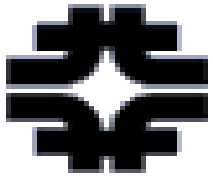
dCache, managed storage

Patrick Fuhrmann

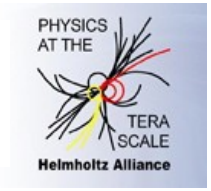
dCache.org

Partners



Picture stolen from Flavia Donno

additional funding, support or contributions by



d-grid
DGI II

Content

dCache.org

The LHC computing grid.

Quick introduction

dCache, managed storage

In a nutshell

Deployment

Commercial

Going for standardization

NFS 4.1

We are not discussion

dCache.org

How to get such a project funded.

How to produce professional software contributed by 3 independent partners with different objectives.

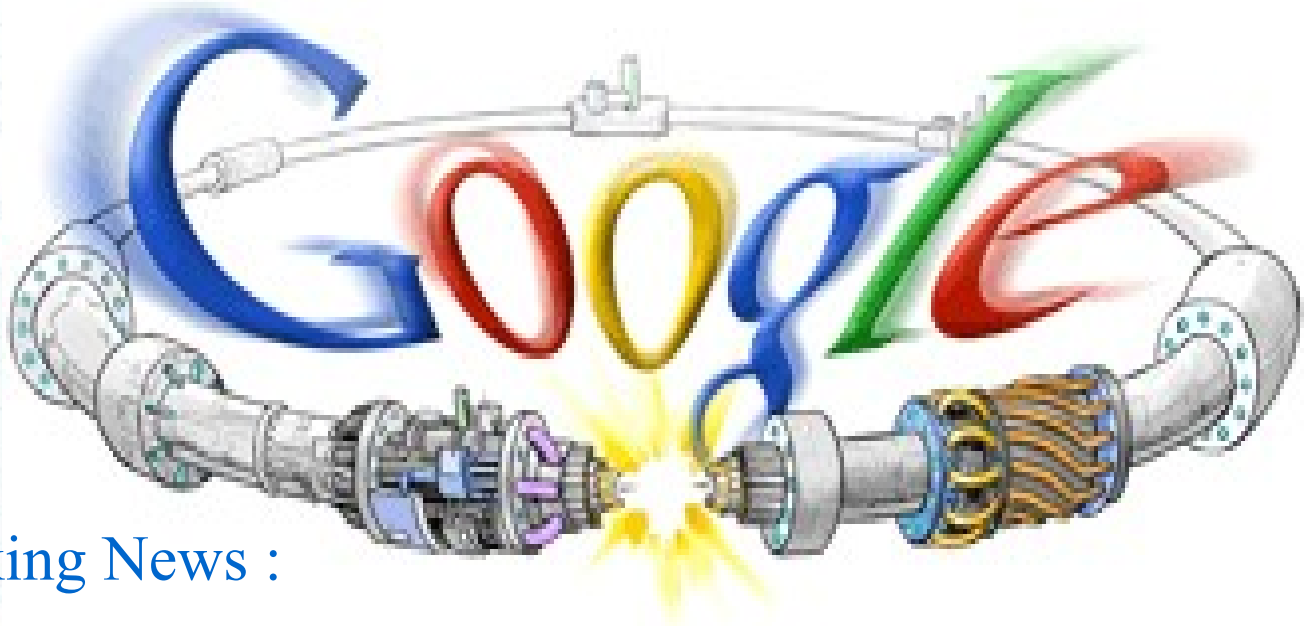
Serving a set of huge communities which believe data management has something to do with USB disks.

Doing business with CERN.

dCache.org

We are indeed lucky,
We only have technical problems.

Maybe you remember ...



Breaking News :

10 Sep 2008 : Physicists are launching worlds largest experiment

Which will be the last one in case LHC creates a black hole

20 Sep 2008 : LHC shutdown due to overheated magnet

8 Dec 2008 : LHC relaunches in July, 22 Million Euros ...

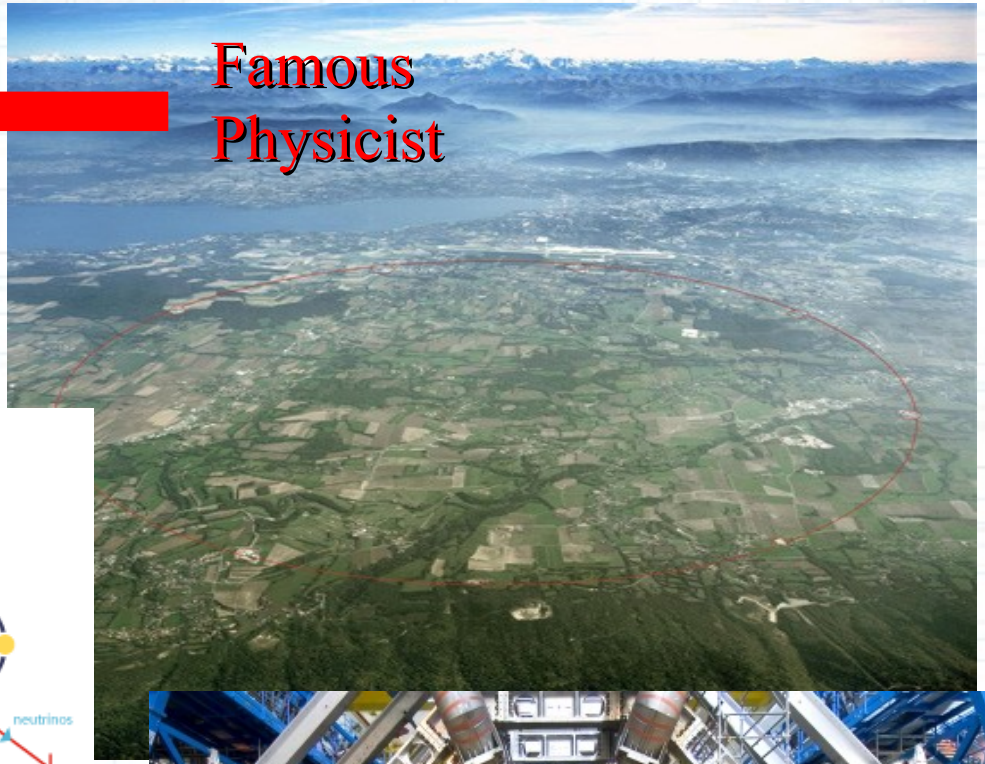
23 Oct 2009 : The Large Hadron Collider is cool! (1.9Kelvin)

To be continued ...

