Jaidon Lybbert
Aut. 22 GPU
Homework 4
11/16/2022

NSight Profiling Analysis

Summary Analysis

The global memory kernel sent no requests to shared memory, while the shared memory and corner turning kernels did. This is as expected. Despite not using shared memory, the global kernel outperformed the other two kernels by making good use of the L1/TEX cache, which saw as high as a 94% hit rate with the global kernel. The other two kernels saw a nearly 0% hit rate in the L1 cache, as they sent requests to shared memory instead.

The corner turning kernel performed the slowest, in all cases, with the lowest AI and FLOPs. This is entirely due to a high number of bank conflicts. The transpose operation on the second matrix operand necessitated memory accesses in the corner turning kernel which were column aligned with respect to successive threads in each thread block in order to achieve the same output as the other kernels given the same original matrices. The source of the bank conflicts is most easily recognized with 32x32 thread blocks, where successive threads in the same warp land exactly on the next 4 bytes stored in the same memory bank on my compute capability 8.6 device. This causes each warp to have 31 bank conflicts, and serializes the accesses, for a total of 31x4x128x128=2,031,616 bank conflicts.

I believe the corner turning kernel I implemented is counter-productive because the input matrices I started with were not stored in a column major format to begin with. It would make sense to use this method had I transposed a column-major matrix into row-major, and then applied the shared memory kernel, instead of transposing a row-major matrix into a column-major matrix, then written a new kernel to make the output match the output of my other kernels without the transpose. What I accomplished here illustrates why it's undesirable to do the matrix multiply operation on a matrix stored in a column-major way.

I saw the same memory patterns when using the 16x16 thread blocks, although the corner turning matrix performed better than the worst case, with the new alignment of threads with memory banks.

L2 cache for all kernels saw an order of magnitude fewer requests than L1 or shared memory. And device memory further served an order of magnitude fewer sectors than L2 cache. The hit rate remained high for L2 cache, and the pattern seems to be that the 'misses' for cache memory are propagated to the next higher memory. Since the hit rate is ~90% for L1 and L2 cache, it makes sense that the next higher level of memory sees an order of magnitude fewer (or 10% of) the requests seen by the current level of memory.

Shared Memory Data

32x32 CTA

Kernel	Requests	Wavefronts	% Peak	Bank Conflicts
Global Mem.	0	0	0	0
Shared Mem.	135,168	136,320	12.77	0
Corner Turning	135,168	2,162,688	33.27	2,031,616

16x16 CTA

Kernel	Requests	Wavefronts	% Peak	Bank Conflicts
Global Mem.	0	1,024	0.16	0
Shared Mem.	139,264	141,312	20.35	0
Corner Turning	139,264	600,064	48.74	458,752

8x32 CTA

Kernel	Requests	Wavefronts	% Peak	Bank Conflicts
Global Mem.	0	1,024	0.16	0

32x8 CTA

Kernel	Requests	Wavefronts	% Peak	Bank Conflicts
Global Mem.	0	1,024	0.15	0

L1/TEX Data

32x32 CTA

Kernel	Requests	Wavefronts	Sectors	Sectors/Req	% Hit Rate	Bytes
Global Mem.	131,584	131,584	329,728	2.51	94.41	10,551,296
Shared Mem.	4,608	4,608	18,432	4	0	589,824
Corner Turning	4,608	4,608	18,432	4	0	589,824

16x16 CTA

Kernel	Requests	Wavefronts	Sectors	Sectors/Req	% Hit Rate	Bytes
Global Mem.	131,584	131,584	264,192	2.01	86.82	8,454,144
Shared Mem.	8,704	8,704	34,816	4	0	1,114,112
Corner Turning	8,704	8,704	34,816	4	0	1,114,112

8x32 CTA

Kernel	Requests	Wavefronts	Sectors	Sectors/Req	% Hit Rate	Bytes
Global Mem.	131,584	131,584	329,728	2.51	86.96	10,551,296

32x8 CTA

Kernel	Requests	Wavefronts	Sectors	Sectors/Req	% Hit Rate	Bytes
Global Mem.	131,584	131,584	329,728	2.51	86.96	10,551,296

L2 Cache Data

32x32 CTA

Kernel	Requests	Sectors	Sectors/Req	% Hit Rate	Bytes
Global Mem.	11,610	20,285	1.75	71.22	649,120
Shared Mem.	5,874	23,849	4.06	78.90	763,168
Corner Turning	10,174	39,917	3.92	38.16	1,277,344

16x16 CTA

Kernel	Requests	Sectors	Sectors/Req	% Hit Rate	Bytes
Global Mem.	26,462	35,949	1.36	85.20	1,150,368
Shared Mem.	17,408	34,816	2	88.24	1,114,112
Corner Turning	18,270	37,821	2.07	86.64	1,210,272

8x32 CTA

Kernel	Requests	Sectors	Sectors/Req	% Hit Rate	Bytes
Global Mem.	17,986	45,053	2.50	84.88	1,441,696

32x8 CTA

Kernel	Requests	Sectors	Sectors/Req	% Hit Rate	Bytes
Global Mem.	43,886	47,981	1.09	90.66	1,535,392

Device Memory Data

32x32 CTA

Kernel	Sectors	% Peak	Bytes	Throughput (GB/s)
Global Mem.	6,820	3.54	218,240	15.189
Shared Mem.	8,040	3.83	257,280	16.577
Corner Turning	15,476	1.21	495,232	5.246

16x16 CTA

Kernel	Sectors	% Peak	Bytes	Throughput (GB/s)
Global Mem.	8,636	6.77	276,352	29.275
Shared Mem.	4,100	2.99	131,200	12.773
Corner Turning	4,100	1.69	131,200	7.270

8x32 CTA

Kernel	Sectors	% Peak	Bytes	Throughput (GB/s)
Global Mem.	5,716	4.42	182,912	18.439

32x8 CTA

Kernel	Sectors	% Peak	Bytes	Throughput (GB/s)
Global Mem.	4,132	3.01	132,224	13.286