# EMI Data, preparing for the finish line

Patrick Fuhrmann, DESY/dCache.org
EMI Data Area Leader

*EGI Community Forum, Munich, Mar 26, 2012*

EMI INFSO-RI-261611

EMI is partially funded by the European Commission under Grant Agreement RI-261611

# Content

Introduction

Deployment Status

EMI Data Tasks
"Percentage done"

Tasks
- Done
- In Progress
- Cancelled

EMI Data Lib

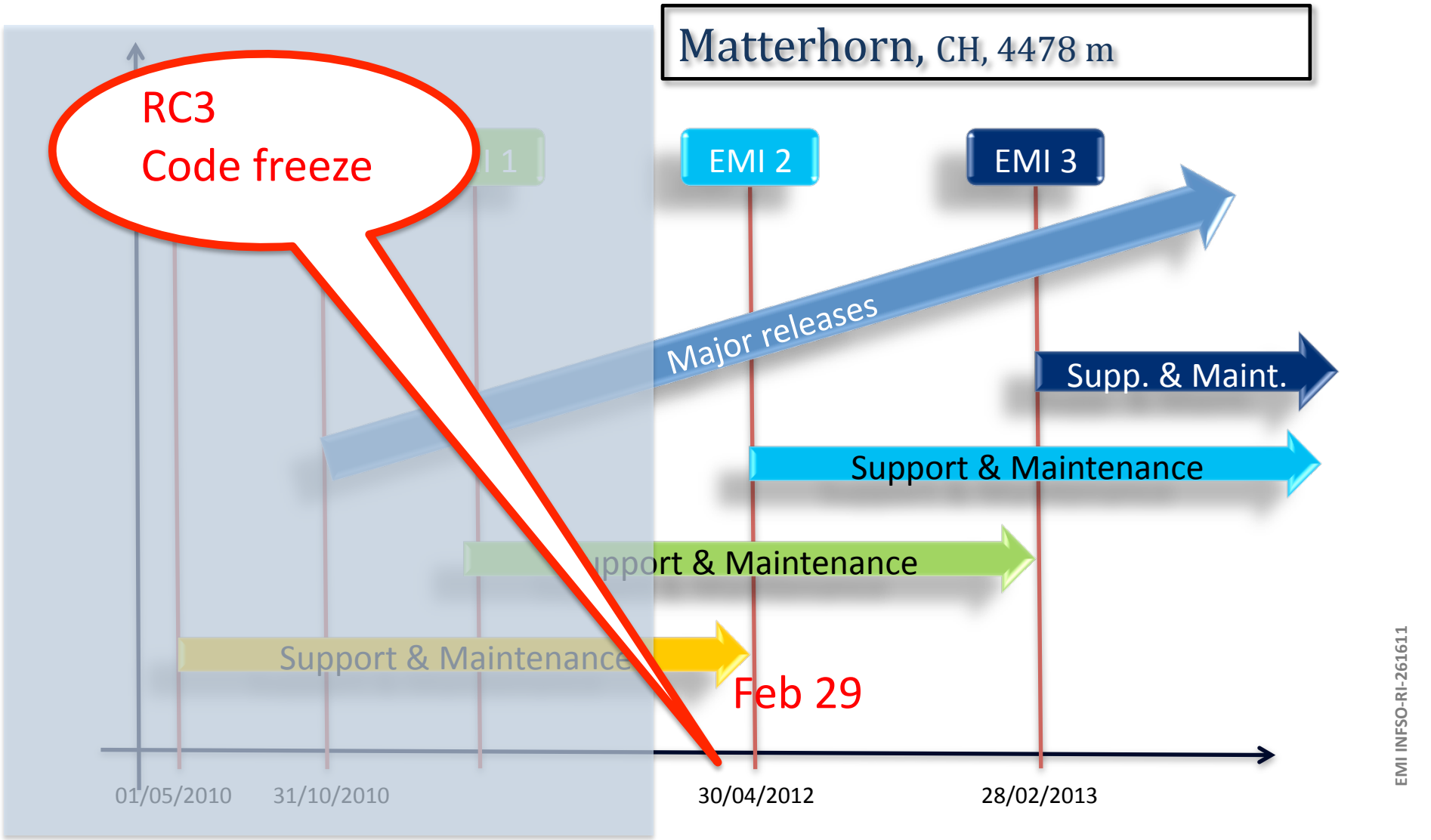
Storage Accounting

Catalogue Synchronization

Standards
- NFS 4.1 / pNFS
- WebDAV
- WebDAV for LFC/SE

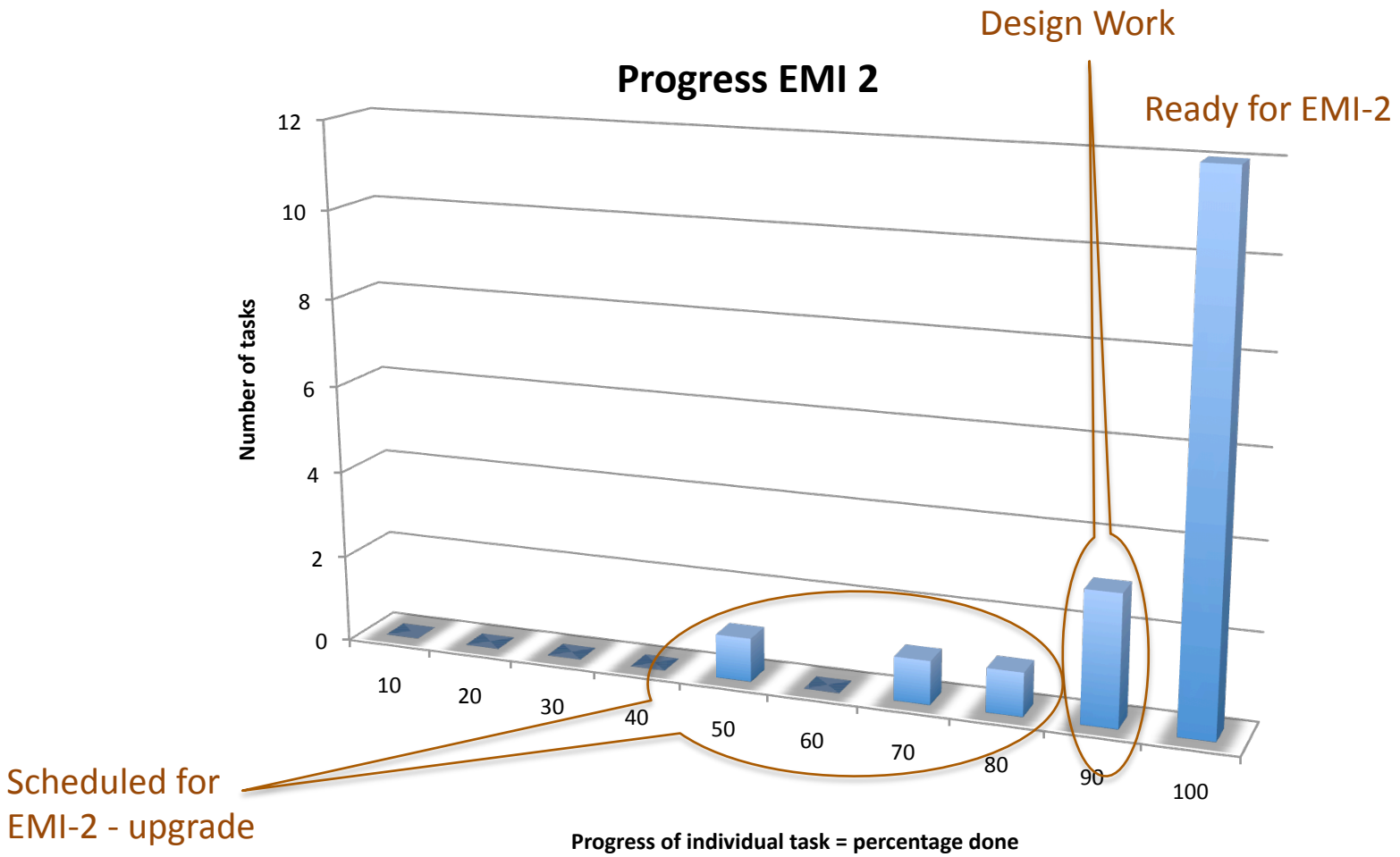
Something new

# Project status (Well known overview)




EMI INFSO-RI-261611

# Status of EMI 2 Data Tasks



# Some objectives/tasks needed to be cut

EUROPEAN MIDDLEWARE INITIATIVE

- ✓ Implementing cloud strategies
  - This might be worked on based on user requirements independently of EMI
- ✓ Persistent Data ID's (Evaluation)
- ✓ Deploying SRM with ssl/https instead of GSI
  - Could be shown to work
  - Delegation agreement will be used for WebDAV 3rd party copy (e.g. with FTS)
  - Watching the WLCG TE-Group(s)
- ✓ Permission synchronization (2nd part of cat-sync initiative)

# Objectives already achieved

- ✓ GLUE 2.0 in servers and clients
  - Publishing EMI version numbers
- ✓ ARGUS integration (Blacklisting) in SE's
- ✓ WebDAV for storage elements.
- ✓ WebDAV for LFC
- ✓ POSIX access to EMI storage elements.
  - NFS 4.1/pNFS for dCache and DPM
  - Native through GPFS/Lustre for StoRM
- ✓ Storage accounting record defined and introduced to OGF (working group)
- ✓ Agreement on delegation and introduction to OGF (working group)

