

DELIVERABLE D3.1.1

ARCHITECTURE

DESIGN OF THE

TIER-1 SYSTEM

VERSION 1

TeRABIT

AUGUST 2023

Mirko Cestari, Giuseppa Muscianisi (CINECA)
Giorgio Bolzon, Stefano Salon (OGS)

Eccetto dove diversamente specificato, il contenuto di questo sito è
concesso in licenza con Attribuzione Creative Commons - 4.0 internazionale.



Executive Summary

The deliverable describes the architecture co-design process for the upgrade of PRACE-Italy Tier-1 system. Specifically, the document presents a synthesis of the users' requirements, which has been elaborated through the analysis of the responses to a public survey sent to PRACE-Italy users during the second quarter of 2023. The aim of the survey was i) to collect feedback on the use of Cineca cloud infrastructure and the users' expectations, and ii) to draw a plan to improve the services offered by the upgraded PRACE-Italy Tier-1 system, part of the TeRABIT roadmap. The co-design process has highlighted the need of extending the capacity and capabilities of the Tier-1 computational cloud services, specifically cloud computing (with a target of 200.000 vCPUs) and multi-protocol storage for interactive analysis and processing and mid-term repository (with a target of 40 PB).

The deliverable also includes some of the specifications enabling WP2, WP3 and WP4 infrastructure integration.

Table of Content

1. <i>CO-DESIGN PROCESS</i>	2
1.1. PRACE-Italy Tier-1 evolution survey results	2
1.1.2. Respondents' statistics.....	3
1.1.3. Cloud infrastructure users	4
1.1.4. Non-cloud infrastructure users	7
1.1.5. Conclusions.....	9
2. <i>INFRASTRUCTURE INTEGRATION</i>	10

List of Figures

Figure 1 - Answers to: "Are you already using Cineca cloud resources?"	3
Figure 2 - Scientific domains divided by "already users of cloud resources" (Yes) and "not users of cloud resources" (No)	3
Figure 3 - Cloud infrastructure was in line with expectations.....	4
Figure 4 - Easiness to use of the cloud infrastructure.	4
Figure 5 - Overall satisfaction with the cloud infrastructure.	5
Figure 6 - Importance of connection with HPC storage; (1= not important at all; 5 = very important).	6
Figure 7 - Importance of connection with HPC compute infrastructure; (1= not important at all; 5 = very important).	6
Figure 8 - Preference for future resources.	7
Figure 9 - # of requests for the different ADACloud flavours.	7
Figure 10 - Distribution of requests for additional permanent storage.	8
Figure 11 - # of requests for different cloud features.....	9
Figure 12: Targeted high-level architecture of Tier-1 system evolution.	10

List of Tables

Table 1 - Summary of the upgrade projection for Tier-1.	9
--	---

1. CO-DESIGN PROCESS

The main goal of WP3 is providing to the national scientific communities the demonstration of the integration of the PRACE-Italy research infrastructure into the TeRABIT multi-disciplinary network through a significant upgrade of the present PRACE-Italy Tier-1 system, Galileo100¹. Following a co-design approach, the new HPC system will provide users with the ability to interactively elaborate, process and visualize data, as well as the ability to drive non-standard HPC workloads (e.g., continuous processing of data coming from the edge-to-cloud-to HPC, as in digital twins). The heterogeneous data processing capabilities provided by the Tier-1 system will represent enabling technologies by allowing researchers to scale-up to Tier-0 system resources (provided by the ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing) thus tackling higher scientific challenges, fully exploiting the EuroHPC dimension. For this reason, a coordinated evolution roadmap of the computing technologies involved in the Tier-0 and Tier-1 architectures will be strategic, so that the latter bridges and complements the former.

The Tier-1 system will be based on industry grade computing racks, integrating state-of-the-art computing CPU- or GPU-based blades, I/O drives, data network switches, and ancillary equipment, and will result in an energy-efficient, scalable, and highly-dense solution capable to provide mid-range petascale system performance according to EuroHPC classification, and a peak performance in the order of 5 PFlops. The exact configuration of the Tier-1 system is the subject of an open procurement procedure aimed at maximizing the solution in terms of performance of the computing and storage services provided. The co-design process has been built through a public survey organized by Cineca, distributed to all Cineca and PRACE-Italy users through the HPC-News newsletter mailing list², cloud project users, Cineca HPC staff. Section 1.1 presents the details of the surveys and the results.

