



# **SNS COLLEGE OF ENGINEERING**

Kurumbapalayam (Po), Coimbatore – 641 107

**An Autonomous Institution**

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade  
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai



## **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**COURSE NAME :19IT301 COMPUTER ORGANIZATION AND  
ARCHITECTURE**

**II YEAR /III SEMESTER**

**Unit 4- Memory system**



# UNIT 4 MEMORY SYSTEM



Basic concepts

RAM memories

ROM memory

Speed, size and cost

Cache memory

Performance

Virtual memory

Memory management requirements

Secondary storage



# Basic concepts



MR operation

MW operation

Memory access time: The time that elapses between the initiation of an operation and the completion of that operation

Memory cycle time: Time delay required between the initiation of two successive memory operations (two read operations).



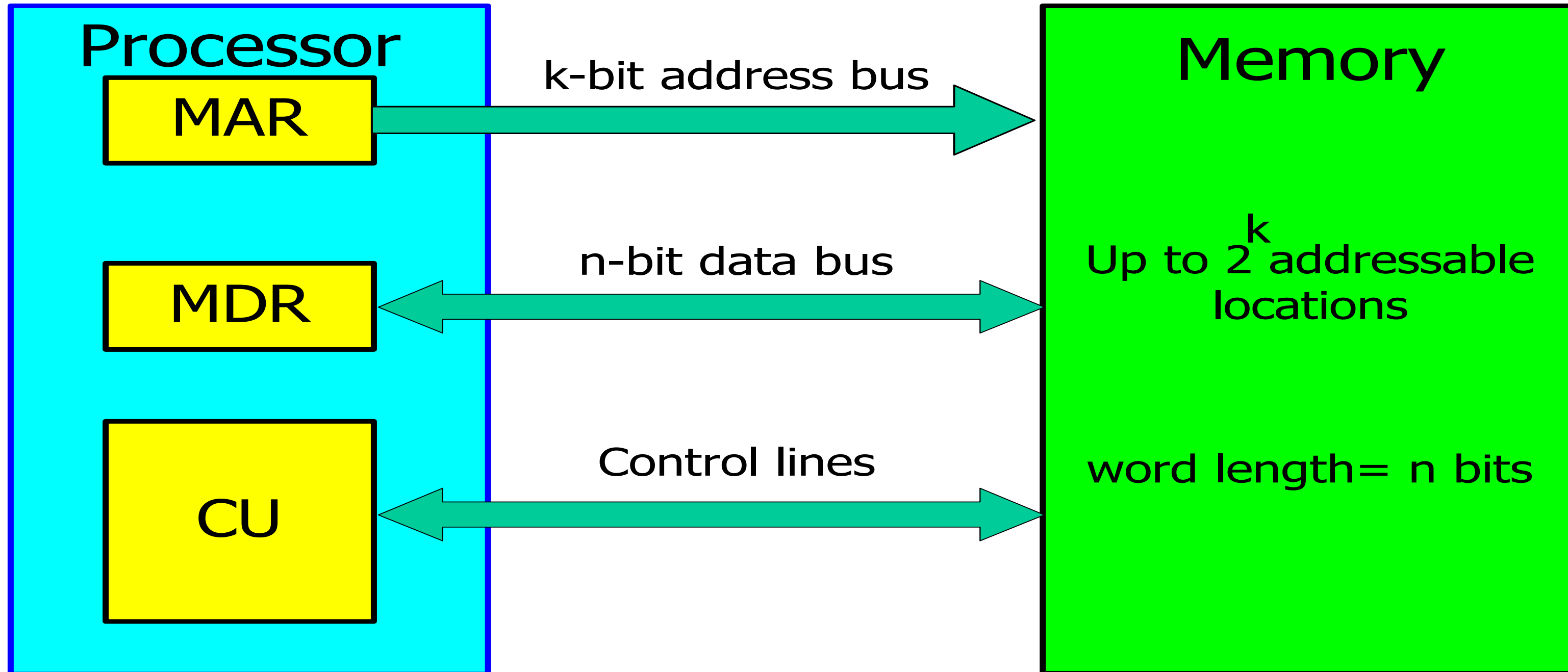
- ✓ Cache memory: A special buffer storage ,smaller and faster than main storage, that is used to hold a copy of instructions and data in main storage that are likely to be needed next by the processor and that have been obtained automatically from main storage
- ✓ It is small, fast memory placed between a processor and main memory.
- ✓ It reduces memory access time.



Virtual memory: It is a concept that an user can view the computer as having a single addressable memory of essentially unlimited size to which he alone has access.



# Connection of the memory to the CPU





# SEMICONDUCTOR RAM MEMORIES



INTERNAL ORGANIZATION OF MEMORY CHIPS  
STATIC MEMORIES  
ASYNCHRONOUS DRAMS  
SYNCHRONOUS DRAMS  
STRUCTURE OF LARGER MEMORIES  
MEMORY SYSTEM CONSIDERATIONS  
RAMBUS MEMORY



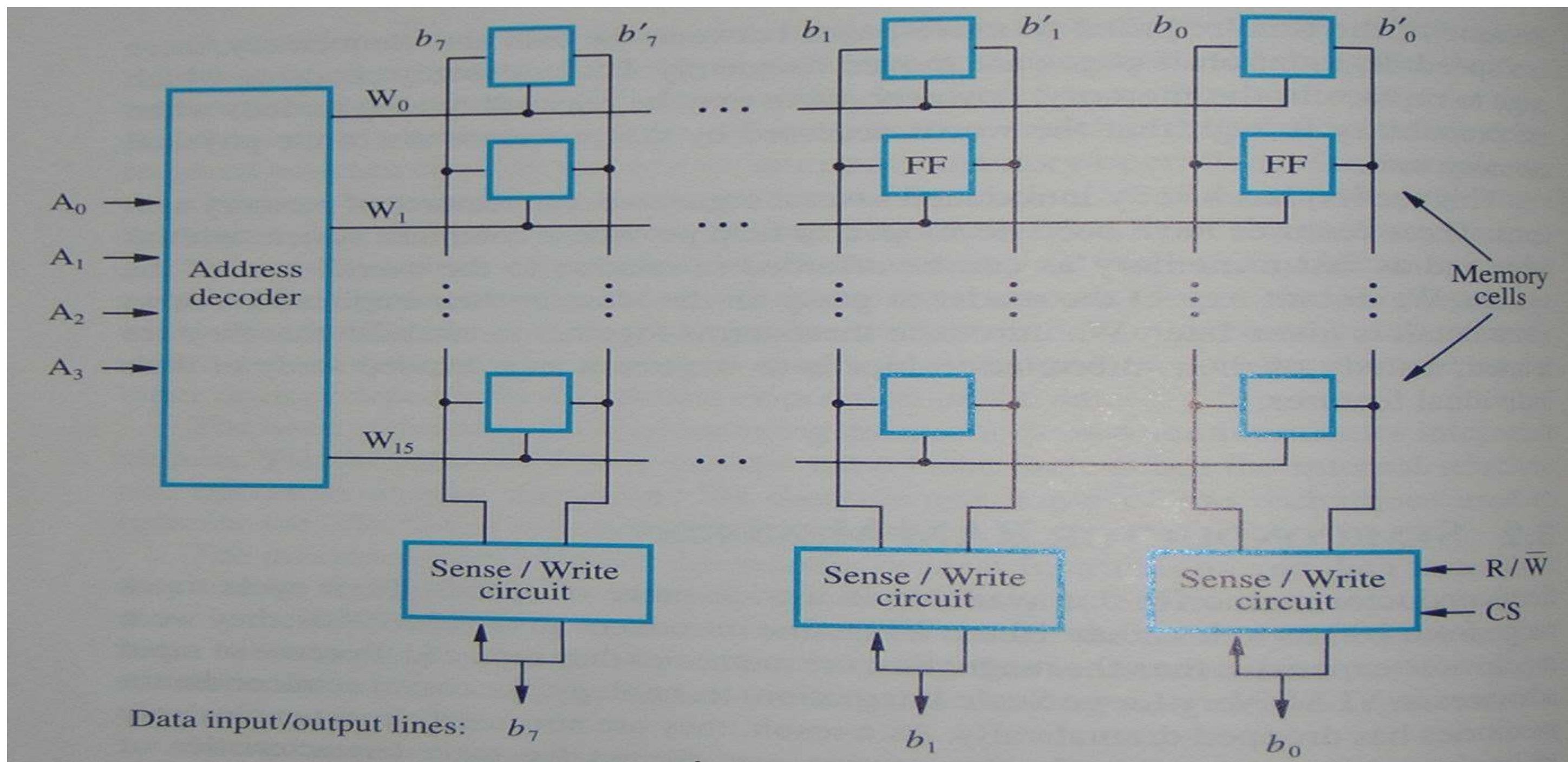


# INTERNAL ORGANIZATION OF MEMORY CHIPS





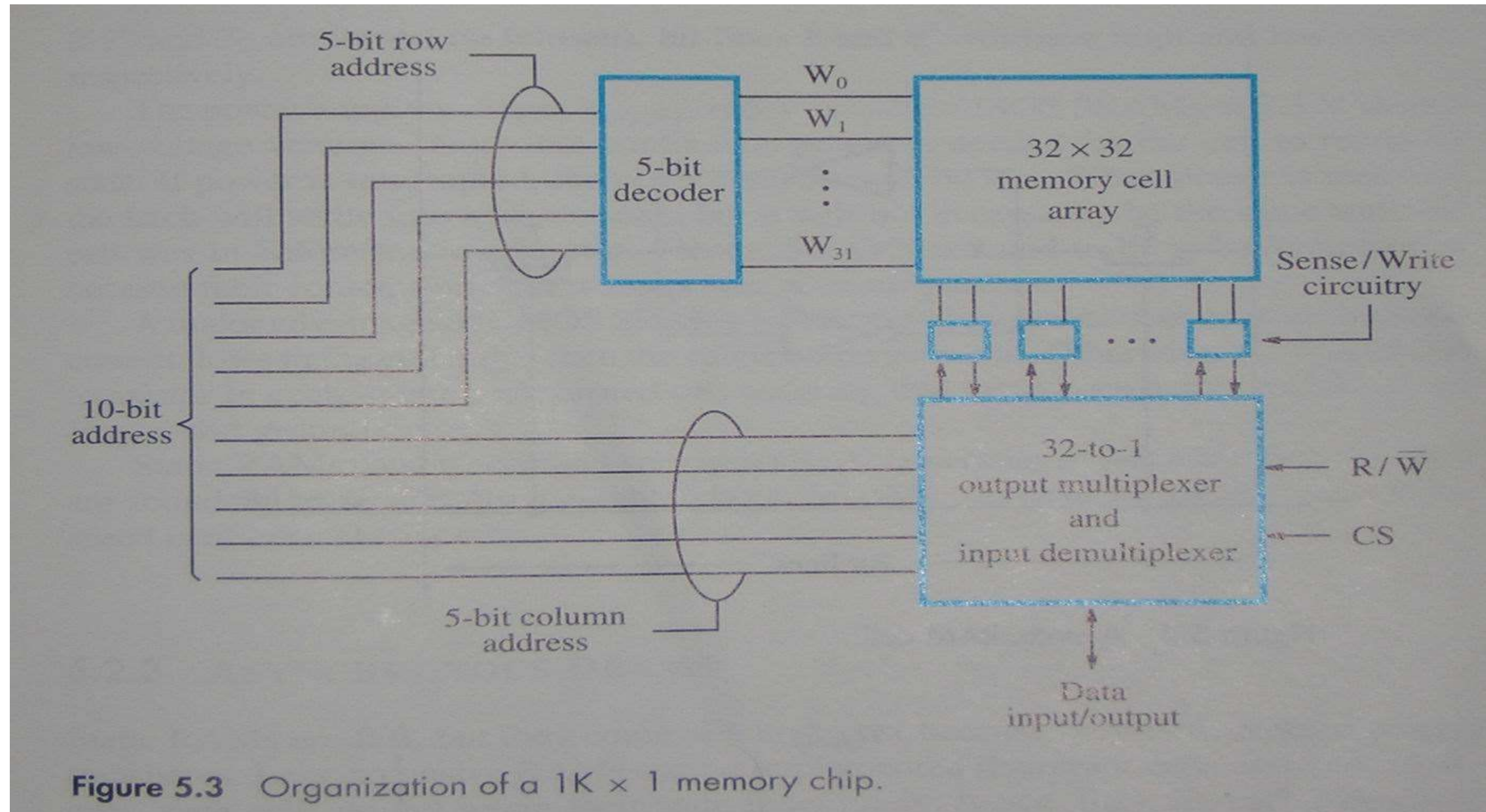
# Organization of bit cells in a memory chip [16\* 8]



From Figure 5.2 Page 296 of "Computer Organization", Carl Hamacher, 5<sup>th</sup> edition, McGraw Hill pub.



# Organization of a 1Kx1



From Figure 5.3 Page 297 of "Computer Organization", Carl Hamacher, 5<sup>th</sup> edition, McGraw Hill pub.



# Static memories

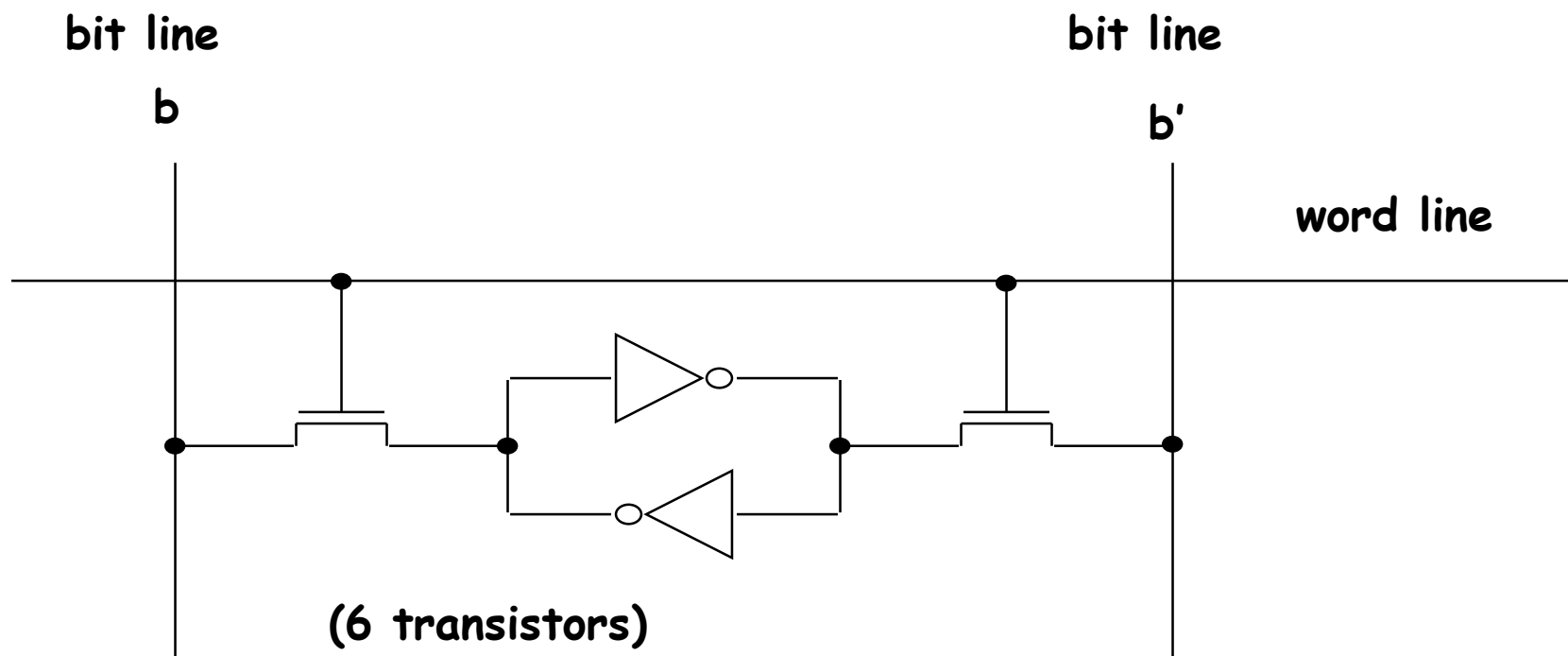


Memories that consist of circuits capable of retaining their state as long as power is applied are known as static memories.

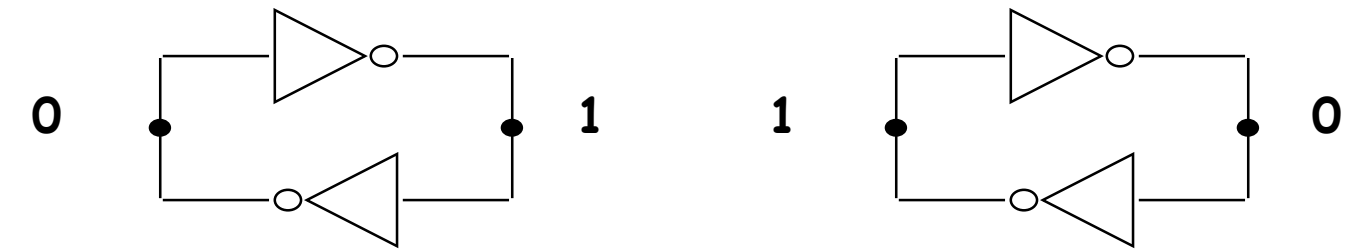
MR

MW

# Anatomy of an SRAM Cell



## Stable Configurations



### Terminology:

*bit line:*

carries data

*word line:*

used for addressing

## Write:

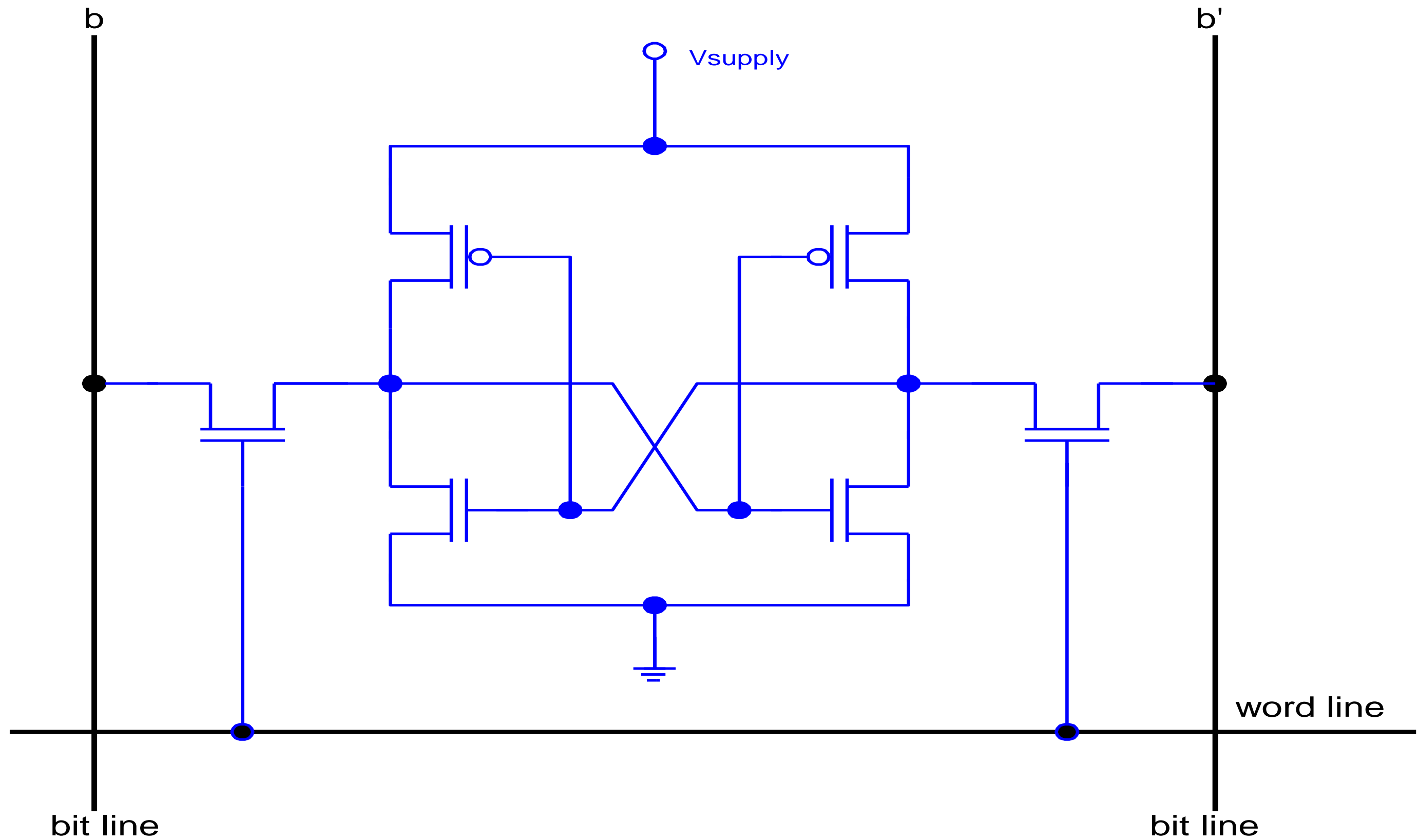
1. set bit lines to new data value
    - **b'** is set to the opposite of **b**
  2. raise word line to "high"
- ⇒ sets cell to new state (may involve flipping relative to old state)

## Read:

1. set bit lines high
2. set word line high
3. see which bit line goes low



# CMOS SRAM cell







# Asynchronous DRAMs



Information is stored in a dynamic memory cell in the form of a charge in a capacitor and cells do not retain their state indefinitely, hence they are called DRAM. The timing of the memory device is controlled by a specialized memory controller which provides the necessary control signals, RAS, CAS that governs the timing. The processor must take into account the delay in the response of the memory (asynchronous).



