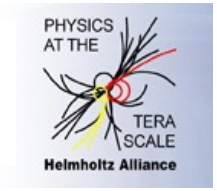
dCache.ORG

dCache.ORG

dCache, ready for the LHC production and analysis ?

Patrick Fuhrmann et al.

additional funding, support or contributions by



d-grid
DGI II

The people

dCache.ORG

dCache.ORG

Head of dCache.ORG

Patrick Fuhrmann

Core Team (Desy, Fermi, NDGF)

Andrew Baranovski

Gerd Behrmann

Bjoern Boettcher

Ted Hesselroth

Alex Kulyavtsev

Iryna Koslova

Tanya Levshina

Dmitri Litvintsev

David Melkumyan

Paul Millar

Owen Synge

Neha Sharma

Vladimir Podstavkov

Tatjana Baranova

N.N. : Hiring .....

Head of Development FNAL :

Timur Perelmutov

Head of Development DESY :

Tigran Mkrtchyan

Head of Development NDGF:

Gerd Behrmann

External

Development

Abhishek Singh Rana, SDSC

Jonathan Schaeffer, IN2P3

Support and Help

German HGF Support Team

Flavia Donno, CERN

Akos Frohner

Content

dCache.ORG

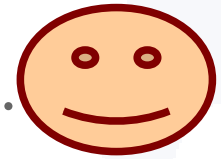
Quick reminder on what dCache.org does.

Quick reminder on dCache highlights (in a nutshell).

Ongoing work to improve dCache within the overall WLCG data management community. Learning by doing.

Ongoing work in terms of standardisation.

Some remarks on analysis

Some fun ... 

A summary

dCache.ORG

How is dCache.org structured ?

dCache.org : sustained and independent funding

dCache.ORG

dCache.org is independent of the LCG, OSG organisational structures and funding.

This made us some kind of 'aliens' in the past but it will be appreciated by dCache users in the future when EGEE X will phase out and the EGI and UMD are still not properly set up.

All three major partners in the dCache collaboration, FERMIlab, NDGF and DESY, of which two are WLCG Tier I's, highly depend on the product themselves and have been building a whole infrastructure around it.

So, whatever happens with other middle-ware components, as long as the dCache technology makes sense, there will be powerful and sustained support.

dCache.org : sustained and independent funding

dCache.ORG

Beside :

8 out of 11 Tier I centers are using dCache as well as some 40 larger Tier II's. The dCache.org funding bodies understand the responsibility. (Follow up MoU in preparation)

The dCache Tier I workshop in January has been very promising in terms that other Tier I's would be willing to contribute if a framework is found in which this can be done.

In Germany a 'Storage Support Group' has been set up, funded by the German Government, which builds up dCache knowledge. They support the German site but contribute to documentation and training. in general, e.g. Chimera and dCache ACL workshop in April.

dCache.ORG : product from one source

dCache.ORG

Although different dCache components are provided by different LAB's there is a central 'virtual' place which makes it a single product.

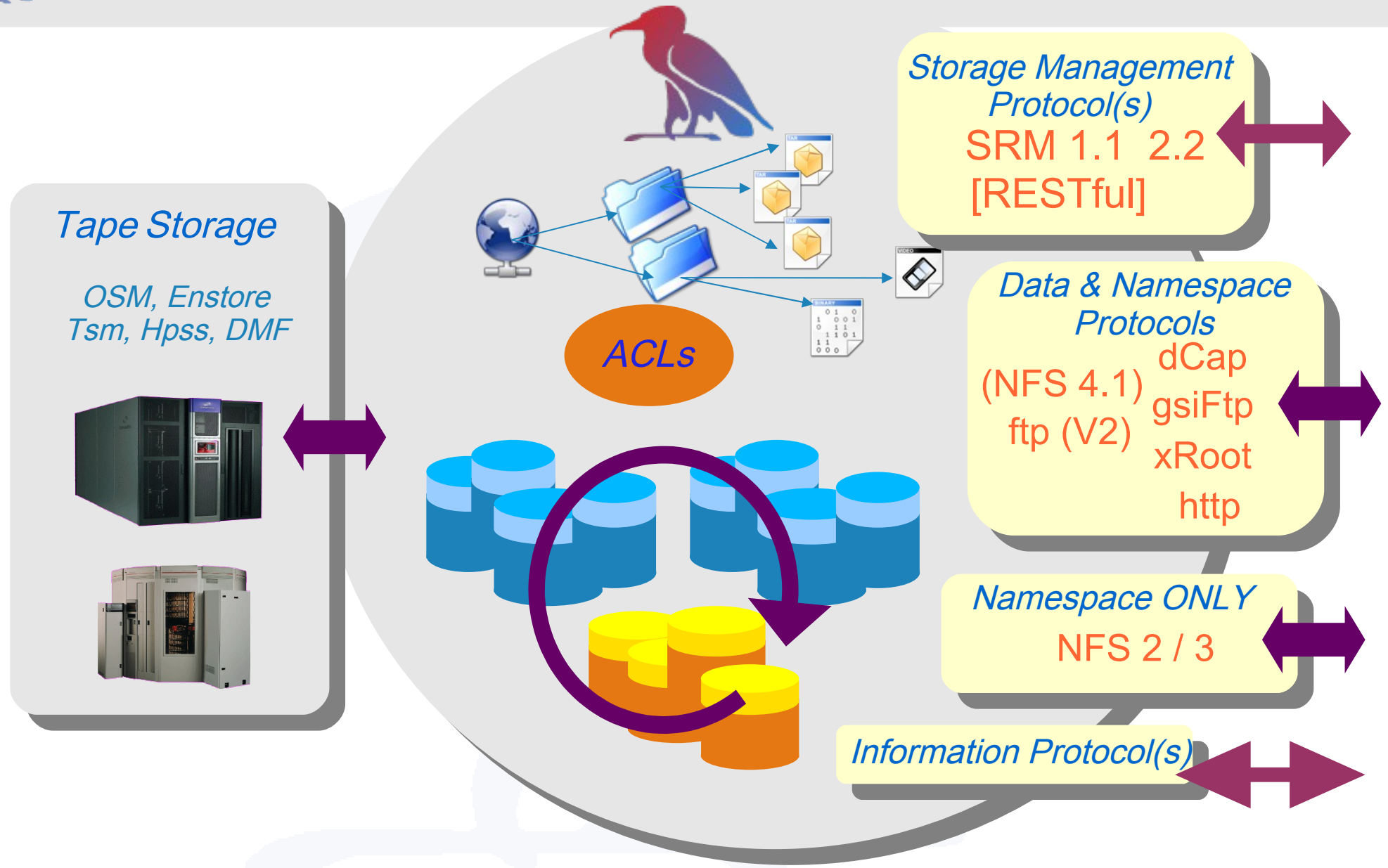
- The consequences are :
  - All components fit together naturally, *no patchwork, one design.*
  - At least once per week phone conferences on *compatibility,* testing, release strategy and internal certification.
  - Although each lab mainly supports its own components we have a *centralised support system*, which find the appropriate person for you.
  - *Single place for download* and documentation.
  - *Consistent release strategy* and release notes on all components at a single place.
  - dCache.org interacts with *gLite and VDT* and the national support teams. (certification and testing)