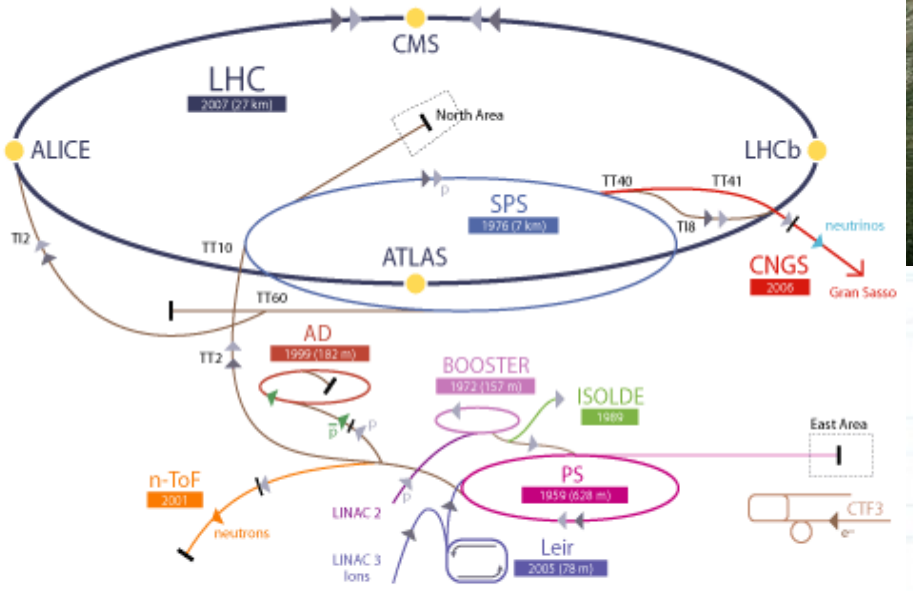
dCache.org

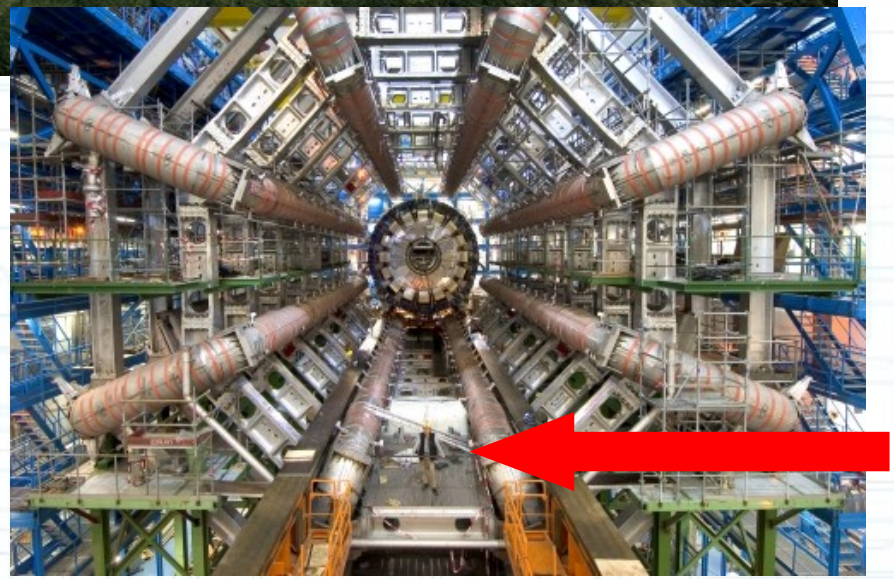
The LHC Tier model and the Grid.

Famous Physicist



CERN Accelerator Complex





Standard Physicist

dCache.org

In the meantime, without anybody noticing ...

The Worldwide LHC Computing GRID was built.

For the data management part this means :

- The LHC data grid has to handle a sustained stream of data in the order of 15 Petabytes per year produced at CERN and being distributed around the world, to
- 11 huge storage sites. (several xx Petabytes per site)
- And nearly 200 smaller sites (reaching into the x Petabyte area)
- 4 huge experiments with individual requirements.
- Thousands of active physicists around the word need to access the data in a timely fashion.

dCache.org

dCache.org

But why is LHC using the Grid for its ?

