The Essentials of Computer Organization and Architecture

Linda Null and Julia Lobur
Jones and Bartlett Publishers, 2003

Chapter 6 Instructor's Manual

Chapter Objectives

Chapter 6, Memory, covers basic memory concepts, such as RAM and the various memory devices, and also addresses the more advanced concepts of the memory hierarchy, including cache memory and virtual memory. This chapter gives a thorough presentation of direct mapping, associative mapping, and set-associative mapping techniques for cache. It also provides a detailed look at overlays, paging and segmentation, TLBs, and the various algorithms and devices associated with each. A tutorial and simulator for this chapter is available.

Lectures should focus on the following points:

- **Types of memory.** There are many types of memory, but the two basic categories are RAM and ROM.
- **The memory hierarchy.** One of the most important considerations in understanding the performance capabilities of a modern computer is the memory hierarchy. The goal of this section is understanding how system memory (registers, cache, and main memory), online memory (hard disk), near line memory (optical disk), and offline memory (tapes and floppy disks) work together to provide acceptable performance at a minimal cost. Locality of reference (or the clustering of memory references) is integral in understanding how a memory hierarchy works.
- **Cache memory.** The purpose of cache is to speed up memory accesses by storing recently used data closer to the CPU (in a memory that requires less access time). It is important to discuss where this data is stored in cache, so direct mapping, fully associative cache, and set associative cache are covered. The effective access time is a good way to measure the performance of cache.
- **Virtual memory.** Virtual memory is a method used to increase the available address space for a process by using the hard disk as an extension of RAM. Both paging and segmentation (including advantages and disadvantages) are covered. In addition TLBs are introduced as a method for improving performance of paging systems.
- **Real-world examples of memory management.** The chapter concepts are studied in the context of the Pentium memory hierarchy.

Required Lecture Time

The important concepts in Chapter 6 can typically be covered in 5 lecture hours. However, if a teacher wants the students to have a mastery of all topics in Chapter 6, 9 lecture hours are more reasonable. If lecture time is limited, we suggest that the focus be on the memory hierarchy and cache memory. Virtual memory is often covered in an operating systems course, but we provide coverage in this textbook because we feel it is important that students see the hierarchy in its entirety.

Lecture Tips

Students often have difficulty understanding why the memory hierarchy works. In particular, many miss the concept of bringing in an entire block when a "miss" occurs (thus using the principle of locality). Instructors should focus on several small examples to make sure students understand the concept of locality. This will also help with the section on paging.

Although the cache mapping schemes are relatively straight forward, students have a tendency to miss *why* the schemes are necessary. (For example, many can work examples involving the various mapping schemes, but they aren't sure exactly what they are doing and why they are doing it.) It is important to stress not only how the main memory address is mapped to a cache location but why.

For the various cache mapping schemes, we have found that working out small examples in detail is a good way to approach these concepts. Examples showing blocks of memory actually put into a small cache, and then that cache getting full (or several blocks mapping to the same location in cache) are very helpful for motivating the need for the tag in cache.

If students are comfortable with cache mapping, paging is much easier for them to deal with, as many concepts are similar. However, these similarities can cause confusion for the students, as some tend to mix up cache and paging. Covering examples using both cache and paging helps to clear up this confusion.

Answers to Exercises

1. Suppose a computer using direct mapped cache has 2^{20} words of main memory, and a cache of 32 blocks, where each cache block contains 16 words.

   a. How many blocks of main memory are there?
   b. What is the format of a memory address as seen by the cache, i.e., what are the sizes of the tag, block, and word fields?
   c. To which cache block will the memory reference $0DB63_{16}$ map?

Ans.

a. $2^{20} / 2^4 = 2^{16}$

b. 20 bit addresses with 11 bits in the tag field, 5 in the block field, and 4 in the word field

c. 0DB63 = 00001100101 10110 0111, which implies Block 22

---

2. Suppose a computer using direct mapped cache has 2^{32} words of main memory, and a cache of 1024 blocks, where each cache block contains 32 words.

a. How many blocks of main memory are there?
b. What is the format of a memory address as seen by the cache, i.e., what are the sizes of the tag, block, and word fields?
c. To which cache block will the memory reference $000063FA_{16}$ map?

