

dCache/XRootD

Dmitry Litvintsev
(DMS/DMD)

What is dCache

- Distributed multi-petabyte scalable disk storage system with single rooted filesystem providing location independent file access
 - Can be interfaced with tertiary storage serving as cache front-end to optimize I/O
- dCache.org is a collaboration between DESY, Fermilab and NDGF
- dCache is an open source (github.com/dCache/dcacha)

dCache worldwide



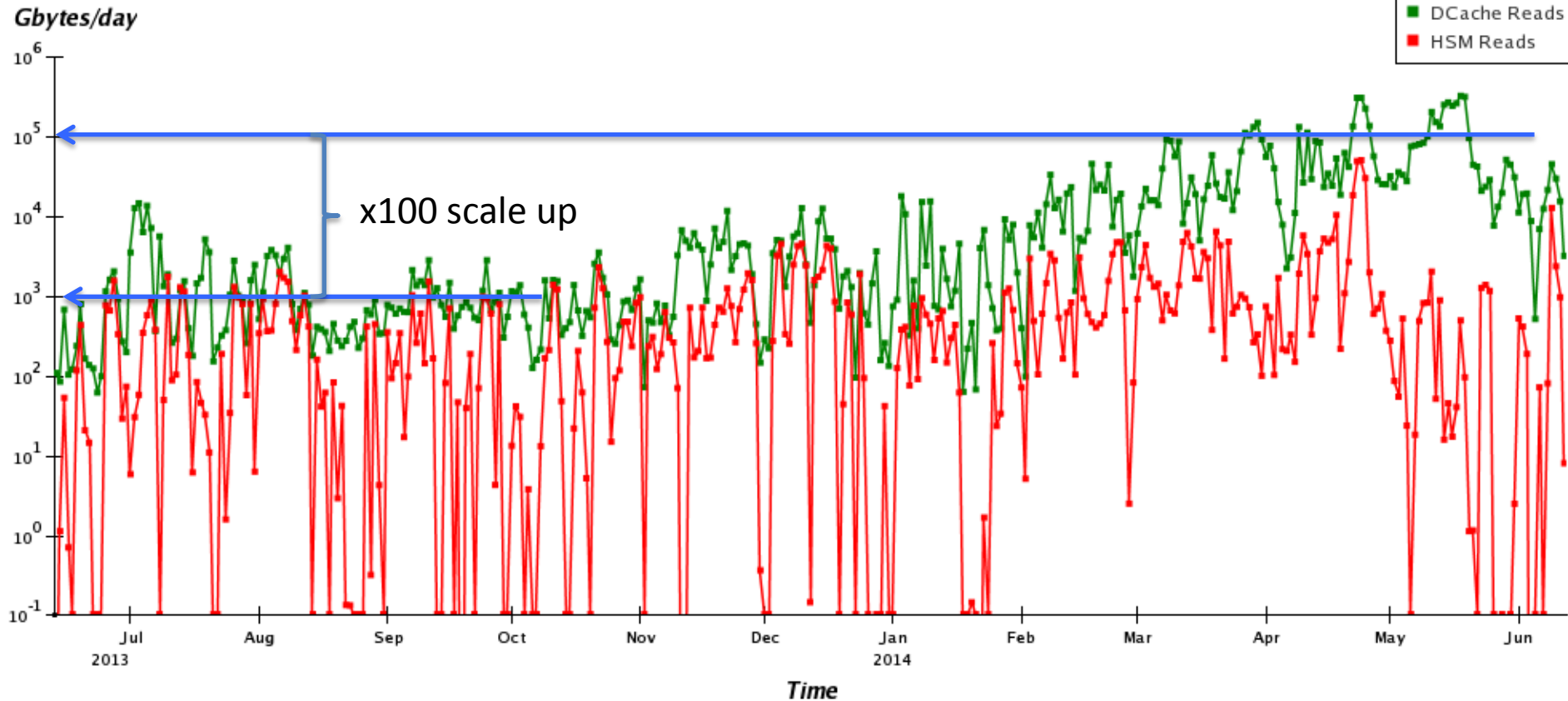
More that 70 installations, with total storage volume of ~ 120 PB

dCache @ Fermilab

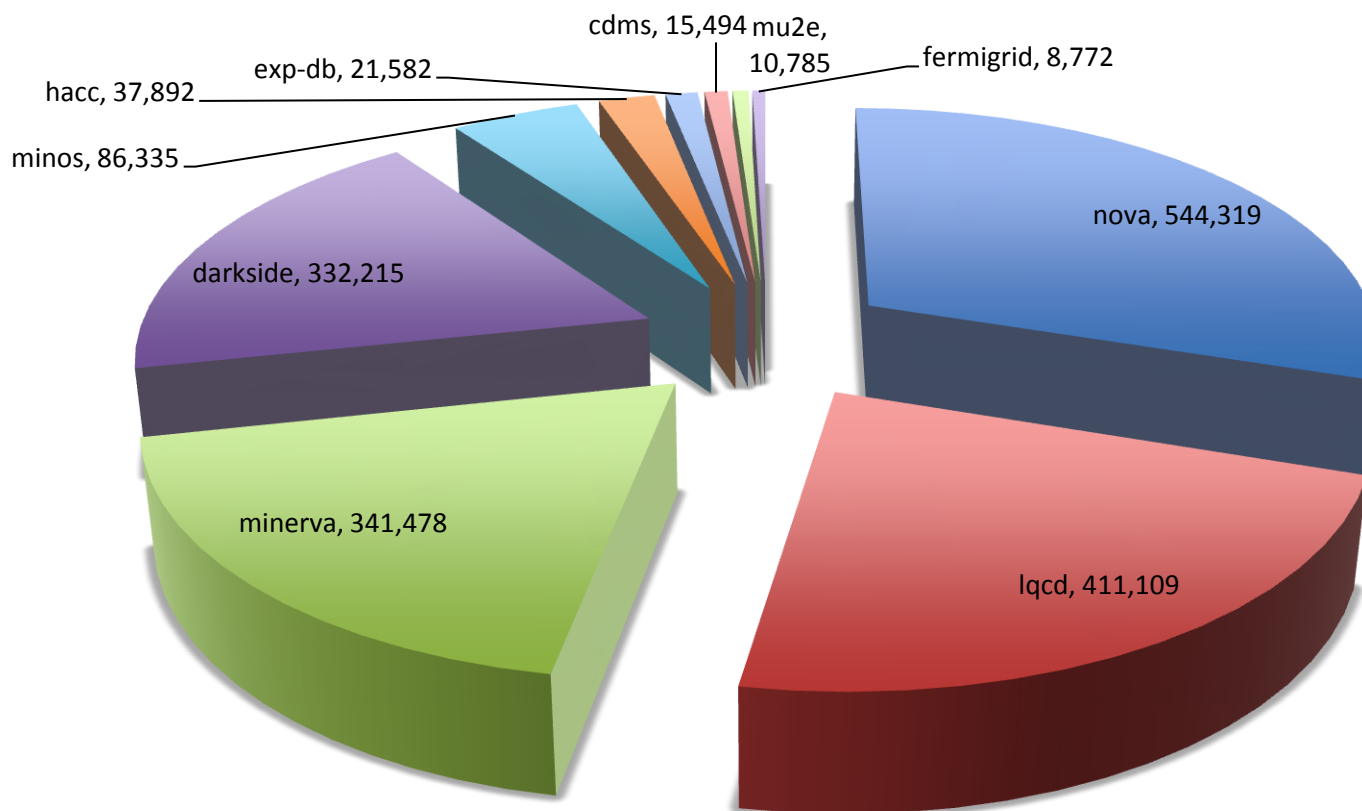
- Long and successful history:
 - Run II : CDF (in production since 2003)
 - General purpose Public dCache (in production since 2004)
 - CMS (since 2005)
 - D0 (in production since 2014)
- Fall 2013: public dCache expanded by a factor of 20 to meet needs of IF experiments (a part of FIFE program)