dCache in a nutshell.

With a focus on 'managed storage'.

dCache.ORG

dCache characteristics, the Overview

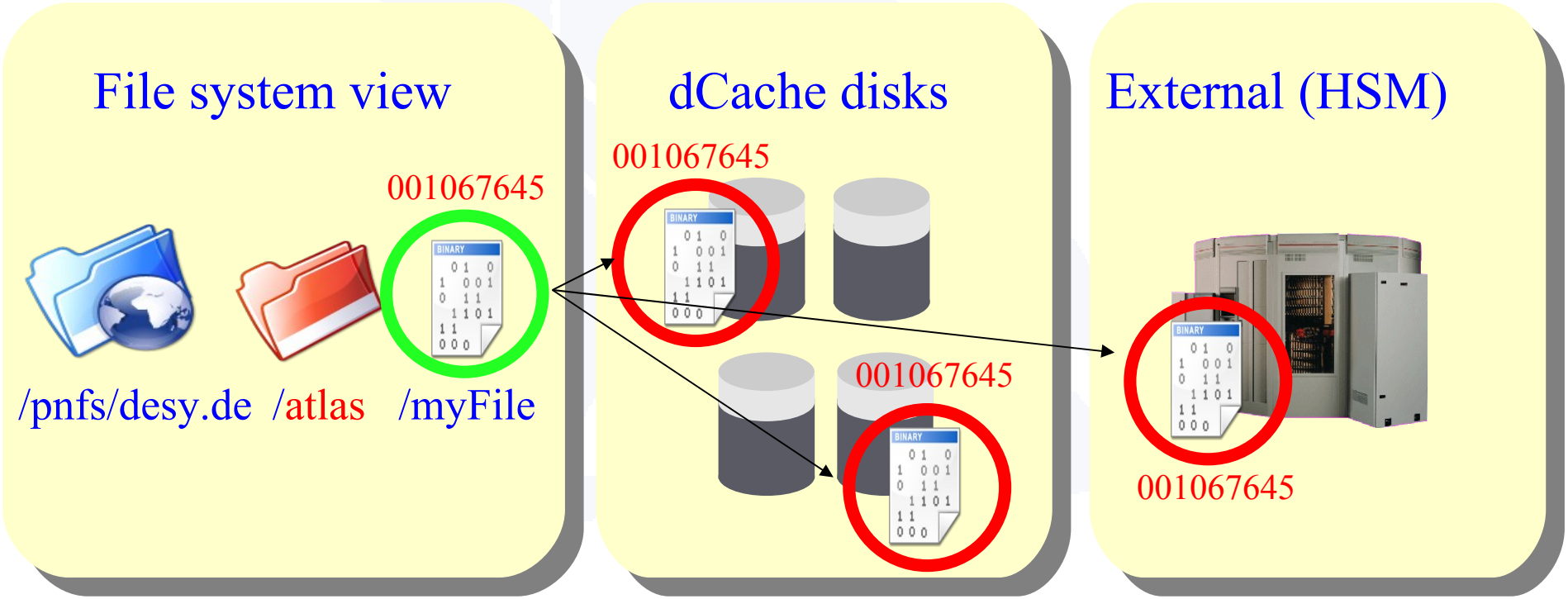


dCache.ORG

dCache.ORG

dCache characteristics : names and files

- 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 ftp, SRM)



dCache.ORG

File hopping

Use a intermediate buffer to optimise tape access.

Reliable (expensive) write pools.

