Page: Details | Launch: 0 - 505 - MM_v2 | Add Baseline | Apply Rules | Save as PDF

	Launch	Time	Cycles	Regs	GPU	SM Frequency	CC	Process
Current	505 - MM_v2 (128, 1, 1)x(16, 16, 1)	4.55 msecond	5,999,993	37	NVIDIA GeForce RTX 3060	1.32 cycle/nsecond	8.6	[22484] MM_v2

GPU Speed Of Light Throughput

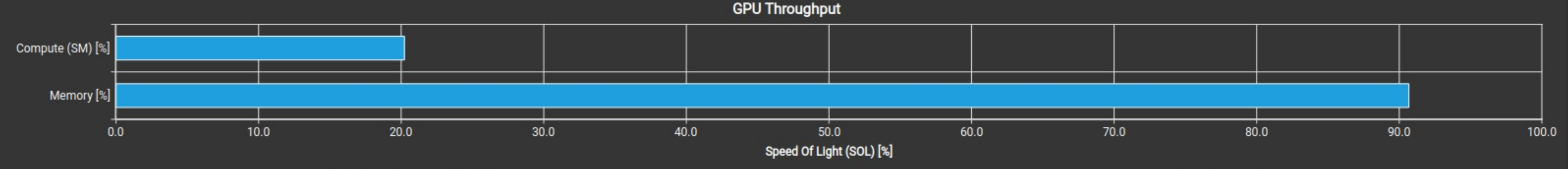
High-level overview of the throughput for compute and memory resources of the GPU. For each unit, the throughput reports the achieved percentage of utilization with respect to the theoretical maximum. Breakdowns show the throughput for each individual sub-metric of Compute and Memory to clearly identify 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 [%]	20.23	Duration [msecond]	4.55
Memory Throughput [%]	90.68	Elapsed Cycles [cycle]	5,999,993
L1/TEX Cache Throughput [%]	99.21	SM Active Cycles [cycle]	5,483,872.43
L2 Cache Throughput [%]	5.29	SM Frequency [cycle/nsecond]	1.32
DRAM Throughput [%]	5.74	DRAM Frequency [cycle/nsecond]	7.29

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 Memory Workload Analysis section.

Roofline Analysis The ratio of peak float (fp32) to double (fp64) performance on this device is 64:1. The kernel achieved 1% of this device's fp32 peak performance and 0% of its fp64 peak performance.

