

Ofer Rind
HEPiX, GSI-Darmstadt
May 3, 2011

BNL RACF Site Report

RHIC/ATLAS Computing Facility Overview

- Located at the Brookhaven National Laboratory on Long Island, New York
- Provide scientific and general computing needs for the RHIC experiments (STAR, PHENIX)
 - RHIC currently in the midst of Run 11
- US Tier-1 center for ATLAS
- Support for other local research efforts (Daya Bay, LBNE, LSST, EIC, ATLAS Tier-3)
- Currently 34 FTEs (2 openings)

This presentation will focus on updates since the Cornell meeting...

Facility Infrastructure

- In late 2010, reached the 920 KW limit of UPS-backed power in old facilities.
- Excess capacity only available in CDCE (30% of 1 MW in use)
 - Add'l PDU and CRAC units being installed for new ATLAS procurement (June)
 - Ongoing migration of all ATLAS servers and switches to CDCE
- Long overdue upgrade of power/temp monitoring system commenced in March for deployment in early Fall
 - Integrated monitoring to include new areas (CDCE, Sigma7) along with BCF
 - Automated action can be taken in a localized manner

Network

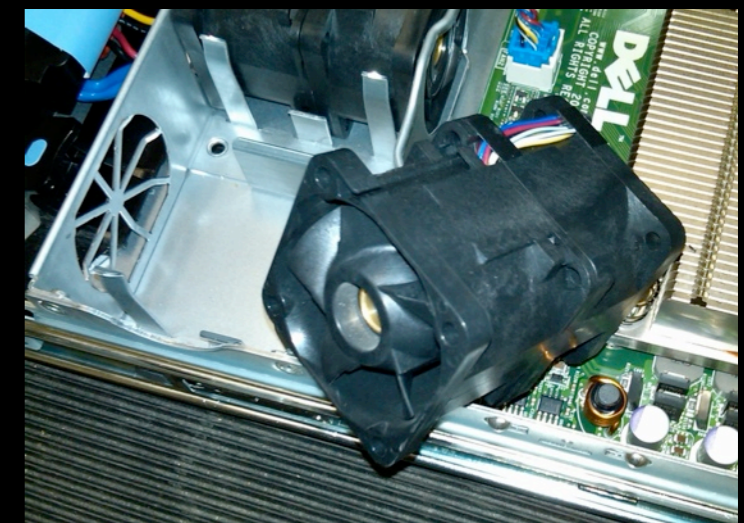
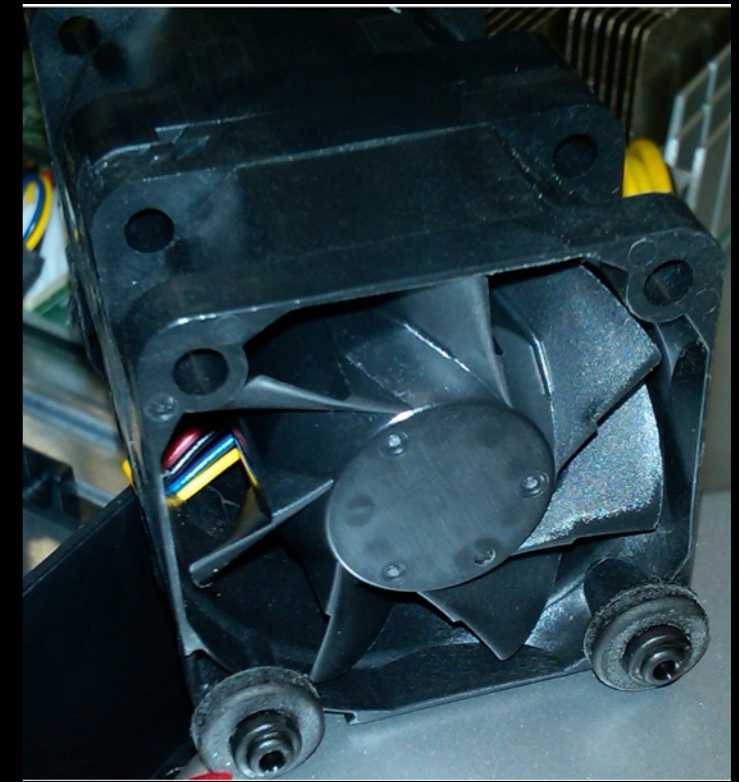
- Major ATLAS network reconfig in January - better utilization of inter-switch links (also, all new inter-switch links are 100 GbE ready)
- Practically all ATLAS systems now line rate connected (1/10 GbE); investigating ramifications of 10 GbE compute nodes
- Eliminated legacy Cisco 6500 ATLAS switches; ongoing retirement for RHIC (down to one by end of '11)
- Considering IPv6 introduction for selected web services and ssh gateways (end FY2012)

