CUDA gpuSD Kernel Execution Time Prediction Report

This report applies the CUDA kernel performance prediction template to the kernel:
gpuSD, 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 gpuSD(const float * vectsA, size_t na, const float * vectsB, size_t nb,
                      size_t dim, const float * means, const float * numPairs, float * sds)
{
    size_t offset, stride, tx = threadIdx.x, bx = blockIdx.x, by = blockIdx.y;
    float a, b, termA, termB;
    __shared__ float meanA, meanB, n, threadSumsA[NUMTHREADS], threadSumsB[NUMTHREADS];
    ...
}
```

1. Parallelism and Workload

Launch configuration:

- Grid = (5,5) → 25 blocks
- Block = (32,32) → 1024 threads/block
- NUMTHREADS = 16 (actual workers along threadIdx.x)

Within each block, only tx = 0..15 threads perform useful work. These 16 threads cooperatively process dim = 100 elements of vectsA and vectsB for each (bx,by) pair. All 25 blocks are active since na = nb = 5.

For each valid element:

- Load a, b
- Compute termA = (a - meanA), termB = (b - meanB)
- Accumulate squared deviations (termA^2, termB^2)

We approximate ≈ 8 FLOP-equivalents per valid element (subtract, multiply, add, and isnan logic).

2. FLOP Count

For each block:

- 100 loop iterations × 8 FLOPs ≈ 800 FLOPs
- Reduction over NUMTHREADS = 16 entries ≈ 30 FLOPs
- Final sqrtf operations ≈ 6 FLOPs

Total per block: ≈ 836 FLOPs
Across 25 blocks:
F_total ≈ 836 × 25 = 2.09 × 10■ FLOPs.

3. Memory Traffic

Per valid element:
- Load a and b → 8 bytes.

Per block:
- Loop: 100 × 8 = 800 bytes
- meanA, meanB, n loading: 12 bytes
- Final write of two sds values: 8 bytes

Total per block ≈ 820 bytes.
Across 25 blocks:
B_total ≈ 820 × 25 = 2.05 × 10■ bytes (~0.02 MiB).

Memory traffic is tiny; kernel runtime is dominated by launch overhead.

4. GPU Specs 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

5. Time Estimates

Using:
- F_total ≈ 2.09 × 10■ FLOPs
- B_total ≈ 2.05 × 10■ bytes

Compute per-GPU:
t_compute = F_total / Peak_FP32
t_mem = B_total / Bandwidth
t_body = max(t_compute, t_mem)
t_total ≈ t_body + 5 μs launch overhead.

GPU	t_compute (μs)	t_mem (μs)	t_body (μs)	t_total (μs)
GTX TITAN Black	0.0041	0.0610	0.0610	≈ 5.061
GTX TITAN X	0.0034	0.0609	0.0609	≈ 5.061
TITAN V	0.0014	0.0314	0.0314	≈ 5.031
RTX 2080 Ti	0.0016	0.0333	0.0333	≈ 5.033
RTX 4070	0.0007	0.0407	0.0407	≈ 5.041

6. Conclusion

The gpuSD kernel performs only ~2 × 10■ FLOPs and moves ~0.02 MiB of data per launch. Both compute and memory-bound times are under 0.1 µs on all GPUs tested.

Thus, execution time is dominated by the fixed CUDA launch overhead (~5 µs).

Predicted per-launch times:

- GTX TITAN Black: ≈ 5.06 µs
- GTX TITAN X: ≈ 5.06 µs
- TITAN V: ≈ 5.03 µs
- RTX 2080 Ti: ≈ 5.03 µs
- RTX 4070: ≈ 5.04 µs

These are consistent with extremely small kernels where arithmetic and memory traffic are negligible compared to the kernel launch cost.