

1. Only the first step executes without divergence.  
Others 9 steps execute with divergence.
2. The first 5 steps execute without divergence.  
The final 5 steps execute with divergence
3. Naïve reduction:

```
kernel_name = _Z14naïvereductomPTS_
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=755576

===== Core cache stats =====
```

Optimized reduction:

```
kernel_name = _Z14naïvereductomPTS_
kernel_launch_uid = 1
gpu_sim_cycle = 89134
gpu_sim_insn = 62024030
gpu_ipc = 695.8516
gpu_tot_sim_cycle = 89134
gpu_tot_sim_insn = 62024030
gpu_tot_ipc = 695.8516
gpu_tot_issued_cta = 0
gpu_stall_dramfull = 5337
gpu_stall_icnt2sh = 37314
gpu_total_sim_rate=826987

===== Core cache stats =====
```

number of cycles for optimized reduction is less than naïve reduction, thus optimized reduction performed better.

4. Naïve reduction:

```
gpu_reg_bank_conflict_stalls = 0
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
: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
1:0 W22:0 W23:0 W24:0 W25:0 W26:0 W27:0 W28:0 W29:0 W30:0 W31:0 W32:2125882
traffic_breakdown_coretoemem[CONST_ACC_R] = 120 {8:15,}
```

Optimized reduction:

```
gpgpu_stall_shd_mem[l_mem_id][wb_rsrv_fail] = 0
gpu_reg_bank_conflict_stalls = 0
Warp Occupancy Distribution:
Stall:92060 W0_Idle:98762 W0_Scoreboard:385670 W1:13678 W2:7816 W3:0 W4:7816 W5:0 W6:0 W7:0 W8:7816 W9:0W
10:0 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
traffic_breakdown_coretoemem[CONST_ACC_R] = 120 {8:15,}
traffic_breakdown_coretoemem[GLOBAL_ACC_R] = 250000 {8:31250,}
```

5. Programs need to follow SIMT fashion of execution and execution of different instructions on different threads lead to different instructions executing in a warp.