Result	Size	Time	Cycles	GPU	SM Frequency	Process	Attributes
Current 1790 - gemm1dBlockTiling	(64, 64, 1)x(512, 1, 1)	85.72 ms	77,151,538	0 - NVIDIA GeForce RTX 3060 Laptop GPU	900.00 Mhz	[105627] 04_1d_block_tiling	

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

GPU Speed Of Light Throughput

All

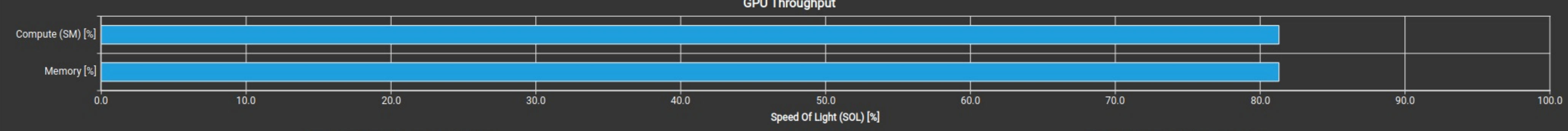
High-level overview of the throughput for compute and memory resources of the GPU. For each unit, the throughput reports the achieved percentage of utilization with respect to the theoretical maximum. Breakdowns show the throughput for each individual sub-metric of Compute and Memory to clearly identify 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 [%]	81.30	Duration [ms]	85.72
Memory Throughput [%]	81.30	Elapsed Cycles [cycle]	77151538
L1/TEX Cache Throughput [%]	81.53	SM Active Cycles [cycle]	76922855.97
L2 Cache Throughput [%]	15.24	SM Frequency [Mhz]	900.00
DRAM Throughput [%]	16.87	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 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 kernel achieved 23% 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.

GPU Throughput



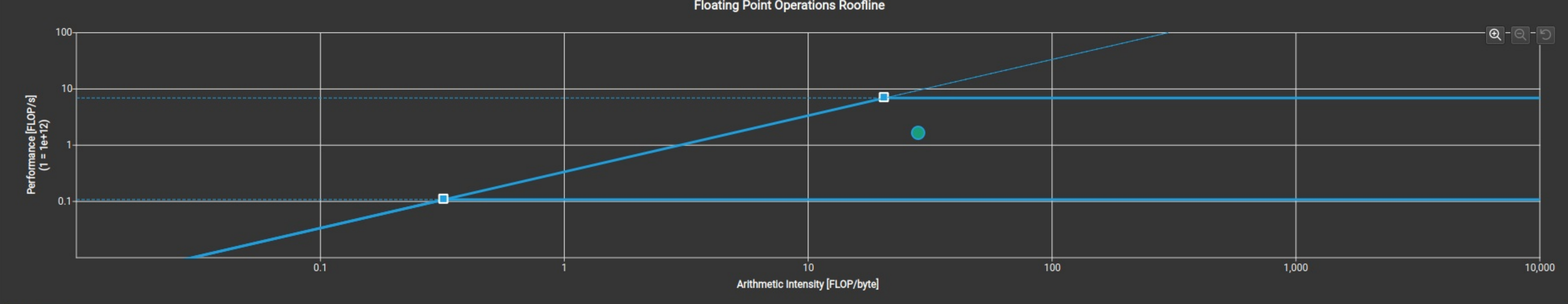
Compute Throughput Breakdown

SM: Inst Executed Pipe Lsu [%]	81.30
SM: Issue Active [%]	38.15
SM: Inst Executed [%]	38.15
SM: Mio2rf Writeback Active [%]	30.46
SM: Mio Inst Issued [%]	29.04
SM: Pipe Fma Cycles Active [%]	23.21
SM: Pipe Fmaheavy Cycles Active [%]	19.23
SM: Inst Executed Pipe Adu [%]	5.81
SM: Pipe Alu Cycles Active [%]	4.44
SM: Inst Executed Pipe Uniform [%]	2.92
SM: Mio Pq Write Cycles Active [%]	2.92
SM: Mio Pq Read Cycles Active [%]	2.92
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.00
SM: Pipe Tensor Cycles Active [%]	0
SM: Pipe Fp64 Cycles Active [%]	0
SM: Memory Throughput Internal Activity [%]	0
IDC: Request Cycles Active [%]	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

