The Virtual Observatory - A New Era for Astronomy

Reinaldo R. de Carvalho⁽¹⁾, Roy Gal⁽²⁾, Haroldo C. Velho⁽³⁾, Hugo V. Capelato⁽¹⁾, Francesco La Barbera⁽⁴⁾, Eduardo Charles⁽³⁾, Renata Rocha⁽³⁾, João L. K. Moreira⁽⁵⁾, Paulo A. Lopes^(6,7) and Marcelle Soares-Santos⁽⁸⁾

ABSTRACT

We present an overview of a new paradigm in astronomy, the Virtual Observatory (VO). We trace it from the early developments only a decade ago to its current state. In a few years, astronomy will have accumulated an unprecedented amount of data, on the order of 100 Pb, and adding 2-4 PB/year. This is an astonishing five orders of magnitude higher than it was in 2000. The VO is a response to the astronomical community's demands for improved and homogenized access to this data, combined with the tools to manipulate and explore them. It is a complex enterprise with a decentralized, webcentric nature, implying that astronomers need to rethink the old ways of conducting their scientific programs. Most projects related to the VO started in the late 90's and today an international effort is coordinated by the International Virtual Observatory Alliance (IVOA). In Brazil the National Institute for Science & Technology (INCT-Astrophysics) recently created by the Ministry of Science & Technology (MCT) is taking the lead in VO development (BRAVO - BRAzilian Virtual Observatory). At the National Institute for Space Research (INPE), we are concentrating our effort (BRAVO@INPE) on three distinct aspects of the VO development: 1) Database Development and Basic Infrastructure; 2) Data Grid & Processing Grid; and 3) Data mining. This paper describes our view in setting a roadmap for VO in Brazil and a few technical developments on which we have already embarked.

- (1) Instituto Nacional de Pesquisas Espaciais (INPE/DAS), Brazil
- (2) Institute for Astronomy, University of Hawaii, USA
- (3) Instituto Nacional de Pesquisas Espaciais (INPE/LAC), Brazil
- (4) Osservatorio Astronomico di Capodimonte, Italy
- (5) Observatório Nacional/MCT, Coordenação de Astronomia, Brazil
- (6) Observatório do Valongo, UFRJ. Brazil
- (7) IP&D Universidade do Vale do Paraíba, Brazil
- (8) Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, Brazil

Keywords: Learning Machine Algorithms, Virtual Observatories, Cyber Infrastructure

1 Introduction

Astronomy is now an enormously data-rich science, and currently produces terabytes of raw data per day, with a few petabytes already in various archives. Both the data volume and data rate are increasing exponentially, with a doubling time of ~ 1.5 years. Even more important is the growth of data complexity (expressed, e.g., as the dimensionality of the parameter space spanned by the measurements of the detected sources) and heterogeneity. These data are now being federated in a global data grid under the umbrella of the Virtual Observatory (VO) paradigm. A complete and effective scientific exploitation and exploration of these large and complex data spaces is a highly non-trivial task, requiring a new generation of software (databases, scalable data mining tools, interfaces), hardware (computing power, storage, network infrastructure), and expertise. The absence of these resources is a key bottleneck in data-rich astronomy: the data are there, but the means of extracting knowledge from them are not.

The diagram below demonstrates the severity of these problems. We see the rapid increase in data volume from only a decade ago, where the Digitized Second Palomar Observatory Sky Survey provided single-epoch observations of half the sky in just 3 bands, to current projects like Pan-STaRRS1, which provides imaging of three-quarters of the sky in 5 filters but at hundreds of epochs. The addition of the time domain not only grows the storage and computational requirements, but challenges the community with the need for new algorithms and tools. Incredibly, we see that astronomy is generating data at the same pace as experiments in particle physics. This is extraordinary, considering that the number of researchers and the worldwide financial investment is much less in astronomy. Figure 1 clearly exhibits the necessity of efficient data storage, data processing and data mining, which are specific topics covered by this project.

Typical research paths taken in the scientific exploitation of large sky surveys are either construction of statistical samples of objects or populations of interest (e.g., normal galaxies, guasars, etc.) and their study (e.g., to probe their evolution, large-scale structure, etc.), or selection of interesting targets (e.g., peculiar galaxies, distant guasars, brown dwarfs, supernovae, etc.) for follow-up observations. The scientific potential of such studies is greatly enhanced by federating data sets (e.g., combining optical, infrared, and radio sky surveys), which often reveal important features and populations of objects not easily distinguishable in any of the data sets taken separately. For example, a typical VO-data-enabled project would be a complete clustering and correlation analysis of combined source catalogs, using a federation of multi-wavelength data from several major astronomical surveys, ranging from radio, through infrared, optical, UV, to X-ray, or even y-ray. Data federation of the source catalogs from these surveys generally results in a parameter space of ~ $10^8 - 10^9$ data vectors in ~ $10^2 - 10^9$ 10³ dimensions. The existing tools and algorithms do not scale well to such hyperdimensional data sets, so we must assemble, test, improve, and deploy the necessary data mining, statistical, and visualization tools for this exploration. Concurrently, we must develop the necessary computational and network infrastructure and human

expertise to develop, implement, and utilize these tools. Examples of specific challenges will be presented later in this paper.



Figure 1 - The rapid increase of astronomical data, considering only the most important optical surveys carried out in the past 20 years. For comparison we show the data rate from the Large Hadron Collider experiment at CERN (for details see <u>http://lhc.web.cern.ch/lhc/</u>).

The main objective of the BRAVO@INPE project is to address these strategic issues. More specifically, the BRAVO@INPE project intends to generate investment in information technology, with particular emphasis on Computational Infrastructure, Data Grid, Data Processing, and Data Mining. We present not only a brief history of what has been done in the recent past but also elucidate the specific needs for the near future. This effort aims to prepare the Brazilian astronomical community for the avalanche of data and massive data processing needs that are a reality now, and which will increase rapidly in the coming years with the advent of the large telescopes and surveys currently under development (GMT, TMT; LSST, Pan-STaRRS, VISTA, VST).

This paper is organized as follows: Section 2 outlines the general concept of the Virtual Observatory, establishing a context for the more specific components described later. Section 3 describes the initial stages taken in generating a roadmap for the VO in Brazil, while Section 4 introduces two of the basic elements of any VO: computational infrastructure and databases. The fundamental concepts of data grids and processing

grids are presented in Section 5. Section 6 describes an image processing pipeline developed by our group, 2DPHOT, and the main characteristics of the Astro-Wise environment. Section 7 provides an overview of the ongoing developments within Brazil in terms of astrophysical applications. Section 8 reviews data mining and describes specific projects we are undertaking in this field. Section 9 focuses on the four main areas in which we plan to invest resources, while Section 10 summarizes the BRAVO@INPE project.

