Proceedings of the Second EELA-2 Conference
R. Mayo et al. (Eds.)
CIEMAT 2009

The São Paulo State University Campus Grid Initiative

Rogério Luiz Iope

Núcleo de Computação Científica
Universidade Estadual Paulista
rogerio.iope@cern.ch

Abstract

UNESP - the São Paulo State University - is in the final stages of setting up one of the largest Campus Grid initiatives in Latin America, with computing resources widely dispersed on seven different cities over the State of São Paulo, Brazil. GridUNESP, as the project is known, will empower University research groups of several areas of scientific investigation, mainly genetic sequencing, weather forecasting, molecular and cellular modeling, medical image reconstruction, development of new materials, quantum chemistry, large-scale numerical simulations, and high-energy physics, allowing them to access state-of-art data processing and storage systems. The central cluster, which is installed at the new UNESP campus in Barra Funda, São Paulo, has 2,048 processing cores, reaching a theoretical peak performance of about 23.2 teraflops (trillions of calculations per second).

The cost of the whole project, over US$ 2 million, has been supported by the Brazilian Research and Projects funding agency FINEP (Research and Projects Financing Agency), also known as the Brazilian Innovation Agency, a public company subordinated to the Ministry of Science and Technology. The computational infrastructure includes 386 servers, over 200 terabytes of storage space and an advanced networking infrastructure for inter-cluster connection. The project is based on a hierarchical 2-tier architecture, which includes a central powerful cluster located in São Paulo and seven secondary clusters at UNESP campuses of Araraquara, Bauru, Botucatu, Ilha Solteira, Rio Claro, São José do Rio Preto and São Paulo. Rio Claro is the nearest site, around 180 km far from Sao Paulo, whereas Ilha Solteira, the farthest one, is over 650 km away from the State capital.

GridUNESP is being connected to the U.S. Internet2 through the KyaTera network - an advanced R&E optical testbed funded by FAPESP (the State of São Paulo Research Foundation), which interconnects the main research institutions in the State of São Paulo - and through the 10 Gbps research network infrastructure between São Paulo and Miami provided by WHREN/LILA - Western Hemisphere R&E Networks / Links Interconnecting Latin America. The interconnection between the main cluster and the secondary ones is also being implemented through the KyaTera network. This might be considered one of the main distinguishing features of the GridUNESP project: a Grid infrastructure built over an underlying optical transport network based on WDM technology which provides fully dedicated 1 Gbps and 10 Gbps lambdas for inter-cluster connection. Araraquara is the main aggregation point, being connected to Sao Paulo through a 10 Gbps fully dedicated lambda. The remaining 6 sites attach to Araraquara through 1 Gbps dedicated links.

The project also includes some high-end L2/L3 network switches that are being installed on strategic positions along the interconnecting paths over the KyaTera network. This infrastructure will allow easily setting up and tearing down on-demand end-to-end optical connections between the participating sites by means of dynamically created virtual LANs. The high-end distributed computing and storage resources, communicating through dedicated channels over an underlying transparent optical transport infrastructure, put GridUNESP in a privileged position, as it should become the first Lambda Grid deployed in Latin America in the near future.

GridUNESP has also established a formal partnership with the Open Science Grid Consortium, an organization composed of service and resource providers, researchers from universities and national laboratories, as well as computing centers across the U.S., Asia and Latin America. GridUNESP, will use the OSG middleware stack to integrate its computational resources and share them with other R&E institutions worldwide.

GridUNESP resources will be accessible to any UNESP researcher that needs computational power for processing, storing and moving large amounts of data.

