Inspired

ISSUE 23
APRIL 2016

news from the EGI community

TOP STORIES

MORE

Engage - Grow - Innovate

www.egi.eu

This issue

In this edition of *Inspired*:

- We announce the Digital Infrastructures for Research 2016 event in Kraków
- Ian Bird tells us how the LHC computing is preparing for Run 2
- Christos Kanellopoulos gives the lates updates on the AAI roadmap of EGI-Engage
- We write about a EGI FedCloud use case published in Nature
- Pasquale Pagano writes about the uses of hybrid data e-infrastructures as VREs-as-a-Service
- Sara Pittonet Gaiarin introduces the INDIGO-DataCloud champions

and

- Magdalena Brus presents a virtual platform for the ENVRI+ community

Your feedback and suggestions are always welcome!

Send an email to Sara & Iulia at: press@egi.eu

EGI's next event will be the Digital Infrastructures for Research 2016 in Kraków Poland. The event is co-organised by EUDAT, GÉANT, OpenAIRE and RDA Europe

Computing centres: CESGA

Rubén Díez Lázaro shows us the CESGA, one of the building blocks of EGI

The Supercomputing Centre of Galicia, CESGA, is the centre for high-performance computing, communications and advanced services used by the scientific community of Galicia, the University System of Galicia and the Spanish National Research Council (CSIC).

The cluster in the photo is made by 40 working nodes (HP ProLiant SL2x170z G6).

Some time ago, the cloud management software (OpenNebula) and the storage host ran in some of the HP ProLiant DL180 present in the rack. Currently, the NAS storage (Network-attached storage) is provided by a specifically designed machine (EMC VNX5700) and OpenNebula also runs in a virtual machine. Recently, new and more powerful working nodes were added to the infrastructure.

Rubén Díez Lázaro, technical officer, CESGA, Spain

About CESGA

CESGA website: http://www.cesga.es/

Digital Infrastructures for Research 2016: 28-30 Sept

The next EGI event will be in Kraków and is co-organised with EUDAT, GÉANT, OpenAIRE and RDA Europe

Digital Infrastructures for Research 2016.
Serving the User Base
www.digitalinfrastructures.eu
28-30 September 2016, Krakow, Poland

In this first jointly organised event, EGI and Europe's other e-infrastructures invite all researchers, developers and service providers for three days of brainstorming and discussions, hosted by ACC Cyfronet AGH - Kraków's academic computing centre.

Conference tracks

The Call for Participation to the DI4R2016 event is live and online submission of abstracts will open in early May. The Programme Committee, chaired by David Foster (CERN and EGI Council), welcomes submissions for full sessions, single presentations, training sessions and demonstrations to:

> Challenges facing users and service providers: emerging needs of research collaborations, the requirements of added value thematic services and the computing needs of data-driven science. For example:

- Working with the research community and industry,
- Community engagement, computing platforms (cloud, HTC, HPC),
- Thematic platforms (science gateways, virtual research environments)

> Services enabling research: services and frameworks needed to enable researchers to securely collaborate and share resources in a federated environment combining geographically distributed services from multiple providers and further the opportunities of Open Science. Submissions for this track should highlight benefits and challenges as seen by researchers when using existing frameworks or present ideas to address the future challenges.

> A changing environment, changing research: The environment in which research is conducted, and digital infrastructures operate, is changing rapidly. Access and provisioning of services require clear governance, engagement rules, policies and funding models. Submissions should focus on the barriers, opportunities and changes in this environment in order to address the non-technical pressures, for example social, financial, legal and policy that influence the present and future opportunities.

> Working with data: requirements of data-driven science and the solutions for finding, accessing, integrating and reusing research data. Papers that highlight requirements and opportunities for a seamless usage of digital infrastructures for data management, storage and curation as well as for linking and publishing all forms of research objects like data, software, tools, pipelines and publications would be particularly welcome.

