King Fahd University of Petroleum and Minerals
Department of Computer Engineering

COMPUTER ARCHITECTURE COE 308

Homework 4

Student Name:…….

Student ID: …….

Problems	Grading
Question 1 / 10	
Question 2 / 10	
Question 3 / 10	
Question 4 / 10	
TOTAL	

1 – Hierarchical Memory Systems

A computer has a processor P with a hierarchical memory system that consists of a 4-ways set-associative cache memory CM and a main-memory MM. P generates a 32-bit memory address ADDR = (TAG, Index, Word). CM has 64-entries. The page size is 32 bytes, where each word addressed by P is one byte. MM has Tmm=18 ns access time. Denote by (Tcm) and (Ph) as CM access time and its hit probability, respectively.

Answer each of the following questions:

1. Find the number of bits in "Index", "Word", and "TAG".
2. Determine the page size in bits.
3. Determine the cache capacity in bits.
4. Determine the main memory capacity in bits.
5. CM can be designed in the following ways: (1) as a 4-way set-associative cache with Tcm = 1.25 ns and Ph = 0.88, (2) as a fully-associative cache with Tcm = 1.5 ns and Ph = 0.97, (3) as a direct-mapped cache with Tcm = 1 ns and Ph = 0.80. Answer each of the above questions:
    a. Which of the above three cache designs has the least conflict misses.
    b. For what reason the direct-mapped cache has the least access time.
    c. For what reason the cache with largest hit probability has the largest access time.
    d. Determine which of the above three cache designs gives the best global performance of hierarchical memory system including CM and MM.

6. The processor is running a benchmark with the following instruction distribution: (1) 55% of R-type, (2) 18% are loads, (3) 12% are stores, and (4) 15% are branching instructions. Suppose the retained cache solution is as a 4-way set-associative cache with Tcm = 1.25 ns and Ph = 0.8 which is used as an instruction cache and as a data cache. Answer each of the following questions:
    a. Evaluate the average access time of the memory.
    b. Evaluate the average stall time per instruction (Tstall).
    c. Evaluate the CPI for this processor if the ideal CPI is 2 clocks and processor clock rate is 1 GHz.

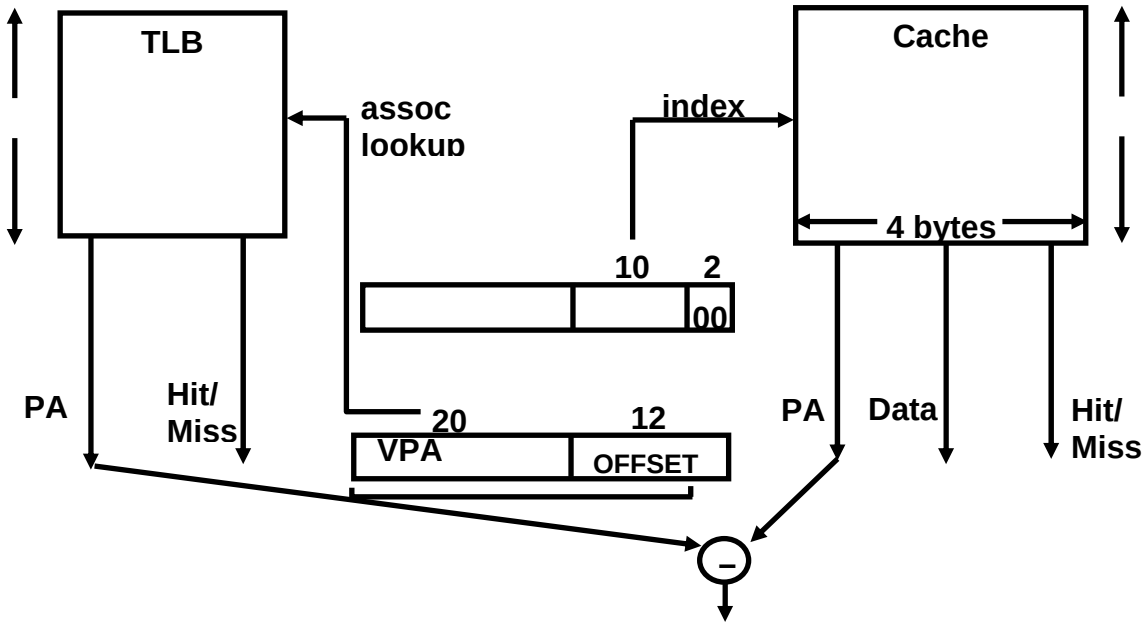
Solution:

1. Since CM has 64-entries, then Index is 6 bits. The page size is 32 bytes, then Word is 5 bits. As ADDR is 32 bits, then 32 - (6+5) = 21 bits which is the number o bits in TAG.
2. The page size is 32 x 8 = 256 bits
3. The cache capacity is 4 x 64 x (21 + 256) = 70912 bits.
4. The main memory capacity is 2^{32} x 8 = 2 ^{35} bits = 32 Gbits.
5. First, considering the three cache designs, the fully-associative cache has the least conflict misses. Second, the direct-mapped cache has the least access time because it has the simplest access hardware. Third, the fully-associative cache has least hit probability and largest access time because it has a 4-to-1 MUX to select one page out of four. Fourth, the above three cache designs have the following average access times: (1) Tav = 1.25+ (1-0.88) x 18= 3.41 ns, (2) Tav = 1.5+ (1-0.97) x 18= 2.04 ns, (3) Tav = 1 + (1- 0.8) x 18= 4.6 ns. Therefore the least Tav corresponds to 2.04 (fully-associative) which give the best global performance of hierarchical memory system.
6. First, the memory average access time is Tav = 1.25+ 0.8 x 18= 4.85 ns. Second, the average stall time per instruction Tstall= 1 x (1 -0.8) x 18 + (0.18 +0.12) x (1-0.8) x 18 = 4.68 ns. Third, the CPI= 2 + Tstall = 2 + 4.68 = 6.68 ns.

2 – Virtual Memory System

A processor generates a Virtual Address VA = (VPA, Offset) where VPA is the virtual page address and Offset is the offset of the addressed word within a given VP. To shorten address translation time, the parallel translation scheme is used by simultaneously sending parts of the VA to a TBL and to a direct mapped cache memory.

a. Give a block diagram of the parallel address translation by referring to CPU, VA, TLB, Cache, Main Memory, and Disk Memory.



IF cache hit AND (cache tag = PA) then deliver data to CPU
ELSE IF [cache miss OR (cache tag = PA)] and TLB hit THEN
access memory with the PA from the TLB
ELSE do standard VA translation

b. Answer only by writing text each of the following issues:

a) What part of VA is used to access the TLB,
The Virtual page address VPA.

b) how a hit or miss if found in TLB,
TLB is an associative memory which is searched using VPA.
A hit means VPA is found.

c) how to service a hit or a miss in TLB
A hit leads to outputting the corresponding physical page address PPA. A miss lead to fetching a VP from Disk, make placement in MM and update the TBL for translation.

d) What is retrieved from TLB
Once VPA is found in TLB, the corresponding physical page address PPA is retrieved from TLB to help computing the physical address.

e) What part of VA is used to access the Cache
Some part of OFFSET is used as index for accessing the cache.

f) how a hit or miss if found in cache,
A Hit if the TAG from (PPA,OFFSET) matches the stored TAG. Else a miss.

g) How to service a hit or a miss in cache
Hit: retrieve page and select word using WORD. Miss: retrieve page from MM, placement in cache, service the CPU and update cache entry.

h) What is retrieved from the cache.
Only when we have a hit we retrieve a page and select a word for the CPU.

3 - Multiprocessor System

1. Shortly describe the main difference between SIMD and MIMD multiprocessors in terms of number of decoding units, and style of communication among processors.
2. Consider the following sequential program (SP):

```
For i=0,N-1
    A(i)= (b(i) - c(i)) * b(i)
Endfor
```

Where N=1000.
Answer each of the following questions:
Translate program SP to an SIMD program (SP-SIMD) targeted to an SIMD multiprocessor with NP=20 as the number of processing elements. Each processing element consists of an ALUs and a small set of registers. Your SP-SIMD program should refer to loading array data, computing, and storing results. Assume local array addresses are already available in local registers.

Solution:

1. The main difference between SIMD and MIMD multiprocessors are (1) SIMD has one Decoding unit (DU) while MIMD has multiple DUs, (2) SIMD operates as a lock step computer where all unit mus start and complete in synchrony while MIMD processors are completely asynchronous, (3) SIMD PEs may communicate by either message passing or through shring memory while in MIMD the processors may share a memory or communicate by exchanging packet messages.

2. Translating the program SP to an SIMD program (SP-SIMD):

```
For I = 0, N/NP-1   ; outer loop over vectorized body
        Load rb, b     ; rb is some vector register having NP components
        Load rc, c     ; rc is some vector register having NP components
        Subtract reg, rb,r c   ; vector subtract
        Multiply reg, reg, rb  ; vector multiply
        Store a, reg           ; vector store
Endfor
```

The array addresses fetched are function of PEs numbers, i.e. vector data is fetched and each PE buffers its part in the vector. The same process is done in storing vector data.