A REGIONAL ANALYSIS CENTER AT THE UNIVERSITY OF FLORIDA

Jorge L. Rodriguez*, Paul Avery, Dimitri Bourilkov, Richard Cavanaugh, Yu Fu, Craig Prescott,
University of Florida, Gainesville, FL 32612, USA

Abstract

The High Energy Physics Group at the University of Florida is involved in a variety of projects ranging from High Energy Experiments at hadron and electron positron colliders to cutting edge computer science experiments focused on grid computing. In support of these activities members of the Florida group have developed and deployed a local computational facility which consists of several service nodes, computational clusters and disk storage servers. The resources contribute collectively or individually to a variety of production and development activities such as: the UFlorida Tier2 centre for the CMS experiment at the Large Hadron Collider (LHC), Monte Carlo production for the CDF experiment at Fermi Lab, the CLEO experiment, and research on grid computing for the GriPhyN and iVDGL projects.

The entire collection of servers, clusters and storage servers is managed as a single facility using the ROCKS cluster management system. Managing the facility as a single centrally managed system enhances our ability to relocate and reconfigure the resources as necessary in support of both research and production activities. In this paper we describe the architecture deployed, including details on our local implementation of the ROCKS systems, how this simplifies the maintenance and administration of the facility.

INTRODUCTION

Scientific research is increasingly more dependent on computational resources as scientists strive to understand problems of greater complexity, sophistication and importance. This is particularly true for high energy physicists working on the Large Hadron Collider (LHC). Understanding the physics at the LHC will require the analysis of peta-scale size data sets with very complicated event topologies these conditions drive the requirement for large amounts of computing resources. Also, the proposed computing model for the LHC calls for the creation of a series of Tier2 centres across the glob to help meet the computing requirement for the experiments. In this paper we present the current status and plans of a facility designed to satisfy these requirement as well as the needs of the local and regional HEP community.

The University of Florida's (UF) Regional Analysis Centre's primary function is to provide the operational support for various organizations from the High Energy Physics (HEP) groups at UF, US national grid computing projects GriPhyN and iVDGL[1] and regional sites within the state of Florida. The facility currently supports Monte Carlo production and physics analyses for the Compact Muon Solenoid (CMS) an experiment at the LHC, the Central Detector Facility (CDF) at Fermi Lab, and the CLEO collaboration an e^+e^- experiment at Cornell[2]. We also support scientific research for other disciplines, including Bio-Science and Astronomy and Astrophysics and other HEP experiments, ATLAS, BTeV through our participation in the Grid3 project[3]. The facility also supports a variety of grid related research and development projects such as: the Sphinx[4] grid scheduling middleware application, CAVES[5], a project designed to explore advanced virtual data concepts for HEP data analysis, and GridCat[6] a site status, monitoring and cataloguing application for the grid. Also, UF personnel are providing guidance and technical support for facilities at other universities across the state of Florida and South America through our participation in iVDGL and the CHEPRO[7] project. CHEPREO is a partnership of US and Brazilian universities dedicated to HEP research and outreach activities in a grid environment.

PHYSICAL INFRASTRUCTURE

Most of the computing hardware for the Regional Centre is physically located in the UF Department of Physics' server room. The department currently provides us with six racks worth of floor space, conditioned power and climate control. It also provides us with our networking feed from the campus WAN. Our computing infrastructure is connected to a 1 Gbps fibre optic line which we expect to upgrade to 10 Gbps by the end of the year. The 10 Gbps equipment is funded through the Florida Lambda Rail initiative an extension of the National Lambda Rail program to institutions across the state of Florida. The equipment for the upgrade is already on order and plans are underway to use some of it in a joint demonstration to saturation a 10 Gbs connection between Jacksonville, the NLR touch down point, and California. The bandwidth challenge demonstration is scheduled to occur during Super Computing 2004.

The computational hardware installed at the Florida Analysis Centre currently consists of 75 dual CPU boxes of mixed PIII/P4 Xeons running the Linux operating system. Most of the servers are connected via a Cisco 4003 switch with a mix of copper Fast Ethernet and GigE fibber ports. Storage is provided by a combination of standalone RAID arrays and large fileservers with built in RAID controllers. The storage devices provide approximately 9.0 Tera Bytes (TB) of space. A little over half (5.4 TB) is made available through our dCache system which consists of one large 2.0 TB dCache pool, and 40 smaller pools co-located on some of our production worker nodes.

