FENIX RESEARCH INFRASTRUCTURE

D4.6

Infrastructure at CINECA

Work package:	WP4 - Procurement, deployment and operation	
Author(s):	Mirko Cestari	CINECA
Reviewer #1	Javier Bartolomé / Michael Mucciardi	BSC
Reviewer #2		
Dissemination Level	Public	
Nature	Other	

Date	Author	Comments	Version	Status
17.05.2021	Mirko Cestari	Initial version	V0.3	Draft
02.07.2021	Javier Bartolomé / Michael Mucciardi / Dirk Pleiter	Internal review	V0.31	Draft
16.07.2021	Mirko Cestari	Updates	V0.4	Draft
23.08.2021	Mirko Cestari	Final revision	V0.6	Draft
31.08.2021	Valentina Armuzza	Final editorial updates	V1.0	Final

The ICEI project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 800858.

Executive Summary

This document describes the infrastructure components operated by CINECA for the Fenix research infrastructure as of August 2021.

Contents

Acronyms

AAI	Authentication and Authorization Infrastructure
ACD	Active Data Repositories
ACL	Access Control List
API	Application Programming Interface
ARD	Archival Data Repositories
BSC	Barcelona Supercomputing Center
CapEx	Capital Expenditure
CDP	Co-design Project
CEA	Commissariat à l'énergie atomique et aux énergies alternatives
CINECA	Consorzio Interuniversitario
CLI	Command Line Interface
CSCS	Centro Svizzero di Calcolo Scientifico
DL	Data Location Service
DM	Data Mover Service
DT	Data Transfer Service
FPA	Framework Partnership Agreement
FURMS	Fenix User and Resource Management Services
GoP	Group of Procurers
GUI	Graphical User Interface
HBP	Human Brain Project
HPAC	High Performance Analytics and Computing
HPC	High Performance Computing
HPDA	High Performance Data Analytics
HPST	High-Performance Storage Tier
IaaS	Infrastructure as a Service
IAC	Interactive Computing Services
ICCP	Interactive Computing Cloud Platform
ICEI	Interactive Computing E-Infrastructure for the Human Brain Project
ICN	Interactive Computing Node
IdP	Identity Provider
IPR	Intellectual Property Rights
JP	Joint Platform
JSC	Jülich Supercomputing Centre
LCST	Large-Capacity Storage Tier
MS	Monitoring Services

NDA	Non-Disclosure Agreement
NETE	External Interconnect
NETI	Internal Interconnect
NMC	Neuromorphic Computing
NVM	Non-Volatile Memory
NVRAM	Non-Volatile Random Access Memory
OIDC	OpenID Connect
OpEx	Operational Expenditure
PaaS	Platform as a Service
PCP	Pre-Commercial Procurement
PI	Principal Investigator
PID	Persistent Identifier
PIE	Public Information Event
PRACE	Partnership for Advanced Computing in Europe
Q&A	Questions and Answers
QoS	Quality of Service
R&D	Research & Development
R&I	Research & Innovation
RBAC	Role-Based Access Control
RFI	Request For Information
SCC	Scalable Computing Services
SGA	Specific Grant Agreement
SIB	Science & Infrastructure Board
SLA	Service Level Agreement
SP	Subproject
TCO	Total Cost of Ownership
TGCC	Très Grand Centre de calcul du CEA
UI	User Interface
US	User Support Services
VM	Virtual Machine Services
G100	Galileo100
SSD	Solid State Disk
NVMe	Non-volatile memory express

1. Introduction

Based on the scientific use case requirements described in D3.6 [1], the ICEI project team set up the common technical specifications described in D3.1 [2]. These specifications were the basis for the tendering technical specifications developed in D4.1 [3], resulting in coordinated procurements led by the Fenix sites.

In 2019 CINECA started the procurement procedure to provision the infrastructure components for the ICEI project. This document describes the result of the procurement process. According to the site specialization defined in D3.1, CINECA has elected to provision an infrastructure particularly devoted to data analysis and storage, scalable computing and cloud computing. The procured infrastructure is well suited for use cases that require large memory servers and intensive I/O workloads, flexible computing environments, and small-to-mid job size scalability.

