FENIX

RESEARCH INFRASTRUCTURE

D4.4

Infrastructure at CEA

Work package:	WP4 - Procurement, deployment and operation	
Author(s):	Thomas Leibovici	CEA
Reviewer #1	Javier Bartolomé	BSC
Reviewer #2	Colin McMurtrie	ETH Zurich /CSCS
Dissemination Level	Public	
Nature	Other	

Date	Author	Comments	Version	Status
24.06.2020	Thomas Leibovici	Initial version	V0.1	Draft
03.07.2020	Thomas Leibovici	Ready for internal review	V0.2	Draft
15.07.2020	Thomas Leibovici	Minor changes after internal review	V0.3	Draft
20.07.2020	Thomas Leibovici	Changes after internal review	V0.4	Draft
22.07.2020	Thomas Leibovici	Layout finalization	V1.0	Ready
23.07.2020	Anne Nahm	Final editorial updates	V1.1	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 CEA for the Fenix research infrastructure as of July 2020.

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
CEA	Commissariat à l'énergie atomique et aux énergies alternatives
CINECA	Consorzio Interuniversitario
CLI	Command Line Interface
CPU	Central Processing Unit
CSCS	Centro Svizzero di Calcolo Scientifico
DM	Data Mover Service
DWPD	Drive-full-Write-Per-Day (measure of disk endurance)
HBP	Human Brain Project
HPC	High Performance Computing
IaaS	Infrastructure as a Service
IAC	Interactive Computing Services
ICEI	Interactive Computing E-Infrastructure for the Human Brain Project
I/O	Input/Output (access to data)
ICN	Interactive Computing Node
JSC	Jülich Supercomputing Centre
OSS	Object Storage Server (Lustre data server)
PRACE	Partnership for Advanced Computing in Europe
QoS	Quality of Service
RAM	Random Access Memory
SSD	Solid State Drive (silicon-based storage device)
TGCC	Très Grand Centre de calcul du CEA
VM	Virtual Machine Services

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.

This document describes the components set up at CEA for the Fenix infrastructure, including in-kind resources and systems acquired as part of the aforementioned procurements.

According to the site specialization defined in D3.1, CEA implementation focuses on efficient handling and processing of large volumes of data. This implementation particularly addresses use cases that require large memory, large storage capacity, and intensive I/O workloads.

2. Summary of infrastructure components at CEA

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

			Quarterly allocation		
Component	**Service type**	**ICEI resources**	**Total**	**HBP (25%)**	**PRACE (15%)**
Interactive Computing Cluster(1)	IAC	32 nodes	49,932 node-hrs	12,483 node-hrs	7,489 node-hrs
OpenStack Cluster	VM	20 servers	20 servers	5 servers	3 servers
Work filesystem(2)	ACD	3.5 PB	287,437 TB-days	71,859 TB-days	43,115 TB-days
Flash filesystem(2)	ACD	970 TB	79,661 TB-days	19,915 TB-days	11,949 TB-days
Store filesystem	ARD	7.5 PB	7,500 TB	1,875 TB	1,125 TB
Swift/OpenIO	ARD	7 PB	7,000 TB	1,750 TB	1,050 TB

Figure 1 - ICEI resources at CEA; Interactive Computing Services (IAC), Virtual Machine Services (VM), Active Data Repositories (ACD) and Archival Data Repositories (ARD).

Details on resource computation:

(1) The calculation of available computation hours is based on:

- 95% of availability rate;

- 75% of average fulfil rate by the job scheduler. Indeed, the job scheduler reserve some idle nodes when it prepares the submission of large jobs. As a result, 75% is the average fulfil rate we state in operation.

(2) To maintain high performance on Lustre filesystems, it is recommended to keep used space below 90% of fulfil rate. Beyond this value, performance degradation and operational issues can occur.

3. Interactive Computing Cluster (IAC)

The Interactive Computing Cluster provides computational resources for pre- and post-treatments like meshing creation, steering computation, remote visualisation but also for running MPI applications. Some of its nodes have particularly large memory (3 Terabytes per node) to overtake usual limitations in neural simulation by allowing to compute larger brain regions.

The cluster totals a compute power of 325 TFlop/s, 1192 cores and 17 Terabytes of memory available for user computations.

