Page: Details | Launch: 0 - 120 - gpu_insert_key | Add Baseline | Apply Rules | Save as PDF

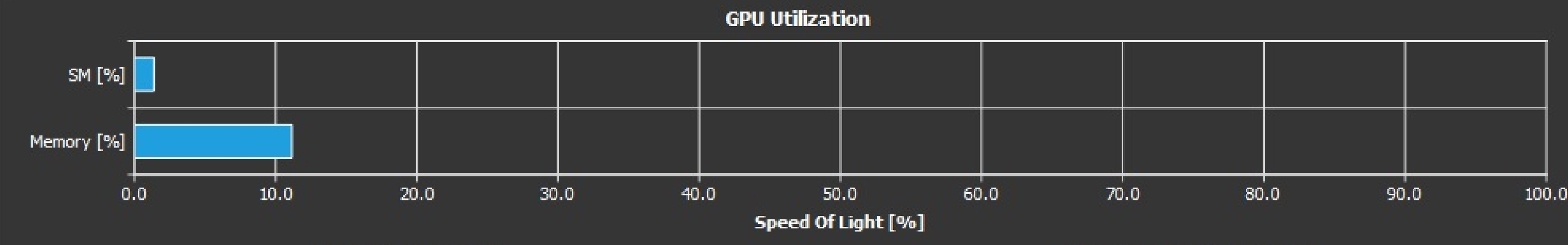
Current 120 - ... **Time:** 2.75 usecond **Cycles:** 5,361 **Regs:** 28 **GPU:** GeForce GTX 1660 Ti **SM Frequency:** 1.94 cycle/nsecond **CC:** 7.5 **Process:** [5972] memcached.exe

GPU Speed Of Light

High-level overview of the utilization for compute and memory resources of the GPU. For each unit, the Speed Of Light (SOL) reports the achieved percentage of utilization with respect to the theoretical maximum. High-level overview of the utilization for compute and memory resources of the GPU presented as a roofline chart.

SOL SM [%]	1.40	Duration [usecond]	2.75
SOL Memory [%]	11.15	Elapsed Cycles [cycle]	5,361
SOL L1/TEX Cache [%]	25.67	SM Active Cycles [cycle]	439.58
SOL L2 Cache [%]	4.61	SM Frequency [cycle/nsecond]	1.94
SOL DRAM [%]	11.15	DRAM Frequency [cycle/nsecond]	7.32

