Date:_____

Quiz for Chapter 5 Large and Fast: Exploiting Memory Hierarchy

Not all questions are of equal difficulty. Please review the entire quiz first and then budget your time carefully.

Name:_____

Course:_____

1. [24 points] Caches and Address Translation. Consider a 64-byte cache with 8 byte blocks, an associativity of 2 and LRU block replacement. Virtual addresses are 16 bits. The cache is physically tagged. The processor has 16KB of physical memory.

(a) What is the total number of tag bits?

(b) Assuming there are no special provisions for avoiding synonyms, what is the minimum page size?

(c) Assume each page is 64 bytes. How large would a single-level page table be given that each page requires 4 protection bits, and entries must be an integral number of bytes.

Name:_____

(d) For the following sequence of references, label the cache misses.Also, label each miss as being either a compulsory miss, a capacity miss, or a conflict miss. The addresses are given in octal (each digit represents 3 bits). Assume the cache initially contains block addresses: 000, 010, 020, 030, 040, 050, 060, and 070 which were accessed in that order.

Cache state prior to access	Reference address	Miss ? Which ?
(00,04),(01,05),(02,06),(03,07)	024	
(00,04),(01,05),(06,02),(03,07)	100	
(04,10),(01,05),(06,02),(03,07)	270	
(04,10),(01,05),(06,02),(07,27)	570	
(04,10),(01,05),(06,02),(27,57)	074	
(04,10),(01,05),(06,02),(57,07)	272	
(04,10),(01,05),(06,02),(07,27)	004	
(10,00),(01,05),(06,02),(07,27)	044	
(00,04),(01,05),(06,02),(07,27)	640	
(04,64),(01,05),(06,02),(07,27)	000	
(64,00),(01,05),(06,02),(07,27)	410	
(64,00),(05,41),(06,02),(07,27)	710	
(64,00),(41,71),(06,02),(07,27)	550	
(64,00),(71,55),(06,02),(07,27)	570	
(64,00),(71,55),(06,02),(27,57)	410	

(e) Which of the following techniques are aimed at reducing the cost of a miss: dividing the current block into sub-blocks, a larger block size, the addition of a second level cache, the addition of a victim buffer, early restart with critical word first, a writeback buffer, skewed associativity, software prefetching, the use of a TLB, and multi-porting.

(f) Why are the first level caches usually split (instructions and data are in different caches) while the L2 is usually unified (instructions and data are both in the same cache)?

Name:_____

2. [6 points] A two-part question.

(Part A)

Assume the following 10-bit address sequence generated by the microprocessor:

Time	0	1	2	3	4	5	6	7
Access	10001101	10110010	10111111	10001100	10011100	11101001	11111110	11101001
TAG								
SET								
INDEX								

The cache uses 4 bytes per block. Assume a 2-way set assocative cache design that uses the LRU algorithm (with a cache that can hold a total of 4 blocks). Assume that the cache is initially empty. First determine the TAG, SET, BYTE OFFSET fields and fill in the table above. In the figure below, clearly mark for each access the TAG, Least Recently Used (LRU), and HIT/MISS information for each access.

Initial			
	Block 0		Block 1
Set 0			
Set 1			

Access 0			
	Block 0		Block 1
Set 0			
Set 1			

Access 1			
	Block 0		Block 1
Set 0			
Set 1			

Access 2			
	Block 0		Block 1
Set 0			
Set 1			

Access 3			
	Block 0		Block 1
Set 0			
Set 1			

Access 4			
	Block 0		Block 1
Set 0			
Set 1			

Access 5			
	Block 0		Block 1
Set 0			
Set 1			

Access 6			
	Block 0		Block 1
Set 0			
Set 1			

Access 7			
	Block 0		Block 1
Set 0			
Set 1			

Name:_____

(Part B)

Derive the hit ratio for the access sequence in Part A.

Name:_____

3. [6 points] A two part question

(a) Why is miss rate not a good metric for evaluating cache performance? What is the appropriate metric? Give its definition. What is the reason for using a combination of first and second- level caches rather than using the same chip area for a larger first-level cache?

(b) The original motivation for using virtual memory was "compatibility". What does that mean in this context? What are two other motivations for using virtual memory?

Name:_____

4. [6 points] A four part question

(Part A)

What are the two characteristics of program memory accesses that caches exploit?

(Part B)

What are three types of cache misses?

Cold misses, conflict misses and compulsory misses

(Part C)

Design a 128KB direct-mapped data cache that uses a 32-bit address and 16 bytes per block. Calculate the following:

(a) How many bits are used for the byte offset?

