OPERATION AND MANAGEMENT OF A HETEROGENEOUS LARGE-SCALE, MULTI-PURPOSE COMPUTER CLUSTER AT BROOKHAVEN NATIONAL LAB

A. Chan*, B. Gibbard, R. Hogue, C. Hollowell, R. Petkus,
R. Popescu, O. Rind, J. Smith, T. Throwe, A. Withers, T. Wlodek
Brookhaven National Laboratory, Upton, NY 11973, USA

Abstract

The operation and management of a heterogeneous large-scale, multi-purpose computer cluster is a complex task given the competing nature of requests for resources by a large, world-wide user base. Besides providing the bulk of the computational resources to experiments at the Relativistic Heavy-Ion Collider (RHIC), this large cluster is part of the U.S. Tier 1 Computing Center for the ATLAS experiment at the LHC, and it provides support to the Large Synoptic Survey Telescope (LSST) project. A description of the existing and planned upgrades in infrastructure, hardware and software architecture that allow efficient usage of computing and distributed storage resources by a geographically diverse user base will be given, followed by a description of near and medium-term computing trends that will play a role in the future growth and direction of this computer cluster.

INTRODUCTION

The RHIC Computing Facility (RCF) is a large scale data processing center established at Brookhaven National Laboratory (BNL) to support the computing needs of the experiments at the Relativistic Heavy Ion Collider (RHIC). The RCF is a full-scale scientific facility, and it provides the bulk of the dedicated data processing, storage and analysis resources for RHIC computing, along with general services such as electronic mail, user account lifecycle management, data backup and archiving, help desk system, web serving and document processing.

In 1997, Brookhaven was also selected as the U.S. Tier 1 computing facility for the ATLAS experiment at the LHC (CERN). The ATLAS Computing Facility (ACF) was established to support the computing needs of the U.S. collaborators in the ATLAS experiment, leveraging the already existing infrastructure and overlapping capabilities of the RCF.

The major components of the RCF/ACF are the 14 TFLOPS Linux Farm (currently with over 4000 processors), the distributed and network-attached disk storage system (1 PB), the robotic tape storage silos (7 PB), the general computing support systems (50 Intel-based servers, 35 Sun Sparc servers and 2 robotic tape libraries), the 3000+ active Gigabit network ports and the grid computing software infrastructure. The hardware is a combination of commodity Intel-based processing servers, enterprise-class UNIX servers and high-specialized mass storage systems connected together by a high-speed network infrastructure.

The transformation of the RCF/ACF from a local into a globally available, distributed computing resource serving over 2400 users and increasingly reliant on Grid-like technologies has resulted in growing design and operational complexity that requires increasing staffing levels (see Fig. 1), changes in facility management procedures and updated hardware & software acquisition procedures.

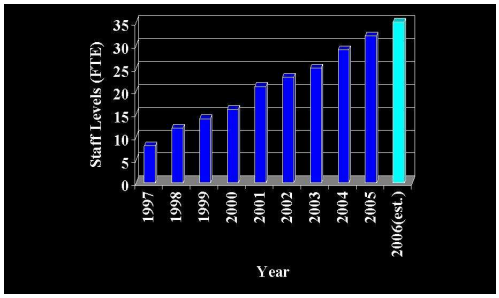

Figure 1: Staffing Levels at the RCF/ACF.

EVOLUTION IN HARDWARE ACQUISITION

The growth of the RCF/ACF (see Fig. 2) from tens of systems to thousands of systems from 1996 to 2005 occurred against the background of falling prices for commodity hardware, as shown in Fig. 3, including cpu's, disks, memory and network cost per port. It not only became affordable to build a large commodity computing cluster, but also a large commodity distributed storage cluster (see Fig. 4) while migrating from 10/100-base to Gigabit-base network connections. These significant improvements will help the RCF/ACF address the computing challenges of the LHC era.

One of the major results of this evolution in hardware performance at the RCF/ACF is that CPU performance is no longer the most important parameter to be considered when designing a cluster or planning hardware purchase. Disk storage capacity and network bandwidth become equally important parameters, specially in the face of large storage and network bandwidth requirements for LHC and RHIC.

