BNL RACF Site Report

HEPIX Fall 2014 – Upton, NY, USA

Shigeki Misawa

RHIC/US Atlas Computing Facility

- Locate at Brookhaven National Laboratory, home of these major US Department of Energy scientific user facilities
  - Relativistic Heavy Ion Collider (RHIC)
  - National Synchrotron Light Source II (NSLS-II)
- RACF hosts the following:
  - Tier 0 computing for experiments at RHIC
  - Tier 1 computing for the ATLAS experiment at CERN
- Supports smaller Nuclear and High Energy Physics groups
  - LSST, Daya Bay, EIC, ...

Changing Times at the RACF

- RACF plays a key role in BNL's C3D (Center for Data Driven Discovery)
  - Proven track record with “Big Data”
  - Can provide key services
    - High performance network
    - Compute Farm (Batch/Grid/Cloud)
    - High performance, multi-petabyte scale “On line” disk storage
    - Near line mass storage
    - Data transfer services
- In the era of “Big Data”, RACF expects dozens of new experiments with GB/sec data rates and multi-Petabyte data volumes scattered throughout the BNL campus in small laboratories.
- In support of these changes, BNL is looking to build a new data center in the former NSLS building (See Imran Latif's talk on Thursday for more details)

Supporting New Groups

- Providing disk storage for the new BNL HPC Cluster
- Making all data accessible from the HPC cluster and the RACF HTC clusters
- Providing compute/storage services for a research group at the Center for Functional Nanomaterials (CFN)
- Providing archival storage for the Collider Accelerator Division (C-AD)
- See Will Strecker-Kellogg's talk on Friday for details

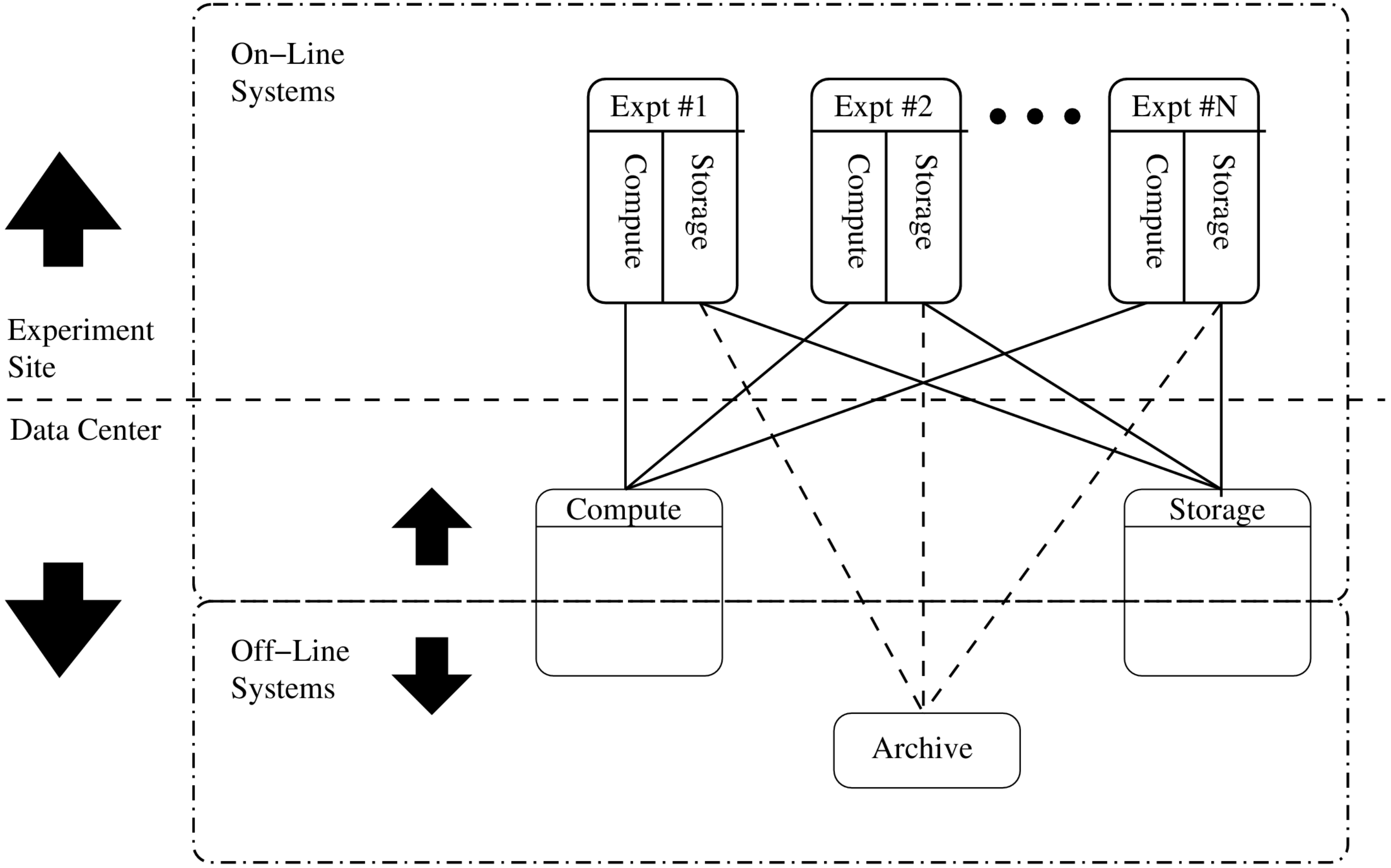
New “HPC Core” Network for “Big Data”

