CS2002-Computer Architecture and Organization

Memory Subsystem

Contents

- Internal organization of a memory chip
- Organization of a memory unit
- Semiconductor memories: SRAM and DRAM cells.
- Error correction
- Read-Only Memories
- Interleaved Memories
- Cache Memories: Concept, Mapping methods, Caches in commercial processors
- Memory management unit: Concept of virtual memory, Address translation,
- Hardware support for memory management,
- Secondary storage: Hard Disks, RAID, Optical Disks, Magnetic Tape Systems.

Von Neumann Architecture	**Harvard Architecture**
• It is ancient computer architecture based on stored program computer concept.	• It is modern computer architecture based on Harvard Mark I relay based model.
• **Same physical memory address is used for instructions and data**.	• **Separate physical memory address is used for instructions and data**.
• There is **common bus** for data and instruction transfer.	• **Separate buses** are used for transferring data and instruction.
• Two clock cycles are required to execute single instruction.	• An instruction is executed in a single cycle.
• It is cheaper in cost.	• It is costly than Von Neumann Architecture.
• CPU can not access instructions and read/write at the same time.	• CPU can access instructions and read/write at the same time.
• It is used in personal computers and small computers (**microprocessors**).	• It is used in **microcontrollers** and signal processing.

Basic Concepts: Memory

- **Maximum size** of the memory that can be used in any computer is determined by the addressing scheme.
- **For example:** Suppose, a computer that generates **16-bit** addresses
  - ✓It is capable of addressing up to $\mathbf{2^{16} = 64K}$ (kilo) memory locations.
  - ✓Similarly, for **32-bit** address computer: $\mathbf{2^{32} = 4GB}$ locations
  - ✓And, for **64-bit** address computer: $\mathbf{2^{64} = 16E}$ (exa) $\approx \mathbf{16 \times 10^{18}}$ **locations**
- The number of locations represents the size of the address space of the computer.

Basic Concepts: Memory

- Digital computer works on stored programmed concept introduced by **Von Neumann**.
- Memory is used to store the **information**, which includes both **program** and **data**.
- Due to several reasons, we have different kind of memories i.e. **at different level different kind of memory** is used.
- **Memory** of computer is broadly categories into two categories:
  - **Internal:** used by CPU to perform task, and
  - **External:** used to store bulk information, including large software and data.

Basic Concepts: Memory Hierarchy

- Programmers want **unlimited amounts** of memory with **low latency**.
- **Fast memory** technology is **more expensive** per bit than slower memory.
- **Solution:** organize memory system into a **hierarchy.**
  - Entire addressable memory space available in largest, slowest memory.
  - Incrementally smaller and faster memories, each containing a subset of the memory below it, proceed in steps up toward the processor.
- The purpose of memory hierarchy is:
  - To **bridge the speed mismatch** between processor and memory at **reasonable cost**.
  - **Minimize the average access time** of entire memory systems

Basic Concepts: Memory Hierarchy

- **Temporal** and **spatial locality** insures that nearly all references can be found in smaller memories.
  - i.e. Gives the allusion of a large, fast memory being presented to the processor.

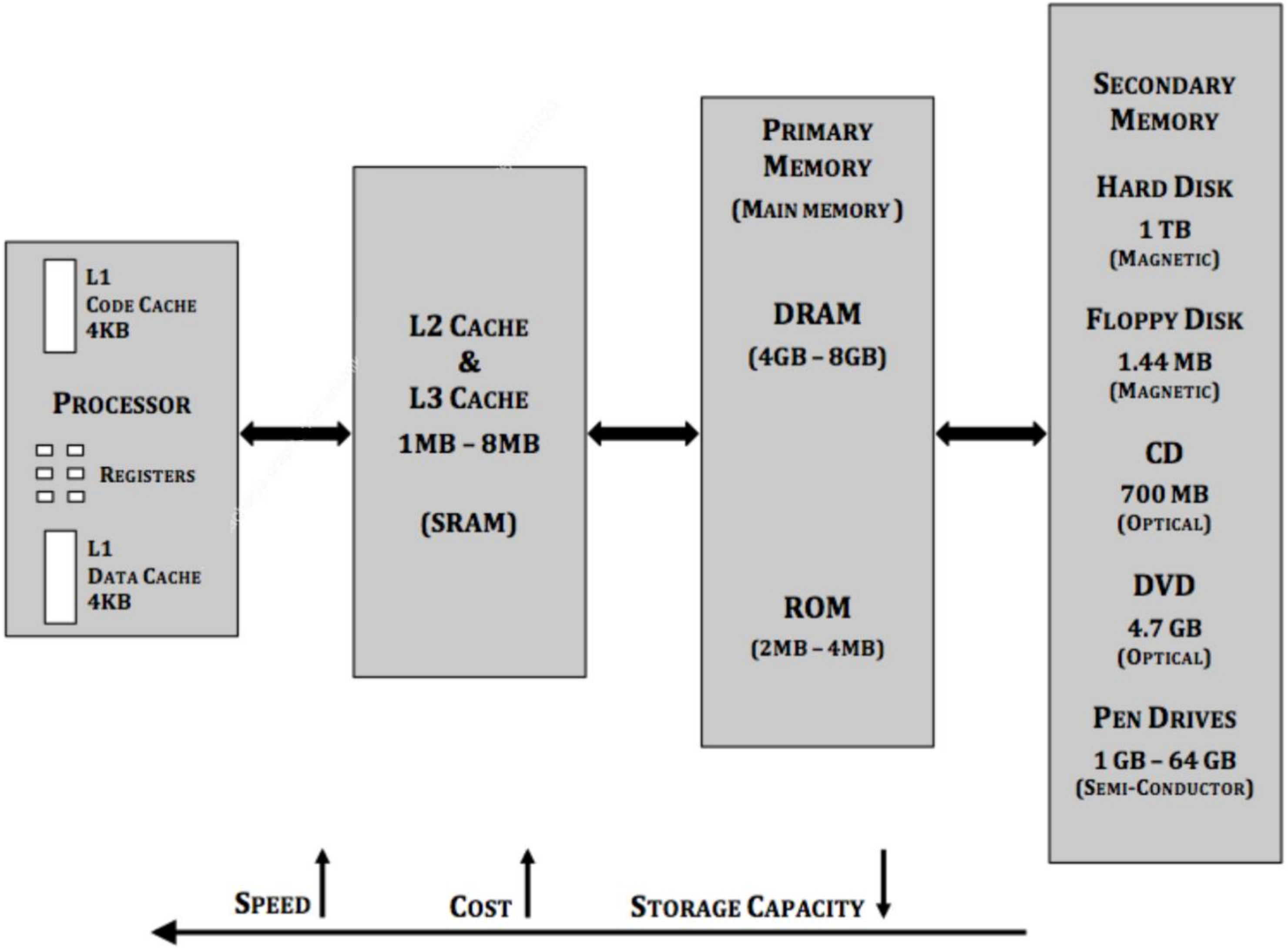
- **Locality**:
  - **Spatial Locality:** Data is more likely to be accessed if neighboring data is accessed.

    e.g., data in a sequentially access array

  - **Temporal Locality:** Data is more likely to be accessed if it has been recently accessed.

