Inspired

ISSUE 18
FEBRUARY 2015

news from the EGI community

(Alain Gavillet / Wikimedia Commons)

TOP STORIES

MORE

European Grid Infrastructure

www.egi.eu

This Issue

In this issue...

> Peter Solagna introduces a pilot project to develop a platform for the long tail of science

> *Inspired* talks to France Grilles to find out how a database of publications can be deployed to demonstrate impactet

> Fotis Psomopoulos and Rafael Jiménez write about the EGI-ELIXIR integration project

> Daniele Bailo writes about the EPOS' architecture, requirements and areas of collaboration with EGI

> Diego Scardaciwrites about the growing FedCloud user community as the use cases reach the 50 mark

> Robert Lovas invites researchers and NGIs to try a new type of computing resource

> And the EGI CSIRT team warns cryptocoin miners that eyes are on them

Send your feedback and suggestions to: sara.coelho@egi.eu
Thanks!

Our next event is in Lisbon!
(Alain Gavillet / Wikimedia Commons)

The registration for the EGI Conference 2015 is open!

The EGI Conference 2015 is hosted by EGI.eu and IBERGRID, a partnership between the Portuguese National Distributed Computing Infrastructure (INCD) and the Spanish National Grid Initiative, with the theme of Engaging the Research Community towards an Open Science Commons.

The Lisbon meeting will be the first opportunity for the community to meet in the post-EGI-InSPIRE era and plan the work for the coming years. The programme will be focused on cross-disciplinary services with thematic days, where research communities and competence centres of different disciplines can join forces to discuss common issues.

The programme will also include the public launch of the EGI-Engage project, which was recently favourably evaluated and is expected to start in March.

Co-locations

The following events will be held in co-location with the EGI Conference 2015:

> EUBrazilCC workshop : Monday, 18 May

> Globus community meeting : Wednesday, 20 May

> Open Grid Forum (OGF44) : Thursday-Friday, 21-22 May

Conference dinner

The will take place at the Palácio da Rocha do Conde d'Óbidos on Wednesday, 20th May in the old Lapa neighbourhood, facing the National Museum of Ancient Art. The palace is the headquarters of the Portuguese Red Cross and was built in the 17th century for the Count of Óbidos. It's one of the few stately homes in Lisbon that survived the 1755 earthquake. The dinner will take place in the Red Cross Council, adjacent rooms and in the terrace with a spectacular view over the river Tejo.

More information

Save the date!
18-22 May 2015 - Lisbon, Portugal

EGI Conference 2015 website
http://conf.egi.eu

Registration
http://go.egi.eu/reg2015

Long tail of science platform: researchers, be our guests!

Peter Solagna introduces a pilot designed to make it easier for researchers within the long tail of science to access EGI's resources

The process of integrating new user communities in EGI has been successful many times, enabling access to EGI resources to hundreds of new users every year.

However, bringing new user communities to EGI involves some overhead effort from the EGI/NGIs side and the user community side.

The advantages of setting up a Virtual Organisation (VO) in EGI are multiple: best support, possibility to have dedicated SLA, detailed accounting and many more technical services, and it is the way to go for long-term collaborations or distributed collaborative communities.

But despite all this, in some cases the overhead of creating a VO and/or getting grid certificate discouraged new users from joining EGI. This has affected especially small research groups or individuals who need to get access to a limited amount of resources for a limited amount of time.

To support these use cases, that are commonly identified as the 'long tail of science', EGI is developing a dedicated platform to make very easy and quick for individual users to access EGI high throughput computing and cloud resources.

The problem

One of the most common barriers to access is the process of requesting a X509 certificate. The EUGridPMA federation has Certification Authorities (CA) in almost every European country to reach as many users as possible. But even with this widespread distribution of certification entities, some users do not have access to them when their home institution has no connection with EUGridPMA at all. Additional effort is needed to register VOs in the EGI Operations Portal, manage the VO on a VOMS server and be enabled at site level.

Features of the long tail of science platform:

Zero-barrier access: any researcher can get a start-up resource allocation
100% coverage: anyone with internet access can become a user
User-centric: User support for platform users is available through the NGIs
Realistic: Reuse existing technology building blocks as much as possible, requiring minimal new developments
Secure: Provide acceptable level of tracking of users and user activities (with no face to face id check)
Scalable: Can scale up to support large number resource providers, technology providers, use cases and users

How can we make this easier?

