Third Trimester Examination

Name: _____

Student Id: _____

This is an open book examination. You are allowed to make use of the Internet, your class notes, etc. But you are not allowed to consult any human being, either in person or remotely. Attempt all questions.

Exam is graded out of 20, maximum grade points is 20.

Solutions are Due by Monday at midnight, Saudi Time. Given the need to submit the grades to the registrar for graduating students, this is not a negotiable deadline.

Question 1 (5 points)

The SpeedMonster (SM) is a 64-bit in-order, single issue processor running at 6GHz. The processor features an 8-stage pipeline. You have been tasked to estimate the performance of an important application.

Preliminary Analysis: A trace-based simulation analysis was performed, and the results show that under ideal memory conditions (no cache or TLB misses), the SM loses about 20% of its cycles due to pipeline bubbles. Thus, the Instruction Per Cycle metric (IPC) was 0.8. A separate study that takes into account the performance of the memory hierarchy showed the following:

L1 instruction cache: 0.95 hit ratio.
L1 data cache: 0.85 hit ratio.
L2 cache: 0.95 hit ratio.
TLB: 0.98 hit ratio.

System Parameters: Access to the L1 cache requires 2 cycles; access to the L2 cache requires 20 cycles; and access to memory requires 120 cycles. The page table is tree-based with 6 levels. Thirty per cent of the instructions executed by the application require a data access.

Your Task: Compute the IPC of the application when caching and TLB effects are considered. Based on your finding, what do you recommend to improve performance?

Show your reasoning and the steps of your derivation.

Question 2 (15 points)

We would like to design a multiprocessor whose basic building block is a chip with eight SM cores. A broadcast bus is chosen as the method of implementation to reduce cost and simplify the design. The bus speed is 6GHz. The unit of transfer is a cache line.

a. Construct a queuing model of the system. Show the various resources and how they are modeled (cores, caches, bus, and memory).
b. Using the system parameters and the memory performance shown in problem 1, construct a simulation to account for queuing effects of the system. Use the simulator to find the maximum number of chips that can be connected to the system while keeping the bus utilization under 80%.

For step b, make any assumption about missing data that you may need (generally not needed). Simplify the queuing system to ensure that the simulator can be written within a reasonable time.

Hint: A process can be simulated as follows:
For any instruction, there is a probability of 20% of a pipeline stall.
For any instruction, there is a probability of 30% that a data access is needed.
For any instruction that use the data cache, the probabilities of L1, L2 and TLB misses apply.
For simplification, use a uniform distribution.
There is one process per core.
A spreadsheet model can be misleading here, because it will not model the queuing effect on the bus and other parts of the system.

You will need to deliver the simulator code along with the answer. Please use either a MacOS or a Linux machine.