Chanpter 4

The Memory System

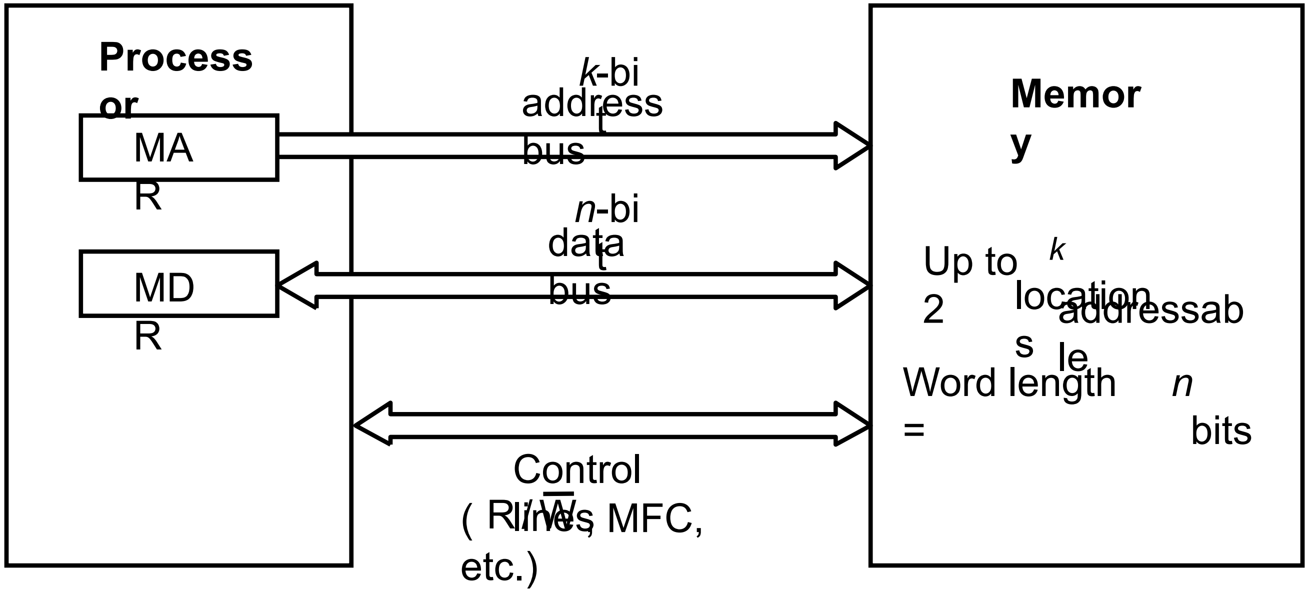
A. M. CHANDRASHEKHAR
Asst. Professor
Computer Science Dept.
S.J. College of Engg. Mysore

Topics to be covered

- Some Basic Concepts:
- Semiconductor RAM Memories,
- Read-only Memories,
- Speed, Size & Cost :
- Cache Memories,
- Performance considerations,
- Virtual memories,
- Secondary Storage,

Some basic concepts

- Maximum size of the Main Memory :address bus length
- byte-addressable (basic unit byte), word-addressable
- CPU-Main Memory Connection


Processor
MAR
MDR
k-bit address bus
n-bit data bus
Control lines ($R/\overline{W}$, MFC, etc.)
Memory
Up to 2^k addressable locations
Word length = n bits

Some basic concepts(Contd.,)

- Measures for the speed of a memory:
  - memory access time + memory cycle time (usually longer than MAT).
  - Min time delay between initiation of two successive memory operation

- An important design issue is to provide a computer system with as large and fast a memory as possible, within a given cost target.

- Several techniques to increase the effective size and speed of the memory:
  - Cache memory (to increase the effective speed).
  - Virtual memory (to increase the effective size).

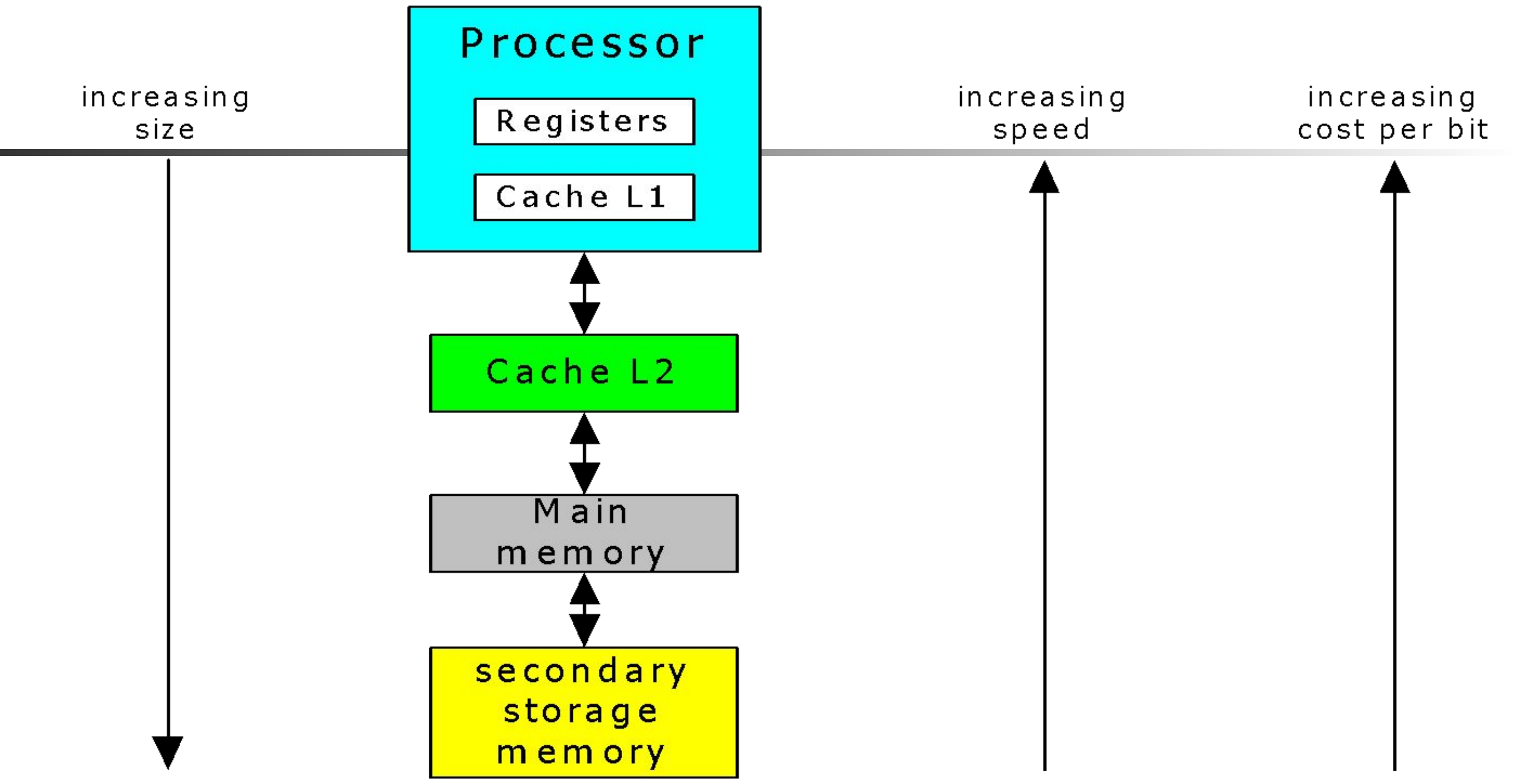
Techniques to improve speed & increase the size of memory

- Cache memory:
  - Faster, smaller, holds currently active segment of program and its data
- Memory interleaving techniques
  - Divide MM into number of memory modules
  - Successive words in memory is placed in different modules
  - Parallel access to different modules is possible
- Virtual memory:
  - Address generated by CPU□ virtual /logical
  - It is different from real physical address
  - Transformation process (mapping): MMU

Speed, Size, and Cost

- A big challenge in the design of a computer system is to provide a sufficiently large memory, with a reasonable speed at an affordable cost.
- Static RAM:
  - Very fast, but expensive, because a basic SRAM cell has a complex circuit making it impossible to pack a large number of cells onto a single chip.
- Dynamic RAM:
  - Simpler basic cell circuit, hence are much less expensive, but significantly slower than SRAMs.
- Magnetic disks:
  - Storage provided by DRAMs is higher than SRAMs, but is still less than what is necessary.
  - Secondary storage such as magnetic disks provide a large amount of storage, but is much slower than DRAMs.

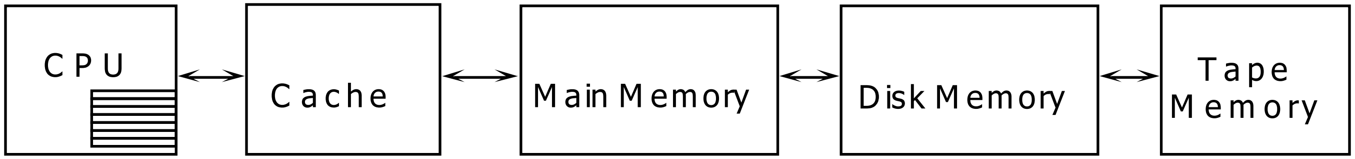
Memory hierarchy



- *Fastest access is to the data held in processor registers.*
- *Relatively small amount of memory that can be implemented on the processor chip. This is processor cache or Level 1 (L1) cache .*
- *Level 2 (L2) cache is in between main memory and processor.*
- *Next level is main memory, Much larger, but much slowerthan cache memory.*
- *Next level is magnetic disks. Huge amount of inexepensive storage.*

The Memory Hierarchy :Cost, Performance

Some Typical Values:



Component	CPU	Cache	Main Memory	Disk Memory	Tape Memory
Access	Random	Random	Random	Direct	Sequential
Capacity	64-1024B	32KB-8MB	< 512GB	> 1TB	many TB
Latency	0.15-0.3ns	0.5-15ns (5 ms)	30-200ns	5,000,000ns	<10s
Block size	1 word	16 words	16 words	4KB	4KB
Bandwidth	100-1000GB/s	10-40GB/s	5-20GB/s	0.05-0.5 GB/s	0.001GB/s
Cost	***** (high)	****	***	**	* (low)

The Memory System

Semiconductor memories

Semiconductor Memories

- Nonvolatile memory
  - Masked ROM
  - ROM
  - PROM
  - EPROM
  - EEPROM
  - Flash memory

- Volatile memory
  - SRAM
  - DRAM
    - Asynchronous
      - DRAM
      - FPM DRAM
    - Synchronous
      - SDRAM
      - DDR SDRAM
      - RDRAM

Introduction to Semiconductor memories …

- Semiconductor integrated circuits are used to implement MM
- Costlier (earlier), available in wide range of speed
- 2 types of semiconductor memories
  - Bipolar memories
  - Metal oxide semiconductor(MOS) memories
    - MOS memories are widely used in MM
    - Bit density is high , Manufacturing process is simpler,
    - low power dissipation
    - Slower operating speed is the disadvantage of MOS memories

Read-Only Memories (ROMs)

- SRAM and SDRAM chips are volatile:

- Many applications need memory devices to retain contents after the power is turned off.
  - For example, computer is turned on, the operating system must be loaded from the disk into the memory.
  - Store instructions which would load the OS from the disk.

