UNIT - 4

The Memory System

Basic Concepts

The maximum size of the memory that can be used in any computer is determined by the addressing scheme (Each memory location is specified by an address). For example, a computer that generates 16-bit addresses is capable of addressing up to 216 = 64K (kilo) memory locations.

Machines whose instructions generate 32-bit addresses can utilize a memory that contains up to 232 = 4G (giga) locations, whereas machines with 64-bit addresses can access up to 264 = 16E (exa) ≈ 16 × 1018 locations.

The number of locations represents the size of the address space of the computer.

The memory is usually designed to **store and retrieve data in word-length quantities**. Memory transfers usually happen in word granularities

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

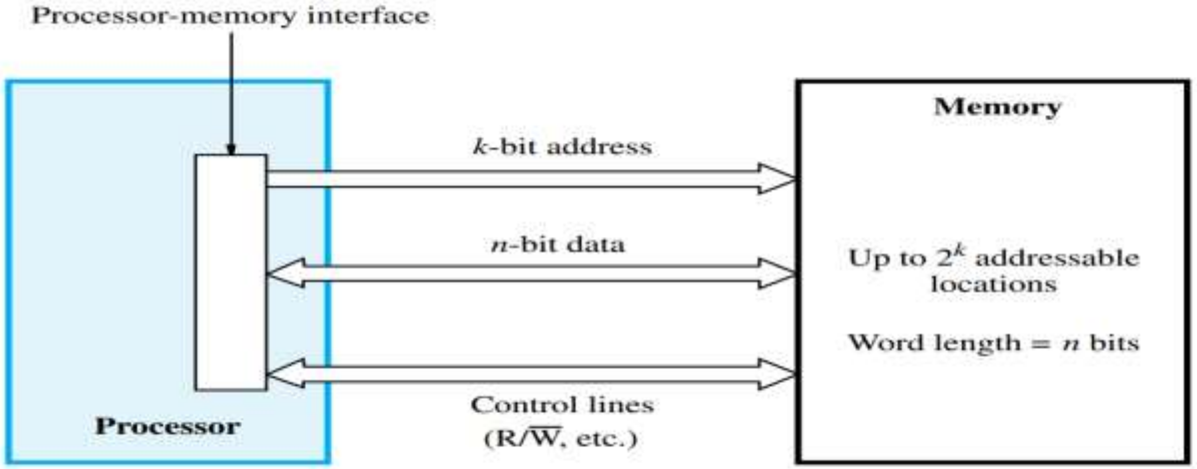

Processor-memory interface
Memory
k-bit address
n-bit data
Up to 2^k addressable locations
Word length = n bits
Control lines ($R/\overline{W}$, etc.)
Processor

Figure 8.1 Connection of the memory to the processor.

- Measures for memory speed:

Memory access time: Time elapsed between the initiation of an operation to transfer data from/to memory and the completion of that operation

Memory cycle time: Time required between initiation of two successive memory accesses for example, the time between two successive Read operations.

- Most important issue in memory systems design is to provide a computer with as large and fast a memory as possible, within a given cost target

- Cost of a memory device depends on both its capacity (total no. of bits) and its density (bits per unit area)

- For a fixed capacity, higher density => less chip area => less cost per bit

- Different memory technologies have different cost & speed characteristics

- Typically memory speed & memory cost are competing constraints

- 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

In non-RAM storage devices, such as magnetic and optical disks, access time depends on the address or position of data

Two types of RAMs: – Static Random Access Memory (SRAM) , Dynamic Random Access Memory

Cache and Virtual Memory

Cache memory is a chip-based computer component that makes retrieving data from the computer's memory more efficient. It acts as a temporary storage area that the computer's processor can retrieve data from easily. This temporary storage area, known as a cache, is more readily available to the processor than the computer's main memory .

The virtual memory is a logical memory. It is a memory management technique handled by the operating system. Virtual memory allows the programmer to use more memory for a program than the available main memory.

For example, assume that a computer has a main memory of 4GB and a virtual memory of 16GB. The user can use this 16GB to execute the program. Therefore, the user can execute programs which require more memory than the capacity of the main memory.

Block Transfers

The 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.

Semiconductor RAM Memories

Semiconductor memories are the **volatile memory** storages that store the program and data until the power supply to the system is ON. The cycle time of these semiconductor memories ranges from 100 ns to 10 ns. The cycle time is the time from the start of one access to the start of the next access to the memory.

Almost all the memory units are made of semiconductor material, especially **silicon. Semiconductor memories are used for storing digital data as they can be accessed faster**.

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

The figure below represents the memory circuit with 16 words (W_0 to W_{15}) where each word has a word length of 8 bits (b_0 to b_7). So, this is referred to as 16×8 memory organization.

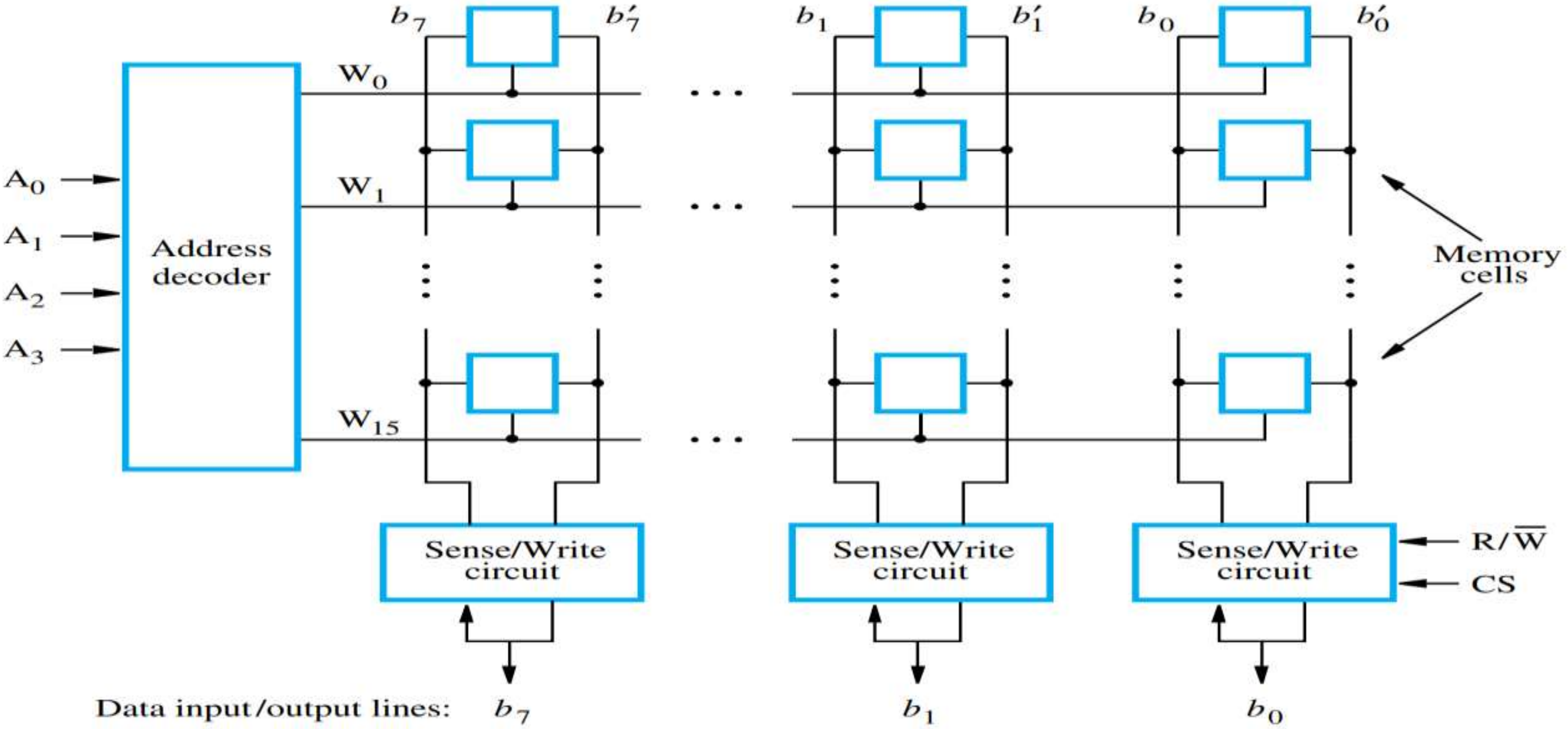


Figure 8.2 Organization of bit cells in a memory chip.

