Chapter - 4 CACHE MEMORY

Characteristics of Computer Memory Systems

① • **Location**
- Refers to whether memory is internal or external to the computer
- Internal (Registers, Cache, main memory)
- External (Optical disks, magnetic disks, tapes)
  → These are peripheral storage device

② • **Capacity**
- ~~It is expr~~ for internal memory it is expressed in terms of bytes (1 byte = 8 bits) or words (16 bits = 2 byte)
- Common word lengths are 8, 16 and 32 bits.
- External memory capacity is typically in terms of bytes.

③ • **Unit of Transfer** : → (This is the no. of bits read out of or written into the memory at a time.)
- Data can be transferred in words or blocks.
(Usually blocks which is much larger units than a word).

for internal the unit of transfer = no. of electrical lines into and out of the memory module. May be equal to word length such as 64, 128 or 256 bits.

Addressable units $= 2^A = N$ (A = length in bits in an address, N = addressable units)

→ Relationship between the length in bits 'A' of an address and the no. 'N' of addressable units is $2^A = N$.

④ • **Access Methods**
→ Sequential (e.g. ~~das~~ tape)
→ Direct (e.g. disk)
→ Random (e.g. RAM)
→ Associative (e.g. Cache).
→ This is a random access type of memory that enables one to make a comparison of desired bit locations within a word for a specified match, and to do this for all words simultaneously.

⑤ • **Performance**
→ Access time
→ Memory Cycle time
→ Transfer Rate

Date _____
Page _____

⑥ • Physical Types
→ Semiconductor (RAM)
→ Magnetic (Disk & tape)
→ Optical (CD & DVD)
→ Magnetic - Optical

⑦ • Physical Characteristics
→ Volatility (Information lost when power is switch off)
→ Non-volatility (eg. magnetics - surface memory)
(Information not lost when power is switch off)
→ Non-erasable (e.g. ROM) / Erasable

⑧ • Organisation
- Physical arrangement of bits into words.
- Not always obvious used
- Eg. - Interleaved
- Memory module

Performance
→ Access time : (latency) ⟶ For random access memory, this is the time it takes to perform a read or write operation.

→ Memory Cycle time : This concept is primarily applied to random access memory and consists of the access time plus any additional time required before a second access can start.

→ Transfer Rate : For RAM, it is equal to 1/(cycle time).
For non-random access memory

$$T_n = T_A + n/R$$

where
T_n = Avg. time to read or write n bits
T_A = Avg. access time

Date _____
Page _____

n = no. of bits
R = Transfer rate, in bits per second (bps)

→ Physical Characteristics

Volatility - In a volatile memory, information decays naturally or is lost ~~while~~ when electrical power is switched off.

Non-Volatile - In a non-volatile memory, information once recorded remains without deterioration until deliberately changed; no electrical power is needed to retain information.

- Semiconductor memory ~~can be~~ may be either volatile or non-volatile.
- Nonerasable memory cannot be altered, except by destroying the storage unit. Semiconductor memory of this type is known as ROM (Read-only-Memory).

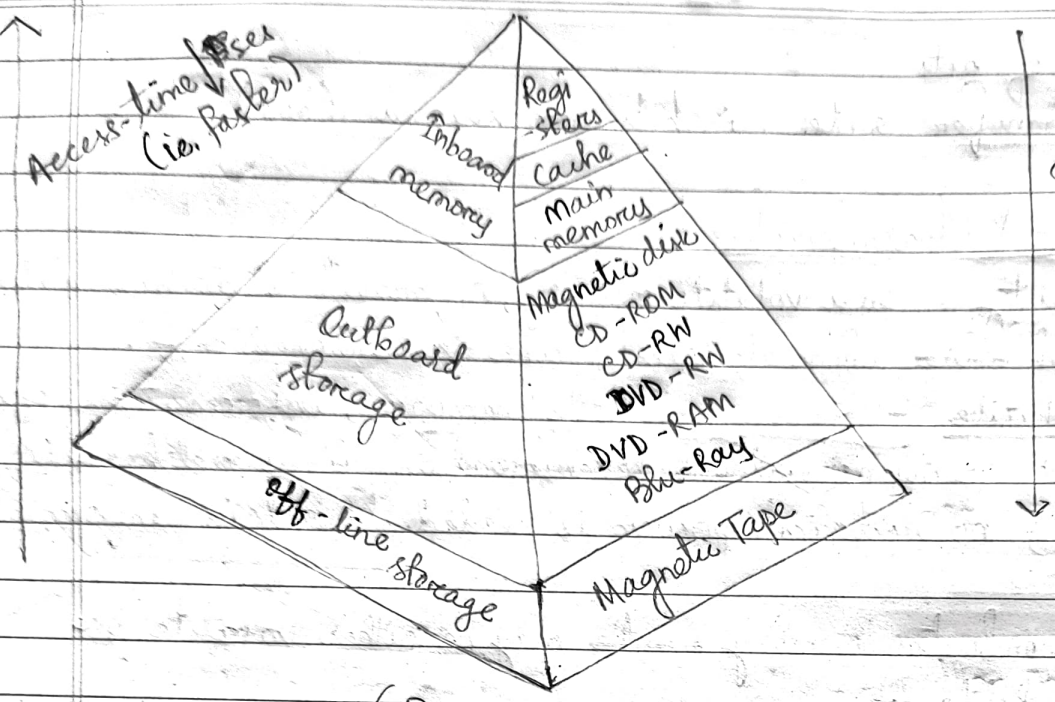
Non-erasable memory must also be non volatile.

