D4.3 2nd Periodical assessment of AI and Infrastructure services

Abstract

The report provides the second-year usage statistics and assessment of all the Artificial Intelligence platform and the underlying Infrastructure services provided under virtual access in WP4.

iMagine receives funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101058625.
https://www.egi.eu/project/imagine/

Document Description

D4.3 2nd Periodical assessment of AI and Infrastructure services			
From: Work Package 4			
Due date	30-09-2024	Actual delivery date:	14-10-2024
Nature of document	Report	Version	1.0
Dissemination level	Public		
Lead Partner	CSIC		
Authors	Álvaro López García (CSIC)		
Reviewers	Gergely Sipos (EGI Foundation)		
Public link	https://zenodo.org/records/14894023		
Keywords	AI, Virtual Access, Infrastructure, Cloud, CPU, GPU, Storage		

Revision history

Issue	Date	Comments	Author/Reviewer
V 0.1	25/09/2024	First draft for ASB meeting	Andrea Anzanello (EGI)
V 0.2	07/10/2024	Updated based on feedback from reviewer and project ASB	Álvaro López García (CSIC)
V 1.0	Submitted version		Andrea Anzanello (EGI)

Copyright and licence info

Table of content

List of Images

1. Executive summary

This report provides an assessment at M24 of the WP4 installations provided by the iMagine project under the Virtual Access (VA) mechanism. This assessment is based on the metrics collected for WP4 installations during the second year of the project, i.e. between September 2023 - August 2024.

WP4 installations can be classified in two groups:

- iMagine AI platform, operated in two interweaved installations:
  - Task 4.1 - iMagine AI **development** service
  - Task 4.2 - iMagine AI **deployment** service
- Infrastructure Services (Task 4.3) where 4 providers (CSIC, LIP, Walton, Tubitak) provide compute and storage services underpinning the iMagine AI platform, alltogether offering 132,000 GPU hours, 6,000,000 CPU and 1,500 TB month during the 36-month long project.

During PY2 the "iMagine AI development service" has been used by the 8 use cases of the project at an increasing scale, due to the fact that they have entered into intensive model training phases for their AI applications. Moreover, during the 2nd project year external use cases have been supported via the iMagine open calls[1].

As an overall result, the GPU, CPU and storage consumption increased, the number of AI modules available in the iMagine marketplace[2] has reached 14, and the installation served more than 60 users from 12 countries.

All the use cases that the project supports are concluding the service development phase (training and validating AI models) with some of them already transitioning to the AI "iMagine AI deployment service". However, the numbers in the deployment service (i.e. where end user applications are delivered to scientists) are low. Noteworthy, these numbers only reflect the real utilization of AI applications, as the underpinning technology allows us to scale to zero resources, when the applications are not being actively used.

[1] https://www.imagine-ai.eu/article/imagine-call-for-use-cases/
[2] https://dashboard.cloud.imagine-ai.eu/marketplace

The growing compute-storage demand was satisfied by the 4 cloud infrastructure providers from Task 4.3 are fully integrated into the iMagine platform. Three of the sites (CSIC-IFCA, INCD and TUBITAK) are powering the iMagine AI development service, whereas the iMagine AI deployment service is backed by CSIC-IFCA (catch-all instance) and WaltonCloud (use-case specific instances). During this period, the providers have delivered 4,197,116 CPU-hours, 183,262 GPU-hours and 495 TByte-month storage to the platform and the supported use cases. Despite the total GPU consumption is higher than was originally anticipated, and already consumed the 132,000 GPU-hours that was budgeted in the project, all the GPU providers are able to continue supporting the project in the 3rd year, especially because the GPU consumption is expected to lower (and CPU consumption increase) as the use cases transition from model training to service delivery.

