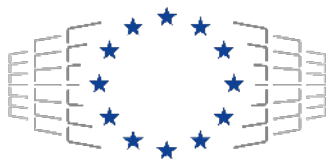
IO-SEA Overall Project Overview

IT4I CORTX Motr Workshop

December 5th, 2022

Sai.Narasimhamurthy@Seagate.com



EuroHPC
Joint Undertaking

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955811. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, the Czech Republic, Germany, Ireland, Sweden, and the United Kingdom.

Project Partners

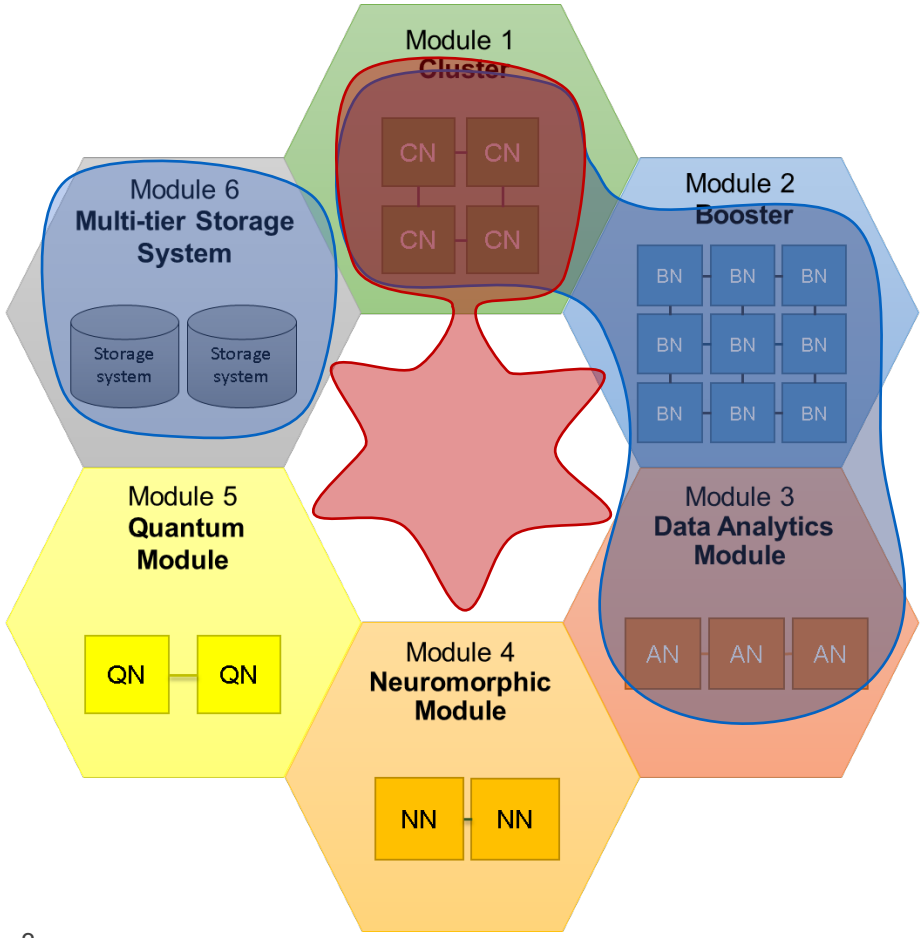
(In alphabatical order) 11 partners, 6 countries

- Atos-Bull (France)
- CEA (France) – **Project Coordinator**
- CEITEC (Czech Republic)
- ECMWF (International)
- Forschungszentrum Jülich (Germany)
- ICHEC (Ireland)
- IT4I (Czech Republic)
- JGU Mainz (Germany)
- KTH (Sweden)
- ParTec (Germany)
- Seagate (UK)

IO-SEA is part of the “SEA project family”
All addressing Modular Supercomputing Architectures

SEA Projects

Funded by the EuroHPC-01-2019 call



DEEP-SEA

Software stack for Exascale heterogeneity

IO-SEA

IO Software stack for the Exascale

RED-SEA

Network solutions for Exascale systems

IO-SEA

IO-SEA to tackle the IO challenges of the Exascale era

1. **Data Scalability:**
   Massive increase of the stored data and metadata
2. **System Scalability:**
   Increase of the number of clients to storage systems
3. **CPU/GPU evolution:**
   Inversion of the memory/core ratio: The operating system will have less available memory
4. **Data Placement:**
   Manage data locality and movements
5. **Data Heterogeneity:**
   Different workloads and different types of resources

Consequence:
currently used paradigms may **not scale** to Exascale.