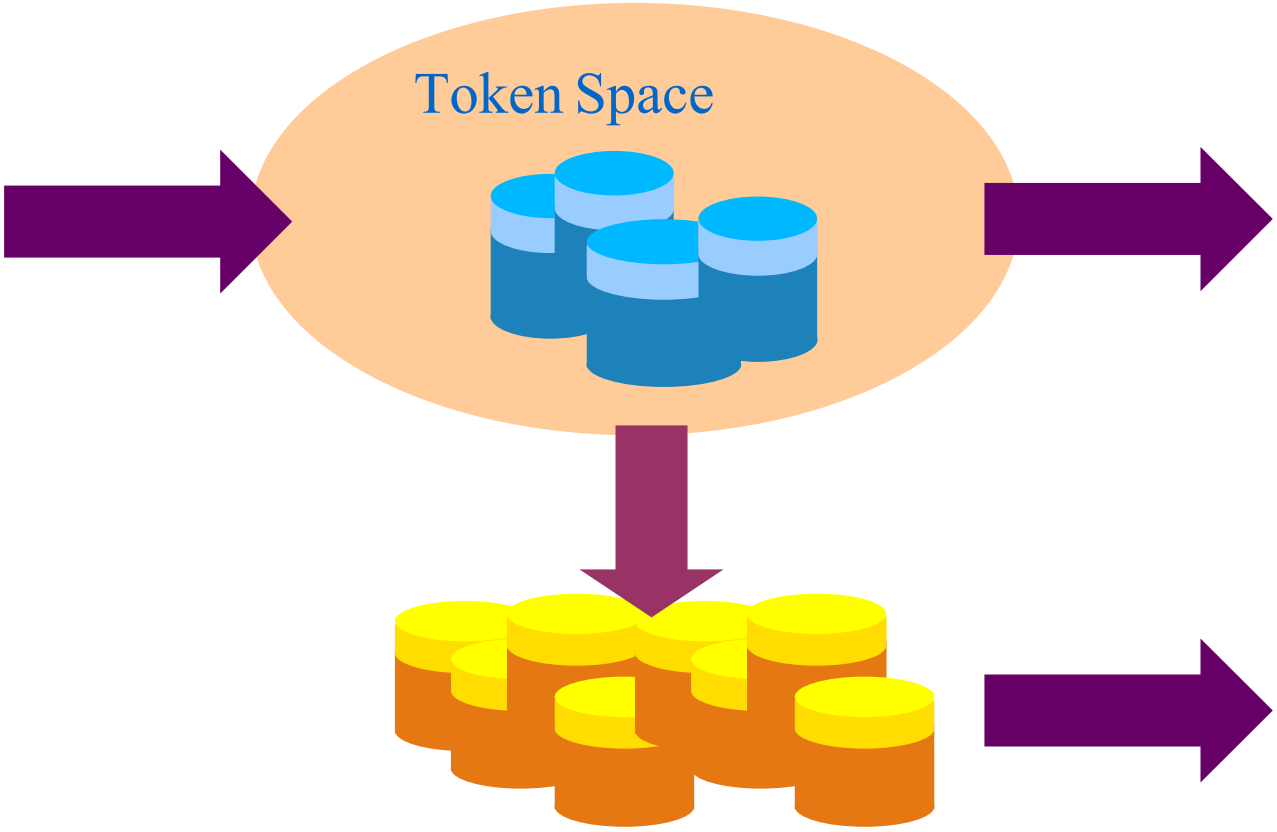
Read pools

Copy selected file-sets to read pools (on arrival)

Any protocol

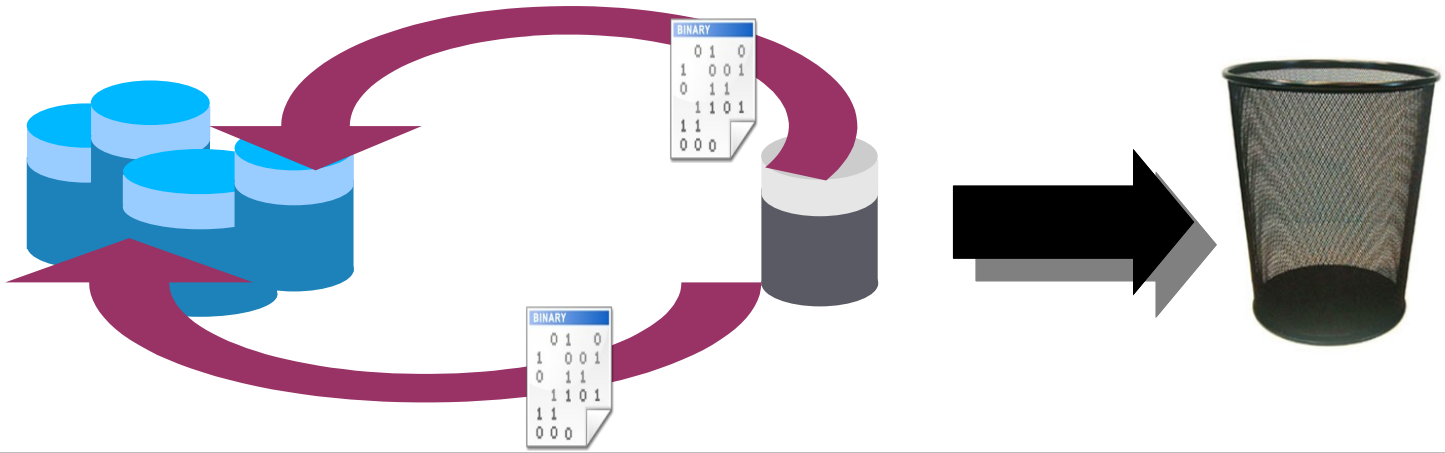
dCache.ORG

Automatic file replication on hot spot detection.

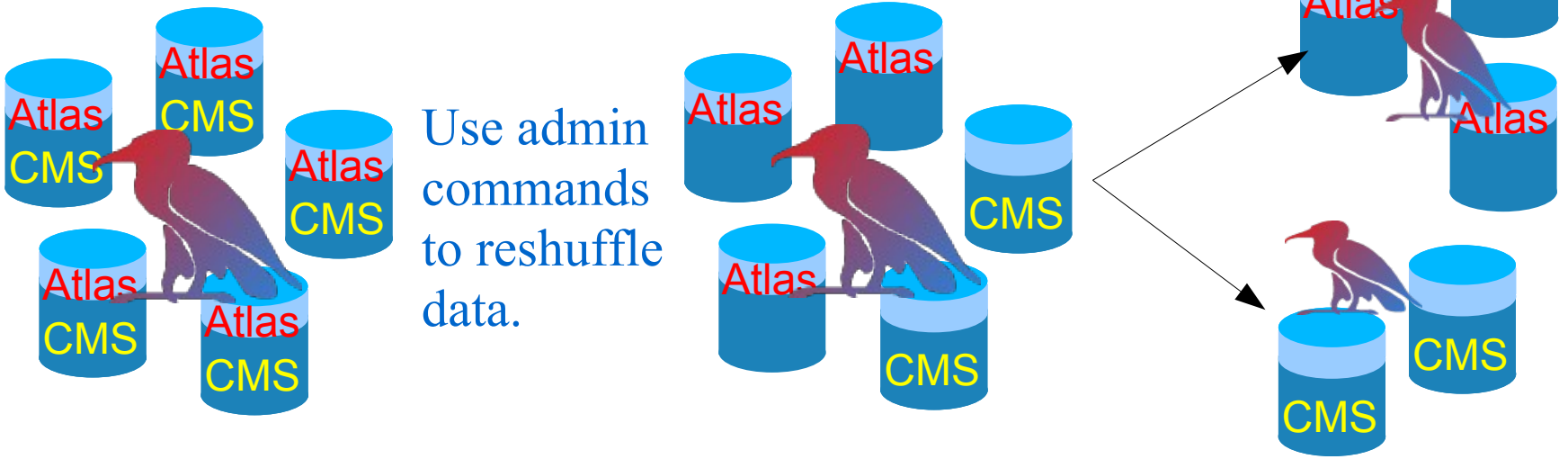


Decommission old hardware

Use admin commands to drain old storage devices.



