Page: Details | Launch: 0 - 508 - shared_mem_kernel | Add Baseline | Apply Duplicate | Save as PSF

	Launch	Time	Cycles	Regs	GPU	SM Frequency	CC	Process
Current	508 - shared_mem_kernel (256, 256, 1)x(32, 32, 1)	17.7 nsecond	19.980.742	26	NVIDIA GeForce GTX 1650	1.27 cycle/nsecond	7.5	21.92.942 ms

GPU Speed Of Light Throughput

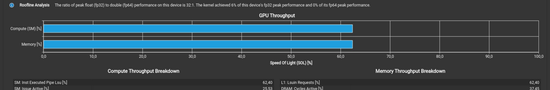
High-level overview of the throughput for compute and memory resources of the GPU. For each unit, the throughput reports the achieved percentage of utilization with respect to the theoretical maximum. Breakdowns show the throughput for each individual sub-metric of Compute and Memory to clearly identify 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 [%]	42,40	
Memory Throughput [%]	90,48	
L1/TEX Cache Throughput [%]	90,51	10.984.743
L2 Cache Throughput [%]	26,45	3.183.874,36
DRAM Throughput [%]	37,45	1,37
	37,45	5,98

Balanced Throughput Compute and Memory are well-balanced. To reduce runtime, both computation and memory traffic must be reduced. Check both the Compute Workload Analysis and Memory Workload Analysis report sections.

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.

GPU Throughput



Speed Of Light (SOL) [%]

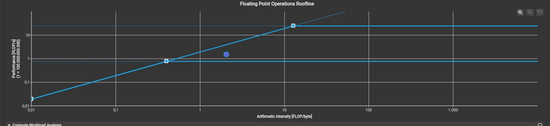
Compute Throughput Breakdown

SM: Instruct Executed Pipe Lsu [%]	42,40
SM: Inst Issued Active [%]	25,33
SM: Inst Executed [%]	25,33
SM: Mem Inst Issued [%]	23,64
SM: Fma Op Cycles Active [%]	17,02
SM: Fma Cycles Active [%]	14,69
SM: Alu0 Writeback Active [%]	11,18
SM: Inst Executed Pipe Adu [%]	8,51
SM: Adu Pipe Busy Cycles Active [%]	6,70
SM: Inst Executed Pipe Lsb [%]	5,29
SM: Mem Pipes Busy Cycles Cbu Pred On Any [%]	1,18
SM: Pipe Tensor Cycles Active [%]	0
SM: Inst Executed Pipe Uniform [%]	0
SM: Pipe Fp64 Cycles Active [%]	0
IDC: Request Cycles Active [%]	0
SM: Inst Executed Pipe Ic [%]	0
SM: Inst Executed Pipe Uniform [%]	0
SM: Inst Executed Pipe Tex [%]	0
SM: Inst Executed Pipe Xu [%]	0
SM: Inst Executed Pipe Ipa [%]	0

Memory Throughput Breakdown

L1: Lsuin Requests [%]	42,40
L1: DRAM Cycles Active [%]	37,45
L1: Lsuin Pipe Lsu Wavefronts [%]	31,33
L1: Sectors [%]	26,45
L1: DRAM: Dram Sectors [%]	22,27
L1: Lsu Writeback Active [%]	19,93
L1: Wavefronts Active [%]	17,60
L2: L1tocSlpr Cycles Active [%]	16,11
L2: L1 Sectors [%]	13,82
L2: L1 Lookup Miss Requests [%]	10,34
L2: L1: Tag Requests [%]	10,46
L1: Tex: L1tocSmsp Cycles Active [%]	9,71
L1: Wr Wbal2Tmm Read Sectors [%]	8,81
L1: Data Bank Reads [%]	7,69
L1: Data Bank Writes [%]	2,61
L1: Tex: L1tocSmsp Cycles Active [%]	0,04
L1: T Wavefronts [%]	0,00
L1: Data Pipe Tex Wavefronts [%]	0
L1: Tex Cache Sectors [%]	0
L2: L2 atomic input Cycles Active [%]	0
L2: Slices Fill Sectors [%]	0