The Memory Hierarchy

The design constraints on a computer's memory can be summed up by three questions:

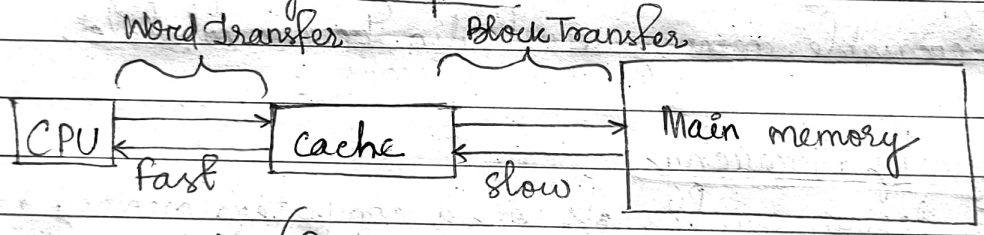
How much? (Capacity) How fast? (Access time) How expensive? → (Cost)

- Faster Access time, greater cost per bit.
- Greater Capacity, smaller cost per bit.
- Greater Capacity, slower access time.

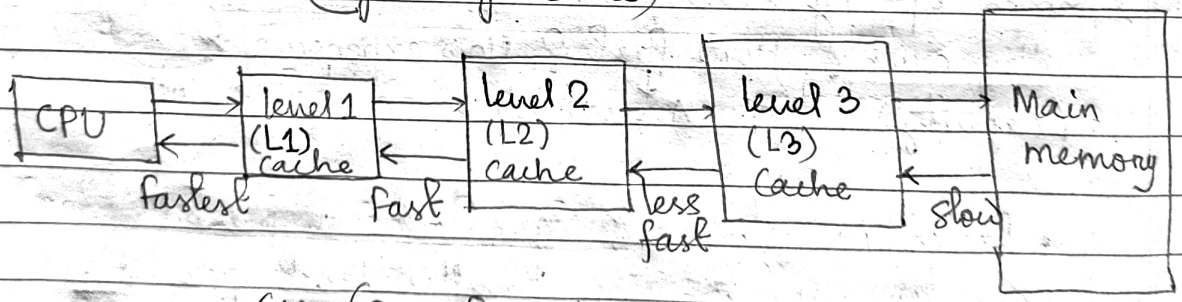


(Fig : The Memory Hierarchy)

