Project #5

CUDA: Monte Carlo Simulation

Marcos Valdez

valdemar@oregonstate.edu

Oregon State University

CS 475 Section 400

Spring Quarter 2023

Submitted:	April 23, 2023
Due:	April 25, 2023

Contents

I. Environment

These results are from a batch run on OSU's DGX via submit-b using nvcc 10.1.

Note: I also ran the same tests on:

- OSU's rabbit using nvcc 10.1
- A self-built machine with a single GeForce RTX 2080 Ti using nvcc 12.1 in Visual Studio 2022 on Windows 11

A summary comparison is included at the end of the document.

II. Probability

Given that the Monte Carlo method yields increased accuracy with more trials, the most accurate estimate of the probability can be found by considering only the results based on the runs with the greatest number of trials (2,097,152).
Tests on DGX were very consistent. The probability converged on 74.68% +/- 0.01%. When compared to the same results from rabbit and my GPU, results converged on 74.69% +/- 0.06%. Given this, the probability, with high confidence, is:

Probability ≈ 74.7%

III. Performance Results

A. Table (probabilities excluded)

Performance (MegaTrials/s)

	Number of Trials						
Block Size	1024	4096	16384	65536	262144	1048576	2097152
8	34.4828	133.3333	457.1428	1465.9985	2975.6629	4067.0224	4396.6187
32	35.7143	121.2121	533.3333	1770.0951	4887.8283	10317.3805	11863.8664
64	32.2581	137.9310	500.0000	1843.3844	5326.3981	9972.0021	12158.8125
128	31.2500	142.8571	516.1290	2072.8746	5357.7502	10132.3440	12268.0640

B. Graphs

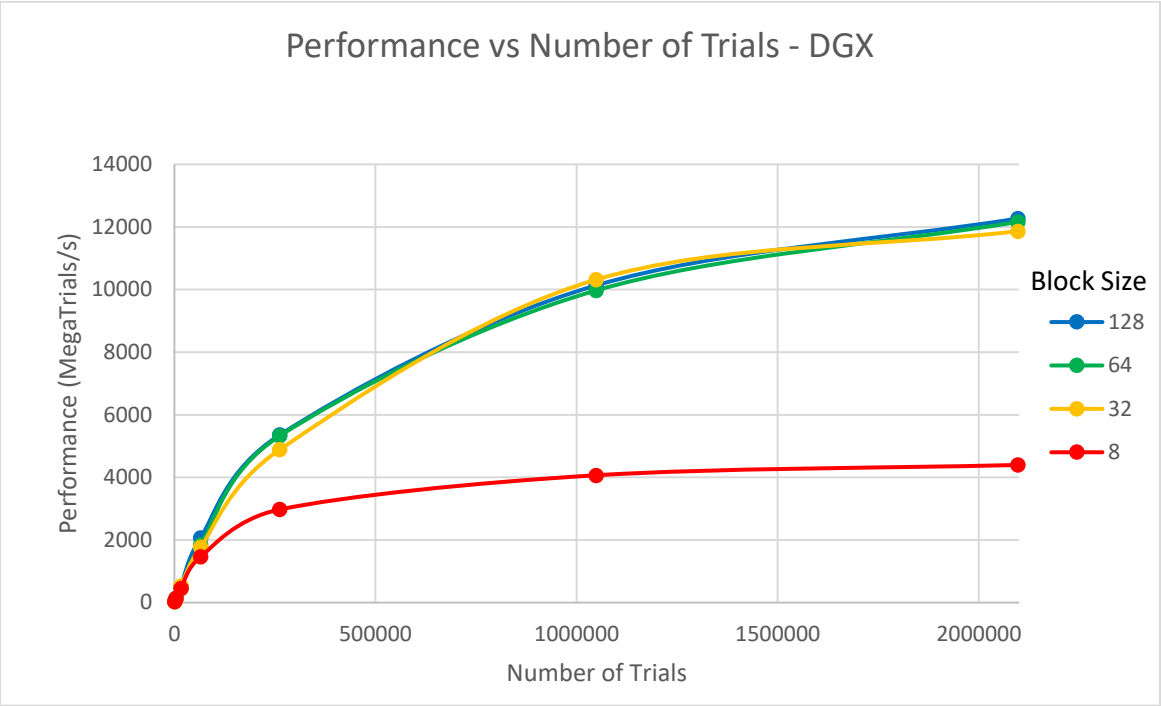

Figure 1: Performance as a Function of Number of Trials for Each Tested CUDA Block Size on DGX

