Data Management in Scientific Gateways

SCI-BUS Data Management Taskforce

EGI Technical Forum, Sep 2012, Prague

Presented by: Mark Santcroos, AMC

SCI-BUS is supported by the FP7 Capacities Programme under contract nr RI-283481

Outline

- SCI-BUS Project
- Considerations
- Use cases
- Goals
- Architecture (WIP)
- Status and plans
- Questions and discussion

SCI-BUS Project

- WS-PGRADE / gUSE
- Various communities, various demands for solutions on data management
- Motivation for this work

Project partners

MTA SZTAKI
COMPUTER AND AUTOMATION RESEARCH INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES

amC

Middle East Technical University

ScaleTools
Strategy Realization

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

TRINITY COLLEGE DUBLIN
COLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH
THE UNIVERSITY OF DUBLIN

E-GROUP
SECURE BUSINESS AUTOMATION

LAUREA
UNIVERSITY OF APPLIED SCIENCES

ISTITUTO NAZIONALE DI ASTROFISICA
INAF
NATIONAL INSTITUTE FOR ASTROPHYSICS

bifi
Instituto Universitario de Investigación
de Biocomputación y Física
de Sistemas Complejos
Universidad Zaragoza

4DSoft

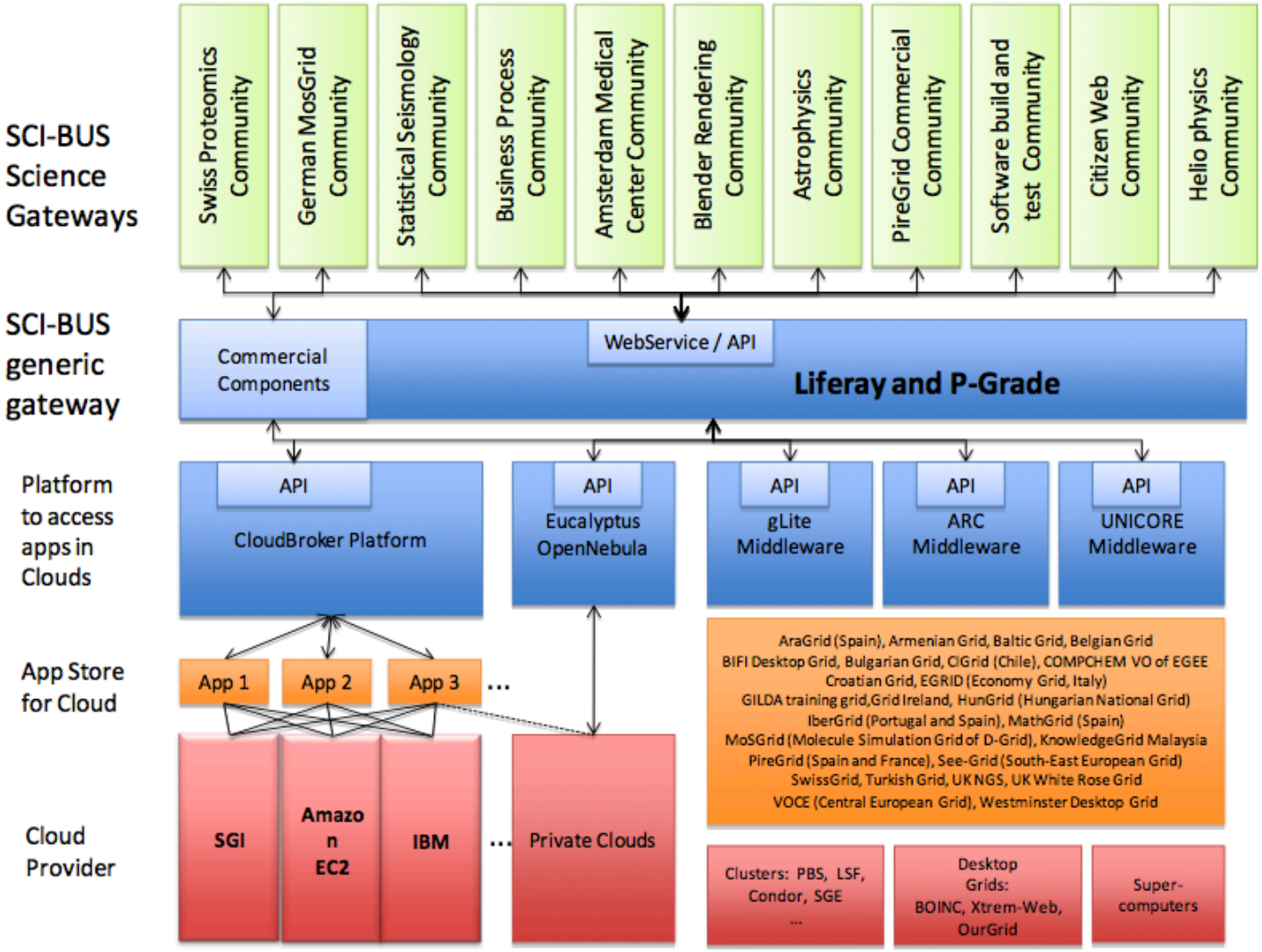
CloudBroker

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

UNIVERSITY OF WESTMINSTER

Simsoft

SCI-BUS Big Picture



Considerations

- Data?

