Implementation of the Fenix Interactive Computing Service

Work package	WP3 Technical specification and coordination
Author(s)	Giuseppe Fiameni - Andrea Pieretti (CINECA)
Dissemination level	ICEI Confidential
Nature	

Date	Author(s)	Comment	Version	Status
19 June 2019	Giuseppe Fiameni - Andrea Pieretti (CINECA)			Draft
2 July 2019	Giuseppe Fiameni - Andrea Pieretti (CINECA)			Draft
19 July 2019	Giuseppe Fiameni (CINECA)	Address DIrk's comments		
21 August 2019	Giuseppe Fiameni (CINECA)	Address all received comments		Final

1. Introduction

The purpose of this document is to describe the Interactive Computing Environment (ICE), a software solution capable to provide interactive access to Fenix resources. It also describes how Cineca plans to tender its development.

2. ICEI Project overview

2.1 The ICEI project and the Fenix research e-Infrastructure

ICEI[1] (Interactive Computing e-Infrastructure) is a collaboration project partially funded by the European Commission involving the leading European HPC Centres (herein referred to as "the five Sites") such as BSC (Spain), CEA (France), CINECA (Italy), CSCS (Switzerland), and FZJ/JSC (Germany). As members of the project, the centres are committed to offer federated compute and storage services.

The federation, named **Fenix Research Infrastructure**[2], will aggregate various resources distributed across Europe and will serve as the primary user, the neuroscience community under the coordination of the **Human Brain Project**[3].

3. Interactive computing

Nowadays, the main usage model for large scale HPC systems is based on the scheduling of batch jobs. Although this approach is the most adopted one, it does not originate from computational requirements, but it reflects the need to maximize the utilization of very expensive computational resources. Conversely, the situation differs when taking into consideration personal workstations or shared memory servers, where timesharing interactive executions are the norm.

Interactive Computing here refers to the capability of a system to spawn interactive sessions on HPC systems nodes or on nodes in the vicinity of an HPC system.

Such capability would, e.g. enable on-the-fly monitoring and steering of distributed computing workloads by the user.

The real-time interaction of a user with a program runtime is motivated by various factors, such as the need to estimate the state of a program or its future tendency, to access intermediate results, and to steer the computation by modifying input parameters or boundary conditions. A typical usage scenario for interactive computing regards the visualization, processing, and reduction of large amounts of data, especially when the processing part cannot be standardized or implemented in a static workflow.

Another scenario is considered even more important: in neuroscience and various other application fields (see for more detail Addendum to D4.15), the computational scientists are starting to use interactive frameworks and scripting languages to integrate the more traditional compute and data processing application running in batch, e.g. the use of R, Stata,

[1] The ICEI project has received funding from the European Union Horizon 2020 research and innovation programme under the grant agreement No 800858
[2] https://fenix-ri.eu
[3] https://www.humanbrainproject.eu

Matlab/Octave or Jupyter Notebook just to name a few. In this way, the time spent in this activity is a non-negligible component in the “time to science”. In that scenario, users will be presented with the possibility to load and visualize data resulting from simulations or collections, series or experimental data, in an interactive way. To properly support interactive supercomputing, the system needs to encompass all the connections and protocols needed to dynamically attach multiple visualization and steering front-ends to a running application and to enable transparent staging of data across multiple storage tiers, from distributed storage participation to node-central memory, to ensure front-ends receive a continuous stream of data.

3.1 Objectives

The Interactive Computing service will provide near-immediate access to computational resources via interactive sessions to allow, e.g. visualization of large data sets, computing via notebook, data manipulation and post-processing, etc. In order to be effective, this service would need to facilitate users’ interaction with the infrastructure, support efficient handling of interactive sessions, maximize resource utilization while executing different workloads, support staging of data across multiple memory and storage tiers, improve energy consumption. To implement this goal, this R&D activity should work to develop a solution to provide users with a single interaction mechanism, e.g. Web, CLI, API capable to handle different utilization requests, from the submission of batch jobs, the creation of interactive visualization sessions, to the transparent set up of notebook environments, e.g. Jupyter Hub Web Server[4].

