IL NUOVO CIMENTO Vol. 32 C, N. 2 Marzo-Aprile 2009
DOI 10.1393/ncc/i2009-10372-1

Colloquia: CSFI 2008

HPC, grid and data infrastructures for astrophysics: An integrated view

F. Pasian

INAF - Information Systems Unit and Osservatorio Astronomico di Trieste
Via G. B. Tiepolo 11, 34131 Trieste, Italy

(ricevuto il 15 Maggio 2009; pubblicato online il 6 Agosto 2009)

Summary. — Also in the case of astrophysics, the capability of performing "Big Science" requires the availability of large HPC facilities. But computational resources alone are far from being enough for the community: as a matter of fact, the whole set of e-infrastructures (network, computing nodes, data repositories, applications) need to work in an interoperable way. This implies the development of common (or at least compatible) user interfaces to computing resources, transparent access to observations and numerical simulations through the Virtual Observatory, integrated data processing pipelines, data mining and semantic web applications. Achieving this interoperability goal is a must to build a real "Knowledge Infrastructure" in the astrophysical domain.

PACS 95.75.Mn – Image processing (including source extraction).
PACS 95.75.Pq – Mathematical procedures and computer techniques.
PACS 95.80.+p – Astronomical catalogs, atlases, sky surveys, databases, retrieval systems, archives, etc.

1. – Recognition of the importance of e-infrastructures at European level

It is to be noted that e-infrastructures are absolutely necessary for the development of science, but they do not come for free, and cannot be given for granted. This is well understood at the European level, and e-infrastructures are an integral part of EU's policies.

ESFRI, which is a multi-disciplinary initiative for the implementation of research infrastructures, has focussed on the need for networking, capability and throughput computing, grid architectures, software, data management and curation as the main priorities.

ASTRONET is another initiative, dedicated to Astronomy and Astrophysics, aimed at defining the scientific priorities in our discipline for the next two decades, and the corresponding priorities in the allocation of resources. The ASTRONET working groups have recognized e-infrastructures as must-haves to tackle the challenges of the future: computing (both capacity and capability), theory and simulations and the Virtual Observatory [1], together with laboratories, must have priority over the rest, because without them science cannot be made.

2. – Requirements for e-infrastructures in Astrophysics

Astrophysics, considering it to be the sum of a set of branches (astronomy, radioastronomy, cosmology, space physics, planetology, solar-terrestrial physics, etc.) is mainly an observational science. The progress in sensor technology has led to an exponential growth in the data volume. The first reason to need computing power is therefore *to process (i.e. reduce) observational data*.

All observations are kept; as a matter of fact, each observation is unique since many of the astrophysical phenomena are time-dependent. As a consequence, there is a continuous increase of the data available. This implies *the need for publicly accessible archives*.

Modern astrophysics is increasingly interested in multi-wavelength studies, and in analyzing time variability, studying supernovae progenitors, etc. There is therefore *the need to harmonize the data archives world wide*, creating what is known as the *international Virtual Observatory*.

An additional reason calling for an increasingly higher computing power is the desire to *perform data mining and statistical analysis on archived data*.

Astrophysical phenomena are usually modelled to be better understood. Computing power is therefore needed *to perform numerical simulations*.

It is then necessary to compare the data obtained from an observation, those retrieved from archives and the simulations. *Matching observed data with numerical simulations* is an additional reason why computing power is necessary.

3. – e-infrastructures for Astrophysics, today and tomorrow

Currently, astrophysicists use a wide range of facilities: PCs, local clusters, advanced computing (Grid, HPC centers), exploiting their computing and storage capacities. Plus, of course, there is a heavy usage of the research network. Astronomical institutions (INAF and Universities) participate actively in the national and international initiatives in the field.

With reference to HPC, it is to be noted that a MoU with CINECA has been active since 2002, and has been renewed in December 2007 until the end of 2010. INAF is one of the contributors to the Italian Supercomputing and Grid Initiative roadmaps, ISI and IGI, both proposed by the Italian research community to the Ministry of University and Research (MiUR). It is furthermore to be noticed that the community has successfully participated in the selection of projects for the DEISA Extreme Computing Initiative (DECI).

The community has also participated in the national PON initiatives for (super)computing resources, and is extensively using campus and geographically distributed Grids for a number of tasks (numerical simulations, stellar evolution models, simulations of the Planck mission, etc.). Italian astronomers also lead the Astronomy & Astrophysics cluster of EGEE-III, and lead the Database WG within the project itself. They are also active in developing mechanisms to integrate databases within the Grid middleware.

In the field of the “data Grid”, Italy participates in the international initiatives for the Virtual Observatory. Italian astronomers hold leadership positions in the International

Virtual Observatory Alliance (IVOA Vice-Chair and Vice-Chair of the Theory Interest Group). They furthermore participate in three projects funded by the EU Framework Programmes: VO-TECH, EuroVO-DCA and EuroVO-AIDA. In particular, the community holds the leadership in the integration of the Virtual Observatory with numerical simulations and the Grid.

