dCache, un système de gestion de données réparties

Mar 27, 2013 a la Séminaire Aristote
Patrick Fuhrmann

Preview

- Some dCache project stuff
  - Funding, partners, deployments
- Software design and features
  - Modules and message passing
  - Namespace and physical location
  - Plug-in services
- Project objectives and consequences
  - Committed to standards
  - Benefits of collaborations
- The dCache labs

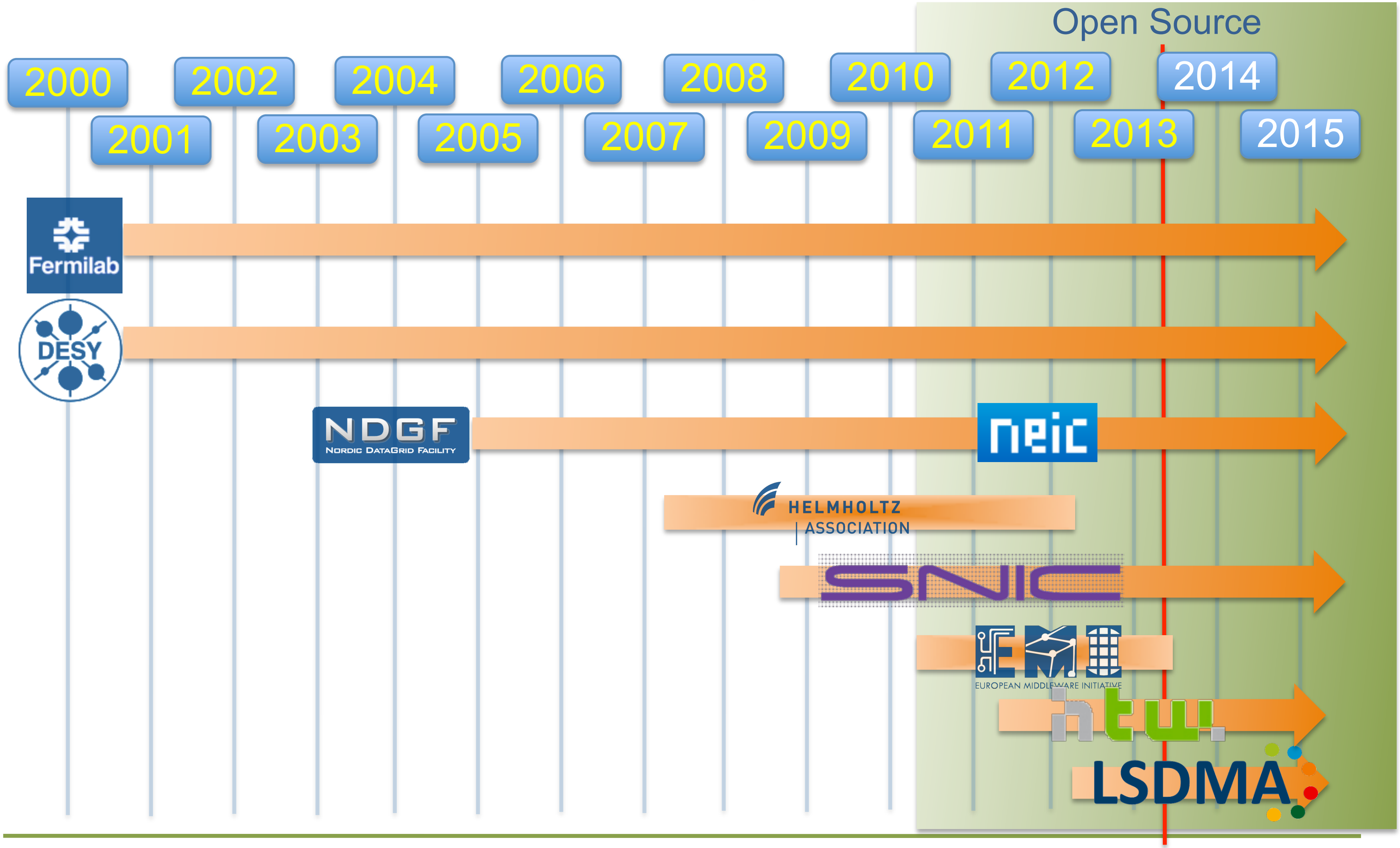
The project … stuff

Projects and funding

Partners and funding

dCache.org

dCache project timeline



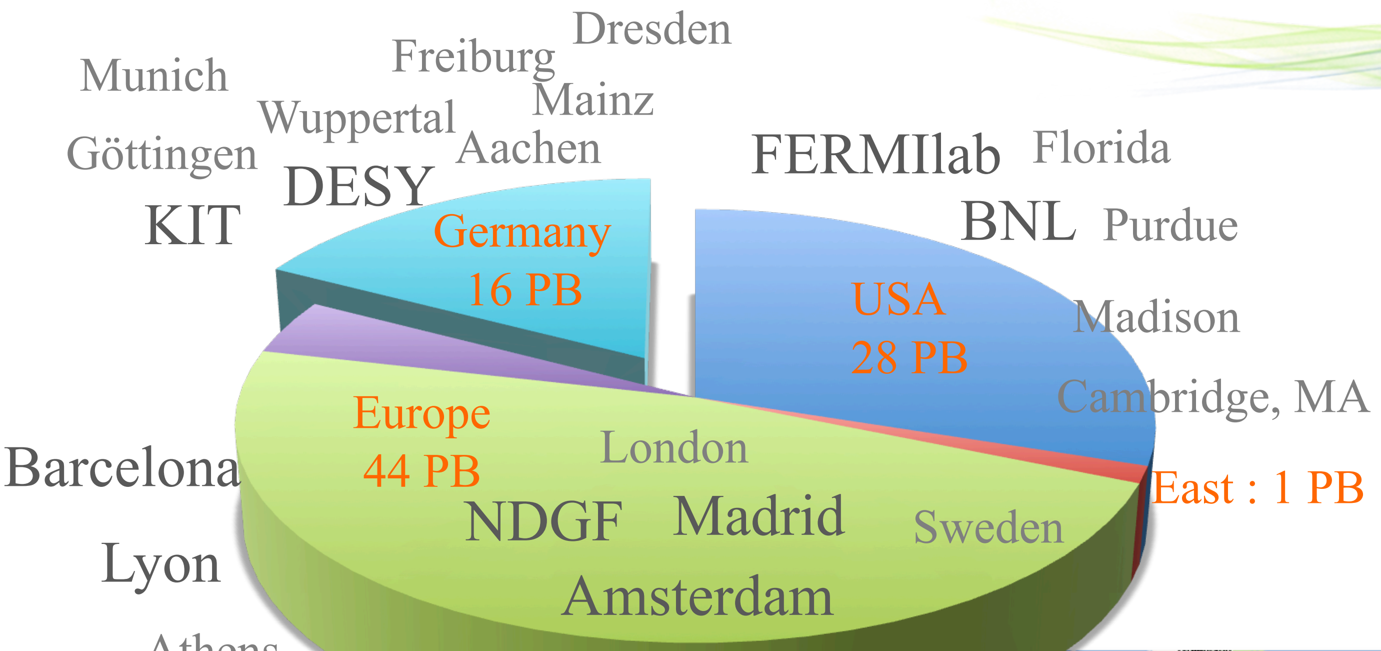
Deployments

WLCG Terminology

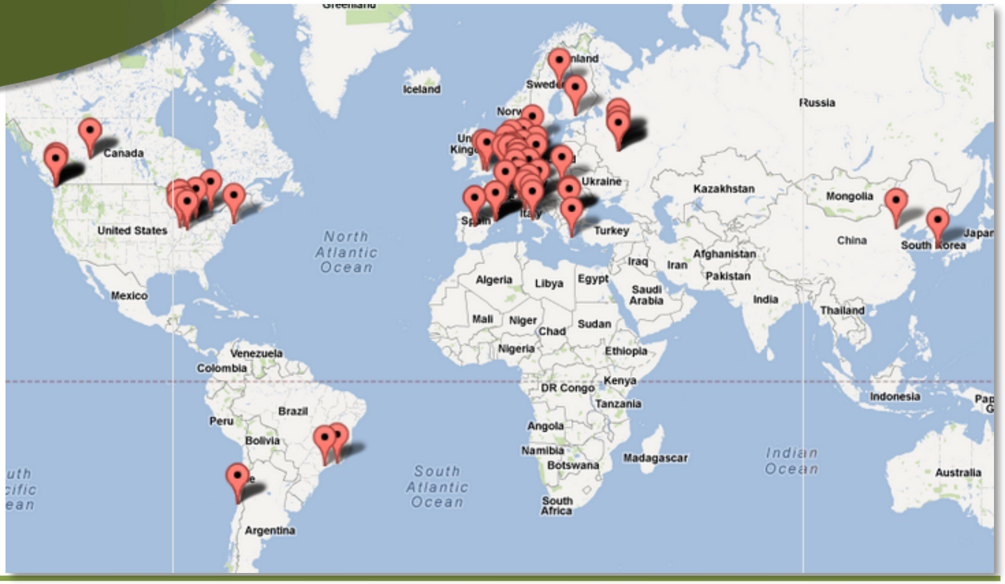
dCache.org

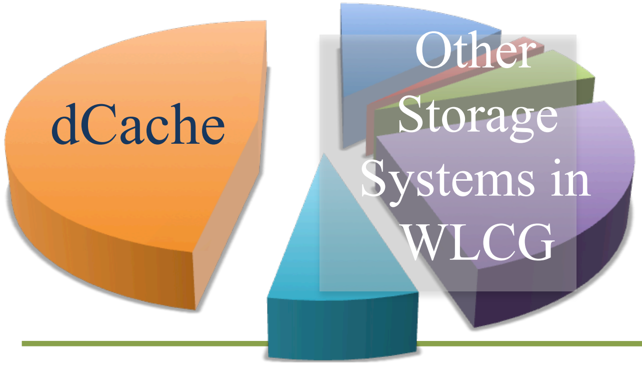
- 100 Pbytes of storage worldwide for WLCG
