

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.

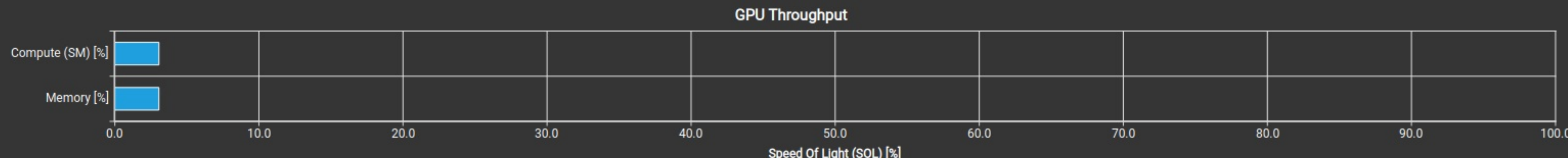
Compute (SM) Throughput [%]	3.06	Duration [ms]	1.25
Memory Throughput [%]	3.06	Elapsed Cycles [cycle]	731,360
L1/TEX Cache Throughput [%]	4.45	SM Active Cycles [cycle]	711,143.22
L2 Cache Throughput [%]	0.69	SM Frequency [Mhz]	584.98
DRAM Throughput [%]	1.53	DRAM Frequency [Ghz]	5.00