Ans.

a. $2^{32} / 2^5 = 2^{27}$

b. 32 bit addresses with 17 bits in the tag field, 10 in the block field, and 5 in the word field

c. 000063FA = 00000000000000000 1100011111 11010, which implies Block 799

---

3. Suppose a computer using fully associative cache has 2^{16} words of main memory, and a cache of 64 blocks, where each cache block contains 32 words.

a. How many blocks of main memory are there?
b. What is the format of a memory address as seen by the cache, i.e., what are the sizes of the tag and word fields?
c. To which cache block will the memory reference F8C9 map?

Ans.

a. $2^{16} / 2^5 = 2^{11}$

b. 16 bit addresses with 11 bits in the tag field and 5 in the word field

c. Since it's associative cache, it can map anywhere

---

4. Suppose a computer using fully associative cache has 2^{24} words of main memory, and a cache of 128 blocks, where each cache block contains 64 words.

a. How many blocks of main memory are there?
b. What is the format of a memory address as seen by the cache, i.e., what are the sizes of the tag and word fields?
c. To which cache block will the memory reference $01D872_{16}$ map?

Ans.

a. $2^{24} / 2^6 = 2^{18}$

b. 24 bit addresses with 18 bits in the tag field and 6 in the word field

c. Since it's associative cache, it can map anywhere

---

5. Assume a system's memory has 128M words. Blocks are 64 words in length and the cache consists of 32K blocks. Show the format for a main memory address assuming a 2-way set associative cache mapping scheme. Be sure to include the fields as well as their sizes.

Ans.

Each address has 27 bits, and there are 7 in the tag field, 14 in the set field and 6 in the word field.

---

6. A 2-way set-associative cache consists of four sets. Main memory contains 2K blocks of eight words each.

a. Show the main memory address format that allows us to map addresses from main memory to cache. Be sure to include the fields as well as their sizes.

b. Compute the hit ratio for a program that loops 3 times from locations 8_{10} to 51_{10} in main memory. You may leave the hit ratio in terms of a fraction.

Ans.

a. $2K * 2^3 = 2^{14}$, so we have 14-bit addresses with 9 bits in the tag field, 2 bits in the set field (since we have four sets), and 3 in the word field

b. First iteration of the loop: Address 8 is a miss, then entire block brought into Set 1. 9-15 are then hits. 16 is a miss, entire block brought into Set 2, 17-23 are hits. 24 is a miss, entire block brought into Set 3, 25-31 are hits. 32 is a miss, entire block brought into Set 0, 33-39 are then hits. 40 is a miss, entire block brought into Set 1 (note we do NOT have to throw out the block with address 8 as this is 2-way set associative), 41-47 are hits. 48 is a miss, entire block brought into Set 2, 49-51 are hits. For the first iteration of the loop, we have 6 misses, and 5*7 + 3 hits, or 38 hits. On the remaining iterations, we have 5*8+4 hits, or 44 hits each, for 88 more hits. Therefore, we have 6 misses and 126 hits, for a hit ratio of 126/132, or 95.45%.

---

7. Suppose a computer using set associative cache has 2^{16} words of main memory, a cache of 32 blocks and each cache block contains 8 words.

   a. If this cache is 2-way set associative, what is the format of a memory address as seen by the cache, i.e., what are the sizes of the tag, set, and word fields?

   b. If this cache is 4-way set associative, what is the format of a memory address as seen by the cache?

Ans.

a. 2^{16} words of main memory implies we have 16 bits in an address. Cache contains 2^5 blocks, but each set must have 2 blocks, so we have $2^5/2=2^4$ sets. Therefore our 16-bit address is divided into 9 bits for the tag field, 4 bits for the set field, and 3 bits for the word field.

b. The 32 blocks in cache must now be divided into sets with 4 blocks each, implying we have only 8 sets. The tag field would now have 10 bits, the set field 3 bits, and the word field 3 bits.

---

8. Suppose a computer using set associative cache has 2^{21} words of main memory, and a cache of 64 blocks, where each cache block contains 4 words.

   a. If this cache is 2-way set associative, what is the format of a memory address as seen by the cache, i.e., what are the sizes of the tag, set, and word fields?

   b. If this cache is 4-way set associative, what is the format of a memory address as seen by the cache?

