# RESULTS

VECTOR ADDITION

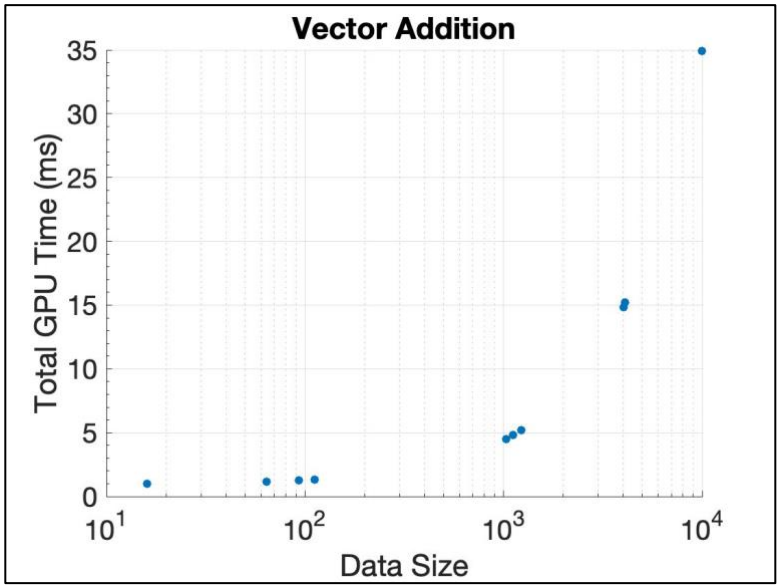

Figure 1: Total GPU Time for addition of varying data size

We plot a line curve rather than histogram because our dataset is small and concentrated in regions. So, a line plot can give us better information about our results. The total time increased with increasing data size mostly because importing data and creating memory on host time increased with data size.
The types of time computed is: Importing Data and creating memory on host, Allocating GPU memory, Copying Input memory to GPU, CUDA computation time, Output memory to CPU copy time, Free GPU memory time, Total Time, Data Transfer time. Because of the small dataset CUDA computation time did not show any pattern. The data transfer time had a parabolic shape.

| All time in millisecond (ms) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Vector Add | | | | | | | | | | |
| Block Size -32 | | | | | | | | | | |
| Dataset | Data Size Inp | Input 2 | Importing Data | Allocating GI | Copying Inpu | CUDA Compu | Output mem | Free GPU Me | Total Time | Data Transfe |
| 0 | 16 | 16 | 0.14 | 0.4 | 0.04 | 0.03 | 0.03 | 0.37 | 1.01 | 0.07 |
| 1 | 64 | 64 | 0.3 | 0.4 | 0.03 | 0.03 | 0.03 | 0.37 | 1.16 | 0.06 |
| 2 | 93 | 93 | 0.4 | 0.4 | 0.033 | 0.03 | 0.025 | 0.37 | 1.258 | 0.058 |
| 3 | 112 | 112 | 0.46 | 0.4 | 0.034 | 0.032 | 0.026 | 0.37 | 1.322 | 0.06 |
| 4 | 1120 | 1120 | 3.94 | 0.4 | 0.034 | 0.026 | 0.025 | 0.37 | 4.795 | 0.059 |
| 5 | 9921 | 9921 | 34.06 | 0.4 | 0.051 | 0.025 | 0.029 | 0.37 | 34.935 | 0.08 |
| 6 | 1233 | 1233 | 4.34 | 0.4 | 0.04 | 0.029 | 0.026 | 0.37 | 5.205 | 0.066 |
| 7 | 1033 | 1033 | 3.66 | 0.4 | 0.039 | 0.025 | 0.026 | 0.37 | 4.52 | 0.065 |
| 8 | 4098 | 4098 | 14.33 | 0.4 | 0.043 | 0.025 | 0.026 | 0.37 | 15.194 | 0.069 |
| 9 | 4018 | 4018 | 13.95 | 0.4 | 0.048 | 0.025 | 0.027 | 0.37 | 14.82 | 0.075 |