Read transfer volume scaling

Bytes Read (365 days since Thu Jun 13 00:00:00 CDT 2013)



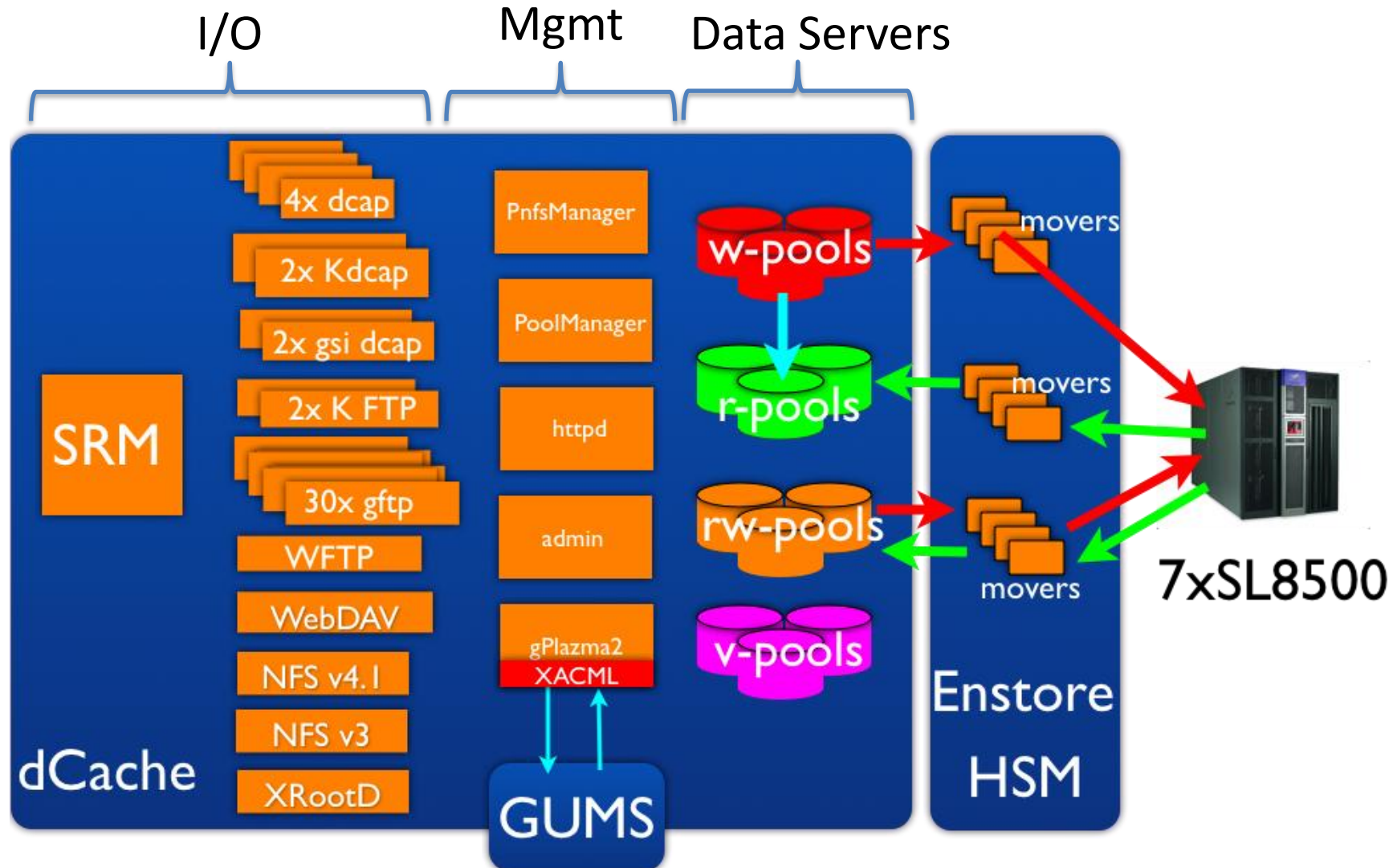
On-line space usage in dCache



On-line space usage in public dCache by storage group (top 10) in GiBs

Total used space currently is about 2PiBs

Public dCache (FNDCA)



