Result	Size	Time	Cycles	GPU	SM Frequency	Process	Attributes
Current 559 - brute_force_coarsening	(32768, 1, 1)x(32, 1, 1)	2.22 ms	1,297,349	0 - Tesla T4	584.99 Mhz	[10345] exe	

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

Compare | Tools | View | Export

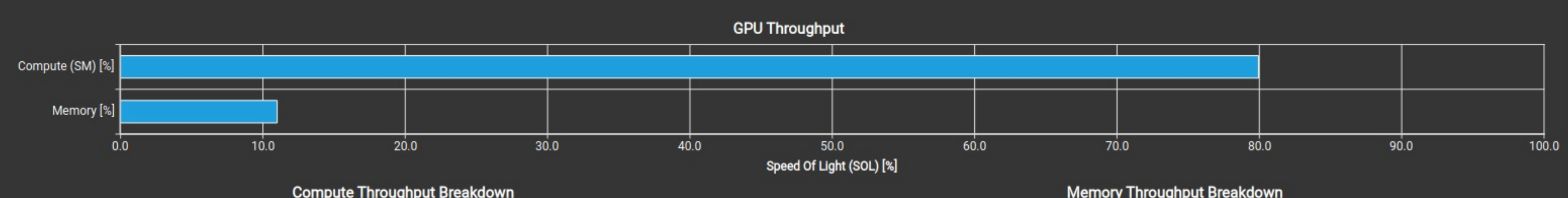
GPU Speed Of Light Throughput

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 [%]	79.94	Duration [ms]	2.22
Memory Throughput [%]	11.00	Elapsed Cycles [cycle]	1,297,349
L1/TEX Cache Throughput [%]	11.85	SM Active Cycles [cycle]	1,292,228.20
L2 Cache Throughput [%]	0.65	SM Frequency [Mhz]	584.99
DRAM Throughput [%]	1.97	DRAM Frequency [Ghz]	4.99

High Compute Throughput Compute is more heavily utilized than Memory: Look at the Compute Workload Analysis section to see what the compute pipelines are spending their time doing. Also, consider whether any computation is redundant and could be reduced or moved to look-up tables.

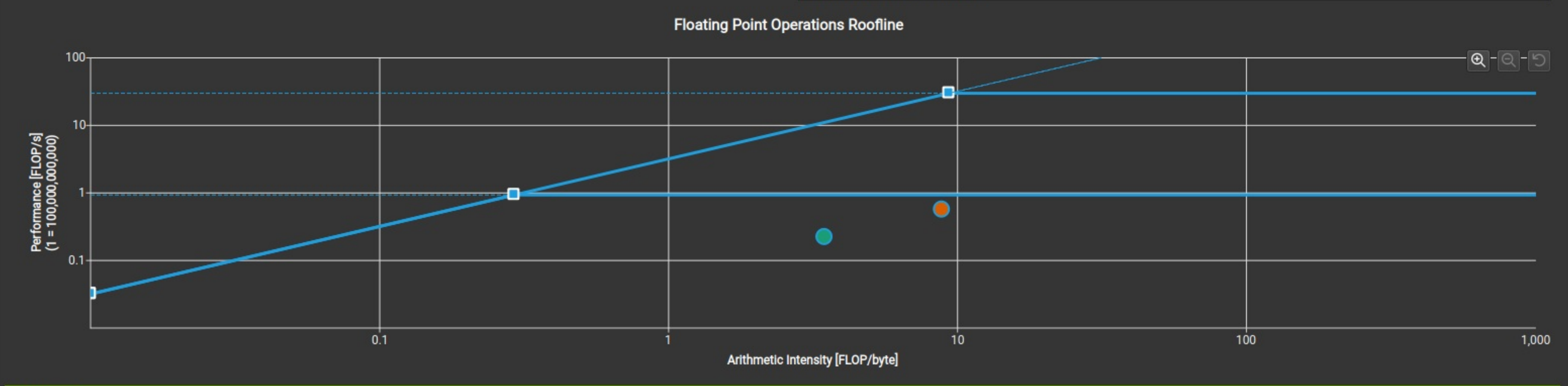


Compute Throughput Breakdown

SM: Pipe Fp64 Cycles Active [%]	79.94
SM: Pipe Alu Cycles Active [%]	48.50
SM: Issue Active [%]	43.31
SM: Inst Executed [%]	43.31
SM: Mio Pq Write Cycles Active [%]	31.18
SM: Inst Executed Pipe Adu [%]	23.44
IDC: Request Cycles Active [%]	15.80
SM: Pipe Fma Cycles Active [%]	15.22
SM: Mio Inst Issued [%]	14.81
SM: Inst Executed Pipe Lsu [%]	11.00
SM: Mio Pq Read Cycles Active [%]	10.69
SM: Inst Executed Pipe Cbu Pred On Any [%]	8.54
SM: Mio2rf Writeback Active [%]	5.29
SM: Inst Executed Pipe Xu [%]	0.13
SM: Inst Executed Pipe Uniform [%]	0.03
SM: Pipe Tensor Cycles Active [%]	0
SM: Inst Executed Pipe Fp16 [%]	0
SM: Pipe Shared Cycles Active [%]	0
SM: Memory Throughput Internal Activity [%]	0
SM: Instruction Throughput Internal Activity [%]	0
SM: Inst Executed Pipe Tex [%]	0
SM: Inst Executed Pipe Ipa [%]	0

