SAND2018-3489C

Methods for Computing Monte Carlo Tallies on the GPU

Kerry L. Bossler, Sandia National Laboratories, Albuquerque, NM, USA

Introduction

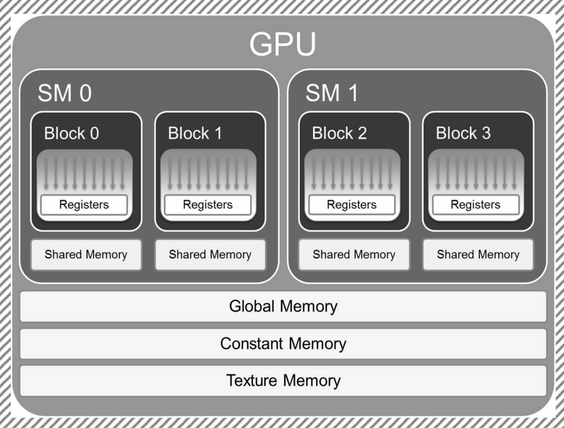
Tally Replication vs. **Atomic Operations**
More Memory Serializes Code

What about NVIDIA's warp shuffle feature?
- minimize memory usage AND limit atomic operations

GOAL: Compare performance of tally based on NVIDIA's warp shuffle feature to tally replication and atomic operations.

GPU Architecture

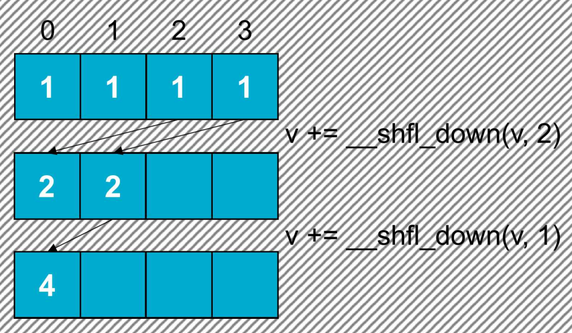
Single-Instruction, Multiple-Thread (SIMT) Model



- Break work down into multiple thread blocks
- Distribute blocks to GPU's streaming multiprocessors
- Execute groups of 32 threads (a.k.a. *warp*)
- Data can exist in many different memory spaces

Parallel Reduction Using Warp Shuffle†

- Introduced with compute capability 3.x
- Enables data exchange between threads in warp
- Thread 0 in warp has access to final sum



† J. Luitjens, https://devblogs.nvidia.com/parallelforall/faster-parallel-reductions-kepler

Tally Methods

Summary of Tally Methods Tested

Method Name	Advantage	Disadvantage	Atomic Updates†
Global Atomics	Larger tallies	Slower atomics	128 Global
Shared Atomics	Faster atomics	Smaller tallies	128 Shared; 1 Global
Warp Shuffle‡	Larger tallies; Limits atomics	One atomic update per warp	4 Global
Block Reduction‡	Larger tallies; Limits atomics	Needs thread synchronization	1 Global
Tally Replication	Eliminates Atomics	Needs more memory	N/A

† Number of atomic operations assuming 128 threads per block
‡ Method uses NVIDIA's warp shuffle feature

Performance Tests

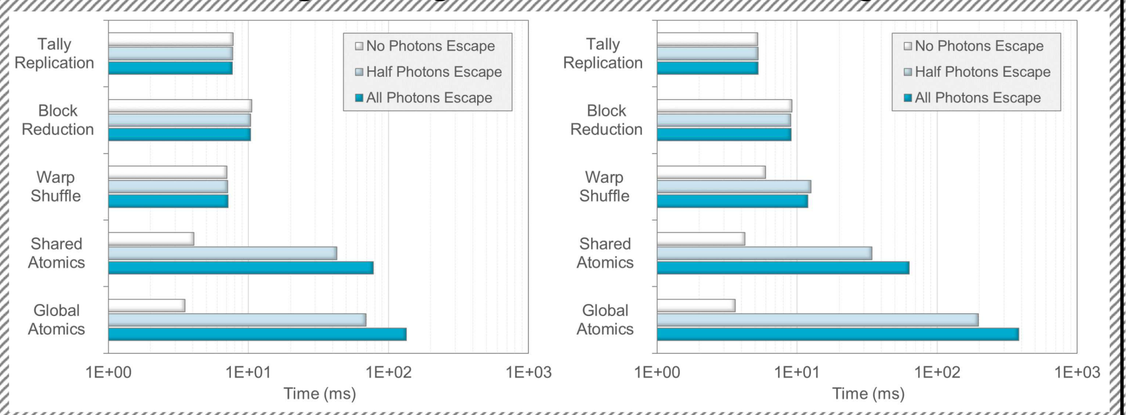
- Photon attenuation with escape tally for 10^8 particle histories
- Considered single and double precision for tallies based on integers and floating-point values
- All timing data is an average of ten independent runs
- Repeated tests on Quadro K5200, Tesla K40, K80, and P100

