FENIX

RESEARCH INFRASTRUCTURE

D4.7

Report on deployed infrastructure

Work package:	WP4 - Procurement, deployment and operation	
Author:	Thomas Leibovici	CEA
Reviewer #1	Colin McMurtrie	ETHZ/CSCS
Reviewer #2	Dirk Pleiter	JUELICH
Dissemination Level	Public	
Nature	Report	

Date	Author	Comments	Version	Status
02.06.2021	Thomas Leibovici	Document structure	V0	Draft
19.07.2021	Thomas Leibovici	Initial Draft	V0.1	Draft
21.07.2021	Thomas Leibovici	All sections complete	V0.2	Draft
05.08.2021	Thomas Leibovici	Added sections 2.2 and 2.3	V0.3	Draft
19.08.2021	Thomas Leibovici	Updated resources	V0.4	Draft
18.09.2021	Colin McMurtrie	Review	V0.4	Draft
19.10.2021	Dirk Pleiter	Review	V1	Draft
20.10.2021	Thomas Leibovici	Submission ready	V1.1	Draft
28.10.2021	Valentina Armuzza	Final editorial updates	V1.2	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 report takes stock of the site specialisation strategy implemented by Fenix, and its few adjustments compared to the initial plan. It summarises the deployed systems and their integration to form the Fenix infrastructure.

Table of Contents

List of Acronyms

AAI	Authentication and Authorization Infrastructure
ACD	Active Data Repositories
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
CSCS	Centro Svizzero di Calcolo Scientifico
DM	Data Mover Service
HBP	Human Brain Project
HPC	High Performance Computing
IAC	Interactive Computing Services
ICEI	Interactive Computing E-Infrastructure for the Human Brain Project
ICN IdP	Interactive Computing Node Trusted Identity Provider
JSC	Jülich Supercomputing Centre
R&D	Research & Development
SCC	Scalable Computing Services
TGCC	Très Grand Centre de calcul du CEA
VM	Virtual Machine Services

1. Introduction

Five High Performance Computing (HPC) centres, namely BSC[1] in Spain, CEA-TGCC[2] in France, CINECA[3] in Italy, ETHZ-CSCS[4] in Switzerland and JUELICH-JSC[5] in Germany, committed to set up and to operate a new set of federated e-infrastructure services within the ICEI project.

After analysing the use case of the HBP (see [D3.6] "Scientific Use Case Requirements Documentation"), common technical requirements were defined in deliverable D3.1 "Common Technical Specifications" [D3.1] that shaped the following services:

- Scalable Compute Services (SCC);
- Interactive Compute Services (IAC);
- VM Services (VM);
- Active Data Repositories based on fast memory and active storage tiers (ACD);
- Archival Data Repositories (ARD);
- Data mover services.

To implement those services, the aforementioned sites conducted procurements to complete their existing infrastructure with required new hardware. In addition to the hardware, R&D services were procured to implement federation services. The tender packages for these new systems and software components have been previously described as part of deliverables [D4.1] and [D4.15].

The deployed infrastructure of each site was described in the following deliverables, submitted as systems were put into service:

- D4.2 Infrastructure at ETHZ/CSCS, submitted in October 2018;
- D4.3 Infrastructure at JUELICH-JSC, submitted in April 2020;
- D4.4 Infrastructure at CEA, submitted in July 2020;
- D4.5 Infrastructure at BSC, submitted in July 2021;
- D4.6 Infrastructure at CINECA, submitted in September 2021.

Besides this, the implementation status of the procured software developments was described in [D4.8].

This document summarizes the deployment strategy, the resources now available in Fenix, and how they are integrated to form the Fenix e-Infrastructure.

2. Deployment strategy

2.1. Initial deployment strategy

Site specialization has been considered to facilitate the integration of applications and workflows that were targeted at the different sites.

[1] https://www.bsc.es/
[2] http://www-hpc.cea.fr/en/complexe/tgcc.htm
[3] https://www.cineca.it/
[4] https://www.cscs.ch/
[5] https://www.fz-juelich.de/ias/jsc/EN/Home/home_node.html

This specialization has been driven by foreseeing for each science case one main site. Thus, the defined sites are primarily responsible for integrating the use-case into the infrastructure. The owners of that science-case or use-case act as lead customers for a given site, which helps realising the local infrastructure and making it more mature.