# UNICORE Access To EMI Storage

By
Christian Löschen

See
Presentation
By
Christian
In this session

# The EMI Data Library

By
Jon Kerr Nielsen

The design

Status and Timeline

Beyond EMI

EMI INFSO-RI-261611

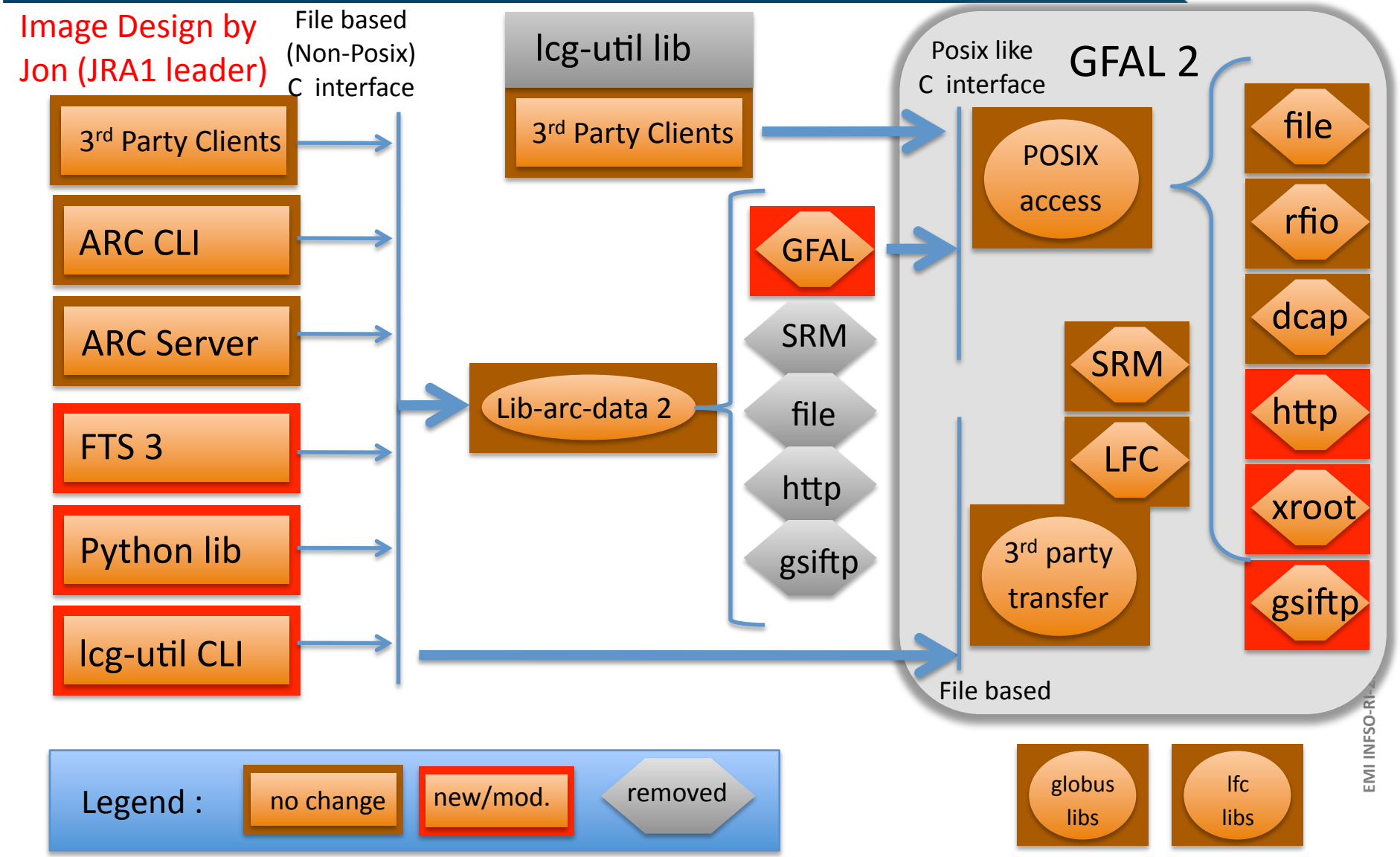
# The EMI Data Lib

EUROPEAN MIDDLEWARE INITIATIVE

## What and why

- ✓ Merging the GFAL (gLite) and libarcdata (ARC) data client libraries
- ✓ Reduces the number of components to support in the future
- ✓ Makes fixes easier, as there is only one library to take into account.

# Design of the EMI data lib



# EMI Data Lib, Status and Timeline

## Status and timeline

Stolen from Jon

- ✓ Design agreement within the EMI data group, waiting for PTB endorsement
- ✓ Main building blocks (libarcdata GFAL2 plugin, GFAL2) already implemented on prototype level
- ✓ Demo/test results to be shown at CHEP 2012 poster session May 2012
- ✓ Will be released in an EMI 2 update in June 2012 (not ready for EMI 2 release in April)
- ✓ Testing and bug fixing during fall 2012

# EMI Data Lib, Beyond EMI

Stolen from Jon

## Beyond EMI

- ✓ lcg_utils and ARC CLIs will still be there after EMI
- ✓ CERN data will continue to support GFAL2 and plug-ins after EMI (BTW: FTS3, the new file transfer service, is based on GFAL2)
- ✓ ARC will support the GFAL2 plug-in as long as it is used
- ✓ EMI_datalib will be supported by ARC and CERN-DM after EMI
  - *If it works as good as or better than current solution*
  - *Until some better solution appears ☺*

# The Storage Accounting Record

By
Jon Kerr Nielsen

Timeline

Current Status

Beyond EMI

# The Storage Accounting Record

Stolen from Jon

## Timeline

- ✓ Design agreed within EMI June 2011
- ✓ Submitted for public hearing within OGF February 2012
  - *Informational document as input to UR 2.0*
  - *Open for comments until OGF34 (Oxford, beginning of March)*
- ✓ Implementing accounting sensors in the EMI storage elements due in May 2012 for EMI 2 update
  - *dCache NDGF already use StAR in production using SGAS*
  - *StoRM will implement in March-April 2012*
  - *Implementation progress will be discussed in EGI Accounting session in EGI-CF/EMI-TF in Munich end of March 2012*
- ✓ Accounting publishers (APEL) to publish storage records in June 2012
- ✓ Storage elements to test and deploy accounting sensors for EMI 3 Monte Bianco RC1 December 2012
- ✓ Testing and bugfixing in EMI 3 RCs January-April 2013
- ✓ Storage elements publishing StAR records released in April 2013

# The Storage Accounting Record

EUROPEAN MIDDLEWARE INITIATIVE

Stolen from Jon

## ❑Current status

- ✓ Not planned for EMI 2 release
- ✓ Status should be clearer after EGI session in Munich (Thursday)
- ✓ Next milestone May 2012 – accounting sensors
  - *Still seems realistic*

## ❑Beyond EMI

- ✓ WLCG TEG sees StAR as the most realistic approach to storage accounting
- ✓ Clear interest from OSG
- ✓ StAR is taken as input to storage part of next generation OGF UR
- ✓ Sustainability through standardization and wide adoption

# Catalogue Synchronization

By
Fabrizio

The problem

What can we solve

How do we solve it

# The Problem

- Various catalogues keep information that is related
  - E.g. LFC keeps info about the content of remote Storage Elements, each one with its own catalogue
    - Dangling References : If a SE looses a file unnoticed by the LFC
    - Dark Data : If a new file is not correctly registered -> dark data
    - ACL Synchronization : A change in the permissions of a file in LFC is not automatically reflected by the peripheral catalogue
- Keeping them in sync is a very hard problem
  - See presentation “Consistency between grid storage elements and file catalogue for the LHCb experiment” by Elisa.
- Namespace scanning for ‘diffs’ is an expensive workaround