The new tracks explored by IO-SEA

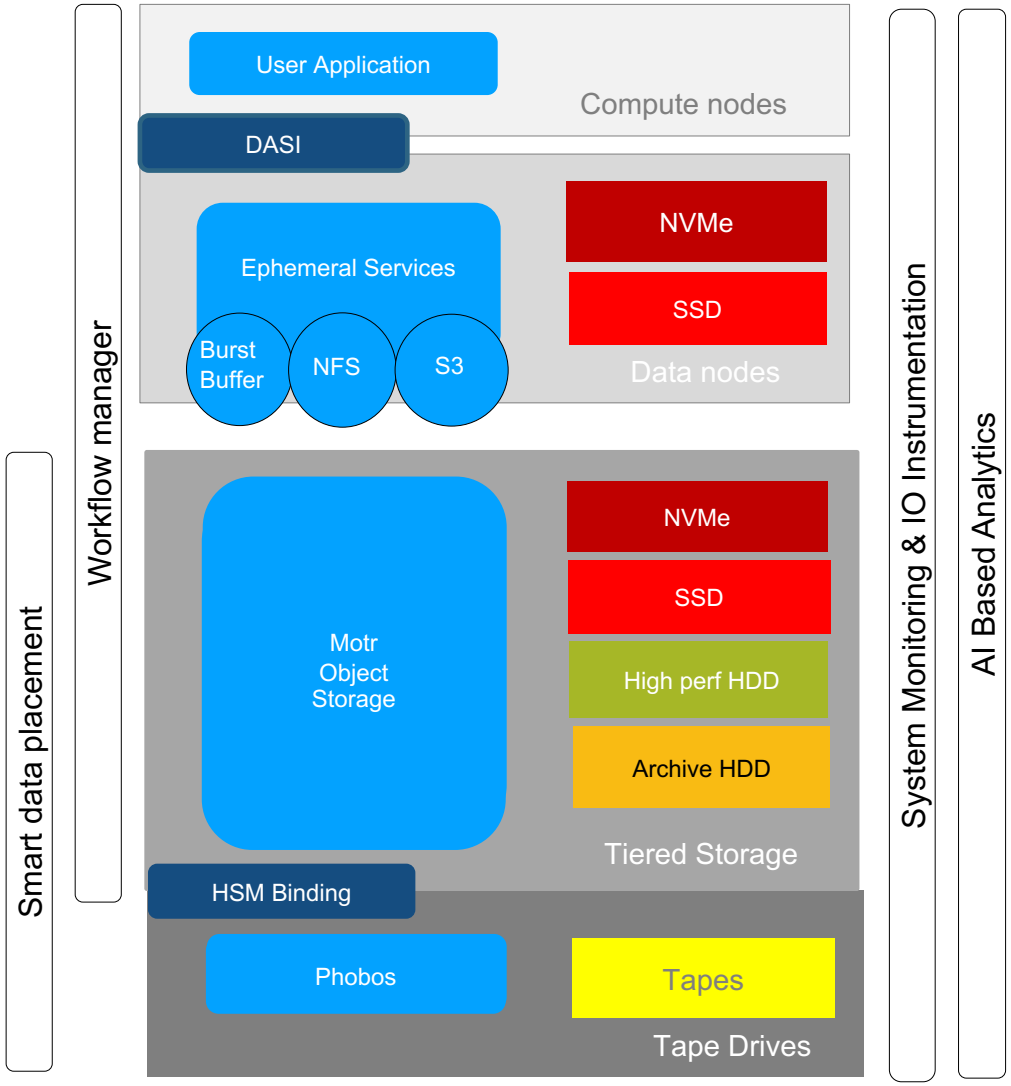
The IO-SEA software stack based on:

- Usage of **Object Stores** to store all data
  - **Hierarchical Storage Management** (HSM) to build an end-to-end storage stack
    - From very fast NVMe devices down to slow but capacitive tapes
  - **On-demand/Ephemeral provisioning** of storage services & **Scheduling**
    - IO servers are scheduled/spawned dynamically and are dedicated to a compute job
      - Running on specialised “data nodes”
      - Built on top of object stores
  - **IO Instrumentation** & AI based telemetry analytics
- Co-design with next generation I/O intensive HPC oriented applications
  - Development of new flexible application Interface (“**DASI**”)

The Big Picture: IO-SEA Architecture

- IO-SEA leverages software from the partners and open-source communities

  - PHOBOS
  - DEIMOS
  - Robinhood Policy Engine
  - CORTX-MOTR
  - NFS-Ganesha
  - Lustre
  - RADOS
  - Llview
  - ATOS IO Instrumentation
  - ATOS IO Pattern Analyzer
  - Parastation HealthChecker
  - Performance Predictor
  - ATOS Flash Accelerator
  - Parastation Management,
  - Ystia
  - IFS
  - MultIO
  - FDB
  - PGEN
  - Kronos
  - TSMP
  - COSMO
  - CLM
  - Parflow
  - HYPRE
  - netCDF
  - HDF5
  - SILO,
  - Lapack
  - GRIBAPI



