Compute Throughput	Memory Throughput
70.60	84.18

62.9 for naive

94.4 for naive

Compute (SM) Throughput [%]	70.60
Memory Throughput [%]	84.18
L1/TEX Cache Throughput [%]	88.80
L2 Cache Throughput [%]	8.95
DRAM Throughput [%]	0.78

slightly less stringent as 21

Compute Throughput Breakdown	Sp
SM: Inst Executed Pipe Lsu [%]	70.60
SM: Mio2rf Writeback Active [%]	40.13
SM: Issue Active [%]	39.28
SM: Inst Executed [%]	39.25
SM: Mio Inst Issued [%]	36.95
SM: Pipe Fma Cycles Active [%]	26.54
SM: Pipe Alu Cycles Active [%]	12.43
SM: Inst Executed Pipe Adu [%]	6.60
SM: Mio Pq Read Cycles Active [%]	3.25
SM: Mio Pq Write Cycles Active [%]	3.25
SM: Pipe Tensor Cycles Active [%]	0.86
SM: Pipe Shared Cycles Active [%]	0.86
SM: Inst Executed Pipe Uniform [%]	0.85
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.05
SM: Inst Executed Pipe Ipa [%]	0
SM: Inst Executed Pipe Fp16 [%]	0
SM: Inst Executed Pipe Tex [%]	0
SM: Pipe Fp64 Cycles Active [%]	0
SM: Inst Executed Pipe Xu [%]	0
IDC: Request Cycles Active [%]	0

up from 63 in naive (due to smem optimization) (L2 we're using smem now)

up from 32% in naive (lots of register file writes/reads)

down from 31% in naive (likely due to

lack of vectorized loads

L1 hit rate

68 → 1.7 b/c of tiling

Mem Busy [%]	84.18
Max Bandwidth [%]	80.15
Mem Pipes Busy [%]	70.60
L2 Compression Ratio	0

down b/c less strain on L1

up b/c more smem + LSU units constantly working

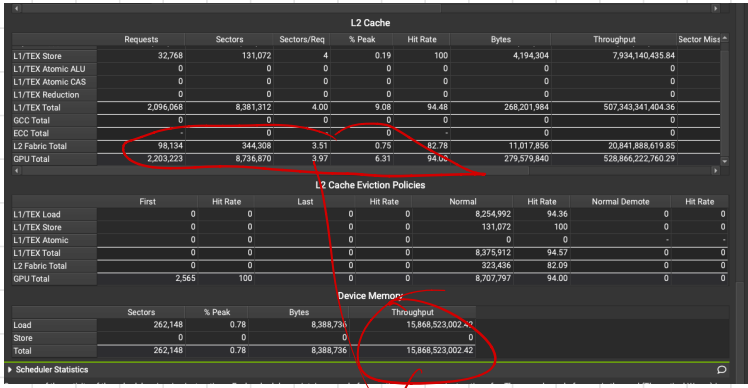
very low bank conflicts

Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	41,943,040	41,943,040	50,365,622	76.90	12,418
Shared Load Matrix	0	0			
Shared Store	2,097,152	2,097,152	2,363,830	3.61	764,385
Shared Store From Global Load	0	0	0	0	4,218
Shared Atomic	0	0	0	0	0
Other	-	-	135,167	0.21	0
Total	44,040,192	44,040,192	52,864,619	80.72	281,021

very good arithmetic intensity store 1x, load/use 20x

very good smem usage for smem → regs



L2 Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes	Throughput	Sector Miss
L1/TEX Store	32,768	131,072	4	0.19	100	4,194,304	7,934,140,435.84	
L1/TEX Atomic ALU	0	0	0	0	0	0	0	
L1/TEX Atomic CAS	0	0	0	0	0	0	0	
L1/TEX Reduction	0	0	0	0	0	0	0	
L1/TEX Total	2,096,068	8,381,312	4.00	9.08	94.48	268,201,984	507,343,341,404.36	
GCC Total	0	0	0	0	0	0	0	
ECC Total	-	0	-	0	-	0	0	
L2 Fabric Total	98,134	344,308	3.51	0.75	82.78	11,017,856	20,841,888,619.85	
GPU Total	2,203,223	8,736,870	3.97	6.31	94.05	279,579,840	528,866,222,760.29	

