dCache.org

dCache, sync'n share for Big Data at DESY

Patrick Fuhrmann

On behave of the project team

INDIGO DataCloud

Fermilab

NDGF Nordic DataGrid Facility

DESY

HELMHOLTZ | ASSOCIATION

Federal Ministry of Education and Research

LSDMA

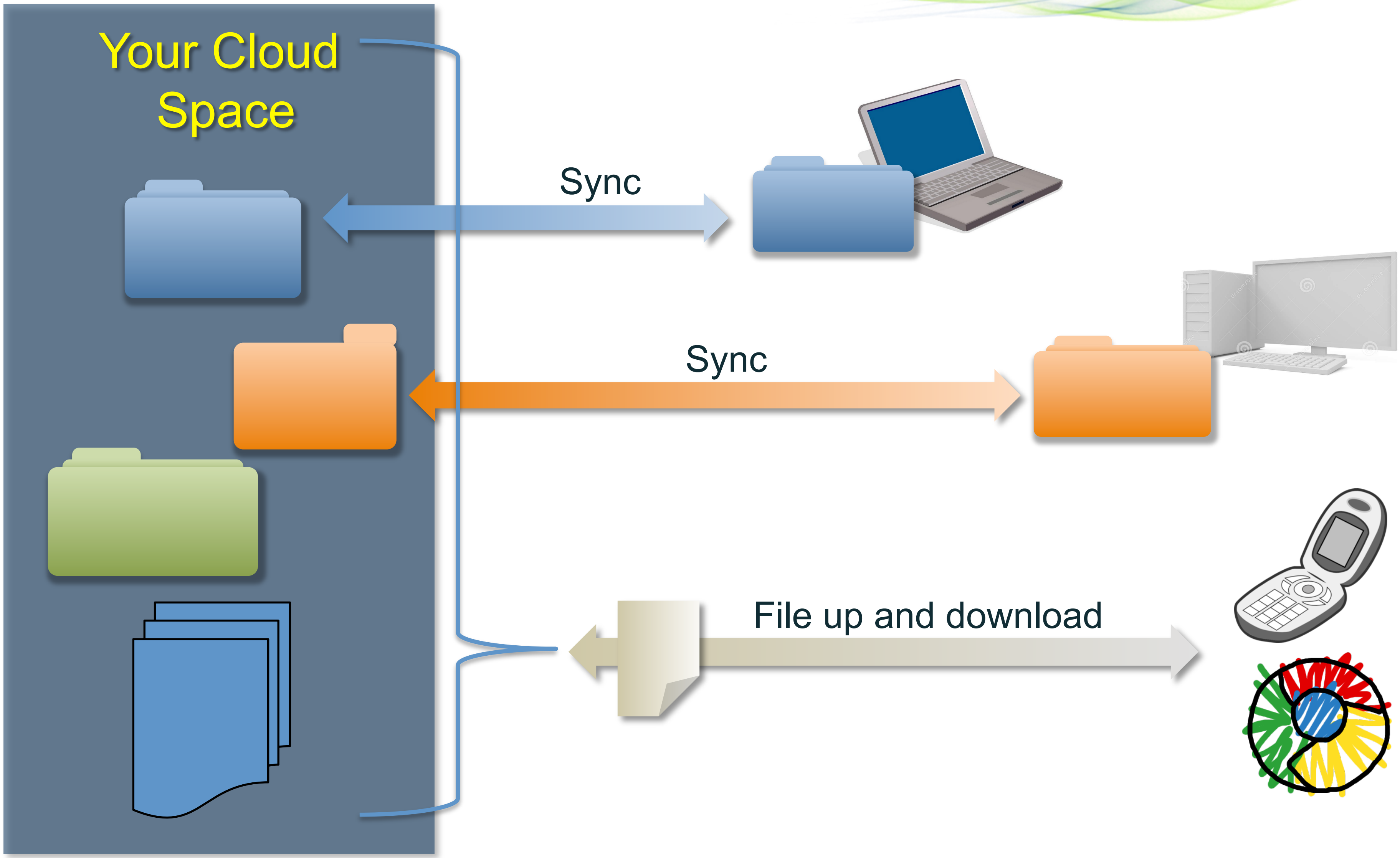
What is this about ?

It’s about on how modern scientists (people) want to manage, access and share their data.

Easy access requirements from DESY users

- New model in accessing data
  - Anytime from everywhere
  - From mobile devices
  - Bidirectional sync’ing between your cloud space and your local devices

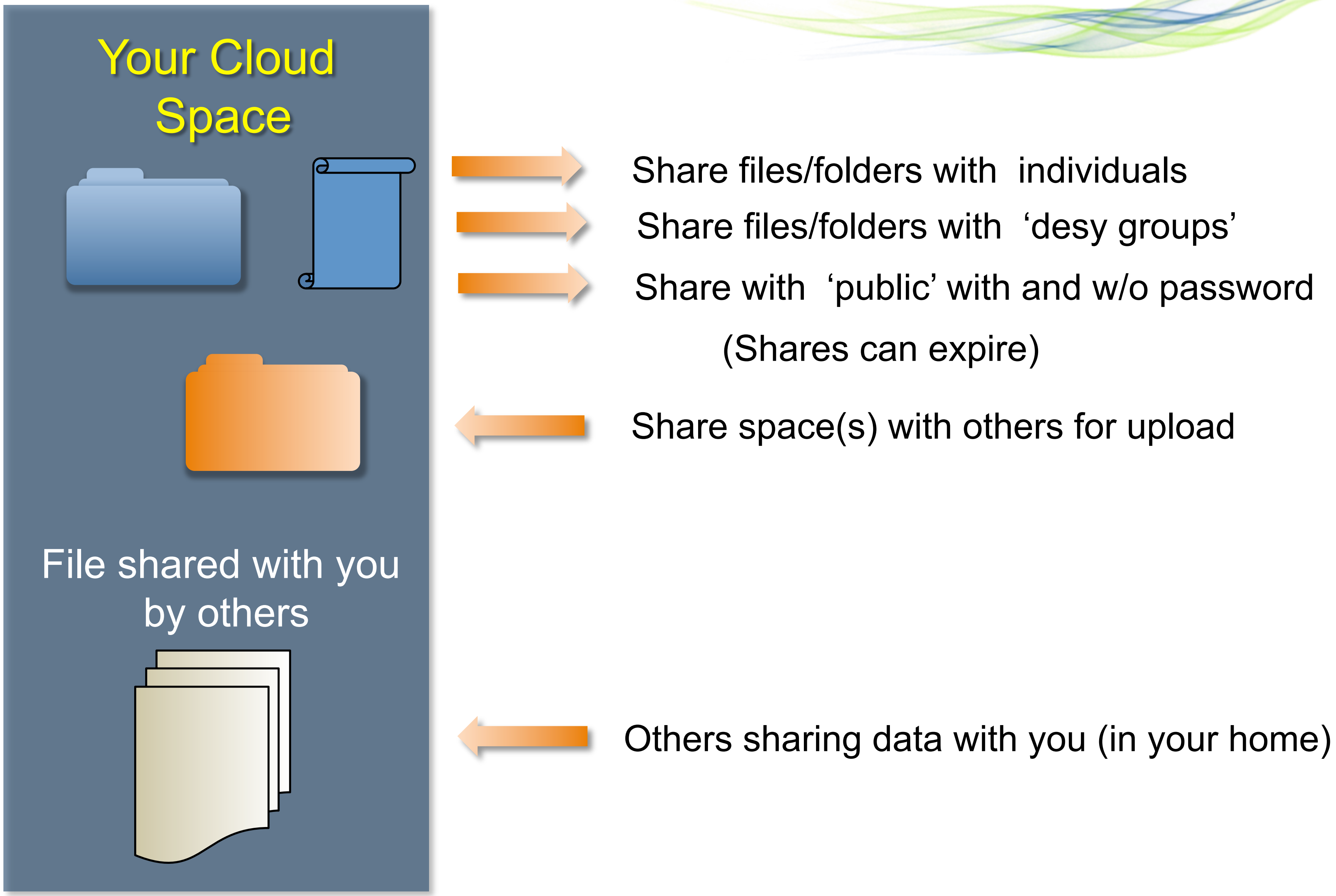
How does that look like



Sharing requirements from DESY users

- Fine grained sharing with individuals and groups.
- Sharing via intuitive Web 2.0 mechanisms (Apps or Browser)
- Sharing with ‘public’ with or w/o password protection
- Sharing of free space (upload)
- Expiration of shares

And the sharingj part



Your Cloud Space

Share files/folders with individuals

Share files/folders with ‘desy groups’

Share with ‘public’ with and w/o password

(Shares can expire)

Share space(s) with others for upload

File shared with you by others

Others sharing data with you (in your home)

Why not using

Dropbox ?

- Because there was this gentleman who decided to leave the US towards Moscow, with a bunch of documents, changing our attitude towards foreign storage services significantly.
- The DESY directorate essentially disallowed storing DESY documents outside of DESY premises.

Evaluation of possible products

PowerFolder

ownCloud

ETC.

CUBEPAD

- Highly secure group-ware system
- Allows sharing encrypted data

Product evaluation (cont.)

