – SRAM)
    - Lower End (lose bits immediately after written – DRAM)

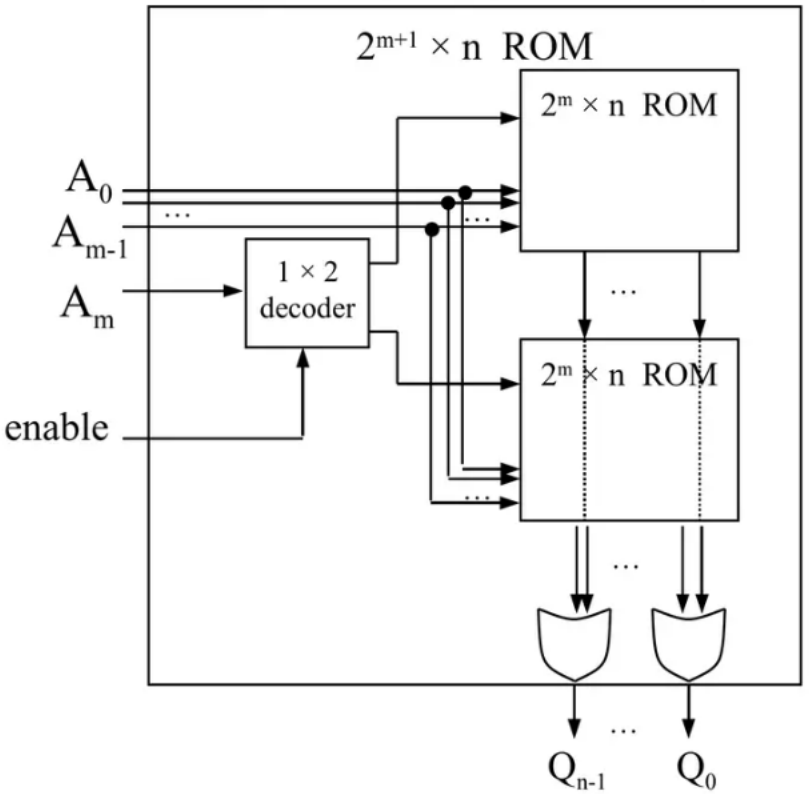
Composing Memory

- Memory size needed often differs from size of readily available memories
- **When available memory is larger, simply ignore unneeded high order address bits and high data lines**
- **When available memory is smaller, compose several smaller memories into one larger memory**
  - Connect side by side to increase width of words
  - Connect top to bottom to increase number of words
    - Added high order address line selects smaller memory containing desired word using a decoder
  - Combine technique to increase number and width of words

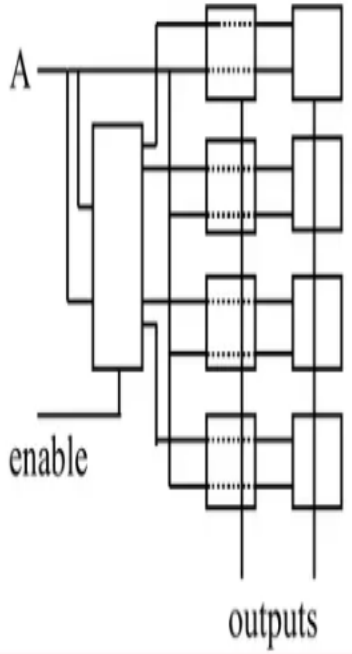
Increase width of words



Increase number of words



Increase number and width of words



- Cache memory is the onboard storage.
  a) True
  b) False
- Which of the following is the fastest means of memory access for CPU?
  a) Registers
  b) Cache
  c) Main memory
  d) Virtual Memory
- The memory implemented using the semiconductor chips is _____
  a) Cache
  b) Main
  c) Secondary
  d) Registers
- Size of the _____ memory mainly depends on the size of the address bus.
  a) Main
  b) Virtual
  c) Secondary
  d) Cache

- Which of the following is independent of the address bus?
  a) Secondary memory
  b) Main memory
  c) Onboard memory
  d) Cache memory
- MAR stands for _____
  a) Memory address register
  b) Main address register
  c) Main accessible register
  d) Memory accessible register
- The standard SRAM chips are costly as _____
  a) They use highly advanced micro-electronic devices
  b) They house 6 transistor per chip
  c) They require specially designed PCB’s
  d) None of the mentioned

- The drawback of building a large memory with DRAM is _____
  a) The large cost factor
  b) The inefficient memory organization
  c) The Slow speed of operation
  d) All of the mentioned
- The memory which is used to store the copy of data or instructions stored in larger memories, inside the CPU is called _____
  a) Level 1 cache
  b) Level 2 cache
  c) Registers
  d) TLB
- The larger memory placed between the primary cache and the memory is called _____
  a) Level 1 cache
  b) Level 2 cache
  c) EEPROM
  d) TLB
- The last on the hierarchy scale of memory devices is _____
  a) Main memory
  b) Secondary memory
  c) TLB
  d) Flash drives

- The memory blocks are mapped on to the cache with the help of _____
  a) Hash functions
  b) Vectors
  c) Mapping functions
  d) None of the mentioned
- During a write operation if the required block is not present in the cache then _____ occurs.
  a) Write latency
  b) Write hit
  c) Write delay
  d) Write miss
- In _____ protocol the information is directly written into the main memory.
  a) Write through
  b) Write back
  c) Write first
  d) None of the mentioned

- The method of mapping the consecutive memory blocks to consecutive cache blocks is called _____
  a) Set associative
  b) Associative
  c) Direct
  d) Indirect
- While using the direct mapping technique, in a 16 bit system the higher order 5 bits are used for _____
  a) Tag
  b) Block
  c) Word
  d) Id
- In associative mapping, in a 16 bit system the tag field has _____ bits.
  a) 12
  b) 8
  c) 9
  d) 10
- The technique of searching for a block by going through all the tags is _____
  a) Linear search
  b) Binary search
  c) Associative search
  d) None of the mentioned

- In set-associative technique, the blocks are grouped into _____ sets.
  a) 4
  b) 8
  c) 12
  d) 6
- A control bit called _____ has to be provided to each block in set-associative.
  a) Idol bit
  b) Valid bit
  c) Reference bit
  d) All of the mentioned
- Data which is not up-to date is called as _____
  a) Spoilt data
  b) Stale data
  c) Dirty data
  d) None of the mentioned