Inspired

ISSUE 22
JANUARY 2016

news from the EGI community

TOP STORIES

MORE

Engage - Grow - Innovate

www.egi.eu

This Issue

In this edition of *Inspired*:

- Ignacio Blanquer presents a solution to deploy custom elastic clusters on the EGI Federated Cloud
- Małgorzata Krakowian explains how SLAs can be used to get services from the providers to the research communities
- Diego Scardaci writes about how an SLA is connecting the BILS research community to three national cloud providers
- Marco Verlato updates us on the latest developments in accelerated computing in EGI
- Roberta Piscitelli summarises the Open Science Cloud that took place in Bari in November

and

- Zhiming Zhao writes about the Data for Science theme of the ENVRI+ project: software and solutions for the environmental sciences

Your feedback and suggestions are always welcome!

sara.coelho@egi.eu

Feeling poorly?
Boid Inclusion Body Disease affects many captive snakes across the world. EGI latest case study is about how the Chipster platform helps virologists to shake the disease.
http://go.egi.eu/BIBD
Image: Jakub Halun/wikicommons

Save the date! EGI Conference 2016: 6-8 April

The **EGI Conference 2016** will take place in Amsterdam Science Park, between 6-8 April.

More information on the event shall follow over the next few weeks and registration is scheduled to open in early February.

http://go.egi.eu/conf2016

Custom elastic clusters to manage Galaxy environments

Ignacio Blanquer presents the EC3 solution against cumbersome protocols and zombie virtual machines

The problem: high-maintenance VMs

The use of clusters - a set of computing nodes orchestrated through a workload management system and normally sharing a disk - is extremely common in many scientific problems. Users submit jobs through command line or through web portals that interact with such computing clusters.

Migrating this approach to cloud computing brings some issues and some benefits. On one side, configuration of multiple Virtual Machines may require system administration skills which are not present at the user level. On the other side, the intrinsic elasticity and availability on demand with cloud computing is extremely beneficial to reduce the number of unnecessary active machines. In a commercial environment, this will lead to a save reduction, and in a scientific environment to a better usage of the resources.

The solution: elastic clusters

Elastic Compute Clusters in the Cloud (EC3) is an open-source technology developed by the Polytechnic University of Valencia, that can be used to configure and monitor computing clusters in EGI Federated Cloud, providing users with a simple-to-use tool to deploy a cluster adaptable to its real usage.

This is also beneficial to EGI as resource provider because it avoids the proliferation of zombie VMs – VMs that remain active consuming resources if a user forgets to shutdown them or if they become inaccessible to the users that deployed them.

EC3 is powered up with Infrastructure Manager - IM, a tool that can be used to automatically configure and fine tune complex applications.

The requirement: INRA needs Galaxy

INRA is a France's national centre for agricultural research and their scientists need a complete execution environment based on Galaxy, a user-friendly interface to launch jobs for researchers who dislike command line.

This requires configuring the front-end, downloading the data, configuring the nodes, booting up the system, and so on. This is a cumbersome list of tasks that requires a deep knowledge of the platform. New users may be discouraged.

Combining EC3 with IM allows to boot an elastic cluster with Galaxy on top and with minimal effort from the user.

The benefits

Scientists can now deploy Galaxy environments on EGI without support – even if their ICT skills are very basic. The set-up also provides higher isolation than the use of a shared public Galaxy installation, a more predictable quality of service and extended configuration capabilities.

EGI and other infrastructure providers can get rid of zombie VMs and shut down unused clusters that remain idle most of the time.

> *It very simple to connect our tools on the EC3 cluster, connect to the EGI Federated Cloud, and have a Galaxy interface deployed.*
>
> *And it was a real pleasure to see the tools easily available for the community.*
>
> Alain Franc, INRA

Who else can use EC3's elastic clusters?

The INRA requirement was part of the technical challenges of the LifeWatch Competence Centre and will be used in the context of this activity.

The system is platform-agnostic and it works with EGI, public clouds, OpenStack and ONE among others. This means that the elastic cloud solution provided for the INRA/LifeWatch CC problem can be applied to any other use case that shares the same requirements.
MoBrain and some of the other Competence Centres are already expressing interest.

More information

EC3 has been developed under GPL v3 license and can be downloaded from github.

EC3 url:
http://servproject.i3m.upv.es/ec3/

IM url:
http://www.grycap.upv.es/im

SLAs: how to get services from the providers to the research communities

Małgorzata Krakowian explains how a simple procedure brings enormous benefits

One of the greatest challenges faced by EGI is to offer an easy way for researchers to discover the services they need and agree on terms of use.