We anticipate the following usage scenarios for interactive access:

1. User connects to ICEI hardware resources via an Interactive Compute Service using secure protocol to, e.g. analyse and visualise data produced by simulations that have been running on Scalable Compute Services, i.e. on a large-scale HPC system. The data may either be directly streamed to the Interactive Compute Node or be staged through an Active Data Repository. This usage scenario stays within a controlled HPC environment. (See Figure 1 for a schematic view on this usage scenario.)
2. User connects through a web protocol to a service running in a cloud environment to analyse and/or visualise data that is made available through an Archival Data Repository. The latter is accessible from both, within an HPC as well as a cloud environment. Optionally, a direct communication channel between, e.g., a simulation running on an HPC system and a cloud service for direct streaming of data for visualisation or steering of the simulation is envisaged. (See Figure 2 for a schematic view on this usage scenario.)

We envisage the Interactive Compute Services to be provisioned through a set of components, clients and servers, that are part of an HPC cluster, i.e. clusters that provide both Interactive Computer Services as well as Scalable Computer Services, and Cloud services will be running on virtual machines provisioned through an OpenStack instance.

Towards enabling these usage scenarios, this tender will focus on realisation of supported solutions through R&D services that address several of the following needs:

[4] https://jupyter.org/hub

1. Fast and easy access to Interactive Compute Services through servers that are part of an HPC cluster.
2. Optimise the use of available compute and storage resources including support for efficient suspend-resume mechanisms for HPC jobs running on Scalable Compute Services to facilitate releasing hardware resources for Interactive Compute Services.
3. Easy-to-use tools for efficient remote visualisation through the Interactive Compute Services.

