

dCache, managing Quality of Service in Cloud Storage Infrastructures



For ISGC '17 in Taipei



Patrick Fuhrmann

On behalf of the project team



That this presentation about ?

- Some remarks on the team, funding and high level objectives.
- How to build reliable cloud systems
 - Full High Availability Functionality
 - Managing underlying storage; delegate storage (CEPH)
 - Storage Federations
- Managing your storage
 - Quality of storage management (Admin View)
 - Web 2.0 access
 - How to integrate sync'n share with traditional storage technologies

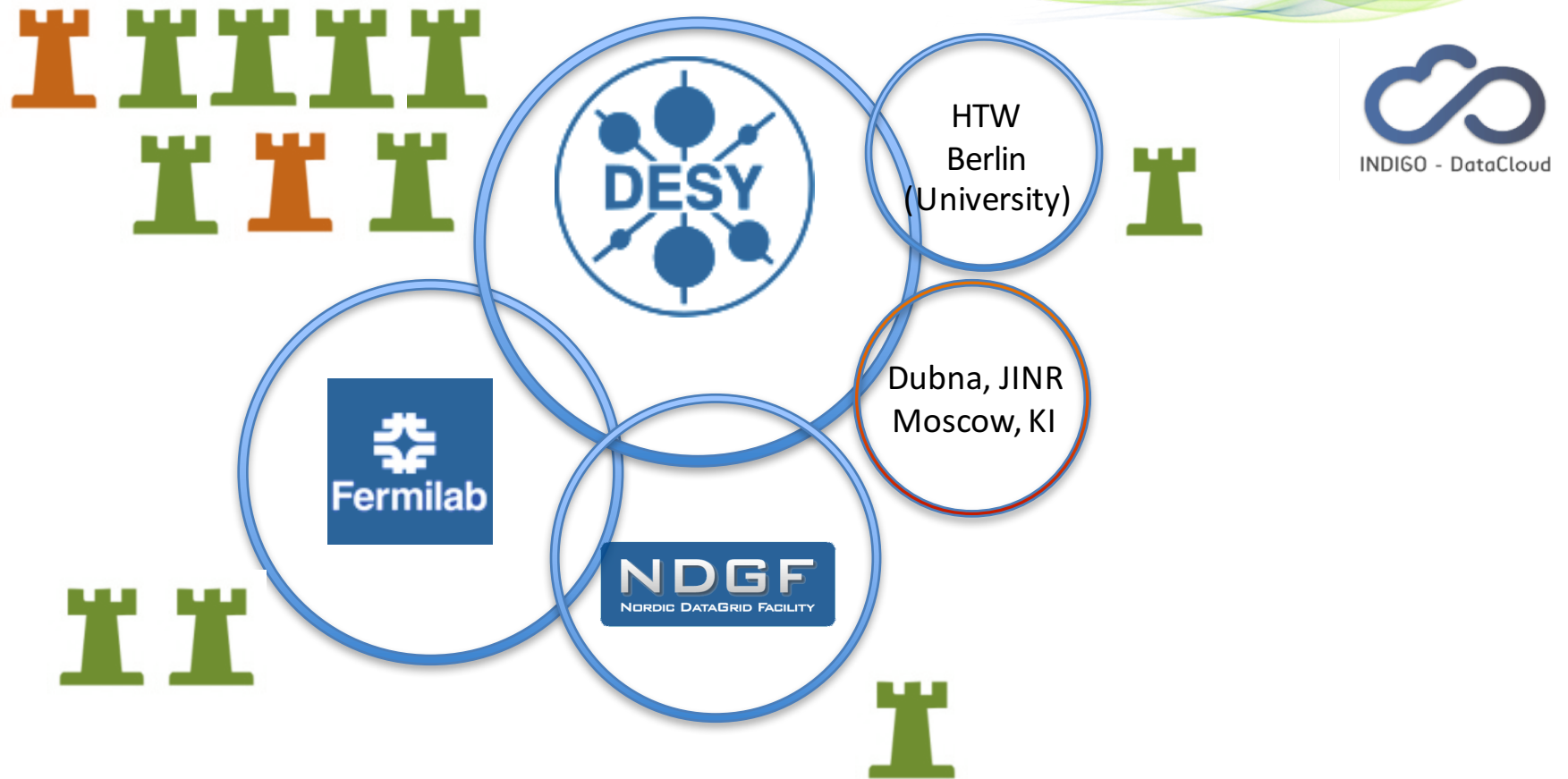
The Technology Cheat Sheet

- Combines heterogeneous storage nodes under a common virtual file system tree and scales into 100PB region.
- Provides access to data via a variety of protocols, e.g. NFS4.1, WebDAV, GridFTP, etc.
- Provides a variety of authentication mechanisms, like User/Pass, X509 Certificates, Kerberos, in preparation SAML and OpenID Connect, Macarons.
- Multi Tier support: moves data around between different media types, like Tape, Spinning Disks and SSDs.
 - By user request.
 - Automatically based on the access profile, hot spot.
- Provides resiliency, e.g. through multiple automatically managed copies.

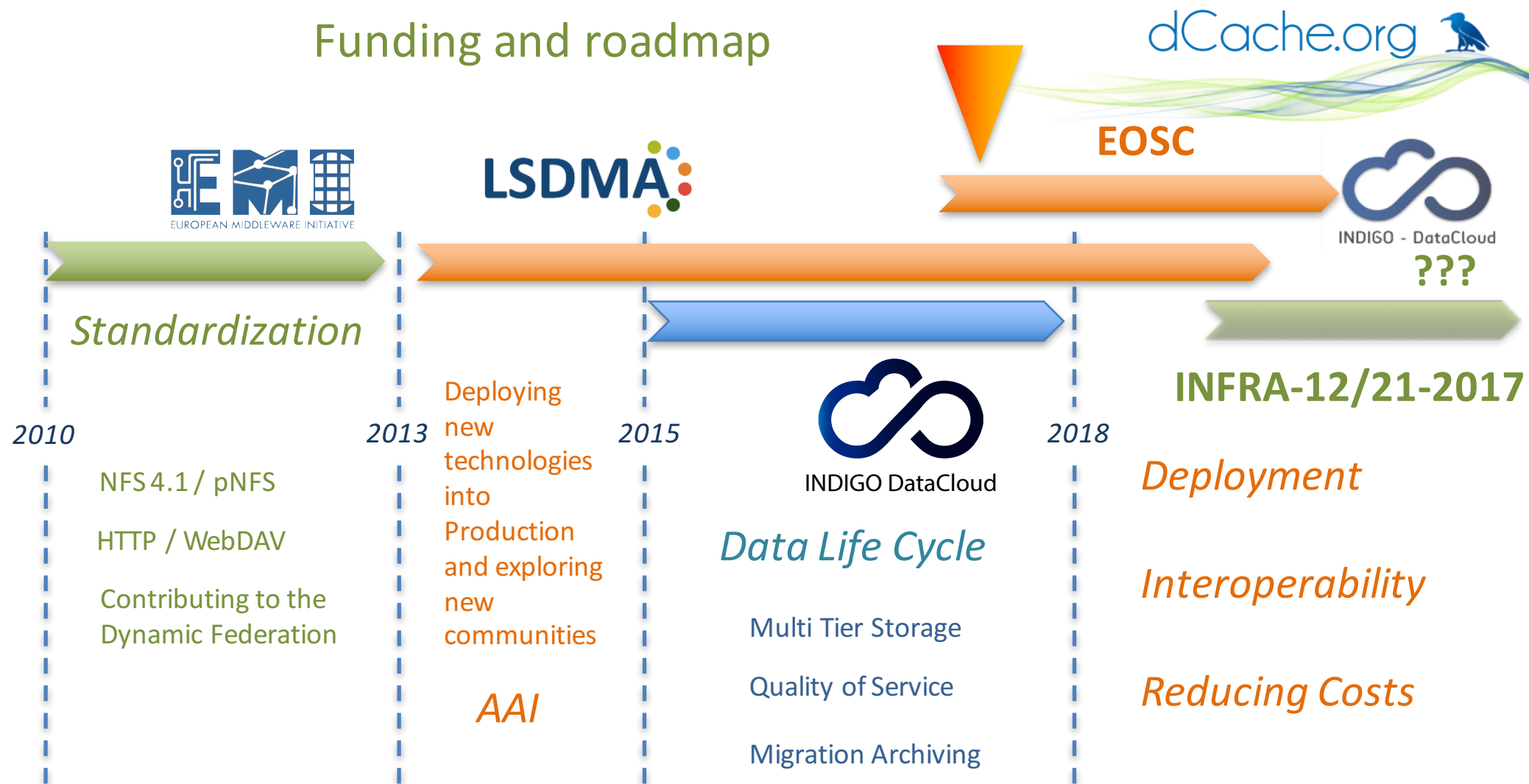


Some strategic considerations

The dCache.org collaboration



Funding and roadmap



Storage Cloud Infrastructures require

- 24/7 availability
- Rolling updates / upgrades
- Operator-less operation for 'dark computer centers'
- Reduced cost for managing low level storage
- Large scale x-site and x-country federations
- User defined quality (price) of storage