Currently, EGI is working with seven pilot communities to make this process easier and integrate the procedure within the IT Service Management framework based on the FitSM standard.

An important first step is to create a reliable, trust-based communication channel between the researchers and the providers to agree on the services, their levels and the types of support. The outcome of this process is a Service Level Agreement (SLA) document. SLAs are not legal contracts but as agreements they outline the clear intentions to collaborate and support research. The SLA provides many benefits:

Benefits for research communities:

> Better communication and clarity on expectations

> Increased confidence that services will be delivered

> Easier future planning of research activities

Benefits for resource providers:

> Direct communication with user communities and clarity on expectations

> Clear responsibilities and rules/policies concerning usage of the resources

> Recognition and greater visibility to the role of the provider by requiring an explicit acknowledgment

Benefits for EGI:

> Promoting the EGI service value with funding agencies and policy makers at national and European levels

> Being seen as mature partner

> Ensuring a foundation of a control process to what is being delivered in the EGI Federation

In practice, the EGI Foundation User Community Support team (UCST) is proactively looking for providers that can fulfil needs of research communities and help in the negotiation of the appropriate terms of use. Once an SLA is agreed, the UCST helps in coordinating the effort between the resource providers to enable the research community on the promised resources.

How do SLAs work? A practical example...

Research Community X is looking for cloud compute, HTC compute and File Storage services and contacts the **USCT team** at the EGI Foundation. The UCST collects the requirements regarding capacity and availability and contacts the **NGIs** that have expressed interest to support Community X's field of research.

Following a call for resources, the UCST team collected expressions of interest from the **NGI from Country Z** and the **NGI from Country Y**.

An **SLA** is then written to cover the delivery of the services provided by NGI Y and NGI Z to Community X, for a given period with possible extension. The **Operations team** at the EGI Foundation will monitor the resource providers' performance to see if they fulfil the promised availability of resources. In return, each resource centres will be mentioned in **acknowledgments** added to all papers published about the research performed on the resources in the SLA.

More information

If you would like to know how to better ensure resources and service levels, please contact us at

sla@mailman.egi.eu

SLAs: Connecting the BILS research community to national cloud providers

Diego Scardaci writes about how the BILS SLA is paving the way to a new form of collaboration

The **Bioinformatics Infrastructure for Life Sciences** (BILS), a distributed national research infrastructure supported by the Swedish Research Council (Vetenskapsrådet), has signed a scientific collaboration agreement with three resource providers of EGI Federated Cloud to guarantee support to life science researchers in Sweden.

The resource providers supporting BILS are:

> **INFN Bari** from Italy

> **TUBITAK** from Turkey

> **IN2P3-IRES** from France

and the agreement will stand for at least two years.

The total amount of resources available to BILS consists of 324 CPU Cores, 648 GB of RAM and around 7 TB of temporary and permanent storage. The greatest part of these resources are offered in a reserved mode (pledged) by INFN Bari (68 cores, 128 GB of RAM and 2 TB of storage) and Tubitak (96 cores, 256 GB of RAM and 5 TB of storage). IN2P3-IRES completed the EGI offer with additional resources (8 cores, 16 GB of RAM and 300 GB of storage) accessible in an opportunistic mode.

BILS is going to deploy its services in the EGI Federated Cloud using the new **VO vo.nbis.se**. The main services that will be soon available for the NBIS users are:

> **Pcons.net**: a Meta server for 3D proteins structure prediction

> **PconsC2**: Web-server for protein residue-residue contact prediction

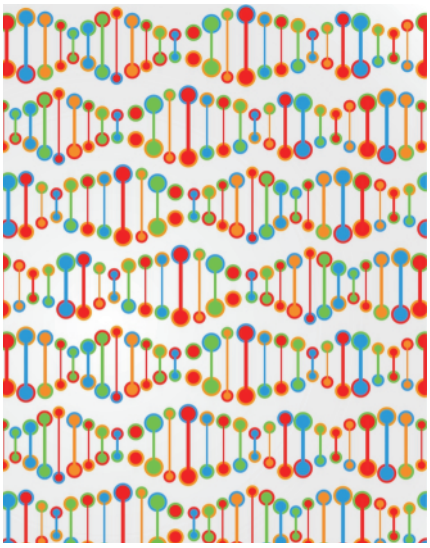
> **TOPCONS**: Consensus prediction of membrane protein topology

> **SCAMPI**: Prediction of membrane protein topology from first principles

The agreement has been formalised through a **Service Level Agreement** (SLA) that specifies service level targets and the responsibilities of both parties. EGI uses SLAs to allow access to cloud and HTC resources for a medium/long-term period, with a defined quality of service for a guaranteed period of time.