Processor Farm Hardware

- Currently, ~16K cores on ~1900 systems
- ATLAS procurement: 142 Dell R410 (dual-Westmere, 12 core), 2.67 Gz, 48 GB, 4x1 TB (expect delivery in June)
 - Tests indicate Westmere uses 37% less power per core (HEP SPEC performance gain is nonlinear)
- RHIC procurement expected in Summer. Smaller procurements for other programs have been placed over past six months

Update On Hardware Issues

- Dell has replaced ~90% of heat sinks to resolve overheating CPU issues reported on in October
- RACF staff finally traced disk failure issue to anomalously high cooling fan vibration
 - Problem reproducible with specialized (HEP SPEC + Bonnie) stress test - problem tracked the fans
 - Dell has replaced the identified problematic fans
 - High fan RPM resulting from temperature/heat sink problems might have caused premature wear of bearings
- Actions to mitigate
 - Rack doors removed to improve airflow
 - Changed BIOS setting from “Maximum Performance” to “Active Power Control,” lowering average fan speed by 20%
 - Add'l BIOS changes (e.g. no turbo boost) reduce power consumption by 15% with only 3% loss in performance.
 - Changes being applied incrementally since April as unobtrusively as possible



Processor Farm Software

- Transition of ATLAS to group quotas in Condor 7.4.2 fairly smooth (PHENIX transitioning soon)
 - Some scheduling priority issues when allowing spillover beyond fixed group quota (production losing resources to analysis) - applied temporary workaround, expect resolution in 7.6?
- Major rewrite of RHIC Central Reconstruction Software (CRS) underway
 - Address long-term PHENIX/STAR requests for asynchronous import/export of files, enhanced job scheduling, feature-rich CLI, redesigned HPSS staging interface, redesigned error handling, upgraded reporting/notification, etc.
 - First stage prototype was available for testing in March
- KSplice Uptrack for centralized rebootless kernel updates
 - Recently deployed on all ATLAS farm nodes and used to patch a kernel (kswapd) bug that was crashing hosts (patch provided in < 2 days)
 - Deployment on RHIC nodes in progress

Distributed Storage

- ATLAS dCache upgrade to 1.9.5-23 in January
 - PNFS: Postgres 9; upgraded fiber-attached storage to increase iops (now 22 drives)
 - SRM: Postgres 9; internal SSD RAID array
 - Added 19 Nexsan Satabeasts for add'l 1.8 PB (9.5 PB total)
 - Retired ~20 remaining 16 TB Thumpers
 - Implemented ATLAS federated XRootd interface
- PHENIX dCache scratch pool hardware upgraded (~70 TB) and write pools reconfigured; Chimera upgrade planning in progress

Distributed Storage

- STAR XRootd upgraded in October (vers. 20100315-1007)
 - New installation and management system; expect eventual migration to Puppet
 - R/W and other functionality to be added
- ATLAS Tier-3 XRootd storage service expanded with procurement of 8 Dell R710's
 - System currently deployed for local use, managed with Puppet
 - Using EPEL repo for XRootd/Proof (for now)
 - Federation and security implementation underway