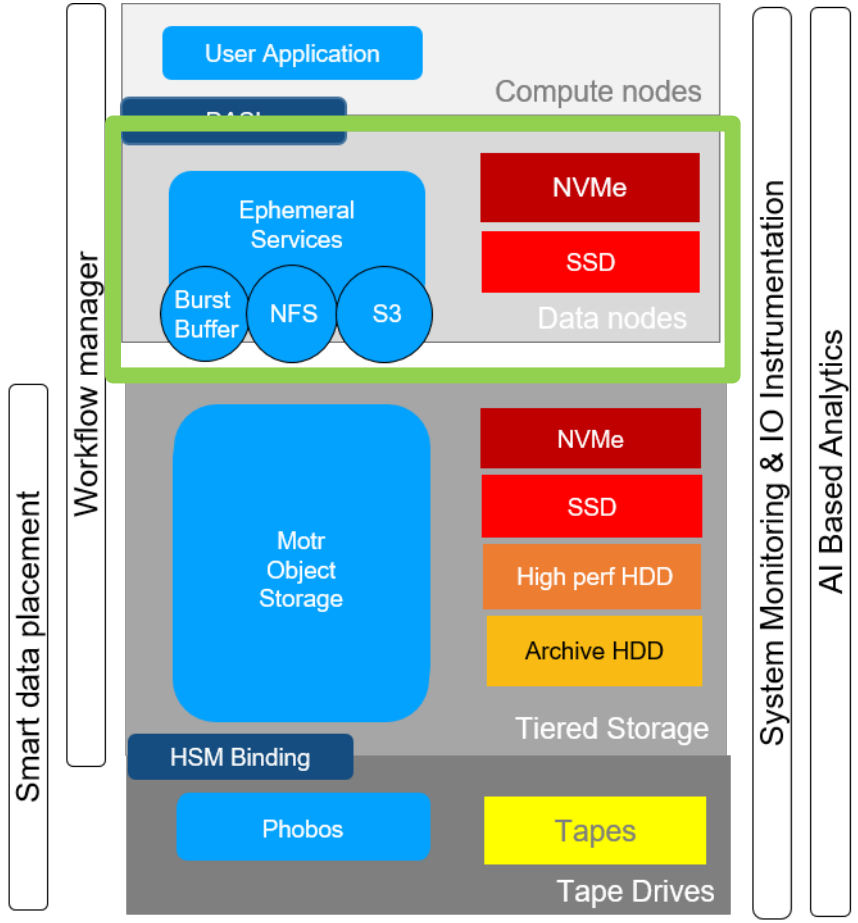
The IT4I platform and the DEEP testcluster

The software is developed using two different infrastructure:

- The IT4I platform offers a cluster of virtual machines used for developing the new software and test it
- The DEEP test cluster is used to perform the integration of all the pieces of software
  - This architecture leverages the prototype used during the Sage2 project, now refurbished as the “IO-SEA prototype”
  - This infrastructure demonstrates the MSA
  - Benchmarking will be done on that architecture after M18