Hmmm

Doesn't matter. Grid is dead anyway.
Long Live Cloud Computing

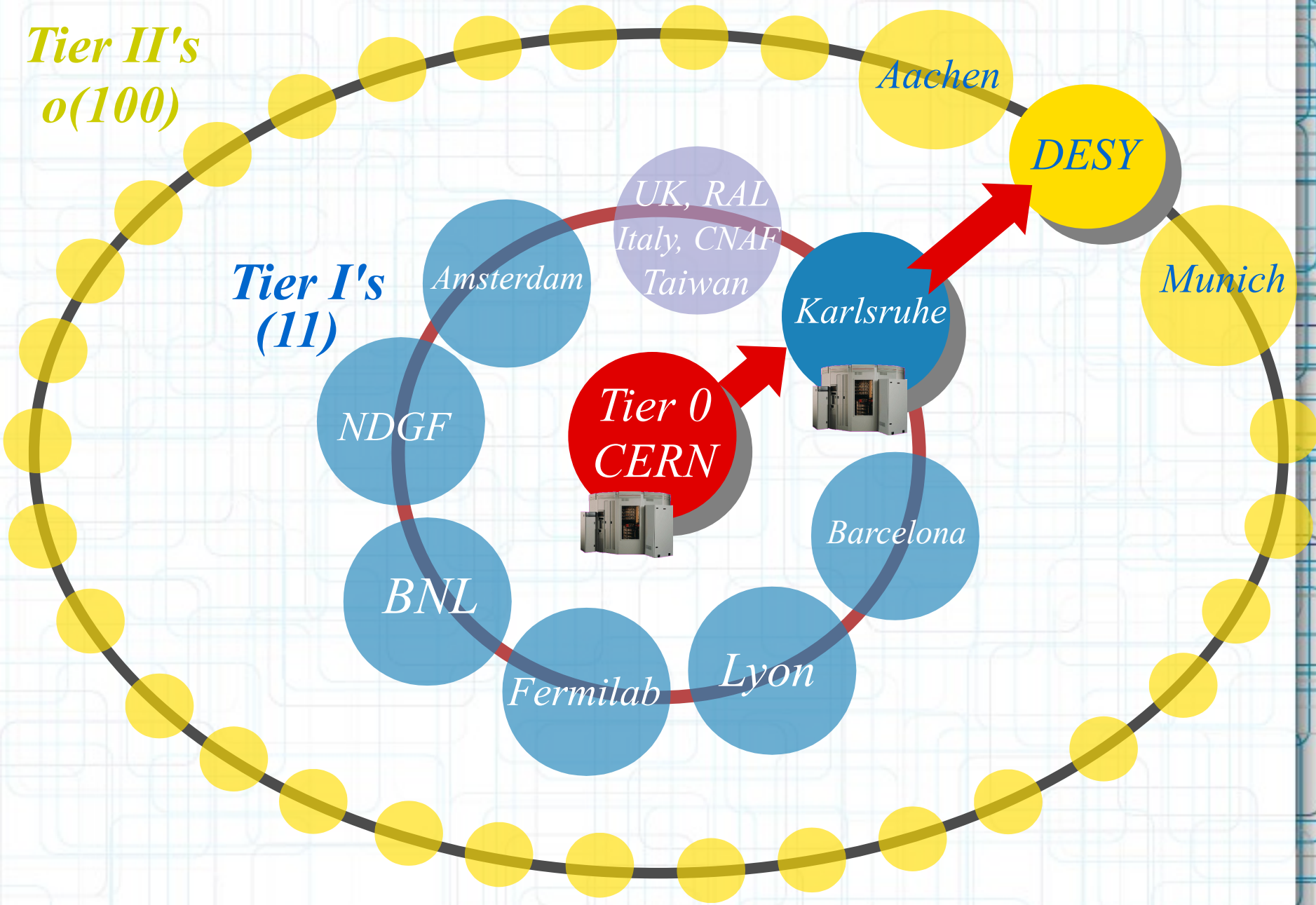
What is needed to run a data grid

Issues to be solved

- Virtual Organization Management (VO)
  - Individual users can't be managed any more.
- Information detection system
  - Using the phone to find space won't work.
- File transfer services
  - Ftp and scp won't do it
  - Transfers need to scheduled.
- Global file location catalogues
  - keep track of locations and replicas
- Remote data management
  - Local (site) sysadmins would be overloaded

dCache.org

The LHC Grid Tier model



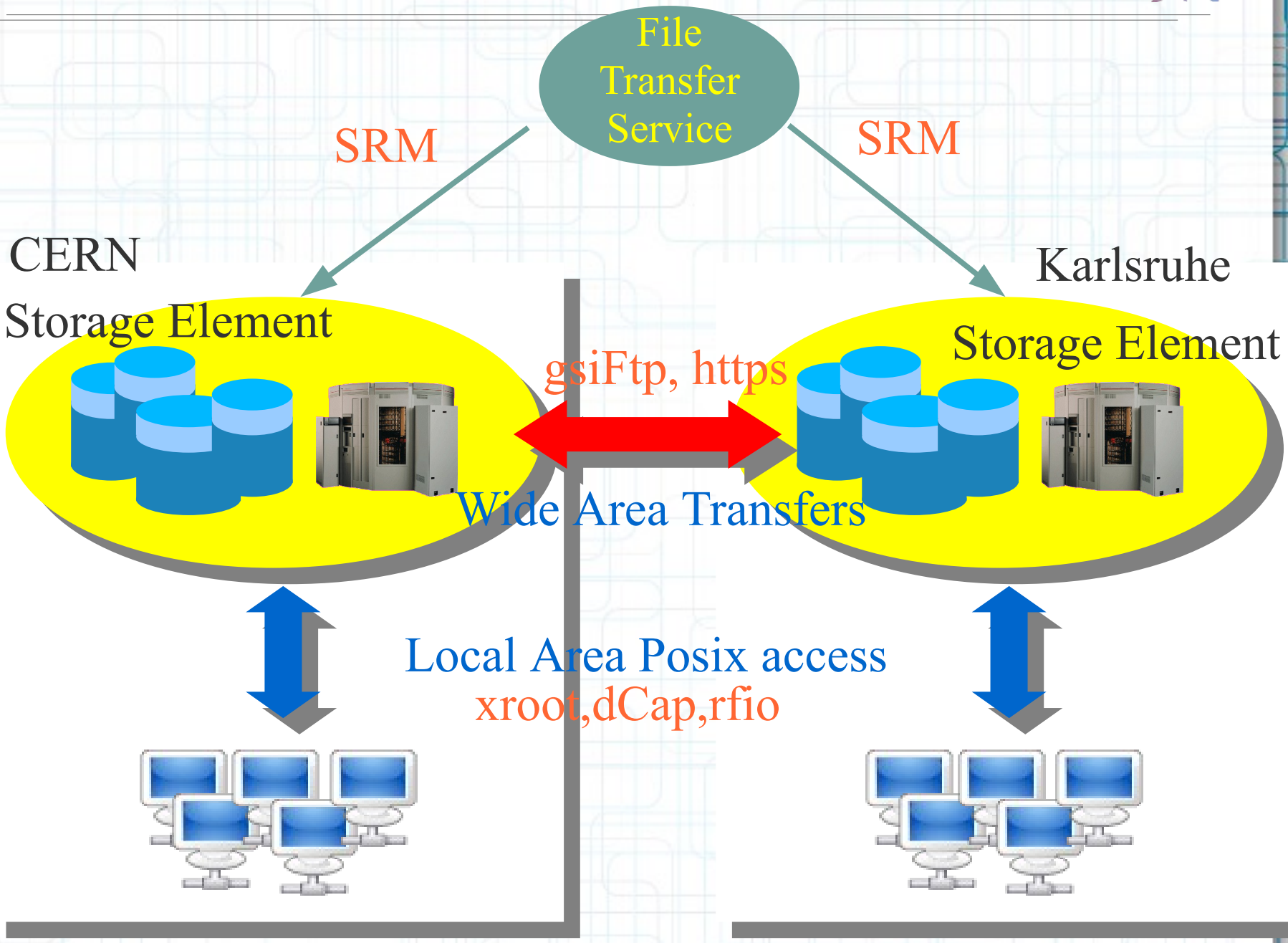
dCache.org

A Storage Element (SE) is a remotely controllable data endpoint.

