FENIX

RESEARCH INFRASTRUCTURE

Especificaciones técnicas para la adquisición de nodos de cómputo para ejecuciones interactivas a usar en FENIX Research e-Infrastructure en el BSC (Interactive compute cluster)

The ICEI project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 800858.

Índice

1. Contexto

1.1 BSC

El Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS), establecido en 2005, sirve como el Centro Nacional de supercomputación en España. El centro alberga MareNostrum, uno de los superordenadores más potentes de Europa.

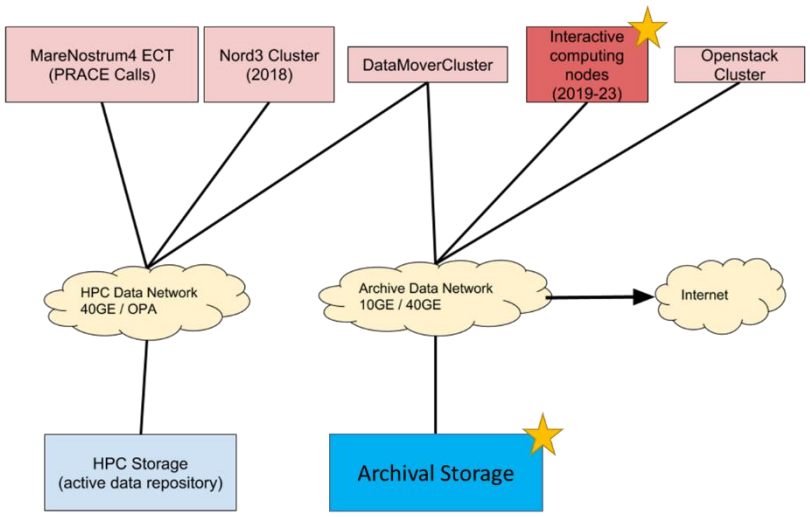
La misión del BSC-CNS es investigar, desarrollar y gestionar tecnologías de la información con un objetivo mayor de proporcionar innovación científica. Trabaja para conseguir estos objetivos en campos de la ciencia de la computación, ciencias de la vida y de la tierra.

BSC es un activo y reconocido participante en diversas iniciativas para la integración y consolidación de la supercomputación y gestión de datos en Europa y España. El BSC gestiona toda la información generado por las simulaciones de altas prestaciones ejecutados en sus clusters (conjunto de nodos de computo). En orden de poder cumplir con los requisitos de datos de los científicos, BSC dispone de más de 20 PB de disco disponible y una librería de 6 PB. A nivel internacional, BSC está trabajando en más de 42 proyectos europeos. A nivel de infraestructura, los principales proyectos de e-infraestructura que el BSC está participando son: EOSC-Hub, PRACE (donde el BSC es uno de los centros Tier-0) y HBP, en todos estos proyectos el BSC está participando como centro de datos como centro de supercomputación.

1.2 Objeto del pliego

El objetivo de este pliego es la adquisición de unos nuevos componentes de una infraestructura que va a ser usada en conjunto que infraestructura actual del BSC, para poder proporcionar los recursos suficientes para poder dar soporte a los requerimientos científicos de diversas comunidades como el Human Brain Project y PRACE.

Siguiendo la referencia de arquitectura técnica que se ha definido dentro del proyecto ICEI, la infraestructura local dentro del BSC tendrá los siguientes componentes (se marcan con una estrella aquellos que se adquirirán a través de este concurso):



Los siguientes componentes se proveerán con infraestructura ya existente en dependencias del BSC:

- **"Scalable compute services"** se proveerá mediante el cluster Nord3, un cluster basado en Intel SandyBridge y red Infiniband
- **"Active Data repository"** se proveerá con el almacenamiento actual de HPC que ya se provee a los diversos clusters de supercomputación
- Actuales BSC "**Data Mover service"** proveerán el servicio que el mismo nombre tiene para ICEI
- **"Openstack cluster"** será proporcionado parcialmente por el cluster Nord3 y algún servidor extra

Nuevos componentes deben ser adquiridos para completar la arquitectura referencia y así poder proveer de los servicios ICEI, concretamente: "Interactive computing node cluster" y el "Archival Storage".