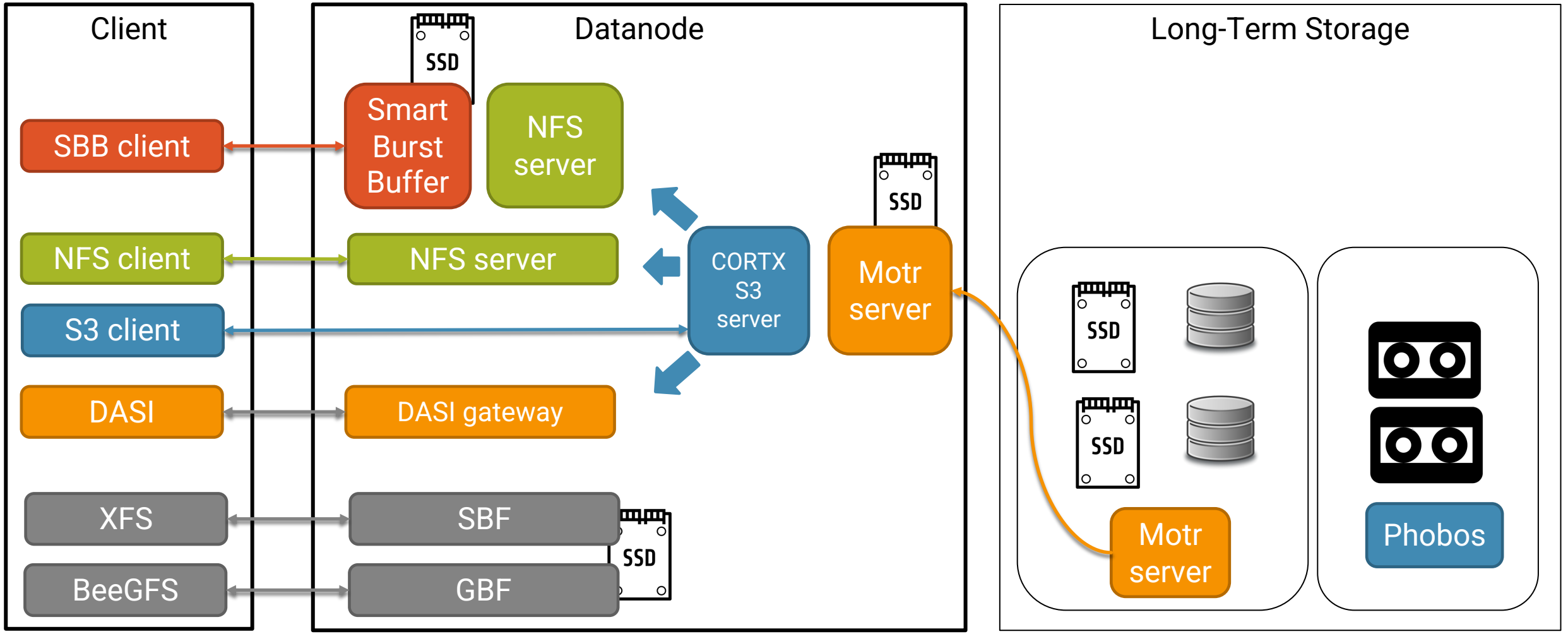
Ephemeral Data Access Environment Exposing data

- Specialised data access environment suitable for applications and workflows
  - Goal: lower the pressure on actual storage system
- Will leverage NVMe resources available on data nodes
- Provides mechanisms to schedule data accesses on demand



