Page: Details | Result: 2 - 131 - 0_bm | Add Baseline | Apply Rules | Occupancy Calculator | Source Comparison | Save as PDF

	Result	Time	Cycles	Regs	GPU	SM Frequency	CC	Process
Current	131 - 0_bm (391, 1, 1)x(256, 1, 1)	13.37 ms(cond)	2,08,10,661	168	0 - NVIDIA GeForce RTX 3090 Ti	1.56 cycle/nsec(cond)	8.6	[9020] Decrypt_Serial

GPU Speed Of Light Throughput

GPU Throughput Chart

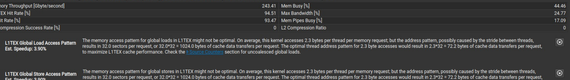
High-level overview of the throughput for compute and memory resources of the GPU. For each unit, the throughput reports the achieved percentage of utilization with respect to the theoretical maximum. Breakdowns show the throughput for each individual sub-metric of Compute and Memory to clearly identify 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 [%]	44.46	Duration [msecond]	13.37
Memory Throughput [%]	44.46	Elapsed Cycles [cycle]	2,08,10,661
L1/TEX Cache Throughput [%]	48.19	SM Active Cycles [cycle]	1,92,01,264.31
L2 Cache Throughput [%]	22.18	SM Frequency [cycle/nsecond]	1.56
DRAM Throughput [%]	24.77	GRAM Frequency [cycle/nsecond]	10.23

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 0% of this device's fp32 peak performance and 0% of its fp64 peak performance. See the Nsight Compute Roofline Analysis guide for more details on roofline analysis.

GPU Throughput



PM Sampling

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

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]	1.36	SM Busy [%]	47.08
Executed Ipc Active [inst/cycle]	1.47	Issue Slots Busy [%]	36.81
Issued Ipc Active [inst/cycle]	1.47		

INFO ALU is the highest-utilized pipeline (47.1%) 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/second]	243.47	Mem Busy [%]	44.46
L1/TEX Hit Rate [%]	94.51	Max Bandwidth [%]	24.77
L2 Hit Rate [%]	93.47	Mem Pipes Busy [%]	17.09
L2 Compression Success Rate [%]	0	L2 Compression Ratio	0

L1TEX Global Load Access Pattern Est. Speedup: 3.95% The memory access pattern for global loads in L1TEX might not be optimal. On average, this kernel accesses 2.3 bytes per thread per memory request, but the address pattern, possibly caused by the stride between threads, results in 32.2 sectors per request, or 32.2*32 = 1024.0 bytes of cache data transfers per request. The optimal thread address pattern for 2.3 byte accesses would result in 2.3*32 = 72.2 bytes of cache data transfers per request, to maximize L1TEX cache performance. Check the L1 Cache Guide section for uncoalesced global loads.

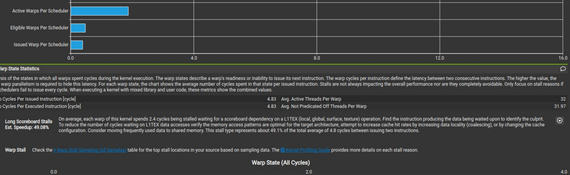
L1TEX Global Store Access Pattern Est. Speedup: 3.95% The memory access pattern for global stores in L1TEX might not be optimal. On average, this kernel accesses 2.3 bytes per thread per memory request, but the address pattern, possibly caused by the stride between threads, results in 32.2 sectors per request, or 32.2*32 = 1024.0 bytes of cache data transfers per request. The optimal thread address pattern for 2.3 byte accesses would result in 2.3*32 = 72.2 bytes of cache data transfers per request, to maximize L1TEX cache performance. Check the L1 Cache Guide section for uncoalesced global stores.

L1TEX Local Load Access Pattern Est. Speedup: 1.83% The memory access pattern for local loads in L1TEX might not be optimal. On average, this kernel accesses 2.3 bytes per thread per memory request, but the address pattern, possibly caused by the stride between threads, results in 4.0 sectors per request, or 4.0*32 = 128.0 bytes of cache data transfers per request. The optimal thread address pattern for 2.3 byte accesses would result in 2.3*32 = 72.2 bytes of cache data transfers per request, to maximize L1TEX cache performance. Check the L1 Cache Guide section for uncoalesced local loads.

L1TEX Local Store Access Pattern Est. Speedup: 1.83% The memory access pattern for local stores in L1TEX might not be optimal. On average, this kernel accesses 2.3 bytes per thread per memory request, but the address pattern, possibly caused by the stride between threads, results in 4.0 sectors per request, or 4.0*32 = 128.0 bytes of cache data transfers per request. The optimal thread address pattern for 2.3 byte accesses would result in 2.3*32 = 72.2 bytes of cache data transfers per request, to maximize L1TEX cache performance. Check the L1 Cache Guide section for uncoalesced local stores.

L2 Local Access Pattern Est. Speedup: 2.16% The memory access pattern for loads from L1TEX to L2 is not optimal. The granularity of an L1TEX request to L2 is a 128 byte cache line. That is 4 consecutive 32-byte sectors per L2 request. However, this kernel only accesses an average of 2.5 sectors out of the possible 4 sectors per cache line. Check the L2 Cache Guide section for uncoalesced loads and try to minimize how many cache lines need to be accessed per memory request.

