dCache.org

The dCache labs

7th International dCache Workshop

Patrick Fuhrmann

neic

Fermilab | NDGF Nordic DataGrid Facility | EMI European Middleware Initiative | DESY | Helmholtz Association | Federal Ministry of Education and Research | LSDMA

Content

- CMS Disk / Tape separation
- dCache supporting federated IdM
- Multi Tier Storage
- Small file support to optimize tape
- Single client performance
- Scientific Storage Cloud

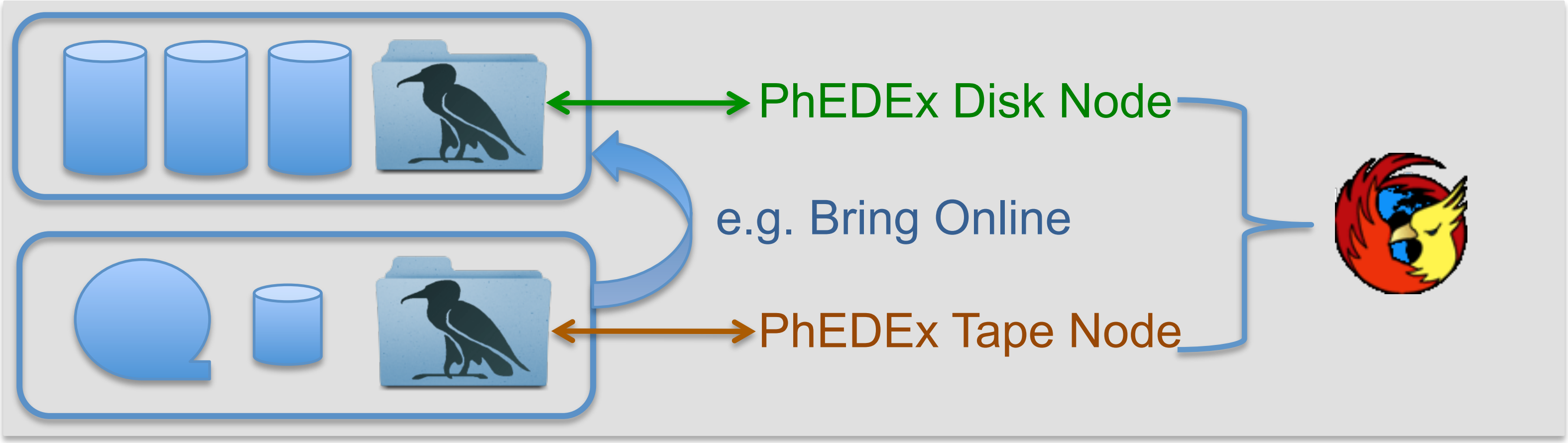
Completed

- gPlamza 2
- NFS 4.1
- WebDAV

CMS Tape Disk Separation

CMS Disk / Tape separation

- CMS is planning to strictly separate disk and tape storage elements at the Tier I level.
  - With the available network bandwidth of the OPN, it should be faster to take data from another Disk-Tier-1, than from Tape.
  - CMS would like to reduce the number of Tier-I's with Tape. (Complex and expensive management)