- Allowing high bandwidth streaming secure data transfers.
- Allowing POSIX like local access e.g. from worker-nodes.
- Providing SE status information
  - Space : used, free space ....
  - Permissions, protocols ....
- Allowing to remotely manage storage. (Storage Resource MGR)
  - Manages spaces
  - Determines access latency (access time threshold)
  - Determines retention policy (probability of data loss, e.g. tape, disk)
  - Provides File name space operations
  - ProvidesTransport protocol negotiation

dCache.org

Storage Element interactions



dCache.org

Storage Element implementations

- **CASTOR** : disk/tape system at the Tier 0 and at Taipei/RAL. Rather complex. It needs special agreement with CERN if used outside CERN.
- **DPM** (gLite) : disk based system (no tape access). Up to medium size sites.
- **StorRM** : SRM 2.2 implementation on top of GPFS or Lustre. Can talk to a tape system through GPFS.
- **BeStMan** : Used in OSG (US) land. Disk only. Small sites.

dCache.org

Storage Element implementations

dCache.org

dCache

dCache in a nutshell

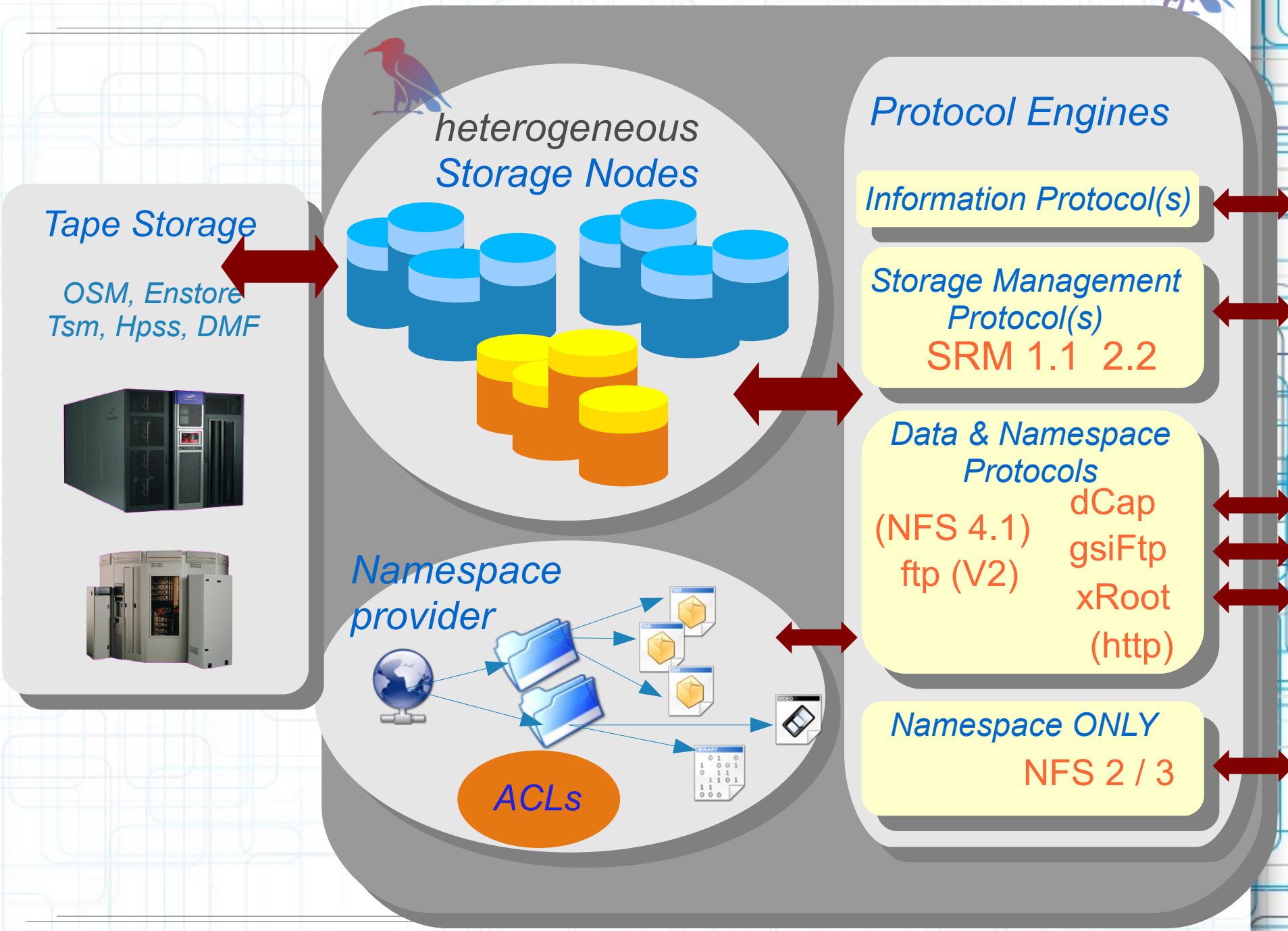
dCache.org

What is dCache; some basics?

