D7.4: Final Report EUDAT/EGI

Author(s)	Michaela Barth (SNIC)
Status	Draft/Review/Approval/Final
Version	v1.1
Date	31/12/2017

Abstract:

This document describes the work undertaken by Task 7.2 "Joint Access to Data, HTC and Cloud Computing Resources" within the EUDAT2020 project focusing on the interoperability between EUDAT and EGI. The main objective of the conducted work was to realize cross-infrastructure services that enable the desired perceived seamless access to a combined infrastructure offering by both EGI and EUDAT services in pairing their data and high-throughput computing resources together.

EUDAT receives funding from the European Union's Horizon 2020 programme - DG CONNECT e-Infrastructures. Contract No. 654065

Document identifier: EUDAT2020-DEL-WP7-D7.4	
Deliverable lead	SNIC
Related work package	WP7
Author(s)	Michaela Barth (SNIC)
Contributor(s)	Ute Karstens (ICOS), Margareta Hellström (ICOS), Matthew Viljoen (EGI). Peter Gille (SNIC), Christian Pagé (CERFACS), Xavier Pivan (CERFACS), Hans van Piggelen (SURFsara)
Due date	31/12/2017 (prolonged from initial 28/02/2017)
Actual submission date	DD/MM/YYYY
Reviewed by	
Approved by	PMO
Dissemination level	PUBLIC
Website	www.eudat.eu
Call	H2020-EINFRA-2014-2
Project Number	654065
Start date of Project	01/03/2015
Duration	36 months
License	Creative Commons CC-BY 4.0
Keywords	Cross-Infrastructure Services, Joint Access to Data, HTC, Cloud Computing

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This deliverable describes the joint access to data and computing resources for three use case pilots, starting from selecting the pilots to the challenges encountered and especially focusing on the relevant recommendations and solution established during the work of EUDAT2020 task T7.2 (Joint Access to Data, HTC and Cloud Computing Resources).

The work described in this document is the continuation of the EUDAT-EGI-Pilot activity work carried out previously and already described in EUDAT2020-DEL-WP7-D7.1 and was again conducted in close collaboration with EGI. In this document the enabling of first two, then three, selected cross-utilization use cases and investigating a joint call between EGI and EUDAT for proposals for further use cases is described.

Work was carried out in an end-user driven approach where EGI and EUDAT worked closely together with EPOS, ICOS and later also ENES research infrastructure (IS-ENES) as use case pilots to test-drive the cross-infrastructure usage of the storage resources managed by EUDAT and the computing resources available through EGI and to finally also validate the results.

Throughout the project, priorities were adjusted to match the aspects most important to the user communities: the user communities put much more emphasis on automated approaches and quality of end user documentation than foreseen in the beginning. Involving the user communities in this way also meant a sometimes steep learning curve on the technological understanding and effective communication using the right terms, which presented a major - but necessary - time investment from their side. This time and trust investment had to be properly administered by not misusing them as free beta-users of any not production-ready and largely undocumented new features, that were eagerly put forward.

Concrete outcomes of the conducted work include e.g. valuable feedback on data-handling support within the EGI DataHub and testing to use the EGI Federated Cloud with automatic submission, data transfer tests between the VMs and B2STAGE instances using both OneData and EGI DataHub to access a common storage for several VMs and evaluating the new B2STAGE HTTP API.

The harmonization of access policies and especially agreeing on the authentication and authorization model to be used has to be seen as part of a wider European infrastructure effort, not only affecting EGI - EUDAT interoperability, but also here our task could again contribute with the user's perspective.

This deliverable shows the final established design of the workflows for each of the followed user communities, highlighting the adaptation to their specific needs and also the cross-fertilization between them.

1. INTRODUCTION

The EGI-EUDAT interoperability collaboration started in March 2015 with the goal to harmonise the two e-infrastructures. In order to create seamless access, and pairing data and computing resources together into one perceived infrastructure offering both EGI and EUDAT services, user communities were identified and selected to bring in their requirements on technical interoperability, authentication, authorisation and identity management, policy and operations.

The work described in this document is the continuation of the EUDAT-EGI-Pilot activity work carried out previously and already described in EUDAT2020-DEL-WP7-D7.1. This pioneer work in cross-infrastructure access concentrated on mapping several candidate research communities on their already established collaborative relations with EUDAT and EGI and their primary requirements on connecting data stored in the EUDAT Collaborative Data Infrastructure (CDI) to high throughput and cloud computing resources provided by EGI. As a major achievement a generic use case for joint access was defined and piloted as described in detail in EUDAT2020-DEL-WP7-D7.1.

Building on those efforts the work was again conducted in close collaboration with EGI and started by following and enabling two pre-selected communities and research infrastructures, ICOS and EPOS and also investigating a joint call for proposals. The main goal after the definition of the universal use case was to test-drive the cross-infrastructure usage of the storage resources managed by EUDAT and the computing resources available through EGI and to finally also validate the results. This was again done in an end-user driven approach together with EPOS, ICOS and later also ENES research infrastructure (IS-ENES) as cross-utilization use case pilots.

While the main result consists of valuable feedback to the service offerings of both of the involved infrastructures, this document will also document possible solutions for different goal settings as coming from the different needs of the accompanied user community research infrastructures.

1.1. Structure of this document

Chapter 1 gives an introduction on the purpose of this document, describes the general objectives of the work conducted as a whole and lays out the structure of the document. Chapter 2 presents the selected use cases and the selection processes used to find them. In section 2.2 it also includes the preparations and the text for the planned joint open call, as well as the decision about it. Chapter 3 is going into the details of the EGI-EUDAT joint pilot study for each use case with listing the requirements, the challenges encountered and the solutions identified. In a separate section (3.4) this chapter also includes non-community specific observations and challenges challenges valid for all use cases as well as resulting recommendations. Chapter 4 covers the activities on harmonization of access policies between EGI and EUDAT and Chapter 5 explains the work methods applied and lists some of the meetings where it was presented. Chapter 6 finally summarizes the conclusions.

2. SELECTION OF THE USE CASES

2.1. Pre-selected/Inherited use cases

As described in EUDAT2020-DEL-WP7-D7.1. pre-selected candidate pilot use cases for a joint pilot study on integrating EGI - EUDAT cross-infrastructure services were approached when the collaboration between EUDAT and EGI started in early 2015, their requirements were collected and a generic use case was defined. The user communities in this phase of the pilot activity were relevant European research infrastructures that were already collaborating with one or desirably both infrastructures and were coming from the fields of Earth Science (EPOS and ICOS), Bioinformatics (BBMRI and ELIXIR) and Space Physics (EISCAT-3D).

As a result of this first phase ICOS and EPOS were chosen as the most mature candidates to participate in the joint pilot study of the second phase and both research infrastructures were willing to act as early adopters of the e-infrastructure services and to provide their feedback. In addition, a joint open call for further early adopters was planned to collect the valuable input from the users of the coupled services and involve them in shaping them according to their needs.

2.1.1. ICOS

The Integrated Carbon Observation System (ICOS) is a pan-European research infrastructure for quantifying and understanding the greenhouse gas balance of the European continent[1]. It collects high-quality observational data relevant to the greenhouse gas budget of Europe and makes them openly and freely available at their ICOS Carbon Portal to all interested parties. The ICOS Carbon Portal can be seen as a one-stop shop for all ICOS data products (e.g. atmospheric, ecosystem and oceanic observations, emission data, meteorological diver fields, outputs of modelling activities based on ICOS observations, ...) and openly promotes the use and reuse of ICOS data for further scientific study. ICOS supports the research community in modelling activities of the greenhouse gas fluxes in time and space and enables the verification of the effectiveness of policies aiming to reduce greenhouse gas emissions.

What ICOS concretely aims to do within their EGI - EUDAT use case is to offer a new web-based service called the "footprint tool" as part of the service offering on the ICOS Carbon Portal. This tool would with atmospheric observations and further ICOS data products as input perform calculations using EGI on-demand computing facilities, to create and visualize the model output. In this case the calculations performed are 3-dimensional Stochastic Time-Inverted Lagrangian Transport (STILT) atmospheric transport model calculations and the output presents time series of climate change indicator concentrations of greenhouse gases and their resulting footprints at selected locations such as atmospheric measurement stations. Figure 1 shows a visualisation of the overall data and computational workflow for the “footprint tool” web service.

