

CS 575 Project #5
CUDA: Monte Carlo Simulation

1. Tell what machine you ran this on

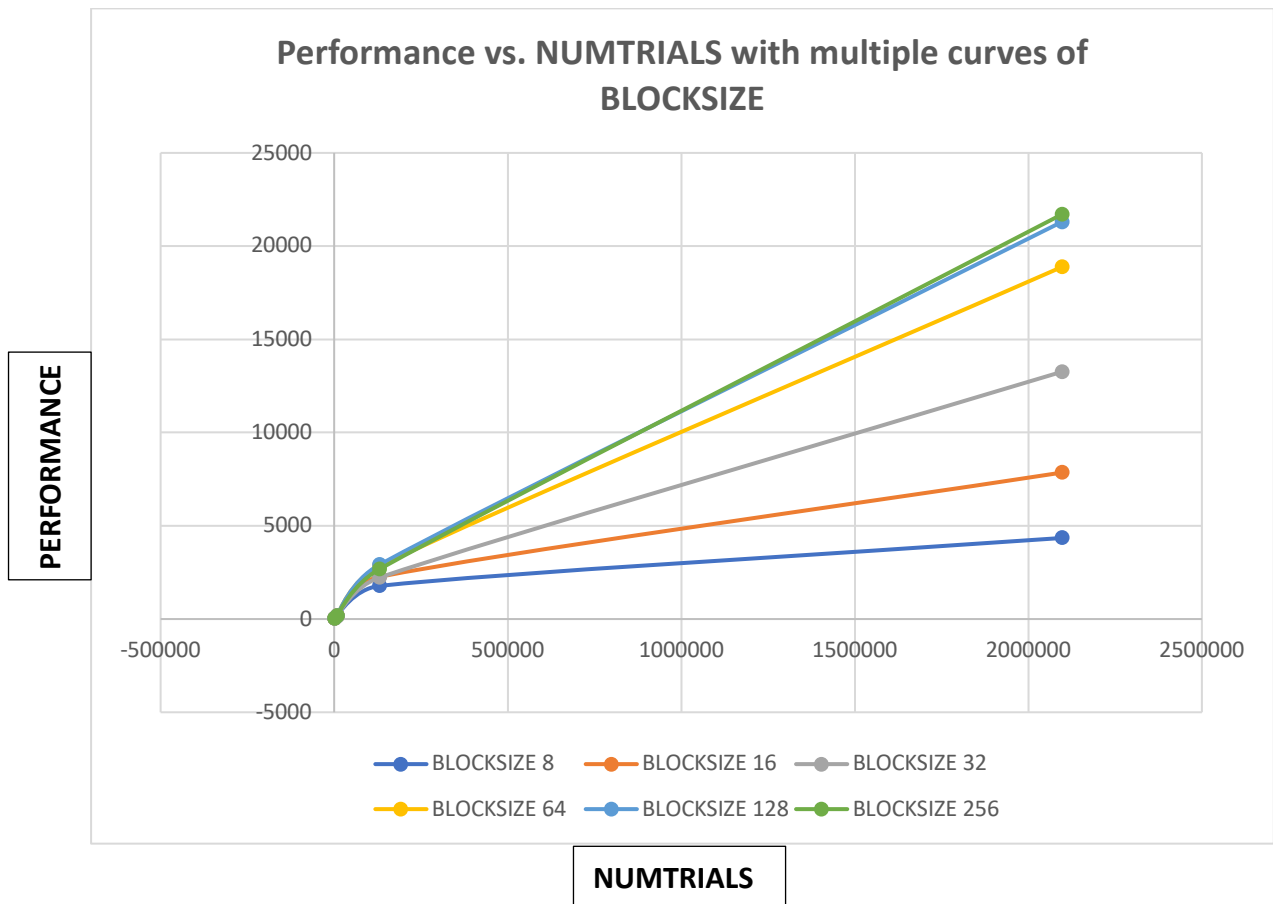
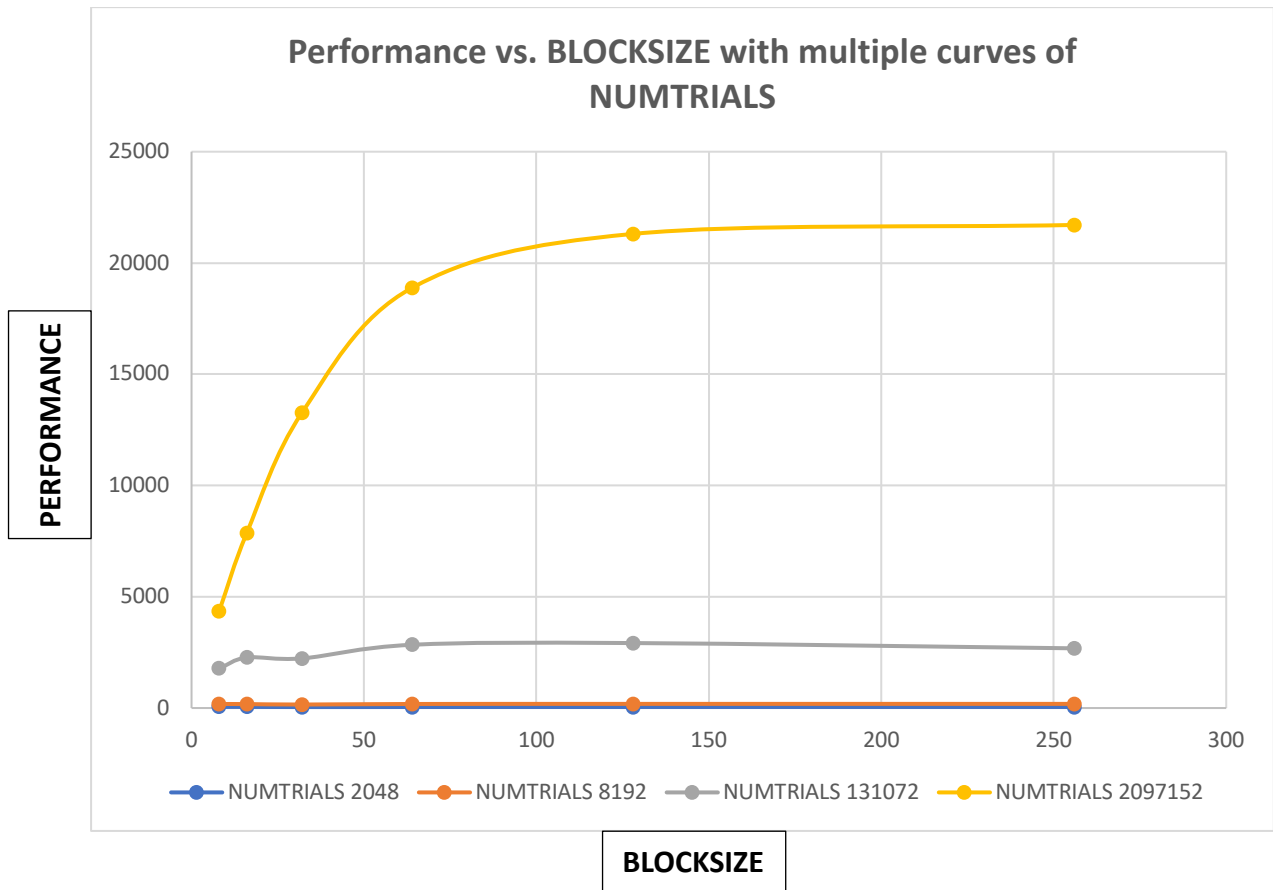
→ DGX

