Setting up an ATLAS PROOF Facility

Ofer Rind - BNL/RACF

US Atlas Tier 2/Tier 3 Workshop, Ann Arbor, MI

28 May 2008

PROOF in ATLAS

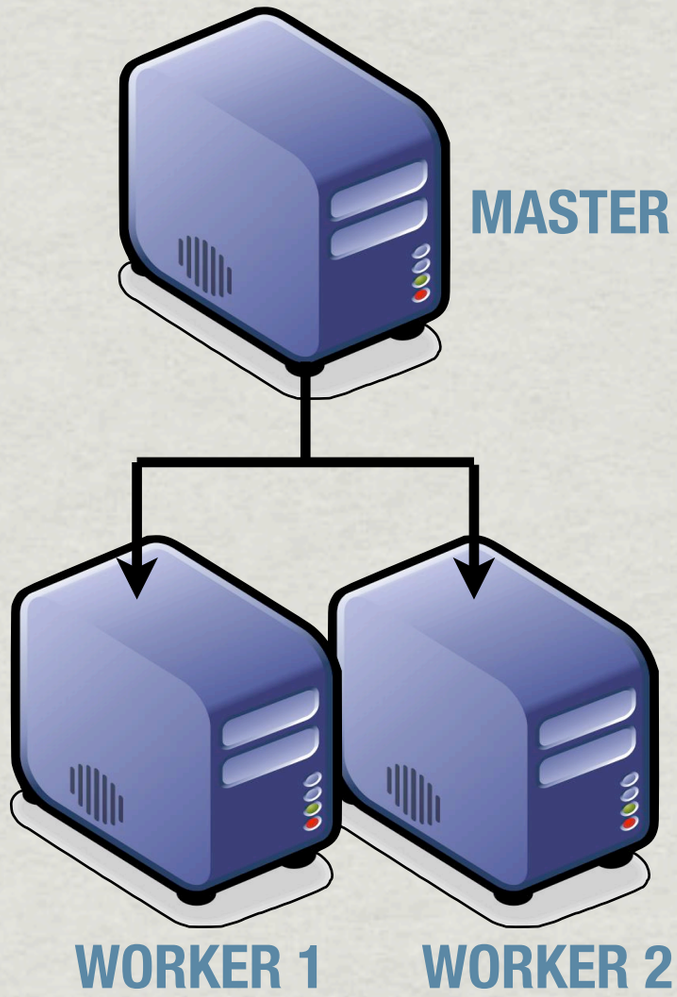
- Promising system for transparently parallelizing interactive or batch processing of large sets of ROOT files on a computing cluster
- Integration with XROOTD for data discovery and file serving
- In use at BNL T1, Munich LMU T2, Wisconsin T3.
  Test farms at UTA and Madrid T2.
- Provided herein is a brief sketch of how to set up a basic PROOF facility, but touching on a slightly advanced topic or two.

Basic Hardware Setup

Master-Slave Architecture

- PROOF Master (Xrootd redirector):
  Modest requirements
  Dual-core, 2G RAM/core should be fine.

- PROOF Workers (Xrootd Servers):
  Multi-core = multi-workers
  2G RAM/core
  Large, high throughput disk configuration
  - With multi-worker, disk contention is the major bottleneck
  - RAID may not be the optimal configuration
  - Solid-State Disk very promising! (see Sergey's talk)



