Page: Details | Launch: 79 - 18181 - ampere_sgemm_128x12 | Add Baseline | Apply Rules | Save as PDF

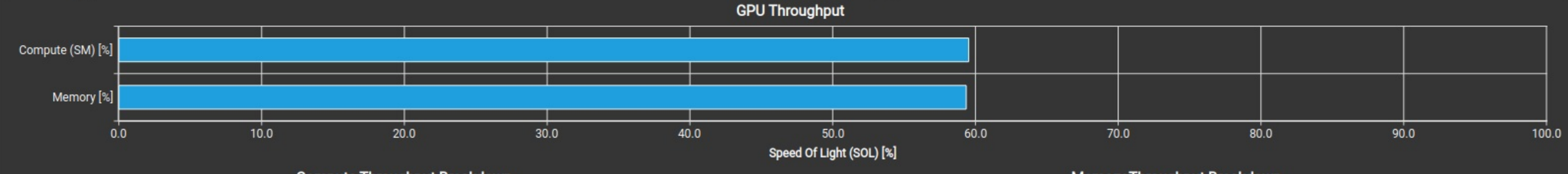
	Launch	Time	Cycles	Regs	GPU	SM Frequency	CC	Process
Current	18181 - ampere_sgemm_128x128_nn (1, 1, 128)x(256, 1, 1)	120.80 usecond	156,402	118	NVIDIA GeForce RTX 3060	1.29 cycle/nsecond	8.6	[24865] cublas

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 [%]	59.53	Duration [usecond]	120.80
Memory Throughput [%]	59.35	Elapsed Cycles [cycle]	156,402
L1/TEX Cache Throughput [%]	49.52	SM Active Cycles [cycle]	142,668.36
L2 Cache Throughput [%]	29.54	SM Frequency [cycle/nsecond]	1.29
DRAM Throughput [%]	59.35	DRAM Frequency [cycle/nsecond]	7.16

