Experience with Xrootd/PROOF farm at BNL: 2010

Sergey Panitkin

BNL

BROOKHAVEN
NATIONAL LABORATORY

Atlas Xrootd/PROOF farm at BNL: Overview

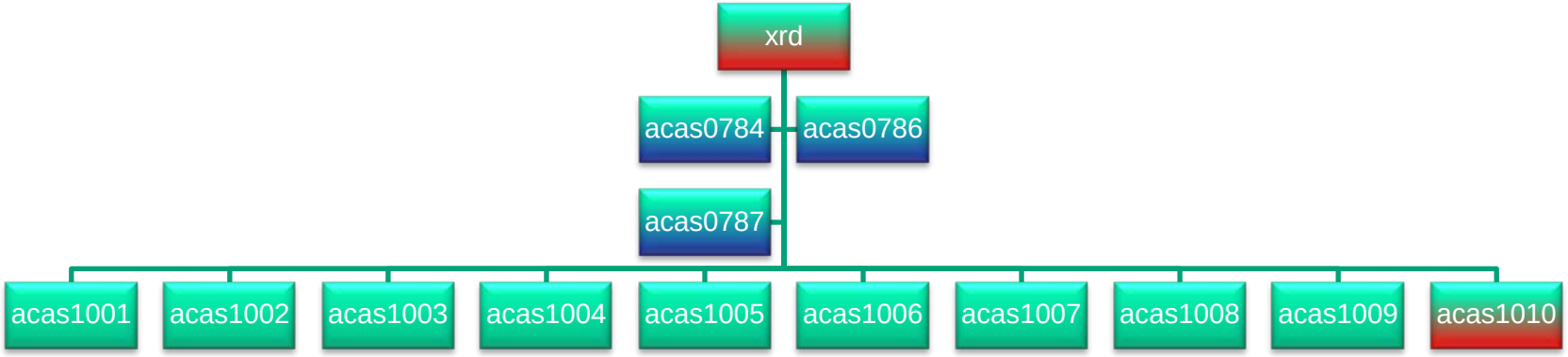
- A new Xrootd/PROOF farm was set up at BNL in summer 2010
- Part of ACF@BNL. Inside ACF T1 security perimeter.
- Accessible from ACF's interactive and batch nodes only
- Tunnel through the ACF gateway should work too
- Redirector node: xrd01.usatlas.bnl.gov (xrd is suggested alias)
- ~20 TB of disk storage distributed over 10 server nodes, 4x500GB per node, in RAID 0 array.
- ~36 TB of disk storage on 3 data vaults - 6x2TB each
- Ganglia monitoring pages for farm nodes exist here.
- Wiki page created and can be found here

ATLAS Xrootd/PROOF farm at BNL: Introduction

- In March 2010 it became clear that physicists run their codes with PROOF-Lite on interactive nodes at BNL
- After testing switch from virtual interactive machines to non-virtualized nodes for PROOF-Lite work
- It also became clear that group will need more disk space
  - BlueArc NFS server was good but too expensive
  - Decided to go for Xrootd with commodity hardware
- Started with hardware testing for Xrootd nodes
  - SAS, SATA, SSD, hardware vs software RAID, etc
- Started operations with two data vaults - machines from testing samples
- Needed to give them back at some point (File renaming!)
- Received 10 more server nodes and a dedicated redirector node
  - Copied data from temp data vaults
- Received 3 data vaults

Xrootd/PROOF Farm at BNL: Organization

- We started with Xrootd redirector and PROOF master running on the same node – xrd.usatlas.bnl.gov
- xrd machine proved to be inadequate as both redirector and master.
  - Small disks did not allow to buffer data transfers
  - Small memory created problems in PROOF histogram merging
  - Changes in /etc/rc.d/init.d/xproofd (defaults: “ulimit -m 1048576 -v 2097152 -n 65000”)
- Moved data transfers responsibilities to acas0787 (that’s where DM scripts run)
  - It has larger disks, works better for data buffering
- PROOF master runs on acas1010 now
  - Larger memory, faster CPUs



ATLAS Xrootd/PROOF farm at BNL: Details

- Xrootd master node xrd01.usatlas.bnl.gov (DNS alias xrd)
  - Old machine close to retirement
  - QuadCore Xeon @3.2GHz, 2 GB RAM, ~400 GB disk
- 10 Server nodes, acas100[1-10] each has:
  - 8 Core Xeon 5560 @ 2.8 GHz
  - 24 GB RAM
  - 1 Gb NIC
  - **4x500 GB** SATA -7200rpm disks in RAID0 (one partition /data)
- 3 data vaults - acas0784, acas0786, acas0787 each has:
  - Dell R710, 8 core Xeon 5560 @ 2.8 GHX GHz
  - 24GB RAM
  - 1Gb NIC
  - **6x2 TB** disks in RAID0 (two partitions: /data0 /data1)

