	Report	Result	Size	Time	Cycles	GPU	SM Frequency	Process
Current	gemm_fp32_warp_spl_8192_profiling	574 - gemm_warp_specialized_16x16	(256, 256, 1)x(16, 16, 1)	869.70 ms	2,09,37,86,245	0 - NVIDIA GeForce RTX 5060 Ti	2.41 Ghz	[825165] gemm_fp32_warp_spl_8192.exe
Baseline 1	gemm_fp32_shared_mem_8192_profiling	574 - gemm_kernel	(256, 256, 1)x(32, 32, 1)	706.62 ms	1,70,11,84,088	0 - NVIDIA GeForce RTX 5060 Ti	2.41 Ghz	[825056] gemm_fp32_shared_mem_8192.exe

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

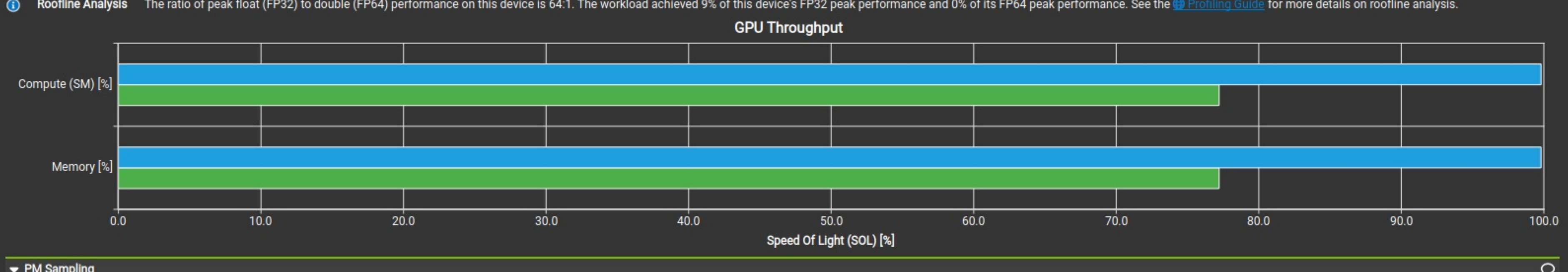
GPU Speed Of Light Throughput

High-level overview of the throughput for compute and memory resources of the GPU. For each unit, the throughput reports the 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 [%]	99.81 (+29.25%)	Duration [ms]	869.70 (+23.08%)
Memory Throughput [%]	99.81 (+29.25%)	Elapsed Cycles [cycle]	2,09,37,86,245 (+23.08%)
L1/TEX Cache Throughput [%]	99.83 (+29.19%)	SM Active Cycles [cycle]	2,09,20,01,743.14 (+23.14%)
L2 Cache Throughput [%]	7.18 (-18.62%)	SM Frequency [Ghz]	2.41 (-0.00%)
DRAM Throughput [%]	18.17 (-18.63%)	DRAM Frequency [Ghz]	13.79 (+0.00%)

High Throughput This workload 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 workloads in the Compute Workload Analysis section.

