D3.5

Consolidated e-Infrastructure Validation and Testing Report-Resubmission

Work package:	WP3 Technical specification and coordination	
Author(s):	Sadaf Alam	ETHZ
Reviewer #1	Thomas Leibovici	CEA
Reviewer #2	Mirko Cestari	CINECA
Dissemination Level	confidential	
Nature	Report	

Date	Author	Comments	Version	Status
01.12.2020	Sadaf Alam	Draft released for internal review	V0.1	Draft
05.12.2020	Dirk Pleiter	Review completed (removed an appendix)	V0.2	Draft
11.12.2020	Mirko Cestari	Review completed	V0.3	Draft
14.12.2020	Sadaf Alam	Addressed internal reviewer comments	V1.0	Final
02.02.2022	Sadaf Alam	Draft for resubmission based on feedback at ICEI Intermediate Review 4 (RV4)	V1.1	Draft
24.02.2022	Mirko Cestari	Review, added new section “CINECA Validation Results”	V1.1	Draft
17.02.2022	Thomas Leibovici	Review	V1.1	Draft
18.03.2022	Sadaf Alam	Updated version based on feedback from internal reviewers	V1.2	Draft

20.03.2022	Anne Nahm	Editorial updates	V1.3	Draft
29.03.2022	Javier Bartolomé, Anne Nahm	Added new section "BSC Validation Results"	V1.4	Draft
31.03.2022	Anne Nahm	Final editorial updates	V2.0	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 provides an update on the implementation and results of the ICEI validation framework that was presented in D3.4 (Validation Framework) [1]. Specifically, details are provided on the execution of the ICEI benchmarks, status of the e-Infrastructure consolidation, status of the KPIs and the ongoing implementation plans. An outlook of the next steps is provided on the research and development (R&D) services that have not been fully operational due to the procedural delays in the procurement and contract execution processes.

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/JUELICH	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

1. Introduction and Background

The objectives of the validation framework are to verify that ICEI infrastructure and federated services allow the expected execution of use cases workflows, and measure the performance of the deployed infrastructure, through a set of tests and benchmarks, and documentation of the results. At the time of submitting this report, the consolidated e-Infrastructure has not been fully materialised primarily due to the procedural delays in the procurement and contract execution processes at different ICEI sites. Notably, services related to scalable and interactive computing, virtual machine and active and archival data repositories are in place for allocation at multiple sites. Furthermore, production Fenix AAI service is in place. Remaining services include R&D services related to Fenix User and Resource Management Service (FURMS), the Active Data Repository to Archival Data Repository Data mover, the Fenix interactive computing service, and additional supported features for the archival data repositories are still under development following their corresponding roadmaps. For them, the awards have been granted and the prototype implementations are expected to be available during the first half of 2022.

In D3.4, the difference between validation, verification and monitoring were identified to avoid possible misunderstandings. The scope of validation is to understand if the requirements can be satisfied through implementation and settings of the infrastructure and services or, in other words, if the infrastructure and services allow these actions to satisfy the use case requirements. The scope of the verification is to check infrastructure and services after that implementation and setting phases have been completed. The scope of monitoring is to regularly check the availability and/or uptime of infrastructure and services to guarantee the correctness of the use cases execution.

1.1 Key Updates for Validation Efforts (2018 to 2020)

To fulfil the scope of validation, steps have been taken

- to understand and document requirements of the use cases,
- to identify mappings of the use cases to the infrastructural elements,
- to carry out procurements, and
- to identify benchmarks and usage scenarios.

A status update on key deliverables and milestones to date confirm that the infrastructure and services in operation have satisfied the use case requirements. These steps will continue as future R&D services are delivered and operational during the remaining timeline of the project, i.e., until September 30 2023. A summary of key activities since the beginning of the ICEI project are listed in table 1.

Table 1: Timeline and key updates

	2018	2019	2020-2021
Documentation (use cases mapping and usage of the infrastructure)	D3.6 (Scientific Use Case Requirements Documentation) [2] D3.1 (Common Technical Specifications) [3]	D3.4 (Validation Framework) D4.9 (Data storage and compute provisioning during M1 - M12) [6]	D3.5 (this document submission) D4.10 (Data storage and compute provisioning during M13 - M24) [9] D4.11 (Data storage and compute provisioning during M25 - M36) [31]
Procurements and technical updates	D4.1 (Tender Documents (Part 1) [4] D4.2 (Infrastructure at ETHZ/CSCS) [5]	D3.2 (Initial Federated AAI Infrastructure) [7] D4.15 (Tender Documents (Part 2) [8]	D4.3 (Infrastructure at JUELICH-JSC) [10] D4.4 (Infrastructure at CEA) [11] D4.5 (Infrastructure at BSC) [29] D4.6 (Infrastructure at CINECA) [30] D4.7 (Report on deployed infrastructure) [32] D4.8 (R&D results) [33]
Validation tests, benchmark results and documentation	Use case mappings in D3.1	Identification of representative benchmarks and use cases in D3.4	Benchmarking results from JUELICH and CEA in D3.5 (this document)

1.2 Status of Fenix Infrastructure Services for Validation

Architectural specifications of the Human Brain Project (HBP) High-Performance Analytics and Computing (HPAC) Platform are outlined in [12][13]. For the scope of this document, these infrastructure services are categorised into three groups based on their readiness and mapping on the ICEI use case requirements:

- Group 1: these infrastructure services have been procured and are in operation. Mapping of use cases and requirements have been identified in D3.1, D3.6 and D3.4 alongside a list of benchmarks. These services include SCC (Scalable Computing Services), IAC resources (Interactive Computing Services), VM (Virtual Machine Services), ACD (Active Data Repositories), ARD (Archival Data Repositories), NETE (External Interconnect) and NETI (Internal Interconnect). Procurements for the necessary equipment have been approved (D4.1).
- Group 2: these are R&D services where the requirements of the use cases and their implication on federation are identified, and procurements are approved (D4.15). These include AAI (Authentication and Authorization Infrastructure), FURMS (Fenix User and Resource Management Services), DM (Data mover Service) and IAC (Interactive Computing Service) resource management and scheduling within a batch environment. Validation tests are defined for these services.
- Group 3: DT (Data Transfer Service), DL (Data Location Service), MONI (Monitoring services for infrastructure services) and SEC (Security services) are not addressed through ICEI but rather HBP SGA3 tasks. An analysis of requirements, design and mapping is needed as a prerequisite for validation.