Floating Point Operations Roofline



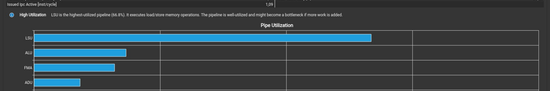
Compute Workload Analysis

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]	1,02	SM Busy [%]
Executed Ipc Active [inst/cycle]	1,02	Issue Slots Busy [%]
Issued Ipc Active [inst/cycle]	1,09	

High Utilization L1tex is the highest-utilized pipeline (84,8%). It executes local/store memory operations. The pipeline is well-utilized and might become a bottleneck if more work is added.

Pipe Utilization



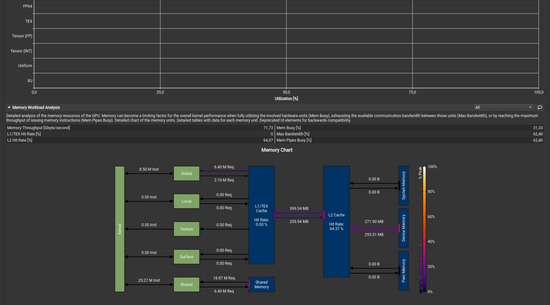
Memory Workload Analysis

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. Expectated kl elements for backwards compatibility.

Memory Throughput [Gbyte/second]	31,33	Mem Busy [%]
L1/TEX Hit Rate [%]	0	Max Bandwidth [%]
L2 Hit Rate [%]	64,57	Mem Pipes Busy [%]

Memory Chart



Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	18.869.760	18.869.760	18.869.760	12,76	0
Shared Load Matrix	0	0	0	0	0
Shared Store	6.404.100	6.404.100	6.404.100	4,33	0
Shared Atomic	0	0	0	0	0
Other			12.416.040	8,40	0
Total	25.273.860	25.273.860	37.691.900	25,50	0

L1/TEX Cache

	Instructions	Requests	Wavefronts	% Peak	Sectors	Sectors/Req	Hit Rate	Bytes	Sector Misses to L2	% Peak to L2	Returns to SM	% Peak to SM
Local Load	0	0	0	0	0	0	0	0	0	0	-	-
Global Load	6.404.100	6.404.100	6.404.100	4,33	13.085.700	2,04	0	418.742.400	13.085.700	8,85	6.404.100	4,33
Surface Load	0	0	0	0	0	0	0	0	0	0	0	0
Texture Load	0	0	0	0	0	0	0	0	0	0	-	-
Global Store	2.096.640	2.096.640	2.096.640	1,42	8.386.560	4	0	268.369.920	8.386.560	5,67	-	-
Local Store	0	0	0	0	0	0	0	0			-	-
Surface Store	0	0	0	0	0	0	0	0	0	0	-	-
Global Reduction	0	0	0	0	0	0	0	0	0	0	-	-
Surface Reduction	0	0	0	0	0	0	0	0	0	0	-	-
Global Atomic ALU	0	0	0	0	0	0	0	0	0	0	see above	see above
Global Atomic CAS	0	0	0	0	0	0	0	0	0	0	see above	see above
Surface Atomic ALU	0	0	0	0	0	0	0	0	0	0	see above	see above
Surface Atomic CAS	0	0	0	0	0	0	0	0	0	0	0	0
Loads	6.404.100	6.404.100	6.404.100	4,33	13.085.700	2,04	0	418.742.400	13.085.700	8,85	6.404.100	4,33
Stores	2.096.640	2.096.640	2.096.640	1,42	8.386.560	4	0	268.369.920	8.386.560	5,67	-	-
Total	8.500.740	8.500.740	8.500.740	5,75	21.472.260	2,53	0	687.112.320	21.472.260	14,52	6.404.100	4,33

