CS420/520 Homework: Memory Hierarchy

7.7 [10] Here is a series of address references given as word addresses: 1, 4, 8, 5, 20, 17, 19, 56, 9, 11, 4, 43, 5, 6, 9, 17. Assuming a direct-mapped cache with 16 one-word blocks that is initially empty, label each reference in the list as a hit or a miss and show the final contents of the cache.

7.8 [10] Using the series of references given in Exercise 7.7, show the hits and misses and final cache contents for a direct-mapped cache with four-word blocks and a total size of 16 words.

7.9 [10] Computer the total number of bits required to implement the cache in the following Figure 7.10. This number is different from the size of the cache, which usually refers to the number of bytes of data stored in the cache. The number of bits needed to implement the cache represents the total amount of memory needed for storing all of the data, tags, and valid bits.

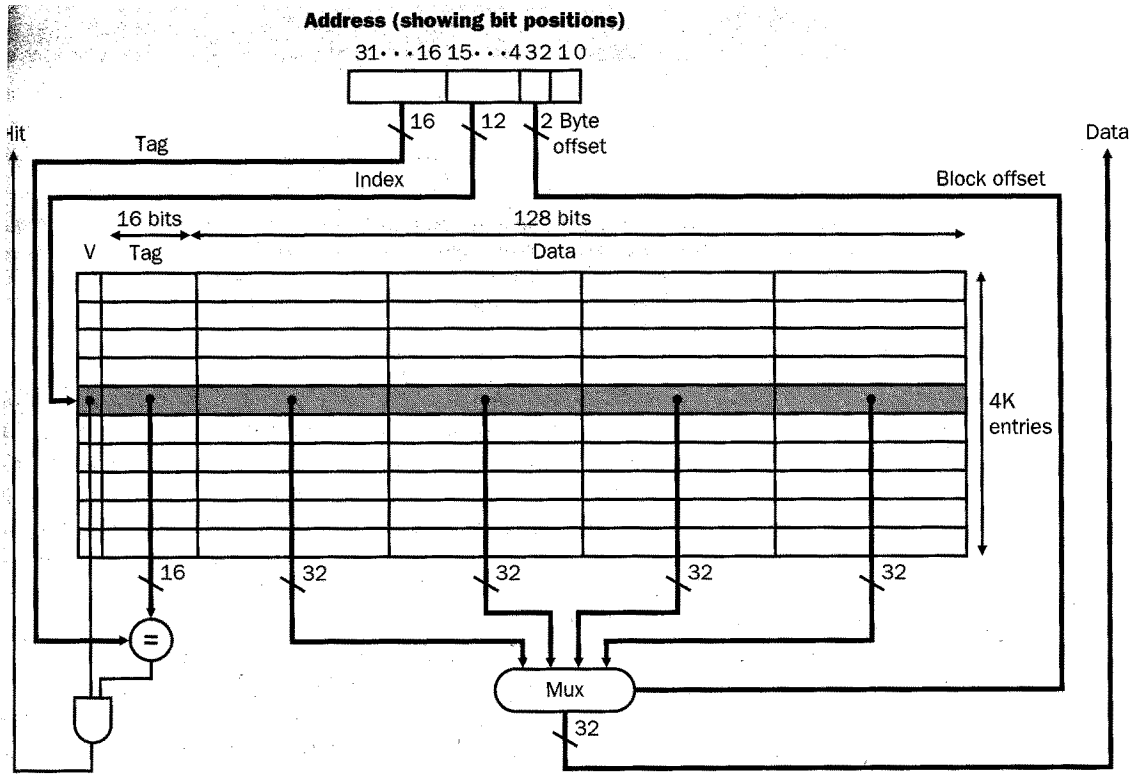

FIGURE 7.10 A 64-KB cache using four-word (16-byte) blocks. The tag field is 16 bits wide and the index field is 12 bits wide, while a 2-bit field (bits 3–2) is used to index the block and select the word from the block using a 4-to-1 multiplexor. In practice, the low-order bits of the address (bits 2 and 3 in this case) are used to enable only those RAMs that contain the desired word, eliminating the need for the multiplexor. Another way to eliminate the multiplexor is to have a large RAM for the data (with the tags stored separately) and use the block offset to supply 2 address bits for the RAM. The RAM must be 32 bits wide and have four times as many words as blocks in the cache.

7.11 [10] Consider a memory hierarchy using one of three organizations in Figure 7.11. The cache block size is 16 words. The bank width of organizations A, B, and C are one word, four words, and one word, respectively. The number of banks in organization C is four. It takes 1 clock cycle to send a bank address to the main memory. If the main memory latency for a new access is 10 cycles and the bus transfer time of a bank data is 1 cycle, what are the miss penalties for each of these three organizations?