Group 1 and 2 services will be covered as part of the D3.5 deliverable (Consolidated e-Infrastructure Validation and Testing Report) as an overall view of work accomplished during the reporting period since D3.4.

1.3 Discussion on coverage of the ICEI Use Cases

As part of D4.1 and D3.4 preparations, a set of benchmarks, including use case driven and synthetic, were identified for each site to validate the procurements. The components of the benchmark suite have been chosen such that it represents the breadth of the HBP science and use cases. The benchmarks are either directly based on applications or are based on micro-benchmarks with parameters chosen such that they reflect the anticipated use of the ICEI infrastructure. All benchmarks have been published under https://wiki.ebrains.eu/bin/view/Collabs/hbp-benchmark-suite-for-technology-trans/ [14].

A brief description of benchmarks (including number of tests per benchmark) is as follows:

- Elephant ASSET (1): This benchmark is based on the Elephant analysis package for analysis of neurophysiological data [15]. Elephant is a Python application, using mpi4py for multi-task parallelism and requiring Numpy for its core functionality.
- NEST (1): NEST is a simulator for spiking neural network models that focuses on the dynamics, size and structure of neural systems rather than on the exact morphology of individual neurons [16].
- Arbor (2): Arbor is a simulation library for networks of morphologically detailed neurons. Two benchmarks are designed to test these two parts of the workflow, a

computationally intensive Ring benchmark and network and memory intensive Proxy cell benchmark [17].

- NEURON/CoreNEURON (2): The MPI enabled benchmarks are based on the NEURON simulator [18]. The compute engine of the NEURON simulator has been extracted and is being optimized as a library called CoreNEURON. NEURON simulator can be configured to use CoreNEURON library for efficient execution. Altogether there are 4 benchmarks for Ring and Traub configuration for NEURON and CoreNEURON.
- Neuroimaging Deep Learning (1): The benchmark is based on TensorFlow (using Horovod for parallelisation) and implements a production use case using human brain images [19].
- TVB-HPC (1): The Virtual Brain (TVB) is a software which has become a validated and popular choice for the simulation of whole brain activity [20].
- TensorFlow (1): The benchmarks are based on TensorFlow and on the CNN (Convolutional Neural Network) benchmarks both running on a single or multiple nodes with synthetic and real data (ImageNet Challenge 2012) [21].
- IOR (1): IOR is a generic benchmark to evaluate performances of file systems. In the context of ICEI, it will be used with specific parameters to reproduce the data access patterns of the neuroscience codes [22].
- Cosbench (1): This is a reference benchmark for object stores. In the context of ICEI, it will be used to evaluate the performance of the archival storage for expected workloads and object sizes in the Fenix infrastructure [23].

In addition to the 11 validation tests listed above that are driven by the ICEI use cases, synthetic tests have been developed for validating R&D services such as Fenix AAI.

Table 2 provides an overview of test cases that were detailed in the earlier ICEI deliverable and their mappings to the ICEI benchmarking suite.

Table 2: Summary of use cases and mapping to representative benchmarks

Case ID	Use case (as identified in D3.6 and D3.1) for coverage	Representative benchmarks (part of D4.1)
1	Data-driven cellular models of brain regions, Olfactory Bulb	NEST, Arbor, Neuron, IOR, Cosbench
3	Learning-to-learn (LTL) in a complex spiking network on HPC and Neuromorphic hardware interacting with NRP	Neuroimaging Deep Learning, TensorFlow
5	Large scale simulations of models: Cerebellum	Neuron, IOR
6	Large scale simulations of models: Hippocampus	Neuron, IOR

7	Elephant big data processing	Elephant ASSET
8	Mouse Brain Atlas Imaging storage	Neuroimaging Deep Learning, IOR, Cosbench
9	Towards a novel decoder of brain cytoarchitecture using large scale simulations	Neuroimaging Deep Learning, TensorFlow
10	Large scale multi-scale co-simulation of the cortex (TVB <-> Nest <-> Arbor)	TVB-HPC, NEST, Arbor
11	Neurorobotics platform, large-scale brain simulations	NEST, IOR
12	BBP collumnar simulation	Neuron, IOR, Cosbench
13	Ilastik as a service on the HPB Collab	IOR, Cosbench
14	Online visualization of multi-resolution reference atlases	IOR
15	Data management and big data analytics for high throughput microscopy	IOR, Cosbench
16	Multi-area macaque Nest simulation with life visualization and interaction	NEST
17	Data management and big data analytics for large cohort neuroimaging	IOR

1.4 Update on D3.4 KPIs

D3.4 contains table 3 as target KPIs for the validation framework:

Table 3: KPIs listed in D3.4

Plans	M27	M30	M33	M36
# of defined tests	3	6	9	12
# of executed tests	0	4	8	12

# of use cases covered	5	11	15	15

Considering the mapping of benchmarks to the use cases (D3.1, D3.6 and D4.1), we provide how the coverage of use cases are determined. In order to understand the mapping, we refer to the concept of clusters introduced in D3.6 for different use cases. These include:

- Simulation: Use of simulation at any point in the processing.
- Multi-scale coupled simulation: Use of multiple simulators at different scales exchanging data at runtime.
- Co-deployment of applications: Processing pipelines needing multiple applications running concurrently and exchanging data at runtime.
- Streaming visualization: In-situ / in-transit visualization of applications running on HPC resources.1
- Machine learning: Machine learning somewhere in the processing pipeline.
- In-the-loop machine learning: Machine learning on data from an online data source.
- Big data processing: Big data collection, pre-processing, curation, processing and storing
- Big data visualization: Visualization of big data sets.