Software Download & Server Setup

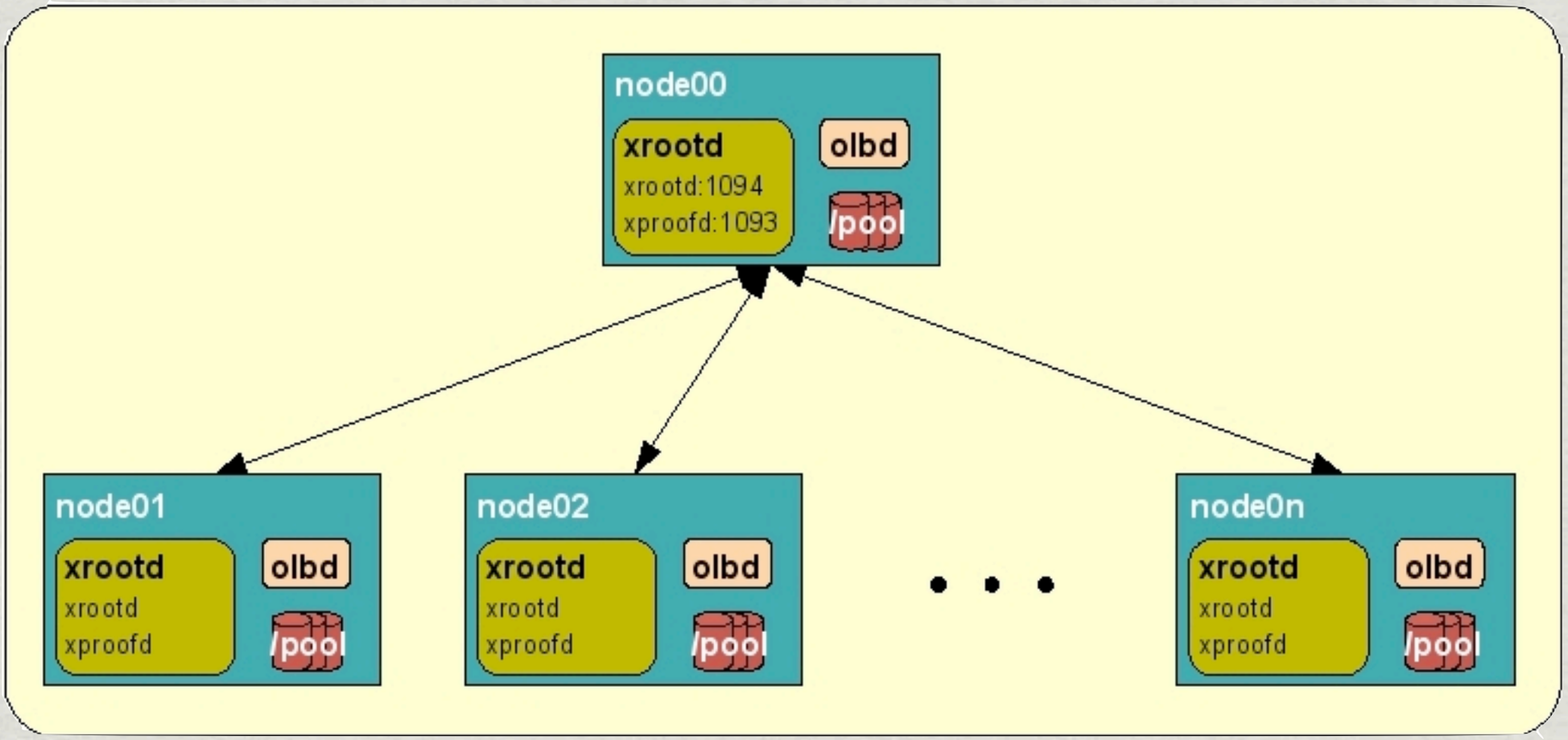
1. Download standard ROOT distribution package and make available on all hosts (AFS?)
   - Binary distros for 32 & 64 bit SLC4, MacOS, ...

2. In our setup, we copy or link binaries and libraries under /opt, but it's all configurable
   /opt/xrootd/bin/arch -> $ROOTSYS/bin
   /opt/xrootd/lib/arch -> $ROOTSYS/lib

3. Xrootd started up by root, but runs as non-privileged user
   - Create "xrdadmin" user
   - Set "xrdadmin" ownership for /opt/xrootd

4. On *workers*, set up local data repository (used for xrootd files + proofserv working directories)
   - In our example, mounted on /data
   - Ownership must be set to "xrdadmin"

Schematic cluster view



FROM HTTP://ROOT.CERN.CH/TWIKI/BIN/VIEW/ROOT/XPDEXAMPLETWO

- xrootd & olbd active on each host
- port 1094 (rootd) must be open to user on master/redirector
- olbd on port 3121 redirector↔server comm

Init Scripts & Sysconfig

Startup scripts: /etc/init.d/xrootd, /etc/init.d/olbd

Xrootd binary → XROOTD=/opt/xrootd/bin/arch/xrootd
Xrootd lib dir → XRDLIBS=/opt/xrootd/arch/lib

+ similar for olbd

Xrootd startup parameters: /etc/sysconfig/xrootd, /etc/sysconfig/olbd

```
XRDUSER="xrdadmin"
XRDLOG="/opt/xrootd/log/xrdlog"   NB:  Log dir must be owned by $XRDUSER
XRDCF="/opt/xrootd/etc/xrootd.cf"
XRDUSERCONFIG="/opt/xrootd/etc/xrd-userconfig.sh”
```

+ similar for olbd

Xrootd env variables: /opt/xrootd/etc/xrd-userconfig.sh

```
export LD_LIBRARY_PATH=/afs/usatlas/project/oracle/instantclient/10.2.0.2/lib::/opt/usatlas/lib:/afs/
usatlas.bnl.gov/cernsw/lcg/external/root/[version]/slc4_ia32_gcc34/root/lib:$LD_LIBRARY_PATH
export ROOTSYS=/afs/usatlas.bnl.gov/cernsw/lcg/external/root/[version]/slc4_ia32_gcc34/root
```

