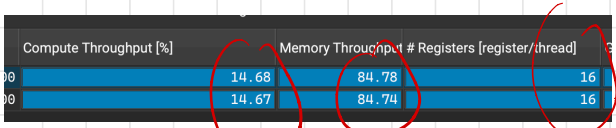
A100 profiling

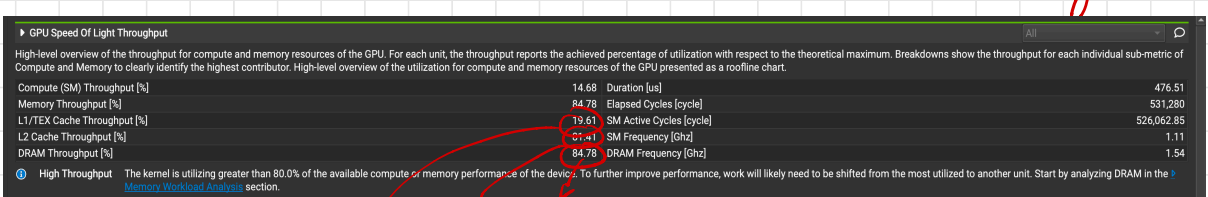
VectorAdd



→ low reg pressure
⇓
high occupancy
⇓
more latency hiding

low util of compute units

close to max util of memory bw



L2 acts as passthru for all dram loads

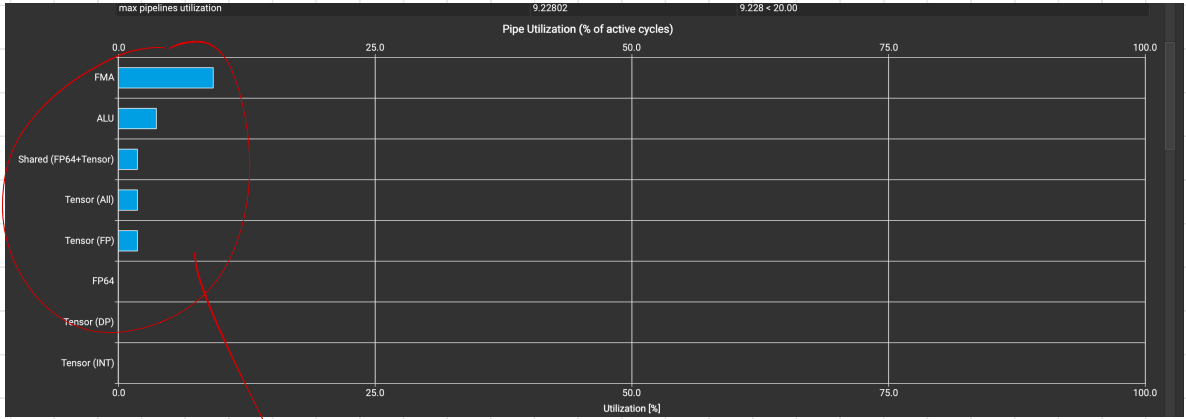
has much higher bandwidth/speed than DRAM so gets underutilized here

Compute Throughput Breakdown

SM: Issue Active [%]	14.68
SM: Inst Executed [%]	14.67
SM: Inst Executed Pipe Lsu [%]	14.67
SM: Mio Inst Issued [%]	9.17
SM: Pipe Fma Cycles Active [%]	9.17
SM: Mio2rf Writeback Active [%]	7.34
SM: Inst Executed Pipe Adu [%]	7.34
SM: Inst Executed Pipe Cbu Pred On Any [%]	4.79
SM: Mio Pq Write Cycles Active [%]	3.67
SM: Mio Pq Read Cycles Active [%]	3.67
SM: Pipe Alu Cycles Active [%]	3.67
SM: Pipe Tensor Cycles Active [%]	1.83
SM: Pipe Shared Cycles Active [%]	1.83
SM: Inst Executed Pipe Uniform [%]	1.83
SM: Inst Executed Pipe Tex [%]	0
SM: Inst Executed Pipe Ipa [%]	0
SM: Inst Executed Pipe Fp16 [%]	0
SM: Pipe Fp64 Cycles Active [%]	0
SM: Inst Executed Pipe Xu [%]	0
IDC: Request Cycles Active [%]	0