Mass Storage

- Added 20 LTO-5 tape drives (10 ATLAS, 10 RHIC)
 - Staying with LTO-4 media for now (ongoing migration off of LTO-3)
- Resolved major read/write incompatibility problem between HP and IBM LTO4 tape drives by replacing all ten HP drives with IBM
- Upgrading ATLAS HPSS disk cache and network to increase throughput (8 GB/s peak, 4 GB/s sustained aggregate I/O)
- One Atlas SL8500 library expanded to full 5 module size (10500 slots)
- Major HPSS upgrade to 7.3.2 planned for August (7.1.1 EOL)
 - Migrating core IBM Power/AIX servers to RHEL servers
 - ACSLS (library control) platform migration from IBM Power/AIX to Solaris x86
 - DB2 storage upgrade
- Looking at T10KC technology, but waiting for more experience in the field

Other Services

■ BlueArc NFS

- Initial deployment of BlueArc FC disk is being replaced after 4 years, to be replaced by SAS with LSI rs12 turbo controllers (25% faster than rc12, up to 96 disks)
- Arranging for pre-beta test of BlueArc NFS 4.1 in next few months - would like to stress test with 4.1 client on the 142 new ATLAS farm servers

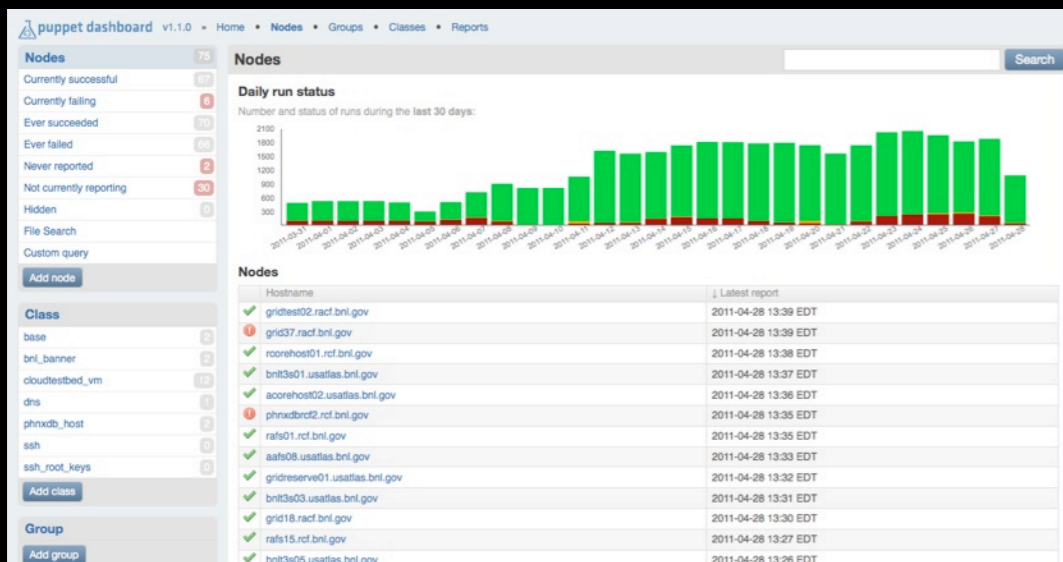
■ Grid/Web services

- Preliminary work on moving from WebAuth web SSO to Shibboleth web SSO for all authenticated pages
- Involved in ATLAS federated identity pilot project
- Established 32-node ATLAS/OSG grid testbed to investigate cloud platform technologies

