PIC
port d'informació
científica

dCache at PIC

Storage Interoperability beyond WLCG

Many slides stolen from Patrick Fuhrman (dCache.org)

patrick.furhman@desy.de

Gerard.Bernabeu@pic.es
Francisco.Martinez@pic.es

PIC
port d'informació
científica

Index

- What is dCache?
  - Enstore
  - NFS 4.1
- Use case examples
  - LOFAR
  - WLCG
- dCache Future Schedule

PIC
port d'informació
científica

What is dCache?

- System for storing and retrieving huge amounts of data
- Distributed among a large number of heterogenous server nodes
- Under a single virtual filesystem tree
- With a variety of standard access methods

PIC
port d'informació
científica

dCache @ PIC in numbers

- 1 FileSystem: /pnfs/pic.es/data ...
  - Few million files
    - 13 million at PIC on may/2011
  - Many PetaBytes
    - Over 7 PB of data behind dCache
      - 4PB on disk
      - 3PB on tape using Enstore

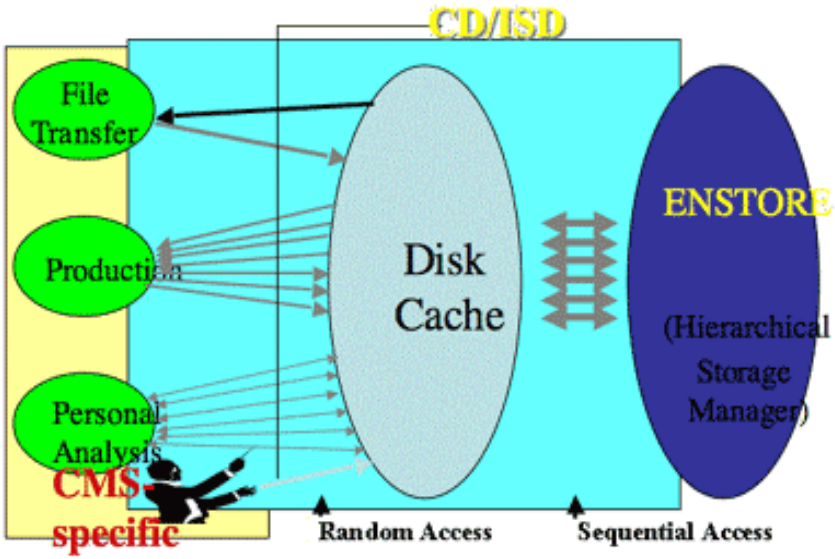
PIC
port d'informació
científica

Enstore

Behind dCache at PIC there is an Enstore Tape backend. Users can use data on Tape as if it was on disk.

Enstore provides distributed access to and management of data stored on tape.

Enstore is developed and maintained by Fermilab.



Fermilab

PIC
port d'informació
científica

How dCache is built - Layers

NON Standard

dCap & xRoot

Standard File Access Protocols

http(s)
WebDav

NFS 4.1

gsiFtp

CDMI (SNIA)
Cloud Data
Management
Interface

Storage
Management

SRM

Common Security Layer

Authentication : Kerberos, X509, Password

Unified ID management

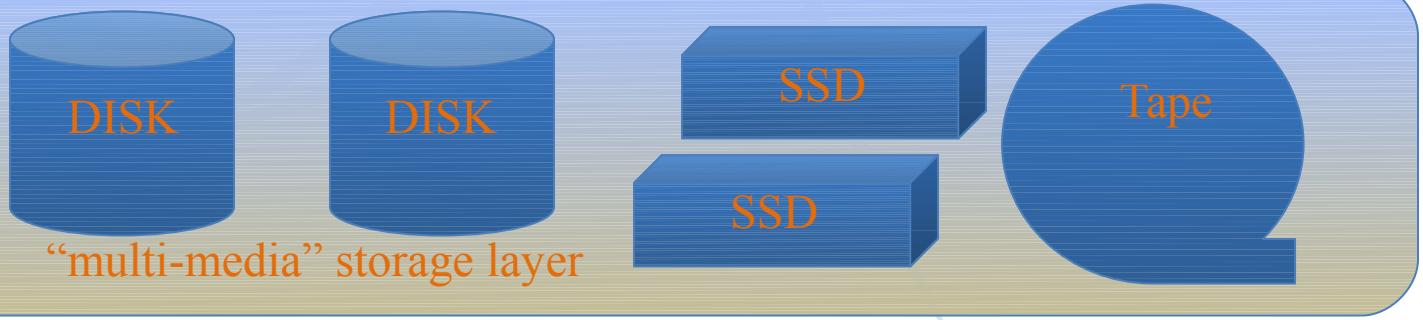
Authorization : ACL's for File system and storage control (SRM)

Common Name Service Layer

Extended Names Service Queries (SQL)

