Amey Thakur
B-50

Mumbai University

Terna Engineering College, Nerul, Navi Mumbai

Department of Computer Engineering

Course Code	CSC403	Program	B.E. (CMPN)
Semester	IV	Year	II
Name of the Faculty	Rohini Palve	Class / Div	A & B
Course Title	Computer Organization and Architecture	Academic year	2019-20

Amey Thakur B-50

ASSIGNMENT 4

Q1) Explain memory interleaving techniques. / Write a short note on interleaved memory

Q2) Explain memory hierarchy in the computer system

Q3) Explain in detail virtual memory, paging and segmentation

Q4) Explain memory segmentation in detail and explain address translation is performed in virtual memory./ Explain the page address translation in case of virtual memory

Q5) Explain cache consistency and coherency with suitable examples. Also, give methods to maintain cache consistency

Q6) Explain how virtual address is translated to a physical address with suitable examples.

Q7) Compare SRAM and DRAM.

Q8) Explain various cache mapping techniques. / What is the necessity of the cache memory? Explain set associative and associative cache mapping techniques

Q9) What is TLB? Explain its working

Q10) Explain various characteristics of memory.

Q11) Consider a cache memory of 16 words, each block consists of 4 words. Size of the main memory is 256 bytes. Draw associative mapping and calculate TAG, and WORD size

Q12) Write a short note on the Principle of the locality of reference

Q13) Calculate the number of page faults and page hits for the page replacement policies FIFO, Optimal & LRU for the given reference string

7, 0 ,1 ,2 ,0 ,3 ,0 ,4 ,2 ,3 ,0 ,3 ,2 ,1 ,2 ,0 ,1 ,7 ,0 ,1(assume three frame size)

Q14) What is associative memory?

(CO5): To demonstrate the memory mapping techniques

Name: Amey Thakur Class: Div: B Roll No: 50

Subject: Topic: Date: Page No: 1

Q1.

Ans:

Interleaved Memory

- Interleaved memory implements the concept of accessing more words in single memory access cycle.
- Memory can be partitioned into N separate memory modules. Thus N accesses can be carried out to the memory simultaneously.
- Once presented with a memory address, each memory module returns one word per cycle. It is possible to present different addresses to different memory modules so that parallel access to multiple words can be done simultaneously or in a pipelined fashion.
- The maximum processor bandwidth in interleaved memory can be equal to the number of modules i.e N words per cycle.
- To achieve the address interleaving consecutive addresses are distributed among N interleaved modules. For ex - If we have consecutive address and 4 interleaved memory modules then 0^{th}, 4^{th}, 8^{th} addresses will be assigned to the first memory module and so on.

0, 4, 8, 12, 16, ... Addresses to memory module

1, 5, 9, 13, 17 ..... Addresses to memory module 1

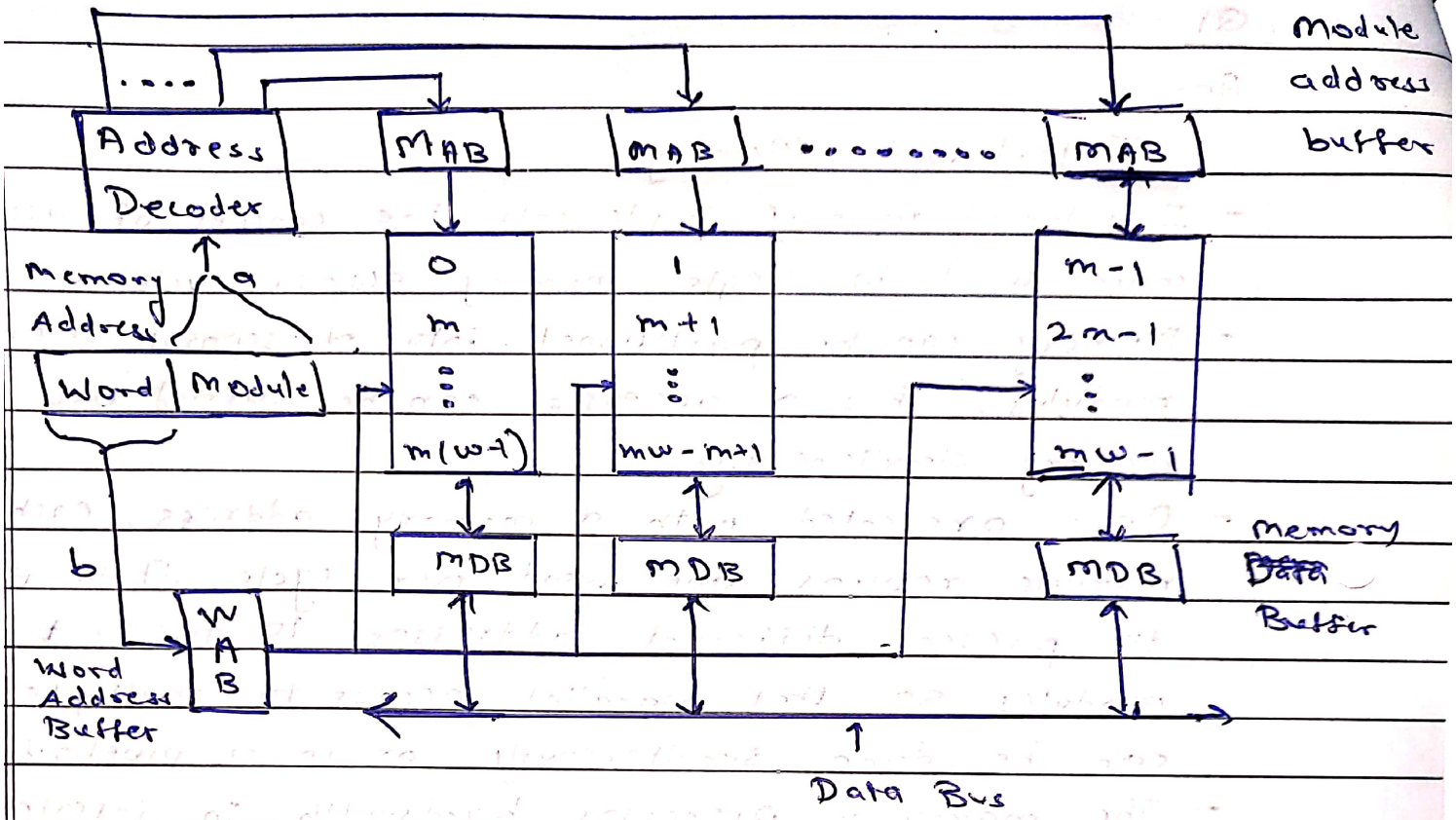
2, 6, 10, 14, 18, ... Addresses to memory module 2

