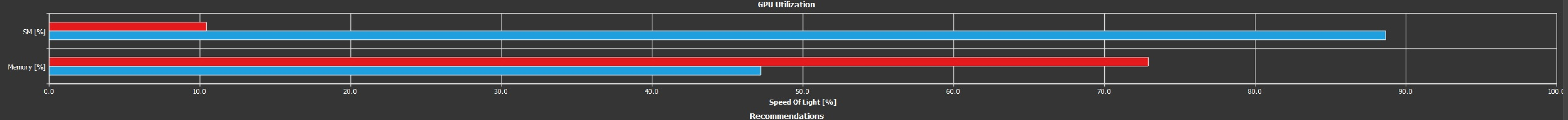
Page: Details | Launch: 0 - 141 - core_encrypt_ecb | Add Baseline | Apply Rules | Save as PDF

Current	141 - core_encrypt_ecb (4194304, 1, 1)	Time: 2.29 msecond	Cycles: 3,085,074	Regs: 61	GPU: GeForce RTX 2080 Ti	SM Frequency: 1.35 cycle/nsecond	CC: 7.5	Process: [4692] cis565_parallel_aes.exe
Unoptimized	141 - core_encrypt_ecb (4194304, 1, 1)	Time: 42.23 msecond	Cycles: 57,036,669	Regs: 42	GPU: GeForce RTX 2080 Ti	SM Frequency: 1.35 cycle/nsecond	CC: 7.5	Process: [1324] cis565_parallel_aes.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.

SOL SM [%]	88.64 (+749.94%)	Duration [msecond]	2.29 (-94.59%)
SOL Memory [%]	47.21 (-35.24%)	Elapsed Cycles [cycle]	3,085,074 (-94.59%)
SOL TEX [%]	47.75 (+4.46%)	SM Active Cycles [cycle]	3,048,355.75 (-94.58%)
SOL L2 [%]	5.67 (-92.22%)	SM Frequency [cycle/nsecond]	1.35 (-0.12%)
SOL FB [%]	9.00 (+1,302.30%)	Memory Frequency [cycle/nsecond]	1.70 (+1.78%)



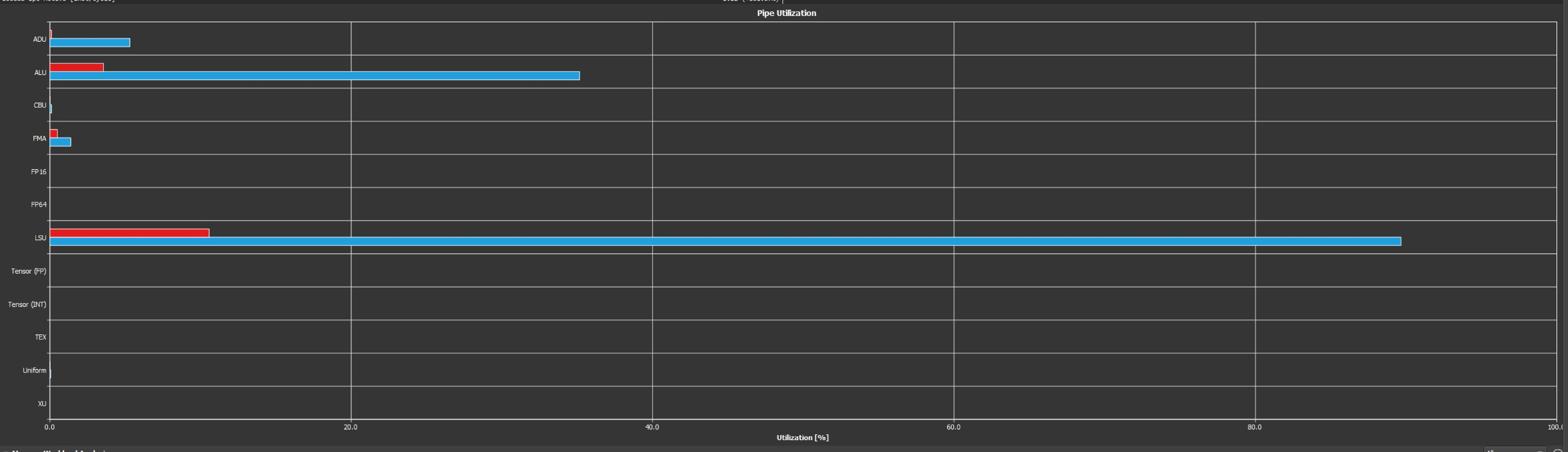
Recommendations

Bottleneck 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.

