Result	Size	Time	Cycles	GPU	SM Frequency	Process	Attributes
Current 1790 - gemmNaive	(128, 128, 1)x(32, 32, 1)	2.65 s	2,387,515,271	0 - NVIDIA GeForce RTX 3060 Laptop GPU	900.00 Mhz	[1566] 01_naive_gemm	

Summary | Details | Source | Context | Comments | Raw | Session — Compare | Tools | View | Export

GPU Speed Of Light Throughput

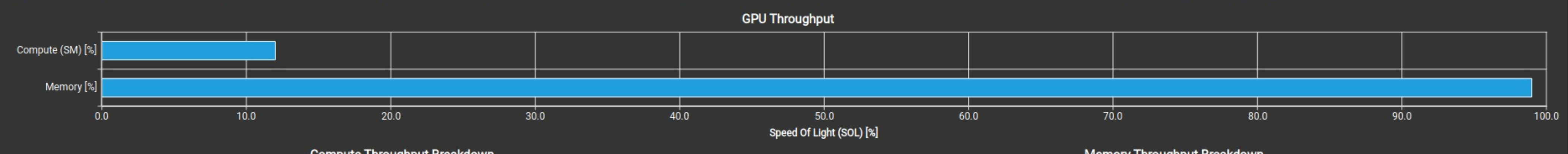
All

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 [%]	12.00	Duration [s]	2.65
Memory Throughput [%]	98.99	Elapsed Cycles [cycle]	2387515271
L1/TEX Cache Throughput [%]	99.11	SM Active Cycles [cycle]	2384490716.17
L2 Cache Throughput [%]	1.20	SM Frequency [Mhz]	900.00
DRAM Throughput [%]	1.07	DRAM Frequency [Ghz]	6.99

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 L1 in the Memory Workload Analysis section.

GPU Throughput



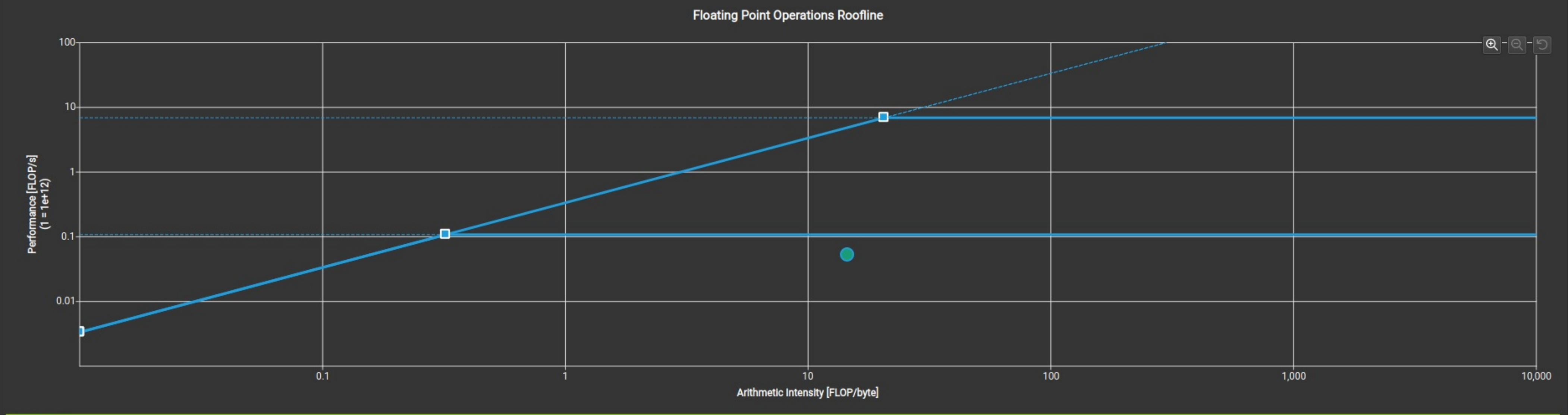
Compute Throughput Breakdown

SM: Inst Executed Pipe Lsu [%]	12.00
SM: Pipe Fmaheavy Cycles Active [%]	6.75
SM: Issue Active [%]	6.19
SM: Inst Executed [%]	6.19
SM: Pipe Fma Cycles Active [%]	4.13
SM: Mio Inst Issued [%]	4.00
SM: Pipe Alu Cycles Active [%]	3.75
SM: Mio2rf Writeback Active [%]	3.01
SM: Mio Pq Read Cycles Active [%]	0.01
SM: Inst Executed Pipe Adu [%]	0.00
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.00
SM: Mio Pq Write Cycles Active [%]	0.00
SM: Inst Executed Pipe Uniform [%]	0.00
SM: Memory Throughput Internal Activity [%]	0
SM: Instruction Throughput Internal Activity [%]	0
SM: Inst Executed Pipe Xu [%]	0
SM: Inst Executed Pipe Tex [%]	0
SM: Inst Executed Pipe Ipa [%]	0
IDC: Request Cycles Active [%]	0
SM: Pipe Fp64 Cycles Active [%]	0
SM: Pipe Tensor Cycles Active [%]	0