Ans.

a. 2^{21} words of main memory implies we have 21 bits in an address. Cache contains 2^6 blocks, but each set must have 2 blocks, so we have $2^6/2=2^5$ sets. Therefore our 21-bit address is divided into 14 bits for the tag field, 5 bits for the set field, and 2 bits for the word field.

b. The 64 blocks in cache must now be divided into sets with 4 blocks each, implying we have only 16 sets. The tag field would now have 15 bits, the set field 4 bits, and the word field 2 bits.

---

*9. Suppose we have a computer that uses a memory address word size of 8 bits. This computer has a 16-byte cache and 256 bytes of main memory. The computer accesses a number of memory locations throughout the course of running a program.

Suppose this computer uses direct-mapped cache. The format of a memory address as seen by the cache is shown below:

Tag 4 bits	Block 2 bits	Word 2 bits

The system accesses memory addresses (in hex) in this exact order: 6E, B9, 17, E0, 4E, 4F, 50, 91, A8, A9, AB, AD, 93, and 94. The memory addresses of the first four accesses have been loaded into the cache blocks as shown below. (The contents of the tag are shown in binary in addition to the entire address in hex.)

	Tag Contents	Cache Contents (represented by address)
Block 0	1110	E0
		E1
		E2
		E3

	Tag Contents	Cache Contents (represented by address)
Block 1	0001	14
		15
		16
		17

	Tag Contents	Cache Contents (represented by address)
Block 2	1011	B8
		B9
		BA
		BB

	Tag Contents	Cache Contents (represented by address)
Block 3	0110	6C
		6D
		6E
		6F

a. What is the hit ratio for the entire memory reference sequence given above?
b. What memory blocks will be in the cache after the last address has been accessed?

Ans.

a.

Address	**Hit or Miss**
6E	Miss, brought into Block 3 with tag 0110 (as shown)
B9	Miss, brought into Block 2 with tag 1011 (as shown)
17	Miss. brought into Block 1 with tag 0001 (as shown)
E0	Miss, brought into Block 0 with tag 1110 (as shown)
4E	Miss, brought into Block 3 with tag 0100
4F	Hit
50	Miss, brought into Block 0 with tag 0101
91	Miss, brought into Block 0 with tag 1001
A8	Miss, brought into Block 2 with tag 1010
A9	Hit
AB	Hit
AD	Miss, brought into Block 3 with tag 1010
93	Hit
94	Hit

Hit ratio is 4 hits out of 14 total accesses, or 28.6% (this is assuming we count the first 4 accesses as misses)

b. Block 0, with tag 1001, contains 90, 91, 92, 93
Block 1, with tag 1001, contains 94, 95, 96, 97

Block 2, with tag 1010, contains A8, A9, AA, AB

Block 3, with tag 1010, contains AC, AD, AE, AF

---

10. A direct-mapped cache consists of eight blocks. Main memory contains 4K blocks of eight words each. Access time for the cache is 22 ns and the time required to fill a cache slot from main memory is 300ns (this time will allow us to determine the block is missing and bring it into cache). Assume a request is always started in parallel to both cache and to main memory (so if it is not found in cache, we do not have to add this cache search time to the memory access). If a block is missing from cache, the entire block is brought into the cache and the access is restarted. Initially, the cache is empty.
    a. Show the main memory address format that allows us to map addresses from main memory to cache. Be sure to include the fields as well as their sizes.
    b. Compute the hit ratio for a program that loops 4 times from locations 0 to 67_{10} in memory.
    c. Compute the effective access time for this program.

Ans.

a. 4K = 2^{12} blocks * 2^3 words each implies 2^{15} words of main memory, so we have 15 bits in an address. Cache contains 2^3 blocks, so the block field requires 3 bits. Each block has 2^3 words, so the word field requires 3 bits. That leaves 9 bits for the tag field. Therefore our 15-bit address is divided into 9 bits for the tag field, 3 bits for the block field, and 3 bits for the word field.

b. Block 0 of main memory (addresses 0 - 7) and Block 8 of main memory (addresses 64 - 67) must share a cache block. The remaining blocks are brought in and are not replaced. So for each access to Block 0, there is one miss and 7 hits. For each access to Block 8, there is one miss and 3 hits. The remaining blocks have one miss each, with all other accesses being hits. If we loop 4 times, we have:

Block 0: 4 misses, 28 hits

Block 1: 1 miss, 31 hits

Block 2: 1 miss, 31 hits

Block 3: 1 miss, 31 hits

Block 4: 1 miss, 31 hits

Block 5: 1 miss, 31 hits

Block 6: 1 miss, 31 hits

Block 7: 1 miss, 31 hits

Block 8: 4 misses, 12 hits for a total of 15 misses, 257 hits, or a hit ratio of 94.49%.

