EFITA WCCA CIGR
2013 Conference TORINO, ITALY
Sustainable Agriculture through ICT innovation

Federating Computation and Storage Resources to support Agricultural Science Communities

Jaakko Lappalainen[1], Dusan Vudragovic[2], Antun Balaz[3], Federico Ruggieri[4], Roberto Barbera[5], Robert Lovas[6], Giannis Stoitsis[7], Kostas Kastrantas[8]

[1] jkk.lapp@uah.es, Computer Science Department, University of Alcalá, Polytechnic Building, Ctra. Barcelona km. 33.6. 28871 Alcalá de Henares (Madrid), SPAIN
[2, 3] Institute of Physics Belgrade
[4, 5] Istituto Nazionale di Fisica Nucleare
[6] Hungarian Academy of Sciences
[7, 8] AgroKnow Technologies

ABSTRACT

Agricultural sciences are one of the largest and most significant examples of Big Data research communities. The development of this research field brings about demands for larger, more powerful and trustable computing infrastructures to support the work of agricultural scientists. On the other hand, Cloud and Grid technologies have emerged in the last decade as a way to integrate large-scale data infrastructures. In the present work, we describe a federation of large computation and storage infrastructures developed within the agINFRA project, and the interfaces to use them. The new data infrastructure and its interfaces allow different user profiles to access and take advantage of applications and services to manage, process, navigate, and visualize trustable information on agricultural science topics.

Keywords: Grid computing, cloud computing, federation, interfaces, big data.

1. INTRODUCTION

Agricultural sciences are one of the largest and most significant examples of Big Data research communities in growth the past decade. As the size of our knowledge grows and the science becomes more and more interconnected, the complexity of the scientific

J. Lappalainen, D. Vudragovic, A. Balaz, F. Ruggieri, R. Barbera, R. Lovas, G. Stoitsis, K. Kastrantas. "Federating Computation and Storage Resources to support Agricultural Science Communities". EFITA-WCCA-CIGR Conference "Sustainable Agriculture through ICT Innovation", Turin, Italy, 24-27 June 2013. The authors are solely responsible for the content of this technical presentation. The technical presentation does not necessarily reflect the official position of the Internation Commission of Agricultural and Biosystems Engineering (CIGR) and of the EFITA association, and its printing and distribution does not constitute an endorsement of views which may be expressed. Technical presentations are not subject to the formal peer review process by CIGR editorial committees; therefore, they are not to be presented as refereed publications.

inquiry increases. Also, science is becoming largely digital - it needs to deal with ever increasing amounts of data and computational power. To engage the technological challenges of the scientific practice, it is necessary to efficiently build computing compounds that bring data analysis and processing to a new level.

The agINFRA data infrastructure aims to enhance the co-operation, data sharing, federation and data exchange between research communities in the agricultural sciences by providing a robust new data infrastructure to compliment the building of trust and federation between agricultural researchers, enhancing the field and creating new opportunities for joint work and research sharing.

2. THE AGINFRA INFRASTRUCTURE

AgINFRA has been conceived as a sustainable data infrastructure for agricultural research. The project addresses implementation of services relevant for a wide range of users: from infrastructure providers, developers of data processing and data management software, data providers, information managers, librarians, developers of high-level portals, to end-users (researchers, educators, citizens). Design of the infrastructure relies on adaptation of the existing resources, their customization, and the development of new components that will ensure balance in the system and its seamless use.

2.1 Science Gateways

On top of the Grid and Cloud infrastructure, the layer of Science gateways (SG) provides seamless access to the distributed infrastructure for software developers. Two different Science gateway architectures are deployed within the agINFRA project. Together, they address different usage patterns and requests, and complement each other in providing advanced grid access mechanisms.

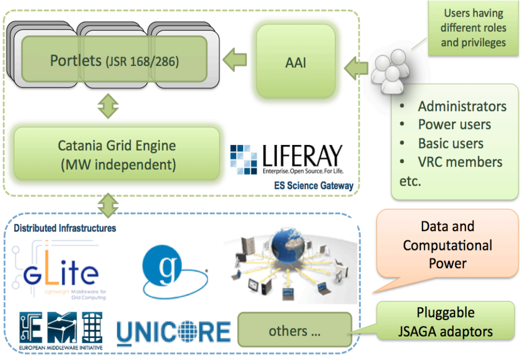
Portlets (JSR 168/286)
AAI
Users having different roles and privileges
Administrators
Power users
Basic users
VRC members
etc.
Catania Grid Engine (MW independent)
LIFERAY
Enterprise. Open Source. For Life.
ES Science Gateway
Distributed Infrastructures
gLite
Lightweight Middleware for Grid Computing
EMI
EUROPEAN MIDDLEWARE INITIATIVE
UNICORE
others ...
Data and Computational Power
Pluggable JSAGA adaptors