1 https://www.icos-ri.eu/

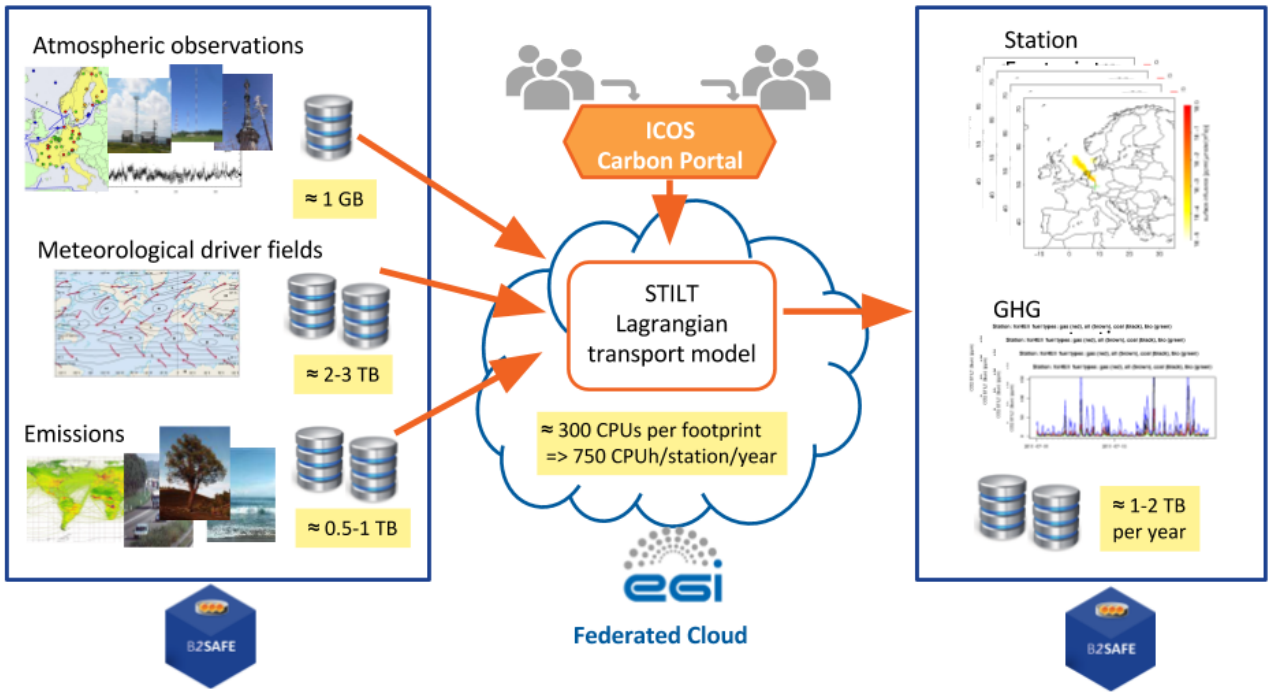

Figure 1: Footprint tool calculations using Lagrangian atmospheric transport model STILT

2.1.2. EPOS

The European Plate Observing System[2] (EPOS) the integrated solid Earth Sciences research infrastructure approved by the European Strategy Forum on Research Infrastructures (ESFRI) and is eager to take full advantage of any new e-science opportunities. Understanding how the Earth works as a geological active system including i.e. tectonic plate movements and volcanic eruptions is critically important to modern society in assessing, forecasting and mitigating the threats posed by those natural hazards. The goal of EPOS is to foster worldwide interoperability in Earth Sciences and provide services to a broad community of users. EPOS plans to achieve this by establishing a *comprehensive multidisciplinary research platform for the Earth sciences in Europe*, which pools together existing and new distributed research infrastructures for solid Earth science and facilitates the integrated use of each other's data, models and facilities allowing both for a long-term solution to do tackle solid Earth grand challenges as well as becoming an effective coordinated monitoring facility for all solid Earth dynamics on the European scale.

2.2. Investigating a joint call for proposals

One of the outlined items in the initial proposal was to have a joint open call for proposals for pilot use cases offering researchers the possibility to couple EUDAT storage capacity to EGI High Throughput Computing and Cloud Computing resources. The main goal of such a joint call for joint EGI computational resources and EUDAT storage services was to expand the activities of the pilot study to other interested research communities with a view to moving to a full production-level cross-infrastructure services that integrate storage resources managed by EUDAT and computing resources available at EGI. Already active pilot use cases were involved to give recommendations on the call and the call text.

2 https://www.epos-ip.org

2.2.1. EGI-EUDAT Call for Participation Text

EGI and EUDAT would like to invite communities interested in using both infrastructures to advance their research for participation in this call.

EGI

EGI, the European infrastructure, is providing advanced computing services for data-intensive research and gives scientists access to more than 650,000 logical CPUs, 500 PB of disk space and tape storage and 21 federated cloud providers in order to drive research and innovation in Europe. EGI provides the digital capabilities needed to advance research and innovation by providing both high throughput computing and cloud compute and storage
capabilities.

Currently supporting more than 45,000 active users, the high throughput, computing and storage services are relied upon from many diverse user communities and domains including natural sciences, physical sciences, health and agriculture. Further details of case studies can be found here:

http://www.egi.eu/case-studies/

EUDAT

EUDAT offers common data services, supporting multiple research communities as well as individuals, through a geographically distributed, resilient network of 35 European organisations. These shared services and storage resources are distributed across 15 European nations and data is stored alongside some of Europe's most powerful supercomputers.

BENEFIT FROM EGI/EUDAT INFRASTRUCTURES

EGI and EUDAT are looking for new communities interested in benefiting from both infrastructures to support them with their data/computing intensive research needs - especially communities who are facing one or several of the following challenges:

- future scaling up of data and computing resources
- bringing data and computing together
- fast access from computing to data storage
- long term preservation of data

By working with you and understanding your use case, we can help you to setup a test-bed using EGI/EUDAT infrastructures and move into production.

Please complete this <template> describing your requirements and send it to the <EGI press office> or <EUDAT press office> by <specify>.

TEMPLATE

Technical contact:
Target user community:
Summary of use case:

Technical requirements (computing, data):
Proposed start date:
Any other miscellaneous information:

2.2.2. Joint EGI-EUDAT decision on the joint open call

After physical meetings in Barcelona and then Amsterdam in September 2016 between the EGI.eu Technical Director and the head of EUDAT discussing the final scope of the pilot and the contributions of resources, it was decided to omit the joint call altogether. The argumentation was that after all the large user research communities already involved in the first phase of the Task 7.2 EGI Pilot Activity and the preselection done there, not many new user groups that were not already considered back then were expected to participate in the joint open call. The current selected user communities were already very productive in providing useful feedback. It made more sense to integrate the user communities EGI and EUDAT already were working with in other aspects into this pilot activity. So the decision was to concentrate on early adopters that ideally were already previously using both EGI and EUDAT services, but not yet in combination. One of them was the ENES research infrastructure (IS-ENES) that was already actively involved within WP8 of EUDAT.

2.2.3. ENES

The infrastructure of the climate community in Europe, the European Network for Earth System Modelling (ENES), is IS-ENES.

From the IS-ENES page

> "IS-ENES is the infrastructure project of the European Network for Earth System Modelling (ENES). IS-ENES combines expertise in Earth system modelling, in computational science, and in studies of climate change impacts. IS-ENES provides services on models and model results both to modelling groups and to the users of model results, especially the impact community. Research activities improve the efficient use of high-performance computers, model evaluation tool sets, access to model results, and prototype climate services for the impact community. Networking activities increase the cohesion of the European ESM community and advance a coherent European Network for Earth System modelling."

One of the major objectives of IS-ENES is to improve access to climate data for end users, such as the climate change impact community, but also to impact modellers and climate researchers. This is especially important in the context of the on-going significant data volume increase.

The main interest for the IS-ENES in the current EGI-EUDAT interoperability task was to enable data reduction and data processing while accessing climate data stored in the Earth System Grid Federation (ESGF) infrastructure. The ESGF Peer-to-Peer (P2P) enterprise system is an international collaboration that develops, deploys and maintains software infrastructure for the management, dissemination, and analysis of model output and observational data.

3. EGI-EUDAT JOINT PILOT STUDY

The objective of this joint pilot study was to define and implement the interfaces between the identified services in the generic use case. The early adopters were the chosen means to drive this implementation in a user-centric approach and then validate its result. By nature this was done in an iterative process of testing, providing feedback, bringing forward new requirements and testing again.

3.1. ICOS

The ICOS use case is to offer a new "footprint tool" web service on the ICOS Carbon Portal (CP) that can perform on-demand calculations based on meteorological air transport data (from ECMWF), greenhouse gas emissions (from the EDGAR inventory) and biospheric fluxes, and visualize the outcome in comparison with ICOS observational and elaborated data products. The calculations are performed on the EGI FedCloud with the data stored in B2SAFE for safe, long-term storage with fast access and B2STAGE used for the transfer of large input and output data sets between storage facilities the EGI FedCloud. In the first steps numerical simulations were run inside a self-contained Docker instance. Subsequently different in-cloud storage solution were added to make it possible to move data in and out.

3.1.1. Requirements

Load-balancing and orchestration when scaling up: Load-balancing is needed for fault tolerance and for scaling up when several VMs are required to provide the service.

'Permanent' IP: The VM hosting the web interface (2a in Figure 2) needs an IP address that is valid/referable throughout its whole existence to embed the service as part of the CP web services. As seen below, this could be solved more dynamically.

Automatization: Both EGI and EUDAT services (like B2ACCESS) need to be accessible via the command line in order to allow web applications like the footprint tool in the ICOS Carbon Portal to take care of all interactions on behalf of the end user.

Reuse of previous model run results: Since the computations with a full STILT run (including computation of particle location, footprint and concentration time series) may require several hours to days, already precomputed particle locations files should be preserved, and the on-demand calculations should only run for new station and/or new time series input data. The availability of particle location and footprint files is checked within the footprint tool.

