Current | Time | Size | Time | Time | GPU | SM Frequency | Process | Attributes
1 | vset_reduce_argmin | 1 | 12.32 s | 12.32 s | 12.32 s | 1.44 GHz | 28.00 GHz | Zero Error

Summary | Details | Source | Context | Comments | Raw | Sessions | Compare | Tools | View | Export

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 percentage of achieved performance 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 SOL Throughput [%]	24.31	Duration [us]	12.32
Memory SOL Throughput [%]	28.01	Elapsed Cycles [cycle]	17,724
DRAM Throughput [%]	28.01	Memory Throughput [Gbyte/s]	66,201.35
L2 Cache Throughput [%]	11.55	SM Frequency [MHz]	1,438.86
SYSMEM Throughput [%]	40.34	SYSMEM Frequency [GHz]	4.99

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.

Roofline Analysis The ratio of peak float (fp32) to double (fp64) performance on this device is 32:1. The kernel achieved 0% of this device's fp32 peak performance and 0% of its fp64 peak performance. See the Roofline Analysis Guide for more details on roofline analysis.

GPU Throughput

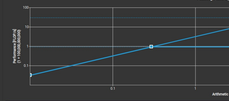
Compute Throughput Breakdown

SM Inst Executed Pipe Lsu [%]	24.31
SM Issue Active [%]	19.99
SM Inst Executed [%]	19.99
SM Mio Inst Issued [%]	9.61
SM Pipe Lsu Cycles Active [%]	24.31
SM Mio Inst Issued Pipe Lsu Active [%]	24.31
SM Mio Fly Piped Cycles Active [%]	9.31
SM Inst Executed Pipe Adu [%]	4.76
SM Mio2rf Writeback Active [%]	4.55
SM Pipe Adu Cycles Active [%]	4.76
SM Inst Executed Pipe Fma Op Heavy [%]	1.19
SM Inst Executed Pipes Unitem [%]	1.19
SM Memory Throughput Internal Activity [%]	1.19
SM Pipe Tensor Cycles Active [%]	0
SM Inst Executed Pipe Tensor [%]	0
SM Inst Executed Pipe Tensor Op Hmma [%]	0
SM Inst Executed Pipe Uniform [%]	0
SM Pipe Tensor Cycles Active [%]	0
SM Inst Executed Pipe Adu [%]	0
SM Inst Executed Pipe Cbu [%]	0
SM Inst Executed Pipe Fma [%]	0
SM Inst Executed Pipe Fp16 [%]	0

Memory Throughput Breakdown

DRAM Cycles Active [%]	40.04
DRAM Client Sectors [%]	29.46
L1 Local Requests [%]	19.01
L1 Data Pipe Lsu Wavefronts [%]	19.01
L1 Mem Shared Pipe Lsu Wavefronts [%]	19.01
L1 Sectors [%]	11.55
L2 L1Tex Cycles Active [%]	11.47
L2 Sectors Fill Device [%]	11.55
L1 Lsu Writeback Active [%]	7.89
L1 Tex Cache Fills [%]	7.89
L1 Data Bank Reads [%]	5.75
L1 Data Bank Reads [%]	3.01
GPC Compute Memory Access Throughput Balance [%]	2.99
L2 T Tag Requests [%]	2.99
L2 Slice Cycles Active [%]	1.89
L2 Slice Fills Sysmem Cycles Active [%]	1.89
L1 Tram Sectors Peak Cycles Active [%]	0.36
L1 T Wavefronts [%]	0.00
L2 X-domain Input Cycles Active [%]	0
L2 Crossbar Cycles Active [%]	0
L2 X-Switch Fill Sysmem [%]	0
L1 Data Pipe Tex Wavefronts [%]	0
GPC Compute Memory Request Throughput Internal Activity [%]	0

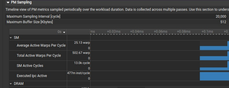
Floating Point Operations Roofline



PM Sampling

Timeline view of PM metrics sampled periodically over the workload duration. Data is collected across multiple passes. Use this section to understand how workload behavior changes over its runtime.

