SCHOOL OF COMPUTER SCIENCES

UNIVERSITI SAINS MALAYSIA

CST433 – Advanced Computer Organization & Architecture

Semester II, 2021/2022 Academic Session

30 June 2022 (Thursday), 12:05 pm – 1:20 pm

**TEST 2**

**Instructions: Answer all questions in the papers provided. Do not forget to write your name on every sheet of your answer paper.**

1. With **two (2)** appropriate examples for each technique, analyse how the following parallelism techniques have influenced the development of modern computer architecture.

(a) Instruction-Level Parallelism. [10/100]

1. Hardware-based speculation: We fetch, issue and execute instructions as if our branch prediction is always correct. Combines three ideas: dynamic branch prediction, speculation, and dynamic scheduling. It is useful for systems if they want to continue doing useful tasks even when the others are not completed.
2. Multiple Issue: Two types of processors, statically scheduled and dynamically scheduled superscalar processors.

(b) Data-Level Parallelism. [10/100]

1. Vector Architectures: It increases performance of simple scalar processors without highly raising the energy demands and design complexity. Programs can run well even in out-of-order efficiently.
2. GPU: It has every type pf parallelism that can found in programming environment (Multithreading, MIMD, SIMD). It handles parallel execution and thread management. GPU is useful for parallel data optimization.

(c) Thread-Level Parallelism. [10/100]

1. Centralized Shared-Memory: Support caching of both shared and private data. Multiprocessor speeds up data access and execution.
2. Distributed Shared-Memory: Processes in this memory is done without inter-process communication. Implement the shared memory model. It speeds up access to data.

2. Vector Architecture, SIMD Multimedia Extension processors and Graphics Processing Units (GPUs) have utilized data-level parallelism in their design and implementation.

1. Evaluate **3 (three)** features/concepts of Vector Architecture that are identical to SIMD Multimedia Extension processors. [15/100]

* Both specify the same operation on vectors of data.
* Vector architecture is a classification of SIMD.
* Both operate on data-level parallel programs.

(b) Critique **3 (three)** features/concepts of Vector Architecture that are identical to Graphics Processing Units (GPUs). [15/100]

* Both only work well with data-level parallel problems even though they take different path, not in instruction-level or thread-level.
* Both use mask registers in conditional branch instructions, which makes unmasked interrupt a problem.
* Both use large register files which act as compiler-controlled buffers.

3. In multiprocessors environment, enforcing cache coherence can be addressed through directory based and snooping protocol.

(a) Evaluate the common implementation of the snooping protocol in a centralized shared-memory architecture. [10/100]

There are two common implementations of the snooping protocols in a centralized shared-memory architecture:

1. Write invalidate protocol

* This protocol invalidates other copies on a write
* Enforces write serialization
* It is also the most common protocol

1. Write update/Write broadcast protocol

* This protocol updates all cached copies of the data items when that item is written
* It also consumes more bandwidth

(b) Evaluate the common implementation of the directory-based protocol in distributed shared-memory architecture. [10/100]

Directory-based protocol in distributed shared-memory architecture involves the coherence protocol, where it must know where to find the directory information for any cached block of memory. The solution is to distribute the directory along with the memory so different coherence requests can go to different directories. There are two primary operations which are handling a read miss and handling a write to a shared, clean cache block.

(c) Discuss **two (2)** synchronization mechanisms that can be used to enforce cache coherence in multiprocessor environment. [10/100]

Synchronization mechanisms that can be used to enforce cache coherence in multiprocessor environment:

1. Atomic exchange primitive. It interchanges the value in a register for a value in the memory. A simple lock is built; 0 for indicating the lock is free while 1 for indicating the lock is unavailable. The key to using it is the operation must be atomic.
2. Locks. Spin locks are implemented once there is an atomic operation. They are used when programmers expect the lock to be held for a very short time. The advantages of caching the lock is the spinning lock process is done on local cache instead of global memory access.

(d) Discuss **two (2)** memory consistency models that can be used enforce cache coherence in multiprocessor environment. [10/100]

1. Sequential consistency. It requires a processor to delay the completion of any memory access until all invalidations caused by that access are completed. Cannot simply place the write in a write buffer and continue with the read.
2. Release consistency. It distinguishes between synchronization operations that are used to acquire access to a shared variable and those that release an object to allow another processor to acquire access. Slightly relax the ordering.