GPU Throughput



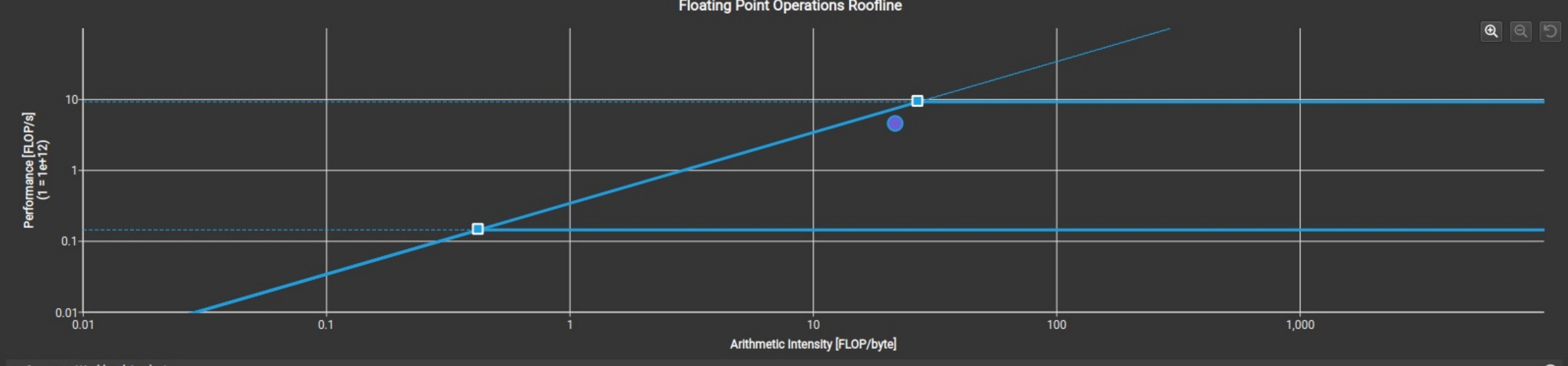
Compute Throughput Breakdown

SM: Issue Active [%]	59.53
SM: Inst Executed [%]	59.50
SM: Pipe Fma Cycles Active [%]	48.39
SM: Inst Executed Pipe Lsu [%]	45.18
SM: Pipe Fmaheavy Cycles Active [%]	40.44
SM: Mio Inst Issued [%]	15.72
SM: Mio2rf Writeback Active [%]	14.76
SM: Pipe Alu Cycles Active [%]	9.60
SM: Mio Pq Write Cycles Active [%]	6.36
SM: Mio Pq Read Cycles Active [%]	6.36
SM: Inst Executed Pipe Adu [%]	2.34
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.40
SM: Inst Executed Pipe Tex [%]	0.05
IDC: Request Cycles Active [%]	0
SM: Inst Executed Pipe Xu [%]	0
SM: Inst Executed Pipe Uniform [%]	0
SM: Inst Executed Pipe Ipa [%]	0
SM: Pipe Fp64 Cycles Active [%]	0
SM: Pipe Tensor Cycles Active [%]	0

Memory Throughput Breakdown

DRAM: Cycles Active [%]	59.35
L1: Lsuin Requests [%]	45.18
DRAM: Dram Sectors [%]	43.21
L1: Data Pipe Lsu Wavefronts [%]	42.94
L2: T Sectors [%]	29.54
L1: Lsu Writeback Active [%]	29.37
L2: D Sectors Fill Device [%]	20.29
L2: Lts2xbar Cycles Active [%]	19.64
L2: Xbar2lts Cycles Active [%]	16.76
L2: T Tag Requests [%]	14.77
L2: D Sectors [%]	14.51
L1: M Xbar2l1tex Read Sectors [%]	11.98
L1: Data Bank Reads [%]	10.53
L1: M L1tex2xbar Req Cycles Active [%]	9.85
L1: Data Bank Writes [%]	4.50
L1: Texin Sm2tex Req Cycles Active [%]	0.11
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