Maximum Sampling Interval [cycle]	20,000	Overall Samples [sample]	1
Minimum Sampling Interval [cycle]	1,000	Total Samples [sample]	1

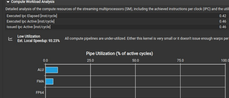


Compute Workload Analysis

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]	0.46	FMA Rate [%]	11.59
Executed Ipc Active [inst/cycle]	0.46	Issue Slots Busy [%]	11.59
Issued Ipc Active [inst/cycle]	0.46		

Low Utilization All compute pipelines are under-utilized. Either this kernel is very small or it doesn't issue enough warps per scheduler. Check the Launch Statistics and Scheduler Statistics sections for further details.



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 these units (Max Bandwidth), or by reaching their latency limits (Mem Pipes Busy). For more detailed analysis of memory bottlenecks, see the Memory Chart section.

Memory Throughput [Gbyte/s]	12.34	Mem Busy [%]	14.41
L1/TEX Hit Rate [%]	0.13	Max Bandwidth [%]	40.34
L2 Hit Rate [%]	0.84	Mem Pipes Busy [%]	24.01

Memory L2 Compression The optimal metric tm_average_gcomp_input_sector_success_rate pct could not be found. Collecting it as an additional metric could enable the kole to provide more guidance.

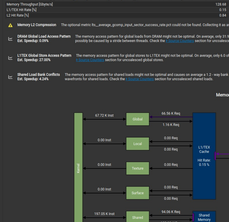
Global Global Load Access Pattern (sm__sass_l1tex_t_sectors_pipe_lsu_mem_global_op_ld.ratio) The memory access pattern for global loads from DRAM might not be optimal. On average, only 31.9 of the 32 bytes transmitted per sector are utilized by each thread. This applies to the 99.6% of sectors missed in L2. This could possibly be caused by a stride between threads. Check the Source Counters section for uncoalesced global loads.

L1TEX Global Store Access Pattern (sm__sass_l1tex_t_sectors_pipe_lsu_mem_global_op_st.ratio) The memory access pattern for global stores to L1TEX might not be optimal. On average, only 8.0 of the 32 bytes transmitted per sector are utilized by each thread. This could possibly be caused by a stride between threads. Check the Source Counters section for uncoalesced global stores.

Shared Load Bank Conflicts (sm__sass_l1tex_t_sectors_pipe_lsu_mem_shared_op_ld.sum.pct_of_peak_sustained_elapsed) The memory access pattern for shared loads might cause on average a 31.1-way bank conflict across all 94060 shared load requests. This results in 1421219 bank conflicts, which represent 12.75% of the overall 11159 wavefronts for shared loads. Check the Source Counters section for uncoalesced shared loads.

Memory Chart

Show All | Transfer Size



Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	94,060	94,060	111,069	8.83	14,219
Shared Load Matrix	0	0	0	0	0
Shared Store	102,992	102,992	103,259	5.17	0
Shared Atomics	0	0	0	0	0
Total	197,052	197,052	214,328	6.84	14,219

L1/TEX Cache

	Instructions	Requests	Wavefronts	% Peak	Sectors	Sectors/Req	Hit Rate	Bytes	Sector Misses to L2	% Peak to L2	Returns to SM
Local Load	0	0	0	0	0	0	0	0	0	0	-
Global Load	66,560	66,560	66,560	2.24	394,240	5.92	0.00	12,615,680	394,230	13.56	99,328
Global Load Ldg	0	0	0	0	0	0	0	0	0	0	-
Texture Load	0	0	0	0	0	0	0	0	0	0	-
Global Store	1,160	1,160	1,160	0.04	1,160	1	50.34	37,120	1,160	0.04	-
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 Atomic Reduction	0	0	0	0	0	0	0	0	0	0	-
Global Atomics	0	0	0	0	0	0	0	0	0	0	-
Surface Reduction	0	0	0	0	0	0	0	0	0	0	-
Global Atomic Add	0	0	0	0	0	0	0	0	0	0	-
Global Store Cast	0	0	0	0	0	0	0	0	0	0	see above
Global Reduction Add	0	0	0	0	0	0	0	0	0	0	-
Surface Atomic Add	0	0	0	0	0	0	0	0	0	0	see above
Loads	66,560	66,560	66,560	2.24	394,240	5.92	0.00	12,615,680	394,230	13.56	99,328
Stores	1,160	1,160	1,160	0.04	1,160	1	50.34	37,120	1,160	0.04	-
Atomics & Reductions	0	0	0	0	0	0	0	0	0	0	-

