**UNIT - IV**

**INPUT-OUTPUT ORGANIZATION**

# Peripheral Devices:

The Input / output organization of computer depends upon the size of computer and the peripherals connected to it. The I/O Subsystem of the computer, provides an efficient mode of communication between the central system and the outside environment

The most common input output devices are:

1. Monitor
2. Keyboard
3. Mouse
4. Printer
5. Magnetic tapes

The devices that are under the direct control of the computer are said to be connected online.

# Input - Output Interface

Input Output Interface provides a method for transferring information between internal storage and external I/O devices.

Peripherals connected to a computer need special communication links for interfacing them with the central processing unit.

The purpose of communication link is to resolve the differences that exist between the central computer and each peripheral.

The Major Differences are:-

1. Peripherals are electromechnical and electromagnetic devices and CPU and memory are electronic devices. Therefore, a conversion of signal values may be needed.
2. The data transfer rate of peripherals is usually slower than the transfer rate of CPU and consequently, a synchronization mechanism may be needed.
3. Data codes and formats in the peripherals differ from the word format in the CPU and memory.
4. The operating modes of peripherals are different from each other and must be controlled so as not to disturb the operation of other peripherals connected to the CPU.

To Resolve these differences, computer systems include special hardware components between the CPU and Peripherals to supervises and synchronizes all input and out transfers

* These components are called Interface Units because they interface between the processor bus and the peripheral devices.

I/O BUS and Interface Module

It defines the typical link between the processor and several peripherals.

The I/O Bus consists of data lines, address lines and control lines. The I/O bus from the processor is attached to all peripherals interface.

To communicate with a particular device, the processor places a device address on address lines.

Each Interface decodes the address and control received from the I/O bus, interprets them for peripherals and provides signals for the peripheral controller.

It is also synchronizes the data flow and supervises the transfer between peripheral and processor.

Each peripheral has its own controller.

For example, the printer controller controls the paper motion, the print timing The control lines are referred as I/O command. The commands are as following:

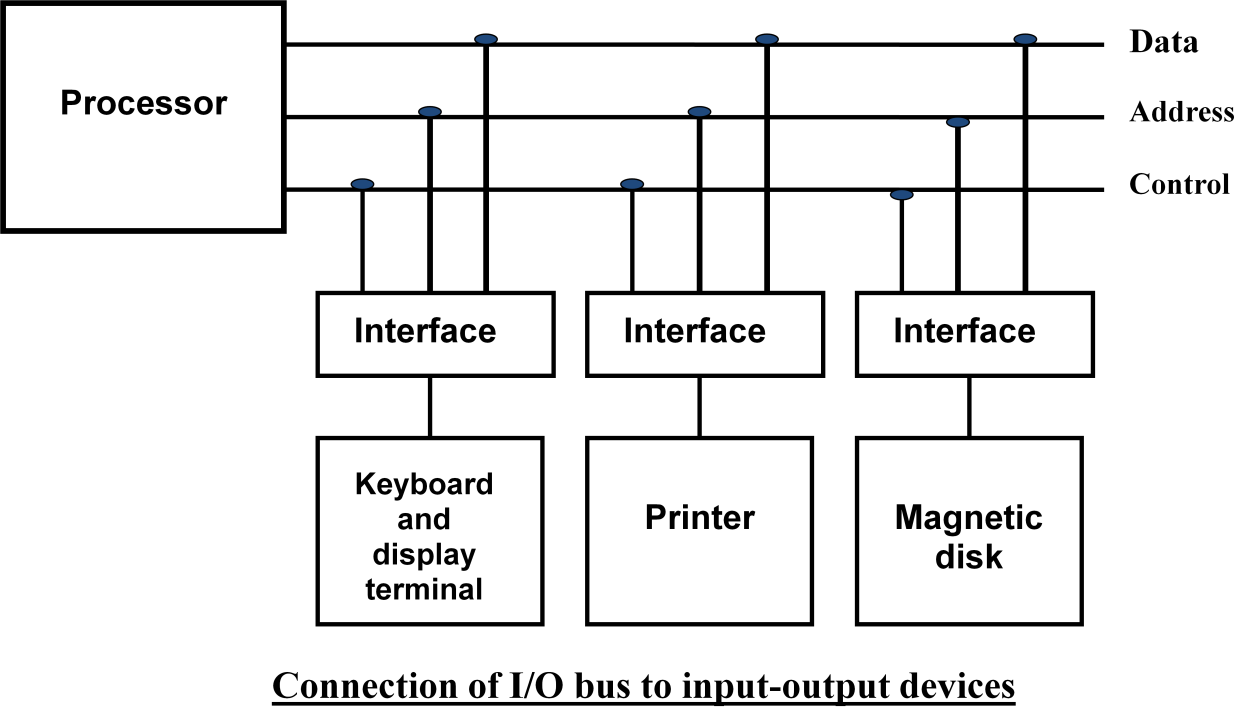
Control command- A control command is issued to activate the peripheral and to inform it what to do.

Status command- A status command is used to test various status conditions in the interface and the peripheral.

Data Output command- A data output command causes the interface to respond by transferring data from the bus into one of its registers.

Data Input command- The data input command is the opposite of the data output.

In this case the interface receives on item of data from the peripheral and places it in its buffer register. I/O Versus Memory Bus



To communicate with I/O, the processor must communicate with the memory unit. Like the I/O bus, the memory bus contains data, address and read/write control lines. There are 3 ways that computer buses can be used to communicate with memory and I/O:

1. Use two Separate buses , one for memory and other for I/O.
2. Use one common bus for both memory and I/O but separate control lines for each.
3. Use one common bus for memory and I/O with common control lines. I/O Processor

In the first method, the computer has independent sets of data, address and control buses one for accessing memory and other for I/O. This is done in computers that provides a separate I/O processor (IOP). The purpose of IOP is to provide an independent pathway for the transfer of information between external device and internal memory.

# Asynchronous Data Transfer :

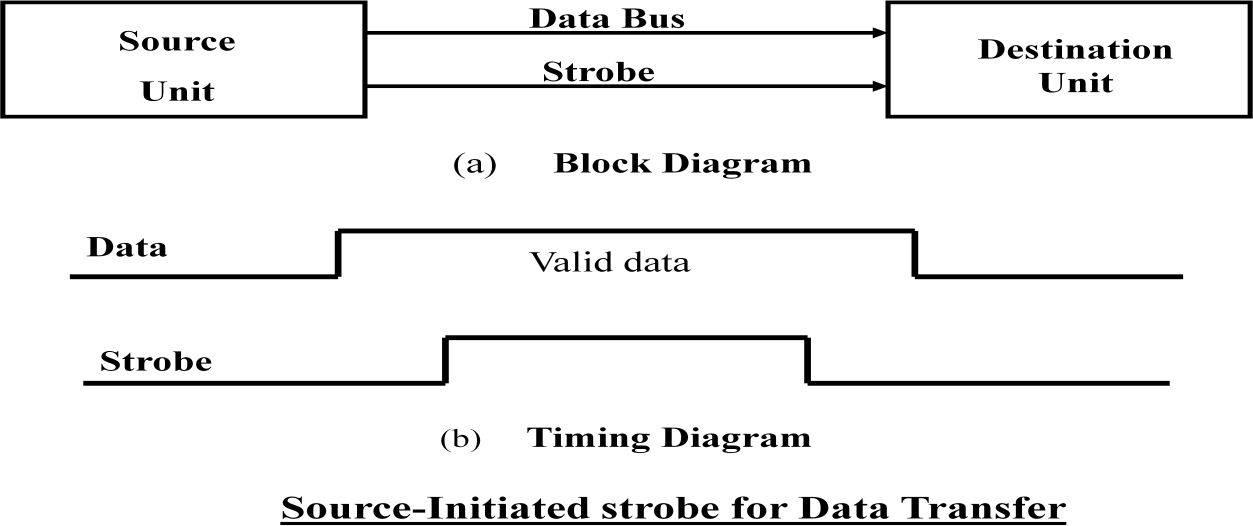
This Scheme is used when speed of I/O devices do not match with microprocessor, and timing characteristics of I/O devices is not predictable. In this method, process initiates the device and check its status. As a result, CPU has to wait till I/O device is ready to transfer data. When device is ready CPU issues instruction for I/O transfer. In this method two types of techniques are used based on signals before data transfer.

* 1. Strobe Control
  2. Handshaking

## Strobe Signal :

The strobe control method of Asynchronous data transfer employs a single control line to time each transfer. The strobe may be activated by either the source or the destination unit.

Data Transfer Initiated by Source Unit:



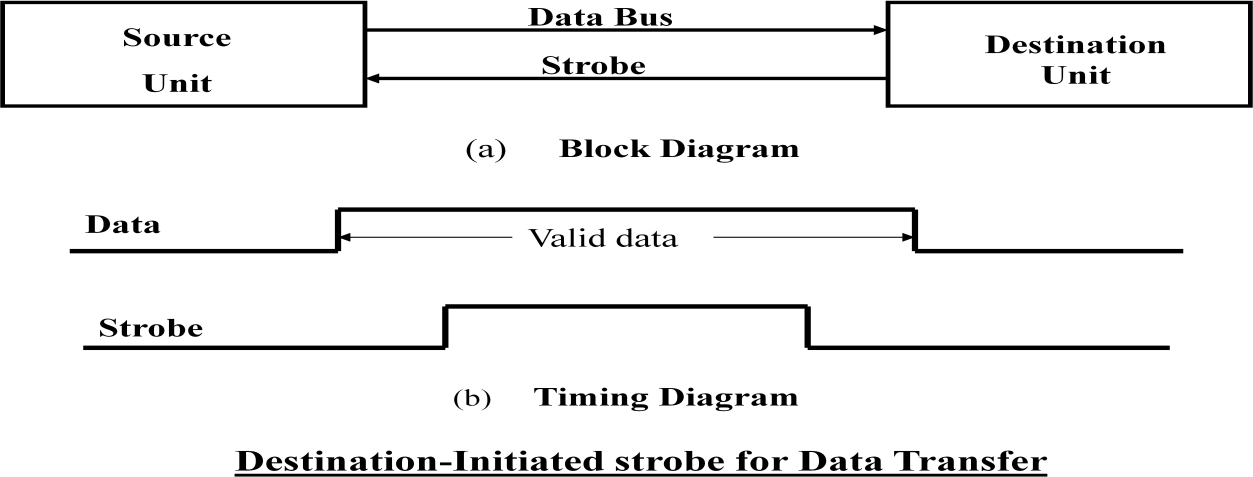
In the block diagram fig. (a), the data bus carries the binary information from source to destination unit. Typically, the bus has multiple lines to transfer an entire byte or word. The strobe is a single line that informs the destination unit when a valid data word is available.

The timing diagram fig. (b) the source unit first places the data on the data bus. The information on the data bus and strobe signal remain in the active state to allow the destination unit to receive the data.

Data Transfer Initiated by Destination Unit:

In this method, the destination unit activates the strobe pulse, to informing the source to provide the data. The source will respond by placing the requested binary information on the data bus.

The data must be valid and remain in the bus long enough for the destination unit to accept it. When accepted the destination unit then disables the strobe and the source unit removes the data from the bus.