1. Introduction

The ubiquitous availability of powerful computers and high-speed networks as low-cost commodity components is changing the way scientists and engineers manipulate information. Together with computer scientists and network engineers, scientists around the world are collaborating to build the necessary middleware and to leverage global-scale high-speed networks aiming to integrate distributed computing resources into unified infrastructures. The ultimate goal is to reach the ability to easily set up secure and controlled environments for collaborative sharing of resources to fostering their research activities [1]. These efforts have led to the development of a new research area in computer science, commonly known as Grid computing. The term Grid was introduced in the middle of 1990s by the pioneers Foster and Kesselman to describe their proposal for a distributed computing infrastructure for advanced science and engineering [2], inspired in other kinds of distributed infrastructures assembled to deliver a certain class of services to the final users, like electric power grids, or to flow through wide areas, like railroad grids. A Grid computing infrastructure is defined by Foster as "a distributed system that integrates and coordinates resources not subject to centralized control, using standard and open protocols and interfaces to deliver nontrivial qualities of service" [3].

Grid computing has thus emerged as the dominant paradigm for wide-area distributed computing. Over the last decade the research efforts devoted towards making this vision a reality can be divided into two classes: (1) efforts addressing the definition and implementation of the middleware core services to securely authenticate Grid users and authorize them to access specific resources at certain sites belonging to multiple administrative domains, allowing the establishment of Virtual Organizations within the shared resources; and (2) efforts to extend the existing programming models for parallel and distributed computing, as well as languages, tools and frameworks for the development of efficient Grid applications [4].

A critical aspect of high performance Grid computing environments is the network support. The rapid evolution and adoption of Grid technologies are occurring mainly due to the significant increase of the speed of networks. The way which Grid environments and applications use the underlying network vary widely. At one extreme we can find the so-called peer-to-peer Grid architectures, where the network infrastructure is the standard Internet, based on a best-effort model. The standard Internet offers no specific support for Grid applications, often resulting in a rather poor environment for Grid applications that need

to run parallel jobs or to handle huge masses of data [5]. At the other extreme, we can find the so-called Lambda-Grids, in which the distributed resources are interconnected through very high performance dedicated optical fiber network infrastructures, based on Wavelength Division Multiplexing technologies that allow placing multiple optical signals within the same fiber [6]. Building a widespread Grid infrastructure over a WDM optical networking core is a very attractive proposition, as optical connections between pairs of resources can be dynamically set up on demand by dedicated lightpaths that traverse the network mesh. Those emerging lightpath-based technologies are appealing to the Grid community because they allow the inclusion of the underlying network infrastructure as a schedulable resource under the control of the Grid middleware, just like it is done with computing and storage resources.

This paper describes the preliminary steps towards the deployment of a production-quality campus Grid infrastructure at the São Paulo State University, with computing resources widely dispersed on seven different cities over the State of São Paulo, Brazil. GridUNESP, as the project is known, will empower University research groups of several areas of scientific investigation, mainly genetic sequencing, weather forecasting, molecular and cellular modeling, medical image reconstruction, development of new materials, quantum chemistry, large-scale numerical simulations, and high-energy physics, allowing them to access state-of-art data processing and storage systems.

The remaining of the paper is organized as follows. In Section 2 we briefly introduce the São Paulo State University and outline the historic development of the GridUNESP project. The underlying hardware that forms the distributed clusters is detailed in Section 3. In Section 4 we describe the networking infrastructure. Software infrastructure and preliminary results are only briefly described in Section 5, as the computing systems have only recently been switched on. Finally, in Section 6 we draw some conclusions and discuss future work.

2. The São Paulo State University and the GridUNESP project

UNESP, acronym for Universidade Estadual Paulista (São Paulo State University), is the second largest and one of the most important universities in Brazil, with distinguished achievements in teaching, research and extension services. UNESP is part of the State of São Paulo public higher education system, which also includes the University of São Paulo (USP) and the State University of Campinas (UNICAMP). A prominent distinction of UNESP among others is its multi-campus structure, with 33 faculties and institutes on 23 campuses distributed throughout the State of São Paulo.

UNESP is in the final stages of setting up a distributed computational system which might be considered one of the largest Campus Grid initiatives in Latin America, with computing resources widely dispersed on seven campuses. The project, officially named “Computational Capacity Integration at UNESP” but commonly known as GridUNESP, started in 2004 with a call for scientific proposals sent to the University researchers, with the goal of supporting a subsequent request for funding. As a result, several research proposals have been submitted, and eight of them, from different areas of scientific investigation, have been chosen. The proposals cover the following areas: genetic sequencing, weather forecasting, molecular and cellular modeling, medical image reconstruction, development of new materials, quantum chemistry, large-scale numerical simulations, and high-energy physics.

