dCache

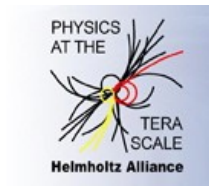## dCache

Future of European Middle-ware(s)

Patrick Fuhrmann et al.

additional funding, support or contributions by

d-grid
DGI II

What is dCache ?

What is dCache.org ?

Who is contributing to dCache ?

Who is is using dCache ?

How is dCache used (The LHC Storage Element) ?

Enhanced usage : The NDGF (NorduGrid) approach !!!

Preparing dCache for the future !
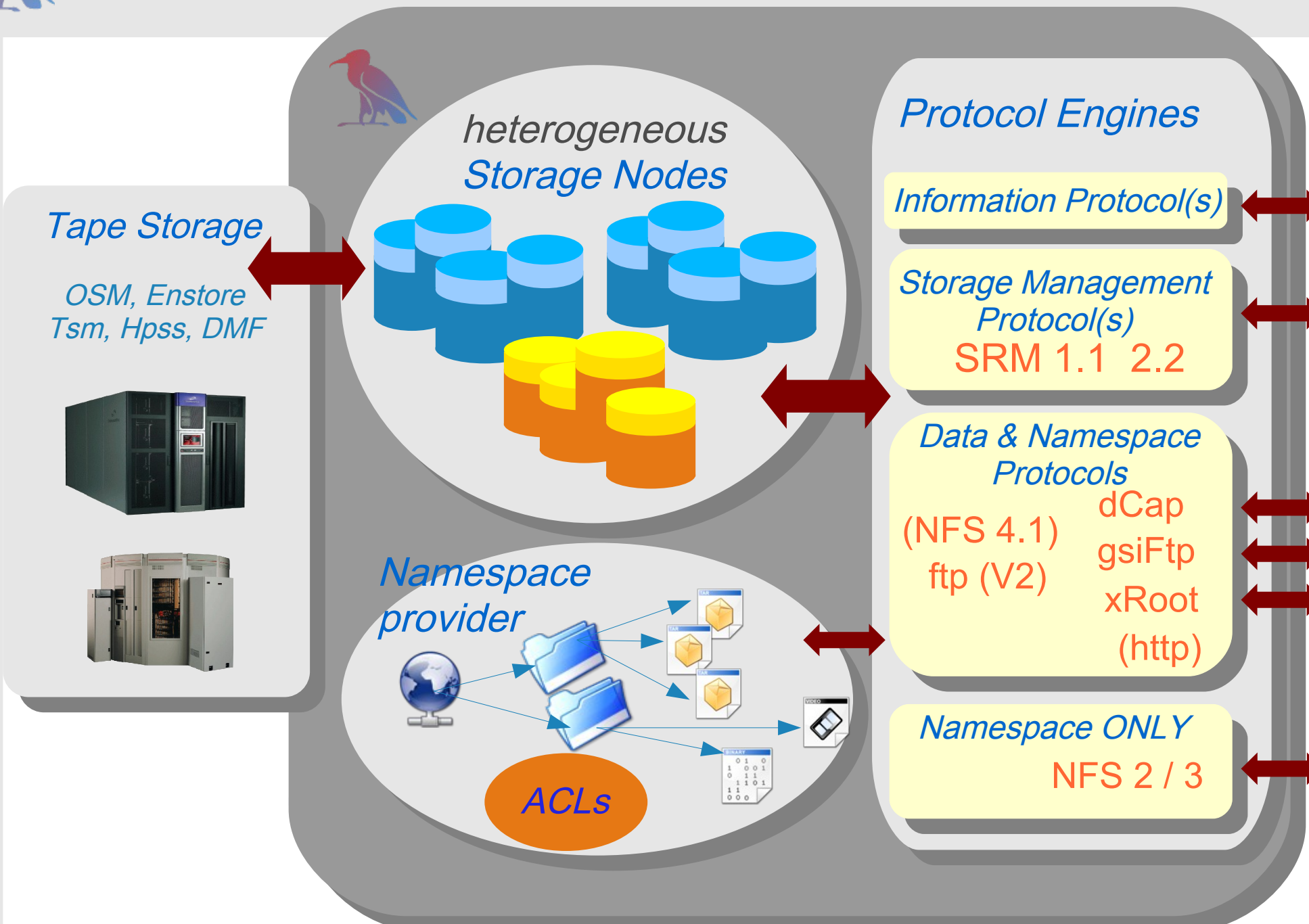
What is dCache ?

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*

What is dCache, some basics?

heterogeneous Storage Nodes

Tape Storage

OSM, Enstore Tsm, Hpss, DMF

Namespace provider

ACLs

Protocol Engines

Information Protocol(s)

Storage Management Protocol(s)
SRM 1.1 2.2

Data & Namespace Protocols
(NFS 4.1)
ftp (V2)

dCap
gsiFtp
xRoot
(http)

Namespace ONLY
NFS 2 / 3

In a Nutshell

* 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

  - xRoot

  - SRM

  - nfs2/3 (name space only)

  - NFS 4.1 with dCache 1.9.4
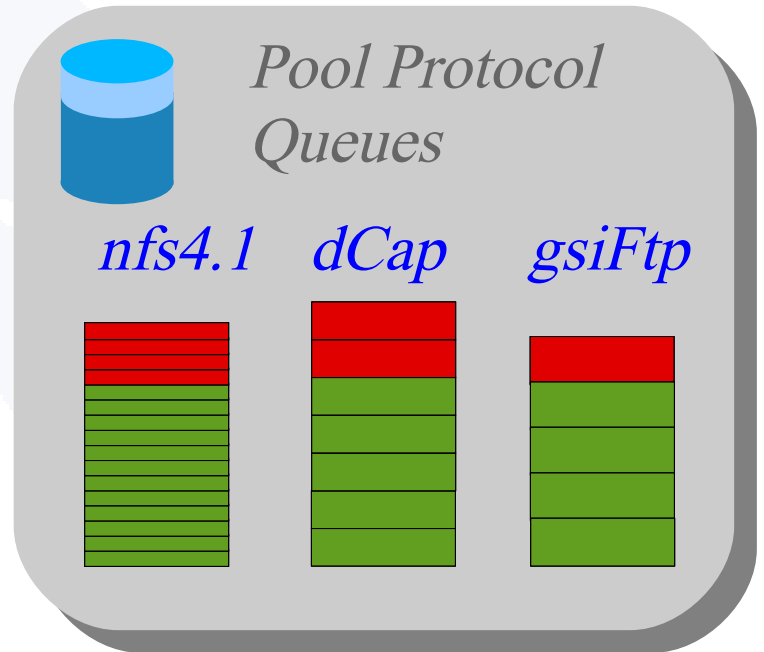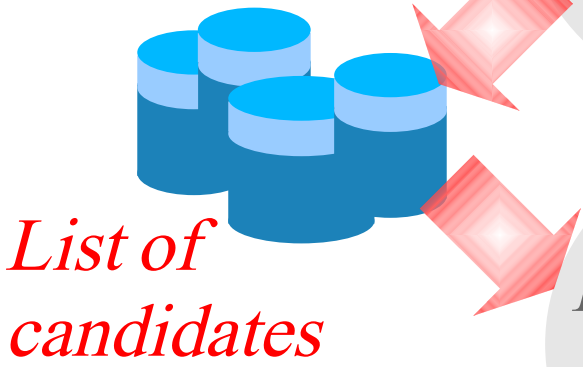
dCache in a Nutshell