A dedicated User Management Portal for the long tail of science researchers is currently under development. The users will be able to register in the portal in just a few minutes with their eduGAIN credentials or an EGI SSO login.

Every user request will be assessed by the EGI/NGIs user support teams, to check that it is acceptable and that the information is correct. The verification will involve a phone call or an email from EGI. The target is to process requests within four weeks.

All our guest users enabled in the User Management Portal, will be assigned to a catch-all VO, enabled in pre-selected resource centres. The users can access EGI resources through dedicated science gateways deployed specifically for this use case and integrated with the User Management Portal.

The development of the platform is progressing and our target is to present it at the EGI Conference in Lisbon, and to open the platform for a trial period during that week.

If you are part of a research group, with no need for long-term agreements and dedicated resources, or if you are worried about the overhead of becoming an EGI user, we will have something for you very soon!

What is the long tail of science?

It refers to individuals or small groups of researchers, with limited resources and/or expertise, that find difficult to exploit modern computational methods, or share valuable data.

More information

http://go.egi.eu/LToSpilot

France Grilles: creating a publications' database to show the infrastructure's impact

The French NGI has an online database of +1,500 scientific papers, published thanks to the computing resources and services they provide. Inspired *talked to Geneviève Romier to find how they created the database and how they deploy it to demonstrate the scientific impact of the French e-Infrastructure.*

Inspired: How did this start? Why did you decided to start collecting the publications in a systematic way?

Geneviève Romier: In May 2011, when I started working with the user communities at with France Grilles, one of my first actions was to set up a database of publications.

This was a response to a request by the France Grilles funders to assess the scientific activity based on grid resources and to measure its impact. The database allowed us to establish statistics to evaluate the use of grids across time. The scope of the database was then extended to include cloud and France Grilles' resources in general.

I: How does it work? What's the system you use?

GR: The main effort was to set up the framework. We needed a simple and the most possible automated way to set up and maintain the database.

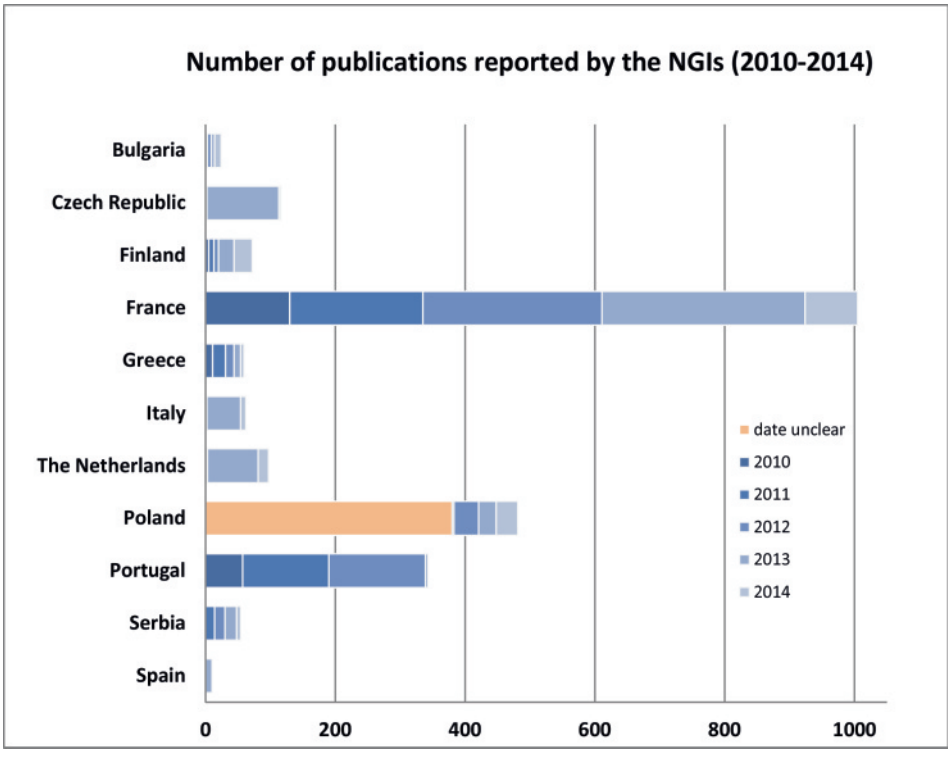
The technical part was almost easy. The obvious choice was HAL (https://hal.archives-ouvertes.fr/), which was set up more than 10 years ago as an open archive where authors can deposit scholarly documents from all academic fields. HAL allows us to build collections with specific stamps. We created a France Grilles stamp to manage the new database.