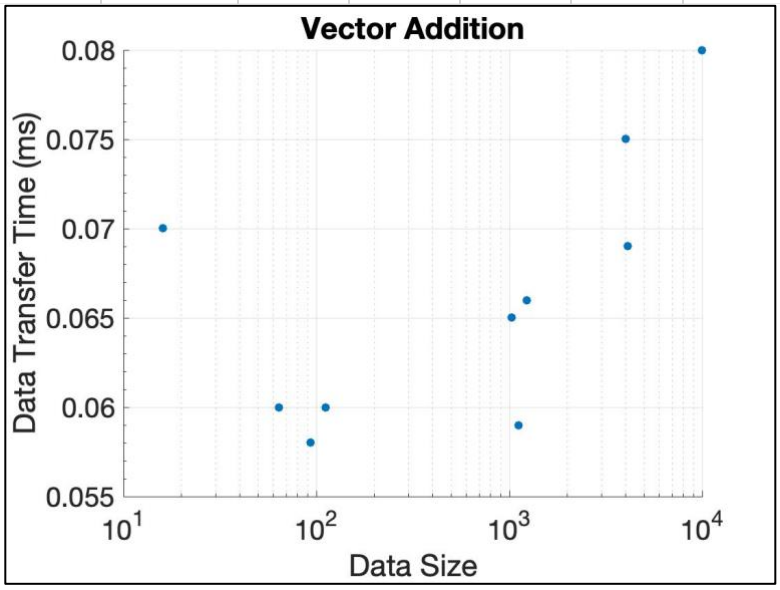

Figure 2: Data Transfer Time for addition of varying data size

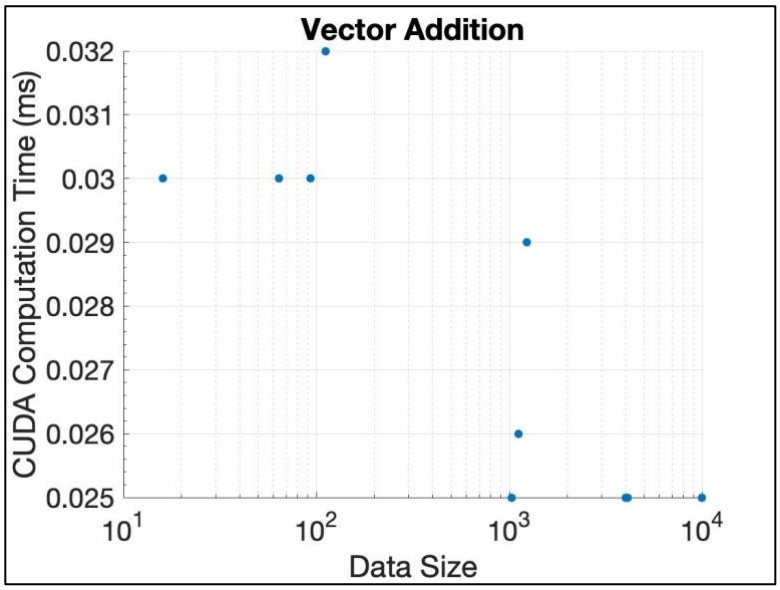

Figure 3: CUDA Computation Time for addition of varying data size

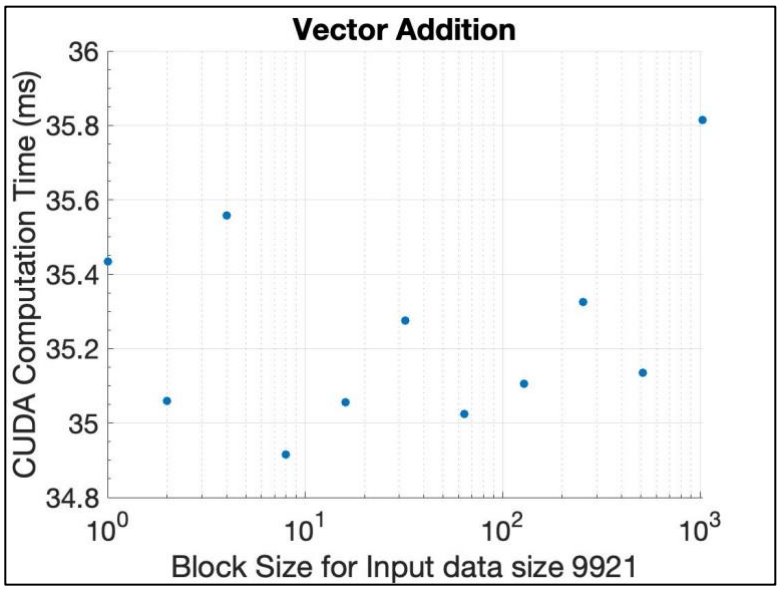

Figure 4: CUDA Computation Time for addition of varying block size for data size of 9921

For varying Block Size, the CUDA computation time was all in the same range with a very large block size of 1024x1x1 being not very favorable.