GPU Utilization



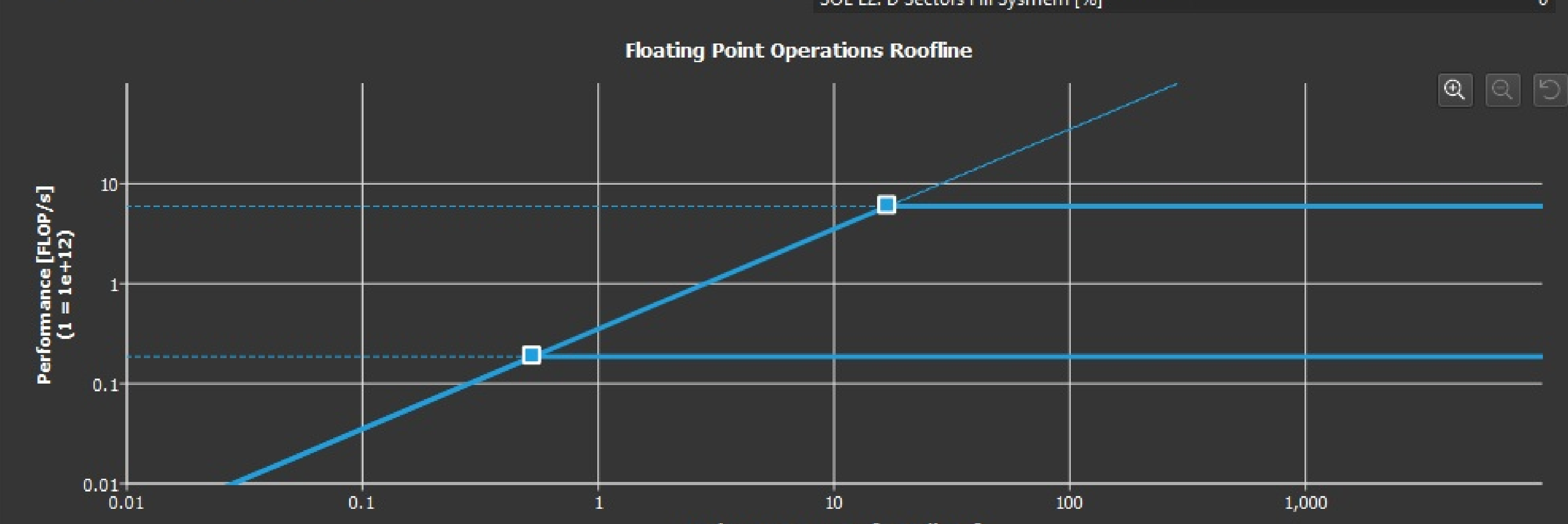
SOL SM Breakdown

SOL SM: Inst Executed Pipe Lsu [%]	1.40
SOL SM: Issue Active [%]	1.22
SOL SM: Inst Executed [%]	1.17
SOL SM: Pipe Alu Cycles Active [%]	1.00
SOL SM: Inst Executed Pipe Cbu Pred On Any [%]	0.60
SOL SM: Mio Inst Issued [%]	0.60
SOL SM: Pipe Fma Cycles Active [%]	0.45
SOL SM: Mio Pq Read Cycles Active [%]	0.42
SOL SM: Inst Executed Pipe Adu [%]	0.40
SOL SM: Mio2rf Writeback Active [%]	0.30
SOL SM: Inst Executed Pipe Xu [%]	0.20
SOL SM: Mio Pq Write Cycles Active [%]	0.20
SOL IDC: Request Cycles Active [%]	0
SOL SM: Inst Executed Pipe Fp16 [%]	0
SOL SM: Inst Executed Pipe Ipa [%]	0
SOL SM: Inst Executed Pipe Tex [%]	0
SOL SM: Inst Executed Pipe Uniform [%]	0
SOL SM: Pipe Fp64 Cycles Active [%]	0
SOL SM: Pipe Shared Cycles Active [%]	0
SOL SM: Pipe Tensor Cycles Active [%]	0

SOL Memory Breakdown

SOL DRAM: Cycles Active [%]	11.15
SOL DRAM: Dram Sectors [%]	5.52
SOL L2: Xbar2lts Cycles Active [%]	4.61
SOL L2: T Sectors [%]	3.23
SOL L2: T Tag Requests [%]	2.86
SOL L2: D Sectors [%]	2.26
SOL L2: D Sectors Fill Device [%]	2.19
SOL L1: M L1tex2xbar Req Cycles Active [%]	2.11
SOL L2: Lts2xbar Cycles Active [%]	1.97
SOL L1: Data Pipe Lsu Wavefronts [%]	1.42
SOL L1: Lsuin Requests [%]	1.40
SOL L1: Lsu Writeback Active [%]	0.61
SOL L1: M Xbar2l1tex Read Sectors [%]	0.55
SOL L1: Data Bank Reads [%]	0.08
SOL L1: Data Bank Writes [%]	0.08
SOL L1: Texin Sm2tex Req Cycles Active [%]	0.06
SOL L1: F Wavefronts [%]	0.06
SOL L1: Data Pipe Tex Wavefronts [%]	0
SOL L1: Tex Writeback Active [%]	0
SOL L2: D Atomic Input Cycles Active [%]	0
SOL L2: D Sectors Fill Sysmem [%]	0

Floating Point Operations Roofline



Recommendations

Bottleneck [Warning] This kernel grid is too small to fill the available resources on this device, resulting in only 0.2 full waves across all SMs. Look at Launch Statistics for more details.

