EGI Applications On Demand Service

Catering for the computational needs of the long tail of science

Gergely Sipos
EGI Foundation
Amsterdam, Netherlands
gergely.sipos@egi.eu

Giuseppe La Rocca
EGI Foundation
Amsterdam, Netherlands
giuseppe.larocca@egi.eu

Diego Scardaci
INFN/EGI Foundation
Catania, Italy
diego.scardaci@egi.eu

Peter Solagna
EGI Foundation
Amsterdam, Netherlands
peter.solagna@egi.eu

***Abstract*—This paper describes the new EGI 'Applications on demand service', that the EGI collaboration specifically designed for individual researchers, small research teams and early phase research infrastructures that do not have dedicated computational and storage resources, online applications and science gateways to perform scientific data analysis. The described service is available at http://access.egi.eu and through a lightweight registration and user identity vetting process allows user-friendly access to a growing number of scientific applications and application hosting frameworks (science gateways, VREs) that are configured to use the dedicated pool of cloud computing and HTC clusters from EGI. The service operates as an open and extensible 'hub' for providers and e-infrastructure user support teams who wish to federated and share applications and services with individual researchers, or small, fragmented communities, typically referred to as 'the long tail of science'.**

Keywords—long-tail; SaaS; e-infrastructures; cloud; HTC/HPC

I. INTRODUCTION

EGI is an e-Infrastructure collaboration that provides advanced computing and data services for research and innovation. The collaboration operates a federated, publicly-funded e-infrastructure that currently comprises more than 300 resource centers from Europe and beyond. Over the last decade this infrastructure was the enabler of digital research conducted by over 40,000 researchers through the whole spectrum of science from High-Energy Physics, to Earth Sciences, Life Sciences, Chemistry, Astrophysics, and Humanities.

EGI resource centers rely on the expertise of the 'EGI Foundation', EGI's coordinating institute located in Amsterdam. EGI members – national compute/data centers (so called NGIs) and Intergovernmental Research Organizations, such as CERN – operate the compute, storage, application and software services that comprise the 'EGI infrastructure'. These compute/storage providers allocate resources to scientific communities via 'Virtual Organisations' (VOs), according to institutional or regional/national priorities. A VO is the online representation of a scientific user group whose members are usually work in similar or related research areas, or are part of the same scientific collaboration, for which reason they need the same applications, software and underlying hardware capabilities. Some of the biggest VOs of EGI represent experiments of the Large Hadron Collider (ALICE, ATLAS, CMS, LHCb) [2], the VIRGO experiment [3], the Cerenkov Telescope Array Observatory [4] or life science researchers from multiple countries and diverse background (biomed VO) [5].

Since its start in 2010, EGI has well defined processes to create and operate VOs for large, structured, international user communities. These communities have well established and long-term presence, moreover they are resourced well enough to sustain skilled IT support teams who can instantiate and operate VO services for the researchers. The most advanced research infrastructures from the ESFRI Research Infrastructure Roadmap [6] are the typical operators of the largest VOs in EGI.

At the same time individual researchers, small research teams and members of early phase research infrastructures often struggle to access applications, compute and data services in EGI. The tools and resource allocation policies that were designed for long-living, structured communities were recognized as unsuitable for these type of users because they typically involve:

- Obtaining and using an X.509 personal digital certificate from Certification Authority (CA) recognized by EGI.
- Joining an existing VO that matches the requester's research subject/goals, or setting up a new VO.
- Integrating relevant scientific applications with the VO and fulfilling operational responsibilities, such as VO

membership management, resource allocation negotiations and community/application specific service monitoring.

These individuals are often referred to as the 'long tail of science' [1] and they share the common characteristic of (1) missing dedicated arrangement to computational and storage resources and online applications and services to manage and analyze large amount of data, and (2) lacking the skills and experience with deploying and scaling applications to distributed computer architectures. Despite some NGIs operate national services for the long tail of science, such national 'catch-all' VOs are not available or not well maintained any more in several cases [21].

Recognizing the needs of this unique group, and trying to compensate for the loss of NGI services in this domain we designed and developed a new service within EGI: the 'Applications on demand service'. The service was designed in late 2014, demonstrated in November 2015, reached Alpha release for early adopters in January 2017, and was opened for the general public in April 2017 as a Beta[1] service. The service is available at http://access.egi.eu.