GPU Throughput



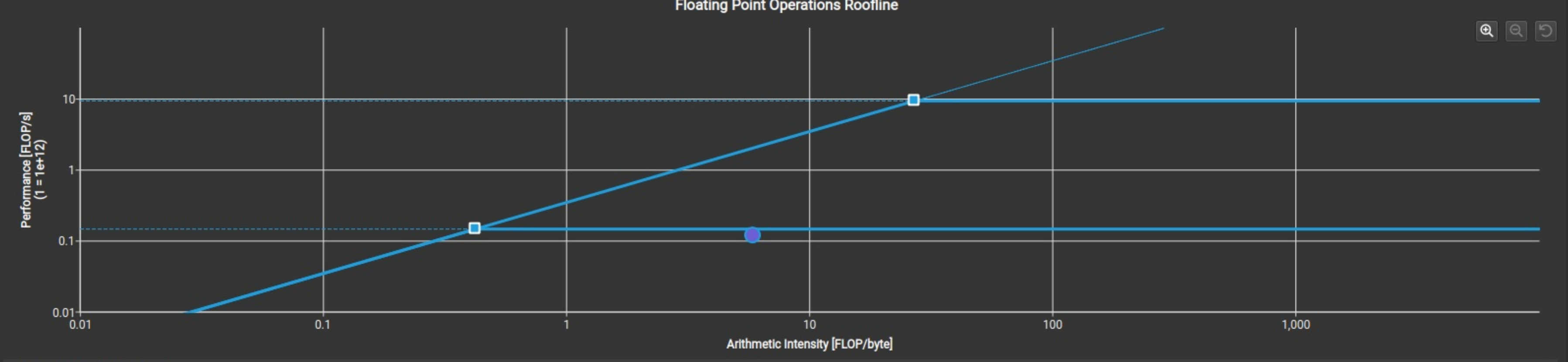
Compute Throughput Breakdown

SM: Inst Executed Pipe Lsu [%]	20.23
SM: Mio Inst Issued [%]	6.74
SM: Mio2rf Writeback Active [%]	5.25
SM: Issue Active [%]	4.56
SM: Inst Executed [%]	4.56
SM: Pipe Fma Cycles Active [%]	1.36
SM: Pipe Alu Cycles Active [%]	1.23
SM: Pipe Fmaheavy Cycles Active [%]	0.24
SM: Mio Pq Read Cycles Active [%]	0.08
SM: Mio Pq Write Cycles Active [%]	0.08
SM: Inst Executed Pipe Uniform [%]	0.00
SM: Inst Executed Pipe Adu [%]	0.00
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.00
SM: Inst Executed Pipe Ipa [%]	0
SM: Inst Executed Pipe Tex [%]	0
SM: Inst Executed Pipe Xu [%]	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 [%]	90.68
L1: Lsuin Requests [%]	20.23
L1: Lsu Writeback Active [%]	10.74
DRAM: Cycles Active [%]	5.74
L2: Xbar2lts Cycles Active [%]	5.29
L2: T Sectors [%]	4.51
DRAM: Dram Sectors [%]	4.18
L1: M L1tex2xbar Req Cycles Active [%]	3.20
L2: T Tag Requests [%]	2.84
L2: Lts2xbar Cycles Active [%]	1.94
L2: D Sectors Fill Device [%]	1.77
L2: D Sectors [%]	1.76
L1: Data Bank Reads [%]	1.48
L1: M Xbar2l1tex Read Sectors [%]	1.14
L1: Data Bank Writes [%]	0.20
L2: D Sectors Fill Sysmem [%]	0.00
L1: Texin Sm2tex Req Cycles Active [%]	0.00
L1: F Wavefronts [%]	0
L1: Data Pipe Tex Wavefronts [%]	0
L1: Tex Writeback Active [%]	0
L2: D Atomic Input Cycles Active [%]	0

