SAND2018-3123C

Methods for Computing Monte Carlo Tallies on the GPU

PRESENTED BY

Kerry L. Bossler

LDRD
Laboratory Directed Research and Development

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

INTRODUCTION

- All variants of Monte Carlo particle transport codes need to frequently update a variety of different tallies

Is there a better alternative for tallying on the GPU?

- Updating tallies on the GPU can be more complicated
  - Best approach depends on multiple factors

Warp Shuffle!

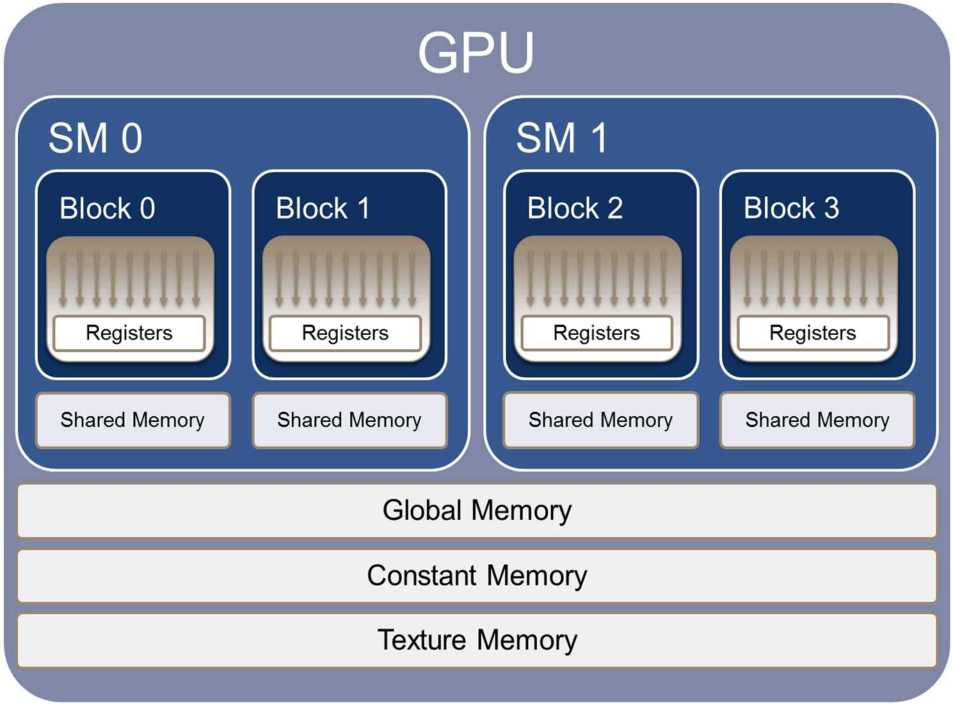
- Two general approaches are used for tallying on the GPU
  - Replicate the tallies across one or more GPU threads **OR**
  - Relying on atomic operations that serialize the code

NVIDIA GPU ARCHITECTURE

NVIDIA GPU architecture uses Single-Instruction, Multiple-Thread (SIMT) technology

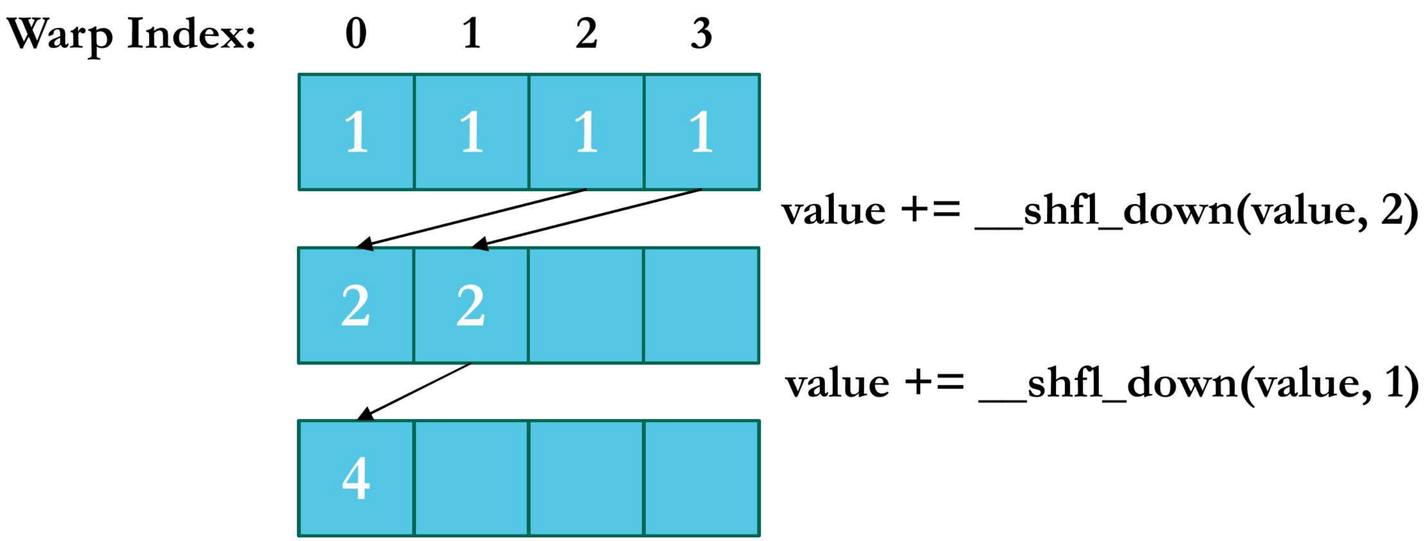
Parallel work initiated by launching CUDA kernel

