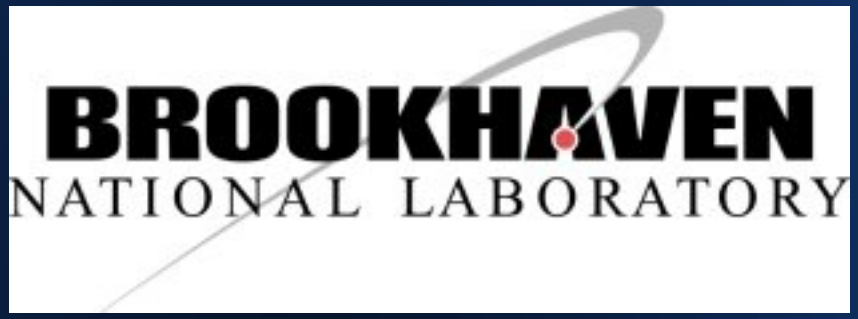
BNL RHIC/ATLAS Computing Facility Site Report

Christopher Hollowell <hollowec@bnl.gov>
RHIC/ATLAS Computing Facility (RACF)
Physics Department
Brookhaven National Laboratory



RHIC/ATLAS Computing Facility (RACF) Overview

Located at Brookhaven National Laboratory (BNL) on Long Island, New York, USA

- One of the primary research facilities at BNL is RHIC – The Relativistic Heavy Ion Collider

Created in the mid 1990's to provide centralized computing services for the RHIC experiments: BRAHMS, PHOBOS, STAR and PHENIX

Expanded our role in the late 1990's to act as the tier1 computing center for ATLAS in the United States

Currently employ 30 FTEs

- Planning on adding several staff members this fiscal year

RACF Overview (Cont.)

LHC 7 TeV run underway for the next 18-24 months
- Expanding resources provided to ATLAS

RHIC Run 10 (high luminosity 200 GeV Gold+Gold) began in late December 2009
- Plan to continue running until June 2010

Mass Storage

Using HPSS as our backend mass storage system
- Upgraded to 7.1.1 in October 2009

~15 PB of data currently in tape

6 StorageTek SL8500 tape libraries
- 4 in production, 2 empty

2 StorageTek Powderhorn 9310 silos
- Plan on retiring these soon

~55 TB total disk cache for HPSS
- Disk cache for RHIC recently upgraded to an IBM DS3400 array

Mass Storage (Cont.)

Custom batch interface - ERADAT

All new data stored on LTO4 tapes - phasing out LTO3 and 9940B tapes/drives

HPSS gateway and core server running AIX 6.1 on IBM P560Q (Power5) hardware

HPSS movers running 64-bit Red Hat Enterprise Linux (RHEL) 5
- Hardware recently upgraded - IBM x3650 servers with dual Xeon X5550 2.67 GHz CPUs and 12 GB of RAM

HPSS mover backend network upgraded from 1 GigE to 10 GigE
- Can now transfer ~900 Megabytes/sec using 9 LTO4 drives

Mass Storage (Cont.)

New STK SL8500 Tape Libraries

NFS Storage

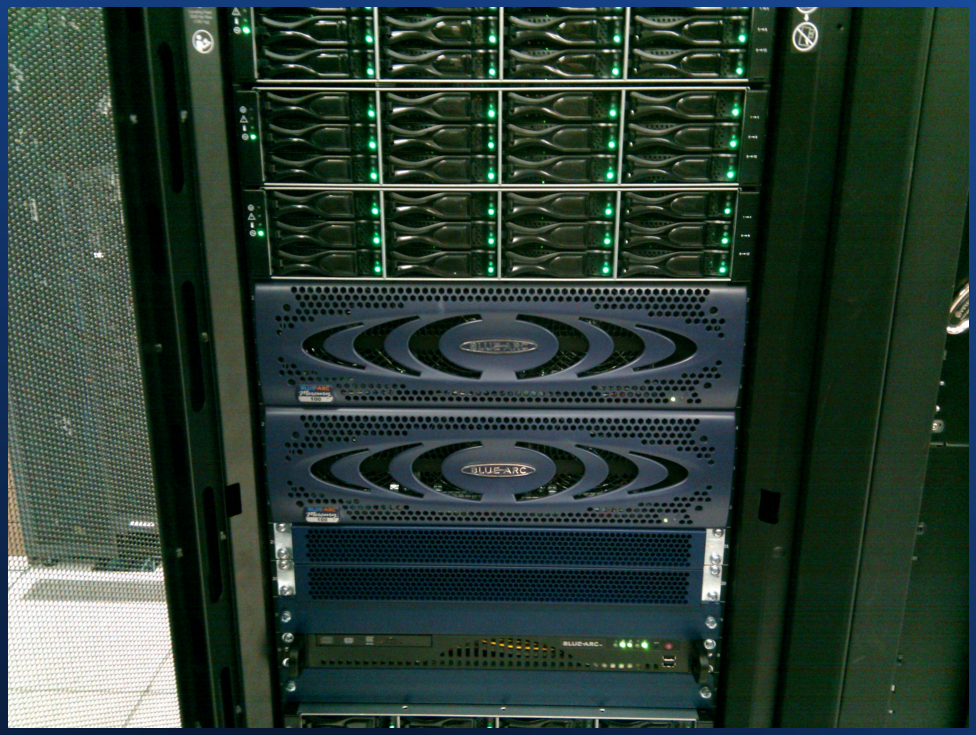
Primarily using BlueArc appliances for NFS service
- Mainly for user home directories and scratch space