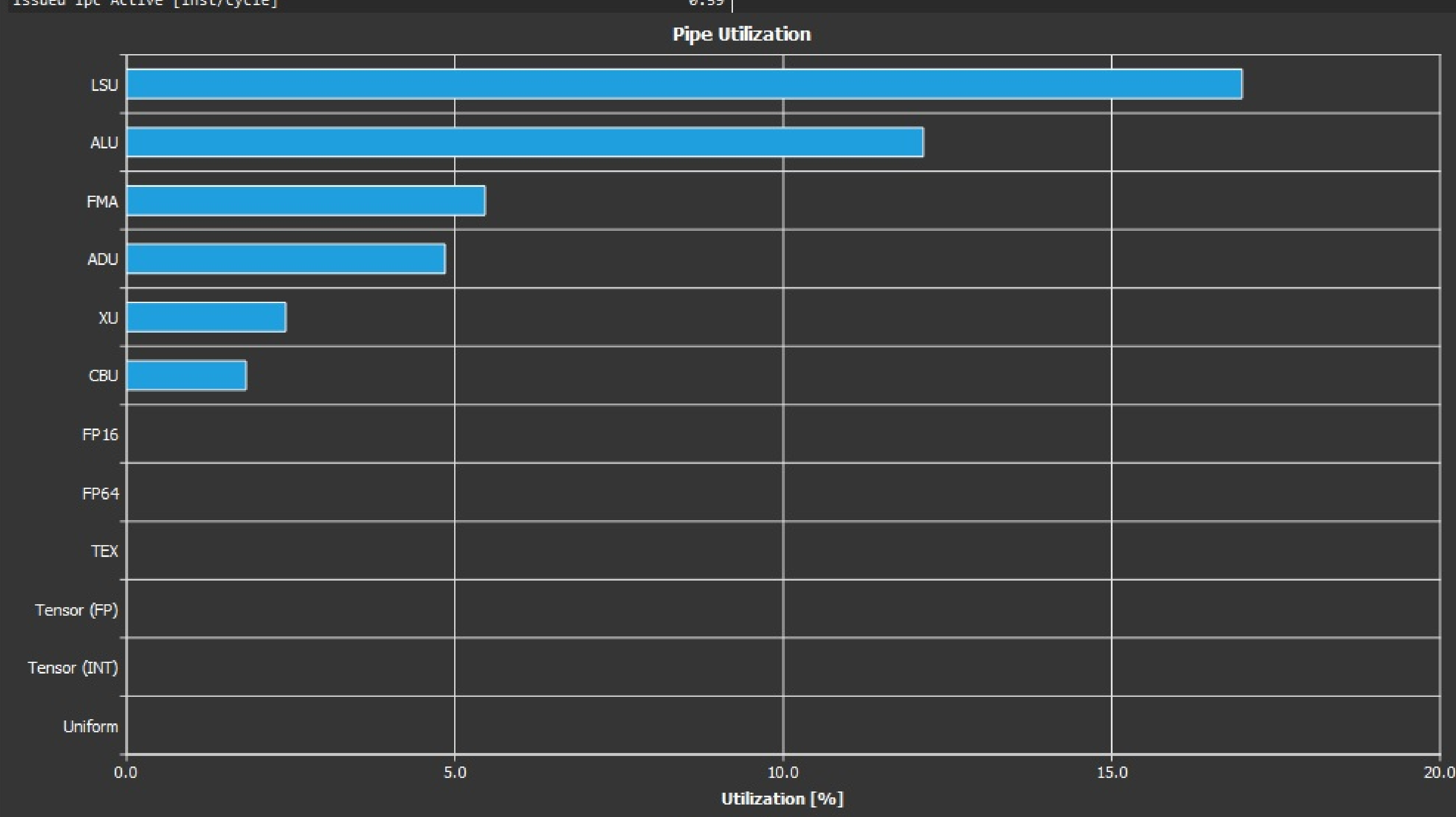
Roofline Analysis The ratio of peak float (fp32) to double (fp64) performance on this device is 32:1. The kernel achieved 0% of this device's fp32 peak performance and 0% of its fp64 peak performance.

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]	0.05	SM Busy [%]	14.86
Executed Ipc Active [inst/cycle]	0.57	Issue Slots Busy [%]	14.86
Issued Ipc Active [inst/cycle]	0.59		