The detailed matching between science-cases and sites is detailed in [D3.1].

The choice for site specialisation has been made based on:

- **Data locality**, to minimize the cost of transferring the large amounts of experimental data from the labs to the Fenix infrastructure;
- **Available Resources**, by assigning projects to sites that provides the most suitable infrastructure for the science-case, also taking into account the balancing of resources across the different Fenix sites;
- **Local expertise**, to help addressing the technical challenges that science-cases may rise.

Based on this, the following table has been established, to drive the kind of use-case each site had to address particularly:

Science-case requirements	BSC	CEA	CINECA	ETHZ	JUELICH
Simulation	✓	✓	✓	✓	✓
Multi-scale coupled simulation				✓	✓
Co-deployment of apps	✓		✓	✓	✓
Streaming visualization	✓			✓	✓
Machine learning		✓	✓	✓	✓
In-the-loop machine learning			✓	✓	
Big data processing		✓	✓		✓
Big data visualization	✓	✓	✓	✓	✓

Table 1 - Targeted use-cases

These requirements have been categorised according to the kind of equipment needed, thus defining the type of service to be implemented by each site, and summarized in the table below:

Service	BSC	CEA	CINECA	CSCS	JUELICH
Scalable Compute Services	✓			✓	✓
Interactive Compute Services	✓	✓	✓	✓	✓
Active Data Repositories	✓	✓	✓	✓	✓

Archival Data Repositories	✓	✓	✓	✓	✓
Data Mover Service	✓	✓	✓		✓
Virtual Machine Service		✓	✓	✓	✓

Table 2 - Planned services per site

2.2. Adjustments in the deployment strategy

During the course of the project, adjustments were made to benefit from technical opportunities to provide extra services, or to adapt the total amount of provisioned resources to the stated needs on sites already in production.

While leading the procurement to acquire hardware for Interactive Computing services, CEA and CINECA had the technical opportunity to select hardware that could also implement Scalable Computing services. Indeed, the procured systems are equipped with a high performance low latency network, so they can also be used to run parallel computations (e.g. based on MPI). Therefore, these two sites can provide Scalable Computing services that were not initially planned in the site specialisation, in addition to Interactive Computing services, which was the originally foreseen focus.

The initial plan was to provide scalable compute services at BSC, but it was identified that those services were already provided via other current infrastructures (PRACE and the national Spanish RES infrastructure). After analysis, virtual machine service was identified as a more complementary service to be provided by BSC to cover new needs from scientific communities and that was not provided at BSC before the ICEI project. The site specialization was therefore supplemented by Virtual machine services.

The Data Mover service is still under development as part of ICEI R&D tenders. The company that develops this service delivers development versions incrementally, but the software is not production-ready at this stage. Therefore, the service is not yet deployed in production at any Fenix site.

The deployment of an Archival Data repository at JSC is in progress. This service will be available at JSC once the XCST storage system has been completed. It should, however, be noted that this is not an ICEI-funded resource and therefore not available through the HBP-ICEI and PRACE-ICEI allocation mechanisms.[6]

The table below summarizes the differences between the planed services, and the services that have been made available to date:

	BSC		**CEA**		**CINECA**		**CSCS**		**JUELICH**	
Provided services	Inital plan	Actually provided	Inital plan	Actually provided	Inital plan	Actually provided	Inital plan	Actually provided	Inital plan	Actually provided
Scalable Compute Services	✓	X		✓		✓	✓	✓	✓	✓
Interactive Compute Services	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

[6] JSC plans to make, however, resources available to selected brain research projects and therefore the integration of this ARD is of value for the HBP community.

Active Data Repositories	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Archival Data Repositories	✓	✓	✓	✓	✓	✓	✓	✓	✓	X
Data Mover Service	✓		✓		✓		✓		✓	
Virtual Machine Service		✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 3 - Provisioned services per site (to date)

2.3. Summary of implemented use-cases to date

The following table summarizes the use-cases implemented on the ICEI infrastructure so far, compared to the initial plan.