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
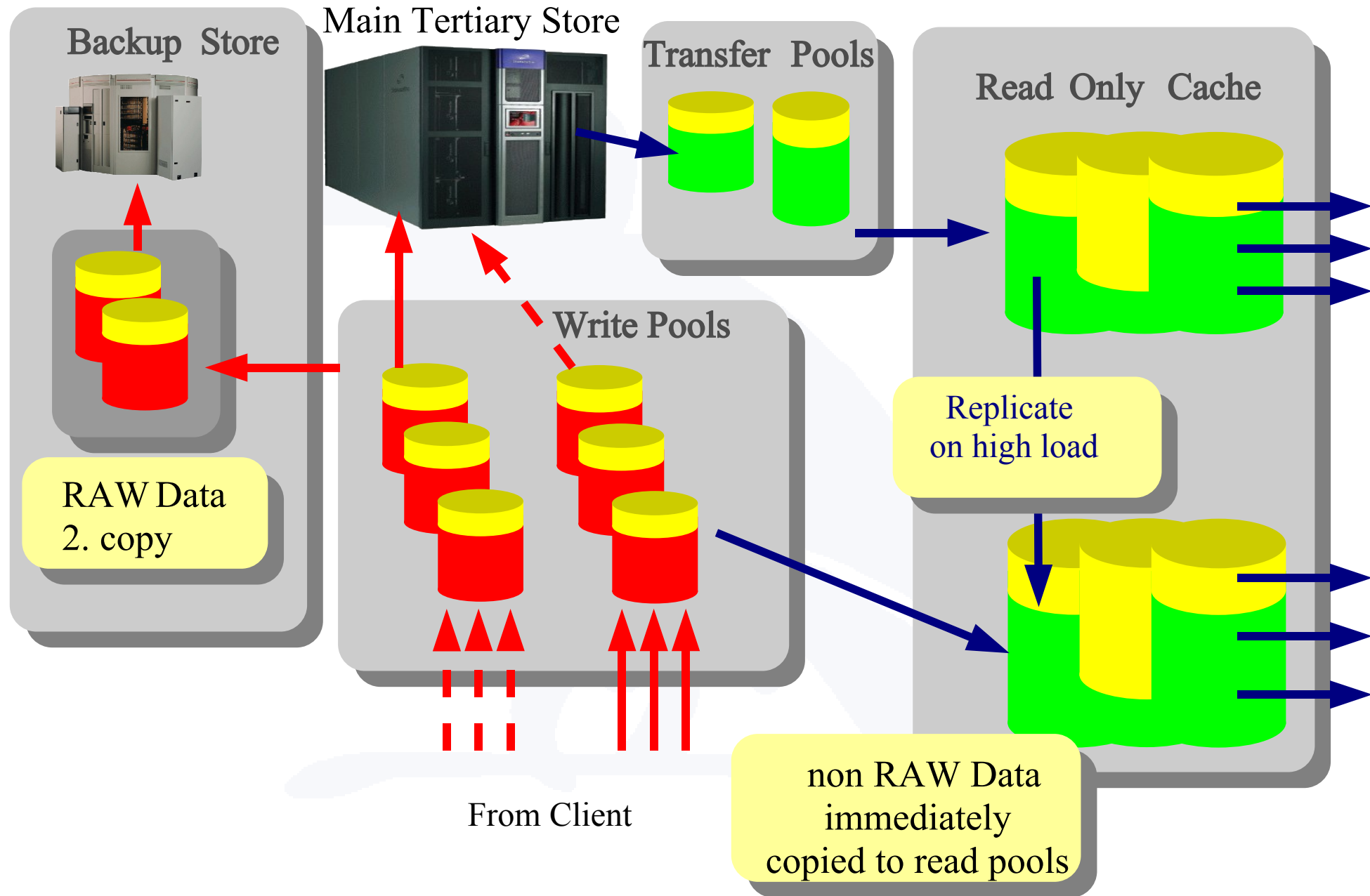
Dispatcher by Pool Cost

Pool Protocol Queues

nfs4.1   *dCap*   *gsiFtp*

dCache.ORG

dCache.ORG

Backup Store

Main Tertiary Store

Transfer Pools

Read Only Cache

RAW Data
2. copy

Write Pools

Replicate
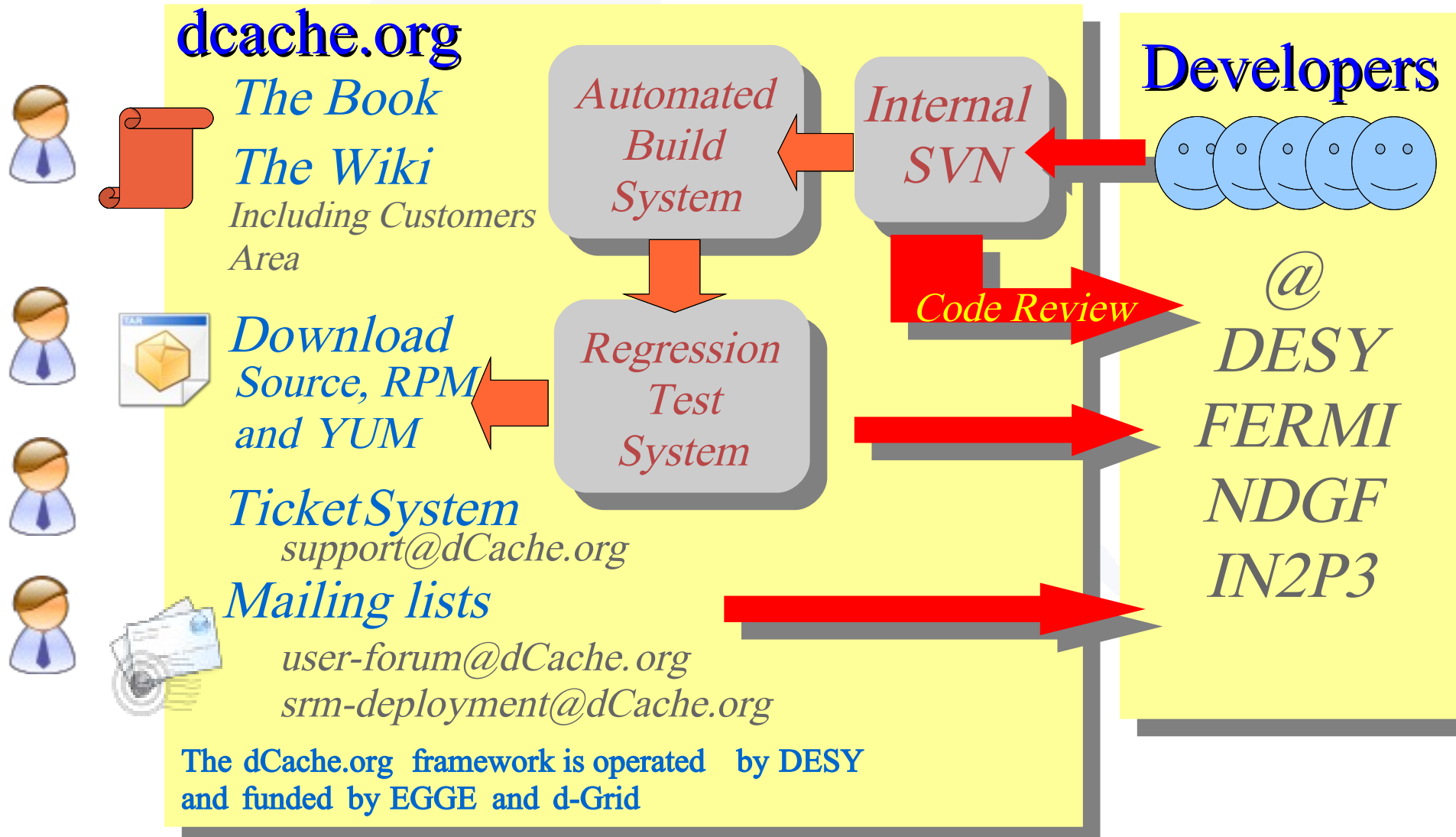on high load

From Client

non RAW Data
immediately
copied to read pools

What is dCache.org ?