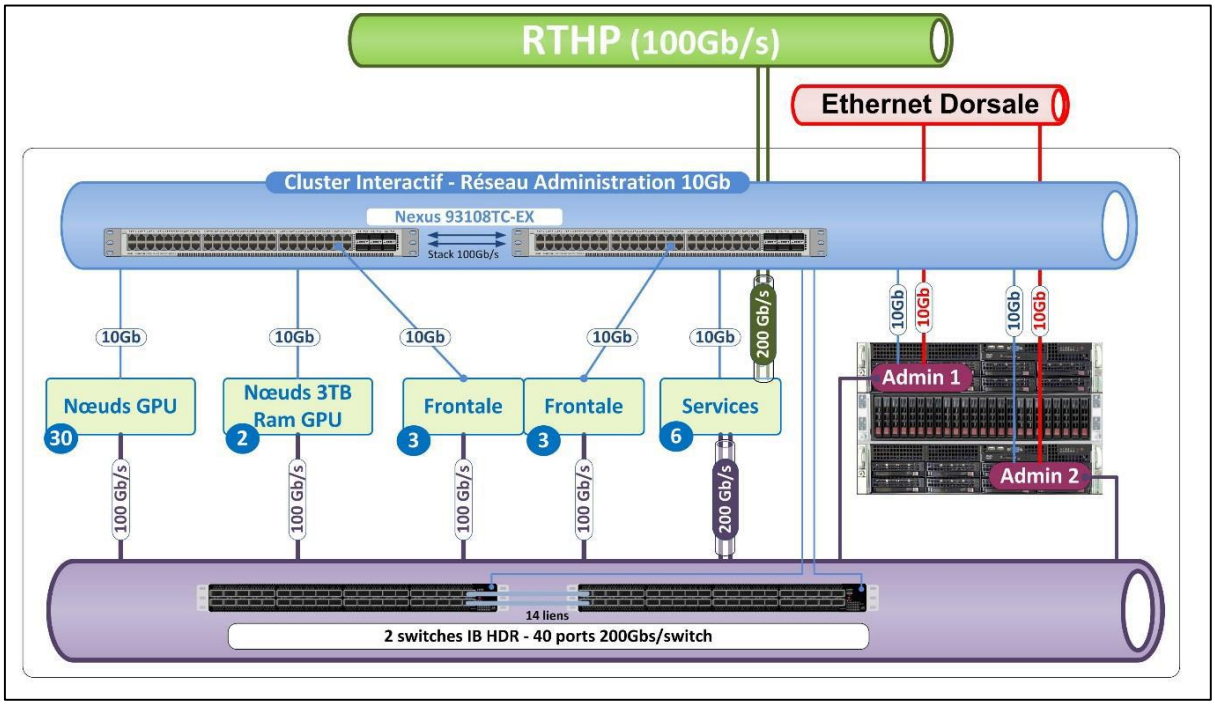

Figure 2 - Internal architecture of the Interactive Computing Cluster

3.1 Hardware

The Interactive Computing Cluster consists of:

- 30 compute nodes, made of:

    - 2 CPUs Intel Cascade Lake G-6240, with 18 cores each, running at a frequency of 2.6 GHz;
    - 384 GB of memory (RAM);
    - 1 GPU Nvidia V100 embedding 32 GB of RAM.
- 2 nodes with large memory, made of:
    - 4 CPUs Intel Cascade Lake G-6240, with 18 cores each, running at a frequency of 2.6 GHz;
    - 3,072 GB of memory;
    - 1 GPU Nvidia V100 embedding 32 GB of RAM.
- 6 login nodes to receive user incoming connections, and to host interactive sessions.
- 6 routers to connect the cluster to data repositories (ACD and ARD) though a 100 Gbits InfiniBand EDR network. Each router is connected to the internal interconnection network of the cluster on one side with 2 ports 100 Gbits, and to the high-performance data network of the compute centre on the other side with 2 other ports 100 Gbits. Thus, the total I/O throughput of the cluster is 6x2x100Gbits = 1200 Gbits (150 GBytes/sec).
- 2 management nodes.
- The internal interconnection network relies on 2 InfiniBand HDR 200Gb switches (2 for redundancy). Nodes are equipped with HDR-EDR 100 Gbits adapters.
- The administration network is Ethernet 10 Gbits.