2 The VO Concept

For more than two decades, the international astronomical community has witnessed an exponentially growing capacity for accumulating astronomical data. Today, information is gathered in large surveys from the ground and from space, covering virtually the entire electromagnetic spectrum, from X-rays through the ultraviolet, optical, infrared, and beyond. Individual projects yield complementary data through specific, targeted scientific programs. Much of these data are made available to the community through public servers, usually in several different formats, and distributed at many institutions. The data quality, metadata, interfaces, and accessibility are heterogeneous, since each project typically curates its own data, presents it in a custom database, and even data formats in astronomy are instrument dependent with little effort made to unify them.

An underlying concept of the VO is that by providing improved and homogenized data access combined with the tools to manipulate and explore the data, the need for new observations will be reduced even as the scientific output is increased. All gathered data, even those that are initially proprietary, can be accessed via the VO, enriching the international community. Large surveys would take precedence over individual, targeted observations, providing added coherence to the VO structure. Therefore, the VO is not an enterprise driven by a single institute or even one country. It is rather a community proposal aimed at the democratization of information that will certainly expand to other scientific areas like meteorology, geophysics, and space science, allowing new interactions and the exchange of methods and technology. Thus, the VO today represents to the astronomical community what the Internet was for the academic world in the 1980s. It is clear today that science, especially in developing countries, would be shockingly different without the Internet in the same way that we envisage in the future saying that astronomy would not be the same without the VO.

The Virtual Observatory (VO) concept is the astronomical community's response to the scientific and technological challenges posed by massive and complex data sets. To exemplify the obstacles to dealing with a modestly large amount of data, its complexity, and the challenge of processing it over a reasonable timescale, we examined the reprocessing of reduced galaxy images from the seventh and final data release from the Sloan Digital Sky Survey (SDSS DR7). These data cover ~8400 square degrees on the sky and providing images in four bands. We consider only the re-measurement of photometric parameters using a custom pipeline (described later) - not the reprocessing of raw data to calibrated images, and ignore the spectroscopic data entirely. Even for this modest task, if we want to process the imaging data in one week, we would need

6577 processors, which is a factor of 15 more computing power than everything currently available to Brazilian astronomers. We use a timescale of one week as an upper limit for what a user would accept to retrieve important information from such a large data set - and this dataset is almost trivial compared to upcoming surveys.

Computational hardware requirements are just one small part of the issues that arise when dealing with such vast datasets. Processing takes a lot of time, so once completed, it is of paramount importance that querying and retrieving data be done quickly. This requires investment in not only database software, but the astronomical and computational expertise to design and implement efficient and scientifically useful data models. This information, once structured in such a database, needs to be retrieved efficiently, demanding high-speed internet connections to which most research centers in Brazil do not have access. For these reasons, our top priorities include implementing grid computing to enable the processing of massive datasets; creating a dedicated network for astronomy to enable access to the resulting data, and training astronomers and computer scientists to develop these tools to produce cutting-edge science. Figure 2 summarizes this critical situation.



Figure 2 - Diagram showing the complexity of implementing a VO structure when considering all the infrastructural elements and attending the demands of the astronomical community, which expands fast and will grow beyond the computational resources currently available, especially in Brazil.

3 BRAVO: Brazil's First Steps Into the VO

The new era of large data sets and the co-requisite data processing needs led to the recognition two years ago that we must modernize the tools for astrophysics in Brazil. In addition to the large photometric and spectroscopic surveys being carried out in both hemispheres, Brazil has committed significant resources to new facilities (including SOAR, Gemini, BDA, etc). As a result, we have access to extraordinary amounts of data in all portions of the electromagnetic spectrum, but without standard techniques for storage, retrieval, distribution, processing or analysis. Thus, the underlying concept of BRAVO@INPE is to federate these resources, using a common framework, standard interfaces, computational infrastructure, and analysis tools.

Before embarking on a major enterprise to develop BRAVO@INPE, we must understand our current hardware/software/personnel resources and their ability to meet our needs both today and in the future. The partners in this project are the thirty-one institutes comprising the INCT-Astronomy (National Institute for Science & Technology) recently created by the MCT (Ministry of Science & Technology). The central repository of knowledge about computational hardware, software and personnel in BRAVO@INPE will be the BNPGA (Brazilian Network for Processing Grid in Astronomy). Appendix A lists the institutes that will compose the BNPGA and their representatives. This new program will allow us to trace the roadmap of what is really needed for the future.

We have performed an initial census of the capabilities of the INCT-Astronomy member institutes. Here we provide a brief synopsis of the results; the complete list of questions and results can be found in Appendix B. We find that most users of our community have access to at least a desktop computer with moderate computational capacity. This conclusion must be seen with caution, as many members of our community use data of low complexity and in small volumes. This situation is changing dramatically with the next generation of large surveys and telescopes. In this context, the current computational facilities may be adequate today, but it is clear that the current cyberinfrastructure will be obsolete when dealing with the extremely large amount of data coming from both stellar and extragalactic projects.

These new programs will often require large computing clusters. We examined access to modern servers with more than 8 processors each (Class A) and to beowulf types, composed of mono-processed nodes and internal networks of 100 Mps (Class B). Only 12 out of the 20 institutes that responded have access to a cluster and only 7 out of these 12 have access to a Class A cluster. It is important to note that in some cases the clusters are shared with researchers from different disciplines like Physics since the small groups of researchers developing Astronomy in Brazil are contained within large Physics departments.

Adding up all the available processors in the different clusters gives, in principle, the total number of processors available for grid processing (see App. B, Table 4). This total, 419, is only 6% of the required number to processing the entire SDSS DR7 in one band, in one week, for example. This is only a crude estimate considering that all the

processors are different, some better than others - fifty are old types of processors that would add little to the total processing capacity. Even more disturbingly, modern surveys like Pan-STARRS use a single 512-node cluster with the latest 3+GHz processors to analyze their data - more powerful than all of the Brazilian astronomical community's computers combined.

Beyond processing power, the total disk storage available to our clusters is approximately 45 Tb. While this satisfies the needs of individual groups, it is clearly incompatible with the needs of the coming decade where telescopes will produce data at a rate of 2 Pb/year. Moving any fraction of such data quantities also requires highspeed network connections, which many of our institutions still do not have.

The results of this census demonstrate the extreme deficiency of the current hardware, software and network infrastructure in Brazil. An often overlooked (and underfunded) aspect of any computational project is the need for personnel with expertise in all aspects of the program. In BRAVO@INPE, we cannot expect a computer scientist with experience in commercial database applications to understand and implement astronomical databases without new training. Similarly, we would not expect an astronomer to develop efficient computational algorithms for, say, clustering analysis, without learning about recent advances in such applications.

To address some of these issues, we organized two workshops in 2007 at INPE where formal presentations were released by well known researchers working on e-science, including Dr. George Djorgovski (Caltech), Roy Williams (Caltech), and Robert Hanisch (STScI). At the second workshop we had specific presentations by researchers from Brazil engaged in VO-related projects, showing the tremendous potential that we have to actively participate in this international effort (see www.ivoa.net). For more information and access to the presentations of the lectures see www.lac.inpe.br/ projetos/bravo/. A major component of this project will be the training of technical staff, which is of paramount importance for us. We have already begun a program of visits by Brazilian astronomers and computer scientists to foreign institutions with extensive astronomy database involvement, including Caltech, Johns Hopkins University, and the Institute for Astronomy in Hawaii.

