classfellow
DATE : / 20
PAGE No

I/O Versus Memory Bus :

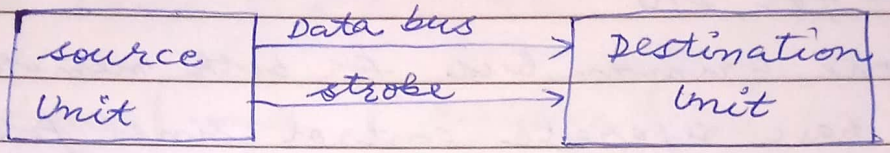
3 ways that computer buses can be used to communicate with memory and I/O

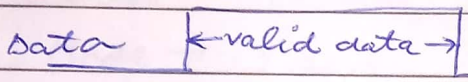
1. Use 2 separate buses, one for memory and other for I/O
2. Use one common bus for both memory and I/O but have separate control lines for each.
3. Use one common bus for memory and I/O with common control lines.

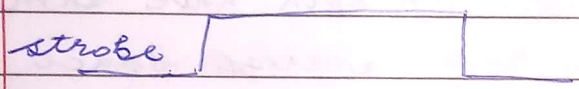
Isolated I/O	Memory Mapped I/O
1. Memory and I/O have separate address space	1. Both have same address space
2. All address can be used by memory	2. Due to addition of I/O addressable memory become less for memory
3. separate instruction control read and write operation in I/O and memory	3. Same instruction
4. In this I/O address are called ports	4. Normal memory address are for both
5. More efficient due to separate buses	5. Lesser efficient
6. Larger in size due to more buses	6. Smaller in size
7. It is complex due to separate logic is used to control both	7. Simpler logic is used as I/O is also treated as memory only.

Asynchronous data Transfer:

1. stroke

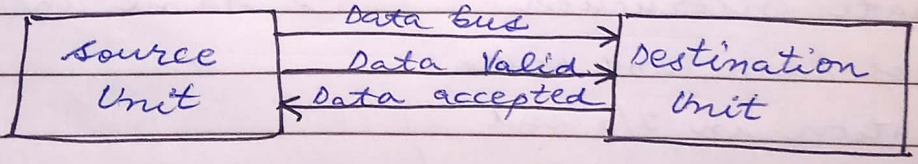


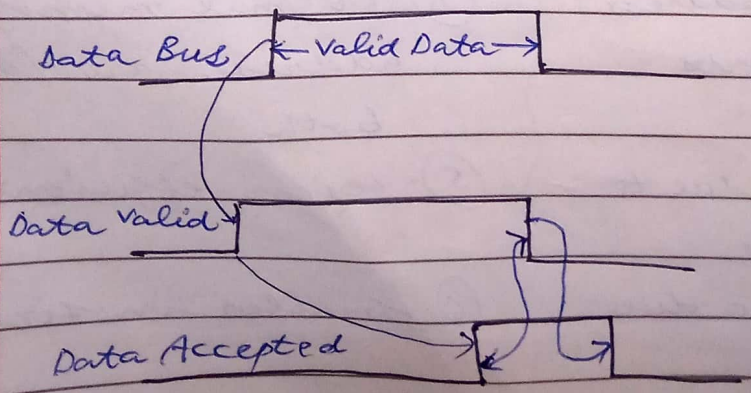




source initiated strobe for Data Transfer

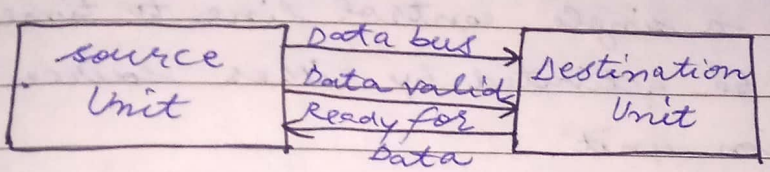
Handshaking: (source-initiated)