The formal proposal was submitted to the Brazilian Research and Projects funding agency FINEP by the end of 2005 and the final approval occurred by the middle of 2006. A Request for Information (RFI) has been distributed to some top-quality vendors for defining the hardware to be acquired, followed by a Request for Proposals (RFP), distributed to global leaders in large-scale clustered systems such as Dell, HP, IBM, SGI, Sun, Unisys, as well as local vendors, for the bidding process. Special attention has been given to the RFP writing,

carefully designed to promote fairness. It clearly described the scope of the project, its goals and needs, the minimum technical requirements, the expectations of the winning vendor, and the responsibilities of all parties involved. To a large extent, the success of the acquisition may be credited to the time our team has invested in the RFP elaboration. The analysis of the technical and commercial bids has been monitored by a multi-institutional commission made up of experienced professionals.

The hardware acquisition involved a complex import process which started by the end of 2007. The equipments were delivered in São Paulo on the third quarter of 2008. Unfortunately an unforeseen situation prevented us from starting the deployment of the systems by that time: the physical construction of the new UNESP building in São Paulo capital, including the area for the new data center, had been severely delayed by several months. All the hardware had to be stored into a warehouse until the finish of the building and data center construction. During this period most of the hardware had been transported to the vendor's data center, where they were tested, and initial deployment (mainly firmware upgrade, storage configuration, and operating system installation) initiated. The construction of the new UNESP Center for Scientific Computing, which includes a well designed data center, finished by the middle of this year, when we finally started the GridUNESP deployment process.

3. Hardware infrastructure

The GridUNESP computational infrastructure includes 386 servers, over 200 terabytes of storage space and an advanced networking infrastructure for inter-cluster connection. The project is based on a hierarchical 2-tier architecture, which includes a central cluster located in São Paulo, capital of the state, and seven secondary clusters at UNESP campuses of Araraquara, Bauru, Botucatu, Ilha Solteira, Rio Claro, São José do Rio Preto and São Paulo. Rio Claro is the nearest site, around 180 km far from Sao Paulo, whereas Ilha Solteira, the farthest one, is over 650 km away from the state capital.

The central cluster have 2048 processing cores (not considering the main servers), based on Intel Xeon quad-core architecture, which leads to a theoretical computing capacity of 23.2 TFlops. It consists of eleven 42U, 19 inch rack enclosures, into which are installed 256 worker nodes, 4 main servers, a central fibre-channel-based SAN storage system, and 4 hybrid data servers. Each worker node consists of dual quad-core Intel Xeon E5440 (Harpertown) 2.83GHz processors, 16 GB of main memory, and a 73 GB SAS disk. The main servers consist of quad quad-core Intel Xeon E7340 (Tigertown) 2.40GHz processors, 32 GB of main memory, and four 146 GB SAS disks. The fibre-channel SAN has 36 TB of storage capacity and two 4 Gbps channels. Each of the four hybrid data servers has forty-eight 500 GB SATA disks, totalizing 96 TB of raw storage space. The worker nodes and the main servers are interconnected through an internal gigabit Ethernet network for usual TCP/IP traffic and a dedicated Fast Ethernet management network for out-of-band server monitoring and control. They are also interconnected through a central Infiniband server switch that can operate in double data rate mode, thus delivering 20 Gbps per port. The Infiniband switch occupies a fully dedicated rack enclosure. The central cluster rack enclosures also hold an independent, general-purpose mini-cluster that is available for development and testing of new applications. It consists of a head node and 4 worker nodes, each composed of dual quad-core Intel Xeon E5335 (Harpertown) 2.00GHz processors, 8 GB of main memory, and a 73 GB SAS disk. It also has an independent 24-port gigabit switch and all servers are attached to the Fast Ethernet management network as well. A high-performance interconnect, provided by a double data rate 24-port Infiniband 24 switch, is also present. Figure 1 shows the central cluster internal connections.

