	Result	Size	Time	Cycles	GPU	SM Frequency	Process	Attributes
Current	562 - reduce_argmin	(1024, 1, 1)x(1024, 1, 1)	1.26 ms	734,902	0 - Tesla T4	584.96 Mhz	[15952] exe	

Summary | Details | Source | Context | Comments | Raw | Session

Compare | Tools | View | Export

GPU Speed Of Light Throughput

All

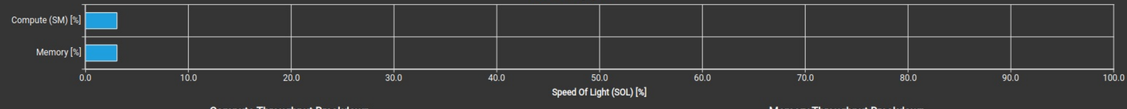
High-level overview of the throughput for compute and memory resources of the GPU. For each unit, the throughput reports the achieved percentage of utilization with respect to the theoretical maximum. Breakdowns show the throughput for each individual sub-metric of Compute and Memory to clearly identify the highest contributor. High-level overview of the utilization for compute and memory resources of the GPU presented as a roofline chart.

Compute (SM) Throughput [%]	3.06	Duration [ms]	1.26
Memory Throughput [%]	3.06	Elapsed Cycles [cycle]	734,902
L1/TEX Cache Throughput [%]	4.45	SM Active Cycles [cycle]	714,154.72
L2 Cache Throughput [%]	0.68	SM Frequency [Mhz]	584.96
DRAM Throughput [%]	1.53	DRAM Frequency [Ghz]	4.98

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 e Scheduler Statistics and e 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 Kernel Profiling Guide for more details on roofline analysis.

GPU Throughput



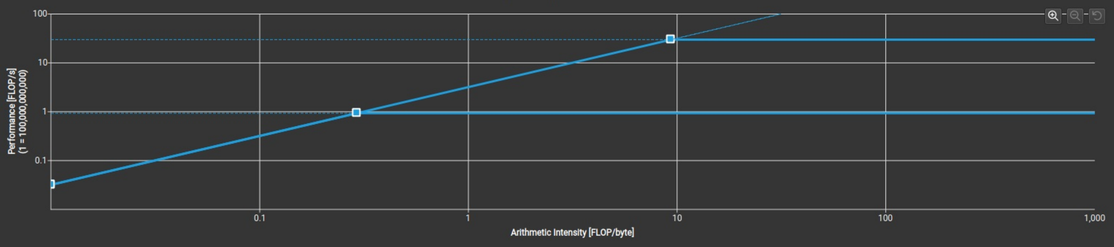
Compute Throughput Breakdown

SM: Inst Executed Pipe Lsu [%]	3.06
SM: Issue Active [%]	2.16
SM: Inst Executed [%]	1.93
SM: Mio Pq Write Cycles Active [%]	1.75
SM: Mio Pq Read Cycles Active [%]	1.44
SM: Mio Inst Issued [%]	1.19
SM: Pipe Alu Cycles Active [%]	1.07
SM: Pipe Fma Cycles Active [%]	0.75
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.56
SM: Inst Executed Pipe Adu [%]	0.50
SM: Mio2rf Writeback Active [%]	0.43
SM: Inst Executed Pipe Uniform [%]	0.43
SM: Inst Executed Pipe Tex [%]	0.01
SM: Memory Throughput Internal Activity [%]	0
SM: Pipe Tensor Cycles Active [%]	0
SM: Pipe Shared Cycles Active [%]	0
SM: Pipe Fp64 Cycles Active [%]	0
IDC: Request Cycles Active [%]	0
SM: Instruction Throughput Internal Activity [%]	0
SM: Inst Executed Pipe Xu [%]	0
SM: Inst Executed Pipe Ipa [%]	0
SM: Inst Executed Pipe Fp16 [%]	0

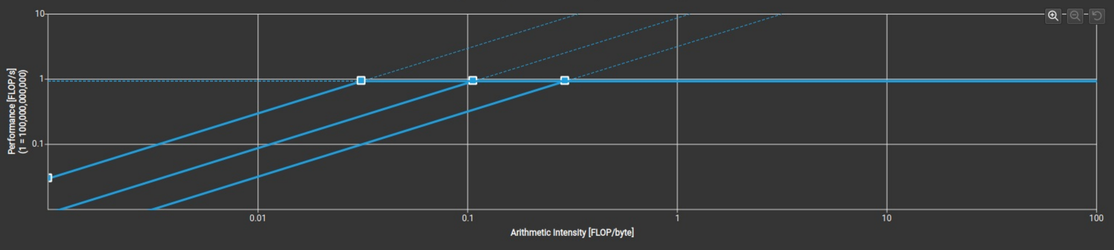
Memory Throughput Breakdown

L1: Lsuin Requests [%]	3.06
L1: Data Pipe Lsu Wavefronts [%]	2.23
DRAM: Cycles Active [%]	1.53
DRAM: Dram Sectors [%]	1.11
L1: Lsu Writeback Active [%]	0.87
L1: M Xbar2l1tex Read Sectors [%]	0.76
L2: T Sectors [%]	0.68
L2: Lts2xbar Cycles Active [%]	0.68
GPU: Compute Memory Access Throughput Internal Activity [%]	0.40
L2: T Tag Requests [%]	0.40
L1: M L1tex2xbar Req Cycles Active [%]	0.39
L2: D Sectors Fill Device [%]	0.38
L2: Xbar2lts Cycles Active [%]	0.37
L1: Data Bank Writes [%]	0.31
L2: D Sectors [%]	0.30
L2: D Atomic Input Cycles Active [%]	0.27
L1: Data Bank Reads [%]	0.24
L1: Texin Sm2tex Req Cycles Active [%]	0.05
L1: F Wavefronts [%]	0.00
L1: Data Pipe Tex Wavefronts [%]	0
L1: Tex Writeback Active [%]	0
L2: D Sectors Fill Sysmem [%]	0
GPU: Compute Memory Request Throughput Internal Activity [%]	0