Computationally a full STILT run needs:

- 3 GB memory per job
- ≈ 300 CPU seconds per footprint
- ≈ 700 CPU hours per station per year

Within the EGI-EUDAT collaboration further requirements and dedicated feedback from the ICOS use case towards the EGI Open Data Platform prototype, developed within the JRA2.1 activity of the EGI-ENGAGE project, could be made. Building on Onedata, an open source distributed virtual filesystem, the EGI Open Data Platform prototype is described in the EGI-Engage Deliverable D4.9 referred to in Chapter 5. The main requirement made after testing the initial prototype solution was to improve the insufficient support of writing a large number of small output files which is needed to support the reuse of calculated data requirement.

Storage Requirements:

- Input datasets (quasi-static, updated every 6-12 months)
  - Emissions (EDGAR, VPRM and more): 0.3 TB
  - Initial and boundary concentration data: 0.3 TB
  - Meteorology: 2 TB
  - Particle location files (precomputed for many stations and dates): up to 20 TB
  - Observation time series: few MB
- Output datasets (model runs access output data from previous runs and eventually add new files)
  - Aggregated Footprints: 300-400 GB
  - Concentration time series: 100 GB
  - Particle location files for new sites (produced in a full STILT run): 20 TB (assuming that users will initiate computation of additional footprints and time series for approx. 100 new sites but only for 1-2 years)

3.1.2. Challenges

- **Support for robot certificates**: In the suggested model it was agreed to use a robot certificate to interact with EGI and EUDAT so the workflow could get automated to a further extend. Initially it was not possible though to specify the service name in the Distinguished Name (DN) of the robot certificate at the default Certification Authority (CA) used. An alternative CA that could do it was identified, but found slow to respond. In the meantime discussions with GÉANT led to the added possibility to add the service name in the certificate DN also for the default CA.
- **Scaling up**: increasing the number of users and model calculations is a challenge in its own since answering a higher workload with automatic creation of VMs requires to define the minimum throughput needed that enables to fine tune for an optimal configuration. It also includes considerations on **load balancing** in general like streamlining the data access schemes and internal communication within the footprint tool between the different VMs. All work on these aspects also serves as good **preparation for** the future usage of **orchestration** tools.
- **Handling large amount of small files as output within EGI OneData**[3]: In order to enable the reuse of already previously calculated data as input for further model runs, a large amount of small files are written in the shared storage, but OneData had a problem with that and at some point just stopped working where the same setup would work nicely when just writing some large files.
  - As a temporary workaround while waiting to be able to employ the EGI Open Data Platform for both input and output data from the workflow, input data was provided via a federated storage while the output data was written to an NFS server.
  - It was hard to give an exact time scale when the issue would be solved. Several new OneData releases that were each supposed to provide a fix were approved within the EGI Change Management process, this can be due to the fact that the issue gets triggered by a combination of factors which make it hard to understand and to reproduce all of them.
  - In one version the functionality within the EGI DataHub had been split up into *aggregating and searching for data* and into *bringing your data to the VMs for computing* (with the "EGI Federated Datamanager" acting as a separate service, but with its functionality accessible through the EGI DataHub) as ICOS had requested.
  - Additionally the additional testing necessary led to delays in the upgrade schedule. However at the same time also still complete new functionality (like the easy import of existing data and coming with a graphical user interface to do this) had been added.

3 https://wiki.egi.eu/wiki/EGI_Federated_Data

- In the latest stage complete tar files of the current test data were sent to the developers to replicate the problem locally (on the very same VMs even if they wish), and - after having put in a lot of time and effort up to this stage - the user community declared themselves no longer willing to continue testing until the OneData service within the EGI DataHub is really production ready.

3.1.3. Solutions and Final Workflow

For load balancing and dynamic deployment within the EGI FedCloud different orchestrators were looked at more closely like Docker Swarm[4] and the Infrastructure Manager[5] (IM) by the INDIGO-Data Cloud, and also Kubernetes. Within the web service akka.io is used for adaptive routing across VM nodes.

Figure 2 visualises the general work and dataflow within the use case with the numbers given in the different subsections below corresponding to those in the figure. Requirements for VM size and storage space are still provisional, but it gives a clear picture of the design decisions made so far.

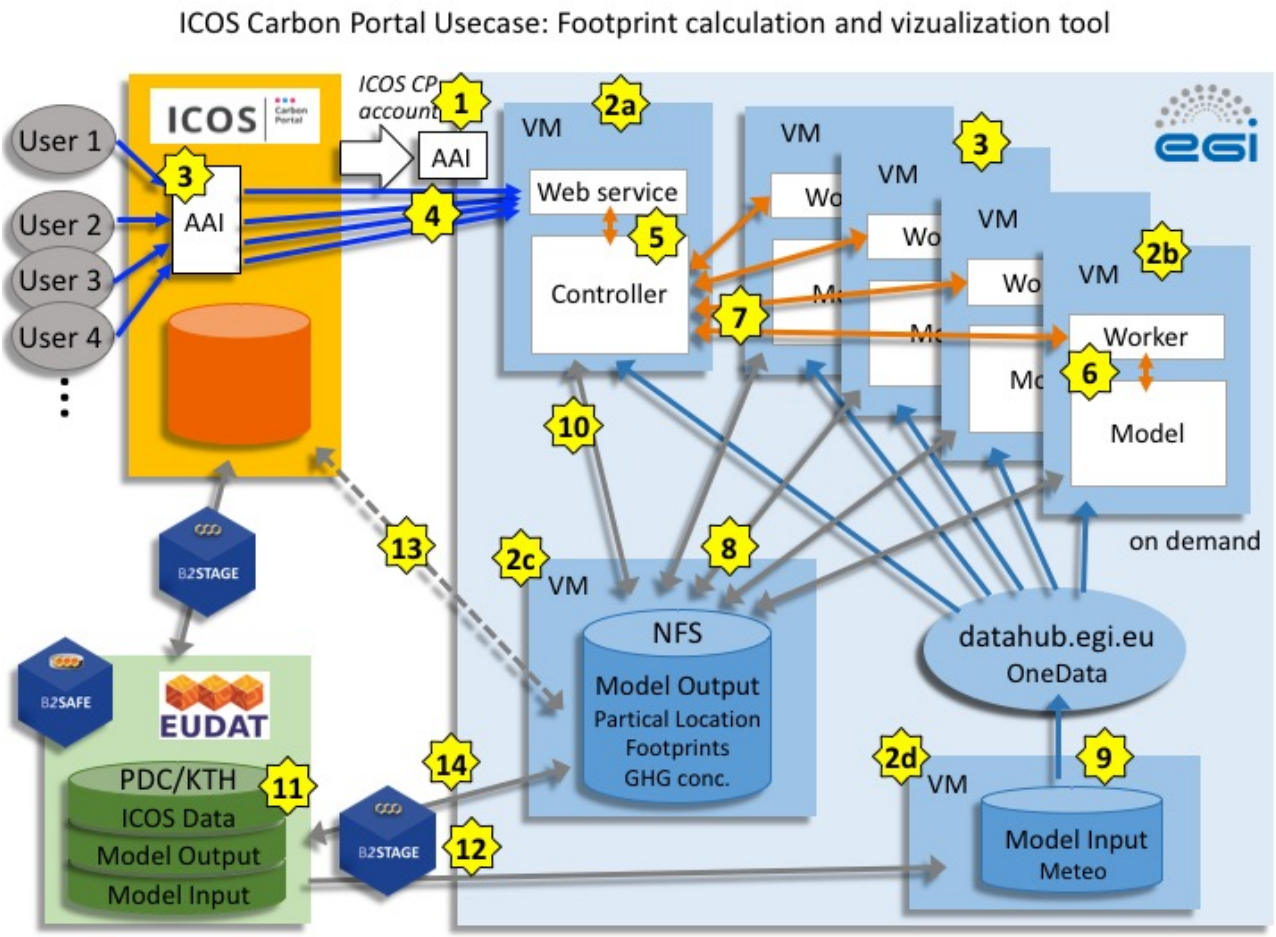

Figure 2: General work and dataflow around the footprint tool service

Workflow for setting up and providing the footprint tool service:

4 https://wiki.egi.eu/wiki/Federated_Cloud_Containers#Clusters

5 http://www.grycap.upv.es/im/index.php

1. ICOS CP instantiates several VMs in the EGI FedCloud to host the different parts of the use case. A robot certificate associated to ICOS CP is used for the authentication and authorization in the EGI FedCloud and for EUDAT B2STAGE/B2SAFE-services.

2. ICOS CP runs containerized versions of the services on these VMs:

- VM hosting the **web interface** for request of model runs and visualization of the results and the **controller** for load balancing and container orchestration (based on Akka Cluster akka.io). Prototype: 8 CPUs, 16 GB RAM
- VM hosting the **worker** for starting model runs and sending back log files and the **transport model** computations. (More VMs might be needed depending on the demand.) Prototype: 8 CPUs, 32 GB RAM (+ 2 TB local storage)
- VM with large block storage attached with **NFS** to store model output data produced on one or more VMs. Prototype: 4 CPUs, 16 GB RAM + 2 TB storage
- VM hosting OneProvider with large block storage attached to store parts of the input data for the computation (4D meteorological reanalysis). Prototype: 4 CPUs, 16 GB RAM, 2 TB storage

The VMs (2a, b, c, d) are running continuously. The ICOS CP DNS record for the service (stilt.icos-cp.eu) does not point directly at an IP address, but redirect to another domain name, which will be resolved by a dynamic DNS provider. The VM (2a) has a dynamic DNS client installed, which updates the dynamic DNS record on VM startup.