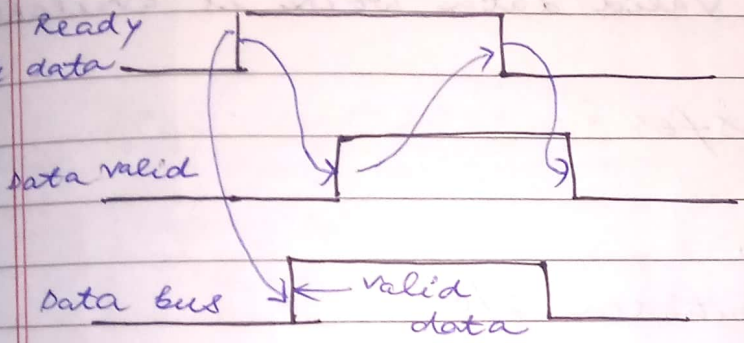




Timing Diagram

Destination Initiated :





Timing diagram

Synchronous → registers share common clock with CPU registers

asynchronous → Internal timings for each unit is independent from other and each uses its own private clock

strobe pulse supplied by one of the units to indicate to the other unit when transfer has to occur.

→ Accompany each data item being transferred with a control signal that indicates presence of data in bus. The unit receiving the data responds with another control signal to acknowledge receipt of data. This agreement is referred to as handshaking

Strobe control:

employs a single control line to time each other. It may be activated by either source or destination unit.

strobe is a single line that informs the destination unit when a valid data word is available.

Modes of Transfer:

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

Example of Programmed I/O:

F = flag bit

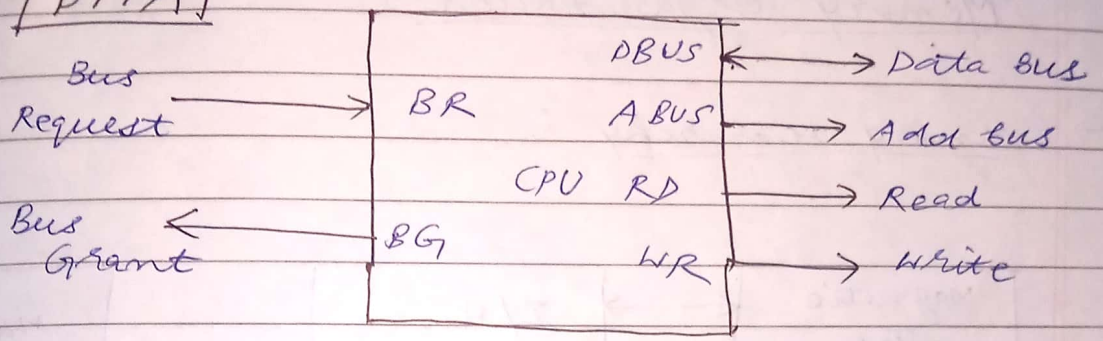
Data Transfer from I/O device to CPU

Interrupt - initiated I/O:

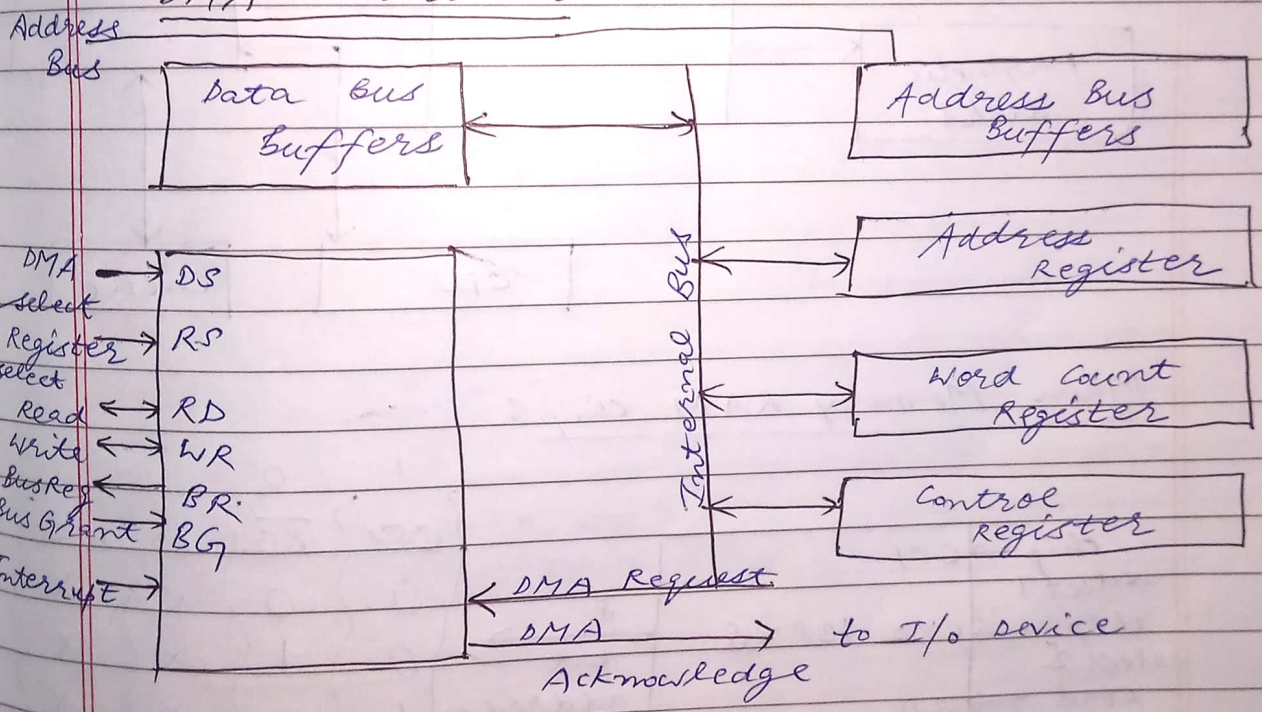
(1) Non-Vectored → branch address is assigned to fixed location in memory.