- Break work down into many thread blocks
- Blocks distributed to streaming multiprocessors (SMs)
- Each SM executes 32 threads concurrently (a.k.a. warp)
- Data can exist in many different memory spaces



WARP SHUFFLE FEATURE

- Introduced for GPUs with compute capability 3.x or higher
- Allows all 32 threads in a warp to simultaneously exchange or broadcast data without using shared memory
- Can use warp shuffle to implement an efficient parallel reduction across the threads in a warp†



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

COMPARISON OF TALLY METHODS

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
No Atomics	Eliminates atomics	Needs more memory	-

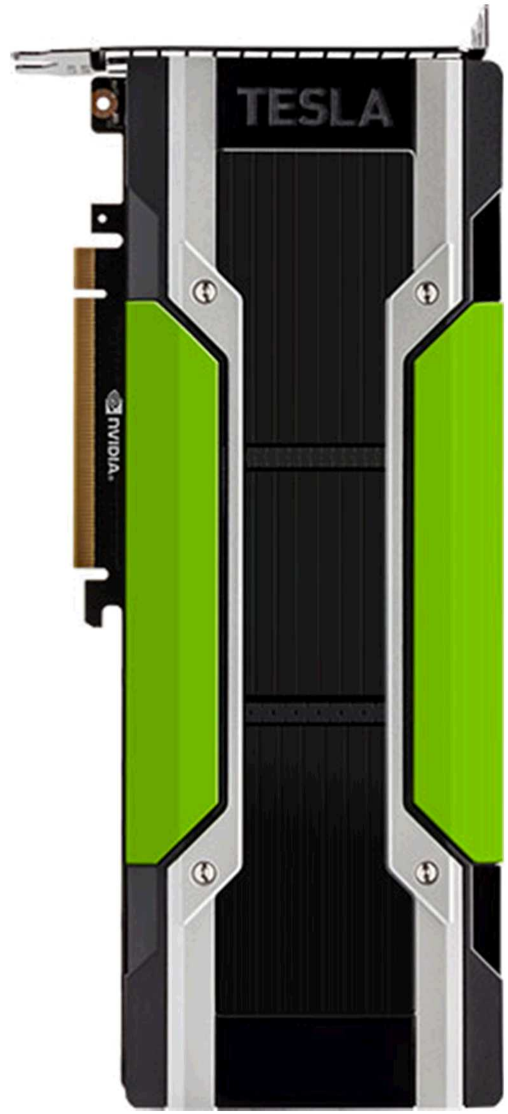
† Number of atomic operations assuming 128 threads per block