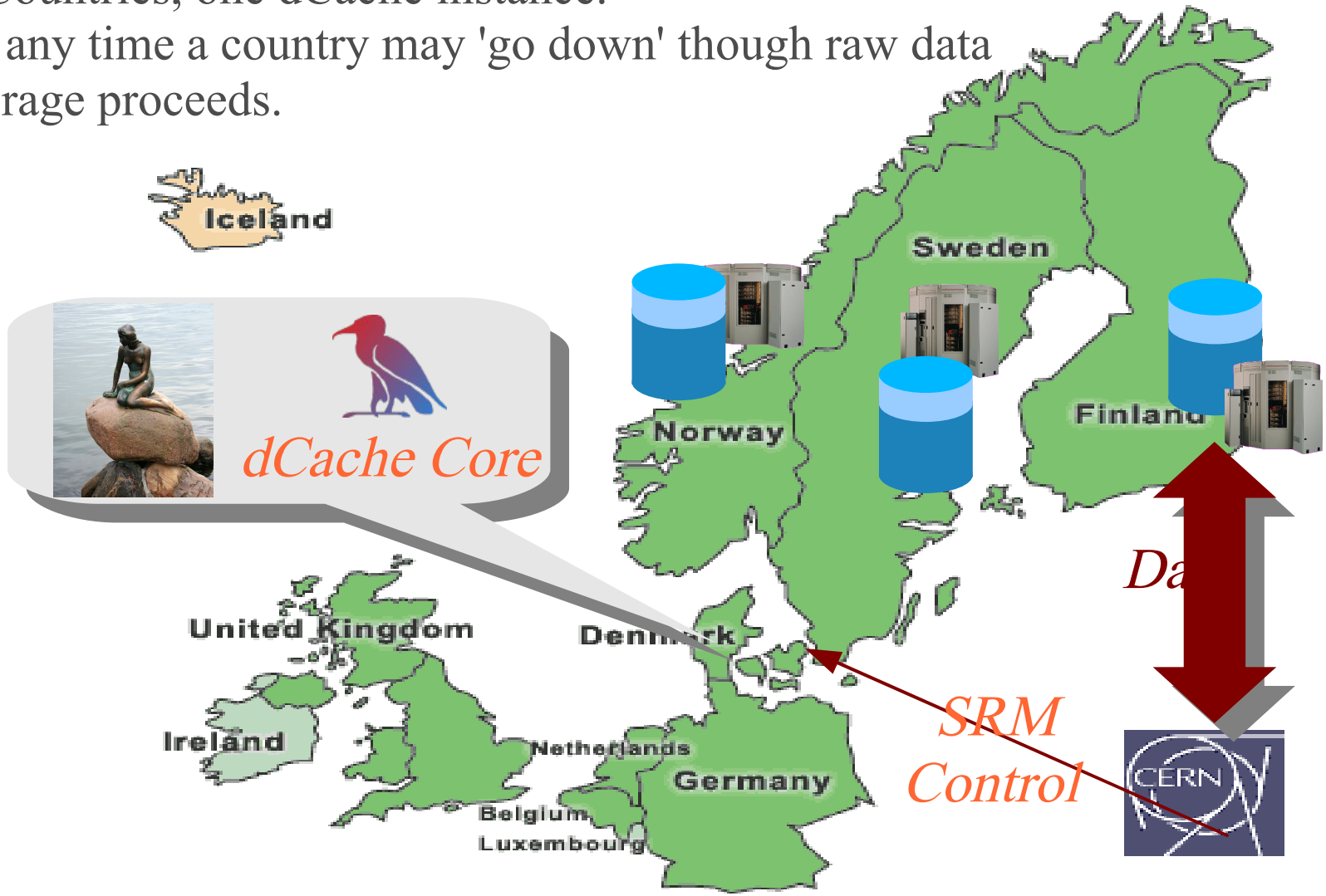
Split off parts of an instance



dCache.ORG

Building a highly distributed system (NDGF)

- 4 Countries, one dCache instance.
- At any time a country may 'go down' though raw data storage proceeds.



Ongoing improvements

with the rest of the WLCG storage crowd.

Thanks to Flavia, Ákos, Andrea, Tanja, Jean-Philippe and many many more.

dCache.ORG

Learning by doing ...

dCache.ORG

This is the first time we built a data grid of that size, composed of a variety of different storage systems.

We had to understand how those components (SE's) interact globally.

The SRM, supposed to solve this problem, has only be a partial success but has been a lesson. (A rather expensive one though).

Short term (pre D-day) improvements

dCache.ORG

- SRM has at least two duties :
  - Serve user requests as fast as possible.
  - Protect back-end storage system from overload.
- And two problems :
  - It doesn't do either.
    - Implementation problem
    - Protocol interaction problem
  - To much of an abstraction (Graeme S.).

dCache.ORG

How do we improve ?