- 9 existing Tier I's
  - New York, Chicago
  - Vancouver
  - Lyon
  - Karlsruhe
  - Barcelona
  - Amsterdam
- 2 new Tier I's in Russia (Moscow and Dubna)
- 60 Tier II's

WLCG Deployments



Stolen from Tigran





Other communities

Stolen from Paul

European XFEL
ilc
USQCD
EMBL
LOFAR
SNIC
HASYLAB
Mu2e
SLOAN DIGITAL SKY SURVEY
ICECUBE
SOUTH POLE NEUTRINO DETECTOR
SciBooNE

Most important for sustainability

For all major partners, dCache is a strategic system, running in production.

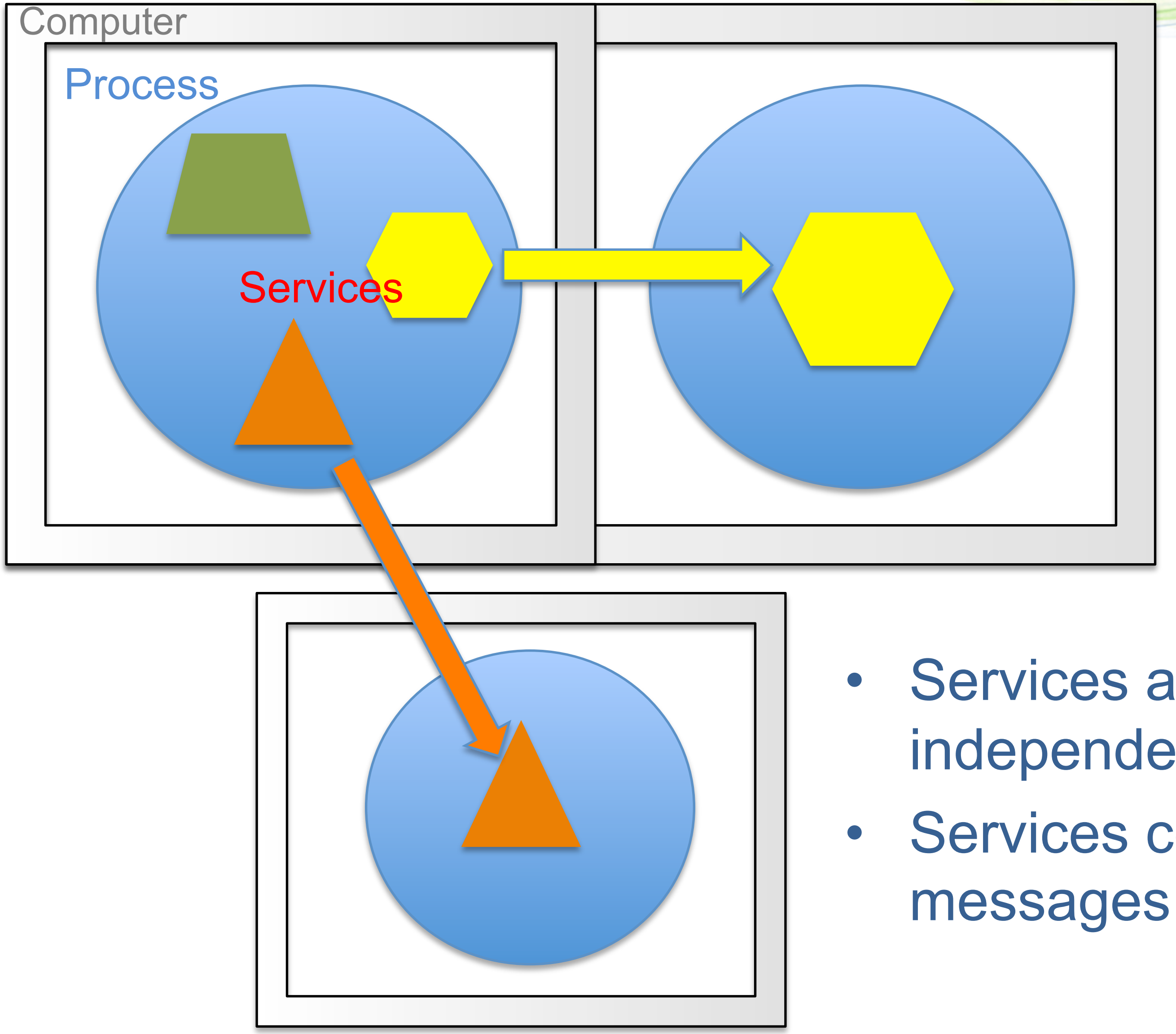
And now for something completely different

Software design and features

Design #1

Service Modules & Message Passing

Scale-out Design



- Services are location independent.
- Services communicate via messages.

