NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER 2 EXAMINATION 2022-2023

CE3001/CZ3001/SC3050 – ADVANCED COMPUTER ARCHITECTURE

Apr/May 2023 Time Allowed: 2 hours

INSTRUCTIONS

1. This paper contains 4 questions and comprises 7 pages.

2. Answer **ALL** questions.

3. This is a closed-book examination.

4. All questions carry equal marks.

5. Information on instruction formats is provided in the Appendix on Page 7.

---

1. (a) Two systems S1 and S2 have clock frequencies of 300 MHz and 200 MHz, respectively. A given program P consists of $2x10^9$ instructions, which are distributed among two instruction types T1 and T2, each contributing $1x10^9$ instructions. Cycles-Per-Instruction (CPI) of T1, while executing on S1 and S2, are 2 and 3, respectively. CPI of T2, while executing on S1 and S2, are 4 and 3, respectively.

(i) Determine the execution time of P on S1 and S2.

(4 marks)

(ii) Comment which system is slower after finding the speed-up of S1 over S2.

(3 marks)

(b) Briefly explain the working of the instructions "STUR X0, [X1, #8]" and "CBZ X2, #8". Figure Q1 shows the LEGv8 architecture, where one line is marked with "X". "X" indicates there is a hardware failure due to a broken bus. Indicate which of the instructions above will have issue in executing. Explain in detail the reason by indicating the path(s) taken by the instruction(s) with issue.

(10 marks)

Note: Question No. 1 continues on Page 2

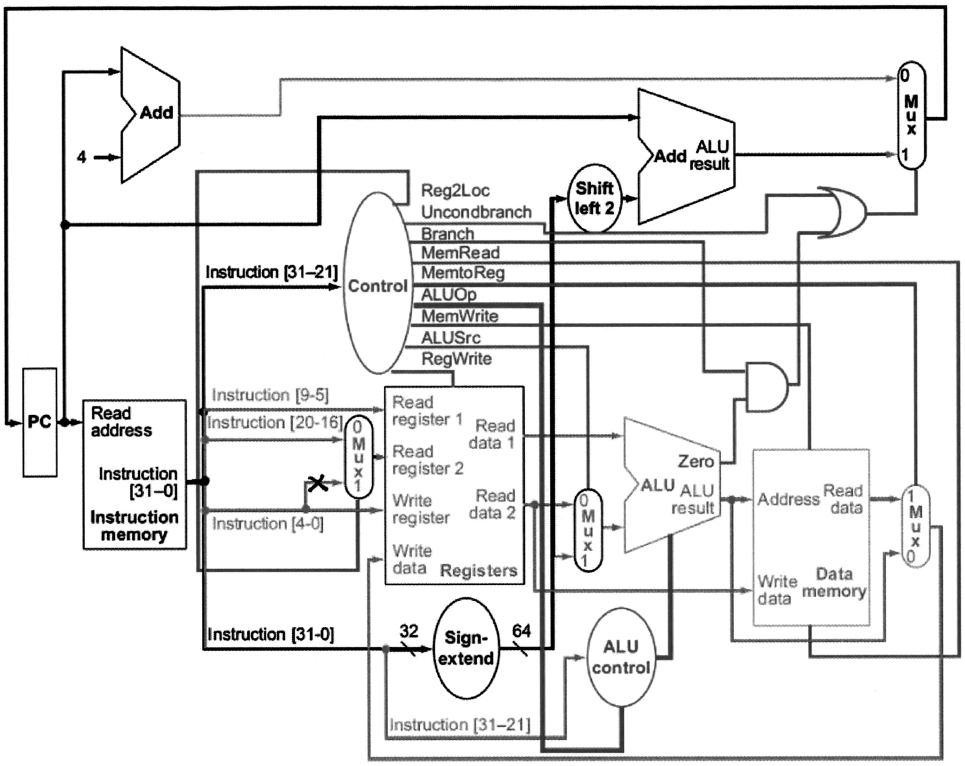


Figure Q1

(c) With the example given in Table Q1, clearly explain the operation of instruction with reference to its addressing mode. You need to clearly emphasize the minimum and maximum range of branching (by suggesting a value for offset).

Table Q1

PC value	Instruction
0X10F0FF0C	B offset

(8 marks)

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. 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 value of `X8` be `0x0000000000001000` in hexadecimal (`CBZ`: *branch if equal to 0*).