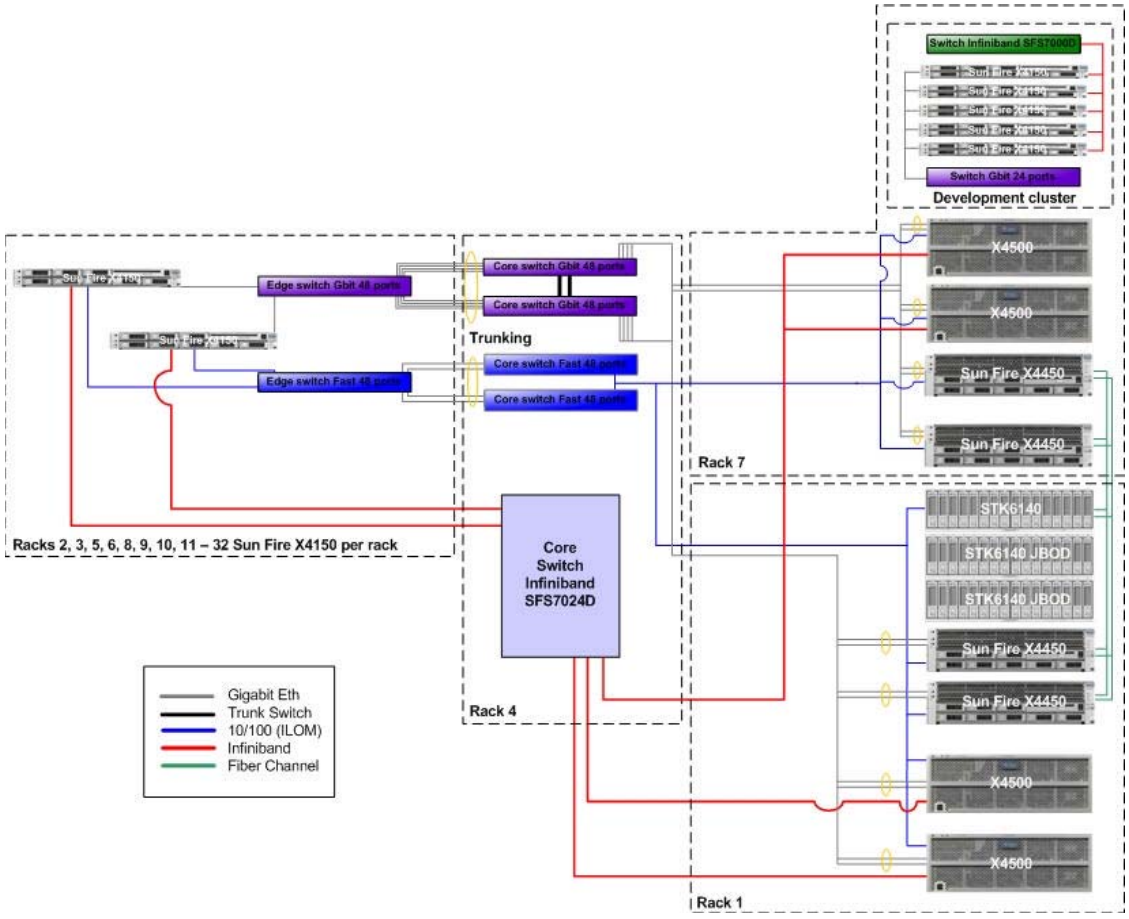

Figure 1. GridUNESP central cluster internal connections

Each of the seven secondary clusters are composed of 16 worker nodes and two head nodes, leading to 128 processing cores (again not considering the main servers). The worker nodes are identical to the corresponding systems of the main cluster. The two head nodes are identical to the servers of the development cluster, except that they have 16 GB of main memory and two 146 GB SAS disks. They are also both connected to a SAN fibre-channel-based storage with 12 TB of capacity. Each secondary cluster also has a 24-port gigabit Ethernet switch for regular TCP/IP traffic and a 24-port Fast Ethernet switch for out-of-band management. Although the management network is not directly accessible externally, it can be accessed by means of the head nodes. This implies that all the servers of all the secondary clusters can be monitored and controlled remotely.