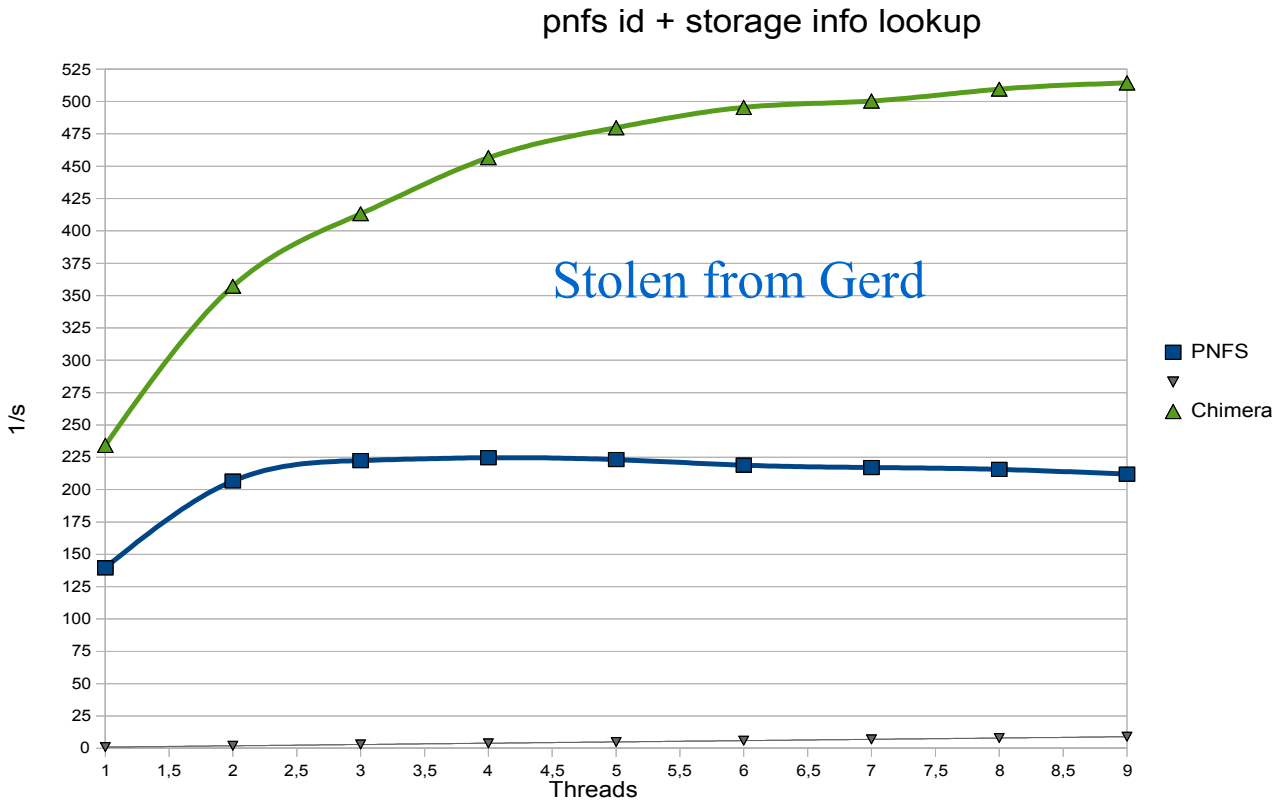
Improvements NOW : Chimera

dCache.ORG

Obvious improvements : Make the back-end faster.

Faster name space : **Chimera** instead of pnfs.

NDGF converted successfully last week.



dCache system can now be modified to support bulk operations on the name space level, which would make better use of SRM bulk requests.

Implementation **independent** improvements.

dCache.ORG

(This is a collaborative effort)

SRM_INTERNAL_ERROR

Inform the client that we are currently really busy and that we would appreciate if it would back off for a moment.

Request Lifetime

If client and server would agree (in advance) on the maximum time before both time-out a request, unnecessary requests wouldn't have to be processed.

Asynchronous SRM ls

The server may queue the request and proceed with light weight requests (e.g. get status)

Short term (pre D-day) improvements

dCache.ORG

Implementation **dependent** modifications :

Faster name space (pnfs to Chimera)

Stolen from Timur

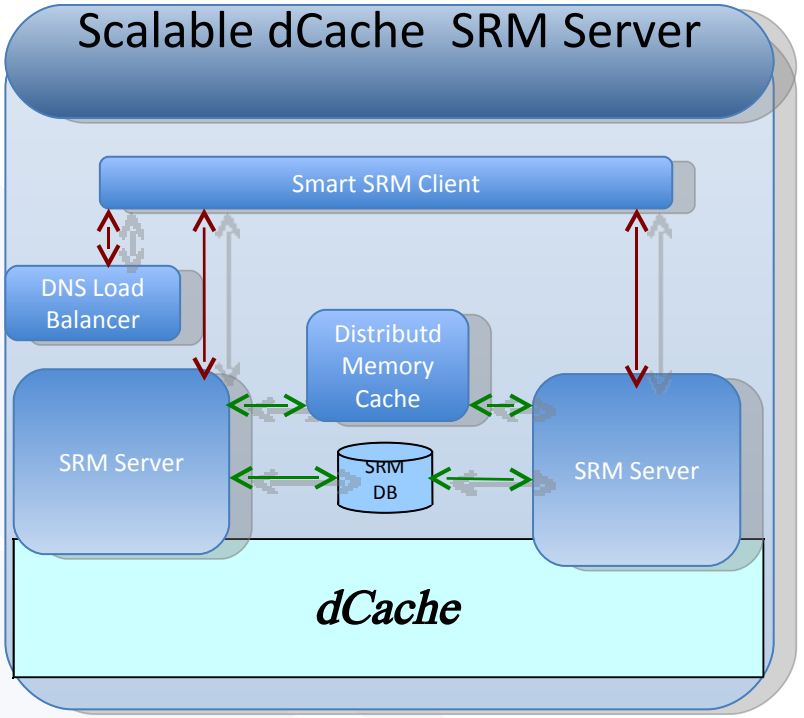
High CPU load due to GSI Authentication and Credential Delegation.

- Cache public and private key pairs used in GSI authentication and handshake.
- Work with Globus on improvements.
- Consider https as a long term solution.

Stolen from Timur

Scalability

- SRM is a single point of entry into a storage
  - Natural bottleneck
  - Single point of failure
- Distributed SRM
  - Scalable
  - More reliable



There is a poster on his topics.

dCache.ORG

File system ACL's will be in 1.9.3

We are working on documentation.