Floating Point Operations Roofline



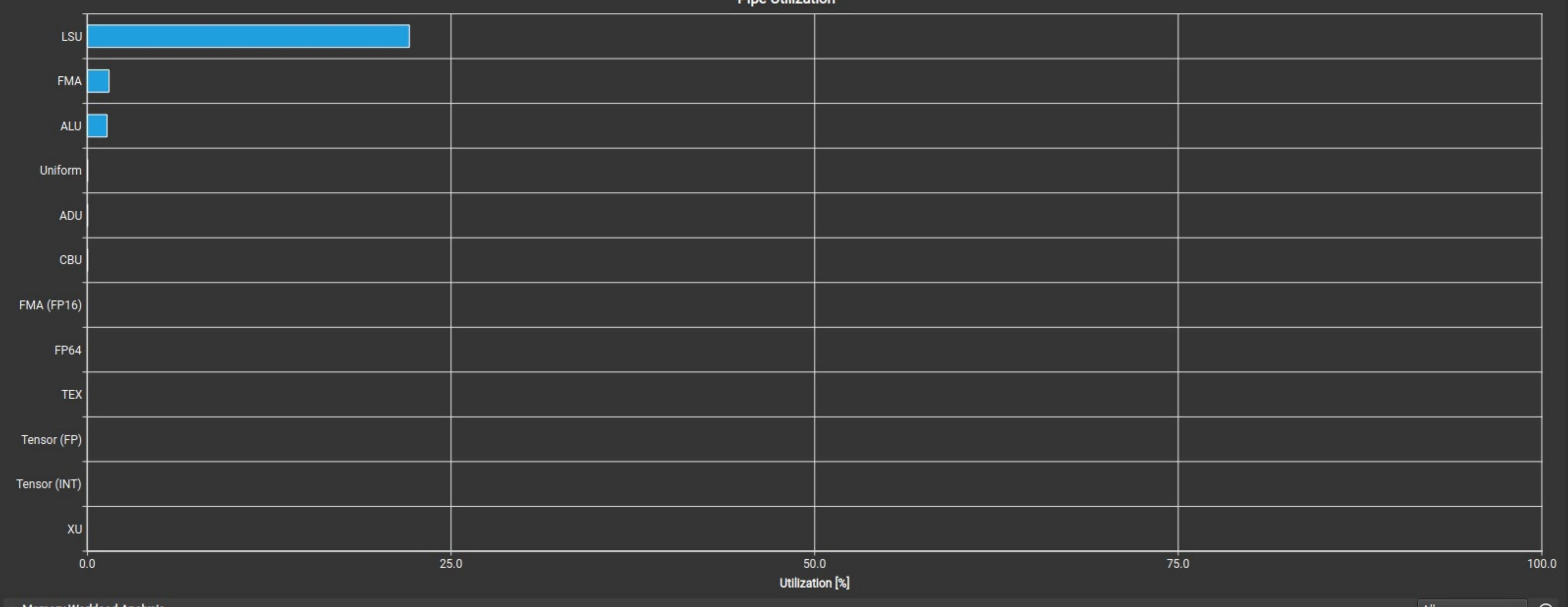
Compute Workload Analysis

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]	0.18	SM Busy [%]	7.38
Executed Ipc Active [inst/cycle]	0.20	Issue Slots Busy [%]	4.99
Issued Ipc Active [inst/cycle]	0.20		

Balanced No pipeline is over-utilized.