* awchan@bnl.gov

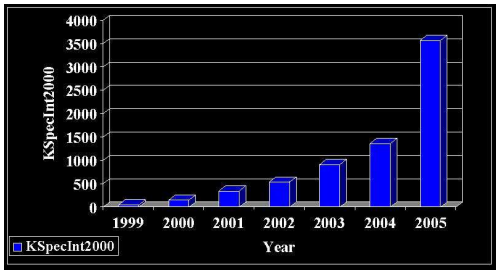

Figure 2: Growth of the Linux Farm processing capacity.

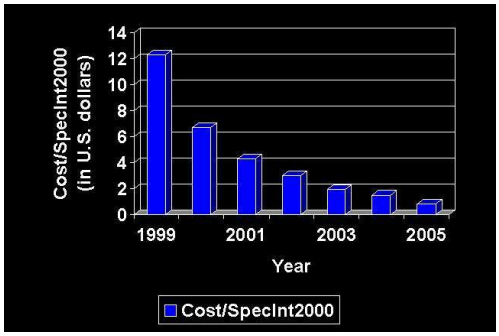

Figure 3: Server cost based on price/performance.

A consequence of the increasing acceptance of the Linux OS in large data centers has been the slow, but steady migration of critical, high availability services such as front-end, disk-caching servers to tape storage systems, network file system (i.e., AFS, NFS) servers, database (Oracle, MySQL) servers and Web servers to commodity hardware. The RCF/ACF has been moving in this direction too, thereby reducing its exposure to expensive, proprietary enterprise-class hardware equiment.

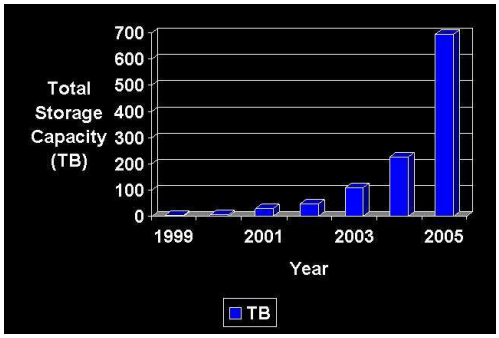

Figure 4: Growth of local storage capacity in the Linux Farm cluster.

SOFTWARE SUPPORT FOR DISTRIBUTED COMPUTING

Gradual changes in software support at the RCF/ACF are also occurring to keep up with changing needs. Currenty, the facility supports numerous software tools that enable a wide array of activities: batch (Condor, LSF), compilers (PGI, Intel, gcc), software development tools (Python, Perl, Java, Totalview, MySQL, Insure++), disk storage solutions (rootd, xrootd, Panasas, dCache, NFS), graphic display software (Grace, GNUplot) and word-processing (TeX).

The overall theme in our software support policy is a migration to widely supported, scalable, open-source software with a high degree of inter-operability suitable for a distributed, Grid-like computing environment that performs well on commodity hardware.

An example of a software package that meets our evolving requirements is the Condor [3] batch system. Condor in 2004 as a replacement for a custom-built batch system and for LSF. Condor's existing features such as support for federation of remote, independent clusters and resource-job requirement matchmaking philosophy meets the requirements of a distributed computing model that the facility is migrating towards. Condor is now available on the 4000+ processor Linux Farm at the RCF/ACF, and it is configured to maximize facility usage by allowing users to schedule jobs on remote idle resources when one's own resources are busy and unavailable. Fig. 5 shows the monitoring interface for Condor that allows the RCF/ACF to track usage of this batch system.

Another example is dCache [4], a grid-enabled distributed disk storage management software, which utilizes the locally-mounted disk systems on the RCF/ACF Linux Farm as a front-end, disk-caching area for the tape storage facility. The dCache software allows us to forgo expensive network-attached storage (NAS) applicances in lieu of cost-effective commodity storage systems while also providing a scalable, fault-tolerant, load-balanced solution for efficient and optimized tape data access.

Table 1 summarizes some of the major transitions in software package support within the RCF/ACF over the last few years.

Table 1: Software Evolution

Package	Old	New	Date
OS	RH Linux	Scien. Linux	2004
Batch	Custom/LSF	Condor	2003
Monitoring	Custom	Ganglia/Nagios	2003/2005
Security	NIS	K5/GSI	2003/2004
Dist. Stor.	–	Rootd/dCache	2001/2004

INFRASTRUCTURE ISSUES

The increasing performance of commodity hardware has been accompanied by rapidly increasing power (and cooling) requirements, and addressing these two needs has become an important issue at various large data centers lately. Figure 6 shows the projected increase in power requirements at the RCF/ACF in the next few years. Options being considered include water-cooled racks, high-efficiency air-cooled racks, low-energy usage cpu's, etc.