This paper introduces this new service offering and details the technologies and components that were developed and integrated to realize the service. The rest of the paper is organized as follows: Section II introduces the user and provider requirements that drove the system development. Section III describes the overall architecture and the main design considerations. Section IV provides details about the different system components that were developed/customized from existing EGI tools to establish the service. Section V in a table the already integrated applications and components of the service. In Section VI we describe the steps that an applications hosting framework provider has to follow to make his framework accessible within the EGI Applications on Demand service. In Section VII we present our plans for the extensions and improvement of the service, while in Section VIII we draw conclusions from the described work.

II. Requirements

The design process for the new service started with a requirement collection and analysis that helped us to understand the needs/preferences of the long tail of science, as well as the constraints from the EGI resource/service providers' side that we have to consider as we are aiming at a sustainable service. These needs and constraints are summaries in this section:

A. *Users' needs:*

1. Zero-barrier access: any user who carries out non-for-profit research should be able to get an account with a 'start-up' resource allocation on EGI to access scalable application services together with the underlying cloud/HTC/storage resources.
2. 100% coverage: anyone with Internet access can become a user, no need to ask for personal travel to obtain special credentials (e.g. X.509 certificate).
3. Extendible and open: The service must be extendible with additional application services to support specialized scientific disciplines and users. The extension should be possible via open interfaces and protocols.
4. User-centric: Support for users should be available in as many EGI member countries as possible.

B. *Service providers' constaints:*

1. Realistic: Define an architecture that's implementable under the available effort levels. Reuse existing EGI technological building blocks as much as possible.
2. Secure: Provide as high quality user identity vetting, and tracking of user activities as possible (close to the existing solutions that are based on personal X.509 certificates).
3. Scalable: Be able to scale to 100s of compute/data/application providers. (The number and size of providers limits the users too).
4. Recognized: Have sufficient policies and tools that enforce the users to acknowledge the use of the service in scientific publications resulted from this use.

III. Implementation

The architecture of the Applications on Demand service is presented in Figure 1. In the heart of the service there is a 'User Registration Portal' (URP). This is where new users enter to the service. The list of applications and underpinning services can be publicly browsed on the URP. Usage requires login. Login is possible with Google, Facebook and EGI Single-Sign-On accounts – All un-vetted accounts that are available for anyone with Internet access. Within the URP the user can setup a personal profile and can submit a resource access requests. The request includes an estimate of the compute and storage capacity that the user would like to use through the application(s) he/she would like to use in the service. The request can use the default capacity allocation (at the time of writing this is 1000 CPUh and 10 TB storage) or can be a customized request.

The access request is forwarded to the distributed service support team. A team member – from the applicant's country or from the EGI Foundation responds to the request and vets the user identity, checks the capacity request. The vetting is done by assessing the validity of the information provided by the user in the profile, including links to departmental websites and scientific publications and projects relating to the requested e-infrastructure use. If needed, the support teams contacts the applicant in email or by phone. Routing the access requests to national EGI members not only allows us to conduct such conversation in local languages, but also to connect long tail users to national EGI support teams, reaching recognition and trust in national e-infrastructure initiatives. If the capacity

[1] EGI Beta service: Service being developed while available for testing publicly.

request exceeds the default allocation then the estimate is double-checked and if needed negotiated resource providers of the service.

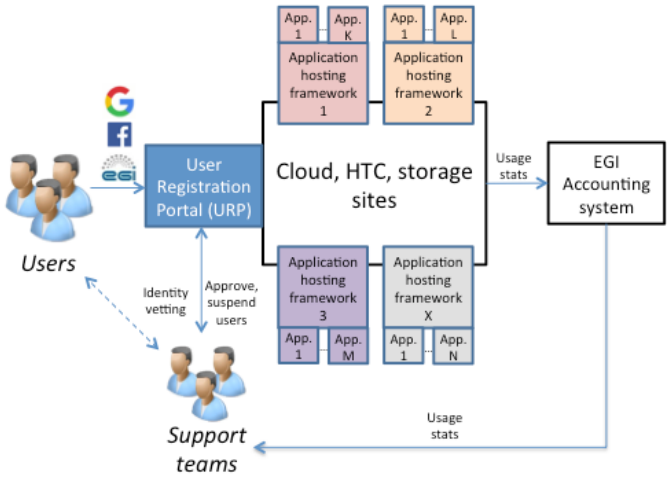

Figure 1. Infrastructure architecture

After the applicant's resource access request is approved, the user profile is set to 'active' in the URP. This information is propagated to the connected application hosting environments (science gateways, Virtual Research Environments or similar). These environments operate as applications hosting frameworks, being equipped with a set of applications and user interfaces, manuals/guides best suited for those applications. For example, some gateways can be suited for workflow type applications, others for parameter study jobs, yet others for applications of a certain scientific discipline. The service is open for any application hosting framework (See Section VI for more details). Science gateway and Virtual Research Environments supported by the European Commission FP7 and H2020 work programmes are for example ideal candidates for integration.