3.2 Software

The administration nodes of the cluster are installed using *Deep Blue*, a cluster management software stack developed by CEA, based on CentOS 7. All other nodes of the cluster are installed with vendor's software *Bull SCS5*, an HPC-optimized Operating System based on RedHat Enterprise 7.

The scheduling of users' compute jobs is managed by SLURM (version 18). Job communications are managed by OpenMPI 4. The code compiler is Intel 19.

4. OpenStack Cluster (VM)

The OpenStack Cluster makes it possible to run virtual machines managed by users, so they can setup community services such as web services, data processing service, data exploration, etc.

With this cluster, a total of 720 cores and 3.8 Terabytes of memory are available to run VM services of Fenix. It also provides a total of 60.8 Terabytes of SSD-based storage for the local storage of virtual machines.

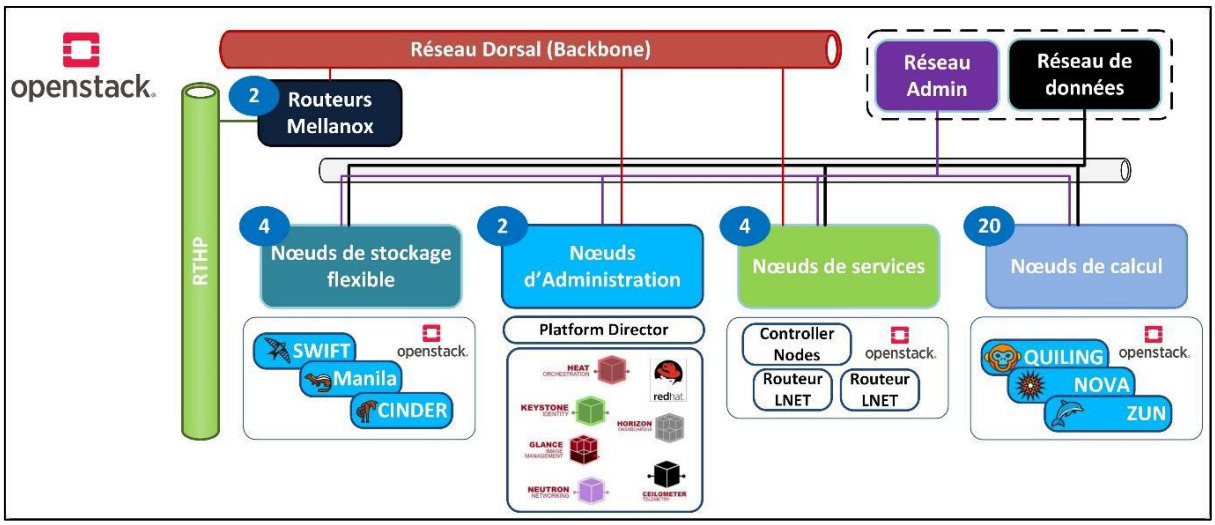

Figure 3 - Internal architecture of the OpenStack Cluster

4.1 Hardware

The OpenStack Cluster consists of:

- 20 hypervisors to run user's virtual machines. Each of these servers is equipped with 2 CPUs Intel Cascade Lake G-6240 (each 18 cores @ 2.6 GHz), and 192 GB of memory. These nodes are managed by OpenStack [4] components like Nova Compute, Nova KVM, Agent Ceilometer, Open vSwitch….
- 4 storage servers to provide a flexible pool of local storage to the virtual machines. Each storage server has 8 SSD drives of 1.9 Terabytes each. This configuration offers a total raw volume of 60.8 Terabytes. This storage will be managed by a CEPH [5] filesystem, exposed by OpenStack Cinder component (and possibly also Manilla, Swift…).
- 4 service nodes to host OpenStack services (Horizon, Keystone, Nova API, Neutron, Open vSwitch, Glance…).
- 2 administration nodes for bare metal systems control and orchestration.
- The network is made of 2 switches Cisco Ethernet 10Gb/s for internal administration and data flows. A QoS can be implemented to manage priorities between administration and user data flow.