(b) How many bits are used for the set (index) field?

(c) How many bits are used for the tag?

Name:_____

(Part D)

Design a 8-way set associative cache that has 16 blocks and 32 bytes per block. Assume a 32 bit address. Calculate the following:

(a) How many bits are used for the byte offset?

(b) How many bits are used for the set (index) field?

(c) How many bits are used for the tag?

Name:_____

5. [6 points] The memory architecture of a machine X is summarized in the following table.

Virtual Address	54 bits
Page Size	16 K bytes
PTE Size	4 bytes

(a) Assume that there are 8 bits reserved for the operating system functions (protection, replacement, valid, modified, and Hit/Miss- All overhead bits) other than required by the hardware translation algorithm. Derive the largest physical memory size (in bytes) allowed by this PTE format. Make sure you consider all the fields required by the translation algorithm.

(b) How large (in bytes) is the page table?

(c) Assuming 1 application exists in the system and the maximum physical memory is devoted to the process, how much physical space (in bytes) is there for the application's data and code.

Name:_____

6. [6 points] This question covers virtual memory access. Assume a 5-bit virtual address and a memory system that uses 4 bytes per page. The physical memory has 16 bytes (four page frames). The page table used is a one-level scheme that can be found in memory at the PTBR location. Initially the table indicates that no virtual pages have been mapped. Implementing a LRU page replacement algorithm, show the contents of physical memory after the following virtual accesses: 10100, 01000, 00011, 01011, 01011,11111. Show the contents of memory and the page table information after each access successfully completes in figures that follow. Also indicate when a page fault occurs. Each page table entry (PTE) is 1 byte.

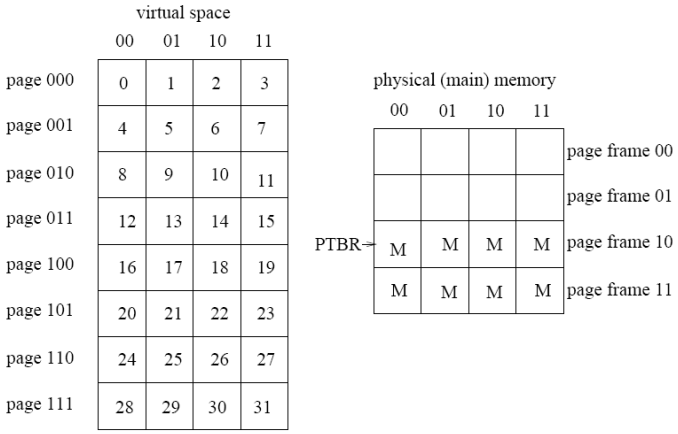


Figure 1: The initial contents of memory.

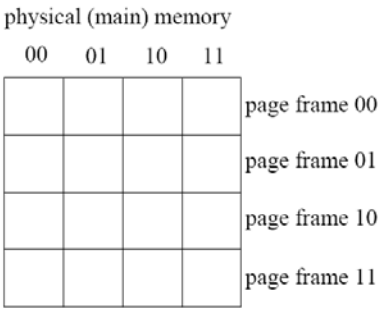


Figure 2: Figure (after access 10100).

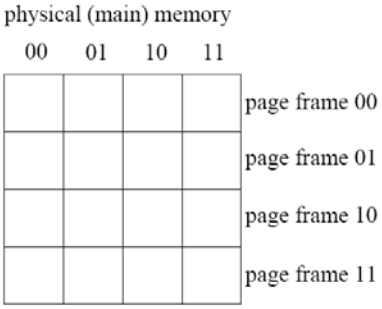


Figure 3: Figure (after access 01000).

Name:_____

physical (main) memory

00	01	10	11	
				page frame 00
				page frame 01
				page frame 10
				page frame 11

Figure 4: Figure (after access 00011).

physical (main) memory

00	01	10	11	
				page frame 00
				page frame 01
				page frame 10
				page frame 11

Figure 5: Figure (after access 01011).

physical (main) memory

00	01	10	11	
				page frame 00
				page frame 01
				page frame 10
				page frame 11

Figure 6: Figure (after access 01011).

physical (main) memory

00	01	10	11	
				page frame 00
				page frame 01
				page frame 10
				page frame 11

Figure 7: Figure (after access 11111).

Name:_____

7. [6 points] A multipart question.

(Part A)

In what pipeline stage is the branch target buffer checked?