MR operation (memory refreshing)

MW operation

High density

Low cost

Low speed

Fast page mode:

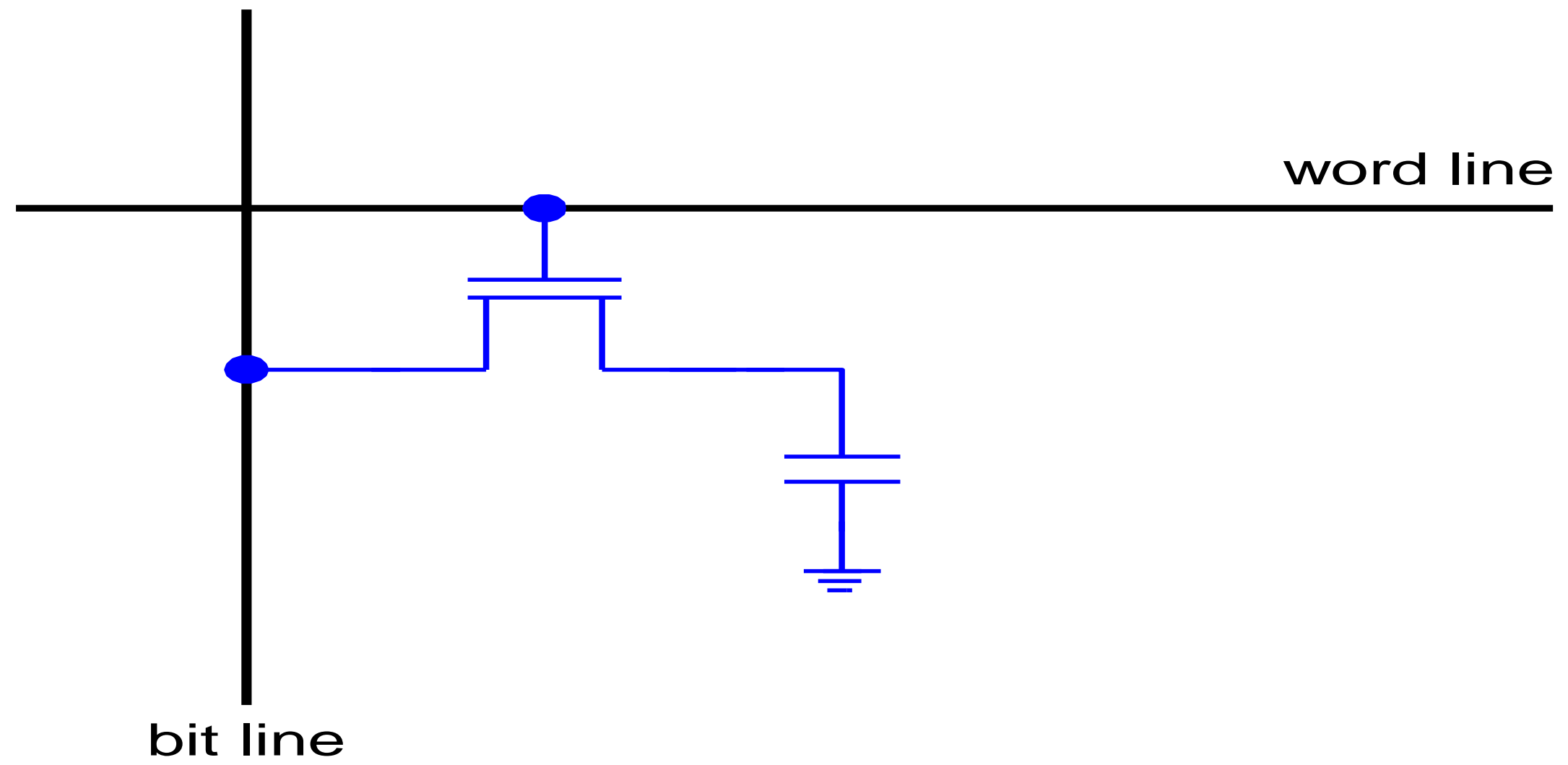
The most useful arrangement is to transfer the bytes in sequential order, which is achieved by applying a consecutive sequence of column addresses under the control of successive CAS signals.

This scheme allows transferring a block of data at a much faster rate than can be achieved for transfers involving random addresses.



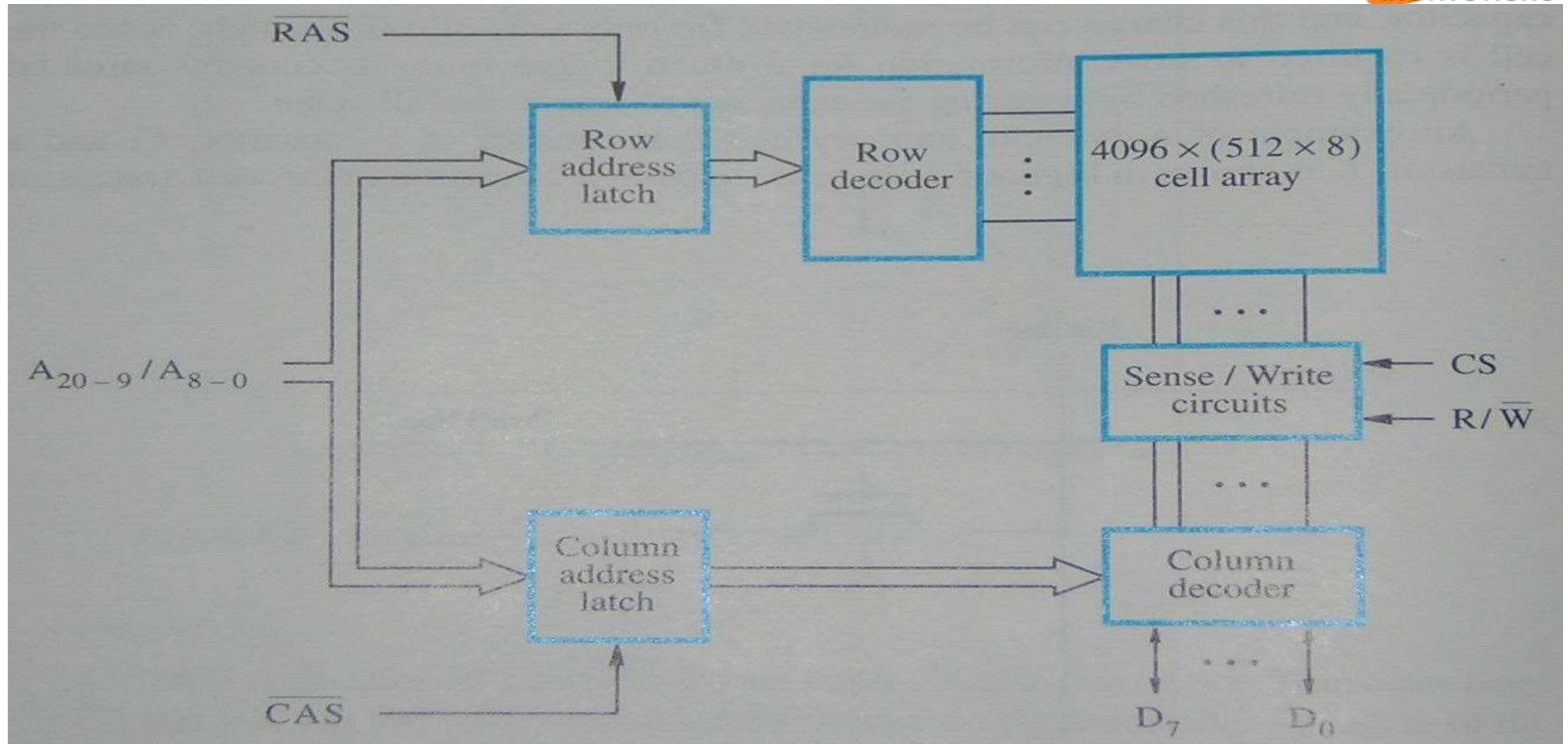


# DRAM cell





# Internal organization of a 2Mx8 DRAM



From Figure 5.7 Page 300 of "Computer Organization", Carl Hamacher, 5<sup>th</sup> edition, McGraw Hill pub.



# DRAM: Multiplexed Row-Column addressing



Reducing Address pins of IC chip

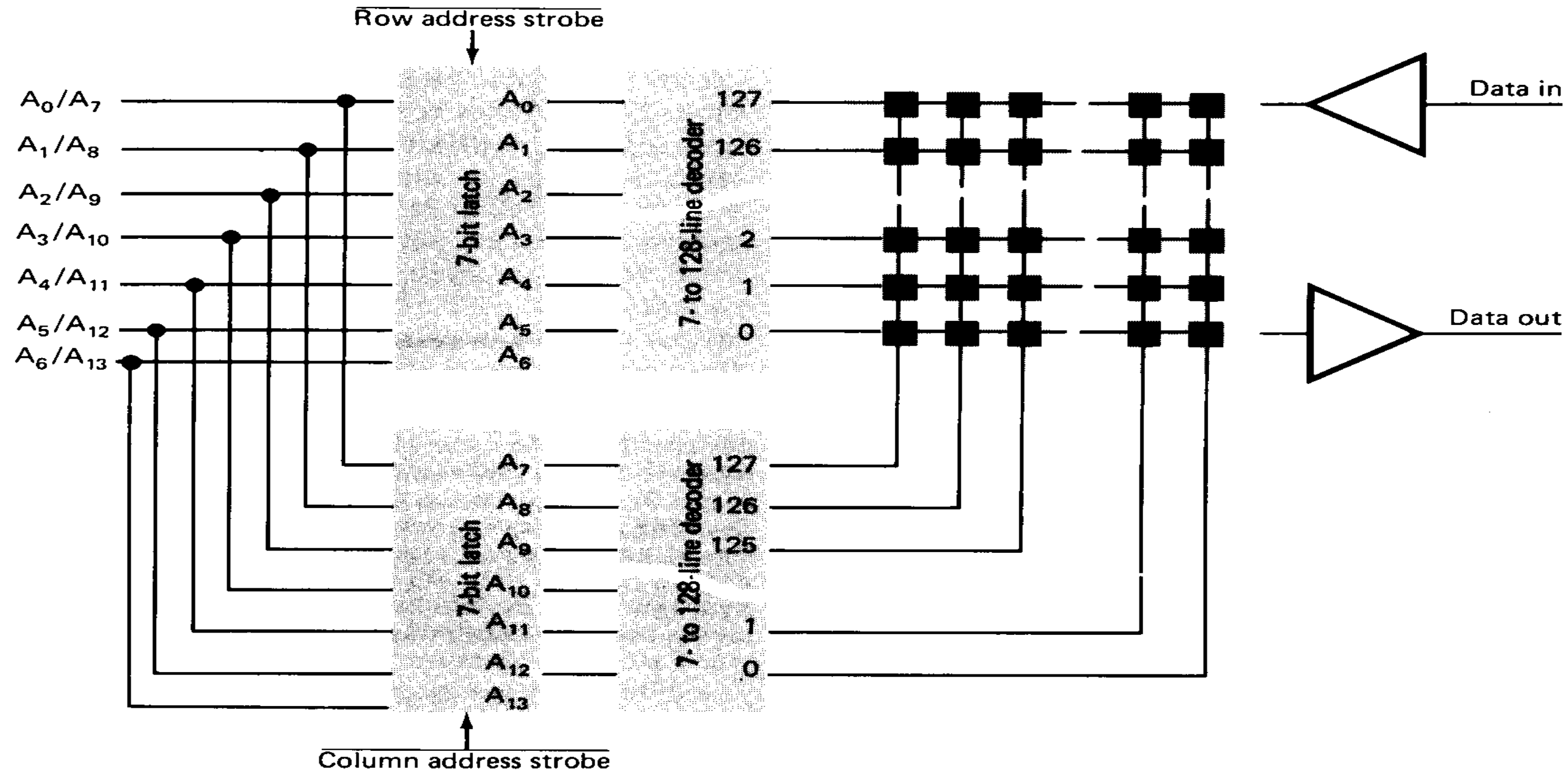
RAS = Row Address Strobe

CAS = Column Address Strobe





# DRAM: Multiplexed Row-Column addressing





# SDRAM



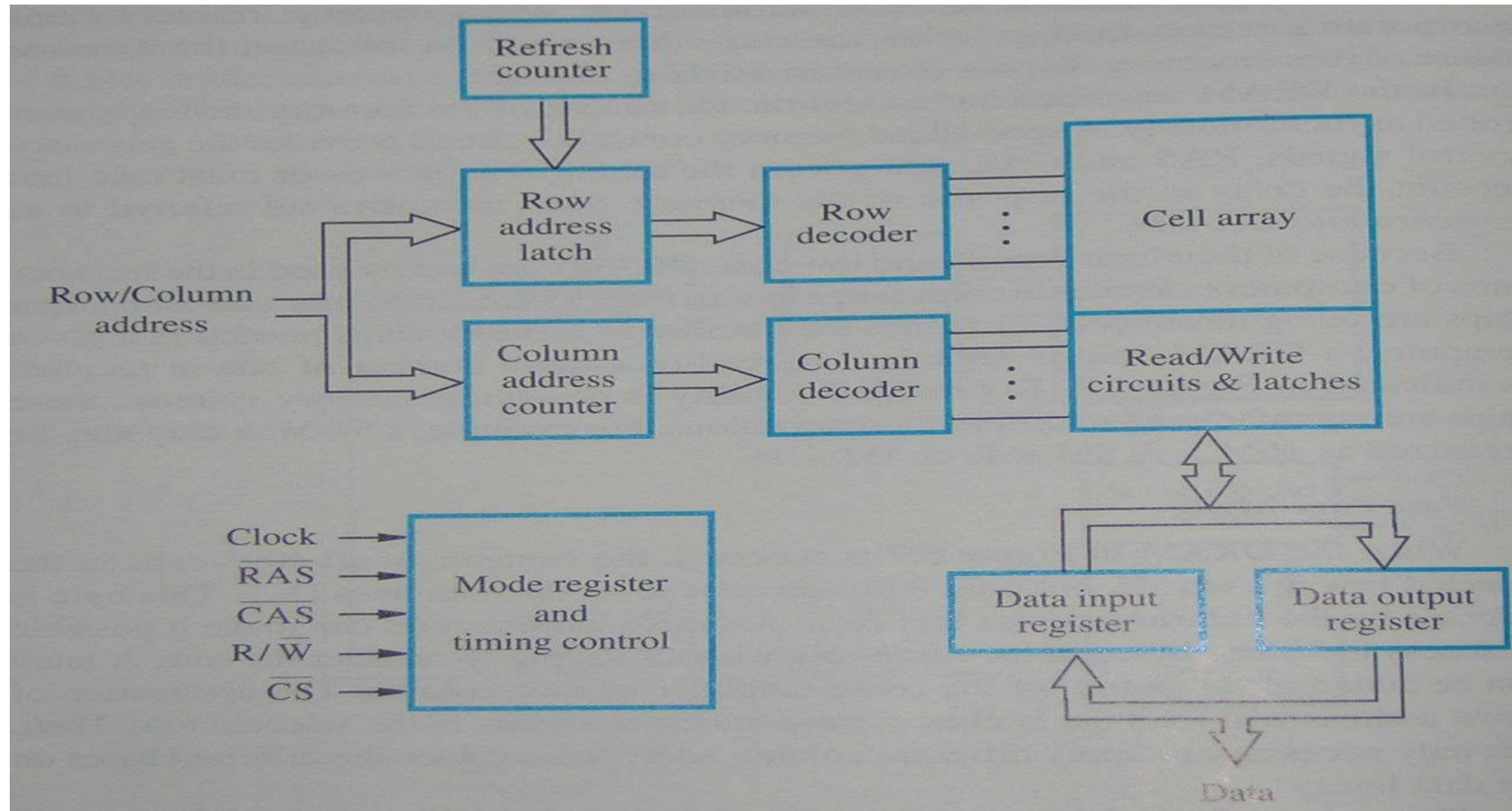
Synchronous DRAM

Need clock signal for synchronize operation

Can be used with clock speed 100 and 133 MHz

Built in refresh circuitry

# Structure of Synchronous DRAM

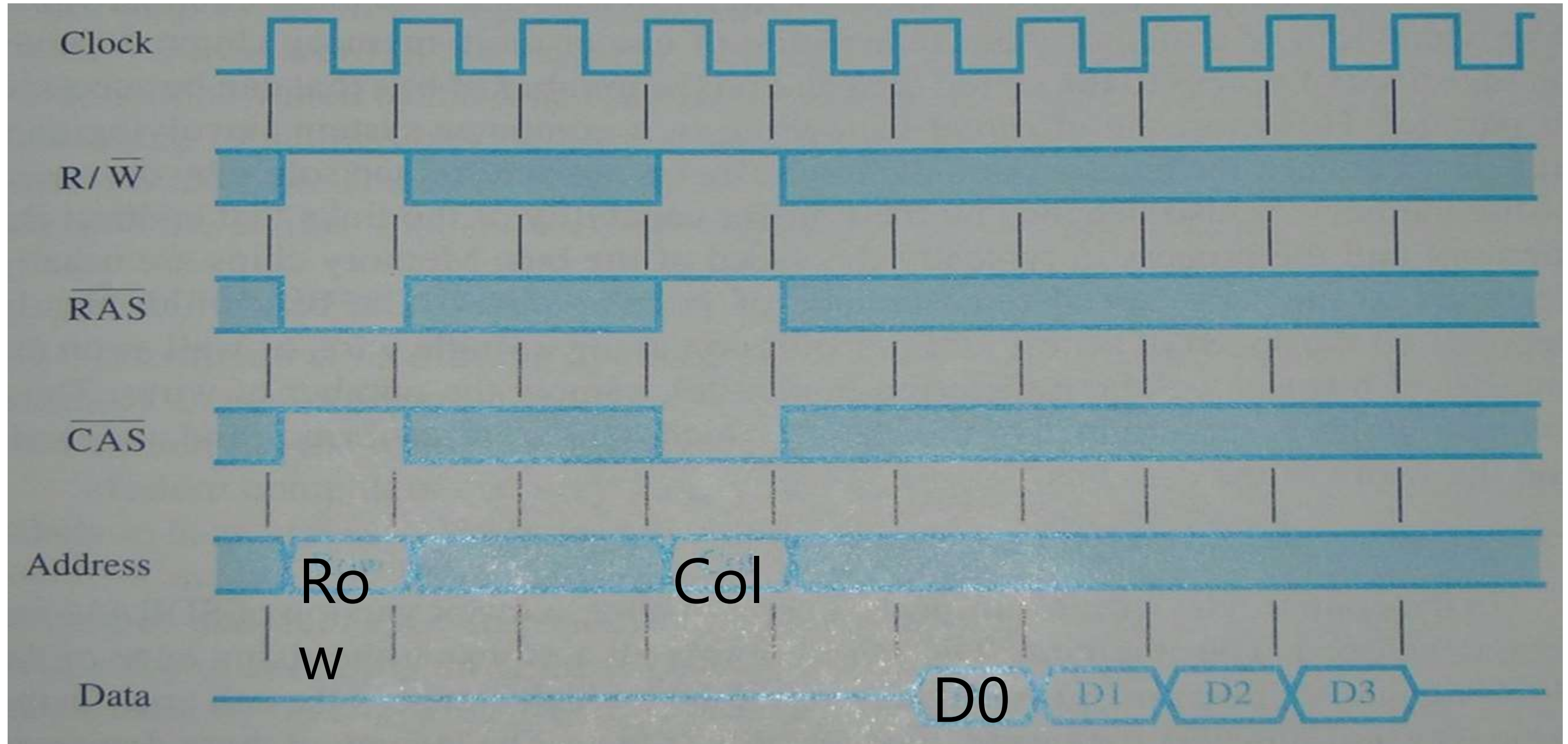


From Figure 5.8 Page 302 of "Computer Organization", Carl Hamacher, 5<sup>th</sup> edition, McGraw Hill pub.





# Burst read of length 4 in an SDRAM



From Figure 5.9 Page 303 of "Computer Organization", Carl Hamacher, 5<sup>th</sup> edition, McGraw Hill pub.





A good indication of the performance is given by two parameters:

Latency

Bandwidth

Latency: The amount of time it takes to transfer a word of data to or from the memory.

Latency=number of clock cycles required for  
data transfer/clock rate

$$5/100\text{Mhz} = 50\text{ns}$$



Bandwidth: it depends on the speed of access and transmission along a wire or wires.

It is the product of the rate at which data are transferred and the width of the data bus.