A very small memory circuit consisting of 16 words of 8 bits each. This is referred to as a **16 × 8 organization.**

The 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.

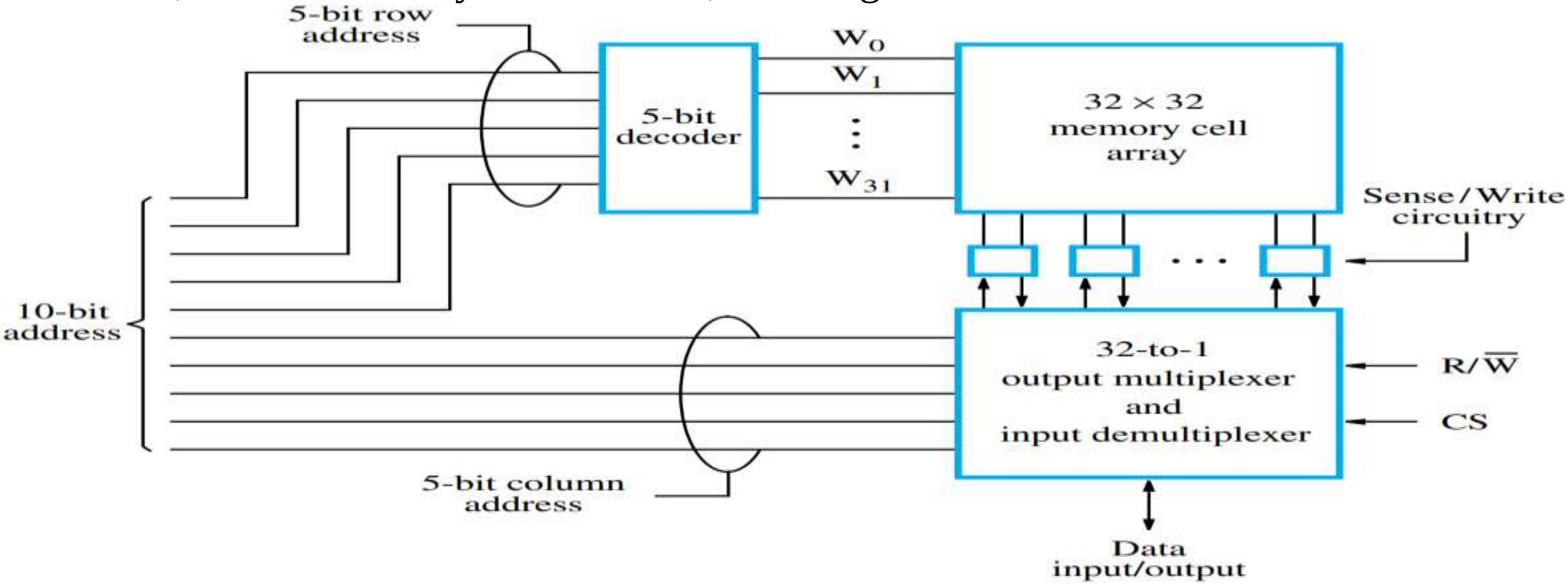
Two control lines, R/W and CS, are provided. The R/W (Read/Write) input specifies the required operation, and the CS (Chip Select) input selects a given chip in a multichip memory system.

stores 128 bits and requires 14 external connections for address, data, and control lines.

It also needs two lines for power supply and ground connections. Consider now a slightly larger memory circuit, one that has 1K (1024) memory cells.

This circuit can be organized as a 128 × 8 memory, requiring a total of 19 external connections.

Alternatively, the same number of cells can be organized into a 1K × 1 format. In this case, a 10-bit address is needed, but there is only one data line, resulting in 15 external connections



The required 10-bit address is divided into two groups of 5 bits each to form the row and column addresses for the cell array.

A row address selects a row of 32 cells, all of which are accessed in parallel. But, only one of these cells is connected to the external data line, based on the column address.

Large chips have essentially the same organization as Figure 8.3, but use a larger memory cell array and have more external connections.

For example, a 1G-bit chip may have a 256M × 4 organization, in which case a 28-bit address is needed and 4 bits are transferred to or from the chip.

Static Memories

- **In static RAM memories or SRAM, the content of the memory cell retains as long as the power supply to the memory chip is ON. In any situation, if the power supply to the memory chip is interrupted then the content of memory cells of static RAM is also lost.**
- When the power is resumed back to the memory chip there is no guarantee that the memory cells may have the same content as they have before the interruption of the power supply. This is the reason the static RAM is volatile in nature

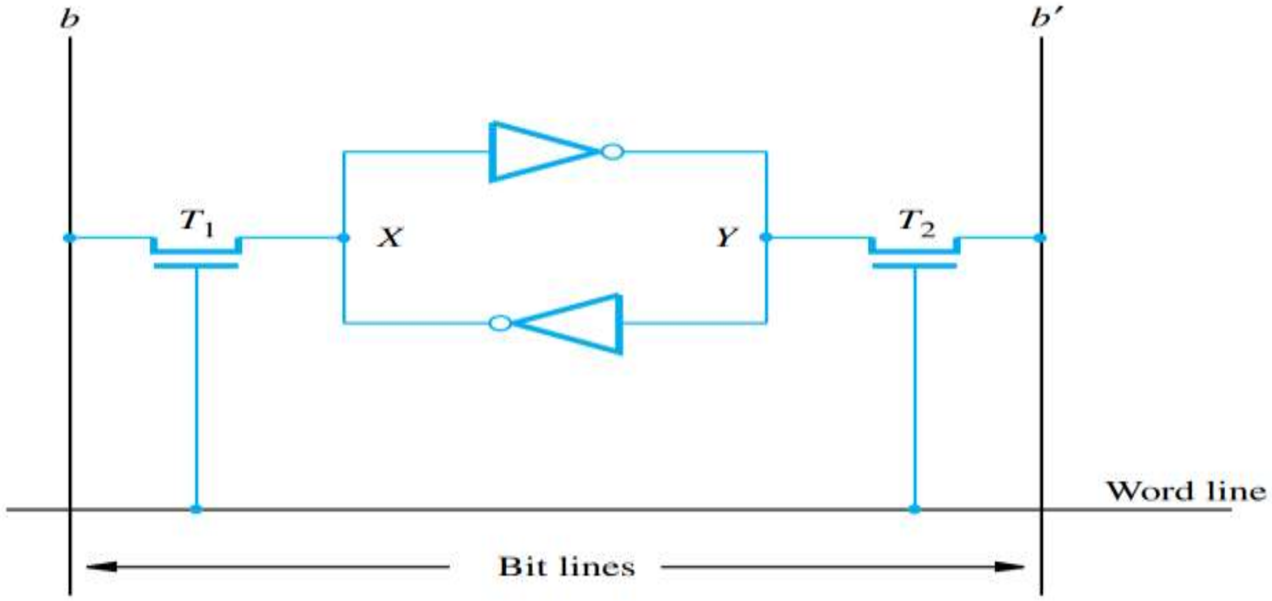


Figure 8.4 A static RAM cell.

Two inverters are cross-connected to form a latch. The latch is connected to two bit lines by transistors T1 and T2. 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.

Read Operation

In order to read the state of the SRAM cell, the word line is activated to close switches T1 and T2. 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.

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.

CMOS Cell

CMOS SRAMs i.e. complementary metal-oxide-semiconductor memory consumes very low memory as the power is supplied through the cells only when the cell is being accessed.