Resulting in …. Fits all sizes

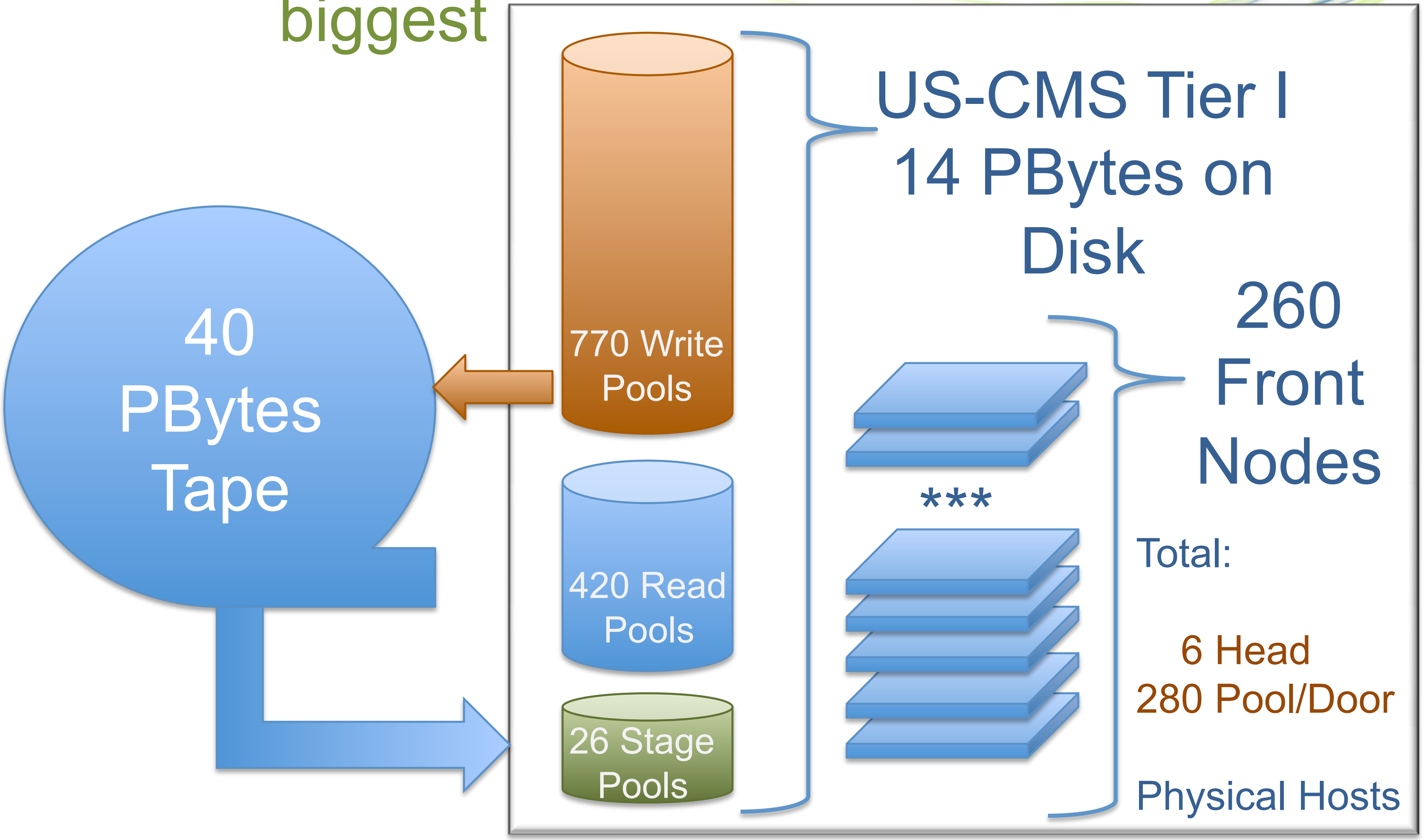
Starting with possibly the biggest

US-CMS Tier I
In FERMIlab
next to
Chicago

Information provided by Catalin Dumitrescu and Dmitry Litvintsev

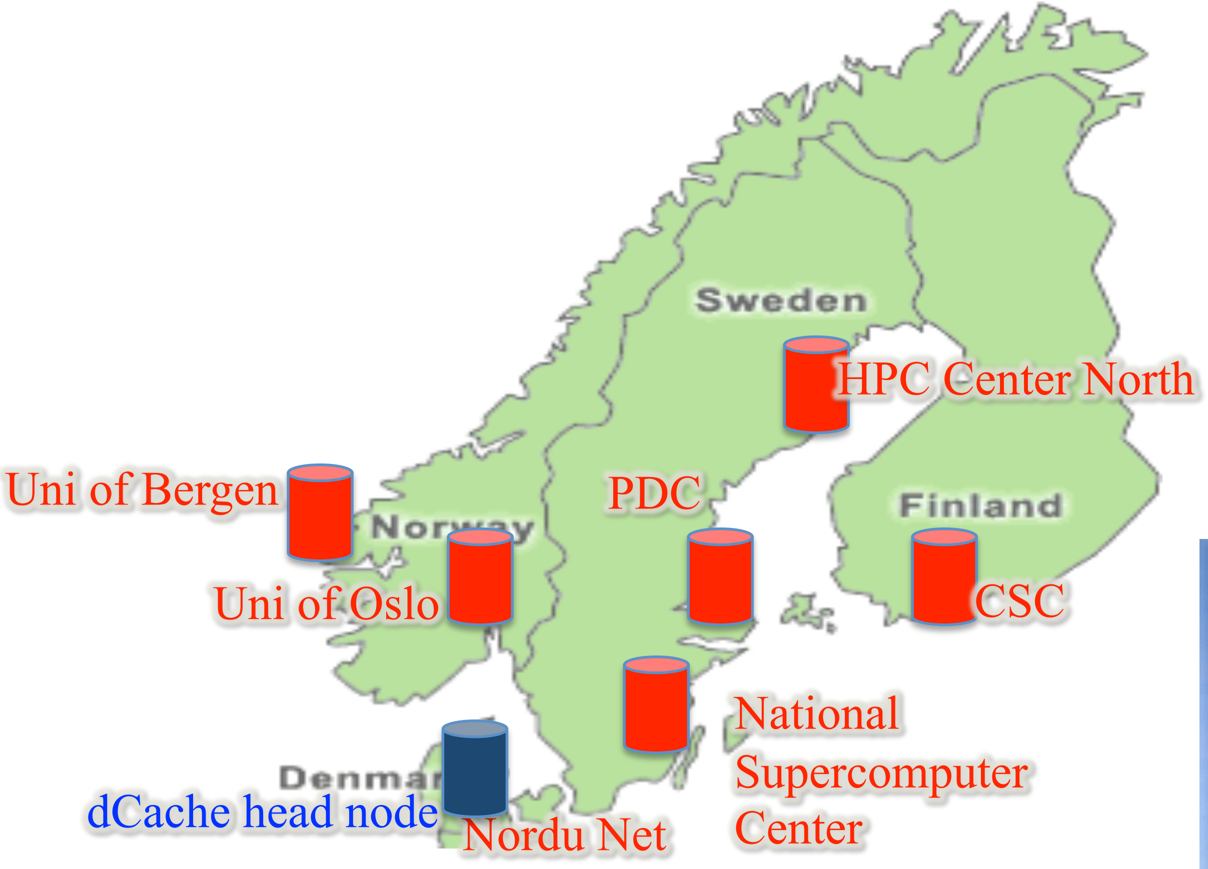
Starting with possibly the biggest

dCache.org



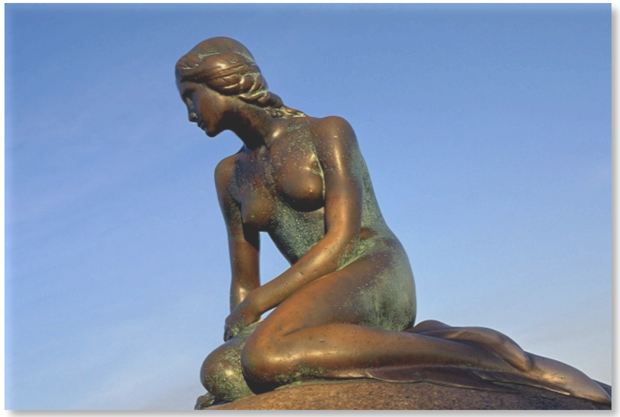
Information provided by Catalin Dumitrescu and Dmitry Litvintsev

