CUDA UpdateHHNorms Kernel Execution Time Prediction Report

This report applies the CUDA kernel performance prediction template to the kernel: UpdateHHNorms, launched with Grid = (5,5) and Block = (32,32).

We estimate the expected per-launch execution time on five NVIDIA GPUs:
• GeForce GTX TITAN Black
• GeForce GTX TITAN X
• NVIDIA TITAN V
• GeForce RTX 2080 Ti
• GeForce RTX 4070

```
__global__ void UpdateHHNorms(int cols, float *dV, float *dNorms) {
    int colIndex = threadIdx.x + blockIdx.x * blockDim.x;
    if (colIndex < cols) {
        float val = dV[colIndex];
        dNorms[colIndex] -= val * val;
    }
}
```

1. Workload Analysis

Launch configuration in main():
• Grid = (5,5) → 25 blocks
• Block = (32,32) → 1024 threads per block
• cols = 5 × 32 × 5 × 32 = 25,600

The kernel uses only blockIdx.x and threadIdx.x to compute colIndex:
colIndex = threadIdx.x + blockIdx.x * blockDim.x;

With gridDim.x = 5 and blockDim.x = 32, colIndex ranges from 0 to 159.
Since cols = 25,600, the condition (colIndex < cols) is always true.

Thus, every launched thread executes the body once:
Total threads = 25 blocks × 1024 threads/block = 25,600
→ N_iters_total = 25,600 iterations

2. FLOP Count

Per iteration:
val = dV[colIndex]; // load
dNorms[colIndex] -= val * val; // one multiply, one add

Approximate FLOPs per iteration = 2 (mul + add).

Total FLOPs:
F_total = 25,600 × 2 = 51,200 ≈ 5.12 × 10^4 FLOPs.

3. Memory Traffic

Per iteration global memory:
• Load dV[colIndex]: 4 bytes
• Load dNorms[colIndex]: 4 bytes
• Store dNorms[colIndex]: 4 bytes
→ 12 bytes per iteration.

Total bytes:
B_total = 25,600 × 12 = 307,200 bytes ≈ 3.07 × 10^5 bytes (~0.29 MiB).

FLOPs are tiny; the kernel is dominated by memory + launch overhead.

4. GPU Specifications Used

Approximate FP32 peak performance and memory bandwidth:

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

5. 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 µs (kernel launch overhead).

For F_total = 5.12 × 10^4 FLOPs and B_total ≈ 3.07 × 10^5 bytes, we obtain:

GPU	t_compute (µs)	t_mem (µs)	t_body (µs)	t_total (µs)
GTX TITAN Black	0.01	0.91	0.91	≈ 5.91
GTX TITAN X	0.01	0.91	0.91	≈ 5.91
TITAN V	0.003	0.47	0.47	≈ 5.47
RTX 2080 Ti	0.004	0.50	0.50	≈ 5.50
RTX 4070	0.002	0.61	0.61	≈ 5.61

6. Conclusion

UpdateHHNorms performs a very small amount of computation (~5.12 × 10^4 FLOPs) and moves only about 0.29 MiB of data per launch.

As a result, execution time is dominated by kernel launch latency (~5 µs). All five GPUs are capable of completing the kernel body in under 1 µs, so the predicted per-launch runtime lies in a narrow band

around 5.5–6.0 µs:

• GTX TITAN Black / TITAN X: ≈ 5.9 µs
• TITAN V: ≈ 5.5 µs
• RTX 2080 Ti: ≈ 5.5 µs
• RTX 4070: ≈ 5.6 µs

These values are consistent with expectations for small, memory-light CUDA kernels where the fixed launch overhead dominates.