dCache, news

HEPiX, Ann Arbor, Autumn 2013

Patrick Fuhrmann et al.

Content

- The project structure
  - Partners and people
  - Our funding
  - Sustainability/Networking
- Deployments
  - WLCG overall
  - News
- Customers Relation
  - Deployment Channels
  - User Support channels
- Work in progress
  - For WLCG
  - For Photon Science
  - Cloud software and service

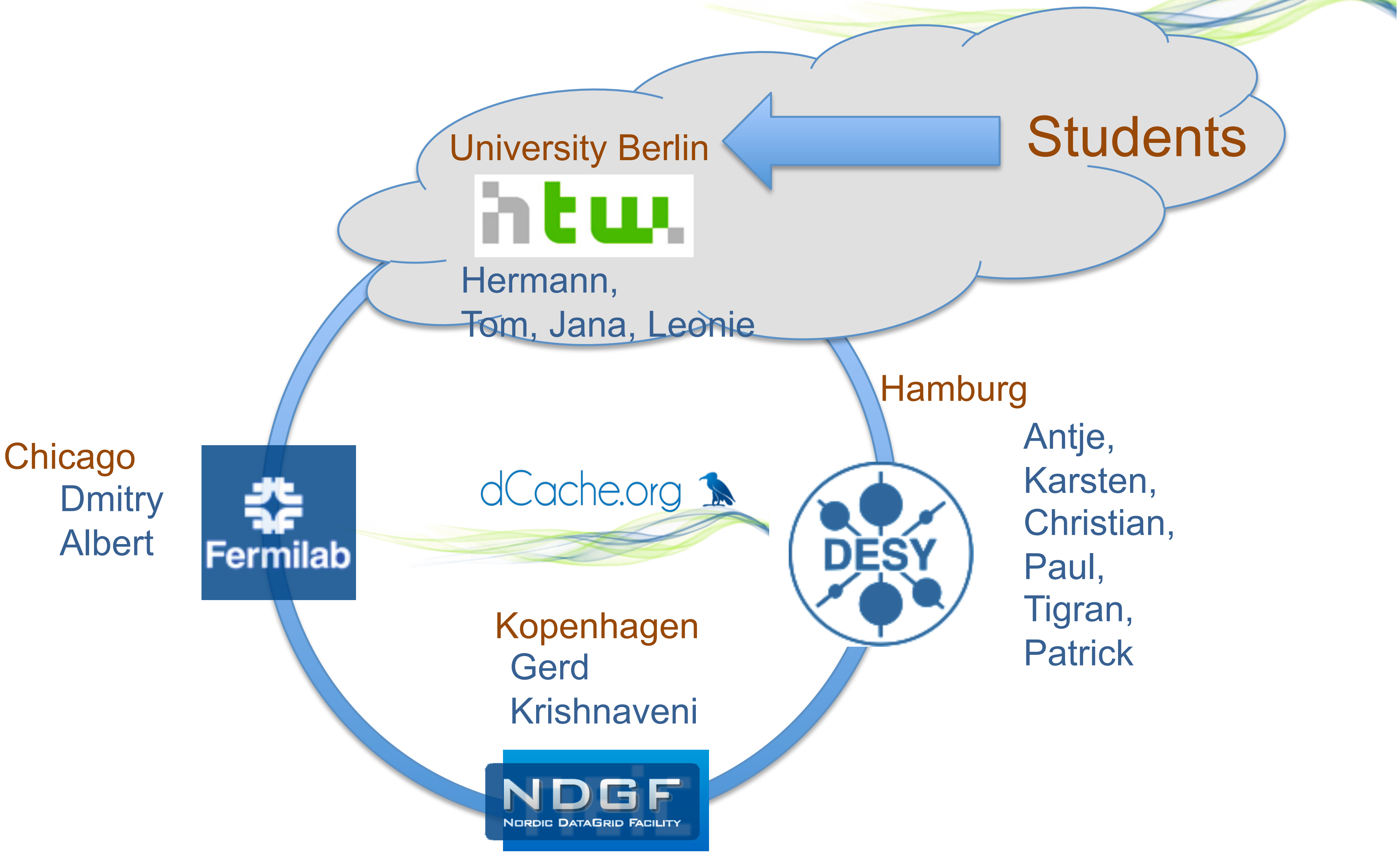
Cheat Sheet

Cheat Sheet

- dCache.org is an international collaboration, developing and distributing storage software (dCache)
- dCache is in production in about 60 places around the world and stores (roughly) about 120 Pbytes in total for WLCG.
- dCache supports different storage media, like disk, SSD and tape and provides mechanisms for manual and automated internal and external replication and transitions.
- dCache storage can be accessed via standard protocols like WebDAV, NFS, and gridFTP and proprietary protocols like dCap and xrootd, and in process of impl. CDMI.
- dCache supports a variety of authentication and mapping mechanisms, e.g. Kerberos, X509, User/Password, LDAP, NIS, NSSWITCH.

Project Structure

The dCache partners and team



Students

University Berlin

Hermann,
Tom, Jana, Leonie

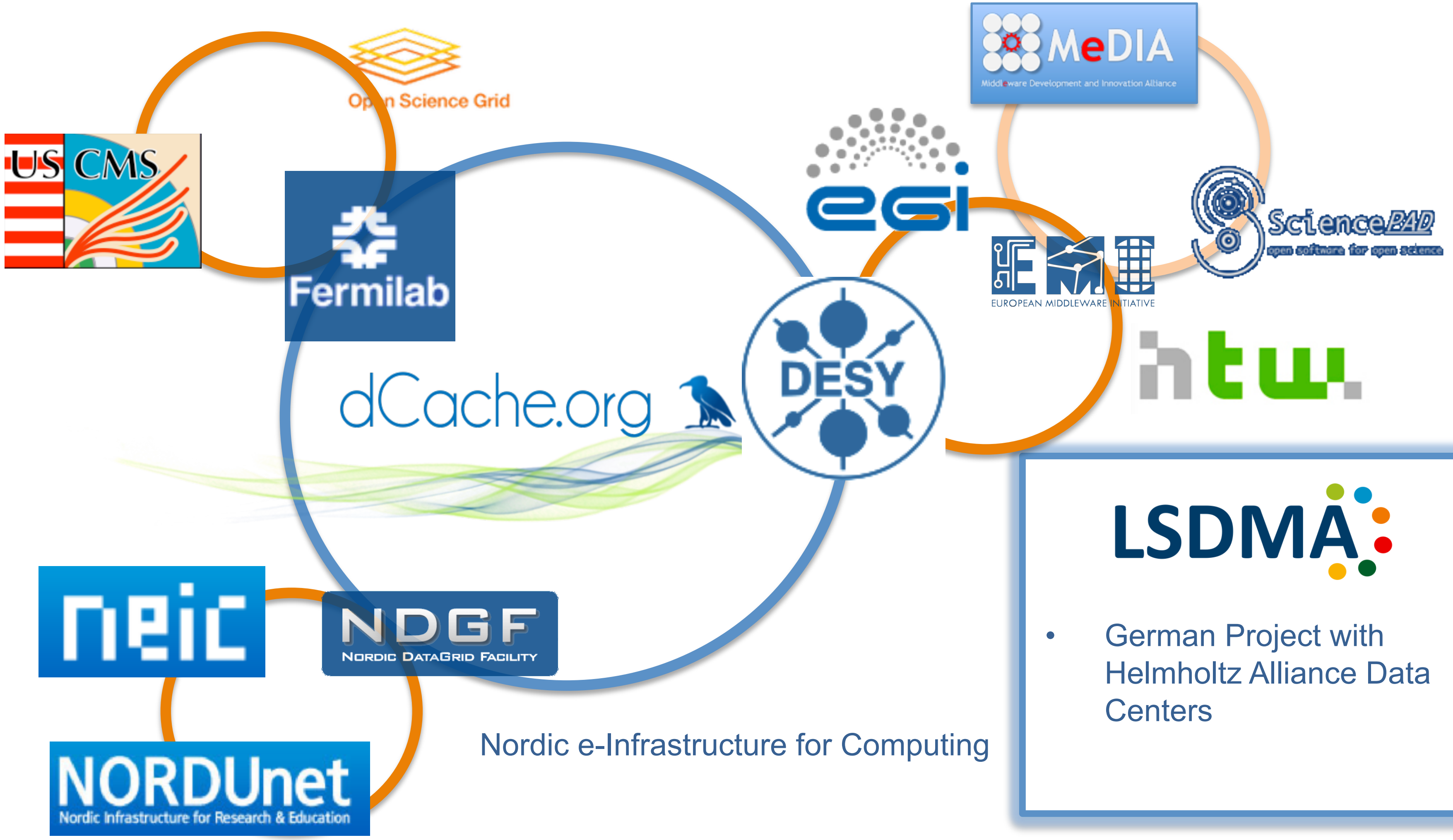
Hamburg

Antje,
Karsten,
Christian,
Paul,
Tigran,
Patrick

Chicago
Dmitry
Albert

Kopenhagen
Gerd
Krishnaveni

dCache partners bridging national projects and activities.