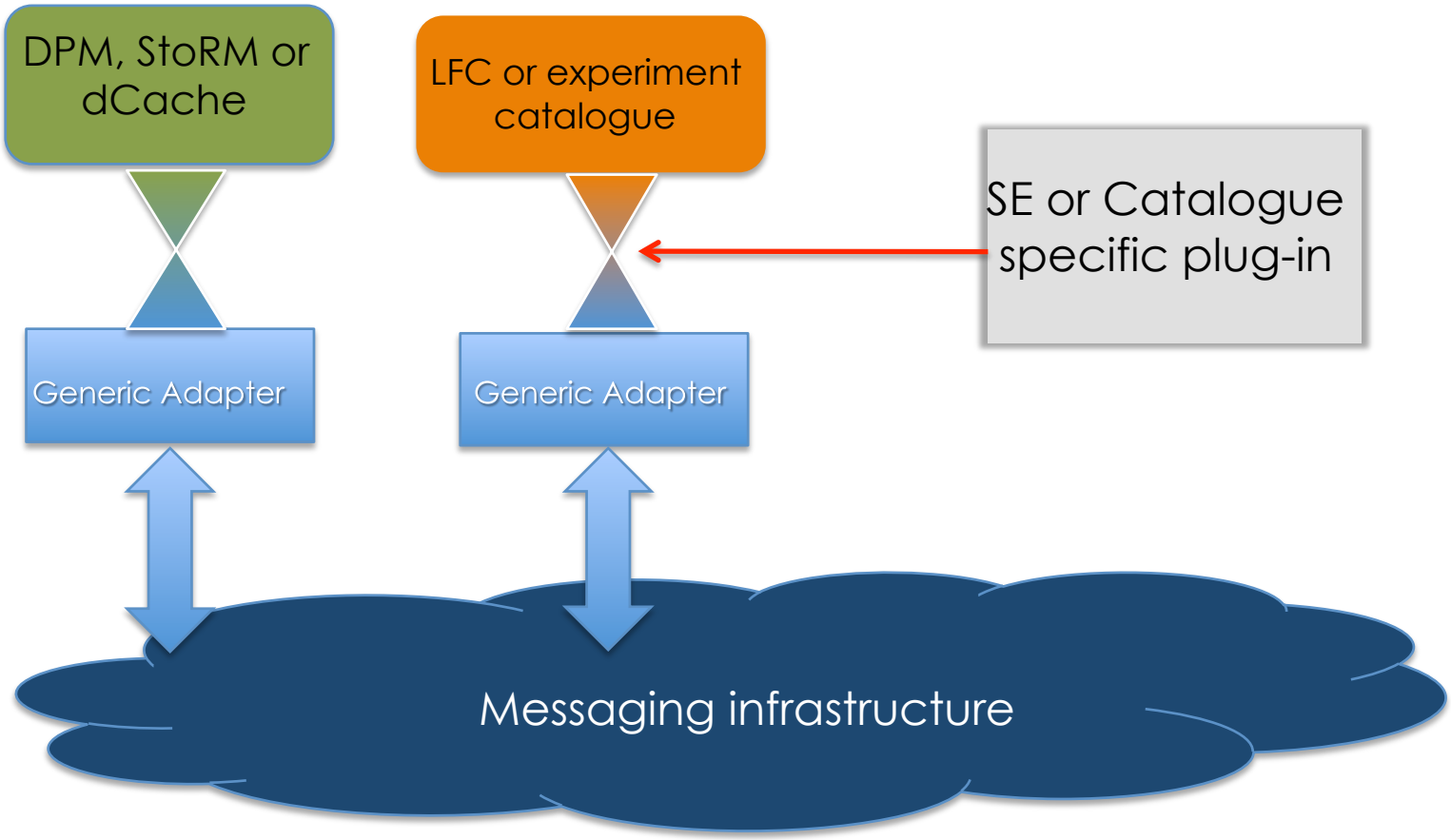
# What can we solve

Stolen from Fabrizio

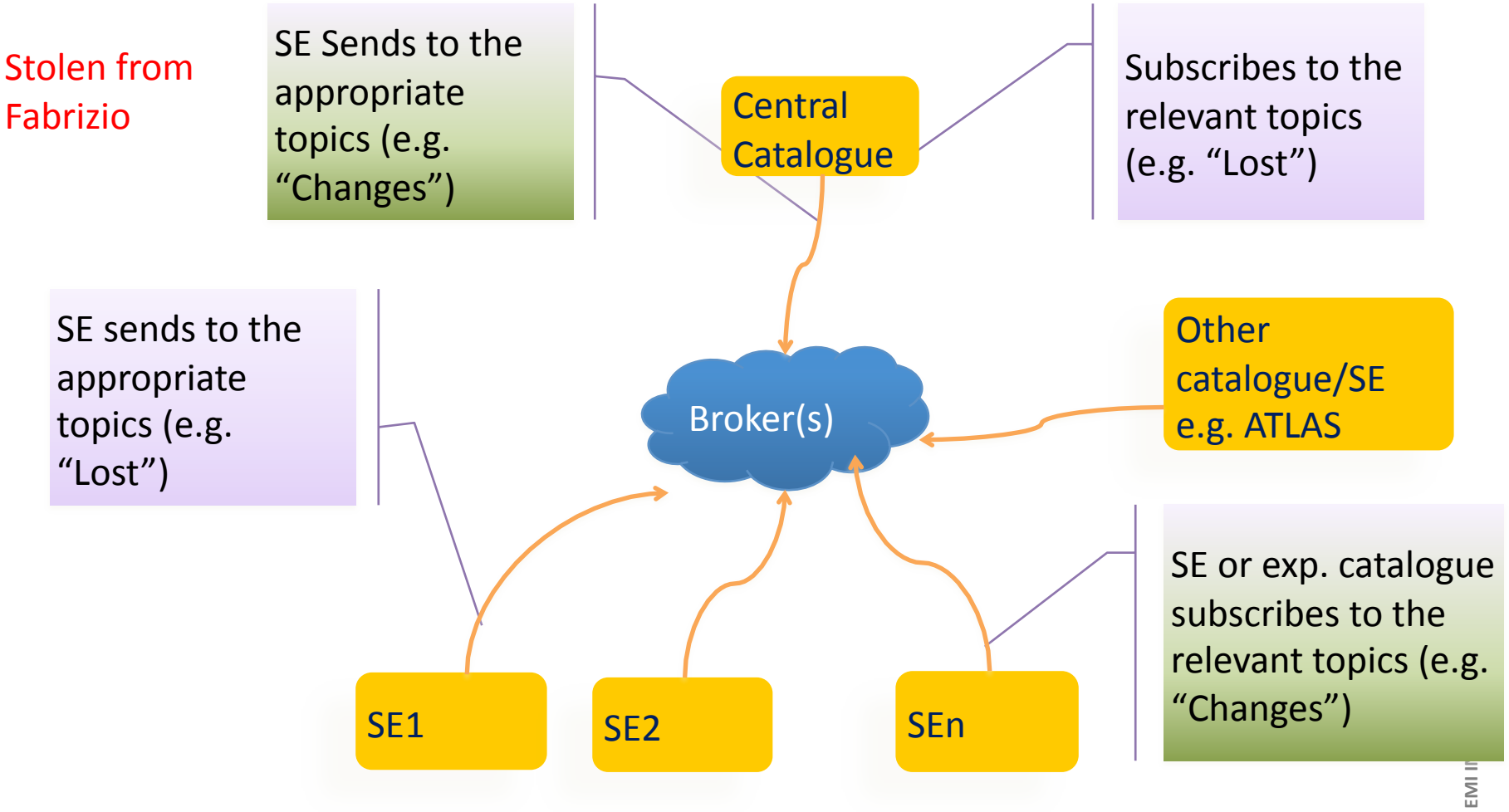
Make the various catalogues/SE able to talk to each other