We have recently ordered several new machines that will improve our capacity and increase the functionality of our facility. New faster hardware will replace our

*jorge@phys.ufl.edu

current analysis cluster which we expect to take delivery by the end of November 2004. The cluster will be dedicated to analysis of HEP data, primarily CMS anticipating a large increase in the number of users as we ramp up towards full Tier2 operations. We will also double our storage capacity with the new purchases. These will also be dedicated to the HEP analysis efforts at UF.

In addition to the computing infrastructure, our facility provides the infrastructure that allows faculty, staff and students to collaborate with colleagues across the globe. We have equipped two video conferencing rooms with Polycom systems each with its own PC and three visitor's offices with workstations, telephones, printers and various other peripherals. We installed an Access Grid (AG) into our largest departmental conference room. The AG is used to broadcast and participate in remote conferences and lectures and we expect it will play a more significant role in our operations as additional AG sites come on line and the AG software matures.

All of the facilities described above are supported by our group with some help from the Physics Department's computing staff.

FACILITY SYSTEM ADMINISTRATION

Facility Design Overview

The computational resources at the Analysis Centre are managed as a single computing system where all of the servers are installed from a centralized RPM repository and cluster management system. The philosophical approach was to have a well organized, secure and flexible systems administration model where operating system (OS) installations can be managed across the entire infrastructure instead of on a server by server basis. This approach allows us to simplify maintenance and streamline upgrades while maintaining the ability to re design the architecture as the needs of our user community develops and evolves.

Another important consideration in the systems management strategy was the desire to support multiple versions of the RedHat (RH) distribution. In HEP, RH is the dominant Linux distribution used and will most likely be for foreseeable future as the community transitions to Scientific Linux which is based on a recent version of RedHat Enterprise Linux (RHEL). The ability to support multiple versions of the OS distribution simultaneously on the same infrastructure is necessary to keep pace with developments of the OS software stack while supporting the needs of the applications community which seldom develops in coincidence with upgrades of OS distributions. This is particularly true in the HEP community.

Facility Design Implementation

To meet the design criteria outline in the previous section we have adopted a cluster management methodology derived on the ROCKS cluster distribution. ROCKS[8] is an open source cluster distribution and management system whose main function is to deploy the RedHat distribution on to a cluster of computers. ROCKS is layered on top of RedHat's kickstart technology and thus draws heavily on the extensive work done by RedHat to deploy Linux on a very large array of hardware configurations.

ROCKS bundles cluster applications, such as MPICH and batch systems like OpenPBS or Sun Grid Engine on top of the RH distribution. It also enhances kickstart by adding machinery to push distributions out to servers connected to a management node via the private LAN. The machinery relies on DHCP to dispense node identity information, https as the transport protocol and cgi scripts and python code to generate the kickstart description files on the fly.

The kickstart generation engine is the heart of the ROCKS cluster management system. It generates kickstart files by combining node specific information stored in a MySQL database with installation instructions organized into a series of XML specification files. The database stores information describing every node in the cluster identifying each server by its unique MAC address, private network id and other node specific information. It also stores global cluster(s) information such as NIS domain names, host name of the batch system head-node servers an anything else that describes general attributes of the cluster or meta-cluster. All of this information is stored in the MySQL database.

The XML specification files contain the set of instructions needed to install the software on the system. This includes the list of RPMs and system commands that configure the individual server. The XML files are organized as modules in an object oriented architecture making it easy to create new deployments of entire clusters or single service nodes by simply rearranging, adding or modifying existing objects and their dependencies. All of the components collaborate to generate a unique kickstart file for every node as it is installed.

ROCKS at UF

The standard ROCKS distribution installs a single cluster where the administration node is also the main login host and runs most cluster wide services. At Florida we have significantly modified the standard architecture to meet our facility design criteria. Our cluster architecture is depicted in Fig. 1. The key difference between a standard ROCKS installation and ours is the multiple cluster architecture which we support from the single administration nodes. We also support multiple versions of the RH distribution from the same administration node, keeping the most up to date version on WAN exposed nodes, nodes which are vulnerable to attack while keeping older versions on nodes that run legacy applications. Also many of the services were moved from the standard frontend node to other servers to load balance services and to isolate critical component from the WAN. The later is done as a security measure and to improve the reliability of critical service components.

