Naive Reduction Output

Total cycles statistics

```
gpu_sim_cycle = 182537
gpu_sim_insn = 82511310
gpu_ipc =      452.0251
gpu_tot_sim_cycle = 182537
gpu_tot_sim_insn = 82511310
gpu_tot_ipc =      452.0251
gpu_tot_issued_cta = 0
gpu_stall_dramfull = 2176
gpu_stall_icnt2sh    = 9177
gpu_total_sim_rate=182144
```

Warp Distribution

```
Warp Occupancy Distribution:
Stall:232023    W0_Idle:66273   W0_Scoreboard:370408

W1:793184      W2:406362      W3:0    W4:406362      W5:0    W6:0    W7:0    W8:406362      W9:0    W10:0

W11:0   W12:0   W13:0   W14:0   W15:0   W16:406362      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   W32:2328958
```

```
gpgpu_simulation_time = 0 days, 0 hrs, 7 min, 33 sec (453 sec)
gpgpu_simulation_rate = 182144 (inst/sec)
gpgpu_simulation_rate = 402 (cycle/sec)
451.745239 s
Copying data from device to host...0.000625 s
```

Optimized reduction output

Total cycles statistics

```
gpu_sim_cycle = 124046
gpu_sim_insn = 99531179
```

```
gpu_ipc =     802.3731
gpu_tot_sim_cycle = 124046
gpu_tot_sim_insn = 99531179
gpu_tot_ipc =     802.3731
gpu_tot_issued_cta = 0
gpu_stall_dramfull = 2618
gpu_stall_icnt2sh    = 13970
gpu_total_sim_rate=242758
```

Warp Distribution

```
Warp Occupancy Distribution:
Stall:153948    W0_Idle:37611

W0_Scoreboard:234291    W1:16609       W2:10747       W3:0    W4:10747       W5:0    W6:0    W7:0

W8:10747       W9:0    W10:0    W11:0    W12:0    W13:0    W14:0    W15:0    W16:10747       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
W32:3193519
```

Overall time statistics

```
gpgpu_simulation_time = 0 days, 0 hrs, 6 min, 50 sec (410 sec)
gpgpu_simulation_rate = 242758 (inst/sec)
gpgpu_simulation_rate = 302 (cycle/sec)
Running Optimized Reduction
408.617401 s
Copying data from device to host...0.000713 s
```

We can see in the optimized output one that only w32 threads are being executed till stride 512 - stride 32, after that warp divergence starts and we see warp of size 16-1 being active.

Hence total divergence in 5 iterations (16, 8, 4, 2, 1)

We see in naive reduction that after the first iteration divergence starts.

Hence total divergence in 9 iterations (256, 128, 64, 32, 16, 8, 4, 2, 1)

We can clearly see from the time statistics that optimized reduction is taking less time than naive optimization. Because it is simulating more instruction per sec.

GPGPU architecture makes them suffer from warp divergence as it is SIMD and change of control flow will make threads in a single warp execute diff behaviour.