Xrootd test farm
at Atlas Computing Facility at BNL.
“To Do” List and Status

Sergey Panitkin

BNL

August 1, 2007

ATLAS

BROOKHAVEN
NATIONAL LABORATORY

Introduction

- This talk is intended as an overview of the work in progress
- It is pretty much my “to do” list from the Wiki page, with a few additional slides.
- Things that were done or progressed significantly are labeled in green.

Hardware and facilities

- Reconfiguration of the network connections to eliminate or reduce current bandwidth limitations.
  - Ticket for network reconfiguration submitted by Robert on July 21st. Outcome unknown. :-(
  - Scalability tests with reconfigured network (Tadashi, Sergey)
- Set up redundant Xrootd redirectors, PROOF master nodes (Ofer?)
- Estimate how much "bytes out" bandwidth is need for the farm for "typical" analysis scenarios.
  - See next two slides.
- Needed input:
  - Number of nodes used for typical Condor submission
  - Jobs per node
  - Typical job input rate (Athena/AODs ~2 GB/s ?)
- What is the final size of the farm? (Michael)
  - See next slides.

Disk I/O tests.

- Current single server disk subsystem properties:
  - 4x500GB SATA disks in RAID 0
  - Single disk properties:
    - Model ST3500630NS, Seagate Barracuda SE family
    - Perpendicular recording
    - 7200 rpm
    - 8.5 ms seek time
    - 16GB cache
    - SATA-3GB interface
    - NCQ capable
    - Maximum sustained transfer rate of 72 MB/s is claimed by the Seagate.
    - Independent tests reported from 44 to 78 Mb/s transfer rate
- Robert Petkus performed *iozone* disk tests on one of the nodes
  - “Torture test”: 4 threads simultaneously reading 4 separate files (8GB each)
  - Measured average transfer rate from the RAID array ~74 MB/s
  - Looks like a reasonable match for 1 Gb network bandwidth per node!

Back of the envelope estimate of the required farm size.

- Somewhat idealized scenario
- ACF batch can run up to ~1500 simultaneous jobs
  - admittedly extreme case
- Assuming that typical Atlas Athena/root job consumes ~3MB/s
  - In tests with tags we've seen ~2 MB/s for loose cuts case and ~4 MB/s for strict cuts
- Total theoretical bandwidth needed is: 1500x 3Mb/s= ~4.5 GB/s
- Assuming that each node can provide ~70 MB/s
- That gives : 4.5 GB/s / 70 Mb/s/node= ~ 64 server nodes
- Perhaps 64 nodes total, with 2-3 redundant xrootd redirectors
- Curiously enough 64 is a “magic” number for Xrootd architecture (scales as 64 B-Tree)
- We will also need a test “mini farm” for new versions tests and validation and other related software development ~3-5 nodes

System Management and Security

- General Farm Management
  - Scripts for full farm xrootd/PROOF: "Start up/Shutdown/Restart". (done, Ofer)
  - Scrips for data movement to from Xrootd farm to NFS and dCache. (done? Sergey)
  - Local farm “file catalog”
    - Web page? Web interface?
  - Scripts to delete datasets or selected files
- Security
  - Test ssh and certificate authentication. Select most appropriate method (convenience, security, etc)
  - PROOF client “conduit”
  - Off-site users issues? Is login to acas necessary?
  - Firewall problems
  - Kerberos authentication ? (Ofer, Edgar)

Farm usage model

- Two aspects: Xrootd and PROOF
  - Different use cases, different load types
  - Different security issues
- How to utilize the farm
  - Open to all, open to selected people, open to group representatives, etc....
  - What is needed? What is sustainable? What is manageable?
  - What is experience in other experiments (BaBar, Alice, CMS, etc) ?
  - Xrootd for all, PROOF for a few?

