National University of Computer and Emerging Sciences, Lahore Campus

Course Name:	Advance Computer Architecture	Course Code:	EE502
Program:	MS (Computer Science)	Semester:	Fall2018
Duration:	3 hours	Total Marks:	70
Paper Date:	26-12-2018	Weight	50
Exam Type:	Final	Page(s):	8

Student : Name:_____ **Roll No.**_____ **Section:** _____

Instruction:
1. Attempt all question in the provided space. You can use rough sheets but it should not be attached.
2. If you think some information is missing, write assumption and solve accordingly.

Question 1a [10+4+6] (Multiple options can be selected in each question)

1. Which of these modifications would you introduce on a cache to decrease the conflict misses?
   A. Increase the cache size
   B. Decrease the cache associativity
   C. Increase the size of the blocks
   D. Add a second level of cache
2. Which of the following statements regarding pipeline hazards arae TRUE?
   A. All data hazards can be avoided by forwarding
   B. An instruction could be dependent on previous store instruction
   C. Data hazards may cause error in execution result
   D. Structural hazards occur whenever an instruction depends on another for operand
3. Which complications are introduced by out-of-order execution and out-of-order completion?
   A. WAR and WAW hazards
   B. Structural hazards
   C. Exception handling
   D. Control hazards
4. Typically during pipelined execution, the read operation on the register file and the memory is performed during _____
   A. First half of the cycle
   B. Later half of the cycle
   C. Any half of the cycle
   D. Throughout the cycle
5. CPU time can be reduced by reducing:
   E. Instruction count
   F. Clocks per instruction
   G. Clock cycle time
   H. Clock frequency
6. Structural hazards can be avoided by
   A. adding stalls
   B. forwarding
   C. adding more hardware
   D. register renaming
7. Vector processors are a type of
   A. SISD processors
   B. SIMD processors
   C. MISD processors
   D. MIMD processors
8. We can flush an instruction by
   A. disabling its control signals
   B. disabling IF/ID register
   C. flushing IF/ID register
   D. All of the above

9. If branch related calculations and decision making is done in execution stage then how many nops instructions are required after the branch instruction to avoid wrong fetching of next instruction.
    A. 1
    B. 2
    C. 3
    D. 4

10. Allowing jumps branches, and ALU instructions to take fewer stages/cycles than the five required by the load instruction will
    A. increase pipeline performance
    B. decrease pipeline performance
    C. increase clock cycle time
    D. decrease clock cycle time

Question 1b:True/False

(Write true or false in front of each statement)

1. The global miss rate represents the fraction of memory accesses going from the processor to the main memory. **True**
2. The introduction of a larger L2-cache reduces both the effective miss penalty and the average memory access time. **True**
3. In MIMD Distributed architecture, the memory address space is shared. **False**
4. We may need to add stall before branch instruction due to control hazard. **False**

Question 1c

Fill in the following table by considering the given values

	Virtual Address bits	**Physical Address bits**	**Page Size**	**VPN bits**	**PPN bits**	**Minimum bits in each Page table entry**
a	32	26	**8kb**	**19**	13	**15**
b	**36**	32	**32kb**	21	**15**	17
c	**40**	**36**	32kb	25	**19**	21

VPN: Virtual page number

PPN: Physical page number

Question 2[10]

Define each of the following cache optimization technique and explain how it improves the performance.

a. Critical word first

- Request missed word from memory first
- Send it to the processor as soon as it arrives
- Effectiveness of this strategy depends on block size and likelihood of another access to the portion of the block that has not yet been fetched

b. Merging arrays

This technique improves spatial locality by single array of compound elements instead of accessing two arrays separately

```
int val[SIZE];
int key[SIZE];

for (i=0; i<SIZE; i++){
  key[i] = newkey;
  val[i]++;
}
```

```
struct record{
int val;
int key;
 };
struct record records[SIZE];

for (i=0; i<SIZE; i++){
records[i].key = newkey;
records[i].val++;
}
```