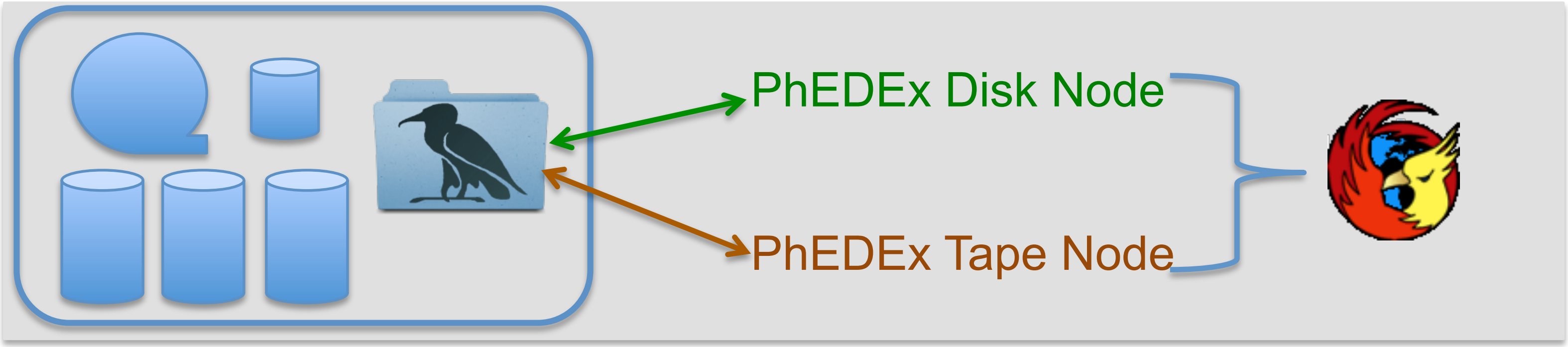


PhEDEx Disk Node

e.g. Bring Online

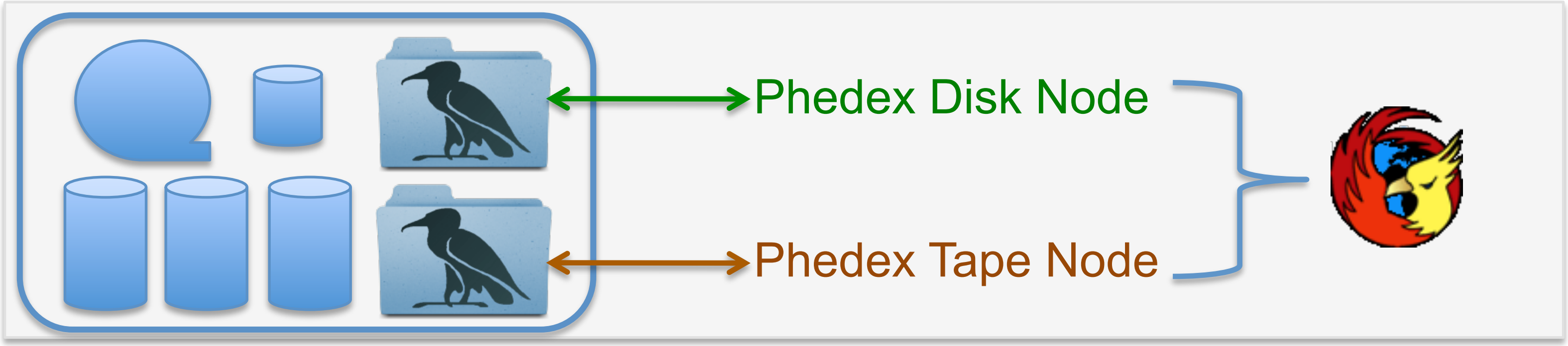
PhEDEx Tape Node

A possible solution



- A single dCache pretends to be ‘two’
- Highly customized PhEDEx Adapter
  - Stat of file has to be replaced by location query
- Transitions (limited selection)
  - ● Get file from tape to disk : -> Done : Bring Online
  - ● Migrate file to tape (selectively)
  - ● Accept file to disk (from other Tier I) which is already on tape locally
  - ● Remove files from tape but keep file on disk

Other solution



Phedex Disk Node

Phedex Tape Node

- A single dCache with two similar name space trees
  - One as tape endpoint and the other as disk endpoint
- PhEDEx Adapter nearly unchanged
- Transitions
  - 🟢 Get file from tape to disk : -> Done : Bring Online
  - 🟠 Migrate file to tape (selectively)
  - 🟢 Accept file to disk which is already on tape locally (different file in dC.)
  - 🟠 Remove files from tape but keep file on disk

CMS Disk / Tape separation (cont)

- Plan
  - PIC (Pepe) is organizing the effort and will help us evaluating solutions. Support from other sites is welcome.
  - We can begin right away with two completely independent name spaces in one dCache.
  - We can work on the optimization gradually.
  - Interesting: flush files to tape individually or conditionally