Memory Throughput Breakdown

L1: Data Pipe Lsu Wavefronts [%]	98.99
L1: Lsuin Requests [%]	12.00
L1: Lsu Writeback Active [%]	6.25
L1: Data Bank Reads [%]	1.55
L2: Xbar2lts Cycles Active [%]	1.20
L2: T Sectors [%]	1.17
GPU: Compute Memory Access Throughput Internal Activity [%]	1.17
L2: T Tag Requests [%]	1.17
L2: Lts2xbar Cycles Active [%]	1.14
DRAM: Dram Sectors [%]	1.07
DRAM: Cycles Active [%]	1.06
L1: M L1tex2xbar Req Cycles Active [%]	0.91
L1: M Xbar2l1tex Read Sectors [%]	0.86
L2: D Sectors Fill Device [%]	0.53
L2: D Sectors [%]	0.43
L1: Data Bank Writes [%]	0.11
L2: D Sectors Fill Sysmem [%]	0.00
L1: Texin Sm2tex Req Cycles Active [%]	0.00
L1: F Wavefronts [%]	0
L1: Tex Writeback Active [%]	0
L2: D Atomic Input Cycles Active [%]	0
L1: Data Pipe Tex Wavefronts [%]	0
GPU: 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.

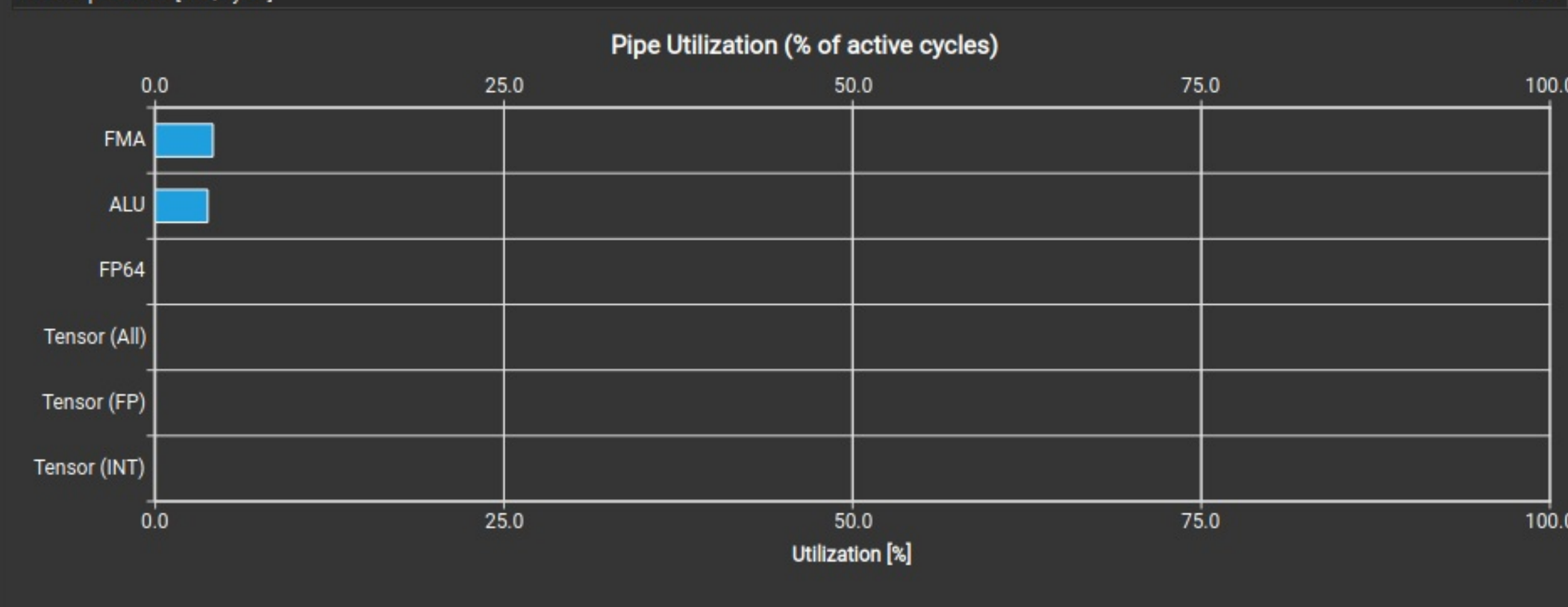
Maximum Sampling Interval [ms]	8.19	# Pass Groups	2
Maximum Buffer Size [Mbytes]	18.88	Dropped Samples [sample]	8

