EGI Use Cases

EGI: advanced computing for research

www.egi.eu

EGI's mission is to create and deliver open solutions for science and research infrastructures by federating digital capabilities, resources and expertise between communities and across national boundaries.

www.egi.eu

About **EGI**

EGI is a federation of almost **300 data and compute centres** worldwide and **21 cloud providers** united by a mission to support research activities. EGI is also committed to supporting business and innovation. The federation is governed by the EGI Council and coordinated by the EGI Foundation, with headquarters in Amsterdam, the Netherlands.

Since its establishment in 2010, the EGI e-infrastructure has been delivering unprecedented data analysis capabilities to more than **tens of thousands of researchers** from over **hundreds virtual organisations** covering many scientific disciplines. This publication showcases their work and highlights the diversity of EGI-supported science.

In March 2017, EGI became **the first European-wide publicly-funded e-infrastructure to be certified to ISO standards**, a sign of our dedication to continuously improve our service offering.

The scientists relying on EGI services work in large international organisations, in research infrastructures, projects, university labs, or as individual researchers.

Today, EGI provides both technical and human services, from integrated and secure distributed **high-throughput computing** and **cloud computing, storage** and data resources to consultancy, support and co-development.

Certifications

go.egi.eu/cert

ISO 9001
Certified Quality Management System
www.tuv-sud.com/ms-cert

ISO 20000
Certified IT Service Management System
www.tuv-sud.com/ms-cert

Key usage figures

(March 2015 – February 2017):

1.390 billion

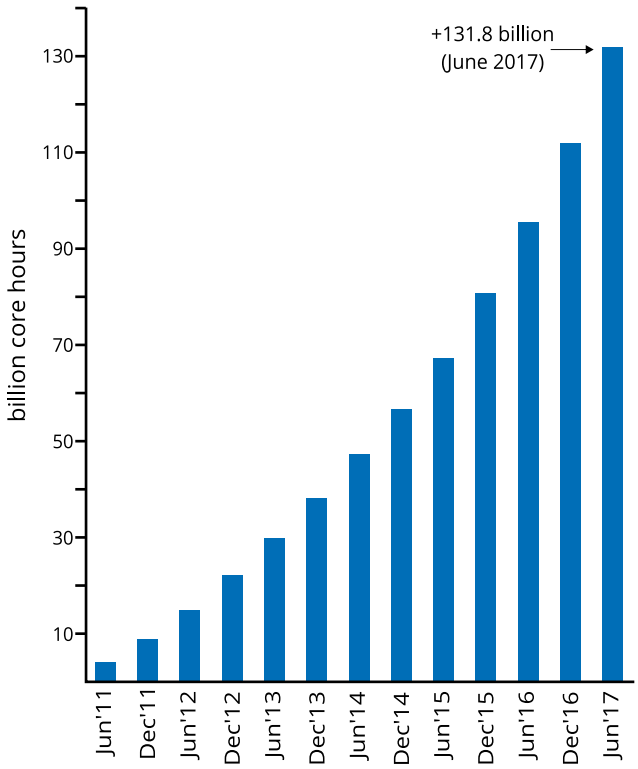
computing jobs submitted

44.800 billion

core hours consumed

Cumulative core hours consumed

at month indicated



Participants of the EGI Council

Belgium

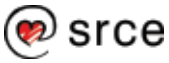
Germany

Slovakia

Bulgaria

Greece

Slovenia



Croatia

International Organisation

Spain

Czech Republic

Italy

Sweden

Estonia

Netherlands

Switzerland

Finland

Poland

Turkey

France

Portugal

United Kingdom

FY Republic of Macedonia

Romania

The almost **300 EGI federated data centres** are located mostly in European countries represented by the EGI Council.

The **EGI Council participants** are organisations representing national e-infrastructures and one European Intergovernmental Research Organisation.

Additional data centres are hosted by other European countries, not part of the EGI Council, and integrated data centres in Canada, USA, Latin America, North Africa and the Asia-Pacific region.

The complete list of EGI federated data centres is available online at: egi.eu/federation/data-centres/

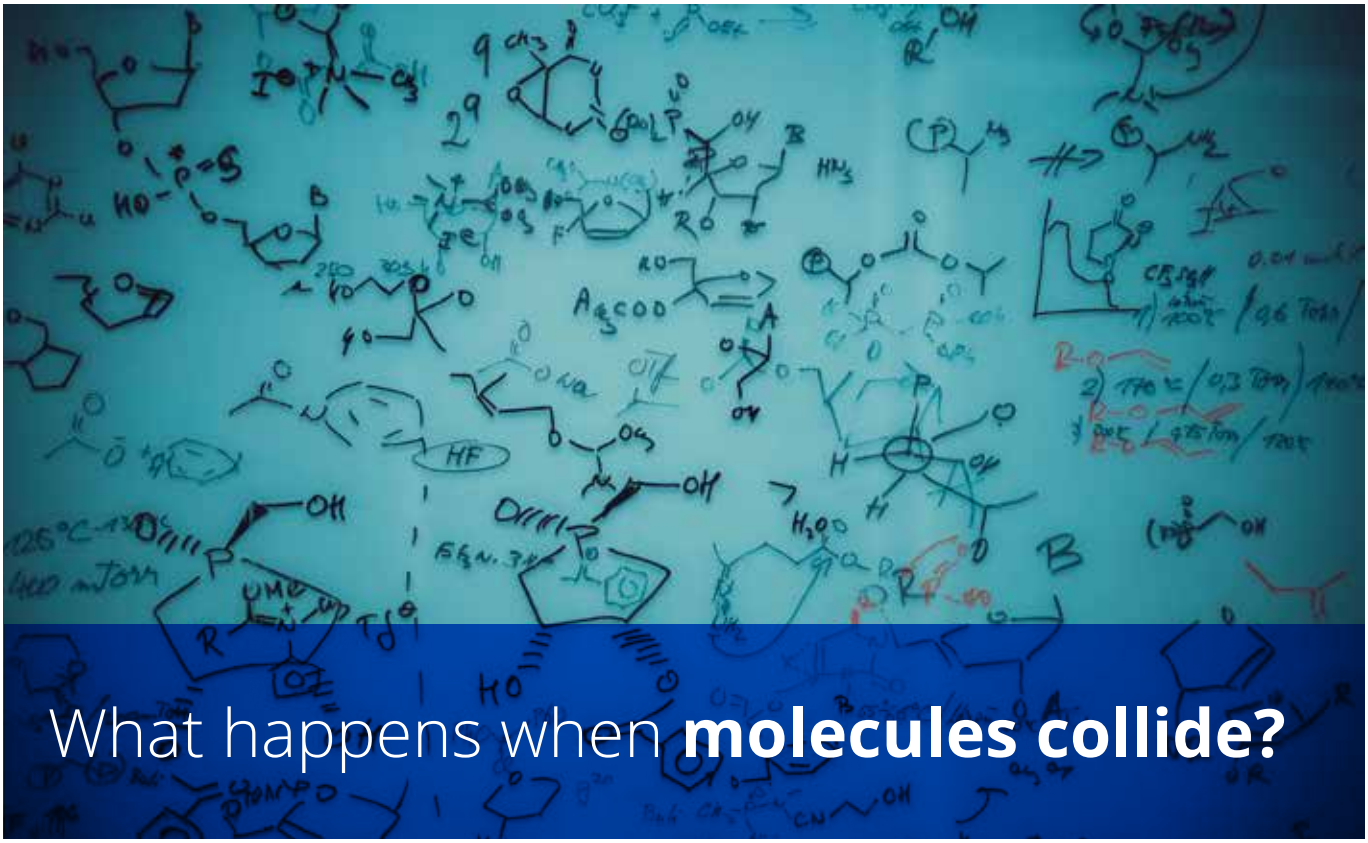
EGI Federated Cloud

The EGI Federated Cloud is a IaaS-type cloud, made of academic private clouds and virtualised resources and built around open standards. Its development is driven by requirements of the scientific community.