Figure 1. Catania SG Architecture

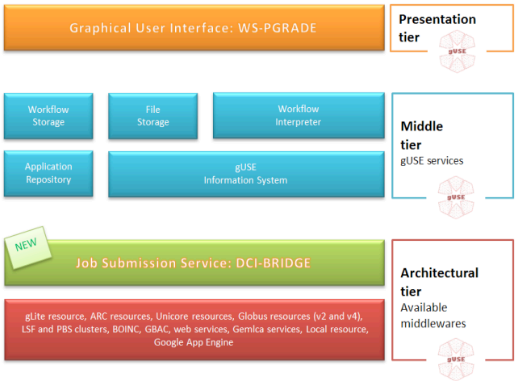
Graphical User Interface: WS-PGRADE
Presentation tier
gUSE
Workflow Storage
File Storage
Workflow Interpreter
Application Repository
gUSE Information System
Middle tier
gUSE services
gUSE
NEW
Job Submission Service: DCI-BRIDGE
gLite resource, ARC resources, Unicore resources, Globus resources (v2 and v4), LSF and PBS clusters, BOINC, GBAC, web services, Gemlca services, Local resource, Google App Engine
Architectural tier
Available middlewares
gUSE

Figure 2. gUSE architecture

The Catania Science Gateway, shown in Figure 1, provides a full-featured environment accessing with single software architecture to many distributed infrastructure resources. It also allows creating high-level user interfaces able to submit and control Grid jobs, manage data and metadata content, manage Cloud nodes, provide both secure and anonymous access to the Cloud hosted services and more in general avoid its user community to deal directly with the technicalities and the complexities of each specific distributed infrastructures.

SZTAKI has developed the gUSE (grid User Support Environment), shown in Figure 2, which is a grid virtualization environment providing a scalable set of high-level Grid services by which interoperation between Grids and user communities can be achieved. Incorporating a more flexible workflow concept and enabling its distribution on clusters and different Grid sites, gUSE is aimed to extend the objectives and features of P-GRADE Portal based on Liferay. SZTAKI is currently developing a generic-purpose gateway technology as a toolset to provide seamless access to major computing, data and networking infrastructures and services in Europe including clusters, supercomputers, and grids, academic and commercial clouds.

The SZTAKI SG implementation focuses on the management of complex job workflows, while the Catania SG implementation focuses on user membership management through the adoption of the Identity Federations and Identity Providers approach, allowing the single sign-on across many supported institutions, but also offering a vertical interface to all user levels.

2.2 AGINFRA STORAGE BACKEND

In parallel to the two general-purpose SG implementations, agINFRA RESTful gateway[9] is provided as well, as a third agINFRA layer. This gateway is developed within the project, and its main responsibilities are cataloging, off-line processing, and management of data. As a catalog facility, agINFRA RESTful gateway keeps user's configurations of Grid-ported applications, datasets' metadata information and locations, arbitrary user-defined metadata information, and allows querying of these information.

The gateway carries out automatic replication of datasets, ensures the existence of the same dataset on different storage systems, and its exposure through the HTTP protocol. Instead of perpetual reorganization of schemas, the database back-end has been migrated to the document-oriented data model. Due to several additional features, such as offline replication, Multi-Version Concurrency Control, incremental replication, fault-tolerance, we have decided to replace MySQL with CouchDB[10] technology. Existing tables and rows are reorganized into CouchDB JSON documents that could be now directly transferred via REST API incorporated in CouchDB.

[9] agINFRA RESTful gateway documentation, http://agro.ipb.ac.rs/
[10] CouchDB official web site, http://couchdb.apache.org/

2.3 Public Clouds

Some of the proposed agINFRA services require an on-demand delivery of service with a sort of interactive response such as agriDrupal and Omeka data components. The grid was initially designed to support long-term job runs, rather than online response capabilities. The integrated approach of agINFRA, requires also that the services (on Virtual Machines) should be available with an user friendly access to the infrastructure. For these reasons, agINFRA is able to connect with Cloud infrastructures, such as CLEVER (TUSA et al., 2010) and OKEANOS (KOUKIS et al. 2013).