4 Database Development and Basic Infrastructure

Many scientists, including those of you reading this document and those of us writing it, think of databases as a just some place to store a simple table of data, like an Excel spreadsheet. Perhaps you could do simple computations on those columns on your personal computer and output the results. Typical tables might have tens to thousands of entries. Even a large food market has only about 50,000 different items available - and we are tempted to imagine that their warehouse database must be large and complex. We would be very wrong.

Astronomical datasets have far surpassed the largest commercial databases in size and complexity. Almost twenty years ago the Digitized Second Palomar Observatory Sky

Survey database contained over 100 million objects, measured in three bands, with a total of ~100 properties per object. Information about the survey (calibration, plate metadata, related CCD imaging, classification schemes) was spread over ~50 different tables, which often had to be cross-referenced. This database never became easily accessible to the public, which would have required the creation of an added layer of interfaces and query tools.

The current "gold standard" of databases in astronomy is the Sloan Digital Sky Survey. The imaging catalog has almost half a billion objects, in five filters, with nearly 500 columns of data on each object. While the volume of this single table (many Tb) is itself daunting, the SDSS database has nearly 100 unique tables, with an additional 50 views offering easy access to scientifically useful subsets of specific tables. The complexity of this database required years of consideration to design a workable schema, decide on which columns to generate indices to speed queries, understand how to load and update tables with new data, and how to provide public access. Just writing a portion of the table documentation was a full time job for a postdoctoral researcher for almost two years. Beyond the nearly 20Tb of catalog data, SDSS also allows users to access a comparable volume of images.

While one is tempted to believe that SDSS is as complicated as astronomical databases could become, it is far from the truth. Upcoming surveys such as Pan-STaRRS and LSST will yield a comparable amount of data - every time they survey the sky. These projects will create a new "SDSS" every few months. Not only do they produce multifilter imaging, which must be processed, cataloged, stored, and distributed, they will also produce time series. Every object detected in one image must be matched to its corresponding detection in all earlier images of that same area. Optimal methods for differencing images must be developed to look for astronomical sources that vary or move. An entire pipeline is necessary to take moving objects, find them at different locations in images taken at different times, associate them, and generate orbits. Light curves for both stationary and moving objects must be created. All of this must be done almost instantaneously, because rare, one-time events such as supernovae must be found and notifications for follow-up observations disseminated before they fade. This means processing one gigapixel image every minute. The resulting database is correspondingly more difficult to model and populate. A "static" sky database must be created with everything detected, and updated as repeated observations allow for the creation of ever deeper images. Variable and moving objects must have all of their detections stored so that light curves and orbits can be derived.

The evolution of these surveys vividly demonstrates that we must contend with a new paradigm in astronomy. We must have the resources to store, disseminate and access large databases. We must have the knowledge of how such databases are structured, and how we can develop our own tools to create novel science. We must have our own databases for Brazilian programs, and enable interoperability with VO tools to maximize the scientific potential. We must also remember that so far we have only discussed large optical surveys. Multi-wavelength and multi-epoch studies demand new tools to cross-identify sources observed across the electromagnetic spectrum, with different

spatial and temporal resolution. This fundamental problem too has been approached but is far from solved. In all of these arenas, Brazil has much to learn, but also much to contribute.

5 Data Grid & Processing Grid

We are in the midst of a revolution in data gathering that encompasses all realms of science. In particular, the volume of data in astronomy, both real and simulated, is growing exponentially. The need for tools to analyze these data is naturally creating a new branch of scientific investigation - *data science*. There are many challenges in this emerging enterprise. We must have methods for extracting knowledge from large amounts of data, which by itself in non-trivial. What is important and what is noise? Which correlations are fundamental and which are secondary? In addition, we must urgently develop the skills and tools for processing these data. These requirements are already being addressed by two areas of computer science: data mining and high performance computing (HPC). We discuss the former in Section 8; here we will focus on the latter.

There are many approaches to HPC. The first direction utilized parallel machines (vector machines, multi-processing machines with shared memory, multi-processing machines with distributed memory, and more recently multi-core processing chips). These solutions aimed to improve the processing capacity of a single, central machine. By the late 1990s, a form of distributed computing was created, using internet connections among geographically distributed processors to spread the computational labor. This is the underlying concept of grid computing, where processors across a city, country, or the whole world can be shared by a single program. This type of grid is a new environment for the science of the current century.

There are many types of grids and generally they can be classified according to:

- 1 the nature of the processing: data grid or processing grid;
- 2 the focus of the processing: open vs. closed or general vs dedicated;
- 3 the hardware components: homogeneous or heterogeneous

Since the accumulation rate of data in astronomy is already reaching an unprecedented level of 10 Pb/year, it is becoming difficult, technically and financially, to centrally store all of the data, and impossible to replicate data for personal use. A data grid provides an environment for distributing, sharing, and modifying large amounts of data. We find applications for such a grid in different fields, such as meteorology. Examples include the Earth System Grid (http://www.earthsystemgrid.org/) and SegGrid (http:// seghidro.lsd.ufcg.edu.br/). In astronomy, the Montage software platform (http:// montage.ipac.caltech.edu) has been prepared to run in a grid environment.

Similarly, increasingly large and complex processing tasks are required to process and analyze these data sets, or to generate large simulations. A processing grid addresses

these issues. Collaborative processing is a type of application that exemplifies this new technology. The SETI@home project (Search for Extra-Terrestrial Intelligence, http:// setiathome.ssl.berkeley.edu/) is the largest popular distributed computing project, with over 3 million users, hosted by the Space Sciences Laboratory (University of California, USA). It is only one of 50 such projects using the BOINC volunteer and grid computing platform (http://boinc.berkeley.edu/). A UK project called AstroGrid has been developing VO-compliant tools that operate on a grid. Processing grids in other fields include protein folding (http://folding.stanford.edu/), climate change (http:// www.climateprediction.net/), and seasonal mesoscale climate prediction GBRAMS (http://www.cptec.inpe.br/brams/gbrams.shtml) and RECLIRS (http:// yule.lacesm.ufsm.br/nucleus332/).

Thus, it is imperative that we take advantage of the computing resources available at different computer centers linked by fast network connections. This could take the form of our own, internally developed grid implementation, or the installation and deployment of existing tools such as BOINC. Today, Brazilian users contribute almost 10 Teraflops of computing power to BOINC projects, the highest in South America. One of the main goals of BRAVO@INPE is to create a processing grid to harness academic computing along with this private processing power, with initial focus on two specific astrophysical applications: image processing with 2DPHOT (described in Section 3); and analysis of cosmological simulations with hundreds of millions of particles using the FoF algorithm (described in Section 4).

