D4.5 Final Periodical Assessment of AI and Infrastructure Services

Abstract

The report provides third-year usage statistics and an assessment of all Artificial Intelligence platforms 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

D4.5 Final periodical assessment of AI and Infrastructure services			
Work Package 4			
Due date	31/08/2025	Actual delivery date:	20/10/2025
Nature of document	Report	Version	1.0
Dissemination level	Public		
Lead Partner	CSIC		
Authors	Álvaro López García (CSIC)		
Reviewers	Gergely Sipos (EGI Foundation) Fernando Aguilar (CSIC)		
Public link	https://zenodo.org/records/16943248		
Keywords	AI, Virtual Access, Infrastructure, Cloud, CPU, GPU, Storage		

Revision history

Issue	Date	Comments	Author/Reviewer
V 0.1	06/08/2025	First draft structure for ASB meeting	Andrea Anzanello (EGI)
V 0.2	29/09/2025	Advanced draft	Álvaro López García (CSIC)
V 0.3	10/10/2025	FInal draft for review	Álvaro López García (CSIC)
V 0.4	11/10/2025	Reviewed version	Fernando Aguilar (CSIC) Gergely Sipos (EGI)
V 1.0		Finalised and submitted version	Andrea Anzanello (EGI)

Copyright and licence info

Table of Contents

List of Figures

List of Tables

Executive summary

This report provides an assessment at M36 (project end) of the WP4 installations provided by the iMagine project under the Virtual Access (VA) mechanism. The primary focus is on the metrics collected for the installations during the 3rd year of the project, i.e. between September 2025 - August 2025, however summary is provided about the overall provisioning during the three project years.

The report covers service installations from two groups:

- iMagine AI platform, operated in two interwoven installations:
  - Task 4.1 - iMagine AI **development** service
  - Task 4.2 - iMagine AI **deployment** service
- Infrastructure Services (Task 4.3), where four providers (CSIC, LIP, Walton, Tubitak) provide compute and storage services underpinning the iMagine AI platform.

During PY3, the "iMagine AI development service" has been continuously used by the eight use cases of the project in order to further refine and fine-tune their models. Moreover, as in the previous reporting period, during the 3rd project year, external use cases have been supported via the iMagine open calls[1].

As a result, the GPU, CPU, and storage consumption of the infrastructure services increased, and the number of AI modules available in the iMagine marketplace has reached 18 (plus five additional generic AI modules). The installation served more than 60 users from 12 countries.

All the mature use cases have reached production status (cf. D5.3), and all the use cases supported by the project have concluded the service development phase (training and validating AI models) and transitioned to the AI "iMagine AI deployment service". This transition is supported by Continuous Integration and Continuous Delivery techniques, allowing a streamlined transition from development to deployment and operations.

The growing compute-storage demand was satisfied by the 4 cloud infrastructure providers from Task 4.3, as they are fully integrated into the iMagine platform. Three of the sites (CSIC-IFCA, INCD and TUBITAK) are powering the iMagine AI development service, delivering both GPU and CPU compute power, whereas the iMagine AI deployment service is backed by CSIC-IFCA (catch-all instance) and WaltonCloud - use-case specific instances, delivering only CPU.

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

During the whole project duration, the providers have delivered 11,113,730 CPU-hours; 365,878 GPU-hours and 831 TByte-month storage to the platform. Despite GPU and CPU consumptions were higher than originally anticipated at proposal time[2], T4.3 providers were able to allocate enough capacity to satisfy the demand and help the project meet and eventually exceed its original objectives.

[2] There were 6,000,000 CPU-hours, 132,000 GPU-hours and 1,500 TB-month capacity budgeted in T4.3.

1. Introduction

Virtual Access (VA) is a financial instrument that reimburses access providers for the costs of provisioning access. 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 two 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 3rd year of the project (Sep 2024 - Aug 2025).

