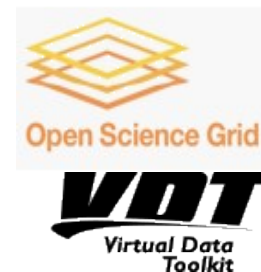




dCache, an update

Patrick
for the dCache Team

support and funding by





Content

dCache.ORG

Project Topology

- The Team
- The Partners
- The Goal
- The Version Timeline

In a nutshell

- Big Picture
- Basic Feature Set
- New Features in 1.7.0
- New Features in 1.8.0

dCache.ORG

Selected Topics

- Managed File Hopping
- The NDGF Challenge
- The xRoot Protocol

Future

- Protocol wise NFS 4.1 (pNfs)



Project Topology

The Team

The Partners

The Goal

The Version Timeline



Project Topology : The Team

dCache.ORG

dCache.ORG

Head of dCache.ORG

Patrick Fuhrmann

Head of Development FNAL :

Timur Perelmutov

Head of Development DESY :

Tigran Mkrtchyan

Core Team (Desy and Fermi)

Andrew Baranovski
Bjoern Boettscher
Ted Hesselroth
Alex Kulyavtsev

External

Development

Gerd Behrmann, NDGF
Abhishek Singh Rana, SDSC
Jonathan Schaeffer, IN2P3



Iryna Koslova
Dmitri Litvintsev
David Melkumyan

Support and Help

Greig Cowan, gridPP
Stijn De Weirdt (Quattor)
Maarten Lithmaath, CERN
Flavia Donno, CERN

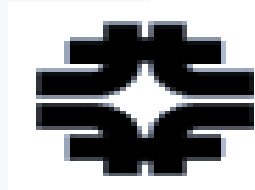
Dirk Pleiter
Martin Radicke
Owen Syngé
Neha Sharma
Vladimir Podstavkov



Project Topology : The Partners

dCache.ORG

dCache.ORG



Code contribution

beside DESY, FERMI

NDGF : ftp (protocol V2)

IN2P3 : HoppingManager



Integration. Verification

- CERN
- Open Science Grid
- d-Grid





Project Topology : The Goal

Because with the start of LHC, dCache will manage the largest share of LHC data, we will concentrate on

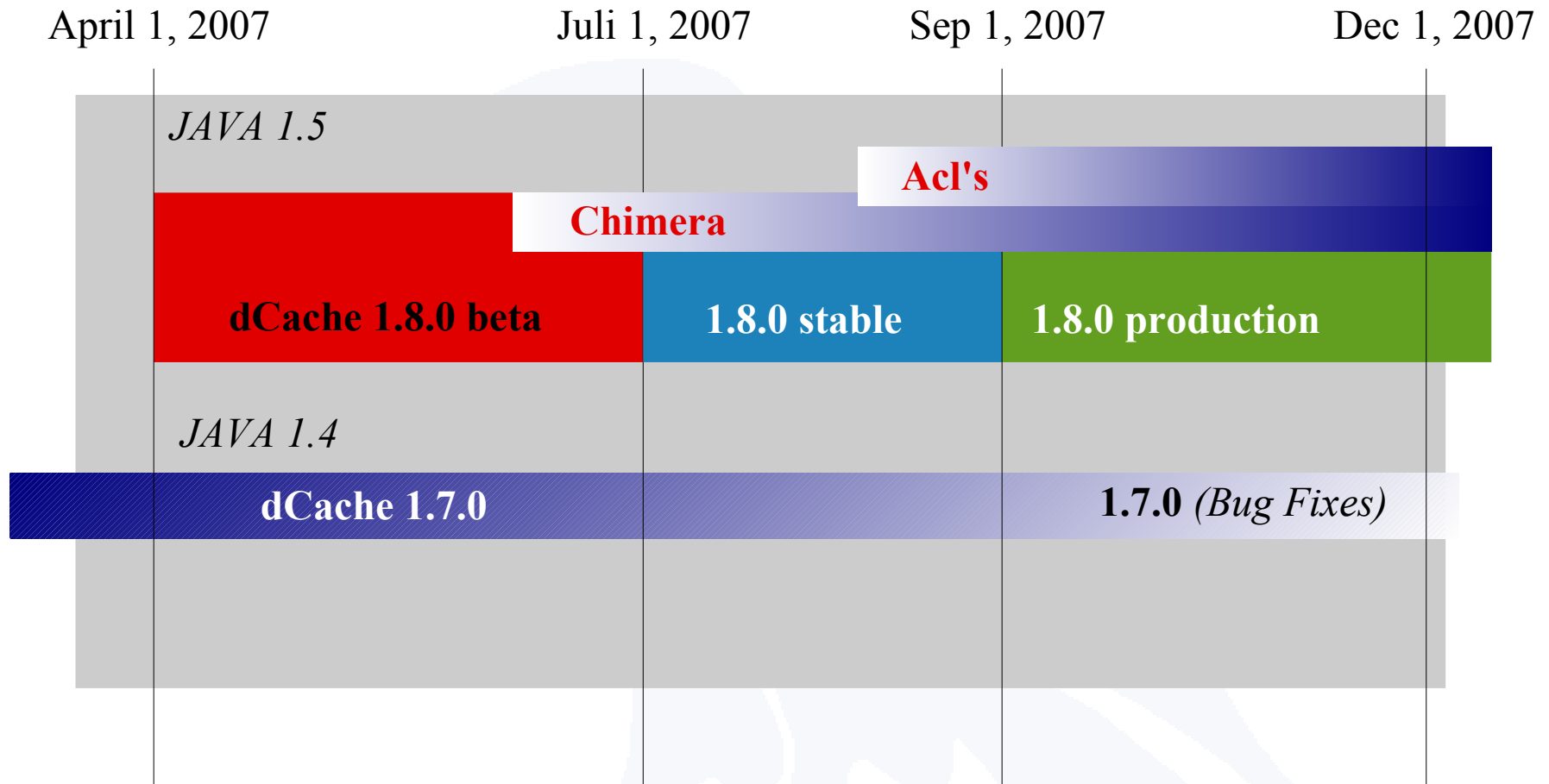
- Stability
- Simplified installation (Yes Miron, it has already significantly improved)
- Documentation (needs lot more)
- Support (maybe special support for Tier I's)