	BSC		CEA		CINECA		ETHZ		JUELICH	
Science-case requirements	initial plan	allocated projects	initial plan	allocated projects	initial plan	allocated projects	initial plan	allocated projects	initial plan	allocated projects
Simulation	✓	✓	✓	✓	✓		✓	✓	✓	✓
Multi-scale coupled simulation							✓	✓	✓	✓
Co-deployment of apps	✓	✓			✓		✓	✓	✓	✓
Streaming visualization	✓						✓	✓	✓	✓
Machine learning			✓	✓	✓		✓	✓	✓	✓
In-the-loop machine learning					✓		✓	✓		
Big data processing		✓	✓	✓	✓			✓	✓	✓
Big data visualization	✓		✓		✓		✓	✓	✓	✓

Table 4 - Use-cases implemented on the ICEI infrastructure

In most cases, the different sites started to be used as expected according to the defined site specialisation.

Although not initially planned at BSC and CSCS, Big Data processing has been requested and could be implemented on those sites.

The systems at CINECA have been made available only very recently, so we cannot provide enough perspective on the project allocation on these systems.

The other boxes that are not checked in the table above represent use cases that are still possible for new science cases.

3. Procurement and deployment timeline

3.1. Procurement phases

The procurement of equipment and R&D services has been implemented in 2 phases:

Part 1 of the procurement covered the following topics:

- CEA procured Interactive compute cluster (IAC), active storage, archival storage, OpenStack cluster, data mover nodes and network extensions;
- CINECA procured IAC, active storage, archival storage, and data movers nodes;
- JUELICH procured an Interactive and Cloud Computing Platform, which is now called JUSUF, and a High-Performance Storage Tier (active storage), which was named JUST-IME.

Part 2 of the procurement covered the following topics:

- BSC procured IAC and archival storage;
- BSC, CEA, CINECA, CSCS and JUELICH procured R&D services.

3.2. Deployment at CSCS

Thanks to a forward planning ETHZ/CSCS was able to provide all co-financed resources from April 2018. Available resources on the Piz Daint supercomputer made it possible to provide scalable computing services to HBP and PRACE from the very beginning of ICEI project. Once the ICEI project was approved, ETHZ/CSCS procured the necessary additional scalable and interactive computing resources. These resources were installed and made available in the mid-September 2018.

Figure 1 - The Piz Daint supercomputer at CSCS

3.3. Deployment at JSC

After the validation of the tender documents as part of deliverable [D4.1], JUELICH-JSC started the procurement of infrastructure components for the ICEI project.

Two major installations, the JUSUF compute and cloud platform and the JUST-IME high-performance storage systems, took place in late 2019 and early 2020. These infrastructure components were made available to the Fenix users from April 2020.