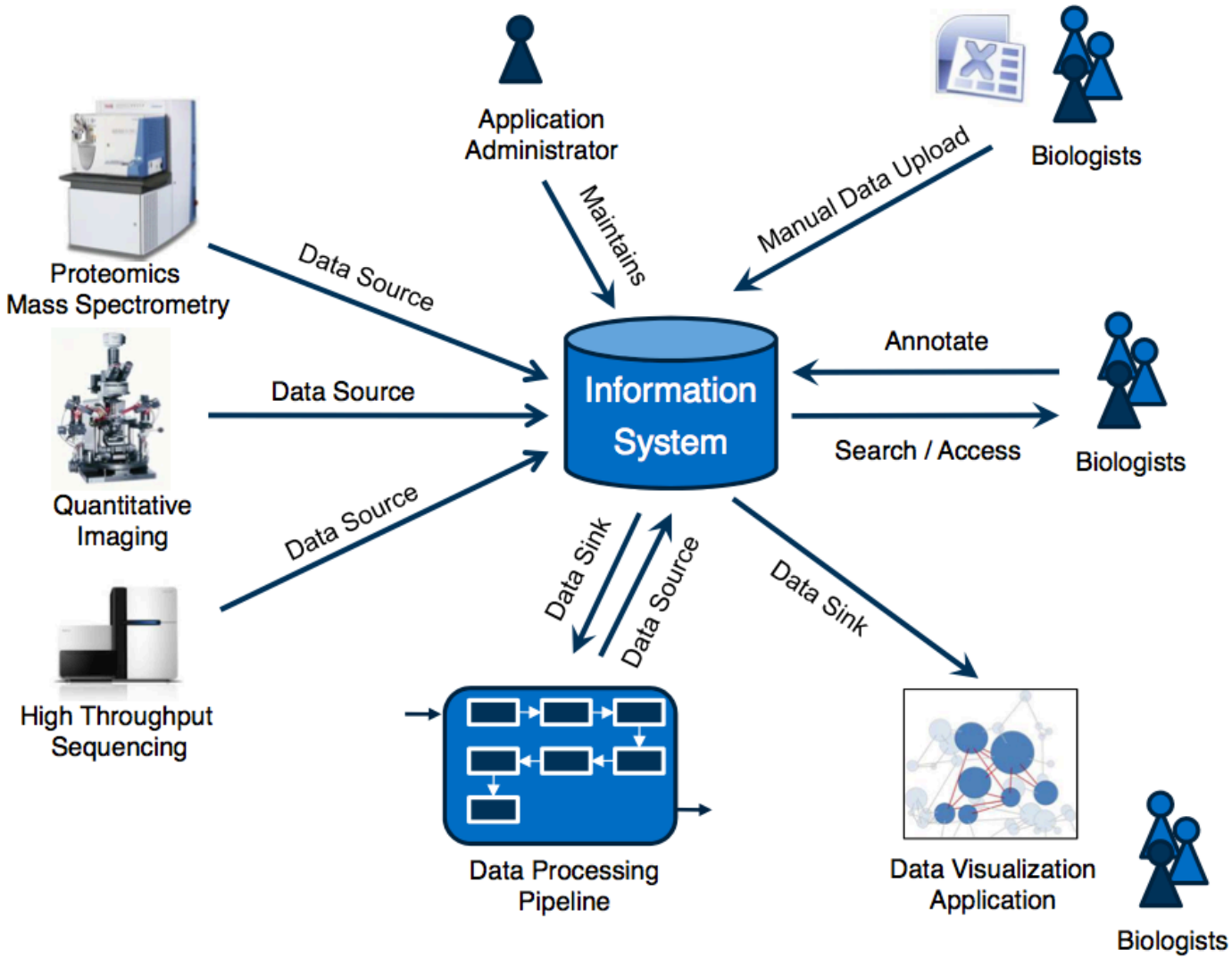
- Meta-data
- Files
- Databases

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

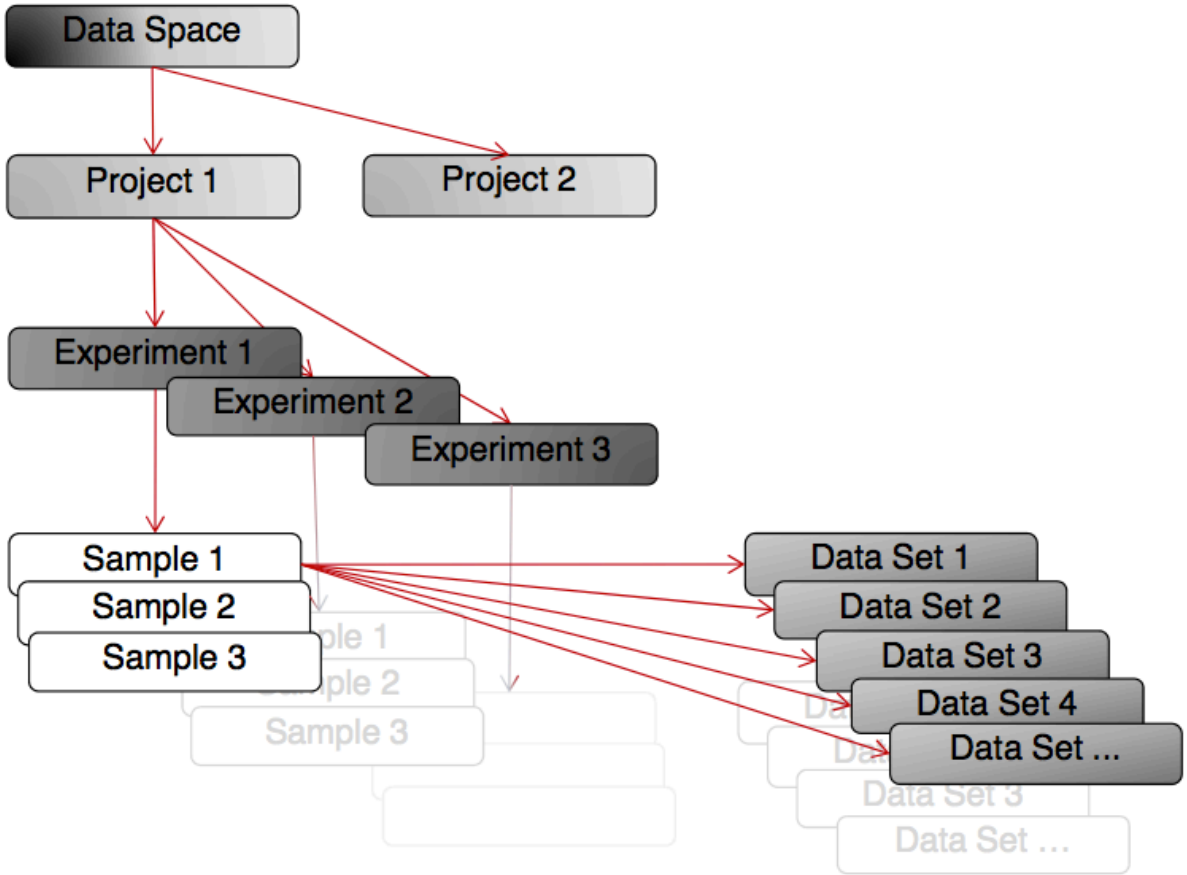
Use case: Proteomics Portal

- Large-scale Mass Spectrometry data analysis
  - Peptide and Protein identification
  - Relative and absolute quantification
  - Targeted proteomics
  - Ability to rerun analysis with same or new parameters
- New instruments produce O(TB)/week/lab
  - We have several labs to support
