Page: Details | Result: 0 - 535 - global_mem_mmult | Clear Baselines | Apply Rules | Occupancy Calculator | Save as PDF

	Report	Result	Time	Cycles	Regs	GPU	SM Frequency	CC	Process
Current	mmul...3000	535 - global_mem_mmult (375, 6...	27.91 msecond	41,855,681	40	0 - NVIDIA GeForce RTX 3070	1.50 cycle/nsecond	8.6	[1288] HW3.exe
t16x16_shared	mmul...3000	538 - shared_mem_mmult (188, 1...	31.21 msecond	46,806,961	27	0 - NVIDIA GeForce RTX 3070	1.50 cycle/nsecond	8.6	[17568] HW3.exe
t16x16_global	mmul...3000	535 - global_mem_mmult (188, 1...	27.41 msecond	41,108,780	40	0 - NVIDIA GeForce RTX 3070	1.50 cycle/nsecond	8.6	[17568] HW3.exe

t32x32_global | t32x32_shared | t8x8_shared | cta32x8_global | cta64x4_global | cta8x32_global

GPU Speed Of Light Throughput

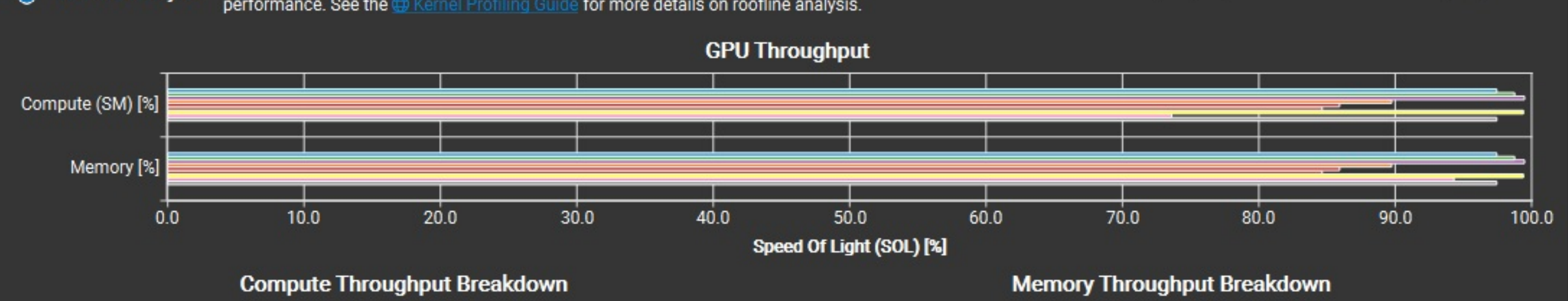
All

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 [%]	97.44 (+6.94%, z=+0.66)	Duration [msecond]	27.91 (-12.91%, z=-0.82)
Memory Throughput [%]	97.44 (+3.98%, z=+0.60)	Elapsed Cycles [cycle]	41,855,681 (-12.78%, z=-0.82)
L1/TEX Cache Throughput [%]	97.51 (+6.85%, z=+0.66)	SM Active Cycles [cycle]	41,828,855.57 (-12.70%, z=-0.82)
L2 Cache Throughput [%]	24.11 (+24.01%, z=+0.59)	SM Frequency [cycle/nsecond]	1.50 (+0.12%, z=+0.32)
DRAM Throughput [%]	17.38 (-55.81%, z=-0.76)	DRAM Frequency [cycle/nsecond]	6.79 (+0.12%, z=+0.31)

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 6% 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.