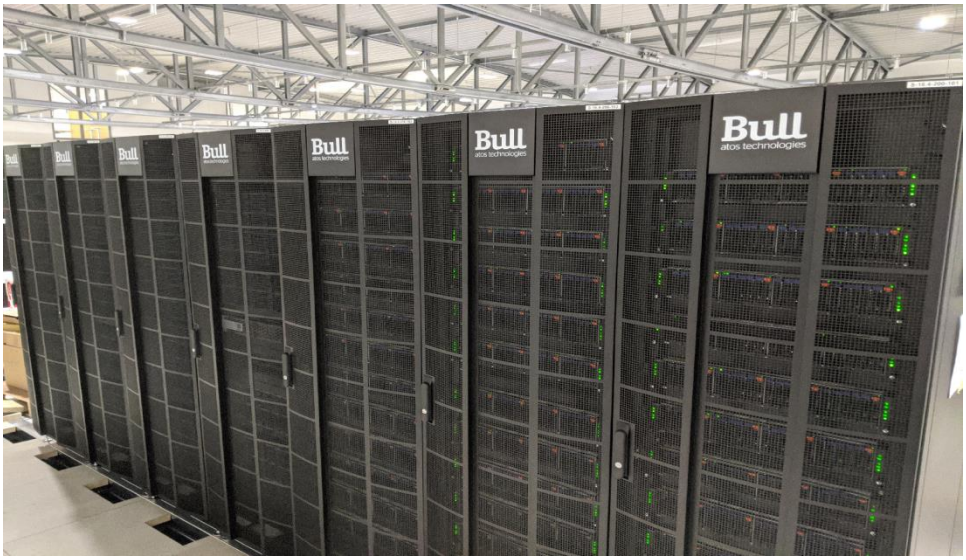

Figure 2 - The JUSUF compute and cloud platform at JSC

3.4. Deployment at CEA

After the validation of the tender documents as part of deliverable [D4.1], CEA led a competitive dialog in 2019 to acquire the required systems for Fenix, including an interactive computing cluster, an OpenStack cluster, as well as active and archive storage.

Delivery took place in the first half of 2020 and systems were made available in July 2020 (with some delays due to COVID-related measures).

Figure 3 - Fenix systems at CEA

3.5. Deployment at BSC

Tender documents for Fenix systems at BSC were validated as part of [D4.15] in September 2019.

The tender procedure took place late 2019 and was divided in 2 lots:

- Interactive compute cluster
- Archival Data repository

The contract for lot 2 was awarded first, in January 2020. The delivery and installation of this lot started in June 2020 (with a delay of 3 months due to COVID measures), and was opened to users in June 2021.

The contract for lot 1 was awarded later, in December 2020, and the corresponding hardware was delivered and installed starting in January 2021 and was opened to first users on June 2021.

Figure 4 - Nord3 compute racks at BSC

3.6. Deployment at CINECA

The tender documents for the Fenix systems at CINECA were validated as part of [D4.15] in September 2019. CINECA started the procurement of the system for the ICEI project late 2019 with the goal of provisioning an infrastructure particularly devoted to data analysis and storage, scalable computing and cloud computing. The procurement procedure was completed by the end of 2020, with a delay primarily due to the tender suspension caused by the COVID19 pandemic. Delivery of the system started in the first half of 2021 and was made available in September 2021.

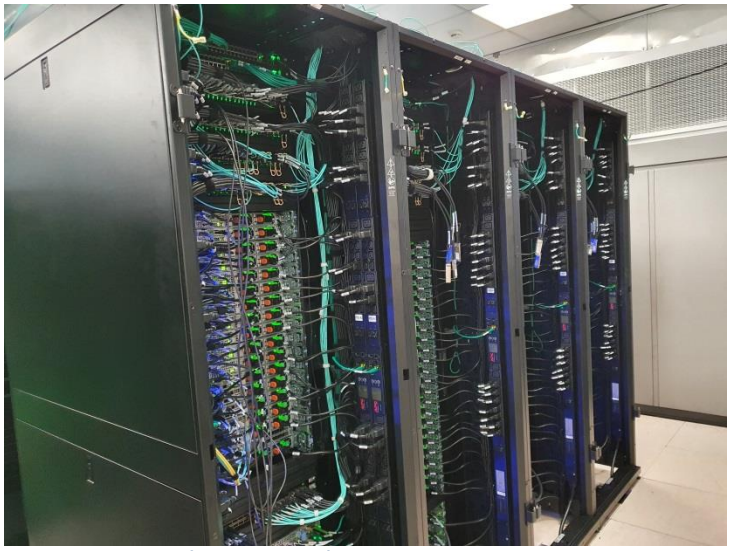

Figure 5 - Fenix systems at CINECA

4. Fenix Resources

Depending on the site specialisation recalled in section 2 and the result of the procurements, the provisioned resources have various characteristics and sizing on the various sites. This section summarizes the resources now available in Fenix and the specific hardware available at each site.