Example: Xrootd Config File

```
xrd.port 1094
olb.port 3121

xrootd.fslib /opt/xrootd/lib/arch/libXrdOfs.so

xrootd.export /data r/w  # Valid path prefix for file requests.

if acas0420.usatlas.bnl.gov   # Define manager/server roles
  all.role manager
  ofs.forward all
else
  all.role server
fi

all.manager acas0420.usatlas.bnl.gov 3121    # Manager location (ignored by managers)

oss.cache public /data/cache*          # Storage path pointers
oss.path /data/proofpool r/w
oss.path /data/proofbox r/w

olb.path w /data          # Paths handled by server
olb.delay startup 30    # Delay client requests at manager startup
```

Example: Xrootd Config File (cont.)

```
# PROOF part

if exec xrootd
xrd.protocol xproofd /opt/xrootd/lib/arch/libXrdProofd.so    # Load the XrdProofd protocol
fi

# location of root distribution
xpd.rootsys /afs/usatlas.bnl.gov/cernsw/lcg/external/root/5.14.00/slc4_ia32_gcc34/root

xpd.workdir /data/proofbox   # User sandboxes

xpd.resource static /opt/xrootd/etc/proof.cf   # Where to find node list

if acas0420.usatlas.bnl.gov    # Define master/worker roles
   xpd.role master
else
   xpd.role worker
   xpd.allow acas0420.usatlas.bnl.gov
fi

xpd.poolurl root://acas0420.usatlas.bnl.gov   # URL for data storage
xpd.namespace /data/proofpool
```

MORE ON XPD DIRECTIVES: HTTP://ROOT.CERN.CH/TWIKI/BIN/VIEW/ROOT/XPDDIRECTIVES

PROOF config file & Startup

Define # of workers (1 per core) →

```
master acas0420.usatlas.bnl.gov

worker acas0421.usatlas.bnl.gov
worker acas0421.usatlas.bnl.gov
worker acas0421.usatlas.bnl.gov
worker acas0421.usatlas.bnl.gov
...
```

Now, ready to start up! On each node, as root user:

```
> /etc/init.d/xrootd start
> /etc/init.d/olbd start
```

To copy in data:

```
> xrdcp  yourfile.root root://acas0420.usatlas.bnl.gov/data/test/yourfile.root
```

PROOF is ready!

Try connecting to PROOF master:

```
> root -b
 ******************************************
 *
 *       W E L C O M E  to  R O O T
 *
 *   Version  5.18/00c       19 May 2008
 *
 *  You are welcome to visit our Web site
 *          http://root.cern.ch
 *
 ******************************************

ROOT 5.18/00c (branches/v5-18-00-patches@23940, May 21 2008, 13:10:27 on linux)

CINT/ROOT C/C++ Interpreter version 5.16.29, Jan 08, 2008
Type ? for help. Commands must be C++ statements.
Enclose multiple statements between { }.
root [0] TProof *p = TProof::Open("acas0420")
Starting master: opening connection ...
Starting master: OK
Opening connections to workers: OK (36 workers)
Setting up worker servers: OK (36 workers)
PROOF set to parallel mode (36 workers)
root [1]
```

Some further topics...

- Authentication
- Administration
- Data management
- Resource sharing

PROOF with authentication

- Leverages well-developed xrootd security plug-in infrastructure - pwd, krb, afs, gsi
- Auth can be required at multiple levels - PROOF master, PROOF slaves, xrootd file access
- Authentication enabling directives (we tested pwd and krb5):

Use PWD file ➡
-*or*- KRB5 service principal ➡

```
xpd.seclib libXrdSec.so
xpd.sec.protocol pwd
xpd.sec.protocol krb5 /etc/krb5.keytab.xrd.xrootd xrd/xrootd
```

- Generating password file:

```
> xrdpwdadmin add -host acas0420.usatlas.bnl.gov -email xrdadmin@bnl.gov
> xrdpwdadmin add [user]
```

- More information: http://root.cern.ch/twiki/bin/view/ROOT/XpdAuthConfig
  + xrootd security reference

Administrative Tools

- Life Support - Users all over the machines now, what if daemons crash?
  - Robust - xrootd/olbd restart always seems quick and easy
  - At BNL, we run simple cron check/restart/notify every 3 minutes
- Admin Scripts - necessary to move data or other files around on the cluster
  - Some offerings in $ROOTSYS/etc/proof/utils/
  - At BNL, we use Tentakel + sudo for limited nonprivileged user admin capabilities
  - Other approaches? (cfengine+capify at ALICE-GSI)

