Page: Details | Launch: 3 - 634 - MM_v1 | Add Baseline | Apply Rules | Save as PDF

	Launch	Time	Cycles	Regs	GPU	SM Frequency	CC	Process
Current	634 - MM_v1 (128, 1, 1)x(3...	8.41 msecond	11,097,721	37	NVIDIA GeForce RTX 3060	1.32 cycle/nsecond	8.6	[9136] MM_v1

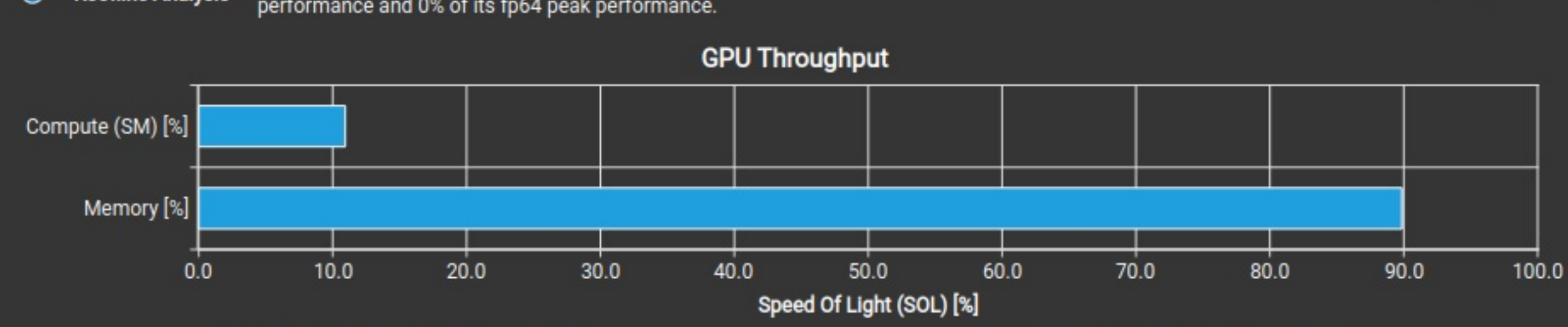
GPU Speed Of Light Throughput

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 [%]	10.94	Duration [msecond]	8.41
Memory Throughput [%]	89.86	Elapsed Cycles [cycle]	11,097,721
L1/TEX Cache Throughput [%]	98.29	SM Active Cycles [cycle]	10,145,569.54
L2 Cache Throughput [%]	3.10	SM Frequency [cycle/nsecond]	1.32
DRAM Throughput [%]	2.33	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 close to 1% of this device's fp32 peak performance and 0% of its fp64 peak performance.