dCache.ORG

- *dCache.ORG is an infrastructure*
- *dCache.ORG is the door into the dCache team*

dcache.org

The Book

The Wiki
Including Customers
Area

Download
Source, RPM
and YUM

Ticket System
support@dCache.org

Mailing lists
user-forum@dCache.org
srm-deployment@dCache.org

Automated Build System

Internal SVN

Developers

Regression Test System

Code Review

@
DESY
FERMI
NDGF
IN2P3

The dCache.org framework is operated by DESY
and funded by EGGE and d-Grid

dCache distribution

- dCache is distributed (YUM at DESY and CERN) and configured through YAIM for the Tier II's in Europa.
- dCache is distributed and configured through VDT for OSG supported Tier II's.
- dCache is distributed through dCache.org for the Tier I's and configured manually due to the complexity of the Tier I setups.
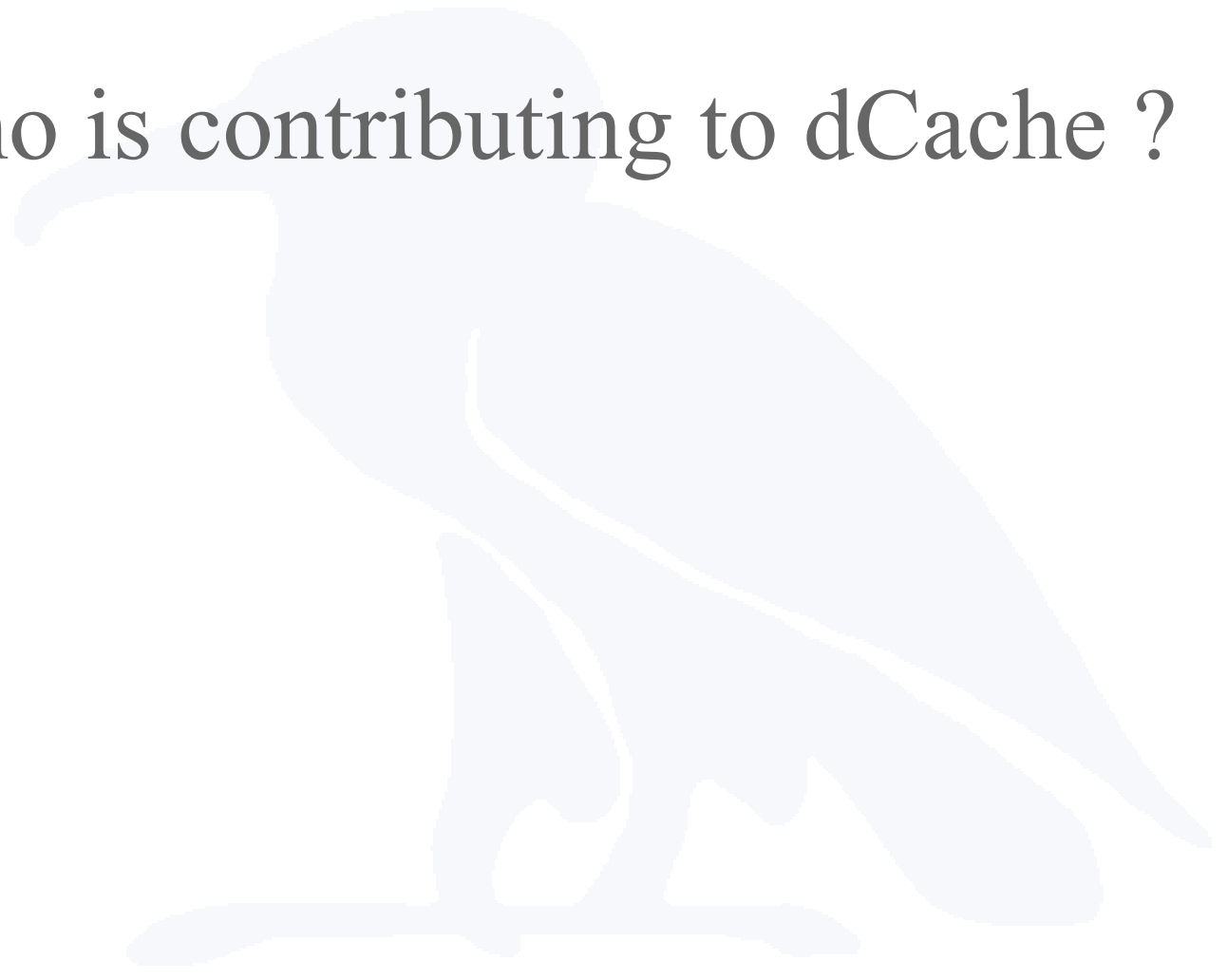
The source of the dCache software is always dCache.ORG.

- Redistribution is ok.
- Support is ok.
- Repackaging is not permitted.

The Role of UMD, related to dCache, is not clear.

Who is contributing to dCache ?

Who is contributing (People)

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

2 * N.N.

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

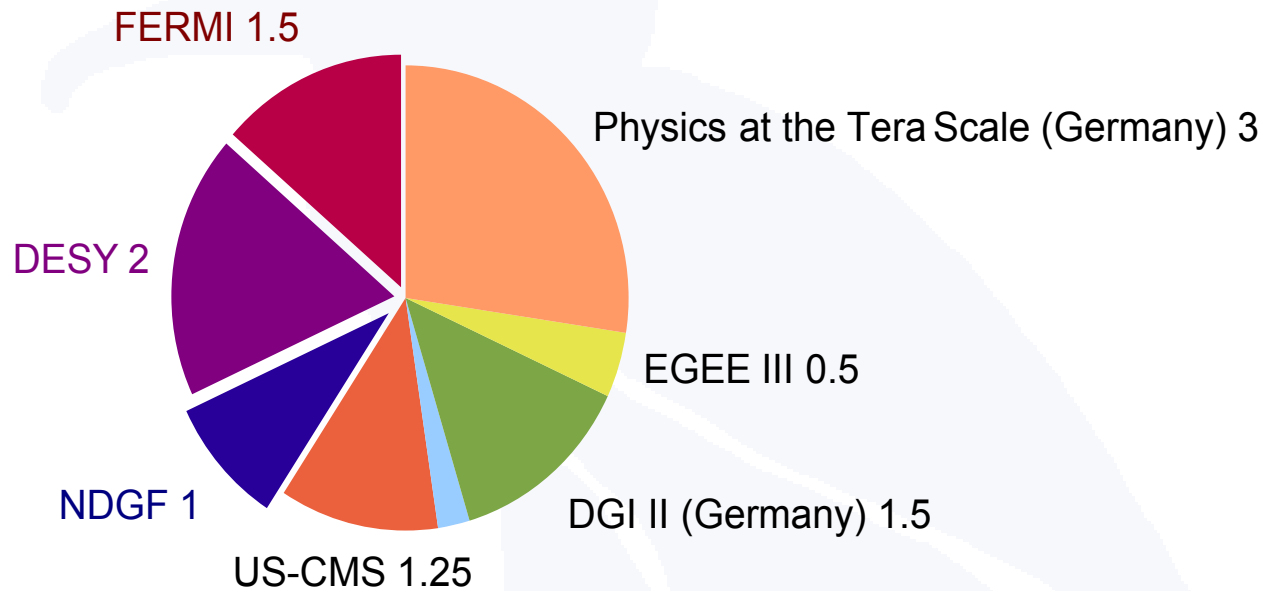German HGF Support Team
Greig Cowan, gridPP
Flavia Donno, CERN