- (Grid) Storage Software.
- Combines 1000's of independent heterogeneous *storage nodes* to a single system.
  - A *storage node* is a bunch of disks, some CPU, an OS and a network interface.
  - Only restriction : you need to be able to run Java on that box.
- Provides a single 'rooted' file system view. (/pnfs/mydomain/...)
- Name space is independent of the physical location(s) of the data.
- Support of physical data location outside of dCache. (Tape)
  - Currently used back-ends : Tsm, Hpss, DMF, Enstore, OSM
- Support of multiple internal and external copies of the same file system entry.
- Overall system is resistant against failures of single *Storage Nodes*.
- Support of all necessary storage control, data transport and information provider protocols for grid applications. (eg SRM, GLUE, gsiFtp....)
- dCache is an implementation of an *LCG Storage Element*

dCache.org

What is dCache, some basics?



In a Nutshell

dCache.org

- Strict name space and data storage separation, allowing
  - mv, rm, mkdir e.t.c without moving data
  - create, remove replicas or tape copies without changes in the name space.
  - convenient name space management by nfs (or http)

- File hopping (no user interaction required)
  - automated hot spot detection
  - configuration (read only, write only, stage only pools)
  - on arrival (configurable)
  - outside / inside firewalls

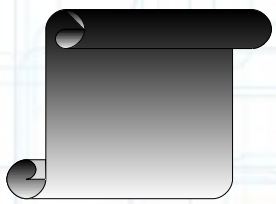
In a Nutshell

- Overload and meltdown protection
  - Request Scheduler.
  - Primary Storage pool selection by protocol, IP, directory, IO direction
  - Secondary selection by system load and available space considerations.
  - Separate I/O queues per protocol (load balancing)
- Supported protocols :
  - (gsi)ftp
  - (gsi)dCap
  - SRM 2.2
  - Nfs 3 (name space only)
  - NFS 4.1 with dCache 1.9.5 (Golden Release)
  - LHC Information Provider Protocol (GLUE 1.3)

dCache.org

dCache in a Nutshell

Scheduler and I/O queues and meltdown protection



I/O Request

Space Manager

Dispatcher by request Attributes

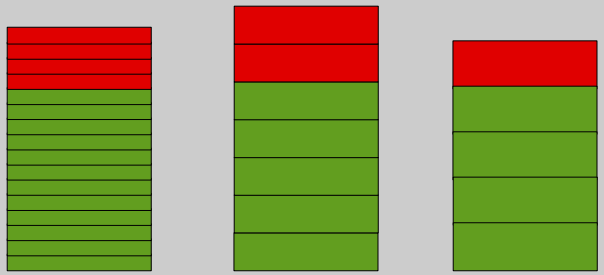
Pool Candidates selected by
Protocol
Client IP number/net
Data Flow Direction
Name Space Attributes (Directory)
SRM Spaces

