Demo Session

Introduced by Massimo Lamanna
NA4 HEP coordinator - CERN

EGEE-II 1st EU Review (CERN)
15-16 May 2007

www.eu-egee.org

Information Society and Media

Welcome to the demo session!

- **Why a demo session?**
  - A good demo is an important tool to
    - Exchange views and experience on the grid added value for applications
    - Attract new users by clear exposition of the advantages of the grid for given applications

Three demos today

- **Intelligent Distributed Data Management**
  Enabling consistent collaborative data use in Earth Science
  - Earth Sciences community
    - Kerstin Ronneberger and Stephan Kindermann (DKRZ, Hamburg)

- **Wisdom Production Environment**
  Search of drugs against Malaria and BirdFlu
  - Biomedical community
    - Jean Salzemann and Vincent Bloch (IN2P3 Clermont-Ferrand)

- **Dashboard**
  Integrated monitor of large grid communities
  - High-Energy Physics community
    - Julia Andreeva and Pablo Saiz (CERN)

Messages

- **Grid added value**
  - Ease daily workflows and support collaborative work (Climate Data Management)
  - Access a new scale of processing and data sharing (WISDOM)
  - Monitoring of applications contributing to the grid evolution (Dashboard)
- **Relevance**
  - All activities are part of leading-edge research in the corresponding fields
- **Examples of cross-applications collaboration**
  - Sharing experience, tools and actual software solutions
    …enabled by EGEE!

Intelligent Distributed Data Management in Earth system science

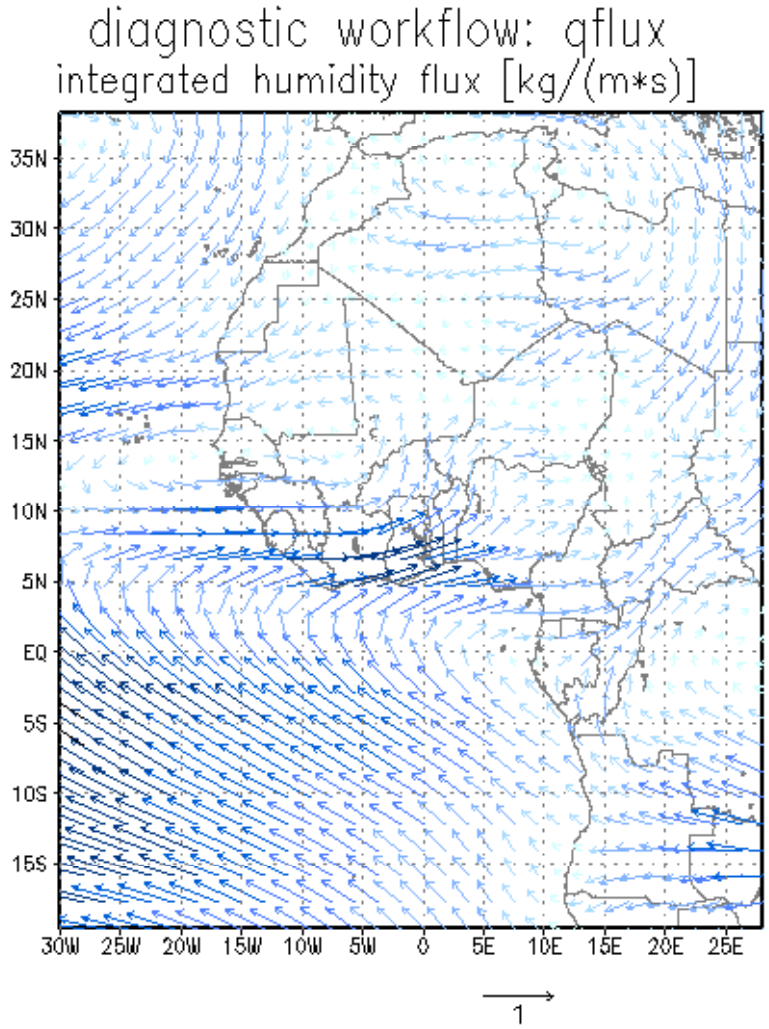
K. Ronneberger, DKRZ, Germany
S. Kindermann, DKRZ,Germany
B. Bräuer, AWI, Germany
T. Brücher, ZAIK, Germany
H. Ramthun, M&D, Germany
M. Stockhause, MPI-Met, IFM-GEOMAR, Germany

www.eu-egee.org

Information Society and Media

QFLUX: Humidity flux calculation

Enabling Grids for E-sciencE



Tim Brücher, ZAIK, bruecher@uni-koeln.de

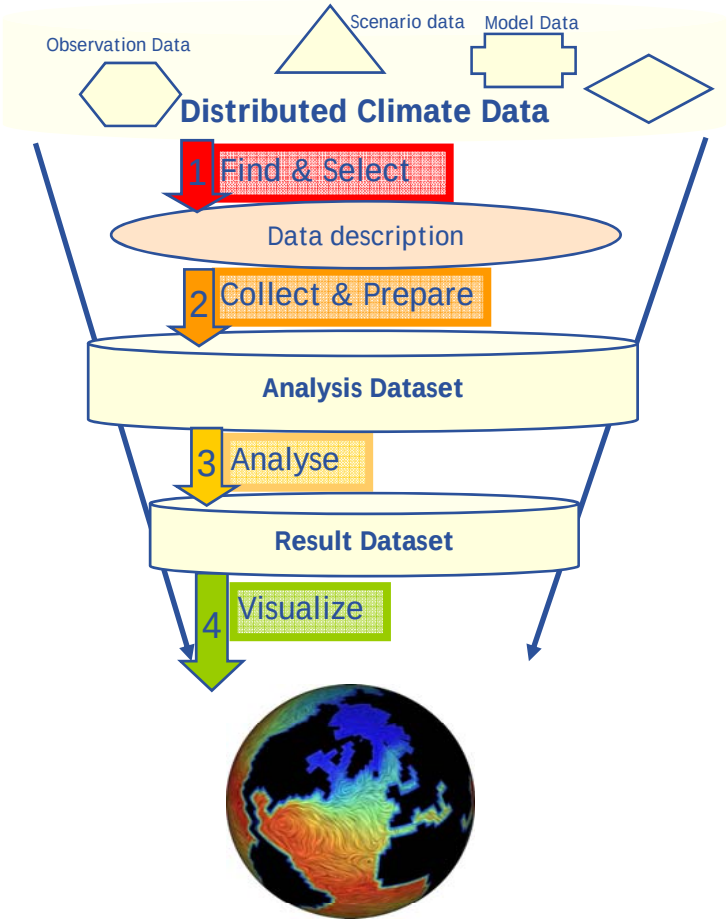
Structure

- What is Earthsystem Science about?
  - Typical workflows
  - Traditional infrastructure
- Where can grid-technology help?
  - Limits of the current practice
- How do we use this technology?
  - Demo of the portal
  - Demo of an example workflow
  - Outline of the developed infrastructure
- What is still missing
  - Next steps and challenges