---

11. Consider a byte-addressable computer with 24-bit addresses, a cache capable of storing a total of 64K bytes of data and blocks of 32 bytes. Show the format of a 24-bit memory address for:

    a. direct mapped
    b. associative
    c. 4-way set associative

Ans.

a. 64K = $2^6 2^{10} = 2^{16}$; $2^{16} / 2^5 = 2^{11}$ blocks in cache, so 11 bits are needed for the block field. 5 bits are needed for the word field, leaving 8 for the tag.

b. Again, 5 bits are needed for the word field, leaving 19 for the tag.

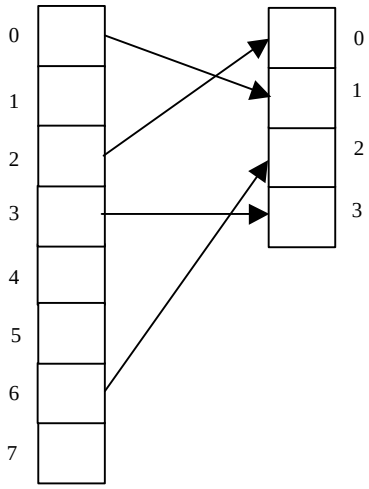
c. There are $2^{11}/2^2 = 2^9$ sets in cache, so 9 bits are needed for the set field. We still need 5 bits for the word field, leaving 10 for the tag field.

---

12. Suppose a process page table contains the entries shown below. Using the format shown in Figure 6.15a, indicate where the process pages are located in memory.

Frame	Valid Bit
1	1
--	0
0	1
3	1
--	0
--	0
2	1
--	0

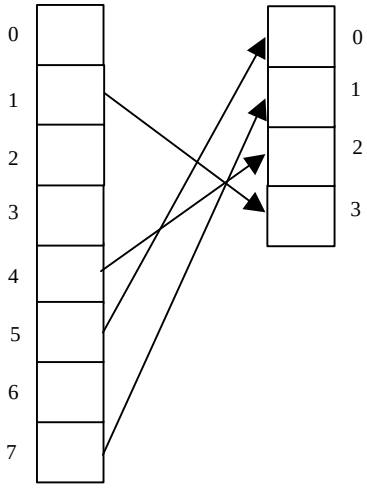
Ans.



♦13. Suppose a process page table contains the entries shown below. Using the format shown in Figure 6.15a, indicate where the process pages are located in memory.

Frame	Valid Bit
--	0
3	1
--	0
--	0
2	1
0	1
--	0
1	1

Ans.



---

*14. You have a virtual memory system with a two-entry TLB, a 2-way set associative cache and a page table for a process P. Assume cache blocks of 8 words and page size of 16 words. In the system below, main memory is divided up into blocks, where each block is represented by a letter. Two blocks equals one frame.

0	3
4	1

TLB

Set 0	tag	C	tag	I
Set 1	tag	D	tag	H

Cache

	Frame	Valid
0	3	1
1	0	1
2	-	0
3	2	1
4	1	1
5	-	0
6	-	0
7	-	0

Page Table

Frame		Block
0	C	0
	D	1
1	I	2
	J	3
2	G	4
	H	5
3	A	6
	B	7

Main Memory

Page		Block
0	A	0
	B	1
1	C	2
	D	3
2	E	4
	F	5
3	G	6
	H	7
4	I	8
	J	9
5	K	10
	L	11
6	M	12
	N	13
7	O	14
	P	15

Virtual Memory
For Process P

Given the system state as depicted above, answer the following questions:

a. How many bits are in a virtual address for process P? Explain.
b. How many bits are in a physical address? Explain.
c. Show the address format for virtual address 18_{10} (specify field name and size) that would be used by the system to translate to a physical address and then translate this virtual address into the corresponding physical address. (Hint: convert 18 to its binary equivalent and divide it into the appropriate fields.) Explain how these fields are used to translate to the corresponding physical address.

d. Given virtual address 6_{10} converts to physical address 54_{10}. Show the format for a physical address (specify the field names and sizes) that is used to determine the cache location for this address. Explain how to use this format to determine where physical address 54 would be located in cache. (Hint: convert 54 to binary and divide it into the appropriate fields.)
e. Given virtual address 25_{10} is located on virtual page 1, offset 9. Indicate exactly how this address would be translated to its corresponding physical address and how the data would be accessed. Include in your explanation how the TLB, Page Table, Cache and Memory are used.