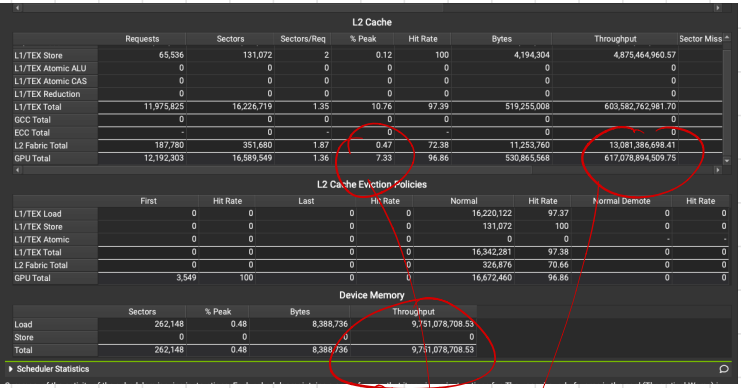
L2 Cache Eviction Policies

	First	Hit Rate	Last	Hit Rate	Normal	Hit Rate	Normal Demote	Hit Rate
L1/TEX Load	0	0	0	0	8,254,992	94.36	0	0
L1/TEX Store	0	0	0	0	131,072	100	0	0
L1/TEX Atomic	0	0	0	0	0	0	-	-
L1/TEX Total	0	0	0	0	8,375,912	94.57	0	0
L2 Fabric Total	0	0	0	0	323,436	82.09	0	0
GPU Total	2,565	100	0	0	8,707,797	94.00	0	0

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	262,148	0.78	8,388,736	15,868,523,002.42
Store	0	0	0	0
Total	262,148	0.78	8,388,736	15,868,523,002.42

Scheduler Statistics



L2 Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes	Throughput	Sector Miss
L1/TEX Store	65,536	131,072	2	0.12	100	4,194,304	4,875,464,960.57	
L1/TEX Atomic ALU	0	0	0	0	0	0	0	
L1/TEX Atomic CAS	0	0	0	0	0	0	0	
L1/TEX Reduction	0	0	0	0	0	0	0	
L1/TEX Total	11,975,825	16,226,719	1.35	10.76	97.39	519,255,008	603,582,762,981.70	
GCC Total	0	0	0	0	0	0	0	
ECC Total	-	0	-	0	-	0	0	
L2 Fabric Total	187,780	351,680	1.87	0.47	72.38	11,253,760	13,081,386,698.41	
GPU Total	12,192,303	16,589,549	1.36	7.33	96.86	530,865,568	617,078,894,509.75	

L2 Cache Eviction Policies

	First	Hit Rate	Last	Hit Rate	Normal	Hit Rate	Normal Demote	Hit Rate
L1/TEX Load	0	0	0	0	16,220,122	97.37	0	0
L1/TEX Store	0	0	0	0	131,072	100	0	0
L1/TEX Atomic	0	0	0	0	0	0	-	-
L1/TEX Total	0	0	0	0	16,342,281	97.38	0	0
L2 Fabric Total	0	0	0	0	326,876	70.66	0	0
GPU Total	3,549	100	0	0	16,672,460	96.86	0	0

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	262,148	0.48	8,388,736	9,751,078,708.53
Store	0	0	0	0
Total	262,148	0.48	8,388,736	9,751,078,708.53