3, 7, 11, 15, 19, ... Addresses to memory module 3

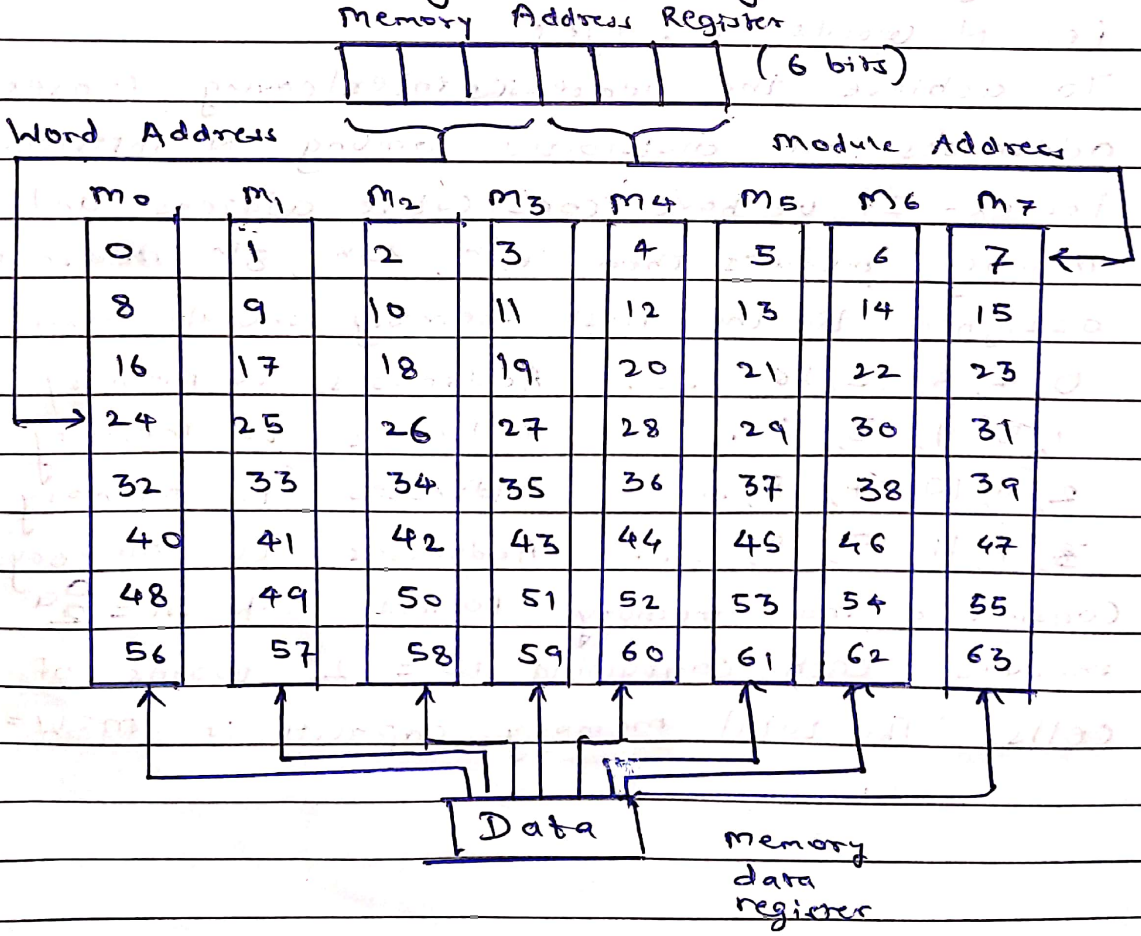
- Consider a main memory formed with $m = 2^a$ memory modules, each containing $w = 2^b$ words of memory cells. The total memory capacity is $m \cdot w = 2^{a+b}$ words.

Amey Thakur

B-50



Low-order m-way interleaving (the c-access memory scheme)



Name: Amey Thakur Class: Div: B Roll No: 50

Subject: Topic: Date: Page No: 3

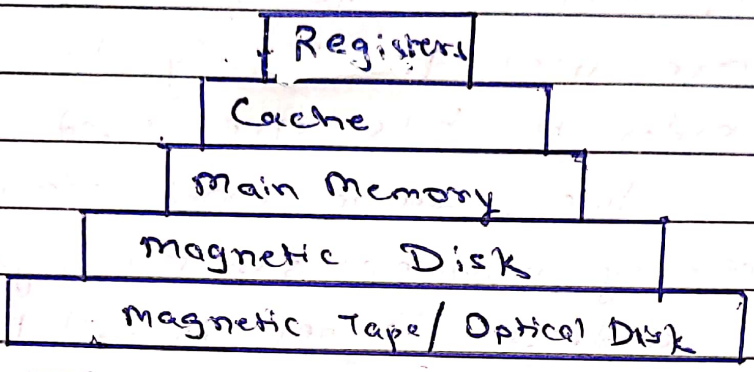
- Figure shows memory format for memory interleaving. Interleaving spreads contiguous memory locations across m modules horizontally.
- This implies that the low order a bits of the memory address are used to identify the memory module.
- The high order b bits are used to address a word inside a module, same word address is applied to all memory modules simultaneously.
- A module address decoder is used to distribute module address.

Amey Thakur
B - 50

Q.2.

Ans:

Memory Hierarchy



The Memory Hierarchy

- Memory in computer is required for storage and subsequent retrieval of instructions or data.
- A memory system of a computer must be able to keep up with the CPU. i.e. as the CPU is executing instructions, it should not wait for the operands or instructions to be read from the memory.
- For practical system the cost of the memory must be reasonable.
- A memory system has 3 basic characteristics.
  - → Cost per bit
  - → Capacity
  - → Access Time
- There found to be a trade off among the three characteristics of memory.
- The following relationship hold:
  - → Smaller access time, greater cost per bit
  - → Greater capacity, smaller cost per bit
  - → Greater capacity, Greater access time

Name: Amey Thakur Class: Div: B Roll No: 50

Subject: Topic: Date: Page No: 5

- In typical memory hierarchy, as we move from top to bottom, the following occurs -
  - → Decreasing cost per bit
  - → Increasing capacity
  - → Increasing Access Time
  - → Decreasing frequency of access of the memory by the CPU.
- The memory hierarchy will work only if the frequency of access to slower memories is significantly less than the faster memories

Amey Thakur
B-50

Q3.

Ans:

Virtual Memory

- In most computer systems the physical main memory is not as large as the address space of processor. For ex- if the processor uses 32 bit address then its address space will be.