The Federated Cloud architecture is based on the concept of an abstract Cloud Management Framework that supports a set of cloud interfaces to communities. Each Federated Cloud provider operates an instance of this framework according to its own technology preferences and integrates it with the federation by interacting with EGI core components (e.g. AAI, service registry, accounting, monitoring).

Integration is performed by using public interfaces of the supported CMFs, thus minimising the impact on site operations. Providers are organised into the Open Standards and OpenStack realms, each realm exposing a homogeneous interface.

Federated Cloud Providers:



What happens when **molecules collide?**

How High-Throughput Compute helps researchers to simulate chemical reactions

Chemical reactions are at the core of everything that happens in the Universe. From the thermonuclear fusion that powers the Sun, to how antibiotics help to fight pneumonia, everything depends on what happens when molecules collide and interact to form new compounds.

Chemist Ernesto García, based at the University of the Basque Country in Vitoria (Spain), creates computational models to describe chemical reactions. Having good theoretical models to predict molecular behaviour means that simulations will be realistic and useful to tackle research problems in the real world.

Accurate models take into account many parameters (e.g., shapes of molecules or thermal properties). García uses the **GEMS workflow** to streamline the computational work of the calculations.

GEMS was developed by the team of Antonio Laganà at the University of Perugia in Italy and is powered by High-Throughput Compute resources made available by the **compchem virtual organisation.**

García and his collaborators have published eight papers in two years in journals covering many scientific fields, from astronomy to industrial processes or theoretical chemistry.

One example is the chemical evolution of interstellar clouds - amalgamations of gas, plasma and dust scattered across the Universe. In Rampino et al. 2016, García and his team modelled the formation of C2+ from one atom of carbon and CH+, a radical common in interstellar space. They found something surprising: the C2+ rates of formation are several orders of magnitude different from the values used in current astronomical models.

SELECTED OUTPUT

Rampino et al. 2016. doi: 10.1093/mnras/stw1116
Garcia et al. 2016. doi:10.1021/acs.jpcb.5b06423.

31 million core hours

2.5 million compute jobs

Providers:
compchem VO, supported by 17 federated data centres in France, Greece, Italy, Poland and Spain.

The genetics of *Salmonella* infections

How Cloud Compute helped scientists to understand what happens when a human cell meets Salmonella

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.

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.

Konrad Förstner, a bioinformatician working at the University of Würzburg in Germany, and his colleagues analysed the RNAs produced by the Salmonella and the host at the same time, in the same experiments.

First, the team infected human cells with Salmonella. Then they analysed the combined RNA from the both organisms with READemption, **a cloud compute pipeline** designed to process the computational tasks.

The 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.

OUTPUT

AJ Westermann et al. 2016. Nature. doi:10.1038/nature16547

"EGI Cloud Compute helped us to handle computational demand peaks when new data sets arrived and that sped up the whole process significantly."

Providers:
GWDG (Germany) and IFCA (Spain), part of the EGI Federated Cloud

Small settlements coalesce into **larger cities**

How the OpenMOLE platform and High-Throughput Compute helped to validate a long-held theory in geography

Cities have a life of their own, governed by interactions between people, resources and other settlements. These balances are delicate and can change: what looked like a humble village many centuries ago can today be a huge metropolis, while some of the great cities of the past are now nothing more than ruins.

Understanding the dynamics of cities is key to plan the future and for that we need accurate models, able to take into account all the complexities of an ever-changing system. But how good are these models?

Denise Pumain and her team from the ERC Geodivercity have focused on SimpopLocal – a computer model that simulates the evolution of agriculture-based villages under strong environmental constraints. The model considers six parameters to account for population, resources and innovation. Each simulation covers the equivalent of 4,000 years.

The problem is that the historical record is not complete. We don't have precise economic or demographic data covering the last 4,000 years of a city so an alternative had to be found to validate the SimpopLocal model.

The team gave 10 different values to each of the six parameters, covering a wide range of possible scenarios. This created a bottleneck because the number of combinations quickly escalated to millions and manual checking became impossible.

They used the **EGI High-Throughput Compute** service to handle the lion share of the work and the **OpenMOLE platform** to manage the workflow of the calculations.

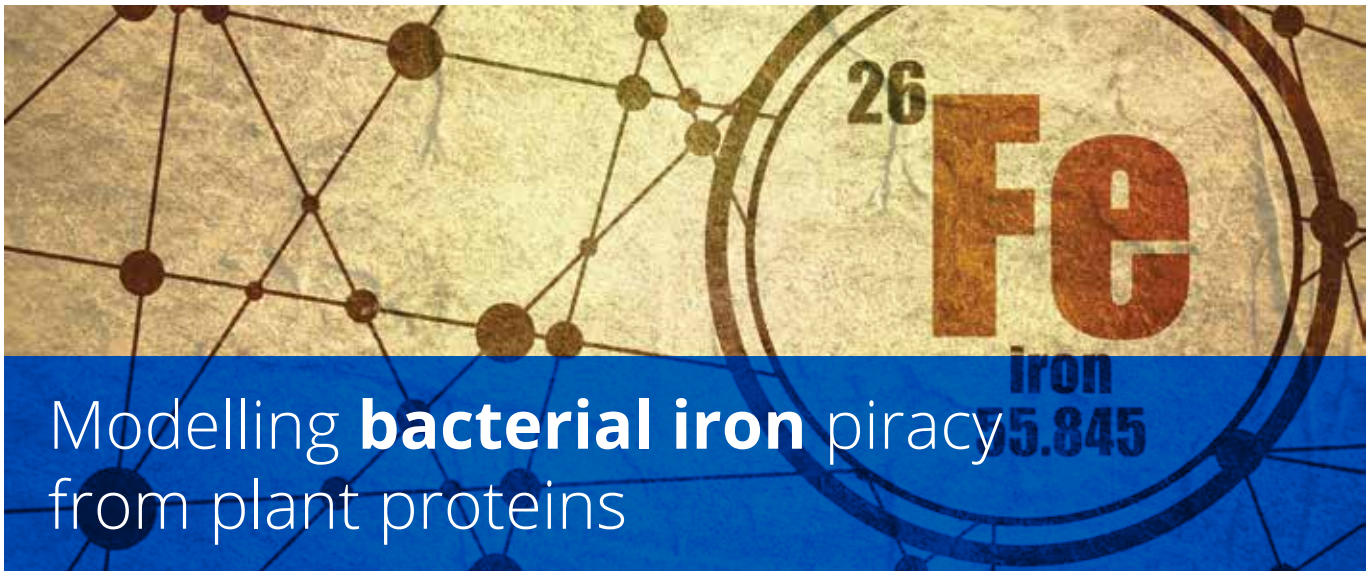
The work published as a book (Urban Dynamics and Simulation Models) concludes that the coalescing small settlements theory is a good framework to understand the development of cities.

OUTPUT
D Pumain et al. Springer. ISBN 978-3-319-46495-4.

200 thousand jobs

50 million core hours

Providers:
complex-systems VO, supported by 10 federated data centres in France and Greece



Modelling **bacterial iron** piracy from plant proteins

The HADDOCK portal helps to show that bacteria have a special drawbridge to steal iron from plant proteins

Iron is an essential element for life. Mammals, for example, use it to carry oxygen in red blood cells, plants need it to transport electrons. Bacteria are not an exception and they also need iron to infect other organisms and survive. But iron is not readily available in their environment. Animal and plant cells 'hide' their iron inside proteins to prevent bacteria from getting it and stopping infections before they start. So bacteria have to come up with mechanisms to get the iron they need.

Rhys Grinter, a microbiologist based at Monash University in Australia, investigated how Pectobacterium sources iron from the plants it infects.

He and his colleagues found that Pectobacterium cells have a receptor – dubbed FusA – specially adapted to grab the iron from ferrodoxin proteins.

They first determined the molecular structure of FusA from Nuclear Magnetic Resonance (NMR) and X-ray crystallography data. Then the team used the **HADDOCK docking tool** to simulate how FusA binds to plant ferredoxins.

