CUDA Kernel Execution Time Analysis on GTX TITAN Black

1. Work per Kernel Launch

- Grid: (5,5) = 25 blocks
- Block: (32,32) = 1024 threads
- Each block computes one Euclidean distance.
- k = 100 dimensions.

Work:

- threadIdx.x covers offsets 0–99 once.
- 32 threadIdx.y values repeat work → 32× redundancy.
- Per threadIdx.y row: 200 float loads (100 from A, 100 from B).
- Per block: 200 * 32 = 6400 float loads.
- 25 blocks: 160,000 float loads = 640,000 bytes ≈ 0.61 MB.
- FLOPs: 200 * 32 * 3 ≈ 19,200 FLOPs/block → 480,000 FLOPs total.

2. GTX TITAN Black Capabilities

- Peak compute: 5.1 TFLOPS (single precision)
- Peak memory bandwidth: 336 GB/s

3. Theoretical Lower Bounds

Compute-bound:
480k FLOPs / (5.1e12 FLOPs/s) ≈ 0.094 µs

Memory-bound:
640k bytes / (336e9 bytes/s) ≈ 1.9 µs

4. Expected Real Performance

- Kernel extremely small → dominated by launch overhead.
- Kepler launch overhead = several microseconds.
- With only 25 blocks, GPU underutilized.

5. Final Prediction

Average per-kernel time:
≈ 10–20 µs
Best estimate ≈ 15 µs