D4.5

Infrastructure at BSC

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

Date	Author	Comments	Version	Status
09.03.2021	Javier Bartolomé	Initial version	V0.1	Draft
14.04.2021	Javier Bartolomé	Whole text included	V0.2	Draft
01.06.2021	Javier Bartolomé	Draft version for internal review	V0.3	Draft
21.06.2021	Alex Upton	1st internal review, minor revisions	V0.31	Draft
22.06.2021	Javier Bartolomé	Accept/review 1st internal review	V0.4	Draft
29.06.2021	Mirko Cestari	Internal review	V0.41	Draft
30.06.2021	Javier Bartolomé	Accept/review 2nd internal review	V0.5	Draft
10.07.2021	Dirk Pleiter	Internal review	V0.55	Draft
14.07.2021	Javier Bartolomé	Correct Dirk´s comments	V0.6	Draft
20.07.2021	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 BSC for the Fenix research infrastructure as of June 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
SSD	Solid State Disk
NVMe	Non-volatile memory express

1. Introduction

Based on the scientific use case requirements described in deliverable D3.6 [1], the ICEI project team set up the common technical specifications described in deliverable D3.1 [2]. These specifications were the basis for the tendering technical specifications developed in deliverable D4.1 [3] and for BSC infrastructure and R&D services in deliverable D4.15 [4], resulting in coordinated procurements led by the Fenix sites.

This document describes the services set up at BSC for the Fenix infrastructure, including the underlying components and systems acquired as part of the aforementioned procurements, as well as in-kind resources to support the project objectives.

According to the site specialization defined in D3.1 and D4.15, BSC implementation focuses on providing modern computing technologies: virtual machines, container technologies, jupyter hub jobs (https://jupyter.org/hub), workflow, visualization and steering, in conjunction with efficient handling and processing of large volumes of data.

2. Summary of infrastructure components at BSC

The table below gives a summary of the ICEI resources available at BSC, the type of service they provide, and the corresponding quarterly allocation to HBP and PRACE users:

				Quarterly allocation		
Component	**Service type**	**New/ In-kind**	**ICEI resources**	**Total**	**HBP (25%)**	**PRACE (15%)**
Nord3 cluster	SCC / VM	In-kind	84 nodes	84 servers	21 servers	12 servers
Interactive Computing Cluster[1]	IAC	New	3 nodes	5832 node-hrs	1486 node-hrs	874 node-hrs
HPC Storage	ACD	In-kind	70 TB	70 TB	17.5 TB	10.5 TB
Agora Archive Storage	ARD	New	10 PB	10 PB	2.5 PB	1.5 PB

Table 1: ICEI resources at BSC; Interactive Computing Services (IAC), Virtual Machine Services (VM), Active Data Repositories (ACD) and Archival Data Repositories (ARD).

3. Nord3 Cluster (SCC/VM)

Nord3 is a supercomputer based on Intel SandyBridge processors, iDataPlex Compute Racks, a Linux Operating System and an Infiniband interconnection.

[1] Note that the calculation of available computation hours is based on a 90% availability rate

Figure 1: Image of Nord3 compute racks

The current Peak Performance of the system is 292 Gigaflops. In total there are 12,096 Intel SandyBridge-EP E5–2670 cores at 2.6 GHz, distributed across 756 compute nodes, with at least 24.2 TB of main memory.

The compute nodes of Nord3 can be booted with two types of operating system image:

- HPC image: The node will be automatically added to Slurm and will be available to run HPC batch workloads
- Cloud image: The node will be added to the OpenStack [5] infrastructure allowing the execution of VMs

3.1 Hardware

The complete Nord3 cluster is composed of the following hardware components:

- 9 iDataPlex compute racks. Each of these is composed of:
  - 84 IBM dx360 M4 compute nodes
  - 4 Mellanox 36-port Managed FDR10 IB Switches
- All IBM dx360 M4 compute nodes contain:
  - 2x E5–2670 SandyBridge-EP 2.6GHz cache 20MB 8-core
  - 500GB 7200 rpm SATA II local HDD
- Interconnection Networks:
  - Infiniband Mellanox FDR10
  - 1 and 10 Gigabit Ethernet
- Operating System: Linux - SuSE Distribution 11 SP3 (being upgraded to RHEL8)

There are 3 types of nodes, each one with a different amount of memory available:

- 500 Default nodes: 32 GB/node
- 128 Medium memory nodes: 64 GB/node
- 128 High memory nodes: 128 GB/node

Note that only a subset of the Nord3 cluster is devoted to ICEI resources, which are one full compute rack with each 84 compute nodes.