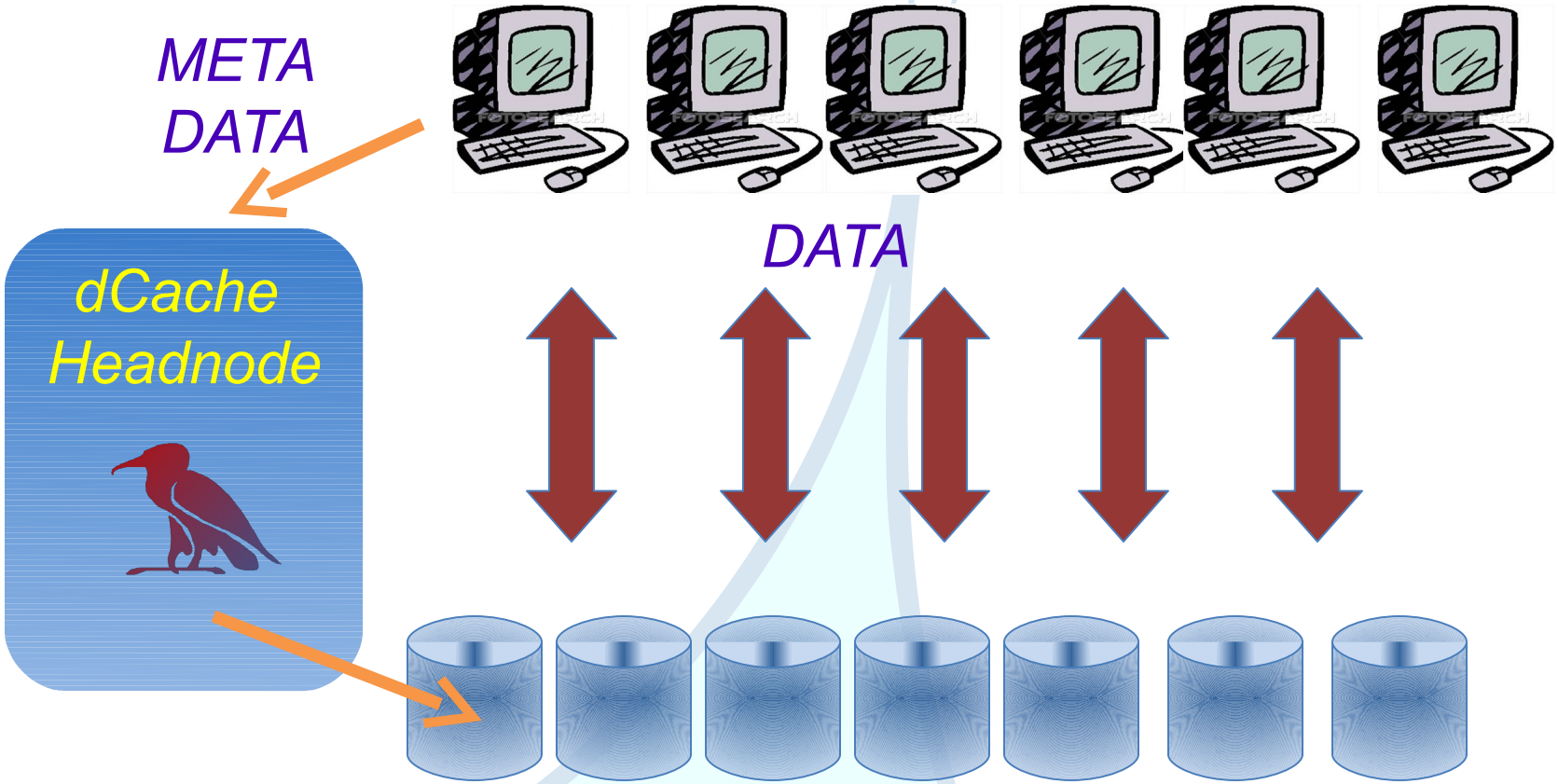
Data Movement Layer

Tape <-> Disk ; Disk <-> Disk ; Pool Draining,
File replication on hot-spot detection, Resilient dataset manager,
Replication on arrival



PIC
port d'informació
científica

How dCache is built – Data flow



PIC
port d'informació
científica

What is the NFS 4.1 initiative?

Industry initiative between all the major storage and OS vendors.

Coordinated by CITI at the University of Michigan

It is an WLCG demonstrator.

Funded effort within the European Middleware Initiative

Major effort in dCache

For non LCG communities

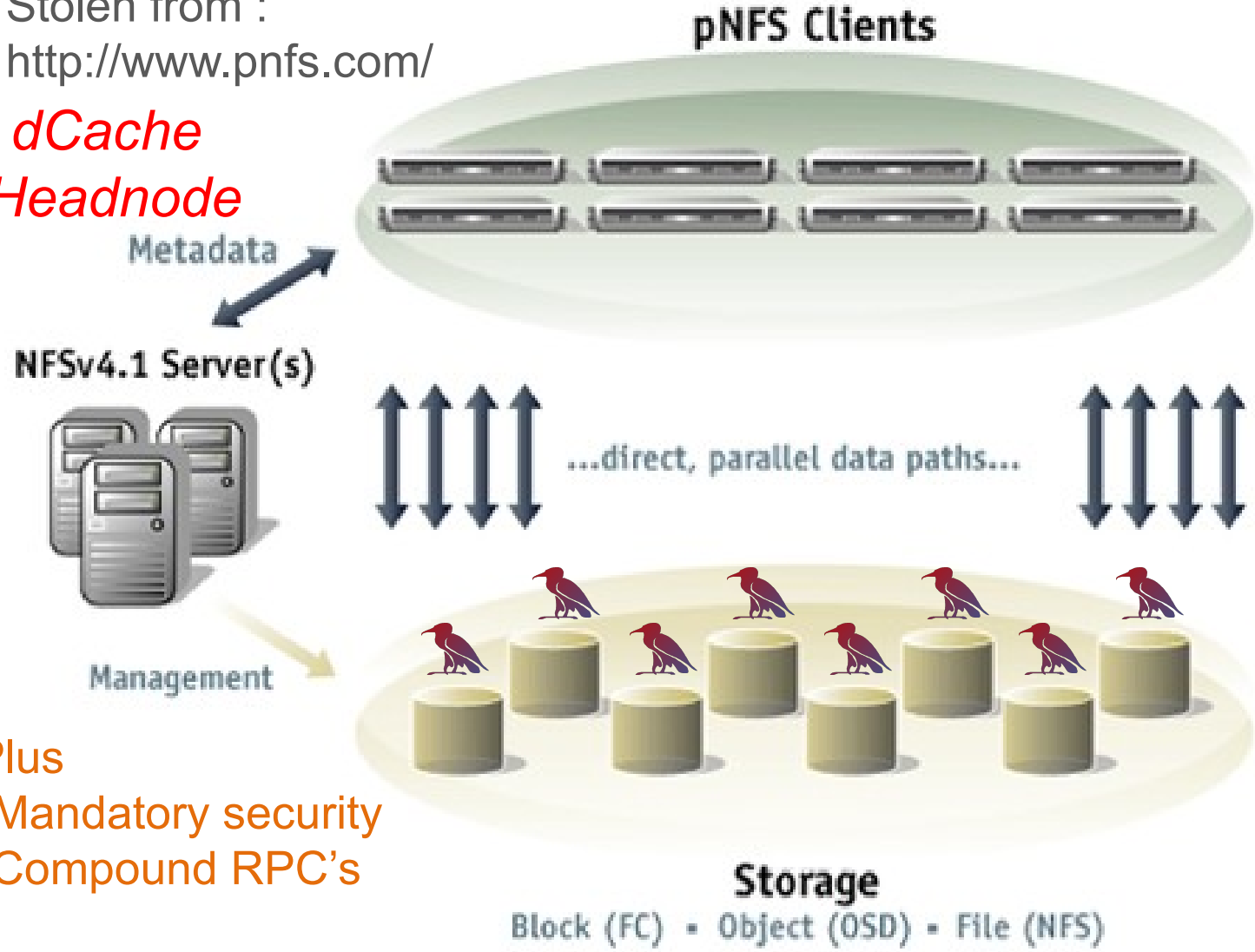
Hopefully for HEP as well

PIC
port d'informació
científica

How does NFS 4.1 work?

Stolen from :
http://www.pnfs.com/

dCache
Headnode



Plus
- ✓Mandatory security
- ✓Compound RPC's

PIC
port d'informació
científica

Why is industry interested?

Stolen from : http://www.pnfs.com/

Benefits of Parallel I/O

- Delivers Very High Application Performance
- Allows for Massive Scalability without diminished performance

Benefits of NFS (or most any standard)

- Ensures Interoperability among vendor solutions
- Allows Choice of best-of-breed products
- Eliminates Risks of deploying proprietary technology

PIC
port d'informació
científica

Why is HEP interested?

- Don’t have to care about client software anymore.
  - No specific ROOT drivers (dCap,rfio,xroot). Just ‘open /foo/blah’
- Less software components to maintain.
- Can be used by unmodified applications (e.g. Mathematica®)
  - regular mount-point as any other FS e.g. /afs, /pnfs
- File/Block caching algorithms provided by professional computer scientists within the OS kernel.