Earthsystem Sciences

- **Goal:** learn about the past, the present, and possible futures of the earth system
- **Community:** internationally and interdisciplinary distributed but strongly interconnected
- **Method:** Analysing, comparing and processing data
- **Input:** data from observations and/or other modelling studies

Typical workflow



An example workflow: “qflux”

Location		Datavolume
Various data centers & portals	1 Find & Select relevant & available datasets: Specific humidity, Temperature, Wind speed → Distributed Climate Data	Several PB ~3,1TB (300-500 files)
Institutional storage & computing facilities	2 Collect & Prepare a temporal and spatial subset of the data → Analysis Dataset	~10,3GB (28 files) ~76 MB
local facilities	3 Analyse the integrated, transport of humidity between selected levels → Result Dataset	~6MB
Personal Computer	4 Visualize selected result	~66KB

Potential use of grid technology

Current issues

- **Search & select**
  - Different portals with **different authentications and data descriptions**
- **Collect & prepare**
  - **Different access mechanisms** of the different providers
  - Pre-processing requires **sufficient local facilities**
- **Analyse**
  - Existing **tools** and already processed **data** are **available locally** and **miss proper description**
- **Visualize**
  - **Detached** from the remaining workflow

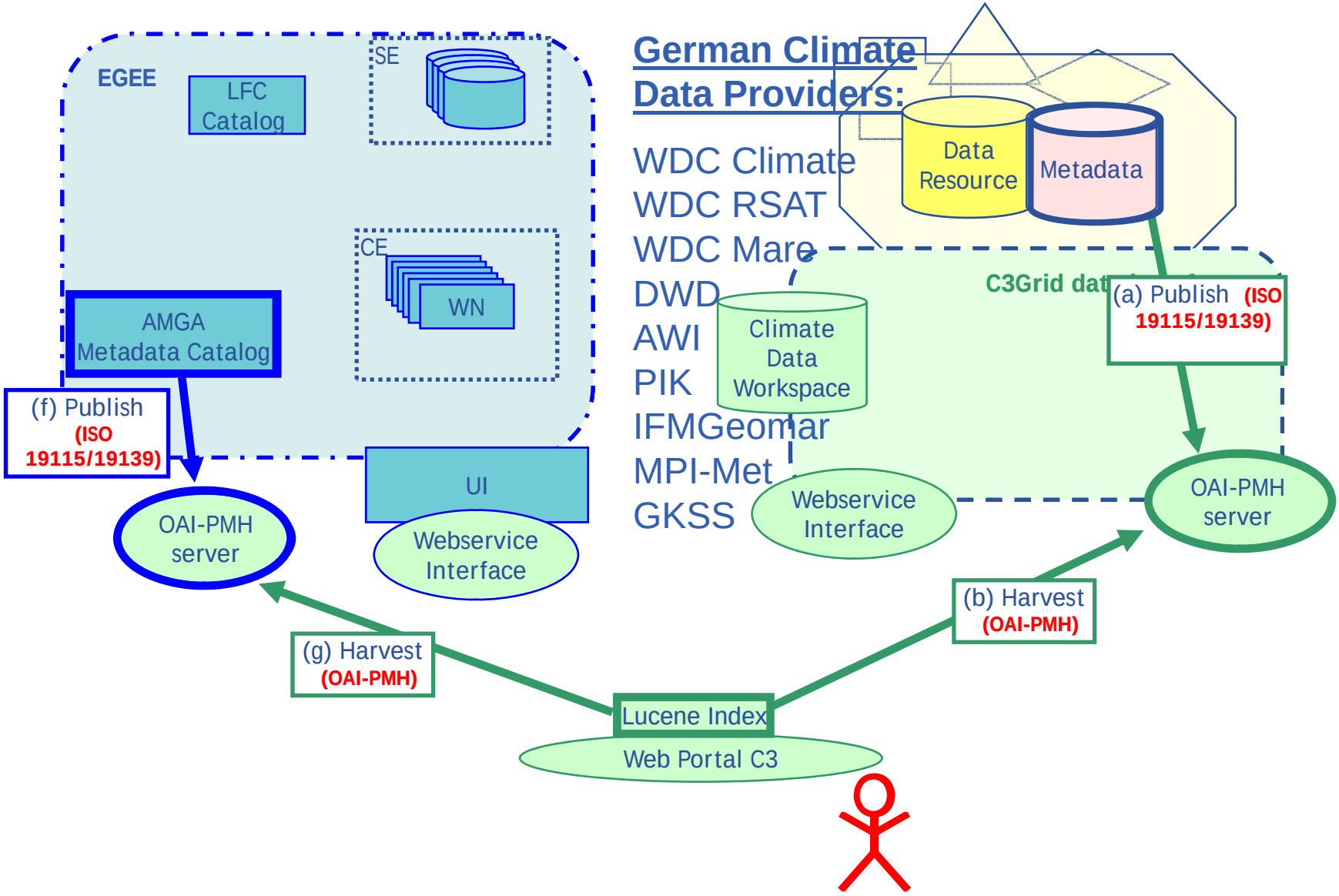
- **Central unique authentication** to a common catalogue with **standardized metadata**
- **Shared resources** with **standardized access** hiding proprietary access mechanisms
- Commonly defined **tool description**
- **Log** processing steps and **automatically republish** processed data
- Integrate basic visualization (**first peep**) into the workflow

