Quiz Questions for Module 7

1. Assume that each atomic operation in a DRAM system has a total latency of 100ns. What is the maximal throughput we can get for atomic operations on the same global memory variable?

    a. 100G atomic operations per second
    b. 1G atomic operations per second
    c. 0.01G atomic operations per second
    d. 0.0001G atomic operations per second

    Answer: (C)
    Explanation: No other atomic operation can touch the same variable for the entire duration of 100ns. The maximal rate is 1/100n = 0.01G

2. For a processor that supports atomic operations in L2 cache, assume that each atomic operation takes 4ns to complete in L2 cache and 100ns to complete in DRAM. Assume that 90% of the atomic operations hit in L2 cache. What is the approximate throughput for atomic operations on the same global memory variable?

    a. 0.225G atomic operations per second
    b. 2.75G atomic operations per second
    c. 0.0735G atomic operations per second
    d. 100G atomic operations per second

    Answer: (C)
    The average latency is 4ns * 90% + 100ns *10% = 13.6ns. The average throughput is approximately 1/13.6 = 0.0735G atomic operations per second

3. In question 1, assume that a kernel performs 5 floating-point operations per atomic operation. What is the maximal floating-point throughput of the kernel execution as limited by the throughput of the atomic operations?

    a. 500 GFLOPS
    b. 5 GFLOPS
    c. 0.05 GFLOPS
    d. 0.0005 GFLOPS

    Answer: (C)
    The maximal is 5 operations every 100ns
    5 floating-point operations every 13.6 ns, approximately 5/13.6 = 0.368 GFLOPS

4. In Question 1, assume that we privatize the global memory variable into shared memory variables in the kernel and the shared memory access latency is 1ns. All original global memory atomic operations are converted into shared memory atomic operation. For simplicity, assume that the additional global memory atomic operations for accumulating privatized variable into the global variable adds 10% to the total execution time. Assume that a kernel performs 5 floating-point operations per atomic operation. What is the maximal floating-point throughput of the kernel execution as limited by the throughput of the atomic operations?

    a. 4500 GFLOPS
    b. 45 GFLOPS
    c. 4.5 GFLOPS
    d. 0.45 GFLOPS

    Answer: (C)
    The effective throughput without the final accumulation to the global variable is 5/1ns = 5 GFLOPS. Since the time is stretched by 10%, the final effective throughput is approximately 5/1.1 = 4.5 GFLOPS

5. To perform an atomic add operation to add the value of an integer variable Partial to a global memory integer variable Total. Which one of the following statements should be used?

    a. atomicAdd(Total, 1);
    b. atomicAdd(&Total, &Partial);
    c. atomicAdd(Total, &Partial);
    d. atomicAdd(&Total, Partial);

    Answer: (D)
    Explanation: The first argument should be a pointer to the variable to be updated and the second argument should be the variable whose value is to be added to the global variable.