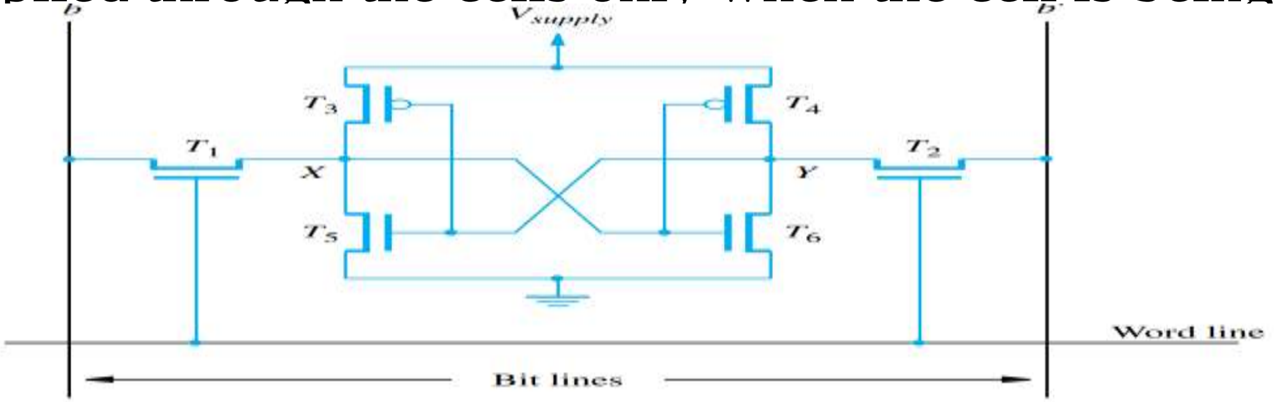


Figure 8.5 An example of a CMOS memory cell.

Transistor pairs (T3, T5) and (T4, T6) form the inverters in the latch.

The state of the cell is read or written as just explained. For example, in state 1, the voltage at point X is maintained high by having transistors T3 and T6 on, while T4 and T5 are off.

If T1 and T2 are turned on, bit lines b and b' will have high and low signals, respectively

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, T1, T2, and one transistor in each inverter are turned off, ensuring that there is no continuous electrical path between Vsupply and ground.

Dynamic RAMs

Though the static RAM is faster its memory cells require several transistors which makes it expensive. So, to design a less expensive and higher density RAM we have to implement it using simpler cells.

But, the fact with the simpler cell is that the simpler cell does not hold data for a long period until the data is accessed from the cell frequently either for read or write operation. The memory circuit implemented using such simpler cells is referred to as *dynamic RAMs*.

In dynamic RAM the information is stored in the memory cell by imposing the charge on the capacitor.

But the capacitor can hold charge only for tens of milliseconds and to hold the content for a longer time the capacitor must be charged to its full value.

The capacitor is charged while its content is refreshed either by reading the contents from the cell or writing new information to the cell.

An example of a dynamic memory cell that consists of a capacitor, C, and a transistor, T, is shown in Figure.

To store information 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.

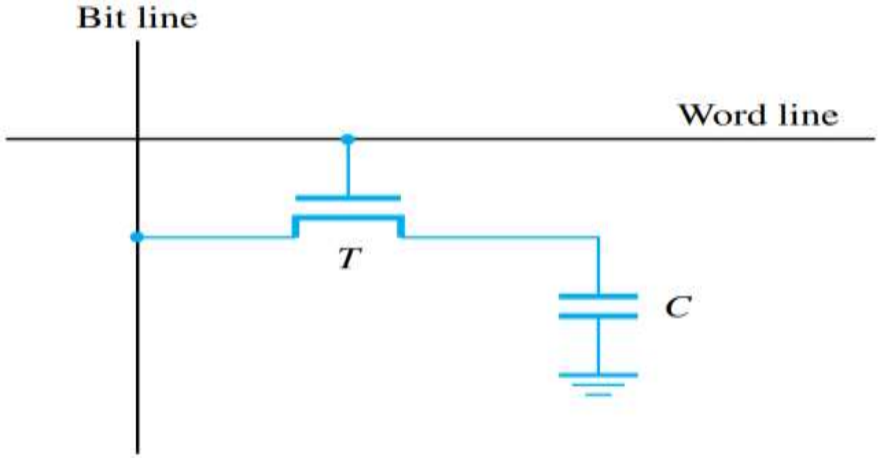


Figure 8.6 A single-transistor dynamic memory cell.

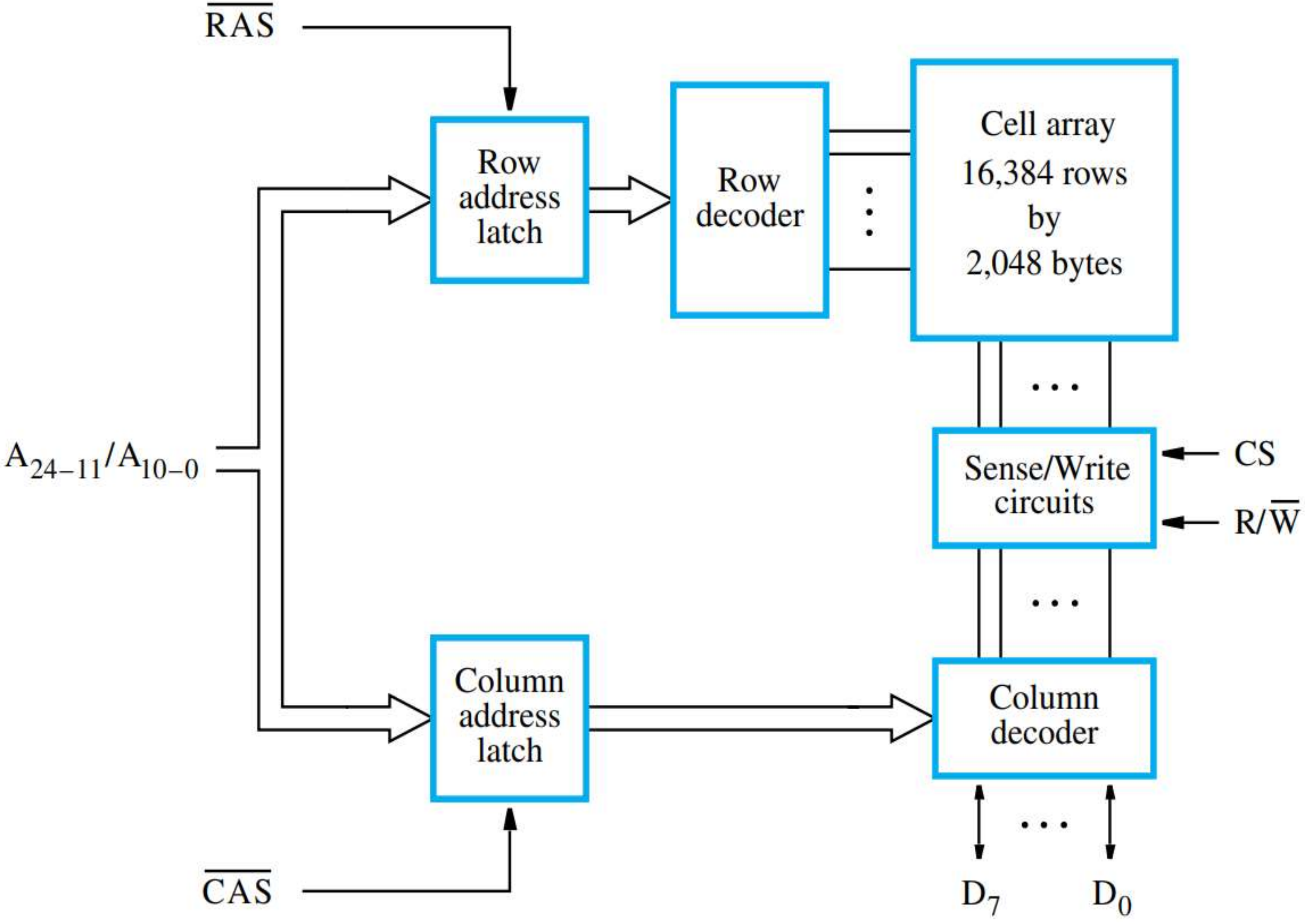


Figure 8.7 Internal organization of a 32M × 8 dynamic memory chip.