We went for Own Cloud

- Open Source plus Enterprise version
- Most popular solution:
  - – Reduces likelihood for ‘product disappearing’
  - – Possibly building a user-community
    - TU-Berlin, FZ-Jülich, TU-Dresden ****
    - CERN, United Nations
- CERN is evaluating a similar approach and we are in contact anyway (WLCG)

Inevitable RP activities

- Collaboration with HTW Berlin (LSDMA)
- Pre-evaluation of cloud solutions by “InFa” -> Q3/2013
  - Erarbeiten und Umsetzen eines firmeninternen Online-Speicherdienstes in einer Teststellung. (Quirin Buchholz)
- Presenting the concept at HEPIX.
- Information exchange with CERN. (CHEP’13) Oct 13
- Berlin Cloud Event, (mostly OwnCloud and PowerFolder) in Mai 14 (we published first paper)
- Participating the CERN Cloud Event (Nov ‘14) including a presentation of our proposed solution.
- Various papers submitted and accepted at ISGC in Taipei in March and CHEP’15 in Japan.

However, as we do scientific computing and to just storing and sharing images, there is more to consider.

More requirements

- Request for *unlimited, indestructible storage*.
- Request for *different quality of services* (SLA), coming with different price tags and controlled by customer.
  - *Data Loss Protection* (non-user introduced), e.g.:
    - One copy.
    - Two copies on independent systems.
    - Two copies in different buildings.
    - Two copies at different sites (e.g. Hamburg and Zeuthen)
    - Some of above plus ‘n’ tape copies.
  - *Access latency* and max data rate, e.g.:
    - Regular sync and web access.
    - Worker-node access: High throughput
    - Low latency (e.g. on SSD) for HPC.
- User defined *Data Life Cycle*
  - Move data to tape after ‘n’ months.
  - Remove from random access media after ‘m’ months.
  - Make public after ‘x’ month.
  - Remove completely after ‘y’ months.
- Controlled by Web or API (*Software defined storage*)

And not to forget

- Access to the same data via different transport mechanisms
  - GridFTP for wide area bulk transfers
  - http/WebDAV for Web applications
  - NFS 4.1/pNFS for low latency, high speed access (e.g. HPC)
- Access with different credentials
  - Username / password
  - X509 Certificates
  - SAML (Single Sign On)
  - Kerberos
  - Macaroons

Our solution

- Non of the Web 2.0 sync and share software products cover the additional requirements.
- So we went for *dCache* as the actually *storage backend*.
- Which is not really a surprise as we are part of the dCache collaboration.

Now … what’s a dCache

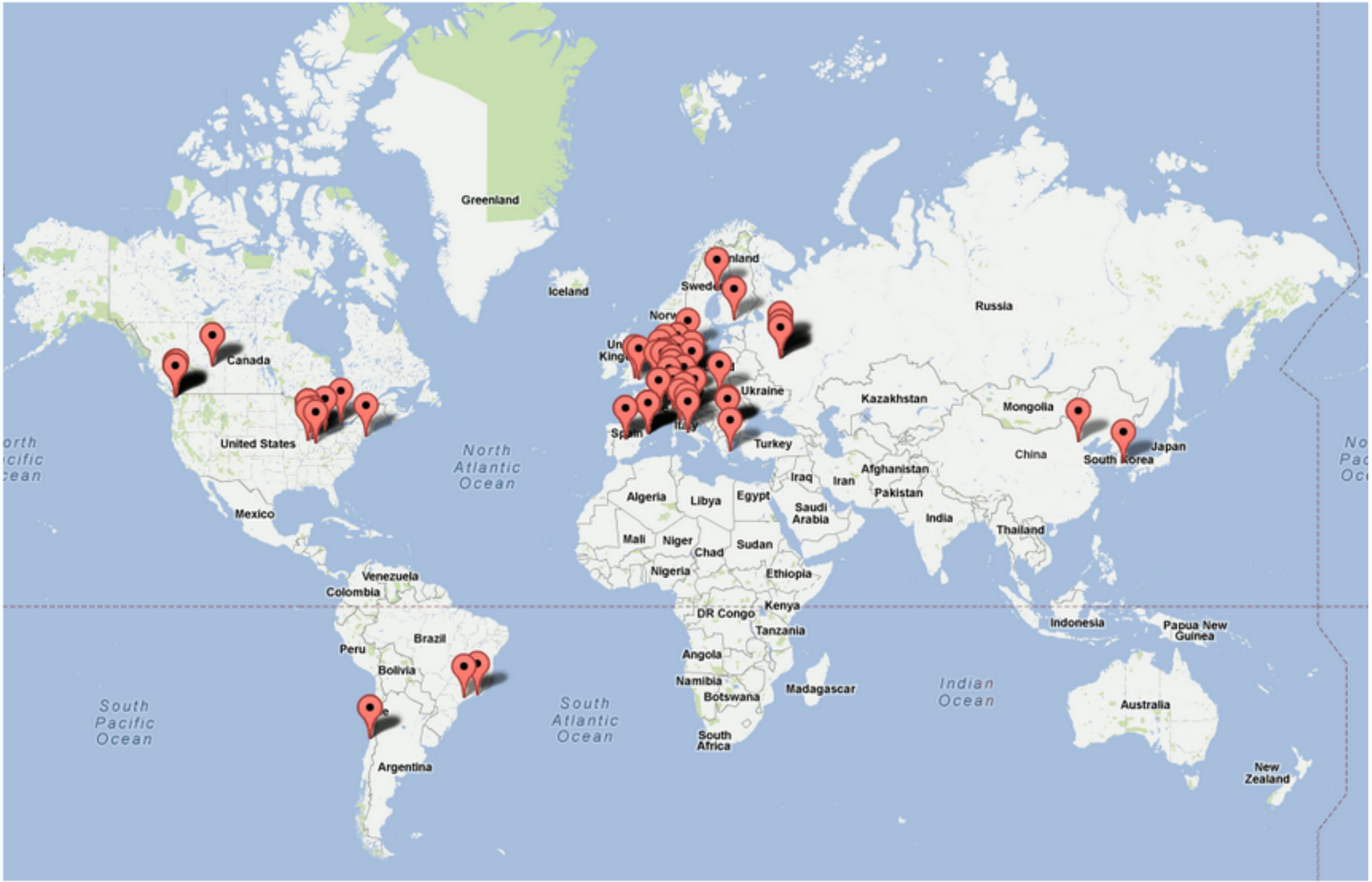
dCache

dCache Cheat - sheet

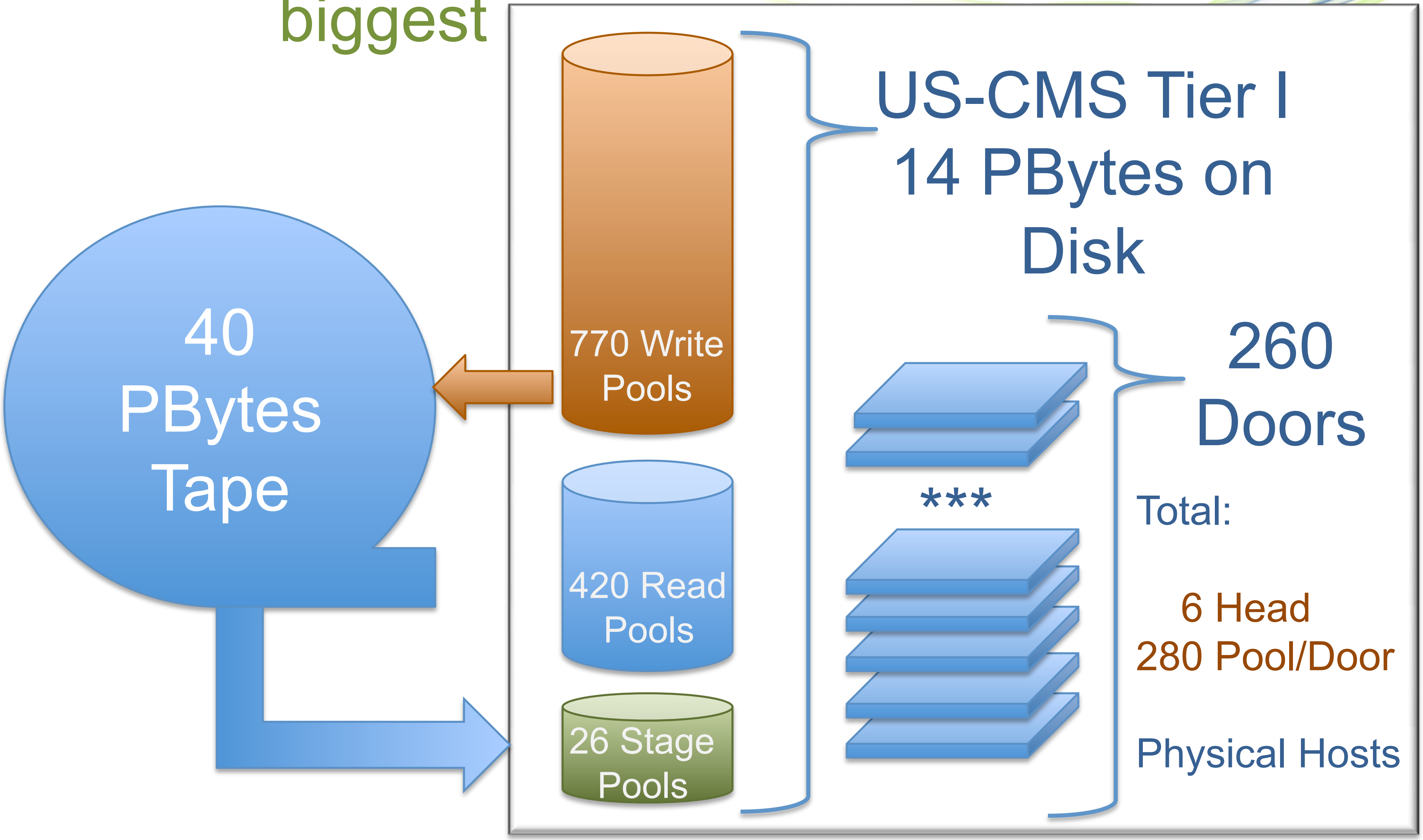
- dCache is a horizontally scaling ‘data management system’ looking like a file system, providing various data access and data management protocols.
- dCache is operated on about 70 sites around the world.
- Total space approaching 200 Petabytes.
  - We store 50 % of the entire WLCG storage.
- Biggest dCache holds about 50 Petabytes on disk and table.
- Larges dCache spans 4 countries.
- dCache is provided by dCache.org

