Page: Details Result: 1 - 538 - shared_mem_mmult Clear Baselines Apply Rules Occupancy Calculator Save as PDF

	Report	Result	Time	Cycles	Regs	GPU	SM Frequency	CC	Process
🟦 **Current**	m...0	538 - shared_mem_m...	31.20 msecond	46,802,969	27	0 - NVIDIA GeForce RTX 3070	1.50 cycle/nsecond	8.6	[16620] HW3.exe
🟩 **100x100_global**	m...0	535 - global_mem_mm...	1.86 msecond	2,791,525	40	0 - NVIDIA GeForce RTX 3070	1.50 cycle/nsecond	8.6	[15504] HW3.exe
🟪 **100x100_shared**	m...0	538 - shared_mem_m...	2.14 msecond	3,215,351	27	0 - NVIDIA GeForce RTX 3070	1.50 cycle/nsecond	8.6	[15504] HW3.exe

🟧 **1000x1000_global** 🟥 **1000x1000_shared** 🟨 **2000x3000_global** 🟫 **2000x3000_shared**

GPU Speed Of Light Throughput

All

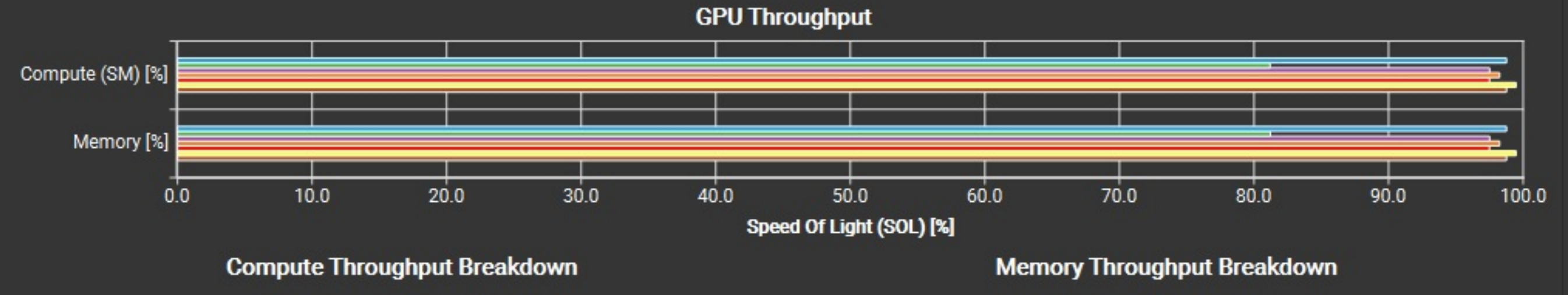
High-level overview of the throughput for compute and memory resources of the GPU. For each unit, the throughput reports the achieved percentage of utilization with respect to the theoretical maximum. Breakdowns show the throughput for each individual sub-metric of Compute and Memory to clearly identify 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 [%]	98.76 (+3.47%, z=+0.47)	Duration [msecond]	31.20 (+181.06%, z=+1.24)
Memory Throughput [%]	98.76 (+3.47%, z=+0.47)	Elapsed Cycles [cycle]	46,802,969 (+181.02%, z=+1.24)
L1/TEX Cache Throughput [%]	98.81 (-0.09%, z=-0.19)	SM Active Cycles [cycle]	46,781,669.26 (+182.67%, z=+1.24)
L2 Cache Throughput [%]	17.29 (-2.75%, z=-0.27)	SM Frequency [cycle/nsecond]	1.50 (-0.05%, z=-0.29)
DRAM Throughput [%]	29.81 (+112.82%, z=+1.05)	DRAM Frequency [cycle/nsecond]	6.79 (-0.06%, z=-0.34)

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 5% 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