There will be a work shop on ACL's (and chimera) beginning of April in Aachen.

dCache.ORG

Standardisation efforts

Standardisation efforts

dCache.ORG

Collaborative effort within WLCG

SynCat

Standardised SE name space dump for synchronising file catalogues.

dCache.org took a leading role in this effort.

There is a poster on his or just find Paul.

GLUE 1.3 and 2.0

dCache.org is actively participating.

Standardisation efforts : NFS 4.1 (pNFS)

dCache.ORG

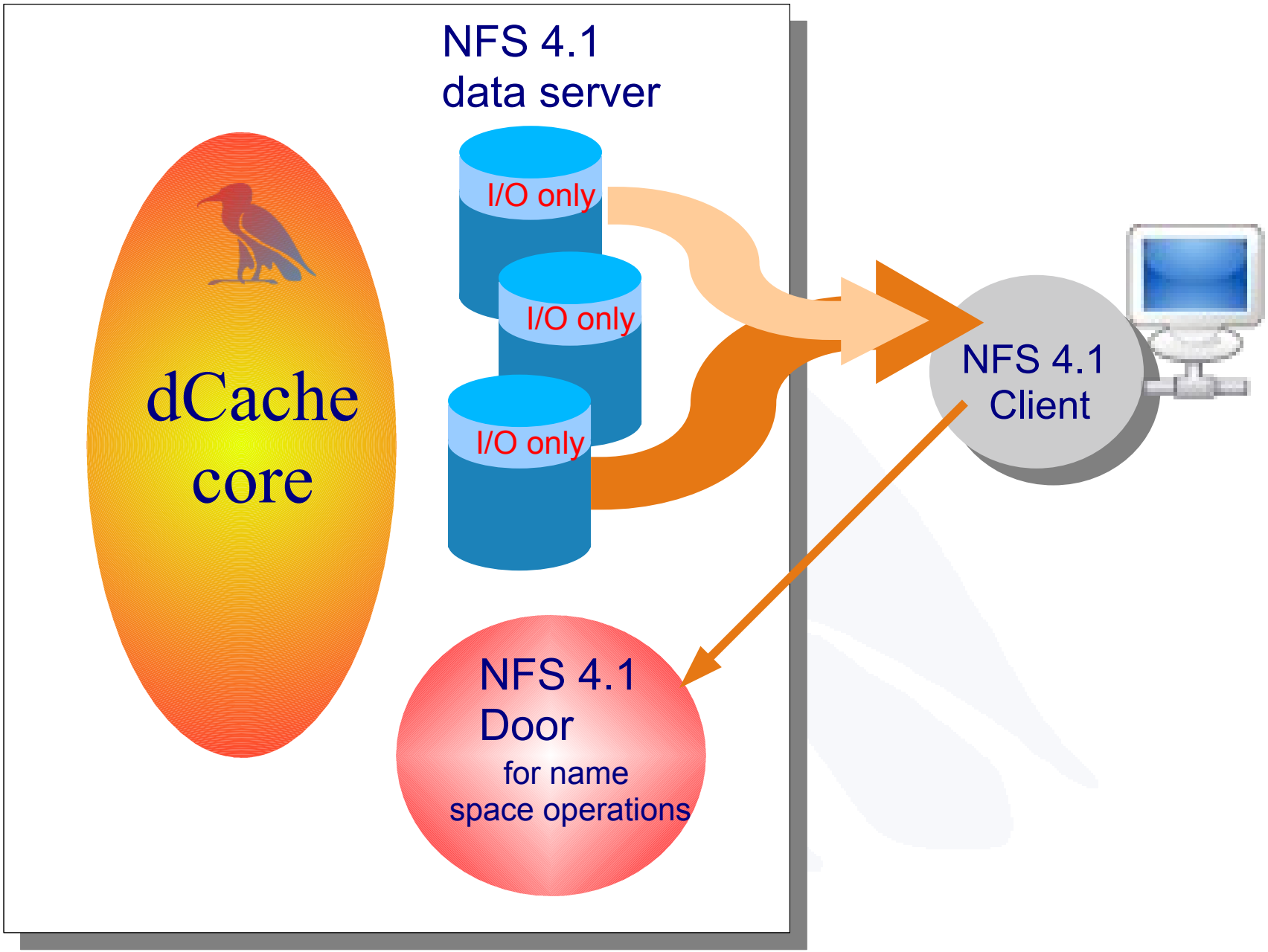
- NFS 4.1(pNFS) is aware of distributed data.
- NFS 4.1 (pNFS) is an IETF standard.
- POSIX Clients are coming for free. No preload, no relinking. (provided by all major OS vendors).
- Widely adopted by major storage hardware vendors.
- Will make dCache attractive to other (non-LHC) applications and communities.
- LCG could consider to drop LHC specific protocols, to avoid manoeuvring ourselves into a technological corner.

NFS 4.1 : technical perspective

dCache.ORG

- NFS 4.1 is aware of distributed data
- Faster (optimized) e.g.:
  - Compound RPC calls
  - e.g. : 'Stat' produces 3 RPC calls in v3 but only one in v4
- GSS authentication
  - Built-in mandatory security on file system level
- ACL's
- dCache can keep track on client operations
  - OPEN / CLOSE semantic (so system can keep track on open files)
  - 'DEAD' client discovery (by client to server pings)
- smart client caching.

dCache.ORG



dCache.ORG

About analysis ...

dCache.ORG

Are we ready for analysis ?

I don't know.

We couldn't find reliable requirements yet.