Currently have 3 BlueArc clusters serving ~975 TB of storage

RHIC – 6 Titan 3200 heads, split into 2 clusters
- 18 LSI arrays
- 1-2 10 GigE connections per head
- ~850 TB raw storage

ATLAS – Recently put 2 Mercury 100 heads into production
- 4 LSI arrays
- 1 10 GigE connection per head
- ~128 TB raw storage

NFS Storage (Cont.)



BlueArc Mercury 100 Cluster

BlueArc Titan 3200 Cluster

AFS Storage

Running OpenAFS 1.4.11 on most servers

Primarily used for experiment software repositories

Using Teradactyl TiBS for backups

RHIC - ~4.5 TB of space served by 2 file servers
ATLAS - ~3 TB of space served by 2 file servers

Scheduling downtime to upgrade OpenAFS on production file servers is difficult with our current setup

- Plan on purchasing a spare server and storage
- Use of “*vos move*” to relocate data off a server before upgrade

dCache

Maintaining dCache installations for two experiments

ATLAS
- Running version 1.9.4
- ~4.5 PB total disk space
- Testing new DataDirect Networks (DDN) storage: ~2 PB
- Processor farm nodes no longer locally storing pool data
- ~100 Sun X4500/X4540 (“Thumper”/“Thor”) pool servers
  - Running Solaris 10
  - All Sun X4540 systems serving a Fibre channel Nexsan array as well as locally attached storage
  - Up to 120 TB per host

dCache (Cont.)

- PHENIX
  - Running version 1.9.0
  - ~1.1 PB total disk space
    - Most of this storage capacity is provided by the local disk in the processor farm nodes
    - The majority of the data is also in tape

Network

Replaced Cisco core routers with Force10 chassis

Core routing now fully redundant for both RHIC and ATLAS

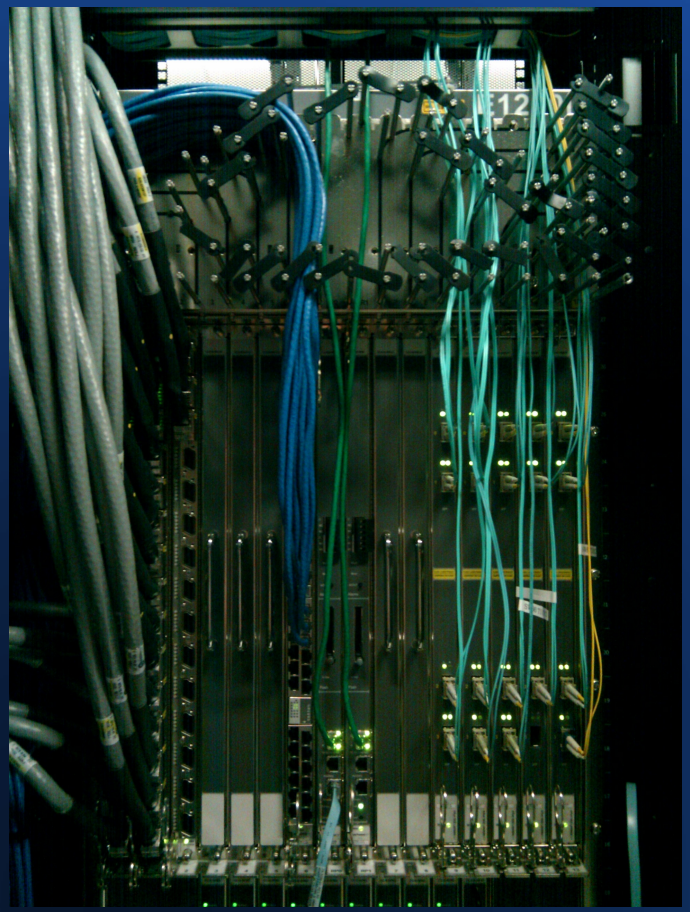
Split PHENIX and STAR subnets into 2 separate switches