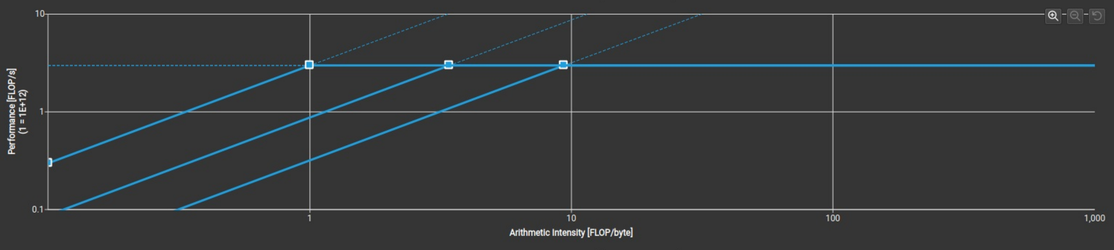
Floating Point Operations Roofline



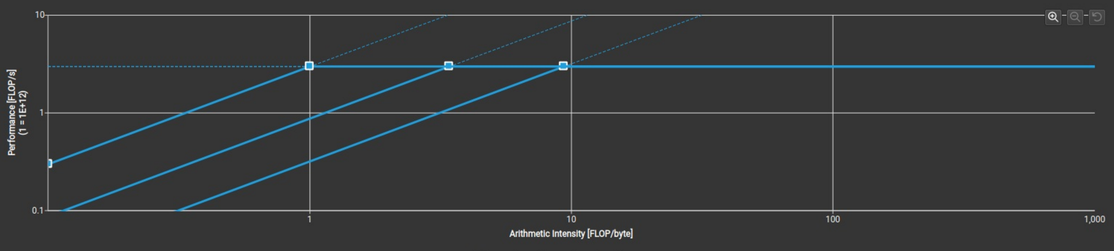
Floating Point Operations Roofline (Double Precision)



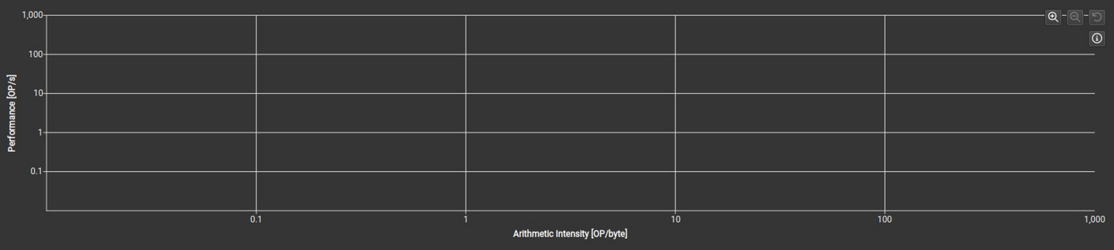
Floating Point Operations Roofline (Half Precision)



Floating Point Operations Roofline (Single Precision)



Tensor Core Operations Roofline



	# Operations	# Operations / Cycle	# Operations / s	Peak % ?eak Operations / Cycle	Peak Operations / s
Src:fp16,bf16,tf32 Dst:fp32	0	0	0	0	40,960	23,900.05
Src:fp16 Dst:fp16	0	0	0	0	40,960	23,900.05
Src:int1	0	0	0	0	81,920	47,800.10
Src:int4	0	0	0	0	81,920	47,800.10
Src:int8	0	0	0	0	81,920	47,800.10

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]	714,154.72	Average L1 Active Cycles [cycle]	714,154.72
Average L2 Active Cycles [cycle]	112,565.66	Average SMSP Active Cycles [cycle]	199,932.30
Average DRAM Active Cycles [cycle]	95,501	Total SM Elapsed Cycles [cycle]	29,322,376
Total L1 Elapsed Cycles [cycle]	29,322,376	Total L2 Elapsed Cycles [cycle]	34,370,752
Total SMSP Elapsed Cycles [cycle]	117,289,504	Total DRAM Elapsed Cycles [cycle]	50,094,080

SMSPs Workload Imbalance
Est. Speedup: 15.83%
One or more SMSPs have a much lower number of active cycles than the average number of active cycles. Maximum instance value is 58.05% above the average, while the minimum instance value is 78.22% below the average.

L2 Slices Workload Imbalance
Est. Speedup: 9.33%
One or more L2 Slices have a much higher number of active cycles than the average number of active cycles. Maximum instance value is 89.02% above the average, while the minimum instance value is 30.04% below the average.

Workload Distribution

	Average	Min	Max	Sum
SM Active Cycles	714,154.72	700,143	728,273	28,566,189
SMSP Active Cycles	199,932.30	43,554	476,642	31,989,168
L1 Active Cycles	714,154.72	700,143	728,273	28,566,189
L2 Active Cycles	112,565.66	78,752	1,024,968	3,602,101
DRAM Active Cycles	95,501	95,012	96,540	764,008