CINECA ICEI Infrastructure is provided through a highly integrated system named Galileo100, in short G100. G100 provides services that exploit the synergy of having infrastructure with tightly coupled computing and storage resources

2. Overview of infrastructure components at CINECA

The table below outlines the Fenix services and the underlying resources provided by CINECA through the ICEI G100 system, as well as the corresponding quarterly allocation to HBP and PRACE users:

			Quarterly allocation		
G100 Component	**Service type**	**ICEI resources**	**Total**	**HBP (25%)**	**PRACE (15%)**
Thin nodes	SCC	340 nodes	601,264 node-hrs	150,316 node-hrs	90,189 node-hrs
Fat nodes + GPU nodes	IAC	214 nodes	378,443 node-hrs	94,611 node-hrs	56,766 node-hrs
OpenStack Cluster	VM	77 servers	77 servers	19 servers	11 servers
IME + Exascaler	ACD	10.5 PB	4,200 TB	2,625 TB	1,575 TB
Object Storage	ARD	10 PB	4,000 TB	2,500 TB	1,500 TB

Table 1 – Summary of ICEI services and underlying resources provided by CINECA.

Total resources availability is calculated assuming a typical resource provisioning (a 95% availability is assumed to take maintenance into account, and due to scheduling inefficiencies, a maximum usage of 85% is considered realistic).

G100 is an integrated system that provides all the services summarized in Table 1. Figure 1 reports a simplified schema of the computing infrastructure.

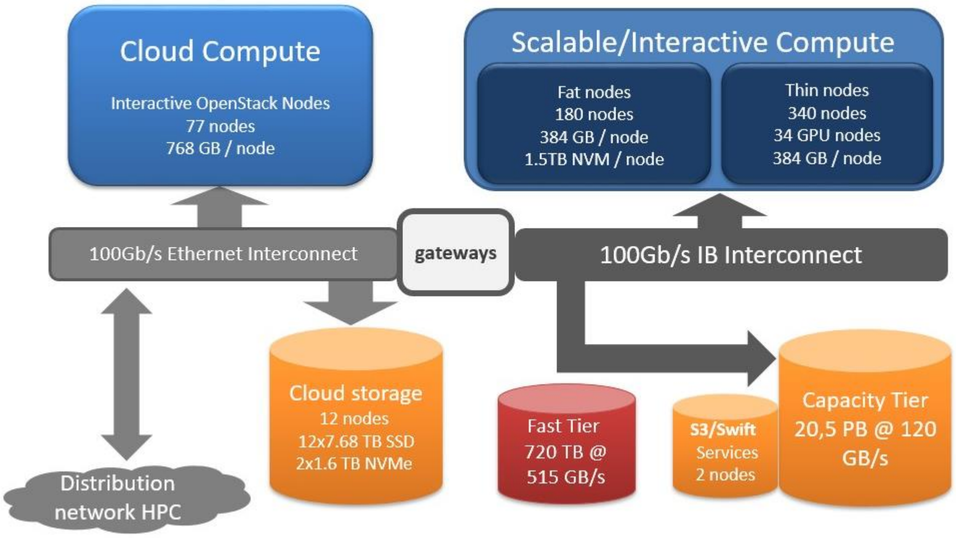


Figure 1 - Internal architecture of G100 Computing Cluster

3. Scalable computing (SCC) and interactive computing (IAC)

The Scalable/Interactive computing partition of the G100 system will be mainly devoted to support mid-size parallel jobs and interactive computing session for data analysis and visualization. In this regard a variety of different workloads can be addressed. This range covers, IO/RAM intensive (via fat nodes), accelerated workloads (via GPU nodes) and traditional scalable workloads (via thin nodes).

From the usage perspective, CINECA foresees the users to interact mainly through the batch job workload manager (Slurm [4]) exploiting its capability to utilize diverse computing resources thus granting the user the ability to compose the desired workflow.

The SCC and IAC partition features a computational peak power of roughly 2 PFlop/s, over 27,000 cores and 200 TB of system memory available for user computations.

More information on the system can be found here: https://wiki.u-gov.it/confluence/display/SCAIUS/UG3.3%3A+GALILEO100+UserGuide#UG3.3:GALILEO100UserGuide-SystemArchitecture

3.1 Hardware configuration

The scalable computing and interactive computing partition consist of 564 cluster nodes, interconnected via a NVIDIA Mellanox HDR100 InfiniBand, with a fat-tree topology and a 2:1 blocking ratio. The resulting interconnect throughput was deemed to be sufficient to support small-to-mid-size job parallelism, while keeping the costs of the InfiniBand network low.