Compute Workload Analysis

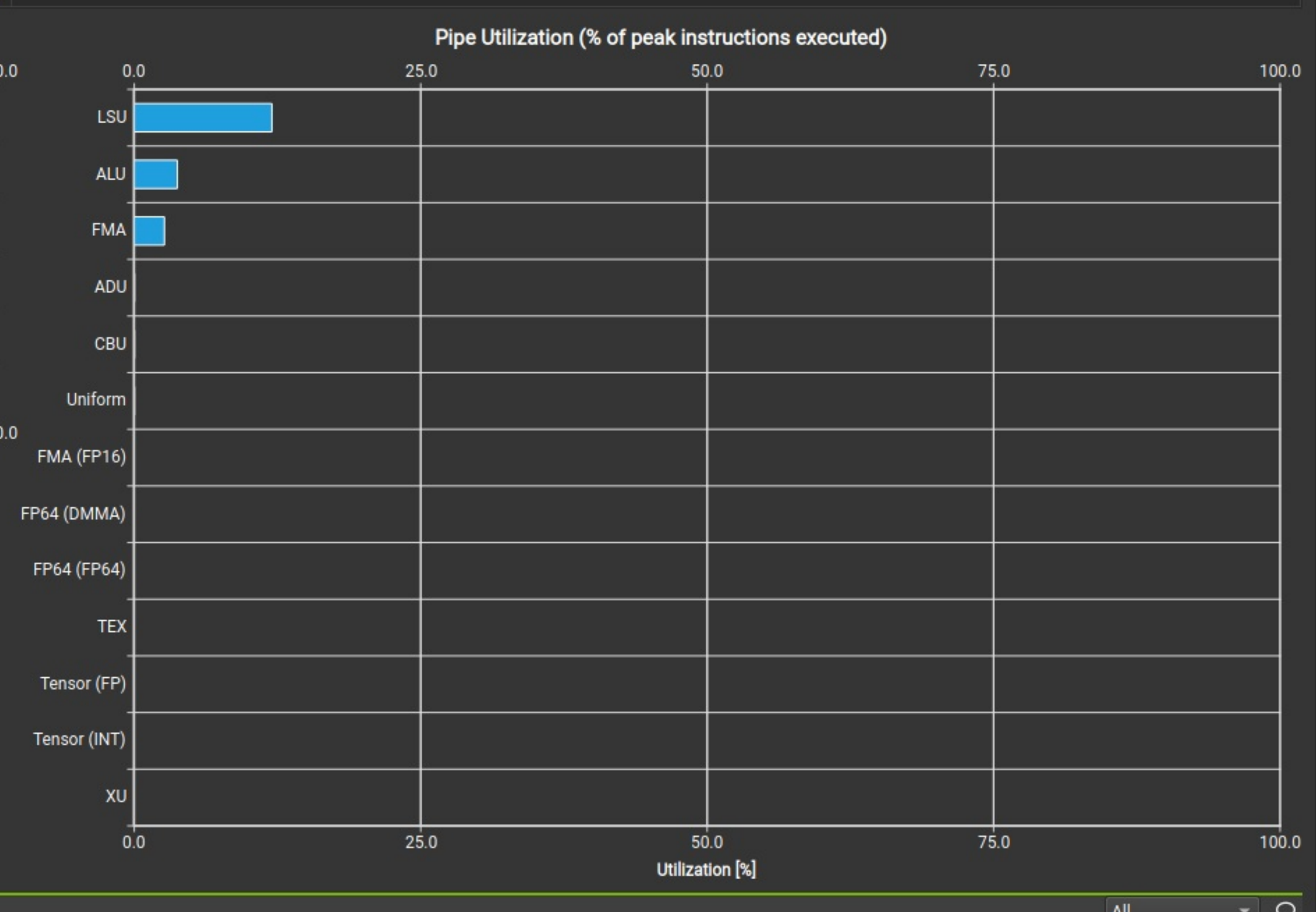
Detailed analysis of the compute resources of the streaming multiprocessors (SM), including the achieved instructions per clock (IPC) and the utilization of each available pipeline. Pipelines with very high utilization might limit the overall performance.

