

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.

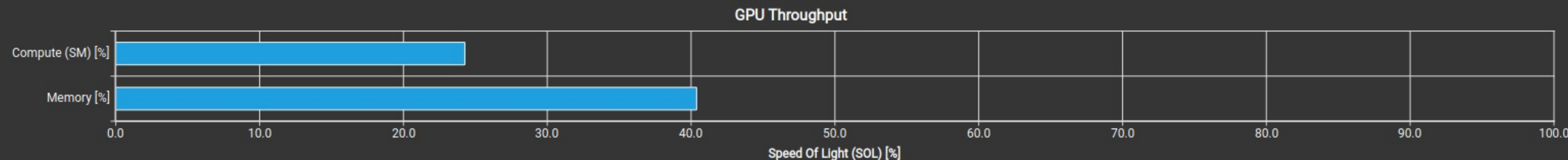
Compute (SM) Throughput [%]	24.27	Duration [us]	124.64
Memory Throughput [%]	40.38	Elapsed Cycles [cycle]	72,891
L1/TEX Cache Throughput [%]	33.31	SM Active Cycles [cycle]	66,064.80
L2 Cache Throughput [%]	11.64	SM Frequency [Mhz]	584.77
DRAM Throughput [%]	40.38	DRAM Frequency [Ghz]	4.96