Where do you find dCache’s

Worldwide distribution

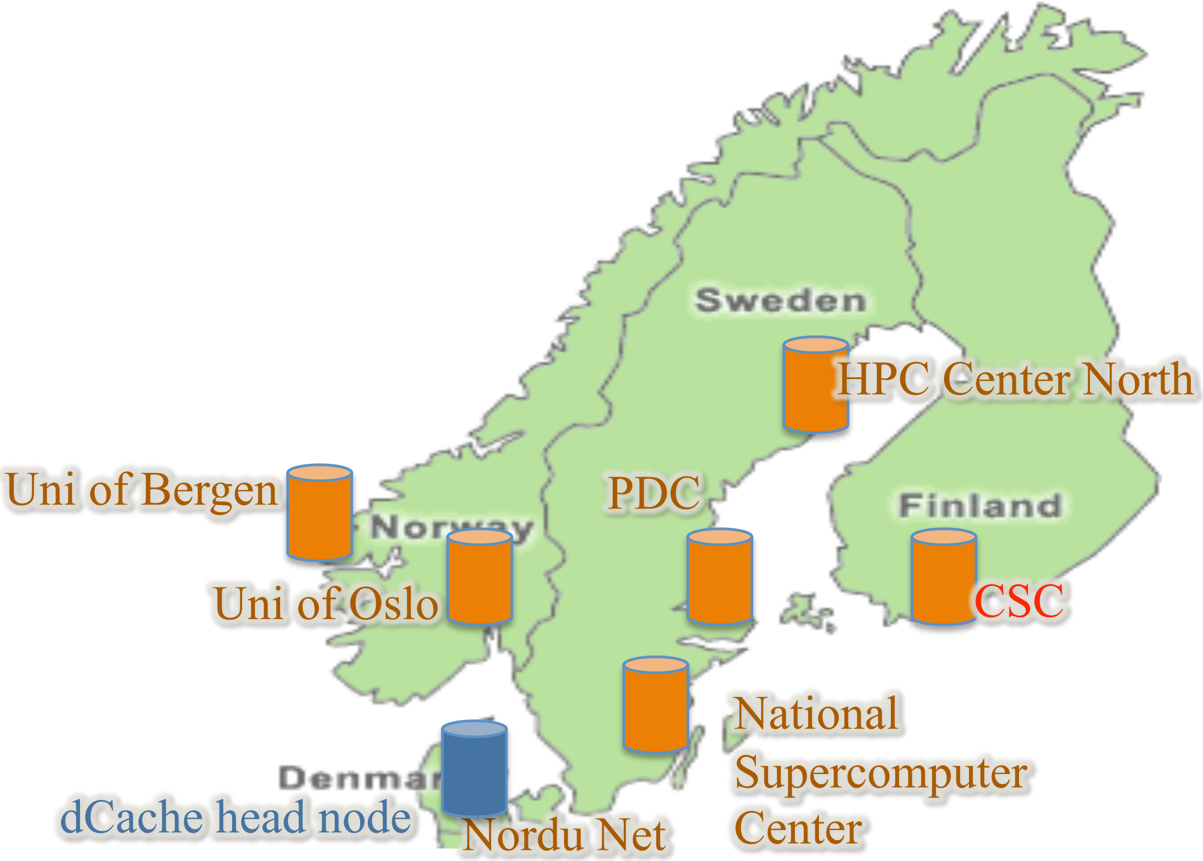


Starting with possibly the biggest



Information provided by Catalin Dumitrescu and Dmitry Litvintsev

To certainly the most widespread



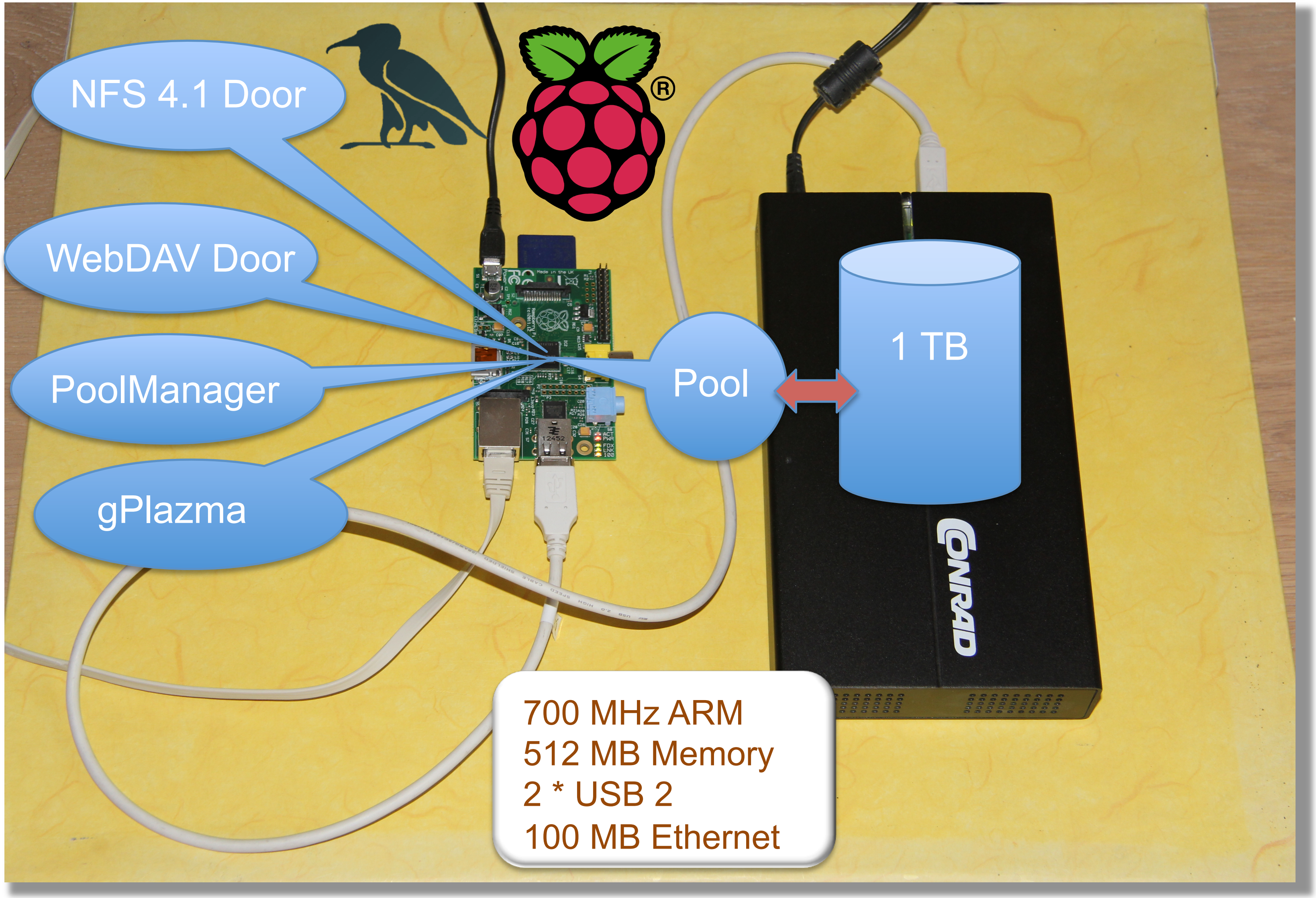
4 Countries

One dCache

Slide stolen from Mattias Wadenstein, NDGF

To very likely the smallest

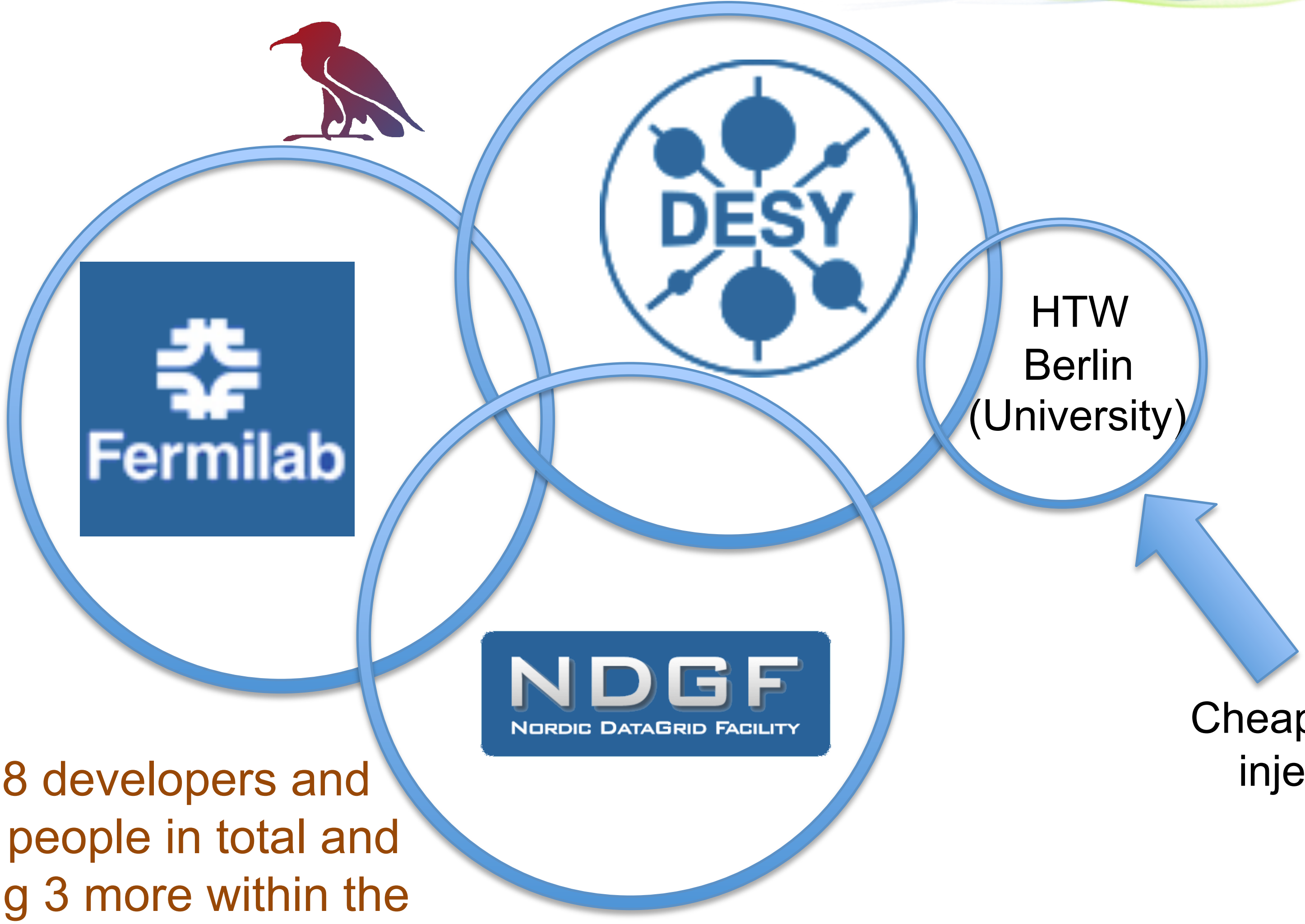
One Machine – One Process