4.2 Software

Like for the Interactive Computing Cluster, administration nodes are deployed with *Deep Blue*, based on CentOS7. The deployed OpenStack software stack is the community release "Ussuri", that can run on CentOS8.

5. Active Data Repositories (ACD)

The purpose of Active Data Repositories (ACD) is to handle intensive I/O workloads, in particular data extractions that are often needed to process experimental data.

The Fenix ACD at CEA are implemented by two Lustre [6] filesystems: *work* and *flash*. These filesystems are assigned with regards to user's needs in terms of capacity, throughput, and data lifecycle.

Work is a high performance parallel filesystem based on HDDs. It is adapted to general-purpose data access during computations.

The *flash* filesystem provides a higher performance in terms of I/O operations per second and is particularly adapted to non-sequential accesses. It however offers a lower capacity than the *work* filesystem.

5.1 *Work* filesystem

The *work* file system is a parallel filesystem (Lustre) based on classical spinning hard disk drives (HDDs).

It provides a storage capacity of 3.5 PB and a throughput up to 70 GB/s.

It is based on a Seagate ClusterStor L300 appliance, made of:

- 2 SSU blocks, each including 2 Lustre servers;
- 492 HDDs of 10 Terabytes (SAS), evenly distributed between SSUs and ESUs;
- 4 SSDs of 3.2 TB to speed-up small I/Os using the Nytro-XD feature of ClusterStor.

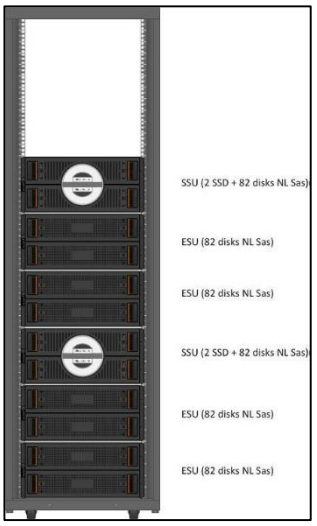

Figure 4 - Hardware of the 'work' filesystem

5.2 *Flash* filesystem

The *flash* filesystem is a parallel filesystem (Lustre) fully made of flash devices (SSDs).

It provides a storage capacity of 970 TB and a throughput of 110 GB/s. It allows low latency accesses up to 1,500,000 non-sequential read operations per seconds (read IOPS), and 260,000 non-sequential write operations per seconds (write IOPS).