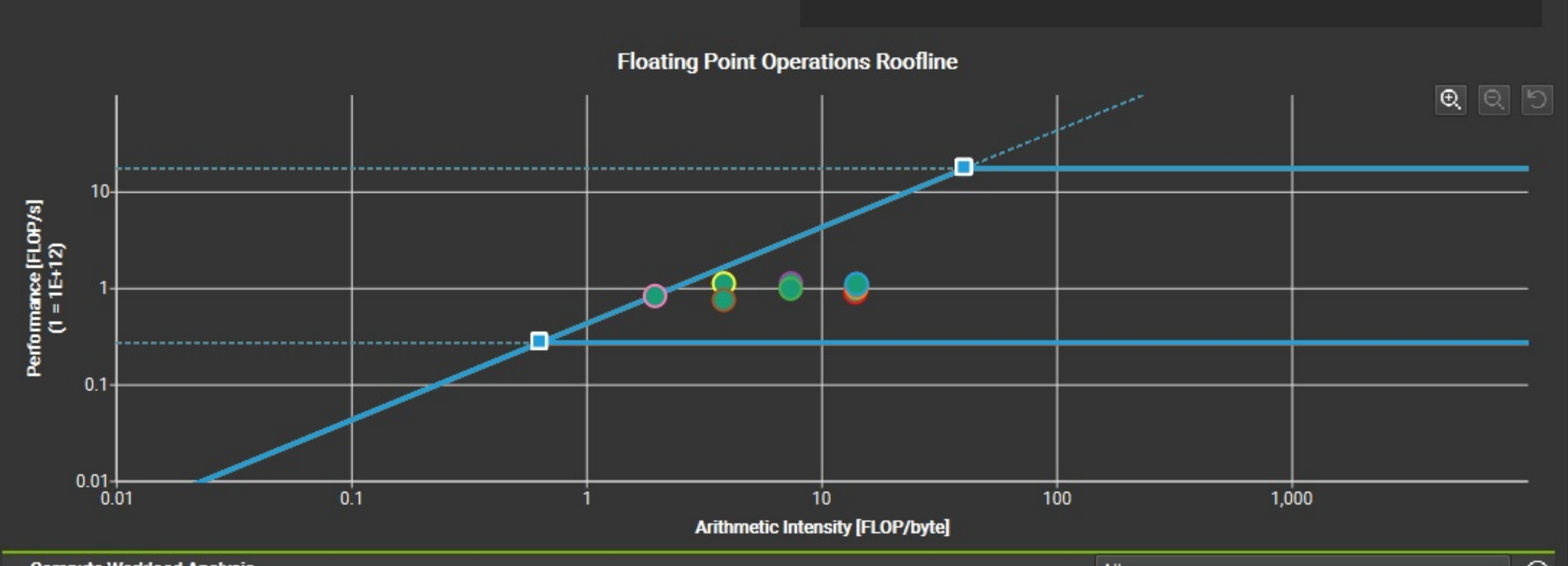


Compute Throughput Breakdown

SM: Inst Executed Pipe Lsu [%]	97.44	(+6.94%, z=+0.66)
SM: Mio Inst Issued [%]	32.49	(+4.58%, z=+0.45)
SM: Mio2rf Writeback Active [%]	30.70	(+28.27%, z=+1.27)
SM: Issue Active [%]	26.75	(-3.88%, z=-0.24)
SM: Inst Executed [%]	26.75	(-3.88%, z=-0.24)
SM: Pipe Fmaheavy Cycles Active [%]	25.54	(+43.71%, z=+0.81)
SM: Pipe Fma Cycles Active [%]	18.28	(+31.36%, z=+0.85)
SM: Pipe Alu Cycles Active [%]	3.94	(-64.99%, z=-0.68)
SM: Inst Executed Pipe Adu [%]	0.04	(-98.12%, z=-0.62)
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.01	(+18.12%, z=+0.95)
SM: Mio Pq Read Cycles Active [%]	0.01	(-99.04%, z=-0.62)
SM: Mio Pq Write Cycles Active [%]	0.01	(-99.06%, z=-0.62)
SM: Inst Executed Pipe Uniform [%]	0.00	(-97.82%, z=-0.52)
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: Inst Executed Pipe Xu [%]	0	(-100.00%, z=-0.70)
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 [%]	97.44	(+6.94%, z=+0.66)
L1: Data Pipe Lsu Wavefronts [%]	76.72	(+42.96%, z=+1.56)
L1: Lsu Writeback Active [%]	76.68	(+50.21%, z=+1.61)
L2: T Sectors [%]	24.11	(+24.01%, z=+0.59)
L2: Lts2xbar Cycles Active [%]	24.07	(+24.15%, z=+0.59)
L2: T Tag Requests [%]	23.85	(+73.77%, z=+1.29)
L2: Xbar2lts Cycles Active [%]	23.84	(+85.35%, z=+1.33)
DRAM: Cycles Active [%]	17.38	(-55.81%, z=-0.76)
L1: M L1tex2xbar Req Cycles Active [%]	15.54	(+98.17%, z=+1.36)
L1: M Xbar2l1tex Read Sectors [%]	15.50	(+24.56%, z=+0.59)
DRAM: Dram Sectors [%]	10.47	(-55.79%, z=-0.76)
L2: D Sectors [%]	7.16	(-6.35%, z=-0.13)
L2: D Sectors Fill Device [%]	5.14	(-56.43%, z=-0.76)
L1: Data Bank Reads [%]	4.57	(-47.05%, z=-1.23)
L1: Data Bank Writes [%]	1.94	(-6.24%, z=-0.13)
L2: D Sectors Fill Sysmem [%]	0.00	(-47.38%, z=+0.00)
L1: Texin Sm2tex Req Cycles Active [%]	0.00	(-99.99%, z=-0.42)
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)