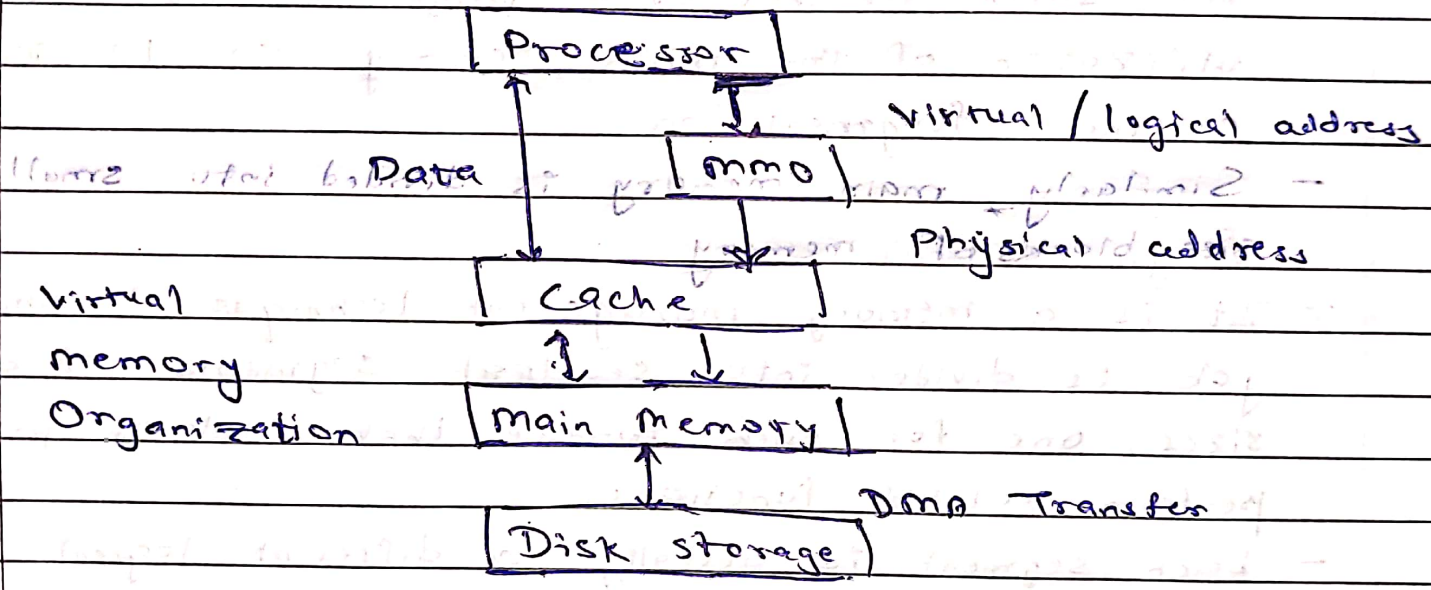
$$2^{32} = 2^{30} \times 2^{2} = 4 \text{ GB}$$

- In reality the physical memory that we use in a computer system is in megabytes and not in Gigabytes.
- When we try to run a program that will not completely fit into the main memory, the parts of it currently being executed are stored in main memory and the remaining portion is stored in secondary storage device such as hard disk.
- Of course, all parts of a program which are needed for execution are first brought into the main memory. When a new segment of the program is to be brought and the memory is full, it must replace another segment already in the memory. Thus an application program run without any limitation imposed by available main memory.
- The techniques that automatically move program and data blocks into the physical main memory when they are required for execution are called as Virtual memory techniques.
- Program or processor references an instruction and data space that is independent of available physical memory space.
- The address issued by the processor is called virtual / logical address.

Name: Amey Thakur Class: Div: B Roll No: 50

Subject: Topic: Date: Page No: 7

- The virtual address is translated into physical address by a combination of hardware and software components
- If a virtual address refers to a part of the program or data space that is currently in the physical memory, then it is accessed immediately.
- If the referenced address is not in the main memory, its content must be brought into main memory before it can be used.



- Special hardware unit called the memory management unit (MMU) translates virtual address into physical address.

Amey Thakur
B - 50

Paging

- It is a memory management technique in which process address space is broken into blocks of the same size called pages (size is power of 2, between 512 bytes and 8192 bytes).
- The size of the process is measured in the number of pages. Similarly, main memory is divided into small fixed size blocks of physical memory called frames. and the size of a frame is kept the same as that of a page to have optimum utilization of the main memory and to avoid external fragmentation.

Segmentation

- It is a memory management technique in which each job is divided into several segments of different sizes, one for each module that contains pieces that perform related functions.
- Each segment is actually a different logical address space of the program.
- When a process is to be executed, its corresponding segmentation are loaded into non-contiguous memory though every segment is loaded into contiguous memory.
- segmentation memory management works very similar to paging but here segments are of variable length wheas in paging pages are of fixed size.
- A program segment contains the program's main function, utility functions, data structures and so on.
- The OS maintains a segment map table for every process and a list of free memory blocks along with segment numbers, their size and corresponding memory locations in main memory.

Name: Amey Thakur Class: Div: B Roll No: 50

Subject: Topic: Date: Page No: 9

- For each segment, the table stores the starting address of the segment and the length of the segment.
- A reference to a memory location includes a value that identifies a segment and an offset.

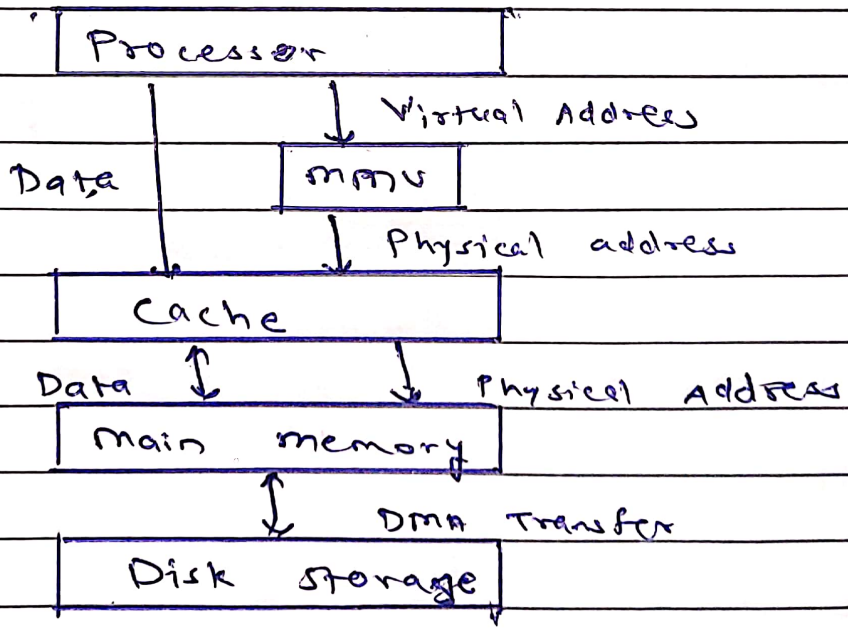
Amey Thakur
B-50

Q4.

