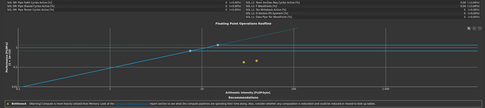
Start: 2026-05-17 17:24:01 +0200 (UTC+02:00)
Launch: /home/ubuntu/Downloads/Nsight Compute-2026.1.1/host/linux-desktop-glibc_2_11_3-x64/ncu --config-file off --export /home/ubuntu/Downloads/Nsight Compute-2026.1.1/host/linux-desktop-glibc_2_11_3-x64/profile --force-overwrite --replay-mode application --app-replay-match grid --cache-control all --clock-control base --set full --target-processes all /home/ubuntu/Downloads/Nsight Compute-2026.1.1/host/linux-desktop-glibc_2_11_3-x64/../target/linux-desktop-glibc_2_11_3-x64/QtWebEngineProcess
Target: 1 process terminated
Profiling: 1 process profiled
Report: /home/ubuntu/Downloads/Nsight Compute-2026.1.1/host/linux-desktop-glibc_2_11_3-x64/profile.ncu-rep
Time: 256.57 (+overhead) Cycles: 1,013,992,482 Regs: 72 SMs: 76.81 (+overhead) Cycles: 1,013,297,992 Regs: 72 DRAM: 0.00 (+overhead) Cycles: 1,013,297,992 Regs: 72 GPU: 76.81 (+overhead) Cycles: 1,013,297,992 Regs: 72 Process: 2103960 (profile) 0 / 0

The kernel compute-bound in most cases.

GPU Speed Of Light [1]

High-level overview of the utilization for compute and memory resources of the GPU. For each unit, the Speed Of Light (SOL) reports the achieved percentage of utilization with respect to the theoretical maximum. High-level overview of the utilization for compute and memory resources of the GPU presented as a roofline chart.

SM Frequency	1.34 cycle/nsec	Duration	755.97 msec
Elapsed Cycles	1,013,992,482	Memory Frequency	6.81 cycle/nsec
SM Active Cycles	1,013,297,992	L1/TEX Cache Throughput	76.81 %
L2 Cache Throughput	76.81 %	DRAM Throughput	0.00 %



SOL SM Breakdown

SOL SM Issue Active [%]	76.81
SOL SM Issue Inactive [%]	23.19
SOL SM Peak Fma Cycles Active [%]	20.00
SOL SM Peak Fma Cycles Elapsed [%]	20.00
SOL SM Peak Fma Cycles Active [%]	20.00
SOL SM Peak Fma Cycles Elapsed [%]	20.00
SOL SM Max Pipe Active Cycles [%]	76.81
SOL SM Max Pipe Elapsed Cycles [%]	76.81
SOL SM Fma Instructions [%]	20.00
SOL SM Fma Instructions [%]	20.00
SOL SM Fma Instructions Peak [%]	20.00
SOL SM Fma Instructions Peak [%]	20.00
SOL SM Fma Instructions Active [%]	20.00
SOL SM Fma Instructions Active [%]	20.00
SOL SM Fma Instructions Peak [%]	20.00
SOL SM Fma Instructions Peak [%]	20.00
SOL SM Fma Instructions Active [%]	20.00
SOL SM Fma Instructions Active [%]	20.00

SOL Memory Breakdown

SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81
SOL L1/TEX Cache Throughput [%]	76.81

Floating-Point Operations Roofline

Arithmetic Intensity [FLOP/byte] Recommendations

Workload Memory/Compute is more heavily utilized than Memory: 1.00 s

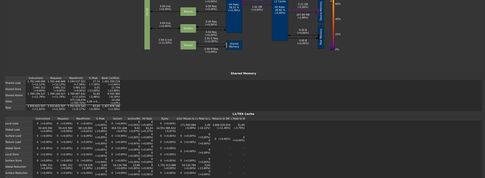
Roofline Analysis The rate of peak floating-point operations (FLOP/s) is the maximum that can be achieved by the kernel on the device. Achieved performance can be compared against this rofline chart.

Compute Workload Analysis

Detailed analysis of the compute resources of the streaming multiprocessors (SM), including the achieved instructions per clock (IPC) and the utilization of each available pipeline. Pipelines with very high utilization might limit the overall performance.