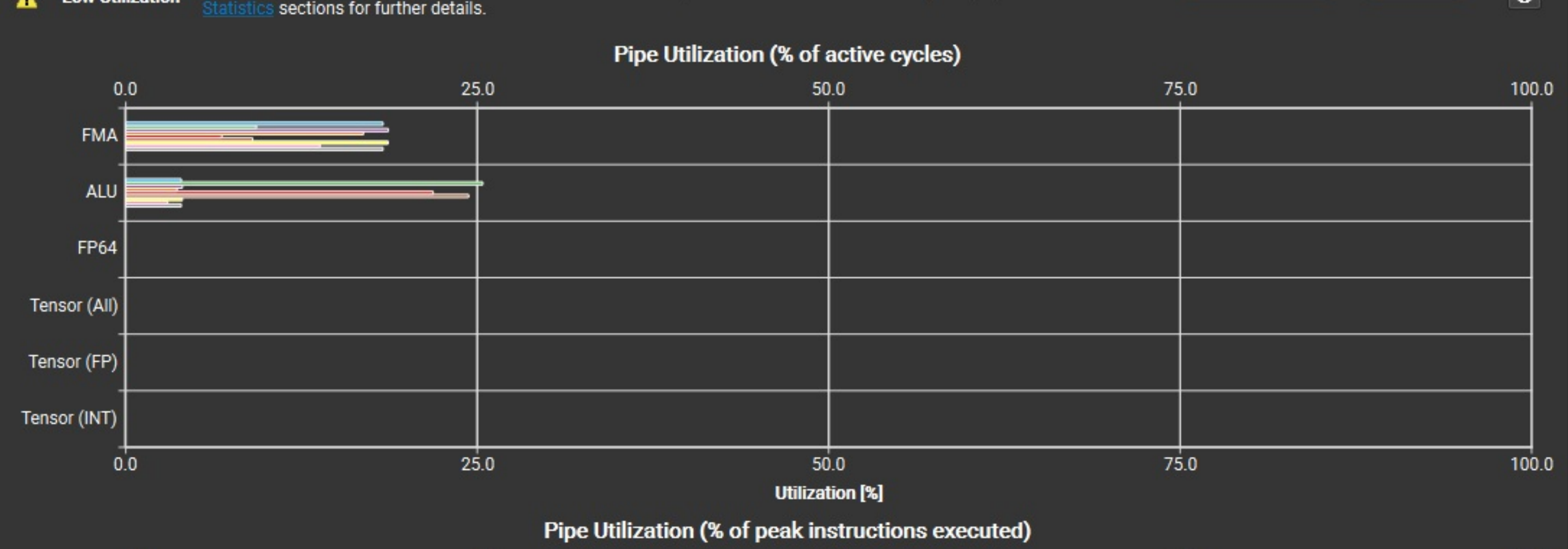
Compute Workload Analysis

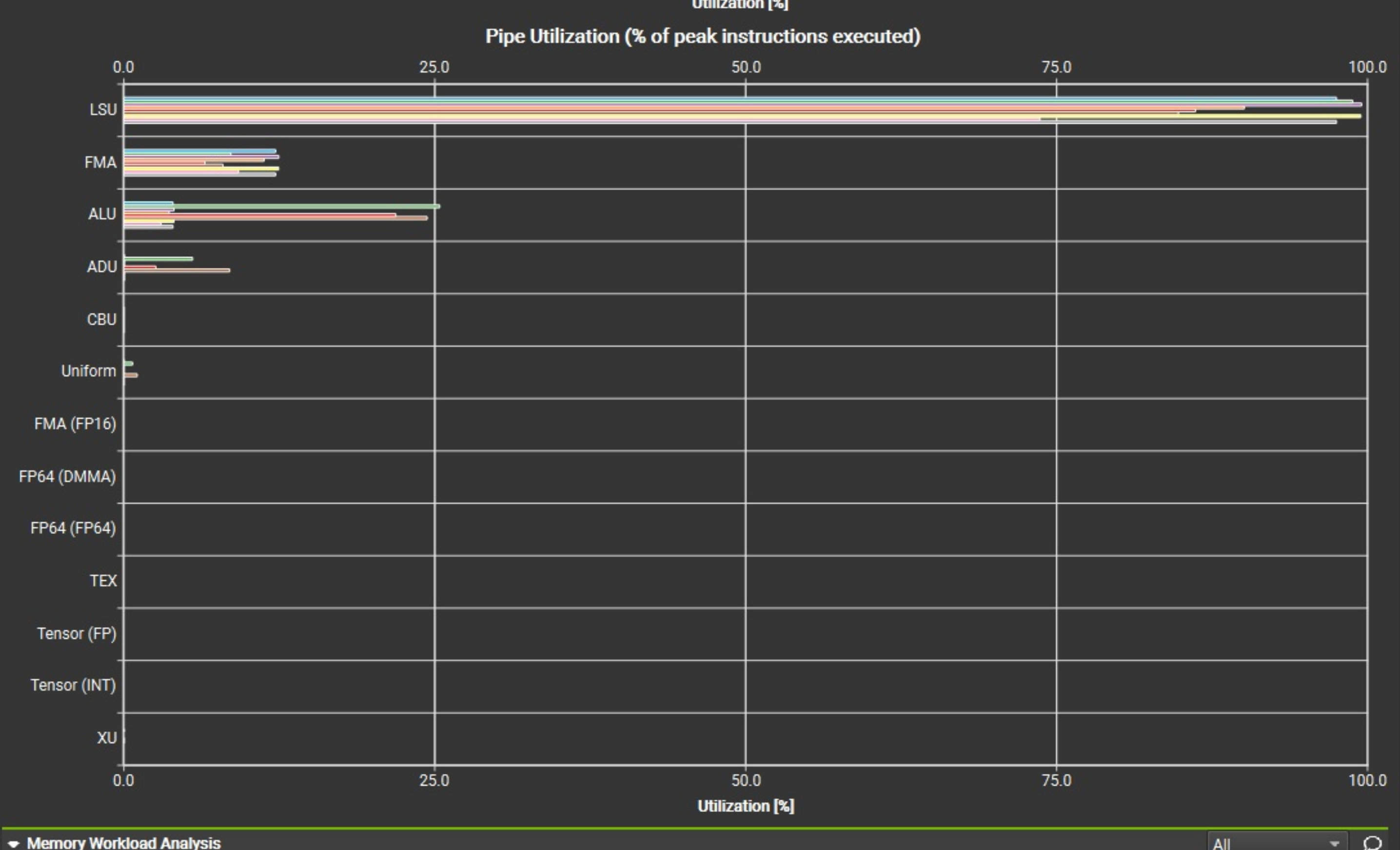
All

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]	1.07 (-3.88%, z=-0.24)	SM Busy [%]	32.51 (+3.64%, z=+0.35)
Executed Ipc Active [inst/cycle]	1.07 (-3.96%, z=-0.24)	Issue Slots Busy [%]	26.77 (-3.96%, z=-0.24)
Issued Ipc Active [inst/cycle]	1.07 (-3.96%, z=-0.24)		

Low Utilization All compute 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