PIC
port d'informació
científica

The question is

Will the WLCG/EGEE storage middleware stack, as provided to EGI through the European Middleware Initiative (EMI), be able to satisfy the needs of new data intensive communities ?

PIC
port d'informació
científica

Use Case Examples

Would like to use the SARA storage facility, which is currently serving as WLCG Tier.

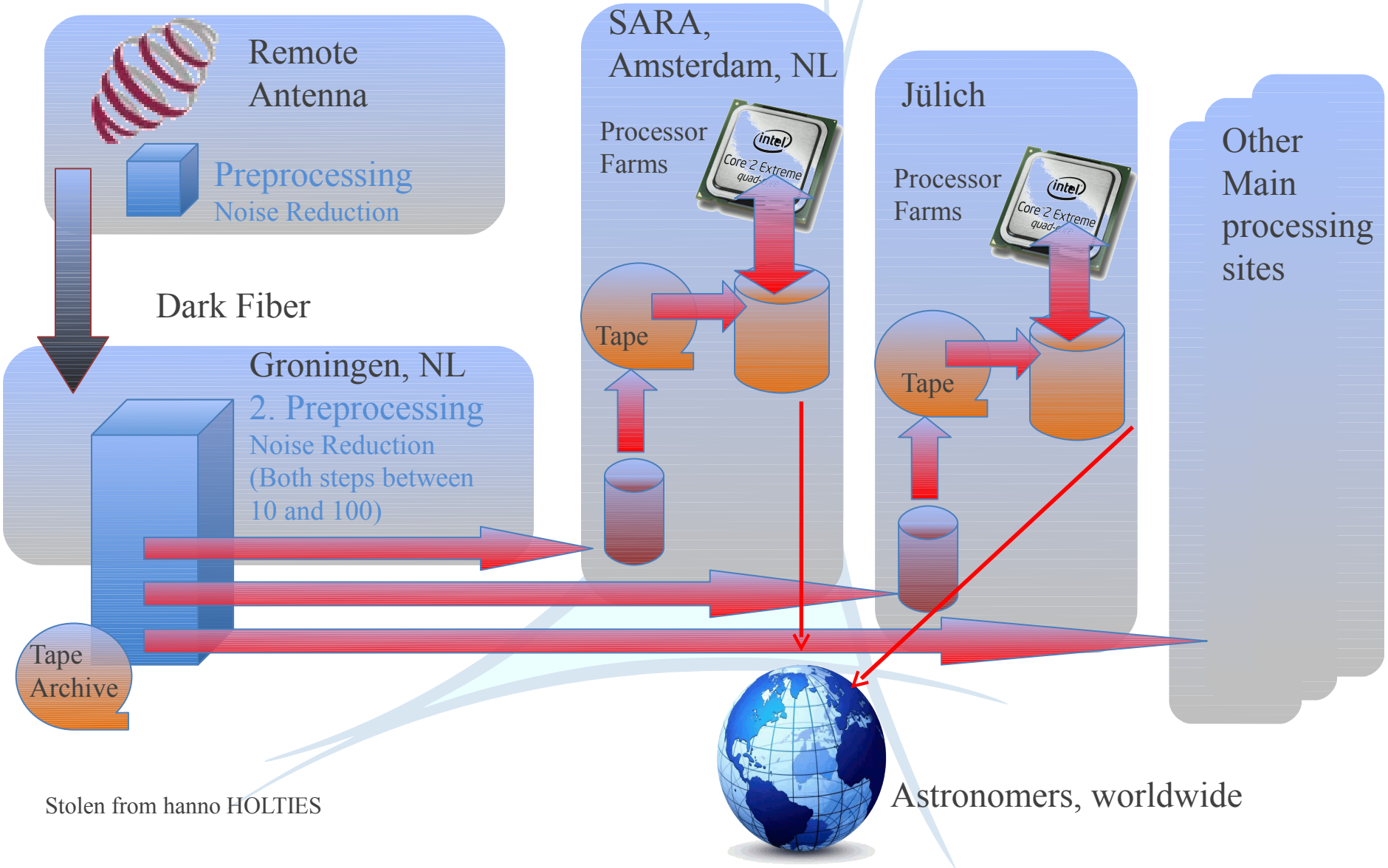
WLCG ATLAS, CMS and LHCb use PIC as Tier-1

At PIC dCache is used in production for many other projects: MAGIC, T2K, CTA, PAUS, TURBO

And some other projects will join soon...