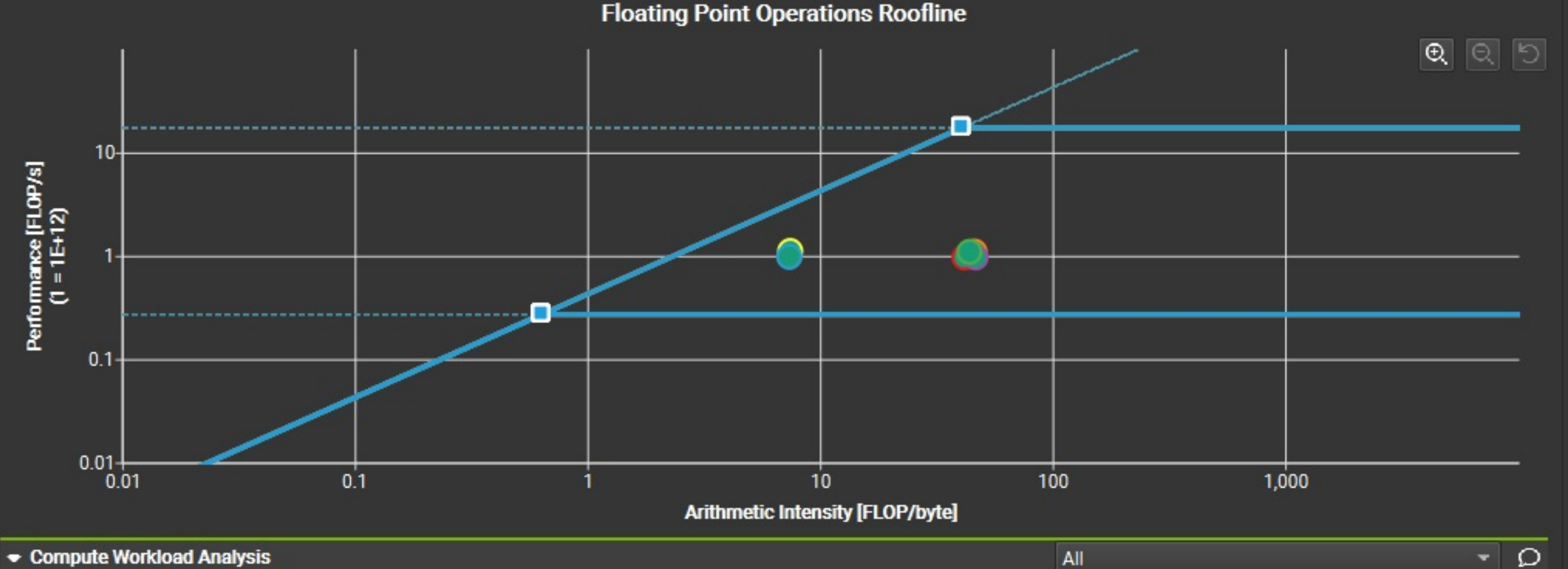
Compute Throughput Breakdown

SM: Inst Executed Pipe Lsu [%]	98.76 (+3.47%, z=+0.47)
SM: Issue Active [%]	35.02 (+16.00%, z=+0.86)
SM: Inst Executed [%]	35.02 (+16.00%, z=+0.86)
SM: Mio Inst Issued [%]	34.76 (+6.15%, z=+0.68)
SM: Pipe Alu Cycles Active [%]	25.38 (+74.78%, z=+0.88)
SM: Mio2rf Writeback Active [%]	23.38 (-3.94%, z=-0.27)
SM: Pipe Fma Cycles Active [%]	9.28 (-30.42%, z=-0.83)
SM: Pipe Fmaheavy Cycles Active [%]	7.80 (-51.43%, z=-0.85)
SM: Inst Executed Pipe Adu [%]	5.52 (+97.92%, z=+0.87)
SM: Mio Pq Read Cycles Active [%]	2.74 (+100.58%, z=+0.88)
SM: Mio Pq Write Cycles Active [%]	2.74 (+100.58%, z=+0.88)
SM: Inst Executed Pipe Uniform [%]	0.69 (+96.76%, z=+0.86)
SM: Inst Executed Pipe Xu [%]	0.02 (+1.40%, z=+0.01)
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.01 (-53.69%, z=-1.22)
IDC: Request Cycles Active [%]	0 (+0.00%, z=+0.00)
SM: Inst Executed Pipe Ipa [%]	0 (+0.00%, z=+0.00)
SM: Inst Executed Pipe Tex [%]	0 (+0.00%, z=+0.00)
SM: Pipe Fp64 Cycles Active [%]	0 (+0.00%, z=+0.00)
SM: Pipe Tensor Cycles Active [%]	0 (+0.00%, z=+0.00)

Memory Throughput Breakdown

L1: Lsuin Requests [%]	98.76 (+3.47%, z=+0.47)
L1: Data Pipe Lsu Wavefronts [%]	53.83 (+0.52%, z=+0.04)
L1: Lsu Writeback Active [%]	47.75 (-5.32%, z=-0.35)
DRAM: Cycles Active [%]	29.81 (+112.82%, z=+1.05)
DRAM: Dram Sectors [%]	17.96 (+113.37%, z=+1.05)
L2: T Sectors [%]	17.29 (-2.75%, z=-0.27)
L2: Lts2xbar Cycles Active [%]	17.24 (-0.65%, z=-0.06)
L1: M Xbar2l1tex Read Sectors [%]	11.02 (-2.33%, z=-0.23)
L2: T Tag Requests [%]	10.57 (-15.91%, z=-0.66)
L2: Xbar2lts Cycles Active [%]	9.71 (-19.01%, z=-0.72)
L2: D Sectors Fill Device [%]	8.89 (+117.86%, z=+1.04)
L1: Data Bank Reads [%]	7.51 (+7.81%, z=+0.77)
L2: D Sectors [%]	6.38 (+20.06%, z=+0.81)
L1: M L1tex2xbar Req Cycles Active [%]	5.92 (-17.21%, z=-0.65)
L1: Data Bank Writes [%]	2.74 (+34.10%, z=+0.98)
L1: Texin Sm2tex Req Cycles Active [%]	0.34 (+101.58%, z=+0.88)
L2: D Sectors Fill Sysmem [%]	0.00 (+69.27%, z=+0.00)
L1: Data Pipe Tex Wavefronts [%]	0 (+0.00%, z=+0.00)
L1: F Wavefronts [%]	0 (+0.00%, z=+0.00)
L1: Tex Writeback Active [%]	0 (+0.00%, z=+0.00)
L2: D Atomic Input Cycles Active [%]	0 (+0.00%, z=+0.00)