Executed Ipc Elapsed [inst/cycle]	0.25	SM Busy [%]	6.76
Executed Ipc Active [inst/cycle]	0.25	Issue Slots Busy [%]	6.20
Issued Ipc Active [inst/cycle]	0.25		

Pipe Utilization (% of active cycles)



Pipe Utilization (% of peak instructions executed)



Memory Workload Analysis

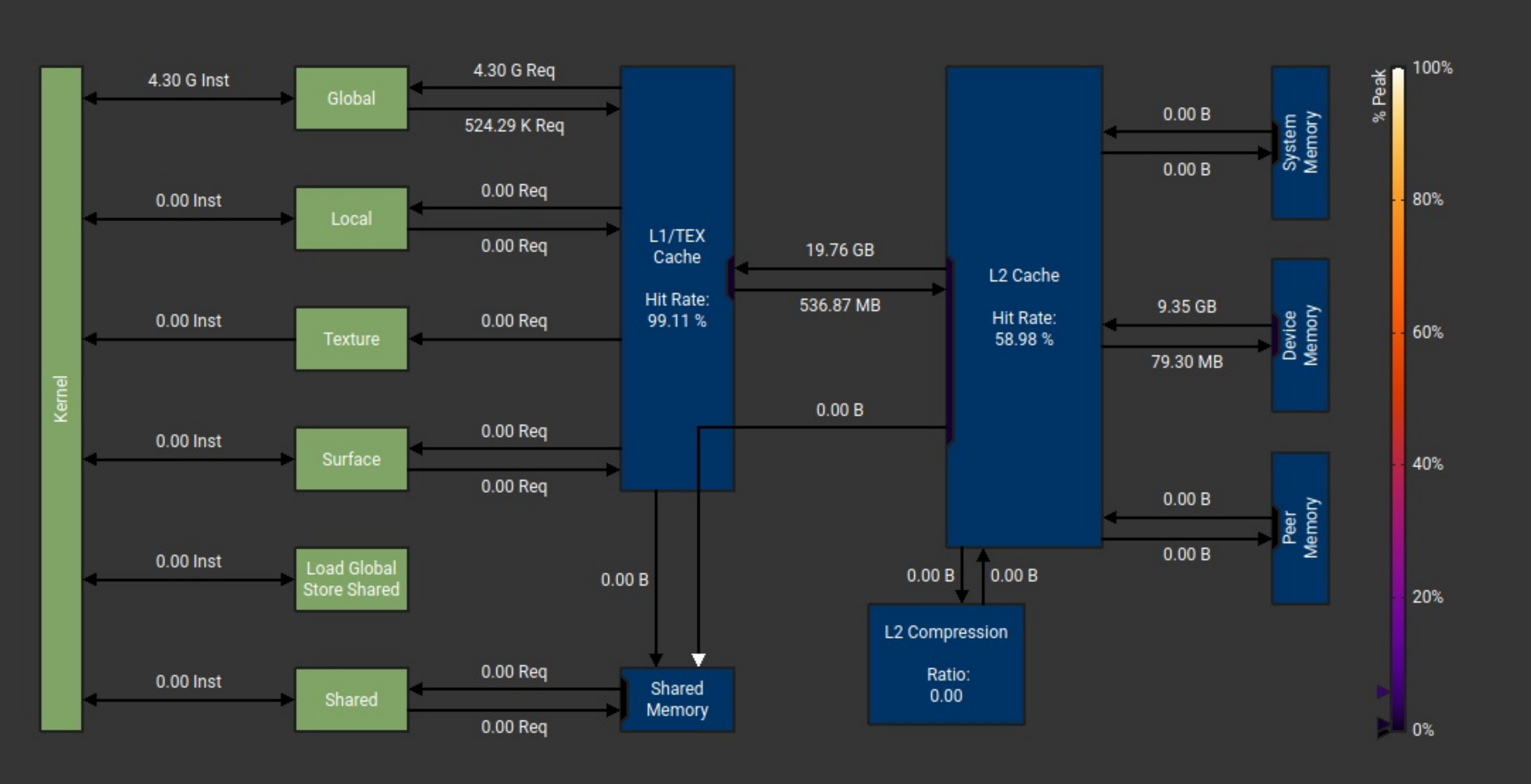
All

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/s]	3.56	Mem Busy [%]	98.99
L1/TEX Hit Rate [%]	99.11	Max Bandwidth [%]	12.00
L2 Hit Rate [%]	58.98	Mem Pipes Busy [%]	12.00
L2 Compression Success Rate [%]	0	L2 Compression Ratio	0

Memory Chart

Show As: Transfer Size



Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	0	0	0	0	0
Shared Load Matrix	0	0			
Shared Store	0	0	0	0	0
Shared Store From Global Load	0	0	0	0	0
Shared Atomic	0	0	0	0	0
Other	-	-	1064960	0.00	0
Total	0	0	1064960	0.00	0