## Disadvantage of Strobe Signal :

The disadvantage of the strobe method is that, the source unit initiates the transfer has no way of knowing whether the destination unit has actually received the data item that was places in the bus. Similarly, a destination unit that initiates the transfer has no way of knowing whether the source unit has actually placed the data on bus. The Handshaking method solves this problem.

## Handshaking:

The handshaking method solves the problem of strobe method by introducing a second control signal that provides a reply to the unit that initiates the transfer.

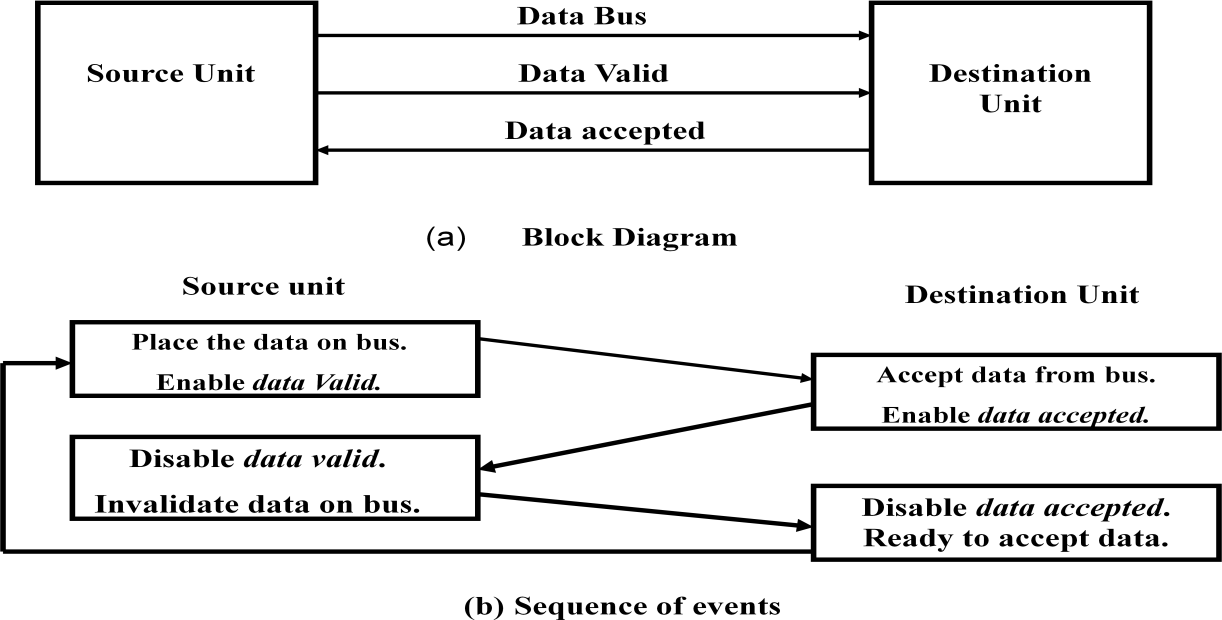
Principle of Handshaking:

The basic principle of the two-wire handshaking method of data transfer is as follow:

One control line is in the same direction as the data flows in the bus from the source to destination. It is used by source unit to inform the destination unit whether there a valid data in the bus. The other control line is in the other direction from the destination to the source. It is used by the destination unit to inform the source whether it can accept the data. The sequence of control during the transfer depends on the unit that initiates the transfer.

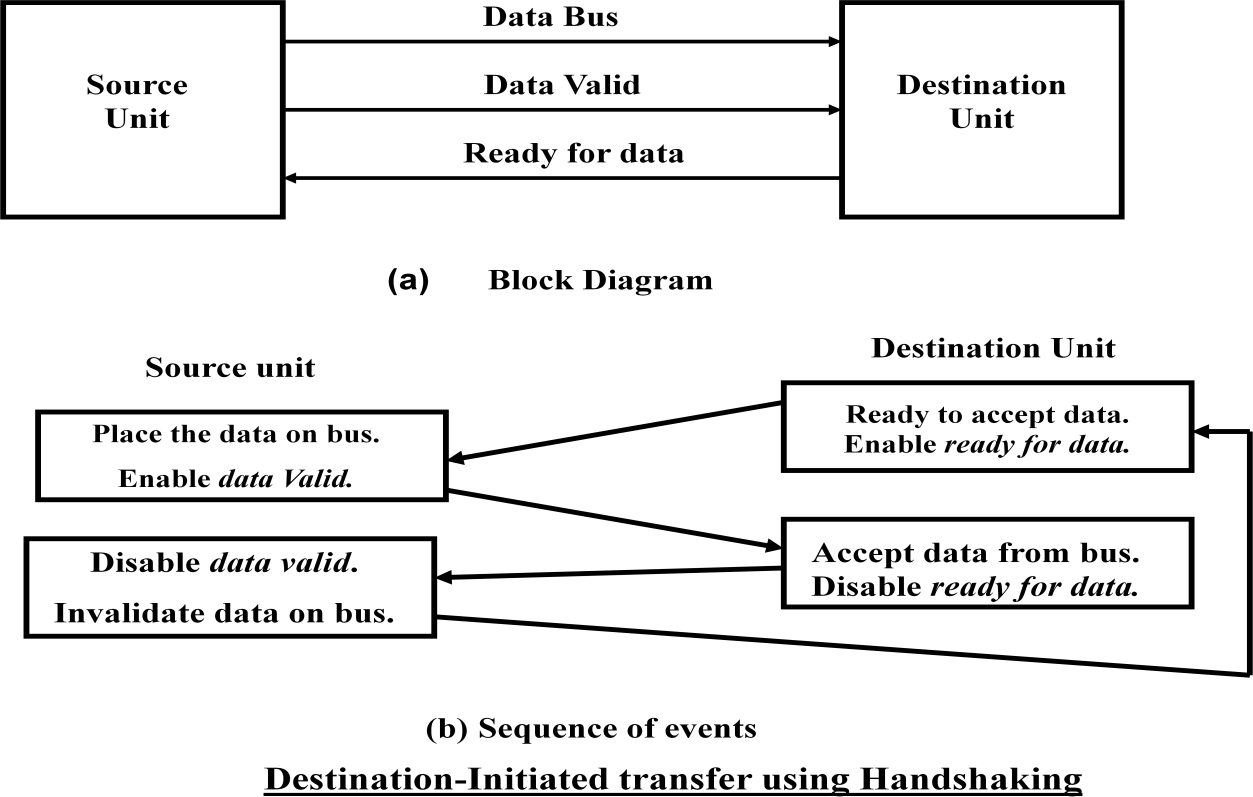
Source Initiated Transfer using Handshaking:

The sequence of events shows four possible states that the system can be at any given time. The source unit initiates the transfer by placing the data on the bus and enabling its *data valid* signal. The *data accepted* signal is activated by the destination unit after it accepts the data from the bus. The source unit then disables its *data accepted* signal and the system goes into its initial state.



Destination Initiated Transfer Using Handshaking:

The name of the signal generated by the destination unit has been changed to *ready for data* to reflects its new meaning. The source unit in this case does not place data on the bus until after it receives the *ready for data* signal from the destination unit. From there on, the handshaking procedure follows the same pattern as in the source initiated case.

The only difference between the Source Initiated and the Destination Initiated transfer is in their choice of Initial sate.

Advantage of the Handshaking method:

* The Handshaking scheme provides degree of flexibility and reliability because the successful completion of data transfer relies on active participation by both units.
* If any of one unit is faulty, the data transfer will not be completed. Such an error can be detected by means of a *Timeout mechanism* which provides an alarm if the data is not completed within time.

## Asynchronous Serial Transmission:

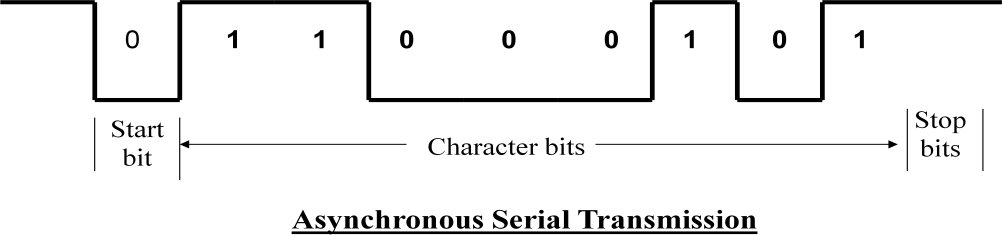
The transfer of data between two units is serial or parallel. In parallel data transmission, n bit in the message must be transmitted through n separate conductor path. In serial transmission, each bit in the message is sent in sequence one at a time.

Parallel transmission is faster but it requires many wires. It is used for short distances and where speed is important. Serial transmission is slower but is less expensive.

In Asynchronous serial transfer, each bit of message is sent a sequence at a time, and binary information is transferred only when it is available. When there is no information to be transferred, line remains idle.

In this technique each character consists of three points :

1. Start bit
2. Character bit
3. Stop bit
4. Start Bit- First bit, called start bit is always zero and used to indicate the beginning character.
5. Stop Bit- Last bit, called stop bit is always one and used to indicate end of characters. Stop bit is always in the 1- state and frame the end of the characters to signify the idle or wait state.
6. Character Bit- Bits in between the start bit and the stop bit are known as character bits. The character bits always follow the start bit.



Serial Transmission of Asynchronous is done by two ways:

1. Asynchronous Communication Interface
2. First In First out Buffer

## Asynchronous Communication Interface:

It works as both a receiver and a transmitter. Its operation is initialized by CPU by sending a byte to the control register.

The transmitter register accepts a data byte from CPU through the data bus and transferred to a shift register for serial transmission.

The receive portion receives information into another shift register, and when a complete data byte is received it is transferred to receiver register.

CPU can select the receiver register to read the byte through the data bus. Data in the status register is used for input and output flags.

## First In First Out Buffer (FIFO):

A First In First Out (FIFO) Buffer is a memory unit that stores information in such a manner that the first item is in the item first out. A FIFO buffer comes with separate input and output terminals. The important feature of this buffer is that it can input data and output data at two different rates.

When placed between two units, the FIFO can accept data from the source unit at one rate, rate of transfer and deliver the data to the destination unit at another rate.

If the source is faster than the destination, the FIFO is useful for source data arrive in bursts that fills out the buffer. FIFO is useful in some applications when data are transferred asynchronously.

# Modes of Data Transfer :

Transfer of data is required between CPU and peripherals or memory or sometimes between any two devices or units of your computer system. To transfer a data from one unit to another one should be sure that both units have proper connection and at the time of data transfer the receiving unit is not busy. This data transfer with the computer is Internal Operation.

All the internal operations in a digital system are synchronized by means of clock pulses supplied by a common clock pulse Generator*.* The data transfer can be

* 1. Synchronous or
  2. Asynchronous

When both the transmitting and receiving units use same clock pulse then such a data transfer is called Synchronous process. On the other hand, if the there is not concept of clock pulses

and the sender operates at different moment than the receiver then such a data transfer is called Asynchronous data transfer.