Project Topology : Versions Timeline

dCache.ORG

dCache.ORG





In a nutshell

Managed Storage

Basic Feature Set

New Features in 1.7.0

New Features in 1.8.0





- Strict name space and data storage separation, allowing
 - *consistent name space operations (mv, rm, mkdir e.t.c)*
 - *consistent access control per directory resp. file*
 - *managing multiple internal and external copies of the same file*
 - *convenient name space management by nfs (or http)*
- Automated file replication on access hot spot detection
- HSM connectivity (enstore,osm,tsm,hpss, dmf)
- Automated HSM migration and restore (optimizing HSM operations).
- Handles data in Peta-byte range on 1000's of pools
- Supported protocols : (gsi)ftp , (gsi)dCap, xRoot, SRM, nfs2/3
- Separate I/O queues per protocol
- Supports resilient dataset management (worker-node support)
- Sophisticated command line interface and graphical interface



- dCache partitioning for very large installations
- File hopping on
 - automated hot spot detection
 - configuration (read only, write only, stage only pools)
 - on arrival (configurable)
- gPlazma (authentication, authorization, GUMS connectivity)
- Passive dCap
- xRoot support (with *Alice* authorization)
- Central HSM FLUSH manager
- Maintenance module (draining pools)
- improved GUI
- Jpython interface for all kind of configuration (e.g.used by quattor)
- Easy installation (Yaim and VDT)



- SRM 2.2 following WLCG agreement
 - Details : see Timurs talk
- xRoot protocol
 - vector read
 - currently working on async I/O
- Chimera (new namespace provider) included (optional)
- working on ACL's
- support of multiple, non overlapping HSM systems (NDGF approach)



Selected Topics

Controlled File Hopping

The NDGF Challenge

The xRoot protocol



Selected Topics

Controlled File Hopping

The NDGF Challenge

The xRoot protocol



Why file hopping and pool queues per protocol ?

