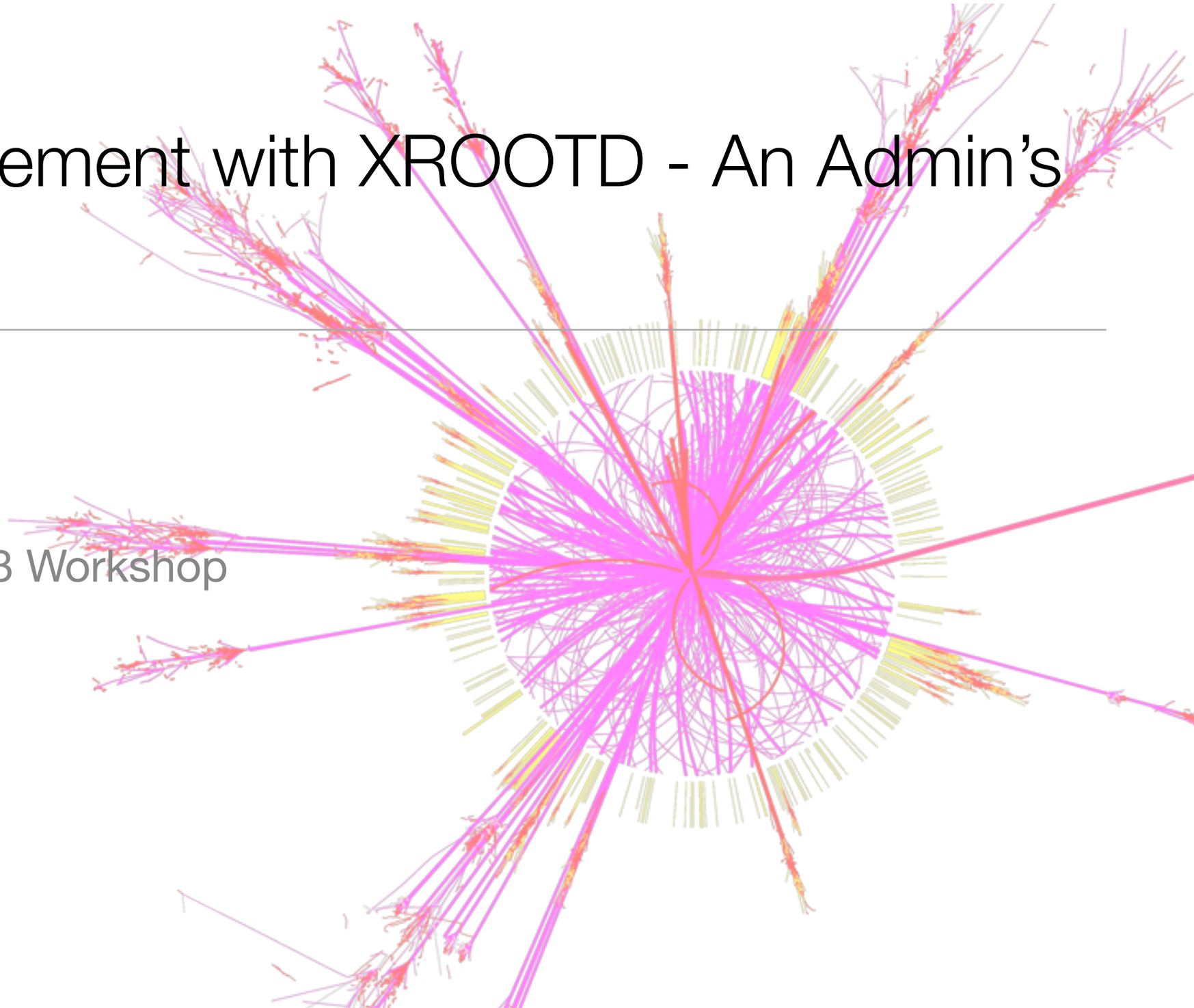
Storage Management with XROOTD - An Admin's Perspective...

Ofer Rind
BNL RACF

US ATLAS Tier 2 & Tier 3 Workshop
SLAC
November 29th, 2007



Outline

- Quick intro to xrootd and interest for ATLAS
- Installation procedures and configuration
- BNL setup
- Management
- Monitoring
- Future & Usage Model

What is Scalla?

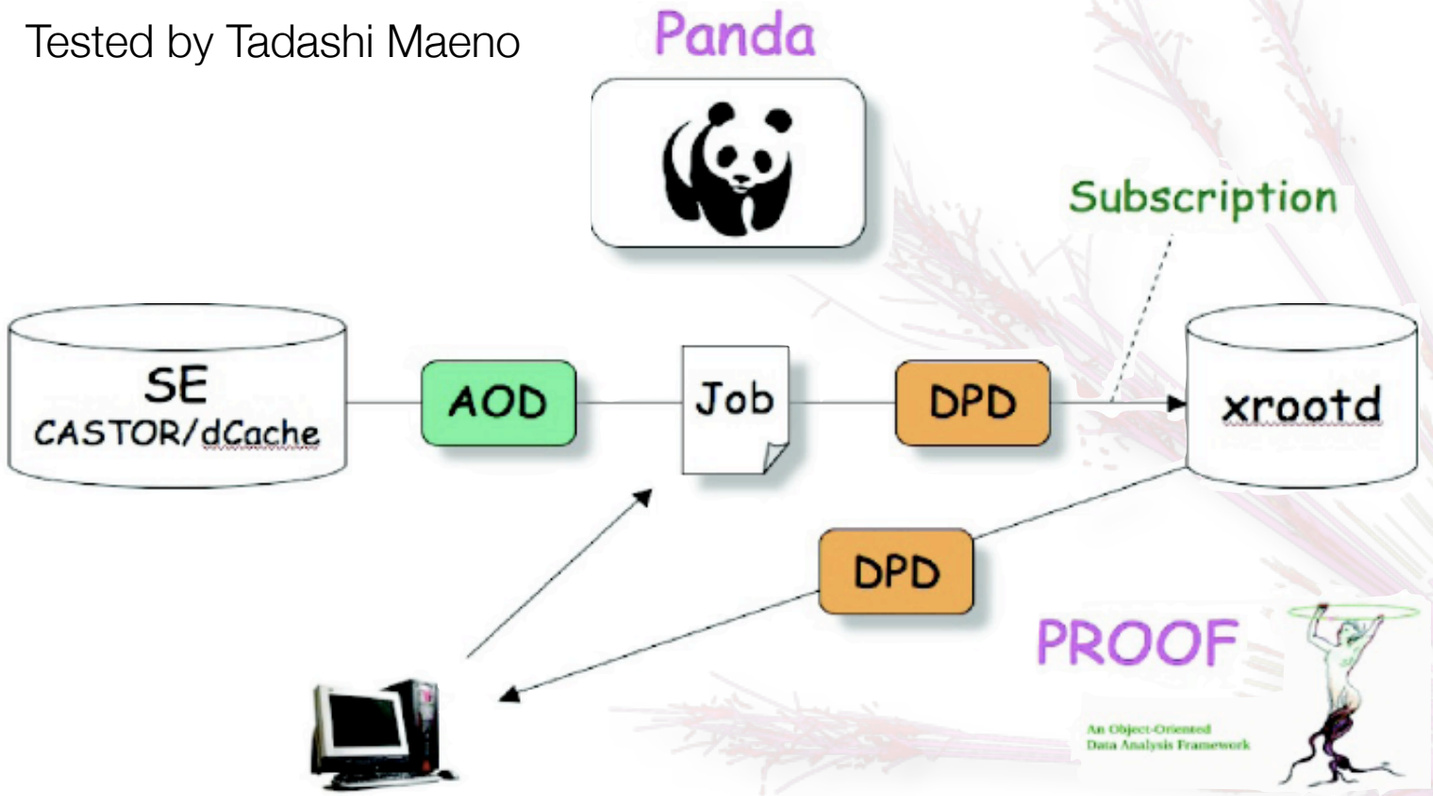
- Scalla (**S**tructured **C**luster **A**rchitecture for **L**ow **L**atency **A**ccess)=xrootd+olbd
- xrootd: file server with low latency access
  - ROOT files or other, including POSIX support
- olbd: clusters xrootd servers into a common name space
  - Scalable and self-organizing
  - Manager:
    - Keeps track of file paths and redirects client to file
    - Configurable server selection
  - Server:
    - Track xrootd health and utilization
    - Report statistics to manager