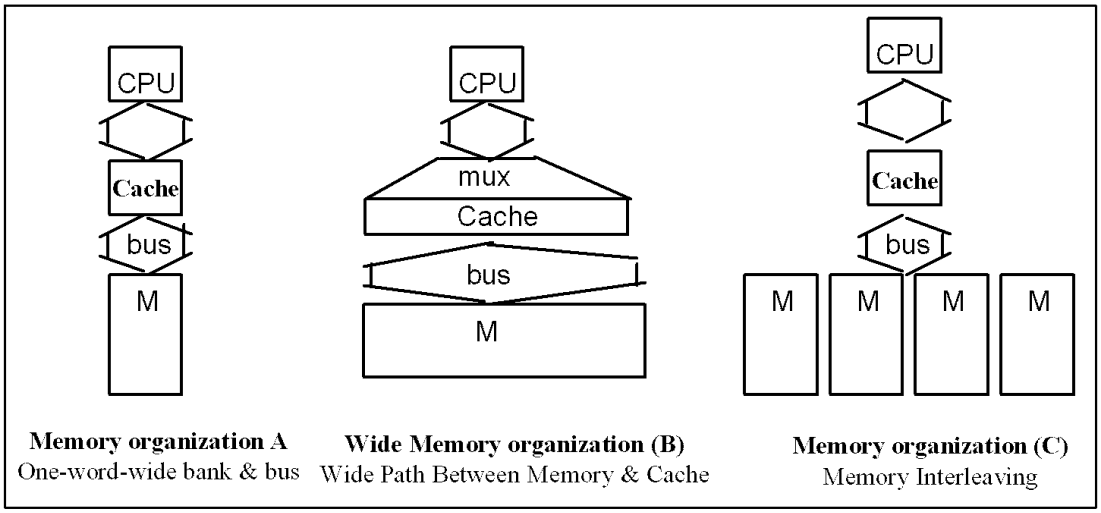

Figure 7.11 Memory Bus & Bank Organizations

7.12 [5] Suppose a processor with a 16-word block size has an effective miss rate per instruction of 0.5%. Assume that the CPI without cache misses is 1.2. Using the memory organizations described in Figure 7.11, how much faster is this processor when using the wide memory (B) than when using narrow (A) or interleaved memory organization (C)?

7.19 [10] Using the series of references given in Exercise 7.7, show the hits and misses and final cache contents for a two-way set-associative cache with *one-word* blocks and a total size of 16 words. Assume LRU replacement.

7.20 [10] Using the series of references given in Exercise 7.7, show the hits and misses and final cache contents for a two-way set-associative cache with *two-word* blocks and a total size of 16 words. Assume LRU replacement.

7.21 [10] Using the series of references given in Exercise 7.7, show the hits and misses and final cache contents for a fully associative cache with *one-word* blocks and a total size of 16 words. Assume LRU replacement.

7.22 [10] Using the series of references given in Exercise 7.7, show the hits and misses and final cache contents for a fully associative cache with *four-word* blocks and a total size of 16 words. Assume LRU replacement.

7.23 [5] Associativity usually improves the miss ratio, but not always. Please create a short series of addresses references for which a two-way set-associative cache with LRU replacement would experience more misses than a direct-mapped cache of the same size.

7.24 [15] Suppose a computer's address size is *k* bits (using byte addressing, for example, *k* = 32 in 32 MIPS instruction set architectures), the cache size is *S* bytes, the block size is *B* bytes, and the cache is *A*-Way set-associative. Assume that *B* is a power of two, so $B = 2^b$. Figure out what the following quantities are in terms of *S*, *B*, *A*, *b*, and *k*:

(a) the number of sets in the cache;
(b) the number of index bits in the address;
(c) the number of total bits needed to implement the cache (similar to Exercise 7.9)

7.27 [15] Consider three machines with different cache configurations:

Cache 1: direct-mapped with one-word blocks
Cache 2: direct-mapped with four-word blocks
Cache 3: two-way set associative with four-word blocks

The following miss rate measurements have been made:

Cache 1: Instruction miss rate is 4%; data miss rate is 8%.
Cache 2: Instruction miss rate is 2%; data miss rate is 5%.
Cache 3: Instruction miss rate is 2%; data miss rate is 4%.

For these machines, one-half of the instructions contain a data reference. Assume that the cache miss penalty is 6 + (Block size in terms of the number of words per block). E.g., if the block size is 2 words, the cache miss penalty is 6 + 2 = 8. The CPI for this workload was measured on a machine with cache 1 and was found to be 2.0. Determine which machine spends the highest percentage of cycles on cache misses.

7.33 [5] Give three techniques/options that can reduce the miss rate.
7.34 [5] Give two techniques/options that can reduce the conflict miss.
7.35 [5] Give two techniques/options that can reduce the miss penalty.
7.36 [5] Give three techniques/options that can reduce the average access time of a memory hierarchy.