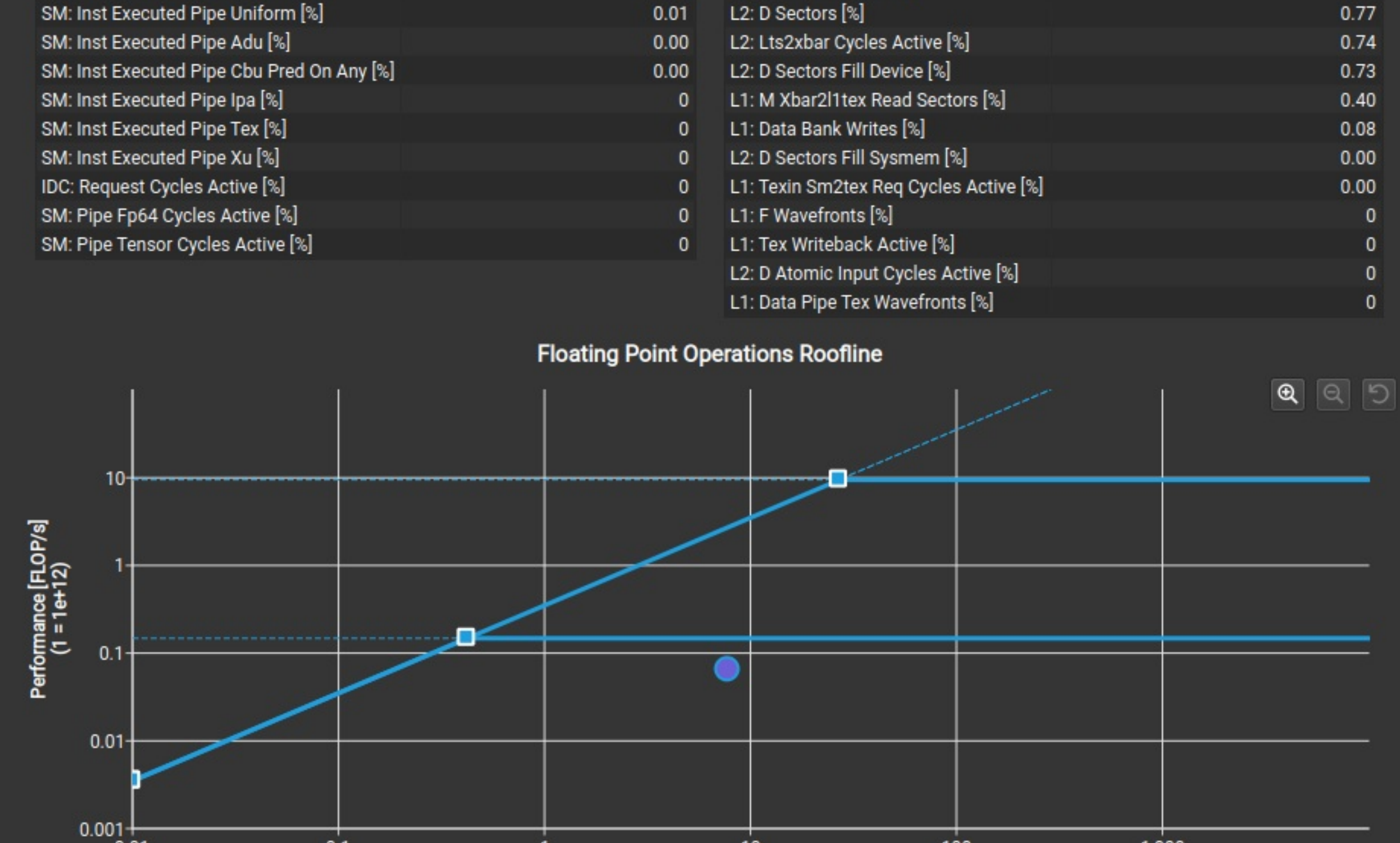


Compute Throughput Breakdown

SM: Inst Executed Pipe Lsu [%]	10.94
SM: Mio Inst Issued [%]	3.65
SM: Mio2rf Writeback Active [%]	2.71
SM: Issue Active [%]	2.47
SM: Inst Executed [%]	2.47
SM: Pipe Fma Cycles Active [%]	0.73
SM: Pipe Alu Cycles Active [%]	0.67
SM: Pipe Fmaheavy Cycles Active [%]	0.13
SM: Mio Pq Read Cycles Active [%]	0.05
SM: Mio Pq Write Cycles Active [%]	0.04
SM: Inst Executed Pipe Uniform [%]	0.01
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 [%]	89.86
L1: Lsuin Requests [%]	10.94
L1: Lsu Writeback Active [%]	5.60
L2: Xbar2lts Cycles Active [%]	3.10
DRAM: Cycles Active [%]	2.33
L2: T Sectors [%]	2.15
L1: M L1tex2xbar Req Cycles Active [%]	1.82
L2: T Tag Requests [%]	1.80
DRAM: Dram Sectors [%]	1.70
L1: Data Bank Reads [%]	1.43
L2: D Sectors [%]	0.77
L2: Lts2xbar Cycles Active [%]	0.74
L2: D Sectors Fill Device [%]	0.73
L1: M Xbar2l1tex Read Sectors [%]	0.40
L1: Data Bank Writes [%]	0.08
L2: D Sectors Fill Sysmem [%]	0.00
L1: Texin Sm2tex Req Cycles Active [%]	0.00
L1: F Wavefronts [%]	0
L1: Tex Writeback Active [%]	0
L2: D Atomic Input Cycles Active [%]	0
L1: Data Pipe Tex Wavefronts [%]	0