Further analysis done for D3.6 showed mapping of the feature sets of a use case to clusters:

- Simulation: 10 (use cases 1, 3, 5, 6, 7, 9, 10, 11, 12, 16)
- Multi-scale coupled simulation: 2 (use cases 1,10)
- Co-deployment of applications: 8 (use cases 1, 3, 5, 6, 7, 10, 11, 16)
- Streaming visualization: 7 (use cases 1, 6, 7, 10, 11, 13, 16)
- Machine learning: 5 (use cases 1, 3, 8, 9, 15)
- In-the-loop machine learning: 2 (use cases 1, 3)
- Big data processing: 5 (use cases 8, 9, 14, 15, 17)
- Big data visualization: 7 (use cases 8, 9, 12, 13, 14, 15, 17)

Arbor, NEST, Neuron and TVB-HPC benchmarks provide coverage for the simulation use cases including some multi-scale coupled simulation instances. Visualisation and machine learning use cases are covered by Bcfind, neuroimaging deep learning and IOR for stress tests. Big data processing and in-the-loop machine learning use cases are represented by the Elephant ASSET benchmark. TensorFlow benchmarks for machine learning and big data processing capabilities. The only benchmark that has not been executed on any ICEI platform so far is bcfind and this will not be attempted, since the development has been dropped by the research team that owns the application.

Table 4 shows the progression of the KPIs. The step after identifying an application for benchmarks, was to identify and package a test case based on the benchmark. Initial test cases were from the simulation based test cases. Later, machine learning was introduced. Additional test cases were added for storage performance. The last one to be added was for the Fenix AAI federation.

Table 4: Updates on D3.4 KPIs

Status	**M27** actual (target)	**M30** actual (target)	**M33** actual (target)	**M36** actual (target)
# of defined tests	3 (3)	4 (6)	11 (9)	12 (12)
# of executed tests	0 (0)	0 (4)	11 (8)	12 (12)
# of use cases covered	10 (5)	11 (11)	15 (15)	15 (15)

The work is ongoing to fulfil R&D requirements and to cover for all sites and will be reported as part of the status reports of individual services.

2. Use Cases Validation Results

Individual results gathered from procurements acceptance and validation are presented in this section to validate whether the use case requirements are being fulfilled by the procured ICEI hardware equipment and its configuration. It is relevant to note that for each procured infrastructure, the benchmarks reported in the subsequent sections were an integral part of the procurement process. Application suites with expected benchmark values were provided to the procurement candidates, with the request to assess the offered solution against the benchmarks. This was part of the evaluation process of the proposed bids. The winning bids have largely outperformed the benchmark set values, as assessed during the evaluation of the offers, and in in line with the state-of-the-art technology in the offered solutions.

Insights into parameters influencing performance and scaling characteristics of individual benchmarks have been reported from the study titled*: Performance Comparison for Neuroscience Application Benchmarks* [28]. The targeted HPC systems in this study namely JURON and JUWELS are composed of Power8 and Intel SkyLake based multi-core platforms respectively.

2.1 Characteristics influencing results of benchmarks

We briefly outline results reported for neuroscience and synthetic ICEI benchmarks that have been targeted for procurements and identify parameters of hardware and software configuration that can influence achievable performance results on the procured hardware.

- Elephant ASSET
  - Reported results: simulation time reported in seconds
  - Parameters influencing performance results: number of nodes, MPI tasks per node, OpenMP threads per node. A higher memory bandwidth ratio per task within a node is expected to result in better performance and scaling efficiencies.
- NEST
  - Reported results: simulation time reported in seconds
  - Parameters influencing performance results: number of nodes, number of tasks per node, number of threads per tasks. This is a compute-intensive, highly scalable application that does not exploit SIMD parallelism but benefits from a large number of threads per server.
- Arbor
  - Reported results: simulation time reported in seconds
  - Parameters influencing performance results: number of nodes, number of tasks per node, number of threads per tasks. This compute-intensive benchmark has been optimized for exploiting the SIMD instructions on multi-core processors and has been ported for GPU accelerators, therefore, it can benefit from nodes with a high number of cores with SIMD instructions and GPUs.
- NEURON/CoreNEURON
  - Reported results: simulation time reported in seconds
  - Parameters influencing performance results: type of test case (ring or traub), number of cores, number of nodes. NEURON is sensitive to the memory sub-system configuration, while CoreNEURON has been optimised SIMD optimization and GPU accelerators.
- Neuroimaging Deep Learning
  - Reported results: number of k-sample per second
  - Parameters influencing performance results: number of nodes, number of tasks per node, and the usage of GPU devices. This benchmark exhibits key characteristics of widely used frameworks like TensorFlow and

Horvord. As a result, it is compute intensive and is sensitive to vectorization, memory bandwidth and inter-node bandwidth.

- TVB-HPC
  - Reported results: simulation time reported in seconds
  - Parameters influencing performance results: Number of nodes and number of MPI tasks per node. The application demonstrates high inter and intra node scaling on multicore architectures. The benchmark does not exploit SIMD vectorization effectively.
- TensorFlow
  - Reported results: number of images per second
  - Parameters influencing performance results: these are rather versatile sets of benchmarks that can be therefore impacted by several system and software configuration parameters.
- IOR
  - Reported results: read and write bandwidth reported as MB/sec
  - Parameters influencing performance results: this is a versatile benchmark with several configuration parameters enabling evaluation of file systems under different payload. Several system and software stack configurations can impact performance such as the high speed network, the storage system configuration, bandwidth to individual storage servers and load balancing across parallel file system targets.
- Cosbench
  - Reported results: throughput as operations per second and bandwidth as GB/sec
  - Parameters influencing performance results: a versatile benchmark that can be configured for read and write operations of different object sizes. The hardware and object storage configurations also impact achievable performance such as network configuration, number of clients, storage backends, and storage hardware.

2.2 JUELICH Validation Results

The results are collected from the JUSUF Cluster partition, which is used for SCC and IAC workloads. JUSUF Cluster and Cloud are composed of 205 compute nodes in total, each of them equipped with two AMD EPYC Rome 7742 64-core CPUs, 256 GB DDR4 memory, and one 800GB NVMe used for local scratch. A partition of 61 of these nodes provides also an additional Nvidia Volta V100 GPU with 16 GB high-bandwidth memory. The nodes in the cluster are interconnected with a 100 Gb/s Nvidia Mellanox HDR100 high-speed interconnect in a full fat-tree topology. In addition, a 40 Gigabit Ethernet connection is used for access to JSC's central storage infrastructure.

The benchmarks have been run by the vendor as part of the acceptance tests for the JUSUF Cluster.

Elephant Asset

Comparing the former results from JURON and JUWELS presented in [28] to JUSUF Cluster, the AMD EPYC Rome 7742 64-core CPUs available in the nodes demonstrate a higher scaling efficiency, resulting in lower simulation runtime, with 128 tasks per node. The application particularly benefits from the high core count and high memory bandwidth ratios of the x86 based, 64 core processors.