Memory Throughput Breakdown

DRAM: Cycles Active [%]	84.78
DRAM: Dram Sectors [%]	61.69
L2: D Sectors [%]	44.05
L2: D Sectors Fill Device [%]	41.63
L2: T Sectors [%]	41.32
L2: T Tag Requests [%]	31.10
L2: Xbar2lts Cycles Active [%]	29.26
L2: Lts2xbar Cycles Active [%]	20.83
L1: M L1tex2xbar Req Cycles Active [%]	19.49
L1: Data Pipe Lsu Wavefronts [%]	16.75
L1: M Xbar2l1tex Read Sectors [%]	14.67
L1: Lsuin Requests [%]	14.67
L1: Lsu Writeback Active [%]	11.00
L1: Data Bank Reads [%]	5.50
L1: Data Bank Writes [%]	5.50
L1: F Wavefronts [%]	0.00
L1: Texin Sm2tex Req Cycles Active [%]	0.00
L2: D Atomic Input Cycles Active [%]	0
L2: D Sectors Fill Sysmem [%]	0
L1: Tex Writeback Active [%]	0
L1: Data Pipe Tex Wavefronts [%]	0



I can clearly see 14-1. Compute util here

(9-1. FMA + 3-1. ALU)

▼ Compute Workload Analysis

Detailed analysis of the compute resources of the streaming multiprocessors (SM), including the achieved instructions per clock (IPC) ar

Executed Ipc Elapsed [inst/cycle]	0.59
Executed Ipc Active [inst/cycle]	0.59
Issued Ipc Active [inst/cycle]	0.59

Maximal IPC is 4 (1 per warp scheduler)

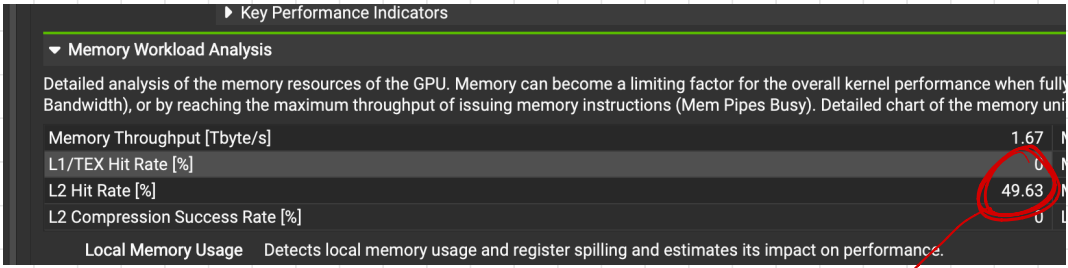
⚠ Low Utilization All compute pipelines are under-utilized. Either this kernel is very small or it doesn't issue enough warps per schec

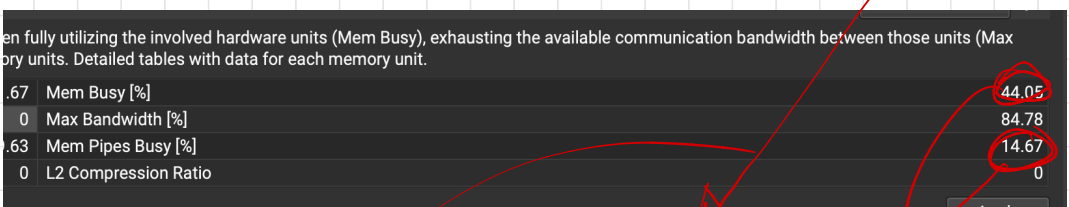
All

) and the utilization of each available pipeline. Pipelines with very high utilization might limit the overall performance.

9	SM Busy [%]	14.77
9	Issue Slots Busy [%]	14.77
9		

heduler. Check the ▷ Launch Statistics and ▷ Scheduler Statistics sections for further details.





Data only used 4x in vector add so L1 = 0%

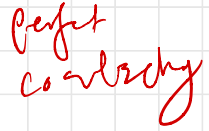
LD/ST has finite amt of requests they can put into the q

