d4science

EGEE08 Conference

Scientific Data Infrastructure Ecosystem

Wednesday 24 September, Istanbul, Turkey

Scientific e-Infrastructures: the D4Science perspective

Pasquale Pagano
CNR-ISTI
Pasquale.pagano@isti.cnr.it

George Kakaletris
University of Athens
gkakas@di.uoa.gr

SOI

Is a Virtualised IT Infrastructure which

1. exposes **a catalog of WS** instead of running service instances,
2. supports **SOA Application**, and
3. includes infrastructure resources such as **compute**, **storage**, and **networking** hardware and software (middleware) to support the running of services.

D4Science is building a production quality scientific e-Infrastructures (SOI) empowered by a collaborative environment

D4Science vision

calls for the realization of scientific e-Infrastructures that will remove technical concerns from the minds of scientists, *hide all related complexities from their perception*, and **enable users to focus on their science and collaborate on common research challenges**

gCube is

a framework to manage distributed **e-infrastructures** where it is possible to define, host, and maintain dynamic **virtual environments** capable to satisfy the collaboration needs of distributed **Virtual Organizations** (VOs)

The Technological Requirement

A Software framework

- to support ON-DEMAND virtual collaborations* among remote parties
  - cost-effective, secure,
  - dynamic, both **short** and **long** lived
  - overcome ad-hoc systems alike
- to make discoverable and accessible
  - computing, storage,
  - data and service resources
- to promote and/or contribute to data and service integration

* Research Environment

The Technological Complexity

Software framework

- needs a ‘middleware’ (typically distributed)
- is open by definition
  - new resource types and/or new resource instances can be de/registered at any time

- is powerful if it supports application scope
  - the portion of the infrastructure in which a resource exists
  - the portion of the infrastructure in which a resource can act, operate, or has power or control
- is powerful if it supports sharing scope (controlled resource sharing)
  - machines, storage, data and services resources