Dates and deadlines

> **1 May 2016**: Submission system opens

> **3 June 2016**: Deadline for
* Full sessions
* Single presentations
* Training sessions & demos

> **20 July 2016**: Deadline for
* Lightning talks
* Posters

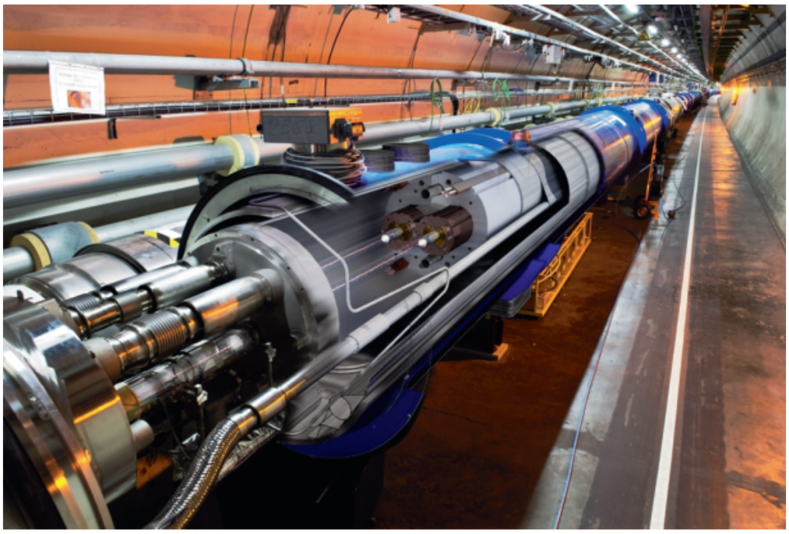
More information

Digital Infrastructures for Research 2016
www.digitalinfrastructures.eu/

Full text of the Call for Participation
http://go.egi.eu/DI4R-CfP

LHC computing ramps up for Run 2

Ian Bird writes about the challenges ahead for WLCG



(c) CERN, Maximilien Brice

The Large Hadron Collider at CERN has recently been brought back online after a two-year programme of upgrades to enable precision measurements and hopefully lead to further discoveries.

At higher energy and intensity, collision events are more complex to reconstruct and to analyse and computing requirements will increase accordingly. Starting in April 2016 there will be a three-year run anticipated to produce twice the data produced in the first run (about 50Pb/year).

Shut down time

The shutdown gave us time to install enhancements to the core software of the experiments, enabling WLCG to manage this increased data rate and complexity by 'only' a doubling of the CPU and storage capacity. Today WLCG has access to some 600,000 cores and 500 PB of storage, provided by the 170 collaborating sites in 42 countries.

During the shutdown, we also updated the computing models to take advantage of the fantastic performance of the global networking infrastructure, which continues to provide increasing bandwidth and reliability. These changes include a move away from the original hierarchical model of data flow from Tier 0, to Tiers 1 and 2, to a peer site model. The new model gives us the ability to use the full capabilities of a computing site rather than its strict hierarchical role. This is also enhanced by the introduction of 'data-federations' allowing access to data across the network, and reducing the need for pre-placement or additional duplicates of datasets.

Future developments

As we look further into the future, there are two phases of upgrades planned for the LHC. The first phase (2019-2020) will see major upgrades of ALICE and LHCb, as well as increased luminosity of the LHC. The second phase – the ESFRI High Luminosity LHC project (HL-LHC) in 2024-2025 - will upgrade the LHC to a much higher luminosity and the ATLAS and CMS detectors to provide higher precision.

The requirements for data and computing will grow dramatically over this time, with rates of 500 PB/year expected for HL-LHC. The needs for processing are expected to increase more than ten times over and above what technology evolution will provide.

This is why we are embarking on a programme of R&D to investigate how the computing models could evolve to address these needs. We will focus on three points:

> apply more intelligence into filtering and selecting data as early as possible;

> investigate the distributed infrastructure itself (the grid) and how we can best make use of available technologies and opportunistic resources (grid, cloud, HPC, volunteer etc.);

> improve software performance to optimize the overall system.