(2) Vectored → the source that interrupts supplies the branch information. This info is called interrupt vector.

DMA



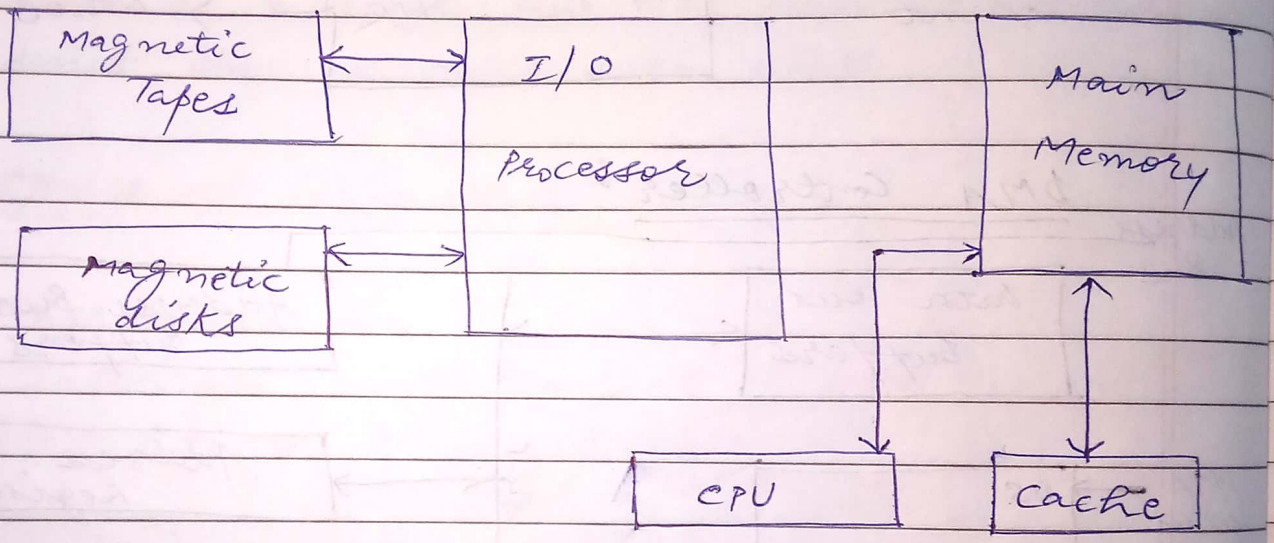
DMA Controller:



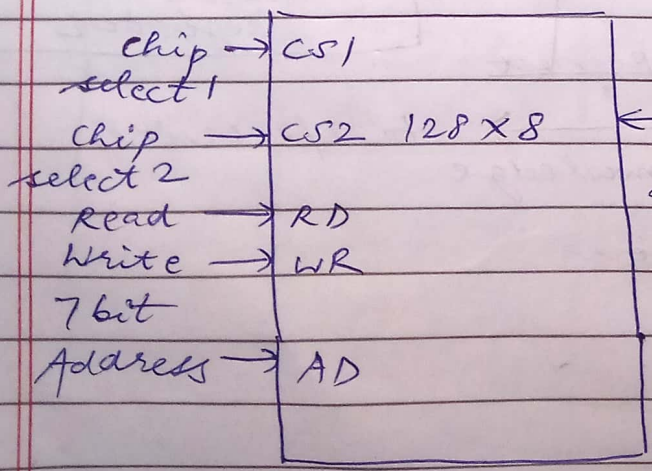
I/O Communication

Memory Organization :

→ ## Memory Hierarchy :



Main Memory RAM chip :



1	0		
CS1	$\overline{CS2}$	RD	WR
0	0	x	x
0	1	x	x
1	0	x	x
1	0	0	1
1	0	1	x
1	1	x	x
∄			

[Working of chip]

F

classfollow
DATE : / /20
PAGE No

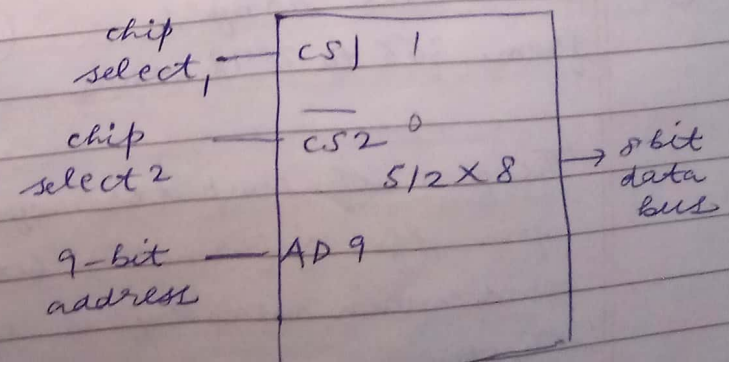
Memory	Address Hexa-decimal Address	Address Map 10 9 8 7 6 5 4 3 2 1
RAM 1	0000–007F	0 0 0 x x x x x x x
RAM 2	0080–00FF	0 0 1 x x x x x x x
RAM 3	0100–017F	0 1 0 x x x x x x x
RAM 4	0180–01FF	0 1 1 x x x x x x x
ROM	0200–03FF	1 x x x x x x x x x

(represents address space for each chip)

512 bytes of RAM
128 x 8 RAM
$\frac{512 \times 8}{128 \times 8}$ = 4 RAM

512 bytes of ROM
512 x 8 RAM
= 1 ROM

Memory function	State of bus
Inhibit	High Impedance
Inhibit	"
Inhibit	"
Write	Input data to RAM
Read	Output data from RAM
Inhibit	High impedance



classfellow
DATE : / 20
PAGE No

→ Memory unit communicates directly with CPU is called Main Memory. Devices that provides backup storage are called auxiliary memory.

→ Magnetic tapes are used to store removable files.
Magnetic disks are used as backup storage

→ Cache is used for storing segments of programs currently being executed in CPU and temporary data frequently needed in present calculations. Cache increases performance rate.

→ I/O processor manages data transfers b/w auxiliary memory and main memory.

Memory access speed increases, relative cost increases.

	Auxiliary Memory	Cache Memory
(1)	Holds parts of data that are not presently used by CPU	(1) holds parts of program and data heavily used
(2)	Access time → 100 ns	(2) → 1 to 16 words
(3)	Block size ranges from 256 to 2048 words	(3)

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

Main Memory:

central storage unit, fast memory, relatively large.

RAM chips

- static: consists of internal flip flops, easier to use
- dynamic: stores binary info in form of electric charges applied to capacitors, capacitors are recharged by refreshing dynamic memory, it offers reduced power consumption and larger storage capacity.

Main memory consists of RAM and ROM

- RAM: used for storing bulks of program, volatile
- ROM: store programs that are permanent resident

ROM portion of main memory is needed for storing an initial program called a bootstrap loader.

↓

start computer OS

RAM chips:

used for communication. It has 1 or more control inputs that select the chip only when needed. It has bidirectional bus which allows the transfer of data either from memory to CPU during a read operation, or from CPU to memory during write operation. 3-state buffer output placed whose values are 0, 1, or a high impedance state.