Floating Point Operations Roofline



Compute Workload Analysis

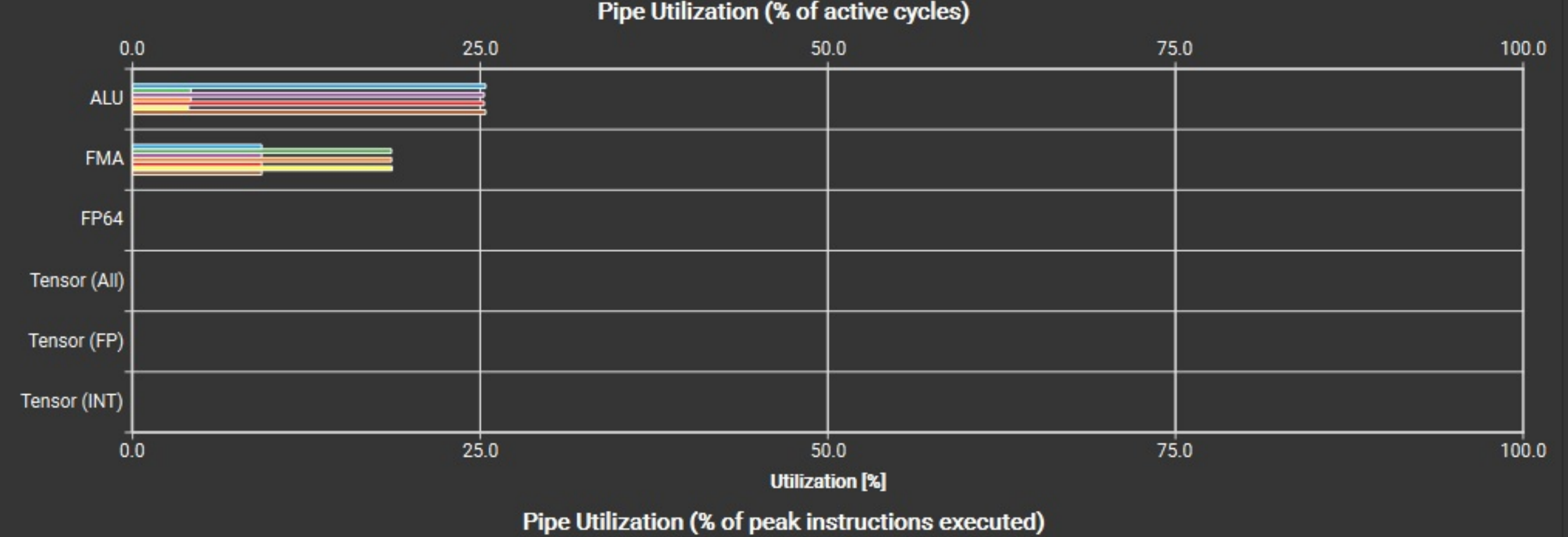
All

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.

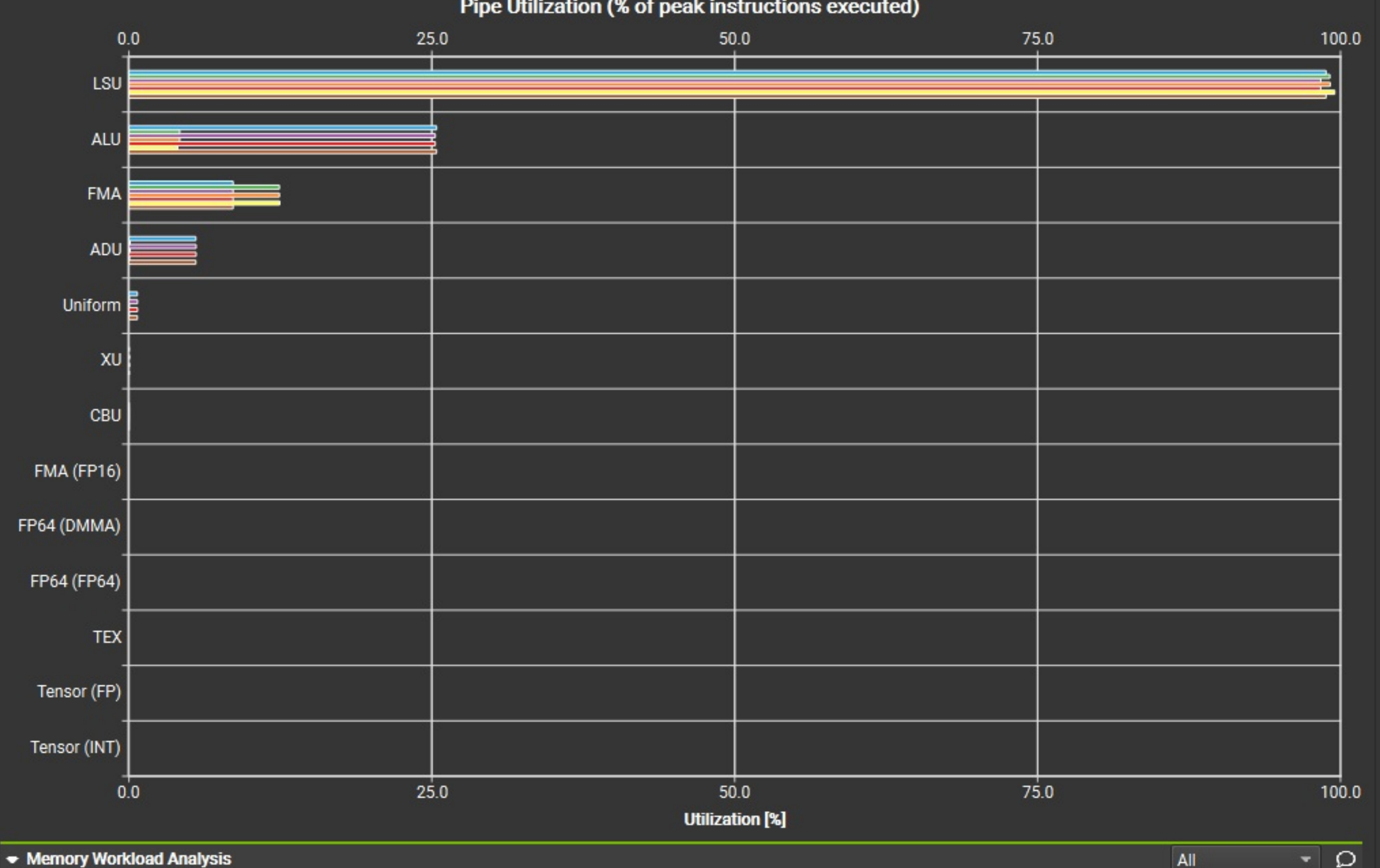
Executed Ipc Elapsed [inst/cycle]	1.40 (+16.00%, z=+0.86)	SM Busy [%]	35.04 (+2.91%, z=+0.91)
Executed Ipc Active [inst/cycle]	1.40 (+12.42%, z=+0.88)	Issue Slots Busy [%]	35.04 (+12.42%, z=+0.88)
Issued Ipc Active [inst/cycle]	1.40 (+12.42%, z=+0.88)		

Balanced ALU is the highest-utilized pipeline (25.4%) based on active cycles, taking into account the rates of its different instructions. It executes integer and logic 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/second]	129.62 (+112.76%, z=+1.05)	Mem Busy [%]	53.83 (+0.52%, z=+0.04)
L1/TEX Hit Rate [%]	11.00 (-77.23%, z=-0.82)	Max Bandwidth [%]	98.76 (+3.47%, z=+0.47)
L2 Hit Rate [%]	48.49 (-36.02%, z=-1.15)	Mem Pipes Busy [%]	98.76 (+3.47%, z=+0.47)
L2 Compression Success Rate [%]	0 (+0.00%, z=+0.00)	L2 Compression Ratio	0 (+0.00%, z=+0.00)