60 Gbps total inter-switch links for ATLAS, 20 Gbps for RHIC
Upgrading ATLAS to 80 Gbps

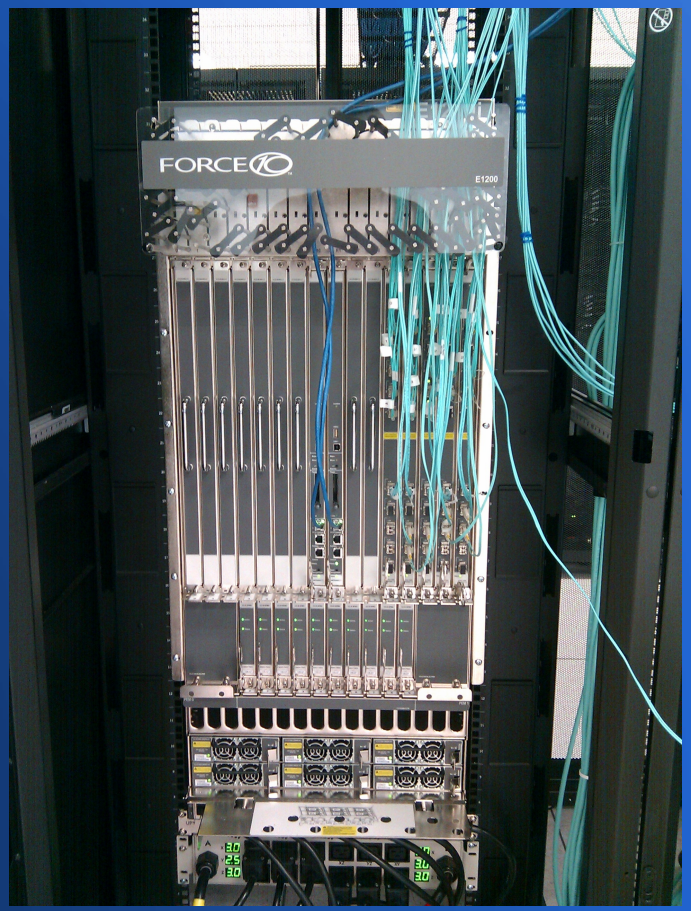
20 Gbps total Internet bandwidth

All ATLAS processor farm 1 GigE ports now non-blocking

Network (Cont.)



Force10 TeraScale Chassis



Force10 ExaScale Chassis

Grid Activities

Two experiments we support are members of OSG, and run jobs at our facility via the grid: US-ATLAS and STAR

OSG released software stack version 1.2
- Feature complete and reliable

In the process of upgrading grid gatekeeper hosts to 64-bit RHEL5

OSG released an official glexec RPM
- In the process of testing the use of glexec at our site

Using Condor-G 7.4.2 for grid job submission
- Scheduling performance is significantly better than in 7.3.x

General Services

- WWW servers, DB servers, SSH/FTP gateways, centralized monitoring, DNS servers, mail servers, LDAP servers, testbeds, etc.
  - Currently ~250 machines
- All hosts running RHEL4 or RHEL5: system deployment and management via Red Hat Network (RHN) Satellite
  - This solution works, but not optimal for such varied services
  - Investigating other system management tools
    - Configuration management
      - Puppet
      - Cfengine
    - Asset management
      - OCS Inventory
      - GLPI

General Services (Cont.)