Memory Throughput Breakdown

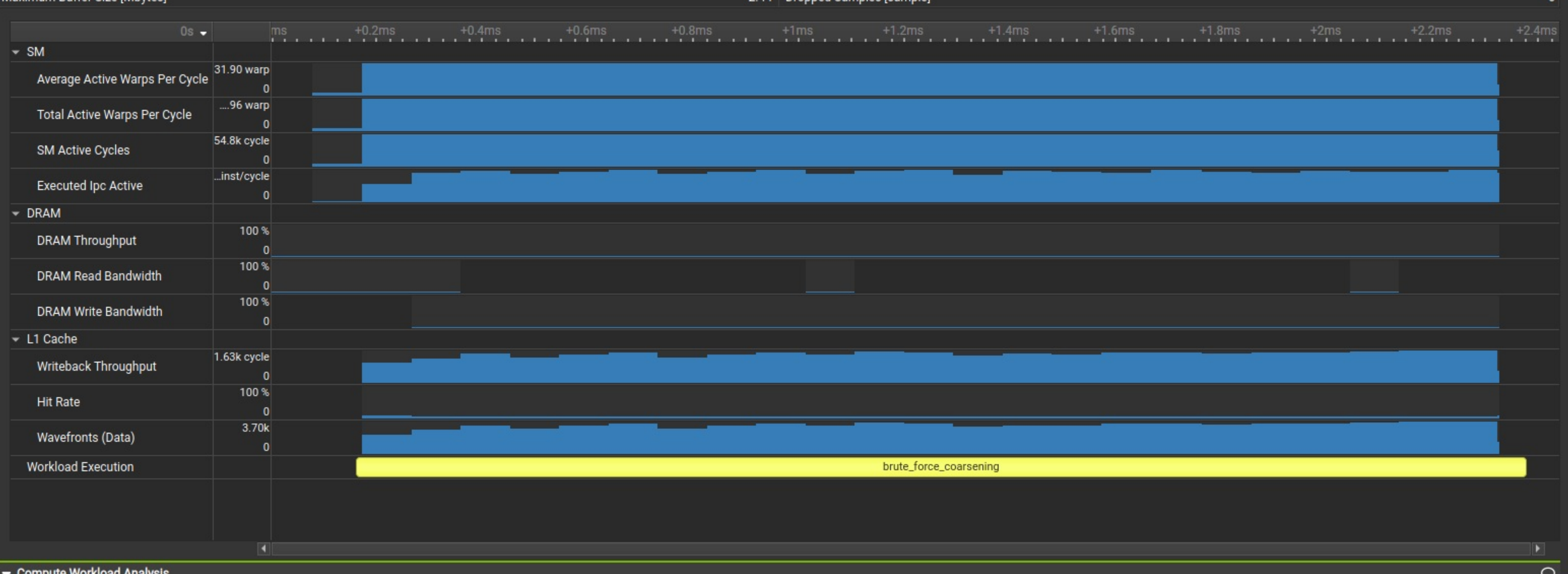
L1: Lsuin Requests [%]	11.00
L1: Data Pipe Lsu Wavefronts [%]	5.93
L1: Lsu Writeback Active [%]	2.62
DRAM: Cycles Active [%]	1.97
DRAM: Dram Sectors [%]	1.44
L1: M L1tex2xbar Req Cycles Active [%]	0.76
L1: Data Bank Reads [%]	0.69
L2: T Sectors [%]	0.65
L2: Xbar2lts Cycles Active [%]	0.65
L1: Data Bank Writes [%]	0.40
L2: D Sectors [%]	0.32
GPU: Compute Memory Access Throughput Internal Activity [%]	0.17
L2: T Tag Requests [%]	0.16
L2: Lts2xbar Cycles Active [%]	0.01
L2: D Sectors Fill Device [%]	0.00
L1: Texin Sm2tex Req Cycles Active [%]	0.00
L1: F Wavefronts [%]	0.00
L2: D Atomic Input Cycles Active [%]	0
L1: Tex Writeback Active [%]	0
L2: D Sectors Fill Sysmem [%]	0
L1: M Xbar2l1tex Read Sectors [%]	0
GPU: Compute Memory Request Throughput Internal Activity [%]	0
L1: Data Pipe Tex Wavefronts [%]	0



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.