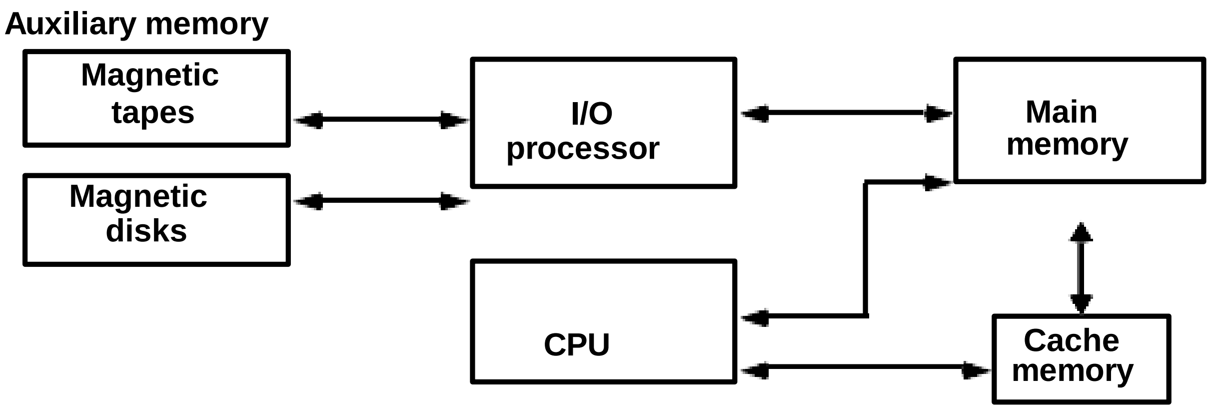
    e.g. code within a loop

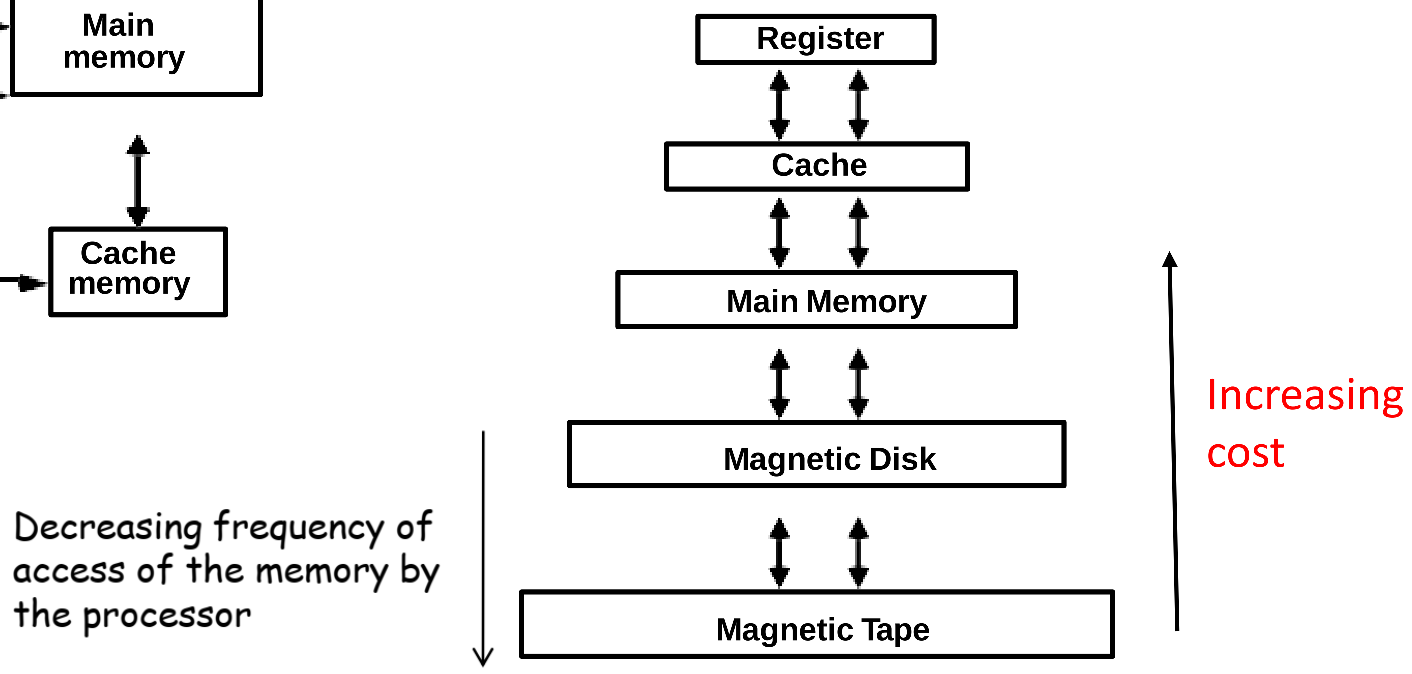
Basic Concepts: Memory Hierarchy



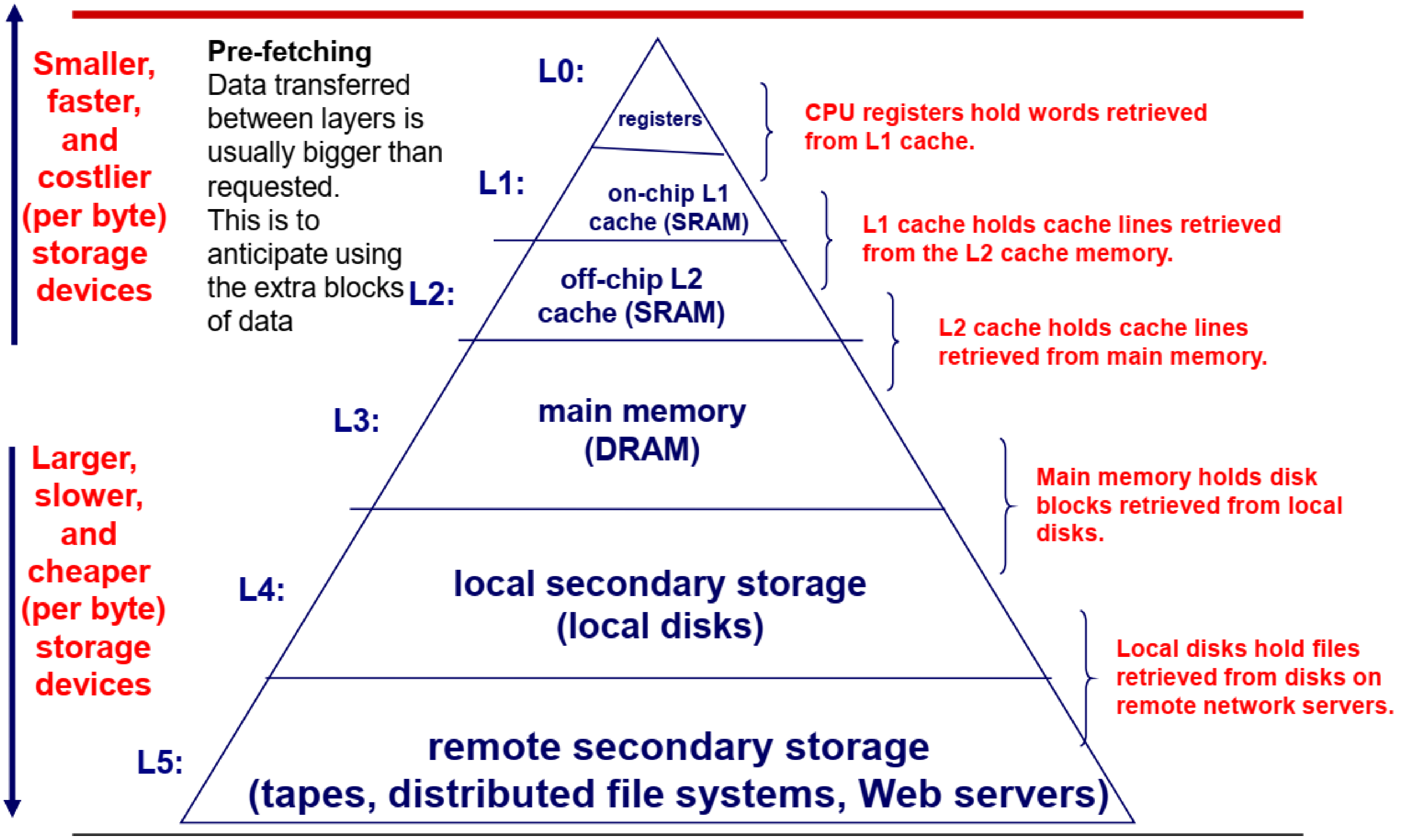
Basic Concepts: Memory Hierarchy

- Memory Hierarchy is to obtain the **highest possible access speed while minimizing the total cost** of the memory system
- Speed of memory access is critical, the idea is to **bring instructions and data** that will be used in the near future **as close to the processor as possible**.





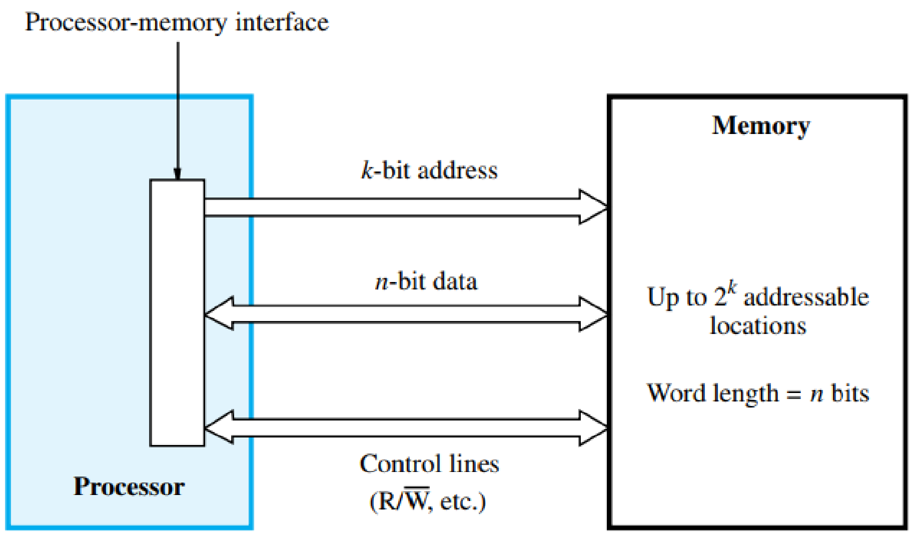
Basic Concepts: Memory Hierarchy Example



Main Memory

- The connection between the processor and its memory consists of **address**, **data**, and **control lines**.

