dCache, the Challenge

Patrick
for the dCache Team

support and funding by

DESY

NDGF
Nordic DataGrid Facility

HEP CG
D-Grid

Open Science Grid

VDT
Virtual Data Toolkit

egee
Enabling Grids
for E-sciencE

Topics

Technical Introduction or What is a dCache ?

Black Box View
In a nutshell
Technical Example : the Request View

and of course : What is dCache.ORG ?

and not to forget : What is this Chimera thing ?

What are we currently working on ?

The NDGF Challenge (Very short term)

NFS 4.1 (Mid Term)

SRM (Example : the space token)

dCache 1.8 Deployment time-line / schedule

Project Topology : The Team

Head of dCache.ORG

Patrick Fuhrmann

Core Team (Desy and Fermi)

Andrew Baranovski
Bjoern Boettscher
Ted Hesselroth
Alex Kulyavtsev
Iryna Koslova
Dmitri Litvintsev
David Melkumyan
Dirk Pleiter
Martin Radicke
Owen Synge
Neha Sharma
Vladimir Podstavkov

Head of Development FNAL :

Timur Perelmutov

Head of Development DESY :

Tigran Mkrtchyan

External Development

Gerd Behrmann, NDGF
Jonathan Schaeffer, IN2P3

Support and Help

Abhishek Singh Rana, SDSC
Greig Cowan, gridPP
Stijn De Weirdt (Quattor)
Maarten Lithmaath, CERN
Flavia Donno, CERN

Responsibilities

DESY

- dCache.ORG infrastructure
- Cell Communication System
- dCache core Services : PoolManager, Pnfs/ChimeraManager, Pools, (gsi)dCap doors and mover ....
- File Systems : Pnfs , Chimera
- Upcoming : NFS 4.1, HSM controller
- Building, Regression Tests and Publishing
- Yaim Integration : sl3/sl4 32/64 bit ; In Progress : Solaris
- LCG gLite Integration

FERMILab

- SRM 1.1 and SRM 2.2
- Space Management
- Authorization, Authentication gPlazma
- gsiFtp doors and movers
- resilient Manager
- OSG VDT Integration

Responsibilities

NDGF

- Multi Hsm Support
- gsiFtp Protocol Version 1 & 2 doors and movers
- Code Review
- Various

BNL

- Horizontally scaling of SRM
- BNL Specific Issues

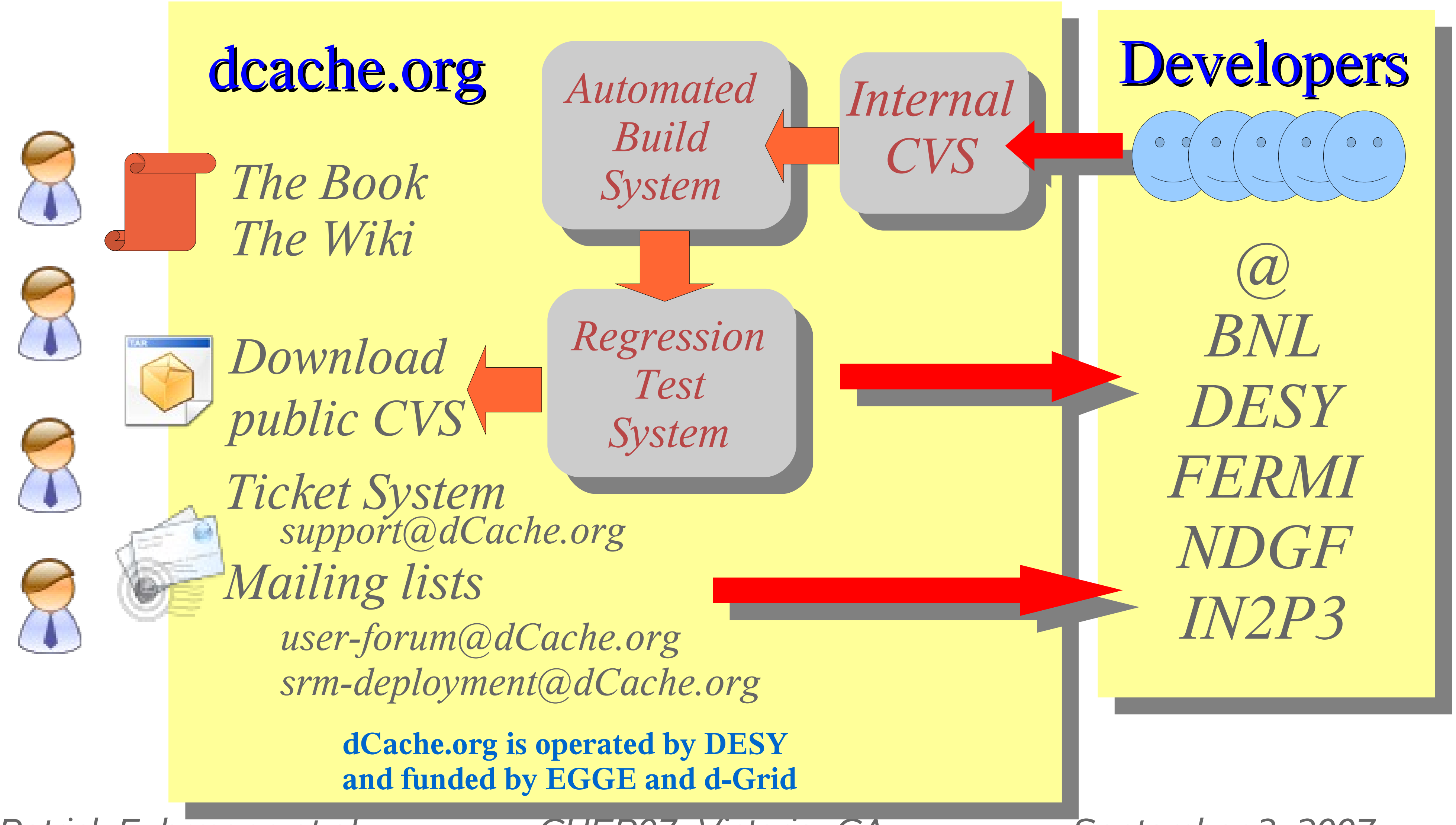
IN3P3

- Various

What is dCache.ORG ?

