NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER 2 EXAMINATION 2021-2022

CE3001/CZ3001 – ADVANCED COMPUTER ARCHITECTURE

Apr/May 2022 Time Allowed: 2 hours

INSTRUCTIONS

1. This paper contains 4 questions and comprises 5 pages.
2. Answer **ALL** questions.
3. This is a closed-book examination.
4. All questions carry equal marks.

---

1. (a) If fraction T of a program is enhanced by a factor of F when executed on an enhanced machine, derive the expression of speedup over the original machine. Based on that expression, find the value of speedup with T=0.4, for F=1000.

(4 marks)

(b) State the addressing mode for conditional branching. Determine the minimum and maximum possible values to which an instruction "CBZ X31, address" can branch for a LEGv8 architecture, given that the content of the 64-bit program counter value of the branch instruction is 0xF000 0000 000C.

(9 marks)

(c) Use a neat diagram to show the datapath of a single-cycle architecture that supports the execution of both "STUR X0, [X1, #8]" and "ADDI X4, X3, #5". Note that the datapath needs to have only minimal number of multiplexers (control signals can be simplified). Also briefly explain the working of the instructions "STUR X0, [X1, #8]" and "ADDI X4, X3, #5".

(12 marks)

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

2. Listing Q2 shows a code segment that is intended to be executed in a 5-stage pipelined LEGv8 processor. The program counter is updated with the branch target address at the Decode stage. Note that, no data forwarding is allowed but write-back and register-read operations of different instructions can be performed in the same clock cycle.
Let the initial values be `X4= 0x0000000000000000`, `X7=0x0000000010000000` and `X8=0x0000000000001000`.

Listing Q2

```
I1   loop:  LDUR  X1,  [X7,#0]
I2          LDUR  X2,  [X8,#0]
I3          ADD   X3,  X2,  X1
I4          ADD   X4,  X4,  X3
I5          SUBI  X7,  X7,  #8
I6          SUBI  X8,  X8,  #8
I7          CBNZ  X8,  loop

    Finish
```

(a) Spot all the flow dependencies in Listing Q2. Also find the total number of loop iterations.

(4 marks)

(b) The code segment shown in Listing Q2 is now intended to be executed in a two-way superscalar processor. In the superscalar processor, one way is exclusively for load and store instructions whereas the other way can execute all instructions except load and store. Find the CPI achieved for the code segment shown in Listing Q2 using superscalar architecture.

(7 marks)

(c) Perform loop unrolling by a factor of 2 for the code segment in Listing Q2 and do the necessary reordering of instructions to reduce the number of stall cycles to the minimum. You are allowed to use new temporary registers to get rid of hazards.

(6 marks)

(d) The unrolled and reordered code segment derived in Q2(c) is now intended to be executed in a two-way superscalar processor. As mentioned earlier, one way is exclusively for load and store instructions whereas the other way can execute all instructions except load and store.

Note: Question No. 2 continues on Page 3

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Find the CPI achieved for the unrolled and reordered code segment derived in Q2(c) using superscalar architecture. Comment on the performance enhancement.

(8 marks)

3. (a) Name two basic cache write policies.

(4 marks)

(b) Consider a hierarchical memory system composed of an 8 GB Byte-addressable main memory, and a 16 MB eight-way set-associative cache with block size of 256 Bytes. Determine the lengths (in number of bits) of the tag, index, and offset fields in the address, respectively. (1 GB = 1024 MB; 1 MB = 1024 KB; 1 KB = 1024 Bytes)

(8 marks)

(c) Following Q3(b), what is the miss rate of the cache when a program accesses the main memory in a totally random fashion, that is, random memory addresses are uniformly accessed for a sufficiently long time?

(6 marks)

(d) Tables Q3a and Q3b show the results of cache miss rate and average memory access time when a system is configured with different cache block sizes. Assume a cache hit takes 1 clock cycle. The cache miss penalties with different block sizes are provided in Table Q3b.