To certainly the most widespread



4 Countries

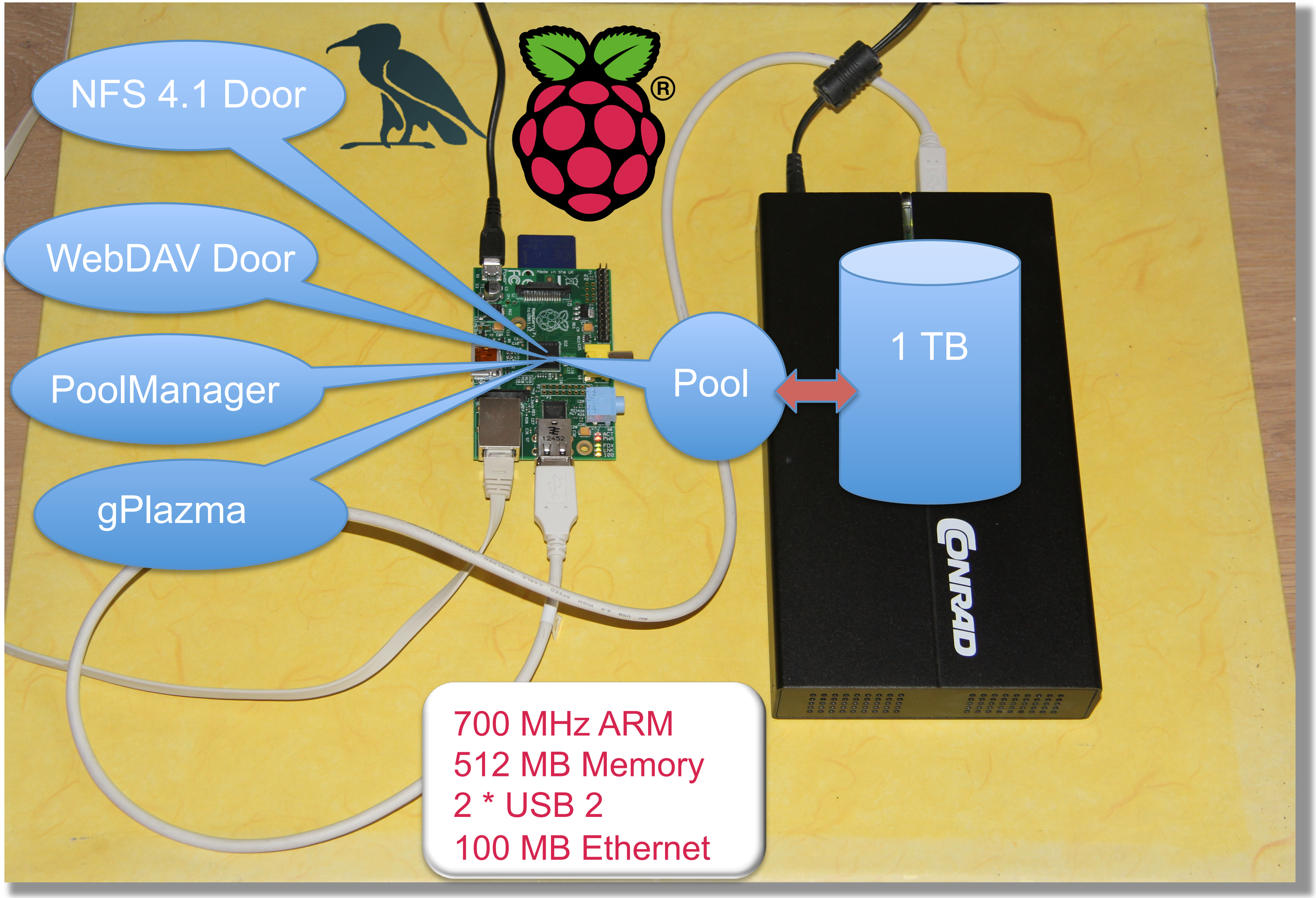
One dCache



Slide stolen from Mattias Wadenstein, NDGF

To very likely the smallest

One Machine – One Process



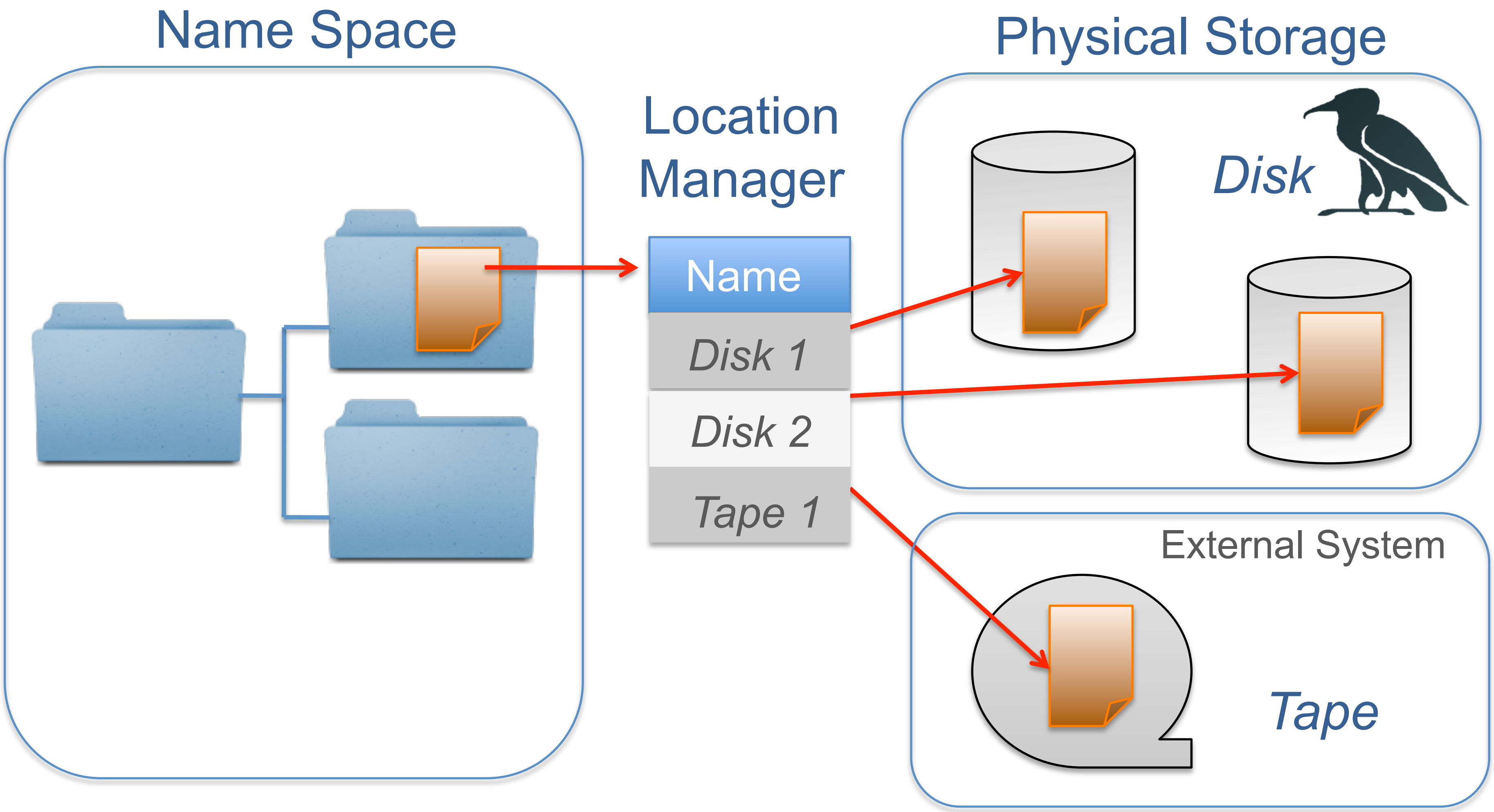
700 MHz ARM
512 MB Memory
2 * USB 2
100 MB Ethernet

