Parallel and distributed Systems
4th assignment

Erika Koro

October 13, 2022

Objective

The objective of this assignment is to calculate the K-Nearest neighbours for all points of a dataset. Specifically, in the first serial source code is constructed the vantage point tree which sequentially is used in the KNN algortihm. In the KNNSearch the construction of the VP Tree per point is used in order to sort the array with the points, according to their median distance from the pivot and chooses the first-k elements as neighbours. In order to parallelize this algorithm, MPI specification for C is used to optimize performance and give a distributed solution. The code for this assignment can be found in here.

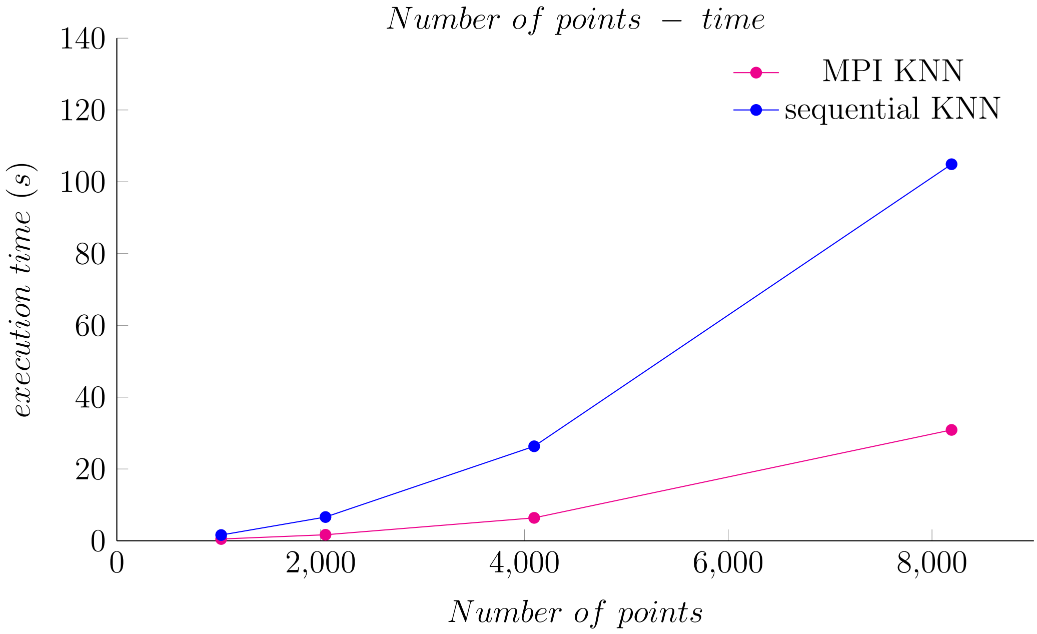
Experiments

The experiments were executed in an AMD Ryzen 7 5800H, 8 cores/ 16 threads, 3200.000 MHz. and the results are based on the number of points and dimension of the dataset and the number of searching neighbours is a power-of-two value with k = $2^{[1:8]}$.

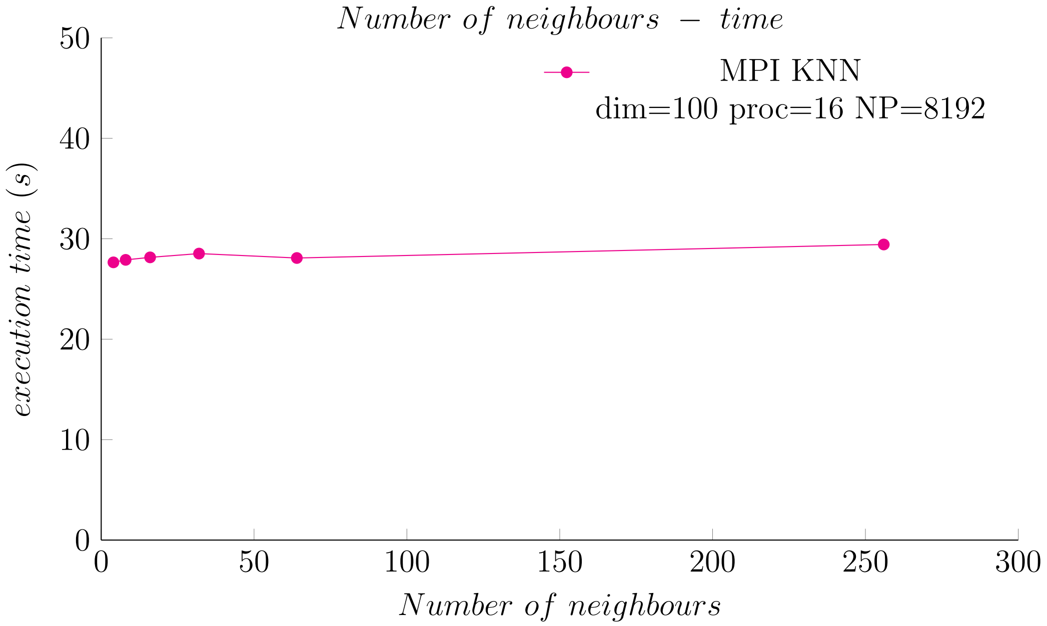
Observations

First of all, by keeping stable the dimension and the number of processes (optimal=16) and increasing the number of points it causes 5 times slower KNN algortihm than MPI on average. Secondly, keeping the number of points, the dimension and the processes stable and increasing the number of k-nearest neighbours keeps MPI's execution time constant.Meanwhile, increasing the number of k increases KNN sequential's execution time. Also, increasing the number of dimension by a factor of two, caused 2-times slower performance in MPI. At last, as it is expected by raising the number of processes from 2 to 16 is observed the optimized performance, a number that can be predicted by the number of available cores.

