MEMORY & I/O SYSTEMS

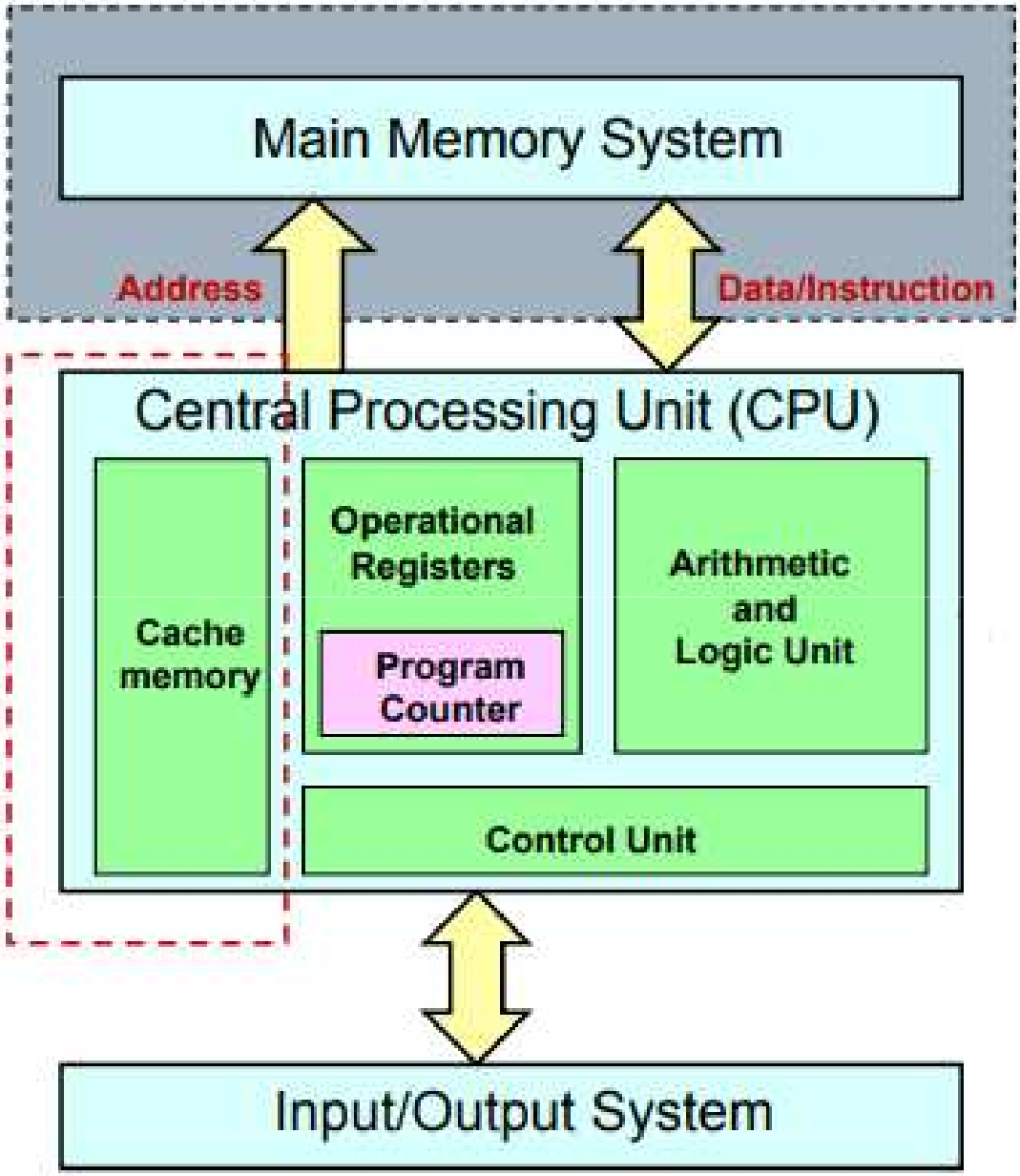
UNIT IV

Syllabus

- Memory Hierarchy – memory technologies – cache memory – measuring and improving cache performance – virtual memory
- TLB‘s – Accessing I/O Devices – Interrupts – Direct Memory Access – Bus structure – Bus operation – Arbitration – Interface circuits – USB.

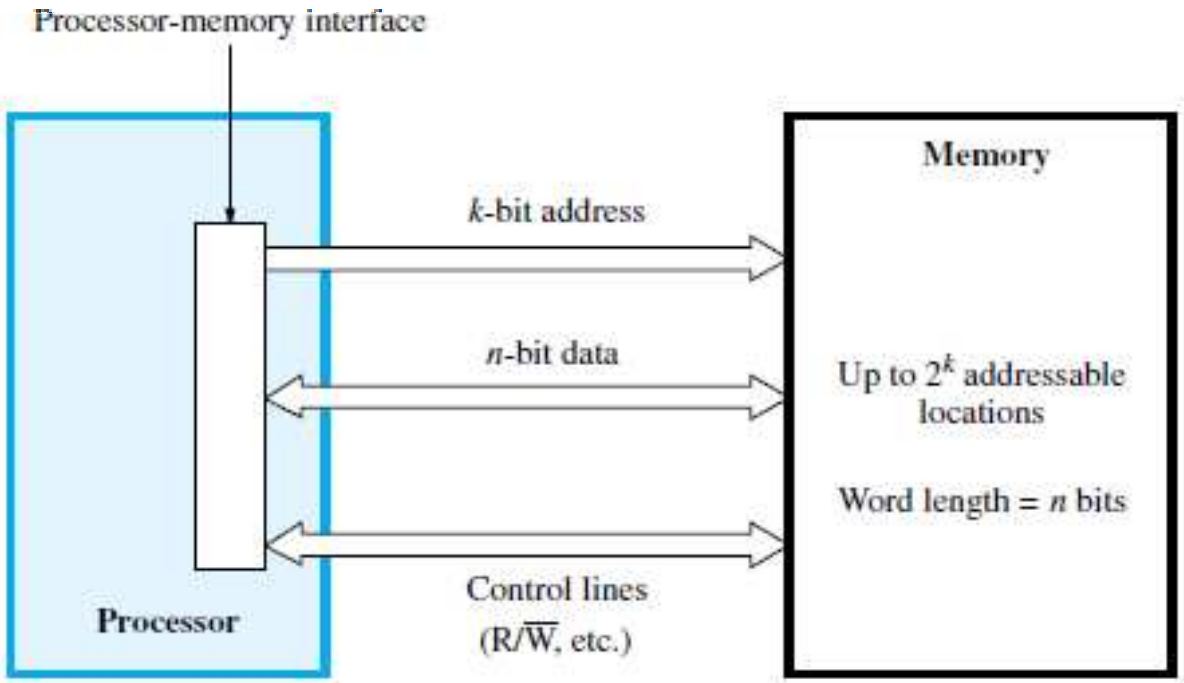
Memory

- Memory is a big array of bytes
- *Technologies*
  - Writable vs. non-writable
    - ROM (Read only memory)
    - RAM (Read write memory but known as Random Access memory)
  - Volatility
    - storage that only maintains its data while the device is powered
    - if memory retain values when powered off (Non-volatile)
  - Technology
    - Main Memory
      - ROM: Data masked at fabrication time
        - » PROM: Once programmable (specialized programmers)
        - » EPROM: Erasable and then programmable
        - » EEPROM: Electrically (selectively) erasable
      - RAM
        - » access time to any location is the same, independent of the location's address.
        - » cache memory:
          - SRAM
        - » Main Memory
          - DRAM Dynamic: Slow but cheap. DRAM, EDORAM, SDRAM, DDR-RAM
    - Flash Memory
      - Non-volatile memory
      - secondary memory in Personal Mobile Devices
    - Magnetic disk
    - Virtual Memory
      - program sees a memory that is much larger than the computer's physical main memory