Memory Throughput Breakdown

L1: Lsuin Requests [%]	81.30
L1: Data Pipe Lsu Wavefronts [%]	69.07
L1: Lsu Writeback Active [%]	60.93
DRAM: Dram Sectors [%]	16.87
DRAM: Cycles Active [%]	16.79
L2: T Sectors [%]	15.24
L2: Lts2xbar Cycles Active [%]	15.13
L1: M Xbar2l1tex Read Sectors [%]	11.69
L1: Data Bank Reads [%]	10.17
GPU: Compute Memory Access Throughput Internal Activity [%]	9.44
L2: T Tag Requests [%]	9.44
L2: Xbar2lts Cycles Active [%]	9.33
L2: D Sectors Fill Device [%]	8.23
L1: M L1tex2xbar Req Cycles Active [%]	7.22
L2: D Sectors [%]	5.92
L1: Data Bank Writes [%]	2.92
L1: Texin Sm2tex Req Cycles Active [%]	0.18
L2: D Sectors Fill Sysmem [%]	0.00
L1: Tex Writeback Active [%]	0
L2: D Atomic Input Cycles Active [%]	0
L1: F Wavefronts [%]	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]	1.02	# Pass Groups	2
Maximum Buffer Size [Mbytes]	19.19	Dropped Samples [sample]	0

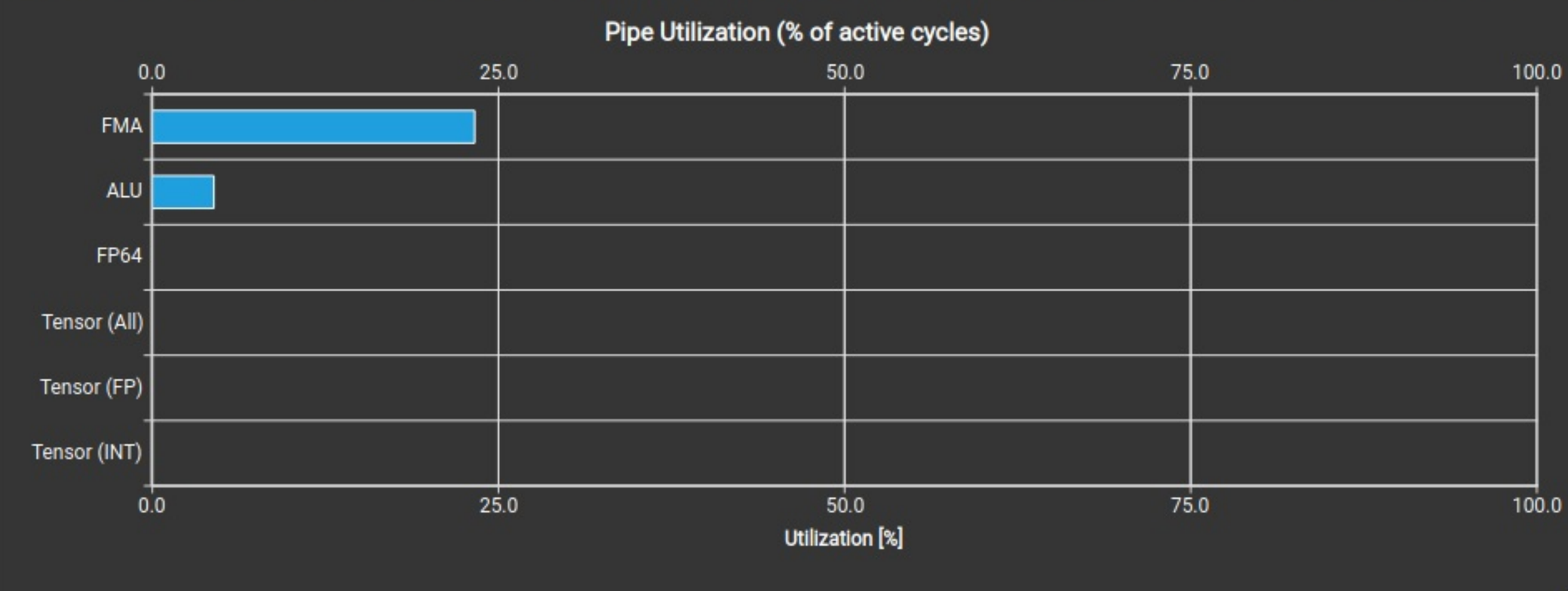
Compute Workload Analysis

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.

