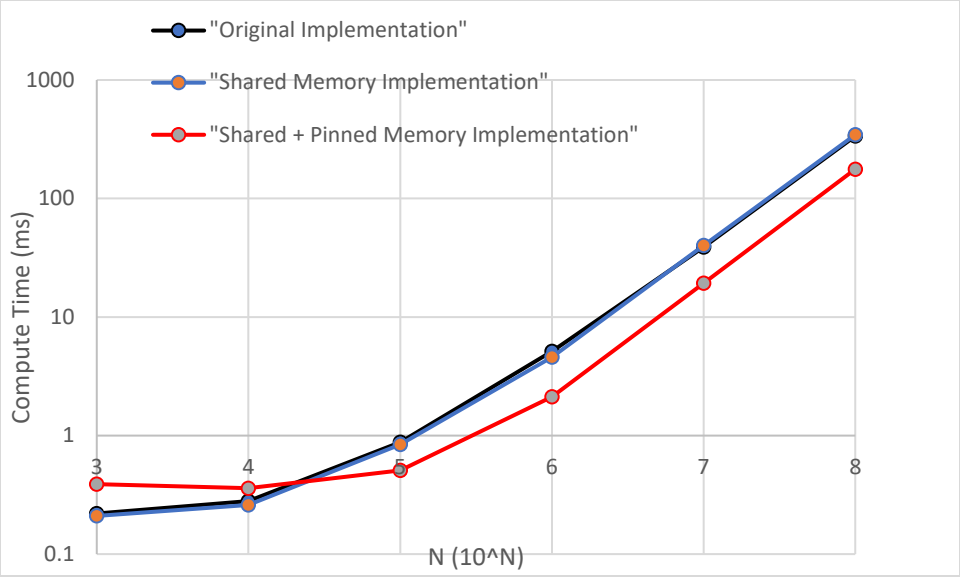

N	orig time	1a time	1b time
3	0.22	0.21	0.39
4	0.28	0.26	0.36
5	0.88	0.84	0.51
6	5.13	4.58	2.12
7	38.86	40.17	19.35
8	336.07	345.1	176.88

In this homework my shared memory implementation was better than the original implementation for small values of N. However when N increased and many blocks were used, there were more overlaps between the blocks so performance suffered. My shared memory implementation was size of 512 floats, and each thread referenced the values in it. However when there were overlaps between neighboring blocks, the thread still has to go to global memory to get the value. In this case the value is pulled twice from global memory so it is not very efficient. It would have been better to use more threads for my shared implementation to avoid having threads reference values in adjacent blocks as often – minimal amount of blocks would be best in this case.. Using pinned memory improved the result even further. It cut the process time in half on the shared memory implementation.