Ans.

a. $2^3*2^4 = 2^7$, so there are 7 bits in a virtual address
b. $2^2*2^4 = 2^6$, so there are 6 bits in a virtual address
c. 18 = 001 0010 (where 001 is the page field and 0010 is the offset). Using the page table, and going to entry 1, we see that page 1 maps to frame 0, so the actual physical address would be 00 0010, or 2.
d. 54 = 11 0 110, where 11 is the tag, 0 is the set, and 110 is the offset. Therefore, this would map to Set 0 in cache. Once there, if the tag is found, the block is in cache. If not, it is a miss.
e. 25 = 001 1001, where 001 is the virtual page, and 1001 is the offset. (This maps to physical page 00 with offset 1001.) The TLB would first be checked to see if the pair (1,0) was present (virtual page 1, physical frame 0). If so, 00 can be substituted for 001, giving the actual physical address 00 1001. If the entry is not found in the TLB, the Page Table must be accessed, at which time the physical address 00 1001 is determined. At this point (regardless of whether we found the physical address using the TLB or the Page Table), the cache is checked to see if the block containing this physical address is currently cached. The memory address 001001 is divided up into 00 1 001, where 00 is the tag, 1 is the set, and 001 is the offset. Set 1 in cache is then checked for the tag 00 (it would be found as address 25 is in block D). The value in cache is used. (If it had not been found in cache, main memory would be accessed.)

---

15. Given a virtual memory system with a TLB, a cache, and a page table. Assume the following:

- A TLB hit requires 5ns
- A cache hit requires 12ns
- A memory reference requires 25ns
- A disk reference requires 200ms (this includes updating the page table, cache, and TLB)
- The TLB hit ratio is 90%
- The cache hit rate is 98%
- The page fault rate is .001%
- On a TLB or cache miss, the time required for access includes a TLB and/or cache update, but the access is NOT restarted
- On a page fault, the page is fetched from disk, all updates are performed but the access IS restarted
- All references are sequential (no overlap, nothing done in parallel)

For each of the following, indicate whether or not it is possible. If so, specify the time required for accessing the requested data.

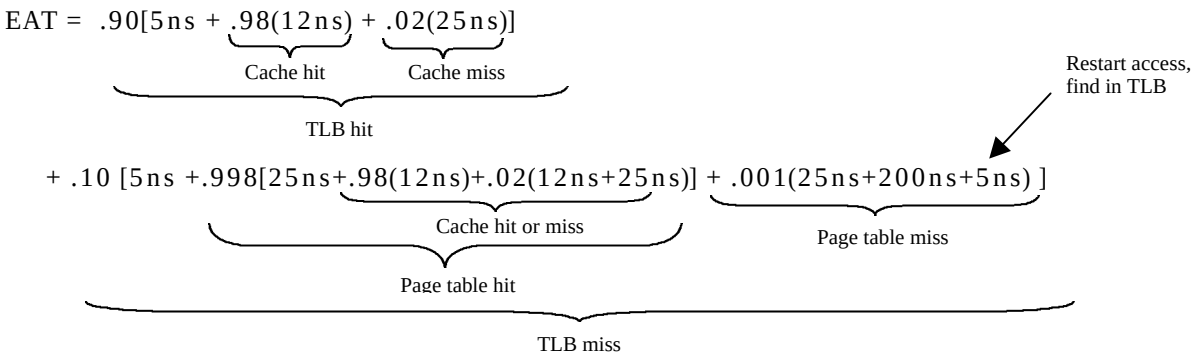
a. TLB hit, cache hit
b. TLB miss, page table hit, cache hit

c. TLB miss, page table hit, cache miss
d. TLB miss, page table miss, cache hit
e. TLB miss, page table miss

Write down the equation to calculate the effective access time.