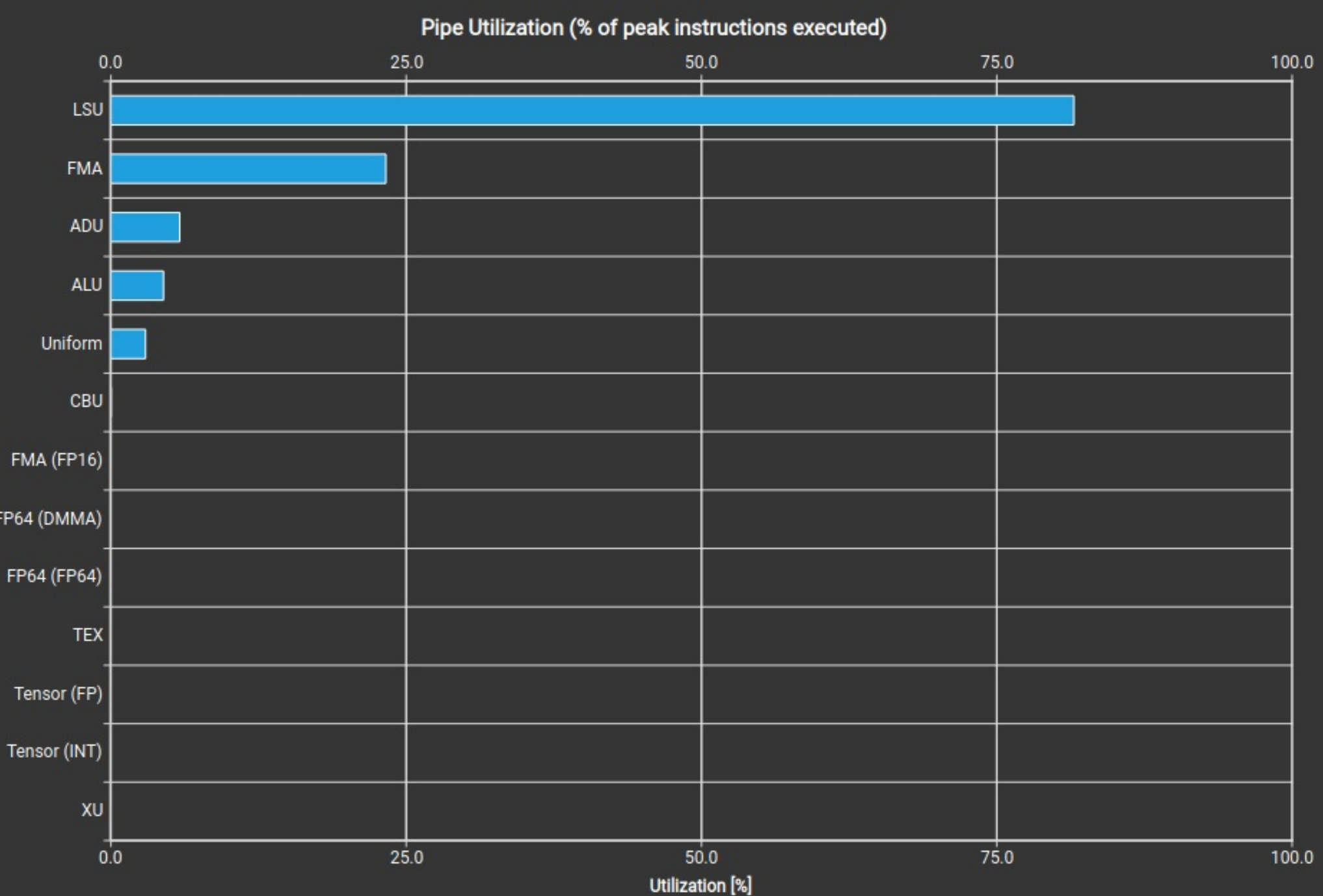
Executed Ipc Elapsed [inst/cycle]	1.53	SM Busy [%]	38.25
Executed Ipc Active [inst/cycle]	1.53	Issue Slots Busy [%]	38.25
Issued Ipc Active [inst/cycle]	1.53		

Balanced FMA is the highest-utilized pipeline (23.3%) based on active cycles, taking into account the rates of its different instructions. It executes 32-bit floating point (FADD, FMUL, FMAD, ...) and integer (IMUL, IMAD) operations. It is well-utilized, but should not be a bottleneck.