4.1. Scalable Computing Services

Site	Component local name	Provisioned ICEI resources	Description
CSCS (CH)	Piz Daint Multicore	250 nodes	• Memory per node: 64 GB, 128 GB • Compute nodes/processors: 1813 Cray XC40 nodes with Two Intel® Xeon® E5-2695 v4 @ 2.10GHz (2 x 18 cores) CPUs • Interconnect configuration: Cray Aries
JSC (DE)	JUSUF	195 nodes	• CPU: 2x AMD EPYC "ROME" 7742 (2x64 cores, 2.2 GHz base clock) • Memory: 256 GByte DDR4-3200 • GPU (some nodes): 1x NVidia Volta V100 with 16 GByte memory • Local storage: 960 GByte NVMe drive • Interconnect: HDR100
CINECA (IT)	Galileo100	340 nodes	• 2x CPU 8260 Intel Cascade Lake, 24 cores, 2.4 GHz base frequency (3,90 GHz Turbo) • 384 GB RAM • 480 GB SSD local disk

Total Scalable Computing resources provisioned by ICEI: 785 nodes

4.2. Interactive Computing Services

Site	Component local name	Provisioned ICEI resources	Description
CSCS (CH)	Piz Daint Hybrid	400 nodes	• Memory per node: 64 GB • GPU memory: 16 GB CoWoS HBM2 • Compute nodes/processors: 5704 Cray XC50 nodes with Intel® Xeon® E5-2690 v3 @ 2.60GHz (12 cores) CPUs and NVIDIA® Tesla® P100 GPUs • Interconnect configuration: Cray Aries
JSC (DE)	JUSUF	5 nodes [1]	• CPU: 2x AMD EPYC "ROME" 7742 (2x64 cores, 2.2 GHz base clock) • CPU memory: 256 GBytes • GPU (some nodes): 1x NVidia Volta V100 with 16 GBytes memory • Local storage: 960 GBytes NVMe • Interconnect: HDR100
CEA (FR)	Interactive Computing Cluster	32 nodes	• 30 nodes with 2 CPUs (18 cores @ 2.6 GHz each), 1 GPU (NVidia V100, 32GB), 384GB of RAM. • 2 nodes with extra large memory (3072GB) 4 CPUs and 1 GPU (Nvidia V100, 32GB). • Infiniband HDR100 interconnect
CINECA (IT)	Galileo100	214 nodes	• 2x CPU 8260 Intel Cascade Lake, 24 cores, 2.4 GHz base frequency (3,90 GHz Turbo) • 384 GB RAM DDR4 • 2 TB SSD local disk • 34 nodes equipped with 2x NVIDIA GPU V100
BSC (ES)	Interactive Computing Cluster	3 nodes	• IBM Power System AC922 System • 2x IBM POWER9 16-core 2.6GHz • Memory 1024 GB DDR4 • 2x NVIDIA Tesla V100 SXM2 16GB Accelerator • 2x 960 GB SSD Disks

Total Interactive Computing resources provisioned by ICEI: 654 nodes

[1] As they are based on the same hardware, scalable computing and interactive computing nodes of JUSUF can be used for both purposes. The indicated share between scalable computing and interactive usage can be adjusted depending on the project allocation requests.

4.3. VM services

Site	Component local name	Provisioned ICEI resources	Description
CSCS (CH)	OpenStack Cluster	35 servers	2 types of compute nodes: Type 1: • CPU: 2x Intel E5-2660 v4 14C • RAM: 512 GBCPU Type 2: • CPU: 2x Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20GHz 8C • RAM: 768 GB VMs can be of various flavours and use several cores up to 16.
JSC (DE)	JUSUF	5 servers	• CPU: 2x AMD EPYC "ROME" 7742 (2x64 cores, 2.2 GHz base clock) • CPU memory: 256 GBytes • GPU (some nodes): 1x NVidia Volta V100 with 16 GByte memory • Local storage: 960 GBytes NVMe • Interconnect: Ethernet 40Gb
BSC (ES)	Nord3	84 servers [2]	Intel Sandybridge cluster being able to be used as VM host or scalable cluster. 84 nodes dx360m4. Each node has the following configuration: • 2x Intel SandyBridge-EP E5-2670/1600 20M 8-core at 2.6 GHz • 32 GBytes of RAM.
CEA (FR)	Openstack Cluster	20 servers	Each server has the following configuration: • 2 CPUs (18 cores @ 2.6GHz each) • 192GBytes of RAM
CINECA (IT)	Galileo100 Openstack cluster	77 servers	Each server has the following configuration: • 2x CPU 8260 Intel CascadeLake, 24 cores, 2.4 GHz • 768 GBytes of RAM • Local storage : 2 TB SSD

Total VM resources provisioned by ICEI: 221 servers

[2] The servers of Nord3 can also be used for scalable computing.