Scalla Features

- Performance enhancements in rootd protocol
- Dynamically built namespace (just-in-time query) – no persistent state
- High efficiency, low footprint
  - Suitable for heterogeneous commodity and legacy hardware
- No 3rd party software dependencies, plugin architecture (e.g. auth, monitoring, generic MSS interface)
- Fault tolerant (self-organizing, no DB requirements)
- ”Low” administration cost, at least at a basic level
- Easy interface to dCache via xrootd door (xrdcp)
- Ability to federate inexpensive hardware locally and cross-site (tie-in with PROOF)

US ATLAS Interest in xrootd?

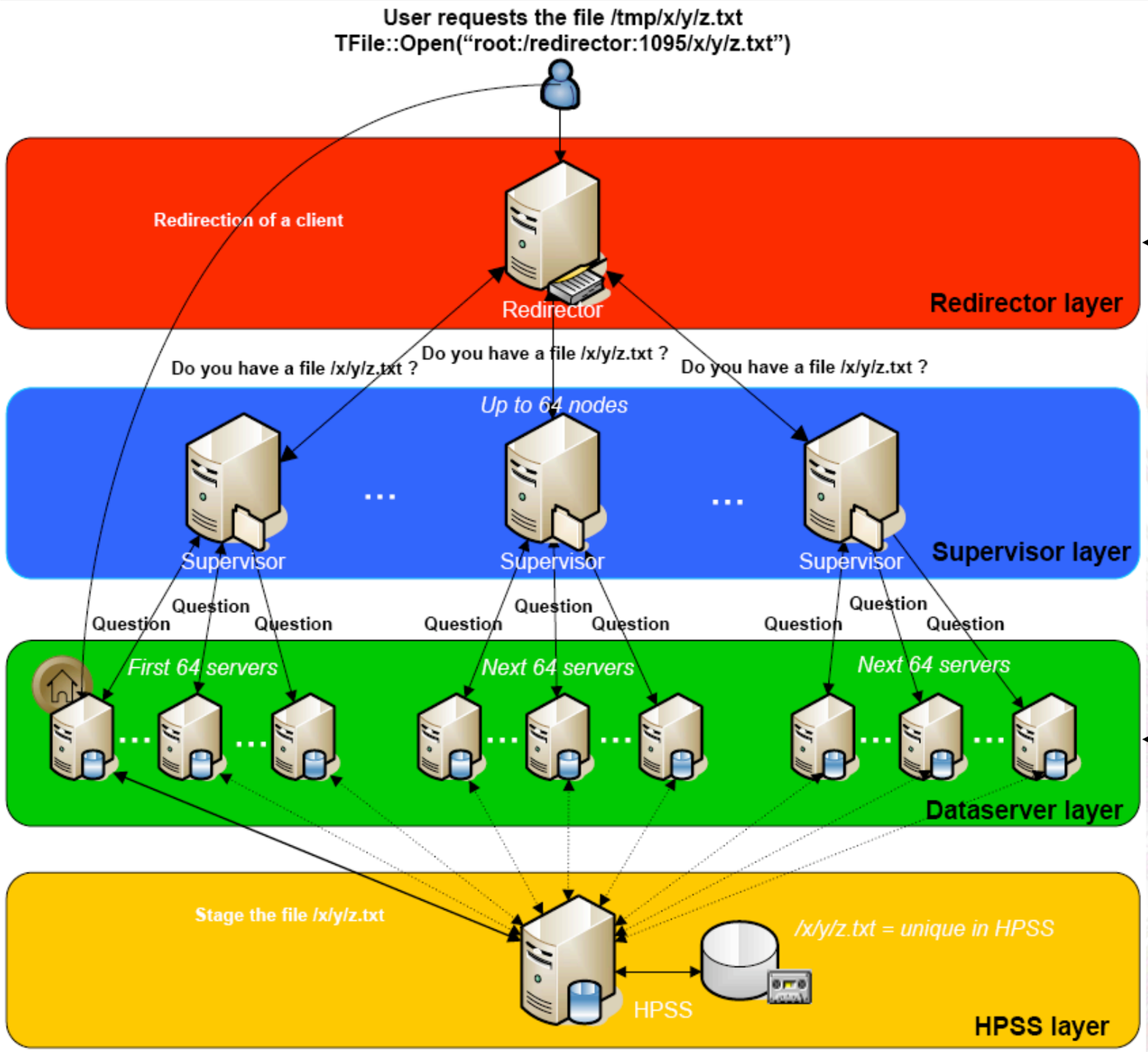
Tested by Tadashi Maeno



- Currently employ a two-stage analysis model:
  1) (p)athena for ESD & AOD datasets
  2) ROOT-based analysis of DPD
- Some perceived issues with dCache throughput under certain access models
- Suitability for smaller sites (Tier-3)
- Xrootd integration with PROOF

We can also use dCache as an endpoint storage element in Panda and then pull DPDs from dCache via xrootd door

