Page: Details | Launch: 0 - 123 - kernel | Add Baseline | Apply Rules | Save as PDF

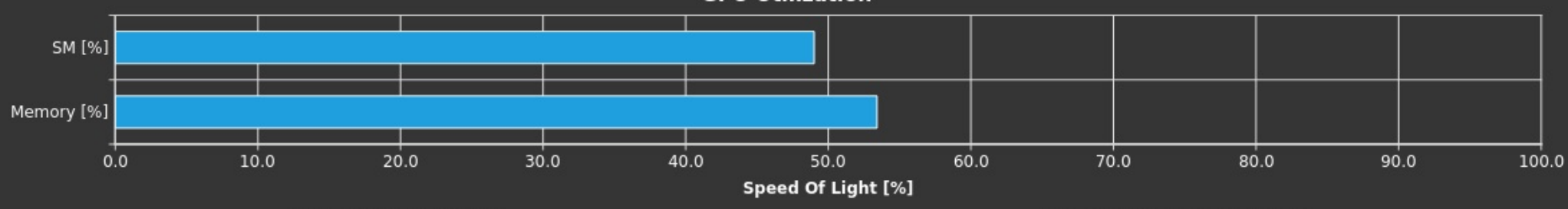
Current 12... **Time:** 336.70 usecond **Cycles:** 505,649 **Regs:** 18 **GPU:** NVIDIA GeForce RTX 3070 **SM Frequency:** 1.50 cycle/nsecond **CC:** 8.6 **Process:** [2063114] main

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. High-level overview of the utilization for compute and memory resources of the GPU presented as a roofline chart.

SOL SM [%]	49.01	Duration [usecond]	336.70
SOL Memory [%]	53.42	Elapsed Cycles [cycle]	505649
SOL L1/TEX Cache [%]	49.45	SM Active Cycles [cycle]	502786.11
SOL L2 Cache [%]	18.59	SM Frequency [cycle/nsecond]	1.50
SOL DRAM [%]	53.42	DRAM Frequency [cycle/nsecond]	6.81