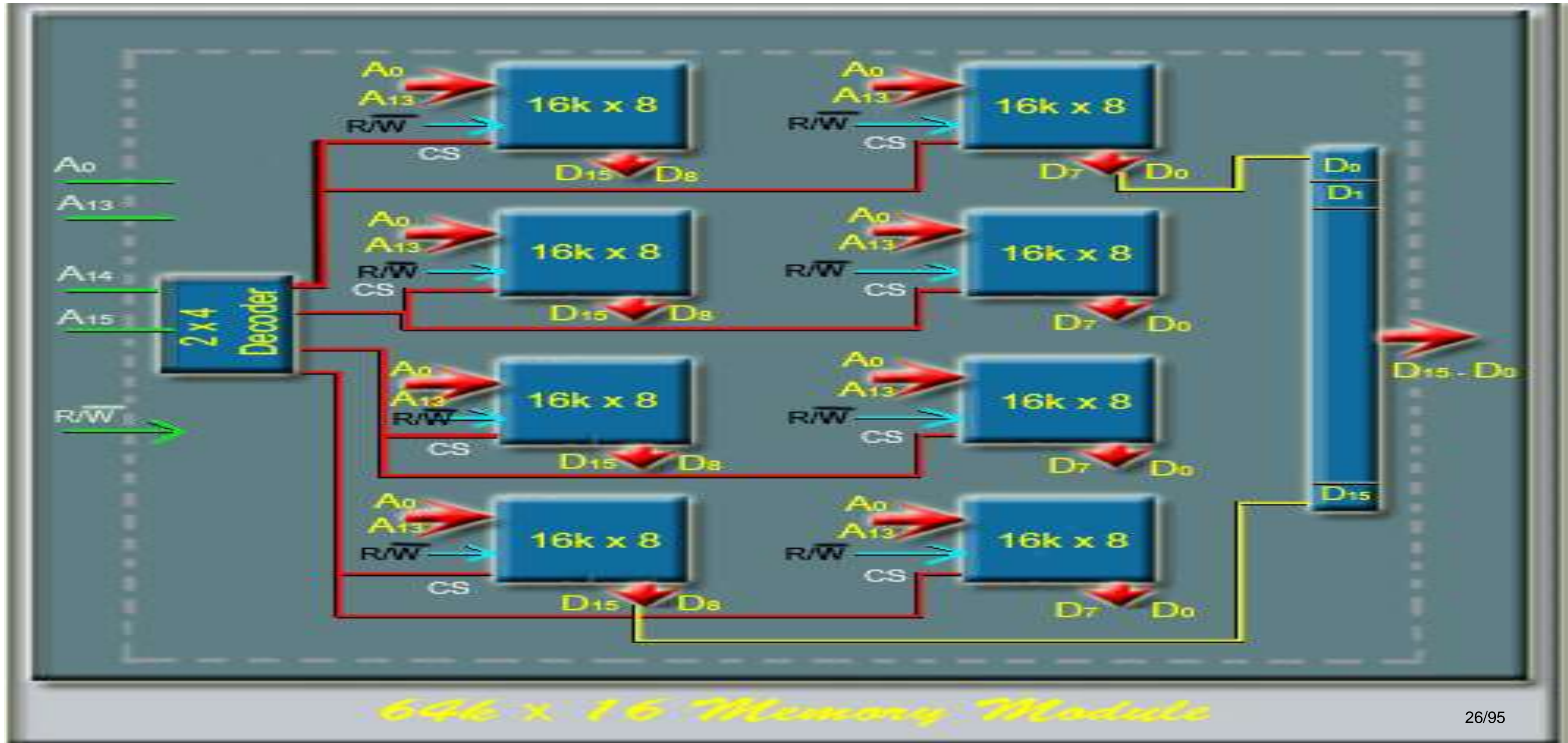
Data rate: Number of bits that can be transferred in one second.



Double Data Rate SDRAM(DDR SDRAM):

The standard. SDRAM perform all actions on the rising edge of the clock signal. Here data transfer on both edges of the clock (same latency and bandwidth doubled).

# Structure of larger memories







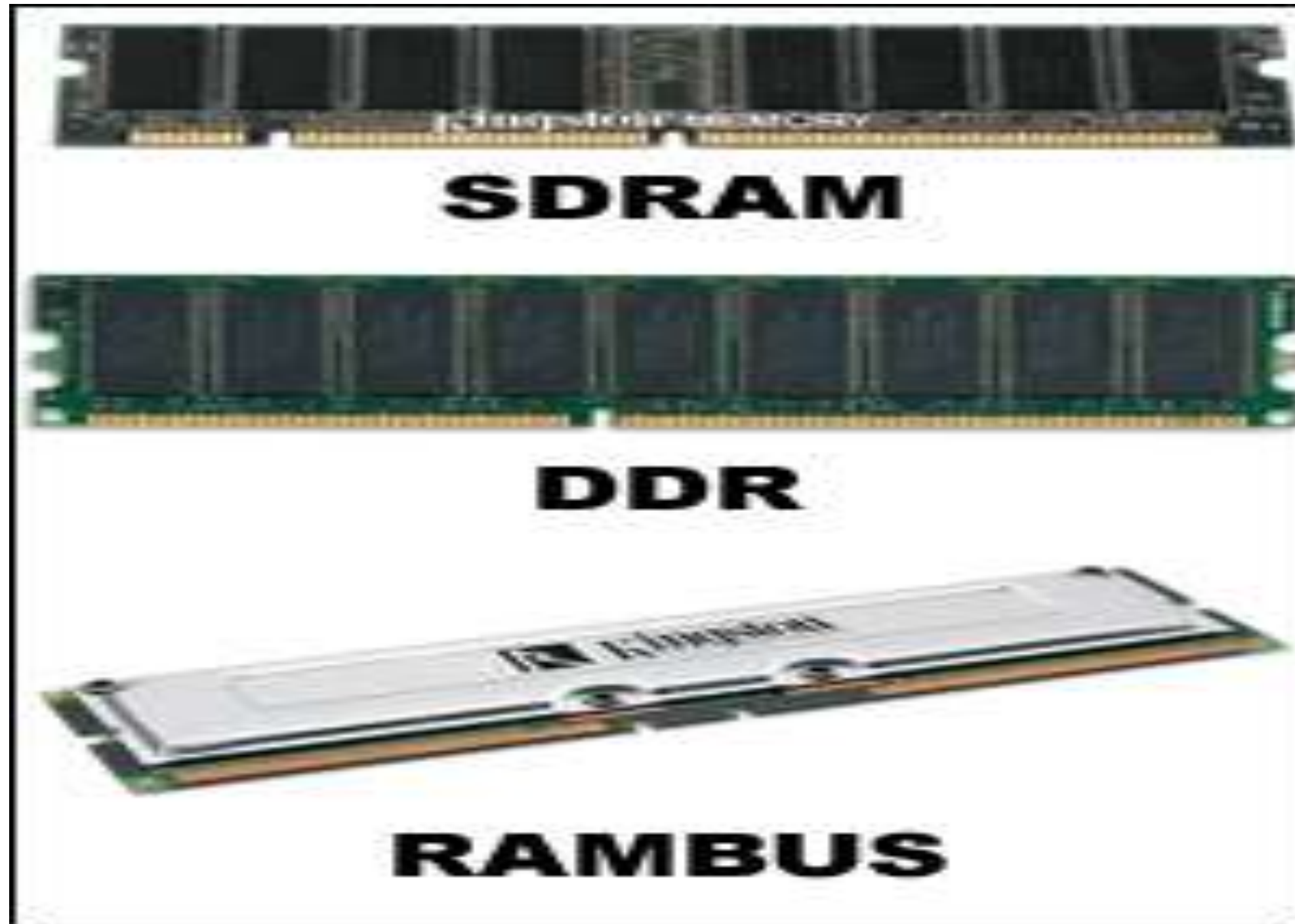
# Memory Module



From [www.oamao.com/Matos/ordi/guide.htm](http://www.oamao.com/Matos/ordi/guide.htm)



# 3 Types of RAM modules

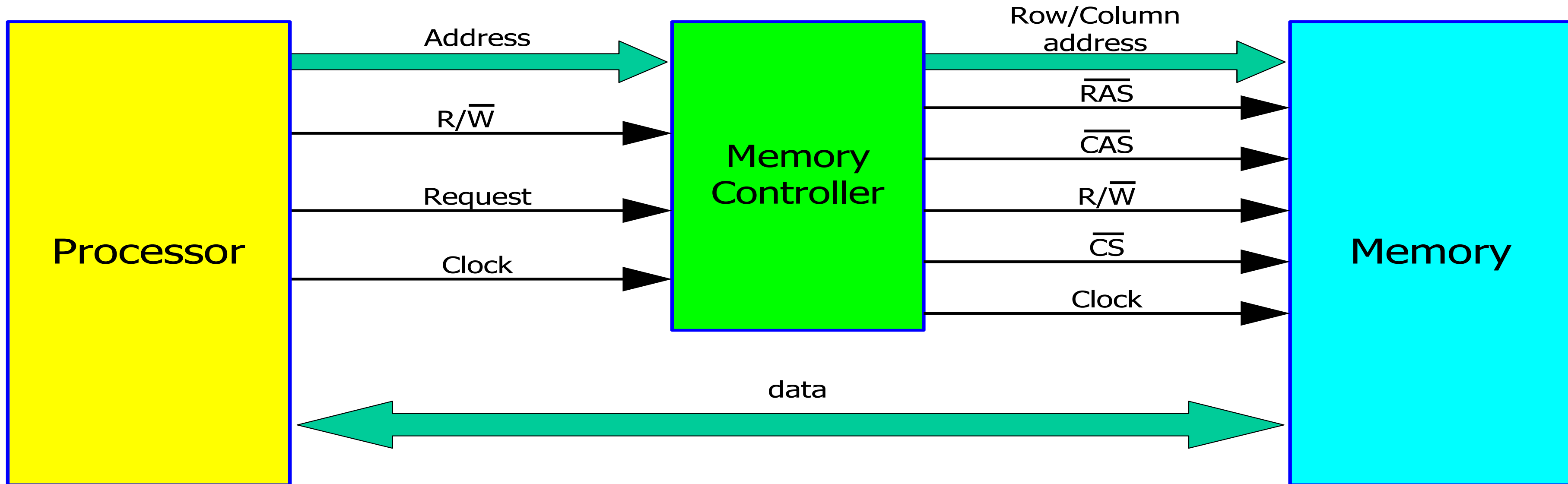




# Memory system considerations



# The use of Memory controller





# Rambus memory



The key feature of the rambus technology is a fast signaling method used to transfer information between chips.

Differential signaling: the two logic values are represented by .3v swings above and below  $V_{ref}=2v$



# Semiconductor Memories



## Nonvolatile memory

ROM

PROM

EPROM

EEPROM

Flash memory

## Volatile memory

SRAM

DRAM

Asynchronous

DRAM

FPM DRAM

Synchronous

SDRAM

DDR SDRAM

RDRAM



# ROM



ROM : Read Only Memory

Programmed when manufacturing is in process.

PROM : Programmable Read Only Memory

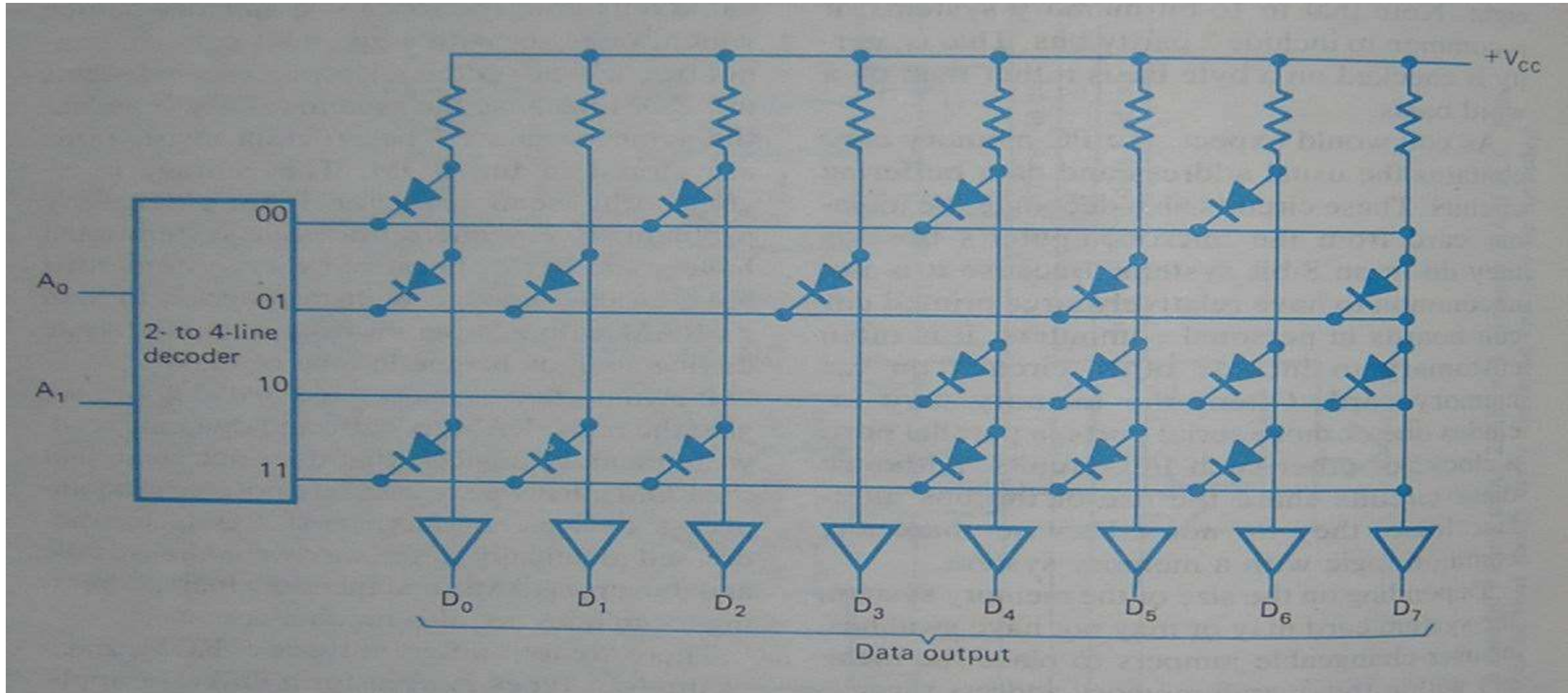
Programmable by user only once

Flexible and convenient compared to ROM

Programmed by burning the fuse using high current pulse



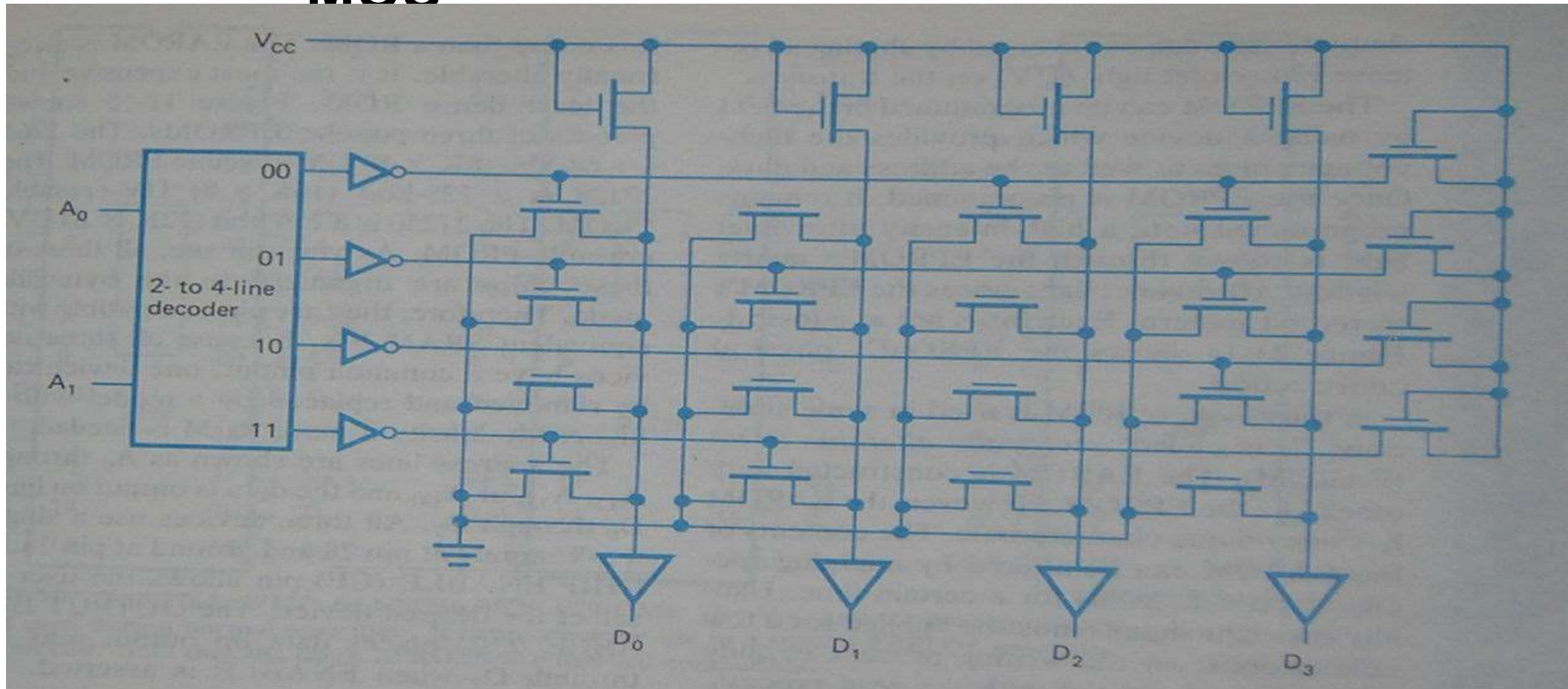
# A simple 4-word ROM



From Figure 11-12 Page 298 of "Microprocessors: principles and applications", Charles M. Gilmore, McGraw Hill pub.



# A simple 4-word ROM using MOS



From Figure 11-13 Page 299 of "Microprocessors: principles and applications", Charles M. Gilmore, McGraw Hill pub.