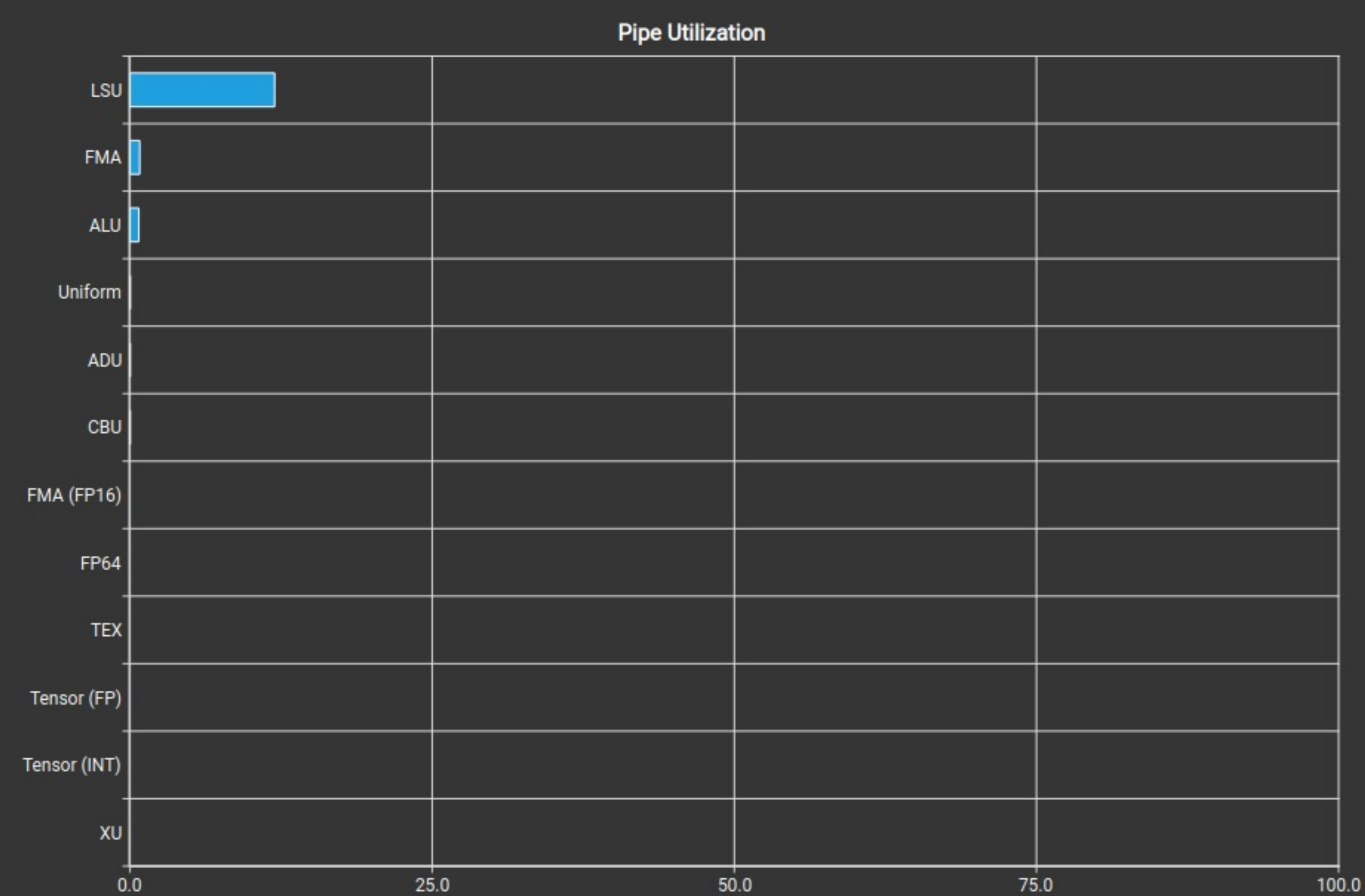


Compute Workload Analysis

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]	0.10	SM Busy [%]	3.99
Executed Ipc Active [inst/cycle]	0.11	Issue Slots Busy [%]	2.70
Issued Ipc Active [inst/cycle]	0.11		

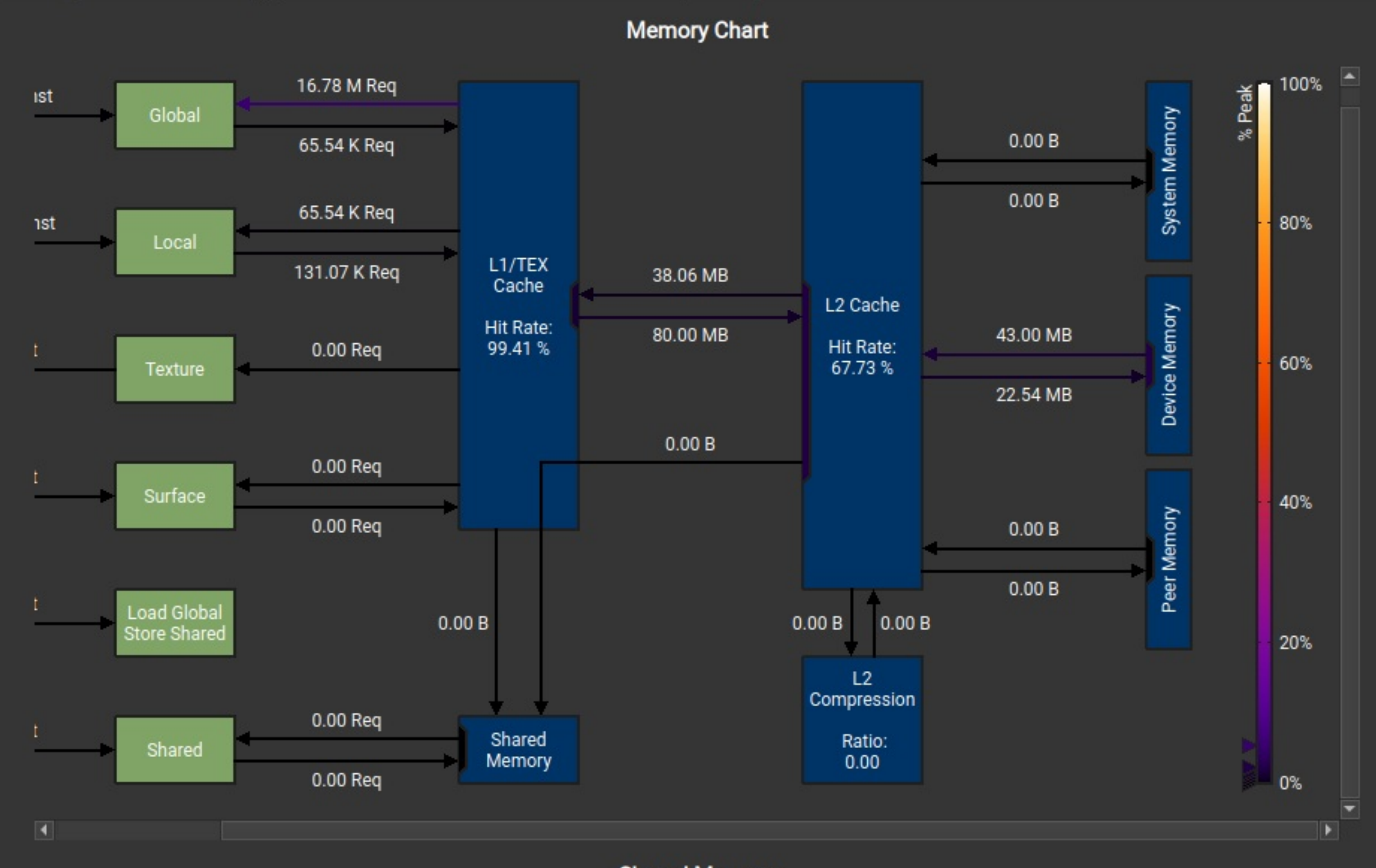
Low Utilization All pipelines are under-utilized. Either this kernel is very small or it doesn't issue enough warps per scheduler. Check the Launch Statistics and Scheduler Statistics sections for further details.