Our team is currently strongly engaged in the use of grid technologies and web services within the VO context. Specifically, our focus is data modeling, within the scope of BRAVO@INPE, to develop a framework for the metadata describing both observed and simulated data. We examine the logical relationships between these metadata, with the intent of establishing a general architecture for retrieving, processing, and interpreting data from different branches of spatial science and in particular from astronomy. This is an important step for constructing protocols that will guide VO applications.

6 Data Processing

As described above, critical issues in VO development include the large amount of data and how it is to be processed - taken from raw images to reduced data suitable for further analysis. Here, we describe two concrete steps to address these problems undertaken in Brazil: the installation and operation of the first Astro-Wise (AW) node in South America and the insertion of our photometry environment (2DPHOT) into AW. 2DPHOT is an automated tool to obtain both integrated and surface photometry of galaxies in an image, to perform reliable star-galaxy separation with accurate estimates of contamination at faint flux levels, and to estimate the completeness of the image catalog. A 2DPHOT graphical user interface (named 2DGUI) is also under development, allowing the user to easily set 2DPHOT input options and detection parameters. More details can be found in La Barbera et al. 2008a. We show a schematic representation of the 2DPHOT environment in Figure 3.



Figure 3 - Schematic representation of the 2DPHOT environment

The main tasks of 2DPHOT are:

- 1 Producing a cleaned catalog of the image.
- 2 Performing reliable star/galaxy classification.

3 - Estimating the completeness of the galaxy catalog and the contamination due to star/galaxy misclassification.

4 - Constructing an accurate model of the Point Spread Function (PSF) of the input image, taking into account possible spatial variations of the PSF as well as deviations of stellar isophotes from circularity.

5 - Deriving structural parameters of galaxies by fitting galaxy images with twodimensional PSF-convolved Sérsic models.

6 - Measuring galaxy isophotes by fitting them with Fourier-expanded ellipses, and deriving one-dimensional surface brightness profiles of galaxies.

7 - Measuring the growth curve of seeing corrected aperture magnitudes for galaxies.

The image analysis flow of 2DPHOT is presented in Figure 4.



Figure 4 - Image Analysis flow of 2DPHOT.

2DPHOT is being applied in several projects conducted by researches involved in this project and others. In a spectroscopic and photometric study of a rich cluster at intermediate redshift, it is used to measure global properties of cluster galaxies (Mercurio et al. 2008); and on a fundamental plane study based on SDSS and UKIDSS data (La Barbera et al. 2008b). The analysis of internal color gradients in early-type systems have been recently published in La Barbera & de Carvalho (2009). We also used 2DPHOT in a recent study of Fossil Groups (La Barbera et al. 2009). We have also begun a large-scale study (SPIDER; Spheroid's Panchromatic Investigation in Different Environment Regime of the general properties of early-type galaxies (ETGs) combining SDSS and UKIDSS data). This project makes extensive use of 2DPHOT to properly measure the seeing corrected structural parameters for 40356 ETGs.

2DPHOT is only a starting point in preparation for the avalanche of data in the next few decades. 2DPHOT requires as input an already processed image. Thus, we must be able to process raw images either on an individual basis or in a pipeline. To do so, we are taking advantage of the Astro-Wise (AW) system, developed by a consortium of European astronomy research institutes, coordinated by Dr. Edwin Valentijn of Kapteyn institute in Groningen. The AW environment consists of hardware and software

federated over five institutes in Europe. It was designed to scientifically exploit the increasing amount of data produced by experiments in different fields. AW is a general information system which was initially geared towards astronomy, but is now also used in other branches of science. This is an essential trait of AW in the context of a unified environment for data processing at INPE. It allows a user to archive raw data, calibrate data, and perform post-calibration scientific analysis. All results are stored in one environment that links together all the discrete steps in data analysis. This complete linking, including the input, output, and the software code used to get from one to the other, for arbitrary data volumes, has only been feasible thanks to a novel paradigm devised by the creators of AW. The algorithms included in the software have been developed to include arbitrary optical wide field imagers. This aspect is of major importance for BRAVO@INPE, since we will be developing software that enables us to ingest data from instruments available at SOAR and in the future from LSST.

AW was designed and implemented as a fully scalable and distributed information system to properly handle the huge amount of data that will be produced by large area surveys in the near future. By allowing the end-user to trace the data products, following all dependencies from the final catalog back to the raw data and, it becomes possible to re-derive the result with better calibration and/or improved analysis tools. This represents perhaps the first time that astronomers could truly reproduce each others' results.

To achieve these goals, the structural functions of AW include:

- * A data model that is translated to an object model and stored in a database;
- * I/O residing in a distributed database, containing all metadata for bulk data, parameter values used for processing/calibration, and the resulting catalogues;
- * A federated file server connecting to the synchronized databases and storing hundreds of terabytes of data

* An AW processing grid which sends jobs to parallel clusters, which then request data from the database.

The database architecture supports rapid trend analysis, complex queries and fast searching in terabyte-sized catalogues. All system components are distributed throughout Europe, enabling research groups to collaborate on shared projects. The web portal includes data viewing, quality labeling and compute services.

Some of the attractive features of the AW information system include:

- * fully scalability
- * can work with data from any imaging camera
- * allows sharing of results with collaborators globally
- * provides a web-based archive (called DB Viewer) to view an arbitrary subset of data of an arbitrary project at each stage of the processing
- * based on the python programming language
- * easy to add your own analysis code in python

- * wrapping of code written in fortran, C, etc. is straightforward
- * access to computing power via computer clusters (compute-grid infrastructure is part of the system)

These properties mean that AW overcomes the limitations of traditional analysis tools, which typically reside on a user's own computer or cluster. Reduction processes would usually be run by a single user, and saving sufficient metadata to reproduce every step is up to that individual. These behaviors are simply not sufficient for the new data volumes, collaborations and complexity in modern astronomy. The Astro-Wise system will be connected to the Virtual Observatory via the Euro-VO. Hundreds of terabytes of data will start entering the system when SOAR starts operating with the complete suite of instruments that were planned.

7 Data Analysis

The processing of raw data from a telescope into images, spectra or other products suitable for further analysis is only the first computationally intensive step on the path from photons to science. The processed data must be analyzed to detect, classify and characterize individual objects and groups of objects, and obtain physically meaningful measurements.

Within BRAVO@INPE, we are focusing on a few distinct data analysis projects:

- 1. Implementation of a decision tree for star/galaxy separation in the faint magnitude regime for wide field images;
- Development of a parallelized Friends-of-Friends (FoF) algorithm, with application to galaxy catalogs from the SDSS Stripe 82 project (<u>http://www.sdss.org/drsn1/</u> <u>DRSN1_data_release.html</u>)
- 3. Automatic morphological analysis of images in Stripe 82 using both traditional tools for structural parameter estimation (e.g. concentration/ asymmetry, Hernandez-Toledo, 2008) and advanced methods for image analysis such as the Euler characteristic and gradient spectral analysis (Rosa et al., 2007).
- 4. Development of a cluster finding algorithm based on Voronoi-Tesselation code, but considering a more realistic background distribution instead of the usual poissonian assumption. Preliminary results were presented in Soares Santos et al. (2008).
- 5. Virial analysis of galaxy clusters, allowing us to measure the most important dynamical quantities including total mass, based on the gapper technique described below.

