dCache meets SARA

Patrick Fuhrmann

With contributions by
Gerd Behrmann
Tigran Mkrtchyan
Mattias Wadenstein

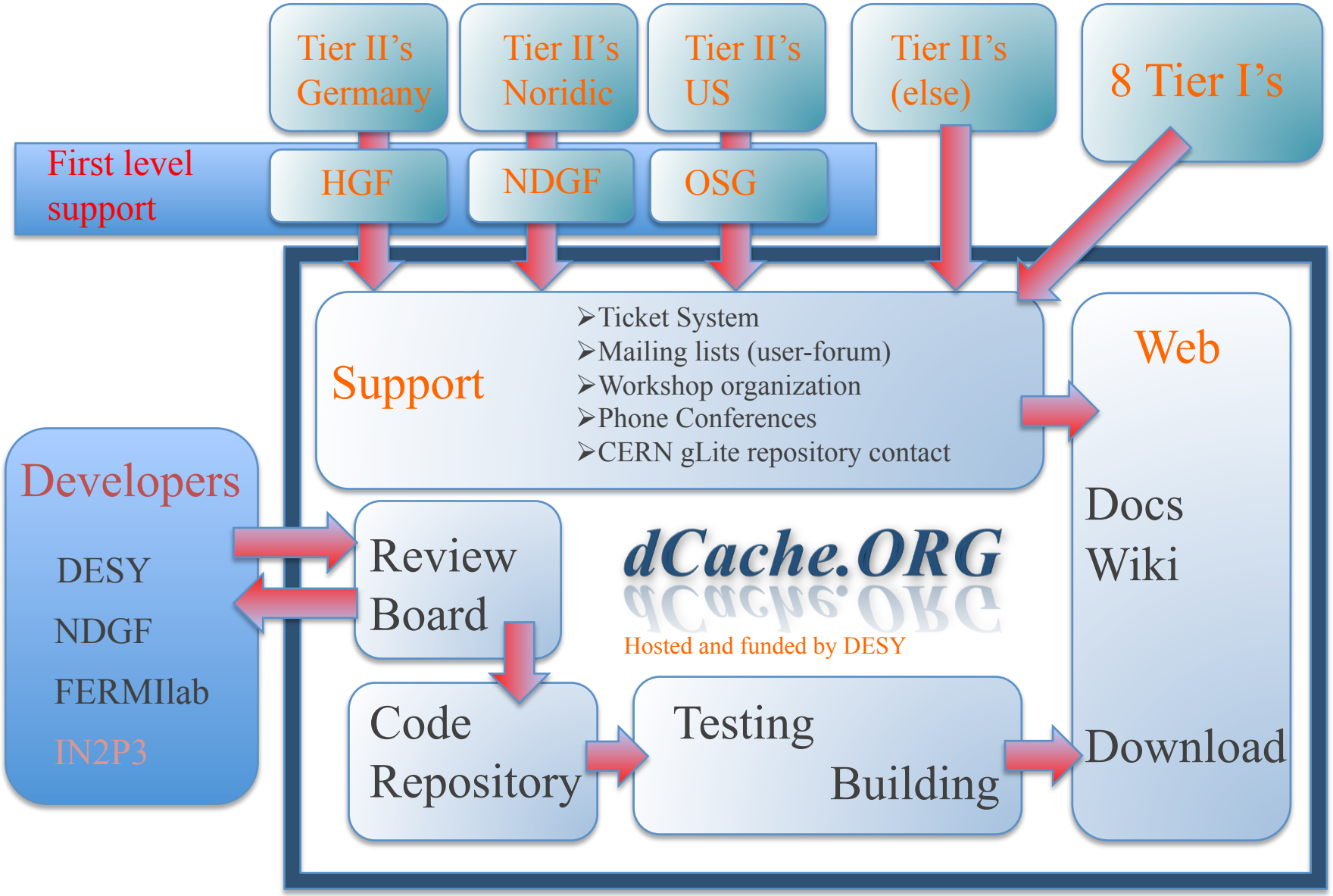
These slides are a result of a meeting at SARA, NL with
SARA team,
BioMed,
Long Term Storage and
LOFAR

Content

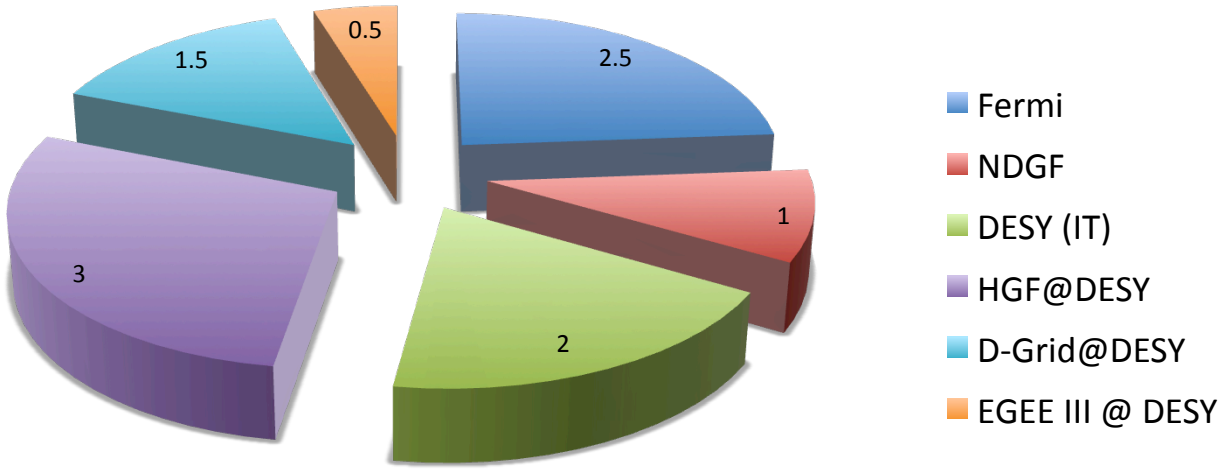
- ✧ The dCache organization
- ✧ dCache spec's
- ✧ dCache deployment
- ✧ dCache plans

The dCache Organization

What is dCache.ORG ?



