CUDA makeHVector Kernel Execution Time Prediction Report

This report analyzes the CUDA kernel makeHVector under the launch configuration Grid = (5,5) and Block = (32,32), estimating performance on five NVIDIA GPUs.

1. Workload Analysis

rows = 25,600
1024 threads/block × 25 blocks = 25,600 threads total.

Loop 1:
50 iterations per thread → 1,279,200 iterations
Bytes = 10.2336M, FLOPs = 2.5584M

Reduction:
24,800 FLOPs

Loop 2:
1,280,000 iterations → 10.24M bytes, 1.28M FLOPs

Total:
FLOPs_total ≈ 3.86M
Bytes_total ≈ 20.47M bytes

2. GPU Specs

GPU	Peak FP32	Bandwidth
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

Compute t_compute = FLOPs / PeakFP32
Memory t_mem = Bytes / Bandwidth
t_total = max(t_compute, t_mem) + 5 µs

GPU	t_compute (µs)	t_mem (µs)	t_total (µs)
GTX TITAN Black	0.75	60.93	≈ 65.9
GTX TITAN X	0.63	60.84	≈ 65.8

TITAN V	0.26	31.36	≈ 36.4
RTX 2080 Ti	0.29	33.24	≈ 38.2
RTX 4070	0.13	40.62	≈ 45.6

4. Conclusion

makeHVector performs ~3.86M FLOPs and moves ~20MB of data per launch.
It is strongly memory-bound.

Final predicted times:
- TITAN Black/X: ~66 µs
- TITAN V: ~36 µs
- RTX 2080 Ti: ~38 µs
- RTX 4070: ~46 µs