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]	1.21 (+803.35%)	SM Busy [%]	35.16 (+887.07%)
Executed Ipc Active [inst/cycle]	1.22 (+801.12%)	Issue Slots Busy [%]	30.56 (+801.37%)
Issued Ipc Active [inst/cycle]	1.22 (+801.37%)	-	-

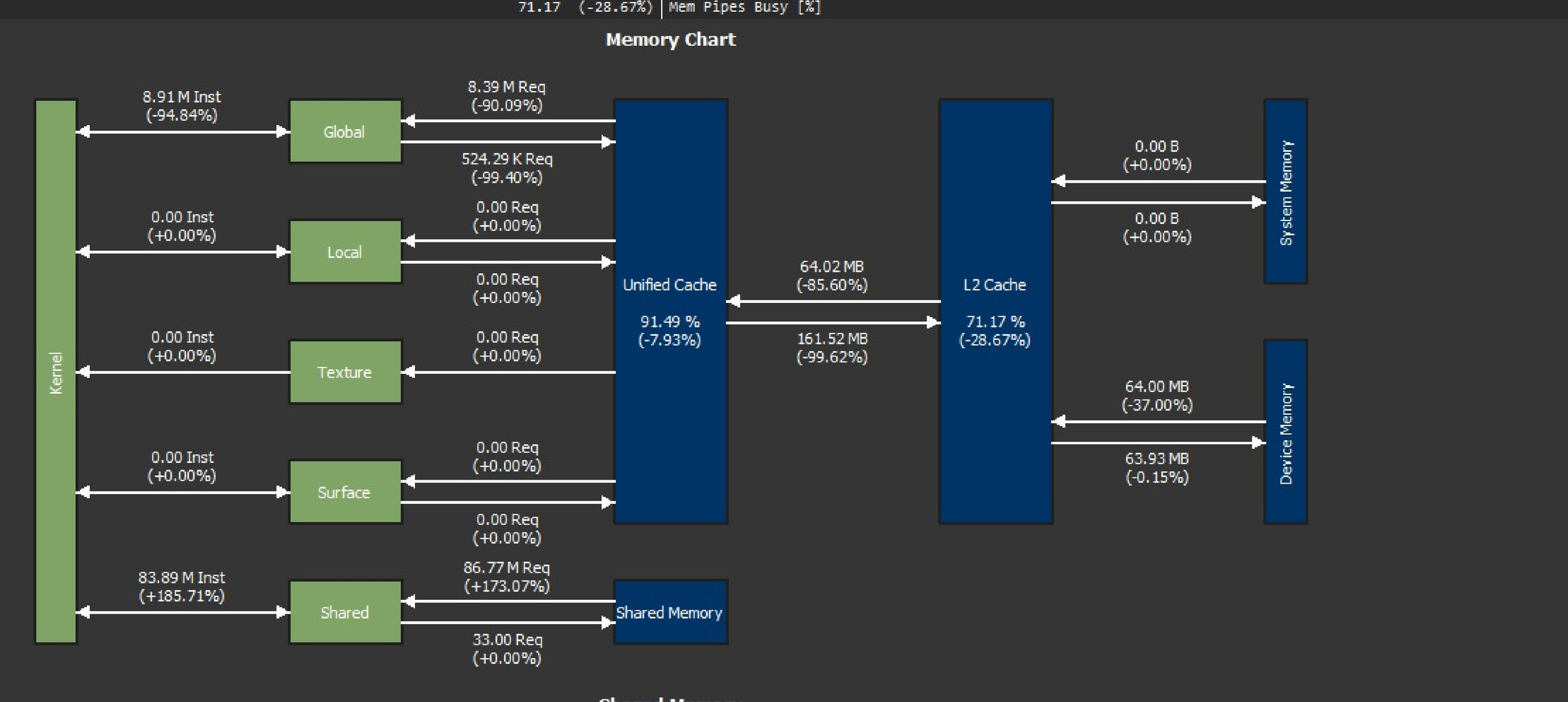


Memory Workload Analysis

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]	58.69 (+1,327.19%)	Mem Busy [%]	47.21 (-19.80%)
L1 Hit Rate [%]	91.49 (-7.93%)	Max Bandwidth [%]	44.07 (-39.55%)
L2 Hit Rate [%]	71.17 (-28.67%)	Mem Pipes Busy [%]	88.64 (+749.94%)