dCache team by funding agency



About 10 team members in total.

dCache Specification

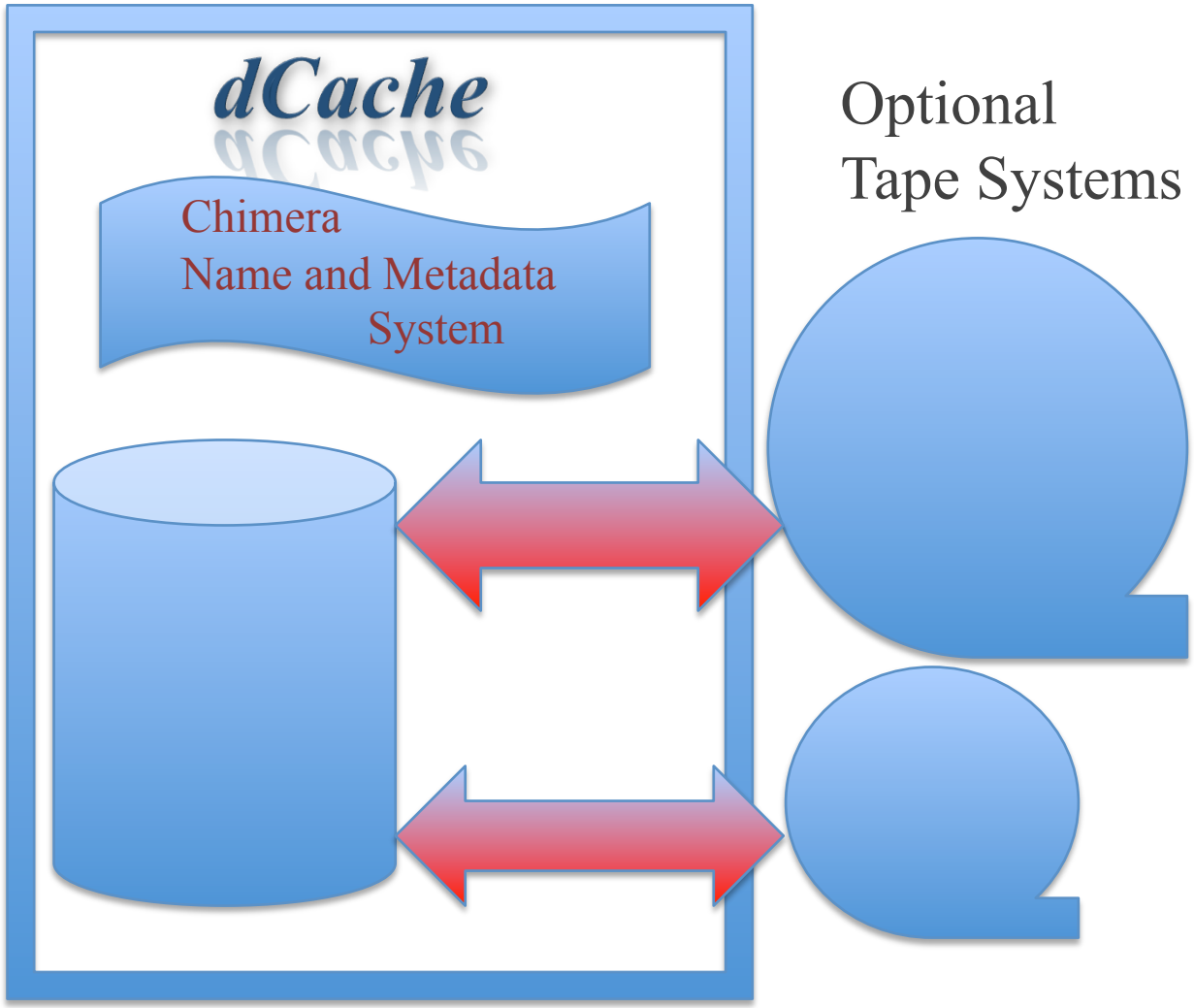
dCache BOX View

Storage Control
SRM

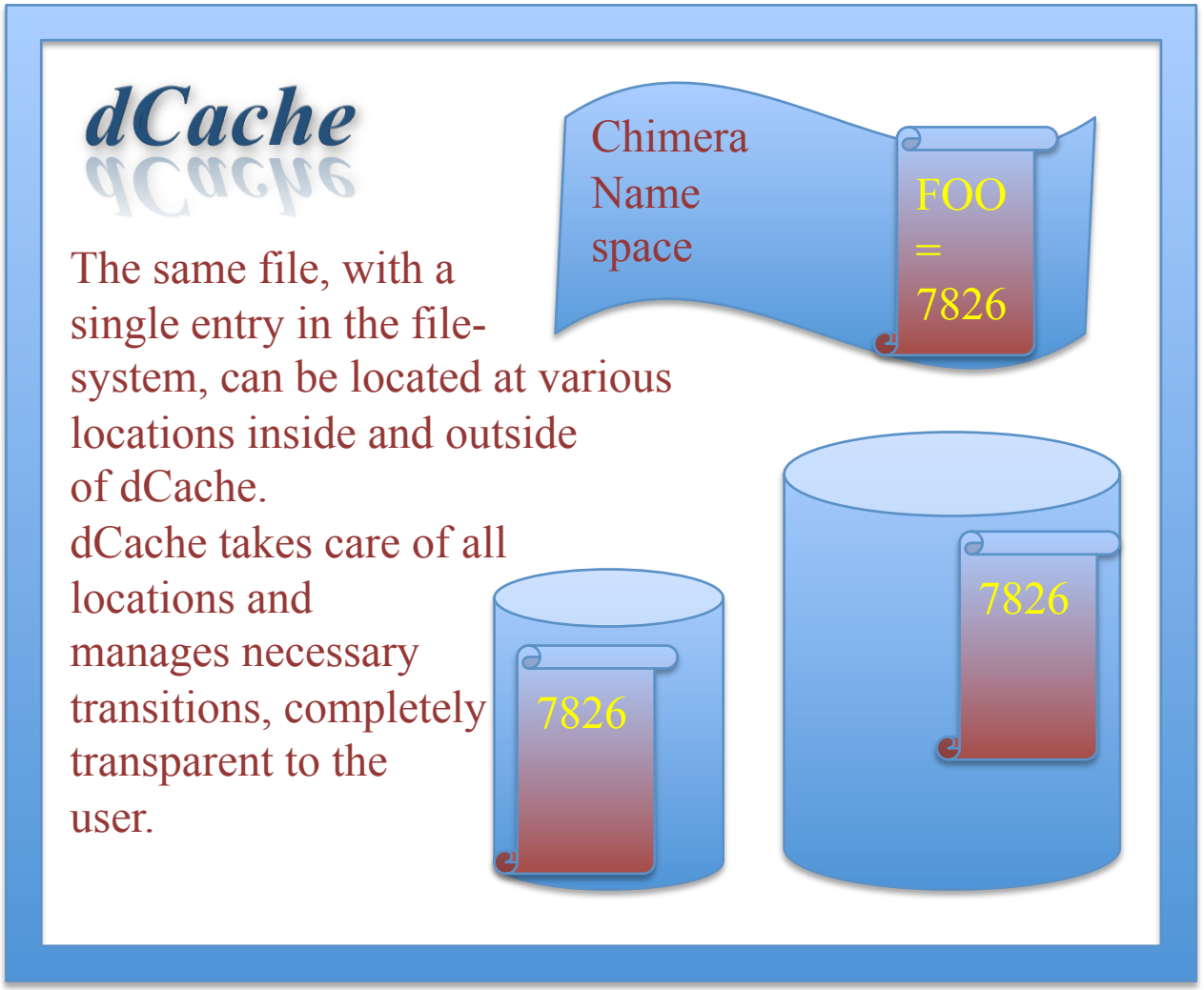
Wide AreaTransport
(gsi)Ftp
http(s) / WebDav