Cache Memory Principles

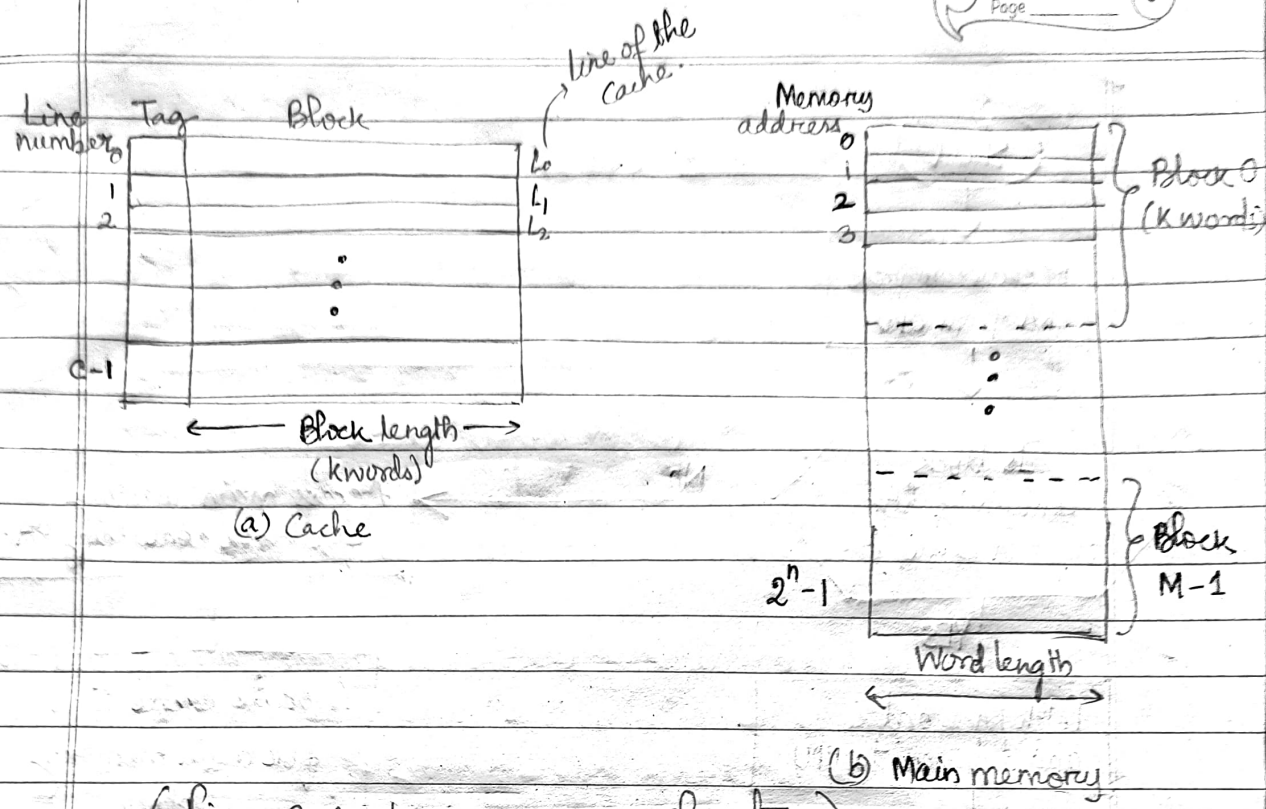


(a) (Fig. Single Cache)



(b) (Fig. Three-level Cache Organisation)

(Fig. Cache and Main memory)



(Fig. Cache/Main Memory Structure)

(Blocks get stored in Caches.)