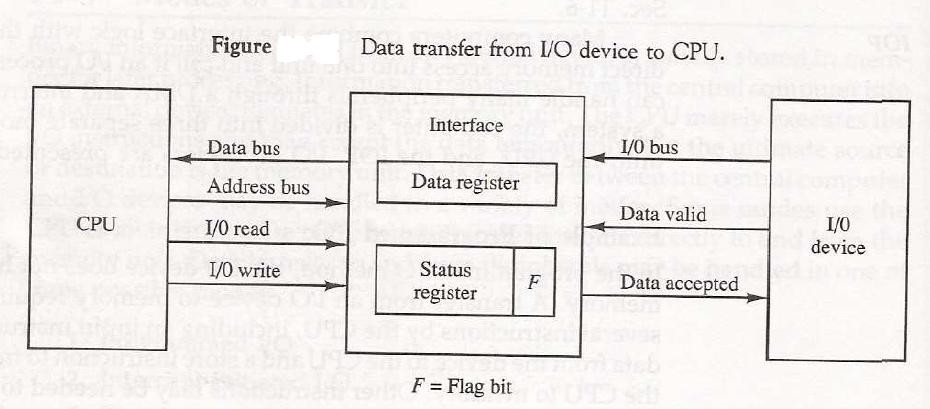
The data transfer can be handled by various modes. some of the modes use CPU as an intermediate path, others transfer the data directly to and from the memory unit and this can be handled by 3 following ways:

1. Programmed I/O
2. Interrupt-Initiated I/O
3. Direct Memory Access (DMA)

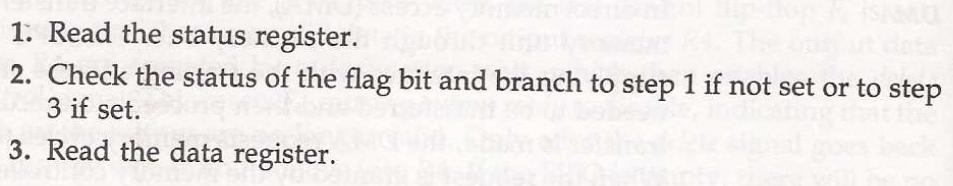
## Programmed I/O Mode:

In this mode of data transfer the operations are the results in I/O instructions which is a part of computer program. Each data transfer is initiated by a instruction in the program. Normally the transfer is from a CPU register to peripheral device or vice-versa.

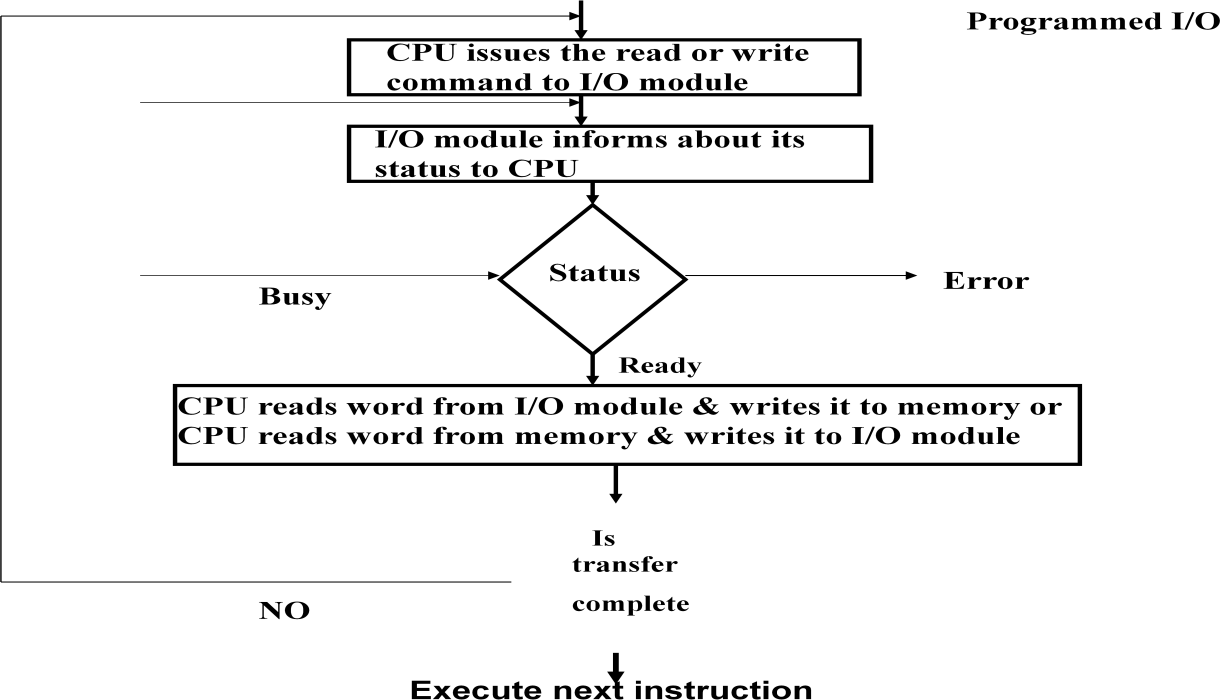
Once the data is initiated the CPU starts monitoring the interface to see when next transfer can made. The instructions of the program keep close tabs on everything that takes place in the interface unit and the I/O devices.



* The transfer of data requires three instructions:



In this technique CPU is responsible for executing data from the memory for output and storing data in memory for executing of Programmed I/O as shown in Flowchart-:



Drawback of the Programmed I/O :

The main drawback of the Program Initiated I/O was that the CPU has to monitor the units all the times when the program is executing. Thus the CPU stays in a program loop until the I/O unit indicates that it is ready for data transfer. This is a time consuming process and the CPU time is wasted a lot in keeping an eye to the executing of program.

To remove this problem an Interrupt facility and special commands are used.

## Interrupt-Initiated I/O :

In this method an interrupt facility an interrupt command is used to inform the device about the start and end of transfer. In the meantime the CPU executes other program. When the interface determines that the device is ready for data transfer it generates an Interrupt Request and sends it to the computer.

When the CPU receives such an signal, it temporarily stops the execution of the program and branches to a service program to process the I/O transfer and after completing it returns back to task, what it was originally performing.

* In this type of IO, computer does not check the flag. It continue to perform its task.
* Whenever any device wants the attention, it sends the interrupt signal to the CPU.
* CPU then deviates from what it was doing, store the return address from PC and branch to the address of the subroutine.
* There are two ways of choosing the branch address:
  + Vectored Interrupt
  + Non-vectored Interrupt
* In vectored interrupt the source that interrupt the CPU provides the branch information. This information is called interrupt vectored.
* In non-vectored interrupt, the branch address is assigned to the fixed address in the memory.

## Priority Interrupt:

* There are number of IO devices attached to the computer.
* They are all capable of generating the interrupt.
* When the interrupt is generated from more than one device, priority interrupt system is used to determine which device is to be serviced first.
* Devices with high speed transfer are given higher priority and slow devices are given lower priority.
* Establishing the priority can be done in two ways:
  + Using Software
  + Using Hardware
* A pooling procedure is used to identify highest priority in software means.

## Polling Procedure :

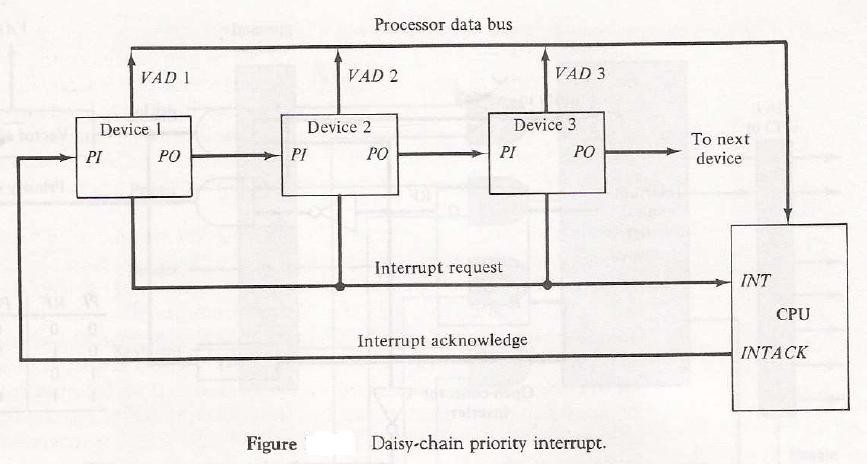
* There is one common branch address for all interrupts.
* Branch address contain the code that polls the interrupt sources in sequence. The highest priority is tested first.
* The particular service routine of the highest priority device is served.
* The disadvantage is that time required to poll them can exceed the time to serve them in large number of IO devices.

Using Hardware:

* Hardware priority system function as an overall manager.
* It accepts interrupt request and determine the priorities.
* To speed up the operation each interrupting devices has its own interrupt vector.
* No polling is required, all decision are established by hardware priority interrupt unit.
* It can be established by serial or parallel connection of interrupt lines.

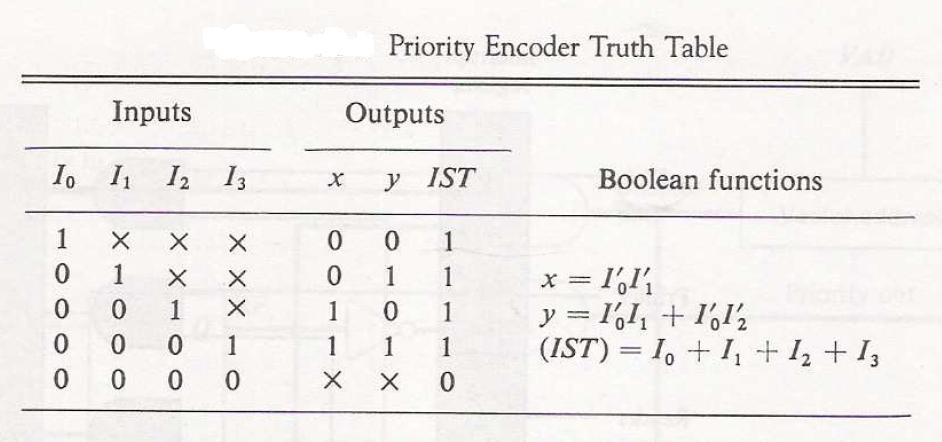
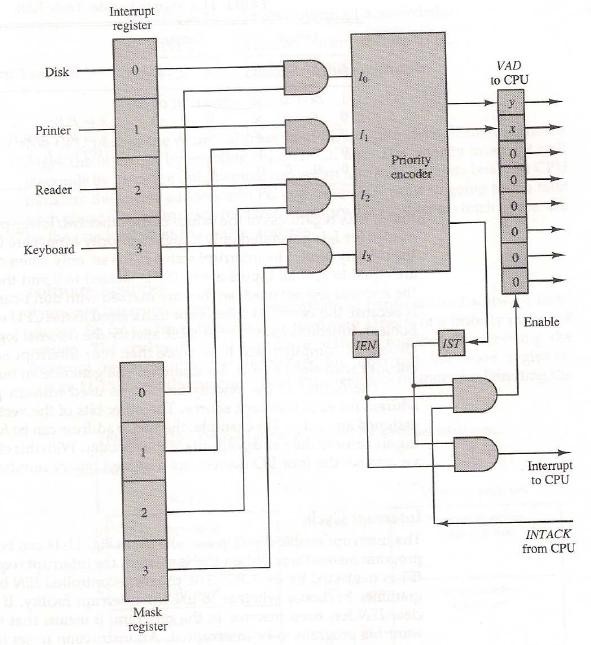
## Serial or Daisy Chaining Priority:

* Device with highest priority is placed first.
* Device that wants the attention send the interrupt request to the CPU.
* CPU then sends the INTACK signal which is applied to PI(priority in) of the first device.
* If it had requested the attention, it place its VAD(vector address) on the bus. And it block the signal by placing 0 in PO(priority out)
* If not it pass the signal to next device through PO(priority out) by placing 1.
* This process is continued until appropriate device is found.
* The device whose PI is 1 and PO is 0 is the device that send the interrupt request.