Main Memory System
Address
Data/Instruction
Central Processing Unit (CPU)
Cache memory
Operational Registers
Program Counter
Arithmetic and Logic Unit
Control Unit
Input/Output System

Memory

- 16-bit addresses 2^{16} = 64K memory locations
- 32-bit addresses 2^{32} = 4G memory locations
- Memory store and retrieve - data in word length
- MFC (Memory Function Complete) (Control Signal)
  - asserted when memory R/W has been completed
- Chip selects (Control Signal)
  - Select chips from an array within a memory device



Memory

- *Memory Access Time:*
  - time that elapses between **the initiation** of an operation to transfer a word of data and the **completion** of that operation
- *Memory Cycle Time:*
  - minimum time delay required between the initiation of two successive memory operations,
  - time between two successive Read operations

- Block transfer – bulk data transfer

Principle of Locality

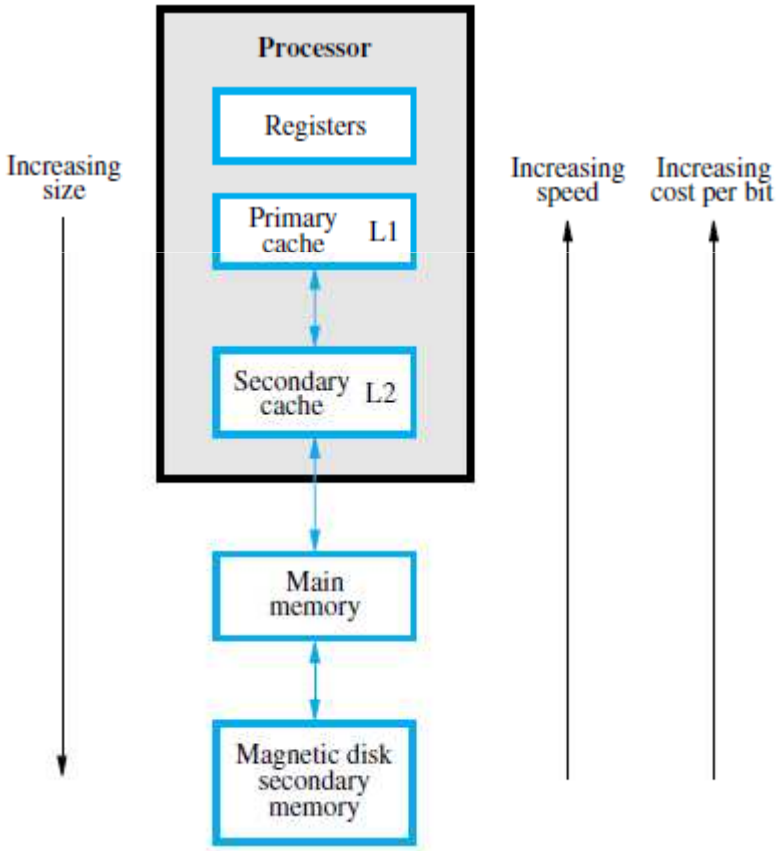
- Programs access a small proportion of their address space at any time
- Temporal locality (time)
  - Items accessed recently are likely to be accessed again soon
  - e.g., instructions in a loop
- Spatial locality (space)
  - Items near those accessed recently are likely to be accessed soon
  - E.g., sequential instruction access, array data

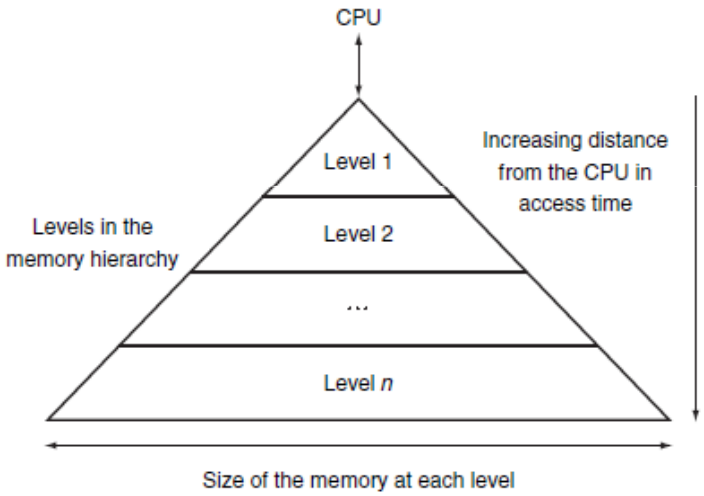
Taking Advantage of Locality

- Memory hierarchy
- Store everything on disk
- Copy recently accessed (and nearby) items from disk to smaller DRAM memory
  - Main memory
- Copy more recently accessed (and nearby) items from DRAM to smaller SRAM memory
  - Cache memory attached to CPU

MEMORY HIERARCHY

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





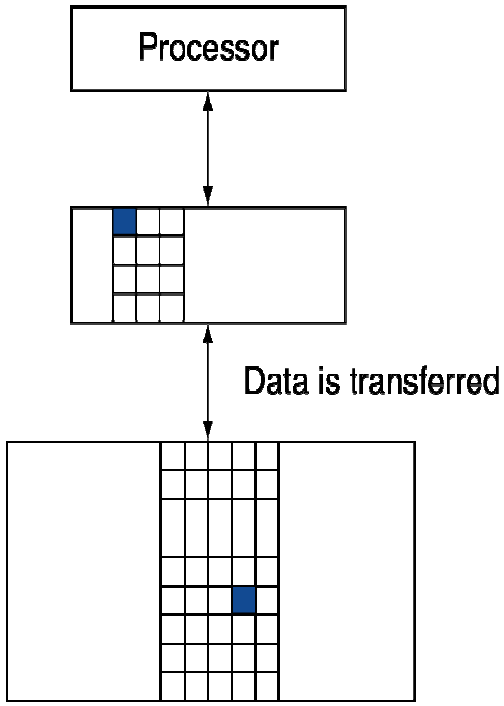
Speed, Size, and Cost

- 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.**
  - Goal is to present the user with as much memory as is available in the cheapest technology, while providing access at the speed offered by the fastest memory.
    - 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.
      - Storage provided by DRAMs is higher than SRAMs, but is still less than what is necessary.
    - Magnetic disks:
      - Secondary storage such as magnetic disks provide a large amount of storage, but is much slower than DRAMs.

Memory Hierarchy

- Block : unit of copying
- Hit: If accessed data is present in upper level
  - **Hit ratio:** hits/accesses
- Miss: If accessed data is absent
  - Then the block is copied from lower level
    - Time taken: miss penalty
    - **Miss ratio:** misses/accesses
      = 1 – hit ratio
- **Hit time: is the time to access the upper level**
  - which includes the time needed to determine whether the access is a hit or a miss
- **Miss penalty: is the time to replace a block in the upper level with** the corresponding block from the lower level, plus the time to deliver this block to the processor

The **hit rate, or *hit ratio, is the fraction of memory accesses*** found in the upper level; it is often used as a measure of the performance of the memory hierarchy.
The **miss rate (1–hit rate) is the fraction of memory accesses** not found in the upper level.