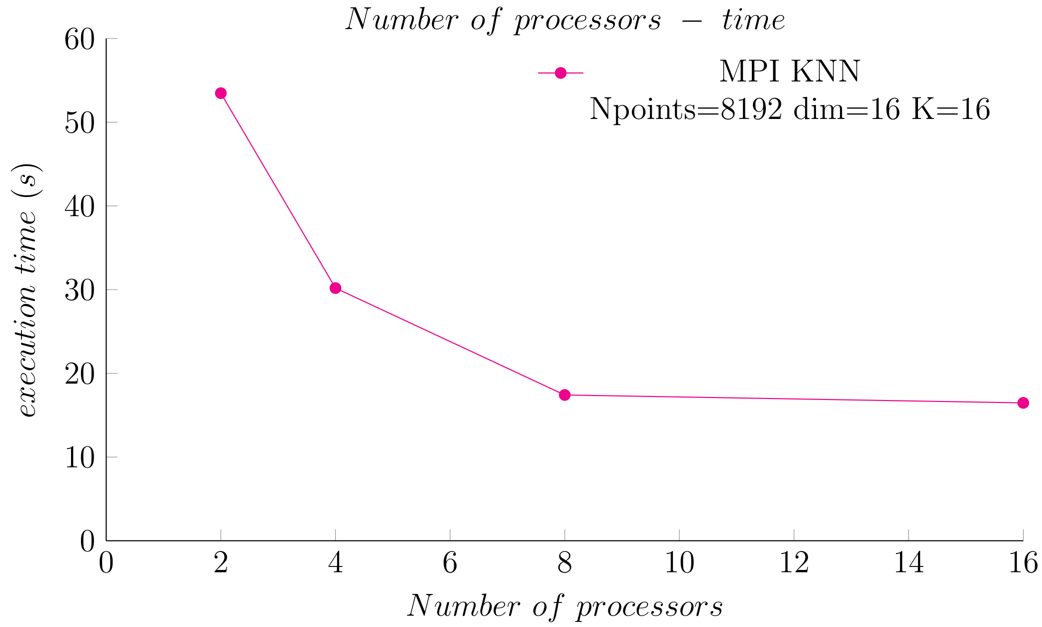
Plots

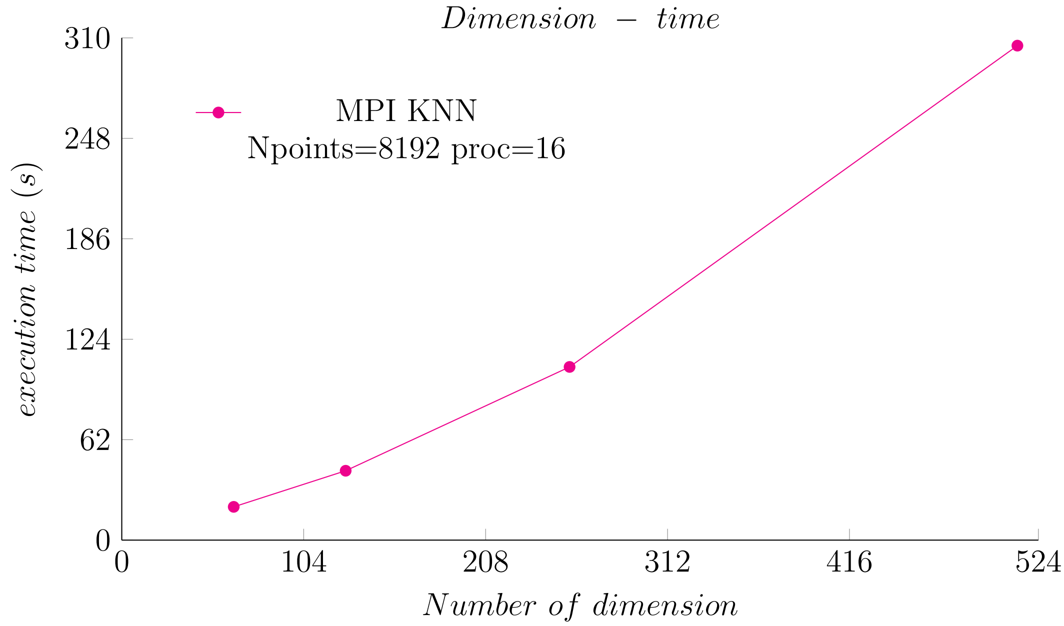


Number of points − time
MPI KNN
sequential KNN
execution time (s)
Number of points
0
20
40
60
80
100
120
140
0
2,000
4,000
6,000
8,000



Number of neighbours − time
MPI KNN
dim=100 proc=16 NP=8192
execution time (s)
Number of neighbours
0
10
20
30
40
50
0
50
100
150
200
250
300


Number of processors − time
60
50
40
30
20
10
0
execution time (s)
MPI KNN
Npoints=8192 dim=16 K=16
0
2
4
6
8
10
12
14
16
Number of processors


Dimension − time
310
248
186
124
62
0
execution time (s)
MPI KNN
Npoints=8192 proc=16
0
104
208
312
416
524
Number of dimension

Problems encountered

Initially, the most important issue I faced was debbuging the MPI code amd eliminating the memory leaks in order to give a bigger dataset but MPI's memory leaks filled my memory. Another problem I encountered, was trying to parallelize the sequential construction of VP tree by using the OpenCilk library but the execution time was equal to the sequential's. I also tried to give it a threshold size which optimized the parallel code for number of points / 4 and then the sequential code was executed. The reason why OpenCilk did not achieve better performance than the serial code was due to the workload per thread although the CILK WORKERS were set at the number of cores.