Ephemeral I/O Services

- **NFS, BB-NFS, S3, DASI, SBF, GBF**



Datasets/Namespaces Implementation

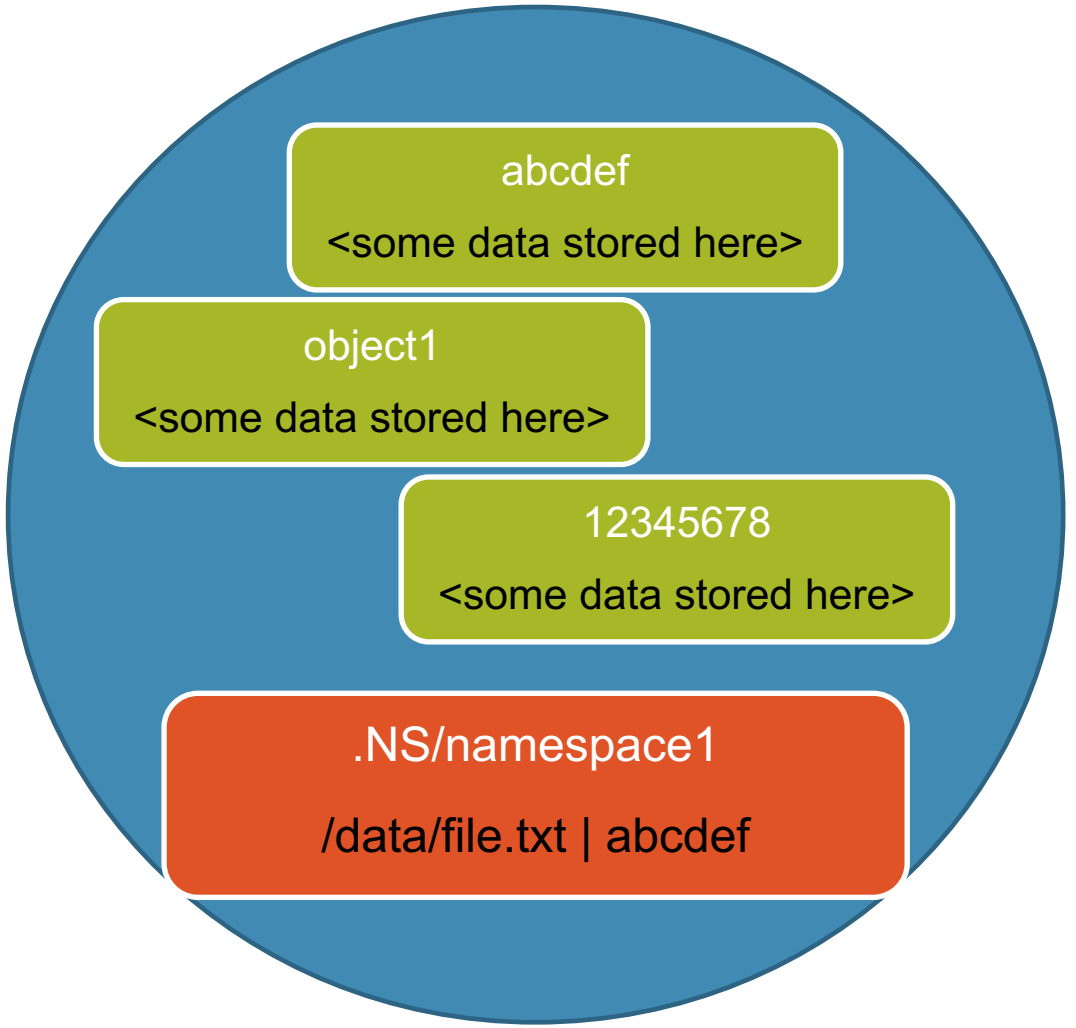
A Dataset is stored as 1 S3 Bucket

Files stored as S3 object

Namespaces stored in « metadata » objects

Think of it as a set of directories and symbolic links to objects in dataset

- .NS/namespace1
- .NS/namespace2

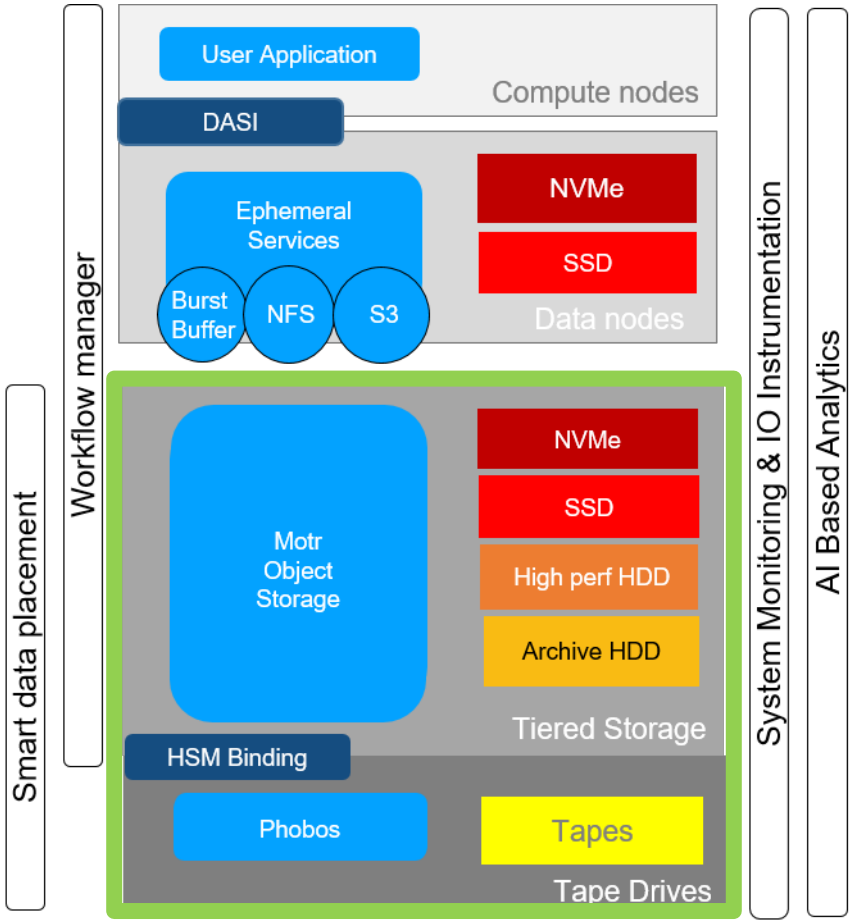


My_dataset

HSM Features:
The right data at the right place

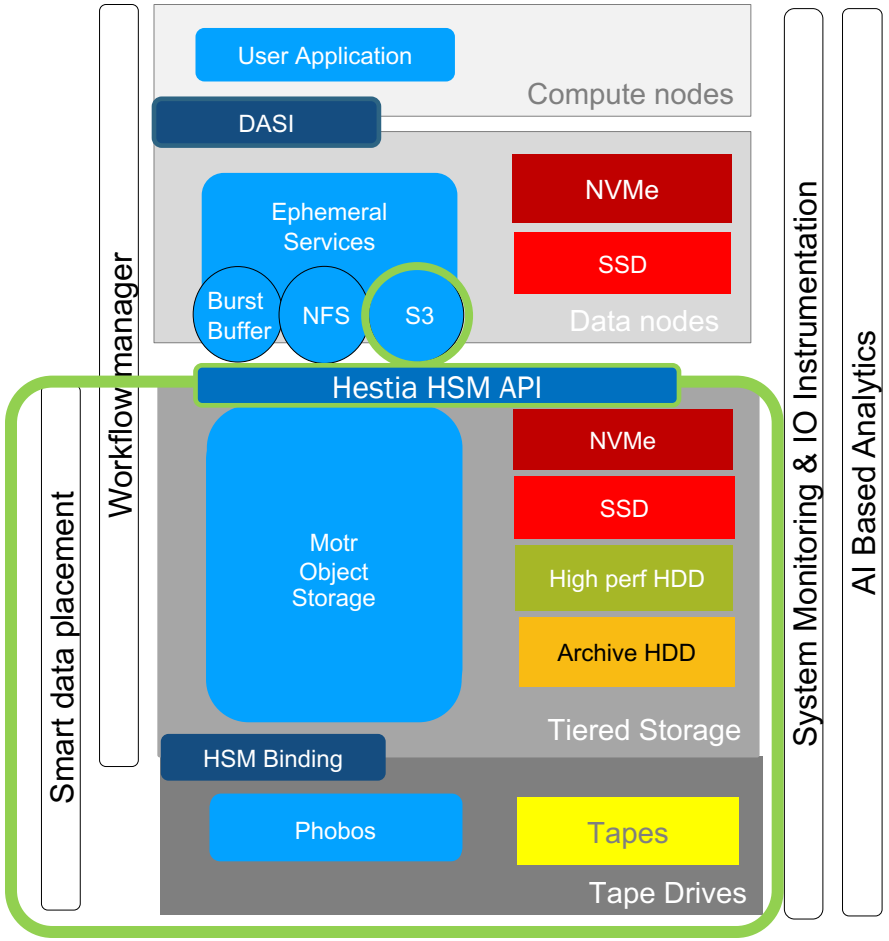
HSM Mechanism for managing data movements between multiple tiers of Persistent Storage tiers, such as:

- NVMe
- SSD
- Disk
- Tape

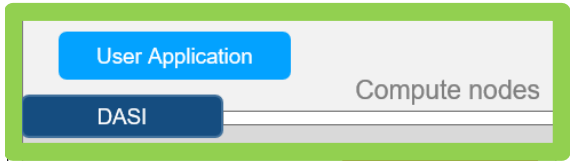


HSM Interfaces

- 2 interfaces to applications and higher layers:
  - Hestia HSM API
    - Low-level API (C++) to manage data placement between the storage tiers
    - Usable on a Motr client (data node)
  - An S3 server (HTTP-based) provides a higher level interface including:
    - Namespace management
    - User authentication
    - Control of access rights
    - Remote access from any client (data node or compute node)