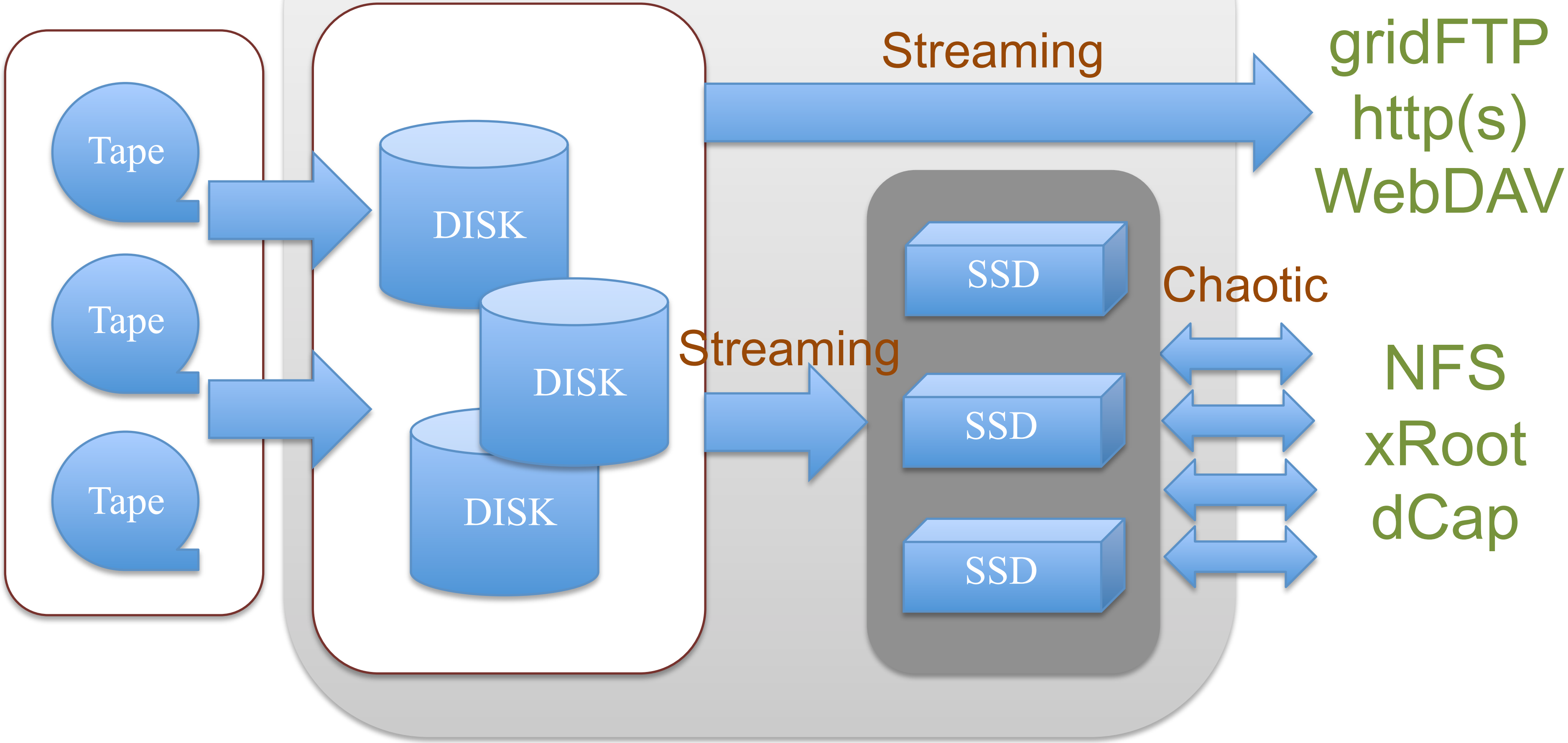
Federated Identities

Federated Identities

- General issue:
  - Use credentials from site-A to access data at site-B.
- Plenty of possible combinations
  - SAML or X509 including conversion (e.g. STS)
  - Web-based (including ECP Profile)
  - Generic (no portals involved)
  - And all possible combinations
- We will agree on an example setup
  - “Relying Party”: dCache for sure.
  - Likely SAML support
  - Details need to be negotiated in LSDMA WP1
- Goal for dCache :
  - Accept (federated) Identity Providers
    - OpenID (Google, Facebook), Shibboleth , SAML, Umbrella