NVIDIA GPU OPTIONS

All tally methods tested on four NVIDIA GPUs

Quadro K5200

- 1 GPU per card
- 3.5 Compute Capability
- 2304 CUDA cores
- 12 SMs



Tesla K40

- 1 GPU per card
- 3.5 Compute Capability
- 2880 CUDA cores
- 15 SMs

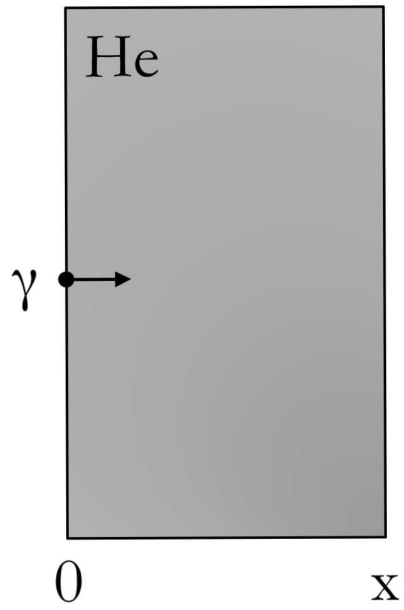
Tesla K80

- 2 GPUs per card
- 3.7 Compute Capability
- 2496 CUDA cores
- 13 SMs

Tesla P100

- 1 GPU per card
- 6.0 Compute Capability
- 3584 CUDA cores
- 56 SMs

PERFORMANCE TESTS



Fraction of γ escaped

$$\frac{N}{N_o} = e^{-6.59936E-3\,x}$$

Test scenarios considered

- All photons escape (x = 0 m)
- Approximately half of the photons escape (x = 100 m)
- No photons escape (x = 10,000 m)