Floating Point Operations Roofline

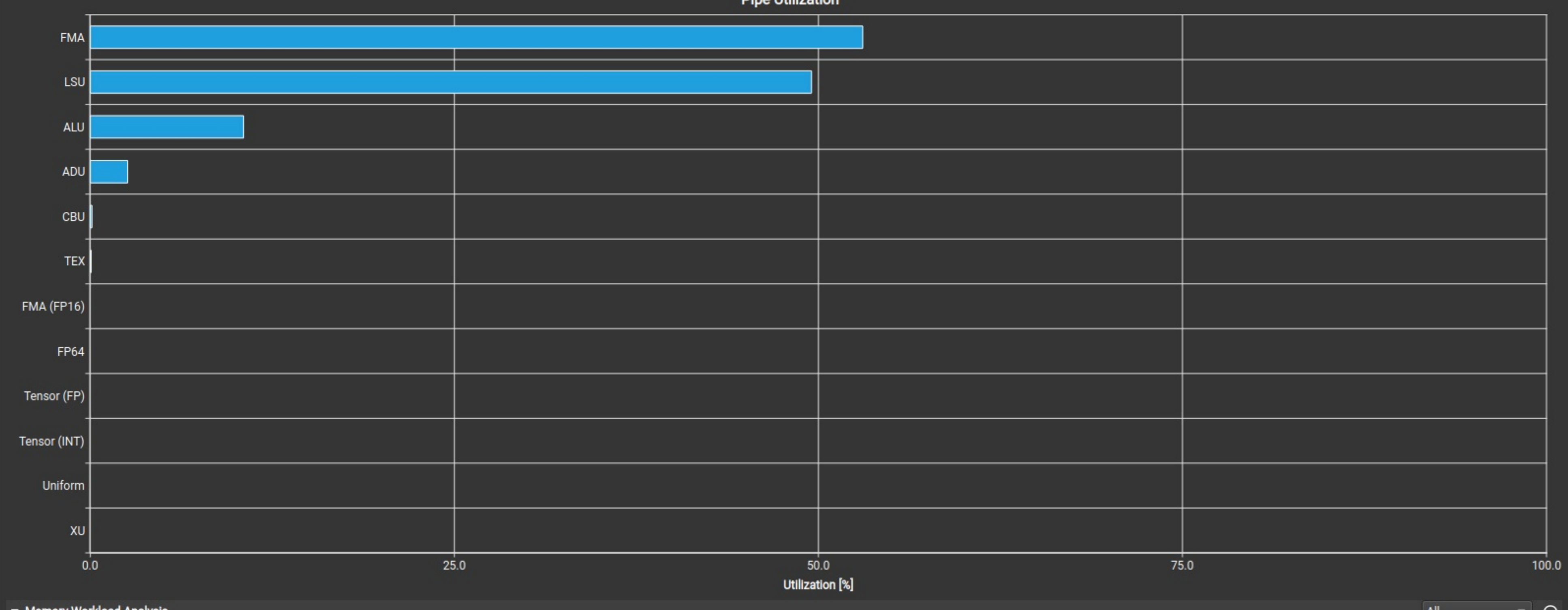


Compute Workload Analysis

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]	2.38	SM Busy [%]	65.25
Executed Ipc Active [inst/cycle]	2.61	Issue Slots Busy [%]	65.25
Issued Ipc Active [inst/cycle]	2.61		