Workflow for a user of the footprint tool service:

(blue arrows)

3. The user accesses the service at the ICOS CP. Authentication and authorization of the user is handled at the ICOS CP with HTTPS cookie-based authentication. The ICOS CP takes care of all interactions with the EGI FedCloud on behalf of the user.

4. The user is directed to the web service running on VM (2a).

Information flow inside the footprint tool:

(orange arrows)

5. The web service provides the user interface to initiate model runs and visualize results. Input parameters (geographic coordinates and time span) are selected on the web interface and passed to the controller. The controller distributes the jobs to the required number of cores/VMs. Input parameter are passed to the worker node(s) on the VM(s) for the model runs. The controller also initiates the creation of additional VMs if required.

The web service serves several users in parallel and the controller launches separate specific model runs.

6. Model computations are hosted in a separate VM. As the model runs only on a single CPU, parallel model runs are possible. Each model run is further split into small jobs to allow easy distribution over many cores/VMs to efficiently serve multiple users.

7. Log files from the model run are transferred back to the controller and displayed on a dashboard in the user interface.

Data flow inside the footprint tool:

(grey arrows)

8. Model output is written to a common Network File System. The model output data might be re-used in further model runs if the same station and (partly) the same time slots are requested. Therefore output data from all previous runs should be available to the model run. Model output (particle location files, footprints, concentration time series) consists of a large number (~1 Mio) of small files (1-5 MB).

9. Model input (4D meteorological reanalysis, emissions maps) files are stored on block storage attached to a VM hosting OneProvider and access is managed through the EGI DataHub[6]. The access of data from multiple providers might be a useful application.

10. Model output is displayed on the web interface and the user can also download the results to her/his local computer.

Data storage and transfer:

(light blue arrows)

11. Quasi-static input data (incl. metadata) are stored on the ICOS CP server and for long-term storage in B2SAFE (at PDC/KTH), data transfer using B2STAGE.

12. A copy of the input data is transferred using B2STAGE from B2SAFE to the storage attached to the VM hosting OneProvider and to the storage attached to the VM hosting the NFS. Regular updates are needed when new datasets become available.

13. Datasets not stored in B2SAFE (eg. intermediate versions of model output) are directly transferred from the ICOS CP server the VM hosting the NFS.

14. Model output data (incl. in the future also metadata) are transferred regularly to B2SAFE using B2STAGE and archived in B2SAFE for long-term storage.

Note: The strategy to attach a PID (or DOI) to the model output (and user-specified request) is not yet included. ICOS is setting up its proper facility for dataset publication associating PID/DOI. Integration with EGI/EUDAT cataloging systems has to be further discussed.

In parallel also the internal workflow inside the footprint tool had to be updated and improved:

Workflow inside the footprint tool:

- User selects station location (either pre-defined coordinates of existing station or latitude/longitude of new location) and time range for calculation in web service.
- Availability of footprints and particle location files for all time slots in the requested time range is checked.
- If footprints or particle location files are missing, initiate parallel STILT computations for single time slots.
- The user gets an estimate of the required computation time and receives a notification as soon as the calculations are ready.
- Concentrations for all time slots are computed based on particle location files and emissions in a final STILT run and results are combined.
- Display time series and footprints in web service.

Work performed in detail:

6 https://wiki.egi.eu/wiki/EGI_Opendata_platform

Work started with the creation of a VM (with an operating system, NGINX and Docker installed) on the EGI FedCloud. The docker container (hosting both web service and model computations at that time) running in the VM was setup and computations using data manually copied into the block storage attached to the VM could be performed. Within AAI it was agreed on the suggested model using a certificate to interact with EGI and EUDAT. An action plan was agreed on. The communication channel between EGI and EUDAT was tested according to the EGI-EUDAT integration pilot[7] based on the universal use case. For that the contextualisation script provided by EGI was used to deploy the test software appliance when configuring the ICOS VM. From within this VM data could then be successfully transferred to the B2STAGE at PDC-HPC, KTH. Storing of ICOS data has been tested on iRODS system at PDC-HPC, KTH.

In the second step, the usage of EGI OpenData Platform (ODP)[8] OneData to share data between VMs by providing access to a common storage for several VMs has been tested. Training events on how to use The EGi DataHub in general have been attended. Furthermore, input for the joint open pilot call has been provided as well as feedback on not up-to-date documentation (e.g. Docker Swarm and B2STAGE). As a step towards automation, ICOS applied for a robot certificate and finally installed it (see also Section 3.1.2). A step by step document from user perspective on how to setup the ICOS use case in all technical detail has been written.

After the completion of the first round of tests with OneData an NFS-based work-around for the small files issue had to be developed (any distributed or network filesystem like NFS or Lustre could be used to share the files, the usage of OneData is not obligatory within this use case).

Internally ICOS had to work on (1) the adaptation of STILT to split the model runs into the separate small entities, and on (2) the communications between different VMs within the footprint tool to prepare for orchestration options and improve load balancing. Also further internal improvements streamlining access to data schemes, like i.e. an alternative with having the large meteorology input files only on one machine and splitting the input data on a non-shared storage and sending the data separately to the VMs has been discussed. In the end the identified solution was to use the AKKA.io cluster deployed in the EGI FedCloud to manage the Docker container orchestration in the FedCloud allowing for the creation of VMs at increased workload to distribute computations and user requests to several VMs. Work on automating the creation of VMs at high workload is still ongoing. A ticket about a related bug within rOCCI has been submitted.

Presentations and webinars about the new B2STAGE HTTP API have been attended, and the ICOS community declared willingness to test it. Later this has been tested at the B2STAGE instance at CINECA. In order to automatize the B2STAGE transfer between the ICOS data storage in B2SAFE (at KTH) and the storage used by the VMs, first the transfer from the ICOS CP from iRODs to B2STAGE had to be organised. For these tests an additional test system at KTH was set up. The tests are not satisfactory or complete yet (see also Section 3.4). In an ongoing effort documentation for EGI and EUDAT services is being improved and checked for old or invalid links and examples.

Different tools for load balancing and orchestration (automatically creating/destroying VMs according to workload) have been investigated: 1) EC3[9] 2) Occupus[10] 3) Docker Swarm[11] 4) INDIGO-DataCloud's Infrastructure Manager (IM)[12] and 5) Kubernetes. However, no decision has been made yet.

Following the experience within the ENES use case also ICOS RI would now again be interested to test GEF again which would allow a faster and lighter creation of VMs without the need to install them and while using a lighter Docker container.

7 https://appdb.egi.eu/store/swappliance/egi.eudat.integration.pilot
8 https://wiki.egi.eu/wiki/EGI_Opendata_platform
9 http://servproject.i3m.upv.es/ec3/
10 http://occopus.lpds.sztaki.hu/
11 EGI step-by-step guide https://wiki.egi.eu/wiki/Federated_Cloud_Containers#Clusters
12 http://www.grycap.upv.es/im/index.php

Implementation of the visualization of footprints and time series at ICOS Carbon Portal:

https://data.icos-cp.eu/stilt/

Test implementation of on-demand calculation and visualization of footprints and time series - still under development:

https://stilt.icos-cp.eu/worker/

3.2. EPOS

The main objectives in the pilot activity were to identify and validate authentication and authorisation services, to test cloud resources and usage models, and to provide knowledge transfer services between e-infrastructure and EPOS communities.

3.2.1. Requirements

EPOS is currently in its implementation phase. One of their key principles is not to reinvent the wheel. Following that principle EUDAT services of interest to the EPOS research infrastructure have been identified. EPOS then decided on a gradual approach in taking up and integrating those EUDAT services, starting with using B2SAFE for long-term preservation of seismological datasets that are enriched with persistent identifiers (PIDs) and replicated onto external data facilities. From EGI the federated cloud computing resources have been identified as desirable computational resources within their portal.

3.2.2. Challenges

The two main challenges were:

1. Defining the best strategy for Web Services Parallel Grid Runtime and Developer Environment Portal (WS-PGRADE) to access the Federated Cloud and
2. Identification of secure and efficient data transfer protocols towards the iRODS system in Virtual Earthquake and seismology Research Community in Europe (VERCE).

3.2.3. Solutions and Final Workflow

Test portals used were a test instance not connected to the Fedcloud, the official production website [13] and the one used to test the integration with the Fedcloud [14].

The final workflow hasn't changed compared to the one presented in EUDAT2020-DEL-WP7-D7.1. Most effort on this use case was done by EGI within the EGI Engage project to ensure a working cloud integration (access and file transfer to the Federated Cloud) and is reported in the EGI Engage Deliverable D4.8. Relevant bugs on e.g. the file transfer issues have been correctly reported. Migration has been completed for the HPC part of the workflow, there are currently no open tasks or issues, but with EPOS' gradual approach there might still be untested parts. The workflow has been tested along with its implementation with backend support. Test parts include: job submission to the EGI FedCloud via gUse portal, data migration to IRODS v4 along with data provenance catalogue, GSIFTP. Information on how to use B2STAGE in VERCE was sent to the users. B2STAGE documentation for using the API has been reviewed by this use case as sufficient but with more examples desirable. Further usage examples have been provided as input for an updated B2STAGE API documentation.

