Lab 2 Reduction

Dekang Zeng 862188852

1. For the naive reduction kernel, how many steps execute without divergence? How many steps execute with divergence?

The block size is 512. So we need 9 steps, cause 2^9 = 512.

There is only first step without divergence. The other 8 steps execute with divergence.

2. For the optimized reduction kernel, how many steps execute without divergence? How many steps execute with divergence?

The first 4 steps execute without divergence, the last 5 steps execute with divergence.

3. Which kernel performed better? Use profiling statistics to support your claim.

For optimized version:

```
kernel_name = _Z9reductionPfS_j
kernel_launch_uid = 1
gpu_sim_cycle = 88707
gpu_sim_insn = 62024030
gpu_ipc =     699.2011
gpu_tot_sim_cycle = 88707
gpu_tot_sim_insn = 62024030
gpu_tot_ipc =     699.2011
gpu_tot_issued_cta = 0
gpu_stall_dramfull = 3890
gpu_stall_icnt2sh    = 38325
gpu_total_sim_rate=424822
```

For naive version:

```
kernel_launch_uid = 1
gpu_sim_cycle = 126806
gpu_sim_insn = 71024154
gpu_ipc =     560.1009
gpu_tot_sim_cycle = 126806
gpu_tot_sim_insn = 71024154
gpu_tot_ipc =     560.1009
gpu_tot_issued_cta = 0
gpu_stall_dramfull = 2686
gpu_stall_icnt2sh    = 11024
gpu_total_sim_rate=441143
```

So, the cycle of naive on GPGPU-Sim is 126806, but the cycle of optimized version is 88707. So optimized version performed better.

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

For naive version:

Warp Occupancy Distribution:
Stall:146546 W0_Idle:56748 W0_Scoreboard:315050
W1:369306 W2:187584 W3:0 W4:187584 W5:0 W6:0 W7:0 W8:187584
W9:0 W10:0 W11:0 W12:0 W13:0 W14:0 W15:0 W16:187584 W17:0 W18:0
W19:0 W20:0 W21:0 W22:0 W23:0 W24:0 W25:0 W26:0 W27:0 W28:0 W29:0
W30:0 W31:0 W32:2125882

For optimized version:

Warp Occupancy Distribution:
Stall:93304 W0_Idle:97096 W0_Scoreboard:378426
W1:13678 W2:7816 W3:0 W4:7816 W5:0 W6:0 W7:0 W8:7816
W9:0 W10:0 W11:0 W12:0 W13:0 W14:0 W15:0 W16:7816 W17:0 W18:0
W19:0 W20:0 W21:0 W22:0 W23:0 W24:0 W25:0 W26:0 W27:0 W28:0 W29:0
W30:0 W31:0 W32:2024274

5. Why do GPGPUs suffer from warp divergence?

Because warp divergence has significant performance impact on GPUs as it can bring more conflicts to shared resources.