PIC
port d'informació
científica

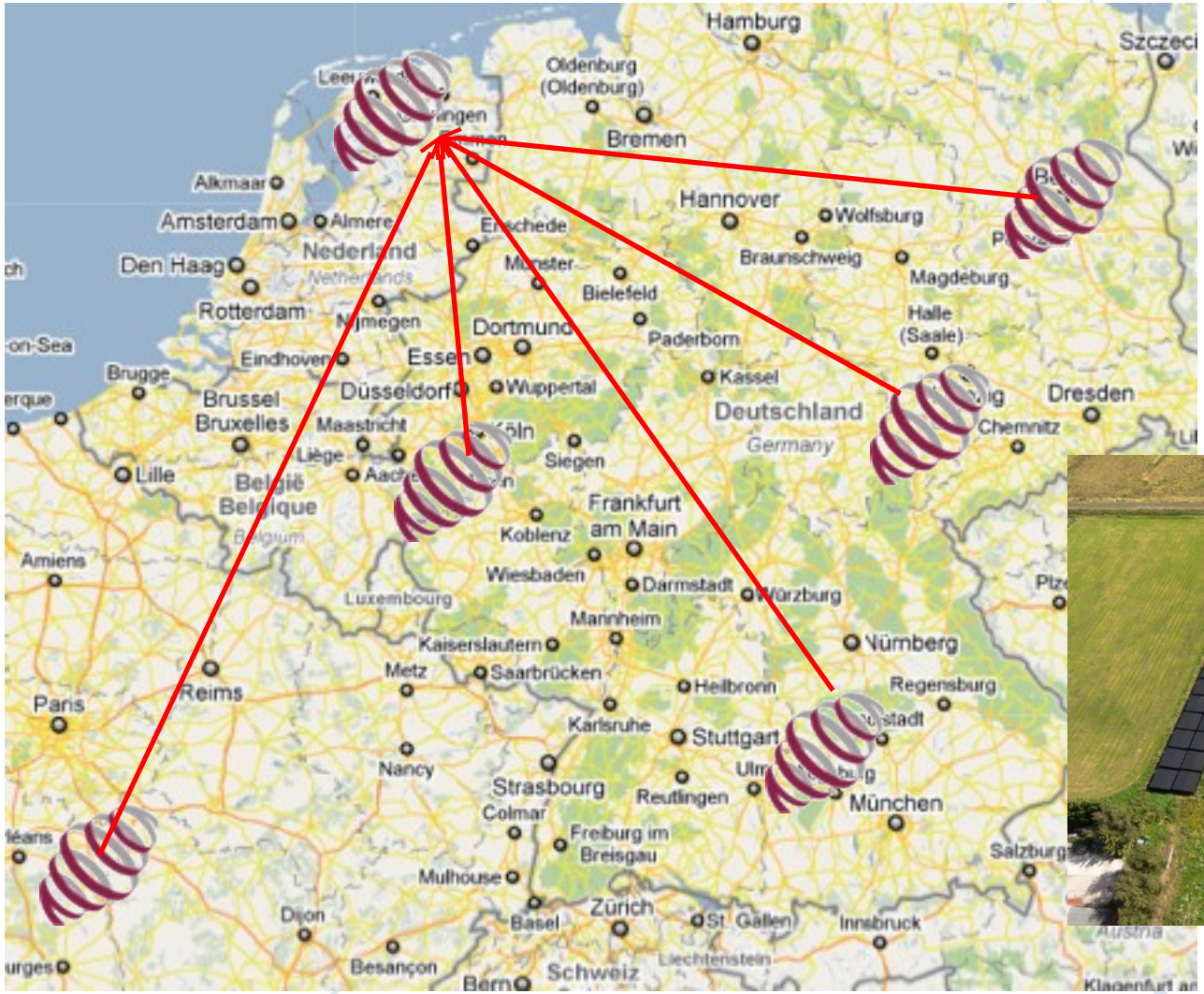
LOFAR (Simplified) Data Flow Model



Stolen from hanno HOLTIES

PIC
port d'informació
científica

The International LOFAR Radio Telescope



As of Feb 24, 2010 :

- 21 Complete Stations
- 10 In Progress
- 13 Planned
- NL, DE, UK, FR, SE

Stolen from hanno HOLTIES

PIC
port d'informació
científica

LOFAR Requirements

Low threshold data retrieval

- Access only by registered LOFAR members.
- CERTS are not desirable for all members.
- Owner of data needs to disable directory browsing.
- Common protocols : Mounted file system, http/WebDav

Roles

- OPERATIONS can put data into permanent storage.
- USER may retrieve data from permanent storage.
- Quotas on ‘tape backend usage’.
- Groups storage areas for read/write

Integration with external (non-EGEE) identity management system.

Accounting

- Per VO, user, directory
- Quotas

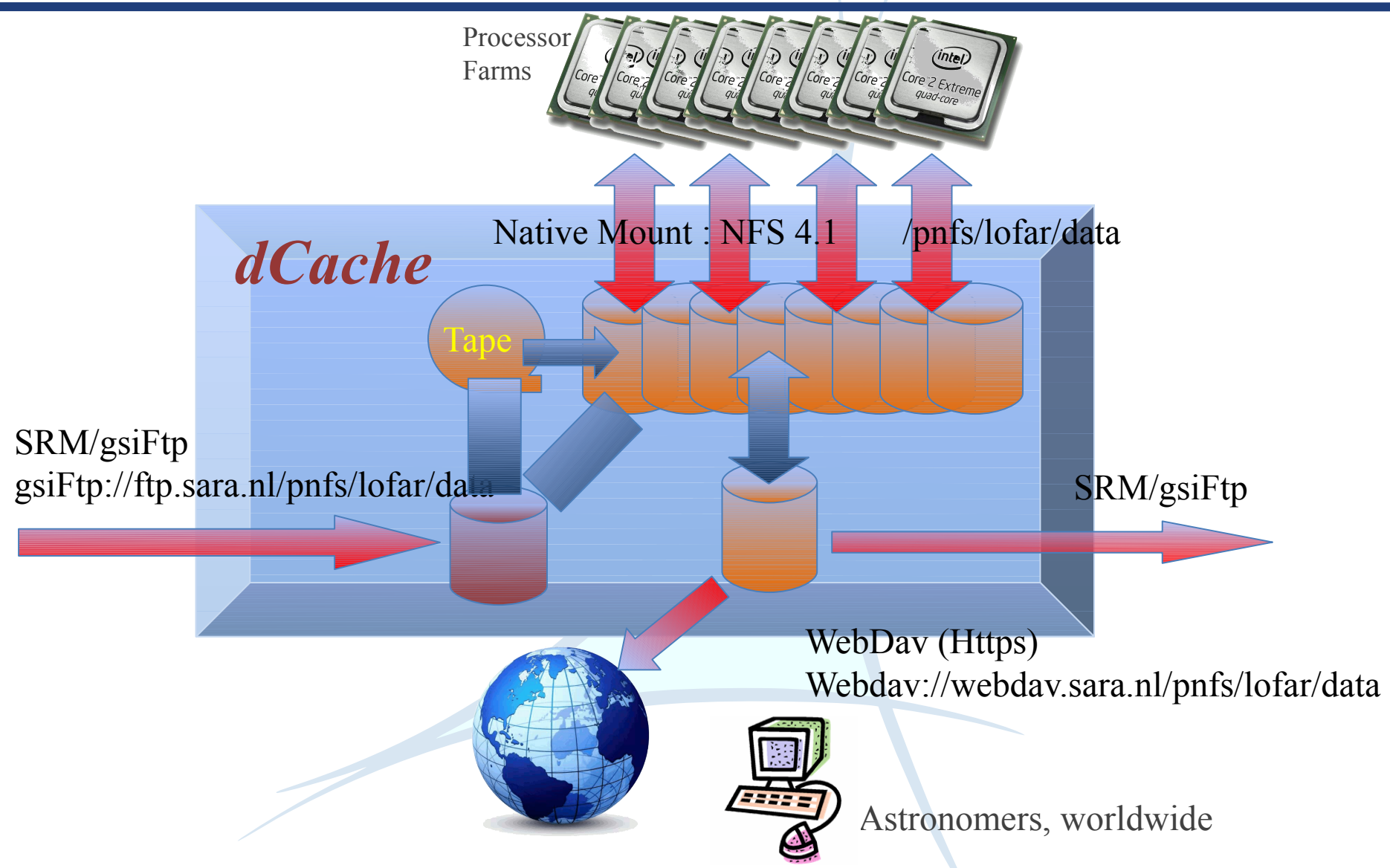
Data integrity

Fixed URLs (to support external catalogues)