Shared Memory

	Instructions	Requests	% Peak	Bank Conflicts
Shared Load	83,886,080 (+185.71%)	86,771,328 (+173.07%)	41.86 (+4,938.50%)	2,885,248 (+19.40%)
Shared Store	24 (+0.00%)	33 (+0.00%)	0.00 (+1,745.15%)	0 (+0.00%)
Shared Atomic	0 (+0.00%)	-	-	-
Total	83,886,104 (+185.71%)	86,771,361 (+173.07%)	41.86 (+4,938.49%)	2,885,248 (+19.40%)

First-Level (Unified) Cache

	Instructions	SM->TEX Requests	% Peak	Hit Rate	TEX->L2 Requests	% Peak	L2->TEX Returns	% Peak	TEX->SM Returns	% Peak
Global Load Cached	8,388,608 (-90.09%)	8,388,608 (-90.09%)	4.05 (+82.80%)	87.09 (-11.45%)	-	-				
Global Load Uncached	-	-	-	-	-	-	2,097,696 (-85.60%)	1.01 (+165.77%)	92,406,524 (-19.14%)	44.58 (+1,391.92%)
Local Load Cached	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	-	-				
Local Load Uncached	-	-	-	-	-	-				
Surface Load	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	-	-	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Texture Load	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	-	-	0 (+0.00%)	0 (+0.00%)		
Global Store	524,288 (-99.40%)	524,288 (-99.40%)	0.25 (-89.02%)	100 (+0.00%)	5,292,728 (-99.62%)	2.55 (-93.01%)	-	-	-	-
Local Store	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)			-	-	-	-
Surface Store	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	-	-	-	-
Global Reduction	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	-	-	-	-
Surface Reduction	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	-	-	-	-
Global Atomic	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	see above	see above
Global Atomic Cas	0 (+0.00%)									
Surface Atomic	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	see above	see above
Surface Atomic Cas	0 (+0.00%)									
Loads	8,388,608 (-90.09%)	8,388,608 (-90.09%)	4.05 (+82.80%)	87.09 (-11.45%)	-	-	2,097,696 (-85.60%)	1.01 (+165.77%)	92,406,524 (-19.14%)	44.58 (+1,391.92%)
Stores	524,288 (-99.40%)	524,288 (-99.40%)	0.25 (-89.02%)	100 (+0.00%)	5,292,728 (-99.62%)	2.55 (-93.01%)	-	-	-	-
Total	8,912,896 (-94.84%)	8,912,896 (-94.84%)	4.30 (-4.80%)	91.49 (-7.93%)	5,292,728 (-99.62%)	2.55 (-93.01%)	2,097,696 (-85.60%)	1.01 (+165.77%)	92,406,524 (-19.14%)	44.58 (+1,391.92%)

Second-Level (L2) Cache

	TEX->L2 Requests	% Peak	L2->TEX Returns	% Peak	Total Bytes	Total Throughput
Global Load Cached	-	-			67,126,272 (-85.60%)	29,367,156,656.87 (+166.13%)
Global Load Uncached	-	-	2,097,696 (-85.60%)	1.01 (+165.77%)	-	-
Local Load Cached	-	-			-	-
Local Load Uncached	-	-			-	-
Surface Load	-	-	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Texture Load	-	-	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Global Store	5,292,728 (-99.62%)	2.55 (-93.01%)	-	-	169,367,296 (-99.62%)	74,096,710,065.80 (-93.00%)
Local Store			-	-		
Surface Store	0 (+0.00%)	0 (+0.00%)	-	-	0 (+0.00%)	0 (+0.00%)
Global Reduction	0 (+0.00%)	0 (+0.00%)	-	-	0 (+0.00%)	0 (+0.00%)
Surface Reduction	0 (+0.00%)	0 (+0.00%)	-	-	0 (+0.00%)	0 (+0.00%)
Global Atomic	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Global Atomic Cas						
Surface Atomic	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Surface Atomic Cas						
Loads	-	-	2,097,696 (-85.60%)	1.01 (+165.77%)	67,126,272 (-85.60%)	29,367,156,656.87 (+166.13%)
Stores	5,292,728 (-99.62%)	2.55 (-93.01%)	-	-	169,367,296 (-99.62%)	74,096,710,065.80 (-93.00%)
Total	5,292,728 (-99.62%)	2.55 (-93.01%)	2,097,696 (-85.60%)	1.01 (+165.77%)	236,493,568 (-99.48%)	103,463,866,722.67 (-90.32%)

