Page: Details | Launch: 1 - 133 - sgemm_fast_kernel | Add Baseline | Apply Rules | Save as PDF

Current 133 ⋯ **Time:** 6.29 msecond **Cycles:** 4,898,524 **Regs:** 109 **GPU:** GeForce RTX 3070 Laptop GPU **SM Frequency:** 779.33 cycle/usecond **CC:** 8.6 **Process:** [24388] hpclab6.exe

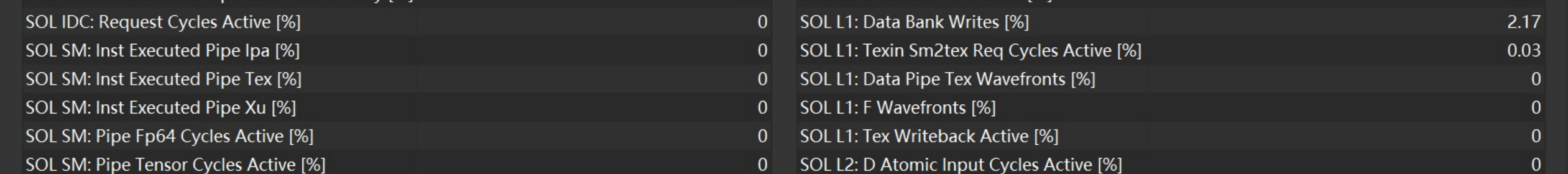
GPU Speed Of Light

All

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.

SOL SM [%]	37.73	Duration [msecond]	6.29
SOL Memory [%]	86.27	Elapsed Cycles [cycle]	4,898,524
SOL L1/TEX Cache [%]	95.03	SM Active Cycles [cycle]	4,447,106.70
SOL L2 Cache [%]	10.70	SM Frequency [cycle/usecond]	779.33
SOL DRAM [%]	5.61	DRAM Frequency [cycle/nsecond]	6.11

GPU Utilization



SOL SM Breakdown

SOL SM: Issue Active [%]	37.73
SOL SM: Inst Executed [%]	37.73
SOL SM: Pipe Fmaheavy Cycles Active [%]	34.48
SOL SM: Pipe Fma Cycles Active [%]	34.45
SOL SM: Inst Executed Pipe Lsu [%]	22.52
SOL SM: Mio2rf Writeback Active [%]	14.02
SOL SM: Mio Inst Issued [%]	7.78
SOL SM: Mio Pq Read Cycles Active [%]	2.27
SOL SM: Mio Pq Write Cycles Active [%]	2.21
SOL SM: Inst Executed Pipe Adu [%]	0.81
SOL SM: Pipe Alu Cycles Active [%]	0.75
SOL SM: Inst Executed Pipe Uniform [%]	0.00
SOL SM: Inst Executed Pipe Cbu Pred On Any [%]	0.00
SOL IDC: Request Cycles Active [%]	0
SOL SM: Inst Executed Pipe Ipa [%]	0
SOL SM: Inst Executed Pipe Tex [%]	0
SOL SM: Inst Executed Pipe Xu [%]	0
SOL SM: Pipe Fp64 Cycles Active [%]	0
SOL SM: Pipe Tensor Cycles Active [%]	0

SOL Memory Breakdown

SOL L1: Data Pipe Lsu Wavefronts [%]	86.27
SOL L1: Lsu Writeback Active [%]	31.57
SOL L1: Lsuin Requests [%]	22.52
SOL L2: T Sectors [%]	10.70
SOL L1: Data Bank Reads [%]	10.20
SOL L2: Lts2xbar Cycles Active [%]	10.07
SOL L1: M Xbar2l1tex Read Sectors [%]	8.51
SOL L2: Xbar2lts Cycles Active [%]	5.85
SOL GPU: Dram Throughput [%]	5.61
SOL L2: T Tag Requests [%]	5.23
SOL L1: M L1tex2xbar Req Cycles Active [%]	4.95
SOL L2: D Sectors [%]	3.32
SOL L2: D Sectors Fill Device [%]	2.29
SOL L1: Data Bank Writes [%]	2.17
SOL L1: Texin Sm2tex Req Cycles Active [%]	0.03
SOL L1: Data Pipe Tex Wavefronts [%]	0
SOL L1: F Wavefronts [%]	0
SOL L1: Tex Writeback Active [%]	0
SOL L2: D Atomic Input Cycles Active [%]	0
SOL L2: D Sectors Fill Sysmem [%]	0

Recommendations

Bottleneck The kernel is utilizing greater than 80.0% 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 Memory Workload Analysis section.

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	256	Registers Per Thread [register/thread]	109
Block Size	256	Static Shared Memory Per Block [Kbyte/block]	16.38
Threads [thread]	65,536	Dynamic Shared Memory Per Block [byte/block]	0
Waves Per SM	3.20	Driver Shared Memory Per Block [Kbyte/block]	1.02
		Shared Memory Configuration Size [Kbyte]	65.54

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 [%]	33.33	Block Limit Registers [block]	2
Theoretical Active Warps per SM [warp]	16	Block Limit Shared Mem [block]	5
Achieved Occupancy [%]	31.85	Block Limit Warps [block]	6
Achieved Active Warps Per SM [warp]	15.29	Block Limit SM [block]	16