CMPT 334 **Spring 2021**

Assignment 2

Due: Tuesday March 02, 2021 – 8:00 AM

Submit a single pdf/word file in Moodle.

Provide short answers for the following question (100 points):

Q1. Describe how the communications between computer components are handled.

The most common communication structures are (1) the bus and various multiple-bus structures, and (2) point-to-point interconnection structures with packetized data transfer.

A system bus consists data lines. These data lines provide a path for moving data among system modules. These lines, collectively, are called the data bus. A key characteristic of a bus is that it is a shared transmission medium.

A point-to-point intercommunication, such as QPI, is a communication structure where multiple components within the system enjoy direct pairwise connections to other components. This eliminates the need for arbitration found in shared transmission systems.

Q2. What are two different approaches of dealing with interrupts?

(1) Disable all interrupts while an interrupt is being processed.

(2) Define priorities for interrupts and to allow an interrupt of higher priority to cause a lower-priority interrupt handler to be interrupted.

Q3. What is the advantage of using QPI inter-communication as opposed to the system bus inter-communication?

QPI eliminates the need for arbitration found in shared transmission systems since it is point-to-point. There is no single bus that all the processors must share and contend with each other. It also improves scalability.

Q4. Consider a 32-bit processor having 32-bit instructions. The instructions are composed of: (1) first byte representing the opcode and (2) the remainder representing the memory address of an operand.

- **What is the maximum addressable memory capacity in bytes?**
  Since the opcode is one byte (8-bits), the remainder of the instruction is 24-bits. So, the maximum addressable memory is 2^{24}

- **How many bits are needed for the program counter (PC) register?**
  The program counter must be at *least* 24 bits since it contains the address of the next instruction to be executed. It 32-bit processors it also can be 32-bit program counter.

- **How many bits are needed for the instruction register?**
  If the instruction register is to contain the whole instruction, it will have to be 32-bits long; if it will contain only the op code (called the op code register) then it will have to be 8 bits long.

Q5. Describe the different between sequential access, direct access and random access.

Sequential access: Memory is organized into units of data, called records. Access must be made in a specific linear sequence.

Direct access: Individual blocks or records have a unique address based on physical location. Access is accomplished by direct access to reach a general vicinity plus sequential searching, counting, or waiting to reach the final location.

Random access: Each addressable location in memory has a unique, physically wired-in addressing mechanism. The time to access a given location is independent of the sequence of prior accesses and is constant.

Q6. Why is the "principle of locality" critical to a hierarchical (multi-layer) memory architecture?

It is possible to organize data across a memory hierarchy such that the percentage of accesses to each successively lower level is substantially less than that of the level above. Because memory references tend to cluster, the data in the higher-level memory need not change very often to satisfy memory access requests.

Q7. What are the differences between spatial and temporal locality?

Spatial locality refers to the tendency of execution to involve a number of memory locations that are clustered. Temporal locality refers to the tendency for a processor to access memory locations that have been used recently.

Q8. Consider the following code:

```
for (int j = 0 ; j < 20; j++){
        for (int i = 0; i < 10; i++){
                arr[i] = arr[i] * arr[i];
        }
}
```

- **What is an example of spatial locality in the code?**
  A reference to the first instruction is immediately followed by a reference to the second.

- **What is an example of temporal locality in the code?**
  The ten accesses to “arr[i]” within the inner for loop which occur within a short interval of time.

Q9. Assume you have a computer with a main memory, one cache memory and a main single processor (not multi-core). A miss penalty is 1 clock cycle to send an address to the main memory and a total of 4 cycles to transfer a 32-bit word from the main memory to the processor and cache. What is the miss penalty of a cache block, if the cache block size is 4 words?

miss penalty = 4 × (1 + 4) = 20 clock cycles

Q10. Consider the following programs:

Program A:	**Program B:**

for (int i = 1; i < n; i++){
z[i] = a[i] – b[i];
z[i] = z[i] * z[i];
}
	for (int i = 1; i < n; i++){
	z[i] = a[i] – b[i];
	}
	for (int i = 1; i < n; i++){
	z[i] = z[i] * z[i];
	}

The given two programs do the same function. Which program performs better? Explain why?

Program A performs better because it has better data locality. This is because program A access array “z” sequentially one cell at a time, while program B access array “z” twice. If array “z” is large (doesn’t fit in a single block), program B may need to transfer blocks containing array “z” more times than program A.