In Fig. 1 each of the servers are represented by boxes labelled with their appliance names. The uflorida-frontend node is the management node for the entire facility. It

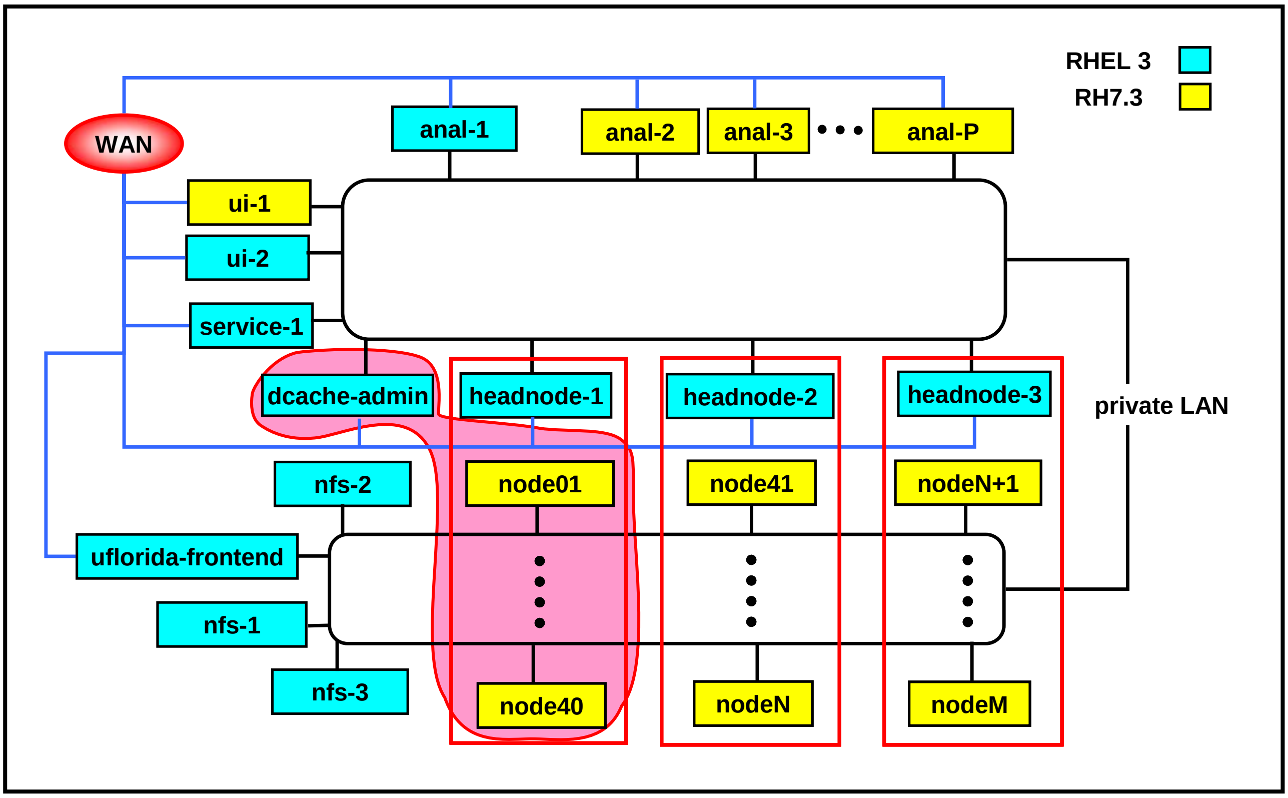

Figure 1: The diagram describes the architecture of the computing facility at UF. There are currently four clusters defined: two of them are Grid3 production clusters one of them is a development cluster all three outline by the red boxes. The fourth cluster, anal-3 through anal-P at the top of the figure, is the analysis cluster . The servers contributing to the dCache installation are outlined by the pink shaped shaded region. The diagram displays other service nodes such as the nfs servers, service node service-1 where we run our web and CVS servers and two grid User Interfaces machines ui-1 and ui-2. The public and private network topology is also shown in the figure.

runs all of the ROCKS specific services required to push distributions out to the servers and is protected from the WAN by very aggressive firewall rules and severely restrictive user access. The figure displays our two production clusters, both of them assigned to Grid3 and each with its own resource allocation policy. The large 80 CPU cluster is the US CMS production facility for our prototype Tier2 centre. These are indicated by the box outlines. Our new CMS analysis cluster is represented by the three servers, towards the top of the figure (anal1-anal3). These machines have been configured to support the CMS analysis community. They allow normal user login ins are equipped with AFS system configured with access to CERN repositories and the CMS application development environment plus they have an assortment of user applications. All of the analysis servers are currently running RH7.3 since this is the only platform fully supported by the CMS application development environment. We try to maintain all other WAN exposed servers installed with latest and most up to date version of the OS available. At the time of this writing these machines are deployed with RHEL3 compiled by the ROCKS team and patched by us with the latest updates from White Box Enterprise Linux[8].