Open Science Grid
US CMS
Fermilab
dCache.org
DESY
egi
MeDIA
Middleware Development and Innovation Alliance
EMI
EUROPEAN MIDDLEWARE INITIATIVE
ScienceBAD
open software for open science
htw
neic
NDGF
NORDIC DATAGRID FACILITY
NORDUnet
Nordic Infrastructure for Research & Education

Nordic e-Infrastructure for Computing

LSDMA

- German Project with Helmholtz Alliance Data Centers

LSDMA according to dCache

dCache.org

Data Lifecycle Labs (Customers)

- Energy
  - smart grids, battery research, fusion research
- Earth and Environment
- Heath
- Key Technologies
  - synchrotron radiation, nanoscopy, high throughput microscopes, electron-microscope imaging techniques
- Structure of Matter

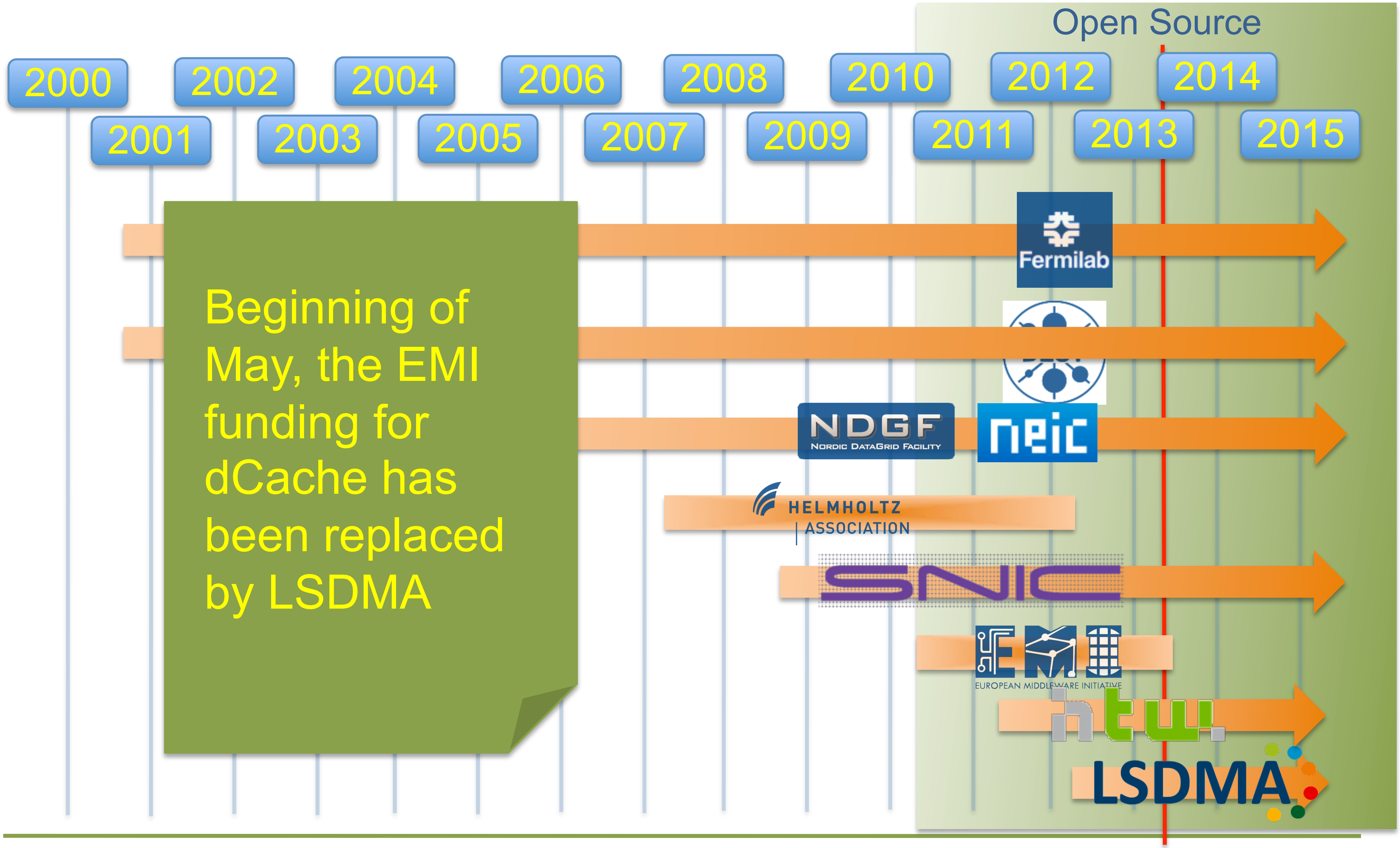
Data Service Integration Team

dCache.org

- **Federated Identity**
- Federated Data Access
- Metadata Management
- Archiving

Funding and Partners

dCache project timeline



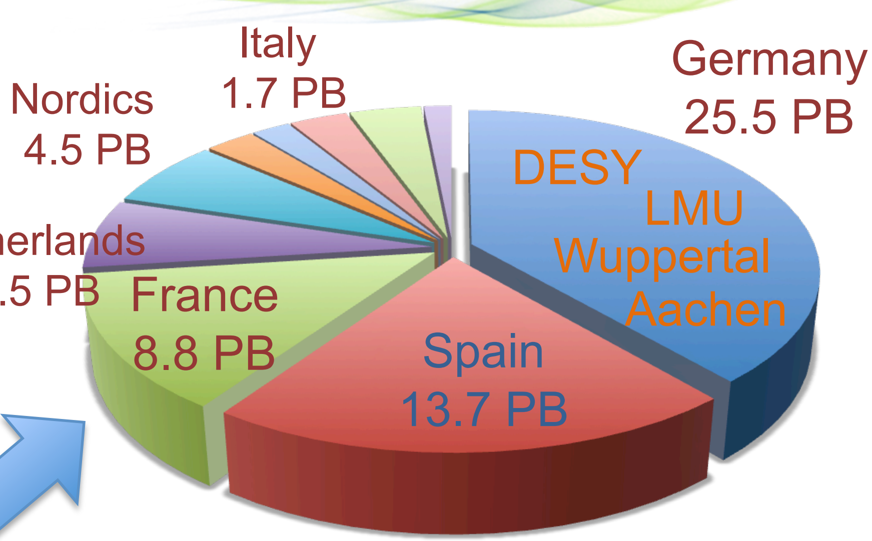
Beginning of May, the EMI funding for dCache has been replaced by LSDMA dCache.org