Latency Issue

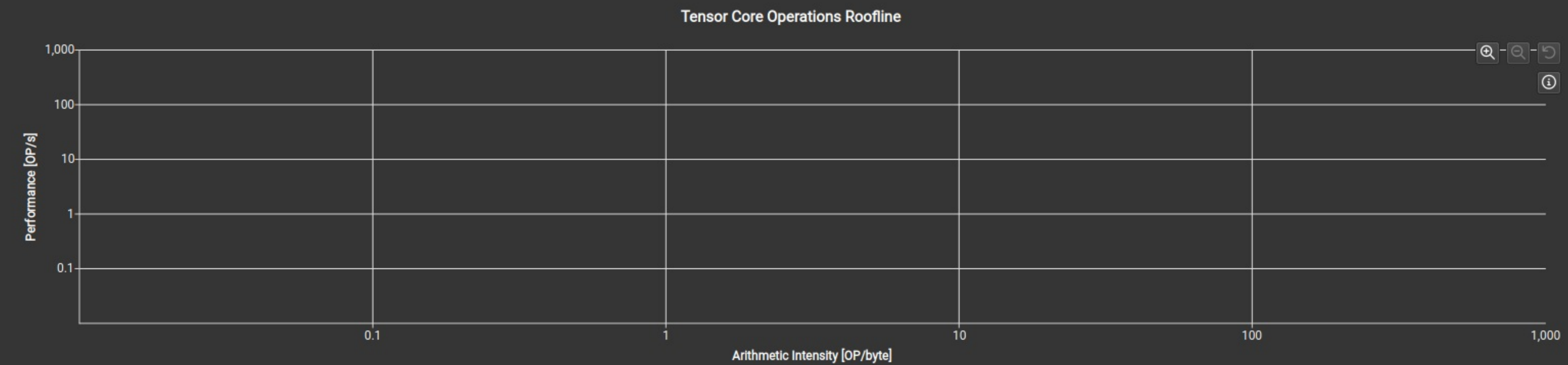
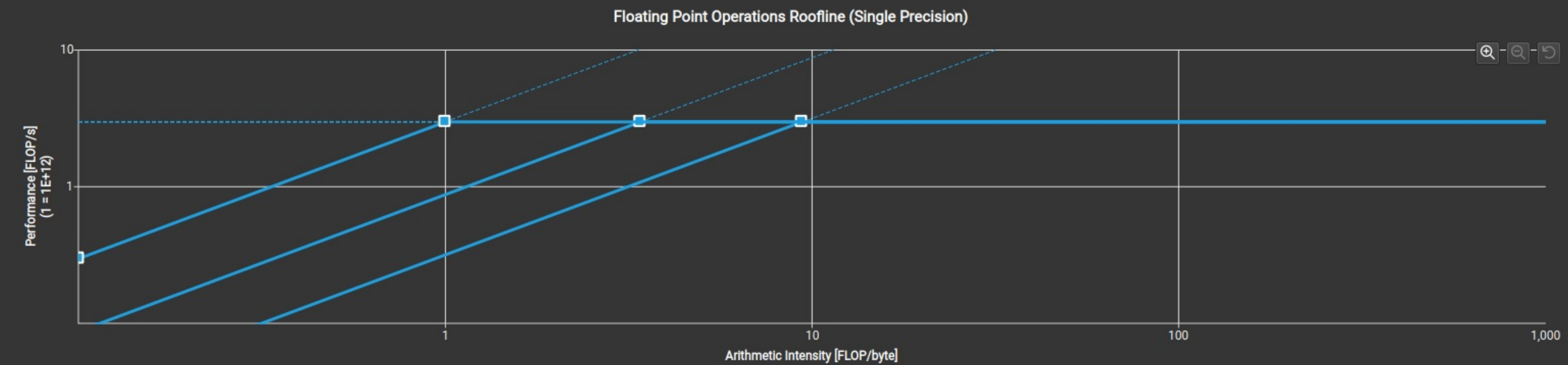
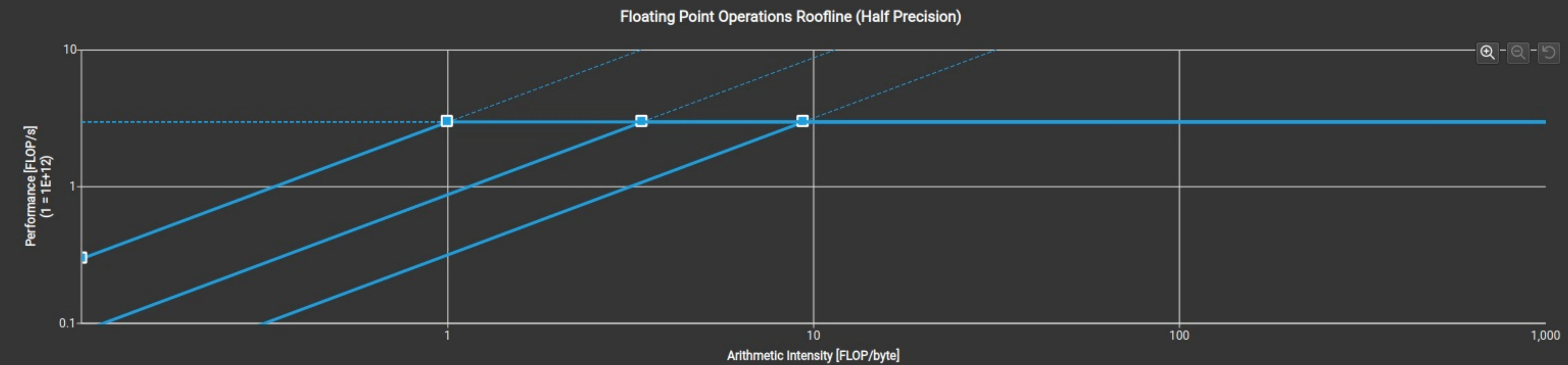
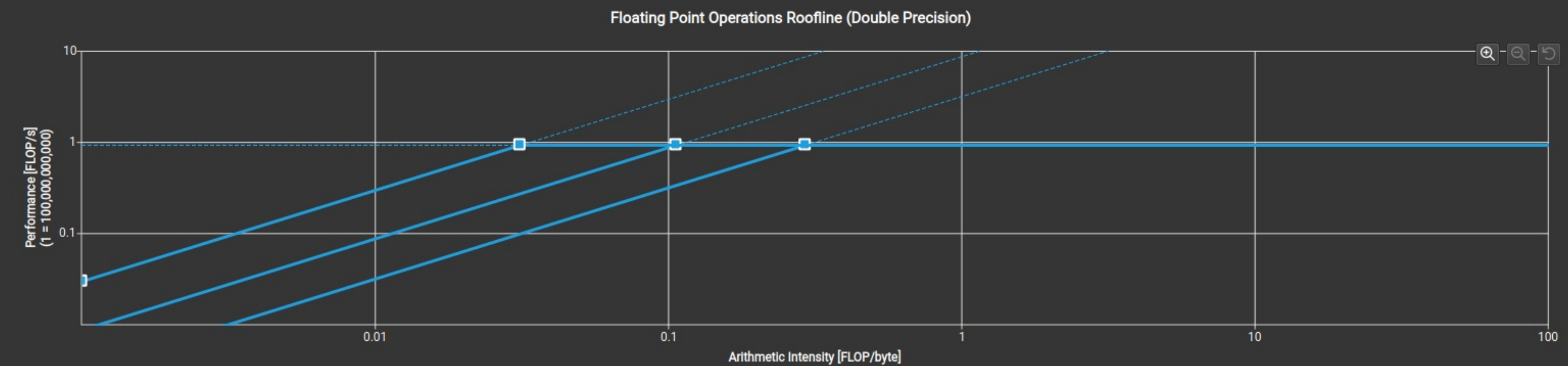
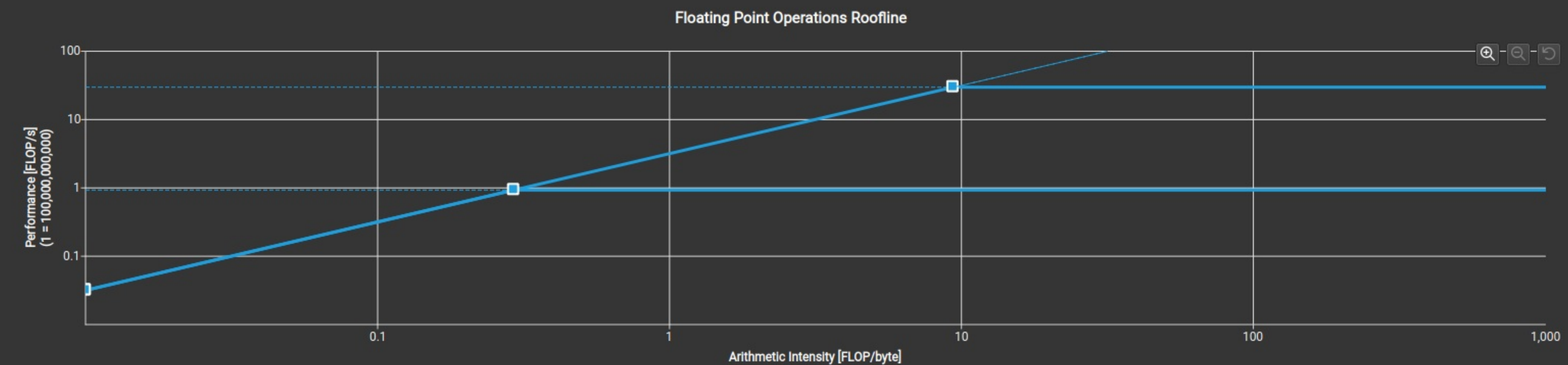
This kernel exhibits low compute throughput and memory bandwidth utilization relative to the peak performance of this device. Achieved compute throughput and/or memory bandwidth below 60.0% of peak typically indicate latency issues. Look at [Scheduler Statistics](#) and [Warp State Statistics](#) for 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 [%]	24.27	DRAM: Cycles Active [%]	40.38		
SM: Issue Active [%]	10.53	DRAM: Dram Sectors [%]	29.28		
SM: Inst Executed [%]	10.47	L1: Lsuin Requests [%]	24.27		
SM: Mio Inst Issued [%]	9.69	L1: Data Pipe Lsu Wavefronts [%]	16.66		
SM: Pipe Alu Cycles Active [%]	6.18	L1: M Xbar2l1tex Read Sectors [%]	13.59		
SM: Mio Pq Write Cycles Active [%]	5.39	L2: T Sectors [%]	11.64		
SM: Mio Pq Read Cycles Active [%]	5.32	L2: Lts2xbar Cycles Active [%]	11.59		
SM: Inst Executed Pipe Adu [%]	4.80	L2: D Sectors Fill Device [%]	11.55		
SM: Mio2rf Writeback Active [%]	4.52	L1: Lsu Writeback Active [%]	7.91		
SM: Pipe Fma Cycles Active [%]	3.85	L2: D Sectors [%]	6.06		
SM: Inst Executed Pipe Cbu Pred On Any [%]	2.58	L1: Data Bank Writes [%]	3.77		
SM: Inst Executed Pipe Uniform [%]	1.13	L1: Data Bank Reads [%]	3.07		
SM: Memory Throughput Internal Activity [%]	0	GPU: Compute Memory Access Throughput Internal Activity [%]	3.01		
SM: Pipe Tensor Cycles Active [%]	0	L2: T Tag Requests [%]	2.98		