- To improve tape system performance (keep streaming)
- Overcome disk deficiencies, read versus write access
- Balance low bandwidth versus high bandwidth applications
- With dCap we can even distinguish between applications.
- Protect disk systems from overload
- Overcome firewall issues



Selected Topics

Controlled File Hopping

dCache.ORG

dCache.ORG

Back-end Tape Storage
OSM, Enstore, Tsm, Hpss, DMF

Queue(s)
p2p dCp xRoot

dCap, xRoot
(local access)

Queue(s)
p2p gsiFtp

gsiFtp
(Tier II download)





Selected Topics

Controlled File Hopping

The NDGF Challenge

The xRoot protocol



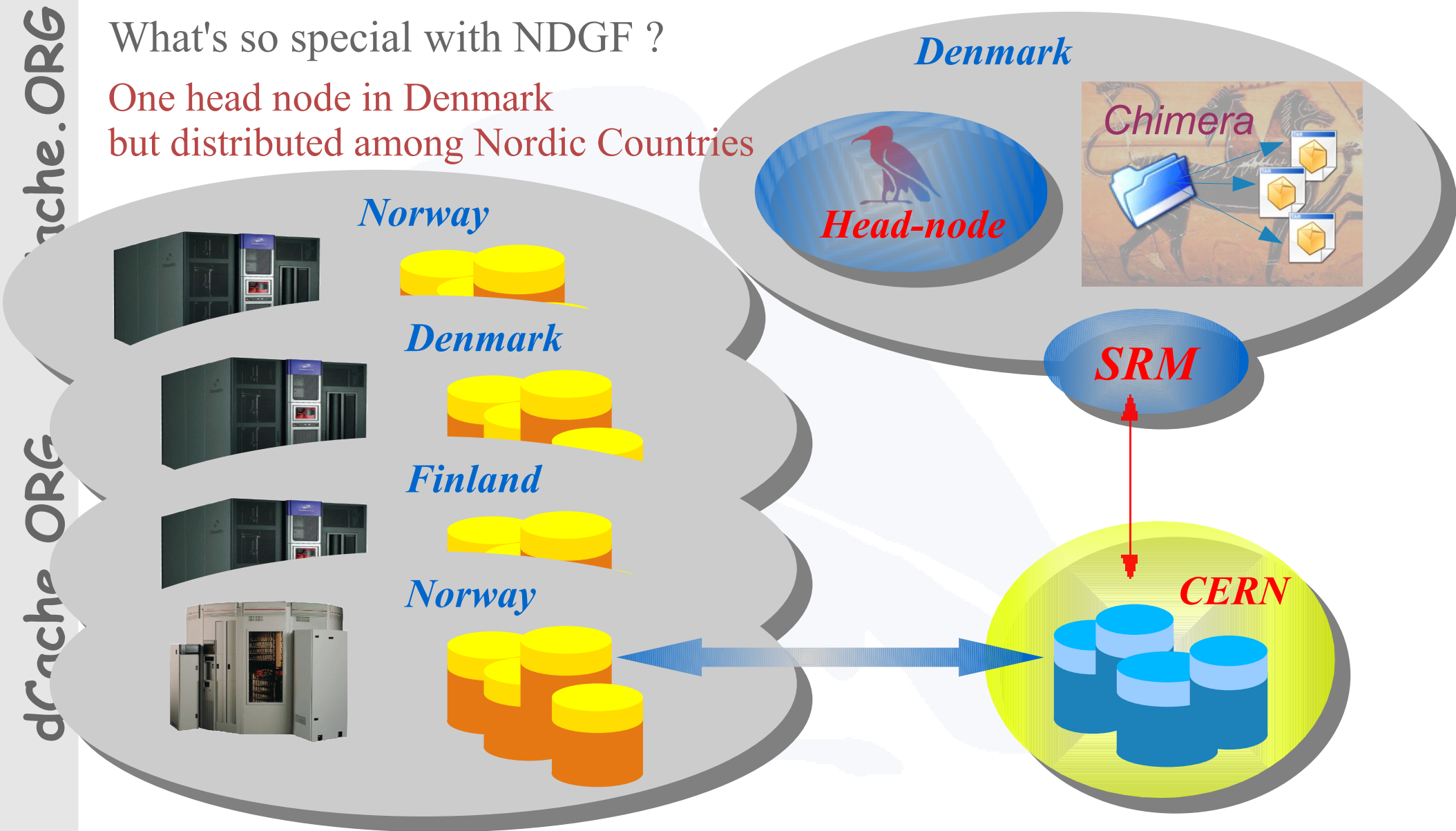
Selected Topics

The NDGF Challenge

dCache.ORG

What's so special with NDGF ?

One head node in Denmark
but distributed among Nordic Countries





What's needed ?