The Scalla Architecture

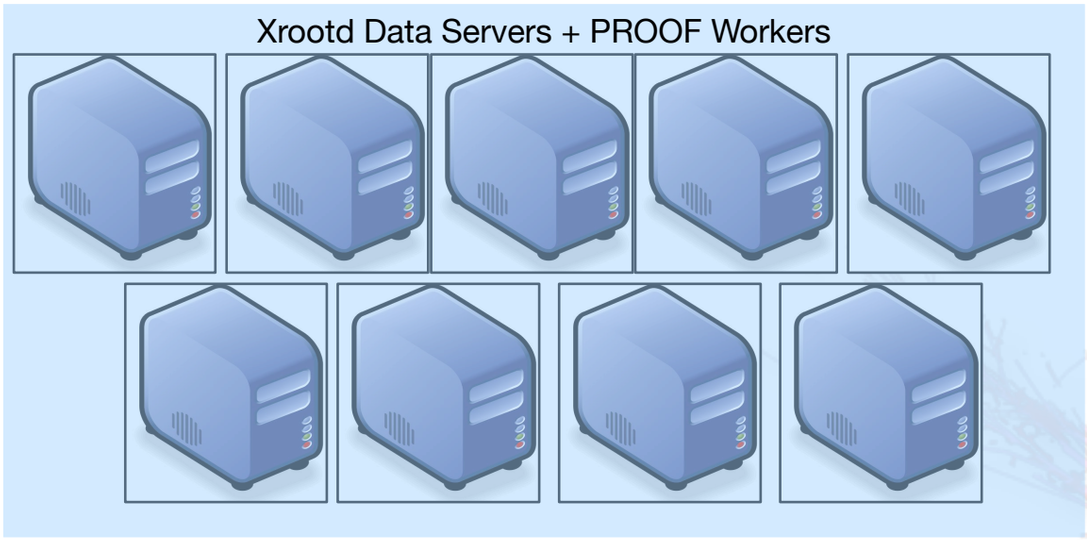


Currently implemented at BNL USATLAS Testbed

STAR Architecture - See e.g. P. Jakl ROOT Workshop 2007

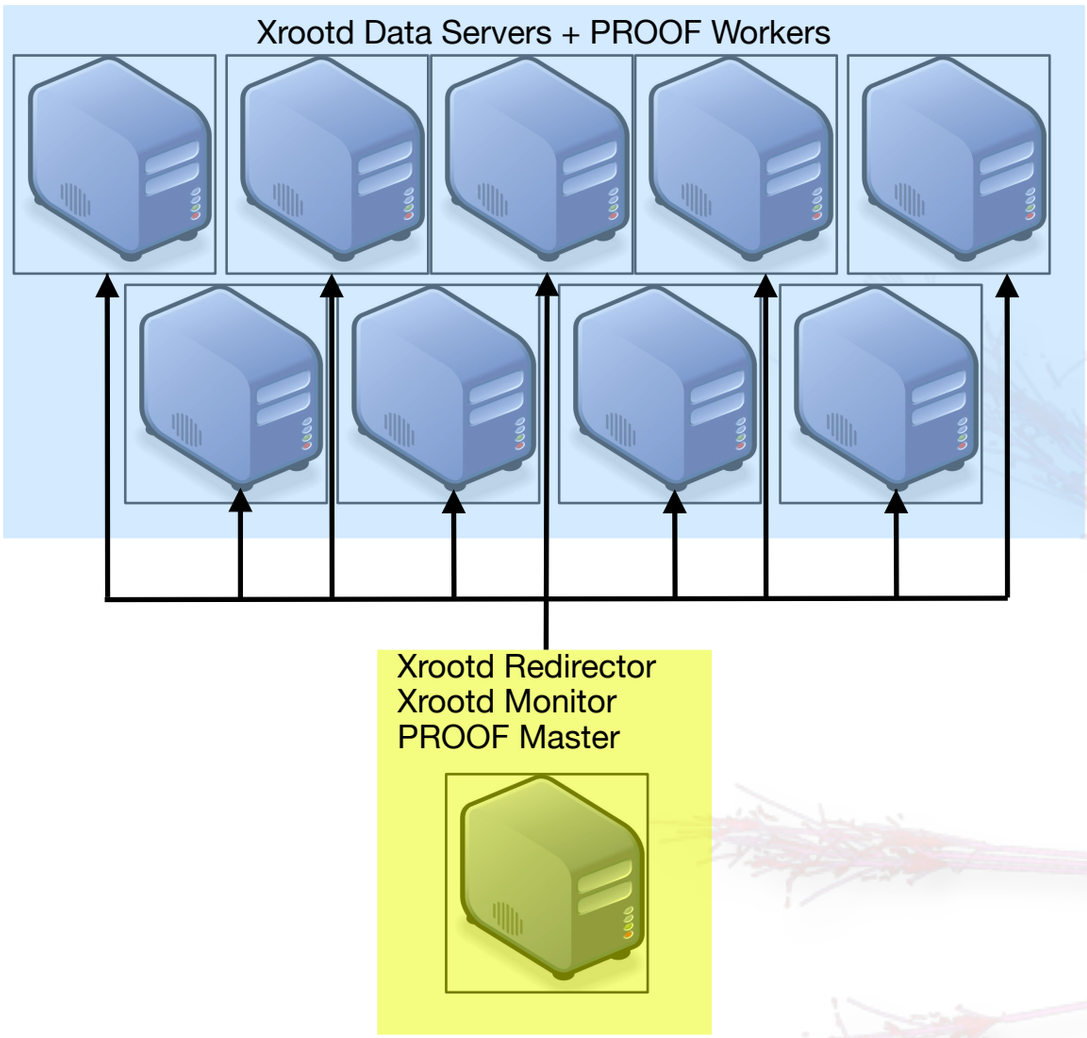
BNL ATLAS Testbed Configuration

BNL ATLAS Testbed Configuration



(9) Data Servers + PROOF Workers each with:
- (2) dual-core 1.8 GHz Opteron 265 processors
- (4) 500 GB SCSI disks (1.8TB) configured RAID0
- Scientific Linux 4.4
- xrd v.20071001-0000a