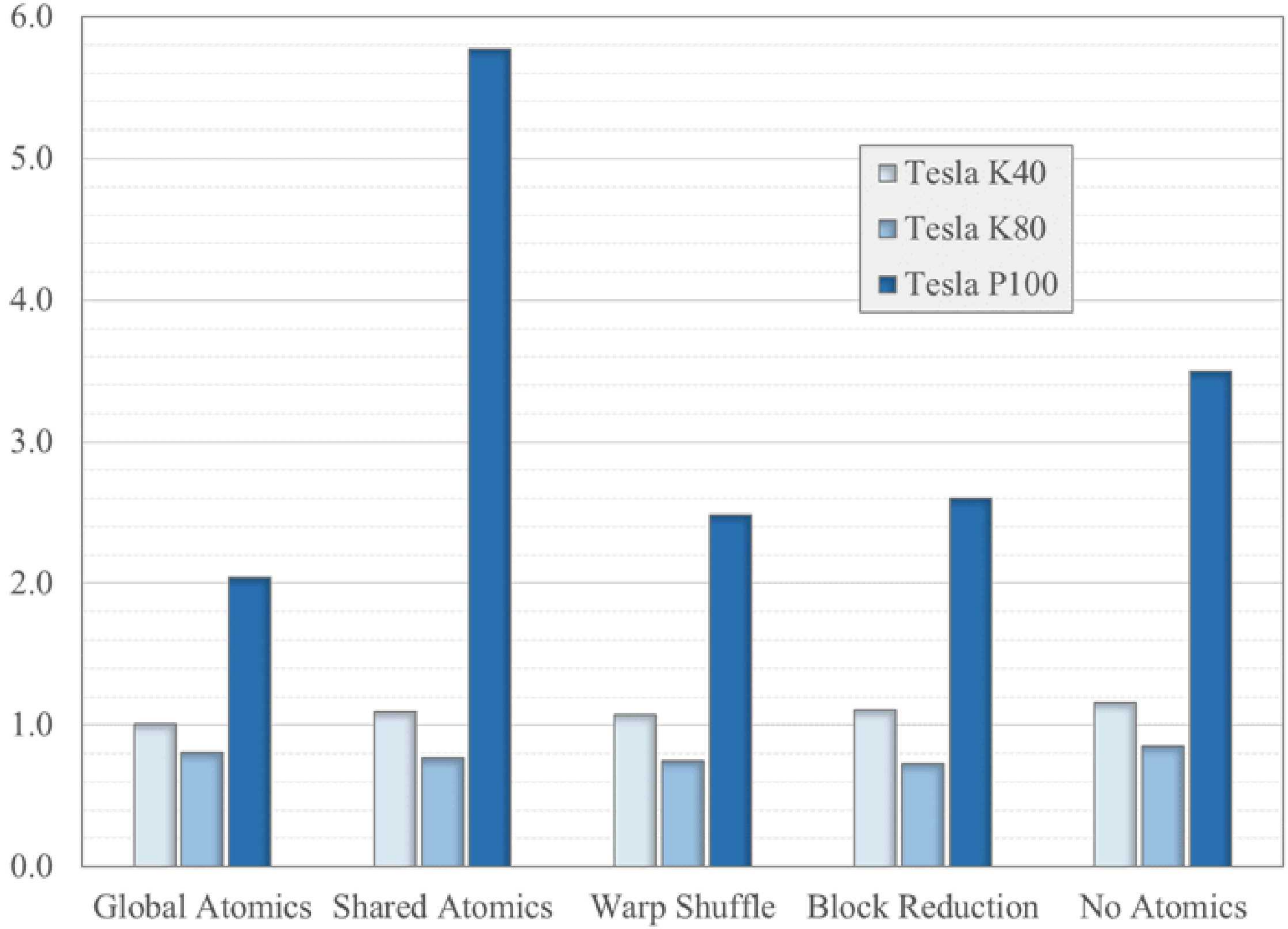
RESULTS: OVERVIEW

- Each test scenario was run with
  - 10^8 particle histories
  - 128 threads per block
- All timing data is an average of ten independent runs
  - Measured contribution of tally updates
- Considered multiple data types
  - 32-bit integers
  - 64-bit unsigned integers
  - 32-bit floating-point type
  - 64-bit floating-point type (Tesla P100 only)

RESULTS: QUADRO K5200

Test Scenario	Global Atomics (ms)	Shared Atomics (ms)	Warp Shuffle (ms)	Block Reduction (ms)	No Atomics (ms)
INTEGER TYPE (32-bit)					
1	5.48 (1.3)	7.57 (0.5)	6.64 (0.5)	9.34 (0.6)	5.26 (0.2)
2	71.0 (4.7)	34.6 (1.9)	6.58 (0.4)	9.30 (0.6)	5.22 (0.2)
3	3.44 (0.1)	4.05 (0.2)	6.12 (0.4)	9.04 (0.6)	5.31 (0.3)
UNSIGNED INTEGER TYPE (64-bit)					
1	134 (5.0)	78.1 (4.9)	7.15 (0.4)	10.4 (0.6)	7.70 (0.3)
2	69.2 (2.5)	42.9 (2.0)	7.13 (0.4)	10.4 (0.6)	7.73 (0.3)
3	3.53 (0.1)	4.08 (0.3)	7.01 (0.4)	10.6 (0.7)	7.78 (0.3)
FLOATING-POINT TYPE (32-bit)					
1	384 (4.0)	63.1 (3.8)	11.9 (< 1%)	9.07 (0.5)	5.27 (0.2)
2	197 (0.3)	34.3 (1.8)	12.6 (0.8)	9.05 (0.5)	5.26 (0.2)
3	3.61 (0.2)	4.23 (0.3)	5.96 (< 1%)	9.18 (0.6)	5.22 (0.2)