- Non-volatile memory (ROM) is read in the same manner as volatile memory.
  - Separate writing process is needed to place information in this memory.

ROM

- ROM : Read Only Memory
  - Data are written / programmed into a ROM when it is manufactured (MROM□Masked ROM)
- PROM : Programmable ROM
  - Allow the data to be loaded / programmed by user only once.
  - Process of inserting the data is irreversible.
  - Programmed by burning the fuse using high current pulse
  - Flexible and convenient compared to ROM

Erasable PROM (EPROM)

- Erasable & Reprogrammable ROM.
- Erasure requires exposing the ROM to UV light.
- Need to remove from circuit for Erasing.
- The whole chip is erased and selective / block erase is not allowed.
- Flexibility, useful during the development phase of digital systems.

Electrically Erasable PROM (EEPROM)

- EEPROMs the contents can be stored and erased electrically.
- No requirement of physically removed from the circuit for reprogramming
- Use special voltage level to erase data(21V)
- Any cell contents can be delete selectively
- different voltage levels required for read, write and erase □ Disadvantage.

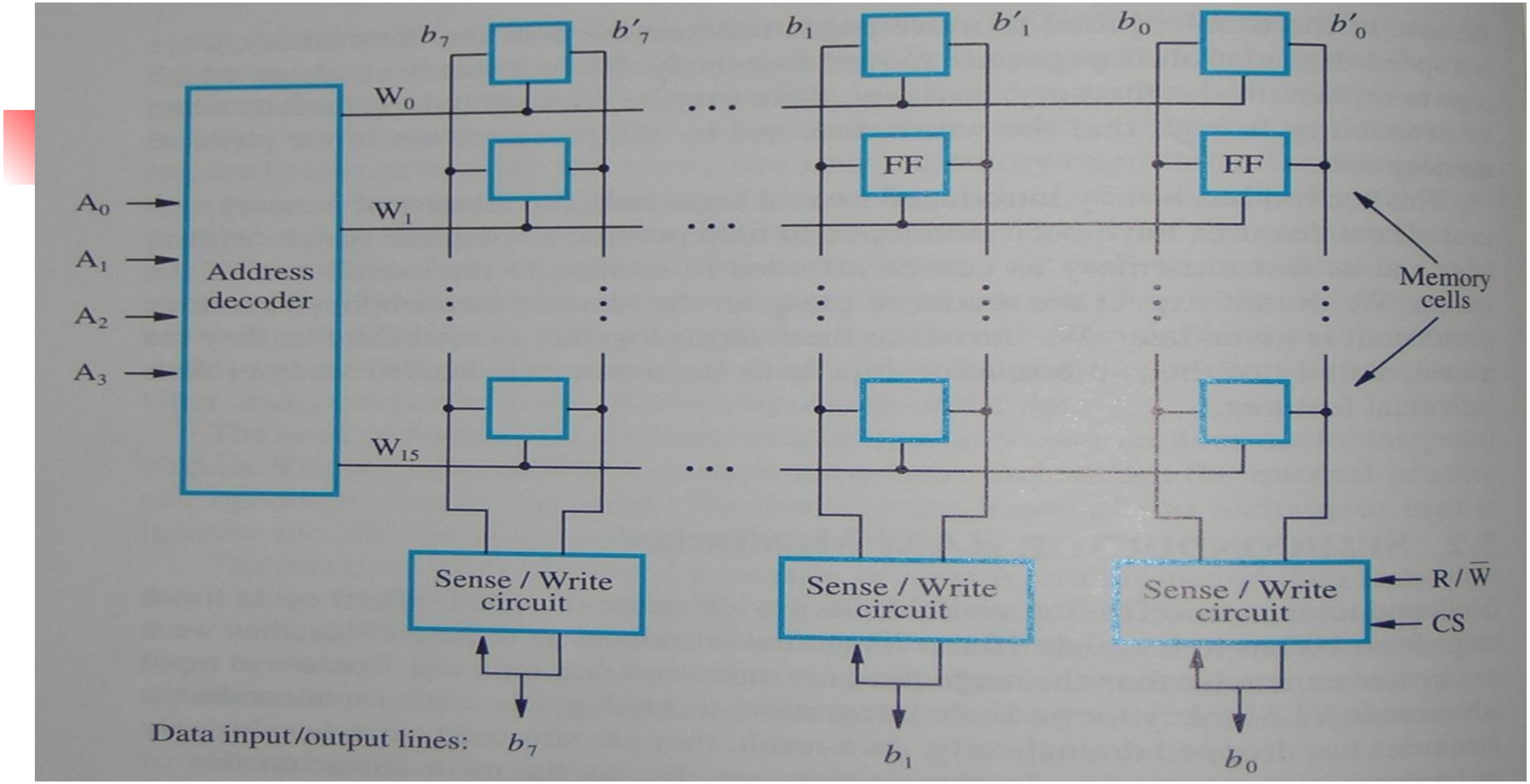
Flash Memory

- Similar to EPROM cell (slightly higher capacity, greater density, not complex, cheaper than EPROM)
- Rapid erase & write time than EEPROM
- Single cell can be read but can be written only an entire block of cells.
- Prior to writing, the previous of the block are erased (bulk erase possible).
- Power consumption is low, , making it attractive for use in equipment that is battery-driven. (fit Battery driven)
- Single flash chips are not sufficiently large, so   larger memory modules are implemented using flash cards and flash drives.
- Suitable for used as solid state disk such as CompactFlash, MemoryStick, SD, MD etc.

Internal organization of memory chips

- Each memory cell can hold one bit of information.
- Memory cells are organized in the form of an array.
- One row is one memory word.
- All cells of a row are connected to a common line, known as the “word line”.
- Word line is connected to the address decoder.
- Sense/write circuits are connected to the data input/output lines of the memory chip.

Organization of bit cells in a memory chip



16 words of 8 bits each: usually referred as 16x8 memory org.. It has 16 external connections: addr. 4, data 8, control: 2, power/ground: 2

1K memory cells: can be organized as 128x8 memory, external connections: ? 19(7+8+2+2)

If the same 1K is organized as 1Kx1: external connections: ?? 15 (10+1+2+2)

Organization of a 1Kx1 memory chip

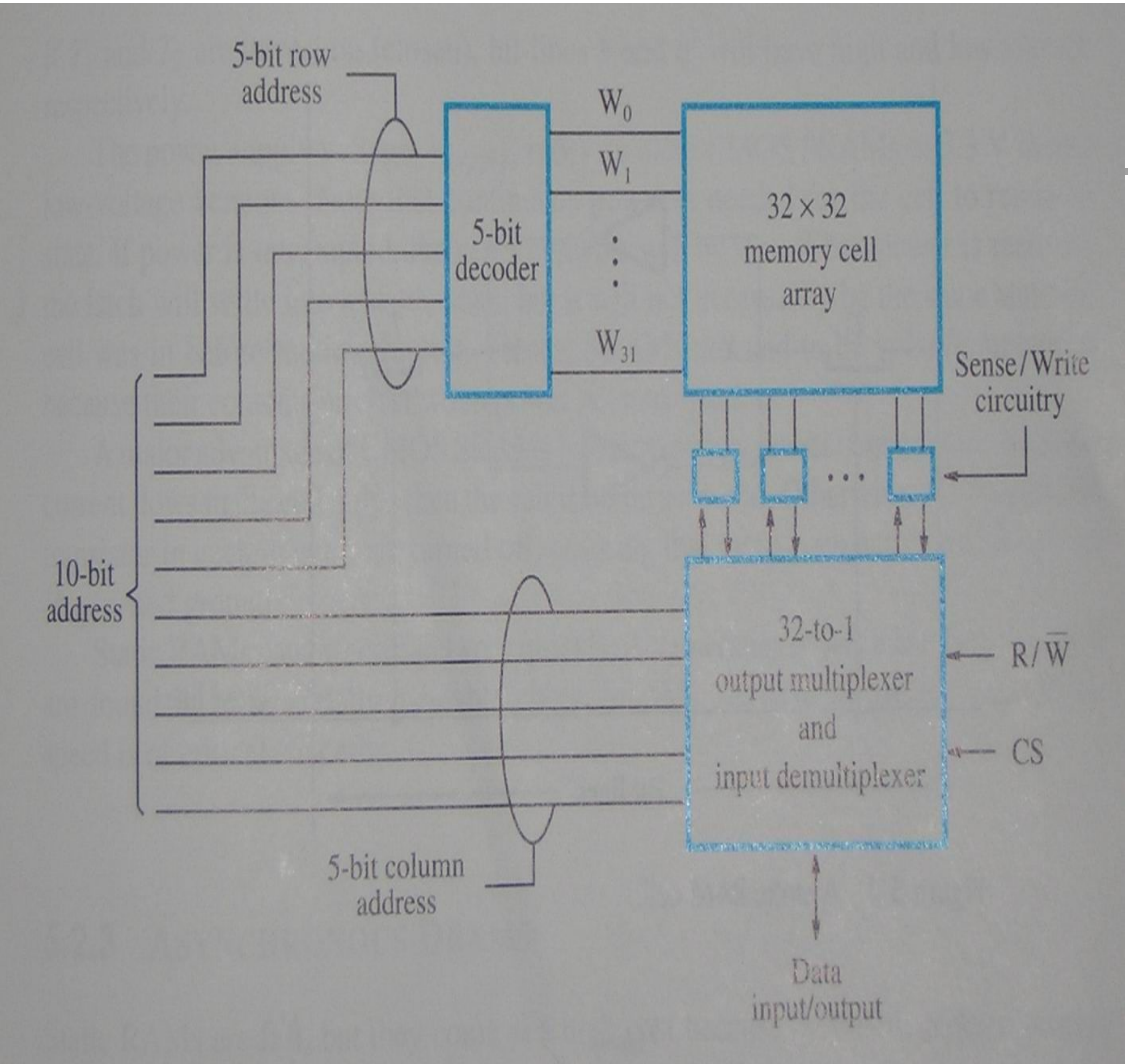


Figure 5.3 Organization of a 1K × 1 memory chip.

- 1K =1024;2^10
- 15pin (10ad+1data+2con+2PS)

- Alternatively this can be Organized as 128X8 which requires 19 pins (7ad+8data+2con+2PS)

- 1M can be organized as 1MX1
  2^20=1M so 25 pins (20+1+2+2)

- 4M can be organized as 1MX4
  28 pins (20+4+2+2)

Types of Semiconductor RAMs

- Static memories (SRAMs):
  - Retaining their state as long as the power is applied.
  - Uses latches and Transistors
  - Volatile memories □ Contents are lost when power is interrupted.
  - Access times of static RAMs -□ few nanoseconds.
  - However, the cost is usually high.
  - Example : Bipolar and MOS cells
- Dynamic memories (DRAMs):
  - Information is stored in the form of charge in the capacitors
  - Do not retain their state indefinitely.
  - Contents must be periodically refreshed.
  - Contents may be refreshed while accessing them for reading.

