Analysis Report

three_nonDivergentKernel(float*, int)

Duration	52.17954 ms (52,179,539 ns)
Grid Size	[2097152,1,1]
Block Size	[256,1,1]
Registers/Thread	9
Shared Memory/Block	1 KiB
Shared Memory Executed	8 KiB
Shared Memory Bank Size	4 B

[0] Quadro P4000

GPU UUID	GPU-55bcc7c2-f527-bf35-99ec-b54f8392185d
Compute Capability	6.1
Max. Threads per Block	1024
Max. Threads per Multiprocessor	2048
Max. Shared Memory per Block	48 KiB
Max. Shared Memory per Multiprocessor	96 KiB
Max. Registers per Block	65536
Max. Registers per Multiprocessor	65536
Max. Grid Dimensions	[2147483647, 65535, 65535]
Max. Block Dimensions	[1024, 1024, 64]
Max. Warps per Multiprocessor	64
Max. Blocks per Multiprocessor	32
Half Precision FLOP/s	41.44 GigaFLOP/s
Single Precision FLOP/s	5.304 TeraFLOP/s
Double Precision FLOP/s	165.76 GigaFLOP/s
Number of Multiprocessors	14
Multiprocessor Clock Rate	1.48 GHz
Concurrent Kernel	true
Max IPC	6
Threads per Warp	32
Global Memory Bandwidth	243.328 GB/s
Global Memory Size	7.929 GiB
Constant Memory Size	64 KiB
L2 Cache Size	2 MiB
Memcpy Engines	2
PCIe Generation	3
PCIe Link Rate	8 Gbit/s
PCIe Link Width	16

1. Compute, Bandwidth, or Latency Bound

Use the button below to collect the profiling data needed for this analysis.

2. Instruction and Memory Latency

Instruction and memory latency limit the performance of a kernel when the GPU does not have enough work to keep busy. Use the button below to collect the profiling data needed for this analysis.

3. Compute Resources

GPU compute resources limit the performance of a kernel when those resources are insufficient or poorly utilized. Use the button below to collect the profiling data needed for this analysis.

4. Memory Bandwidth

Memory bandwidth limits the performance of a kernel when one or more memories in the GPU cannot provide data at the rate requested by the kernel. Use the button below to collect the profiling data needed for this analysis.