IS-ENES3 Milestone M5.1

Draft Architecture Design

Reporting period: 01/01/2019 – 30/06/2020

Authors: Philip Kershaw, UKRI; Sandro Fiore, CMCC; Stephan Kindermann, DKRZ

Release date: 10 July 2020

ABSTRACT

This document describes draft architectural plans for the ENES software infrastructure with specific consideration to its interaction and evolution alongside external collaborations in the European context and broader international efforts. These include the European Open Science Cloud, European Space Agency related initiatives, the Copernicus Climate Data Store and the Earth System Grid Federation. These plans will be developed over the course of the project and will feed into the deliverable D5.3 Architecture Design Plans at month 36.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824084

Table of contents

1 Applicable Documents

AD1 ANNEX 1 (part A), Research and Innovation action, number - 824084 - IS-ENES3, EU Directorate-General Research and Innovation, Ref. Ares (2018) 6225460 - 04/12/2018

AD2 D10.1 Architectural document of the ENES CDI software stack, Sandro Fiore

AD3 ESGF Future Architecture Report, v1.0, Philip Kershaw, Ghaleb Abdulla, Sasha Ames, Ben Evans, https://doi.org/10.5281/zenodo.3928222

2 Reference Documents

RD1 Earth Observation Exploitation Platform Common Architecture, Use Case Analysis Document: EOEPCA.TN.005, TVUK System Team, Version 1.0, 02/08/2019, https://eoepca.github.io/use-case-analysis/published/v1.0/

RD2 Earth Observation Exploitation Platform Common Architecture, Master System Design Document: EOEPCA.SDD.001, TVUK System Team Version 1.0, 02/08/2019, https://eoepca.github.io/master-system-design/published/v1.0/

RD3 OGC Testbed-13 EP Application Package Engineering Report, Pedro Gonçalves, Ref. OGC 17-023, 30/01/2018, http://www.opengis.net/doc/PER/t13-ES001

RD4 OGC Testbed-14: ADES & EMS Results and Best Practices Engineering Report, Paulo Sacramento, Ref. OGC 18-050r1, 08/02/2019, http://www.opengis.net/doc/PER/t14-D009

3 Objectives

Work package 5 in the proposal document AD1, states for this milestone, “Produce and communicate a draft architectural plan. Design diagrams enabling scalable, resilient, easy to operate and cost-effective infrastructure highlighting IS-ENES connection with European e-infrastructure”. M5.1 then concerns the architecture with a specific focus on the interface with external programmes of work in the European context and internationally. We present these as a set of different *collaboration views*, one for each related external programme to ENES:

1) Earth System Grid Federation. The future architecture work for ESGF including the long-term vision and work with external partners. We include the ESGF Future Architecture Report, the findings from a meeting of ESGF partner organisations to convened in order to develop a new system architecture
2) European Space Agency – specifically the Climate Change Initiative Open Data Portal which has re-used components and aspects of the software architecture from ESGF. Also, the Earth Observation Exploitation Platform Common Architecture, a programme of work which concerns a federation infrastructure much in the same way as ESGF
3) Copernicus Programme Climate Data Store - Services to deliver access to global and regional climate projections data provided by partners from the ENES consortium and built using technologies from ESGF
4) EOSC – The PID system, EOSC-Hub, future aspects - infra-EOSC and Regional clusters

This document should be considered in the context of the other milestones and deliverables for the project. D10.1 Architectural document of the ENES CDI software stack [AD2] is related but focuses on the technical implementation and internal work to IS-ENES3. M5.1 forms the basis for later deliverables D5.3 Architecture Design Plans and D5.4 IS-ENES3 Involvement in ESGF.

4 Collaboration Views

4.1 Earth System Grid Federation

The Earth System Grid Federation (ESGF) is a distributed software infrastructure for the dissemination of climate model data developed as part of an international collaboration effort between national research institutions working in the climate sciences. ENES makes a major contribution to this initiative both to the development and operations of the software services and ESGF software underpins the ENES-CDI. An effort is currently underway to re-architect and re-engineer the ESGF system. This is described in the ESGF Future Architecture report [AD3] which forms an integral part of this milestone and provides insights about the ESGF architecture for the next decade.

4.2 European Space Agency

4.2.1 Climate Change Initiative Open Data Portal

1.1.1.1 Introduction

The ESA Climate Change Initiative (CCI) is a programme of work whose goal is to provide stable, long-term, satellite-based essential climate variable (ECV) data products for climate modellers and researchers. As part of the CCI, ESA has funded the Open Data Portal project and its successor the Knowledge Exchange project to establish a central repository to bring together the data from these multiple sources and make it available in a consistent and harmonized form. The Portal provides a single point of access for the data to enable its dissemination to the user community. It consists of a web front-end, but also a back-end which provides the underpinning service layer and hosting infrastructure. ENES partner CEDA, worked on the latter with responsibility for the development of the metadata catalogue, data archive and data discovery and access services.

The development of this system has illustrated the ability to apply technologies from ESGF to a new set of climate data: observations predominately from satellite-based instruments. Experiences from this project have in turn provided important input into the future architecture for ESGF and the development of an infrastructure for the ENES community.

1.1.1.2 Data Management

Currently, 19 Essential Climate Variables (ECVs)[1] are hosted resulting in an output of around 300TB. These are produced by a range of different institutions representing a varied set of data

[1] http://cci.esa.int/data

both in terms of the associated science and spatio-temporal characteristics. Data heterogeneity is a major challenge for hosting and archiving the data. The establishment and adherence to strict data standards[2] was a critical factor in enabling the effective integration of data and services. This follows the philosophy of other ESGF projects. The majority of datasets are in netCDF format and conform to the CF (Climate and Forecast) metadata conventions.