SM: Pipe Shared Cycles Active [%]	0	L1: M L1tex2xbar Req Cycles Active [%]	1.92		
SM: Pipe Fp64 Cycles Active [%]	0	L2: Xbar2lts Cycles Active [%]	1.87		
IDC: Request Cycles Active [%]	0	L1: Texin Sm2tex Req Cycles Active [%]	0.36		
SM: Instruction Throughput Internal Activity [%]	0	L1: F Wavefronts [%]	0.00		
SM: Inst Executed Pipe Xu [%]	0	L2: D Atomic Input Cycles Active [%]	0		
SM: Inst Executed Pipe Tex [%]	0	L1: Tex Writeback Active [%]	0		
SM: Inst Executed Pipe lpa [%]	0	L2: D Sectors Fill Systemem [%]	0		
SM: Inst Executed Pipe Fp16 [%]	0	L1: Data Pipe Tex Wavefronts [%]	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	23,828.31
Src:fp16 Dst:fp16	0	0	0	0	40,960	23,828.31
Src:int1	0	0	0	0	81,920	47,656.62
Src:int4	0	0	0	0	81,920	47,656.62
Src:int8	0	0	0	0	81,920	47,656.62

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]	66,064.80	Average L1 Active Cycles [cycle]	66,064.80		
Average L2 Active Cycles [cycle]	73,801.78	Average SMSP Active Cycles [cycle]	44,353.59		
Average DRAM Active Cycles [cycle]	249,466	Total SM Elapsed Cycles [cycle]	2,900,352		
Total L1 Elapsed Cycles [cycle]	2,900,352	Total L2 Elapsed Cycles [cycle]	3,408,576		
Total SMSP Elapsed Cycles [cycle]	11,601,408	Total DRAM Elapsed Cycles [cycle]	4,941,824		

Workload Distribution				
	Average	Min	Max	Sum
SM Active Cycles	66,064.80	64,985	67,247	2,642,592
SMSP Active Cycles	44,353.59	42,265	46,176	7,096,575
L1 Active Cycles	66,064.80	64,985	67,247	2,642,592
L2 Active Cycles	73,801.78	72,264	75,459	2,361,657
DRAM Active Cycles	249,466	248,464	250,428	1,995,728