L2 Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes	Throughput	Sector Misses to Device	Sector Misses to System	Sector Misses to Peer
L1/TEX Load	6.404.100	13.085.700	2,04	16,18	37,50	418.742.400	94.509.228.547,24	8.508.805	0	0
L1/TEX Store	2.096.640	8.386.560	4	10,62	100	268.369.920	34.806.974.268,04	0	0	0
L1/TEX Atomic ALU	0	0	0	0	0	0	0	0	0	0
L1/TEX Atomic CAS	0	0	0	0	0	0	0	0	0	0
L1/TEX Reduction	0	0	0	0	0	0	0	0	0	0
L1/TEX Total	8.500.740	21.472.260	2,53	26,42	60,81	687.112.320	99.115.742.815,80	8.508.805	0	0
GPU Total	8.500.974	21.472.297	2,53	26,42	60,81	687.181.504	99.120.646.549,57	8.508.913	0	0

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	8.509.577	19,31	285.106.484	39.977.177.648,29
Store	8.372.710	19,14	267.926.720	34.789.032.986,07
Total	17.882.287	37,45	553.033.184	77.726.210.634,35

L1/TEX Cache (Deprecated)

	Instructions	L1/TEX Requests	% Peak	Hit Rate	L2 Requests	% Peak	L2 Returns	% Peak	L1/TEX Returns	% Peak
Global Load Cached	6.404.100	6.404.100	4,33	0	-	-	-	-	-	-
Global Load Uncached	-	-	-	-	-	-	13.085.700	8,85	6.404.100	4,33
Local Load Cached	0	0	0	0	-	-	-	-	-	-
Local Load Uncached	-	-	-	-	-	-	-	-	-	-
Surface Load	0	0	0	0	-	-	0	0	0	0
Texture Load	0	0	0	0	-	-	0	0	-	-
Global Store	2.096.640	2.096.640	1,42	0	8.386.560	5,67	-	-	-	-
Local Store	0	0	0	0	-	-	-	-	-	-
Surface Store	0	0	0	0	-	-	-	-	-	-
Global Reduction	0	0	0	0	-	-	-	-	-	-
Surface Reduction	0	0	0	0	-	-	-	-	-	-
Global Atomic	0	0	0	0	-	-	-	-	see above	see above
Global Atomic Cas	0	0	0	0	-	-	-	-	-	-
Surface Atomic	0	0	0	0	-	-	0	0	see above	see above
Surface Atomic Cas	0	0	0	0	-	-	-	-	-	-
Loads	6.404.100	6.404.100	4,33	0	-	-	13.085.700	8,85	6.404.100	4,33
Stores	2.096.640	2.096.640	1,42	0	8.386.560	5,67	-	-	-	-
Total	8.500.740	8.500.740	5,75	0	8.386.560	5,67	13.085.700	8,85	6.404.100	4,33

L2 Cache (Deprecated)

	L2 Requests	% Peak	L2 Returns	% Peak	Total Bytes	Total Throughput
Global Load Cached	-	-	-	-	418.742.400	94.509.228.547,24
Global Load Uncached	-	-	13.085.700	8,85	-	-
Local Load Cached	-	-	-	-	-	-
Local Load Uncached	-	-	-	-	-	-
Surface Load	-	-	0	0	0	0
Texture Load	-	-	0	0	-	-
Global Store	8.386.560	5,67	-	-	268.369.920	34.806.974.268,04
Local Store	-	-	-	-	-	-
Surface Store	0	0	-	-	0	0
Global Reduction	0	0	-	-	0	0
Surface Reduction	0	0	-	-	0	0
Global Atomic	0	0	-	-	0	0
Global Atomic Cas	-	-	-	-	-	-
Surface Atomic	0	0	-	-	0	0
Surface Atomic Cas	-	-	-	-	-	-
Loads	-	-	13.085.700	8,85	418.742.400	94.509.228.547,24
Stores	8.386.560	5,67	-	-	268.369.920	34.806.974.268,04
Total	8.386.560	5,67	13.085.700	8,85	687.112.320	99.115.742.815,80