1.1. PRACE-Italy Tier-1 evolution survey results

To understand the requirements of the users with respect to the infrastructure resources and related services, the survey was structured with questions to understand:

- User scientific domain and use case.
- For users already making use of the Cineca ADA Cloud infrastructure³ - part of the PRACE-Italy Tier-1 system -, the current level of satisfaction with the service and the provided resources, and their needs for the future.

¹ <https://www.hpc.cineca.it/hardware/galileo100>

² <https://www.hpc.Cineca.it/content/stay-tuned>

³ <https://wiki.u-gov.it/confluence/pages/viewpage.action?pageId=436932577>

- For respondents not already using ADA Cloud, the expectation in terms of cloud resources and services.

1.1.2. Respondents' statistics

A total of 56 replies were collected with 27 (48%) of them from already ADA Cloud users (Figure 1).

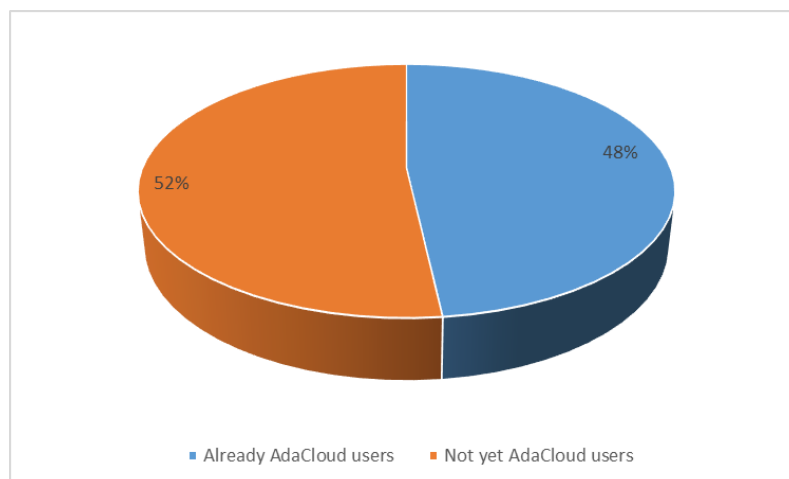


Figure 1 - Answers to: "Are you already using Cineca cloud resources?"

Looking at the scientific domains, 8 domains were covered by the replies, 5 of which with replies from both users of cloud resources and potentially new users (Figure 2). The "Earth and Climate Science" and "Computational Engineering" domains had the first two largest number of replies, 14 and 12 respectively; "CFD" domain was the domain with just one reply.

