

ID •	Estimated Speedup	Function Name	Demangled Name	Duration (8.2008e+07)	Runtime Improvement (4.26926e+07)	Compute Throughput	Memory Throughput	# Registers	Grid Size
	52.47	checkSortedRowWi	checkSortedRowWi	0.05 (+81.99%)	0.02	5.05 (-78.52%)	8.20 (-87.30%)	16 (+0.00%)	16, 512,
	1 65.18	checkSortedColum	checkSortedColum	0.03 (+28.97%)	0.02	6.11 (-73.98%)	6.11 (-90.54%)	16 (+0.00%)	16, 512,
	2 23.28	sortRowsKernelInt	sortRowsKernelInt(22.63 (+88,966.75%)	5.27	89.48 (+280.87%)	89.48 (+38.55%)	18 (+12.50%)	512, 1,
	89.13	transposeKernelInt	transposeKernelInt(0.05 (+111.96%)	0.05	11.49 (-51.10%)	35.52 (-45.01%)	16 (+0.00%)	16, 16,
	4 52.18	checkSortedRowWi	checkSortedRowWi	0.08 (+198.99%)	0.04	7.15 (-69.55%)	8.15 (-87.38%)	16 (+0.00%)	16, 512,
	65.06	checkSortedColum	checkSortedColum	0.03 (+29.35%)	0.02	6.12 (-73.96%)	6.12 (-90.53%)	16 (+0.00%)	16, 512,
	6 28.37	sortRowsKernelInt	sortRowsKernelInt(22.25 (+87,467.76%)	6.31	91.03 (+287.46%)	91.03 (+40.95%)	18 (+12.50%)	512, 1,
	7 86.92	transposeKernelInt	transposeKernelInt(0.06 (+121.41%)	0.05	11.42 (-51.38%)	35.31 (-45.33%)	16 (+0.00%)	16, 16,
	87.19	checkSortedRowWi	checkSortedRowWi	32.41 (+127,439.17	28.25	3.12 (-86.73%)	49.82 (-22.87%)	16 (+0.00%)	16, 512,
	59.99	checkSortedColum	checkSortedColum	4.43 (+17,318.14%)	2.66	23.11 (-1.63%)	40.01 (-38.05%)	16 (+0.00%)	16, 512,

Est. Speedup: 89.13%

locations. The @ CUDA Best Practices Gui has an example on optimizing shared memory accesses.

The memory access pattern for shared loads might not be optimal and causes on average a 32.0 - way bank conflict across all 8192 shared load requests. This results in 253952 bank conflicts, which represent 96.88% of the overall 262144 wavefronts for shared loads. Check the Source Counters section for uncoalesced shared loads.

0

0

Est. Speedup: 10.84%

Est. Speedup: 68.81%

The difference between calculated theoretical (100.0%) and measured achieved occupancy (89.2%) can be the result of warp scheduling overheads or workload imbalances during the kernel execution. Load imbalances can occur between warps within a block as well as across blocks of the same kernel. See the @ CUDA Best Practices Guid for more details on optimizing occupancy.