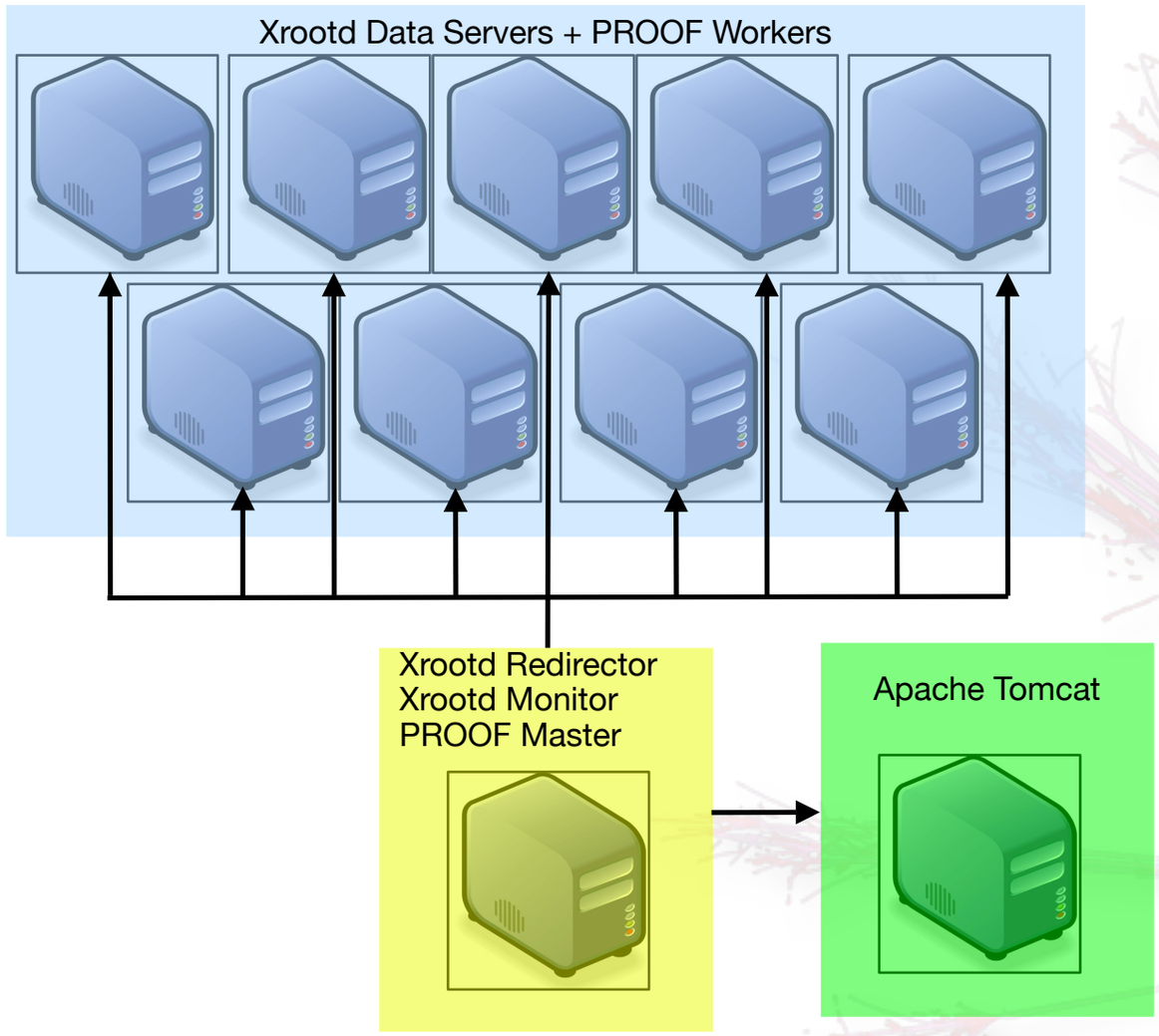
BNL ATLAS Testbed Configuration



(9) Data Servers + PROOF Workers each with:
- (2) dual-core 1.8 GHz Opteron 265 processors
- (4) 500 GB SCSI disks (1.8TB) configured RAID0
- Scientific Linux 4.4
- xrd v.20071001-0000a