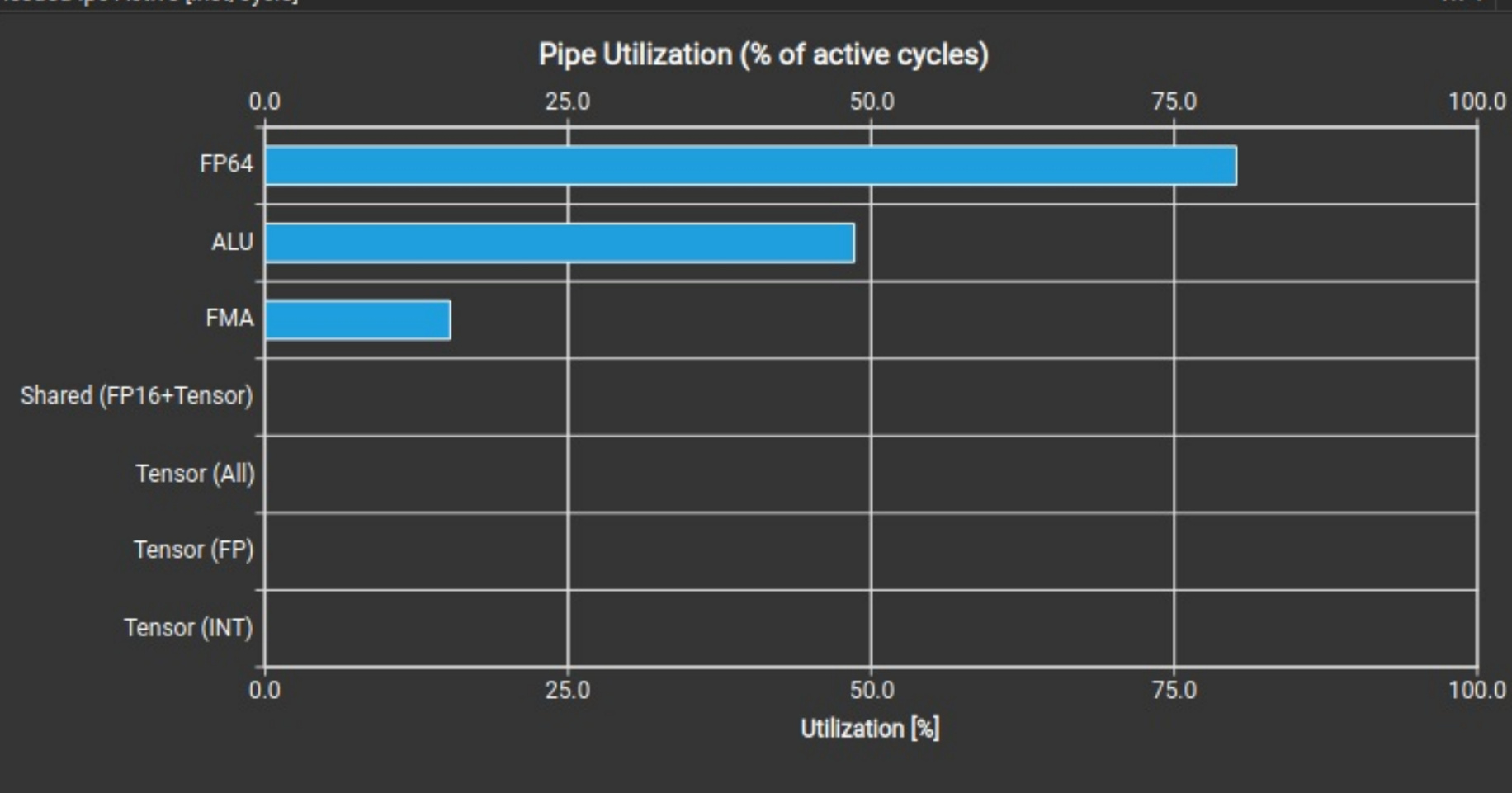
Maximum Sampling Interval [cycle]	80,000	# Pass Groups	1
Maximum Buffer Size [Mbytes]	2.44	Dropped Samples [sample]	0