"HADDOCK consistently ranks at the top of protein prediction experiments and is one of the best programs available for molecular docking," says Grinter. "This allowed us to place a high degree of confidence in its predictions and to present a credible representation of the FusA/ferredoxin complex for our paper."

The team learned that FusA is a glove-like structure able to grab plant ferredoxins and squeeze them across the outer membrane of the bacteria. The structure of the bacterial plant ferredoxin receptor FusA has been published in Nature Communications.

OUTPUT
Grinter et al. 2016 Nature Communications
doi:10.1038/ncomms13308

HADDOCK facts

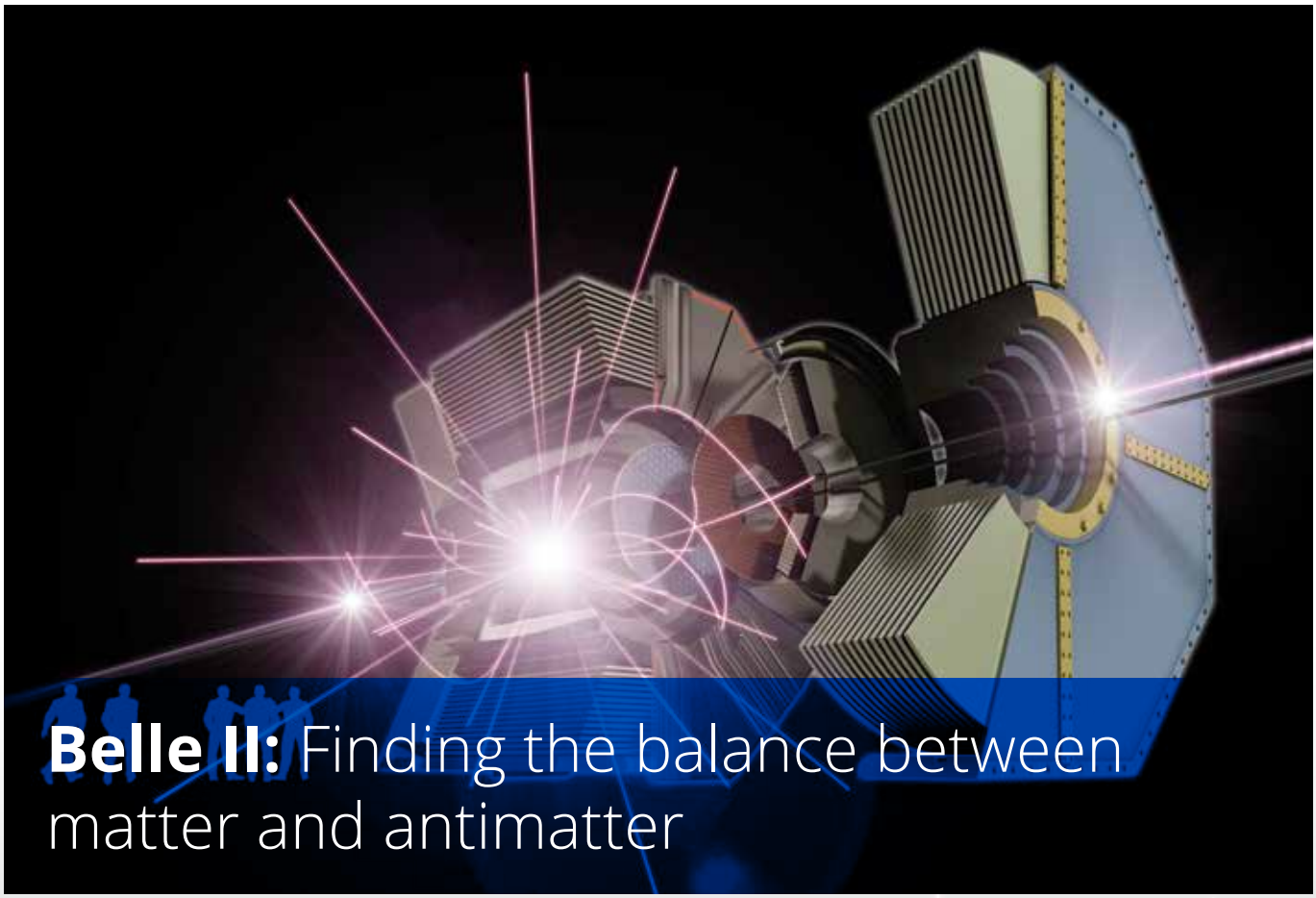
130,000 submissions

+8,000 users

8 million jobs/year

"HADDOCK consistently ranks at the top of protein prediction experiments and is one of the best programs available for molecular docking."

Providers:
HADDOCK is supported by the national e-Infrastructures of Belgium, France, Germany, Italy, the Netherlands, Poland, Portugal, Spain, UK, and other international organisations.



Belle II: Finding the balance between matter and antimatter

Credit: KEK/Rey.Hori

High-Throughput Compute and Storage solutions for Belle II's computational challenges

Belle II is an international collaboration put together to record and analyse the experimental data collected by the SuperKEKB accelerator in Japan. The experiment involves around 700 scientists from 23 countries and regions and builds on the success of Belle, a previous phase that ran between 1999 - 2010.

Belle II is looking into the imbalance of matter and antimatter in our Universe and relies on EGI High-Throughput Compute and storage services to analyse and share their data.

The science case

In the beginning of times, right after the Big Bang, the amount of matter and antimatter in the Universe was balanced. But today everything we observe on Earth and in space is made of matter only.
"If the behavior of the matter and antimatter were the same, the present imbalance will not happen," explains Takanori Hara, the Computing Coordinator of the Belle II collaboration. "So the imbalance implies something tilted this balance."

The original Belle experiment was designed to understand the difference in the behaviour of matter and antimatter, first postulated in 1973 by Japanese physicists Makoto Kobayashi and Toshihide Maskawa as the so-called CP-violation.

Their theory predicted that the effect of difference between matter and antimatter emerges in "B" meson pair system. The work of the Belle team confirmed the prediction with experimental results and Kobayashi and Maskawa won the Nobel Prize for Physics in 2008.

"However, the Kobayashi-Maskawa theory is not enough to explain the present matter-dominated universe," says Hara. "To realize the current universe, we need a new mechanism beyond the Kobayashi-Maskawa theory and, although we have many theories now, we do not know which one is right."

The Belle II experiment at an upgraded SuperKEKB accelerator plans to search for the new mechanism of the CP violation beyond the Kobayashi-Maskawa theory.

The computational challenges

The SuperKEKB accelerator is expected to generate an amount of data similar to the ATLAS detector of Cern. The number of institutes part of the Belle II experiment is however much smaller and "because of this, we are always facing the problem of resources," explains Hara.

Belle II uses **EGI High-Throughput Compute and storage resources** made available by the **Belle Virtual Organisation** and provided by 24 federated data centres.

Since 2009, Belle II has consumed over **1.6 billion CPU hours** (HEPSPEC, elapsed time) of compute time and submitted more than **16 million compute jobs.** The team manages the workload with DIRAC, a system originally developed for LHCb to guarantee the interoperability of heterogeneous computing systems. Belle II also uses GGUS as the issue tracker and GOCDB as the downtime information system.

In addition to EGI Federation resources, the Belle II experiment also benefits from Cloud, HPC, local clusters.

Why EGI?

The Belle II team needs to establish a distributed computing infrastructure with a limited manpower. "The EGI infrastructure has been already proven for the stable operation and scalability," says Hara. "That is the established technology, which is a very important feature for us."

Website: *www.belle2.org*

Resource Providers:
The data centres providing the most computing and storage resources to Belle experiments are:

Australia-T2
Hephy-Vienna
CA-MCGILL-CLUMEQ-T2
prague_cesnet_lcg2
FZK-LCG2
DESY-HH
INFN-T1
INFN-COSENZA
INFN-FRASCATI
INFN-LNL-2
INFN-NAPOLI-ATLAS
GRISU-UNINA
UNINA-EGEE
RECAS-NAPOLI
INFN-PISA
INFN-ROMA3
INFN-TORINO
JP-KEK-CRC-02
KR-KISTI-GSDC-01
CYFRONET-LCG2
TW-NTU-HEP
TW-NCHC
TR-10-ULAKBIM
UA-ISMA

Belle II experiment also benefits from Cloud, HPC, local clusters.