The organizational part was trickier. Legally, the deposit of a publication in an archive must be made in agreement with the co-authors and in the respect for the policy of the publishers. Depositing papers in HAL is part of the policies of several organisations and laboratories. But unfortunately this is not always the case.

I: What are the challenges?

GR: The first issue is to convince scientists to deposit their articles in HAL or at least to reference them. I use all possible means of dissemination: advertisements and also presentation of the statistics of the collection during the France Grilles scientific days, specific messages on our website and email lists, personal emails, personal discussions, presentation of the collection in national workgroups and EGI VTs... and it works even if it is not enough from my point of view.

The second issue is to identify the relevant publications already in HAL. I often need to read the text or the acknowledgements of an article to find if a paper is relevant. The 'Acknowledgment Statements' provided by the EGI Scientific Publications Repository Implementation Virtual Team is very helpful.



"We can now present statistics that are very helpful to show our funders that the our infrastructure and our services are a real tool for research."

I: How much effort did you spend?

GR: The technical effort to setup the collection took two or three meetings with HAL team and several days to customize a website for the database and to understand how to find relevant publications.

For dissemination, I take every opportunity to ask for publications and to show the statistics. Populating the database is a question of regular work and it takes a few hours per month. Personally I think that the return of investment is great.

I also want to acknowledge the help of the HAL team, all the communities that disseminate my numerous emails in their own circles, and all researchers and the laboratories librarians who reference their publications in HAL.

I: What is the involvement of the scientists?

GR: The scientists have to reference or deposit their publications in HAL. Depending on where they work, this may be done by librarians.

I: Do you ask scientists to acknowledge France Grilles?

GR: France Grilles set up an acknowledgement that is part of the Acceptable Use Policy of our national VO and that is advertised on our website.

I: What are the advantages of collecting publications?

GR: France Grilles knows now a lot more about its users and the different communities. We can present statistics that are very helpful to show our funders that the infrastructure and our services are a real tool for research. This complements very well the accounting information we have on CPU hours consumed. It shows the diversity of the disciplines and we can see the future generation of researchers emerging through the PhD thesis. The database gives a living image of France Grilles.

I: What is there to be improved?

GR: In my wish list the European one is the interconnection of HAL with OpenAire and all national repositories to mutually share at least the publications references. This interconnection could help all of us to mutualize our work and to better know who our users are and how to better support them.

More information

Geneviève Romier is NGI International Liaison (NIL) of France Grilles.

France Grilles' archive of publications: *https://hal.archives-ouvertes.fr/FRANCE-GRILLES*

France Grilles - http://www.france-grilles.fr/

France Grilles, the French NGI, was created in 2010 as a Scientific Interest Group by eight organizations including the French ministry of Research and Higher Education. It aims at building and operating a multidisciplinary national Distributed Computing Infrastructure open to all sciences and to developing countries. France Grilles provides an open space for collaborations within and across disciplines and organizations.

France Grilles is operated by CNRS' Institut des Grilles et du Cloud and has three main objectives:
> establish and operate a national production grid and cloud infrastructure
> contribute with the other countries involved to the EGI European e-infrastructure
> strengthen synergies and collaborations between teams using grids and clouds for scientific production and doing research on the grid and cloud

EGI–ELIXIR project: Integrating datasets for bioinformatics

Fotis Psomopoulos and Rafael Jiménez write about a joint project that will open new door for life science research

Image: ELIXIR-Europe

The complexity of datasets

Bioinformatics is an interdisciplinary domain, combining computational methods and statistical approaches to advance research in life sciences. As a field, bioinformatics relies heavily on public reference datasets and benefits from increasing compute capabilities to run algorithms. But the ways to turn all the available capacity and data into knowledge have not grown as fast and users struggle to discover, download and compute with the constantly expanding volume of data.

