{ Shengxin Qian
Xinquan Liu

Question 1:

A 16-byte cache has 4-byte blocks, has 2 sets, and is 2-way set-associative. The cache initially is empty (all valid bits are off: indicated by a blank box in the table below). The cache receives requests in the sequence listed below. For each address in the sequence (a) split it into the tag, index, and offset; (b) categorize the access as a *hit*, a *compulsory miss*, a *conflict miss*, or a *capacity miss* (You can abbreviate hit=H, Compulsory=O, Conflict=F, Capacity=P); (c) show the new contents of the cache after the access—write the tags for each way, and note which way is LRU. The first one is done for you:

	Address	Split Address				Set 0			Set 1		
		Tag	Index	Offset	Result	Way 0	Way 1	LRU Way	Way 0	Way 1	LRU Way
	FF	1F	1	3	O			0	1F		1
00100010	22	04	0	2	O	04		1	1F		1
01000010	42	08	0	2	O	04	08	0	1F		1
00100001	31	06	0	1	O	06	08	1	1F		1
01000011	43	08	0	3	H	06	08	0	1F		1
00110000	30	06	0	0	H	06	08	1	1F		1
00100011	23	04	0	3	F	06	04	0	1F		1
1111 1100	FC	1F	1	0	H	06	04	0	1F		1
0000 1000	08	01	0	0	O	01	04	1	1F		1
0001 0000	10	02	0	0	O	01	02	0	1F		1
0000 1000	08	01	0	0	H	01	02	1	1F		1
0011 0011	33	06	0	3	P	01	06	0	1F		1

Question 2:

You are designing the memory hierarchy for a processor which has a memory access latency of 120 ns.

1. The Level 3 cache can be accessed in 30 ns on a hit. What hit rate does it need to acheive an average access latency of 60ns?

   Miss rate: X ; hit rate: $1-X$

   L_3: $120 \times X\% + 30 = 60$

   $\therefore X = 25\%$

   hit rate: $1 - 25\% = 75\%$

2. Assuming the Level 3 cache acheives the hit rate in part 1 (thus its average access latency is 60ns), and that the Level 2 cache has a 80% hit rate, what hit latency does the L2 cache need in order to acheive an average access time of 25 ns?

   Hit latency: t

   $60 \times 20\% + t = 25$

   $\therefore t = 13\ ns$

3. You have two choices for the Level 1 data cache design:

   Option A 64KB, 8-way set associative, 2.5 ns hit latency, 95% hit rate

   Option B 32KB, 4-way set associative, 1 ns hit latency, 90% hit rate

   Assuming the 60ns L3 and 25ns L2 average access times above, which option would you pick? Why?

   A: $25 \times 5\% + 2.5 = 3.75ns$

   B: $25 \times 10\% + 1 = 3.5ns$ ✓ choose B.

   → so choose B.

   the average access time of B is smaller than the average access time of A,

4. Some processors support the ability to dynamically increase their clock frequency (and voltage so that the logic runs faster) while they run. Suppose that the above data assumes the processor is running at 2.0GHz, but the processor can increase its frequency up to 4GHz. If the processor were to spend most of its time running this application at 4.0GHz, would your answer to part (3) change? Why or why not?

   L_1, L_2: half the latency time of before.

   L_3: same latency time as before.

   A: $\{[(120 \times 25\% + 30) \times 20\% + 6.5\} \times 5\% + 1.25 = 2.175ns$ ✓

   B: $\{[(120 \times 25\% + 30) \times \frac{20\%}{2} + 6.5\} \times 10\% + 0.5 = 2.35ns$

   the average access time of A is smaller than the average access time of B, so choose A.