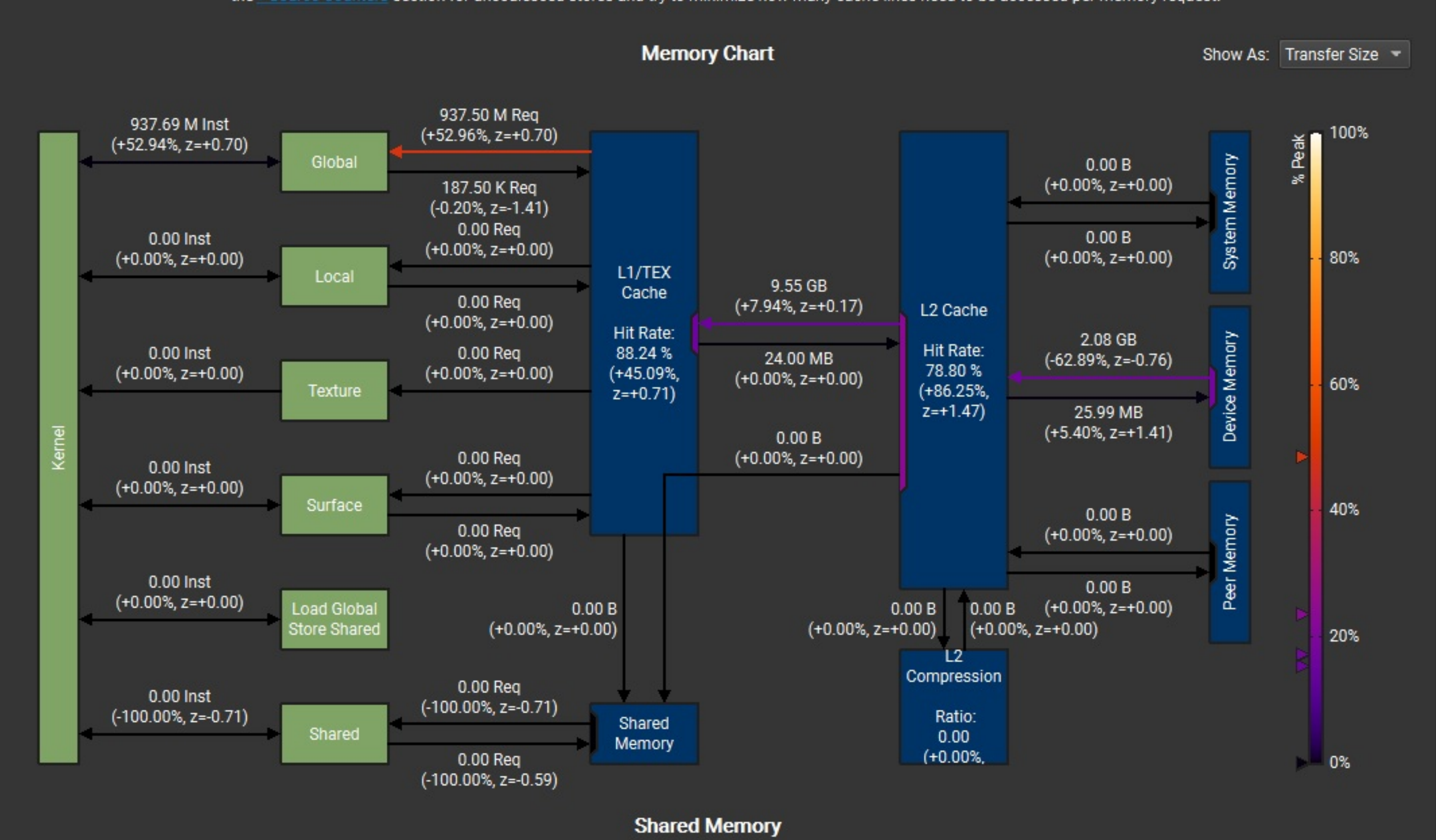
All

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]	75.58 (-55.75%, z=-0.76)	Mem Busy [%]	76.72 (+42.96%, z=+1.56)
L1/TEX Hit Rate [%]	88.24 (+45.09%, z=+0.71)	Max Bandwidth [%]	97.44 (+3.98%, z=+0.60)
L2 Hit Rate [%]	78.80 (+86.25%, z=+1.47)	Mem Pipes Busy [%]	97.44 (+6.94%, z=+0.66)
L2 Compression Success Rate [%]	0 (+0.00%, z=+0.00)	L2 Compression Ratio	0 (+0.00%, z=+0.00)

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 1.0 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 1.0 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.



Shared Memory

	Instructions		Requests		Wavefronts		% Peak		Bank Conflicts	
Shared Load	0	(-100.00%, z=-0.71)	0	(-100.00%, z=-0.71)	0	(-100.00%, z=-0.71)	0	(-100.00%, z=-0.70)	0	(-100.00%, z=-0.50)
Shared Load Matrix	0	(+0.00%, z=+0.00)	0	(+0.00%, z=+0.00)						
Shared Store	0	(-100.00%, z=-0.59)	0	(-100.00%, z=-0.59)	0	(-100.00%, z=-0.59)	0	(-100.00%, z=-0.62)	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		-		-	378,000	(-97.02%, z=-0.54)	0.02	(-98.45%, z=-0.61)	0	(-100.00%, z=-0.60)
Total	0	(-100.00%, z=-0.71)	0	(-100.00%, z=-0.71)	378,000	(-99.90%, z=-0.71)	0.02	(-99.88%, z=-0.70)	0	(-100.00%, z=-0.60)

L1/TEX Cache

	Instructions		Requests		Wavefronts		% Peak		Sectors		Sectors/Req		Hit Rate	
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)	0	(+0.00%, z=+0.00)
Global Load	937,500,000	(+52.96%, z=+0.70)	937,500,000	(+52.96%, z=+0.70)	1,477,408,507	(+103.68%, z=+1.27)	76.74	(+117.63%, z=+1.33)	2,317,143,804	(+52.89%, z=+0.76)	2.47	(-20.05%, z=-0.59)	88.27	(+45.09%, z=+0.71)
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)	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)	0	(+0.00%, z=+0.00)
Global Store	187,500	(-0.20%, z=-1.41)	187,500	(-0.20%, z=-1.41)	379,282	(+47.90%, z=+1.36)	0.02	(+67.18%, z=+1.67)	750,000	(+0.00%, z=+0.00)	4	(+0.20%, z=+1.41)	0.01	(-98.31%, z=-0.40)
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)	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)	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)	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)	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)	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)	0	(+0.00%, z=+0.00)
Surface Atomic CAS	0	(+0.00%, z=+0.00)												
Loads	937,500,000	(+52.96%, z=+0.70)	937,500,000	(+52.96%, z=+0.70)	1,477,408,507	(+103.68%, z=+1.27)	76.74	(+117.63%, z=+1.33)	2,317,143,804	(+52.89%, z=+0.76)	2.47	(-20.05%, z=-0.59)	88.27	(+45.09%, z=+0.71)
Stores	187,500	(-0.20%, z=-1.41)	187,500	(-0.20%, z=-1.41)	379,282	(+47.90%, z=+1.36)	0.02	(+67.18%, z=+1.67)	750,000	(+0.00%, z=+0.00)	4	(+0.20%, z=+1.41)	0.01	(-98.31%, z=-0.40)
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)	0	(+0.00%, z=+0.00)
Total	937,687,500	(+52.94%, z=+0.70)	937,687,500	(+52.94%, z=+0.70)	1,477,787,789	(+103.66%, z=+1.27)	76.75	(+117.62%,	2,317,893,804	(+52.86%, z=+0.76)	2.47	(-20.04%,	88.24	(+45.12%,