Results

Timing Data for Quadro K5200

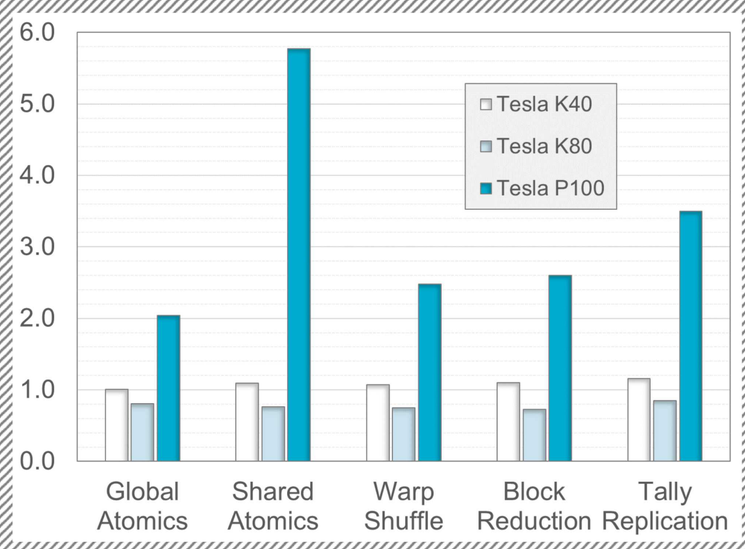
- Best option depends on size, type, and update frequency
- More photons escaping means a higher update frequency

64-bit Unsigned Integers **32-bit Floating-Point Values**



Speedup over Quadro K5200 for 10^8 32-bit Integer Updates

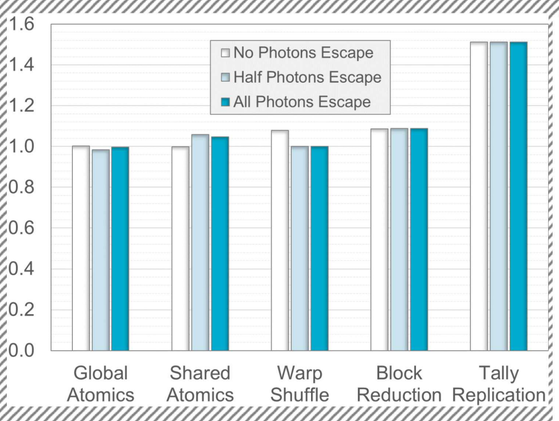
- Tesla P-100 is generally 2-6 times faster than other GPUs
- Similar behavior is seen for most data types and scenarios



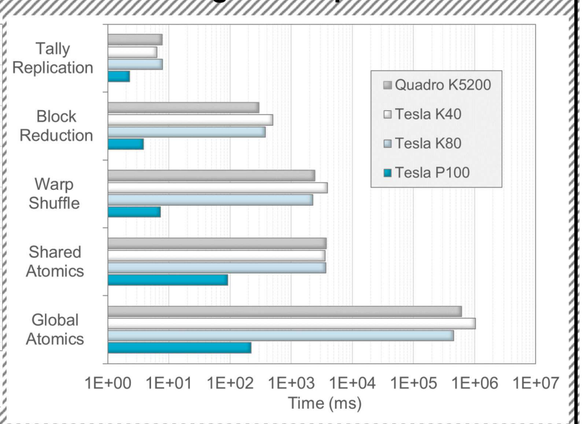
Single vs. Double Precision for Floating-Point Values

- Tesla P100 only GPU tested with native support for 64-bit floating-point atomic operations
- Penalty for double precision atomics on Tesla P100 is low

Speedup of Single Precision over Double Precision on Tesla P100



Timing Data for 10^8 64-bit Floating-Point Updates



Conclusions

Use Method…	When…
Global Atomics	*Large* tally with *low* update frequency
Shared Atomics	*Small* tally with *low* update frequency
Warp Shuffle	*Large* tally with *high* update frequency (*integers*)
Block Reduction	*Large* tally with *high* update frequency (*floating-point*)
Tally Replication	*Small* tally with *high* update frequency

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration
Supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND No. _____
LDRD Laboratory Directed Research and Development
Sandia National Laboratories