- ✓ In order to exchange messages that keep them synchronized in real-time
- ✓ Two problems fixed :
  - Central Catalogue->SE (downstream)
    - e.g. to propagate changes in the permissions
  - SE->Central Catalogue (upstream)
    - e.g. to propagate info about lost and missing files
- ✓ No fix for : dark data

# How is it solved



# How is it solved



# Catalogue synchronization

## Status

- ✓ Starting with “File lost” message from SEs
- ✓ DPM: Publishing ‘file lost’ is ready.
- ✓ LFC: Sets the reported entry to ‘temp. unavailable’.
- ✓ Prototype for one experiment catalogue, but waits for dCache to be ready as well.
- ✓ Code available in dCache for WebDAV door, but still waiting go be merged with EMI-2 branch. Other doors following soon.
- ✓ Permission synchronization is on hold until sufficient interest has been demonstrated by customers.

# Standard Protocols EMI

## NFS 4.1 /pNFS

By
CERN-IT-GT
dCache.org

# Standard protocols : NFS 4.1 / pNFS

EUROPEAN MIDDLEWARE INITIATIVE

- Reminder NFS 4.1 / pNFS
  - ✓ Industry Standard
  - ✓ Allows direct connection between client and the data source for distributed storage systems (First open NFS providing this)
  - ✓ Provides build-in security (part of the spec. not on top)
  - ✓ Mounts into your file system as easy as your memory stick
  - ✓ Data clients are provided by the OS providers similar to xfs/ext3/…
  - ✓ Allows to prevent vendor locks as the storage system can be easily expanded to a heterogeneous setup w/o changing the client nodes setup.
  - ✓ It's really cool

# Standard protocols : NFS 4.1 / pNFS

EUROPEAN MIDDLEWARE INITIATIVE

## ❑ Current status

- ✓ DPM and dCache servers are ready to serve data with NFS 4.1 / pNFS
- ✓ Vendors now start to provide ‘test’ NFS4.1/pNFS machines to ‘friends’.
  - ✓ E.g. NetApp
- ✓ Authentication : Kerberos included (client and server)
- ✓ Authentication : X509 : some attempts made but still evaluating
- ✓ Clients ( = linux kernel module) are available by now

| | |
|---|---|
| SL 6 | Kernel 2.6.32 – 220+ |
| SL C 6 | ? |
| Fedora 16 | Kernel 3.2.x |
| Currently in ebian unstable, Will be in “Wheezy” | Kernel 3.2.0-1 |
| And more. E.g. ORACLE Unbreakable Enterprise K. | 3.0.16 reads as 2.6.39 |

# Standard protocols : NFS 4.1 / pNFS

EUROPEAN MIDDLEWARE INITIATIVE

- NFS deployed ….

  - ✓ dCache NFS 4.1 already in production at DESY for Photon Science for about a year.
  - ✓ dCache NFS 4.1 evaluated at FERMIlab by “Running Experiments Department, Grid Support Group” for their “Fermilab Intensity Frontier experiments” customer.
  - ✓ DPM, NFS 4.1 evaluation cluster in Taipei
  - ✓ NFS 4.1/pNFS is a done deal. Nice success story.
  - ✓ New communities are smart enough to start evaluation/production now.
  - ✓ Wide area
    - Dima (DESY) mounted a dCache/NFS4.1 system, located in Taipei at A. S during ISGC., from an NFS client at DESY, copied a binary into it and executed it. Worked like a charm. However: found problems with WAN.
    - In general we don’t have enough experience yet with NFS/pNFS WAN access. That still needs to be evaluated.