Pipe Utilization



Recommendations

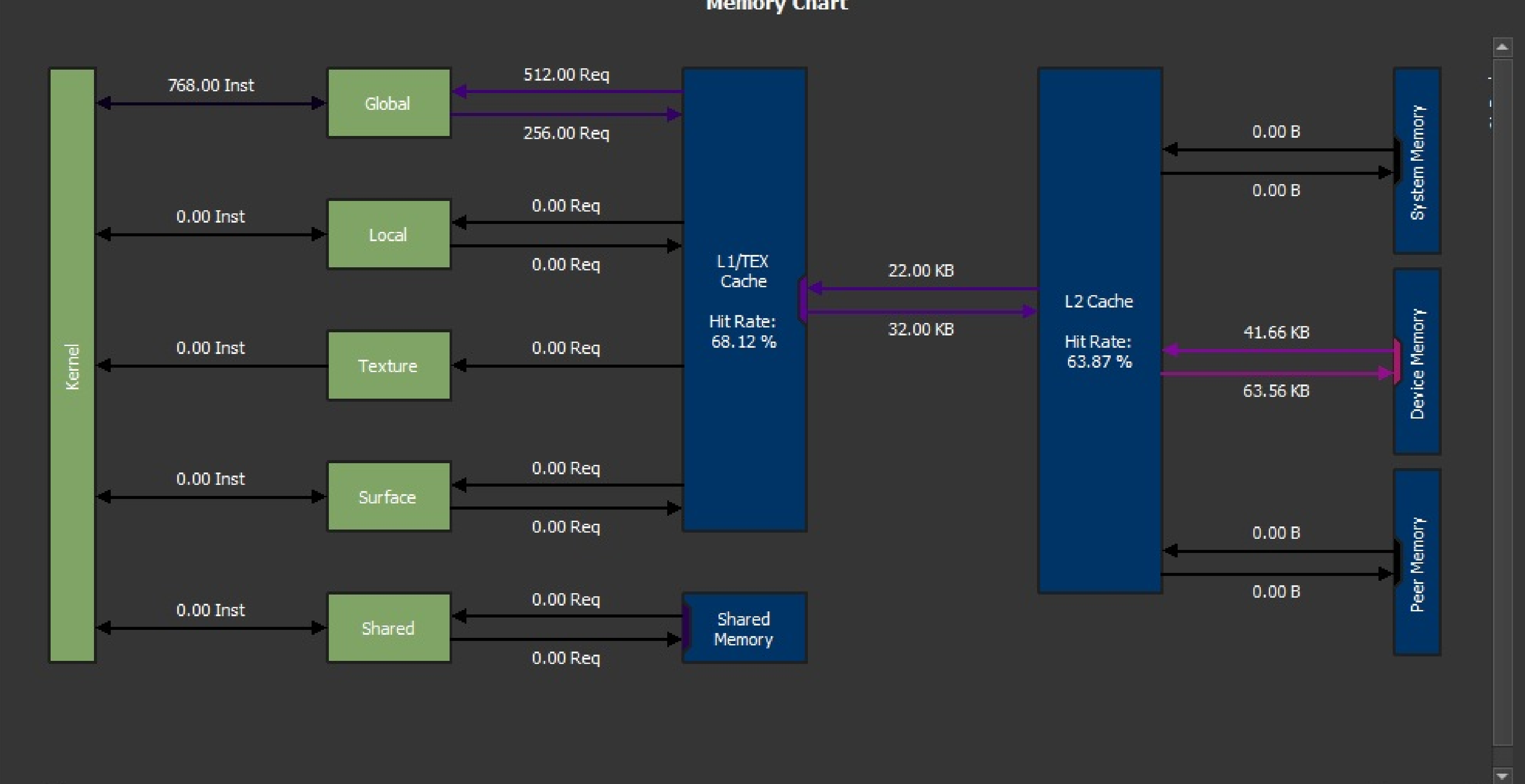
High Pipe Utilization [Warning] All pipelines are under-utilized. Either this kernel is very small or it doesn't issue enough warps per scheduler. Check the Launch Statistics and Scheduler Statistics sections for further details.

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]	39.15	Mem Busy [%]	3.23
L1/TEX Hit Rate [%]	68.12	Max Bandwidth [%]	11.15
L2 Hit Rate [%]	63.87	Mem Pipes Busy [%]	1.40

Memory Chart



Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	0	0	0	0	0
Shared Store	0	0	0	0	0
Shared Atomic	0	0	0	0	0
Other	-	-	132	0.10	0
Total	0	0	132	0.10	0

L1/TEX Cache

	Instructions	Requests	Wavefronts	% Peak	Sectors	Sectors/Req	Hit Rate
Local Load	0	0	0	0	0	0	
Global Load	512	512	659	0.51	1,280	2.50	44
Surface Load	0	0	0	0	0	0	
Texture Load	0	0	0	0	0	0	
Global Store	256	256	538	0.42	1,024	4	97
Local Store	0	0	0	0	0	0	
Surface Store	0	0	0	0	0	0	
Global Reduction	0	0	0	0	0	0	
Surface Reduction	0	0	0	0	0	0	
Global Atomic ALU	0	0	0	0	0	0	
Global Atomic CAS	0						
Surface Atomic ALU	0	0	0	0	0	0	
Surface Atomic CAS	0						
Loads	512	512	659	0.51	1,280	2.50	44
Stores	256	256	538	0.42	1,024	4	97

L2 Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes	Throu
L1/TEX Load	577	708	1.23	1.16	0	22,656	
L1/TEX Store	1,020	1,024	1.00	1.68	100	32,768	
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	1,576	1,728	1.10	2.83	45.71	55,296	

