HW5

Question 1:

Below is a list of 32-bit memory address references, given as **word addresses**:
3, 180, 43, 2, 191, 88, 190, 14, 181, 44, 186, 253

a. For each of the above, identify the binary address, the tag, and the index given a direct-mapped cache with 16 one-word blocks. Also list if each reference is a hit or a miss, assuming the cache is initially empty.

b. For each of the above, identify the binary address, the tag, and the index given a direct-mapped cache with 2-word blocks and a total size of 8 blocks. Also list if each reference is a hit or a miss, assuming the cache is initially empty.

c. In terms of miss rate, which of the following three direct-mapped cache design is best, all with 8 words of data: C1 (1-word blocks), C2 (2-word blocks), or C3 (4-word blocks), if the miss stall time is 25 cycles, and C1 has an access time of 2 cycles, C3 takes 3 cycles and C3 5 cycles, which is the best design?

For the following questions, consider a direct-mapped cache with the following parameters:

Cache Data Size: 32 KiB
Cache Block Size: 2 words
Cache Access Time: 1 cycle

d. What is the total number of bits required for the above cache? Assuming a 32-bit address and that the only overhead bit needed is validity. Find the total size of the closest direct-mapped cache with 16-word blocks of total equal size or greater. Explain why the second cache, despite its larger size, might provide slower performance compared to the first cache.

e. Generate a series of read requests that have a lower miss rate on a 2KiB 2-way set associative cache than the cache listed above.

f. In class, you saw that the method to index a direct-mapped cache is (Block address) modulo (Number of blocks in the cache). Assuming a 32-bit address and 1024 blocks in the cache, consider a different indexing function: (Address[31:27] XOR Address[26:22]). Is it possible to use it to index a direct-mapped cache? If so, explain why and discuss any changes that might need to be made to the cache. If it is not possible, explain why.

Question 2:

Consider a video streaming workload that accesses a 512 KiB working set sequentially with the following address stream:

0, 2, 4, 6, 8, 10, 12, 14, 16, …

a. Assume a 64 KiB direct-mapped cache with 32-byte block. What is the miss rate for the address stream above? How is this miss rate sensitive to the size of the cache or working set? How would you categorize the misses this workload is experiencing, based on the 3C model?

b. Re-compute the miss rate when the cache block size is 16 bytes, 64 bytes, and 128 bytes. What kind of locality is the workload exploiting?

c. "Prefetching" is a technique that leverages predictable address patterns to speculatively bring additional cache blocks when a particular cache block is accessed. One example of prefetching is a stream buffer that prefetches sequentially adjacent cache blocks into a separate buffer, it is considered as a hit and moved into the cache and the next cache block is prefetched. Assume a 2-entry stream buffer and assume that the cache latency is such that a cache block can be loaded before the computation on the previous cache block is completed. What is the miss rate for the address stream above?

Cache block size (B) can affect both miss rate and miss latency. Assuming a 1-CPI machine with an average of 1.35 references (both instruction and data) per instruction, help find the optimal block size given the following miss rates for various block sizes.

8: 4%, 16: 3%, 32: 2%, 64: 1.5%, 128 1%

d. What is the optimal block size for a miss latency of 20xB cycles?

e. What is the optimal block size for miss latency of 24+B cycles?

f. For the constant miss latency, what is the optimal block size?

Question 3:

This question compares direct-mapped caches to associative caches.

Below is a list of 32-bit memory address references, given as **word addresses**:

3, 180, 43, 2, 191, 88, 190, 14, 181, 44, 186, 253

a. Using the sequence of references, show the final cache contents for a 3-way set associative cache with 2-word blocks and a total size of 24 words. Use LRU replacement policy. For each reference identify the index bits, the tag bits, the block offset bits, and if it is a hit or a miss.

b. Using the sequence of references, show the final cache contents for a fully associative cache with 1-word blocks and a total size of 8 words. Use LRU replacement policy. For each reference identify the index bits, the tag bits, the block offset bits, and if it is a hit or a miss.

c. Using the sequence of references, what is the miss rate for a fully associative cache with 2-word blocks and a total size of 8 words? Use LRU replacement policy.