# Standard protocols : http / WebDAV

EUROPEAN MIDDLEWARE INITIATIVE

## Standard Protocols EMI

## WebDav

By
CERN-IT-GT
dCache.org

webdav.dcache.org
Connected as: WebDAV
Disconnect

EMI INFSO-RI-261611

# Standard protocols : http / WebDAV

## That's what you already know :

- ITEF Standard
- Everybody, using the internet, has a http/WebDAV client at his/her fingertips
- Allows "File system like" access with
  - Mac OS
  - Linux
  - Windows
- Either supported by OS or Browsers.
- Authentication : x509 Certificates, User/Password

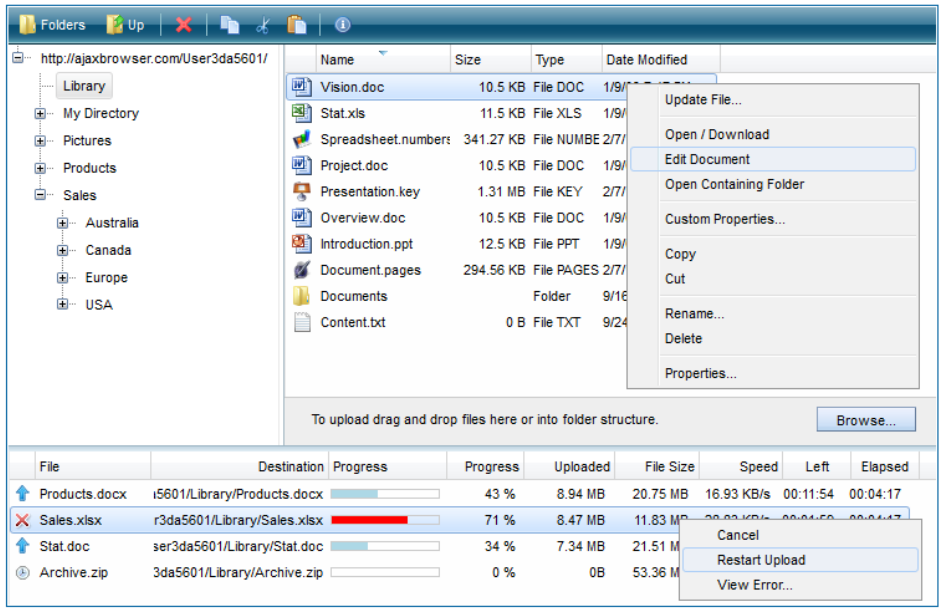
# Standard protocols : http / WebDAV

## More important :

- Ready for DPM and dCache in EMI-2.
- StoRM will follow soon.
- Experiments are now seriously considering to use http(s)/WebDAV for data access and transfers.
  - See presentation “DDM Site Services: A Solution for Global Replication of HEP Data” by Fernando.
- WebDAV is in the development plan for FTS.

# Standard protocols : http / WebDAV

- In addition, dCache provides a 3rd party drag&drop javascript interface for browsers.
- Another proof that standards allow easy integration of 3rd party software



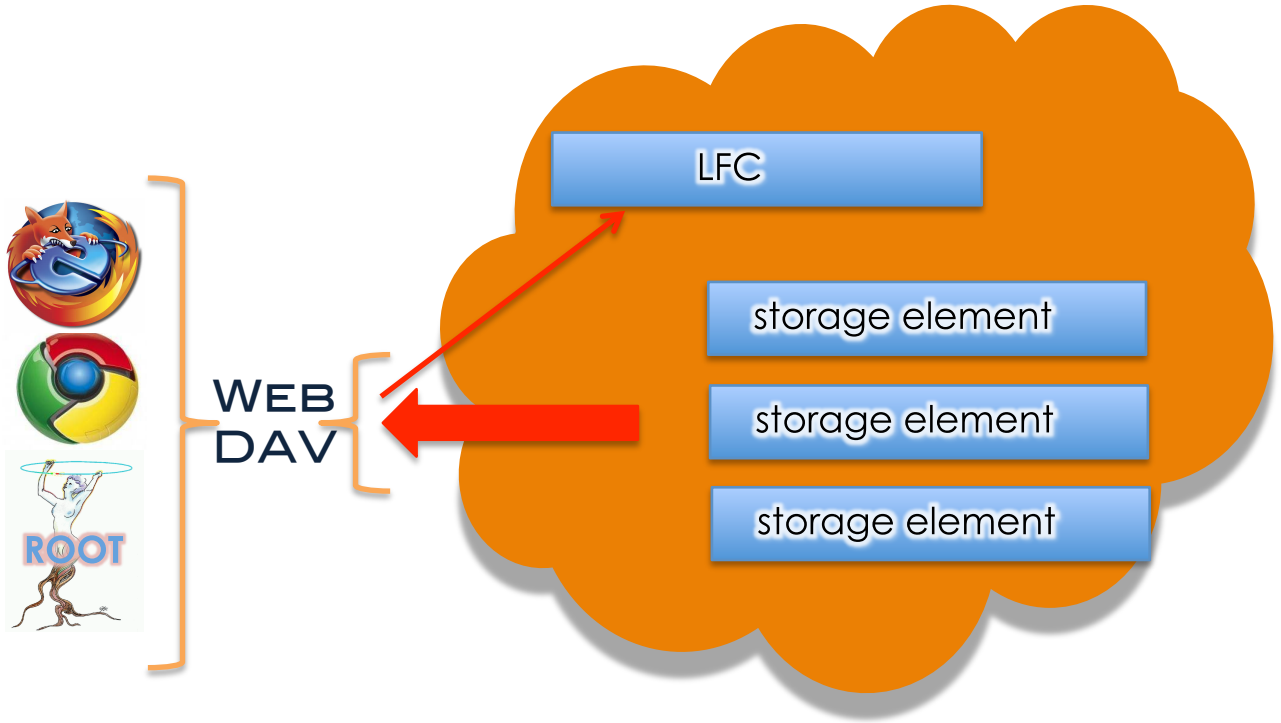

EMI INFSO-RI-261611

# Common WebDAV Frontend to LFC and SE's

By
CERN-IT-GT
dCache.org
(StoRM)

# Common WedDAV frontend for LFC/SE

- Goal
  - ✓ Provide transparent access to data through catalogues, using standard protocols : http(s), WebDAV
  - ✓ Redirection from catalogues to the final data source doesn't require intermediate steps by the user but is part of the protocol.