Synchronous DRAMs

In the early 1990s, developments in memory technology resulted in DRAMs whose operation is synchronized with a clock signal. Such memories are known as synchronous DRAMs (SDRAMs).

Their structure is shown in Figure. The cell array is the same as in asynchronous DRAMs. The distinguishing feature of an SDRAM is the use of a clock signal, the availability of which makes it possible to incorporate control circuitry on the chip that provides many useful features.

For example, 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.

The address and data connections of an SDRAM may be buffered by means of registers

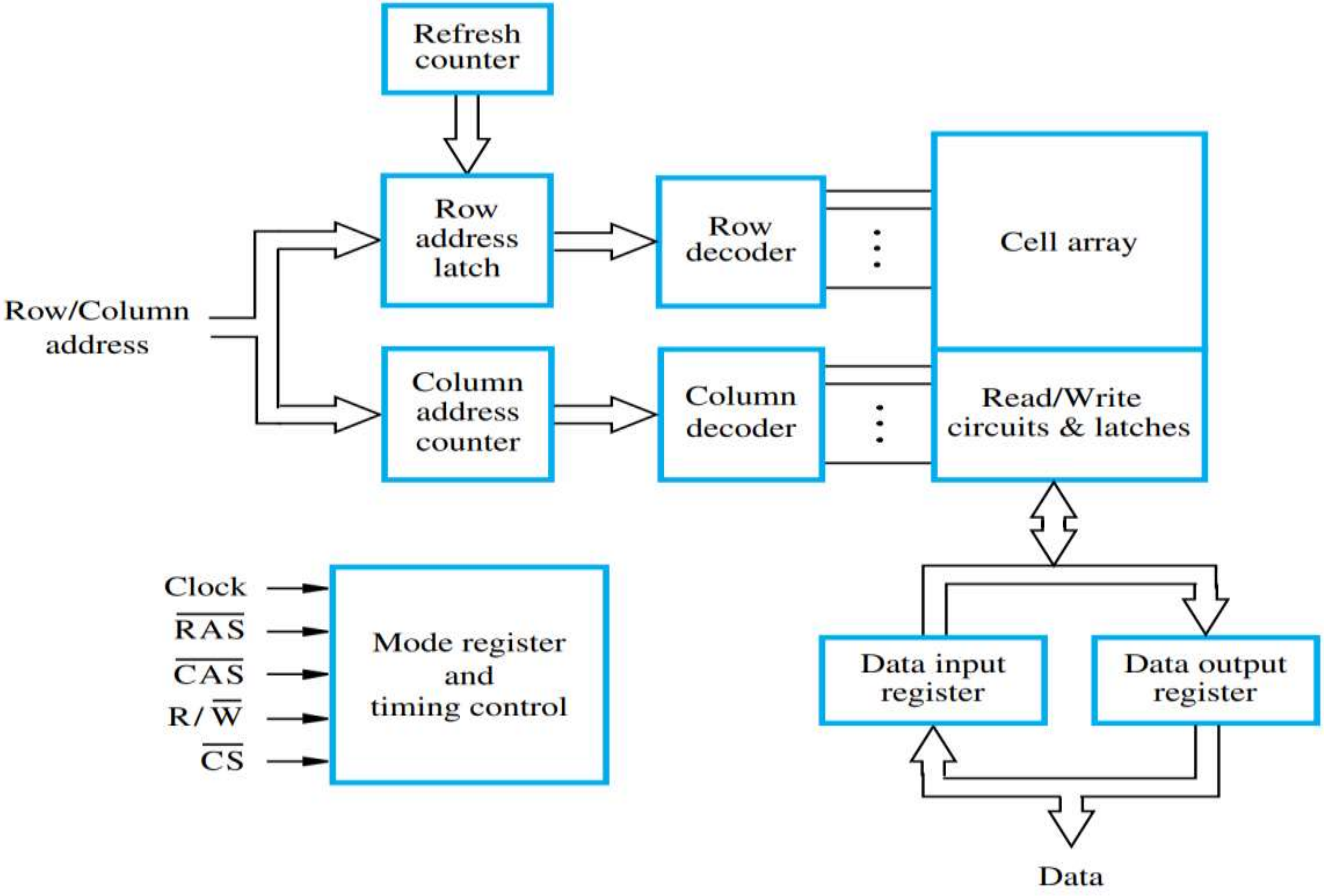


Figure 8.8 Synchronous DRAM.

Read-only Memories

Both static and dynamic RAM 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.

What is ROM?

- ROM, which stands for read only memory, is a memory device or storage medium that stores information permanently.
- It is also the primary memory unit of a computer along with the random access memory (RAM). It is called read only memory as we can only read the programs and data stored on it but cannot write on it.
- It is restricted to reading words that are permanently stored within the unit.

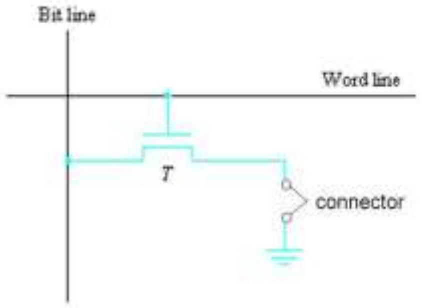
Types of ROM:

- Masked Read Only Memory (MROM)
- Programmable Read Only Memory (PROM)
- Erasable and Programmable Read Only Memory (EPROM)
- Electrically Erasable and Programmable Read Only Memory (EEPROM)
- FLASH ROM(MEMORY)

Masked Read Only Memory (MROM):

- It is the oldest type of read only memory (ROM). It has become obsolete so it is not used anywhere in today's world.
- It is a hardware memory device in which programs and instructions are stored at the time of manufacturing by the manufacturer. So it is programmed during the manufacturing process and can't be modified, reprogrammed, or erased later.
- The MROM chips are made of **integrated circuits**. Chips send a current through a particular input-output pathway determined by the location of fuses among the rows and columns on the chip.
- The current has to pass along a fuse-enabled path, so it can return only via the output the manufacturer chooses. This is the reason the rewriting and any other modification is not impossible in this memory.

A ROM CELL:



- leave connected to store a 0
- disconnect to store a 1

- a possible configuration for a ROM cell. A logic value 0 is stored the cell if the transistor is connected to ground at point P; otherwise, a I 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.
- Data are written into a ROM when it is manufactured.

Programmable Read Only Memory (PROM):

- PROM is a blank version of ROM. It is manufactured as blank memory and programmed after manufacturing. We can say that it is kept blank at the time of manufacturing.
- In the chip, the current travels through all possible path ways. The programmer can choose one particular path for the current by burning unwanted fuses by sending a high voltage through them.
- The user has the opportunity to program it or to add data and instructions as per his requirement. Due to this reason, it is also known as the user-programmed ROM as a user can program it.
- To write data onto a PROM chip; a device called PROM programmer or PROM burner is used.
- The process or programming a PROM is known as burning the PROM. Once it is programmed, the data cannot be modified later, so it is also called as one-time programmable device.

- PROMs provide flexibility and convenience not available with ROMs. The latter are economically attractive for storing fixed programs and data when high volumes of ROMs are produced.
- PROMs provide a faster and considerably less expensive approach because they can be programmed directly by the user

Uses: It is used in cell phones, video game consoles, medical devices, RFID tags, and more.

