GPU Speed Of Light Throughput

Section Overview 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. High-level overview of the utilization for compute and memory resources of the GPU presented as a roofline chart.

Compute (SM) Throughput [%]	85.93	Duration [nsecond]	116.22
Memory Throughput [%]	20.87	Elapsed Cycles [cycle]	122769874
L1/TEX Cache Throughput [%]	29.45	SM Active Cycles [cycle]	122264182.56
L2 Cache Throughput [%]	10.46	SM Frequency [cycle/nsecond]	1.06
DRAM Throughput [%]	6.80	DRAM Frequency [cycle/nsecond]	1.51

High Throughput The kernel 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.

Roofline Analysis The ratio of peak float (fp32) to double (fp64) performance on this device is 2:1. The kernel achieved 40% of this device's fp32 peak performance and 0% of its fp64 peak performance. See the Kernel Profiling Guide for more details on roofline analysis.

GPU Throughput

Speed Of Light (SOL) [%]

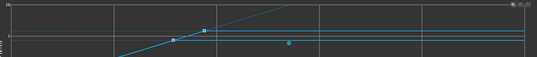
Compute Throughput Breakdown

SM: Instn Active [%]	85.93
SM: Inst Executed [%]	85.93
SM: Pipe Alu Cycles Active [%]	81.01
SM: Pipe Fma Cycles Active [%]	81.29
SM: Inst Executed Pipe Lsu [%]	14.17
SM: Mio2rf Writeback Active [%]	11.31
SM: Mio Inst Issued [%]	7.28
SM: Sysldst Cycles Active [%]	7.28
SM: Mio Pq Read Cycles Active [%]	2.93
SM: Inst Executed Pipe Adu [%]	2.93
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.94
SM: Inst Executed Pipe Uniform [%]	0.00
SM: Pipe Tensor Cycles Active [%]	0.00
SM: Pipe Shared Cycles Active [%]	0.00
SM: Inst Executed Pipe Tex [%]	0.00
SM: Inst Executed Pipe Ipa [%]	0.00
SM: Inst Executed Pipe Ipc [%]	0.00
SM: Pipe Fp64 Cycles Active [%]	0.00
SM: Inst Executed Pipe Xu [%]	0.00
ISC: Request Cycles Active [%]	0.00

Memory Throughput Breakdown

L1: Data Pipe Lsu Wavefronts [%]	29.07
L1: Data Pipe Tex Wavefronts [%]	22.61
L1: Lsuin Requests [%]	14.17
L1: Hit Requests [%]	10.46
L1: Sm2tex Cycles Active [%]	8.76
L1: Tex Hit Sum Rounds [%]	5.35
L2: Xbar Cycles Active [%]	6.59
L2: Lts Cycles Active [%]	6.59
L2: Lts2xbar Avg Cycles Active [%]	4.19
L2: D Sectors [%]	6.18
DRAM: Cmd Cycles Active [%]	6.80
L1: M Xbar2l1wrt Read Sectors [%]	5.07
DRAM: Dram Sectors [%]	4.45
L2: D Sectors Fill Device [%]	4.19
L2: Lts Tag Hit Rate [%]	3.93
L1: Tex Sectors Read Hits Active [%]	0.00
L1: F Wavefronts [%]	0.00
L2: D Sectors Fill System [%]	0.00
L1: Data Pipe Tex Wavefronts [%]	0
L2: D Atomic Input Cycles Active [%]	0
L1: Tex Writeback Active [%]	0

Floating Point Operations Roofline



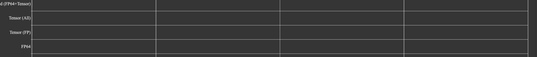
Compute Workload Analysis

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.