Multi Tier Storage

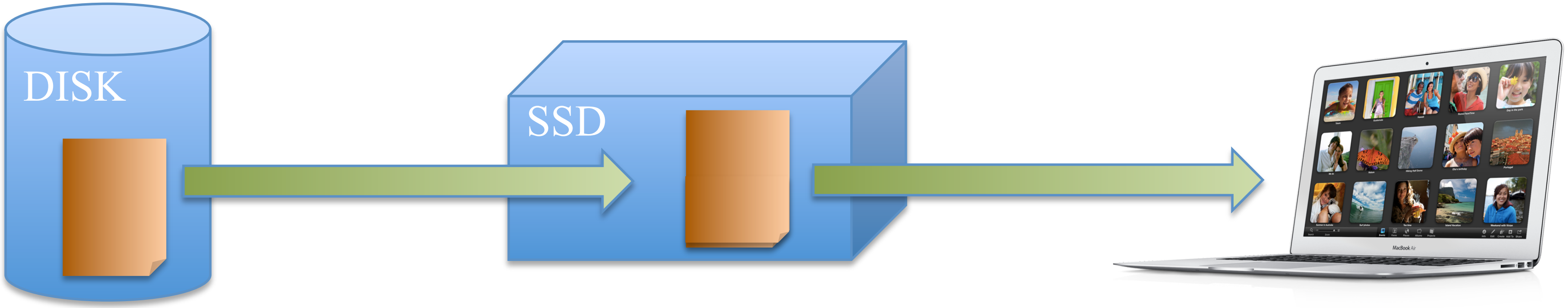
Multi Tier Storage



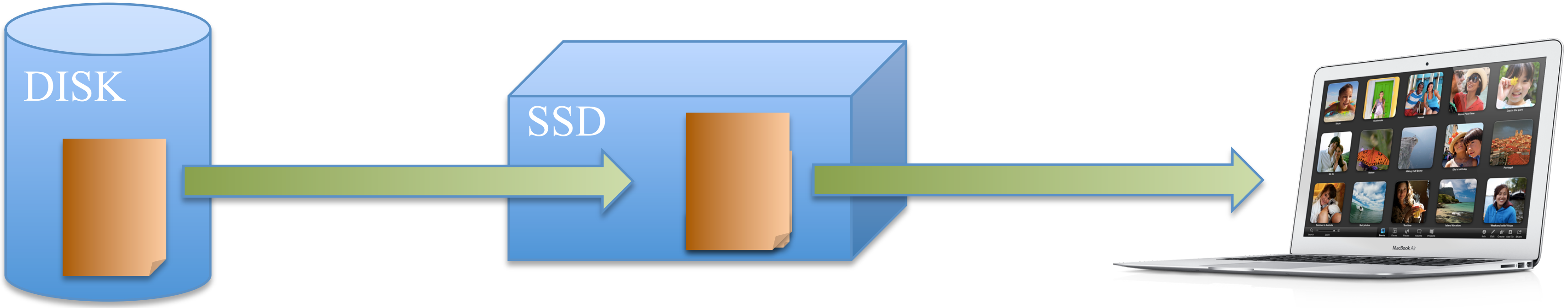
This you can already do with dCache, BUT