7 bit address and 8 bit bidirectional data bus

Auxiliary Memory:

characteristics → transfer rate, access mode, access time, capacity, cost

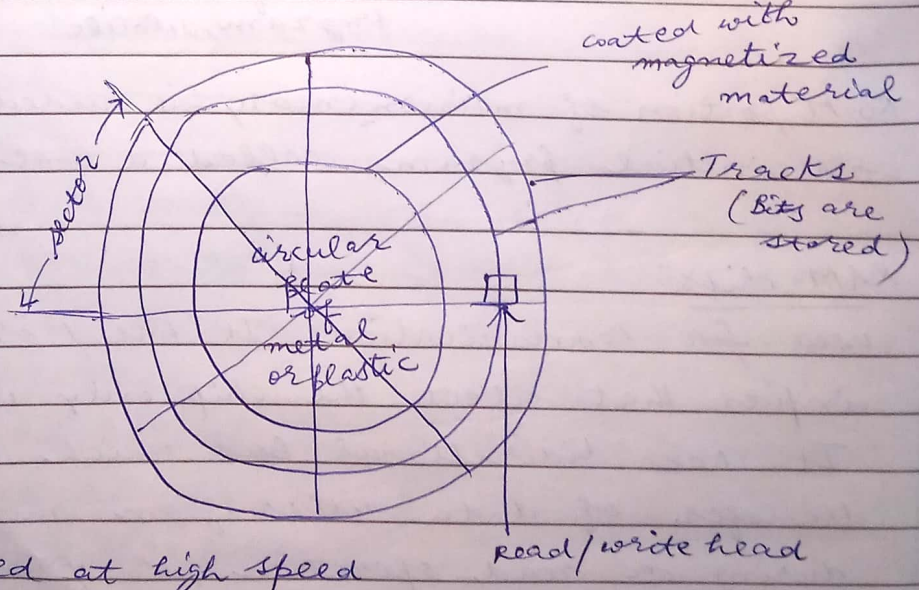
→ Average time to reach a storage location in memory and obtain its contents is called access time.

→ seek time required to position read-write head to a location

→ Transfer rate is number of characters or words that device can transfer per second.

Auxiliary Memory is organized in blocks.

Magnetic Disks:



Rotated at high speed

Tracks are divided into sections called sectors.

Track Address bits move read write head

(Read from book)

Magnetic Tape

classfollow
DATE : / 20
PAGE No

Asynchronous Data Transfer :

Asynchronous data Transfer b/w 2 independent units requires the control signals be transmitted b/w communicating units to indicate time at which data is being transmitted.

1. strobe pulse supplied by one unit to other to indicate transfer
2. accompany each data item with a control signal

strobe control :

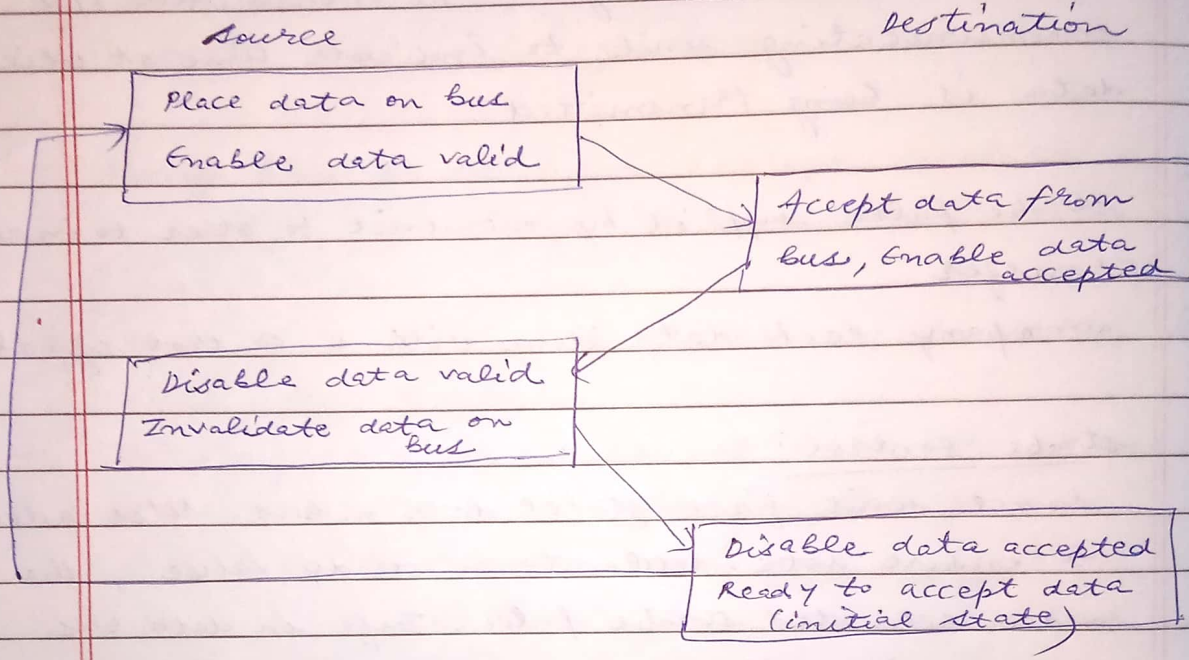
Source unit first places data on bus. After a delay to ensure data settle to a steady value, the source activates strobe pulse. Info in data bus remains in active state.