# EEPROM



• No requirement of physically removed from the circuit for reprogramming

• Use special voltage level to erase data

• Any cell contents can be delete selectively

Electrically Erasable PROM



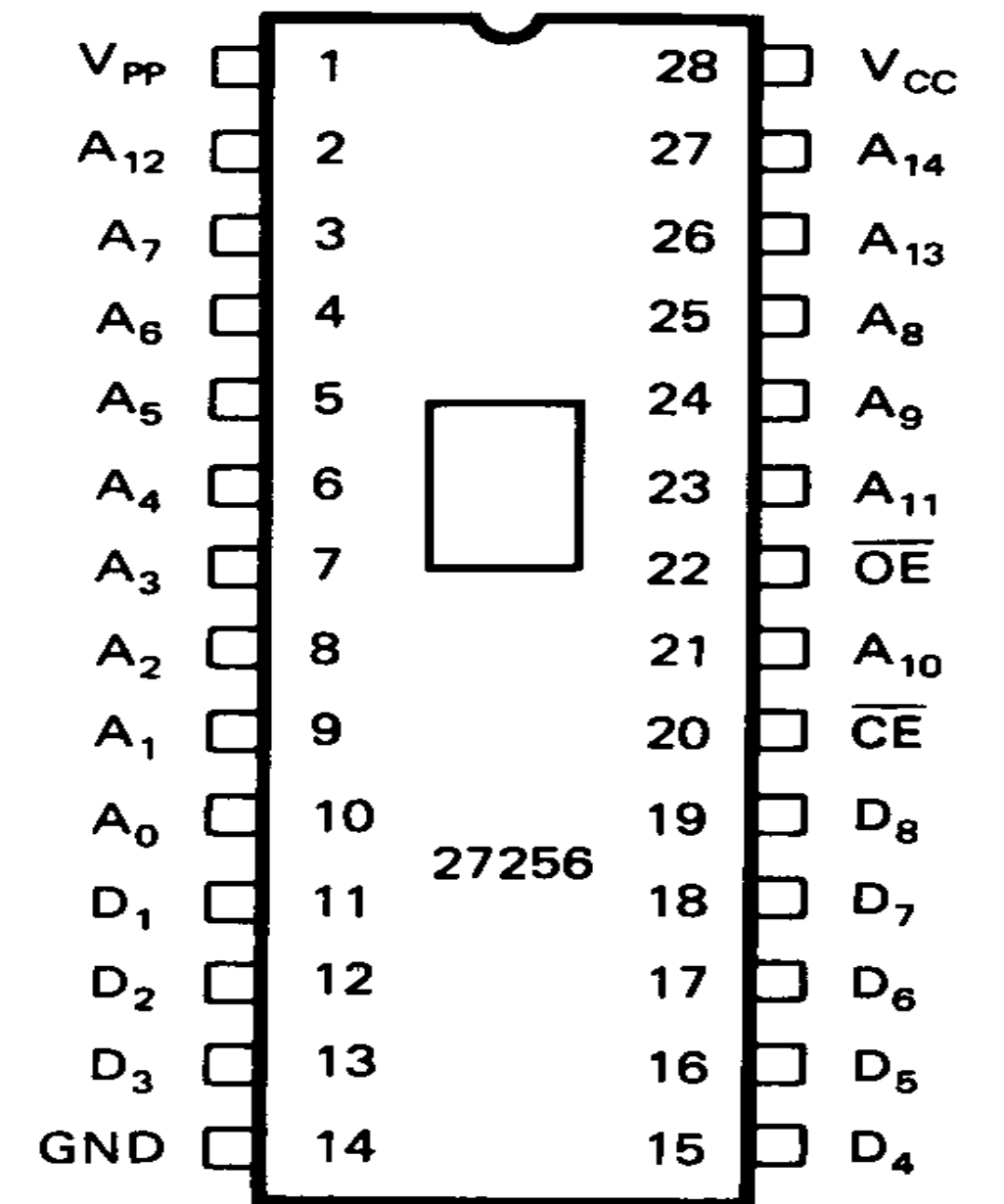
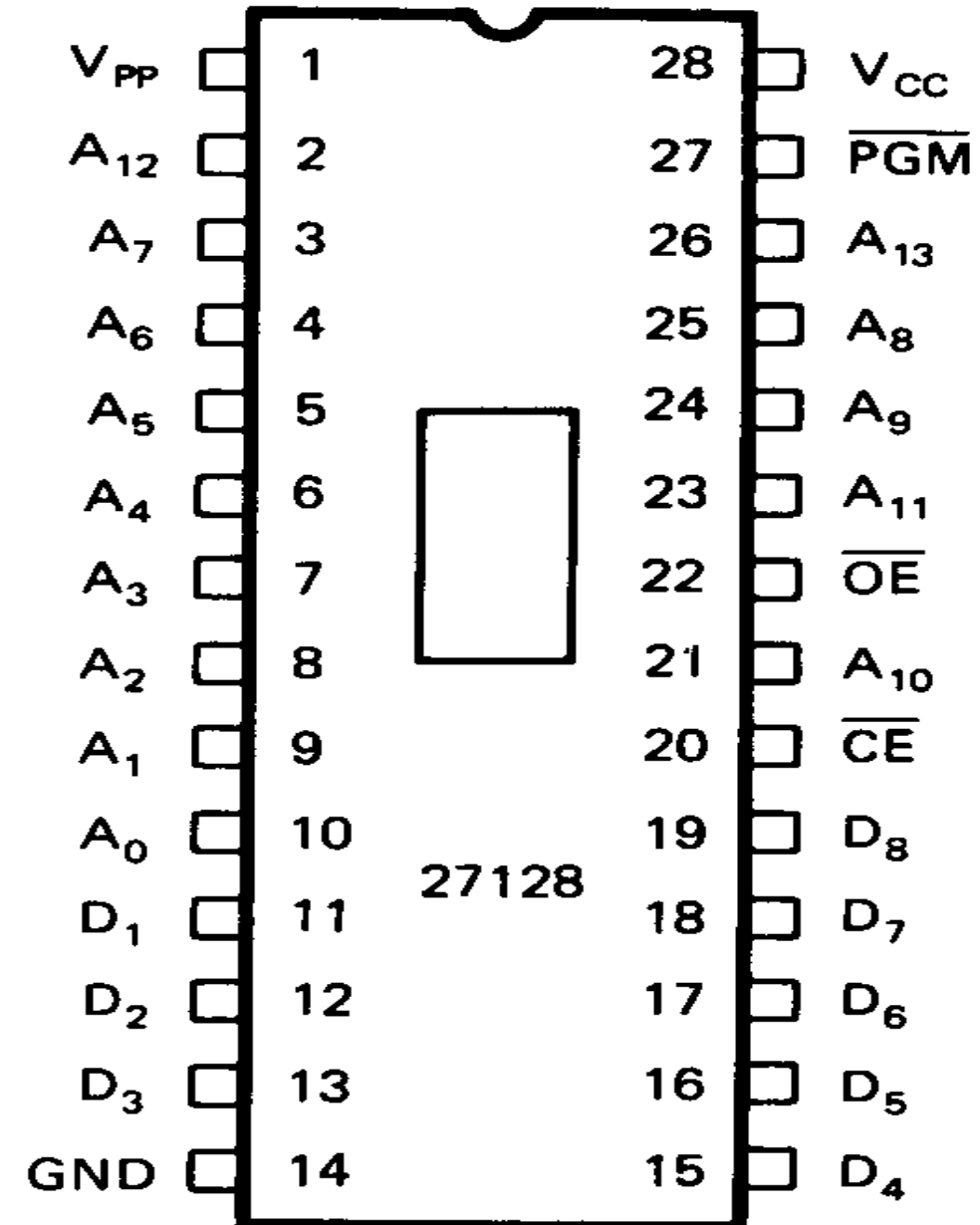
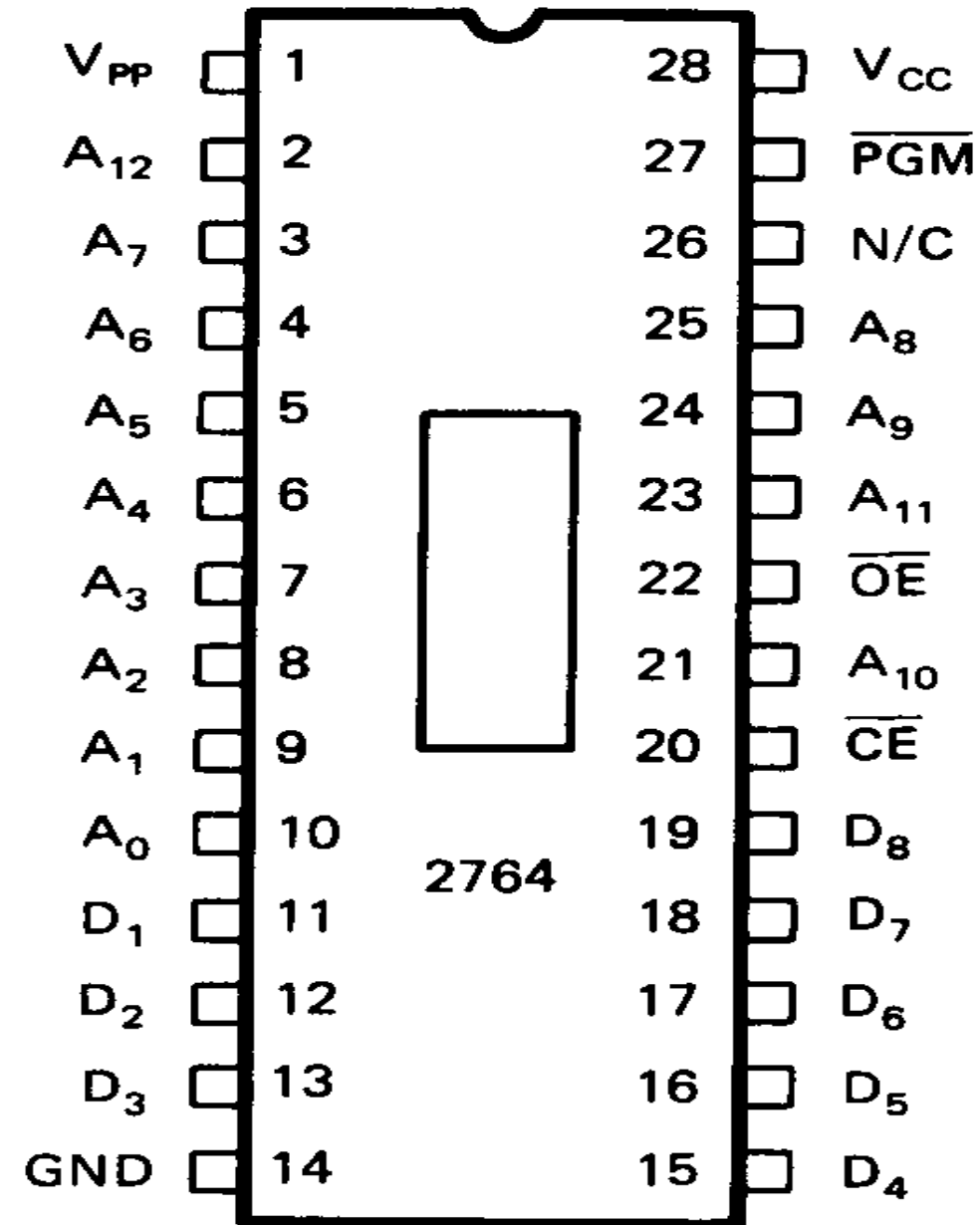


- Reprogrammable
- Erased by UV light
- Example EPROM chips
  - 27C64 : 8KB
  - 27C128 : 16KB
  - 27C256 : 32KB
  - 27C512 : 64KB





# EPROM 2764, 27128, 27256

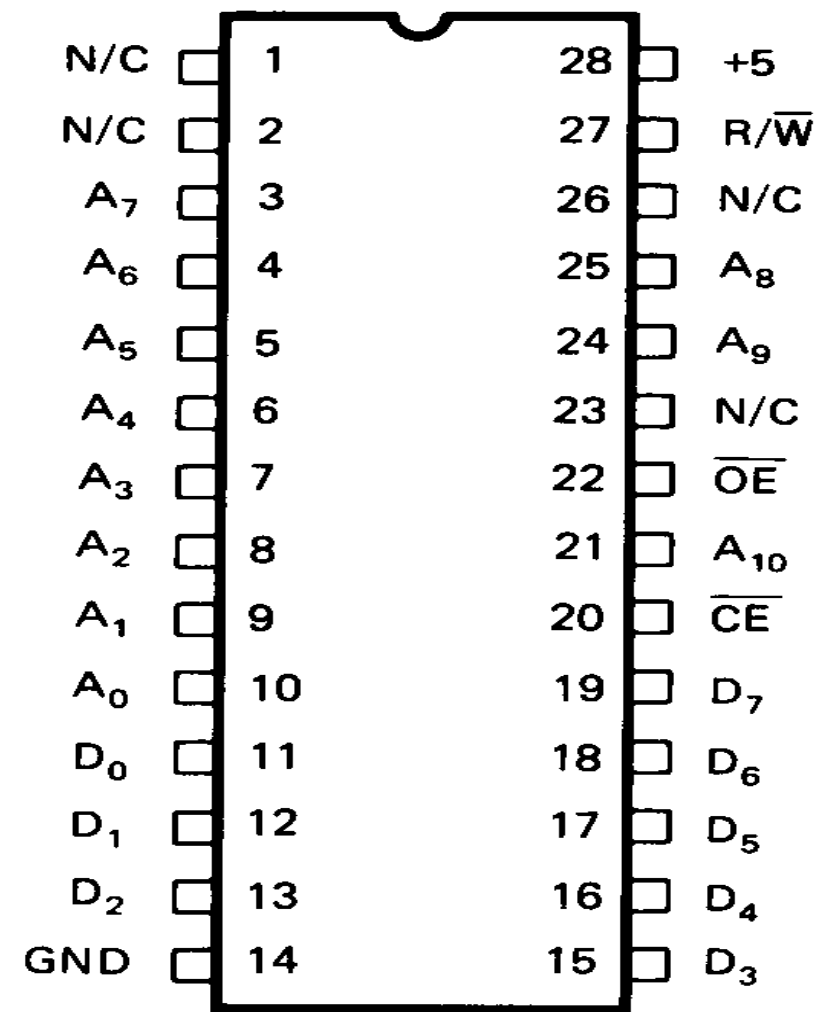




# Static RAM

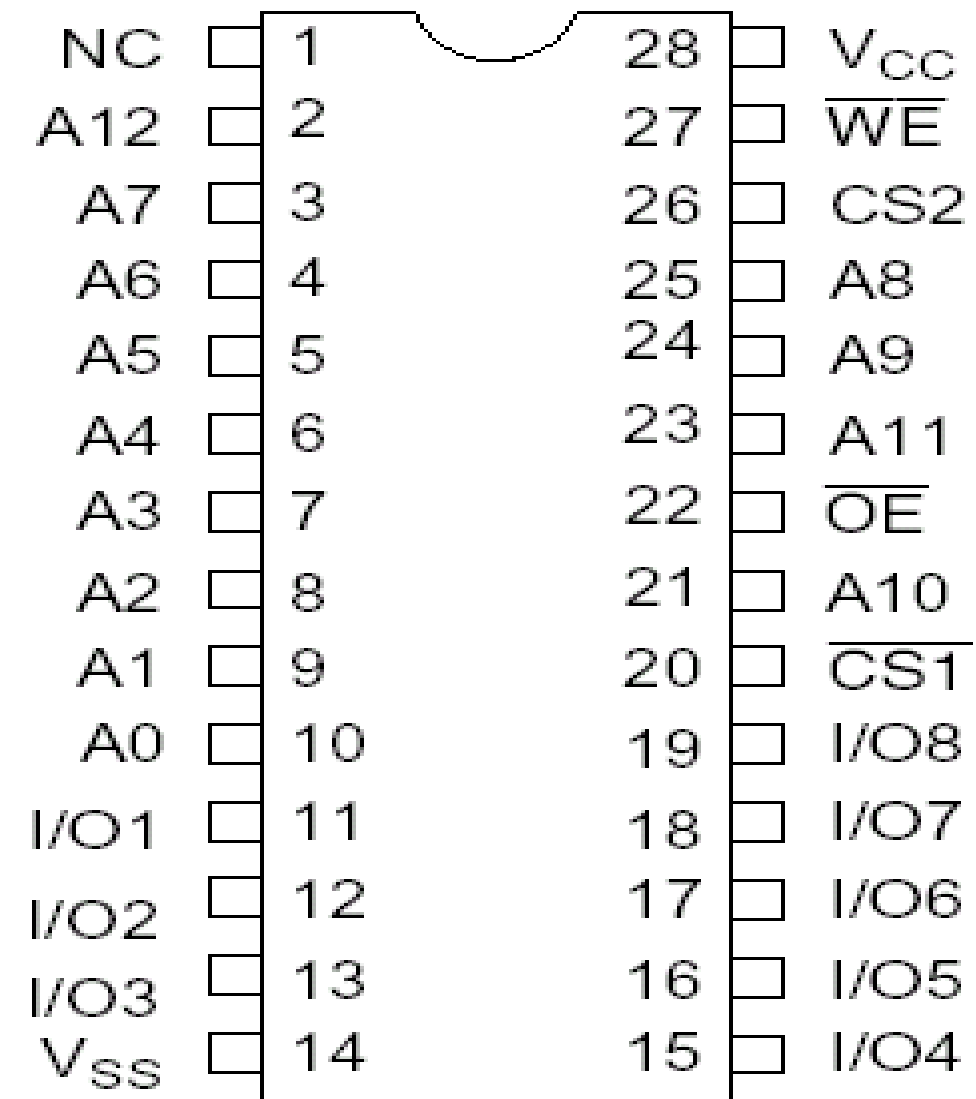


- 2Kx8



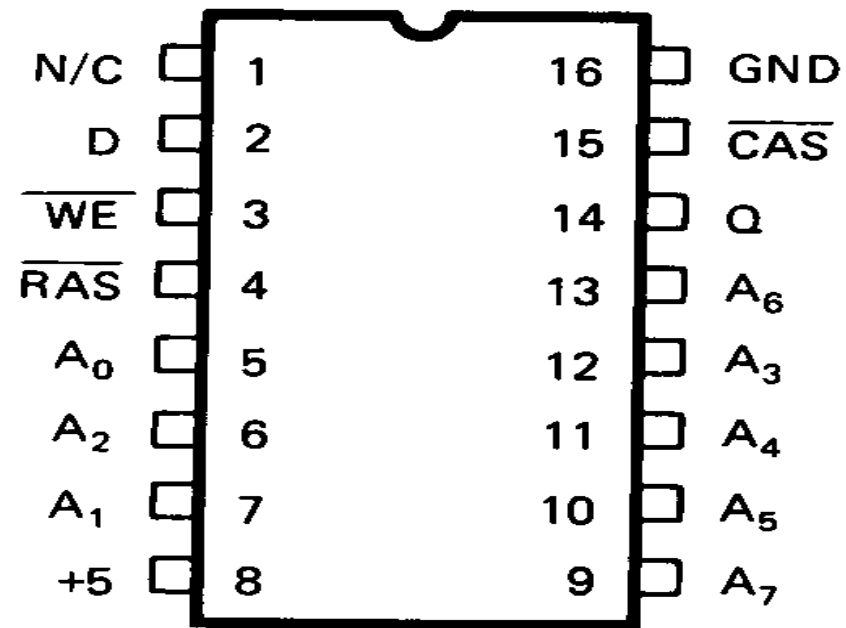
(a)

- 8Kx8

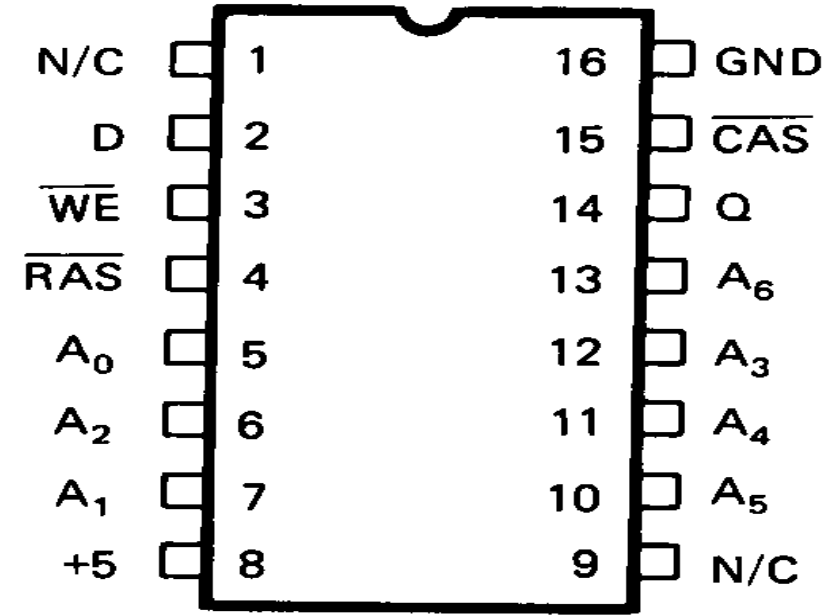




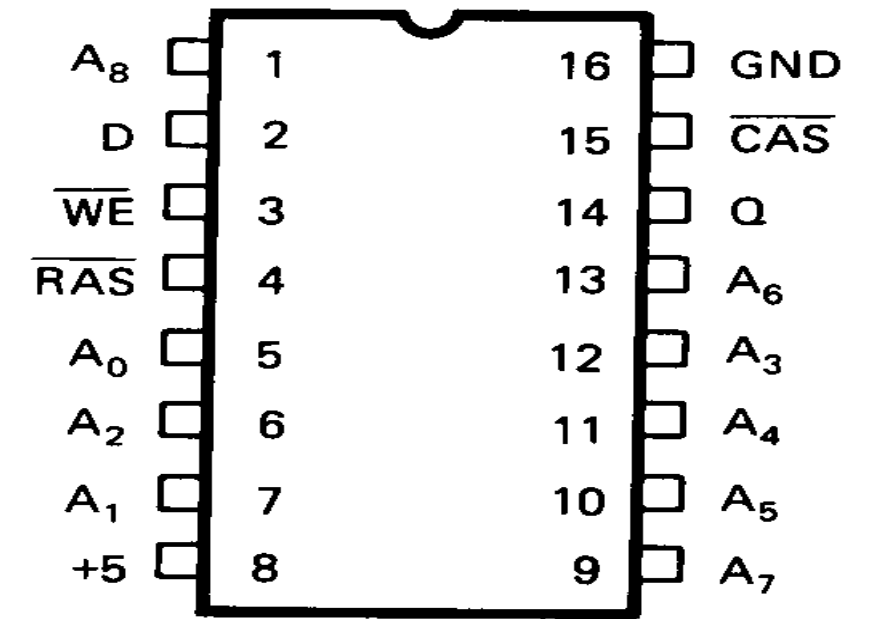
# Dynamic RAM chip: Example



(b)



(a)



(c)





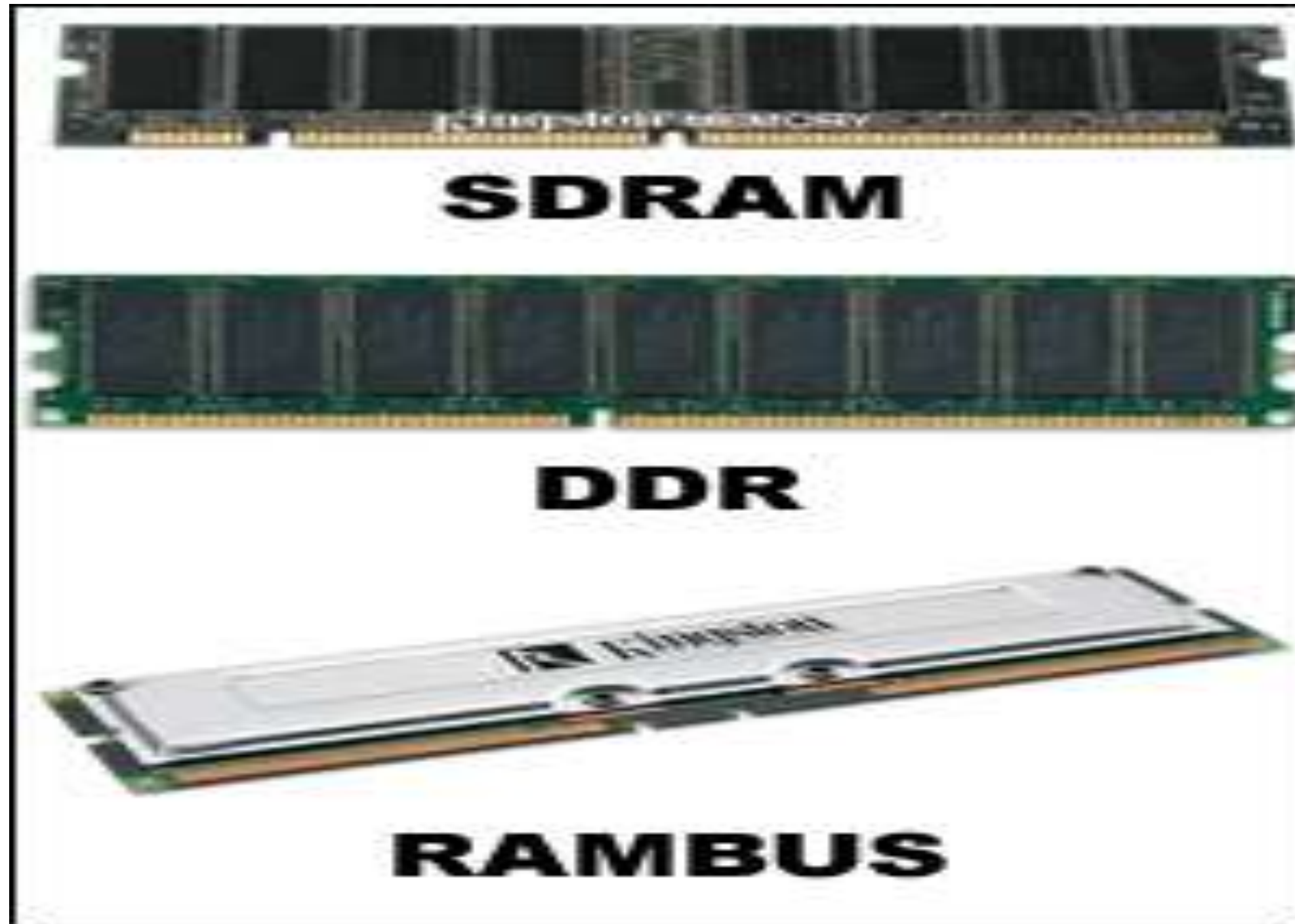
# Memory Module



From [www.oamao.com/Matos/ordi/guide.htm](http://www.oamao.com/Matos/ordi/guide.htm)



# 3 Types of RAM modules





# The role of Memory controller



Is the North-bridge chip in typical PC

Activate/Deactivate signal RAS and CAS timing for DRAM

Interposed between Processor and Memory

Refresh DRAM if required



# The role of Memory controller



Is the North-bridge chip in typical PC  
Activate/Deactivate signal RAS and CAS timing for DRAM  
Interposed between Processor and Memory  
Refresh DRAM if required





# 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



# Flash Memory



Electrically erasable

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.