1 Introduction

Virtual Access (VA) is financial instruments to reimburse the access provisioning costs to access providers. This instrument is provided by the European Commission to increase the sharing of research infrastructures and services that otherwise would not be available to international user groups.
In VA, the services – also called "installations" – must be made available 'free of charge at the point of use' for European or International researchers. VA access is open and free access to services through communication networks to resources needed for research, without selecting the researchers to whom access is provided.
Virtual Access to services of the iMagine catalogue applies to the following 2 categories:

1. AI platform and compute infrastructure services in WP4
2. Imaging data and analysis service in aquatic sciences in WP5

This document provides Virtual Access metrics and assessment for WP4 during the 2nd year of the project (Sep 2023 - Aug 2024).

In the 1st project year WP4 worked on the establishment of the iMagine AI platform, serving the 8 use cases that are part of the consortium, and participated in the setup of the open call to attract further users from Q3 2023. During the 2nd project year, WP4 has worked on streamlining the integration and failover mechanisms for new sites into the platform, and the final integration with the application delivery services, namely OSCAR.

1.1 WP4 Installations

Within iMagine project 6 installations are part of Virtual Access work package 4. These installations support the baseline computing infrastructure of iMagine as part of the following services and their usage metrics:

- iMagine Platform AI Application Development Service (formally called DEEP): Is for development and validation of AI models. The service was used during PY1 and PY2 by the 8 use cases. The usage is monitored with the following metrics:
    - ML training cycles measured in CPU/GPU hours
    - Number of AI models trained

    - The total number of AI models developed both in the marketplace and private
    - Names of the countries reached over last year (users' location)
    - Number of the countries reached over last year (users' location)
- iMagine Platform AI Application as a Service (formally called DEEP): Is for the delivery of validated models 'as services' for external users. During PY1 the service was not delivered, as the use cases were still under development. During PY2 we started delivering this service to the use cases, that have started using it during the last period phase. The usage is monitored with the following metrics:
    - ML application usage cycles measured in CPU/GPU hours
    - Number of AI applications hosted via the iMagine platform
    - Names of the countries reached over last year (users' location)
    - Number of the countries reached over last year (users' location)
- Cloud compute and storage infrastructures underpinning the previous two platform services:
    - IFCA CSIC Scientific Cloud (Spain)
    - INCD cloud (Portugal).
    - TR-FC1-ULAKBIM (Turkey).
    - WaltonCloud (Ireland).

  All resource providers are integrated in the platform and are being used by the platform services introduced previously, except WaltonCloud that is used exclusively for the iMagine Platform AI Application as a Service. The usage is monitored with the following metrics:
    - Number of users
    - CPU/GPU node-hours served
    - Storage served
    - Names of the countries reached (users' location)
    - Number of countries reached over the period (users' location)

  There is 132,000 GPU-hours, 6,000,000 CPU-hours and 1,500 TB-month capacity budgeted in WP4.

2 Installations

2.1 iMagine - AI Application Development Service

Description	The iMagine AI Application Development Service allows Artificial Intelligence developers to prototype, build and train AI applications, exploiting resources from EU e-Infrastructures. The installation allows the prototyping of AI models and applications through the train-test-evaluation cycle on underlying GPU-CPU-Storage. Once a model has been initially built, the Dashboard allows users to interact with resources and with the Open Catalogue. The service can store the history of all the performed training sessions for the monitoring the status of training directly from the training Dashboard. The development environment is based on JupyterLab instances, where users have access to major data science, artificial intelligence, machine learning and deep learning frameworks and various tools, with corresponding user support.
Task	T4.1
URL	https://dashboard.cloud.imagine-ai.eu/
Service Category	Infrastructure service
Service Catalogue	https://marketplace.eosc-portal.eu/services/imaging-ai-platform-for-aquatic-science
Providers	CSIC, LIP, UPV, KIT, IISAS
Location	Spain, Slovakia, Germany, Portugal
Duration	M1-M36
Modality of access	API and Web GUI based access (M1-M36) Additional terms: https://confluence.egi.eu/display/IMPAIP/Acceptable+Use+Policy
Support offered	Support is offered via the EGI Helpdesk. Detailed documentation about service, APIs, user guides, tutorials, etc. available.