- Data files tightly controlled
  - Users cannot access original data, only copy them out
  - Data management and metadata stored in **OpenBIS**

Open Biology Information System openBIS



Spaces, Projects, Experiments, Samples & Data Sets



Data Space:
Visibility rules

Project:
A plan for a number of experiments

Experiment:
Empirical approach to acquiring data represented by a set of protocols

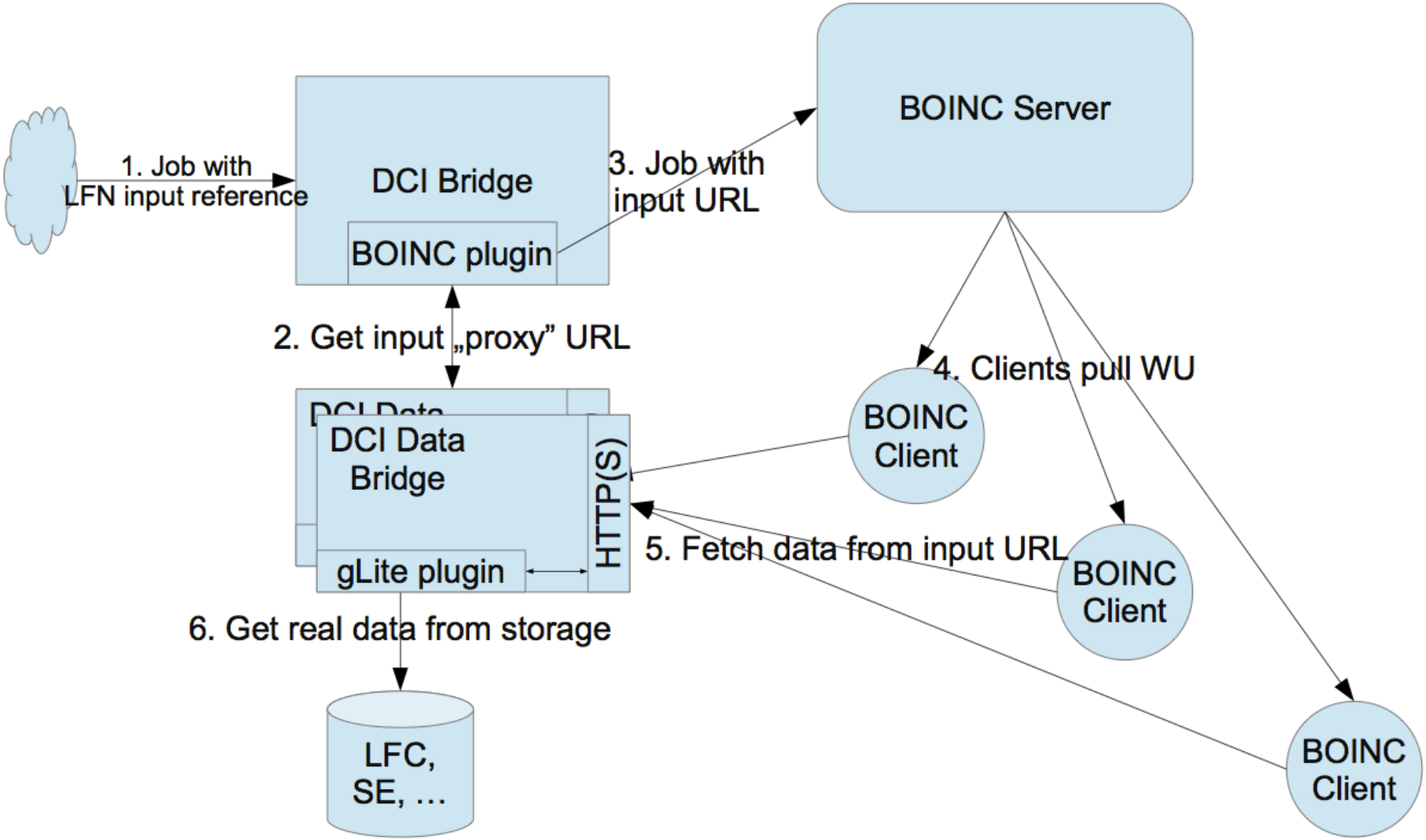
Sample:
Object being observed/ measured and compared to each other