Executed Ipc Active [inst/cycle]	76.81
Executed Ipc Elapsed [inst/cycle]	76.81
Issue Slots Busy [%]	76.81
Issued Ipc Active [inst/cycle]	1.00

Pipe Utilization



Pipe Utilization Recommendations

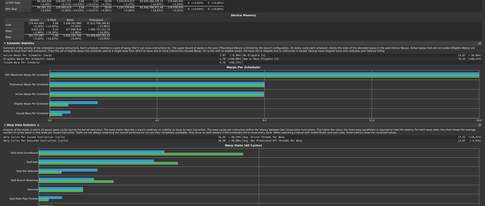
Pipe Utilization No pipeline is over-utilized.

Memory Workload Analysis

Detailed analysis of the memory resources of the GPU. Memory can become a limiting factor for the overall kernel performance when fully utilizing the involved hardware units (Mem Busy), exhausting the available communication bandwidth between those units (Max Bandwidth), or by reaching the maximum throughput of issuing memory instructions (Mem Pipe Busy). Detailed chart of the memory units. Detailed tables with data for each memory unit.

Memory Throughput [Gbyte/second]	0.00	L1/TEX Hit Rate [%]	0.00
L2 Hit Rate [%]	0.00	Mem Busy [%]	76.81
Max Bandwidth [%]	0.00	L2 Compression Success Rate [%]	0.00

Memory Chart



Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
Shared Store	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
Shared Atomic	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
Other	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
Total	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)

L1/TEX Cache

	Instructions	Requests	Wavefronts	% Peak	Sectors	Sectors/Req	Hit Rate	Sectors to L2	Read to L2	Returns to SM	Clients to SM
Local Load	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	1.28	0.84	0	0
Local Store	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	1.28	0.84	0	0
Global Load	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	-
Texture Load	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	-
Global Store	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	-	-
Local Store	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	-	-	-
Surface Store	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	-	-	-
Global Reduction	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	-	-	-
Surface Reduction	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	-	-	-
Global Atomic Red.	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	see above	see above	-
Surface Atomic Red.	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	-	-	-
Surface Atomic Op.	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	see above	see above	-
Loads	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	-
Stores	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	-	-
Total	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)

L2 Cache

	Requests	Sectors	Sectors/Req	Hit Rate	Sectors	Throughput	Hit Rate	Sectors to SM	Returns to SM
L2 Cache	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
L2 Read	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
L2 Write	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
L2 Atomic (All)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
L2 Atomic (Reduction)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
L2 Evict First	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
L2 Evict Last	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
L2 Lookup Hit	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)
L2 Lookup Miss	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)

Device Memory

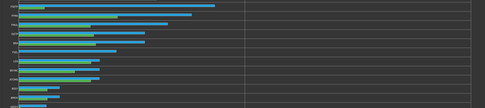
	Sectors	% Peak	Byte	Throughput
Read	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second
Write	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second
Total	0 (=0.00%)	0 (=0.00%)	0 (=0.00%)	0 byte/second

Scheduler Statistics

Summary of the activity of the schedulers issuing instructions. Each scheduler maintains a pool of warps that it can issue instructions for. The upper bound of warps in the pool (Theoretical Warps) is limited by the launch configuration. On every cycle each scheduler checks the state of the allocated warps in the pool (Active Warps). Active warps that are not stalled (Eligible Warps) are ready to issue their next instruction. From the set of eligible warps the scheduler selects a single warp from which to issue one or more instructions (Issued Warps). On cycles with no eligible warps, the issue slot is skipped and no instruction is issued. Having many skipped issue slots indicates poor latency hiding.

One or More Eligible [%]	76.81
Issued Warp Per Scheduler	1.00
No Eligible [%]	23.19
Active Warps Per Scheduler	11.98
Eligible Warps Per Scheduler	1.00

Warps Per Scheduler

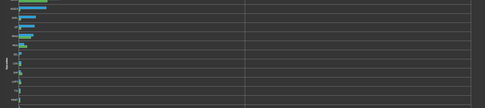


Warp State Statistics

Analysis of the states in which all warps spent cycles during the kernel execution. The warp states describe a warp's readiness or inability to issue its next instruction. The value is calculated by dividing the sum of cycles spent in a given state by the number of active cycles in which at least one warp was active. Stalled warp cycles are the inverse of issue cycles. The chart shows the average number of cycles spent in each state for all executed instructions. Stalls can be reduced by improving the memory access pattern, reducing the number of synchronization points, or by increasing the amount of parallel work that can be executed while waiting for data.

