9 GPUs and SIMD [45 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 32 threads, and there are 32 SIMD lanes in the 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. Both B and C are arrays of integers and each integer in these arrays has an absolute value of less than 10 (i.e., $|B[i]| < 10$ and $|C[i]| < 10$, for all i).

```
for (i = 0; i < 1024; i++) {
    A[i] = B[i] * C[i];     // instruction 1
    if (/* Condition */) { // instruction 2
        // instruction 3
        // instruction 4
        .
        .
        .
        // instruction k + 2
    }
    C[i] = C[i] - 1;       // instruction k + 3
}
```

Please answer the following four questions.

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

32 Warps.

Explanation:
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 = 32 = 2^5 (given) Warps = $2^{10}/2^5 = 2^5$

(b) [20 points] Assume that the condition for the if statement is (i % 16 == 0). What is the number of instructions (k) in the body of the conditional block given a SIMD utilization of $\frac{11}{32}$? Assume that there are **no** control flow instructions in the body of the if statement. Show your work.

7 Instructions.

Explanation:
Two of the 32 threads go inside of the conditional block. This pattern is homogeneous through all warps.

$\frac{2\times(3+k)+30\times3}{32\times(3+k)} = \frac{11}{32} \rightarrow k = 7$ instructions.

(c) [20 points] Assume that the condition for the `if` statement is `(i % 16 == 0 && i < 512)`. What is the number of instructions (`k`) in the body of the conditional block given a SIMD utilization of $\frac{5}{8}$? Assume that there are **no** control flow instructions in the body of the `if` statement. Show your work.

4 Instructions.

Explanation:
Two of the 32 threads **only within the first 16 warps** go inside of the conditional block. In the rest of the warps no thread goes inside of the conditional block.

$\frac{16(32\times(3))+16(2\times(k+3)+30\times3)}{16(32\times(3+k))+16(32\times3)} = \frac{5}{8} \rightarrow k = 4$ instructions.