Data Set:
Collection of data containing the values of the actual observations/measurements

Use case: Desktop Grid

- Parameter sweep execution (sometimes even 1M+ job instances)
- Typically small input and output files, available on HTTP(S) servers
- Sometimes with some fixed input files (i. e. the same file is used by each job)
- Different types of clients operating in pull mode:
  - Windows
  - OS X
  - Linux

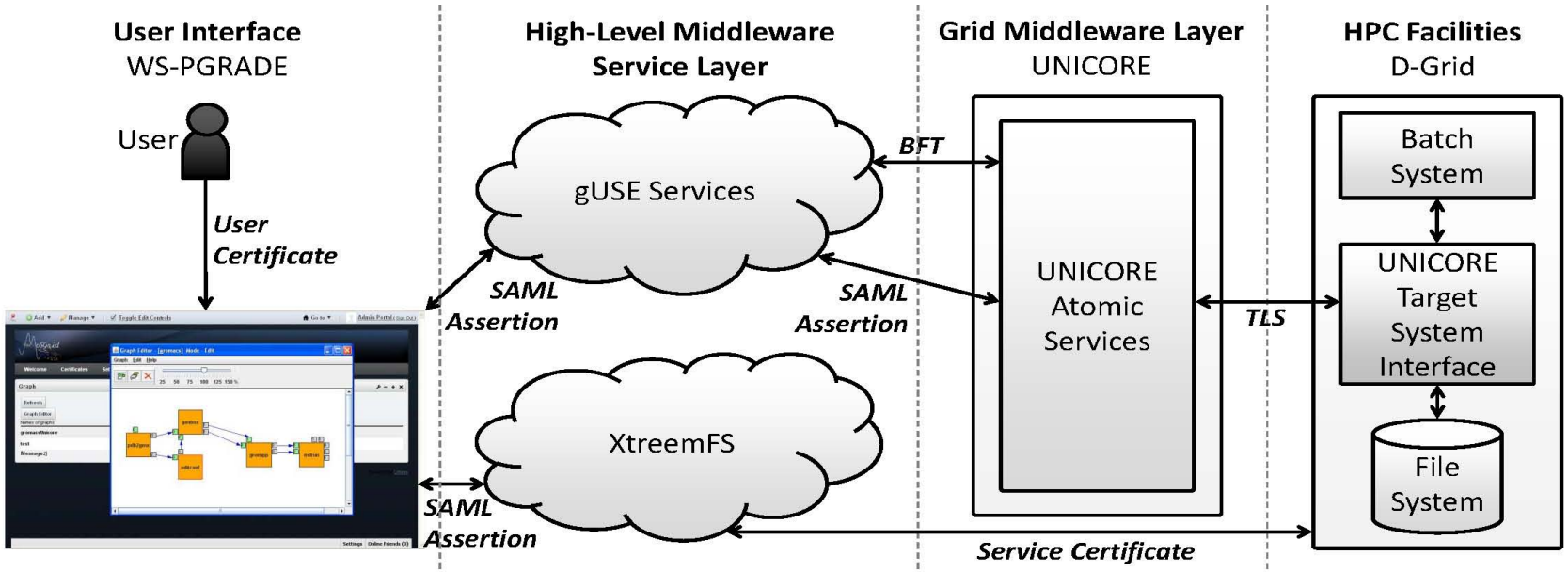
Desktop Grid



Use case: Molecular Simulation

- Multiple different domains e.g. quantum chemical, molecular dynamics, …
  with different formats and requirements
- Typically small input files (molecular structure and job description) generating huge output (number of files and data volume)
- Metadata annotation of simulation results