Cycles Per Issued Instruction	23.19
Cycles Per Issue Active	23.19
Cycles Per Issue Eligible	76.81
Cycles Per Issue No Eligible	76.81

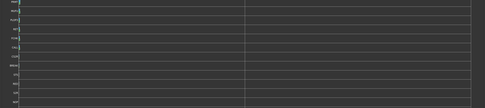
Warp Per Scheduler



Warp State (Instructions)

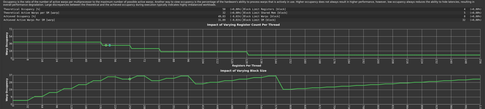
Analysis of the states in which all warps spent cycles during the kernel execution. The warp states describe a warp's readiness or inability to issue its next instruction. The value is calculated by dividing the sum of cycles spent in a given state by the number of active cycles in which at least one warp was active. Stalled warp cycles are the inverse of issue cycles. The chart shows the average number of cycles spent in each state for all executed instructions. Stalls can be reduced by improving the memory access pattern, reducing the number of synchronization points, or by increasing the amount of parallel work that can be executed while waiting for data.

Warp State (All Cycles)



Warp State (Instructions)

Cycles per Instruction



Scheduler Statistics Recommendations

Thread Divergence Warps are executed in lanes. Each lane groups of 32 threads. Optimal execution throughput is achieved if all threads in a warp execute the same instruction. The chart shows the distribution of the number of active threads in a warp per cycle. The number of active threads is limited by the launch configuration. On every cycle each scheduler checks the state of the allocated warps in the pool. Active warps that are not stalled (Eligible Warps) are ready to issue their next instruction. From the set of eligible warps the scheduler selects a single warp from which to issue one or more instructions (Issued Warps). On cycles with no eligible warps, the issue slot is skipped and no instruction is issued. Having many skipped issue slots indicates poor latency hiding. A detailed analysis of the sampled warp stalls breaks down the reasons for latency issues.

Avg. Active Threads Per Warp	32
Avg. Not Predicated Off Threads Per Warp	32

Off Stall 'Barrier' Warp stall analysis for 'Barrier' issues

Off Stall 'Branch Resolving' Warp stall analysis for 'Branch Resolving' issues

Off Stall 'Dispatch Hold' Warp stall analysis for 'Dispatch Hold' issues

Off Stall 'Drain' Warp stall analysis for 'Drain' issues

Off Stall 'IMC Miss' Warp stall analysis for 'Immediate constant cache (IMC) misses

Off Stall 'LG Throttle' Warp stall analysis for 'LG Throttle' issues

Off Stall 'Long Scoreboard' Warp stall analysis for 'Long Scoreboard' issues

Off Stall 'Math Pipe Throttle' Warp stall analysis for 'Math Pipe Throttle' issues

Off Stall 'Membar' Warp stall analysis for 'Membar' issues

Off Stall 'MIO Throttle' Warp stall analysis for 'MIO Throttle' issues

Off Stall 'Misc' Warp stall analysis for 'Misc' issues

Off Stall 'No Instruction' Warp stall analysis for 'No Instruction' issues

Off Stall 'Not Selected' Warp stall analysis for 'Not Selected' issues

Off Stall 'Short Scoreboard' Warp stall analysis for 'Short Scoreboard' issues

Off Stall 'Sleeping' Warp stall analysis for 'Sleeping' issues

Off Stall 'TEX Throttle' Warp stall analysis for 'TEX Throttle' issues

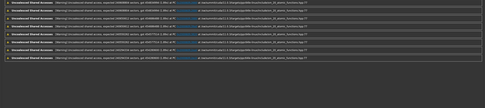
Off Stall 'Wait' Warp stall analysis for 'Wait' issues

Instruction Statistics

Statistics of the executed assembly instructions. The executed may not provide insight into the type and frequency of the executed instructions. It extends only of instruction types relative a dependency on the instruction pipelines, while others remain unused. Using multiple pipelines allows hiding latencies and increases utilization and throughput efficiently and can diverge if cycles are spent in unused cores.

Executed Instructions	16,908,352,512	Issued Instructions	16,908,352,512
Executed Instructions / Issued Instructions Ratio	1.00	Executed Instructions / Thread	264,192,227
Thread Instructions	1,081,270,202,112	Executed Instructions / Warp	8,254,078,375

Executed Instruction Mix



Executed Instructions/Opcode

NVLink Topology

No NVLink data was collected.