What is dCache.ORG

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



dcache.org

The Book
The Wiki

Download
public CVS

Ticket System
support@dCache.org

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

Automated Build System

Internal CVS

Regression Test System

Developers
@
BNL
DESY
FERMI
NDGF
IN2P3

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

Technical Introduction or What is a dCache ?

dCache complies to the definition of an WLCG Storage Element.

dCache will store the largest share of the LHC data for the first years of data-taking.

at 7 Tier I's : NDGF, IN2P3, SARA, FERMI, BNL, FZK, PIC and numerous Tier II

Technical Introduction
Black Box View

High Level Services

Resilient Manager

Admin Module (ssh, jpython)

Maintenance Module

Flush Manager

Hopping Manager

dCache Core

Information Protocol(s)

Storage Management Protocol(s)

SRM 1.1 2.2

Tape Storage

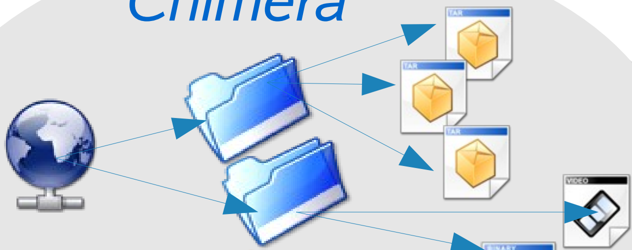
OSM, Enstore
Tsm, Hpss, DMF

Data & Namespace Protocols

(NFS 4.1)
ftp (V2)

dCap
gsiFtp
xRoot
(http)

Chimera



ACLs

Namespace ONLY

NFS 2 / 3

In the Nutshell Basic Feature Set

- 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 http)*
- Automated file replication on access hot spot detection
- HSM connectivity (enstore,osm,tsm,hpss, dmf)
- Automated HSM migration and restore.
- Handles data in Peta-byte range on 1000's of pools
- Supported protocols : (gsi)ftp , (gsi)dCap, xRoot, SRM, nfs2/3
- Separate I/O queues per protocol
- Supports resilient dataset management (worker-node support)
- Sophisticated command line interface and graphical interface

In the Nutshell New Features in 1.7.0

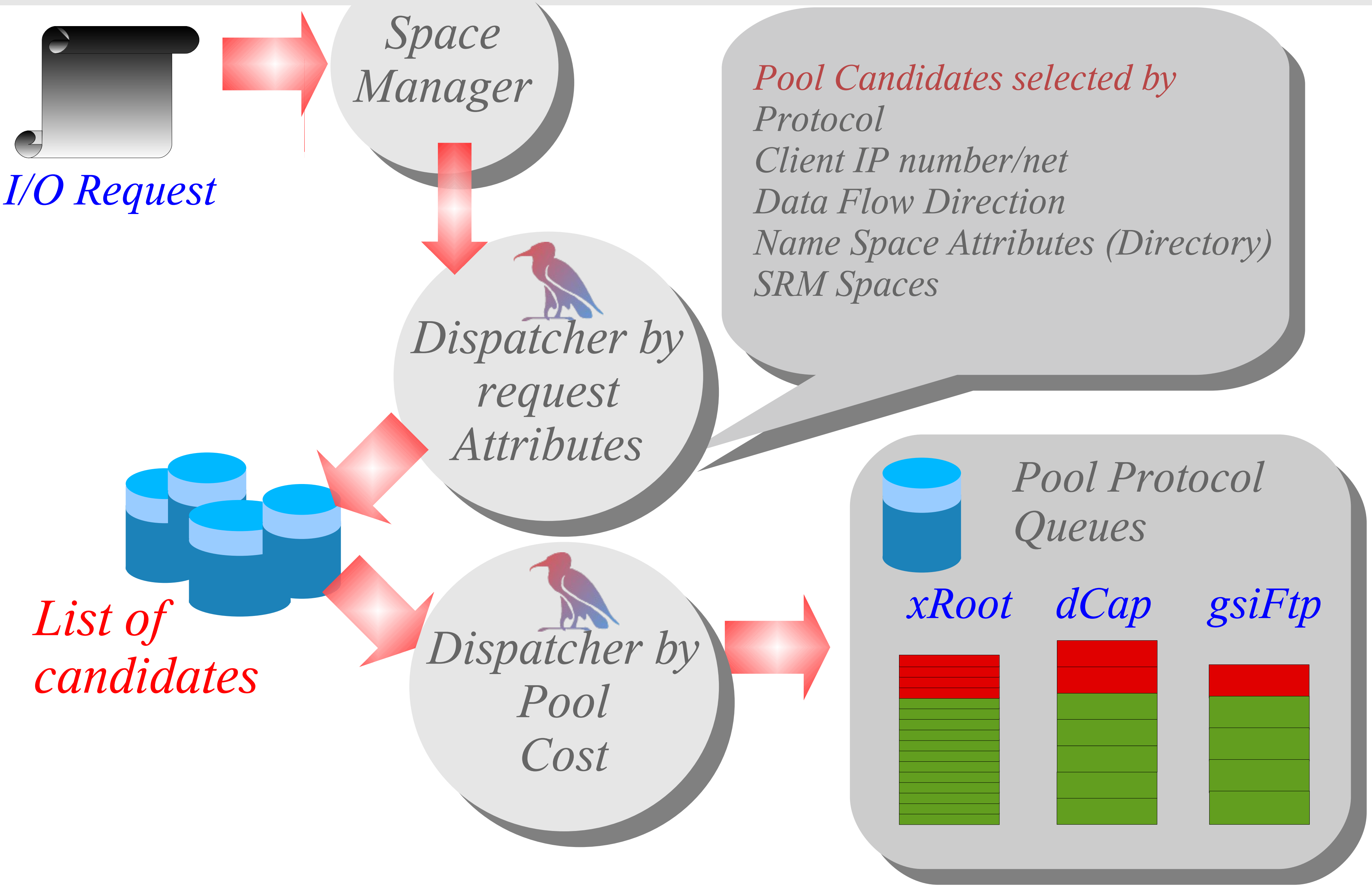
- dCache partitioning for very large installations
- File hopping on
  - automated hot spot detection
  - configuration (read only, write only, stage only pools)
  - on arrival (configurable)
- gPlazma (authentication, authorization, GUMS connectivity)
- Passive dCap
- xRoot support (with *Alice* authorization)
- Central FLUSH manager
- Maintenance module (draining pools)
- improved GUI
- Jpython interface for all kind of configuration (e.g.used by quattor)
- Easy installation (Yaim and VDT)