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]	8.17	Mem Busy [%]	89.86
L1/TEX Hit Rate [%]	99.41	Max Bandwidth [%]	10.94
L2 Hit Rate [%]	67.73	Mem Pipes Busy [%]	10.94
L2 Compression Success Rate [%]	0	L2 Compression Ratio	0



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	-	-	8,192	0.00	0
Total	0	0	8,192	0.00	0

L1/TEX Cache

	Instructions	Requests	Wavefronts	% Peak	Sectors	Se
Global Load	16,777,216	16,777,216			276,824,063	
Global Load To Shared Store (access)	0	0	75,497,644	24.30	0	
Global Load To Shared Store (bypass)	0	0			0	
Surface Load	0	0	0	0	0	
Texture Load	0	0	0	0	0	
Global Store	65,536	65,536	524,300	0.17	2,097,152	
Local Store	81,920	131,072	131,072	0.04	524,376	
Surface Store	0	0	0	0	0	
Global Reduction	0	0	0	0	0	
Surface Reduction	0	0	0	0	0	
Global Atomic ALU	0	0	0	0	0	
Global Atomic CAS	0					
Surface Atomic ALU	0	0	0	0	0	
Surface Atomic CAS	0					
Loads	16,842,752	16,842,752	75,563,180	24.32	277,085,791	
Stores	147,456	196,608	655,372	0.21	2,621,528	
Total	16,990,208	17,039,360	76,218,552	24.53	279,707,319	

L2 Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes
L1/TEX Store	2,228,224	2,621,440	1.18	1.38	100	83,886,080
L1/TEX Atomic ALU	0	0	0	0	0	0
L1/TEX Atomic CAS	0	0	0	0	0	0
L1/TEX Reduction	0	0	0	0	0	0
L1/TEX Total	3,299,636	3,869,364	1.17	2.04	70.15	123,819,648
GPU Total	3,317,044	4,074,596	1.23	2.15	69.11	130,387,072

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	1,408,900	1.53	45,084,800	5,360,111,699.11
Store	738,532	0.80	23,633,024	2,809,719,648.92
Total	2,147,432	2.33	68,717,824	8,169,831,348.04

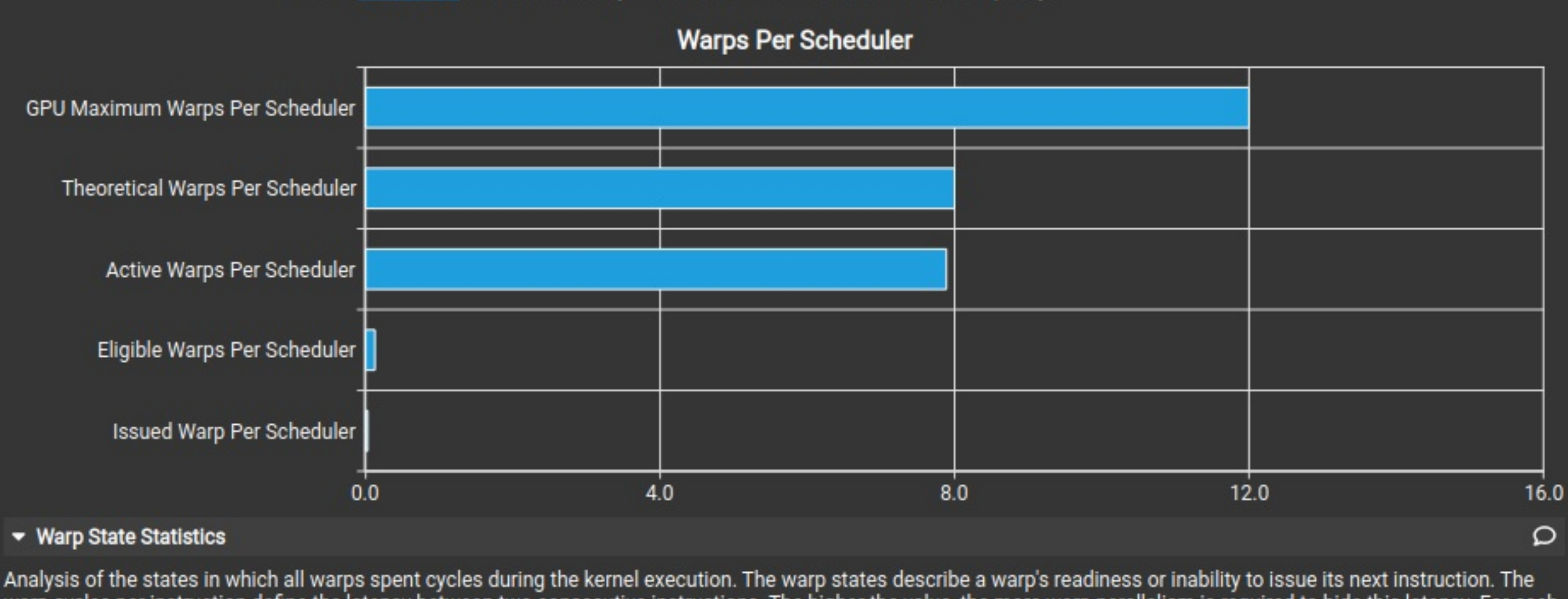
Scheduler Statistics

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]	7.89	No Eligible [%]	97.30
Eligible Warps Per Scheduler [warp]	0.13	One or More Eligible [%]	2.70
Issued Warp Per Scheduler	0.03		

Issue Slot Utilization Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 37.0 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 7.89 active warps per scheduler, but only an average of 0.13 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.