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

3 slides on dCache.org

What’s dCache.org



HTW
Berlin
(University)

Cheap labor
injector

About 8 developers and support people in total and expecting 3 more within the next 3 months.

dCache.org networking

EGI
European Grid Infrastructure

OSG
Open Science Grid (US)

RDA
Research Data Alliance

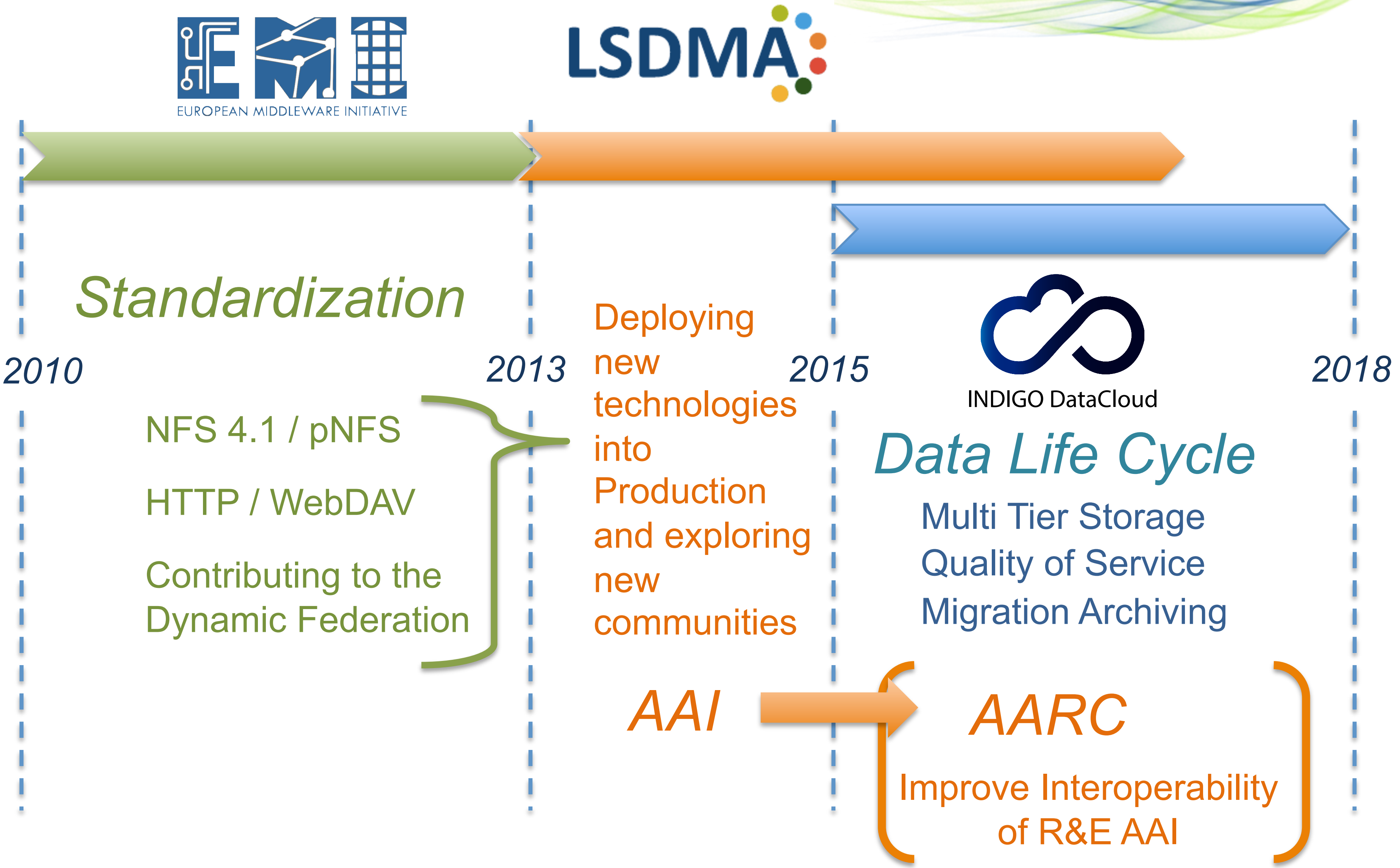
NeiC
Nordic e-Infrastructure Collaboration