All of the servers of each of the eight clusters (main cluster plus seven secondary ones) can be monitored and controlled remotely. Thus any server of the entire infrastructure can be powered on or off, its internal temperatures can be checked, and its console messages can be viewed, from a single central point. The central point is the Center for Scientific Computing control room at São Paulo, which can then be considered the GridUNESP GOC (Grid Operations Center). Figure 2 shows the secondary clusters internal connections.

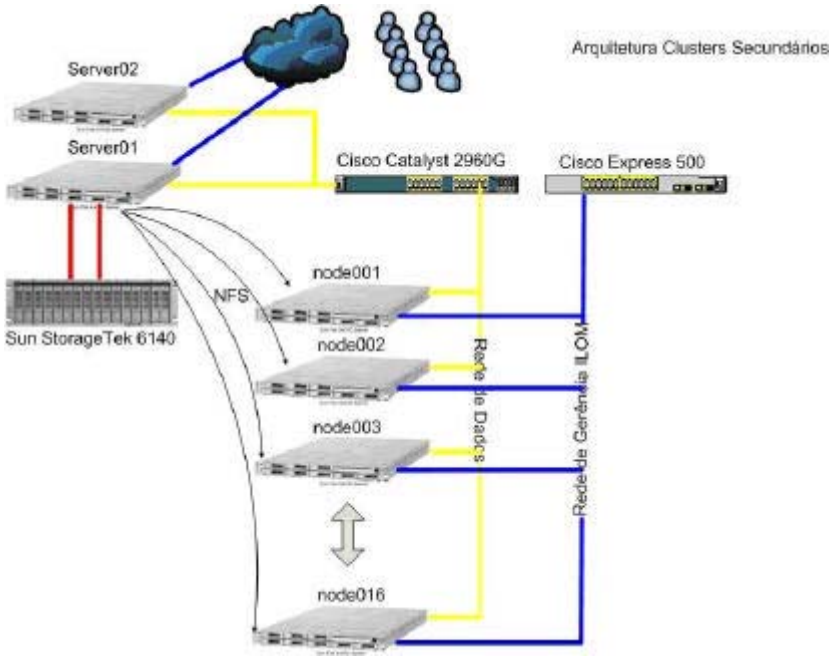

Figure 2. GridUNESP secondary clusters internal connections

The secondary clusters are installed in the main IT service rooms of the following UNESP faculties and institutes:

- Instituto de Química (Araraquara)
- Faculdade de Ciências (Bauru)
- Instituto de Biociências (Botucatu)
- Faculdade de Engenharia (Ilha Solteira)
- Instituto de Geociências e Ciências Exatas (Rio Claro)
- Instituto de Biociências, Letras e Ciências Exatas (São José do Rio Preto)

4. Network infrastructure

High performance Grid architectures in general, and the GridUNESP in particular, are generally designed to provide services that cannot be supported by traditional network infrastructures. The commodity Internet, based on a best-effort delivery model, is too slow to allow efficient and effective use of remote resources and to handle the huge masses of data being generated in emerging e-Science applications. The UNESP network infrastructure still uses leased E3 lines to interconnect its campuses, making things even worse.