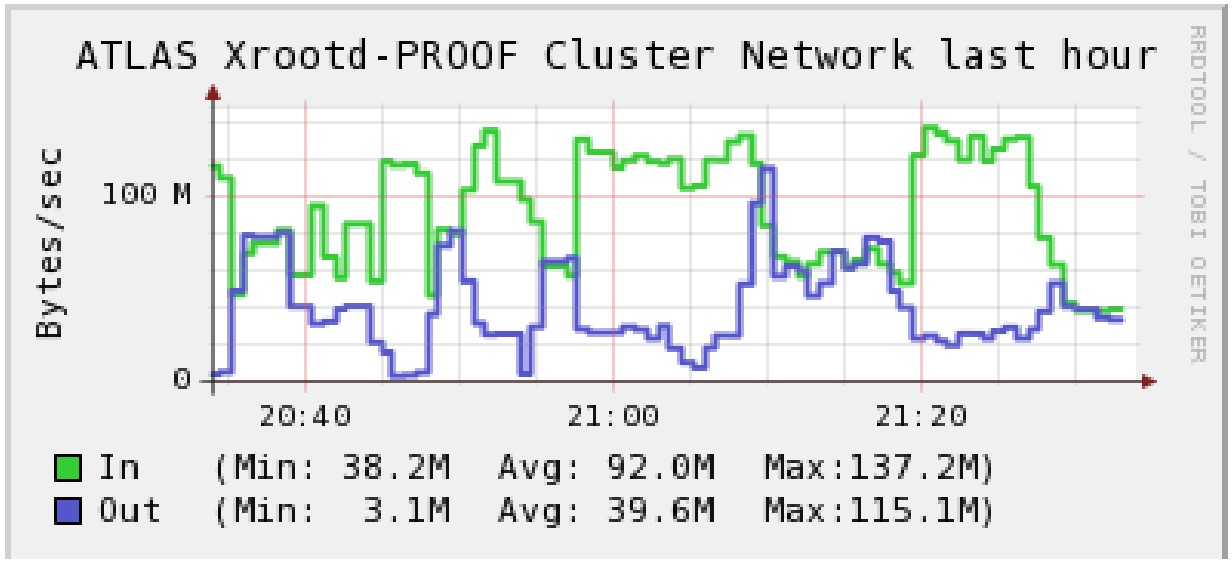
Data Management

- Name space convention: root://xrd//data/datasetname/filename
- The whole DM scheme is dataset centric – similar to ATLAS DDM
- That affects file system layout as well as bookkeeping
- At copy time we generate lists of files in a dataset with proper names for xrootd
- This files lists can be found at: ~xrdadmin/xrd_copied dataset/datasetname.clist
- We also tested and will be using PQ2 tools from PROOF for dataset registration
  - Waiting for root 5.28 with new PQ2 functionality
- If you need to copy your data to Xrootd - contact Hong Ma or me (panitkin@bnl.gov)

Data Copy algorithm

- If container – get list of datasets for the container
- Loop over datasets
- Check if the dataset is already on Xrootd
  - If “yes” skip to next dataset
- Get information about dataset files with dq2-ls
  - Filter on root files, skip logs, etc
  - Number of files, size, etc
- dq2-get brings data to a local buffer
- Compare what we got with what is in dq2 catalog
- xrdcp copies data to Xrootd farm
- Create bookkeeping records
  - Plain ASCII files with lists of files (most popular)
  - Register files in PQ2
  - If any step fails record what failed
- Delete local buffer
- Get next dataset

Data Transfer in action



- We can pull in data at 100+MB/s using multi-stream and proximity features of dq2-get

Issues with DM

- Main drawback of the current DM scheme is that its not fully automatic and require an “operator” to initiate and supervise data transfer
- Users seem to like dq2 subscription type of mechanism
  - They don’t like delays
- Install SRM and become a grid site?
- Modify current scheme to be more operator independent?

PROOF

- Farm runs PROOF in conjunction with Xrootd
- PROOF farm master is acas1001.usatlas.bnl.gov
  - To start Proof session: TProof *p = TProof::Open("acas1010")
  - Farm runs root v.5.27.04, will be switching to v.5.27.06, waiting for v5.28
    - /afs/usatlas/sw/lcg/external/root/5.27.04/x86_64-slc5-gcc43-opt/root
  - Your Proof-Lite code should run right away (not always, but hopefully true!)
  - Try to run with 86 workers, 192 workers max.
- Currently we have about 10 users, running a whole spectrum of analyzes
- You are welcome to run on the farm.
- If you have an account at ACF you are good to go!

The End

- Questions?