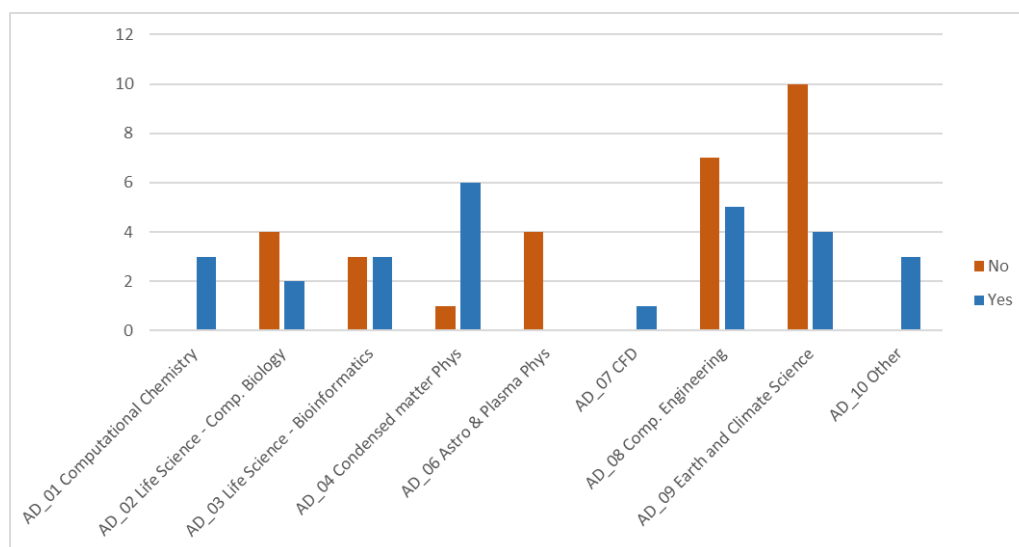


Figure 2 – Scientific domains divided by "already users of cloud resources" (Yes) and "not users of cloud resources" (No)

The users are making or would make use of cloud resources for different reasons: testing/exploration, become a direct user, make available the service to my community/research group, integrate the service with my (or my community) research workflow, without a clear preference.

1.1.3. Cloud infrastructure users

As mentioned before, 27 respondents are already using the Cineca ADA Cloud infrastructure: 77% of the users replied that the cloud infrastructure is in line with their expectations (Figure 3) and 55% of the respondents found it easy to use (Figure 4).

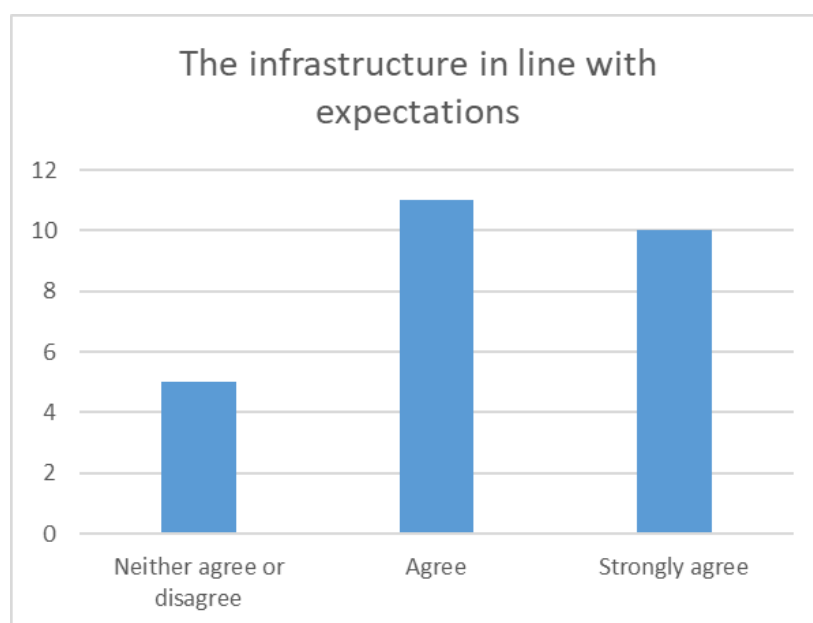


Figure 3 – Cloud infrastructure was in line with expectations.

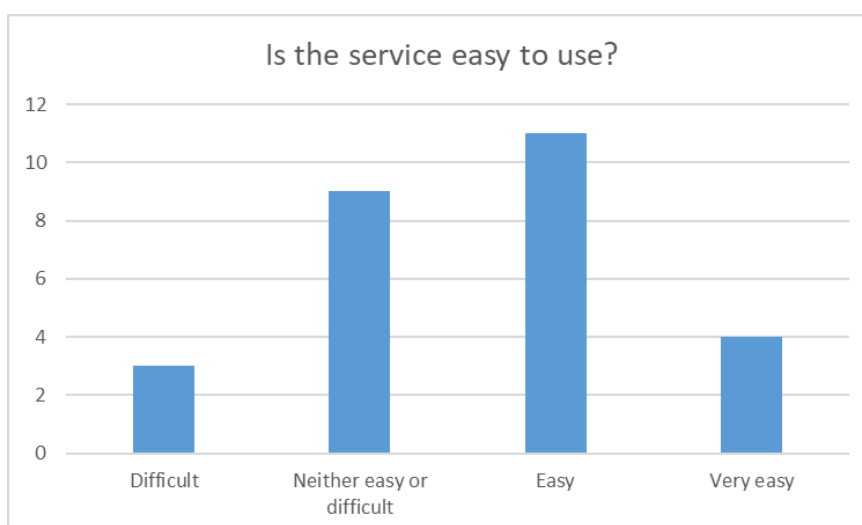


Figure 4 – Easiness to use of the cloud infrastructure.

