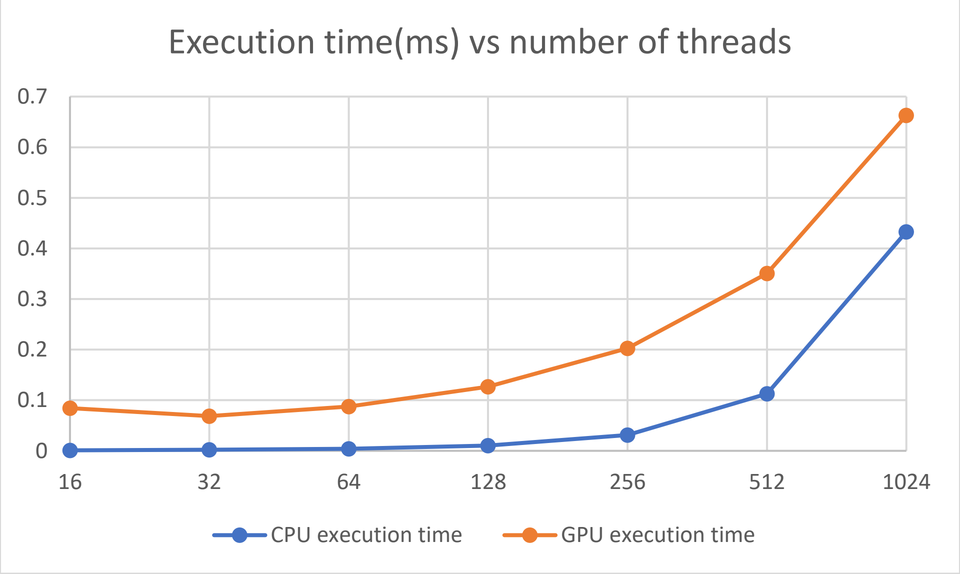
CSCE 435

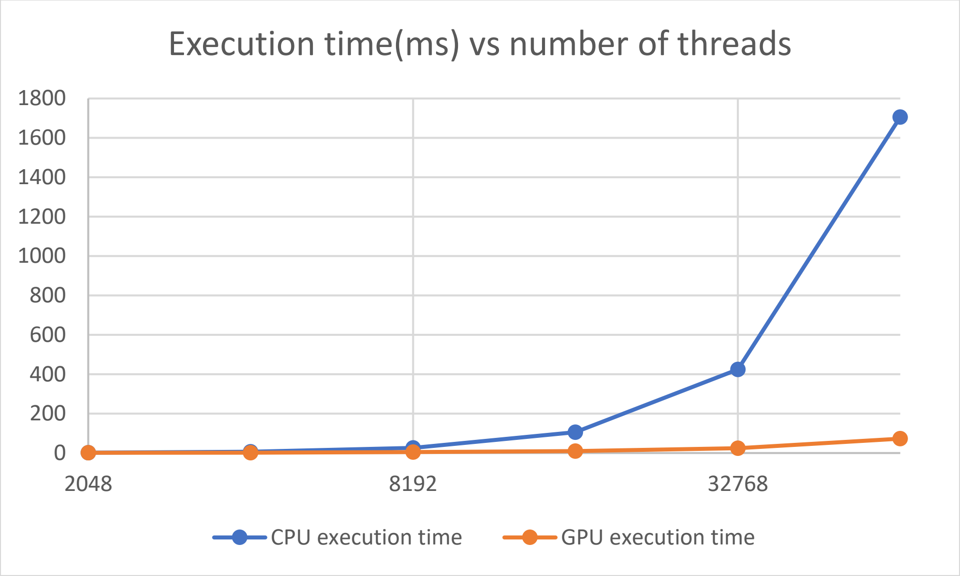
Sahil Palnitkar

1) I added a minimum_distance_kernel function to use CUDA code to calculate the minimum distance between points parallelly. The number of threads running equal the total number of points. Each thread will calculate the distance between its point and all the points after it. The total number of points are divided into blocks of 1024 points. Every block will then hold the smallest distance between the points in it and any of the other n-1 points. After all the blocks finish calculation, the last thread goes through all the minimums and calculates the final minimum and sends it back to the host. The function is based off the host code, but it divides the code into blocks to allow to monitor thread ids to run many threads at the same time.
2)

n	CPU execution time(ms)	GPU execution time(ms)
16	0.000997	0.084384
32	0.001861	0.06864
64	0.003762	0.087552
128	0.010131	0.126624
256	0.031405	0.202624
512	0.112861	0.350496
1024	0.432856	0.662752



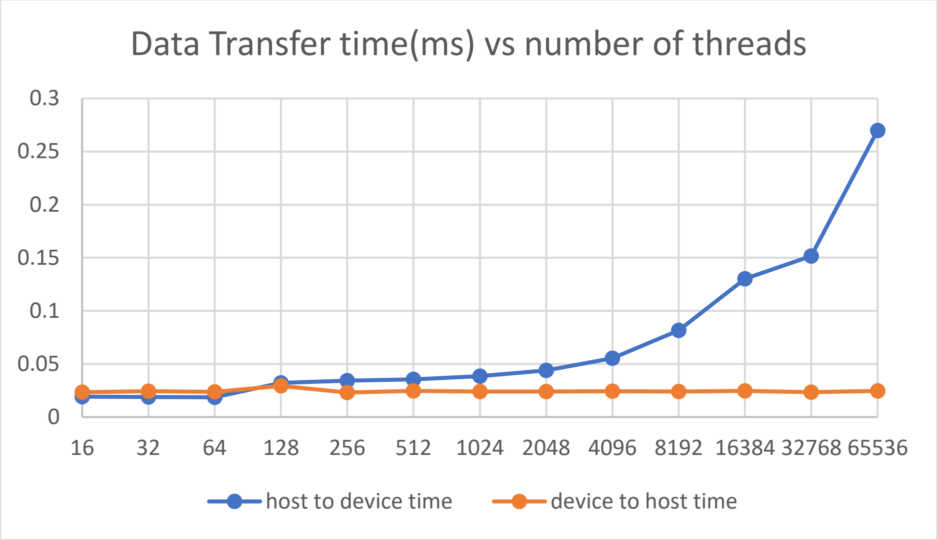
n	CPU execution time(ms)	GPU execution time(ms)
2048	1.687422	1.28272
4096	6.651351	2.520736
8192	26.498434	5.005792
16384	105.700233	10.505472
32768	423.909119	23.902369
65536	1705.446899	72.88797



From 2048 points onwards, the GPU execution time becomes faster than the CPU time.

3)

n	host to device time(ms)	device to host time(ms)
16	0.019168	0.023392
32	0.018976	0.024288
64	0.018592	0.023776
128	0.032064	0.02928
256	0.034304	0.023104
512	0.035584	0.024576
1024	0.038496	0.024032
2048	0.043968	0.023904
4096	0.05552	0.02432
8192	0.081568	0.024
16384	0.130176	0.024576
32768	0.151584	0.023264
65536	0.269984	0.024704


Data Transfer time(ms) vs number of threads
0.3
0.25
0.2
0.15
0.1
0.05
0
16
32
64
128
256
512
1024
2048
4096
8192
16384
32768
65536
host to device time
device to host time