- Joint RACF/BNL Networking project
- Provides high performance network for multiple purposes
  - Internal RACF connectivity
  - Connectivity to existing RACF customers (RHIC counting houses)
  - Connectivity for new RACF customers (CFN/C-AD)
  - Connectivity to the new BNL HPC cluster
  - Connectivity for any other scientific organization at BNL requiring high performance network connectivity

HPC Core Characteristics

- Scalable connectivity
  - 10GbE/40GbE/100GbE (Initial port counts ~1000/~300/~100)
  - Data center (SR) /Campus (LR) distances
  - Initial deployment 60Tb/s system throughput
  - Ports and Bandwidth can scale by adding components
- Line rate capability
  - Connectivity between end points is at full bandwidth
  - No firewalls or data transfer nodes (no “DTN”)
- Security
  - Able to selectively enable connectivity between network regions
  - No router ACL's needed

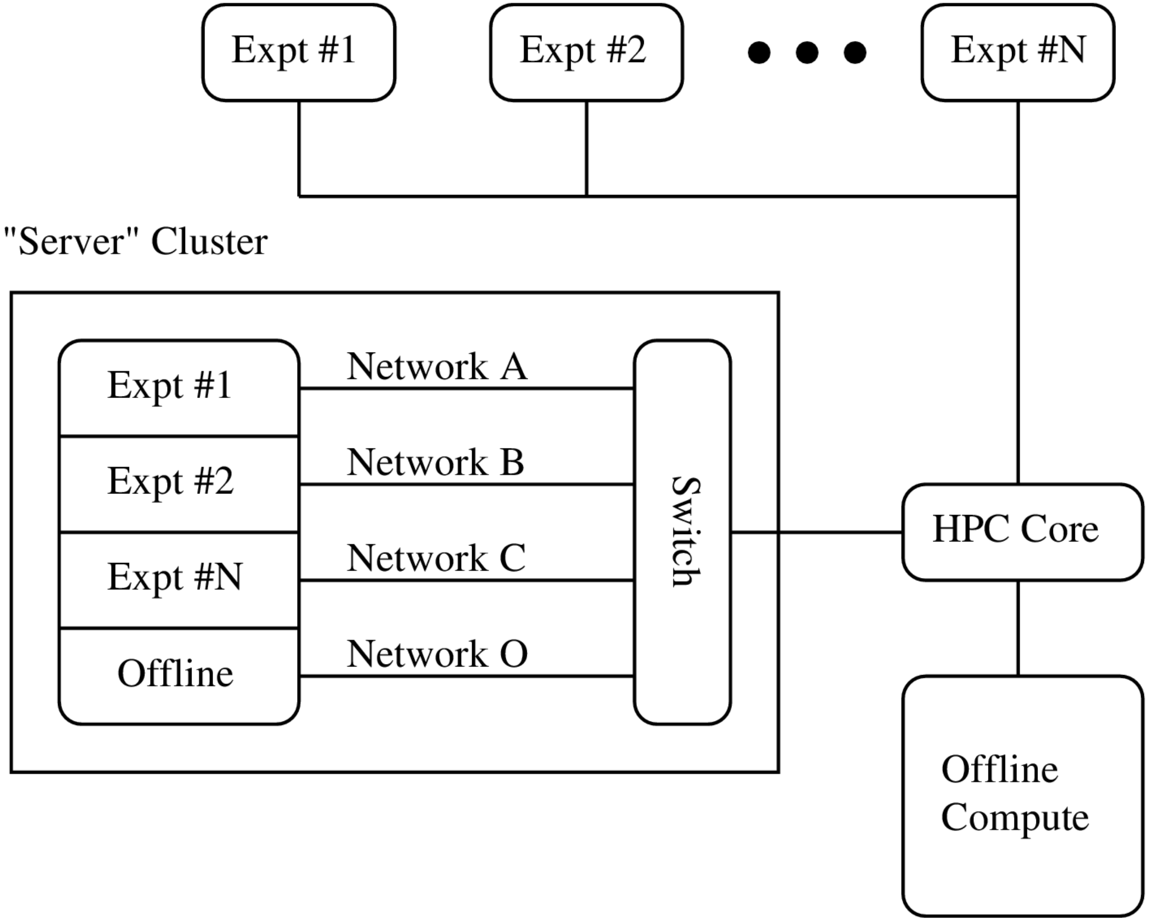
Application of HPC Core Capabilities



HPC Core Capabilities

- “Move” storage and compute resources between “logical domains”
  - Between on-line and off-line
  - Between different experiments
- Capabilities complements the rapid compute server provisioning capabilities of VM environments (OpenStack/Vmware) and bare metal provision environments (Puppet/Chef)
- Enables separating compute and storage from the physical location of an experiment