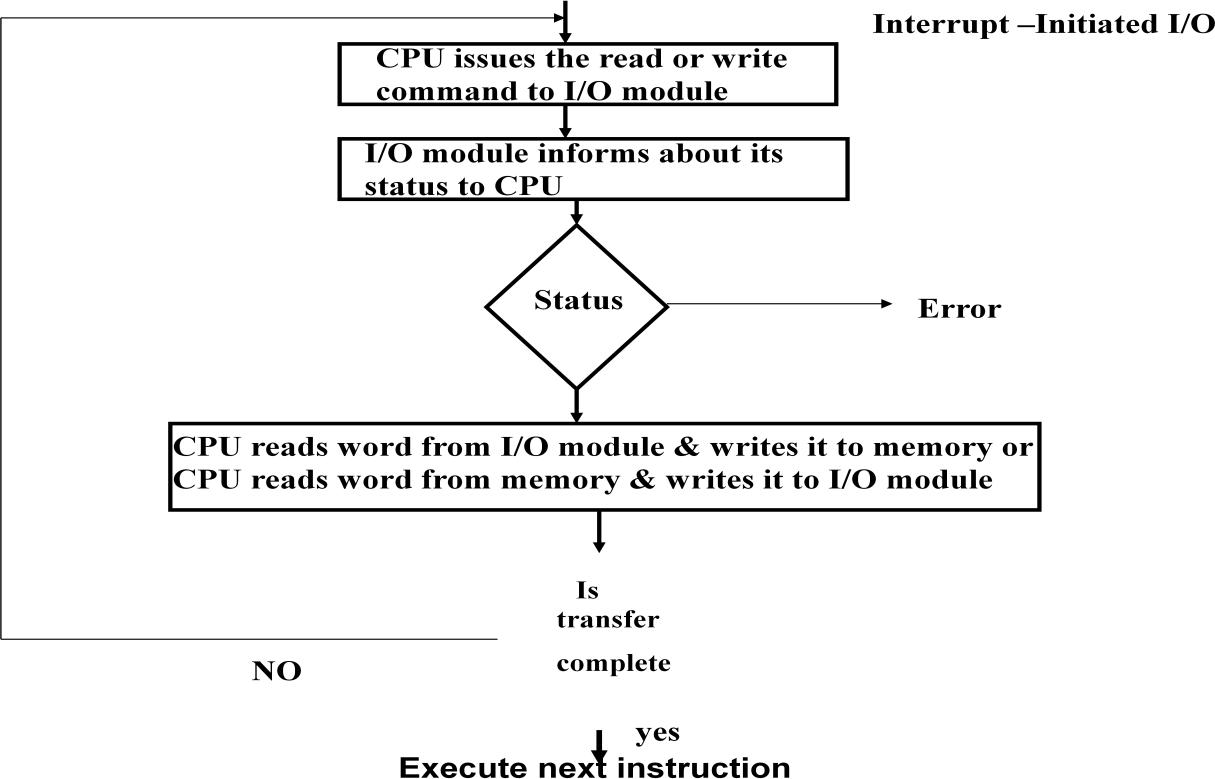


## Parallel Priority Interrupt :

* It consist of interrupt register whose bits are set separately by the interrupting devices.
* Priority is established according to the position of the bits in the register.
* Mask register is used to provide facility for the higher priority devices to interrupt when lower priority device is being serviced or disable all lower priority devices when higher is being serviced.
* Corresponding interrupt bit and mask bit are ANDed and applied to priority encoder.
* Priority encoder generates two bits of vector address.
* Another output from it sets IST(interrupt status flip flop).



The Execution process of Interrupt–Initiated I/O is represented in the flowchart:



## Direct Memory Access (DMA):

In the Direct Memory Access (DMA) the interface transfer the data into and out of the memory unit through the memory bus. The transfer of data between a fast storage device such as magnetic disk and memory is often limited by the speed of the CPU. Removing the CPU from the path and letting the peripheral device manage the memory buses directly would improve the speed of transfer. This transfer technique is called Direct Memory Access (DMA).

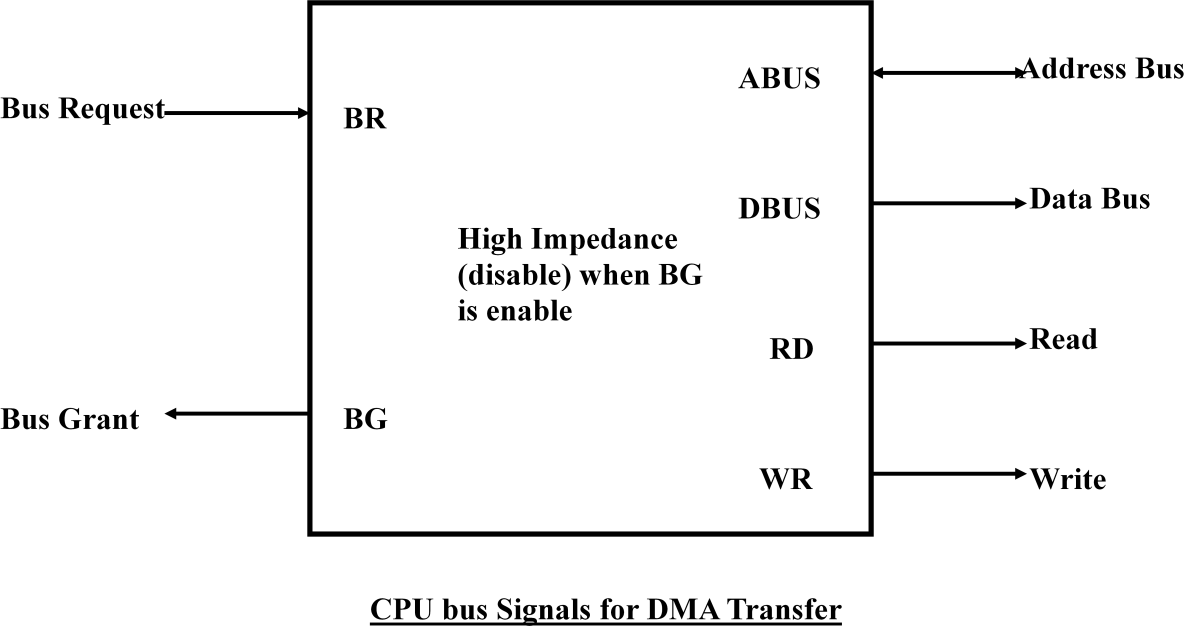
During the DMA transfer, the CPU is idle and has no control of the memory buses. A DMA Controller takes over the buses to manage the transfer directly between the I/O device and memory.

The CPU may be placed in an idle state in a variety of ways. One common method extensively used in microprocessor is to disable the buses through special control signals such as:

* Bus Request (BR)
* Bus Grant (BG)

These two control signals in the CPU that facilitates the DMA transfer. The *Bus Request (BR)* input is used by the *DMA controller* to request the CPU. When this input is active, the CPU terminates the execution of the current instruction and places the address bus, data bus

and read write lines into a *high Impedance state.* High Impedance state means that the output is disconnected.



The CPU activates the *Bus Grant (BG)* output to inform the external DMA that the Bus Request (BR) can now take control of the buses to conduct memory transfer without processor.

When the DMA terminates the transfer, it disables the *Bus Request (BR)* line. The CPU disables the *Bus Grant (BG)*, takes control of the buses and return to its normal operation.

The transfer can be made in several ways that are:

1. DMA Burst
2. Cycle Stealing
3. DMA Burst :- In DMA Burst transfer, a block sequence consisting of a number of memory words is transferred in continuous burst while the DMA controller is master of the memory buses.
4. Cycle Stealing :- Cycle stealing allows the DMA controller to transfer one data word at a time, after which it must returns control of the buses to the CPU.

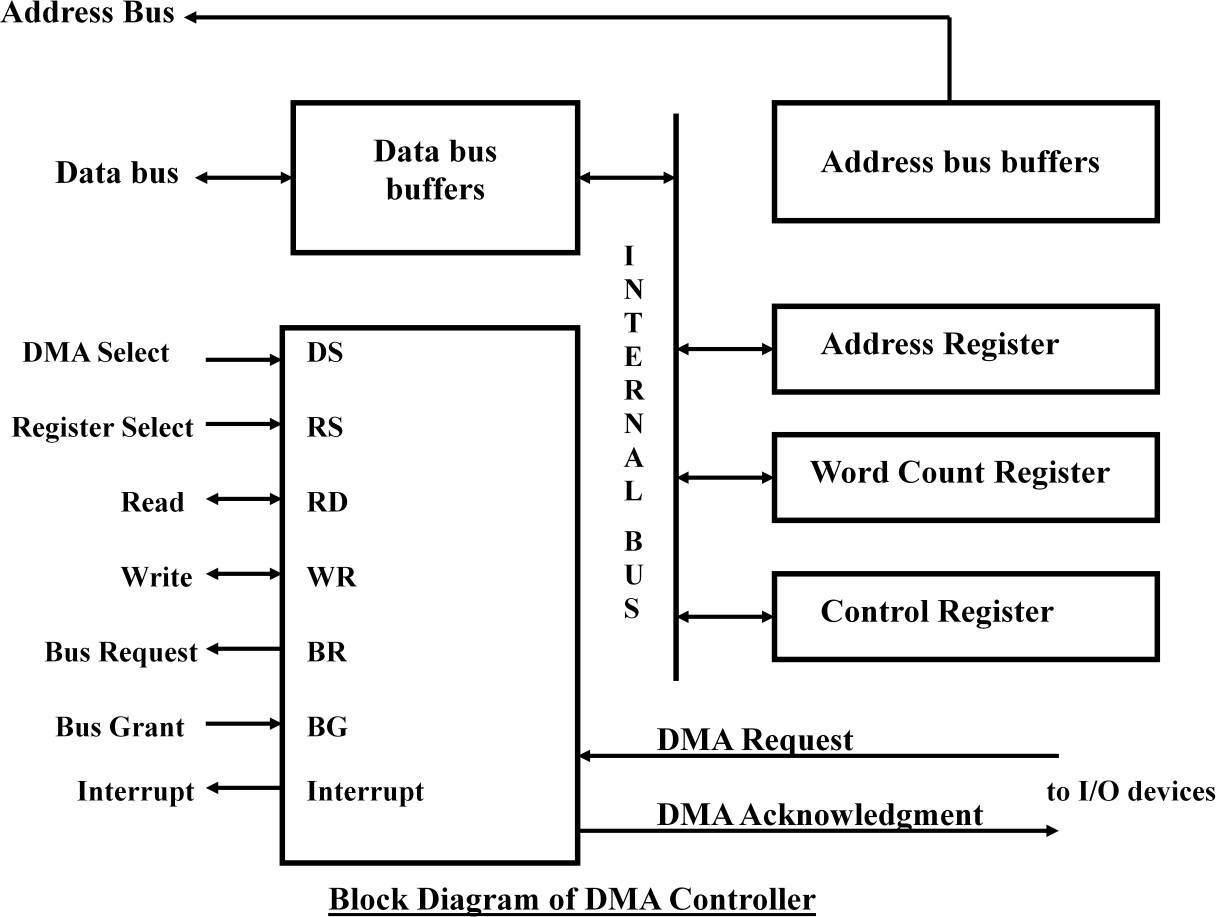
DMA Controller:

The DMA controller needs the usual circuits of an interface to communicate with the CPU and I/O device. The DMA controller has three registers:

* 1. Address Register
  2. Word Count Register
  3. Control Register

1. Address Register :- Address Register contains an address to specify the desired location in memory.
2. Word Count Register :- WC holds the number of words to be transferred. The register is incre/decre by one after each word transfer and internally tested for zero.
3. Control Register :- Control Register specifies the mode of transfer

The unit communicates with the CPU via the data bus and control lines. The registers in the DMA are selected by the CPU through the address bus by enabling the DS (DMA select) and RS (Register select) inputs. The RD (read) and WR (write) inputs are bidirectional.