SRAM VS DRAM

SRAM

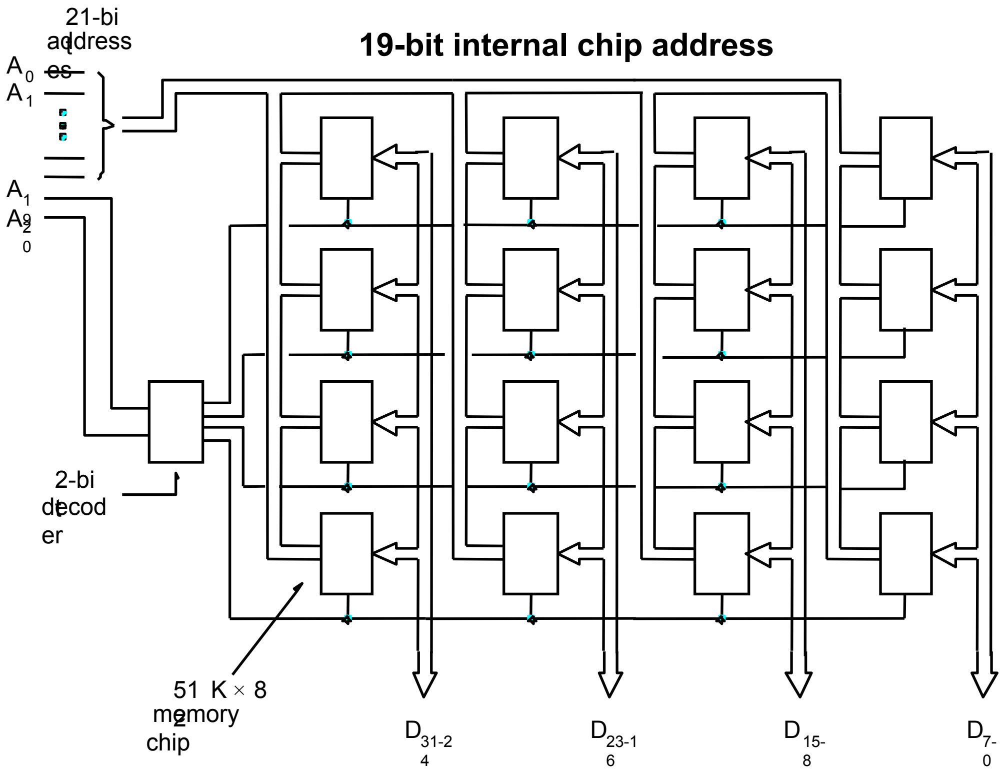
- Very fast
- Very Expensive
- Used in Cache memory and CPU register

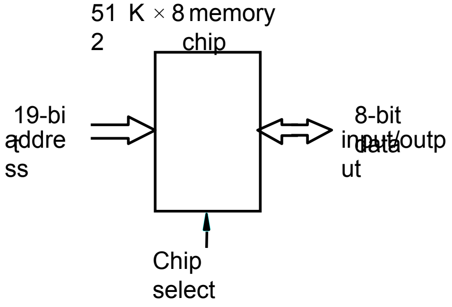
DRAM

- Slower than SRAM
- Cheaper than SRAM
- Used in most computer as main memory
- Need to be refreshed periodically

Large memory systems : Static memories

Implement a memory unit of 2M words of 32 bits each.Use 512x8 static memory chips.





Each column consists of 4 chips. Each chip implements one byte position.

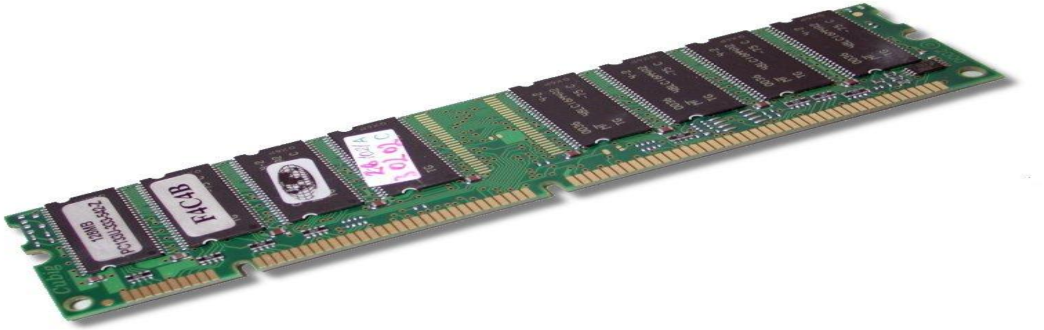
A chip is selected by setting its chip select control line to 1.

Selected chip places its data on the
data output line, outputs of other chips are in high impedance state.

21 bits to address a 32-bit word. High order 2 bits are needed to select the row, by activating the four Chip Select signals.

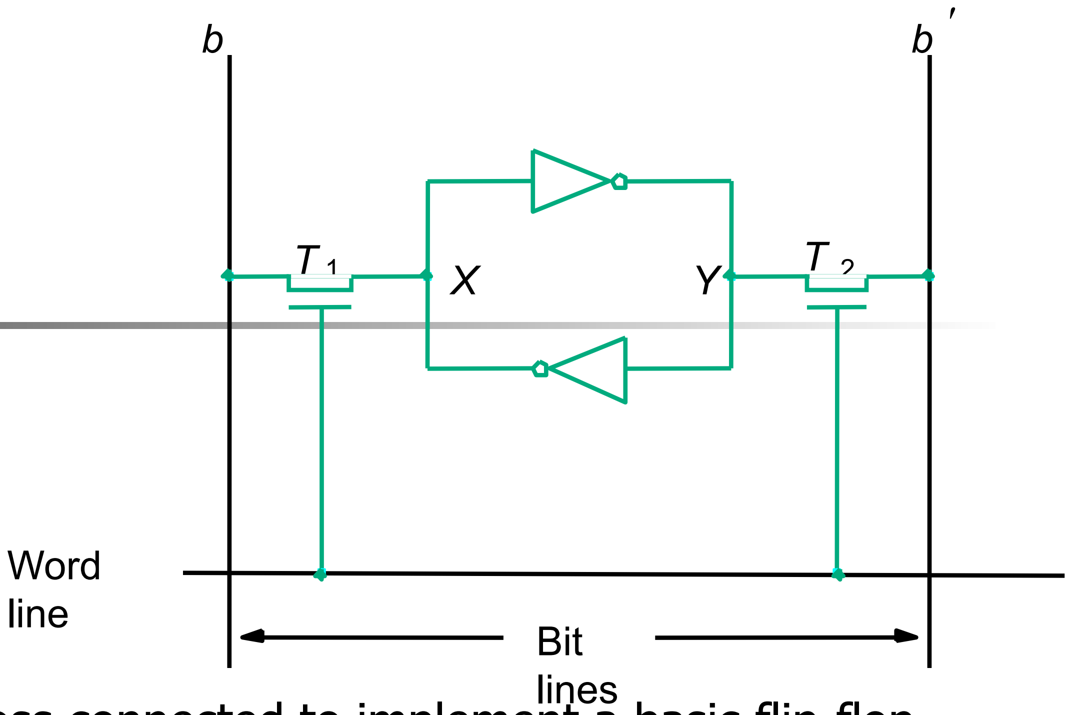
19 bits are used to access specific byte locations inside the selected Chip.

Dynamic memories



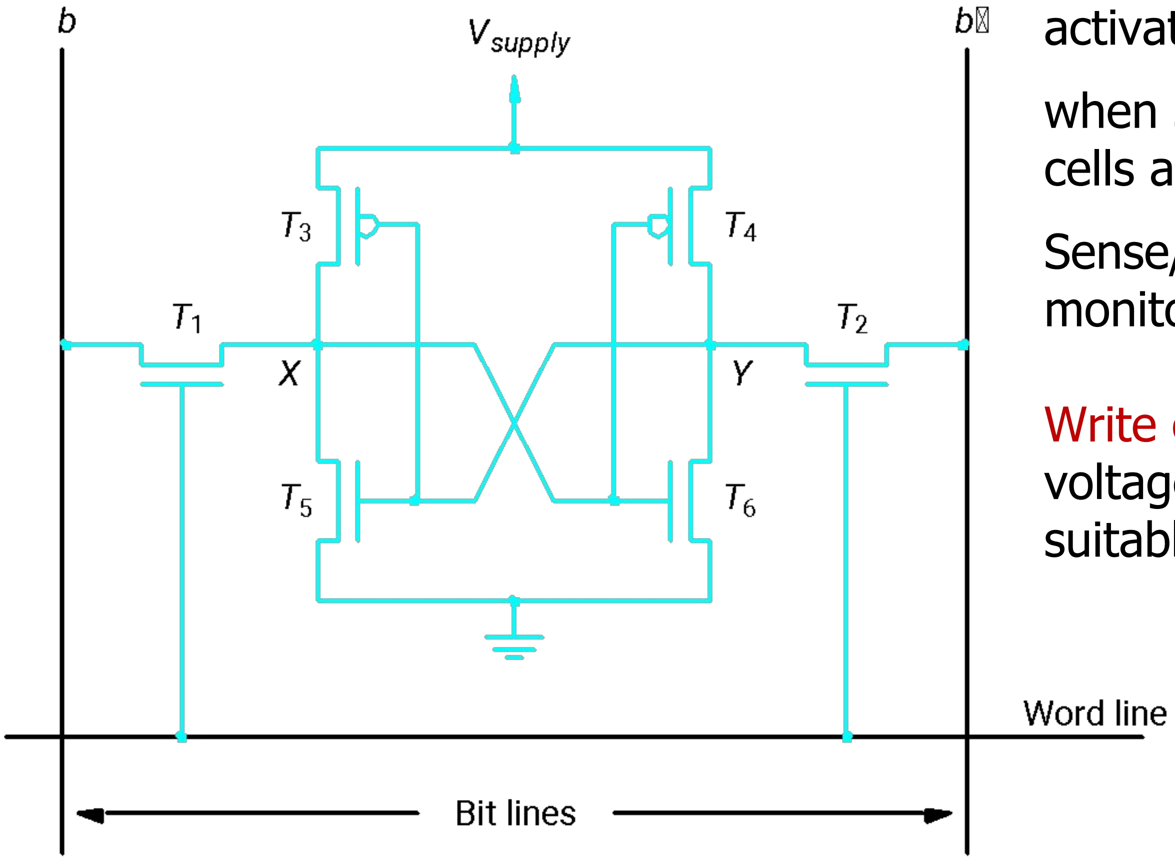
- Large dynamic memory systems can be implemented using DRAM chips in a similar way to static memory systems.
- Placing large memory systems directly on the motherboard will occupy a large amount of space.
  - Also, this arrangement is inflexible since the memory system cannot be expanded easily.
- Packaging considerations have led to the development of larger memory units known as SIMMs (Single In-line Memory Modules) and DIMMs (Dual In-line Memory Modules).
- Memory modules are an assembly of memory chips on a small board that plugs vertically onto a single socket on the motherboard.
  - Occupy less space on the motherboard. & Allows for easy expansion by replacement.