Ans:

Virtual Memory -

- Virtual memory was introduced to the system in order to increase size of memory
- A hardware unit called Memory Management Unit (MMU) translates virtual addresses into physical addresses
- If CPU wants data from main memory and it is not present in main memory then MMU causes operating system to bring the data into the memory from disk.
- As the disk limit is beyond the main memory address, the desired data address has to be translated from virtual to physical address.
  MMU does the address translation
- Figure for Virtual Memory Organization



Name: Amey Thakur Class: Div: B Roll No: 50

Subject: Topic: Date: Page No: 11

Paging

- Virtual memory space is divided into equal size pages
- main memory space is divided into equal size page frames each frame can hold any page from virtual memory.
- When CPU wants to access page, it first looks into main memory.
  If it is found in main memory then it is called Hit and page is transfer from main memory to CPU.
- If CPU needs page that is not present in main memory then it is called as page fault.
  The page has to be loaded from virtual memory to main memory
- There are different page replacement schemes such as FIFO, LRU, Random, etc
- During page replacement, if the old page has been modified in the main memory, then it needs to be first copied into the virtual memory and then replaced.
  CPU keeps track of such updated pages by maintaining dirty bit for each page.
  When page is updated in main memory dirty bit is set then this dirty page first copied into Virtual memory & then replaced.
- Pages are loaded into main memory only when required by CPU, then it is called demand paging
  Thus pages are loaded only after page faults.

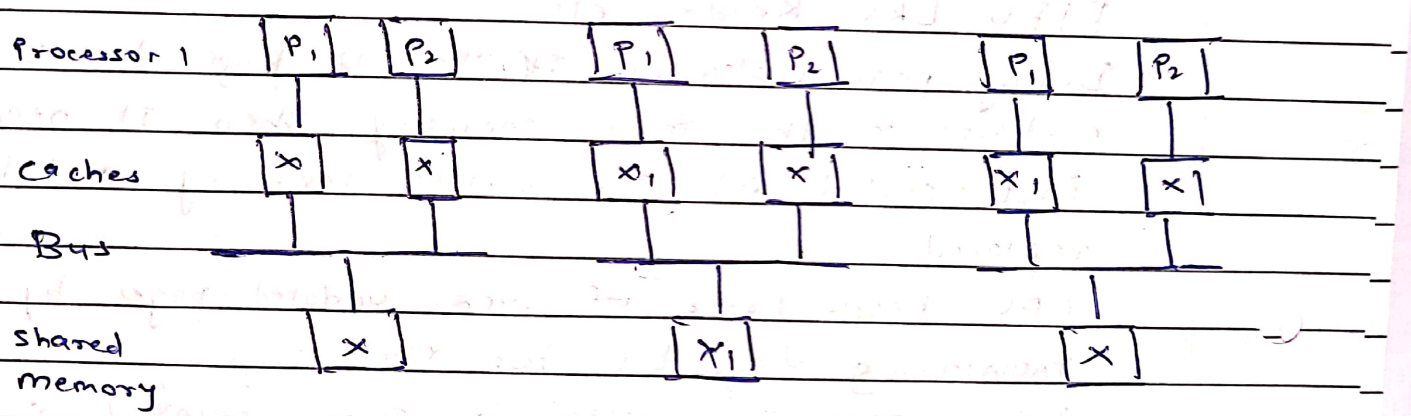
Amey Thakur
B-50

Q 5.

Ans:

The Cache Coherence Problem

- In a multiprocessor system, data inconsistency may occur among adjacent levels or within the same level of memory hierarchy.
- For ex- the cache and the main memory may have inconsistent copies of the same object.
- As multiple processor operate in parallel and independently multiple caches may posses different copies of the same memory block, this creates cache coherence problem. Cache coherence schemes help to avoid this problem by maintaining a uniform state for each cached block of data.



- Let x be an element of shared data which has been referenced by 2 processors P_1 & P_2.
- In the beginning, 3 copies of x are consistent.
- If the processor P_1 writes a new data x_1 into the cache, by using write-through policy, the same copy will be written immediately into the shared memory.

Name: Amey Thakur Class: Div: B Roll No: 50

Subject: Topic: Date: Page No: 13

- In this case, inconsistency occurs between cache memory and the main memory.
- When a write-back policy is used, the main memory will be updated when the modified data in the cache is replaced or invalidated.
- In general, there are 3 sources of inconsistency problem
  - → Sharing of writable data
  - → Process Migration
  - → I/O Activity
- Maintaining cache coherency is a problem in a multiprocessor system when the processor contain local cache memory.
  Data inconsistency between different caches easily occurs in the system

Amey Thakur
B-50

Q.6.

Ans:

- When a program accesses memory, It does not know or care where the physical memory backing the address is stored. It knows it is up to operating system and hardware to work together to map / locate the right physical address and thus provide access to data it wants. Thus turns the address a program is using to access memory a virtual address. A virtual address consist of two parts page and offset

Page:

- Since the entire possible address space is divided up into regular sized pages, every possible address resides in a page.
- The page component of the virtual address acts as an index into page table.
- Since the page is smallest unit of memory allocation within the system, there is a trade-off between making pages very small and there having very many pages for the OS to manage and making pages larger but potentially wasting memory.

Offset:

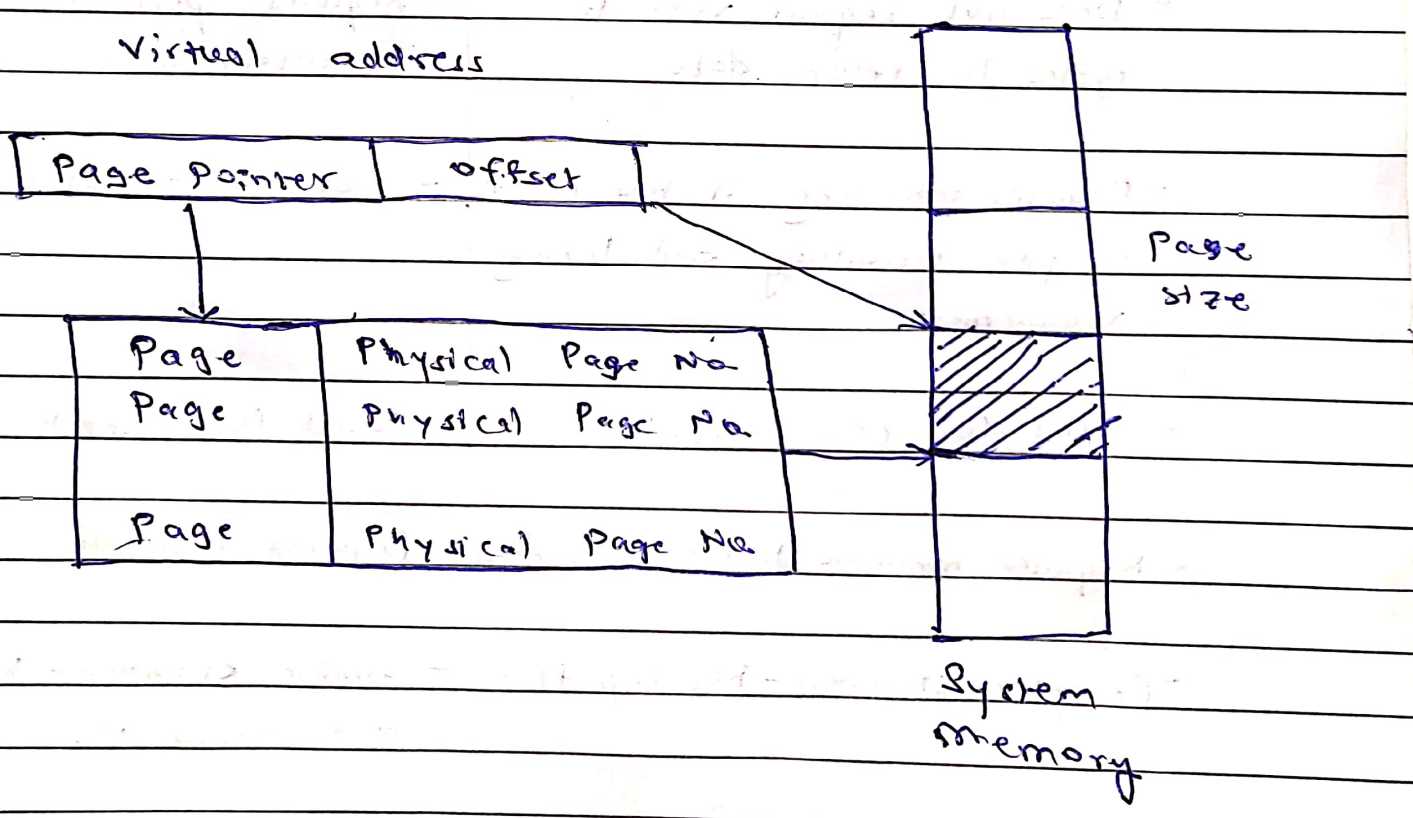
- The last bit of virtual address are called offset which is the location difference between byte address you want and the start of page. You require enough bits in the offset to be able to get to any bite in the page. For a 4k page, you require 12 bits of offset.
- Remember that the smallest amount of memory that OS or hardware deals with is a page, so each of these 4096 bytes resides within a single page or are dealt with as "one".

Name: Amey Thakur Class: Div: B Roll No: 50

Subject: Topic: Date: Page No: 15

Virtual Address Translation

- It refers to the process of finding out which physical page maps to which virtual page.
- When translating a virtual address to a physical address, we only deal with page number. The essence of procedure is to take page number of the given address to look it up from the virtual address to find a pointer to a physical address, to which offset from virtual address is added, giving the actual location in system memory.
- Since the page tables are under control of the OS doesn't exist in the page tables then the OS knows the process is trying to access memory that has not been allocated to it and the access will not be allowed.



Amey Thakur

B-50

Q.7.

Ans.

SRAM	DRAM
- A type of semiconductor memory that uses bi-stable latching circuitary (flip flop) to store each bit-	- A type of random access semiconductor memory that stores each bit of data in a separate tiny capacitor within integrated circuit.
- Stands for static Random Access memory	- Stands for Dynamic Random Access memory
- Very Fast	- Not as fast as SRAM
- Does not require refresh cycles to retain data	- Requires periodical refresh cycles to retain data
- Requires refreshing, it has more complex circuitary and timing requirements	- Not as complex as SRAM
- Used for CPU cache	- Used for computer's main memory
- Requires minimum time to access data	- Requires more time to access data
- Complex structure - has flip flops	- Simple structure - has transistor and capacitor
- Has a lower density	- Has a higher density
- Expensive	- Less expensive

Name: Amey Thakur Class: Div: B Roll No: 50

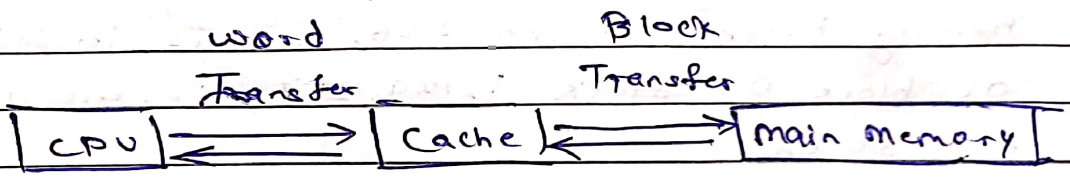
Subject: Topic: Date: Page No: 17

Q8.

Ans:

Cache Memory

- It is very high speed memory used to increase speed of program by making current program and data available to the CPU at a rapid rate
- Access time to cache memory is less compared to main memory. It contains a copy of portions of main memory.
- When CPU attemps to read a word from main memory, check is made to determine if the word is in cache. If so, then word is delivered from cache.
- If word is not there in cache then a block of main memory consisting some word along with that word is read into cache and the required word is delivered to CPU. This is called Principle of Locality of Reference.
- During a miss if there are no empty blocks in the cache then such replacement policies such as FIFO, LRU, LFU, etc. are used