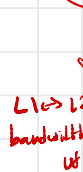
overall mem circuitry not just DRAM

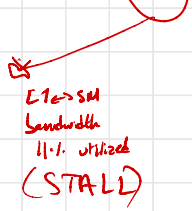
L2 likely prefetching which L1 does NOT

L1/TEX Cache

	Instructions	Requests	Wavefronts	% Peak	Sectors	Sectors/Req	Hit Rate	Bytes	Sector Misses to L2	% Peak to L2	Returns to SM	% Peak to SM
Local Load	0	0	0	0	0	0	0	0			6,291,456	11.00
Global Load	4,194,304	4,194,304			16,777,216	4	0	536,870,912	16,777,216	14.67		
Global Load To Shared Store (access)	0	0	4,194,304	7.34	0	0		0			-	-
Global Load To Shared Store (bypass)	0	0			0	0	-	0	0	0	-	-
Surface Load	0	0	0	0	0	0	0	0	0	0	0	0
Texture Load	0	0	0	0	0	0	0	0	0	0		
Global Store	2,097,152	2,097,152	2,097,152	3.67	8,388,608	4	0	268,435,456	8,388,608	14.67	-	-
Local Store	0	0	0	0	0	0	0	0			-	-
Surface Store	0	0	0	0	0	0	0	0	0	0	-	-
Global Reduction	0	0	0	0	0	0	0	0	0	0	-	-
Surface Reduction	0	0	0	0	0	0	0	0	0	0	-	-
Global Atomic ALU	0	0	0	0	0	0	0	0	0	0	see above	see above
Global Atomic CAS	0											
Surface Atomic ALU	0	0	0	0	0	0	0	0	0	0	see above	see above
Surface Atomic CAS	0											
Loads	4,194,304	4,194,304	4,194,304	7.34	16,777,216	4	0	536,870,912	16,777,216	14.67	6,291,456	11.00
Stores	2,097,152	2,097,152	2,097,152	3.67	8,388,608	4	0	268,435,456	8,388,608	14.67	-	-
Atomics & Reductions	0	0	0	0	0	0	0	0	0	0	-	-
Total	6,291,456	6,291,456	6,291,456	11.00	25,165,824	4	0	805,306,368	25,165,824	29.34	6,291,456	11.00







L2 Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes	Throughput	Sector Misses to Device	Sector Misses to System	Sector Misses to Peer
L1/TEX Load	4,194,304	16,777,216	4	20.82	0	536,870,912	1,126,668,188,838.90	16,777,216	0	0
L1/TEX Store	2,097,152	8,388,608	4	13.88	100	268,435,456	563,334,094,419.45	0	0	0
L1/TEX Atomic ALU	0	0	0	0	0	0	0	0	0	0
L1/TEX Atomic CAS	0	0	0	0	0	0	0	0	0	0
L1/TEX Reduction	0	0	0	0	0	0	0	0	0	0
L1/TEX Total	6,291,456	25,165,824	4	31.23	33.33	805,306,368	1,690,002,283,258.34	16,777,216	0	0
GCC Total	0	0	0	0	0	0	0	0	0	0
ECC Total	-	3	-	0.00	-	96	201,463.97	3	-	-
L2 Fabric Total	6,229,877	24,777,236	3.98	61.49	66.13	792,871,552	1,663,906,789,335.84	8,401,984	0	0
GPU Total	12,529,594	49,955,691	3.99	41.32	49.63	1,598,582,112	3,354,757,303,068.97	25,161,179	0	0

L2 Cache Eviction Policies

	First	Hit Rate	Last	Hit Rate	Normal	Hit Rate	Normal Demote	Hit Rate
L1/TEX Load	0	0	0	0	16,777,216	0	0	0
L1/TEX Store	0	0	0	0	8,388,608	100	0	0
L1/TEX Atomic	0	0	0	0	0	0	-	-
L1/TEX Total	0	0	0	0	25,165,824	33.33	0	0
L2 Fabric Total	0	0	0	0	12,585,136	33.35	0	0
GPU Total	2,573	99.73	0	0	37,740,305	33.35	0	0

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	16,777,228	57.22	536,871,296	1,126,668,994,694.78
Store	8,083,400	27.57	258,668,800	542,837,955,812.24
Total	24,860,628	84.78	795,540,096	1,669,506,950,507.02

SM issues Store
b/c L2 is write-back,
SM sees it as success

L2 much faster

loads more exp than stores
(plus vector add has 2x loads than stores)

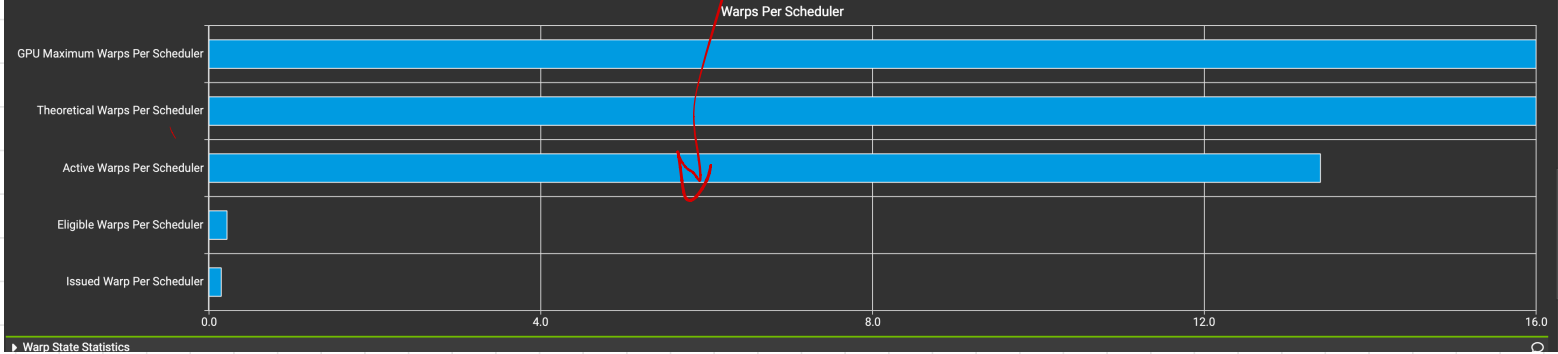
LOL
just waiting

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]	13.40	No Eligible [%]	85.26
Eligible Warps Per Scheduler [warp]	0.22	One or More Eligible [%]	[illegible]
Issued Warp Per Scheduler	0.15		