In addition there is a lot of interest in investigating new ways of data analysis: global queries, machine learning, and many more.

These are all significant and exciting challenges, but it is clear that our 'grid' will continue to evolve and grow and that in 10 years it will look very different from what it is today, while retaining the features that enable global collaboration.

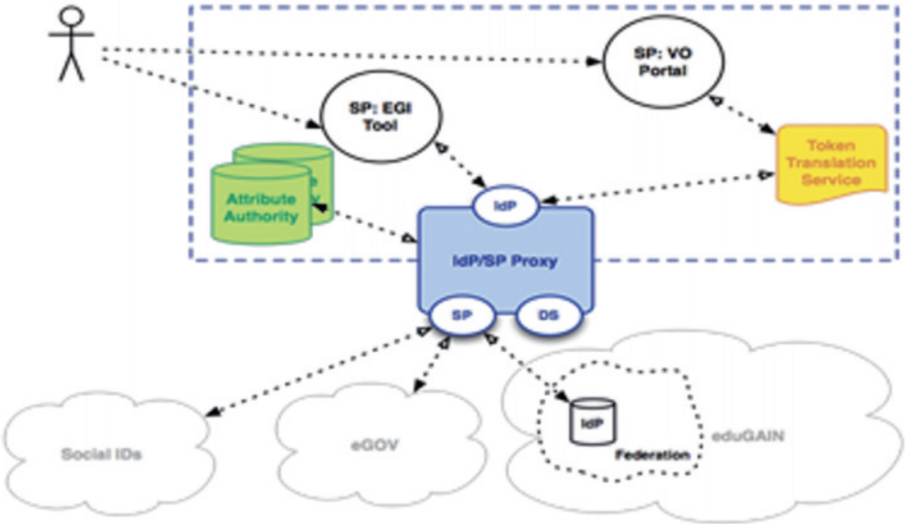
More information

Ian Bird is a senior scientist at CERN, and Project Leader of the WLCG (Worldwide LHC Computing Grid)

http://cern.ch/wlcg-public

Latest updates on the AAI roadmap in EGI-Engage

Christos Kanellopoulos presents a short guide to scientific requirements and next steps



The EGI AAI architecture

The Authorisation and Authentication Infrastructure (AAI) activity in EGI-Engage started in May 2015. During the first months we worked together with the AARC project to identify the requirements of the scientific communities. This work resulted in a ***set of principles to guide the AAI roadmap and architecture***:

- Users should be able to access the EGI services using the credentials they have from their home organisations through eduGAIN when possible, but alternate methods should be available
- The EGI platform expects to receive at least an identifier that uniquely identifies the user within her organisation.
- Within the EGI environment, a user should have one persistent, non-reassignable and non-targeted unique identifier.
- EGI should define a set of minimum mandatory attributes for all users within the EGI environment
- EGI should attempt to retrieve these attributes from the user's home organisation. If this is not possible, then an alternate process should exist in order to acquire and verify the missing user attributes.
- There should be a distinction between self-asserted attributes and the attributes provided by the home organisation/VO (virtual organisation).
- Access to the various services should be granted based on the VO/EGI roles the user has.
- EGI services should not have to deal with the complexity of multiple IdPs/federations / attribute authorities / technologies. This complexity should be handled centrally.

AAI: future plans

By the end of the first year of EGI-Engage, the EGI AAI will be fully functional in terms of core features and we can start on-boarding scientific communities. The recent introduction of the pilot CILogon service enables all users to access even the non-web EGI services through the EGI AAI.

During April 2016 it is expected that the EGI AAI will join eduGAIN as service provider in the Research & Scholarship entity category. Through eduGAIN, the EGI services will automatically become available to more than 2000 universities and institutes that are connected to the 38 eduGAIN Federations. Complementary to this, users without an account on a federated identity provider will be able to use their Google, Facebook, LinkedIn and ORCID accounts to access EGI services that do not require a substantial level of assurance.