Furthermore, SLAs offer greater visibility to the service providers by including the possibility of being explicitly acknowledged in research communications and publications and help EGI as a whole in promoting its service value proposition with funding agencies and policy makers.

The BILS team has expertise in protein bioinformatics, mass spectrometry (MS), next generation sequencing (NGS), large-scale data handling, metagenomics, systems biology, biostatistics and RNA sequencing. BILS is predominantly offering support in various projects, depending on the user needs, as well as infrastructure and tools for bioinformatics analyses.



BILS works together with **SNIC** (Swedish National Infrastructure for Computing, who represents Sweden in the EGI Council) developing systems and strategies for long-term large-scale storage of bioinformatics data (MS proteomics data, NGS sequence data).

BILS is also the Swedish node in ELIXIR - the European infrastructure for biological information.

From 2016, BILS is now part of the National Bioinformatics Infrastructure Sweden (NBIS), a new consortium that puts together BILS, SciLifeLab Wallenberg Advanced Bioinformatics Infrastructure and Bioinformatics Platform and SILS - Systems Biology Infrastructure for the Life Sciences.

More information

BILS https://bils.se/

Accelerated computing in EGI: state-of-the-art

Marco Verlato updates us on the latest technical GPGPU developments

The Accelerated Computing task of EGI-Engage started its activity in March 2015, with the goal of enhancing the existing EGI grid and cloud platforms by supporting the capabilities of the most popular hardware acceleration cards (for example: GPGPUs, MIC coprocessors).

After having collected the requirements from many EGI user communities (MoBrain and LifeWatch Competence Centres, MolDynGrid Virtual Laboratory, Virgo and LHCb physics experiments) at the Lisbon EGI Conference in May, the intermediate achievements of this activity (led by INFN, IISAS and CIRMMP researchers) were presented at the EGI Community Forum at Bari in November 2015.

Grid platform testbed

The testbed consists of three nodes (2x Intel Xeon E5-2620v2) with two NVIDIA Tesla K20m GPGPUs per node was made available at CIRMMP and installed with a Torque/Maui Local Resource Management System (LRMS) version natively supporting GPGPU resources. A modified CREAM-CE prototype was developed on top of this LRMS in order to allow the jobs requiring the use of one or more GPGPUs to be dispatched towards Worker Nodes which are connected to this kind of resources. This GPGPU-enabled CREAM-CE was opened to enmr.eu VO and used at the end of November by the UU research team to execute the DisVis application, a structural biology program that calculates the reduced accessible interaction space of distance-restrained binary protein complexes. This allows direct visualization and quantification of the information content of the distance restraints. The performance of the system was in line with the one provided by the local servers.

Cloud platform testbed

The testbed, build with a master node and two IBM dx360 M4 servers with two NVIDIA Tesla K20 accelerators, was made available at IISAS and installed with the OpenStack (Kilo version) Cloud Management Framework. The PCI pass-through virtualization model (i.e. the physical GPGPU card is tied to a single VM) was adopted to create VM images based on Ubuntu 14.04 with the needed GPGPU driver and libraries.

We used a molecular dynamics package as a test application. Comparison between runs on a VM and on a physical server showed a performance degradation of about 2-3% slower. The IISAS Cloud site was fully certified and integrated into the EGI Federated Cloud in October, and is now open to users of moldyngrid, enmr.eu and vo.lifewatch.eu VOs.

A guide on how users of EGI FedCloud can create and deploy their own GPGPU server is publicly available and a live demonstration of the MolDynGrid use case was shown at the Bari Community Forum.