During the design of the GridUNESP project, we carefully devised the feasibility of deploying a dedicated network infrastructure for cluster interconnection. Of course doing so from scratch would require an amount of financial resources several times larger than the budget for the entire project. But this was not really necessary as the state of São Paulo already has an advanced optical network testbed that interconnects its main research laboratories. The KyaTera network, as it is known, is a large, geographically distributed laboratory facility, available to the research community for field trials of optical components and equipments, for fundamental and applied research in optical transmission and networking technologies, and for the development of advanced Internet applications. The network is based upon the concept of dark fibers whose endpoints are deployed directly to the research laboratories, and the inter-nodal points are connected through DWDM multiplexers and demultiplexers. KyaTera is part of TIDIA, acronym for Tecnologia da Informação no

Desenvolvimento da Internet Avançada (Information Technology for the Development of Advanced Internet), a special funding program created by FAPESP - the São Paulo State Research Funding Agency - for high impact cooperative projects having the Internet as the main subject of research. The KyaTera is essentially a transport network that uses equipments based on the ITU G.709 Optical Transport Network protocol, which enables multiple network services (SONET/SDH, IP/Ethernet at 1 and 10 Gbps, SAN, Infiniband, video streamming, and so on) to flow on the infrastructure.

After a series of meetings with the KyaTera administrators, we concluded that the use of the KyaTera network as an underlying infrastructure for interconnecting the GridUNESP clusters would be possible, but some arrangements should me made:

- deficiencies along some paths should be corrected by adding new optical amplifiers
- a dedicated lambda should be deployed to all sites, to prevent traffic congestion from other experiments running on the network
- the KyaTera metropolitan ring at São Paulo capital should be expanded to include the UNESP Center for Scientific Computing (and consequently, the main cluster)
- the network infrastructure should be extended to Bauru, Botucatu, and Ilha Solteira through the deployment of new dark fibers; the remaining campuses could be easily inserted into the network
- the 10 Gbps KyaTera backbone should be extended from São Carlos up to Araraquara (50 Km far from each other) and make this point the central distribution point to all GridUNESP secondary sites

We have also devised that it should be important to have dedicated L2/L3 switch/routers at two strategic points in the network: (1) at Araraquara as it has been chosen as the aggregation point of the secondary sites, and (2) at NAP of Brazil in Barueri (30 Km far from São Paulo capital) for the interconnection to the 10 Gbps research network infrastructure between São Paulo and Miami provided by WHREN-LILA - Western Hemisphere R&E Networks - Links Interconnecting Latin America [7].

After a series of negotiations we concluded that the costs for those arrangements were within the reach of the project budget. This indeed can be considered one of the main distinguishing features of the GridUNESP project: a Grid infrastructure built over an underlying optical transport network based on WDM technology which provides fully dedicated 1 Gbps and 10 Gbps lambdas for inter-cluster connection. Araraquara is the main aggregation point, being connected to Sao Paulo through a 10 Gbps fully dedicated lambda. The remaining 6 sites attach to Araraquara through 1 Gbps dedicated links. Figure 3 shows the network infrastructure through the KyaTera network. It also shows how the Center for Scientific Computing at São Paulo capital is going to use the MetroSampa network, a metro ring made up of dark fibers provided by Eletropaulo Telecom and Cisco switch-routers that interconnects several Research & Education institutions in the metropolitan region of São Paulo capital. MetroSampa is part of the Redecomep initiative, supported by Brazil's Ministry of Science and Technology and conducted by RNP (acronym for Rede Nacional de Ensino e Pesquisa, the Brazilian National Research and Education Network), which aims to install high-speed networks throughout the metropolitan regions around the 27 Brazilian state capitals. MetroSampa is a shared facility used by R&E institutions to flow Internet commodity traffic and its main ring is made up of 1 Gbps links. Traffic throughout MetroSampa flows in Layer 2, so it is fully based on virtual LANs.