Todos los requerimientos de este pliego han sido definidos teniendo en cuenta el uso futuro de esta infraestructura dentro de Fenix y las comunidades de usuarios y sus necesidades dentro del ámbito del Human Brain Project, PRACE como cualquier otra comunidad científica Europea con altas necesidades de computación y uso de datos.

El "Archival Data repository" ya fue licitado en el expediente: CONSU02019003OP, este pliego se encarga de licitar el "Interactive computing cluster", es decir, el cluster de computo para ejecuciones interactivas que quedó desierto en el pliego antes indicado.

El cluster de cómputo (grupo de servidores) para ejecuciones interactivas se describe a continuación:

Interactive Computing Cluster (Grupo de servidores para ejecuciones interactivas)

Cluster de computo consiste en un grupo de servidores que tienen de ser capaces de soportar ejecución de jobs HPC y al mismo tiempo ser capaz de soportar sesiones interactivas, por ejemplo, de visualización o de análisis de ejecuciones que los usuarios puedan interactuar durante la ejecución de su simulación. Para esta licitación se piden la provisión de un mínimo de dos de este tipo de servidores.

Después del estudio de los casos de uso de las diversas comunidades científicas que harán uso de este cluster, los nodos de cómputo o servidores se espera que tengan:

- GPUs : Para ayudar en la ejecución de diversas aplicaciones de inteligencia artificial y visualización
- Alta capacidad de memoria (diversos TB)
- Alta densidad de almacenamiento local, para complementar la alta densidad de memoria principal
- Procesadores capaces de ejecutar de forma efectiva nuevos paradigmas de programación como contenedores, virtualizaciones, códigos de inteligencia artificial.

1.3 BSC Data Center

Todo el "Interactive Compute cluster" se instalará en el nuevo centro de procesadimento (CPD) del BSC localizado en el sótano del edificio TG. En el siguiente esquema se presenta la sala y en verde están marcado los armarios que pueden ser destinados a esta instalación.

El licitador deberá instalar sus equipos en uno de los armarios de archivo activo existentes o en alguno de los que proporcione la empresa que fue la adjudicataria del lote2 del concurso CONSU02019003OP.

2. Especificaciones técnicas

En este capítulo, vamos a describir la especificación técnica de este concurso.

Esta licitación se basa en la adquisición, instalación y puesta en marcha de un conjunto de servidores, donde el BSC implementará el servicio de “Interactive computing cluster” descrito en el apartado anterior.

En las siguientes tablas se describen el número y las características técnicas que dichos servidores deben cumplir.

Todo componente informático deberá ser debidamente integrado e instalado con el resto de la infraestructura del BSC para su uso en producción, en un régimen de llaves en mano.

2.1 Hardware

	Description	Category
“Interactive Compute cluster” / Cluster de nodos de cómputo	Mínimo de 2 nodos de cómputo (servidores) que deben ser capaz de ejecutar aplicaciones de análisis de datos de forma eficiente. • Mínimo de 1 TB de memoria principal volátil, configurada en un modo balanceado entre todos los canales disponibles y máximo rendimiento	MRQ

	• Mínimo 2 GPUs por nodo para inteligencia artificial, cómputo y visualización de la tecnología más avanzada • Conectividad para montar el filesystem de HSM y conectividad hacia internet de mínimo 2 links de 25/40Gbit Ethernet por nodo	
	El número de nodos de computo (servidores) ofrecidos por encima del mínimo será evaluado	TC-1
	Las características hardware de los servidores ofrecidos serán comparados y evaluadas a nivel de: • Capacidad de computación; • Memoria ofrecida por nodo; • Incorporación de NVMRAM en los nodos; • Almacenamiento local ofrecido.	TC-2
	Los nodos deben soportar cualquier tipo de virtualización y sus extensiones, como poder configurar las GPUs en modo “pass-through” o poder ofrecerlas a las máquinas virtuales de forma directa.	MRQ
	Los servidores deberán poder ser administrados via out-of-line via una interfaz ethernet (ipmi, openbmc, o similar)	MRQ
	Estos nodos deberán integrarse en la red de administración del archive storage system, su instalación de sistema operative será gestionada por dicho cluster. Cualquier modificación de estos nodos puedan necesitar para adaptarse a dicha infraestructura deberá estar incluido (hardware/software).	MRQ
Red	Es necesario proveer todos los componentes de red necesarios para poder conectar estos componentes a las diversas redes del sistema de HSM. Esas conexiones se deberán realizar de manera que se garantice la redundancia y balanceo de carga, siempre que la red no sea un factor limitante en el cluster de interactive computing.	MRQ
	Los componentes de red ofertados serán evaluados y comparados entre las diversas ofertas recibidas, tanto a nivel de redundancia, balanceo de carga de red y rendimiento de la red.	TC-1