Stolen from hanno HOLTIES

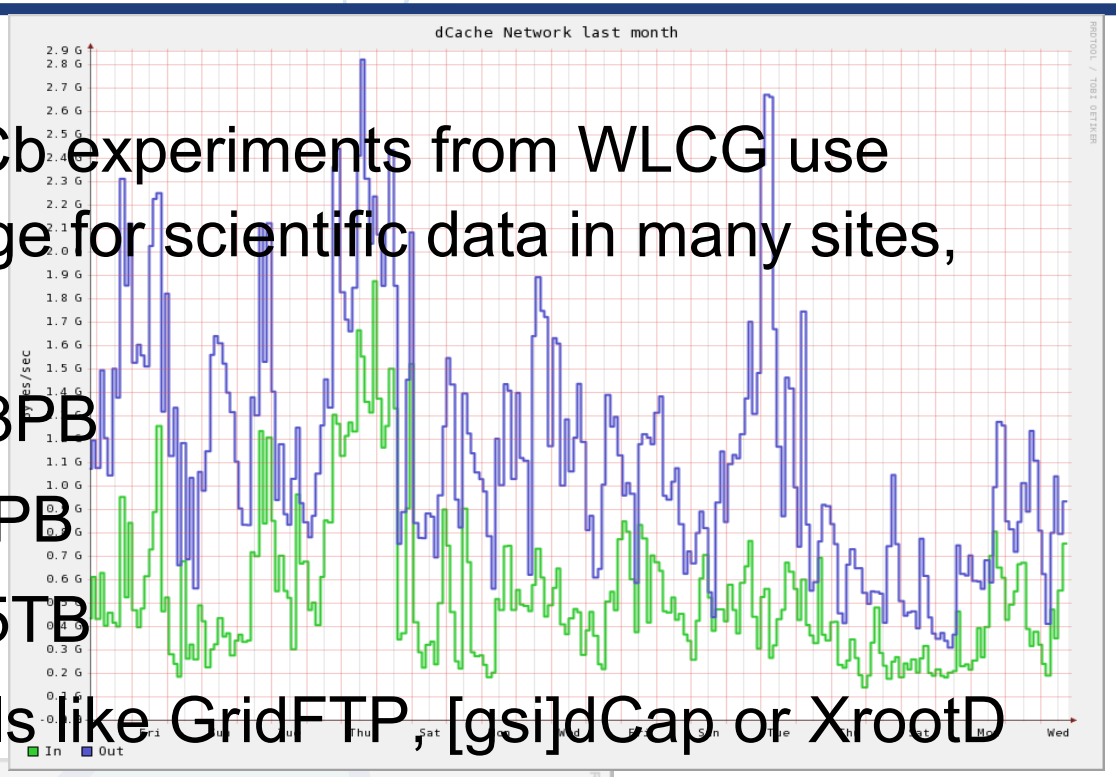
PIC
port d'informació
científica

LOFAR Processing Site



PIC
port d'informació
científica

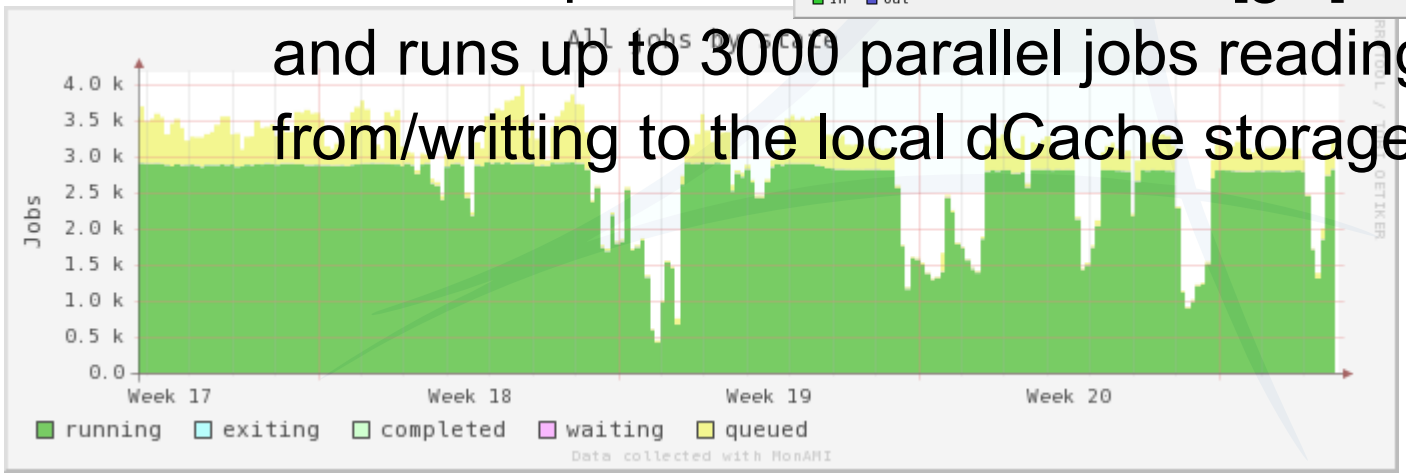
WLCG at PIC



ATLAS, CMS and LHCb experiments from WLCG use dCache as its storage for scientific data in many sites, among them PIC

- ATLAS: 2,3PB
- CMS: 1,35PB
- LHCb: 0,35TB

WLCG uses protocols like GridFTP, [gsi]dCap or XrootD and runs up to 3000 parallel jobs reading from/writting to the local dCache storage 24x7x365.

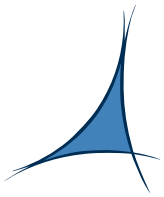


PIC
port d'informació
científica

dCache Future Schedule

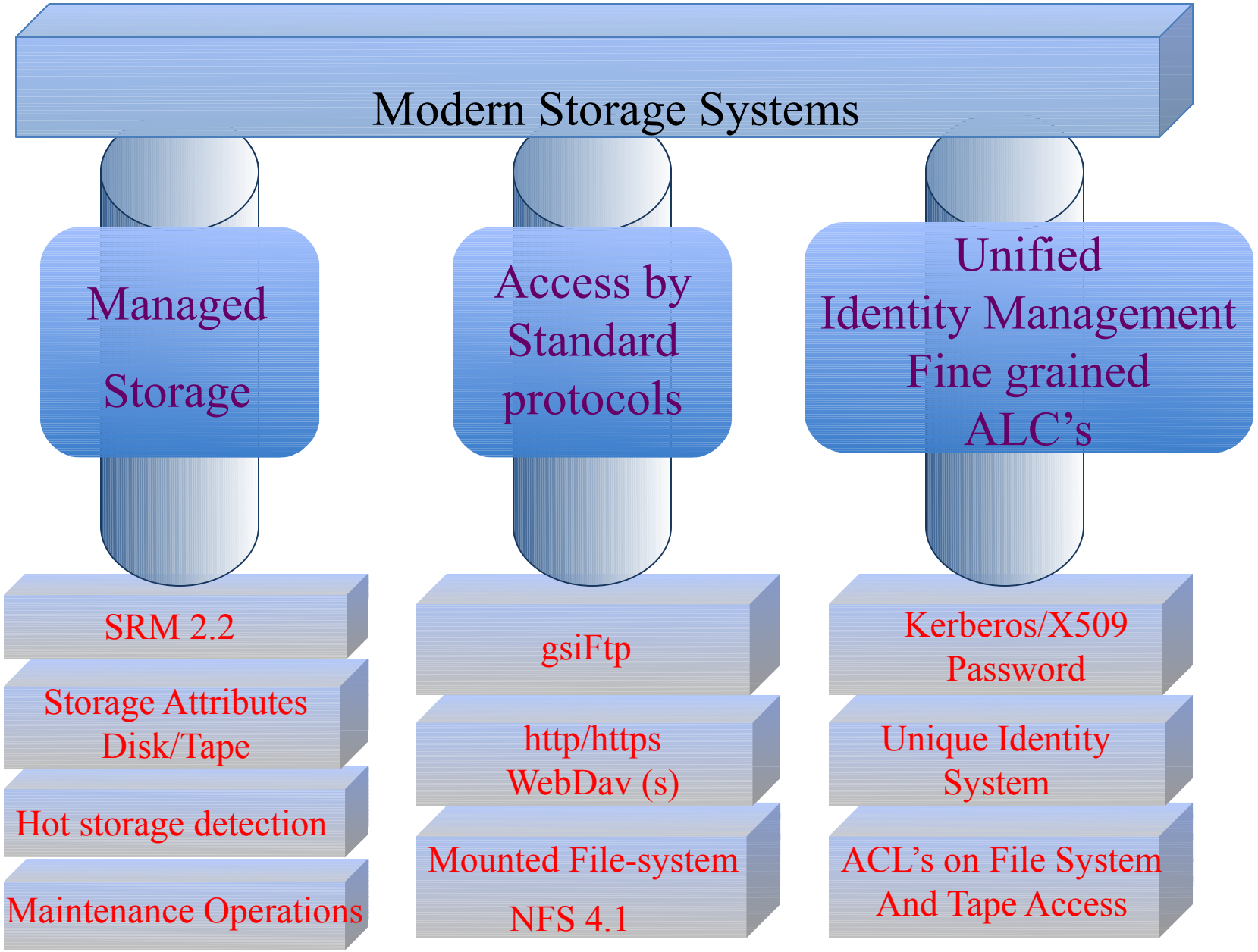
The involvement in EMI allows dCache to integrate the needs and requirements of non-HEP communites into their 2 years planning of new functionalies