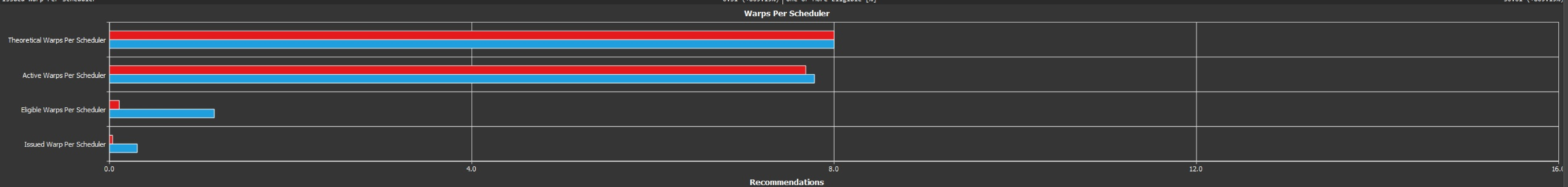
Device Memory (FB)

	L2<->FB Sectors	% Peak	Bytes	Throughput
Load	2,097,188 (-37.00%)	4.50 (+1,043.68%)	67,110,016 (-37.00%)	29,360,044,799.10 (+1,063.98%)
Store	2,094,775 (-0.15%)	4.50 (+1,712.67%)	67,032,800 (-0.15%)	29,326,263,474.73 (+1,744.85%)
Total	4,191,963 (-22.76%)	9.00 (+1,302.30%)	134,142,816 (-22.76%)	58,686,308,273.83 (+1,327.19%)

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]	7.78 (+1.25%)	Instructions Per Active Issue Slot [inst/cycle]	1 (+0.00%)
Eligible Warps Per Scheduler [warp]	1.16 (+984.07%)	No Eligible [%]	69.39 (-28.19%)
Issued Warp Per Scheduler	0.31 (+809.19%)	One or More Eligible [%]	30.61 (+809.19%)