Multi Tier Storage

- Can already be configured, but



- Tigran : Better would be



- Will be done, if we find resources

Small file support for tape

What’s the issue

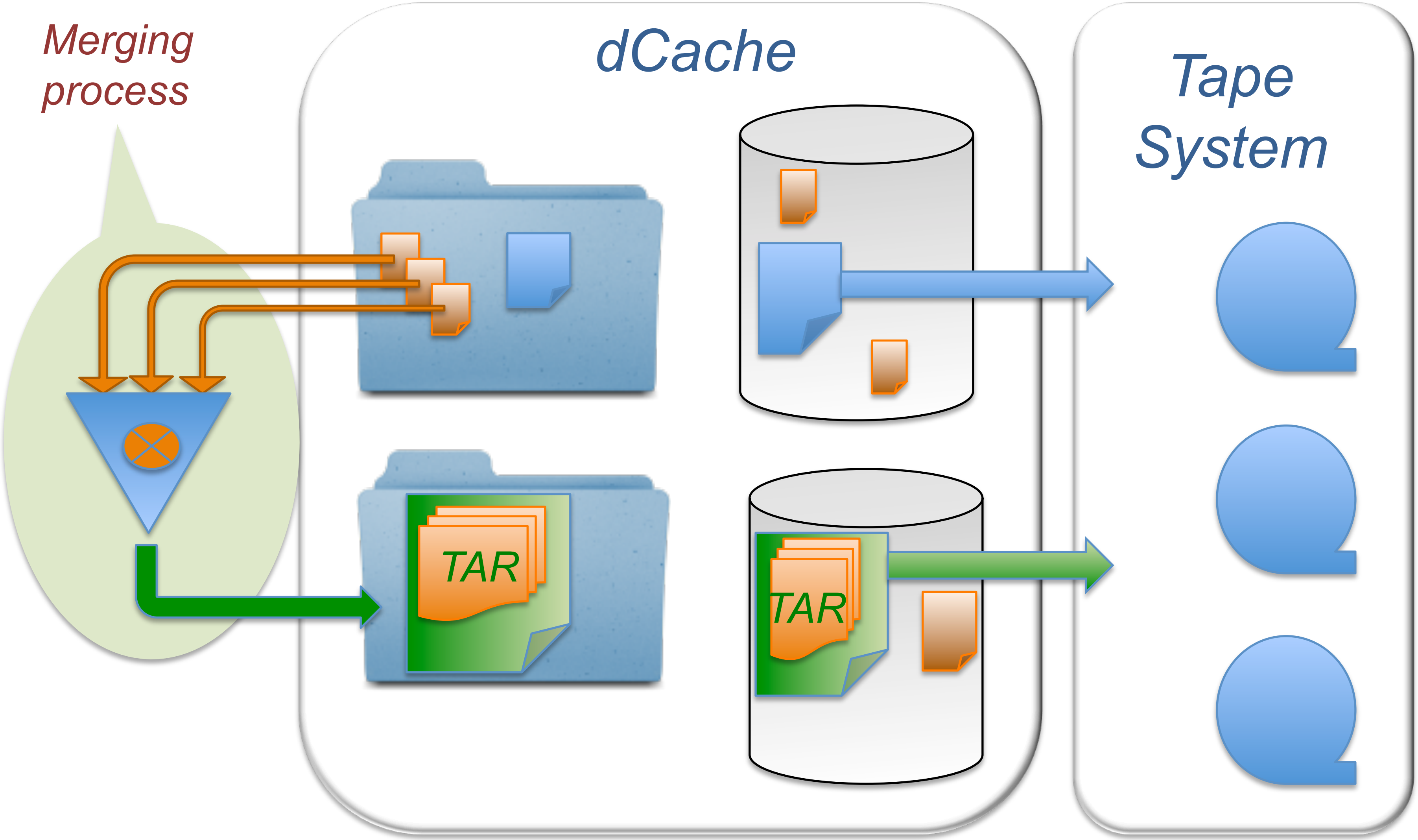
Or, Why do small files kill tape systems ?