Nodes	Number of tasks per node	Number of threads per task	Simulation time [sec]	Power Envelope [Watt]
1	128	1	57.24	559

NEST

Comparing these results from JURON and JUWELS presented in [28], JUSUF Cluster with its AMD EPYC processors is significantly faster than JUWELS. Power efficiency has been reported in addition to the performance results. The performance also benefits from a memory sub-system that supports irregular memory access.

Nodes	Number of tasks per node	Number of threads per task	Simulation time [sec]	Power Envelope [Watt]
2	128	2	6.27	729

TVB-HPC

Since the benchmark has been under active development, a direct comparison cannot be made with the results from JURON and JUWELS presented in [28]. For comparison, the runtime of this benchmark is 32 seconds on JUWELS when using up to 96 tasks per node. The result below indicates a higher performance on JUSUF (14.73 sec of runtime) due to a higher available core count and thus more tasks.

Nodes	Number of tasks per node	Number of threads per task	Simulation time [sec]	Power Envelope [Watt]

6	128	1	14.73	646

Arbor

Comparing these results from JURON and JUWELS presented in [28], two AMD EPYC Rome 7742 64-core CPUs and feature 256 GB memory nodes demonstrate a higher scaling efficiency, resulting in lower simulation runtime on 2 nodes with 128 tasks per node. The results benefit from multiple characteristics, particularly taking advantage of SIMD vectorization on CPU cores of the x86 based multi-core nodes and high memory bandwidth.

Nodes	Number of tasks per node	Number of threads per task	Simulation time [sec]	Power Envelope [Watt]
2	128	2	44.69	644

Neuro-imaging Deep Learning Benchmark

Comparing these results from JURON and JUWELS presented in [28], we observe expected performance gains when using the new generation of GPU devices, P100 as compared to V100 for the benchmark results. An additional contributing factor to these performance gains could be a higher internode bandwidth for the ICEI system.

Nodes	Number of tasks per node	Number of threads per task	Time per kilo sample [sec]	Power Envelope [Watt]
2	1	1	4.82	575

IOR Results

Here we try to measure the overall systems IO bandwidth that JUSUF has with its potentially 205*40GE (1,05TB/s) Ethernet based network connection in direction of our central storage system JUST. JUST itself is able to handle up to 500GB/s.

The following results were collected on 180 nodes, with 12 MPI tasks per node and 3 repetitions.

Nodes	Number of tasks per node	Number of threads per task	Write MiB/s	Read MiB/s

180	12	1	381362	239341

The focus was on testing the overall bandwidth, using 16MB blocksize, the POSIX interface, one file per process with independent sequential writes.

2.3 CEA Validation Results

Validation of Interactive Computing Resources

ICEI resources at CEA include Interactive Computing with Linux servers.

These servers are equipped with:

- NVidia V100 GPU embedding 32GB of memory;
- 2 processors Intel Cascade Lake G-6240 at 2,6GHz, each processor has 18 cores;
- 384 GB of DDR4 memory (2933 MT/s).

Additionally, 2 large memory nodes are available. They include 4 processors (same model as described above) and 3072 GB of DDR4 memory (2933 MT/s). These nodes are made of 2 bi-socket modules (Bull X804 chassis).

All nodes are interconnected with an Infiniband 100Gbits network.

The NEURON and CoreNEURON results demonstrate advantages of large memory node configurations and contemporary multi-core processors resulting in high scaling efficiencies, from a single core to the full node. Thus, NEURON shows an excellent scalability on this system: when running on all 36 cores of a node, it reaches up to 88% of the theoretical scalability. CoreNEURON already takes advantage of the GPU when running on a single core, hence a lesser speed-up when running on the full node (still 57% improvement due to the CPU speed-up).

Similar to CoreNEURON, the neuroimaging deep learning benchmark benefits from HPC multi-core and GPU node configuration.

NEST results of 4 nodes with dual-socket multi-core servers (18 cores) demonstrate high performance efficiency, as well as the efficient network communication between compute nodes.

Table 5: ICEI application benchmark validation results from the CEA infrastructure

Computational benchmarks				
Code	**Test case**	**Resources**	**Execution time**	**Performance compared to perfect scalability**
NEURON	Ring	1 core	0.58s	-
NEURON	Ring	1 node	0.02s	81%
NEURON	Traub	1 core	4.74s	-
NEURON	Traub	1 node	0.15s	88%
CoreNEURON	Ring	1 core	0.464s	-
CoreNEURON	Ring	1 node	0.0226443s	57%
CoreNEURON	Traub	1 core	1.42728s	-
CoreNEURON	Traub	1 node	0.0690794s	57%
NEST		4 nodes	7.259s	-
NeuroImaging deep learning		2 nodes	4.768s/ksamples	-

Validation of the Active Data Repository

The active data repository (ACD) at CEA consists of a Lustre parallel file system fully made of SSD devices.

It is composed of 2 DDN SFA18KXe module. Each module is made of two controllers for parallelism and high-availability. Each of these modules embeds 84 SSD drives of 15.36TeraBytes. Drives are connected through SAS 12Gbits interfaces. Drives can endure 1 full Drive-Write-Per-Day (DWPD).

Each module can run up to 8 Lustre servers to provide a high level of I/O parallelism.

The IOR benchmark results show high throughput at different scales, especially demonstrating efficiencies when I/O tasks are spread across a large number of nodes.

This will ensure that the I/O system can handle throughput of large I/O intensive jobs as well as multiple smaller jobs running concurrently on the system. The results are repeated using the same number of MPI tasks with a different number of nodes with different IOR configurations. Details of IOR configurations are available from [22]. Results are shown for different interfaces such as POSIX and HDF5 formats to evaluate the performance for various I/O middleware. Some configurations also evaluate the performance for small non-aligned accesses, which has been stated for many of the neuroscience use cases. Here is a brief description of a couple of IOR options or runtime flags:

- -a api – API for I/O [POSIX|MPIIO|HDF5|HDFS|S3|S3_EMC|NCMPI|RADOS]
- -E useExistingTestFile – do not remove test file before write access
- -C reorderTasksConstant – changes task ordering to n+1 ordering for readback
- -Q taskPerNodeOffset for read tests use with -C & -Z options
- -g intraTestBarriers – use barriers between open, write/read, and close
- -t transferSize – size of transfer in bytes (e.g.: 8, 4k, 2m, 1g)
- -b blockSize – contiguous bytes to write per task (e.g.: 8, 4k, 2m, 1g)
- -k keepFile – don't remove the test file(s) on program exit

Together these flags enable controlling overheads for reading and writing to files, payload sizes and total file sizes.