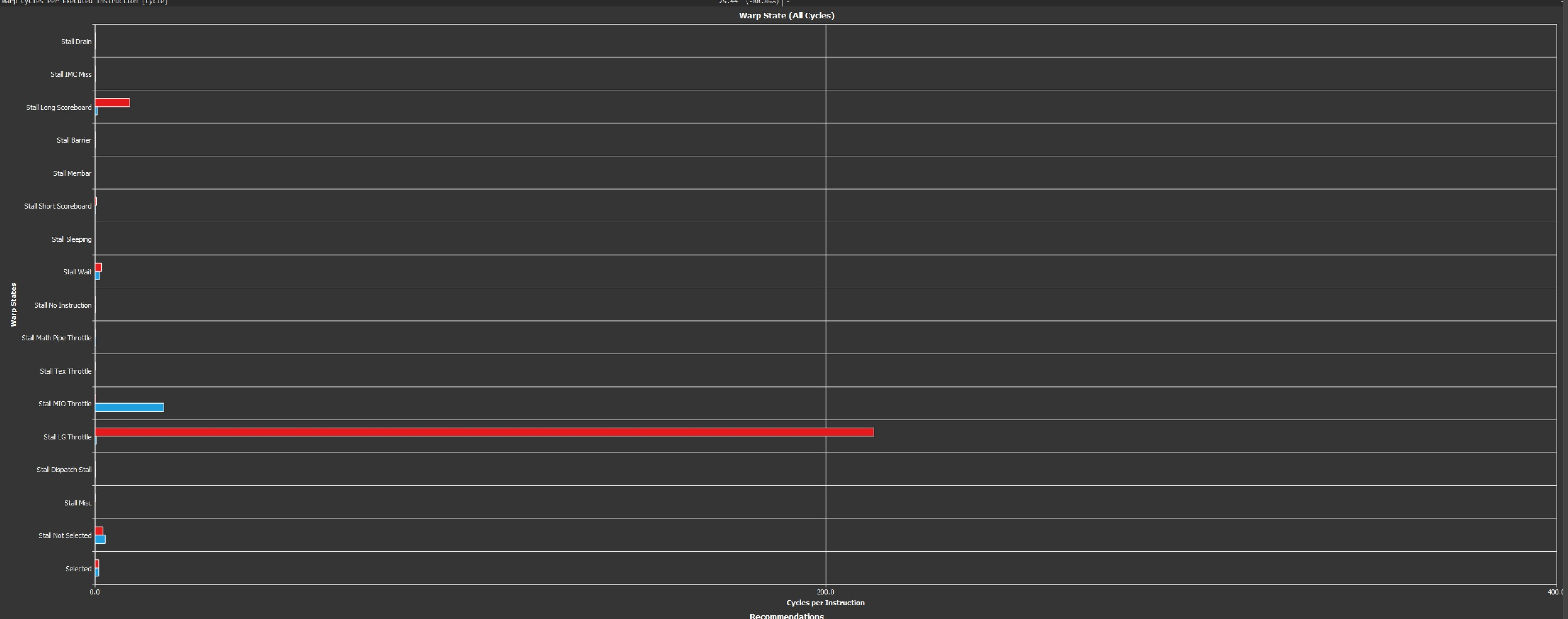
Recommendations

⚠ **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 3.3 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 7.78 active warps per scheduler, but only an average of 1.16 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.

Warp Cycles Per Issued Instruction	25.43 (-88.86%)	Avg. Active Threads Per Warp	32 (+0.00%)
Warp Cycles Per Issue Active	25.43 (-88.86%)	Avg. Not Predicated Off Threads Per Warp	31.92 (+5.38%)
Warp Cycles Per Executed Instruction [cycle]	25.44 (-88.86%)	-	-