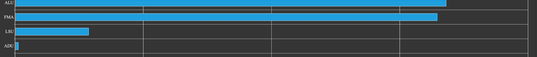
Executed Ipc Elapsed [inst/cycle]	3.44	SM Busy [%]
Executed Ipc Active [inst/cycle]	3.48	Issue Slots Busy [%]
Issued Ipc Active [inst/cycle]	3.48	Issue Slots Per Cycle [%]

Very High Utilization ALU is the highest-utilized pipeline (84.0%) based on active cycles, taking into account the rates of its different instructions. It executes integer and logic operations. The pipeline is over-utilized and likely a performance bottleneck. Based on the number of executed instructions, the highest utilized pipeline (84.0%) is ALU. It executes integer and logic operations. Comparing the two, the overall pipeline utilization appears to be caused by frequent, low-latency instructions. See the Kernel Profiling Guide or hover over the pipeline name to understand the workloads handled by each pipeline. The Instruction Statistics section provides a breakdown of instructions by unit.

Pipe Utilization (% of active cycles)



Pipe Utilization (% of peak instructions executed)



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 Pipes Busy). Detailed chart of the memory units. Detailed tables with data for each memory unit.

Memory Throughput [Gbyte/second]	116.06	Mem Busy [%]	29.07
L1/TEX Hit Rate [%]	10.46	Max Bandwidth [%]	22.61
L2 Hit Rate [%]	44.67	Mem Pipes Busy [%]	14.17
L2 Compression Success Rate [%]	0	L2 Compression Ratio	0

L2 Load Across Partners The memory access pattern for loads from L1/TEX to L2 is not optimal. The granularity of an L1/TEX request to L2 is a 128 byte cache line. That is 4 consecutive 32-byte sectors per L2 request. However, this kernel only accesses an average of 1.8 sectors out of the possible 4 sectors per cache line. Check the L2 Cache Locality section for uncoalesced loads and try to minimize how many cache lines need to be accessed per memory request.

Memory Chart

Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	1344798720	1344798720	2687044811	20.10	47719
Shared Load Matrix	0	0	0	0	0
Shared Store	2103785568	2103785568	3411804062	2.35	3245880
Shared Atomics Non Global Load	0	0	0	0	2,514
Shared Atomic	0	0	0	0	0
Other			5368355	0.04	0
Total	3448584288	3448584288	8099160298	22.50	3293246

L1/TEX Cache

	Instructions	Requests	Wavefronts	% Peak	Sectors	Sectors/Req	Hit Rate	Sector Misses to L2%	Peak to L2	Returns to SM	% Peak to SM
Local Load	335544320	335544320	335542116	2.51	1342443229	3.94	0.01	42510247328	154201797[2	5.02	335631887	2.51
Global Load To Shared Byte (access)	0	0	335542116	2.51	0	0	0	0	0	0	-	-
Global Load To Shared Byte (bypass)	0	0	0	0	0	0	0	0	0	0	-	-
Global Load Cache	0	0	0	0	0	0	0	0	0	0	0	0
Texture Load	0	0	0	0	0	0	0	0	0	0	-	-
Global Store	2621440	2621440	2621440	0.02	10485760	4	0	335544320	1048576[0	0.08	-	-
Local Store	0	0	0	0	0	0	0	0	0	0	-	-
Surface Store	0	0	0	0	0	0	0	0	0	0	-	-
Global Reduction	0	0	0	0	0	0	0	0	0	0	-	-
Surface Reduction	0	0	0	0	0	0	0	0	0	0	-	-
Global Atomic ALU	0	0	0	0	0	0	0	0	0	0	see above	see above
Global Atomic CAS	0	0	0	0	0	0	0	0	0	0	see above	see above
Surface Atomic ALU	0	0	0	0	0	0	0	0	0	0	see above	see above
Surface Atomic CAS	0	0	0	0	0	0	0	0	0	0	see above	see above
Loads	335544320	335544320	335542116	2.51	1342443229	3.94	0.01	42510247328	154201797[2	5.02	335631887	2.51
Stores	2621440	2621440	2621440	0.02	10485760	4	0	335544320	1048576[0	0.08	-	-
Atomics & Reductions	0	0	0	0	0	0	0	0	0	0	-	-
Total	338165760	338165760	338163556	2.53	1353929989	3.99	0.01	42845594608	155203373[2	5.10	335631887	2.51