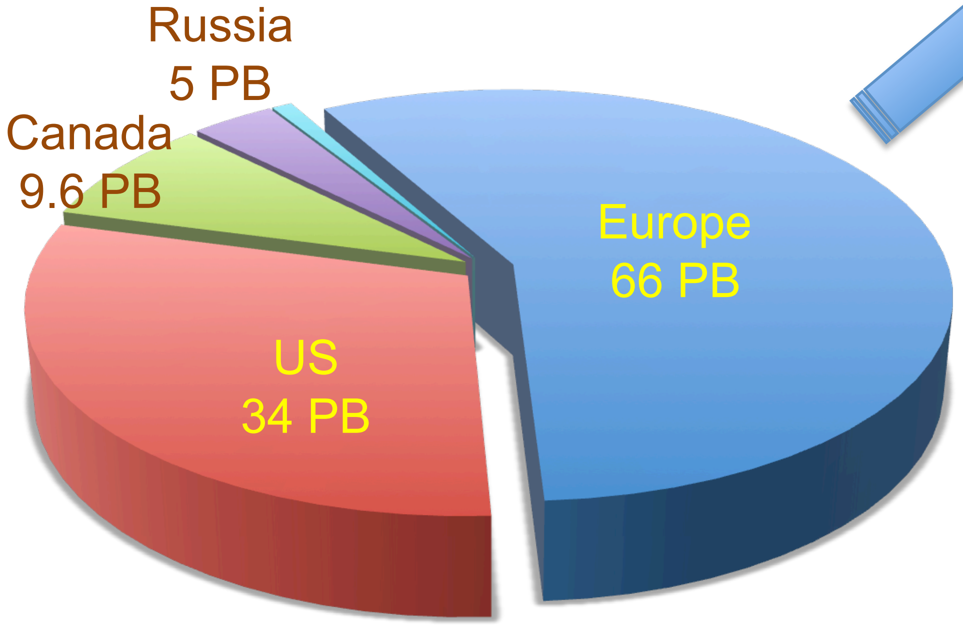
Deployments

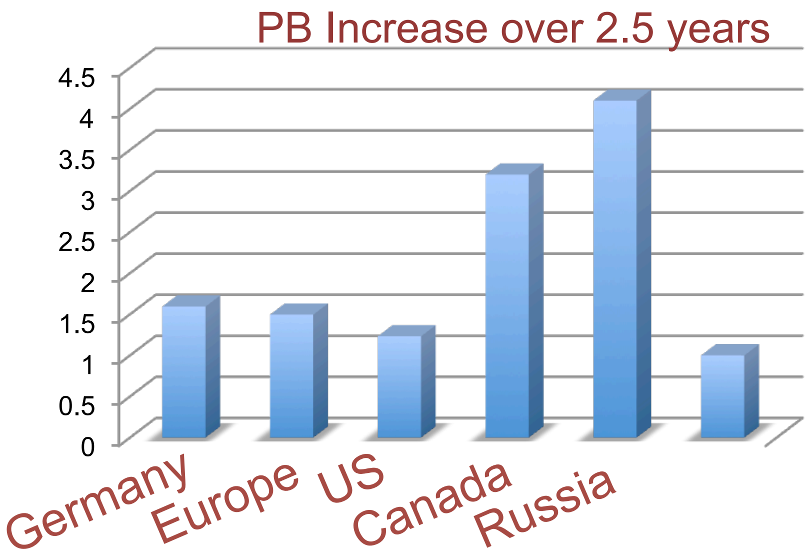
dCache storage for WLCG

dCache.org

- About 115 PBytes just for WLCG
- In 8(+2) out of 11(+3) Tier 1 centers
- And about 60 Tier 2's, which is
- about ½ of the entire WLCG data.







dCache.org

Recent deployment news:

US CMS Nearline System will be a dCache

Poster @ CHEP'13

Evaluating Tier-1 Sized Online Storage Solutions, by Ian Fisk

And Lisas presentation (this morning)

FERMIlab, Intensity Frontier

3 Pbytes of new dCache storage at FERMIlab

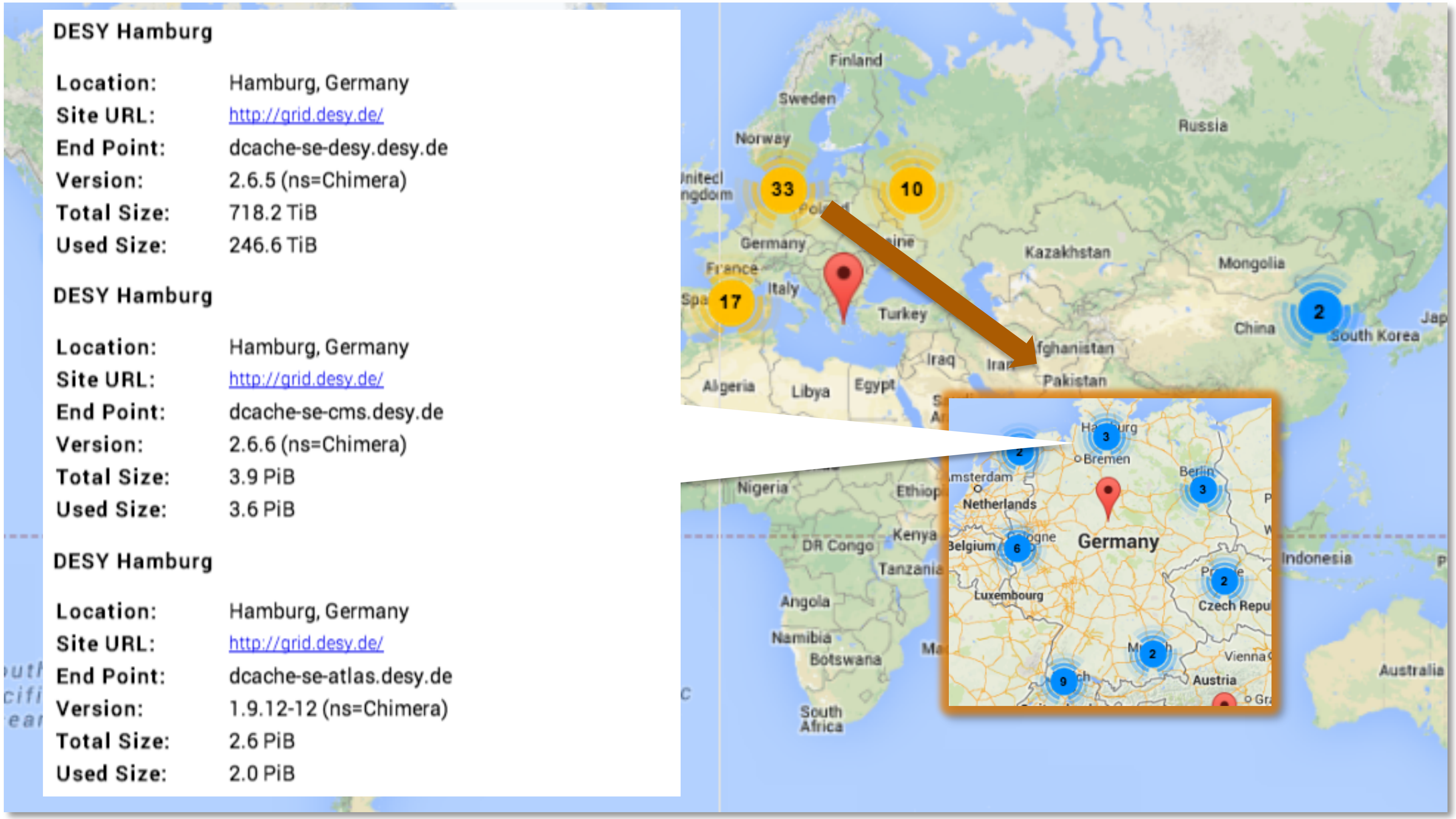
Moscow: 2 * Tier I's

Kurtschatow & Dubna

dCache and Enstore

DPHEP: DESY Data Preservation storage system

Tigrans new dCache world map



DESY Hamburg

Location:	Hamburg, Germany
Site URL:	http://grid.desy.de/
End Point:	dcache-se-desy.desy.de
Version:	2.6.5 (ns=Chimera)
Total Size:	718.2 TiB
Used Size:	246.6 TiB

DESY Hamburg

Location:	Hamburg, Germany
Site URL:	http://grid.desy.de/
End Point:	dcache-se-cms.desy.de
Version:	2.6.6 (ns=Chimera)
Total Size:	3.9 PiB
Used Size:	3.6 PiB

DESY Hamburg

Location:	Hamburg, Germany
Site URL:	http://grid.desy.de/
End Point:	dcache-se-atlas.desy.de
Version:	1.9.12-12 (ns=Chimera)
Total Size:	2.6 PiB
Used Size:	2.0 PiB