L2 Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes	Throughput	Sector Misses to Device	Sector Misses to System	Sector Misses to Peer
L1TEX Load	99,552	394,259	3.97	11.46	0.00	12,615,616	195.138 GB/s, 174.149	393,550	0	0
L1TEX Store	1,160	1,160	1	0.03	99.66	37,120	294.204 GB/s, 0.29	120	0	0
L1TEX Atomic Add	0	0	0	0	0	0	0	0	0	0
L1TEX Atomic Cas	0	0	0	0	0	0	0	0	0	0
L1TEX Atomic Red	0	0	0	0	0	0	0	0	0	0
L1TEX Total	100,480	395,419	3.93	11.49	0.21	12,652,736	195.137 GB/s, 294.204	393,670	0	0
EDC Total	-	-	-	-	-	-	4.160	15,346,391	0	0
DRAM Total	393,300	394,704	3.91	11.46	0.84	12,630,528	195.276,294.204	393,344	0	0

Device Memory

	Sectors	% Peak	Bytes	Throughput
DRAM	393,773	11.46	12,618,336	194.604 GB/s
PCIe Read	1,278	0.18	40,896	0.631 GB/s
PCIe Write	1,278	0.18	40,896	0.631 GB/s

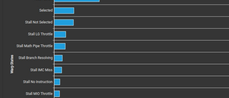
Scheduler Statistics

Summary of the activity of the schedulers issue instructions. Each scheduler maintains a pool of warps that can issue instructions for. The upper bound of warps in the pool (Theoretical Warps) is limited by the launch configuration. In 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.

Active Warps Per Scheduler [warp]	4.39	No Eligible [%]	63.16
Eligible Warps Per Scheduler [warp]	0.46	One Warp Eligible [%]	15.24
Issued Warps Per Scheduler	0.46		

Issue Slot Utilization Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 2.9 cycles. This might leave hardware resources underutilized and may lead to less optimal performance. Out of the maximum of 4 warps per scheduler, this kernel allocates, on average 4.39 active warps per scheduler, but only an average of 0.46 warps were eligible per cycle. Eligible warps are those warps which are ready to issue their next instruction. Every cycle with no eligible warps results in no instruction being issued and the issue slots remains unused. To increase the number of eligible warps, reduce the time the active warps are stalled by inspecting the top stall reasons on the Warp State Statistics and Source Counters sections.

Warps Per Scheduler



Warps Per Scheduler

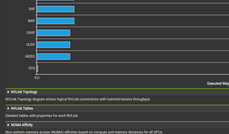
Summary of the activity of the schedulers issue instructions. Each scheduler maintains a pool of warps that can issue instructions for. The upper bound of warps in the pool (Theoretical Warps) is limited by the launch configuration. In every cycle each scheduler checks the state of the allocated warps 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.

Cycles With One Active Warp [%]	30.21	Avg. Active Threads Per Warp	30.31
Cycles With One Eligible Warp [%]	26.45	Avg. Not Predicated Off Threads Per Warp	26.45

Warp Cycles Per Issued Instruction On average, each warp of this kernel spends 13.9 cycles being stalled waiting for a scoreboard dependency on a L1TEX (local, global, surface, texture) operation. Find the instruction producing the data being waited upon to identify the culprit. To reduce the number of cycles waiting on L1TEX data accesses verify the memory access patterns are optimal for the target architecture, attempt to increase cache hit rates by increasing data locality (coalescing), or by changing the cache configuration. Consider staging frequently used data in shared memory. This stall type represents about 93.2% of the total average of 29.1 cycles between issuing two instructions.