- **Address lines** to specify the **memory location** involved in a data transfer operation
- **Data lines** to transfer the data.
- At the same time, **control lines** carry the command indicating a Read or a Write operation and whether a byte or a word is to be transferred.
- **Control lines** also provide the necessary **timing information** and are used by the memory to indicate when it has completed the requested operation.



How to measure Speed of memory units?

- ***Memory access time:***
  - It is the **time** that elapses between the **initiation of an operation** to transfer a word of data and the **completion of that operation**.
  - This is referred to as the ***memory access time***.

- ***Memory cycle time***
  - It is the **minimum time delay** required between **the initiation of two successive memory operations**.
    - For example, the time between two successive Read operations.

- The cycle time is usually slightly longer than the access time, depending on the implementation details of the memory unit.

Random-Access Memory **(RAM)**

- A memory unit is called a ***Random-access Memory*** (RAM) **if the access time to any location is the same**, independent of the location's address.
- This distinguishes such memory units from **serial, or partly serial**, access storage devices such as **magnetic and optical disks**.
  - Access time of the these devices depends on the address or position of the data.
- The technology for implementing computer memories uses **semiconductor integrated circuits**.

Cache Memory

- In general, **a processor can process instructions and data faster than they can be fetched from the main memory**.
- Hence, the **memory access time is the bottleneck in the system**.
    - **Solution**: use a ***cache memory.***
- ***Cache memory:***
    - a **small, fast memory** inserted between the larger, slower **main memory** and the **processor**.
    - It holds the currently active portions of a program and their data.

Reading Assignment: **What is Virtual Memory**

Virtual Memory

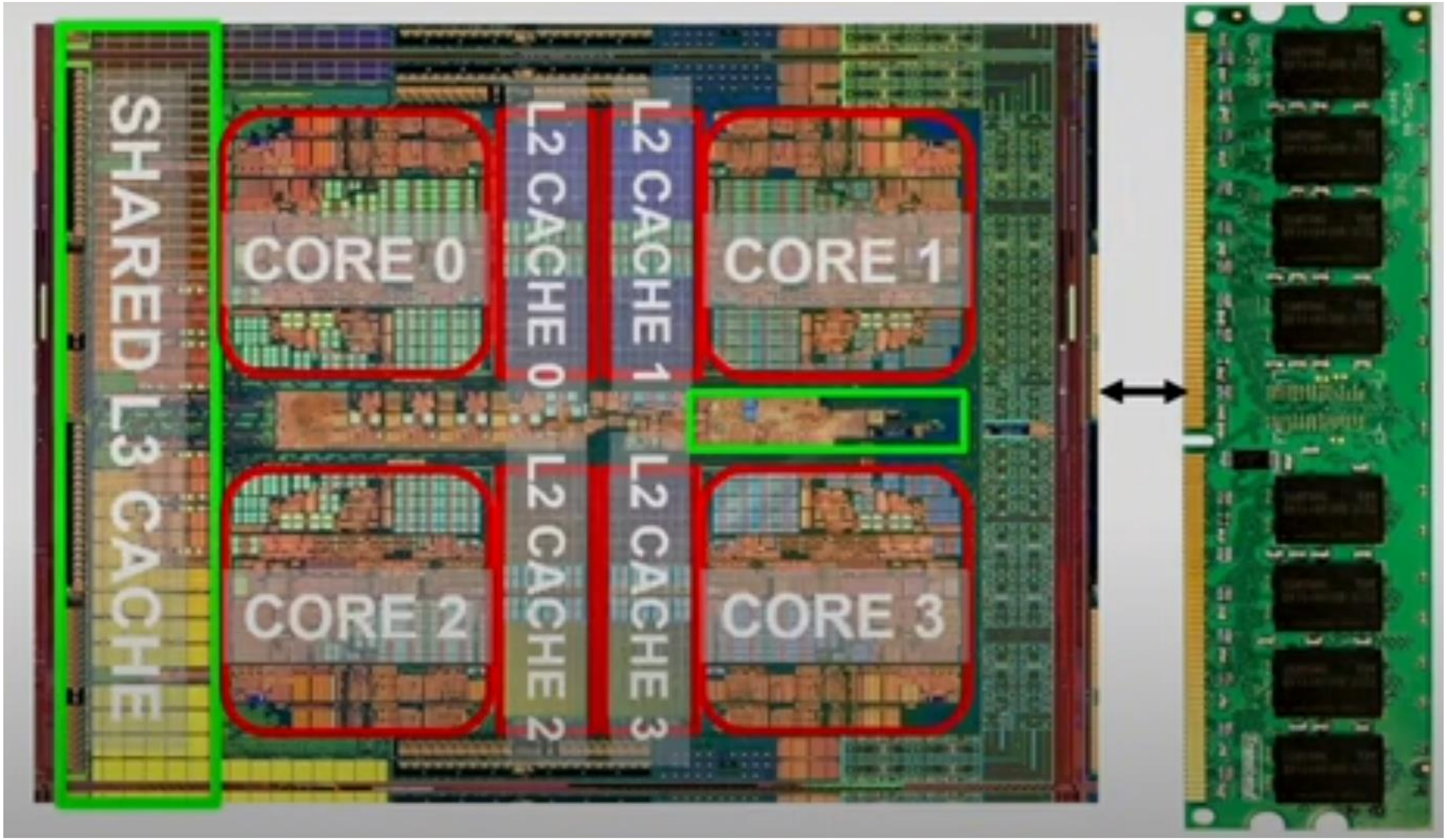
- *Virtual memory* is another important concept related to memory organization.
- **With this technique:**
  - only the **active portions of a program** are stored in the **main memory**, and **remainder** is stored on the much larger **secondary** storage device.
  - Sections of the program are transferred back and forth between the main memory and the secondary storage device in a manner that is transparent to the application program.
  - As a result, the application program sees a memory that is much larger than the computer's physical main memory.

Block Transfers

- As we know, **data move frequently** between the **main memory and the cache** and between the **main memory and the disk**.
- These transfers **do not occur one word at a time**.
- Data are always transferred in contiguous blocks involving tens, hundreds, or thousands of words.
  - Data transfers between the main memory and high-speed devices such as a graphic display or an Ethernet interface also involve large blocks of data.
- Hence, a critical parameter for the **performance of the main memory is its ability to read or write blocks of data at high speed**.
- This is an important consideration that we will encounter repeatedly as we discuss memory technology and the organization of the memory system.

Semiconductor RAM Memories

Semiconductor random-access memories (RAMs) are available in a **wide range of speeds**. Their cycle times range from **100 ns** to less than **10 ns**.