Approved users can login to any of the connected application hosting frameworks and use the embedded applications. These applications are already configured to scale to the distributed compute and storage systems that are contributed to the service by EGI members. The resources are made accessible for the applications in the form of Infrastructure as a Service clouds joint into the 'EGI Federated Cloud' [22], or cluster resources federated with grid middleware into the EGI 'HTC Service' [23].

The application use generates load on these resources and this is reported by the resources into the EGI accounting system. From this accounting system the user support teams can obtain statistics about individual users, as well as about application hosting frameworks, cloud/HTC/HPC sites. When a user exceeds the amount of compute/storage/network capacity that was allocated for him/her, the account can be suspended in the URP, blocking the user from further consumption through this service. The continued use requires a new justification from the user and a corresponding evaluation by the support team.

IV. Enabling Technologies

The service was created by customizing various existing EGI components, and by developing a few new ones in order to glue together the required elements into a single service. The following customizations and developments were made:

- The User Registration Portal (URP) was developed by CYFRONET as an extension of the e-GRANT EGI resource allocation tool [7]. The URP includes forms that guide users through the profile setup, resource request and application selection workflow. The tool was extended with web forms and an email notification subsystem to help the distributed user support team during the user identity vetting and user approval process. The URP provides an identity federation for the whole service, i.e. a user can authenticate to this portal and after his/her account is validated he/she will be able to login to any of the connected application hosting frameworks and respective hosted applications.
- A resource pool was formed from EGI cloud and HTC sites. The participating sites feel institutionally or nationally responsible for supporting the long tail of science, and therefore contribute with capacity into this pool. The sites are joined together into an EGI Virtual Organization called 'vo.access.egi.eu' [25]. The sites in this pool accept special X.509 proxy certificates that are unique for the Applications on demand service and which identify user workloads from the application hosting frameworks (See details under next point). At the time of writing the resource pool includes cloud resources from Italy (INFN-Catania and INFN-Bari) and Spain (BIFI, CESGA) and HTC clusters from Belgium (VUB), Italy (INFN-Catania and INFN-Bari), Poland (CYFRONET) and Spain (CESGA).
- Access to EGI resources requires short-living X.509 proxy certificates from the client side. The traditional method is to generate such proxies either from a long-term personal certificate, or from a robot certificate [8] that is used by an application hosting framework. Unfortunately neither of these approaches were sufficient for the Applications on Demand service, because (1) personal certificates are found difficult to handle or impossible to obtain by certain long tail users, and (2) robot-proxies do not include any information about the individual end-users, hiding all users workload under a single identity, making impossible for the identification of excessive or harmful use by certain individuals. To overcome these limitations a new, so-called 'Per-User Sub-Proxy' (PUSP) mechanism was developed. PSUPs are short-term proxies that are generated from robot certificates in a special way: The 'distinguished name' (DN) field of the proxy includes a unique string that is specific to the requester user. The DN is the same for a specific user for multiple user sessions, even if those sessions are initiated by different application hosting frameworks. The compute and storage resources report the proxy DN into the EGI accounting system, and based on the user-specific DNs

we are able to trace back every e-infrastructure operation to the initiating user. The user-specific strings are generated during the identity vetting and approval process, and they are propagated to every participating application hosting framework. This tracing process is graphically presented in Figure 2.

- The connected application hosting frameworks must generate PSUPs from robot certificates. This is possible in two ways: (1) from a robot certificate that is deployed either locally on a USB smartcard on the server that hosts the framework. (2) Because robot certificates are not available from EGI-recognised CAs in every country, we setup an 'eToken server' at INFN-Catania that can serve any participating application hosting frameworks with PSUPs. The application hosting frameworks can send a PSUP generation request to the server via its network API, including the userID as a parameter. The eToken server responds with the short-term proxy certificate that can be used by the framework to interact with the VO resources (See step 5 and 6 in Figure 2.).