Available at dCache.org

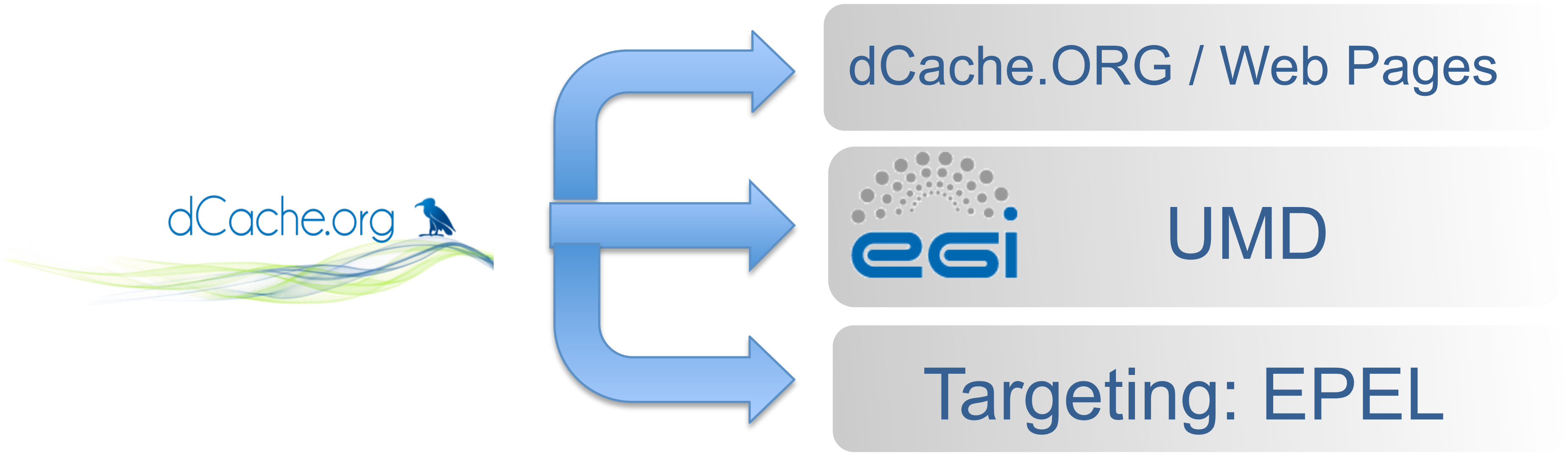
Interesting installations

The raspberry dCache

700 MHz ARM
512 MB Memory
2 * USB 2
100 MB Ethernet

Customer Relations

Deployment Channels



dCache.ORG / Web Pages

UMD

Targeting: EPEL

Reminder of Support Channels

support@dCache.org (security@dCache.org)

- for all bug reports, feature requests and requests for help. Tickets are distributed to all dCache partners.

German Support Group:

- Group composed of German dCache sites, helping each other with monitoring and daily operational work and organizing the dCache tutorial of the annual GridKA school of computing

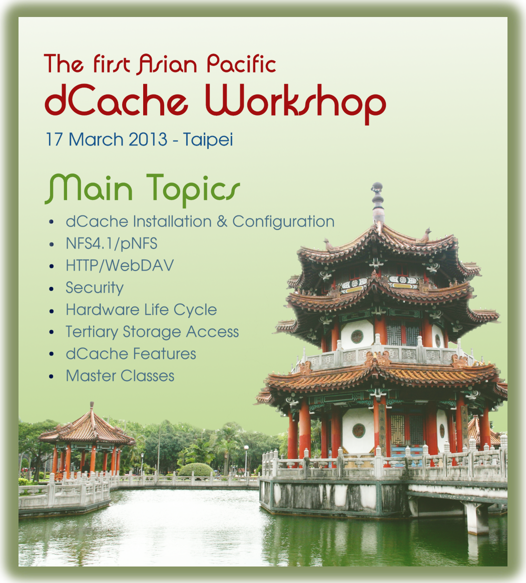
EGI.eu:

- Frist level support for dCache packages taken from UMD.

- Weekly customer phone meetings
- 2 dCache workshops/year (Europe & Asia) plus GridKa School

Asian and European Workshops

This Year



The first Asian Pacific

dCache Workshop

17 March 2013 - Taipei

Main Topics

- dCache Installation & Configuration
- NFS4.1/pNFS
- HTTP/WebDAV
- Security
- Hardware Life Cycle
- Tertiary Storage Access
- dCache Features
- Master Classes

Next Year

Asian Workshop

22 or 23 or 24 March'14

European Workshop

Trying close to Spring HEPIX

Work in progress

WLCG

NFS
Federations
CMS Disk Tape Separation

Moving-on with NFS 4.1 / pNFS

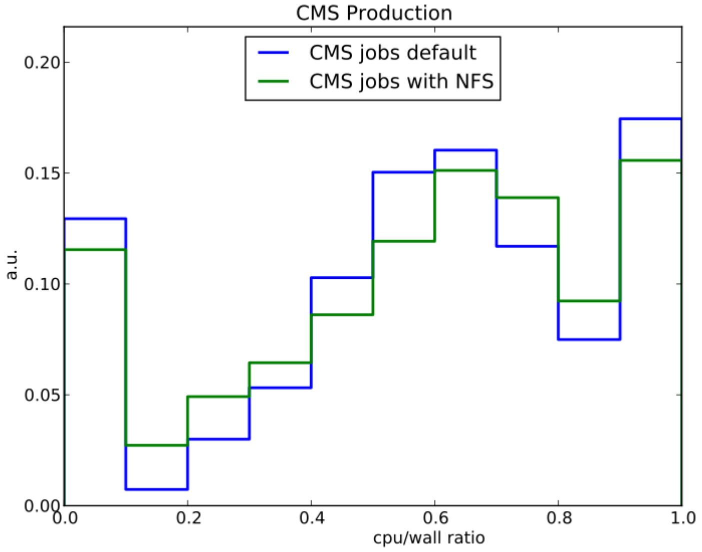
- Quick reminder:
  - pNFS allows storage elements (e.g. dCache and DPM) to be mounted like regular disk systems.
  - Other than 'fuse', It provides scaling by letting the client directly exchanging data with the individual storage node.
  - Photon Science and BELLE (1&2) are already accessing their data via NFS at DESYs dCaches for 1-2 years.
- As SL6 is now ready for WLCG, NFS 4.1/pNFS clients are available on work group servers and worker nodes.
- CMS and ALTAS dCache at DESY have been upgraded, supporting latest NFS4.1/pNFS server.
- DESY is now evaluating NFS for CMS (many thanks to Christoph Wissing and DOT Team)

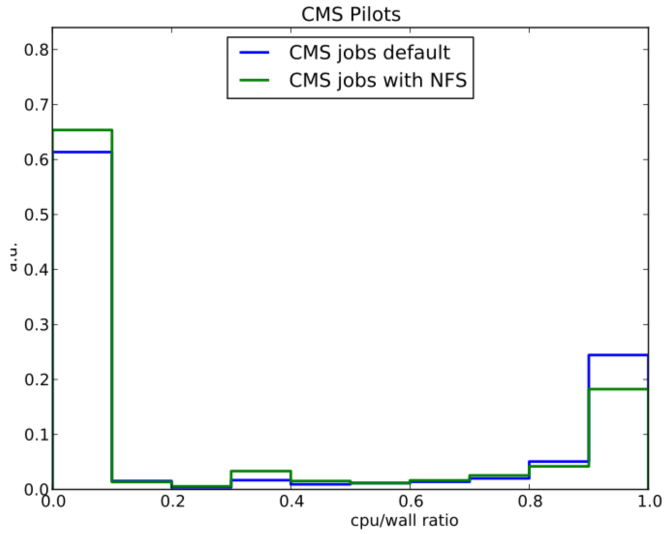
NFS 4.1 Setup

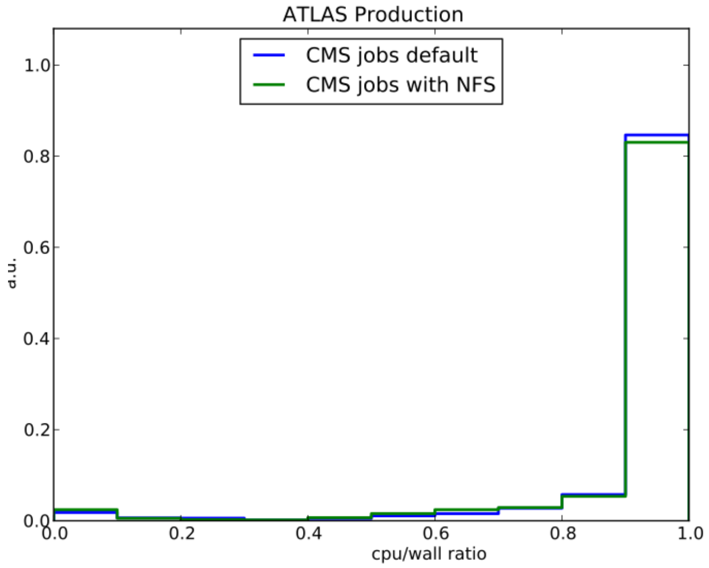
- We configured about 60 generic WN’s (1000 Job Slots) with NFS access to the CMS dCache.
- Only CMS jobs on those machines are using NFS4.1.
- CMS jobs on the other WN’s are still using dCap.
- ATLAS jobs on all machines are still using dCap.