When the BG (Bus Grant) input is 0, the CPU can communicate with the DMA registers through the data bus to read from or write to the DMA registers. When BG =1, the DMA can communicate directly with the memory by specifying an address in the address bus and activating the RD or WR control.

DMA Transfer:

The CPU communicates with the DMA through the address and data buses as with any interface unit. The DMA has its own address, which activates the DS and RS lines. The CPU initializes the DMA through the data bus. Once the DMA receives the start control command, it can transfer between the peripheral and the memory.

When BG = 0 the RD and WR are input lines allowing the CPU to communicate with the internal DMA registers. When BG=1, the RD and WR are output lines from the DMA controller to the random access memory to specify the read or write operation of data.

Memory Organization

* 1. **MEMORY HIERARCHY**

**Memory hierarchy in a computer system :**

Memory hierarchy system consists of all storage devices employed in a computer system from the slow but high capacity auxiliary memory to a relatively faster main memory, to an even smaller and faster cache memory accessible to the high speed processing logic.

* + - **Main Memory**: memory unit that communicates directly with the CPU (RAM)
    - **Auxiliary Memory**: device that provide backup storage (Disk Drives)
    - **Cache Memory**: special very-high-speed memory to increase the processing speed (Cache RAM)

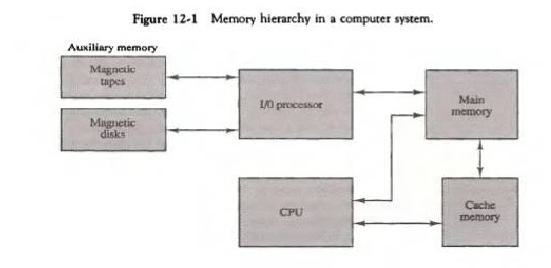


Figure12- 1 illustrates the components in a typical memory hierarchy. At the bottom of the hierarchy are the relatively slow magnetic tapes used to store removable files. Next are the Magnetic disks used as backup storage. The main memory occupies a central position by being able to communicate directly with CPU and with auxiliary memory devices through an I/O process. Program not currently needed in main memory are transferred into auxiliary memory to provide space for currently used programs and data.

The cache memory is used for storing segments of programs currently being executed in the CPU. The I/O processor manages data transfer between auxiliary memory and main memory. The auxiliary memory has a large storage capacity is relatively inexpensive, but has low access speed compared to main memory. The cache memory is very small, relatively expensive, and has very high access speed. The CPU has direct access to both cache and main memory but not to auxiliary memory.

# Multiprogramming:

Many operating systems are designed to enable the CPU to process a number of independent programs concurrently.

Multiprogramming refers to the existence of 2 or more programs in different parts of the memory hierarchy at the same time.

# Memory management System:

The part of the computer system that supervises the flow of information between auxiliary memory and main memory.

**12 – 2 MAIN MEMORY**

Main memory is the central storage unit in a computer system. It is a relatively large and fast memory used to store programs and data during the computer operation. The principal technology used for the main memory is based on semi conductor integrated circuits. Integrated circuits RAM chips are available in two possible operating modes, static and dynamic.

* Static RAM – Consists of internal flip flops that store the binary information.
* Dynamic RAM – Stores the binary information in the form of electric charges that are applied to capacitors.

Most of the main memory in a general purpose computer is made up of RAM integrated circuit chips, but a portion of the memory may be constructed with ROM chips.

* Read Only Memory –Store programs that are permanently resident in the computer and for tables of constants that do not change in value once the production of the computer is completed.

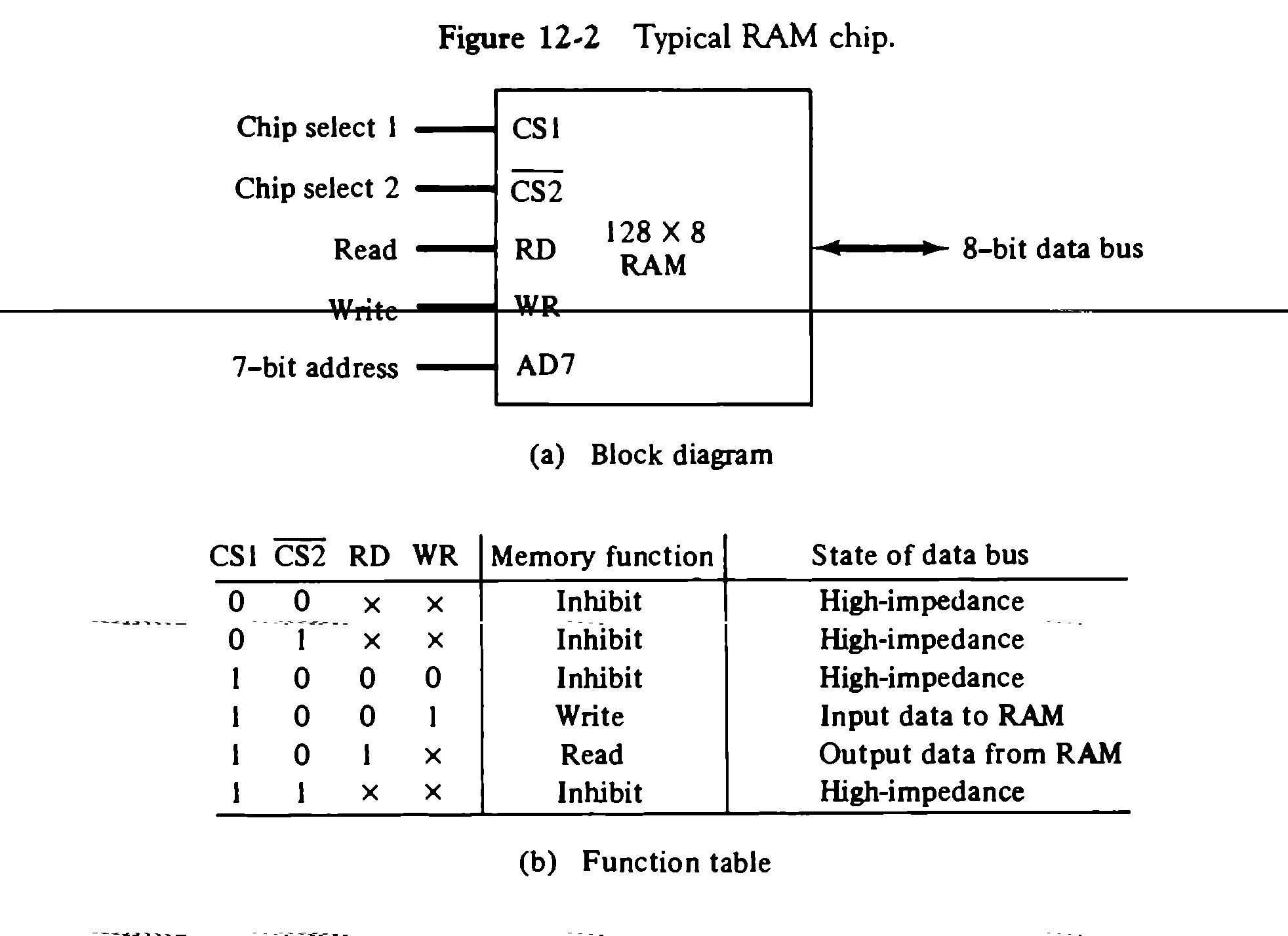
loader.

The ROM portion of main memory is needed for storing an initial program called a Bootstrap

* Boot strap loader –function is start the computer software operating when power is turned on.
* Boot strap program loads a portion of operating system from disc to main memory and control is then transferred to operating system.

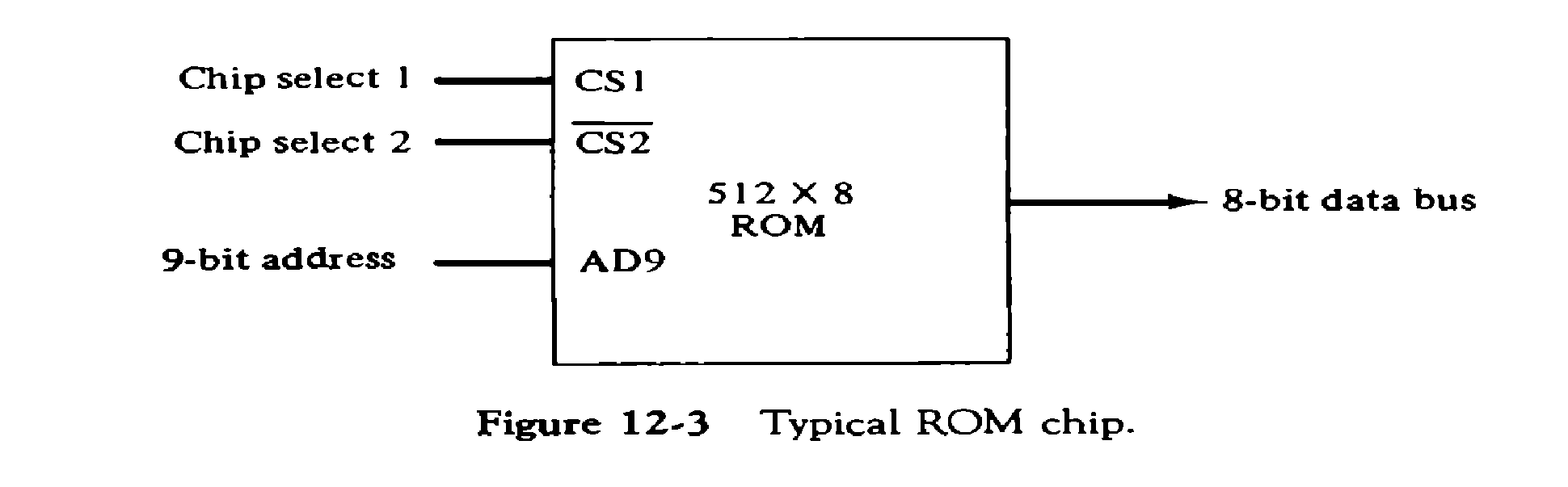
# RAM and ROM CHIP

* + RAM chip –utilizes bidirectional data bus with three state buffers to perform communication with CPU



The block diagram of a RAM Chip is shown in Fig.12-2. The capacity of memory is 128 words of eight bits (one byte) per word. This requires a 7-bit address and an 8-bit bidirectional data bus. The read and write inputs specify the memory operation and the two chips select (CS) control inputs are enabling the chip only when it is selected by the microprocessor. The read and write inputs are sometimes combined into one line labelled R/W.