- Increased use of virtualization to reduce operating costs and make optimal use of modern multicore hardware
  - Currently using Xen 3.0.3 as shipped with RHEL5
  - Considering switching to alternative hypervisor/container technology
    - KVM
    - Parallels Containers
- Need a virtualization management toolkit to simplify administration/deployment of heterogeneous virtual machines on multiple host/Dom0 machines
  - RHEV - Expensive
  - Convirt

Processor Farm

~2,000 systems, providing ~10,000 CPU cores

Purchasing 250 new Dell R410 servers for ATLAS
- 24 GB of DDR3-1333 RAM each
- Dual 2.8 GHz Xeon X5560 processors
- 2,000 additional cores

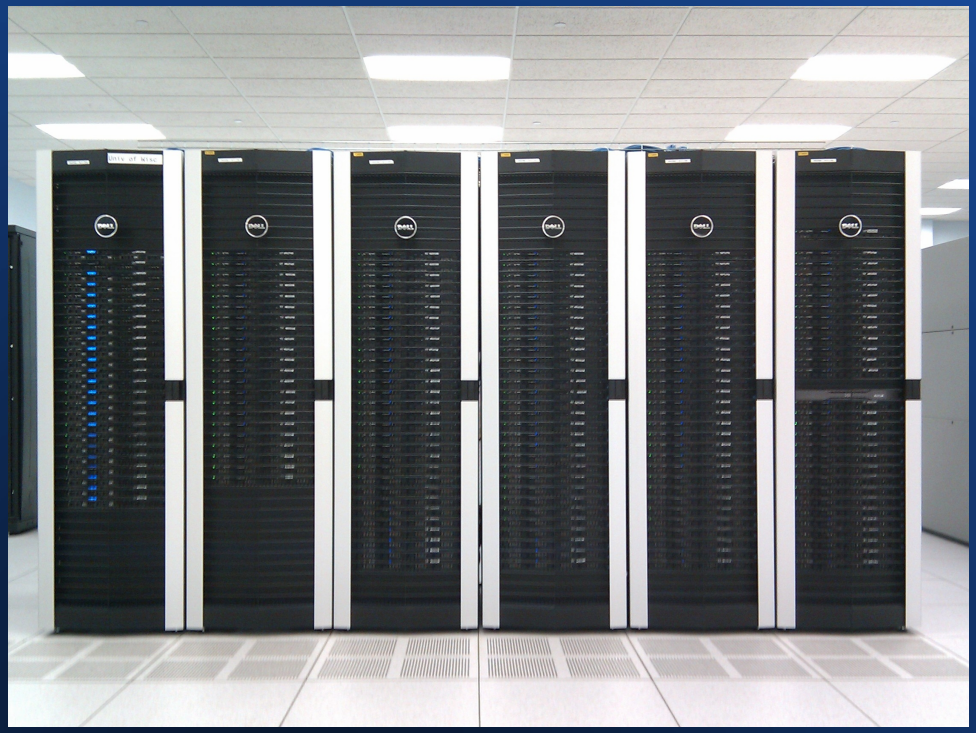
RHIC nodes are disk heavy
- Used for storing copies of data in tape for fast access
- Up to 6 TB of storage per system
- ~2.7 PB of total local node storage space
- Managed by dCache, Rootd, or XRootd

Virtualizing interactive/submit hosts with Xen

Processor Farm (Cont.)

- OS upgrade to 64-bit Scientific Linux (SL) 5.3 Summer/Fall 2009
  - Previously running 32-bit SL 4.4
  - Most experiments still running 32-bit code
    - Had to ensure all necessary 32-bit compatibility libraries were installed
- Automated/mass OS deployment via SL Kickstart and PXE
  - In-house developed asset management software
  - Custom PXE installation control software
  - Packages installed from a local HTTP repository
- Using Condor as our batch system
  - In the process of upgrading to 7.4.x
  - 7.4.1 already running on our central manager hosts

Processor Farm (Cont.)