(1) Redirector + PROOF Master + Xrootd monitor (Perl, MySQL) configured as above

BNL ATLAS Testbed Configuration

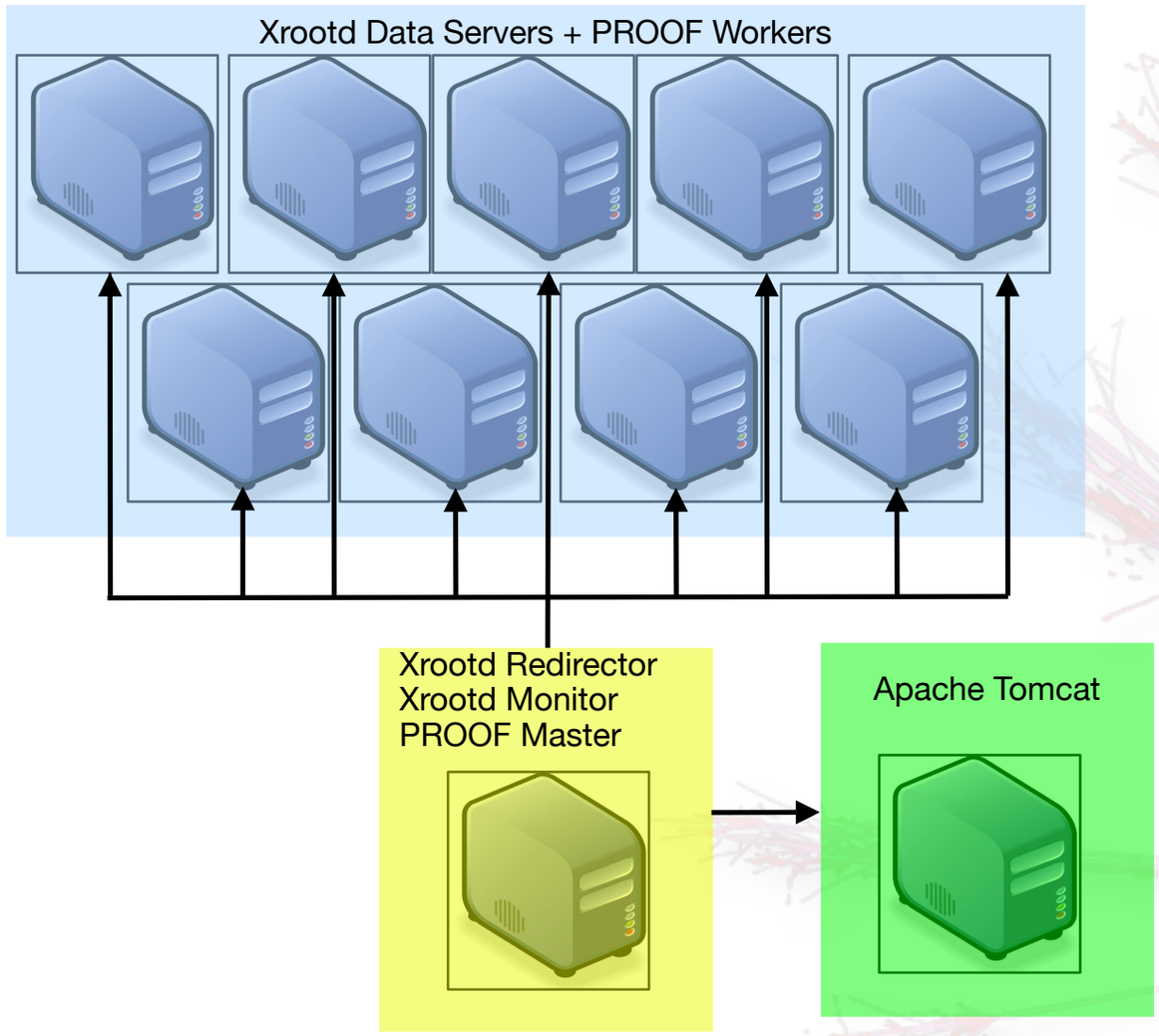


(9) Data Servers + PROOF Workers each with:
- (2) dual-core 1.8 GHz Opteron 265 processors
- (4) 500 GB SCSI disks (1.8TB) configured RAID0
- Scientific Linux 4.4
- xrd v.20071001-0000a

(1) Redirector + PROOF Master + Xrootd monitor (Perl, MySQL) configured as above

(1) Apache Tomcat server for monitoring display

BNL ATLAS Testbed Configuration



(9) Data Servers + PROOF Workers each with:
- (2) dual-core 1.8 GHz Opteron 265 processors
- (4) 500 GB SCSI disks (1.8TB) configured RAID0
- Scientific Linux 4.4
- xrd v.20071001-0000a

(1) Redirector + PROOF Master + Xrootd monitor (Perl, MySQL) configured as above

(1) Apache Tomcat server for monitoring display

- Non-privileged “xrdadmin” user (run-mode, file ownership)
- No authentication or authorization, but has been tested with K5 plugin.
- On-site PROOF access through RACF firewall + offsite through gateway

Installation & Startup

- Standard prod and dev versions available from the Scalla home page
  - Simple installation and configuration for basic deployment
- Also packaged into a special ROOT branch by Gerri Ganis at CERN
  - Needed to ensure xrootd-Proof interoperability since Proof is under active development
    - Current at BNL: /afs/cern.ch/sw/lcg/contrib/proof/root/5.17.05-PROOF.00/
  - Source code on subversion server https://root.cern.ch/svn/root/trunk
- Startup
  - Need to start both xrootd and olbd (if clustering)
  - Standard scripts provided with Scalla package - need some configuration
  - Standard init.d scripts provided with ROOT branch - need to start as root with effective user in order to run Proof
- Configuration Files: mainly xrootd.cf

Starting Scalla

Standard scripts to start Scalla from init.d (must be started this way by root in order to use PROOF)

**And similar for olbd*

```
# chkconfig: 345 20 80
# description: The xrootd daemon is used to as file server and starter of
#              the PROOF worker processes.
~~~~
XROOTD=/opt/xrootd/bin/arch/xrootd
XRDLIBS=/opt/xrootd/arch/lib
~~~~
# Get xrootd config
[ -f /etc/sysconfig/xrootd ] && . /etc/sysconfig/xrootd
# Read user config
[ ! -z "$XRDUSERCONFIG" ] && [ -f "$XRDUSERCONFIG" ] && . $XRDUSERCONFIG
~~~~
start() {
        echo -n $"Starting $prog: "
        # Options are specified in /etc/sysconfig/xrootd .
        # See $ROOTSYS/etc/daemons/xrootd.sysconfig for an example.
        # $XRDUSER *must* be the name of an existing non-privileged user.
        export LD_LIBRARY_PATH=$XRDLIBS:$LD_LIBRARY_PATH
        daemon $XROOTD -b -l $XRDLOG -R $XRDUSER -c $XRDCF $XRDDEBUG
```

Define xrootd parameters used below

Set ROOTSYS and LD_LIBRARY_PATH

xrootd.cf config file (some basics from BNL system)

xrootd.cf config file (some basics from BNL system)

```
xrd.port 1094
olb.port 3121
```

xrootd.cf config file (some basics from BNL system)

```
xrd.port 1094
olb.port 3121

xrootd.fslib /afs/usatlas.bnl.gov/cernsw/lcg/external/root/5.17.05-PROOF.00/slc4_ia32_gcc34/root/lib/libXrdOfs.so
```

xrootd.cf config file (some basics from BNL system)

```
xrd.port 1094
olb.port 3121

xrootd.fslib /afs/usatlas.bnl.gov/cernsw/lcg/external/root/5.17.05-PROOF.00/slc4_ia32_gcc34/root/lib/libXrdOfs.so

xrootd.export /data                    path prefix
```

xrootd.cf config file (some basics from BNL system)

```
xrd.port 1094
olb.port 3121

xrootd.fslib /afs/usatlas.bnl.gov/cernsw/lcg/external/root/5.17.05-PROOF.00/slc4_ia32_gcc34/root/lib/libXrdOfs.so

xrootd.export /data                    path prefix

olb.path w /data                       paths handled by server
```

xrootd.cf config file (some basics from BNL system)

```
xrd.port 1094
olb.port 3121

xrootd.fslib /afs/usatlas.bnl.gov/cernsw/lcg/external/root/5.17.05-PROOF.00/slc4_ia32_gcc34/root/lib/libXrdOfs.so

xrootd.export /data                 path prefix

olb.path w /data                    paths handled by server

oss.cache public /data/cache*       define disk cache partition
```

xrootd.cf config file (some basics from BNL system)

```
xrd.port 1094
olb.port 3121

xrootd.fslib /afs/usatlas.bnl.gov/cernsw/lcg/external/root/5.17.05-PROOF.00/slc4_ia32_gcc34/root/lib/libXrdOfs.so

xrootd.export /data                 path prefix

olb.path w /data                    paths handled by server

oss.cache public /data/cache*       define disk cache partition

if acas0420.usatlas.bnl.gov         redirection for clustered servers
all.role manager
else
all.role server
fi
```

xrootd.cf config file (some basics from BNL system)