Pipe Utilization

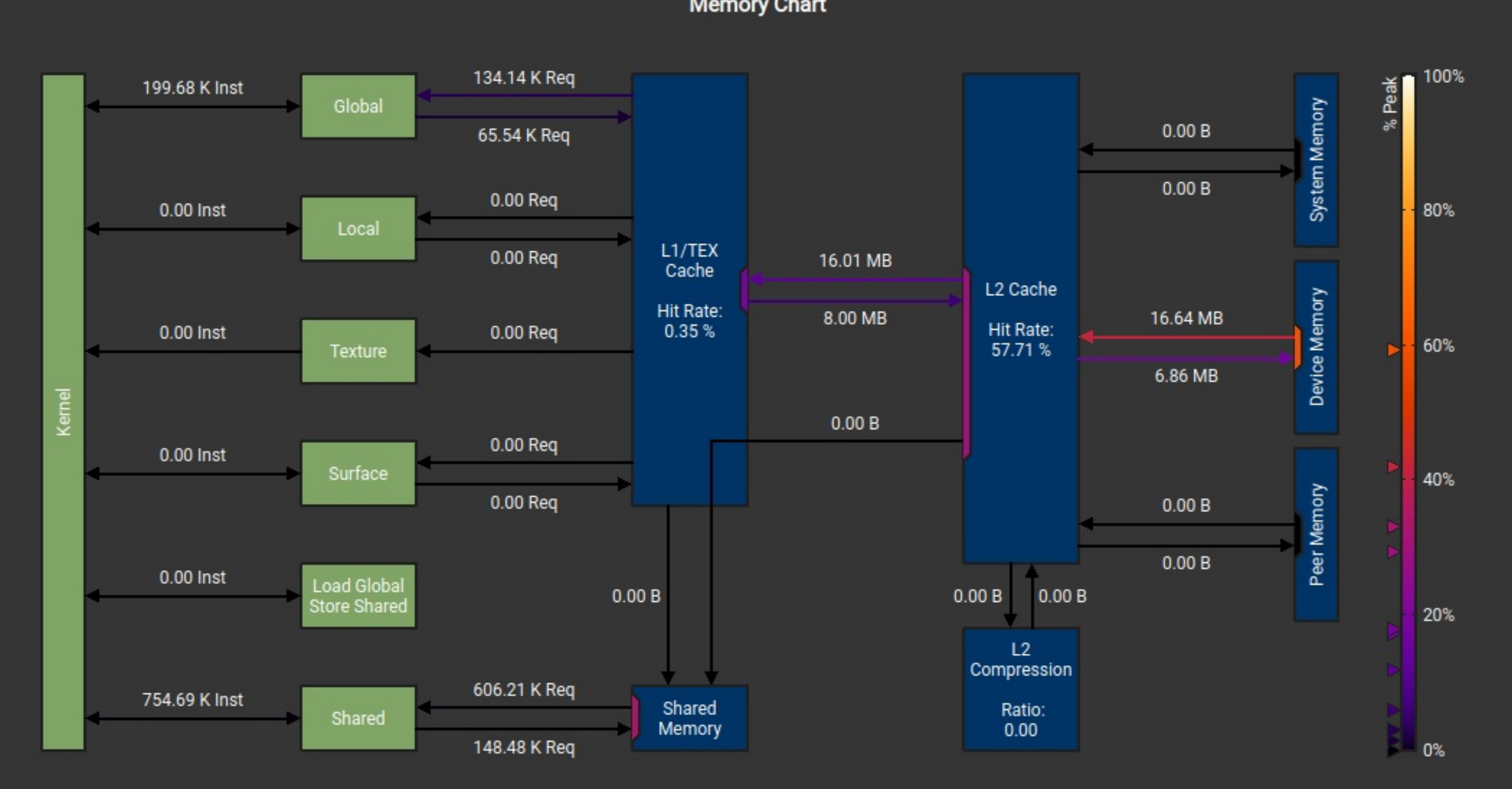


Memory Workload Analysis

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]	203.95	Mem Busy [%]	42.94
L1/TEX Hit Rate [%]	0.35	Max Bandwidth [%]	59.35
L2 Hit Rate [%]	57.71	Mem Pipes Busy [%]	45.18
L2 Compression Success Rate [%]	0	L2 Compression Ratio	0

Memory Chart



Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	606,208	606,208	1,147,117	26.20	0
Shared Load Matrix	0	0	0	0	0
Shared Store	148,480	148,480	263,168	1.50	65,536
Shared Store From Global Load	0	0	0	0	0
Shared Atomic	0	0	0	0	0
Other	-	-	42,294	5.47	0
Total	754,688	754,688	1,452,579	33.17	65,536

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	0	0	0	0	0	0	0	0			
Global Load	134,144	134,144			513,922	3.83	0.53	16,445,504	524,613	11.98	13
Global Load To Shared Store (access)	0	0	134,196	3.06	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	65,536	1.50	262,144	4	0	8,388,608	262,144	5.99	
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	
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	134,144	134,144	134,196	3.06	513,922	3.83	0.53	16,445,504	524,613	11.98	13
Stores	65,536	65,536	65,536	1.50	262,144	4	0	8,388,608	262,144	5.99	

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	328,004	524,613	1.60	19.63	26.13	16,787,616	138,970,331,125.83	541,096	0	0
L1/TEX Store	65,536	262,144	4	9.81	100	8,388,608	69,442,119,205.30	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	393,540	786,754	2.00	29.44	45.63	25,176,128	208,411,655,629.14	541,940	0	0
GPU Total	394,492	789,632	2.00	29.54	45.65	25,268,224	209,174,039,735.10	541,324	62	0

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	545,268	42.04	17,448,576	144,441,854,304.64
Store	224,628	17.32	7,188,096	59,504,105,960.26
Total	769,896	59.35	24,636,672	203,945,960,264.90

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]	3.69	No Eligible [%]	35.13
Eligible Warps Per Scheduler [warp]	1.57	One or More Eligible [%]	64.87
Issued Warp Per Scheduler	0.65		