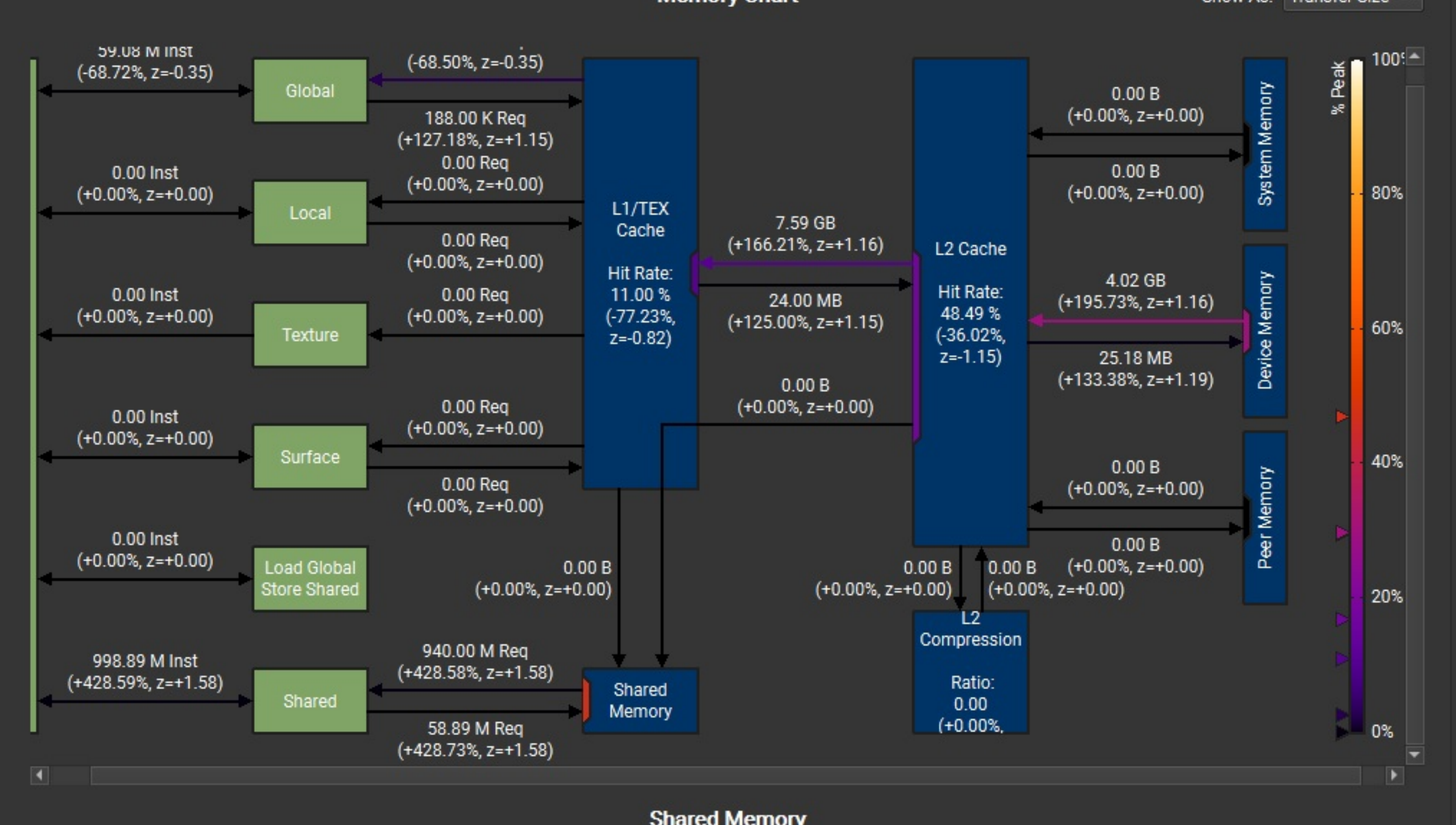
L1TEX Global Load Access Pattern The memory access pattern for global loads in L1TEX might not be optimal. On average, this kernel accesses 4.0 bytes per thread per memory request; but the address pattern, possibly caused by the stride between threads, results in 4.1 sectors per request, or 4.1*32 = 132.4 bytes of cache data transfers per request. The optimal thread address pattern for 4.0 byte accesses would result in 4.0*32 = 128.0 bytes of cache data transfers per request, to maximize L1TEX cache performance. Check the Source Counters section for uncoalesced global loads.

L2 Load Access Pattern The memory access pattern for loads from L1TEX to L2 is not optimal. The granularity of an L1TEX request to L2 is a 128 byte cache line. That is 4 consecutive 32-byte sectors per L2 request. However, this kernel only accesses an average of 1.6 sectors out of the possible 4 sectors per cache line. Check the Source Counters section for uncoalesced loads and try to minimize how many cache lines need to be accessed per memory request.

L2 Store Access Pattern The memory access pattern for stores from L1TEX to L2 is not optimal. The granularity of an L1TEX request to L2 is a 128 byte cache line. That is 4 consecutive 32-byte sectors per L2 request. However, this kernel only accesses an average of 1.6 sectors out of the possible 4 sectors per cache line. Check the Source Counters section for uncoalesced stores and try to minimize how many cache lines need to be accessed per memory request.