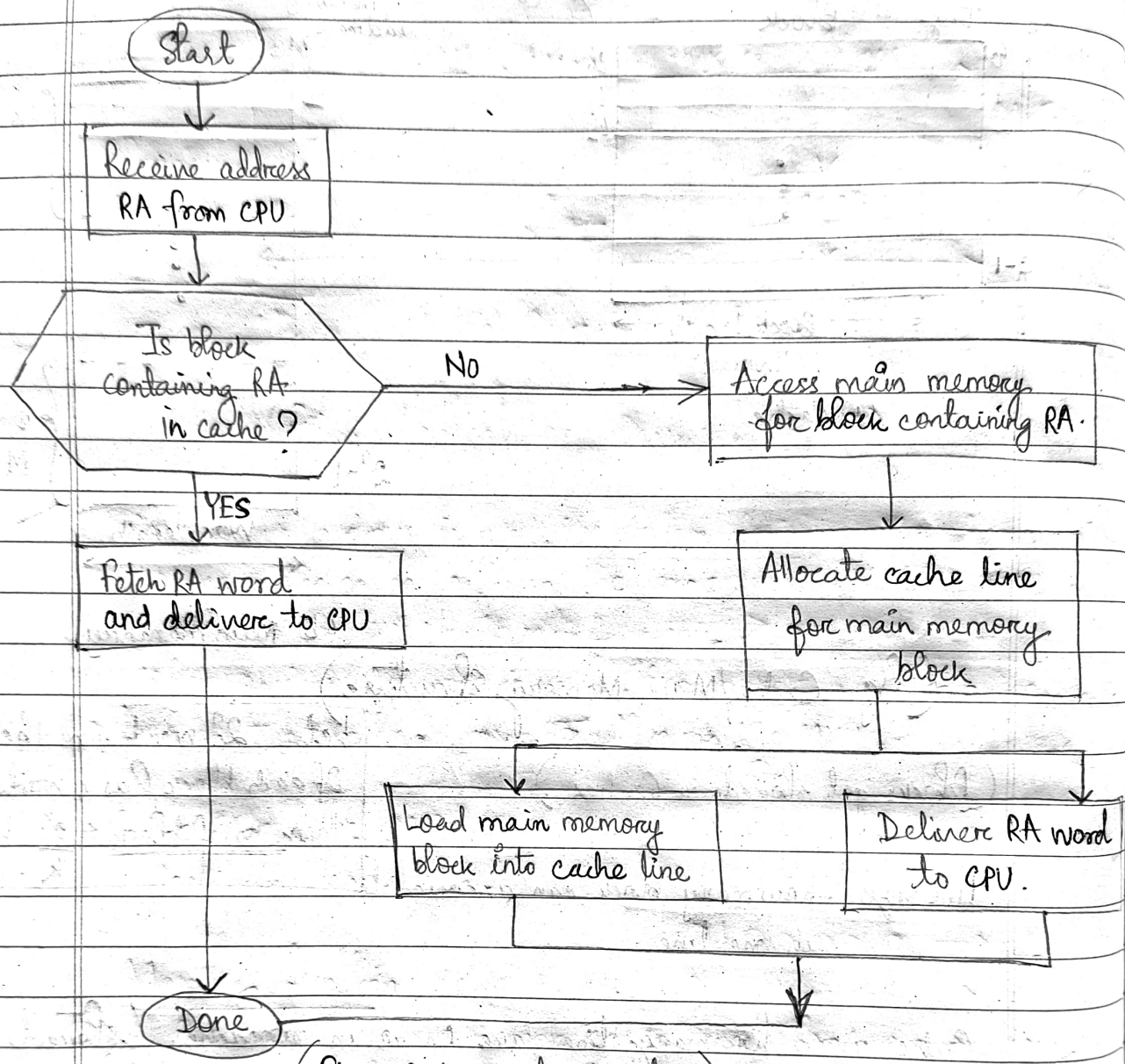
Let, 2^n words are there.
If each block has k words
then no. of block $= \frac{2^n}{k}$ blocks.

Line size - how many block can accomodate in one line.

Cache size - how many lines are there in a cache. / stored in a cache.

Cache Operation - Overview

→ CPU requests contents of memory location
→ Check Cache for this data
→ If present, get from Cache (fast)
→ If not present, read required block from main memory to cache (Cache Organization)
→ Then deliver from cache to CPU
→ Cache includes tags to identify which block of main memory is in each cache slot.



(Fig. Cache Read Operation)

Cache Organization - Overview

→ The cache connects to the processor via data, control and address lines.

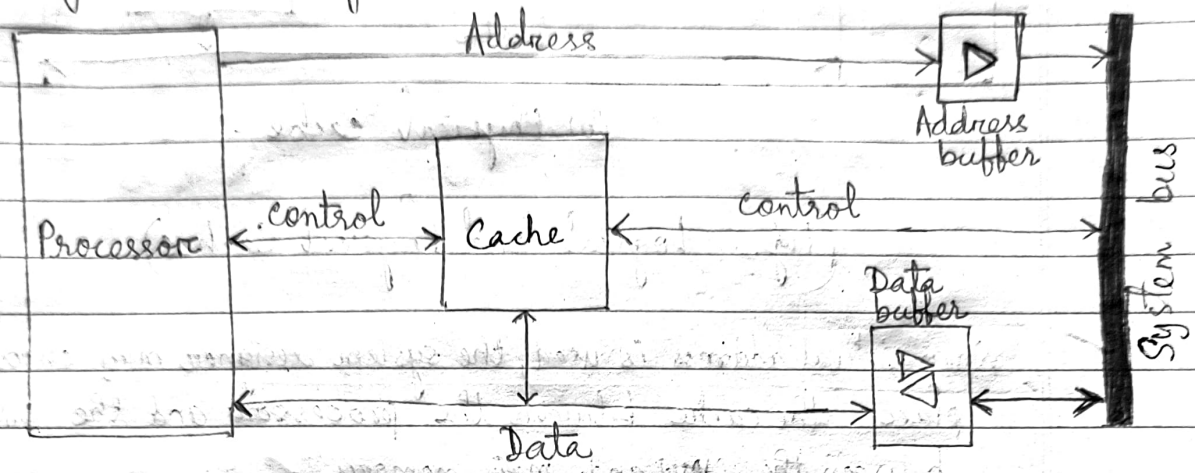
→ The data and address lines attach to data and address buffers through system bus to reach main memory.