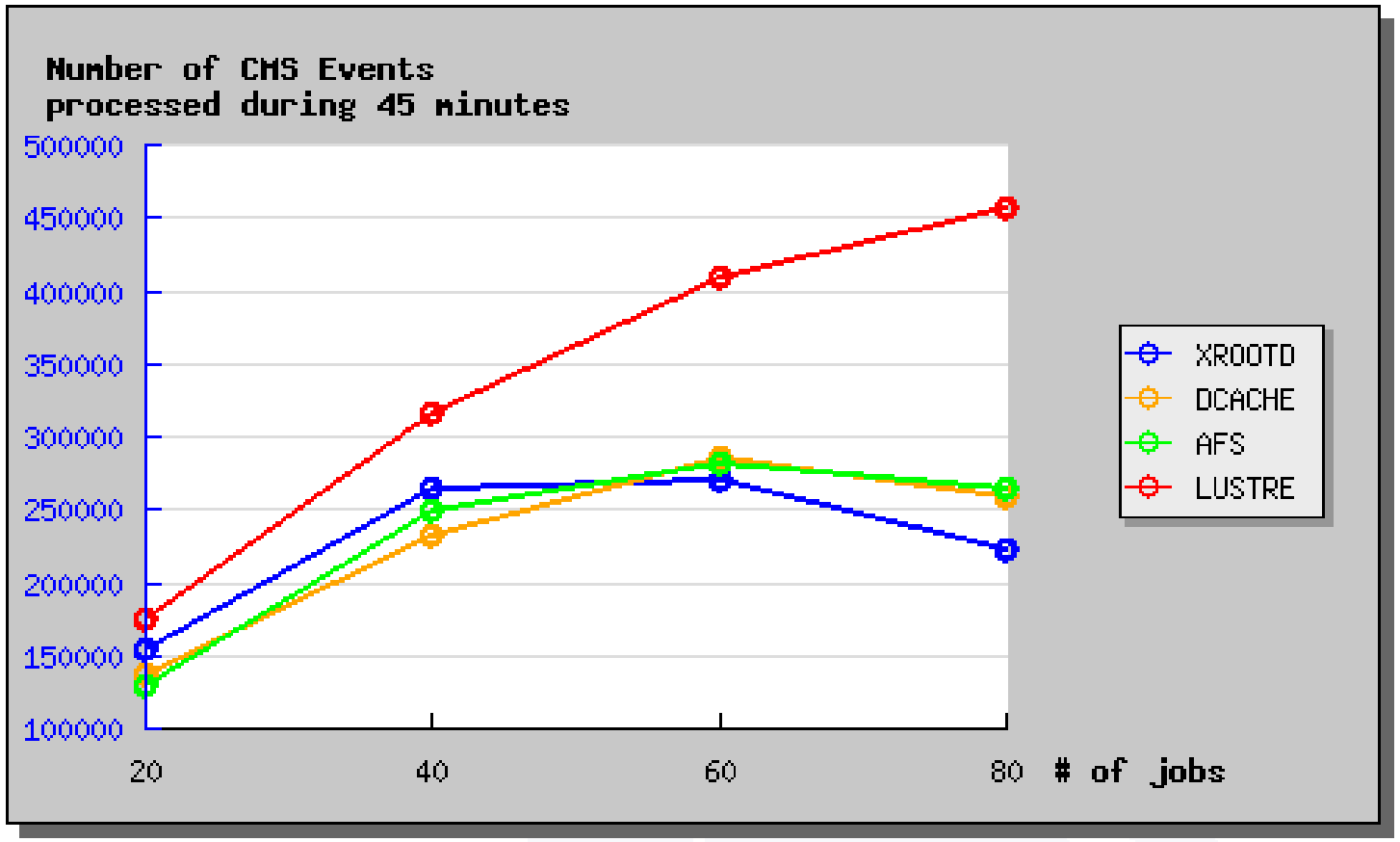
We need sample analysis jobs for creating a matrix.

It doesn't look bad as far as we can say now.

We'll concentrate on this, but we need feedback from sites.

dCache.ORG

Stolen from Andrei Maslennikov's presentation at the Fall HEPIX 2008



Quote Andrei "*In this summary graph, Lustre seems to be almost twice as efficient compared to the other methods. As AFS, dCache and xroot seem to be very close to each other, they may have had a common blocking factor such as the local file system (and NOT the data access protocol)*"

Atlas Hammer Cloud results

dCache.ORG

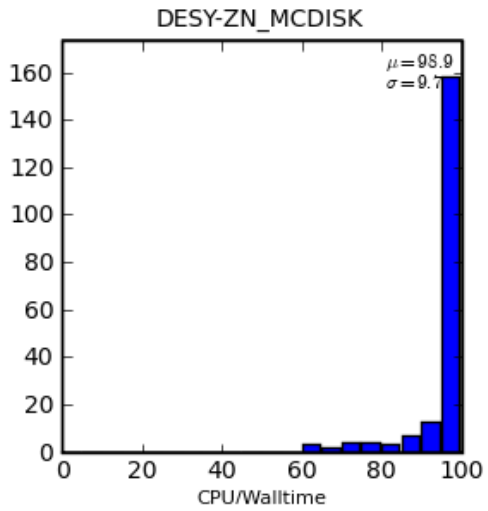
Running dCache/dccp/dCap at DESY Zeuthen

dccp

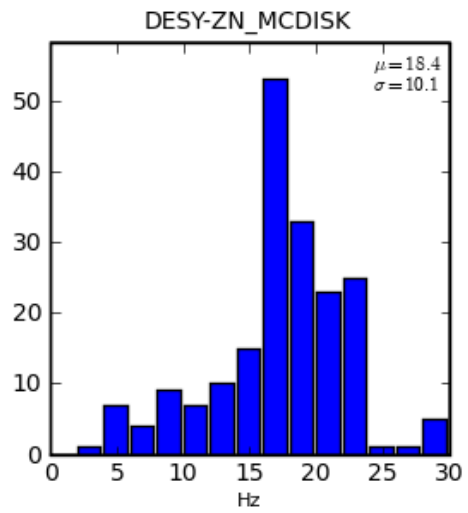
dCap

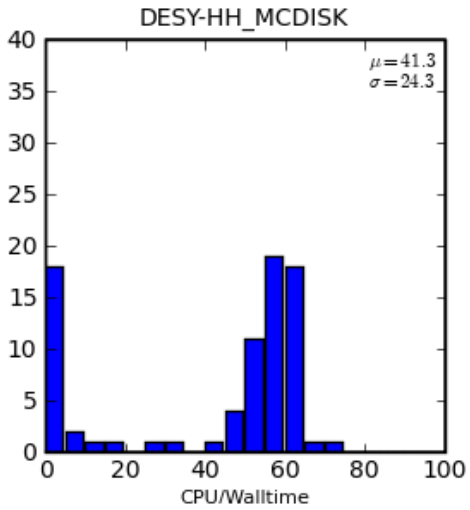
(Room for improvement :-))

