dCache.ORG

dCache.ORG

Chimera and NFS 4.1 in dCache

Patrick Fuhrmann
Tigran Mkrtchyan

presented by
Peter van der Reest, DESY
at HEPiX, Fall 2007

Content

Motivation

dCache.ORG

- *Pnfs is the current name space and meta data provider for dCache.*
- *With the increasing demands on dCache instances concerning the number of file operations per second, especially at Tier I centers, we expect Pnfs to become a bottleneck with the start of LHC.*
- *Chimera is the replacement of Pnfs, which targets the problems described in this presentation.*
- *The presentation gives some technical details on Pnfs and Chimera.*

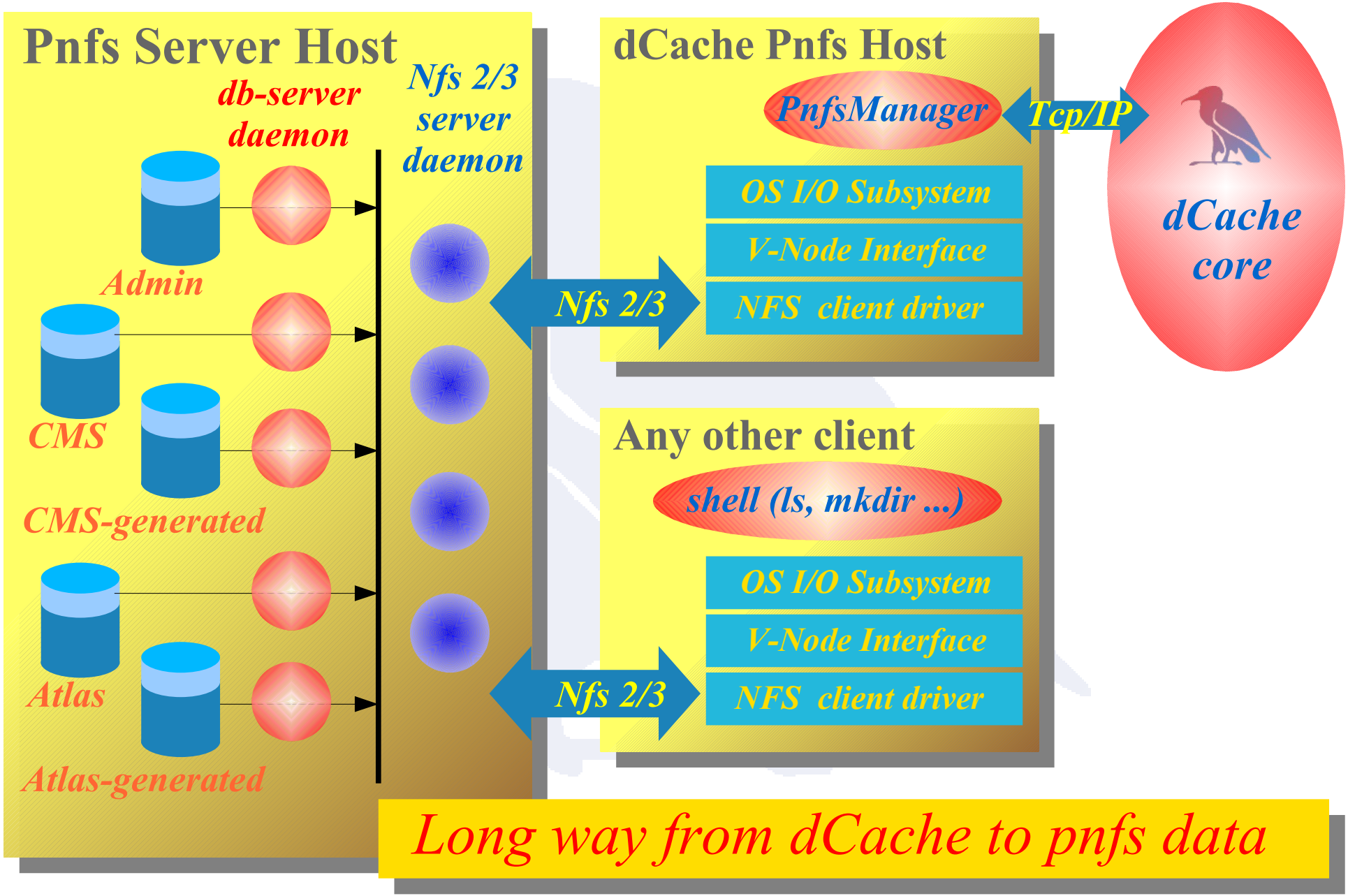
What is Pnfs doing in dCache ?

