Problem 5: GPUs (42 pts)

We define the **SIMD utilization** of a program running on a GPU as the fraction of SIMD lanes that are kept busy with active threads during the run of a program.

The following code segment is running 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 5 instructions in each thread as labled below.) A warp in the GPU consists of 64 threads, and there are 64 SIMD lanes in the GPU.

```
for (i = 0; i < 16384; i++) {
  if (A[i] > 0) {           //Instruction 1
    A[i] = A[i] * C[i];     //Instruction 2
    B[i] = A[i] + B[i];     //Instruction 3
    C[i] = B[i] + 1;        //Instruction 4
    D[i] = C[i] * B[i];     //Instruction 5
  }
}
```

A) [2 pts] How many warps does it take to execute this program?

$16384/64 = 256$

B) [10 pts] As shown below, assume array A has a repetitive pattern which has 32 ones followed by 96 zeros repetitively and array B has a different repetitive pattern which has 64 zeros followed by 64 ones repetitively. What is the SIMD utilization of this program?

A:	1	1	...29 1s...	1	0	0	...93 0s...	0	...32 1s...	96 0s...	...
B:	0	0	...61 0s...	0	1	1	...61 1s...	1	...64 0s	...64 1s...	...

When a warp is working on a segment of array A that has 64 0s, none of the threads in the warp will take the branch, which yields no branch divergence of the warp. Hence, the SIMD utilization of this particular input set is $(64 + 64 + 32 * 4)/(64 + 64 * 5)$ = 66.7%

Continued . . .

C) [10 pts] Is it possible for this program to yield a SIMD utilization of 25%?

CIRCLE ONE: **YES** **NO**

If YES, what should be true about arrays A and B for the SIMD utilization to be 25%? Be precise and show your work. If NO, explain why not.

Yes. For example, if only 4 elements in **every 64** elements of A are positive, we can have a SIMD utilization of $(64 + 4 * 4)/(64 * 5) = 25\%$.

D) [10 pts] Is it possible for this program to yield a SIMD utilization of 20%?

CIRCLE ONE: **YES** **NO**

If YES, what should be true about arrays A and B for the SIMD utilization to be 20%? Be precise and show your work. If NO, explain why not.

No. The smallest SIMD utilization one can get is to have one and only one element in every 64 elements of A to be positive, which yields a minimal SIMD utilization of $(64 + 1 * 4)/(64 * 5) = 21.25\%$, which is still greater than 20%.

E) [10 pts] During an execution with a particular input array A, which has exactly 24 positive elements in every 64 elements, Hongyi finds that the SIMD utilization of the program is 50%. Based on this observation, Hongyi claims that any input array that has an **average** of 24 out of 64 elements positive would yield a 50% SIMD utilization. Is Hongyi correct?

CIRCLE ONE: **YES** **NO**

If YES, show your work. If NO, provide a counterexample.

Hongyi is incorrect. If A has a repetitive pattern of 48 contiguous 1s followed by 80 contiguous 0s, in which case 37.5% of the elements are positive on average, then the SIMD utilization of the program will be 83.3% rather than 50%.