Improved System Management I

Full High Availability



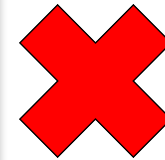
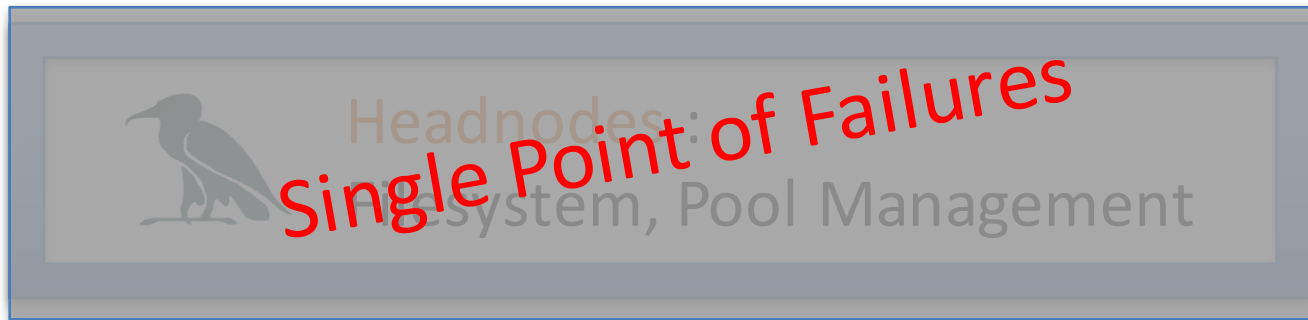
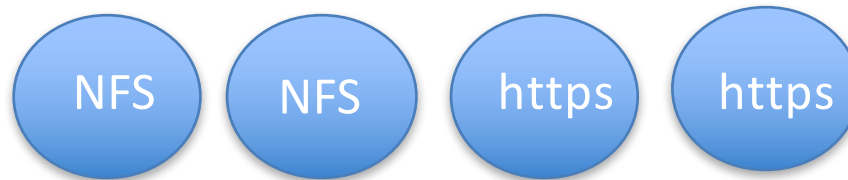
With the

- increasing size of dCache installations,
- the usage beyond WLCG, with more challenging SLAs and with
- the pressure to run 'operator less' for days
 - Weekends
 - Dark Data Centers
 - Reduce Costs

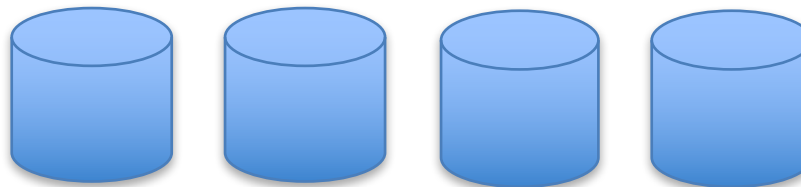
High Available functionality became inevitable.

High Availability

Redundant **Doors**
(Protocol Engines)



Redundant **Pools**
(Replica Manager)
Multiple File Copies



High Availability

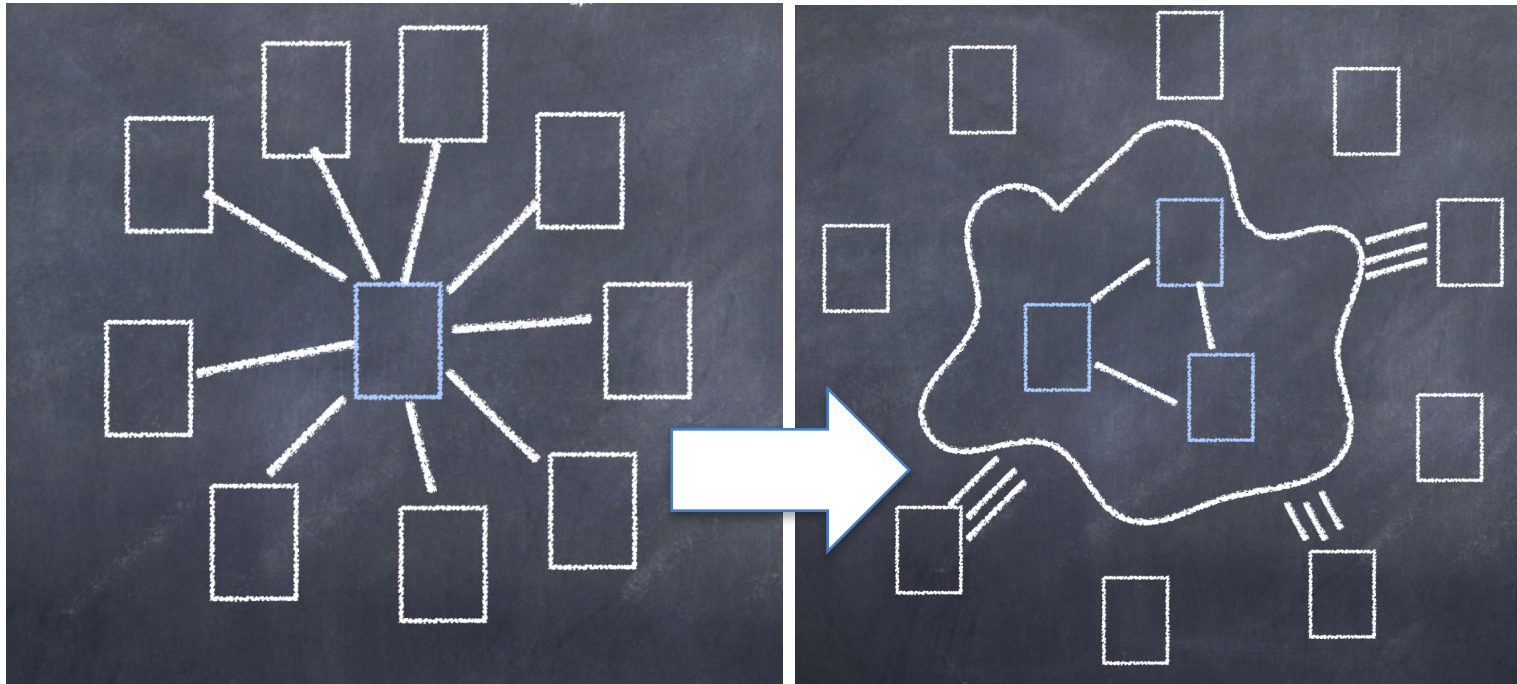
dCache.org 

Mainly the following issues need to be fixed for dCache HA



- The dCache core “message passing system” had to be fixed to overcome failures on single path segments: ‘rerouting’
- dCache sub-services needed to become redundant
 - Single point of failure services
 - state-less services

High Availability: Message Passing

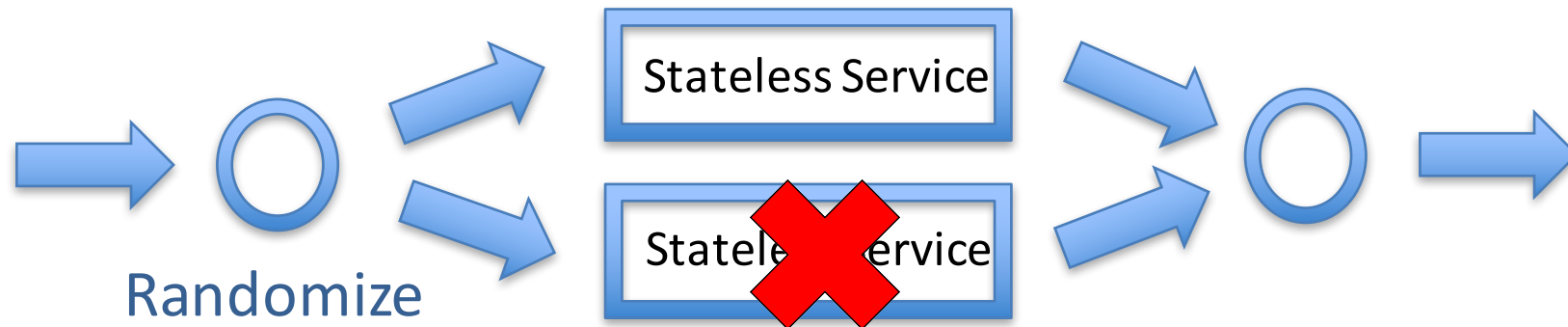


Any single component can fail, w/o breaking the service

Stolen from Gerd Behrman, NDGF

High Availability

- Stateless services : use publish subscribe



- Singletons (build quorums, e.g. using Zookeeper)



Result (With 3.0)

At any point in time, one internal service (node)
can fail without consequences for the overall
service.