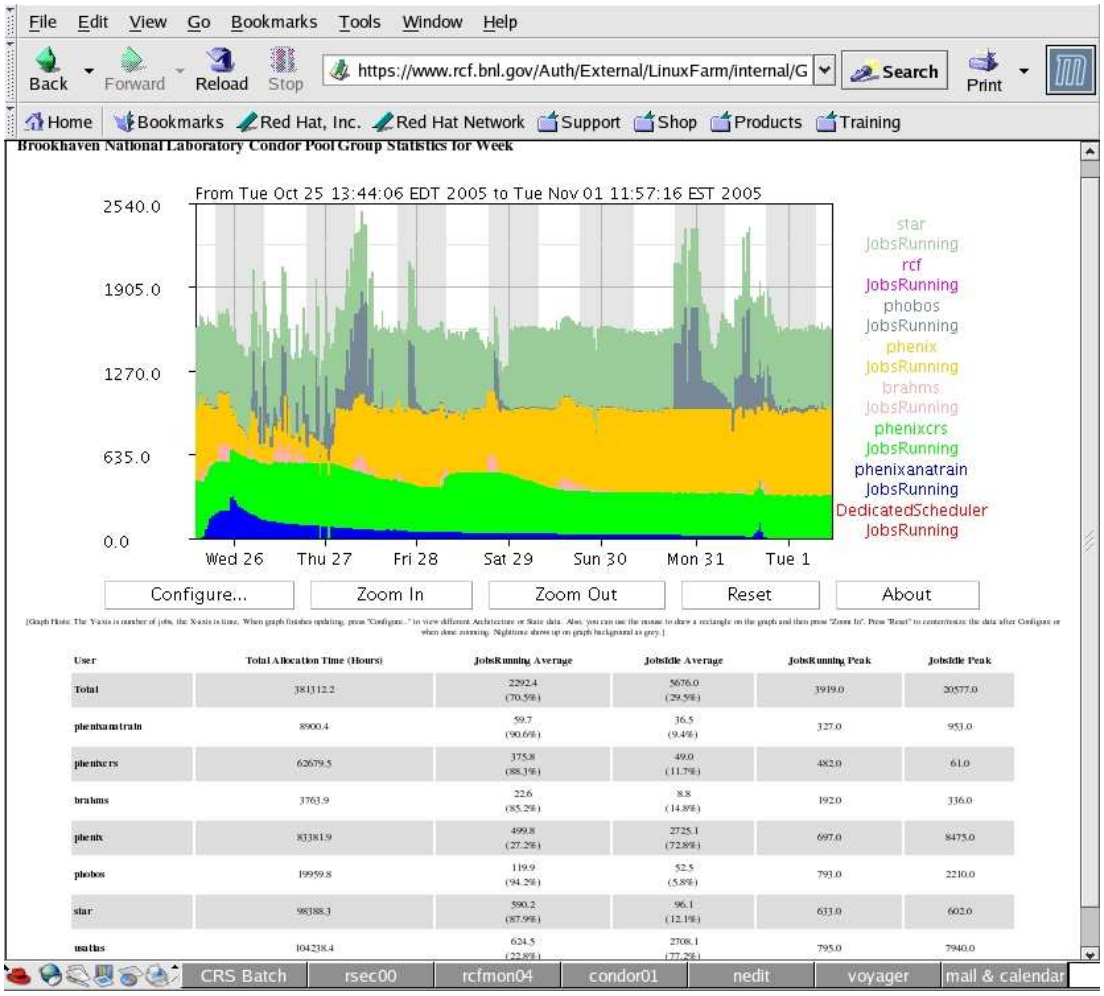

Figure 5: The graphical monitoring interface for the condor batch system.

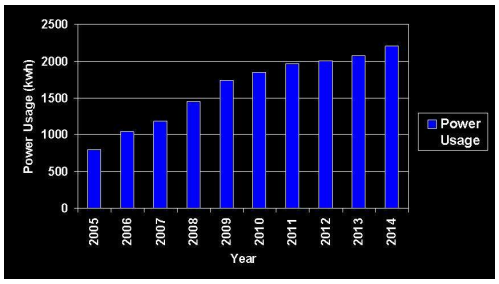

Figure 6: Long-range estimate of RCF/ACF power requirements.