Pipe Utilization (% of active cycles)



Pipe Utilization (% of peak instructions executed)



Memory Workload Analysis

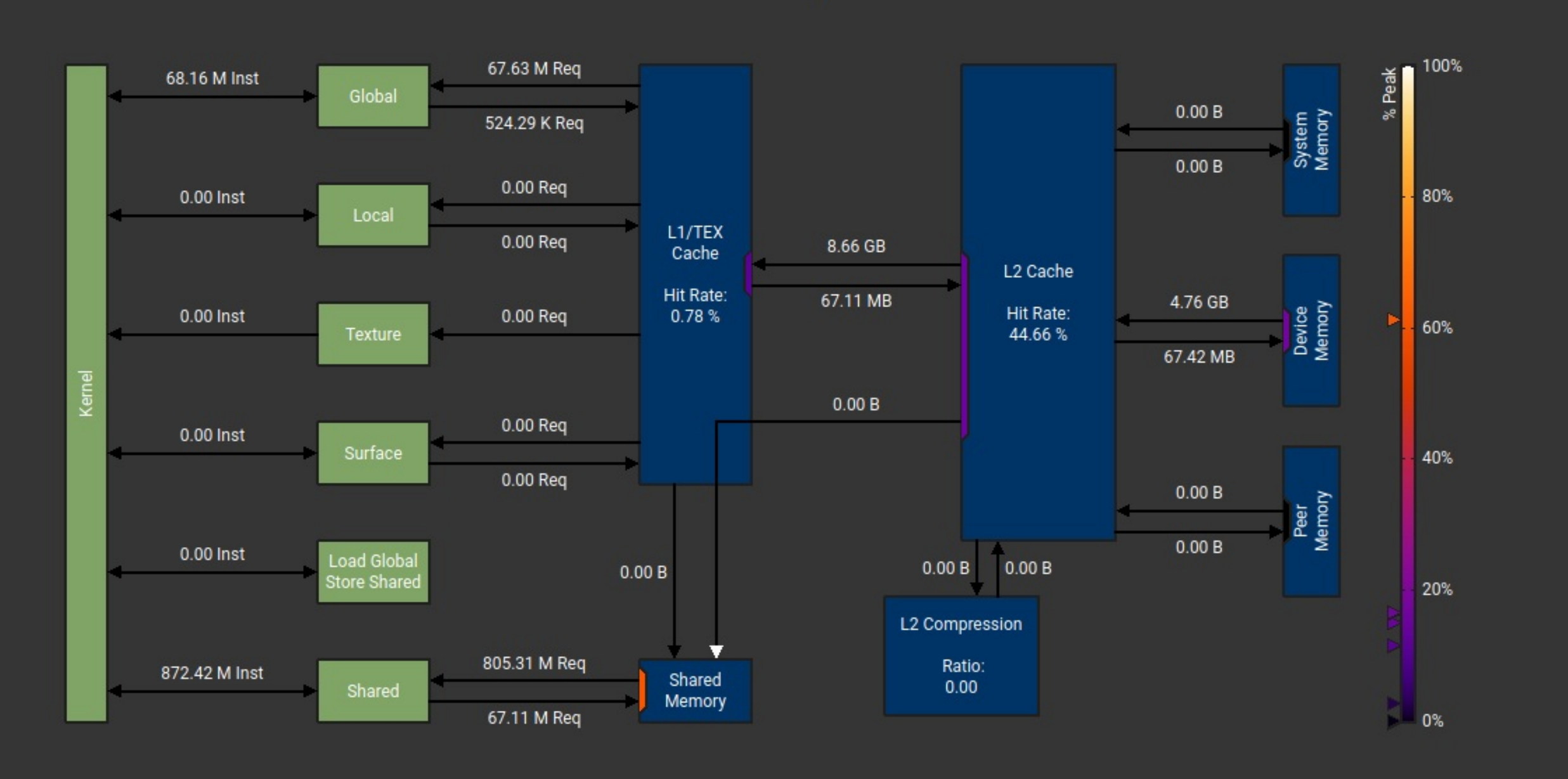
All

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]	56.36	Mem Busy [%]	69.07
L1/TEX Hit Rate [%]	0.78	Max Bandwidth [%]	81.30
L2 Hit Rate [%]	44.66	Mem Pipes Busy [%]	81.30
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	805306368	805306368	1342301236	58.00	79789
Shared Load Matrix	0	0			
Shared Store	67108864	67108864	67108864	0.72	0
Shared Store From Global Load	0	0	0	0	0
Shared Atomic	0	0	0	0	0
Other	-	-	11630599	2.68	7296252
Total	872415232	872415232	1421040699	61.41	7376041