2.2 Software

	Description	**Category**
General	La solución debe proveer con todo el software necesario para que funcionen los servidores ofertados como cada uno de los componentes por los que están formados.	MRQ
OS software	El sistema operativo ha de ser Linux y debe estar soportado por el resto de componentes hardware o software.	MRQ

3. Benchmarks

Una lista de benchmarks han sido seleccionados para ayudar a evaluar el rendimiento técnico de las soluciones presentadas.

Los benchmarks seleccionados se pueden organizar en 2 grupos:

- Benchmarks genéricos o sintéticos que miden el rendimiento general
- Benchmarks de aplicación derivados del grupo denominado: ICEI application Benchmark suite. Este grupo se usará para medir el rendimiento de aquellas aplicaciones y casos de uso que esta infraestructura proveerá servicio en un futuro a través de la Fenix-RI.

El input, software e instrucciones de ejecución para los benchmarks se derivan del ICEI application benchmark suite, que será proporcionado durante el procedimiento de este concurso, aunque se describe en la siguiente página web:

https://indico-jsc.fz-juelich.de/event/87/material/slides/0.pdf

Cada proveedor debe proveer resultados reales o una estimación de rendimiento que producirá su solución en los benchmarks seleccionados. Esos valores serán confirmados luego con la infraestructura durante las pruebas de aceptación.

Para el “cluster interactive computing” un grupo de benchmark sintéticos ha sido seleccionado, se deben de proveer valores con ejecuciones reales en nodos como los ofertados o una estimación de los valores que se espera de ellos en los nodos ofertados.

En caso de presentar una estimación, se deberá razonar los valores presentados, dichos valores se deberán luego validar en la fase de validación de la solución presentada.

- HPL pico
- HPL real
- HPCG
- Stream
- IOR (medir rendimiento de la memoria no-volátil)

Un Segundo grupo de benchmark de aplicaciones de usuario ha sido seleccionado; estos benchmarks representa las aplicaciones que se espera que sean ejecutados en estos nodos.

- Neuron
- Nest
- Neuroimaging Deep Learning
- Elephant ASSET

4. Mantenimiento

Mantenimiento debe ser proporcionado (correctivo y preventivo) para cualquiera de los componentes de la solución que se provee (hardware o software), como de la solución como una unidad.

En el evento de un fallo, una respuesta deberá proporcionarse dentro de las siguientes 4 horas dentro del horario laboral del BSC (08:00 – 17:00) y se deberá de proveer de un servicio de mantenimiento de 'next business day'.
En el evento de un fallo crítico que implique un fallo global de toda la infraestructura se deberá realizar un seguimiento 24x7 hasta que el incidente quede resuelto.

Durante el periodo de mantenimiento, el proveedor tendrá la responsabilidad de todas las tareas relacionadas con la sustitución de cualquier componente hardware o software que falle.

El proveedor proveerá al BSC con el acceso a todas las actualizaciones de los diversos componentes software que la solución está compuesta, como pueden ser: paquetes de sistema operativo, software de datos, sistema de ficheros paralelo, firmware, etc.

Deberá existir un único punto de contacto para que el BSC pueda reportar problemas de cualquier de los componentes que forma la solución ofertada, se debe proveer de un teléfono y una web para poder abrir dichos casos.

Mantenimiento pro-activo es obligatorio, donde el proveedor proporcionará al BSC recomendaciones sobre actualizaciones firmware/software, indicando también que mejoras o fallos solventan dichas actualizaciones.

El candidato debe de proveer una lista de riesgos que pueden afectar negativamente a la instalación del sistema. Para cada riesgo, el candidato deberá dar una indicación de probabilidad que pase y proveer de una descripción del impacto esperado como las acciones de mitigación que se implementarían, todo esto como parte del contrato.