BASIC MATRIX MULTIPLICATION

To plot the total time vs the matrix size, we determine the number of multiplication and addition operations that need to be performed for matrix multiplication. For multiplication of two, $k * l$ and $l * m$ matrix, the total number of addition and multiplication operations is given by $2klm - km$. We set that as our x-axis for matrix multiplication and is a good indicator of data size.

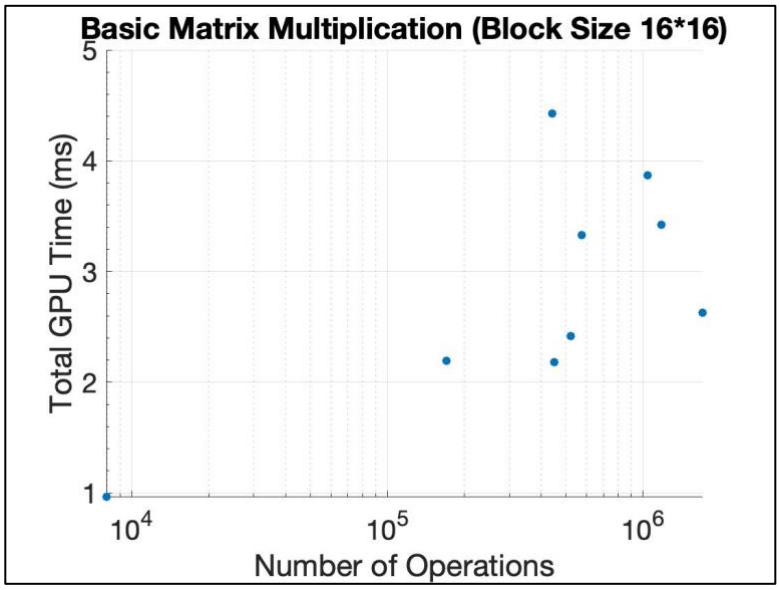

Figure 5: Total GPU Time for basic matrix multiplication for varying number of operations

The total time as seen increased with increasing number of operations. The data size of matrix can be seen in the table.

| Basic Matrix Multiplication | | | K*L x L*M matrix requires KM(L + L -1) multiplication and addition operations. | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Block Size - 16X16 | | | | | | | | | | |
| Dataset | Data Size Inp | Input 2 | Importing Data | Allocating GI | Copying Inpu | CUDA Compu | Output mem | Free GPU Me | Total Time | Data Transfe |
| 0 | 16X16 | 16x16 | 0.22 | 0.33 | 0.032 | 0.035 | 0.025 | 0.328 | 0.97 | 0.057 |
| 1 | 64X64 | 64x64 | 1.57 | 0.401 | 0.037 | 0.034 | 0.027 | 0.346 | 2.415 | 0.064 |
| 2 | 64X128 | 128x64 | 2.98 | 0.409 | 0.043 | 0.034 | 0.027 | 0.378 | 3.871 | 0.07 |
| 3 | 112x48 | 48x16 | 1.38 | 0.373 | 0.037 | 0.03 | 0.026 | 0.347 | 2.193 | 0.063 |
| 4 | 84x84 | 84x84 | 2.56 | 0.381 | 0.046 | 0.033 | 0.028 | 0.376 | 3.424 | 0.074 |
| 5 | 80x99 | 99x28 | 3.61 | 0.366 | 0.051 | 0.035 | 0.03 | 0.332 | 4.424 | 0.081 |
| 6 | 67x53 | 53x64 | 1.35 | 0.339 | 0.047 | 0.032 | 0.027 | 0.387 | 2.182 | 0.074 |
| 7 | 29x117 | 117x85 | 2.43 | 0.422 | 0.042 | 0.035 | 0.027 | 0.371 | 3.327 | 0.069 |
| 8 | 191x19 | 19x241 | 1.71 | 0.361 | 0.043 | 0.029 | 0.085 | 0.399 | 2.627 | 0.128 |

