(1991), Stallings (2001), Tanenbaum and Woodhull (1997), or Silberschatz, Galvin, and Gagne (2001) books. Hennessy and Patterson (1996) discuss issues involved with determining cache performance. For an online tutorial on memory technologies, see www.kingston.com/king/mg0.htm. George Mason University also has a set of workbenches on various computer topics. The workbench for virtual memory is located at cne.gmu.edu/workbenches/vmsim/vmsim.html.

## REFERENCES

Davis, W. *Operating Systems, A Systematic View*, 4th ed., Redwood City, CA: Benjamin/Cummings, 1992.

Flynn, I. M., & McHoes, A. M. *Understanding Operating Systems.* Pacific Grove, CA: Brooks/Cole, 1991.

Hamacher, V. C., Vranesic, Z. G., & Zaky, S. G. *Computer Organization*, 5th ed., New York: McGraw-Hill, 2002.

Hennessy, J. L., & Patterson, D. A. *Computer Architecture: A Quantitative Approach*, 2nd ed., San Francisco: Morgan Kaufmann, 1996.

Mano, Morris. *Digital Design*, 2nd ed., Upper Saddle River, NJ: Prentice Hall, 1991.

Silberschatz, A., Galvin, P., & Gagne, G. *Operating System Concepts*, 6th ed., Reading, MA: Addison-Wesley, 2001.

Stallings, W. *Computer Organization and Architecture*, 5th ed., New York: Macmillan Publishing Company, 2000.

Stallings, W. *Operating Systems*, 4th ed., New York: Macmillan Publishing Company, 2001.

Tanenbaum, A. *Structured Computer Organization*, 4th ed., Englewood Cliffs, NJ: Prentice Hall, 1999.

Tanenbaum, A., & Woodhull, S. *Operating Systems, Design and Implementation*, 2nd ed., Englewood Cliffs, NJ: Prentice Hall, 1997.

## REVIEW OF ESSENTIAL TERMS AND CONCEPTS

1. Which is faster, SRAM or DRAM?
2. What are the advantages of using DRAM for main memory?
3. Name three different applications where ROMs are often used.
4. Explain the concept of a memory hierarchy. Why did your authors choose to represent it as a pyramid?
5. Explain the concept of locality of reference and state its importance to memory systems.
6. What are the three forms of locality?
7. Give two noncomputer examples of the concept of cache.
8. Which of L1 or L2 cache is faster? Which is smaller? Why is it smaller?
9. Cache is accessed by its _____, whereas main memory is accessed by its _____.

**10.** What are the three fields in a direct mapped cache address? How are they used to access a word located in cache?

**11.** How does associative memory differ from regular memory? Which is more expensive and why?

**12.** Explain how fully associative cache is different from direct mapped cache.

**13.** Explain how set associative cache combines the ideas of direct and fully associative cache.

**14.** Direct mapped cache is a special case of set associative cache where the set size is 1. So fully associative cache is a special case of set associative cache where the set size is ___.

**15.** What are the three fields in a set associative cache address and how are they used to access a location in cache?

**16.** Explain the four cache replacement policies presented in this chapter.

**17.** Why is the optimal cache replacement policy important?

**18.** What is the worst-case cache behavior that can develop using LRU and FIFO cache replacement policies?

**19.** What, exactly, is effective access time (EAT)?

**20.** Explain how to derive an effective access time formula.

**21.** When does caching behave badly?

**22.** What is a dirty block?

**23.** Describe the advantages and disadvantages of the two cache write policies.

**24.** What is the difference between a virtual memory address and a physical memory address? Which is larger? Why?

**25.** What is the objective of paging?

**26.** Discuss the pros and cons of paging.

**27.** What is a page fault?

**28.** What causes internal fragmentation?

**29.** What are the components (fields) of a virtual address?

**30.** What is a TLB and how does it improve EAT?

**31.** What are the advantages and disadvantages of virtual memory?

**32.** When would a system ever need to page its page table?

**33.** What causes external fragmentation and how can it be fixed?

## 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, that is, what are the sizes of the tag, block, and word fields?
- **c)** To which cache block will the memory reference $0DB63_{16}$ map?

**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, that is, what are the sizes of the tag, block, and word fields?

**c)** To which cache block will the memory reference $000063FA_{16}$ map?

- **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, that is, what are the sizes of the tag and word fields?
  - **c)** To which cache block will the memory reference $F8C9_{16}$ map?

**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, that is, what are the sizes of the tag and word fields?

**c)** To which cache block will the memory reference $01D872_{16}$ map?

- **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.

**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 to 51 in main memory. You may leave the hit ratio in terms of a fraction.

**7.** Suppose a computer using set associative cache has $2^{16}$ words of main memory and 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, that is, 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?

**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, that is, 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?

***9.** Suppose we have a computer that uses a memory address word size of 8 bits. This computer has a 16-byte cache with 4 bytes per block. 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 here:


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 and the cache "contents" are simply the address stored at that cache location.)

| | 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?

**10.** A direct-mapped cache consists of eight blocks. Main memory contains 4K blocks of eight words each. Access time for the cache is 22ns and the time required to fill a cache slot from main memory is 300ns. (This time allows 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.

**11.** Consider a byte-addressable computer with 24-bit addresses, a cache capable of storing a total of 64KB 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

**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 |

◆ **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 |

***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 into blocks, where each block is represented by a letter. Two blocks equal one frame.