Memory Technologies

- Four primary technologies
  1. *cache memory:* SRAM
  2. Main Memory: DRAM
  3. Flash memory
  4. Magnetic disk

Memory technology	Typical access time	$ per GiB in 2012
SRAM semiconductor memory	0.5–2.5 ns	$500–$1000
DRAM semiconductor memory	50–70 ns	$10–$20
Flash semiconductor memory	5,000–50,000 ns	$0.75–$1.00
Magnetic disk	5,000,000–20,000,000 ns	$0.05–$0.10

Flash memory

- Type of *electrically erasable programmable read-only memory* (**EEPROM**).
- Writes can **wear out** flash memory bits
  - To cope with such limits, most flash products include a controller to spread the writes by remapping blocks that have been written many times to less trodden blocks.
  - *wear leveling*
    - *With wear leveling, personal* mobile devices are very unlikely to exceed the write limits in the flash.
    - Such wear leveling lowers the potential performance of flash, but it is needed unless higher level software monitors block wear.

Disk Memory

- collection of platters
- rotate on a spindle at 5400 to 15,000 revolutions per minute
- metal platters are covered with magnetic recording material on both sides
- a movable *arm* containing a small electromagnetic coil called a ***read-write head***
  - *is located just above* each surface to read and write information on a hard disk
  - The disk heads for each surface are connected together and move in conjunction, so that every head is over the same track of every surface.
- The entire drive is permanently sealed
  - to control the environment inside the drive
  - allows the disk heads to be much closer to the drive surface.
- Each disk surface is divided into concentric circles, called **tracks.**
  - There are typically tens of thousands of tracks per surface
  - Each track is in turn divided into **sectors** that contain the information
  - each track may have thousands of sectors
  - Sectors are typically 512 to 4096 bytes in size.
  - The sequence recorded on the magnetic media is a sector number, a gap, the information for that sector including error correction code, a gap, the sector number of the next sector, and so on.
- The term ***cylinder*** *is used* to refer to all the tracks under the heads at a given point on all surfaces.