L1/TEX Cache

	Instructions	Requests	Wavefronts	% Peak	Sectors	Sectors/Req	Hit Rate	Bytes	Sector Misses to L2	% Peak to L2	Returns to SM	% Peak to SM
Local Load	0	0	0	0	0	0	0	0			4473239771	6.25
Global Load	4295491584	4295491584			70883373390	16.50	99.13	2268267948480	617100240	0.86		
Global Load To Shared Store (access)	0	0	19331629032	26.99	0	0		0			-	-
Global Load To Shared Store (bypass)	0	0			0	0	-	0	0	0	-	-
Surface Load	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	524288	524288	4243116	0.01	16777216	32	1.08	536870912	16777216	0.02	-	-
Local Store	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	-	-
DSMEM Reduction						-	-		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											
Surface Atomic ALU	0	0	0	0	0	0	0	0	0	0	see above	see above
Surface Atomic CAS	0											
Loads	4295491584	4295491584	19331629032	26.99	70883373390	16.50	99.13	2268267948480	617100240	0.86	4473239771	6.25
Stores	524288	524288	4243116	0.01	16777216	32	1.08	536870912	16777216	0.02	-	-
Atomics & Reductions	0	0	0	0	0	0	0	0	0	0	-	-
Total	4296015872	4296015872	19335872148	27.00	70900150606	16.50	99.11	2268804819392	633877456	0.89	4473239771	6.25

L2 Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes	Throughput	Sector Misses to Device	Sector Misses to System	Sector Misses to Peer
L1/TEX Load	617329381	617342272	1.00	1.12	52.85	19754952704	7446845336.42	290818375	0	0
L1/TEX Store	24102091	16777216	0.70	0.03	100	536870912	202379358.09	0	0	0
L1/TEX Atomic ALU	0	0	0	0	0	0	0	0	0	0
L1/TEX Atomic CAS	0	0	0	0	0	0	0	0	0	0
L1/TEX Reduction	0	0	0	0	0	0	0	0	0	0
L1/TEX Total	633322810	633874587	1.00	1.15	54.05	20283986784	7646270515.03	290885939	0	0
ECC Total	-	0	-	0	-	0	0	0	-	-
GPU Total	642200911	643764739	1.00	1.17	56.51	20600471648	7765572943.58	292156557	40927	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	617290996	52.80	0	0
L1/TEX Store	0	0	0	0	16777216	100	0	0
L1/TEX Atomic	0	0	0	0	0	0	-	-
L1/TEX Total	0	0	0	0	633798132	54.07	0	0
GPU Total	9960216	88.46	0	0	633729702	54.08	0	0

Device Memory

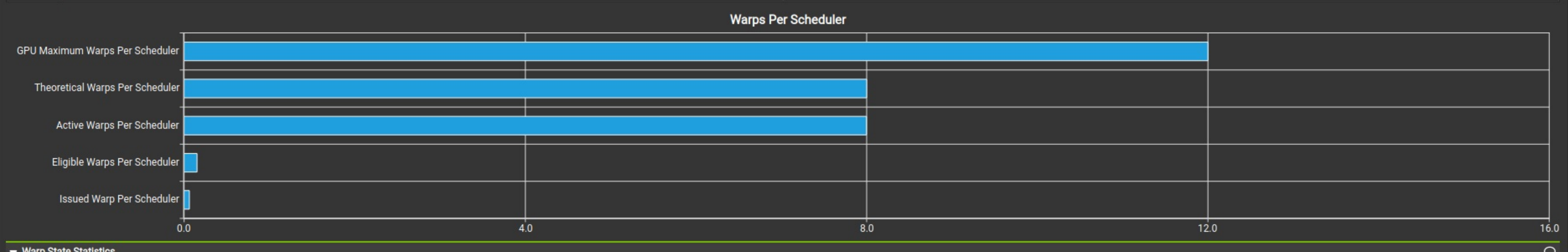
	Sectors	% Peak	Bytes	Throughput
Load	292235640	1.05	9351540480	3525165392.96
Store	2478252	0.01	79304064	29894533.69
Total	294713892	1.06	9430844544	3555059926.66

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]	8.00	No Eligible [%]	93.80
Eligible Warps Per Scheduler [warp]	0.15	One or More Eligible [%]	6.20
Issued Warp Per Scheduler	0.06		