3. LAYERED ARCHITECTURE

The conceptual architecture of agINFRA is depicted in Figure 3. On the top of the infrastructure lay the integrated services, existing portals that provide to end-users with high quality information on a specific topic or sub-domain of an agricultural science. Considered integrated service examples in agINFRA are Organic.Edunet (MANOUSELIS et al., 2009) portal, AGRIS (SUBIRATS et al., 2008) and CIARD RING (PESCE et al., 2011).

In a layer conceptually below integrated services are the data components, software packages that provided advanced data processing, management, visualization and navigation capabilities to users and integrated service managers.

If we look at the grid infrastructure conceptually in terms of APIs, the Figure 3, depicts its main elements. Web-based components and integrated services use APIs to access the infrastructure service layer.

In turn, these services use existing generic Grid resources. It should be noted that here the term “API” is used in a broad sense, as it includes any kind of interface exposed by the infrastructure, including REST APIs, exposure in the form of RDF LD, as well as any other mechanism for exposing data management and processing services.

In Figure 5 there is a detail of the agINFRA infrastructure showing how IS and data components can access two kinds of services Conceptually, there are four subsystems inside the infrastructure that fulfill different missions. The functional data flow and interactions with data components is shown in Figure 5.

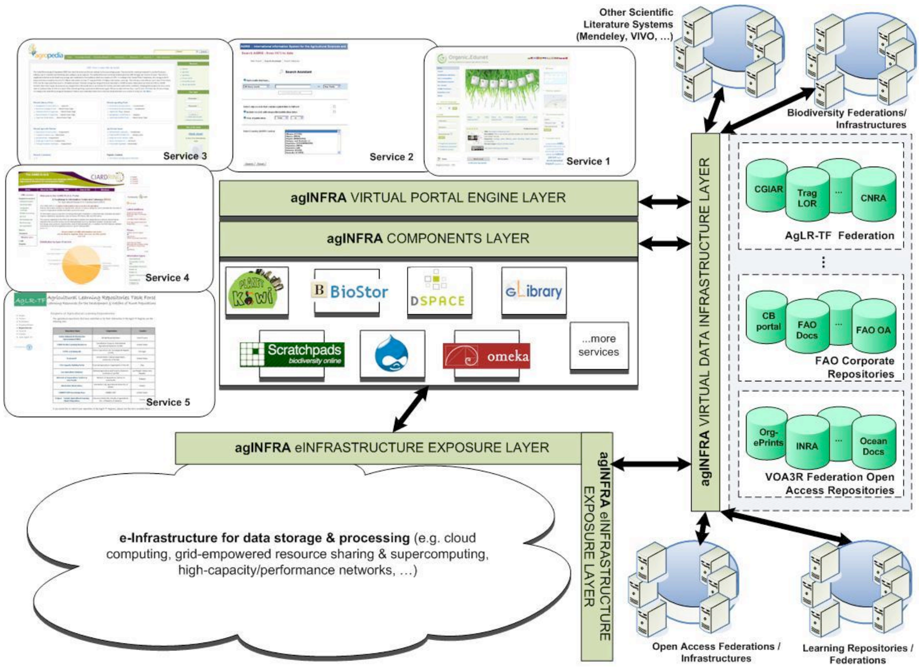
Other Scientific Literature Systems (Mendeley, VIVO, ...)
Biodiversity Federations/ Infrastructures
Service 3
Service 2
Service 1
Service 4
Service 5
agINFRA VIRTUAL PORTAL ENGINE LAYER
agINFRA COMPONENTS LAYER
BioStor
DSPACE
GLibrary
Scratchpads biodiversity online
omeka
...more services
agINFRA VIRTUAL DATA INFRASTRUCTURE LAYER
CGIAR
Trag LOR
...
CNRA
AgLR-TF Federation
CB portal
FAO Docs
FAO OA
FAO Corporate Repositories
Org-ePrints
INRA
Ocean Docs
VOA3R Federation Open Access Repositories
agINFRA eINFRASTRUCTURE EXPOSURE LAYER
agINFRA eINFRASTRUCTURE EXPOSURE LAYER
e-Infrastructure for data storage & processing (e.g. cloud computing, grid-empowered resource sharing & supercomputing, high-capacity/performance networks, ...)
Open Access Federations / Infrastructures
Learning Repositories / Federations