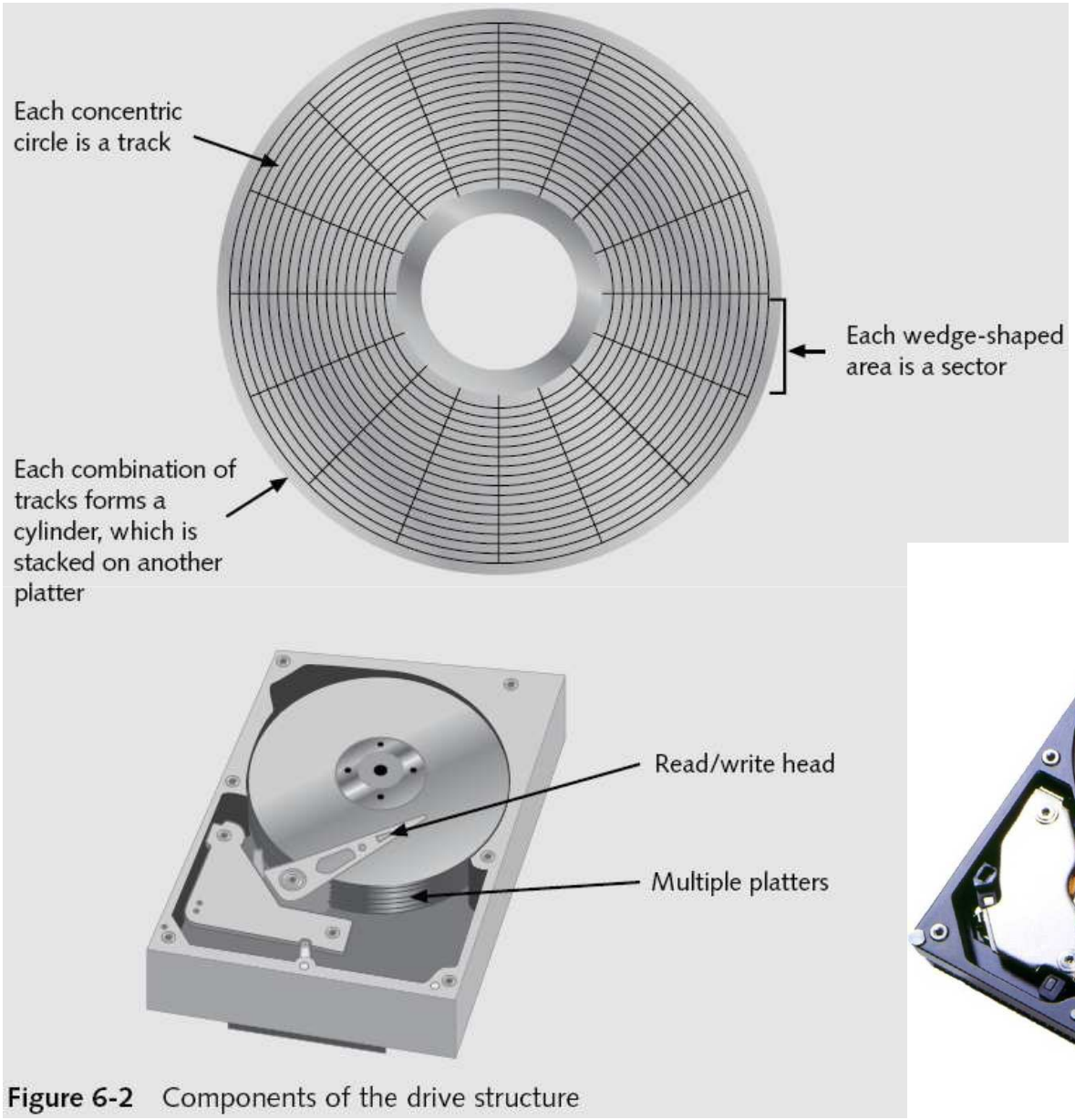


Figure 6-2 Components of the drive structure

https://www.youtube.com/watch?v=NtPc0jI21i0

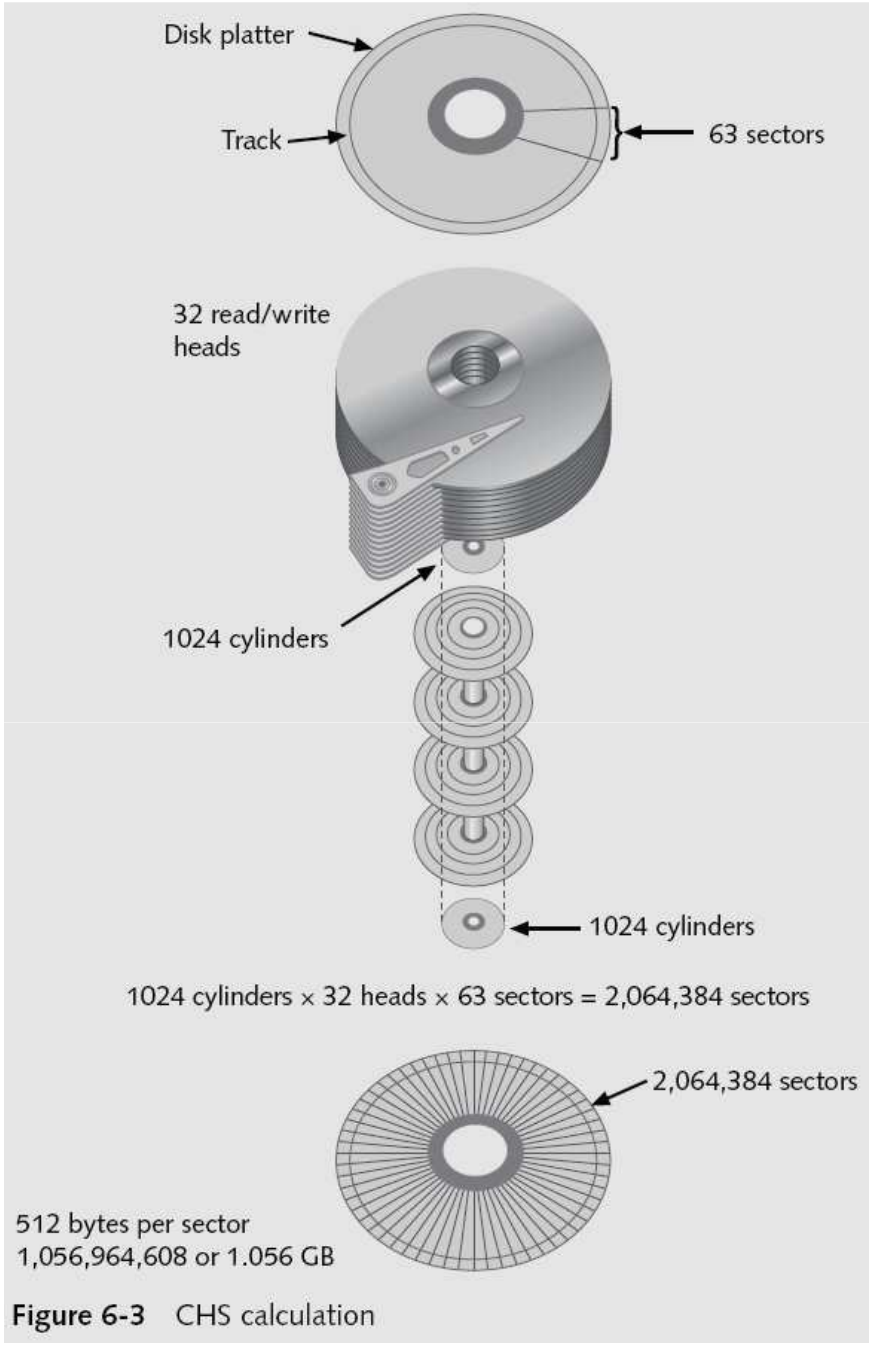


Figure 6-3 CHS calculation

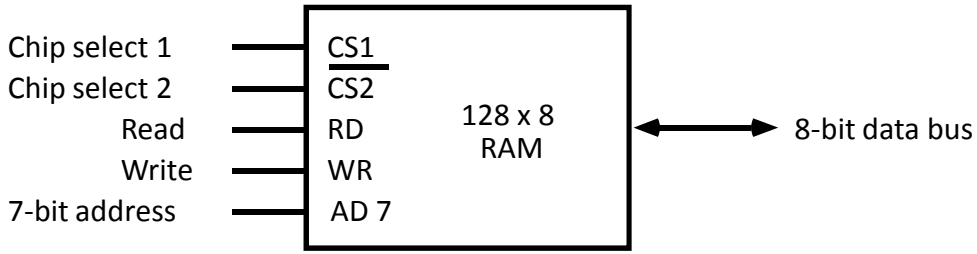
Disk Memory

- To access data - three-stage process
- The first step is to position the head over the proper track
  - Th is operation is called a **seek, and the time to move the head to the desired track is called the *seek time.***
  - Disk manufacturers report minimum seek time, maximum seek time, and average seek time in their manuals. Th e fi rst two are easy to measure, but the average is open to wide interpretation because it depends on the seek distance.
  - Th e industry calculates average seek time as the sum of the time for all possible seeks divided by the number of possible seeks.
- Once the head has reached the correct track, we must wait for the desired sector to rotate under the read/write head.
  - This time is called the **rotational latency or rotational delay**
  - **The average latency to the desired information is halfway around** the disk.
- The last component of a disk access, *transfer time, is the time to transfer a block* of bits.

MAIN MEMORY

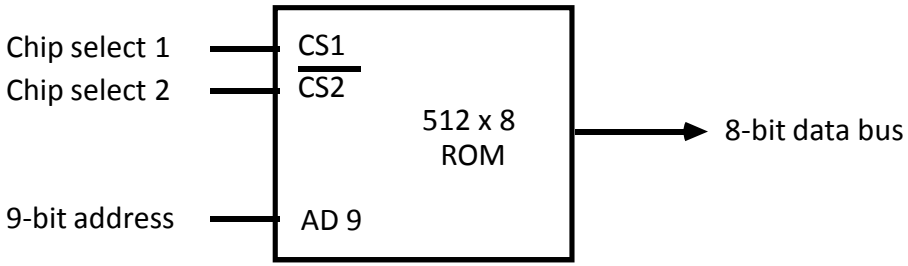
RAM and ROM Chips

Typical RAM chip



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

Typical ROM chip

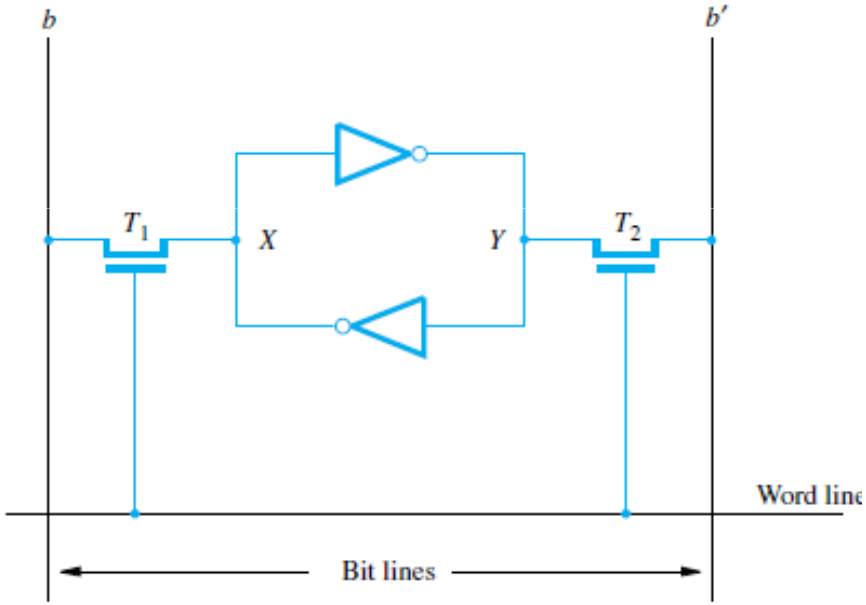


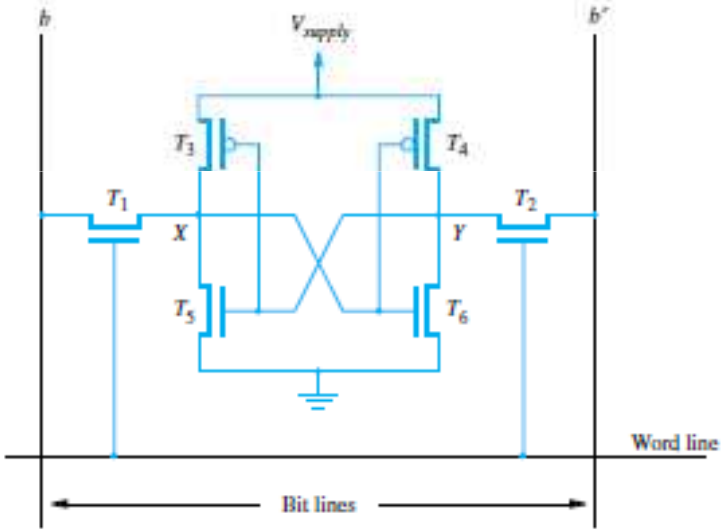
Static RAMs

- Static RAMs (SRAMs):
  - Integrated circuits - memory arrays
  - single access port - provide either a read or a write
  - Fixed access time to any datum
  - The read and write access times may differ
  - don’t need to refresh - so the access time is very close to the cycle time - few nanoseconds.
  - use six to eight transistors per bit to prevent the information from being disturbed when read
  - Needs only minimal power to retain the charge in standby mode
  - Consist of circuits that are capable of retaining their state as long as the power is applied.
  - Volatile memories, because their contents are lost when power is interrupted.
  - the cost is usually high

