High resolution simulations for Precision Cosmology

The SIMCODE-DS project deals with the need of high resolution simulations in view of the advent of what is known as the epoch of "Precision Cosmology". The latter term indicates the huge quality leap in the accuracy of observational data expected for the next decade (mostly through large galaxy surveys as the European satellite mission Euclid) that will allow to test the cosmological model to percent precision. As a robust interpretation of such high-quality data will require a large number of cosmological simulations, the community will face in the next years a serious issue of big data storage and sharing.

The Scientific Challenge

Cosmological simulations are an essential ingredient for the success of the next decade of "Precision Cosmology" observations, including also large and costly space missions as e.g. the Euclid satellite. Since the required precision and the need to test for statistical anomalies, astrophysical contamination, parameter degeneracies, etc will require a large number of such simulations, the community is about to face the issue of storing and sharing big amounts of simulated data through a Europe-wide collaboration. In fact, cosmological simulations are getting progressively cheaper as computing power increases, and even for the exquisite accuracy and the huge dynamical range that are required for Precision Cosmology, the main limitation will be determined by data handling rather than by computational resources. Also, while large simulations can now be run in a relatively short time taking advantage of highly optimised parallelisation strategies and of top-ranked supercomputing facilities, their information content might require years of post-processing work to be fully exploited. A typical example is given by the Millennium Simulation (Springel et al. 2005) that is now more than 10 years old but is still employed for scientific applications.

The present Pilot aims at testing possible strategies to make large amounts of simulations data available to the whole cosmological community and to store the data for a timescale comparable with the duration of collaboration such as Euclid (~10 years). The main idea behind the project is that various types of simulations (differing by size, dynamical range, physical models implemented, astrophysical recipes, etc) can be safely stored on a central longterm repository and their content made easily available through metadata and indexing procedures to the community at large, which can range from a small group of collaborators to the whole Euclid Consortium (> 1000 people) depending on the specific nature of the stored simulations.

Who benefits and how?

The Pilot is targeted to the research group of the SIMCODE project led by Dr. Marco Baldi at the University of Bologna. The SIMCODE project is funded with half a million Euro for a period of 3 years to develop simulations mostly targeted to the Euclid collaboration. Therefore, the present Pilot call is mostly relevant to the community of computational cosmologists working in the field of cosmological simulations of structure formation. This kind of research aims at using large supercomputer simulations to investigate how structures in the universe form and evolve, and to make predictions on how such observables might carry information on the underlying cosmological model. In particular, cosmological N-body simulations will be a primary and necessary tool for the broad community working on the preparation of future large survey such as the satellite mission Euclid. The Euclid collaboration is made by more than 1000 members (scientists, software developers and engineers). However, the numerical simulations for the collaboration are performed by a more restricted group of scientists (about 80 people) that collaborate through the Cosmological Simulations Working Group.

The P.I. of the present Pilot call is a member of the latter, and is the coordinator for the implementation of non-standard cosmological models in the Euclid simulations pipeline. The main benefit of a long-term and large storage capacity through a high-level infrastructure will consist in the possibility to store, share, and exchange

simulations data within the Simulations group of the Euclid collaboration, thereby allowing an easier achievement of the collaboration's preparation requirements

Technical Implementation

At the moment we have mostly worked on the data production on various supercomputing facilities in Europe and on the data ingestion into the dedicated storage space provided for the Pilot on the PICO machine at Cineca. So far, about 50 TB of simulations data coming from different simulations suites and different supercomputing centres have been moved to PICO. In particular, we have collected data from the Hydra cluster at the RechenZentrum Garching, from the C2PAP cluster at the Leibniz RechenZentrum, from the Sciama cluster at the University of Portsmouth, from the CNAF computing centre in Bologna, and from the former Tier-0 machine Fermi at Cineca.

We are still running new simulations and we will keep moving data into the machine for the whole duration of the project. The current activity related to data storage and sharing is focusing on devising appropriate ways to pack the data into archive files of a manageable size in order to allow a direct access to the data and on the creation of metadata for these archive files. Ideally, this would lead to the development of a specific pipeline (in shell scripting language) that can be run on various simulations formats and produce the relative archive files in a flexible way. This is under development.

Preliminary Results

The main goal of the SIMCODE-DS project consists in the implementation of a pipeline capable of organising large amounts of data of cosmological simulations in an indexed structure so to allow an easy browsing of the data, a fast path to the relevant portion of data, and most importantly to a dedicated platform to share and distribute the data to a large community of potential users. This is partly already happening with several collaborative projects that are making use of the simulations data which are transferred to specific users in different countries for post-processing analysis. Nonetheless, at the moment this is done on the native data format, which means that individual files have to be selected and transferred (normally by the person who produced them). Ideally, this process should be made easier once the archiving pipeline is in place.

Contacts

- Marco Baldi, Bologna University, marco.baldi5(a)unibo.it

Furhter Information

Scientific Challenges behind the Pilot
To learn more about the EUDAT/SIMCODE-DS collaboration visit here

Read more