Erasable and Programmable Read Only Memory (EPROM)

- EPROM is a type of ROM that can be reprogramed and erased many times. The method to erase the data is very different; it comes with a quartz window through which a specific frequency of ultraviolet light is passed for around 40 minutes to erase the data.
- So, it retains its content until it is exposed to the ultraviolet light. You need a special device called a PROM programmer or PROM burner to reprogram the EPROM.
- Another type of ROM chip allows the stored data to be erased and new data to be loaded. Such an erasable, reprogrammable ROM is usually called an EPROM. It pro- vides considerable flexibility during the development phase of digital systems.

- Since EPROMs are capable of retaining stored information for a long time, they can be used in place of ROMs while software is being developed. In this way, memory changes and updates can be easily made.
- The important advantage of EPROM chips is that their contents can be erased and reprogrammed. Erasure requires dissipating the charges trapped in the transistors of memory cells; this can be done by exposing the chip to ultraviolet light. For this reason. EPROM chips are mounted in packages that have transparent windows.

Uses: It is used in some micro-controllers to store program, e.g., some versions of Intel 8048

Electrically Erasable and Programmable Read Only Memory (EEPROM):

- ROM is a type of read only memory that can be erased and reprogrammed repeatedly, up to 10000 times. It is also known as Flash EEPROM as it is similar to flash memory.
- It is erased and reprogrammed electrically without using ultraviolet light. Access time is between 45 and 200 nanoseconds.
- The data in this memory is written or erased one byte at a time; byte per byte, whereas, in flash memory data is written and erased in blocks.
- So, it is faster than EEPROM. It is used for storing a small amount of data in computer and electronic systems and devices such as circuit boards.

Uses: The BIOS of a computer is stored in this memory.

FLASH MEMORY

- It is an advanced version of EEPROM. It stores information in an arrangement or array of memory cells made from floating-gate transistors. The advantage of using this memory is that you can delete or write blocks of data around 512 bytes at a particular time.
- Whereas, in EEPROM, you can delete or write only 1 byte of data at a time. So, this memory is faster than EEPROM
- It can be reprogrammed without removing it from the computer. Its access time is very high, around 45 to 90 nanoseconds. It is also highly durable as it can bear high temperature and intense pressure.
- An approach similar to EEPROM technology has more recently given rise to flash memory devices. A flash cell is based on a single transistor controlled by trapped charge, just like an EEPROM cell.
- Flash devices have greater density, which leads to higher capacity and a lower cost per bit. They require a single power supply voltage, and consume less power in their operation.
- The low power consumption of flash memory makes it attractive for use in portable equipment that is battery driven. Typical applications include hand-held computers cell phones, digital cameras, and MP3 music players. In hand-held computers and cell phones, flash memory holds the software needed to operate the equipment, thus obviating the need for a disk drive.
- In digital cameras, flash memory is used to store picture image data. In MP3 players, flash memory stores the data that represent sound Cell. phones, digital cameras, and MP3 players are good examples

Flash Cards

- One way of constructing a larger module is to mount flash chips on a small card. Such flash cards have a standard interface that makes them usable in a variety of products.
- A card is simply plugged into a conveniently accessible slot. Flash cards come in a variety of memory sizes. Typical sizes are 8, 32, and 64 Mbytes. A minute of music can be stored in about I Mbyte of memory. using the MP3 encoding format Hence, a 64-MB flash card can store an hour of music.

Flash Drives

- Larger flash memory modules have been developed to replace hard disk drives, and hence are called flash drives. They are designed to fully emulate hard disks, to the point that they can be fitted into standard disk drive bays.
- However, the storage capacity of flash drives is significantly lower. Currently, the capacity of flash drives is on the order of 64 to 128 Gbytes.
- In contrast, hard disks have capacities exceeding a terabyte. Also, disk drives have a very low cost per bit.
- The fact that flash drives are solid state electronic devices with no moving parts provides important advantages over disk drives. They have shorter access times, which result in a faster response.
- They are insensitive to vibration and they have lower power consumption, which makes them attractive for portable, battery-driven applications.

Direct Memory Access

Direct Memory Access (DMA) transfers the block of data between the memory and peripheral devices of the system, without the participation of the processor. The unit that controls the activity of accessing memory directly is called a DMA controller.

- We have two other methods of data transfer, programmed I/O and Interrupt driven I/O. Let’s revise each and get acknowledge with their drawbacks.
- In programmed I/O, the processor keeps on scanning whether any device is ready for data transfer. If an I/O device is ready, the processor fully dedicates itself in transferring the data between I/O and memory. It transfers data at a high rate, but it can’t get involved in any other activity during data transfer. This is the major drawback of programmed I/O.
- In Interrupt driven I/O, whenever the device is ready for data transfer, then it raises an interrupt to processor. Processor completes executing its ongoing instruction and saves its current state. It then switches to data transfer which causes a delay. Here, the processor doesn’t keep scanning for peripherals ready for data transfer. But, it is fully involved in the data transfer process. So, it is also not an effective way of data transfer.

The above two modes of data transfer are not useful for transferring a large block of data. But, the DMA controller completes this task at a faster rate and is also effective for transfer of large data block.

Figure shows an example of the DMA controller registers that are accessed by the processor to initiate data transfer operations. Two registers are used for storing the starting address and the word count.

The third register contains status and control flags. The R/W bit determines the direction of the transfer. When this bit is set to 1 by a program instruction, the controller performs a Read operation, that is, it transfers data from the memory to the I/O device.

Otherwise, it performs a Write operation. Additional information is also transferred as may be required by the I/O device.

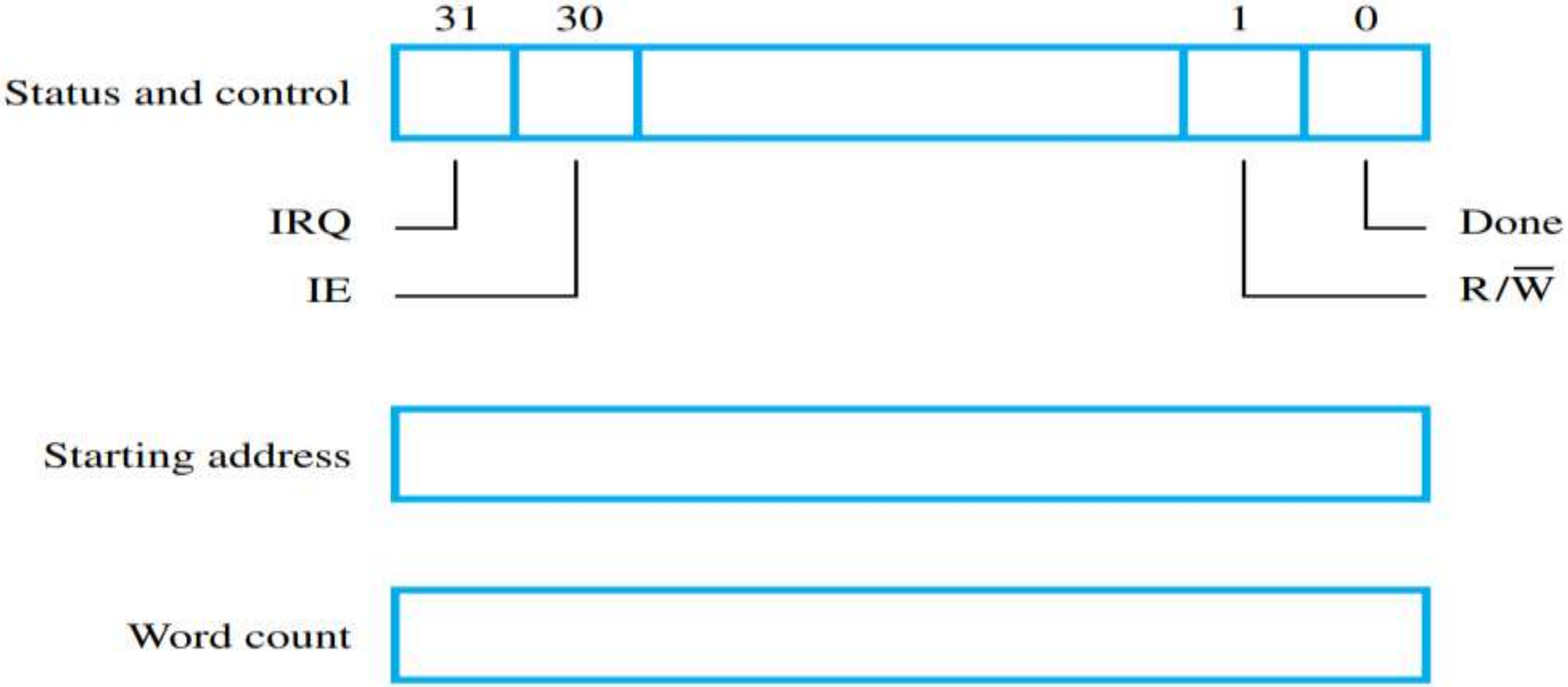


Figure 8.12 Typical registers in a DMA controller.