Issue Slot Utilization The 8.00 theoretical warps per scheduler this kernel can issue according to its occupancy are below the hardware maximum of 12. Use the Occupancy section to identify what limits this kernel's theoretical occupancy.



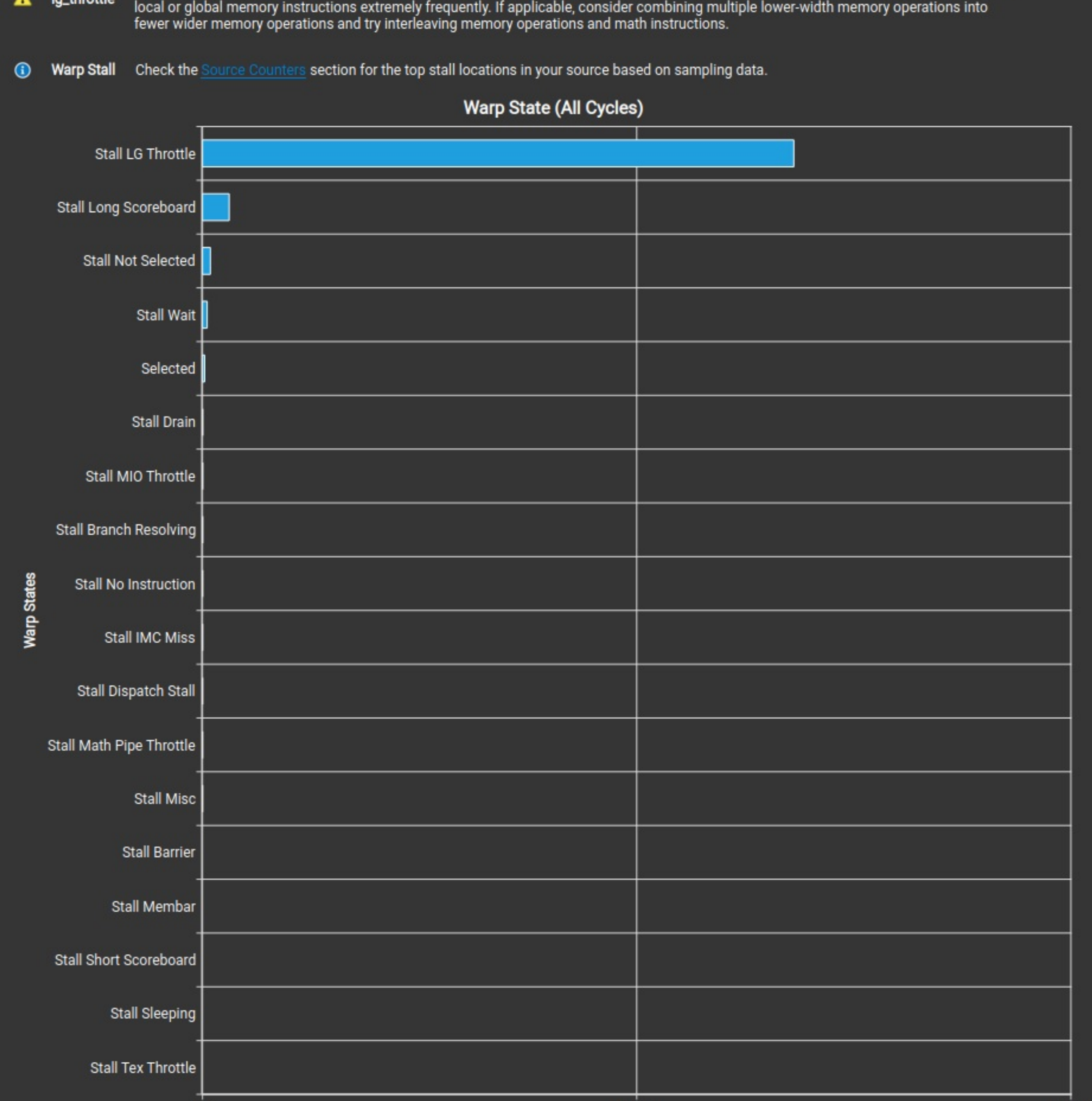
Warp State Statistics

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]	291.82	Avg. Active Threads Per Warp	32
Warp Cycles Per Executed Instruction [cycle]	291.89	Avg. Not Predicated Off Threads Per Warp	31.93

lg_throttle On average, each warp of this kernel spends 272.4 cycles being stalled waiting for the local/global instruction queue to be not full. This represents about 93.3% of the total average of 291.8 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.

