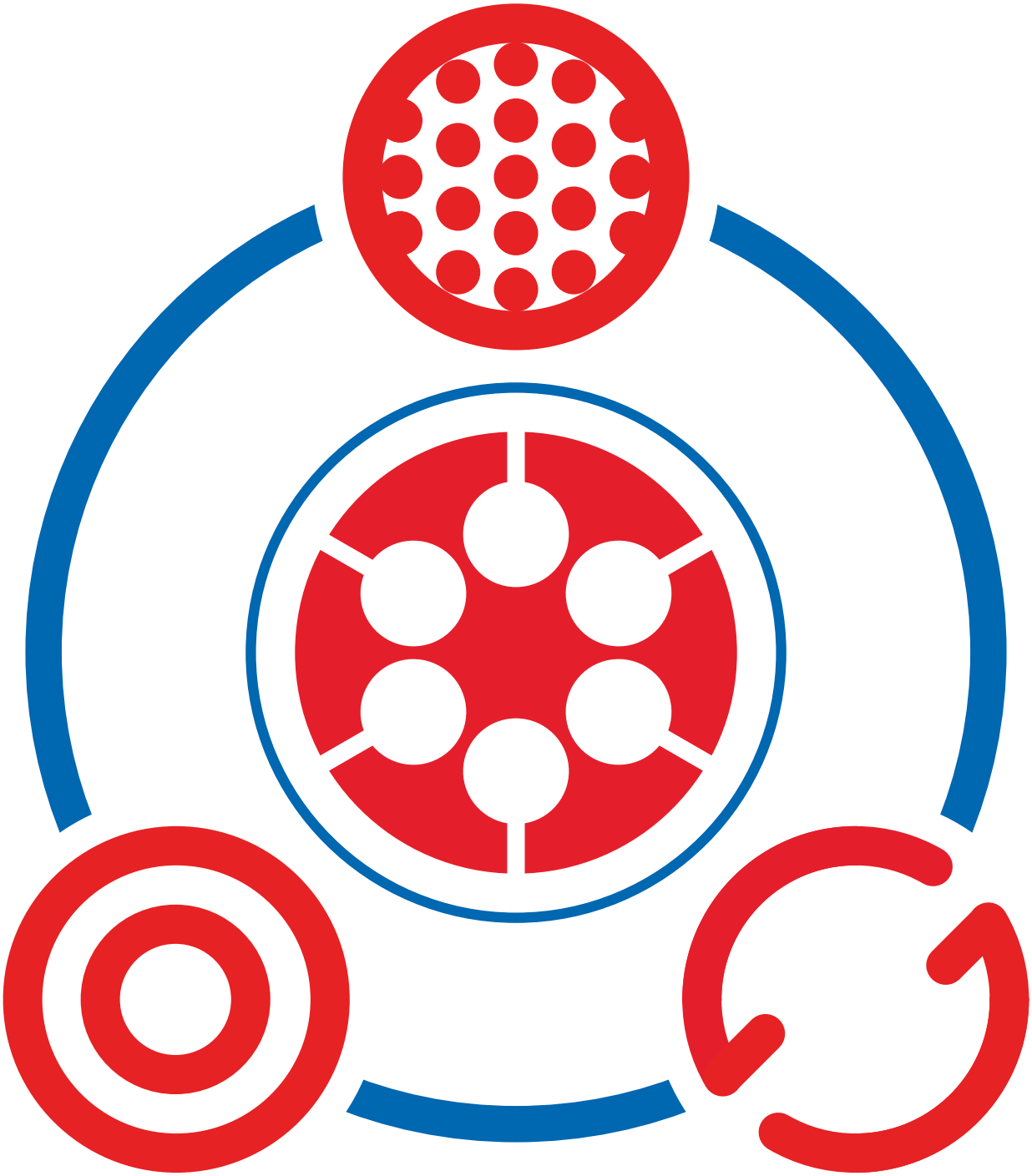
EUROPEAN GRID INFRASTRUCTURE

EGI SOLUTIONS



- HIGH-THROUGHPUT -

DATA ANALYSIS

WWW.EGI.EU

EGI receives funding from EGI-InSPIRE, a project co-funded by the Seventh Framework Programme of the European Union for research, technological development and demonstration under grant agreement no RI-261323

Table of Contents

Introduction

The "High-throughput Data Analysis" solution is aimed to help individual researchers, and research communities that have large scale data management and computational capacity requirements. The challenges that they typically face are limitations in available computational resources, the technical and administrative problems of sharing their data sets, and a need to access different facilities in different locations.