SHARED L3 CACHE
CORE 0
L2 CACHE 0
L2 CACHE 1
CORE 1
CORE 2
L2 CACHE 2
L2 CACHE 3
CORE 3

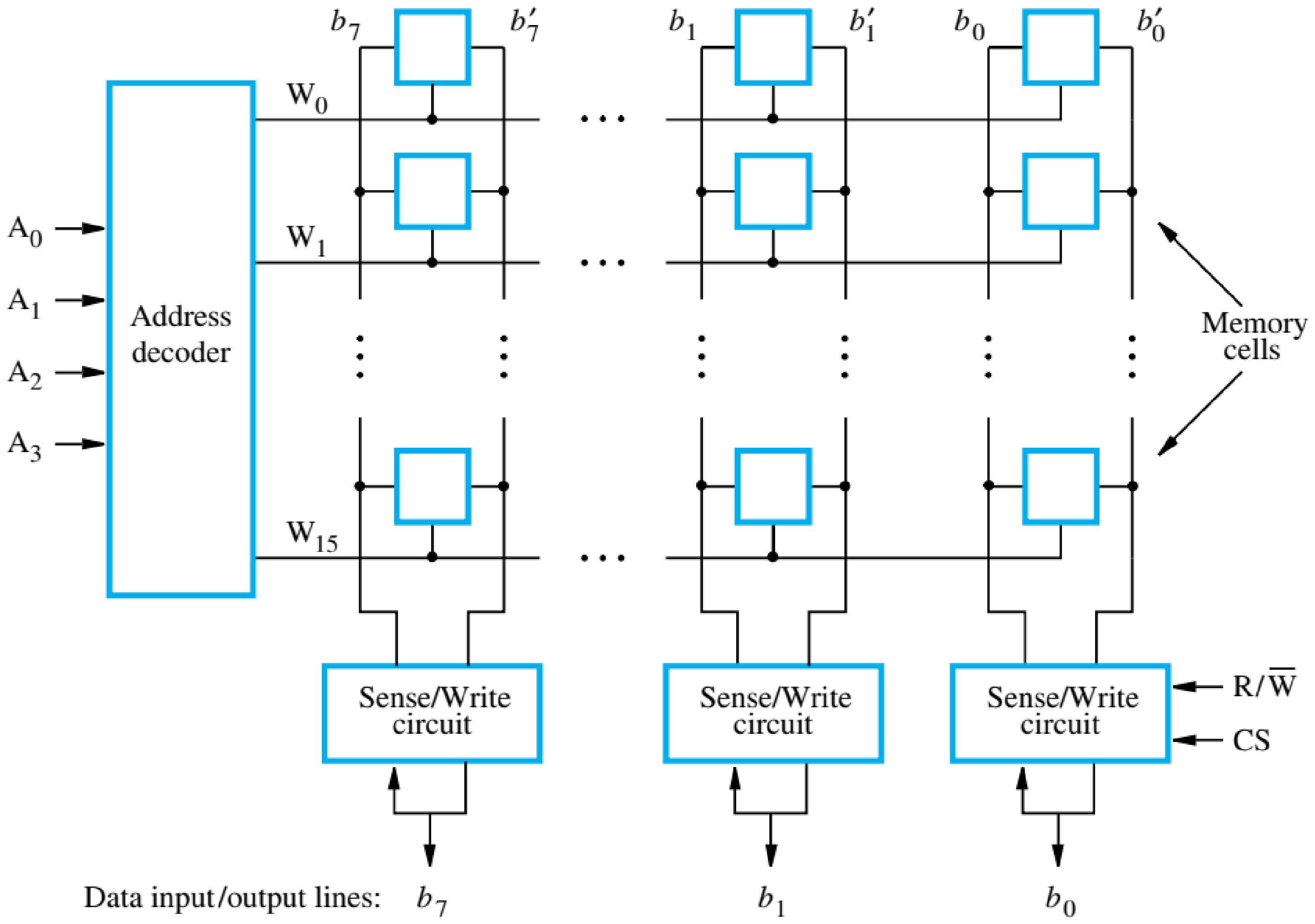
Internal Organization of Memory Chips

- **Memory cells** are usually organized in the form of an **array**, in which **each cell is capable of storing one bit** of information.
- **Each row** of cells constitutes **a memory word** and all **cells of a row** are connected to a common line referred to as the **word line**, which is driven by the address decoder on the chip.
- The **cells in each column** are connected to a **Sense/Write circuit by two bit lines**, and the Sense/Write circuits are connected to the **data input/output lines** of the chip.
- During a **Read operation**, these **circuits sense, or read**, the information stored in the cells selected by a word line and place this information on the output data lines.
- During a **Write operation**, the Sense/Write circuits receive input data and store them in the cells of the selected word.

Internal Organization of Memory Chips

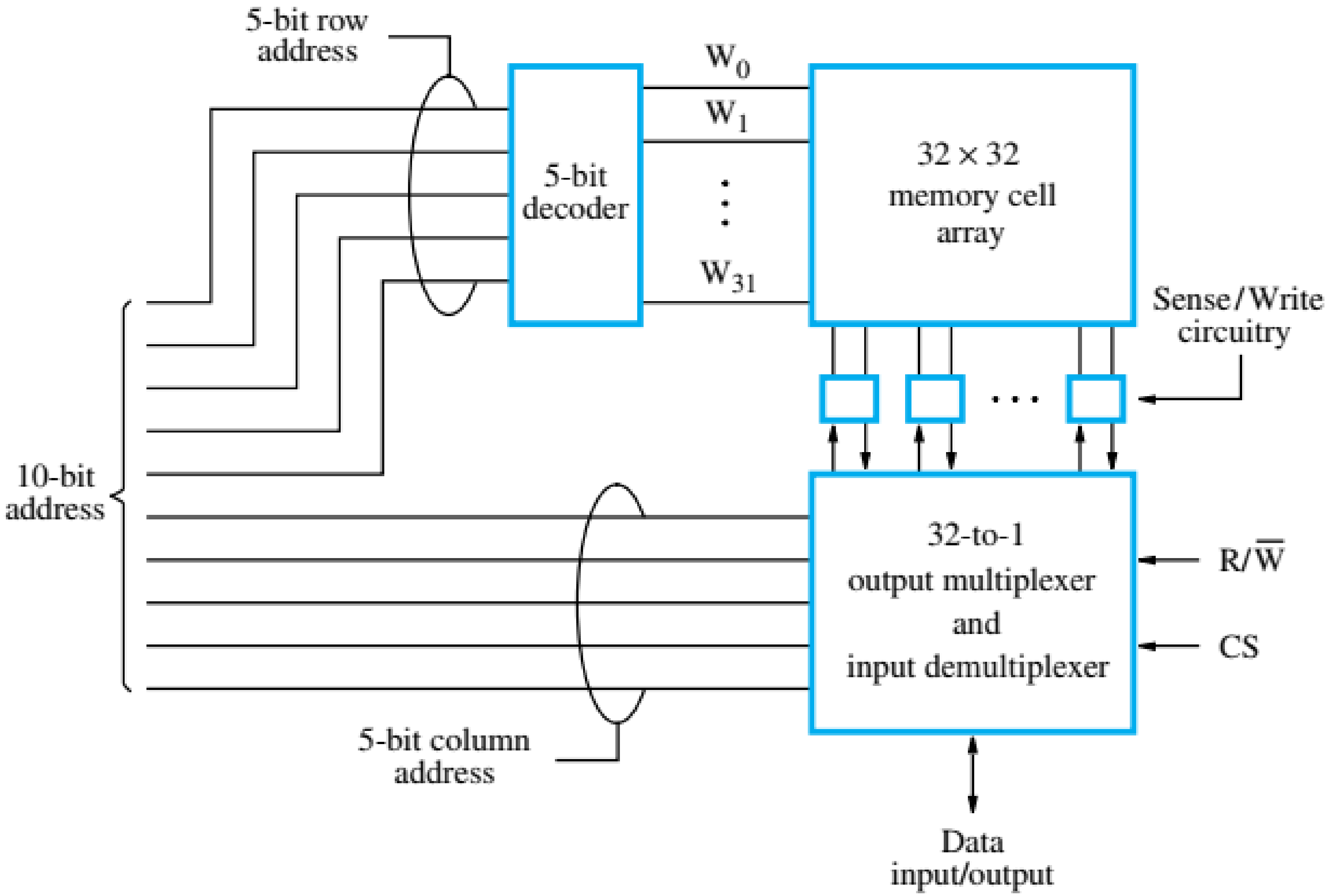
- **An example** of a very small memory circuit consisting of **16 words of 8 bits** each, referred to as a **16 × 8 organization**.

- Data input and the data output of each Sense/Write circuit are connected to a single bidirectional data line that can be connected to the data lines of a computer.
- Two control lines, R/W and CS, are provided.
- The **R/W (Read/Write)** input specifies the required operation, and the **CS (Chip Select)** input selects a given chip in a multichip memory system.



Internal Organization of Memory Chips

Organization of a
1K × 1 memory chip.



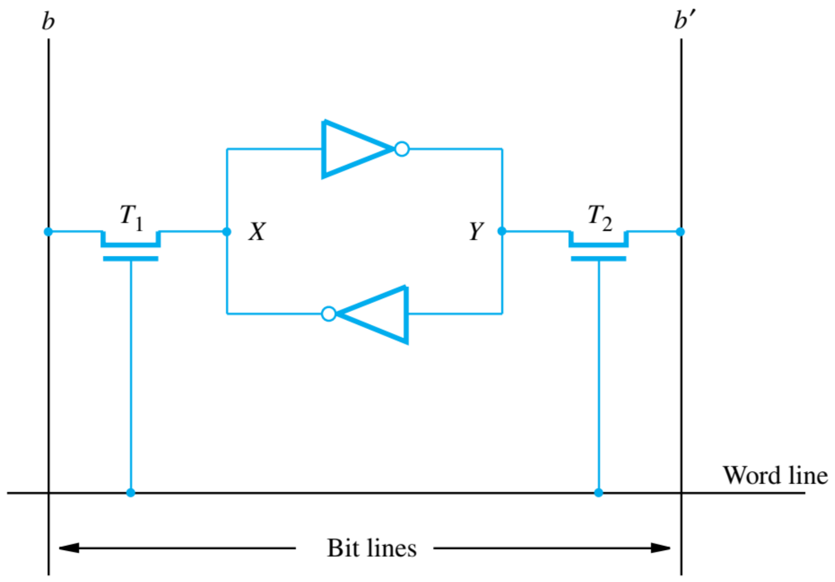
Reading Assignments

- What is CMOS (P-type and N-type transistors)? Working Principle & Its Applications.
- What is field effect transistor (FET)? Working Principle & Its Applications.

Memory Cell: SRAM

Memories that consist of circuits capable of retaining their state as long as power is applied are known as ***static memories***.

- Two **inverters** are cross-connected to form a latch and it is connected to two bit lines by transistors T_1 and T_2.
- These transistors act as switches that can be opened or closed under control of the word line.
- When the word line is at ground level, the transistors are turned off and the latch retains its state.
- **For example,**
    - if the logic value at point X is 1 and at point Y is 0, this state is maintained as long as the signal on the word line is at ground level.
    - Assume that this state represents the **value 1**.