700 scientists
23 countries and regions

Belle's EGI services

High-Throughput Compute

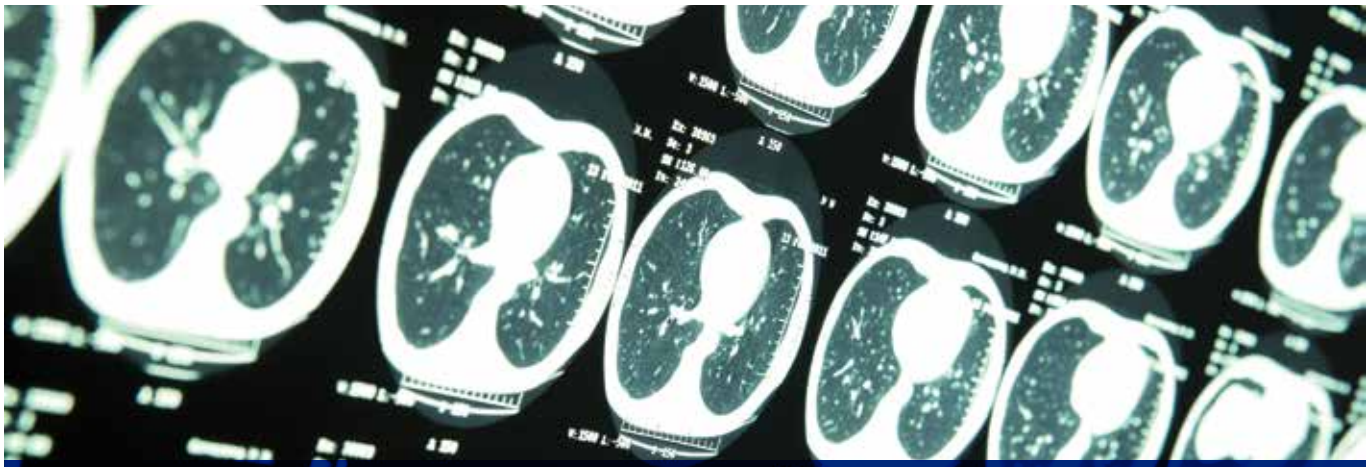
Online Storage

Belle's EGI usage
(2009-2016)

1.61 billion
core hours

16.4 million
computation jobs

"The EGI infrastructure has been already proven for the stable operation and scalability. That is the established technology, which is a very important feature for us."



Reducing the incidence of **epilepsy**

How HTC helps scientists to predict the development of epilepsy

Epilepsy is a chronic disorder that affects 2.4 million of people per year according to the World Health Organisation. Current research focuses on identifying markers that can predict the development of the disease before symptoms emerge.

Massimo Rizzi and his colleagues at the Mario Negri Institute for Pharmacological Research studied the problem using mice, which can develop epilepsy just as humans do.

They started by looking at the behaviour of mice and performed epidural electrocorticograms (ECoGs) to record what happened in their brain's electrical activity after exposure to a risk factor.

Then for the analysis, Rizzi and his colleagues used the High-Throughput Compute (HTC) and storage resources made available by the **Italian Grid Infrastructure** (IGI) and the **biomed virtual organisation.**

Using **High Throughput Compute** meant shortening the time required for the calculations. Rizzi estimates that using a single PC to complete the calculations of 25,000 jobs, they would have needed 54 days. Given that in total the team submitted about **200,000 jobs,** the time saved was in the order of years.

The analysis revealed an oscillation pattern – also called a dynamic intermittency – in the ECoGs of mice developing epilepsy. Applying an experimental anti-epileptogenic treatment successfully reduced the rate of the event. The results, published in Scientific Reports, confirmed that high rates of dynamic intermittency are a marker for the onset of epilepsy.

OUTPUT
Rizzi et al. 2016. Scientific Reports. doi:10.1038/srep31129

"The only reasonable way to describe the role of EGI resources in my daily work is that they make possible what is practically impossible to achieve using ordinary computing power."

200
thousand
compute
jobs

Providers:
biomed VO, supported by 60 federated data centres, managed by 11 EGI Council members
IGI, the Italian national e-infrastructure for research

Does ice water change the optical properties of **cirrus clouds?**

High-Throughput Compute helps scientists to better understand the impact of cirrus clouds on climate

Cirrus clouds, with their dramatic strokes of white against the blue sky, are more often associated with spectacular scenery than with weather either good or bad. But despite their feather light looks, cirrus clouds have a greenhouse effect and their contribution needs to be accounted for in climate models.

Scientists usually assume that cirrus clouds are planar and homogeneous. Céline Cornet, an atmospheric scientist based at the University of Lille in France, and her colleagues used this premise to investigate what kind of errors can be introduced from observing cirrus clouds from space.

To do this the team used a model based on the statistical Monte-Carlo method to trace how photons are reflected by clouds with different ice water distributions.

"For the study," explains Cornet, "we computed radiances for eleven cases of cirrus clouds and three wavelengths." On a single core computer it would have taken about three years to compute the necessary information for each example for a single wavelength. "For all cases, this leads to more than hundreds of years, which is obviously not possible," she adds.

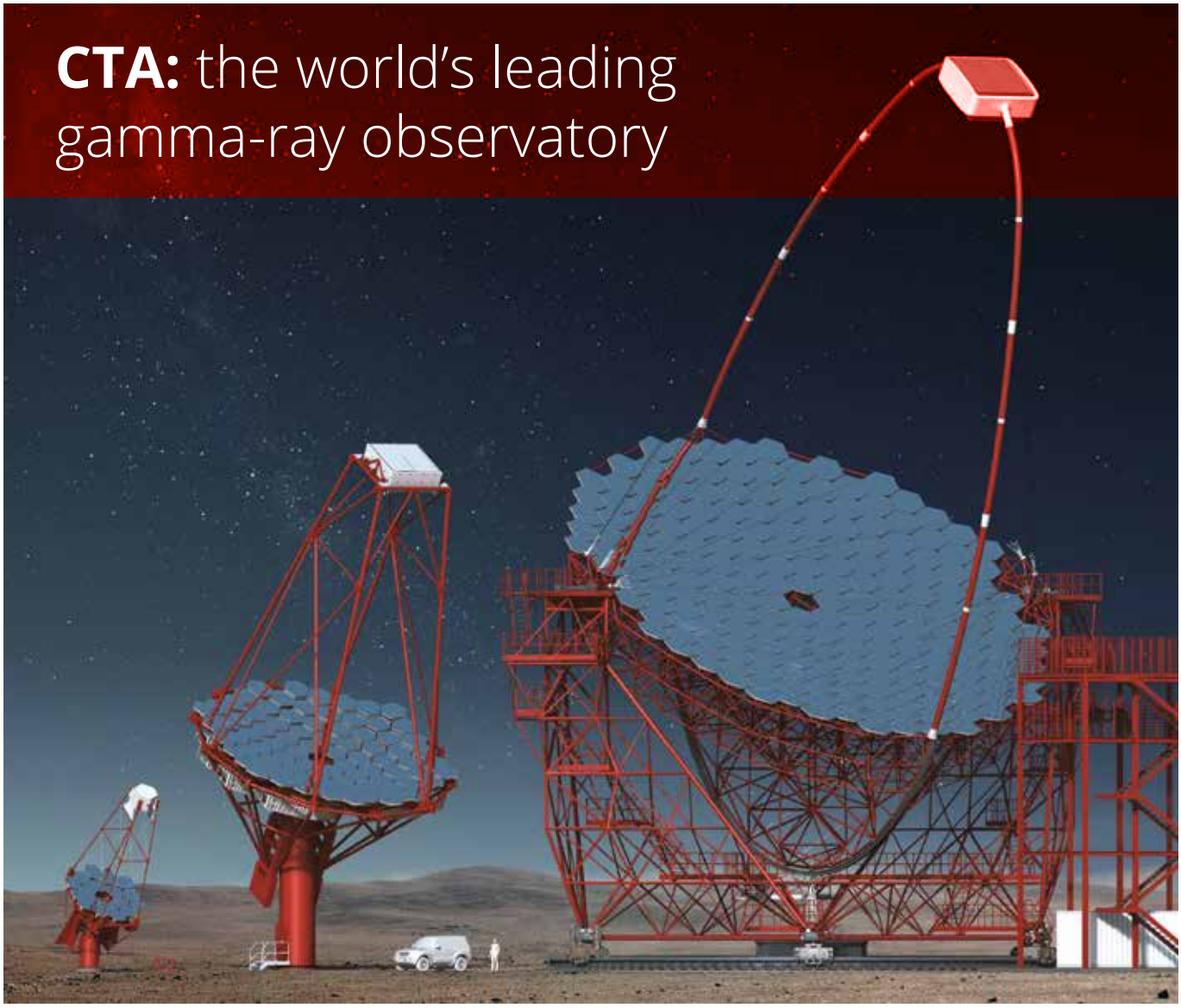
The team used the **High-Throughput Compute resources** made available by **France Grilles** (the French representative in the EGI Council) to run different batch simultaneously and expedite the analysis.