Cache Mapping Techniques

→ Direct Mapping

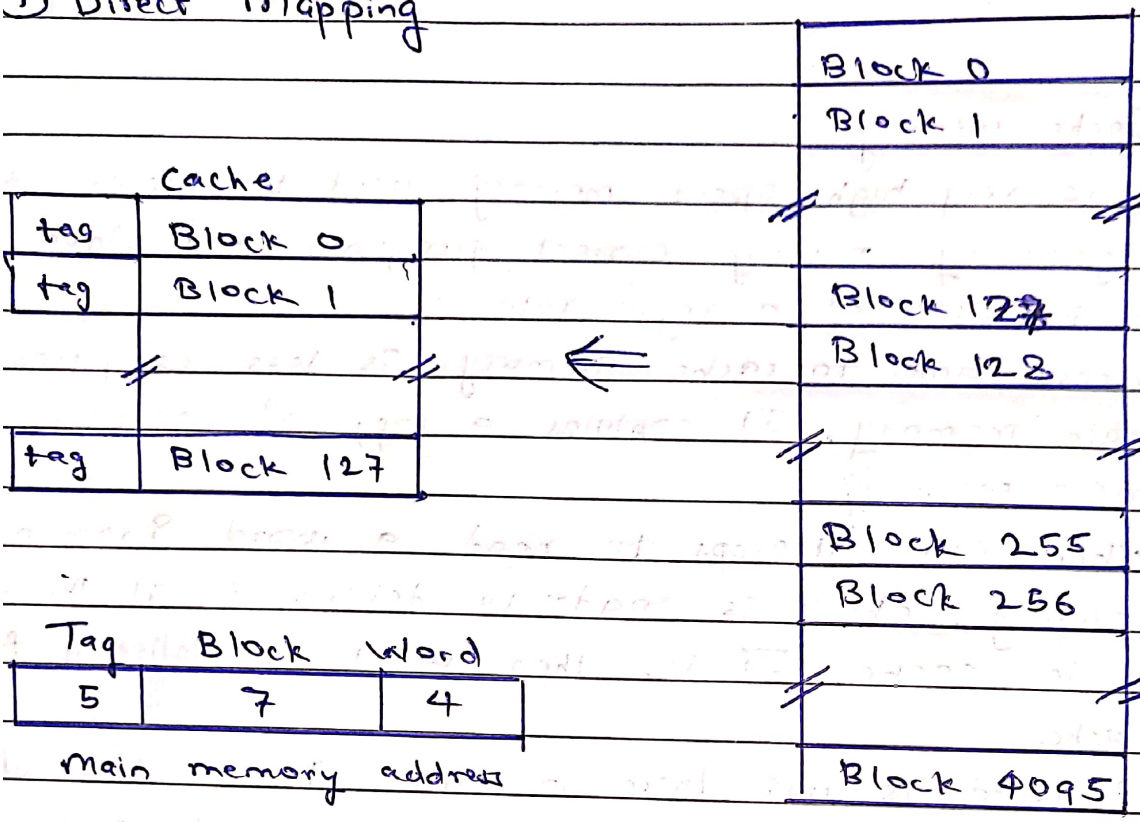
→ Associative Mapping

→ Set Associative Mapping

Consider a cache consisting of 128 blocks of 16 words each, for total of 2048 (2K) words and assume that the main memory is addressable by 16 bit address. Main memory is 64 K which will be viewed as 4K blocks of 16 words each

Amey Thakur
B-50

1) Direct Mapping



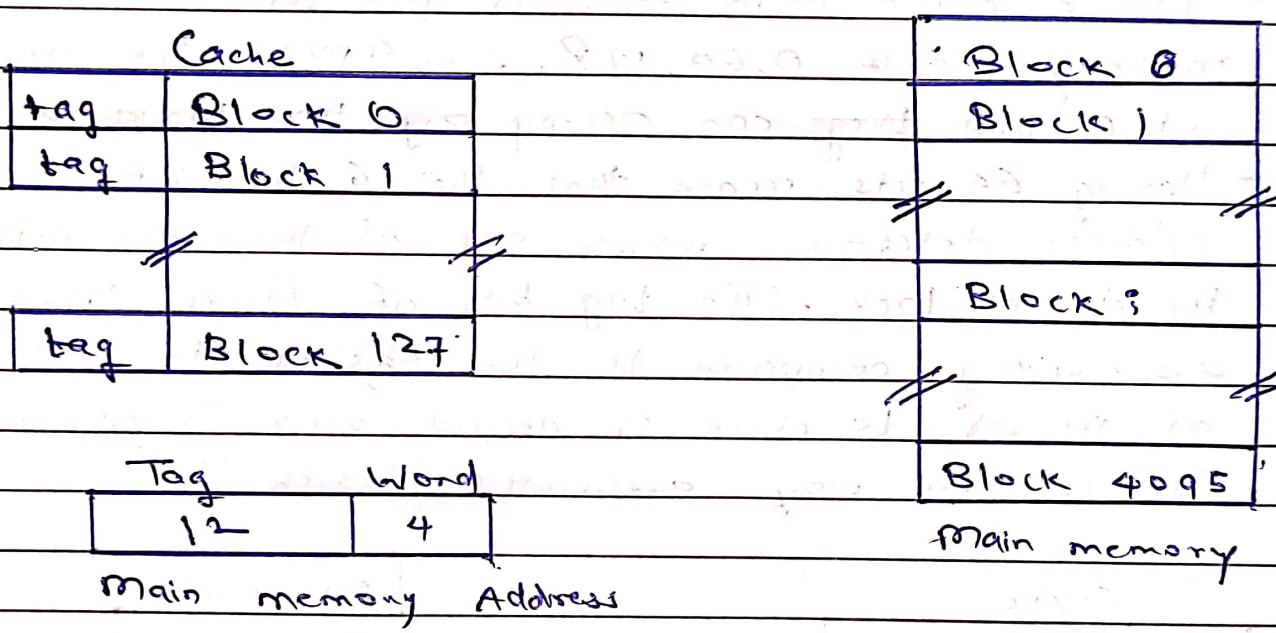
Tag	Block	Word
5	7	4

Main memory address

- The simplest way to determine cache location in which stores memory blocks is direct mapping technique.
- In this block J of the main memory maps on to block J modulo 128 of the cache. Thus main memory blocks 0, 128, 256, ... is loaded into cache at block 0. Blocks 1, 129, 257, ... are stored at block 1 and so on
- Placement of the block in the cache is determined from memory address. Memory address is divided into 3 fields, the lower 4 bits select one of the 16 words in a block
- When new block enters the cache, the 7 bit cache block field determines the cache position in which this block must be stored

- The higher order 5 bits of the memory address of the block are stored in 5 tag bits associate with its location in cache. They identify which of the 32 blocks that are mapped into this cache position are currently resident in the cache.
- It is easy to implement but not flexible.