SRAM Cell



- Two transistor inverters are cross connected to implement a basic flip-flop.
- The cell is connected to one word line & two bits lines by transistors T1 & T2
- T1 & T2 are acting like switches under the control of word line.
- When word line is at ground level, the transistors are turned off and the latch retains its state.
- Read operation: the word line is activated to close switches T1 and T2. Sense/Write circuits at the bottom monitor the state of b and b′.
- Write operation : Reduce word line voltage from 2.5V to 0.3V
  - **Apply positive voltage (~ 3V) on b′ to store 1 and b to store 0**

SRAM cell : N channel MOS memory cell (CMOS)



Read operation: the word line is activated to close switches T1 and T2.

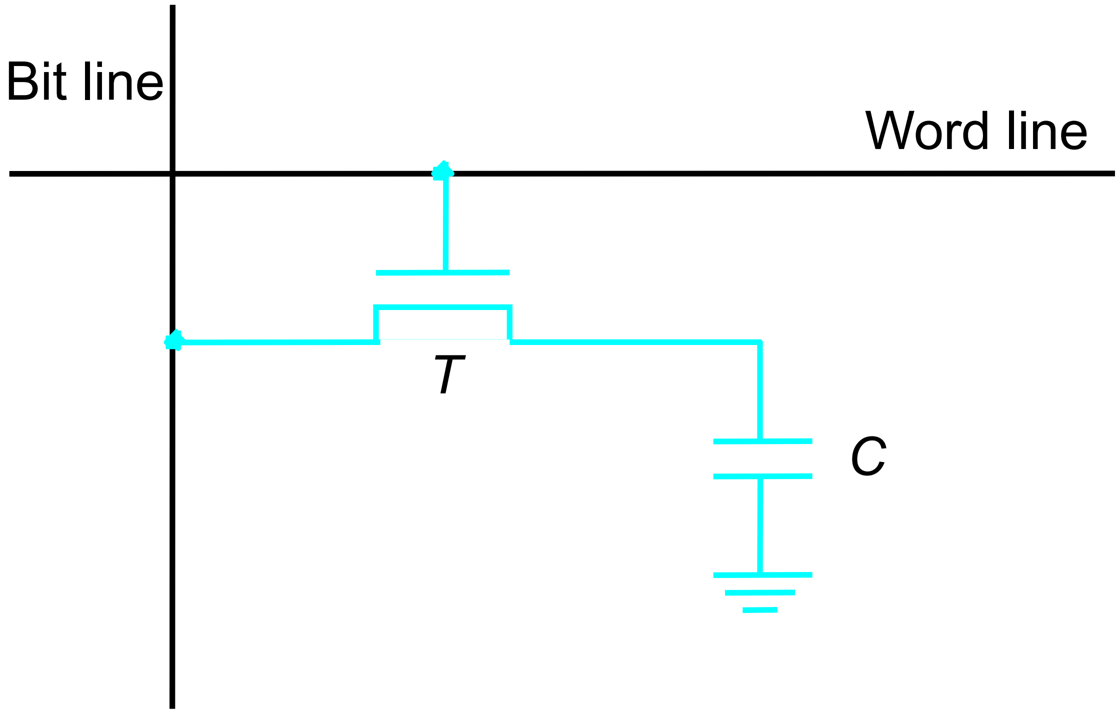
when switches are closed contents of cells are transformed to bit lines

Sense/Write circuits at the bottom monitor the state of b and b′.

Write operation : Reduce word line voltage from high to low, Apply suitable voltages on b and b′ to write

Asynchronous DRAMs

- Static RAMs are fast, but they cost more area and are more expensive.
- Dynamic RAMs (DRAMs) are cheap and area efficient, but they can not retain their state indefinitely – need to be periodically refreshed.

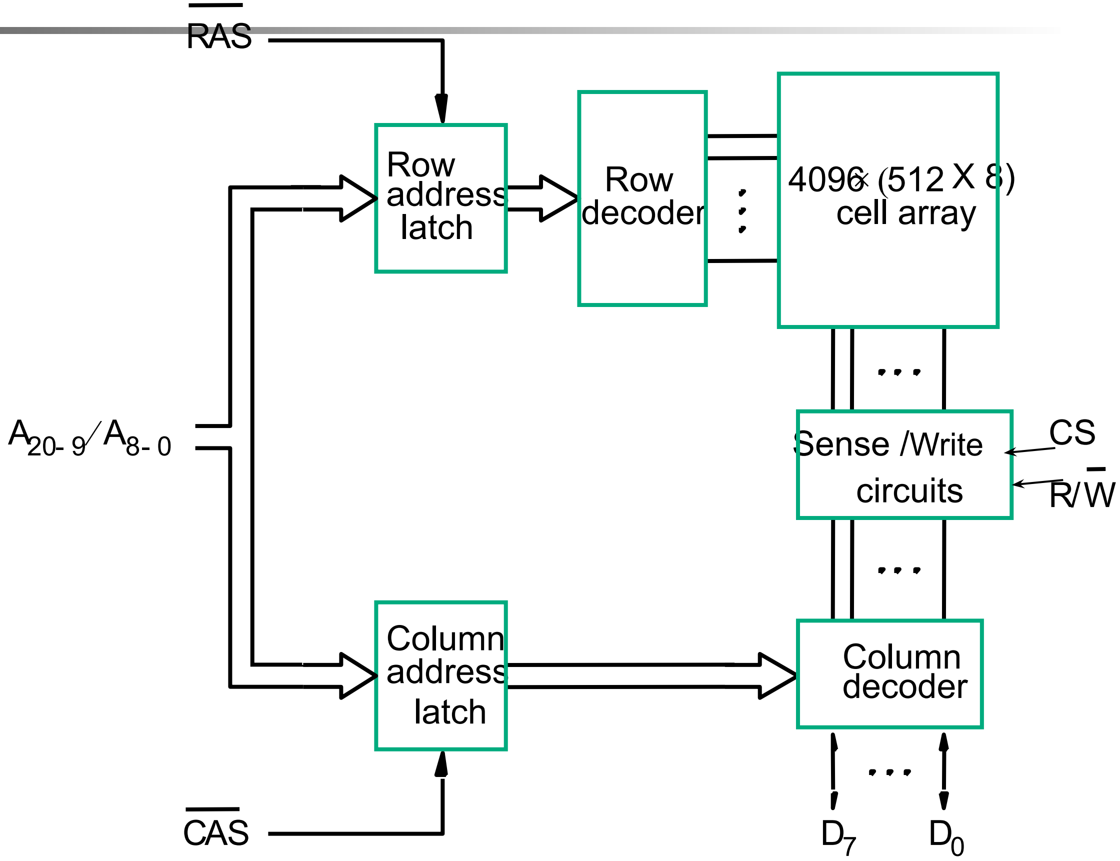


A single-transistor dynamic memory cell

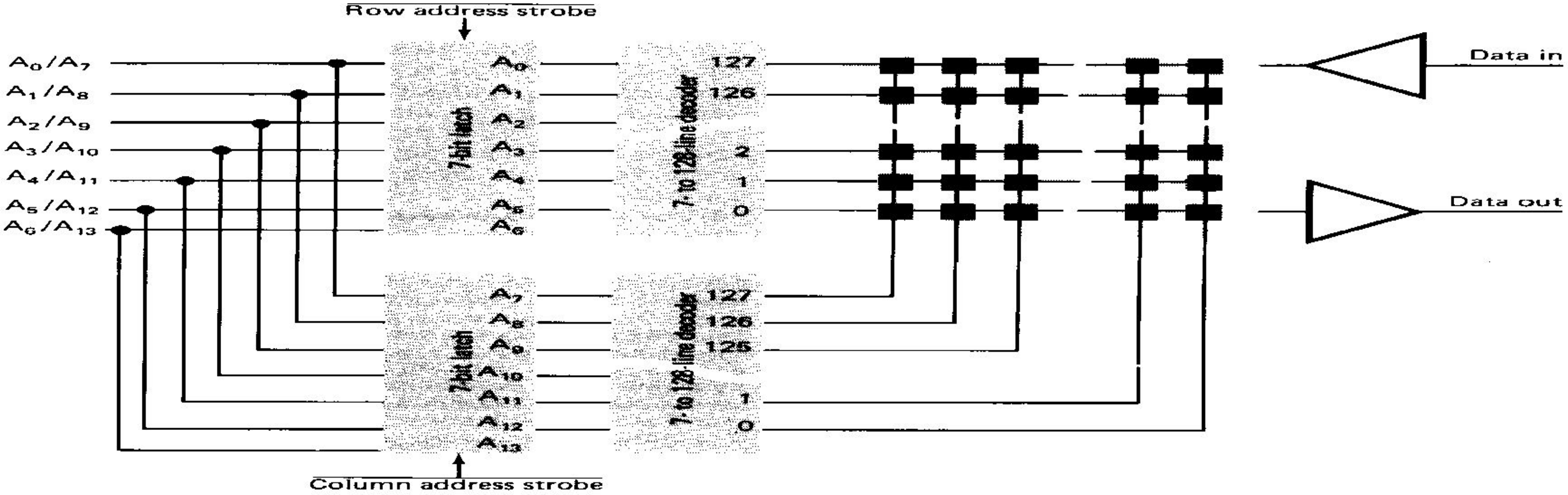
Asynchronous DRAMs

Internal organization 2MX8 Dynamic memory chip

- *Each row can store 512 bytes. 12 bits to select a row, and 9 bits to select a group in a row. Total of 21 bits.*
- *First apply the row address, RAS signal latches the row address. Then apply the column address, CAS signal latches the address.*
- *Timing of the memory unit is controlled by a specialized unit which generates RAS and CAS.*