The "big science" challenges in astrophysics call for an expansion of the computing infrastructures—and of network and data management infrastructures as well. The community, and INAF in particular, is interested in using, and participating in defining, competitive computing infrastructures so to let its know-how in the field grow. But there is furthermore the desire to integrate the network, data and computing infrastructures, or at least to let them interoperate.

4. – Integrated/interoperable e-infrastructures

To offer scientists a useful service, all of the components of the informatics infrastructure need to be thought as integrated, or at least fully interoperable. In other words, the various infrastructure components (applications, computing, data) should interact seamlessly exchanging information, and be based on a strong underlying network component [2].

Considering first the computing infrastructure, the facilities available to the scientists can be very different: from the laptop or desktop PC to the HPC center, passing through local clusters and "the Grid". From the user perspective, ideally, all facilities should be seen homogeneously; in reality, they all tend to have different access modes. As a result, users find obstacles to their ideal exploitation.

As for the data infrastructure, again from the user perspective, there is the need to transparently and homogeneously access a wide variety of data (multi-frequency and multi-instrument observations and numerical simulations). The World-wide Astronomical Virtual Observatory is progressively fulfilling this requirement by providing seamless access to data centers and facilities through its standards, but this is only the beginning of the story. User queries could be expressed in natural language, or a user query may imply, besides a data infrastructure, the implicit use of applications and computing resources. This, again, requires tight integration of the various infrastructure components.

The first thing to do is to enhance the basic infrastructure: network and computing resources. The GARR Network needs to be improved by increasing the bandwidth and eliminating the bottlenecks; in particular, last mile connections to the individual local sites need to be upgraded. At the same time, resources must be found to allow Italy to harmonise with the European plan for HPC by allowing to build at very minimum a national Tier-1 facility, aiming at a PetaFlop Tier-0 system.

Not in contrast with this approach, the Grid concept should become far more widespread than currently. The old perception of the Grid re-using the idle cycles of a set of PCs (as was in the SETI@home experiment that originated the Grid concept) is nowadays completely misleading—from the computing and storage viewpoints, the Grid has now the possibility of linking powerful clusters and even HPC hardware, provided it uses the proper middleware.

The long-term goal is to allow complete interoperation of HPC centres: initially, Tier-2 (regional) facilities shall be integrated with the Grid, followed by full integration of local clusters ("Tier 3") and, as final goal, achieving inclusion of Tier-0/1 within a common scheme. In any case, a necessary step is to achieve as soon as possible the full compatibility between HPC and Grid at least at the User Interface level.

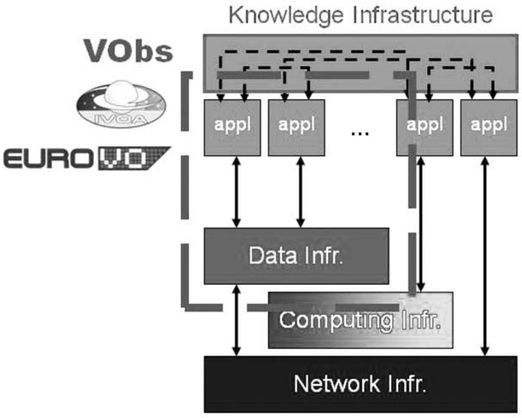

Fig. 1. – The different layers of the integrated ICT infrastructure for astrophysics. Scientific applications may interact with the Data Infrastructure, or with the Computing Infrastructure, or directly with the Network Infrastructure (each interacting with each other). The interoperation of specific applications creates a Knowledge Infrastructure. The dashed box defines the area where the Virtual Observatory is expected to act.

There is a further point. Up to now, the Grid has mainly delivered computing power, the main issue that its implementers, mostly involved in large High-Energy Physics experiments (*e.g.* at CERN), needed to solve. Accessing data, and databases using the Grid paradigm is a step still to be improved. Some activities are being carried out, also with the participation of INAF, within EGEE. This is of course a further step in the direction of interoperability.

To fulfill this requirement applications, computing power, data repositories and databases holding metadata or catalogues should be accessed as a single utility. We may call it Grid, or Cloud (as is now fashionable), or Virtual Observatory (which may be more familiar to our community), ... It does not really matter, provided the goal is the same.

The far-end goal, in any case, is achieving the full interoperation of the relevant applications (data processing, analysis, mining, bibliography, semantic web) to start building a Knowledge Infrastructure for Astronomy, as shown in fig. 1.

* * *

The INAF-CINECA agreement is funded by INAF and managed by the Information Systems Unit of the INAF Projects Department. The support of its Director, G. Vettolani, to the participation of INAF in the IGI and HPC-ISI proposals to the MIUR roadmap is acknowledged. The author wishes to thank M. Mazzucato of INFN for information on Grid initiatives, G. Erbacci and C. Gheller of CINECA for information on the status of HPC.

REFERENCES

[1] Solano E., *Lect. Notes Essays Astrophys.*, **2** (2006) 71.

[2] Pasian F., *(Super)computing within integrated e-Infrastructures*, in *Computational Astrophysics in Italy: Results and Perspectives*, edited by Becciani U. (*Mem. Soc. Astron. It., Suppl.*, **13** (2009) 117).