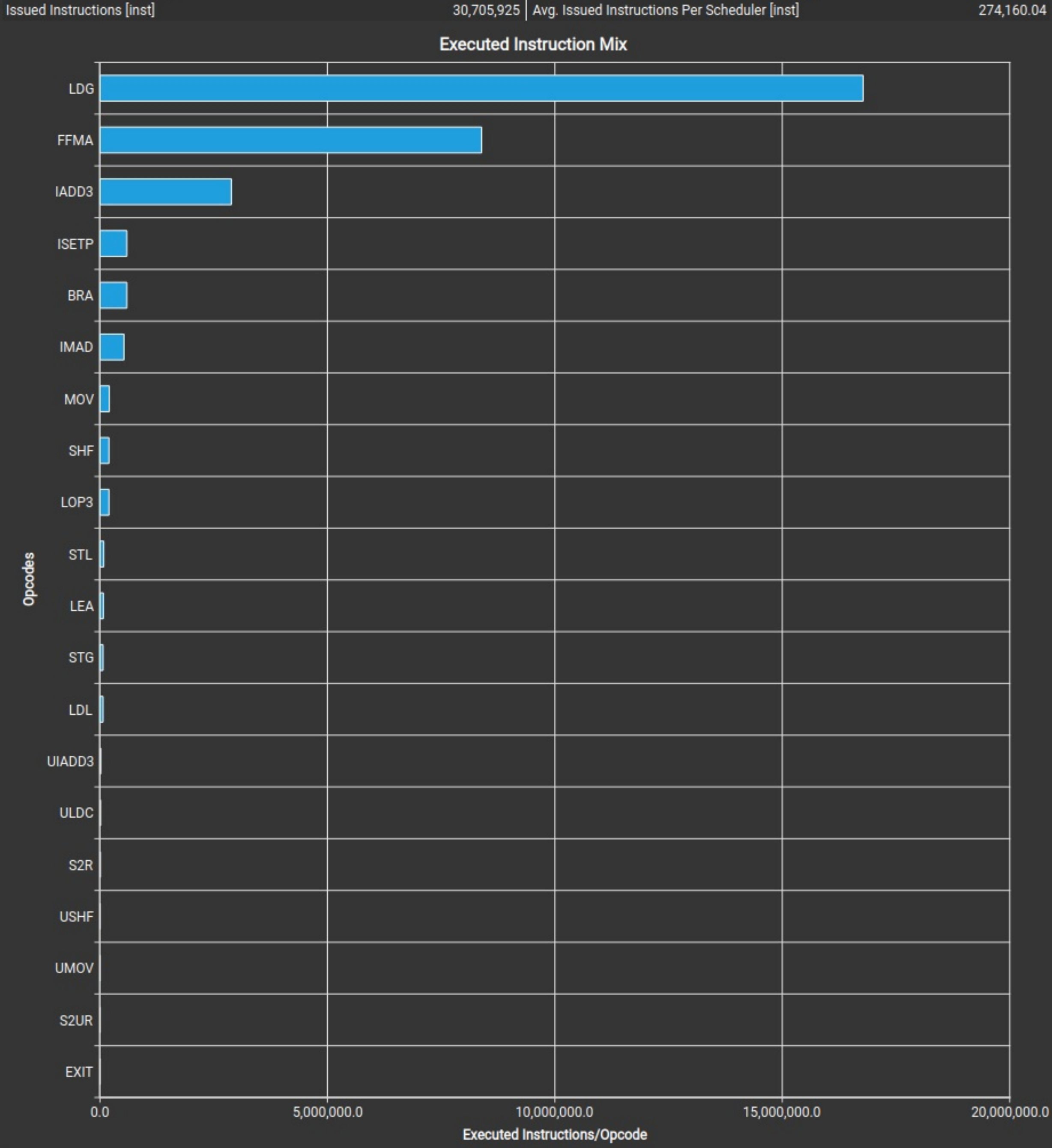
Warp Stall Check the Source Counters section for the top stall locations in your source based on sampling data.



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]	30,699,520	Avg. Executed Instructions Per Scheduler [inst]	274,102.86
Issued Instructions [inst]	30,705,925	Avg. Issued Instructions Per Scheduler [inst]	274,160.04



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.