LSDMA
Large Scale Data Management And Analysis

WLCG
World Wide LHC Computing Group

Funding and Objectives

dCache.org



EMI EUROPEAN MIDDLEWARE INITIATIVE

LSDMA

2010

2013

2015

2018

Standardization

NFS 4.1 / pNFS

HTTP / WebDAV

Contributing to the Dynamic Federation

Deploying new technologies into Production and exploring new communities

AAI

INDIGO DataCloud

Data Life Cycle

Multi Tier Storage
Quality of Service
Migration Archiving

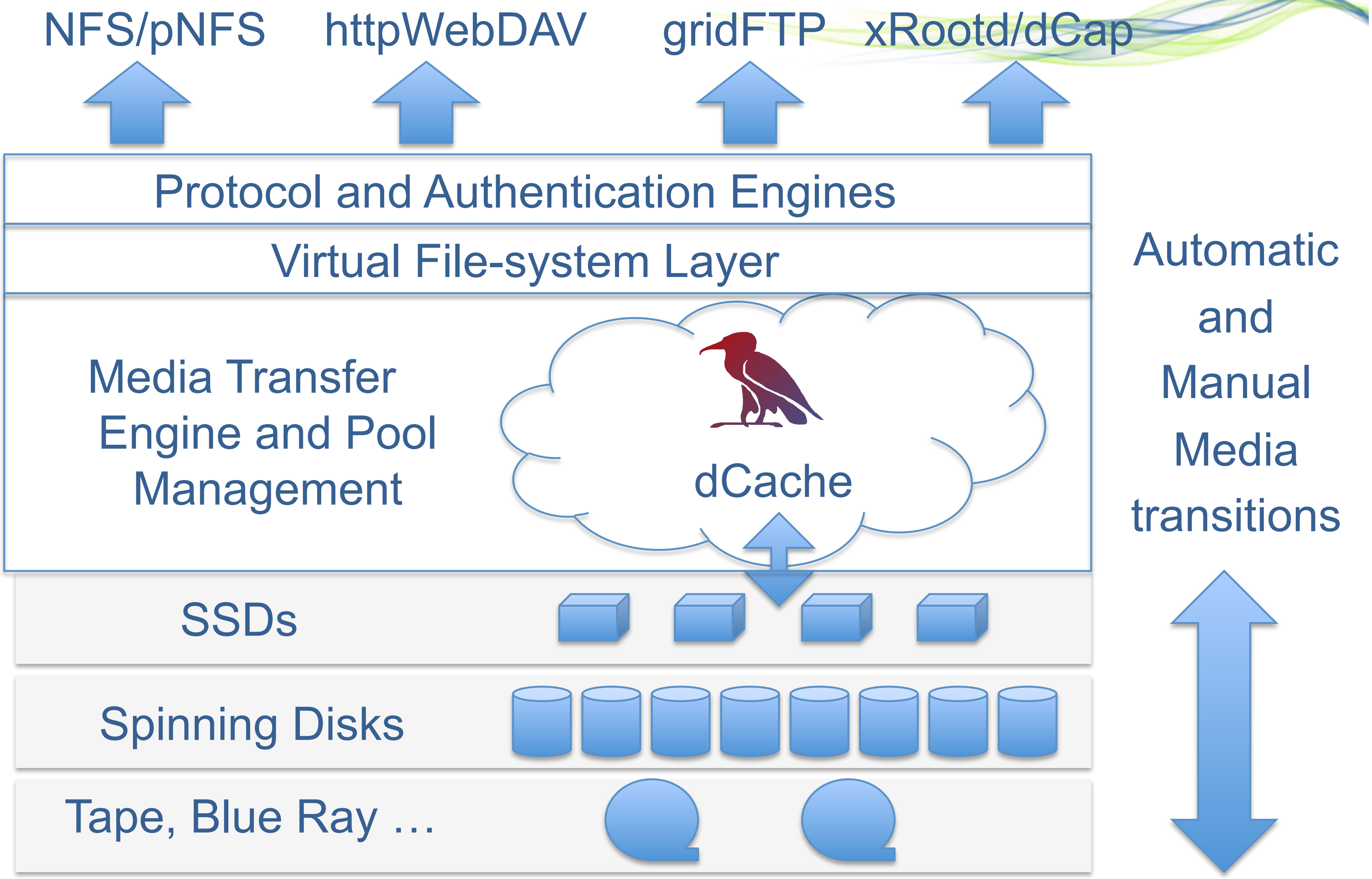
AARC

Improve Interoperability of R&E AAI

Back to technology

dCache spec for Dummies

dCache.org



NFS/pNFS

httpWebDAV

gridFTP

xRootd/dCap

Protocol and Authentication Engines

Virtual File-system Layer

Media Transfer Engine and Pool Management

dCache

SSDs

Spinning Disks

Tape, Blue Ray …

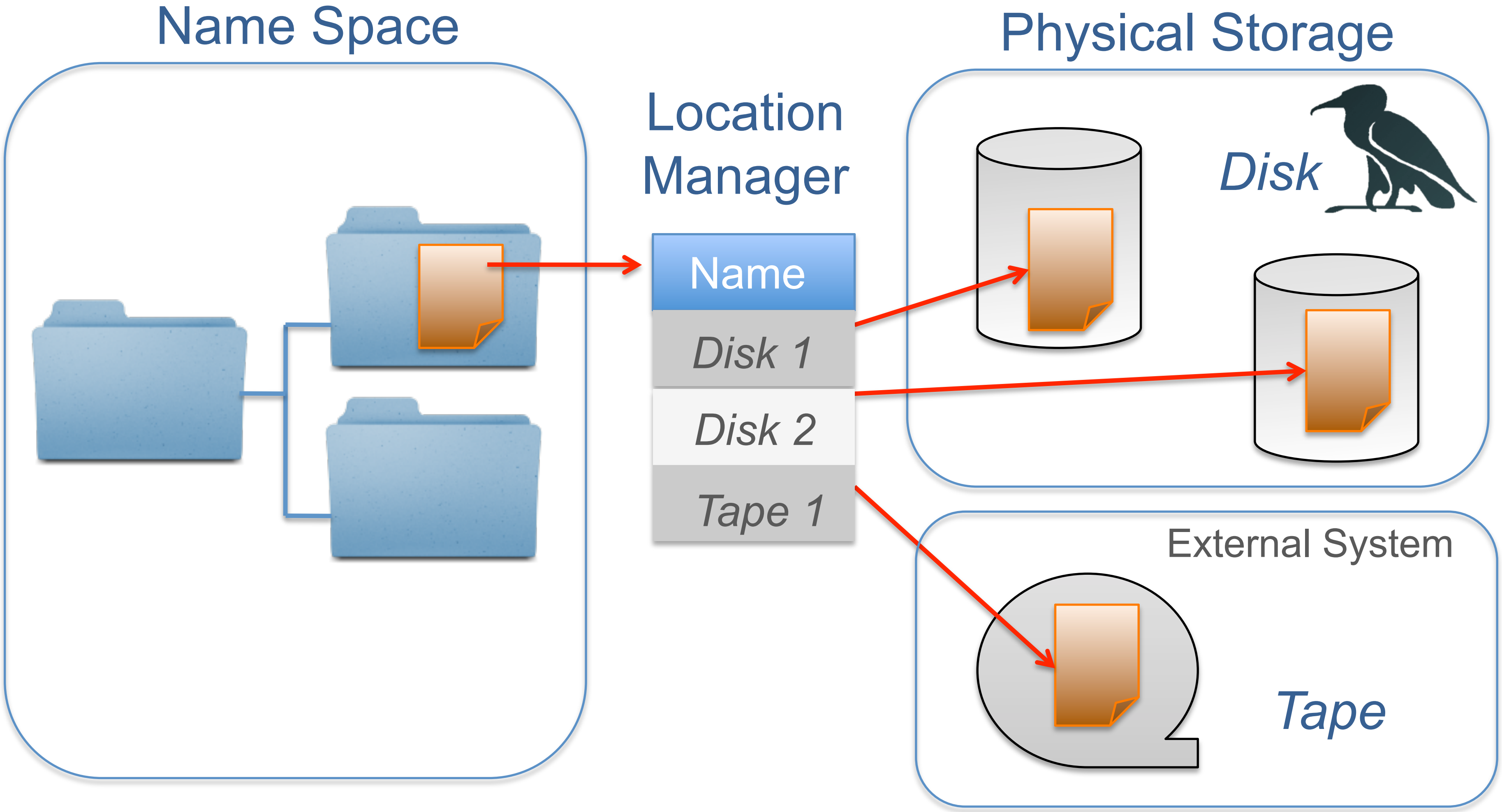
Automatic and Manual Media transitions

In other words

- Files are stored as objects on various data back-ends (Hardsdisk, SSD, Tape)
- Back-ends can be highly distributed, even beyond country bounderies.
- The File namespace engine is independent of the data storage itself.
- File object location manager keeps track of copies on the various media.

Design

Namespace – Storage separation



Resulting Features

- Hot Spot detection
  - Files are copied from ‘hot’ to ‘cold’ pools
- Multi Media Support
  - File location is based on access profile and storage media type/properties
    - Fast streaming from spinning disks
    - Fast random I/O from SSD’s
- Migration Module(s)
  - Files can be manually/automatically moved or copied between pools.
  - Rebalancing of data after adding new (empty) pools.
  - Decommission pools.
- Resilient Manager
  - Keeps max ‘n’ min ‘m’ copies of a file on different machines.
  - System resilient against pool failures.
- Tertiary System connectivity (Tape systems)
  - Data is automatically migrating to tape.
  - Data is restored from tape if no longer on disk

And what ?

