Rosen Center
for Advanced
Computing

Storage at Purdue

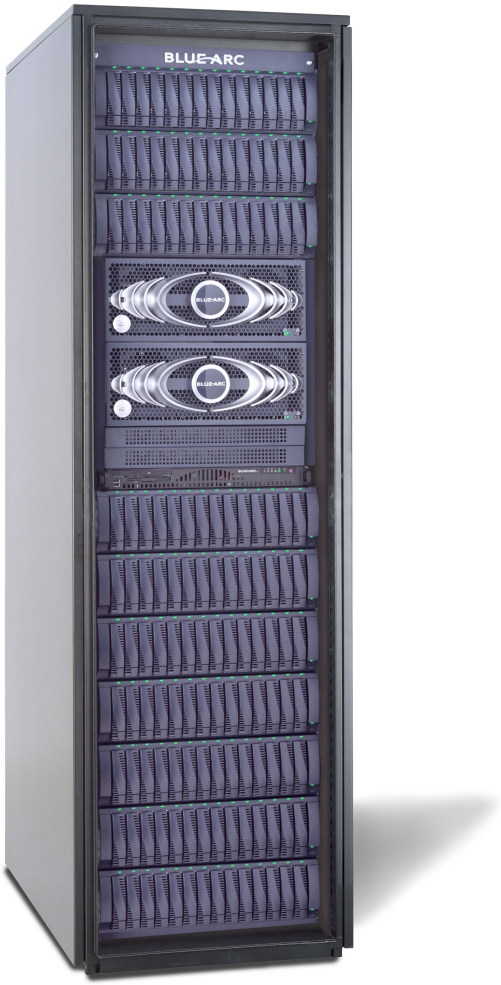
Sept 22, 2010

Preston Smith psmith@purdue.edu

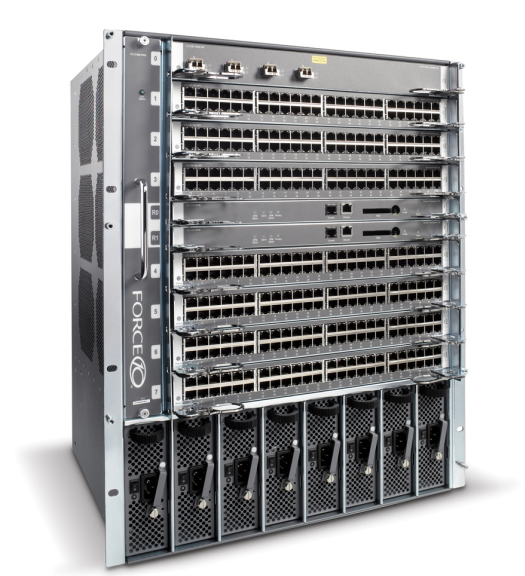
- HPC Resources at Purdue
  - *"Community Clusters"*
    - Central acquisition and operation of HPC resources, with nodes funded by faculty funds (grants, startup packages)
    - Facilities, system administration, storage, networking, etc. all centrally provided by the University
    - 3 systems in production today:
      - **Steele** (893 nodes, dual 2.33 GHz Quad-Core Intel E5410, GigE interconnect)
      - **Coates** (974 nodes, dual 2.5 GHz Quad-Core AMD 2380, 10GbE interconnect)
      - **Rossmann** (384 nodes, dual 2.1 GHz 12-Core AMD 6172, 10GbE interconnect)

Central Storage

- Central Storage
  - Home directories:
    - All homes in RCAC served by 60TB BlueArc Titan NAS
      - Local CMS users and users from OSG all get BlueArc space
  - General-purpose scratch:
    - Second 120TB BlueArc Titan NAS provides scratch for Steele
      - As well as shared application space
    - 200 TB Lustre Filesystem provides scratch for Coates and Rossmann
  - Archive
    - 1.3 PB DXUL archival system available to users. Upgrade to HPSS planned for late 2010.



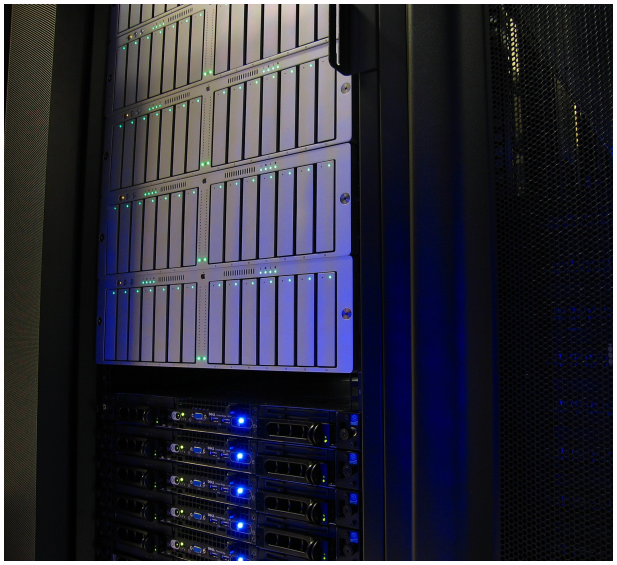
- All dCache nodes are on public IP space
  - Research network core provided by Cisco 6509s, maintained by campus data network staff
    - dCache connected Force10 C300, which is connected to core at 20 Gb/s
    - 10 GbE nodes connected to Cisco Nexus switch, connected to C300 at layer 2
    - 9000-byte MTUs everywhere