	https://confluence.egi.eu/display/IMPAIP/User+guide#Userguide-Gettingaccess
Operational since	2020
User definition	Single researchers, collaborations of any size, citizen scientists

2.1.1 Metrics

Metric name	Baseline	Define how measurement is done	M13-M18	M19-M24
Number of AI models developed	30	logs	12	14
Number of AI models trained	500	logs	12	14
ML training cycles (CPU+GPU-hours)	4.000.000	logs	166,296	509,435
Number of countries reach	10	logs	12	12
Names of countries reach	SP, PT, FR, US, DE, UK, SK, CZ, CH, AU	logs	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT

2.1.2 Assessment

The numbers reported for the iMagine AI platform refer to the specific usage of the system for developing AI-based image models and tools for aquatic science.

As it can be seen, the platform usage has increased during the second project year, an expected fact since it is when use cases have started the training at scale of their AI based applications. Moreover, more resources are available for users, as all cloud sites have been integrated into the platform.

2.2 iMagine - AI Applications as a Service

Description	The iMagine AI Applications as a Service allows the transitioning of developed and trained AI/ML models into online services, following a serverless architecture.This installation allows the deployment of the AI models as an application to be offered to end users (i.e. not the application developers in the project, but for researchers outside), making it possible to build imaging data tools as production services. With the serverless approach the service can exploit the full potential of this computing model (i.e. function composition, event-based processing). Served models will exploit the DEEPaaS API to expose the underlying functionality.
Task	T4.2
URL	Not yet deployed. Will be available under https://services.imagine-ai.eu or similar location
Service Category	Infrastructure service
Service Catalogue	https://marketplace.eosc-portal.eu/services/imaging-ai-platform-for-aquatic-science
Providers	CSIC, LIP, UPV, KIT, IISAS
Location	Spain, Slovakia, Germany
Duration	36 months
Modality of access	API and Web GUI based access (M1-M36) Additional terms: https://confluence.egi.eu/display/IMPAIP/Acceptable+Use+Policy
Support offered	Support is offered via the EGI Helpdesk. Detailed documentation about service, APIs, user guides, tutorials, etc. available.
Operational since	2020
User definition	Single researchers, collaborations of any size, citizen scientists

2.2.1 Metrics

Metric name	Baseline	Define how measurement is done	M13-M18	M19-M24
ML application usage cycles, measured in CPU/GPU hours	0	Logs	1	33
Number of AI applications hosted via the iMagine platform	15	Logs	2	3
Number of countries reached over last year	0	Logs	3	4
Names of the countries reached over last year	N/A	Logs	ES, PT, CZ, IE	ES, PT, CZ, IE

2.2.2 Assessment

The numbers are as expected. Although they may seem low, these numbers reflect only the real resources utilization by the applications, as the serverless approach allows the platform to scale to zero (i.e. without resource utilization). Moreover, deployment as applications started around M20 for three use cases, and it is expected that these numbers increase during this last project year, when more models are promoted as application services.

2.3 IFCA-CSIC Scientific Cloud - CPU

Description	A federated compute environment based on the EGI Cloud Compute services, with multiple IaaS providers that offer compute resources (CPUs and GPUs) as VMs, alongside associated storage (Block/Object storage) for storing and accessing datasets. The service is suitable for hosting and processing of large datasets in a scalable way and is used as the major element of the EOSC Compute Platform.
Task	T4.3
URL	https://www.egi.eu/services/cloud-compute/
Service Category	Infrastructure service
Service Catalogue	https://www.egi.eu/services/cloud-compute/
Location	Spain
Duration	36 months
Modality of access	Services are free at the point of use. Access to the service require registration as an EGI user on Check-in and enrolment into a Virtual Organisation for authorisation
Support offered	Technical support is provided via the helpdesk central support team, and by the support team at the installation. EGI provides central documentation, trainings, webinars and hands-on sessions during conferences and events.
Operational since	2012
User definition	Single researchers, small communities, large collaborations