Pipe Utilization

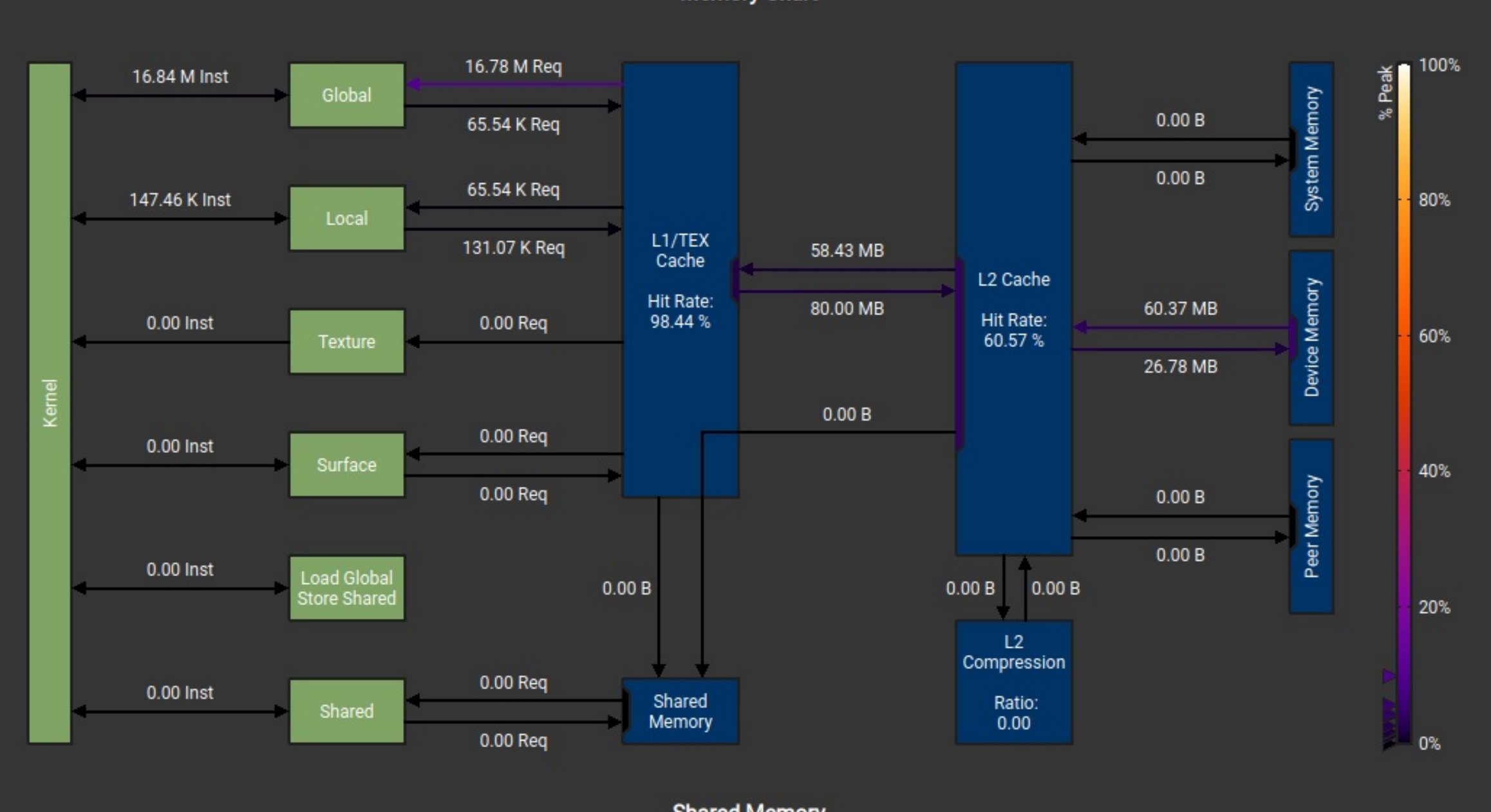


Memory Workload Analysis

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/second]	20.08	Mem Busy [%]	90.68
L1/TEX Hit Rate [%]	98.44	Max Bandwidth [%]	20.23
L2 Hit Rate [%]	60.57	Mem Pipes Busy [%]	20.23
L2 Compression Success Rate [%]	0	L2 Compression Ratio	0

Memory Chart



Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	0	0	0	0	0
Shared Load Matrix	0	0	0	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	-	-	2,048	0.00	0
Total	0	0	2,048	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
Local Load	65,536	65,536	65,536	0.04	261,924	4.00	0.81	8,381,568			18,04
Global Load	16,777,216	16,777,216			150,994,679	9.00	98.90	4,831,829,728	1,915,934	1.14	
Global Load To Shared Store (access)	0	0	41,943,068	24.97	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	
Texture Load	0	0	0	0	0	0	0	0	0	0	
Global Store	65,536	65,536	262,144	0.16	2,097,152	32	90.63	67,108,864	2,621,440	1.56	
Local Store	81,920	131,072	131,072	0.08	524,200	4.00	44.08	16,774,400			
Surface Store	0	0	0	0	0	0	0	0	0	0	
Global Reduction	0	0	0	0	0	0	0	0	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 a
Global Atomic CAS	0										
Surface Atomic ALU	0	0	0	0	0	0	0	0	0	0	see a
Surface Atomic CAS	0										
Loads	16,842,752	16,842,752	42,008,604	25.01	151,256,603	8.98	98.73	4,840,211,296	1,915,934	1.14	18,04