Device Memory

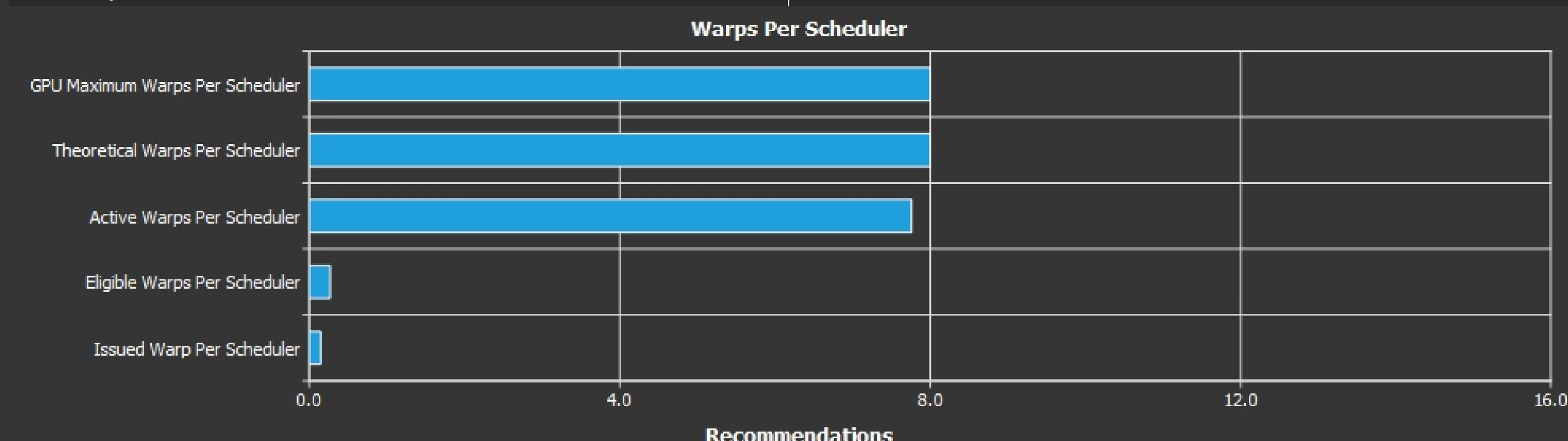
	Sectors	% Peak	Bytes	Throughput
Load	1,333	4.41	42,656	15,500,000,000
Store	2,034	6.73	65,088	23,651,162,790.70
Total	3,367	11.15	107,744	39,151,162,790.70

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]	7.76	No Eligible [%]	85.17
Eligible Warps Per Scheduler [warp]	0.27	One or More Eligible [%]	14.83
Issued Warp Per Scheduler	0.15		

Warps Per Scheduler



Recommendations

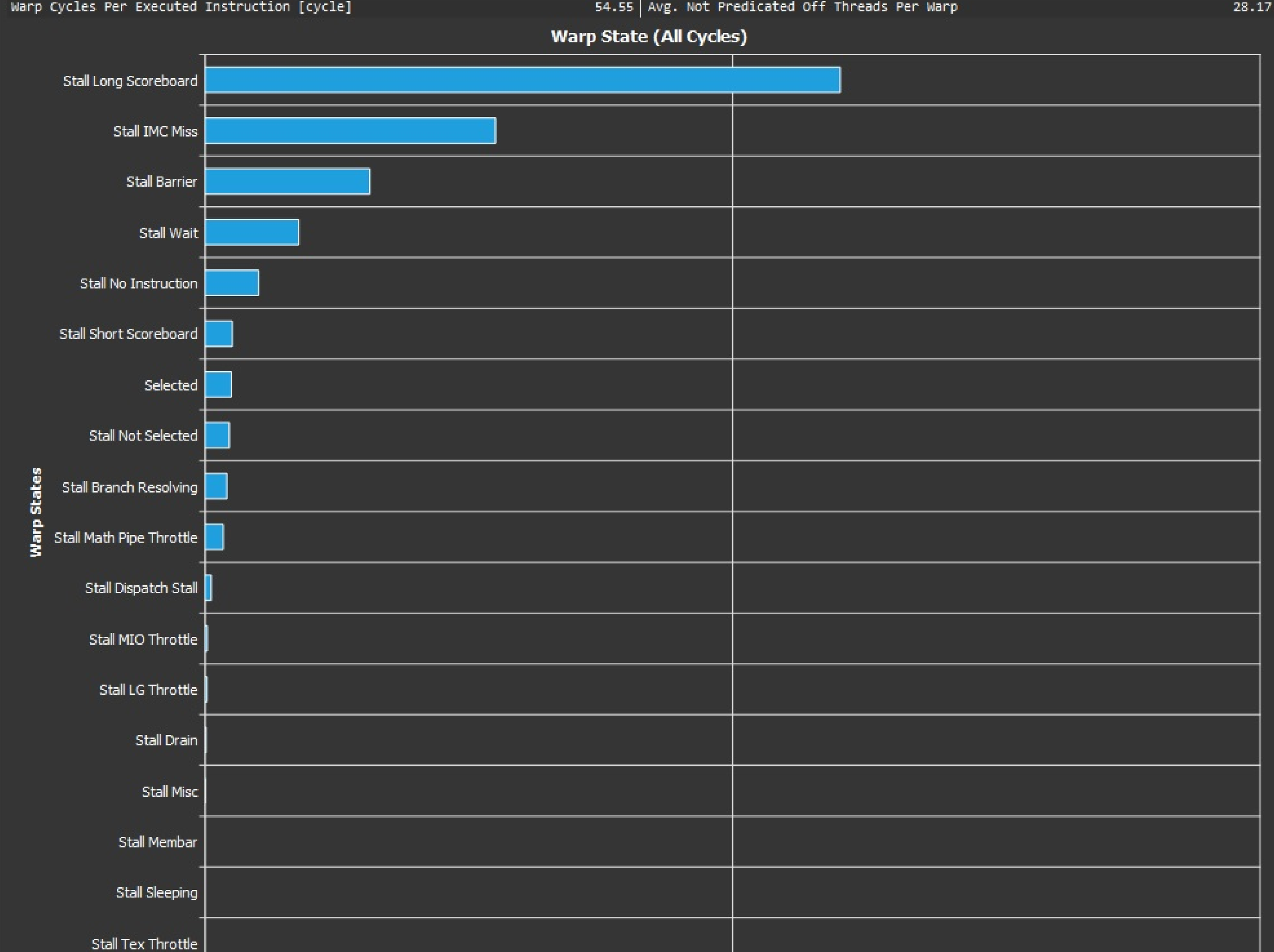
Issue Slot Utilization [Warning] Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 6.7 cycles. This might leave hardware resources underutilized and may lead to less optimal performance. Out of the maximum of 8 warps per scheduler, this kernel allocates an average of 7.76 active warps per scheduler, but only an average of 0.27 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 either increase the number of active warps or reduce the time the active warps are stalled.