New Features in 1.8.0

- SRM 2.2 following WLCG agreement
  - Details : see Timurs talk
- xRoot protocol
  - vector read
  - currently working on async I/O
- Chimera (new name space provider)   (optional)
- ACL's ready but not yet in distribution ("Team Test Phase")
- support of multiple, non overlapping HSM systems (NDGF approach)

Technical Introduction

Request View



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

What is Chimera ?
(for details see Tigrans Poster)

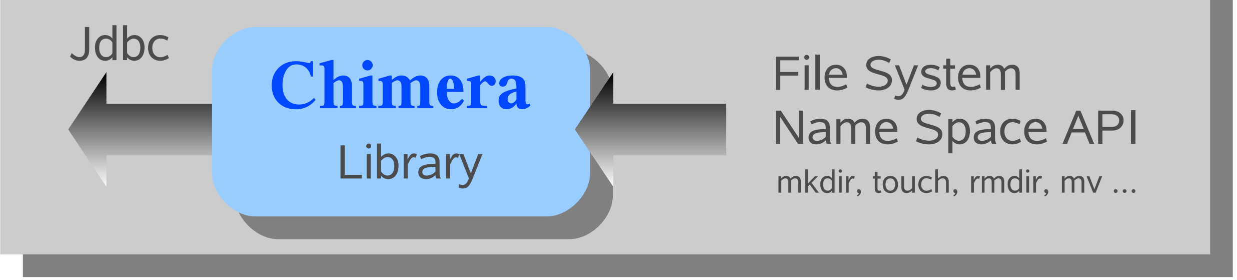
Chimera (See Tigrans Poster)

Why do we need a new Name Space System

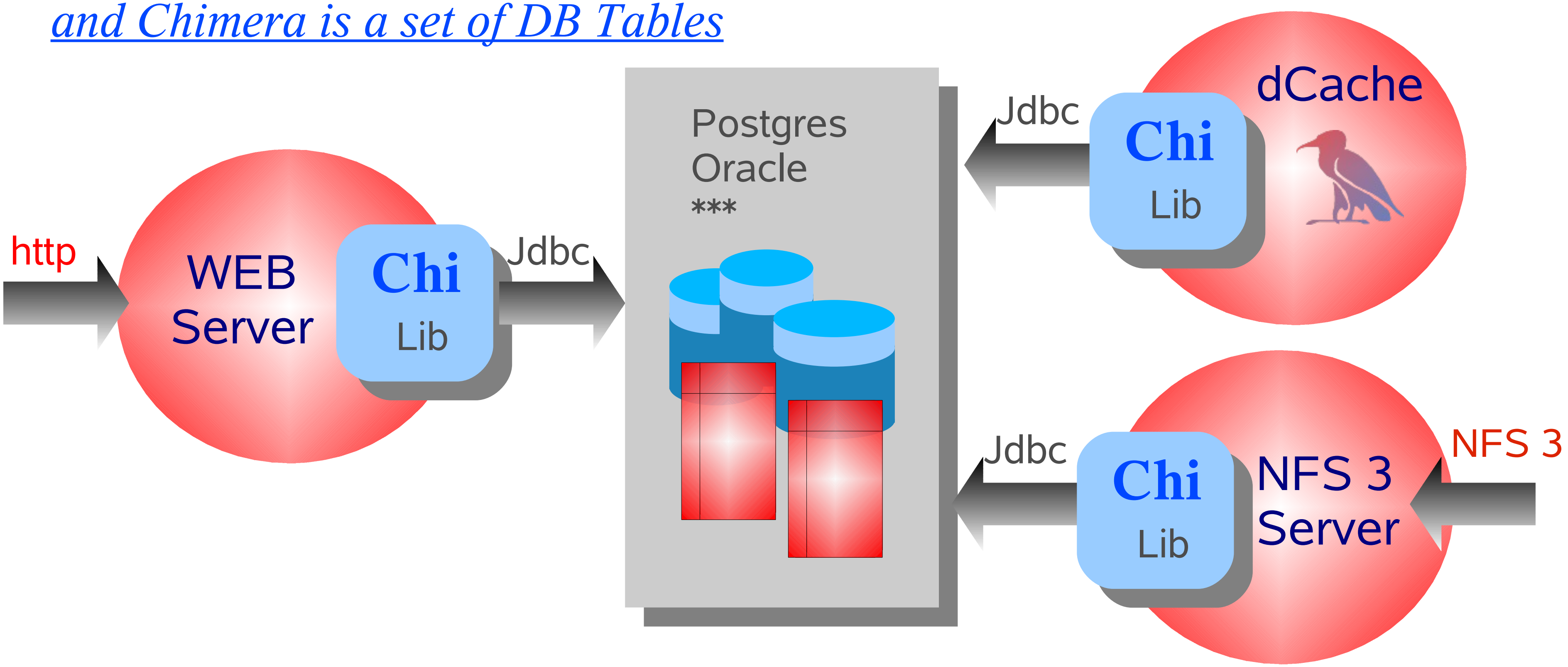
- File sizes may exceed 2G limit.
- dCache does no longer need to mount PNFS (security)
- Acl's can be plugged in. (One ACL implementation already exists)
- Real use of underlying DB features.
- Allows for user defined queries. (Quotas, billing)
- Chimera doesn't add additional lock mechanisms. So it can be as fast a underlying database.