(a) Fetch
(b) Decode
(c) Execute
(d) Resolve

What needs to be stored in a branch target buffer in order to eliminate the branch penalty for an unconditional branch?

(a) Address of branch target
(b) Address of branch target and branch prediction
(c) Instruction at branch target.

The Average Memory Access Time equation (AMAT) has three components: hit time, miss rate, and miss penalty. For each of the following cache optimizations, indicate which component of the AMAT equation is improved.

- Using a second-level cache
- Using a direct-mapped cache
- Using a 4-way set-associative cache
- Using a virtually-addressed cache
- Performing hardware pre-fetching using stream buffers
- Using a non-blocking cache
- Using larger blocks

Name:_____

(Part B)

(a) (True/False) A virtual cache access time is always faster than that of a physical cache?

(b) (True/False) High associativity in a cache reduces compulsory misses.

(c) (True/False) Both DRAM and SRAM must be refreshed periodically using a dummy read/write operation.

(d) (True/False) A write-through cache typically requires less bus bandwidth than a write-back cache.

(e) (True/False) Cache performance is of less importance in faster processors because the processor speed compensates for the high memory access time.

(f) (True/False) Memory interleaving is a technique for reducing memory access time through increased bandwidth utilization of the data bus.

Name:_____

8. [12 points] A three-part question. This question covers cache and pipeline performance analysis.

(Part A)

Write the formula for the average memory access time assuming one level of cache memory:

(Part B)

For a data cache with a 92% hit rate and a 2-cycle hit latency, calculate the average memory access latency. Assume that latency to memory and the cache miss penalty together is 124 cycles. Note: The cache must be accessed after memory returns the data.

(Part C)

Calculate the performance of a processor taking into account stalls due to data cache and instruction cache misses. The data cache (for loads and stores) is the same as described in Part B and 30% of instructions are loads and stores. The instruction cache has a hit rate of 90% with a miss penalty of 50 cycles. Assume the base CPI using a perfect memory system is 1.0. Calculate the CPI of the pipeline, assuming everything else is working perfectly. Assume the load never stalls a dependent instruction and assume the processor must wait for stores to finish when they miss the cache. Finally, assume that instruction cache misses and data cache misses never occur at the same time. Show your work.

- Calculate the additional CPI due to the icache stalls.
- Calculate the additional CPI due to the dcache stalls.
- Calculate the overall CPI for the machine.

Name:_____

9. [12 points] A three-part question.

(Part A)

A processor has a 32 byte memory and an 8 byte direct-mapped cache. Table 0 shows the current state of the cache. Write hit or miss under the each address in the memory reference sequence below. Show the new state of the cache for each miss in a new table, label the table with the address, and circle the change:

Addr	10011	00001	00110	01010	01110	11001	00001	11100	10100
H/M									

0. Initial state

Index	V	Tag	Data
000	N		
001	Y	00	Mem(00001)
010	N		
011	Y	11	Mem(11011)
100	Y	10	Mem(10100)
101	Y	01	Mem(01101)
110	Y	00	Mem(00110)
111	N		

1.

Index	V	Tag	Data
000			
001			
010			
011			
100			
101			
110			
111			

2.

Index	V	Tag	Data
000			
001			
010			
011			
100			
101			
110			
111			

3.

Index	V	Tag	Data
000			
001			
010			
011			
100			
101			
110			
111			

4.

Index	V	Tag	Data
000			
001			
010			
011			
100			
101			
110			
111			

5.

Index	V	Tag	Data
000			
001			
010			
011			
100			
101			
110			
111			

6.

Index	V	Tag	Data
000			
001			
010			
011			
100			
101			
110			
111			

7.

Index	V	Tag	Data
000			
001			
010			
011			
100			
101			
110			
111			

Name:_____

(Part B)

Do the same thing as in Part A, except for a 4-way set associative cache. Assume 00110 and 11011 were the last two addresses to be accessed. Use the Least Recently Used replacement policy.

Addr	10011	00001	00110	01010	01110	11001	00001	11100	10100
H/M									

0.

Set	Tag	Data	Tag	Data	Tag	Data	Tag	Data
0			0011	Mem(00110)			1010	Mem(10100)
1	0000	Mem(00001)	1101	Mem(11011)	0110	Mem(01101)		

1.

Set	Tag	Data	Tag	Data	Tag	Data	Tag	Data
0								
1								

2.

Set	Tag	Data	Tag	Data	Tag	Data	Tag	Data
0								
1								

3.