- All standard protocols
- Support for small files in connection with Tape.
- Improved load balancing for write (no admin tuning should be necessary)
- following IBM and NetApp with the multi tier storage model. Tape only if disk breaks, next level is high condensed disk (which could be a set of dcache pool on top of a hadoop fs system or GPFS) and fast satellites (nfs 4.1/pNFS on SSDs).



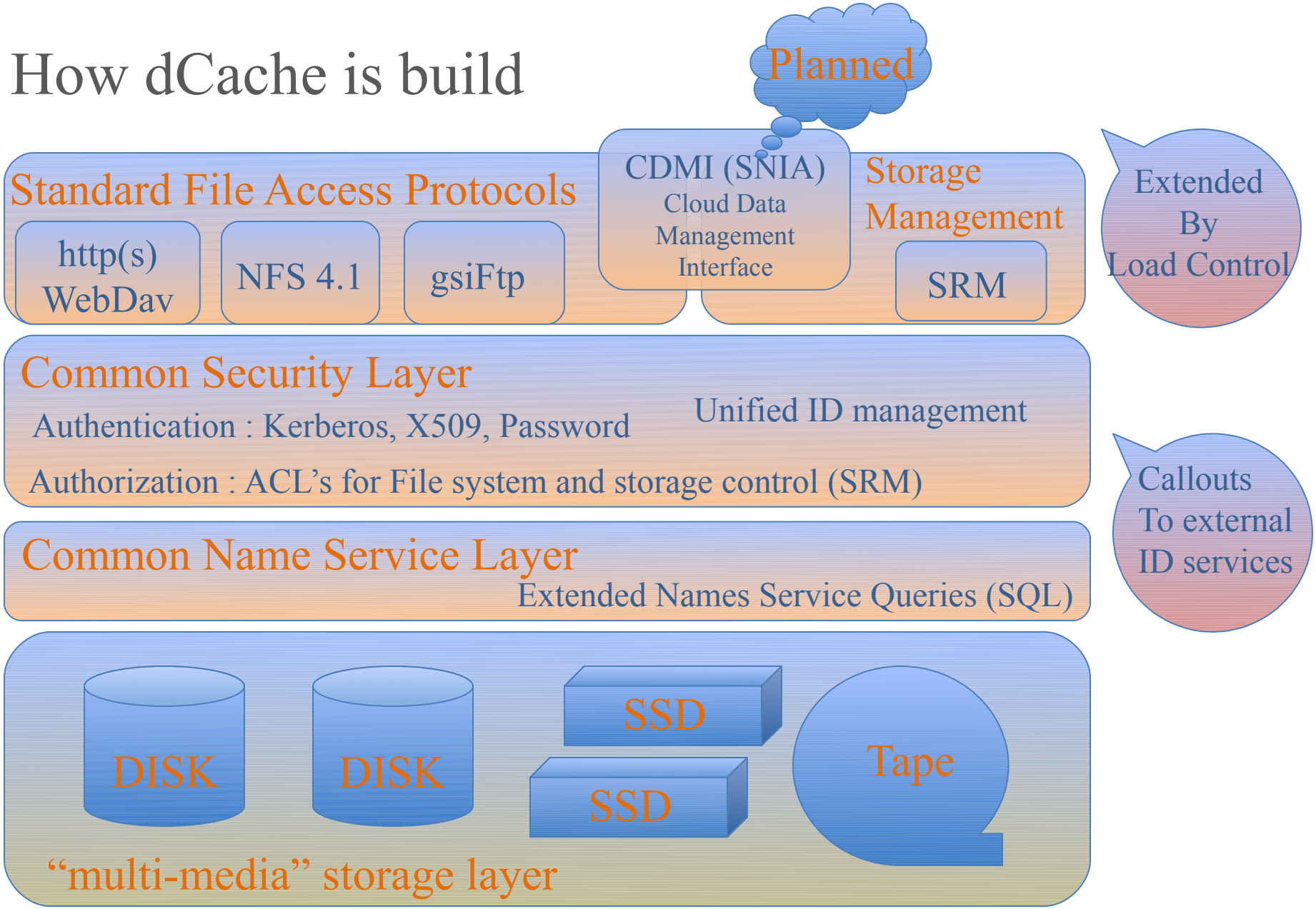
PIC
port d'informació
científica

BACKUP SLIDES

