CS320 – Spring 2017
Homework assignment 5

Due: Frioday, May 5th, in class

1. Consider a pipelined processor with just one level of cache. Assume that in the absence of memory delays, the baseline CPI of this processor is 2. Now assume that the percentage of memory instructions in a typical program executed on this CPU is 50% and the memory access latency is 150 cycles. Assuming that the I-cache delays are already accounted for in the baseline CPI, consider the following two alternatives for the D-cache design:

   Alternative 1: A small D-cache with a hit rate of 94% and a hit access time of 1 cycle (assume that no additional cycles on top of the baseline CPI are added to the execution on a cache hit in this case).

   Alternative 2: A larger D-cache with a hit rate of 98% and the hit access time of 2 cycles (assume that every memory instruction that hits into the cache adds one additional cycle on top of the baseline CPI).

   a) [10%] Estimate the CPI metric for both of these designs and determine which of these two designs provides better performance. Explain your answers!

   b) [5%] Repeat part (a), but now assume that the memory access latency is reduced to 50 cycles.

   c) [5%] Repeat part (b), but now assume that the L2 cache is also added to the system with the hit rate of 75% and access latency of 10 cycles.

2. Consider the use of multi-level hot-cold bits in implementing a replacement algorithm for a set-associative cache with 2^p ways and S sets. This is an approximation of the LRU policy that marks the most recently assessed way as hot, and the other one as cold (for a 2-way set-associative cache). This mechanism can be extended to more than two ways by providing multiple levels of hot-cold bits. For example, for a 4-way cache the victim is determined as the cold way of the cold pair.

   a) [5%] What is the total number of hot-cold bits for each set in this cache?

   b) [5%] What is the total number of hot-cold bits in the cache?

   c) [5%] Assume that p = 2 and accesses to a specific set result to the following ways in sequence and no misses occur:

      Way 0, Way 3, Way 2, Way 0, Way 1, Way 1, Way 2, Way 3, Way 2, Way 0

      At the end of the fifth access in this sequence (to Way 1), what are the settings of the hot-cold bit? Do they point to the same victim that would have been selected by a true-LRU algorithm?

d) [5%] Repeat Part (c), assuming that the all accesses shown in the sequence given in Part (c) have been completed with the last access being made to Way 0.

3. Consider a system with physically-addressed caches, and assume that 40-bit virtual addresses and 32-bit physical addresses are used, and the memory is byte-addressable. Further assume that the cache is 4-way set-associative, the cache line size is 64 Bytes and the total size of the cache is 64 KBytes. Answer the following questions, providing adequate explanations in all cases:

   a) [5%] What should be the minimum page size in this system to allow for the overlap of the TLB access and the cache access?

   b) [5%] Repeat part (a) assuming that the cache associativity is increased to 8. Assume that the total cache size and the cache line size remain the same.

   c) [10%] Assuming that the memory page size in this system is as calculated in your answer to Part (b), compute the total size of the page table in bytes. Assume that a simple linear page table is used and only the page translation information is stored in each entry, with no additional bits.

   d) [10%] Repeat part (b) assuming that the OS can guarantee that the two least significant bits of the page number would not change during the address translation.

   e) [10%] Assume that an access to a virtual address 0x7FFF5A3B resulted in a page fault. When the page containing this address is swapped in to physical memory to service the page fault, determine which frames in the physical memory can be used to store the contents of this page to support the OS guarantee described in part (d).

4. [20%] Consider two cache designs, each with a total cache capacity of 1KB. The first design is a 2-way set associative cache with the line size of 64 bytes. The second design is a 4-way set associative cache with the line size of 4 bytes. Assume that both caches use the LRU replacement policy and that the memory is byte-addressable. Trace the cache hit/miss outcomes of the following stream of byte addresses by filling the table shown below. Also indicate the set number of each access. Explain how you arrived at your answers.

Byte Address	2-Way Hit/Miss?	Set index	4-Way Hit/Miss?	Set Index
0				
2				
4				
0				
40				
86				
0				