13 http://portal.verce.eu
14 https://verce-portal-test.scai.fraunhofer.de/

On a site note we would also like to mention a similar use case using also both WSPGRADE and gUSE: the project EUROARGO/ENVRIPLUS is using VM deployment on the FedCloud via HADOOP. A successful proof of concept was demonstrated with assistance from Carlos Blanco from UNICAN.

3.3. ENES

The overproportional increase in the climate data volume as shown in Table 1 has forced the European Network for Earth System modeling (ENES) to develop a new approach to process and analyze data. The ENES use case investigates the interoperability between the EUDAT Generic Executive Framework (GEF[15]), the B2services on data storage and staging such as B2SHARE or B2STAGE, and computing resources provided by the EGI Federated Cloud. The GEF, based on Docker technology, allows to encapsulate a scientific workflow and download data selected by a researcher inside a Docker volume. After the containerized calculation ends successfully, the post-processing result can be downloaded from the GEF user interface. The ENES use case deployed the GEF workflow on EGI computing resources to perform calculations with strong computing resources extracting large data from B2 data storage services. Through this use case, a simple temporal average on Temperature At Surface (TAS) time series has been completed on typical Coupled Model Intercomparison Project phase 5/6 (CMIP5/CMIP6) input data ranging from 300MB to 500GB. Execution of the ENES use case workflow, including both data transfer and the average calculation using GEF docker rule engine, was found to be significantly faster when deployed on EGI compared to on a localhost. In a nutshell, we successfully proved that the interoperability between the three infrastructures will speed up climate analysis and represents a reliable and more sustainable solution for the future challenge.

Table 1: The expected development of the climate impact data volume hosted at ESGF: the data volumes will reflect more complex models, finer spatial resolutions and larger ensembles

Format	Total Size	# of data sets	in # files	Notes
CMIP5	1.8 PB	59000	4.3 EE06	in 23 ESGF data nodes (about 50 times the volume for CMIP3)
CMIP6	90 PB	No est. yet	215 EE06	(This is an extrapolation)

The first challenge for ENES use case was to establish a close collaboration between the GEF team from EUDAT WP5 and the climate community from EUDAT WP8 to set up a road map and identify necessary changes from both sides to deploy the GEF backend on EGI. Since the GEF is based on Docker, the best solution to fit the GEF-EGI interoperability consisted of building a Docker image able to instantiate EGI computing resources with EGI appDb input. After initially having had instantiated VMs with rOCCI, the best and fastest solution was found in the java based jOCCI API. From this API we built and dockerized a java app to integrate the VM instantiation capacity to the GEF UI as an internal GEF service repository. Figure 3 represents the GEF service repository[16] whose purpose is to generate VM machines on EGI and bind the GEF and the new EGI VM together. The workflow works as follows: 1) The authenticated researcher selects an EGI endpoint and a resource template (CPU, RAM) from the EGI appDB and pastes the URL in the preconfigured JSON file that keeps track of timestamp and general referrals for the VMs and GEF. 2) The user starts the Docker service to instantiate the VM. The operation is complete once the EGI VM is active and ready for computation. 3) Generation of TLS certificate based on the new VM IP as shown in Figure 4. The TLS client certificate are stored on GEF backend while we copy the TLS server certificate onto the EGI VM to reboot the Docker daemon. At the end of this workflow we are able to perform encapsulated calculations using the EGI resource as long as TLS client certificate is valid.

15 https://github.com/EUDAT-GEF/GEF

16 https://github.com/EUDAT-GEF/GEF/tree/master/services/_internal/maven-EGI

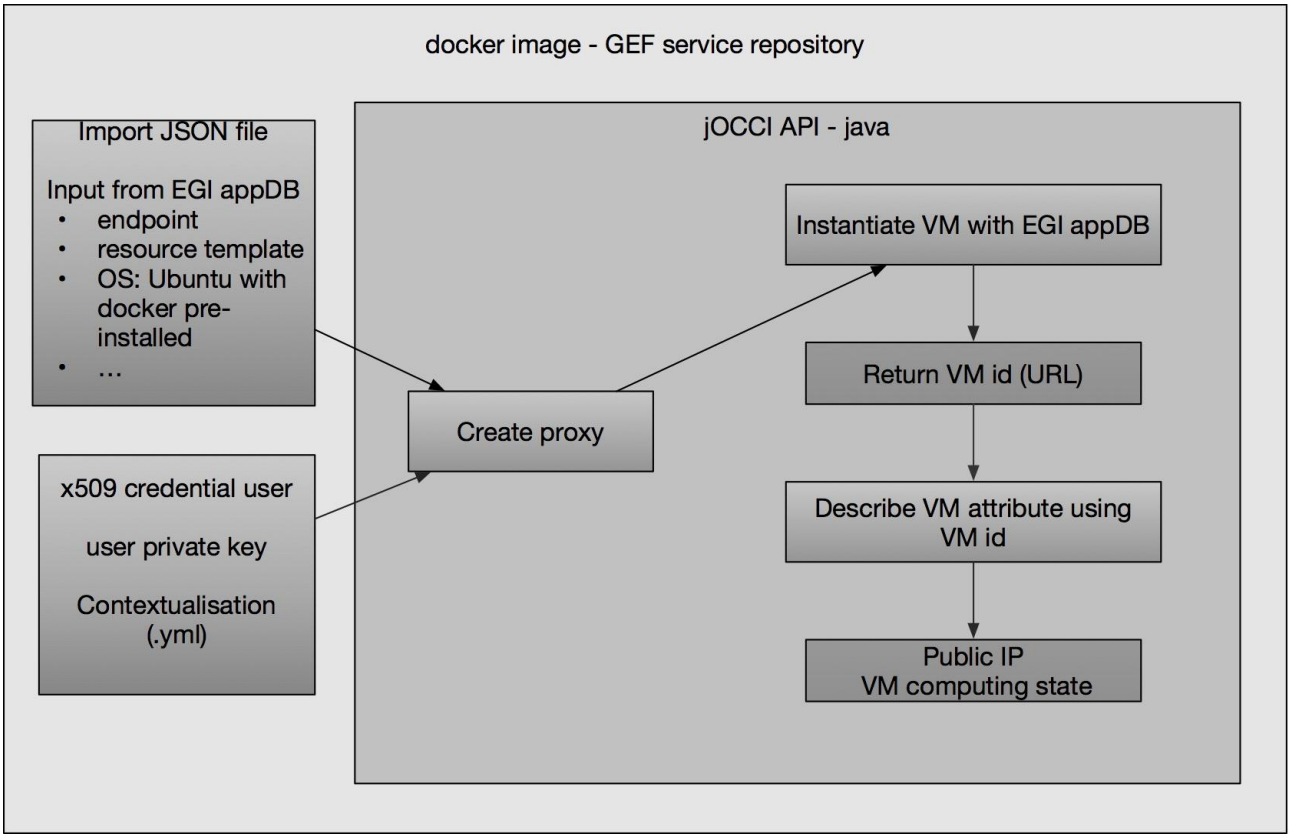

Figure 3: GEF internal docker service repository to instantiate computing resources on EGI