Stores	147,456	196,608	393,216	0.23	2,621,352	13.33	81.32	83,883,264	2,621,440	1.56	

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	1,719,437	1,915,543	1.11	1.87	8.24	61,297,376	13,471,903,395.50	1,760,778	0	0
L1/TEX Store	1,166,381	2,621,440	2.25	2.56	100	83,886,080	18,436,436,267.48	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	2,895,306	4,538,662	1.57	4.43	60.77	145,237,184	31,920,147,973.11	1,757,120	0	0
GPU Total	2,891,743	4,625,449	1.60	4.51	61.45	148,014,368	32,530,515,936.65	2,006,056	433	0

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	1,978,136	3.98	63,300,352	13,912,116,353.00
Store	877,472	1.76	28,079,104	6,171,209,947.39
Total	2,855,608	5.74	91,379,456	20,083,326,300.39

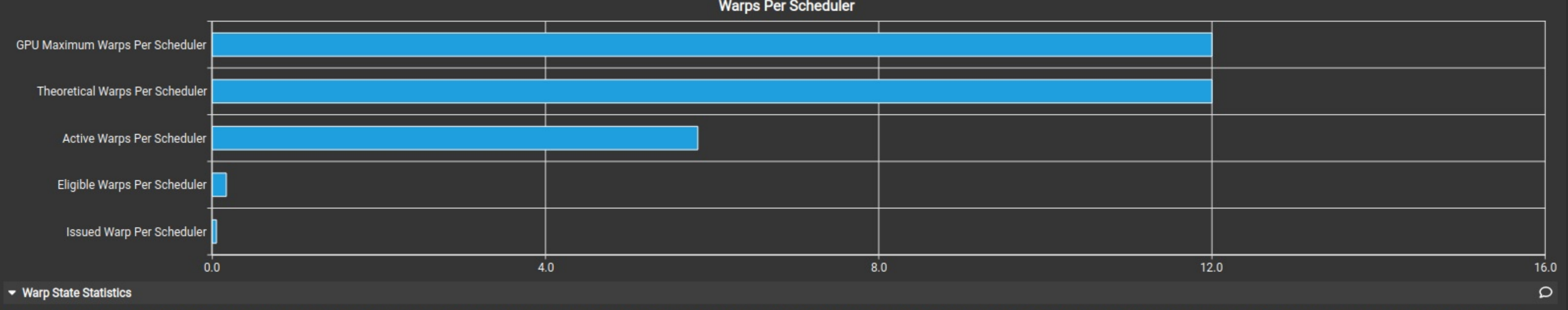
Scheduler Statistics

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]	5.83	No Eligible [%]	95.01
Eligible Warps Per Scheduler [warp]	0.17	One or More Eligible [%]	4.99
Issued Warp Per Scheduler	0.05		

Issue Slot Utilization Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 20.0 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 5.83 active warps per scheduler, but only an average of 0.17 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, reduce the time the active warps are stalled by inspecting the top stall reasons on the Warp State Statistics and Source Counters sections.