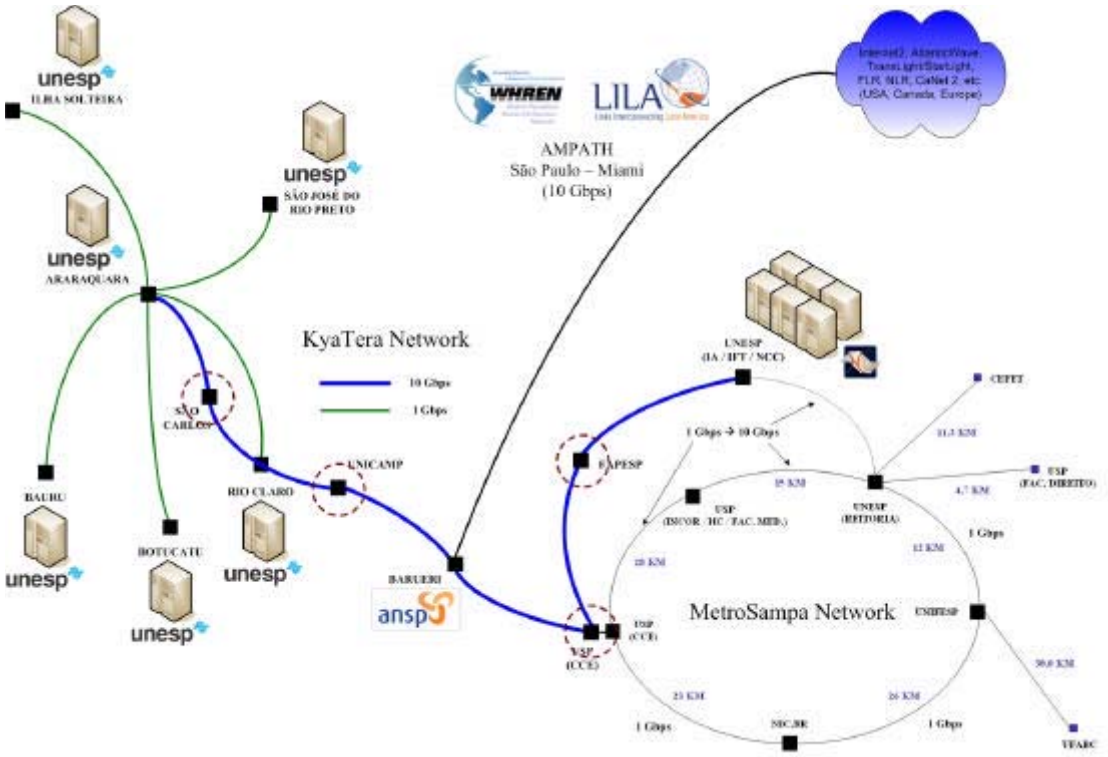

Figure 3. A simplified view of the GridUNESP network infrastructure

Such a dedicated network infrastructure will allow easily the setting up and tearing down of on-demand end-to-end optical connections between the clusters by means of dynamically created virtual LANs. The high-end distributed computing and storage resources, communicating through dedicated channels over an underlying transparent and fully dedicated optical transport infrastructure, put GridUNESP in a privileged position, as it should become the first Lambda Grid deployed in Latin America in the near future.

It should also be considered that even if all the secondary sites simultaneously generates intense traffic by means of continuous streams that fill up their respective links to Araraquara, the aggregated traffic at the central network point for the secondary clusters will reach at most 60% of the pipe from Araraquara to São Paulo, where the central cluster is located. This means that the infrastructure is not only robust enough for the most demanding applications - the ones that, of course, require no more than a 1 Gbps pipe (and it is not so easy to fill up a 1 GbpE link with only a bunch of common applications) -, but it can accommodate extra traffic from other locations. The points marked with dashed circles in Figure 3 are special positions of the dedicated network where the 10G transponders are connected back-to-back. It means that we can easily install active network elements between the transponders for having other entry points in the network.

5. Software infrastructure and preliminary results

The GridUNESP software infrastructure is being built from the ground up. The support team have developed and deployed a set of tools for accessing directory services over the Internet based on OpenLDAP, which provides an excellent framework for distributed enterprise user management and single sign-on. All information regarding the identity of the users (usernames, passwords, privileges, ssh keys, and so on) and hosts (MAC addresses, IP addresses, DHCP services, DNS services, etc) are being organized and stored into a central

repository installed in the central cluster. All the information stored in this central repository is regularly replicated to corresponding repositories within all secondary clusters, assuring high availability and easy access to the most critical services. We have combined OpenLDAP, OpenSSL, e-mail, and other open technologies to produce a top-class resource access management system with an integrated single sign-on mechanism. One immediate benefit is that we can have several administrators with root privileges without having to share a common password.