Table 6: Storage performance results with IOR

Test	# nodes	# MPI tasks	Write MB/s	Read MB/s	Used IOR options
Test_1 (measured)	24	192	59271	56034	-E -C -Q 1 -g -t 4m -b 8g -a POSIX -k
Test_2 (measured)	4	192	2475	9460	-E -C -Q 1 -g -t 47008 -b 47008 -s 100000 -a POSIX -k
Test_3 (measured)	32	320		133783	-E -a POSIX -C -Q 16 -g -t 4m -b 32g -r -k
Test_4 (measured)	16	704		23416	-E -a POSIX -C -Q 1 -g -t 47008 -b 47008 -s 100000 -r -k

Test_5 (measured)	16	320		138317	-E -a POSIX -C -Q 1 -g -b 22G -t 128k -F -r -k
Test_6 (measured)	20	320	20714	39020	-E -a HDF5 -C -Q 1 -g -b 8G -t 2048 -k
Test_7 (measured)	16	384		1631	-E -g -a POSIX -b 4656M -t 2328 -r -z -k
Test_110 (measured)	16	256	118253	136524	-E -a POSIX -C -Q 8 -g -b 64G -t 1M -F -r -w -vv

Validation of the Archival Data Repository

The archival data repository at CEA runs OpenIO, an open-source parallel object store with Swift gateways.

This system is made of a DDN SFA18KXe module with 10 extension drawers to hold the 650 hard drive disks of the system. Each disk has a capacity of 14TeraBytes.

3 bare-metal servers are dedicated to run the metadata services of OpenIO as well as the Swift gateways. This allows cumulating the bandwidth of the three servers (100Gbits each) to increase the total throughput of the system.

Archival data repository results have been collected by running the **Cosbench** test cases.

These results show a balance throughput for object accesses with an average object size of 1 Gbyte. This shows that the system can reach a high throughput of 22GBytes/sec for write operations and 34GBytes/sec for read operations. This is close to the maximum bandwidth that can be obtained using 3 Swift gateways.

This performance is much higher than the throughput allowed by the network connections between the Fenix sites (high speed connexions of 10Gbits allow an effective throughput of 1GByte/sec). Thus, remote access to data from the ARD can be done using the full speed of the physical link to the Internet. Besides this bandwidth is also suitable with the high-speed network in the compute centre, which is 100Gbits (allows an effective throughput of 10GBytes/sec per client).

Write throughput:

Configuration: 300 workers distributed on 3 client machines

Op-Type	Op-Count	Byte-Count	Avg-ResTime	Avg-ProcTime	Throughput	Bandwidth
write	886 ops	886 GB	13008.97 ms	385.18 ms	22.3 op/s	22.3 GB/S

Read throughput:

This test is done in 2 steps: 1) create test data ("prepare-write" step); 2) read the test data ("read" step).

Configuration: 400 workers distributed on 3 client machines

Op-Type	Op-Count	Byte-Count	Avg-ResTime	Avg-ProcTime	Throughput	Bandwidth
prepare - write	300 ops	300 GB	21012.42 ms	324.39 ms	14.73 op/s	14.73 GB/S
read	1.86 kops	1.86 TB	8765.05 ms	48.72 ms	34.01 op/s	34.01 GB/S

2.4 CINECA Validation Results

CINECA procured an integrated system to provide SCC, IAC, VM, ARD and ACD services, named Galileo100. Results were collected using the Galileo100 partition providing SCC and IAC services, featuring a total of 554 computing nodes, each with two CPU Intel CascadeLake 8260, 24 cores at a base nominal frequency of 2.4 GHz (3.90 GHz turbo), and 384GB RAM. This node partition is divided in:

- 340 standard nodes with 480 GB SSD;
- 180 data processing nodes with 2TB SSD, 3TB Intel Persistent Memory (Optane);
- 34 GPU nodes with 2x NVIDIA GPU V100 with 100Gbs InfiniBand interconnection and 2TB SSD.

The procurement procedure provided a benchmark suite to assess the offered solution. The suite was composed of the following test cases:

Application	**Download link**
IOR	https://b2drop.eudat.eu/s/96r6fojp55nyN9c/download
NEURON	https://b2drop.eudat.eu/s/bXQyxrB6rHyNqqB/downloadhttps://b2drop.eudat.eu/s/bXQyxrB6rHyNqqB/download
NEST	https://b2drop.eudat.eu/s/M44Edt5oo4i4F4Y/download

Each procurement candidate provided an assessment of the benchmarks for their offered solution, committing on the benchmark values obtained. The values of the winning solution were validated during acceptance of the system. The validation results are reported below. Please note that the procurement application suite included also ALIQUIS, a test case not part of the ICEI benchmarks. This test case was deemed relevant for the neuroscience community, thus motivating its inclusion in the application suite for the procurement. Results are not shown for this test case since they are out of scope of this document.

NEURON results

Benchmarks demonstrated adequate scaling efficiencies for the use case, from a single to two nodes. The NEURON results demonstrate advantages of large memory node configurations and contemporary multi-core processors resulting in high scaling efficiencies, from a single core to the full node.

Nodes	Tasks per Node	Elapsed Time [s]	Real Time [s]
1	48	122.77	129.36
2	48	66.91	75.81

NEST results

Benchmarks tested the efficiency of the system and interconnection. The table below shows the best configuration (2 MPI processes per node, 24 task per process). The scaling efficiency and performance of the benchmark demonstrate that the targeted infrastructure benefits the memory access patterns of the application.

Nodes	Tasks per Node	Thread per Task	Elapsed /s
1	2	24	52.24
2	2	24	24.41
4	2	24	12.26
8	2	24	6,83
16	2	24	4.2
32	2	24	2.96

IOR results

The ACD at CINECA is provided with a LUSTRE parallel file system. Benchmarks tested the I/O efficiency. Five tests have been performed to measure the high throughput at different scales. IOR use cases were set to mimic patterns of HBP scientific applications.