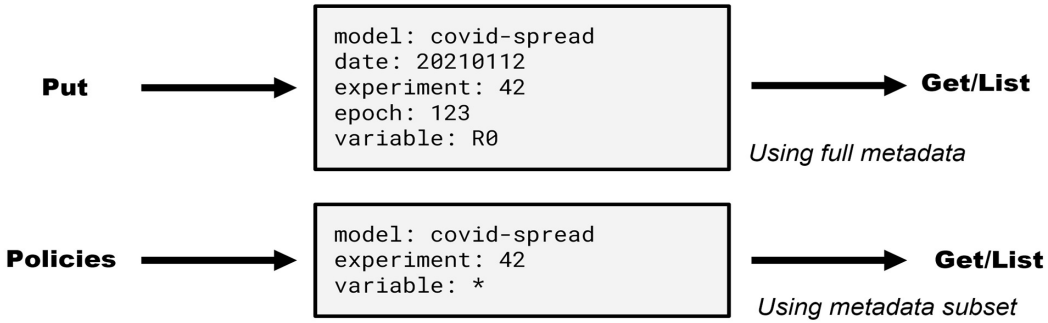


Application Interfaces and DASI: Where users meet their data



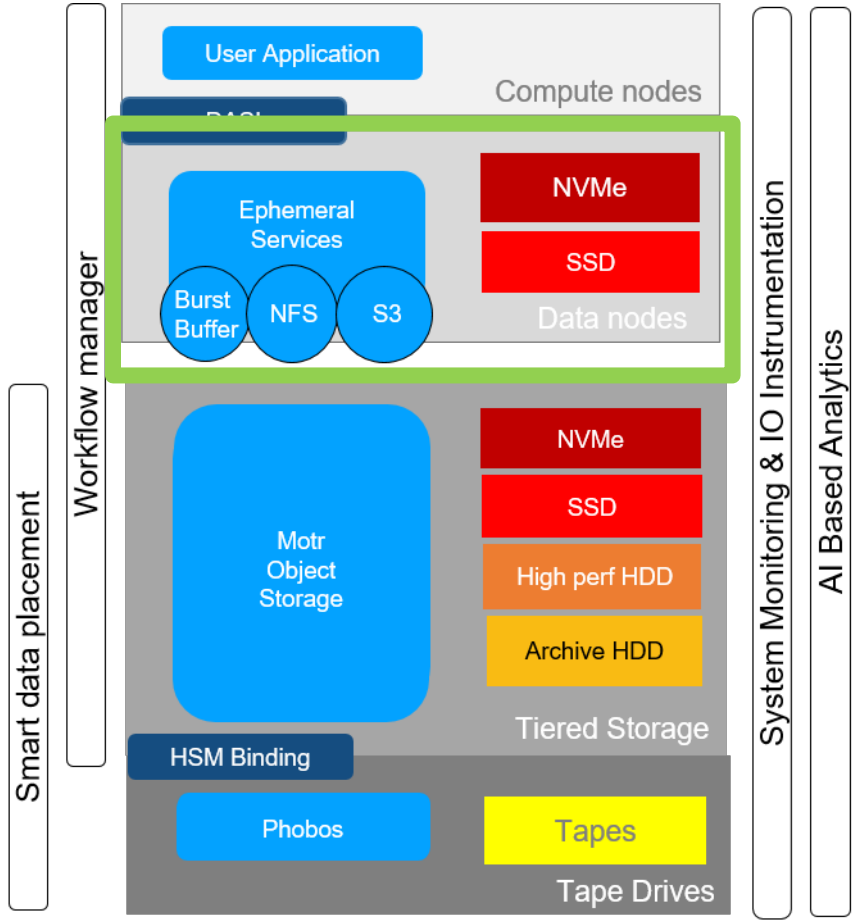
Development of the ephemeral services:

- Data Access and Storage Interface (DASI) Layer (Language) will be built to abstract the complex storage layer
  - Using semantic description of data – speaking the language of the scientific domain



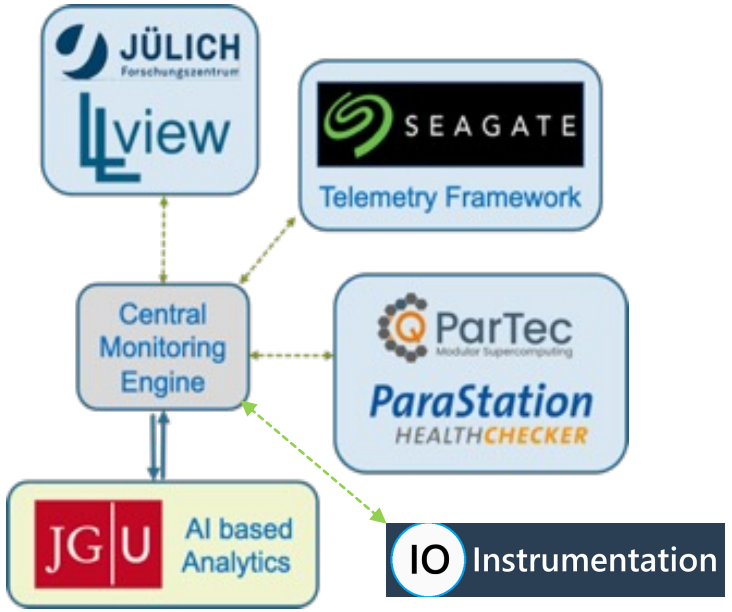
Ephemeral Data Access Environment Exposing data

- Specialised data access environment suitable for applications and workflows
  - Goal: lower the pressure on actual storage system
- Will leverage NVMe resources available on data nodes
- Provides mechanisms to schedule data accesses on demand



Instrumentation & Monitoring: Knowing what happens in the system

- Gathering knowledge on I/O behaviour of applications & workflows
  - Analyse collected data using AI based techniques
- Knowledge will feed algorithms that will allocate I/O services & data nodes resources
- Gathering knowledge about infrastructure resources to make efficient scheduling decisions
  - AI algorithms will complement scheduling decisions made by users
- I/O & instrumentation tools will be adapted to each protocol (S3, NFS, POSIX, etc)