Scheduler Statistics

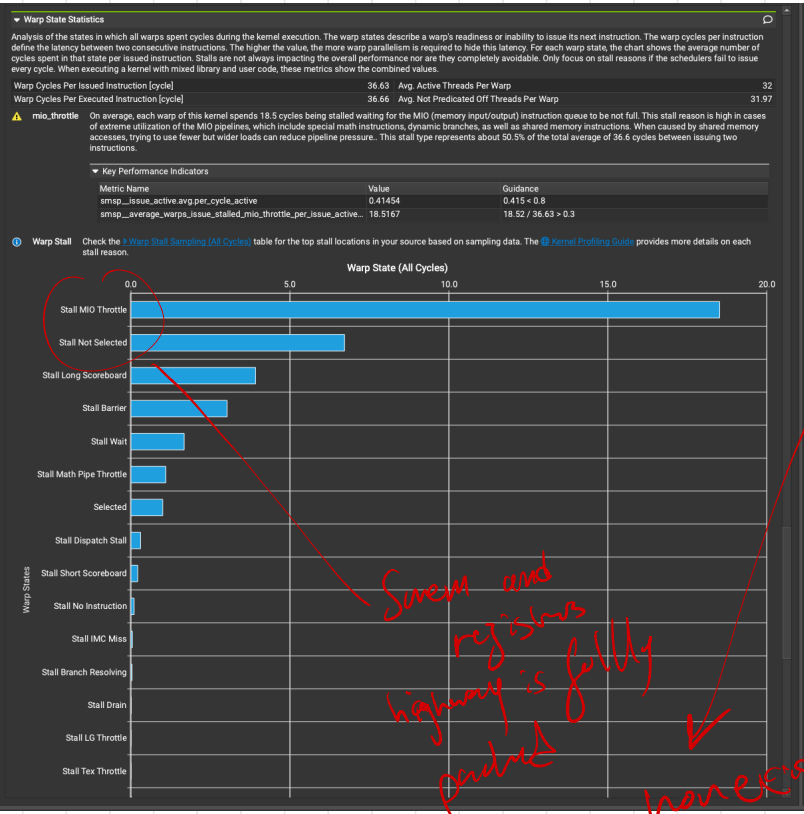
in tiled
Smem is now
fully on the L2
responsibility

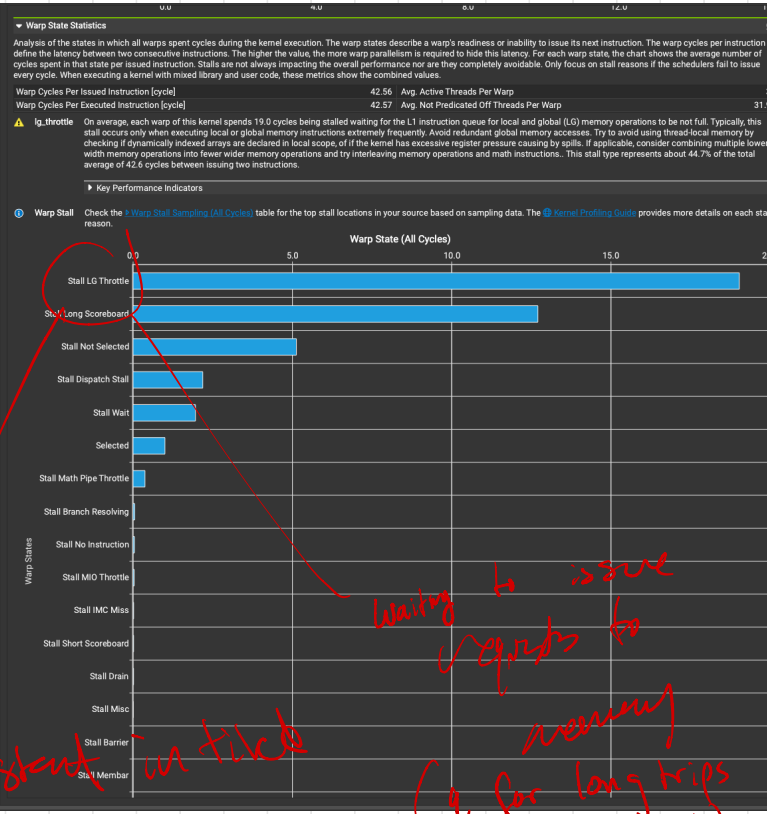
tiled is faster and scales given inputs
faster

In naive L2 saved us

No eligible ↓ from 65 naive to 57 in filled

(due to latency hiding)