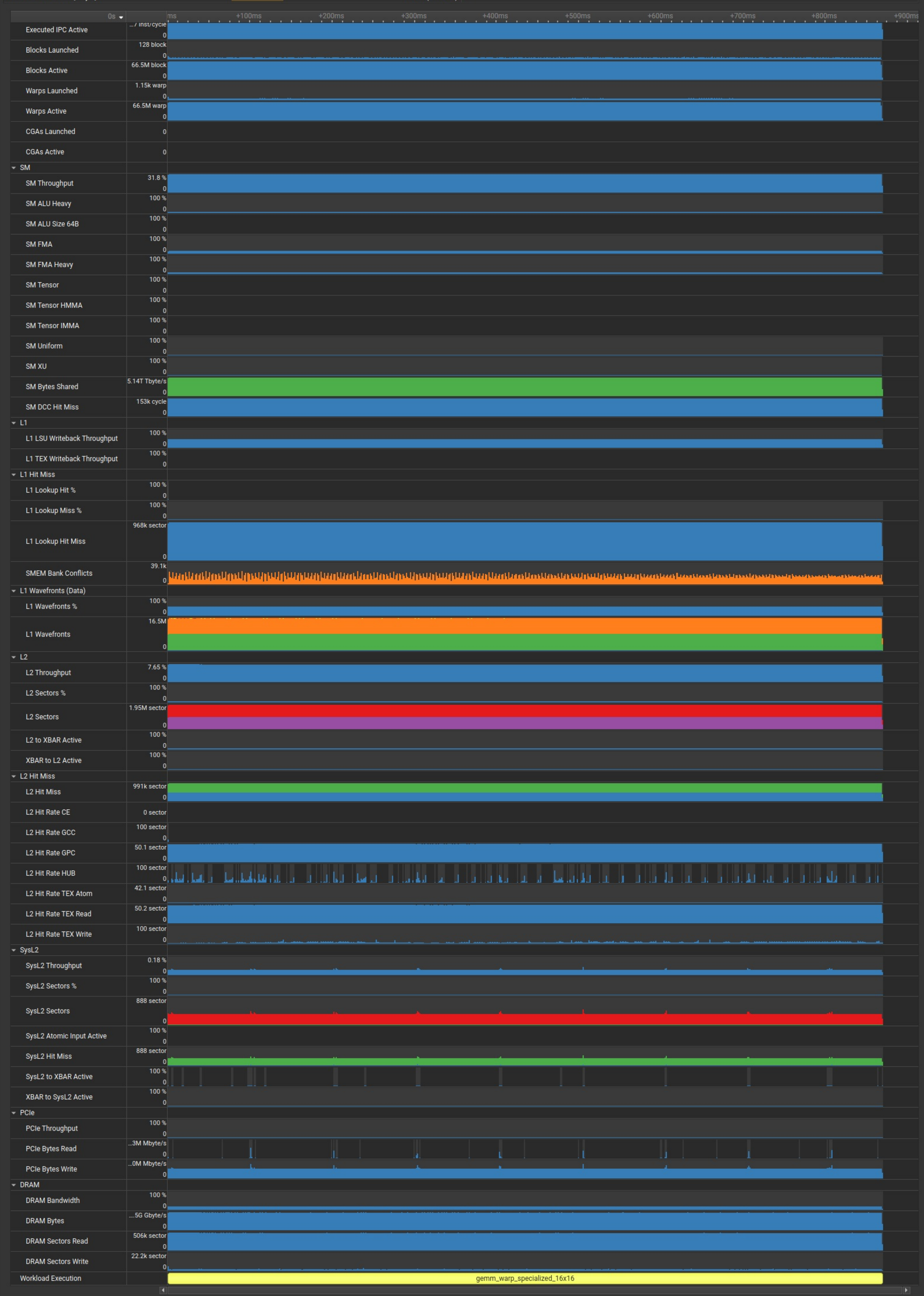
Roofline Analysis The ratio of peak float (FP32) to double (FP64) performance on this device is 64:1. The workload achieved 9% of this device's FP32 peak performance and 0% of its FP64 peak performance. See the Profiling Guide for more details on roofline analysis.



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 [us]	192 (+0.00%)	# Pass Groups	1 (+0.00%)
Maximum Buffer Size [Mbyte]	53.28 (+23.00%)	-	-



