D-Grid HEP CG
Work Package 1
Data-Management

Final Results

Patrick Fuhrmann

The Team

stephan freitag
david melkumyan
dirk pleiter
martin radicke
lars schley
owen synge

Gefördert vom

Bundesministerium
für Bildung
und Forschung

Reminder

Sub-work-packages :

The Extensible Metadata Catalogue

dCache, The scalable Storage Element

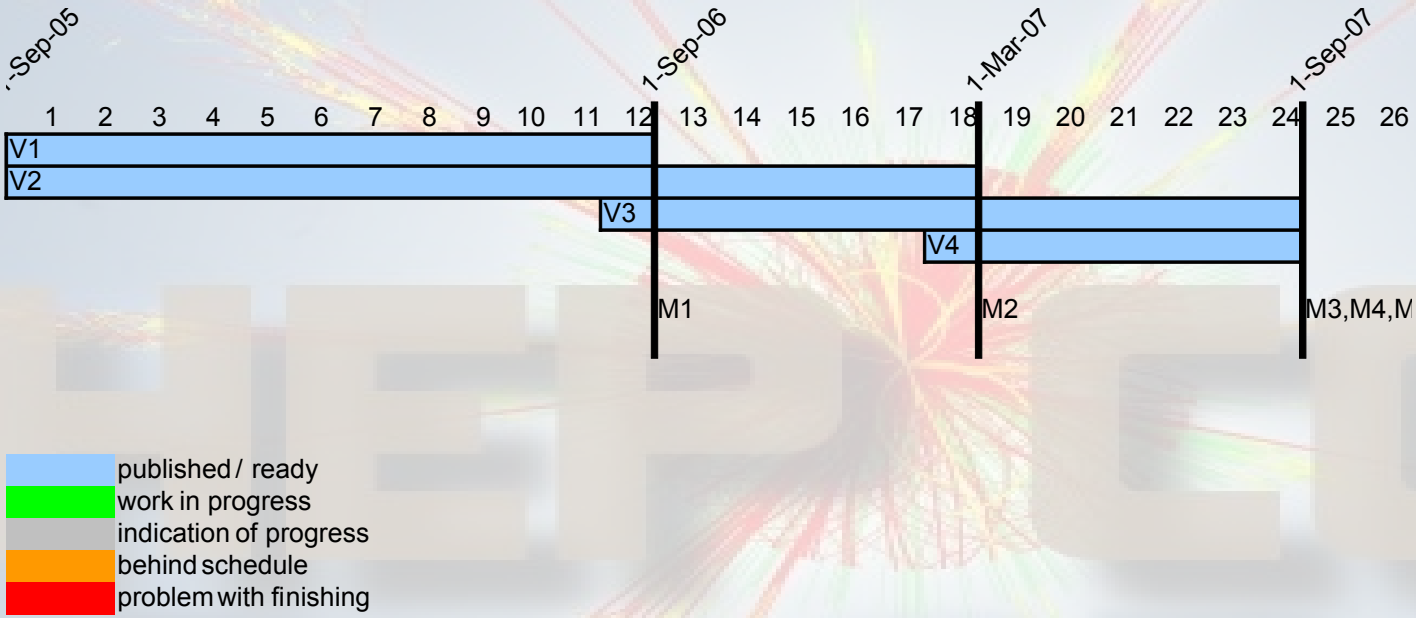
Co-scheduling / Smart Job Scheduling

Extensible Metadata Catalogue

Dirk Pleiter
David Melkumyan

Extensible Metadata Catalogue

AP1: Arbeitspaket 1 – Datenmanagement:Entwicklung eines erweiterbarenMetadaten-Katalogesfür semantischen Dater



Extensible Metadata Catalogue

V1: Erstellung einer Markup-Beschreibungphysikalischer Daten (XML-Schema)
Aufwand: 12 PM
Zeitraum: M1-M12

V2: Entwicklung eines Metadaten-Katalogs
- Entwicklung einer SQL-basierten Schnittstelle zu relationalen Datenbanken
- Auswahl und Test von Software zur automatisierten Abbildung von XML-Schemata auf SQL- Schemata
- Erstellen von Software für das Einfügen, Modifizieren und Lesen der in relationale Datenbanken gespeicherten XML-Dokumenten
- Definition einer Schnittstelle für Benutzer-Operationen
- Einbettung der Daten in ein dCache basiertes Storage Element, Schaffung einer Schnittstelle zwischen den Metadaten-Katalogen und den LCG-Replikakatalogen
- Implementieren von Web-Services für den Zugriff auf den Metadaten-Katalog

Aufwand: 36 PM
Zeitraum: M1-M18

V3: Entwicklung des Prototyps eines Benutzer-Interfaces
- Fertigstellung eines einfachen command line Interfaces
- Testphase mit realen Daten

Aufwand: 12 PM
Zeitraum: M12-M24

V4: Entwicklung eines Access Control Services.
- Spezifikation der Schnittstelle zum Replika-Katalog
- Implementierung des Service innerhalb des Metadaten-Katalogs