Access Protocols

Protocol	URL	Authentication	Client(s)
dcap	dcap://fndca1.fnal.gov:{24125,24136,24137,24138}	None	dccp, dcap, root
	dcap://fndca1.fnal.gov:{24525,24536}	GSI	
	dcap://fndca1.fnal.gov:{24725,24736}	Kerberos(GSS)	
FTP	fndca1.fnal.gov:24126	passwd	ftp
	fndca1.fnal.gov:24127	kerberos(GSS)	
gFTP	fndca1.fnal.gov:{2811,2812}	GSI	globus-url-copy, srmcp, uberftp
SRM	srm://fndca1.fnal.gov:8443	GSI	lcg-cp, srmcp...
NFS v4.1	stkensrv1n.fnal.gov:2049	None	POSIX I/O
NFS v3	stkensrv1n.fnal.gov:2049	None	POSIX metadata ops, POSIX I/O with dcap preload library
WebDAV	https://fndca1.fnal.gov:2880	GSI	Browsers, curl, cadaver, davfs, root
XRootD	(x)root://fndca1.fnal.gov:1094	GSI	xrdcp, xrd, POSIX preload library, FUSE, root

Access from root

- Provided the proper plugin the access to files in dCache from root looks like:
 - dcap (no auth) on mounted namespace (NFS v3) using preload library
`mount -o nfsvers=3 stkensrv1n:/minos /pnfs/minos`
`export LD_PRELOAD=/usr/lib64/libpdcap.so.1`
`root [0] f=TFile::Open("/pnfs/minos/path/to/file");`
 - dcap (no auth)
`root [0] f=TFile::Open("dcap://fndca1.fnal.gov:24125/full/path/to/file");`
 - dcap (GSI Auth)
`{grid,voms}-proxy-init ...`
`export DCACHE_IO_TUNNEL=/usr/lib64/dcap/libgsiTunnel.so`
`root [0] f=TFile::Open("dcap://fndca1.fnal.gov:24525/full/path/to/file");`
 - dcap (kerberos Auth)
`kinit <username>@FNAL.GOV`
`export DCACHE_IO_TUNNEL=/usr/lib64/dcap/libgssTunnel.so`
`root [0] f=TFile::Open("dcap://fndca1.fnal.gov:24725/full/path/to/file");`

Access from root

- NFS v4.1:

```
mount -t nfs4 -o minorversion=1 stkensrv1n:/minos /pnfs/minos  
root [0] f=TFile::Open("/pnfs/minos/path/to/file");
```

- WebDAV (TDavixFile plugin)

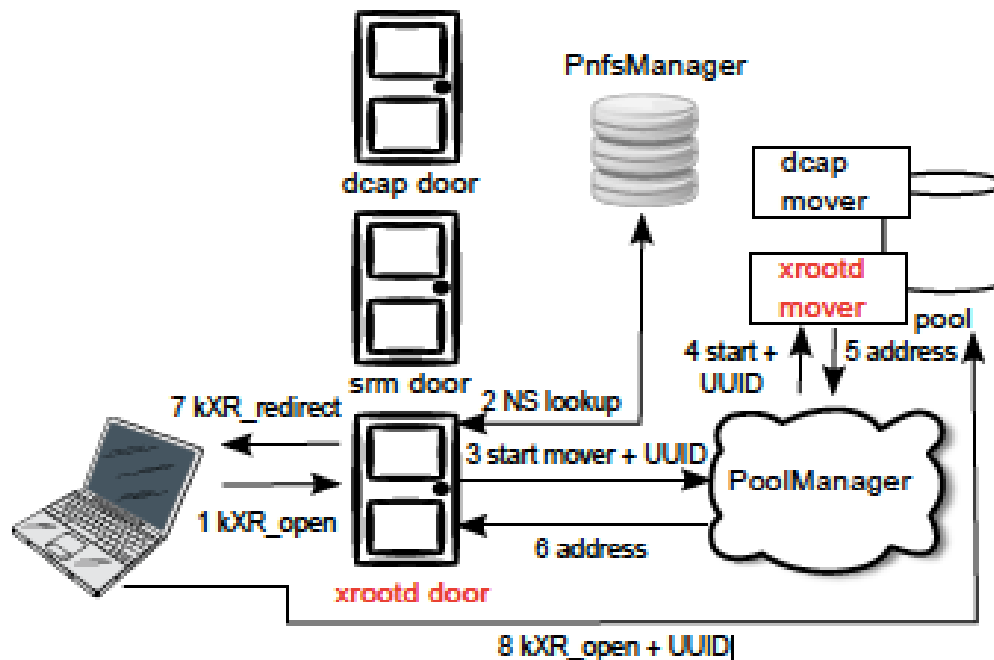
```
{grid,voms}-proxy-init ..  
root [0] f=TFile::Open("https://fndca1.fnal.gov:2880/path/to/file");
```

XRootD

- Extended rootd (eXtended root daemon)
- File access and data transfer protocol
 - POSIX-like random access to arbitrary data organized in files of any type
- A reference implementation of
 - xrootd daemon provides access to distributed data servers
 - cmsd daemon clusters xrootd daemons
- Provides fault tolerant, low latency, high bandwidth access to data

dCache XRootD

- dCache provides fully functional xRootD server
 - Native Java implementation of xrootd protocol
 - Acts as any other dCache door. dCache door is xrootd redirector
 - Mover and pools are xrootd data servers



XRootD door

- On public dCache we run XRootD door:

`root://fndca1.fnal.gov:1094`

- How to use:

- Create {grid,voms} proxy

`{grid,voms}-proxy-init ...`

- With root files:

`root [0]`

`f=TFile::Open("root://fndca1.fnal.gov/pnfs/fnal.gov/usr/fermigrid/volatile/fermilab/litvinse/xib.root");`