The Technological Solution: gCube

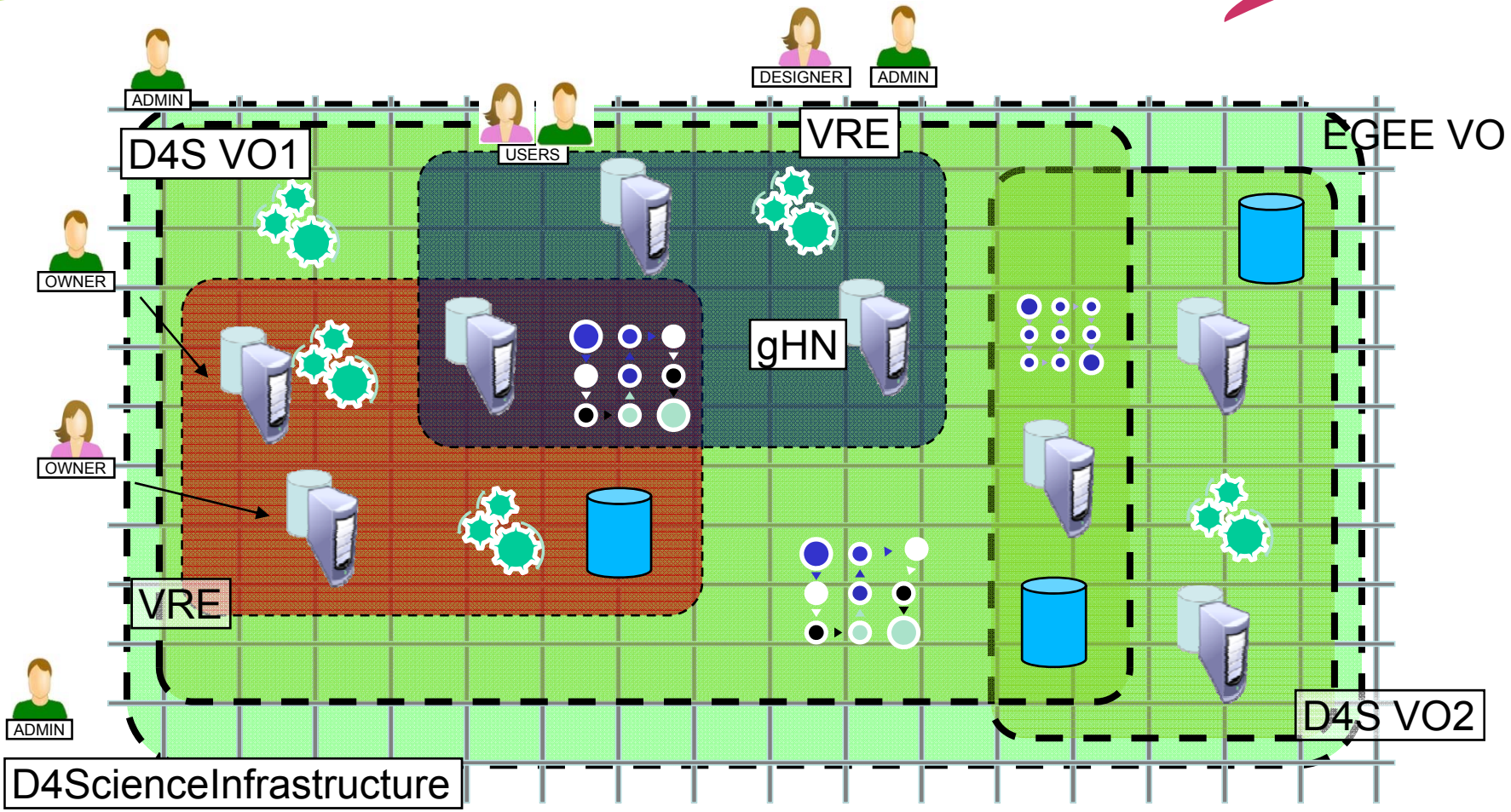
- By relying on gLite, gCube is an e-Infrastructure enabling system to share
  - computing, storage, data and service resources → g3
- gCube system allows collaborations in eScience
  - strongly content-oriented, potentially data and processing intensive
  - within the sharing scope of **Virtual Organizations** (VOs)
    - broader and longer lived
    - may stretch across the whole infrastructure
    - or else over significant subsets thereof
  - take place in **Virtual Research Environments** (VREs) scope
    - interactively created, managed, defined, and used: system administrators, application designers, researchers
    - typically short to medium lifespan

VRE applications are **designed**, **dynamically deployed**, and **operated** as a set of cooperating resources:

- grid resources (computing, storage)
- enabling services (the SOI middleware)
- VRE services
  - content and storage management, discovery and access, …
- standalone applications
  - mostly provided by the VOs
- collections of raw data, content, and metadata
  - enriched with schemas, mappings, transformation programs, relationships, …
- processes defined to manage such resources

- gCube services run onto **gCube Hosting Nodes** (gHNs)
  - **runtime containers** bound to logical ports
  - provide access to local hardware resources
    - storage, systems, instruments, CPU cycles...
  - **grant lifetime management**
  - mediate service2service interactions
    - route requests to target service
    - **enforce security and scope policies**

D4Science Infrastructure



D4Science VREs provide access to a workspace where users can

- share:
  - Private data
  - Data process results
  - Annotation
  - Process definition
  - Derived data
- collaborate to
  - define new processes,
  - tune applications and processes
  - compare execution results
- opens unique opportunities for virtual collaborations

Contain both objects owned by the workspace owner and objects the workspace owner has been allowed to see, e.g. *group* objects;

D4Science adds Service Management capabilities

- by providing solutions for system administrators to
  - eliminate manual deployment overheads,
  - eliminate manual environment configuration overheads,
  - ensure optimal placement of services within the infrastructure
  - support user community services orchestration
- opens unique opportunities for outsourcing state-of-the-art service implementations

Interoperability: Key concept in data management and processing facilities of D4Science Infrastructure

Data Management stages in D4Science VREs:

- Acquisition (import)
- Hosting (storage)
- Processing (transformation, …)
- Retrieval (search, …)

Information Object: the basic unit of payload stored in gCube data space.