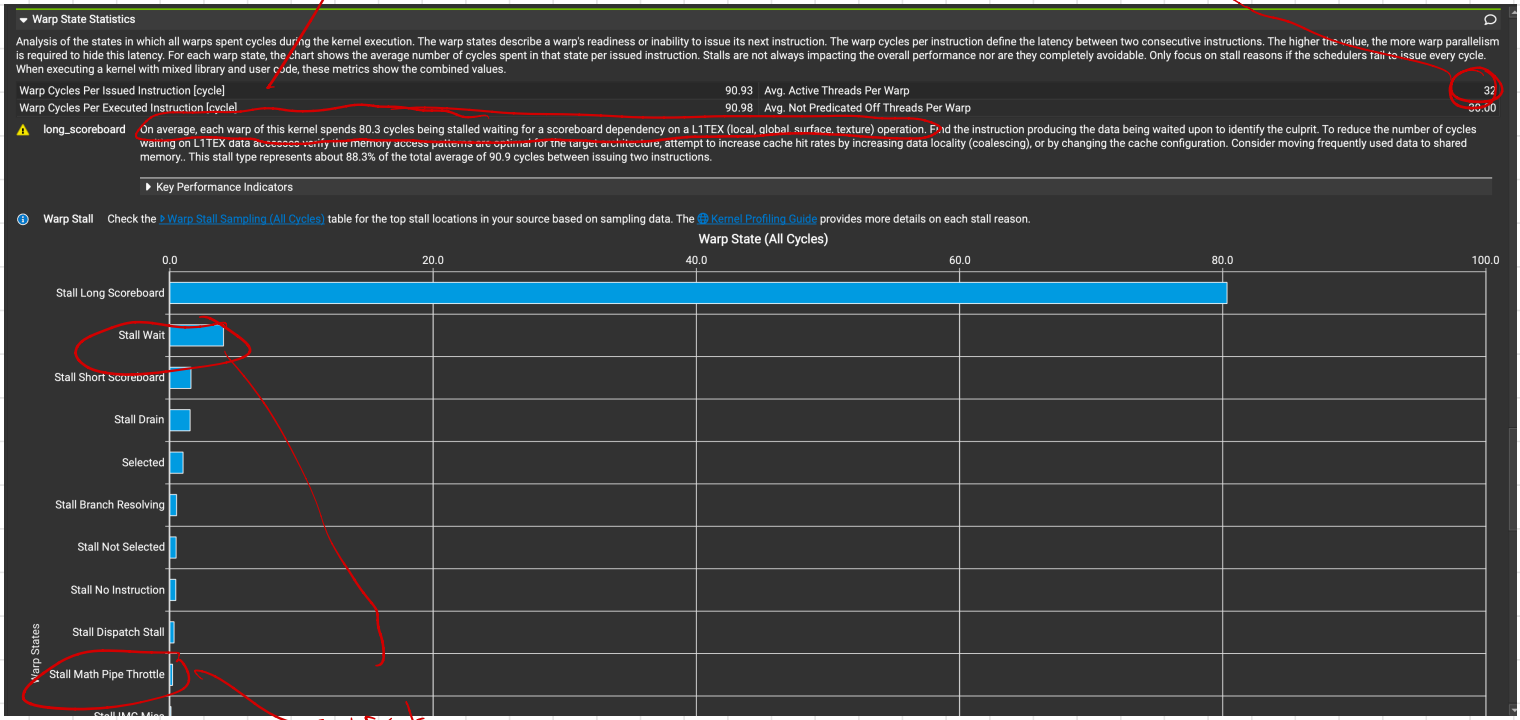
⚠ **Issue Slot Utilization** Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 6.8 cycles. This might leave hardware resources underutilized and may lead to less optimal performance. Out of the maximum of 16 warps per scheduler, this kernel allocates an average of 13.40 active warps per scheduler, but only an average of 0.22 warps were eligible per cycle. Eligible warps are the subset of active warps that are ready to issue their next instruction. Every cycle with no eligible warp results in no instruction being issued and the issue slot remains unused. To increase the number of eligible warps, avoid possible load imbalances due to highly different execution durations per warp. Reducing stalls indicated on the Warp State Statistics and Source Counters sections can help, too.



Warp State Statistics

just waiting for data to come in from memory

no divergence



Not compute bound.