(2) Associative Mapping

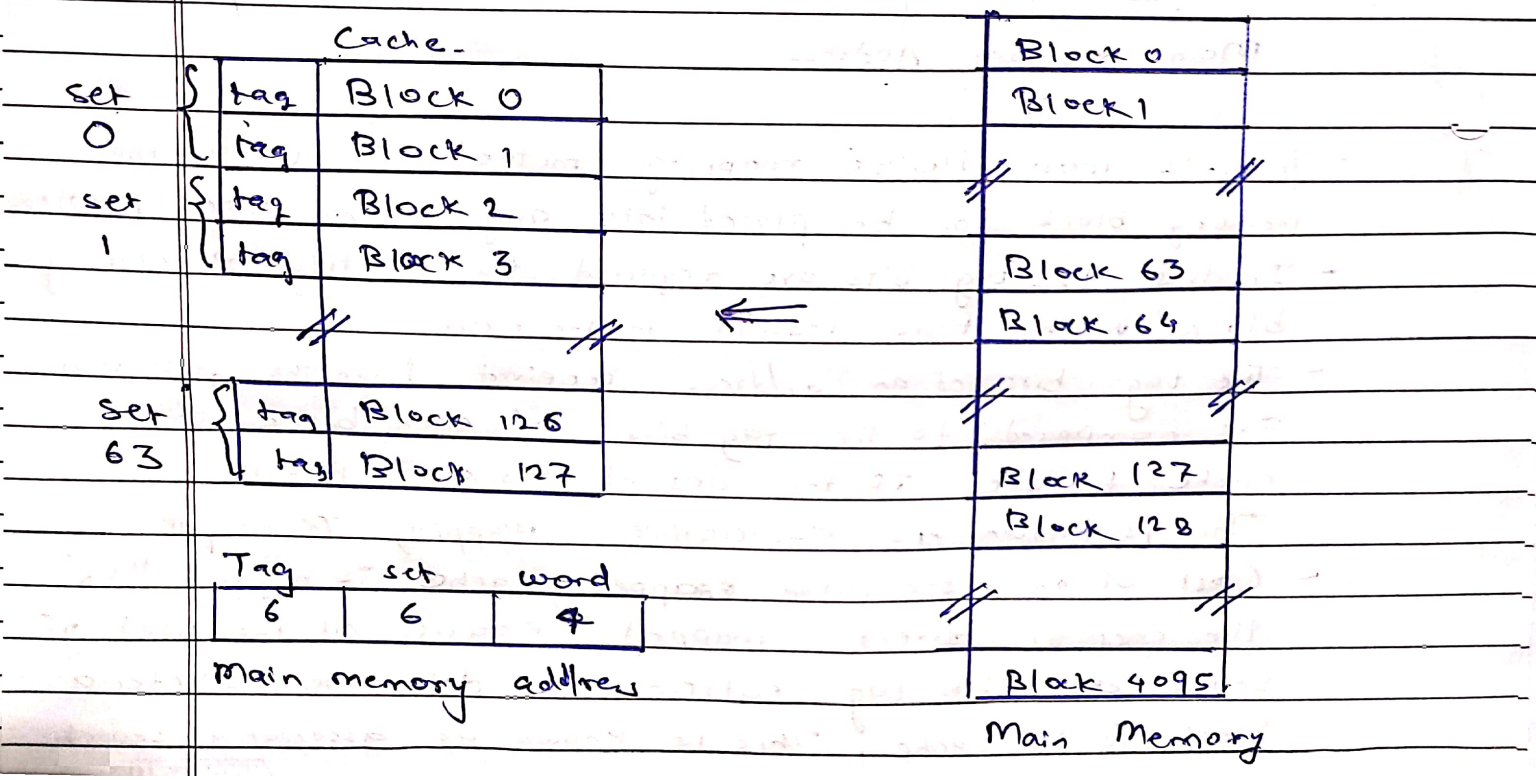


- This is more flexible mapping method, in which main memory block can be placed into any cache block position.
- In this, 12 tag bits are required to identify a memory block when it is resident in the cache.
- The tag bits of an address received from the processor are compared to the tag bits of each block of the cache to see, if the desired block is present.
  This is known as associative mapping Technique.
- Cost of an associated mapped cache is higher than the cost of direct mapped because of the need to search all 128 tag patterns to determine whether a block is in cache. This is known as associative search.

Amey Thakur
B-50

(3) Set Associated Mapping

- It is the combination of direct and associative mapping technique
- Cache blocks are grouped into sets and mapping allows block of main memory reside into any block of a specific set. Henle contention problem of direct mapping is eased, at the same time, hardware cost is reduced by decreasing the size of associative approach
- For a cache with 2 blocks per set. In this case, memory blocks 0, 64, 128, ----, 4032 map into cache set 0 and they can occupy any two blocks within set
- Having 64 sets means that the 6 bit set field of the address determines which set of the cache might contain the desired lock. The tag bits of address must be associatively compared to the tags of the two blocks of the set to check if desired block is present. This is Two way associative search



Name: Amey Thakur Class: Div: B Roll No: 50

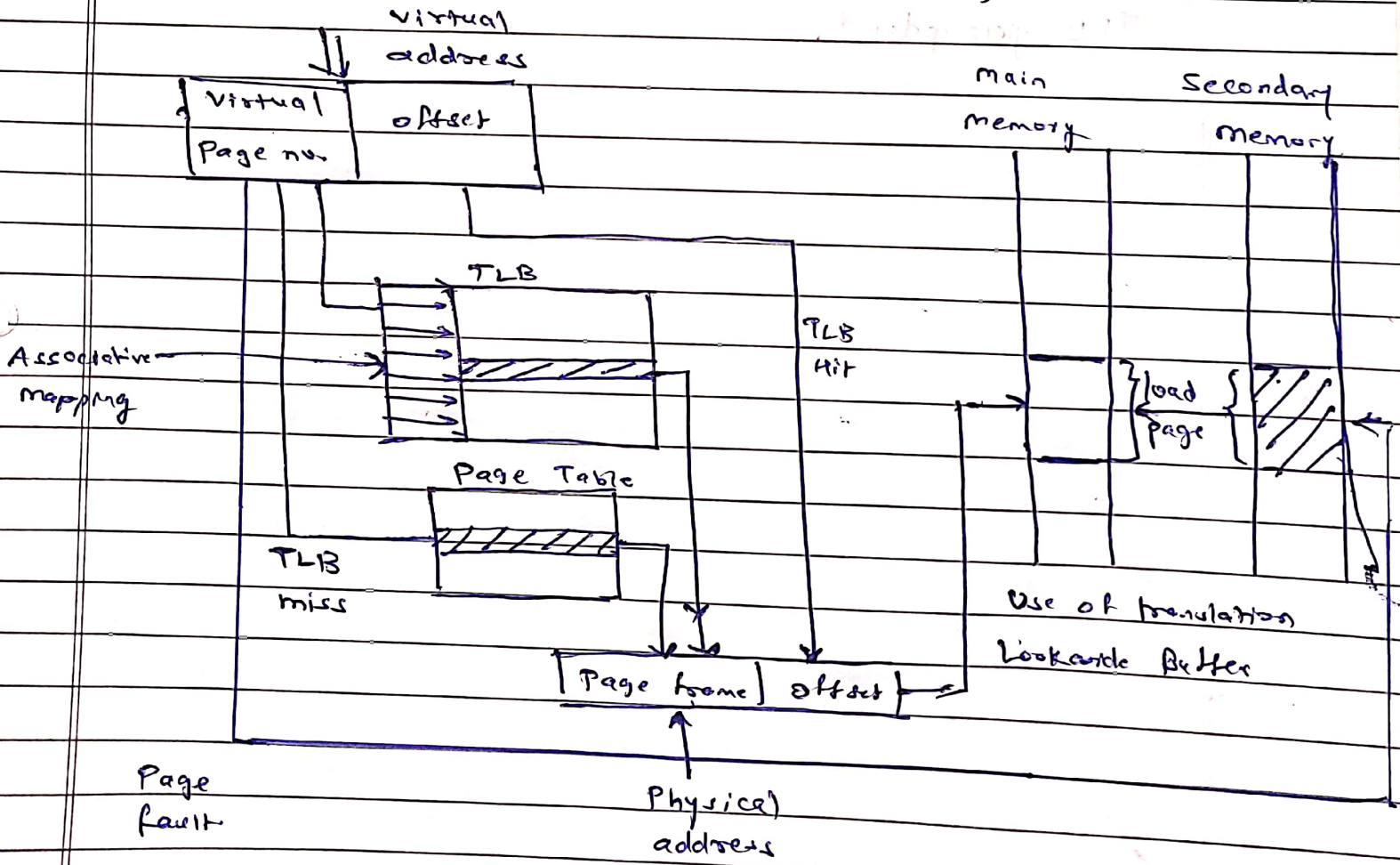
Subject: Topic: Date: Page No: 21

Q9.