Note: Question No. 2 continues on Page 3

Listing Q2

```
I1   loop:  LDUR  X1, [X8,#0]
I2          ANDI  X2, X1, 0xFFFF
I3          STUR  X2, [X8,#0]
I4          SUBI  X8, X8, 0x0010
I5          CBZ   X8, loop
I6          B     Finish

     Finish
```

(a) Calculate the steady-state CPI of the code segment in Listing Q2 with the help of a reservation table. Show the forwarded paths and the dependencies. Also find the total number of loop iterations.

(7 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, both ways can be used for all the instructions, and instruction reordering is allowed. Find the CPI achieved by the superscalar architecture.

(6 marks)

(c) Perform unrolling by a factor of 4 and do the necessary reordering. Use the unrolled and reordered code segment to be executed in a two-way superscalar machine. Find the CPI achieved by the superscalar architecture. Comment on the change in CPI when compared to Q2(b).

(10 marks)

(d) Briefly comment on the impact of applying static branch prediction of "Always Not Taken" in Q2(a).

(2 marks)

3. (a) Name two different write policies for cache systems. Which policy has better performance considering the huge difference in speed between the cache and the main memory?

(5 marks)

(b) Consider a memory system with Byte-addressable main memory and 32-bit physical addresses. The cache system configuration is as follows:

Note: Question No. 3 continues on Page 4

- L1 Instruction Cache: 64 KB (1 KB = 1024 Bytes) in size, 128-Byte blocks, 4-way set associative cache, indexed and tagged with physical addresses.
- L1 Data Cache: 32 KB in size, 64-Byte blocks, direct mapped cache, indexed and tagged with physical addresses.
- Both caches keep the following information bits for each cache line: a dirty bit, a reference bit, and 3 permission bits.

Specify the number of offset, index, and tag bits for each of these structures, and compute the total size in number of bits for each of the tag arrays (including both information bits and tag bits) in Table Q3a.

Table Q3a

Structure	Tag Bits	Index Bits	Offset Bits	Size of Tag Array in Bits
L1 Instruction Cache				
L1 Data Cache				

(12 marks)

(c) Tables Q3b and Q3c show the results of cache miss rates and average memory access times when a system is configured with different cache sizes and block sizes. Assume a cache hit takes 2 clock cycles. The cache miss penalties with different block sizes are provided in Table Q3c.

Table Q3b: Cache Miss Rate v.s. Block Size

Block Size (Bytes)	Cache Size			
	4KB	16KB	64KB	256KB
16	9.37%	5.59%	2.86%	1.36%
64	X%	3.87%	1.43%	0.77%
256	7.52%	4.23%	1.58%	0.62%

Note: Question No. 3 continues on Page 5

Table Q3c: Average Memory Access Time v.s. Block Size

Block Size (Bytes)	Miss Penalty	Cache Size			
		4KB	16KB	64KB	256KB
16	42	5.94	4.35	3.20	Y
64	48	4.95	3.86	2.69	2.37
256	72	7.41	Z	3.14	2.45

(i) Compute the missing entries X, Y and Z in Tables Q3b and Q3c.

(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 system? Briefly justify your answer.

(3 marks)

4. (a) Assume a branch predictor uses a two-bit prediction scheme as shown in Figure Q4a. The predictor is initialized to the state of "00". Consider a branch instruction with the following repeating sequence of actual outcome: 'T T T N T T", where T indicates the branch is taken and N indicates the branch is not taken.

