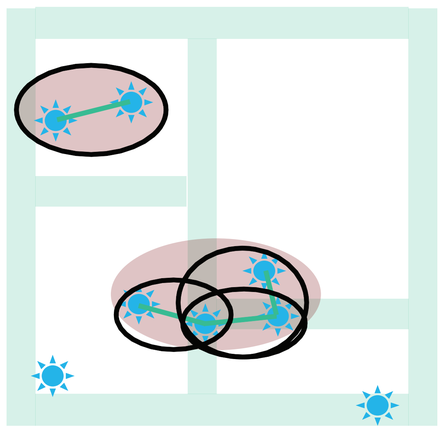
DISC-Finder: A Distributed Algorithm for Identifying Galaxy Clusters

Bin Fu, Kai Ren, Julio López, Eugene Fink, Garth Gibson

We have developed a distributed version of the Friends-of-Friends technique, which is a standard astronomical application for analyzing clusters of galaxies. The distributed procedure can process tens of billions of galaxies, which makes it sufficiently powerful for modern astronomical datasets and cosmological simulations.

Friends-of-Friends Algorithm

- Two galaxies are "friends" if they are close to each other; that is, the distance between them is within a specific global threshold
- The algorithm analyzes an undirected graph, where galaxies are vertices and their "friendships" are edges. It identifies the connected components of the graph, which serve as an approximation of gravitationally bound clusters
- The time complexity is $O((n * \log n)^{1.5})$ for the exact computation, and O(n) for an approximate algorithm

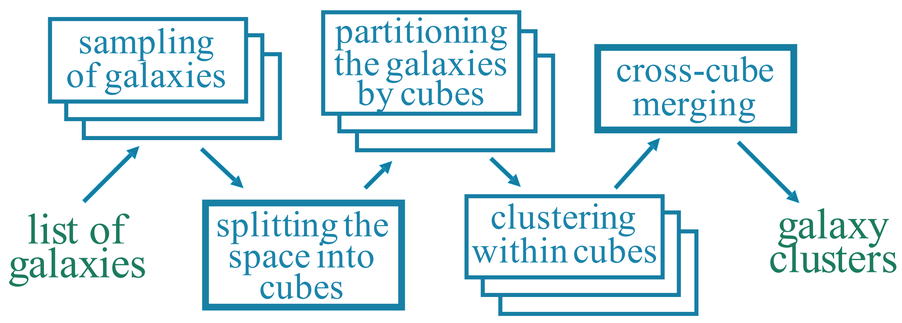


Galaxy clusters and space partitioning

Distributed Procedure

We have developed a Map-Reduce "wrapper" that distributes the Friends-of-Friends computation among multiple cores.

- Divide the space into cubes, where each cube includes about the same number of galaxies, by applying the kd-tree construction to a randomly selected sample of galaxies
- Apply a sequential Friends-of-Friends procedure to find the clusters within each cube
- Identify cross-cube "friendships" and merge the respective clusters, using the union-find algorithm

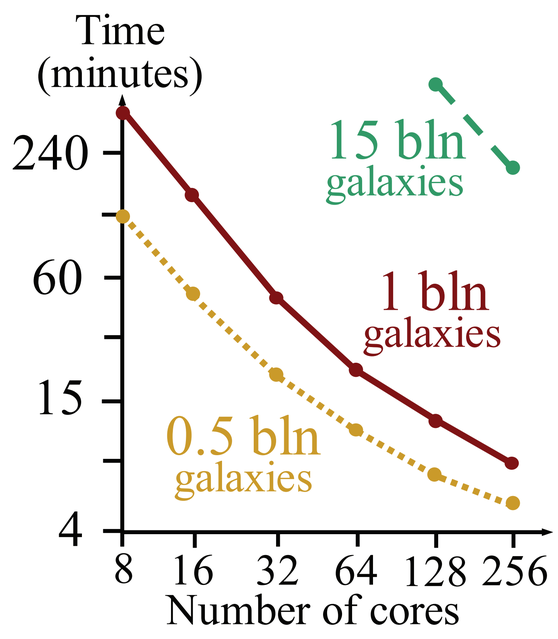


Map-reduce wrapper

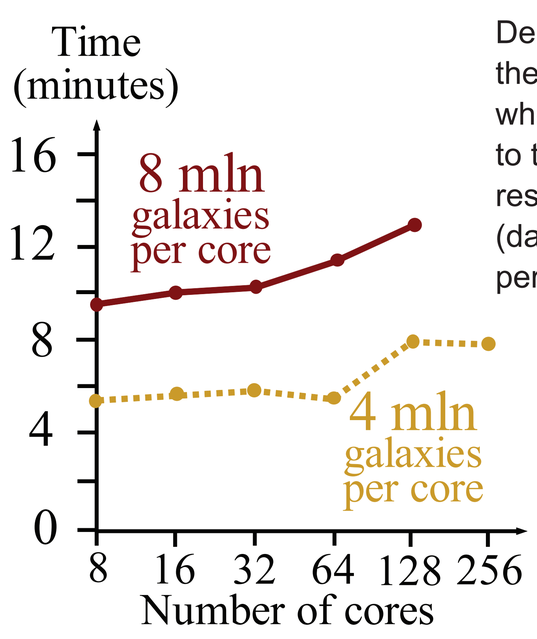
Performance

STRONG SCALABILITY

Dependency of the running time on the number of available cores for 0.5 billion galaxies (dotted line), 1 billion galaxies (solid line), and 15 billion galaxies (dashed line).



WEAK SCALABILITY



Dependency of the running time on the number of available cores, where the input size is proportional to the number of cores. We show results for 4 million galaxies per core (dashed line) and 8 million galaxies per core (solid line).

Carnegie Mellon

Parallel Data Laboratory

SVD10