2.3.1 Metrics

Metric name	Baseline	Define how measurement is done	M13-M18	M19-M24
CPU node/hours served over the period	3M	Collected from local accounting	674,850	1,247,363
Names of the countries reached over the period	ES, PT, FR, UK, IT, GE, BE, SK, PL	Collected from local AAI system	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT
Number of countries reached over the period	9	Collected from local AAI system	12	12
Number of users	200	Collected from local AAI system	55	64

2.3.2 Assessment

The resources delivered by this installation include the following two different utilizations:

- Deployment of the control-plane of the iMagine platform (e.g. API, dashboard, MLOps services, etc.), quality assurance components (i.e. Jenkins), test platform (i.e. where preview functionalities are thoroughly tested before being rolled out) and storage services (e.g. NextCloud).
- Deployment of the platform nodes used to deliver part of the computing power for the iMagine AI platform and AI as a Service (catch-all instance) installations.

As it can be seen from the given numbers, there has been an increase in the usage in the second project year, due to the inclusion of new resources both for the application serving and development.

2.4 IFCA-CSIC Scientific Cloud - GPU

Description	A federated compute environment based on the EGI Cloud Compute services, with multiple IaaS providers that offer compute resources (CPUs and GPUs) as VMs, alongside associated storage (Block/Object storage) for storing and accessing datasets. The service is suitable for hosting and processing of large datasets in a scalable way and is used as the major element of the EOSC Compute Platform.
Task	T4.3
URL	https://www.egi.eu/services/cloud-compute/
Service Category	Infrastructure service
Service Catalogue	https://www.egi.eu/services/cloud-compute/
Location	Spain
Duration	36 months
Modality of access	Services are free at the point of use. Access to the service require registration as an EGI user on Check-in and enrolment into a Virtual Organisation for authorisation
Support offered	Technical support is provided via the helpdesk central support team, and by the support team at the installation. EGI provides central documentation, trainings, webinars and hands-on sessions during conferences and events.
Operational since	2012
User definition	Single researchers, small communities, large collaborations

2.4.1 Metrics

Metric name	Baseline	Define how measurement is done	M13-M18	M19-M24
GPU node/hours served over the period	1M	Collected from local accounting	14,808	57,316.00
Names of the countries reached over the period	ES, PT, FR, UK, IT, GE, BE, SK, PL	Collected from local AAI system	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT
Number of countries reached over the period	9	Collected from local AAI system	12	12
Number of users	200	Collected from local AAI system	55	64

2.4.2 Assessment

These resources are being used solely by the iMagine AI platform to deliver computing power to the use cases to develop their AI models. As it can be seen, there has been a substantial increase in the usage, due to two different facts:

- During PY1 the platform was affected by the transition to the new underlying software and resources (hence the low usage in the 1st period).
- During PY2 more use cases have started to consume GPU resources in order to be able to perform training of the models at scale.

The overall numbers are aligned with the expectations.

2.5 IFCA-CSIC Scientific Cloud - Storage

Description	A federated compute environment based on the EGI Cloud Compute services, with multiple IaaS providers that offer compute resources (CPUs and GPUs) as VMs, alongside associated storage (Block/Object storage) for storing and accessing datasets. The service is suitable for hosting and processing of large datasets in a scalable way and is used as the major element of the EOSC Compute Platform.
Task	T4.3
URL	https://www.egi.eu/services/cloud-compute/
Service Category	Infrastructure service
Service Catalogue	https://www.egi.eu/services/cloud-compute/
Location	Spain
Duration	36 months
Modality of access	Services are free at the point of use. Access to the service require registration as an EGI user on Check-in and enrolment into a Virtual Organisation for authorisation
Support offered	Technical support is provided via the helpdesk central support team, and by the support team at the installation. EGI provides central documentation, trainings, webinars and hands-on sessions during conferences and events.
Operational since	2012
User definition	Single researchers, small communities, large collaborations