dCache.ORG

dCache.ORG

FTE's by organisation.

FERMI 1.5

Physics at the Tera Scale (Germany) 3

DESY 2

EGEE III 0.5

NDGF 1

DGI II (Germany) 1.5

US-CMS 1.25

Focus moving from development to support.

Who is using dCache ?

dCache is primarily used by LHC/LCG Tier I , II and III sites.

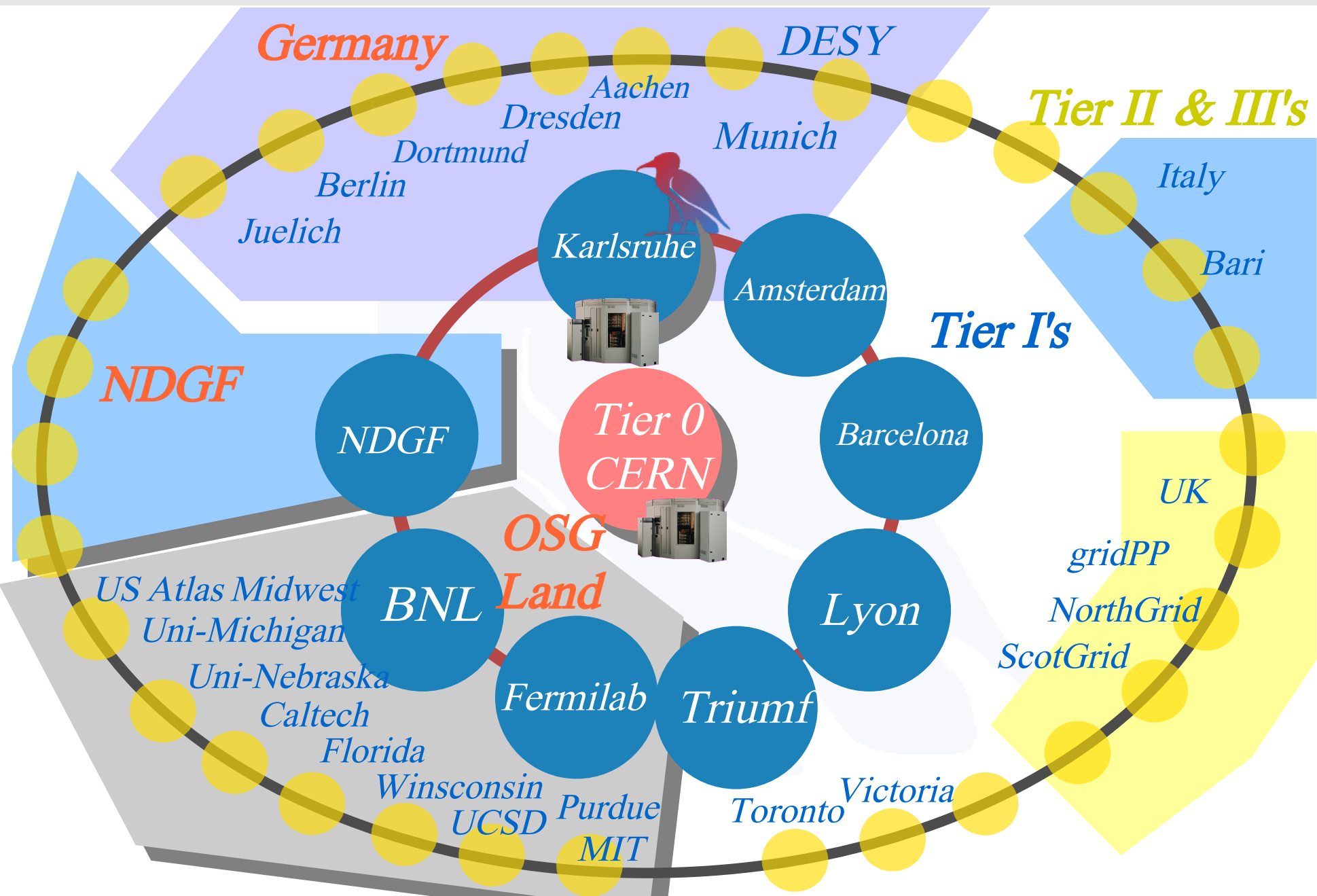In this context, dCache will hold the largest share of LHC data

outside CERN

Total Incoming Traffic into UCSD (TierII) Sep 10, 2008
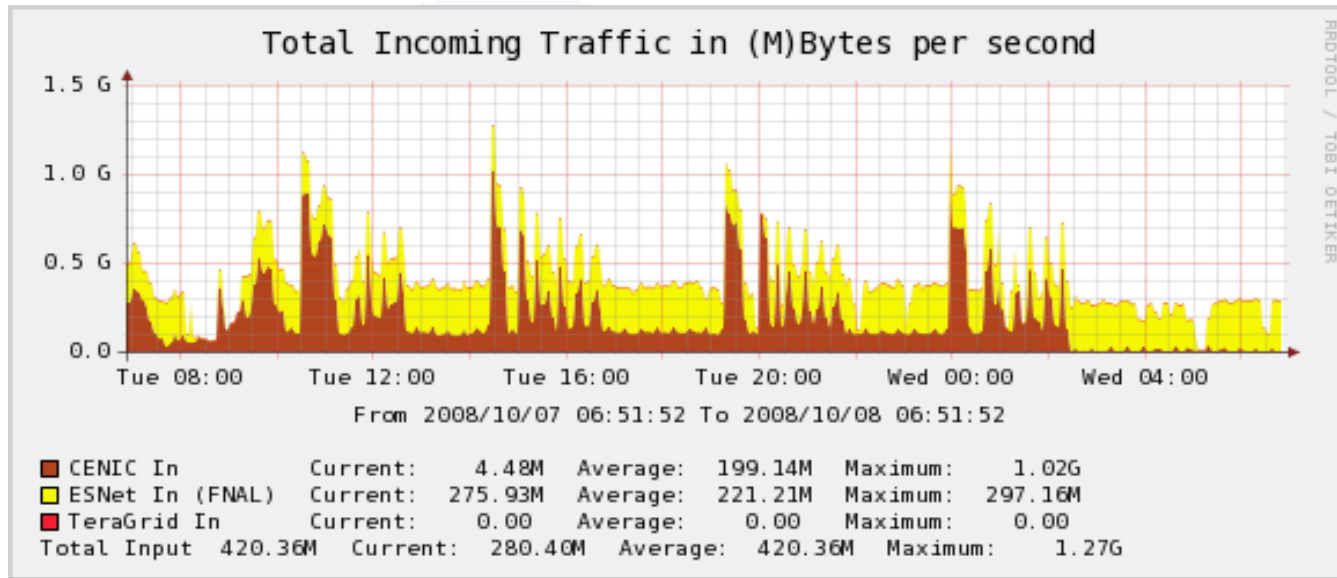
FERMILab US-CMS dCache is running on more than 600 pools and stores about 3 PB on disk and 2 PB on tape in a single dCache instance.