It reduces conflicts between val & key and improves spatial locality

c. Loop interchange

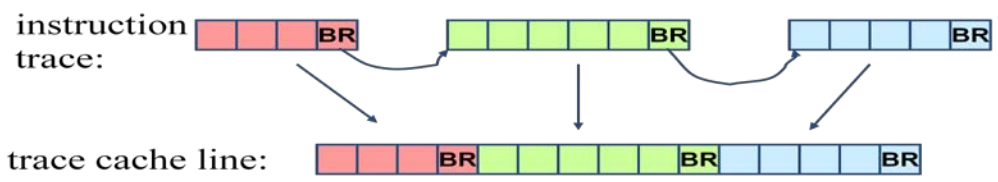
This implies that we do sequential accesses instead of striding through memory every 100 words which improves spatial locality.

```
for (col=0; col<100; col++)
      for (row=0; row<5000; row++)
        X[row][col] = X[row][col+1];

    for (row=0; row<5000; row++)
      for (col=0; col<100; col++)
        X[row][col] = X[row][col+1];
```

d. Trace caches

This technique packs multiple non-contiguous basic blocks into one contiguous trace cache line



- **Single fetch brings in multiple basic blocks**
- **Trace cache indexed by start address *and* next *n* branch predictions**

Question 3 [10]

Assuming a cache of 512 blocks, a 2-word block size, and a 20‑bit memory address. Calculate:

1. The total number of sets for a direct-map cache: ____512_____

2. The total number of sets for an8-way set associate cache: _____64_____

3. How many total bits of storage(tag+ data +valid)are required for an8-way set associate cache: __512*76_____

4. How many total bits of storage(tag+ data +valid)are required for a fully associative cache: _____512*82_____

5. In case of an 8-way set associate cache, Into what set would bytes with each of the following addresses be stored

	Set number (decimal)
▪ 0001 0001 0001 1011 1111	55
▪ 1100 1111 0011 0011 0100	38

Question 4 [10]

Let us consider a computer with a L1 cache and L2 cache memory hierarchy with the following parameters: Processor Clock Frequency = **1 GHz** ; Hit Time L1= 1 clock cycle ; Hit Rate L1= 90%;
Hit Time L2 = 15 clock cycles; Hit Rate L2= 80% ; Miss Penalty L2= 50 clock cycles;
Total Memory Accesses Per Instruction = 1.5 ; CPI exec = 2

1. How much is the AMAT?

$= 1.5(1 + 0.1(15 + 0.2(50)))$

$= 1.5(3.5)$ $\quad (15 + 10)$

$= 5.25$

2. If we want to improve the AMAT by 20 % by improving hit rate of L2 then what should be the miss rate of L2?

$4.2 = 1.5(1 + 0.1(15 + \times(50)$

miss rate = 6 % $\quad$ hit rate 94 %

3. How much is the impact on CPU time of the L2 cache with respect to an IDEAL L2 cache that never miss

Ideal cache

$= 1.5(1 + 0.1(15)$

$= 3.75 + 2 \quad = 5.75$

$= 5.25 + 2 \quad = 7.25 \quad = 1.26 \quad 26\ \%$

Let us assume to introduce an L3 cache with Hit time L3 = 25 clock cycles and Hit Rate L3 = 92%.