Memory Cell: SRAM

Read Operation:

- In order to **read the state of the SRAM cell**, the word line is activated to close switches $\boldsymbol{T_1}$ and $\boldsymbol{T_2}$.
- If the **cell is in state 1**, the **signal on bit line *b* is high** and the signal **on bit line *b*' is low.**
- The **opposite is true if the cell is in state 0**.
- Thus, ***b*** and ***b'*** are always complements of each other.
- The Sense/Write circuit at the end of the two bit lines monitors their state and sets the corresponding output accordingly.

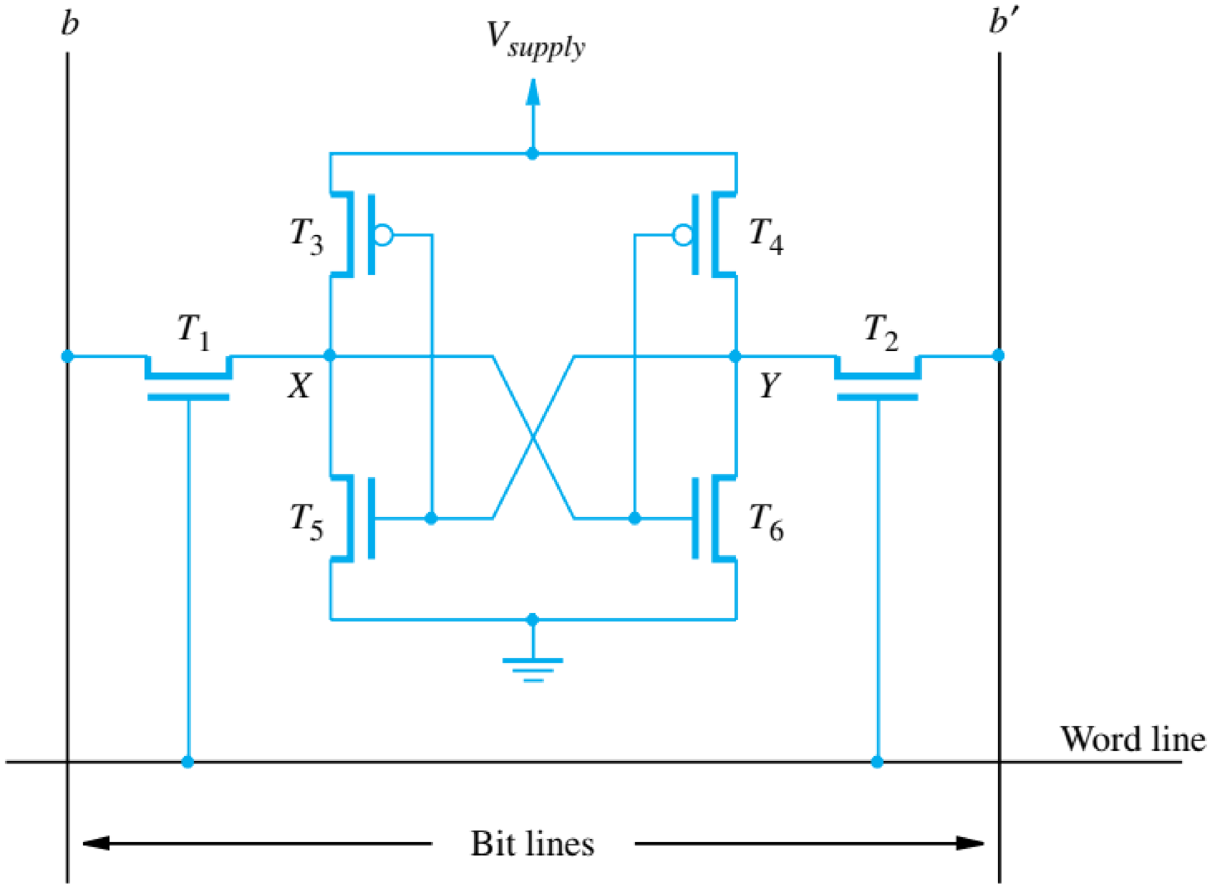
Memory Cell: SRAM

Write Operation:

- During a **Write operation**, the Sense/Write circuit drives **bit lines b** and **b'** instead of sensing their state.
- It places the appropriate value on **bit line b and its complement on b'** and activates the word line.
- This forces the cell into the corresponding state, which the cell retains when the word line is deactivated.

Memory Cell: SRAM CMOS Cell

- Transistor pairs (***T*₃, *T*₅**) and (***T*₄, *T*₆**) form the inverters in the latch.
- For example, **in state 1** the **voltage at point X is maintained high** by having transistors:
  - **T_3** and **T_6 ON**, while **T_4** and **T_5** are **OFF**.
  - If **T_1** and **T_2** are turned **ON**, bit lines **b** and **b'** will have high and low signals, respectively.



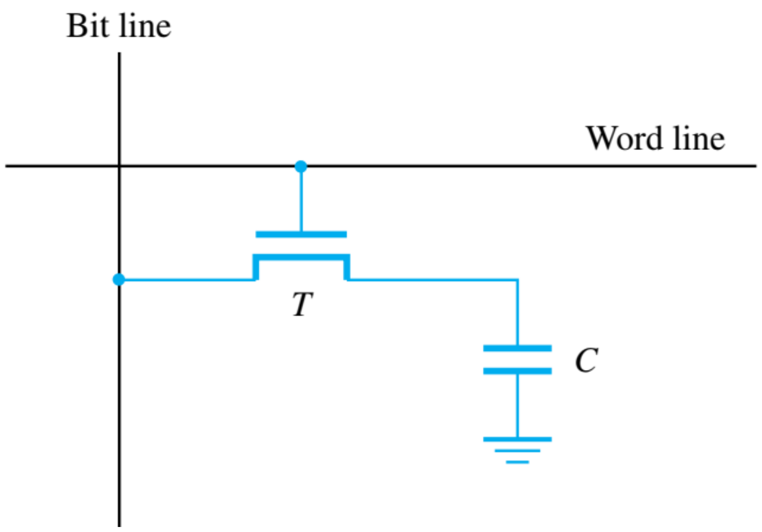
Memory Cell: SRAM CMOS Cell

- **Continuous power is needed for the cell to retain its state**.
- If power is interrupted, the cell's contents are lost.
- When power is restored, the latch settles into a stable state, but not necessarily the same state the cell was in before the interruption.
- Hence, SRAMs are said to be **volatile memories** because their contents are lost when power is interrupted.
- A major **advantage of CMOS SRAMs** is their **very low power consumption**, because current flows in the cell only when the cell is being accessed.
- Otherwise, $\boldsymbol{T_1}$, $\boldsymbol{T_2}$, and **one transistor in each inverter are turned off**, ensuring that there is no continuous electrical path between $\boldsymbol{V_{supply}}$ and **ground**.
- Static **RAMs can be accessed very quickly**. Access times on the order of a few nanoseconds are found in commercially available chips.
- **SRAMs are used in applications where speed is of critical concern**.

Memory Cell: DRAM

- Static RAMs are fast, but their cells require several transistors → **Expensive**
- **Less expensive** and higher density RAMs can be implemented with simpler cells.
- But, these **simpler cells do not retain their state for a long period**, unless they are accessed frequently for Read or Write operations.
- Memories that use such cells are called ***dynamic RAMs*** **(DRAMs)**.

- A DRAM **memory cell** consists of a single **field effect transistor (FET)** and a **capacitor**.
- **Bit stored in a cell in the form of a charge on a capacitor**.
- To store a **bit** in this cell, transistor ***T*** is turned **ON** and an appropriate voltage is applied to the bit line.
- This causes a known amount of charge to be stored in the capacitor.