Design #2

Namespace – Physical Storage separation

Design

Namespace – Storage separation

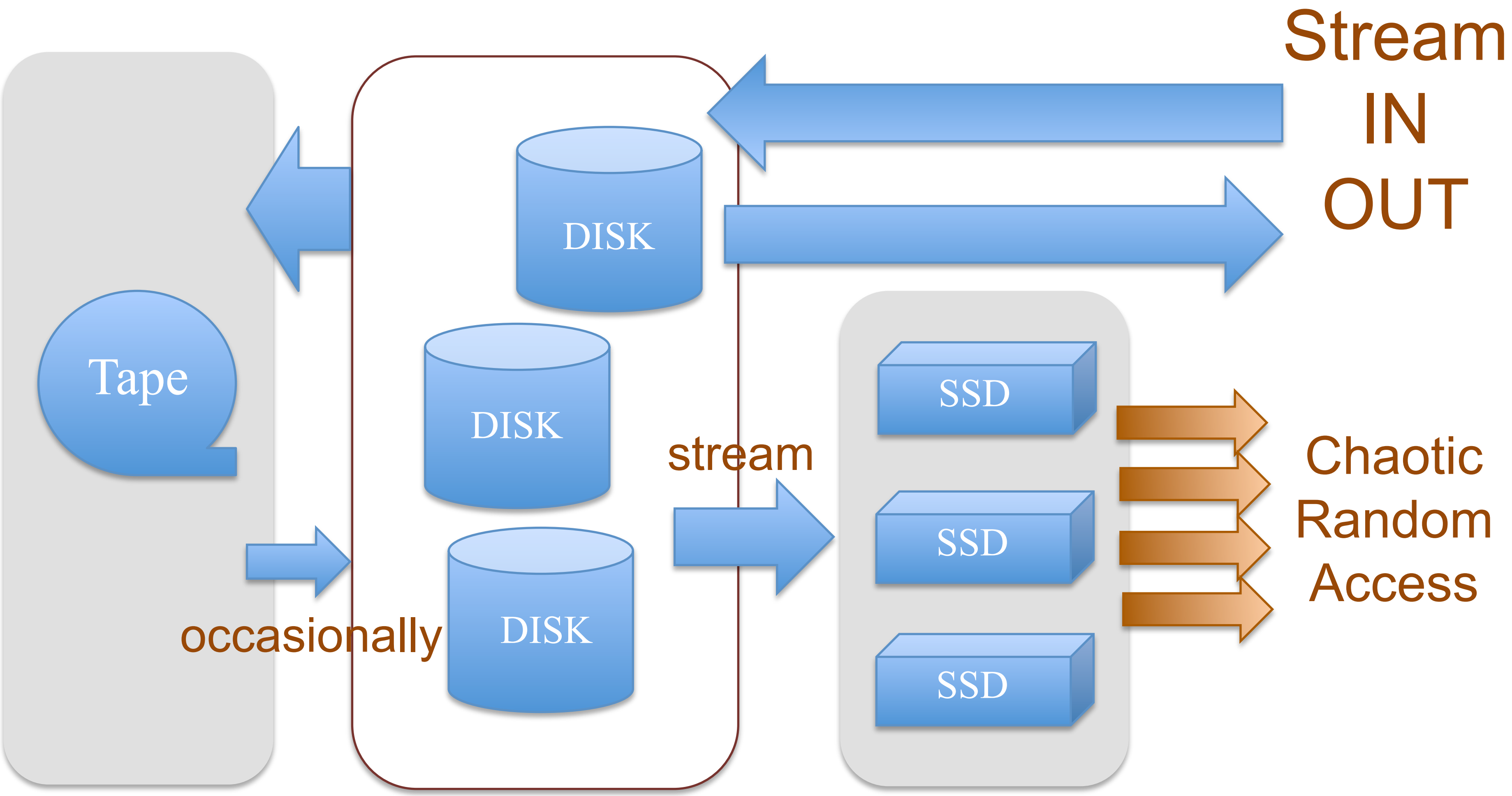


Resulting in …. Replica Management

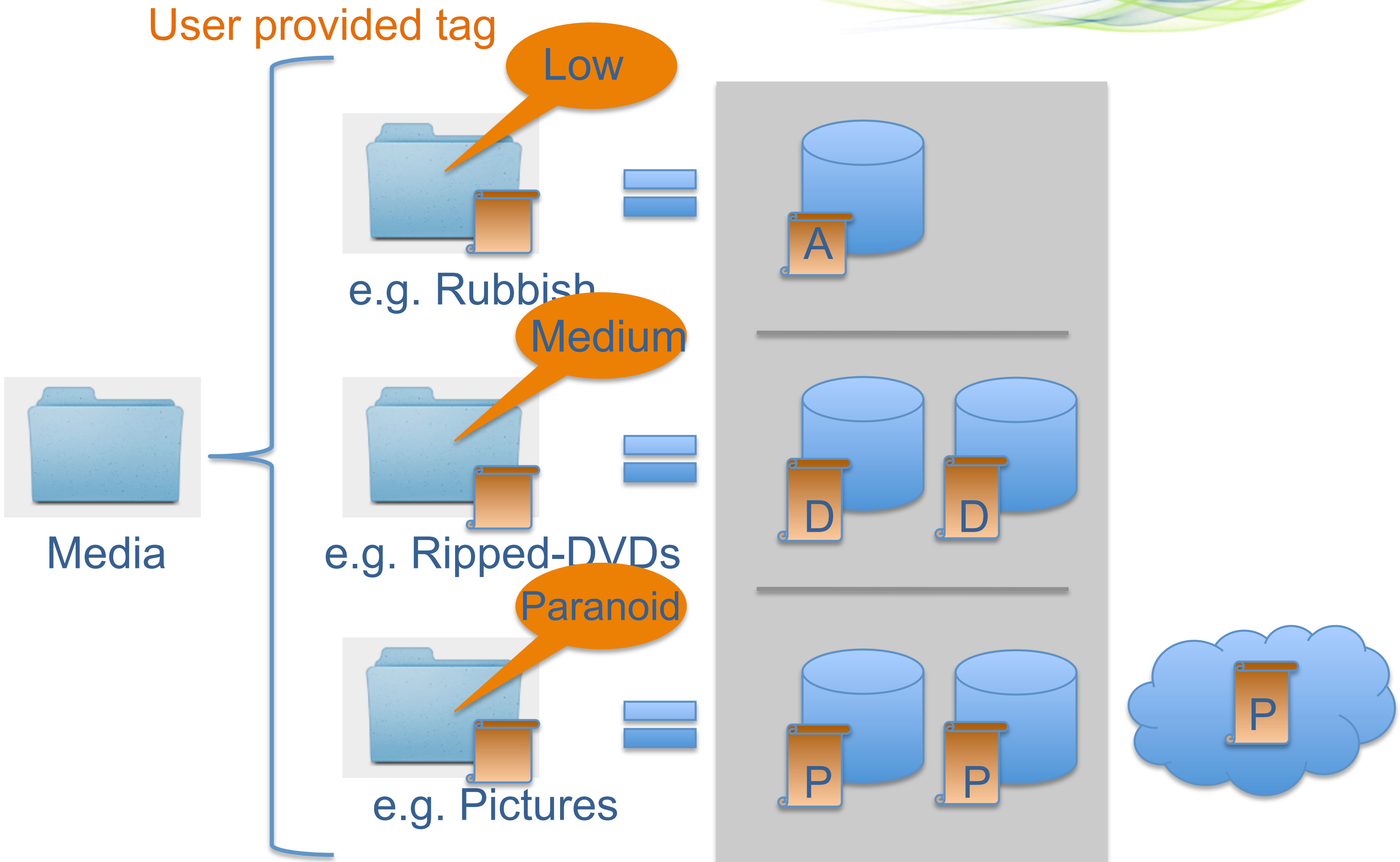
Replica Management

- Hot Spot detection
  - Files are copied from ‘hot’ to ‘cold’ pools
- Multi Media Support
  - File location is based on access profile and storage media type/properties
    - Fast streaming from spinning disks
    - Fast random I/O from SSD’s
- Migration Module(s)
  - Files can be manually/automatically moved or copied between pools.
  - Rebalancing of data after adding new (empty) pools.