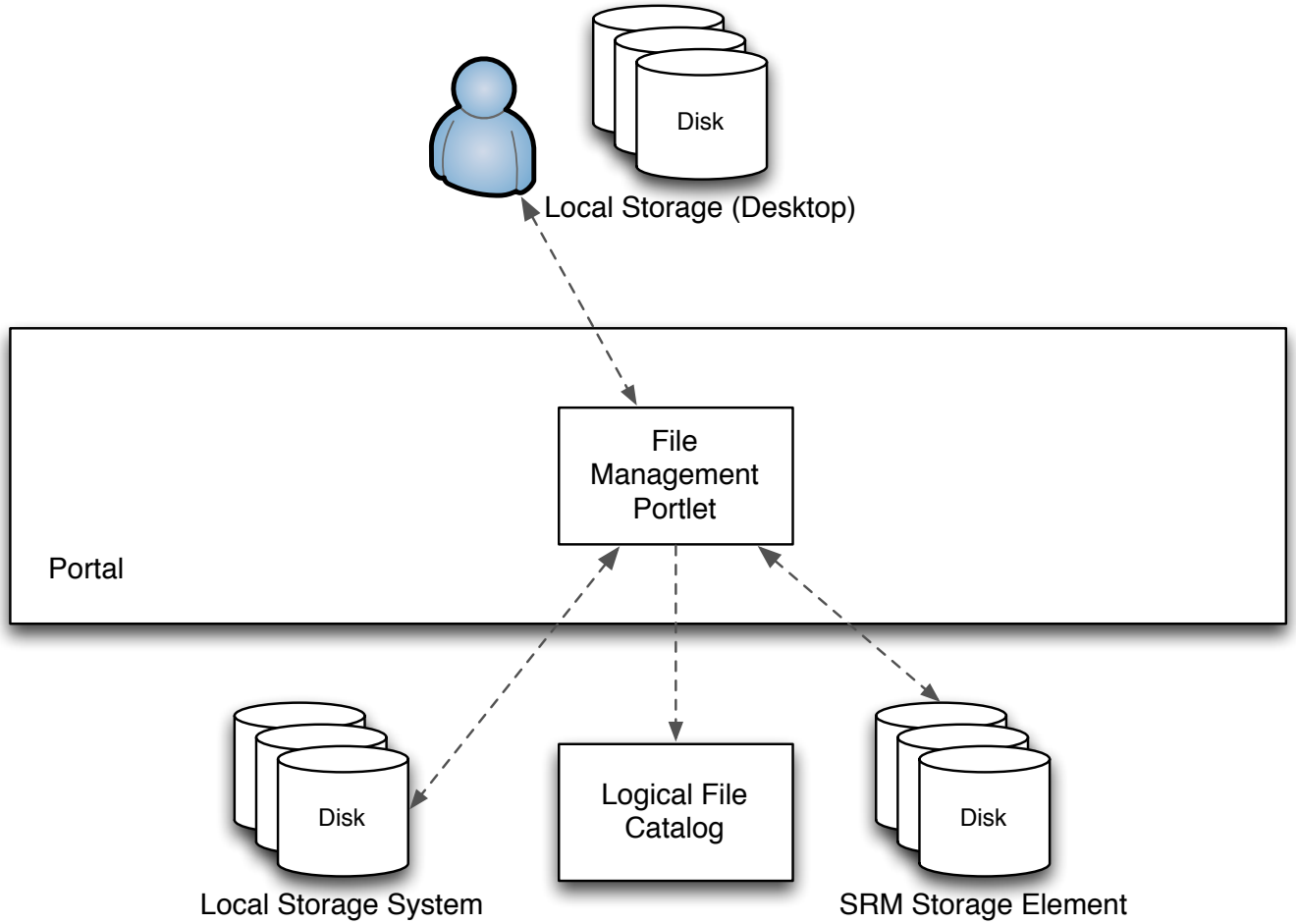
Mosgrid


User Interface
WS-PGRADE
High-Level Middleware
Service Layer
Grid Middleware Layer
UNICORE
HPC Facilities
D-Grid
User
User
Certificate
gUSE Services
BFT
SAML
Assertion
SAML
Assertion
UNICORE
Atomic
Services
TLS
Batch
System
UNICORE
Target
System
Interface
File
System
XtreemFS
SAML
Assertion
Service Certificate