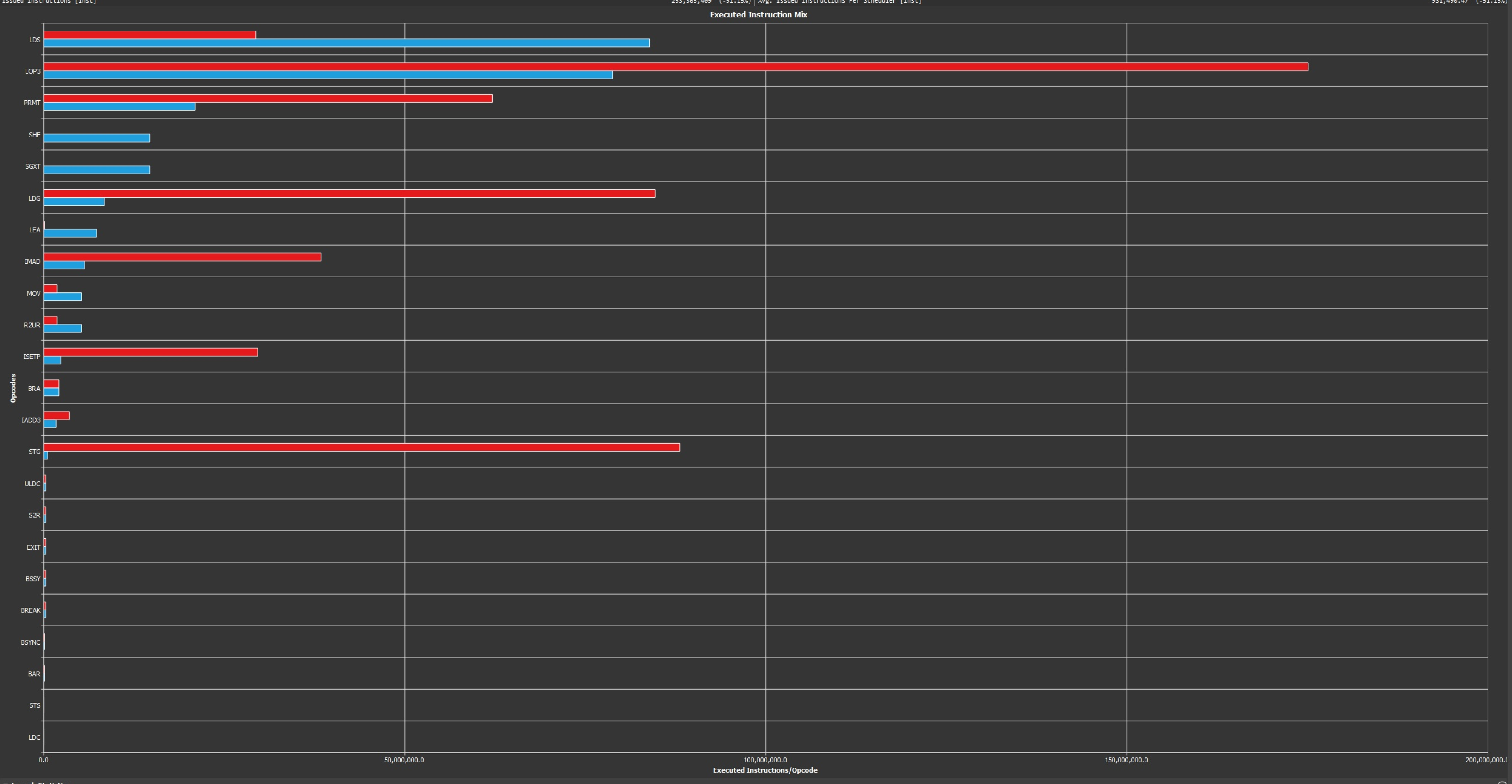
Recommendations

⚠ **CPI Stall 'MIO Throttle'** [Warning] On average each warp of this kernel spends 18.8 cycles being stalled waiting for the MIO instruction queue to be not full. This represents about 73.9% of the total average of 25.4 cycles between issuing two instructions. This stall reason is high in cases of extreme utilization of the MIO pipelines, which include special math instructions, dynamic branches, as well as shared memory instructions.

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.

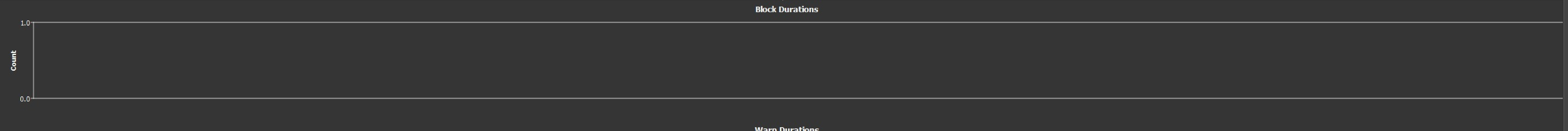
Executed Instructions [inst]	253,231,224 (-51.16%)	Avg. Executed Instructions Per Scheduler [inst]	930,997.15 (-51.16%)
Issued Instructions [inst]	253,365,409 (-51.15%)	Avg. Issued Instructions Per Scheduler [inst]	931,490.47 (-51.15%)



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	4,096 (+0.00%)	Registers Per Thread [register/thread]	61 (+45.24%)
Block Size	1,024 (+0.00%)	Static Shared Memory Per Block [byte/block]	768 (+0.00%)
Threads [thread]	4,194,304 (+0.00%)	Dynamic Shared Memory Per Block [byte/block]	0 (+0.00%)
Waves Per SM	60.24 (+0.00%)	Shared Memory Configuration Size [Kbyte]	32.77 (+0.00%)



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%)	Block Limit Registers [block]	1 (+0.00%)
Theoretical Active Warps per SM [warp/cycle]	32 (+0.00%)	Block Limit Shared Mem [block]	85 (+0.00%)
Achieved Occupancy [%]	97.15 (+0.63%)	Block Limit Warps [block]	1 (+0.00%)
Achieved Active Warps Per SM [warp]	31.09 (+0.63%)	Block Limit SM [block]	16 (+0.00%)