New ATLAS Dell R410 Servers

Liebert XD Overhead Cooling

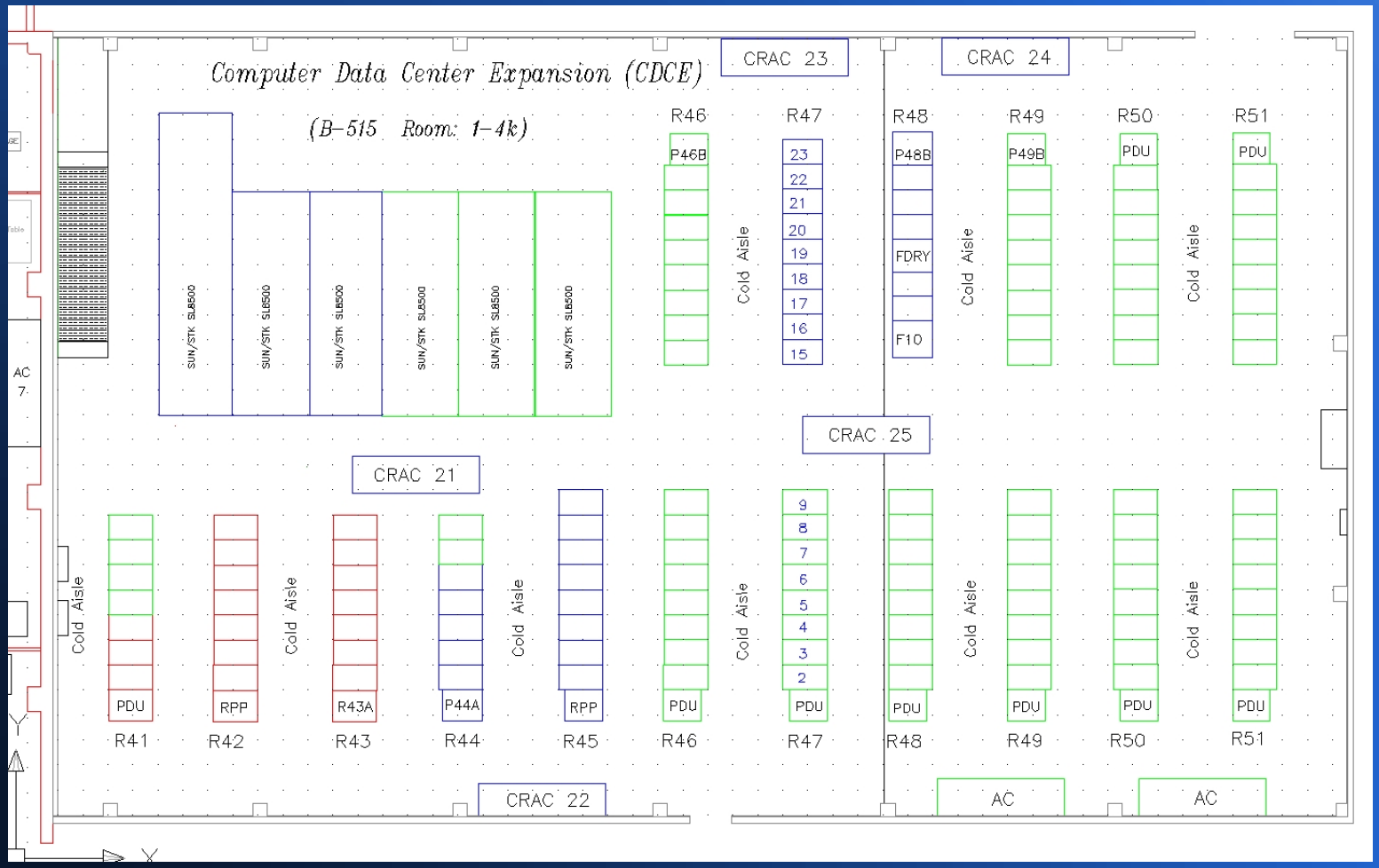
Computer Data Center Expansion (CDCE) Project

Completed Fall 2009

Added ~6,000 square feet of space to our data center

Various new features integrated
- 3 foot raised floors
- Timed lights
- Sloped floor with drain
- Underfloor power cable trays
- Overhead network wiring
- Ample space for on-floor A/Cs

CDCE Layout


Computer Data Center Expansion (CDCE)
(B-515 Room: 1-4k)
CRAC 23
CRAC 24
CRAC 25
CRAC 21
CRAC 22
AC
AC
AC 7
SUN/STK SL8500
Cold Aisle
PDU
RPP
R43A
P44A
P46B
P48B
P49B
FDRY
F10
R41
R42
R43
R44
R45
R46
R47
R48
R49
R50
R51

Questions?

Thanks to the following people at BNL for contributing some of the information presented:

Maurice Askinazi, Tony Chan, Dave Free, Richard Hogue, John Hover, John McCarthy, Shigeki Misawa, James Pryor, Ofer Rind, Pedro Salgado, David Yu, Xin Zhao