ECE 786 Advanced Computer Architecture: Data Parallel Processors
Spring 2023 Huiyang Zhou HW 2

1

In this exercise, we analyze how the instructions of a Vector-Add kernel function run on SIMT pipelines. The following is the instruction sequence in the PTX format:

```
ld.global.f32   %f1, [%r14+0]    // f1 = mem[r14+0]
ld.global.f32   %f2, [%r16+0]    // f2 = mem[r16+0]
add.f32         %f3, %f1, %f2    // f3 = f1 + f2
st.global.f32   [%r18+0], %f3    // mem[r18+0] = f3
```

Using the pipeline timing diagram to show how these instructions are executed on one SM. The following assumptions are made:

1. The SM has 16 SPs
2. The load/store latency is 5 cycles (i.e., 5 MEM stages) with AGEN included in the 5 stages, and the add latency is 3 cycles (i.e., 3 EX stages)
3. The SM can run 6 warps concurrently and the warp scheduling policy is round robin
4. There is no scoreboard to support more than 1 instruction from the same warp to be issued to the pipeline.
5. IF and ID are 2 cycles each
6. WB can handle 16 threads at a time.

How many cycles does it take to execute all the 6 warps, counting from when the first instruction enters the IF stage to when the last instruction enters the WB stage?

It takes x cycles to execute all the 6 warps.