With this solution users gain access to a vast amount of distributed resources, which are allocated via a central process and made accessible with uniform interfaces. EGI provides a single entry point to a federated pool of resources that can be allocated to new or existing user communities who need resources to perform their research activity. User communities are enabled to perform their investigations in an effective, cost-efficient collaborative way, which otherwise would not have been possible.

The solution is built with a combination of services already provided by the EGI.eu organisation, such as Operations, Technology and Security Coordination, and Technical Consultancy and Support, but also using the resources provided by the federation that comprises the European Grid Infrastructure.

This solution is one of the ways in which EGI attends to the needs of the researchers and research communities not only within the European Research area but also worldwide.

1. Target Groups And Specific Challenges

1.1. Target Groups

The "High-throughput Data Analysis" solution targets groups of researchers who work on the same topic, share data or use the same application software. The groups can be very small, such as individual researchers, or very large international collaborations. The solution is aimed at research activity that analyses or produces large datasets through the execution computational tasks, which can be either single or parallel processes. It is also for communities that need to access distributed resources or datasets in a collaborative way.

1.2. Specific Challenges

The challenges can be described as follows:

- Users do not have access to enough resources within their institution.
- The user community has resources distributed in different resource centres, and they need to have uniform access to them.
- The access to the distributed resources and datasets must facilitate collaboration among the researchers.

If these challenges are not properly met, the researchers will not be able to produce the results in the expected timeline.

2. EGI Solution

2.1 .Objectives

EGI offers the "High-throughput Data Analysis" solution to enable users to:

- Access transparently distributed resources with uniform interfaces.
- Be authenticated in an uniform way in different sites.
- Autonomously manage their communities structure and to regulate access to services and data throughout the infrastructure.
- Access resources assigned through a central allocation process.

Users are able to access distributed resources through common standard interfaces uniformly available in the different resource centres. The user can manage his data and execute and control the computational tasks using common services and APIs independently from the Local Resource Management Services chosen by the administrators of the different sites.

Users' identity is uniformly recognised in the whole infrastructure. This reduces workload, for example, by making it possible for a computational task running in one resource centre to access data stored by the user in another centre.

2.2 Accessing the Solution

Users can access resources through common interfaces and authentication methods.

For uniform authentication all the EGI services users need to have grid credentials, which are x509 certificates . Users usually keep grid certificates password protected on their User Interface (UI). Users without a certificate can get one from their local Certification Authority (CA), or the generic EGI CA. More information is available on the EGI website.

Users usually access the computing and storage services from a user interface

(UI), which is a machine (virtual or physical) through which they access the infrastructure. A UI installation contains a client for most of the services deployed in EGI, this client allow users to submit computing tasks and retrieve results as well as store and manage data. EGI provides the necessary virtual images and installation tools to deploy a user interfaces on a wide range of hardware.

Authorisation of access to resources is normally regulated by Virtual Organization (VO) membership. Users who are not members of any VO can join an existing VO that supports the research topics of the user, or create a new one. For individual users, or small groups joining an existing VOs reduces their administrative overhead.

Users who have federated their resources in EGI can, with the credentials and the VO, start accessing their distributed resources.

Users who do not own resources can contact the NGI or NGIs individually to ask for opportunistic usage of the available resources, in practice this will result in enabling the VO in the services operated by the NGI. A second option is to ask for resources directly from EGI . These resources reserved and allocated centrally through a call open to all users.

2.3 Building the Solution

The "High-throughput Data Analysis" solution builds upon the middleware services that enable access to distributed resources.

For user authentication EGI services enable grid certificates issued by the Certification Authorities (CA) that are members of the EUGridPMA organisation. EGI collaborates with EUGridPMA to deploy and test the usage of CA certificates in their production services, to ensure uniform user authentication.