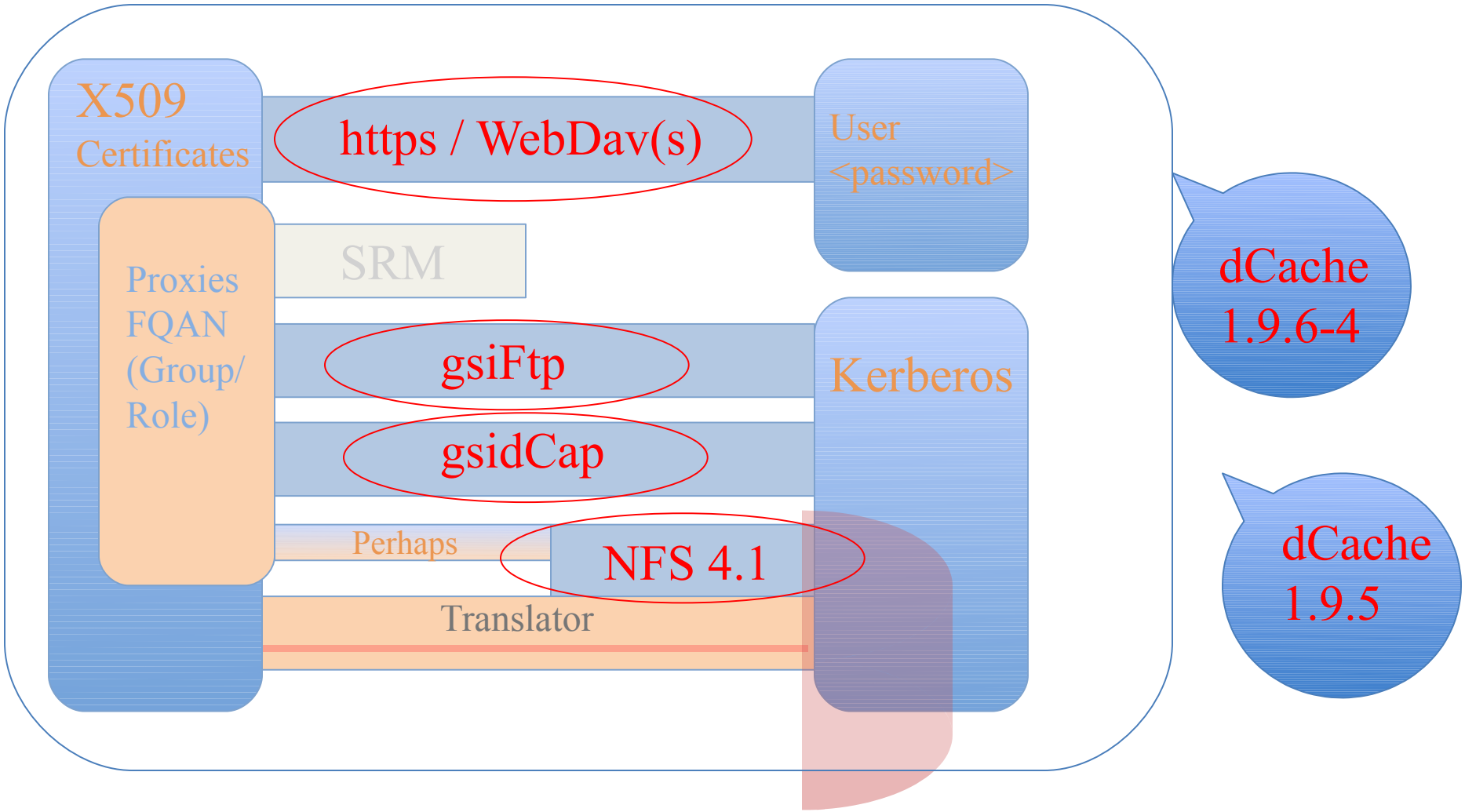

Modern Storage Systems
Managed Storage
Access by Standard protocols
Unified Identity Management Fine grained ALC's
SRM 2.2
Storage Attributes Disk/Tape
Hot storage detection
Maintenance Operations
gsiFtp
http/https WebDav (s)
Mounted File-system NFS 4.1
Kerberos/X509 Password
Unique Identity System
ACL's on File System And Tape Access

Can we solve this with dCache ?

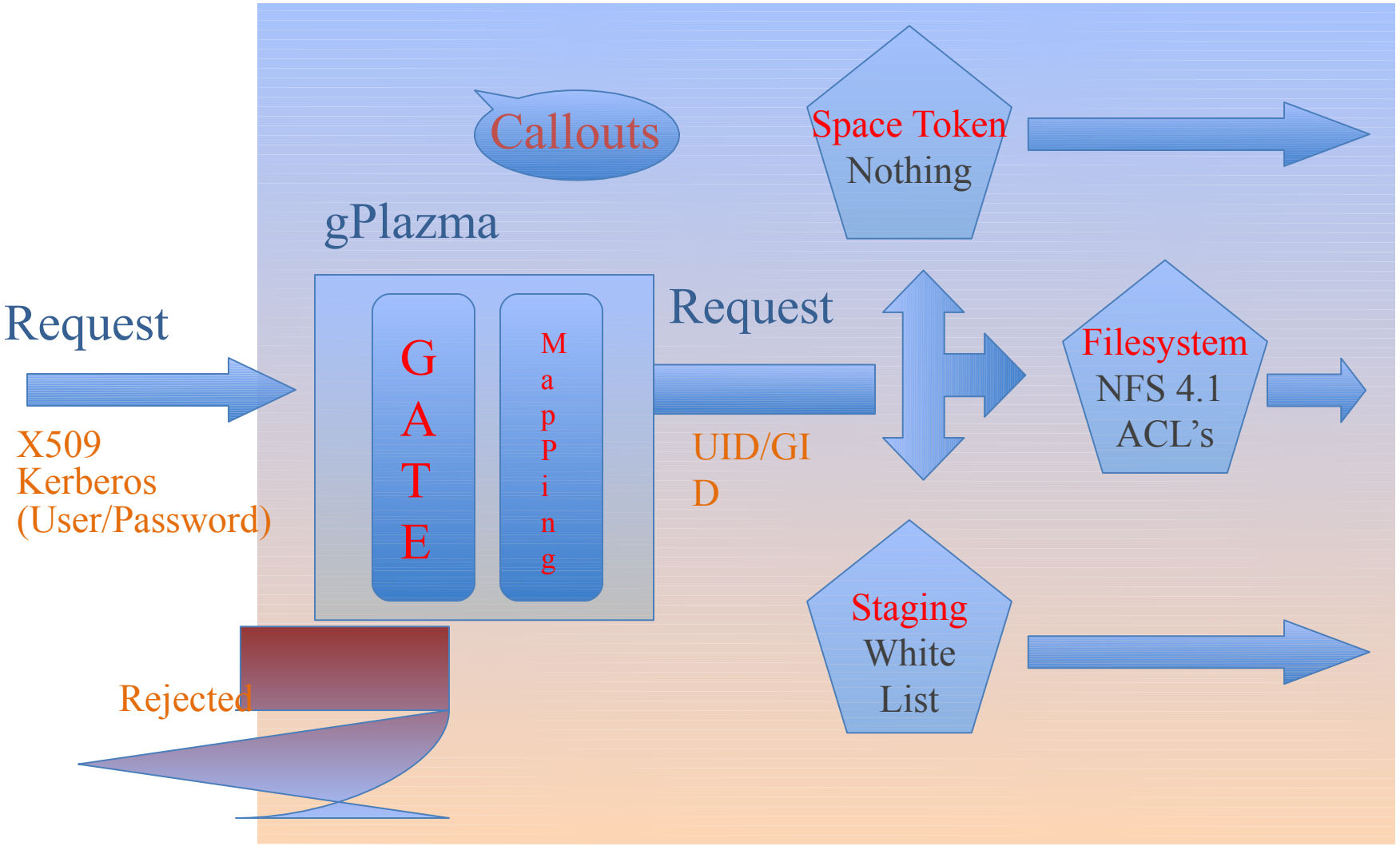
How dCache is build



dCache supported data access protocol suite.