DRAM: Multiplexed Row-Column addressing



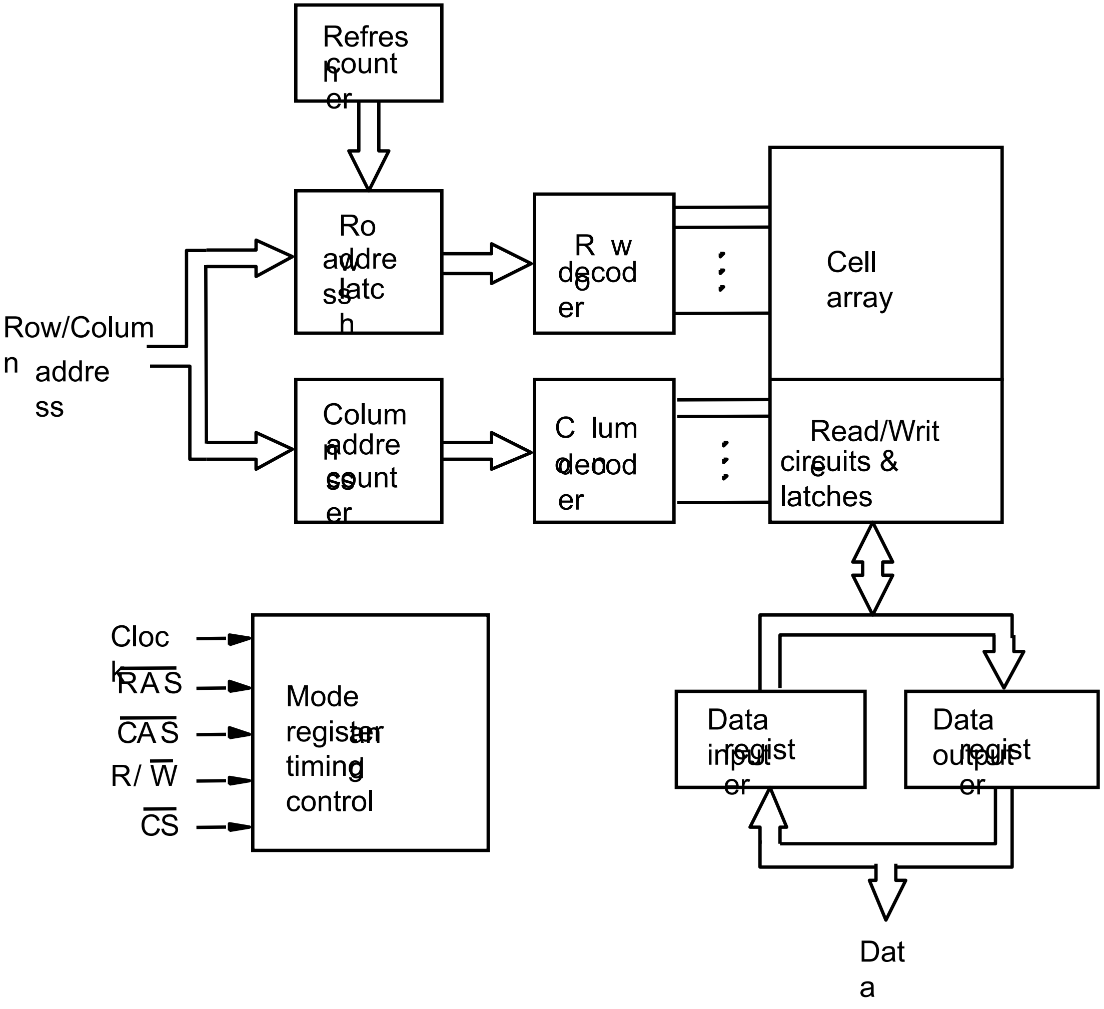
- AIM: Reducing Address pins of IC chip
- RAS = Row Address Strobe CAS = Column Address Strobe

- Address is divided into two parts:
  - High-order address bits select a row in the array.
  - They are provided first, and latched using RAS signal.
  - Low-order address bits select a column in the row.
  - They are provided later, and latched using CAS signal.

Synchronous DRAM (SDRAM)

- Need clock signal for synchronize operation
- Can be used with clock speed 100 and 133 MHz
- Has Built in refresh circuitry

Structure of Synchronous DRAMs



- *Operation is directly synchronized with processor clock signal.*
- *The outputs of the sense circuits are connected to a latch.*
- *During a Read operation, the contents of the cells in a row are loaded onto the latches.*
- *During a refresh operation, the contents of the cells are refreshed without changing the contents of the latches.*
- *Data held in the latches correspond to the selected columns are transferred to the output.*
- *For a burst mode of operation, successive columns are selected using column address counter and clock.*

CAS signal need not be generated externally. A new data is placed during raising edge of the clock

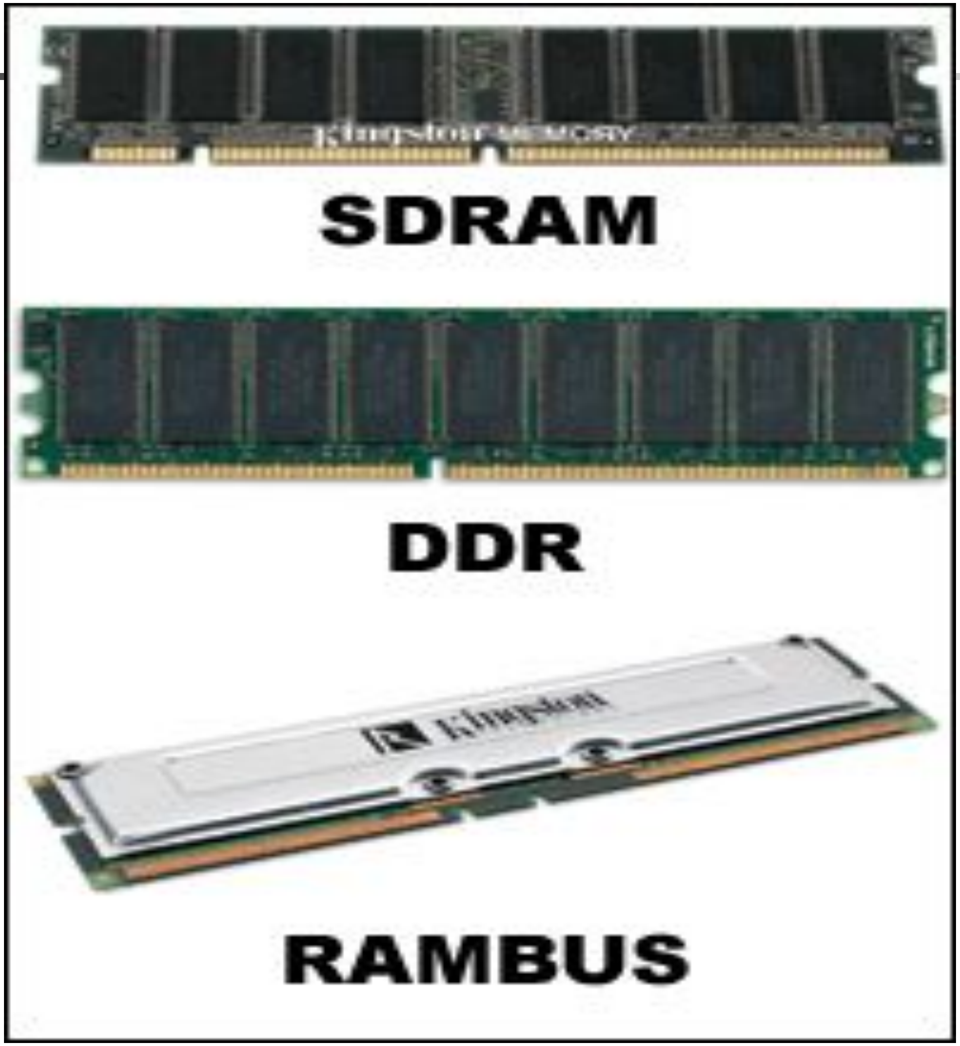
Latency and Bandwidth

- The speed and efficiency of data transfers among memory, processor, and disk have a large impact on the performance of a computer system.
- Memory latency – the amount of time it takes to transfer a word of data to or from the memory.
- Memory bandwidth – the number of bits or bytes that can be transferred in one second. It is used to measure how much time is needed to transfer an entire block of data.
- Bandwidth is not determined solely by memory. It is the product of the rate at which data are transferred (and accessed) and the width of the data bus.

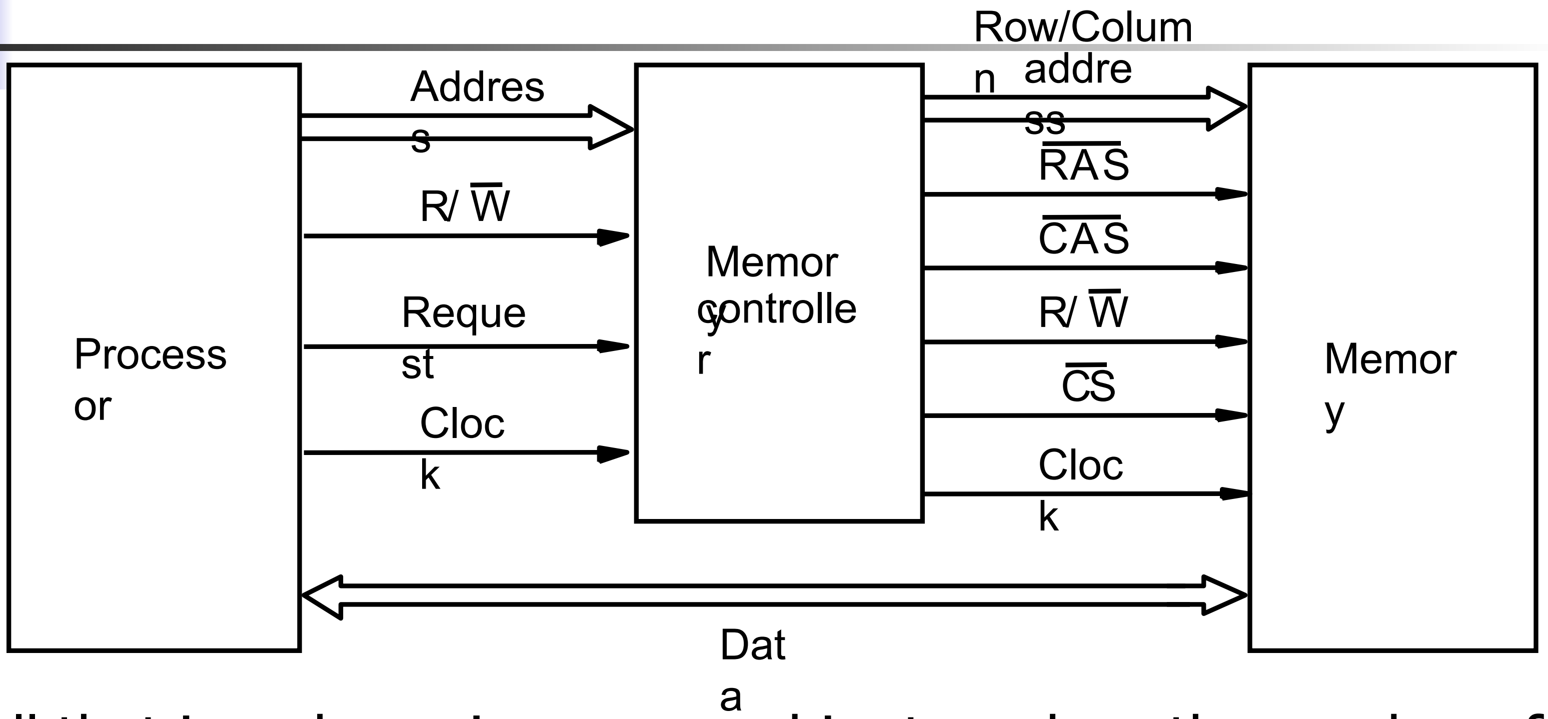
Double-Data-Rate SDRAM (DDR SDRAM)

- Standard SDRAM performs all actions on the rising edge of the clock signal.
- DDR SDRAM accesses the cell array in the same way, but transfers the data on both edges of the clock.
- The cell array is organized in two banks. Each can be accessed separately.
- DDR SDRAMs and standard SDRAMs are most efficiently used in applications where block transfers are prevalent.