The 564 nodes are subdivided into the following groups:

- 340 thin nodes, each equipped with:
  - 2x CPU 8260 Intel Cascade Lake, 24 cores, 2.4 GHz base frequency (3,90 GHz Turbo)
  - 384 GB RAM DDR4 2933MT/s

  - 480 GB SSD local disk

- 180 Data processing Fat nodes, each equipped with:
  - 2x CPU 8260 Intel Cascade Lake, 24 cores, 2.4 GHz base frequency (3,90 GHz Turbo)
  - 384 GB RAM DDR4 2933MT/s
  - 2 TB SSD local disk
  - 1,5 TB Intel Optane (12x128GB Intel Optane DCPMM)

- 34 GPU nodes, each equipped with
  - 2x CPU 8260 Intel Cascade Lake, 24 cores, 2.4 GHz base frequency (3,90 GHz Turbo)
  - 384 GB RAM DDR4 2933MT/s
  - 2 TB SSD local disk
  - 2x NVIDIA GPU V100

Complete the scalable and interactive computing cluster partition 5 management and service nodes, a management network 1 GbE and a cluster service network 2x25 GbE for operations such as bare metal/OS installation, monitoring and metering. Furthermore, the system features 2 redundant NVIDIA Mellanox gateway devices to bridge the Ethernet (see OpenStack cluster partition in section 4.) and the InfiniBand sub-nets.

3.2 Software configuration

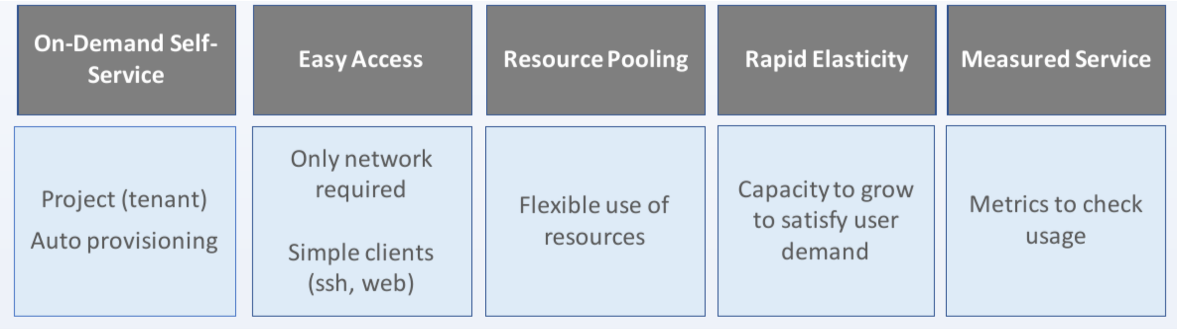
The computing nodes will be deployed with CentOS 8, while management and service nodes will host CentOS 7. Following the announcement that CentOS 8 will be discontinued at end of the year (2021), CINECA has performed a preliminary analysis to assess the current options available on the market. The most promising solutions are RedHat and Ubuntu.

The selected workload manager is Slurm. This is consistent with all the other HPC systems hosted by CINECA, and with the partners of the Fenix-ICEI project. HPC application software stack is provided as standard modules, spack packages [5] and containers, to ease users from building their own productivity environment. Given the Intel-based architecture of the computing nodes, the programming environment will be mainly based on the Intel compiler stack [6].

4. OpenStack Cluster (VM)

As reported in Figure 1, a partition of G100 will be devoted to providing cloud resources. The Cloud infrastructure integrates and completes the G100 system by providing a tightly integrated infrastructure that is able to serve high-performance computation through a flexible computing environment. We expect this flexibility to better adapt to the diversity of user workloads, while still providing high-end computing power.

CINECA offers cloud services via an OpenStack [6] infrastructure. Key features of a cloud infrastructure are:



On-Demand Self-Service	Easy Access	Resource Pooling	Rapid Elasticity	Measured Service
Project (tenant) Auto provisioning	Only network required Simple clients (ssh, web)	Flexible use of resources	Capacity to grow to satisfy user demand	Metrics to check usage

Figure 2 – Key features of the Cloud sevice.