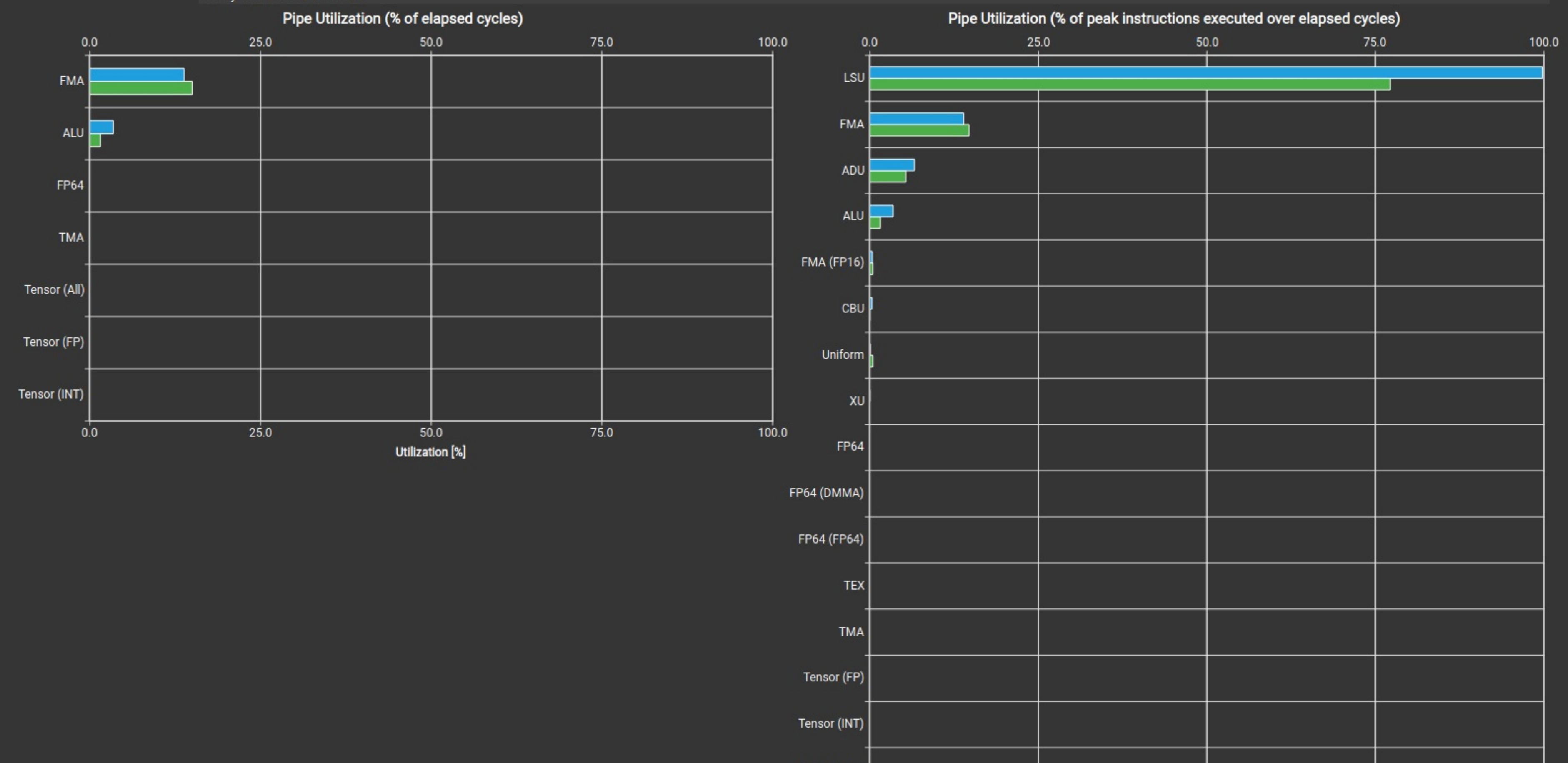
Compute Workload Analysis

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]	1.27 (+16.50%)	SM Busy [%]	35.47 (+28.93%)
Executed Ipc Active [inst/cycle]	1.27 (+16.44%)	Issue Slots Busy [%]	31.72 (+16.50%)
Issued Ipc Active [inst/cycle]	1.27 (+16.44%)		

Low Utilization
Est. Local Speedup: 86.20%
All compute pipelines are under-utilized. Either this workload is very small or it doesn't issue enough warps per scheduler. Check the Launch Statistics and Scheduler Statistics sections for further details.