C3 Grid and EGEE - the components

- **Central web-portal:** unique entrance point to common central metadata catalogue (Lucene index) and access facility
- **Standardized Metadata:** hierarchical description of discovery- and some use-aspects of the data (ISO 19115/ISO 19139)
- **Standardized data request interface:** hide the complexity of specific data access mechanisms and pre-processing functionality (webservice technology)
- **Automatic update and republishing of metadata:** metadata of data processing is updated, managed and can be harvested (AMGA + java extension, OAI-PMH server)

Find & select
Collect & prepare
analyse
visualize

Bridging EGEE and C3



Demo

(1) Search-, discover-, and select-functionality of the portal

(2) Upload and register data to EGEE

(3) Trigger the example workflow qflux from the portal

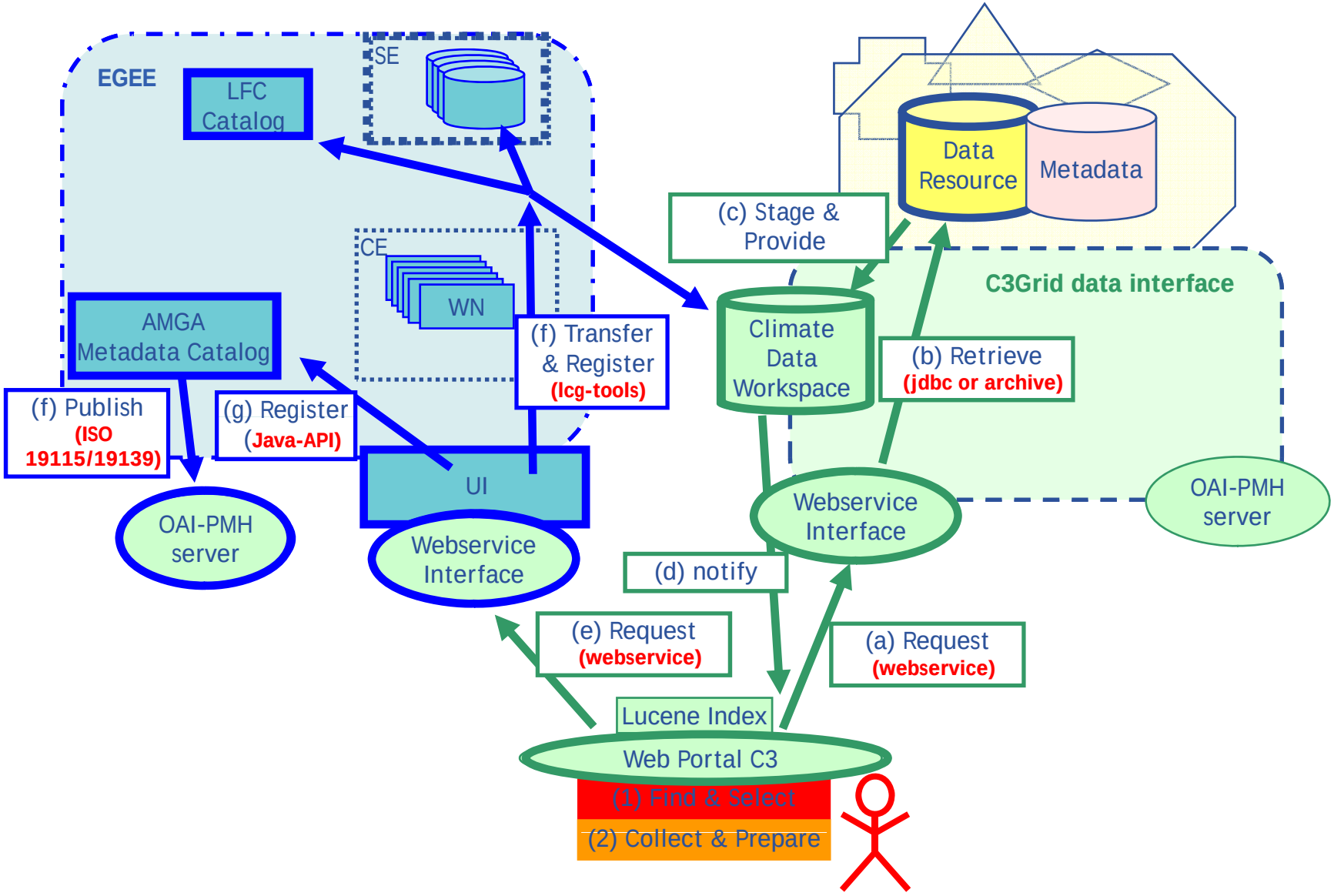
Demo

(1) Search-, discover-, and select-functionality of the portal

(2) Upload and register data to EGEE

(3) Trigger the example workflow qflux from the portal

Upload pre-processed data to EGEE



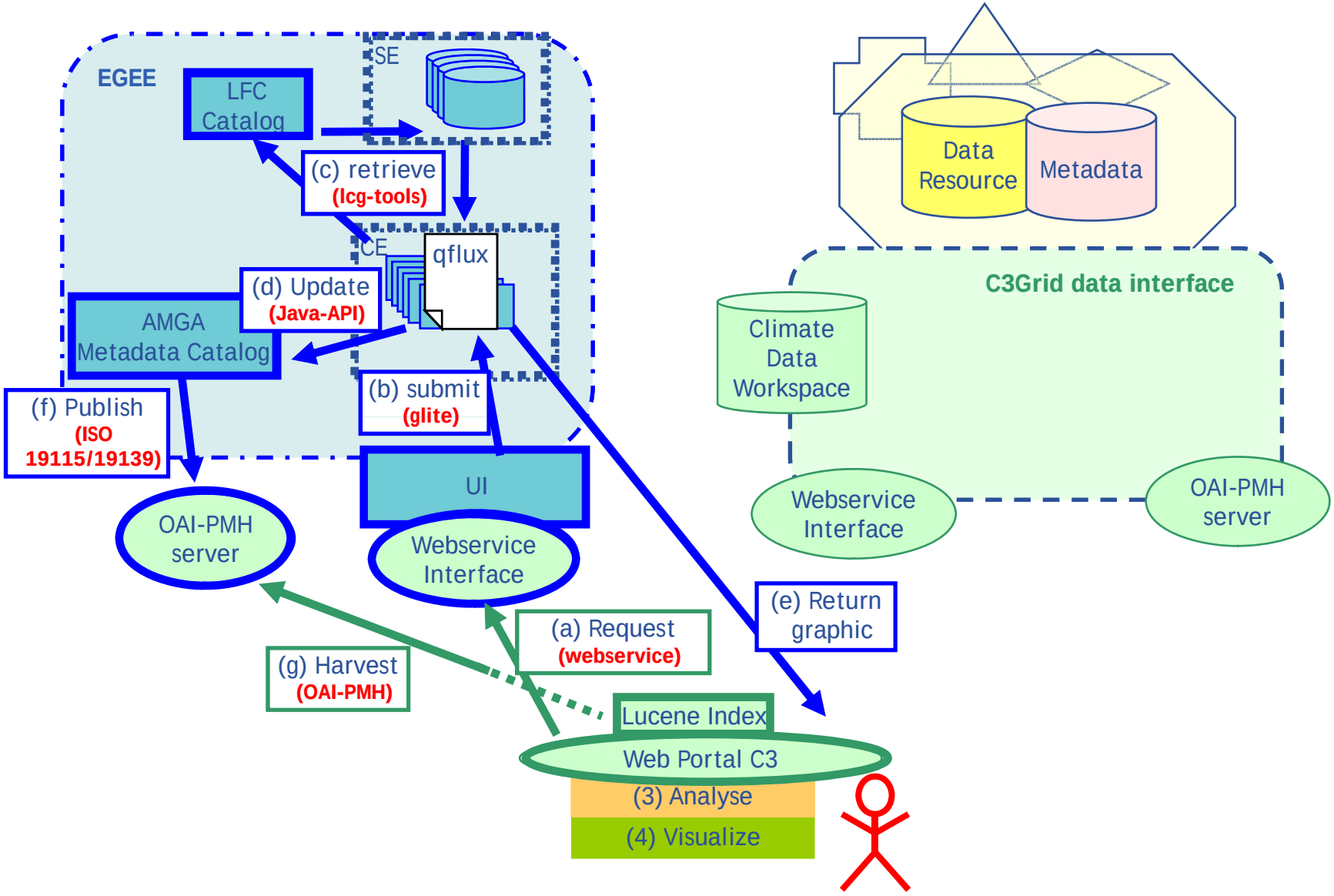
Demo

(1) Search-, discover-, and select-functionality of the portal

(2) Upload and register data to EGEE

(3) Trigger the example workflow qflux from the portal

Trigger qflux workflow



Data access in ESR grid projects

Enabling Grids for E-sciencE

	Earth System Grid project (USA)	C3 Grid/(EGEE) (Germany)	NERC data grid (UK)
Scope (project)	High performance access of climate model data	Uniform & effective discovery and access of data of various disciplines & types	Harmonized & detailed search and access of data of various disciplines & types
Data stock (status)	• Homogenous • Flat-file storage	• Heterogeneous • Databases & flat-file storage	• Heterogeneous • Databases & flat-file storage
Data description (solution)	• Use aspect of data, tools and models • E.g. NcML for netCDF data	• Discovery and some use aspects • ISO 19115/ISO 19139	• Content of the data in great detail • Semantic datamodel (CSML, based on GML)