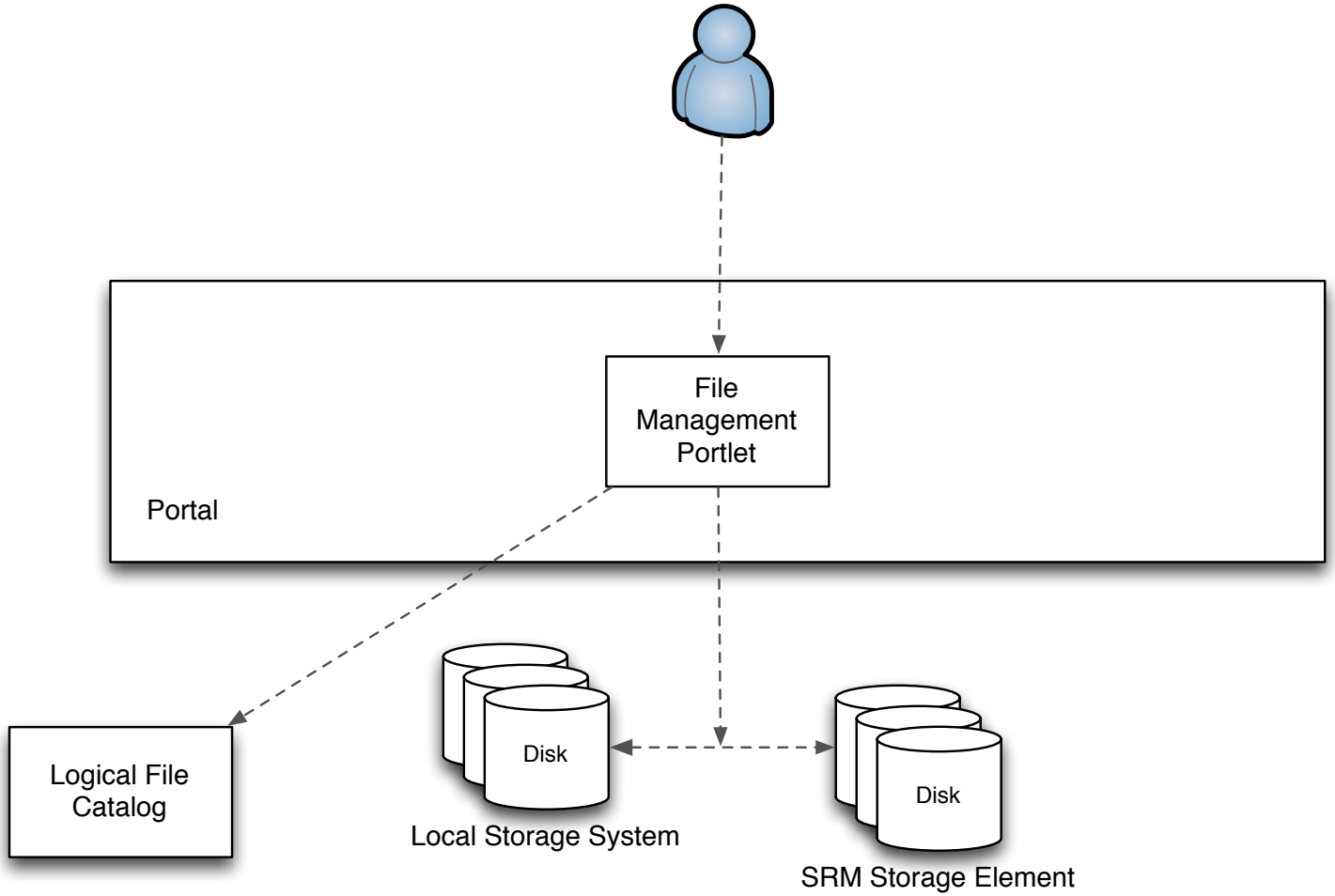
Use case: AMC

- Biomedical workflows
- Input data exist:
  - on various locations
  - accessible through different protocols
- Mixed use of compute resources based on application
- Track provenance