Destination unit uses falling edge to indicate the transfer of contents into one of its internal registers.

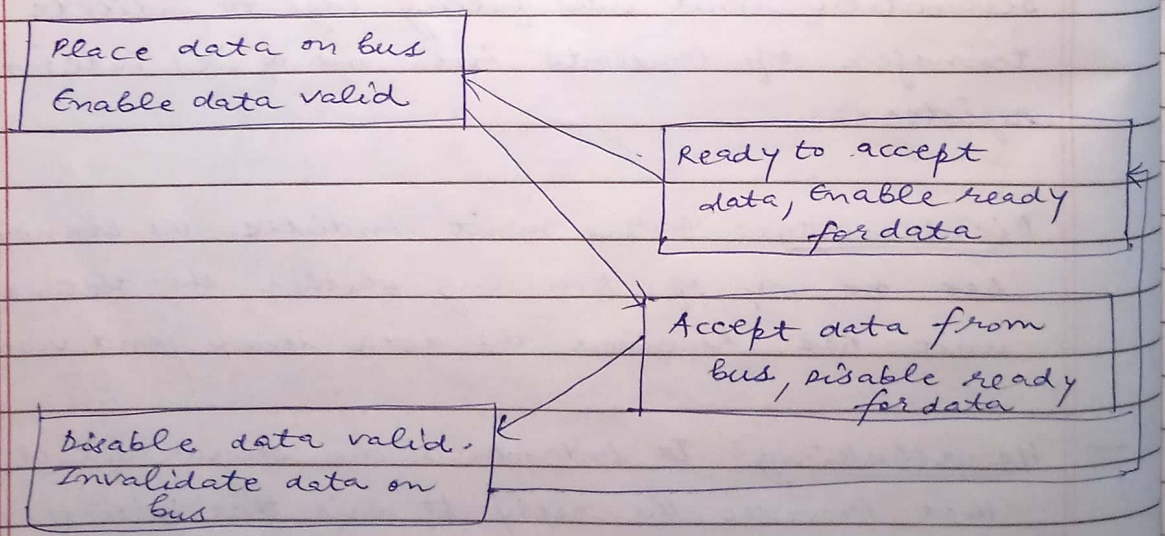
Disadvantage: source unit initiates the transfer has no way of knowing whether the destination unit has received the data item and vice versa.

→ Handshaking: It introduces the second control signal that provides the reply to unit that initiates the transfer.

Source initiated transfer using Handshaking:



Destination initiated transfer using Handshaking



Timeout Mechanism: produces an alarm if data transfer is not completed within a predetermined time.

Modes of Transfer:

(1) Programmed I/O: Each data item initiated by an instruction. Transferring data under program control requires constant monitoring of peripheral by CPU. Once data transfer is initiated, CPU monitors interface to see when a transfer can again be made. CPU stays in a program until I/O unit indicates that it is ready for data Transfer.

(2) I/O initiated: Interrupt facility is used to inform the interface to issue an interrupt signal when the data are available from device. In meantime CPU execute another program. When interface determines that device is ready for data transfer it generates a interrupt request to the computer.