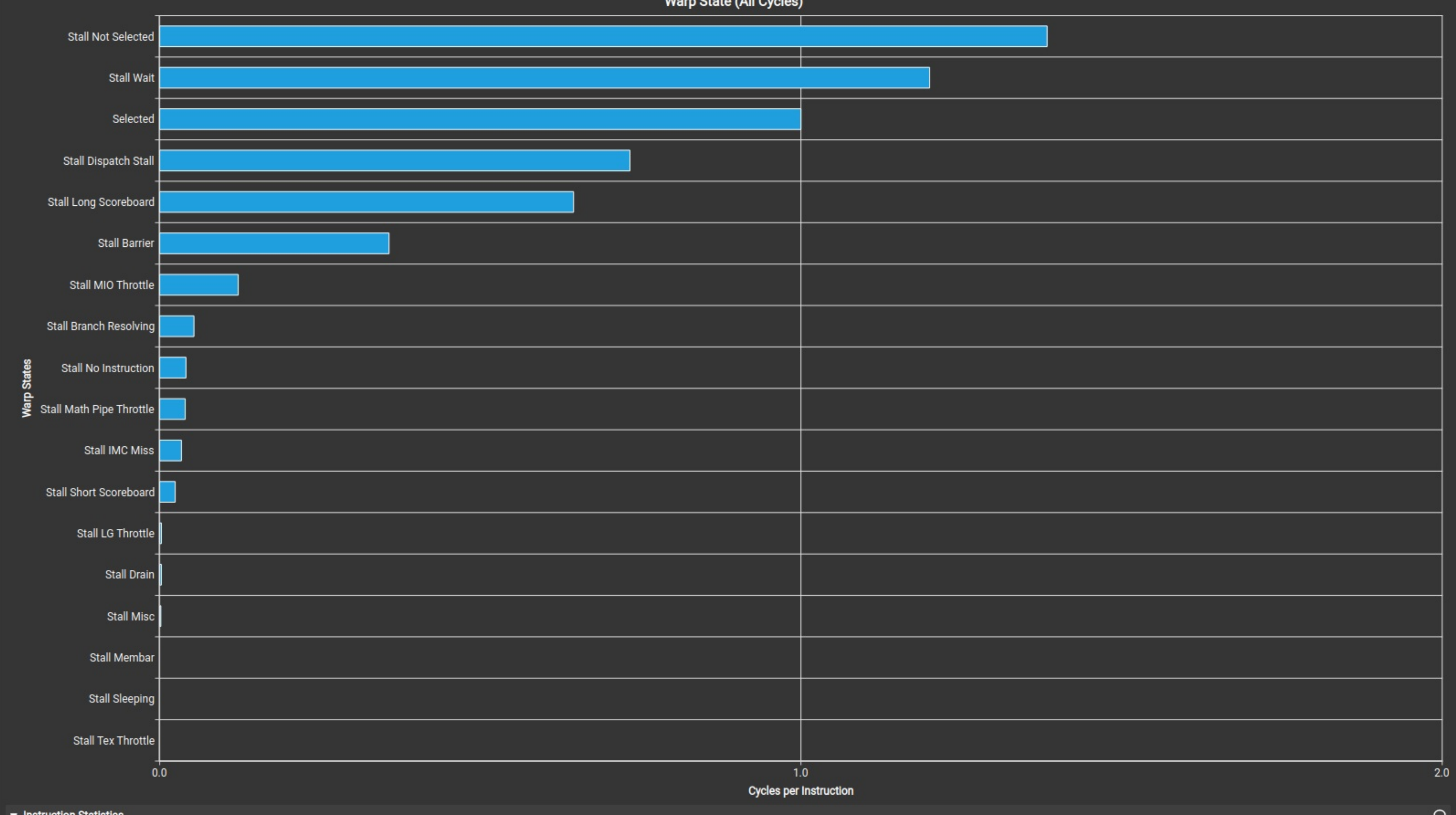
Warps Per Scheduler

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]	5.69	Avg. Active Threads Per Warp	32
Warp Cycles Per Executed Instruction [cycle]	5.69	Avg. Not Predicated Off Threads Per Warp	31.63

Warp State (All Cycles)