Warps Per Scheduler



Warp State Statistics

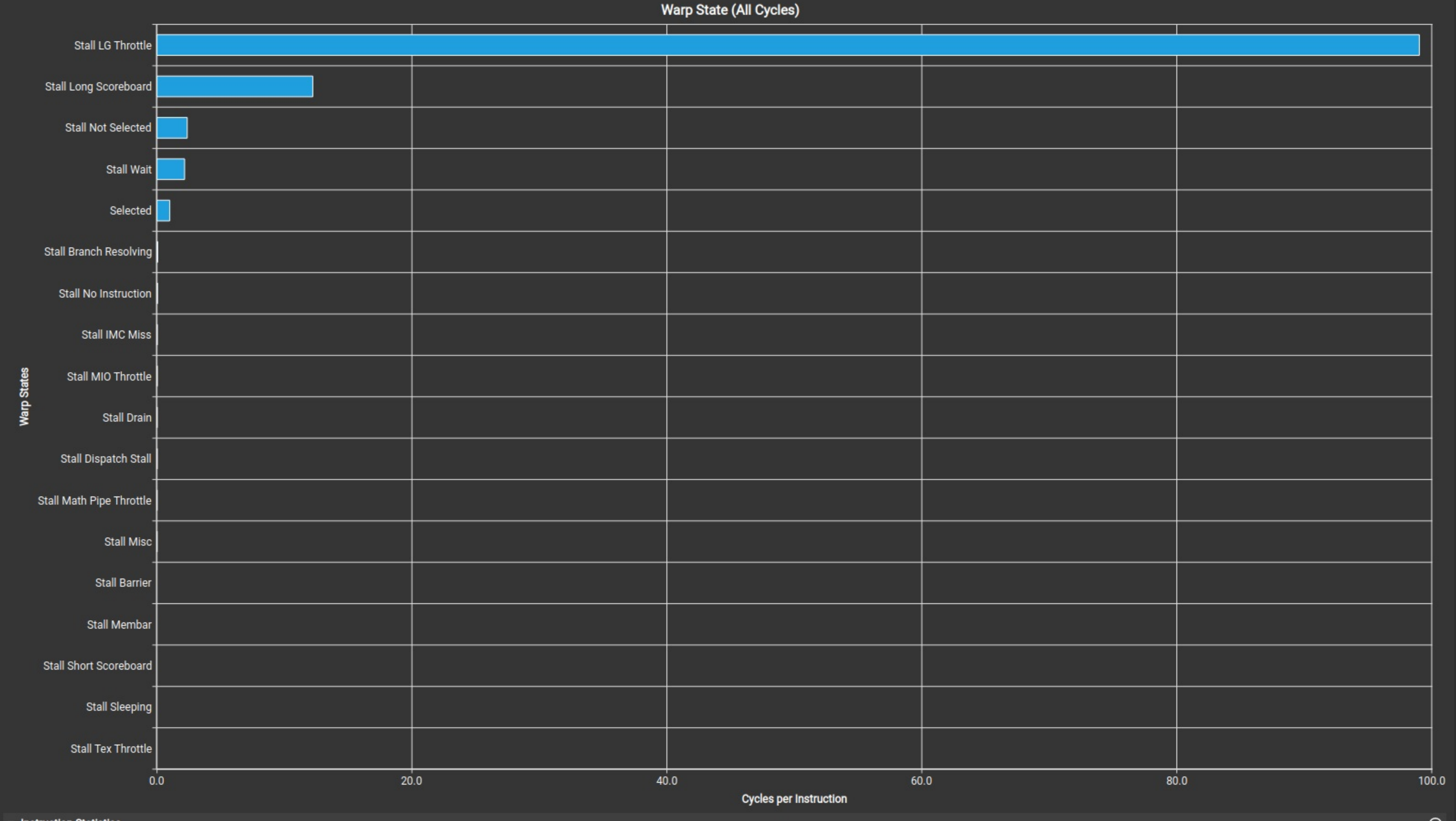
Analysis of the states in which all warps spent cycles during the kernel execution. The warp states describe a warp's readiness or inability to issue its next instruction. The warp cycles per instruction define the latency between two consecutive instructions. The higher the value, the more warp parallelism 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]	116.76	Avg. Active Threads Per Warp	32
Warp Cycles Per Executed Instruction [cycle]	116.79	Avg. Not Predicated Off Threads Per Warp	31.93