In the 1st project year, WP4 worked on establishing the iMagine AI platform to serve the eight use cases within 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. In PY3, the last project period, WP4 has worked in supporting WP5 to deliver the use case applications (resulting in a significant usage increase in the application delivery services), together with continuous support for model refinement and updates.

1.1 WP4 Installations

Within the iMagine project, six installations are part of the 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 the development and validation of AI models. The eight use cases used the service during PY1 and PY2. The usage is monitored with the following metrics:
    - ML training cycles are 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 the last year (users' location);
    - Number of countries reached over the 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 are measured in CPU/GPU hours;
    - Number of AI applications hosted via the iMagine platform;
    - Names of the countries reached over the last year (users' location);
    - Number of countries reached over the 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 were 132,000 GPU-hours, 6,000,000 CPU-hours and 1,500 TB-month capacity budgeted in WP4.

2. Analysis of compute services

The table below provides a summary of the compute-storage capacity that is budgeted in the project for the cloud installations (i.e., T4.3), and the consumption status with 6-month intervals:

Table 1 Cumulative resource consumption during project lifetime

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)	USED BY M30 (Feb 2025)	USED BY M36 (Aug 2025)
Storage (TB-month) - Accumulative	**1,500**	0	34	72	495	782	843
GPU-compute (GPU-h) - Accumulative	**132,000**	10,800	14,400	104,798	183,262	274,486	365,878
CPU-compute (CPU-h) - Accumulative	**6,000,000**	94,656	662,547	1,745,910	4,197,116	7,460,311	11,113,730

The following Figures represent the consumption graphically:

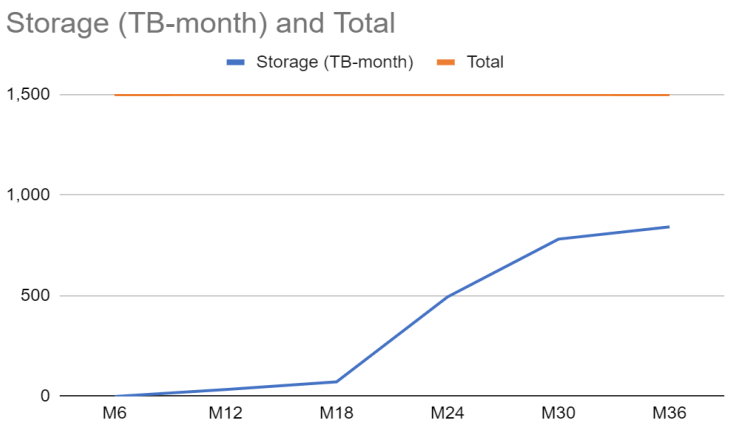

Figure 1 Storage usage status and trend

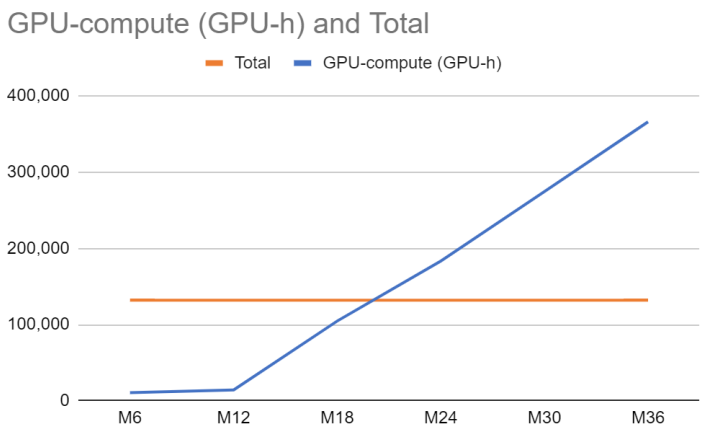

Figure 2 GPU usage status and trend

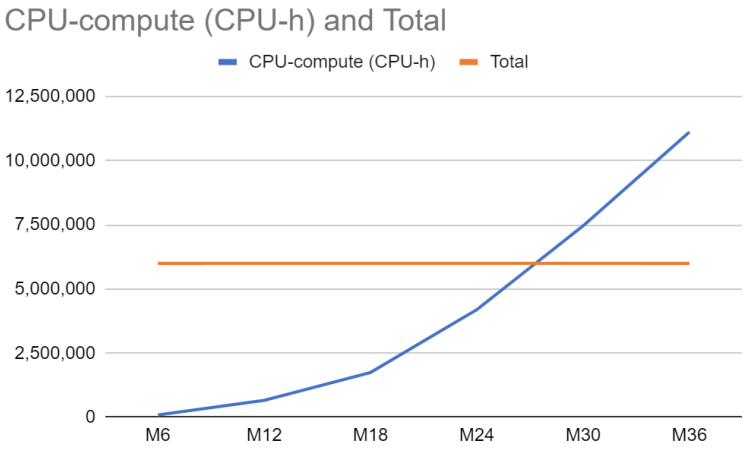

Figure 3 CPU usage status and trend

From an analysis of **Table 1** and **Figure 1** to **Figure 3** we can state that:

- The demand for CPU (**Figure 3**) and GPU (**Figure 2**) capacity was underestimated during the proposal elaboration time. The high GPU usage is partly due to intensive model training, which requires GPUs. A significant contributor to the high number is that GPUs are non-sharable resources; once a GPU is assigned to a user, the user must manually release it for the following user. Unfortunately, this does not always happen, so idle, but non-released GPUs appear as 'GPU-hour consumed' in the statistics. We have educated users about fair sharing and releasing GPUs when they are not in use, and we have implemented batch jobs, allowing users to run long-running tasks non-interactively, thus utilising the resources only for the exact time when computations are running. However, this functionality was only available in the last project quarter; therefore, it was insufficient to see the real impact.
- Irrespective of the above, the 3 GPU providers (LIP, TUBITAK, CSIC) were able to continue supporting the project in the 3rd year despite sustained GPU demand.
- The CPU usage is also high because additional services have been introduced on the platform, such as try-me endpoints, image annotation tools (CVAT), MLOps tracking services (MLFlow), among others. Moreover, it should be taken into account that we are operating two different sets of platform services (i.e., the iMagine production instance at **https://dashboard.cloud.imagine-ai.eu/** and the iMagine preview instance at **https://dashboard.dev.imagine-ai.eu/**) to deliver early features to the users without interrupting the production services, therefore, additional services (consuming CPU resources) have been deployed.
- On the contrary, storage consumption (**Figure 1**) remains below the expected threshold. This is due to the fact that the training datasets were not as large as expected, and that data is only copied into the platform when it needs to be processed.

3. Installations

3.1 iMagine - AI Application Development Service

Description	The iMagine AI Application Development Service enables Artificial Intelligence developers to prototype, build, and train AI applications, leveraging 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 monitoring the status of training directly from the training Dashboard. The development environment is based on JupyterLab instances, where users have access to significant 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://www.imagine-ai.eu/services/imagine-ai-platform
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

3.1.1 Metrics

Metric name	Baseline	Define how measurement is done	M1-M12	M13-M24	M25-M36	TOT
Number of AI models developed	30	logs	10	16	14	**40**
Number of AI models trained	500	logs	10	16	14	**40**
ML training cycles (CPU+GPU-hours)	4.000.000	logs	554,136	675,731	1,676,636	**2,801,047**
Number of countries reach	10	logs	12	0	0	**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	FR, BE, ES, IE, DE, SK, NL, USA, UK, DK, TR, PT	

3.1.2 Assessment

The numbers reported for the iMagine AI platform refer to the system's specific use in developing AI-based image models and tools for aquatic science. As can be seen, the platform usage has increased over the last project year, as use cases

continued to work to fine-tune and further refine their AI-based applications. Resources have been continuously available for users, with no significant disruptions.

