

Fast Monte Carlo Photon Transport Simulations for Heterogeneous Computing Systems

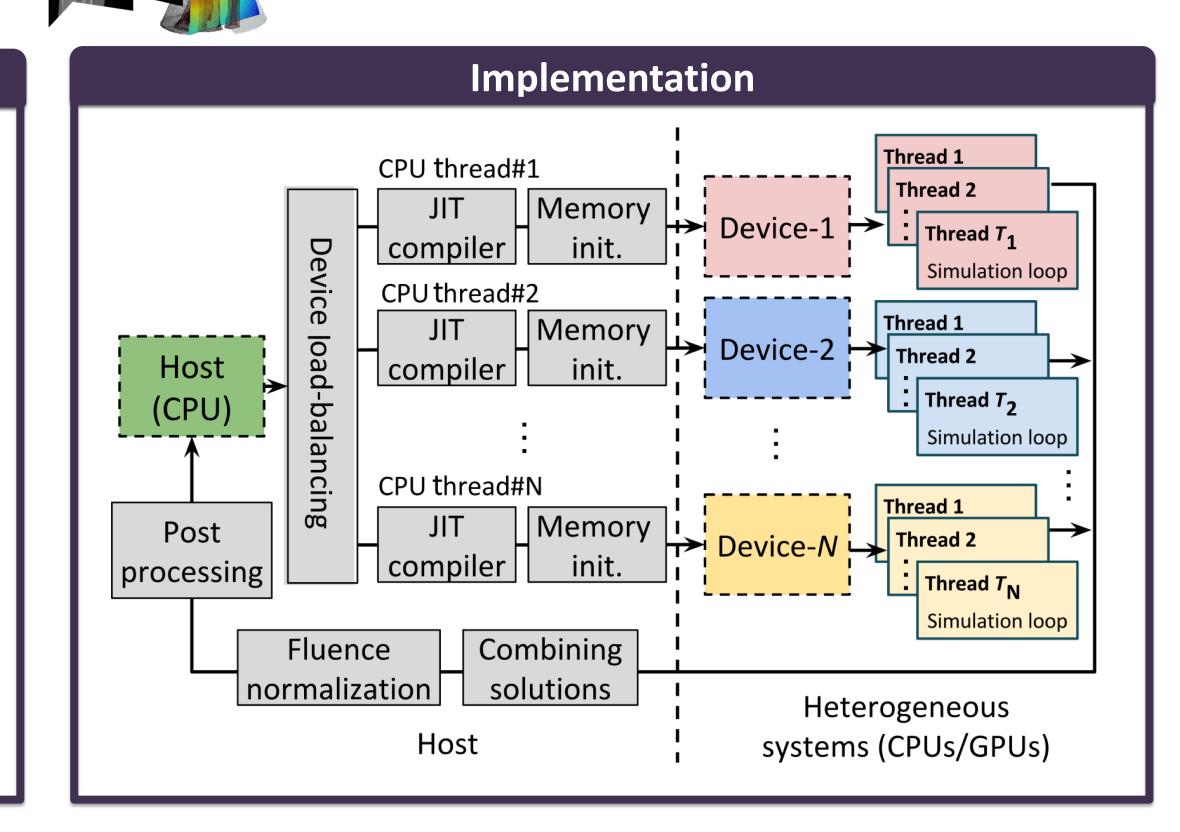
Leiming Yu, Fanny Nina-Paravecino, David Kaeli and Qianqian Fang