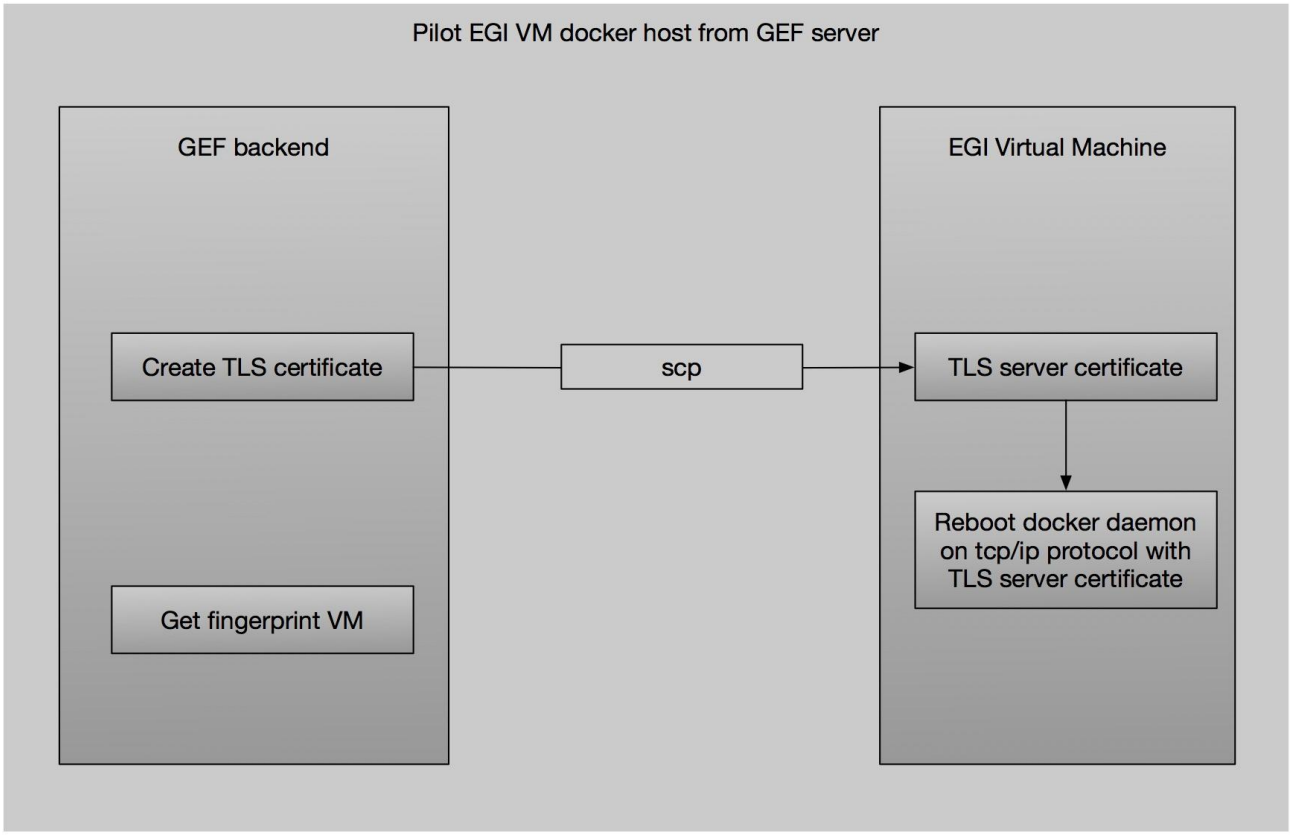

Figure 4: Step to safely pilot the EGI VM docker daemon where GEF is installed

3.3.1. Requirements

- **Fast bandwidth between data storage and compute facilities**. This is needed in order to perform efficient data analysis with a large number of realizations (ensemble of scenarios), process higher spatial and temporal resolutions and allow for easy sharing of intermediate results with collaborators.
- **Support a more flexible and robust data life cycle.** All stages within the research data lifecycle (creation, processing, analysing, preservation, authorization, re-usage) should be covered to allow for reproducible experiments and more robust setups so several experiment configurations can be explored to answer scientific questions.
- **Automatic instantiation of VMs.** The the climate impact researcher as final end user is not supposed to deal with the details of the VM setup. An automated workflow is expected to run more efficiently.
- For the prototype, the storage requirement for the test data is only at around **500 GB** of storage for each of B2SAFE and B2STAGE.

3.3.2. Challenges

- The connection to three different infrastructures (ESGF, EUDAT and EGI) needs special emphasize and care in the AAI setup, this needs to be revisited compared to the current prototype. Especially in the beginning to get a working setup with (partially still self-signed) certificates and working proxy generation was a challenge.
- The VMops dashboard as a way of managing VMs on the FedCloud is not relevant for this use case, since the VMs have to be instantiated automatically from GEF, indicating that the jOCCI method is to prefer.
- Integration with B2 services is still up to improval: The data is currently accessed via EGI resources, direct download from EUDAT is still missing. Data in B2STAGE is only updated manually (via gridFTP) to provide the input for the workflow in the prototype phase.
- Security design was not the main focus at this stage, secure proxy generation with JOCCI API (e.g. via myproxy) should be reviewed in the next stage when going from the prototype into a production service.
- In parallel also a prototype using the ESGF CWT API processing service has been co-developed as a calculation delegation. This part was initiated by the GEF and is covered in WP8. Work within this task had to by synchronized with that development.

3.3.3. Solutions and Final Workflow

Simplified view of the steps of the current adapted use case as also depicted in Figure 5:

1. VM instantiation on EGI e-infrastructure, see Figure 4. GEF docker endpoint run on EGI VM IP.
2. Researcher configures a Dockerfile to encapsulate its calculation related to the selected data.
3. The Dockerfile is uploaded, built on the VM docker repository and available to run as a service from the GEF UI. Researcher's calculation is ready to be containerized on EGI.
4. Researcher submits the service, providing the data URL and any additional data as input. The container downloads the data and performs the analysis.
5. Workflow ends and the researcher can download the resulting post-processing from the GEF UI.

The goal of the use case was to develop a prototype that will be connected to the ESGF production services later. In the current prototype stage the automation of all steps is completed. Input data comes from the ENES/ESGF data nodes, but eventually from B2SHARE. It uses the dockerized jOCCI API for automatic instantiation of VMs. During the work in the use case support tickets have been successfully filed to stabilize the interaction between EGI FedCloud[17] and the jOCCI API from a Docker. ENES got access to B2STAGE and B2SAFE services at STFC and on CINECA. Several different adaptations of the use case have been discussed. After a further planned change of the GEF API backend, a new section can be added to the frontend designed within WP5 and WP8 giving the user access to more options. The design for a concrete implementation at CINECA is shown in Figure 6.

In the final production version the raw (CMIP5/CMIP6) data will made available and processed through the ESGF Computing API and post-processing results (like temporal averages on TAS time series) will be sent back to be displayed and further processed at the IS-ENES climate4impact.eu platform. There they can also be downloaded in different and more common data formats as tailored products via a simple website interface.

Documentation and tutorials on how to use the GEF backend to create the VMs in the cloud faster are available on the GEF github[18] so other user communities can also test this method.

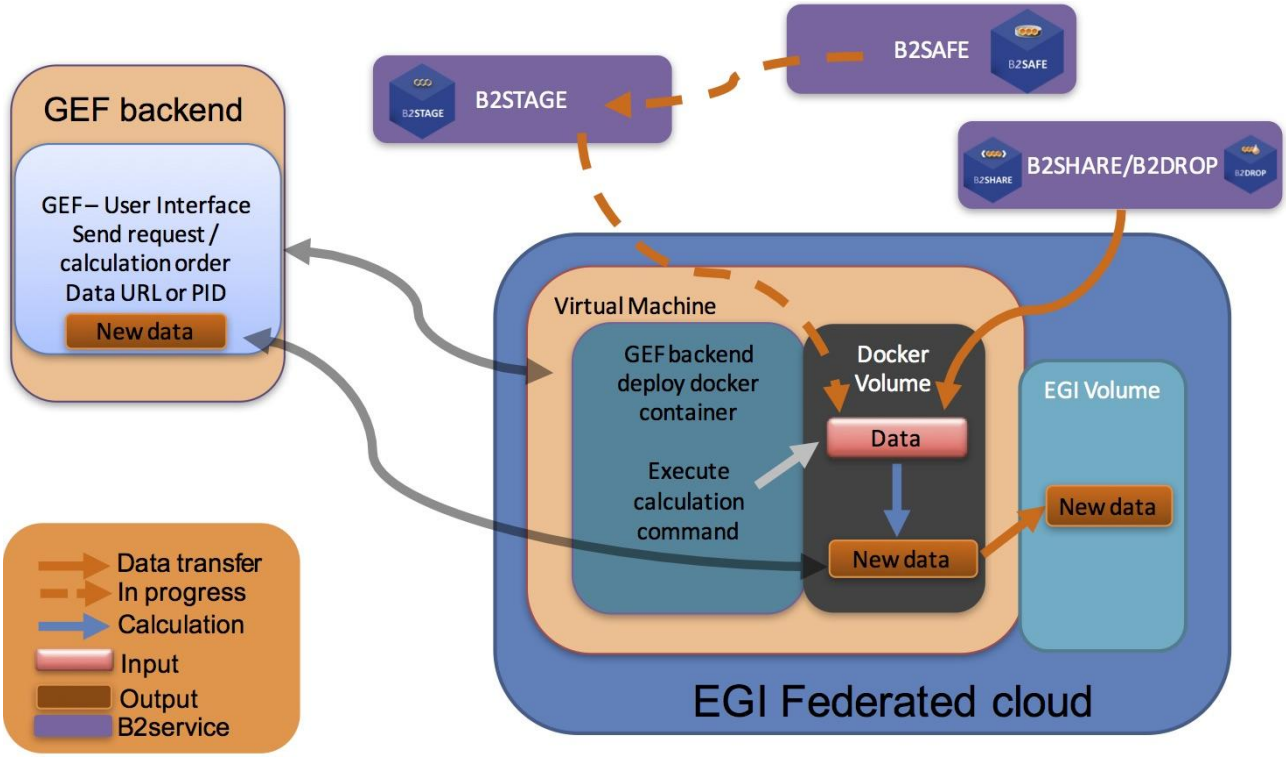

Figure 5: ENES use case prototype workflow of deploying GEF execution on EGI FedCloud

17 https://github.com/enolfc/fedcloud-userinterface
18 https://github.com/EUDAT-GEF/GEF