Take the areas of epigenetics and metagenomics, for example. For both, the bottlenecks in the analysis workflow are comparisons against reference datasets or computing with multiple datasets. Given the amount of data involved, users tend to opt for local datasets, often within the home institution. This means that researchers end up preferring limited studies, missing out in the process significant data patterns and motifs that could emerge from the inclusion of Big Data techniques and methods.

The bottom line is that we have an increasing demand to compute the data where the datasets are. Some EGI members already host some biological reference datasets across the infrastructure. EGI, however, does not yet provide discovery capabilities, nor guidelines to replicate additional datasets onto EGI sites.

About ELIXIR

ELIXIR is a pan-European research infrastructure for biological information, uniting life science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. One of its early services is the ELIXIR Registry, which lists all its available science databases and analytics tools.

url: http://www.elixir-europe.org/

A combined EGI-ELIXIR solution

EGI will join forces with ELIXIR to solve this problem through a 9-month project to develop ways to facilitate discovery of existing reference datasets in EGI. The project will also develop and deploy new services to allow replication of life science reference datasets by data providers, resource providers and researchers.

The joint project will involve ELIXIR, the NGIs from Greece, Italy, Poland and Slovenia and the EGI.eu team in Amsterdam.

The project will:

> Identify key datasets suitable for replication in EGI, including datasets already existing in the infrastructure, and set out the corresponding standards and procedures;

> Develop a dataset registry for the Applications Database;

> Evaluate and propose analysis

tools to work with data replicas;

> Promote integration with the ELIXIR Registry.

Benefits

The project will benefit individual researchers the most. By combining the computational infrastructure with the most frequently-accessed reference datasets, the project will enable users to execute highly detailed analysis workflows on a much greater scale than previously possible, and at a reasonable time frame.

For ELIXIR's members and NGIs, the project will provide:

> A set of tools to allow for more balanced load on storage resources across their sites;

> A pilot infrastructure with key datasets for life science analysis workflows and a registry with information for users about the available reference datasets and tools, and

> The technical knowledge to guide the setup of production infrastructures based on the pilot infrastructure.

For the EGI community as a whole, the project will:

> Increase what we know about tools, methods and solutions to replicate large datasets onto the e-infrastructures (grid, cloud);

> Develop tools and best practices that are reusable by other scientific domains, and

> Will broaden EGI's scope from a compute infrastructure to a data infrastructure.

More information

EGI-ELIXIR project: *https://wiki.egi.eu/wiki/Integrating_Reference_Datasets*

Fotis Psomopoulos is a researcher in bioinformatics at the Aristotle University of Thessaloniki and an EGI Champion.

Rafael Jiménez is ELIXIR's Chief Technical Officer.

How to participate in the TNC15

Karim Mostafi

TNC is the largest networking conference for research and education, attracting more decision makers, managers, network and collaboration specialists, and identity and access management experts from research organisations, universities and industry every year.

Today, national research and education networks are at an exciting crossroads in society. They provide the basic infrastructure on which big science and big data is build, but they also address the needs of increasingly diverse user communities. Many questions around cloud computing, software defined networking or mobility remain a challenge. All of this in a time when privacy and security are not a given.

This year's TNC will address these topics through keynote speeches, technical sessions, lightning talks and demos.

You can participate in the TNC2015 by:

> submitting a proposal for a 5-minute lightning talk or a poster presentation (deadline 15 April);

> sending demonstration proposals until 15 April to: horvath@terena.org ;

> registering to the event;

> following live streams online;

> following #TNC15 on social media.

TNC15 will be held in Porto, Portugal from 15-18 June.

The conference programme is now available and the keynote speakers confirmed:

> John Day – The missing layer

> Manfred Laubichler – Detecting innovation in networks of collaboration: a graph theoretical approach

> Sarah Kenderdine – Digitising UNESCO heritage / data analysis

> Timo Lüge – Disaster Response in a Connected World

> Avis Yates Rivers – Evolution of diversity in information technology

> João Paulo Cunha – Smart cities

More information

TNC 2015 website: https://tnc15.terena.org

EPOS: an e-Infrastructure to integrate Solid Earth Science data

Daniele Bailo writes about requirements and possible areas of collaboration with EGI

The European Plate Observing System (EPOS) is an ambitious long-term integration plan addressing the major solid earth research infrastructures in Europe. For its large scale and extent, EPOS is a unique initiative which will foster new scientific discoveries and enable scientists to investigate the solid earth system in unprecedented ways.