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]	52.33	Avg. Active Threads Per Warp	28.94
Warp Cycles Per Executed Instruction [cycle]	54.55	Avg. Not Predicated Off Threads Per Warp	28.17

Warp State (All Cycles)



Recommendations

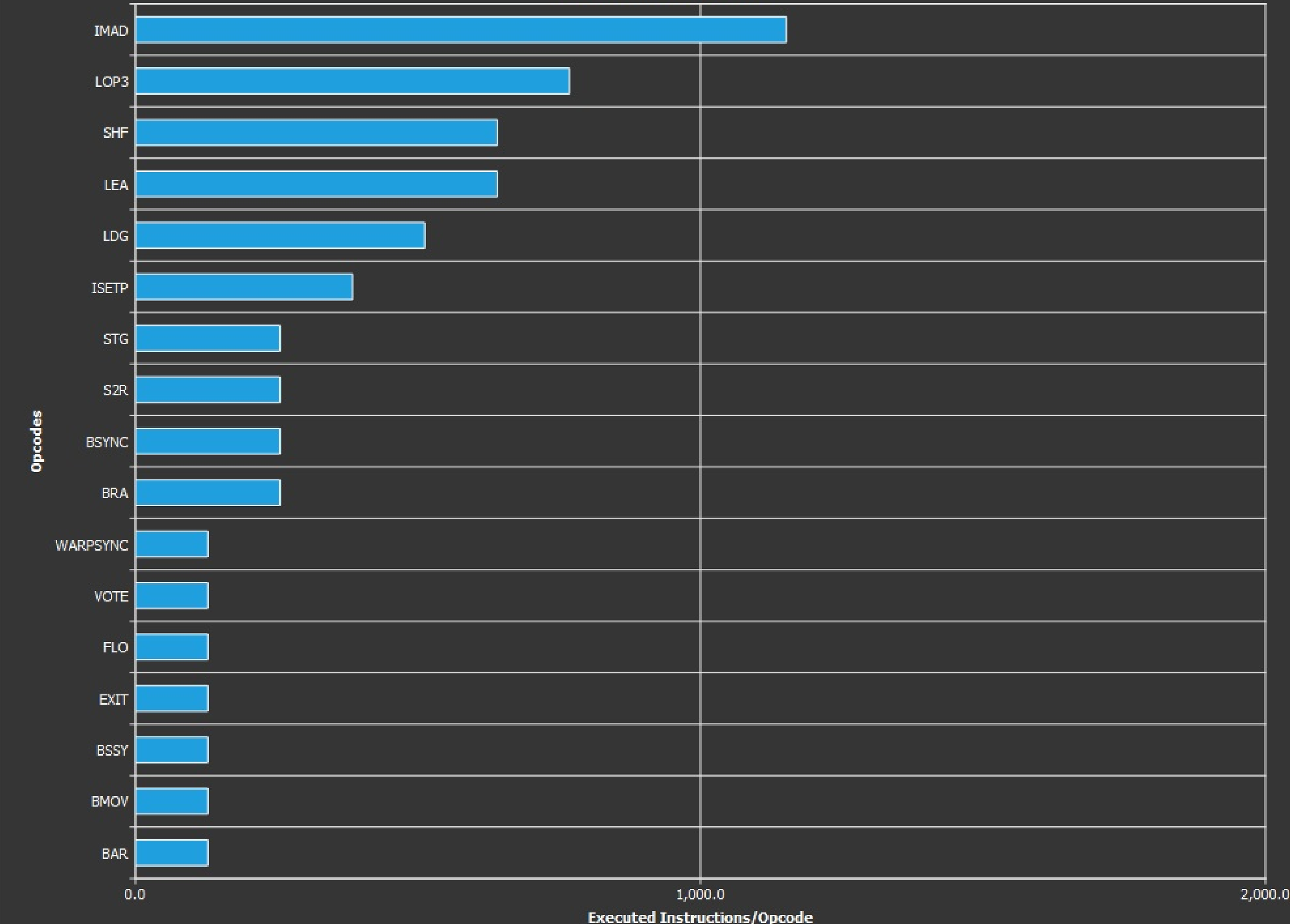
CPI Stall 'Long Scoreboard' [Warning] On average each warp of this kernel spends 24.1 cycles being stalled waiting for a scoreboard dependency on a L1TEX (local, global, surface, texture) operation. This represents about 46.0% of the total average of 52.3 cycles between issuing two instructions. To reduce the number of cycles waiting on L1TEX data accesses verify the memory access patterns are optimal for the target architecture, attempt to increase cache hit rates by increasing data locality or by changing the cache configuration, and consider moving frequently used data to shared memory.