1.1.1.3 Evolution of the System Architecture

A range of data serving applications were required to meet the needs of different user communities and maximise dissemination of the data. Technology from ESGF was re-used directly to implement the system: the ESGF Data Node to provide data access and the Index Node, and search services. The project re-used the ESGF concept of a *Data Reference Syntax* (DRS), a set of controlled vocabularies of terms used to organize the data and provide search facets. However, it also made the additional innovation of expressing these vocabularies in a machine-readable form using Linked Data[3] technologies. These can be queried from a vocabulary web service providing a single authoritative source for different components in the system – data publishing and search services.

However, the ESGF system alone did not provide the capability required by ESA. A different web front-end was needed and notably an additional search service was required based on the OGC CSW standard. This and ESGF's bespoke search system were fundamentally incompatible with one another, highlighting the need for standardisation and adoption of community standards for search interfaces and metadata. Recognising this problem, the ESGF Search was dropped in the second phase of the work (Knowledge Exchange project) in favour of OpenSearch. The latter is widely used in the Earth Observation community[4] and is capable of providing the required capability covered by the OGC CSW and the ESGF search system.

A second challenge encountered in the first phase of work was the ability to publish data into the system given the variety and volume of ECV datasets. Some of them contained vast numbers of files which put a strain on the ESGF data publishing system. This demonstrated the need to re-engineer and improve the latter. For the Knowledge Exchange, a new publishing system was adopted which supports parallel execution for ingestion processes and can cater for files formats other than netCDF encountered for some CCI datasets. The resulting architecture is shown in the following Figure 4-1.

[2] http://cci.esa.int/sites/default/files/CCIDataStandards_v2-1_CCI-PRGM-EOPS-TN-13-0009.pdf

[3] https://www.w3.org/standards/semanticweb/data

[4] http://ceos.org/document_management/Working_Groups/WGISS/Projects/OpenSearch/CEOS_OpenSearch_Best_Practice_Doc-v.1.0.1_Jun2015.pdf

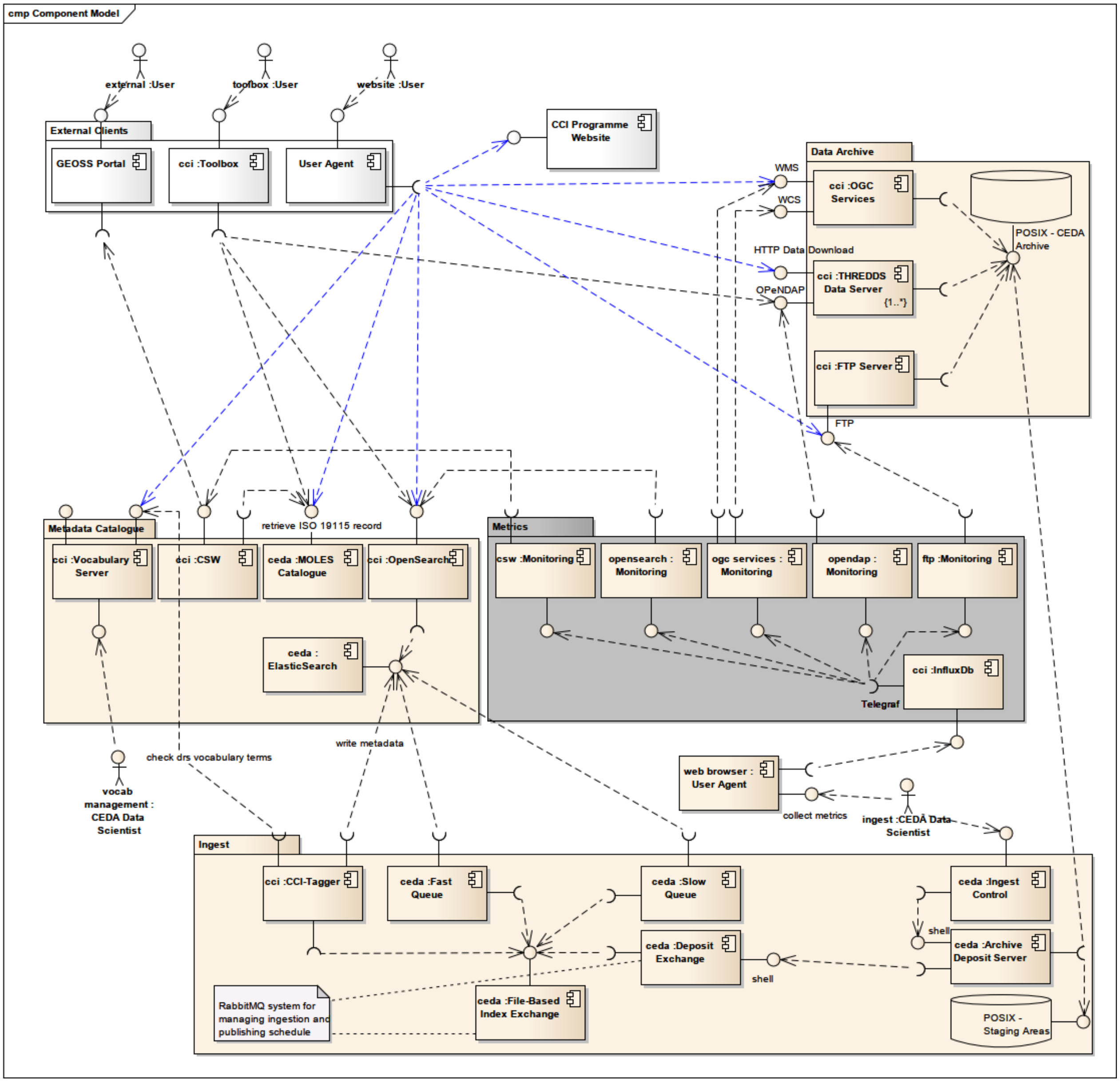

Figure 4-1: System Architecture for the ESA CCI Open Data Portal. The initial system was based on ESGF Data and Index Node software components. This was improved upon and redeveloped, adopting OpenSearch as the standard search interface and incorporating a new system for data ingestion and publishing