7.1 Decision Tree (DT)

A decision tree is a computational method for splitting data into distinct classes, either based on pre-existing knowledge of the subgroups (supervised) or on inherent characteristics (unsupervised). Let a data set be described by a collection of attributes for each object in the data set. Each attribute is a measurement of some characteristic of an object (such as magnitude or size). These objects could belong to different classes or clusters (such as stars and galaxies). Imagine a data set for training, where the class of each object is already known. Our task is to develop a classification rule to

determine the class of an object based on its various attributes. If two objects have the same attributes, but they belong to different classes, then it is impossible to separate these objects based on this set of attributes. In this case, the data set with these attributes is not appropriate for a training set for the induction task. Thus, we must also determine the appropriate attributes to separate the objects into the desired classes.

Unlike other techniques for clustering analysis, the DT does not rely on distance metrics but instead makes a series of branching decisions based solely on numerical values of the attributes. A DT is a simple structure, where the final leaves define to which cluster an object with a specific set of attributes belongs. The nodes represent tests on a given attribute, with a branch for each possible output. For classifying an object, the starting point is the root of the tree; a test is applied to one attribute and the appropriate output branch is determined. The process is repeated using other attributes until the last leaf. Therefore, the object will belong to the cluster represented by that leaf.

There are many induction algorithms for decision trees. The ID3 algorithm, developed by Quinlan (1986), is the most popular. The algorithm was improved, allowing continuous parameters (Quinlan 1993). A package has been developed, called WEKA (Waikato Environment for Knowledge Analysis) where several standard machine learning techniques were incorporated into a "workbench". Several decision trees were designed for classifying objects detected in the SDSS (Sloan Digital Sky Survey) data for 5 passbands (u, g, r, i, z), employing WEKA (see Ruiz et al. 2008; Suchkov 2005; Ball et al. 2006). Our main goal is to provide a VO service to deal generally with the problem of star-galaxy separation - for whatever training set provided by the user, allow the generation of an appropriate DT using different methods and cross-validate the final obtained tree.

7.2 Parallel Friends-of-Friends algorithm

The friend-of-friends (FoF) algorithm is commonly used to join galaxies within a linking volume around each galaxy. This method has several attractive features, like being independent of the particular geometry of the galaxy distribution. For a given linking volume a unique group catalog is defined. One of the main problems in using this algorithm is the time it takes to process large numbers of objects, scaling with $N^2 \log N$. It is necessary to weaken this dependence on the total number of objects and thus be able to treat the hundreds of millions of particles found in current large cosmological simulations.

First experiments on reducing the dependence on N have shown that after a domain decomposition (subdividing the data in redshift shells) combined with a post-processing step we have already reduced the scaling to N log N², a considerable improvement. A simple domain decomposition can be implemented in a purely parallel manner, but it is insufficient because some objects artificially separated by sub-domain boundaries could in reality belong to the same group. Therefore, a post-processing procedure is applied to examine objects close to a boundary but with a valid friend in an adjacent sub-domain. Our parallel version has fully reproduced previous results (Caretta et al.

2008) for computing the potential gravitational energy spectrum for galaxies and clusters of galaxies at many redshifts. A VO service will be made available allowing the user to run the FoF algorithm over the most important cosmological simulations available to date and those inputed by the user.

7.3 Advanced tools for morphological analysis

As spatial information becomes ever more accessible through high resolution digital images, the need for robust techniques for complex pattern characterization is obvious. An obvious example is the mathematical description of galaxy images. Considerable attention has been paid to morphological classification of E/SO/Sa/Sab/Sm/Irr galaxy morphologies using Sloan Digital Sky Survey imaging. The data to be analyzed usually are (1) sky-subtracted, cleaned and log scaled g-band images; (2) filtered-enhanced versions of the g-band images; (3) the corresponding RGB composite images; and (4) a set of measured parameters, including surface brightness, position angle, ellipticity and spectral coefficients. In this sense, some useful mathematical and statistical approaches have been proposed (e.g. Lots et al. 2004) to estimate the CAS (concentration, asymmetry and clumpiness) structural parameters. Motivated by the data analysis challenges in the context of BRAVO@INPE, we have developed an alternative and complementary approach for characterization of inhomogeneity and radial asymmetry in galaxy images. Inhomogeneity is calculated using the Euler characteristic from the Minkowski functional. Radial asymmetry is obtained by applying gradient pattern analysis to 2D wavelet multi-resolution samples of the image. The combination of both structural characteristics is proposed as an effective measurement for galaxy morphology. The main objective here is to implement a VO service to deal with morphological analysis in general and in particular to analyze the entire SDSS (DR7) and explore the relationships between morphology and stellar population parameters, for instance.

7.4 A Modified Voronoi Tesselation code to search for clusters of galaxies

We are currently developing a cluster finder algorithm in 2+1 dimensions based on Voronoi tesselation (VT). The method is non-parametric and does not smooth the data, making the detection independent of the cluster shape. It uses all of the available galaxies, going as far down in the luminosity function as the input catalog permits. It does not rely on the existence of features such as a unique brightest cluster galaxy or a tight ridgeline in color-magnitude space. It works in shells of redshift, treating each shell as an independent 2-dimensional field. The core of the VT algorithm is the background above which an overdensity must rise to be identified as a cluster. In contrast to earlier implementations of the VT algorithm, we do not assume a Poissonian background. We use a more realistic assumption that the angular two-point correlation function of the background distribution has a power-law shape. In a given redshift shell, we build a Voronoi diagram and compare the distribution of cell areas with the distribution expected from a background-dominated field. We set as a threshold the cell size below which the distribution starts to increase faster than its background counterpart. The

clumps of contiguous cells found with density significantly above their respective cells are flagged as potential clusters.

The Voronoi diagram of a 2-dimensional distribution of points is a unique, non-arbitrary and non-parametric fragmentation of the area into polygons. A simple pseudo-algorithm to perform such fragmentation is the following: starting from any point P1, we label its nearest neighbor P2 and walk along the perpendicular bisector between those points. We stop when we reach for the first time a point Q1 equidistant from P1, P2 and any third point P3. We now walk along the perpendicular bisector between P1 and P3 until we reach the point Q2 and identify the next point P4 by the same criterion. Successive repetition of this process will eventually brings us back to Q1 after a finite number of steps. The set of points Q_i are the vertices of a polygon, the Voronoi cell, associated with P1. If this process is repeated for each point P_i we will have built the VT corresponding to this point field. There are several robust and efficient computational algorithms to build a Voronoi diagram from a given distribution. In our code we use the so-called divide-and-conquer algorithm implemented in the Triangle library (Shewchuk 1996).