RESULTS: TESLA GPUS

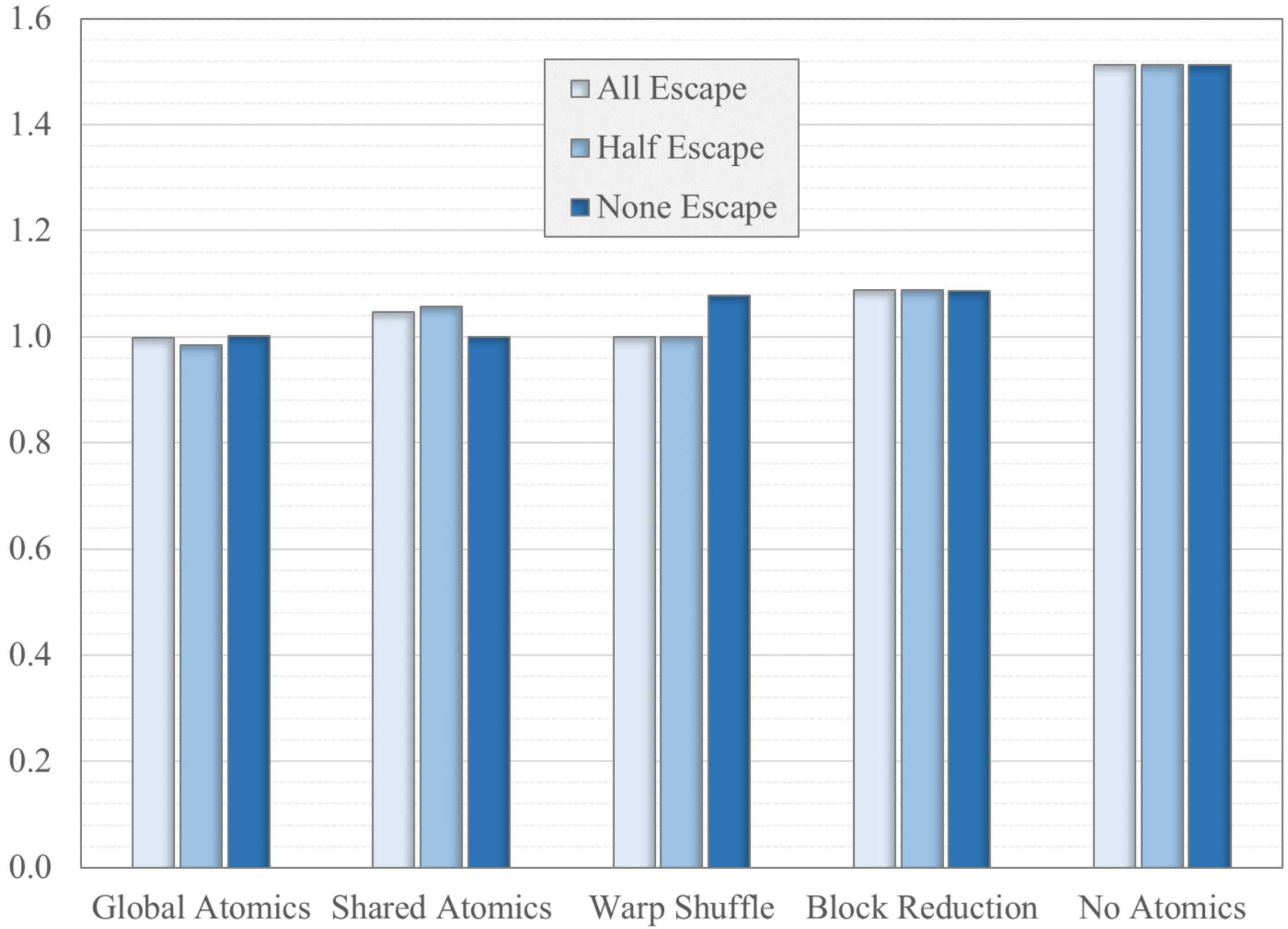


Speedup over Quadro K5200 for 10^8 tally updates using 32-bit integer type

RESULTS: TESLA P100

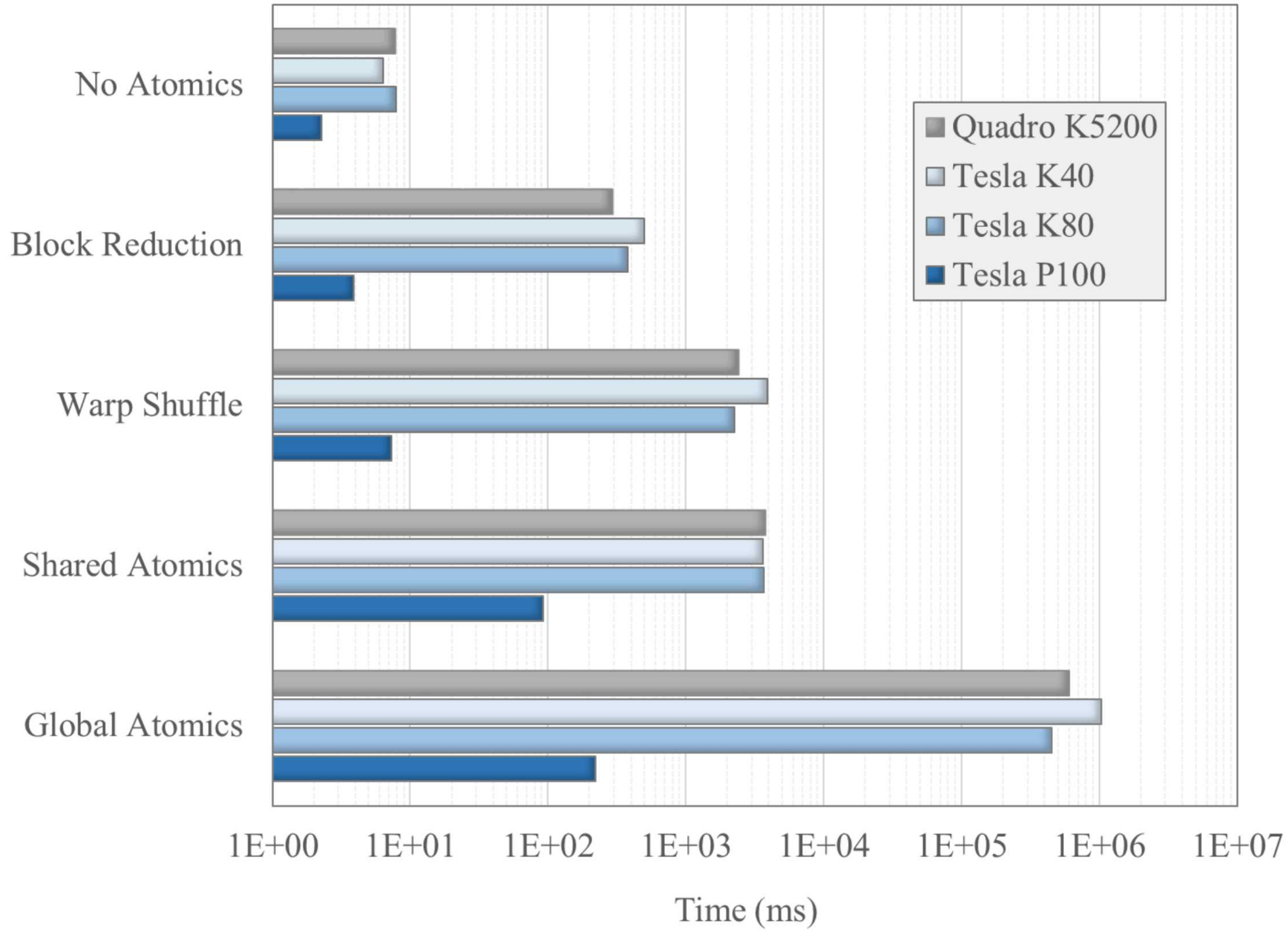
Test Scenario	Global Atomics (ms)	Shared Atomics (ms)	Warp Shuffle (ms)	Block Reduction (ms)	No Atomics (ms)
INTEGER TYPE (32-bit)					
1	2.67 (<1%)	1.31 (<1%)	2.68 (<1%)	3.59 (<1%)	1.50 (<1%)
2	2.69 (<1%)	1.31 (<1%)	2.68 (<1%)	3.59 (<1%)	1.50 (<1%)
3	1.31 (<1%)	1.31 (<1%)	2.23 (<1%)	3.54 (<1%)	1.50 (<1%)
UNSIGNED INTEGER TYPE (64-bit)					
1	77.0 (1.7)	92.6 (0.8)	2.68 (<1%)	3.92 (<1%)	2.27 (<1%)
2	40.1 (0.5)	25.3 (0.2)	2.68 (<1%)	3.92 (<1%)	2.27 (<1%)
3	1.31 (<1%)	1.31 (<1%)	2.40 (<1%)	3.90 (<1%)	2.27 (<1%)
FLOATING-POINT TYPE (32-bit)					
1	222 (6.6)	88.2 (2.8)	7.28 (<1%)	3.56 (<1%)	1.50 (<1%)
2	117 (2.9)	24.0 (0.06)	7.28 (<1%)	3.56 (<1%)	1.50 (<1%)
3	1.31 (<1%)	1.31 (<1%)	2.23 (<1%)	3.55 (<1%)	1.50 (<1%)

SINGLE OR DOUBLE PRECISION?



Speedup of using single precision over double precision on a Tesla P100

DOUBLE PRECISION ATOMIC UPDATES



Timing data for 10^8 tally updates using 64-bit floating point type

CONCLUSIONS

- Five methods for tallying photon escape on the GPU were compared on four different architectures
- Tesla P100 is the best GPU architecture to use for tallying
  - Process tally updates 2-6 times faster than other architectures
  - Native support for 64-bit floating-point atomic operations
- Tally replication is the most performant method for frequent updates on the GPU if there is sufficient memory available
- Using the warp shuffle feature for tallying on the GPU is often more effective than relying only on atomic operations
  - Warp shuffle method was better for integers
  - Block reduction method was better for floating-point values

ACKNOWLEDGEMENTS

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.

Sandia National Laboratories

LDRD
Laboratory Directed Research and Development

U.S. DEPARTMENT OF ENERGY

NNSA
National Nuclear Security Administration