4. How much is Miss Penalty of L2?

$25 + 0.08(50)$

29

5. How much is the new AMAT?

$= 1.5(1 + 0.1(15 + 0.2(25 + 0.08(50))$

$25 + 4$

29

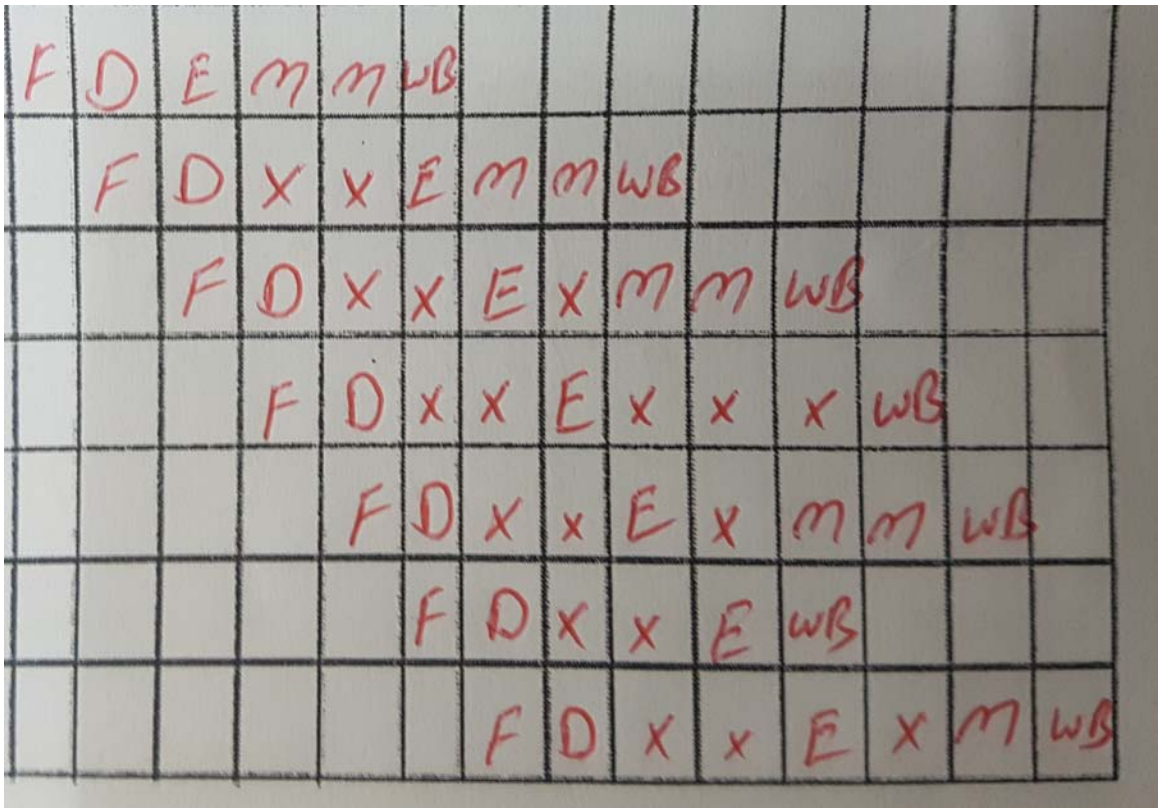
1.58

$(3.08)1.5$

Question 5 [6+6+8]

Code	Instructions
Loop: ld R2,100(R6) **st R2,200(R6)** **ld R2,300(R6)** **addi R2,R2,10** **st R2,300(R6)** **addi R6,R6,4** **beq R6, R7, Loop**	o ALU operations take 1 cycle in execution and no cycle in memory. Beq takes one cycle in each stage o Memory operations take 2 cycles in memory stage o There is no register renaming o Instructions are fetched in order however decode, execution, memory and writeback can be out of order. o An instruction will only enter the execution stage if it does not cause a RAW, WAR or WAW hazard o There is only one unit in each stage so only one instruction can enter a stage at a time, and in case multiple instructions are ready, the oldest one will go first o Forwarding is implemented