Figure 3. Conceptual infrastructure

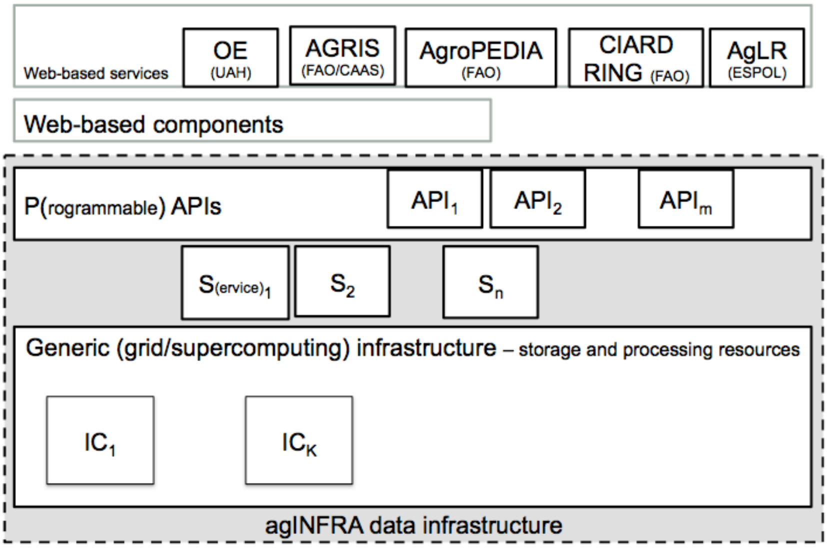

Figure 4. Layered view

The Data Management layer (DML) is responsible for the creation, update and management of datasets, including their metadata and actual contents when required. The DML offers a Dataset API externally and also to the rest of the internal elements, and acts as a central repository of datasets, its versions and derivatives. The Data Processing Layer (DPL) carries out the execution of offline jobs over the data. These jobs are scheduled using grid middlewares and provide users with non-interactive powerful computation services. The Data Transformation Layer (DTL) deals with metadata schema transformation services. These functions are key to interlink datasets and ensure interoperability of data. This adds an important value to the available content. Finally, the Data Exposure Layer (DEL) is in charge of exposing content and metadata in LD, allowing machines to consume high-quality integrated data and services.

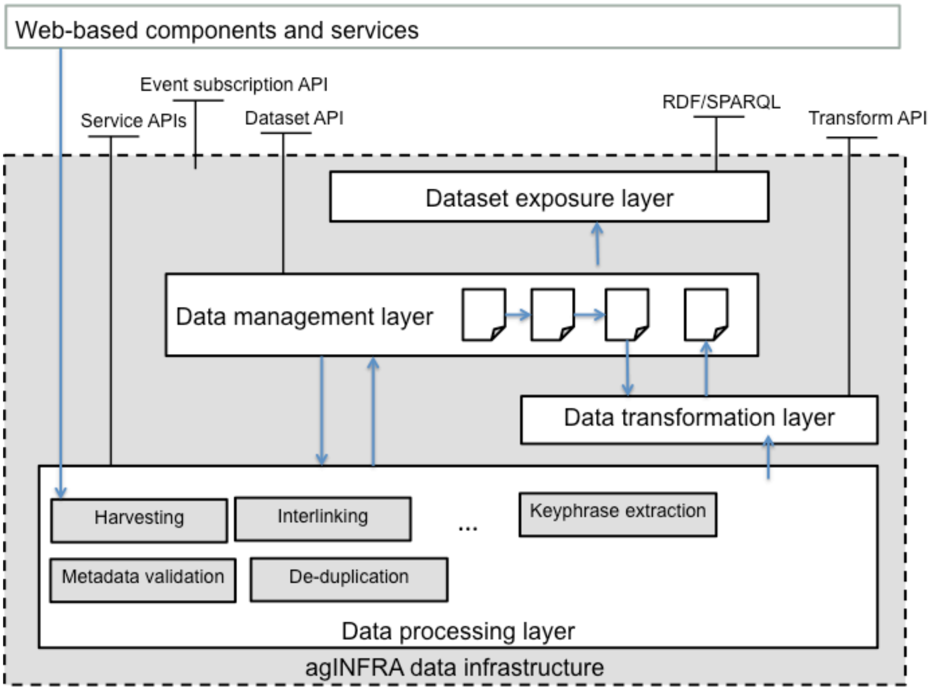

Figure 5. Functional data flow

4. RELATED WORK

Among the wide range of agINFRA related projects, the iPlant Collaborative (GOOF et al., 2011) is one of excellent examples of e-Infrastructures designed to solve challenges, here in particular for plant science. It successfully merges the plant biology tools and services for data analysis, mining and visualization, workflow and education resources. The project develops applications to support remote collaborations and virtual organizations.