  - Decommission pools.
- Resilient Manager
  - Keeps max ‘n’ min ‘m’ copies of a file on different machines.
  - System resilient against pool failures.
- Tertiary System connectivity (Tape systems)
  - Data is automatically migrating to tape.
  - Data is restored from tape if no longer on disk

3 Tier Storage



Data resiliency

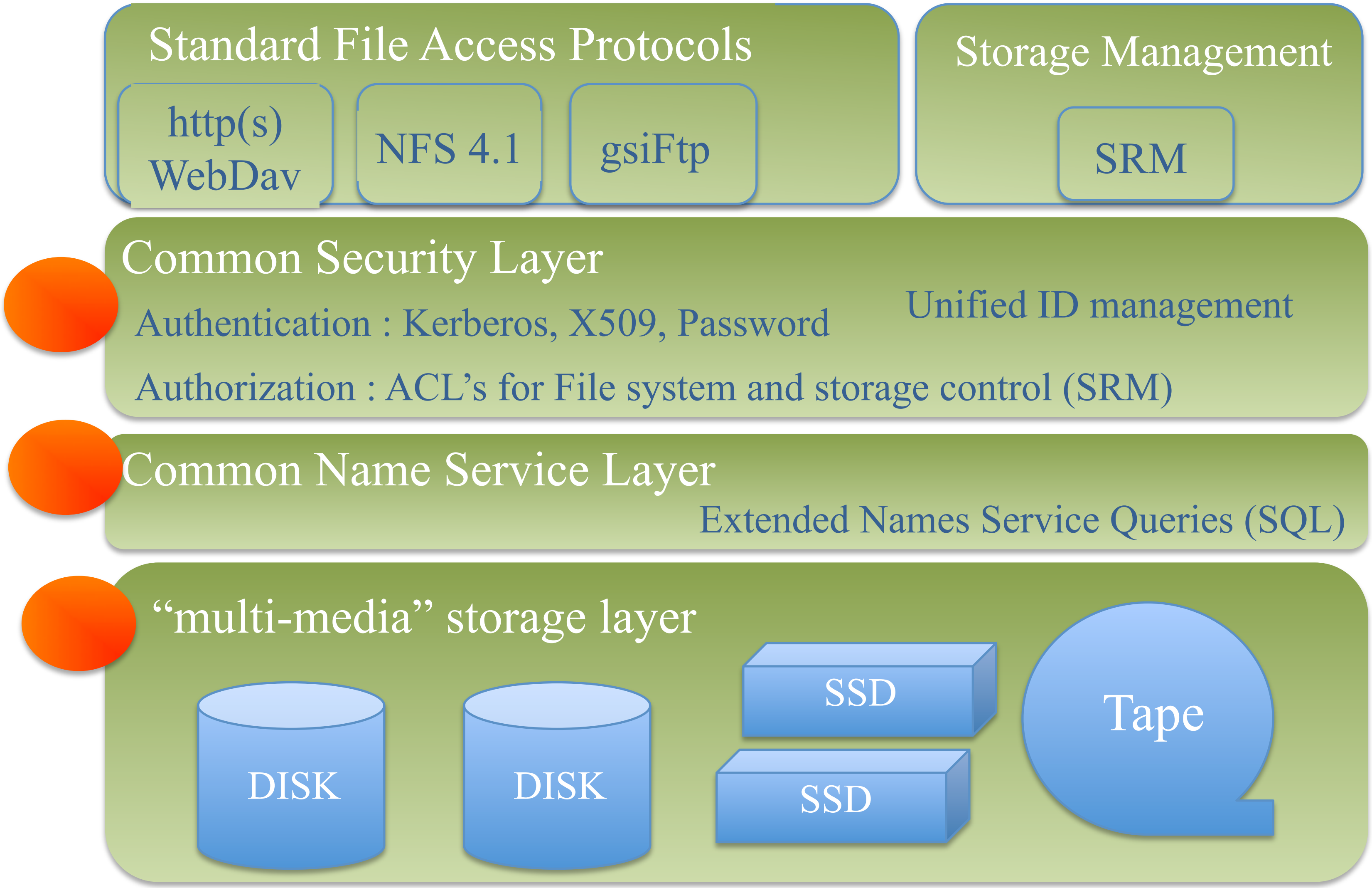


Design #3

Services allow plug-ins

Resulting in … customizable behavior

Plug-in Facility



Standard File Access Protocols

http(s) WebDav

NFS 4.1

gsiFtp

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)