Static RAM (SRAM)

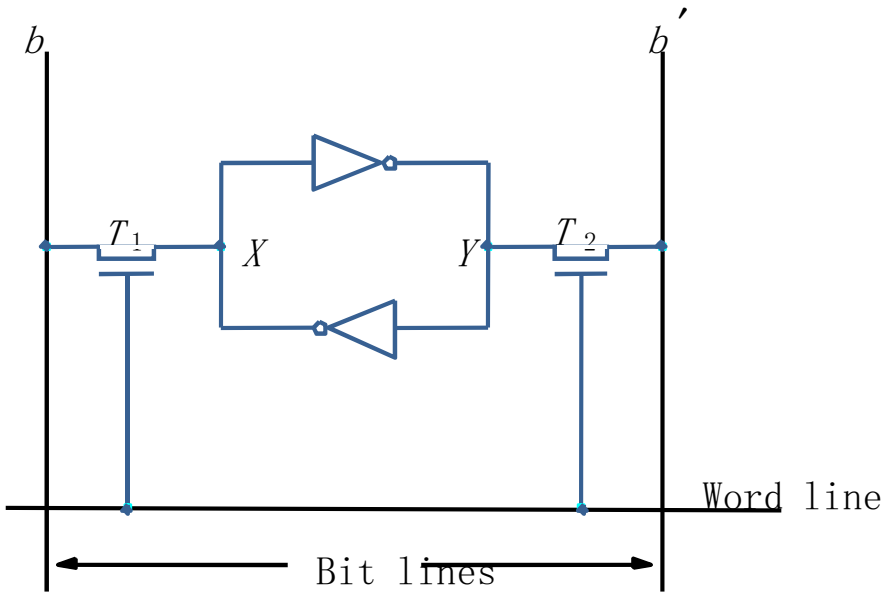
- Retaining their state as long as power is applied





SRAM Cell

- Two transistor inverters are cross connected to implement a basic flip-flop.
- The cell is connected to one word line and two bits lines by transistors T1 and T2
- When word line is at ground level, the transistors are turned off and the latch retains its state
- Read operation: In order to read state of SRAM cell, the word line is activated to close switches T1 and T2. Sense/Write circuits at the bottom monitor the state of b and b’

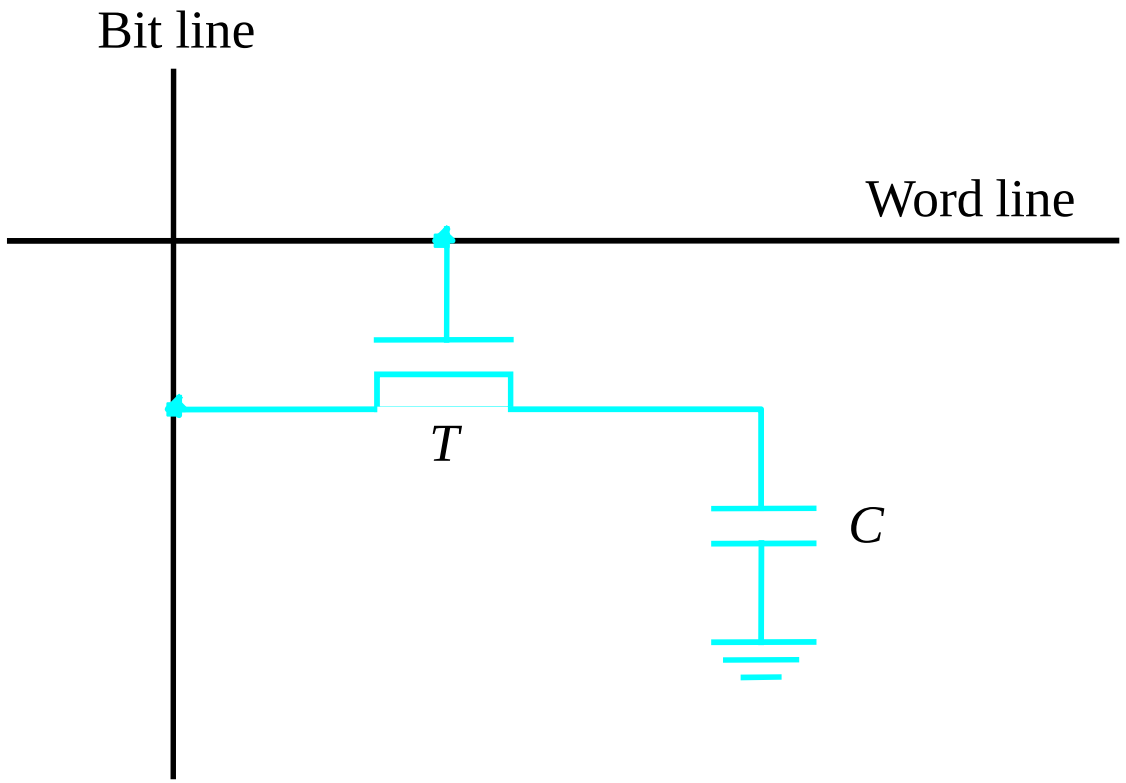


Dynamic RAMs

- Dynamic RAMs (DRAMs):
  - Do not retain their state indefinitely.
  - Contents must be periodically refreshed.
  - read its contents and write it back. refresh an entire row
    - with a read cycle followed immediately by a write cycle.
  - Contents may be refreshed while accessing them for reading.
  - value kept in a cell is stored as a charge in a capacitor
  - use only a single transistor per bit of storage
  - denser and cheaper per bit than SRAM
  - uses significantly less area per bit of memory
  - have larger capacity for the same amount of silicon

Dynamic RAMs

- 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.



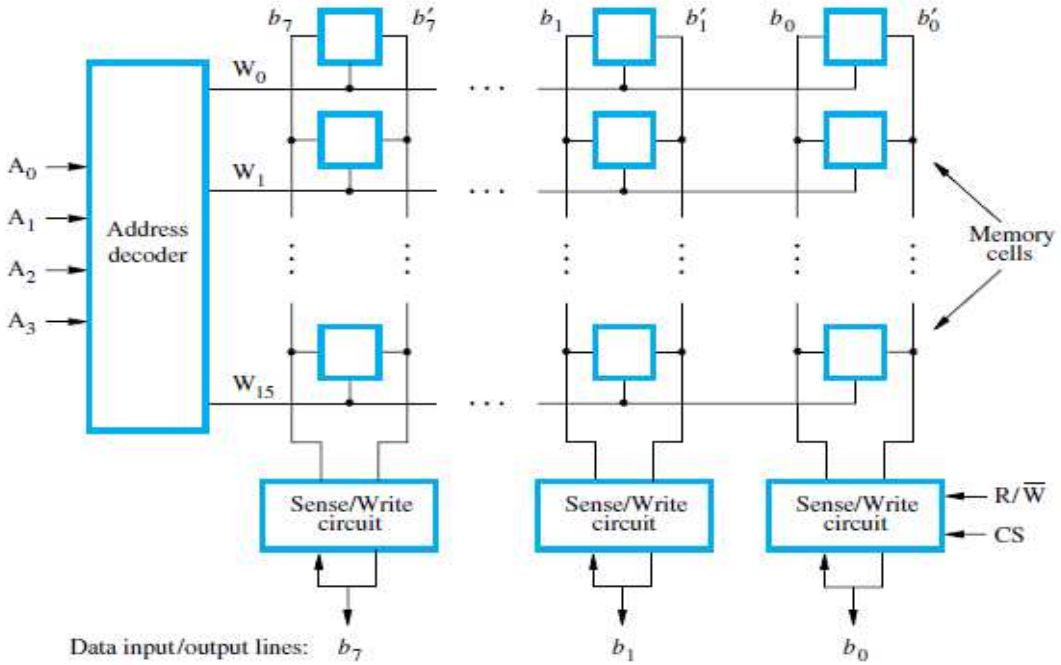
Dynamic RAMs

- organization *inside the DRAM*
  - Instead of just a faster row buffer, the DRAM can be internally organized to read or write from multiple *banks, with each having its own row buffer. Sending an address* to several banks permits them all to read or write simultaneously. For example, with four banks, there is just one access time and then accesses rotate between the four banks to supply four times the bandwidth. This rotating access scheme is called *address interleaving.*
- *dual inline memory modules (DIMMs).* 4–16 DRAMs