BNL will grow beyond the 10 Pbytes storage range starting 2010/11

How is dCache used ?

or

Standards in the LHC Storage Element

dCache.ORG

dCache.ORG

Tier 0 Storage Element

Wide area transfers

Tier I Storage Element

Local Area Access

Compute Cluster

Tier II Storage Element

Local Area Access

Compute Cluster
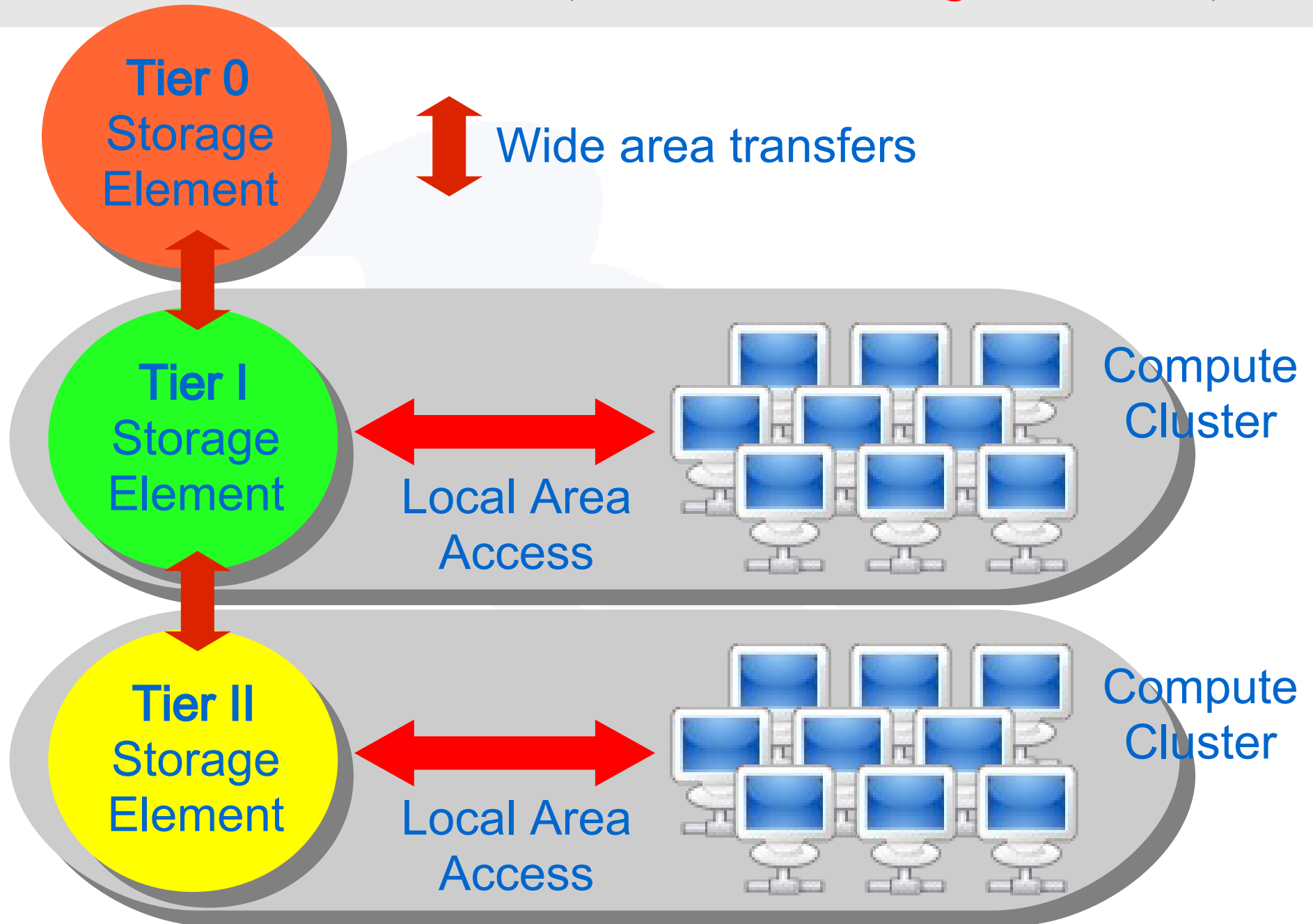
A Storage Element is a Grid entity.

Grid is about standards (power plug model)

How about Storage Element standards ?

dCache.ORG

dCache.ORG

Storage Element

Space/Protocol Management

SRM Storage Resource Management

OGF

Wide Area Transport Protocol
In use : gsiFtp
Discussed : http(s)

IETF

Information Service Protocol
Transport : LDAP
Content : GLUE Schema

IETF
OGF

Local Access Protocol
(gsi)dCap or rfio and xRoot

These is not at all a standard

dCache.org is proposing the NFS 4.1 protocol as the local data access to Grid Storage Elements !

- ➢ NFS 4.1 supports highly distributed data within the same file system tree.

- ➢ NFS 4.1 is becoming an IETF standard.

- ➢ Standards give industry a fair chance to compete.

- ➢ Other grid communities refuse to adopt special HEP solutions.

- ➢ They prefer real posix I/O. (Direct file system access)

- ➢ Client software is provided by OS. Makes grid deployment people happy.

Grid Standards : Breaking News

NFS 4.1 Bakeathons last month :

- ➢ dCache server can talk to all known NFS4.1 clients

- ➢ Some limitations : no modify, no striping but not a problem for clients

- ➢ NFS4.1 will be in official standard linux kernel Q1 2009

- ➢ full IETF approval till end of 2008

- ➢ Client versions
  - ➢ Linux
  - ➢ SUN (Solaris)
  - ➢ CITI will work on Mircosoft client very soon