Data access (solution)	• Intelligence at portal • Different protocols	• Uniform access interface • Intelligence at data provider / grid	• link from portal to data provider • Different protocols

Potential Impact

A framework to easily and consistently exchange and manage ES-data and tools between EGEE and traditional ES data-storage-systems

- **Potential impact on current and potential EGEE ES-community**

A framework to connect further portals or infrastructures to EGEE

- **Potential impact on international ES-community**

Built on international standards thus easy adaptable/expandable by other disciplines and by further partners

- **Potential impact on other disciplines**

Next steps

- Expand the demonstrated prototype to a reliable and stable system
- Porting further workflows and some pre-processing functionality to EGEE
- Enlarge the user community

Future challenges or missing bricks

- Comprehensive and consistent security context to control access to (restricted) data with a single sign-on
  - **Approach: federated AA infrastructure based on Shibboleth**
- Analysis-services description to improve discovery, use and share possibilities
  - **Approach: adapt ISO19119/19139 as a common metadata format for analysis-tool description**
- Modularized workflows to increase the flexibility and enable intelligent scheduling
  - **Approach: implement a workflow information service**

Thank you!

WISDOM: a large scale docking application on the grid

Vincent Bloch, Hurng-Chun Lee, Jean Salzemann

LPC Clermont-Ferrand IN2P3/CNRS

Academia Sinica, Taipei

www.eu-egee.org

Information Society and Media

Goals

- **WISDOM stands for World-wide In Silico Docking On Malaria**
- **Goal: find new drugs for neglected and emerging diseases**
  - Neglected diseases lack R&D
  - Emerging diseases require very rapid response time
- **Method: grid-enabled virtual docking**
  - Cheaper than in vitro tests
  - Faster than in vitro tests

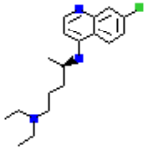
Grid-enabled virtual docking

Millions of potential drugs to test against interesting proteins!

~~High Throughput Screening ~10$/compound, several hours~~

Too costly for neglected disease!