Memory Chart

Show As: Transfer Size



Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	940,000,000 (+428.58%, z=+1.58)	940,000,000 (+428.58%, z=+1.58)	940,090,625 (+428.57%, z=+1.58)	43.67 (+102.22%, z=+0.88)	70,722 (+345.91%, z=+1.56)
Shared Load Matrix	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)			
Shared Store	58,891,000 (+428.73%, z=+1.58)	58,891,000 (+428.73%, z=+1.58)	58,891,000 (+428.73%, z=+1.58)	0.68 (+102.56%, z=+0.89)	0 (+0.00%, z=+0.00)
Shared Store From Global Load	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Shared Atomic	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Other	-	-	14,195,245 (+395.49%, z=+1.58)	2.71 (+90.51%, z=+0.81)	1,646,414 (+426.00%, z=+1.58)
Total	998,891,000 (+428.59%, z=+1.58)	998,891,000 (+428.59%, z=+1.58)	1,013,176,870 (+428.09%, z=+1.58)	47.06 (+101.51%, z=+0.88)	1,717,136 (+422.13%, z=+1.58)

L1/TEX Cache

	Instructions	Requests	Wavefronts	% Peak	Sectors	Sectors/Req
Local Load	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Global Load	58,891,000 (-68.81%, z=-0.35)	58,891,000 (-68.50%, z=-0.35)	82,286,780 (-61.64%, z=-0.32)	3.82 (-86.13%, z=-0.85)	243,724,281 (-39.00%, z=-0.22)	4.14 (+40.91%, z=+1.15)
Global Load To Shared Store (access)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)			0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Global Load To Shared Store (bypass)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)			0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Surface Load	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Texture Load	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Global Store	188,000 (+124.70%, z=+1.15)	188,000 (+127.18%, z=+1.15)	251,745 (+129.91%, z=+1.15)	0.01 (-48.53%, z=-1.20)	750,000 (+125.00%, z=+1.15)	3.99 (-3.04%, z=-0.37)
Local Store	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Surface Store	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Global Reduction	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Surface Reduction	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Global Atomic ALU	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Global Atomic CAS	0 (+0.00%, z=+0.00)					
Surface Atomic ALU	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Surface Atomic CAS	0 (+0.00%, z=+0.00)					
Loads	58,891,000 (-68.81%, z=-0.35)	58,891,000 (-68.50%, z=-0.35)	82,286,780 (-61.64%, z=-0.32)	3.82 (-86.13%, z=-0.85)	243,724,281 (-39.00%, z=-0.22)	4.14 (+40.91%, z=+1.15)
Stores	188,000 (+124.70%, z=+1.15)	188,000 (+127.18%, z=+1.15)	251,745 (+129.91%, z=+1.15)	0.01 (-48.53%, z=-1.20)	750,000 (+125.00%, z=+1.15)	3.99 (-3.04%, z=-0.37)
Atomics & Reductions	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)
Total	59,079,000 (-68.72%, z=-0.35)	59,079,000 (-68.42%, z=-0.35)	82,538,525 (-61.54%, z=-0.32)	3.83 (-86.10%, z=-0.85)	244,474,281 (-38.86%, z=-0.22)	4.14 (+40.83%,

L2 Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes	Throughput
L1/TEX Load	146,947,985 (+130.03%, z=+0.93)	237,317,986 (+166.18%, z=+1.16)	1.61 (+13.98%, z=+0.92)	16.86 (-1.67%, z=-0.16)	49.36 (-36.76%, z=-1.15)	7,594,175,552 (+166.18%, z=+1.16)	243,370,113,994
L1/TEX Store	469,500 (+124.82%, z=+1.15)	750,000 (+125.00%, z=+1.15)	1.60 (+0.21%, z=+1.15)	0.05 (-51.88%, z=-1.19)	100 (+0.00%, z=+0.00)	24,000,000 (+125.00%, z=+1.15)	769,126,641.
L1/TEX Atomic ALU	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0
L1/TEX Atomic CAS	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0
L1/TEX Reduction	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0
L1/TEX Total	147,416,237 (+130.01%, z=+0.94)	238,064,896 (+166.03%, z=+1.16)	1.61 (+13.93%, z=+0.93)	16.91 (-1.96%, z=-0.19)	49.52 (-36.64%, z=-1.15)	7,618,076,672 (+166.03%, z=+1.16)	244,136,071,834
ECC Total	-	0 (+0.00%, z=+0.00)	-	0 (+0.00%, z=+0.00)	-	0 (+0.00%, z=+0.00)	0 (+0
GPU Total	148,634,618 (+130.41%, z=+0.94)	243,411,907 (+165.71%, z=+1.16)	1.64 (+12.51%,	17.29 (-2.75%, z=-0.27)	48.45 (-35.66%,	7,789,181,024 (+165.71%, z=+1.16)	249,619,443,316

L2 Cache Eviction Policies

	First	Hit Rate	Last	Hit Rate	Normal	Hit Rate	Normal D
L1/TEX Load	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	237,315,812 (+166.17%, z=+1.16)	49.36 (-36.75%, z=-1.15)	0 (+0.0
L1/TEX Store	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	750,000 (+125.00%, z=+1.15)	100 (+0.00%, z=+0.00)	0 (+0.0
L1/TEX Atomic	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	
L1/TEX Total	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	238,071,245 (+166.01%, z=+1.16)	49.52 (-36.62%, z=-1.15)	0 (+0.0
GPU Total	4,782,904 (+220.84%, z=+1.27)	0.90 (-88.60%,	0 (+0.00%, z=+0.00)	0 (+0.00%, z=+0.00)	238,066,602 (+166.02%, z=+1.16)	49.52 (-36.63%,	0 (+0.0

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	125,606,892 (+195.73%, z=+1.16)	29.63 (+117.10%, z=+1.06)	4,019,420,544 (+195.73%, z=+1.16)	128,810,142,626.84 (+117.04%, z=+1.05)
Store	786,860 (+133.38%, z=+1.19)	0.19 (-48.69%, z=-1.17)	25,179,520 (+133.38%, z=+1.19)	806,926,651.98 (-48.73%, z=-1.17)
Total	126,393,752 (+195.24%, z=+1.16)	29.81 (+112.82%, z=+1.05)	4,044,600,064 (+195.24%, z=+1.16)	129,617,069,278.83 (+112.76%, z=+1.05)

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]	11.94 (+1.54%, z=+1.15)	No Eligible [%]	64.96 (-5.62%, z=-0.88)
Eligible Warps Per Scheduler [warp]	1.48 (-1.17%, z=-0.20)	One or More Eligible [%]	35.04 (+12.42%, z=+0.88)
Issued Warp Per Scheduler	0.35 (+12.42%, z=+0.88)		

Issue Slot Utilization Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 2.9 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 11.94 active warps per scheduler, but only an average of 1.48 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