- Allows for shared access to storage services (and other shared services) while still maintaining isolation of clients

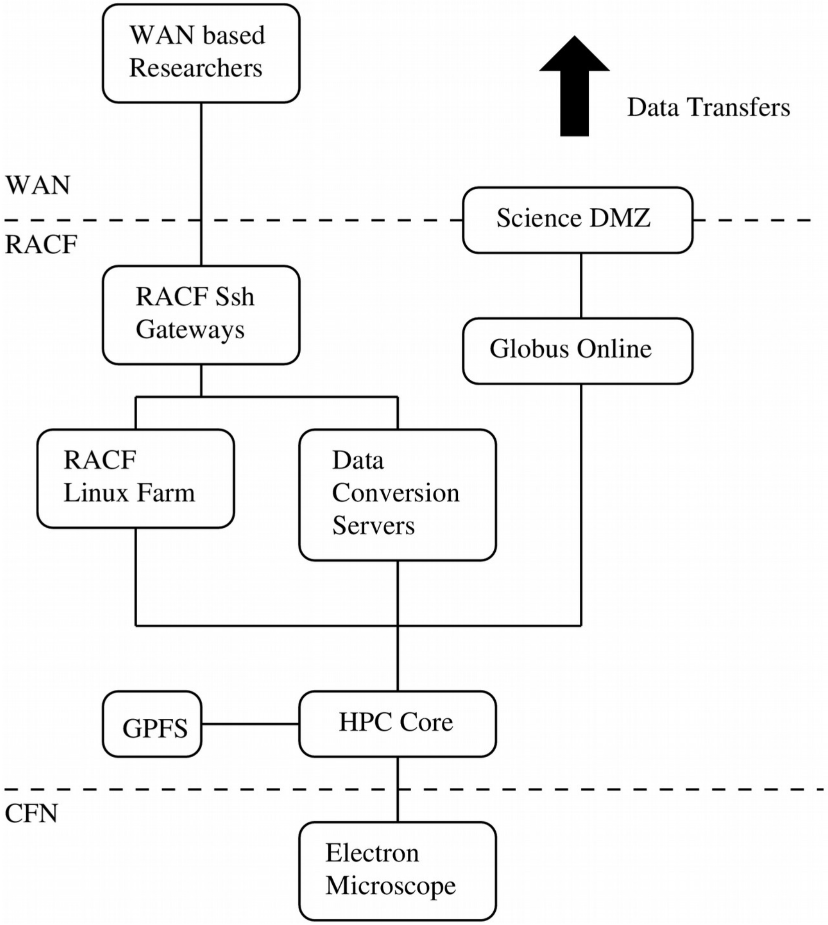
Dynamic Allocation of Compute



First Application of HPC Core

- CFN Electron Microscopy Group
  - K2 Imaging system capable of generating 4TB of data in 15 minutes
  - Experiments limited by
    - Time needed to move data off the system
    - Data conversion took over two days
    - Disk storage limited to 32TB
    - No physical space for additional data processing/storage equipment
    - All components of the system managed by one PC
      - Imaging system control
      - Data acquisition
      - Data storage
      - Data processing

CFN Big Data Solution



CFN Solution

- GPFS replaced local ISCSI disk array. Moving data off imaging system now capped by 10GbE connectivity of control PC
- GPFS capacity is 288TB (9x more storage than before)
- GPFS server physically located in the RACF data center
- GPFS file system is visible to all compute nodes within the RACF facility
- Compute resources of the RACF Linux Farm now available to the CFN group
- WAN transfers of data via RACF Globus Online endpoint augments traditional “FedEx/UPS/DHL/Post office” method of moving data
- Network security of the Imaging system is mostly intact
- CFN access to HPSS mass storage system is now possible for archive storage.