Internal Organization of Memory Chips

- Array of cell
- Cell store 1 bit
- 16 x 8

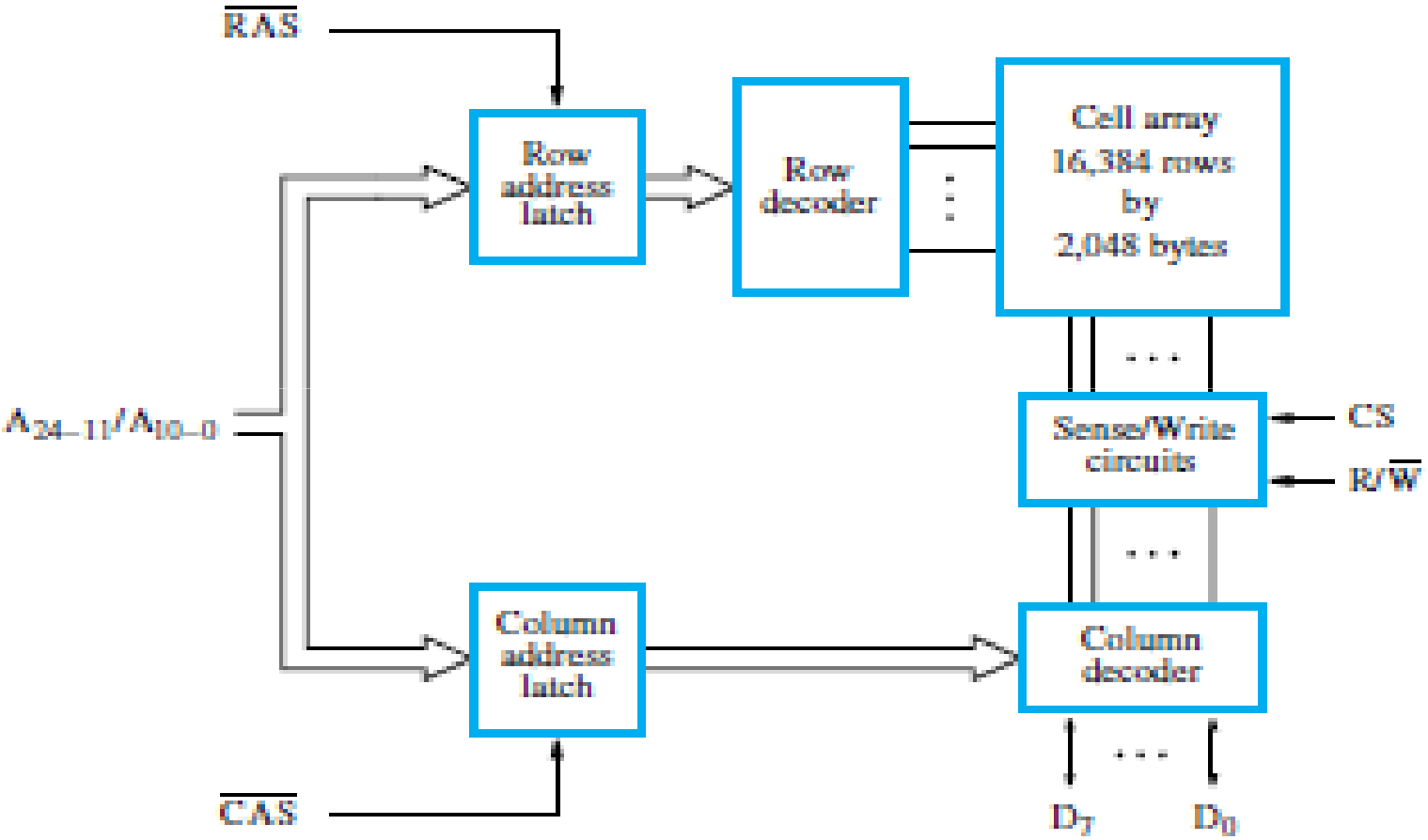
•Each memory cell can hold one bit of information.
•Memory cells are organized in the form of an array.
•One row is one memory word.
•All cells of a row are connected to a common line, known as the “word line”.
•Word line is connected to the address decoder.
•Sense/write circuits are connected to the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



Fast Page Mode

- contents of all 16,384 cells in the selected row are sensed
- only 8 bits are placed on the data lines, D7–0
- 8 bits are selected by column address
- it **possible to access the other bytes in the same row without having to reselect the row, under the control of successive CAS signals**
- Block of data can be transferred at a much faster rate than can be achieved for transfers involving random addresses