Essential for a huge 24/7 installations.



Improved System Management II

Delegating low level storage functionality

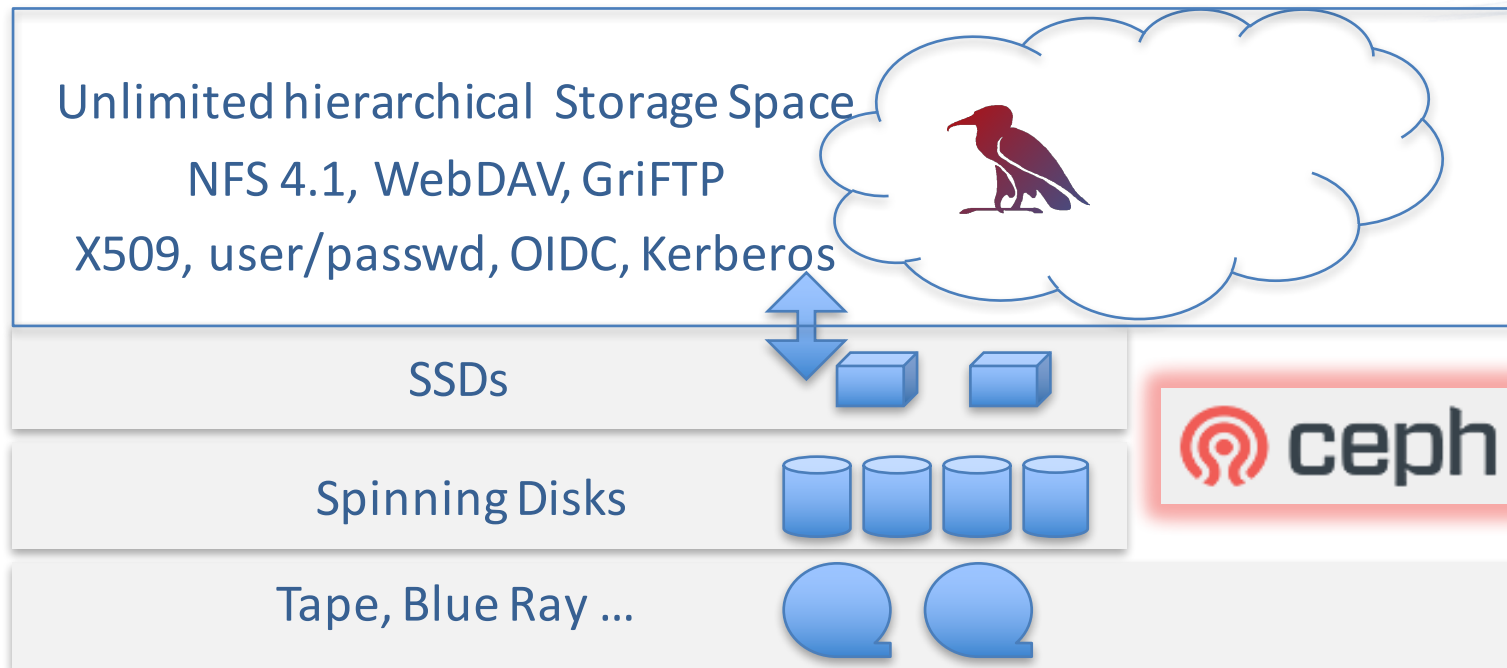
Delegating work to external Storage Layers

- Provides a single-rooted namespace.
- Metadata (namespace) and data locations are independent.
- Uniquely handles different Authentication mechanisms, like x509, Kerberos, login+password,
- Provides access to a wide variety of access protocols (WebDAV, NFSv4.1/pNFS, xxxFTP, DCAP, Xrootd, DCAP).
- Provides data migration between multiple tiers of storage (DISK, SSD, TAPE).
- Aggregates multiple storage nodes into a single storage system.
- Manages data movement, replication, integrity.

Can be delegated

Slide stolen from Tigran Mkrtchyan

Integrating low level storage technologies

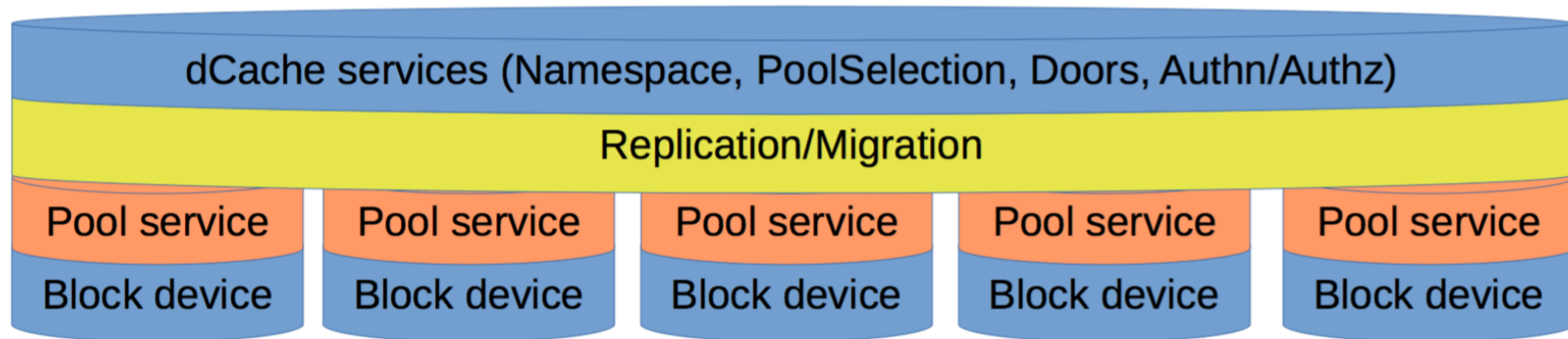


Benefits: Best of two worlds

- Low maintenance CEPH functionality
- High level dCache functionality : access protocols and authentication mechanisms

CEPH in dCache, what we have now

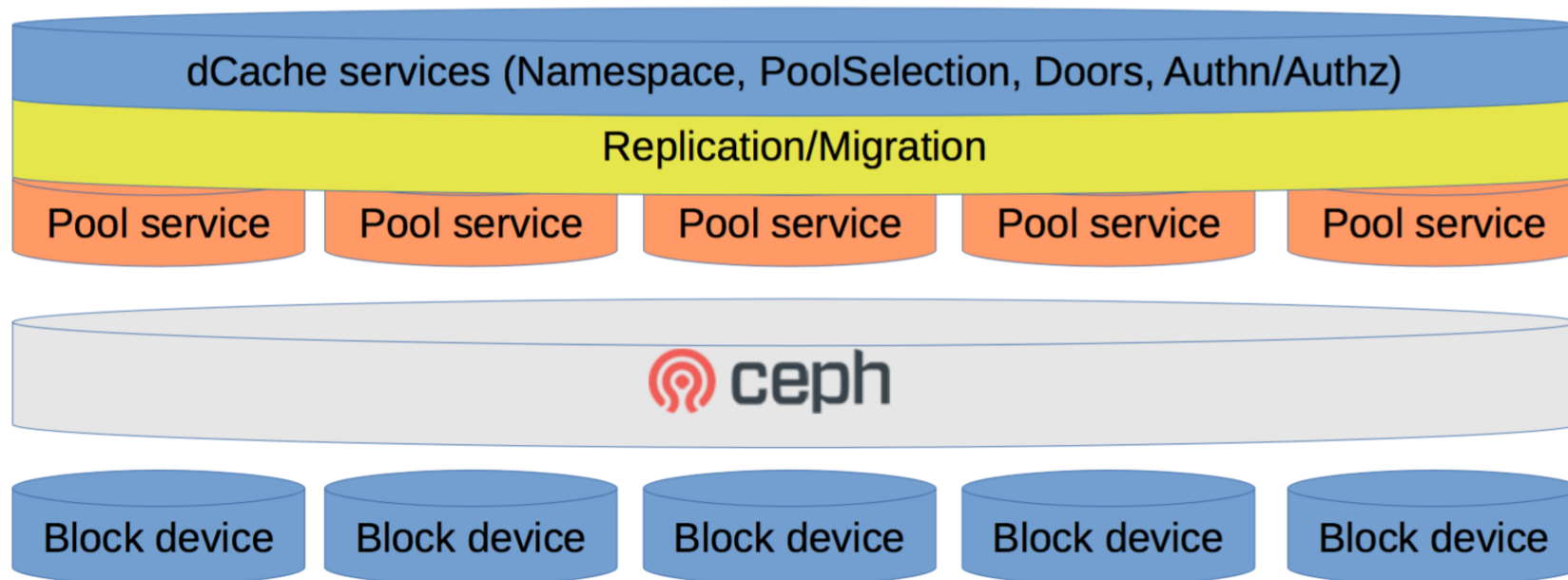
- dCache provides high level service
- Data replication and storage management is a core dCache service
- Each pool is attached to its own disk.