Posix LIKE Access
(gsi)dCap
xRoot

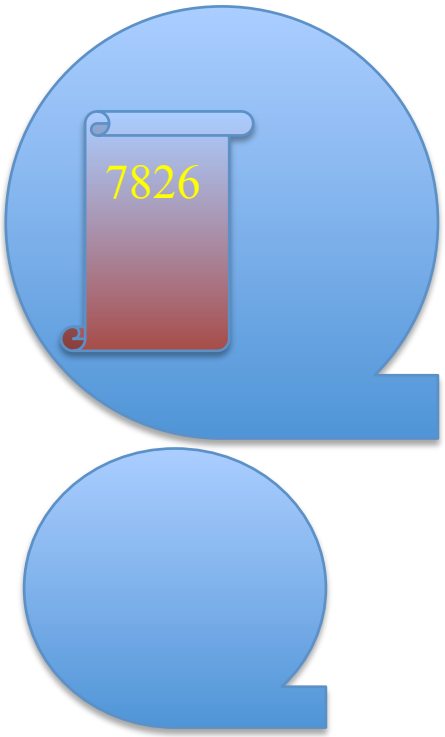
Posix native Access
NFS 4.1



dCache Idea



Optional
Tape Systems

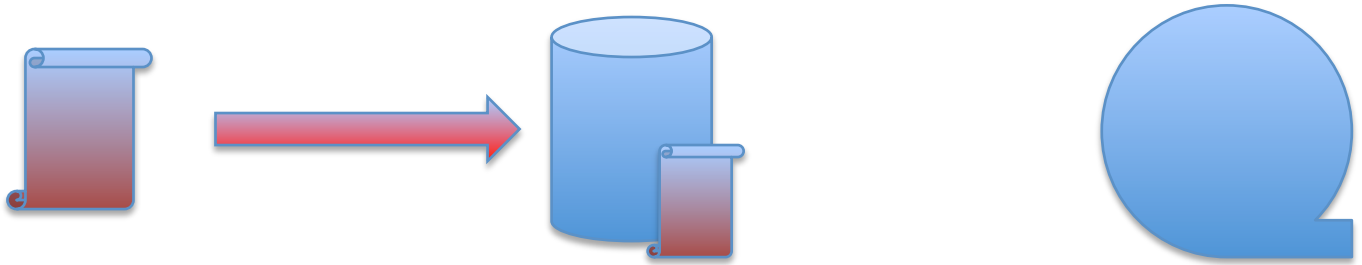


The consequence

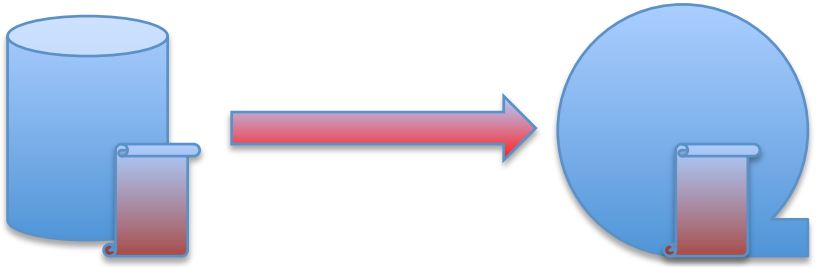
- ✧ Data is automatically replicated on detection of access hotspots.
- ✧ Data can be replicated on arrival. (second copy prior to tape backup)
- ✧ Data is migrated to tape if configured and restored if necessary.
- ✧ Data can be scheduled for replication for maintenance operations.
- ✧ Configuration can enforce a second or third copy of each file.