Content: a document oriented Information Object

Metadata: Content describing other content

- *In the following no clear separation is assumed among content and metadata apart from the relationship "describes" that binds "metadata" to "content"*

Porting Content to a D4Science VRE: Challenges and principles

- Avoidance of any schema or format restrictions
- Exploitation the liberal gCube Information Model
  - Information Objects
  - Relationships
- Ability to handle standard and non-standard sources
  - OAI-PMH
  - XML-based sources
  - “file-system” like sources
  - Tabular data source (time-series, tables)
- Openness that matches the capacities of the gCube data processing pipeline and Information Retrieval services
  - “feeding” of corresponding entities

gCube Archive Import Service

- Provides language (AISL) for describing import operations
  - Resolves multiple levels of redirection for metadata / content relationships resolution
  - Offers extensibility for capturing future sources and import procedures
- Modular : supports pluggable components for access protocols, processing commands, and processing stages
- Allows partial / incremental imports
  - Supported by Storage Management Services
- Supports multiple protocols for accessing external sources : http, ftp, file, gridftp, etc (can be extended with plugins)

Storage Management Services

- Host payload (IOs) across storage providers
- Maintain low-level metadata and relations among IOs

Metadata Management Services

- Offers Value Added Services on top of SMS for metadata
- Employs XML for representation of metadata

Fundamental requirements for transforming content in D4Science

D4Science VREs raise requirements for

- Arbitrary transformation
- Homogenization

Of

- Content
- Metadata

For

- Information Retrieval
- Presentation
- Processing
- Exporting

Supporting

- Structured and semi-structured content
- Textual and binary content

Transforming Content in D4Science VREs: gDTS

Transforms Information Objects among formats

- Format = mimetype + detailed information

Captures cases of:

- XML to XML transformations (XSL/XSLT)
- Typical binary content transformations
  - Image, video, (audio), …
- Complex content transformations
  - Merging multiple sources to one new.

Special features:

- Detects formats automatically or guided
- Offers extensibility through pluggin-based architecture
- Performs automatic detection of transformation paths
- Exploits of grid capacities for processing
  - Jobs

VRE: Processing and Retrieval Requirements

- Process any kind of information (data, content, …) hosted in D4Science VREs
- Exploit the capacities of the distributed D4Science and Grid infrastructures
- Provide all typical (Distributed) Information Retrieval facilities
- Allow arbitrary, scientific domain-specific processing of data and information
- Allow sufficient customization and personalization for matching domain specific needs

Operation Principle

- Preprocess and transform queries into workflows of operations (operators) and execute the resulting graph by invoking appropriate services (operators) corresponding to nodes, passing data from node to node (via gRS)

Essential Elements

- Query analyzer and execution planner (Search Service)
- Distributed Information Retrieval components (DIR)
- Execution engine (Process Execution Service)
- Data processing nodes ((Search) Operators)
- Data transport (gRS)

The gCube Data Processing PipeLine (features)

Standards-based approach

- SOA / WSRF processing nodes
- BPEL Workflows
- Evolving to WS-DAI

Highly distributed, compliant with the grid paradigm

- Employs workers (operators) wherever appropriate based on optimization policy
- Employs the gRS mechanism for efficient data moves

Is schema agnostic

- No assumptions on metadata schemas
  - Projects among schemas for presentation and qualified searches, if needed
  - Application-level configuration steers the processing pipeline
- Common schemas improve performance

Supports several IR methods

- Geo-temporal lookups (with ranking capacities)
- Full text searches
- (semi) Qualified searches over XML metadata
- Similarity searches over binary content (though Feature Extraction and Indices)
- Tabular data lookups (incl. time-series)