Computing resources of this cluster will be devoted to ICEI calls to offer Scalable Computing (SCC) and Virtual Machine (VM) services.

In September 2021, a set of 8 servers will be added to the OpenStack VM service, to cover those requests that may require a large amount of main memory. Here are the hardware characteristics of this extension are as follows:

- Number of Nodes: 8
- 2x Intel(R) Xeon(R) Gold 6152 22 cores CPU @ 2.10GHz
- 256 GB main memory
- 4x 25 Gbit Ethernet network connection

3.2 Software

The Nord3 cluster is deployed using xCAT software. Depending on the use-cases and the underlying resources requested, nord3 compute nodes can be configured from running the HPC operating system image to running the cloud operating system image, and by doing so provide adequate computing resources to the allocated use-cases.

The HPC operating system is currently based on SLES11 SP3, with LSF as the batch scheduling system, and is in the process of being upgraded to RHEL 12 and Slurm. The Cloud operating system is based on RHEL 8, and OpenStack version train as the software to provide all VM services.

4. Interactive Computing Cluster (IAC)

The interactive computing cluster provides computational resources with large memory (1 TB), and will be used for pre- and post-processing tasks, steering computation, remote visualisation and for running MPI applications. Two of the three nodes of the cluster will have GPUs, in order to support for visualisation workloads.

The cluster has a compute power of 30 TFlop/s peak, 2.5 TB of memory and 13 TB NVMe storage. Each of the nodes uses ports at 25 and 100 Gbit Ethernet for their interconnection.

4.1 Hardware

The Interactive Computing Cluster consists of:

- 2 compute nodes, composed of:
    - IBM Power System AC922 System
    - 2x IBM POWER9 16-core 2.6GHz Processor
    - 16x 64 GB DDR4 2666 MHz DDR4 RDIMM Memory
    - 2x NVIDIA Tesla V100 SXM2 16GB Accelerator
    - 1x Ethernet 25Gb Adapter with two ports
    - 2x 960 GB SSD Disks
- 1 compute node, composed of:
    - IBM Power system AC922 System

- 2x 20-core 2.4 GHz (3.0 GHz Turbo) POWER9 Processor
- 512GB memory in 16x 32GB DDR4 DIMMs
- 2x 960 GB 2.5in SATA/SSD Disk Drive
- 1x PCIe3 LP 2-port 100GbE (NIC& RoCE) QSFP28 Adapter x16
- 2x PCIe3 LP 6.4 TB SSD NVMe adapter

4.2 Software

All of the interactive compute cluster is installed via xCAT using Red Hat Enterprise Linux. The latest version of Slurm will be installed which will be used to distribute user workloads between the 3 nodes.

Current software versions to be installed are collated in the following list:

- Red Hat Enterprise Linux 8.2 4.18.0-193.37.1.el8_2.ppc64le
- xCAT 2.16.1
- IBM Spectrum MPI 10.4.0.3
- IBM Spectrum Scale 5.0.5-4
- NVIDIA CUDA 11 Toolkit 11-0-11.0.3-1
- NVIDIA Driver 450.80.02

5. Active Data Repositories (ACD)

Active Data Repositories (ACD) is the POSIX based storage infrastructure, which is mounted in all SCC resources, and being able to handle intensive I/O workloads.

BSC implements ACD for Fenix using their main HPC Storage infrastructure, which is described below.