4.2.2 Earth Observation Exploitation Platform Common Architecture

1.1.1.4 Introduction and Context

The Earth Observation Exploitation Platform Common Architecture (EOEPCA) forms part of a broader initiative to facilitate better exploitation of satellite-based observations, especially data from the Sentinel missions. EOEPCA follows on from ESA's EO Exploitation Platforms a programme which has sought to develop Virtual Research Environments (VREs) consisting of a complete software stack for processing, analysis and visualisation of EO data products as a hosted solution such that rather than users downloading data to their own computers, they access a common platform where data and computing resources are co-located. This approach exploits cloud computing technology to provide a shared hosting environment. Exploitation platforms have been commissioned on the basis of thematic areas (Thematic Exploitation Platforms – TEPs). Examples include the Forestry and Polar TEPs. They have also been developed around regions (multi-thematic platforms and MEPs (Mission/Sensor Exploitation Platforms) e.g. Proba-V. Exploitation platforms have some commonality with the work around the ENES Climate4Impact Portal and also exploitation of ESGF technologies in the development of the ESA Climate Change Initiative Open Data Portal described in the previous section.

The EOEPCA, relates directly to the federated characteristics of ESGF. It concerns the development of an agreed common architecture and interfaces to enable distributed EO resources to interoperate with one another for the sharing of data and computing resources. This encompasses exploitation platforms but also the ESA EO Network of Resources. The latter concerns the resource tier – ICT providers and data hosting facilities which underpin exploitation platforms. Examples include the ESA DIASs (Data and Information Access Services) and national research infrastructure providers such as JASMIN[5].

1.1.1.5 System Architecture

The EOEPCA sets out a series of use cases [RD1] and proposes an overall system architecture [RD2]. The project will create a reference implementation of this architecture and deploy and test this on a selected hosting infrastructure. The work is being executed by Telespazio-VEGA UK under contract to ESA. Sub-contracts are awarded to individual supplies to deliver the various aspects of the system. The OGC is a funded partner in the work and the technical direction is strongly informed by the outcomes of recent OGC Testbeds – 13 [RD3] and 14 [RD4]. The use cases include, access to a given platform, discovery of data products and processing services and execution of the latter. These are then addressed in the system design under the domain areas: *User Management*, *Processing and Chaining* and *Resource Management* (Figure 4-2).

[5] http://www.jasmin.ac.uk/

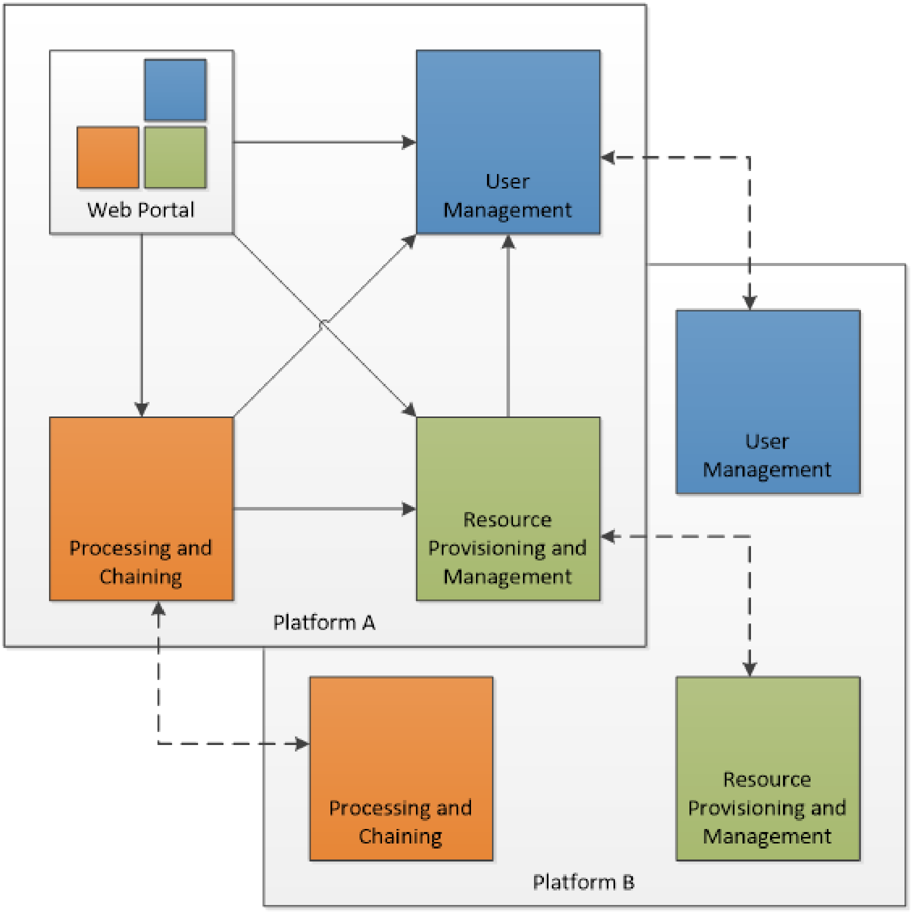

Figure 4-2: EO Exploitation Platform Common Architecture high-level functional domains

1.1.1.6 User Management

EOEPCA proposes an architecture for single sign-on (SSO) and federated authorisation. Sign-on is mediated through an IdP Proxy following the AARC Architectural Blueprint[6] and likewise selected for the ESGF Future Architecture. OpenID Connect is adopted as the standard for the SSO interface.

For authorisation, the system draws from ESGF's existing architecture with its system of Attribute Authorities to manage role-based access entitlement in VOs (Virtual Organisations). However, it also proposes the concept of delegation of authorisation to third party Policy Decision Points (PDPs) and suggests the use UMA[7] and XACML 3.0[8] as means to implement delegated authorisation.