Is extensible

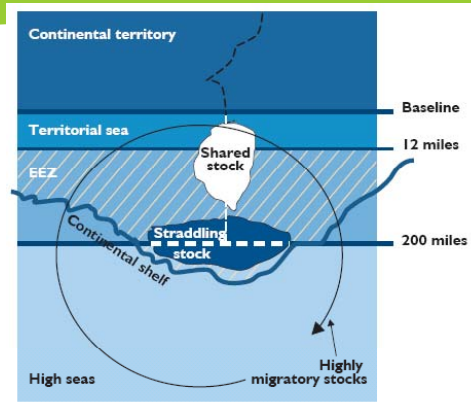
- Arbitrary search operator registration and invocation is supported
  - Though IS, profiles and gCube Query Language

Offers several other features

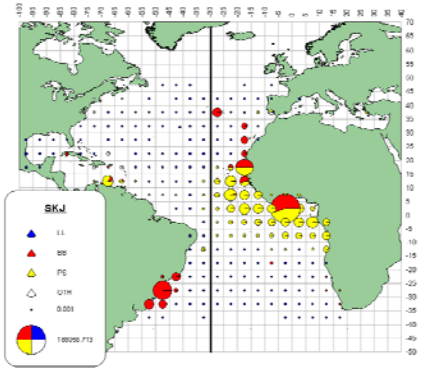
- Multi-linguality, Personalisation, …

Conclusion

- D4Science complements EGEE by providing support for new resource types
- D4Science reduces the costs to manage a complex multi-domain Service Oriented Infrastructure.
- D4Science offers an horizontal solution to manage and enrich on-demand created VREs on the EGEE infrastructure.
- D4Science is equipped with data and metadata management facilities that allows to make interoperable heterogeneous data sources.



www.d4science.eu



- gCube is compliant with consolidated and emerging standards.

- gCube offers an open family of frameworks that can be easily customised

- gCube is a working horizontal solution
  - composed by more than 200 software components
    - > 60 WSs, >50 independent libraries, and >30 portlets
  - most components widely tested and certified
    - Public and Stable Release (November 2008)

- **gCube Software Documentation**
  - https://technical.wiki.d4science.research-infrastructures.eu/
- **gCube Core Software Documentation**
  - https://wiki.gcore.research-infrastructures.eu/
- **gCube System web site**
  - http://www.gcube-system.org

gCube & Standards for communication

- gCube exploits the Java WSCore, Apache Axis, GridForum specifications (and implementation if any):
  - WS-Notification, WS-Addressing, WS-Security,
  - WSRF
    - WS-ResourceProperties (WSRF-RP)
    - WS-ResourceLifetime (WSRF-RL)
    - WS-BaseFaults (WSRF-BF)
    - WS-ServiceGroup (WSRF-SG)
  - WS-DAI, WS-DAIR, WS-DAIX

gCube & Standards for communication

- Mutual authentication based on GSI secure conversation (through delegation and renewal)
- Business Process Execution Language for Web Services (WS-BPEL)
- GridFTP and SRM support
  - VOMS for users and groups management
  - GWT and JSR168(JSR268 is coming)

gCube Standard web page:
https://quality.wiki.d4science.research-infrastructures.eu/quality/index.php/Standards

gCube & Standardization Bodies

- ISO: data representation (e.g. ISO3166 for countries, ISO4217 for currencies) and metadata (ISO19115 for GIS)
- OGF: Standards related to Architecture (e.g. OGSA), Data (e.g. DAIS, GridFTP), Management (e.g. GLUE, Resources Usage), Applications (e.g. DRMAA), Compute (e.g. JSDL)
- OAI: Resources Exposure/Harvesting (OAI-PMH) Resources Exchange (OAI-ORE)
- OASIS: Standards related to stateful web services (e.g. WSRF), process management (BPEL), remote user interfaces (WSRP), A&A (SAML / XACML)
- W3C: All the standards related to Web Architecture (e.g., URI/URL, HTTP), Service Oriented Architectures (e.g. SOAP, WSDL, WS-Addressing) and data representation and manipulation(e.g. XML*)
- Others: Classification systems (e.g. ISSCAAP, ISSCFV, ISSCFG), features representation (e.g. GML for GIS), metadata (e.g. AgMES for Agricultural, SDMX for Statistics)