lg_throttle On average, each warp of this kernel spends 99.0 cycles being stalled waiting for the local/global instruction queue to be not full. This represents about 84.8% of the total average of 116.8 cycles between issuing two instructions. Typically this stall occurs only when executing local or global memory instructions extremely frequently. If applicable, consider combining multiple lower-width memory operations into fewer wider memory operations and try interleaving memory operations and math instructions.

Warp Stall Check the Source Counters section for the top stall locations in your source based on sampling data.

Warp State (All Cycles)

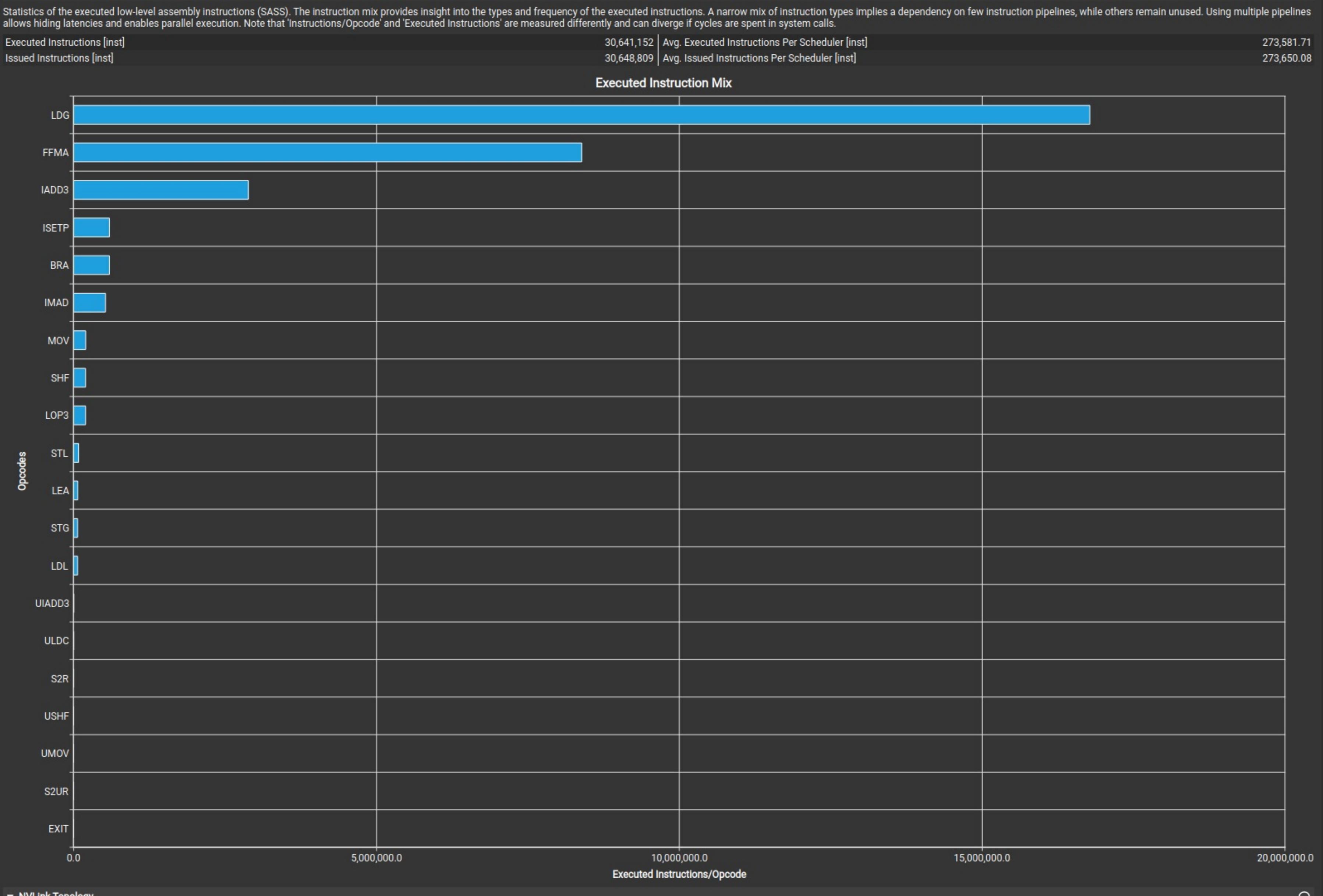


Instruction Statistics

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]	30,641,152	Avg. Executed Instructions Per Scheduler [inst]	273,581.71
Issued Instructions [inst]	30,648,809	Avg. Issued Instructions Per Scheduler [inst]	273,650.08

Executed Instruction Mix



NVLink Topology

NVLink Topology.

NVLink Topology

The system does not have any NVLink connections.

NVLink Tables

Detailed tables with properties for each NVLink.

Logical NVLink Properties

The system does not have any NVLink connections.

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	128	Registers Per Thread [register/thread]	37
Block Size	256	Static Shared Memory Per Block [byte/block]	0
Threads [thread]	32,768	Dynamic Shared Memory Per Block [byte/block]	0
Waves Per SM	0.76	Driver Shared Memory Per Block [Kbyte/block]	1.02
Function Cache Configuration	cudaFuncCachePreferNone	Shared Memory Configuration Size [Kbyte]	8.19

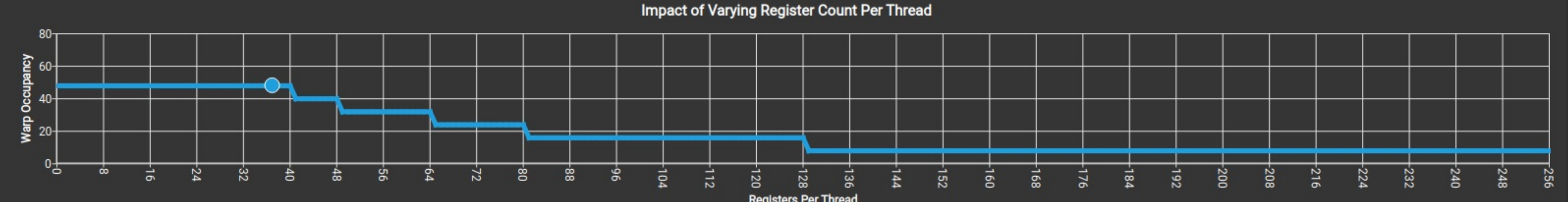
Occupancy

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.

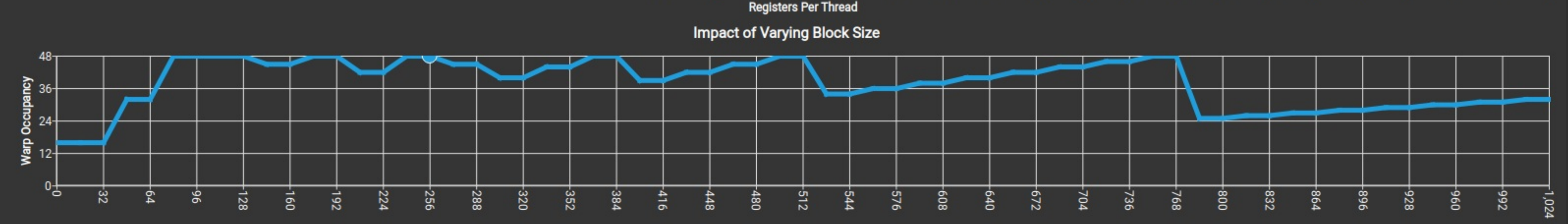
Theoretical Occupancy [%]	100	Block Limit Registers [block]	6
Theoretical Active Warps per SM [warp]	48	Block Limit Shared Mem [block]	100
Achieved Occupancy [%]	48.57	Block Limit Warps [block]	6
Achieved Active Warps Per SM [warp]	23.31	Block Limit SM [block]	16

Occupancy Limiters This kernel's theoretical occupancy is not impacted by any block limit. The difference between calculated theoretical (100.0%) and measured achieved occupancy (48.6%) can be the result of warp scheduling overheads or workload imbalances during the kernel execution. Load imbalances can occur between warps within a block as well as across blocks of the same kernel.