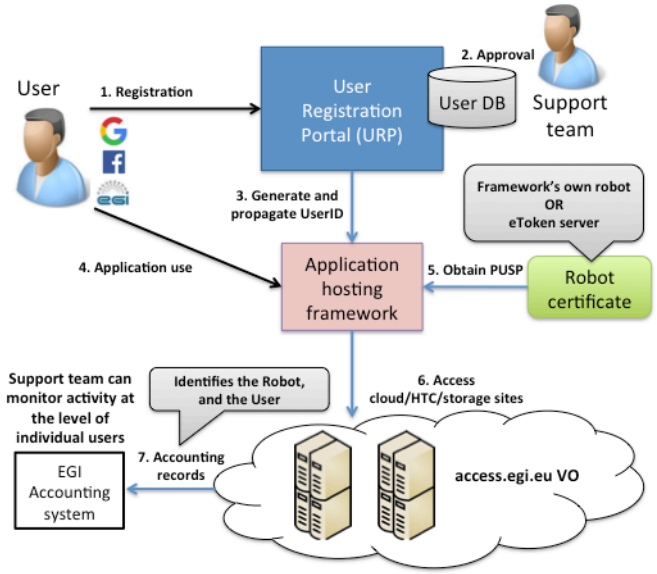

Figure 2. Tracing user activity on VO resources

- Application hosting frameworks that can provide user-friendly interfaces to conduct scientific applications on VO platform resources in cloud or HTC/HPC clusters. The gateways use the identity federation of the URP to allow access to approved users, and user the PUSP mechanism to interact with cloud and HTC resources. The service currently includes three applications hosting frameworks: the WS-PGRADE [9], the EC3/IM [10] and the CSG [11]. The service is open for any additional framework that wish to make applications and application development/hosting services available at the European/worldwide scale. Technical instructions to integrate a new applications hosting framework to the service are provided in Section VI.

- 17 applications from different scientific areas and tools have been already integrated into the existing applications hosting frameworks and are offered 'as services' to users:
  - Molecular Docking, Workflow and parameter study tool (WS-PGRADE portal).
  - Galaxy, Docker, Octave, Apache Tomcat, GnuPlot, NAMD, Hadoop, Marathon, Chronos, Jupyter Notebook, Cloud orchestrator (in the EC3/IM portal).
  - Chipster, ClustalW2, Semantic Search, the Statistical R for Computing (in the Catania Science Gateway).

  Additional applications will be integrated into the frameworks following cost-value assessment (more impactful, more broadly relevant applications have priority).
- Two policies were developed: A security policy for resource centers offering cloud/HTC/HPC/storage and an Acceptable Use Policy (AUP) for users.
  - The first policy is compulsory to accept and implement by participating Resource Centers. The policy defines that offering resources in this service shall not negatively affect the security or change the security risk of any other VO. In particular, security incidents originating in the Applications on Demand service should not impact the IT Infrastructure in ways that are incompatible with the operational model of other VOs. This document also provides guidelines on the implementation of security procedures and controls to facilitate offering of the Service by Resource Centers and Science Gateways. The Guidelines contain normative information on how to implement the Policy.
  - The user AUP defines the conditions of use, and responsibilities of the users – such as using the services only for activities that relate to the work that was described in the access request form; or what text to use in scientific publications to acknowledge the use of the infrastructure.

V. Elements of the Service

The following Table 1 provides a summary of the components that are currently operated within the EGI Applications on Demand service.

Table 1. Already connected services and resources.

Type of resource/service		Providers
IaaS clouds (compute and storage)		BIFI (Spain)
		CESGA (Spain)
		INFN-Bari (Italy)
		INFN-Catania (Italy)
HTC clusters (compute and storage)		CESGA (Spain)
		CYFRONET (Poland)
		INFN-Bari (Italy)
		INFN-Catania (Italy)
		ULB-VUB (Belgium)
Applications	Molecular Docking	WS-PGRADE (SZTAKI, Hungary)
	Workflow and parameter study tool	
	Galaxy	EC3/IM (UPV, Spain)
	Docker	
	Octave	
	Apache Tomcat	
	GnuPlot	
	NAMD	
	Hadoop	
	Marathon	
	Chronos	
	Jupyter Notebook	
	Cloud orchestrator tool	
	Chipster	Catania Science Gateway (INFN-Catania, Italy)
	ClustalW2	
	Semantic Search	
	Statistical R	

VI. How to Join as a Gateway/VRE Provider

This section provides a short overview of the integration steps that gateway/VRE providers must complete to contribute to the service. Interested providers should consult with the online manual for details [12]. Compute and storage providers (cloud, HTC) can join the service by federating into the 'vo.access.egi.eu' Virtual Organisation [25], following the regular EGI guides for resource providers.