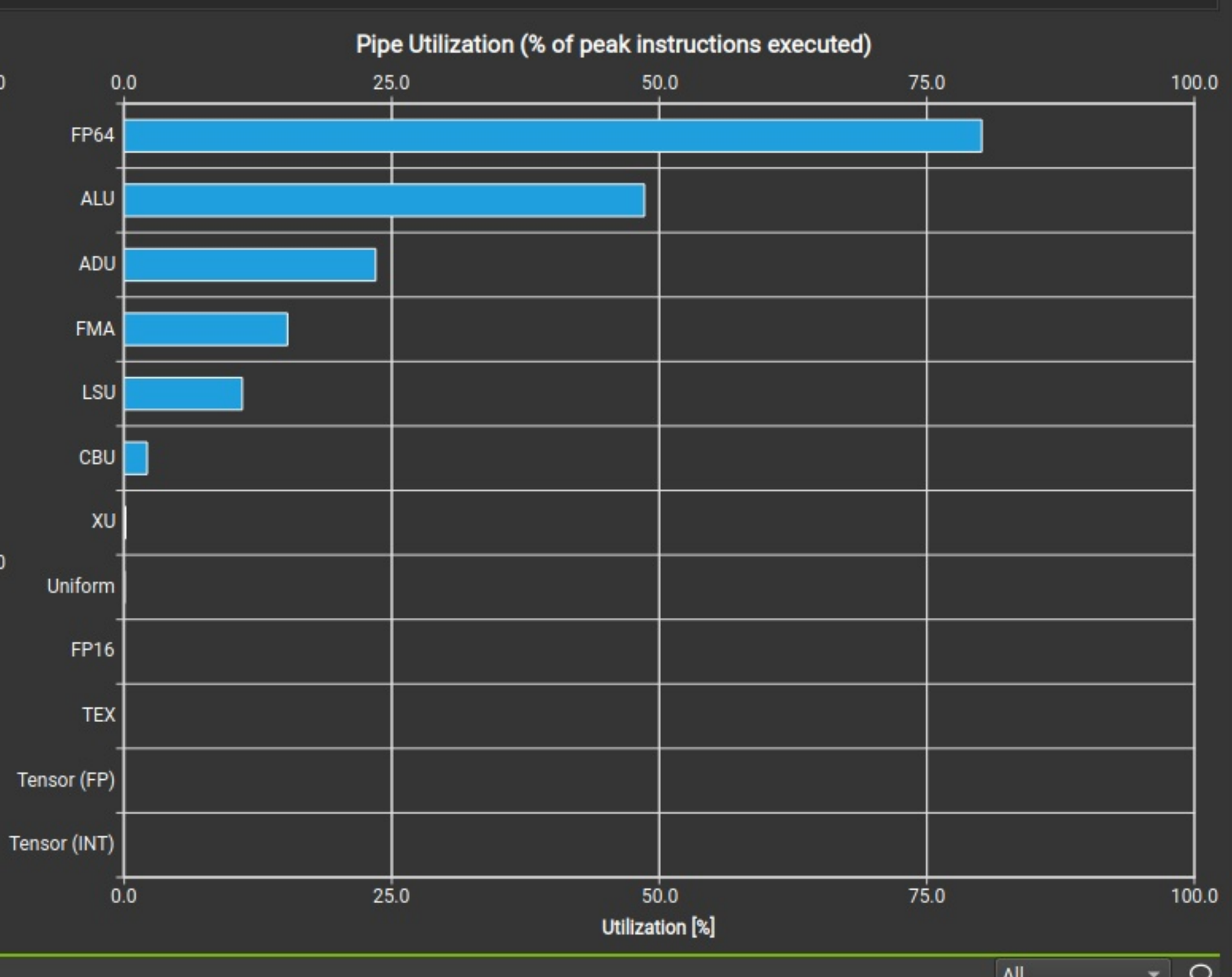


Compute Workload Analysis

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]	1.73	SM Busy [%]	80.13
Executed Ipc Active [inst/cycle]	1.74	Issue Slots Busy [%]	43.41
Issued Ipc Active [inst/cycle]	1.74		