They concluded that uneven distributions of ice water in cirrus cloud cannot be neglected because they have an "impact on the retrieved parameters such as optical thickness, cloud emissivity and ice effective radius that are directly linked with the ice water content," says Cornet.

OUTPUT
Fauchez et al. 2015 Atmospheric Measurement Techniques doi:10.5194/amt-8-633-2015

Providers:
France Grilles, the French national distributed computing e-infrastructure

CTA: the world's leading gamma-ray observatory



Credit: G. Pérez, IAC, SMM

High-Throughput Compute and Storage solutions for CTA's computational challenges

The Cherenkov Telescope Array (CTA) brings together 1350 scientists and engineers from 32 countries with the goal of building the world's largest and most sensitive very high-energy gamma-ray observatory.

The CTA will be used to understand the role of high-energy particles in the most violent phenomena of the Universe and to search for annihilating dark matter particles.

The computational challenges

The CTA will be a distributed array of more than 100 telescopes built in La Palma in Spain and the European Space Observatory site in Paranal, Chile, and is expected to produce up to 27 petabytes per year for long term archive.

The computational challenges start right there: how to transfer all this data from the telescopes to scientists across the world? Then, the CTA infrastructure needs space to archive the data and enough processing power for data reduction and large-scale Monte Carlo simulations.

CTA aims to be the first ground-based gamma-ray public observatory. This means that a fraction of the observation time will be opened to the whole scientific community. Any scientist in the world will be able to submit an observation proposal to CTA and will have access to the corresponding data, which after a proprietary period, will become public.

This collaborative model, with the participation of the scientific community, implies other major challenges: "we need to make sure we have the

capability to provide a unified and efficient access to data, which will follow the common standards of the Virtual Observatory", explains Luisa Arrabito, the CTA computing grid technical coordinator, based at the Laboratoire Univers et Particules de Montpellier.

Why EGI services?

CTA relies on **EGI's High-Throughput Compute** and **Online Storage services** to manage its computational challenges during the project's preparatory phase.

According to Arrabito **the benefits are:**

- The possibility to aggregate resources from several sites, benefiting also of opportunistic resources
- The long term reliability of storage resources provided by the main sites supporting CTA
- Common and transparent data access for all CTA members

The compute and storage services are provided to the consortium via the CTA virtual organisation (vo.cta.in2p3.fr), one of the most active user groups of EGI.

Since 2012 the consortium has used EGI services to guide the choice of the best sites to host CTA telescopes in the North and in the South hemispheres. Once La Palma and Paranal sites were selected, "CTA also performed additional Monte Carlo simulation campaigns to determine the optimal array geometry" concludes Arrabito.

Website: *www.cta-observatory.org/*

1350 scientists and engineers from 32 countries

CTA's EGI services

High-Throughput Compute

Online Storage

CTA EGI usage
(2013-2016)

360 million
core hours

11 petabytes
data transferred

2 petabytes
in storage

11 million
computation jobs

Resource Providers:
The data centres providing the most computing and storage resources to CTA are:

CC-IN2P3
GRIF
IN2P3-LAPP
CYFRONET-LCG2
DESY-ZN
INFN-T1

The CTA also benefits from resources provided by the national e-Infrastructures of the Czech Republic, France, Germany, Italy, Poland and Spain.

New viruses implicated in **fatal snake disease**

Credit: Jakub Halun/wikicommons

How the Chipster platform is helping virologists to make sense of millions of virus genomes

Boid Inclusion Body Disease is a threat to boas and pythons in homes and zoos around the world and is often fatal. So far scientists know that the disease is somehow related to viruses from the arenavirus family, which are usually carried by mice and may cause haemorrhagic fever or meningitis if transmitted to humans.

Jussi Hepojoki, a virologist based at the University of Helsinki in Finland, and his team set out to sequence the genome of the viruses found in the samples of six snakes. The first step was to isolate the RNA of the viruses and sequence the samples using Next Generation Sequencing (NGS) techniques.

Hepojoki used the **Chipster platform** to handle the millions of sequences generated by the NGS analysis.

To assemble the genomes, Hepojoki ran the MIRA software using the computing resources provided by **CSC** – who represents Finland in the EGI Council.

"I think that without CSC and Chipster, I would not have been able to carry out the sequence assembly," said Hepojoki. "The power of 'normal' computers is just not enough for this type of work, or it takes a ridiculous amount of time."

The surprising conclusion is that all samples had more than one arenavirus, some of them from entirely new species. It seems that the viruses previously linked to this disease are only the tip of the iceberg.

OUTPUT

J. Hepojoki et al. 2015, Journal of Virology. doi:10.1128/JVI.01112-15

"The power of 'normal' computers is just not enough for this type of work, or it takes a ridiculous amount of time."

Providers:
Chipster is supported by CESNET (Czech Republic), RECAS-BARI (Italy) and CESGA (Spain), part of the EGI Federated Cloud. CSC is the national e-infrastructure of Finland

How to predict **social media trends?**

Cloud Compute helps to test new ways to detect trends on social networks

Social networks nowadays are big data production engines. Their analytics can produce insights on trending topics that can be used in various domains, from advertising to politics and even emergency situations. However, the prediction of a social network's topic as a trend needs to be first declared a trend by the social network itself (e.g. Twitter), and this can count as a classification problem. Another challenge is to manage massive data volumes and extract valuable information in a real-time fashion.

To address these problems, Athena Vakali and her colleagues at the Aristotle University of Thessaloniki in Greece worked on a new model of detecting social media trends in a near-real world context.

They started by using actual Twitter large-scale data threads and employed trend prediction under a framework designed in Lambda architecture.

Lambda architecture is a data-processing model capable of handling massive quantities of data using both batch-processing and stream-processing methods to provide views of online data. The team chose to use this model because it tackles the manipulation problems of both the volume and the velocity of data.