Basic file life cycle (all protocols)

File written to dCache

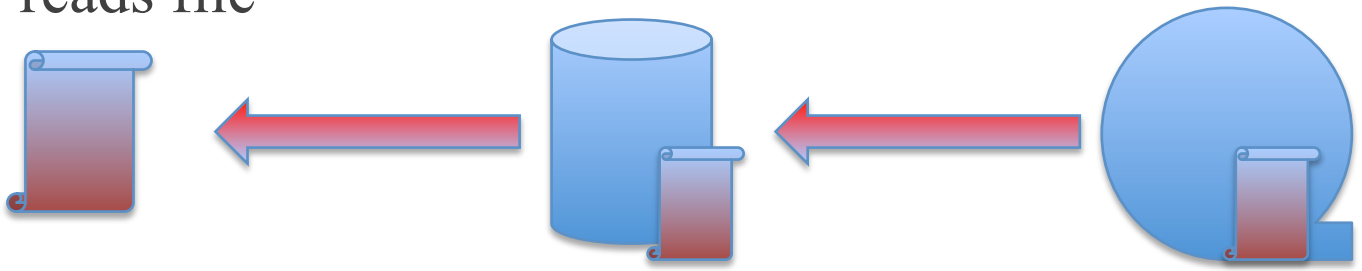


After awhile
(file is flushed to tape)

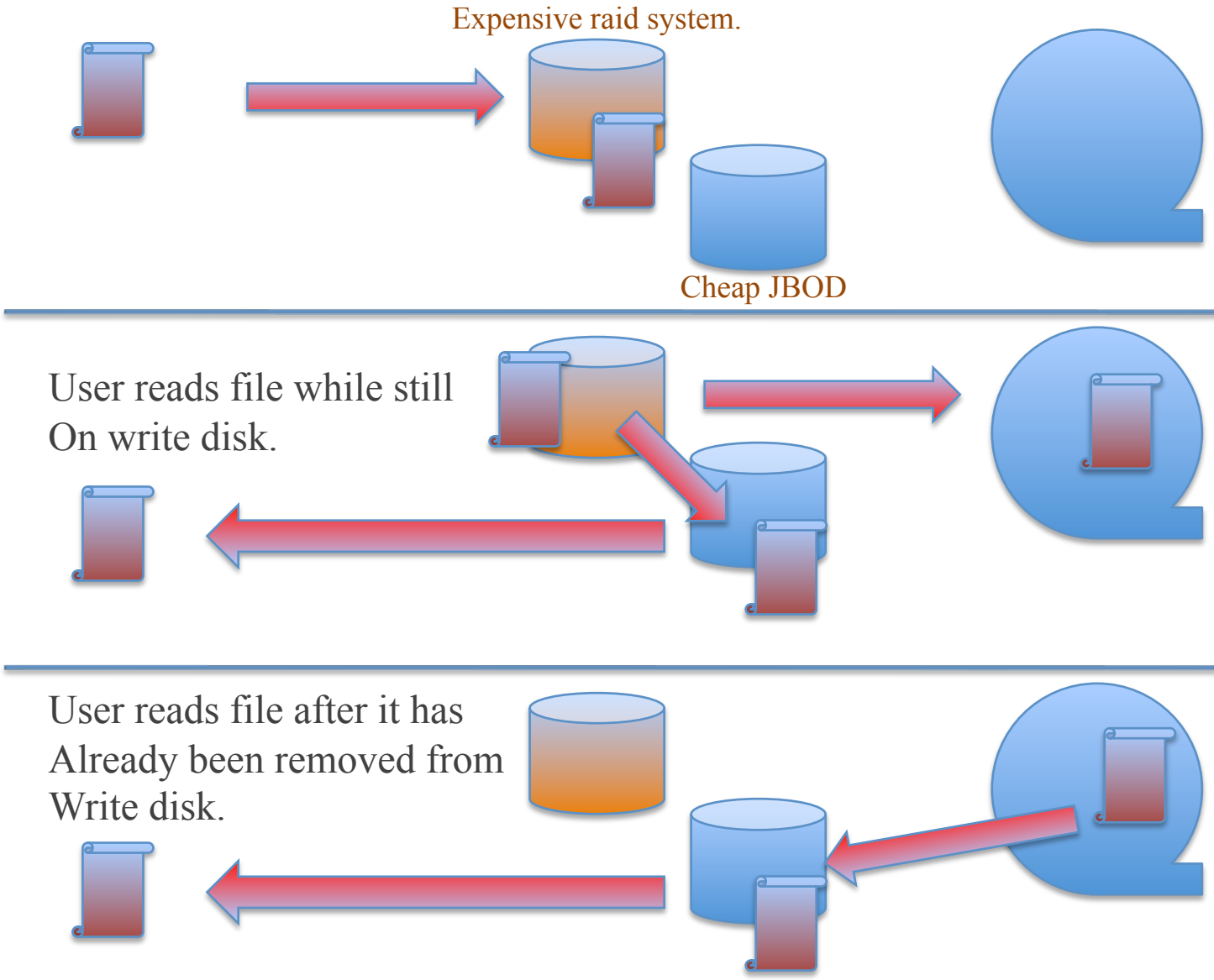


Space is running short
(File is removed from disk)

User reads file

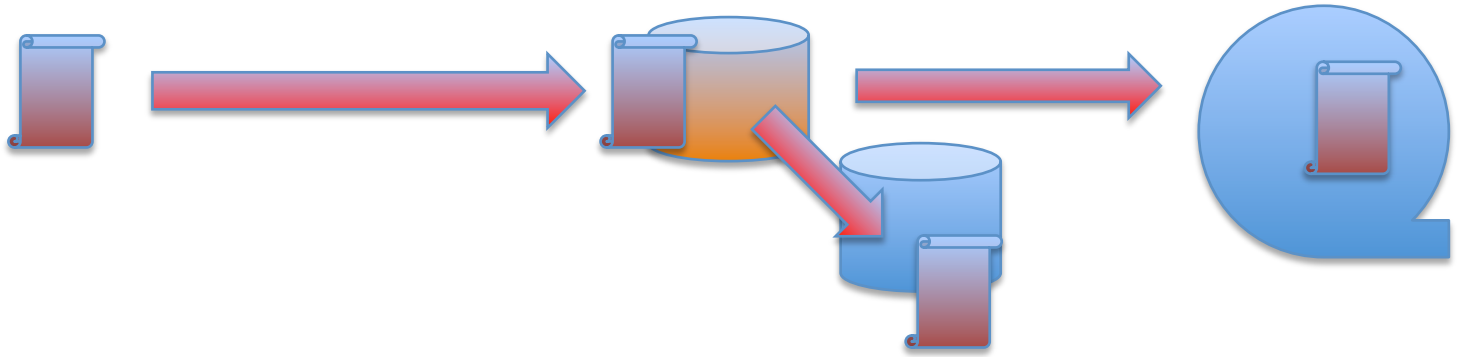


Basic file life cycle (technical view)