All the servers are already running CentOS version 5.3, an Enterprise-class Linux distribution derived from sources freely provided by Red Hat. An initial deployment of the Condor scheduler has been done with support to MPI jobs. Our team has recently succeeded in submitting small MPI jobs to Condor that can be scheduled to run on all 2048 cores of the central cluster.

Some preliminary runs of Linpack have also been accomplished. Initial results point to performance of 17.6 TFlops for the central cluster, which is approximately 75% of the theoretical peak performance.

The Grid middleware infrastructure deployment is in its initial stages. A formal partnership has been established with the Open Science Grid Consortium, an organization composed of service and resource providers, researchers from universities and national laboratories, as well as computing centers across the U.S., Asia and Latin America. GridUNESP will use the OSG middleware stack to integrate its computational resources and share them with other R&E institutions worldwide. Our technical team has jut started to focus efforts on the middleware deployment. Regular phone and video conferences between GridUNESP staff and the OSG liaison in U.S. is occurring every week, and the first basic middleware services are being deployed and tested.

6. Summary, conclusions, and future work

The GridUNESP is a computing infrastructure that provides the means for the researchers from the São Paulo State University to work on scientific projects that demand a great deal of computing power. Itmust be considered a long term project that intends to leverage the scientific research activities at UNESP in the coming years by deploying an e-infrastructure composed of a central powerful cluster and a total of seven smaller clusters distributed throughout the state of São Paulo.

Hard work is central to the future success of GridUNESP. We are about to move from a period of time where the underlying physical infrastructure, hardware deployment and testing have been the main focus, to a new period where we must focus on the middleware services towards the goal of building a production-quality Grid that needs to fulfil the requirements of UNESP researchers. We have jut started to focus our efforts on the middleware deployment: regular phone and video conferences between our technical team and the OSG liaison is occurring every week, and the first basic middleware services are being deployed and tested.

During the design of the project, almost 3 years in the past, we envisaged three phases: (1) deployment of the underlying hardware with the appropriate levels of robustness and stability towards a production-quality Grid infrastructure; (2) deployment of the middleware services to integrate the resources into a unified infrastructure that will enable the establishment of a GridUNESP Virtual Organization and the integration with the Open Science Grid; (3) development of an operational model to help our support team in handling not only the daily administrative tasks to keep the robustness and stability of the infrastructure intact during long periods, but also training them to work on user and application support.

Phase 1 is essentially done; phase 2 is on the way.

Acknowledgments

The author wishes to thank the whole project team for their valuable contribution to the deployment of GridUNESP: Marco A. F. Dias, José Roberto B. Gimenez, Sergio M. Lietti, Jadir M. da Silva, Allan Szu, Gabriel von Winckler.

References

[1] G. Fraser, editor (2006). "The New Physics for the twenty-first century". Cambridge University Press.

[2] I. Foster and C. Kesselman, editors (1999). The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers.

[3] I. Foster (2002). "What is the Grid? A three-point checklist". Grid Today, Vol 1, No. 6.

[4] M. Parashar and C. A. Lee (2005). Scanning the Issue: Special Issue on Grid Computing. Proceedings of the IEEE, 93(3):479-484, March 2005.

[5] D. Anderson, "SETI@home", in A. Oram, editor, Peer-to-Peer: Harnessing the benefits of a Disruptive Technology, chapter 5. O'Reilly, 2001.

[6] T. de Fanti, C. De Laat, J. Mambretti, K. Neggers, and B. St. Arnaud, TransLight: A global-scale LambdaGrid for e-science, Comm. of the ACM, 47(11), 2003.

[7] The Western Hemisphere Research and Networking - Links Interconnecting Latin America (WHREN-LILA) initiative. Retrieved October 16, 2009, from http://www.whren-lila.net/