2. What do you think this new probability is?

→ The new probability is approximately 83.8%

3. Show the rectangular table and the two graphs

NUMTRIALS	BLOCKSIZE	Performance	Probability
2048	8	50	83.35
2048	16	51.2821	83.84
2048	32	44.4444	82.96
2048	64	46.5116	83.45
2048	128	47.619	83.84
2048	256	47.619	83.84
8192	8	181.8182	83.87
8192	16	177.7778	84.01
8192	32	156.8627	84.09
8192	64	181.8182	83.08
8192	128	186.0465	83.84
8192	256	186.0465	83.67
131072	8	1778.5497	83.83
131072	16	2266.7404	83.89
131072	32	2229.7224	83.84
131072	64	2842.4704	83.96
131072	128	2919.4584	83.87
131072	256	2682.3838	83.79
2097152	8	4348.1953	83.75
2097152	16	7856.1499	83.78
2097152	32	13263.7114	83.84
2097152	64	18886.4547	83.81
2097152	128	21305.5923	83.81
2097152	256	21707.8507	83.8



4. What patterns are you seeing in the performance curves?

- In the performance vs. BLOCKSIZE graph, performance generally increases with larger BLOCKSIZE values up to a certain point, then plateaus or slightly decreases.
- In the performance vs. NUMTRIALS graph, performance increases linearly with larger NUMTRIALS for all BLOCKSIZE values. Larger BLOCKSIZE values achieve higher performance.

5. Why do you think the patterns look this way?

- The increase in performance with larger BLOCKSIZE values is due to better utilization of GPU resources, which allows more parallel computations. However, after a certain point, additional threads may not improve performance significantly due to overhead and resource contention.
- The linear increase in performance with larger NUMTRIALS is expected because more trials mean more data to process, which scales the workload and the achieved performance.

6. Why is a BLOCKSIZE of 8 so much worse than the others?

- A BLOCKSIZE of 8 is worse because it underutilizes the GPU resources. GPUs are designed to handle thousands of threads concurrently, and a small BLOCKSIZE does not provide enough threads to keep the GPU busy, leading to inefficiency.

7. How do these performance results compare with what you got in Project #1? Why?

- The performance results in Project #5 significantly surpass those in Project #1, with Project #5 achieving up to 21,707.85 Megatrials/Second compared to Project #1's 307.52 Megatrials/Second. This stark contrast is due to Project #5 utilizing CUDA on a GPU, which is optimized for handling massive parallel tasks with thousands of threads concurrently, unlike the CPU used in Project #1.

8. What does this mean for what you can do with GPU parallel computing?

- GPU parallel computing allows for significant performance improvements in tasks that can be parallelized, such as Monte Carlo simulations. By optimizing the number of threads per block and the workload, we can achieve high throughput and efficiency, making GPUs ideal for large-scale simulations and data processing tasks.