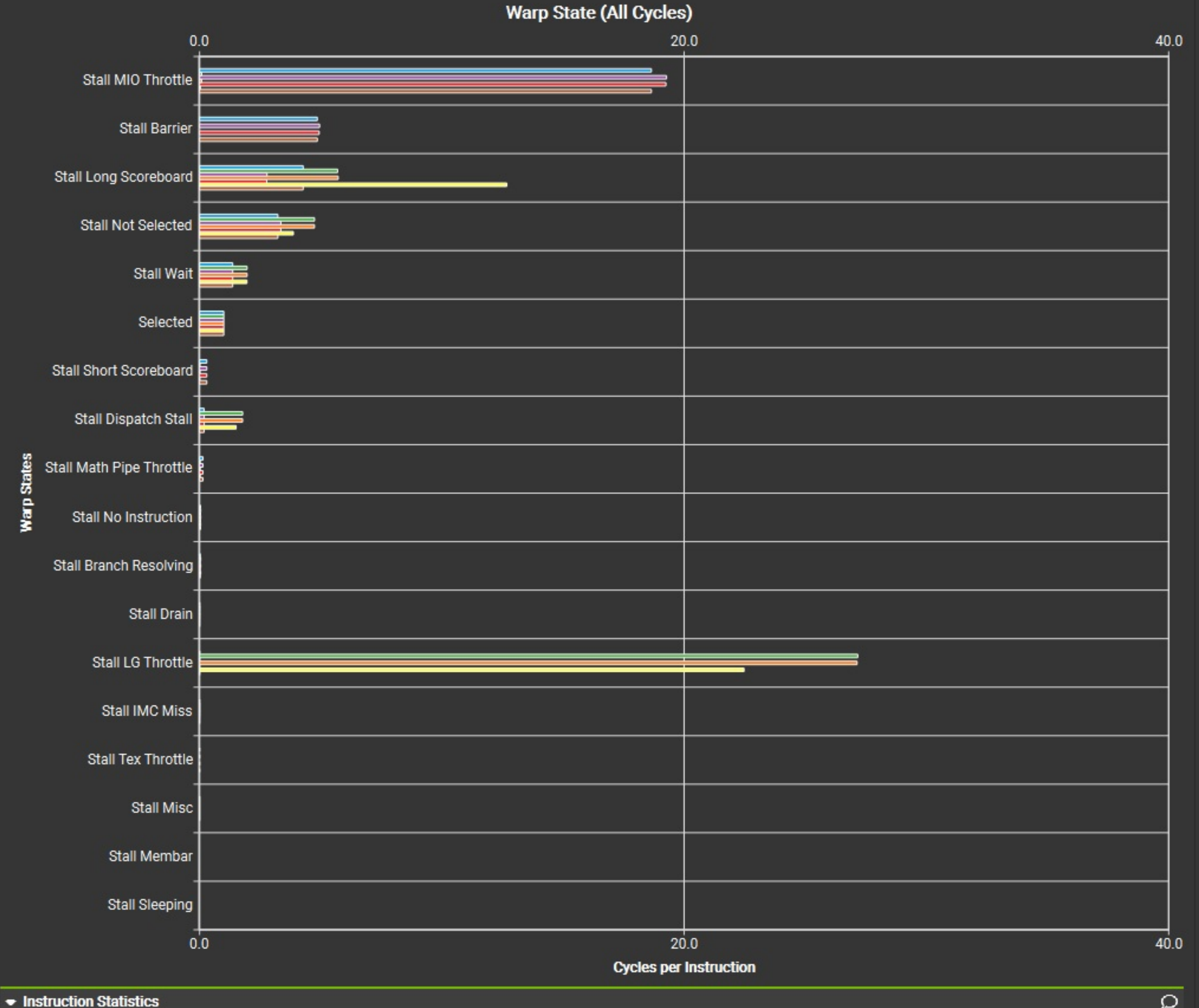
Analysis of the states in which all warps spent cycles during the kernel execution. The warp states describe a warp's readiness or inability to issue its next instruction. The warp cycles per instruction define the latency between two consecutive instructions. The higher the value, the more warp parallelism 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]	34.09 (-11.01%, z=-0.79)	Avg. Active Threads Per Warp	32.00 (+0.31%, z=+0.77)
Warp Cycles Per Executed Instruction [cycle]	34.09 (-11.01%, z=-0.79)	Avg. Not Predicated Off Threads Per Warp	31.91 (+0.32%, z=+1.00)

mio_throttle On average, each warp of this kernel spends 18.6 cycles being stalled waiting for an MIO instruction queue to be not full. This represents about 54.7% of the total average of 34.1 cycles between issuing two instructions. This stall reason is high in cases of 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.

Warp Stall Check the Source Counters section 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)

