CS/CoE 1541 – Exam 2 (Spring 2019). Name: _____

Question 1 (5+5+5=15 points): Show the content of each of the caches shown below after the two memory references

35, 44

Use the notation [tag, M(address),...] to describe the content of each entry. For example [4,M(46)] indicates that the entry contains tag=4 and the data from memory location 46. Similarly, [4,M(46),M(47)] indicates that the entry contains a block of two words from locations 46 and 47.

(a) An 8-words, direct mapped cache with block size = 1 word	
Index	Content of cache
0	
1	
2	
3	[4, M(35)]
4	[5, M(44)]
5	
6	
7	

(b) A 16-words, direct mapped cache with block size = 2 words	
Block index	Content of cache
0	
1	[2, M(34), M(35)]
2	
3	
4	
5	
6	[2, M(44), M(45)]
7	

(c) A 64-words, 2-ways set associative cache with Block size = 4 words		
Block Index	Way 1	Way 2
0	[1, M(32), M(33), M(34),M(35)]	
1		
2		
3	[1, M(44), M(45), M(46), M(47)]	
4		
5		
6		
7		

Question 2 (8+4+3=15 points):

(a) Consider a 256 KB, 4-way set-associative L1 cache with 16-Bytes blocks in a 32-bit byte addressable system. Complete the following sentences assuming that a byte address is of the form $b_{31}, b_{30}, \ldots, b_1, b_0$:

(i) The total number of blocks in the cache is $2^{18} / 2^4 = 2^{14} = 16$ K blocks

(ii) Of the 32 bits of the address, The number of bits used for indexing the cache is 12 and the number of bits used for tagging is 16

(iii) The address bits used for indexing a block in the cache are b__ , b__ , … , b __ , b__ (Identify the subscripts)

b15, b14, … , b5, b4

(iv) The address bits that are used as a tag for a block are b__ , b__ , … , b __ , b__ (Identify the subscripts)

b31, b30, … , b17, b16

(v) The total number of bits used as tags for all the blocks in the cache is

16 K blocks * 16 bits = 256K bits.

(b) Assume that the miss rate for this cache is 5% and that the miss penalty (number of cycles needed to access a block in memory) is 90 cycles. What is the average memory access time (in cycles) for the memory hierarchy (composed of the cache + the memory)?

1 + 0.05 * 90 = 1 + 4.5 = 5.5 cycles

(c) Assume that an L2 is added to the memory hierarchy and that the access time for the L2 is 6 cycles. What is the he average memory access time (in cycles) for the memory hierarchy if 80% of the accesses to the L2 (the misses from the L1) are hits.

1 + 0.05 * (6 + 0.2 * 90) = 1 + 0.05 * 24 = 2.2 cycles

Question 3 (6+6+3 = 15 points):

a) Assume that the time to open a row in a DRAM bank is 100ns, the time to close the row is 80 ns and the time to access a column in the row buffer is 20 ns. Assume also that the memory controller accesses two bytes X and Y in succession (issues a request for Y as soon as it receives X) and neither of them is in the row buffer. Compute the delay from the time the request to X is issued to the time Y is returned to the memory controller in the following cases:

(i) X and Y are in the same DRAM row and the open row policy is applied

80 + 100 + 20 + 20 = 220 (note that the row is open before X is accessed)
100 + 20 + 20 = 140 (this answer is OK if you assume that the row is initially closed)

(ii) X and Y are in the same DRAM row and the closed row policy is applied

100 + 20 + 80 + 100 + 20 = 320

(iii)X and Y are not in the same DRAM row and the open row policy is applied

80 + 100 + 20 + 80 + 100 + 20 = 400 (the row is open before X is accessed)
100 + 20 + 80 + 100 + 20 = 320 (if you assume that the row is initially closed)

b) Assuming that each of the two nxn arrays A and B is stored row-wise in memory (A[0][0], A[0][1], A[0][2], …. and B[0][0], B[0][1], B[0][2], ….), complete the following sentences pertaining to the cache miss rate during the execution of the following loop which stores the transpose of array A into array B

```
for(i=0;  i<n ; i++)
   for(j=0; j<n; j++)
      B[j][i] = A[i][j] ;
```

(i) When all the elements of A and B can fit into the cache and the cache block size = 4 words, the cache miss rate during the execution of the above loop is 25% for array A and 25% for array B.

(ii) When the cache can only fit 4n elements (of either A or B) while the cache block size is 4 elements, the cache miss rate during the execution of the above loop is 25% for array A and 100% for array B.

(iii)When the cache can only fit 4n elements of A and 4n elements of B while the cache block size is 4 elements, the cache miss rate during the execution of the above loop is 25% for array A and 25% for array B.

c) What is the least recently used reference in the sequence of references 3, 4, 2, 1, 1, 4, 3, 4, 2, 4, 3?