Warps Per Scheduler

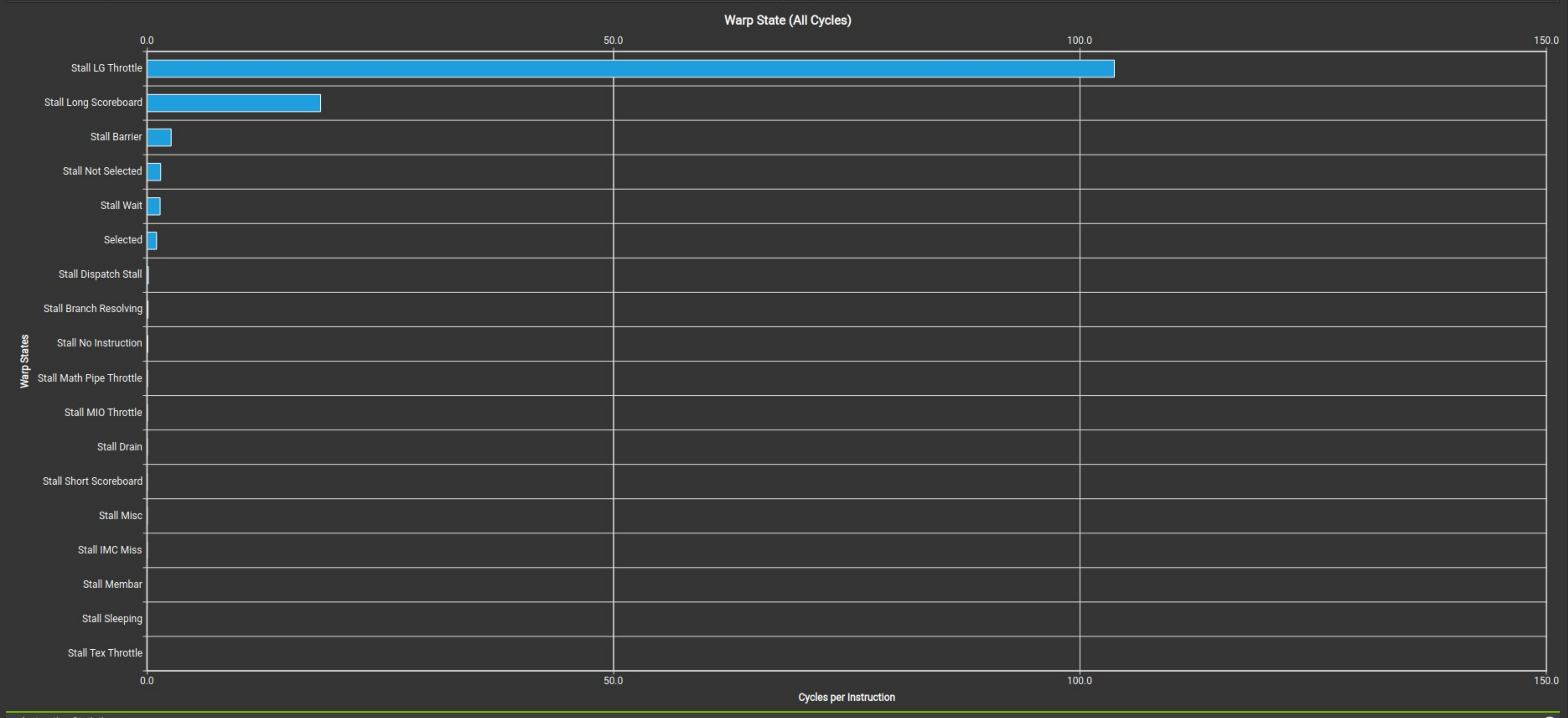


Warp State Statistics

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 schedulers fail to issue every cycle. When executing a kernel with mixed library and user code, these metrics show the combined values.

Warp Cycles Per Issued Instruction [cycle]	129.01	Avg. Active Threads Per Warp	32
Warp Cycles Per Executed Instruction [cycle]	129.01	Avg. Not Predicated Off Threads Per Warp	32.00

Warp State (All Cycles)

