Xrootd, XrootdFS and BeStMan

Wei Yang

2009-10-30 US ATALS Tier 3 meeting, ANL

- ✧ **Xrootd Storage components**

- ✧ **How does Xrootd works**
- ✧ **What is XrootdFS**

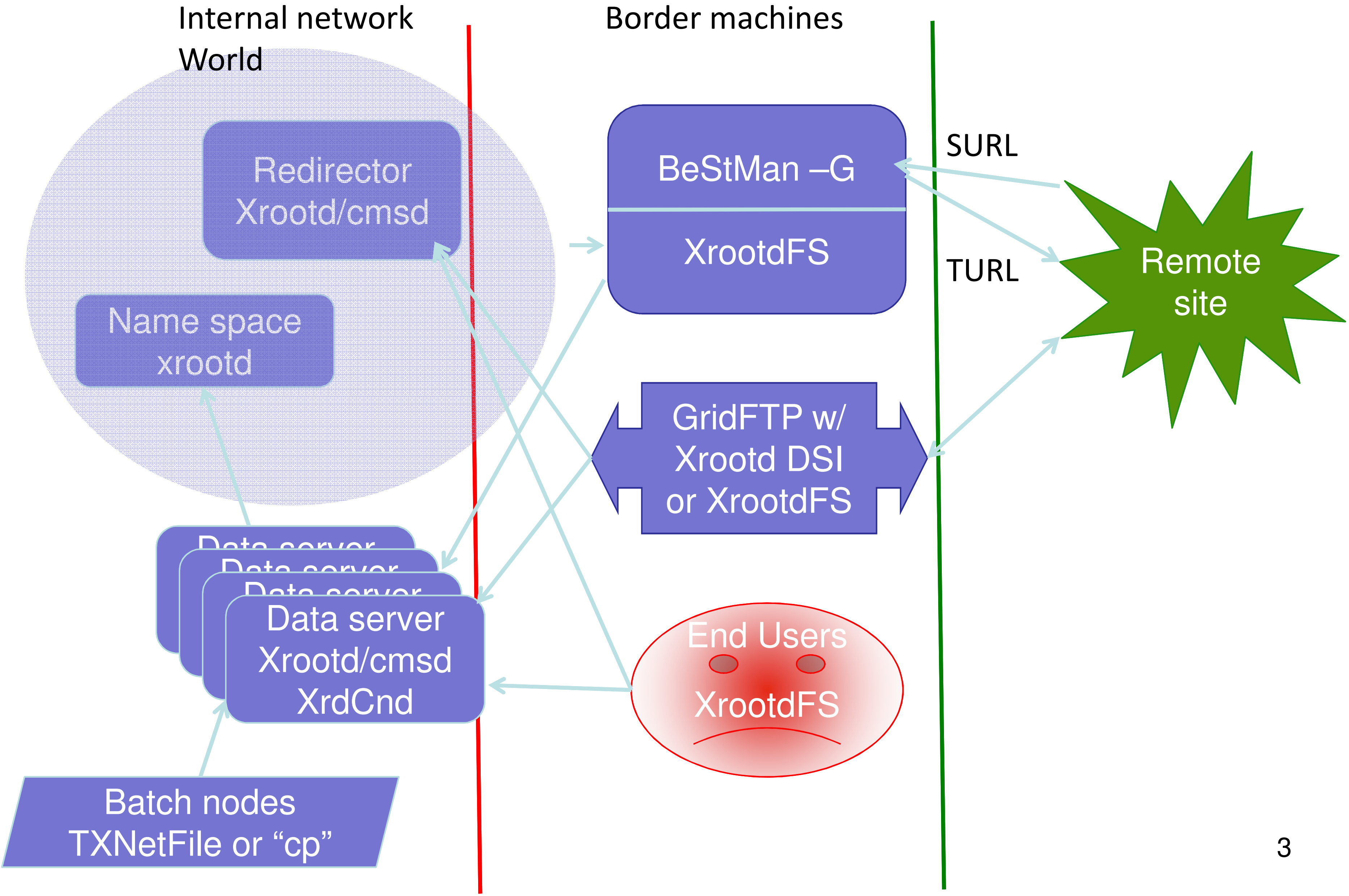
- ✧ **How to access Xrootd storage**

  Interactive

  From ATLAS jobs

- ✧ **BeStMan**

Storage Architecture



Storage Components

❑ **Bestman Gateway** ← T2/T3g

◆ **XrootdFS** ← For users and minimum T3g

- Usage is like NFS
- Based on Xrootd Posix library and FUSE
- BeStMan, dq2 clients, and Unix tools need it