Reliability

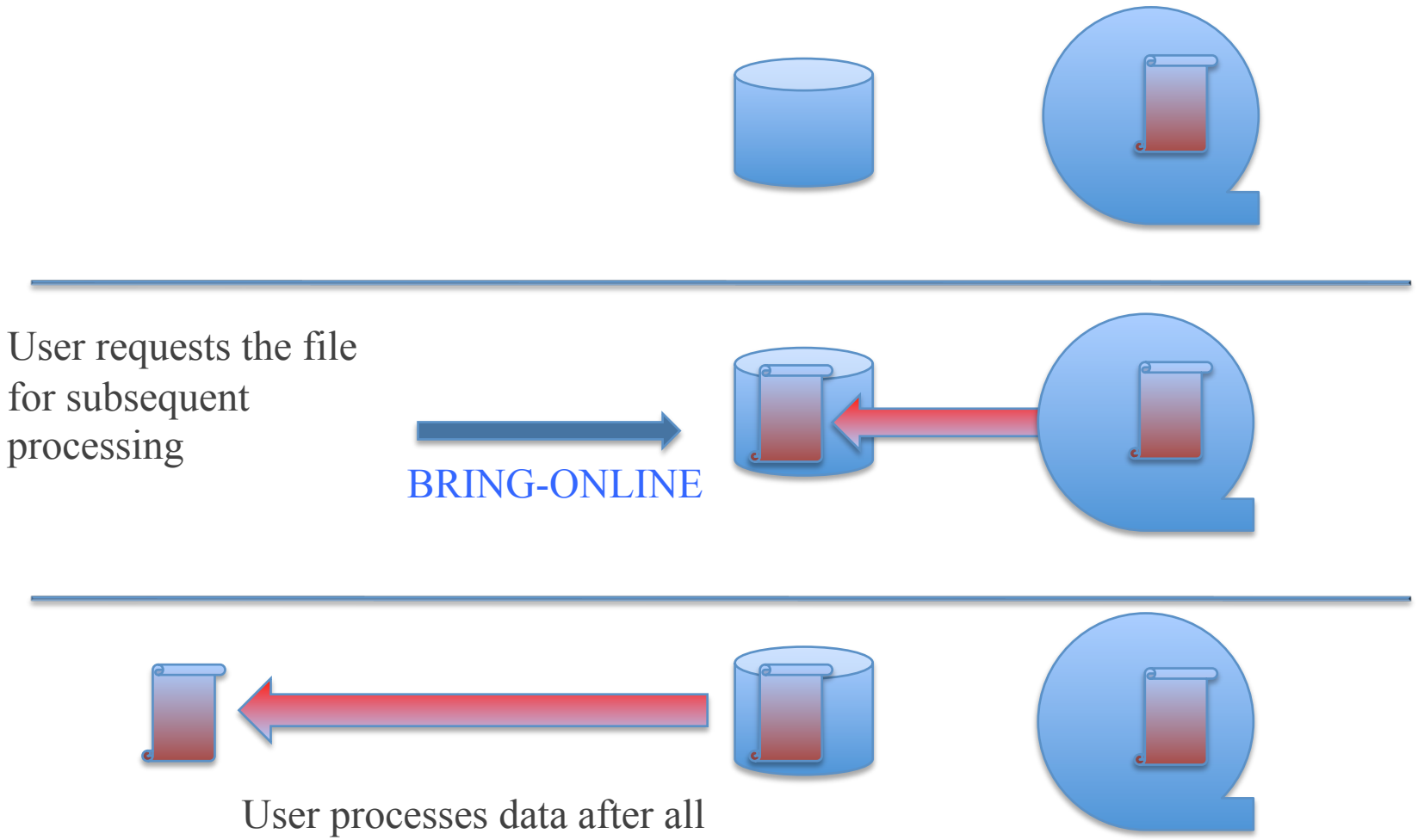
Checksums are calculated on all transfers (except for reading)



What is storage control ?

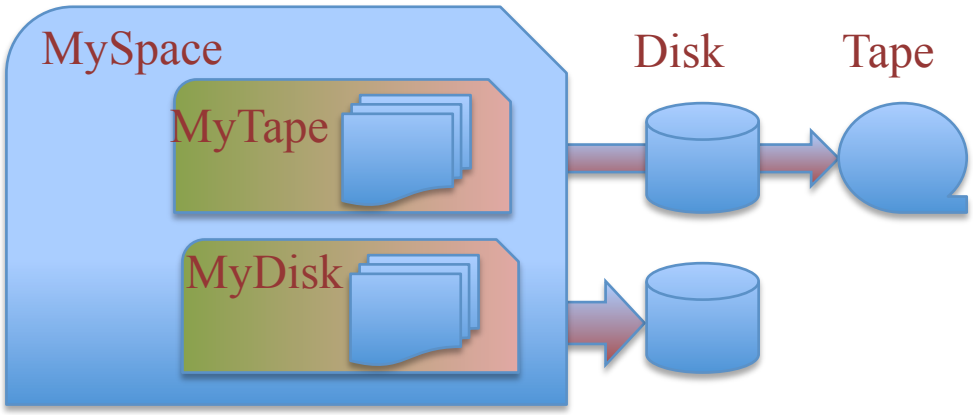
- ✧ dCache supports both : manual and automatic storage control
- ✧ Data is directed to pool-groups based on directory, client IP, protocol …
- ✧ Data can be directed to disk-only or disk-tape (Storage attributes)
    - ✧ Directory based storage attributes for all protocols
    - ✧ File based attributes for SRM only (Storage Resource Manager)
- ✧ Files can be pinned to disk (forever or for a fixed time) using SRM.
- ✧ Files can be restored to disk to schedule subsequent access.
- ✧ Automatic restore (tape -> disk) can be protected to avoid tape disaster.

Basic file life cycle and storage control (User)



Another example for User-Storage-Control

User may specify whether a file should end up on tape or on disk only.