Suitable for used as solid state disk such as Compact Flash, Memory Stick, etc.

Applications: cell phones digital camera,MP3 players



# Flash Memories



## ☐ **Pros**

- Small, light-weight, robust
- Low power consumption
- Fast read access times comparable to those of DRAM

## ☐ **Cons**

- Much slower write access times
- No in-place-update: need an erase operation.
  - Erase operations can only be performed in a much larger unit than the write operation.
- Limited lifetime
- Bad blocks

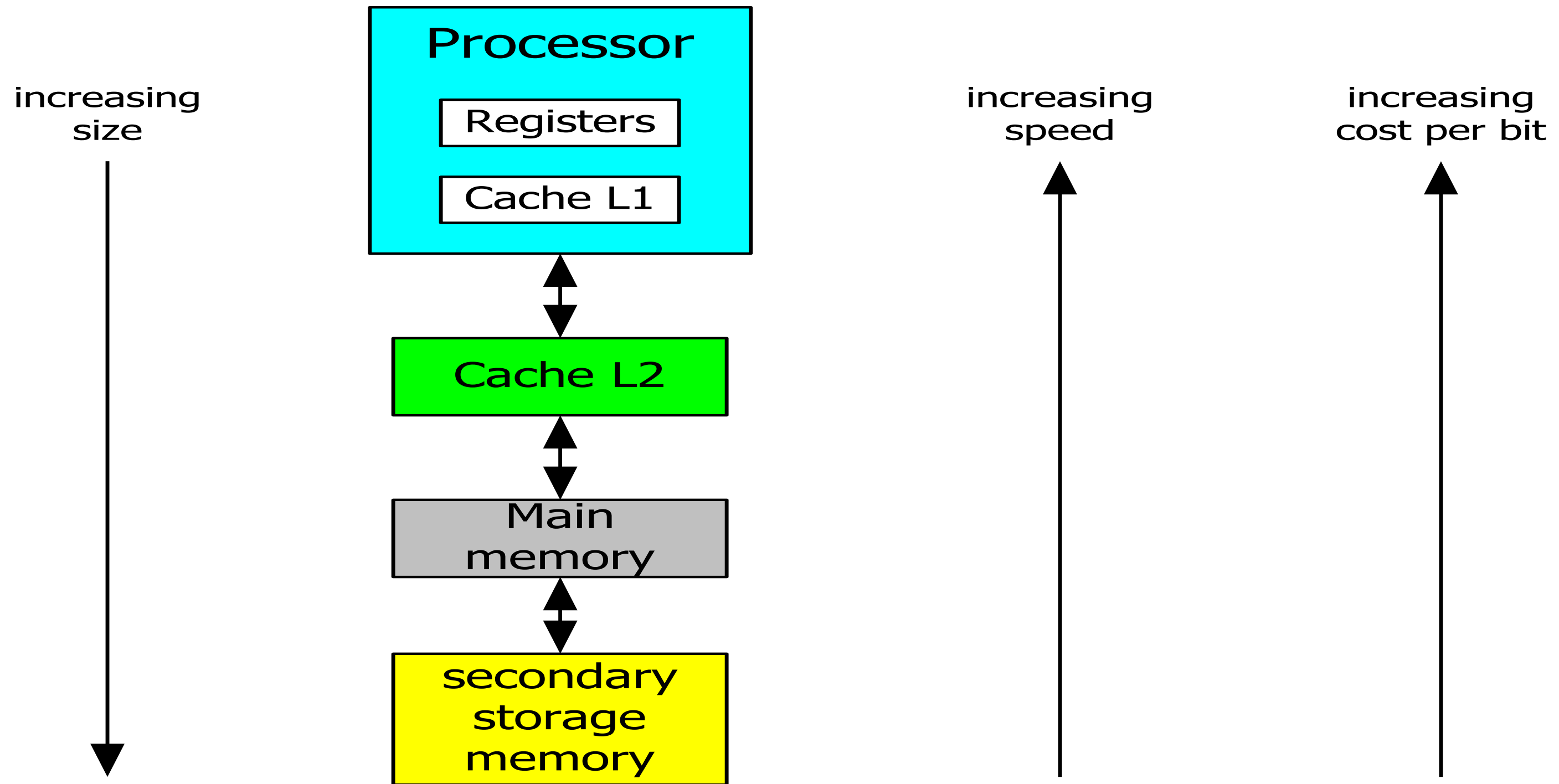


# Speed, size and cost





# Memory hierarchy



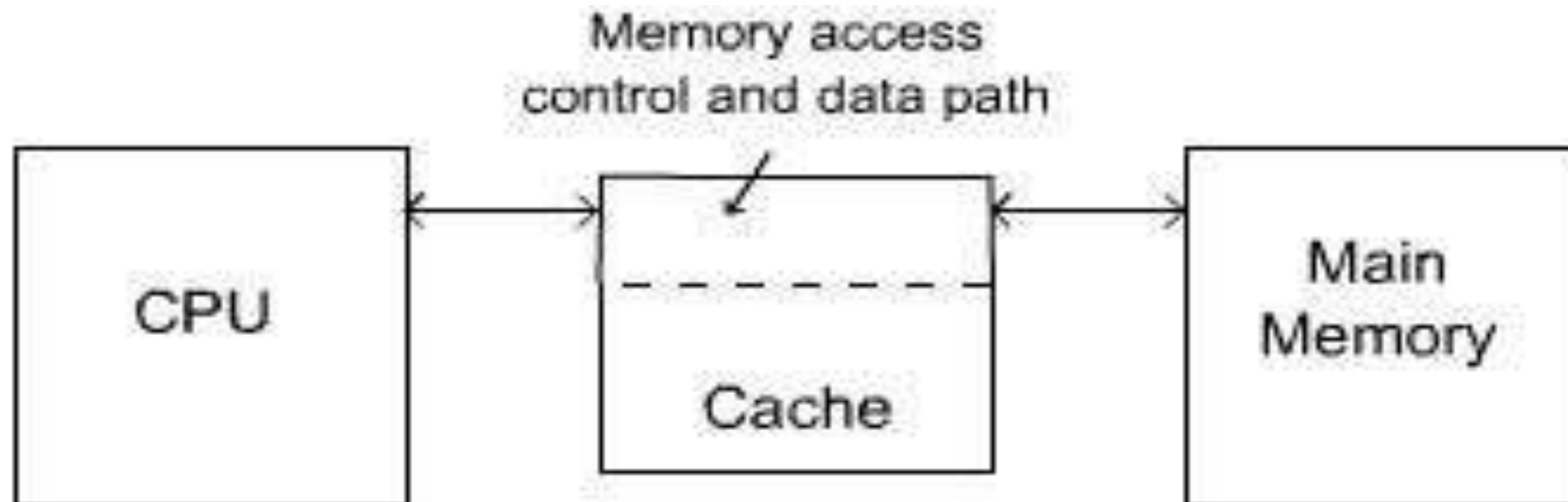


# Cache Memory

Analysis of large number of programs has shown that a number of instructions are executed repeatedly.

This may be in the form of a simple loops, nested loops, or a few procedures that repeatedly call each other.

It is observed that many instructions in each of a few localized areas of the program are repeatedly executed, while the remainder of the program is accessed relatively less. This phenomenon is referred to as locality of reference.





# CACHE MEMORIES



The correspondence between the main memory blocks and those in the cache is specified by a mapping function.

When the cache is full and a memory word (instruction or data) that is not in the cache is referenced, the cache control hardware must decide which block should be removed to create space for the new block that contains the referenced word.

The collection of rules for making this decision constitutes the **replacement algorithm**.



***Read or write hit:*** read or write operation is performed on the appropriate cache location.

***Write-through protocol:*** the cache location and the main memory location are updated simultaneously.

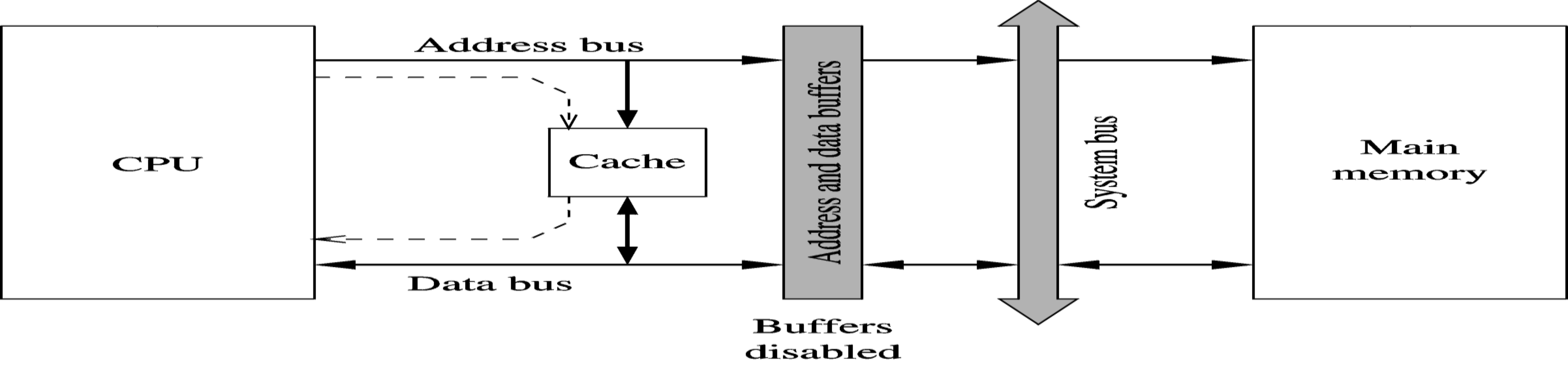
***Write-back or copy-back protocol:***

update cache location and main memory later (in terms of cache block).

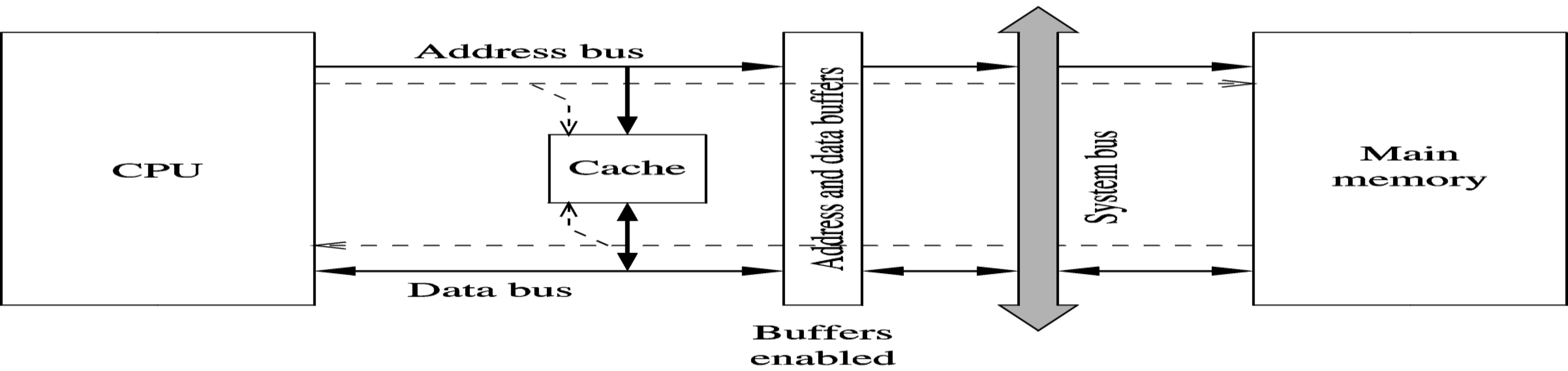
***Read miss:*** when the addressed word in a read operation is not in the cache.

***Write miss:*** during write operation word is not in the cache.

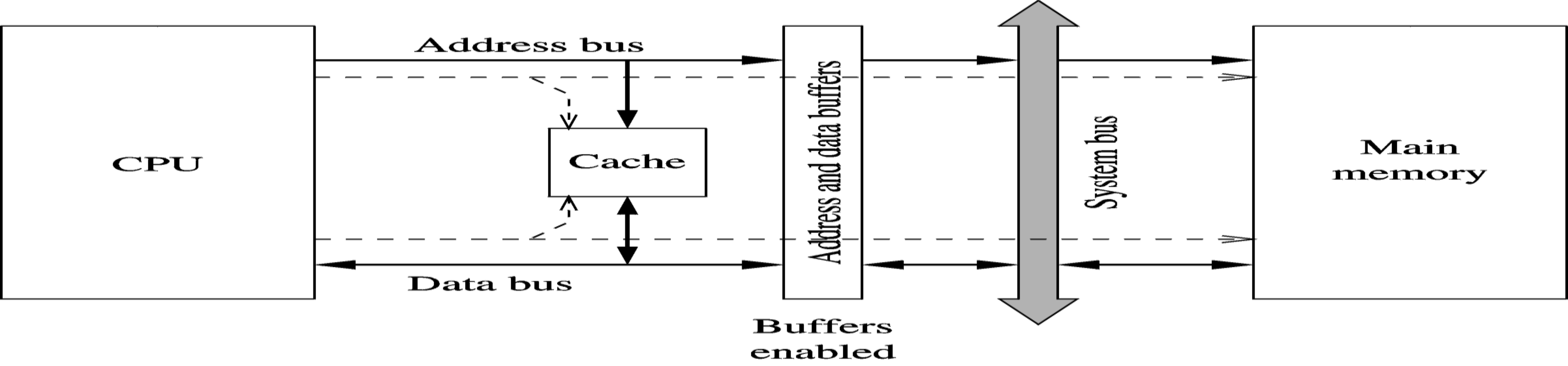




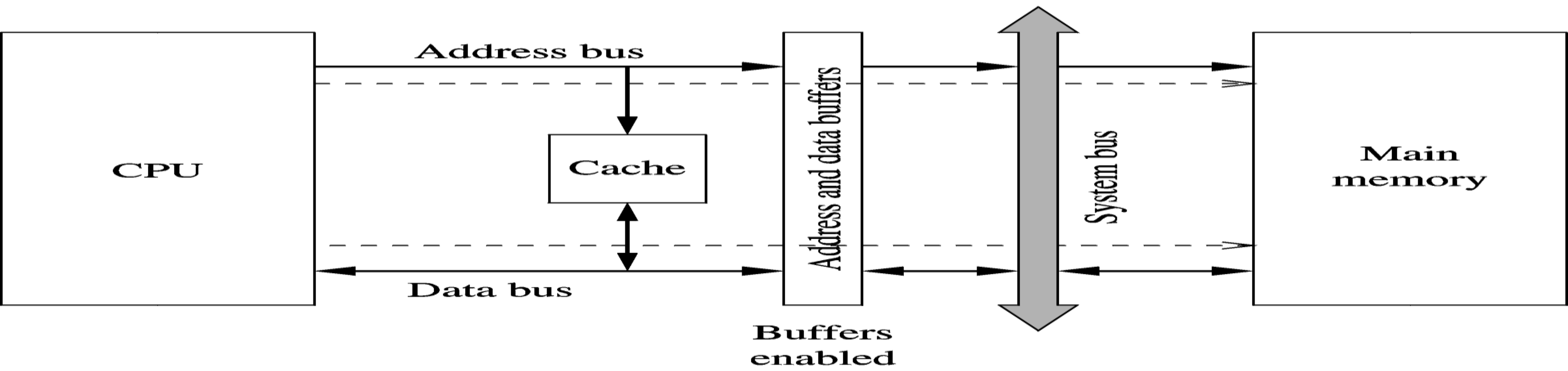
(a) Read hit



(b) Read miss



(a) Write hit



(b) Write miss



# Mapping Function



Determines how memory blocks are mapped to cache lines

Three types

## 1. Direct mapping

Specifies a single cache line for each memory block So, it is not flexible .



## 2. Set-associative mapping

Specifies a set of cache lines for each memory block

## 3. Associative mapping

No restrictions

Any cache line can be used for any memory block



**Mapping Functions :** The mapping functions are used to map a particular block of main memory to a particular block of cache. This mapping function is used to transfer the block from main memory to cache memory. Three different mapping functions are available:

**Direct mapping:** A particular block of main memory can be brought to a particular block of cache memory. So, it is not flexible.

**Associative mapping:** In this mapping function, any block of Main memory can potentially reside in any cache block position. This is much more flexible mapping method.

**Block-set-associative mapping:** In this method, blocks of cache are grouped into sets, and the mapping allows a block of main memory to reside in any block of a specific set. From the flexibility point of view, it is in between to the other two methods.

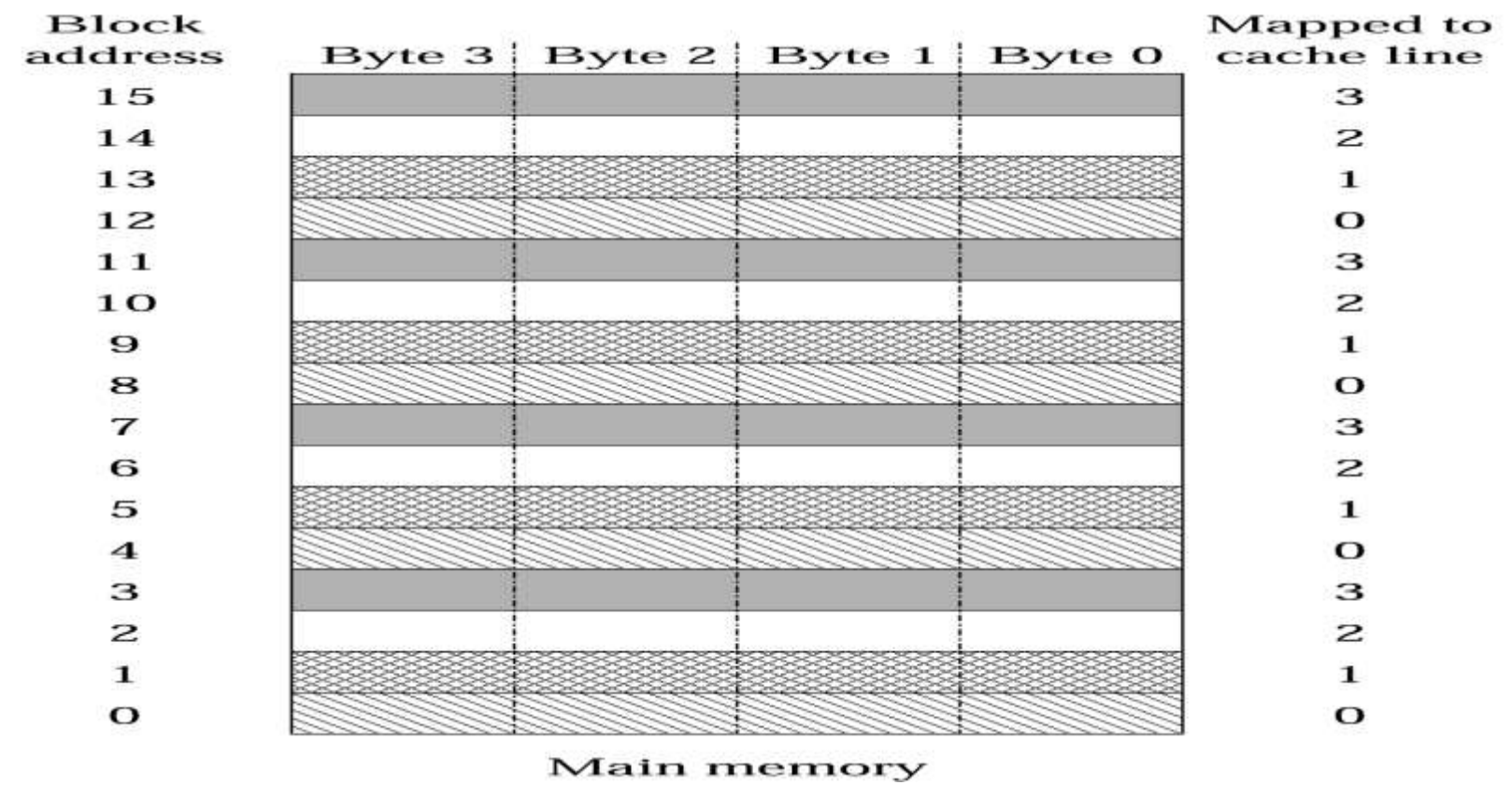
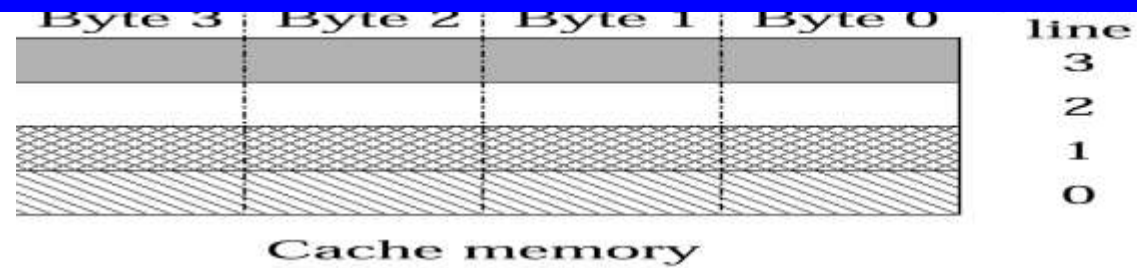