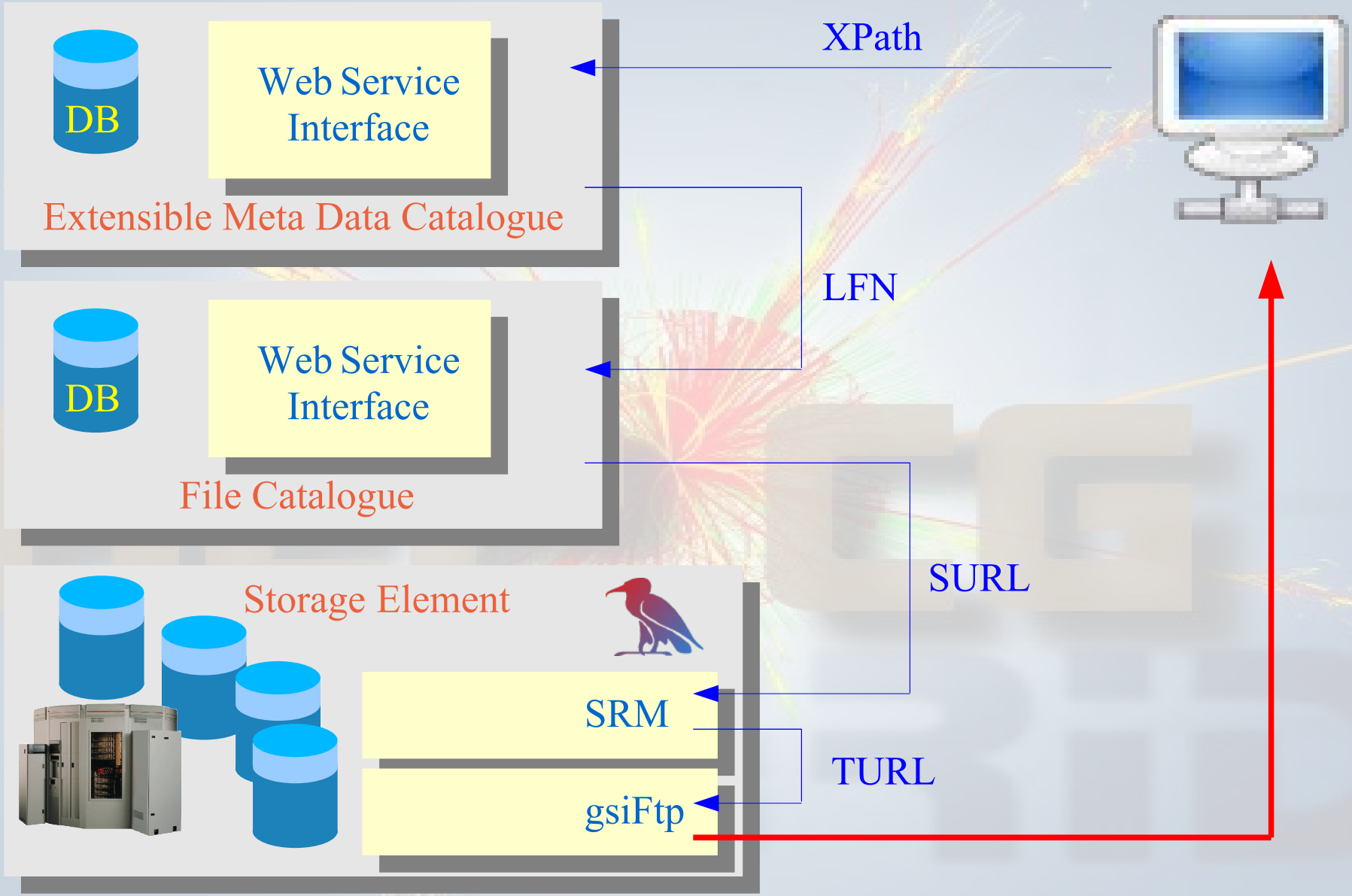
Aufwand: 9 PM
Zeitraum: M18-M24

Extensible Metadata Catalogue

Goals of the International Lattice Data Grid

- Establish an international data-grid infrastructure
- Long-term storage and global sharing of data

Extensible Metadata Catalogue



Extensible Metadata Catalogue

Successful completion of major milestones:

- Production level MDC (M18)
- Web service interface providing authenticated access to MDC services (M24)
- User tools for easy command line access to MDC, FC and storage elements (M24)
- Access control service (M24)

Extensible Metadata Catalogue

More results

- Simple deployment of GUI including gLite data management tools for various Linux flavours
- Installed by users in: Australia, China, Cyprus, France, Germany, Italy, Japan, Spain, UK, US
- Strong visibility with the International Lattice Data Grid

Extensible Metadata Catalogue

Developed software is successfully used within the community of lattice gauge theory physicists

- Infrastructure mature enough to provide added value instead of extra burden

Requirements from this community are far from unique

- Metadata which conform to XML schema
- Focus on datagrid
- Storage infrastructure which includes HPC centres
- Simple deployment channels for client SW on different Linux flavours

Sustainability

ILDG Meta Data Catalogue

Developed software is successfully used within the community of lattice gauge theory physicists