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			67721428	2.93
Global Load	67633152	67633152			270462061	4.00	0.01	8654785952	270506016	11.69		
Global Load To Shared Store (access)	0	0	67634962	2.92	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	524288	0.02	2097152	4	100.00	67108864	2097152	0.09	-	-
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	67633152	67633152	67634962	2.92	270462061	4.00	0.01	8654785952	270506016	11.69	67721428	2.93
Stores	524288	524288	524288	0.02	2097152	4	100.00	67108864	2097152	0.09	-	-
Atomics & Reductions	0	0	0	0	0	0	0	0	0	0	-	-
Total	68157440	68157440	68159250	2.95	272559213	4.00	0.78	8721894816	272603168	11.78	67721428	2.93

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	168289815	270506168	1.61	15.11	44.64	8656197376	100977525267.14	150177716	0	0
L1/TEX Store	524288	2097152	4	0.12	100	67108864	782848023.89	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	168814170	272603064	1.61	15.23	45.06	8723298048	101760277728.52	151057640	0	0
ECC Total	-	0	-	0	-	0	0	0	-	-
GPU Total	169031714	272851371	1.61	15.24	44.65	8731243872	101852968503.57	151038592	15072	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	270458500	44.78	0	0
L1/TEX Store	0	0	0	0	2097152	100	0	0
L1/TEX Atomic	0	0	0	0	0	0	-	-
L1/TEX Total	0	0	0	0	272781844	45.30	0	0
GPU Total	251956	81.01	0	0	273772373	44.84	0	0

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	148864684	16.55	4763669888	55569850776.91
Store	2106928	0.23	67421696	786497316.97
Total	150971612	16.79	4831091584	56356348093.88

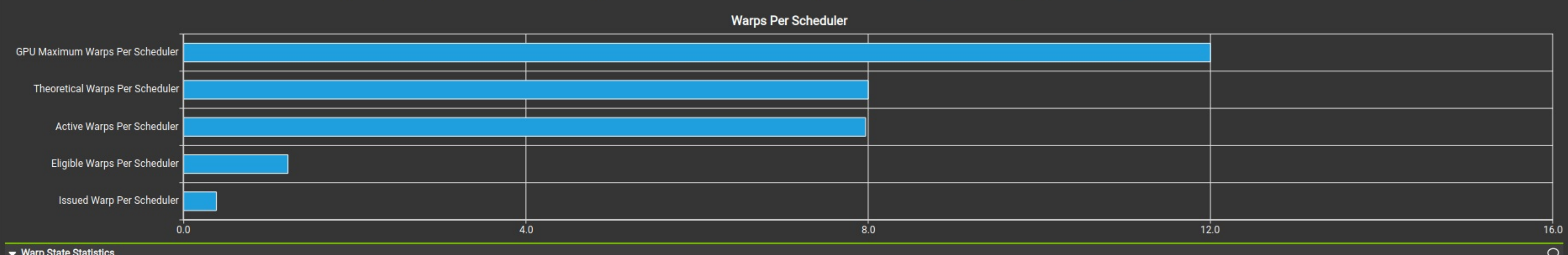
Scheduler Statistics

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]	7.97	No Eligible [%]	61.75
Eligible Warps Per Scheduler [warp]	1.22	One or More Eligible [%]	38.25
Issued Warp Per Scheduler	0.38		

Issue Slot Utilization
Est. Local Speedup: 18.70%
Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 2.6 cycles. This might leave hardware resources underutilized and may lead to less optimal performance. Out of the maximum of 12 warps per scheduler, this kernel allocates an average of 7.97 active warps per scheduler, but only an average of 1.22 warps were eligible per cycle. Eligible warps are the subset of active warps that are ready to issue their next instruction. Every cycle with no eligible warp results in no instruction being issued and the issue slot remains unused. To increase the number of eligible warps, avoid possible load imbalances due to highly different execution durations per warp. Reducing stalls indicated on the Warp State Statistics and Source Counters sections can help, too.