3.2 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, using 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	Available under https://inference.cloud.imagine-ai.eu/.
Service Category	Infrastructure service
Service Catalogue	https://www.imagine-ai.eu/services/image-analysis-services-for-aquatic-sciences
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

3.2.1 Metrics

Metric name	Baseline	Define how measurement is done	M1-M12	M13-M24	M25-M36	**TOT**
ML application usage cycles, measured in CPU/GPU hours	0	Logs	0	34	24,810	**24,843**
Number of AI applications hosted via the iMagine platform	15	Logs	0	5	111	**116**
Number of countries reached over last year	0	Logs	0	4	1	**5**
Names of the countries reached over last year	N/A	Logs	/	ES, PT, CZ, IE	ES, CZ, BE, DE, FR	

3.2.2 Assessment

The numbers are as expected. Although they may seem low, these numbers reflect only the real resource utilisation by the applications, as the serverless approach allows the platform to scale to zero (i.e. without resource usage), with the control plane (i.e., the OSCAR services) being accounted as CPU usage in the different cloud installations (T4.3).

3.3 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 large datasets in a scalable way and is a significant 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 requires 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

3.3.1 Metrics

Metric name	Baseline	Define how measurement is done	M1-M12	M13-M24	M25-M36	**TOT**
CPU node/hours served over the period	3M	Collected from local accounting	488,544	1,922,213	4,424,245	**8,381,165**
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	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	7	5	0	**12**
Number of users	200	Collected from local AAI system	36	80	60	**140**

3.3.2 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 the cumulative numbers show, there has been an increase in usage in the last project phase, due to the inclusion of new resources both for the application serving and development.

3.4 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 large datasets in a scalable way and is a significant 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 requires 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

3.4.1 Metrics

Metric name	Baseline	Define how measurement is done	M1-M12	M13-M24	M25-M36	**TOT**
GPU node/hours served over the period	1M	Collected from local accounting	14,400	72,124	121,296	**207,820**
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	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	7	5	0	**12**
Number of users	200	Collected from local AAI system	36	83	21	**140**

3.4.2 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 the following 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.
- During PY3, use cases have continued exploiting GPU resources in order to fine-tune and to continue improving their AI models.

The overall numbers align with expectations.

3.5 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 large datasets in a scalable way and is a significant 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 requires 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

3.5.1 Metrics

Metric name	Baseline	Define how measurement is done	M1-M12	M13-M24	M25-M36	**TOT**
Names of the countries reached over the period	ES, PT, FR, UK, IT, GE, BE, SK, PL	Collected from local AAI system	7	5	0	**12**
Number of countries reached over the period	9	Collected from local AAI system	FR, BE, ES, IE, DE, SK, NL	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 users	> 200	Collected from local AAI system	36	83	21	**140**

Storage served over the period	1 PB	Collected from local accounting	30	430	295	**755**

3.5.2 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 migration into a new NextCloud storage, and during PY3 due to the increase in storage utilisation and requirement from the use cases.

3.6 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 large datasets in a scalable way and is a significant 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

3.6.1 Metrics

Metric name	Baseline	Define how measurement is done	M1-M12	M13-M24	M25-M36	**TOT**
CPU/hours served over the period	3,900,000	openstack accounting	174,003	1,552,267	2,270,601	**3,996,871**
Names of the countries reached over the period	ES, PT	country of tenant email	FR, BE, ES, IE, DE, SK, NL	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	7	5	0	**12**
Number of users	50	openstack tenant	16	39	0	**55**

3.6.2 Assessment

INCD Cloud was initially used for testing and integration of the iMagine AI platform in the second half of PY1, with integration completed in PY2. Resources are being utilised 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 align with expectations.

3.7 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 large datasets in a scalable way and is a significant 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

3.7.1 Metrics

Metric name	Baseline	Define how measurement is done	M1-M12	M13-M24	M25-M36	**TOT**
GPU node/hours served over the period	> 15000	openstack accounting	0	1,752	8,760	**10,512**
Names of the countries reached over the period	PT, ES	country of tenant email	FR, BE, ES, IE, DE, SK, NL	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	7	5	0	**12**
Number of users	50	openstack tenant	16	39	0	**55**

3.7.2 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 utilisation increased during PY3, since use cases have consumed more resources to fine-tune and further develop their models.