Other Network Changes

- Migration of US ATLAS Tier 1 WAN presence to BNL Science DMZ
- RHIC presence in the Science DMZ for high bandwidth WAN transfers of data to OSG sites
- End of Support for 1GbE Linux Farm switches
  - Dell/Force 10 Exascale CY2020
  - Brocade RX-16 CY2018
- Moving to Brocade SX-1600 switches for 1GbE Atlas Linux Farm connectivity
- RHIC already transitioning to 10GbE (Arista)
- Last Dell/Force 10 TeraScale being replaced with Brocade SX-1600

Linux Farm

- Hardware
  - ~52K Cores/2300 servers
  - RHIC FY2015 – No compute node purchases
  - Atlas FY2015 – 88 Dell R430 servers, 2 x Xeon 2660v3 96GB DDR4, 4x2GB 7.2K RPM SAS drives
  - Tested NVMe SSD (see Chris Hollowell's talk on Thurs)
- Software
  - Running SL-6
  - Still testing SL-7, no plans to upgrade at this time.
  - Testing Docker (See Chris Hollowell's other talk on Wed.)

HPSS Mass Storage

- Completed transition to LTO-6 for all storage classes
- Actively migrating data on LTO-3/LTO-4 cartridges to dual copy LTO-6/T10K-D
- Tape Drives in the system
  - 23 LTO-3
  - 45 LTO-4
  - 42 LTO-5
  - 49 LTO-6
  - 6 T10K
- 9 SL-8500 tape libraries
  - ~57K tapes
  - 63.4 PBytes (10^15) of data, 99.7K files

HPSS Mass Storage

- RHIC HPSS mover (i.e. server) refresh in progress for RHIC
- Atlas HPSS mover refresh completed in FY2015
- HPSS disk cache upgrade in progress
  - 1.6 PB of disk cache (8x increase)
  - Expect 2x to 2.5x increase in disk performance
- HPSS network upgrade next week (move to HPC core)
- HPSS software upgrade to version 7.4.3p2 in November
  - Support for RHEL/SL 7 clients

Disk Storage

- “Legacy” NFS
  - Four Hitachi Data Systems HNAS 4100 heads in production
  - Remaining two BlueArc/HDS Titan-3200 retiring CY2015
  - BlueArc/HDS Mercury retired
  - Retiring NFS storage will be replaced with GPFS
- GPFS
  - Three NSD clusters, 12 NSD total
  - Three client clusters
  - Capacity and performance expansion in the works
  - NSD expansion under consideration
  - In use for on line data collection by the CFN Electron Microscopy Group
  - JBOD and HW RAID in use on NSD's

Disk Storage (cont'd)

- ATLAS dCache
  - Moving to asymmetric file replication (primary/secondary)
  - Primary copy on Hitachi Data Systems HW RAID systems
  - Secondary copy on new JBOD storage with 8TB Seagate SMR disks (6.336 PB usable space) and other older HW RAID systems
  - 12.2PB of usable primary storage, ~10PB of usable secondary storage
  - Current issues include “right sizing” Java memory settings and RAM on Dcache storage nodes (128GB of DRAM not sufficient)
  - For more details see Zhenping Liu talk on Thurs

Disk Storage (cont'd)

- Production Ceph cluster
  - Running Ceph 0.94.3
  - Used for ATLAS Event Service
  - RAM upgraded in cluster
  - CephFS with GridFTP front end in production for 3 months
  - RAM upgrade for production cluster
  - For more details see Alexandr Zaytsev''s talk on Wed

Disk Storage (cont'd)

- PHENIX dCache
  - All disks are on the PHENIX compute nodes
  - 7.6 PB of storage space
  - New 2x10GbE Dcache door in Science DMZ for event reconstruction jobs on OSG resources
- STAR Xrootd
  - All disks on the STAR compute nodes
  - 8.2 PB of storage space
- OpenStack Swift implemented on RACF Cloud – no production use at this time.

Amazon Pilot Project

- Since Sept 2014, pilot project to demonstrate ATLAS at full scale on Amazon EC2
- Most recent test (Sept 2015)
  - Ran medium scale test ~50K cores /~6000 Spot instances for ~5 days
  - 3 instance types (PV and HVM) in US East region
  - All VM”s programmatically created/contextualized vi Imagefactory and Puppet
  - Scheduling automatically managed by AutoPyFactory
  - Successfully completed large, useful physics task for Atlas
  - Cost of the run $23K, but covered by the Amazon research grant.

OpenStack in Production

- Current production environment
  - Running Ice House
  - 47 compute nodes
  - Customers include
    - Atlas
    - Atlas Tier 3
    - BNL Biology
    - BNL Computer Science Group
- Next generation environment
  - “New” servers (compute nodes being retired by the Linux Farm)
  - Will move to Kilo

Questions ?

- Note that we will have two tours of the RACF data center
  - ~5:50PM Today (Wednesday)
  - ~5:50PM Tomorrow (Thursday)
- Each tour is limited to about 15 people
- Send mail to hepix2015@bnl.gov to join the tour