[6] https://aarc-project.eu/architecture/

[7] https://tools.ietf.org/html/draft-hardjono-oauth-umacore-14

[8] http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html

1.1.1.7 Processing and Chaining

The Processing and Chaining functionality for EOEPCA is worth focussing on as it is a distinguishing element of the architecture from ESGF. A broad range of processing use cases are examined including bulk and systematic processing and notably the ability to build new processing services and chains of processing services. This builds on the work of Testbed 13 which introduces the concept of an Application Package (AP) and Application Deployment and Execution Service (ADES). Algorithms can be packaged as Docker containers and together with associated metadata for their execution (AP), can be registered, installed and run at a third-party site using an ADES. As initially designed, the AP only supports the packaging of a single container image. This restricts processing capability to the number of cores available on the given host node. However, the MSDD [RD2] also proposes the use of AP definitions based on CWL[9] workflows also.

The ADES wraps the OGC WPS standard and uses this as the interface by which a client can run the newly deployed algorithm. This overall concept provides a means to realise the model of bringing the compute to the data since a client can deploy an algorithm in situ where data is hosted. The chaining together of processing services in workflows is also considered drawing on Testbed 14, which introduces the concept of an EMS (Execution Management Service) itself along with the ADES are derived from the TEP architecture. The EMS provides overall management of processing and chaining and delegates execution to individual ADES instances.

1.1.1.8 Resource Provisioning and Management

This domain covers resource discovery, access and visualisation. Metadata associated with processing is a first-class resource alongside resources more traditionally associated with metadata catalogues, datasets. Metadata is organised around the Browse, Discovery and Archive classes[10]. It considers different models for data discovery and federation: a) Federated Catalogue Gateway, b) Centralised Federated Catalogue and c) Distributed Federated Catalogue. ESGF currently supports c) but b) is proposed for the future architecture. The FedEO[11] system follows a) linking together catalogues at different sites.

For the management of underlying computing resources Kubernetes is proposed as an abstraction and a *Data Access Gateway* in order to provide a wrapper layer to account for the expected data access protocols for the processing algorithm to be executed versus the access protocols available on the host environment. For example, an algorithm may expect a POSIX file system for data in and out, but the host environment may have data only available via an S3 interface to object storage. The data would need to be staged from an object store to a local POSIX file system cache or else a POSIX wrapper provided over the top of the S3 interface.

[9] https://www.commonwl.org

[10] https://doi.org/10.1098/rsta.2008.0237

[11] http://wiki.services.eoportal.org/tiki-index.php?page=FEDEO

The document does not consider a common data model for data itself within the system. For example, the semantics for handling of different data formats and the consistency of data in terms of structure and compliance with standards.

4.3 Copernicus Programme Climate Data Store

4.3.1 Introduction

The Climate Data Store (CDS)[12] is part of the Copernicus Climate Change Service (C3S). C3S is operated by ECMWF on behalf of the European Union. It aims to provide key indicators of climate change drivers, supporting all sectors. The CDS provides a single, freely available interface to a range of climate-related observations and simulations. These are sourced from many participating organisations. ENES partners have worked together on two different contracts, to provide quality-controlled subsets of CMIP5 and CORDEX global and regional climate projections data respectively. These have built on work together in the ENES collaboration exploiting software from Earth System Grid Federation (ESGF). The architecture of these systems was driven largely by the key requirement from the CDS for an operational uptime of 98%. This led to a solution which took advantage of elements of ESGF distributed architecture but also public cloud.

4.3.2 High-level Interfaces

CDS uses a plugin architecture whereby different data providers comply with a standard set of interfaces to make their data available. Figure 4-3 shows the arrangement for providing the CDS with access to CMIP5 data. An ICD (Interface Control Document) defines the interfaces provided by the data provider to CDS.

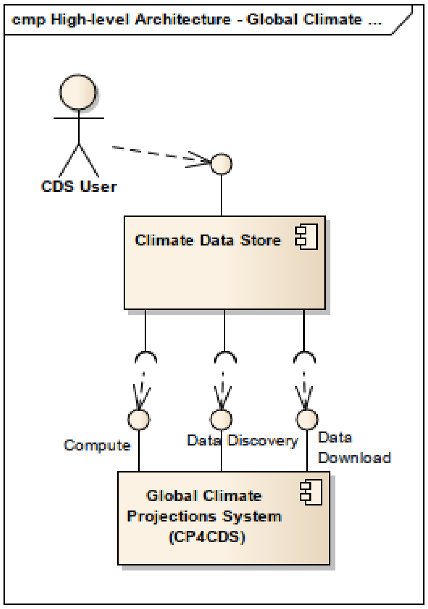
cmp High-level Architecture - Global Climate ...
CDS User
Climate Data Store
Compute
Data Discovery
Data Download
Global Climate Projections System (CP4CDS)

Figure 4-3: Interface between Climate Data Store and data providers, in this case Global Climate Projections data from CMIP5

[12] https://cds.climate.copernicus.eu/#!/home

The solution leverages software components and architectural elements from ESGF.

- Data access is provided using OPeNDAP[13], a web service for sub-setting gridded data products;
- ESGF provides a web service API[14] for data discovery;
- Finally, although ESGF does not provide a standard compute service, work of the ENES project partners was built upon to implement this. This is based on the Web Processing Service (WPS)[15] standard from the OGC (Open Geospatial Consortium).

4.3.3 Operational requirements

The CDS requires a high level of availability for services stipulating a 98% uptime. This contrasts with 95% typical for research infrastructure in general, representative for the project partners normal operating model.

To satisfy this requirement two overall options were explored:

1) Leverage the redundancy available by virtue of the ESGF federated architecture
2) Use Public (commercial) Cloud to host the required data and services off-premise in an infrastructure that can be configured to support the required resilience.