EPOS ended its Preparatory Phase in October 2014 and, being included in the top three infrastructures in the prioritization list by Competitiveness Council, is – at the moment of writing this article – in the process of submitting the proposal for Implementation Phase.

EPOS architecture

A key goal of EPOS is to provide end-users with homogeneous access to services and multidisciplinary data collected by monitoring infrastructures and experimental facilities, as well as access to software, processing and visualization tools. Such a complex system requires a solid, scalable and reliable architecture in order to accommodate innovative features and to meet the evolving expectations of the heterogeneous communities involved.

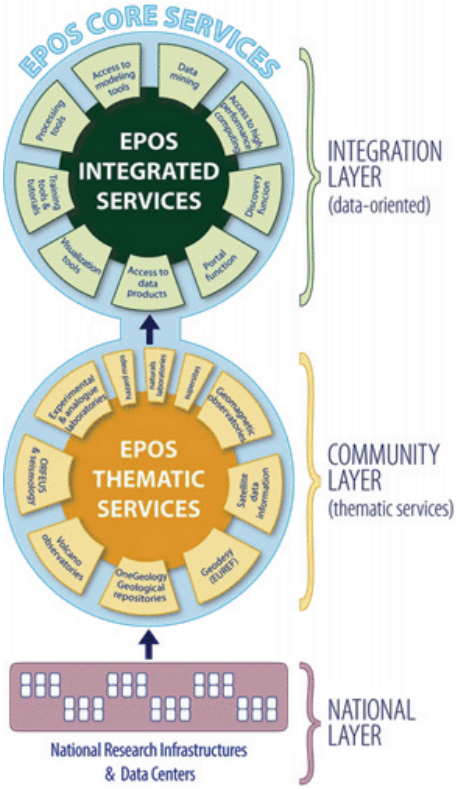
The overarching technical objective is to design and implement an Integrated Core Service (ICS) platform, supporting the standardised and transparent access to data, data products and services.

The EPOS community is organised in Thematic Core Services (TCS) - e-Infrastructures (often distributed) directly connected with national research infrastructures and data centres and providing data, data products, metadata and services related to one discipline only (e.g. seismology). TCS, therefore, are involved both as users and data/services providers.

One of the main components of the ICS system is the metadata catalogue based on the CERIF model, used to manage access and exploitation of users, software, data and resources.

Because one of the aims of EPOS is the integration and optimization of resources at European level, the project relies on IT initiatives which can provide computational services to users and implement fundamental modules, for instance AAAI software to manage secure access to resources. Such services are called Computational Earth Sciences (CES).

In this framework, organisations such as EGI can strongly contribute in the construction of EPOS: this will enable EU to optimise resources and to start tackling the challenging issue of resources procurement management, and EPOS to build on the basis of well tested and reliable technologies, thus creating a new, robust and innovative solid earth science e-Infrastructure.



EPOS architecture

EPOS is numbers

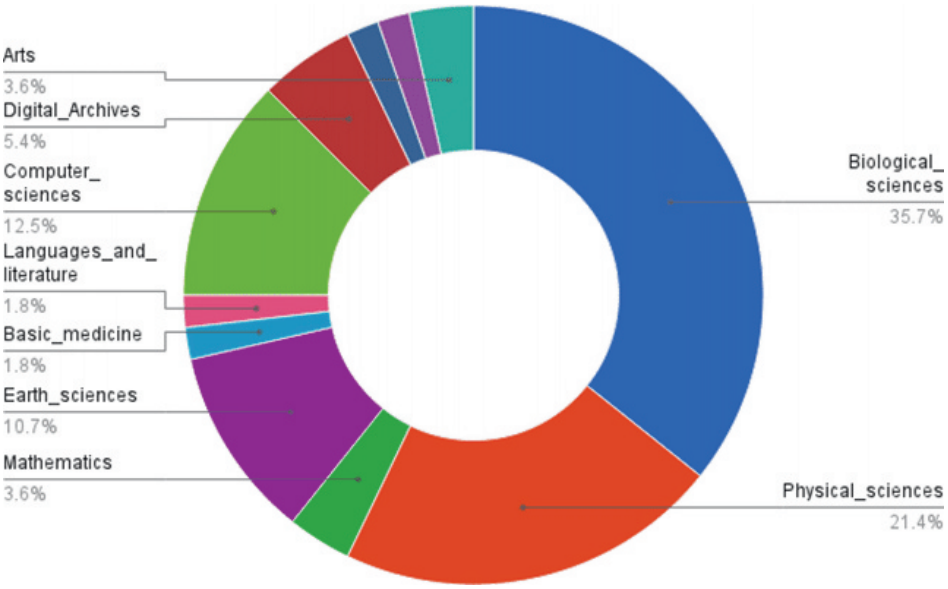
25 countries
256 national research infrastructures
+140 Institutions
+120 Experimental Labs
+170 Monitoring Networks
+7000 seismic and GNSS stations