Ans:

Translation Lookaside Buffer

- Every virtual memory reference can cause two physical memory accesses
  1. One to fetch the appropriate page table entry
  2. One to fetch the desired data
- Thus the virtual memory scheme would have the effect of doubling the memory access time.
- To overcome this problem most virtual memory management schemes make use of a special high speed cache for page table entries, usually called a Translation Lookaside Buffer (TLB)



Amey Thakur

B - 50

- Here associative mapping is used.
- TLB contains those page entries that have been most recently used.
- An entry of TLB contains
  - → Virtual Page number
  - → Control Bits
  - → Page frame number in memory
- Address translation proceeds as follows given a virtual address, the processor looks in TLB for reference page. If the page table entry for this page is found in TLB, the physical address obtains immediately.
- If there is miss in the TLB, then the required entry is obtained from the page table in main memory and TLB gets updated.

Name: Amey Thakur Class: Div: B Roll No: 50

Subject: Topic: Date: Page No: 23

Q10.

Ans:

Characteristics of memory

Type of Memory	Access mode	Permanance of storage	Physical Nature of storage
· Semiconductor memories	Random	Volatile	Electronic
Magnetic Disk	Direct	Non-Volatile	Magnetic
Magnetic Tape	Sequential	Non-volatile	Magnetic
Compact Disk ROM	Direct	Non-Volatile	Optical

Amey Thakur

B - 50

Q11.

Ans:

Given:

Cache memory = 16 words

Block consists of 4 words

Main memory = 256 bytes

Main Memory = 256 * 16 words $= 2^8 * 2^4 = 2^{12}$

TAG field $= 2^{12}$ = 12 bits consists of both set fields and TAG field

SET + TAG = 12

4 + TAG = 12

TAG = 12 - 4

TAG = 8 bytes

Word size = As there are 8 blocks and each block consists of 4 words

Hence, 8 * 4 = 32

$= 2^5$

TAG field	SET field	Word Field
4 bytes	8 bytes	5 bytes

Name: Amey Thakur Class: Div: B Roll No: 50

Subject: Topic: Date: Page No: 25

Q12.

Ans:

Principle of locality of reference

- Locality of reference is the term used to explain the characteristics of program that run in relatively small loops in consecutive memory locations.
- The locality of reference principle comprises of two components
  - → Temporal locality
  - → Spatial locality

① Temporal locality

- Since the program have loops, the same instructions are required frequently, i.e the program tend to use the most recently used information again and again
- If for a long time an information in cache is not used then it is less likely to be used again
- This is known as principle of temporal locality

② Spatial locality

- Programs and the data accessed by the processor mostly reside in consecutive memory location
- This means that the processor is likely to need code or data that are close to locations already accessed
- This is known as principle of spatial locality

Amey Thakur
B-50

Q.13.

Ans:

① FIFO Page Replacement:

In FIFO page replacement, when a page is needed to be replaced, we select the oldest page.

String	7	0	1	2	0	3	0	4	2	3	0	3	2	1	2	0	1	7	0	1
1	7	7	7	2	2	2	2	4	4	4	0	0	0	0	0	0	0	7	7	7
2		0	0	0	0	3	3	3	2	2	2	2	2	1	1	1	1	1	0	0
3			1	1	1	1	0	0	0	3	3	3	3	3	2	2	2	2	2	1
F	F	F	F		F	F	F	F	F	F			F	F			F	F	F	

Page Hit: If the file is already present, then it is a page Hit.

Page miss: If an entry is not found, then it is a page miss.

No. of page hit = 5

No. of page miss = 15.

Name: Amey Thakur Class: Div: B Roll No: 50

Subject: Topic: Date: Page No: 27

② Optimal page replacement.

Here, when a page replacement is needed, it look ahead in the input queue for the page frame which will be referenced only after a long time. The page with the longest referenced is swapped.

String	7	0	1	2	0	3	0	4	2	3	0	3	2	1	2	0	1	7	0	1
1	7	7	7	2	2	2	2	2	2	2	2	2	2	2	2	2	2	7	7	7
2		0	0	0	0	0	0	4	4	4	0	0	0	0	0	0	0	0	0	0
3			1	1	1	3	3	3	3	3	3	3	3	1	1	1	1	1	1	1
F	F	F	F		F		F			F			F				F			

No. of page Hit = 11

No. of page miss = 9

③ LRU page replacement

This method uses the recent past as an approximation of near future. We replace the page which has not been referenced for a long time in the past.

String	7	0	1	2	0	3	0	4	2	3	0	3	2	1	2	0	1	7	0	1
1	7	7	7	2	2	2	2	4	4	4	0	0	0	1	1	1	1	1	1	1
2		0	0	0	0	0	0	0	0	0	3	3	3	3	3	3	0	0	0	0
3			1	1	1	3	3	3	2	2	2	2	2	2	2	2	1	7	7	7
F	F	F	F		F		F	F	F	F			F		F		F			

No. of page Hit = 8

No. of page miss = 12

Amey Thakur
B - 50

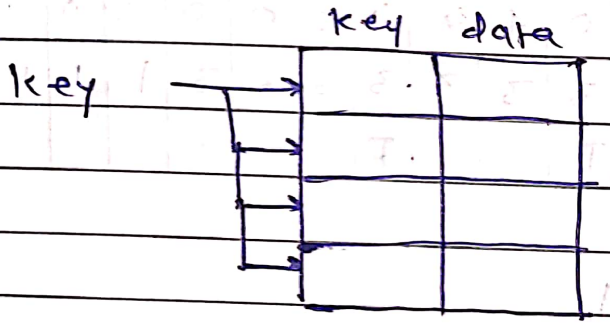
Q14.

Ans:

Associative Memory

- In Associative memory, any stored items can be accessed by using the contents of items
- Items are stored in an associative memory have two field format.

Key, Data



- Associative searching is based on simultaneous matching of key to be searched with stored key associated with each line of data
- A word is retrieved based on a portion of its content rather than its address