Compounds:

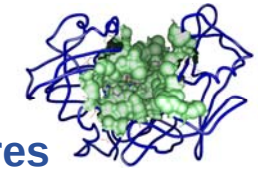


ZINC: 4.3M

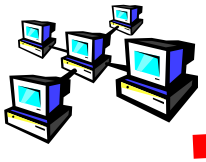
Chembridge: 500 000

Molecular docking (FlexX, Autodock)
~1 to 15 minutes

Targets:



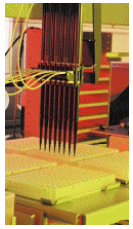
PDB: 3D structures



Data challenge on EGEE
~ 2 to 30 days on ~5000 computers

~~Cheap and fast!~~

Selection of the best hits



Hits screening using assays performed on living cells

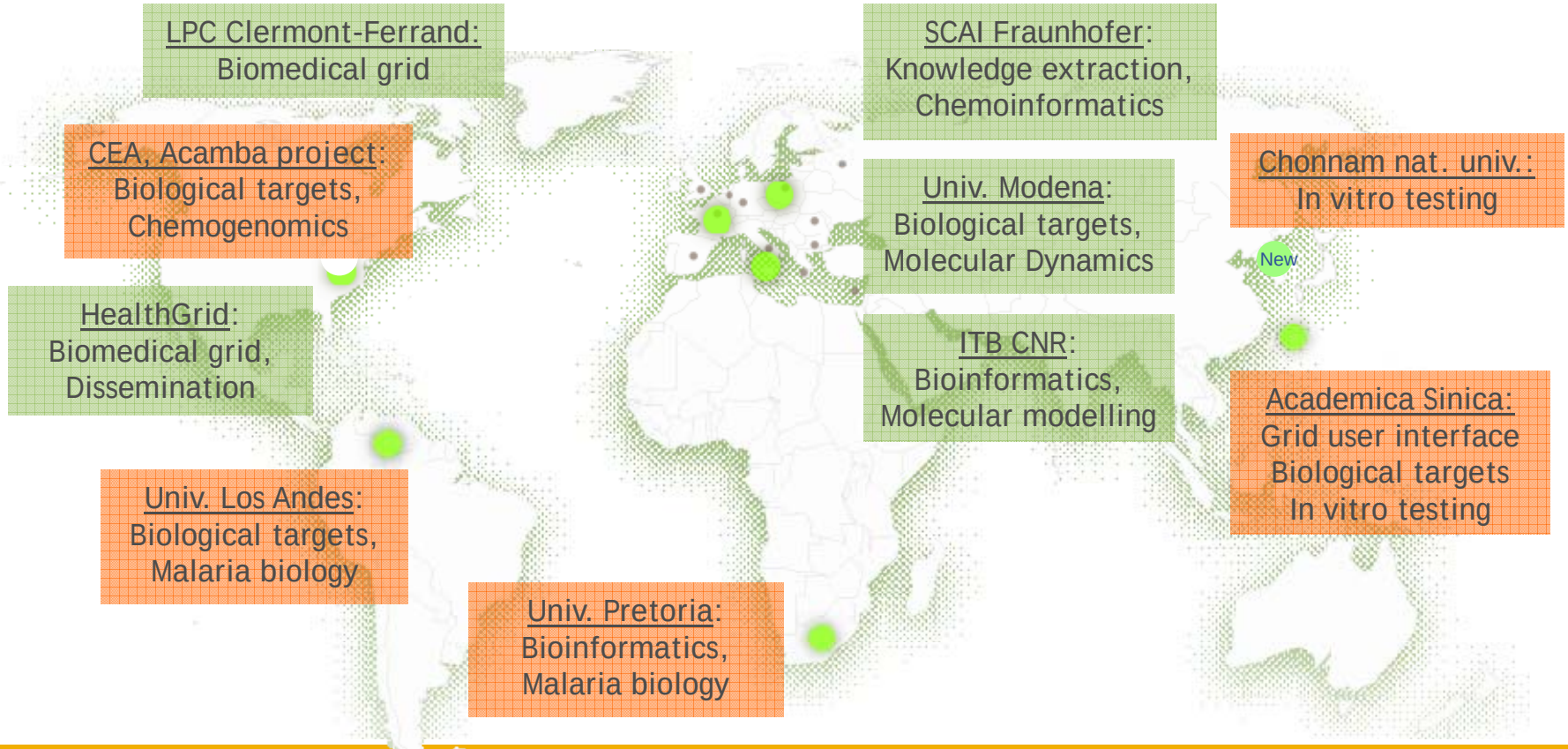
Leads

Clinical testing

Drug

The grid added value

- The grid provides the centuries of CPU cycles required on demand
- The grid provides the reliable and secure data management services to store and replicate the biochemical inputs and outputs
- The grid offers a collaborative environment for the sharing of data in the research community on avian flu and malaria



Objective of the WISDOM data challenges

- **Objective**
  - To dock a whole compound database in a limited time with a minimal human involvement
- **Need for an optimized environment**
  - To achieve production in a limited time
  - To optimize performances
- **Need for a fault tolerant environment**
  - To handle Grid heterogeneity and dynamics
  - To collect and store critical data
- **Need for user-friendly high-level interfaces**
  - To ease the execution
  - To offer a service to the non grid experts

Statistics of deployment

- **First Data Challenge: July 1st - August 15th 2005**
  - Target: malaria
  - 80 CPU years
  - 1 TB of data produced
  - 1700 CPUs used in parallel
  - 1st large scale docking deployment world-wide on a e-infrastructure

- **Second Data Challenge: April 15th - June 30th 2006**
  - Target: avian flu
  - 100 CPU years
  - 800 GB of data produced
  - 1700 CPUs used in parallel
  - Collaboration initiated on March 1st: deployment preparation achieved in 45 days