There are no arbitrary parameters in constructing the VT for a given dataset. The cell edges are segments of the perpendicular bisectors between neighbor points and each vertex is an intersection of two bisectors. This implies that the cells will be smaller in the high-density regions and since each cell contains one and only one point, the inverse of the cell area gives the local density. The VT cluster finder takes advantage of this fact in the process of detection. We plan to implement a VO service where the user can input a galaxy catalog over a given area of the sky and receive a cluster catalog as output.

7.5 The Virial analysis tool for understanding cluster dynamics

Removal of interlopers and proper selection of galaxy cluster members is an essential step in the dynamical modeling of clusters and investigations of environmental effects affecting bound galaxies. There are several different approaches for interloper removal available in the literature. A recent comparison of the performance of many different methods applied to N-body cosmological simulations is given by Wojtak et al. (2007). In particular, they found that differences in mass estimates may be explained by the number of interlopers a given method selects or rejects. These could also explain the discrepant estimates from other methods of mass estimation (e.g., based on X-ray observations or lensing analysis). The shifting gapper method has two main advantages: (i) it is based on combined information for both position and velocity; (ii) it is independent of any hypotheses regarding the dynamical state of the cluster. The procedure we consider is similar to the approach adopted by Fadda et al. (1996). The input data consists of the radial and velocity offsets of each galaxy from the cluster center, being visualized as a phase-space diagram. It works through the application of the gap technique (Katgert et al. 1996; Olsen et al. 2005) in radial bins from the cluster center. This technique is used to identify gaps in the redshift distribution, resulting in the identification of groups in z-space. The bin size we consider for the shifting gapper is 0.60 Mpc or larger to force the selection of at least 15 galaxies (consistent with Fadda et al. 1996). Galaxies not associated with the main body of the cluster are eliminated. This procedure is repeated until the number of cluster members is stable (no more galaxies are rejected as interlopers). After a final list of members is reached, they can be used to measure the cluster velocity dispersion, from which we can estimate the cluster mass through virial analysis. While other procedures are based on physical assumptions about the cluster mass profile, the shifting gapper makes no physical hypotheses about the cluster's dynamical state. Further details of this method can be found in Lopes et al. (2009). This technique will be integrated in the VO service described in the previous subsection and will allow the user to carry out a dynamical analysis for the clusters detected with VT and having sufficient redshift measurements.

8 Data Mining

After the completion of image processing and derivation of meaningful quantities through data analysis, we are now confronted with an enormous collection of numerical quantities describing our data. An extant example is SDSS imaging, with 500 million objects, each with nearly 500 measured attributes. Which of these parameters are connected to fundamental physical properties? How do different types of objects cluster in high-dimensional parameter spaces? How do we find rare classes of objects, especially in the presence of errors or catastrophic mismeasurements? The 21st century will be a period of data-driven science, with the development of techniques to uncover the hidden knowledge in these kinds of massive datasets. This is the primary concern of a new branch of computer science - data mining (DM). It embraces a set of techniques for dealing with:classification, clustering, rule induction, visualization, pattern recognition and statistical analysis of massive data sets with extremely high dimensionality.

In the space science domain, although there are extensive archival data resources available over the web, the ability of scientists to access and analyze this content is becoming more and more limited. The large data volumes cannot be moved to a personal workstation to be processed by an individual's own software, while the software cannot be placed on the data host. Thus, DM in the context of this project, refers to specific computational methodologies, working in a logical system, to extract information and find hidden patterns embedded in the large amounts of data from space science surveys. Generally speaking, any computational methodological tool performed to transform data into information is calling a Data Mining System (DMS). It is notable that in space science (astronomy, astrophysics, cosmology and solar system studies) many existing data archives are unsuitable for DM because key pieces of metadata are missing. Hence, our goal in the first part of this project is to outline the major components of such a DMS, logically connected to the data processing and data grid requirements described previously.

9 A New Era for BRAVO

In the past two years we have gained important experience and knowledge of VO development, and our project was realigned to be fiscally feasible. New collaborations were established in all aspects of our planned investment. Within the context of the

INCT-Astronomy, the priority is to devise a roadmap for the near future to coherently invest in hardware and software that can meet our researchers' needs. BRAVO@INPE aims to create this synergy and contribute in strategic areas of the global VO.

Below we list the main strategic points of this enterprise. We emphasize that these are the overarching items defining this project and can be seen as pillars of a consistent investment in VO:

9.1 Network Infrastructure

From the results of our IT census we see the level of insufficiency of the network used by the scientific community in Brazil, especially for astronomers. High speed and secure network connections are of paramount importance not only for simple tasks in our daily work but also for establishing a national grid processing facility, such as the one we are developing with the BNPGA. Discussions are already in place with people from RNP (Brazilian National Research and Education Network) on development of a plan to elevate the accessibility of the astronomical community in Brazil to a higher level. This is one of the main points of this project - to conduct a study of the current situation and move to a modern network infrastructure.

9.2 Astro-Wise as a national environment for data reduction and analysis

Several pipelines were developed in recent years to address the demand of large area surveys like SDSS. In these cases, users do not have to worry about data reduction. However, more and more sophisticated algorithms for object detection, star-galaxy separation, photometric redshift estimates, morphological analysis and more are flourishing and there is an obvious need for reprocessing in some cases. As presented before, we are implementing AW as the environment for large amounts of data processing, which was developed under the VO rules. It will be the first AW node in South America. AW is currently extant only in Europe in compliance with IVOA standards.

9.3 Creating the BNPGA

The census we did within the INCT-Astronomy community indicated that we not only need to upgrade our network infrastructure but also invest in creating a Grid Processing facility that can meet the growing demands of the astronomy community, not only because of the increase in the amount of data but also due to its increasing complexity. The BNPGA is the response to community's need for processing a large amount of data and reliably publishing the results in an environment meeting VO standards. BNPGA will start as an exercise of processing the entire SDSS in one band and by doing so we will be able to implement the environment before we can move to better and more powerful clusters with thousands of modern processors.

9.4 The Virtual Lab for Advanced Data Analysis (VLADA)

The Virtual Laboratory for Advanced Data Analysis is a project initiated at the Lab for Computing and Applied Mathematics-INPE which aims to provide a new virtual environment for scientific analysis tools to extract statistical and physical information from times series, images and hypercube data. Its preliminary version consists of a PHP user interface through which the user can input the data and receive specific measures characterizing the data (ex. statistical moments, power spectra, generalized dimensions, Euler characteristics, asymmetry coefficients, etc). In the context of BRAVO@INPE, VLADA might be considered a virtual tool box for data analysis in general.

