	Result	Size	Time	Cycles	GPU	SM Frequency	Process	Attributes
Current	584 - multiplyMatricesSharedMemoryKernel	(313, 313, 1)x(32, 32, 1)	2,88 s	4.624.815.568	0 - NVIDIA GeForce RTX 4060 Laptop GPU	1,61 Ghz	[10708] main_v2.exe	

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 [%]	77,89	Duration [s]	2,88
Memory Throughput [%]	77,89	Elapsed Cycles [cycle]	4.624.815.568
L1/TEX Cache Throughput [%]	78,24	SM Active Cycles [cycle]	4.599.648.130,71
L2 Cache Throughput [%]	6,01	SM Frequency [Ghz]	1,61
DRAM Throughput [%]	25,56	DRAM Frequency [Ghz]	5,55

Balanced Throughput Compute and Memory are well-balanced: To reduce runtime, both computation and memory traffic must be reduced. Check both the Compute Workload Analysis and Memory Workload Analysis sections.

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	97.969	Function Cache Configuration	CachePreferNone
Registers Per Thread [register/thread]	36	Static Shared Memory Per Block [Kbyte/block]	8,19
Block Size	1.024	Dynamic Shared Memory Per Block [byte/block]	0
Threads [thread]	100.320.256	Driver Shared Memory Per Block [Kbyte/block]	1,02
Waves Per SM	4.082,04	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 [%]	66,67	Block Limit Registers [block]	1
Theoretical Active Warps per SM [warp]	32	Block Limit Shared Mem [block]	1
Achieved Occupancy [%]	66,49	Block Limit Warps [block]	1
Achieved Active Warps Per SM [warp]	31,92	Block Limit SM [block]	24

Theoretical Occupancy
Est. Local Speedup: 33.33%

The 8.00 theoretical warps per scheduler this kernel can issue according to its occupancy are below the hardware maximum of 12. This kernel's theoretical occupancy (66.7%) is limited by the number of required registers, and the number of warps within each block.

Key Performance Indicators

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]	4.599.648.130,71	Average L1 Active Cycles [cycle]	4.599.648.130,71
Average L2 Active Cycles [cycle]	3.609.121.939,50	Average SMSP Active Cycles [cycle]	4.599.541.353,09
Average DRAM Active Cycles [cycle]	4.082.801.148	Total SM Elapsed Cycles [cycle]	110.886.408.480
Total L1 Elapsed Cycles [cycle]	110.886.408.480	Total L2 Elapsed Cycles [cycle]	68.054.594.000
Total SMSP Elapsed Cycles [cycle]	443.545.633.920	Total DRAM Elapsed Cycles [cycle]	63.896.032.256

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.