Set	Tag	Data	Tag	Data	Tag	Data	Tag	Data
0								
1								

4.

Set	Tag	Data	Tag	Data	Tag	Data	Tag	Data
0								
1								

5.

Set	Tag	Data	Tag	Data	Tag	Data	Tag	Data
0								
1								

6.

Set	Tag	Data	Tag	Data	Tag	Data	Tag	Data
0								
1								

7.

Set	Tag	Data	Tag	Data	Tag	Data	Tag	Data
0								
1								

Name:_____

(Part C)

(a) What is the hit and miss rate of the direct-mapped cache in Part A?

(b) What is the hit and miss rate of the 4-way set associative cache in Part B?

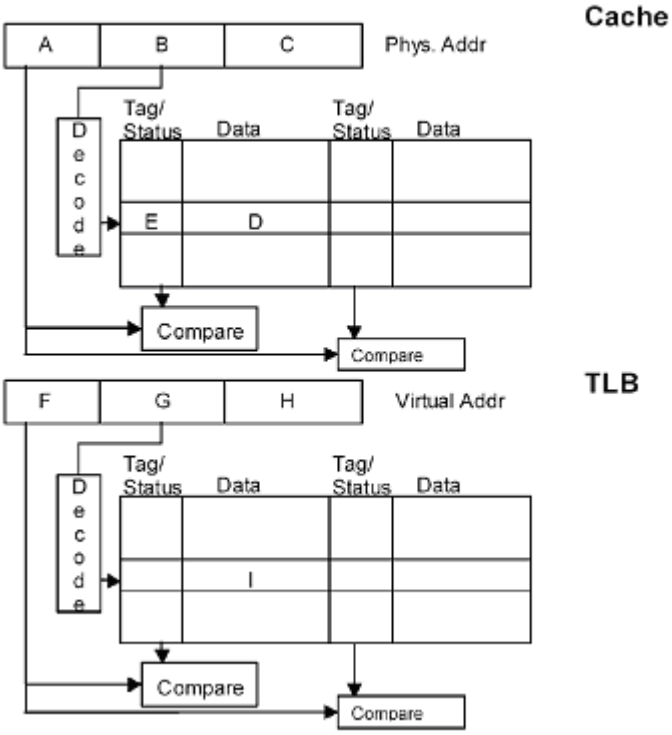
(c) Assume a machine with a CPI of 4 and a miss penalty of 10 cycles. Ignoring writes, calculate the ratio of the performance of the 4-way set associative cache to the direct-mapped cache. In other words, what is the speedup when using the machine with the 4-way cache?

Name:_____

10. [6 points] Consider a memory system with the following parameters:

- Translation Lookaside Buffer has 512 entries and is 2-way set associative.
- 64Kbyte L1 Data Cache has 128 byte lines and is also 2-way set associative.
- Virtual addresses are 64-bits and physical addresses are 32 bits.
- 8KB page size

Below are diagrams of the cache and TLB. Please fill in the appropriate information in the table that follows the diagrams:



L1 Cache		TLB	
A =	bits	F =	Bits
B =	bits	G =	Bits
C =	bits	H =	Bits
D =	bits	I =	Bits
E =	bits		

Name:_____

11. [3 points] Invalidation vs. Update-based Protocols.

(a) As miss latencies increase, does an update protocol become more or less preferable to an invalidation-based protocol? Explain.

(b) In a multilevel cache hierarchy, would you propagate updates all the way to the first-level cache or only to the second-level cache? Explain the trade-offs.

Name:_____

12. [6 points] How many total SRAM bits will be required to implement a 256KB four-way set associative cache. The cache is physically-indexed cache, and has 64-byte blocks. Assume that there are 4 extra bits per entry: 1 valid bit, 1 dirty bit, and 2 LRU bits for the replacement policy. Assume that the physical address is 50 bits wide.

Name:_____

13. [6 points] TLB's are typically built to be fully-associative or highly set-associative. In contrast, first-level data caches are more likely to be direct-mapped or 2 or 4-way set associative. Give two good reasons why this is so.

Name:_____

14. [6 points] Caches: Misses and Hits

```
int i;
int a[1024*1024];
int x=0;

for(i=0;i<1024;i++)
{
x+=a[i]+a[1024*i];
}
```

Consider the code snippet in code above. Suppose that it is executed on a system with a 2-way set associative 16KB data cache with 32-byte blocks, 32-bit words, and an LRU replacement policy. Assume that int is word-sized. Also assume that the address of a is 0x0, that i and x are in registers, and that the cache is initially empty. How many data cache misses are there? How many hits are there?