Warps Per Scheduler



Warp State Statistics

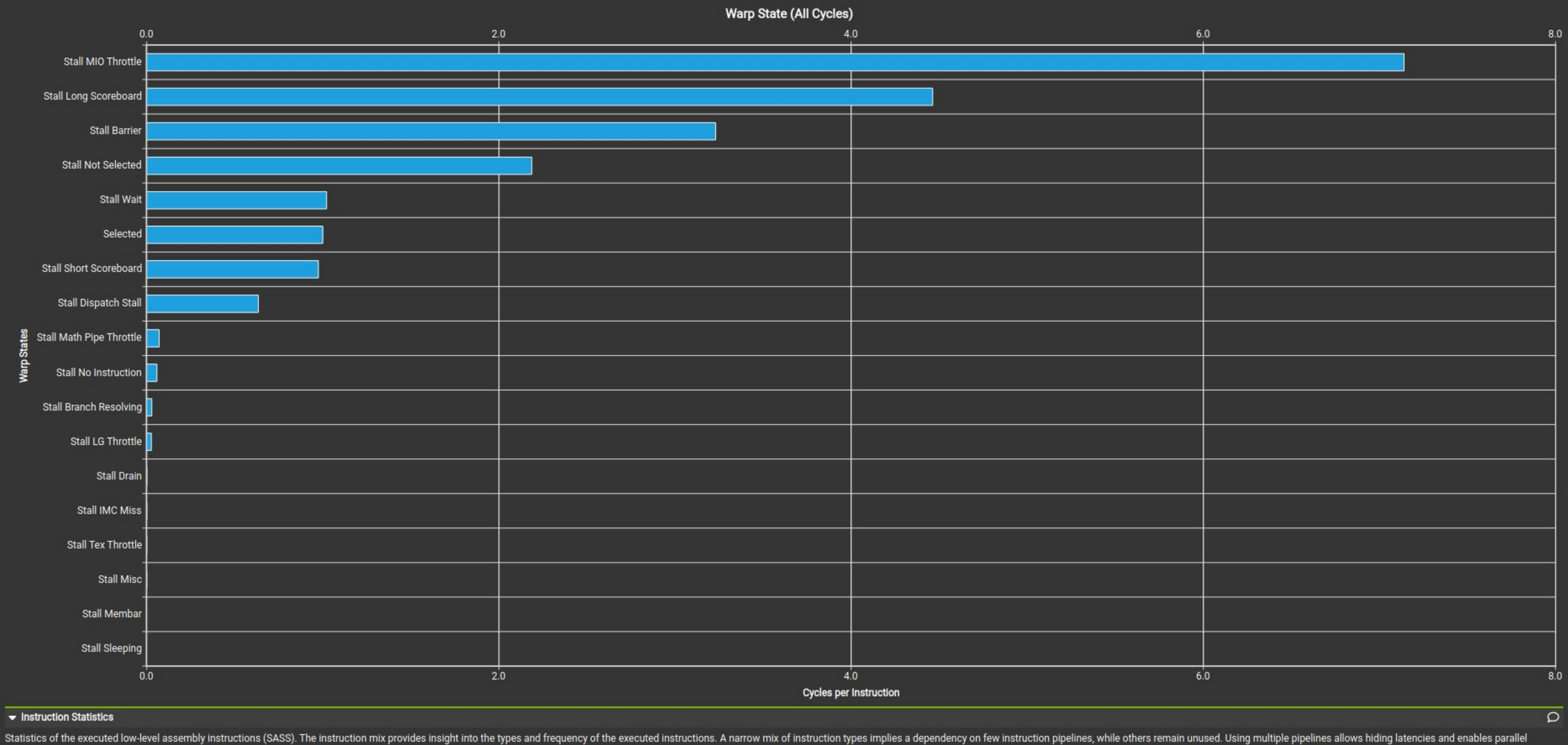
Analysis of the states in which all warps spent cycles during the kernel execution. The warp states describe a warp's readiness or inability to issue its next instruction. The warp cycles per instruction define the latency between two consecutive instructions. The higher the value, the more warp parallelism 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]	20.84	Avg. Active Threads Per Warp	32
Warp Cycles Per Executed Instruction [cycle]	20.84	Avg. Not Predicated Off Threads Per Warp	32.00

Mio Throttle Stalls
Est. Speedup: 18.70%
On average, each warp of this kernel spends 7.1 cycles being stalled waiting for the MIO (memory input/output) instruction queue to be not full. This stall reason is high in cases of extreme utilization of the MIO pipelines, which include special math instructions, dynamic branches, as well as shared memory instructions. When caused by shared memory accesses, trying to use fewer but wider loads can reduce pipeline pressure. This stall type represents about 34.3% of the total average of 20.8 cycles between issuing two instructions.

Warp Stall Check the Warp Stall Sampling (All Samples) table for the top stall locations in your source based on sampling data. The Kernel Profiling Guide provides more details on each stall reason.