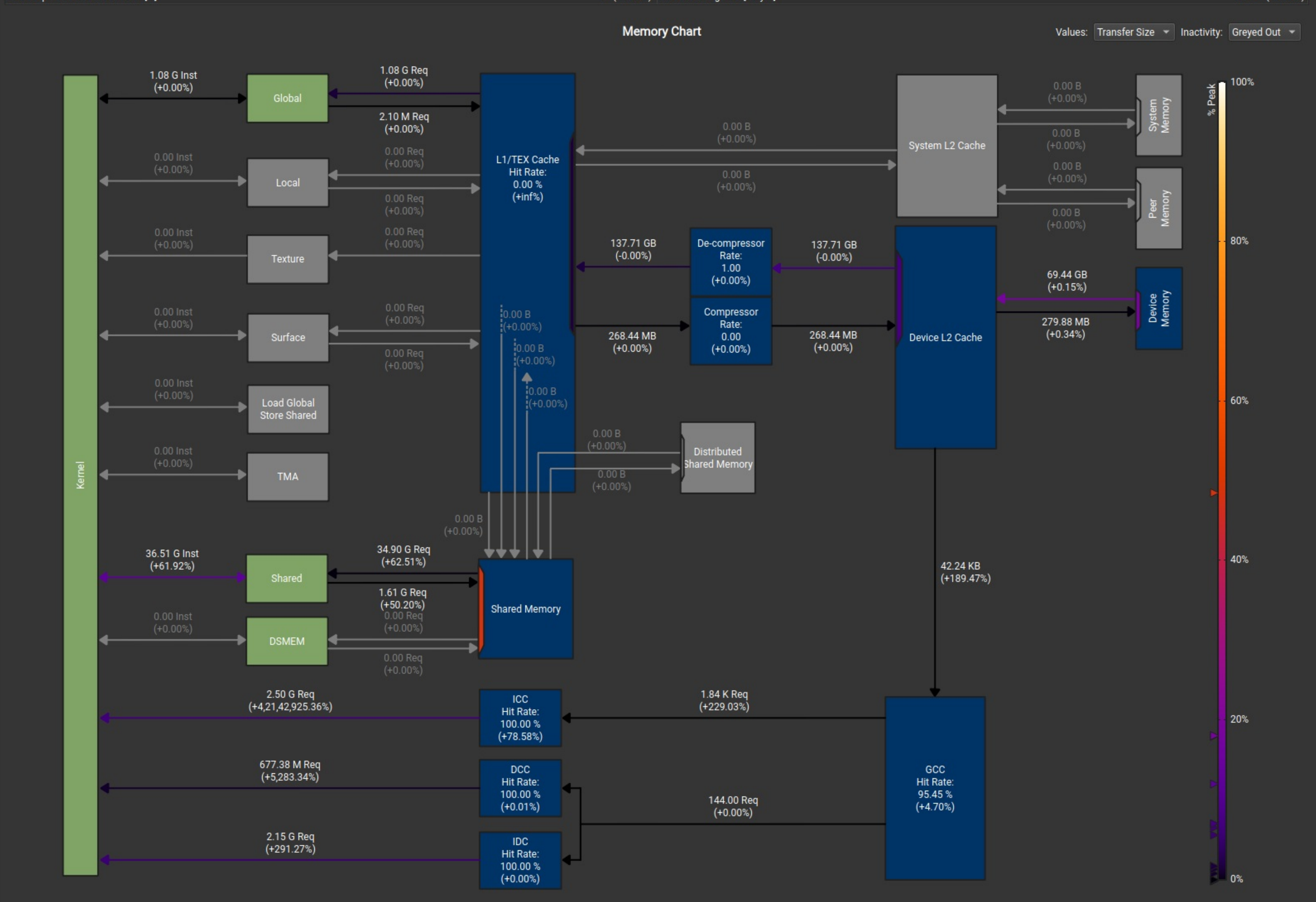
Key Performance Indicators



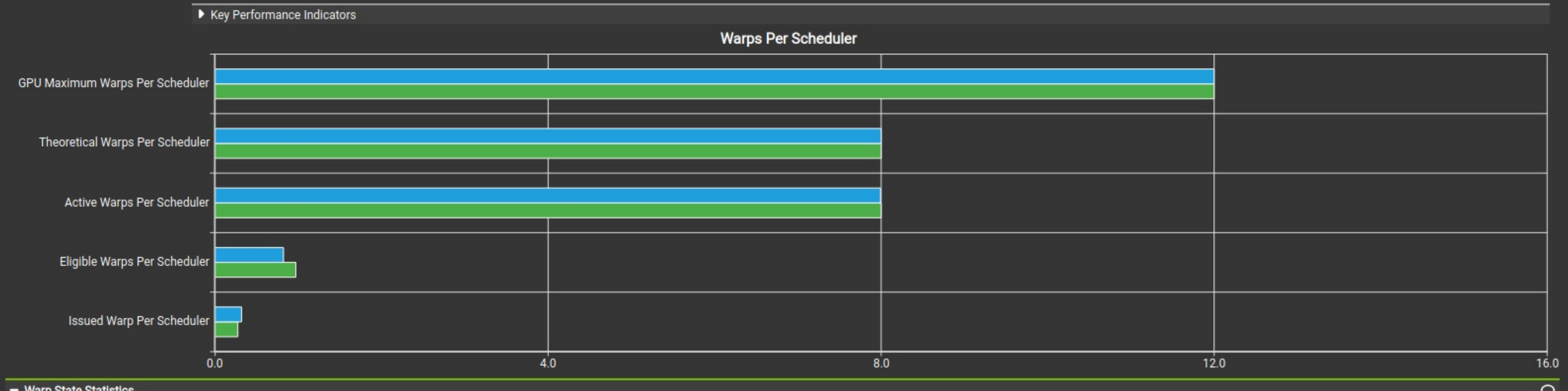
Memory Workload Analysis

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 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.

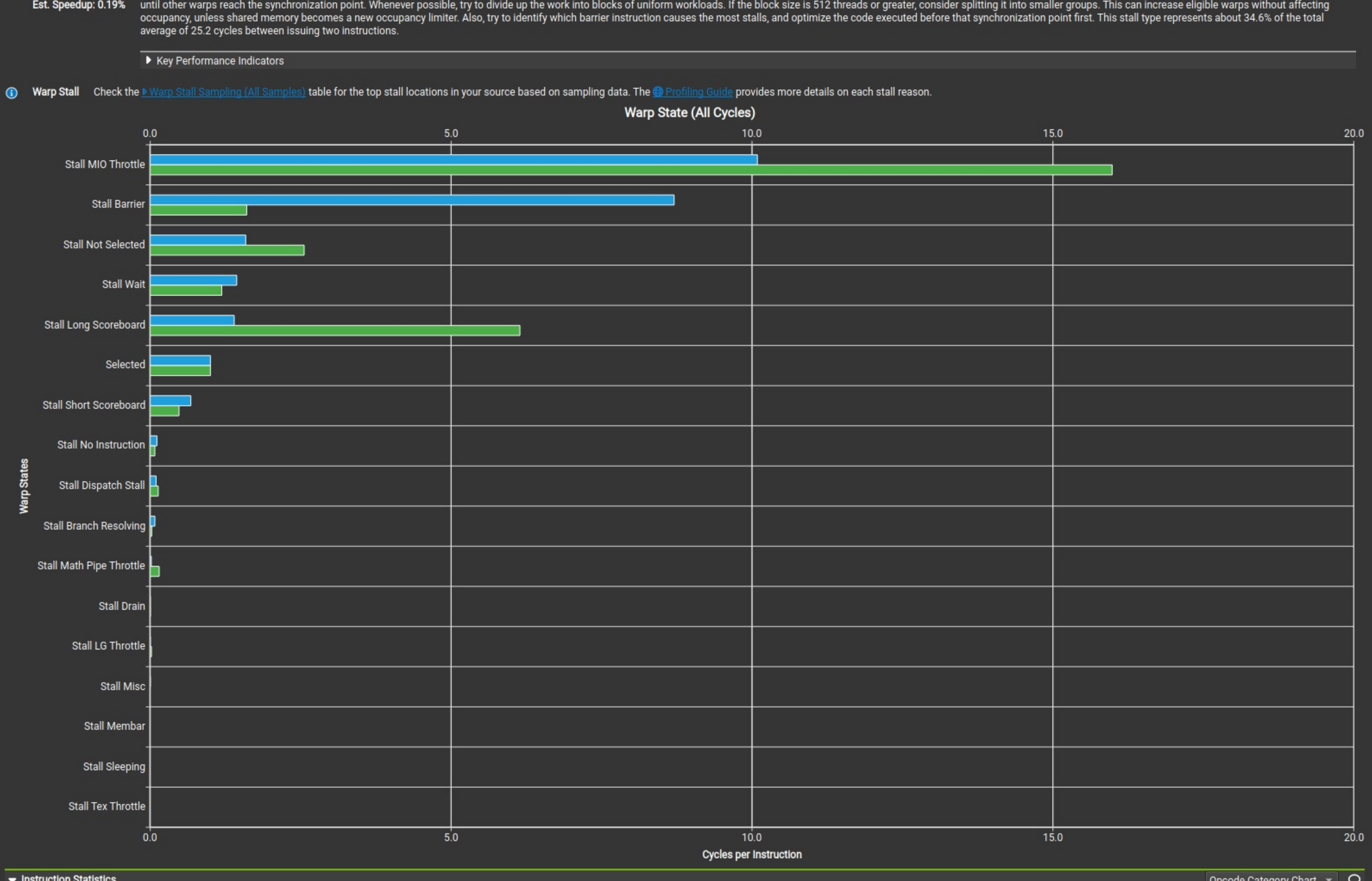
Memory Throughput [Gbyte/s]	80.17 (-18.63%)	Mem Busy [%]	50.04 (+8.45%)
L1/TEX Hit Rate [%]	0.00 (+inf%)	Max Bandwidth [%]	99.81 (+29.25%)
L2 Hit Rate [%]	49.48 (-0.26%)	Mem Pipes Busy [%]	99.81 (+29.25%)
L2 Compression Input Sectors [sector]	0 (+0.00%)	Local Memory Spilling Requests	0 (+0.00%)
L2 Compression Ratio	0 (+0.00%)	Local Memory Spilling Request Overhead [%]	0 (+0.00%)
L2 Compression Success Rate [%]	0 (+0.00%)	L2 Persisting Size [Mbyte]	6.29 (+0.00%)