The issue of power reliability is an important consideration for large computing facilities. Sudden loss of electrical power not only disrupts computing activities, but also requires significant manpower effort to restore services. For this reason, the RCF/ACF has a Uninterruptible Power Supply (UPS) unit coupled with a software system that can shut down the facility automatically in an orderly fashion in case of an unexpected power loss. Experience has shown that an orderly shutdown reduces considerably the amount of time and effort needed to power up the facility when power is restored.

The RCF/ACF role as a U.S. Tier 1 computing center for ATLAS requires an upgrade in the network infrastructure at Brookhaven to meet the high network bandwidth needs in the LHC era. Brookhaven recently completed an upgrade to OC-48 external link to handle increasing WAN traffic. The RCF/ACF internal network supports multi-gigabit throughput, which will be upgraded in 2006 to 20 Gbps with full redundancy to match WAN network connectivity.

OTHER CHANGES IN FACILITY MANAGEMENT PHILOSOPHY

The on-going transition of the RCF/ACF from a local into a globally available, distributed computing resource has produced other changes. The facility has migrated from custom-built monitoring system to a combination of scalable, highly configurable and automated monitoring software such as **nagios** [1] (see Fig. 7) and **ganglia** [2] to manage the facility efficiently. It allows the staff to de-

termine the status of the facility quickly and accurately for prompt problem resolution and service restoration. Whenever possible, automated responses to facility status conditions have been implemented to increase efficiency and decrease response time.

Within the 4000+ processor Linux Farm, server systems are no longer treated as individuals but as part of a larger collective. High availability is based on redundancy, not on 24-hour availability of individual systems. Increasingly larger clusters can be managed by a small number of people productively if servers are identically built. For this reason, building a large number of specialized, custom servers within the facility is discouraged.

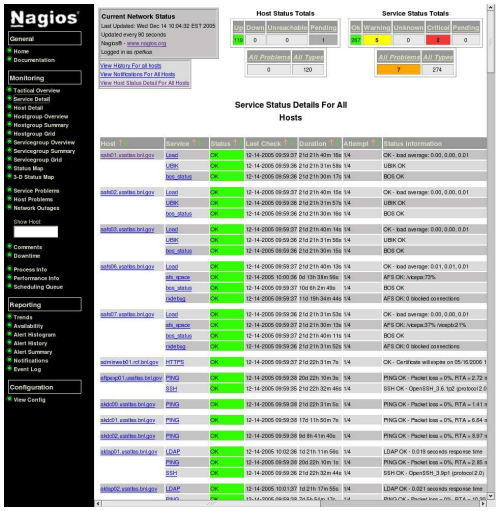

Figure 7: The nagios [1] monitoring software at the RCF/ACF.

NEAR-TERM COMPUTING TRENDS

The appearance of multi-core (2 or more) processors to address cooling and performance issues promises to revolutionize computing in many ways, by providing more power-efficient, high-performance computing cores per unit server.

In similar ways, faster, high capacity RAM and SATA/SAS disks will continue to allow data centers to scale up in size as we approach the LHC turn-on date later this decade. The trend at the RCF/ACF is for distributed storage to continue to diminish the role of expensive NAS appliances as a provider of affordable storage. In addition, with the cost of disk storage per MB approaching the cost of tape storage per MB, it is also likely that disk storage will play an increasingly more prominent role in low-latency, cost-sensitive applications during the LHC operational lifetime.

For a more seamless interaction between the RCF/ACF Tier 1 center and U.S. Tier 2 sites, we expect closer coordination in hardware acquisition, resource allocation & sharing and better integration of distributed computing software support. The ability of Grid middleware to integrate different software packages from different sites will become a more critical element in the distributed computing model of the LHC era at the RCF/ACF.

ACKNOWLEDGEMENTS

The RCF/ACF would like to thank BNL's Physics Department, Information Technology Division (ITD) and the United States Department of Energy (DOE) for their support.

REFERENCES

[1] http://www.nagios.org/

[2] http://ganglia.sourceforge.net/

[3] http://www.cs.wisc.edu/condor

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