Page: Details | Launch: 15 - 11x15 - ampere_sgemm_128x128_nn | Kernel: 11x15 - ampere_sgemm_128x128_nn (64, 64, 1)x(256, 1, 1) | Add Baseline | Apply Baline | Occupancy Calculator | Save as PDF

Session: Duration: 1.14 s | Date: 2021-Jan-20 | Time: 16:19:14 | Device: NVIDIA A100-SXM4-40GB | Process: [1940]0[0] mamba1

GPU Speed Of Light Throughput

High-level overview of the throughputs 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 instances of the GPU presented as a roofline chart.

Compute (SM) Throughput [%]	92.98	Duration [msecond]	1.14
Memory Throughput [%]	92.98	Elapsed Cycles [cycle]	1,427,968
DRAM Throughput [%]	1.19	Memory [%]	92.98
L1/TEX Cache Throughput [%]	93.11	SM Active Cycles [cycle]	1,426,518.00
L2 Cache Throughput [%]	93.11	SM Frequency [cycle/usecond]	1,254.00
Device Throughput [%]	1.19	Dram Frequency [cycle/usecond]	1.00

High Throughput This kernel's utilizing greater than 60.0% 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.

Roofline Analysis The ratio of peak float (fp32) to double (fp64) performance on this device is 2:1. The kernel achieved 94% of this device's fp32 peak performance and 0% of its fp64 peak performance. See the Kernel Profiling Guide for mode details on roofline analysis.

GPU Throughput

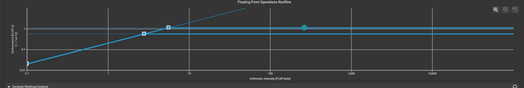
Compute Throughput Breakdown

SM Inst Executed [%]	92.98	L1 Cache Fill Active [%]	93.11
SM Pipe Fma Cycles Active [%]	92.98	L1 Cache Evictions Active [%]	93.11
SM Pipe Lsu Cycles Active [%]	92.98	L1 Cache Requests [%]	93.11
SM Pipe Fma Fp64 Cycles Active [%]	0.00	L1 Cache Hit Rate [%]	0.00
SM Issued Warpsched Active [%]	13.66	L1 T Sectors [%]	0.00
SM Warp Issue Stalled [%]	13.66	L1 Read Bytes Active [%]	13.75
SM Warp Issue Stalled Lsu [%]	13.66	L1 Write Bytes Active [%]	93.11
SM Wbc Fp Wbfdp Cycles Active [%]	0.13	L1 Scoreboard Cycles Active [%]	0.08
SM Wbc Fp Scoreboard Cycles Active [%]	0.13	L1 Scoreboard Wait Cycles Active [%]	0.00
SM Inst Executed Pipe Fma Cycles Active [%]	92.98	L1 Tag Requests Global Sectorized [%]	0.00
SM Inst Executed Pipe Fma Fp64 Ops [%]	0.00	L1 Sectors [%]	0.75
SM Inst Executed Pipe Fma [%]	92.98	L1 Data Bank Wrfdp [%]	0.00
SM Issued Cycles Active [%]	13.66	DRAM Cycles Active [%]	0.08
SM Issued Dual Issue Cycles [%]	0.00	L1 Data Bank Conflicts [%]	0.00
SM Inst Executed Pipe Fp64 [%]	0.00	L1 Sectors Hit Pipe [%]	0.00
SM Inst Executed Pipe Fp32 [%]	0.00	L1 Tag Stalls Mio Cycles Active [%]	0.00
SM Inst Executed Pipe Lsu [%]	0.00	L1 Tag Stalls Lg Cycles Active [%]	0.00
SM Pipe Shared Cycles Active [%]	0.00	L1 Wavefronts Pipe Tex Cycles [%]	0.00
SM Pipe Shared Cycles Active [%]	0.00	L1 Wavefronts Pipe Lsu Cycles [%]	0.00
SM Pipe Tensor Cycles Active [%]	0.00	L1 Wavefronts Pipe Shared Cycles [%]	0.00
SM Pipe Tensor Cycles Active [%]	0.00	L1 Tag Requests Global [%]	0.00