- ➔ gsiFtp Protocol Version II
- ➔ Different HSM systems in different countries
 - ➔ Pool are selected based on the secondary location of the data
- ➔ Secure internal cell communication
- ➔ Fine grained command authorization



Selected Topics

Controlled File Hopping

The NDGF Challenge

The xRoot protocol



Basic xRoot Protocol is implemented in dCache. We are not using the original server.

Integrated as any other protocol, so :

- Makes use of the *pool selection mechanism*
- Uses internal *cost mechanism*
- Allows real name-space operations
- Can make use of gPlazma authorization (except for Alice)

Progress

- Basic functions in 1.7.0
- Vector read in 1.8.0
- Following soon
 - Asynchronous I/O
 - Pre-stage request



First Results

“Johannes Elmsheuser”, LMU, Munich, ATLAS

- *Comparison of dCache-xRoot and dCache-dCap*
- *Same results if dCap “read ahead” set to high numbers*
- *Remark : TDCap Root driver file is not well supported by dCache.ORG*

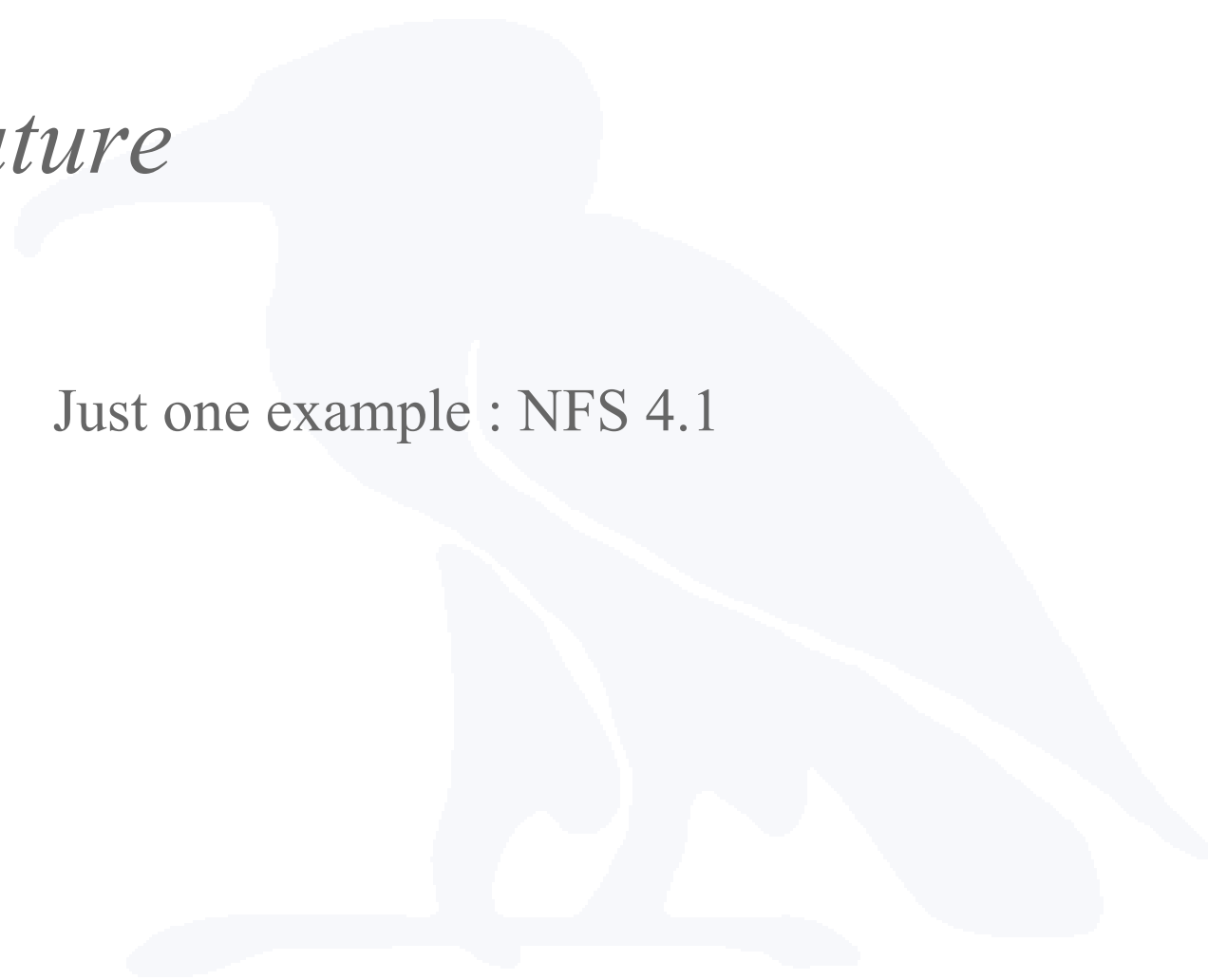
“Sergey Panitkin”, BNL, ATLAS

- *Comparison of dCache-xRoot native xRoot (both non secure)*
- *Loose cuts. AOD on dCache : 910 events / min*
- *Loose cuts. AOD on xRootd : 1000 events / min*
- *Tight cuts. AOD on dCache : 62 events / min*
- *Tight cuts. AOD on xRootd : 259 events / min*



Future

Just one example : NFS 4.1





Future

Stolen from Tigrans talk :

We are currently putting significant efforts in the NFS 4.1 protocol

Deployment Advantages :

Clients are coming for free ...

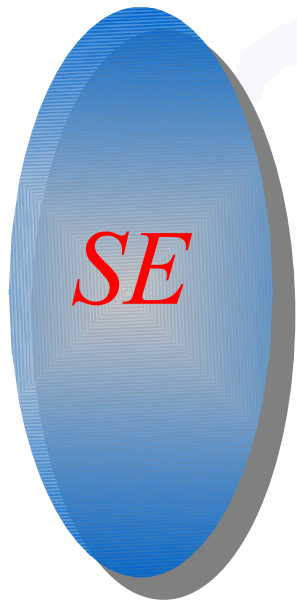
Technical Advantages :

- ◆ NFS 4.1 (pNFS) design perfectly matches the dCache design
- ◆ Faster (optimized) e.g.:
 - Compound RPC calls
 - 'Stat' produces 3 RPC calls in v3 but only one in v4
- ◆ GSS authentication
 - Built in mandatory security on file system level
- ◆ ACL's
- ◆ OPEN / CLOSE semantic (so can keep track on open files)
- ◆ 'DEAD' client discovery (by client to server pings)