- With any files

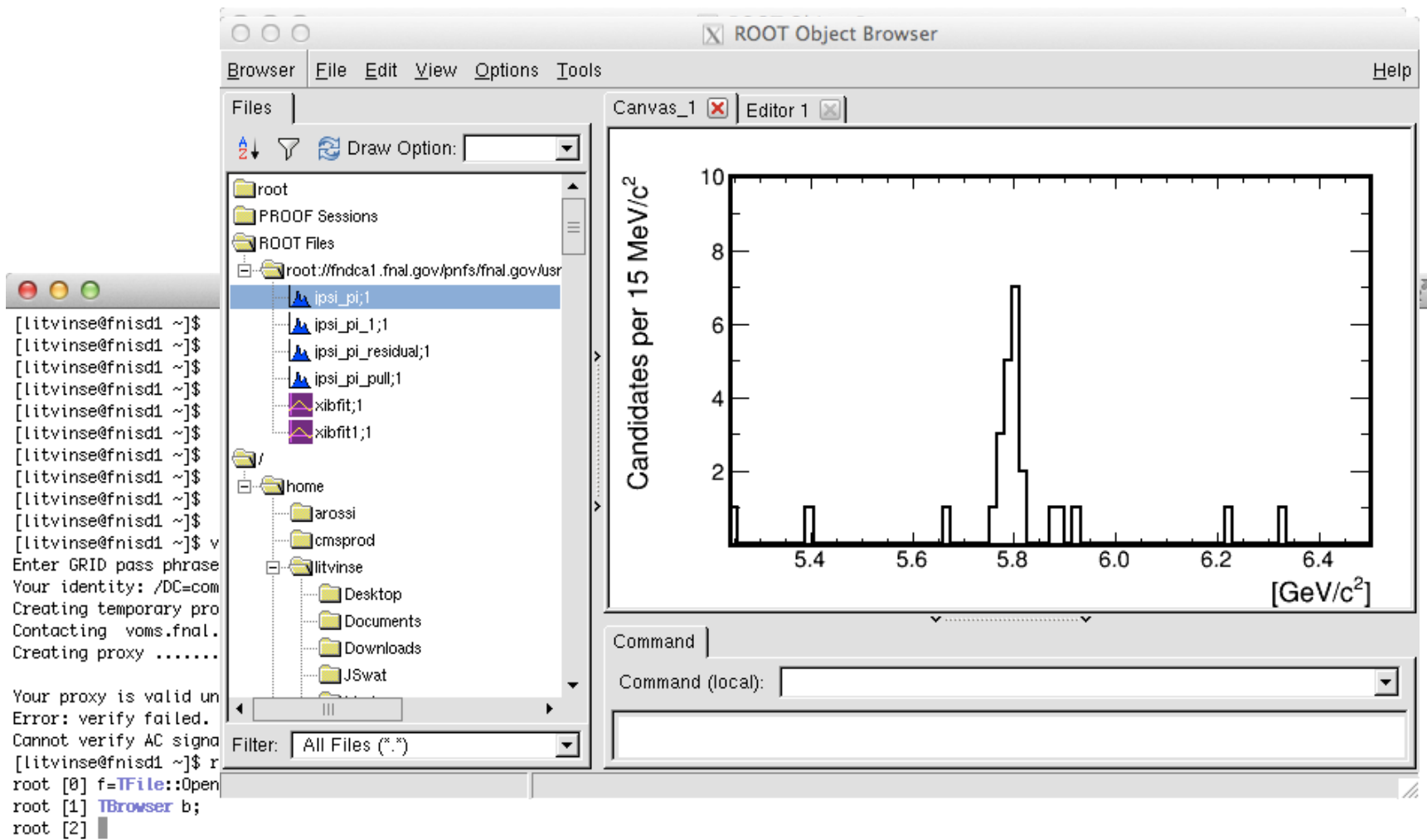
`xrdcp`

`xroot://fndca1.fnal.gov/pnfs/fnal.gov/usr/fermigrid/volatile/fermilab/litvinse/gums1.txt .`

`[xrootd] Total 0.00 MB |=====| 100.00 % [0.0 MB/s]`

- Use xrd - xrootd file and directory meta-data utility

XRootD usage from root



xrd client

- Invoke it like so

X

- ```
root://fndca4a.fnal.gov:1094/> cat /pnfs/fnal.gov/usr/fermigrd/volatile/fermilab/litvinse/gums1.txt
```

```
dmitrylitvintsev — litvinse@fnisd1:~ — ssh — 139x47
[litvinse@fnisd1 ~]$ xrd fndca1.fnal.gov
(C) 2004-2010 by the Xrootd group. Xrootd version: v3.3.6
Welcome to the xrootd command line interface.
Type 'help' for a list of available commands.
root://fndca4a.fnal.gov:1094/> help
(C) 2004-2010 by the Xrootd group. Xrootd version: v3.3.6
```

```
workshop — litvinse@fnisd1:~ — ssh — 124x25
[litvinse@fnisd1 ~]$ xrd fndca1.fnal.gov
(C) 2004-2010 by the Xrootd group. Xrootd version: v3.3.6
Welcome to the xrootd command line interface.
Type 'help' for a list of available commands.
root://fndca4a.fnal.gov:1094/> cat /pnfs/fnal.gov/usr/fermigrd/volatile/fermilab/litvinse/gums1.txt
#
/etc/fstab
Created by anaconda on Wed Sep 26 13:39:49 2012
#
Accessible filesystems, by reference, are maintained under '/dev/disk'
See man pages fstab(5), findfs(8), mount(8) and/or blkid(8) for more info
#
UUID=4f92564e-56b9-4628-8afe-a3cfcdcd0918a / ext3 defaults 1 1
UUID=a41c036b-5e11-4140-94ca-27f2dd8311cd /boot ext3 defaults 1 2
UUID=4aa8938b-99da-46af-9898-2b6ee0bd11d8 /data ext3 defaults 1 2
UUID=3586d5bc-e689-4579-bd9d-53d97d2405c1 /home ext3 defaults 1 2
UUID=90a46e33-f043-44a6-b2d2-e98e6d74e33a swap swap defaults 0 0
tmpfs /dev/shm tmpfs defaults 0 0
devpts /dev/pts devpts gid=5,mode=620 0 0
sysfs /sys sysfs defaults 0 0
proc /proc proc defaults 0 0
cat returned 0

root://fndca4a.fnal.gov:1094/>
```

```
rmrmdir <dirname> Removes a directory.
prepare <filename> <options> <priority> Stage a file in.
query <reqcode> <parms> Obtain server information.
queryspace <logicalname> Obtain space information.
truncate <filename> <length> Truncate a file.

For further information, please read the xrootd protocol documentation.
```

```
root://fndca4a.fnal.gov:1094/> █ FIFE workshop
```

