FENIX

RESEARCH INFRASTRUCTURE

Workload Management Requirements for an Interactive Computing e-Infrastructure

25-26.09.2018

ICEI team

BSC, CEA, CINECA, CSCS, JUELICH

Presenter: Sadaf Alam, CSCS

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

Introduction (I)

- Interactive Computing E-Infrastructure
  - Federated infrastructure for data, compute and additional services by five leading European HPC data centres (BSC in Spain, CEA in France, CINECA in Italy, CSCS in Switzerland and Juelich in Germany) (https://fenix-ri.eu/)

Introduction (II)

- Human Brain Project (HBP)
  - a research infrastructure to help advance neuroscience, medicine and computing (https://www.humanbrainproject.eu/)
- PRACE (Partnership for Advanced Computing in Europe)
  - enable high-impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness for the benefit of society (http://www.prace-ri.eu)

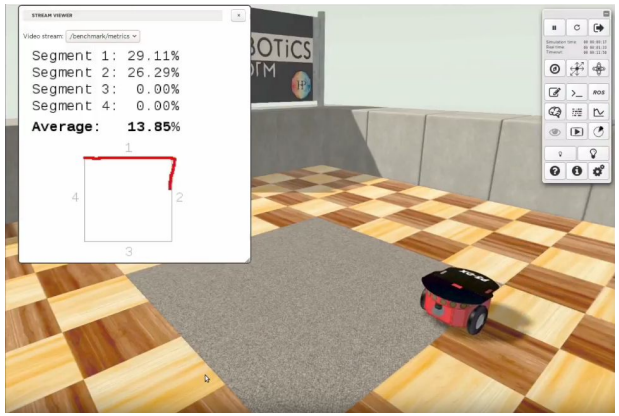
Fenix Services Implemented by ICEI

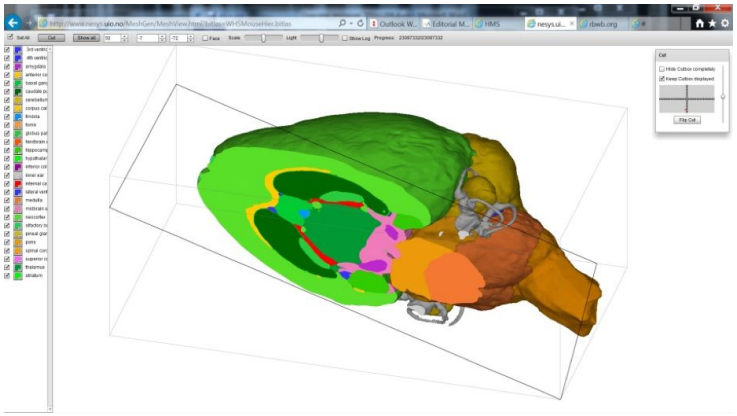
- Consumable and accountable services
  - **Interactive Computing Services**
  - Scalable Computing Services
  - Virtual Machine Services
  - Active Data Repositories
  - **Archival Data Repositories**

- Underlying and building block services
  - Internal interconnect
  - External interconnect
  - **Authentication/Authorization Services**
  - **Data Mover Services**
  - Data Transfer Services

Fenix Services Implemented by ICEI

- User and customer support services
  - Fenix User and Resource Management Service (FURMS)
  - Monitoring Services
  - User Support Services