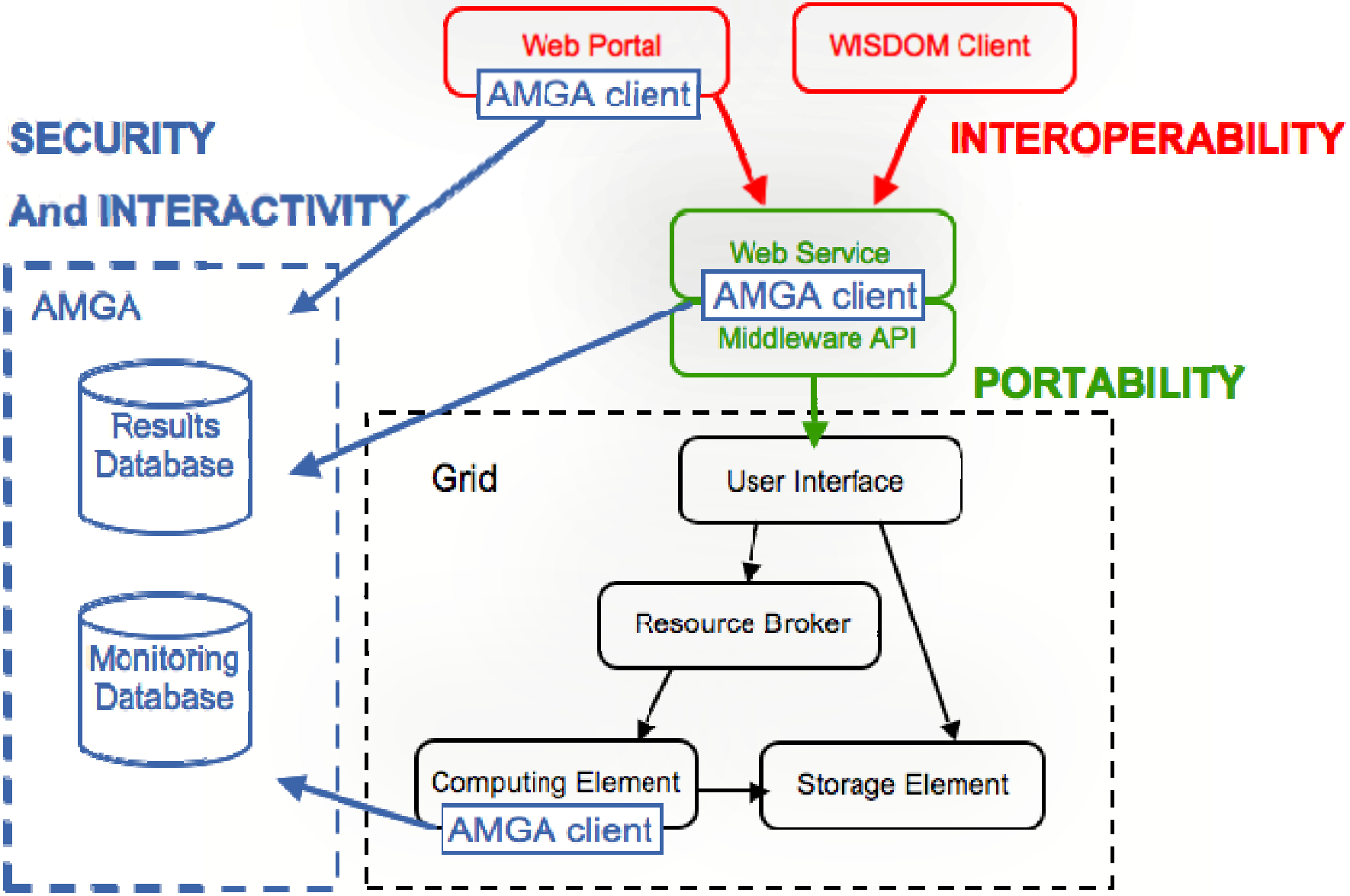
- **Third Data Challenge: October 1st - 15th December 2006**
  - Target: malaria
  - 400 CPU years
  - 1,6 TB of data produced
  - Up to 5000 CPUs used in parallel
  - Very high docking throughput: > 100.000 compounds per hour

Grid Statistics Portal

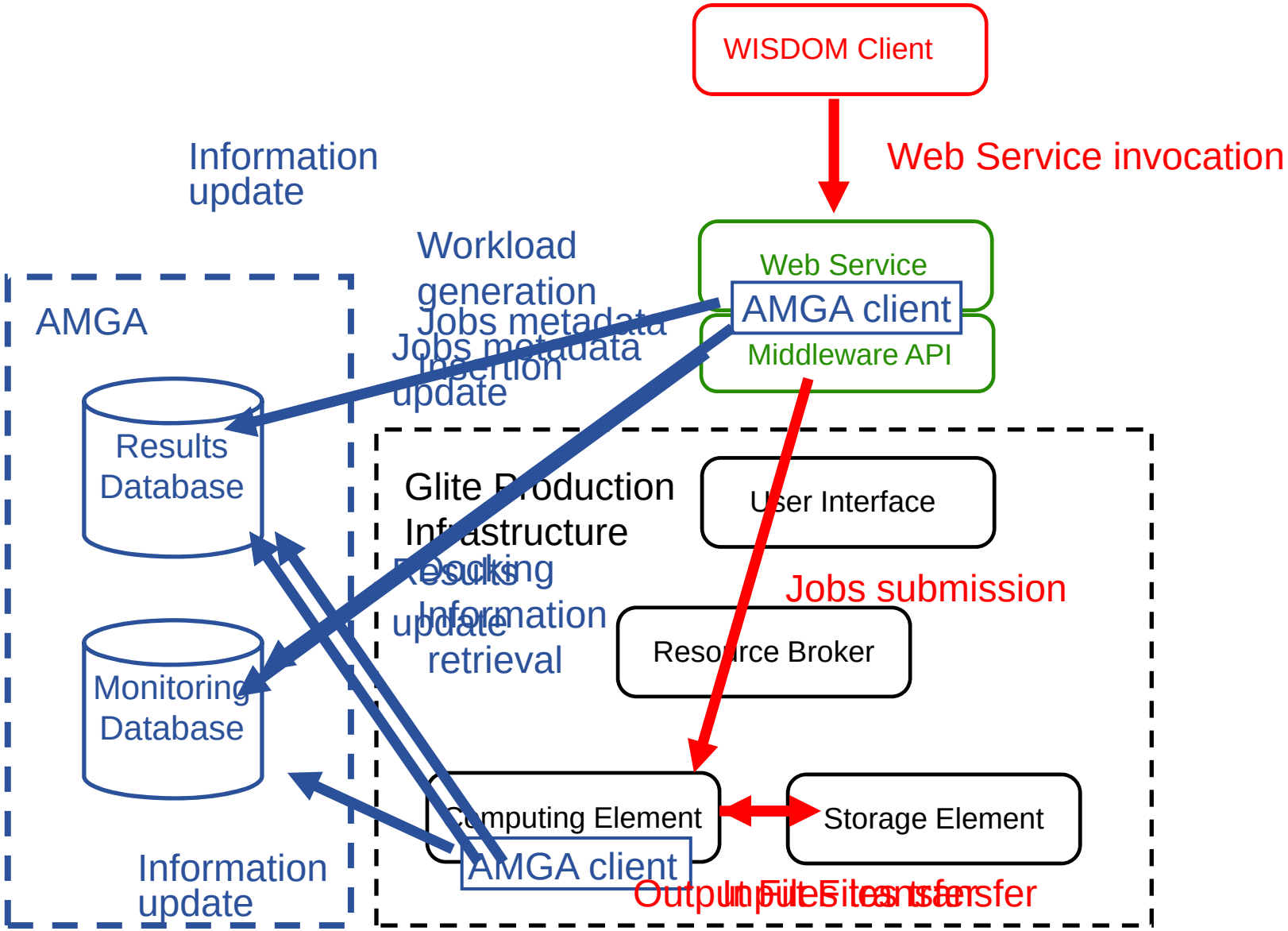
Available at http://wisdom-demo.healthgrid.org