2.5.1 Metrics

Metric name	Baseline	Define how measurement is done	M13-M18	M19-M24
Names of the countries reached over the period	ES, PT, FR, UK, IT, GE, BE, SK, PL	Collected from local AAI system	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT
Number of countries reached over the period	9	Collected from local AAI system	12	12
Number of users	> 200	Collected from local AAI system	55	64
Storage served over the period	1 PB	Collected from local accounting	30	400

2.5.2 Assessment

The numbers indicate the storage used through the Nextcloud Cloud storage deployed for the iMagine AI platform. As it can be seen, the storage has been increased during PY2 due to the fact to the migration of a new NextCloud storage, and also due to the increase in storage utilization and requirement from the use cases.

2.6 INCD - CPU

Description	A federated compute environment based on the EGI Cloud Compute services, with multiple IaaS providers that offer compute resources (CPUs and GPGPUs) as VMs, alongside associated storage (Block/Object storage) for storing and accessing datasets. The service is suitable for hosting and processing of large datasets in a scalable way and is used as the major element of the EOSC Compute Platform.
Task	T4.3

URL	https://www.egi.eu/services/cloud-compute/
Service Category	Infrastructure service
Service Catalogue	https://www.egi.eu/services/cloud-compute/
Location	Portugal
Duration	36 months
Modality of access	Modality of access (Duration): API and Web GUI based access (M01-M36)
Support offered	Helpdesk, support for deployment and usage of ML applications
Operational since	2018
User definition	Mostly user communities both big and small that correspond to openstack tenants

2.6.1 Metrics

Metric name	Baseline	Define how measurement is done	M13-M18	M19-M24
CPU/hours served over the period	3,900,000	openstack accounting	408,513	1,143,754
Names of the countries reached over the period	ES, PT	country of tenant email	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT
Number of countries reached over the period	2	country of tenant email	12	12
Number of users	50	openstack tenant	55	55

2.6.2 Assessment

INCD Cloud was initially used for testing and integration of the iMagine AI platform over the second half of the PY1, being integrated into the platform in PY2. Resources are being utilized for:

- Deployment of the platform nodes used to deliver part of the computing power for the iMagine AI platform and AI as a Service installations.
- Deployment of part of the control-plane for the iMagine services (i.e. iMagine container registry).
- Testing of new components and functionalities (e.g. MLOps, FAIR-EVA) before rolling them out into production.

The overall numbers are aligned with the expectations.

2.7 INCD - GPU

Description	A federated compute environment based on the EGI Cloud Compute services, with multiple IaaS providers that offer compute resources (CPUs and GPGPUs) as VMs, alongside associated storage (Block/Object storage) for storing and accessing datasets. The service is suitable for hosting and processing of large datasets in a scalable way and is used as the major element of the EOSC Compute Platform.
Task	T4.3
URL	https://www.egi.eu/services/cloud-compute/
Service Category	Infrastructure service
Service Catalogue	https://www.egi.eu/services/cloud-compute/
Location	Portugal
Duration	36 months

Modality of access	Modality of access (Duration): API and Web GUI based access (M01-M36)
Support offered	Helpdesk, support for deployment and usage of ML applications
Operational since	2018
User definition	Mostly user communities both big and small that correspond to openstack tenants

2.7.1 Metrics

Metric name	Baseline	Define how measurement is done	M13-M18	M19-M24
GPU node/hours served over the period	> 15000	openstack accounting	0	1752
Names of the countries reached over the period	PT, ES	country of tenant email	/	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT
Number of countries reached over the period	2	country of tenant email	0	12
Number of users	50	openstack tenant	0	64

2.7.2 Assessment

The INCD GPU resources are being used solely by the iMagine AI platform to deliver computing power to the use cases to develop their AI models. The utilization increased during PY2, since use cases have started to consume more resources.