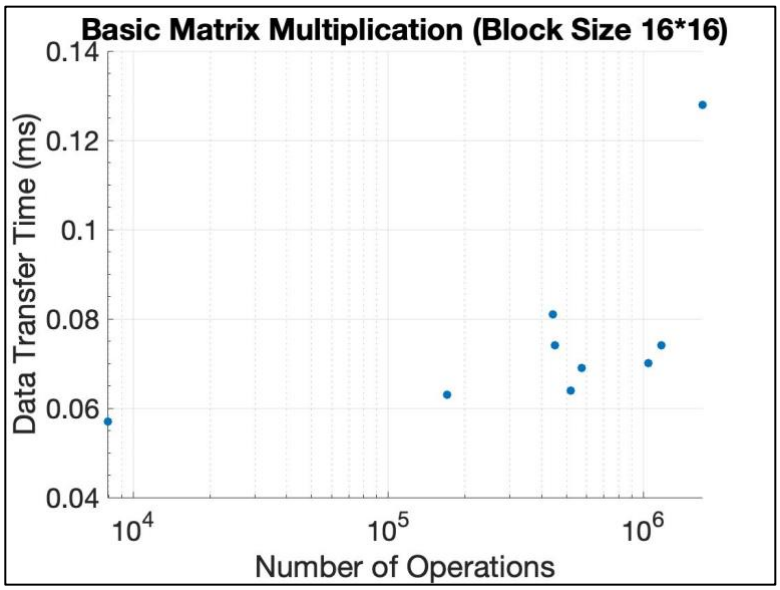


Figure 6: Data Transfer Time for basic matrix multiplication for varying number of operations

The data transfer time showed no specific pattern, but CUDA computation time generally increased with number of operations.

We then varied the block size for 64x128 and 128x64 dataset. We captured the CUDA computation time varying square block size of 1*1, 2*2, 4*4, 8*8, 16*16, 32*32. 1x1 and 32x32 took largest computation time.

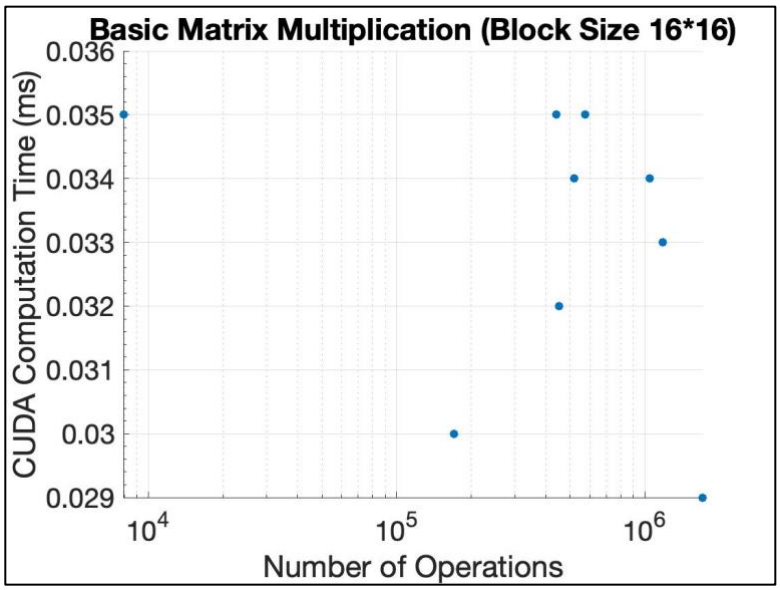

Figure 7: CUDA Computation Time for basic matrix multiplication for varying number of operations

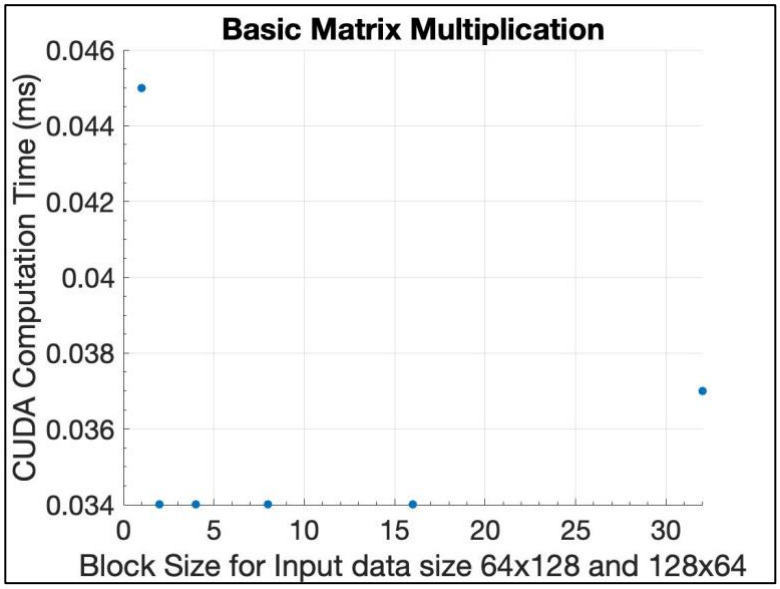

Figure 8: CUDA Computation Time for basic matrix multiplication for varying square block size