Future (hopefully)

dCache.ORG
dCache.ORG



Information Protocol(s)

Storage Management
Protocol(s)
SRM 1.1 2.2

Data & Namespace
Protocols

rfio dCap
gsiFtp
xRoot http

Namespace ONLY
NFS 2 / 3

Finally



Information Protocol(s)

Storage Management
Protocol(s)
SRM 2.2 (3.0)

Data & Namespace
Protocols

NFS 4.1
http(s)



Further reading

www.dCache.ORG





Deployment and distribution

Automated testing procedure

Deployment process



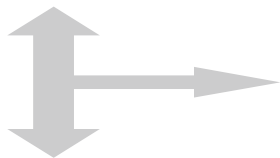


Automated testing process

dCache.ORG

dCache.ORG

CVS check-in



Full Compilation and RPM creation

CVS Tag



Full Compilation
RPM creation

Results on web page and e-mailed to developers

In Progress

RPM from developer repository



(Regression) Test suite



OK Web Site
ATP repository
(s13/4 ; 32/64 bit)



OK or Failed to developer

Test Suite is becoming a dCache.ORG product as well



Deployment and feedback Process

Feedback from user community

- *support @ dCache.org* for bug reports
- *user-forum @ dCache.org* for 'users helping users'

Deployment/Announcement of new versions resp. sub-versions

- * New subversions are announced at
user-forum and *announce @dcache.org*
(and RSS feed in the future)
- * and are published on the *dCache.ORG* web page
- * and are published in the '*stable*' APT repository
- * RPM will always have the corresponding 'change log'



Ongoing Development

dCache.ORG

dCache.ORG

SRM 2.2

Main features

Milestones

Status

SRM version interoperability issues

SRM evaluation deployment plan

Chimera



NFS 4.1



Storage Classes

Administrator determines 'retention policy' and 'access latency'

Retention policy REPLICATION, CUSTODIAL

Access Policy ONLINE, NEARLINE

Tape1-Disk0 : NEARLINE + CUSTODIAL

Tape1-Disk1 : ONLINE + CUSTODIAL

Tape0-Disk1 : ONLINE + REPLICATION

Storage Class Transitions foreseen (not high priority)

Space Tokens

To guarantee space for incoming transfers.