- 20 Installations around the world
- Germany : DESY, Jülich and Berlin
- In the order of 100 users participating in e.g. China, Netherlands, Cyprus and Spain

dCache, the scalable Storage Element

Patrick Fuhrmann
Martin Radicke
Owen Synge

dCache, the scalable Storage Element

Major Topics

- Scalability
- xRoot protocol integration
- dCache in 10 minutes w and w/o Yaim
- Automated build and test System
- SRM 2.2 integration with dCache 1.8
- HSM improvements
- Information Publishing with GLUE 1.3 (2.0 ?)
- Building german support community (DGI II and HGF)

dCache, the scalable Storage Element

Scalability

- Several thousands of pool nodes
- 20 Pbytes per site 2011
- Several 100 simultaneous transfers

- dCache To Go :-)
- checkout trac.dcache.org

dCache Storage Element

xRoot protocol integration

xRoot protocol integration

- Initially required by Alice only.
- Now all LHC experiments are interested
- dCache transfer performance excellent.
- 'file open' latency as fast/slow as will other protocols.
- Missing : security
  - Alice only pseudo security
  - root people are implementing PSEUDO gsi
  - 'no security' may be decided by sites/experiments
  - xRoot for wide area : security mandatory.

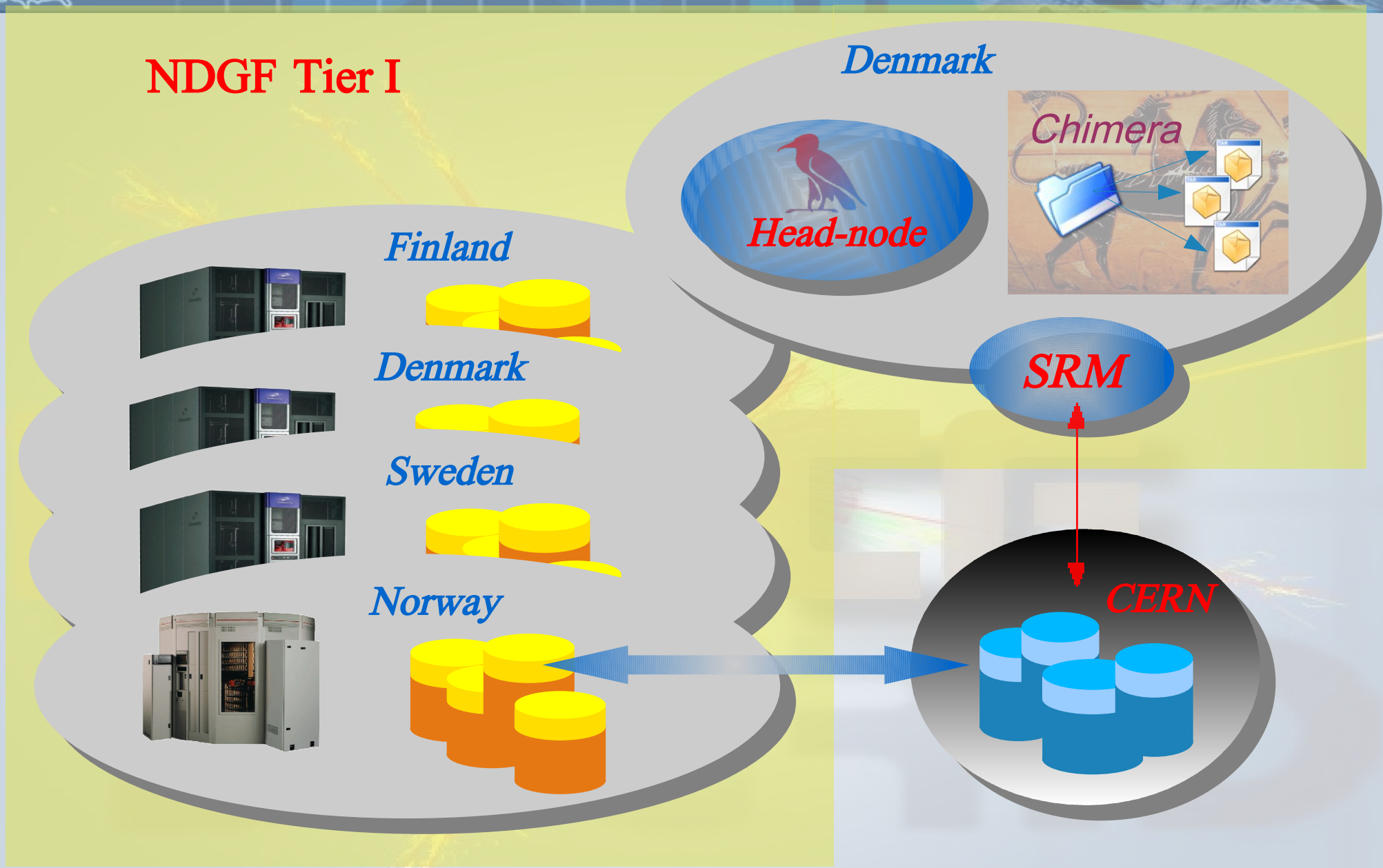
dCache, the scalable Storage Element

The NDGF approach

- Tier I center distributed among 4 countries.
- Using a single distributed dCache installation
- Head-node in Denmark
- Pools and HSM's in Finland, Sweden, Norway and Denmark
- Problems to solve :
  - Security
  - gsiFtp protocol Version II
  - HSM access (hepcg proposal)