Memory Cell: DRAM

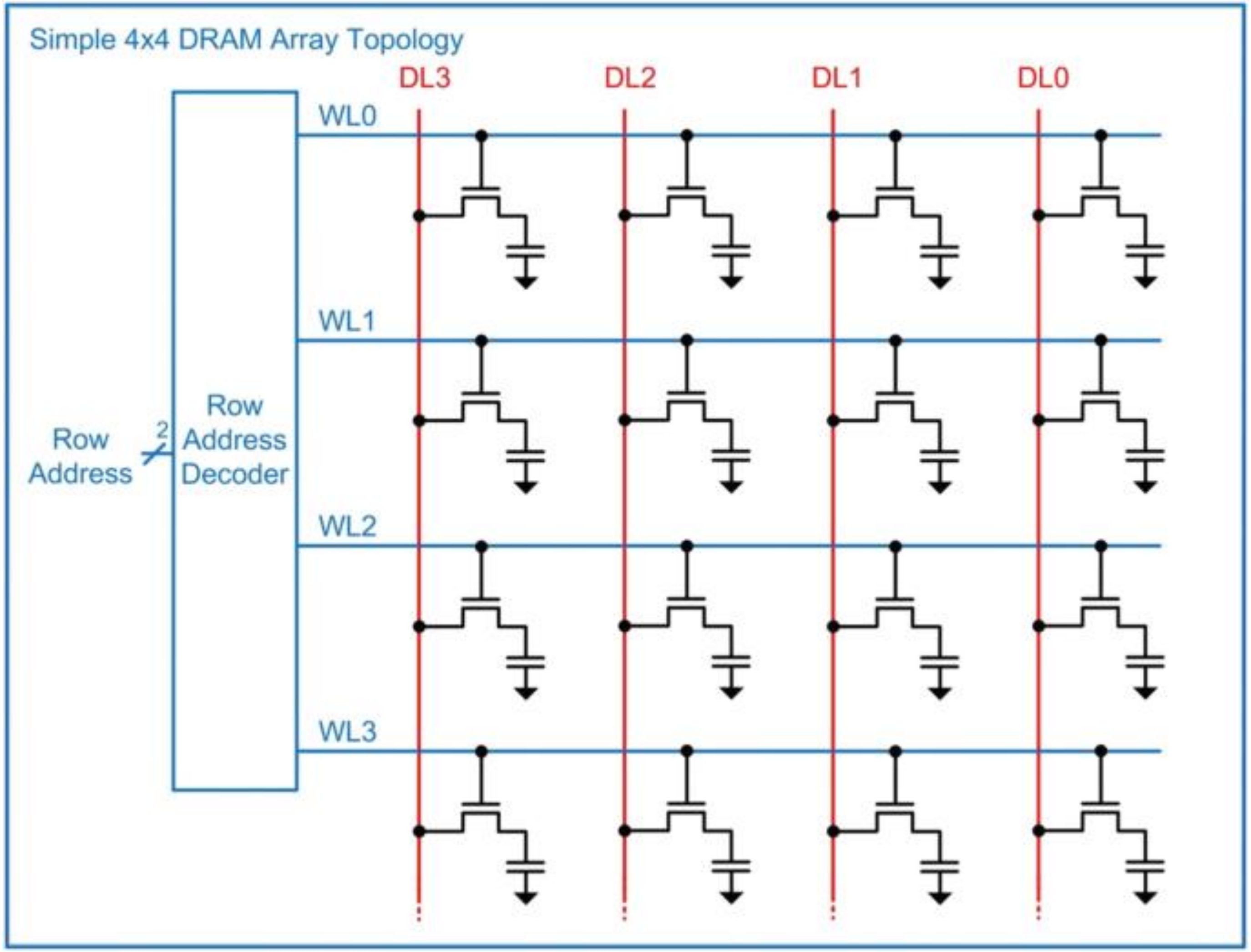
- After **the transistor is turned off**, the charge remains stored in the capacitor, but not for long.
- The capacitor begins to discharge. charge can be maintained for only tens of milliseconds.
- This is because the transistor continues to conduct a tiny amount of current, measured in **picoamperes**, after it is turned off.
- Cell is required to store data for longer time, its content must be periodically ***refreshed*** by restoring **capacitor charge to its full value**.

Memory Cell: DRAM

- Information stored in the cell can be retrieved correctly only if it is read before the charge in the capacitor drops below some threshold value.
- A sense amplifier connected to the bit line detects whether the **charge stored in the capacitor** is **above or below the threshold value**.
- If the **charge is above the threshold**:
  - Sense amplifier drives the bit line to **the full voltage representing the logic value 1**.
  - As a result, the capacitor is recharged to the full charge corresponding to the **logic value 1**.
- If the **charge in the capacitor is below the threshold value**:
  - It pulls the bit line to ground level to discharge the capacitor fully (**logic value 0**).
  - Thus, reading the contents of a cell automatically refreshes its contents.

Since the word line is common to all cells in a row, all cells in a selected row are read and refreshed at the same time.

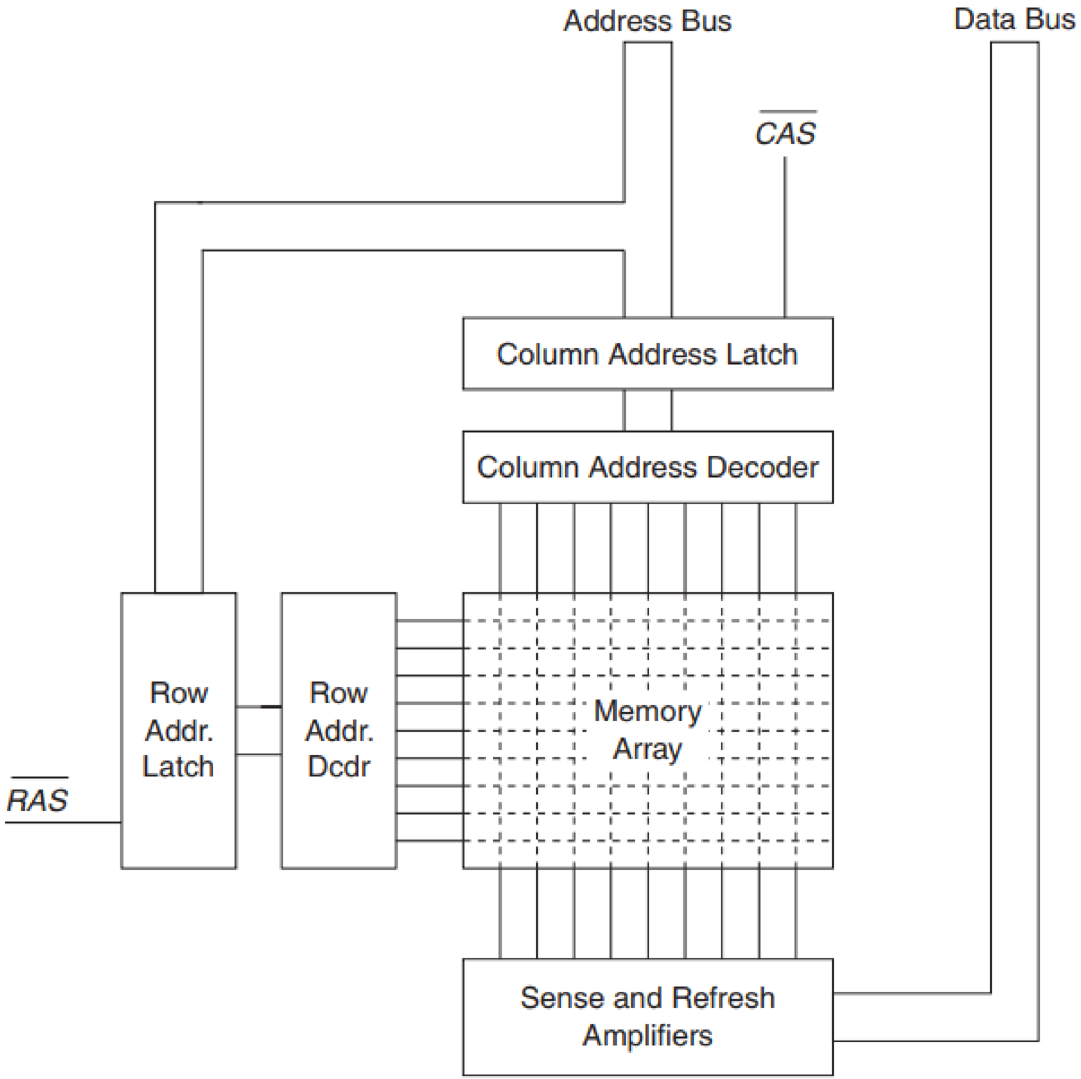
Memory Cell: DRAM



Internal diagram of a DRAM chip

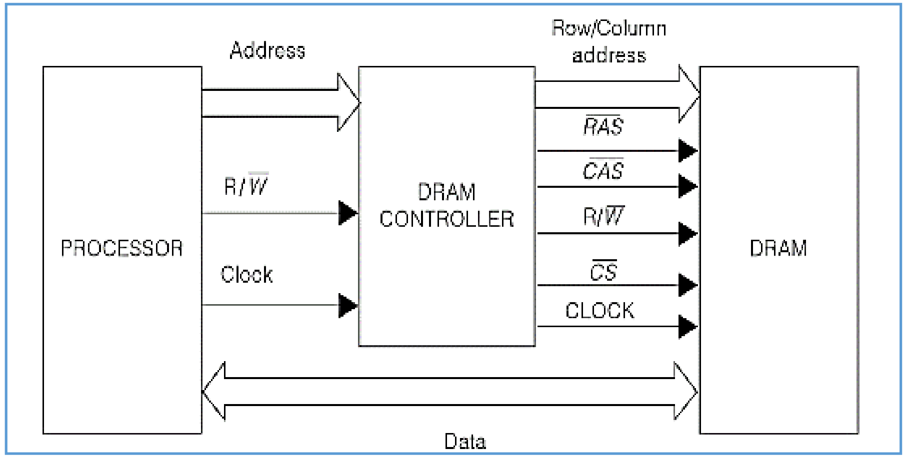
Each cell has a unique location or address defined by intersection of row and a column.

i) The row address is placed on the rows and given sufficient time to stabilize and be latched.
ii) The row address strobe $\overline{RAS}$ signal is then activated.
iii) The row address decoder selects the proper row.
iv) Next, the column address is placed on the same address lines and allowed to stabilize and be latched.
v) The column address strobe $\overline{CAS}$ signal is then activated.
vi) The $\overline{CAS}$ pin also serves as the output enable, so once the $\overline{CAS}$ signal has been stabilized, the sense amps place the data from the selected row and column, on the data bus.
vii) With this, the data in the selected address is available at the output buffers of the chip, and it is transferred to the data bus.