L2 Cache

	Requests		Sectors		Sectors/Req		% Peak		Hit Rate		Bytes		Throughput	
L1/TEX Load	298,483,929	(+51.38%, z=+0.86)	298,457,913	(+7.93%, z=+0.17)	1.00	(-35.73%, z=-1.07)	23.71	(+24.62%, z=+0.59)	79.84	(+85.20%, z=+1.46)	9,550,653,216	(+7.93%, z=+0.17)	342,230,179,670.38	(+24.73% z=+0.60
L1/TEX Store	750,000	(+59.91%, z=+1.35)	750,000	(+0.00%, z=+0.00)	1	(-44.18%, z=-1.25)	0.06	(+12.71%, z=+0.87)	100	(+0.00%, z=+0.00)	24,000,000	(+0.00%, z=+0.00)	859,996,078.42	(+12.66% z=+0.85
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	(+0.00%, z=+0.00
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	(+0.00%, z=+0.00
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	(+0.00%, z=+0.00
L1/TEX Total	299,221,670	(+51.28%, z=+0.85)	299,254,585	(+7.95%, z=+0.17)	1.00	(-35.72%, z=-1.07)	23.77	(+24.64%, z=+0.59)	79.90	(+84.44%, z=+1.45)	9,576,146,720	(+7.95%, z=+0.17)	343,143,692,731.43	(+24.74% z=+0.60
ECC 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
GPU Total	300,247,534	(+50.88%, z=+0.85)	303,525,376	(+7.41%, z=+0.16)	1.01	(-35.93%,	24.11	(+24.01%,	78.61	(+85.90%,	9,712,812,032	(+7.41%, z=+0.16)	348,040,844,080.42	(+24.11% z=+0.59

L2 Cache Eviction Policies

	First		Hit Rate		Last		Hit Rate		Normal		Hit Rate		Normal Demote	
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)	298,465,661	(+7.98%, z=+0.17)	79.86	(+85.32%, z=+1.46)	0	(+0.00%, z=+0.00)
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	(+0.00%, z=+0.00)	100	(+0.00%, z=+0.00)	0	(+0.00%, z=+0.00)
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)	299,291,601	(+7.98%, z=+0.17)	79.90	(+84.58%, z=+1.46)	0	(+0.00%, z=+0.00)
GPU Total	4,363,312	(-19.69%, z=-0.94)	0.18	(-11.52%,	0	(+0.00%, z=+0.00)	0	(+0.00%, z=+0.00)	299,221,597	(+7.97%, z=+0.17)	79.91	(+84.64%,	0	(+0.00%, z=+0.00)

Device Memory

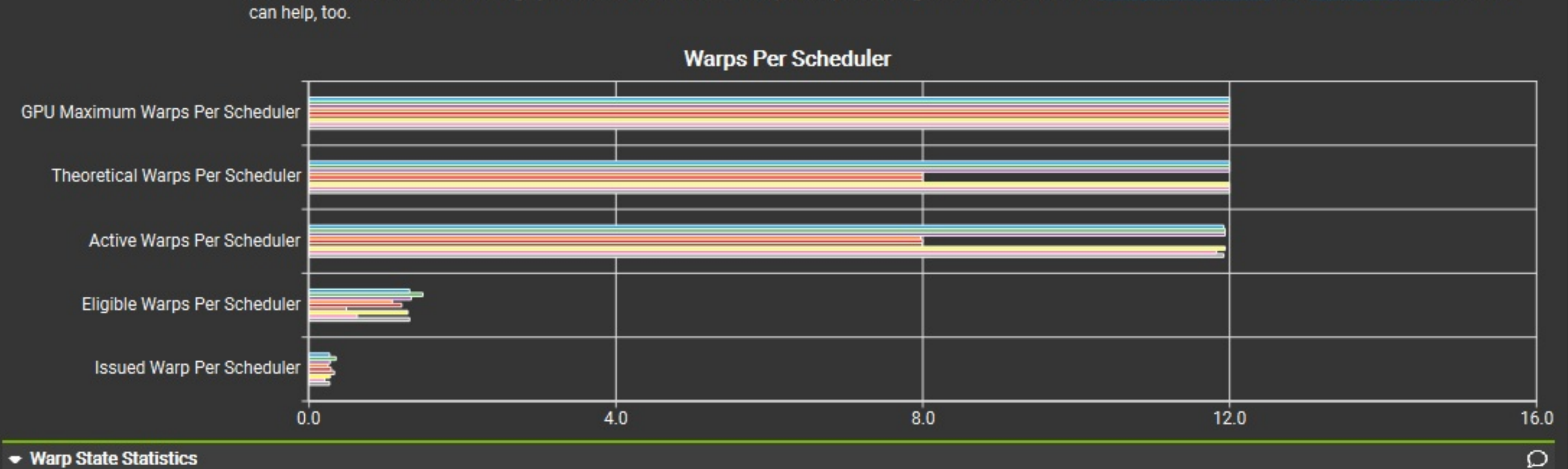
	Sectors		% Peak		Bytes		Throughput	
Load	65,098,848	(-62.89%, z=-0.76)	17.17	(-56.16%, z=-0.76)	2,083,163,136	(-62.89%, z=-0.76)	74,646,338,652.70	(-56.09%, z=-0.76)
Store	812,112	(+5.40%, z=+1.41)	0.21	(+18.59%, z=+1.16)	25,987,584	(+5.40%, z=+1.41)	931,217,513.65	(+18.69%, z=+1.16)
Total	65,910,960	(-62.59%, z=-0.76)	17.38	(-55.81%, z=-0.76)	2,109,150,720	(-62.59%, z=-0.76)	75,577,556,166.34	(-55.75%, z=-0.76)

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]	11.92 (+14.18%, z=+0.71)	No Eligible [%]	73.23 (+1.53%, z=+0.24)
Eligible Warps Per Scheduler [warp]	1.31 (+18.85%, z=+0.57)	One or More Eligible [%]	26.77 (-3.96%, z=-0.24)
Issued Warp Per Scheduler	0.27 (-3.96%, z=-0.24)		

Issue Slot Utilization Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 3.7 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 11.92 active warps per scheduler, but only an average of 1.31 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.