Worker Tables

Detailed tables with properties for each Worker.

Streaming Multiprocessor

Summary of the configuration used by 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.

Block Size	128	Grid Size	79,168
Threads	10,133,504	Waves Per SM	79.04
Block Limit SM	16	Block Limit Registers	4
Block Limit Shared Mem	16	Block Limit Warps	4
Theoretical Active Warps per SM	16	Theoretical Occupancy	25 %
Achieved Active Warps Per SM	11.98	Achieved Occupancy	18.72 %

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 memory latencies, resulting in overall performance degradation. Large discrepancies between the theoretical and the achieved occupancy during execution typically indicate highly imbalanced workloads.

Theoretical Occupancy [%]	25	Achieved Occupancy [%]	18.72
Theoretical Active Warps per SM [warp]	16	Achieved Active Warps Per SM [warp]	11.98

Theoretical Occupancy

Impact of Varying Register Count Per Thread

Impact of Varying Block Size

Impact of Varying Shared Memory Usage Per Block

Occupancy

Occupancy: Observing: This kernel's theoretical occupancy is limited by the number of required registers.

Source Counters [2]

Source metrics for memory, branch and warp stall sampling. Sampling data metrics are periodically sampled over the kernel runtime. They indicate which areas were active and could be optimized. See the documentation for a description of all stall reasons. Only focus on stalls if the scheduler's fail to issue every cycle.

Branch Instructions Ratio	0.007,938,932	Branch Instructions	134,217,728
Branch Instructions	134,217,728	Branch Efficiency [%]	100
Avg. Divergent Branches	0		

Sampling Data (All)

Sampling Data (Not Issued)

Most Instructions Executed

Location	Value
[0x0]	268,435,456
[0x10]	268,435,456
[0x20]	268,435,456
[0x30]	268,435,456
[0x40]	268,435,456
[0x50]	268,435,456
[0x60]	268,435,456
[0x70]	268,435,456
[0x80]	268,435,456
[0x90]	268,435,456

Recommendations

Uncoalesced Shared Accesses Observing: Uncoalesced global access, expected 33139072 sectors, got 66126886 (1.99x) at /root/home/ubuntu/Downloads/Nsight Compute-2026.1.1/host/target-desktop-glibc_2_11_3-x64/../../include/cuda/std/__cccl/cuda_toolkit.h:32:1 in void cutlass::device_kernel<cutlass::gemm::kernel::GemmUniversal<cutlass::gemm::GemmShape<128, 128, 32>, cutlass::gemm::GemmShape<64, 64, 32>, cutlass::gemm::GemmShape<16, 8, 16>, cutlass::gemm::threadblock::MmaPipelined<cutlass::gemm::GemmShape<128, 128, 32>, cutlass::transform::threadblock::PredicatedTileIterator<cutlass::MatrixShape<128, 32>, float, cutlass::layout::RowMajor, 1, cutlass::transform::PitchLinearStripminedThreadMap<cutlass::layout::PitchLinearShape<32, 128>, 128, 4>, 4, cutlass::NoPermute>, cutlass::transform::threadblock::RegularTileIterator<cutlass::MatrixShape<128, 32>, float, cutlass::layout::ColumnMajor, 1, cutlass::transform::PitchLinearStripminedThreadMap<cutlass::layout::PitchLinearShape<32, 128>, 128, 4>, 4>, cutlass::Array<float, 4, true>, cutlass::layout::RowMajor, cutlass::gemm::threadblock::MmaPolicy<cutlass::gemm::warp::MmaSimt<cutlass::gemm::GemmShape<32, 64, 32>, float, cutlass::layout::ColumnMajor, float, cutlass::layout::RowMajor, float, cutlass::layout::RowMajor, cutlass::gemm::warp::MmaSimtPolicy<cutlass::MatrixShape<4, 8>, cutlass::layout::RowMajorInterleaved<2>, cutlass::gemm::GemmShape<4, 4, 1>>>, cutlass::MatrixShape<0, 0>, cutlass::MatrixShape<0, 0>, 1>, cutlass::NumericArrayConverter<float, float, 4, (cutlass::FloatRoundStyle)2>, cutlass::gemm::threadblock::GemmIdentityThreadblockSwizzle<1>, 3>>>(T1::Params)