3.8 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 large datasets in a scalable way and is a significant 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

3.8.1 Metrics

Metric name	Baseline	Define how measurement is done	M1-M12	M13-M24	M25-M36	**TOT**
Names of the countries reached over the period	ES, PT	country of tenant email	FR, BE, ES, IE, DE, SK, NL	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	7	5	0	**12**
Number of users	50	openstack tenant	16	39	0	**55**
TB/month served over the period	> 100	openstack accounting	8	31	41	**80**

3.8.2 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.

3.9 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 large datasets in a scalable way and is a significant 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 communities

3.9.1 Metrics

Metric name	Baseline	Define how measurement is done	M1-M12	M13-M24	M25-M36	**TOT**
GPU node/hours served over the period with 2 CPU, 40 core and 4 GPU (V100)	29433.6	Local Accounting	0	94,986	52,560	**147,546**
Names of the countries reached over the period	TR	Turkey, National HPC Centre	/	FR, BE, ES, IE, DE, SK, NL, USA,	CN, DE, IT, RO, ES, SE, TR, GB,	

				UK, DK, TR, PT	US	
Number of countries reached over the period	1	Turkey, National HPC Centre		12	0	**12**
Number of unique users	269	Local Accounting	0	64	19	**83**

3.9.2 Assessment

The TR-FC1-ULAKBIM started to be used in PY2 due to the fact that the iMagine AI platform was transitioning (in PY1) 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.

3.10 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 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

3.10.1 Metrics

Metric name	Baseline	Define how measurement is done	M1-M12	M13-M24	M25-M36	**TOT**
CPU/hours served over the period	11,586,723	OpenStack builtin statistics for reference period	0	60,089	221,768	**281,857**
Names of the countries reached over the period	0	To be developed	/	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	0	To be developed	0	12	0	**12**
Number of users	105	Checked against User logins over 12 month period	0	54		**54**

3.10.2 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 (i.e. for use cases not self-deploying their own solution, or for testing purposes), whereas use case specific installations are deployed in WaltonCloud (hence the increase in the CPU usage).

3.11 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 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

3.11.1 Metrics

Metric name	Baseline	Define how measurement is done	M1-M12	M13-M24	M25-M36	**TOT**
Names of the countries reached over the period	0	To be developed	/	ES, PT, CZ, IE	ES, CZ, BE, DE, FR	
Number of countries reached over the period	0	To be developed	0	4	1	**5**
Number of users	105	Checked against User logins over 12 month period	0	5	111	**116**
TB/month served over the period	1188	OpenStack builtin statistics for reference period	0	0	12	**12**

3.11.2 Assessment

The WaltonCloud storage has been used in the 3rd year for the inference runs carried out by the OSCAR component of the "AI applications as a service" installation.

4. Conclusion

This document provides the final periodical assessment of AI and infrastructure services, including detailed insights into resource utilisation throughout the project.

While the initial budget for storage, GPU, and CPU capacity was established at proposal time and delivered via T4.3, the actual consumption trends for GPU and CPU have exceeded the estimated projections.
The GPU excess was due to (1) the intensive model training by the scientific use cases from the consortium and beyond (onboarded through the open call) and (2) due to the suboptimal use of GPUs by the users (manual resource release).
The CPU usage has gone beyond expectations, mainly due to the fact of an increased service offer for the user communities, with additional services providing added value that were not considered at proposal time.
The four compute providers were able to allocate additional compute resources for the consortium, and therefore the extra CPU and GPU demand did not cause a limitation in achieving and eventually exceeding the initial project ambition.
Moving forward, we consider that it is essential to implement stricter GPU usage monitoring, encourage fair resource-sharing practices, and explore batch job solutions for GPUs for optimised resource utilisation. These measures will ensure economically sustainable and efficient use of computational resources, aligning with the long-term goals and operational needs of aquatic sciences.