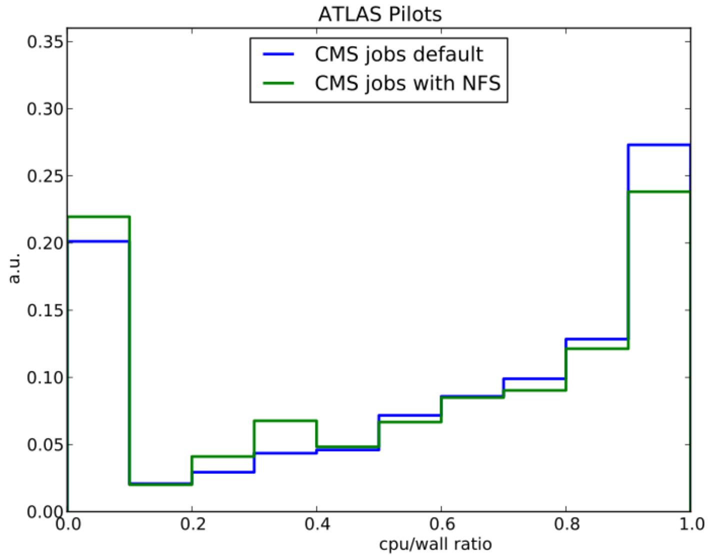
Very first results (cont.)

Thanks to Friederike Novak for the analysis



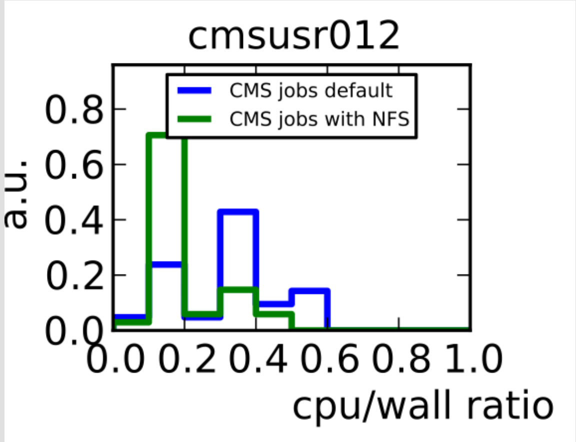






NFS results and issues

Worst we could find



- We need more statistics to get a clear statement.
- We have to have a closer look into ‘very bad cases’.
- We need to make sure the nfs vector-read (fadvise) makes it back into the ROOT “file:” driver.
- Found a ‘protocol incompatibility’, which is now fixed.

Next Steps

- Extend the NFS mount to the entire WN space (automount)
- Extend the usage of NFS to other WLCG VOs
- FERMIlab looking into NFS mounts for Intensity Frontier.

Federations

dCache.org

xRootd federation

- Rob Gardner initiated a closer relationship between dCache and the Atlas Federation People (Andy, Lukazs, Illya)
- One f2f meeting at DESY followed by regular phone meetings
- A list of issues dCache needs to solve to simplify pure dCache systems to become part of the Atlas federation.
- With Gerd now again heavier involved in dCache (see Mattias W. slides) this should now move on faster.

Federations (cont.)

HTTP/WebDAV federation

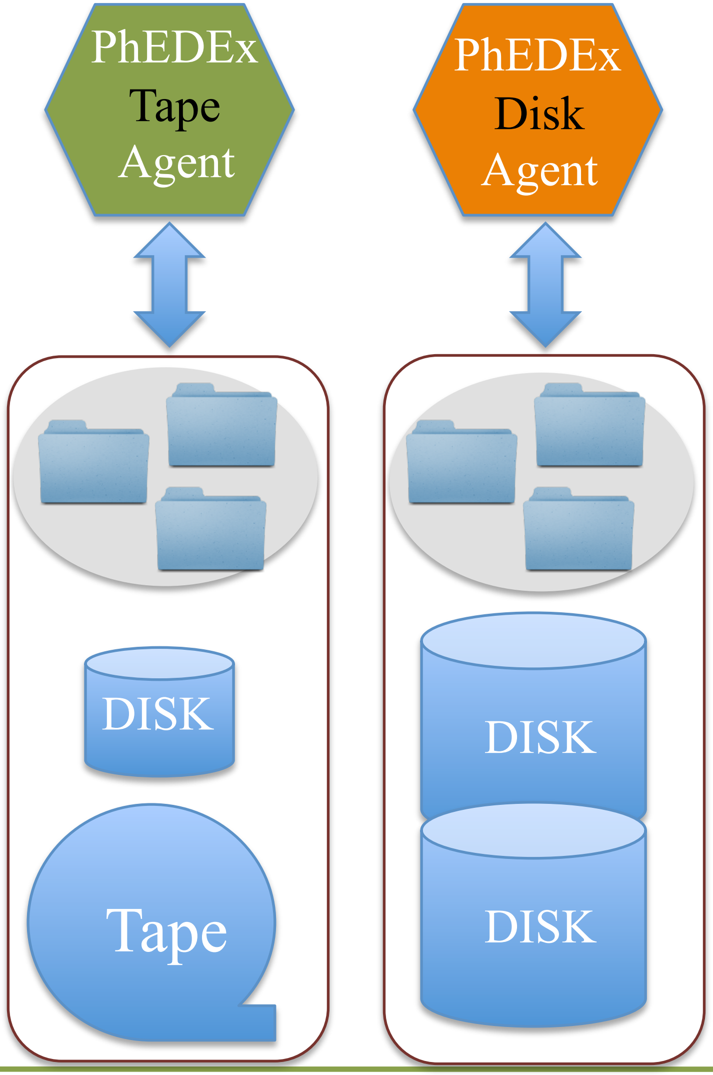
- Close collaboration with the folks at CERN (Fabrizio F.) on building a Dynamic Http Federation.
- The prerequisites
  - The “HTTP eco system” is on a good way. We’ll start to check for http/WebDAV endpoints with SAM Jobs.
  - ATLAS file renaming is already done with WebDAV
  - Http plug-in available for xrootd storage elements.
- The actual Dynamic Federation System is already running as prototype at CERN and DESY.

CMS Disk Tape separation

- CMS decided to no longer assume that a storage system can do disk – tape transitions.
- Therefore the PhEDEx agents assume to talk either to a disk-only or the tape-only (?) system.
- Transitions between disk and tape are done through PhEDEx by just copying files between the two types of systems.
- As all Tier 0/I storage systems support internal transitions (except EOS), we have different options to fulfill that requirement.

CMS Disk Tape Separation

Two independent dCache's
(FERMIlab)