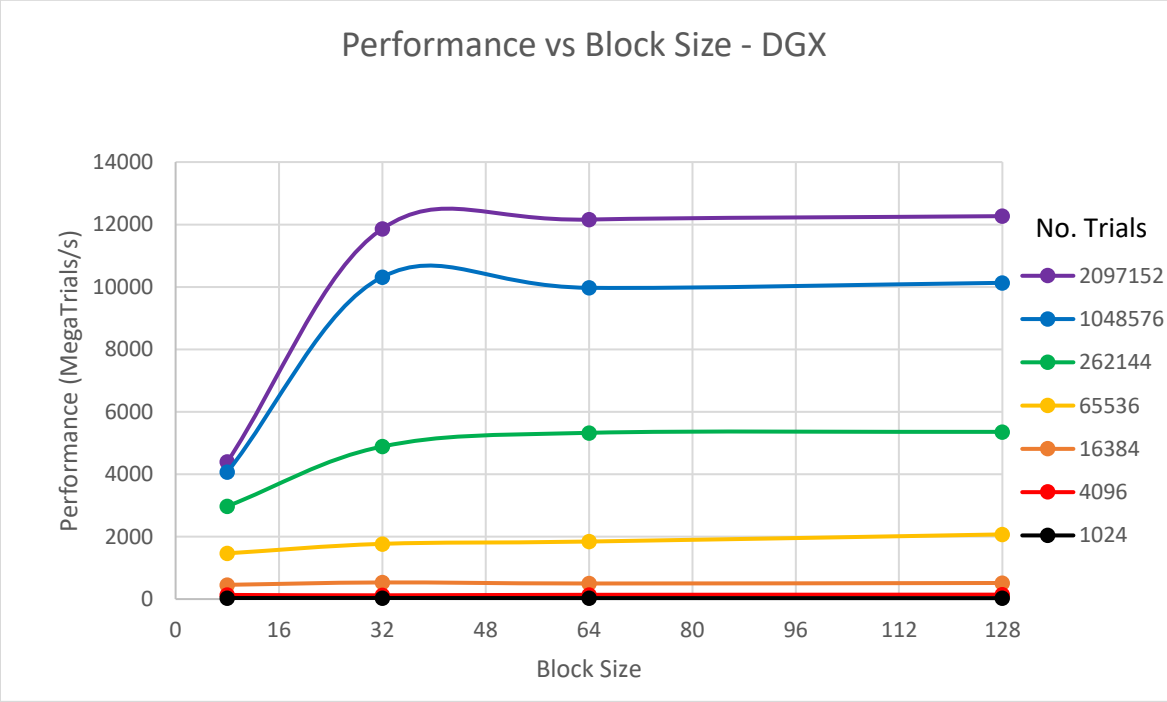

Figure 2: Performance as a Function of CUDA Block Size for Each Tested Number of Trials on DGX

IV. Commentary

A. Performance Graph Patterns

When performance is plotted vs number of trials, block size of 8 near their maximum performance by 250,000 trials and reach it by 1,000,000 trials. All block sizes 32 and above perform similarly: performance increases steadily and appears to begin leveling off near 2,000,000 trials.
When performance is plotted vs block size, numbers of trials less than 250,000 are similarly performant regardless of block size. When there are more than 250,000 trials, performance is lower for block sizes under 32 and reaches its maximum at 32.

B. Pattern Explanation

Performance increases for greater numbers of trials as overhead becomes a smaller fraction of the work being performed. This trend levels off as the parallel work eclipses the overhead and the GPU approaches its maximum throughput. A block size 32 is already making full use of the GPU's warps. Multiples of this value are not functionally different. This probably indicates that Monte Carlo code never really leaves warp idle.