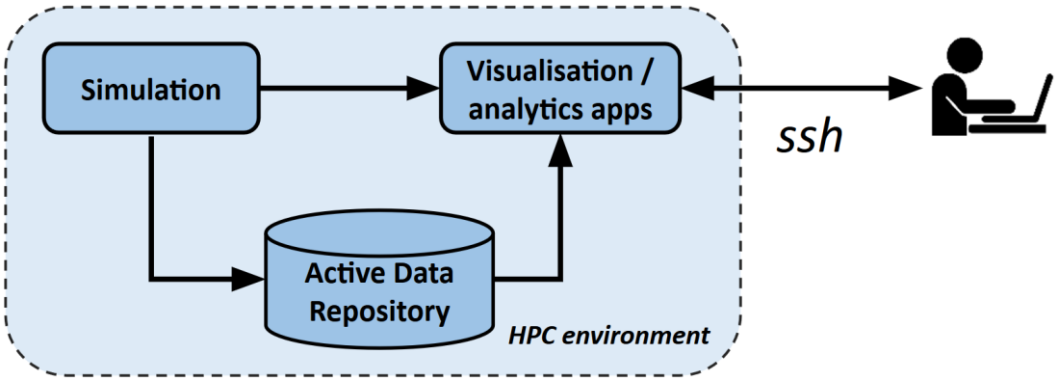

Figure 1 – Interaction with the system via a SSH client

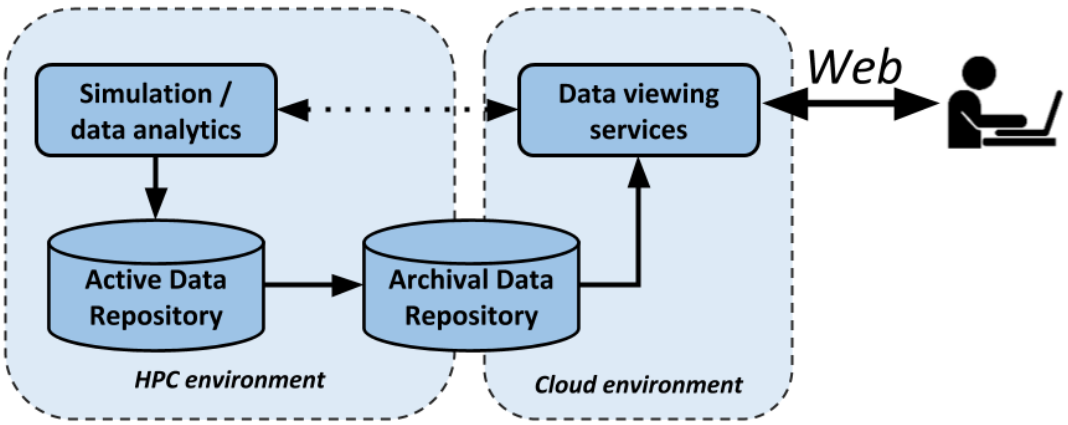

Figure 2 – Interaction with the system via a Web application

4. Glossary

Term	Description
CINECA	Consorzio interuniversitario per il calcolo automatico
Fenix	e-infrastructure providing access to Interactive Computing Services, elastic access to scalable compute resources and federated data
Fenix AAI	Fenix Authentication and Authorization Infrastructure
Proxy	A Proxy server is a server (a computer system or an application) that acts as an intermediary for requests from clients seeking resources from other servers
ICE	Interactive Computing Environment
IdP	The Identity Provider is a system entity that creates, maintains, and manages identity information for principals while providing authentication services to relying applications within a federation or distributed network
Fenix Central Proxy IdP	Is the Central Identity Provider proxying authentication requests across the federation
SP	Service Provider
SSO	Single Sign On
OIDC	‘OpenID connect’ authentication protocol
OIDC Access token	Access tokens are used as bearer tokens. A bearer token means that the bearer (who hold the access token) can access authorized resources without further identification.
SAML	‘Security Assertion Markup Language’
Introspection	The Token Introspection extension in OIDC defines a mechanism for resource servers to obtain information about access tokens.
LoA	Level of Assurance
Account Linking	Account linking lets you connect the identity of the user with a user account in a different system.
Hosting site	A centre providing services and resources and part of the Fenix federation (CINECA,CSCS,BSC,JSC,CEA)
FURMS	Fenix User and Resource Management Service

5. Technical specifications

5.1 Categories of requirements

The requirements and features within the documents are categorised as follows.

Categories	Description
MRQ	**Important requirement:** The described requirement is considered a necessary requirement which all final proposals must fulfil.
TC-1	**Target capability (priority 1):** The described target capability, i. e. functional feature or performance target, is considered an important requirement for the operation and functionality of the service. Failure to meet the requirement does not lead to exclusion, but results in a lower evaluation score.
TC-2	**Target capability (priority 2):** The described requirement is considered optional. A proposal that fulfills the requirement is considered an advantage for the bid.
DOC	Documentation that must be included in the proposal. All documentation items are mandatory and must be provided by all candidates in their proposal.

5.2 The Fenix Interactive Computing Service

Table below includes all the base requirements.

#	Project outcomes	Category
1	Specification and implementation of a software solution, HW agnostic, preferably based on Open Source components, capable to: ● provide users with a single interaction mechanism, e.g. Web, CLI, API capable to handle different utilization requests, from the submission of batch jobs, the creation of interactive visualization sessions, to the transparent set up of notebook environments, i.e. Jupyter Hub Web Server[5]; ● facilitate users' interaction with the infrastructure while supporting efficient handling of interactive sessions; ● maximize resources utilization while executing different workloads; ● support staging of data across multiple memory and storage tiers; ● optimize energy consumption. ICEI sites will deploy interactive computing resources and develop software components that are in line with the project goals.	**MRQ**
2	Specification and implementation of a software solution, HW agnostic, preferably based on Open Source components, **preferably** capable to provide: ● support for efficient suspend-resume mechanisms for HPC jobs running on Scalable Compute Services to facilitate releasing hardware resources for Interactive Compute Services. In order to be effective, the solution may consider extending resource manager/operating system capabilities;	**TC-1**
3	Training and documentation about developed software solution.	**MRQ**
4	Maintenance/support services for the whole project duration.	**MRQ**

[5] https://jupyter.org/hub

5.3 Licenses and Intellectual Property Rights

#	Requirement description	Category
1	The supplier must grant a free non-revocable, non-exclusive right of use for the developed interactive computing environment (ICE), to the following entities: ● The Public Procurer; ● All ICEI Sites (BSC, CEA, CINECA, CSCS, JSC); ● All partners of the Human Brain Project as identified in the Framework Partnership Agreement	**MRQ**
2	Access to relevant third-party software components must be available at fair and reasonable conditions.	**MRQ**
3	In case the source code of the developed software is not published, the supplier may claim the access to it.	**MRQ**
4	The developed software and documentation will be released under an open-source license and be made available online.	**TC-1**
5	The following rules concerning the intellectual property rights are foreseen. Details are subject to negotiations: ● The Supplier background intellectual property used in the execution of the contract are not affected; ● The Supplier and the Public Procurer are required to exchange background information inasmuch as required for the execution of the contract; ● Sole inventions remain the property of the inventor.	**MRQ**

5.4 Time Schedule and Deliverable plan

#	Requirement description (functional)	Category
1	The development project will be structured in multiple Deliverables. Project plans with lower project duration and early delivery of important features will be preferred. The Public Procurer intends to impose a maximum project duration and latest possible submission time for the last Deliverable. This deadline is tentatively set to 31. December 2021. In order to simplify the procurement process, the project schedule should initially be organized such that at least two and at most five (development) phases and associated Deliverables are foreseen. A more fine-grained plan with the Supplier may be negotiated at a later point.	**MRQ**
2	The Proposal describes the timeline and scope of the entire project.	**MRQ**
3	The Proposal will document the envisioned payment plan matching the time schedule.	**MRQ**
4	A development project could also be provided and it should contain: ● Estimated date of readiness; ● Early deliveries of “beta” versions related to the implementation of important features (if applicable)	TC-1