TILED MATRIX MULTIPLICATION

We then repeated the above tasks for tiled matrix multiplication. The total time increase with increasing number of operations on expected lines. We did not observe any trends for data transfer time like for vector addition and basic matrix multiplication. The CUDA computation time increased with number of operations.
When we varied the square block size, just like basic matrix multiplication, the CUDA computation time was high for lowest and highest block size.

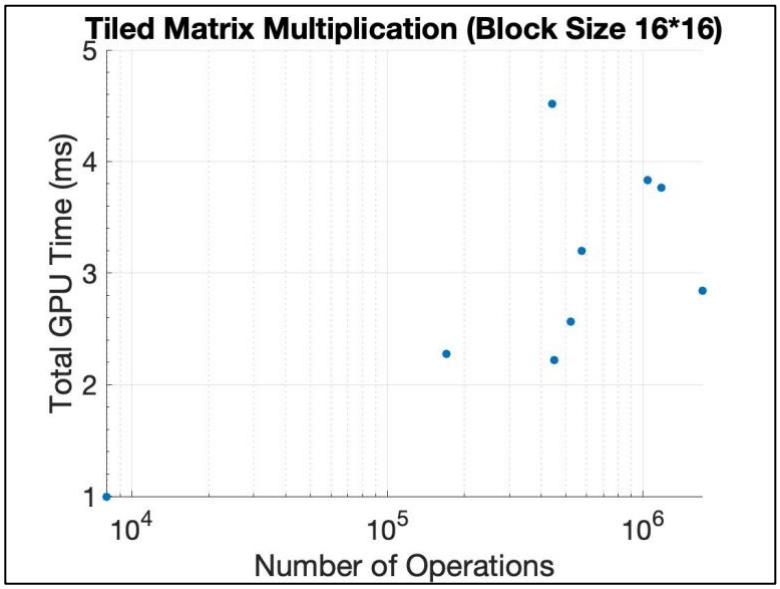

Figure 9: Total GPU Time for basic matrix multiplication for varying number of operations

| Tiled Matrix Multiplication | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Tile Size | 16 | Block Size | 16x16 | | | | | | | |
| Dataset | Data Size Inp | Input 2 | Importing Data | Allocating Gl | Copying Inpu | CUDA Compu | Output mem | Free GPU Me | Total Time | Data Transfe |
| 0 | 16X16 | 16x16 | 0.215 | 0.362 | 0.033 | 0.03 | 0.025 | 0.333 | 0.998 | 0.058 |
| 1 | 64X64 | 64x64 | 1.745 | 0.383 | 0.042 | 0.027 | 0.026 | 0.342 | 2.565 | 0.068 |
| 2 | 64X128 | 128x64 | 2.961 | 0.363 | 0.048 | 0.03 | 0.027 | 0.404 | 3.833 | 0.075 |
| 3 | 112x48 | 48x16 | 1.288 | 0.429 | 0.068 | 0.047 | 0.044 | 0.401 | 2.277 | 0.112 |
| 4 | 84x84 | 84x84 | 2.578 | 0.407 | 0.054 | 0.299 | 0.028 | 0.398 | 3.764 | 0.082 |
| 5 | 80x99 | 99x28 | 3.598 | 0.402 | 0.056 | 0.029 | 0.029 | 0.402 | 4.516 | 0.085 |
| 6 | 67x53 | 53x64 | 1.354 | 0.357 | 0.069 | 0.044 | 0.043 | 0.353 | 2.22 | 0.112 |
| 7 | 29x117 | 117x85 | 2.406 | 0.363 | 0.042 | 0.031 | 0.027 | 0.328 | 3.197 | 0.069 |
| 8 | 191x19 | 19x241 | 1.68 | 0.401 | 0.038 | 0.27 | 0.089 | 0.361 | 2.839 | 0.127 |