# dCache best practices

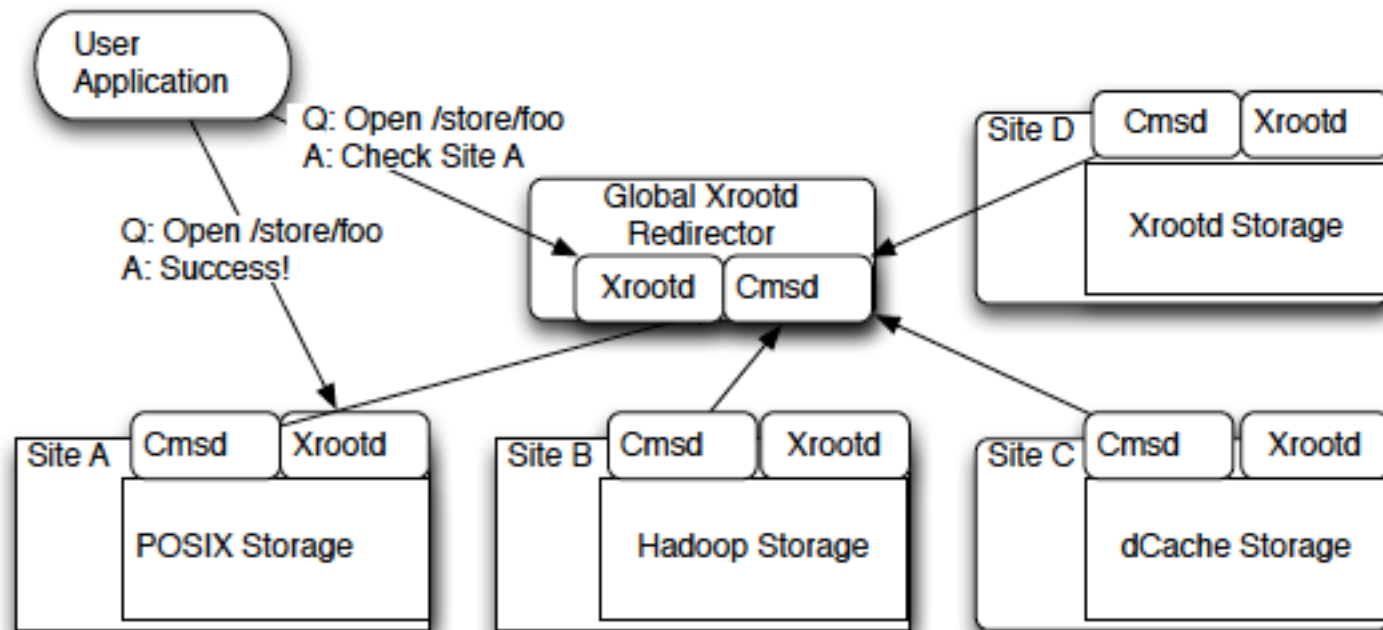
- For temporary files use scratch area  
/pnfs/<experiment>/scratch/users/<username>
- For permanent files:
  - Use large files (few GBs). Concatenate small files before putting them to dCache if possible
  - Use SFA (Small File Aggregation) with small files
- Pre-stage data before running on the grid



# Storage Federation

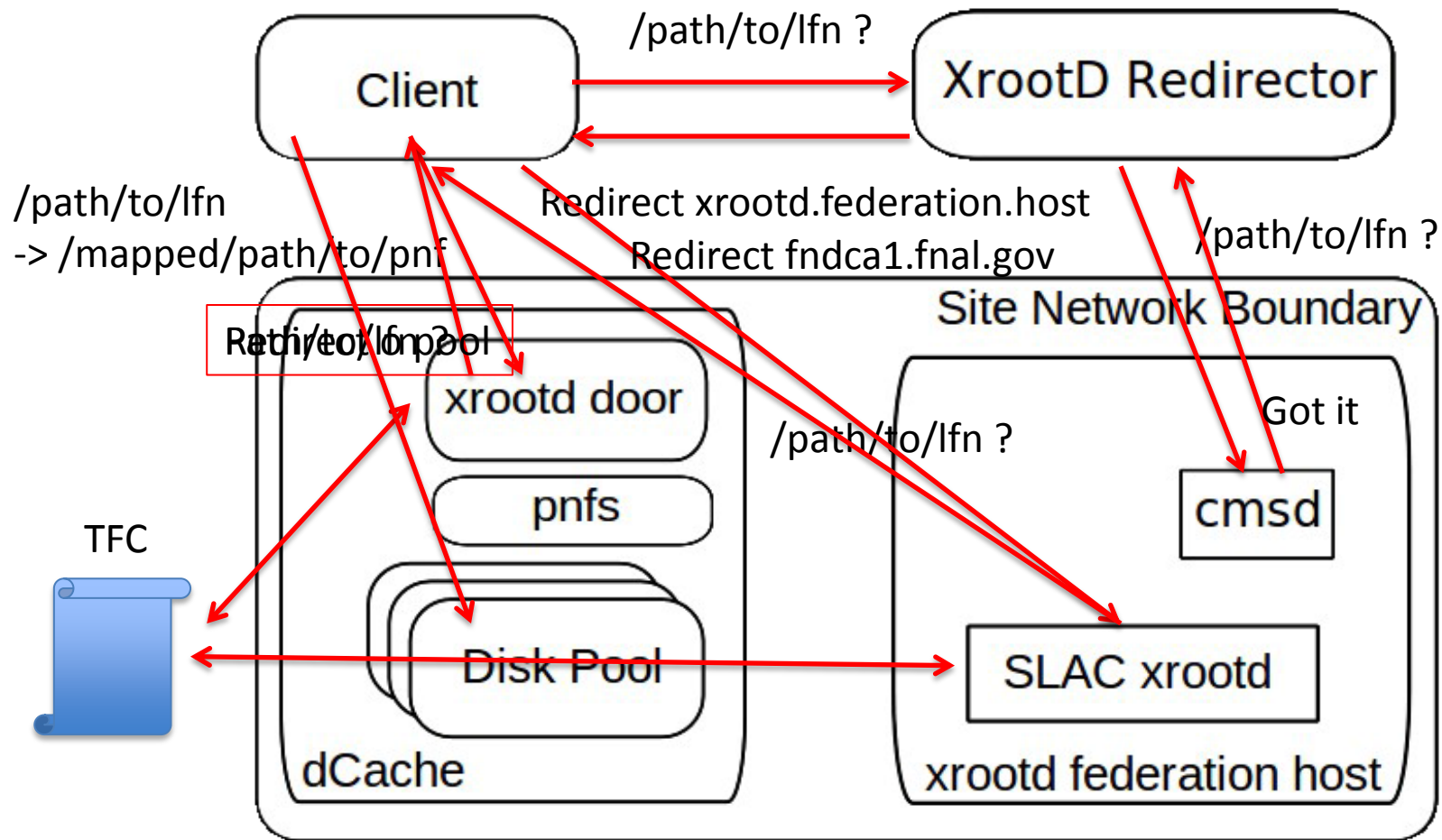
- Collection of heterogeneous storage resources managed by co-operating but independent administrative domains transparently accessible via a common namespace.