There are two fundamental prerequisites of integrating an application hosting framework: (1) the framework must be a mature technology[2] with demonstrated use within publicly funded science and (2) the framework must be already able to use cloud, HTC or data services from EGI, or be ready to do so (to enable the scale-out of the hosted applications).

As reported in the previous Sections, the URP provides the identity federation that enables users to authenticate in any of the connected application hosting frameworks with either social credentials or EGI SSO accounts. In the current implementation, this identity federation is implemented with Unity [13], an authentication & authorization management solution that uses OpenID Connect as standard interface. The first requirement to integrate a new applications hosting framework into the service is to register the framework in Unity to get a *clientID* and *secretKey*. These credentials will be used by Unity to identify the new provider and implement a secure connection with it.

For applications hosting frameworks based on Liferay technology INFN-Catania has developed an OpenID Connect module [14] that enables Liferay-based gateways to authenticate with OpenID Connect providers. This module is adopted by the WS-PGRADE and CSG frameworks within this service.

A second step of integration is using the userID provided by Unity to generate Per-User Sub-Proxies (PUSP) to secure user interaction with the EGI resources. For this, the provider can rely on the eToken server that was already presented in the previous Section.

Lastly, the technical integration is complete when the service support team together with the framework provider 1.) registers the framework in the EGI service registry (GOCDB) to activate the availability and reliability monitoring for the framework with the EGI ARGO service [15], 2.) setup a framework-specific support unit in EGI Helpdesk [16] (this is for example tickets will be opened when the framework is noticed inaccessible by the monitor system and 3.) sign an Operation Level Agreement (OLA) defining for example availability and reliability targets, helpdesk ticket response times.

VII. Future Work

The 'EGI Applications on Demand service' was opened for public use in April 2017. In the next few months we are working on promoting the system to potential users, mainly through the user support teams of EGI member states. Promotion will be focused on countries where national support is lacking or insufficient for the long tail of science.

In parallel with broadening the uptake of the new service we are also planning to improve/expand the technical setup. This work will cover the following areas:

1. Replacing PUSPs with short-term proxies generated by the RC Auth service [17]. RC Auth was designed by the AARC H2020 project [18] to have an open, European proxy factory service that can be used by any e-infrastructure and Research Infrastructure that needs X.509 proxies for service access. The use of RC Auth in the EGI Applications on Demand service would improve the compatibility of our architecture with other European initiatives, simplifying the integration of additional applications and hosting environments. Besides, the change would improve the sustainability of our setup, by eliminating components that EGI currently has to sustain alone.
2. In the current architecture every application framework and application should implement its own tools to manage the users' scientific data (for example importing

[2] For example at 'Technology Readiness Level 8 or 9 [24]

data into the infrastructure from external storage systems, exporting results into external repositories). We are planning to integrate a data management service (or layer) into the EGI Applications on Demand service that could be used by all applications hosting frameworks for data management. If properly designed, the service could also help users curate and archive application outputs, tackling the problem of 'dark data in the long tail' [19]. The details of this data management layer are yet to be defined.

3. The current EGI flagship project, EGI-Engage [20], is close to establishing an 'EGI Marketplace'. This marketplace would offer a one-stop-shop for individual researchers, research communities and industry to browse and request services from EGI. The EGI Applications on Demand service and/or its individual services will have to be connected/integrated into this marketplace to make them visible and accessible for potential users and customers. This will require some sort of merge of the URP into the EGI Marketplace.

VIII. Conclusions

In this paper we presented a new EGI Applications on Demand service. The service is specifically designed to cater for the needs of individual researchers, small research collaborations and early-phase Research Infrastructures. The service provides easy to use environments for these user communities to request and access scalable scientific applications and application porting environments. The service includes also distributed computing and storage resources, and eliminates the need for users to form community-specific agreements with EGI providers, and to sustain skilled IT teams to operate VO services.