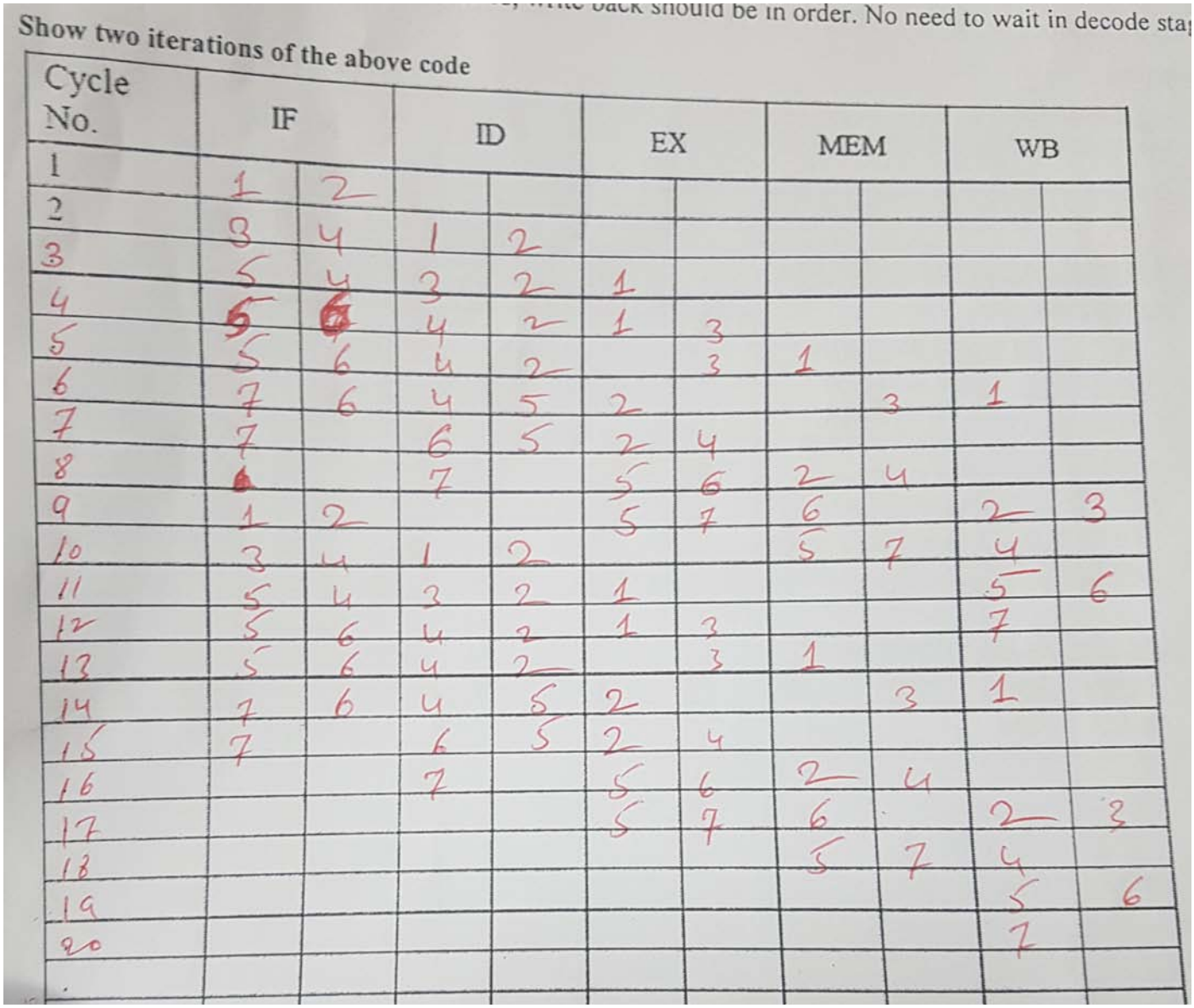
a. Fill in the following table pointing for each instruction, the pipeline stages activated in each clock cycle. For example instruction 1 is fetched in 1st cycle, decoded in 2nd as shown below



b. Consider the given program to be executed on 2-issue superscalar processor with following specification.

- There are two generic execution units
- Memory type instructions take 2 cycles in execution
- Forwarding is implemented
- In case of false dependencies, write back should be in order. No need to wait in decode stage

Show two iterations of the above code



c. Your task is to do dynamic scheduling based on Tomasulo algorithms. There are:

- 2 LOAD/STORE Functional Units (LDU1, LDU2) with latency of 4 cycles
- 2 ALU/BR Function Units (ALU1, ALU2) with latency 2 cycles
- 2 RESERVATION STATIONS (add1, add2) for ALU/BR operation
- 2 Buffers for LOAD/STORE (Load1, Load2)

Moreover you have to identify:

- Structural hazards for RS in ISSUE phase
- RAW hazards in START EXECUTE phase

Show a single iteration of the loop using Tomsula