5.1 Hardware

HPC Storage infrastructure at BSC were acquired in conjunction with the MareNostrum4 supercomputer, and is based on the IBM Elastic Storage Server GL6 model. The complete solution uses 7 modules for data storage and 2 modules for metadata storage, and has the following hardware configuration:

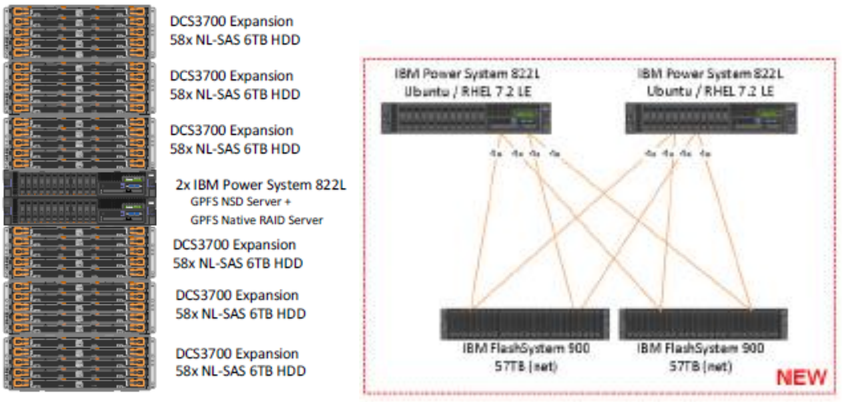

Figure 2: Data and Metadata modules for ACD

The complete HPC storage provides a total capacity of 14 PB, and an overall performance of 150 GB/s. This storage infrastructure can be accessed by all supercomputing clusters available at BSC.

All the servers use 10/40 Gbit Ethernet and OPA interconnectivity to provide the connection to the different BSC supercomputing clusters.

5.2 Software

All storage hardware for ACD is using IBM Spectrum Scale/GPFS version 4 as a parallel file system to be mounted in all BSC HPC computing nodes.

In order to ensure reliability on the different hardware components, IBM Spectrum Scale Native RAID is used for data storage, and normal RAID controllers for metadata storage.

IBM Spectrum Scale Native RAID provides and integrates storage controller functionality inside a GPFS server. Instead of defining LUNs (storage volumes) in an external storage controller, IBM Spectrum Scale Native RAID software takes care of creating and managing those LUNs using all hard drives inside the JBODs expansions.

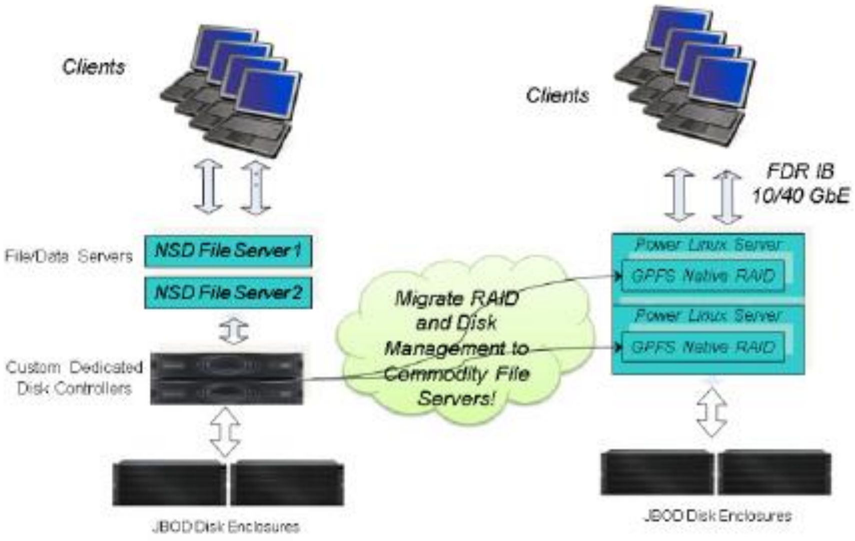

Figure 3: IBM Spectrum Scale Native RAID vs Normal controllers

6. Archive Agora Storage - Archival Data Repositories (ARD)

The purpose of Archival Data Repositories (ARD) is the long-term storage of large amounts of data (experimental data, simulation results…). In the context of Fenix, this storage must be accessible through the OpenStack Swift protocol; as such, this service will be implemented by an infrastructure called Agora. In addition, this storage is used to store any information needed by the VM Service (OpenStack).

Movement of data between ARD and ACD will be performed via the data mover service, at BSC this will be implemented via 6 servers which are part of the Agora hardware.

The Agora system is a hierarchical storage that uses 15 PB based on Flash and hard drive storage as a first-tier of storage, and tape storage of 100PB as a second-tier.

6.1 Hardware

The Agora Archive infrastructure is a modular storage solution able to scale to more than 100 PB, with a first level of high performance storage based on Flash/SSD for metadata and HDD for data, and a second level, cost-effective, multi-petabyte storage based on tapes.