Monitoring and Documentation

- Farm Monitoring.
  - Ganglia pages. (done. Jason)
  - Adapt SLAC monitoring package?(Edgar, Ofer) Done.
- Web page/TWIKI with general farm information, help, examples, tips, talks, links to Ganglia page, etc. (Robert?).
  - Done (Kyle)
- Think of appropriate mail list/hypernews. Have a separate one?
-

Integration with Atlas DDM

- DDM
  - Integration with current (US?) Atlas distributed analysis model. Panda/pAthena/DQ2 (tadashi?) In , sprogress, see Tadashi's talk
  - data movement scenarios
    - PANDA to xrootd directly with registering in DQ2 ?
    - Two stages: PANDA to dCache, then to Xrootd ?
  - Interaction with dCache at BNL
  - Tests with Xrootd “door” on dCache
  - Who will be in charge of moving/removing datasets
- Interaction with remote Xrootd servers
  - Proof of the principle

Software and Analyses Tests

- Analysis Software Tests
  - PROOF analysis with TSelector, Analysis with libraries in PAR format. (Kyle?). Tselector done, PAR format I'm not sure.
  - Examples for Tutorials. Done (Kyle, Tadashi)
  - More analyses from different physics groups (Kyle?). Kevin Black
  - Heavy Ion program
    - datasets – currently mostly ntuples on dCache
    - Try optimization of dCache read ahead, etc. Different size here!
      - Done. Factor of 5 improvement!
    - Mostly root based analyses, interactive and batch, try with PROOF (Mark Baker, Sergey?)
  - Try Scott's “AODs in root” with PROOF (Scott ?)
  - Analyses with several "post-AOD" root based formats(DPD, etc) using PROOF (Diego, Kevin)
- Xrootd scalability tests with reconfigured network (Tadashi, Sergey?)

Software and Analyses Tests

- Proof scalability tests
  - How many slaves per box (3, 4, 5…?) with given hardware?(Sergey)
    - Note that farm will be heterogeneous
    - In root 5.14 observed PROF crashes with 4 slaves per node. The reason is unknown. Crashes with 3 slaves per node.
    - Will resume testing with root 5.16
  - How many simultaneous users can work on the farm? (Kyle?)
  - How robust is PROOF system anyways?
    - Gain experience with long jobs, slave crashes, crash recovery, etc
    - Observed some PROOF instability. Introduced daemon auto-restart scripts (Ofer).
    - 27 daemon crashes over 2 days

T3 aspects

- T3 Aspect of Xrootd/PROOF
- Shall we think of and provide recommendations for:
    - Use cases
    - Usage examples
    - Management tools
    - Security issues and recipes
    - Hardware/Networking recommendations
        - Wiki page is up. Examples, tips, etc
        - Performed disk transfer rate tests.
        - Monitoring framework is running
        -
- There is already Atlas T3 task force

Xrootd/PROOF developments

- Gerri Ganis set up at CERN an afs area for PROOF development snapshots
    - PROOF development is more rapid than root releases
    - Root/xrootd/PROOF versions synchronized and validated
        - Important from our experience!
    - Targeted for Alice, Atlas (us+Wisconsin?) users
    - /afs/cern.ch/sw/lcg/contrib/proof/root
    - Latest version based on root 5.16
    - Source code and a few binary distributions
    - slc4_ia32_gcc344 distro is installed in BNL root directory (Thanks to Alex Undrus!) and on the test farm yesterday.
    - Please test.

New in ROOT 5.16

- Added support for processing datasets 'by name', i.e. by just sending the name of a dataset known by the PROOF master node:
  - root[] proof->Process("dc2006", "TrackAnalysis.C")
- Added first infrastructure to support for scheduling and quota controls:
  - possibility to define groups of users
  - simple dynamic scheduler assigning nodes to the sessions based on the cluster load;
  - worker-level priority-based throttling of sessions
- Reduced memory footprint of proofserv at startup by a factor 2
- Added possibility to load a macro or a class (TProof::Load(...))
  - root [6] p = TProof::Open("ganis@lxb6043")
  - root [7] p->Load("h1analysis.C+")
- Added full support for authentication for the xrootd-based communication layer with automatic credential forwarding to worker nodes; supported protocols: password-based, Kerberos, GSI.
- Added support for the acquisition of AFS tokens within the proofserv sessions; this allows to access private AFS areas from any worker or master node.
- Several improvements in the new packetizer - TAdaptivePacketizer - which is now the default packetizer.
- Several bug fixes and internal optimizations
- More in v5.16 release notes!