◆ **GridFTP for Xrootd** ← WT2 for a while

- Globus GridFTP + Data Storage Interface (DSI) module for Xrootd/Posix

✧ **Xrootd Core** ← All Babar needed is this layer

Redirector, data servers, xrdcp

More functionality

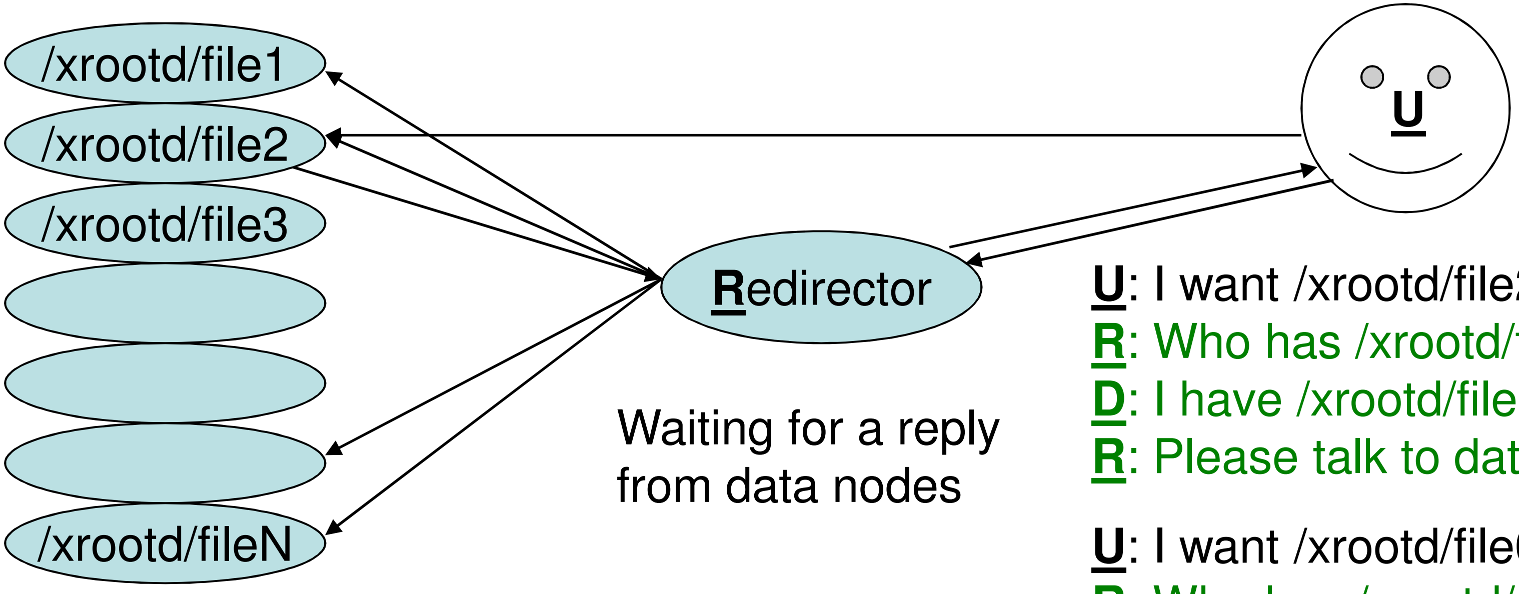
How Xrootd works

Glue file servers together by a redirector

User only need to know XROOT path: root://redirector:port//path/file

Simple, low overhead

- No complex features such as locking
- Good for reading dominated environment, e.g. HEP data analysis



Waiting for a reply
from data nodes

Data node 1,2,3,…,N

U: I want /xrootd/file2
R: Who has /xrootd/file2
D: I have /xrootd/file2
R: Please talk to data node 2

U: I want /xrootd/file0
R: Who has /xrootd/file0
Time out!
R: Nobody has /xrootd/file0

Xrootd Export Path, Disk Cache and Space Token

$VDT_LOCATION/xrootd/etc/xrootd.cfg

- Xrootd Export Path is what user will use to access file

```
all.export = /xrootd          =>          root://host:port//xrootd/file
```

- Xrootd Disk Caches are hard disk partitions storing data files

```
Filesystem            Size  Used Avail Use% Mounted on
/dev/sdb               12G  6.0G  5.0G  55% /xrdcache01
```

```
oss.cache   public   /xrdcache01
```