Rosen Center for Advanced Computing

dCache at Purdue

- CMS storage is provided by dCache, version 1.9.5-19
- **~960TB of usable space**
- Hybrid of resilient dCache and non-resilient RAID pools
  - 321 TB of resilient space
  - ~635 TB of nonresilient space



dCache Hardware

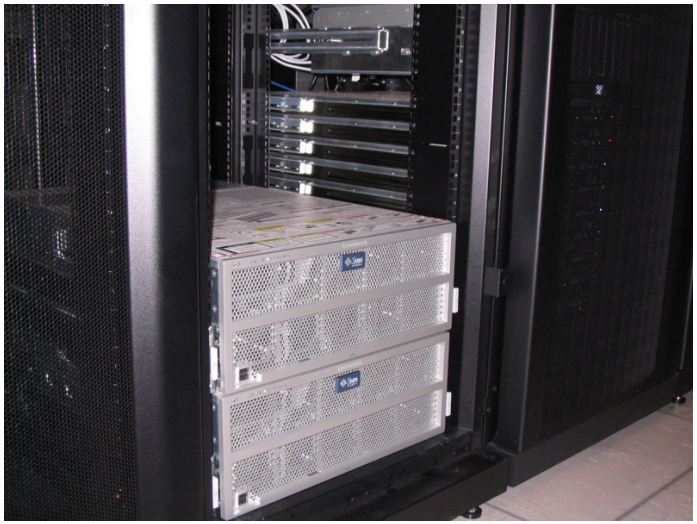
- Core servers
  - Admin node (PoolManager, admin interface, gPlazma, replica manager)
    - Sun x2200 (Dual-socket, dual-core Opteron 2218, 4GB RAM)
  - SRM node
    - HP ProLiant DL185 G5 (Dual-socket, quad core Opteron 2380, 16 GB RAM)
  - Chimera
    - HP ProLiant DL185 G5 (Dual-socket, quad core Opteron 2380, 16 GB RAM)

dCache Hardware

- Resilient Nodes
  - 70 Sun X2200 Compute nodes
    - Sun x2200 (Dual-socket, quad-core Opteron 2376, 8GB RAM)
    - 2 750 GB disks
  - 154 Dell 1950 Compute Nodes
    - Dell 1950 (Dual-socket, quad core Xeon E5410, 16 GB RAM)
    - 2 1 TB disks
  - All resilient nodes are connected at 1Gbit

dCache Hardware

- RAID Servers
  - 2 Sun x4500 “Thumpers”
    - 24TB ZFS
    - 4-way LACP bond (4 Gbit)
  - 2 newer models of Sun x4500 “Thumpers”
    - 48 TB ZFS
    - 4-way LACP bond (4 Gbit)
  - 3 Sun X4540 “Thors”
    - 48 TB ZFS
    - 10 GbE
  - 11 White-box storage servers
    - 48 TB RAID-6 XFS, hotspares
    - 10 GbE



dCache Doors

- dCache Doors
  - As described already – dedicated server for SRM
  - Dcap doors on 5 servers
    - CMS TFC customized to select different door based on filename, to spread dcap load around
  - GridFTP doors on 8 servers
    - RAID storage servers double as GridFTP servers
  - Xrootd doors
    - Used by PROOF cluster and distributed xrootd service

Now, where are the bodies buried?

- dCache experiences
  - Despite what some may say, we are not dCache fanboys
    - We've operated a large installation stably for 5 years – the enemy we know is better than one we don't!
  - Some things work really well
    - Chimera is great!
    - System is fast overall
    - We can implement powerful storage polices with combining replica manager, path-based rules, and some scripting
  - Some are less good
    - Had isses with dCache respecting secondary groups
    - ACLs
    - Have had issues in the past with doors getting stuck

Useful Tools

- Tools and scripts that we use
  - Write-protect nearly full pools
    - Bad things happen when pool filesystems fill to 100%
  - chimera-dump.py
    - http://www-zeuthen.desy.de/~leffhalm/chimera-dump.html
    - Can do a lot of things with output from this tool
  - Path-based replication policies
    - Used UNL's PFM previously
    - Now have similar functionality with perl-based script that uses chimera-dump.py
      - » Can specify policies like:
        - » /store/user gets 2 copies of each file, on resilient nodes
        - » /store/unmerged gets 2 copies of each file
        - » /store/generator (Pileup) gets 5 copies of each file
      - » Watch dashboard for popular datasets, and replicate more copies while it's in demand.

Other Efforts

- Other storage projects at Purdue
  - Have two Hadoop clusters
    - Used mostly by Map-Reducing groups.
  - Lustre
    - Operate large Lustre filesystems in production today
  - Distributed Filesystem Testing
    - Ceph, MooseFS
  - Distributed xrootd service
    - Wrote xrootd -> dcap plugin to tie dCache to service