

The following performance optimization opportunities were discovered for this result. Follow the rule links to see more context on the Details page.

Note: Speedup estimates provide upper bounds for the optimization potential of a kernel assuming its overall algorithmic structure is kept unchanged.

L1TEX Global Load Access Pattern Est. Speedup: 62.51% The memory access pattern for global loads from L1TEX might not be optimal. On average, only 11.5 of the 32 bytes transmitted per sector are utilized by each thread. This could possibly be caused by a stride between threads. Check the Source Counters section for uncoalesced global loads.

 $\odot$ 

Uncoalesced Global Accesse Est. Speedup: 58.57% This kernel has uncoalesced global accesses resulting in a total of 2883680 excessive sectors (61% of the total 4755920 sectors). Check the L2 Theoretical Sectors Global Excessive table for the primary source locations. The CUDA Programming Guide has additional information on reducing uncoalesced device memory accesses.



## Lg Throttle Stall

Est. Speedup: 51.36%

On average, each warp of this workload spends 14.6 cycles being stalled waiting for the L1 instruction queue for local and global (LG) memory operations to be not full. Typically, this stall occurs only when executing local or global memory instructions extremely frequently. Avoid redundant global memory accesses. Try to avoid using thread-local memory by checking if dynamically indexed arrays are declared in local scope, of if the kernel has excessive register pressure causing by spills. If applicable, consider combining multiple lower-width memory operations into fewer wider memory operations and try interleaving memory operations and math instructions. This stall type represents about 57.0% of the total average of 25.5 cycles between issuing two instructions.