- **Real-Time monitoring of the Grid**
- **Customizable interface**
- **Drag and drop components**

Production Environment



Environment (2/2)



Interactive Web Portal

Available at http://t-ap41.grid.sinica.edu.tw:8088/WisdomPortal

- **User-friendly Interface for biologists**
- **Real Time output of the results**
  - 3D views of the docking poses and structures
- **Resubmission and monitoring of docking jobs**

Encouraging biochemical results

- **Avian flu data challenge: in the selection of 2250 compounds out of initial 308585 compounds:**
  - 5 out of 6 known effective inhibitors were found.
  - enrichment factor of 111 was observed.

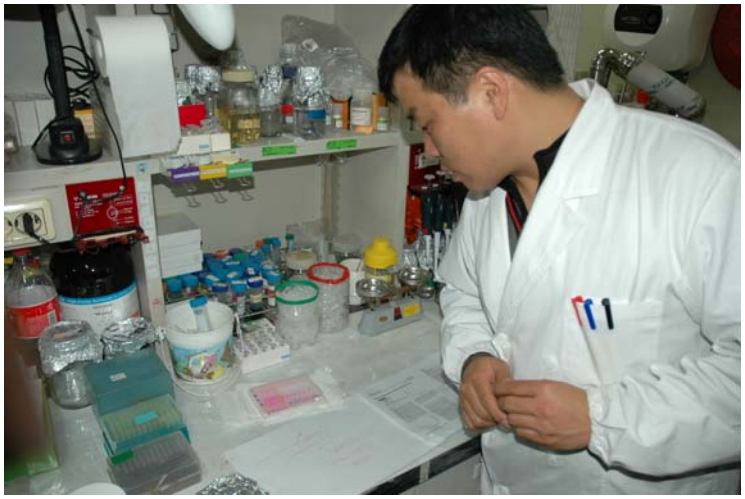
Global effectiveness:
$(Hits_{sampled}/N_{sampled})/(Hits_{total}/N_{total})$
Pearlman & Charifson, JMC, 2001

Pre-sceening (AUTODOCK)
over collection and sample first 15%
EF^1
$= (5/6)/15\% = 5.5$

Re-ranking (SDDB) first 15% and
sample first 5%
$EF^2 = (5/6)/(5\%*15\%) = 111$

- **Experimental assay confirms 7 active out of 123 purchased "potential hits"**

- **Data challenges on malaria: the 25 most promising compounds out of 500.000 are now being purchased and will be tested in vitro at Chonnam National University, South Korea**



Conclusions

- **WISDOM proposes a new approach to drug discovery thanks to the grid**
  - Rapid deployment of very large scale virtual screening
  - Collaborative environment for the sharing of data in the research community
- **WISDOM fully exploits EGEE services, APIs and resources.**
  - AMGA allows to store securely results and statistics immediately
  - Web Service Interface using WS-I profile guarantees interoperability
- **First biochemical results demonstrate grid relevance to the drug discovery community**
  - Grid is a superior tool to discover new drugs

Credits

- **Development of the WISDOM environment**
  - ASGC: Yu-Hsuan Chen, Li-Yung Ho, Hurng-Chun Lee
  - ITB-CNR: G. Trombetti
  - CNRS-IN2P3: V. Bloch, M. Diarena, J. Salzemann
  - HealthGrid: B. Grenier, N. Spalinger, N. Verhaeghe

- **Biochemical preparation and analysis**
  - ASGC: Y-T Wu
  - Chonnam National University: D. Kim & al
  - CNRS-IN2P3: A. Da Costa, V. Kasam
  - ITB-CNR: L. Milanesi & al

LHC Experiments Dashboard demo

Julia Andreeva and Pablo Saiz (CERN)

On behalf of the Dashboard development group :
Catalin Cirstoiu, Benjamin Gaidioz, Juha Herrala,
Gerhild Maier, Ricardo Rocha, Pablo Saiz, Julia Andreeva