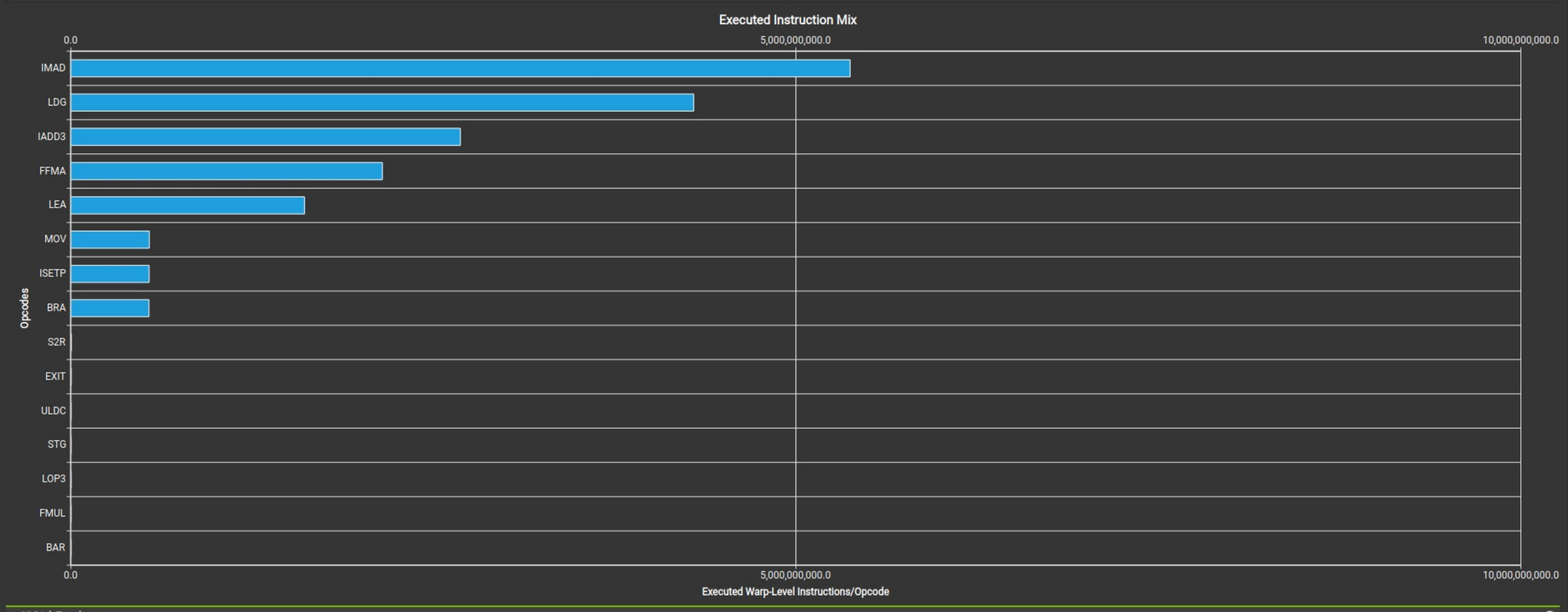


Instruction Statistics

Statistics of the executed low-level assembly instructions (SASS). The instruction mix provides insight into the types and frequency of the executed instructions. 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]	17738760192	Avg. Executed Instructions Per Scheduler [inst]	147823001.60
Issued Instructions [inst]	17738765523	Avg. Issued Instructions Per Scheduler [inst]	147823046.03

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	16384	Function Cache Configuration	CachePreferNone
Registers Per Thread [register/thread]	36	Static Shared Memory Per Block [byte/block]	0
Block Size	1024	Dynamic Shared Memory Per Block [byte/block]	0
Threads [thread]	16777216	Driver Shared Memory Per Block [Kbyte/block]	1.02
Waves Per SM	546.13	Shared Memory Configuration Size [Kbyte]	8.19
Uses Green Context	0	# SMs [SM]	30

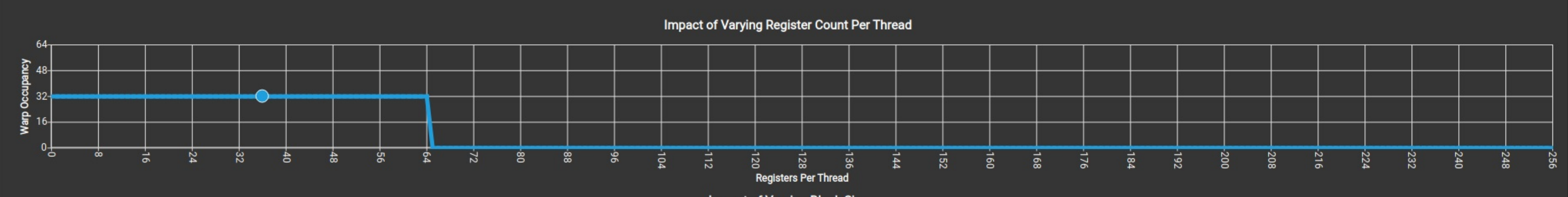
Occupancy

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.

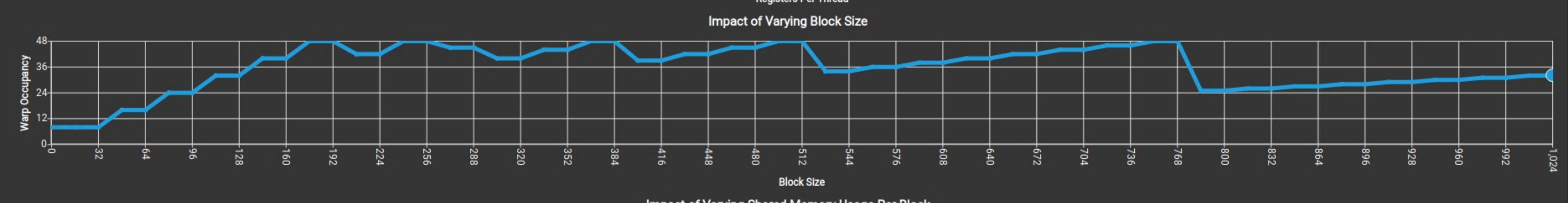
Theoretical Occupancy [%]	66.67	Block Limit Registers [block]	1
Theoretical Active Warps per SM [warp]	32	Block Limit Shared Mem [block]	8
Achieved Occupancy [%]	66.65	Block Limit Warps [block]	1
Achieved Active Warps Per SM [warp]	31.99	Block Limit SM [block]	16

Theoretical Occupancy
Est. Local Speedup: 33.33%
The 8.00 theoretical warps per scheduler this kernel can issue according to its occupancy are below the hardware maximum of 12. This kernel's theoretical occupancy (66.7%) is limited by the number of required registers. This kernel's theoretical occupancy (66.7%) is limited by the number of warps within each block.