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]	6,016	Avg. Executed Instructions Per Scheduler [inst]	62.67
Issued Instructions [inst]	6,272	Avg. Issued Instructions Per Scheduler [inst]	65.33

Executed Instruction Mix



NVLink

High-level summary of NVLink connection status.

Physical Links	0	Logical Links	0

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	4	Registers Per Thread [register/thread]	28
Block Size	1,024	Static Shared Memory Per Block [byte/block]	0
Threads [thread]	4,096	Dynamic Shared Memory Per Block [byte/block]	0
Waves Per SM	0.17	Driver Shared Memory Per Block [byte/block]	0
Function Cache Configuration	cudaFuncCachePreferNone	Shared Memory Configuration Size [Kbyte]	32.77

Recommendations

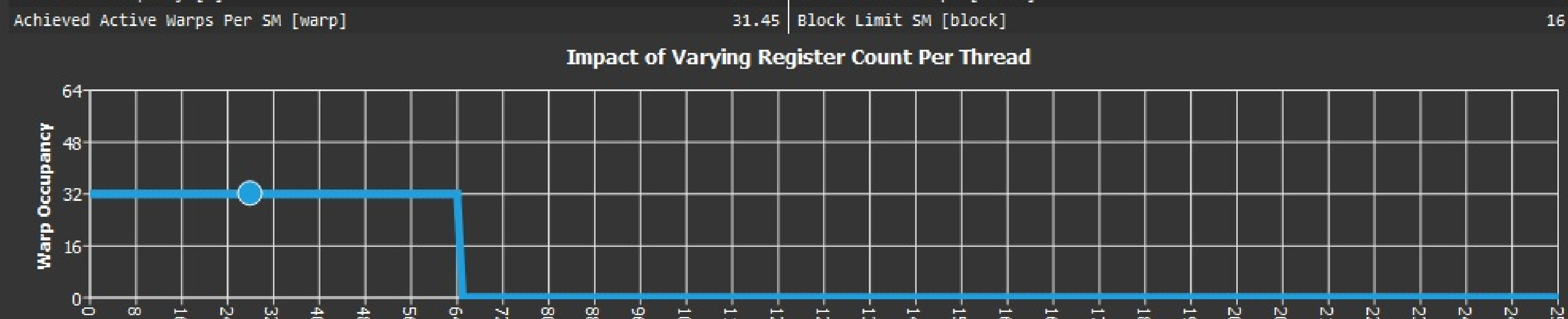
Launch Configuration [Warning] The grid for this launch is configured to execute only 4 blocks, which is less than the GPU's 24 multiprocessors. This can underutilize some multiprocessors. If you do not intend to execute this kernel concurrently with other workloads, consider reducing the block size to have at least one block per multiprocessor or increase the size of the grid to fully utilize the available hardware resources.