Fast Page Mode

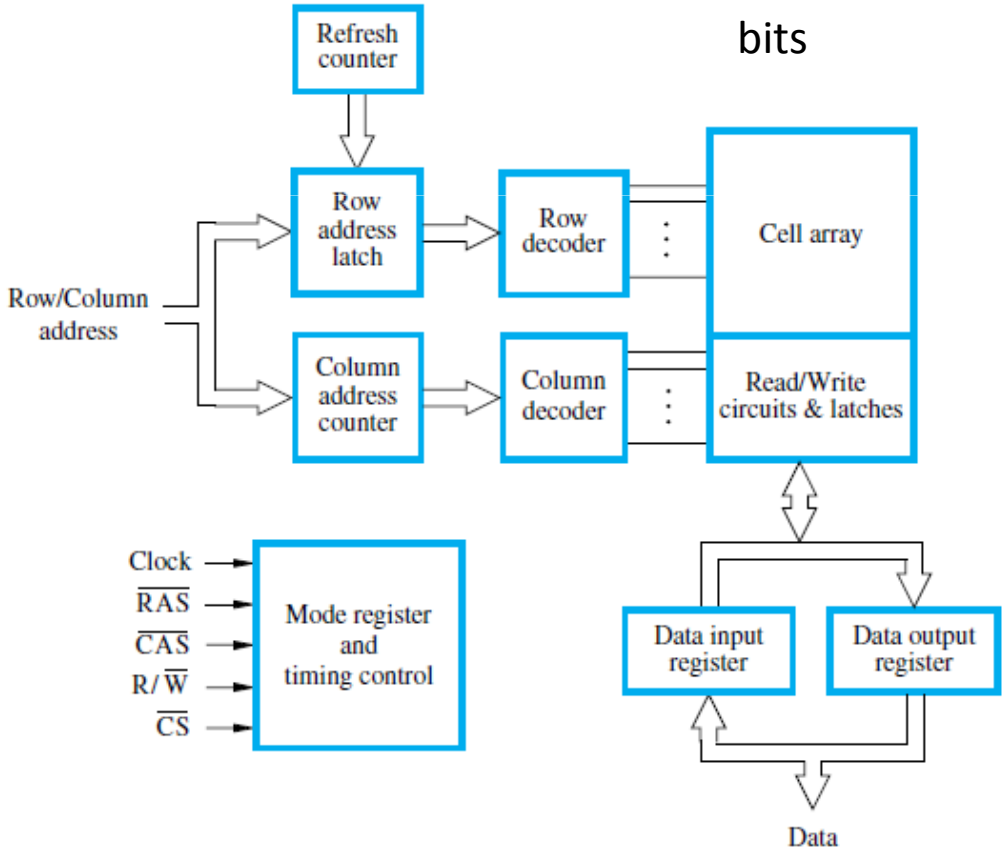


Synchronous DRAMs

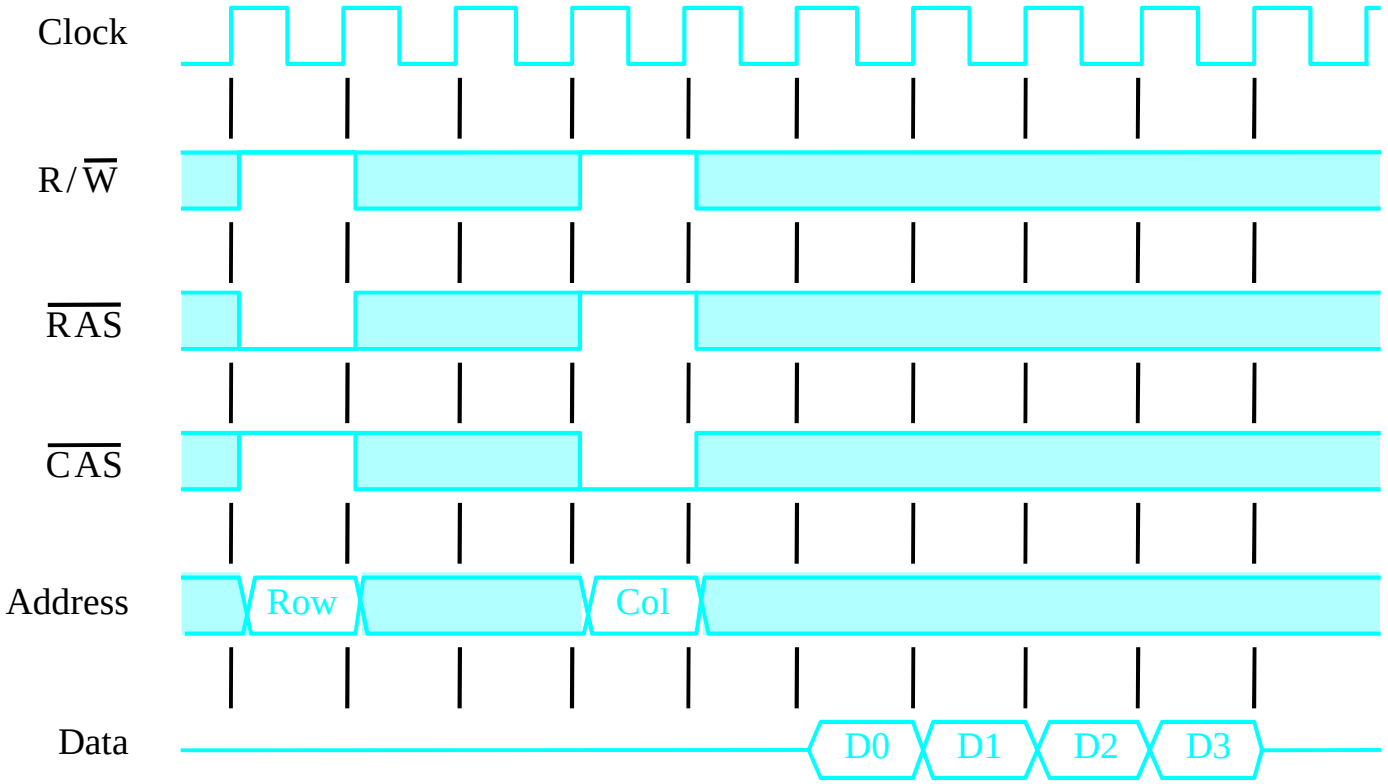
- The operations of SDRAM are controlled by a clock signal.

•DRAMs with added clock - Synchronous DRAMs or SDRAMs

- •the use of a clock eliminates the time for the memory and processor to synchronize.
- •The speed advantage of synchronous DRAMs comes from the ability to transfer the bits in the burst without having to specify additional address bits



Synchronous DRAMs



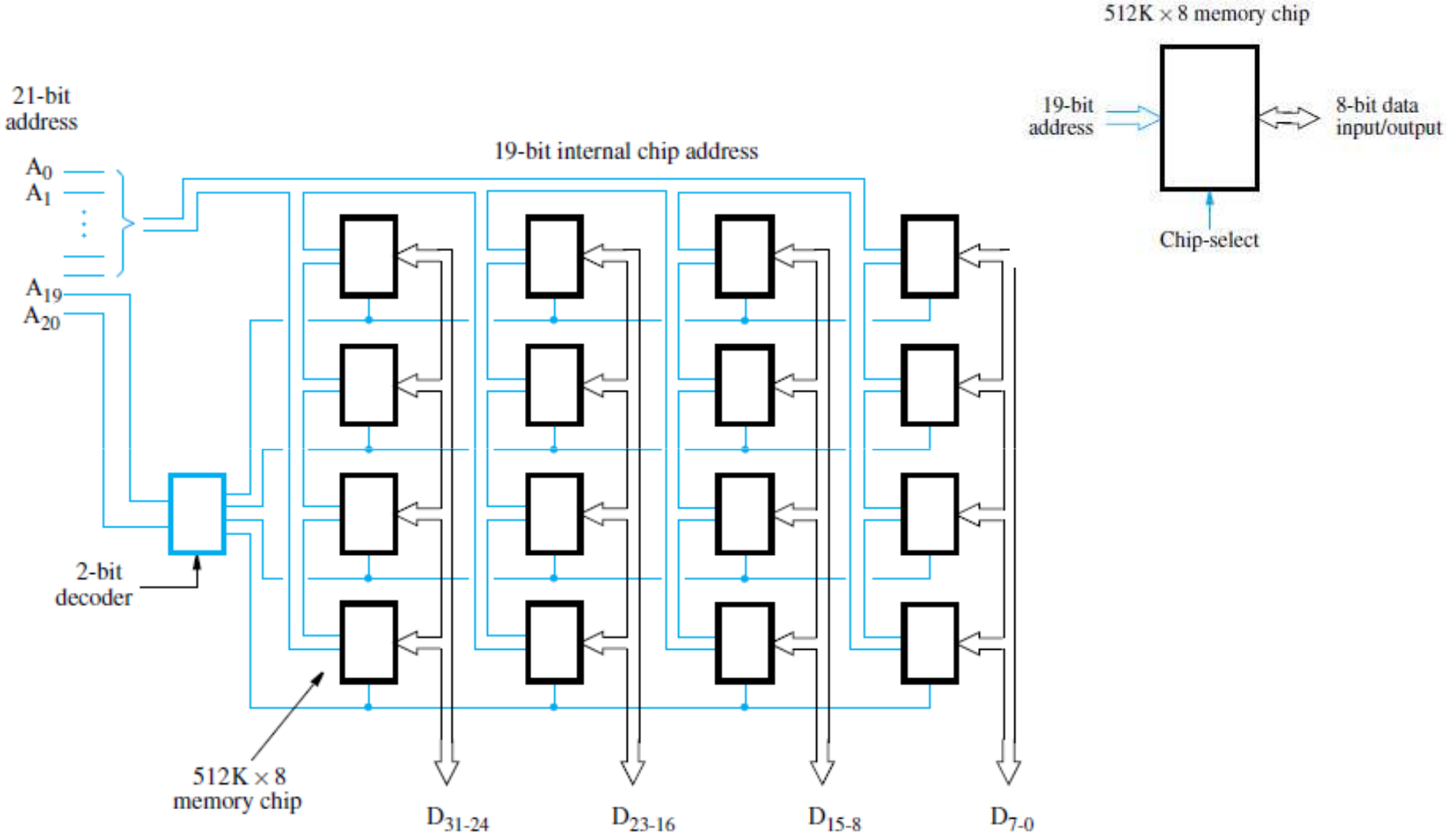
Latency, Bandwidth

- Memory latency is the time it takes to transfer a word of data to or from memory
- Memory bandwidth is the number of bits or bytes that can be transferred in one second

DDR SDRAM

- The fastest version
- Double-Data-Rate 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.**
- *getting twice as much bandwidth*
- 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.