The new service does not replace the 'traditional', community/project specific and national VOs of EGI. Structured scientific collaborations and mature Research Infrastructures continue to require dedicated VOs in EGI because only those can host fully customized, community-specific services and only they can offer dedicated compute and storage capacity at extreme scales. National 'long tail VOs' are also here to remain in those NGIs that have effort to support local users with national services. The EGI Foundation continues to negotiate and secure services for community-specific VO through Service Level and Operational Level Agreements (SLAs, OLAs).

Acknowledgment

The work presented in this paper has been supported by the EGI-Engage H2020 project (Grant number 654142). The authors would like to thank the contribution to the implementation of the described service to the following institutes: BIFI, CESGA, CNRS, CYFRONET, INFN-Catania, INFN-Bari, NIKHEF, SZTAKI, ULB-VUB, UPV.

References

[1] Unwinding the 'Long Tail' of Science, https://www.ci.uchicago.edu/blog/unwinding-long-tail-science, Accessed: 8/Apr/2017.

[2] LHC experiments: https://home.cern/about/experiments, Accessed: 8/Apr/2017.

[3] VIRGO experiment: http://www.virgo-gw.eu, Accessed: 8/Apr/2017.

[4] Cerenkov Telescope Array Observatory: https://www.cta-observatory.org/, Accessed 8/Apr/2017.

[5] Biomed Virtual Organisation: http://lsgc.org/biomed.html, Accessed 8/Apr/2017.

[6] European Strategy Forum on Research Infrastructures (ESFRI). Infrastructure roadmap 2016: http://www.esfri.eu/roadmap-2016

[7] EGI e-GRANT resource allocation tool: https://e-grant.egi.eu, Accessed 8/Apr/2017.

[8] IGTF Robot certificates: https://www.eugridpma.org/guidelines/robot/, Accessed 8/Apr/2017.

[9] P. Kacsuk, Farkas, Z., Kozlovszky, M., Hermann, G., Balasko, Á., Karóczkai, K., and Márton, I., "WS-PGRADE/gUSE Generic DCI Gateway Framework for a Large Variety of User Communities", Journal of Grid Computing, vol. 10, no. 4, pp. 601 - 630, 2012.

[10] M Caballer, D Segrelles, G Moltó, I Blanquer, "A platform to deploy customized scientific virtual infrastructures on the cloud", Concurrency and Computation: Practice and Experience 27 (16), 4318-4329, 2015.

[11] V. Ardizzone, R. Barbera, A. Calanducci, M. Fargetta, E. Ingrà, I. Porro, G. La Rocca, S. Monforte, R. Ricceri, R. Rotondo, D. Scardaci, A. Schenone, "The DECIDE Science Gateway", Journal of Grid Computing vol. 10, issue 4, pp. 689-707, 2002.

[12] Manual on how to connect an application hosting framework to the infrastructure: https://wiki.egi.eu/wiki/Long-tail_of_science_-_information_for_providers, Accessed 8/Apr/2017.

[13] https://www,unity-idm.eu

[14] OpenId Connect for Liferay: https://github.com/csgf/OpenIdConnectLiferay

[15] http://argo.egi.eu/

[16] https://wiki.egi.eu/wiki/FAQ_GGUS-New-Support-Unit-or-VO

[17] RCauth: The white-label Research and Collaboration Authentication CA Service for Europe. https://www.rcauth.eu/, Accessed: 8/Apr/2017.

[18] AARC H2020 project: https://aarc-project.eu/, Accessed 8/Apr/2017.

[19] Heidorn PB, "Shedding light on the dark data in the long tail of science", Library Trends vol. 57, issue 2, pp.280–289., 2008.

[20] Engaging the EGI Community towards an Open Science Commons (EGI-Engage) H2020 project: https://wiki.egi.eu/wiki/EGI-Engage, Accessed 8/Apr/2017.

[21] EGI Accounting data about 'catch-all' Virtual Organisations: https://accounting.egi.eu/discipline/Miscellaneous/normcpu/VO/DATE/2014/5/2017/5, Accessed 9/May/2017.

[22] EGI Cloud Compute service: https://www.egi.eu/services/cloud-compute, Accessed 9/May/2017.

[23] EGI HTC service: https://www.egi.eu/services/high-throughput-compute, Accessed 9/May/2017.

[24] Technology Readiness Levels by the European Commission: https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf, Accessed 9/May/2017.

[25] vo.access.egi.eu Virtual Organisarion in the EGI Operations Portal: https://operations-portal.egi.eu/vo/view/voname/vo.access.egi.eu, Accessed 9/May/2017.