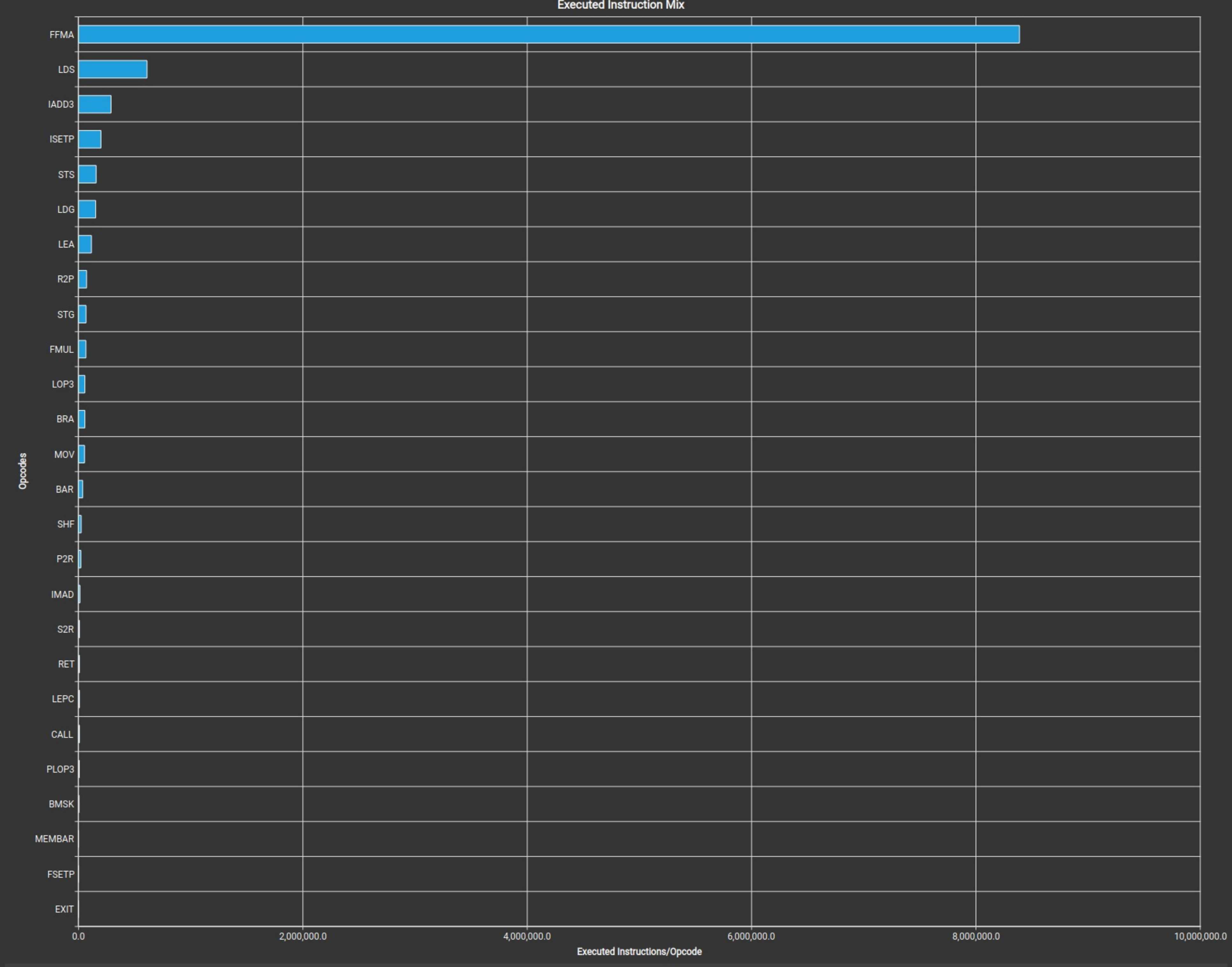


Instruction Statistics

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]	10,421,248	Avg. Executed Instructions Per Scheduler [inst]	93,046.86
Issued Instructions [inst]	10,426,456	Avg. Issued Instructions Per Scheduler [inst]	93,093.36

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.

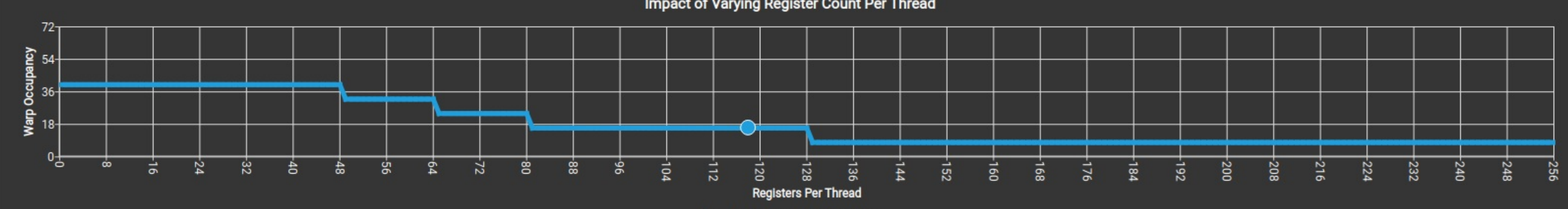
Grid Size	128	Registers Per Thread [register/thread]	118
Block Size	256	Static Shared Memory Per Block [Kbyte/block]	16.64
Threads [thread]	32,768	Dynamic Shared Memory Per Block [byte/block]	0
Waves Per SM	2.29	Driver Shared Memory Per Block [Kbyte/block]	1.02
Function Cache Configuration	cudaFuncCachePreferNone	Shared Memory Configuration Size [Kbyte]	65.54