List of candidates

Dispatcher by Pool Cost

Pool Protocol Queues

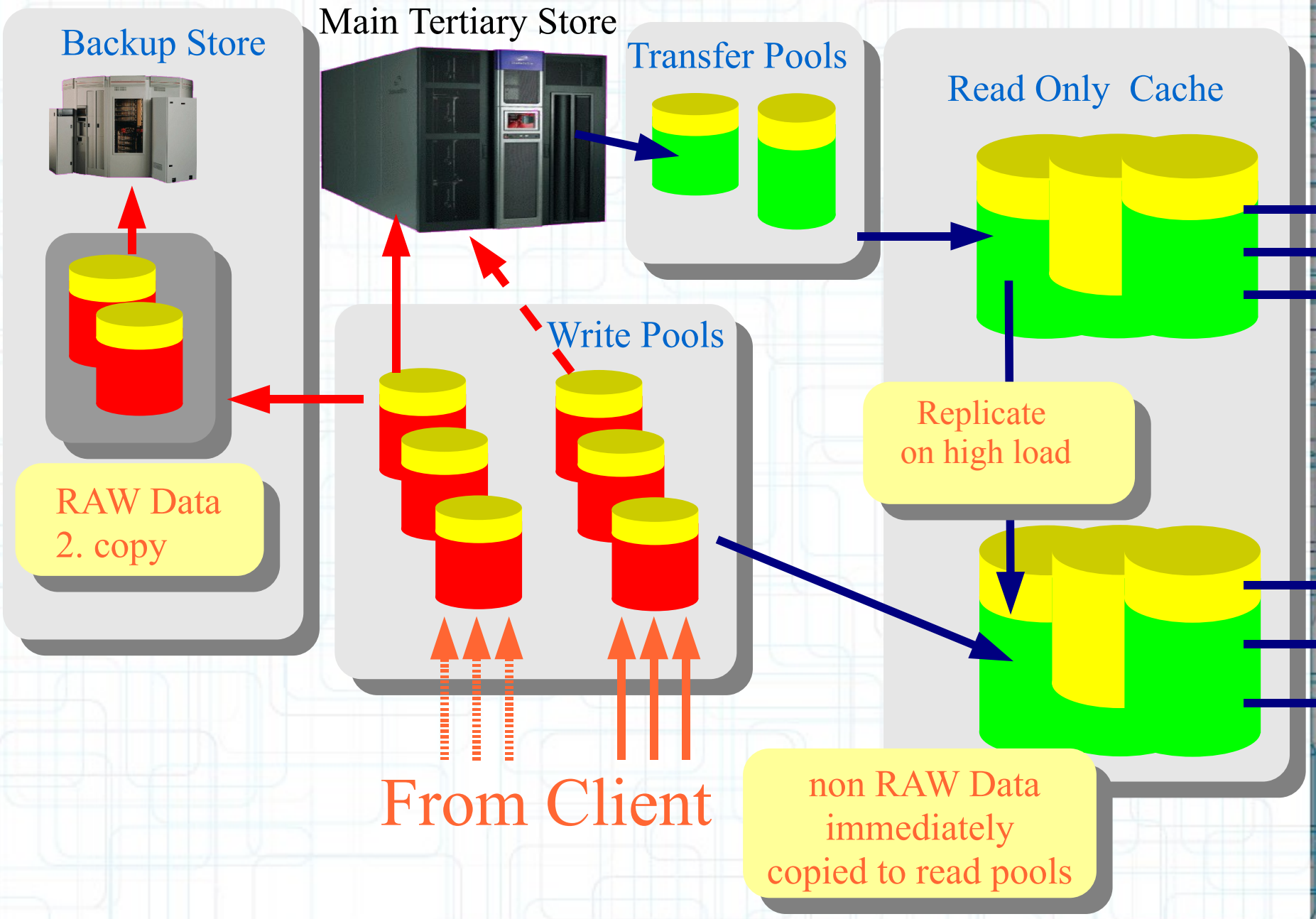
nfs4.1 *dCap* *gsiFtp*



dCache.org

In a Nutshell

File Hopping



dCache.org

dCache deployment

Get your dCache from

www.dCache.org

EMI
European Middleware Initiative

dCache.org

dCache deployment

8 out of 11 LHC Tier I centers

- Brookhaven National Lab (New York, US)
- Fermi National Lab (Chicago, US)
- gridKa (Karlsruhe, BRD)
- SARA (Amsterdam, NL)
- IN2P3 (Lyon, FR)
- PIC (Barcelona, SP)
- NDGF (Kopenhagen, Finland,Norway, Sweden,Denmark)
- Triumf (Vancouver, CA)

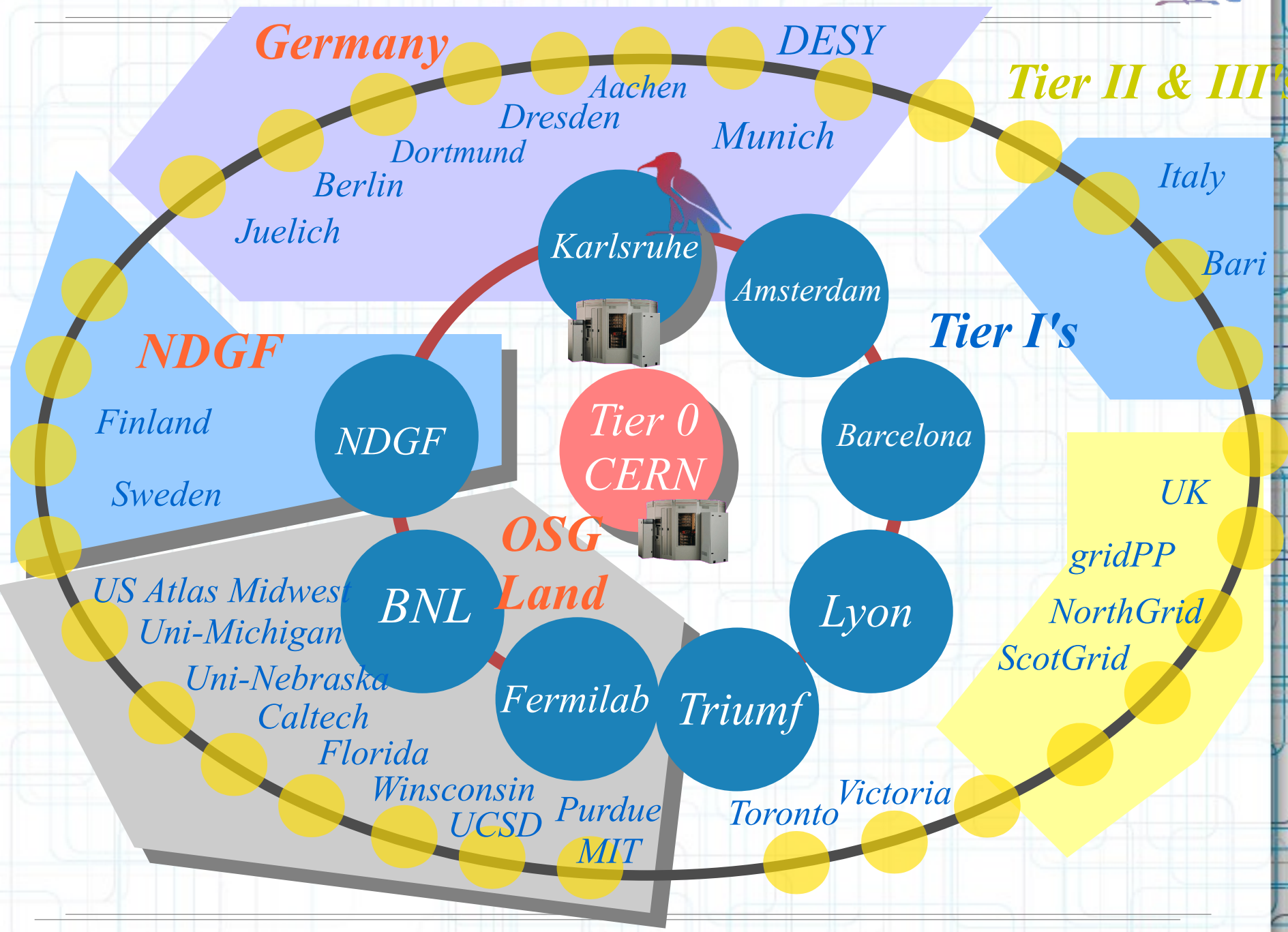
40 Tier II centers around the world

- Germany
- US
- UK
- Italia
- Australia

dCache will manage the largest share of LHC data outside CERN.