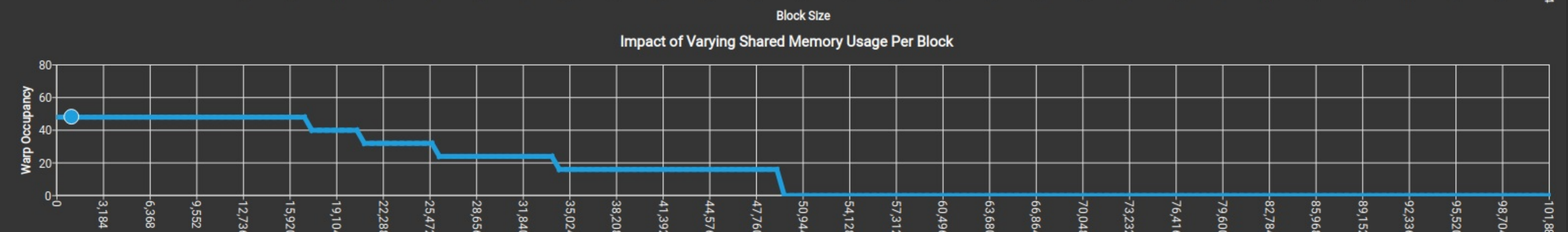
Impact of Varying Register Count Per Thread



Impact of Varying Block Size



Impact of Varying Shared Memory Usage Per Block



Source Counters

Source metrics, including branch efficiency and sampled warp stall reasons. Sampling Data 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]	590,848	Branch Efficiency [%]	100
Branch Instructions Ratio	0.02	Avg. Divergent Branches	0

Uncoalesced Global Accesses Uncoalesced global access, expected 2068229 sectors, got 8243993 (3.99x) at PC 0x7ff272fafd80

Uncoalesced Global Accesses Uncoalesced global access, expected 2068199 sectors, got 8243843 (3.99x) at PC 0x7ff272fafe90

Uncoalesced Global Accesses Uncoalesced global access, expected 2068133 sectors, got 8243513 (3.99x) at PC 0x7ff272fafd70

Uncoalesced Global Accesses Uncoalesced global access, expected 2068109 sectors, got 8243393 (3.99x) at PC 0x7ff272fafcf0

Uncoalesced Global Accesses Uncoalesced global access, expected 2068082 sectors, got 8243258 (3.99x) at PC 0x7ff272fafdd0

Uncoalesced Global Accesses Uncoalesced global access, expected 2067827 sectors, got 8241983 (3.99x) at PC 0x7ff272fafc70

Uncoalesced Global Accesses Uncoalesced global access, expected 2067794 sectors, got 8241818 (3.99x) at PC 0x7ff272fafcd0

Uncoalesced Global Accesses Uncoalesced global access, expected 2067698 sectors, got 8241338 (3.99x) at PC 0x7ff272fafe60

Uncoalesced Global Accesses Uncoalesced global access, expected 2067668 sectors, got 8241188 (3.99x) at PC 0x7ff272fafd10

Uncoalesced Global Accesses Uncoalesced global access, expected 2067656 sectors, got 8241128 (3.99x) at PC 0x7ff272fafc90

Sampling Data (All)

Location	Value	Value (%)
0x7ff272fafc60 in MM_v2	8,701	6
0x7ff272fafde0 in MM_v2	8,375	6
0x7ff272fafdd0 in MM_v2	5,744	4
0x7ff272fafda0 in MM_v2	5,505	4
0x7ff272fafe80 in MM_v2	5,286	4

Sampling Data (Not Issued)

Location	Value	Value (%)
0x7ff272fafc60 in MM_v2	8,290	6
0x7ff272fafde0 in MM_v2	7,958	6
0x7ff272fafdd0 in MM_v2	5,468	4
0x7ff272fafda0 in MM_v2	5,266	4
0x7ff272fafe80 in MM_v2	5,047	4

Most Instructions Executed

Location	Value	Value (%)
0x7ff272fafd30 in MM_v2	524,288	2
0x7ff272fafe80 in MM_v2	524,288	2
0x7ff272fafe70 in MM_v2	524,288	2
0x7ff272fafe60 in MM_v2	524,288	2
0x7ff272fafe50 in MM_v2	524,288	2