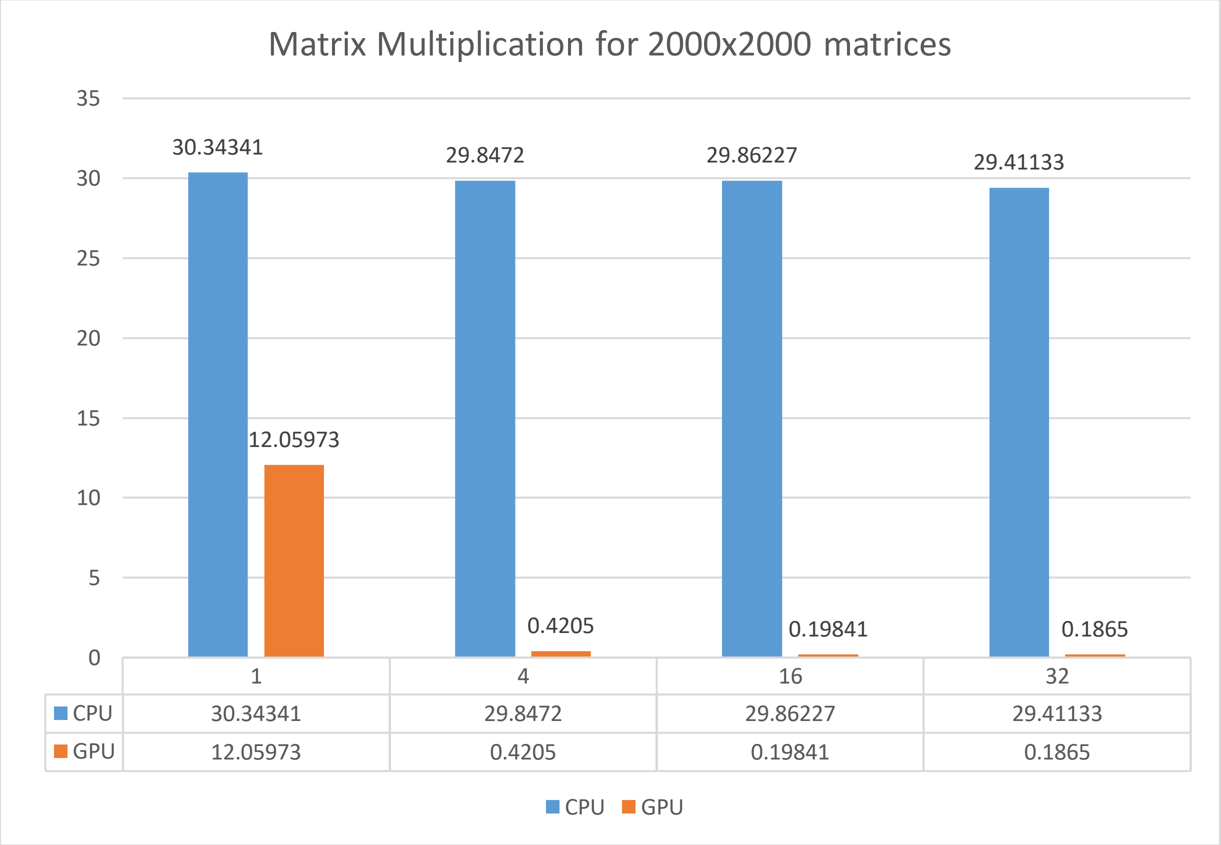
CS398 Assignment 2 report

Analysis of CPU vs GPU usage for different grid size

The below figure shows a bar graph plotting time taken for 2000 by 2000 matrix multiplication for double precision floating point values and speedup from CPU to GPU. The data table is included as certain values are unable to be seen due to the difference in values.



	1	4	16	32
🟦 CPU	30.34341	29.8472	29.86227	29.41133
🟧 GPU	12.05973	0.4205	0.19841	0.1865

Figure 1. Bar graph plotting time taken for matrix multiplications for different shared memory grid sizes for GPU with comparison between CPU and GPU.

As seen from the above figure, as the grid size increases, the speed increase is considerably larger than the previous, with the best speed being at grid size 32. This is due to the grid size being aligned to the warp size, allowing each thread to have a 1 to 1 relationship to each of the shared memory and fully utilizing it when doing matrix multiplication.

The grid size of 64 is not included due to the hardware limitations of 48kB of shared memory while the shared memory of 2 64x64 shared memory matrices puts the shared memory required at 65kB.

We can also see that having a shared memory grid size of 1 is considerable slow and can be compared to not using shared memory at all. This allows us to see the speed up between having shared memory and not having shared memory at all.

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.