The main hardware components of Agora are:

- 3x IBM Elastic Storage Server GL6S providing data storage for the first level of storage
- 2x IBM Elastic Storage Server GS4S providing metadata storage for the first level of storage
- 1x IBM Tape library TS4500 with a total of 80 LTO8 drives and almost 100PB storage
- Set of servers to provide the different data services:
    - 8x Cloud Servers to host virtual machines, which could be offered on-demand out of the OpenStack installation for those cases which cannot be covered under the VM service
    - 8x Archive Servers that will implement the movement from the different tiers (HDD and tape) inside the Agora storage infrastructure
    - 6x Data Mover Servers that will hold all data transfer operations between different storages (ACD, ARD)
    - 4x Export servers that will provide accessibility to data through other protocols such as, Swift, NFS or SMB
    - 2x Monitoring Servers that will take care of monitoring and alerting for all components Agora is composed of
    - 2x Backup servers that will take care of backup tasks for different services

All components will be connected using 10/25/40/100 Gbit Ethernet connectivity using 2 switches as a central point based on Mellanox SN3800 switches, which also connects this infrastructure to the rest of the BSC LANs and external Internet.

The following figure shows all Agora hardware components and their function:

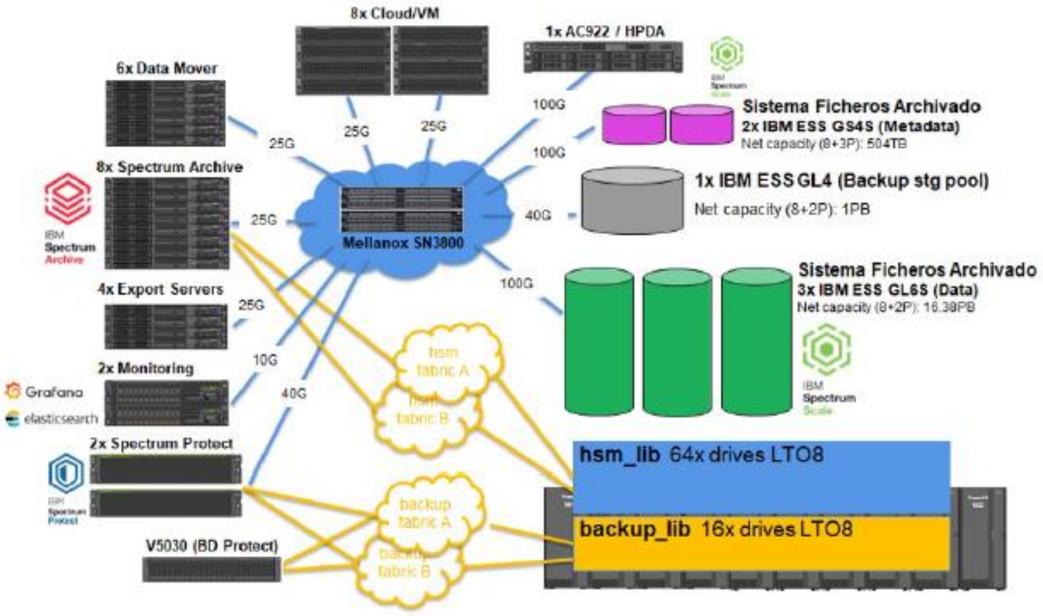

Figure 4: Agora (ACD) hardware overview

6.2 Software

Several software components work together to provide Agora storage services. Mainly we will focus on the IBM Spectrum Scale and IBM Spectrum Archive, which provide parallel file system and HSM functionalities respectively.

IBM Spectrum Scale is a widely used parallel file system in supercomputing and other type of IT services, it provides a high performance file system that is accessible by thousands of clients via different type of networks.

In order to provide reliability against hardware failures, IBM Spectrum Scale Native RAID is used. This is a software technology that provides RAID technologies, distributing the set of RAID information in a declustered fashion to improve recovery time in case of hardware issues (Hard drive or SSD failures).