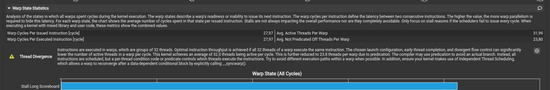
Scheduler Statistics

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 Warps). 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]	7,75	No Eligible [%]
Eligible Warps Per Scheduler [warp]	0,93	One or More Eligible [%]
Issued Warp Per Scheduler	0,28	

Issue Slot Utilization Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 3,6 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 7,75 active warps per scheduler, but only an average of 0,93 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, an instruction was issued and the issue slot remains unused. To increase the number of eligible warps, avoid possible long stalls or latency hiding by highly different execution durations per warp. Reducing stalls indicated on the Warp State and Scheduler Statistics pages might increase performance cue hints. Low occupancy may also help, too.

Warps Per Scheduler



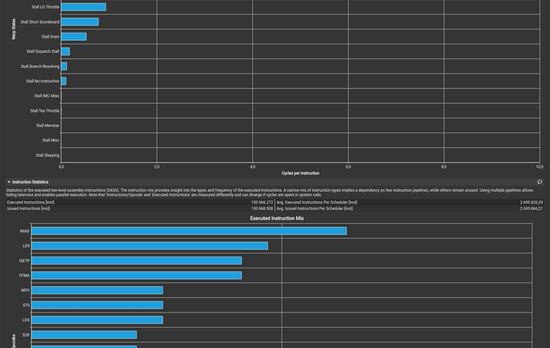
Warp State Statistics

Analysis of the states in which all warps spend 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 number of cycles is 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 scheduler fail to issue every cycle. Warps spending most cycles in a long stall state, more metrics show the underlying issues.

Warp Cycles Per Issued Instruction [cycle]	27,01	Avg. Active Threads Per Warp
Warp Cycles Per Executed Instruction [cycle]	27,97	Avg. Not Predicated Off Threads Per Warp

Thread Divergence Instructions are executed in warps, which are groups of 32 threads. Optimal instruction throughput is achieved if all 32 threads of a warp execute the same instruction. The chosen launch configuration, early thread completion, and divergent flow control can significantly lower the number of active threads in a warp per cycle. This kernel achieves an average of 32,0 threads being active per cycle. This is a lower value than 32,0 because one warp stalls indefinitely. The compiler may use predication to avoid an warp stall issue, instead all instructions are scheduled, but a per-thread condition code or predicate controls which threads execute the instructions. Try to avoid different execution paths within a warp where possible. In addition, ensure your kernel makes use of Independent Thread Scheduling, which allows a warp to reconverge after a data-dependent conditional block by explicitly calling __syncwarp().

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 opcodes are spent in spoken calls.

Executed Instructions [inst]	190.966.272	Avg. Executed Instructions Per Scheduler [inst]	2.695.826,29
Issued Instructions [inst]	190.966.189	Avg. Issued Instructions Per Scheduler [inst]	2.695.824,21

Executed Instruction Mix

NVLink

High-level summary of NVLink utilization. It shows the total received and transmitted (sent) memory, as well as the overall link peak utilization. NVLink Topology. Detailed tables with properties for each NVLink.

Physical Links	0	Logical Links

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	65.536	Registers Per Thread [register/thread]	26
Block Size	1.024	Shared Memory Per Block [Byte/block]	4,62
Threads [thread]	67.108.864	Dynamic Shared Memory Per Block [Byte/block]	0
Waves Per SM	2.048,00	Static Shared Memory Per Block [Byte/block]	0
Function Cache Configuration	cudaFuncCachePreferNone	Shared Memory Configuration Size [Kbyte]	32,77

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]	10
Achieved Occupancy [%]	96,51	Block Limit Warps [block]	1
Achieved Active Warps Per SM [warp]	30,88	Block Limit SM [block]	16

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

Source Counters

Summary statistics, 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]	4.195.792	Branch Efficiency [%]	0
Branch Instructions Ratio [%]	0,03	Avg. Divergent Branches	0