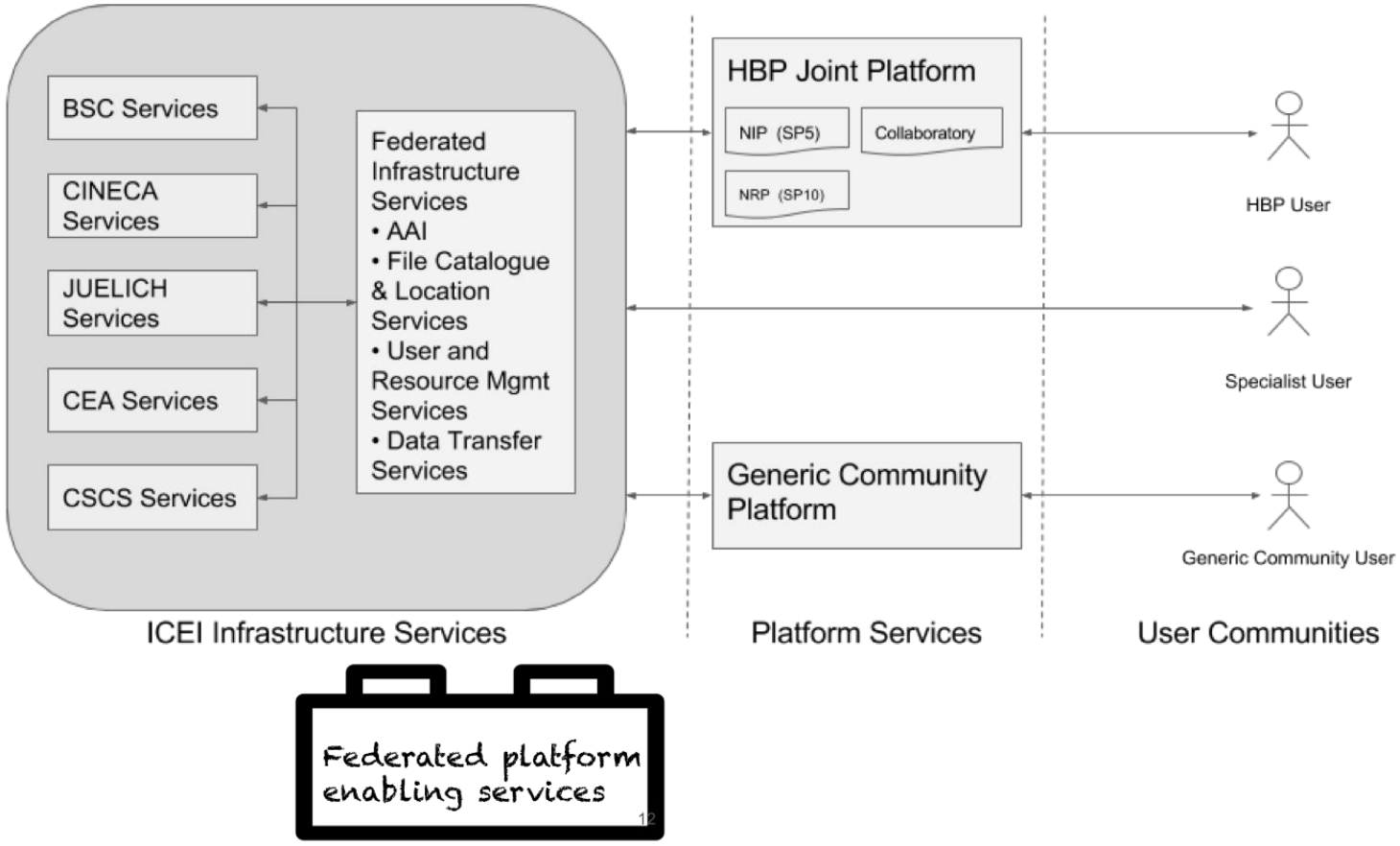




ICEI Services for PRACE 18th CFP

Component	Site (Country)	Total ICEI (100%)	PRACE (15%)	Minimum request
Scalable computing services				
Piz Daint Multicore	CSCS (CH)	250 nodes	38 nodes	1 node
Interactive computing services				
ICCP@JUELICH	JSC (DE)	175 nodes	26 nodes	1 node
Interactive Computing Cluster	CEA (FR)	60 nodes	9 nodes	1 node
Piz Daint Hybrid	CSCS (CH)	400 nodes	60 nodes	1 node
T.B.D.	CINECA (IT)	350 nodes	50 nodes	1 node
T.B.D.	BSC (ES)	6 nodes	1 node	1 node
VM services				
ICCP@JUELICH	JSC (DE)	25 nodes	4 nodes	1 VM
Openstack compute node	CEA (FR)	600 VM (20 nodes)	90 VM (3 nodes)	1 VM
Pollux Openstack compute node	CSCS (CH)	35 nodes	5.25 nodes	1 VM
Nord3	BSC (ES)	84 nodes	12.60 nodes	1 node
Archival data repositories				
Archival	CEA (FR)	7000 TB	1050 TB	0
Archival Data Repository	CSCS (CH)	4000 TB	600 TB	1 TB

