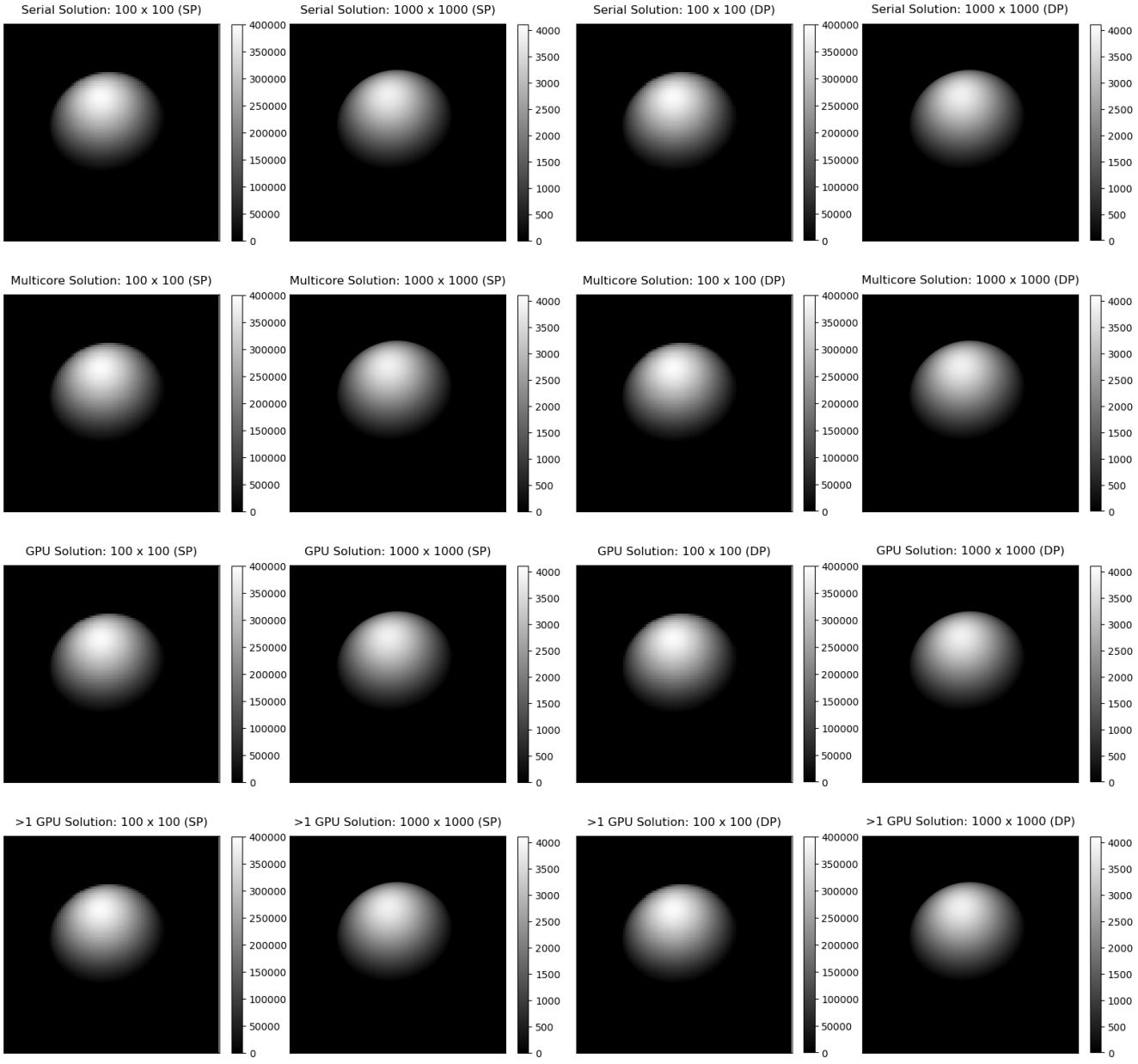
Project 2 — Final Specification

Introduction

Name: Annabelle Huang
Email: ahuang02@uchicago.edu

Validation



Performance

For the serial and OpenMP versions, the key optimizations include using xoshiro as the PRNG, storing vectors on the stack instead of the heap, and caching the intermediate computations in variables to avoid redundant recalculations. In another version, Taylor's expansion is applied for sine and cosine calculations to further optimize performance.

For the GPU version, the main optimizations include offloading the computations to the GPU to maximize parallel processing and utilizing CUDA's built-in math functions to improve numerical efficiency. We also tested different block size and threads per block configurations to get the fastest total and kernel execution times.

For the multi-GPU version, the same kernel-side optimizations were applied to maximize parallel efficiency. On the host side, we also optimized the communication between GPUs using `MPI_Reduce` instead of `MPI_Send` and `MPI_Recv` or `MPI_Gather` to handle data aggregation. Similarly, we also tested different block size and threads per block configurations to get the fastest total and kernel execution times.

Proc	Grid	Time (SP)	KTime (SP)	Time (DP)	KTime (DP)	Blk/TPB	Cores	Samples
A100	1000^2	350.76 ms	276.04 ms	471.78 ms	297.21 ms	2048/1024	–	14,928,181,512
A100	100^2	211.17 ms	208.43 ms	302.63 ms	298.54 ms	2048/1024	–	14,928,181,512
V100	1000^2	386.57 ms	199.20 ms	525.85 ms	336.41 ms	2048/1024	–	14,928,181,512
V100	100^2	202.37 ms	198.98 ms	339.76 ms	336.42 ms	2048/1024	–	14,928,181,512
RTX6000	1000^2	1353.76 ms	1163.29 ms	5447.86 ms	5257.33 ms	2048/1024	–	14,928,181,512
RTX6000	100^2	1167.08 ms	1163.81 ms	5258.23 ms	5251.45 ms	2048/1024	–	14,928,181,512
CPU Serial	1000^2	345.12 s	–	312.60 s	–	–	1	14,928,091,394
CPU Serial	100^2	336.02 s	–	290.88 s	–	–	1	14,928,091,394
CPU OMP	1000^2	21.51 s	–	18.72 s	–	–	16	14,927,951,260
CPU OMP	100^2	21.37 s	–	18.31 s	–	–	16	14,927,951,260
>1 GPU*	1000^2	418.21 ms	227.72 ms	561.29 ms	370.10 ms	1024/512	2	14,922,612,038
>1 GPU*	100^2	235.53 ms	225.30 ms	371.31 ms	367.26 ms	1024/512	2	14,922,612,038

Table 1: Fill out Table 1 with your best performance for the hardware and problem size specified. The Total Time (Time) and Kernel Time (KTime) columns must include both single and double precision performance (SP/DP). Assume one billion rays, xorwow RNG in curand, with problem parameters set as in Milestones 1 and 2. The CPU must be a Midway 3 Cascade Lake node. For the multi-GPU runs, the "cores" column should be used to denote the number of MPI ranks. Samples refers to the total number of random numbers drawn in the simulation. It is included as a sanity check.

Proc	Grid	Time (SP)	KTime (SP)	Time (DP)	KTime (DP)	Blk/TPB	Cores	Samples
CPU Serial	1000^2	133.57 s	–	125.98 s	–	–	1	14,928,091,394
CPU Serial	100^2	130.65 s	–	107.57 s	–	–	1	14,928,091,394
CPU OMP	1000^2	8.86 s	–	9.09 s	–	–	16	14,927,951,260
CPU OMP	100^2	8.55 s	–	7.01 s	–	–	16	14,927,951,260

Table 2: Best performance for the hardware and problem size specified with Taylor's expansion for sine and cosine.