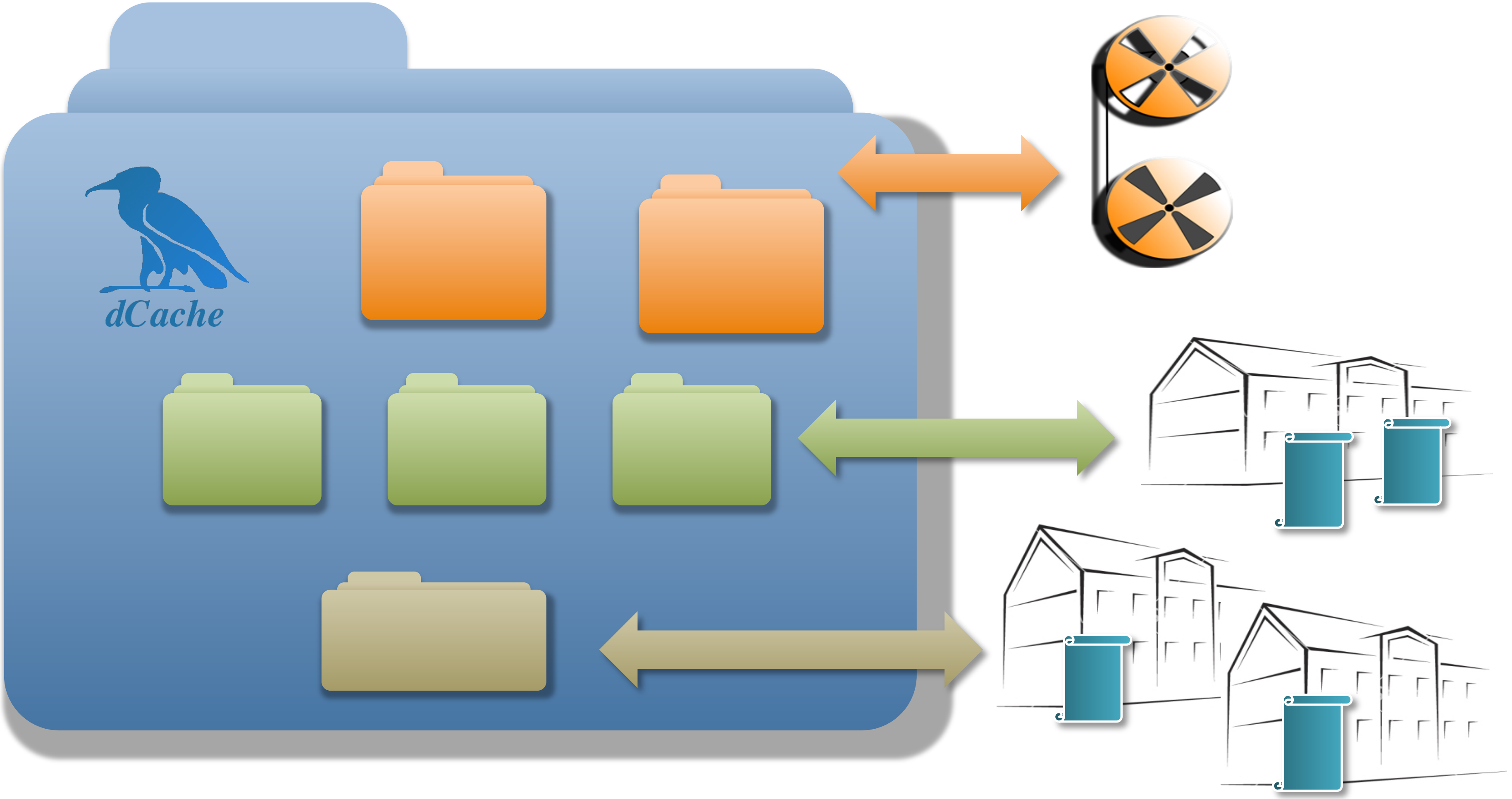
- Why do we need those features ??
- They are the basis for
  - Software defined Storage
  - Quality of Service Management
    - Defining data access latency
    - Defining data retention policies
  - Data Life Cycle support

So, what do we get ?

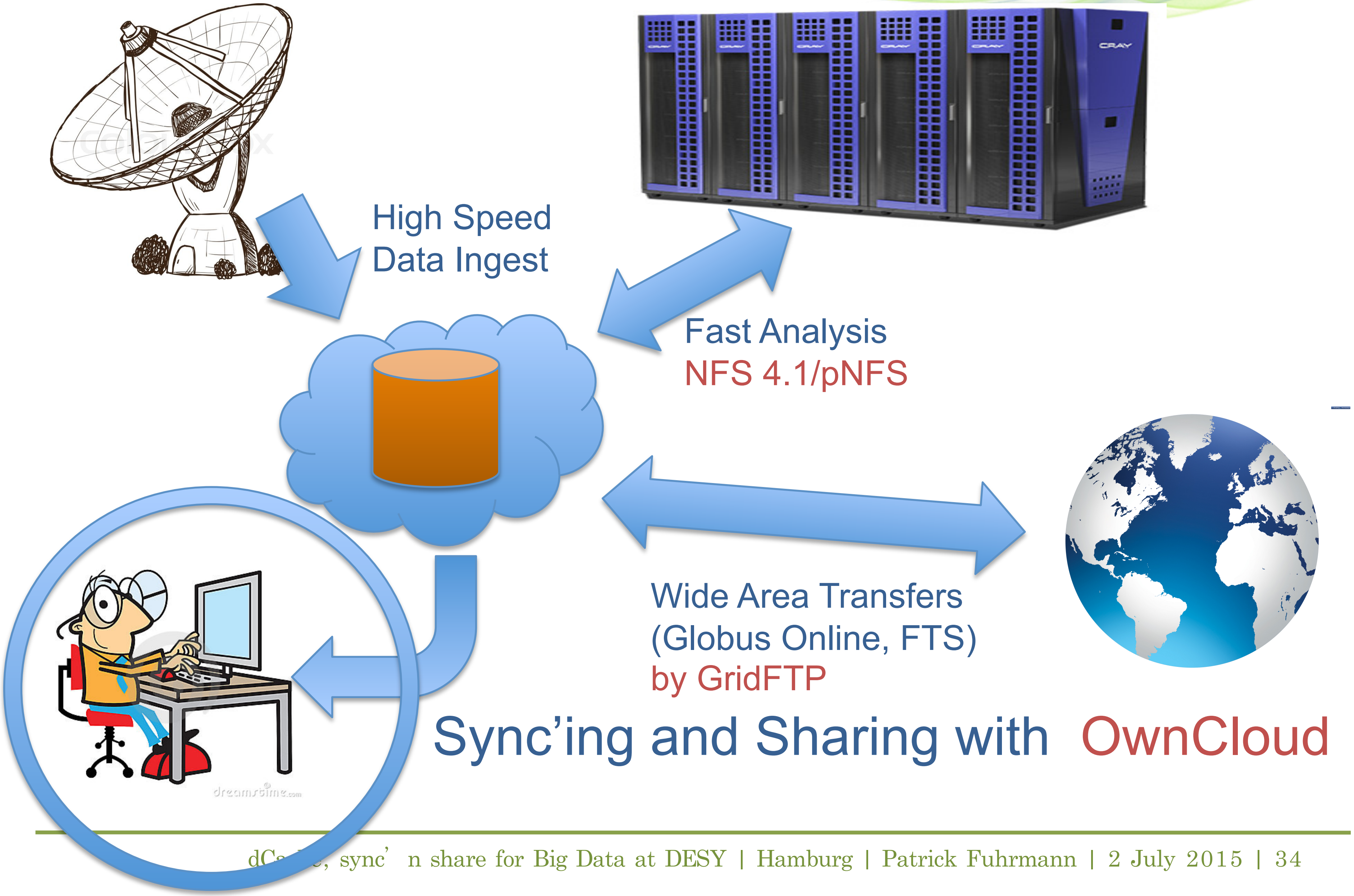
- Through Own Cloud
  - Sync'ing
  - Sharing
- Through dCache
  - Multi protocol support
  - Quality of service (Software defined storage)