Federated IaaS and Sites' Autonomy

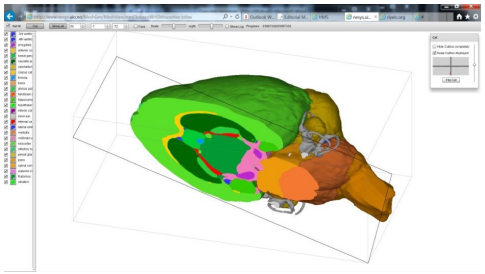


R&D for ICEI Services

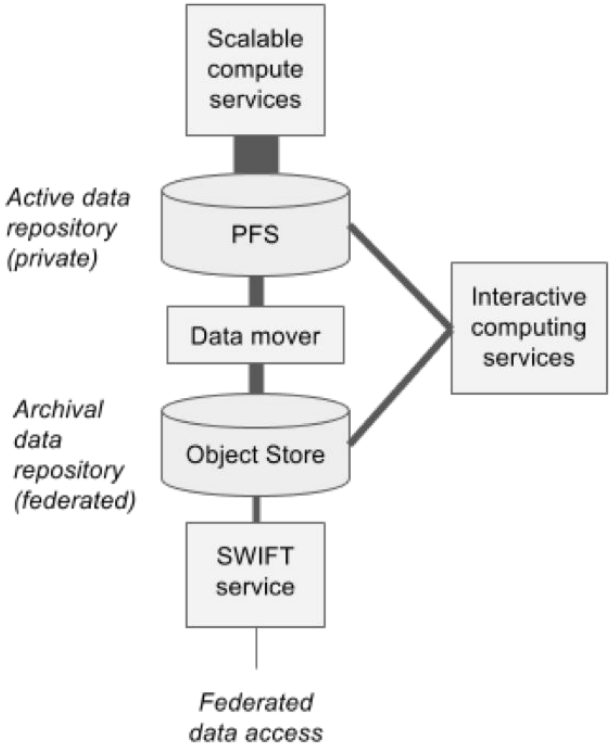
- Interactive Computing Service
  - Job preemption, suspension, etc. for multiple resource types
- Data mover service
  - Data movement (within one site) from POSIX to Object (OpenStack Swift)
- Authentication and Authorization Protocols Support
  - Secure access (REST API) for web services and OpenStack APIs

Interactive Computing Service

- Provide access to resources for interactive jobs, i.e. visualization of very large data sets
  - Brain atlas viewer
  - Neurorobotics platform
  - …



- Use cases
  - Interactive and in-situ visualization
  - JupyterNotebook elastic jobs
  - Post-processing
  - Computational steering



Interactive Computing Service

- Requirements
  - Support multiple resource types i.e. cores, GPUs, storage tiers, etc.
  - Balance interactive and batch jobs to maximize resources utilization via jobs suspension
  - Suspension and preemption of resources with QoS
  - NVM flush to free resources to create interactive sessions

- Extend SLURM gang scheduling and preemption e.g. NVMe unload/reload
  - Support interactive workflows similar to data science workflows

Data Mover Service

- Data movement terminology in Fenix
  - Within active storage “POSIX” file system (file transfer)
  - Data mover: Between active and archive within a site
  - Data transfer: Between archival storage of ICEI sites
- Use cases
  - Collaboratory (an HBP platform)
  - Neurorobotics platform
- Requirements
  - Swift (OpenStack object storage) for federate archival data repositories ⟵⟶ active data repositories with POSIX interface (e.g. HPC PFS storage)
  - Integration to slurm job by users
  - Maximize usage of compute and storage resources