10 Summary

The Virtual Observatory is rapidly becoming a reality. The combination of growing data volumes and data complexity, coupled with computational and algorithmic advances, has made the VO a necessity. We have described some of the ongoing projects to implement databases, general-purpose computational algorithms, grid networks, and other VO-enabling technologies in Brazil. A common theme among all of these developments is the dire need for computational resources (CPUs, storage and network), software, and the expertise to design, install, and bring to life these complex systems. The international nature of astronomy implies that everyone can benefit and everyone should contribute to this enterprise. We have provided a basic and certainly not exhaustive outline of the components of the VO, and described the specific contributions that the Brazilian astronomical and computer science communities have made and will be making to this effort. Our growing partnerships in large telescopes and unfettered access to large public datasets demands that we develop our own tools and expertise to leverage these investments and strengthen our scientific output. Finally, we have described the necessary next steps in terms of hardware, software and personnel to advance BRAVO from an incipient program to a fully functioning project

REFERENCES

Ball et al. 2006, ApJ. 650, 497 Caretta et al. 2008, A&A. 487, 445 Fadda D. et al. 1996, ApJ, 473, 670 Gargiulo et al. 2009, arXiv0902.4383 Hernandez-Toledo 2008, A&A 136, 2115 Katgert P. et al. 1996, A&A, 310, 8 La Barbera et al. 2008a, PASP, 120, 681 La Barbera et al. 2008b, ApJL, 689, 913 La Barbera et al. 2009, arXiv0812.2929 La Barbera, F. & de Carvalho, R.R. 2009, ApJL, 699, 76 Lopes P.A.A. et al. 2009, MNRAS, 392, 135 Lots et al. 2004, AJ, 128, 163 Mercurio et al. 2008, MNRAS, 387,1374 Olsen L.F. et al. 2005, A&A, 435, 781 Quinlan 1986, Machine Learning, 1, 81 Quinlan 1993, Programs for Machine Learning. San Mateo, CA: Morgan Kaufman Rosa et al. 2007, Physica A 386, 666-673 Ruiz et al. 2008, National Congress on Computing and Applied Mathematics, www.sbmac.org.br/eventos/cnmac/xxxi cnmac/PDF/251.pdf Shewchuk 1996, Applied Computational Geometry: Towards Geometric Engineering, Eds. Ming C. Lin and Dinesh Manocha, Vol 1148 of Lecture Notes in Computer Science, pages 203-222, Springer-Verlag, Berlin Soares Santos et al. 2008, arXiv:0810.3689. Suchkov 2005, AJ, 130, 2439 Wojtak R. et al. 2007, A&A, 466, 437

Appendix A

Universities and Institutes Associated to INCT-A:

São Paulo

USP - Universidade de São Paulo

INPE - Instituto Nacional de Pesquisas Espaciais

UPM - Universidade Presbiteriana Mackenzie

UNICSUL – Universidade Cruzeiro do Sul

UNIVAP- Universidade do Vale do Paraíba

UNESP – Universidade Estadual Júlio de Mesquita Filho

UNIFESP - Universidade Federal de São Paulo

UFABC - Universidade Federal do ABC

FSA - Fundação Santo André

Rio Grande de Sul

UFRGS – Universidade Federal do Rio Grande do Sul

UFSM – Universidade Federal de Santa Maria

UFPel – Universidade Federal de Pelotas

Unipampa – Universidade Federal do Pampa

UCS – Universidade de Caxias do Sul

Rio de Janeiro

ON - Observatório Nacional

UFRJ – Universidade Federal do Rio de Janeiro

CBPF - Centro Brasileiro de Pesquisas Físicas

Minas Gerais

UFMG – Universidade Federal de Minas Gerais

LNA - Laboratório Nacional de Astrofísica

UNIFEI – Universidade Federal de Itajubá

UFJF – Universidade Federal de Juiz de Fora

Santa Catarina

UFSC – Universidade Federal de Santa Catarina

Unochapecó – Universidade Comunitária Regional de Chapecó

Bahia

UESC – Universidade de Santa Cruz

Distrito Federal

UNB – Universidade de Brasília **Paraná** UEL – Universidade Estadual de Londrina **Pernambuco** UNIVASF – Universidade Federal do Vale do São Francisco

Appendix B

We conducted a census with all these 31 institutes associated to INCT-Astronomy, asking specifically:

1 - What is the total number of users they have in their Department ?

2 - How many computers they have access to, including Desktops, and what are their main characteristics ?

3 - Do they have access to cluster systems ? if so, what are the characteristics ?

Although the questionary might not be very objective, the main idea was to collect as much info as possible and then try to organize it accordingly. Twenty (66%) of the institutes participating of the INCT-Astronomy responded to the questions. The remaining institutes, which did not answer, represent small groups (2-3 researchers in located in Physics Departments) still involved in implementing basic infrastructure. Therefore, it would be fair to consider the data presented in the Tables 1-4 (see appendix) as representative of the cyber infrastructure of the Brazilian astronomical community.

Three types of information were requested and we present them in Tables 1-4:

i) Number of users, including researchers and graduate students (Table 1)

ii) Cyber infrastructure available in terms of Desktops (Table 2)

iii) Cyber infrastructure available in terms of clusters (Tables 3a and 3b)

From Table 1 we see that 66% of the institutes composing the INCT-Astronomy contribute with 277 users. The remaining 12 institutes contribute with ~30 users, making a total of ~310 users who participate directly or indirectly of the INCT-Astronomy. The figure below shows the distribution of the number of users per institute from where we can see that although the Brazilian community had a very significant growth in the last 20 years, most of the main power working on astronomy in Brazil is still concentrated in a few places. This is something to be addressed in the near future within the context of

a broad program being prepared by the INCT-Astronomy but it goes beyond the scope of this project.

From Table 2, we can conclude that most users of our community has access to at least a Desktop with moderate computational capacity. This conclusion must be seen with caution. It happens that the Brazilian astronomical community is dominated by stellar astrophysicists (~70%) doing important and competitive research but using data of low complexity (1D spectroscopy and/or 1D photometry etc). It is important to remember that even this situation is changing dramatically and will keep changing in the near future with the large telescopes coming up. In this context, we understand that the current computational facilities available seem to be adequate and fulfill the present demand. However, it is clear that the current Cyber infrastructure will be obsolete when dealing with the extremely large amount of data coming from either stellar or extragalactic projects.

Tables 3a and 3b refer to the info about clusters available, allowing high performance processing. As we can see, researchers from these 20 institutes have access to modern servers with more than 8 processors each (Class A) and to beowulf types, composed of mono-processed nodes and internal networks of 100 Mps (Class B). Only 12 out of the 20 institutes listed in Tables 3a and 3b have access to a cluster and only 7 out of these 12 have access to a Class A cluster. These numbers will not change considerably if we include the remaining 12 institutes which did not provide information.

In essence, 50% of the institutes composing the INCT-Astronomy have access a cluster, regardless of which class. It is important to note that in some cases the clusters are shared with researchers from different disciplines like Physics since the small groups of researchers developing Astronomy in Brazil are inserted in large Physics Departments.

Adding up all the available processors in the different clusters as listed in Tables 3a and 3b, we would have in principle the total number of processors for grid processing (see Table 4). This total, 419, is only 6% of the required number mentioned in the figures presented previously for processing the entire DR7 in one band, in one week, for example. This is only a crude estimate considering that all the processors are different, some better than others - fifty are old type of processors that would add little to the total processing capacity.