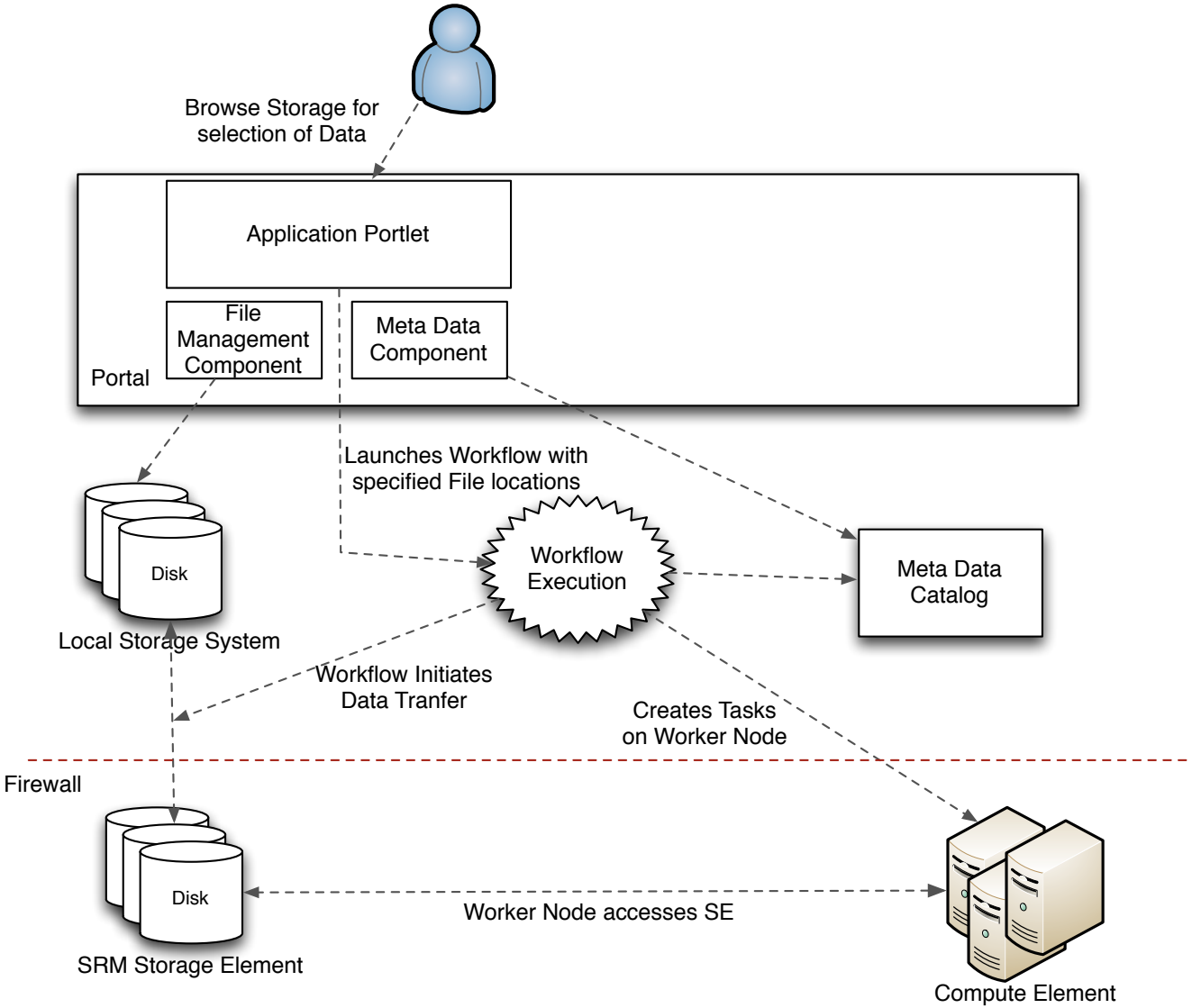
File Management



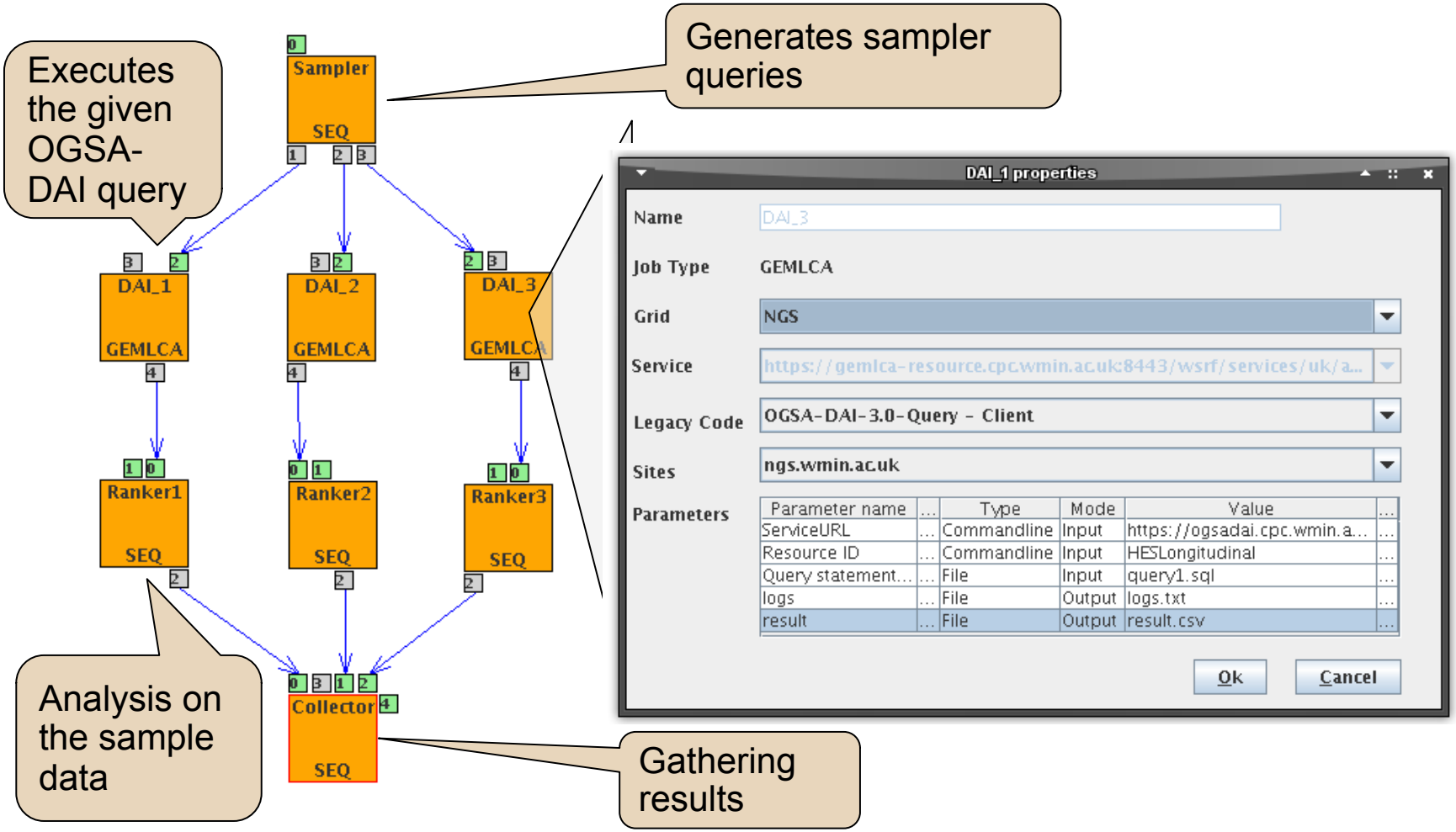
File Management (3rd party)