EGI in collaboration with the NGIs deploys a number of Virtual Organisations Management Services (VOMS) that allow VOs to register and manage their members.

Several Technology Providers (TP) provide all the services described in this section. Each TP releases their software independently and through different

distribution channels. EGI collects the most relevant components and release them in dedicated repositories as part of the UMD distribution. UMD releases are tested at several levels including a pre-production deployment. Sites and users have the added value of having all the components that build the infrastructure accessible in a single repository and with an additional quality assurance process on top of what is performed by a TP.

The services provided by the Technology Providers are:

- Operations Coordination is a set of management and coordinating activities ensuring that operational activities across the federated infrastructure work seamlessly, without fragmentation. The coordination binds the infrastructure so that the services are delivered at an agreed service level.

- Technology Coordination ensures continuous technological innovation through sourcing of software components from diverse technology providers to meet the current and emerging needs of both researchers and Resource Centres.

- Security Coordination ensures a secure and stable infrastructure to mitigate threats, enhance services, and give users the protection and confidence they demand from a service.

- Technology Coordination ensures continuous technological innovation through sourcing of software components from diverse technology providers to meet the current and emerging needs of both researchers and Resource Centres.

- Technical consultancy and support offers tailored technical and management advice to help partners and clients make the most out of e-Infrastructure technologies.

3. Value Proposition

Users can access in a uniform way distributed resources provided by different resource providers. The research workflows do not need to be adapted to the single providers, but use standard interfaces. The resources are easily shared among user communities maximising usage efficiency.

PROBLEM	PROVIDED SOLUTION	ADDED VALUE
Users do not have access to enough resources within their institution.	Central resource allocation service. A platform based on standards, common interfaces and protocols	- Transparent access to distributed computational infrastructure beyond the local capacity restrains - Flexible use of data storage and computation across disciplines and borders - Easier management of access to services and data throughout the whole infrastructure - Access to resources assigned through a central allocation process - More efficient use of available resources, both computational and human - Time and effort saving, more efficient research process - Improved user experience
The user community has resources distributed in different resource centres, and they need to have uniform access to them.	Centrally-provided expertise and streamlined best practices on how to set up and manage federations. Common Core Infrastructure to build the federation on.	
The access to the distributed resources and datasets must facilitate the collaboration among the researchers.	Tools for Virtual Organisation management. Uniform authentication and authorization mechanism among different resource providers. Federated service management best practices, cost-effective sharing of services (support, processes, policies, activities), community expertise & re-use of tools/output from public funded projects	

4. Success Stories

High-throughput data analysis services have already being used by the scientific community including the Large Hadron Collider, civil disaster mitigation and mathematics.

The Large Hadron Collider

The high energy physics community working on the Large Hadron Collider (LHC) produced one of the brightest scientific results of 2013, announcing the experimental evidence of the Higgs Boson. The community includes thousands of researchers distributed across the world, who need access to hundreds of resource centres in Europe and beyond. The LHC ran more than four hundred million jobs that consumed more than one and a half billions of CPU/hours during 2013. These huge numbers were made possible by the highly integrated and distributed high throughput infrastructure provided by EGI.

Civil disaster mitigation

The application of the services provided by EGI and facilitated by this solution allowed applying a 3-D model system to study the effects of a toxic spill in a real tailing dam. The research involved complex calculations based on 15,000 different sets of input parameter values, which was at their time considered very demanding in terms of computation resources. The results were considered extremely satisfactory by the researcher and other stakeholders.

Mathematics

A very computational demanding study was designed to investigate the Goldbach conjecture, which still remains as one of the biggest challenge among mathematicians. This project submitted 173,816 jobs to the European Grid Infrastructure adding up to 869,080 tasks, consuming 1 CPU-hour each, which took 7 months to process. The same task in a normal laptop would have needed about 99 years.

5. Conclusion

The “High-throughput Data Analysis” solution is a key part of EGI’s solution portfolio. It is aimed specifically at helping researchers to manage their data and improve access to computing resources.

With this solution researchers and research communities gain seamless, transparent access to greater computational capacity. This allows them to concentrate on their own research, obtaining scientific results securely and quickly. They are also enabled to work collaboratively with other groups located remotely, across countries, and even continents.