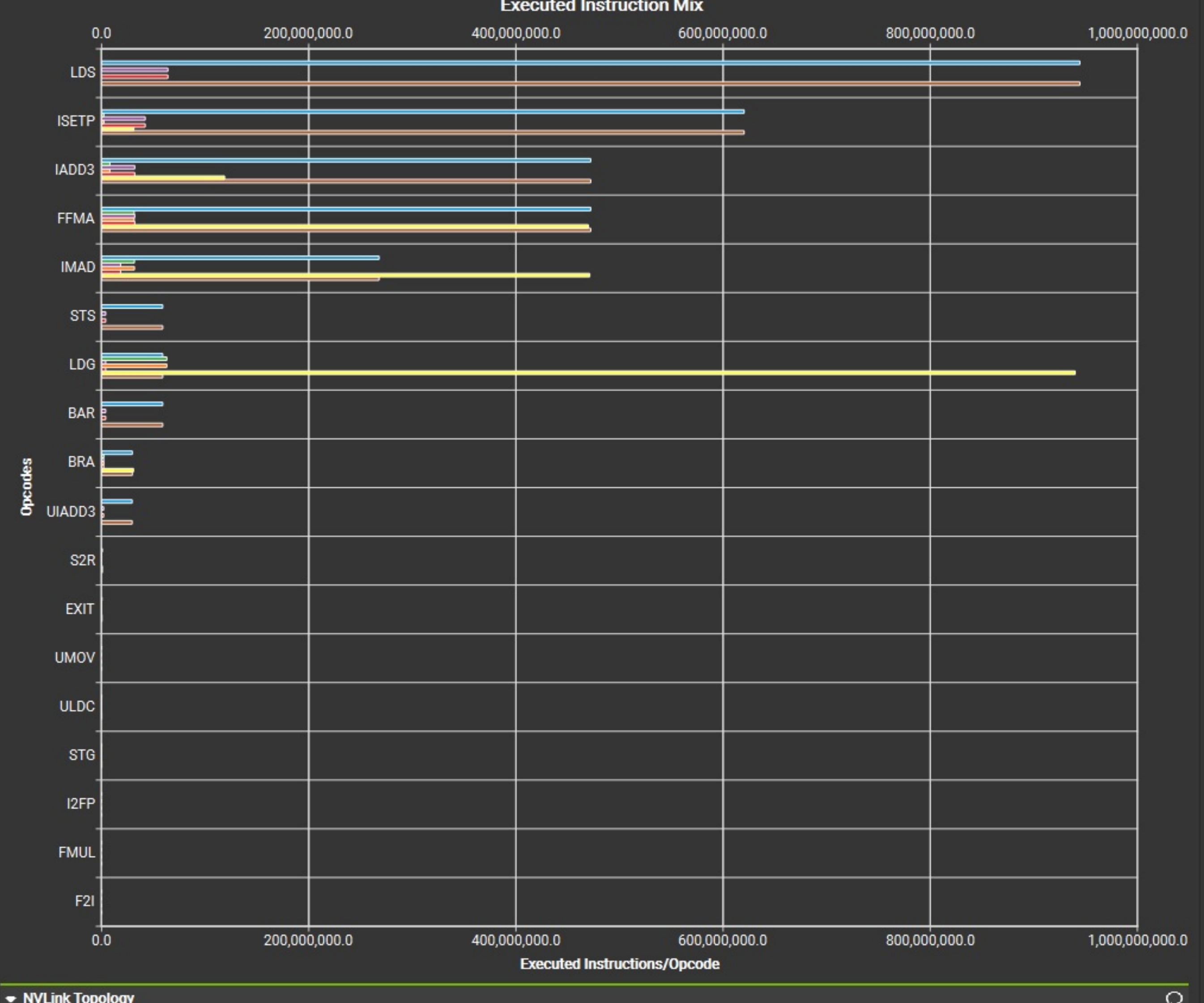


Instruction Statistics

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]	3,015,896,000 (+214.91%, z=+1.37)	Avg. Executed Instructions Per Scheduler [inst]	16,390,739.13 (+214.91%, z=+1.37)
Issued Instructions [inst]	3,015,916,076 (+214.91%, z=+1.37)	Avg. Issued Instructions Per Scheduler [inst]	16,390,848.24 (+214.91%, z=+1.37)

Executed Instruction Mix



NVLink Topology

NVLink Topology diagram shows logical NVLink connections with transmit/receive throughput.

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	23,500 (+124.25%, z=+1.15)	Function Cache Configuration	cudaFuncCachePreferNone (cudaFuncCachePreferNone)
Registers Per Thread [register/thread]	27 (-19.40%, z=-0.87)	Static Shared Memory Per Block [Kbyte/block]	2.05 (+100.00%, z=+0.87)
Block Size	256 (+0.00%, z=+0.00)	Dynamic Shared Memory Per Block [byte/block]	0 (+0.00%, z=+0.00)
Threads [thread]	6,016,000 (+124.25%, z=+1.15)	Driver Shared Memory Per Block [Kbyte/block]	1.02 (+0.00%, z=+0.00)
Waves Per SM	85.14 (+124.25%, z=+1.15)	Shared Memory Configuration Size [Kbyte]	32.77 (+60.00%, z=+0.87)

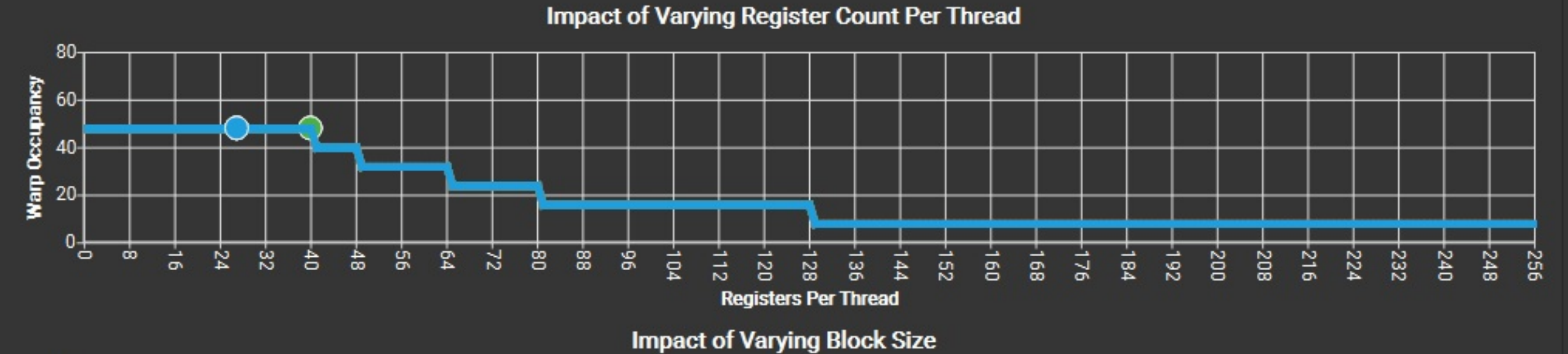
Occupancy

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.

