Current	Result	Size	Time	Cycles	GPU	Start	Frequency	Process	Attributes
	0 (0 - applyMultiheadSoftmaxKernel)	1	2.62 s	1,626,5..1	1	2.62 s	620.4 MHz	[3.0.0] Cudalib	
	1	1	2.62 s	1,626,4..1	0	248.384 s	0 - Tesla T4	616.79 MHz	[3.0.0] Cudalib

Details Source [1] Context [1] Comments Tips Statistics

Compare Node View Export

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. Compute and Memory throughput are likely 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 [%]	33,85
Memory Throughput [%]	85,74
L1/TEX Cache Throughput [%]	36,69
L2 Cache Throughput [%]	50,05
DRAM Throughput [%]	85,74
Duration [us]	425,44
Elapsed Cycles [cycle]	248,924
Int Active Cycles [cycle]	245,500,32
FP Active Cycles [cycle]	535,74
Tensor Active Cycles [cycle]	0,54

High Throughput This kernel is utilizing greater than 85.0% 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 DRAM in the Memory Workload Analysis section.

Compute Bottleneck Detect bottlenecks arising from compute capabilities — Apply

Memory Bottleneck Detect bottlenecks arising from memory capabilities — Apply

Roofline Analysis The ratio of peak float (fp32) to double (fp64) performance on this device is 32:1. The kernel achieved 1% 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.

GPU Throughput

Compute Throughput Breakdown

SM: Inst Executed Pipes cycle [%]	33,85
SM: Pipe Tensor cycles active [%]	0,54
SM: Inst Issued [%]	16,29
SM: Mem Issue Issued [%]	13,34
SM: Mem Pq Write Cycles Active [%]	10,26
SM: Mem Pq Read Cycles Active [%]	10,26
SM: Pipe Alu Cycles Active [%]	9,49
SM: Pipe Fma Cycles Active [%]	8,71
SM: Mem Pipe Busy Cycles [%]	8,71
SM: Mad Pipe Busy Cycles [%]	3,84
SM: Inst Executed Pipes Cbu Pred On Any [%]	5,14
SM: Inst Executed Pipes Lsu Pred On Any [%]	4,71
SM: Inst Executed Pipe Lsu [%]	3,08
SM: Inst Executed Pipe Uniform [%]	0,51
SM: Inst Executed Pipe Cbu Pred On Any [%]	0
SM: Pipe Shared Cycles active [%]	0
SM: Pipe Fp64 Cycles Active [%]	0
SM: Inst Executed Pipes Tex [%]	0
SM: Inst Executed Pipe Ipa [%]	0
SM: Inst Executed Pipe Fp16 [%]	0

Memory Throughput Breakdown

DRAM: Cycles Active [%]	85,74
DRAM: Read Bandwidth [%]	85,74
DRAM: Write Bandwidth [%]	0,07
L2: Sectors [%]	29,09
L2: Tag Requests [%]	50,05
L1: Tex Cache Busy Cycles Active [%]	79,30
L1: Data Pipe Lsu Wavefronts [%]	18,37
L2: Xbar Cycles Active [%]	17,23
L1: Sf Xbar2Tex Bwd Bytes [%]	16,41
L1: Data Bank Conflicts Pipe [%]	15,62
L2: Sectors Fill Device [%]	14,04
L1: Sectors [%]	13,95
L2: Sectors Hit Device [%]	8,72
L2: Tag Requests [%]	7,55
L1: Data Bank Writes [%]	4,16
L1: Sectors Miss Pipe Lsu [%]	3,08
L1: Texs Sm2Tex Req Cycles Active [%]	0,51
L1: F Wavefronts [%]	0,51
L1: Data Pipe Tex Wavefronts [%]	0
L1: Tex Writeback Active [%]	0
L2: Atomic Input Cycles Active [%]	0
L2: Sectors Fill Fb Device [%]	0

Floating Point Operations Roofline

Achieved Value
Peak Fp32 Fmaa
Peak Fp32
Peak Fp32 Fmaa
Peak Fp32 Add
Peak Fp32 Mul
Peak Fp32 Add
Peak Fp32 Mul
Peak Fp32 Add
Peak Fp32 Mul