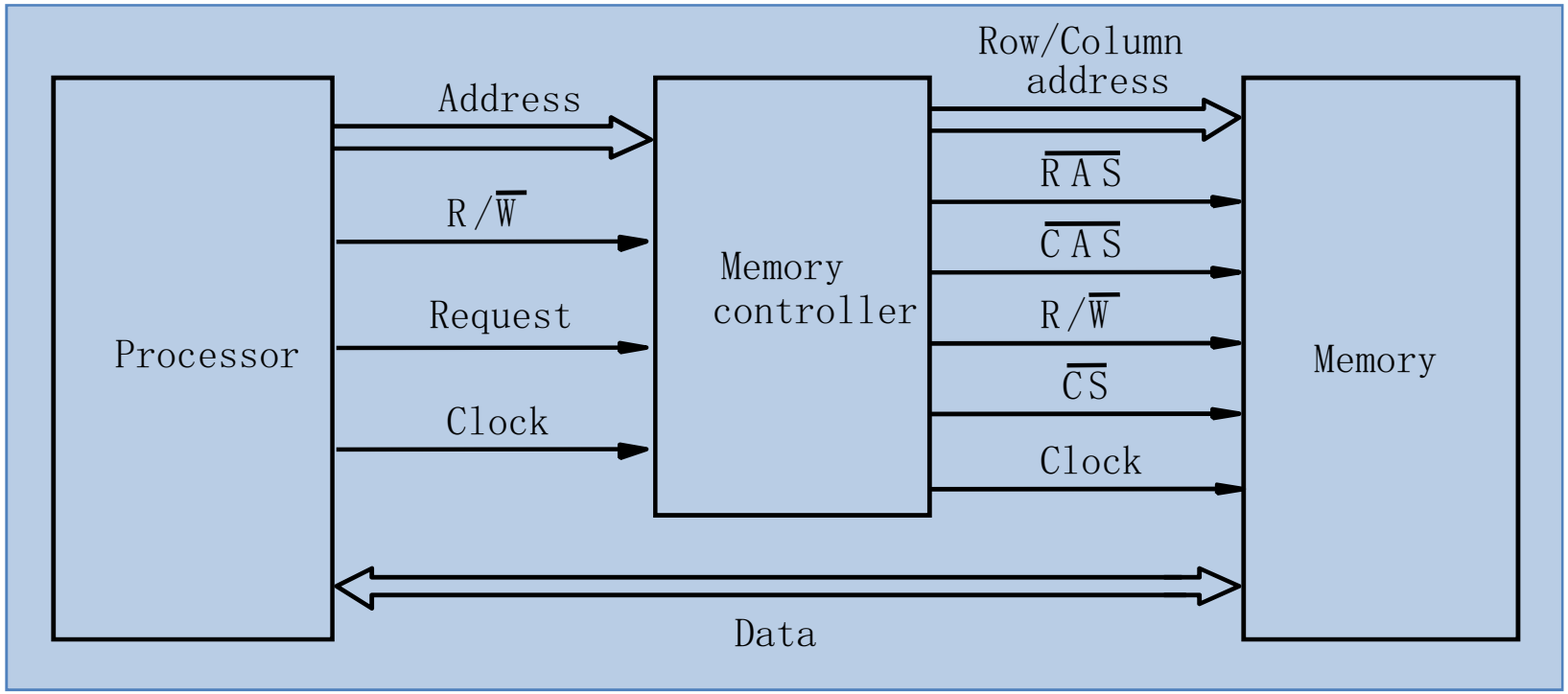
Structures of Larger Memories



Memory controller

- Dynamic memory chip, use multiplexed addresses
- 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
- 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.

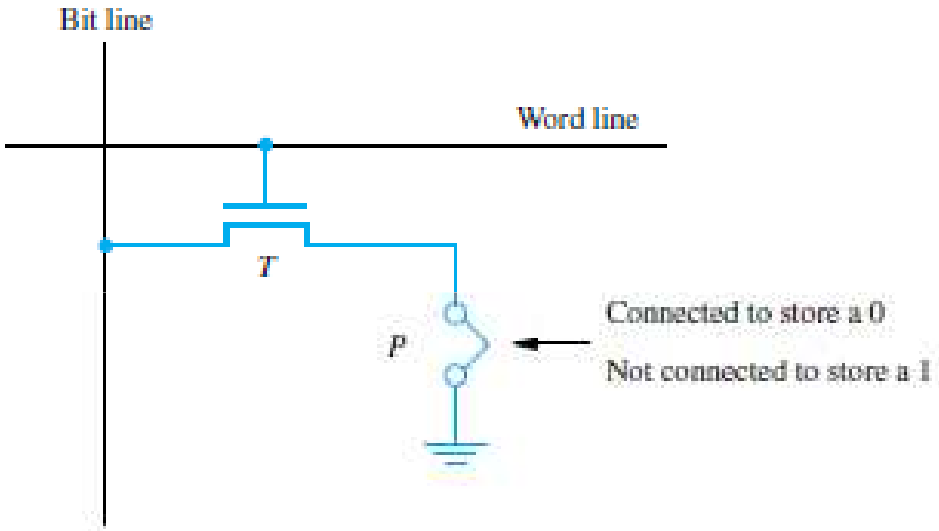
Memory controller



Read-Only Memories (ROMs)

- Non 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.
  - Need to store these instructions so that they will not be lost after the power is turned off.
  - We need to store the instructions into a non-volatile memory.
- Non-volatile memory is read in the same manner as volatile memory.
  - Separate writing process is needed to place information in this memory.
  - Normal operation involves only reading of data, this type of memory is called Read-Only memory (ROM).

Read-Only Memories (ROMs)



Read-Only Memories

- Read-Only Memory:
  - Data are written into a ROM when it is manufactured.
- Programmable Read-Only Memory (PROM):
  - Allow the data to be loaded by a user.
  - Process of inserting the data is irreversible.
  - Storing information specific to a user in a ROM is expensive.
  - Providing programming capability to a user may be better.
- Erasable Programmable Read-Only Memory (EPROM):
  - Stored data to be erased and new data to be loaded.
  - Flexibility, useful during the development phase of digital systems.
  - Erasable, reprogrammable ROM.
  - Erasure requires exposing the ROM to UV light.

Read-Only Memories

- Electrically Erasable Programmable Read-Only Memory (EEPROM):
  - To erase the contents of EPROMs, they have to be exposed to ultraviolet light.
  - Physically removed from the circuit.
  - EEPROMs the contents can be stored and erased electrically.
- Flash memory:
  - Has similar approach to EEPROM.
  - Read the contents of a single cell, but write the contents of an entire block of cells.
  - Flash devices have greater density.
    - Higher capacity and low storage cost per bit.
  - Power consumption of flash memory is very low, making it attractive for use in equipment that is battery-driven.
  - Single flash chips are not sufficiently large, so larger memory modules are implemented using flash cards and flash drives.