Figure shows how DMA controllers may be used in a computer system.

One DMA controller connects a high-speed Ethernet to the computer's I/O bus. The disk controller, which controls two disks, also has DMA capability and provides two DMA channels. It can perform two independent DMA operations, as if each disk had its own DMA controller.

The registers needed to store the memory address, the word count, and so on, are duplicated, so that one set can be used with each disk.

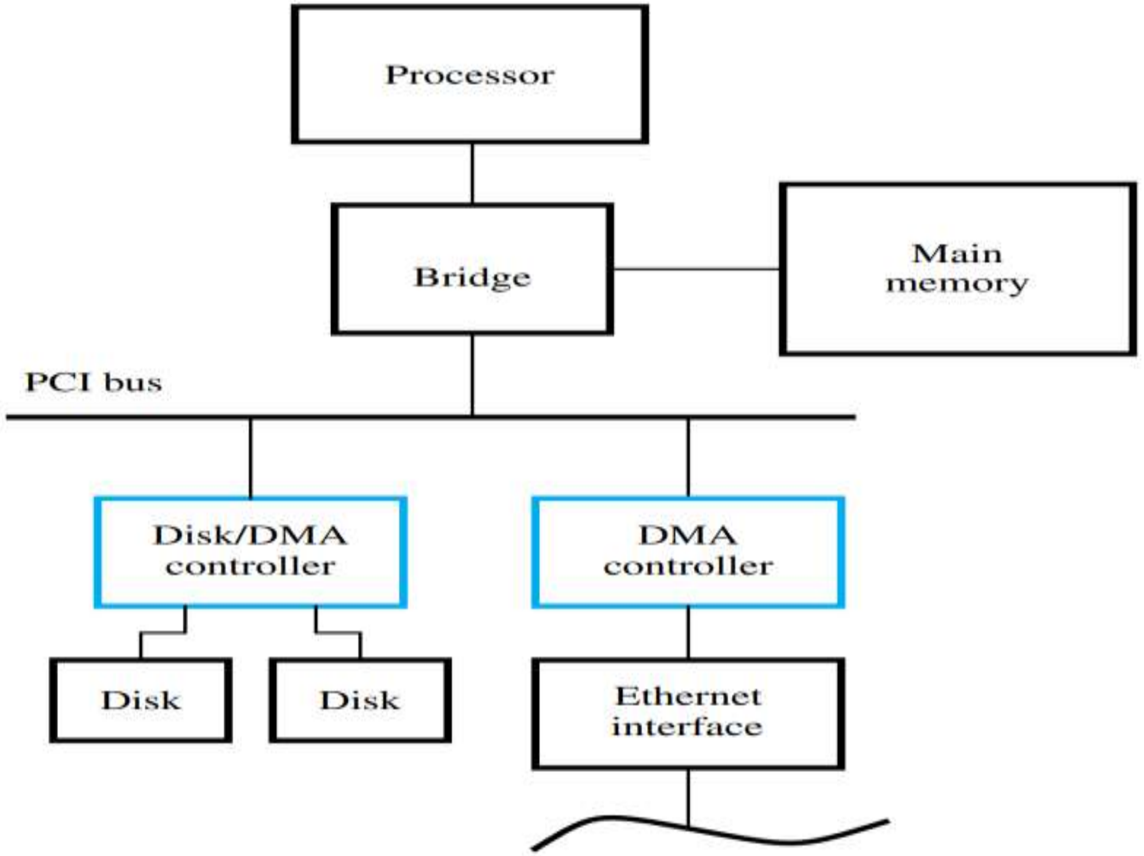


Figure 8.13 Use of DMA controllers in a computer system.

Cache memory

- A very high speed memory called a **Cache** is used to increase the speed of processor
-
- Cache memory reducing the total execution time of the program
- Cache memory is placed b/w the main memory and CPU
- Whenever CPU needs to access the data Cache memory is checked if the word is found in Cache memory it is need at the fast speed then transfer to CPU

Cache memory

- If the word is not found in Cache memory go to main memory and access the data
- If the Cache memory has no vacant space for storing the new information then create vacant space using page replacement algorithm
- The performance of Cache memory is frequently measured in terms of quality is called **Hit Ratio**
- Transfer of data from main memory to Cache memory is referred as mapping process

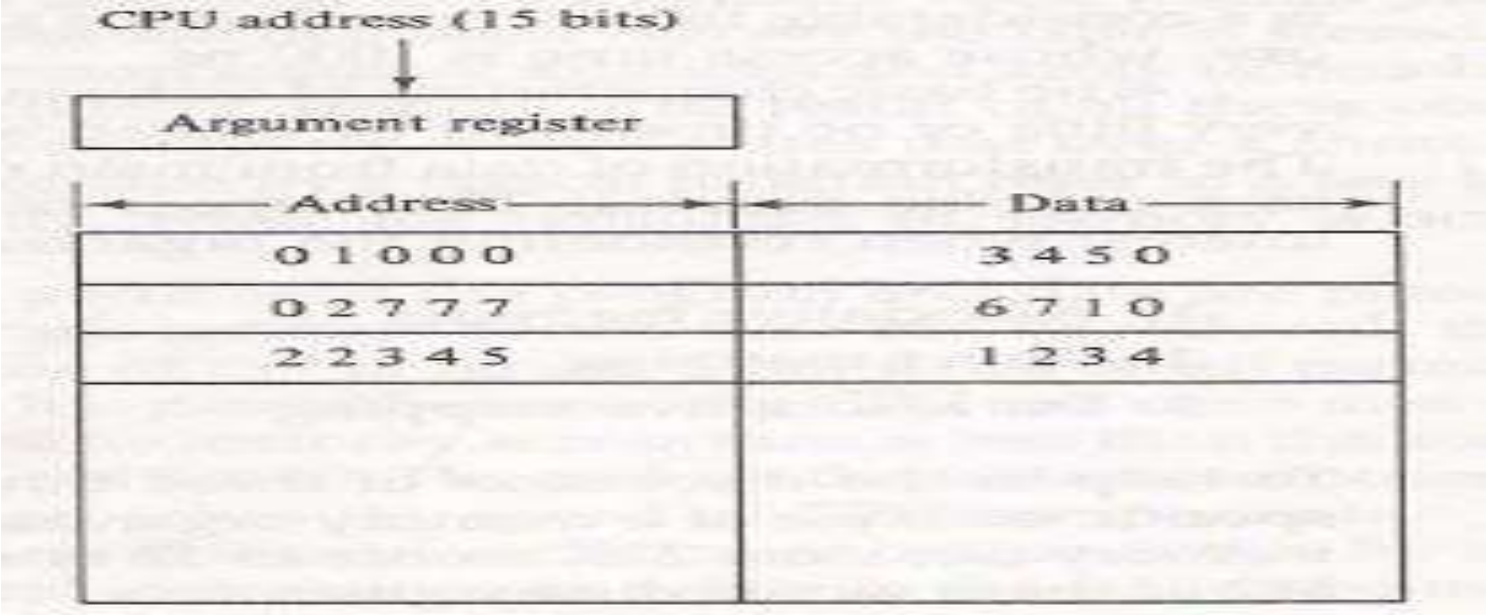
Cache memory

- There are three different types of mapping processes as shown below:

- 1. Associative mapping
- 2. Direct mapping
- 3. Set-associative mapping

Associative mapping

- **Associative mapping:** The fastest and most flexible cache organization uses an associative mapping.



