Problem Set #2: Cache Coherence

DUE MONDAY, MAY 8 (no late days)

This problem set requires you to solve problems from the *Computer Architecture: A Quantitative Approach* book. The questions reproduced here are drawn from *Parallel Computer Architecture: A Hardware/Software Approach* by Culler, Singh, and Gupta.

1) Write-through and write-back caches:

Assume the following average data memory traffic for a bus-based shared memory multiprocessor:

private reads: 70%
private writes: 20%
shared reads: 8%
shared writes: 2%.

Also assume that 50% of the instructions (32b each) are either loads or stores. With a split instruction/data cache of 32KB total size, we get hit rates of 97% for private data, 95% for shared data, and 98.5% for instructions. The cache line size is only 16B.

We want to place as many processors as possible on a bus that has 64 data lines and 32 address lines. The processor clock is twice as fast as that of the bus, and the processor CPI is 2.0 before considering memory penalties. How many processors can the bus support without saturating if we use **(a) write-through caches with write-allocate strategy? (b) write-back caches?** Ignore cache consistency traffic and bus contention. The probability of having to replace a dirty block in the write-back caches on a miss that fetches a new block is 0.3. For reads, memory responds with data 2 cycles after being presented the address. For writes, both address and data are presented to memory at the same time. Assume that the bus is atomic and that processor miss penalties are equal to just the number of bus cycles required for each miss.

2) MESI vs. Dragon:

For each of the memory reference streams given in the following, compare the cost of executing it on a bus-based machine that supports **(a) the Illinois MESI protocol** and **(b) the Dragon protocol**. Explain the observed performance differences in terms of the characteristics of the streams and the coherence protocols.

Stream 1: r1 w1 r1 w1 r2 w2 r2 w2 r3 w3 r3 w3
Stream 2: r1 r2 r3 w1 w2 w3 r1 r2 r3 w3 w1
Stream 3: r1 r2 r3 r3 w1 w1 w1 w1 w2 w3

All of the references in the streams are to the same location: r/w indicates read or write, and the digit refers to the processor issuing the reference. Assume that all caches are initially empty, and use the following cost model: read/write cache hit—1 cycle; misses requiring a simple transaction on bus (bus upgrade, bus update) —60 cycles; and misses requiring whole cache block transfer—90 cycles. Assume all caches are write allocated.

3) Alternatives for invalidation:

H&P 6.6

4) Invalidate and update cache coherence schemes:

H&P 6.7