“multi-media” storage layer

DISK

DISK

SSD

SSD

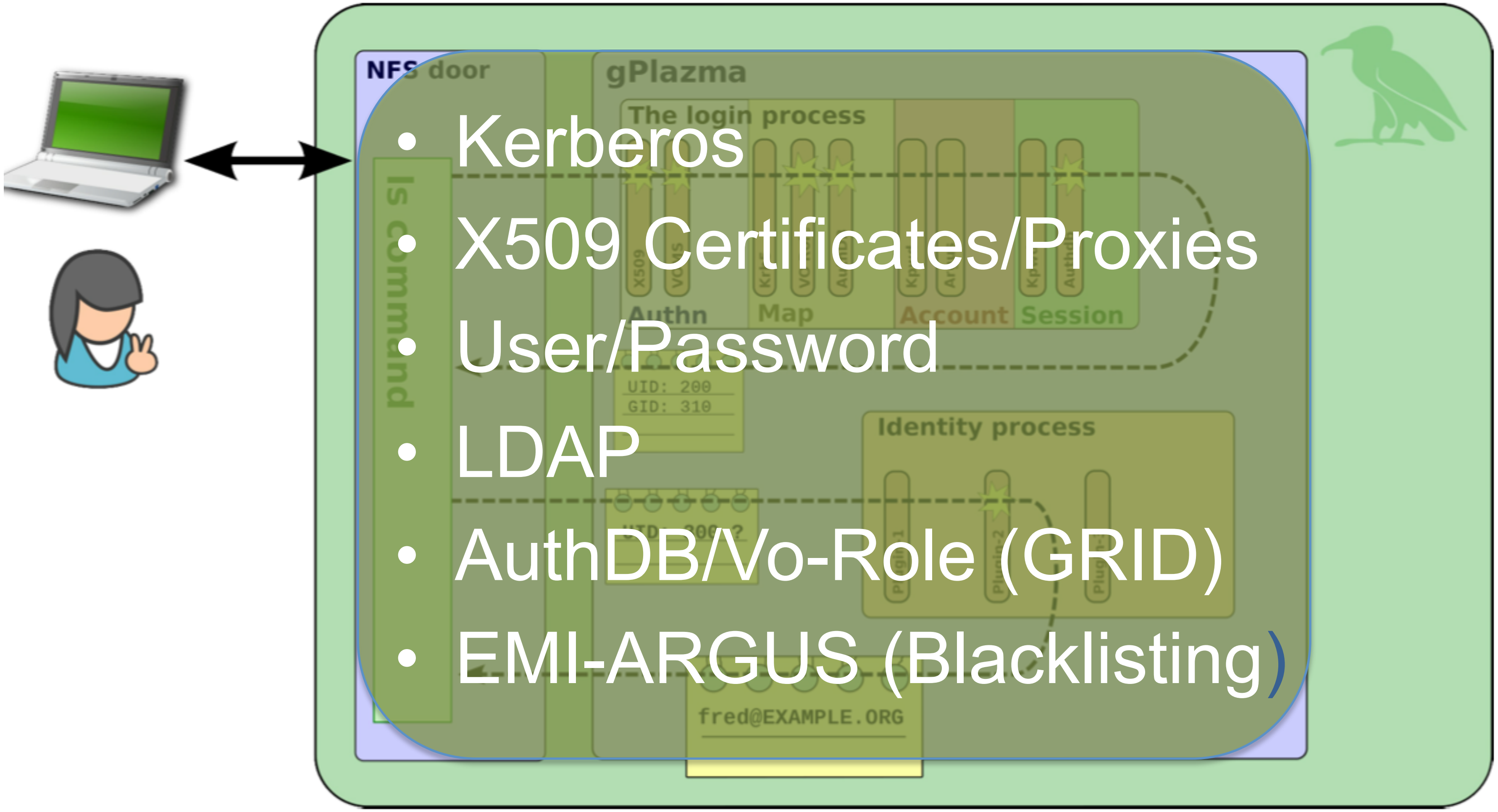
Tape

Plug-in Facility

- gPlazma / Authentication system
  - Authentication
  - Mapping (user names and UID/GID)
  - Actually in the door:
    - LFN to PFN mapping for CMS and Atlas
- Name space provider (PNFS -> chimera)
- File System back end (Hadoop, GPFS , …)
- File distribution / reshuffling system

gPlazma plug-ins (e.g. NFS4.1)

Slide stolen from Paul Millar



- Kerberos
- X509 Certificates/Proxies
- User/Password
- LDAP
- AuthDB/Vo-Role (GRID)
- EMI-ARGUS (Blacklisting)

Now … about some project objectives

Objective #1

Committed to standards

Resulting in …

- Support of
  - – GLUE 2
  - – SRM
  - – WebDAV
  - – NFS 4.1 / pNFS
  - – The Storage Accounting Record (StAR)
- Working on Cloud protocols

Makes dCache an Open Source competitor to expensive industry solutions and attracts non WLCG communities.

Objective #2

We believe in the power of collaborations

Resulting in

- European Middleware Initiative (EMI)
  - Funding for very interesting development
  - Learning about the storage needs of non HEP communities
- CERN Datamanagement
  - HTTP Dynamic Federation
- Globus-Online
  - gridFTP and staging
- **L**arge **S**cale **D**ata **M**anagement and **A**nalysis
  - about ‘federated identity and storage access’

The dCache labs

The small file versus tape issue.

dCache labs
Small File Support for Tape

Or, Why do small files kill tape systems ?

- 0 Byte files occupy between .5 and 1.6 Mbytes on tape. So, small files are wasting space.
- Writing file marks forces the drive to synchronize tape writing (halts streaming)
- LTO Spec :
  - 80 Seconds max seek time
  - 50 Seconds average
  - Which means: For reading files from tape, which are not exactly in order, each transfer takes about 50 Seconds minimum.
- If data is not on same tape, mount/dismount has to be added (30 – 60 Seconds)
- Tape systems consist of 3 non-shareable units :
  - Robot (Arm and gripper)
  - Drive
  - Tape