(3) DMA (Direct Memory Access): Interface transfers data in and out of the memory unit through memory bus. CPU initiates the transfer by suppling the interface with starting address and number of words need to be transferred and execute other tasks. When the transfer is made, DMA requests memory cycles through memory bus. When the request is granted by memory controller, the DMA transfers data directly into memory.

CPU can be idle in a variety of ways. Two control signals in CPU that facilitate the data DMA transfer.

classfollow
DATE : / 20
PAGE No

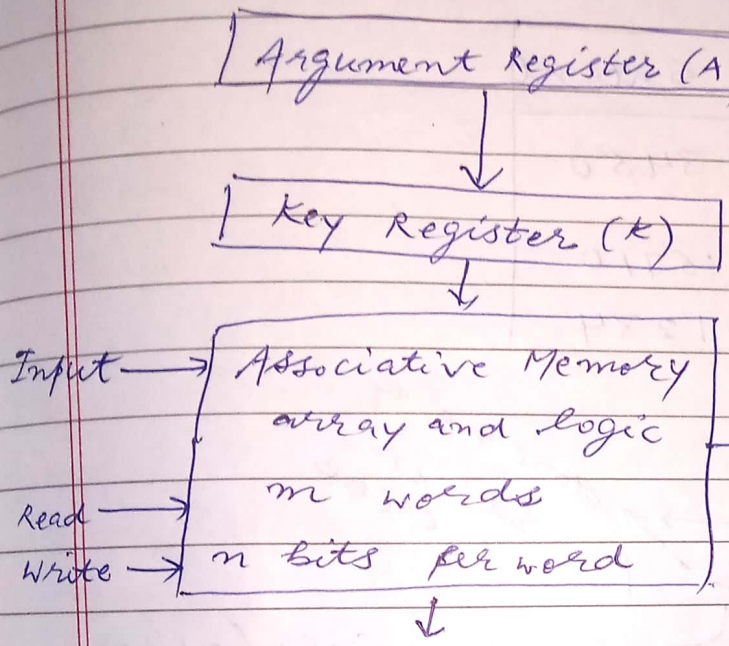
input ① BR (Bus Request) : Used by DMA controller to request the CPU to give control of buses. When this input is active CPU terminates the current instruction and places the data bus, address bus and read-write lines in high impedance state.

output ② Bus Grant (BR) : CPU activates this to inform the external DMA that the buses are in high-impedance state. DMA that originates the bus request can now take control of buses to conduct memory transfers. When DMA terminates the transfer, it disables the bus request line. CPU disables the bus grant, take control of buses and returns to its normal execution.

Burst Transfer : A block sequence consisting of a number of memory words is transferred in a continuous burst while DMA controller is the master of memory buses.

Cycle stealing : Allows the DMA controller to transfer one data word at a time after which it must return control of buses to CPU.

Associative Memory :



A = 10111111 10
k = 11100000 00
word 1 = 10011110 0
word 2 = 10100000 1
word 3 = 10011111 1
word 4 = 10100001

word 1 = no match → means 0
word 2 = match → 1
word 3 = no match → 0
word 4 = match → 1

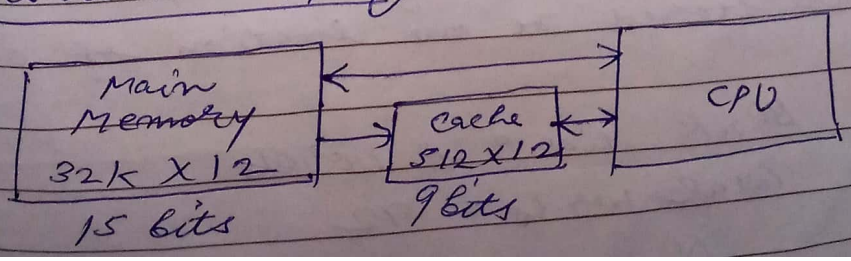
match means 1

Cache Memory : Hit Ratio = $\frac{\text{Hit}}{\text{Hit + Miss}}$

Mapping : Transfer data from main memory to cache memory.