Occupancy

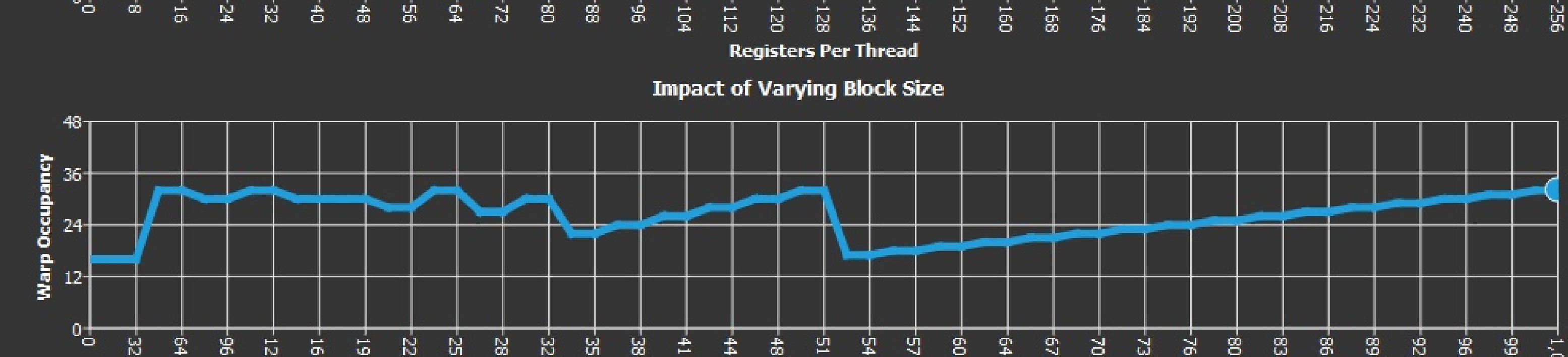
Occupancy is the ratio of the number of active warps per multiprocessor to the maximum number of possible active warps. Another way to view occupancy is the percentage of the hardware's ability to process warps that is actively in use. Higher occupancy does not always result in higher performance, however, 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]	2
Theoretical Active Warps per SM [warp]	32	Block Limit Shared Mem [block]	16
Achieved Occupancy [%]	98.27	Block Limit Warps [block]	1
Achieved Active Warps Per SM [warp]	31.45	Block Limit SM [block]	16

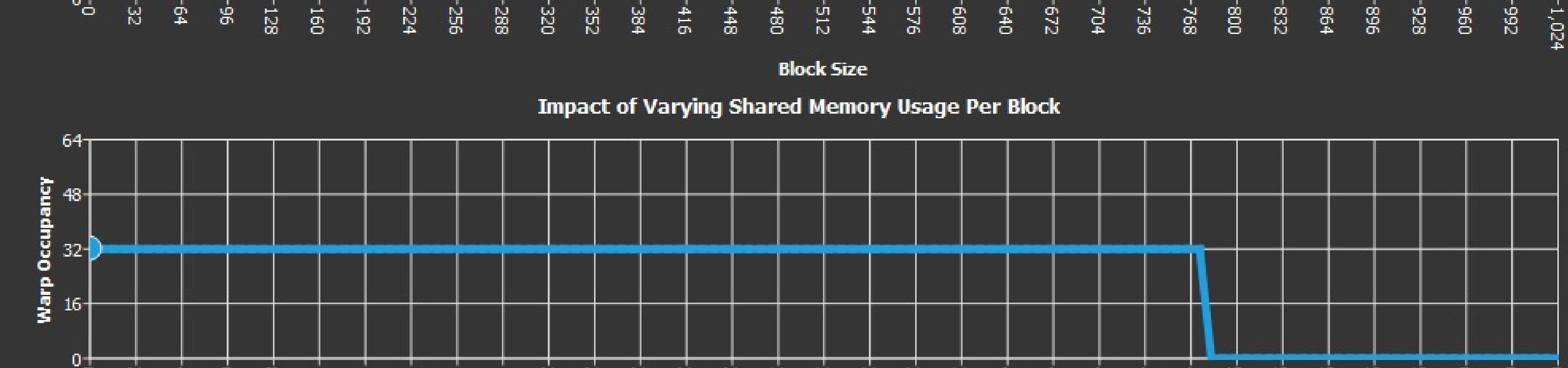
Impact of Varying Register Count Per Thread



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]	768	Branch Efficiency [%]	50
Branch Instructions Ratio [%]	0.13	Avg. Divergent Branches	1.33

Sampling Data (All)

Location	Value	Value (%)
No data available		

Sampling Data (Not Issued)

Location	Value	Value (%)
No data available		

Most Instructions Executed

Location	Value	Value (%)
0x700b8fba0 in gpu_insert_key	256	4
0x700b8f910 in gpu_insert_key	128	2
0x700b8f920 in gpu_insert_key	128	2
0x700b8f930 in gpu_insert_key	128	2
0x700b8f940 in gpu_insert_key	128	2

Recommendations

Uncoalesced Global Accesses [Warning] Uncoalesced global access, expected 128 sectors, got 512 (4.00x) at PC 0x700b8fb70

Uncoalesced Global Accesses [Warning] Uncoalesced global access, expected 128 sectors, got 512 (4.00x) at PC 0x700b8fb90

Uncoalesced Global Accesses [Warning] Uncoalesced global access, expected 128 sectors, got 256 (2.00x) at PC 0x700b8fb60

Uncoalesced Global Accesses [Warning] Uncoalesced global access, expected 128 sectors, got 256 (2.00x) at PC 0x700b8fb80