In terms of total storage, these clusters do not go over 45 Tb, and although it satisfies the needs of individual groups, is clearly incompatible with the needs of the coming decade where large telescopes will produce data on a 2 Pb/year rate.

Finally, we want to stress that this census although may not represent the entire Brazilian astronomical community, it shows how deficient it is the current hardware/ software and network infrastructure.

Table 1	
---------	--

INSTITUTE	Users	Researchers	Visitors + Postdocs	Students: Msc+PhD +Grad	
1	2	3	4	5	
USP/IAG	67	22	10	35	
USP/EACH	1	1	0	0	
LNA	[6]	[0]	?	0	
USP/IF	1	2	?	?	
UFSC/DF	23	4	1	18	
UNIVAP/ IP&D	13	6	0	7	
INPE	20	8	0	12	
UFMG/DF	22	8	2	12	
UNICSUL/ NucAstro	7	7	0	0	
UESC/DCET	6	7	1	0	
UFRGS/IF	49	10	4	35	
UFRJ/OV	25	14	1	15	
UNIMACK	8	8	?	?	
UFRJ/IF	21	4	1	16	
CAXIAS S.	1	1	0	1	
UNIFESP	1	2	?	?	
UFABC	5	2	?	3	
UNIFEI	4	4	?	?	
UNIPAMPA	3	3	0	0	
ON		9?	?	?	

Table 2

INSTITUTE	Desktops	Processors	Total RAM (Gb)	Storage (Tb)	OS
1	2	3	4	5	6
USP/IAG	?	185	[>185]	29,6	LWM
USP/EACH	2	2	3	3	?
LNA	?	?	?	?	L[WM]
USP/IF	?	?	?	?	L[WM]
UFSC/DF	18	35	16	2,5	LW
UNIVAP/ IP&D	7	14	7x1	>7x0,2	LW
INPE	21	30	30	4	LW
UFMG/DF	12	17	25	2,6	LW
UNICSUL/ NucAstro	6 +{1}	?	18	2	LW
UESC/DCET	21	28	20	1	LW
UFRGS/IF	10 - 15	22	13	1	LW
UFRJ/OV	13	17	10+?	0,8+?	LW
UNIMACK	12	12	18,5	9	L
UFRJ/IF	21	21	30	1	LW
CAXIAS S.	5 {+10}	8	5	1,5	W
UNIFESP	3	3	6	0,75	LW
UFABC	5	?	10	6	?
UNIFEI	6	11	30	2	LW
UNIPAMPA	?	?	?	?	?
ON	10	15	25	4	LW

Table 3a

INSTITUTE	CLUSTER	Description	Total Disk (Tb)	Internal Network	OS	Parallel Software	Installed Software
1	2	3	4	5	6	7	8
	HYDRA	20x2 Xeon ; 3Ghz ; Itautec	20 x 0.2	Gigabit [=Infiniband?]	Ganglia+ [Linux]	MPI	
USP/IAG	HPC	22x1 AMD64 ; 3Ghz	22 x 0.16	Ethernet 100Mb	Ganglia+ [Linux]	MPI	
	BEETHOVEN	12x1 Intel/ AMD 3Ghz	1 x 0.12	Ethernet 100Mb	Ganglia+ [Linux]		
USP/EACH							
LNA	HP- DL380R05	1xQuad-Core Xeon E5405 ; 2.0GHz	12		[Linux]		
USP/IF	PMC /DFMa	48x2 ??? ; > 1Ghz ; 1x48Gb/RAM + Servidor 10Gb/RAM	[<10Tb]	[Ethernet 100Mb]	Linux	[VPM]	mathematica
UFSC/DF	MINERVA	6xHP(Xeon E5405) +9xAMD64_3 Ghz + 4xCore2quad + 3x storage	3,5	Gigabit [?]	Ganglia+ cacti + pbs		
UNIVAP/ IP&D	[noname]	7xDual- Core ; 3Ghz	7x0.08	Trendnet	Linux	[MPI]	[gadget]
INPE	CPAD /Inpe	23x2+2x4 Opteron 2core + servidor + storage	4,5	Infiniband 2.5Gbps	Linux RedHat	MPI	gadget ; 2DPhot

Table 3b

INSTITUTE	CLUSTER	Description	Total Disk (Tb)	Internal Network	OS	Parallel Software	Installed Software
UFMG/DF	{BULL / University}	{50x4 ?; 50x32Gb/ RAM}	{}	{Infiniband?}	8		
	estação1	8x ? ; 16Gb/RAM	1,5		Linux RedHat		
UNICSUL/ NucAstro	estação2	8x ? ; 32Gb/RAM	1,5		Linux RedHat		
	{}						
UESC/DCET	[noname]	16xAthlon ; 1.7Ghz ; 8Gb/RAM					
	{BULL / University}	{160x Xeon ; 2.66Ghz ; 320Gb/RAM ; storage}	{10}	{infiniband}	{Linux ; PBS-pro ; Bull}		{Oracle}
UFRGS/IF	CPADA	6x2*Athlon; 1.4GHz; 1GB RAM	0,24	Ethernet 100Mb	Linux	MPI	Gadget
UFRJ/OV	[noname]	4xCeleron ; 2.4Ghz ; 4x1Gb RAM	0,32	Ethernet 100Mb	Rock's +_Linux		
	DELL	4xXeon ;	0,25		Ubuntu		
UNIMACK	[noname]	15x2+1x2 Opteron ; 1.4Ghz; #Gb RAM	1	?	Debian		
UFRJ/IF							
CAXIAS S.	[noname]	11xPentium4 ; 2.6GHz; 1Gb RAM +1xPentium4 (servidor) 3Ghz	12x0.08	Ethernet 100Mb	Oscar-Linux		

Table 4

Institute	Users	No. Cluster Processors	Cluster Storage (Tb)	
USP/IAG	67	74	7,64	
USP/EACH	1	0	0	
LNA	6	4	12	
USP/IF	1	96	10	
UFSC/DF	23	31	3,5	
UNIVAP/ IP&D	13	14	0,56	
INPE	20	108	4,5	
UFMG/DF	22	0	0	
UNICSUL/ NucAstro	7	16	3	
UESC/DCET	6	16	1	
UFRGS/IF	49	12	0,24	
UFRJ/OV	25	8	0,57	
UNIMACK	8	32	1	
UFRJ/IF	21	0	0	
CAXIAS S.	1	12	0,96	
UNIFESP	1	0	0	
UFABC	5	0	0	
UNIFEI	4	0	0	
UNIPAMPA	3	0	0	
ON	[9]	0	0	
Totals	283	423	44,97	

Notes to Tables

- Items within braces ({ }) correspond to projected acquisitions.
 Items within brackets ([]) are our own estimates of not informed items.
 In Table 1 column (2) totalizes the content of the ensuing columns.
 In Column (6) of Table 4: L = Linux ; W = Ms-Windows ; M = Mac OS.
 In Tables 3a and 3b, column (2) gives the nicknames of the corresponding equipment; a bar (/) indicates that this is a shared resource. Notice that multi-processed (N > 4) storage servers have also been included here.