Occupancy

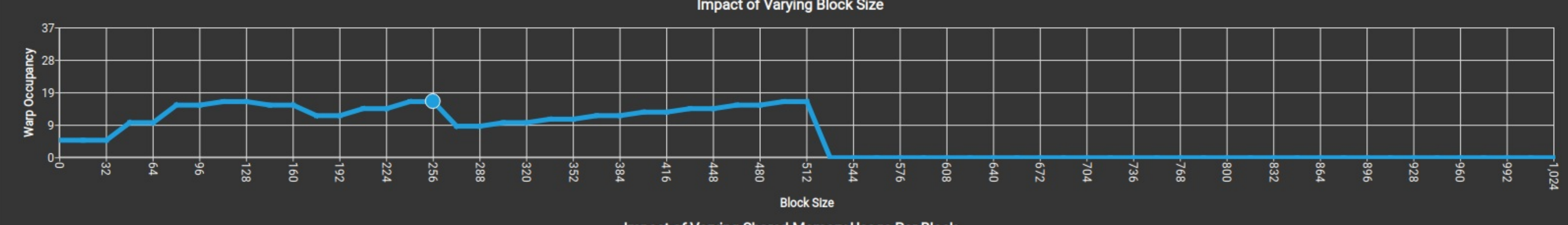
Occupancy is the ratio of the number of active warps per multiprocessor to the maximum number of possible active warps. Another way to view occupancy is the percentage of the hardware's ability to process warps that is actively in use. Higher occupancy does not always result in higher performance, however, 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 [%]	33.33	Block Limit Registers [block]	2
Theoretical Active Warps per SM [warp]	16	Block Limit Shared Mem [block]	5
Achieved Occupancy [%]	31.02	Block Limit Warps [block]	6
Achieved Active Warps Per SM [warp]	14.89	Block Limit SM [block]	16

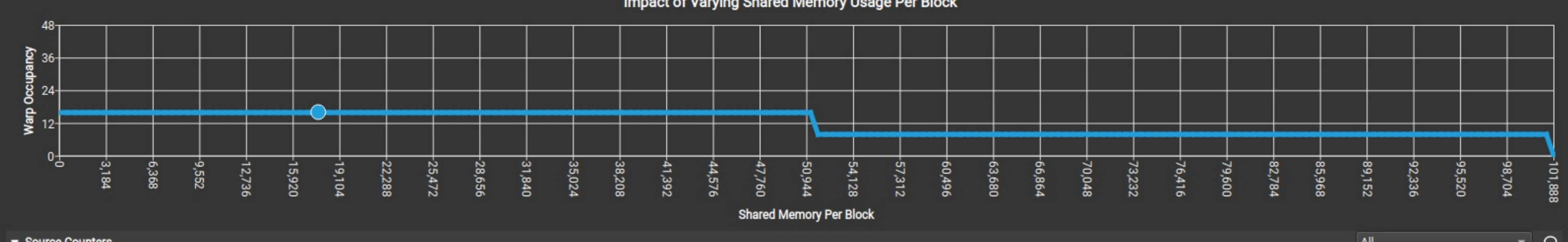
Impact of Varying Register Count Per Thread



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]	73,728	Branch Efficiency [%]	100
Branch Instructions Ratio	0.01	Avg. Divergent Branches	0

Sampling Data (All)

Location	Value	Value (%)
0x7fbf237e70b0 in ampere_sgemm_128x128_nn	154	4
0x7fbf237e56c0 in ampere_sgemm_128x128_nn	153	4
0x7fbf237e7810 in ampere_sgemm_128x128_nn	138	3
0x7fbf237e92b0 in ampere_sgemm_128x128_nn	95	2
0x7fbf237e6180 in ampere_sgemm_128x128_nn	75	2

Sampling Data (Not Issued)

Location	Value	Value (%)
0x7fbf237e56c0 in ampere_sgemm_128x128_nn	124	18
0x7fbf237e70b0 in ampere_sgemm_128x128_nn	98	14
0x7fbf237e5070 in ampere_sgemm_128x128_nn	44	6
0x7fbf237e7810 in ampere_sgemm_128x128_nn	42	6
0x7fbf237e5250 in ampere_sgemm_128x128_nn	34	5

Most Instructions Executed

Location	Value	Value (%)
0x7fbf237e67c0 in ampere_sgemm_128x128_nn	16,384	0
0x7fbf237e6730 in ampere_sgemm_128x128_nn	16,384	0
0x7fbf237e6740 in ampere_sgemm_128x128_nn	16,384	0
0x7fbf237e6750 in ampere_sgemm_128x128_nn	16,384	0
0x7fbf237e6760 in ampere_sgemm_128x128_nn	16,384	0