Memory Workload Analysis

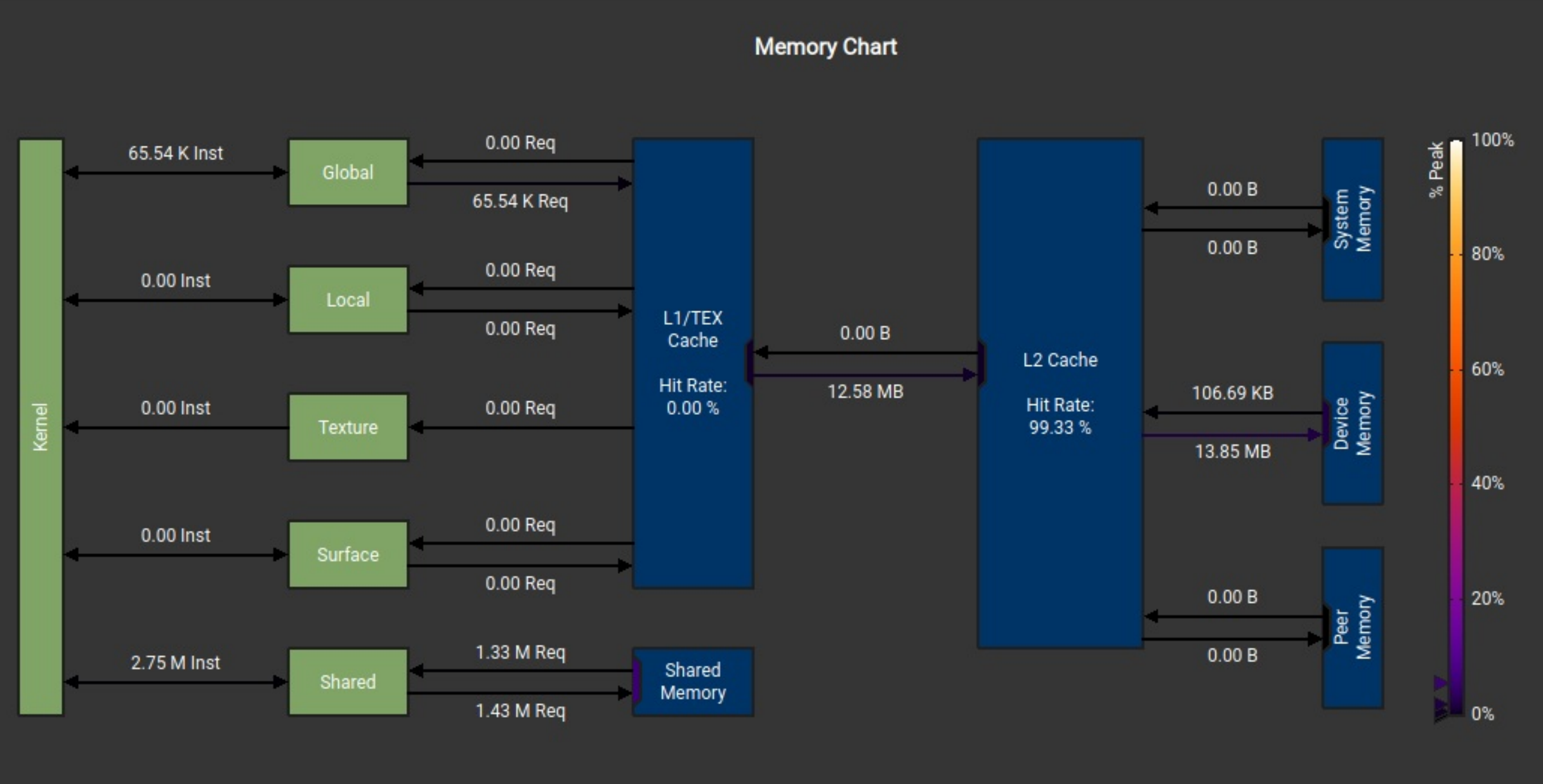
All

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]	6.29	Mem Busy [%]	5.93
L1/TEX Hit Rate [%]	0	Max Bandwidth [%]	11.00
L2 Hit Rate [%]	99.33	Mem Pipes Busy [%]	31.18

Memory Chart

Show As: Transfer Size



Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	1,327,104	1,327,104	1,327,104	2.56	0
Shared Load Matrix	0	0			
Shared Store	1,425,408	1,425,408	1,600,727	3.09	0
Shared Atomic	0	0	0	0	0
Other	-	-	32,513	0.06	0
Total	2,752,512	2,752,512	2,960,344	5.71	0

L1/TEX Cache

	Instructions	Requests	Wavefronts	% Peak	Sectors	Sectors/Req	Hit Rate	Bytes	Sector Misses to L2	% Peak to L2	Returns to SM
Global Load	0	0	0	0	0	0	0	0	0	0	0
Surface Load	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	65,536	65,536	65,536	0.13	393,216	6	0	12,582,912	393,216	0.76	-
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
Global Atomic CAS	0										
Surface Atomic ALU	0	0	0	0	0	0	0	0	0	0	see above
Surface Atomic CAS	0										
Loads	0	0	0	0	0	0	0	0	0	0	0
Stores	65,536	65,536	65,536	0.13	393,216	6	0	12,582,912	393,216	0.76	-
Atomics & Reductions	0	0	0	0	0	0	0	0	0	0	-
Total	65,536	65,536	65,536	0.13	393,216	6	0	12,582,912	393,216	0.76	0

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	0	0	0	0	0	0	0	0	0	0
L1/TEX Store	98,304	393,216	4	0.65	100	12,582,912	5,673,785,062.91	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	98,304	393,216	4	0.65	100	12,582,912	5,673,785,062.91	0	0	0
ECC Total	-	104	-	0.00	-	3,328	1,500,634.88	104	-	-
GPU Total	99,882	397,354	3.98	0.65	99.75	12,715,328	5,733,493,016.28	66	0	0

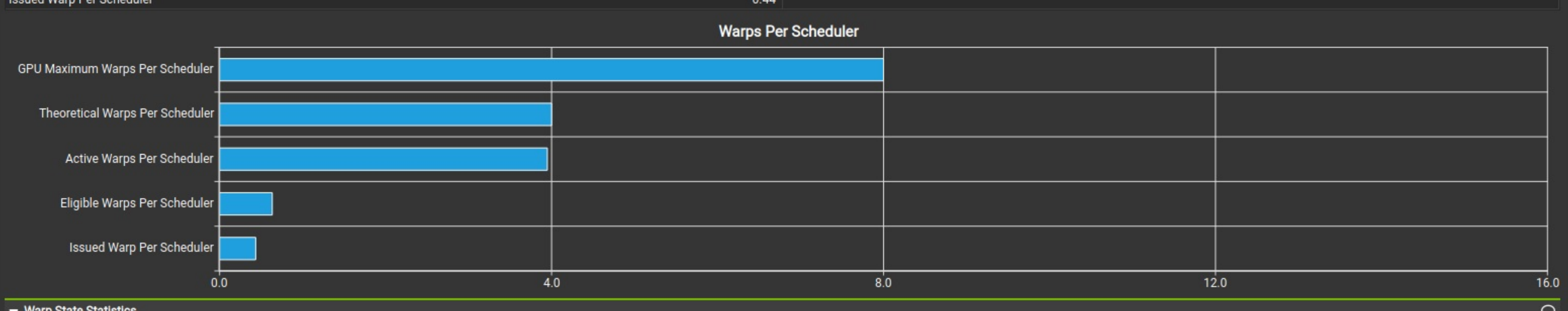
Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	3,334	0.02	106,688	48,106,891.38
Store	432,932	1.96	13,853,824	6,246,854,438.42
Total	436,266	1.97	13,960,512	6,294,961,329.79

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.