How does chimera work ?

Chimera is a Library



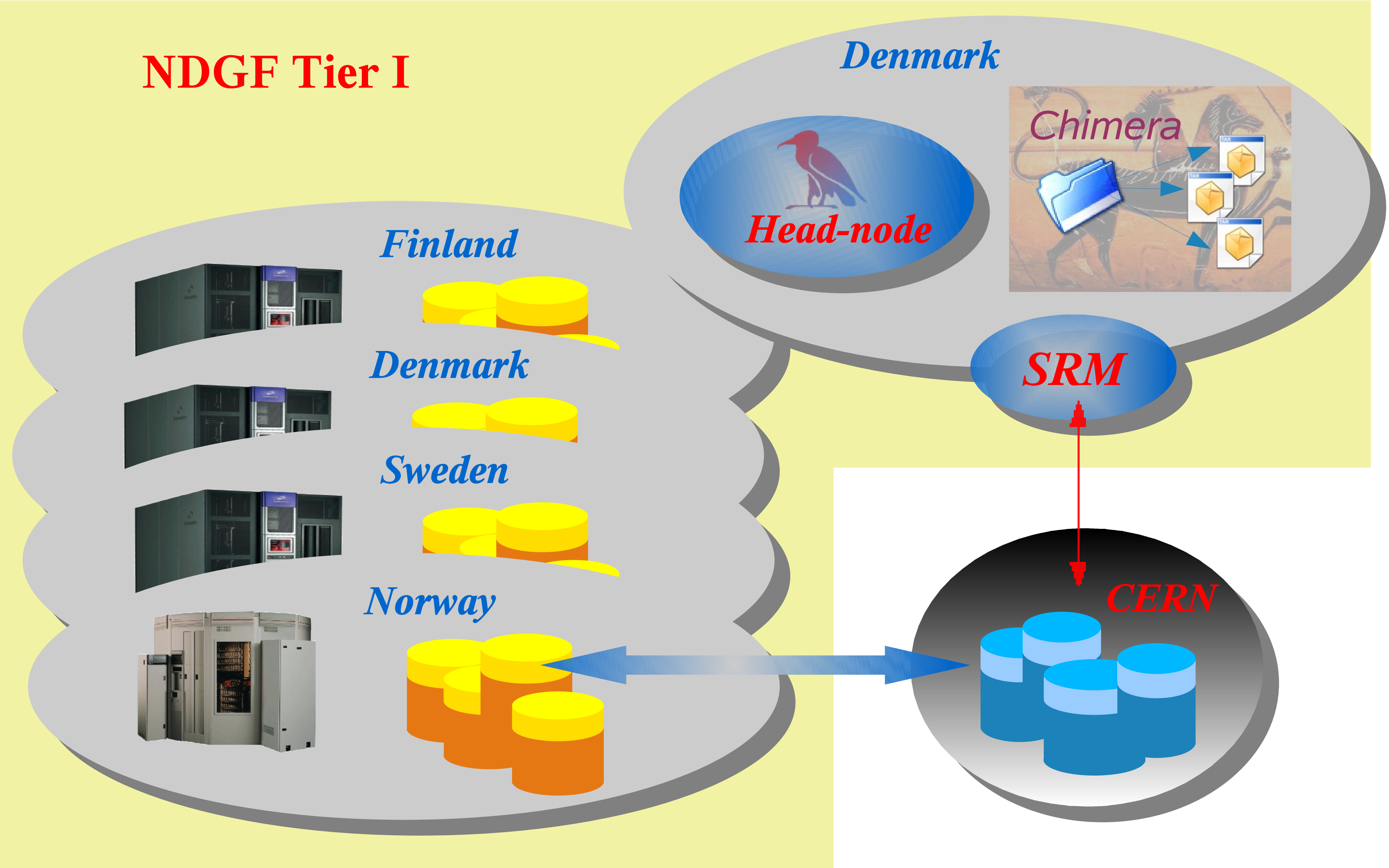
and Chimera is a set of DB Tables



What are we currently working on ?

