# Lab 0 Report

Quan Fan
862099688
qfan005@ucr.edu

## Questions

Input size is 1,000,000 and 1M % 512 = 64 so all warps are fully active at the beginning.

1. For the naive reduction kernel, how many steps execute without divergence? How many steps execute with divergence?
Answer:
Given the BLOCKSIZE is 512 (a.k.a. 2^9), so 9 steps will be executed. Only the first step is without divergence since all the threads follow the same instruction to compute the sum. All other 8 steps are executed with divergence because 1 thread out of 2 will have nothing to do after the first step.

2. For the optimized reduction kernel, how many steps execute without divergence? How many steps execute with divergence?
Answer:
Before the number of effective threads becomes smaller than a warp, there is no divergence at all. Each warp contains 32 threads therefore first 4 steps are without divergence and last 5 steps have divergence.

3. Which kernel performed better? (for both real GPUs and GPGPU-Sim)
Answer:

Table 1 Execution Time (seconds)

| | Naïve | Optimized |
|---|---|---|
| Bender | 0.000281 | 0.000154 |
| GPGPU-Sim | 113 | 89 |

The optimized version is better in performance.

4. How does the warp occupancy distribution compare between the two Reduction implementations?
Answer:

Table 2 Warp Occupancy Distribution

| | Stall | W0_Idle | W0_SB | W1 | W2 | W3 | W4 | W5 | W6 |
|---|---|---|---|---|---|---|---|---|---|
| Naïve | 146546 | 56748 | 315050 | 369306 | 187584 | 0 | 187584 | 0 | 0 |
| Opt-ed | 93304 | 97096 | 378426 | 13678 | 7816 | 0 | 7816 | 0 | 0 |
| | W7 | W8 | W9 | W10 | W11 | W12 | W13 | W14 | W15 |
| Naïve | 0 | 187584 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Opt-ed | 0 | 7816 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | W16 | W17 | W18 | W19 | W20 | W21 | W22 | W23 | W24 |
| Naïve | 187584 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Opt-ed | 7816 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 2 Warp Occupancy Distribution (Cont.)

| | W25 | W26 | W27 | W28 | W29 | W30 | W31 | W32 | |
|---|---|---|---|---|---|---|---|---|---|
| Naïve | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2125882 | |
| Opt-ed | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2024274 | |

This table can reveal the reason of performance gap between two solution. The optimized version has less stalled warps and the warps that are not fully occupied. It means that the utilization rate is higher in optimized version.

5. Why do GPGPUs suffer from warp divergence?

Answer:

The nature of SIMT/SIMD model is that all data are processed under the same code. If there are two possible paths to go through, which means divergence, they will be serialized. The serialization undermines the performance.