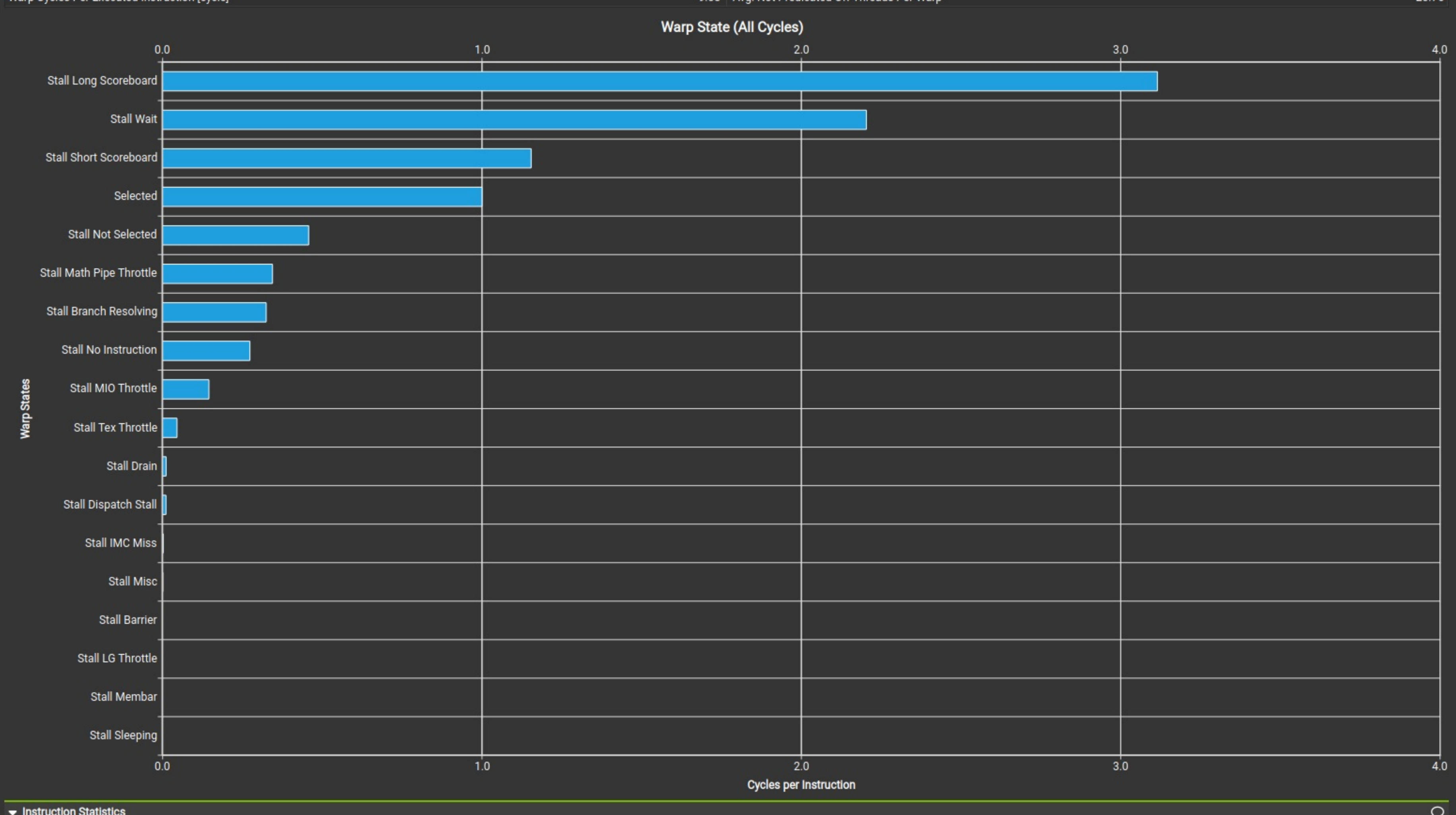
Active Warps Per Scheduler [warp]	3.95	No Eligible [%]	56.50
Eligible Warps Per Scheduler [warp]	0.63	One or More Eligible [%]	43.50
Issued Warp Per Scheduler	0.44		



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.