# xRootD Federated Storage (AAA take)



An XrootD based Framework built on top of existing storage

# dCache + XrootD Federation



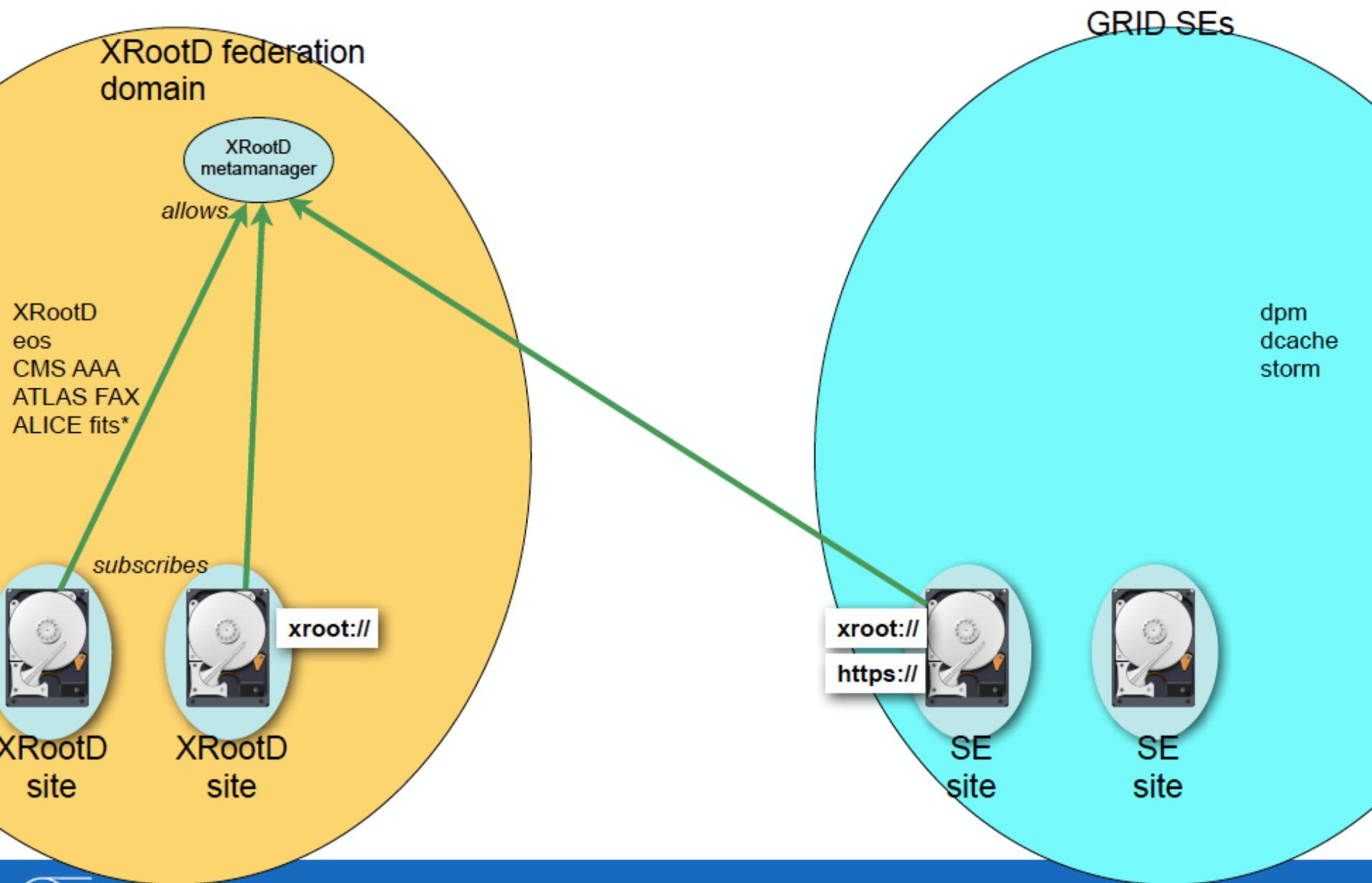
# HTTP Plugin for XRootD

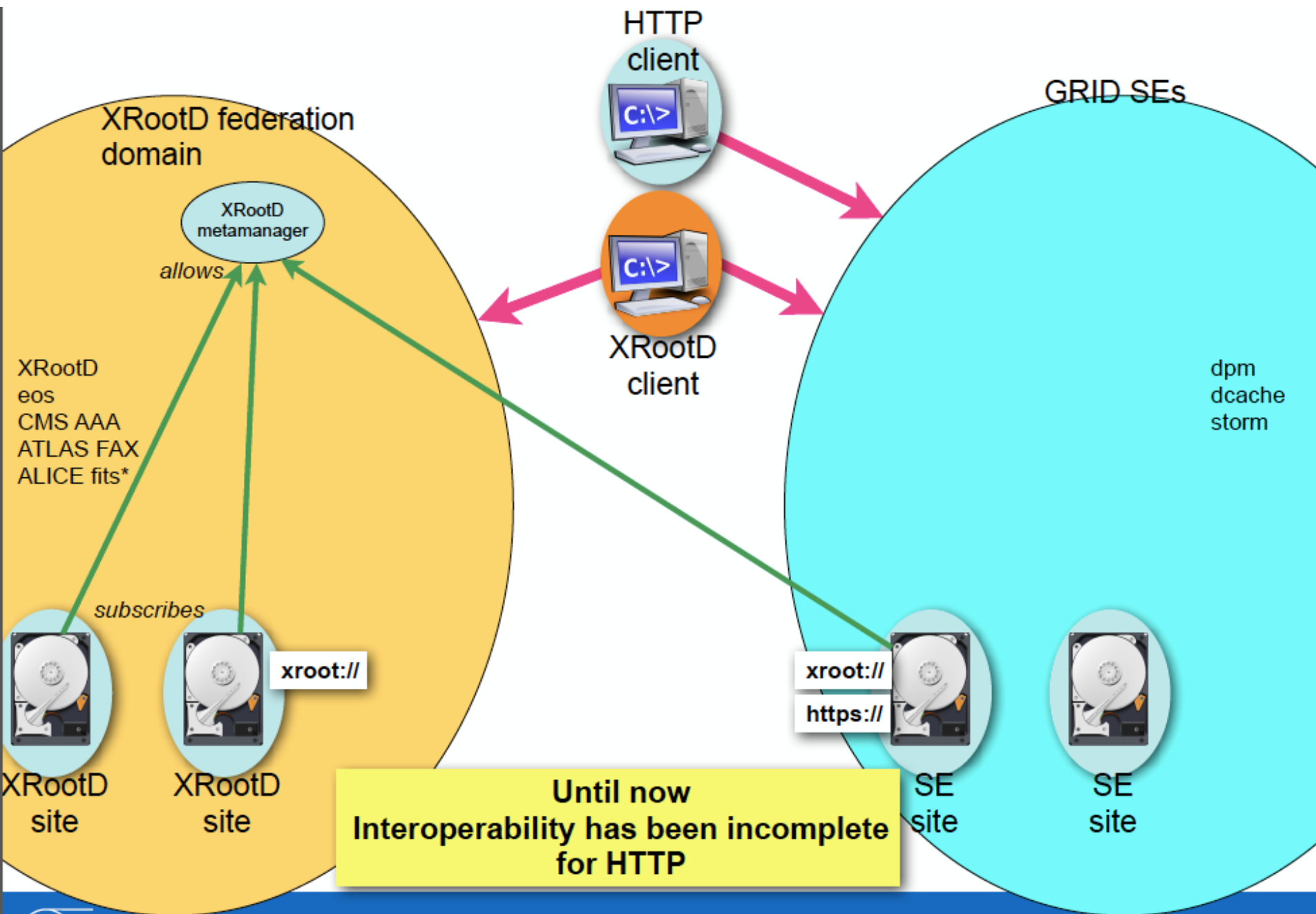
## (Fabrizio Furano CERN IT-SDC/ID)

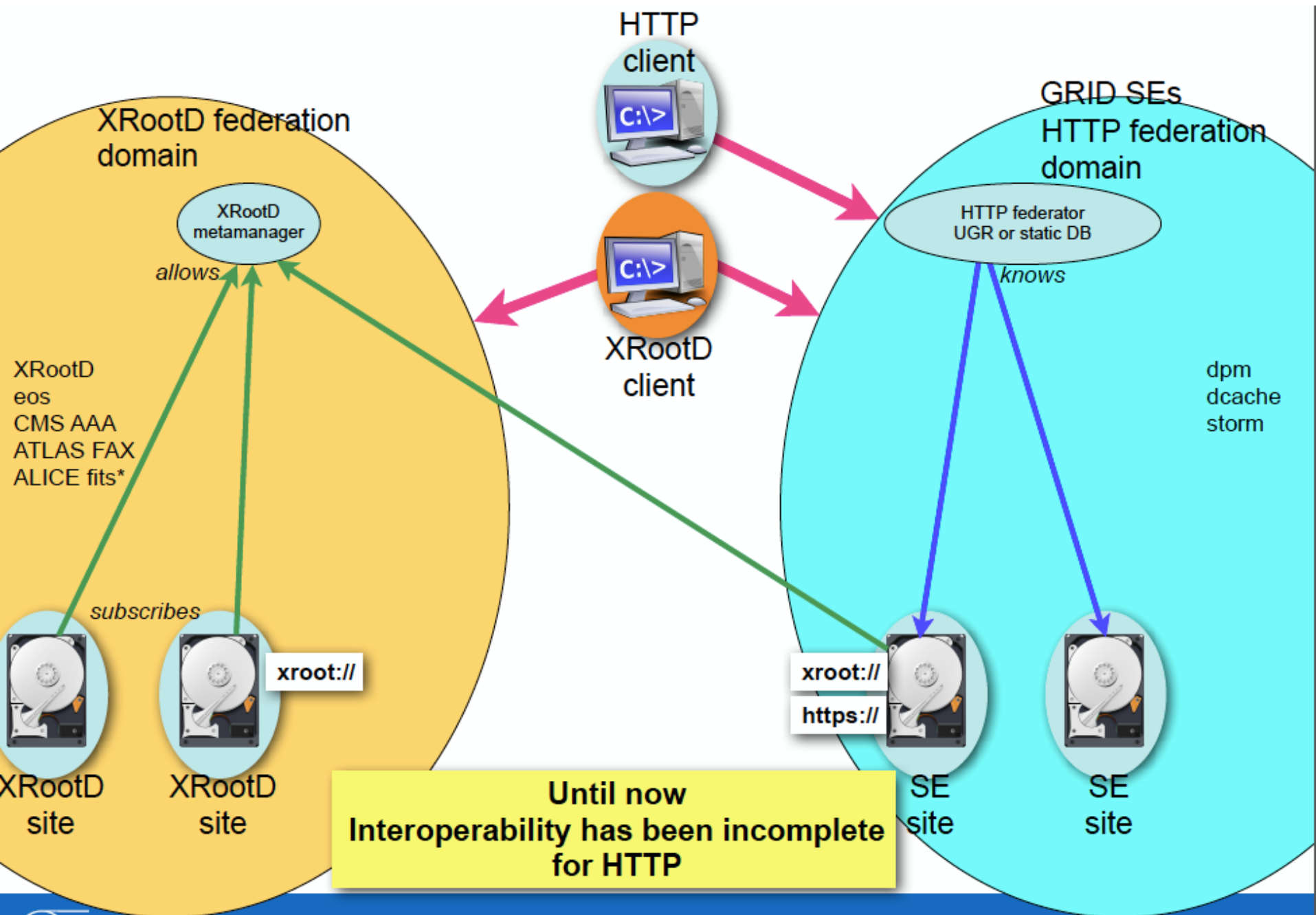
- XrdHTTP plugin provides support for HTTP(s) and WebDAV in XRootD framework
- Can share same port as XRootD protocol. Clients automatically detected.
- Support X509 auth., proxy certs. and VOMS extensions
- Any XRootD server keeps its advances functionality plus gets HTTP compliance
- IPv6 compatible