ESGF inherently supports redundancy and replication capabilities:

- Identical copies of data can be served independently at nodes hosted by different host organisations
- The search system is federated such that users can discover data hosted at any one site *and* find out the location of replicas at other sites
- There is an established system for bulk copying of large volumes of data between sites
- It represents an existing community with established operational procedures and governance
- Data access services rely on underlying storage based on traditional POSIX file systems
- Costs for compute and storage are less than public cloud. With the latter the cost margin is much greater.
- Capital (hardware purchase) and recurrent (operations) cost model

Public cloud has built-in capability for resilience:

- Supports the concept regions and availability zones to allow data and services to be duplicated and so provide resilience
- Supports elasticity enabling services to scale to meet demand
- Solutions for shared POSIX storage available. Object storage is more cost-effective but the existing ESGF software was not adapted to use it
- Built-in tooling to support deployment, operations, monitoring and metrics reducing overheads
- Storage and compute costs greater than for on-premise equivalent
- Recurrent costing model

[13] https://www.opendap.org/

[14] https://github.com/ESGF/esgf.github.io/wiki/ESGF_Search_REST_API

[15] https://www.ogc.org/standards/wps

4.3.4 Hybrid on-premise / public cloud solution

Public cloud provides an attractive solution but considering the large volume of data associated with climate model outputs, the hosting cost would be large. Analysis showed however that by replicating services and data across the three sites and load balancing the three it was possible to provide an aggregate uptime meeting the 98% requirement. Existing ESGF capability could be leveraged to replicate datasets and search metadata between the three sites CEDA, DKRZ and IPSL.

DNS-based and proxying methods were explored in order to provide load balancing. The latter was rejected since it involves data traffic being funnelled through a single point. Here public cloud was exploited taking advantage of Amazon Web Services Route53[16] for DNS-based load-balancing. Through a set of configurable tests, it is possible to check the health of the load balanced services removing unhealthy instances should the tests fail and re-instating them when they are restored to the healthy state. See Figure 4-4.

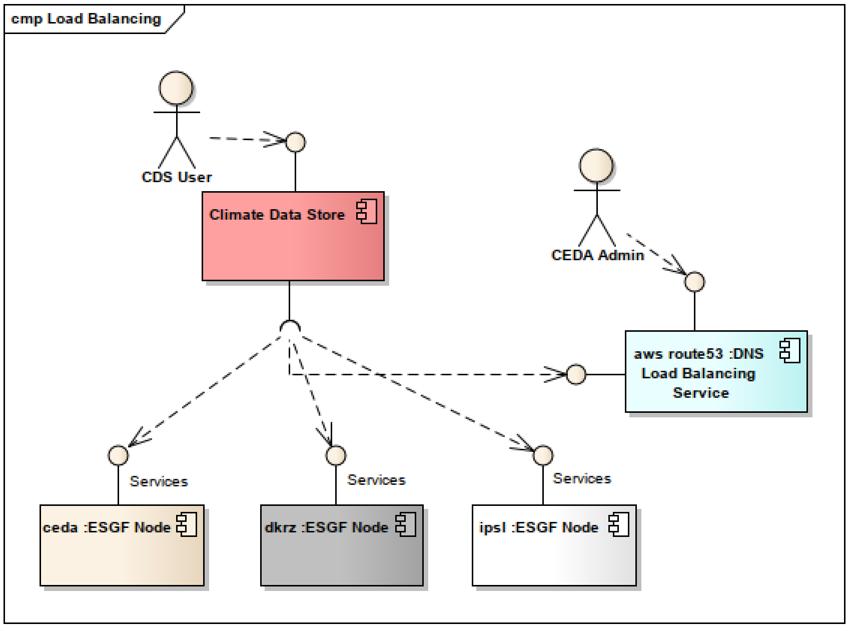

Figure 4-4: DNS-based load balancing services replicated at three sites

4.3.5 Publication and Synchronisation of Data

With a solution established for load balancing, a system needed to be developed to publish data and synchronise it across sites so that clients obtain a consistent response for a given query through the service. For this the CEDA site acted as the reference point with a master copy of CMIP5 data archive and the search index. The ESGF publishing software creates a manifest of all data files published and this forms the basis for the replication process to the other two sites

[16] https://aws.amazon.com/route53/

synchronising using GridFTP[17]. ESGF's search system based on Apache Solr NoSQL database uses its sharding system to automatically search index replicas from CEDA to search services at DKRZ and IPSL.

Figure 4-5, shows the final architecture for the system with load balancing in place for each of the three replicated services: search (Index Node), OPeNDAP (Data Node) and WPS (Compute Node).