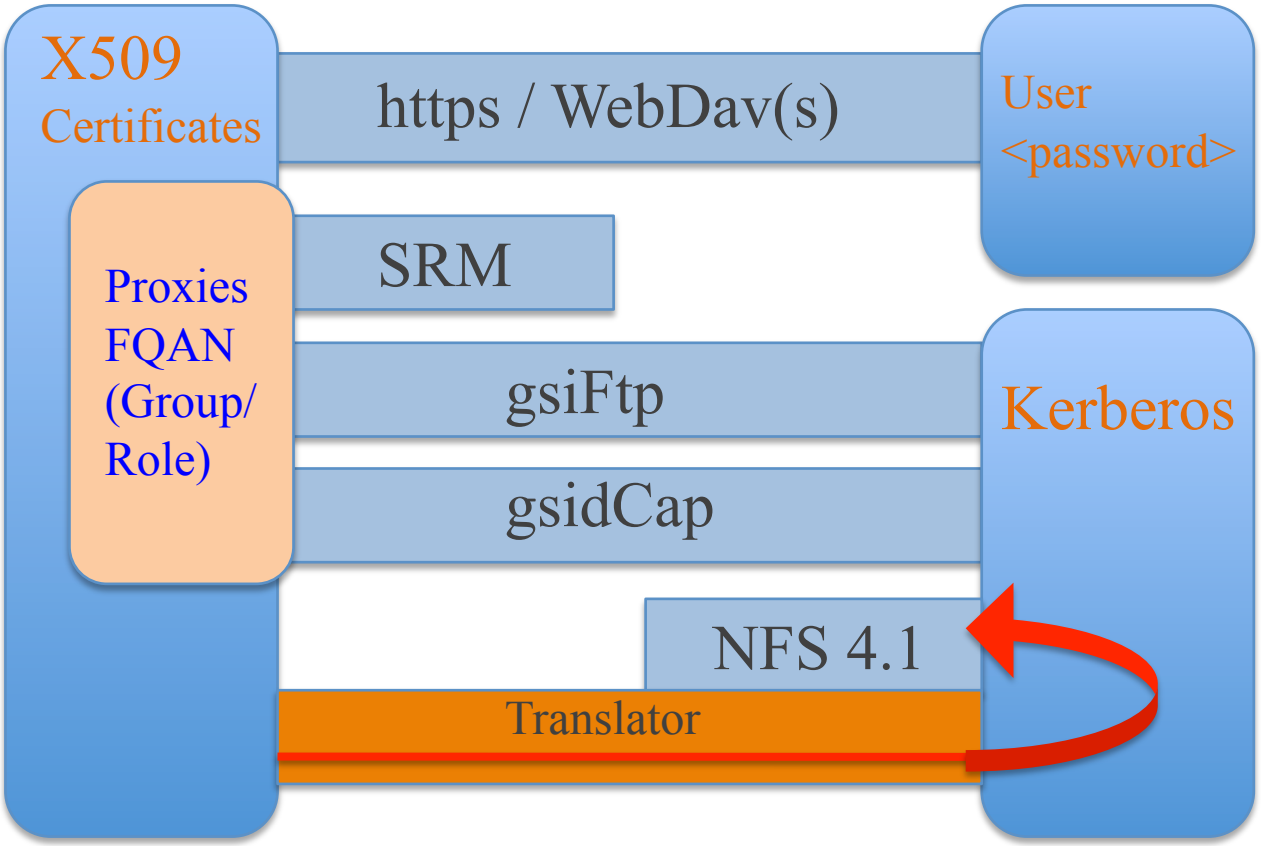


/users/x/MySpace/MyTape/Foo

/users/x/MySpace/MyDisk/Foo

Security

Authentication



Security

Authorization

File system, all protocols : full NFS 4.1 ACLs

Tape Protection : simple FQAN/DN based

Space tokens : indirect through file system and link groups

The dCache Customers

dCache is in production at :

WLCG (Europe plus OSG)