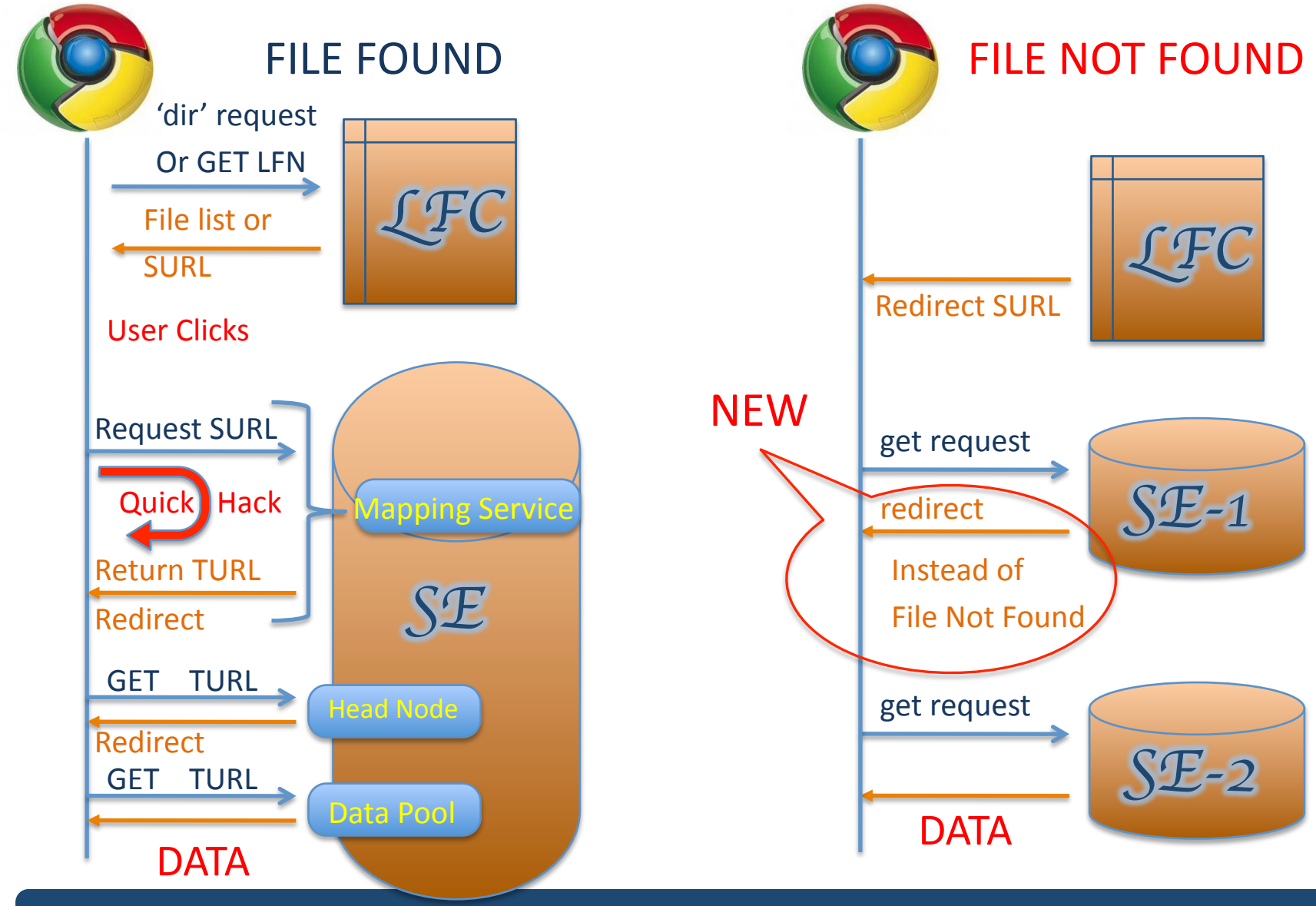
# WedDAV frontend: Progress

EUROPEAN MIDDLEWARE INITIATIVE

- Progress

  - ✓ First functional prototype introduced by CERN-IT-GT for the 'EMI all hands meeting' in Padova, Oct 2011, using LFC and DPM.
  - ✓ Semi-Final design document provided by Ricardo, circulated and improved, circulated and improved … , circulated and approved.
  - ✓ dCache developer worked at CERN for 6 weeks to integrate the design to dCache.
  - ✓ Unfortunate issue: As LFC contains SRM-SURLS, some implicit assumptions need to be made to translate to SE-TURL, resp. to find the WebDAV endpoint.
  - ✓ Proposed solution: SRM-light SURL->TURL mapping service (pure http)

# WedDAV frontend: Some Insight



# New Objective for EMI Y 3

**The Dynamic Federation Project**

**By**
**CERN-IT-GT**
**dCache.org**
**Slides by Fabrizio**

EUROPEAN MIDDLEWARE INITIATIVE

# The Idea

Accessing federated/replicated data with a standard protocol (http/WebDAV), using a single (possibly replicated) endpoint.