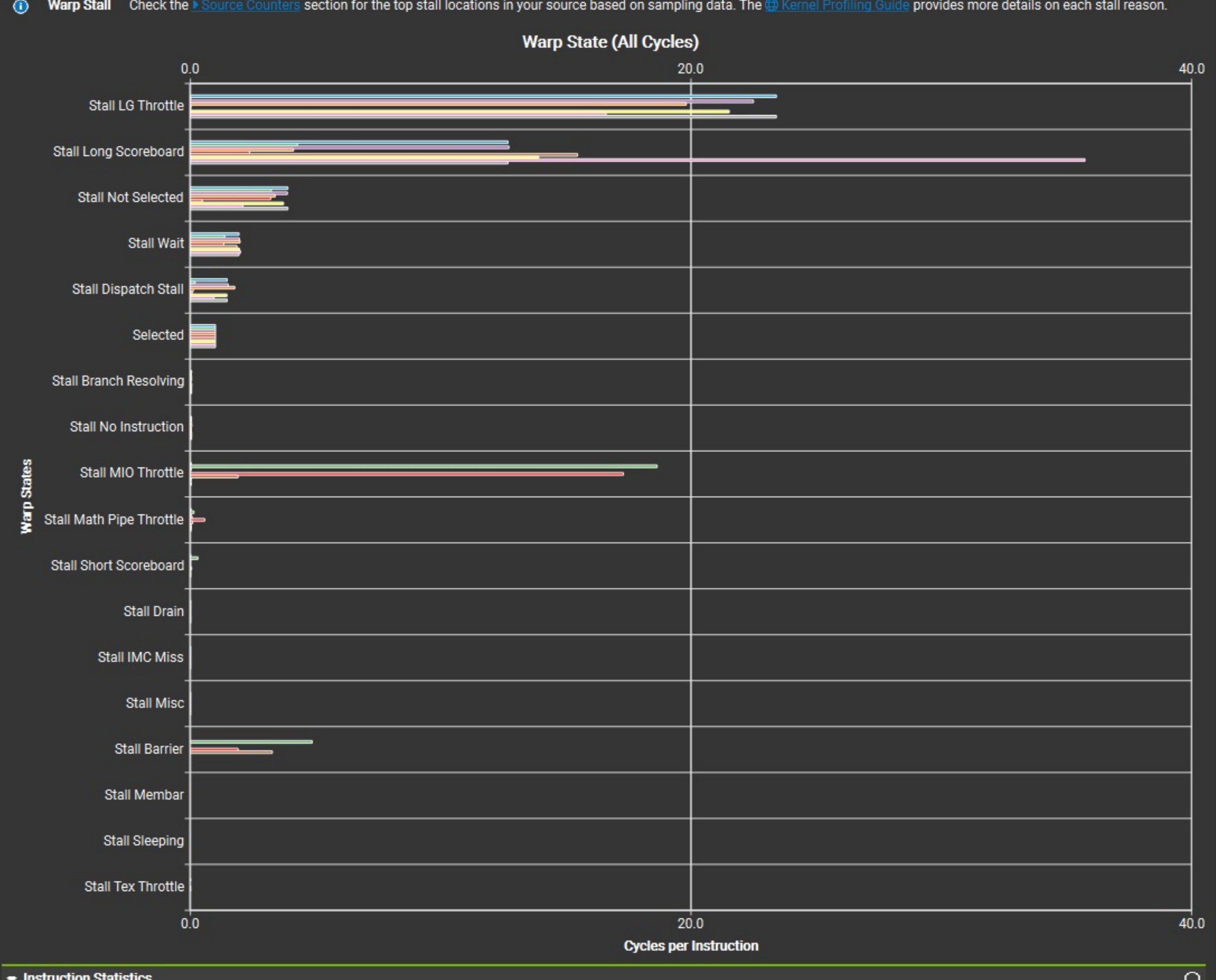
Warp State Statistics

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]	44.54 (+15.33%, z=+0.53)	Avg. Active Threads Per Warp	32 (+0.13%, z=+0.89)
Warp Cycles Per Executed Instruction [cycle]	44.54 (+15.33%, z=+0.53)	Avg. Not Predicated Off Threads Per Warp	31.98 (+0.24%, z=+1.17)

lg_throttle On average, each warp of this kernel spends 23.4 cycles being stalled waiting for the local/global instruction queue to be not full. This represents about 52.6% of the total average of 44.5 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.