dCache.ORG

Pnfs is the dCache name space and meta data provider

- Generates a posix like virtual file system name space
- Maps file-system names to unique ID's (pnfsID)
  - dCache (internally) only uses pnfsIDs (never file names)
- Stores posix meta data with file object
  - e.g.: Size, Permissions, Access Timestamps, etc
- Stores arbitrary meta data with the pnfsID's
  - User meta data in /pnfs/.../.(use)(3-7)(<filename>)
  - dCache related data (File Location, HSM information, etc...)
- **Pnfs provides its services through the nfs2/3 interface only**
- **Pnfs doesn't store any 'real' data**
- **The Pnfs software doesn't know anything about dCache**

How does dCache interact with Pnfs ?



What is so wrong with Pnfs ?

Major Pnfs Flaws

- Very long way from dCache to the pnfs name service
  - PnfsManager
  - Local I/O subsystem
  - nfs 2 (client driver)
  - nfs 2 server (pnfs) daemon
  - db server of pnfs sub partition
  - postgres database
- Only one read/write lock per database (blocks whole database e.g. CMS)
- Pnfs can not distinguish between dCache and other clients
- Pnfs can only run on a single host
- ('ls' on) Pnfs extremely slow if > 2000 files per directory.
- Nfs 2 : Maps file system operations to too many nfs ops
  - Some dCache operations are mapped to > 200 nfs ops
- Nfs 2 : File size limit < 2 Gbytes

