Dept. of Computer Science and Engineering, National Sun Yat-sen Univ.
Spring Semester of 2021 PhD Qualifying Exam
Subject: Computer Architecture

Problem 1: 1. All the following questions are multiple-choice questions. Please explain your reasons for the choices, which you don't select. (10%)

____ (1) (2%) A computer has 128MB of memory. Each word in the computer is eight bytes. How many bits at least are needed to address any single word in memory?
(a) 8 bits (b) 16 bits (c) 24 bits (d) 32 bits

____ (2) (2%) In modern computer architectures, TLB is usually involved to improve the efficiency of the memory hierarchy. If we meet a situation of TLB miss, what is the following description true?
(a) The data requested by the CPU must be not in the cache.
(b) The data requested by the CPU must be not in the main memory.
(c) The CPU is failed to get the physical address of the required data.
(d) The CPU must send a request to access the main memory immediately.

____ (3) (2%) To solve the cache coherence problem, Snoopy protocol is the simplest. Regarding the Snoopy coherence protocol, what is the following description true?
(a) Any CPU should invalidate the cache block if it receives a write miss signal to the data in the local cache block from the bus.
(b) Any CPU should invalidate the cache block if it receives a read miss signal to the data in the local cache block from the bus.
(c) The Snoopy coherence protocol is an atomic operation.
(d) The Snoopy coherence protocol is proper to apply to a single-core system.

____ (4) (2%) Regarding the cache miss, what is the following description true?
(a) Compulsory miss indicates that the cache cannot contain all blocks required for the execution.
(b) Coherence miss only happens in multi-core systems.
(c) The miss rate will go down while the block size is made very large.
(d) Increasing the associativity decreases miss rate due to lower compulsory miss.

____ (5) (2%) Regarding the cache design, what is the following description true?
(a) With the criteria of the identical entries, we need more tag bits in a multi-word cache than in set-accusative cache.
(b) Through interchanging the loops in a code, the cache miss rate cannot be degraded.
(c) Through the way prediction technique, the miss penalty during the cache access can be reduced.
(d) If we pipeline the cache access, the cache bandwidth can be improved.

Problem 2: (30 points)

Suppose we have a processor with a base CPI of 1.0, assuming all reference hit in the primary cache, and a clock rate of 5GHz. Assume a main memory access time of 100 ns, including all the miss handling. Suppose the miss rate per instruction at the primary cache is 2%.

A. (10%) What is the CPI if this single-core processor only has one level of cache?

B. (20%) To speed up the process, we increase the CPU cores from 1-core CPU to 2-core CPU and retain other design settings. We assume 60% of instructions must be executed sequentially. Please estimate the speedup ratio by using the new architecture.

Problem 3: (20 points)

Assume a GPU architecture that contains 10 SIMD processors. Each SIMD instruction has a width of 32 and each SIMD processor contains 8 lanes for single-precision arithmetic and load/store instructions, meaning that each non-diverged SIMD instruction can produce 32 results every 4 cycles. Assume a kernel that has divergent branches that cause on average 80% of threads to be active. Assume that 70% of all SIMD instructions executed are single-precision arithmetic and 20% are load/store. Since not all memory latencies are covered, assume an average SIMD instruction issue rate of 0.85. Assume that the GPU has a clock speed of 1.5 GHz. Please compute the throughput, in GFLOP/sec, for this kernel on this GPU.

Problem 4: (40 points)

You are a system engineer. Today, you need to design a memory hierarchical system, including one CPU, one or two-level cache, one main memory, and one hard disk. Currently, you have the following design policies for the cache-level design.

Policy 1: With only L1 cache by using 2-way associativity

Policy 2: With direct mapping L1 cache and 2-way associative L2 cache

Policy 3: With 2-way associative L1 and L2 caches

Assume that there is one embedded TLB in each cache level and one embedded page table in the main memory (*i.e.*, we do not need extra memory to store the TLB and page table). On the other hand, the specifications of this memory hierarchical system are

- This is a 32-bit machine.
- The base CPI is 1.0 and the clock rate is 5GHz.
- Each cache block is a single-word block.
- The L1 cache can contain 8KB data.
- The L2 cache can contain 16KB data.
- The page size is 2^{12} bytes.
- In the TLB and page table, we need to involve an extra one dirty bit to implement the write-back policy; one reference bit to approximate the LRU replacement policy; one valid bit to judge the data hit/miss.
- The number of the TLB entries in L1 and L2 caches are 10 and 20 respectively. Besides, the number of entries in the page table is 30. To reduce the miss rate, the fully associative policy is adopted to implement the TLB and page table.

A. (10%) Please determine the number of bits required in the page table, TLB in the L1 cache, and TLB in the L2 cache.

B. (10%) Please determine the number of bits required if
   a. (5%) L1 cache is implemented by using 2-way associative mapping strategy.
   b. (5%) L2 cache is implemented by using 2-way associative mapping strategy.

C. (20%) Without considering the data transferring time, we assume the access time of L2 cache is 5 ns including all the miss handling; the access time of the main memory is 100 ns including all the miss handling; the access time of the hard disk is 1 us including all the miss handling. According to the data transference time between each memory level, we ignore the data transference time between the L1 and L2 cache and the data transference time between the main memory and the lowest level cache and disk are both 50 ns including all the miss handling. In this system, the TLB will be located at the lowest level cache. When a data request comes, the TLB must be accessed first. If the TLB miss happens, we need to spend 10ns to handle the TLB-miss exception. When we adopt direct mapping strategy, the miss rate of the L1 cache and the embedded TLB are both 2%; the miss rate of the L2 cache and the embedded TLB are both 0.5%. If the 2-way mapping strategy is applied, the

miss rate of the L1 cache and the embedded TLB are both 1%; the miss rate of the L2 cache and the embedded TLB are both 0.1%. At last, the miss rate of the main memory is 0.1%. During manufacturing, we need to spend 0.01 USD to handle one bit in each kind of memory. Please provide a design suggestion, including how many cache level you suggest and what kind of mapping strategy for each cache level you suggest, to your customer by considering the system performance and the manufacturing simultaneously cost.