More information

EPOS website
http://www.epos-eu.org/

EGI Federated Cloud: 26 communities, 50 use cases

Diego Scardaci writes about the growing usage of cloud resources among the EGI user community



Federated Cloud use cases per scientific fields (according to the EGI classification of scientific disciplines)

The EGI Federated Cloud was officially launched a year ago in Helsinki at the EGI Community Forum 2014.

Since then, EGI has devoted a big effort to setup a strategy to support the many communities interested to exploit this new capability.

The first task was to produce documentation and tutorials to allow scientists to understand the cloud paradigm and try the services. We then created a catchall VO to speed-up the time between the first contact with a community interested to use the FedCloud and their first real tests on the e-Infrastructure. The catchall VO, fedcloud.egi.eu, has been enabled in all FedCloud sites and can be used for application prototyping and validation for up to six months. This hugely reduced the effort to access the infrastructure, allowing user communities to quickly check if the infrastructure can be useful for them in the long-term.

Thanks to the VO and a clear process to bring the use cases from the initial tests to a full production status, we are currently supporting 26 communities and 50 use cases coming from different scientific disciplines. Nine of these are now in production and nine have pre-production status.

FedCloud use cases

A third of the FedCloud use cases come from the biological sciences. The Swedish (BILS) and Finnish (CSC) ELIXIR nodes are working to integrate some of their services on the FedCloud and the BioVel project is running four applications in production. The READemption use case, a next generation sequencing application developed by the Würzburg University in Germany, recently moved to production too.

Physical sciences represent about a fifth of the use cases including CERN's ATLAS, CMS and LHCb communities evaluate running their jobs in the FedCloud.

Earth sciences use cases include large communities as DRIHM and VERCE running, respectively, hydrology and hydraulic models and seismology applications. The well-known WRF climate prediction application has been integrated by the CHAIN-REDS project and Jena University successfully tested the FedCloud to run JAMS, a framework to build up complex hydrological models.

In addition, some private companies are showed interest to the EGI Federated Cloud services too. For example, Engineering, an Italian SME, is currently testing the FedCloud to run the HAPPI toolkit, a platform for data preservation developed in the context of SCIDEP-ES project, and the backend services of INERTIA, an application to retrieve energetic data of final-occupant of tertiary building.

The user base of the FedCloud is not yet complete. We are currently supporting use cases coming from basic medicine, arts, language and architecture, mathematics and computer sciences.

More information

http://go.egi.eu/fcc

Crowd computing resources at your fingertips

Robert Lovas invites EGI scientists and operators to try a new technology

In September, the Amsterdam region hosted a number of events related to Big Data and scientific computing, including the open meeting of Research Data Alliance and the EGI conference on Big Data and Clouds. The Crowd Computing 2014 meeting in Almere was a part of these events, with a focus on crowd computing, citizen science and desktop grid technology integrated in the mainstream of distributed e-Infrastructures.

The International Desktop Grid Federation (IDGF) hosted the event and invited David Wallom, from the Oxford e-Science Centre, to give a keynote talk addressing the integrated use of different types of computing resources. Their Climate Prediction programme, for example, involves citizens, crowd computing platforms and the EGI Federated Cloud.

The HADDOCK portal of WeNMR community is another instance of complementary use of crowd computing resources. HADDOCK is now sending a significant fraction of its compute-intensive jobs to IDGF's volunteer resources at production level. According to the EGI Accounting Portal, nearly 10,000 jobs have been processed in a few months period through IDGF.