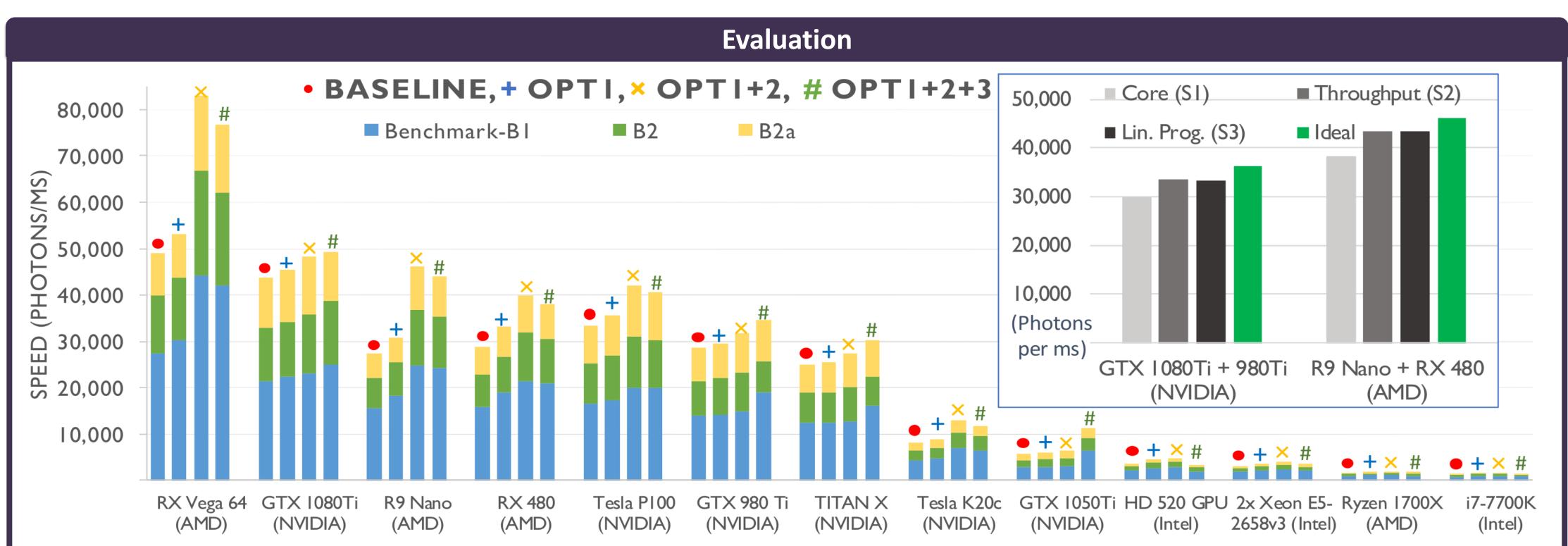


Motivation

- Monte Carlo (MC) method is considered as the gold standard for modeling light propagation inside complex media, such as human brains or bones.
- Due to its computational complexity, sequential execution can easily take up to several hours.
- Leveraging Graphics Processing Units (GPUs), we can significantly reduce the simulation time.
 - For scalability and portability, we have developed a fast Monte Carlo photon transport simulation framework in OpenCL for heterogeneous computing systems.

Algorithm Start Thread_{i+1} Thread, eed GPU RNG with Launch a new photon Global Compute attenuation based on absorption Repetition Compute a new complete? Accumulate photon energy loss to the direction vector TΝ **Retrieve solution** Normalize & save solution **Terminate** thread **GPU**





												_	
Benchmark1	RX Vega 64	GTX 1080Ti	R9 Nano	RX 480	Tesla P100	GTX 980 Ti	Titan X	Tesla K20c	GTX 1050 Ti	HD 520 GPU	2x Xeon E5 2658v3	Ryzen 1700X	i7-7700K
Cores	4096	2816	4096	2304	3584	2816	3072	2496	768	24	24	8	4
Throughput(photons/ms)	44307	24994	24845	21372	19968	18875	16069	6897	6472	2757	2391	1151	1024
Throughput(photons/ms) Per Core	11	9	6	9	6	7	5	3	8	115	100	144	256
Thermal Design Power(Watt)	295	250	175	150	300	250	250	225	75	15	210	95	91
Throughput(photons/ms) Per Watt	150	100	142	142	67	76	64	31	86	184	11	12	11

Optimizations

- ☐ OPT1: Native Math
- ☐ OPT2: Balanced Threads
- ☐ OPT3: Simplify control flow

Benchmarks

- ☐ B1: No reflection
- ☐ B2: Reflection with nonatomic operations
- ☐ B2a: Reflection with atomic operations

Load-partitioning strategies

- ☐ S1: Number of stream-processors
- ☐ S2: Estimated device throughput
- ☐ S3: Linear-programming solution

Results

- GPUs excel in MC simulations with many less powerful cores. The Intel HD
 520 GPU reports the highest power efficiency.
- Optimization schemes achieve a 56% performance improvement on average on AMD GPUs, 20% on Intel CPUs/GPUs, and 10% on NVIDIA GPUs.
- Efficient load-partitioning strategies, based on the device throughput and linear-programming models, achieve higher throughput vs. core-based approach.

References

- L Yu, F Nina-Paravecino, D Kaeli and Q Fang. "Scalable and massively parallel Monte Carlo photon transport simulations for heterogeneous computing platforms." Journal of Biomedical Optics 23, no. 1 (2018): 010504.
- Source code for MCXCL, http://mcx.space/mcxcl

Acknowledgement

• This research is supported by the National Institutes of Health (NIH) grants # R01-GM114365 and R01-CA204443.