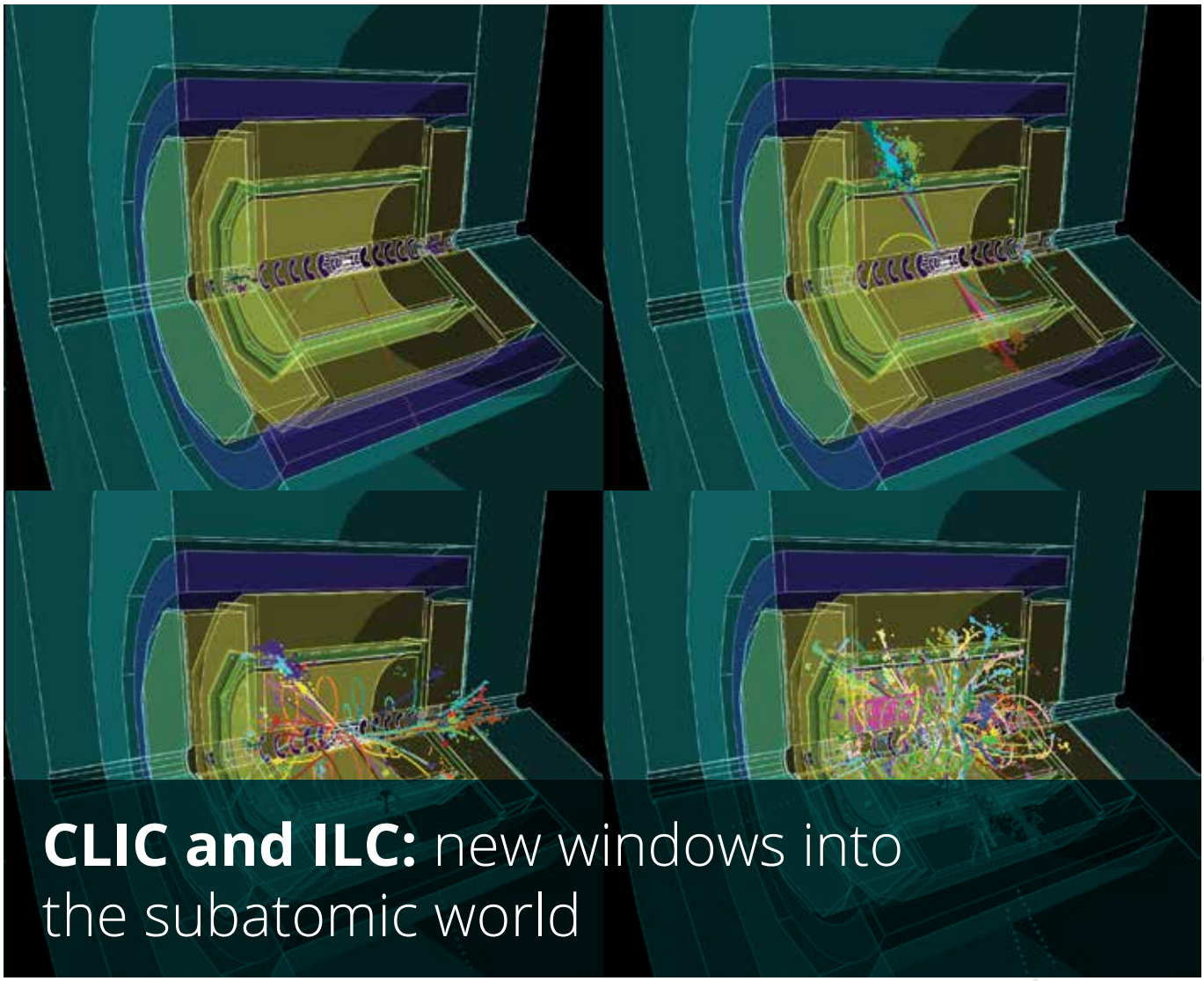
The **cloud resources** needed for the project were made available by **GRNET** through their **Okeanos cluster.**

Vakali and her colleagues found that almost 80% of the actual trending topics were classified as potential trending topics. The results, published in Advances in Big Data, validate the performance of the proposed research framework and emphasise its ability to early detect trending topics.

OUTPUT

Vakali et al. 2016. Advances in Big Data. doi: 10.1007/978-3-319-47898-2_20

Providers:
GRNET, the Greek national e-infrastructure for research
Okeanos is part of the EGI Federated Cloud.



CLIC and ILC: new windows into the subatomic world

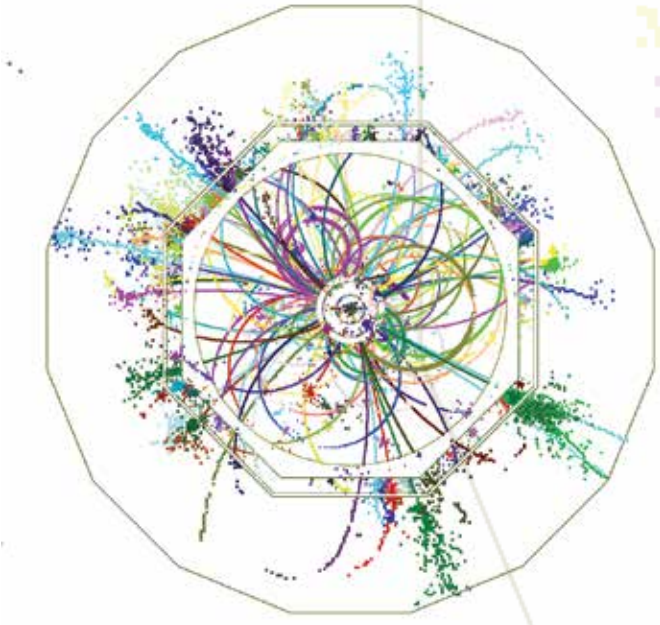
Credit: Akiya Miyamoto

High-Throughput Compute solutions for linear colliders

The Large Hadron Collider (LHC) operated by CERN has been home to some of the most interesting discoveries in high-energy physics. Thousands of scientists work on the data collected by LHC detectors and their successes are well known, with the Higgs particle as highlight.

But there is more physics to be discovered and the LHC is not the only window into the subatomic world.

The Compact Linear Collider Study (CLIC) and the International Liner Collider (ILC) are two collaborations set up to explore what happens when electrons and positrons (which are antielectrons) collide at high-energy. Using electrons and their antiparticles instead of protons (as in the LHC experiments) will allow scientists to collect a new range of high-precision measurements of the Higgs boson and to get a different view on high-energy physics.



Both collaborations bring together hundreds of scientists from multiple institutions and countries. They have succeeded in designing detectors that are now validated and ready for construction.

CLIC is one of the options for the next collider to be built near CERN at the border between France and Switzerland. The ILC is proposed by an international collaboration with Japan as a host country and is currently negotiated on the political level for a timely start-up.

110 scientists from about 20 countries

EGI + CLIC and ILC

The ILC virtual organisation (VO) was set up ten years ago to provide the computing and storage resources for the ILC collaboration. Both CLIC and ILC have used **EGI High-Throughput Compute** to simulate and reconstruct tens of millions of collision events during the design, testing and validation phases. The simulations helped the teams to evaluate and fine-tune the physics capabilities of the detectors.

In 2016, the ILC VO ran about **8 million computing jobs** and consumed over **220,000,000 core hours.**

The biggest challenge at the moment is to optimise the processing and analysis of the collision events. Every collision simulation consumes a lot of computing time to simulate and reconstruct. The ideal scenario is to have large files to minimize the number of High-Throughput Compute jobs. But on the other hand, large files may lead to problems with queue length limitations at the data centre level. Improving these splitting and merging processes is one area of improvement.

Website: *ILC: www.linearcollider.org/ILC*
Website: *CLIC: clic-study.web.cern.ch/*

CLIC & ILC EGI services

High-Throughput Compute

ILC VO EGI usage
(2016)

core hours **compute jobs**

Resource Providers:
Important contributions were also made by the national e-infrastructures of: France, Germany, Israel, Poland, Portugal, Spain, United Kingdom as well as the US and the Asia-Pacific region.

The EGI federated data centres that provide compute and storage resources to the ILC VO are:

CERN-PROD
CYFRONET-LCG2
DESY-HH
DESY-ZN
GRIF
IFCA-LCG2
IL-TAU-HEP
IN2P3-CC
IN2P3-LAPP
JP-KEK-CRC-02
RAL-LCG2
TECHNION-HEP
UKI-LT2-Brunel
UKI-LT2-IC-HEP
UKI-LT2-QMUL
UKI-LT2-RHUL
UKI-NORTHGRID-LIV-HEP
UKI-NORTHGRID-MAN-HEP
UKI-SCOTGRID-DURHAM
UKI-SCOTGRID-ECDF
UKI-SCOTGRID-GLASGOW
UKI-SOUTHGRID-BHAM-HEP
UKI-SOUTHGRID-BRIS-HEP
UKI-SOUTHGRID-CAM-HEP
UKI-SOUTHGRID-OX-HEP
UKI-SOUTHGRID-RALPP
UNI-FREIBURG
WEIZMANN-LCG2

How to **reduce flood risks** in the Netherlands

High-Throughput Compute power can help to improve dyke stability and safety

Geoscientist Yajun Li and the team led by Michael Hicks and Philip Vardon at the Delft University of Technology studied the Dutch earth structures and analysed their properties and failure threats. Their work led to the design of a new risk assessment framework that defines ways of evaluating the stability and failure consequences of the Dutch soil constructions.