L2 Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes	Throughput	Lost Misses to Tex	Lost Misses to Sys	Lost Misses to Peer	Misses to Peer
L1/TEX Load	830781959	1342018082	1.60	7.08	83.85	42944570624	369509192250.00	4913131834	0	0	0
L1/TEX Store	2621440	10485760	4	0.07	100	335544320	2887100871.22	0	0	0	0
L1/TEX Atomic ALU	0	0	0	0	0	0	0	0	0	0	0
L1/TEX Atomic CAS	0	0	0	0	0	0	0	0	0	0	0
L1/TEX Reduction	0	0	0	0	0	0	0	0	0	0	0
L1/TEX Total	841323399	1353503842	1.61	7.15	84.02	43280132944	372397221380.35	4934049[2	0	0	0
L2 Hits	1494812366	5910337661	3.88	6.13	64.29	189397775[2	1599836036.28	0	0	0	0
L2 Fabric Total	1494812366	5910337661	3.88	6.13	64.29	189397775[2	1599836036.28	0	0	0	0
GPU Total	4911721518	9326485608	1.90	6.57	68.11	416456176288	3504882350.13	0	0	0	0

L2 Cache Eviction Policies

	First	Hit Rate	Last	Hit Rate	Normal	Hit Rate	Normal Demote	Hit Rate
L1/TEX Load	0	0	0	0	1340779264	63.55	0	0
L1/TEX Store	0	0	0	0	10485760	100	0	0
L1/TEX Atomic	0	0	0	0	0	0	0	0
L1/TEX Total	0	0	0	0	1350264024	64.12	0	0
L2 Fabric Total	0	0	0	0	2962968148	50.82	0	0
GPU Total	40502	15.06	0	0	2427687488	37.46	0	0

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	411148172	8.67	1315629350	1132605388.13
Store	10399956	0.15	332799392	286320699.37
Total	421548128	8.88	1348428742	1160431087.50

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 Warp). 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.

Active Warps Per Scheduler [warp]	3.97	No Eligible [%]	13.61
Eligible Warps Per Scheduler [warp]	2.18	One or More Eligible [%]	86.39
Issued Warps Per Scheduler	0.87		

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 warp cycles per instruction define the latency between two consecutive instructions. The higher the value, the more warp parallelism is required to hide this latency. For each warp state, the chart shows the average number of cycles spent in that state per issued instruction. Stalls are not always impacting the overall performance nor are they completely avoidable. Only focus on stall reasons if the scheduler fails to issue every cycle. When a kernel is latency or instruction bound, of stalls, these metrics show the combined values.

Warp Cycles Per Issued Instruction [cycle]	4.58	Avg. Active Threads Per Warp
Warp Cycles Per Executed Instruction [cycle]	4.58	Avg. Not Predicated Off Threads Per Warp

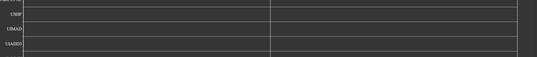
Instruction Statistics

Statistics of the executed low-level assembly instructions (SASS). The instruction mix provides insight into the types and frequency of the executed instructions. Analysis of memory instruction types implies a dependency on few instruction pipelines, while others remain unused. Using multiple pipelines effectively will improve the achievable parallel execution. Note that 'Instructions/Opcode' and 'Executed Instructions' are measured differently, see the Kernel Profiling Guide for details.

Executed Instructions [inst]	4996279200	Avg. Executed Instructions Per Scheduler [inst]	1963454276
Issued Instructions [inst]	4596264672	Avg. Issued Instructions Per Scheduler [inst]	1806345705.38

FP32/64 Instructions This kernel executes 0 fused and 2147483648 non-fused FP32 instructions. By converting pairs of non-fused instructions to their fused equivalent, the achieved FP32 performance could be increased by up to 50% (relative to its current performance). Check the Source page to identify where this kernel executes FP32 instructions.

Executed Instruction Mix



NVLink Topology

NVLink Topology diagram shows logical NVLink connections with transmit/receive throughput.

