National Institute of Technology, Tiruchirappalli

Department of Computer Science and Engineering

End Semester
CSPC31 – Computer Architecture

Branch/Semester/ Section : CSE/ V/ A

Time : 10:00AM to 01:00 PM

Date : 02.12.2021 Max Marks: 60

Answer All Questions

1.

a) Let there are two operations to be performed: one is a product of 3 scalar variables, and one is a matrix sum of a pair of two-dimensional arrays, with dimensions 9 by 9. For now let's assume only the matrix sum is parallelizable;
   i. What speed-up do you get with 10 versus 40 processors? **3**
   ii. Calculate the speed-ups assuming the matrices grow to 20 by 20. **3**

b) The power consumption of several computer system components is presented in the figure below. **4**

Component type	Product	Performance	Power
Processor	Sun Niagara 8-core	1.2 GHz	72–79 W peak
	Intel Pentium 4	2 GHz	48.9–66 W
DRAM	Kingston X64C3AD2 1 GB	184-pin	3.7 W
	Kingston D2N3 1 GB	240-pin	2.3 W
Hard drive	DiamondMax 16	5400 rpm	7.0 W read/seek, 2.9 W idle
	DiamondMax 9	7200 rpm	7.9 W read/seek, 4.0 W idle

Assuming the maximum load for each component, and a power supply efficiency of 80%, what wattage must the server's power supply deliver to a system with an Intel Pentium 4 chip, 2 GB 240-pin Kingston DRAM, and one 7200 rpm hard drive? How much power will the 7200 rpm disk drive consume if it is idle roughly 60% of the time?

2.

a) Consider a 32-bit microprocessor that has an on-chip 16-kB four-way set-associative cache. Assume that the cache has a line size of four 32-bit words. Draw a block diagram of this cache showing its organization and how the different address fields are used to determine a cache hit/miss. Where in the cache is the word from memory location ABCDE8F8 mapped? **6**

b) Consider a cache with a line size of 32 bytes and a main memory that requires 30 ns to transfer a 4-byte word. For any line that is written at least once before being swapped out of the cache, what is the average number of times that the

line must be written before being swapped out for a write-back cache to be more efficient that a write-through cache? How does the answer change if the main memory uses a block transfer capability that has a first-word access time of 30 ns and an access time of 5 ns for each word thereafter?

4

3.

a. Identify the RAW, WAR, and WAW dependencies in the following instruction sequence:

I1: R1 = 100
I2: R1 = R2 + R4
I3: R2 = r4 – 25
I4: R4 = R1 + R3
I5: R1 = R1 + 30

4

b. (i) what are the limitations of the basic 3 step tomasulo's algorithm

3

(ii) Using an example explain how the limitations are overcome by hardware based speculation using commit step

3

4. Suppose the branch frequencies (as percentages of all instructions) are as follows:

Conditional branches: 10%

Jumps and calls: 2%

Taken conditional branches 55% are taken

a. We are examining a five-deep pipeline where the branch is resolved at the end of the third cycle for unconditional branches and at the end of the fourth cycle for conditional branches. Assuming that only the first pipe stage can always be done independent of whether the branch goes and ignoring other pipeline stalls, how much faster would the machine be without any branch hazards?

5

b. Now assume a high-performance processor in which we have a 15-deep pipeline where the branch is resolved at the end of the fifth cycle for unconditional branches and at the end of the tenth cycle for conditional branches. Assuming that only the first pipe stage can always be done independent of whether the branch goes and ignoring other pipeline stalls, how much faster would the machine be without any branch hazards? **5**

5.

a) Assume a hypothetical GPU with the following characteristics: **2+1**

■ Clock rate 1.5 GHz
■ Contains 16 SIMD processors, each containing 16 single-precision floating-point
units
■ Has 100 GB/sec off-chip memory bandwidth
Without considering memory bandwidth, what is the peak single-precision floating-point throughput for this GPU in GLFOP/sec, assuming that all memory latencies can be hidden? Is this throughput sustainable given the memory bandwidth limitation?

b) Consider the following code, which multiplies two vectors that contain single-precision complex values:

```
for (i=0;i<100;i++) {
c_re[i] = a_re[i] * b_im[i] – a_im[i] * b_re[i];
c_im[i] = a_re[i] * b_re[i] + a_im[i] * b_im[i];
}
```

Assume that the processor runs at 700 MHz and has a maximum vector length of
64. The load/store unit has a start-up overhead of 10 cycles; the multiply unit, 6 cycles; and the add/subtract unit, 4 cycles.

i. Convert this loop into VMIPS assembly code. **4**

ii. Assuming chaining and a single memory pipeline, how many chimes are required? How many clock cycles are required per complex result value, including start-up overhead? **2+1**

6.

a) Examine the effect of the interconnection network topology on the clock cycles per instruction (CPI) of programs running on a 64-processor distributed-memory multiprocessor. The processor clock rate is 3.3 GHz and the base CPI of an application with all references hitting in the cache is 0.5. Assume that 0.2% of the instructions involve a remote communication reference. The cost of a remote communication reference is (100 + 10h) ns, where h is the number of communication network hops that a remote reference has to make to the remote processor memory and back. Assume that all communication links are bidirectional. **4**

i. Calculate the worst-case remote communication cost when the 64 processors are arranged as a ring, as an 8×8 processor grid, or as a hypercube. (Hint: The longest communication path on a 2n hypercube has n links.)

ii. Compare the base CPI of the application with no remote communication to the CPI achieved with each of the three topologies in part (a).

b) (i) For which kind of applications the distributed cache coherence protocol is suited. Justify the same. **3**

(ii) Using a state diagram elaborate on the functioning of the distributed cache coherence protocol **3**

****