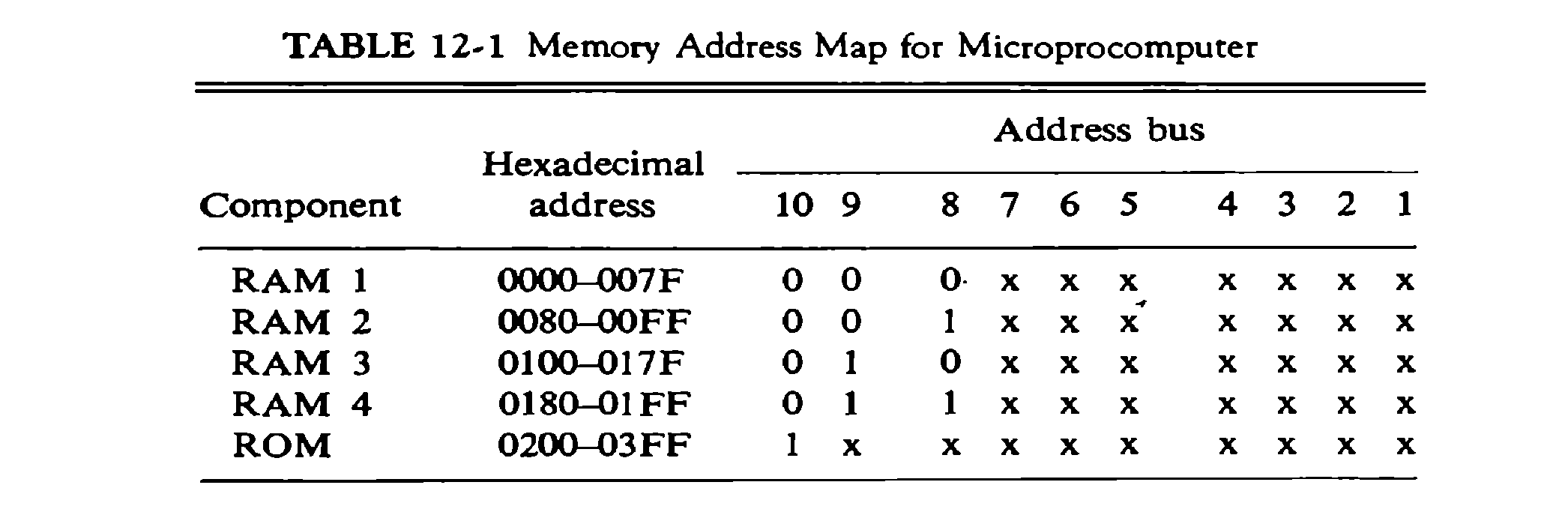
The function table listed in Fig.12-2(b) specifies the operation of the RAM chip. The unit is in operation only when CS1=1 and CS2=0.The bar on top of the second select variable indicates that this input is enabled when it is equal to 0. If the chip select inputs are not enabled, or if they are enabled but the read or write inputs are not enabled, the memory is inhibited and its data bus is in a high-impedance state. When CS1=1 and CS2=0, the memory can be placed in a write or read mode. When the WR input is enabled, the memory stores a byte from the data bus into a location specified by the address input lines. When the RD input is enabled, the content of the selected byte is placed into the data bus. The RD and WR signals control the memory operation as well as the bus buffers associated with the bidirectional data bus.



A ROM chip is organized externally in a similar manner. However, since a ROM can only read, the data bus can only be in an output mode. The block diagram of a ROM chip is shown in fig.12-3. The nine address lines in the ROM chip specify any one of the512 bytes stored in it. The two chip select inputs must be CS1=1 and CS2=0 for the unit to operate. Otherwise, the da~~ta bu~~s is in a high-impedance state.

# Memory Address Map

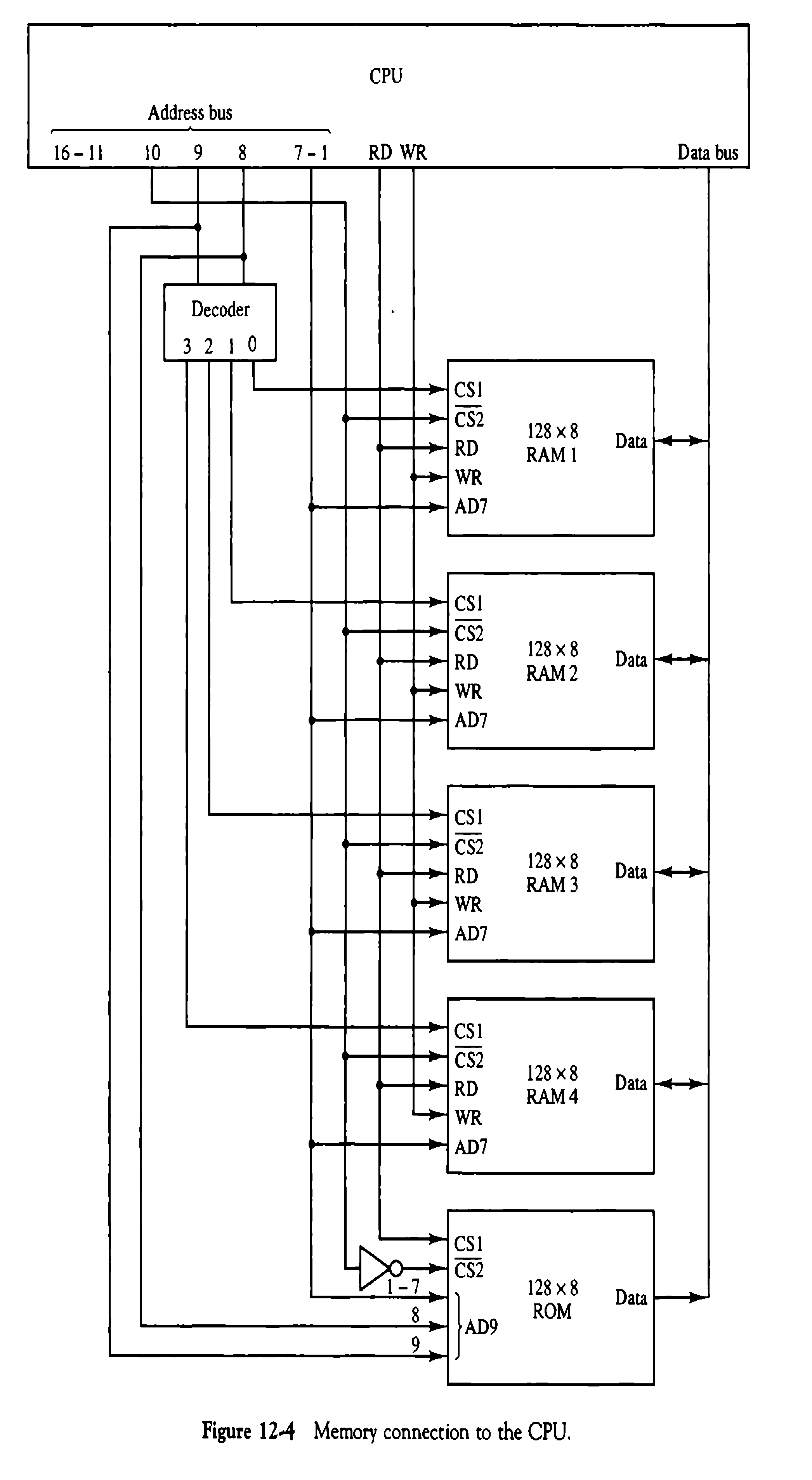
The interconnection between memory and processor is then established from knowledge of the size of memory needed and the type of RAM and ROM chips available. The addressing of memory can be established by means of a table that specify the memory address assigned to each chip. The table called Memory address map, is a pictorial representation of assigned address space for each chip in the system.



The memory address map for this configuration is shown in table 12-1. The component column specifies whether a RAM or a ROM chip is used. The hexadecimal address column assigns a range of hexadecimal equivalent addresses for each chip. The address bus lines are listed in the third column. The RAM chips have 128 bytes and need seven address lines. The ROM chip has 512 bytes and needs 9 address lines.

# Memory Connection to CPU:

RAM and ROM chips are connected to a CPU through the data and address buses. The low order lines in the address bus select the byte within the chips and other lines in the address bus select a particular chip through its chip select inputs.