dCache, the scalable Storage Element

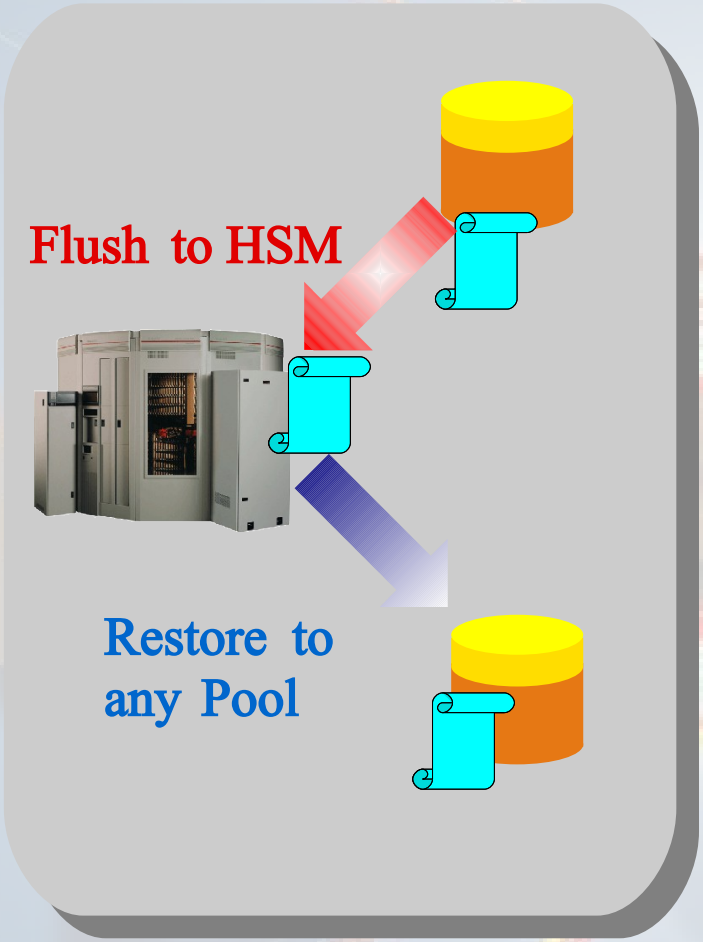
HSM Improvements



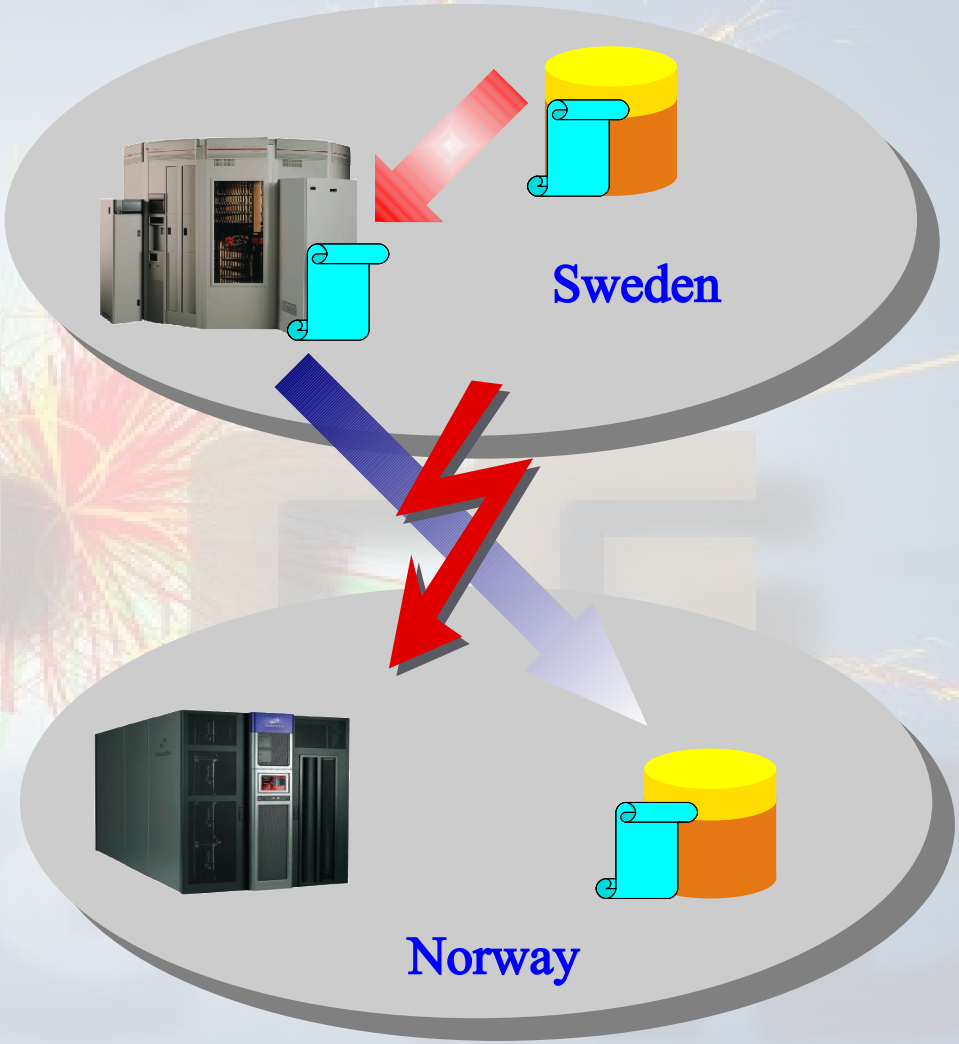
dCache, the scalable Storage Element