```
xrd.port 1094
olb.port 3121

xrootd.fslib /afs/usatlas.bnl.gov/cernsw/lcg/external/root/5.17.05-PROOF.00/slc4_ia32_gcc34/root/lib/libXrdOfs.so

xrootd.export /data                 path prefix

olb.path w /data                    paths handled by server

oss.cache public /data/cache*       define disk cache partition

if acas0420.usatlas.bnl.gov         redirection for clustered servers
all.role manager
else
all.role server
fi

if acas0420.usatlas.bnl.gov         metadata command forwarding
ofs.forward all
else
ofs.redirect target
fi
```

xrootd.cf config file (some basics from BNL system)

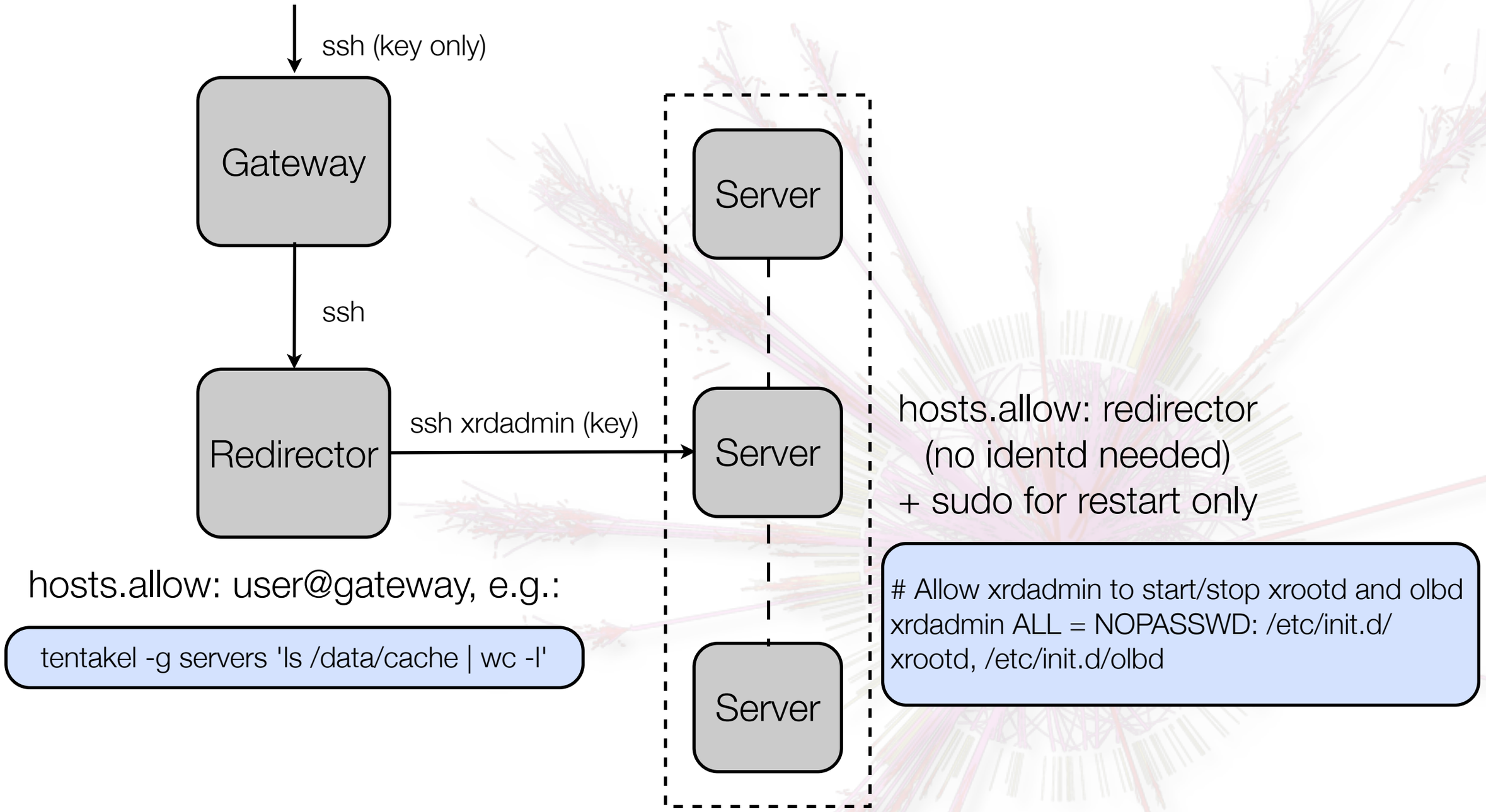
```
xrd.port 1094
olb.port 3121

xrootd.fslib /afs/usatlas.bnl.gov/cernsw/lcg/external/root/5.17.05-PROOF.00/slc4_ia32_gcc34/root/lib/libXrdOfs.so

xrootd.export /data                 path prefix

olb.path w /data                    paths handled by server

oss.cache public /data/cache*       define disk cache partition

if acas0420.usatlas.bnl.gov         redirection for clustered servers
all.role manager
else
all.role server
fi