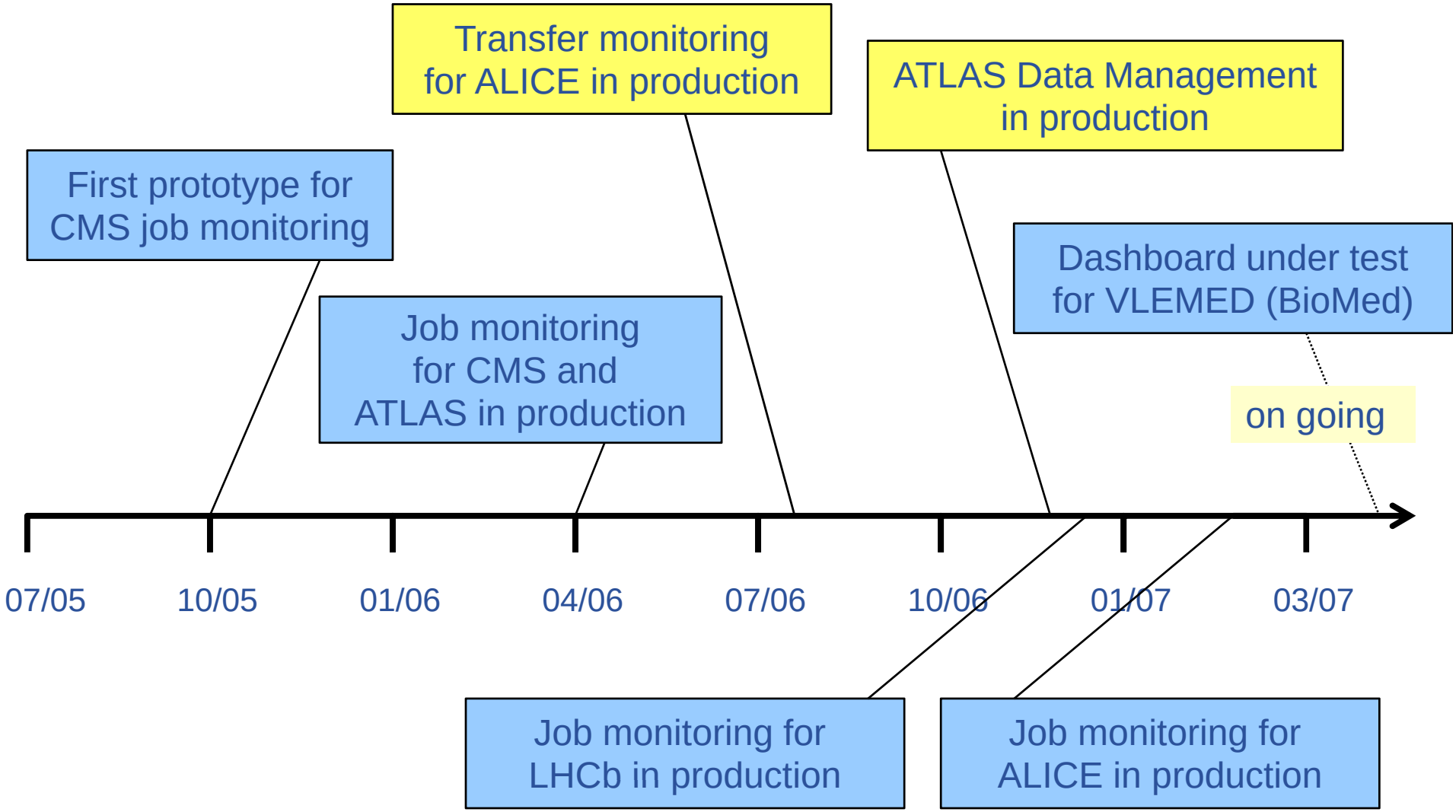
Outline

- **Introduction**
- **Main Monitoring Applications**
  - Job monitoring
  - Site reliability
  - Data management monitoring
- **Conclusions**

Why one more monitoring tool, what does make it different from the others?

- *Independent of the Grid flavor*
- *Covering different areas and various aspects of the VO activities*
- *Combining Grid job status and service status information with the specific data of the experiment/application*
- *Flexible enough to allow rapid integration with the new requirements*
- *Reliable and scaling well*

Evolution of the project



Developers

- *The tool is developed by ARDA (CERN; NA4 HEP) team in collaboration with MonAlisa (Caltech) developers and participation of ASGC (Taiwan), MSU and JINR (Russia) and LAL (France, JRA2)*
- *Recently a collaboration was set up with the EDS company within the CERN OpenLab project*
- *The future evolution of the project is being discussed in conjunction with other EGEE activities and in particular with SA1*

Experiment Dashboard link

http://dashboard.cern.ch

Experiment Dashboard concept

INPUT

Multiple sources of information

- Increasing the reliability
- Providing both global and very detailed view

OUTPUT

Providing output in various formats

(Web pages, xml, csv, image formats)

Collect data of VO interest coming from various sources
Store it in a single location
Provide UI following VO requirements
Analyze collected statistics
Define alarm conditions

Can satisfy users with various roles:

- Generic user running his jobs on the Grid
- Site administrator
- VO manager, production or analysis group coordinator, data transfer coordinator…

Can be used by various clients
both users and applications

report monitoring information

- **Potentially other**
- **Clients:**
- **PANDA, ATLAS production**
- **<XML,CSV, image formats>**

Monitoring applications

- Job Monitoring
- Site Reliability
- Data Management Monitoring
- Data Transfer Monitoring
- Monitoring of the distributed DBs (developed by 3D project)

Job monitoring

- **What is the status of the jobs**
  - belonging to an individual user/group/VO
  - submitted to a given site or Grid flavor or via a given resource broker
  - reading a certain data sample, running a certain application…
- **If they are pending or running**
  - for how long, where?
- **If they are finished**
  - Did they fail or run properly?
- **If they failed – why?**
- **How resources are shared?**
- **Are the available resources are efficiently?**

Site reliability

- Job Exit Reason returned by Logging and Bookkeeping System is often not enough to understand what happened to the job
- Job can be resubmitted multiple time and all “job attempts” not necessary happen at the same site
- Dashboard analyses the reliability of the job processing and calculates success rate by splitting the job flow in “job attempts”
- Detailed failure reason analysis and error ranking
- Where possible the troubleshooting recipes are provided

Data management monitoring

- Data management is a key component of the computing models of the LHC experiments
- Very strong requirements regarding data safety, consistency of the bookkeeping and large-scale data replication
- Example of the Data Management monitoring application is the Dashboard monitoring of the ATLAS Distributed Data Management (DDM)

Plans

- **User Interface**
  - User Task monitoring
- **Improvement of data completeness and reliability**
  - Adding new information sources (GridIce, SAM, APEL, condor-g)
- **Improvement of effectiveness for troubleshooting**
  - Sending alarms. Need to define alarm conditions with the experiments for various use cases
  - Collecting and analyzing failure information
  - Collecting troubleshooting recipes, making them available at the dashboard UI
  - Correlating where relevant failures with the results of the SAM tests

Conclusions

- Experiment Dashboard is used by all 4 LHC experiments and evolving very fast to match their requirements
- Currently is evaluated by VO outside LHC community
- The tool proved to provide reliable and useful VO-oriented monitoring data, with needed level of details, available in various formats.

Demo session conclusions

- **Grid added value**
  - Ease daily workflows and support collaborative work (Climate Data Management)
  - Access a new scale of processing and data sharing (WISDOM)
  - Monitoring of applications contributing to the grid evolution (Dashboard)
- **Relevance**
  - All activities are part of leading-edge research in the corresponding fields
- **Examples of cross-applications collaboration**
  - Sharing experience, tools and actual software solutions

  …enabled by EGEE!