Memory Control of DRAM

- **Word line** and **bit line** are connected as shown to select the required bit within memory to be read or written to.
- Multitudes of Such cells form word consisting of bits.
- Memory addresses are decoded and converted as rows and columns of matrix that memory elements are arranged in.
- Processor when address memory sends the complete address on its address pins.
- Between **processor and DRAM** chip there is a **memory controller** whose function is to split the address into two as columns and rows.
- The **memory controller** should also generate the signals necessary for reading or writing to DRAM.

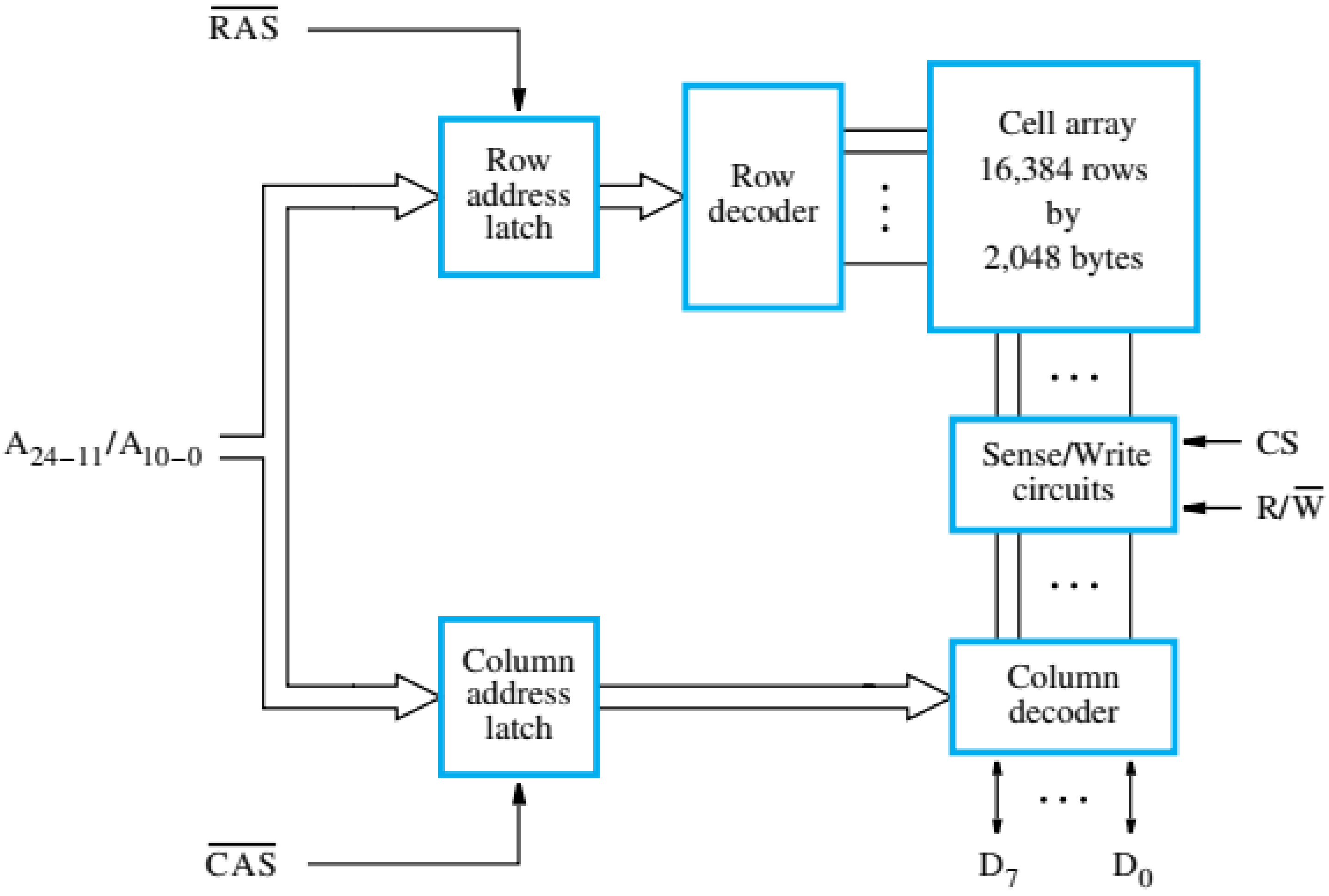


DRAM: 256-Megabit chip

Internal organization of a **32M × 8** DRAM chip.

In commercial DRAM chips, the **RAS** and **CAS** control signals are **active when low**.

RAS: row address strobe
CAS: column address strobe



DRAM: 256-Megabit chip

- **256-Megabit DRAM** chip, configured as **32M×8**, is shown in the above Figure
- The cells are organized in the form of a **16K×16K** array.
- The **16,384** cells in each row are divided into **2,048** groups of 8, forming 2,048 bytes of data.
- Therefore, **14 address bits are needed to select a row**, and another **11 bits are needed to specify a group of 8 bits** in the selected row.
- In **total, a 25-bit** address is needed to access a byte in this memory.
- The high-order 14 bits and the low-order 11 bits of the address constitute the row and column addresses of a byte, respectively.

DRAM: Refreshing

- The refresh rate of DRAM depends on the temperature and the DRAM standard:
  - DDR5 and LPDDR5: refresh period of **32 milliseconds at 85°C**.
- **JEDEC standard**:
  - **refresh rate of 64 milliseconds** at normal temperatures (<85°C)
  - **32 milliseconds** at high temperatures (>85°C)

How is refreshing done?

- There are many methods for refresh and one commonly used method is **ROR (RAS only Refresh). By activating each row using RAS.**

DRAM: Refreshing

- **DRAM controller** takes care of scheduling the refreshes and making sure that
  they do not interfere with regular reads and writes.
- So to keep the data in DRAM chip from leaking away, the **DRAM controller periodically sweeps** through **all of the rows by cycling repeatedly** and placing a series of row addresses on the address bus.
- This method is designated as **ROR** or **RAS Only Refresh**.
- To **reduce the number of refresh cycles**, one method of design is to split the address such that there are fewer rows and more columns.
- So, the DRAM array is then a **rectangular array**, rather than **a square one**.

Synchronous DRAM

- In **Asynchronous DRAM**, access timing is **not related to the system clock** at all.
- In **Synchronous DRAM**, access are **synchronized with system clock** and SDRAM is currently the RAM that is used as primary memory in general purpose computer systems.
- **Synchronization with system clock easier control of the memory access operations.**
- **SDRAMs have built-in refresh circuitry**, with a refresh counter to provide the addresses of the rows to be selected for refreshing.
- As a result, the dynamic nature of these memory chips is almost invisible to the user.

Synchronous Vs Asynchronous DRAM

Asynchronous

- Does **not share any common clock with CPU**, the controller chips have to manipulate the DRAM's control pins based on all sorts of timing considerations.
- For Accessing Memory, toggling of the external control inputs has a direct effect on internal memory array.