Quality of service

My dCache XXL Home



Scientific Data Flow



High Speed Data Ingest

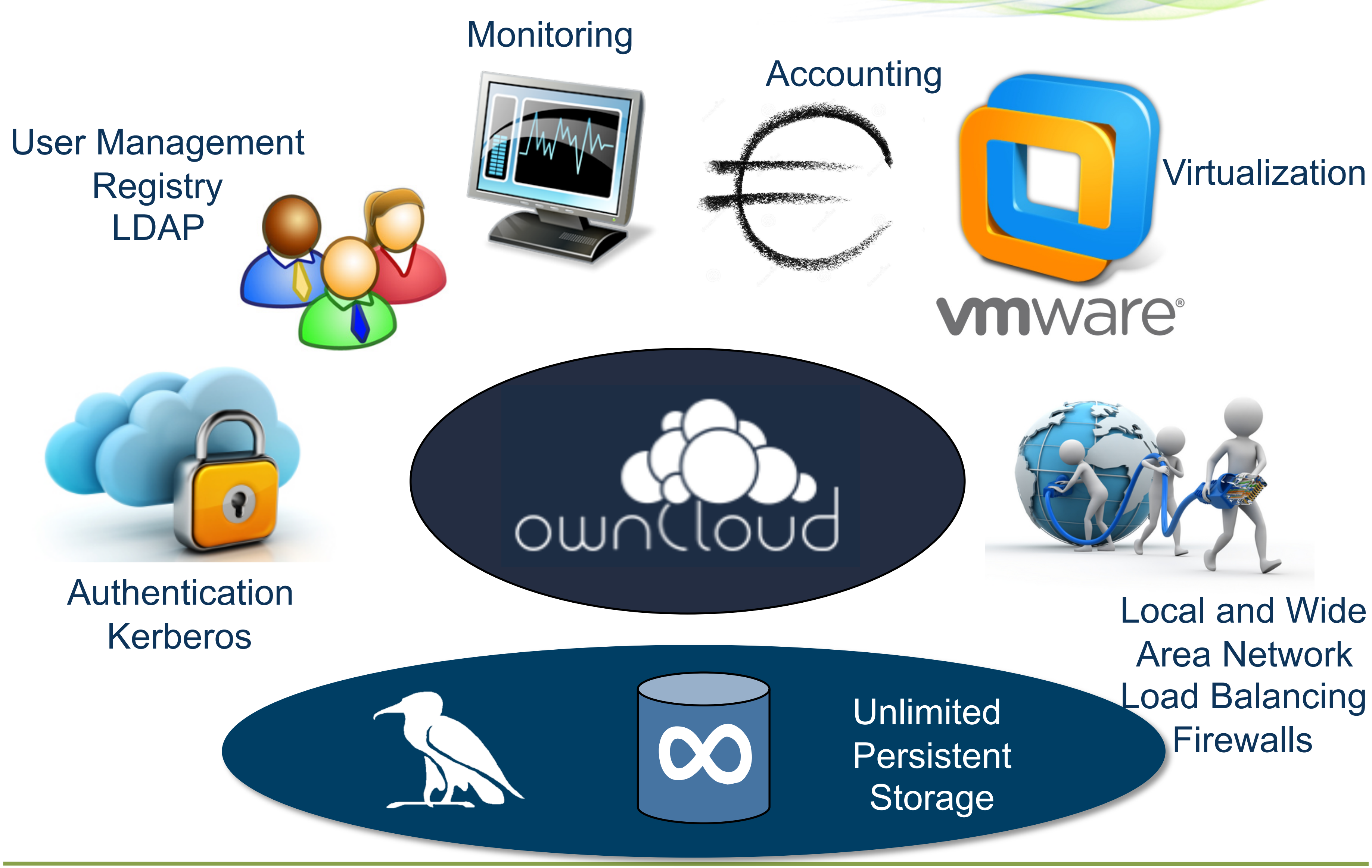
Fast Analysis
NFS 4.1/pNFS

Wide Area Transfers
(Globus Online, FTS)
by GridFTP

Sync’ing and Sharing with OwnCloud

How is that implemented at DESY ?

Integration into the DESY infrastructure



Monitoring

Accounting

Virtualization

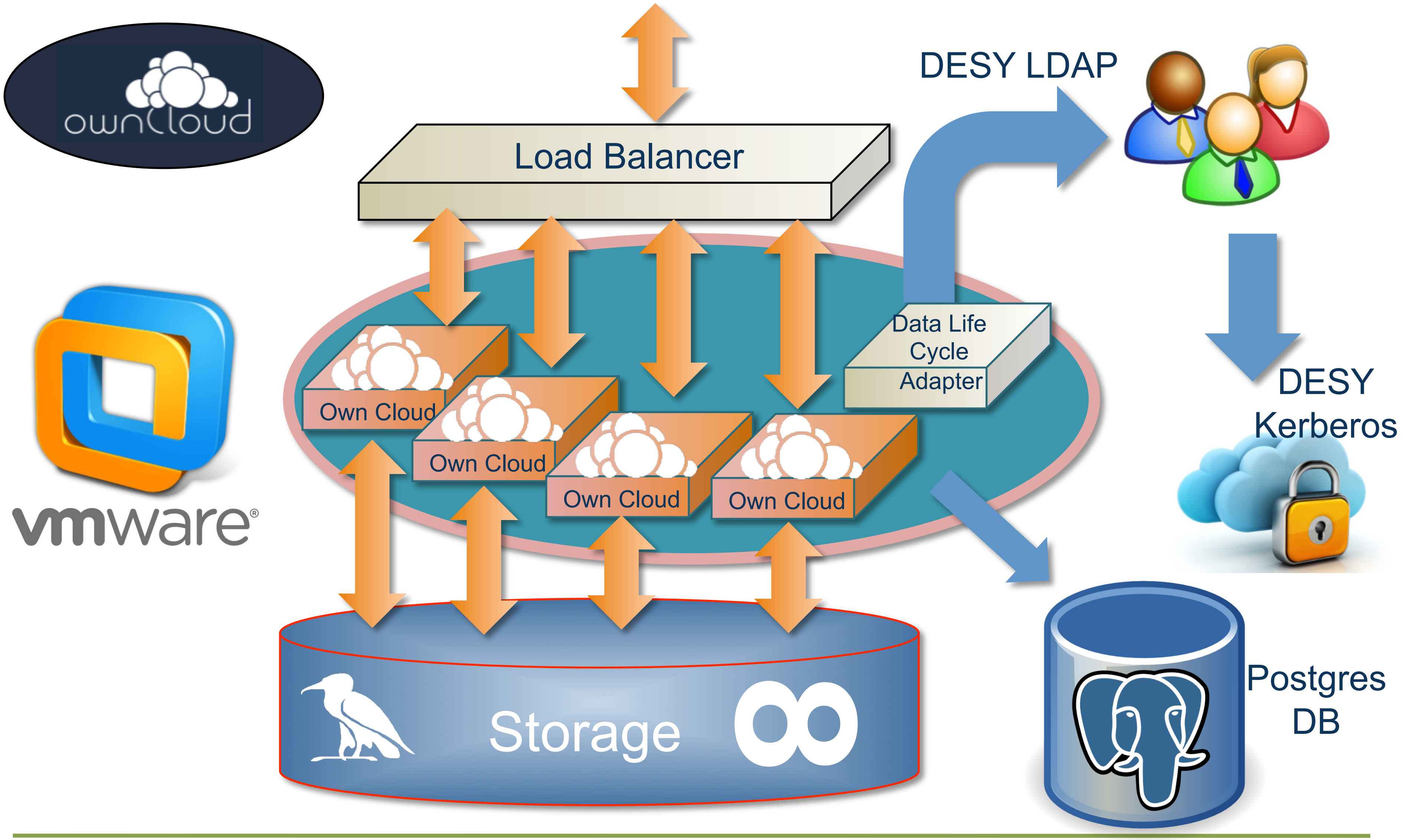
User Management
Registry
LDAP

Authentication
Kerberos

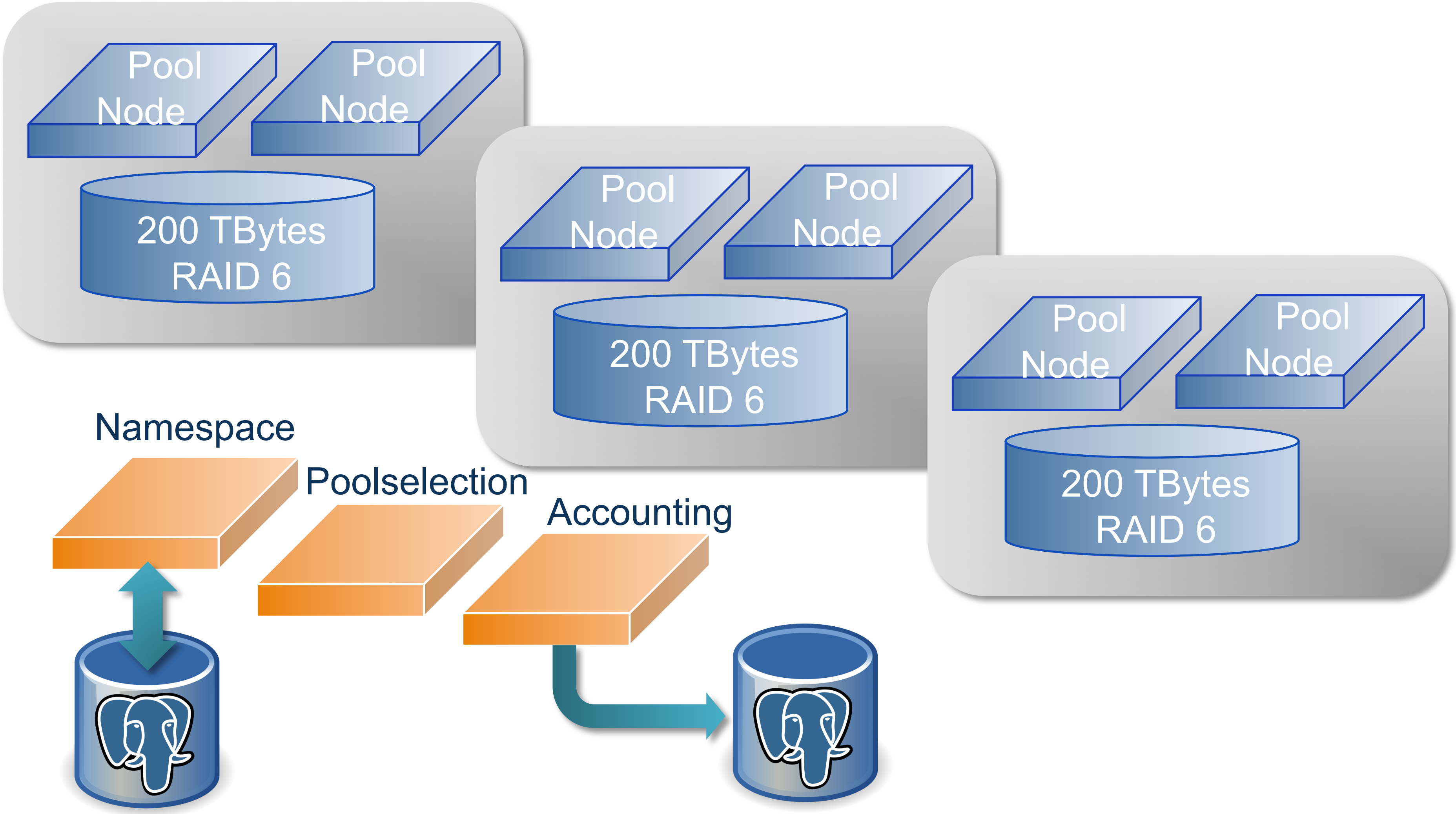
Local and Wide
Area Network
Load Balancing
Firewalls

