CUDA gpuFindMax Kernel Execution Time Prediction Report

This report applies the CUDA kernel performance prediction template to the kernel:
gpuFindMax

The kernel is evaluated on five GPUs:
- GTX TITAN Black
- GTX TITAN X
- NVIDIA TITAN V
- GeForce RTX 2080 Ti
- GeForce RTX 4070

1. Kernel Work Analysis

Code structure:
Each thread processes:
start = threadWorkLoad * threadIdx.x; end = start + threadWorkLoad; for (i = start+1; i < end; i++) { if (i >= n) break; if (data[i] > data[localMaxIndex]) localMaxIndex = i; } Launch configuration:
- Grid = (5,5) → 25 blocks
- Block = (32,32) → 1024 threads per block
- Total threads = 25 × 1024 = 25,600
- n = 25,600
- threadWorkLoad = 10

Each thread performs 9 iterations of the loop, so:
N_iters_total = 25,600 × 9 = 230,400

FLOPs per iteration:
1 float comparison → approx 1 FLOP-equivalent

FLOPs_total ≈ 230,400 × 1 = 2.304 × 10■

Reduction phase:
Additional ≈ 24,800 FLOPs across all blocks

Total FLOPs ≈ 2.55 × 10■

Memory Traffic:
Each loop iteration loads one float (4 bytes)

Bytes_loop = 230,400 × 4 = 921,600 bytes
Reduction adds ≈ 198,400 bytes
Final stores ≈ 100 bytes

Total bytes ≈ 1.12 × 10■ bytes

2. GPU Specifications Used

GPU	Peak FP32 (FLOPs/s)	Bandwidth (bytes/s)
GTX TITAN Black	5.12e12	3.36e11
GTX TITAN X	6.14e12	3.365e11
TITAN V	1.49e13	6.528e11
RTX 2080 Ti	1.345e13	6.16e11
RTX 4070	2.9e13	5.04e11

3. Time Estimates

We compute:
- t_compute = FLOPs_total / Peak_FP32
- t_mem = Bytes_total / Bandwidth
- t_body = max(t_compute, t_mem)
- t_total = t_body + 5 microseconds launch latency

GPU	t_compute (µs)	t_mem (µs)	t_body (µs)	t_total (µs)
GTX TITAN Black	0.050	3.33	3.33	≈ 8.33
GTX TITAN X	0.042	3.33	3.33	≈ 8.33
TITAN V	0.017	1.72	1.72	≈ 6.72
RTX 2080 Ti	0.019	1.82	1.82	≈ 6.82
RTX 4070	0.009	2.22	2.22	≈ 7.22

4. Conclusion

The gpuFindMax kernel performs very little computational work (~2.55×10■ FLOPs) and moderate memory traffic (~1.12 MB). Execution time is dominated by memory bandwidth and CUDA kernel launch latency (~5 µs).

Expected time ranges:
- Older GPUs: 8.3 µs
- TITAN V and 2080 Ti: ~6.7–6.8 µs
- RTX 4070: ~7.2 µs

These values match typical behavior for small, reduction-based CUDA kernels.