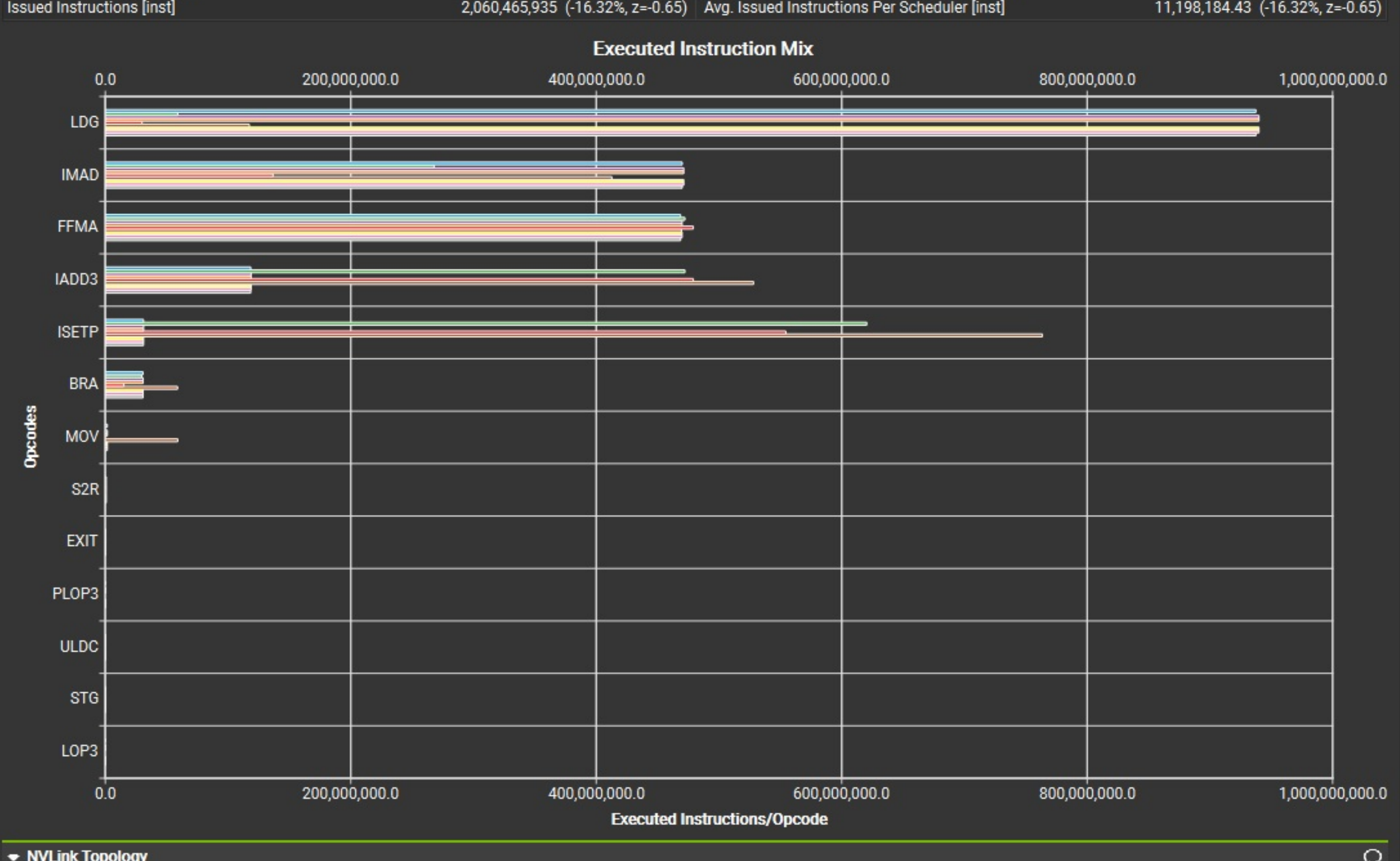
Warp Stall Check the Source Counters section for the top stall locations in your source based on sampling data. The Kernel Profiling Guide provides more details on each stall reason.



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]	2,060,452,500 (-16.32%, z=-0.65)	Avg. Executed Instructions Per Scheduler [inst]	11,198,111.41 (-16.32%, z=-0.65)
Issued Instructions [inst]	2,060,465,935 (-16.32%, z=-0.65)	Avg. Issued Instructions Per Scheduler [inst]	11,198,184.43 (-16.32%, z=-0.65)



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.

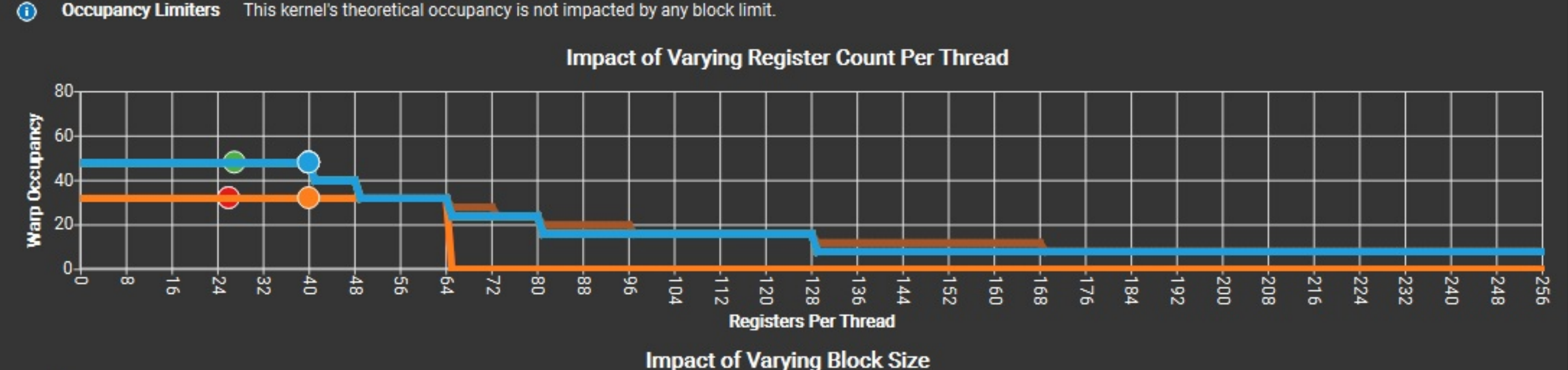
Grid Size	23,625 (-15.33%, z=-0.16)	Function Cache Configuration	cudaFuncCachePreferNone (cudaFuncCachePreferNone)
Registers Per Thread [register/thread]	40 (+14.70%, z=+0.71)	Static Shared Memory Per Block [byte/block]	0 (-100.00%, z=-0.47)
Block Size	256 (-39.62%, z=-0.44)	Dynamic Shared Memory Per Block [byte/block]	0 (+0.00%, z=+0.00)
Threads [thread]	6,048,000 (+0.30%, z=+0.70)	Driver Shared Memory Per Block [Kbyte/block]	1.02 (+0.00%, z=+0.00)
Waves Per SM	85.60 (-15.57%, z=-0.69)	Shared Memory Configuration Size [Kbyte]	8.19 (-46.67%, z=-0.63)