Address	Data
01000	3450
02777	6710
22345	1234

The associative memory stores both the address and content of the memory word.
The address value of 15 bits is shown as a five digit octal number and its corresponding 12 bit word is shown as a four digit octal number

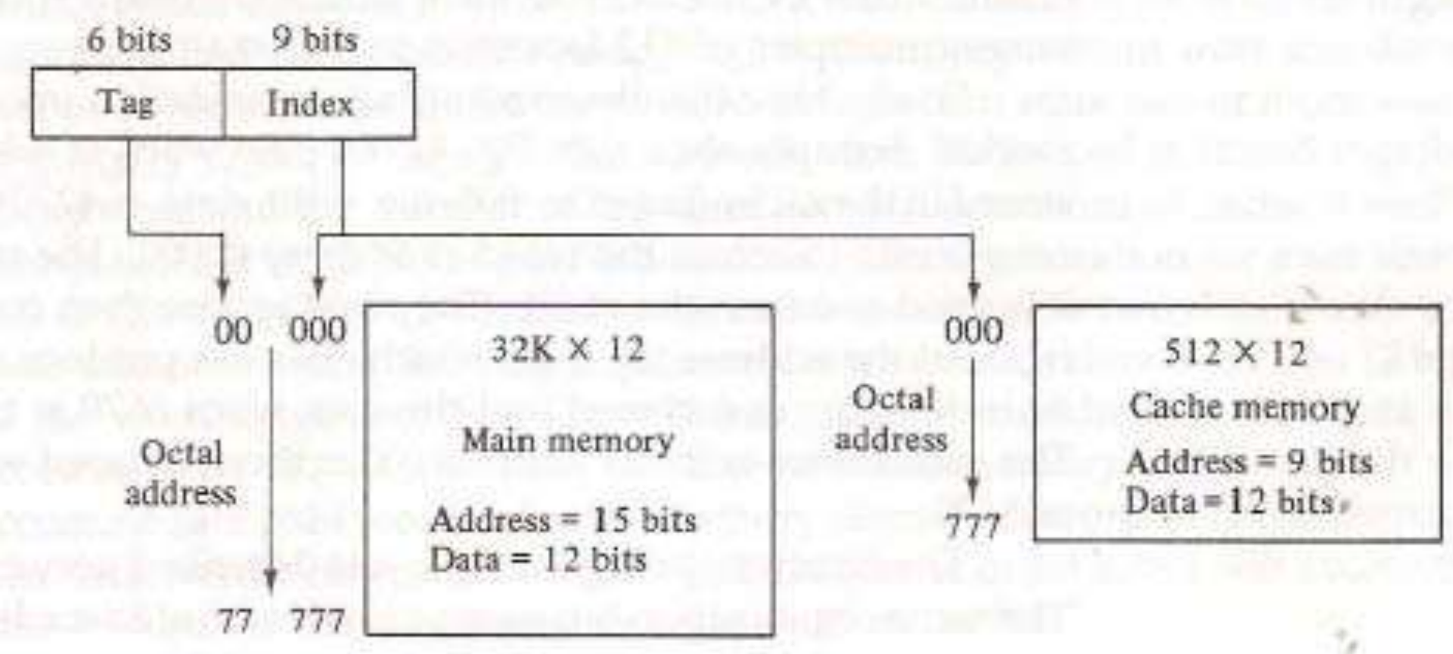
Associative mapping

- A CPU address of 15 bits is placed in the argument register and the associative memory is searched for a matching address.
- If the address is found, the corresponding 12 bit data is read and sent to the CPU.
- If no match occurs, the main memory is accessed for the word.

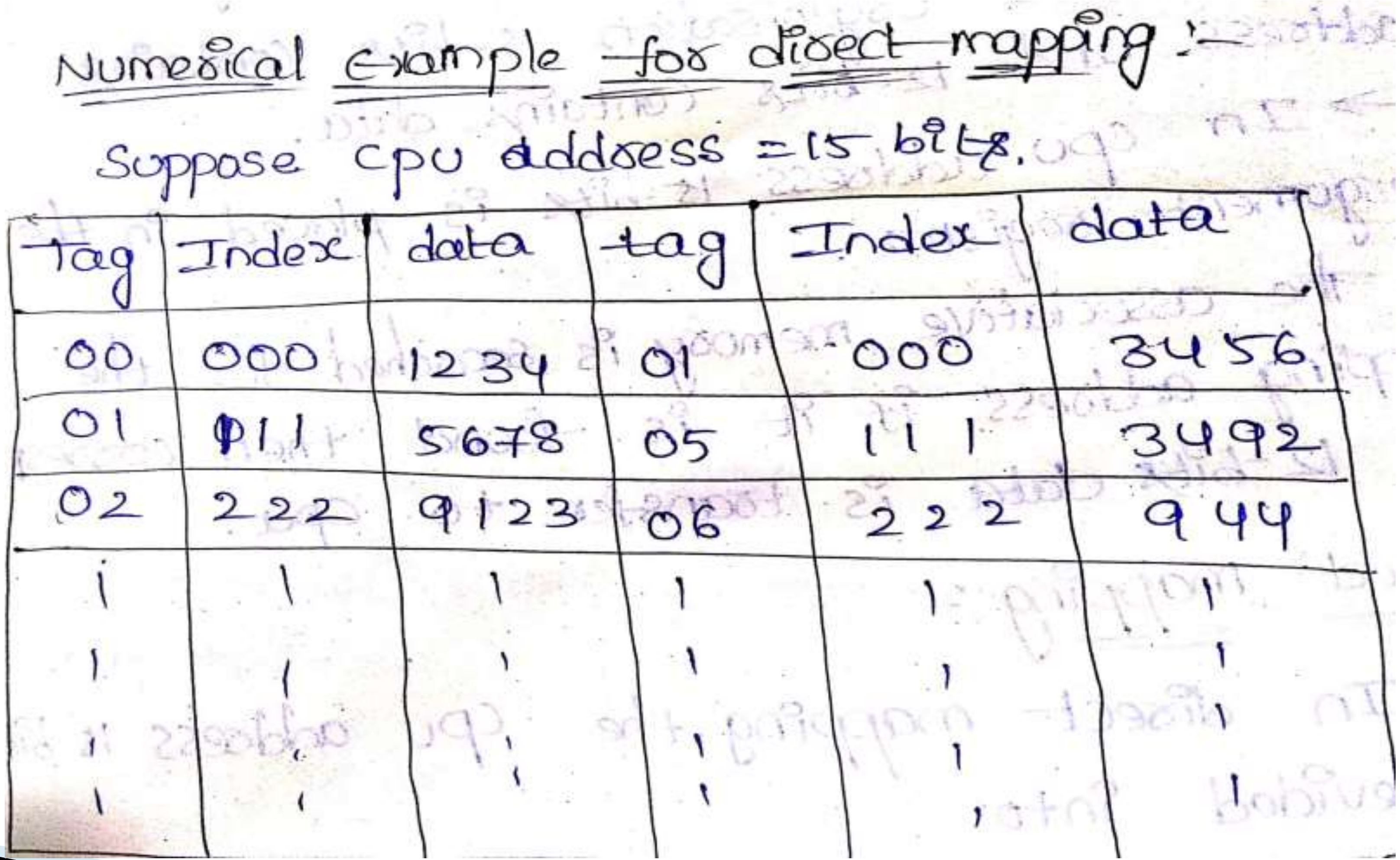
Direct mapping

- In this, the CPU address of 15 bits is divided into 2 fields. The nine least significant bits constitute the *index field and the remaining six bits form the tag field.*

Figure 12-12 Addressing relationships between main and cache memories.