→ Hit occurs - communication only between processor and cache (disable data and address buffer)

→ Miss occurs - the data are return through the data buffer

to both cache and the processor (desired address is loaded onto the system bus).

Typical Cache Organization

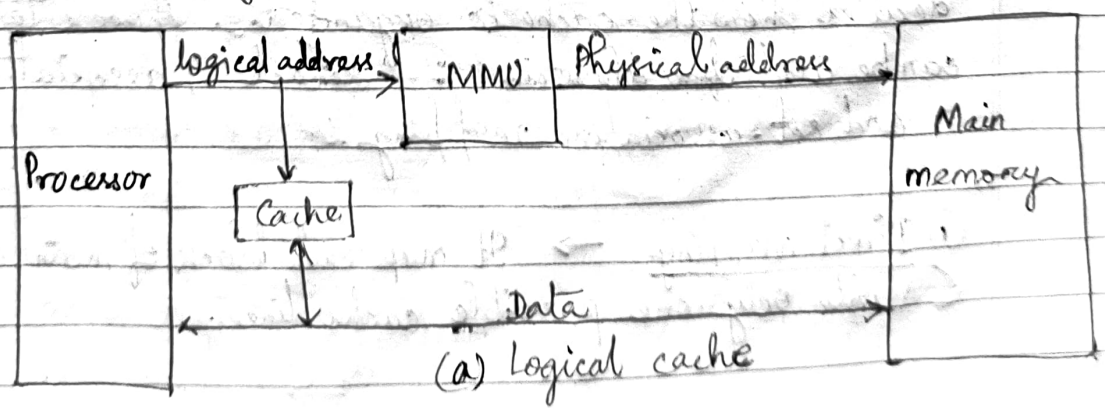


Elements of Cache Design

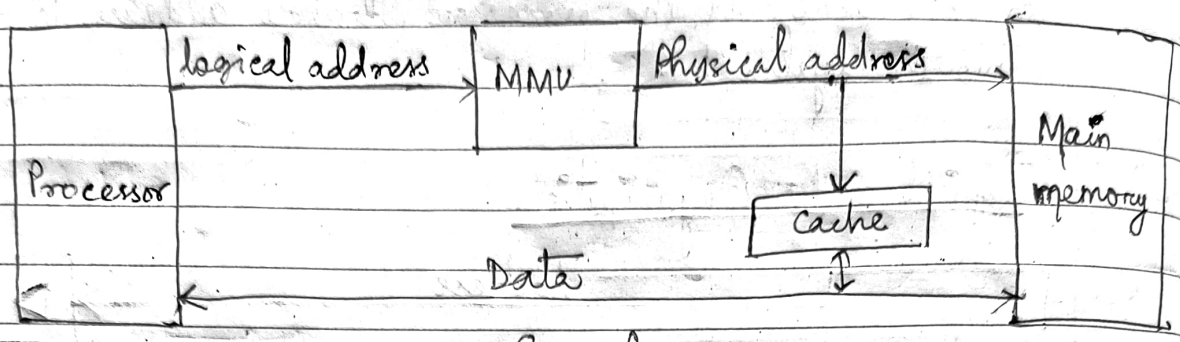
Cache Addresses

Virtual memory is a facility that allows programs to address memory from a logical point of view, without regard to the amount of main memory physically available.
When virtual memory is used, the address fields of machine instructions contain virtual addresses.

A hardware is used called as memory management unit (MMU) that translates virtual address in to a physical address in main memory.



(a) Logical cache



(b) Physical cache -

(Fig. Logical and Physical Caches)