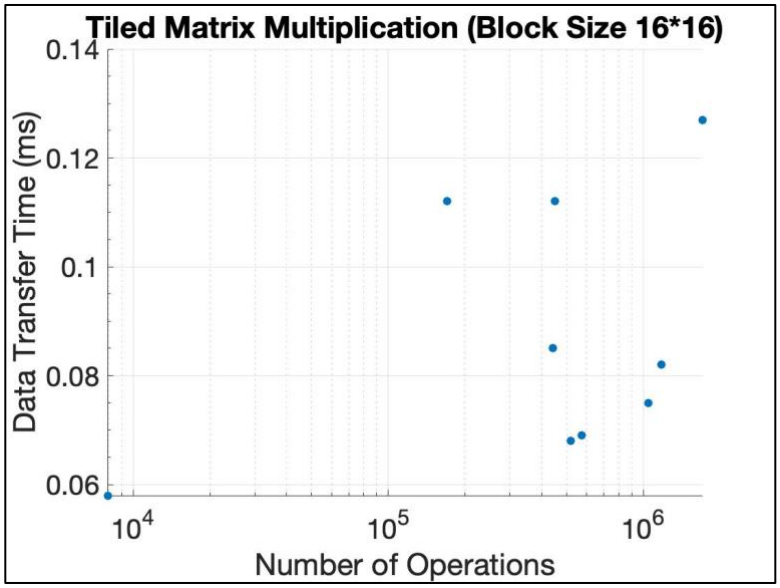

Figure 10: Data Transfer Time for basic matrix multiplication for varying number of operations

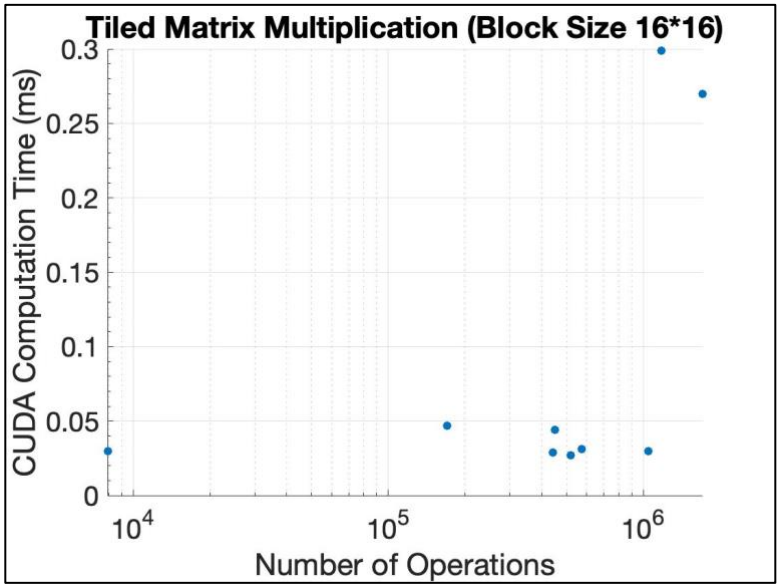


Figure 11: CUDA computation Time for basic matrix multiplication for varying number of operations

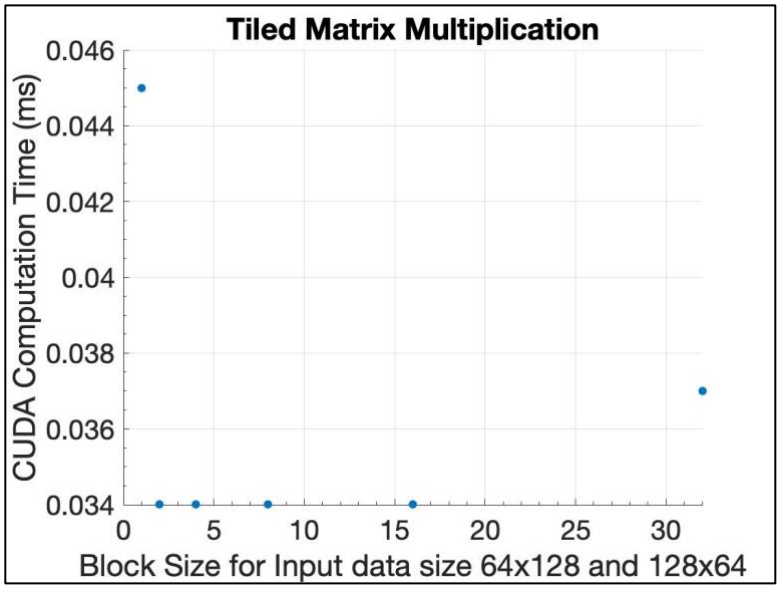


Figure 12: CUDA Computation Time for basic matrix multiplication for varying square block size