In the second quarter of 2016, we will be working on the first phase of the pilot with the EGI Competence Centres to connect them to the EGI AAI. This is going to be an interactive process, which will allow us to shape the EGI AAI to the needs of our customer base.

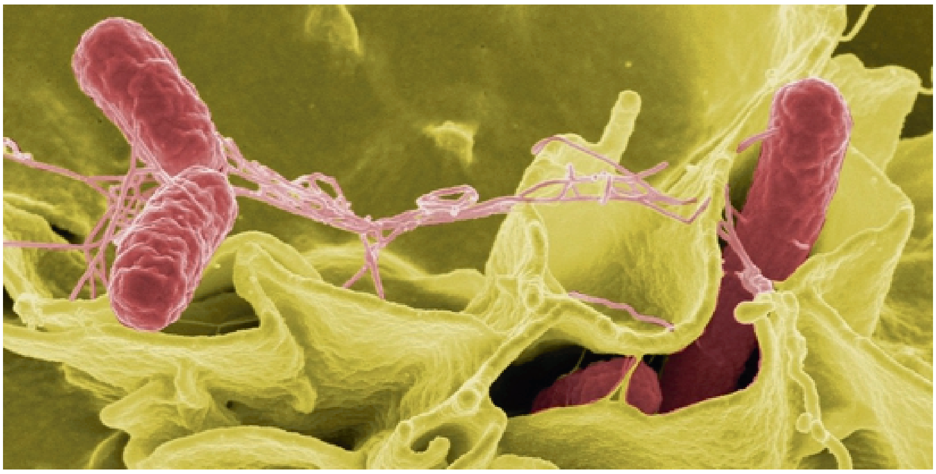
In the third quarter of this year, we will continue with the second phase of the pilot and we expect to have all the EGI scientific communities on board at the end of Summer. We will also introduce the new OpenID Connect interface, which will enable us to introduce new services to the EGI platform in a faster and friendlier way.

More information

Christos Kanellopoulos leads the AAI task of the EGI-Engage project. Christos is based at GRNET.

EGI FedCloud in Nature: the genetics of infections

How EGI Federated Cloud resources helped scientists to understand what happens during when a human cell meets Salmonella.



Colour-enhanced scanning electron micrograph showing *Salmonella* (in red) invading human cells.
Credit: Rocky Mountain Laboratories, NIAID, NIH (Wikimedia Commons)

Salmonella infections end up with many unpleasant symptoms and a likely trip to the hospital. What happens inside the cells is an invisible genetic battle between the bacteria Salmonella cells and the unfortunate host.

Every living cell has a generic blueprint with instructions for all possible scenarios –the DNA. When the opportunity comes, the bacteria activate the part of the DNA they need to start the infection. This DNA is translated into mRNA and the mRNA is used to make proteins - the ammunition of the attack. On the other side of the barricade, the host cells activate the mRNA they need to make proteins for the defence.

Small RNAs

Konrad Förstner, from the University of Würzburg, and his colleagues took a new approach to look at this dynamics: they analysed the RNAs produced by the *Salmonella* and the host at the same time, in the same experiments.

First, the team infected human HeLa cells with *Salmonella*. Then they analysed the combined RNA from the both organisms by using so called high-throughput sequencing.

Analysing dual RNA-Seq data is complex and time consuming. To streamline all the work, Förstner ran the READemption tool on the EGI Federated Cloud to process most of the computational tasks.

Using the FedCloud

Förstner says that FedCloud resources "accelerated our analysis dramatically." He also added that "the FedCloud helped us to focus on scientific solutions and results instead of resource management and system administration."

Thanks to this novel approach, the research team found that a piece of *Salmonella* RNA called PinT is heavily involved in what happens right after the infection. In Förstner's words, PinT "fine-tunes the transition from the infection stage, which requires a large set of genes, to the survival stage, that needs a different set of genes."

When PinT activity was shut down in a follow up experiment, the team saw the changes in the bacteria and in the host cell as well. This shows that PinT also influences what happens in the host and is "of high relevance for understanding the infection process as well as the immune response," Förstner concludes.