In the context of cloud HPC resources provisioning, CINECA acts accordingly to the following division of roles:

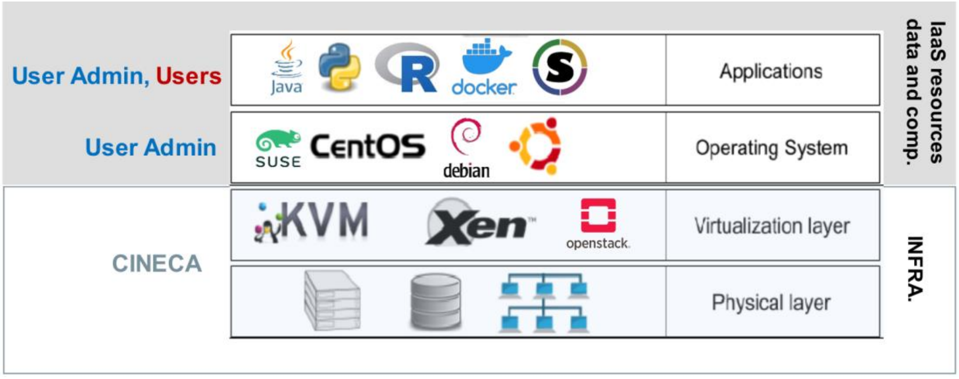


Figure 3 – CINECA typically acts as IaaS resource provider.

In brief:

- CINECA is responsible for administering the physical infrastructure and providing the virtualization layer.
- “User Admins” and “Users” are roles typically assumed by people external to CINECA. User Admins can create VM instances and configure the resources via dashboard; “Users” do not access directly the dashboard and are local to each VM instance (for example those added via *adduser* linux command).

Any user (“User Admins or “Users”) with administration privileges on IaaS resources (VMs) have the responsibility to maintain the security (security patch, fix) on those resources. From the project management perspective, CINECA will interact only with “User Admins".

More info on the VM service can be found here: https://wiki.u-gov.it/confluence/display/SCAIUS/ADA+CLOUD+CINECA+HPC

4.1 Hardware configuration

The OpenStack Cluster consists of:

- 77 OpenStack compute servers, featuring 2x CPU 8260 Intel CascadeLake, 24 cores each at 2.4 GHz, for a total of 48 cores per server and 96 vCPUs. This amounts to a total of 7400 vCPUs for the cloud infrastructure. Each computing server features also 768 GB of DDR4 RAM.
- 12 OpenStack storage nodes providing a total of 720 TB of dedicated CEPH [8] storage full flash (NVMe/SSD) for high IOPS. This storage will host VM root disks, disk volumes and VM snapshots.
- 6 nodes for storage management metadata (CEPH).
- 1024 floating IPs for external (public) connection.
- 3 service nodes to host OpenStack services.

The nodes are interconnected via a 100 Gb Ethernet, with a fat-tree topology.

4.2 Software

To provide cloud resources CINECA has relied on the OpenStack [7] software stack. It is a well-known and widely used software enabling to build private and public cloud instances. CINECA deployed the first OpenStack instance in 2015 with the version code-named Liberty. The OpenStack version currently in production, and deployed via Kolla Docker containers and Ansible playbooks, is code-named Train. The deployment through containers and Ansible playbooks allows easier maintenance (e.g., patch fixing) and greater flexibility easing installation of new (OpenStack) compute resources. CINECA cloud dashboard can be reached at the following url: https://adacloud.hpc.cineca.it. Authorized users will use the dashboard to control the resource allocation and perform common VM operations.

OpenStack is a modular application, providing multiple features. CINECA continuously evaluates those that are most promising for scientific communities use cases. Most notably:

- Encryption of cinder volumes. This feature in conjunction with the key manager service (barbican in OpenStack jargon) was identified as a key feature to host workloads processing sensitive data. Volume encryption can be directly managed by authorized users via the dashboard (feature available since OpenStack version Kilo) and is supported by the key manager. Volume encryption relies on standard LUKS which is supported by Linux kernel.
- Manila, file system shared between VMs. This feature will allow to easily setup and manage shared storage between multiple VMs through the OpenStack dashboard. The feature is still under evaluation. Alternative solutions to Manila can be built at user level. For example, an NFS service could be deployed to share to multiple client VMs a mounted volume attached to a server VM.
- Magnum, engine for container orchestration. Still under evaluation.

5. Active Data Repositories (ACD)

The purpose of Active Data Repositories (ACD) is to handle intensive I/O workloads and store hot data. For this purpose, G100 is equipped with a 2-tier storage system providing a single namespace to users in order to preserve ease of use while optimizing available performance.