NVLink Tables

Detailed tables with properties for each NVLink.

NUMA Affinity

Non-uniform memory access (NUMA) affinities based on compute and memory distances for all GPUs.

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	256	Block Size	128
Registers Per Thread [register/thread]	128	Static Shared Memory Per Block [Kbyte/block]	24.58
Block Size	256	Dynamic Shared Memory Per Block [Kbyte/block]	0
Threads [thread]	1310720	Driver Shared Memory Per Block [Kbyte/block]	1.02
Warps Per SM	23.70	Shared Memory Configuration Size [Kbyte]	102.40

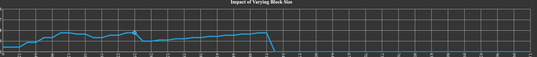
Occupancy

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.

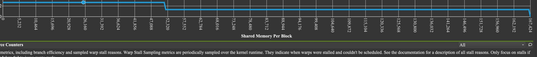
Theoretical Occupancy [%]	25	Block Limit Registers [block]	4
Theoretical Active Warps per SM [warp]	16	Block Limit Shared Mem [block]	4
Achieved Occupancy [%]	24.81	Block Limit Warps [block]	4
Achieved Active Warps Per SM [warp]	15.88	Block Limit SM [block]	32

Occupancy Limiters This kernel's theoretical occupancy (25.0%) is limited by the number of required registers. See the NVIDIA Best Practices Guide for more details on optimizing occupancy.

Impact of Varying Register Count Per Thread



Impact of Varying Block Size



Impact of Varying Shared Memory Usage Per Block

Source Counters

Source metrics, including branch efficiency and sampled warp stall reasons. Warp Stall Sampling metrics are periodically sampled over the kernel runtime. They indicate when warps were stalled and could not be scheduled. See the documentation for a description of all stall reasons. Only focus on stalls if the scheduler fails to issue every cycle.

Branch Instructions [inst]	160389360	Branch Efficiency [%]	100
Branch Instructions Ratio [%]	0.00	Avg. Divergent Branches	0

Uncoalesced Shared Access This kernel has uncoalesced shared accesses resulting in a total of 262140 excessive wavefronts (0% of the total 8027867200 wavefronts). Check the L1 Wavefronts Shared Excessive table for the primary source locations. The L1/TEX Hint Projection Table lists an example on optimizing shared memory accesses.

L1 Wavefronts Shared Excessive

Location	Value	Value (%)
0x7f0275d0f0f0 in sgemm_128x128nt_kernel_nsmall2	163,840	
0x7f0275d0f170 in sgemm_128x128nt_kernel_nsmall2	163,840	
0x7f0275d0f1e0 in sgemm_128x128nt_kernel_nsmall2	163,840	
0x7f0275d0f250 in sgemm_128x128nt_kernel_nsmall2	163,840	
0x7f0275d0f2c0 in sgemm_128x128nt_kernel_nsmall2	163,840	

Warp Stall Sampling (All Cycles)

Location	Value	Value (%)
0x7f0275d0f0a0 in sgemm_128x128nt_kernel_nsmall2	155,648	
0x7f0275d0f2d0 in sgemm_128x128nt_kernel_nsmall2	145,114	
0x7f0275d0f0e0 in sgemm_128x128nt_kernel_nsmall2	81,673	
0x7f0275d0f0c0 in sgemm_128x128nt_kernel_nsmall2	72,406	
0x7f0275d0f0a0 in sgemm_128x128nt_kernel_nsmall2	65,526	

Most Instructions Executed

Location	Value	Value (%)
0x7f0275d0fa20 in sgemm_128x128nt_kernel_nsmall2	42,902,080	
0x7f0275d0fa60 in sgemm_128x128nt_kernel_nsmall2	42,902,080	
0x7f0275d0fa40 in sgemm_128x128nt_kernel_nsmall2	42,902,080	
0x7f0275d0fa80 in sgemm_128x128nt_kernel_nsmall2	42,902,080	
0x7f0275d0fa00 in sgemm_128x128nt_kernel_nsmall2	42,902,080	