Meta-data and staging



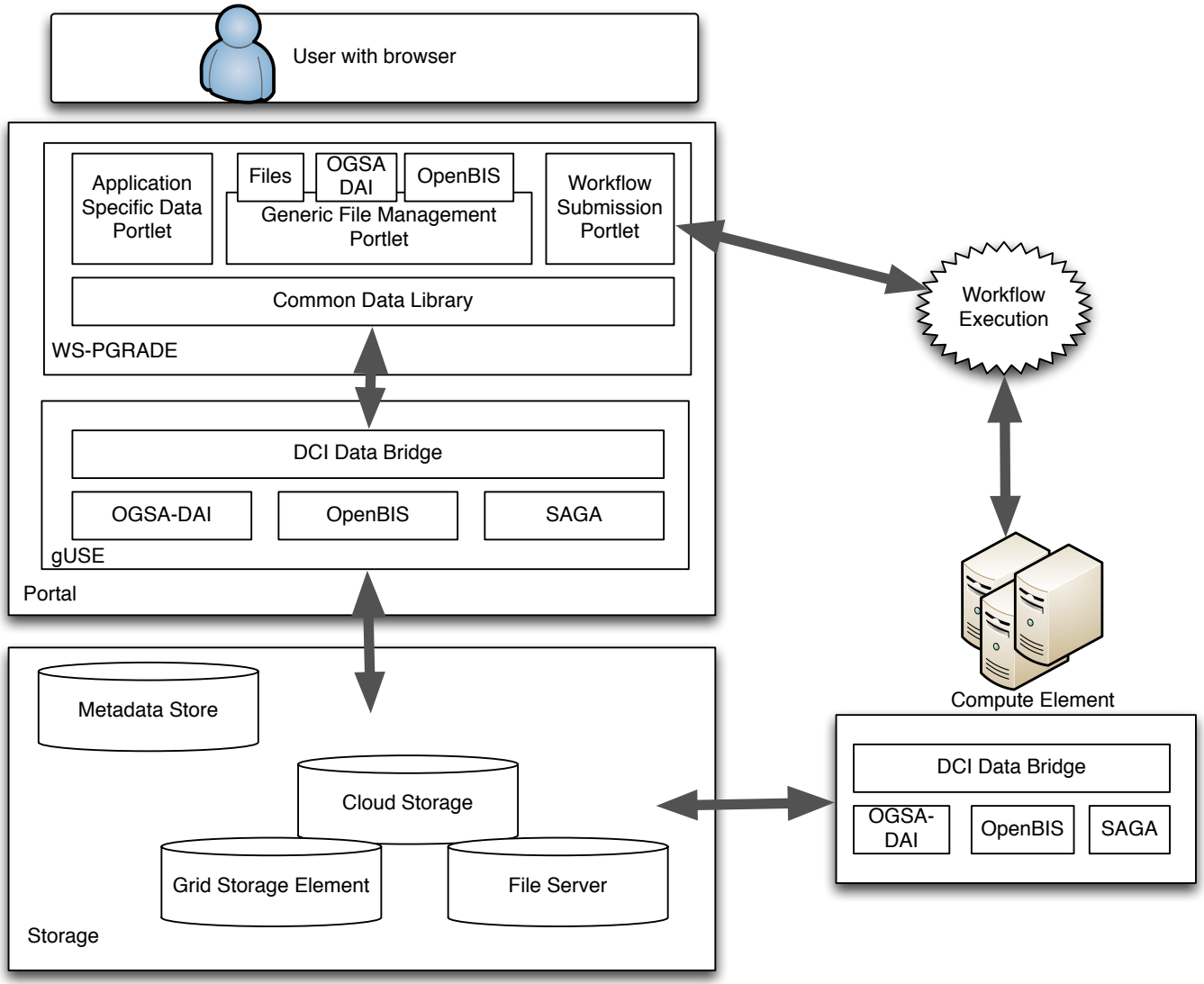
Use case: Performance rating framework for UK hospitals



Goals

- Generic layer
- Application-specific layer
- Workflow input selection
- Workflow execution

Architecture (WIP)



Status and plans

- Analyzing requirements from communities
- Exploring technologies in solution-space

- Define generic layer (but also non-generic)
- Implementation

- Develop application specific portlets

Questions and discussion

Visit the SCI-BUS booth!