C. Block Size of 8

A block size of 8 is ¼ of the GPU warp size and thus leaves a large portion of the hardware idle during all parallel processing. When most of the work is parallel (by 250,000 trials), this results in a drastic reduction in performance.

D. Comparison to Project 1

I ran Project 1 on an Intel Core i9-9900K. It has 8 cores with hyperthreading and was running at its stock clock speed of 3.60 GHz. Its maximum performance was achieved when using 16 OpenMP threads and running any Number of Trials greater than or equal to 100,000.
The maximum performance for that CPU was 285 MegaTrials/s. For a similar number of trials (250,000), excluding a block size of 8, DGX sits at an average of 5,191 MegaTrials/s. By 2,000,000 trials, it achieves 12,096 MegaTrials/s with a trend line that suggests a bit more room for improvement.
Even my single RTX 2080 Ti performed significantly better than the CPU (see graph at end of document). For 250,000 trials, it reached 2,793 MegaTrials/s. For 2,000,000 trials, it reached 4,205 MegaTrials/s.

E. Implications of GPU Parallel Computing

For applications that do not need to handle irregular data, perform branch prediction, or perform out-of-order execution, GPUs make parallel computing vastly more efficient. A single GPU can process regular data many times faster than a single multicore CPU.

V. Appendix – Hardware Comparison

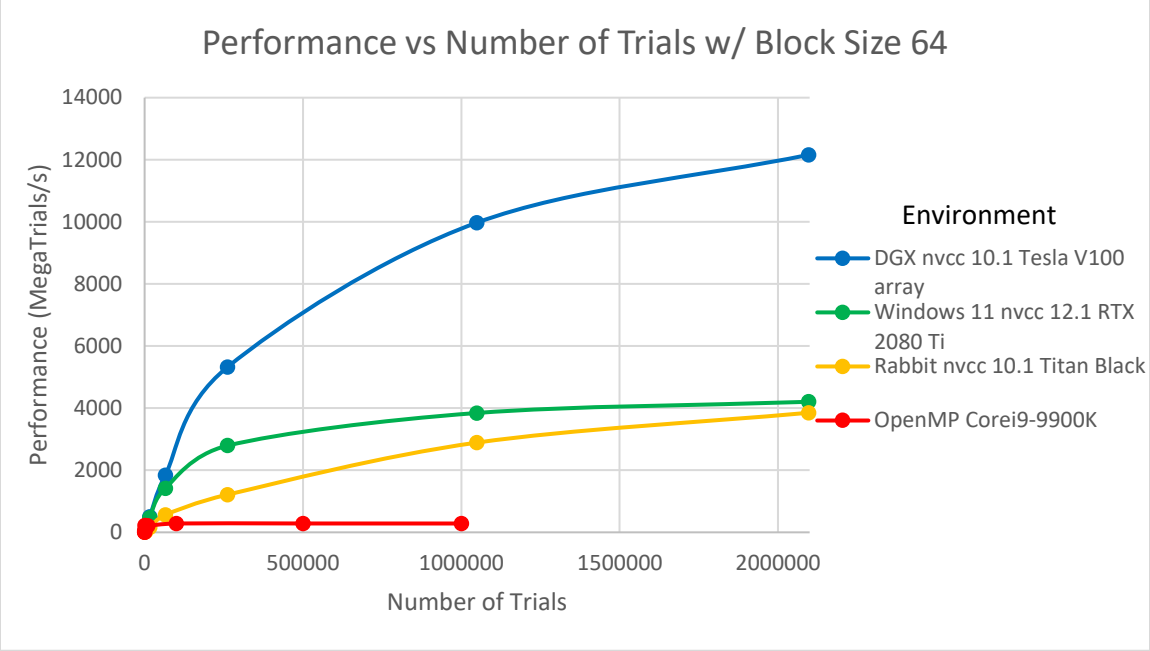

Figure 3: Performance as a Function of Number of Trials for a CUDA Block Size of 64 on Different Hardware