# Mapping Function (cont'd)



## Direct mapping example





# Mapping Function (cont'd)



Implementing direct mapping

Easier than the other two

Maintains three pieces of information

Cache data

Actual data

Cache tag

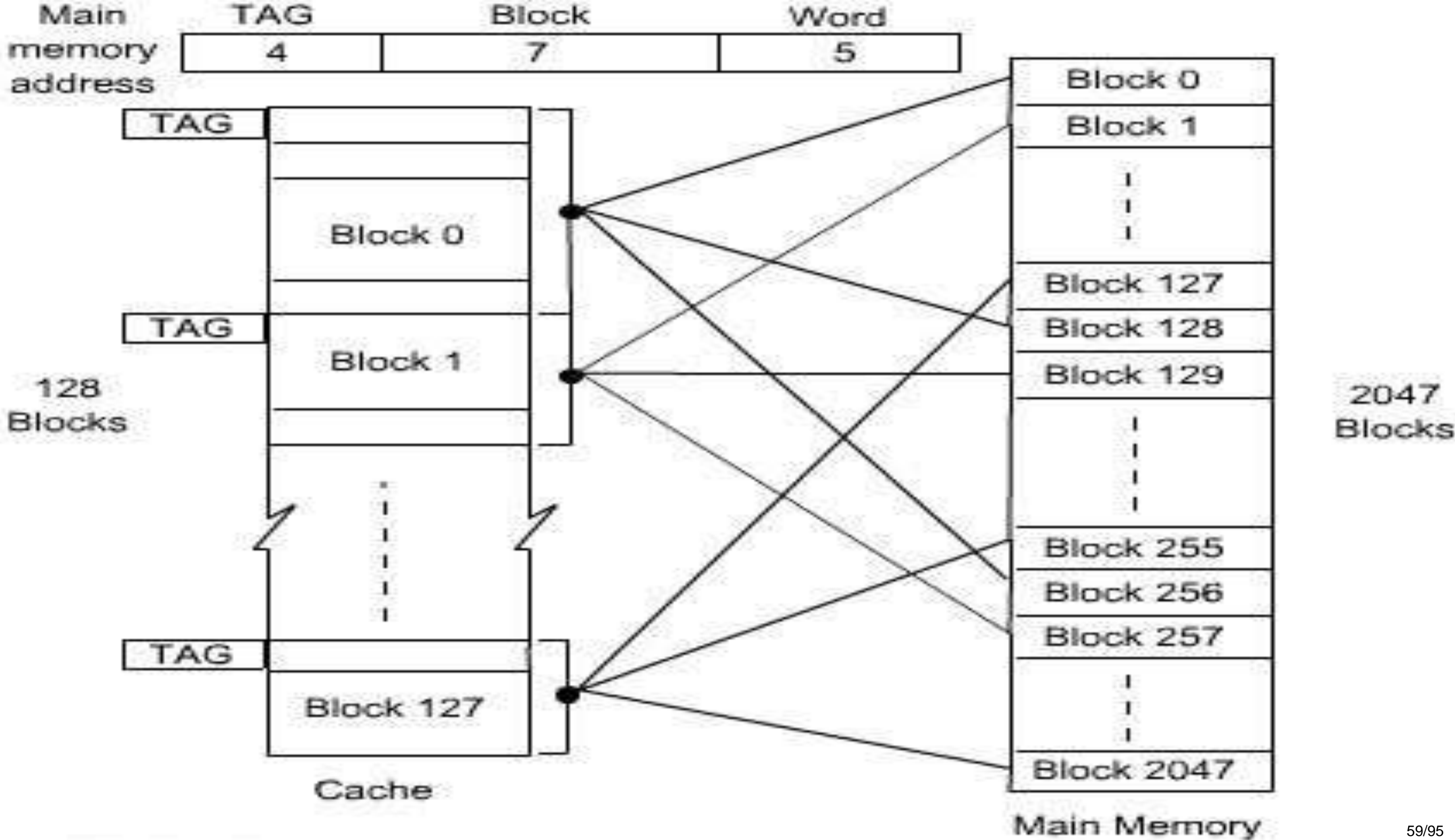
Problem: More memory blocks than cache lines

Several memory blocks are mapped to a cache line

Tag stores the address of memory block in cache line

Valid bit

Indicates if cache line contains a valid block





## Cache Memory

All these three mapping methods are explained with the help of an example.

Consider a cache of 4096 (4K) words with a block size of 32 words. Therefore, the cache is organized as 128 blocks. For 4K words, required address lines are 12 bits. To select one of the block out of 128 blocks, we need 7 bits of address lines and to select one word out of 32 words, we need 5 bits of address lines. So the total 12 bits of address is divided for two groups, lower 5 bits are used to select a word within a block, and higher 7 bits of address are used to select any block of cache memory.

Let us consider a main memory system consisting 64K words. The size of address bus is 16 bits. Since the block size of cache is 32 words, so the main memory is also organized as block size of 32 words. Therefore, the total number of blocks in main memory is 2048 (2K x 32 words = 64K words). To identify any one block of 2K blocks, we need 11 address lines. Out of 16 address lines of main memory, lower 5 bits are used to select a word within a block and higher 11 bits are used to select a block out of 2048 blocks. Number of blocks in cache memory is 128 and number of blocks in main memory is 2048, so at any instant of time only 128 blocks out of 2048 blocks can reside in cache memory. Therefore, we need mapping function to put a particular block of main memory into appropriate block of cache memory.



## Cache Memory

**Direct Mapping Technique:** The simplest way of associating main memory blocks with cache block is the direct mapping technique. In this technique, block  $k$  of main memory maps into block  $k \bmod m$  of the cache, where  $m$  is the total number of blocks in cache. In this example, the value of  $m$  is 128.

In direct mapping technique, one particular block of main memory can be transferred to a particular block of cache which is derived by the modulo function. Since more than one main memory block is mapped onto a given cache block position, contention may arise for that position.

The detail operation of direct mapping technique is as follows:

The main memory address is divided into three fields. The field size depends on the memory capacity and the block size of cache. In this example, the lower 5 bits of address is used to identify a word within a block. Next 7 bits are used to select a block out of 128 blocks (which is the capacity of the cache). The remaining 4 bits are used as a TAG to identify the proper block of main memory that is mapped to cache. When a new block is first brought into the cache, the high order 4 bits of the main memory address are stored in four TAG bits associated with its location in the cache. When the CPU generates a memory request, the 7-bit block address determines the corresponding cache block. The TAG field of that block is compared to the TAG field of the address. If they match, the desired word specified by the low-order 5 bits of the address is in that block of the cache.



# Cache Memory



**Associated Mapping Technique:** In the associative mapping technique, a main memory block can potentially reside in any cache block position.

In this case, the main memory address is divided into two groups, low-order bits identifies the location of a word within a block and high-order bits identifies the block.

In the example here, 11 bits are required to identify a main memory block when it is resident in the cache , high-order 11 bits are used as TAG bits and low-order 5 bits are used to identify a word within a block.



# Associated Mapping Technique



Main  
memory  
address

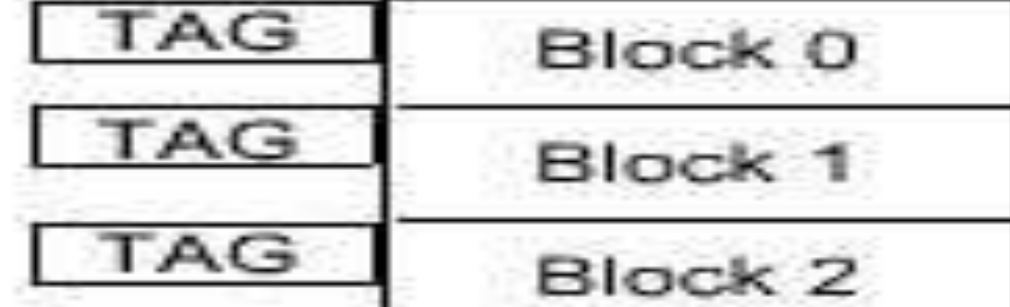
TAG

11

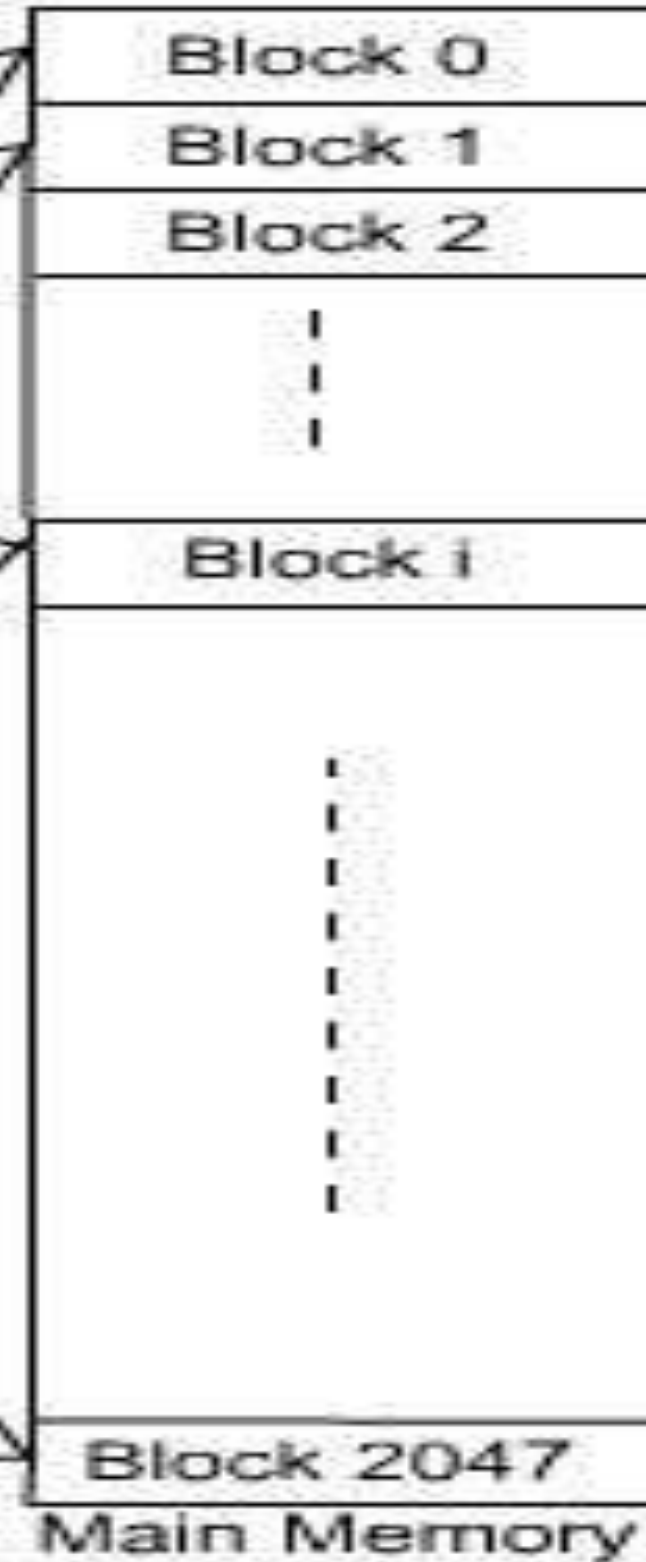
Word

5

←11 bits→



128  
Blocks



2047  
Blocks



# Cache Memory



**Block-Set-Associative Mapping Technique:** This mapping technique is intermediate to the previous two techniques. Blocks of the cache are grouped into sets, and the mapping allows a block of main memory to reside in any block of a specific set.

This also reduces the searching overhead, because the search is restricted to number of sets, instead of number of blocks.

Consider the same cache memory and main memory organization of the previous example. Organize the cache with 4 blocks in each set. The TAG field of associative mapping technique is divided into two groups, one is termed as SET bit and the second one is termed as TAG bit. Each set contains 4 blocks, total number of set is 32.

The main memory address is grouped into three parts: low-order 5 bits are used to identifies a word within a block. Since there are total 32 sets present, next 5 bits are used to identify the set. High-order 6 bits are used as TAG bits. The 5-bit set field of the address determines which set of the cache might contain the desired block.

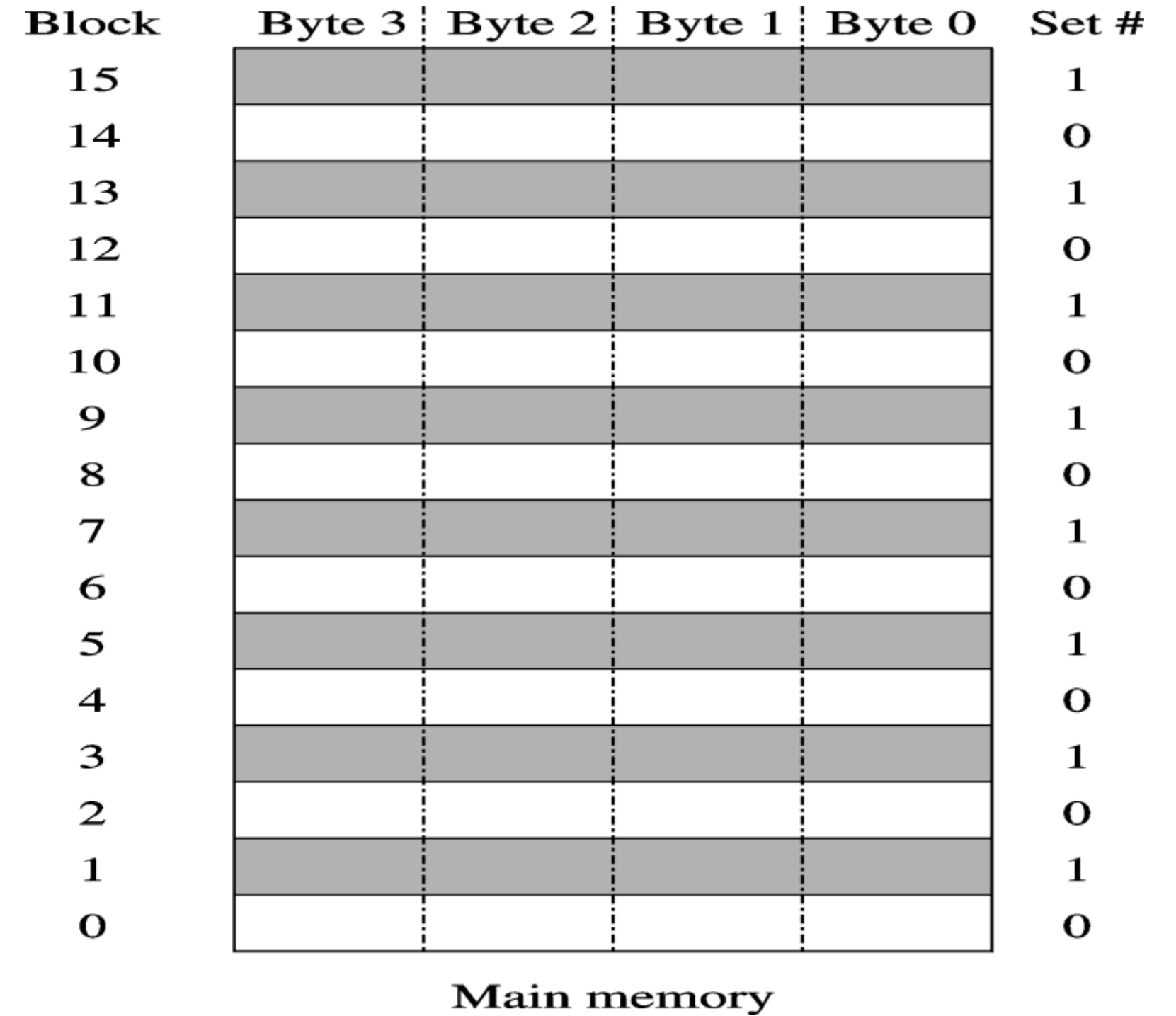
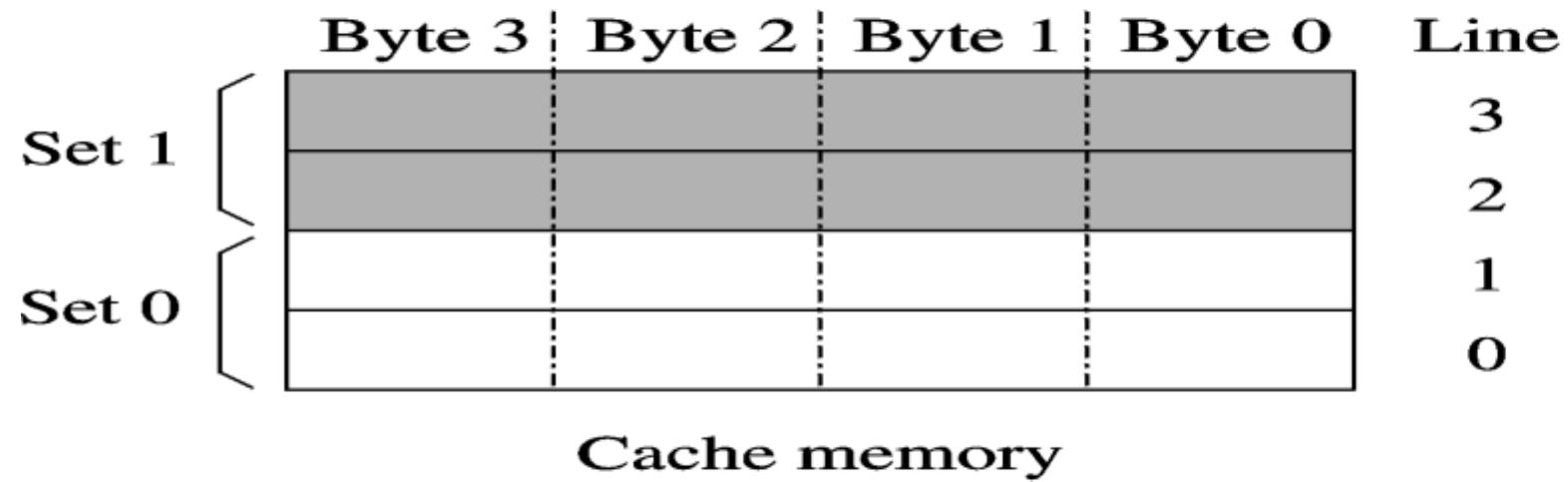


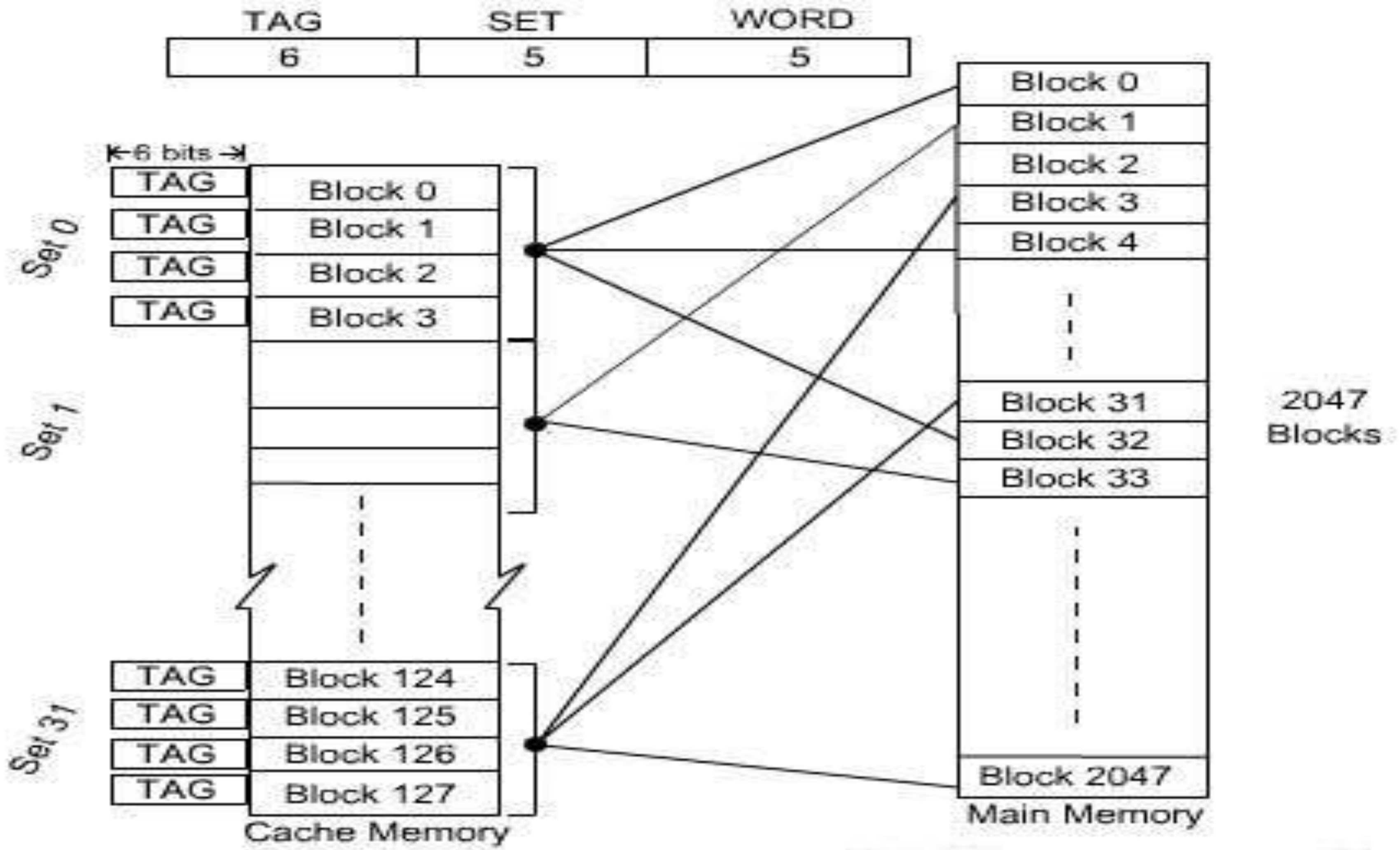


# Mapping Function (cont'd)



## Set-associative mapping







# Replacement algorithm

When the cache is full and a memory word (instruction or data) that is not in the cache is referenced, the cache control hardware must decide which block should be removed to create space for the new block that contains the referenced word.