Slide stolen from Tigran Mkrtchyan

CEPH in dCache, phase 1

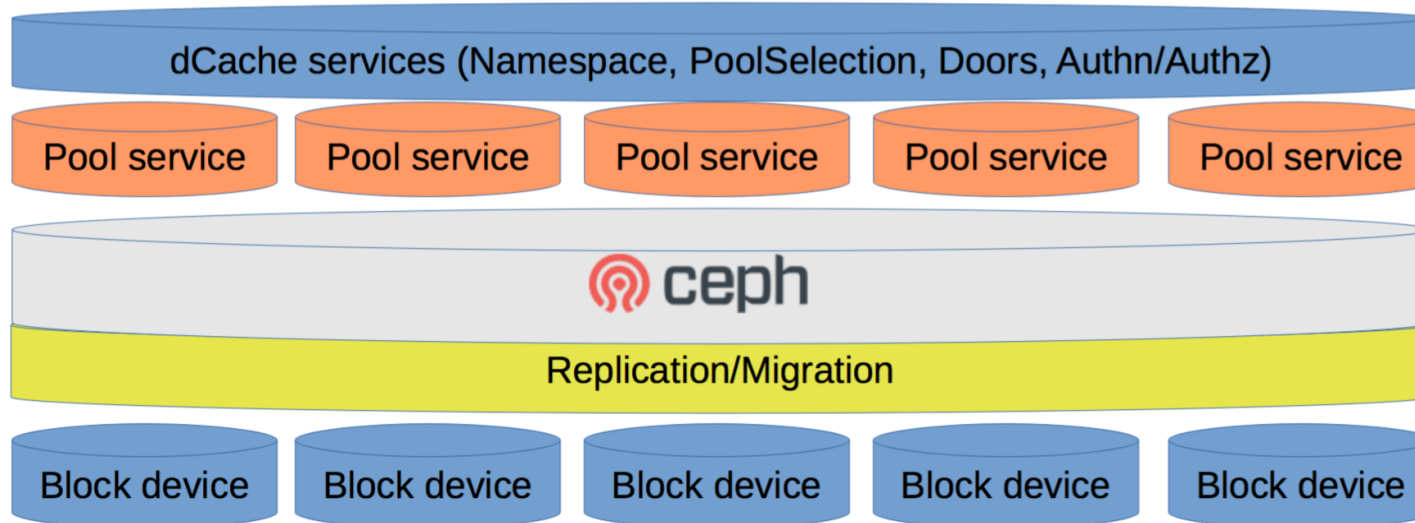
- dCache provides high level service
- Data replication and storage management is a core dCache service
- Each pool has its own partition on a shared storage system.



Slide stolen from Tigran Mkrtchyan

CEPH in dCache, phase 2

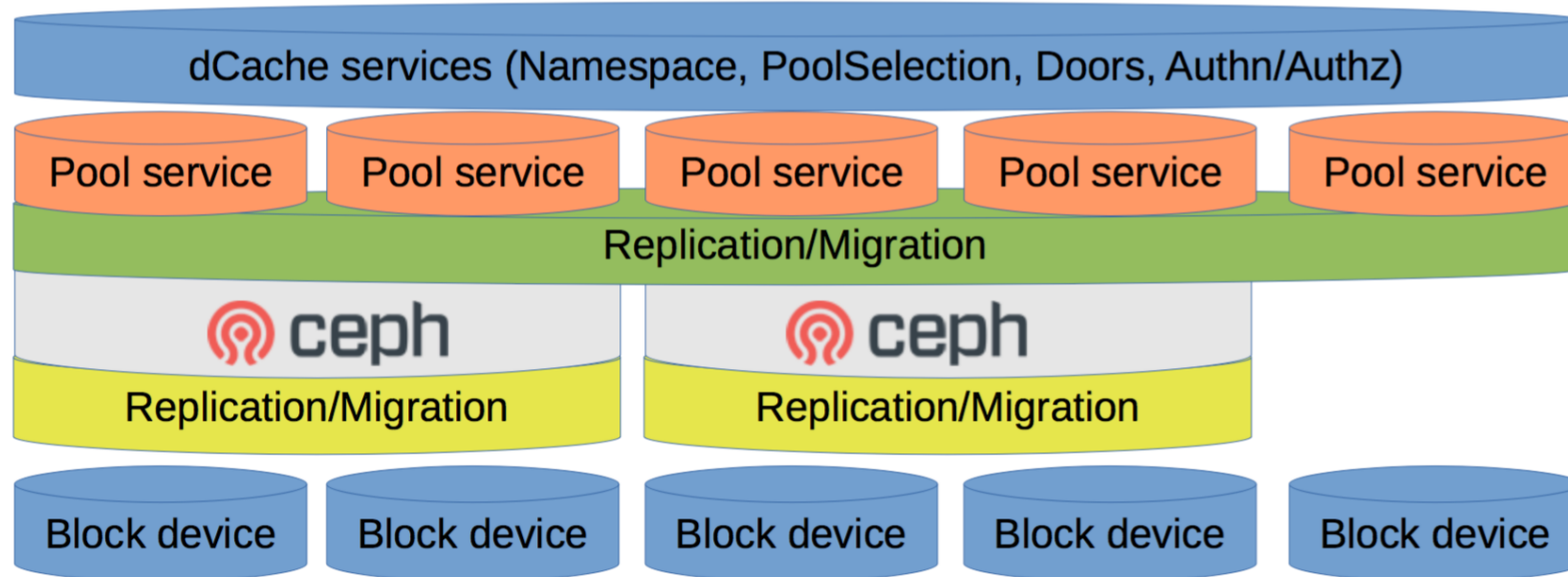
- dCache provides high level service
- All pools can see all partitions on the shared storage
- Any pool can deliver any data from any partition.
- The object storage takes care of replication and data resiliency



Slide stolen from Tigran Mkrtchyan

CEPH in dCache, phase 2

- dCache provides high level service
- dCache can move data between 'regular' and Object Store pools.



Slide stolen from Tigran Mkrtchyan

Managed Storage or Quality of Service in Storage

Why ?

Why do we need to talk about Quality of Storage ?