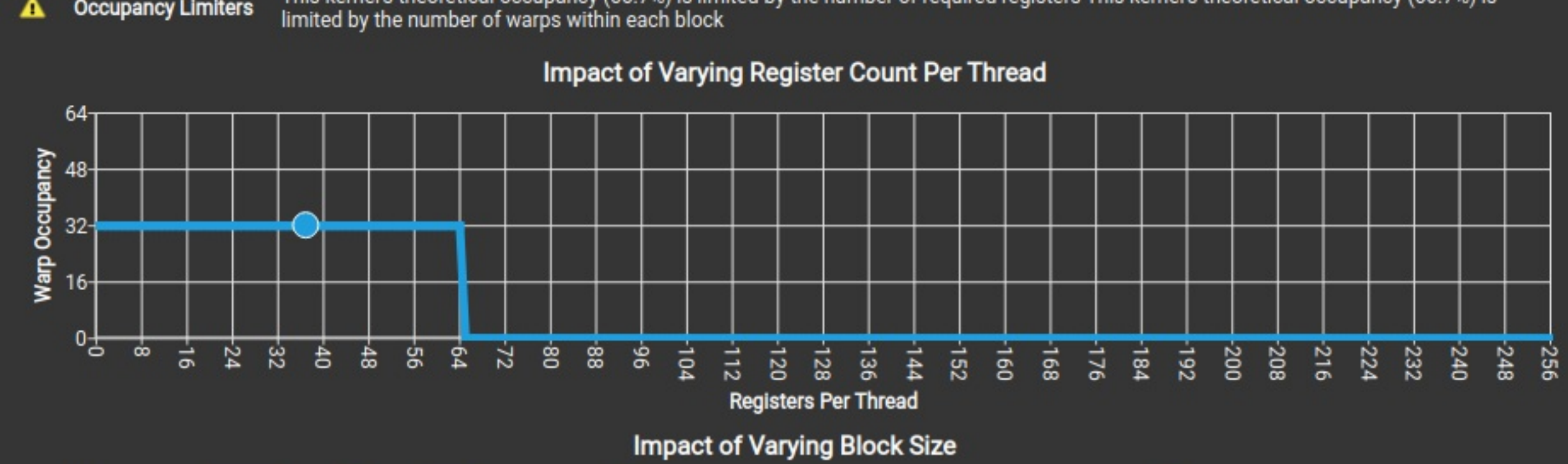
Grid Size	128	Registers Per Thread [register/thread]	37
Block Size	1,024	Static Shared Memory Per Block [byte/block]	0
Threads [thread]	131,072	Dynamic Shared Memory Per Block [byte/block]	0
Waves Per SM	4.57	Driver Shared Memory Per Block [Kbyte/block]	1.02
Function Cache Configuration	cudaFuncCachePreferNone	Shared Memory Configuration Size [Kbyte]	8.19