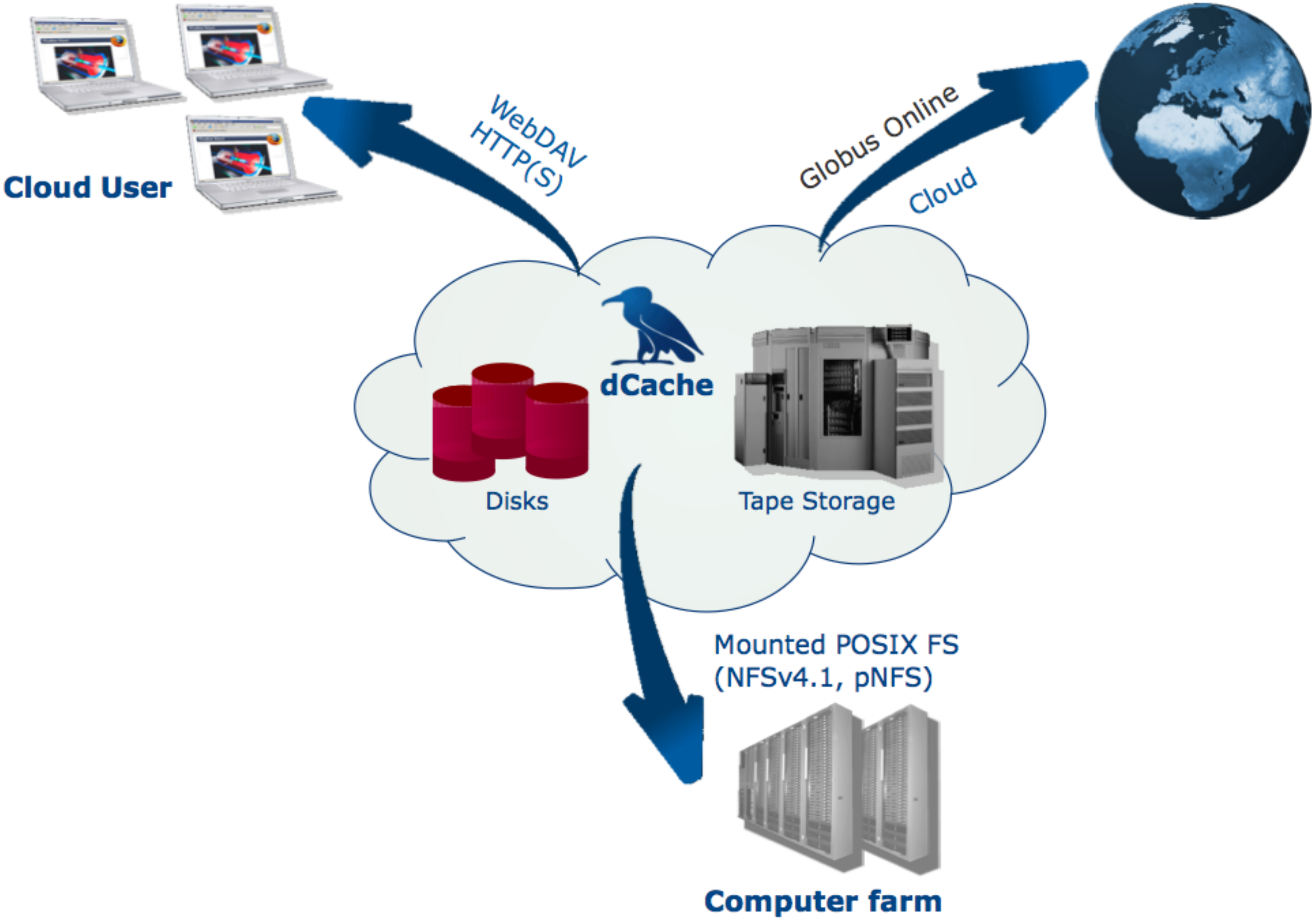
Small Files versus tape

- Transparent for the user:
  - We ‘tar’ or ‘cpio’ files before they are flushed to tape.
  - We extract the correct file from the archive if needed.
- Options:
  - Only the requested file is extracted, or
  - when the first file of a container is requested, dCache could extract all files of the container.
- As the container file is still on disk for awhile after the first file has been extracted (depending on space availability), subsequent requests for small files will be handled w/o further tape access.
- We could even pin recalled containers for some time.

TOM HANKS HALLE BERRY JIM BROADBENT HUGO WEAVING JIM STURGESS DOONA BAE BEN WHISHAW
JAMES D'ARCY ZHOU XUN KEITH DAVID DAVID GYASI MIT SUSAN SARANDON UND HUGH GRANT
CLOUD ATLAS
VERGANGENHEIT.GEGENWART.ZUKUNFT.
ALLES IST VERBUNDEN.
EIN WACHOWSKI / TYKWER FILM
AB 15. NOVEMBER IM KINO
XFILME
cloudatlas-derfilm.de
XVERLEIH

The scientific cloud

dCache.org



About Cloud Protocols

- Cloud storage protocols S3 versus CDMI ?
  - We got a student from the University of Applied Sciences Berlin
  - She will initially implement CMDI (for her master)
  - As a side effect we will hopefully support S3 sooner or later.
  - But: we are still trying to understand if this is really necessary

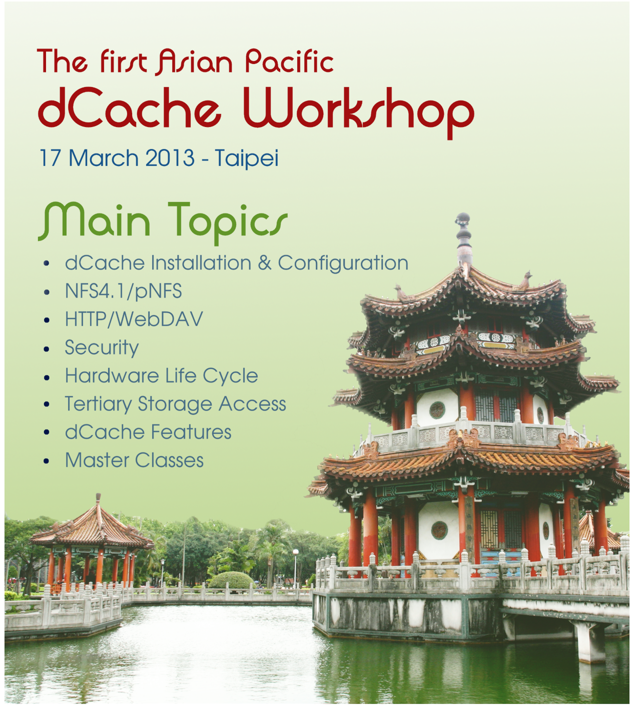
More dCache labs

- Enhanced 3 Tier storage
  - e.g. scheduling of data location changes
  - Migration of data based on access count
- Adopting more standard identity mechanisms, IdP (e.g. Shibboleth, OpenID)

And nearly done

Where to learn more about dCache ?

- One workshop per year in Europe.
- One dCache day during the GridKA school.
- First Asian Pacific dCache in Taipei (last week).



The first Asian Pacific

dCache Workshop

17 March 2013 - Taipei

Main Topics

- dCache Installation & Configuration
- NFS4.1/pNFS
- HTTP/WebDAV
- Security
- Hardware Life Cycle
- Tertiary Storage Access
- dCache Features
- Master Classes

With participants from

○ Australia
○ Taiwan
○ Thailand
○ Japan
○ India
○ Germany

Summary

- dCache is a professional Open Source project, with a large developers base and significant community support.
- Funding is provided by a variety of sources.
- dCache is committed to standards
  - To ease customer acceptance for storage
  - Simplifies system administrators life.
- The dCache system evolves, following
  - Community requirements (SRM, GLUE2, StaR …)
  - Technology changes (NFS 4.1, SSD, Hadoop FS, …)

Next European dCache Workshop
27 May – 29 May
In Berlin

further reading

www.dCache.org