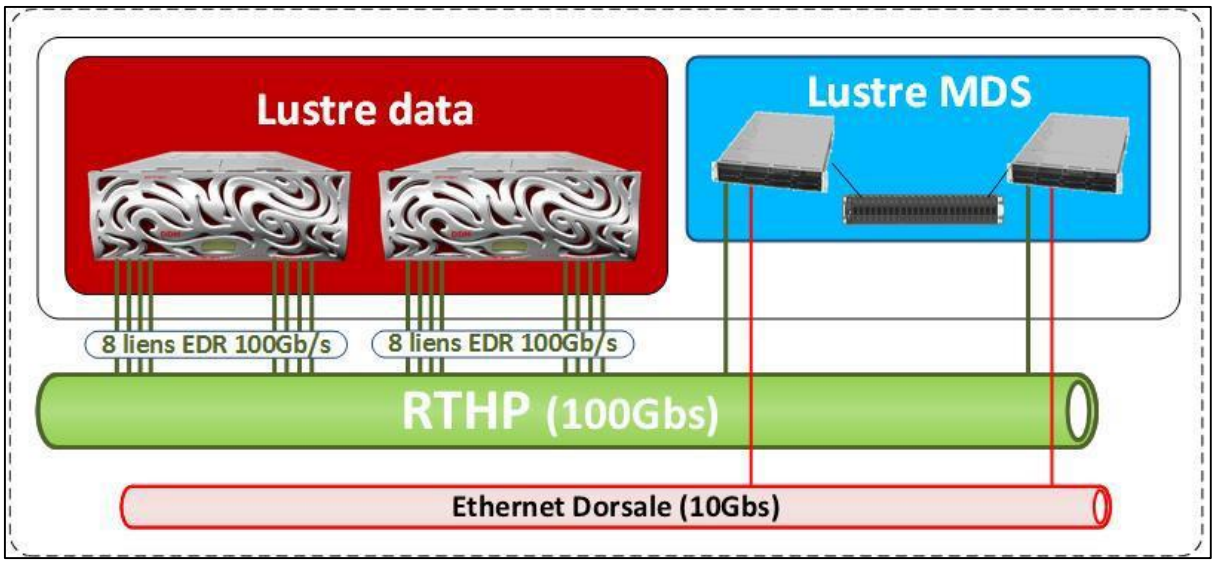

Figure 5 - Architecture of the 'flash' Lustre filesystem

This data storage is implemented by two DDN SFA 18KXe systems. Each of these systems is connected to the data network with 8 InfiniBand EDR links (100 Gbits).

Each DDN SFA 18KXe embeds:

- 8 to 16 virtual machines to run Lustre filesystems servers (OSS);
- 42 SSD drives of 15.38 Terabytes with an endurance of 1 DWPD;

  These drives are connected through a 12 Gb/s SAS interface.

Filesystem metadata is handled by 2 dedicated servers for both parallelism and high availability (HA) redundancy. Each server has the following characteristics:

- 2 CPUs Intel Cascade Lake G-6240 (18 cores @ 2.6GHz)
- 96 GB of DDR4 memory

These servers are connected to a NetApp E2824 disk array, including 4 SSDs of 800 GB with an endurance of 3 DWPD. Disks are configured in RAID10, thus providing a 1.6 TB capacity for storing metadata, which is enough to store about 1 billion files.

6. Archival Data Repositories (ARD)

The purpose of Archival Data Repositories (ARD) is to store large amounts of data in the long-term (experimental data, simulation results...). In the context of Fenix, this storage must be accessible through the OpenStack Swift protocol.

Two ARD systems are available at CEA to store such data:

- The *store* filesystem is a hierarchical storage that combines a filesystem (Lustre) as first level of storage on disks, and a long-term storage on tapes. User and applications can access this repository through a filesystem interface. In the near future, it will also be available through a Swift interface, thanks to the outcome of the ICEI R&D tender "Swift over open-source parallel filesystem".
- An object-store of 7 PB that natively provides an OpenStack Swift interface.

6.1 *Store* filesystem

The *store* filesystem is a high capacity and extensible storage implemented as a hierarchical storage. It combines a Lustre filesystem for the first storage level, and a set of tape libraries managed by IBM HPSS for the long-term storage.

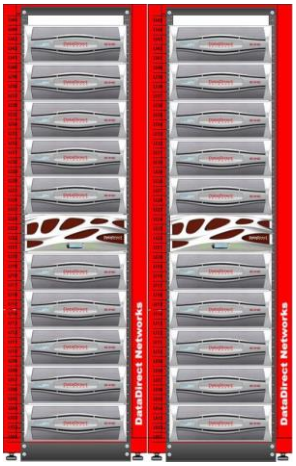

Figure 6 - Hardware of the 'store' filesystem

The Lustre filesystem is managed by 2 controllers DDN SFA 14KXe. Each controller is connected to 10 drawers SS8462 including 410 disks Hitachi HE10 of 8 TB. Each controller also runs 8 VMs that host the Lustre servers. The filesystem totals a usable volume of 4.8 Petabytes, and provides a bandwidth of 70 GB/s.

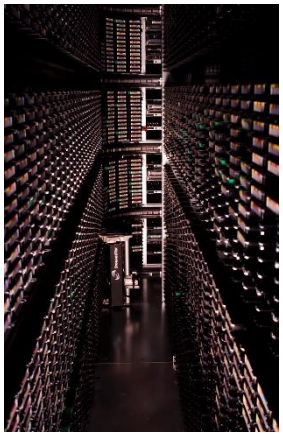

Figure 7 - Inside view of an Oracle SL8500 tape library

The backend storage is managed by IBM HPSS [7] software. It consists itself of two levels:

- one disk cache of 2.4 Petabytes stored in a DDN SFA 14KXe managed by HPSS disk movers;
- 3 tape libraries Oracle SL8500. The library complex contains 90 tape drives and has a capacity of more than 30,000 tapes. Tapes are accessed by HPSS tape movers.

Data is automatically and transparently migrated between the Lustre filesystem and HPSS using the "HSM" feature of Lustre, and the open-source software "RobinHood Policy Engine" [8].

