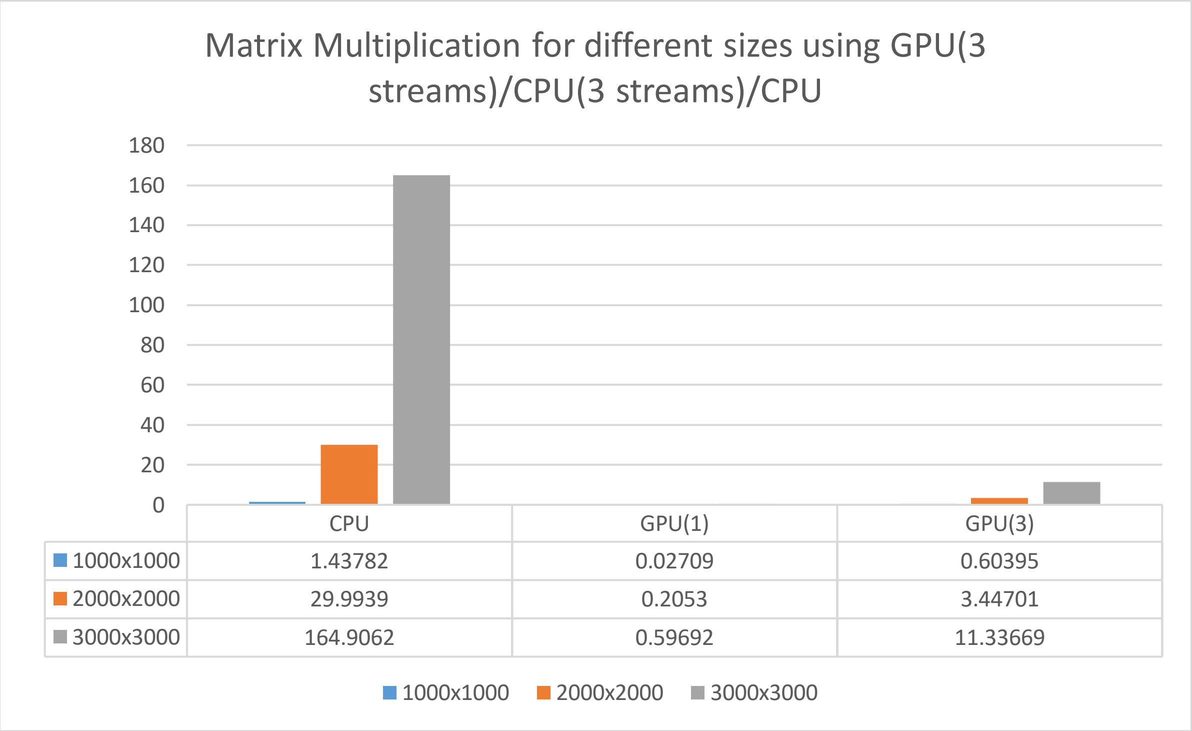
CS398 Report A5

Analysis of CPU vs GPU usage for both single stream and triple stream

The below figure shows a bar graph plotting time taken for each matrix size multiplication for double precision floating point values and speedup from CPU to GPU, comparing GPU streams with 32 block size, 256 tile size and 3 streams. The data table is included as certain values are unable to be seen due to the difference in values.



	CPU	GPU(1)	GPU(3)
🟦 1000x1000	1.43782	0.02709	0.60395
🟧 2000x2000	29.9939	0.2053	3.44701
⬜ 3000x3000	164.9062	0.59692	11.33669

Figure 1. Bar graph plotting time taken for matrix multiplications for GPU single stream, GPU triple stream and CPU.

As seen from the figure above, the single stream GPU version is significantly faster than the triple stream version on the GPU. Both are significantly faster than the CPU version. However, the triple stream version takes significantly lesser memory than the GPU version as they are split into 3 tiles of smaller sizes.

That being said, the block sizes and tile sizes of the streamed version could have caused a performance bottleneck and to further test our theory we have to consider different combinations of the block and tile sizes.

Analysis of GPU tile size and block size usage for triple stream

The below table takes into account different block sizes and tile sizes for stream number of 3.

	Block size | tile size						
	32|256	256 |32	256 | 256	256 |1024	1024 | 256	512 | 512	1024|1024
1000x1000	0.60395	0.27676	0.20532	0.20248	0.22401	0.20496	0.23913
2000x2000	3.44701	1.03992	0.40964	0.42592	0.39318	0.37037	0.40761
3000x3000	11.33669	2.86101	1.08854	0.95404	0.94757	0.97424	0.8974

Table 1. Comparison of different block sizes and tile sizes.

As we can see from the table above. The original tile size of 32 and 256 is significantly slower with a small block size. It is highlighted to compare with other block sizes. We can see a trend of speed increase as the block size increases. However, as the block size becomes significantly large, the speed starts to drop. Based on the different tile sizes, the tile size does affect the speed a little but not to a very big extend as compared to the block size.

The graph below show the trend.

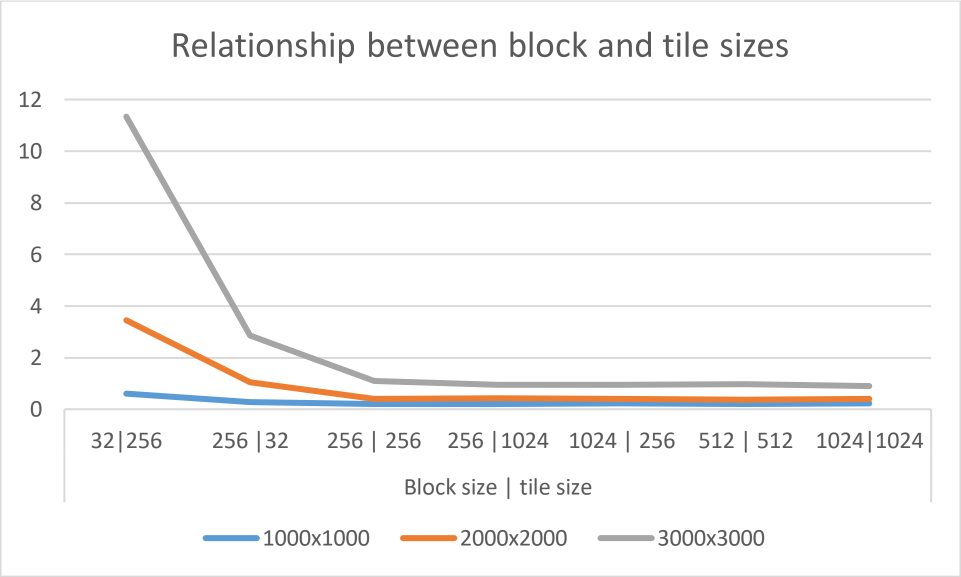


Figure 2. Relationship of different block sizes and tile sizes.

We can also see that it never hits the timing of the single stream GPU. This may be due to the time taken to allocate each stream buffer.

Final notes

The performance requirements were not being able to be met due to the performance throughput of using a double precision floating point value rather than a single precision floating point value. However, it is required to have a double precision floating point value to match the epsilon requirement.