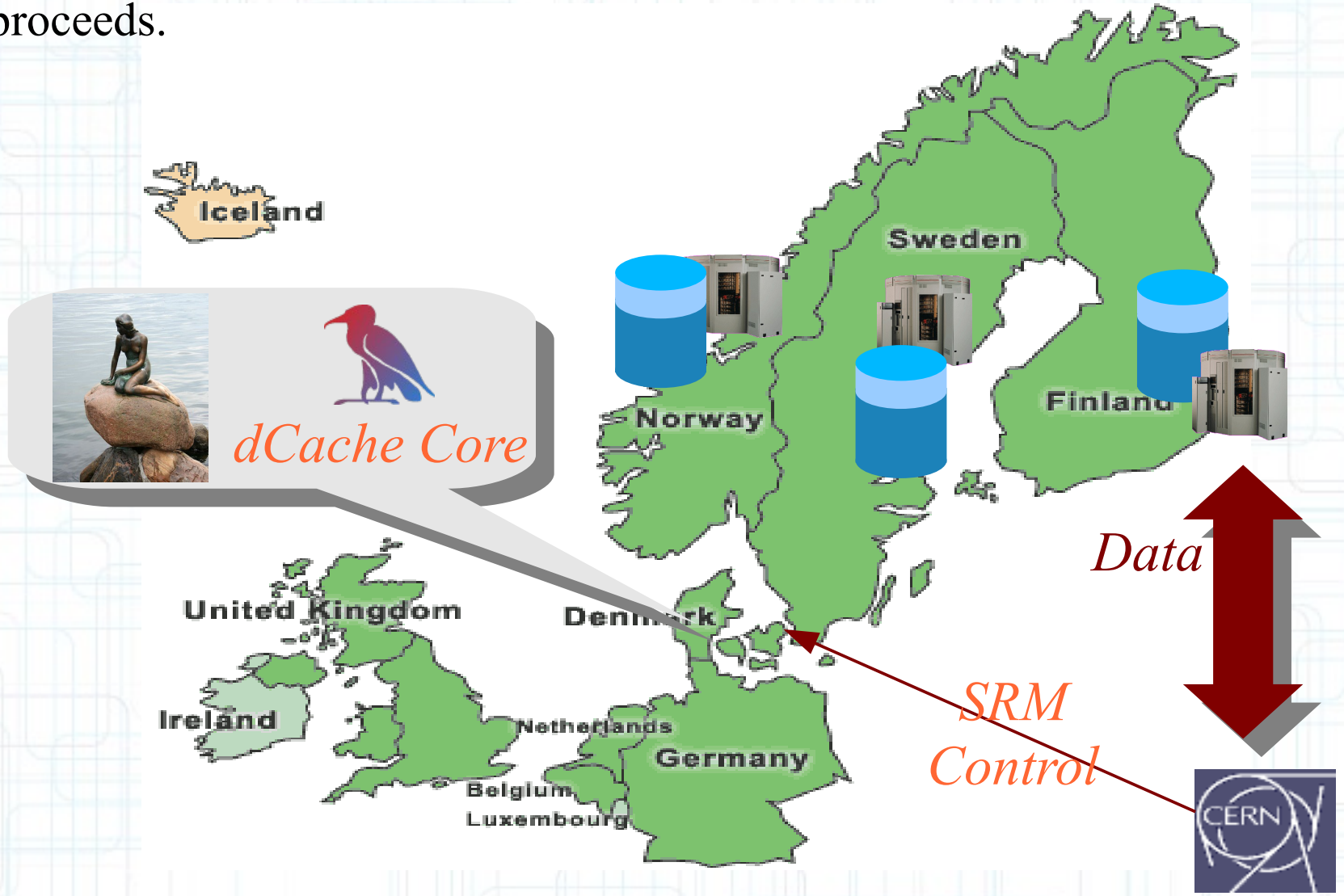
dCache.org

8 out of 11 Tier I's and many Tier II/III's using dCache



Most prominent dCache users (NDGF)

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



dCache.org

dCache Future

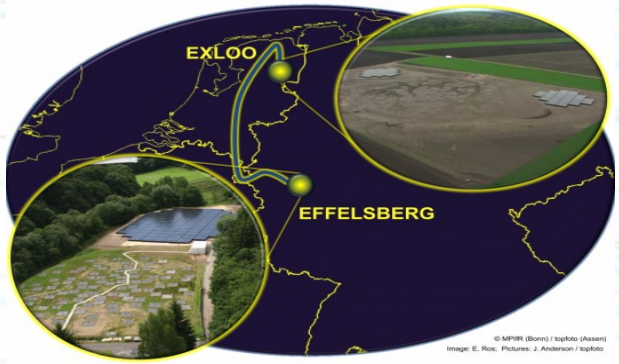
dCache.org

Challenges

New large experiments are on the horizon

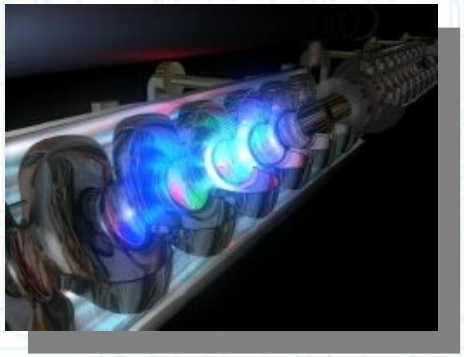
Low Frequency Array (LOFAR)

European Radio Telescope



European X-Ray Free Electron Laser (X-FEL)