They analysed the soil properties of dykes to see if they can predict their reliability (i.e., the likelihood that a dyke is functional for a given period of time) with the 3D random finite element method.

Then they tested the model using the computing resources provided by **SURFsara** (the Dutch national distributed computing e-Infrastructure) and the expertise of Natalie Danezi, grid & cloud services advisor.

"For assessing the reliability of longer dykes, typically 1000 computing jobs are needed to ensure a converged solution. As each job is able to be run independently on a single serial computer, High-Throughput Compute is ideally suitable for such a task and can reduce the computational time dramatically", explains Li.

For his research calculations, Li used **500,000** core hours of compute time, **500 GB** of disk storage and **500 GB** of tape available via the projects.nl virtual organisation, supported by SURFsara.

One of the main results of Li's work is that the longer the dyke, the larger the failure probability.

He recommends that safety standards are made according to different scales – national or regional. The results of his work are detailed in his PhD thesis, published this February 2017.

OUTPUT
Li, Y. 2017. Doctoral Thesis, Delft University

500 thousand core hours

500 gigabytes disk storage

"High-Throughput Compute is ideally suitable for such a task and can reduce the computational time dramatically."

Providers:
This research was supported by SURFsara, the Dutch national distributed computing e-Infrastructure, with resources made available via the projects.nl virtual organisation.

How to predict water conditions along the **Iberian coasts**

The computional challenges of the OPENCoastS forecast mission

Seas and oceans are important drivers for the European economy and they need to be preserved and developed in a sustainable way.

In support of this view, the project **OPENCoastS,** or On-demand Operational Coastal Circulation Forecast Service aims at boosting the progress of the European marine sector by forecasting water levels, 2D velocities and wave parameters along the North Atlantic coast.

The service was developed by the National Laboratory for Civil Engineering (LNEC) in 2010 as WIFF (Water Information Forecast Framework).

Since then, the system has been producing 48-hour forecasts on a daily basis for the Portuguese coast and is running on High-Throughput Compute and storage resources provided by the NCG-INGRID-PT data centre, which is part of the **Portuguese National Distributed Computing Infrastructure (INCD)** and the EGI federation.

LNEC provides this service to Portuguese-based researchers free of charge, but LNEC researchers Anabela Oliveira and Alberto Azevedo are working on making OPENCoastS a service available at an international scale.

The INCD computational and storage capacity needed to run the water forecast service every day uses three typical configurations:

- Reference run: 60h CPU + 3GB storage / 72h period daily forecast
- Large run : 600h CPU + 30GB storage / 72h period daily forecast
- Very large reference run: 1200h CPU + 60GB storage / 72h period daily forecast

These values can amount to up to **2,500 cores and 2 terabytes storage per year** for a total of 100 site deployments. In the future, the platform could use even more computing resources to facilitate the access to circulation forecasts to biologists and geologists, who have strong needs in understanding the impact of water dynamics and ecology.

2,500 cores per year

2 terabytes storage per year

"The development of forecast systems requires not only a strong knowledge of coastal processes but also of information technology, along with access to significant computational and storage resources."

Providers:
INCD is the Portuguese national distributed computing e-infrastructure

H.E.S.S.: a window to the high energy universe

High-Throughput Compute services to unlock the puzzle of cosmic rays

Earth is bombarded every day with high-energy cosmic rays – a type of small, electrically charged particles first discovered in 1912. The H.E.S.S. experiment (High Energy Stereoscopic System) is an array of five telescopes built in the Gamsberg Mountains of Namibia to identify cosmic ray sources and investigate how these tiny particles accelerate and travel through space.

The H.E.S.S. experiment started in 2003 and, so far, the team has found more than 70 sources of cosmic rays, most of them pulsar wind nebulae and supernova remnants. They also discovered a Peta-electronVolt (PeV) particle accelerator in the Milky Way – a crucial piece of the puzzle of high-energy cosmic rays.

The collaboration is an effort of about 200 scientists, from 43 scientific institutions across 14 countries.

Together they have published more than 100 papers and they won, in 2006, the Descartes Prize of the European Commission – the highest recognition for collaborative research.

Besides cosmic rays, "H.E.S.S. is also participating in the quest for dark matter, and also tries to answer fundamental questions, such as whether the speed of light is the same at all energies or not," adds Mathieu de Naurois, spokesperson of the consortium.

Computational challenges

The data collected by the telescopes in Namibia is stored in two computing centres, the CCIN2P3 in France, and the Max Planck Institut für Kernphysik in Heidelberg. At the CCIN2P3 alone, H.E.S.S. has about 2.1 petabytes of data stored in tape.

This data by itself means very little and the H.E.S.S. relies on Monte Carlo simulations to extract meaningful information about the properties of cosmic rays. They use two chains of ongoing simulations and one of them has been running on **EGI High-Throughput Compute resources** since 2012.

The consortium also uses **EGI storage services** to share calibrated telescope data between users who do not have access to the CCIN2P3 and the Max Planck computing centres.

The distributed computing operations of the H.E.S.S. experiment are managed by Jean-Philippe Lenain, an astrophysicist based at the Laboratory of Nuclear and high-energy physics in Paris.

"EGI is a great framework to perform our simulations on a larger scale**,** better than what could be achievable in local computing centres," says Lenain. "It also allows us to transparently share our data among the collaboration."

Website: *www.mpi-hd.mpg.de/hfm/HESS/*

200 scientists, from 43 institutions across 14 countries

H.E.S.S.'s EGI services

High-Throughput Compute

H.E.S.S. usage
(2016-2017)

38 million

core hours

2,000 jobs

at any given time

"EGI is a great framework to perform our simulations on a larger scale, better than what could be achievable in local computing centres. It also allows us to transparently share our data among the collaboration."

Resource Providers:
The data centres providing the most computing and storage resources to H.E.S.S. are:

GRIF,
OBSPM
M3PEC Bordeaux
LAPP
CCIN2P3
DESY-Zeuthen
CAMK
SWEGrid

Only two pigments form the colours of bird eggs

Birds' eggs are famously colourful and display varied and striking patterns, capturing the attention of many scientists and artists since time immemorial. But how diverse are avian eggshell colours after all?

Daniel Hanley, Long Island University, and his colleagues looked into the colour palettes of birds' eggs to investigate if the variety we see can be created by combinations of blue-green and brown pigments that exist in birds' eggshells.

They started by measuring the colour of eggs from more than 600 different types of birds found all over the world. Then, the team deployed a model to generate the colours that would be produced by mixing variable amounts of blue-green and brown pigments. They used the computing resources provided by **Metacentrum** to generate predictions of all possible combinations of pigments.

Their conclusions, published in Biology Letters, confirm that the egg diversity we see is due to combinations of blue-green and brown pigments.

Credit: D.Hanley / L.Vaicenbacher

OUTPUT
Hanley et al. doi: 10.1098/rsbl.2015.0087

Providers:
MetaCentrum is the Czech national distributed computing e-Infrastructure and represents the Czech Republic in the EGI Council.

When did **whales and dolphins** become adapted to life underwater?

Whales and dolphins make up the cetaceans, a special group of mammals adapted to life underwater. Their transition from dry land to an aquatic environment is one of the greatest examples of evolutionary adaptation.