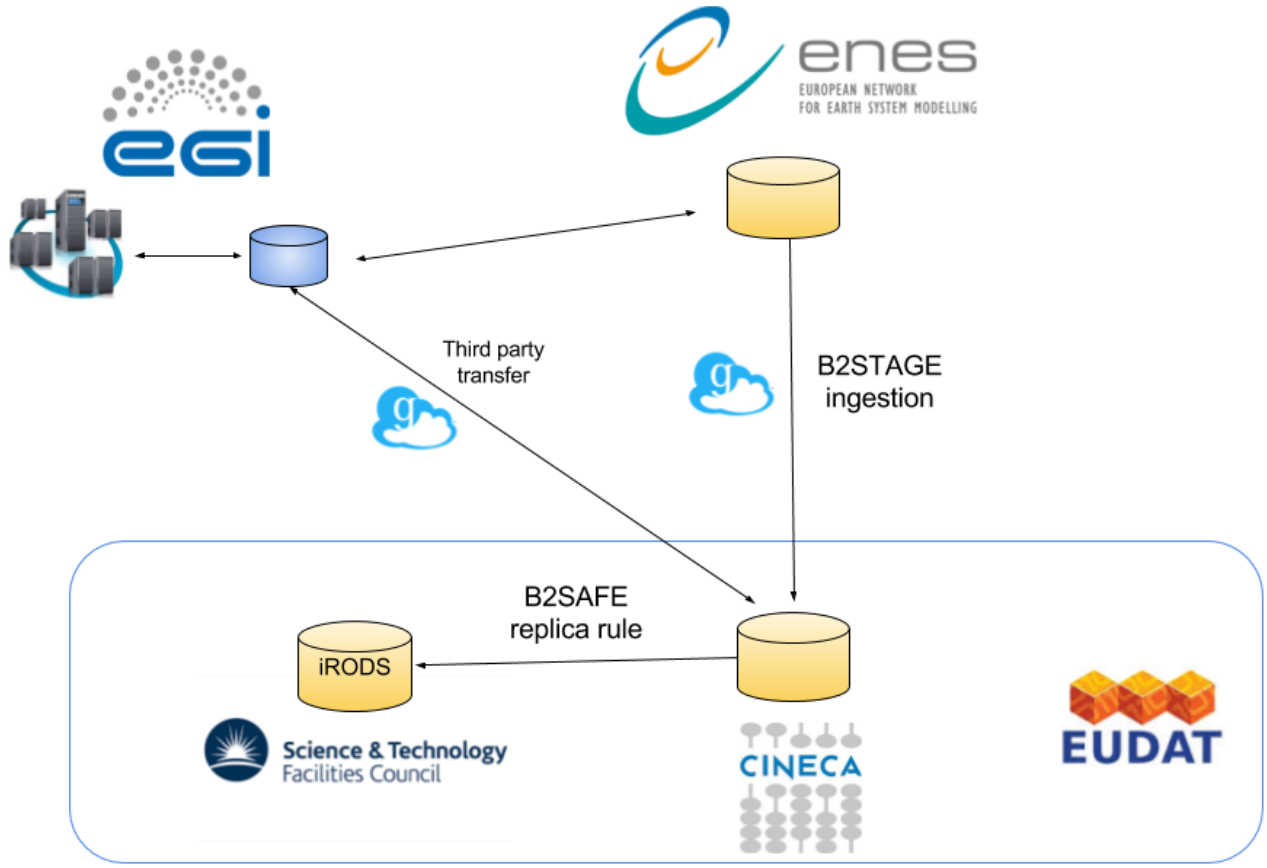

Figure 6: B2STAGE/B2SAFE architecture example in the ENES use case using CINECA (with SNFC instance as replica)

3.4. Use case independent challenges and observations

Some challenges were affecting more than only one use case:

Automatization freedom was obviously one unforeseen requirement. EGI cloud services were not at all intended to be used by automated scripts, they were just focusing on individual users. Formally within EUDAT B2ACCESS already allowed robot certificates, but still personal passwords were needed which created a potential security issue and bottleneck. Thinking about enabling and facilitating automatization changes the whole design process for the provided e-infrastructure services. The upcoming requirement on e.g. the EGI Federated Data Manager to be directly accessible is a good example on clear user feedback influencing the design towards providing added value to the research communities as end-users of the e-infrastructure services.

The user-driven approach comes with many advantages and will likely produce one of the best possible results, but it is also were demanding for all parties. First of all, it is very time-consuming since it incorporates an iterative process of testing, giving feedback, developers reacting on the feedback, releasing new beta version, rewriting documentation, production-ready release and finally testing again. Second it is very demanding for all parties: For communication to work, each others' terms have to be understood, problems and wishes have to be clarified and documented. This consists of an often steep learning curve for the user communities where not everyone is an expert in using e-infrastructure services and also more general ICT professionals might have been missing in the beginning of this interoperability task. Since the

user communities are very engaged they often invest more time in testing and providing their valuable feedback compared to their funding share. **The substantial time and trust investment by the user communities has to be matched by a better discipline of the e-infrastructure providers in creating the right expectation level on the actual status of the services and providing up-to-date documentation**.

For the user communities especially the initial orientation period was problematic as the actual implemented versions of the B2 services turned out to not yet have all the features available compared to the versions advertised to them creating a mismatch between expectations and reality. This misconception was even further fueled by the fact that the documentation available rather depicted the final vision for a service than its current state (e.g. B2ACCESS). Also confusing, or sometimes even wrong, information about events and meetings on the eudat.eu website has caused problems. Our early adopter core communities could access the Confluence wiki system used within EUDAT, but perceived it difficult to navigate as well as poorly organized and definitely not helping in finding up-to-date information about WPs, Deliverables and services.

Besides all that, the overall impression of the pilot users on the engagement between them and the experts within EGI and EUDAT was quite positive and the personal contacts were very much appreciated.

On top of that there were the more expected and technical parts of paving the path towards smooth interoperability between the two e-infrastructures:

- **Managing co-existing support systems and channels:** Not per-se a problem at this stage, but it adds to the perceived overhead by the user, especially in the orientation phase in the beginning
- **Multiplied policy and trust chains:** More harmonization on the operational policy level is still desirable. E.g. the setup of B2STAGE for a use case requires extra registration procedures at the site beyond the standard EUDAT registration to establish the responsibility chain. This has to be done only once, but it interferes with the idea of seamless access.
- **Consequences from 3rd party dependencies**: Globus Toolkit support will not be free of charge as of January 2018[19]. Since the tools that are part of that toolkit like GridFTP were quite central to early interoperability designs between EGI and EUDAT, getting aware of this change had consequences and introduced time constraints also for our interoperability work. EUDAT WP5 was working on developing a new B2STAGE HTTP API that allows users to ingest and retrieve data via a standard RESTful HTTP interface to replace gridFTP, lower the entry barrier and simplify integration into existing workflows while hiding the underlying technology.
- **Missing technical features and delayed development**:
  - Needed support of metadata handling within B2SAFE (GraphDB) still missing
  - The rushed development of B2STAGE HTTP API: the user communities got informed about it, waited for the final release, but had to wait for benchmarking tests to complete until external tests by the user pilots could start. Then there was also a technical problem to get people outside CINCECA test access, new documentation will still be provided[20], a new test system setup at PDC-KTH works with the local database, but not satisfactory yet with an actual iRODS system as backend.

3.4.1. Recommendations for service development and further harmonization

Up-to-date valid documentation written with the users point of view in mind

19 https://github.com/globus/globus-toolkit/blob/globus_6_branch/support-changes.md
20 https://gitter.im/EUDAT-B2STAGE/http-api

From a user perspective still more effort on the documentation of all the services is needed, there especially on more focused documentation that concentrates on simple recipes like e.g. for linking B2STAGE and B2SAFE with giving up-to-date examples as the users would like to see them. This means that checking the validation of the documentation should be part of the release cycle in order to keep the documentation **in-sync with the current real service offering** (not a desired vision or an outdated state) and remove old invalid links.

On-boarding further user communities

Before integrating new user pilots it is absolutely necessary to **establish a true common understanding** of

- What is to be achieved?
- What tools are already there?
- What resources will be needed (people, time,...)?
- What are the technical "boundary conditions" (platforms, software)!?

For all services, platforms and applications it should also be noted:

- How do they work?
- How do they interact?
- What are the benefits of using them (and not something else)?

In analogy to the main EUDAT Data Pilots, a **project enabler** should be appointed (in this case one each from both EGI and EUDAT) and a **direct contact person** acting as consulting expert for each of the offered services.

In the next step **simple entry-point step-by-step guides for typical alternative scenarios** marked by a different collection of underlying services answering the user's needs should be provided. This could really save the time of freshly onboarded research infrastructure communities in orienting themselves in the 'jungle' of all the different service offerings.

4. HARMONIZATION OF ACCESS POLICY

A real practical cross-infrastructure offering of services with a perceived seamless access requires harmonization on all levels. On the technical level this covers the interoperability of the services allowing for a seamless workflow execution, data discoverability and provenance, but also Authentication, Authorization and Identity (AAI) management as well as the combination of the respective service catalogues. On the policy level this includes access policies defining who should get access granted to which resources, the long-term perspective as well as needed operational policies. On the operational level harmonization also includes the operational tools (especially concerning user support), technologies and best practices, as well as the applied security framework and needed service level agreements (SLAs).

In general, the definition of the generic use case can in hindsight be considered as the major milestone in building the ground for a seamless common cross-infrastructure offering by EGI and EUDAT. It allowed the two infrastructures to understand and tackle the main technical parts of the challenge in interoperability.

For the security framework EGI, PRACE and EUDAT already have a long lasting collaboration with joint security trainings and workshops for the site security officers exchanging best practices and experiences with the SCI, the joint Security for Collaborating Infrastructures Trust Framework which is a collaborative activity within the Wise Information Security for e-Infrastructures (WISE) trust community endorsed by EGI, PRACE, EUDAT, GEANT and many more and entered now recently version 2.0 with SCIv2[21].

Concerning data and data policy management, ENES was approached to also act as a pilot user in testing a Data Management Plan (DMP) tool adoption by EUDAT2020 WP5 based on a Norwegian tool facilitating the creation of DMPs. The creation of DMPs follows an increasing requirement towards the researchers to attach them when submitting project proposals, both for National and International projects.