El candidato debe describir los roles de responsabilidad de todos los elementos implicados durante la instalación y pre-producción del sistema en la forma de RACI- (Responsible, Accountable, Consulted, Informed)-model.

Mantenimiento ha de ser de 3 años desde el momento que el material ha sido aceptado y se encuentra en producción.

5. Transferencia de conocimientos

Como parte de la instalación del Proyecto, el proveedor debe proporcionar la documentación describiendo el diseño y el log de instalación de todo el proyecto, explicando las diversas decisiones de diseño durante la fase de instalación.

El proveedor debe realizar reuniones regulares durante la instalación y trabajar en conjunto con el equipo de operaciones del BSC en las tareas de instalación.

El proveedor deber proporcionar sesiones de formación durante la fase de instalación y trabajará de forma conjunta en la instalación con el grupo de operaciones del BSC.

Se debe de proveer un mapa físico de los armarios a proveer, junto con el cableado que va a ser conectado a cada armario y a cada conmutador.

Mapas físicos de todas las redes se deberán de proveer, indicando claramente que está conectado en cada puerto de toda la solución.

La documentación también debe incluir todos los procedimientos operacionales que se necesitan para mantener la infraestructura en funcionamiento óptimo.

Los documentos proporcionados han de ser en formato editable (Office).

6. Instalación y aceptación

Instalación complete se espera que se complete en 4 meses después de la notificación del contrato, la calificación de pre-producción tendrá una duración mínima de 1 mes.

La instalación se divide en las siguientes fases:

1. Instalación Hardware y software;
2. Aceptación provisional;
3. Cualificación de pre-producción;
4. Aceptación final

Instalación Hardware y software

La instalación software y hardware se completa una vez el proveedor ha entregado e instalado todos los elementos que conforman el sistema entero de acuerdo con la oferta presentada en este concurso.

Aceptación provisional

Después de la finalización de la instalación de hardware y software, y siguiendo a la declaración de buena disposición por el proveedor, el sistema se validará en esta fase.

El proveedor reproducirá los valores de rendimiento comprometidos en su oferta. También deberá demostrar los diversos “Target Capabilities” (TC-1 y TC-2) propuestos en su oferta.

Las siguientes pruebas serán realizadas:
- Comprobar concordancia con el contrato a nivel de hardware y software entregado;
- Pruebas realizadas por los representantes del proveedor y personas técnicas del BSC para poder validar el funcionamiento adecuado del sistema y su entorno;
- Resultado de los benchmarks.

Aceptación provisional será proporcionada después de esta fase.

Cualificación de pre-producción

El objetivo de la cualificación de pre-producción es comprobar que el sistema se comporta como es esperado durante una fase inicial de operativa. Elementos clave son la estabilidad, fiabilidad y rendimiento del sistema. Durante este periodo, el funcionamiento adecuado de los mecanismos necesarios del sistema en producción será validados. Durante este periodo, el proveedor deberá realizar los ajustes necesarios para hacer que la disponibilidad del Sistema sea la esperada según el pliego técnico.

La estabilidad de la máquina se verificará usando la metodología proporcionada en este pliego técnico. Actividades como la transferencia de conocimiento se podrán realizar durante esta fase también.

Aceptación final

La aceptación final validará el funcionamiento adecuado del sistema a nivel global durante el periodo de cualificación de pre-producción.

Independientemente de estas fases, los siguientes requerimientos se deben de tener en cuenta:

- La infraestructura se debe entregar completamente instalada, y lista para ser usada (llaves en mano), de acuerdo con el diseño proporcionado por el BSC;
- Toda instalación y configuración debe realizarse en las dependencias del BSC; no se permitirá el acceso remoto para trabajar en las tareas de instalación. La instalación se deberá realizar de forma conjunta con el equipo de operaciones del BSC para facilitar la transferencia de conocimientos durante la instalación;
- Cualquier decisión que se deba realizar o plan a implementar durante la instalación deberá ser aprobada por el grupo de operaciones del BSC antes de implementarse.