IOR command argument to write a file is *-w*, while the one to read a file is *-r*. The performed tests were set to use barriers between open, write/read and close status (-g option). Tests performed 1 iteration only (-i option) with POSIX interface (-a option). The block size (-b option) set the contiguous bytes to write/read per task, while the transfer size (-t option) set the size of transfer in byte. All tests to measure the write performance avoided to remove the test file (-k option), as all tests to measure the read performances avoided to remove the test existing file before the write access (-E option).

Test 1 – The test measured MiB/s to write/read files with a block size of 2 GB setting the transfer size at 4 MB. Every MPI process perform I/O to a unique file (-F option).

- write: ior -g -i 1 -a POSIX -w -k -F -b 2g -t 4m
- read: ior -g -F -i 1 -a POSIX -E -r -b 2g -t 4m

Nodes	1	2	4	8	16	32
Write (MiB/s)	8432,7	16001,7	27954,4	18738,9	18738,9	52830,2
Read (MiB/s)	10583,8	19307	37049,4	51382,5	55657,9	59729,7

Test 2 – The test measured MiB/s to write/read files with a block size of 2 GB setting the transfer size at 4 MB.

- write: ior -g -i 1 -a POSIX -w -k -b 2g -t 4m
- read: ior -g -i 1 -a POSIX -E -r -b 2g -t 4m

Nodes	1	2	4	8	16	32
Write (MiB/s)	701	2596,3	2871,4	10440,1	18496,8	20345,6
Read (MiB/s)	2832,8	3427,2	11556,2	12945,7	24430,1	37542,4

Test 3 – The test measured MiB/s to write/read files with a block size of 48 KB setting the transfer size at 48 KB, and the number of segments (-s option) equal to 5462.

- write: ior -g -i 1 -a POSIX -w -k -s 5462 -b 48k -t 48k
- read: ior -g -i 1 -a POSIX -E -r -s 5462 -b 48k -t 48k

Nodes	1	2	4	8	16	32
Write (MiB/s)	15396,1	30930,8	52598,8	50827	58284,4	58789,2
Read (MiB/s)	9614,7	15722,2	24225,2	44664,1	53968,9	55691,3

Test 4 – The test measured MiB/s to write/read files with a block size of 2 GB setting the transfer size at 128 KB.

- write: ior -g -i 1 -a POSIX -w -k -b 2g -t 128k
- read: ior -g -i 1 -a POSIX -E -r -b 2g -t 128k

Nodes	1	2	4	8	16	32
Write (MiB/s)	1367,9	2797,9	4916,8	10334,1	20218,1	40885,8
Read (MiB/s)	2974,9	5704,1	11474,9	21148,5	35618,4	37771,8

Test 5 – The test measured MiB/s to write/read files with a block size of 582 MB setting the transfer size at 4656 bytes.

- write: ior -g -i 1 -a POSIX -w -k -b 582m -t 4656
- read: ior -g -i 1 -a POSIX -E -r -b 582m -t 4656

Nodes	1	2	4	8	16	32
Write (MiB/s)	449,4	881,1	1672,7	3267,8	6524,9	9764
Read (MiB/s)	45,1	85,9	188,8	296,2	173	333,4

2.5 BSC Validation Results

The results are collected from the archive storage system and the BSC-CNS interactive computing partition, which is composed of 2 nodes that are equipped with two Power9 CPUs and feature 512 GB DDR3 memory and with 2 Nvidia Volta V100 GPU with a 16 GB high-bandwidth memory available.

For each of the applications that were described in the tender for the infrastructure acquisition by BSC, those were executed to validate expected results. In the following we report on the results and provide a short analysis.

NEURON shows an improvement in efficiency thanks to the large memory node configuration of the BSC system. In addition, the multi-thread capability available in the power9 processors (up to 4 threads per core) increases the performance of the NEURON benchmark when running with 80 threads.

For the Elephant benchmark and NEST, the benefit of using the BSC system comes from its good memory bandwidth and the large memory capacity of the system, but is limited in this case by the number of cores available per node and the limited number of nodes on the cluster. As the system was configured to maximize the I/O features and the usage of large size problems thanks to the 512GB of main memory, the system will be able to perform these kind of simulations with good performance and efficiency.

In terms of IOR benchmarks, it can be seen that the interactive nodes are able to fully utilize all the bandwidth available for them: 2x100GB per each of the interactive nodes. Main components of the BSC tender were focused on the acquisition of a big and performant archive storage system, which was validated by the results of IOR benchmarks, shown below. It is out of the scope of this benchmark validation tests to validate the performance of the tape tier, as it cannot be triggered directly from an IOR execution.

NEURON

	Input	Nodes	Number of threads per node	Simulation time [sec]
Baseline node	ring	1	1	1,02
1 node - 40 threads	ring	1	40	0,15
1 node -	ring	1	80	0,075

80 threads				
1 node - 160 threads	ring	1	160	0,14

NEST

	Nodes	Number of tasks per node	Number of threads per task	Elapsed time [sec]
Baseline	1	40	4	54
2 nodes	2	40	4	36,12

Neuro-imaging Deep Learning Benchmark

	Nodes	Number of tasks per node	Time per kilo sample [sec]
Baseline	1	2	9,33

Elephant ASSET

	Nodes	Number of tasks per node	Elapsed time [sec]
Baseline	1	40	77,53

IOR Results

Results are collected on 2 nodes, with 24 MPI tasks per node.

Summary of tests: Max Write: 4465.98 MiB/s Max Read: 50256 MiB/s

2.6 CSCS Validation Results

CSCS resources have been made available since 2018 through an allocation mechanism, therefore, these did not follow the procurement and validation process. Instead, the mapping of use cases was assessed through the production allocation to the neuroscience projects. Table 7 lists a subset of ICEI project allocations for HBP applications and teams alongside with the coverage of use cases. Details of usage are available in internal reports [6][9].