In the following, the fast tier and capacity tier are described in more details.

5.1 *Fast tier*

It is based on DDN Infinite Memory Engine (IME) [9] and acts as an I/O performance enhancer layer between compute and parallel filesystem storage. IME is a software-hardware appliance that receives fragmented IO requests from compute nodes, stores data buffers and metadata in their NVMe disks, and optimally writes/reads coalesced data to the capacity tier backend so that the parallel filesystem can operate at maximum efficiency. Since IME keeps the underlying POSIX filesystem semantic, end users will not be required to adapt the application source code.

G100 features a fast tier based on IME-140 appliances with the following characteristics:

Storage Fast Tier	
Net capacity	720 TB
Disk technology	full flash (NVMe and SSD)
Bandwidth:	Io500 [10] bandwidth (easy and hard): 500 GiB/s read; 400 GiB/s write.

5.2 *Capacity tier*

The capacity storage will provide the POSIX parallel filesystem based on Lustre [11]. Some key security features that will be evaluated are: i) Lustre multi-tenancy, available from version 2.10, aiming to improve security and isolation so that only authenticated users can access a selected portion of the storage namespace: ii) Lustre encryption at rest, available from version 2.14, based on cryptofs, transparent to the parallel filesystem and able to handle file multiple access (multiple client).

The capacity tier storage is implemented with 6 DDN ES7990X appliance each connected to the InfiniBand data network via 4 ports HDR100, for a total aggregated bandwidth of 300 GB/s.

G100 features a capacity tier with the following characteristics:

Storage Capacity Tier	
Net capacity	20.5 PB
Disk technology	NVMe and HDD
Bandwidth:	Io500 [9] bandwidth (easy): 120 GiB/s

6. Archival Data Repositories (ARD)

Within the Fenix-ICEI context the Archival Data Repositories (ARD) is a logical abstraction referring to an object storage service aiming to preserve large amounts of data over time. In the context of the project, this storage must be accessible via HTTP(S) through the OpenStack Swift.

CINECA implements the archival data repository service via the performance optimized capacity storage tier, that will store the data and metadata, and through 2 dedicate DDN service nodes that offer the S3/Swift export service (see Figure1 in section 2). The ARD Service is currently in an early operational/pilot phase and available for interested users. The integration with the Fenix AAI will be performed once the necessary prerequisites by the AAI are met. The Archival Data Repository service must be provided with full support for OAuth2

authentication and, at the same time, must provide compatibility with the common standard authentication protocols, as e.g. OIDC and SAML. The Archival Data Repository of course needs to be configured to provide username mapping functionality locally at the site, this to correctly map the Fenix username to the local hpc user account on the machine.
Finally, for what concerns the storage resources: up to 2.5 PB of storage are available for users of the Human Brain Project and 1.5 PB for PRACE users in the context of the ICEI-PRACE joint calls.

7. Concluding remarks

This document provides an overview of the ICEI resources operated by CINECA from August 2021. From the initial planning system availability was delayed due to unforeseen storage infrastructure issues that required a full reconfiguration on CINECA site.

CINECA expects the system in pre-production from August 2021 and in full production after completion of the pre-production phase. Computing and storage resources are provided to the PRACE-ICEI joint Calls for Proposals and are available to interested users of HBP and PRACE.

8. References

[1] ICEI Deliverable 3.6: Scientific Use Case Requirements Documentation
https://drive.ebrains.eu/smart-link/79b8717a-f730-43e3-a747-61797181077e/

[2] ICEI Deliverable 3.1: Common Technical Specifications
https://drive.ebrains.eu/smart-link/1149da88-5dc4-4195-b364-80d8a71fe668/

[3] ICEI Deliverable 4.1: Tender Documents (Part 1)
https://drive.ebrains.eu/smart-link/bc0a44df-bf60-4f9f-87ad-8853958ee3d4/

[4] Slurm: https://slurm.schedmd.com/

[5] Spack; https://spack.io/

[6] oneAPI: https://software.intel.com/content/www/us/en/develop/tools/oneapi/all-toolkits.html#gs.6a2kt0

[7] OpenStack: http://www.openstack.org

[8] Ceph filesystem: http://ceph.io

[9] IME: https://www.ddn.com/products/ime-flash-native-data-cache/

[10] io500: https://www.vi4io.org/io500/start

[11] Lustre filesystem: http://lustre.org