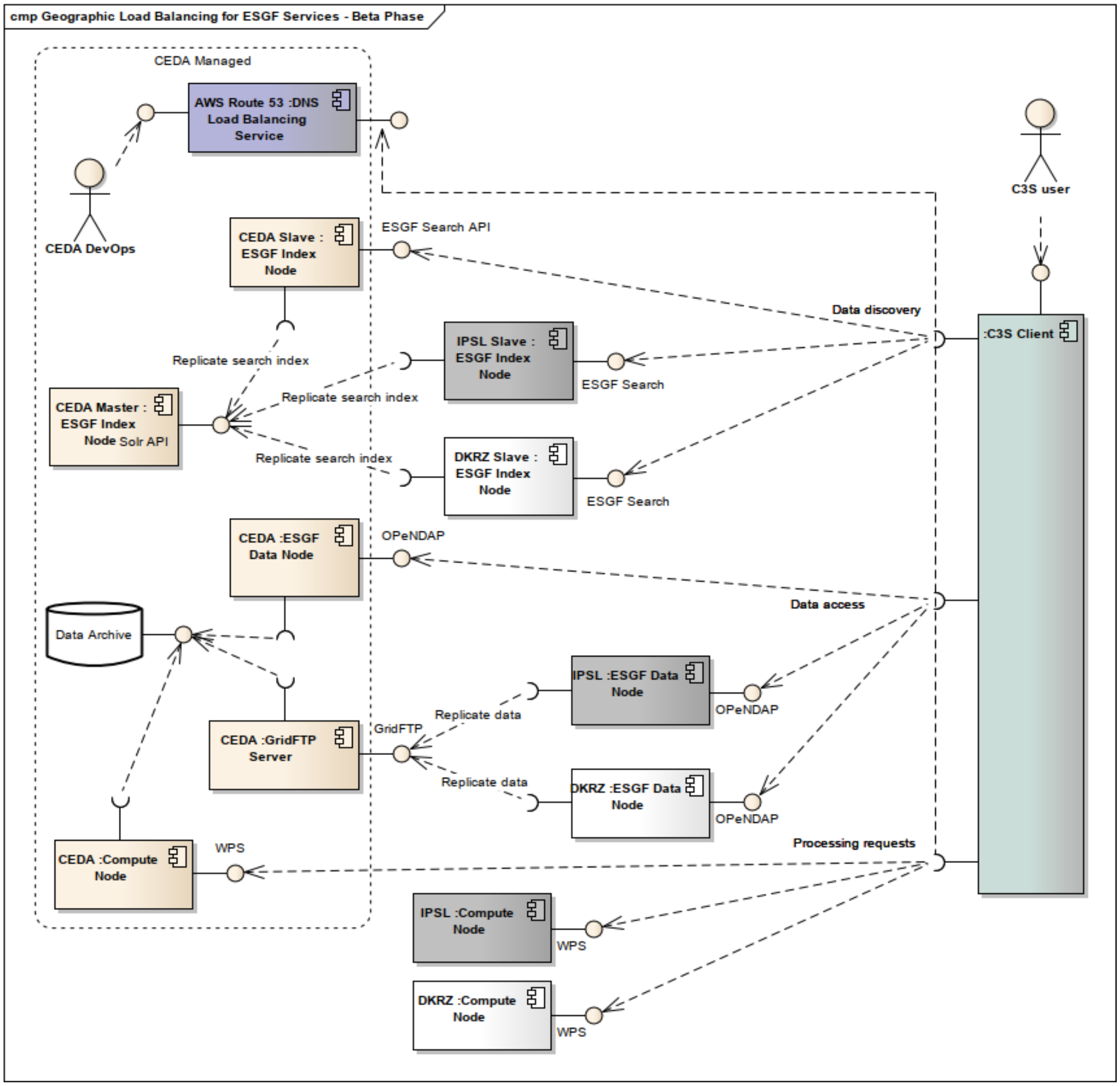

Figure 4-5: Complete deployment showing replicated services at CEDA, DKRZ and IPSL in load balanced configuration for use by the CDS

[17] https://github.com/globus/globus-toolkit

4.3.6 Containerised Deployment

An additional dimension to the project was the ability to scale services to meet demand. As part of a separate initiative in ESGF, a containerised version of the software components had been developed. When deployed with the container orchestration system Kubernetes, this allows deployment in an elastic configuration using a Horizontal Pod Autoscaler (See Figure 4-6). A test deployment was made at the CEDA site successfully demonstrating this arrangement for THREDDS Data Server, the software used to provide OPeNDAP. When exposed to more client requests, the Kubernetes cluster spawned additional THREDDS containers to manage the additional load and likewise, destroyed unused capacity when the client load reduced.

A deployment of the containerised ESGF was also made on Google Kubernetes Engine (GKE)[18] and used to provide search services. The deployment was a trivial process by virtue of the containers, Helm Charts used to specify the deployment configuration and the ready-made Kubernetes cluster provided by GKE.