Warp State (All Cycles)

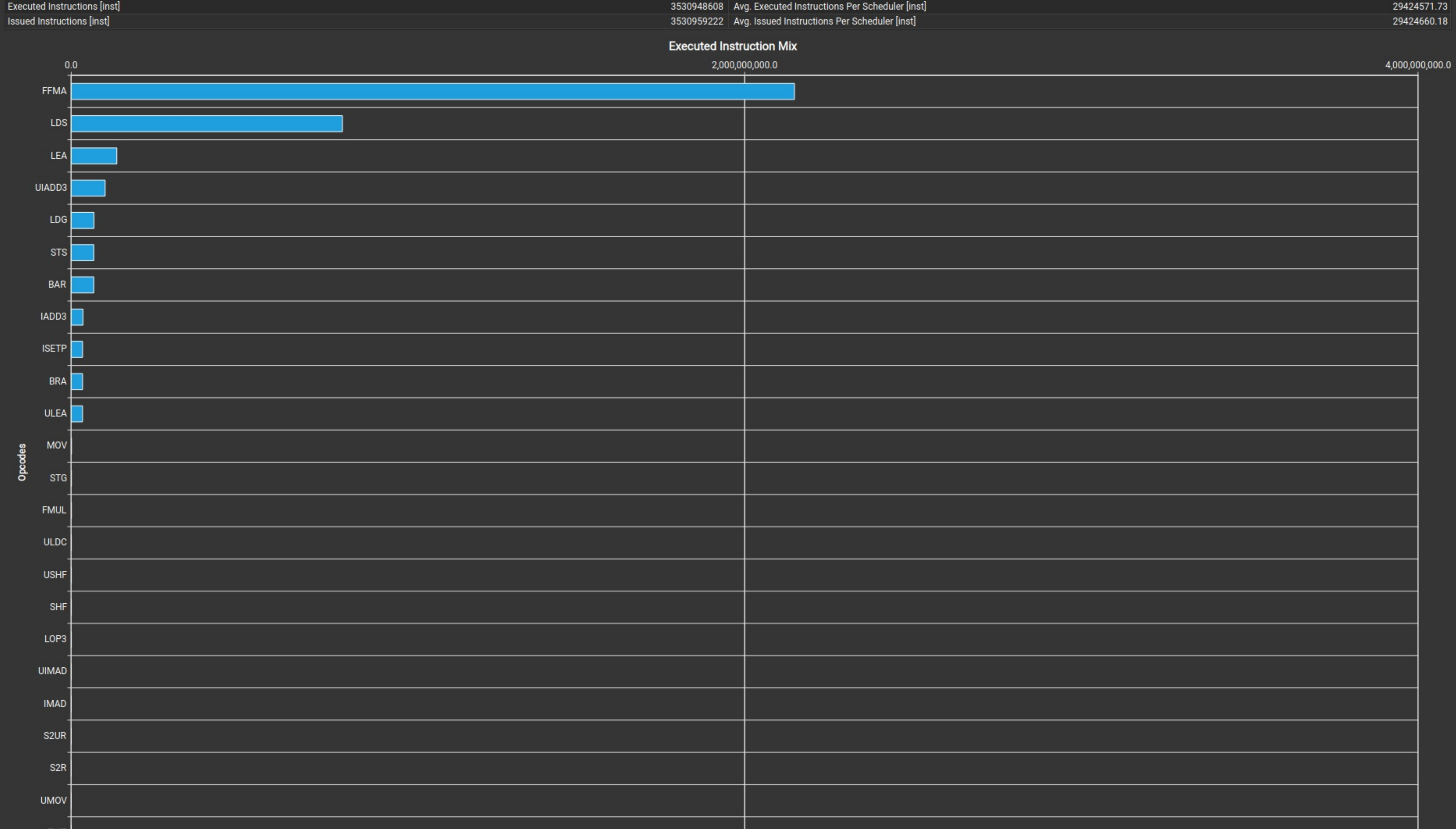


Instruction Statistics

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]	3530948608	Avg. Executed Instructions Per Scheduler [inst]	29424571.73
Issued Instructions [inst]	3530959222	Avg. Issued Instructions Per Scheduler [inst]	29424660.18

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	4096	Function Cache Configuration	CachePreferNone
Registers Per Thread [register/thread]	48	Static Shared Memory Per Block [Kbyte/block]	4.10
Block Size	512	Dynamic Shared Memory Per Block [byte/block]	0
Threads [thread]	2097152	Driver Shared Memory Per Block [Kbyte/block]	1.02
Waves Per SM	68.27	Shared Memory Configuration Size [Kbyte]	16.38
Uses Green Context	0	# SMs [SM]	30