- 5 Tier I's in Europe
- 3 Tier I's in North America
- 40 Tier II's worldwide

HEP

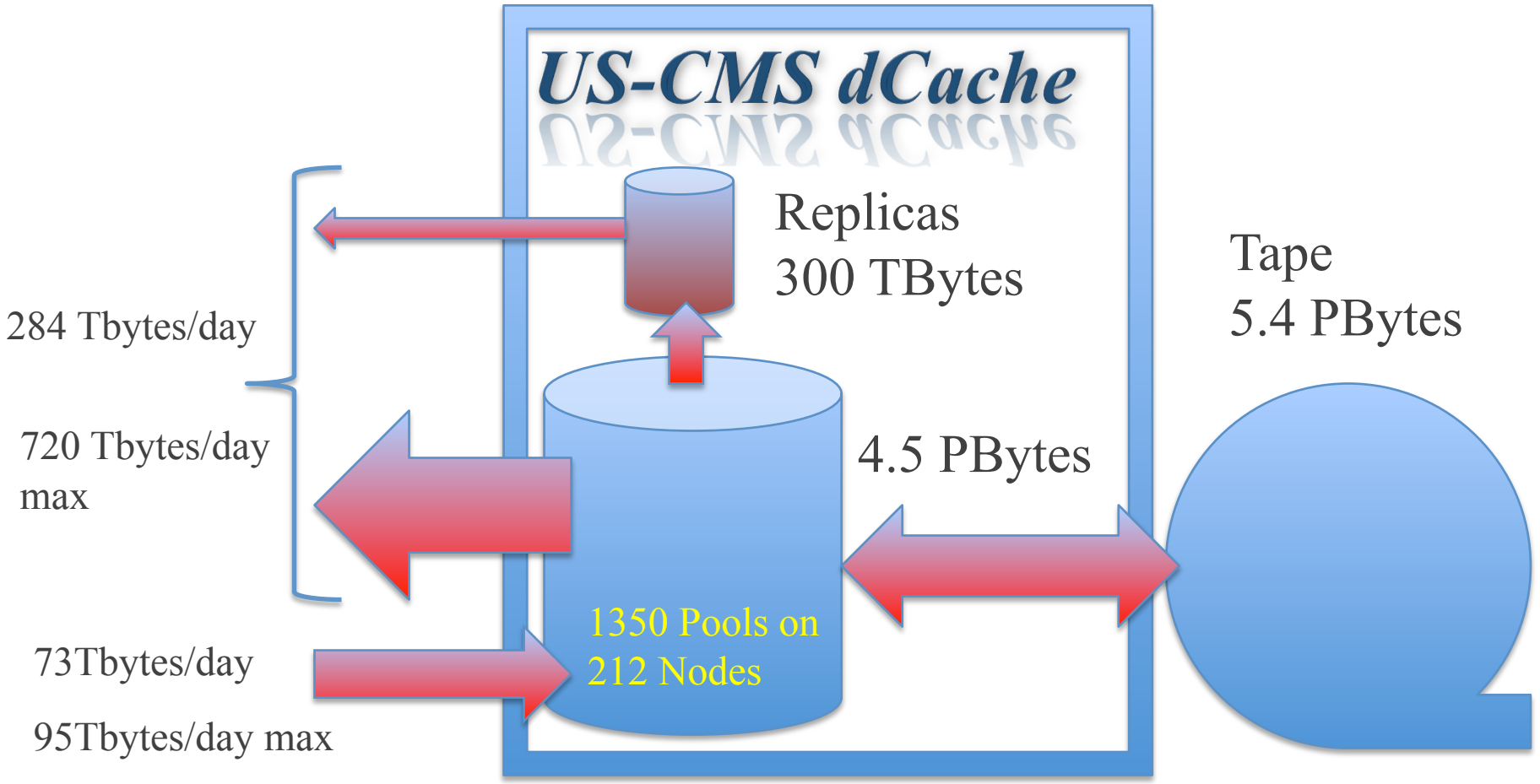
- Hera Tier 0
- ILC

Other communities

- Bio Med (NDGF)
- Photon Science (DESY)

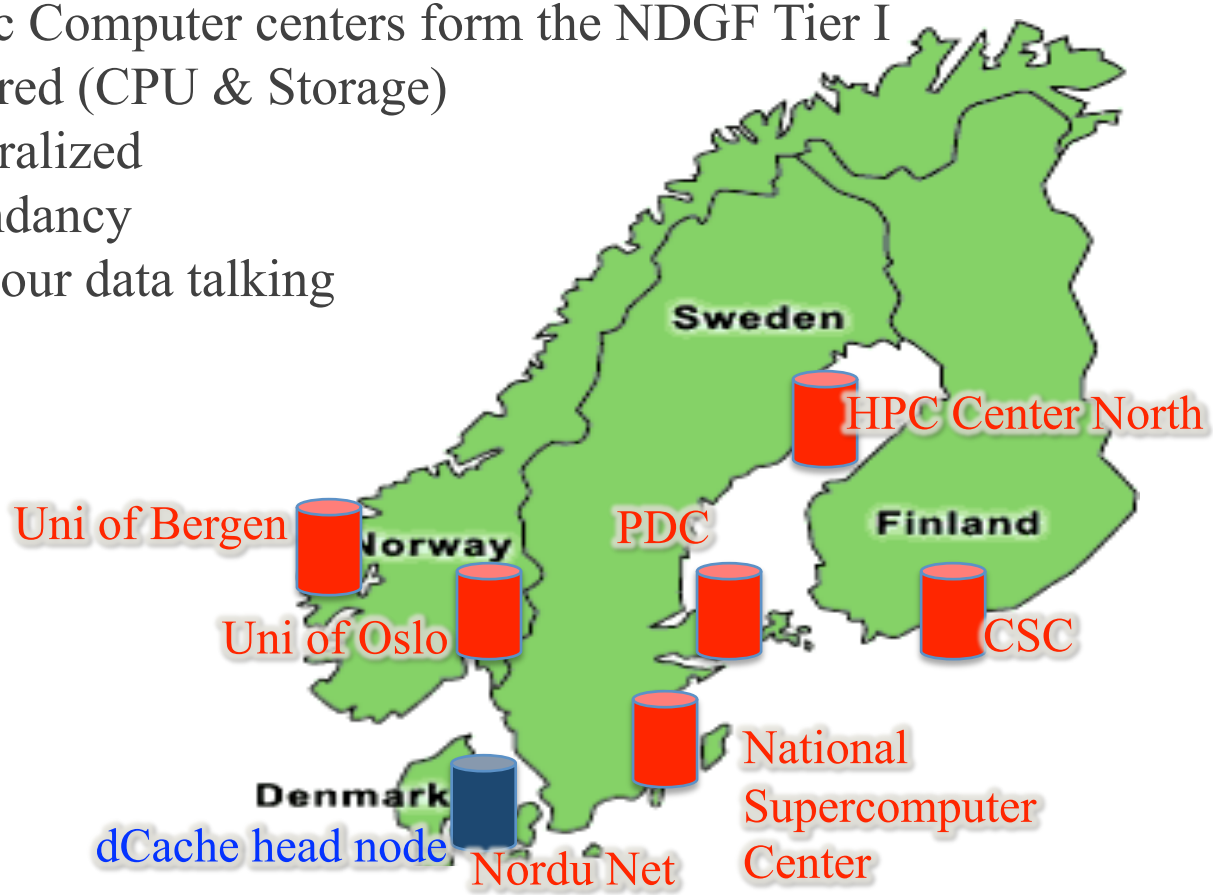
The largest dCache (as far as I know)

(Information provided by Jon Bakken, FEMILab)



The most complex dCache (for sure)

- ✓The 7 biggest Nordic Computer centers form the NDGF Tier I
- ✓Resources are scattered (CPU & Storage)
- ✓Services can be centralized
- ✓Advantages in redundancy
- ✓Especially in 7*24 hour data talking



Slide stolen from Mattias Wadenstein, NDGF

Further roadmap (Sysadmin only)

- Integrated monitoring
  - Information provided in xml format
  - Already done for all GLUE values.
- Simplified component location configuration
  - Single file replaces node/pool config
  - Easy parameter setting per domain/host

Further roadmap (Sysadmin & User)

- Unifying of 'User Representation' (May workshop)
  - File system, tape protection and space tokens will us the same user representation.
- Improved data distribution on bulk transfers
    - Already done for pool to pool transfer
    - Next for write into dCache
- Moving from manual to automatic redistribution of data

Further roadmap (User)

- https : User/Password authentication
- https : support of Proxy/FQAN/Groups/Roles
- ACL's : setting ACLs by user and not only sysadmin
- NFS 4.1 : secure (Kerberos, Certs by modified KDC)

Further roadmap : Going standard

- Already supported standards :
  - gsiFtp (IETF)
  - SRM (OGF)
  - Unsecure http (IETF)
- In beta testing
  - NFS 4.1
  - WebDav (s)

Further roadmap : NFS 4.1

Why not already NFS 2/3 for data access ?

dCache uses NFS 2/3 for name space operations (ls,mv..) only, as it doesn't support data of a single instance being distributed among different storage hosts.
NFS 4.1 (with parallel NFS) is the first standard posix access protocol allowing this.