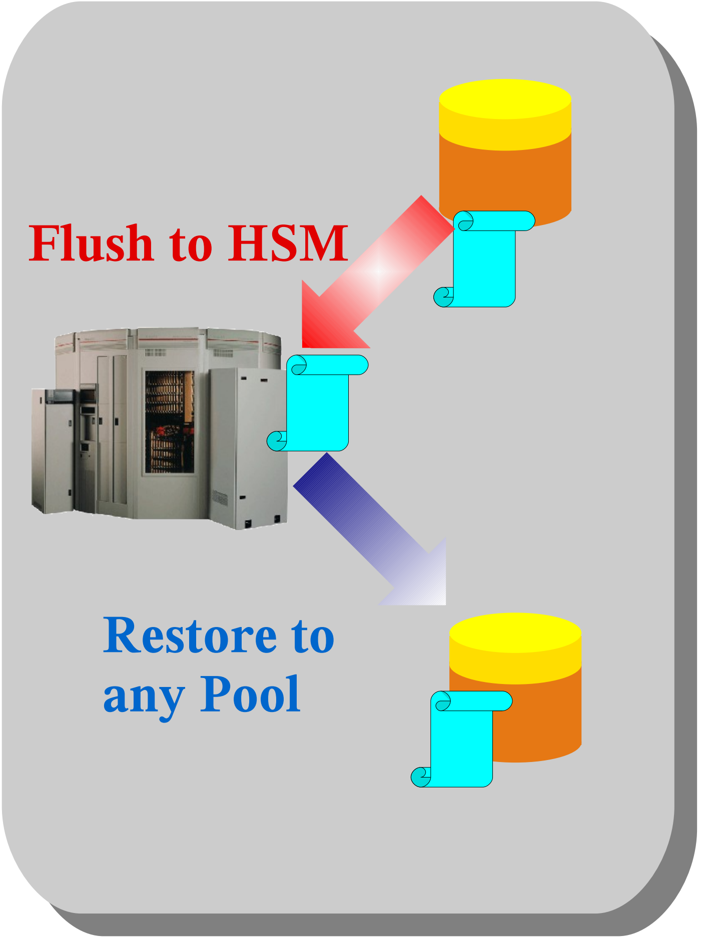
- *The NDGF Challenge (Very short term)*
- *NFS 4.1 (Mid Term)*
- *SRM 2.2 (Now)*

The NDGF Challenge : gsiFtp Protocol Version II

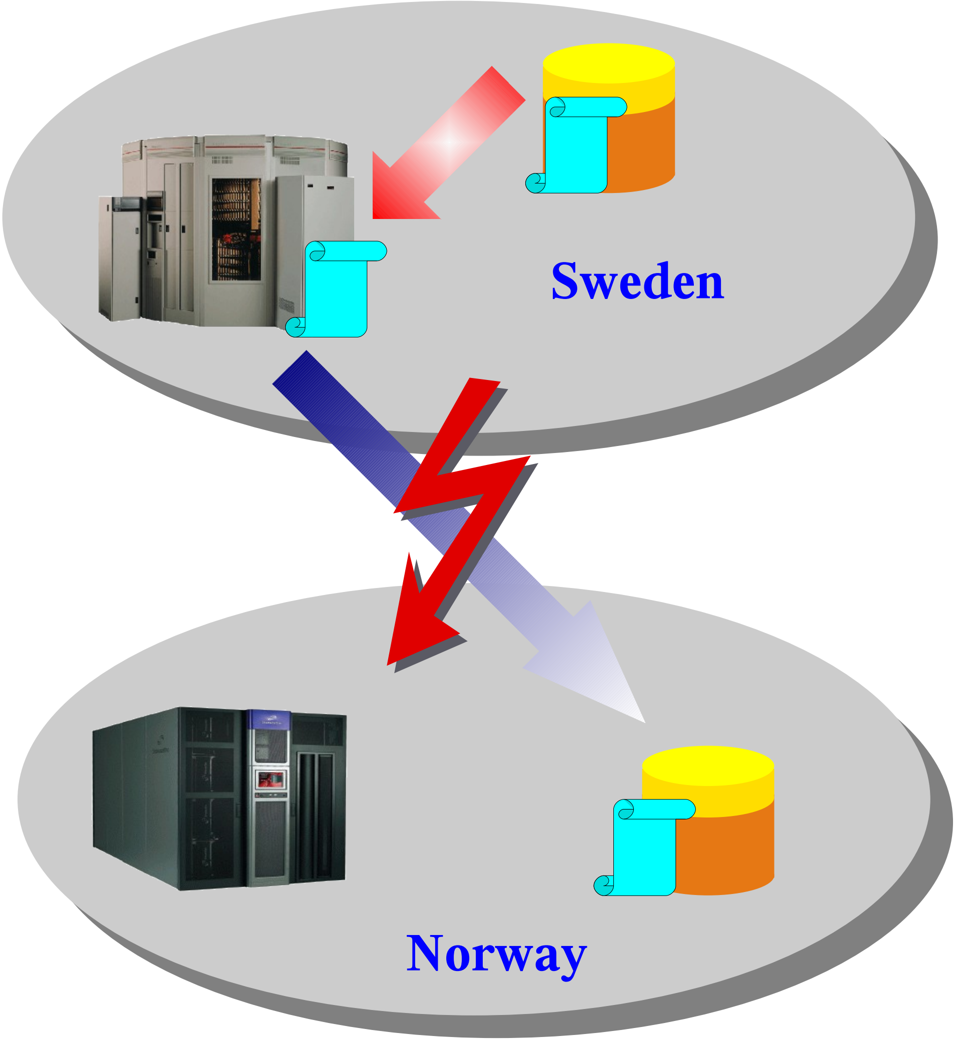


The NDGF Challenge : Multi Site HSM support

Single Site approach



Multi Site approach



Not all pools can access all HSM systems

What's needed for NDGF ?

- PNFS/Chimera doesn't need to be mounted by the pools nodes.
- gsiFtp Protocol Version II to avoid unnecessary data hopping
- While for single site dCaches, all pools are connected to the same HSM instance, for NDGF, files can only be recalled from those pools which are connected to the HSM where the files have been written to. (Sophisticated bookkeeping)
  - Pool are selected based on the secondary location of the data
- Secure internal cell communication
- Fine grained command authorization

What are we currently working on ?

