Challenge #13: Benchmarking Different SAXPY Problem Sizes

Objective:

To profile and visualize the execution time of a CUDA SAXPY implementation for increasing problem sizes.

Approach:

- CUDA implementations of SAXPY were written and executed for matrix sizes from 2^15 to 2^25.
- Execution times were recorded and plotted using Python's Matplotlib.
- Initial execution for small size (2^15) showed higher time likely due to setup overhead.
- Larger matrix sizes had nearly constant low execution time, suggesting efficient GPU handling or timing inaccuracies.

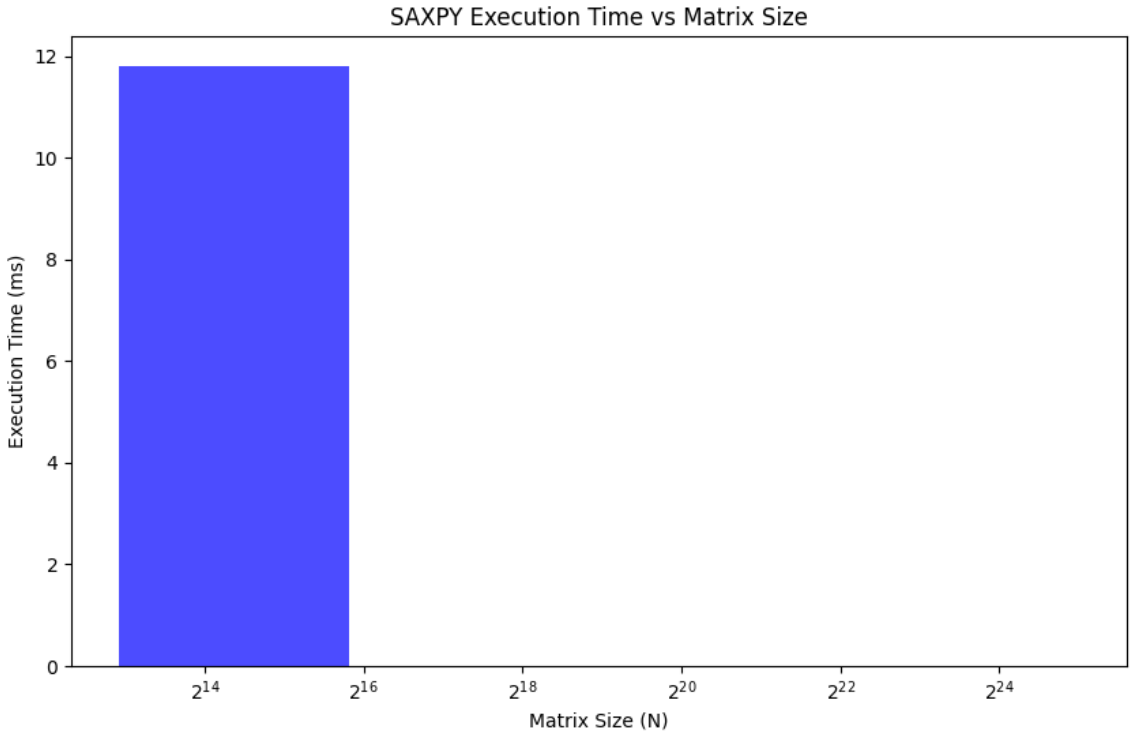
Observations:

- The first run showed a spike in execution time, possibly due to initialization overhead or cold cache effects.
- Subsequent runs remained consistently fast, possibly indicating insufficient problem size to stress the GPU.
- Precision in timing for such fast operations can be limited without cudaEvent timing.

Next Steps:

- Integrate cudaEventRecord() in the CUDA code to separately measure:
    * Total execution time (memory + kernel)
    * Pure kernel time
- Re-run benchmarks to compare and analyze further.

Performance Chart:


SAXPY Execution Time vs Matrix Size
Execution Time (ms)
0
2
4
6
8
10
12
2^{14}
2^{16}
2^{18}
2^{20}
2^{22}
2^{24}
Matrix Size (N)