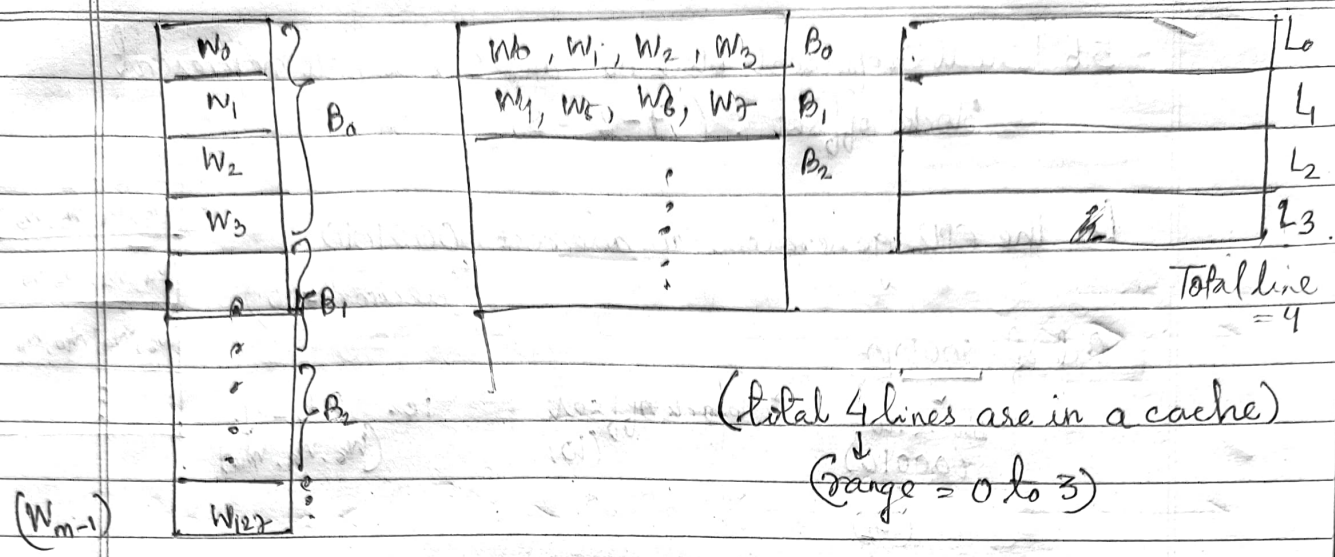
When virtual address is used, the system designer may choose to place the cache between the processor and the MMU or between the MMU and main memory.
When physical address is used, the processor accesses the cache directly, without going through the MMU.

<u>Cache Size</u> : It is impossible to arrive at a single "optimum" cache size.

<u>Mapping function</u>
Because there are fewer cache lines than main memory blocks, an algorithm is required for mapping main memory blocks into cache lines. Further, a technique is required for determining which main memory block currently occupies a cache line. The choice of the mapping decides how the cache is organised. Three techniques can be used ~~for these~~ : direct, associative and set-associative mapping.

(i) <u>Direct Mapping</u> → It maps each block of main memory into only one possible cache line.

128 words



(total 4 lines are in a cache)
(range = 0 to 3)

The direct mapping is expressed as $\boxed{i = j \text{ modulo } m.}$
where i = cache line number
j = main memory block number
m = number of lines in the cache.

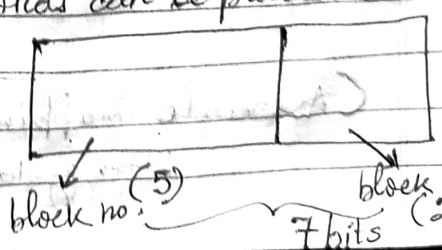
Pg 150

Ex block no = 10 = j
m = no. of lines = 4.
i = cache line = 10 mod 4 = 2.

How to generate address in direct mapping?

- If total is 128 words then range is 0-127
- To represent this in binary we need ~~to have~~ 7 bits (as $2^7 = 128$)
- Physical address will be represented by 7 bits.
- Again, this physical address is divided into 2 parts.

Block offset / Block Size
- It means how many words can be placed in a block, ie 4
- To represent 4 (range 0-3) we needs 2 bits.



- 5 bits will represent block no and 2 bits will represent block offset. (7 − 2 = 5)

Let the CPU generated an address 0001010 (decimal = 10)

W_0, W_1, W_2, W_3
W_4, W_5, W_6, W_7
W_8, W_9, W_{10}, W_{11}
⋮

Out of 0001010

- 00010 → block no. (00010) → (B_2)
- 10 → block offset (10)

ie. 0, 1, 2 → (W_8, W_9, W_{10})

How the address will be generated in cache?

5	2	→	3	2	2
			Tag		Line no.

→ Line L_0 can have neither B_0 or B_4 or B_8 or B_{12} but which block is present in that ~~block~~ tag will indicate

→ Let the tag is 010 and line no. is 0

→ This address means B_8 is present in this location of this cache.

0100000	→ 32 words
0100001	→ 33 words
0100010	→ 34 words
0100011	→ 35 words

Q) If ~~the~~ the tag is 001 in Line no. 1 then which block is present in the cache?

