CS 282 – Computer Architecture and Organization – Assignment 3 (50 points)

Memory Hierarchy

Q1. You are building a system around a processor with in-order execution that runs at 1 GHz and has a CPI of 1.2 excluding memory accesses. The only instructions that read or write data from memory are load (20% of all instructions) and stores (5% of all instructions).

The memory system for this computer is composed of an I-cache and a D-cache with a 1 cycle hit time. Both caches are direct mapped and hold 32 KB each. The I-cache has a 2% miss rate and 32-byte blocks, and the D-cache is write through with a 5% miss rate and 16-byte blocks. There is a write buffer on the D-cache that eliminates stalls for 95% of all writes.

The memory system has a 512 KB write-back L2 cache, with 64-byte blocks and an access time of 10 ns after which a block can be transferred to the L1 caches. It is connected to the L1 cache by a 128-bit data bus that runs at 500 MHz and can transfer 128 bits per bus cycle. Of all memory references sent to the L2 cache in this system, 80% are satisfied without going to main memory.

The 128-bit-wide main memory has an access latency of 60 ns, after which any number of bus words may be transferred at the rate of one per cycle on the 200 MHz main memory bus.

a. What is the miss penalty in the I-cache (time to transfer a block from L2 to I-cache), the miss penalty in the D-cache, and the miss penalty in the L2 cache (time to transfer a block from memory)? Show the answers in nanoseconds and in clock cycles.

b. What is the average memory access time for instruction accesses (nsec and clock cycles)?

c. What is the average memory access time for data reads (nsec and clock cycles)?

d. What is the average memory access time for data writes (nsec and clock cycles)?

e. What is the overall CPI in the presence of memory stall cycles?

f. You are considering replacing the 1 GHz CPU with one that runs at 2 GHz, but is otherwise identical. How much faster does the system run with a faster processor? Assume the L1 caches have a hit time of 1 cycle (faster clock), and that the speed of the L2 cache, main memory, and buses remains the same in absolute terms.

Q2. Smith and Goodman [1983] found that a small direct-mapped instruction cache consistently outperformed a fully associative instruction cache of the same size using LRU replacement.

a. Explain how this would be possible focusing on the replacement policy and a small cache size.

b. Explain where replacement policy fits into the three C's model, and explain why this means that misses caused by a replacement policy cannot in general be definitively classified by the three C's model.

Q3. McFarling [1989] found that the best memory hierarchy performance occurred when it was possible to prevent some instructions from entering the cache.

a. Explain why McFarling's result could be true.

b. The four memory hierarchy questions form a model for describing cache designs. Where does a cache that does not always *read-allocate* fit or not fit into this model?

Q4. As a block is referenced inside the cache, the next block in memory can be prefetched into the cache. This prefetching technique attempts to eliminate some of the cache misses.

a. As caches increase in size, blocks often increase in size as well. If a large instruction cache has large data blocks, is there a need for prefetching? Explain the interaction between prefetching and increased block size in instruction caches.

b. Is there a need for data prefetch instructions when data blocks get larger in a data cache? Explain.

Q5. Some memory systems handle TLB misses in software (as an exception), while others use hardware for TLB misses.

a. What are the tradeoffs between these two methods for handling TLB misses? Are there page table structures that would be difficult to handle in hardware, but possible in software?

b. Use the data in the following tables to calculate the penalty of TLB misses on the CPI on the following two workloads, assuming hardware TLB handlers require 100 cycles per miss and software TLB handlers take 300 cycles per miss.

First Workload

Program	Weight	TLB misses/1000 instructions
gcc	50%	0.30
perl	25%	0.26
ijpeg	25%	0.10

Second Workload

Program	Weight	TLB misses/1000 instructions
swim	30%	0.10
wave5	30%	0.89
hydro2d	20%	0.19
gcc	20%	0.30