2.8 INCD - Storage

Description	A federated compute environment based on the EGI Cloud Compute services, with multiple IaaS providers that offer compute resources (CPUs and GPGPUs) as VMs, alongside associated storage (Block/Object storage) for storing and accessing datasets. The service is suitable for hosting and processing of large datasets in a scalable way and is used as the major element of the EOSC Compute Platform.
Task	T4.3
URL	https://www.egi.eu/services/cloud-compute/
Service Category	Infrastructure service
Service Catalogue	https://www.egi.eu/services/cloud-compute/
Location	Portugal
Duration	36 months
Modality of access	Modality of access (Duration): API and Web GUI based access (M01-M36)
Support offered	Helpdesk, support for deployment and usage of ML applications
Operational since	2018
User definition	Mostly user communities both big and small that correspond to openstack tenants

2.8.1 Metrics

Metric name	Baseline	Define how measurement is done	M13-M18	M19-M24
Names of the countries reached over the period	ES, PT	country of tenant email	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT
Number of countries reached over the period	2	country of tenant email	12	12
Number of users	50	openstack tenant	55	64
TB/month served over the period	> 100	openstack accounting	8	23

2.8.2 Assessment

In the case of INCD Cloud, storage refers to the local storage consumed by the control-plane components requiring local and persistent storage (i.e. container registry). Numbers are aligned with expectations.

2.9 TR-FC1-ULAKBIM

Description	A federated compute environment based on the EGI Cloud Compute services, with multiple IaaS providers that offer compute resources (CPUs and GPUs) as VMs, alongside associated storage (Block/Object storage) for storing and accessing datasets. The service is suitable for hosting and processing of large datasets in a scalable way and is used as the major element of the EOSC Compute Platform.
Task	T4.3
URL	https://www.egi.eu/services/cloud-compute/

Service Category	Infrastructure service
Service Catalogue	https://www.egi.eu/services/cloud-compute/
Location	TURKEY
Duration	36 months
Modality of access	Modality of access (Duration): API and Web GUI based access (M01-M36)
Support offered	Technical support is provided via the helpdesk central support team, and by the support team at the installation. EGI provides central documentation, trainings, webinars and hands-on sessions during conferences and events.
Operational since	2014
User definition	Single researchers, small and big communtities

2.9.1 Metrics

Metric name	Baseline	Define how measurement is done	M13-M18	M19-M24
GPU node/hours served over the period with 2 CPU, 40 core and 4 GPU (V100)	29433.6	Local Accounting		19396
Names of the countries reached over the period	TR	Turkey, National HPC Centre		FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT
Number of countries reached over the period	1	Turkey, National HPC Centre		12
Number of unique users	269	Local Accounting		64

2.9.2 Assessment

The TR-FC1-ULAKBIM was not used in this period until PY2 due to the fact that the iMagine AI platform was transitioning from the old software stack to the new one and the deployment effort focused on this transition, transparent for the use cases. During PY2, TR-FC1-ULAKBIM also suffered from a major upgrade causing an interruption in the service delivery. However, once recovered, the installation has been steadily used, and numbers are aligned with the expectations.

2.10 WaltonCloud - CPU

Description	A federated compute environment based on the EGI Cloud Compute services, with multiple IaaS providers that offer compute resources (CPUs and GPUs) as VMs alongside associated storage (Block/Object storage) for storing and accessing datasets. The service is suitable for hosting and processing of large datasets in a scalable way.
Task	T4.3
URL	https://www.egi.eu/services/cloud-compute/
Service Category	Infrastructure service
Service Catalogue	https://www.egi.eu/services/cloud-compute/
Location	Waterford, Ireland
Duration	36 months
Modality of access	A federated compute environment based on the EGI Cloud Compute services, with multiple IaaS providers that offer compute resources (CPUs and GPGPUs) as VMs, alongside associated storage (Block/Object storage) for storing and accessing datasets.
Support offered	User onboarding

Operational since	2016
User definition	Single researchers, small and big communities

