	Result	Size	Time	Cycles	GPU	SM Frequency	Process	Attributes
Current	573 - image_RGB2BW_kernel	(120, 68, 1)x(16, 16, 1)	238.59 us	344,607	0 - NVIDIA GeForce GTX 1660 Ti	1.44 Ghz	[26565] image	

Summary | Details | Source | Context | Comments | Raw | Session

Compare | Tools | View | Export

GPU Speed Of Light Throughput

GPU Throughput Chart

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 [%]	88.29	Duration [us]	238.59
Memory Throughput [%]	12.54	Elapsed Cycles [cycle]	344607
L1/TEX Cache Throughput [%]	17.15	SM Active Cycles [cycle]	341564.83
L2 Cache Throughput [%]	10.45	SM Frequency [Ghz]	1.44
DRAM Throughput [%]	12.54	DRAM Frequency [Ghz]	6.00

High Throughput This workload 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.

FP64/32 Utilization
Est. Speedup: 85.77%
The ratio of peak float (fp32) to double (fp64) performance on this device is 32:1. The workload achieved 0% of this device's fp32 peak performance and 32% of its fp64 peak performance. If Compute Workload Analysis determines that this workload is fp64 bound, consider using 32-bit precision floating point operations to improve its performance. See the Kernel Profiling Guide for more details on roofline analysis.

PM Sampling

Timeline view of PM metrics sampled periodically over the workload duration. Data is collected across multiple passes. Use this section to understand how workload behavior changes over its runtime.

Maximum Sampling Interval [cycle]	20000	# Pass Groups	1
Maximum Buffer Size [Kbytes]	192	Dropped Samples [sample]	0

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]	0.24	SM Busy [%]	88.53
Executed Ipc Active [inst/cycle]	0.25	Issue Slots Busy [%]	6.15
Issued Ipc Active [inst/cycle]	0.25		

Very High Utilization FP64 is the highest-utilized pipeline (88.5%) based on active cycles, taking into account the rates of its different instructions. It executes 64-bit floating point operations. The pipeline is over-utilized and likely a performance bottleneck. Based on the number of executed instructions, the highest utilized pipeline (88.5%) is FP64. It executes 64-bit floating point operations. Comparing the two, the overall pipeline utilization appears to be caused by frequent, low-latency instructions. See the Kernel Profiling Guide or hover over the pipeline name to understand the workloads handled by each pipeline. The Instruction Statistics section shows the mix of executed instructions for this workload. Check the Warp State Statistics section for which reasons cause warps to stall.

Memory Workload Analysis

Memory Chart

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/s]	36.09	Mem Busy [%]	9.91
L1/TEX Hit Rate [%]	57.33	Max Bandwidth [%]	12.54
L2 Hit Rate [%]	66.10	Mem Pipes Busy [%]	12.61

L1TEX Global Load Access Pattern
Est. Speedup: 12.86%
The memory access pattern for global loads from L1TEX might not be optimal. On average, only 8.0 of the 32 bytes transmitted per sector are utilized by each thread. This could possibly be caused by a stride between threads. Check the Source Counters section for uncoalesced global loads.

L1TEX Global Store Access Pattern
Est. Speedup: 8.58%
The memory access pattern for global stores to L1TEX might not be optimal. On average, only 16.0 of the 32 bytes transmitted per sector are utilized by each thread. This could possibly be caused by a stride between threads. Check the Source Counters section for uncoalesced global stores.

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]	7.33	No Eligible [%]	93.83
Eligible Warps Per Scheduler [warp]	0.10	One or More Eligible [%]	6.17
Issued Warp Per Scheduler	0.06		

Issue Slot Utilization
Est. Local Speedup: 11.71%
Every scheduler is capable of issuing one instruction per cycle, but for this workload each scheduler only issues an instruction every 16.2 cycles. This might leave hardware resources underutilized and may lead to less optimal performance. Out of the maximum of 8 warps per scheduler, this workload allocates an average of 7.33 active warps per scheduler, but only an average of 0.10 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]	118.70	Avg. Active Threads Per Warp	32
Warp Cycles Per Executed Instruction [cycle]	118.85	Avg. Not Predicated Off Threads Per Warp	30.97

Long Scoreboard Stalls
Est. Speedup: 11.71%
On average, each warp of this workload spends 60.9 cycles being stalled waiting for a scoreboard dependency on a L1TEX (local, global, surface, texture) operation. Find the instruction producing the data being waited upon to identify the culprit. 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 (coalescing), or by changing the cache configuration. Consider moving frequently used data to shared memory. This stall type represents about 51.3% of the total average of 118.7 cycles between issuing two instructions.