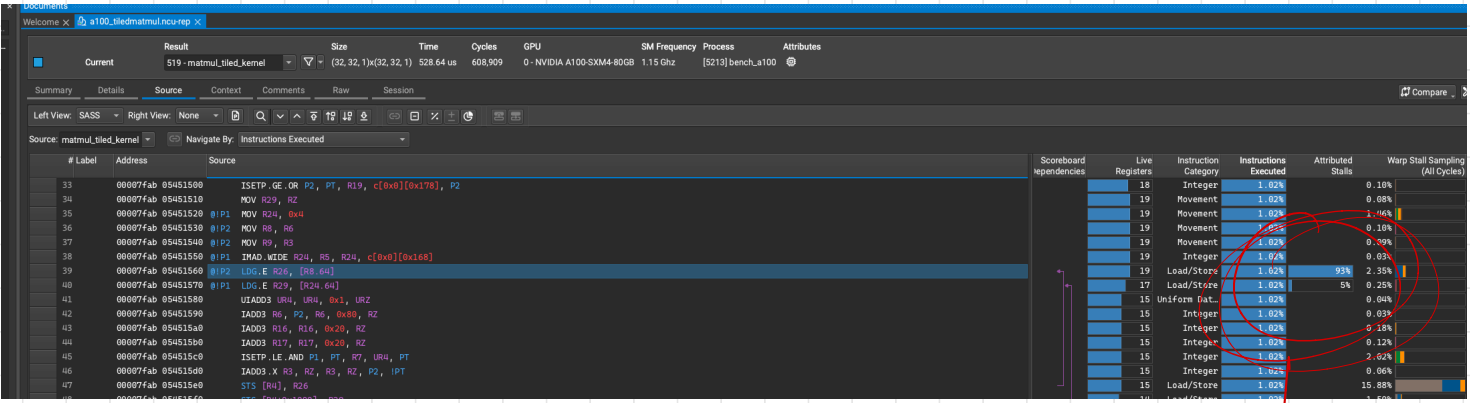


Smem and registers highway is fully packed

waiting to issue (groups to memory for long trips was full)

b/c of hiding Smem latency

Nonetheless, untied


Documents
Welcome
a100_tiledmatmul.ncu-rep
Result
Size
Time
Cycles
GPU
SM Frequency
Process
Attributes
Current
519 - matmul_tiled_kernel
(32, 32, 1)x(32, 32, 1)
528.64 us
608,909
0 - NVIDIA A100-SXM4-80GB
1.15 Ghz
[5213] bench_a100
Summary
Details
Source
Context
Comments
Raw
Session
Compare
Left View: SASS
Right View: None
Source: matmul_tiled_kernel
Navigate By: Instructions Executed
Label
Address
Source
Scoreboard Dependencies
Live Registers
Instruction Category
Instructions Executed
Attributed Stalls
Warp Stall Sampling (All Cycles)
33 00007fab 05451500 ISETP.GE.OR P2, PT, R19, c[0x0][0x178], P2 18 Integer 1.02% 0.10%
34 00007fab 05451510 MOV R29, RZ 19 Movement 1.02% 0.08%
35 00007fab 05451520 @!P1 MOV R24, 0x4 19 Movement 1.02% 1.46%
36 00007fab 05451530 @!P2 MOV R8, R6 19 Movement 1.02% 0.10%
37 00007fab 05451540 @!P2 MOV R9, R3 19 Movement 1.02% 0.09%
38 00007fab 05451550 @!P1 IMAD.WIDE R24, R5, R24, c[0x0][0x168] 19 Integer 1.02% 0.03%
39 00007fab 05451560 @!P2 LDG.E R26, [R8.64] 19 Load/Store 1.02% 93% 2.35%
40 00007fab 05451570 @!P1 LDG.E R29, [R24.64] 17 Load/Store 1.02% 5% 0.25%
41 00007fab 05451580 UIADD3 UR4, UR4, 0x1, URZ 15 Uniform Dat… 1.02% 0.04%
42 00007fab 05451590 IADD3 R6, P2, R6, 0x80, RZ 15 Integer 1.02% 0.03%
43 00007fab 054515a0 IADD3 R16, R16, 0x20, RZ 15 Integer 1.02% 0.18%
44 00007fab 054515b0 IADD3 R17, R17, 0x20, RZ 15 Integer 1.02% 0.12%
45 00007fab 054515c0 ISETP.LE.AND P1, PT, R7, UR4, PT 15 Integer 1.02% 2.62%
46 00007fab 054515d0 IADD3.X R3, RZ, R3, RZ, P2, !PT 15 Integer 1.02% 0.06%
47 00007fab 054515e0 STS [R4], R26 15 Load/Store 1.02% 15.88%

Jeez