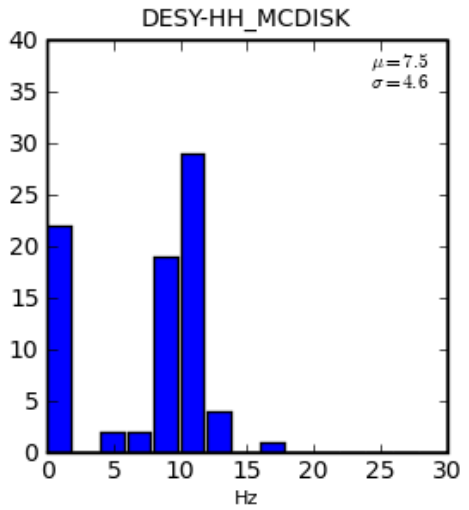
Site CPU/Walltime



Site Events/Second







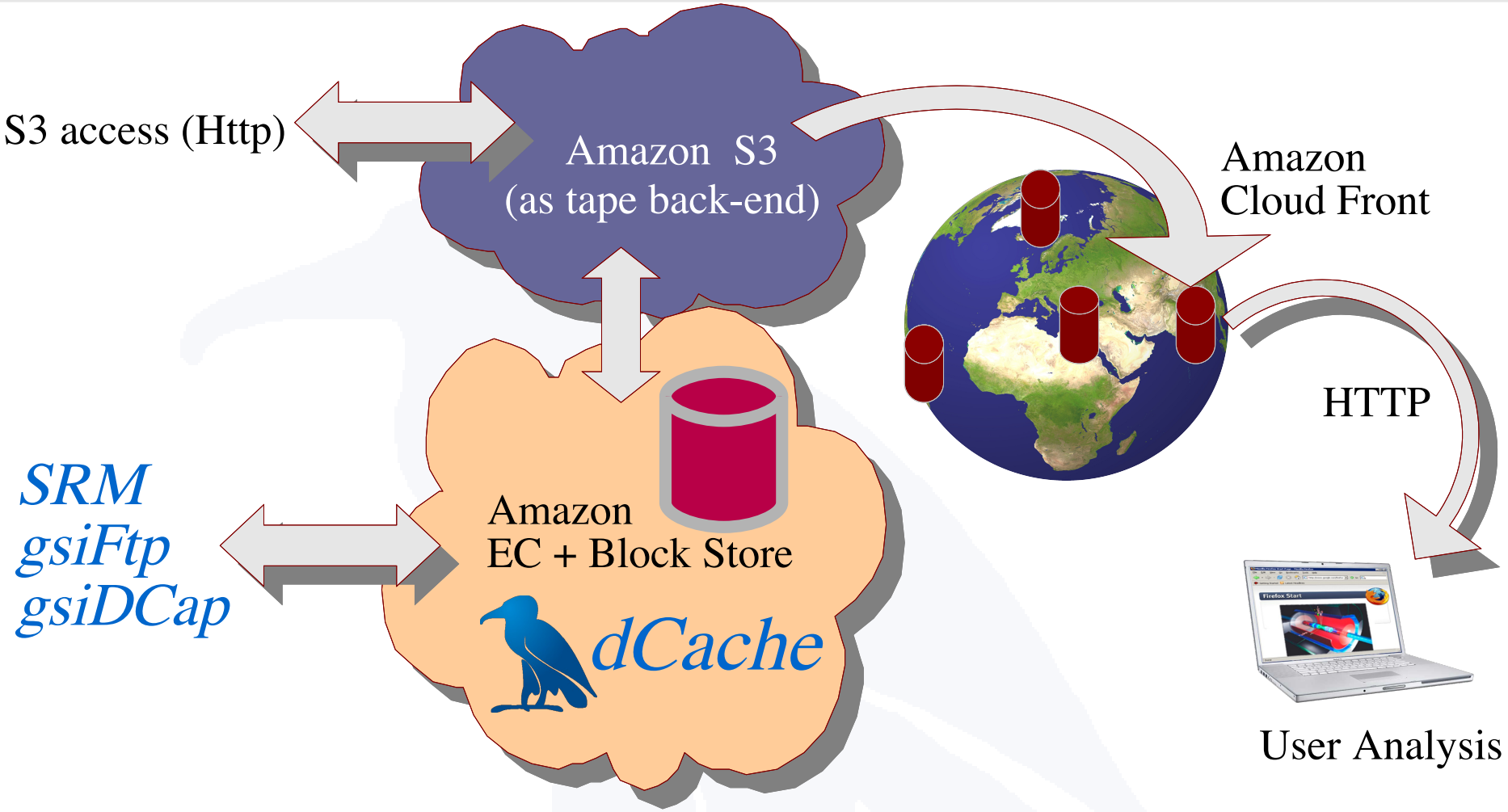
dCache.ORG

dCache.ORG

dCache @ Amazon

Fun : Tier I/II/III @ amazon ?

dCache.ORG



srm://se01.dcache.org/data/myfile → *http://store.dcache.org/data/myfile*

1- 2 Million Dollars for the DESY Altas Tier II (SE) per year.

Summary

- dCache.org organisational structure is well suited for long term usage.
- dCache is an integrated solution with a broad spectrum of activities.
- dCache is managed storage on the large scale.
- In various data challenges we have shown that we can sustain the requested data rates.
- We are very active in standardisation efforts. We believe this is the only way, LHC doesn't end up in a technological corner.
- Analysis hasn't been our focus up to know. It doesn't look bad but there is certainly room to improve.

dCache.ORG

dCache.ORG

Further reading

www.dCache.ORG