- ➢ Server vendors : IBM, SUN, Panasas, netApp, LSI, EMC, dCache

Enhanced dCache usage

or

The NDGF (NorduGrid) challenge !!!
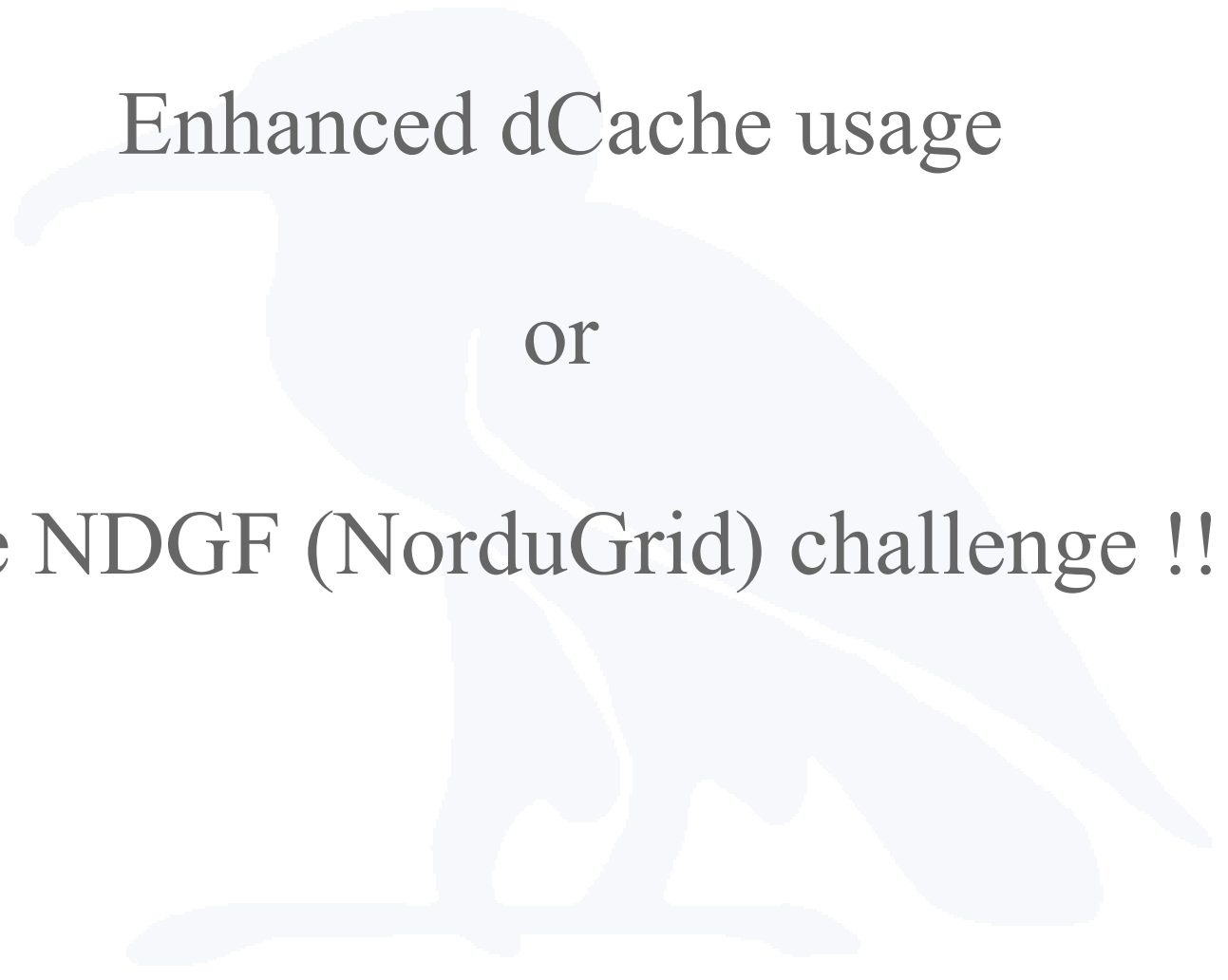
- 4 Countries, one dCache instance.
- At any time a country may 'go down' though raw data storage proceeds.

dCache Core
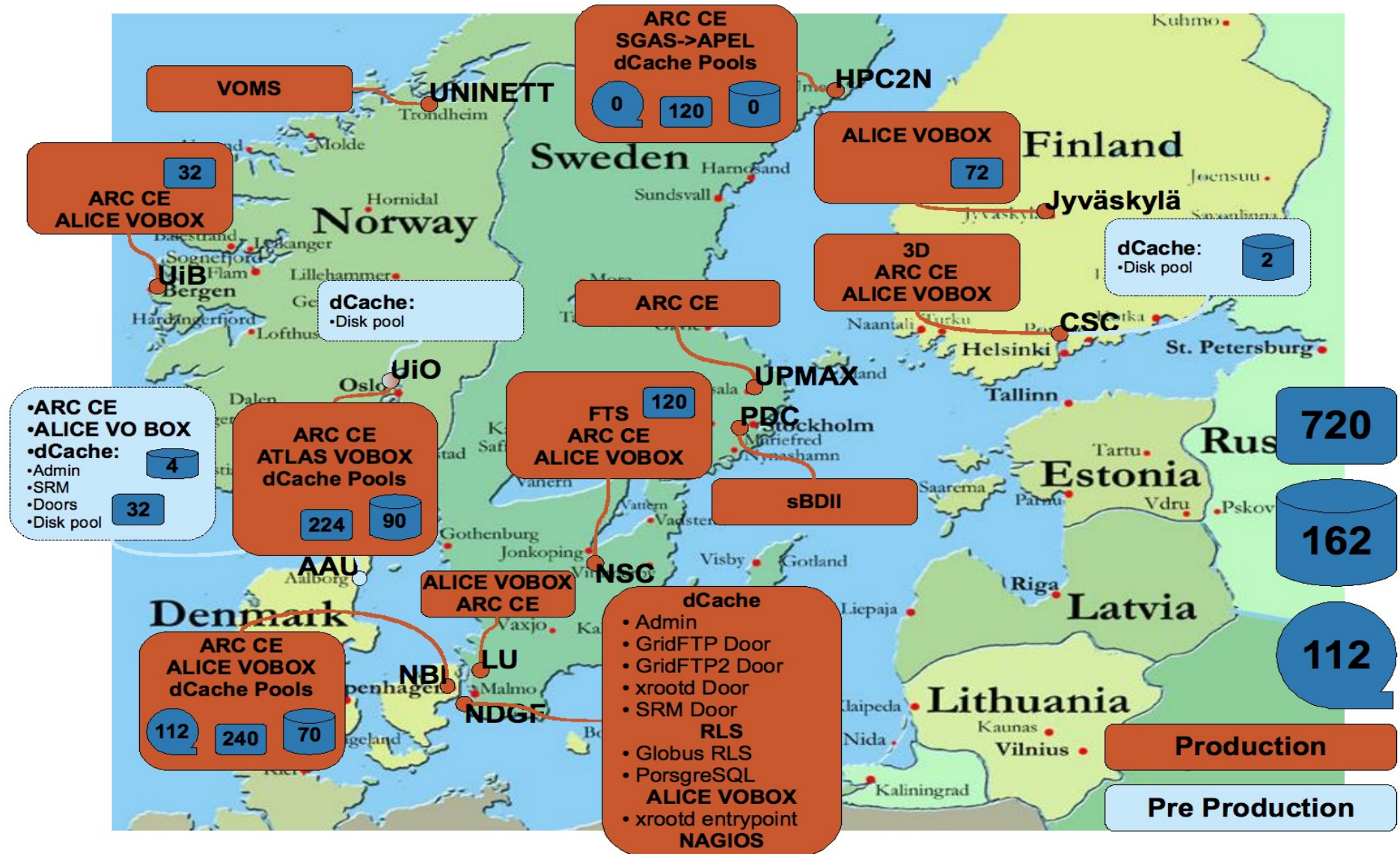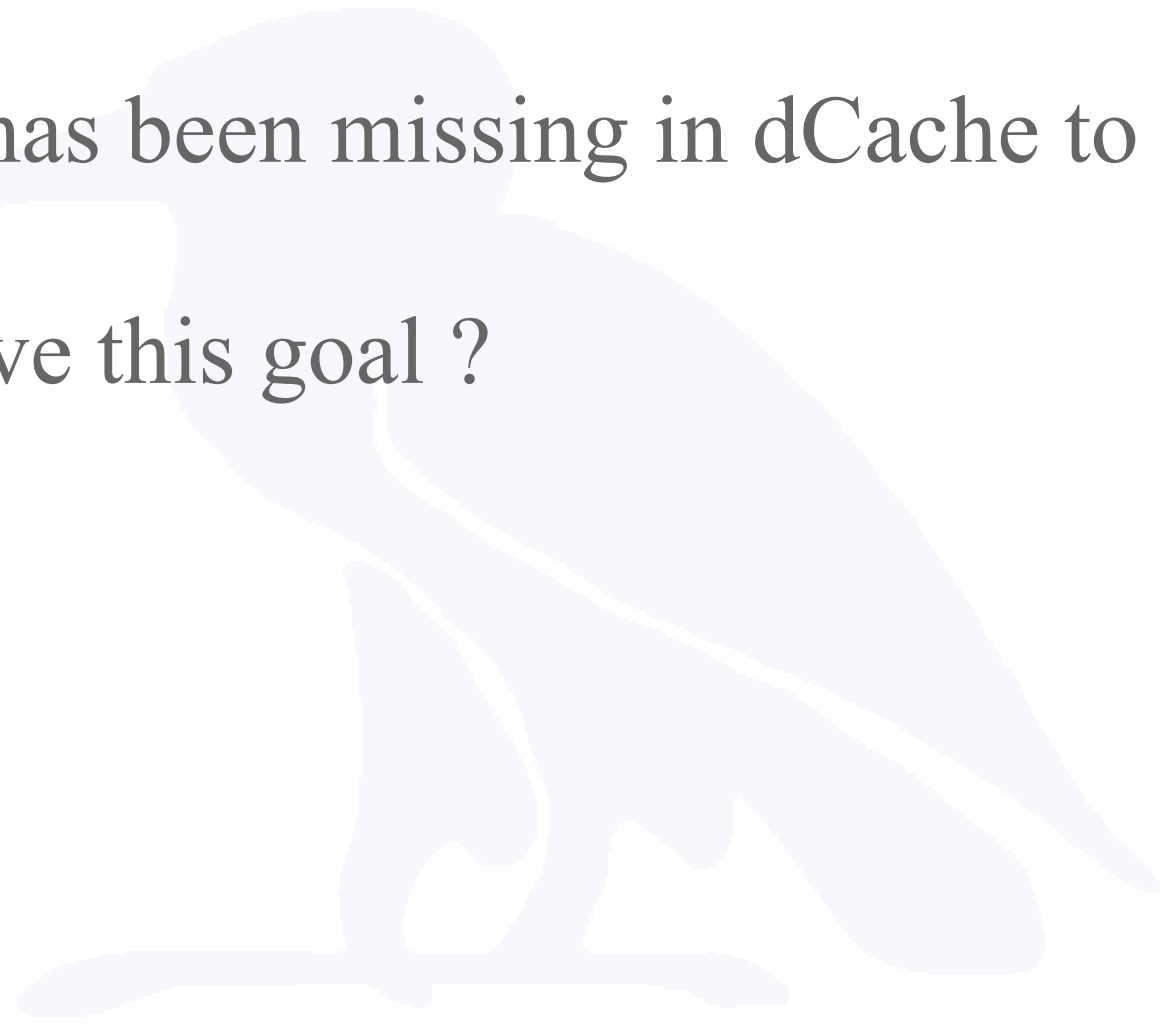
Iceland