Direct mapping example



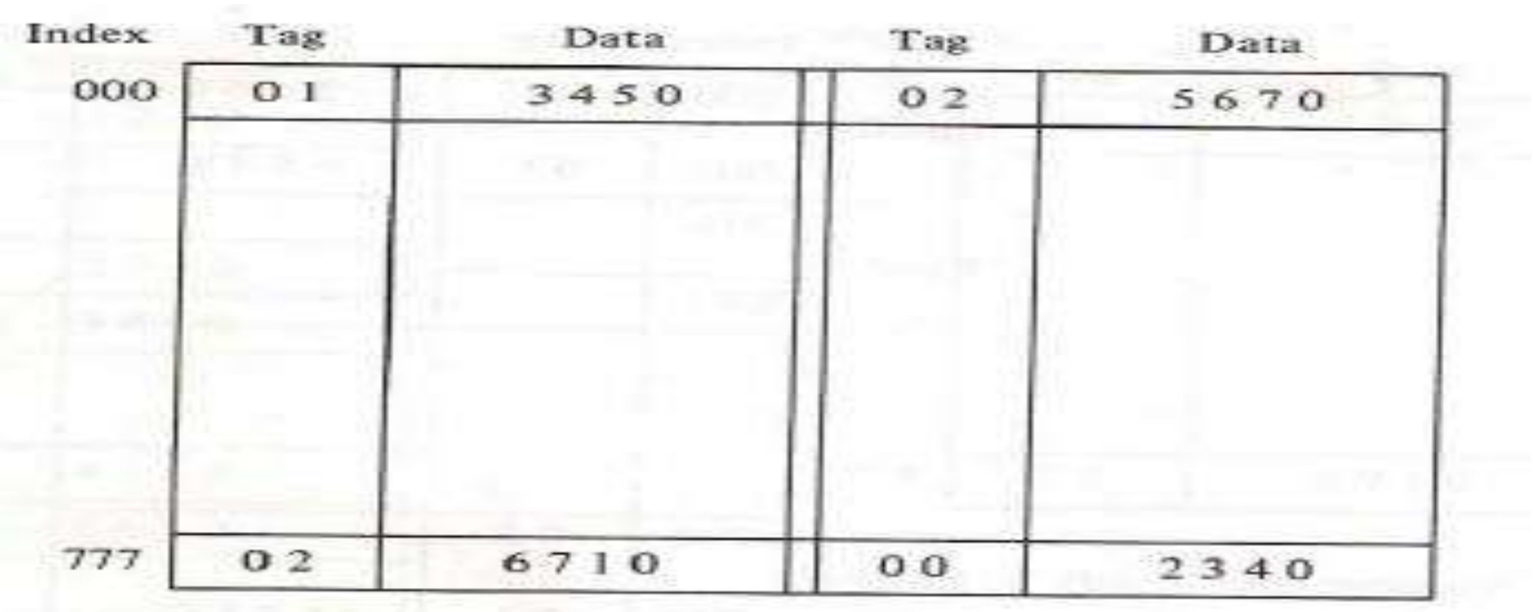
Numerical Example for direct mapping :-

Suppose cpu address = 15 bits.

Tag	Index	data	tag	Index	data
00	000	1234	01	000	3456
01	011	5678	05	111	3492
02	222	9123	06	222	944
⋮	⋮	⋮	⋮	⋮	⋮

Set-associative Mapping

- In this, each word of cache can store two or more words of memory under the same index address.
- Each data word is stored together with its tag.
- An example of set-associative cache organization is shown below.



Index	Tag	Data	Tag	Data
000	01	3450	02	5670
777	02	6710	00	2340

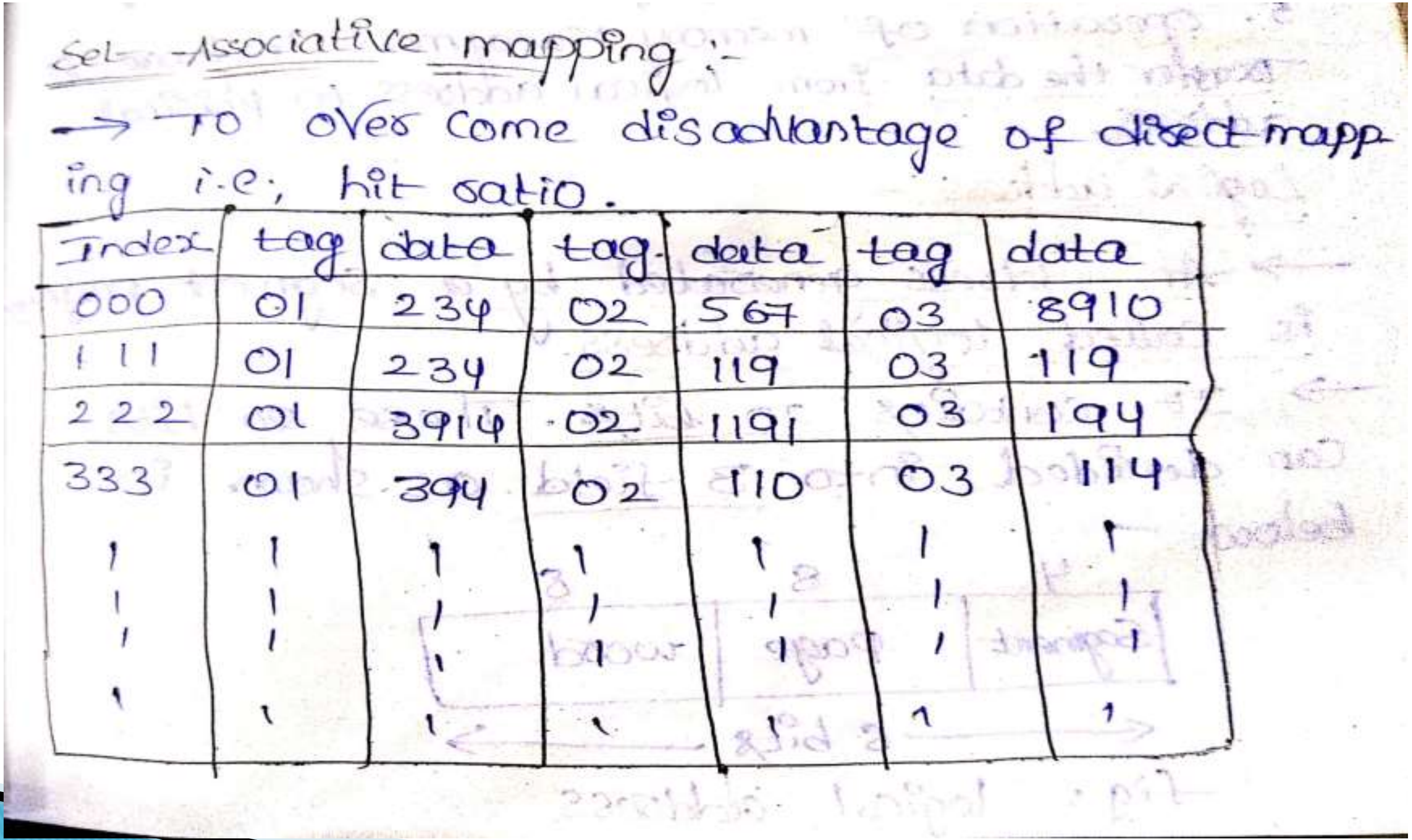
Figure 12-15 Two-way set-associative mapping cache.

Set-associative Mapping

- tag field of the CPU address is then compared with both tags in the cache to determine if a match occurs.

- The hit ratio will improve as the set size increases.

Set-associative Mapping example



Set-Associative mapping :-

→ To over come disadvantage of direct mapping i.e; hit ratio.

Index	tag	data	tag	data	tag	data
000	01	234	02	567	03	8910
111	01	234	02	119	03	119
222	01	3914	02	1191	03	194
333	01	394	02	110	03	114