Name:_____

15. [6 points] Virtual Memory

(a) 32-bit Virtual Address Spaces. Consider a machine with 32-bit virtual addresses, 32-bit physical addresses, and a 4KB page size. Consider a two-level page table system where each table occupies one full page. Assume each page table entry is 32 bits long. To map the full virtual address space, how much memory will be used by the page tables?

(b) 64-bit Virtual Address Spaces. A two-part question.

(Part 1)

Consider a machine with a 64-bit virtual addresses, 64-bit physical addresses, and a 4MB page size. Consider a two-level page table system where each table occupies one full page. Assume each page table entry is 64 bits long. To map the full virtual address space, how much memory will be used by the page tables? (Hint: you will need more than 1 top-level page table. For this question this is okay.)

(Part 2)

Rather than a two-level page table, what other page table architecture could be used to reduce the memory foot print of page tables for the 64-bit address space from the last question? Assume that you do not need to map the full address space, but some small fraction (people typically do not have 264 bytes of physical memory). However, you should assume that the virtual pages that are mapped are uniformly distributed across the virtual address space (i.e. it is not only the low addresses or high addresses that are mapped, but rather a few pages from all ranges of memory).

Name:_____

16. [6 points] Consider an architecture that uses virtual memory, a two-level page table for address translation, as well as a TLB to speed up address translations. Further assume that this machine uses caches to speed up memory accesses. Recall that all addresses used by a program are virtual addresses. Further recall that main memory in the microarchitecture is indexed using physical addresses. The virtual memory subsystem and cache memories could interact in several ways. In particular, the cache memories could be accessed using virtual addresses. We will refer to this scheme as a virtually indexed, virtually tagged cache. The cache could be indexed using virtual addresses, but the tag compare could happen with physical addresses (virtually indexed, physically tagged). Finally, the cache could be accessed using only the physical address. Describe the virtues and drawbacks for each of these systems. Be sure to consider the case where two virtual addresses map to the same physical address.

	Virtually Indexed, Virtually Tagged	Virtually indexed, physically tagged	Physically indexed, physically tagged
Advantages			
Disadvantages			

Name:_____

17. [6 points] Describe the general characteristics of a program that would exhibit very little temporal and spatial locality with regard to instruction fetches. Provide an example of such a program (pseudo-code is fine). Also, describe the cache effects of excessive unrolling. Use the terms static instructions and dynamic instructions in your description.

Name:_____

18. [6 points] You are given an empty 16K 2-way set-associative LRU-replacement cache with 32 byte blocks on a machine with 4 byte words and 32-bit addresses. Describe in mathematical terms a memory read address sequence which yields the following Hit/Miss patterns. If such a sequence is impossible, state why. Sample sequences:

address(N) = N mod 2^{32} (= 0, 1, 2, 3, 4...)
address = (7, 12, 14)

(a) Miss, Hit, Hit, Miss
(b) Miss, (Hit)*
(c) (Hit)*
(d) (Miss)*
(e) (Miss, Hit)*

Name:_____

19. [3 points] Assume an instruction cache miss rate for gcc of 2% and a data cache miss rate of 4%. If a machine has a CPI of 2 without any memory stalls and the miss penalty is 40 cycles for all misses, determine how much faster a machine would run with a perfect cache that never missed. Assume 36% of instructions are loads/stores.

Name:_____

20. [12 points] Caching. "One of the keys to happiness is a bad memory." –Rita Mae Brown

Consider the following piece of code:

```
int x = 0, y = 0; // The compiler puts x in r1 and y in r2.
int i; // The compiler put i in r3.
int A[4096]; // A is in memory at address 0x10000
...
for (i=0;i<1024;i++) {
x += A[i];
}
for (i=0;i<1024;i++) {
y += A[i+2048];
}
```

(a) Assume that the system has a 8192-byte, direct-mapped data cache with 16-byte blocks. Assuming that the cache starts out empty, what is the series of data cache hits and misses for this snippet of code. Assume that ints are 32-bits.

(b) Assume that an iteration of a loop in which the load hits takes 10 cycles but that an iteration of a loop in which the load misses takes 100 cycles. What is the execution time of this snippet with the aforementioned cache?

(c) Repeat part A except assume that the cache is 2-way set associative with an LRU replacement policy and 16-byte sets (8-byte blocks).

(d) Repeat part B using the cache described in part C. Is the direct-mapped or the set-associative cache better?