Ans.

a. Possible, 5ns (TLB access) + 12ns (cache access)
b. Possible, 5ns (TLB access) + 25ns (page table reference) + 12ns (cache access)
c. Possible, 5ns (TLB access) + 25ns (page table reference) + 12ns (cache access) + 25ns (main memory reference)
d. Not possible (the cache value would be stale if the page no longer resides in page table)
e. Possible, 5ns (TLB access) + 25ns (page table reference) + 200ms (disk reference) + 5nx (TLB, since access is restarted) + 12ns (cache hit)



---

16. A system implements a paged virtual address space for each process using a one-level page table. The maximum size of virtual address space is 16MB. The page table for the running process includes the following valid entries (the → notation indicates that a virtual page maps to the given page frame, that is, it is located in that frame):

Virtual page 2 → page frame 4	Virtual page 4 → page frame 9
Virtual page 1 → page frame 2	Virtual page 3 → page frame 16
Virtual page 0 → page frame 1	

The page size is 1024 bytes and the maximum physical memory size of the machine is 2MB.

a. How many bits are required for each virtual address?
b. How many bits are required for each physical address?
c. What is the maximum number of entries in a page table?
d. To which physical address will the virtual address 1524_{10} translate?
e. Which virtual address will translate to physical address 1024_{10}?

Ans.

a. There are 16MB, or $2^4 x 2^{20}$ addresses, so we need 24 bits for a virtual address.
b. Main memory is 2MB, or $2^1 x 2^{20}$, so we need 21 bits for a physical address.
c. There are $2^{24} / 2^{10}$ pages in virtual memory, so the page table can have 2^{14} entries.
d. 1524 is on page 1 (page 0 contains addresses 0=1023, page 1 contains 1024 - 2047), located at offset 500. Page 1 maps to frame 2, so 1524 maps to physical address 2548. You can also find the binary equivalent of 1524 (000000000000010111110100) and

divide it into two pieces: the first 14 bits are the page, and the last 10 are the offset. Replace the first 14 bits (00000000000001) by (00000000000010), to get the physical address 000000000000100111110100, or 2548.

e. Physical address 1024 is at offset 0 in frame 1. Virtual page 0 maps to frame 1, so this is virtual address 0.

---

17. a. If you are a computer builder trying to make your system as price-competitive as possible, what features and organization would you select for its memory hierarchy?

b. If you are a computer buyer trying to get the best performance from a system, what features would you look for in its memory hierarchy?

Ans.

a. To be competitive, the system would need cache, but it could be the cheaper, non-associative cache (direct-mapped). To keep the price low, you light leave off level 2 cache entirely. There could also be a very small TLB. Cheaper main memory could be used, but putting a large hard drive would interest buyers.

b. You should look for a machine with two-level cache (level one being some type of associative memory). Note that a larger cache isn't necessarily better. Up to a certain point, more cache can help raise the hit ratio, but after a point, the hit ratio doesn't change much (so don't pay for what you don't need). A good-sized TLB will help with performance as well A large, fast main memory is a definite plus.

---

*18. Consider a system that has multiple processors where each processor has its own cache, but main memory is shared among all processors.

a. Which cache write policy would you use?

b. **The Cache Coherency Problem.** With regard to the system just described, what problems are caused if a processor has a copy of memory block *A* in its cache and a second processor, also having a copy of *A* in its cache, then updates main memory block *A*? Can you think of a way (perhaps more than one) of preventing this situation, or lessening its effects?

Ans.

a. Write through should be used to maintain consistency. If some processors have a value cached, and one processor changes that value, the other processors would not know about that value. With a write through cache (in addition to a broadcast "invalidate" message) the processors would not be using stale values.

b. As mentioned in the answer for part a, the problem is stale data. One way to solve this problem is to invalidate stale entries. A way to prevent the situation would be to require processors to specify whether the values were for writing or reading. Shared reading would be ok; however, when a processor wanted to write, it would have to either wait until the readers were done, or send an invalidate message to those readers. Exclusive access by writers would then be allowed.

---

*19. Pick a specific architecture (other than the one covered in this chapter). Do research to find out how your architecture approaches the concepts introduced in this chapter, as was done for Intel's Pentium.

Ans.

No answer given.