[18] https://cloud.google.com/kubernetes-engine/

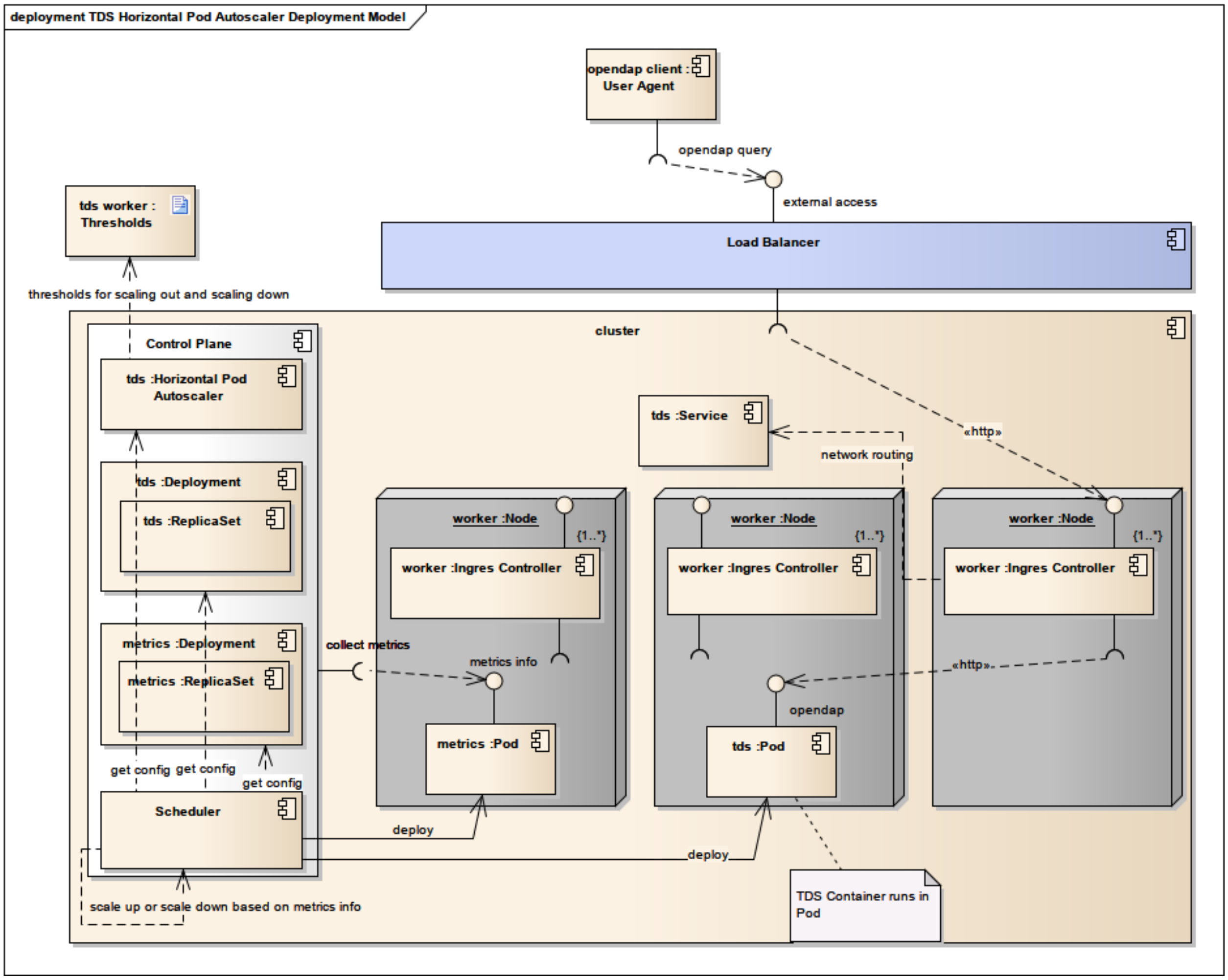

Figure 4-6: Kubernetes cluster with THREDDS Data server deployed in a configuration with horizontal autoscaling

4.3.7 Conclusions and next steps

The collaboration demonstrated the ability of ENES partners to develop effectively a European ESGF super node with a high operational uptime. ESGF's inherent distributed architecture provided a head start in developing a system with built-in redundancy through replicated services and data. Public cloud was utilised to provide a load balancing system and together with the container based ESGF deployment system demonstrated the potential for rapid deployment of an ESGF site, normally a more complex and lengthy process. This has provided an impetus for the redesign of the current ESGF system through the Future Architecture initiative.

Replication of data and the maintaining of data consistency between sites has proved to be operationally burdensome. Even so, the load balanced system has paid dividends with the ability to continue operations uninterrupted on service failure at any one site and also the ability to perform zero-downtime maintenance. Since the original contracts to build the system have been completed, project sponsors and operators of the CDS, ECMWF have simplified the required interface. Search services are no longer required and OPeNDAP is replaced with simple HTTP file serving. This reduces the operational burden for the overall system.

4.4 European Open Science Cloud

4.4.1 Introduction

The European Open Science Cloud (EOSC) initiative has been proposed in 2016 by the European Commission as part of the European Cloud Initiative to build a competitive data and knowledge economy in Europe. The EOSC will offer a virtual environment with open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines by federating existing scientific data infrastructures, currently dispersed across disciplines and the EU Member States. ENES partners are contributing to EOSC-related projects by representing the ENES community and bringing requirements related to the climate domain. Prior to that, over the last decade, ENES partners have been involved in several e-infrastructure projects related to EGI and EUDAT (two key European infrastructural players) by providing inputs and developing use cases targeting infrastructural aspects, for the climate community.

4.4.2 EUDAT Collaborative Data Infrastructure (EUDAT CDI)

ENES partners established in support of CMIP6 an ESGF infrastructure extension to enable highly scalable persistent identifier (PID) registration as part of the ESGF data publication process. The central backend for this is hosted at DKRZ and relies on the handle PID system (see http://handle.net). To enable sustainable operation of such a service at the scale needed for CMIP DKRZ joined the EUDAT CDI and also participates in the European Open Science Cloud. Thus e.g. the PID client library code base can be evolved as part of EOSC and interdisciplinary use cases. By joining the European PID Consortium (ePIC) it was also ensured that the CMIP/ESGF PIDs are replicated and thus can be resiliently and efficiently resolved. Architectural design aspects of this service are described as part of the D10.1 deliverable.

4.4.3 European Grid Infrastructure (EGI)

The EGI Federation is an international e-Infrastructure set up to provide advanced computing and data analytics services for research and innovation. It comprises hundreds of data centres and cloud providers spread across Europe and worldwide. ENES partners have been collaborating with EGI in several e-infrastructure projects, especially with respect to large-scale data analytics in the cloud. The participation in such projects (i.e. EOSC-Hub) has been beneficial to expand the capabilities of the proposed climate software/services towards other communities (cross-community fertilization) as well as their ability to be deployed into federated cloud environments (FedCloud). Additionally, it has also helped to (i) have a better understanding of the EOSC landscape, (ii) represent the ENES community bringing requirements related to the climate domain into a multidisciplinary EU context and (iii) work more synergistically at European level with a key actor like EGI on infrastructural aspects. In this respect, to strengthen the link with EGI, CMCC has recently applied to become an EGI member representing the IS-ENES data infrastructure, under the endorsement of the ENES Data Task Force. Such a membership application has been approved by the EGI Council in June 2020.

4.4.4 EOSC-Hub

The EOSC-hub project (January 2018-December 2020) creates the integration and management system of the future European Open Science Cloud that delivers a catalogue of services, software and data from the EGI Federation, EUDAT CDI, INDIGO-DataCloud and major research e-infrastructures. This integration and management system builds on mature processes, policies and tools from the leading European federated e-Infrastructures to cover the whole life-cycle of services, from planning to delivery. EOSC-Hub provides a comprehensive service portfolio delivered through its marketplace[19].

ENES partners are participating in EOSC-Hub in the context of three service categories:

- Data search and catalog service: ENES partners host and co-develop the EOSC B2Find catalog service
- Persistent identifier service: ENES partners contribute to the EOSC B2Handle service (see 4.4.2)
- Compute service: ENES partners contribute to the ENES Climate Analytics Service (ECAS) by hosting two instances, respectively at CMCC and DKRZ. In such a context, ECAS has been integrated with other EGI[20] and EUDAT services to tackle more advanced, distributed data & computing scenarios in the cloud as well as cross-community aspects to extend the service adoption beyond the climate domain. Figure 4-7 provides architectural insights about the proposed design and shows the link between ECAS and the following services: B2DROP, B2SHARE, B2HANDLE, IAM, Onedata, EGI Check-in, Infrastructure Manager and JupyterHub.