Who is supporting NFS 4.1 (pNFS)

All major vendors :
EMC, IBM, Linux, NetApp, Panasas, Solaris server.
Coming soon : Windows client.

Further roadmap : NFS 4.1 (pNFS) in dCache

- Name server and i/o protocol fully implemented.
- No security yet
  - Soon : Kerberos.
  - X509 unlikely : Solution : modified KDC
- No automatic recall from tape to protect tape system.
  - Soon : part of the standard tape protection mech.
- Full support of NFS Access Control List (ACLs)
  - Right now only by system administrator
  - Soon : through NFS4 ‘setacl’ call by all users.
    - (NFS4 is already part of SL5 dist)
- Fully supports storage control (tape/disk) on directory bases.

Roadmap : NFS 4.1 (pNFS) linux clients

- NFS 4.1 and the linux kernel
  - NFS 4 already in SL5
  - NFS 4.1 in 2.6.32
  - NFS 4.1 plus pNFS in 2.6.34
- Kernel 2.6.34 will be in Fedora 13 and RH6 Enterprise (summer)
- Windows Client expected 4Q10.
- We are testing with :
  - SL5 and 2.6.34 plus some special RPM. (mount tools)
  - See our wiki for further information

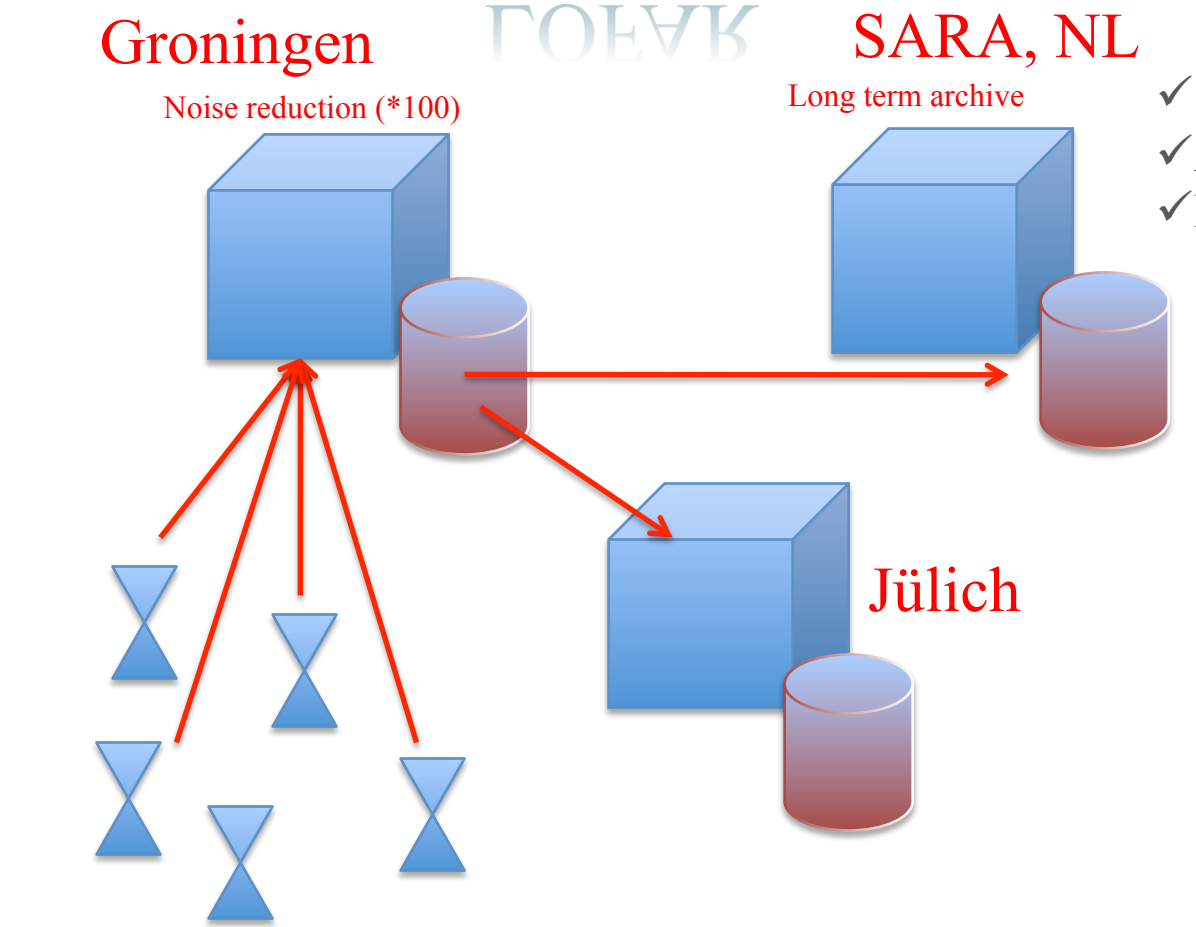
Roadmap : WebDav (s)

- Requested by
  - Bio Grid and other communities at NDGF
  - Light sources (Petra3 and XFEL) at DESY
- Beta release in 1.9.6 (3)
- Tested with Max OS, Windows(XP), SuSE11.2 (Gnome, KDE)
- Supports read and write
- Write via ‘redirect’ or if not supported by client via ‘proxy’.
- Security
  - Plain or x509
  - On redirect, only control line is encrypted.

Further Reading

www.dCache.org

LOFAR



- ✓ 1.5 P-Bytes first year on tape
- ✓About 20% on disk.
- ✓Restage unknown.

- ✓6 Key Science Projects
- ✓5 centrally coordinated
- ✓1 is individual user access.