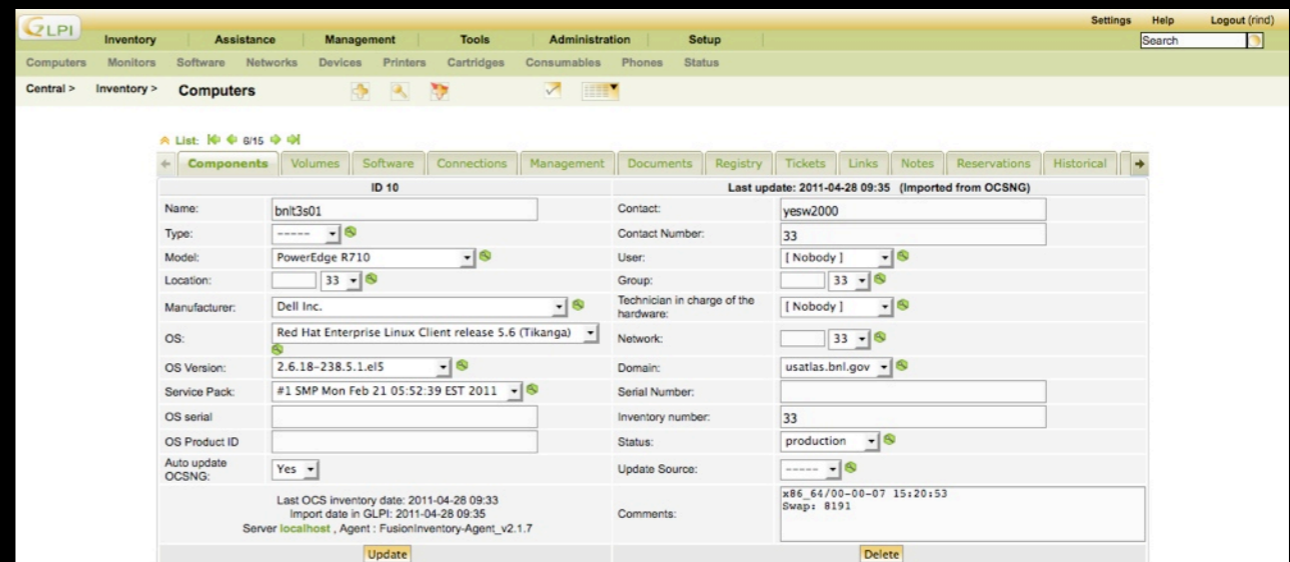
Configuration Management

- In final testing stage and initial production deployment of a new Puppet-based configuration management model
 - Cobbler for base system provisioning and post-install bootstrap of puppet configuration
 - Git for distributed admin access to puppet manifest catalog and historical change management
 - Automatic sync from git to puppet with web-based approval system for production environment
 - GLPI or puppet-dashboard for asset management - used to define puppet node classes & parameters
- Basic modules & classes complete - currently deployed on ~70 hosts, with more added each week, mostly on new system builds

Puppet Dashboard



GLPI



Thanks...and I hope to see you at CHEP 2012! (If not before)

Many thanks to Costin Caramarcu, Chris Hollowell, John Hover, Hiro Ito, Shigeki Misawa, Ognian Novakov, Jason Smith, Will Strecker-Kellogg, Tony Wong, and David Yu for their contributions to this presentation.

The screenshot shows the homepage of the CHEP 2012 website. At the top left is the CHEP 2012 logo, which features a stylized city skyline with the Statue of Liberty and the text 'CHEP 2012'. To the right of the logo is the text 'CHEP 2012 Computing in High Energy and Nuclear Physics 2012 • New York • United States'. In the top right corner, there is a search bar and a notification icon. Below the header is a navigation menu with links for Home, About, History, Registration, Lodging, Program, Submissions, Sponsorship, and Contact. The main content area features a large heading 'Welcome to CHEP 2012!' followed by a paragraph: 'It is our pleasure to announce that the International Conference on Computing in High Energy and Nuclear Physics (CHEP) will be held at New York University's Kimmel Center on May 21-25, 2012.' To the left of this text is a smaller version of the CHEP 2012 logo. Below the welcome message are two columns of text. The left column states: 'The conference will take place at the Helen and Martin Kimmel Center, New York University's hub of campus activity for students, faculty, staff, and alumni. The Kimmel Center is located at 60 Washington Square South.' The right column states: 'One of ten national laboratories overseen and primarily funded by the Office of Science of the U.S. Department of Energy (DOE), Brookhaven National Laboratory conducts research in the physical, biomedical, and environmental sciences, as well as in energy technologies and national security.' To the right of these columns is a 'Conference News' section with links for 'International Advisory Committee', 'CHEP 2013 Proposals', and 'WLCG 2012 Workshop'. Below the news section is a 'Social Program' section featuring a Facebook 'Like' button for 'CHEP 2012' (with 2 likes) and a 'twitter' button. At the bottom of the page, there is a footer with the text 'Home | © 2010-2012 The RHIC and ATLAS Computing Facility at Brookhaven National Laboratory'.

<http://www.chep2012.org>