GPU Utilization



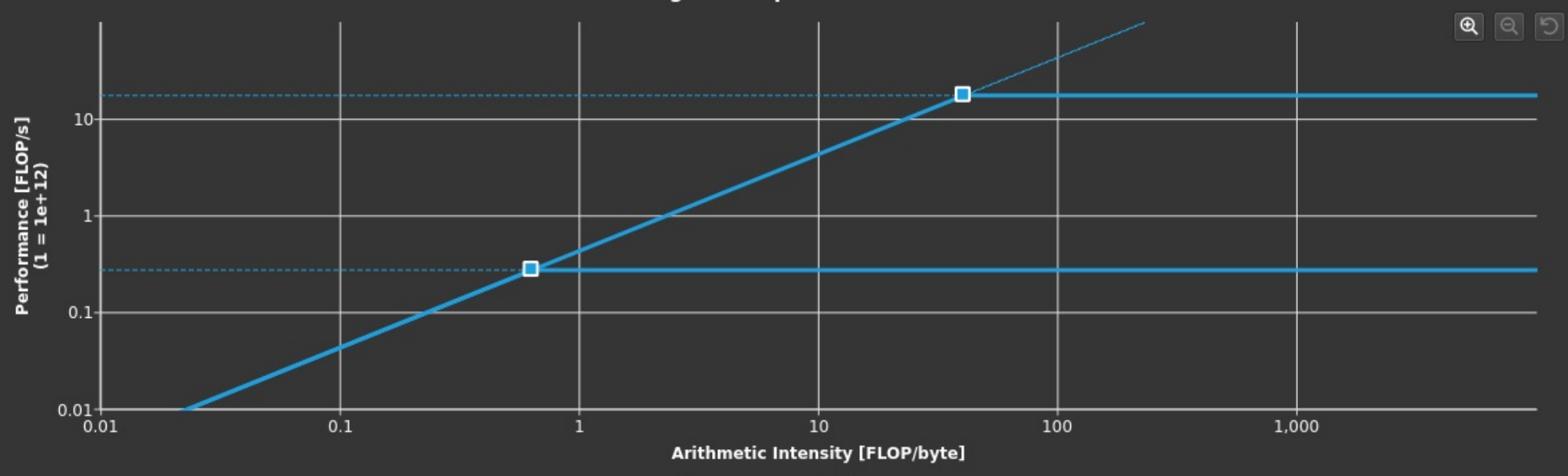
SOL SM Breakdown

SOL SM: Inst Executed Pipe Lsu [%]	49.01
SOL SM: Issue Active [%]	30.97
SOL SM: Inst Executed [%]	30.85
SOL SM: Pipe Alu Cycles Active [%]	22.14
SOL SM: Mio Inst Issued [%]	21.81
SOL SM: Inst Executed Pipe Adu [%]	16.42
SOL SM: Inst Executed Pipe Cbu Pred On Any [%]	14.24
SOL SM: Pipe Fmaheavy Cycles Active [%]	12.00
SOL SM: Mio Pq Write Cycles Active [%]	11.06
SOL SM: Mio Pq Read Cycles Active [%]	10.81
SOL SM: Mio2rf Writeback Active [%]	7.41
SOL SM: Pipe Fma Cycles Active [%]	6.00
SOL SM: Inst Executed Pipe Uniform [%]	1.01
SOL SM: Inst Executed Pipe Xu [%]	0.17
SOL SM: Inst Executed Pipe Ipa [%]	0
SOL SM: Inst Executed Pipe Tex [%]	0
SOL IDC: Request Cycles Active [%]	0
SOL SM: Pipe Fp64 Cycles Active [%]	0
SOL SM: Pipe Tensor Cycles Active [%]	0

SOL Memory Breakdown

SOL DRAM: Cycles Active [%]	53.42
SOL L1: Lsuin Requests [%]	49.23
SOL DRAM: Dram Sectors [%]	31.84
SOL L1: Data Pipe Lsu Wavefronts [%]	26.51
SOL L2: T Sectors [%]	18.59
SOL L1: Lsu Writeback Active [%]	14.86
SOL L2: Xbar2lts Cycles Active [%]	11.35
SOL L2: Lts2xbar Cycles Active [%]	10.08
SOL L2: D Sectors [%]	8.63
SOL L2: D Sectors Fill Device [%]	8.13
SOL L1: M L1tex2xbar Req Cycles Active [%]	7.06
SOL L1: M Xbar2l1tex Read Sectors [%]	6.67
SOL L2: T Tag Requests [%]	5.20
SOL L1: Data Bank Writes [%]	1.63
SOL L1: Data Bank Reads [%]	1.61
SOL L1: Texin Sm2tex Req Cycles Active [%]	0.51
SOL L2: D Atomic Input Cycles Active [%]	0.13
SOL L2: D Sectors Fill Sysmem [%]	0.00
SOL L1: F Wavefronts [%]	0
SOL L1: Data Pipe Tex Wavefronts [%]	0
SOL L1: Tex Writeback Active [%]	0

Floating Point Operations Roofline



Recommendations

Bottleneck [Warning] 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.

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]	1.23	SM Busy [%]	31.11
Executed Ipc Active [inst/cycle]	1.24	Issue Slots Busy [%]	31.11
Issued Ipc Active [inst/cycle]	1.24		

Memory Workload Analysis

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]	232.65	Mem Busy [%]	26.51
L1/TEX Hit Rate [%]	0.77	Max Bandwidth [%]	53.42
L2 Hit Rate [%]	56.54	Mem Pipes Busy [%]	49.01
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]	10.85	No Eligible [%]	69.37
Eligible Warps Per Scheduler [warp]	0.58	One or More Eligible [%]	30.63
Issued Warp Per Scheduler	0.31		

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]	35.41	Avg. Active Threads Per Warp	31.84
Warp Cycles Per Executed Instruction [cycle]	35.55	Avg. Not Predicated Off Threads Per Warp	27.03

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]	28669446	Avg. Executed Instructions Per Scheduler [inst]	155812.21
Issued Instructions [inst]	28780893	Avg. Issued Instructions Per Scheduler [inst]	156417.90

NVLink

High-level summary of NVLink connection status.

Physical Links	0	Logical Links	0

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	19532	Registers Per Thread [register/thread]	18
Block Size	512	Static Shared Memory Per Block [byte/block]	144
Threads [thread]	10000384	Dynamic Shared Memory Per Block [byte/block]	0
Waves Per SM	141.54	Driver Shared Memory Per Block [Kbyte/block]	1.02
Function Cache Configuration	cudaFuncCachePreferNone	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 [%]	100	Block Limit Registers [block]	5
Theoretical Active Warps per SM [warp]	48	Block Limit Shared Mem [block]	80
Achieved Occupancy [%]	94.03	Block Limit Warps [block]	3
Achieved Active Warps Per SM [warp]	45.13	Block Limit SM [block]	16

Source Counters

Source metrics, 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]	4113225	Branch Efficiency [%]	98.87
Branch Instructions Ratio [%]	0.14	Avg. Divergent Branches	106.54