Data Mover Service

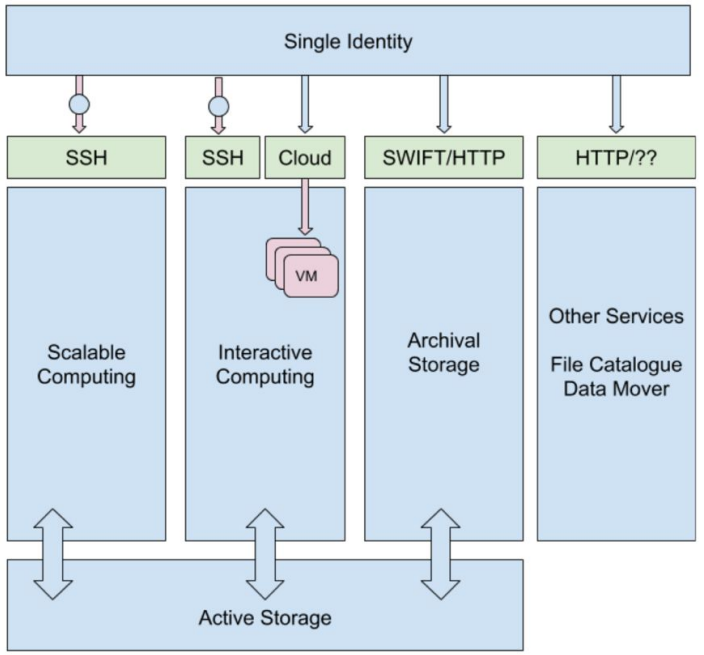
- Data movement terminology in Fenix
  - Data mover: Between active and archive within a site
  - Data transfer: Between archival storage of ICEI sites
- Data mover: Scalable hardware platform + software
  - Note: Different APIs/semantics
- Use cases
  - Data movement triggered via CLI program
  - Data movement triggered from app./workflow code
  - Data movement triggered by Slurm
    - Scheduling and execution of stage-in/stage-out (similar to BB)

Data Mover Service

- Requirements
  - Swift (OpenStack object storage) for federate archival data repositories ⟷ active data repositories with POSIX interface (e.g. HPC PFS storage)
  - Integration with Slurm
    - Handling of credentials/tokens and delegation necessary
    - Challenge: API, AAI credential/token management and delegation
  - Maximize usage of compute and storage resources
    - Policy engines and tuning/optimization algorithms for compute and storage resources

OpenStack Swift AAI Protocol Support

- Use cases
  - Scientific gateways web services → HPC + research/teaching tools
  - HBP Collaboratory and Neuroinformatics Platform
  - HBP Neurorobotics platform (http://www.neurorobotics.net)
  - Materialscloud (https://www.materialscloud.org)
- Requirements
  - APIs for OpenStack Swift authorization e.g. OAuth 2.0, OIDC
  - RESTful interfaces



- Extend semantics of something similar to the burst buffer technology with Swift API authentication & authorization constraints

Summary

- SLURM roadmap alignment
  - Maturity of existing solution (community input helpful)
  - Extension of solutions such as gang scheduling and preemption of multiple types of resources
- R&D efforts
  - Community driven (other sites enabling cloud based science gateways to HPC resources)
  - Use cases driven (identify communities and success stories)
  - Infrastructure driven (SLURM for an HPC cluster co-exists with OpenStack and Kubernetes services)

Abstract (for reference only)

The European ICEI project is co-funded by the European Commission and is formed by the leading European Supercomputing Centres BSC (Spain), CEA (France), CINECA (Italy), ETH Zürich/CSCS (Switzerland) and Forschungszentrum Jülich/JSC (Germany). The ICEI project plans to deliver a set of e-infrastructure services that will be federated to form the Fenix Infrastructure. The distinguishing characteristic of this e-infrastructure is that data repositories and scalable supercomputing systems will be in close proximity and well-integrated. The participating supercomputing centers have SLURM resource management and scheduling systems available on a diverse range of systems including high-end clusters and systems with accelerators. In this talk, we present key requirements by ICEI for the site workload managers, including features such as support for interactive supercomputing, integration of resources such as storage hierarchies, support for RESTful interfaces and an ability to handle credentials such as OAuth and/or SAML.