Rubrics_Mid Semester Exam- 2023
CSE/ECE 511 Computer Architecture

Q1. A CPU has a 3-level cache hierarchy with L1, L2, and L3 caches such that L1 sits closest to the CPU and L3 sits closest to the RAM. The hit-rate, the cycles required to check if there is a hit, the cycles required to access the cache (for both Direct mapped and 2-way set associative) in case of a hit, and the RAM access latency are given below in the table. Note that each level of the cache will have to pay the latency cost to check for hits regardless of whether the access is a hit or not.

Note: You have to assume the same values for various parameters for RAM in Direct-mapped and 2-way cases. The same has been done in the table.

Memory Type	Hit-Rate -- Direct mapped	Hit-rate -- 2-way	Hit check latency -- Direct mapped (Cycles)	Hit check latency -- 2-way (Cycles)	Access latency -- Direct mapped (Cycles)	Access latency -- 2-way (Cycles)	Avg. access latency -- Direct mapped	Avg. access latency -- 2-way
L1	80.00%	84.00%	100	110	50	55		
L2	90.00%	92.70%	1000	1100	500	550		
L3	95.00%	96.90%	10000	11000	5000	5500		
RAM	100.00%	100.00%	0	0	50000	50000		

(a) Compute the expected (average) data access latency for each level of this memory system for both the cache types: direct mapped and 2-way set-associative. The values to be computed correspond to the last two empty columns in the table. You may round off to the nearest integer. *(Assume only one mapping type exists for all levels at a time; do not mix and match the two mappings while calculating expected latencies.)* **[8 Marks]**

(b) What should the following conditions' impact be on expected average latencies for various hierarchical CACHE levels of memory? Also, mention the reasons. (*Assume only one condition applies at a time.)*

(i) Increase in hit-rate [2 Marks]

(ii) Increase in hit-check and access latencies [2 Marks]

(c) Does changing the cache mapping from direct mapped to 2-way have the same impact on expected access latencies for all the three cache levels, i.e., if increment or decrement in values calculated in part (a) across the last two columns yield the same trend for all the L1, L2 and L3 caches? Write the reason for your conclusion with proper analysis. Also, state the cache level(s) that show inconsistency, if any. [4 Marks]

(d) As a designer, using the calculated results and other data in this question, you have to choose cache from two options: direct mapped and 2-way associative cache to optimize the system for expected access latency for each cache level. State which type of associativity (Direct-mapped or 2-way) you will consider for all three levels of the cache. Also, mention the reason. [4 Marks]

Ans:

(a) <u>Cases for evaluation:</u>

(i) If <u>*ONLY*</u> Correct formula is written and no calculations have been done ⇒ **[2 Marks]**
This will NOT change Maximum marks, i.e., 8 Marks.
Formula:
Expected Latency = (hit_rate)*(hit_check_latency + access_latency) +
(1-hit_rate)*(hit_check_latency + expected_latency_of_next_or_higher_level)

(ii) If final values are wrong, but corresponding correct-calculations/correct-formula is shown ⇒ **[0.5 Marks for every correct calculation shown, Zero otherwise.]**

(iii) For correct table entries: [1*8 = 8 Marks]

Note: Provided the CORRECT formula is applied and calculated values are VERY close to that in this table, they will be treated as correct, and the TREND shown should MATCH with the following table.

Memory Type	Hit-Rate -- Direct mapped	Hit-rate -- 2-way	Hit check latency -- Direct mapped (Cycles)	Hit check latency -- 2-way (Cycles)	Access latency -- Direct mapped (Cycles)	Access latency -- 2-way (Cycles)	Avg. access latency -- Direct mapped	Avg. access latency -- 2-way
L1	80.00%	84.00%	100	110	50	55	775	622.60856
L2	90.00%	92.70%	1000	1100	500	550	3175	2915.0535
L3	95.00%	96.90%	10000	11000	5000	5500	17250	17879.5
RAM	100.00%	100.00%	0	0	50000	50000	50000	50000

(iv) All other possibilities will be dealt with on a case-to-case basis only.

(b) Impact:

(i) Avg. Latency should reduce **[1 Mark]**

Reason: **Avg, latency is directly proportional to hit-rate.** If hit-rate is increased, **higher levels of cache** will be accessed **fewer number of times** as the **probability** of **data** being **available in present level of cache** will **increase**, **avoiding the need to go to lower levels having higher latencies.**

(ii) Avg. Latency should increase **[1 Mark]**

Reason: Hit-check and access latencies are **directly proportional to the avg. access latency**. Increasing both means **compulsory hit-check time increases,** having **direct impact** and **access latency has both direct and indirect** impact in case of **hit (access present level) or miss (access next level)** respectively.