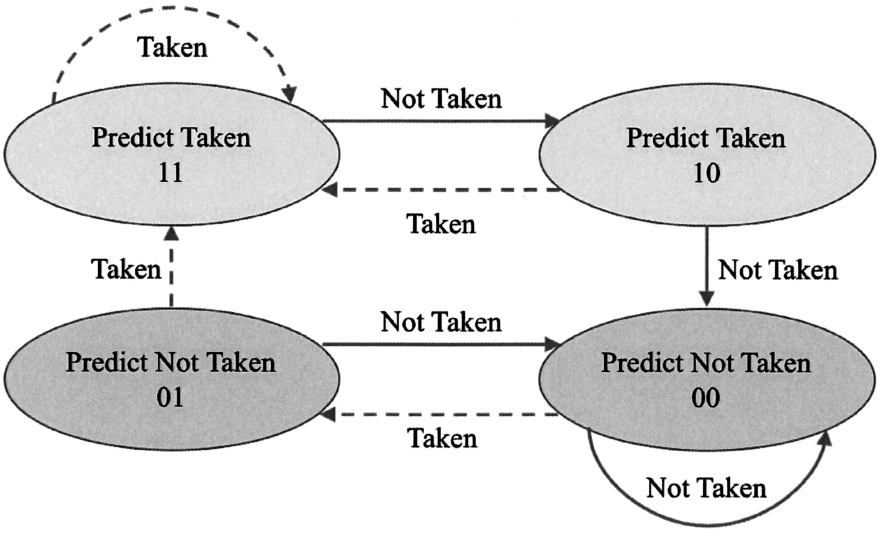


Figure Q4a

Note: Question No. 4 continues on Page 6

(i) Find the prediction accuracy of the two-bit predictor if the repeating sequence of actual outcome is repeated infinitely. Complete Table Q4 below to indicate the prediction decision in each step. The first column has been filled as a reference. What is the prediction accuracy?

Table Q4

Repeating Sequence	1						2					
Predictor State	00											
Prediction Decision	N											
Actual Outcome	T	T	T	N	T	T	T	T	T	N	T	T
Correct Prediction? (Yes/No)	N											

(7 marks)

(ii) Compare the prediction accuracy with a static "Always Taken" predictor. Briefly comment on the better choice of predictor in this case using no more than 2 sentences.

(3 marks)

(b) Briefly analyze how GPUs are designed to hide memory access latencies in no more than 4 sentences.

(4 marks)

(c) Figure Q4b shows a CUDA kernel that runs on a GPU to compute the dot product of two vectors **A** and **B** and outputs a scalar value C:

$$C = \mathbf{A} \bullet \mathbf{B} = \sum_{i=1}^{N} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_N b_N$$

Note: Question No. 4 continues on Page 7

```
Line
1   __global__ void
    dot_prod(int N, int *a, int *b, int *c) {
2     __shared__ int temp[N];
3     int i = blockIdx.x;
4     temp[i] = a[i]*b[i];
5
6     // Thread 0 sums the pairwise products
7     if (i == 0) {
8       int sum = 0;
9       for (int j = 0; j < N; j++)
10        sum += temp[j];
11      *c = sum;
12    }
```

Figure Q4b

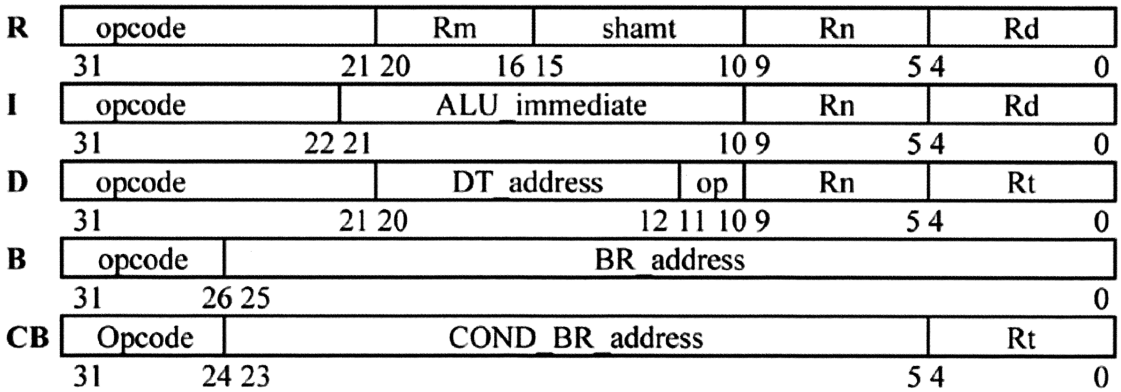
(i) Identify two mistakes in the CUDA C code in Figure Q4b that are related to GPU programming only (ignore other general programming bugs, if any). Indicate the correct CUDA C code to fix the mistakes and make the kernel function correctly.

(6 marks)

(ii) If N = 32, indicate the CUDA C code to launch the kernel with the number of block(s) and thread(s) to be created. Briefly justify how the number of blocks and threads are determined in no more than 4 sentences.

(5 marks)

Appendix - Instruction Formats



END OF PAPER

CE3001 ADVANCED COMPUTER ARCHITECTURE
CZ3001 ADVANCED COMPUTER ARCHITECTURE
SC3050 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.