HSM Improvements

Single Site approach



Multi Site approach



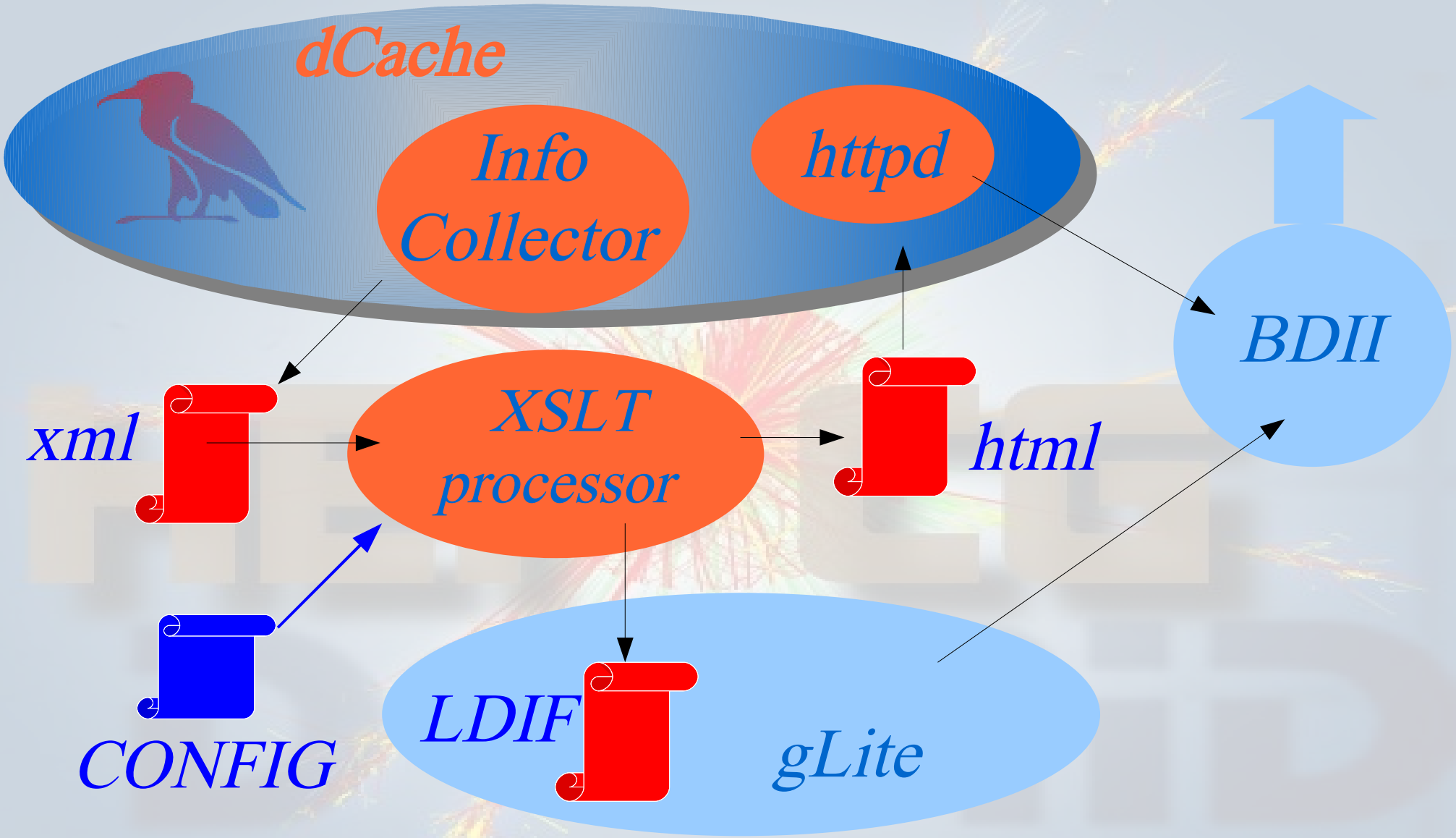
Not all pools can access all HSM systems