Georgia Tsagkogeorga and her colleagues at Queen Mary University of London were interested in how cetaceans evolved in comparison with their closest relatives. In particular, they wanted to know if the molecular adaptations related to aquatic lifestyle appeared before or after whales and dolphins split from the hippos.

Tsagkogeorga used **GridPP's** computing resources to run large-scale molecular evolution analyses. In total, they performed about **110,000 computations with High-Throughput Compute**. "These analyses would have taken months to complete without access to parallel computing," says Tsagkogeorga.

The conclusions, published in Royal Society Open Science, show that the most significant molecular adaptations to aquatic life appeared in cetaceans after the split with the hippos, 55 million years ago.

OUTPUT
Tsagkogeorga et al. 2015 doi: 10.1098/rsos.150156

Providers:
The QMUL GridPP cluster is one of EGI's federated data centres. GridPP is a community of particle physicists and computer scientists based in the United Kingdom and at CERN.

New biomarkers for multiple sclerosis

Multiple sclerosis, also known by its acronym MS, is an inflammatory brain disease that causes progressive physical and cognitive disabilities over many years. MS kicks in early and remains as the number one neurological problem affecting young adults.

Tobias Granberg, from the Karolinska Institutet in Sweden, used the **Virtual Imaging Platform (VIP)** to analyse the results of a long-term study of MS effects in the corpus callosum – an area of the brain highly sensitive to the disease. VIP is science gateway designed in France to provide access to grid computing and storage resources for medical imaging simulation.

Granberg used VIP to run brain tissue segmentations in MS patients and controls. In total there were 83 examinations to segment, and VIP greatly helped in making the volumetric analysis quick and feasible.

The results show that the volume of the corpus callosum is feasible as a quantitative biomarker for cognitive and physical disability in MS research and in clinical practice.

OUTPUT
T. Granberg et al. 2014. Multiple Sclerosis Journal. doi: 10.1177/1352458514560928

Providers:
VIP platform is supported by the biomed VO, with resources from 60 federated data centres of 11 EGI Council members.

The **genetics and diversity** of mussels

Atlantic mussels are very unusual from a genetic point of view: they have mitochondrial DNA from both their fathers and their mothers. Every other organism, from birds to humans, inherits mitochondrial DNA only from the mother's side.

Artur Burzyński, a molecular biologist based at the Institute of Oceanology of the Polish Academy of Sciences, and his colleagues wanted to know why mussels are so diverse.
The team looked into how the last glacial period affected mussels and collected 985 mitochondrial DNA samples to look for polymorphisms, which are distinct features occurring naturally in a species.

Burzyński and the team used parallel computing resources provided by **PL-Grid** (the Polish national e-Infrastructure) for their analysis. They ran hundreds of serial jobs, each taking a few days. "We would still be running those comparisons if only a single computer could have been used," he says.

The team found that all modern mussels coalesced into one single group around 10,000 years ago – at the end of the last glaciation period. And after that, this population underwent rapid expansion genetic diversification. The results were published in the journal Heredity.

OUTPUT
Śmietanka et al. 2014 doi:10.1038/hdy.2014.23

Providers:
PL-Grid is the Polish national e-Infrastructure

Colofon

This publication was prepared by the EGI Foundation Communications Team (Sara Coelho and Iulia Popescu)

We would like to thank all the scientists involved in the preparation of this publication for their support and cooperation.

Throughout this publication we use “core hours” to refer to the logical CPU elapsed time normalised by multiplying the time by logical CPU computing power expressed in HEPSPEC06. (http://w3.hepix.org/benchmarks/doku.php)

For more information about usage, check the **EGI Accounting Portal.** (accounting.egi.eu)

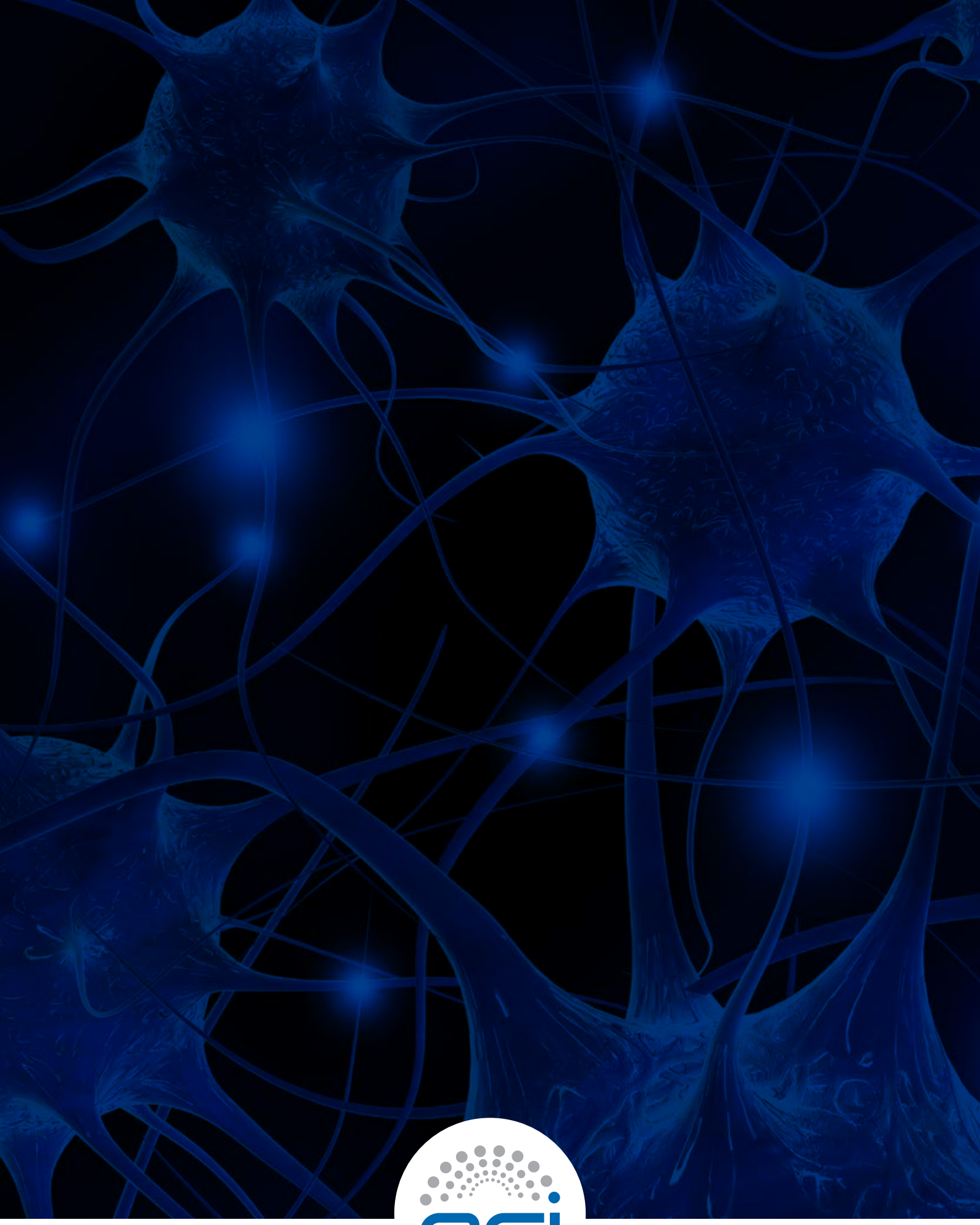
The content of this publication is correct as of June 2017.

Design: Tom Jansen

The EGI-Engage project is co-funded by the European Union (EU) Horizon 2020 program under grant number 654142. The EGI-Engage project (Engaging the Research Community towards an Open Science Commons) started in March 2015, co-funded by the European Commission for 30 months, as a collaborative effort involving more than 70 institutions in over 30 countries.

In EGI's vision, researchers from all disciplines have easy, integrated and open access to the advanced computing capabilities, resources and expertise they need to collaborate in data/compute-intensive challenges.

www.egi.eu



EGI Foundation • Science Park 140 • 1098 XG Amsterdam • The Netherlands
+31 (0)20 89 32 007 • egi.eu