EMI Data, the last year

Patrick Fuhrmann, data area lead, for the EMI data team.

Taipei, TW, Feb 30, 2012
Content

Introduction

Deployment Status

EMI Data Tasks
“Percentage done”

EMI Data Lib

Storage Accounting

Catalogue Synchronization

Standards
  NFS 4.1 / pNFS
  WebDAV
  WebDAV for LFC/SE

Something TOP SECRET
Project status (Well known overview)

RC3 Code freeze

Matterhorn, CH, 4478 m

EMI 1

EMI 2

EMI 3

Support & Maintenance

Major releases

Supp. & Maint.

Support & Maintenance

Support & Maintenance

Feb 29

01/05/2010 31/10/2010 30/04/2012 28/02/2013
Status of EMI 2 Data Tasks

Progress EMI 2

- Design Work
- Ready for EMI-2

Number of tasks

- Progress of individual task = percentage done

Scheduled for EMI-2 - upgrade

EMI, the second year
The following topics are selected out of a variety of tasks in EMI-Data, with the goal to illustrate advantages of collaborated development work, which is made possible through EMI.
The EMI Data Library

By Jon Kerr Nielsen

The design
Status and Timeline
Beyond EMI
The EMI Data Lib

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

Image Design by Jon (JRA1 leader)

File based (Non-Posix) C interface

3rd Party Clients
ARC CLI
ARC Server
FTS 3
Python lib
lcg-util CLI

Legend:
- no change
- new/mod.
- removed

Posix like C interface

GFAL 2

POSIX access
SRM
file
http
gsiftp

File based

3rd party transfer

SRM
LFC
xroot

POSIX
file
rfio
dcap
http

3rd Party Clients

Lib-arc-data 2

LCG
lib

ARC CLI

Python lib

3rd Party Clients

Legend:
- no change
- new/mod.
- removed

2/30/12
EMI, the second year
Status and timeline

✓ 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
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 data 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

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

Current status

- Not planned for EMI 2 release
- Status should be clearer after EGI session in Munich
- 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 loses 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
- Namespace scanning for ‘diffs’ is an expensive workaround
What can we solve

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

DPM, StoRM or dCache

LFC or experiment catalogue

Generic Adapter

SE or Catalogue specific plug-in

Generic Adapter

Messaging infrastructure

Generic Adapter

DPM, StoRM or dCache

LFC or experiment catalogue

SE or Catalogue specific plug-in

Generic Adapter

Messaging infrastructure

Generic Adapter

DPM, StoRM or dCache

LFC or experiment catalogue

SE or Catalogue specific plug-in

Generic Adapter

Messaging infrastructure

Generic Adapter
How is it solved

Stolen from Fabrizio

SE Sends to the appropriate topics (e.g. “Changes”)

Central Catalogue

Subscribes to the relevant topics (e.g. “Lost”)

SE sends to the appropriate topics (e.g. “Lost”)

SE Sends to the appropriate topics (e.g. “Changes”)

SE1

SE2

SEn

Other catalogue/SE e.g. ATLAS

SE or exp. catalogue subscribes to the relevant topics (e.g. “Changes”)
Catalogue synchronization

Status

- Starting with “File lost” message from SEs
- DPM: Publishing ‘file lost’ 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.
Standard Protocols
EMI

NFS 4.1 /pNFS
By
CERN-IT-GT
dCache.org
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
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’.
- Authentication : Kerberos included (client and server)
- Authentication : X509 : some attempts made but still evaluating
- Clients ( = linux kernel module) are available by now

<table>
<thead>
<tr>
<th></th>
<th>Kernel version</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL 6</td>
<td>2.6.32 – 220+</td>
</tr>
<tr>
<td>SL C 6</td>
<td>?</td>
</tr>
<tr>
<td>Fedora 16</td>
<td>3.2.x</td>
</tr>
<tr>
<td>Currently in ebian unstable, Will be in “Wheezy”</td>
<td>3.2.0-1</td>
</tr>
<tr>
<td>And more</td>
<td></td>
</tr>
</tbody>
</table>
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 here in Taipei
- NFS 4.1/pNFS is a done deal. Nice success story.
- New communities are smart enough to start evaluation/production now.
- Just for fun: Yesterday we (Dima) mounted a dCache/NFS4.1 system, located here at A. S., from an NFS client at DESY, copied a binary into it and executed it. Worked like a charm.
- However, in general we don’t have enough experience yet with NFS/pNFS WAN access. That still needs to be evaluated.
Standard protocols: http / WebDAV

Standard Protocols
EMI

WebDAV
By
CERN-IT-GT
dCache.org
Standard protocols: http / WebDAV

- 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
- Ready for DPM and dCache and ready in EMI-2.
- StoRM will follow soon.

2/30/12  EMI, the second year
• In addition, dCache provides a 3\textsuperscript{rd} party drag\&drop javascript interface for browsers.
• Another proof that standards allow easy integration of 3\textsuperscript{rd} party software
Common WebDAV Frontend to LFC and SE’s

By CERN-IT-GT
dCache.org (StoRM)
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.
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 now 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)
FILE FOUND

‘dir’ request
Or GET LFN
File list or
SURL
User Clicks

Request
SURL
Quick Hack
Return TURL
Redirect
GET TURL
Redirect

DATA

FILE NOT FOUND

Redirect SURL

get request
redirect
Instead of File Not Found

NEW

DATA
The Dynamic Federation Project
by CERN-IT-GT
dCache.org
Slides by Fabrizio

For your eyes only
The Dynamic Federation Project

- 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
- A new thing, 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

Stolen from Fabrizio
The Dynamic Federation Project

- 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

You remember this picture?

You remember this picture?
Coming back to the question: **How do we benefit from EMI**

The work on the previous introduced topics, which requires a good understanding of the entire system, is possible, because EMI provides funding, but at least as important: Tt provides the infrastructure framework, allowing different PTs to work on a common goal.
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 is using ‘friendly’ competition between the different product teams for the benefit for the customer by strictly sticking with standard interfaces and protocols.

- For the time 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 is partially funded by the European Commission under Grant Agreement INFSO-RI-261611
EMI Data Lib

- lcg_util_lib
- arc cli
- arc server
- python lib
- 3rd party clients
- fts3

Legend:
- no change
- new/modify
- remove

EMI, the second year
FTS 3
Next Generation File - Transfer Service

By
CERN-IT-GT
Zsolt Molnár
FTS 3 (next generation file transfer)

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