ECE 587/687 Final Exam

Name: ____________________________

Time allowed: 80 minutes
Total Points: 60
Points Scored: _____

Problem No. 1 (15 points)

Consider a computer system with a two level cache hierarchy, consisting of split L1 instruction and data caches and a unified L2 cache. The cache parameters and statistics for a program are as follows:

(i) L1 data cache hit ratio is 95%
(ii) L1 instruction cache hit ratio is 98%
(iii) L2 cache miss rate is 10 misses per 1000 instructions,
(iv) L2 cache hit latency is 10 cycles
(v) Average L2 miss latency = 100 cycles
(vi) Fetching instructions for the program requires one L1 cache access per instruction
(vii) 40% of the instructions in the program are load/store instructions. Each data read (or write) caused by a load (or store) instructions requires one access to the L1 data cache
(viii) The program’s CPI assuming a perfect L2 cache with no misses, is 1.0
(ix) The L2 cache is a blocking cache that processes only one miss at a time, and blocks all other accesses until the current miss returns from memory

Based on the above statistics, answer the following questions:

(a) (3 points) Calculate the actual CPI of the program.

(b) (6 points) Calculate the hit ratio of the L2 cache.
(c) **(6 points)** Assume that we added compression to the L2 cache. This change led to reducing the L2 miss rate to 4 misses per 1000 instructions. However, it introduced 5 cycles of additional latency (due to decompression) on every L2 cache access. Does this design change help or hurt performance, and by how much?

**Problem No. 2 (8 points)**

A computer architect in the processor research team for a microprocessor company has proposed adding a prefetcher to the L2 cache. He implements the prefetcher in a simulator and reports the simulator results to his manager. The following questions relate to the conversation between the researcher and his manager:

(a) **(3 points)** The researcher says: “My results show that the prefetcher has an accuracy of 50%”. Upon hearing that, the manager responds: “That is great. This means that by adding this prefetcher, we can eliminate half of the L2 misses”. Is the manager correct? Why or why not?

(b) **(5 points)** The researcher then reports: “I found that despite getting rid of many misses, the prefetcher was not able to obtain any improvements in processor performance”. Provide two possible explanations for this result.
Problem No. 3 (13 points)
A program is run on a Multiscalar architecture with 4 processing units (cores). The Multiscalar sequencer divides up the program into four tasks of length 1200, 1800, 500, and 2000 instructions, respectively. Each task can commit 2 instructions per cycle on average. All inter-task branches are predicted correctly, and all inter-task dependences are resolved in time and don’t cause any additional latency.

(a) (4 points) Calculate the execution time of this program in terms of number of cycles.

(b) (4 points) Calculate the overall IPC of this program.

(c) (5 points) Assume that one of the cores is unavailable and therefore the program is run on only 3 cores. Will that change the execution time of the program? If yes, what is the new execution time in terms of number of cycles?

Problem No. 4 (24 points)
(Note: For the following problems, please provide concise answers. Unnecessarily long answers will waste more time than is worth)

(a) (3 points) Pentium-M designers established a criteria that a performance optimization would be considered only if it incurs a power overhead of at most 3% for every 1% performance increase. Why did they use the 3% power overhead limit? Why not 2%, 4% or any other number?
(b) **(3 points)** State ONE advantage and ONE disadvantage of SMT:dual issue as compared to SMT:single issue.

(c) **(2 points)** State TWO potential benefits of runahead execution.

(d) **(3 points)** Consider a load instruction in the runahead mode, which accesses the runahead cache and the store buffer in parallel. Assume that there is a hit in the runahead cache. Also, assume that the load address matches one of the preceding stores in the store buffer. Should the load consume the data read from the runahead cache? Why or why not?

(e) **(2 points)** State ONE advantage and ONE disadvantage of PCM compared to DRAM.

(f) **(3 points)** In the Alloy cache paper, the miss latency of the DRAM cache was assumed to be 2x of the hit latency. If the miss latency were to become 8x of hit latency, do you expect Alloy cache’s performance advantage over LH-cache to increase, decrease, or stay the same? Why?
(g) (2 points) Which of the following techniques is the most effective in reducing compulsory misses, (i) Stream buffer, (ii) Victim cache, (iii) Compression?

(h) (3 points) Can the destination physical register used by an instruction be freed after the instruction has been committed? Why or why not?

(i) (3 points) Consider a design upgrade in which the simple 1-bit branch predictor used in a processor is being replaced by a more sophisticated PAp predictor with 10 bits of branch history. How does that change impact the effectiveness of each of the following processor features, (i) Trace cache, (ii) Runahead execution, (iii) Wakeup logic? For each feature, simply state whether the upgrade would make it MORE effective, LESS effective, or the SAME?