Reference 1

Question 4: (8+4+3=15 points):

(a) For each entry in the following table, indicate if the statement is true or false by marking the appropriate column.

		true	**false**
1	In a write through cache with write-allocate, every store operation to a location causes a write to that location both in the cache and in memory	X	
2	Having larger cache block sizes explores the property of temporal locality		X
3	The CPU pipeline stalls if there is a miss in the instruction cache but not if there is a miss in the data cache.		X
4	Larger cache block sizes always result in a lower cache miss rate		X
5	Each process in the system has its own page table	X	
6	For the same virtual and physical memory sizes, the size of the page table increases when the page size is smaller	X	
7	The Page Table walker is invoked when a page fault is detected		X
8	With interleaved memory, the delay of fetching a block of K words is smaller than K times the delay of fetching one word.	X	

(b) The (8,12) Hamming code encodes 8-bits data words into 12-bits code words using 4 parity bits (p1, p2, p3 and p4) using the following template.

Encoded data bits		p1	p2	d1	p4	d2	d3	d4	p8	d5	d6	d7	d8
Parity bit coverage	p1	x		x		x		x		x		x	
	p2		x	x			x	x			x	x	
	p4				x	x	x	x					x
	p8								x	x	x	x	x

Complete the following sentences:

(i) When encoding 00100100 the parity bits are:

p1 = 0 p2= 0 p4= 1 p8= 1

and the corresponding code word is: 000101010100

(ii) Assuming that there are no errors in the code word 011001000011, the corresponding data word is

10100011

Question 5 (10+5=15 points):

Consider a computer system that supports a 1MB virtual address space (byte addressable) with 4KB pages, 512KB physical memory and a 4-entries, **2-way associative** TLB. Also, assume that the TLB configuration and the (partial) content of the page table are as shown:

TLB					
v	**Tag**	**Physical page #**	**v**	**Tag**	**Physical page #**
1	000 0011	010 0100	0		
1	000 1111	010 0101	1	000 0111	000 0101

	Page Table	
	v	**Physical page #**
. . .	. . .	. . .
0000 0011	1	111 0000
. . .	. . .	. . .
0000 0111	0	
. . .	. . .	. . .
0000 1110	1	000 1100
	. . .	. . .
0010 0000	0	
	. . .	. . .

a) Given the following sequence of references to virtual byte addresses, specify for each reference the virtual page number and indicate if the reference results in a TLB miss or hit. Also, for each reference, indicate the physical page number that results after the virtual-to-physical address translation process. If any reference results in a page fault, do not provide the physical page number.

Virtual byte address	→ **Virtual page #**	→ **TLB hit/mis**	→ **Physical page #**
0000 1111 0000 0100 0001	→ 0000 1111	→ hit	→ 000 0101
0000 0011 1011 1110 1001	→ 0000 0011	→ miss	→ 111 0000
0000 0110 0001 1111 1100	→ 0000 0110	→ hit	→ 010 0100
0010 0000 0001 0000 1100	→ 0010 0000	→ miss	→ page fault

b) Assuming that physical page 000 1000 is the page that is replaced in the physical memory in case of a page fault, show the content of the TLB after the addresses of part (a) are referenced in the give order and all the TLB misses and page faults are serviced.

TLB					
v	**Tag**	**Physical page #**	**v**	**Tag**	**Physical page #**
1	000 0011	010 0100	1	0010 000	000 1000
1	000 0001	111 0000	1	0000 111	000 0101

Question 6 (12+3=15 points): Consider 4 processors, P0, P1, P2 and P3, each with its private cache, with a snoopy bus connecting the caches to memory. Assume a cache block size of two words and let (x,y) and (z,w) be two cache blocks that map to the same cache location, L. As shown in the tables below, assume that initially, the caches of P0 and P1 contain (x,y) in the shared state (S), the cache of P2 contains (z,w) in the exclusive state (E) while location L in the cache of P3 is invalid (I). For each of the actions indicated in (a), (b) and (c**), start from the initial state** and list the bus activities and the final state of location L in all the caches. Use one or more of the following to specify the activities on the bus:

- Pi requests (,) to read
- Pi requests (,) to write
- Pi posts "invalidate" (,)
- Pi writes back (,)
- Memory returns the value of (,)

(a) P1 stores (writes) 50 into y

	State of L in P0	State of L in P1	State of L in P2	State of L in P3
Initial state	x=10, y=20 (S)	x=10, y=20 (S)	z=30, w=40 (E)	I
Bus activity(ies)	P1 posts invalidate block (x , y)			
Final state	I	x=10, y=50 (E)	z=30, w=40 (E)	I

(b) P2 stores (writes) 30 into x

	State of L in P0	State of L in P1	State of L in P2	State of L in P3
Initial state	x=10, y=20 (S)	x=10, y=20 (S)	z=30, w=40 (E)	I
Bus activity(ies)	P2 writes back block (z,w) P2 requests block (x,y) to write Memory returns the value of block (x,y)			
Final state	I	I	x=30, y=20 (E)	I

(c) P3 loads (reads) w

	State of L in P0	State of L in P1	State of L in P2	State of L in P3
Initial state	x=10, y=20 (S)	x=10, y=20 (S)	z=30, w=40 (E)	I
Bus activity(ies)	P3 requests block (z,w) to read P2 supplies the value of block (z,w)			
Final state	x=10, y=20 (S)	x=10, y=20 (S)	z=30, w=40 (S)	z=30, w=40 (S)

(d) Complete the following sentences:

- The state of a block is "shared" when the valid bit = 1 and the dirty bit = 0
- The state of a block is "exclusive" when the valid bit = 1 and the dirty bit = 1
- The state of a block is "Invalid" when the valid bit = 0 and the dirty bit = 0 or 1