The collection of rules for making this decision constitutes the **replacement algorithm**.



## Cache Memory

Least

Recently Used (LRU) Replacement policy:

Since program usually stay in localized areas for reasonable periods of time, it can be assumed that there is a high probability that blocks which have been referenced recently will also be referenced in the near future. Therefore, when a block is to be overwritten, it is a good decision to overwrite the one that has gone for longest time without being referenced. This is defined as the least recently used (LRU) block.





# Virtual Memory

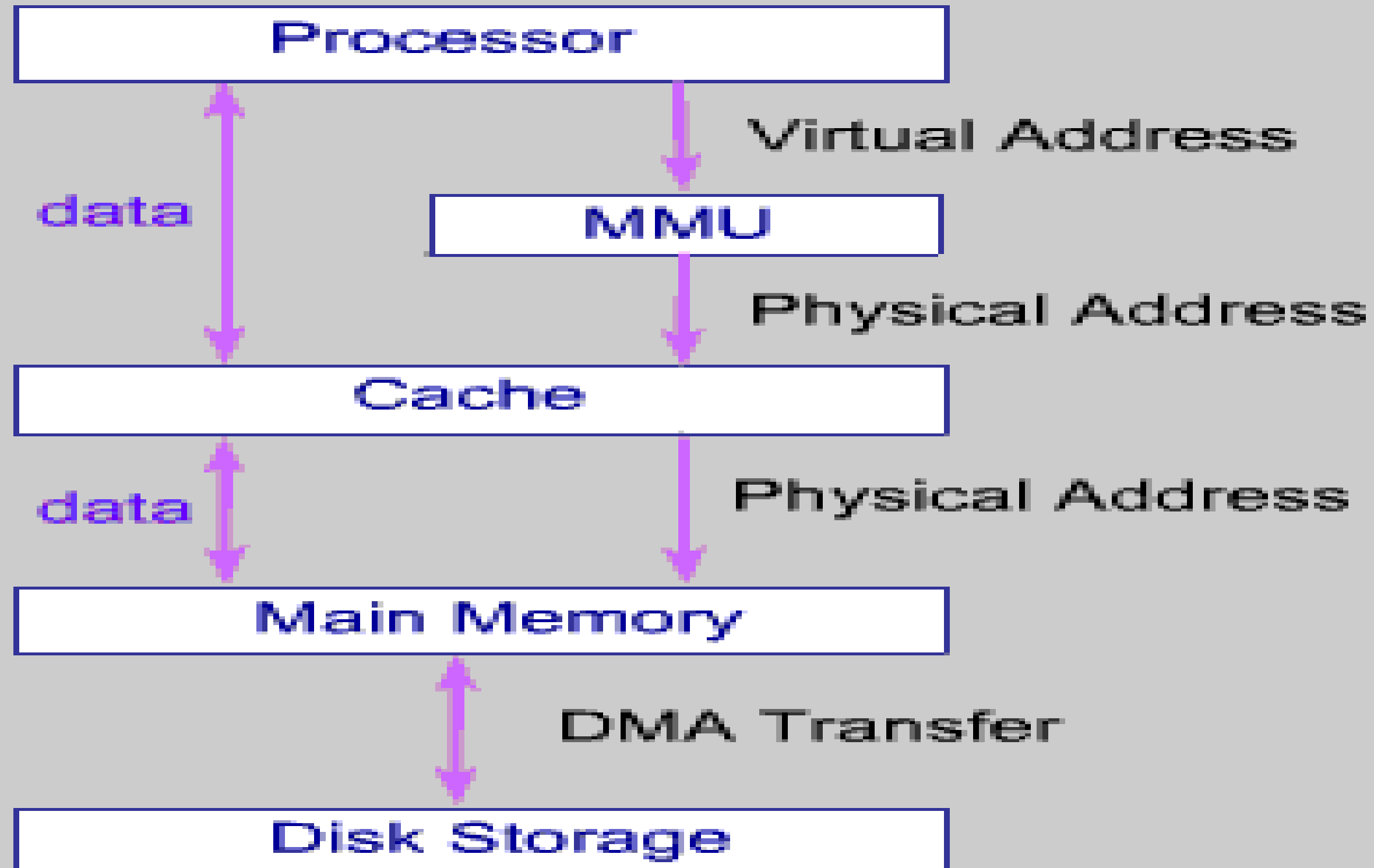


Virtual (or logical) memory is a concept that, when implemented by a computer and its operating system, allows programmers to use a very large range of memory or storage addresses for stored data.

The computing system maps the programmer's virtual addresses to real hardware storage addresses. Usually, the programmer is freed from having to be concerned about the availability of data storage.



# Virtual Memory Organization





# Address Translation



The basic mechanism for reading a word from memory involves the translation of a virtual or logical address, consisting of page number and offset, into a physical address, consisting of frame number and offset, using a page table.

Page table base register contains the starting address of the page table.

Page frame is an area in main memory which can hold a page.

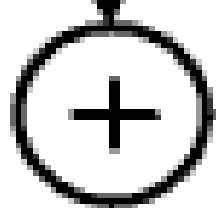
Control bits: validity bit, modify, read and write permission etc.

Page table base register

Page table address

Virtual address from processor

Virtual page no    offset



PAGE TABLE

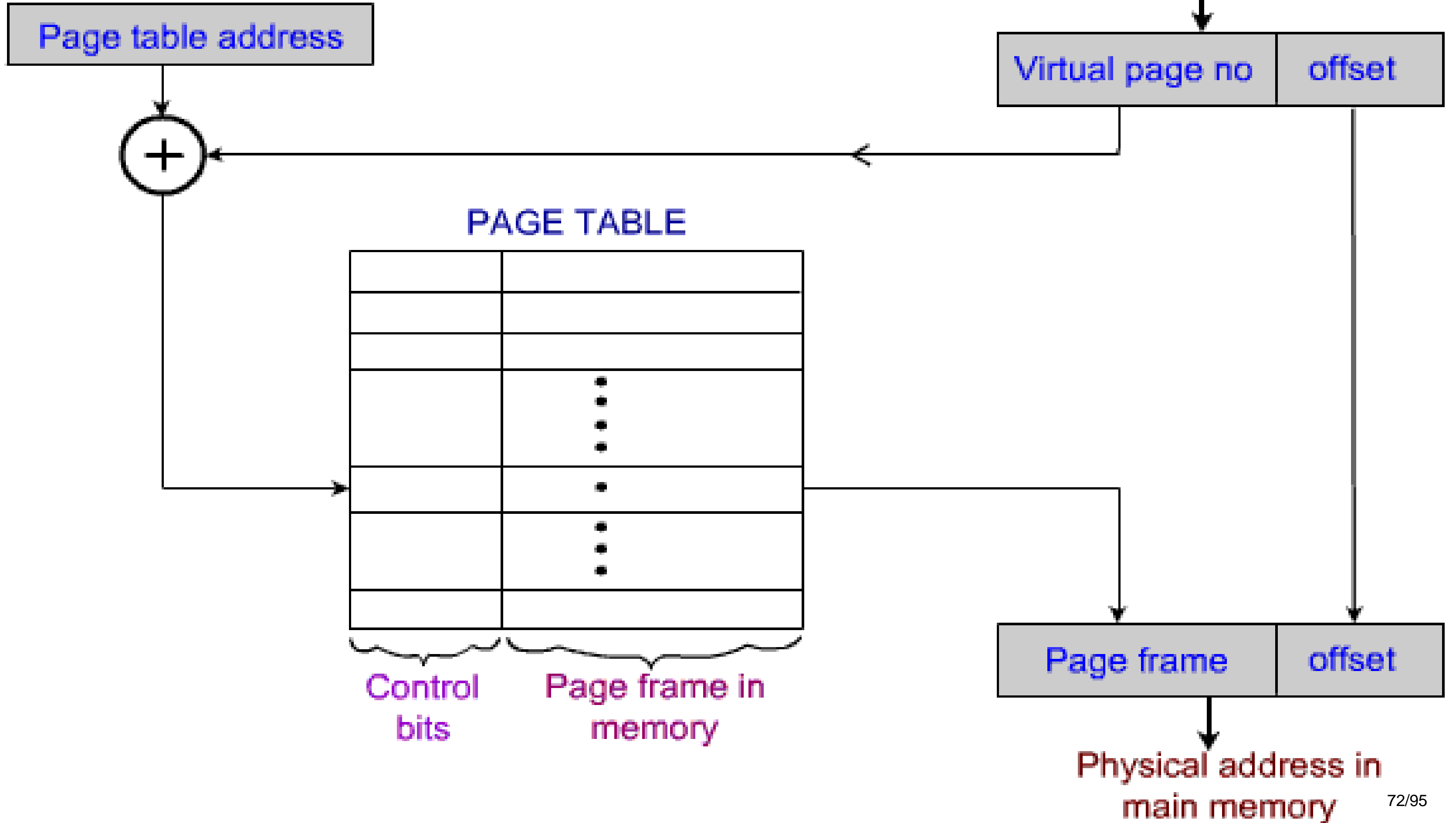
	⋮
	⋮
	⋮
	⋮

Control bits

Page frame in memory

Page frame    offset

Physical address in main memory







## Translation Lookaside Buffer (TLB)

Every virtual memory reference can cause two physical memory accesses.

One to fetch the appropriate page table entry

One to fetch the desired data.

Thus a straight forward virtual memory scheme would have the effect of doubling the memory access time.

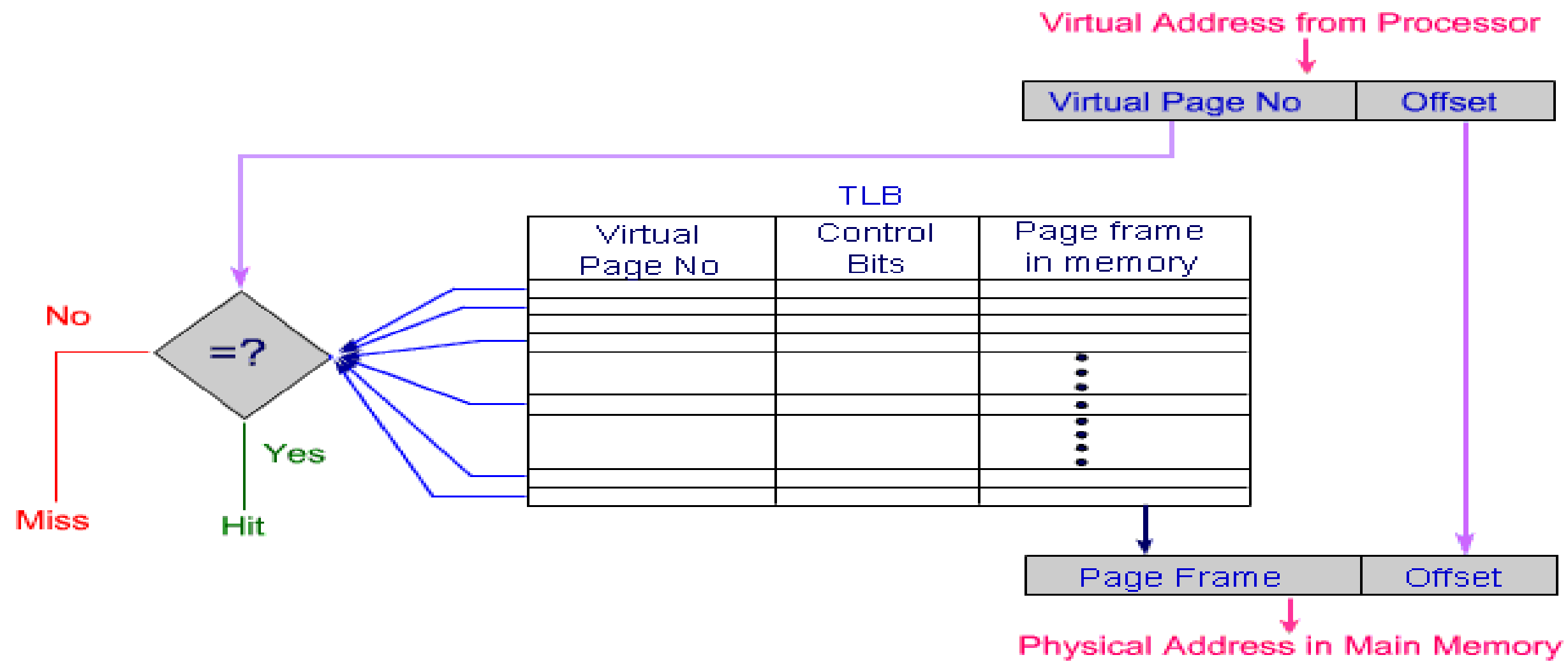
To overcome this problem, most virtual memory schemes make use of a special cache for page table entries, usually called *Translation Lookaside Buffer (TLB)*.

This cache functions in the same way as a memory cache and contains those page table entries that have been most recently used.

In addition to the information that constitutes a page table entry, the TLB must also include the virtual address of the entry.

The Figure 3.27 shows a possible organization of a TLB where the associative mapping technique is used.







# Memory Management

In an uniprogramming system, main memory is divided into two parts : one part for the operating system and the other part for the program currently being executed.

In multiprogramming system, the user part of memory is subdivided to accommodate multiple processes.

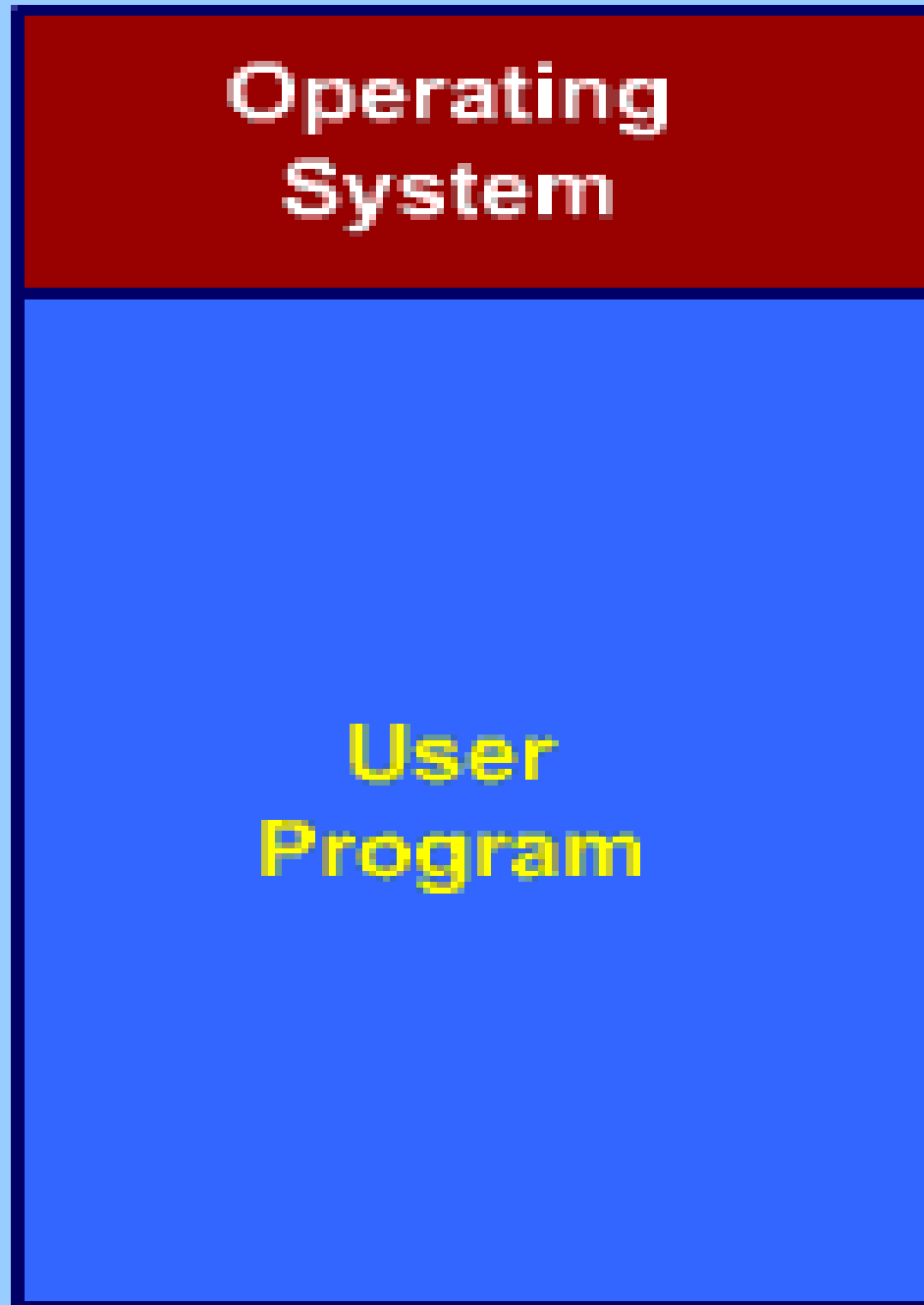
The task of subdivision is carried out dynamically by the operating system and is known as ***memory management***.

In uniprogramming system, only one program is in execution. After compilation of one program, another program may start.

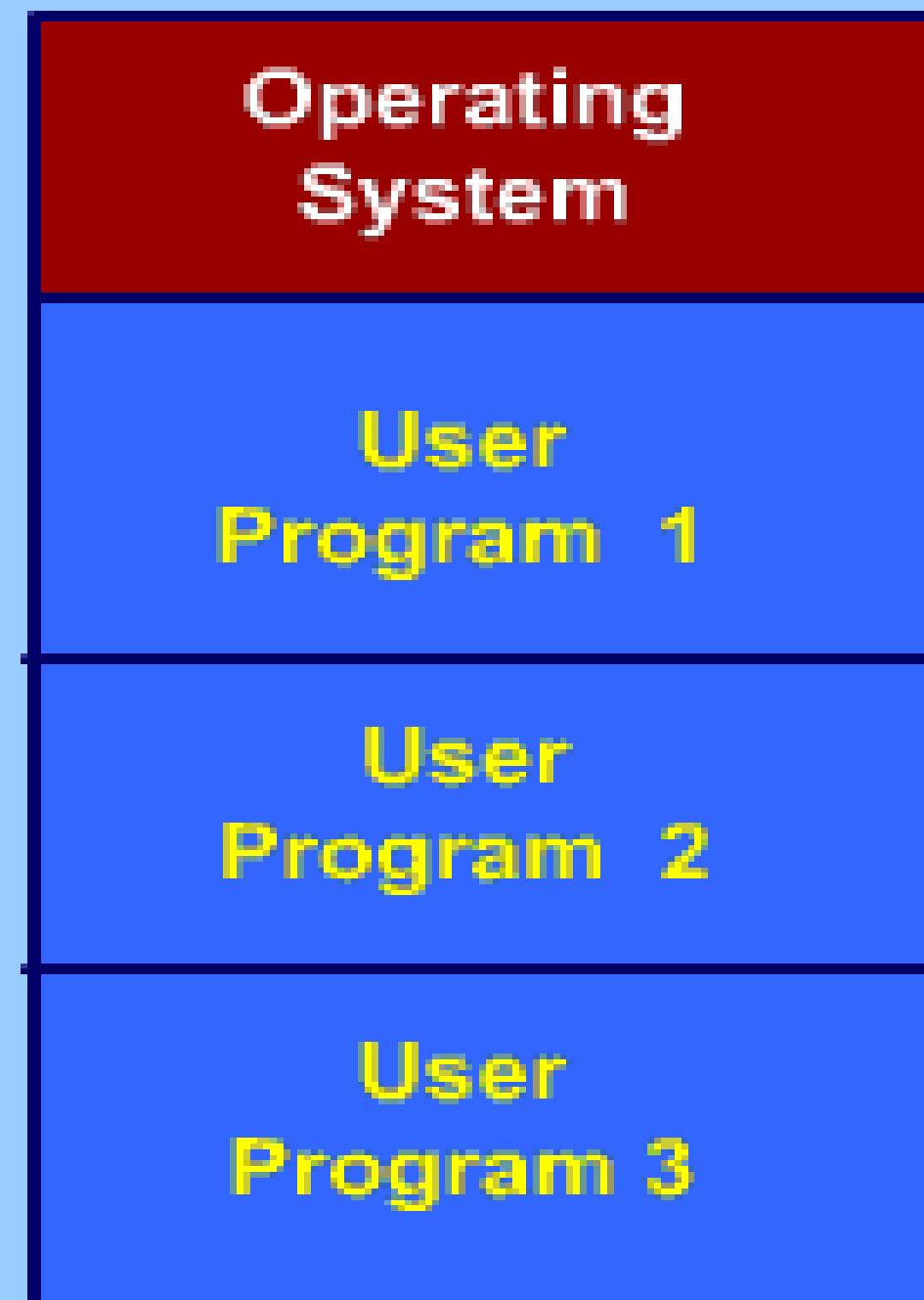
In general, most of the programs involve I/O operation. It must take input from some input device and place the result in some output device.