2.10.1 Metrics

Metric name	Baseline	Define how measurement is done	M13-M18	M19-M24
CPU/hours served over the period	11,586,723	OpenStack builtin statistics for reference period		60089
Names of the countries reached over the period	0	To be developed		FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT
Number of countries reached over the period	0	To be developed		12
Number of users	105	Checked against User logins over 12 month period		54

2.10.2 Assessment

The WaltonCloud installation started to be used during PY2 as the main resource provider for the iMagine custom deployments of the OSCAR component allowing it to deliver applications as a service tailored for the use cases (a catch-all instance is being executed on IFCA-CSIC Cloud installation, whereas use case specific installations are deployed in WaltonCloud). Numbers are as low, but it is expected that usage will increase during the last PY, as more use cases will deploy their own instances.

2.11 WaltonCloud - Storage

Description	A federated compute environment based on the EGI Cloud Compute services, with multiple IaaS providers that offer compute resources (CPUs and GPUs) as VMs alongside associated storage (Block/Object storage) for storing and accessing datasets. The service is suitable for hosting and processing of large datasets in a scalable way.
Task	T4.3
URL	https://www.egi.eu/services/cloud-compute/
Service Category	Infrastructure service
Service Catalogue	https://www.egi.eu/services/cloud-compute/
Location	Waterford, Ireland
Duration	36 months
Modality of access	A federated compute environment based on the EGI Cloud Compute services, with multiple IaaS providers that offer compute resources (CPUs and GPGPUs) as VMs, alongside associated storage (Block/Object storage) for storing and accessing datasets.
Support offered	User onboarding
Operational since	2016
User definition	Single researchers, small and big communities

2.11.1 Metrics

Metric name	Baseline	Define how measurement is done	M13-M18	M19-M24
Names of the countries reached over the period	0	To be developed		0

Number of countries reached over the period	0	To be developed		0
Number of users	105	Checked against User logins over 12 month period		0
TB/month served over the period	1188	OpenStack builtin statistics for reference period		0

2.11.2 Assessment

The WaltonCloud storage has not yet been used as it has not yet been necessary as the AI applications as a service deployments running in the WaltonCloud installation have not required it so far.

2.12 Analysis of compute services

The below table provides a summary of the compute-storage capacity that is budgeted in the project, and the consumption status with 6 month intervals:

UNIT	TOTAL BUDGETED IN THE PROJECT	USED BY M6 (Feb 2023)	USED BY M12 (Aug 2023)	USED BY M18 (Feb 24)	USED BY M24 (Aug 2024)
Storage (TB-month) - Accumulative	**1,500**	0	34	72	495
GPU-compute (GPU-h) - Accumulative	**132,000**	10,800	14,400	104,798	183,262
CPU-compute (CPU-h) -	**6,000,000**	94,656	662,547	1,745,910	4,197,116

Accumulative					

The following Fig 1-3 represent the consumption graphically:

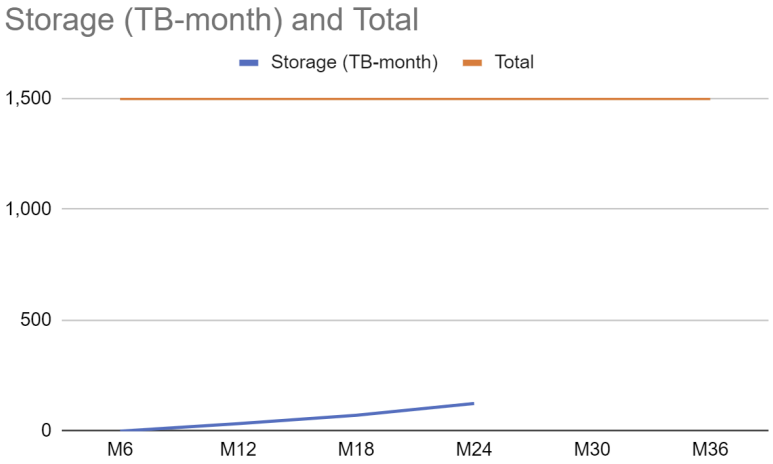

Fig 1. Storage usage status and trend

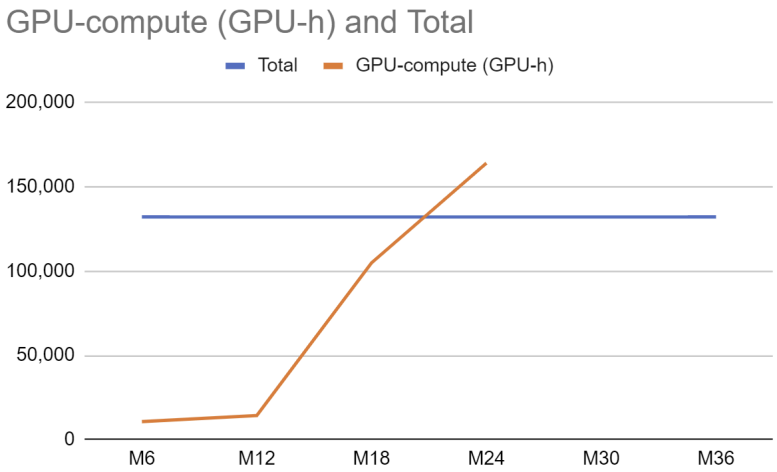

Fig 2. GPU usage status and trend

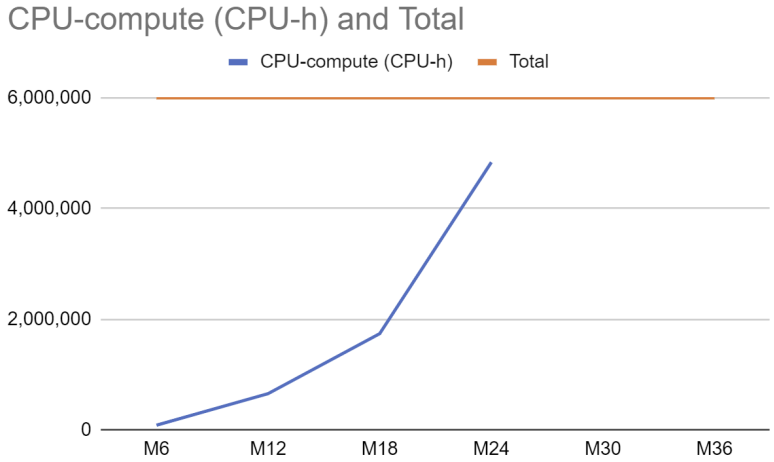

Fig 3. CPU usage status and trend

From this analysis we can state that:

2. The storage consumption (**Fig. 1**) is still modest and the project, due to the fact that so far the training datasets were the only real data ported to the infrastructure. End users data is expected to be larger, but we estimate there is still enough capacity to serve in the 3rd year.
3. The GPU (**Fig. 2**) consumption is higher than anticipated.
    a. The high usage is partly due to the fact that the use cases are in intensive model training and this requires GPUs. A significant contributor to the high number is the fact that GPUs are non-sharable resources, and when a GPU is assigned to a user the user needs to manually release it for the next user. Unfortunately this does not always happen, so idle, but non-released GPUs appear as 'GPU-hour consumed' in the statistics.
    b. The project is making an effort to educate users about the importance of releasing GPUs and we expect improvements on this in the 3rd year.

   c. The overall GPU usage is expected to be lower in the 3rd year anyway, because use case 1-5 will transition to service delivery from model training and service delivery requires CPUs instead of GPUs.
   d. Irrespective of all the above, the 3 GPU providers (LIP, TUBITAK, CSIC) are able to continue supporting the project in the 3rd year with its GPU demand.

4. The CPU usage is as expected, with more use during the model training phase than in the service delivery phase (inference requires a small amount of CPU compared to training.)