Most of the difficulties in the use of the infrastructure are reported to be associated with unclear sections of the user documentation; however, only 2 respondents find user documentation and user support “poor” with 67% of them finding it good or above. In particular, a clear interest was reported towards having more tutorials and/or workshops, and best practices for cloud resources management. These and other training request will be addressed in WP5 as foreseen by the project workplan.

In terms of resources, all respondents, except 2, consider the currently provided resources (vCPUs, RAM, storage, floating IPs) enough to achieve their current research objectives. However, 63% of the respondents expect to need more resources in the future. Of those, 4 mentioned that it is difficult to make an estimate on the precise quantity of resources that will be needed. The others would request in total about 966 vCPUs.

In general, 85% of the respondents are overall satisfied with the offered service (with an average score of 4.4 over 5), as reported in Figure 5.

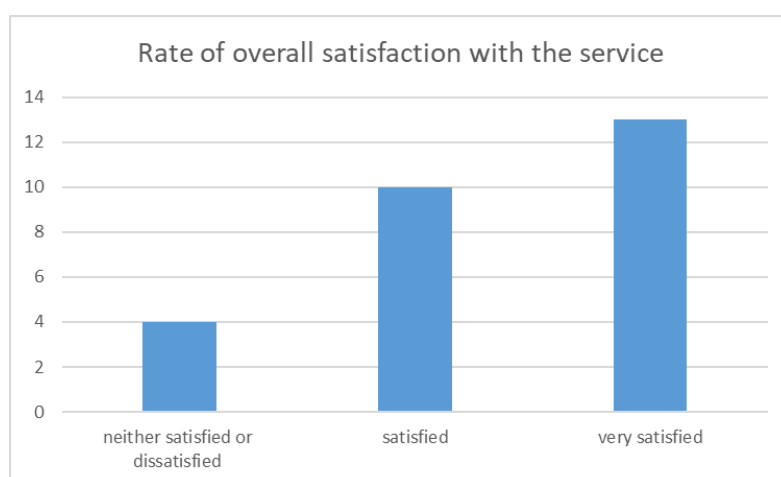


Figure 5 – Overall satisfaction with the cloud infrastructure.

Among the extra features that can be of use, the users mentioned the need for additional storage with the possibility of accessing it across Virtual Machines and HPC systems. The connection between the cloud and the HPC infrastructure has been in fact confirmed as very important by about 79% of the users both for the computational resources (Figure 6) and for the storage (Figure 7).

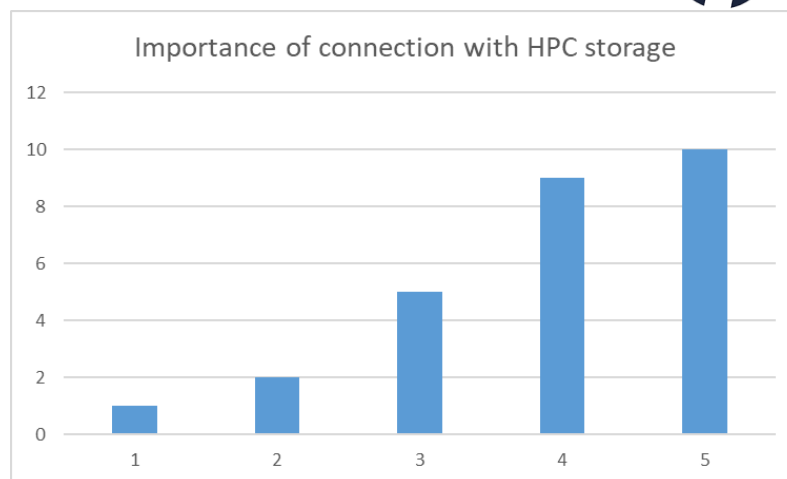


Figure 6 – Importance of connection with HPC storage; (1= not important at all; 5 = very important).