Export Path contains directories and symlinks, pointing to data files OSS Cache

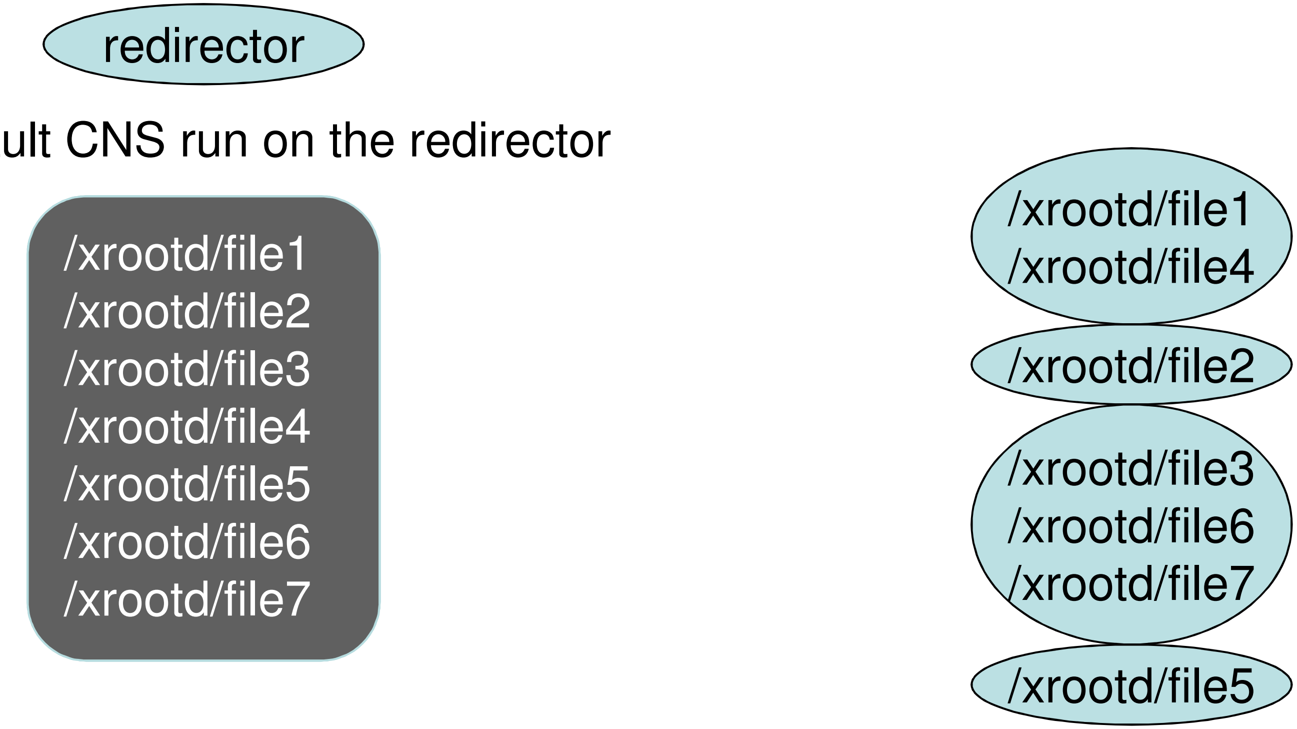
- To support WLCG static space tokens, add more cache groups

```
oss.cache   public   /xrdcache01     xa      # "xa": extend attributes
oss.cache   tokenA  /xrdcache01     xa
```

User create a file using   root://host:port//xrootd/file?oss.cgroup=tokenA

Composite Name Space (CNS)

A standalone Xrootd instance, not part of the main Xrootd cluster



redirector

By default CNS run on the redirector

/xrootd/file1
/xrootd/file2
/xrootd/file3
/xrootd/file4
/xrootd/file5
/xrootd/file6
/xrootd/file7

/xrootd/file1
/xrootd/file4

/xrootd/file2

/xrootd/file3
/xrootd/file6
/xrootd/file7

/xrootd/file5

Empty files (with the “right size”). All in one
Standalone Xrootd node.