Information Publishing

- Schema
  - GLUE 1.3 for now
  - GLUE 2.0 for near future (specification phase)
- Transport
  - LDAP
  - HTTP
- Further use
  - Monitoring
-

dCache, the scalable Storage Element

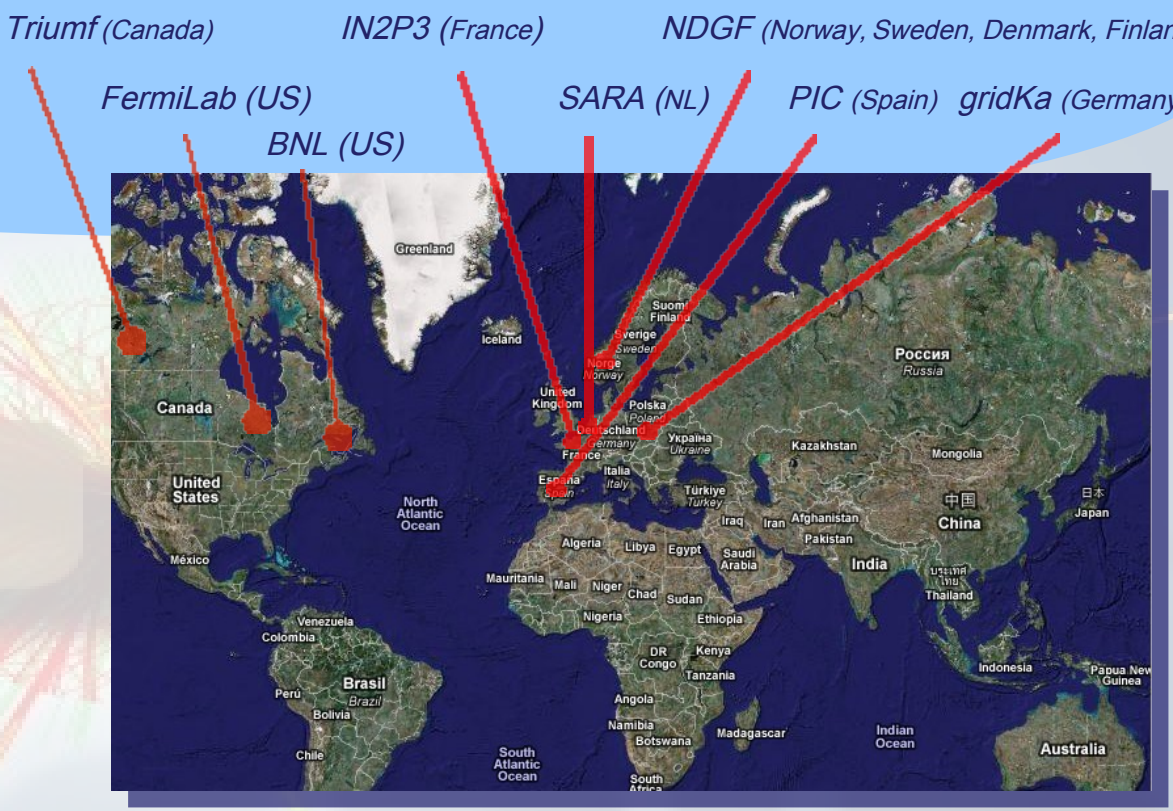
Information Publishing



dCache Sustainability

dCache, the scalable Storage Element

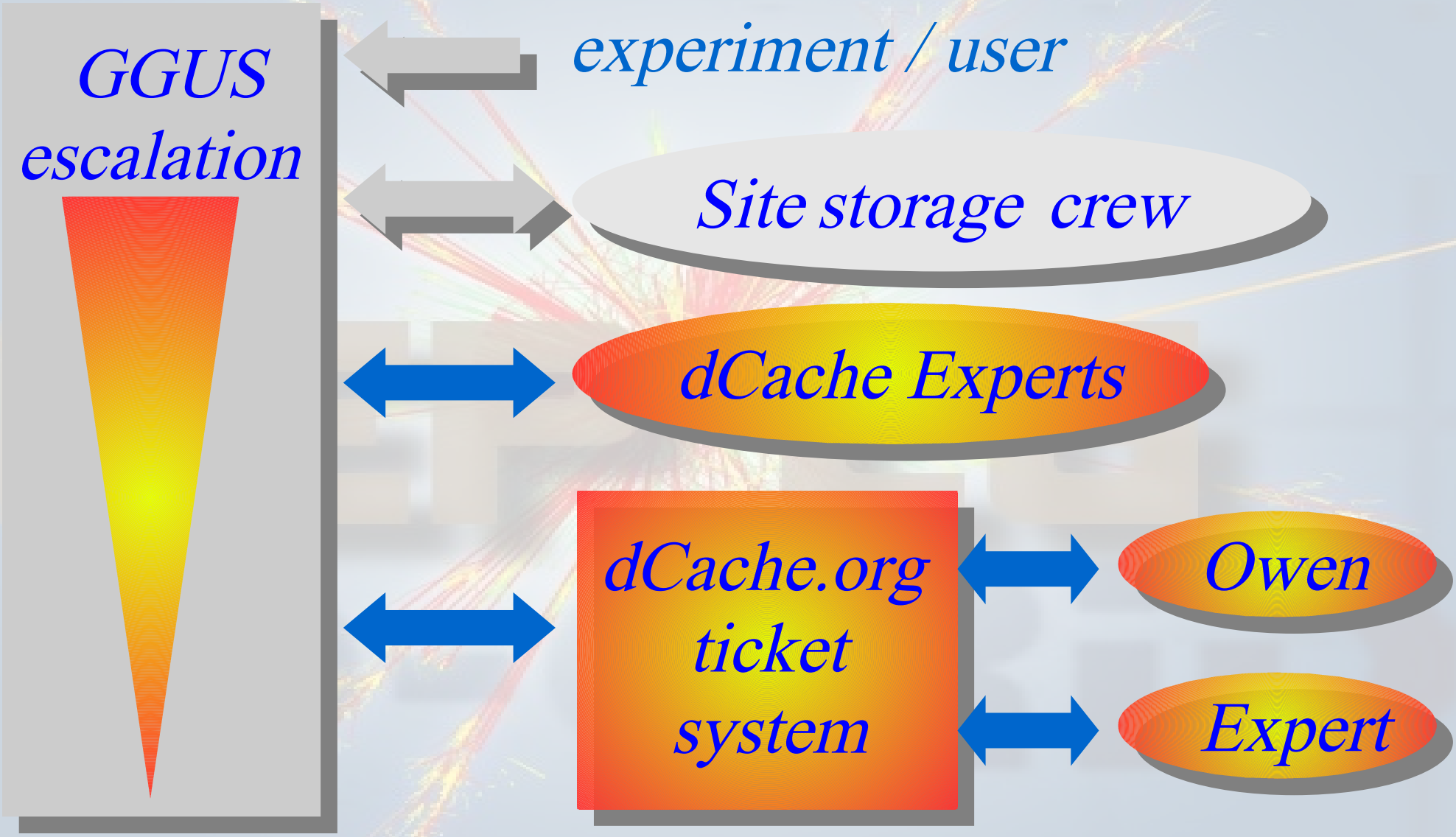
dCache LHC Tier I Centers