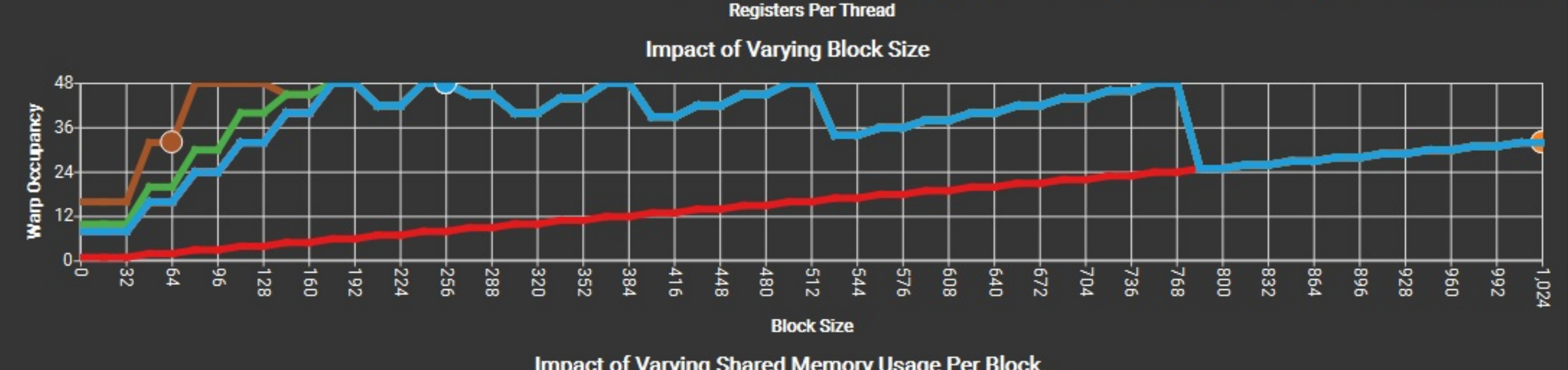
Occupancy

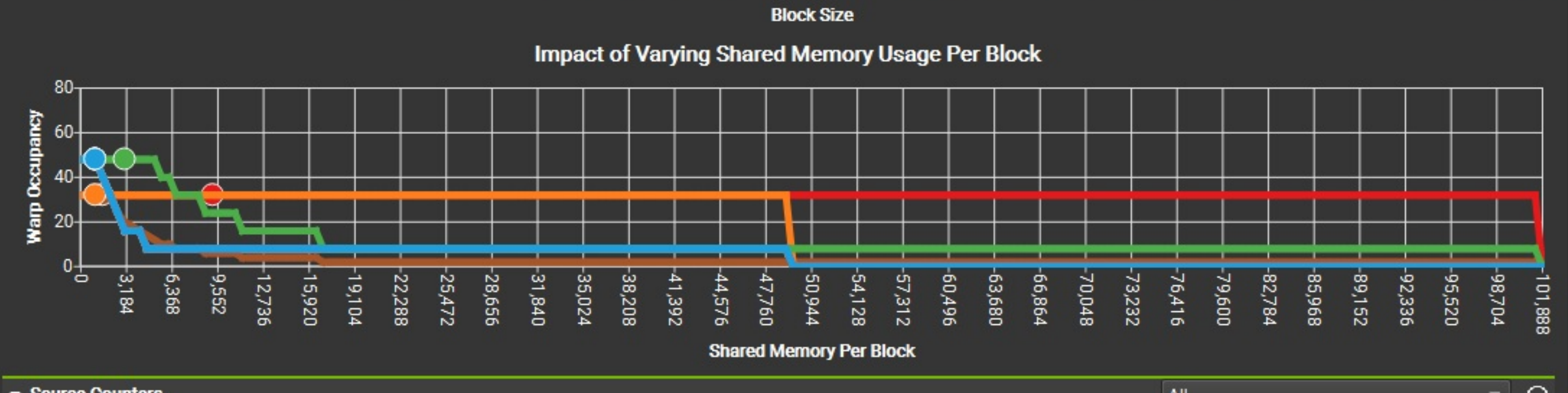
Occupancy is the ratio of the number of active warps per multiprocessor to the maximum number of possible active warps. Another way to view occupancy is the percentage of the hardware's ability to process warps that is actively in use. Higher occupancy does not always result in higher performance, however, 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 (+14.29%, z=+0.71)	Block Limit Registers [block]	6 (-28.36%, z=-0.24)
Theoretical Active Warps per SM [warp]	48 (+14.29%, z=+0.71)	Block Limit Shared Mem [block]	8 (-11.11%, z=-0.18)
Achieved Occupancy [%]	99.37 (+14.13%, z=+0.71)	Block Limit Warps [block]	6 (-14.29%, z=-0.14)
Achieved Active Warps Per SM [warp]	47.70 (+14.13%, z=+0.71)	Block Limit SM [block]	16 (+0.00%, z=+0.00)

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







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]	31,126,500 (-4.53%, z=-0.12)	Branch Efficiency [%]	100 (+0.00%, z=+0.00)
Branch Instructions Ratio [%]	0.02 (+12.43%, z=+0.45)	Avg. Divergent Branches	0 (+0.00%, z=+0.00)

Warp Stall Sampling (All Cycles)

Location	Value	Value (%)
0x700bb3920 in global_mem_mmult	176,201	9
0x700bb3540 in global_mem_mmult	175,459	9
0x700bb37c0 in global_mem_mmult	149,584	8
0x700bb3930 in global_mem_mmult	111,643	6
0x700bb37e0 in global_mem_mmult	78,182	4

Warp Stall Sampling (Not-issued Cycles)

Location	Value	Value (%)
0x700bb3540 in global_mem_mmult	126,848	10
0x700bb3920 in global_mem_mmult	120,539	9
0x700bb37c0 in global_mem_mmult	105,844	8
0x700bb3930 in global_mem_mmult	73,068	6
0x700bb37e0 in global_mem_mmult	57,092	4

Most Instructions Executed

Location	Value	Value (%)
0x700bb3970 in global_mem_mmult	29,250,000	1
0x700bb3960 in global_mem_mmult	29,250,000	1
0x700bb3950 in global_mem_mmult	29,250,000	1
0x700bb3940 in global_mem_mmult	29,250,000	1
0x700bb3930 in global_mem_mmult	29,250,000	1