Table 7: An overview of selected ICEI allocations on CSCS resources (2018-2020)

Example Allocated Project on CSCS Resources	Use Case Coverage
Full-scale hippocampus model	6
Cerebellum single cell optimizations	5
Neurorobotics Platform (NRP) development	11
ilastik as a service	13
Virtual Brains Projects	10
End-to-end learning to grasp in the Neurorobotics Platform	3
Biological Deep Learning	16
Atomistic Molecular Dynamics simulations for relevant signal transduction proteins (MoDEL_CNS)	New use case
Basal Ganglia Circuits	New use case
Computational models of multisensory integration	New use case
QUINT workflow	New use case

Overall, projects allocated at CSCS are using applications that are represented in the ICEI benchmarks:

- NEST – 3 projects
- TVB-HPC – 3 projects
- NEURON/CoreNeuron – 5 projects
- Neuroimaging Deep Learning/TensorFlow – 5 projects

Table 8 presents results for Arbor benchmarks using two ICEI resources namely Piz Daint GPU partition and Piz Daint multi-core (mc) partitions plus a Xeon Phi partition [24].

Table 8: Arbor benchmarking results from different platforms. Single node results. Daint-mc 2 MPI ranks with 36 threads, Daint-gpu 1 rank with 24 threads, Tave-knl cache mode with4 MPI ranks with 64 threads (4 per core)

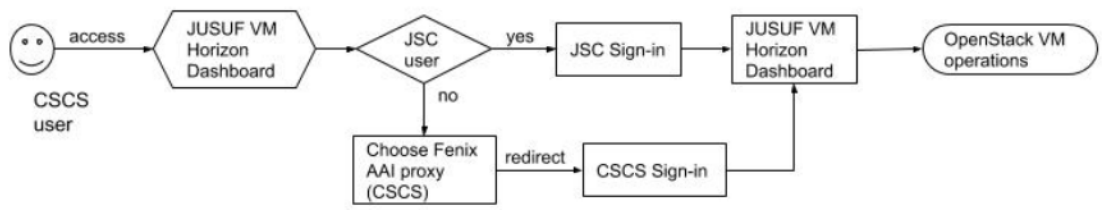
	wall time (s)				energy (kJ)		
cells	mc	gpu	knl	nrn	mc	gpu	knl
32	**0.35**	2.06	1.13	1.73	**0.04**	0.25	0.17
64	**0.39**	2.10	1.29	2.61	**0.05**	0.25	0.22
128	**0.75**	2.44	1.71	8.27	**0.11**	0.33	0.34
256	**1.42**	2.97	2.28	32.92	**0.26**	0.43	0.55
512	**2.66**	4.19	3.36	67.33	**0.58**	0.67	0.97
1024	**5.12**	6.50	6.15	135.52	1.24	**1.14**	1.81
2048	**10.04**	11.11	12.27	272.87	2.53	**2.11**	3.63
4096	**19.93**	19.96	24.39	555.34	5.16	**3.96**	7.24
8192	39.66	**37.24**	48.65	1234.70	10.38	**7.72**	14.45
16384	79.22	**71.65**	97.19	–	20.85	**15.11**	28.99

Additional results on Piz Daint storage are available from an evaluation paper [25]. These fulfil the requirements identified for the ICEI use cases for being able to execute and scale the application on ICEI multi-core and GPU scalable resources.

2.7 Fenix AAI Validation Results

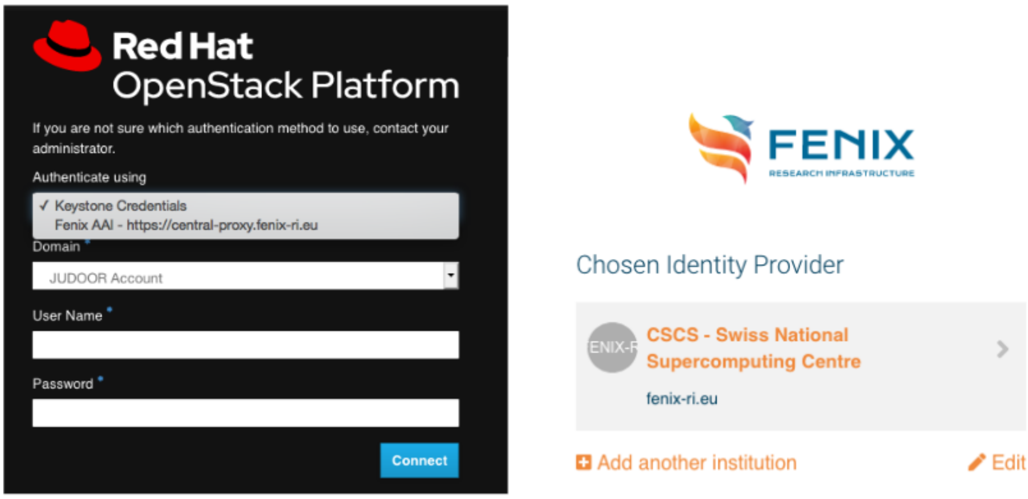
Workflow for validating Fenix AAI implementation, which is an R&D service that has been developed as part of the ICEI project, is shown in figure 1. The use case considered here is implementing a high availability virtual machine where a service can be replicated, load balanced or migrated in case of scheduled and unscheduled downtimes.

Figure 1: Workflow for Fenix AAI validation. This workflow assumes an active quota (resources) to access JUELICH OpenStack system called JUSUF (https://jusuf-cloud.fz-juelich.de/dashboard/auth/login)



A demonstration of this workflow has been presented at the HBP SGA3 Fenix AAI and FURMS workshop (November 5, 2020). Figure 2 demonstrates screen capture of the workflow where a user from one site accesses OpenStack VM resources to another Fenix site. No new user registration is required. As shown in figure 2, a user simply selects an authentication mechanism and is then redirected to the Fenix AAI proxy implementation site where the user selects the primary site for their IdP.

Figure 2: Screen views of the workflow presented in figure 1



3. Next Steps and Updates (2022- Q3/2023)

While this report is a snapshot of the ICEI infrastructure as of January 2022, it highlights the best practices and next steps to continue for the remainder of the project. The best practices include:

- Packaging of self-contained tests
- Documentation of tests and metrics
- Availability of a test registry (with access control)
- Continuous development, testing and deployment (usage of software development and deployment tools)

These approaches allow an independent comparison of capabilities of the existing and upcoming ICEI infrastructure. At the same time, these tests can ensure and track capabilities throughout the lifetime of the infrastructure as it goes through the upgrades of the user and programming environment such as the operating system, compilers, parallel libraries like MPI, etc. Likewise, upgrades to the application environment such as new releases of the software can be incorporated and distributed for continuous testing and development. Individual sites, platform teams and application developers can exploit these best practices for regression testing to proactively mitigate any impact to scheduled and unscheduled changes to the software stacks of applications and systems. Where feasible, we try aligning and leveraging best practices, such as the toolchain by the US Exascale Computing Project (ECP) for software ecosystem and delivery for scientific workflows across diverse IT infrastructure including HPC [26]. Individual ICEI sites have adopted tools for continuous integration and testing, which were reported in a HBP deliverable [27].