Subscribing and Advertising Your Data

- Not specifically a PROOF issue, but an important consideration for its viability within ATLAS is the associated data management
- Beginning with FDR-1, LRC was used to provide catalogue services
  - Data movement handled by Sergey using a set of scripts that
    1. Pull files from dCache via xrdcp door
    2. Update LRC on successful file transfer

Check if a dataset is available on Xrootd farm

```
> dq2-list-dataset-replicas fdr08_run1.0003052.StreamJet.merge.AOD.o1_r6_t1
 INCOMPLETE: LNF
 COMPLETE: TAIWAN-LCG2_DATADISK,SLACXRD,WISC,IFICDISK,TOKYO-LCG2_DATADISK,UKI-SOUTHGRID-
BHAM_DATADISK,NDGF-T1_DATADISK,MWT2_IU,LIP-LISBON,MANC,IN2P3-CC_DATADISK,TIER0TAPE,TRIUMF-
LCG2_DATADISK,INFN-T1_DATADISK,BU_DDM,JINR,CYF,RAL-LCG2_DATADISK,PIC_DATADISK,SWT2_CPB,SARA-
MATRIX_DATADISK,BNL-OSG2_DATADISK,IJS[illegible],FZK-LCG2_DATADISK,DESY-ZN,DESY-HH,AGLT2_SRM,PNPI,TORON,TW-
FTT,UKI-SOUTHGRID-OX-HEP_DATADISK,BNLXRDHDD1,MWT2_UC,OU
```

"SE NAME" = BNLXRDHDD1

Get PFN for your files on Xrootd farms

```
> dq2_ls -fp -s BNLXRDHDD1 fdr08_run1.0003052.StreamJet.merge.AOD.o1_r6_t1

fdr08_run1.0003052.StreamJet.merge.AOD.o1_r6_t1   Total: 2  - Local: 2
   root://acas0420.usatlas.bnl.gov//data/fdr08_run1.0003052.StreamJet.merge.AOD.o1_r6_t1/
fdr08_run1.0003052.StreamJet.merge.AOD.o1_r6_t1._0001.1
   root://acas0420.usatlas.bnl.gov//data/fdr08_run1.0003052.StreamJet.merge.AOD.o1_r6_t1/
fdr08_run1.0003052.StreamJet.merge.AOD.o1_r6_t1._0002.1
```

Subscribing and Advertising Your Data

dq2_ls –fp –s BNLXRDHDD1 “my_dataset”

Analysis

acas0420.usatlas.bnl.gov

DQ2

T0

Panda

GRID transfers

xrdcp with dq2 registration

dCache

/DATA

BNLXRDHDD1

Xrootd/PROOF Farm

COURTESY: S. PANITKIN

NOTE: This is not a grid end point! You can not copy files to or from here using dq2 tools!

➡ See Neng’s talk for developments in automating data mgmt

How to share resources with a batch system (Condor)?

- PROOF needs to claim resources opportunistically from batch. (At least) 3 approaches:
  - CPU priority: No suspension, but batch jobs run at higher nice level
    - \+ Simple to configure; no disruptive suspension of batch (external db conns, in-process file transfer, etc.)
    - \- Memory resources not released
  - Condor directed suspension: COD claim (or higher priority job slot) causes UNIX suspension of batch and no new jobs starting
    - \+ CPU, mem cleared quickly; easy to turn on COD; experience with both COD and priority suspension (PHOBOS)
    - \- Proofserv must be started and managed by Condor; admin overhead in managing COD users

Resource sharing (cont.)

- Load-based suspension: Condor detects load on system triggering UNIX suspension of batch and no new jobs starting

  + CPU, mem cleared as needed (same effect as COD suspension); PROOF startup independent of Condor; simple host ClassAd

  - Other load sources could cause suspension

Example from STAR:

```
HighLoad = 2.0
NonCondorLoadAvg = (LoadAvg - CondorLoadAvg)
IsOwner  = ($(NonCondorLoadAvg) >= $(HighLoad))
Suspend = ($(NonCondorLoadAvg) >= $(HighLoad))
Continue = ($(NonCondorLoadAvg) < $(HighLoad))
```

A sampling of PROOF resources

- PROOF installation reference: http://root.cern.ch/twiki/bin/view/ROOT/ProofInstallation
- Neng's quick installation guide: http://wisconsin.cern.ch/~nengxu/xrootd_install/Install_PROOF.ppt
- Sergey's PROOF user tutorial from the BNL FDR-1 analysis jamboree: http://indico.cern.ch/materialDisplay.py?contribId=11&sessionId=4&materialId=slides&confId=27915
- Atlas PROOF users hypernews forum: hn-atlas-proof-xrootd@cern.ch https://hypernews.cern.ch/HyperNews/Atlas/get/proofXrootd.html
- ROOT forum: http://root.cern.ch/phpBB2/