(c) No. The combined effect is that average latency reduces for L1 & L2 but increases for L3, as we increase associativity to 2-way. **[1 Mark]**

Reason(s):

[2 Marks if all three points are correctly mentioned or their meaning is clearly implied in the answer]

- The effect of **increased hit-rate** is more pronounced in **L1 & L2** than the **increase in hit-check and access latencies.**
- The effect of **increased hit-check and access latencies** in **L3** is more pronounced than the **increase in hit rate.**
- This is why the average access latency for 2-way in L3 is higher than direct mapped, and vice-versa is observed in L1 & L2.

Inconsistency occurs in **L3** as **compared** with **L1 & L2** cases. **[1 Mark]**

Note: If you observe correct inconsistency and your calculations and/or formula are wrong, a maximum of 3 Marks are awarded.

(d) Note: Some students have given the same associativity for all three cache levels, so marks have been given on the basis of overall correctness, and uniform grading across all students has been ensured.

L1 & L2 → 2-way associative **[1 Mark]**

L3 → Direct-mapped **[1 Marks]**

The above decisions are due to **lower average access latencies** for **direct-mapped vs 2-way set associative** for 3 levels of caches. **[1+1 Marks for correct reasoning using underlined and bold keywords]**

Note: If you have selected the correct cache type(s), but your calculations and/or formula are incorrect, a maximum of 2 marks are awarded.

Q2. Suppose a program is run on two different compilers, A and B. Execution of the program using compiler A results in a dynamic instruction count of $1x10^9$ and execution times of 1.1s. Execution of the same program using compiler B results in a dynamic instruction count of $1.2x10^9$ and an execution time of 1.5s.

(a) Find the average CPI for each case, given that the processor has a 1GHz clock rate. [4 Marks]

(b) The two compiled programs run on two different processors respectively. If the execution times on the two processors are the same. How much faster is the clock of the processor running compiler A's code compared to the clock of the processor running compiler B's code? [2 Mark]

(c) If a new compiler is developed and it uses only $6 X 10^8$ instructions and has an average CPI of 1.1. What is the speedup of using the new compiler compared to using compilers A and B on the original processor from (a)? [3 Marks]

(d) Complete the table: [6 Marks]

(Binary Marking: Zero marks for every cell which is incorrectly/partially/excessively filled.)

Parameters	**Depends On-** (hardware/ISA/compiler/CPU Design/program used)
Clock cycle time	
CPI	
Instruction Count	

Ans:

NOTE: For d), marks will be given only to those students who fill up ALL the dependencies correctly.

a) Clock cycle time = 1/clock rate

=1/(1GHz) = $1\mathbf{0^{-9}}$ s

Since,

CPU time = Instruction count x CPI x Clock cycle time

Compiler A CPI = $\frac{1.1}{10^9 X\ 10^{-9}}$ =1.1 [2 Marks]

Compiler B CPI= $\frac{1.5}{1.2\ X\ 10^9 X\ 10^{-9}}$ =1.25 [2 Marks]

b) For same execution time on both processors,

Clock cycle time of A/Clock cycle time of B = (Instruction count(B) × CPI(B)) / (Instruction count (A) × CPI(A)) = 1.36

[2 Marks]

c)Speedup compare to A =(CPU time of compiler A) / (CPU time of NEW compiler)

= (Instruction count×CPI) / (Instruction count×CPI)

= (1 x $1\mathbf{0^9}$ x 1.1) / (6 $x10^8$ x 1.1)

= 1.67 [1.5 Marks]

Speedup compare to B =(CPU timeof compiler B) / (CPU time of NEW compiler)

= (Instruction count×CPI) / (Instruction count×CPI)

= (1.2 x 10^9 x 1.25) / (6 x10^8 x 1.1)
= 2.27 [1.5 Marks]

Note: All the calculations done in inverse will also be given marks for speedup/comparison cases.

d) 2 *3 = 6 Marks

Parameters	Depends On-(hardware/ISA/compiler/CPU Design/program used)
Clock cycle time	Hardware, CPU Design
CPI	CPU Design,ISA, compiler, Program used
Instruction Count	Program used, ISA and compiler

Q3. For a given in-order processor with 5 stages (Fetch, Decode, Execute, Memory, Writeback) and full by passing, determine the number of clock cycles needed to execute the following code. Note that L1 is a label and not a part of the instruction. [15 Marks]

ADD R1, R2, R3
LW R2, 4(R3)
SUB R5,R2, R1
LW R4, 4(R3)
BEQ R2,R4,L1
ADD R6,R0,R1
SUB R5,R6,R0
L1: ADD R7,R8,R9

Ans.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ADD R1, R2, R3	F	D	E	M	W									
LW R2, 4(R3)		F	D	E	M	W								
SUB R5,R2, R1			F	D	D	E	M	W						
LW R4, 4(R3)				F	F	D	E	M	W					
BEQ R2,R4,L1						F	D	D	E	M	W			
ADD R6,R0,R1							F	F	D	-	-			
SUB R5,R6,R0									F	-	-			
L1: ADD R7,R8,R9										F	D	E	M	W

(1.5 + 1.5 + 2.5 + 1.5 +2.5 + 2 + 2 +1.5)