3 Types of RAM modules



Memory controller



- Recall that in a dynamic memory chip, to reduce the number of pins, multiplexed addresses are used.
- However, a processor issues all address bits at the same time.
- In order to achieve the multiplexing, memory controller circuit is inserted between the processor and memory.

Cache Memories

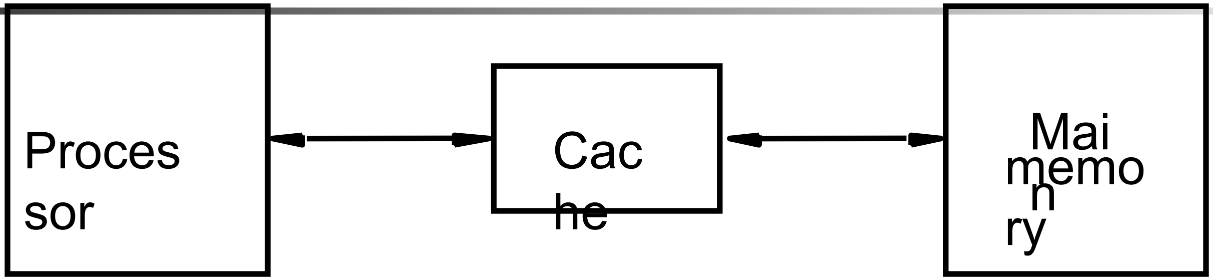
Cache Memories

- Processor is much faster than the main memory.
  - As a result, the processor has to waite for memory function complete
  - This is the Major obstacle towards achieving good performance.
- Speed of the main memory cannot be increased beyond a certain point.
- Cache memory is an architectural arrangement which makes the main memory appear faster to the processor than it really is.
- Cache memory is based on the property of computer programs known as "locality of reference".

Locality of Reference

- Analysis of programs indicates that many instructions in localized areas of a program are executed repeatedly during some period of time, while the others are accessed relatively less frequently.
  - These instructions may be the ones in a loop, nested loop or few procedures calling each other repeatedly.
  - This is called "locality of reference".
- Temporal locality of reference:
  - Recently executed instruction is likely to be executed again very soon.
- Spatial locality of reference:
  - Instructions with addresses close to a recently instruction are likely to be executed soon.

Cache memories



- ***Processor issues a Read request, a block of words is transferred from the main memory to the cache, one word at a time.***

- ***At any given time, only some blocks in the main memory are held in the cache. Which blocks in the main memory are in the cache is determined by a "<u>mapping function".</u>***

- ***When the cache is full, and a block of words needs to be transferred from the main memory, some block of words in the cache must be replaced. This is determined by a <u>"replacement algorithm".</u>***

Cache hit

- When processor request for the data, If the data is present in the cache Then it is called a Read or Write hit.

- Read hit:
  - The data is obtained from the cache.

- Write hit:
  - Cache has a replica of the contents of the main memory.
  - Contents of the cache and the main memory may be updated simultaneously. This is the write-through protocol.
  - Update the contents of the cache, and mark it as updated by setting a bit known as the dirty bit or modified bit. The contents of the main memory are updated when this block is replaced. This is write-back or copy-back protocol.

Cache miss

- *If the data is not present in the cache, then a Read miss or Write miss occurs.*

- **Read miss:**
  - *Block of words containing this requested word is transferred from the memory.*
  - *After the block is transferred, the desired word is forwarded to the processor.*
  - *The desired word may also be forwarded to the processor as soon as it is transferred without waiting for the entire block to be transferred. This is called load-through or early-restart.*

- **Write-miss:**
  - *If Write-through protocol is used, then the contents of the main memory are updated directly.*
  - *If write-back protocol is used, the block containing the addressed word is first brought into the cache. The desired word is overwritten with new information.*

Cache Coherence Problem

- A bit called as “valid bit” is provided for each block.
- If the block contains valid data, then the bit is set to 1, else it is 0.
- Valid bits are set to 0, when the power is just turned on.
- When a block is loaded into the cache for the first time, the valid bit is set to 1.
- Data transfers between main memory and disk occur directly bypassing the cache.
- When the data on a disk changes, the main memory block is also updated.
- However, if the data is also resident in the cache, then the valid bit is set to 0.
- What happens if the data in the disk and main memory changes and the write-back protocol is being used?
- In this case, the data in the cache may also have changed and is indicated by the dirty bit.
- The copies of the data in the cache, and the main memory are different. This is called the cache coherence problem.
- One option is to force a write-back before the main memory is updated from the disk.

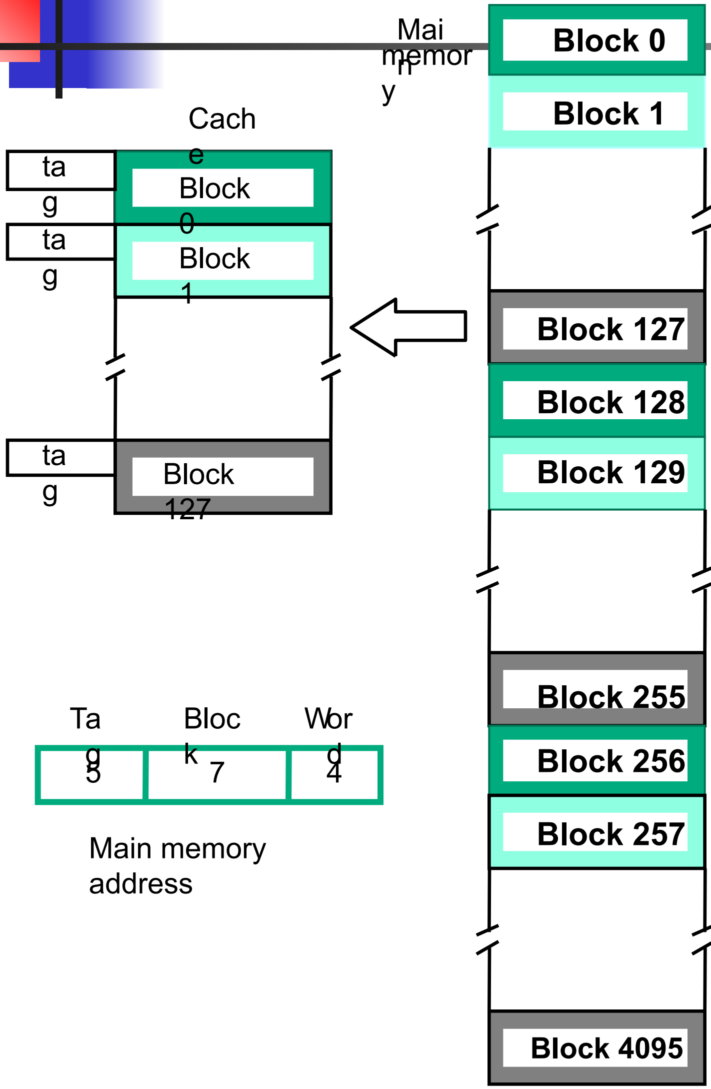
Mapping functions

- Mapping functions determine how memory blocks are placed in the cache.
- A simple processor example:
  - Cache consisting of 128 blocks of 16 words each.
  - Total size of cache is 2048 (2K) words.

  - Main memory is addressable by a 16-bit address.
  - Main memory has 64K words.

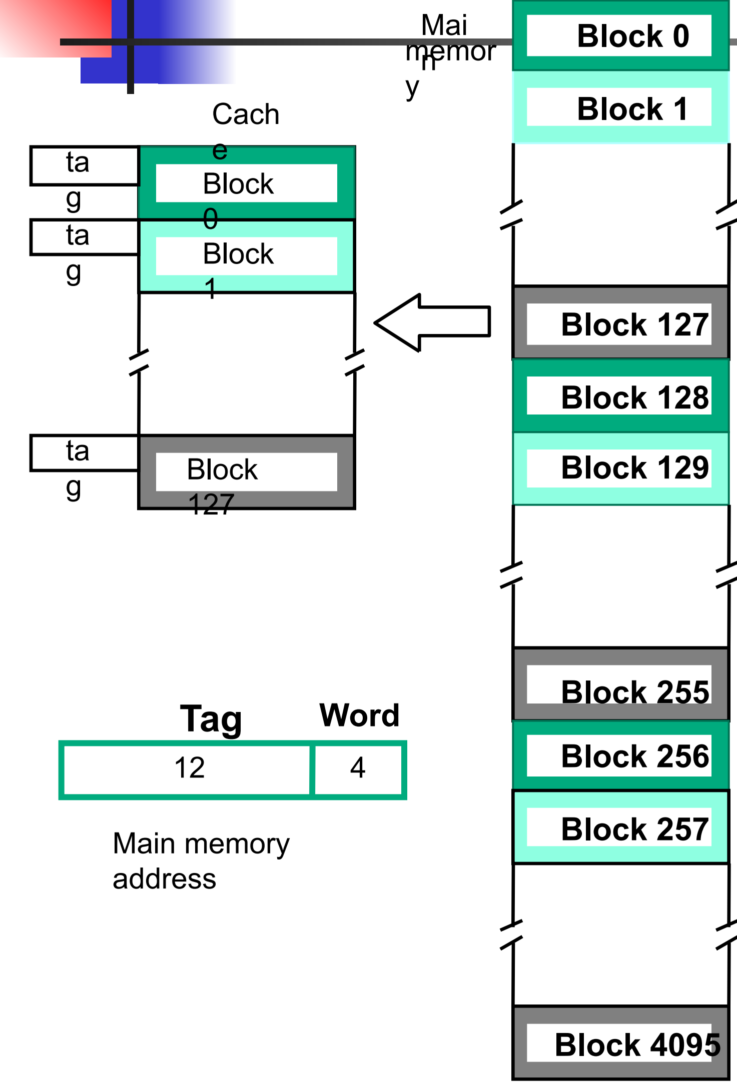
  - Main memory has 4K blocks of 16 words each.
- Three mapping functions:
  - Direct mapping
  - Associative mapping
  - Set-associative mapping.

