Page: Details | Result: 4 - 133 - prefixSumScanKernel | Add Baseline | Apply Rules | Occupancy Calculator | Save as PDF

	Result	Time	Cycles	Regs	GPU	SM Frequency	CC	Process
Current	133 - prefixSumScanKernel (4, 6000, 1)x(1024, 1, 1)	1.82 msecond	٢٬٥٣٩٬٣١٦	١٦	0 - NVIDIA GeForce RTX 3090	1.39 cycle/nsecond	8.6	[258893] task2

GPU Speed Of Light Throughput

High-level overview of the throughput for compute and memory resources of the GPU. For each unit, the throughput reports the achieved percentage of utilization with respect to the theoretical maximum. Breakdowns show the throughput for each individual sub-metric of Compute and Memory to clearly identify the highest contributor.

Compute (SM) Throughput [%]	25.52	Duration [msecond]	1.82
Memory Throughput [%]	23.64	Elapsed Cycles [cycle]	2,539,316
L1/TEX Cache Throughput [%]	24.55	SM Active Cycles [cycle]	2,445,121.15
L2 Cache Throughput [%]	10.15	SM Frequency [cycle/nsecond]	1.39
DRAM Throughput [%]	23.01	DRAM Frequency [cycle/nsecond]	9.49

Latency Issue This kernel exhibits low compute throughput and memory bandwidth utilization relative to the peak performance of this device. Achieved compute throughput and/or memory bandwidth below 60.0% of peak typically indicate latency issues. Look at Scheduler Statistics and Warp State Statistics for potential reasons.

Launch Statistics

Summary of the configuration used to launch the kernel. The launch configuration defines the size of the kernel grid, the division of the grid into blocks, and the GPU resources needed to execute the kernel. Choosing an efficient launch configuration maximizes device utilization.

Grid Size	24,000	Function Cache Configuration	cudaFuncCachePreferNone
Registers Per Thread [register/thread]	16	Static Shared Memory Per Block [Kbyte/block]	8.19
Block Size	1,024	Dynamic Shared Memory Per Block [byte/block]	0
Threads [thread]	24,576,000	Driver Shared Memory Per Block [Kbyte/block]	1.02
Waves Per SM	292.68	Shared Memory Configuration Size [Kbyte]	16.38

Occupancy

Occupancy is the ratio of the number of active warps per multiprocessor to the maximum number of possible active warps. Another way to view occupancy is the percentage of the hardware's ability to process warps that is actively in use. Higher occupancy does not always result in higher performance, however, low occupancy always reduces the ability to hide latencies, resulting in overall performance degradation. Large discrepancies between the theoretical and the achieved occupancy during execution typically indicates highly imbalanced workloads.

Theoretical Occupancy [%]	66.67	Block Limit Registers [block]	4
Theoretical Active Warps per SM [warp]	32	Block Limit Shared Mem [block]	1
Achieved Occupancy [%]	63.49	Block Limit Warps [block]	1
Achieved Active Warps Per SM [warp]	30.48	Block Limit SM [block]	16

Occupancy Limiters This kernel's theoretical occupancy (66.7%) is limited by the required amount of shared memory This kernel's theoretical occupancy (66.7%) is limited by the number of warps within each block See the CUDA Best Practices Guide for more details on optimizing occupancy.