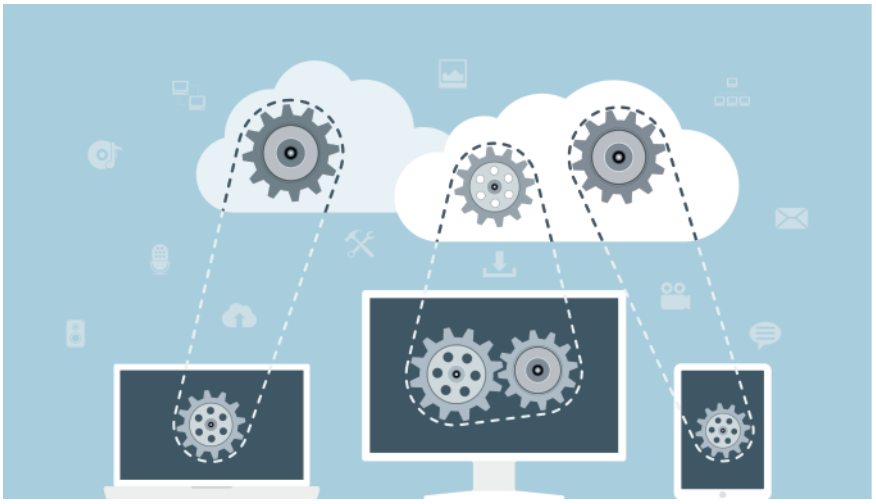
The next steps of the activity will be focused on the proper integration of the acceleration cards capabilities in the Information System, based on the future 2.1 version of the OGF GLUE standard schema, and in the Accounting System of EGI.

More information

Accelerated Computing is task JRA2.4 of the EGI-Engage project
url: *http://go.egi.eu/gpu*

Shaping the Open Science Cloud of the future

Roberta Piscitelli summarises the take-away messages of the Bari workshop



In November 2015, the EGI Community Forum hosted the workshop "Shaping the Open Science Cloud of the Future" co-organised with EGI, EUDAT, GEANT and OpenAIRE.

Building on the joint position paper "European Open Science Cloud for Research", the event attracted more than 100 participants from funding agencies, e-Infrastructures, Research Infrastructures, research communities, service providers and EIROs.

The workshop was an opportunity to reflect on the experience gathered through the development of the EGI Federated Cloud and on the new challenges and requirements from the seven Research Infrastructures cooperating with EGI in the context of EGI-Engage. What has been achieved so far and what is missing to make the Open Science Cloud come true?

Integration is a core objective for Open Science. Research Infrastructures (RIs) need to strongly cooperate to ensure a common strategy, common standards, tools and data validity and quality. A closer collaboration with industry and governments is a high priority to establish sustainable business plans.

Session 1 - Mission and vision

How to meet researchers' needs for Open Science?

The most important components of open science are:

> One-stop-shop

> Open access to scientific publications

> Open access to repositories and information

> Open access to datasets

> Access to shared computing resources

> Funding to support Open Science

> Openness of infrastructures

> Interoperability of tools, open source tools and software

> Standard procedures / policy

> Closer collaboration with industrial partners

A strong interaction between RIs is needed to define common standards and services.

Session 2 - Developing the Open Science Cloud

Towards a unified roadmap

The unified roadmap aims to provide services that address the following priorities:

> Introducing co-designed approaches to engage scientists and user communities.

> Promoting the culture of Open Science; this can be achieved through training actions/summer schools, engaging the new generation of scientists, and defining new educational programmes such as data science programs.

> Engaging governments to secure adequate investments to support digital science including data preservation and access.

> Ensuring interoperability among existing Open Science Cloud providers.

> Incentivising data sharing through European projects, creating an ecosystem that awards data sharers and producers.

> Establishing a credit system based on quality evaluation, to ensure proper authorship recognition in subsequent open data usage and associated reward mechanism.

What is Open Science?

Open Science can be described as:
"*the opening of the creation and dissemination of scholarly knowledge towards a multitude of stakeholders, from professional researchers to citizens*".

> Compatibility to enable validation, verification and reproducibility to ensure trust in science.

> Address standards and policies related to data sharing and privacy.

> Develop new skills in data science.

> Invest in open tools for data analysis and visualisation

Session 3 - Governing the Open Science Cloud

The last session focused on governance models for the Open Science Cloud (OSC). Governance ensures a good balance between different interests and effective decision-making. It can be defined as the set of structures, processes and policies by which the functions within an organization are directed and controlled so as to ensure adherence to the principles, yield business value and to mitigate risk. In the OSC, governance should be lightweight and responsible for defining the mission and values, the rules of engagement, conventions, standards and policies.

> ***For all and with collective empowerment***. There was consensus about the governance principles proposed by the e-Infrastructure joint position paper, calling for an OSC which is publicly funded and governed to ensure that its outcomes are driven by public good and scientific excellence. Governance should be also participatory, one of the foundation principles of the Commons, to make sure its services are accessible to all.

> ***Inclusive***. The stakeholders group should include besides the funding bodies, consumers, service providers/composite service providers (those who aggregate services into complex, value added solutions for their customer), integrators (offer data, application, and system integration services for both customers and IT service platform providers), regulators and standardization bodies. Involvement of the users was recognized to be one of the greatest challenges.