The FedCloud resources used in this research were provided by the following sites:

> ***GWDG,*** the university computing centre for the Georg-August-Universität Göttingen and

> ***IFCA,*** the Institute of Physics of Cantabria

More information

READemption is a pipeline for the computational evaluation of RNA-Seq data. It has a consistent interface to perform different tasks, takes care of parallel processing, statistics and visualisation of results. *http://pythonhosted.org/READemption/*

Reference

AJ Westermann et al. 2016. Nature. doi:10.1038/nature16547 *http://go.egi.eu/salm*

Hybrid data e-infrastructures: VREs as-a-Service

Pasquale Pagano writes about how VREs are helping 900 users in the fisheries community

Credit: Nick Hobgood (Wikimedia Commons)

Science is increasingly global, multipolar, and networked. Data continue to grow in volume and variety and research now crosses the boundaries of single institutions, disciplines, and countries. We need innovative working environments to give researchers the facilities they need (data, services, computing resources) while allowing them to save time and money without compromising research quality. This is the concept behind Virtual Research Environments as-a-Service (VRE aaS).

The iMarine project (2011-2014) started experimenting VREs aaS in 2011. Thanks to D4Science (the underlying hybrid data infrastructure), the iMarine project was able to serve Data Analysts, Managers & Providers, Scientists, Researchers and SMEs working in the field of fisheries and marine living resources with a set of tailor made VREs. After the end of the project, iMarine became a self-sustained initiative jointly supported by the Italian National Research Council (CNR) and the Food and Agriculture Organisation (FAO) together with a series of donors.

Today iMarine has 16 VREs serving around 900 users in the fisheries community.

Building on new requirements from the Blue Growth sector, we started the BlueBRIDGE project in September 2015, and as of April 2016 we have 12 new VREs serving already more than 300 users.

BlueBRIDGE users span from fisheries to the aquaculture communities (with a particular focus on SMEs), but also supports educators in setting up scientific training environments that require execution of data-intensive processes.

The new VREs implemented in BlueBRIDGE will enrich the iMarine portfolio and will be built on the D4Science infrastructure expanding its user base (D4Science connects over 2400 scientists in 44 countries belonging to more than 10 international initiatives).

Joining forces with the EGI FedCloud

The flexibility and capability to serve thousands of users will increase in May 2016 when we extend the array of exploitable computing resources with the EGI Federated Cloud (FedCloud). The EGI resources will help BlueBRIDGE to create a VRE capable of transforming a set of Copernicus in situ marine observations of an acidification parameter into one uniform global distribution map.

VREs under development

BlueBRIDGE is working in close contact with aquaculture and fisheries practitioners to develop a set of innovative VREs to serve their specific needs. Among the others, by the end of this year, three new VREs will be released:

> Stock Assessment VRE: an environment to support stock assessment teams in uploading, harmonizing, analysing and reporting data

> Aquafarming performance modelling VRE: an environment to estimate and assess farm performance indicators

> Spatial planning VRE: an environment to generate holistic maps to support spatial planning

More information

Pasquale Pagano is the technical director of BlueBRIDGE

BlueBRIDGE project
www.bluebridge-vres.eu

BlueBRIDGE VREs
https://i-marine.d4science.org

iMarine VRE Portfolio
http://go.egi.eu/iMarine

Better software for better science

Sara Pittonet Gaiarin presents the INDIGO-DataCloud Champions

In cloud computing, both the public and private sectors are offering cloud resources as IaaS (Infrastructure as a Service). However, there are numerous areas of interest to scientific communities lacking cloud computing uptake, especially at the PaaS (Platform as a Service) and SaaS (Software as a Service) levels.

The **INDIGO-DataCloud** project (Integrating Distributed Data Infrastructures for Global Exploitation) aims at developing a data and computing platform targetting scientific communities, and allows execution of applications on cloud and grid based infrastructures, as well as on HPC clusters.