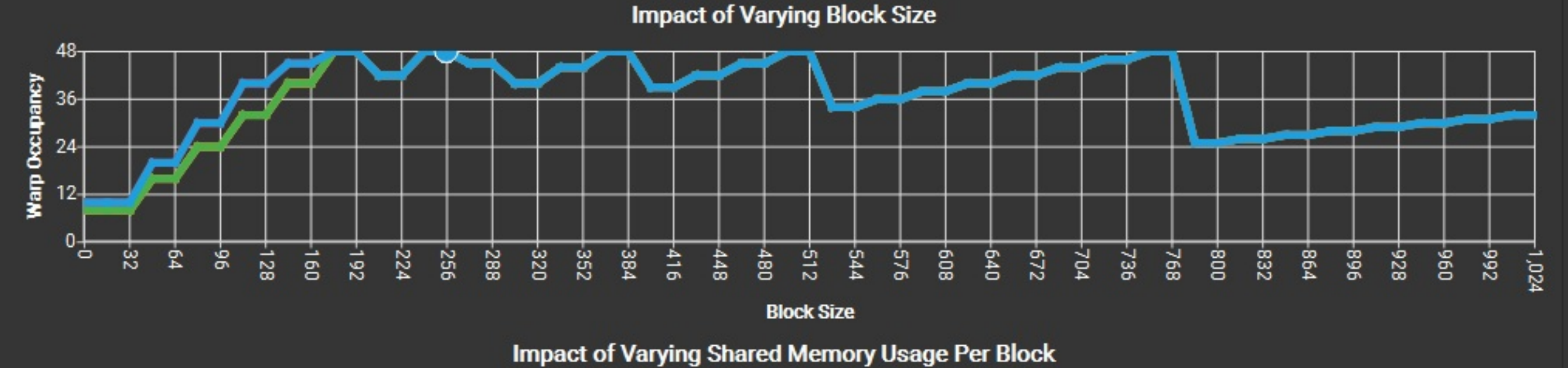
Theoretical Occupancy [%]	100 (+0.00%, z=+0.00)	Block Limit Registers [block]	8 (+14.29%, z=+0.87)
Theoretical Active Warps per SM [warp]	48 (+0.00%, z=+0.00)	Block Limit Shared Mem [block]	10 (+11.11%, z=+0.87)
Achieved Occupancy [%]	99.52 (+1.55%, z=+1.15)	Block Limit Warps [block]	6 (+0.00%, z=+0.00)
Achieved Active Warps Per SM [warp]	47.77 (+1.55%, z=+1.15)	Block Limit SM [block]	16 (+0.00%, z=+0.00)

Occupancy Limiters This kernel's theoretical occupancy is not impacted by any block limit.

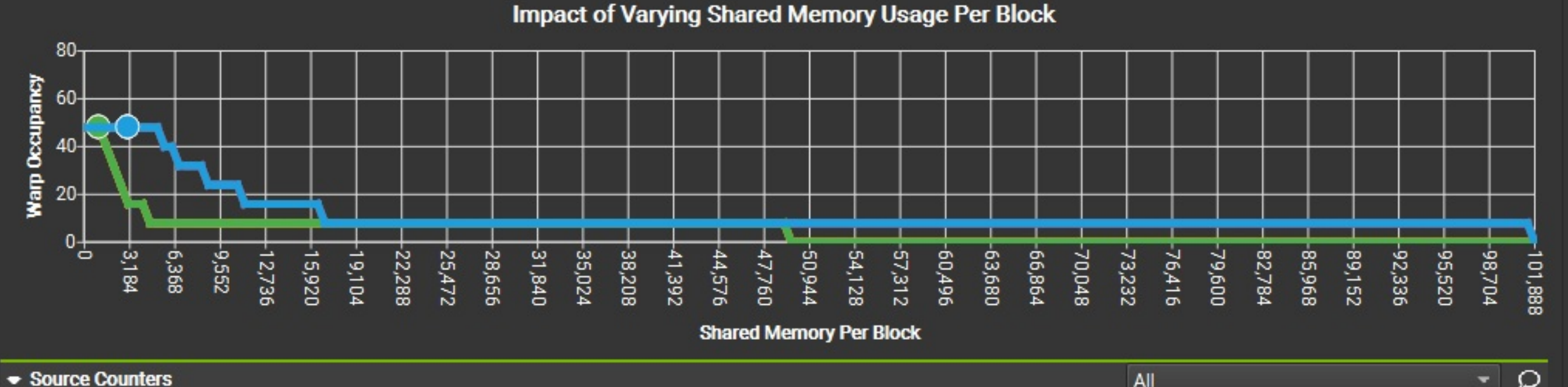
Impact of Varying Register Count Per Thread



Impact of Varying Block Size



Impact of Varying Shared Memory Usage Per Block



Source Counters

All

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]	30,080,000 (+159.45%, z=+1.13)	Branch Efficiency [%]	100 (+0.00%, z=+0.00)
Branch Instructions Ratio [%]	0.01 (-22.92%, z=-0.91)	Avg. Divergent Branches	0 (+0.00%, z=+0.00)

Uncoalesced Global Accesses This kernel has uncoalesced global accesses resulting in a total of 29516000 excessive sectors (11% of the total 264953500 sectors). Check the L2 Theoretical Sectors Global Excessive table for the primary source locations. The CUDA Programming Guide had additional information on reducing uncoalesced device memory accesses.

L2 Theoretical Sectors Global Excessive

Location	Value	Value (%)
0x700bb2310 in shared_mem_mmult	29,516,000	100

Warp Stall Sampling (All Cycles)

Location	Value	Value (%)
0x700bb23b0 in shared_mem_mmult	268,689	13
0x700bb23e0 in shared_mem_mmult	237,997	11
0x700bb23c0 in shared_mem_mmult	226,587	11
0x700bb28b0 in shared_mem_mmult	133,155	6
0x700bb2310 in shared_mem_mmult	91,190	4

Warp Stall Sampling (Not-issued Cycles)

Location	Value	Value (%)
0x700bb23b0 in shared_mem_mmult	176,249	13
0x700bb23c0 in shared_mem_mmult	147,522	11
0x700bb23e0 in shared_mem_mmult	146,925	11
0x700bb28b0 in shared_mem_mmult	82,540	6
0x700bb2310 in shared_mem_mmult	58,632	4

Most Instructions Executed

Location	Value	Value (%)
0x700bb28b0 in shared_mem_mmult	29,516,000	1
0x700bb28a0 in shared_mem_mmult	29,516,000	1
0x700bb2890 in shared_mem_mmult	29,516,000	1
0x700bb2880 in shared_mem_mmult	29,516,000	1
0x700bb2870 in shared_mem_mmult	29,516,000	1