- 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

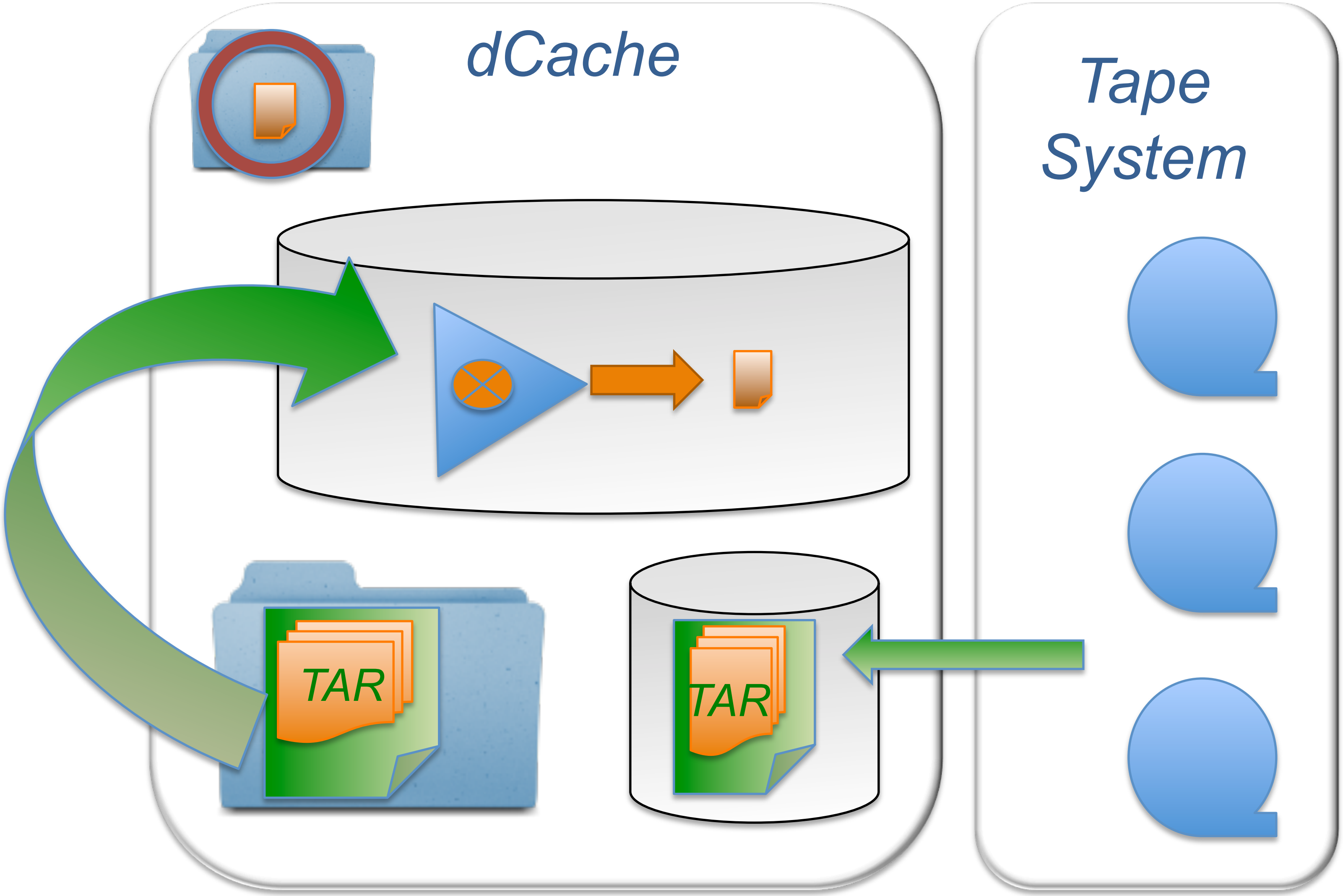
Our suggestion

- Decision on whether files are “large” or “small” will be initially based on directories.
- 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.
- “On top service” Runs on already supported dCache versions.