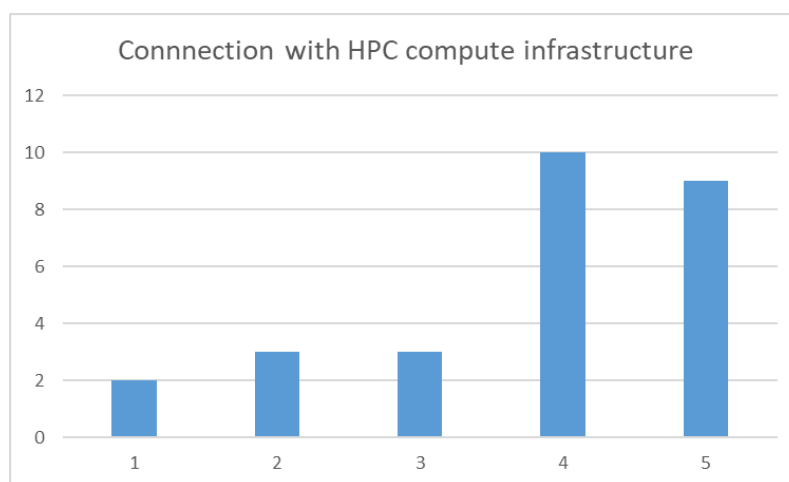


Figure 7 – Importance of connection with HPC compute infrastructure; (1= not important at all; 5 = very important).

Of the 27 ADA Cloud users, 19 are also using HPC resources and almost half of them would prefer to have more resources on the HPC side (Figure 8).

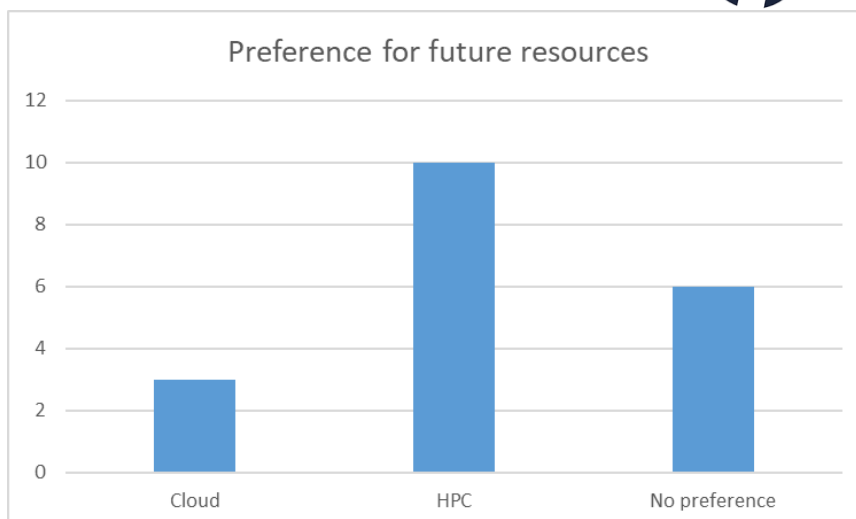


Figure 8 – Preference for future resources.

1.1.4. Non-cloud infrastructure users

A Total of 29 respondents are not already making use of the Cineca cloud infrastructure, and the survey asked them to quantify the amount of resources potentially needed.

Summing up all the requests, there is a need for a total of over 1000 vCPUs, almost 7 TB of RAM and 1 TB of local storage⁴ with the distribution shown in Figure 9.