7. Contexto del proyecto

7.1 Proyecto global

7.1.1 The Human Brain Project (HBP)

El **Human Brain Project** (HBP) es un Proyecto estratégico H2020 (H2020 FET Flagship) financiado por la comisión Europea. Su objetivo es acelerar los campos de la neurociencia, computación y medicina relacionada con el cerebro. Éste aglutina una federación de programas de investigación en neurociencia fundamental, simulación avanzada y modelado a multi-escala con la construcción de una infraestructura de investigación.

7.1.2 El Proyecto ICEI y Fenix Research Infrastructure

7.1.2.1 Presentación

El Proyecto ICEI (Interactive Computing e-Infrastructure) está financiado por la comisión europea y está formado por los centros de supercomputación punteros en Europa como BSC (España), CEA (Francia), CINECA (Italia), CSCS (Suiza), and FZJ/JSC (Alemania).

El plan de este Proyecto es crear un conjunto de e-infraestructuras que será federadas para formar la **Fenix Research Infrastructure**. El Human Brain Project será el primer Usuario de esta plataforma.

7.1.2.2 Servicios de la Fenix Infrastructure

La figura siguiente muestra como los usuarios de las diversas comunidades interactuaran con la Fenix infrastructure.
En esta infraestructura federada, cada centro deberá proveer los siguientes servicios y

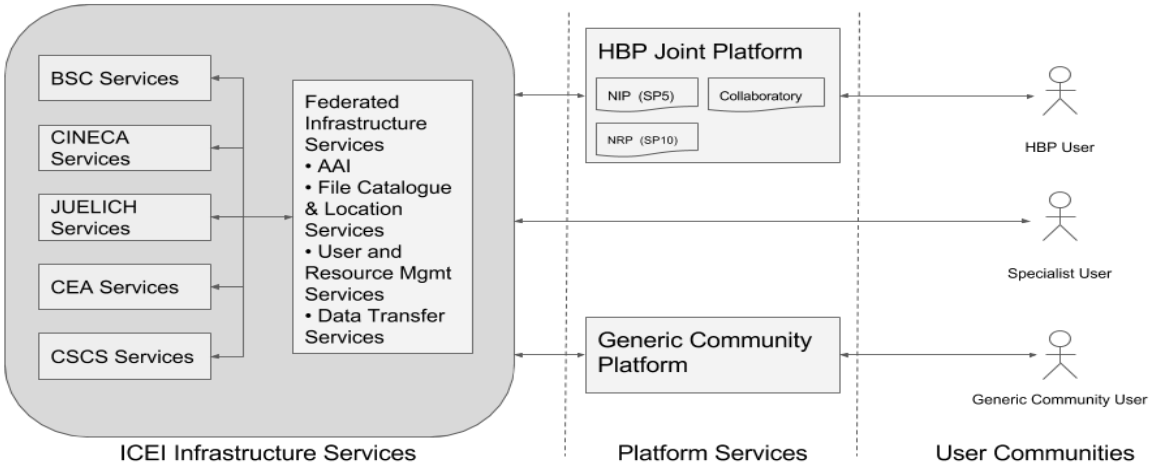


Figure 1 – Federación de la FENIX infrastructure

recursos:

- interactive computing services (IAC);
- scalable computing resources (SCC);
- virtual machines services (VM);
- active data repository (ACD);
- archive data repository (ARD).

Dentro de cada centro, estos servicios interactuarán según el siguiente gráfico:

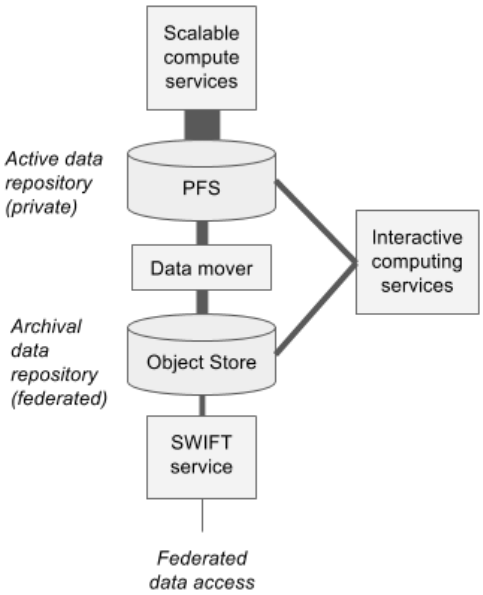


Figure 2 - Interacción entre componentes

7.1.3 Implementación

La Fenix e-infrastructure será adquirida mediante unos concursos coordinados. El resultado del global de la infraestructura se debe a un diseño conjunto entre los diversos miembros del Proyecto ICEI.