Unlimited
Persistent
Storage

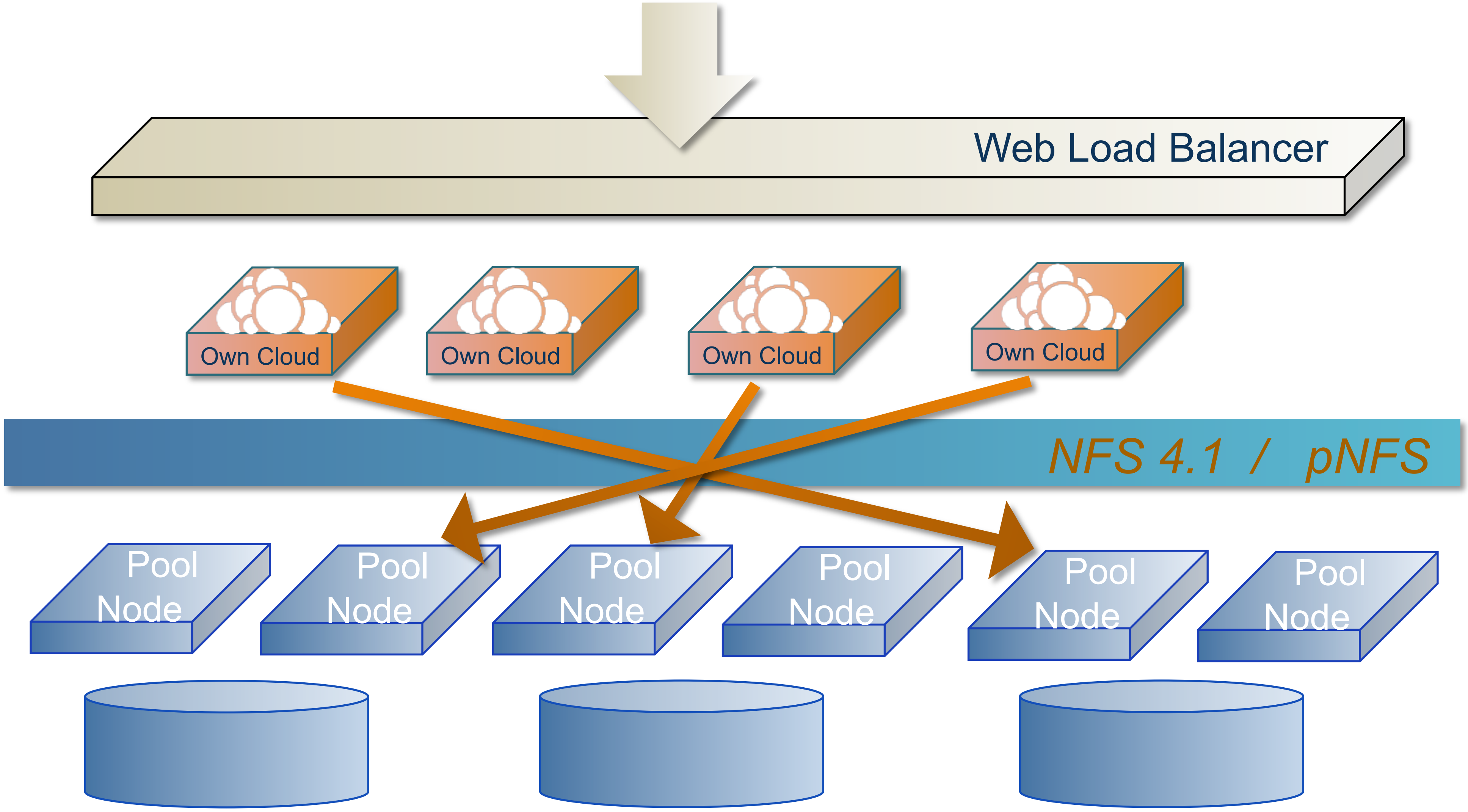
The Own Cloud Part



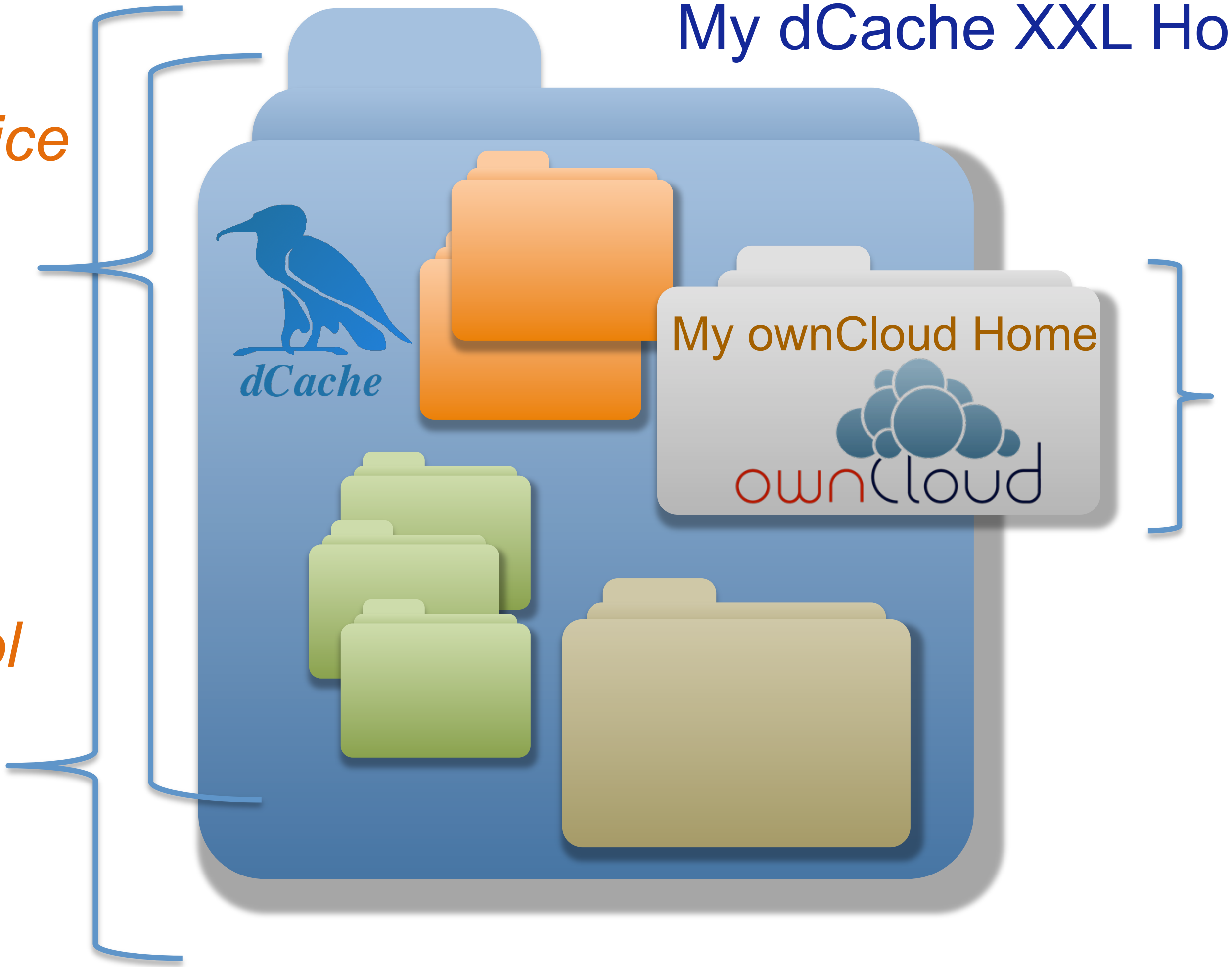
The dCache part



The horizontal scaling



‘HOME’ from user perspective



Quality of Service

Tape
Low Latency
Archive
Scratch

Multi Protocol

NFS 4.1/pNFS
GridFTP
WebDAV
SRM

Sync
Share
Web 2.0

Summary

- With dCache and OwnCloud, DESY offers a first prototype of a Scientific Cloud service, providing:
  - User specified Storage Properties (QoS)
    - Access Latency, Retention Policies
  - A variety of access protocols
    - Http/WebDAV, GridFTP, SRM, NFS 4.1 (CDMI)
  - Multiple Authentication mechanism
    - X509 Certificates, Kerberos, User/Password (SAML)
  - Sync and share
  - Web Browser access

The END

further reading

www.dCache.org

Response to ceph

- CEPH complements dCache perfectly.
  - Simplifies operating dCache disks.
  - dCache accesses data as object-store anyway already.
- dCache is evaluating a ‘two step approach’.
  - Each pools sees it own object space in CEPH
  - All pools have access to the entire space, which is a slight change of dCache pool semantics.
- Would merge CEPH and dCache advantages
  - Multi Tier (Tape, Disk, SSD)
  - Multi protocol support for a common namespace.
    - All protocols see the same namespace
  - All the dCache AAI features
    - Support for X509, Kerberos, username/password