Memory Chart



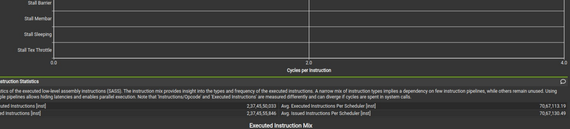
Scheduler Statistics

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	1.87	No Eligible [%]	61.23
Eligible Warps Per Scheduler [warp]	0.48	One or More Eligible [%]	38.77
Issued Warp Per Scheduler	0.39		

Issue Slot Utilization Est. Local Speedup: 55.54% Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 2.6 cycles. This might leave hardware resources underutilized and may lead to less optimal performance. Out of the maximum of 16 warps per scheduler, this kernel allocates an average of 1.87 active warps per scheduler, but only an average of 0.48 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 Scheduler Statistics sections can help, too.

Warps Per Scheduler



Warp State Statistics

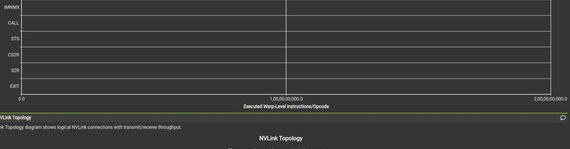
Analysis of the states in which all warps spent cycles during the kernel execution. The warp states describe a warp's readiness or inability to issue its next instruction. The warp cycles per instruction define the latency between two consecutive instructions. The higher the value, the more warps parallelism 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]	4.83	Avg. Active Threads Per Warp	32
Warp Cycles Per Executed Instruction [cycle]	4.83	Avg. Not Predicated Off Threads Per Warp	31.97

Long Scoreboard Stalls Est. Speedup: 49.08% On average, each warp of this kernel spends 2.4 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 49.1% of the total average of 4.8 cycles between issuing two instructions.

Warp Stall Check the Warp Stall Sampling (All Samples) table for the top stall locations in your source based or sampling data. The Nsight Compute Profiling Guide provides more details on each stall reason.

Warp State (All Cycles)

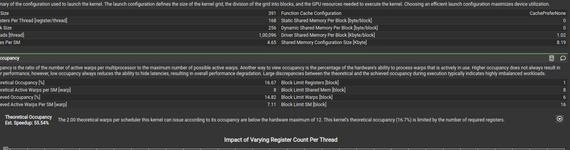


Instruction Statistics

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 cycles are spent in system calls.

Executed Instructions [inst]	2,37,45,50,015	Avg. Executed Instructions Per Scheduler [inst]	75,67,113.19
Issued Instructions [inst]	2,37,45,50,846	Avg. Issued Instructions Per Scheduler [inst]	75,67,130.49

Executed Instruction Mix



NVLink Topology

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

NVLink Topology

The system does not have any NVLink connections.

NVLink Tables

Detailed tables with properties for each NVLink.

Logical NVLink Properties

The system does not have any NVLink connections.

NUMA Affinity

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

NUMA ID Table

GPU ID	GPU Name	CPU Affinity
0	NVIDIA GeForce RTX 3090 Ti	0-15

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	391	Function Cache Configuration	CachePreferNone
Registers Per Thread [register/thread]	168	Static Shared Memory Per Block [Kbyte/block]	0
Block Size	256	Dynamic Shared Memory Per Block [Kbyte/block]	0
Threads [thread]	1,00,096	Driver Shared Memory Per Block [Kbyte/block]	1.02
Waves Per SM	4.69	Shared Memory Configuration Size [Kbyte]	8.19

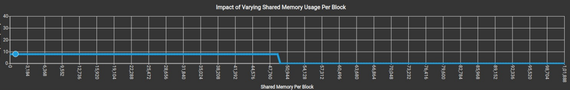
Occupancy

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.

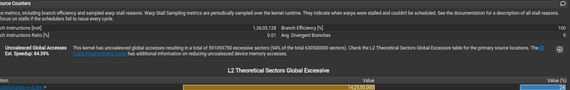
Theoretical Occupancy [%]	16.67	Block Limit Registers [block]	1
Theoretical Active Warps per SM [warp]	8	Block Limit Shared Mem [block]	8
Achieved Occupancy [%]	16.69	Block Limit Warps [block]	6
Achieved Active Warps Per SM [warps]	7.11	Block Limit SM [block]	16

Theoretical Occupancy Est. Speedup: 83.34% The 2.00 theoretical warps per scheduler this kernel can issue according to its occupancy are below the hardware maximum of 12. This kernel's theoretical occupancy (16.7%) is limited by the number of required registers.

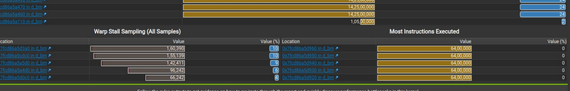
Impact of Varying Register Count Per Thread



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. 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]	1,36,03,139	Branch Efficiency [%]	100
Branch Instructions Ratio [%]	0.01	Avg. Divergent Branches	0

Uncoalesced Global Accesses Est. Speedup: 84.31% This kernel has uncoalesced global accesses resulting in a total of 501050750 excessive sectors (84% of the total 630500000 sectors). Check the L2 Theoretical Sectors Global Excessive table for the primary source locations. The Nsight Compute Profiling Guide has additional information on reducing uncoalesced device memory accesses.

L2 Theoretical Sectors Global Excessive

Location	Value	Value (%)
0x7f1d8ba4e9e0 in Calc_Kernel(int, int, int,