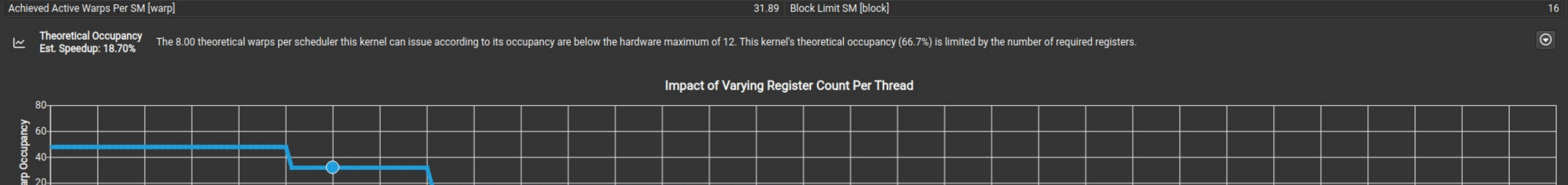
Occupancy

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.

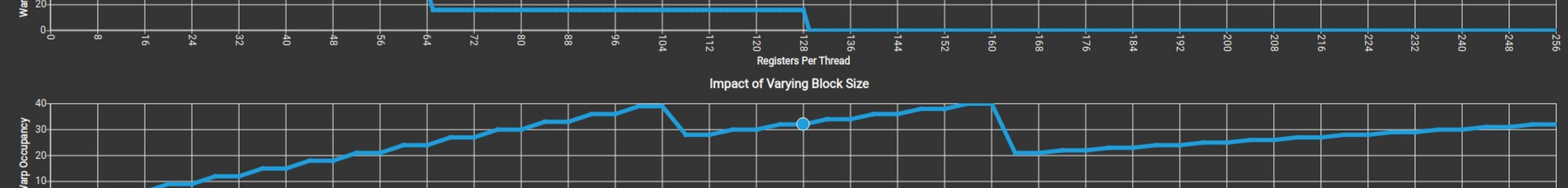
Theoretical Occupancy [%]	66.67	Block Limit Registers [block]	2
Theoretical Active Warps per SM [warp]	32	Block Limit Shared Mem [block]	3
Achieved Occupancy [%]	66.43	Block Limit Warps [block]	3
Achieved Active Warps Per SM [warp]	31.89	Block Limit SM [block]	16

Theoretical Occupancy
Est. Speedup: 18.70%
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.

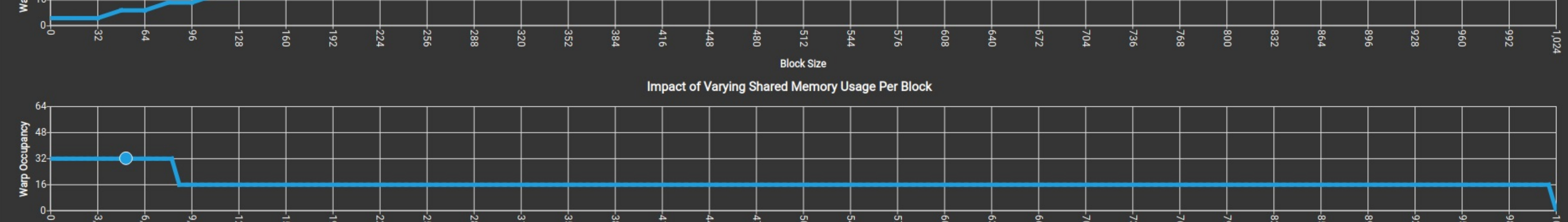
Impact of Varying Register Count Per Thread



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]	76922855.97	Average L1 Active Cycles [cycle]	76922855.97
Average L2 Active Cycles [cycle]	74450511.17	Average SMSP Active Cycles [cycle]	76933249.24
Average DRAM Active Cycles [cycle]	100647741.33	Total SM Elapsed Cycles [cycle]	2314126540
Total L1 Elapsed Cycles [cycle]	2314126540	Total L2 Elapsed Cycles [cycle]	1789915632
Total SMSP Elapsed Cycles [cycle]	9256506160	Total DRAM Elapsed Cycles [cycle]	3597320192

Workload Distribution

	Average	Min	Max	Sum
SM Active Cycles	76922855.97	76644324	77111749	2307685679
SMSP Active Cycles	76933249.24	76619978	77153992	9231989909
L1 Active Cycles	76922855.97	76644324	77111749	2307685679
L2 Active Cycles	74450511.17	74440041	74464226	1786812268
DRAM Active Cycles	100647741.33	100575616	100692592	603886448

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]	33685504	Branch Efficiency [%]	100
Branch Instructions Ratio [%]	0.01	Avg. Divergent Branches	0

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.