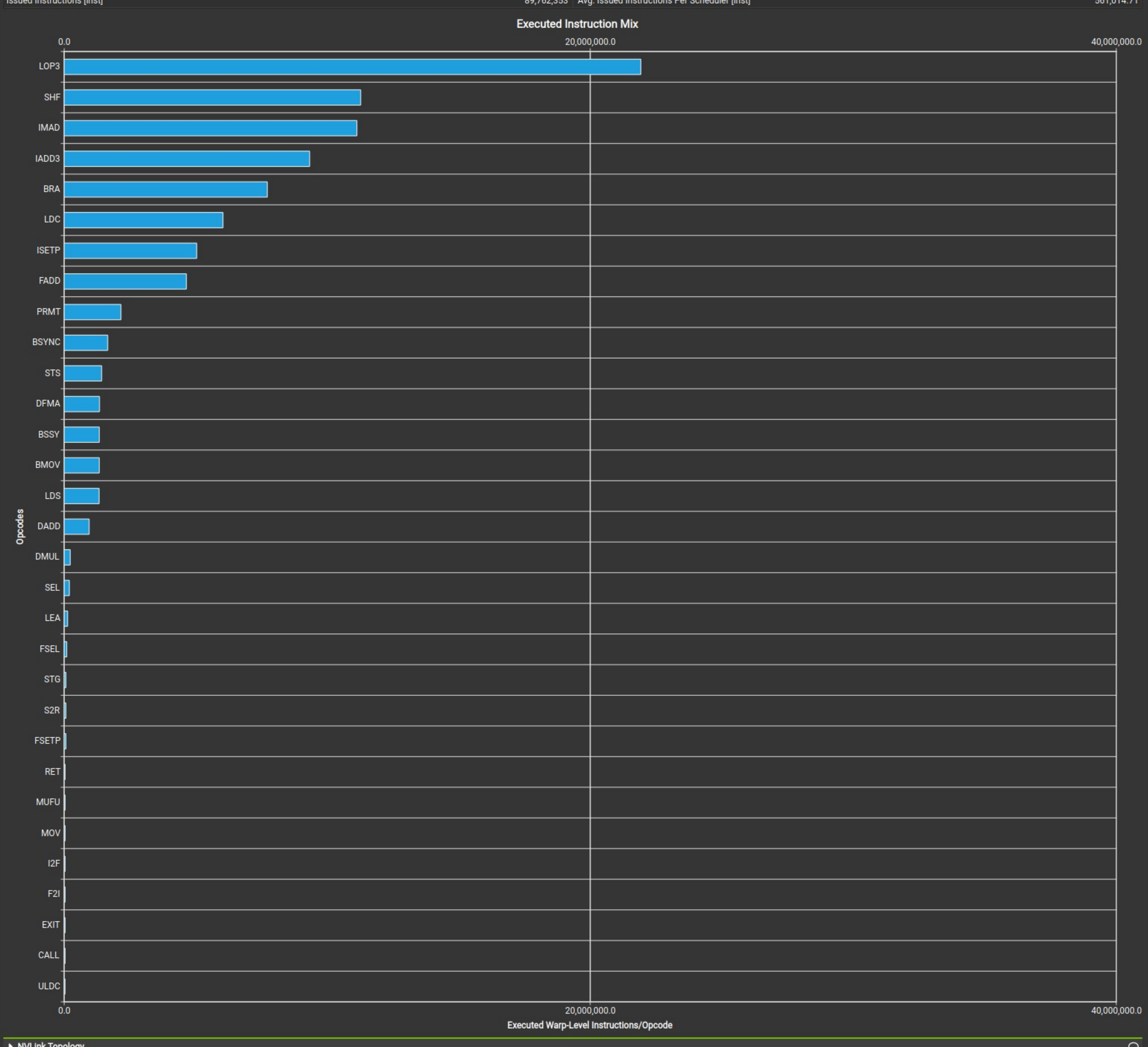
Warp Cycles Per Issued Instruction [cycle]	9.08	Avg. Active Threads Per Warp	22.66
Warp Cycles Per Executed Instruction [cycle]	9.08	Avg. Not Predicated Off Threads Per Warp	20.76



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]	89,751,072	Avg. Executed Instructions Per Scheduler [inst]	560,944.20
Issued Instructions [inst]	89,762,353	Avg. Issued Instructions Per Scheduler [inst]	561,014.71



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.

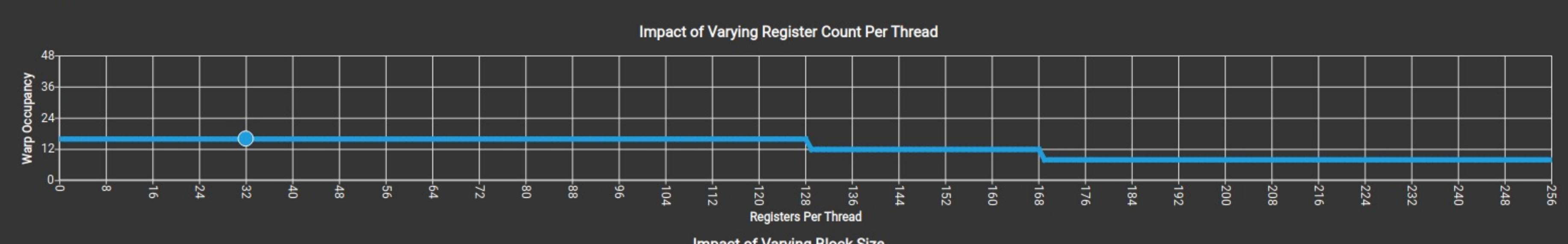
Grid Size	32,768	Function Cache Configuration	CachePreferNone
Registers Per Thread [register/thread]	32	Static Shared Memory Per Block [byte/block]	0
Block Size	32	Dynamic Shared Memory Per Block [byte/block]	96
Threads [thread]	1,048,576	Driver Shared Memory Per Block [byte/block]	0
Waves Per SM	51.20	Shared Memory Configuration Size [Kbyte]	32.77
Uses Green Context	0	# SMs [SM]	40