Memory Throughput Breakdown

L1 Data Pipe Lsu Wavefronts [%]	13.66	Mem Busy [%]	13.66
L1 Lsu Wavefronts [%]	13.66	Max Bandwidth [%]	27.75
L1 Lsu Requests [%]	13.66	Mem Pipes Busy [%]	13.66
L1 Hit Rate [%]	0.00	L2 Cache Throughput [%]	93.11
L1 T Sectors [%]	0.00	L2 Read Bytes Active [%]	13.75
L1 Read Bytes Active [%]	13.75	L2 Write Bytes Active [%]	93.11
L1 Write Bytes Active [%]	93.11	L2 Lookup Cycles Active [%]	0.08
L1 Lookup Cycles Active [%]	0.08	L2 Lookup Wait Cycles Active [%]	0.00
L1 Lookup Wait Cycles Active [%]	0.00	L2 Sectors Hit Pipe Tex Sectors [%]	0.00
L1 Sectors Hit Pipe Tex Sectors [%]	0.00	L2 Sectors [%]	0.75
L1 Sectors [%]	0.75	L2 Data Bank Wrfdp [%]	0.00
L1 Data Bank Wrfdp [%]	0.00	DRAM Cycles Active [%]	0.08
L1 Data Bank Conflicts [%]	0.00	L2 Sectors Hit Pipe [%]	0.00
L1 Sectors Hit Pipe [%]	0.00	L2 Sectors Hit Pipe Lsu [%]	0.00
L1 Tag Stalls Mio Cycles Active [%]	0.00	L2 Tag Stalls Mio Cycles Active [%]	0.00
L1 Tag Stalls Lg Cycles Active [%]	0.00	L2 Tag Stalls Lg Cycles Active [%]	0.00
L1 Wavefronts Pipe Tex Cycles [%]	0.00	L2 Wavefronts Pipe Tex Cycles [%]	0.00
L1 Wavefronts Pipe Lsu Cycles [%]	0.00	L2 Wavefronts Pipe Lsu Cycles [%]	0.00
L1 Wavefronts Pipe Shared Cycles [%]	0.00	L2 Wavefronts Pipe Shared Cycles [%]	0.00
L1 Tag Requests Global [%]	0.00	L2 Tag Requests Global [%]	0.00

Floating Point Operations Roofline



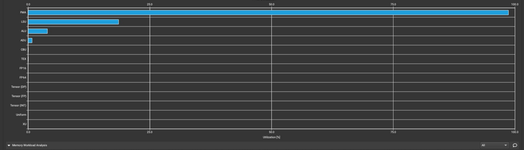
Compute Workload Analysis

Detailed analysis of the compute resources of the streaming multiprocessor (SM), including the achieved instructions per clock (IPC) and the utilization of each available pipeline. Pipelines with very high utilization often limit the overall performance.

Executed Ipc Active [inst/cycle]	1.00	SM Busy [%]	93.66
Executed Ipc Elapsed [inst/cycle]	0.99	Issue Slots Busy [%]	24.73
Issued Ipc Active [inst/cycle]	1.01		

Issue Slot Utilization FMA is the highest utilized pipeline (100.1%). It executes 32-bit floating point (FMA, FMUL, FADD,...) and integer (IMAD, IMUL, IMNMX) operations. The pipeline is a core unit that can cause performance bottlenecks. Start the Instruction Statistics Guide for the workloads handled by each pipeline.

Pipe Utilization



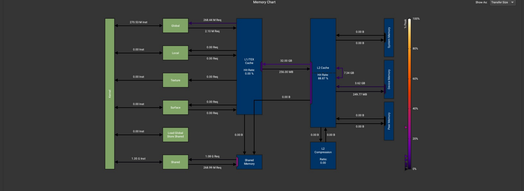
Memory Workload Analysis

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 running the maximum throughput of issuing memory instructions (Mem Pipes Busy). Detailed chart of the memory units, detailed tables with key KPIs for each hardware unit.

Memory Throughput [Gbyte/s]	1,555.20	Mem Busy [%]	13.66
L1/TEX Hit Rate [%]	0.00	Max Bandwidth [%]	27.75
L2 Hit Rate [%]	50.07	Mem Pipes Busy [%]	13.66
L2 Compression Success Rate [%]	0.00	L2 Compression Ratio	0.00

L2 Local Access Pattern The memory access pattern for Gayle Point L1/TEX hit (0.0%) is not optimal. The granularity of an L1/TEX request to L2 is 128 bytes vs the ideal 32 bytes. This is a consequence 32 byte sectors per L2 request. However, this kernel only accesses an average of 1.4 sectors out of this possible 4 sectors per cache line. Check the Memory Chart section for uncoalesced loads and 32 B transactions from/to cache lines caused by the excessive L2 sectors request.

Memory Chart


Show As:
Transfer Size
Device
Uniform
Shared
Local Global
Texture
Surface
Global Reduction
Stack
System Memory
Unified Memory
276.70 GB/sec
0.00 GB/sec
0.00 GB/sec
0.00 GB/sec
0.00 GB/sec
0.00 GB/sec
1.05 GB/sec
0.00 GB/sec
0.00 GB/sec
0.00 GB/sec
0.00 GB/sec
0.00