8 GPUs and SIMD [40 points]

We define the *SIMD utilization* of a program that runs on a GPU as the fraction of SIMD lanes that are kept busy with *active threads* during the run of a program. As we saw in lecture and practice exercises, the SIMD utilization of a program is computed across the *complete run* of the program.

The following code segment is run on a GPU. A warp in the GPU consists of 64 threads, and there are 64 SIMD lanes in the GPU. Each thread executes **a single iteration** of the shown loop. Assume that the data values of the arrays `A` and `B` are already in vector registers so there are no loads and stores in this program. Both `A` and `B` are arrays of integers. (Hint Notice that there are 6 instructions in each thread.)

```
for (i = 0; i < 4096; i++) {
    if (B[i] < 8888) {      // Instruction 1
        A[i] = A[i] * C[i]; // Instruction 2
        A[i] = A[i] + B[i]  // Instruction 3
        C[i] = B[i] + 1;    // Instruction 4
    }
    if (B[i] > 8888) {      // Instruction 5
        A[i] = A[i] * B[i]; // Instruction 6
    }
}
```

Please answer the following four questions.

(a) [2 points] How many warps does it take to execute this program?

Warps = (Number of threads) / (Number of threads per warp) Number of threads = 2^{12} (i.e., one thread per loop iteration) Number of threads per warp = 64 = 2^6 (given) Warps = $2^{12}/2^6 = 2^6$

(b) [10 points] When we measure the SIMD utilization for this program with one input set, we find that it is 134/320. What can you say about arrays A,B, and C? Be precise. (Hint: Look at the "if" branch).

A. Nothing.
B. 2 in every 64 consecutive elements of B are less than 8888, the rest are exactly 8888.
C. Nothing.

(c) [10 points] What needs to be true about array `B` to achieve 100% utilization? Show your work. Be precise and complete. (Hint: The warp scheduler does not issue instructions where no threads are active).

Every 64 consecutive elements of B are either:
(1) equal to 8888,
(2) less than 8888,
(3) greater than 8888.

(d) [8 points] What is the minimum possible SIMD utilization of this program?

132/384.

(e) [10 points] What needs to be true about array `B` to achieve the minimum possible SIMD utilization? Show your work. (Please cover all cases in your answer.)

1 in every 64 of B's elements are greater than 8888, and 1 in every 64 of B's elements are less than 8888, and the rest of the elements are 8888.