Latency Issue

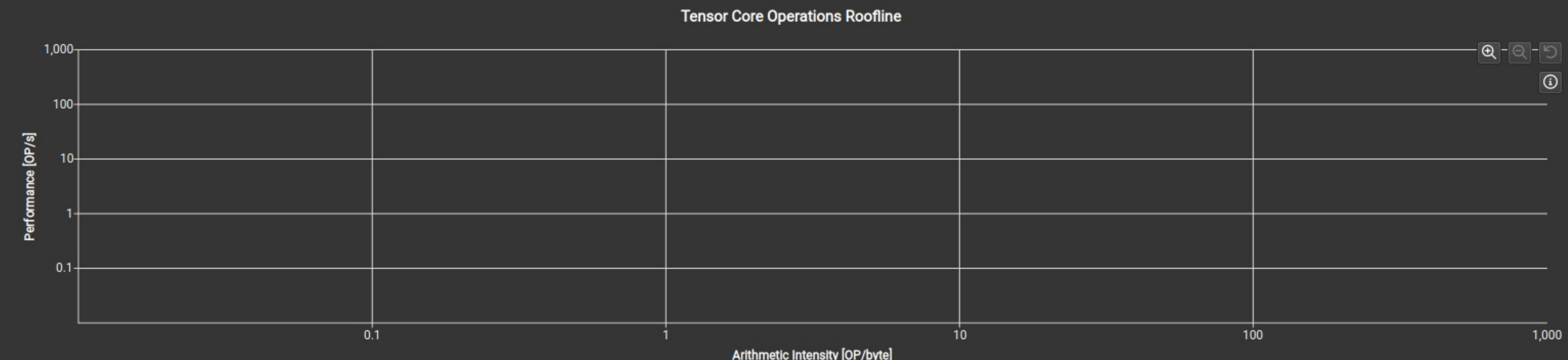
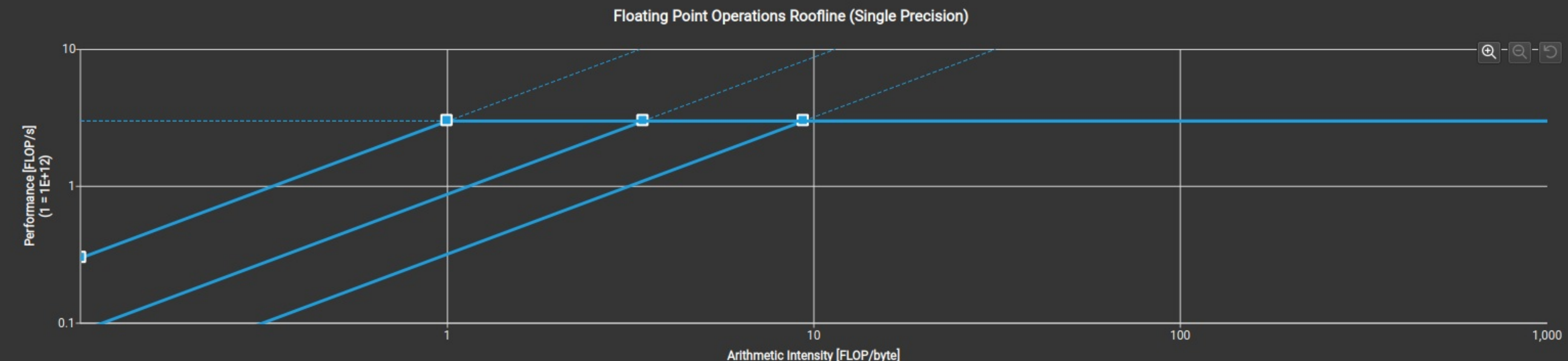
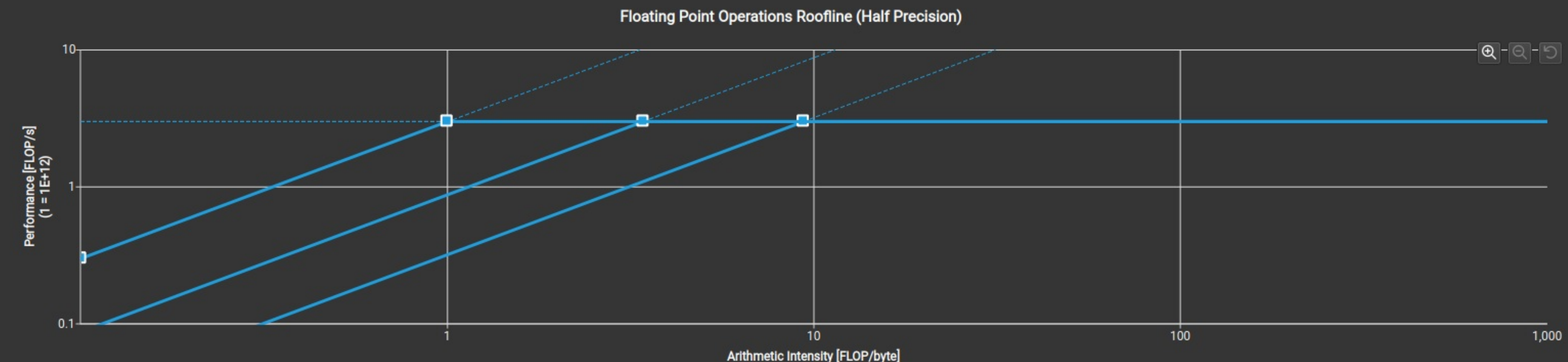
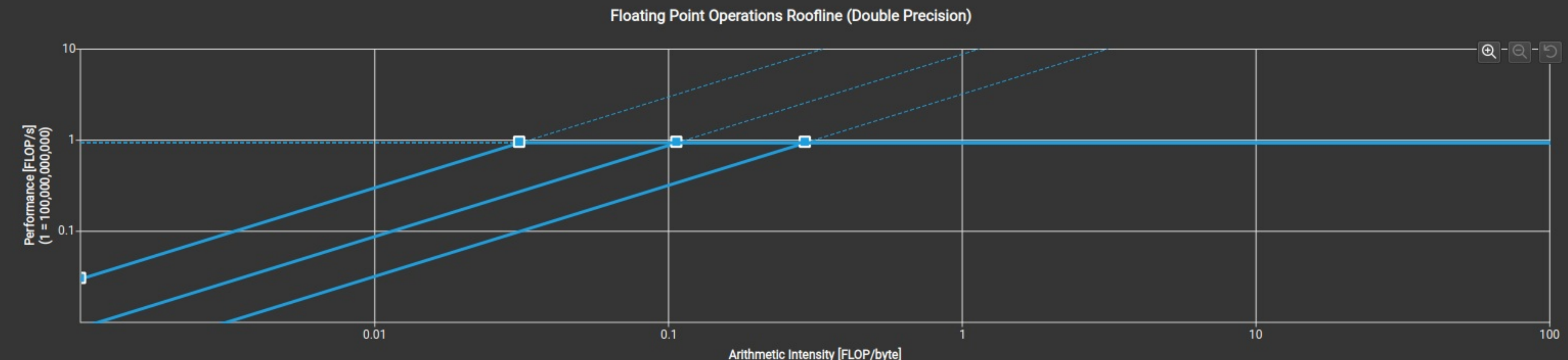
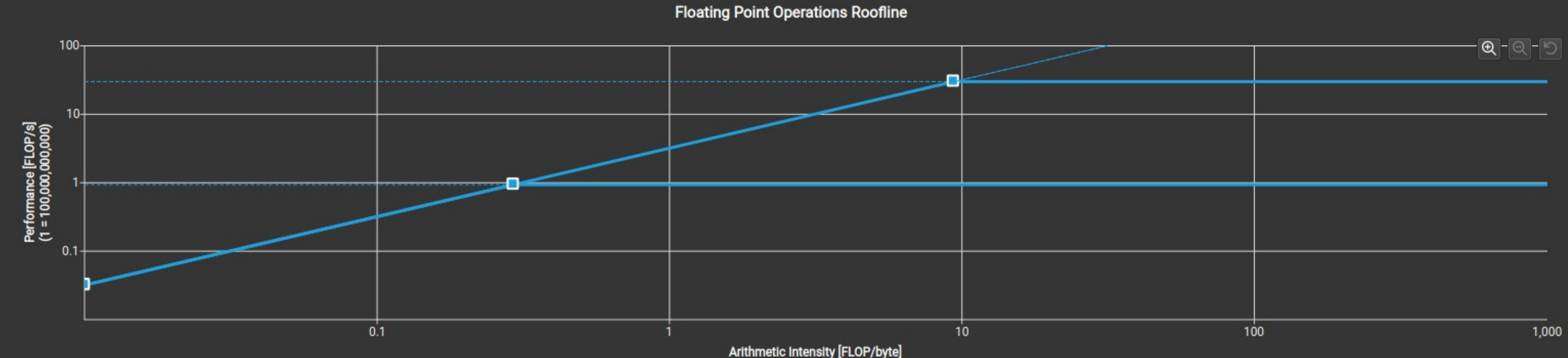
This kernel exhibits low compute throughput and memory bandwidth utilization relative to the peak performance of this device. Achieved compute throughput and/or memory bandwidth below 60.0% of peak typically indicate latency issues. Look at [Scheduler Statistics](#) and [Warp State Statistics](#) for 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 [Kernel Profiling Guide](#) for more details on roofline analysis.