Sweden

Norway

Finland

United Kingdom

Denmark

Ireland

Netherlands

Germany

Belgium

Luxembourg

Data

SRM Control

CERN

dCache.ORG

dCache.ORG

dCache.ORG

dCache.ORG

VOMS

UNINETT

ARC CE SGAS->APEL dCache Pools
0 120 0

HPC2N

ALICE VOBOX
72

Finland

Jyväskylä

32

ARC CE ALICE VOBOX

UiB

3D ARC CE ALICE VOBOX

dCache:
•Disk pool
2

dCache:
•Disk pool

ARC CE

CSC

UPMAX

UiO

•**ARC CE**
•**ALICE VO BOX**
•**dCache:**
•Admin
•SRM
4
•Doors
•Disk pool
32

ARC CE ATLAS VOBOX dCache Pools
224 90

FTS ARC CE ALICE VOBOX
120

PDC

720

162

112

sBDII

AAU

ALICE VOBOX ARC CE

NSC

dCache
• Admin
• GridFTP Door
• GridFTP2 Door
• xrootd Door
• SRM Door
  RLS
• Globus RLS
• PorsgreSQL
  ALICE VOBOX
• xrootd entrypoint
  NAGIOS

Production

ARC CE ALICE VOBOX dCache Pools
112 240 70

NBI

LU

NDGF

Pre Production

What has been missing in dCache to

achieve this goal ?

Requirements for the NDGF dCache usage ?

- ➤ GridFtp 2 protocol support in dCache and in the LCG storage toolkit.

- ➤ Location context dependent HSM support.

- ➤ Minimise synchronised component updates.

- ➤ Secure component communication.

- ➤ Independent admin domains.

- ➤ Network topology awareness of dCache.

Beside presenting the list of requirements,

the provided help.