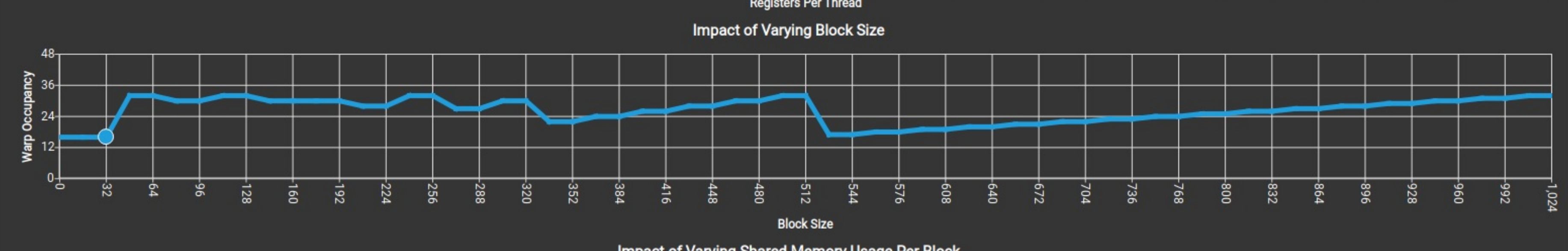
Occupancy

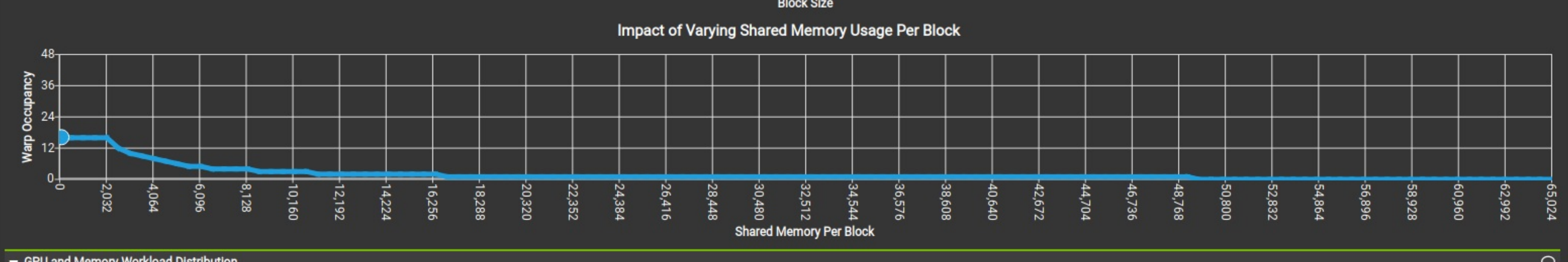
Occupancy is the ratio of the number of active warps per multiprocessor to the maximum number of possible active warps. Another way to view occupancy is the percentage of the hardware's ability to process warps that is actively in use. Higher occupancy does not always result in higher performance, however, 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 [%]	50	Block Limit Registers [block]	64
Theoretical Active Warps per SM [warp]	16	Block Limit Shared Mem [block]	128
Achieved Occupancy [%]	49.36	Block Limit Warps [block]	32
Achieved Active Warps Per SM [warp]	15.80	Block Limit SM [block]	16

Theoretical Occupancy
Est. Local Speedup: 50.00% The 4.00 theoretical warps per scheduler this kernel can issue according to its occupancy are below the hardware maximum of 8. This kernel's theoretical occupancy (50.0%) is limited by the number of blocks that can fit on the SM.







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]	1,292,228.20	Average L1 Active Cycles [cycle]	1,292,228.20
Average L2 Active Cycles [cycle]	110,013.47	Average SMSP Active Cycles [cycle]	1,289,550.25
Average DRAM Active Cycles [cycle]	218,133	Total SM Elapsed Cycles [cycle]	51,811,304
Total L1 Elapsed Cycles [cycle]	51,811,304	Total L2 Elapsed Cycles [cycle]	60,675,744
Total SMSP Elapsed Cycles [cycle]	207,245,216	Total DRAM Elapsed Cycles [cycle]	88,504,320

Workload Distribution

	Average	Min	Max	Sum
SM Active Cycles	1,292,228.20	1,288,438	1,295,930	51,689,128
SMSP Active Cycles	1,289,550.25	1,284,866	1,293,940	206,328,040
L1 Active Cycles	1,292,228.20	1,288,438	1,295,930	51,689,128
L2 Active Cycles	110,013.47	109,320	110,665	3,520,431
DRAM Active Cycles	218,133	217,076	219,440	1,745,064

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]	9,478,336	Branch Efficiency [%]	95.88
Branch Instructions Ratio [%]	0.11	Avg. Divergent Branches	2,004.40

Warp Stall Sampling (All Samples)

Location	Value	Value (%)
kernels.cu:82 (0x7930d70335b0 in brute_f...	1,089	2
kernels.cu:82 (0x7930d7033640 in brute_f...	903	2
kernels.cu:120 (0x7930d7034800 in brute_...	861	2
kernels.cu:82 (0x7930d70336c0 in brute_f...	856	2
kernels.cu:120 (0x7930d70347a0 in brute_...	823	2

Most Instructions Executed

Location	Value	Value (%)
kernels.cu:118 (0x7930d7034810 in brute_...	1,021,952	1
kernels.cu:120 (0x7930d7034800 in brute_...	1,021,952	1
kernels.cu:120 (0x7930d70347f0 in brute_...	1,021,952	1
kernels.cu:120 (0x7930d70347e0 in brute_...	1,021,952	1
kernels.cu:118 (0x7930d70347d0 in brute_...	1,021,952	1

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.