Week 6: Assignment Solutions

1. Which of the following is true for a memory hierarchy?
    a. It tries to bridge the processor-memory speed gap.
    b. The speed of the memory level closest to the processor has the highest speed.
    c. The capacity of the memory level farthest away from the processor is the largest.
    d. It is based on the principle of locality of reference.

    All the answers are true.
    This follows from the basic definition of memory hierarchy design. The fastest and smallest memory module is closest to the processor. The overall access time of the memory system is close to the access time of the fastest memory level. This is achieved by exploiting locality of reference.

2. Which of the following statements are false?
    a. Temporal locality arises because of loops in a program.
    b. Spatial locality arises because of loops in a program.
    c. Temporal locality arises because of sequential instruction execution.
    d. Spatial locality arises because of sequential instruction execution.

    Correct answers are (b) and (c).
    Temporal locality says that an accessed word will be accessed again in the near future, and this is due to the presence of loops.
    Spatial locality says that if a work is accessed, then words in the neighborhood will also be accessed in the near future. This happens due to sequential program execution.

3. Assume that a read request takes 50 nsec on a cache miss and 5 nsec on a cache hit. While running a program, it is observed that 80% of the processor's read requests result in a cache hit. The average read access time is …….. nsec.
    Correct answer is 14.
    Average read time = 0.80 x 5 + (1 – 0.80) x 50 = 14 nsec

4. In a two-level cache system, the access times of L1 and L2 caches are 1 and 8 clock cycles respectively. The miss penalty from L2 cache to main memory is 18 clock cycles. The miss rate of L1 cache is twice that of L2. The average memory access time of the cache system is 2 cycles. The miss rates of L1 and L2 caches respectively are:
    a. 0.130 and 0.065
    b. 0.056 and 0.111
    c. 0.0892 and 0.1784
    d. 0.1784 and 0.0892

    Correct answer is (a).
    Let the miss rate of L2 cache be x.
    So, miss rate of L1 cache = 2x.
    Thus, average memory access time
    AMAT = (1-2x).1 + 2x. [(1-x).8 + x.18] = 2 (given)

Solving, we get

x = 0.065

5. The memory access time is 1 nsec for a read operation with a hit in cache, 5 nsec for a read operation with a miss in cache, 2 nsec for a write operation with a hit in cache, and 10 nsec for a write operation with a miss in cache. The execution of a sequence of instructions involves 100 instruction fetch operations, 60 memory operand read operations, and 40 memory operand write operations. The cache hit ratio is 0.9. The average memory access time (in nanoseconds) in executing the sequence of instructions is:
   a. 1.26
   b. 1.68
   c. 2.46
   d. 4.52

   Correct answer is (b).
   Total number of read = 100 + 60 = 160
   Total number of write = 40.
   So, fraction of reads = 160 / (160 + 40) = 0.8
   And, fraction of writes = 40 / (160 + 40) = 0.2
   Average access time = 0.8 (0.9 x 1 + 0.1 x 5) + 0.2 (0.9 x 2 + 0.1 x 10)
   = 1.68

6. Consider a two-level memory hierarchy with separate instruction and data caches in level 1, and main memory in level 2. The clock cycle time in 1 ns. The miss penalty is 20 clock cycles for both read and write. 2% of the instructions are not found in I-cache, and 10% of data references not found in D-cache. 25% of the total memory accesses are for data, and cache access time (including hit detection) is 1 clock cycle. The average access time of the memory hierarchy will be …….. nanoseconds.
   Average access time = 0.75 (0.98 x 1 + 0.02 x 20) + 0.25 (0.90 x 1 + 0.10 x 20) = 1.76 ns

7. Consider a direct-mapped cache with 64 blocks and a block size of 16 bytes. Byte address 1200 will map to block number …….. of the cache.
   Correct answer is 11.
   We first find out the memory block number that byte address 1200 belongs to. Since the size of a block is 16 bytes.
   Byte address 0 to 15: block 0
   Byte address 16 to 31: block 1
   Byte address 32 to 47: block 2, and so on.
   Byte address 1200 will belong to block number: floor(1200/16) = 75.
   For direct mapped cache,
   Cache block no. = (Memory block no.) MOD (No. of cache blocks)
   = 75 MOD 64 = 11.

8. A cache memory system with capacity of N words and block size of B words is to be designed. If it is designed as a direct mapped cache, the

length of the TAG field is 10 bits. If it is designed as a 16-way set associative cache, the length of the TAG field will be ……. bits.
Correct answer is 14.
For 16-way set associative cache, 4 more bits will be required in the TAG as compared to direct mapping, since $2^4 = 16$.

9. Which of the following statements is true:
    a. The implementation of direct mapping technique for cache requires expensive hardware to carry out division.
    b. The set associative mapping requires associative memory for implementation.
    c. A main memory block can be placed in any of the sets in set associative mapping.
    d. None of the above.

    Correct answer is (b).
    Direct mapping is the easiest to implement. Both fully associative and set associative mappings require an associative memory. (c) is false as in set associative mapping, a main memory block can be mapped to any of the blocks in the selected set.

10. Which of the following statements is false for cache misses?
    a. Compulsory miss can be reduced by decreasing the cache block size.
    b. Capacity miss can be reduced by decreasing the total size of the cache.
    c. Conflict miss can be reduced by decreasing the value of cache associativity.
    d. Compulsory miss can be reduced by prefetching cache blocks.

    Correct answers are (a), (b) and (c).
    This follows from the definitions of compulsory and capacity misses.

11. How can the cache miss rate be reduced?
    a. By using larger block size
    b. By using larger cache size
    c. By reducing the cache associativity
    d. None of the above

    Correct answers are (a) and (b).
    If the cache block size is increased, number of cache misses will be reduced in general. Similar will be the case for larger cache memories, where more number of cache blocks can be stored. However, if the associativity is reduced, we are reducing the choice for cache block placement that can result in an increase in the number of misses.

12. Suppose that in 1000 memory references there are 40 misses in L1 cache and 10 misses in L2 cache. If the miss penalty of L2 is 200 clock cycles, hit time of L1 is 1 clock cycle, and hit time of L2 is 15 clock cycles, the average memory access time will be …….. clock cycles.
    L1 hit ratio = (1000 – 40) / 1000 = 0.96
    L2 hit ratio = (1000 – 10) / 1000 = 0.99

$$\text{Average access time} = 0.96 \times 1 + 0.04\,[0.99 \times 15 + 0.01 \times 200] = 1.634$$