A professional developer for NDGF related
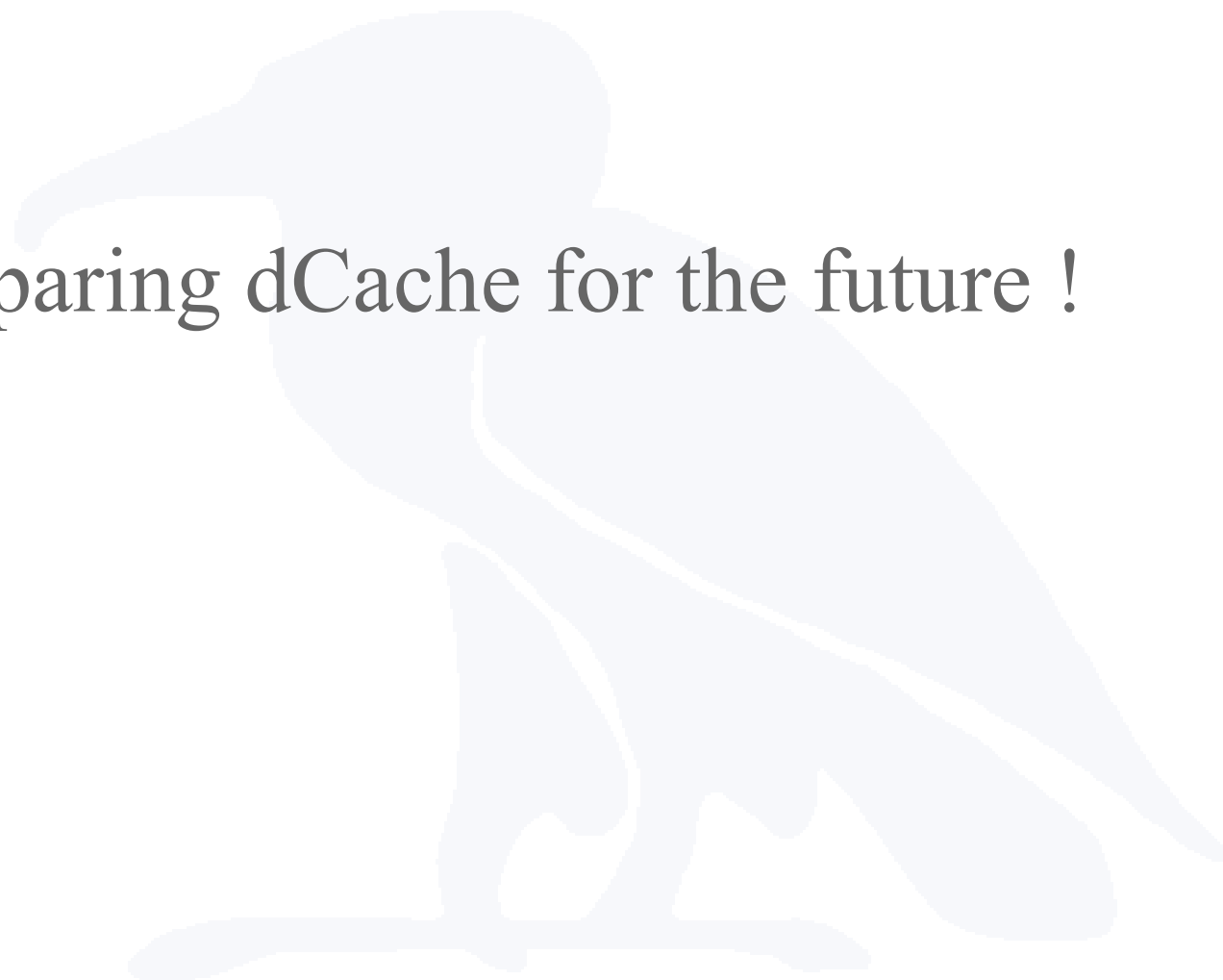
dCache development as part of the team.

This turned out to be a very successful approach.

Preparing dCache for the future !

dCache.ORG

dCache.ORG

Two equally important tasks to be prepared

for the future

* Implementing new requirements.

* Re-factoring existing code.

dCache.ORG

dCache.ORG

2009

Re-factored component

New component/feature

Name space provider.

Information service

NFS 4.1

ACL for file system

Protect Robotics

dCache prepared for LHC analysis

ACL for space tokens
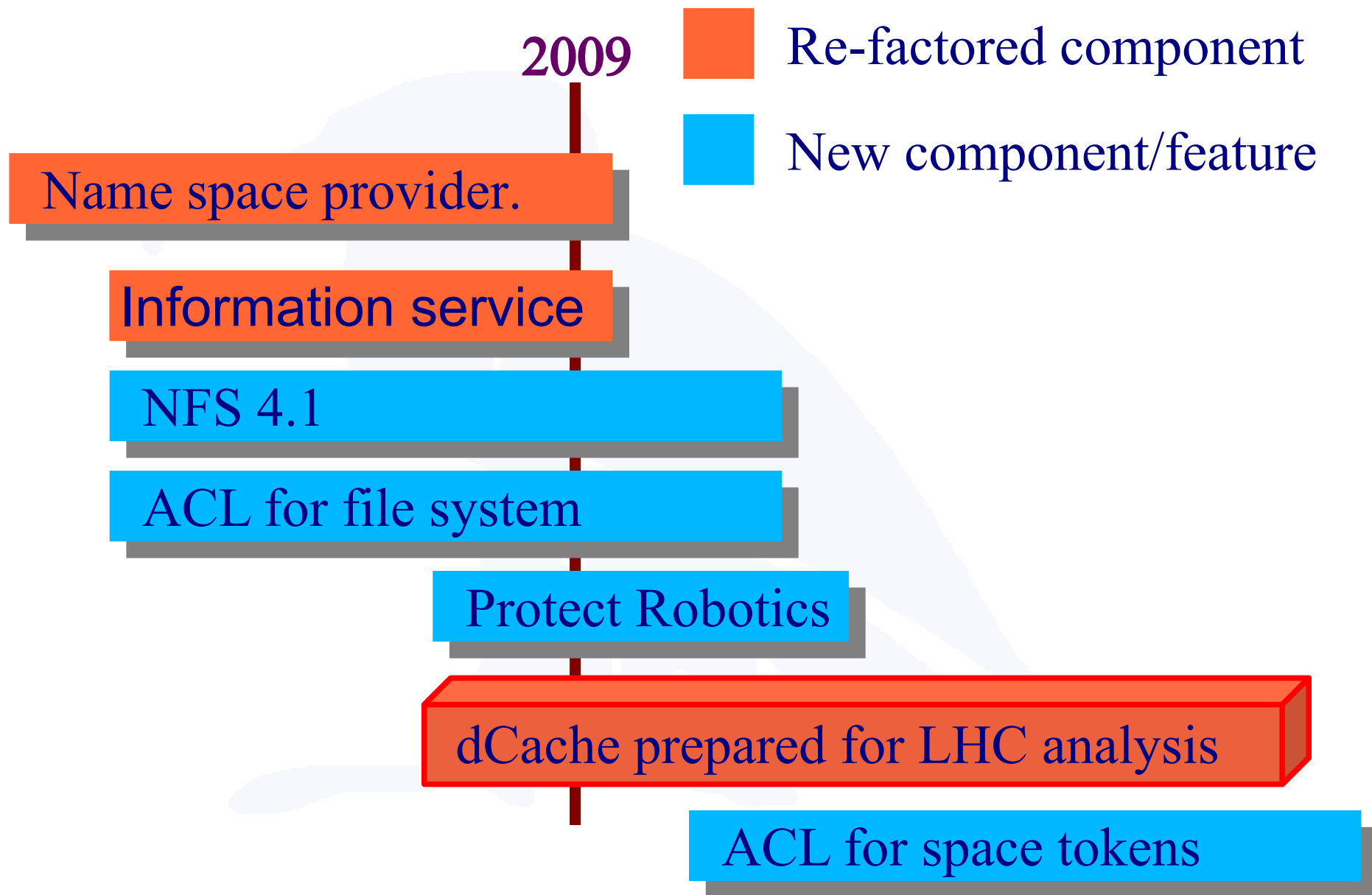
dCache prepared for LHC analysis

What does this mean.

- ➤ Huge amount of files are opened simultaneously.

- ➤ Only a small fraction of a file is actually read.

Possible solution

- ➤ Define areas in dCache with little file fluctuation

- ➤ Caching file location(s) within those areas

- ➤ This should allow rapid file 'opens'.

Conclusion

- The dCache software is doing well and covers LHC requirements especially concerning data volume and speed.

- dCache will hold most of the LHC data outside CERN.

- The dCache project model seems to be successful. Contributors, like OSG and NDGF, are welcome.

- dCache tries to attract non HEP communities.

- dCache tries to push for more industry standards.

2 job positions offered at DESY

(Europe,Germany, Hamburg),

starting end of October.

WE NEED YOU

Further reading

www.dCache.ORG