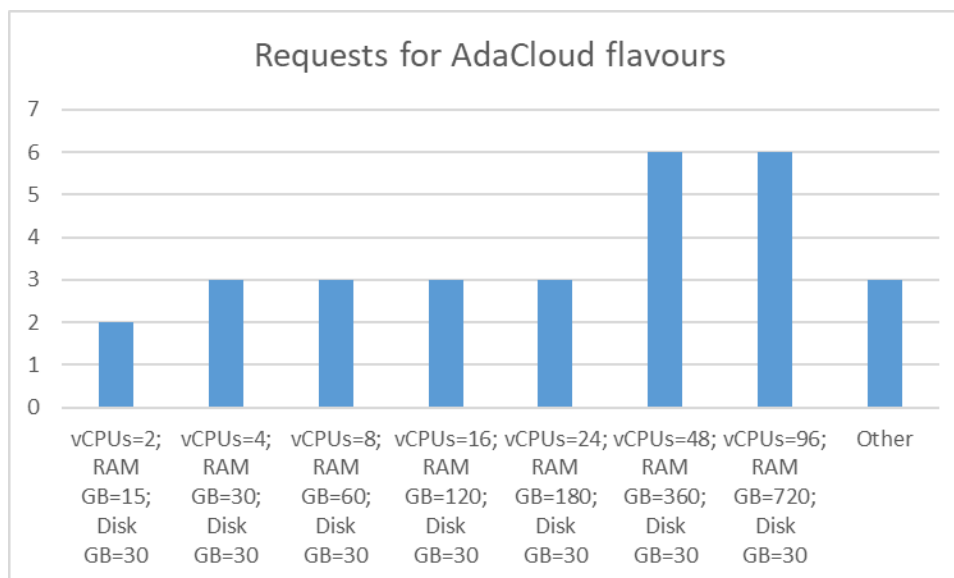


Figure 9 - # of requests for the different ADACloud flavours.

⁴ Typically this is used to host the operating system and user software stack

Moreover, the request for additional permanent storage⁵ reported from the respondents amounted to more than 90 TB (Figure 10).

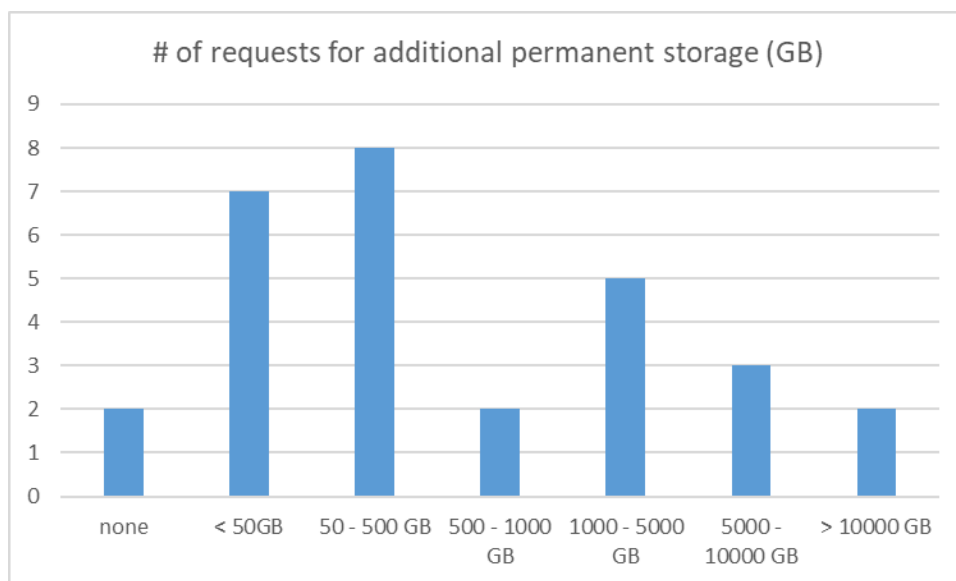


Figure 10 – Distribution of requests for additional permanent storage.

For Floating IPs⁶, there is a request for 399 fIPs with an average of about 13 fIPs per user.

In terms of requested features, the most needed one is access to GPUs together with the shared file system and the possibility to execute hybrid cloud-HPC workflows (Figure 11, please note that multiple replies were allowed for this question).

⁵ Permanent storage, also referred to as virtual machines volumes, typically host all the data to be processed and the outcomes.

⁶ Floating IPs are the *public* IPs available to access virtual machines. Each cloud project needs at least one public IP to be reached from the outside.