Add-on: Best replica is picked by algorithms considering

1. Geo Data resp. network topology
2. Availability and/or performance of SE endpoint
3. Configuration

# The Dynamic Federation Project

Stolen from Fabrizio

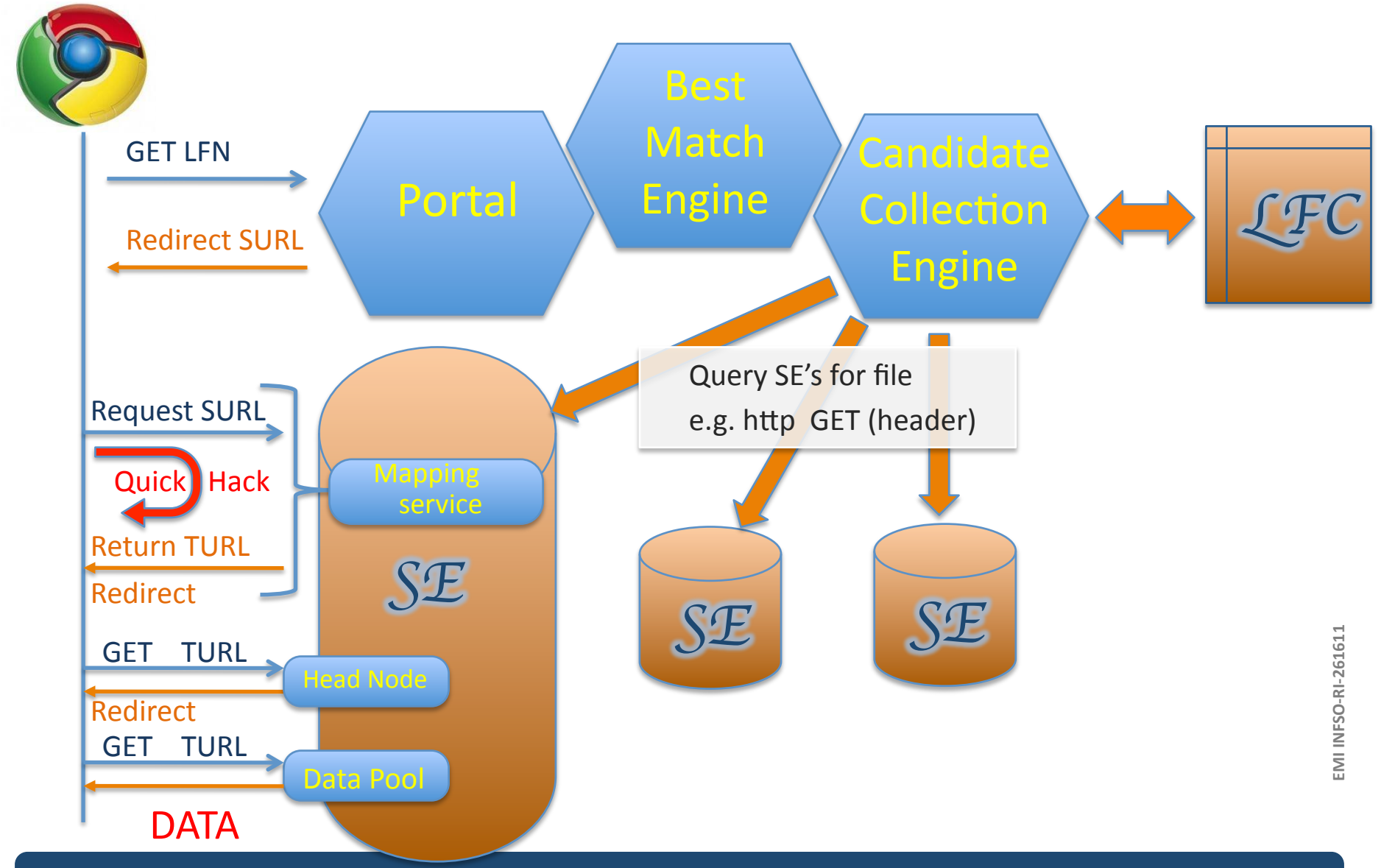
- Technically “loosely coupled storage systems”
- Idea: a single entry point for a federation of endpoints
  - single storage elements (e.g. dCache, DPM, plain HTTP servers)
  - site/VO catalogues (e.g. LFCs) pointing to storage elements
- This entry point knows its endpoints
- An approach with many interesting possibilities
  - Federate third party outsourced HTTP/DAV servers (also clouds) Federate the content of SQUID caches
  - Federate them together with the information of some experiment’s DB
    - When clicking on a file we would download it from an endpoint that is good for us, it could be a cache or a non-cache one
  - See as one experiment’s DBs (e.g. LFC), also considering what’s in the SQUID caches worldwide
    - Direct access to the official replicas AND the cached ones as well

# The Dynamic Federation Project

Stolen from Fabrizio

- The endpoints are a federation, hence they are homogeneous
  - ✓ Same access protocol (e.g.HTTP/WebDAV)
  - ✓ Same name space (file content consistency problem)
  - ✓ The same file / replica has the same (or compatible) path/name (mapping problem)
  - ✓ They grant access to the same groups of users (permission problem)
- This entry point learns dynamically, automatically about their metadata content
  - ✓ As clients contact it to get access to files
  - ✓ It can ask the endpoints for information on the fly
- This entry point redirects each client to the proper endpoint
  - ✓ Eventually applying some smart criteria, e.g. proximity
- In principle it would work for any data access protocol that
  - ✓ works over WAN
  - ✓ supports redirections
- Our focus is towards HTTP/WebDAV for now
  - ✓ DPM and dCache are releasing support for it
- Work in progress, priority is read access
  - ✓ As, in general, write access is done in the local site

# Oversimplified Picture



# FTS 3 Next Generation File - Transfer Service

By
CERN-IT-GT
Zsolt Molnár

EMI INFSO-RI-261611

# FTS 3 (next generation file transfer)

EUROPEAN MIDDLEWARE INITIATIVE

## FTS 3 Demo 1

- Rewritten C++ CLI, on top of WS-I compatible WSDL
- Backward compatibility: you can submit to FTS2 servers as well
- FTS server daemon, in C++ (Java removed). Capabilities:
- Working multithreaded C++ web server and FTS agent integrated, capable of handling submit/status commands
- Host config part of "https://svnweb.cern.ch/trac/fts3/wiki/Configuration" implemented
- Store/retrieve job data in Oracle database, using generic database interface and Oracle plug-in

# Summary and outlook

- EMI Data is well on track in terms of the expected tasks to be delivered for EMI-2.
- New requests from TCB are being discussed
  - ✓ *Some of which make sense*
  - ✓ *Others are too challenging for the remaining EMI project time*
- EMI-Data beyond EMI
  - ✓ *EMI Data PT are 'stable'*
    - They existed before EMI
    - Their products are in heavy use and the funding is guarantied for the foreseeable future
    - They will continue to exist after EMI funding ends.
  - ✓ *EMI Data PT's have a long history in collaborating (SRM,GLUE…)*
  - ✓ *Avoiding unnecessary PT interactions by consistently using standards.*


# Thank you

EMI INFSO-RI-261611

**EMI is partially funded by the European Commission under Grant Agreement INFSO-RI-261611**