if acas0420.usatlas.bnl.gov         metadata command forwarding
ofs.forward all
else
ofs.redirect target
fi
```

**Plus a bunch more related to PROOF!*

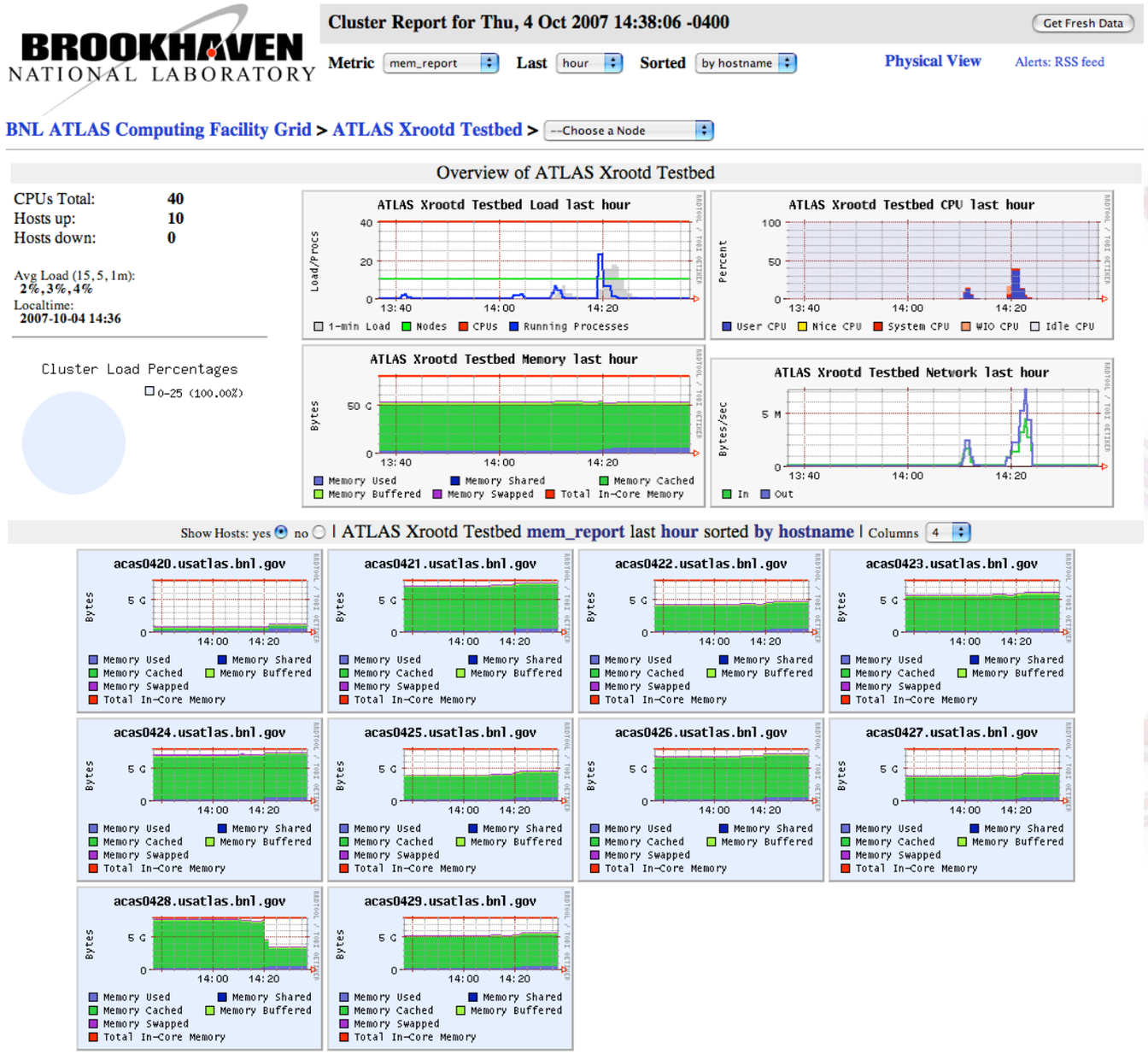
Facility Operation & Management

- Significant experience has been gained
  - Tested an array of Atlas scenarios
    - Integrated with DDM, Panda, dCache,...
  - Administrative management framework established
    - 2 people (Rind, Panitkin) involved in installation, configuration, maintenance
    - Tentakel (using sudo), Daemon life-support
  - Comprehensive Monitoring Setup
    - Ganglia, Nagios + SLAC's native monitoring framework
- Close partnership between facility and users has been invaluable
  - User support and documentation: PROOF Testbed Twiki

User Management Setup



Monitoring - Ganglia

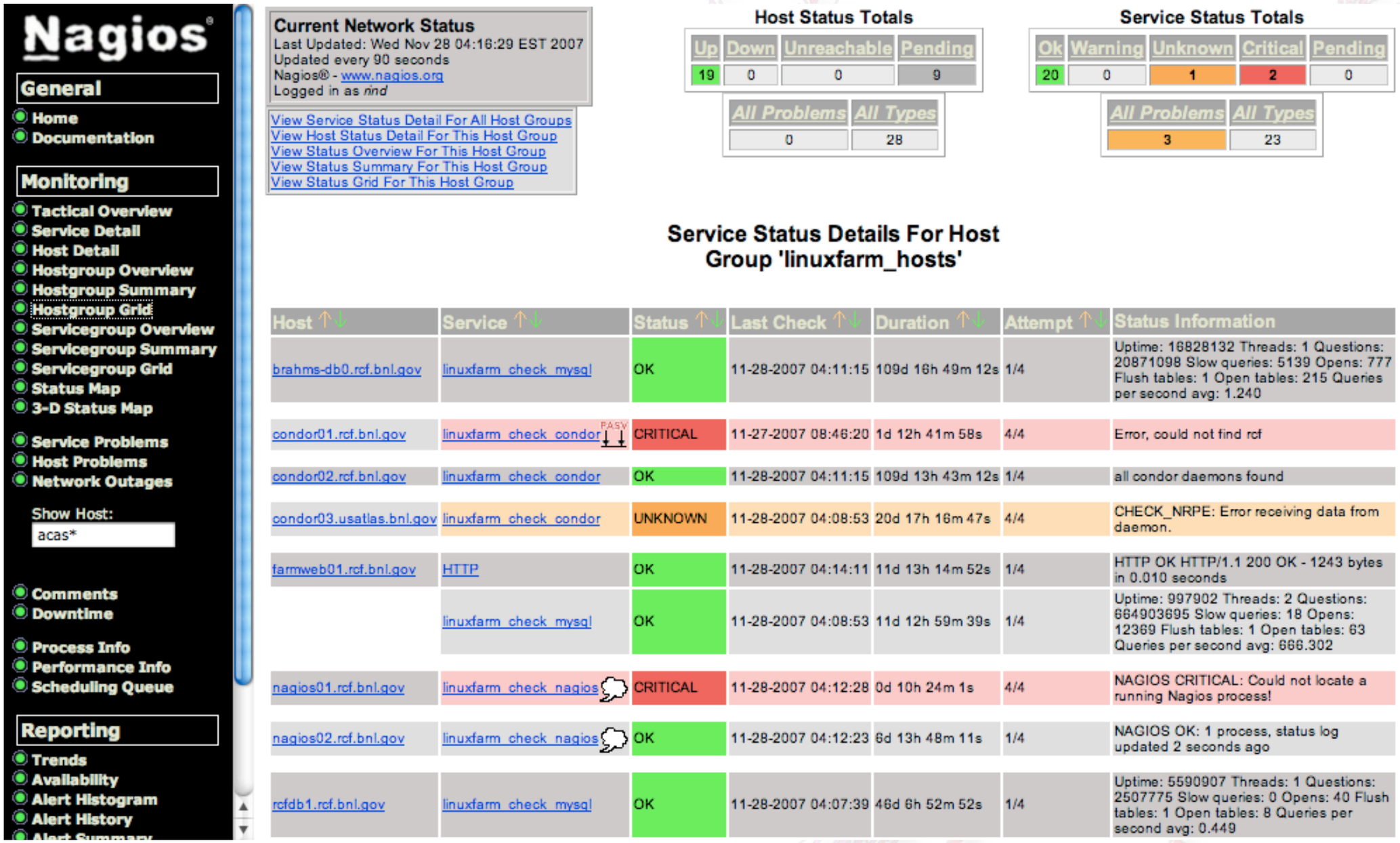


- Farm-wide overview at hardware/ OS level (CPU, memory, disk, network activity, etc.)
- Monitoring, diagnostics, historical accounting
- Customizable via gmetric, e.g.:

```
GMETRIC="/usr/bin/gmetric --mcast_ttl=3 -c$MCHAN -d3000"
for proc in xrootd olbd; do
 stats=( `/bin/ps --no-headers -o rss,vsz,%cpu -C $proc` )
 n=3
 while [ ${#stats[*]} -gt $n ] ; do
  let m=n%3
  stats[$m]=`/bin/echo ${stats[$m]} ${stats[$n]}|/bin/awk '{print $1+$2}'`
  let n+=1
 done
 if [ ${stats} ] || eval "stats=( 0 0 0 )"; then
  $GMETRIC -nmem_${proc} -tuint32 -uKB --value ${stats[0]}
  $GMETRIC -nvsz_${proc} -tuint32 -uKB --value ${stats[1]}
  $GMETRIC -ncpu_${proc} -tfloat  -u%  --value ${stats[2]}
 fi
done
```

Monitoring - NAGIOS

Monitor host status, services + send alerts, open RT tickets (see T. Wlodek talk for details)



Monitoring - XrdMon

Xrootd data access monitoring

Thanks Jacek and Tofigh!

Basic view

Top performers

List all
- users
- dataTypes
- files
- servers
- clients
- jobs

List current
- users
- dataTypes
- files
- servers
- clients
- jobs

Query by
- user
- dataType
- file
- server
- client
- job

Common queries

Xrootd statistics

Table rows: 10 **Time Period:** Last Month **Site:** usatlas Update

Top active users

User Name	Now			Last Month			
	Number of Jobs	Number of Files	File Size [MB]	Number of Jobs	Number of Files	File Size [MB]	MB Read
xrdadmin	0	0	0	5,899	5,834	252,832	0
casadei	0	0	0	489	2,391	131,651	37,130
serp	22	90	0	271	1,685	44	2,555
tarrade	0	0	0	204	223	0	4,524
dladams	0	0	0	28	3	0	3,007
akira	0	0	0	2	1	61	16

Hottest dataTypes

dataType Name	Now				Last Month				
	Number of Jobs	Number of Files	File Size [MB]	Number of Users	Number of Jobs	Number of Files	File Size [MB]	Number of Users	MB Read
HPTV	0	0	0	0	5,020	4,587	256,589	4	37,169
MUON_1	0	0	0	0	997	969	0	2	3,007
HiggsToTauTau-00-00-44	22	90	0	1	836	339	0	3	7,056
serp	0	0	0	0	2	1,519	0	1	0

Hottest files

File Path	File Size [MB]	Now	Last Month	
		Number of Jobs	Number of Jobs	MB Read
/data/cache/HiggsToTauTau-00-00-44/user.TARRADEFabien.trig1_misal1_mc12.005200.T1_McAtNlo_Jimmy.A12.0.6.9.medium_cut_el.AAN.AANT1._00103.root	0	7	19	110
/data/cache/HiggsToTauTau-00-00-44/user.TARRADEFabien.trig1_misal1_mc12.005200.T1_McAtNlo_Jimmy.A12.0.6.9.medium_cut_el.AAN.AANT1._00114.root	0	6	20	50
/data/cache/HiggsToTauTau-00-00-44/user.TARRADEFabien.trig1_misal1_mc12.005200.T1_McAtNlo_Jimmy.A12.0.6.9.medium_cut_el.AAN.AANT1._00117.root	0	6	47	99
/data/cache/HiggsToTauTau-00-00-44/user.TARRADEFabien.trig1_misal1_mc12.005200.T1_McAtNlo_Jimmy.A12.0.6.9.medium_cut_el.AAN.AANT1._00113.root	0	5	24	70
/data/cache/HiggsToTauTau-00-00-44/user.TARRADEFabien.trig1_misal1_mc12.005200.T1_McAtNlo_Jimmy.A12.0.6.9.medium_cut_el.AAN.AANT1._00116.root	0	4	22	60
/data/cache/HiggsToTauTau-00-00-44/user.TARRADEFabien.trig1_misal1_mc12.005200.T1_McAtNlo_Jimmy.A12.0.6.9.medium_cut_el.AAN.AANT1._00107.root	0	4	28	80
/data/cache/HiggsToTauTau-00-00-44/user.TARRADEFabien.trig1_misal1_mc12.005200.T1_McAtNlo_Jimmy.A12.0.6.9.medium_cut_el.AAN.AANT1._00106.root	0	3	22	34

- Example: summary of data read in the past month
- Organized by user, server, client, files, and administratively definable data types
- Searchable; can act like a type of proto-catalog for accessed files

XrdMon Installation

- XrdMon setup is nontrivial
  - One line addition to xrootd.cf:

    xrootd.monitor all dest files info user acas0420.usatlas.bnl.gov:9930
  - Additional xrdmon.cf file (define DB parameters)
  - Data collector process: xrdmoncollector.pl
  - MySQL setup
  - Two scripts to load data from collector and prepare statistics: xrdmonLoadMySQL.pl & xrdmonPrepareStats.pl
  - Tomcat server for display interface (xrdmon.war)
- At each stage, extensive mods and corrections had to be made, some needed files were not publicly available.
- Some instabilities and issues remain
- Nice, useful tool - would love to see a mature available product!

Analysis Facility Plans

- Focus on developing/expanding Proof usage atop xrootd
- Plan to extend the testbed to ~50 nodes (200-400 cores), ~25-100 TB of disk storage in the next couple of weeks.
  - Hardware evaluation (optimal CPU/disk ratio)
- Further test Proof scalability/packetizers/optimal worker distribution
- Look at workload management with increased user base
- Test Proof cluster federation, first locally then offsite (Wisconsin)
  - Evaluate security issues involved with offsite access
- Test AthenaRootAccess with Proof
- Look at FUSE-based XrootdFS?
- Build up shared knowledge-base within ATLAS (T3 Proof farm already established at Wisconsin - see B. Mellado presentation)

Usage Model

- Facility usage model is still evolving
- Xrootd and Proof suggest different use cases, load types, and security issues
- How to manage utilization?
  - Select users? ”Analysis train” approach? COD?
- What particular niche to serve?
  - Attractive option for Tier 3 interactive analysis of smaller root-based datasets
- Need tools for data handling, cataloging, integration, workload management within the ATLAS framework
  - Many aspects will become clear with increased user base doing ”real” analysis
  - Can look at experience in other experiments (BaBar, Alice, CMS, Phobos, Star, ...)

Some useful URLs

- Xrootd Home Page
  - http://xrootd.slac.stanford.edu
  - mailing list: xrootd-l@slac.stanford.edu
- Configuring xrootd to run Proof
  - http://root.cern.ch/twiki/bin/view/ROOT/ProofInstallation
- US ATLAS Proof Testbed Twiki at BNL
  - http://www.usatlas.bnl.gov/twiki/bin/view/AtlasSoftware/ProofTestBed
- Wisconsin Proof/xrootd Test Pages
  - http://wisconsin.cern.ch/~nengxu/proof/
- Ganglia homepage
  - http://ganglia.sourceforge.net/