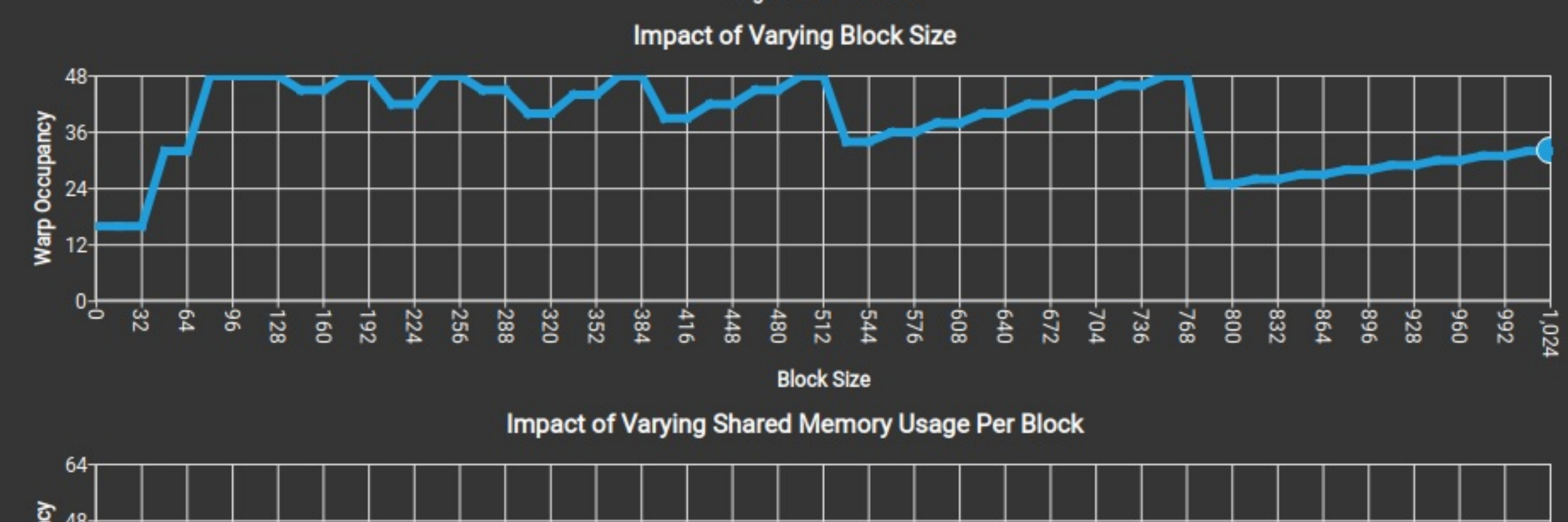
Occupancy

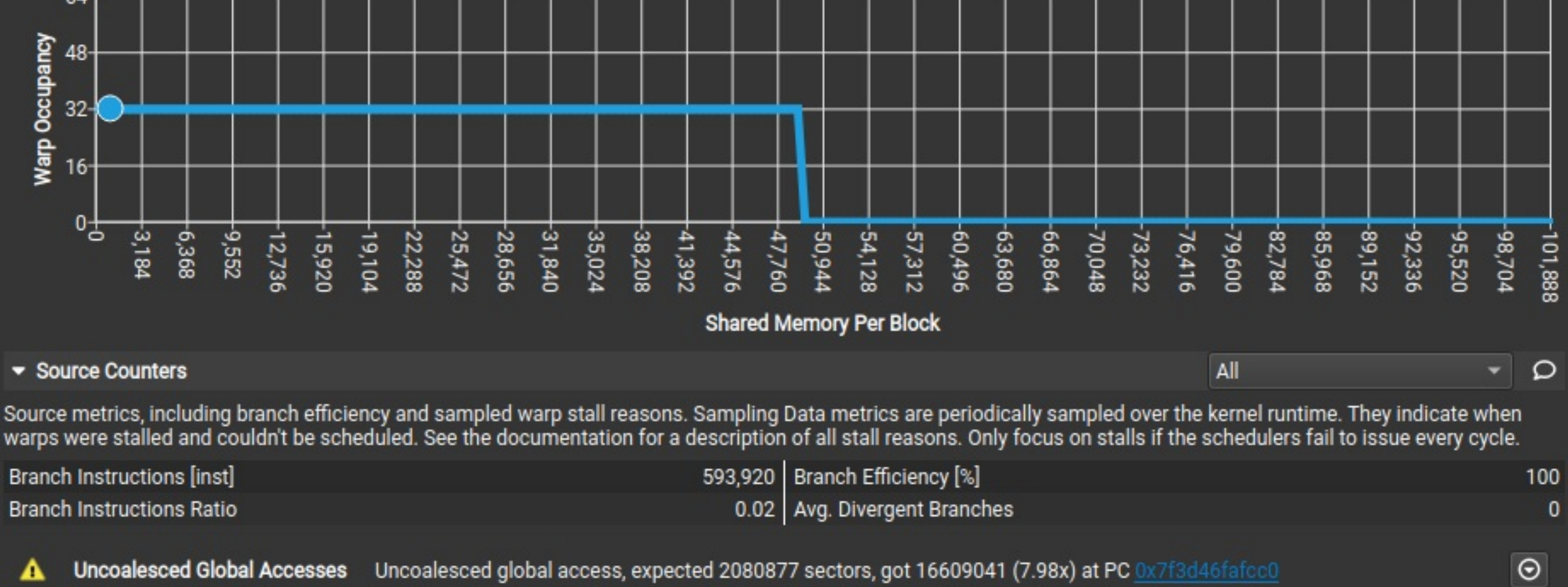
Occupancy is the ratio of the number of active warps per multiprocessor to the maximum number of possible active warps. Another way to view occupancy is the percentage of the hardware's ability to process warps that is actively in use. Higher occupancy does not always result in higher performance, however, 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 [%]	66.67	Block Limit Registers [block]	1
Theoretical Active Warps per SM [warp]	32	Block Limit Shared Mem [block]	100
Achieved Occupancy [%]	65.71	Block Limit Warps [block]	1
Achieved Active Warps Per SM [warp]	31.54	Block Limit SM [block]	16

Occupancy Limiters This kernel's theoretical occupancy (66.7%) is limited by the number of required registers This kernel's theoretical occupancy (66.7%) is limited by the number of warps within each 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]	593,920	Branch Efficiency [%]	100
Branch Instructions Ratio	0.02	Avg. Divergent Branches	0

Uncoalesced Global Accesses Uncoalesced global access, expected 2080877 sectors, got 16609041 (7.98x) at PC 0x7f3d46fafcc0

Uncoalesced Global Accesses Uncoalesced global access, expected 2080787 sectors, got 16608111 (7.98x) at PC 0x7f3d46fafcb0

Uncoalesced Global Accesses Uncoalesced global access, expected 2079290 sectors, got 16592642 (7.98x) at PC 0x7f3d46fafbb0

Uncoalesced Global Accesses Uncoalesced global access, expected 2074721 sectors, got 16545429 (7.97x) at PC 0x7f3d46fafd40

Uncoalesced Global Accesses Uncoalesced global access, expected 2074637 sectors, got 16544561 (7.97x) at PC 0x7f3d46fafbf0

Uncoalesced Global Accesses Uncoalesced global access, expected 2074433 sectors, got 16542453 (7.97x) at PC 0x7f3d46fafdf0

Uncoalesced Global Accesses Uncoalesced global access, expected 2074409 sectors, got 16542205 (7.97x) at PC 0x7f3d46fafda0

Uncoalesced Global Accesses Uncoalesced global access, expected 2074358 sectors, got 16541678 (7.97x) at PC 0x7f3d46fafd70

Uncoalesced Global Accesses Uncoalesced global access, expected 2074355 sectors, got 16541647 (7.97x) at PC 0x7f3d46fafd10

Uncoalesced Global Accesses Uncoalesced global access, expected 2074349 sectors, got 16541585 (7.97x) at PC 0x7f3d46fafc70

Sampling Data (All)

Location	Value	Value (%)
0x7f3d46fafba0 in MM_v1	34,271	12
0x7f3d46fafce0 in MM_v1	11,146	4
0x7f3d46fafd10 in MM_v1	10,682	4
0x7f3d46fafd30 in MM_v1	10,668	4
0x7f3d46fafd60 in MM_v1	10,562	4

Sampling Data (Not Issued)

Location	Value	Value (%)
0x7f3d46fafba0 in MM_v1	33,407	12
0x7f3d46fafce0 in MM_v1	10,869	4
0x7f3d46fafd10 in MM_v1	10,426	4
0x7f3d46fafd30 in MM_v1	10,385	4
0x7f3d46fafd60 in MM_v1	10,286	4

Most Instructions Executed

Location	Value	Value (%)
0x7f3d46fafc40 in MM_v1	524,288	2
0x7f3d46fafce0 in MM_v1	524,288	2
0x7f3d46fafcd0 in MM_v1	524,288	2
0x7f3d46fafcc0 in MM_v1	524,288	2
0x7f3d46fafcb0 in MM_v1	524,288	2