Merging small files

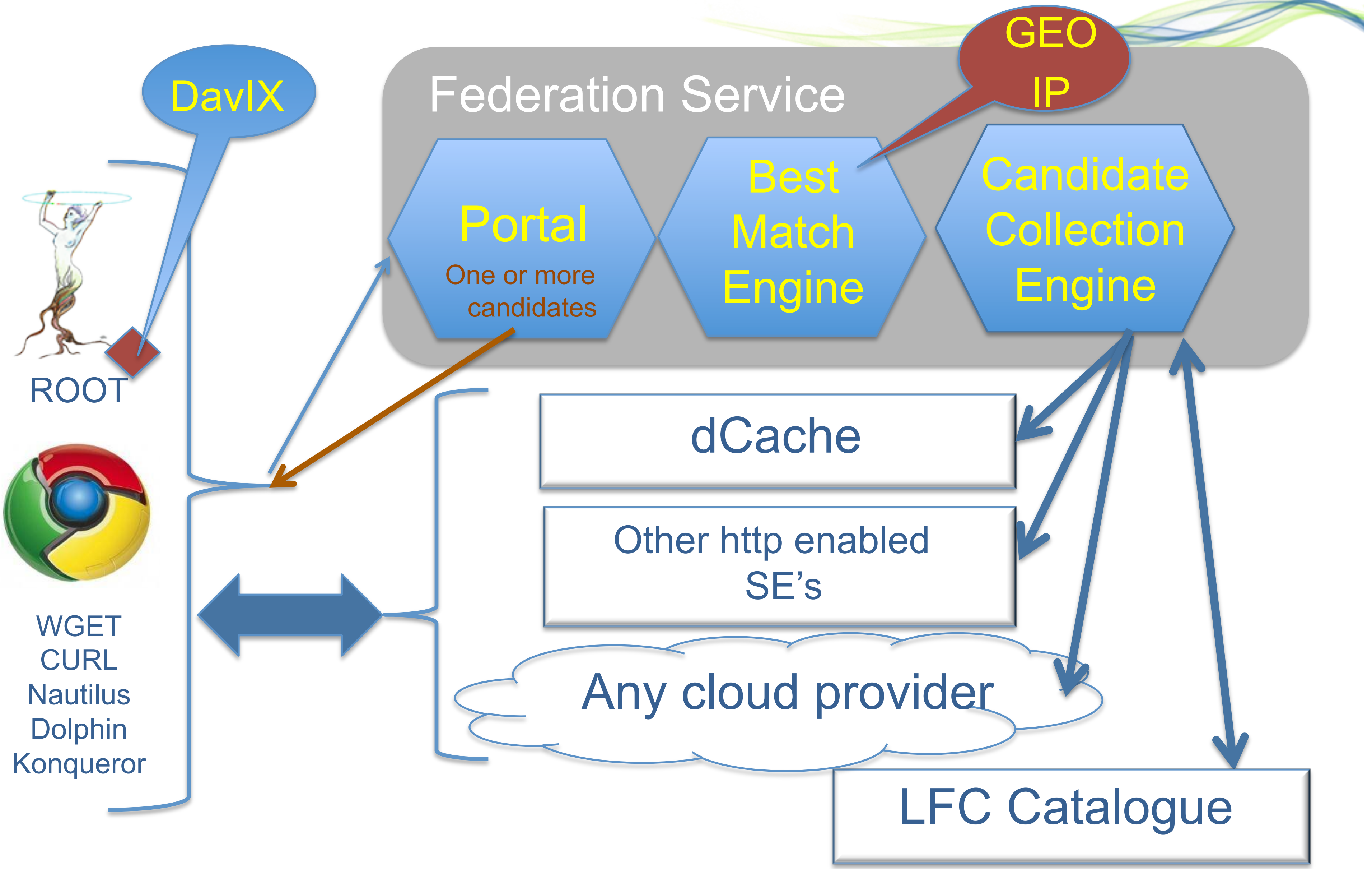


Extracting small file(s)



The Dynamic http/WebDAV federation

Dynamic Federation



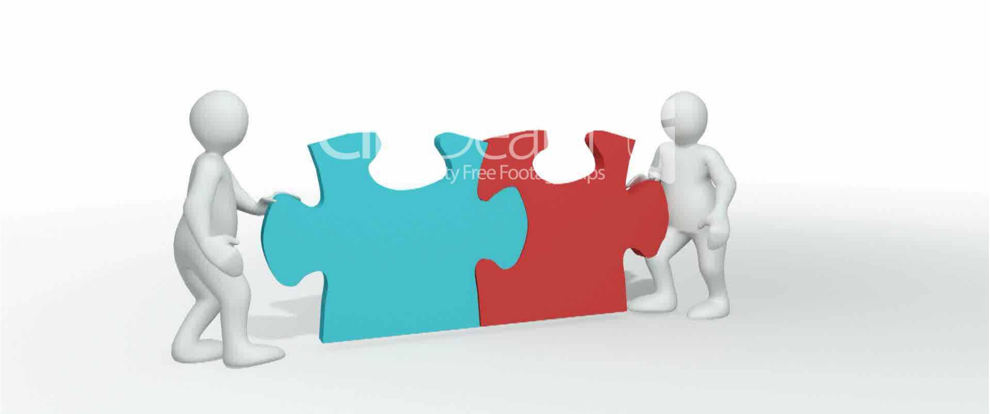
Single access performance

Single client performance

- Up to know, dCache focused on the optimization of overall performance
  - Transaction rates (stats)
  - Transfer speed
- Consequence:
  - Single client transaction time is high compared to high-end systems e.g. GPFS.
- With new requirements from new communities this needs some adjustment.
  - Tigran already started to profile meta-data transactions (open,…)
  - Already clear: Head-room for improvements
  - Work will continue, we'll keep you updated.

How does all this fits together ?

- Supporting individual identity management, remote IdP’s
- Allowing gPlazma to be integrated into the site infrastructure (Ron’s presentation)
- Supporting ‘small’ files for tape
- Supporting individual disk->tape transactions (CMS request)
- Improving single client transaction rate



How does all this fits together ?