5.5 Installation and Acceptance

#	Requirement description (functional)	Category
1	The Deliverable will undergo an evaluation by the Public Procurer with respect to the conformity of the Proposal. In case that the Deliverable does not meet the expected quality or the development scope and expected outcome is insufficient, the Public Procurer may reject the Deliverable. The Supplier will then be requested to provide an updated Deliverable. In the case that the re-submitted Deliverable is judged insufficient (with respect to the above-mentioned criteria), the Public Procurer may terminate the contract. In this case, the Supplier may claim compensation for justified expenses as regulated by the law.	**MRQ**
2	The Deliverable will be undergoing a separate acceptance procedure in which the agreed functionality and performance characteristics are verified. The Deliverable will be tested on all participating Sites who request it, with the Public Procurer acting as the single point of contact. The acceptance of the Deliverable will take place within **30 days** after declaration of readiness by the Supplier. Following successful acceptance, the Public Procurer can report bugs, including missing functionality, within a **3-months** time window outside of the support and maintenance scope.	**MRQ**

5.6 Interactive Computing Environment (ICE)

- Fast and easy access to Interactive Compute Services through servers that are part of an HPC cluster.
- Support for efficient suspend-resume mechanisms for HPC jobs running on Scalable Compute Services to facilitate releasing hardware resources for Interactive Compute Services.
- Easy-to-use tools for efficient remote visualisation through the Interactive Compute Services.

5.6.1 Base features

#	Requested needs	Category
1	The developed solution (hereafter referred to as “ICE”) should deliver a single interaction mechanism capable to handle interactive sessions onto Fenix Interactive Computing resources. This includes the spawning, the execution and the management of interactive sessions (single or multi-node) in order to: ● maximize users experience by permitting them to remotely access computing and visualization facilities through a fast and easy interface. Access to such facilities could be realised through interacting with bastion hosts, login nodes, or gateways, and leverage on resource manager capabilities; ● support staging of data across different storage tiers, either local to the nodes or network attached; ● maximize resource utilization by balancing interactive and batch workloads automatically while optimizing their execution using hardware capabilities, e.g. graphics cards, local storage, interconnectivity features, etc. Detailed specifications for the final solution will be determined during the negotiation phase.	**MRQ**
2	The solution should be able to support multiple interface topologies, including SSH and Web using standard access and transport protocols.	**TC-1**
3	The solution should permit the contextual execution of interactive sessions and batch jobs on the same infrastructure, balancing the use of the resources on the based on the utilization policy. For instance during the normal working hours, interactive sessions may have priority over batch jobs.	**TC-1**

#	Requested needs	Category
4	The solution should be able to accommodate users request in terms of resource capacity adaptively, i.e. dynamically adjust resource allocation (increasing/decreasing) when required during the execution. For instance, to permit the fast execution of an interactive session, the allocation of resources could be gradually increased from a defined minimum to the total requested amount. ● To this end, the solution may introduce functions like: the transparent suspension of running batch jobs to free resources for higher priority workloads, i.e. interactive sessions; ● The flushing of data across available storage tiers to free capacity; ● The co-allocation of compute and storage resources; ● The exploitation of overbooking strategy; ● The automatic adaptation of resources allocation. Initial requests for interactive sessions may require adjustments over the duration of the session to better match workload requirements, e.g. increasing of memory capacity.	**TC-1**
5	The solution should ensure that the interactive sessions are spawned in a reasonable amount of time (order of magnitude is **minutes**). If needed, the session can be started using a pre-defined minimum amount of resources (configurable by the site administrator) while gradually increasing the allocated resources until the requested amount is provided.	**MRQ**
6	The solution may support the execution and/or the orchestration of containers based on commonly used technology, such as Docker and Singularity.	**TC-2**
7	The solution may be able to deal with data that is streamed directly to the Interactive Compute Node or staged through an Active Data Repository.	**TC-2**
8	To ensure efficiency and optimal usage in case of poor network connectivity (low bandwidth, high latency), the solution may introduce optimization mechanisms, e.g. compression of data, auto-scaling of image resolution, color-depth.	**TC-1**
9	The solution must be compatible with the Fenix Authentication and Authorization Infrastructure based on standard protocols. (i.e. OpenID Connect, SAML, SSH, etc.). Further details will be provided during the course of the negotiation.	**MRQ**
10	The solution must support SLURM as the workload manager	**MRQ**
11	The solution must be compatible with Red Hat Enterprise Linux (RHEL) or a binary-compatible distribution (e.g., CentOS).	**MRQ**