Tex Throttle Stalls
Est. Speedup: 11.71%
On average, each warp of this workload spends 50.8 cycles being stalled waiting for the L1 instruction queue for texture operations to be not full. This stall reason is high in cases of extreme utilization of the L1TEX pipeline. Try issuing fewer texture fetches, surface loads, surface stores, or decoupled math operations. If applicable, consider combining multiple lower-width memory operations into fewer wider memory operations and try interleaving memory operations and math instructions. Consider converting texture lookups or surface loads into global memory lookups. Texture can accept four threads' requests per cycle, whereas global accepts 32 threads. This stall type represents about 42.8% of the total average of 118.7 cycles between issuing two instructions.

Warp Stall Check the Warp Stall Sampling (All Samples) table 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]	2013600	Avg. Executed Instructions Per Scheduler [inst]	20975
Issued Instructions [inst]	2016199	Avg. Issued Instructions Per Scheduler [inst]	21002.07

FP64 Non-Fused Instructions
Est. Speedup: 14.76%
This kernel executes 129600 fused and 64800 non-fused FP64 instructions. By converting pairs of non-fused instructions to their fused, higher-throughput equivalent, the achieved FP64 performance could be increased by up to 17% (relative to its current performance). Check the Source page to identify where this kernel executes FP64 instructions.

NVLink Topology

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

NVLink Tables

Detailed tables with properties for each NVLink.

NUMA Affinity

Non-uniform memory access (NUMA) affinities based on compute and memory distances for all GPUs.

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	8160	Function Cache Configuration	CachePreferNone
Registers Per Thread [register/thread]	16	Static Shared Memory Per Block [byte/block]	0
Block Size	256	Dynamic Shared Memory Per Block [byte/block]	0
Threads [thread]	2088960	Driver Shared Memory Per Block [byte/block]	0
Waves Per SM	85	Shared Memory Configuration Size [Kbyte]	32.77
Uses Green Context	0	Stack Size	1024
# SMs [SM]	24	# TPCs	12
Enabled TPC IDs	all	-	-

Occupancy

% Occupancy Graphs

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]	16
Theoretical Active Warps per SM [warp]	32	Block Limit Shared Mem [block]	16
Achieved Occupancy [%]	91.87	Block Limit Warps [block]	4
Achieved Active Warps Per SM [warp]	29.40	Block Limit SM [block]	16

GPU and Memory Workload Distribution

Analysis of workload distribution in active cycles of SM, SMP, SMSP, L1 & L2 caches, and DRAM

Average SM Active Cycles [cycle]	341564.83	Average L1 Active Cycles [cycle]	341564.83
Average L2 Active Cycles [cycle]	284032.75	Average SMSP Active Cycles [cycle]	340206.88
Average DRAM Active Cycles [cycle]	179392	Total SM Elapsed Cycles [cycle]	8220088
Total L1 Elapsed Cycles [cycle]	8220088	Total L2 Elapsed Cycles [cycle]	3935736
Total SMSP Elapsed Cycles [cycle]	32880352	Total DRAM Elapsed Cycles [cycle]	8583168

Source Counters

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]	130080	Branch Efficiency [%]	0
Branch Instructions Ratio [%]	0.06	Avg. Divergent Branches	0

Uncoalesced Global Accesses
Est. Speedup: 61.86%
This kernel has uncoalesced global accesses resulting in a total of 648000 excessive sectors (71% of the total 907200 sectors). Check the L2 Theoretical Sectors Global Excessive table for the primary source locations. The CUDA Programming Guide has additional information on reducing uncoalesced device memory accesses.

L2 Theoretical Sectors Global Excessive

Location	Value	Value (%)
0x5018500 in image_RGB2BW_kernel	194,400	30
0x50185000 in image_RGB2BW_kernel	194,400	30
0x50185000 in image_RGB2BW_kernel	194,400	30
0x50185010 in image_RGB2BW_kernel	64,800	10

Follow the *rules outputs* to get guidance on how to navigate through the report and quickly discover performance bottlenecks in this kernel.
You could also disable individual sections to focus on selected performance aspects and make profiling faster.