Different components for communities

INDIGO-DataCloud delivers open source software components tailored to scientific communities and to e-infrastructures, aimed at increasing ease of practice and effectiveness in their use of cloud resources. INDIGO ready-to-use components can be grouped as:

> User-oriented access services (user interfaces, mobile applications, scientific portals)

> Optimised exploitation of resources across multiple cloud infrastructures

> Seamless and integrated access to geographically distributed data

> Improved functionalities in the popular cloud frameworks OpenNebula and OpenStack.

In addition, the INDIGO Future Gateway (FG) offers a set of easy-to-use RESTful APIs to allow portals, mobile appliances, and desktop applications to exploit different kinds of e-infrastructures (grids, clouds and HPC clusters).

INDIGO Champions @ work

The project identified several 'Champions' from each research community to lead different use cases and test INDIGO technologies. INDIGO Champions have been at work since January 2016 to provide requirements and have contributed to the release of the INDIGO-DataCloud platform.

Some examples?

The 3D structure of molecule fluctuates over time due to the kinetic energy available at room temperature. Computer simulation is the only technique to provide a full view. This use case exploits the INDIGO FG to perform simulations in virtual machines, using web interfaces for set up and analysis.

The INDIGO climate use case demonstrates the capabilities deployed on heterogeneous infrastructures (e.g., HPC clusters and cloud environments), as well as workflow support to run distributed, parallel data analyses.

More information

Sara Pittonet Gaiarin is Digital Communications Strategist & Project Manager at Trust-IT

INDIGO-DataCloud
www.indigo-datacloud.eu/

The virtual platform for ENVRI community

Magdalena Brus on a platform to bring together the environmental science ESFRIs

Every research infrastructure (RI) is special and has its own focus area, but they are all facing common challenges in their construction and are contributing to the wider, trans- and interdisciplinary science questions.

It is therefore important to bring the RIs from the same scientific field together to work towards the interoperability and harmonization of their operations and products.

This cooperation allows utilising the sparse resources and making the environmental RI (ENVRI) landscape as efficient as possible. The Environmental Strategy Forum for Research Infrastructures (ESFRI) has realised the necessity of the RI clusters after the first update of its roadmap in 2008.

The cooperation has since then been boosted by several projects funded by the European Commission. ENVRI (2011-2014) was the first cluster project for the European environmental RIs. Although the main focus of ENVRI was on data and software solutions, it has also built a strong foundation for the ENVRI community mind-set.

It was within this project that the RI communities from different fields of environmental and e-science started to talk together and soon realised that they have many things in common - not just challenges, but also visions towards the future.

The cluster now continues its work in ENVRIplus (2015-2019), a four-year project bringing together 20 environmental RIs in Europe. The effort is organised in six different “Themes” ranging from the development of common technical and data solutions through work on joint policies and guidelines, to transfer of knowledge.

Not all integration is European: COOPEUS project (2012-2015) enabled better data interoperability with RIs’ counterparts in US, and recently funded COOP+ (2016-2018) is aiming at bringing cooperation between environmental RIs on a global level.

ENVRI community platform

But where does all of this come together? How do we use and incorporate the results of all these projects supporting the integrated collaboration between RIs? Where can the community access developed products and solutions and engage in dialogues about future needs? How do we bring on-board new emerging RI networks, e-infrastructures, and other relevant stakeholders?

All of this will be possible within the ENVRI Virtual Community Platform, which is currently being built. The platform will be launched online in May 2016.

It will serve as a meeting point, bringing all the ENVRI community players together to share information, guidelines, products and services developed by current and future projects. One of the most important attributes of the platform is its sustainability for a longer period than the typical lifetime of a project.

The ENVRI platform will be hosted on EGI’s servers. EGI is a trusted partner of the ENVRI community infrastructure, being already involved in ENVRI and ENVRIplus projects and contributing to data solutions, training activities, dissemination and community building.

More information

Magdalena Brus is the project manager of ENVRIplus and leads the development of the virtual ENVRI community platform.

http://www.envriplus.eu/