Later maybe for 'restores from tape' as well.



Jamie Shiers (WLCG)

Services are required for testing in Q2 (two) in preparation for the Dress Rehearsals in Q3 (and the LHC pilot run in Q4)...

- **1st April 2007** - target date for the needed services to be in place at the sites
- **1st June 2007** Ruth (OSG) wants to have SRM 2.2 stable
- **1st July 2007** - start date of Dress Rehearsals (also the date when the WLCG service is commissioned)

dCache

See subsequent slides




Basic WLCG MoU functionality 

Missing 0 out of 25

WLCG MoU functionality due end of 2007 

Missing 2 out of 4

Non MoU functionality 

Missing 6 out of 12

Extended use cases 

Missing 5 out of 40

Flavias stress test started just recently

Up to date information from Flavias 'test page'

<http://grid-deployment.web.cern.ch/grid-deployment/flavia/>



SRM version interoperability (details)

- The initial dCache version with SRM 2.2 included, is **dCache 1.8.0**.
- **dCache 1.8.0** and higher will support **SRM 1.1** and **SRM 2.2** at the same time on the same TCP Port.
- Both SRM protocol versions will run in the same dCache instance, using just one file system instance. (pnfs)
- Both SRM versions will have access to the **same file name space**.
- Files written with 1.1 can be accessed via 2.2 and vice versa.



SRM evaluation deployment plan (Agreement)

- Sites agreed to deploy dCache 1.8 (SRM2.2) in April :
 - FERMILab, DESY
 - BNL
 - gridKa
 - IN2P3
- For those sites we will closely watch the installation and the behavior.
- Systems will have 1-2 head nodes and ≥ 10 TBytes of disk storage.
- Systems will be connected to a Tape Back-end to support all possible storage classes.








SRM evaluation deployment plan (restrictions)

- Full upgrade to 1.8.0 is a prerequisite for the SRM 2.2 activation.
- There is no way to have dCache versions prior to 1.8 running with SRM 2.2
- The following restrictions apply concerning the agreed test systems :
 - It will be a special dCache evaluation instance, and **not part of the production system**.
 - The service is not part of the production monitoring and may be **shut down at any time**, without further notice.
 - All **data** should be regarded as '**not persistent**' and should be copied to the production system in order to become permanent.



SRM evaluation deployment plan (timing)

April *(guided and scheduled deployment)*

1. Week : FERMI – DESY transfers  
2. Week : Installation at BNL 
3. Week : Installation at gridKa 
4. Week : Installation at IN2P3 

Starting May *(regular deployment)*

RPM and Installation are already on dCache.ORG

- Still very good in time
- FERMI, DESY, BNL, gridKA already on Flavias pages
- IN2P3 will follow up this week



SRM evaluation deployment plan (timing)

Further steps depend on the success of the procedures described previously.

Just fair to say :

Although it's certainly our goal to be in production shape in July, we can't yet give advice on whether or not to use dCache SRM 2.2 during the Dress Rehearsal.



Coming Soon

dCache.ORG

dCache.ORG





Chimera





Chimera



Expected Improvements compared to PNFS

- Performance scales with back-end database implementation
 - Small to medium sites with mysql/postgres
 - Really huge sites with oracle cluster (planned for DESY)
- Enables protection against misuse
 - Different 'chimera users' (e.g. nfs, dCache, enstore) may get different doors with different priorities if back-end db allows.
- Simplifies maintenance resp. monitoring tasks
 - By using SQL database
 - Easy to add customized web interfaces.
- Allows ACL plug-ins
 - ACL sub-project started beginning of 2007 (DESY-Zeuthen)



Chimera (cont.)



Current status

- Functional and performance tests in progress
- Ready for testing by external sites : mid of march
- Setting up pnfs -> chimera (de-)migration scenarios
- Production time-line : depends on results of tests;
otherwise as fast as human resources allow.



NFS 4.1

Highlights

- Standardized interface to dCache name-space and data
- 4.1 extension makes use of highly distributed data
- Security (e.g. certificates) is part of spec.
- Clients are provided by OS maintainer(s)

citi.umich.edu is pushing to have the dCache server ready soon