- Different Data Sets have different purposes
 - Scratch (Intermediate Storage)
 - Heavy Processing (HPS analysis access)
 - Long Term preservation for 100 years.
- Different storage devices or brands have different qualities
 - Tape
 - Blue Ray
 - Disk, SSD
- Different Data intensive groups have different budgets
- Different storage technologies have different prices



QoS initiative organized by INDIGO-DataCloud



Involving RDA and SNIA

dCache.org 



- HGF, Germany
 - DESY
 - KIT
- INFN, Italy
 - Bari
 - CNAF, Bologna
- Poland
 - PSNC

Pre-requisite for "Support for Quality of Service in Storage"

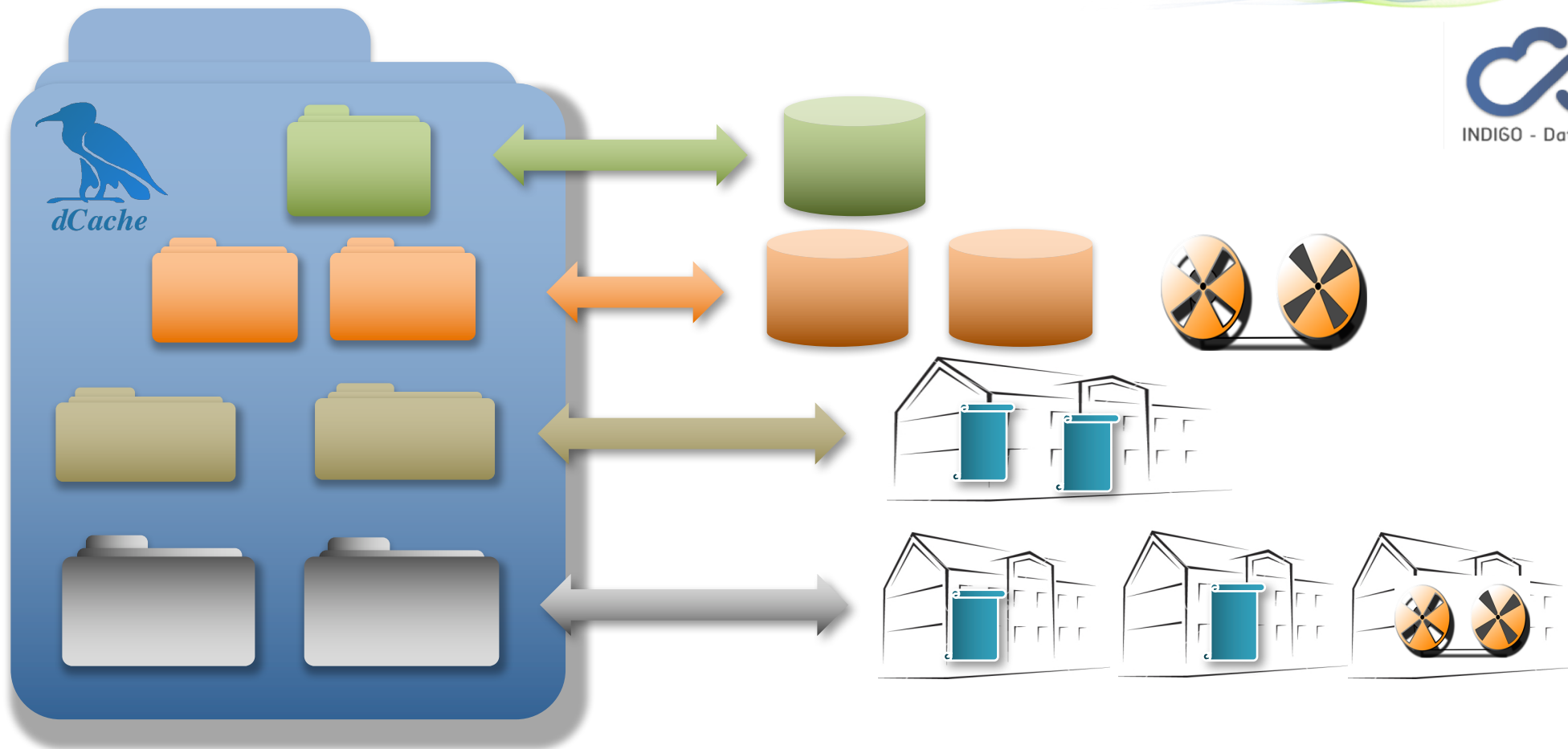


- Integrated support for different storage media
- Transparent transition between media types based on
 - Internal rules
 - APIs for remote "Platform as a Service" systems
 - Manual intervention (Operator , User)
- Consistent and transparent management of replica
 - System must be aware of media attributes (Speed, Latency, Price)
 - System must be aware of locations (Rack, Power supply group, Buildings)

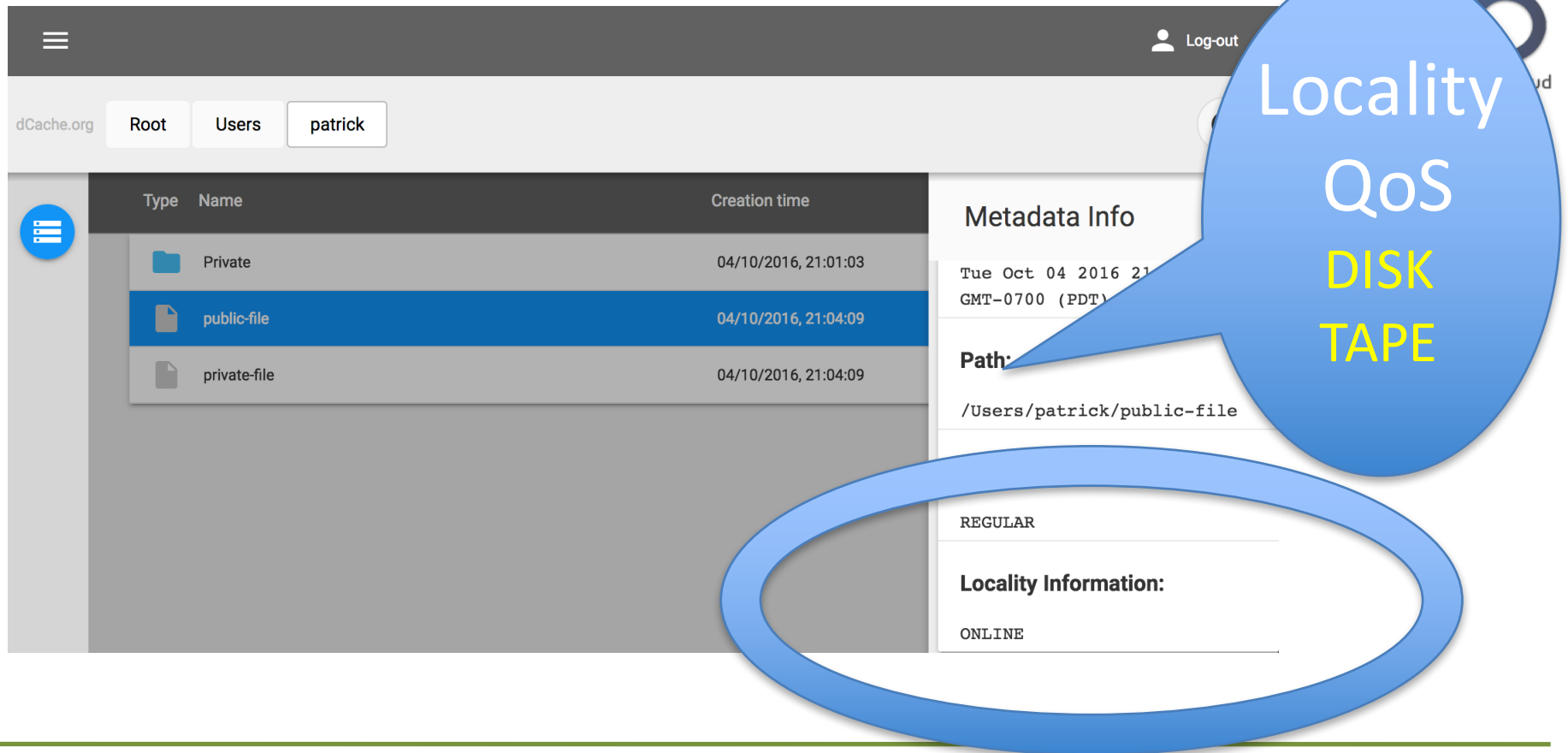
How to steer QoS ?