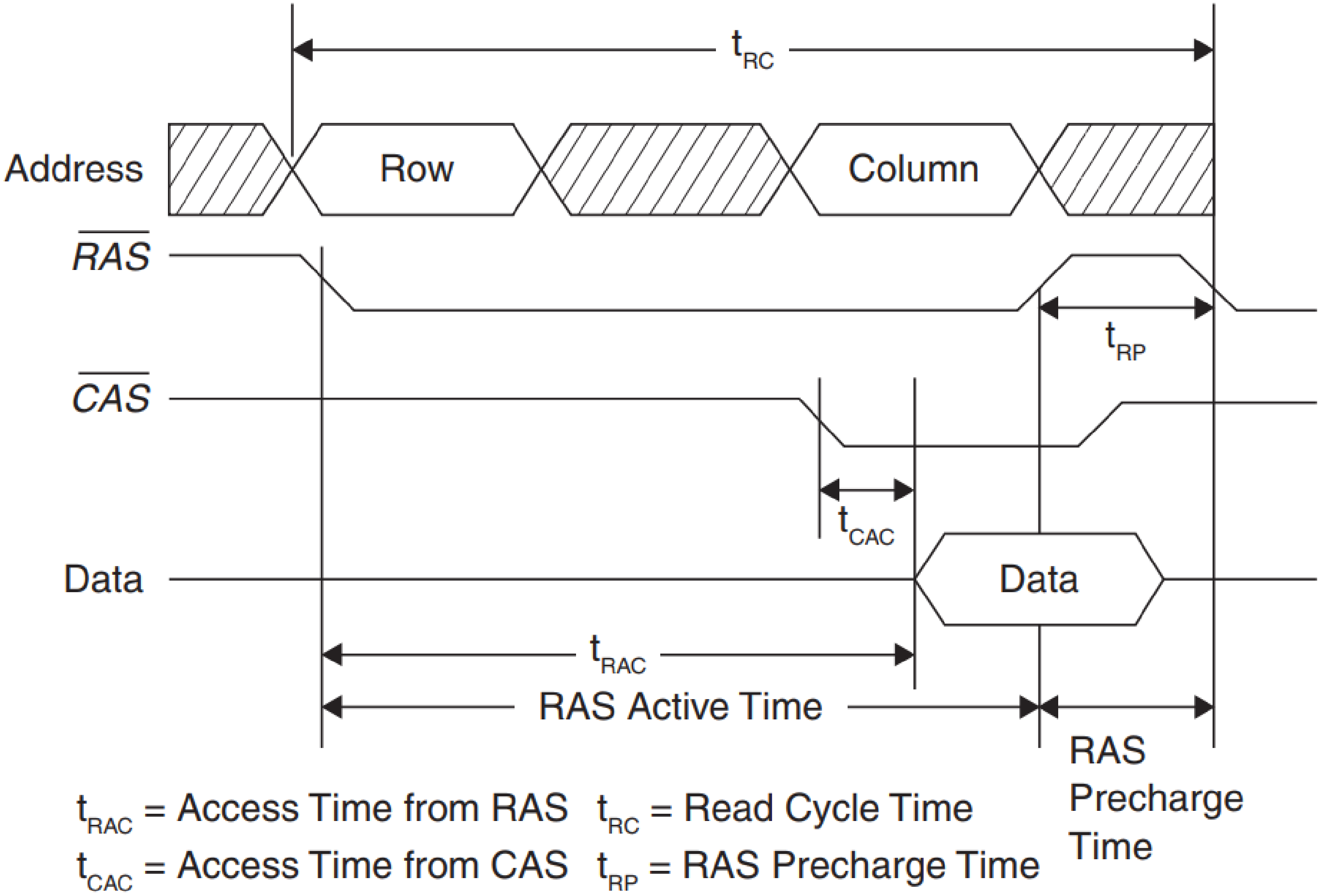
synchronous

- Shares a common clock with CPU, commands can be placed on its control pins on clock edges.
- In SDRAM, the input signals are latched into control logic block which functions as input to a state machine.
- State Machine controls memory Access.
- Read, write and refresh are initiated by loading control commands into device.

DDR (Double Data Rate) SDRAM

- It can be made to transfer data at rising and falling edges of the clock, instead of just at rising edge.
- The key idea is to take advantage of the fact that a **large number of bits are accessed at the same time** inside the chip when a row address is applied.
- To make the best use of the available clock speed, **data are transferred** externally on both the **rising** and **falling edges** of the clock.
- That is why it is called **double the data rate.**
- **Several versions of DDR chips have been developed**:
  - **DDR, DDR2, DDR3**, and **DDR4** with enhanced capabilities in terms **of increased storage capacity, lower power,** and **faster clock speeds**.

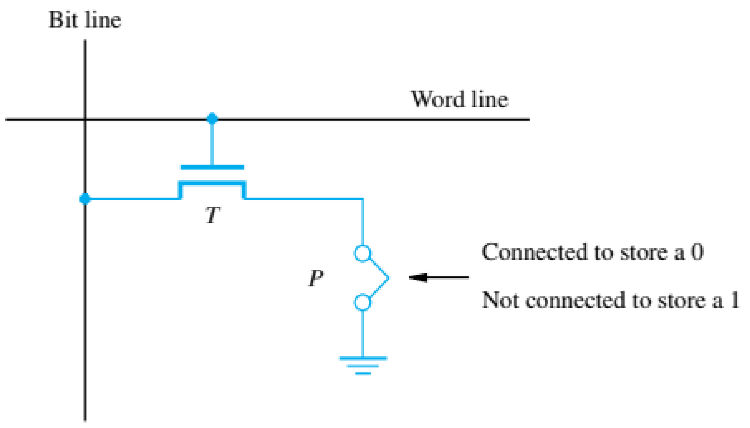
DDR (Double Data Rate) SDRAM



Read-Only Memories

Read-Only Memories

- Both **SRAM** and **DRAM** chips are **volatile**, which means that they retain information only while power is turned on.
- There are many applications requiring memory devices that retain the stored information when power is turned off.



For Example:

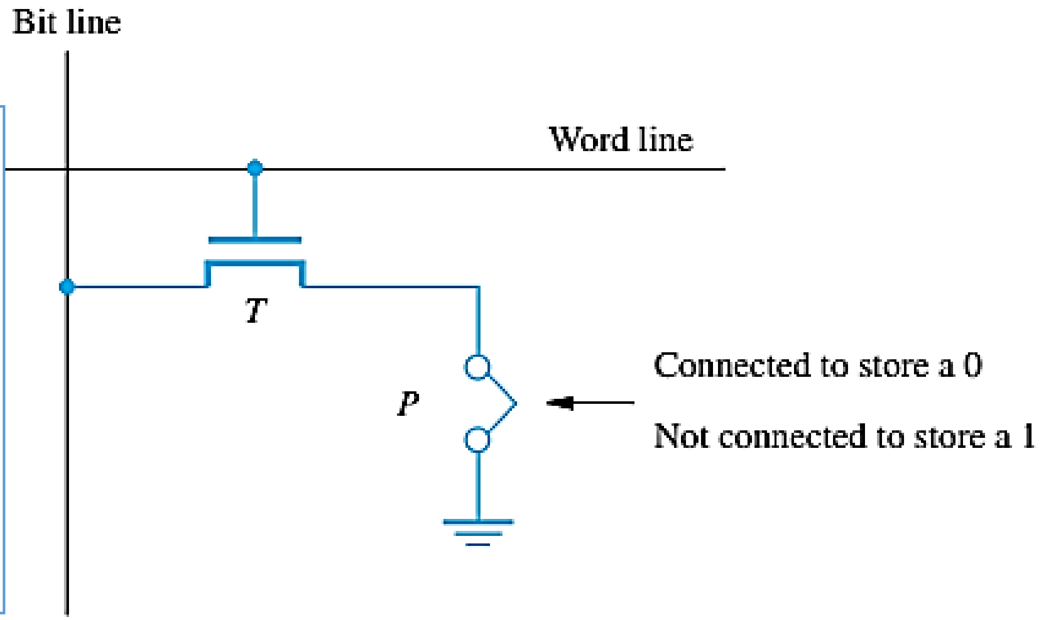
- Booting information of a computer.
- Many embedded systems do not use a hard disk and require nonvolatile memories to store their software. Such as; fire alarm, washing machine etc.

➢ Different types of **nonvolatile** memories have been developed.

➢ Generally, their **contents can be read in the same way** and a **special writing process** is needed to place the information into a nonvolatile memory.

➢ Since its normal operation involves only reading the stored data, a memory of this type is called a ***read-only memory*** **(ROM)**.

ROM



- A **logic value 0** is stored in the cell if the transistor is connected to ground at point *P*; **otherwise, a 1** is stored.
- The bit line is connected through a resistor to the power supply.
- **To read the state of the cell**, the **word line is activated** to **close the transistor switch**.

- As a result, the **voltage on the bit line drops to near zero** if there is a connection between the transistor and ground.
- If there is **no connection to ground**, the **bit line remains at the high** voltage level, indicating a **1**.
- A **sense circuit** at the end of the bit line generates the proper output value.
- The **state of the connection to ground in each cell** is determined when the chip is manufactured, using a mask with a pattern that represents the information to be stored.

PROM

- Some ROM designs allow the data to be loaded by the user, thus providing a ***programmable ROM*** **(PROM).**
- Programmability is achieved by inserting a fuse at **point *P*.**
- Before it is programmed, the memory contains all **0s**.
- The user can insert **1s** at the required locations by burning out the fuses at these locations using high-current pulses. (Of course, this process is irreversible).
- PROMs provide flexibility and convenience not available with ROMs.
- The cost of preparing the masks needed for storing a particular information pattern makes **ROMs cost effective only in large volumes**.
- The alternative technology of **PROMs** provides a more convenient and considerably less expensive approach, because memory chips can be programmed directly by the user.
- **Types of PROMS**: **EPROM** and **EEPROM**

PROM: EPROM and EEPROM

EPROM (Erasable and Programmable ROM):

- Contents can be erased by exposing to **ultraviolet radiation**.
- Such ROMs have a window through which UV light is applied in the chip.

EEPROM (Electrically Erasable PROM):

- Erasure can be done while chip is on circuit board.
- Programmer can change the data one byte at a time, takes long time when erasing.
- EEPROM is non-volatile, but also erasable and reprogrammable
- Used for data storage in small quantities where data have to be read frequently, may not have to be changed normally.

Interleaved Memories