extremely intense X-ray flashes

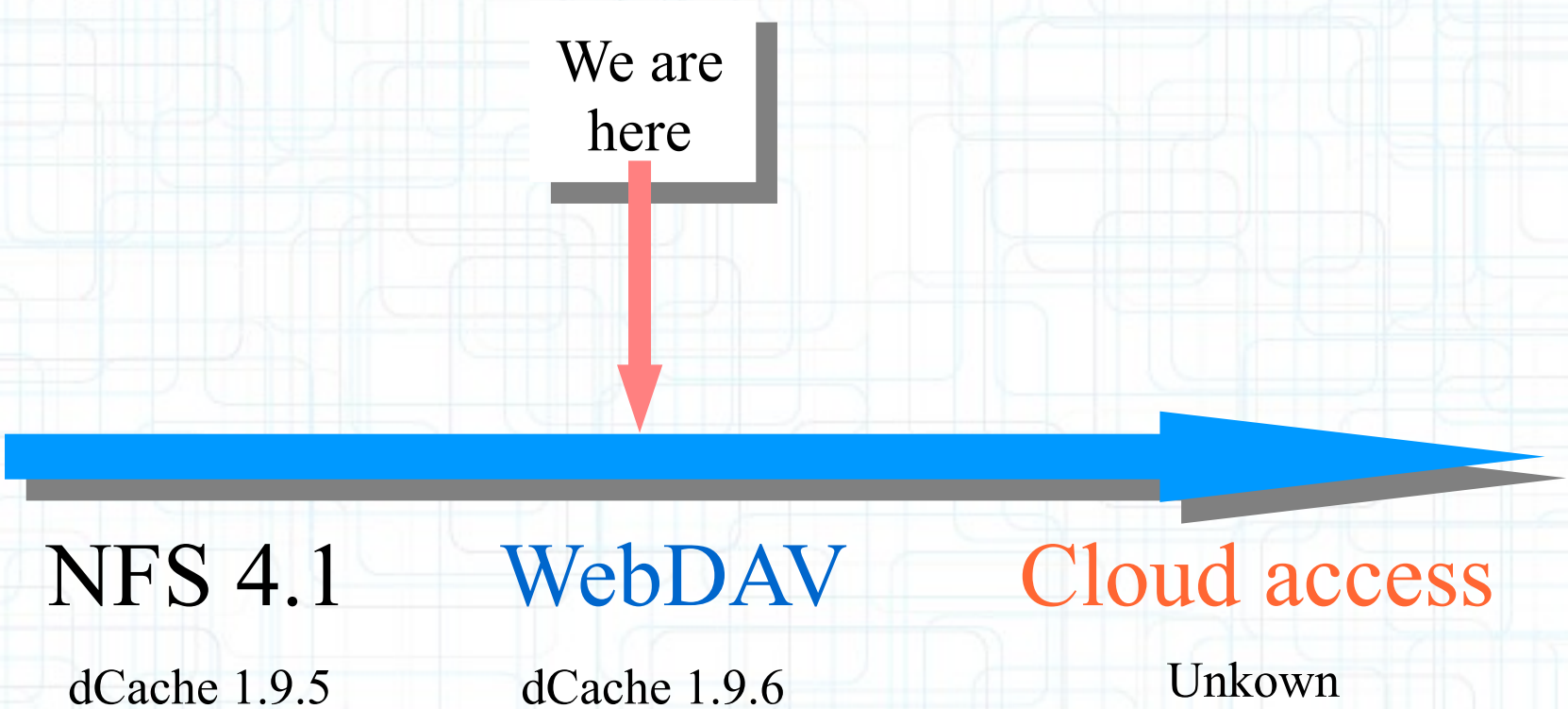


- Expected data rates exceed LHC data rates
- New non-HEP experiments build on standard protocols

dCache.org

How is dCache prepared for this

- No problem with scalability and data rates
- Data Management is briefly solved
- But we need more standards to attract non HEP Communities



dCache.org

dCache.org

Some remarks on NFS 4.1

Advantages

- NFS 4.1 (pNFS) can take advantage of distributed data
- Security is part of the specification (GSSAPI)
- Compound RPCs (faster)
- Client software is provided by the OS distributors/vendors

Solves Prof. Lang's slide
“Reality of data storage ..”

NFS 4.1 Contributors

Coordinated by the Center of Information Technology Integration (U. Michigan)
Slide is stolen from "Lisa Weeks" presentation :
pNFS: Blending Performance and Manageability

Blue Arc
CITI
CMU
EMC
IBM
LSI
OSU
Net App
Ohio SuperComputer
Panasas
Seagate
StorSpeed
Sun Microsystems
Desy

Clients

- Sun (Files)
- Linux (Files / Blocks / Objects)
- Desy / dCache (Java-based / Files)

Servers

- Sun (Files)
- Linux (Files)
- NetApp (Files)
- EMC (Blocks)
- LSI (Blocks)
- Panasas (Objects)
- Desy / dCache (Java-based / Files)

dCache.org

dCache.org

Some remarks on Cloud

Data Clouds

dCache.org

No Cloud data access protocols standardized, yet

Good candidates :

Amazon
CDMI (SNIA, OGF)

dCache, unique features

- dCache manages data through SRM 2.2
  - Remote management
  - Access Latency, Retention Policy (Tape/Disk)
- dCache provides internal file location management through file and client attributes. (IP number, directory, transfer direction, space tokens...)
  - Different Storage Nodes have different duties.
  - Golden Hardware for safe storage (Raid6)
  - Cheap hardware for multiple copies (fast read access)
- Automatic migration and restoring from external tape systems.
- Easy storage hardware maintenance : Adding and Draining pools without system interruption.
- Incorparated industry standards : NFS4.1, SRM2.2, WebDAV

dCache.org

The Team

Head of dCache.ORG

Patrick Fuhrmann

Core Team (Desy, Fermi, NDGF)

Andrew Baranovski
Gerd Behrmann
Bjoern Boettscher
Ted Hesselroth
Alex Kulyavtsev
Iryna Koslova
Tanya Levshina
Dmitri Litvintsev
David Melkumyan
Paul Millar
Owen Synge
Neha Sharma
Vladimir Podstavkov
Tatjana Baranova
Jan Schaefer

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

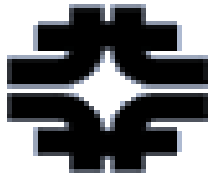
dCache.org

dCache.org

Further reading

www.dCache.ORG

dCache is a collaboration of



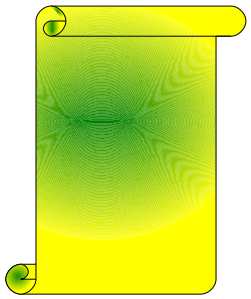
Getting a file out of the GRID

dCache.org

Logical File Name

MyPrecious

Global File Catalogue



Storage (S)URL(s)

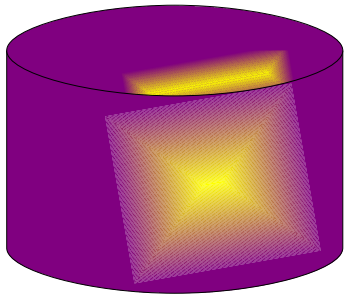
srm://srm.DESY.de/path

srm://srm.CERN.ch/path

Storage URL(s)

srm://srm.DESY.de/path

Storage Element



Transfer (T)URL(s)

gsiftp://dcache.DESY.de/path2

Do finial transfer with

gsiftp://dcache.DESY.de/path2

Return TURL