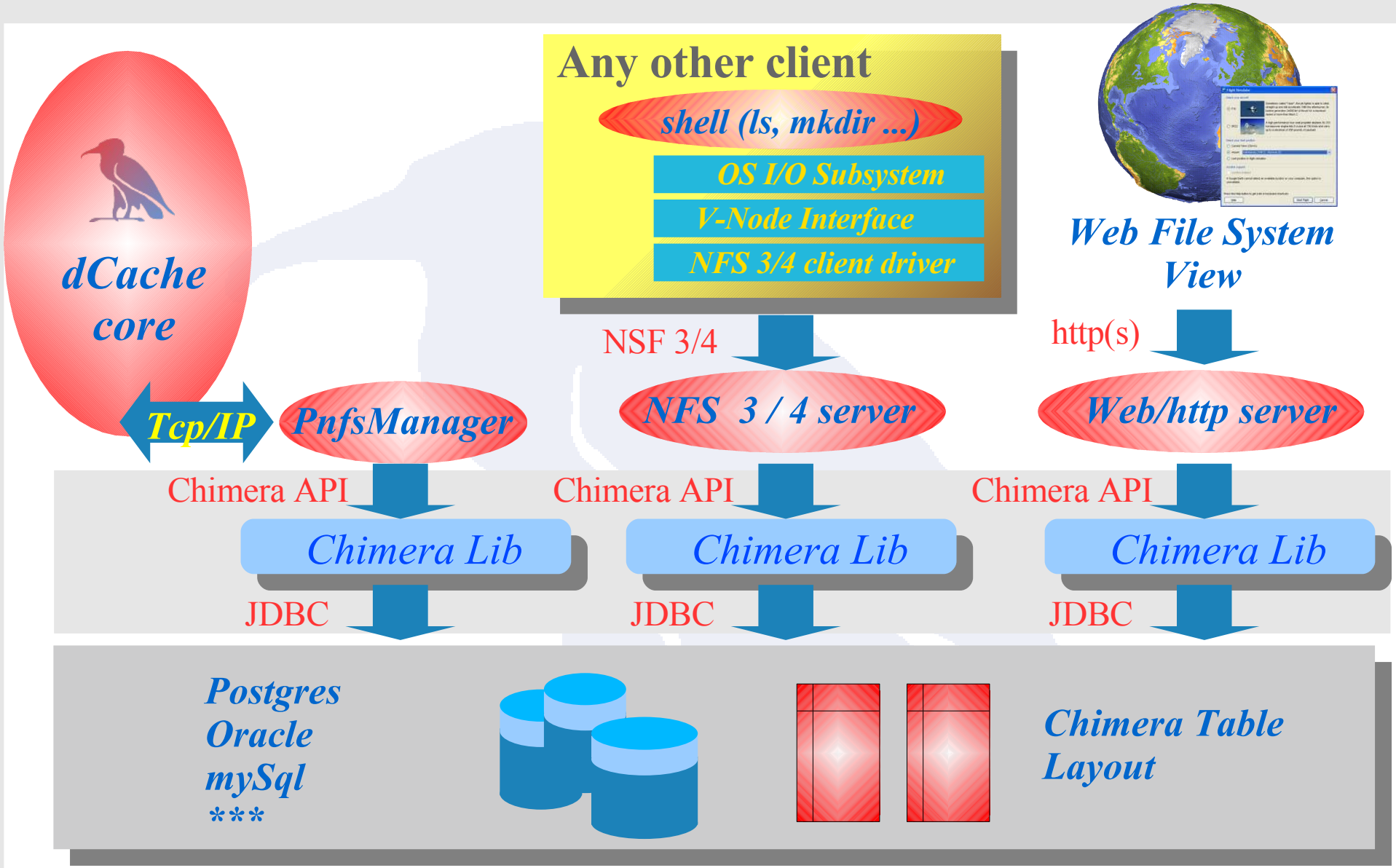
dCache.ORG

What is Chimera (Basics) ?

- Chimera provides the same functionality to dCache as Pnfs does.
- Only the pnfs manager driver within the PnfsManager has to be adjusted.
- Chimera is a Java API, a library and a database table layout.
- Chimera doesn't have any server by itself.
- Consequently it scales with performance of database backend.

dCache.ORG

How does dCache interact with Chimera ?



What does Chimera provide ?

dCache.ORG

- *The dCache PnfsManager talks directly to Chimera. (no intermediate layers).*
- *Chimera can distinguish between dCache core and the various client interfaces.*
- *Chimera allows ACLs to be plugged in (Posix implementation already av.)*
- *Differentiation between read and write locks through DB backend.*
- *Performance independent of number of files per directory.*
- *Arbitrary number of levels for user meta data – space reserved at creation time.*

What does Chimera provide ? (cont'd)

- Chimera takes advantage of the performance of the chosen database back-end.
  - *If the database back-end can span various hosts, Chimera can do as well.*
  - *Partitioning of large tables can help in later stage performance tuning.*
  - *No central database table locks.*
- *Chimera allows at least 10 times more dCache file access operations per second than pnfs (using postgres and similar hardware)*

dCache.ORG

Status of Chimera ?

dCache.ORG

- *Chimera is available at dCache.org with sufficient information on how to setup a dCache 1.8 with Chimera.*
- *Edinburgh(gridPP) and Bari(INFN) are testing Chimera.*
- *OSG, VDT will start investigation mid of December.*
- *dCache development is using Chimera intensively as namespace provider.*
- *Pnfs to Chimera migration mechanisms available.*

What is NFS 4.1

dCache.ORG

- *NFS 4.1 is an NFS 4 extension which is aware of the fact that the back end storage system may have the same file stored on a set of different servers (pNFS, not to be confused with Pnfs).*
- *The specification of NFS 4.1 is in its final phase.*
- *Organizations like CITI, SUN, IBM, EMC, PANASSAS, NETAPP, Linux and dCache.org are active in the specification process.*
- *Regular meetings with all the related developers including dCache.org.*

Major advantages of NFS 4.1

dCache.ORG

Technical Advantages :

- NFS 4.1 is aware of distributed data (as in dCache)
- 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 on file level
- OPEN / CLOSE semantic (so server can keep track of open files)
- 'DEAD' client discovery (client side file lock renew within lease time)

Deployment Advantages :