- SRM : Storage Resource Manager
 - exists for over 10 years
 - no longer “en vogue”
- CDMI : Cloud Data Management Interface
 - Partial support, not great in the standard definition
 - Extensions possible (INDIGO DataCloud and RDA)
 - Concept not sophisticated enough.
- RESTful
 - No standard definition (dCache proprietary)
- GUI or predefined directory structure
 - End user friendly
 - Limited functionality

Directory Based Storage Quality selection



Storage Management GUI



The screenshot shows the dCache Storage Management GUI. At the top, there is a navigation bar with 'Root', 'Users', and 'patrick' buttons. Below this is a table listing files and folders:

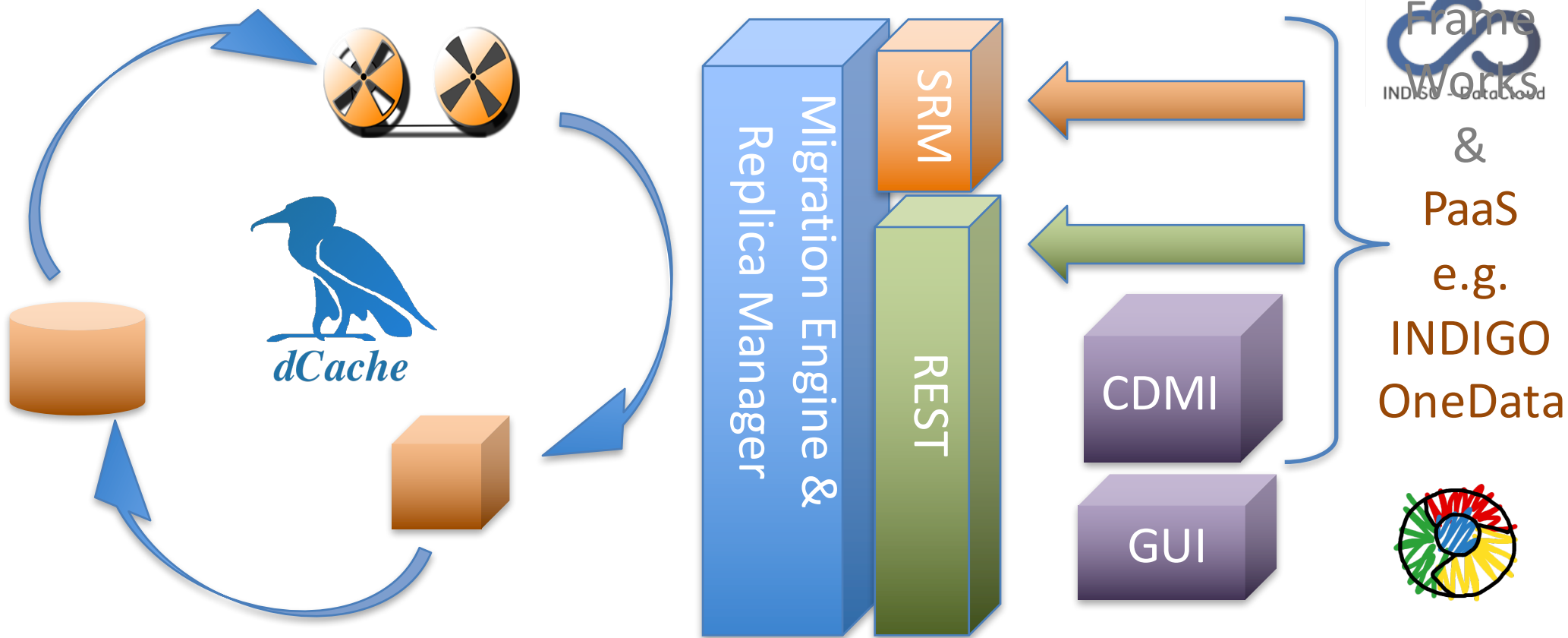
Type	Name	Creation time
Folder	Private	04/10/2016, 21:01:03
File	public-file	04/10/2016, 21:04:09
File	private-file	04/10/2016, 21:04:09

To the right of the table is a 'Metadata Info' panel. It displays the following information:

- Metadata Info
- Tue Oct 04 2016 21:04:09 GMT-0700 (PDT)
- Path: /Users/patrick/public-file
- REGULAR
- Locality Information: ONLINE

A blue speech bubble points to the 'Locality Information' field, containing the text: 'Locality', 'QoS', 'DISK', and 'TAPE'. A blue oval highlights the 'Locality Information' field in the metadata panel.

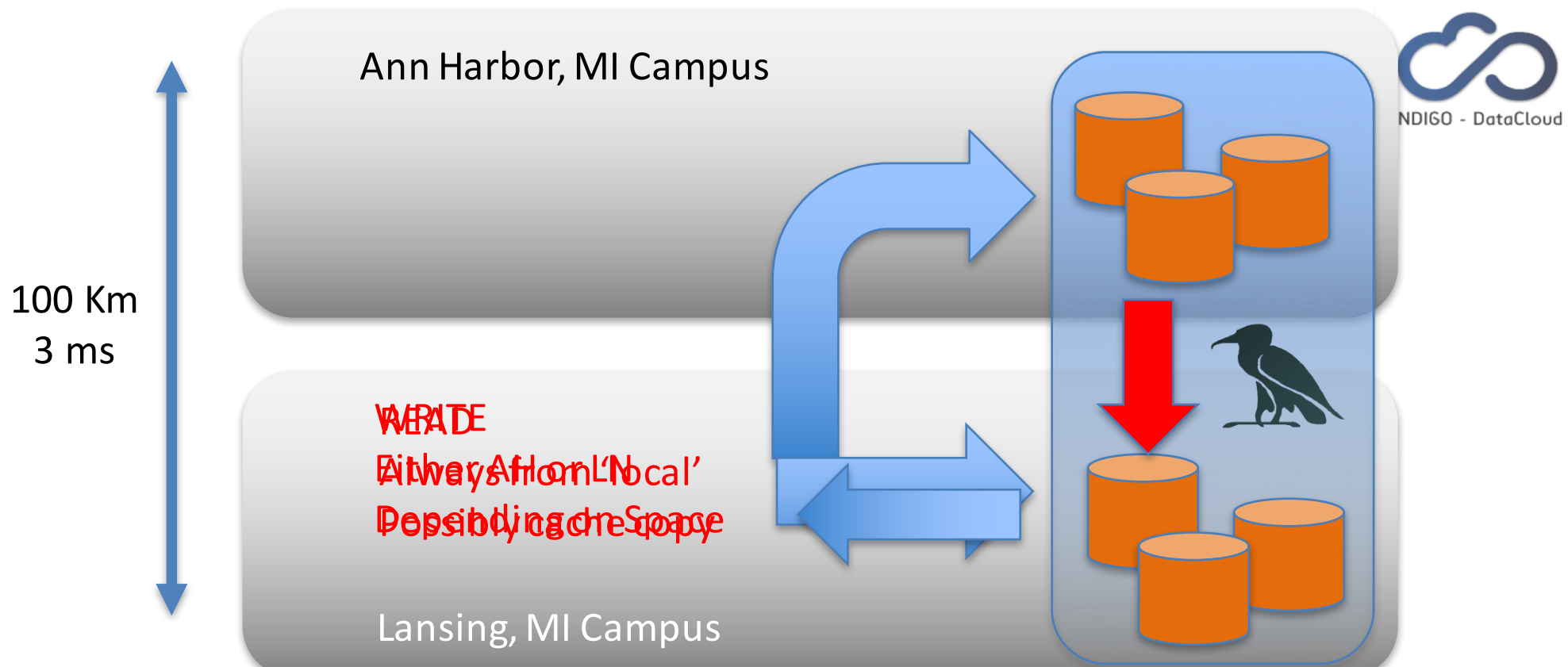
QoS Management interfaces



Managed Storage II

Managing data flows in federated systems.