4.4. Archival Data Repositories

Site	Component local name	Provisioned ICEI resources	Description
CSCS (CH)		4000 TBytes	Stores POSIX and Object, including backup on Tape library (2x).
CEA (FR)	Store	7500 TBytes+	Lustre filesystem with HPSS storage backend on tapes.
CEA (FR)	Swift/OpenIO	7000 TBytes	OpenIO object store with Swift interface
CINECA (IT)	Archival Data Repository	10 000 TBytes	Object storage with Swift/S3 interface
BSC (ES)	Agora	6 000 TBytes	

Total archival storage provisioned by ICEI: 34.5 Peta-Bytes

4.5. Active Data Repositories

Site	Component local name	Provisioned ICEI resources	Description
CSCS (CH)	Low latency storage tier (DataWarp)	80 TBytes	Non-Volatile Memory
JSC (DE)	High Performance Storage Tier (HPST)	2 000 TBytes	Flash-based data cache based on DDN's IME technology
CEA (FR)	Work filesystem	3 500 TBytes	Lustre filesystem
CEA (FR)	WorkFlash filesystem	970 TBytes	Full-flash Lustre filesystem
CINECA (IT)	HPC Storage	10 500 TBytes	Storage for hot data (w DDN IME) accessed from SCC, IAC and VM services
BSC (ES)	HPC Storage	70 TBytes	GPFS Storage accessed from HPC clusters

Total active storage provisioned by ICEI: 17.1 Peta-Bytes

5. Integration of the Fenix resources

The Fenix Authentication and Authorization infrastructure (AAI) is a key requisite for a common user and resource management. This infrastructure aims to provide users with a homogeneous access to the ICEI sites and take advantage of the distributed architecture of Fenix.

To realise the integration of all services in a common AAI and facilitate a common user and resource management, two services have been or will be deployed in the near future:

- A central Identity Provider (IdP), that provides users with a common authentication mechanism to access ICEI resources.

- A Federated User and Resource Management System (FURMS), that implements group, project and resource management in a central interface common to all Fenix sites.

Note that other services that facilitate integration of different resources within the Fenix infrastructure are implemented as part of the HBP SGA3 WP6 work plan (e.g. data transfer service across sites, data location service…). These services are thus outside the scope of this document.

6. Conclusion and next steps

All the systems that were planned to be procured in the framework of the ICEI project are now deployed and available to the targeted user communities (HBP, PRACE and national researchers).

As the new services of federation are implemented and open to the users, these systems will be accessible in an increasingly homogeneous and integrated manner.

Provisioned and consumed resources on these systems will continue being reported in yearly deliverables “Data storage and compute provisioning” (next at months 49 and 63).

Next deliverable D4.14 “Decommissioning strategy” will describe the plan for decommissioning resources deployed within ICEI and will state the options for continuing the service using other funding sources.

7. References

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

[D3.6] 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

[D4.1] 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

[D4.2] ICEI Deliverable D4.2: Infrastructure at ETHZ/CSCS
https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d2722542/ICEI_D4.2_v0.4.pdf

[D4.3] ICEI Deliverable D4.3: Infrastructure at JUELICH-JSC
https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d3149893/ICEI_01--D4p3--v1p0.pdf

[D4.4] ICEI Deliverable D4.4: Infrastructure at CEA
https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d3208434/ICEI_D4.4_v1.1.pdf

[D4.5] ICEI Deliverable D4.5: Infrastructure at BSC
https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d3662581/ICEI_D4.5_v1.0.pdf

[D4.6] ICEI Deliverable D4.6: Infrastructure at CINECA
https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d3693534/ICEI_D4.6_v1.0_final.pdf

[D4.8] ICEI Deliverable D4.8: R&D results
https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d3513263/D4.8_RD_Results_V1.0.pdf

[D4.15] 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