Direct mapping



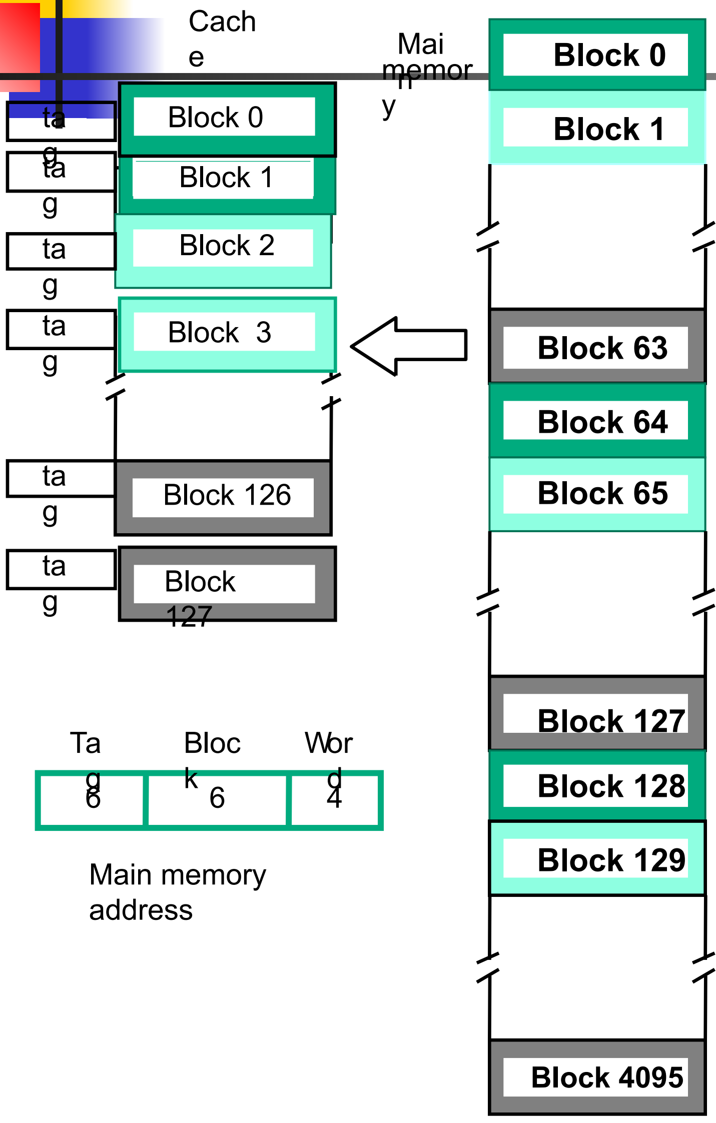
- Block j of the main memory maps to j modulo 128 of the cache. 0 maps to 0, 129 maps to 1.
- More than one memory block is mapped onto the same position in the cache.
- May lead to contention for cache blocks even if the cache is not full.
- Resolve the contention by allowing new block to replace the old block, leading to a trivial replacement algorithm.
- Memory address is divided into three fields:
  - Low order 4 bits determine one of the 16 words in a block.
  - When a new block is brought into the cache, the the next 7 bits determine which cache block this new block is placed in.
  - High order 5 bits determine which of the possible 32 blocks is currently present in the cache. These are tag bits.
- Simple to implement but not very flexible.

Associative mapping



- Main memory block can be placed into any cache position.
- Memory address is divided into two fields:
  - Low order 4 bits identify the word within a block
  - High order 12 bits or tag bits identify a memory block when it is resident in the cache.
- Flexible, and uses cache space efficiently.
- Replacement algorithms can be used to replace an existing block in the cache when the cache is full.
- Cost is higher than direct-mapped cache because of the need to search all 128 patterns to determine whether a given block is in the cache.

Set-Associative mapping



Set-associative mapping combination of direct and associative mapping.

Blocks of cache are grouped into sets.

Mapping function allows a block of the main memory to reside in any block of a specific set.

Divide the cache into 64 sets, with two blocks per set. Memory block 0, 64, 128 etc. map to block 0, and they can occupy either of the two positions.

Memory address is divided into three fields:

- *6 bit field determines the set number.*
- *High order 6 bit fields are compared to the tag fields of the two blocks in a set.*

Number of blocks per set is a design parameter.

- *One extreme is to have all the blocks in one set, requiring no set bits (fully associative mapping).*
- *Other extreme is to have one block per set, is the same as direct mapping.*

Memory systems
Performance considerations

- Performance of a processor depends on:
  - How fast machine instructions can be brought into the processor for execution.
  - How fast the instructions can be executed
- Hit rate & Miss penalty
- Hit rate can be improved by increasing block size, while keeping cache size constant
- Miss penalty can be reduced if load-through approach is used when loading new blocks into cache.
- Block sizes that are neither very small nor very large give best results.

Memory Interleaving

- Divides the memory system into a number of memory modules. Each module has its own address buffer register (ABR) and data buffer register (DBR).
- Arranges addressing so that successive words in the address space are placed in different modules.
- When requests for memory access involve consecutive addresses, the access will be to different modules.
- Since parallel access to these modules is possible, the average rate of fetching words from the Main Memory can be increased.

Other Performance Enhancements

Write buffer

- *Write buffer can be included for temporary storage of write requests.*
- *Fast write buffer can hold the block to be written, and the new block can be read first.*

- **Prefetching**
  - *Prefetch the data into the cache before they are actually needed, or a before a Read miss occurs.*

Replacement Algorithms

- When a new block is to be brought into the cache and all the positions that it may occupy are full, the cache controller must decide which of the old blocks to overwrite.

- In general, the objective is to keep blocks in the cache that are likely to be referenced in the near future.

- But, it is not easy---- The property of locality of reference in programs gives a clue to a reasonable strategy.

- there is a high probability that the blocks that have been referenced recently will be referenced again soon

LRU replacement algorithm.

- when a block is to be overwritten, it is sensible to overwrite the one that has gone the longest time without being referenced.
- This block is called the *least recently used (LRU) block, and the technique is called the LRU replacement algorithm.*
- The LRU algorithm has been used extensively.
- Several other replacement algorithms are also used in practice However, they are generally not as effective as the LRU in choosing the best blocks to remove
- Ex: The simplest algorithm is to randomly choose the block to be overwritten.

Virtual memories

- Physical main memory in a computer is generally not as large as the entire possible addressable space.
- Large programs that cannot fit completely into the main memory have their parts stored on secondary storage devices such as magnetic disks.
  - Pieces of programs must be transferred to the main memory from secondary storage before they can be executed.
  - Operating system automatically transfers data between the main memory and secondary storage.

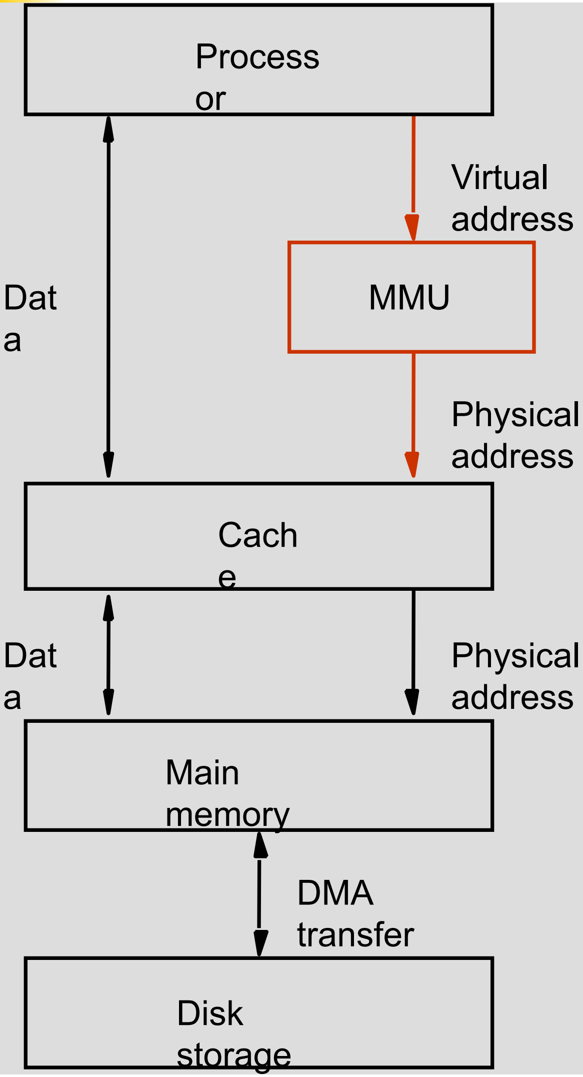
Virtual memories (contd..)

- Architectural solutions to increase the effective speed and size of the memory system.
  - Virtual memory □ to increase the effective size.
  - Cache memories □ to increase the effective speed.
- Techniques that automatically move program and data between main memory and secondary storage when they are required for execution are called virtual-memory techniques.
- Processor issues binary addresses for instructions and data.
  - These binary addresses are called logical or virtual addresses.
- Virtual addresses are translated into physical addresses by a combination of hardware and software subsystems.

Virtual memories ….

- Concepts of virtual memory are similar to the concepts of cache memory.
- Cache memory:
  - Introduced to bridge the speed gap between the processor and the main memory.
  - Implemented in hardware.
- Virtual memory:
  - Introduced to bridge the speed gap between the main memory and secondary storage.
  - Implemented in part by software.

Virtual memory organization



- *Memory management unit (MMU) translates virtual addresses into physical addresses.*

- *If the desired data or instructions are in the main memory they are fetched as described previously.*

- *If the desired data or instructions are not in the main memory, they must be transferred from secondary storage to the main memory.*

- *MMU causes the operating system to bring the data from the secondary storage into the main memory.*