Federating Storage The Michigan Setup

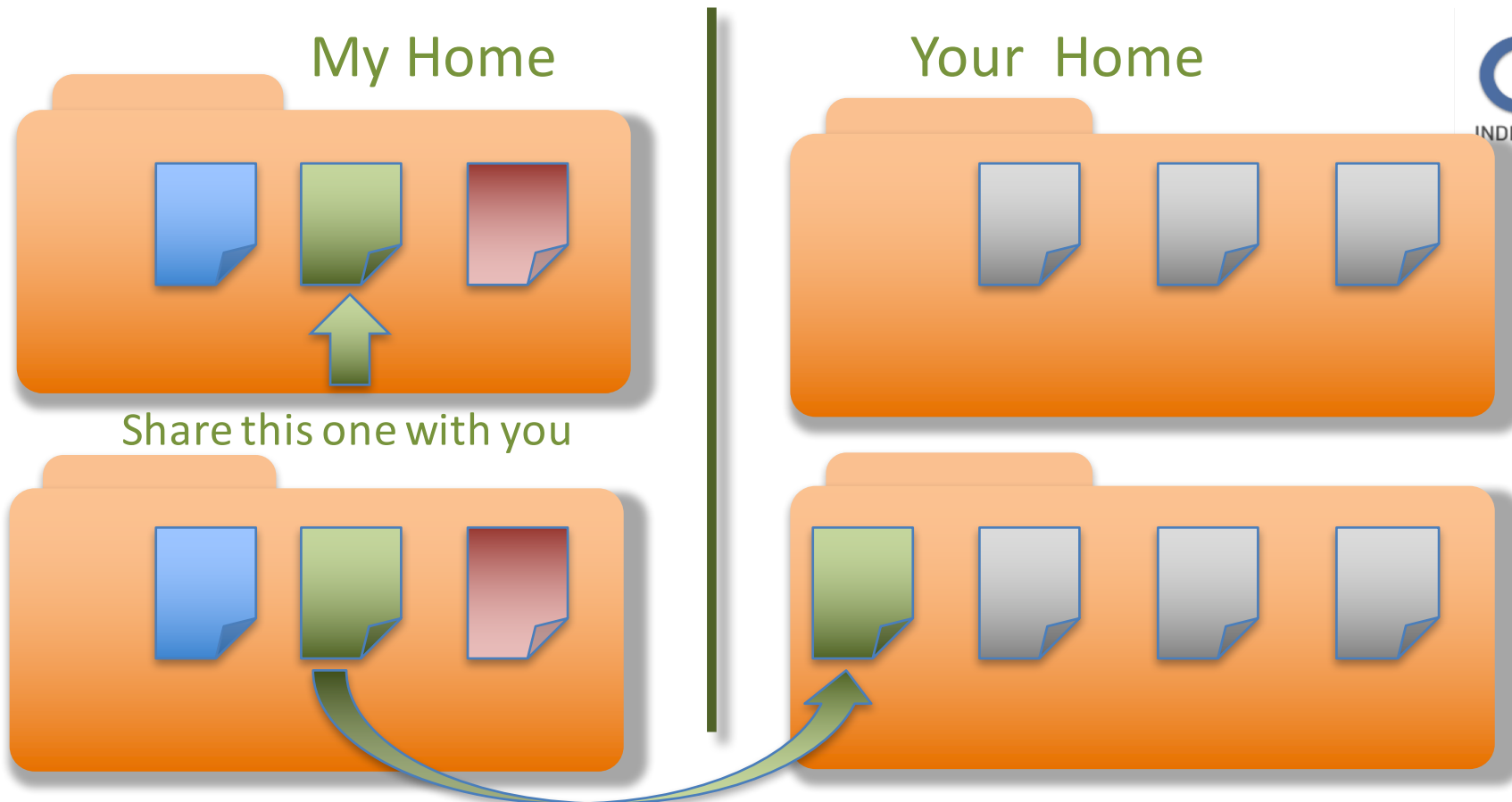


~~WRITE~~
~~Either from 'local'~~
~~positions or space~~

Provided by Shawn McKee

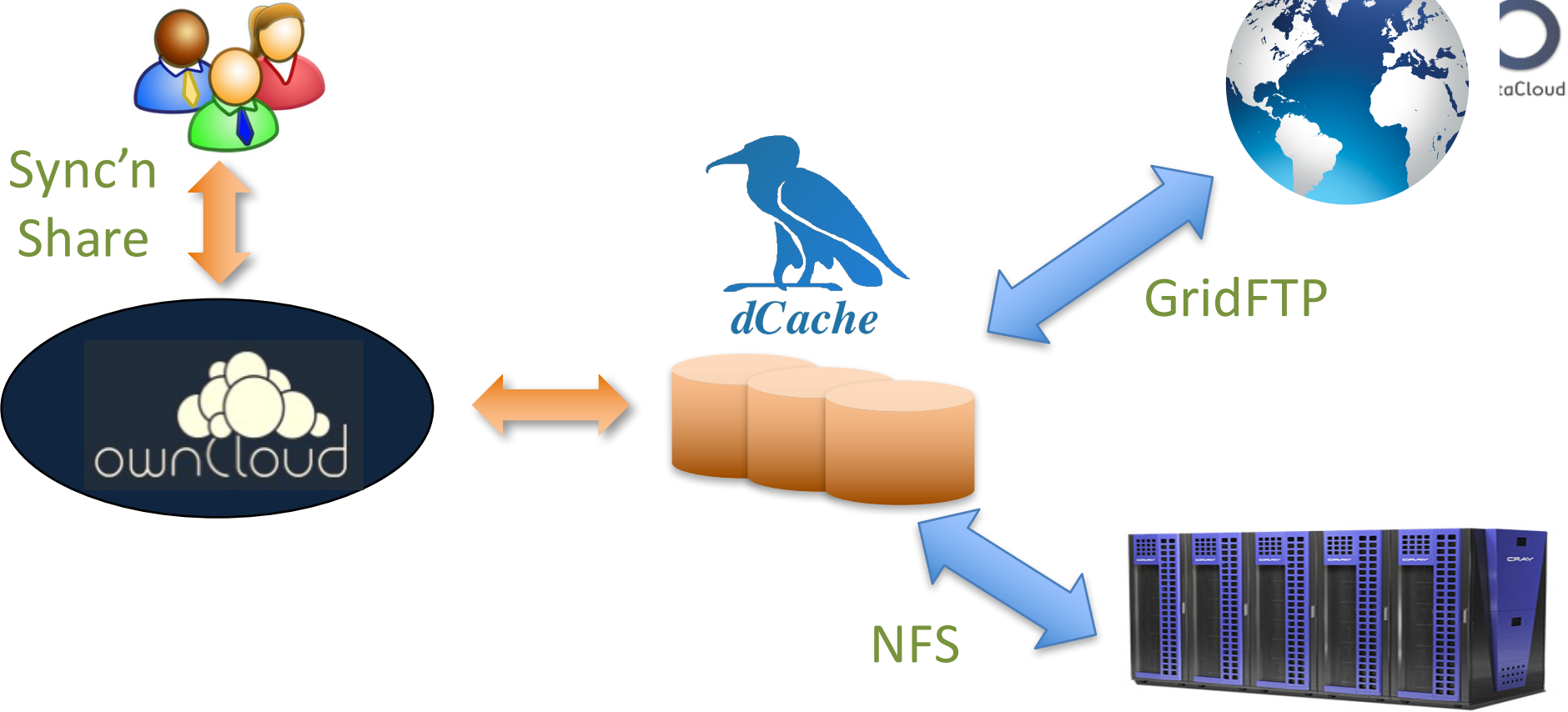
The Cloud Part

The cloud feeling, The **Sharing**



We decided to go for ownCloud © to get this done.

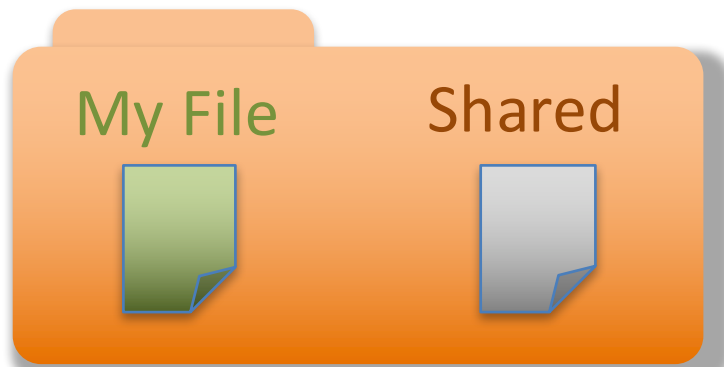
Current Implementation



Missing Bit

- Share only exists through the OwnCloud interface.
- Not available in NFS and GridFTP