- The NDGF Challenge (Very short term)
- NFS 4.1 (Mid Term)
- SRM 2.2 (Now)

What's next --> NFS 4.1 see Tigrans Poster

We are currently putting significant efforts in the NFS 4.1 protocol

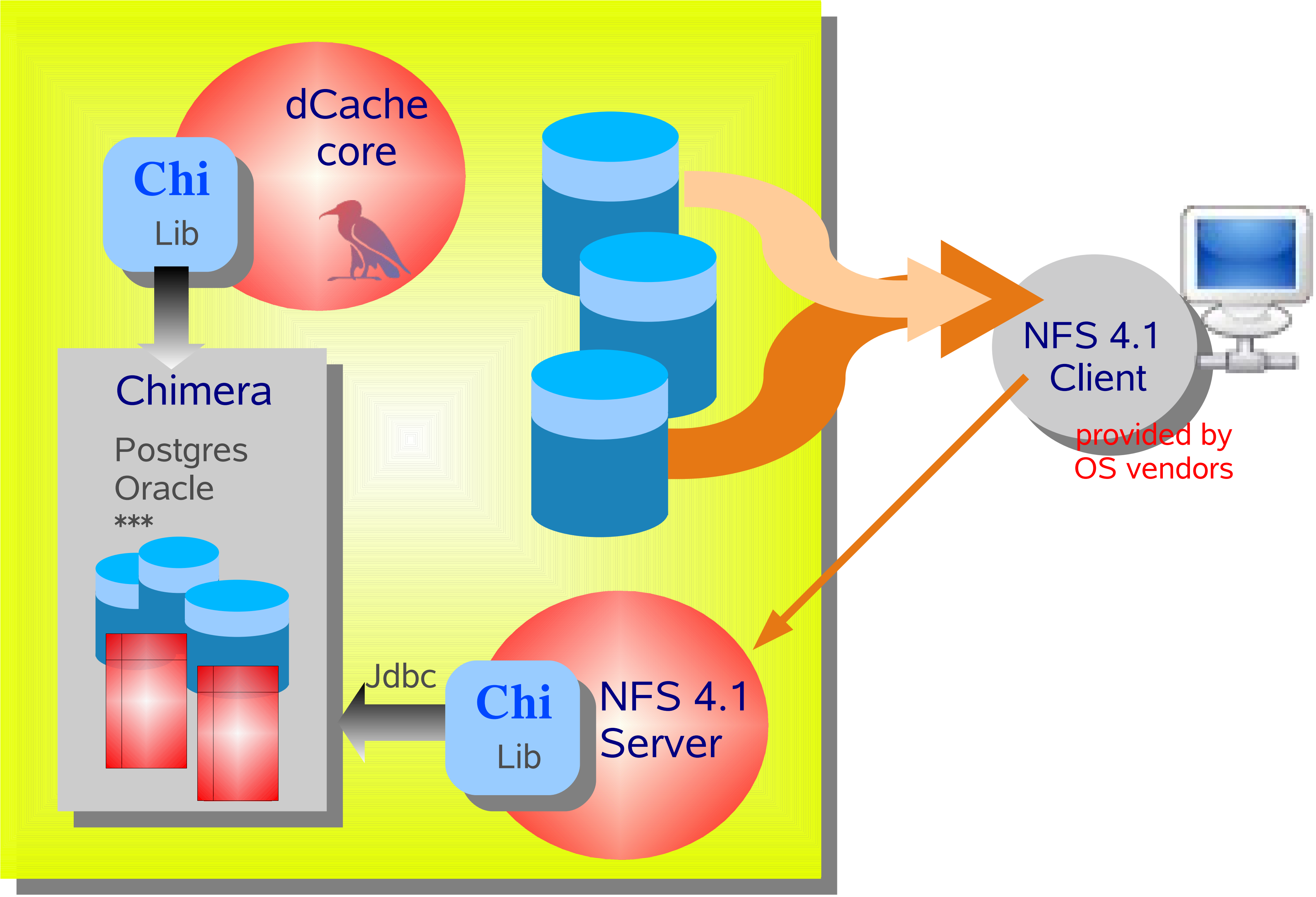
Deployment Advantages :

Clients are coming for free (provided by all major OS vendors).

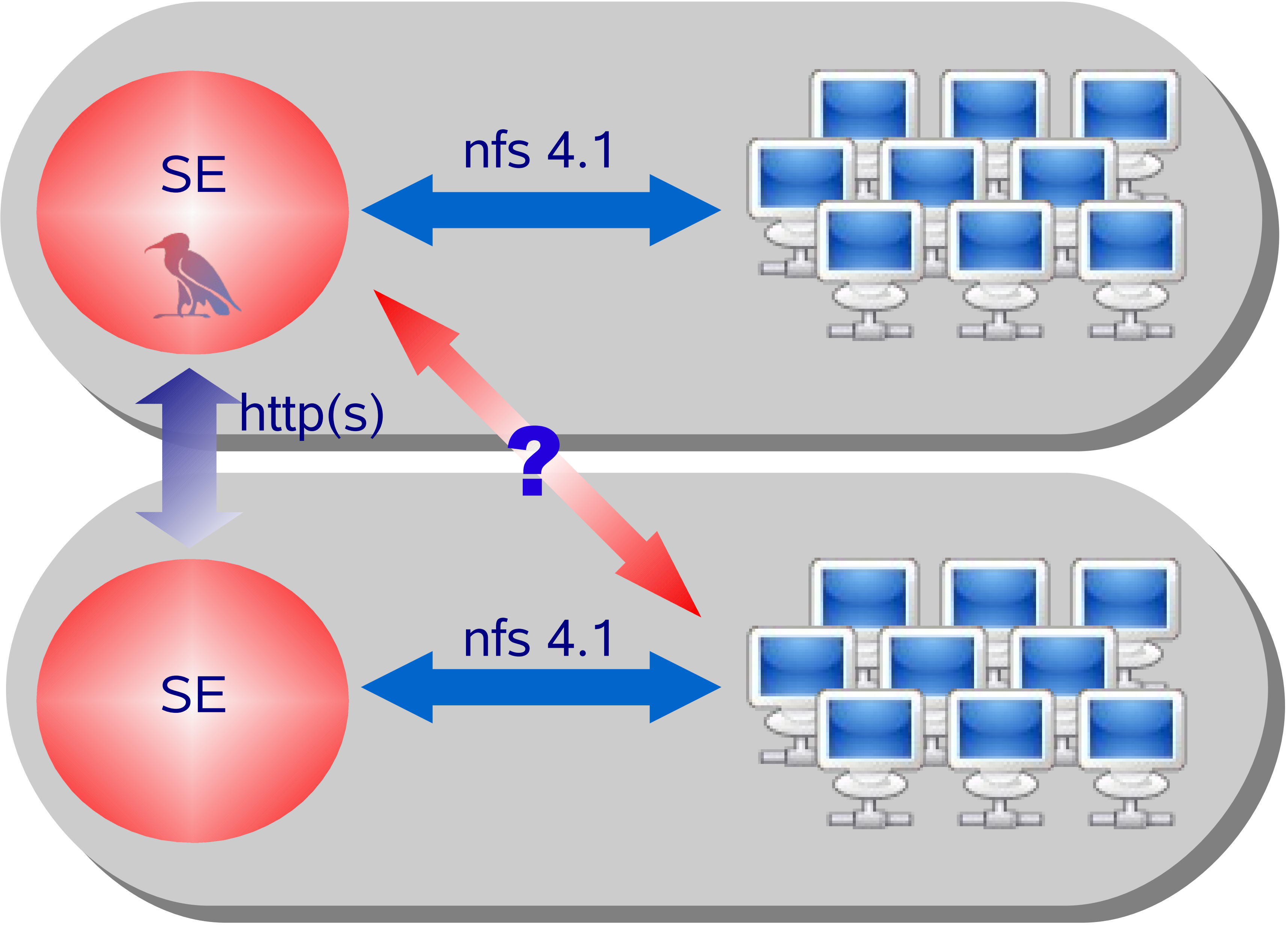
Technical Advantages :