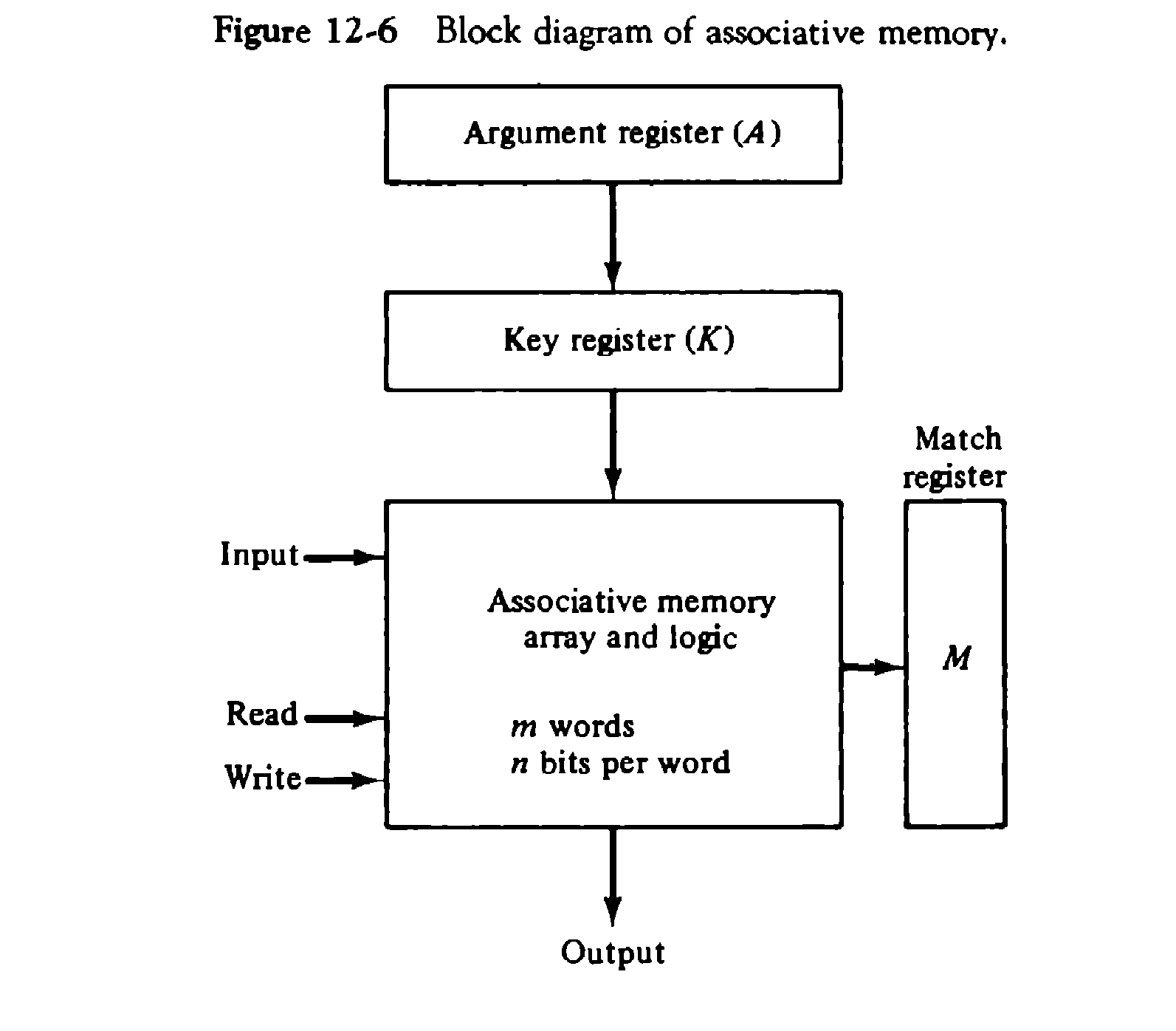
The connection of memory chips to the CPU is shown in Fig.12-4. This configuration gives a memory capacity of 512 bytes of RAM and 512 bytes of ROM. Each RAM receives the seven low-order bits of the address bus to select one of 128 possible bytes. The particular RAM chip selected is determined from lines 8 and 9 in the address bus. This is done through a 2 X 4 decoder whose outputs go to the CS1 inputs in each RAM chip. Thus, when address lines 8 and 9 are equal to 00, the first RAM chip is selected. When 01, the second RAM chip is select, and so on. The RD and WR outputs from the microprocessor are applied to the inputs of each RAM chip. The selection between RAM and ROM is achieved through bus line 10. The RAMs are selected when the bit in this line is 0, and the ROM when the bit is 1. Address bus lines 1 to 9 are applied to the input address of ROM without going through the decoder. The data bus of the ROM has only an output capability, whereas the data bus connected to the RAMs can transfer information in both directions.

* 1. **AUXILIARY MEMORY**

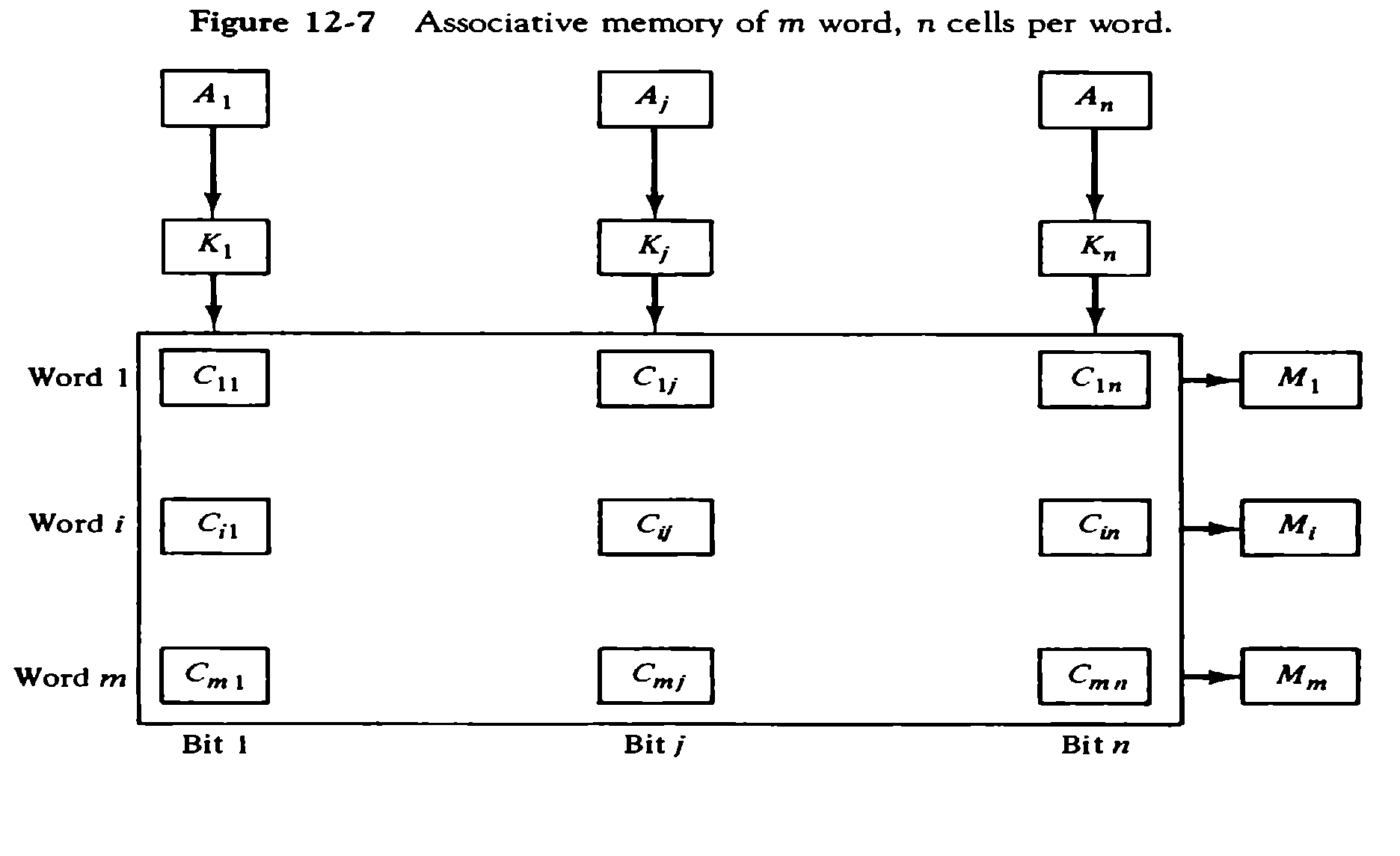
The time required to find an item stored in memory can be reduced considerably if stored data can be identified for access by the content of the data itself rather than by an address. A memory unit accessed by content is called an associative memory or content addressable memory (CAM).

* + - CAM is accessed simultaneously and in parallel on the basis of data content rather than by specific address or location
    - Associative memory is more expensive than a RAM because each cell must have storage capability as well as logic circuits
    - Argument register –holds an external argument for content matching
    - Key register –mask for choosing a particular field or key in the argument word

# Hardware Organization



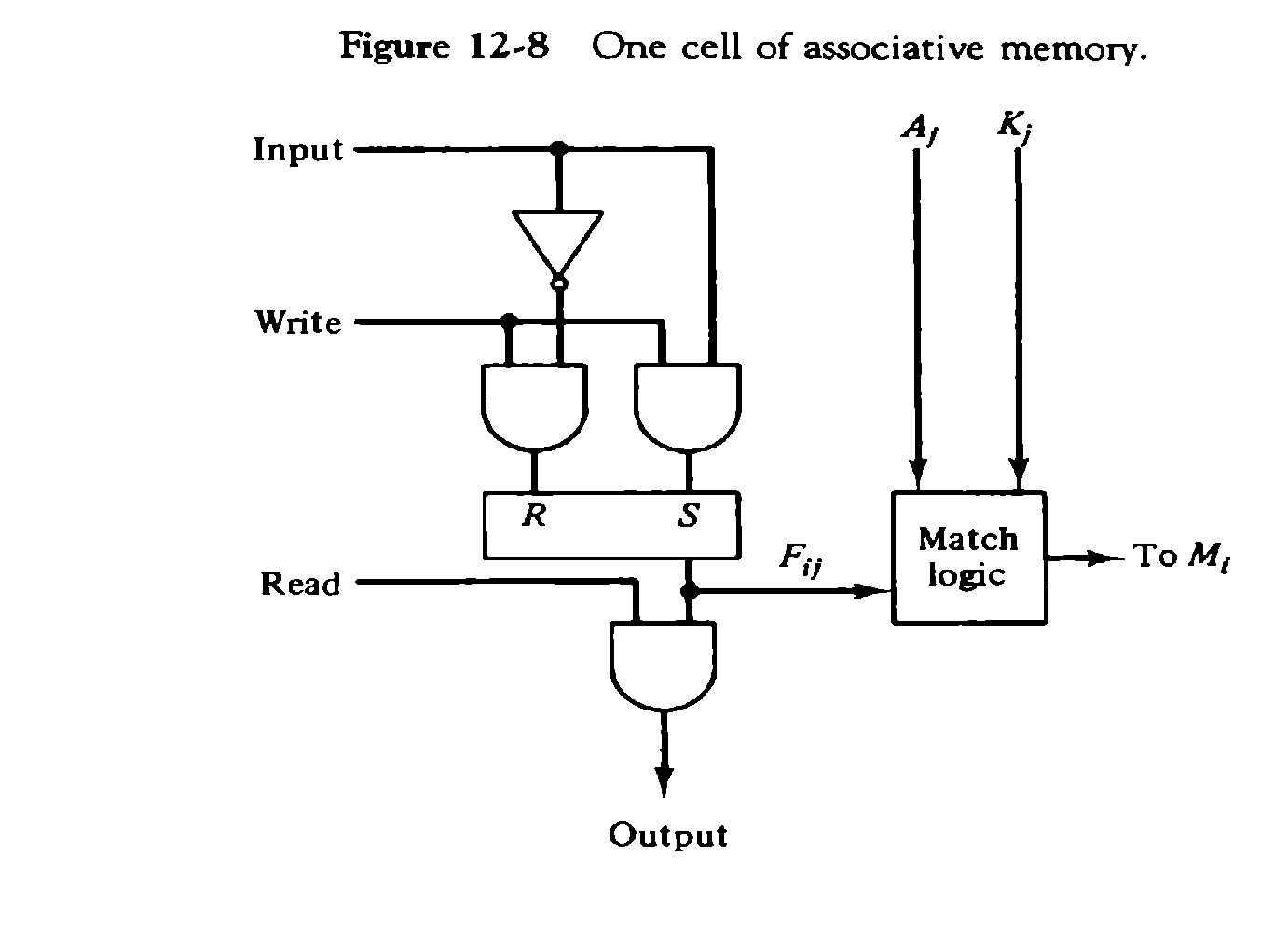
It consists of a memory array and logic for m words with n bits per word. The argument register A and key register K each have n bits, one for each bit of a word. The match register M has m bits, one for each memory word. Each word in memory is compared in parallel with the content of the argument register. The words that match the bits of the argument register set a corresponding bit in the match register. After the matching process, those bits in the match register that have been set bindicate the fact that their corresponding words have been matched. Reading is accomplished by a sequential access to memory for those words whose corresponding bits in the match register have been set.