They are there for directory browsing

Real files, distributed on
Several Xrootd nodes

User interface to Xrootd

TXNetFile class (C++ and ROOT CINT)

Fault tolerance
High performance thought intelligent logics in TXNetFile and server

Command line tools

- **xrdcp**

  simple, native, light weight, high performance

- **Xrootd Posix preload library**

  ```
  export LD_PRELOAD=/…/libXrdPosixPreload.so
  ls/cat/cp/file root://redirector:port//path/file
  ```

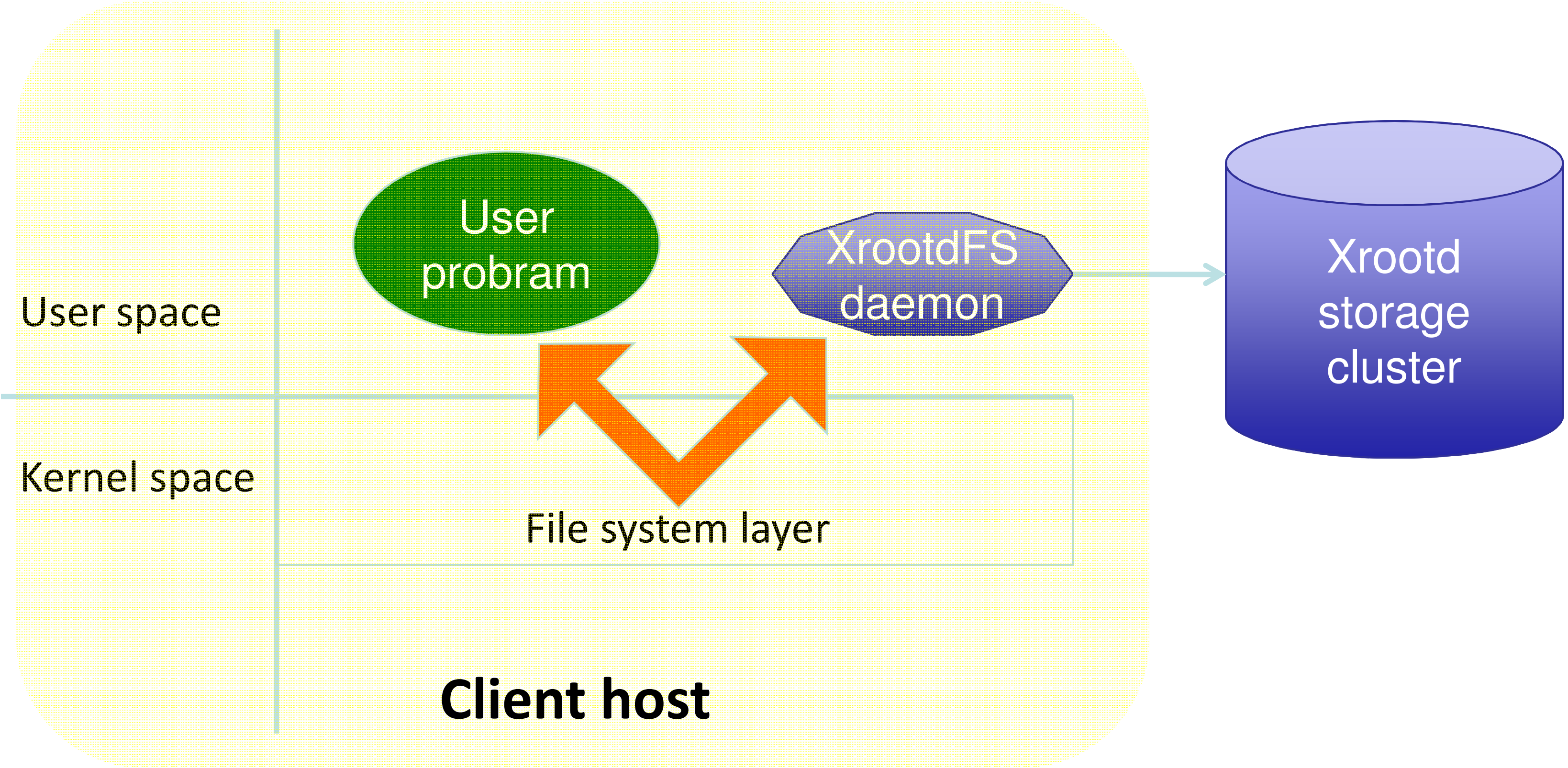
  A subset of UNIX I/O command will work with Posix preload library on files, not on directories

  Some overhead, I/O performance isn't as good as xrdcp

User interface to Xrootd, cont'd

XrootdFS, a client of Xrootd

- Easy to use: NFS like accessing to data in Xrootd.
- Relatively expensive compare to direct accessing