# Why HTTP/WebDAV?

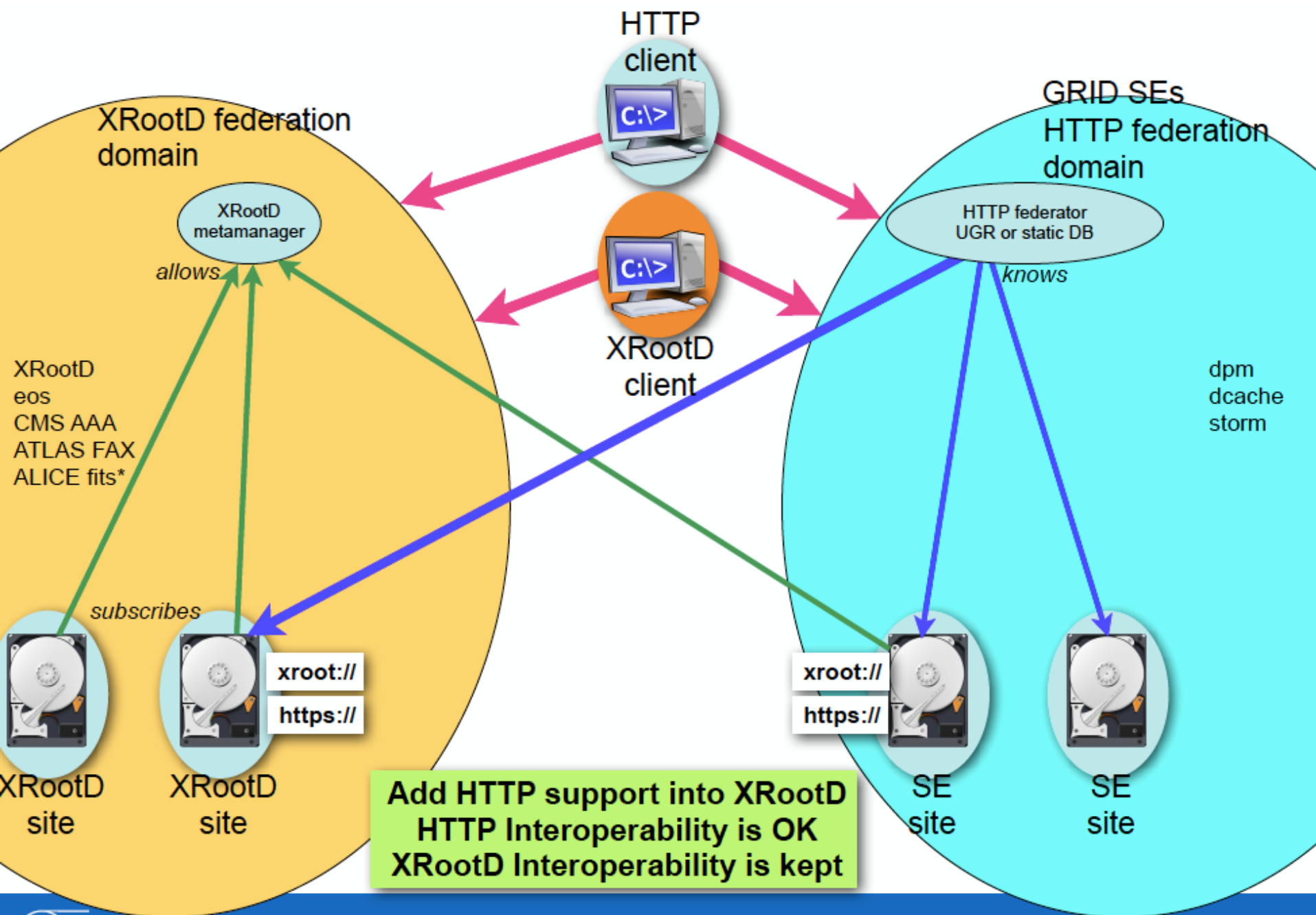
- Covers most existing use cases
- Supports WAN access
- GET/PUT or direct chunked access
- Ubiquitous:
  - Available on all platforms
  - Many standard clients
  - Web browsers
  - Allow integration with Cloud at protocol level
- HTTP moves more data worldwide than HEP
- HTTP + Grid allow a browser to interact with GRID SE











# Conclusion

- dCache is actively used by FIFE experiments
- We run XRootD door
- We are starting to experiment with storage federations
  - XRootD
  - WebDAV