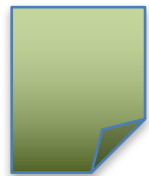
My Home in OwnCloud



My Home in NFS/GridFTP



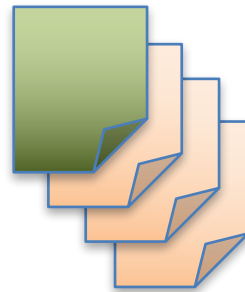
The Solution



Files



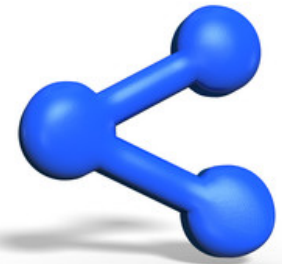
Directories



Levels



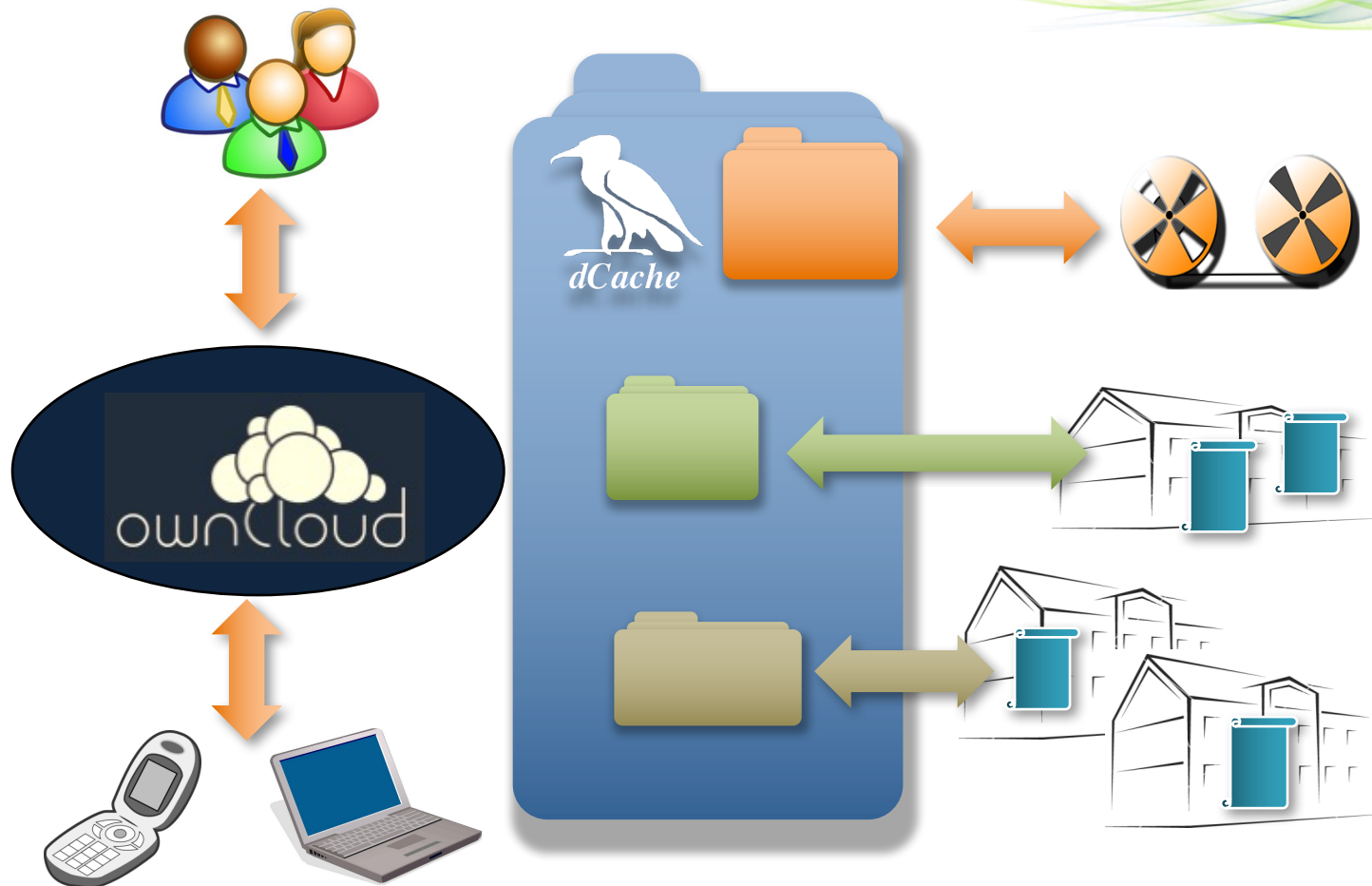
Tags



Shares

- dCache namespace gets a new object : ***The Share***
- The sync'n share interface (e.g. OwnCloud) only refers to the ***Share*** in the dCache namespace.
- ***Share*** is somewhere between hard and soft link.

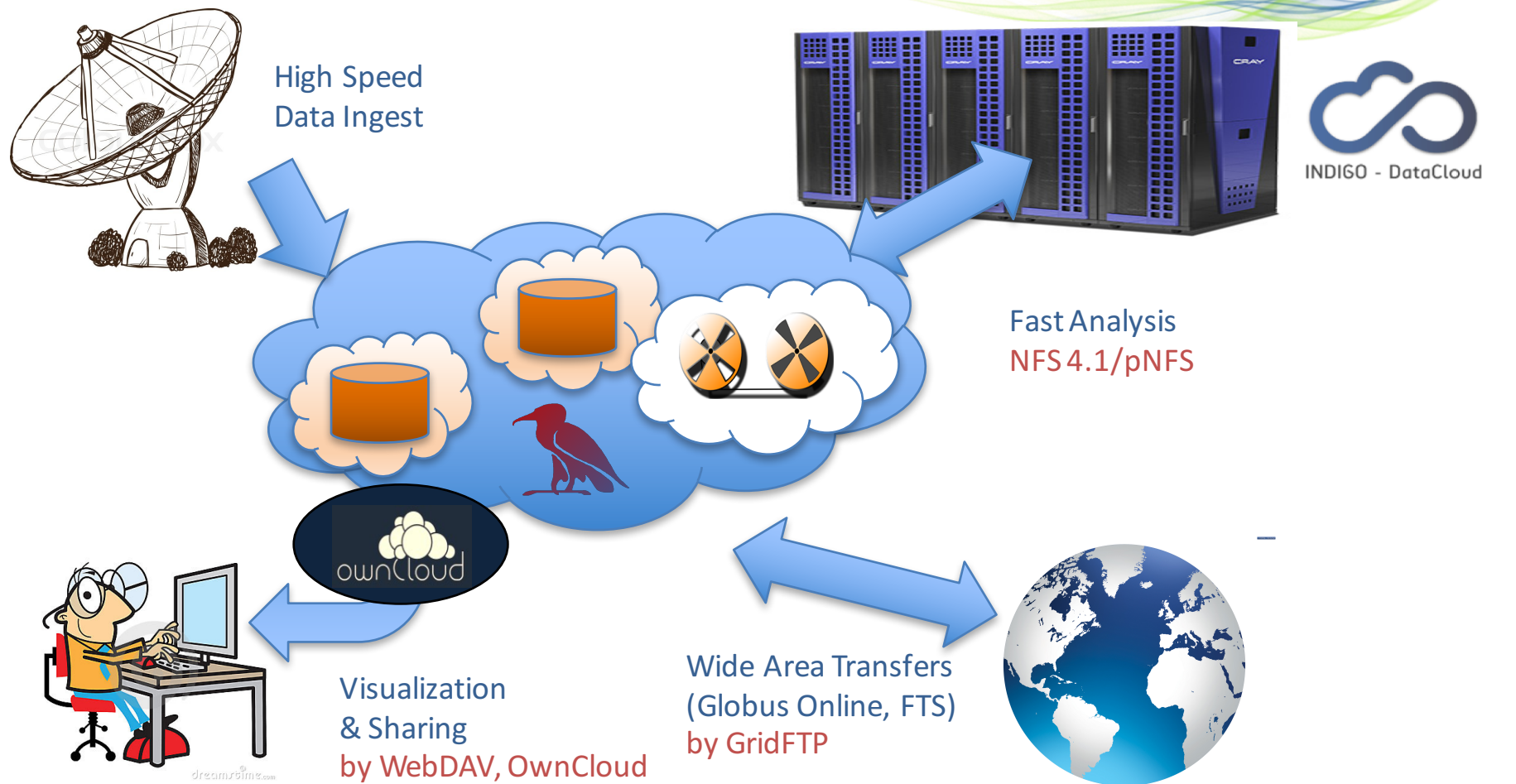
The Hybrid System



Putting it all together

The scientific Data Cloud

dCache.org 



Conclusion

Well on the way for new strategic direction on storage:

- Bringing costs down for operations
 - Using ‘low maintenance’ storage backends (e.g. CEPH)
 - Introducing High Availability mechanisms to support ‘dark data centers’
- Allowing Quality of Service in storage to be defined by sysadmins and users.
 - With GUI
 - Remotely : SRM, REST, CDMI
- Introducing Cloud mechanisms (e.g. sharing) not only for home directory type data but as well for experiment data.
- Implementing the “Scientific Storage Cloud”
- Stick with industry standards
 - Allow sharing resources at sites
 - Benefit from development and fast turnaround in security components and mechanisms.





The END

further reading
www.dCache.org

Integrated Media Technology Support



Technology	Driver
Spinning Disk	POSIX Driver
SSD	POSIX Driver
Shingled (SMR)	POSIX (disable remove)
	TAPE Driver
Removable devices	TAPE Driver
CEPH	New CEPH Block Driver