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

NFS 4.1 in dCache



Goal : Industry standards in HEP ?



What are we currently working on ?

- *The NDGF Challenge (Very short term)*
- *NFS 4.1 (Mid Term)*
- *SRM 2.2 (Now)*

SRM 2.2 (Advertisement)

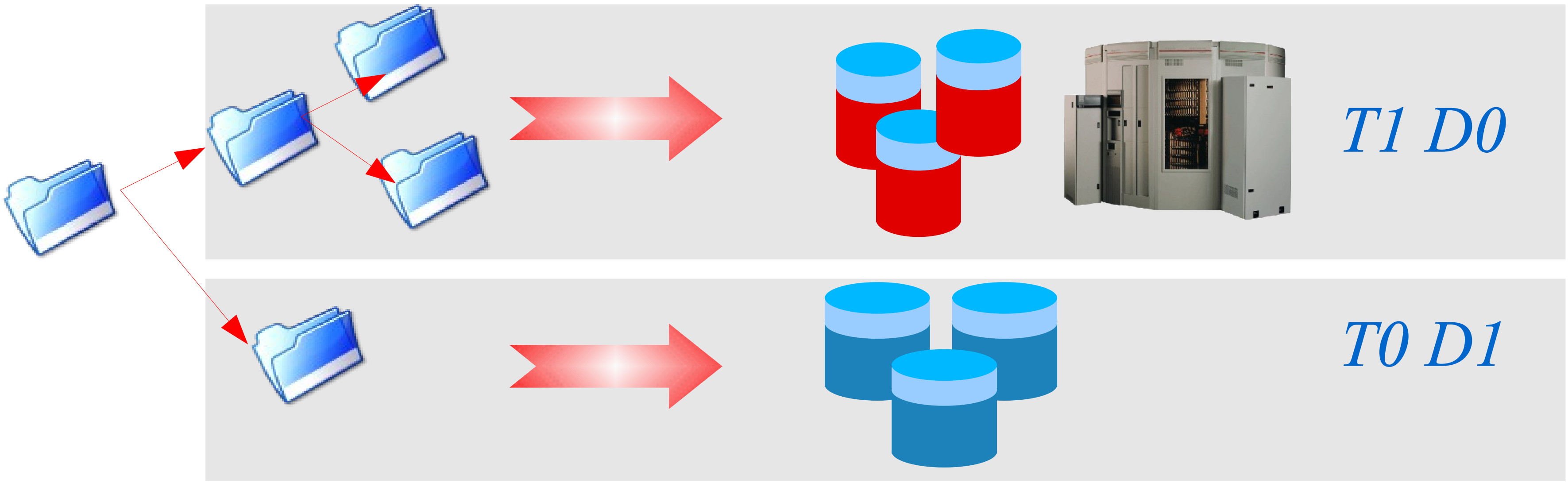
The SRM in dCache supports

- *CUSTODIAL (T1Dx)*
- *NON-CUSTODIAL (T0D1)*
- *Dynamic Space Reservation*
- *late pool binding for spaces*
- *and more*

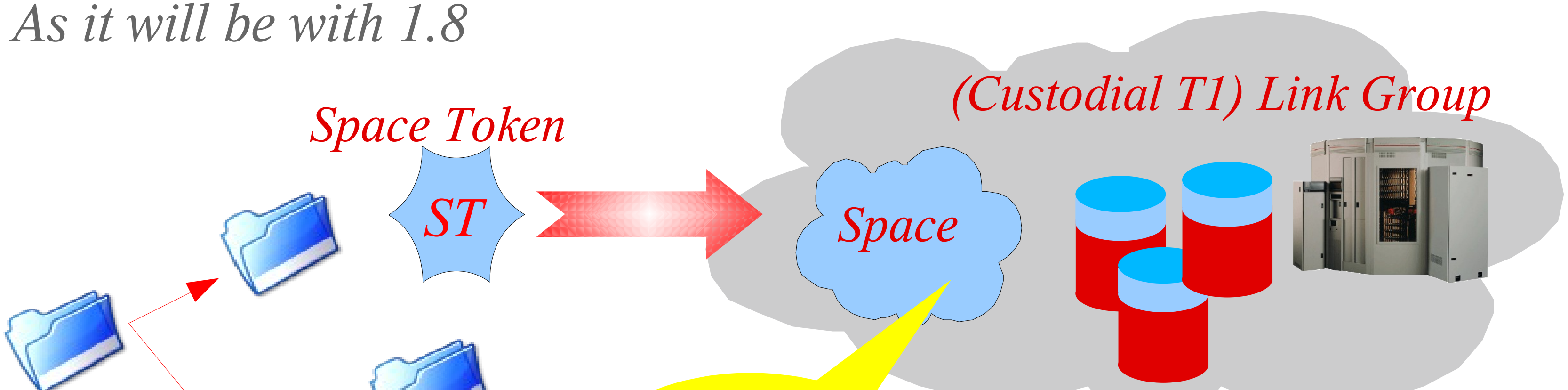
SRM 2.2 (The space token)

