6 GPUs and SIMD

We define the *SIMD utilization* of a program run 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. Each thread executes **a single iteration** of the shown loop. Assume that the data values of the arrays A, B, and C are already in vector registers so there are no loads and stores in this program. (Hint: Notice that there are 6 instructions in each thread.) A warp in the GPU consists of 64 threads, and there are 64 SIMD lanes in the GPU. Please assume that all values in array B have magnitudes less than 10 (i.e., $|B[i]| < 10$, for all i).

```
for (i = 0; i < 1024; i++) {
    A[i] = B[i] * B[i];
    if (A[i] > 0) {
        C[i] = A[i] * B[i];
        if (C[i] < 0) {
            A[i] = A[i] + 1;
        }
        A[i] = A[i] - 2;
    }
}
```

Please answer the following five questions.

(a) [5 points] How many warps does it take to execute this program?

Warps = (Number of threads) / (Number of threads per warp)
Number of threads = 2^{10} (i.e., one thread per loop iteration).
Number of threads per warp = 64 = 2^6 (given).
Warps = $2^{10}/2^6 = 2^4$

(b) [5 points] What is the maximum possible SIMD utilization of this program?

100%

(c) [20 points] Please describe what needs to be true about array B to reach the maximum possible SIMD utilization asked in part (b). (Please cover all cases in your answer)

B: For every 64 consecutive elements: every value is 0, every value is positive, or every value is negative. Must give all three of these.

(d) [10 points] What is the minimum possible SIMD utilization of this program?

Answer: 132/384

Explanation: The first two lines must be executed by every thread in a warp (64/64 utilization for each line). The minimum utilization results when a single thread from each warp passes both conditions on lines 2 and 4, and every other thread fails to meet the condition on line 2. The thread per warp that meets both conditions, executes lines 3-6 resulting in a SIMD utilization of 1/64 for each line. The minimum SIMD utilization sums to $(64 * 2 + 1 * 4)/(64 * 6) = 132/384$

(e) [20 points] Please describe what needs to be true about array B to reach the minimum possible SIMD utilization asked in part (d). (Please cover all cases in your answer)

B: Exactly 1 of every 64 consecutive elements must be negative. The rest must be zero. This is the only case that this holds true.