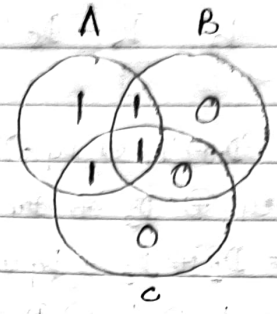
0010100	→ 20	B_5 (Block → 5)
0010101	→ 21	
0010110	→ 22	
0010111	→ 23	

(Associate mapping and Set Associate mapping)
(book pg-160) (self study)

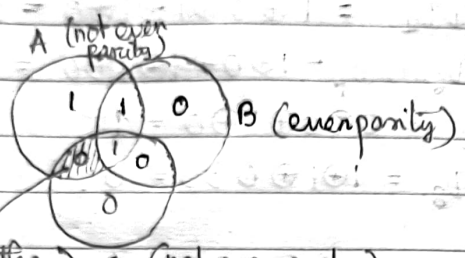
Chapter - 5

INTERNAL MEMORY

(Pg 200) **Error Correction**



→ for even parity



(Here, the error is detected).

Hamming Code.

If $(1110)_2$ then.

Data is M bit wide ⇒ 4

Check bit is k bit wide ⇒ 3

$\boxed{2^k - 1 \geq M + k.}$

Q) If data is having 8 bits ie (M=8) then what will be the k bit?

Ans) $2^k - 1 \geq M + k$

If $k=0$, $2^0 - 1 \geq 8+0$ (not satisfied) ✗

If $k=1$, $2^1 - 1 \geq 8+1$ (✗).

If $k=2$, $2^2 - 1 \geq 8+2$ (✗)

If $k=3$, $2^3 - 1 \geq 8+3$ (✗)

If $k=4$, $2^4 - 1 \geq 8+4$ (15 > 12) (✓).

∴ If M=8 then k=4.

Data = 00111001

Bit Position	12	11	10	9	8	7	6	5	4	3	2	1
Position no:	1100	1011	1010	1001	1000	0111	0110	0101	0100	0011	0010	0001
Data bit:	D_8	D_7	D_6	D_5		D_4	D_3	D_2		D_1		
check bit:					C_8				C_4		C_2	C_1
Word stored as:	0	0	1	1	0	1	0	0	1	1	1	1
Word fetched as:	0	0	1	1	0	1	1	0	1	1	1	1
Position no:	1100	1011	1010	1001	1000	0111	0110	0101	0100	0011	0010	0001
check bit:					0				0		0	1

(Fig. : Check Bit Calculation)

0	0	1	1	1	0	0	1
D_8	D_7	D_6	D_5	D_4	D_3	D_2	D_1

XOR of odd no. 1 = 1 else

checkbit
$$\begin{cases} C_1 = D_1 \oplus D_2 \oplus D_4 \oplus D_5 \oplus D_7 = 1 \oplus 0 \oplus 1 \oplus 1 \oplus 0 = 1 \\ C_2 = D_1 \oplus D_3 \oplus D_4 \oplus D_6 \oplus D_7 = 1 \oplus 0 \oplus 1 \oplus 1 \oplus 0 = 1 \\ C_4 = D_2 \oplus D_3 \oplus D_4 \oplus D_8 = 0 \oplus 0 \oplus 1 \oplus 0 = 1 \\ C_8 = D_5 \oplus D_6 \oplus D_7 \oplus D_8 = 1 \oplus 1 \oplus 0 \oplus 0 = 0 \end{cases}$$

$C_8 C_4 C_2 C_1 = 0111$

New checkbits for words fetched:

$C_1 = D_1 \oplus D_2 \oplus D_4 \oplus D_5 \oplus D_7 = 1 \oplus 0 \oplus 1 \oplus 1 \oplus 0 = 1$
$C_2 = D_1 \oplus D_3 \oplus D_4 \oplus D_6 \oplus D_7 = 1 \oplus 1 \oplus 1 \oplus 1 \oplus 0 = 0$
$C_4 = D_2 \oplus D_3 \oplus D_4 \oplus D_8 = 0 \oplus 1 \oplus 1 \oplus 0 = 0$
$C_8 = D_5 \oplus D_6 \oplus D_7 \oplus D_8 = 1 \oplus 1 \oplus 0 \oplus 0 = 0$

$C_8 C_4 C_2 C_1 = 0001$ (new)

Bitwise XOR of prev and new checkbit =

$$\begin{array}{r} C_8 C_4 C_2 C_1 = 0111 \\ C_8 C_4 C_2 C_1 \oplus 0001 \\ \hline 0110 \end{array}$$

This is called SYNDROME

(i) If syndrome is 0000 → No error
(ii) If there is 1 in 0000 → Error is there but in checkbit (ie. syndrome is three zeros with one 1) (No correction is needed)
(iii) If syndrom is two zeros with two 1's → error is there but in the data bit. (inverted for correction).