3 Types.

1. **Associative Mapping :**



CPU address → Argument Register

Address	Content
01000	3450
02777	6710
22345	1234

5 digits octal address

(2) Direct Mapping :

store no. of bits in cache

← 15 bits →

6 bits	Tag	Index

Cache (9 bits)

Main Memory

00000	1220
	2340
02000	5670
02777	5670

Index	Tag	data
000	00	1220
777	02	5670

Disad → More excess time when index is same

(3) set associative Mapping : 2 or more pairs of tags are stored at one location of cache

Block size = 8 words

Total words = 512

No. of blocks = $\frac{512}{8}$ = 64 blocks

Tag	Data	Tag	Data

32K X 12 memory
512 X 12 cache

Tag	Index

word	Block

Write into cache

① Write Through → simultaneously updation
② write back → status bit = 1,

Ques A digital computer has memory unit of 64K X16 and cache of 1K words

64K X16 → 1K words

The cache uses direct mapping with block size of 4 words.

a) How many bits are there in tag, index, block and word field of add format.
b) How many bits are there in each word of cache include valid bit.
c) How many block cache can accomodate

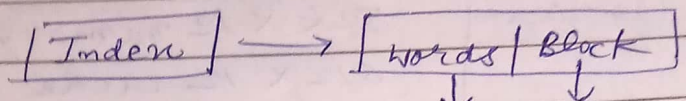
→ Tag - Data is stored in cache

$64K \rightarrow 2^6 \times 2^{10} = 2^{16} = 16$ bits

Tag	Index

16 → 10 bits as 1K words $= 2^{10}$ $= 10$ bits

$16 - 10 = 6$



Block size = 4 words
2^2 = 2 bits for word

(2) Data + Tag
16 + 6 = 22 + 1 = 23
for valid bit

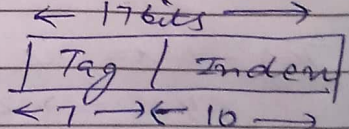
64K x 16
data

(3) 1K words in cache means 1024 locations and block size is of 4 $\Rightarrow \frac{1024}{4} = 256$

Ques A 2 way set associative memory uses blocks of 4 words. The cache can accomodate a total of 2048 words of main memory. size of M.M. is 128K x 32

a) formulate all info required to construct cache Memory.
b) what is size of cache memory.

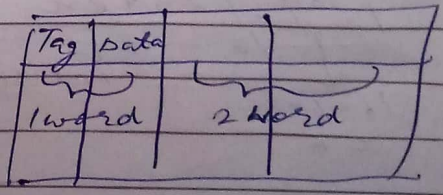
→ a) 128 K x 32



$2^7 \times 2^{10} = 2^{17} = 17$ bits

Tag = 7, Index = 10 → a part

2048 words
$= 1024 \times (2 \times (\text{tag} + \text{data}))$
as 2 way
$= 1024 \times (2 \times (7 + 32))$
$= 1024 \times 78$ → (b) part



If active portions of program and data are placed in a fast small memory, the average memory access time can be reduced, reducing execution time. fast memory is called cache Memory.

Operation: When the CPU needs to access memory, the cache is examined. If the word is found in cache, it is read from fast memory.
Some data is transferred to cache, so that future references to memory find the required words in cache Memory.

If word is found in cache, it is said to be hit else miss.

→ The transformation of data from main memory to cache memory is called Mapping process.

Associative Memory: It stores both address and content of memory word. This permits any location in cache to store any word from main memory.
If address found in main memory then 12 bit data is read and sent to CPU.
The address - data pair is transferred to associative cache memory.

Direct Mapping: The 15 bit address is divided into 9 bit index field and 6 bits form the tag field.

Classfellow
DATE : / 20
PAGE No

Associative Memory:

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 data itself rather than address. A memory unit accessed by content is called an associative memory or content addressable Memory (CAM).

When a word is written in an associative memory no address is given. The memory is capable of finding an empty unused location to store the word. When a word is to be read from an associative memory, the content of word is specified. The memory scans all words which match the specified content and marks them for reading. Expensive

Hardware Organization:

Each word in memory is compared in parallel with content of argument register. The words that match the bits of argument register set a corresponding bit in match register. set indicate corresponding words have been matched

Key register provides a mask for choosing a particular field.

Cache Memory:

References to memory at any given tend to be confined within a few localized areas. This is called locality of reference.