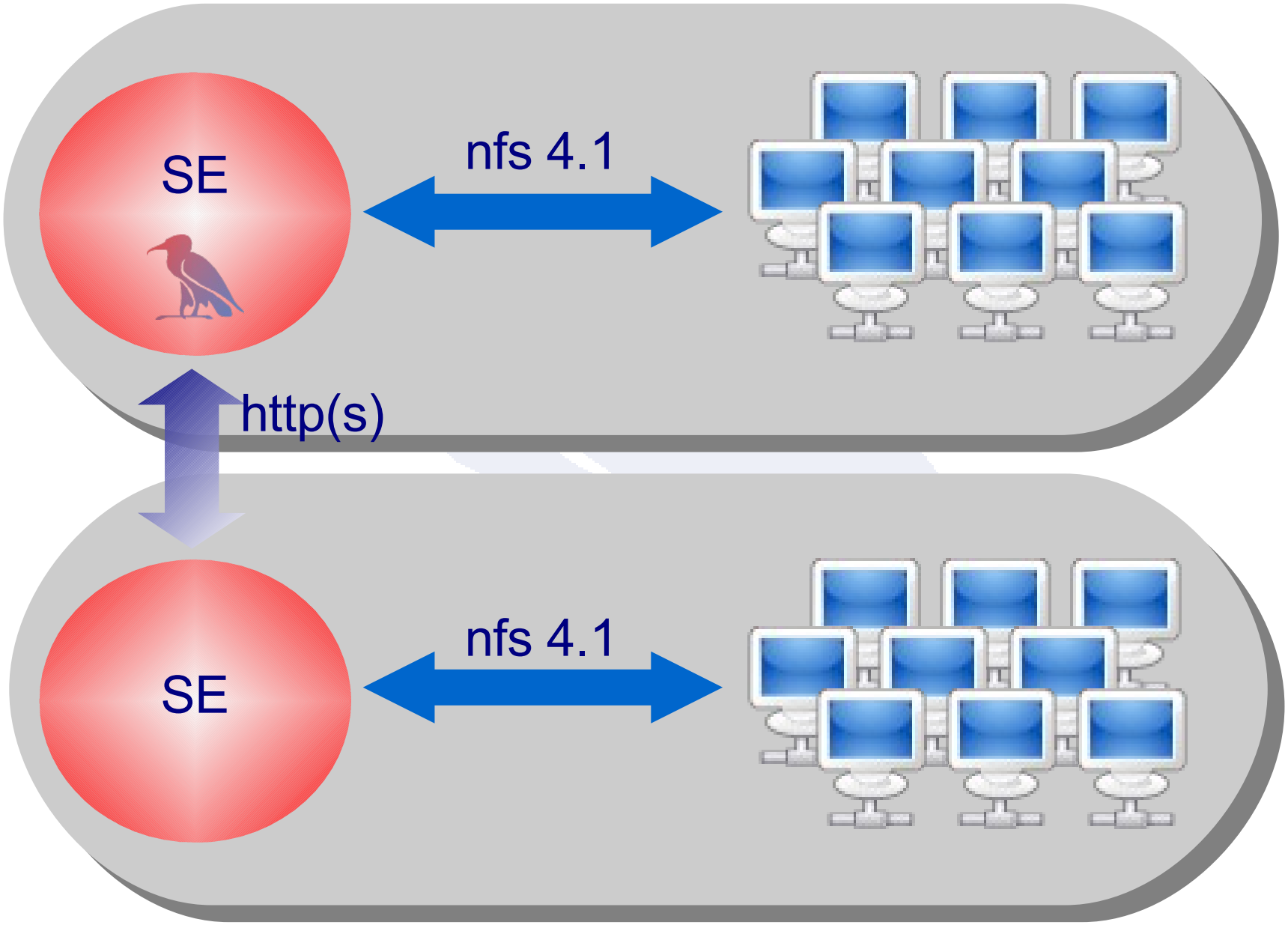
Clients are coming for free (provided by all major OS vendors).

What does NFS 4.1 mean for dCache

dCache.ORG

- We would be able to distribute our data by means of a standard protocol in a posix like manner, without having to offer the client software, which would be provided by the OS providers.
- While NFS 2/3 in dCache only exposes the name space, NFS 4.1 would make the data repository available as well.
- in this perspective, Chimera with a NFSv4.1 door takes the role of MDS, while pools become Storage Devices.
- dccp and srmcp will remain, of course

Industry standards in HEP ?



dCache.ORG

dCache.ORG

Further reading

www.dCache.ORG