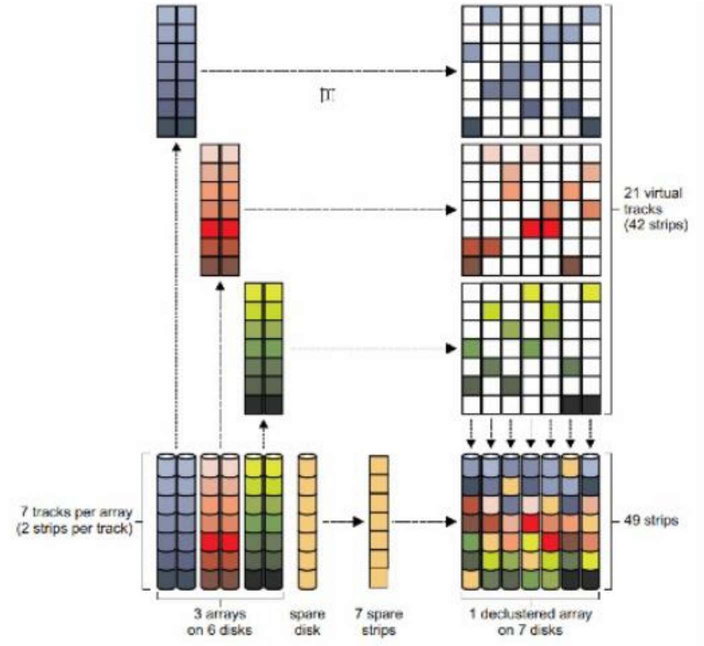
π
21 virtual tracks (42 strips)
7 tracks per array (2 strips per track)
49 strips
3 arrays on 6 disks
spare disk
7 spare strips
1 declustered array on 7 disks

Figure 5: IBM Spectrum Scale Native RAID

The tape-based second tier of storage integrated with the parallel file system as a second-tier (lower cost storage) to store less frequently accessed data. Spectrum Archive is a scalable and innovative solution that transparently moves data to tape when data are not accessed, and is recovered to disk triggered by user access.

Spectrum Archive uses an open format, called Linear Tape File System (LTFS), which permits the export of tape information to other tape robots.

IBM Spectrum Archive includes functionalities for high availability; inside the cluster of 8 servers one acts as a control node responsible for migration and recall task coordination amongst all the servers. The control node functionality can be migrated to other servers in case of any eventuality.

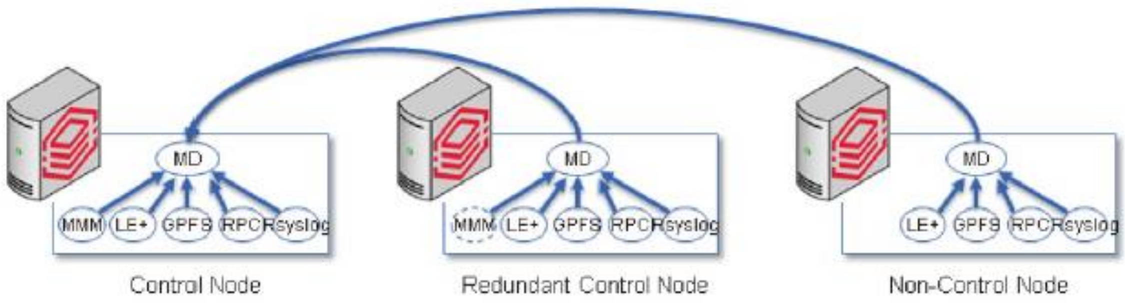

Figure 6: Spectrum Archive Control Node high-availability

7. Concluding remarks

This document provides an overview of the ICEI resources operated by BSC in June 2021.

These resources have been integrated in the PRACE-ICEI joint Calls for Proposals, and are available to interested users of HBP and PRACE.

The first accepted projects started using these resources on June 1st, 2021.

8. References

[1] ICEI Deliverable 3.6: Scientific Use Case Requirements Documentation https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d2749711/ICEI-D3.6-v3.1_clean.pdf

[2] ICEI Deliverable 3.1: Common Technical Specifications https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d2749698/ICEI-D3.1-v3.1_clean.pdf

[3] ICEI Deliverable 4.1: Tender Documents (Part 1) https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d2749723/ICEI-D4.1-v3.1_merged.pdf

[4] ICEI Deliverable 4.15: Tender Documents (Part 2) https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d2995389/ICEI-D4.15-v2.2_merged.pdf

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