Compute Throughput Breakdown			Memory Throughput Breakdown		
SM: Inst Executed Pipe Lsu [%]	3.06	L1: LsuIn Requests [%]	3.06		
SM: Issue Active [%]	2.15	L1: Data Pipe Lsu Wavefronts [%]	2.23		
SM: Inst Executed [%]	1.92	DRAM: Cycles Active [%]	1.53		
SM: Mio Pq Write Cycles Active [%]	1.75	DRAM: Dram Sectors [%]	1.12		
SM: Mio Pq Read Cycles Active [%]	1.44	L1: Lsu Writeback Active [%]	0.87		
SM: Mio Inst Issued [%]	1.19	L1: M Xbar2l1tex Read Sectors [%]	0.76		
SM: Pipe Alu Cycles Active [%]	1.07	L2: T Sectors [%]	0.69		
SM: Pipe Fma Cycles Active [%]	0.75	L2: Lts2xbar Cycles Active [%]	0.68		
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.56	GPU: Compute Memory Access Throughput Internal Activity [%]	0.40		
SM: Inst Executed Pipe Adu [%]	0.50	L2: T Tag Requests [%]	0.40		
SM: Mio2rf Writeback Active [%]	0.43	L1: M L1tex2xbar Req Cycles Active [%]	0.39		
SM: Inst Executed Pipe Uniform [%]	0.43	L2: D Sectors Fill Device [%]	0.39		
SM: Inst Executed Pipe Tex [%]	0.01	L2: Xbar2lts Cycles Active [%]	0.37		
SM: Memory Throughput Internal Activity [%]	0	L1: Data Bank Writes [%]	0.31		
SM: Pipe Tensor Cycles Active [%]	0	L2: D Sectors [%]	0.30		
SM: Pipe Shared Cycles Active [%]	0	L2: D Atomic Input Cycles Active [%]	0.27		
SM: Pipe Fp64 Cycles Active [%]	0	L1: Data Bank Reads [%]	0.24		
IDC: Request Cycles Active [%]	0	L1: Texin Sm2tex Req Cycles Active [%]	0.05		
SM: Instruction Throughput Internal Activity [%]	0	L1: F Wavefronts [%]	0.00		
SM: Inst Executed Pipe Xu [%]	0	L1: Data Pipe Tex Wavefronts [%]	0		
SM: Inst Executed Pipe Ipa [%]	0	L1: Tex Writeback Active [%]	0		
SM: Inst Executed Pipe Fp16 [%]	0	L2: D Sectors Fill Sysmem [%]	0		
		GPU: Compute Memory Request Throughput Internal Activity [%]	0		



	# Operations	# Operations / Cycle	# Operations / s	Peak %	Peak Operations / Cycle	Peak Operations / s
Src:fp16,bf16,tf32 Dst:fp32	0	0	0	0	40,960	24,003.27
Src:fp16 Dst:fp16	0	0	0	0	40,960	24,003.27
Src:int1	0	0	0	0	81,920	48,006.55
Src:int4	0	0	0	0	81,920	48,006.55
Src:int8	0	0	0	0	81,920	48,006.55

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]	711,143.22	Average L1 Active Cycles [cycle]	711,143.22
Average L2 Active Cycles [cycle]	112,999.88	Average SMSP Active Cycles [cycle]	199,163.93
Average DRAM Active Cycles [cycle]	95,557.50	Total SM Elapsed Cycles [cycle]	29,306,496
Total L1 Elapsed Cycles [cycle]	29,306,496	Total L2 Elapsed Cycles [cycle]	34,205,120
Total SMSP Elapsed Cycles [cycle]	117,225,984	Total DRAM Elapsed Cycles [cycle]	49,971,200

SMSPs Workload Imbalance

Est. Speedup: 13.17%

One or more SMSPs have a much lower number of active cycles than the average number of active cycles. Maximum instance value is 48.46% above the average, while the minimum instance value is 72.66% below the average.

L2 Slices Workload Imbalance

Est. Speedup: 9.40%

One or more L2 Slices have a much higher number of active cycles than the average number of active cycles. Maximum instance value is 88.92% above the average, while the minimum instance value is 29.87% below the average.

Workload Distribution

	Average	Min	Max	Sum
SM Active Cycles	711,143.22	697,172	726,353	28,445,729
SMSP Active Cycles	199,163.93	54,448	386,435	31,866,229
L1 Active Cycles	711,143.22	697,172	726,353	28,445,729
L2 Active Cycles	112,999.88	79,249	1,019,747	3,615,996
DRAM Active Cycles	95,557.50	94,824	96,532	764,460