- 8 of 11 LHC Tier I Center are using dCache.
- more than 60 Tier II in 22 Countries
- dCache will hold the largest share of LHC data outside CERN (maybe in total)
- dCache is part of the OSG Virtual Data Toolkit.
- dCache is funded by DGI II and the HGF Alliance

Building Support (DGI II, HGF Alliance)

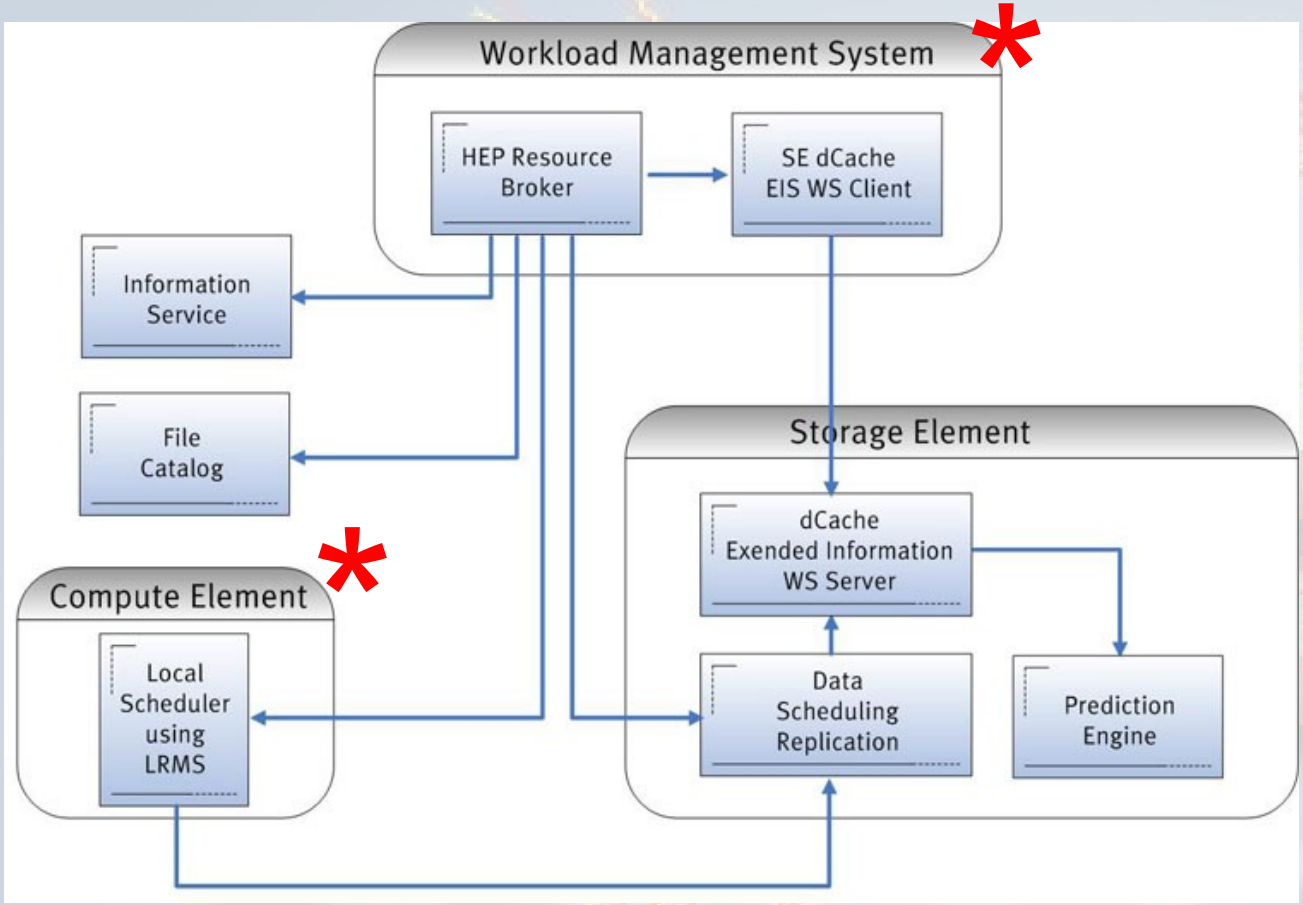
Building German (International) dCache Support



