RESULTS

VECTOR ADDITION

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| 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.



|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 2: Data Transfer Time for addition of varying data size |

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 3: CUDA Computation Time for addition of varying data size |

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| 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, and matrix, the total number of addition and multiplication operations is given by . We set that as our x-axis for matrix multiplication and is a good indicator of data size.

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| 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.



|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| 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.

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 7: CUDA Computation Time for basic matrix multiplication for varying number of operations |

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| 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.

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 9: Total GPU Time for basic matrix multiplication for varying number of operations |
|  |



|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 10: Data Transfer Time for basic matrix multiplication for varying number of operations |

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 11: CUDA computation Time for basic matrix multiplication for varying number of operations |

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 12: CUDA Computation Time for basic matrix multiplication for varying square block size |