IDGF supports local crowd computing initiatives as well. The University of Westminster Local Desktop Grid connects about 1,500 laboratory PCs into a private infrastructure with a high-level science gateway. Application areas include molecular docking simulations and 3D-rendering for computer animations.

Crowd computing is Green

The green policy introduced at the University of Westminster was a good argument to convince decision makers to adopt IDGF's crowd computing resources. The new report from the IDGF member SONY CSL entitled "Evaluate energy footprint of Desktop Grids" provides the evidence we need to demonstrate the green credentials of crowd computing.

To estimate the energy needs, many factors come into play. The most important are the performance/energy ratio of the computing nodes, whether the computation runs on a dedicated machine or as a background task, and whether the waste energy (heat) can be reused. Based on measurements with watt meters, it has been proven that the energy footprint can be extremely low if the infrastructure designer/operator follows the report's recommendations. Crowd Computing can contribute to the sustainability of e-Infrastructures, as acknowledged by the EIROforum's report on e-Infrastructure for the 21st Century: "*Volunteer computing initiatives across Europe have established production structures [...] such as the International Desktop Grid Federation that can support a growing range of application types with very modest operational and coordination overheads*".

The report also emphasises the importance of that "*such structures become an integral part of the e-infrastructure commons*". The new IDGF Operations Centre serves this goal as an umbrella for crowd computing (desktop grid) resources. The operations centre collects and makes such resources available, visible and easily accessible by the users of the EGI infrastructure. The IDGF OC welcomes both users with applications and resource centres (even NGIs) lacking of enough computational resources. A Memorandum of Understanding is to be signed in January to maintain the good relation between the foundations EGI and IDGF in the future.

More information

http://crowdcomputing.eu

http://desktopgridfederation.org/web/working-group-egi-collaboration

http://desktopgridfederation.org/road-map

Cryptocoin miners: the CSIRT is watching you

Leif Nixon and Sven Gabriel, from the EGI Security Team, explain why there is no point in trying to use EGI for mining cryptocoin

In a recent report to Congress from the Inspector General of the US National Science Foundation it was reported that a researcher had had his access to all US government funded compute systems terminated. Why such a drastic measure? This researcher had abused his access to NSF supercomputers by mining bitcoin currency.

Bitcoin, along with other related cryptocoin systems, are virtual currencies that are generated or 'mined' through a computationally expensive process. This makes large scale computing infrastructures very tempting for people looking for a quick profit.

This kind of incident is becoming common also in the EGI infrastructure. The EGI Computer Security Incident Response Team (CSIRT) has increased the levels of alertness and has already handled cases where legitimate users submitted unauthorized cryptocoin mining jobs to EGI systems.

In one especially serious case, substantial amounts of mining jobs were submitted over the 2013 Christmas holidays before being discovered in early January. The user had attempted to masquerade the mining activities as legitimate production jobs and also tried to hide his traces by planting false evidence of external attacks on the job submission machine. He failed and was caught.

The EGI CSIRT takes unauthorized access to the infrastructure very seriously, and the incident prompted an extensive investigation. When it could be shown without reasonable doubt that the mining jobs had been submitted by an insider, rather than an external attacker, the investigation was handed over to the concerned university for further handling and led to filling a complaint.

While it may seem very tempting and harmless to use otherwise idling systems to generate bitcoins, everyone must understand that this is not acceptable behaviour. Cryptocoin mining violates multiple policies and causes direct and significant costs for the hosting institution in the form of increased use of power and cooling capacity.

Besides moral and legal objections, it should also be noted that ordinary compute clusters are a bad choice for illicit mining activities. Clusters are architecturally unsuitable for mining and they number among the few systems where people actually pay attention to the CPU load.

Everyone tempted to do something stupid should also know that cryptocoin mining computations have a very different profile to legitimate scientific applications. This means that they stand out very obviously in the monitoring system. And because cryptocoin mining is a good indicator of illicit activities, intrusion detection systems such as Bro explicitly scan for miners.

Abusing the EGI infrastructure for cryptocoin mining is a low-yield activity with high risk of detection and, when they get caught, with a very negative impact on the offender's career.

To put it simply: kids, if you want to mine coins, don't try to do it on a supercomputer. It gets noticed!

"Everyone tempted to do something stupid should also know that cryptocoin mining computations have a very different profile to legitimate scientific applications."