Chapter 4.

CACHE MEMORY

Problem 1.

When access time of Main Memory is 3500 ns and Cache Memory is 125 ns and miss ratio is 0.3. Calculate the average access time of CPU?

Miss ratio + hit ratio = 1

Average access time = Access time of CM * hit ratio + Access time of MM * miss ratio

=125*0.7 + 3500*0.3

= 1137.5 ns

Example 1.

Suppose that the processor has access to two levels of memory. Level 1 contains 1000 words and has an access time of 0.01 ms; level 2 contains 100000 words and has an access time of 0.1 ms. Assume that if a word to be accessed is in level 1, then the processor accesses it directly. If it is in level 2, then the word is first transferred to level 1 and then accessed by the processor. For simplicity, we ignore the time required for the processor to determine whether the word is in level 1 or level 2. Suppose 95% of the memory accesses are found in the cache. Prove that, for high percentages of level 1 access, the average total access time is much closer to that of level 1than that of level 2.

The average total access time = 0.95x0.01+ 0.05x(0.01+0.1) =0.015=0.01, which is approximately closer to level 1

Problem 4.1

A set-associative cache consists of 64 lines, or slots, divided into four-line sets. Main memory contains 4K blocks of 128 words each. Show the format of main memory addresses.

Answer:

The cache is divided into 16 sets of 4 lines each.

Number of sets = 64/4 = 16
=> The cache is divided into 16(= 2^4) sets

Therefore, 4 bits are needed to identify the set number.

To represent 16 we need = 4 bits

Main memory consists of 4KB = 2^{12} blocks.

Block number = Set number + Tag number

[Set is the part of block; Tag is also a part of block]

Therefore, the set + tag lengths must be 12 bits

Tag length = 12 - 4 = 8

Each block contains 128 words = 2^7 words

Therefore, 7 bits are needed to specify the word field

So, the main memory can be represented as

Main memory address =

TAG	SET	WORD
8	4	7

Problem 4.2

A two-way set-associative cache has lines of 16 bytes and a total size of 8 Kbytes. The 64-Mbyte main memory is byte addressable. Show the format of main memory addresses.

Answer:

- Two way set associative mapping.
- Size of each cache line = 16 bytes
- Size of cache memory = 8 KB
- Number of Blocks at the cache = 8192 byte/16byte=512 blocks (lines in the cache)
- It is two way associative, so the cache consists of 2 sets
- Each set consists of =512/2=256 sets of 2 line each
- Number of bits required for "SET field" in main memory address format is 8-bit (Because 2^8=256)
  Set number is =8

Main memory 64 M byte= 2^ 26 (26-bit address memory)

- Number of Blocks at the Memory= 64 M byte/16 byte=4 M blocks=2^22
- So, the number of blocks is = 22
  We know, Block number = Set number + Tag number

Thus, Tag field =22 - 8=14 bits

The remains will be the Word length or
Block size=16 byte= 2^4 (Word length = 4)

Main memory address=

TAG	SET	WORD
14	8	4

Set:

Main memory = Hard disk and cache memory

To represent CM we use SET

To represent HD we use TAG

Block of main memory = Set + Tag

Problem 4.8

Consider a machine with a byte addressable main memory of 2^16 bytes and block size of 8 bytes. Assume that a direct mapped cache consisting of 32 lines is used with this machine.

a) How is a 16-bit memory address divided into tag, line number, and byte number?

b) Into what line would bytes with each of the following addresses be stored?

0001 0001 0001 1011

1100 0011 0011 0100

1101 0000 0001 1101

1010 1010 1010 1010

c) Suppose the byte with address 0001 1010 0001 1010 is stored in the cache. What are the addresses of the other bytes stored along with it?

d) How many total bytes of memory can be stored in the cache?

e) Why is the tag also stored in the cache?

Ans.

a)
Address length: 16 bits
Word: 3 rightmost bits (as the block size is 8 bytes)
Line: 5 middle bits (as there are 32 lines)
Tag: 8 leftmost bits (as 16 – 3 – 5=8)
1000 1011 1100 0101

b)
0001 0001 0001 1011 → line 3
1100 0011 0011 0100 → line 6
1101 0000 0001 1101 → line 3
1010 1010 1010 1010 → line 21

c)
Bytes with addresses
0001 1010 0001 1000
0001 1010 0001 1001
0001 1010 0001 1010
…
0001 1010 0001 1111
are stored in the cache.

d)
32 lines * 8 bytes = 256 bytes

e)
It is because 2 items with two different memory addresses can be stored in the same place in the cache. The tag is used to distinguish between them.

Problem 4.19

Consider a memory system with the following parameters:

Tc = 100 ns Cc = 10^-4 $/ bit

Tm = 1200 ns Cm = 10^-5 $/ bit

a. What is the cost of 1 MByte of main memory?

b. What is the cost of 1 MByte of main memory using cache memory technology?

c. If the effective access time is 10% greater than the cache access time, what is the hit ratio H?

Ans:

1 MB = 1024 KB = 1024*1024*8 = 8388608 bits*10^-5= 83.88 $

1 Kbit = 1000 bit

1MBytes=8000000 = 8X10^6 bit

1GBytes=8000000000 = 8X10^9 bit

a.

Cost = Cm * 1 MB = 10^**-5** * 8 * 10^6 $ = 80 $ = $80

b.

Cost = Cc * 1 MB = 10^**-4** * 8 * 10^6 $= 800 $ = $800

C.

Effective Access Time:
Te = H x Tc + (1 - H)(Tm + Tc),
Tc access time of cache

Tm access time of main
Tc = 100ns,
Te = 1.1 x Tc
Tm = 1200ns.

Thus,
(1.1) x (100) = H x 100 + (1 - H) (1200 + 100)
110 = 100H + 1300 - 1300H
H = 1190/1200