XrootdFS, cont'd

File system interface for Xrootd

Mount the Xrootd cluster on client host's local file system tree

Provide standard Posix I/O interface to the Xrootd cluster

- open(), close(), read(), write(), lseek(), unlink(), rename()
- opendir(), closedir(), readdir(), mkdir()

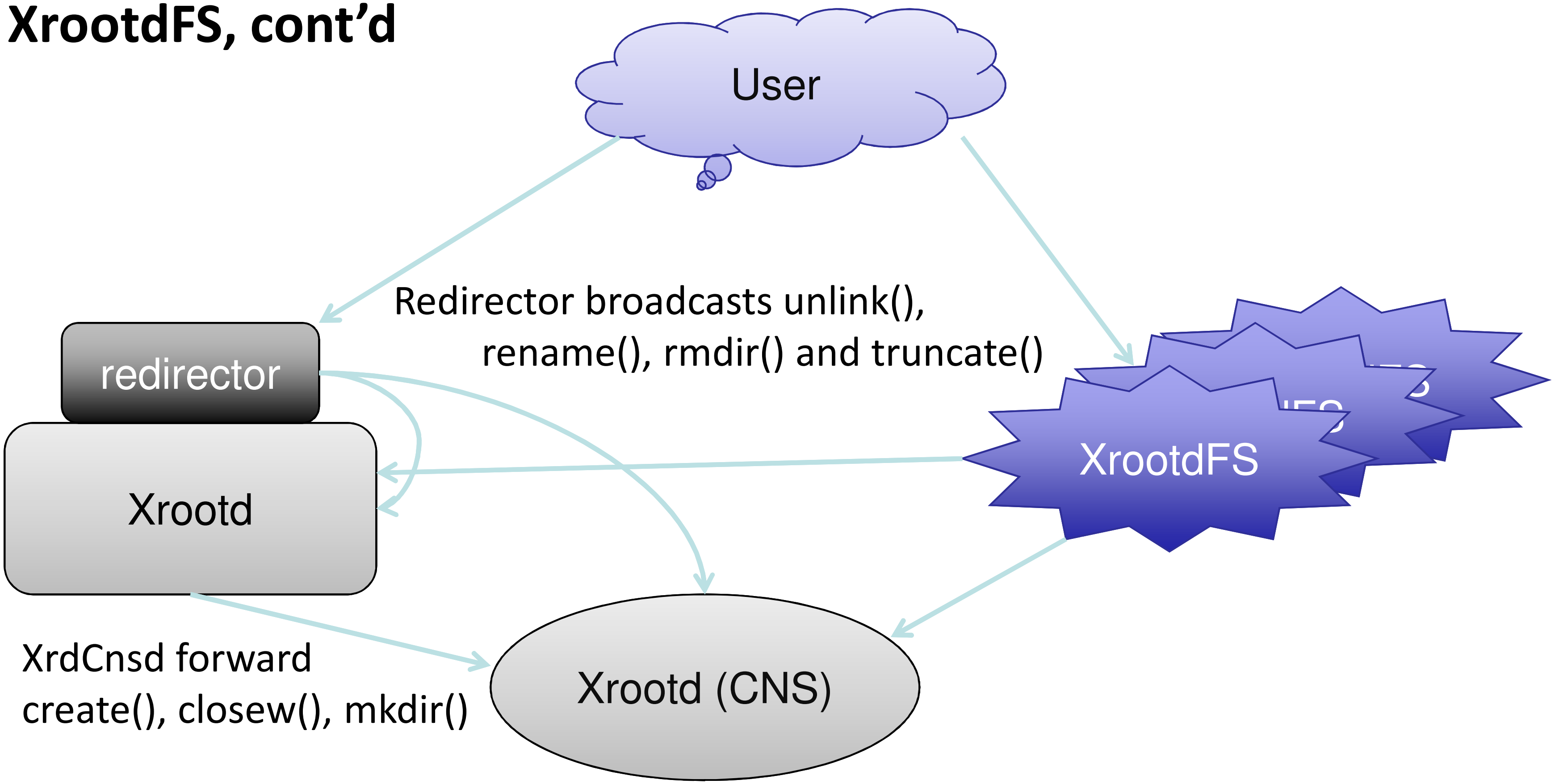
Work with most UNIX commands/tools

- cd, ls, cp, rm, mkdir, cat, grep, find
- ssh/sftp server, gridftp server, SRM, xrootd server
- scp/sftp, gridftp clients, SRM clients, ATLAS **dq2 clients**