Another important component of our installation, highlighted in the diagram by the pink shape, is our dCache[9] storage system. Our deployment consists of 40 pool nodes and a single dCache administrator node dcache-admin. The admin node is installed on a RAID file server and includes a SRM component. The installation is based on RPM packaging provided to us by the joint FNAL/DESY dCache project. We took the RPMs, installation scripts and instructions and deployed the system through ROCKS with our own customisations. This involved the creation of new ROCKS appliances, one for the dcache-admin node and one for the pool nodes. By moving the entire dCache installation into the ROCKS framework we can treat the installation in its entirety instead of on a node by node basis. While we have successfully deployed the dCache system with ROCKS and have moved files in and out of the systems the SRM component has not been fully tested. Effort is underway to fix the problems and finalize the installation.

We have also deployed other cluster wide applications with ROCKS. The condor and OpenPBS systems have been deployed in this manner and we plan to continue deploying other fabric level infrastructure in this way.

Finally, the figure shows our service nodes. These are machines dedicated to provide particular services and need to be isolated from the rest of the infrastructure, to either enhance security or to reduce load on other servers. For example, our CVS server and group webserver are running on the service-1 machine while the grid User Interfaces are located on ui-1 and ui-2 machine each are running on different platforms.

Finally, we want to emphasize that the entire infrastructure depicted in Fig. 1 was deployed and is managed completely with our customized ROCKS system.

FACILITY'S STATUS AND PLANS

We began operation in the Summer of 2000 shortly after the networking and server hardware was assembled onto racks by our colleagues. Since then the cluster architecture has gone through two major revisions and countless upgrades and modifications. During this period we have been a major US contributor to CMS Monte Carlo production activities. Initially productions jobs were submitted manually to the local batch systems. Since end of 2002 all production activities have occurred through grid interfaces to our local resources. The facility has also contributed to CDF analyses by generating a significant amount of Monte Carlo events for members of the local CDF group. For CLEO-C we expect more users now that their software has been ported to Linux. We expect to continue this tradition and expand the level of service with planned upgrades of both hardware and personnel as we ramp up activities as an official Tier2 centre.

We will continue to develop our cluster management strategies to better support our own activities as they evolve. The ROCKS framework, we feel, will greatly facilitate this effort as we have described in this paper. We also want to orient our implementation in such a way as to improve the sharing of fabric level installations with other facilities within the LHC community and beyond.

CONCLUSIONS

We have successfully deployed, from scratch, a computing facility at the University of Florida. The facility's computing infrastructure was designed to accommodate and support a large and varied community of users from local HEP scientists to Grid computing developers to application scientists from a variety of disciplines.

ACKNOWLEDGEMENTS

I would like to thank those who have contributed to the creation of the Regional Analysis Centre at the University of Florida. In particular Holger Stoeck who kicked things off by working long hours assembling the cluster soon after it arrived in early 2000. I would also like to thank the UF Physics Departmental IT staff, Brent Nelson, Dave Hanson and their student helpers who have provided us not only with tools, loaner parts and the occasional muscular aid but most significantly with technical advice and insight. Finally, I'd like to thank our collaborators at Fermi Lab, other US CMS Tier2 centres the ROCKS team for their support.

This work was supported by the National Science Foundation under grants 0086044 and 0122557 and the Department of Energy under grant DE-FG02-97ER41029.

REFRENCES

[1] http://www.griphyn.org and http://www.ivdgl.org

[2] For a survey of HEP activities at UF please visit our HEE webpages http://www.phys.ufl.edu/hee

[3] http://www.ivdgl.org/grid3

[4] http://sphinx.phys.ufl.edu

[5] http://caves.phys.ufl.edu

[6] http://gridcat.phys.ufl.edu

[7] http://www.chepreo.org

[8] http://www.rocksclusters.org

[9] http://www.whiteboxlinux.org

[10] http://www.dcache.org