To generate a 4-bit syndrome for an 8-bit data word with the following characteristics :

(i) If the syndrome contains all 0s, no error has been detected.

(ii) If the syndrome contains one and only one bit set to 1, then an error has occurred in one of the 4 check bits. No correction is needed.

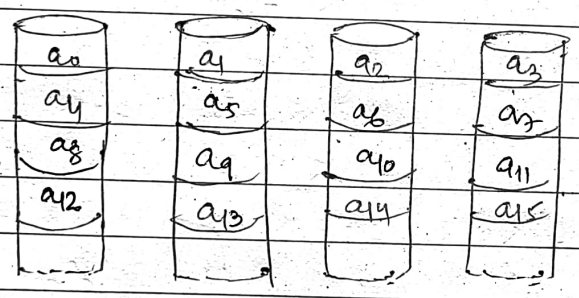
(iii) If the syndrome contains more than one bit set to 1, then the numerical value of the syndrome indicates the position of the data bit in error. This data bit is inverted for correction.

Chapter-6 EXTERNAL MEMORY

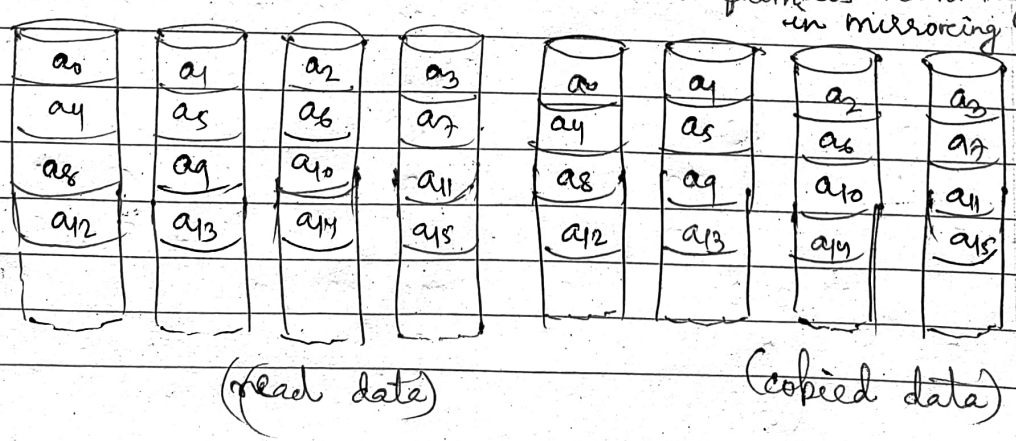
Pg 229

RAID (Redundant Array of Independent Disks)

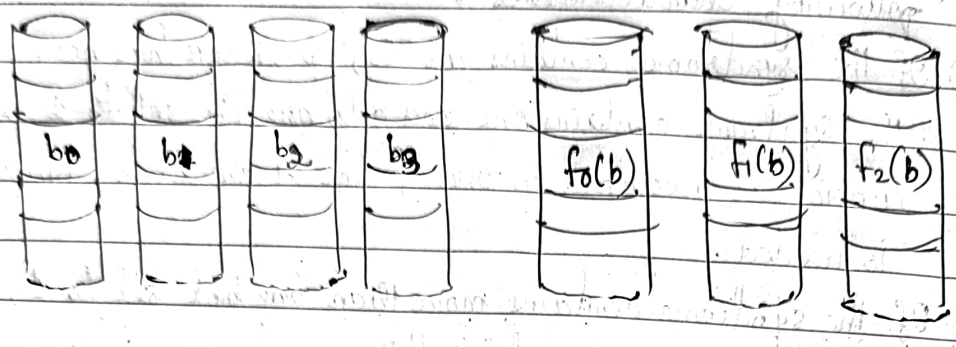
RAID 0 (Nonredundant)



RAID (1) (Mirrored) : If incase the data will be lost then it can be recovered from its hamming code in mirroring



(read data) (copied data)

Date
Page

RAID 2 (Redundancy through hamming code)