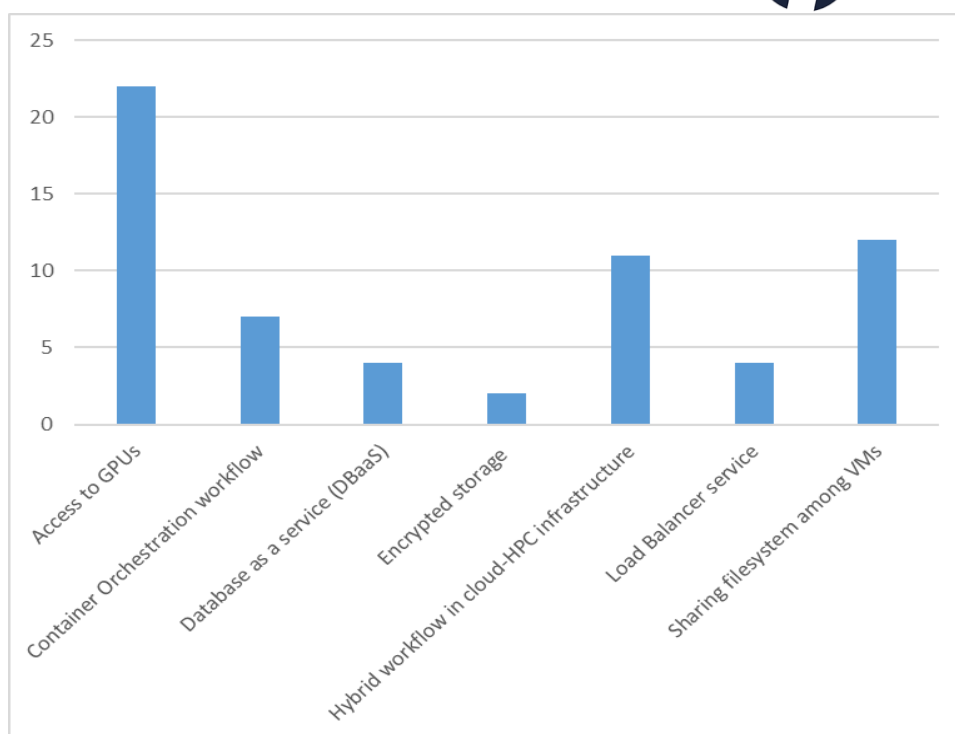


Figure 11 - # of requests for different cloud features.

1.1.5. Conclusions

Overall, considering both cloud and non-cloud users, we can highlight that in terms of resources the total request was around 2000 vCPUs, 7 TB of RAM and 91 TB of storage. If we consider that the total number of users of Tier-1 system is about two orders of magnitude larger than the survey respondents (i.e., around 4500, see D.1.2.4. Capacity Planning”, Section 2.2), we can project a 100x total request of the resources, quantified in 200.000 vCPUs, 700 TB of RAM, and a total storage of 9100 TB (see Table 1).

Table 1 - Summary of the upgrade projection for Tier-1.

	Survey results	Tier-1 upgrade projection	Notes
vCPUs	2000	200000	A target of 100000 cores is attainable. With multi-threading activated 200000 vCPUs can be provided
RAM	7 TB	700 TB	While 700 TB of RAM could be obtained, 400 TB should be enough for the user needs satisfying the common requirement of 2GB per vCPU and complying with an overall better energy efficiency
Storage	91 TB	9100 TB	At least an amount of 8 PB of storage for the cloud must be planned

In terms of features, access to GPUs was the most requested one in the survey and thus the new TeRABIT system will include GPUs nodes.

Considering the feedbacks from the current users, special attention will also be given to the users' documentation preparation together with tutorial and training activities.

Today, Galileo100 offers four main computational services: high-performance computing, cloud computing, interactive analysis and processing, multi-protocol (file and object) data repository. The co-design process has focused on understanding the user needs and projecting them onto the new system architecture, highlighting the need of extending the capacity and capabilities of three of these main Tier-1 computational services.

