(b) Assume that the first load (L1) misses in the cache 20% of the time. In that case comment on whether slice precomputation is beneficial or not to increase overall performance.

Problem 8.8

(a) Consider a simple 5-stage pipeline that is single threaded. The pipeline treats every cache miss as a hazard and freezes the pipeline. While executing a benchmark assume that a L1 cache miss occurs every 100 cycles, and each L1 cache miss takes 10 cycles to satisfy if the block is found in L2 or 50 cycles if L2 misses as well. An L2 cache miss occurs after 200 cycles of computation. Assume that the CPI in the absence of cache misses is one. What is the actual CPI taking into account cache miss latencies?

(b) Now consider the same example above but assume that hardware is now 2-way multithreaded, similar to Figure 8.3. Assume that switching overhead is zero and there are two threads with identical cache miss behavior as described in the first case. What is the CPI of the each of the two programs on the 2-way multithreaded machine? Did the CPI improve? If yes, explain how? If not, explain why one should bother with 2-way multithreaded machine?

(c) Consider the above case but the switching overhead is 5 cycles. Again compute the CPI of each thread and explain why it increases or decreases or stays the same?

(d) Now consider the case that L2 miss latency jumped from 50 cycles to 500 cycles and switching overhead jumped from 5 cycles to 50 cycles. Compute the CPI in this machine?

Problem 8.9

The combination of two enhancements are considered to boost the performance of a chip multiprocessor. The enhancements are: 1) adding more cores or 2) adding more shared level 2 cache. The base chip has three cores and nine L2 cache banks. L2 cache can be added by adding cache banks and each cache bank uses three times the area of a core. Here is what we also know from all kinds of sources:

1) 60% of the workload can be fully parallelized; the rest cannot.
2) The core stall time due to L2 misses accounts for 30% of each core's execution time in the base configuration with four cache banks and four cores.
3) It is suspected that the amount of shared L2 cache per core should remain constant in order to keep the same miss rate.
3) Simulations have also determined that the miss rate of L2 decreases as the square root of its size per core. A conjecture is that the stall time in each core will also decrease as the square root of L2 size per cores.

The company that pays your paycheck has acquired a new technology to build large micro-chips, so that the next generation chips will have four times the area of current chips to dedicate to cores and L2 caches. Given what you know, what kind of best "first cut" design would you propose? A design is characterized by (# of cores, # of L2 cache banks). These numbers can be any integer. The design should be contained in the new chip. Estimate the speedup of your best design that takes advantage of the new chip real estate.

Problem 8.10

This problem is about alternative organizations for a directory maintaining coherence among L1 caches in a CMP with a shared L2 cache. The following is known about the architecture.