Table Q3a: Cache Miss Rate v.s. Block Size

Block Size (Bytes)	Cache Size			
	4KB	16KB	64KB	256KB
16	10.95%	X%	1.95%	0.79%
64	6.82%	2.39%	1.02%	0.47%
256	8.36%	2.93%	0.86%	0.32%

Note: Question No. 3 continues on Page 4

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table Q3b: Average Memory Access Time v.s. Block Size

Block Size (Bytes)	Miss Penalty	Cache Size			
		4KB	16KB	64KB	256KB
16	82	9.98	3.96	2.60	1.65
64	88	7.00	3.10	1.90	Z
256	112	Y	4.28	1.96	1.36

(i) Compute the missing entries X, Y and Z in Tables Q3a and Q3b.

(5 marks)

(ii) Between cache miss rate and average memory access time, which one is the more important performance metric to consider when designing a cache with maximized performance?

(2 marks)

4. (a) List the four types of processor architectures in the processor taxonomy according to Flynn's classification. Which of them are designed to take data-level parallelism into account?

(6 marks)

(b) Briefly discuss the main differences between the typical architecture designs of a CPU and a GPU (in no more than 6 sentences).

(6 marks)

(c) Briefly explain the usage of the specifiers `__global__` and `__shared__` in CUDA C programming (in no more than 4 sentences).

(4 marks)

(d) The code snippet in Figure Q4 shows a C program that uses a CUDA kernel `saxpy()` to compute the SAXPY (Single precision A.X plus Y) operation:

`Y = AX+Y`

where `A` is a scalar, while **`X`** and **`Y`** are vectors each consisting of `N` floating-point numbers.

Note: Question No. 4 continues on Page 5

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
Line
1   __global__
2   void saxpy(int n, float a, float *x, float *y){
3     int i = blockIdx.x * blockDim.x + threadIdx.x;
4     if (i < n)
5        y[i] = a*x[i] + y[i];
6     }
7
8   int main(void){
9     int N = 8191;     // size of vectors X and Y
10    float A = 2.0;
11    X = (float *)malloc(N*sizeof(float));
12    Y = (float *)malloc(N*sizeof(float));
13    // get values of vectors X and Y
:          :
:          :
n     saxpy<<<.............>>>(N, A, d_X, d_Y);
:          :
:     return 0;
n+k   }
```

Figure Q4

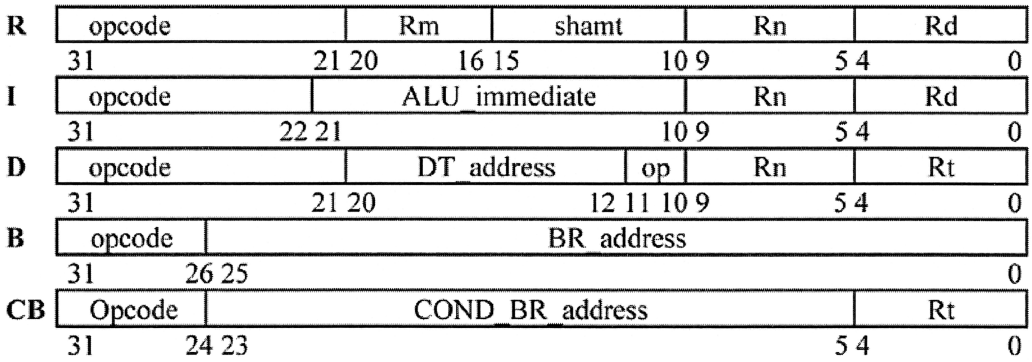
(i) Complete the code shown in Line ***n*** if the number of threads per block is set as 512.

(4 marks)

(ii) Assume a Stream Multiprocessor (SM) in a GPU has sufficient register and shared memory resources to reside all the blocks. What is the total number of warps that will be created by launching the kernel?

(5 marks)

Appendix - Instruction Formats



END OF PAPER

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

CE3001 ADVANCED COMPUTER ARCHITECTURE
CZ3001 ADVANCED COMPUTER ARCHITECTURE

Please read the following instructions carefully:

1. **Please do not turn over the question paper until you are told to do so. Disciplinary action may be taken against you if you do so.**

2. You are not allowed to leave the examination hall unless accompanied by an invigilator. You may raise your hand if you need to communicate with the invigilator.

3. Please write your Matriculation Number on the front of the answer book.

4. Please indicate clearly in the answer book (at the appropriate place) if you are continuing the answer to a question elsewhere in the book.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library