In the following questions consider a processor with the below parameters:

Base CPI: 1.5

Processor Frequency: 2 GHz

Main Memory Access Time: 100 ns

First Level Cache miss rate per instruction: 7%

Second Level Cache – Direct-mapped access time: 12 cycles

Global Miss Rate with Direct-mapped Second Level Cache: 3.5%
Second Level Cache – 8-way Set Associative access time: 28 cycles
Global Miss Rate with 8-way Set Associative Second Level Cache: 1.5%

d. Calculate the Load/Store average CPI for the above processor using:

   Only a first level cache
   A direct-mapped second level cache
   An 8-way set associative second level cache

   How do these numbers change if main memory access time is doubled? If it is cut in half?
e. Will adding a third level cache that takes 50 cycles to access and with a global miss rate of 1.3% provide better performance? Assume a direct-mapped second level cache. In general, what are the advantages and disadvantages of adding a third level cache?
f. Assuming a 512 KiB second level cache has a global miss rate of 4%. If each additional 512 KiB of cache lowered global miss rates by 0.7%, and the cache had a total access time of 50 cycles, how big would the cache have to be to match the performance of the direct-mapped second level cache above? Of the 8-way set associative second level cache above?

Question 4:

The following are parameters of a virtual memory system:
Virtual Address: 43 bits
Physical DRAM Installed: 16 GiB
Page Size: 4 KiB
PTE Size: 4 bytes

a. For a single-level page table, how many page table entries (PTEs) are needed?
b. How much physical memory is needed for storing the page table?

Question 5:

The following data constitutes a stream of virtual addresses as seen on a system. Assume 4 KiB pages, a 4-entry fully associative TLB, and LRU replacement. If the pages must be brought from disk, increment the next largest page number.

Address stream:
4669, 2227, 13916, 34587, 48870, 12608, 49225

TLB:

Valid	Tag	Physical Page #
1	11	12
1	7	4
1	3	6
0	4	9

Page Table:

Valid	Physical Page in Disk
1	5
0	Disk
0	Disk
1	6
1	9
1	11
0	Disk
1	4
0	Disk
0	Disk
1	3
1	12
0	Disk

a. Given the address stream, and the initial TLB and page table states provided above, show the final state of the system. Also list for each reference if it is a hit in the TLB, a hit in the page table, or a page fault.

b. Repeat the previous question, but this time use 16 KiB pages instead of 4 KiB pages. What would be some of the advantages and disadvantages of having a larger page size?

c. Show the final contents of the TLB if it is a 2-way set associative. Also show the contents of the TLB if it is direct-mapped. Discuss the importance of having a TLB to high performance systems. How would virtual memory access be handled if there were no TLB?

There are several parameters that impact the overall size of the page table. Listed below are key page table parameters:
Virtual Address Size: 32 bits
Page Size: 8 KiB

Page Table Entry Size: 4 bytes

d. Given the parameters above, calculate the total page table size for a system running 5 applications that utilize half of the virtual memory available to them.

Question 6:

Assume a 2-way set associative cache with 4 blocks.

Consider the following address sequence (these are block numbers):

0, 2, 4, 8, 10, 12, 14, 16, 0

a. Assuming an LRU replacement policy, how many hits does this address sequence exhibit?

b. Which address should be evicted at each replacement to maximize the number of hits? How many hits does this address sequence exhibit if you follow this policy?

c. Describe why it is difficult to implement a cache replacement policy that is optimal for all address sequences.

d. Assume you could decide upon each memory reference if you want the requested address to be cached or not. What impact could this have on the miss rate?

*Note: In this question, use a table like the below table for the address sequence: 0, 1, 2, 3, 4

Address of Memory Block Accessed	**Hit or Miss**	**Evicted Block**	**Contents of Cache Blocks After Reference**			
			Set 0	Set 0	Set 1	Set 1
0	Miss		Mem[0]			
1	Miss		Mem[0]		Mem[1]	
2	Miss		Mem[0]	Mem[2]	Mem[1]	
3	Miss		Mem[0]	Mem[2]	Mem[1]	Mem[3]
4	Miss	0	Mem[4]	Mem[2]	Mem[1]	Mem[3]
…						