	Result	Size	Time	Cycles	GPU	SM Frequency	Process	Attributes
Current	642 - distribution_elementwise_grid_stride_kernel	(408, 1, 1)x(256, 1, 1)	11.33 us	15,647	0 - NVIDIA GeForce RTX 3080	1.38 Ghz	[4112983] python3.8	

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 [%]	59.52	Duration [us]	11.33
Memory Throughput [%]	45.50	Elapsed Cycles [cycle]	15,647
L1/TEX Cache Throughput [%]	25.10	SM Active Cycles [cycle]	13,279.72
L2 Cache Throughput [%]	21.81	SM Frequency [Ghz]	1.38
DRAM Throughput [%]	45.50	DRAM Frequency [Ghz]	9.20

Latency Issue This kernel exhibits low compute throughput and memory bandwidth utilization relative to the peak performance of this device. Achieved compute throughput and/or memory bandwidth below 60.0% of peak typically indicate latency issues. Look at Scheduler Statistics and Warp State Statistics for potential reasons.

Roofline Analysis The ratio of peak float (fp32) to double (fp64) performance on this device is 64:1. The kernel achieved 10% 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.

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 [us]	2	# Pass Groups	2
Maximum Buffer Size [Mbytes]	32	Dropped Samples [sample]	0

PM Sampling Data Sampling interval is larger than 10% of the workload duration, which likely results in very few collected samples. For better results, use the --pm-sampling-interval option to reduce the sampling interval. Use --pm-sampling-buffer-size to increase the sampling buffer size for the smaller interval, or don't set a fixed buffer size and let the tool adjust it automatically.

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]	2.35	SM Busy [%]	68.92
Executed Ipc Active [inst/cycle]	2.72	Issue Slots Busy [%]	68.92
Issued Ipc Active [inst/cycle]	2.76		

Balanced ALU is the highest-utilized pipeline (55.7%) based on active cycles, taking into account the rates of its different instructions. It executes integer and logic operations. It is well-utilized, but should not be a bottleneck.

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]	334.98	Mem Busy [%]	21.81
L1/TEX Hit Rate [%]	0	Max Bandwidth [%]	45.50
L2 Hit Rate [%]	99.69	Mem Pipes Busy [%]	8.77
L2 Compression Success Rate [%]	0	L2 Compression Ratio	0

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]	8.36	No Eligible [%]	30.99
Eligible Warps Per Scheduler [warp]	3.31	One or More Eligible [%]	69.01
Issued Warp Per Scheduler	0.69		

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]	12.11	Avg. Active Threads Per Warp	32
Warp Cycles Per Executed Instruction [cycle]	12.29	Avg. Not Predicated Off Threads Per Warp	30.04

Not Selected Stalls
Est. Local Speedup: 31.39%

On average, each warp of this kernel spends 3.8 cycles being stalled waiting for the micro scheduler to select the warp to issue. Not selected warps are eligible warps that were not picked by the scheduler to issue that cycle as another warp was selected. A high number of not selected warps typically means you have sufficient warps to cover warp latencies and you may consider reducing the number of active warps to possibly increase cache coherence and data locality. This stall type represents about 31.4% of the total average of 12.1 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]	2,452,416	Avg. Executed Instructions Per Scheduler [inst]	9,016.24
Issued Instructions [inst]	2,489,289	Avg. Issued Instructions Per Scheduler [inst]	9,151.80

FP32 Non-Fused Instructions
Est. Speedup: 6.73%

This kernel executes 349376 fused and 189568 non-fused FP32 instructions. By converting pairs of non-fused instructions to their fused, higher-throughput equivalent, the achieved FP32 performance could be increased by up to 18% (relative to its current performance). Check the Source page to identify where this kernel executes FP32 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	408	Function Cache Configuration	CachePreferNone
Registers Per Thread [register/thread]	38	Static Shared Memory Per Block [byte/block]	0
Block Size	256	Dynamic Shared Memory Per Block [byte/block]	0
Threads [thread]	104,448	Driver Shared Memory Per Block [Kbyte/block]	1.02
Waves Per SM	1	Shared Memory Configuration Size [Kbyte]	8.19
Uses Green Context	0	# SMs [SM]	68

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]	6
Theoretical Active Warps per SM [warp]	48	Block Limit Shared Mem [block]	8
Achieved Occupancy [%]	69.79	Block Limit Warps [block]	6
Achieved Active Warps Per SM [warp]	33.50	Block Limit SM [block]	16

Achieved Occupancy
Est. Local Speedup: 30.21%

The difference between calculated theoretical (100.0%) and measured achieved occupancy (69.8%) can be the result of warp scheduling overheads or workload imbalances during the kernel execution. Load imbalances can occur between warps within a block as well as across blocks of the same kernel. See the CUDA Best Practices Guide for more details on optimizing occupancy.

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]	13,279.72	Average L1 Active Cycles [cycle]	13,279.72
Average L2 Active Cycles [cycle]	9,483.30	Average SMSP Active Cycles [cycle]	13,261.50
Average DRAM Active Cycles [cycle]	47,433.60	Total SM Elapsed Cycles [cycle]	1,045,600
Total L1 Elapsed Cycles [cycle]	1,045,600	Total L2 Elapsed Cycles [cycle]	617,040
Total SMSP Elapsed Cycles [cycle]	4,182,400	Total DRAM Elapsed Cycles [cycle]	1,042,432

L2 Slices Workload Imbalance
Est. Speedup: 5.92%

One or more L2 Slices have a much higher number of active cycles than the average number of active cycles. Maximum instance value is 9.62% above the average, while the minimum instance value is 5.67% below the average.

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]	150,144	Branch Efficiency [%]	100
Branch Instructions Ratio [%]	0.06	Avg. Divergent Branches	0

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.