Galaxy Distribution Problem

András Réka

January 30, 2022

Real data of 100000 galaxies and the same size of randomly positioned synthetic galaxies are given. By calculating the angles between each pair of real galaxies, each pair of real-random galaxies and each pair of random galaxies we can build histograms of the angle distributions. Based on these histograms we can measure the difference (denoted by ω) between the distribution of the real galaxies and synthetic galaxies. The results of our calculations are as many ω values as many bins we have in our histogram. If all ω values are in the range $[-0.5, 0.5]$ we have a random distribution of real galaxies.

Four different solution of the calculations will be given: one sequential and three parallel. The calculations were performed on Dione. Each folder of the repository contains a different solution of the galaxy distribution problem, except the common folder, which contains implementation independent functions (ex. angle distance formula, reading input files). Each solution has it's own c source, makefile, the outputted ω values, and some information regarding the program execution in the log file. The first 5 ω values are:

```
[randras@dione openmp]$ cat omega.out
0.00  : 2.366
0.25  : 1.744
0.50  : 1.418
0.75  : 1.215
1.00  : 1.087
```

We can apply further optimizations by applying the properties of the distance (denoted by D) relation. We know that $D(g, g) = 0$ and we also know that it is a symmetric relation, since $D(g, g') = D(g', g)$, where $g, g' \epsilon G$ and G denotes the set of galaxies. In the consequence of those properties we can reduce the calculations in the DD and RR histograms.

Sequential Version

- run command:

  ```
  srun -N 1 -c 1 ./galaxy_sequential ../common/RealGalaxies_100k_arcmin.txt
  ../common/SyntheticGalaxies_100k_arcmin.txt omega.out >> log
  ```

- run time: 2157.6 secs

OpenMP

- run command:

```
srun -N 1 -c 40 ./galaxy_openmp ../common/RealGalaxies_100k_arcmin.txt
../common/SyntheticGalaxies_100k_arcmin.txt omega.out >> log
```

- run time: 218.2 secs
- Number of threads = 40
- $S_{40} = 9.89$

OpenMPI

- work distribution: there are three main groups of processes based on $ID\%3$. Each group calculates its histogram. Inside the groups the calculations are distributed based on process id's and the total number of processes.
- run command:

```
srun -n 40 --mpi=pmi2 galaxy_mpi ../common/RealGalaxies_100k_arcmin.txt
../common/SyntheticGalaxies_100k_arcmin.txt omega.out
```

- run time: 133.3 s
- Number of processes = 40
- $S_{40} = 16.21$

CUDA

- work distribution: In total we have to perform $2 * N * N$ distance calculations, where N is the total number of galaxies. Threads with $ID < N * N$ calculate the DR histogram. Since we can think of the i, j indexes of two nested loops, where i, j goes from $0 \rightarrow N$ as the Cartesian product of the set $S = \{0, 1, 2, .., N\}$. We can generate the same pairs by using only one index, in this case the thread ID's:

$$\{(id/N, id\%N) \mid id < N * N\}$$

  For the calculation of DD and RR histograms we need $N * N$ times thread in total. We can use the previous process two generate the i, j pairs, but we apply them in a different way. When $i < j$ we calculate the DD histogram, when $i > j$ we calculate the RR histogram, and we terminate the process when $i = j$.

- run command:

```
srun -p gpu -c 1 --mem=10G ./galaxy ../common/RealGalaxies_100k_arcmin.txt
../common/SyntheticGalaxies_100k_arcmin.txt omega.out
```

- run time: 5.3 s
- Thread number in each block:1024
- Size of the blocks in grid: 19531250