The relation between the memory array and external registers in an associative memory is shown in Fig.12-7. The cells in the array are marked by the letter C with two subscripts. The first subscript gives the word number and second specifies the bit position in the word. Thus cell Cij is the cell for bit j in word

1. A bit Aj in the argument register is compared with all the bits in column j of the array provided that kj

=1.This is done for all columns j=1,2,….n. If a match occurs between all the unmasked bits of the argument and the bits in word I, the corresponding bit Mi in the match register is set to 1. If one or more unmasked bits of the argument and the word do not match, Mi is cleared to 0.



It consists of flip-flop storage element Fij and the circuits for reading, writing, and matching the cell. The input bit is transferred into the storage cell during a write operation. The bit stored is read out during a read operation. The match logic compares the content of the storage cell with corresponding unmasked bit of the argument and provides an output for the decision logic that sets the bit in Mi.

# Match Logic

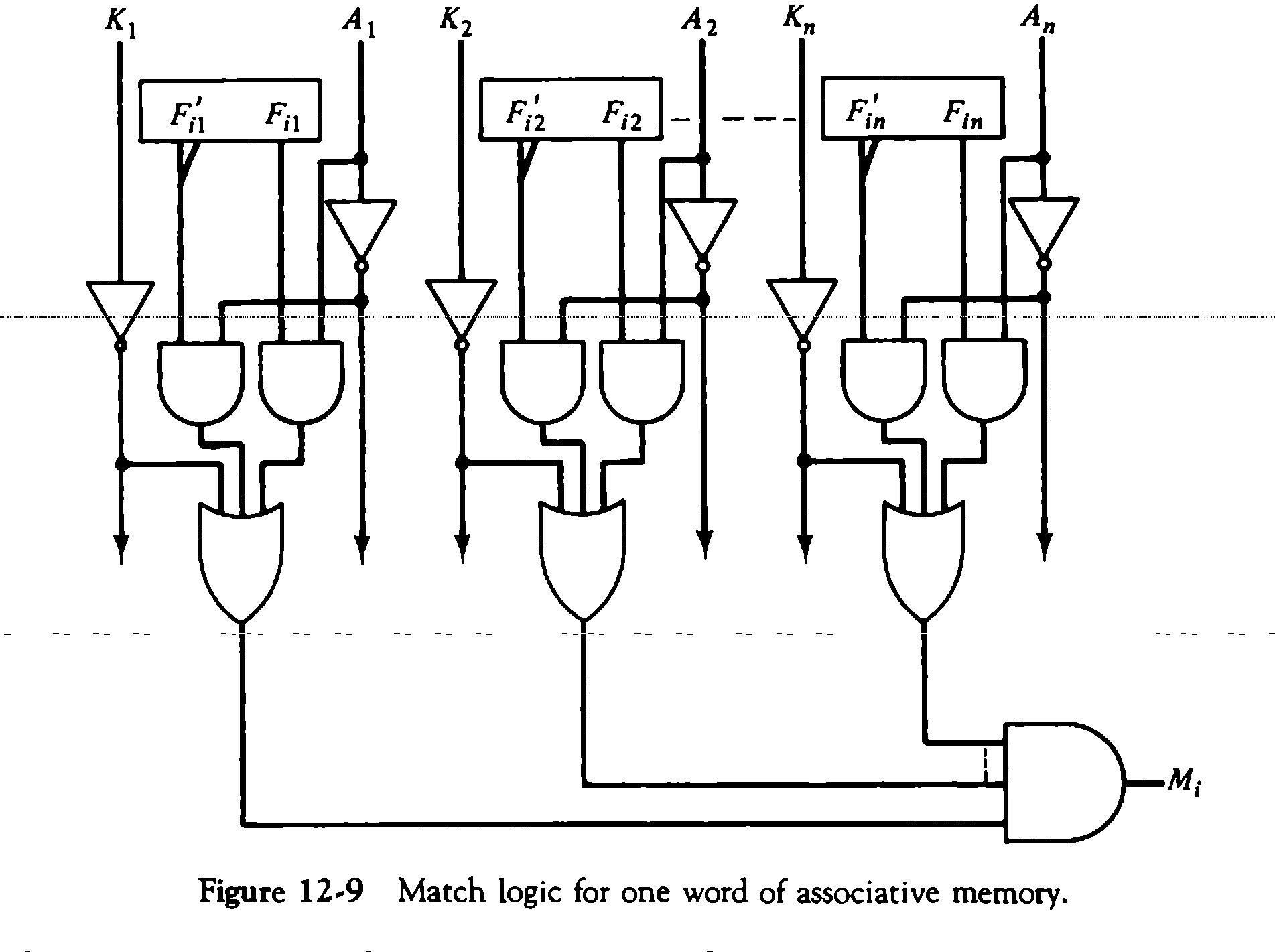
The match logic for each word can be derived from the comparison algorithm for two binary numbers. First, neglect the key bits and compare the argument in A with the bits stored in the cells of the words.

Word i is equal to the argument in A if Aj=F ijfor j=1,2,…..,n. Two bits are equal if they are both 1 or both 0. The equality of two bits can be expressed logically by the Boolean function

xj=Aj Fij + Aj ‘Fij ‘

where xj = 1 if the pair of bits in position j are equal;otherwise , xj =0. For a word i is equal to the argument in A we must have all xj variables equal to 1. This is the condition for setting the corresponding match bit Mi to 1. The Boolean function for this condition is

Mi = x1 x2 x3…… xn

E 

ach cell requires two AND gate and one OR gate. The inverters for A and K are needed once for each column and are used for all bits in the column. The output of all OR gates in the cells of the same word go to the input of a common AND gate to generate the match signal for Mi. Mi will be logic 1 if a match occurs and 0 if no match occurs.

# Read and Write operation Read Operation

* If more than one word in memory matches the unmasked argument field , all the matched words will have 1’s in the corresponding bit position of the match register
* In read operation all matched words are read in sequence by applying a read signal to each word line whose corresponding Mi bit is a logic 1
* In applications where no two identical items are stored in the memory , only one word may match

, in which case we can use Mi output directly as a read signal for the corresponding word

# Write Operation

Can take two different forms

* 1. Entire memory may be loaded with new information 2.Unwanted words to be deleted and new words to be inserted

1. Entire memory : writing can be done by addressing each location in sequence – This makes it random access memory for writing and content addressable memory for reading – number of lines needed for decoding is d Where m = 2 d , m is number of words.
2. Unwanted words to be deleted and new words to be inserted :
   * Tag register is used which has as many bits as there are words in memory
   * For every active ( valid ) word in memory , the corresponding bit in tag register is set to 1
   * When word is deleted the corresponding tag bit is reset to 0
   * The word is stored in the memory by scanning the tag register until the first 0 bit is encountered After storing the word the bit is set to 1.
   1. **CACHE MEMORY**

* Effectiveness of cache mechanism is based on a property of computer programs called **“locality of reference”**
* The references to memory at any given time interval tend to be confined within a localized areas
* Analysis of programs shows that most of their execution time is spent on routines in which instructions are executed repeatedly These instructions may be – loops, nested loops , or few procedures that call each other
* Many instructions in localized areas of program are executed
* repeatedly during some time period and reminder of the program is accessed infrequently This property is called “Locality of Reference”.

# Locality of Reference

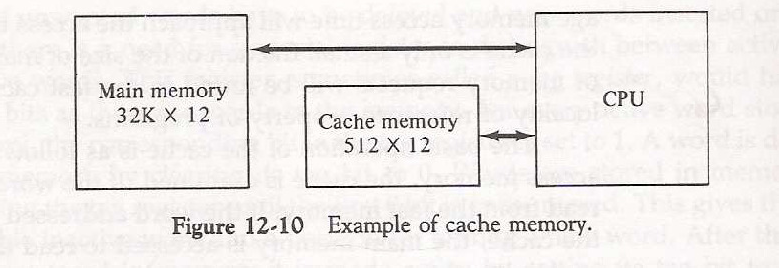
Locality of reference is manifested in two ways :

1. Temporal- means that a recently executed instruction is likely to be executed again very soon.
   * The information which will be used in near future is likely to be in use already( e.g. reuse of information in loops)
2. Spatial- means that instructions in close proximity to a recently executed instruction are also likely to be executed soon
   * If a word is accessed, adjacent (near) words are likely to be accessed soon ( e.g. related data items (arrays) are usually stored together; instructions are executed sequentially

)

1. If active segments of a program can be placed in afast (cache) memory , then total execution time can be reduced significantly
2. Temporal Locality of Reference suggests whenever an information (instruction or data) is needed first , this item should be brought in to cache
3. Spatial aspect of Locality of Reference suggests that instead of bringing just one item from the main memory to the cache ,it is wise to bring several items that reside at adjacent addresses as well ( ie a block of information )

# Principles of cache



The main memory can store 32k words of 12 bits each. The cache is capable of storing 512 of these words at any given time. For every word stored , there is a duplicate copy in main memory. The Cpu communicates with both memories. It first sends a 15 bit address to cahache. If there is a hit, the CPU accepts the 12 bit data from cache. If there is a miss, the CPU reads the word from main memory and the word is then transferred to cache.

* When a read request is received from CPU,contents of a block of memory words containing the location specified are transferred in to cache
* When the program references any of the locations in this block , the contents are read from the cache Number of blocks in cache is smaller than number of blocks in main memory
* Correspondence between main memory blocks and those in the cache is specified by a mapping function
* Assume cache is full and memory word not in cache is referenced
* Control hardware decides which block from cache is to be removed to create space for new block containing referenced word from memory
* Collection of rules for making this decision is called **“Replacement algorithm ” Read/ Write operations on cache**

# Cache Hit Operation

* + CPU issues Read/Write requests using addresses that refer to locations in main memory
  + Cache control circuitry determines whether requested word currently exists in cache
  + If it does, Read/Write operation is performed on the appropriate location in cache **(Read/Write Hit )**

# Read/Write operations on cache in case of Hit

* + In Read operation main memory is not involved.
  + In Write operation two things can happen.

1. Cache and main memory locations are updated simultaneously **(“ Write Through ”)** OR
2. Update only cache location and mark it as “ Dirty or Modified Bit ” and update main memory location at the time of cache block removal **(“ Write Back ” or “ Copy Back ”)** .

# Read/Write operations on cache in case of Miss Read Operation

* + When addressed word is not in cache Read Miss occurs there are two ways this can be dealt with

1. Entire block of words that contain the requested word is copied from main memory to cache and the particular word requested is forwarded to CPU from the cache **( Load Through )** (OR)
2. The requested word from memory is sent to CPU first and then the cache is updated **( Early Restart )**

# Write Operation

* + If addressed word is not in cache Write Miss occurs
  + If write through protocol is used information is directly written in to main memory
  + In write back protocol , block containing the word is first brought in to cache , the desired word is then overwritten.