Please see Timur's talk for the wonderful world of SRM2.2

As it used to be (<= 1.7)



As it will be with 1.8



Remark : The space of a Space Token is assigned to a pool at the moment the file is opened and not when the space token is created.

dCache 1.8 deployment

FAQ

Deployment Status

Deployment : You should know (FAQ)

dCache 1.8 is a prerequisite for SRM 2.2

dCache 1.8 runs SRM 1.1 and SRM 2.2 at the same time

dCache 1.8 can be installed w/o necessarily using SRM 2.2

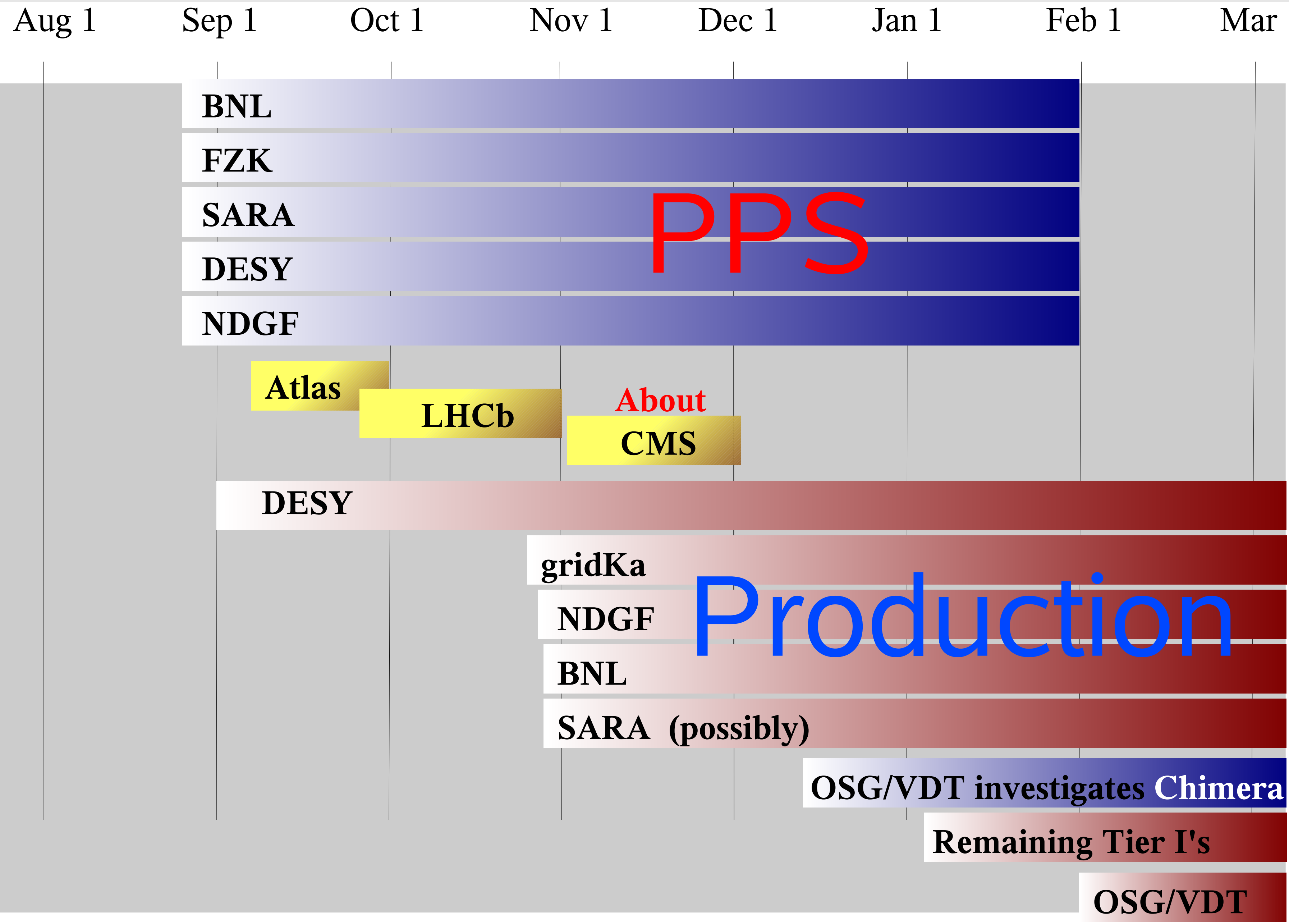
The Chimera and dCache 1.8 deployment runs independently

- 1.7 needs to be operated with PNFS.
- 1.8 can be operated with PNFS and Chimera.
- The default for 1.8 is PNFS, but Chimera is included.

Upgrade Procedure from 1.7 to 1.8

- Smaller sites : Just install and start. May take long.
- Larger sites : Run preparation Job 1-2 days in advance

dCache 1.8 deployment schedule



Further reading

www.dCache.ORG