5.7 Documentation and Training

#	Requirement description (functional)	Category
1	A design document for the ICE solution should be provided as part of the deliverables. The document will need to contain the following sections: ● Technical specification and design architecture of the solution ● Specifications of the supported configuration options as required for deploying ICE at all Sites; ● Specifications for the system software of the base platform (e.g. operating system). The design document will be used by the public procurer and the Sites to procure and deploy the hardware infrastructure for ICE. The Supplier will support ICE installed on the base platform (see 5.2 #12).	**MRQ**
2	Each Deliverable encompasses documentation for the installation, system administration as well as user documentation.	**MRQ**
3	The proposal must include training for the system administrators at the five Sites. The training will provide an introduction in the deployment and operation of ICE. Scope and duration of the training measure should be appropriate for the complexity of the solution. Training may be performed: ● Online; ● **Preferred:** Training course at CINECA. Details are subject to negotiations.	**MRQ**
4	An additional training workshop at a later point in the project is offered.	**TC-2**

5.8 Support, Maintenance and Services

#	Requirement description (functional)	Category
1	The Proposal will include support and maintenance for the ICE software installed at the five Sites on base platforms adhering to the specifications. The support and maintenance period will start with the acceptance of the first Deliverable containing a deployable and functional software and will last until **31.03.2023**. The support and maintenance agreement includes at least: ● Security fixes; ● Fixes for bugs that affect the contracted functionality and performance of the ICE service on the ICEI sites. The detailed scope of the support and maintenance agreement is subject to negotiation.	**MRQ**
2	The Public Procurer can grant the five Sites the authority to report problems directly to the Supplier on its behalf during the maintenance and support period.	**MRQ**
3	Support and maintenance coverage is provided for normal working hours (5×8 hours per week excluding weekends and public holidays in Italy). Selected additional days may be excluded from the support and maintenance coverage to account for foreign holidays.	**MRQ**
4	The Proposal includes **1 month** of deployment support for each of the five Sites after the acceptance of the Deliverable. Each Site can claim for such support until **15.01.2023** (i.e. deployment support will not extend beyond **15.02.2023**) Please note that, independently of this requirement, the deployment mechanisms and documentation have to be sufficiently mature to allow operators at the five Sites to update ICE without assistance.	**MRQ**

5.9 Long term sustainability

#	Requirement description (functional)	Category
1	Options for support and maintenance of the delivered solution at fair and reasonable conditions are available beyond the end of the contract.	**TC-1**

6. Procurement

6.1 Steps

The detailed specifications for the ICE solution will be determined in the negotiation phase. The purpose of this document is to provide a general overview of the expectations concerning the scope and conditions of the development services. The participants are requested to provide a concept of how they can provide a solution that meets the requirements expressed by the public procurer.

6.2 Procedure

Competitive procedure with negotiation under Italian law.

6.3 Budget

A total budget of at most 250 K€ (VAT excluded) is available for this procurement.

6.4 Evaluation criteria

In order to be eligible for contract award, the final Proposal must satisfy all technical requirements (category MRQ) and the total Proposal cost must be lower or equal to the total project budget (see Section 6.3).

Among the final Proposals that fulfil all minimal requirements, the award of the procurement contract will be based on the weighted criteria provided in the following table. A version of this table translated to Italian will be provided separately. The Italian version takes precedence.

Criterion	**Weight**
• Price	30
• Quality of concept (Section 3) • Technical design (Sections 5.2, 5.6)	35
• Documentation and training (Section 5.7) • Maintenance, support and services (Section 5.8) • Sustainability (Section 5.9)	20
• Time schedule and delivery plan (Section 5.4) • Installation and acceptance (Section 5.5) • Use of rights, licenses and intellectual property rights (Section 5.3)	15
Total	**100**

Appendix A: Potential Use Cases

Synthetic use cases to be considered:

- Hardware resources allocated to Interactive Computing services is adjusted dynamically in an automatic manner (like increased during business hours).
- Large-scale simulation running on distributed nodes is suspended to release resources for interactive computing services releasing as much volatile RAM resources as possible.
- User connects to Interactive Computing services and starts Paraview-based visualisation application.