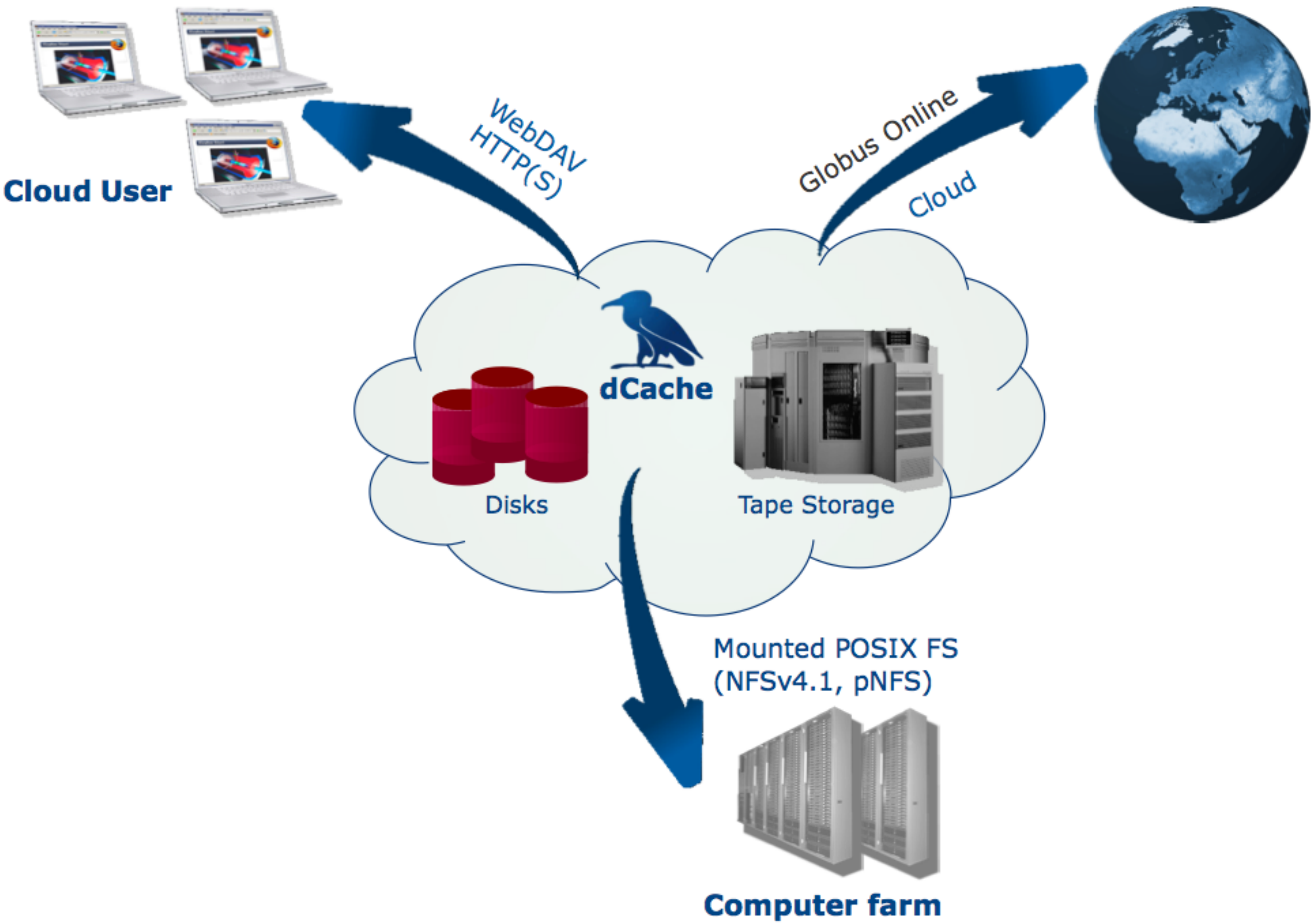
- We are working towards a individualized dCache.
- All supported protocols (WebDAV, nfs, …) will the same view of the repository.
- Various authentication mechanisms (Kerberos, X509, SAML) point to the same identity.
- Authorization is only based on the object (file directory) and the subject (user). -> Protocol independent.

Scientific Storage Cloud

Scientific Storage Cloud

- The same dCache instance can serve
  - Globus-online transfers via gridFTP
  - FTS Transfers for WLCG via gridFTP or WebDAV
  - Private upload and download via WebDAV
  - Public anonymous access via plain http(s)
  - Direct fast access from worker-nodes via NFS4.1
- The same user can use all those access mechanisms using a variety of credentials.
  - User/password
  - Kerberos
  - X509
  - SAML assertions

Scientific Storage Cloud



Questions

further reading

www.dCache.org