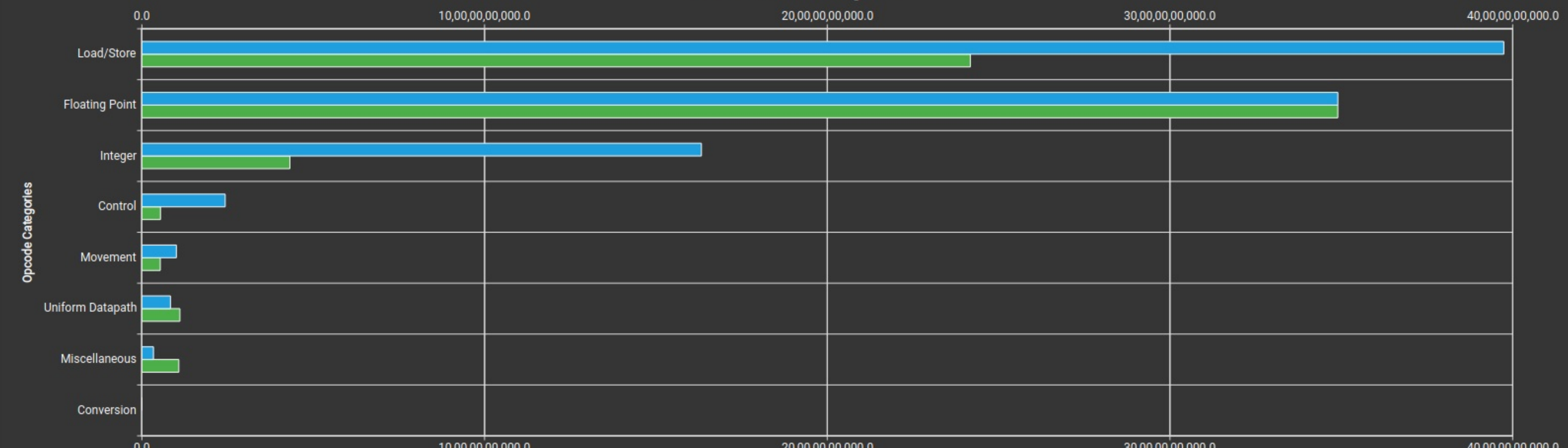
Scheduler Statistics



Warp State Statistics



Instruction Statistics



NVLink Topology

NVLink Tables

NUMA Affinity

Launch Statistics

Occupancy

GPU and Memory Workload Distribution

Source Counters