GPUs Hands On 3

May 15, 2022

J L Gouws

19G4436

Block Size	Kernel Execution Time(ms)	Achieved occupancy(%)	Global Memory Load Throughput(GB/S)	Global Memory Load Efficiency (%)	Global Memory Store Throughput (GB/S)	Global Memory Store Efficiency (%)
			2D Grid			
64 × 8	2.66404	87.8	50.334	100	25.167	100
2 × 64	5.88941	87.2	93	100	46.5	100
			1D Grid			
16 × 16	2.67481	89.3	50.028	100	25.014	100
32 × 32	2.76861	83.9	49.929	100	24.964	100
			1D Grid, 16 Data per thread, unstrided			
2 × 64	7.04664	72.1	156.709	12.5	78.354	12.5
64 × 16	7.41846	91.4	152.988	12.5	76.494	12.5
			1D Grid, 16 Data per thread, unstrided			
32 × 32	2.79042	95.2	48.149	100	24.074	100
16 × 64	2.83848	94.9	48.675	100	24.337	100

Observations:
The table shows that the memory throughput metrics can be misleading. The worst performing kernels have very high memory throughput. This throughput is not always productive, as shown by the memory efficiency metrics.

Achieved occupancy can also be misleading. The, worst performing, unstrided 16 data per thread kernel achieves 91.4% occupancy, yet it has an execution time of ~ 7.4 ms. These considerations of occupancy and throughput show that one particular metric cannot give the full picture. Multiple metrics should be considered to understand the performance of the kernel.

Doing sixteen consecutive data items per thread results in low memory load and store efficiencies. Striding solves this problem and achieves the same memory efficiency as the single datum per thread approach.