Be aware: no file locking, no ownership/protection
file creation delay
some UNIX command are not scalable, e.g. find, ls
cp is slow (due to small I/O block size) 🡐 **no longer true**

Reduce # of network connections to Xrootd data server
More overhead on I/O performanc

XrootdFS, cont’d



Redirector broadcasts unlink(),
rename(), rmdir() and truncate()

XrdCnsd forward
create(), closew(), mkdir()

Each Xrootd data server contains part of the directory tree
CNS has a complete directory tree, with shadow files

http://wt2.slac.stanford.edu/xrootdfs/xrootdfs.html

XrootdFS Configuration

($VDT_LOCATION/xrootdfs/bin/start.sh)

XrootdFS is a Xrootd client. The following script starts XrootdFS

```
export LD_LIBRARY_PATH=${LD_LIBRARY_PATH}:/opt/vdt/xrootd/lib:/opt/fuse/lib

export XROOTDFS_OFSFWD=0
#    export XROOTDFS_USER='daemon’
export XROOTDFS_FASTLS="RDR”
insmod /lib/modules/`uname -r`/kernel/fs/fuse/fuse.ko 2> /dev/null

export XROOTDFS_RDRURL="root://xrootd-redirector:1094//xrootd”
export XROOTDFS_CNSURL="root://CNS:2094//xrootd” (optional for non-interactive machines)

MOUNT_POINT="/xrootd”
xrootdfsd $MOUNT_POINT -o allow_other,fsname=xrootdfs,max_write=131072

$ df -h
Filesystem            Size  Used Avail Use% Mounted on
xrootdfs               55T   34T   22T  62% /xrootd
```

Use “umount /xrootd” to stop XrootdFS

Accessing Xrootd data from ATLAS jobs

- **Copy input data from Xrootd to local disk on WN**

  A wrapper script using xrdcp, or cp + xrootd posix preload library
  Panda production jobs at SLACXRD work this way.

- **Read ROOT files directly from Xrootd storage**

  Identify ROOT file using Unix 'file' command (w/ posix preload library)
  Copy non-ROOT files to local disk on WN
  Put ROOT file's xroot URL (root://…) in PoolFileCatalog.xml
  Athena uses TXNetFile class to read ROOT file
  ANALY_SLAC and ANALY_SWT2_CPB use this mixed accessing mode.

Both need a set of tools for copying, deleting, file id and checksum

- **Mount XrootdFS on all batch nodes**

  All files appear under local file system tree.
  None of the above is needed
  Untested: XrootdFS came out after SLAC sites were established.

BeStMan Full mode and BeStMan Gateway mode

- Full implementation of SRM v2.2
- Support for dynamic space reservation
- Support for request queue management and space management
- Plug-in support for mass storage systems
- Follows the SRM v2.2 specification

- Support for essential subset of SRM v2.2
- Support for pre-defined static space tokens
- Faster performance without queue and space management
- Follows the SRM functionalities needed by ATLAS and CMS

Bestman-Gateway for Xrootd Storage

($VDT_LOCATION/bestman/conf/bestman.rc)

Stable! we tuned a few parameters

Java heap size: (1300MB on a 2GB machine)
Recently increased the # of contains thread from 5 to 25

Make sure BeStMan-G's external dependences are working

When Xrootd servers are under stress

- Xrootd stat() call takes too long:
  result in HTTP time out or CONNECT time out

- Redirector can't locate a file, result in file not found
- Panda jobs (not going though SRM interface) will also suffer

GridFTP configuration

- Globus GridFTP on XrootdFS
  - No additional configuration
  - May have performance penalty
- Data Storage Interface (DSI) module for Xrootd/Posix

Use along with Xrootd Posix preload library

```
$ cat $VDT_LOCATION/vdt/services/vdt-run-gsiftp.sh
#!/bin/sh

. $VDT_LOCATION/setup.sh
export LD_PRELOAD=/opt/xrootd/lib/libXrdPosixPreload.so
export XROOTD_VMP=“xrootd-redirector:port:/xrootd=/xrootd”
# Make sure “libglobus_gridftp_server_posix_gcc32dbg.so” is in LD_LIBRARY_PATH
exec $VDT_LOCATION/globus/sbin/globus-gridftp-server -dsi posix
```

How to access:

root://xrootd-redirector:port//xrootd = gsiftp://gridftpserver/xrootd