PhEDEx Tape Agent

PhEDEx Disk Agent

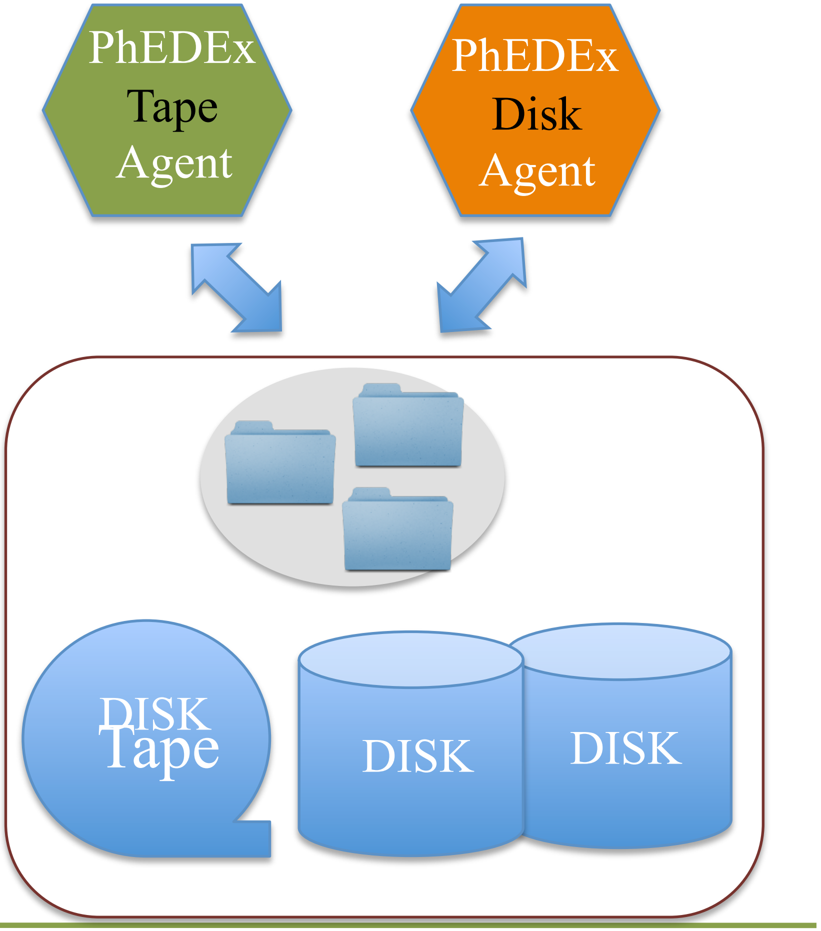
DISK

Tape

DISK

DISK

Best would be



PhEDEx Tape Agent

PhEDEx Disk Agent

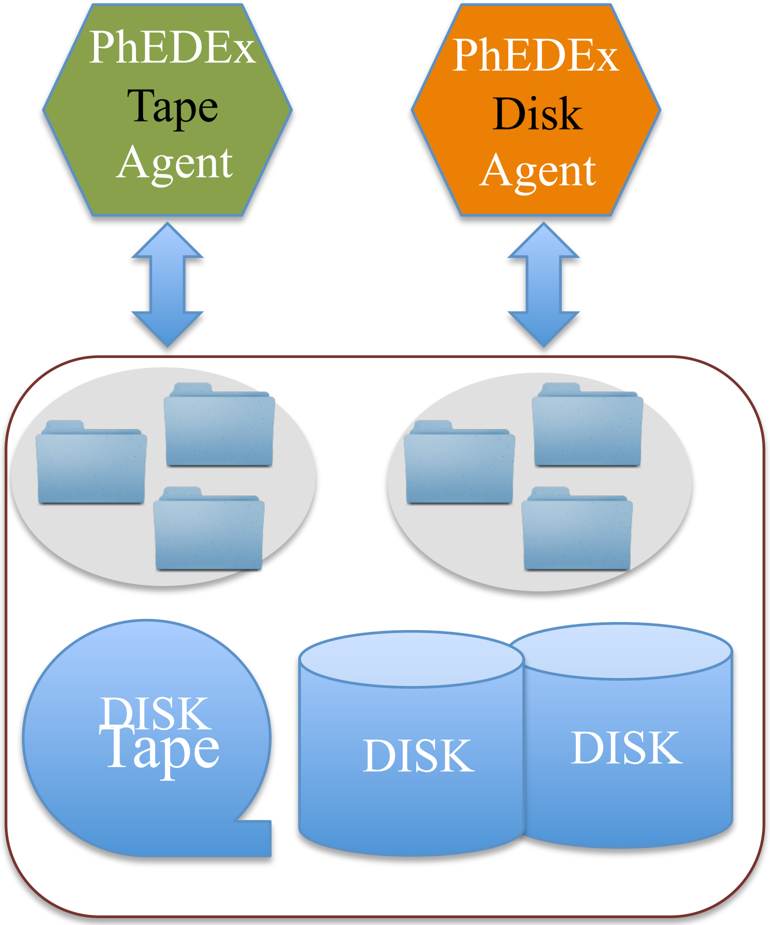
DISK
Tape

DISK

DISK

CMS Disk Tape Separation

One dCache, two namespaces
IN2P2, KIT, RAL,CNAF, (PIC)



- D e change.
- lifferent the
- ache, P ops.
- Has spac provements. Plans are
  - PhEDEx ag s very flexible in talking to the ndpoint, so
  - dCache can p rform internal copies
  - Allowing 'deferred write to tape', w/o external copies.

Completes the dCache data lifecycle operations

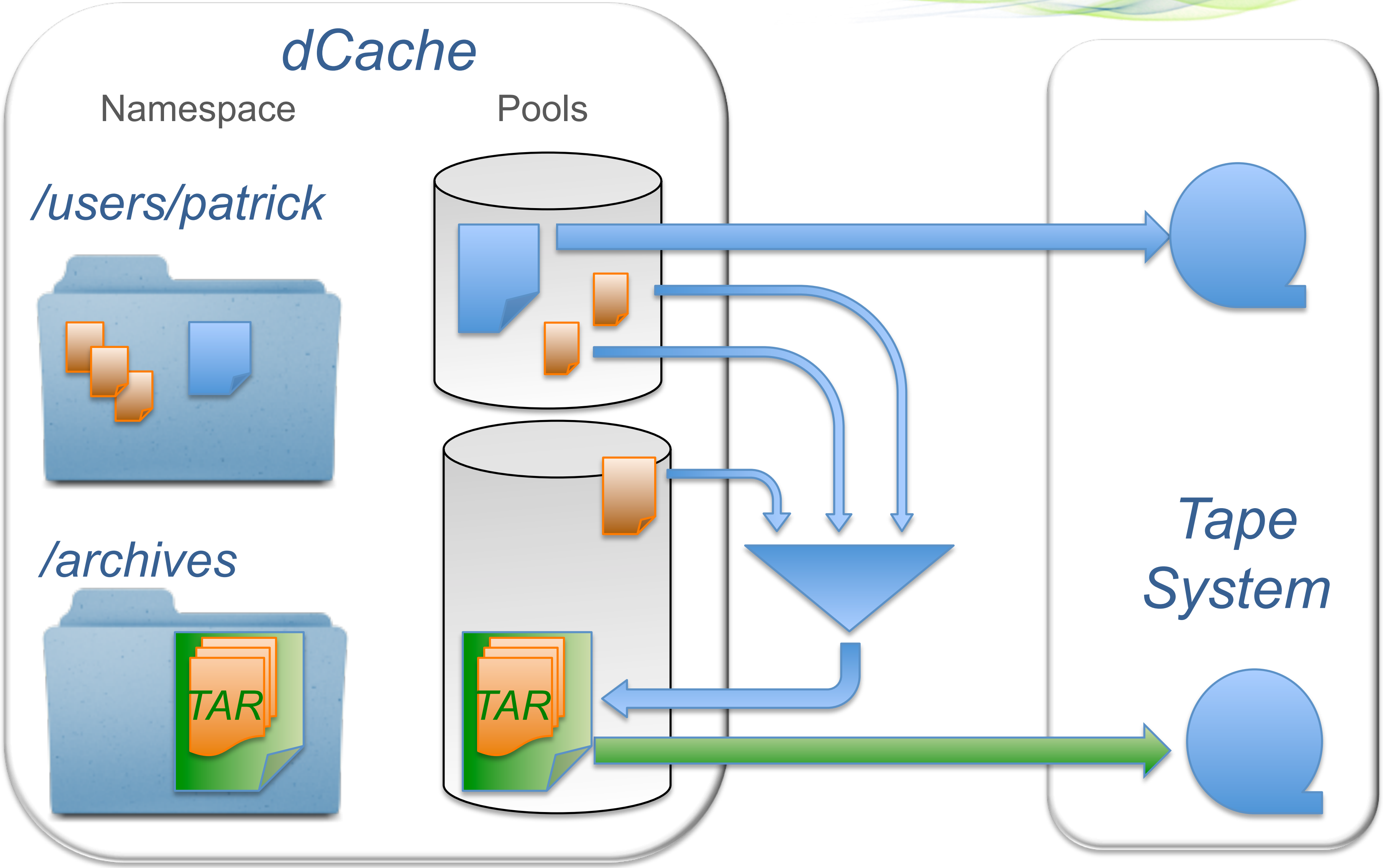
Photon Science and LSDMA

- Small File Support for Tertiary Backends
- Support of HDF5, Nexus file formats
- XAML, Oauth Support
- Cloud Data Management Interface
- Object Stores

Small files support for tertiary storage

- Tape Systems are notoriously bad in handling small files.
  - Waste of space on the tape, due to large file marks
  - Non streaming behavior significantly reduces tape system performance
  - See Sashas presentation from ½ h ago.
- Our take is to fix this in dCache, so that all Tertiary Storage Systems can benefit if they want.