Warp Stall Check the http://docs.nvidia.com/nsight-visual-studio/ sample doc. Check the Source Counters page provides more details on each stall reason.

Warp State (All Cycles)

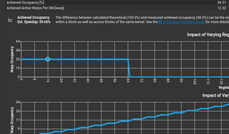


Instruction Statistics

Summary of the executed and issued instructions. Issued. The instruction mix provides insight into the type and frequency of the executed instructions. A narrow mix of instruction types implies a dependency on few instruction pipelines, while others remain unused. Using multiple pipelines allows hiding latencies and enables parallel execution. Note that 'Instructions/Opcode' and 'Executed Instructions' are measured differently and can diverge if cycles are spent in system calls.

Executed Instructions [inst]	1,792,000	Issued Instructions Per Scheduler [inst/cycle]	0.46
Issued Instructions [inst]	1,231,272	Avg. Issued Instructions Per Scheduler [inst]	7.632.95

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	1,024	Function Cache Configuration	CachePreferNone
Registers Per Thread [register/thread]	16	Static Shared Memory Per Block [Kbyte/block]	12.29
Block Size	128	Dynamic Shared Memory Per Block [Byte/block]	0
Number of Threads [thread]	131,072	Driver Shared Memory Per Block [Byte/block]	1.024
Waves Per SM	28.44	Shared Memory Configuration Size [Kbyte]	32.77
Uses Green Context	0	# SMs [SM]	48

Occupancy

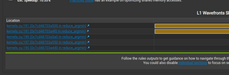
Occupancy is the ratio of the number of active warps per multiprocessor to the maximum number of possible active warps. Another way to view occupancy is the percentage of the hardware's ability to process warps that is actively in use. Higher occupancy does not always result in higher performance, however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

Theoretical Occupancy [%]	100	Block Limit Registers [block]	8
Theoretical Active Warps per SM [warp]	32	Block Limit Shared Mem [block]	7
Achieved Occupancy [%]	50.57	Block Limit Warps [block]	8
Achieved Active Warps Per SM [warp]	12.32	Block Limit SM [block]	16

Achieved Occupancy The difference between calculated theoretical (100.0%) and measured achieved occupancy (50.1%) can be the result of warp scheduling overheads or workload imbalances during the kernel execution. Load imbalances can occur between warps. See the Blocks Launched section 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

GPU and Memory Workload Distribution

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

Average SM Active Cycles [cycle]	66,201.35	Average L1 Active Cycles [cycle]	66,201.35
Average L2 Active Cycles [cycle]	29,208.75	Average SYSMEM Active Cycles [cycle]	16,267.33
Total DRAM Elapsed Cycles [cycle]	26,528,768	Total SM Elapsed Cycles [cycle]	3,177,472
Total L2 Elapsed Cycles [cycle]	2,097,040	Total L1 Elapsed Cycles [cycle]	3,091,456
Total SMSP Elapsed Cycles [cycle]	11,931,360	Total SYSMEM Elapsed Cycles [cycle]	4,940,873

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's fail to issue every cycle.

Branch Instructions [inst]	145,440	Branch Efficiency [%]	92.49
Branch Instruction Ratio [%]	8.12	Avg. Divergent Branches	59

Uncoalesced Shared Accesses (sl: Sampling: 10.5%) This kernel has uncoalesced shared accesses resulting in a total of 3962 excessive wavefronts (10% of the total 38410 wavefronts). Check the L1 Wavefronts Shared Excessive table for the primary source locations. The CUDA Best Practices Guide has an example on optimizing shared memory accesses.

L1 Wavefronts Shared Excessive

Location	Value	Value [%]
kernel(int, int, int,