Co Scheduling / Smart Scheduling

Stefan Freitag / Lars Schley

Co Scheduling
The Idea



Drei natürliche Feinde

- Änderung des WMS Interfaces zu externen Komponenten. (Sollte nicht passieren)
- Änderung der CE Internals. War uns bewusst. (Saurer Apfel)
- Pilot Jobs : Alle unsere Komponenten können weiter benutzt werden, aber neues Gesamtdesign notwendig.

Positive Entwicklung

Die EIS (extended information service) kann möglicherweise völlig vom SRM Protokoll übernommen werden.

Status

Status as of June 2007 :

- *gLite Entwicklunghinsichtlich WMSund gLite-CE*
  - Umstellung auf neue Codebase
  - Erstellung von Patches, um auf Versionsänderungen schneller reagieren zu können (Interface Änderungen)
- *Anpassung des Entwicklungsbetts an die DGI Referenzinstallation bzgl. gLite Nachhaltigkeit*
  (For details please check talks on Friday)
  - Synergieeffekte
  - Verwendung eines gemeinsamen Batchsystems (Torque 2.1.x) Reduzierung der benötigten Maschinen, realitätsnäher
  - Anbindung des gLite-CEs an externen Torque Server (Synergie)
    - Unterscheidet sich von dem Anbindungsmechanismus für lcg-CE

Preliminary Considerations

- The particular LRMS configuration has impacts on the execution of a HEP job
  - If the processing data and the resulting data set are located on a shared file system, the execution of the three phases can be conducted on different nodes
  - If the required data and the resulting data set are located on the local file system, the execution of the three phases has to be conducted on the same node.
  - If the required data or the resulting data set are located on the local system as well as on a shared le system, the execution of the phases can be conducted as a relaxed version of the first to cases.
- Our work will be based on the observations of case 1 and 2.
- We assume that the computational phase of all jobs is at least as long as the sum of stage-in and stage-out times. That is, the computational resource usage is assumed to be the bottleneck for the co-allocation during scheduling.

(provided by Lars Schley)

Preliminary Considerations (cont.)

- Three phases
  - stage-in phase: e.g. utilizing SRM and dCache
  - computational phase: the processing of the pre-staged data
  - stage-out phase: results sets are saved to a remote storage element
- HEP jobs are sequential, exactly one CPU or core is used per job
- Jobs are transformed into a workflow consisting of activities
- Each activity is equivalent to the execution of one of the three phasesThe particular LRMS configuration has impacts on the execution of a HEP job
  - If the processing data and the resulting data set are located on a shared file system, the execution of the three phases can be conducted on **different** nodes
  - If the required data and the resulting data set are located on the local file system, the execution of the three phases has to be conducted on the **same** node.
  - If the required data or the resulting data set are located on the local system as well as on a shared le system, the execution of the phases can be conducted as **a relaxed version of the first to cases.**

(provided by Lars Schley)

Co Scheduling

- For the last 12 month, theoretical models have been evaluated to prove the benefits of co-allocation of concurrent data movement.
- Two investigated approaches
  - Overall utilization improvement
  - Minimized response time (AWRT)
- Conditions for improved utilization
    - I leave this to Lars.

Co Scheduling
Sustainability

- *Wecould prove that the technical framework works and can be integrated into the gLite WMS.*
- *Future of the co-scheduling approach is not clear.*
- *Pilot jobs are taking over.*
- *The integration highly depends on stable interface specification of WMS gLite.*
- *Experiments take full control of data location :*
  - *Experiment framework perform 'prestage'*
  - *User jobs won't have access to tape systems*

Conclusion

- All proposed tasks have been delivered.
- Community requirements are hard to predict over a time of 3 years.
- Some deliverables had to be rescheduled or re-interpreted.
- Lesson learned : Keep design simple.
- Overall success is significantly higher than expected.
- Sustainability is excellent.

FIN

Further reading :

www.dCache.org

www.lqcd.org/ildg

www-zeuthen.desy.de/latfor/ldg