Small files support for tertiary storage (cont.)



Small files support for tertiary storage (cont.)

- Prototype is running in conjunction with our Photon Science web portal.
- For now, this is just a service at DESY and not yet part of the dCache release.
- It’s however rather dCache version agnostic.

Support for HDF5, Nexus files

- Why does the file format matter for dCache?
- Because those files are containers
- They are filled by subsequently running processes
- This means we need to be able to modify a file after it has been closed the first time.

Read/Modify/Write for dCache ?

Support for HDF5, Nexus files(cont)

- Read/modify/write for dCache ?
- Almost ☺
- The plan is:
- Initially for NFS only.
- No replicas and tape copies while in r/m/w mode.
- After mode change to ‘immutable’, the file becomes a regular dCache file with replication and tape copy.
- ‘Immutable mode’ can’t be reversed to ‘r/m/w’ mode.

More Needful Things

- Extend gPlamza to support web based authentication. (IdPs)
  - Your google or twitter account.
  - XAML (e.g. Schibboleth) assertions from federated IdPs
  - Technically possible, but David K is making trouble ☺.
- Implementing CDMI (SNIA standard, HTW Students)
  - For data transfers
  - For meta data storage (replacing Photon Science ICAT interface)
- Using dCache as Object storage
  - Fast access, avoiding unnecessary name space operations
  - Rucio people were very interested
- Evalutions
  - Running dCache pools on DDN storage boxes (KIT, DDN)
  - Running dCache on top of Ceph.

And now for something completely different

The scientific storage cloud

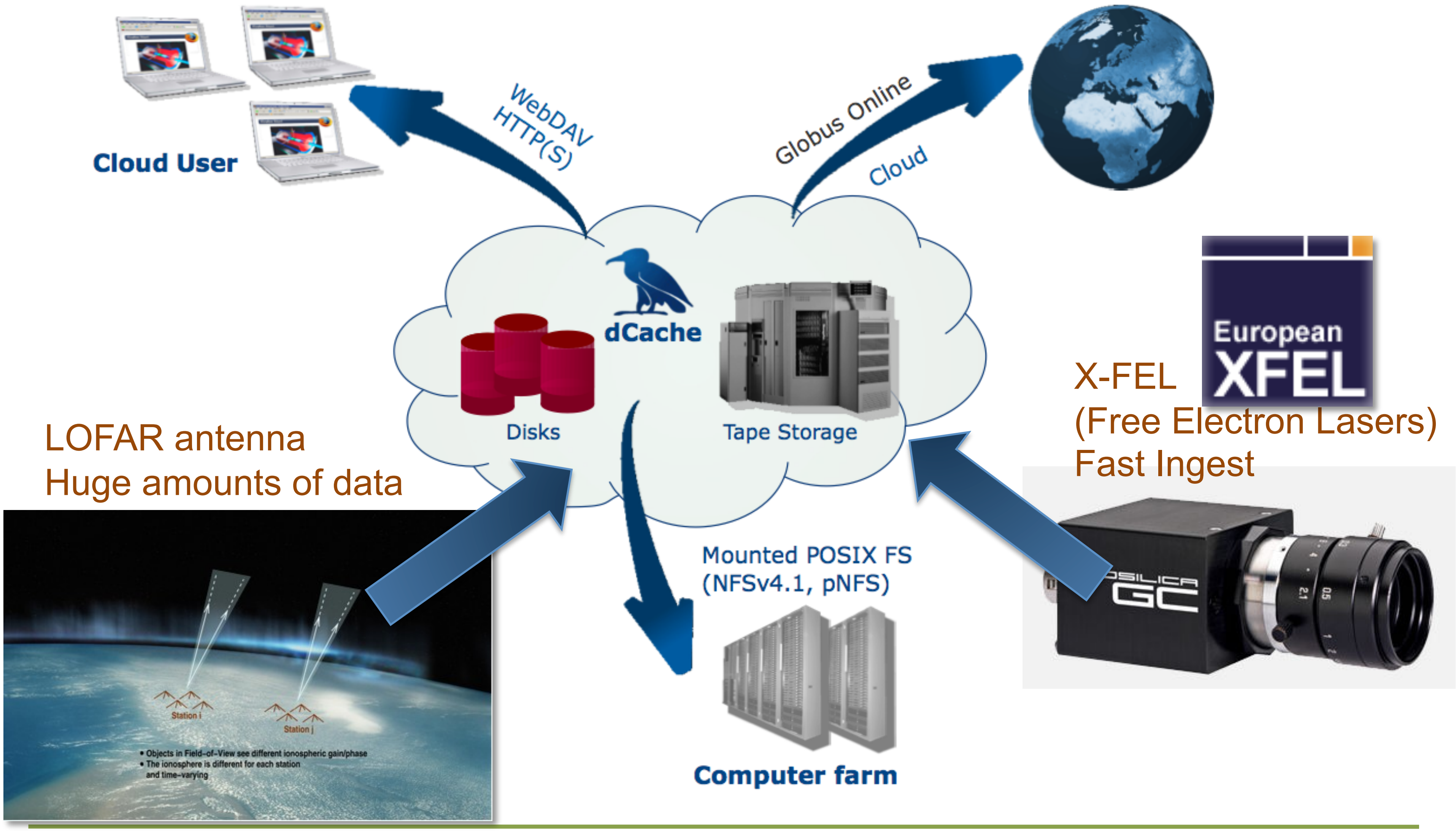
This is about a dCache.org service and not the software

Operated by dCache.org and HTW Berlin under the financial umbrella of the dCache partners and LSDMA

Motivation

- Get students involved
  - They get unlimited storage space and a Master or Bachelor degree
  - We get their time and knowledge on ‘young peoples’ need in terms of storage and sharing
  - Get mobile devices involved
- Provide ‘scientific’ cloud storage
  - Various authentication methods
    - Kerberos, X509, Web 2, Oauth, XAML
  - Various file access methods
    - WebDAV, GridFTP, NFS, CDMI (S3)
  - Various retention properties
    - Scratch, multiple copies, tape

Scientific Storage Cloud



Cloud User

WebDAV
HTTP(S)

Globus Online

Cloud

dCache

Disks

Tape Storage

LOFAR antenna
Huge amounts of data

- Objects in Field-of-View see different ionospheric gain/phase
- The ionosphere is different for each station and time-varying

