

Current

584 - multiplyMatricesKernel

Size

(625, 625, 1)x(16, 16, 1)

Time

3,89 s

Cycles

5.248.579.698

GPU

0 - NVIDIA GeForce RTX 4060 Laptop GPU

SM Frequency

1,35 Ghz

Process

[19344] main_v1.exe

Attributes

Summary

Details

Source

Context

Comments

Raw

Session

Compare

Tools

View

Export

GPU Speed Of Light Throughput

GPU Throughput Chart

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 [%]	98,66	Duration [s]	3,89
Memory Throughput [%]	98,66	Elapsed Cycles [cycle]	5.248.579.698
L1/TEX Cache Throughput [%]	99,94	SM Active Cycles [cycle]	5.212.422.105,25
L2 Cache Throughput [%]	15,88	SM Frequency [Ghz]	1,35
DRAM Throughput [%]	37,28	DRAM Frequency [Ghz]	5,57

High Throughput

This workload is utilizing greater than 80.0% of the available compute or memory performance of the device. To further improve performance, work will likely need to be shifted from the most utilized to another unit. Start by analyzing workloads in the [Compute Workload Analysis](#) section.

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	390.625	Function Cache Configuration	CachePreferNone
Registers Per Thread [register/thread]	40	Static Shared Memory Per Block [byte/block]	0
Block Size	256	Dynamic Shared Memory Per Block [byte/block]	0
Threads [thread]	100.000.000	Driver Shared Memory Per Block [Kbyte/block]	1,02
Waves Per SM	2.712,67	Shared Memory Configuration Size [Kbyte]	16,38
Uses Green Context	0	Stack Size	1.024
# SMs [SM]	24	# TPCs	12
Enabled TPC IDs	all	-	-

Occupancy

% Occupancy Graphs

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 [%]	100	Block Limit Registers [block]	6
Theoretical Active Warps per SM [warp]	48	Block Limit Shared Mem [block]	16
Achieved Occupancy [%]	99,95	Block Limit Warps [block]	6
Achieved Active Warps Per SM [warp]	47,98	Block Limit SM [block]	24

GPU and Memory Workload Distribution

Analysis of workload distribution in active cycles of SM, SMP, SMSP, L1 & L2 caches, and DRAM

Average SM Active Cycles [cycle]	5.212.422.105,25	Average L1 Active Cycles [cycle]	5.212.422.105,25
Average L2 Active Cycles [cycle]	4.958.545.545,25	Average SMSP Active Cycles [cycle]	5.212.415.730,85
Average DRAM Active Cycles [cycle]	8.071.171.256	Total SM Elapsed Cycles [cycle]	126.721.100.792
Total L1 Elapsed Cycles [cycle]	126.721.100.792	Total L2 Elapsed Cycles [cycle]	77.790.823.832
Total SMSP Elapsed Cycles [cycle]	506.884.403.168	Total DRAM Elapsed Cycles [cycle]	86.601.040.384

Missing [Roofline](#) and [Memory Charts](#)? Profile again with the *detailed metric set* to collect all necessary metrics. Also consider collecting all available sections with the *full* metric set.