On the basis of this process, the Tier-1 procurement procedure aims to provide roughly 100.000 CPU cores for virtualization services, 8 PB of cloud dedicated storage and 30 PB of multi-protocol storage for cloud, interactive analysis and processing and mid-term repository. This huge number of state-of-the-art CPU cores⁷ can support low- to mid-parallelism workloads, and a migration from HPC to cloud for these use cases is foreseen. Anyway, it should be noted that the user community will still rely on Galileo100 to continue providing the current services until reaching its planned end-of-life in 2026. Therefore, the high-level architecture resulting from the process and findings described in this Deliverable is reported in Figure 12.

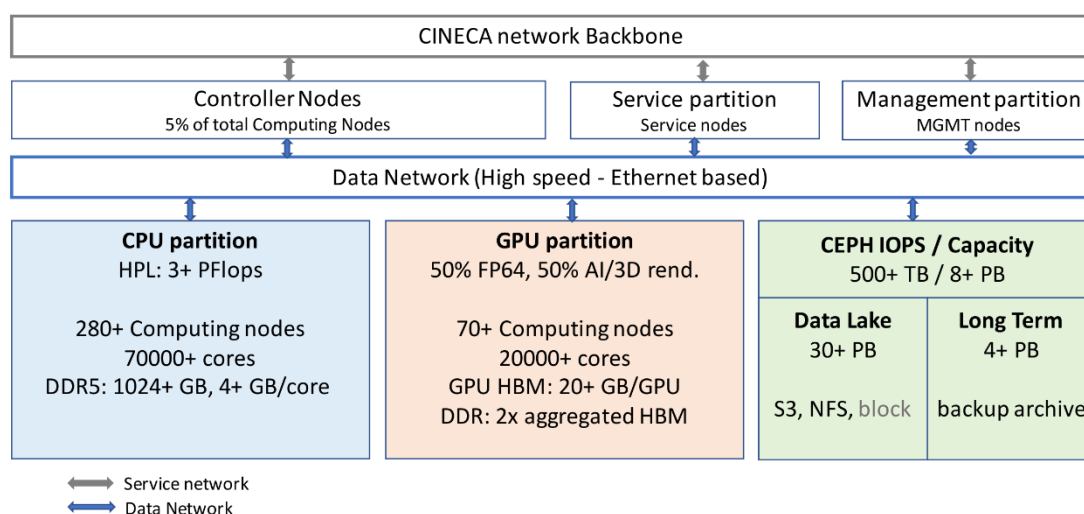


Figure 12: Targeted high-level architecture of Tier-1 system evolution.

2. INFRASTRUCTURE INTEGRATION

The upgraded PRACE-Italy Tier-1 system put in place in WP3 will become a natural computational backend for data coming from edge computing and HPC bubbles (developed in

⁷ The target is new high density CPU production process, such as AMD 5nm or Intel 3.

WP4), allowing researchers to scale their workloads, store processed sensors data, exploiting the fast interconnection provided within the GARR-X network (developed in WP2). We draft here some of the technical specifications that will allow the integration of the three research infrastructures.

Galileo100 and the upgraded Tier-1 system will be interconnected through the CINECA ethernet backbone based on 400 Gb/s switches, allowing high east-west data traffic, and interfaced to the whole GARR network via the point-of-presence BO02-Casalecchio with a bandwidth of 20 Gb/s. This interconnection allows to benefit from any Italian network improvement carried in WP2 of this project. Data can therefore be transferred at high rate to CINECA Tier-1 from institutions and systems at the edge – such as the HPC bubbles developed in WP4 – allowing high-performance, real-time, data processing and long-term data preservation.

The new powerful Cloud infrastructure will provide the required flexibility to be the natural interface layer between HPC Tier-2 systems, spread across Italy and close to data sources, and Tier-1/Tier-0 systems located in CINECA for scale-up or scale-out workloads. Even if the network and processing capability is world-class, the data amount can be such that a processing stage is required to perform a first data reduction. The Tier-2 systems are clearly best suited for this role. Moreover, a Tier-2-Tier-1-Tier-0 supercomputing system pyramid allows high resiliency for those workloads that requires it, providing disaster recovery capability and/or business continuity.