X-FEL
(Free Electron Lasers)
Fast Ingest

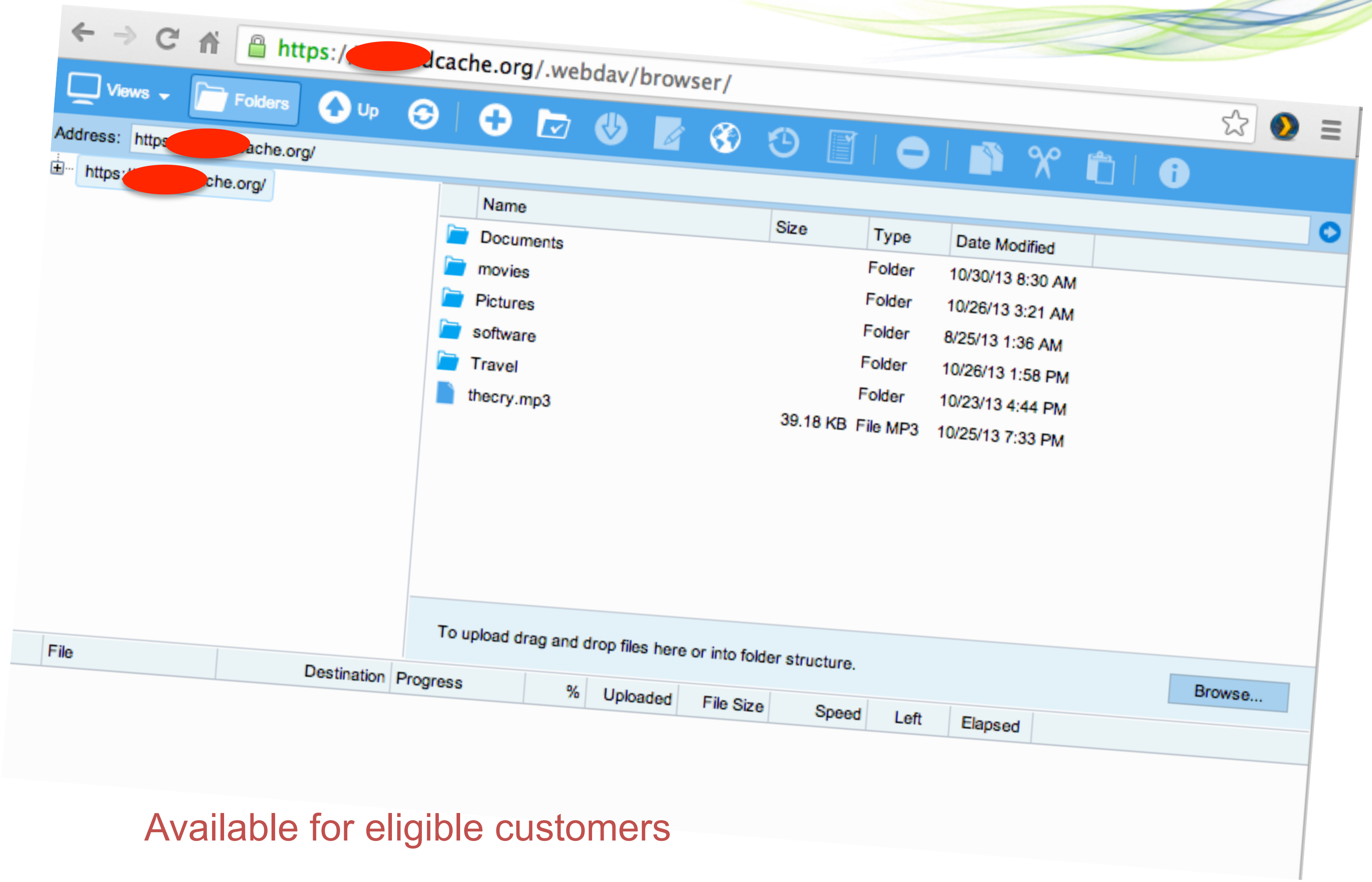
Mounted POSIX FS
(NFSv4.1, pNFS)

Computer farm

Status of S^{2}C

- Service is installed and in use.
- Registration is done through DESY infrastructure (Peter vdR)
- Current access via username/password and WebDAV
- Drivers/Apps for ANDROID already available.
- Registration easier than getting a Google account. (w/o human interaction on our side)
- For convenience, we provide a browser GUI page for up/download and namespace operations.
- You may as well use your OS WebDAV interface.
- Absolutely no sharing for now !!!
  - https://☺.dCache.org/ is always only your home.

Our GUI



https://[redacted]dcache.org/.webdav/browser/

Address: https[redacted]ache.org/

https:[redacted]che.org/

Name	Size	Type	Date Modified
Documents		Folder	10/30/13 8:30 AM
movies		Folder	10/26/13 3:21 AM
Pictures		Folder	8/25/13 1:36 AM
software		Folder	10/26/13 1:58 PM
Travel		Folder	10/23/13 4:44 PM
thecry.mp3	39.18 KB	File MP3	10/25/13 7:33 PM

To upload drag and drop files here or into folder structure.

File	Destination	Progress	%	Uploaded	File Size	Speed	Left	Elapsed

Available for eligible customers

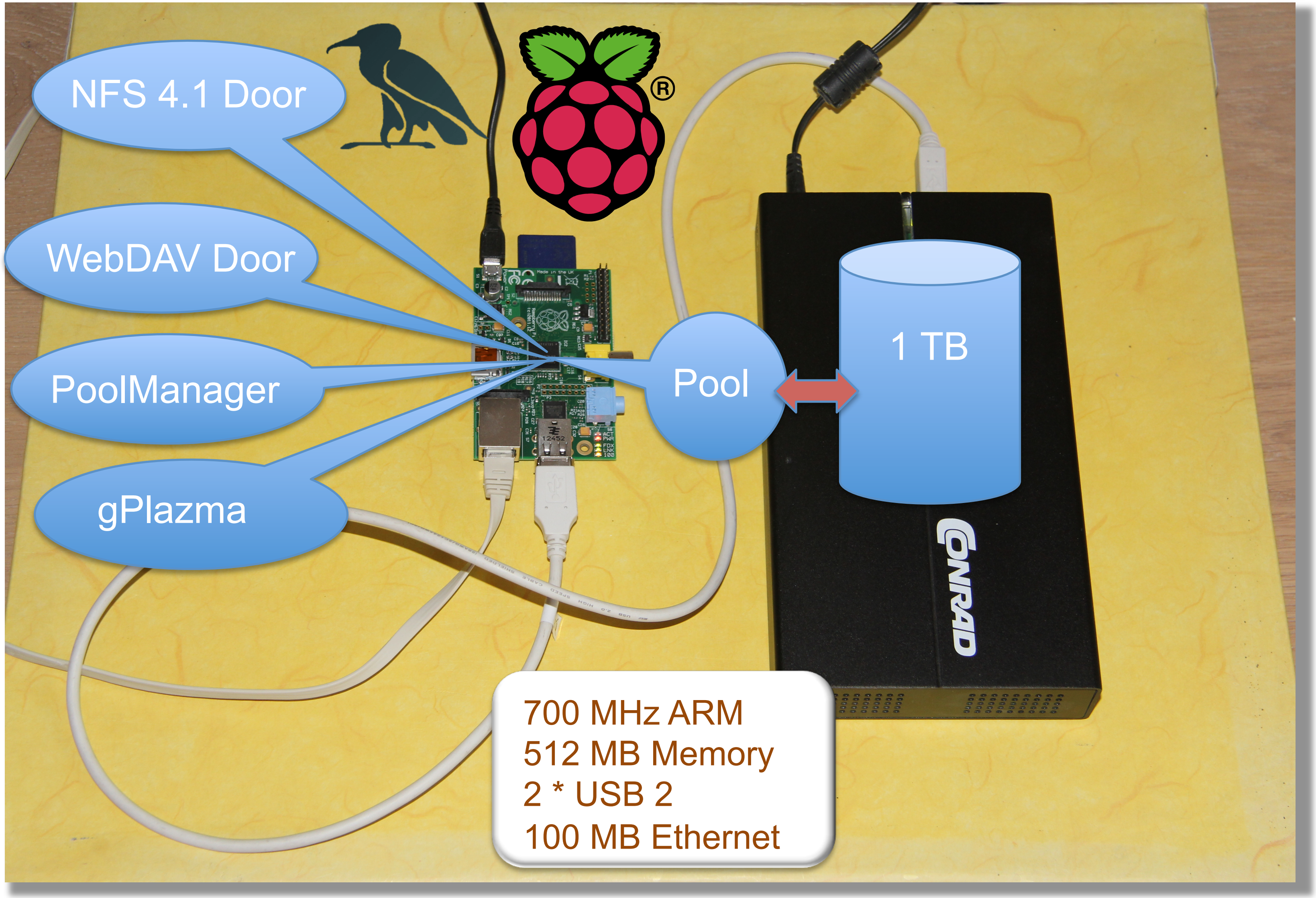
Next Steps S^2C

- Sharing:
  - Public sharing : xxx.dcache.org://public/564-465-765
  - Sharing with users of the same instance
  - Sharing with users of external IdP’s
- Adding more authentication methods:
  - X509
  - XAML (external IdPs)
  - Other web service credentials
- More protocols:
  - Adding GridFTP for transferring data from/to other systems via GlobusOnline or FTS3
  - NFS for local analysis
  - CMDI for cloud applications and data management
- User determined data retention:
  - Scratch space
  - Number of copies > 1
  - Tape copies

In summary

- Due to the broad developers base across international institutions and projects, dCache.org doesn't see any issues in continuous future funding.
- For the same reason, dCache.org is well integrated into the existing infrastructures and communities and keeps on track on upcoming requirements in storage management and access.
- By involving universities and students in the design and development process, dCache is keeping up with the latest developments in computers science and on the requirements of young people in data access and data sharing.

The End



700 MHz ARM
512 MB Memory
2 * USB 2
100 MB Ethernet