[19] EOSC Marketplace https://marketplace.eosc-portal.eu/

[20] Elastic deployment of ECAS on EGI - https://www.egi.eu/about/newsletters/elastic-deployment-of-ecas-on-egi/

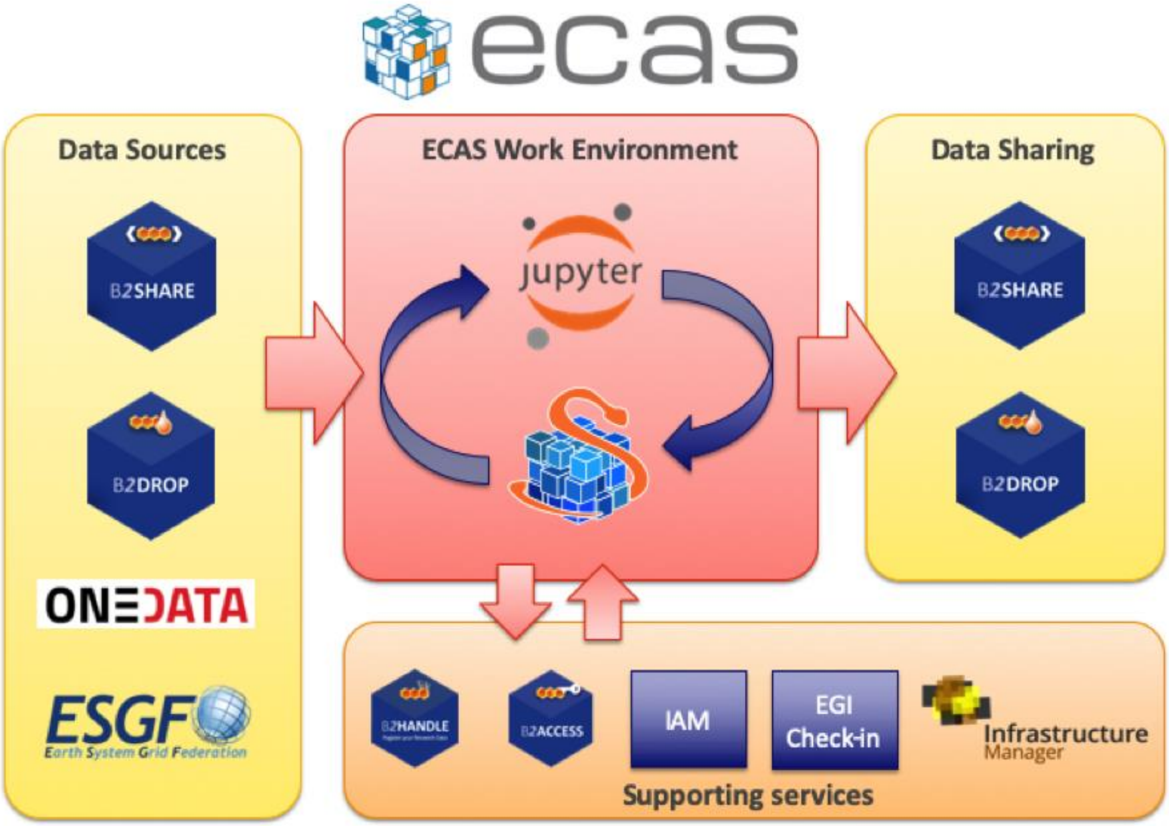

Figure 4-7: ECAS architecture overview with all the integrated services

4.4.5 Regional clusters: EOSC Pillar

EOSC-Pillar[21] (July 2019-June 2022) gathers representatives of the fast-growing national initiatives for coordinating data infrastructures and services in Italy, France, Germany, Austria and Belgium to establish an agile and efficient federation model for open science services covering the full spectrum of European research communities.

As part of the EOSC Pillar project, ENES partners contribute in different use case developments. One objective is to enable interdisciplinary FAIR[22] data handling and processing. In this respect, partners are expected to provide architectural input and prototype data processing (based on the Pangeo software ecosystem) and associated provenance capture, together with other regional research infrastructures in an interdisciplinary European context.

[21] EOSC-Pillar website https://www.eosc-pillar.eu/

[22] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

5 Conclusions

Experiences from the Copernicus climate projections and ESA CCI Open Data Portal projects demonstrate the practical application of the ESGF software stack by ENES partners to new problem domains. These have directly informed the ESGF Future Architecture initiative and consequently the evolution of the architecture for IS-ENES.

The Copernicus projects have demonstrated the inherent flexibility of the existing architecture in meeting new requirements and creating a resilient system delivered through close co-operation between ENES partners. Implementation and operation of the CCI Open Data Portal has however highlighted the need to update the underlying ESGF software system and adopt community standards to enable better integration with other similar systems in the Earth sciences. Notably the Portal has migrated to OpenSearch as its search standard following the CEOS guidelines. This is being investigated as an option for ESGF. The EOEPCA provides an opportunity to engage with ESA activities and directly input into the evolution of a federated system for a related domain – Earth observation – in the broader Earth sciences. Additionally, ESA’s close involvement with OGC Testbeds gives a route for the ENES community to input into the development of new standards for data sharing and processing.

ENES partners DKRZ, CERFACS and CMCC have strong links with elements of EOSC. The PID system developed as part of EUDAT has made an important contribution to CMIP6 enabling the tracking of dataset versions. EOSC-Pillar fosters the co-operation of European partners in the development of shared services relevant to the ENES CDI specifically with regard to compute services and the application of open source software stack from Pangeo. This is also more broadly applicable to the evolution of the ESGF architecture.