Partition of main memory for uni-program and multi program is shown in figure 3.19.





**Memory  
Uni - Program**



**Memory  
Multi - Program**





# PERFORMANCE CONSIDERATIONS

Memory interleaving:

when the main memory of a computer is structured as a collection of modules, memory access operations may proceed in more than one module at the same.

Consecutive addresses are located in successive modules.

Any request for access to consecutive memory locations can keep several modules busy at any one time.

This results in both faster access to a block of data and higher average utilization of the memory as a whole.



Improve Hit rate

Miss penalty: The extra time needed to bring the desired information into the cache. Interleaved memory can reduce miss penalty.



The average access time experienced by the processor is

$$T_{ave} = hc + (1-h)m$$

Where  $h$  = hit rate

$c$  = the time to access information  
in the cache

$m$  = the miss penalty that is the time to access information in  
the main memory.



Caches on the processor chip

Use two levels of cache L1, L2

$T_{ave} = h_1c_1 + (1-h_1)h_2c_2 + (1-h_1)(1-h_2)m$

$h_1$  = hit rate in L1 cache

$h_2$  = hit rate in L2 cache

$c_1$  = time to access information in the L1 cache

$c_2$  = time to access information in the L2 cache

$m$  = the miss penalty that is the time to access information in the main memory





## Other enhancements

Write buffer can be included for temporary storage of write requests. It is possible that a subsequent read request may refer to data that are still in the write buffer.

Prefetching : prefetch the data into the cache before they are needed.



# Magnetic Tape Systems



Off-line storage of large amounts of data

Back-up and archival storage

Data organized into records and files

Highest capacity

Slowest

Cheapest



# Magnetic Tape Systems



Off-line storage of large amounts of data

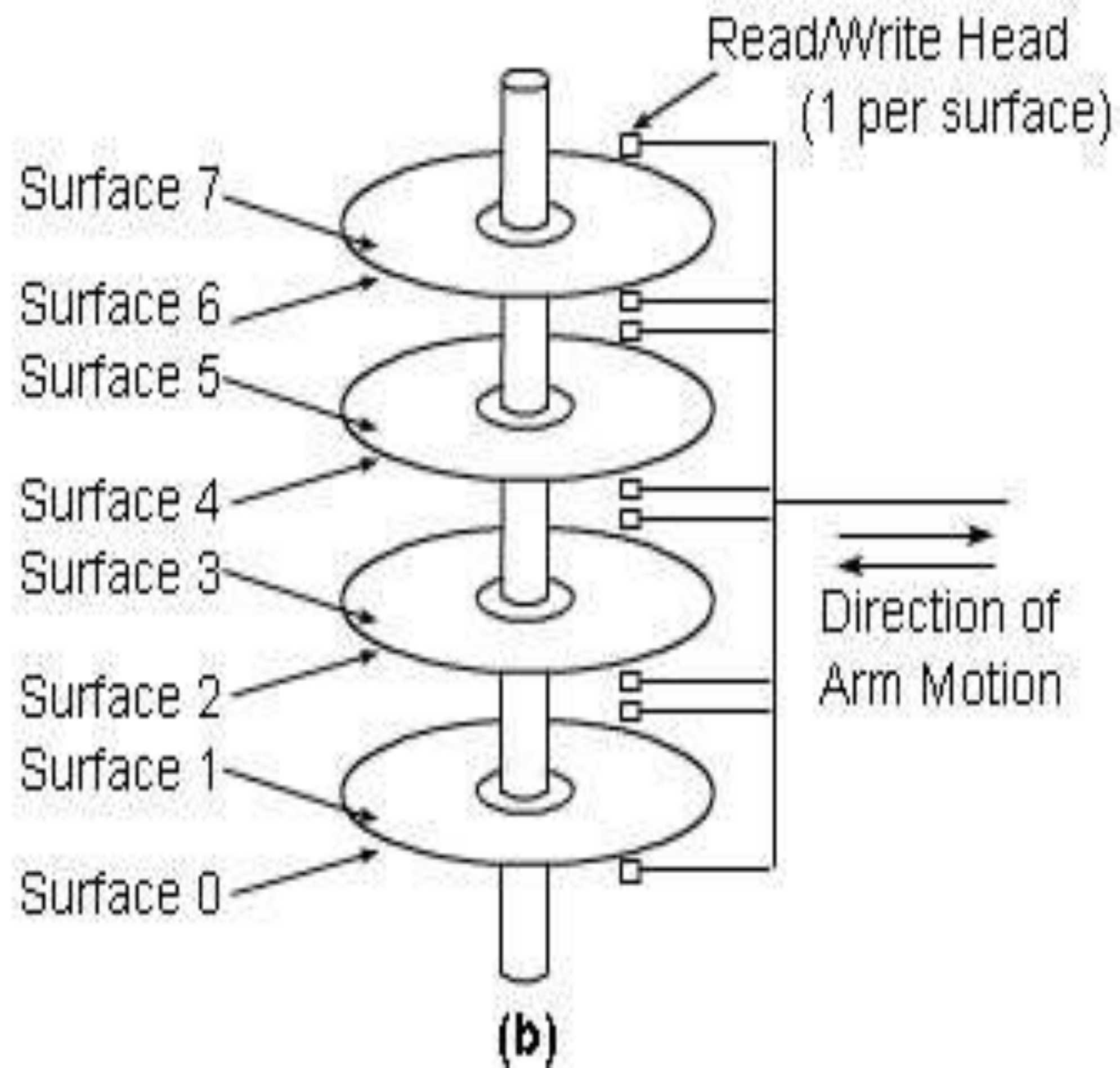
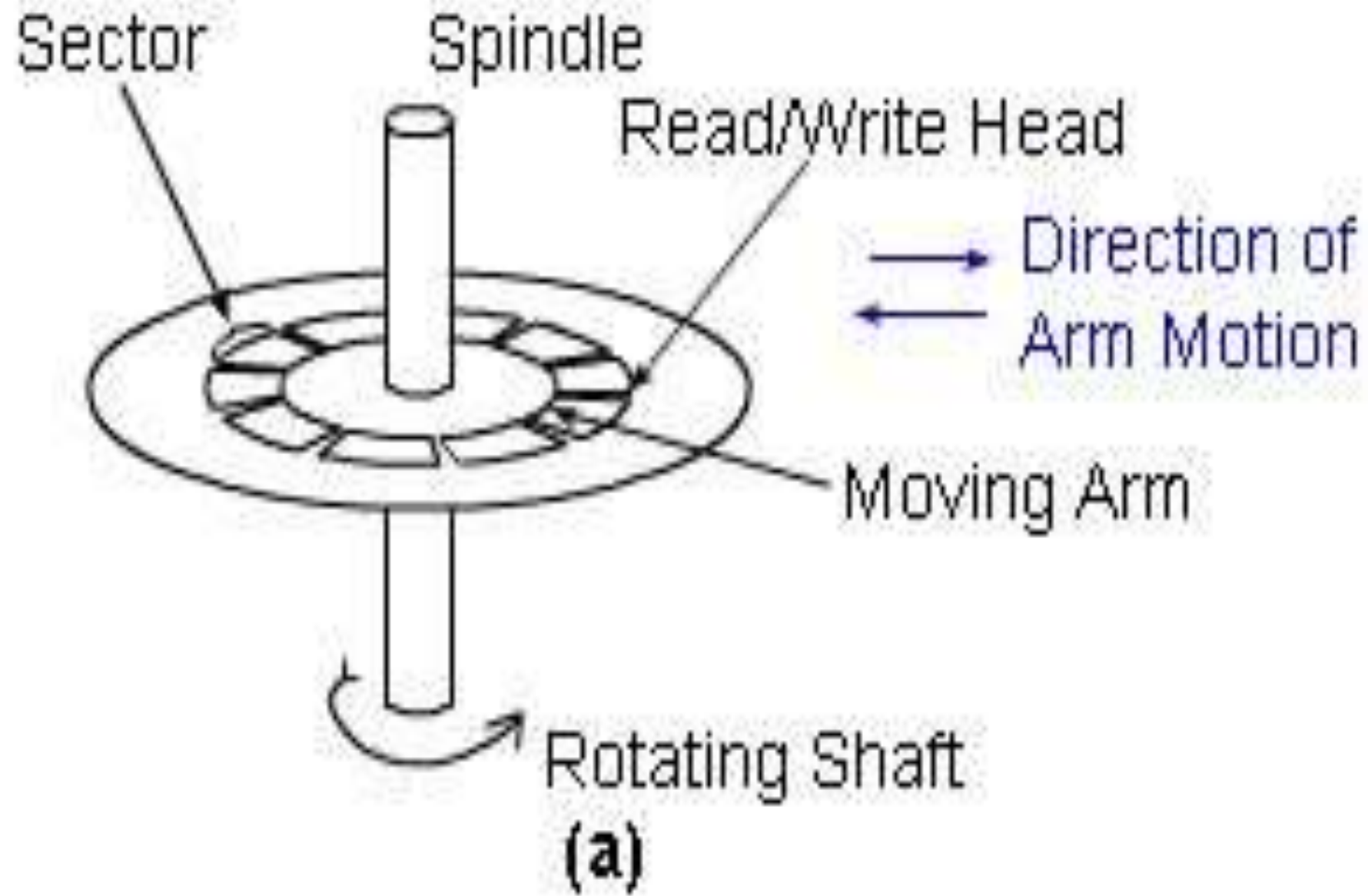
Back-up and archival storage

Data organized into records and files

Highest capacity

Slowest

Cheapest







# Storing the Data



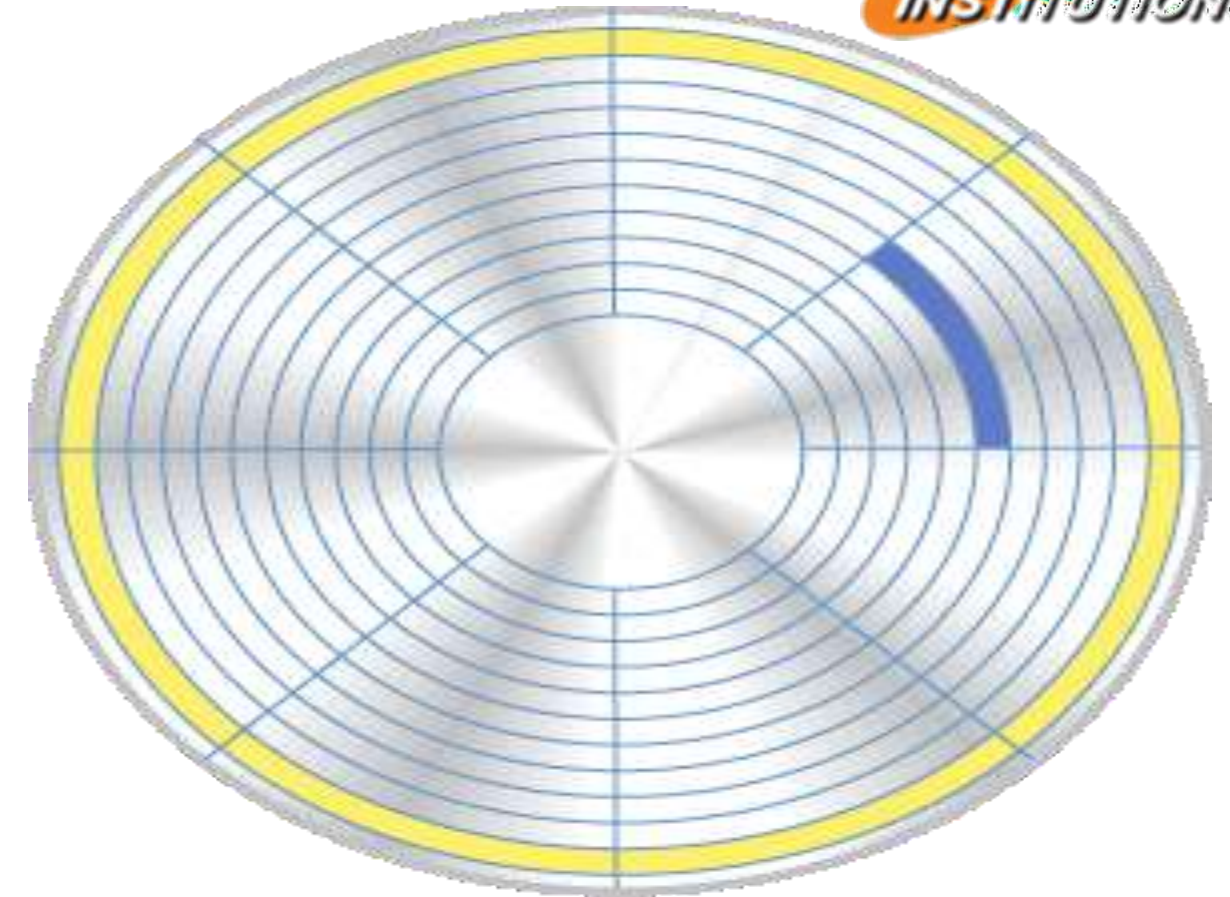
Data is stored on platter surface

**Tracks** -> concentric circles

**Sectors** -> pie-shaped wedges on a track

A sector contains a fixed number of bytes (256, 512 etc.)

Sectors are often grouped together into clusters



©2000 How Stuff Works

## Low level formatting

The drive establishes tracks and sectors on the platter  
Prepares the drive to store blocks of bytes

## High level formatting

Writes the file-storage structures, like file allocation table into sectors  
Prepares the drive to hold files



# Platters and Heads



Multiple platters to increase information storage capacity

This drive has three platters and six read/write heads





# Internal components of Hard disk drive





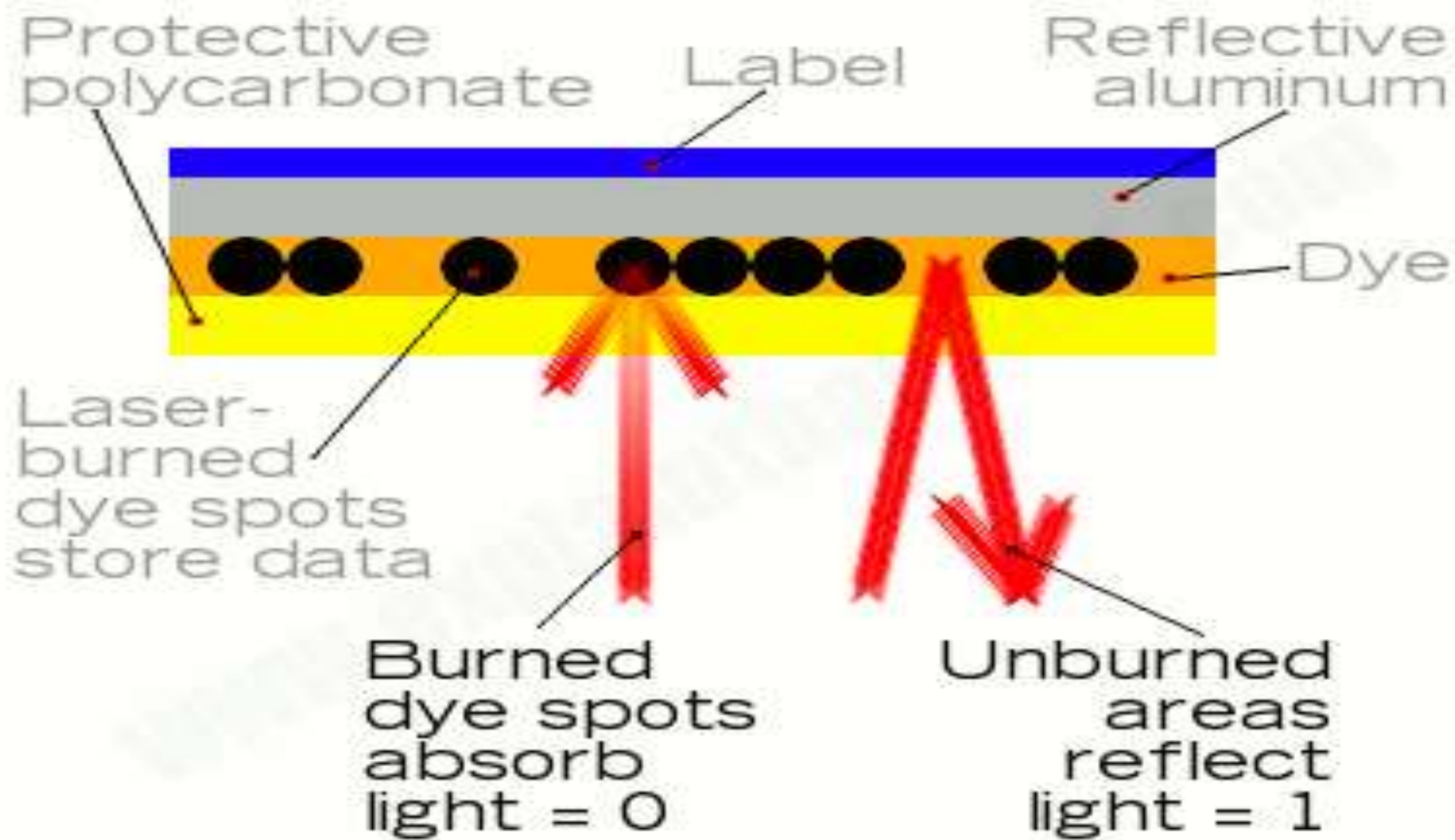
**FIGURE 4.14**

Characteristics of storage media for business consideration

Medium	Storage Capacity	Transfer Rate	Cost (per 1 MB)
Magnetic Hard Disk 	High	Fast	Moderate
Magnetic Tape 	Moderate	Slow	Very Low
Optical Tape 	Very High	Very Slow	Low
CD 	High	Very Slow	Low
DVD 	Very High	Moderate	Very High
Flash Memory 	High	Moderate	Very High



# CD/DVD





## CROSS SECTION OF SMALL PORTION OF CD

- Bottom layer-polycarbonate plastic
- Surface of plastic- programmed
- To store data by indenting with pits
- Unindented parts- lands
- Aluminum layer placed on top of disk
- Covered by protective acrylic layer
- Other layers thin.



## SECONDARY STORAGE

These high-speed storage devices are very expensive and hence the cost per bit of storage is also very high. Again the storage capacity of the main memory is also very limited. Therefore additional memory is required in all the computer systems. This memory is called *auxiliary memory* or *secondary storage*.

In this type of memory the cost per bit of storage is low. However, the operating speed is slower than that of the primary storage.

Huge volume of data are stored here on permanent basis and transferred to the primary storage as and when required.

Most widely used secondary storage devices are *magnetic tapes ,cdrom,dvd,hdd etc*





**Magnetic Tape:** Magnetic tapes are used for large computers like mainframe computers where large volume of data is stored for a longer time.

In PC also you can use tapes in the form of cassettes. The cost of storing data in tapes is inexpensive. Tapes consist of magnetic materials that store data permanently.

The deck is connected to the central processor and information is fed into or read from the tape through the processor. It similar to cassette tape recorder.





## Optical Disk:

With every new application and software there is greater demand for memory capacity. It is the necessity to store large volume of data that has led to the development of optical disk storage medium. Optical disks can be divided into the following categories:

**Compact Disk/ Read Only Memory (CD-ROM):** CD-ROM disks are made of reflective metals. CD-ROM is written during the process of manufacturing by high power *laser beam*. Here the storage density is very high, storage cost is very low and access time is relatively fast. Each disk is approximately 4 1/2 inches in diameter and can hold over 600 MB of data. As the CD-ROM can be *read only* we cannot write or make changes into the data contained in it.

**Write Once, Read Many (WORM):** The inconvenience that we can not write any thing in to a CD-ROM is avoided in WORM. A WORM allows the user to write data permanently on to the disk. Once the data is written it can never be erased without physically damaging the disk. Here data can be recorded from keyboard, video scanner, OCR equipment and other devices. The advantage of WORM is that it can store vast amount of data amounting to gigabytes (10<sup>9</sup> bytes). Any document in a WORM can be accessed very fast, say less than 30 seconds.

**Erasable Optical Disk:** These are optical disks where data can be written, erased and re-written. This also applies a laser beam to write and re-write the data. These disks may be used as alternatives to traditional disks. Erasable optical disks are based on a technology known as *magnetic optical* (MO



# Assessment



a). What is Memory?

---

---

---

b) Mention the purpose

1. RAM \_\_\_\_\_
2. EPROM \_\_\_\_\_
3. Secondary memory \_\_\_\_\_





# Reference



1. Carl Hamacher, Zvonko Vranesic and Safwat Zaky, “Computer Organization”, McGraw-Hill, 6<sup>th</sup> Edition 2012.