> ***With distributed responsibilities***. Funders should be responsible of ensuring the availability of the infrastructure services, while the researchers of the content. Various multi-stakeholders governance examples were discussed, including the internet model, the structure open source organizations allowing participation of different stakeholders, from individual to funding agencies, and a distributed governance model constituted of different structures – each governing different "domains" of the OSC – with overall lightweight coordination.

Possible Business models

> **Sustainable**. The audience agreed on the need of governmental investments to support the services needed by digital science. In modern science the funding of data gathering, curation, preservation, access and re-use are as necessary for scientific excellence as the availability of funding of human resources and other research costs. The increasing need of ICT services clashes with national investments that have been stagnant in the last year and will be further cut in the coming years in various countries.

> **Differentiated**. Multiple providers and different access models to services can coexist in the OSC, which will stimulate B2B relationships between publicly funded and commercial providers in any type of combination according to the rules of cost efficiency in service delivery and excellence of services thanks to co-design involving the users.

> **Challenging**. The majority of national funding and investments for research aims to promote national excellence, and is expected to remain as such in the future. The EC has a key role in promoting a culture of service sharing and exchange that allows cross-border access and enables borderless open science. Similar incentives should be studied to enable the same model at international level.

More information

The **Community Workshop on the Open Science Cloud: Shaping the Open Science Cloud of the Future** was co-organised EGI, EUDAT, GÉANT and OpenAIRE and took place on 13 November 2015 in Bari, as part of the EGI Community Forum. The Steering Committee was:
- Tiziana Ferrari, EGI.eu
- Wouter Los, Independent Expert
- Natalia Manola, University of Athens and OpenAIRE
- Per Öster, CSC and EUDAT2020
- Roberto Sabatino, GÉANT

Joint Open Science Cloud Position Paper: *http://go.egi.eu/OSCPP*

Data for Science: software and solutions for the environmental sciences

Zhiming Zhao on how the ENVRI+ project is making the most of e-Infrastructures

The Data for Science theme in the ENVRI+ project will establish an ICT approach for handling the lifecycle of scientific data based on the latest technologies offered by e-Infrastructure providers such as EUDAT and EGI. This approach will inspire interoperable solutions that can benefit research infrastructures (RI).

To understand the impact on our global environment of societal challenges such as climate change or pollution, scientists need to measure the environment on a large scale, and to understand the interactions between different environmental systems involving the atmosphere, oceans, geosphere and ecosystems. However, the complexity of environmental systems makes this task very difficult. This is despite all the existing ICT tools used to support research on environmental and earth sciences.

Most RIs are constructed to address specific research areas, and so using data and software across different RIs has proven challenging. RIs are now being strongly encouraged to support interdisciplinary research and to contribute to global initiatives such as Copernicus and the Global Earth Observation System of Systems (GEOSS). Many research infrastructures therefore face the same challenges for managing data: how to identify and cite data from different sites; how to catalogue data so as to allow users to search and access data from different infrastructures; how to support experiments conducted using resources from different remote infrastructures, etc. Sharing solutions to these common challenges will not only reduce development costs but also promote solutions across RIs.

The ENVRI+ project has set up the Data for Science theme to break the barriers between RIs and to provide those shared solutions. The theme must address a number of issues:

1) different RIs prioritise problems differently in their own development agenda;

2) RIs do not share a common view on the architecture and constituent components; and

3) there are few standards in consistent and pervasive use.

To tackle these issues, a common reference model plays a pivotal role by helping converge the concept vocabularies used by different RIs and research communities. This reference model will be provided based on a review of existing technologies and will provide the basis for developing solutions that bridge the meta information of data, software tools and resources.

The next step will be to investigate the architecture design of future environmental and earth science RIs and to review available software tools to prototype common solutions.

The Data for Science theme has a duration of four years. In the first six months, the main activities focus on requirement analysis, technology review and gap identification. Then, the Environmental Research Infrastructure Reference Model and the semantic linking framework developed in the previous ENVRI project will be refined and extended based on the evolving requirements and developed facilities of the RIs, and the evolving ICT capabilities and services available.

The Data for Science team is very pleased with the demonstrated involvement of specialists from all RIs and e-Infrastructures such as EUDAT and EGI in efforts to refine and execute the development plan.

More information

Zhiming Zhao is a senior researcher in University of Amsterdam, leading activities in the data for science theme in the ENVRI+ project.

In ENVRI+, EGI contributes to:
> WP: Reference Model
> WP9: Service validation and deployment
> WP10: Governance for sustainability
> WP15: Training and e-Learning
> WP17: Communications and dissemination