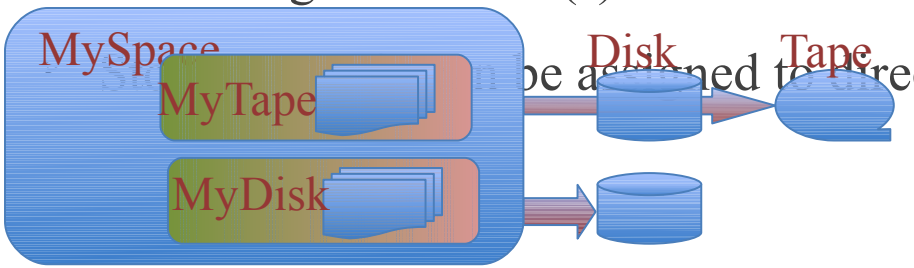
Authentication / Authorization Flow



dCache storage control (Spec)

Manual storage control (aka Managed Storage)

- SRM 2.2 (WLCG & Addendum & Addendum) compatible.
  - Define storage media (Disk/Tape) per file or “Space”.
  - Pin / Unpin files
  - Bring Online file(s)



be assigned to directory (sub) structure.

/users/x/MySpace/MyTape/Foo

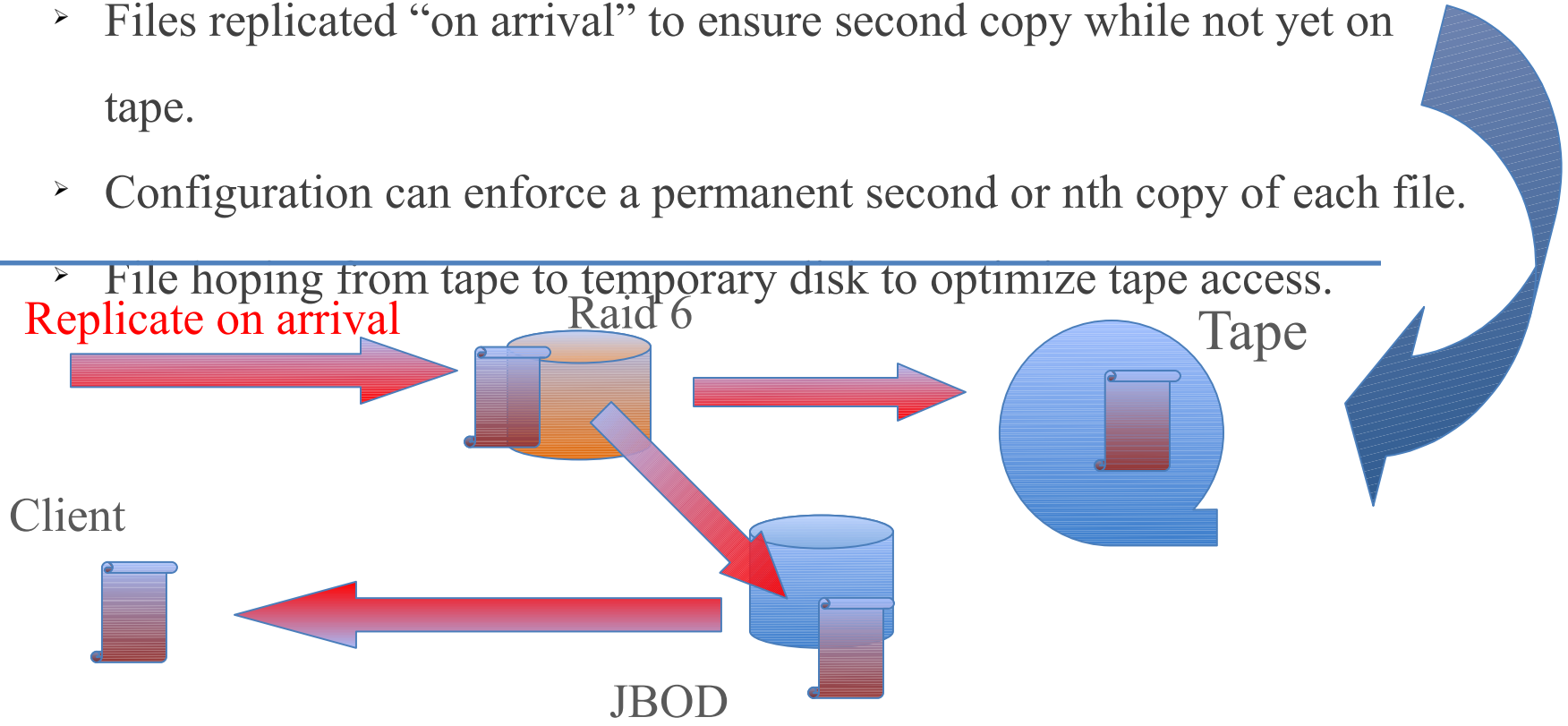
/users/x/MySpace/MyDisk/Foo

- Data can be scheduled for replication for maintenance or performance reasons.
  - Scheduled server downtimes
  - Server decommissioning

dCache storage control (Spec)

Automatic storage control (aka dCache file hopping)

- Data stored to tape and retrieved when needed.
- Files are automatically replicated to cope with high server load.
- Files replicated “on arrival” to ensure second copy while not yet on tape.
- Configuration can enforce a permanent second or nth copy of each file.
- File hoping from tape to temporary disk to optimize tape access.



In summary

dCache combines well known and
standardized data access
mechanisms, e.g. mounted file-system, web access,
browser/WebDav, with a broad
automatic and manual storage control functionality,
under a
common file name space and security umbrella.

With dCache, EMI and with that EGI is well prepared to serve new data intensive communities.

About supporting NFS 4.1

Or

Why is NFS 4.1 more than just file://…

NFS 4.1 in a mini nutshell

- NFS 4.1 (pNFS) is a IETF standard
- NFS 4 defines security standards (gss e.g. Kerberos)
- NFS 4.1 pNFS honors distributed data.
- All important storage vendors (IBM, PANASAS, EMC, NETAPP, dCache) are part of the NFS 4.1 working group under the roof of CITI (University of Michigan) and have an implementation ready.
- NFS 4.1 is available for Solaris and Linux (kernel 2.6.34)
- It will be in RH6 enterprise editions till end of the year.
- Back-ports for SL5 are in discussion.
- No vendor locking (e.g. GPFS, Lustre)
- dCache supports NFS 4.1 since 1.9.5 (Golden Release)

Further Reading

www.dCache.org