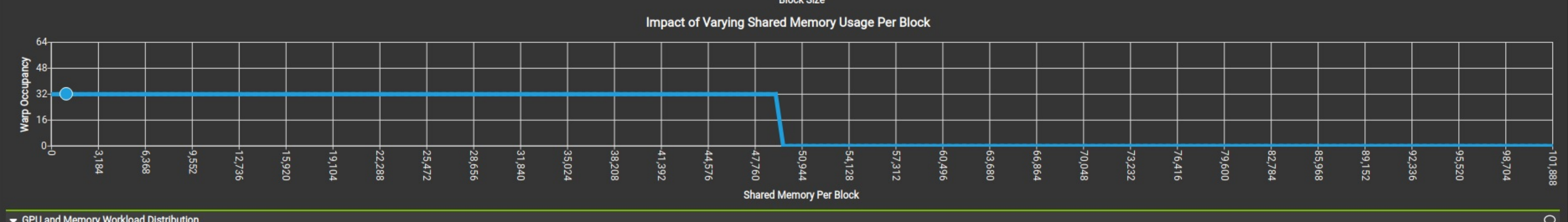
Impact of Varying Register Count Per Thread



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, SMP, SMSP, L1 & L2 caches, and DRAM

Average SM Active Cycles [cycle]	2384490716.17	Average L1 Active Cycles [cycle]	2384490716.17
Average L2 Active Cycles [cycle]	1308009353.38	Average SMSP Active Cycles [cycle]	2384438114.66
Average DRAM Active Cycles [cycle]	196475928	Total SM Elapsed Cycles [cycle]	71623112170
Total L1 Elapsed Cycles [cycle]	71623112170	Total L2 Elapsed Cycles [cycle]	54912851256
Total SMSP Elapsed Cycles [cycle]	286492448680	Total DRAM Elapsed Cycles [cycle]	111321908224

L2 Slices Workload Imbalance
Est. Speedup: 6.49%
One or more L2 Slices have a much higher number of active cycles than the average number of active cycles. Maximum instance value is 11.36% above the average, while the minimum instance value is 2.65% below the average.

Workload Distribution

	Average	Min	Max	Sum
SM Active Cycles	2384490716.17	2383881401	2387403889	71534721485
SMSP Active Cycles	2384438114.66	2383957559	2387579102	286132573759
L1 Active Cycles	2384490716.17	2383881401	2387403889	71534721485
L2 Active Cycles	1308009353.38	1273389669	1475602459	31392224481
DRAM Active Cycles	196475928	196382096	196563088	1178855568

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 couldn't be scheduled. See the documentation for a description of all stall reasons. Only focus on stalls if the schedulers fail to issue every cycle.

Branch Instructions [inst]	539492352	Branch Efficiency [%]	100
Branch Instructions Ratio [%]	0.03	Avg. Divergent Branches	0

Warp Stall Sampling (All Samples)

Location	Value	Value (%)
01_naive_gemm.cu:22 (0x500d8c460 in gemmNaive)	927,359	25
01_naive_gemm.cu:22 (0x500d8c5f0 in gemmNaive)	471,260	13
01_naive_gemm.cu:22 (0x500d8c480 in gemmNaive)	323,830	9
01_naive_gemm.cu:22 (0x500d8c520 in gemmNaive)	322,131	9
01_naive_gemm.cu:22 (0x500d8c4e0 in gemmNaive)	315,645	9

Most Instructions Executed

Location	Value	Value (%)
01_naive_gemm.cu:21 (0x500d8c600 in gemmNaive)	536,870,912	3
01_naive_gemm.cu:22 (0x500d8c5f0 in gemmNaive)	536,870,912	3
01_naive_gemm.cu:22 (0x500d8c5e0 in gemmNaive)	536,870,912	3
01_naive_gemm.cu:22 (0x500d8c5d0 in gemmNaive)	536,870,912	3
01_naive_gemm.cu:21 (0x500d8c5c0 in gemmNaive)	536,870,912	3

Follow the *rules outputs* to get guidance on how to navigate through the report and quickly discover performance bottlenecks in this kernel.
You could also disable individual sections to focus on selected performance aspects and make profiling faster.