The agINFRA science gateways provided by INFN and SZTAKI are an analogy to the iPlant DE. The science gateway technologies provide various modules for web-based graphical user interface creation, user-friendly data management, visual workflow construction, and application deployment in the Grid and HPC environments. In addition, science gateway frameworks, through Grid technology, supply authentication and authorization mechanism that naturally enables easy integration of geographically distributed new computing and storage resources. Furthermore, existing Grid components afford high-level meta-scheduling, workflow execution, task tracking, data and metadata catalogues, replication of data, delegation and renewal of credentials, accounting functionality, etc.

Environment provided by iPlant Atmosphere could be compared with gLite user interface machine (UI). Currently, in agINFRA UI machines can be used remotely at the infrastructure partners' resource centers through command line interfaces, and in some cases by VNC. The other possibility for a user is to install gLite-UI VM image locally at a personal computer. Such gLite-UI image is provided by INFN, and is available for download. Beside the standard gLite interface, this VM image can include particular agINFRA applications and tools. In both cases, user is provided with application programming interfaces (APIs), for software development or for specific data analysis. Furthermore, user is able to access to all agINFRA computing and storage resources, to use all previously installed applications, and to create various tasks workflows.

Under the scope of agINFRA, the SemaGrow[11] project aims to develop data intensive techniques to boost the real-time performance of global agricultural data infrastructures, achiving scalability when querying distributed and big data agricultural data repositories.

5. CONCLUSIONS

This paper describes the technical achievements of the project agINFRA. We have shown how existing technologies can be customized and integrated to set up the foundations to build an advanced data infrastructure to support agricultural science communities. This data infrastructure has been conceived as a set of layers that group different functionalities to manage datasets. These layers provide services that process, retrieve, store, transform and publish heterogeneous data easily.

6. REFERENCES

Foster, I., & Kesselman, C. (1999). The Globus project: A status report. *Future Generation Computer Systems*, *15*(5), 607-621.

Gagliardi, F., Jones, B., Reale, M., & Burke, S. (2002). European DataGrid Project: Experiences of deploying a large scale Testbed for e-Science applications. In *Performance Evaluation of Complex Systems: Techniques and Tools* (pp. 480-499). Springer Berlin Heidelberg.

Gagliardi, F. (2005). The EGEE European grid infrastructure project. In *High Performance Computing for Computational Science-VECPAR 2004* (pp. 194-203). Springer Berlin Heidelberg.

Goff, Stephen A. et al., "The iPlant Collaborative: Cyberinfrastructure for Plant Biology," *Frontiers in Plant Science* 2 (2011), doi: 10.3389/fpls.2011.00034.

[11] The SemaGrow project: http://www.semagrow.eu

Koukis, V., Venetsanopoulos, C., & Koziris, N. (2013). ~ okeanos: Building a Cloud, Cluster by Cluster. *Internet Computing, IEEE*, *17*(3), 67-71.

Manouselis, N., Kastrantas, K., Sanchez-Alonso, S., Cáceres, J., Ebner, H., & Palmer, M. (2009). Architecture of the organic. edunet web portal. *International Journal of Web Portals (IJWP)*, *1*(1), 71-91.

Martin-Flatin, J. P., & Primet, P. V. B. (2005). Guest Editorial: High-speed networks and services for data-intensive Grids: The DataTAG Project. *Future Generation Computer Systems*, *21*(4), 439-442.

Pesce, V., Maru, A., & Keizer, J. (2011). The CIARD RING, an Infrastructure for Interoperability of Agricultural Research Information Services [Article and Abstract]. *Agricultural Information Worldwide*, *4*(1), 48-53.

Subirats, I., Onyancha, I., Salokhe, G., Kaloyanova, S., Anibaldi, S., & Keizer, J. (2008). Towards an architecture for open archive networks in agricultural sciences and technology. *Online Information Review*, *32*(4), 478-487.

Tusa F., Paone M., Villari M. and Puliafito A., "CLEVER: A CLoud-Enabled Virtual EnviRonment", Proceedings of the 15th IEEE Symposium on Computers and Communications, 22 - 25 June 2010, Riccione, Italy.