4.1. EGI/EUDAT AAI interoperability

Shortly after the definition of the generic use case it was realized that more AAI development work is required. In the generic use case pilot X.509 certificates were used for authentication which is a feasible but not optimal approach. Consequently both, the EUDAT and EGI infrastructures were co-designing and now implementing a new AAI infrastructure based on Identity Federation according to the AARC[22] guidelines. This work is not only important to increase the user-friendliness of cross-infrastructure services, it also improves overall usability of any data and computing services offered.

The main goal of EGI/EUDAT AAI interoperability, though, has always been the transparent access where you see the services offered by both EGI and EUDAT as offerings by a unique infrastructure once you are authenticated.

In order to be useful and actionable this general wish had to be broken down into smaller steps:

- Allowing users to access EGI and EUDAT web services with the same credentials
- Allowing users to access EGI and EUDAT non-web services with the same credentials
- Attributes harmonisation
- Enabling EGI services to delegate user's credentials to EUDAT services and vice versa
- Data privacy issues and policy harmonisation

Work done within the field of EGI/EUDAT AAI interoperability started with getting a deeper understanding of each other's AAI layers and breaking the task into smaller steps (see list above). An AAI overview

21 https://wise-community.org/wp-content/uploads/2017/05/WISE-SCI-V2.0.pdf
22 https://aarc-project.eu/

document was created and updated to document the current status and depict the different alternatives of potential desired solutions. Accounts were enabled for technical staff on each other's infrastructures and after some tests of the different alternatives an agreement could be made on which solution is the most desirable one. The preferred working template for a common EGI EUDAT AAI solution is characterized by featuring RCAuth[23] as common link.

Based on the chosen solution a common roadmap has been established by AARC with the timeline being revised by the input of EGI and EUDAT to match the planned update and rollout schedules for all the production services affected. For the policy harmonization it was also decided to rely on AARC which is the project working on the definition of an AAI layer for both e-infrastructures and research infrastructures. Currently, there are 2 pilots within the AARC-2 project that are dealing with EGI-EUDAT AAI interoperability that have been defined based on the input provided.

Since this harmonization work was and is of relevance for many other players as well this work had to be done in close collaboration with EGI, AARC and other work packages and tasks within EUDAT2020. Technically the move to RCAuth within EUDAT was not a problem as shown by the EUDAT RCAuth test instance which was soon available after the decision to move to RCAuth and interoperability tests on allowing users to access EGI and EUDAT web-based applications with the same credentials were successful. The migration of the production services had do be planned properly though. The details (e.g. Level of Assurance (LoA) provided by the Identity provider (IdP) and propagated through EUDAT B2ACCESS too low for some IdPs to validate a switch to RCAuth) are reported within WP5. In the future the EOSC-hub project will take over the activities that guarantee the interoperability between EUDAT's B2ACCESS and EGI's Check-in[24].

4.2. AARC Contribution

The AARC project is interested in authentication and authorization for research and collaboration and wants to enable the access of services and resources in a scalable and secure way by providing tools guidelines for infrastructure operators[25].

In order to collect input to the ongoing EGI-EUDAT pilots within AARC, a questionnaire for gathering the user requirements has been prepared. This questionnaire should help the user communities to determine what type of solution could be the best one for them to handle the registration, deregistration of users and their allocations. Focus will be on improving and advertising tools that are already there like eduTeams, EGI Check-In and EUDAT B2ACCESS. The outcomes of the not too technical questionnaire will be used as input for a common paper which will act as a starting point for later activities with the focus on accessing user portals allowing to submit automated workflows addressing the identified user requirement for more automatization and better integration within their own research infrastructure portals. EPOS is already part of the AARC project. The ICOS step-by-step setup guide was shown to AARC and ICOS and ENES RIs agreed to fill out the questionnaire once it is ready and particularly when EUDAT has rolled-out the new authorization infrastructure agreed on.

23 https://rcauth.eu/

24 https://www.egi.eu/internal-services/checkin/

25 https://www.youtube.com/watch?v=Xpwb6BNxNW4

5. WORK FORMAT AND DISSEMINATION

Most of the work of this second phase of the pilot activity was carried out mainly through virtual (https://indico.egi.eu/indico/category/165/) and physical meetings in close collaboration with the research communities and EGI as well as through written communication via the egi-eudat mailing list (egi-eudat@mailman.egi.eu) or direct email conversations and video meetings. Meetings were held typically once a month per user community with the chair alternating between EUDAT and EGI or on special topics (EGI-EUDAT AAI interoperability, user documentation). The resulting work was presented at:

- DI4R 2016, Krakow, Poland, between 28-30 September 2016[26]
- EUDAT User Forum, Helsinki, Finland, 23-27 January 2017[27]
- ICOS Germany Meeting, Offenbach, Germany, 22-23 March 2017
- EGI Community Forum, Catania, Italy, 9-12 May 2017[28]
- EUDAT summer school, Heraklion, Greece, 3-7 July 2017[29]
- EGI Final Review Meeting, Brussels, Belgium, 23-24 October 2017
- DI4R 2017, Brussels, Belgium, 29 November - 1 December 2017[30]
- 7th annual ESGF F2F Meeting, San Francisco, 4-8 December 2017
- EUDAT 2018 conference, Porto, Portugal, *planned* 23-25 January 2018[31]
- PDC Newsletter, *planned* March 2018, Vol 18 No.1, 2018, M. Barth,

  "Production cross-infrastructure services: towards seamless access" (article)

The EGI part of the work done on ICOS as a use case for the EGI OpenData Platform OneData was also presented in EGI Engage Deliverable D4.9 "Open Data Platform: Demonstrator, Experience Report and Use Cases"[32].

Overall the EGI view of the combined EGI - EUDAT collaboration has been reported in the EGI Engage Deliverable D4.8 "Cross-infrastructure case studies report"[33] with input also from EUDAT staff.

Especially at bigger events with a high proportions of e-infrastructure users like the EUDAT user forum or the EUDAT summer school great interest on the progress and the actual working details and experiences of the EGI EUDAT interoperability early adopters was perceived.

26 https://www.digitalinfrastructures.eu/content/about-di4r-2016
27 https://eudat.eu/events/user-meetings/eudat-helsinki-meeting-23-27-january-2017-helsinki-finland
28 https://indico.egi.eu/indico/event/3249/ https://indico.egi.eu/indico/event/3249/session/20/contribution/211
29 https://eudat.eu/summer-school-programme
30 https://www.digitalinfrastructures.eu/
31 https://eudat.eu/eudat-conference-2018-programme
32 https://documents.egi.eu/public/ShowDocument?docid=3033
33 https://documents.egi.eu/public/ShowDocument?docid=3026

6. CONCLUSIONS

The work conducted within Task 7.2 provided valuable input to both EUDAT and EGI in shaping their services to match the real needs of specific user communities.

Fortunately the user driven approach - even though demanding, time-consuming and constantly requiring a high investment of all parties when it comes to clear communication and learning each others' terms - was introduced early enough to reguide critical design choices that will in the future i.e. allow for automated access to the e-infrastructure services provided.

We have come a long way, however harmonization work is a continuous process including many different aspects and parties and will have to continue within the EOSC-pilot and EOSC-hub projects and activities.

While the technical interoperability challenges can be and are getting tackled, up2date end-user friendly documentation is often one step behind. Just providing simple entry point guides with different alternative service offerings available depending on user needs could save a lot of time for new research infrastructures entering the picture in orienting themselves in the abundance of different service offerings.

ANNEX A. GLOSSARY

Term	Explanation
AAI	Authentication, Authorisation and Identity
AARC	Authentication and Authorisation for Research and Collaboration
API	Application Programming Interface
BBMRI	Biobanking and BioMolecular resources Research Infrastructure
CA	Certificate Authority
CDI	Collaborative Data Infrastructure
CERFACS	Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique
CINECA	CINECA is a consortium acting as the largest Italian computing centre
CP	(ICOS) Carbon Portal www.icos-cp.eu
CPU	Central Processing Unit
DMP	Data Management Plan
ECMWF	European Centre for Medium-Range Weather Forecast
EDGAR	Emissions Database for Global Atmospheric Research
EGI	EGI Foundation also known as EGI.eu
EISCAT-3D	European Incoherent Scatter Scientific Association
ELIXIR	research infrastructure for life sciences
(IS)-ENES	(Infrastructure for the) European Network of Earth System Modelling
EPOS	European Plate Observing System
ESFRI	European Strategy Forum on Research Infrastructures
ESGF	Earth System Grid Federation
GridFTP	Grid File Transfer Protocol
HPC	High Performance Computing
HTC	High Throughput Computing
IaaS	Infrastructure as a Service
ICOS	Integrated Carbon Observation System
IdP	Identity Provider
iRODS	Integrated Rule-Oriented Data System
jOCCI	jOCCI is a suite of Java libraries enabling the OCCI (Open Cloud Computing Interface) protocol as standardized by OGF (the Open Grid Forum)
PID	Persistent Identifier
PRACE	Partnership for Advanced Computing in Europe
RI	Research Infrastructure
SaaS	Software as a Service
SNIC	Swedish National Infrastructure for Computing
STFC	Science and Technologies Facilities Council (in the UK)
STILT	Stochastic Time-Inverted Lagrangian Transport model
TAS	Temperature At Surface
TLS	Transport Layer Security
UI	User Interface
VM	Virtual Machine
VO	Virtual Organisation
VOMS	Virtual Organisation Membership Service