We therefore envision three types of tests going forward:

- Singular tests (manually). This will be similar to the validation tests and results presented in this report for BSC, CEA, CINECA and JUELICH procurements.
- Sanity tests (to be performed after major updates and interventions, manually or automated). ICEI sites run a series of regression tests as part of scheduled and unscheduled interventions. Small scale benchmarks can be included to ensure validation results are reproducible. Some tests can also be involved for debugging and troubleshooting user reported issues for applications such as NEST, CoreNEURON, etc.
- Periodic tests (monitoring tests, automated). Some lightweight tests can be included to monitor essential functionality of a resource. Typically, sites run these scripts (few seconds in most cases) before and after submission of a job to check health of CPU, memory, GPU, etc.

Table 9 lays out a plan for continued efforts for the remainder of the ICEI project time frame. These activities include reporting on resource provisioning, allocation and consumption as well as output of validation for outstanding R&D services.

Table 9: Plans for the remaining time frame for the ICEI project

	2022	**September 2023**
Related Deliverable and Milestone	D4.12 (Data storage and compute provisioning during M37 – M48)	D4.13 (Data storage and compute provisioning during M49 – M60) D4.16 (Data storage and compute provisioning during M61 - M69)
Validation outputs	FURMS validation results Data Mover validation results	

References

[1]. ICEI D3.4: Validation Framework

[2]. ICEI D3.6: Scientific Use Case Requirements Documentation

[3]. ICEI D3.1: Common Technical Specifications

[4]. ICEI D4.1: Tender Documents (Part 1)

[5]. ICEI D4.2: Infrastructure at ETHZ/CSCS

[6]. ICEI D4.9: Data storage and computing provisioning during M1-M12

[7]. ICEI D3.2: Initial Federated AAI Infrastructure

[8]. ICEI D4.15: Tender Documents (Part 2)

[9]. ICEI D4.10: Data storage and computing provisioning during M13-M24

[10]. ICEI D4.3: Infrastructure at JUELICH-JSC

[11]. ICEI D4.4: Infrastructure at CEA

[12]. HBP SGA2 D7.1: HPAC Platform Architecture Specification v1

[13]. HBP SGA2 D7.2: HPAC Platform Architecture Specification v2

[14]. HBP Benchmark suite for technology transfer: https://wiki.ebrains.eu/bin/view/Collabs/hbp-benchmark-suite-for-technology-trans/

[15]. Elephant ASSET benchmark: https://gitlab.version.fz-juelich.de/benchmarks/elephant.git

[16]. NEST benchmarks: https://gitlab.version.fz-juelich.de/benchmarks/NEST.git

[17]. Arbor benchmark: https://gitlab.version.fz-juelich.de/benchmarks/arbor.git

[18]. Neuron/CoreNeuron benchmark: https://gitlab.version.fz-juelich.de/benchmarks/coreneuron.git

[19]. Neuroimaging deep learning benchmark: https://gitlab.version.fz-juelich.de/benchmarks/neuroimg-deeplearning.git

[20]. TVB-HPC: https://www.thevirtualbrain.org/

[21]. TensorFlow benchmarks: https://github.com/tensorflow/benchmarks/tree/master/scripts/tf_cnn_benchmarks

[22]. IOR benchmark: https://github.com/LLNL/ior

[23]. Cosbench benchmark: https://github.com/bissenbay/cosbench

[24]. Akar, Nora & Cumming, Ben & Karakasis, Vasileios & Küsters, Anne & Klijn, Wouter & Peyser, Alexander & Yates, Stuart. (2019). Arbor — A Morphologically-Detailed Neural Network Simulation Library for Contemporary High-Performance Computing Architectures. 274-282. 10.1109/EMPDP.2019.8671560.

[25]. "An Operational Perspective on a Hybrid and Heterogeneous Cray XC50 System", In proceedings of the Cray User Group (CUG) meeting, 2017.

[26]. Exascale Computing Project (ECP) Software Ecosystem and Delivery: https://www.exascaleproject.org/research-group/software-ecosystem/

[27]. HBP SGA2 D7.3.1 Multi-compartment benchmark and validation suites on HPC systems.

[28]. Andreas Herten, Thorsten Hater, Wouter Klijn, and Dirk Pleiter. 2019. Performance Comparison for Neuroscience Application Benchmarks. In *High Performance Computing: ISC High Performance 2019 International Workshops, Frankfurt, Germany, June 16-20, 2019, Revised Selected Papers*. Springer-Verlag, Berlin, Heidelberg, 418–431. DOI: https://doi.org/10.1007/978-3-030-34356-9_31

[29]. ICEI D4.5: Infrastructure at BSC

[30]. ICEI D4.6: Infrastructure at CINECA

[31]. ICEI D4.11: Data storage and computing provisioning during M25-M36

[32]. ICEI D4.7: Report on deployed infrastructure

[33]. ICEI D4.8: R&D results

Appendix 1

Tag Name	Arbor bench
Related Use Case(s)	1 (Data-driven cellular models of brain regions, Olfactory Bulb) and 10 (Large scale multi-scale co-simulation of the cortex (TVB <-> Nest <-> Arbor))
Identifier Number TID#	Arbor-validate-0.1
Release #.#	0.1
Owner(s)	Arbor development team
Component	SCC, IAC, NETI, ACD
Site(s)	JUELICH, CSCS
Description	Arbor is a simulation library for networks of morphologically detailed neurons. Two benchmarks are designed to test these two parts of the workflow, a computationally intensive Ring benchmark and network and memory intensive Proxy cell benchmark
Command/code	https://gitlab.version.fz-juelich.de/benchmarks/arbor.git
Output	Wall time (seconds)
Metric(s)	Runtime