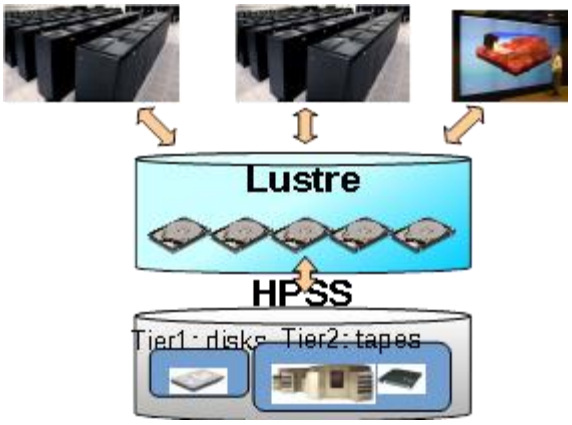

Figure 8 - Overview of the 'store' hierarchical filesystem at TGCC

Users and applications can access this storage system as a file system. It will later be available through a Swift interface too, using the outcome of the ICEI R&D tender "Swift over open-source parallel filesystem".

6.2 Swift/OpenIO

OpenIO [9] is an open-source object storage solution that natively provides S3 and Swift interfaces.

In the context of the ICEI project, OpenIO is deployed on the following hardware:

- 3 metadata servers with SSD disks. These nodes also host Swift servers;
- 1 DDN SFA18KXe for capacitive data storage.

Additionally, 2 servers are installed to run the Data Mover service developed as part of the ICEI project. These servers will be used to move data between ACDs and the ARDs through Swift.

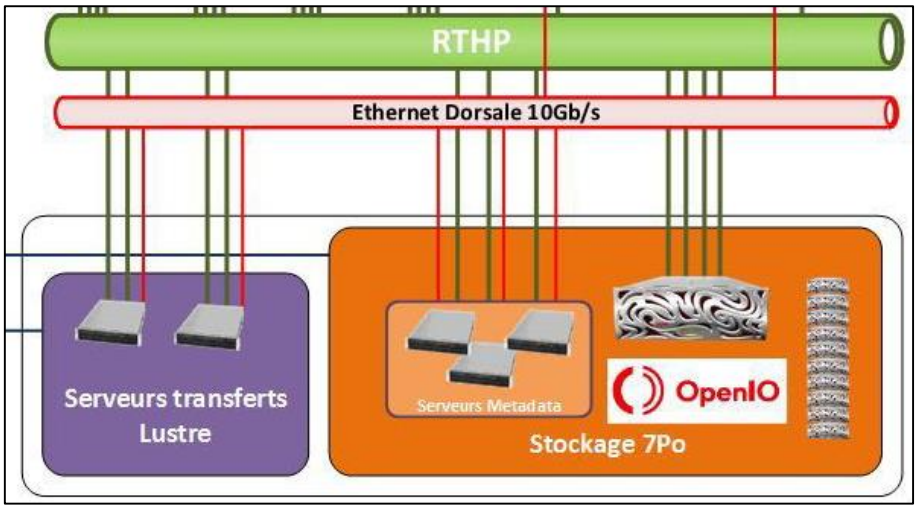

Figure 9 - Architecture of the OpenIO object storage system, and Data Mover service

For fault tolerance and high-availability, object metadata is replicated on three metadata servers. Each of these servers includes 4 SSD of 1.9 TB to store metadata. This will make it possible to store up to 2.5 billion objects.

The capacity data storage is managed by a DDN SFA18KXe controller, connected to 10 drawers SS9012 (up to 90 disks per drawer). The system totals 650 HDDs of 14 TB each.

Data safety is implemented by a de-clustered RAID 8+2 mechanism. Moreover, 10 "hot spare" disks are available to replace faulty disks automatically in case of hardware failure.

This system provides a total usable capacity of 7 Petabytes.

7. Concluding remarks

This document provides an overview of the ICEI resources operated by CEA in July 2020.

These resources have been integrated to 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 July 1st, 2020.

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

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

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

[7] HPSS (High Performance Storage System): http://hpss-collaboration.org

[8] Robinhood Policy Engine: http://robinhood.sf.net

[9] OpenIO: http://www.openio.io