Cada centro es responsable para liderar la adquisición del su hardware y los servicios que van a proporcionar.

8. Definiciones

8.1 Definiciones de procedimiento

8.1.1 Definiciones

Term	Description
Candidato	Empresa participante en este concurso
Proveedor	El candidato que ha sido galardonado con el contrato como parte de esta licitación.

8.1.2 Categoría de requerimientos

Los requerimientos y características en este documento se categorizan de la siguiente forma:

Categoría	Descripción
MRQ	"Mandatory Requirements" o Requisitos obligatorios son considerados esenciales para el Sistema que se está adquiriendo y deben cumplirse por las propuesta final. Mandatory Requirements se verificarán por cada propuesta. Las propuestas finales que no cumplan con todos los Mandatory Requirements serán descalificadas.
TC-1 **TC-2**	"Target Capabilities" son características deseables o niveles de rendimiento del Sistema a adquirir. A diferencia con los "Mandatory Requirements", el no cumplimiento de "Target Capabilities" no implicará la descalificación de la propuesta. Propuestas que provean los "Target Capabilities" o los mejoren recibirán mejor puntuación. "Target Capabilities" están priorizadas. Nivel uno Target Capabilities (**TC-1**) se consideran de más importancia de nivel 2 Target Capabilities (**TC-2**).

8.2 Glosario

Term	Description
Backbone	Red Ethernet (10Gb, 25Gb, 40Gb or 100Gb).
BSC-CNS	Barcelona Supercomputing Center – Centro Nacional de Supercomputación
ESS	Elastic Storage Server
GPFS	IBM Sistema de ficheros paralelo
GPGPU	"General-Purpose computing on Graphics Processing Units" Cómputo de propósito general basado en unidades de visualización
HA	High-Availability. Mecanismo para asegurar la disponibilidad de un servicio en caso de un fallo de un componente.
HBP	Human Brain Project. H2020 FET Flagship financiado por la Comisión Europea
HPC	High-Performance Computing.
HSM	Hierarchical Storage Management. Almacenamiento que puede usar varios niveles de tecnología (discos, cintas,…).
ICEI	Interactive Computing e-Infrastructure.
IOPS	Input Output oPerations per Second
LTFS	Linear Tape Filesystem
OpenStack	Plataforma de código abierto de cloud computing
PDU	Power Distribution Unit. Elemento de distribución eléctrica en un armario.

PRACE	Partnership for Advanced Computing in Europe.
RAID	Redundant Array Of Inexpensive Disks. Mecanismo para prevenir de fallo de un disco, almacenando información redundante en discos adicionales (mirror, paridad, …)
Spectrum Archive	Solución de archivado desarrollada por IBM
SPOF	"Single Point Of Failure". Parte de un Sistema que en caso de fallo, impide el funcionamiento normal del sistema a nivel global.
SPOM	"Single Point Of Management". Servidor(es) que proven de forma centralizada monitorización y servicios de administración.
Swift	Almacenamiento en objetos de la plataforma OpenStack.
UPS	"Uninterruptible Power Supply" o sistema ininterrumpido de alimentación
VM	Virtual machine / Máquina Virtual
Benchmark	Ejecución de una apliación para medir el rendimiento de un sistema informático hardware o software

8.3 Unidades

En este documento, la siguiente convención se ha usado para medir las capacidades de almacenamiento:

- 1 kilo-byte, escrito 1KB son 1,000 bytes;
- 1 mega-byte, escrito 1MB son 1,000 KB;
- 1 giga-byte, escrito 1GB son 1,000 MB;
- 1 tera-byte, escrito 1TB, son 1,000 GB;
- 1 peta-byte, escrito 1PB son 1,000 TB.

Estas unidades también se usan para medir el rendimiento, por ejemplo, 20GB/s corresponde a 20 giga-bytes (20 x 10^9 bytes) por segundo.