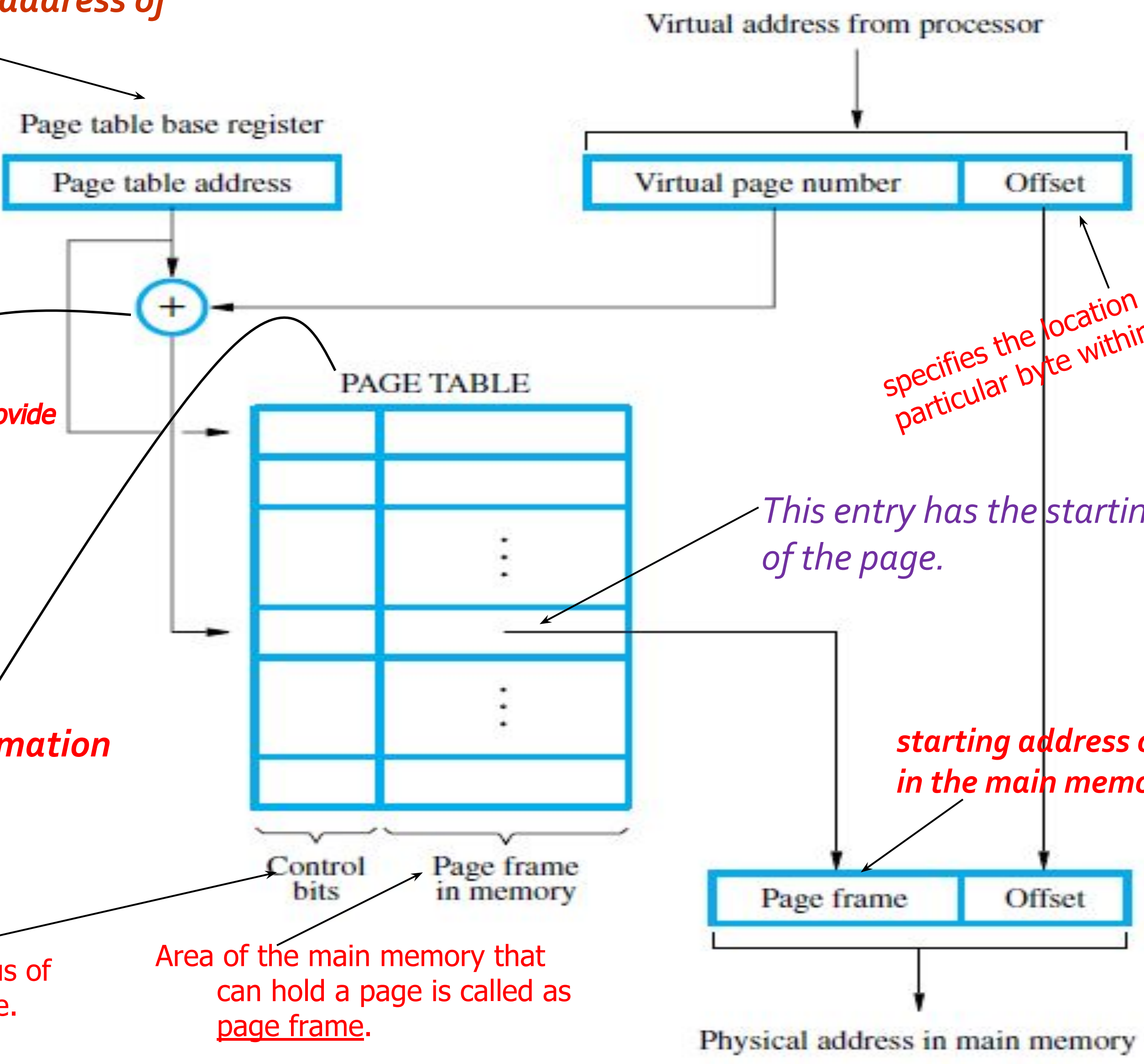
Address translation

Page:

- Assume that program and data are composed of fixed-length units called pages.
- A page consists of a block of words that occupy contiguous locations in the main memory.
- Page is a basic unit of information that is transferred between secondary storage and main memory.
- Size of a page commonly ranges from 2K to 16K bytes.

Address translation (contd..)

- Each virtual or logical address generated by a processor is interpreted as a virtual page number (high-order bits) plus an offset (low-order bits) that specifies the location of a particular byte within that page.
- Information about the main memory location of each page is kept in the page table.
- Area of the main memory that can hold a page is called as page frame.
- Starting address of the page table is kept in a page table base register (PTBR)



PTBR holds the address of the page table.

PTBR + virtual page number provide the entry of the page in the page table.

Page table holds information about each page.

specifies the location of a particular byte within that page

This entry has the starting location of the page.

starting address of the page in the main memory.

Current status of the page.

Area of the main memory that can hold a page is called as page frame.

Figure 8.25 Virtual-memory address translation.

Address translation (contd..)

Translation look aside Buffer (TLB)

- for every read and write access MMU uses The page table information.
- Where should the page table be located?
  - Ideal location for the page table is within the MMU.
- Page table is quite large.
- MMU is implemented as part of the processor chip.
- Impossible to include a complete page table on the chip.

Address translation (contd..)

Translation look aside Buffer (TLB)

- Page table is kept in the main memory.
- A copy of a small portion of the page table can be accommodated within the MMU.
  - Portion consists of page table entries that correspond to the most recently accessed pages.
- A small cache called as Translation Lookaside Buffer (TLB) is included in the MMU.
  - TLB holds page table entries of the most recently accessed pages.

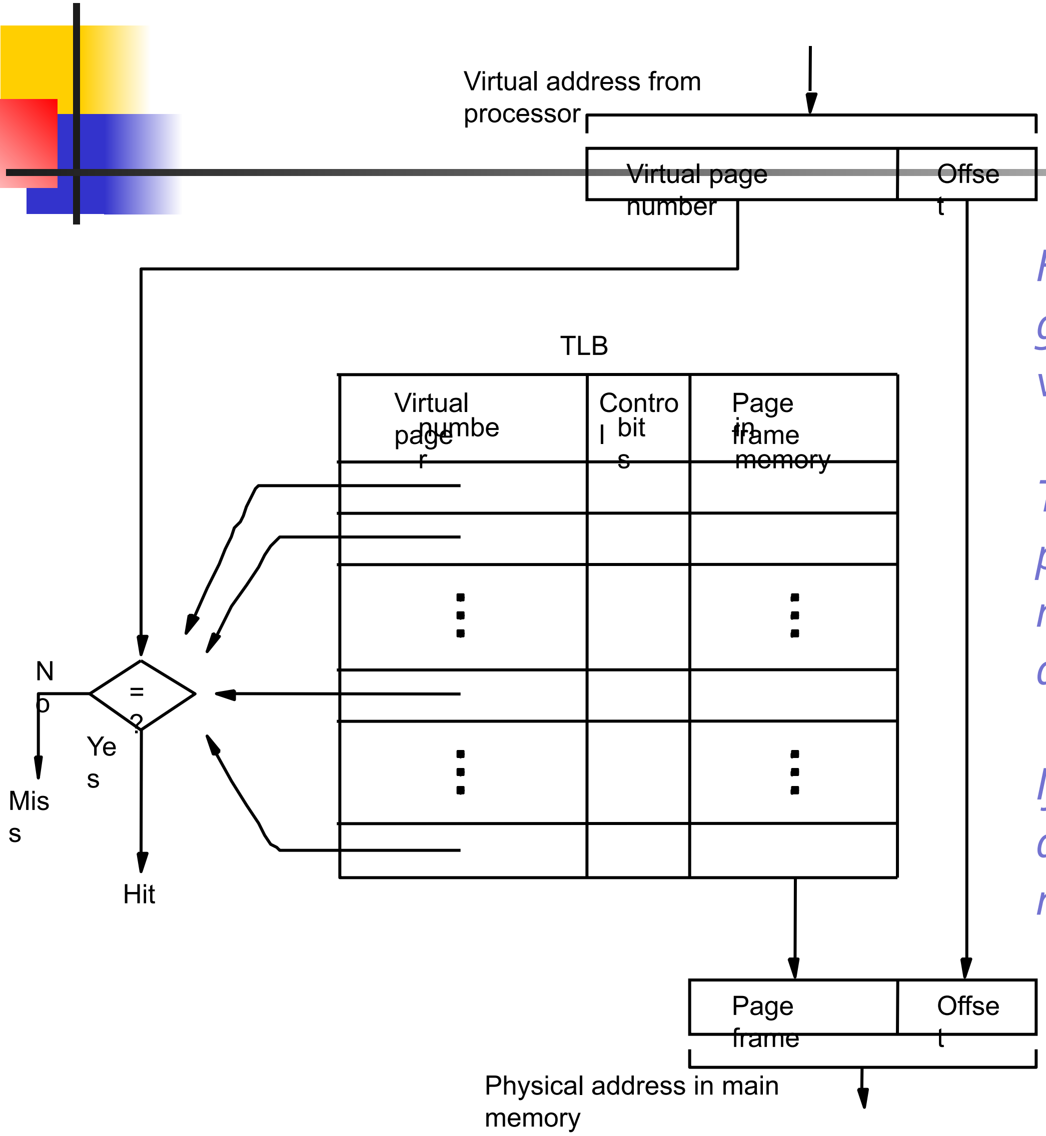
Address translation
TLB (contd..)

- Recall that cache memory holds most recently accessed blocks from the main memory.
  - Operation of the TLB and page table in the main memory is similar to the operation of the cache and main memory.

- Page table entry for a page includes:
  - **Address of the page frame where the page resides in the main memory.**
  - **Some control bits**.

- In addition to the above for each page, TLB must hold the virtual page number for each page.

Address translation TLB (contd..)



Virtual address from processor

Virtual page number

Offset

TLB

Virtual page number

Control bits

Page frame in memory

=?

No

Yes

Miss

Hit

Page frame

Offset

Physical address in main memory

<u>*Associative-mapped TLB*</u>

High-order bits of the virtual address generated by the processor select the virtual page.

These bits are compared to the virtual page numbers in the TLB. If there is a match, a hit occurs and the corresponding address of the page frame is read.

If there is no match, a miss occurs and the page table within the main memory must be consulted.

Set-associative mapped TLBs are found in commercial processors.

Address translation TLB(contd..)

- How to keep the entries of the TLB coherent with the contents of the page table in the main memory?
- Operating system may change the contents of the page table in the main memory.
  - Simultaneously it must also invalidate the corresponding entries in the TLB.
- A control bit is provided in the TLB to invalidate an entry.
- If an entry is invalidated, then the TLB gets the information for that entry from the page table.
  - Follows the same process that it would follow if the entry is not found in the TLB or if a “miss” occurs.

Address translation (contd..)
Page Fault

- What happens if a program generates an access to a page that is not in the main memory?
- In this case, a page fault is said to occur.
  - Whole page must be brought into the main memory from the disk, before the execution can proceed.
- Upon detecting a page fault by the MMU, following actions occur:
  - MMU asks the operating system to mediate by raising an exception.
  - Processing of the active task which caused the page fault is interrupted.
  - Control is transferred to the operating system.
  - Operating system copies the requested page from secondary storage to the main memory.
  - Once the page is copied, control is returned to the task which was interrupted.

Address translation (contd..)

- Servicing of a page fault requires transferring the requested page from secondary storage to main memory.
- This transfer may incur a long delay.
- While the page is being transferred the operating system may:
  - Suspend the execution of the task that caused the page fault.
  - Begin execution of another task whose pages are in the main memory.
- Enables efficient use of the processor.

Address translation (contd..)

- How to ensure that the interrupted task can continue correctly when it resumes execution?
- There are two possibilities:
  - Execution of the interrupted task must continue from the point where it was interrupted.
  - The instruction must be restarted.
- Which specific option is followed depends on the design of the processor.
- When a new page is to be brought into the main memory from secondary storage, the main memory may be full.
  - Some page from the main memory must be replaced with this new page.

Address translation (contd..)

- How to choose which page to replace?
  - This is similar to the replacement that occurs when the cache is full.
  - The principle of locality of reference (?) can also be applied here.
  - A replacement strategy similar to LRU can be applied.
- Since the size of the main memory is relatively larger compared to cache, a relatively large amount of programs and data can be held in the main memory.
  - Minimizes the frequency of transfers between secondary storage and main memory.

Address translation (contd..)

- A page may be modified during its residency in the main memory.
- When should the page be written back to the secondary storage?
- Recall that we encountered a similar problem in the context of cache and main memory:
  - Write-through protocol(?)
  - Write-back protocol(?)
- Write-through protocol cannot be used, since it will incur a long delay each time a small amount of data is written to the disk.

MEMORY

- END