Start: 2024-05-17 23:24:51 +0200 (Measurement 1) Application: ./build/bin/llama-bench -m /home/user/Models/Qwen2.5-Coder-7B-Instruct-Q4_K_M.gguf -p 512 -n 128 -t 16 -ngl 99 -fa 1 -ctk q8_0 -ctv q8_0 -r 3 -o jsonl -v Kernel: 6.8.0-1027-oem Hostname: user-MS-Tower-i9-14900K Profiler: NVIDIA Nsight Compute 2024.2.1.0 (build 34678576) (public-release)

Time: 21.45 minutes Cycles: 43.01 Bil Regs: 32 GPU: Tesla V100-SXM2-16GB SM Frequency: 1.31 cycles/nsec DRAM Frequency: 0.88 cycles/nsec Process: 2107600 python3 Range: 0

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.

SM [%]	79.72	Duration [ms]	21.88 ms
Memory [%]	46.42	L1/TEX Cache Throughput [%]	80.13
DRAM Throughput [%]	46.42	L2 Cache Throughput [%]	46.42
Max Bandwidth [%]	46.42	Mem Busy [%]	46.42
Compute (SM) Throughput [%]	79.72		

GPU Utilization

Compute (SM) [%]
Memory [%]
0
20
40
60
80
100
0.0
20.0
40.0
60.0
80.0
100.0
Duration [ms]

SOL SM Breakdown

SOL SM Issue Active [%]	79.72	(±0.00%)
SOL SM Issue Inactive [%]	20.28	(±0.00%)
SOL SM Issue Inst Per Cycle Max [%]	49.86	(±0.00%)
SOL SM Issue Inst Per Cycle [%]	39.72	(±0.00%)
SOL SM Issue Dual Per Cycle [%]	0.00	(±0.00%)
SOL SM Pipe Fma Cycles Active [%]	79.72	(±0.00%)
SOL SM Pipe Alu Cycles Active [%]	0.00	(±0.00%)
SOL SM Pipe Lsu Cycles Active [%]	0.00	(±0.00%)
SOL SM Pipe Tex Cycles Active [%]	0.00	(±0.00%)
SOL SM Pipe Uniform Cycles Active [%]	0.00	(±0.00%)
SOL SM Pipe Shared Cycles Active [%]	0.00	(±0.00%)
SOL SM Pipe Tensor Cycles Active [%]	0.00	(±0.00%)
SOL SM Pipe Xu Cycles Active [%]	0.00	(±0.00%)

SOL Memory Breakdown

SOL L1/TEX Cache [%]	80.13	(±0.00%)
SOL L2 Cache [%]	46.42	(±0.00%)
SOL Low Utilization Across [%]	29.62	(±0.00%)
SOL DRAM Cycles Active [%]	46.42	(±0.00%)
SOL DRAM Cycles Elapsed [%]	46.42	(±0.00%)
SOL DRAM Cycles Active Rd [%]	46.42	(±0.00%)
SOL DRAM Cycles Active Wr [%]	0.00	(±0.00%)
SOL L1/TEX Throughput [%]	80.13	(±0.00%)
SOL L1/TEX Cycles Active [%]	80.13	(±0.00%)
SOL L1/TEX Cycles Elapsed [%]	80.13	(±0.00%)
SOL L1/TEX Lsu Cycles Active [%]	0.00	(±0.00%)
SOL L1/TEX Tex Cycles Active [%]	0.00	(±0.00%)
SOL L1/TEX Lsu Wb Cycles Active [%]	0.00	(±0.00%)
SOL L1/TEX Tex Wb Cycles Active [%]	0.00	(±0.00%)

Floating Point Operations Roofline

Floating Point Operations/second
Arithmetic Intensity [FLOPs/byte]

Arithmetic Intensity [FLOPs/byte]

Recommendations

Softmax - Warning! Compute is more heavily utilized than Memory. Look at the Compute Workload Analysis section to see what the compute pipelines are spending their time doing. Also, consider whether any computation is redundant and could be reduced or moved to look up tables.

Roofline Analysis - The ratio of peak float (fp32) to double (fp64) performance on this device is 2:1. The kernel achieved 39% 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 Active [inst/cycle]	79.72	(±0.00%)
Executed Ipc Elapsed [inst/cycle]	79.72	(±0.00%)
Issued Ipc Active [inst/cycle]	79.72	(±0.00%)

Pipe Utilization

Pipe Utilization [%]
0
20
40
60
80
100
Tensor (DP)
Tensor (Int)
Tensor (FP)
ADU
CBU
FMA
FP16
FP64
ALU
LSU
XU
Tex
Uniform
Shared
Fma
Fp16
Fp64
Alu
Lsu
Xu
Tex
Uniform
Shared
0.0
20.0
40.0
60.0
80.0
100.0
Duration [ms]

Recommendations

Pipe Fma Utilization - No pipeline is over-utilized.

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]	99.19	(±0.00%)
L1/TEX Hit Rate [%]	80.13	(±0.00%)
L2 Hit Rate [%]	46.42	(±0.00%)
Mem Busy [%]	46.42	(±0.00%)

Memory Chart

Device Memory
L2 Cache
L1/TEX Cache
Shared Memory
SM
Unified L1/TEX Cache
Shared Memory
Texture Memory
Global Memory
Local Memory
Constant Memory
Surface