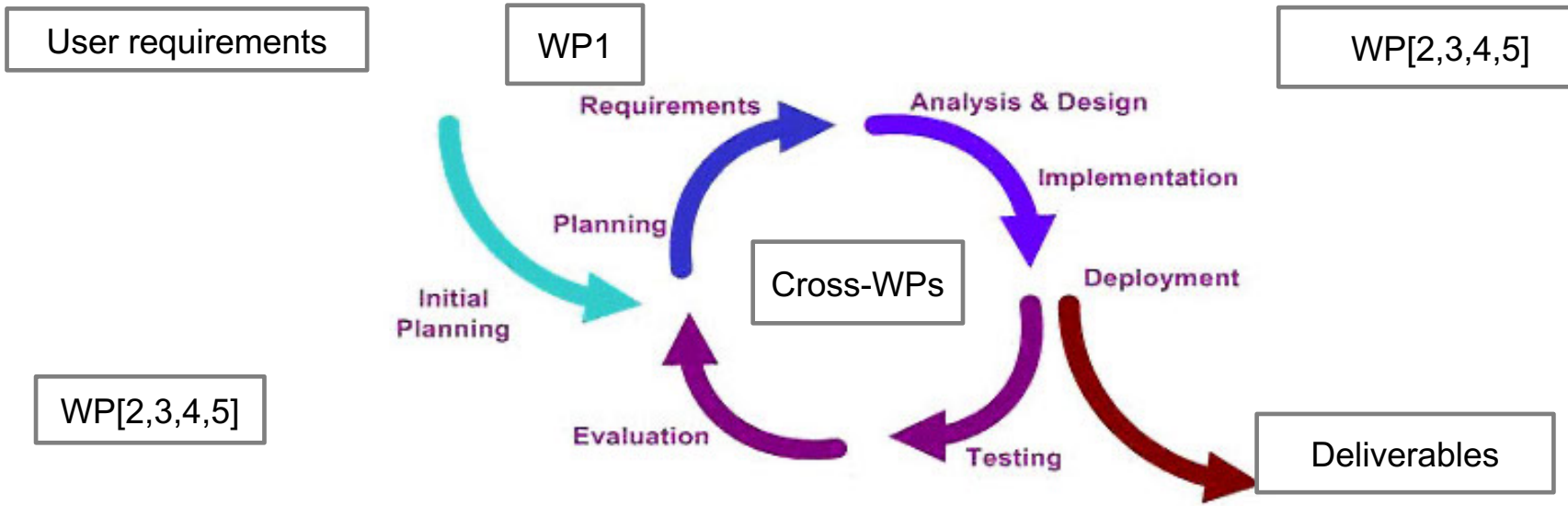


Components of the monitoring infrastructure

IO-SEA

Focus: IO-SEA's organisation: co-design and cross-WP session

- Co-design is a keystone to IO-SEA
  - WP1 provides input about data patterns and user needs
  - WP[2,3,4,5] design concepts and solutions
- Cross-WP sessions to keep this process active
  - Key concepts (dataset, namespace, workflow, ..) were jointly defined via « design meeting »



Focus: Collaboration with other projects within EuroHPC-19-1

- With our SEA-friends
  - All projects rely on the **MSA architecture**
  - Via joint use cases and joint topics on benchmarking, traces, monitoring , …
  - Regular “ALL-SEA” meetings, including meeting with the PO
- Explicit collaboration with ADMIRE
  - IO-SEA and ADMIRE are “IO flagships” in EuroHPC-19-1
  - Collaboration will built a collection of IO Traces in HPC
  - Currently involving 6 out of 11 projects
- The IO-SEA software stack will used and deployed on the EUPEX pilot

The project's main outcomes as of now

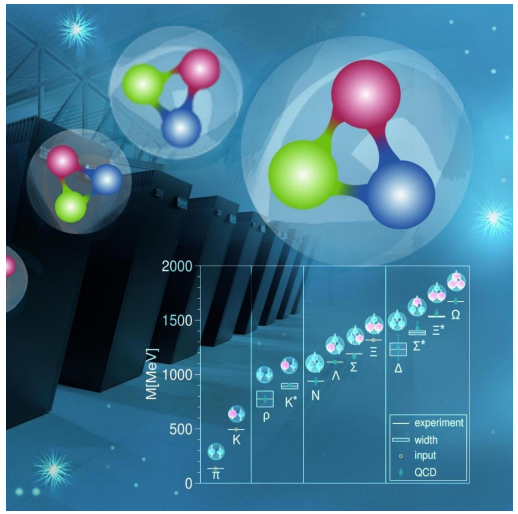
Ready and delivered (as alpha versions):

- Workflow / data nodes and ephemeral services are available on IT4I infrastructure
- The different monitoring tools are integrated and deployed on DEEP cluster
- Hierarchical Storage Management API: HESTIA is available
  - HESTIA integrates HSM in object stores
- First version of the DASI is delivered
- First version of ephemeral services are ready to be spawned by workflow

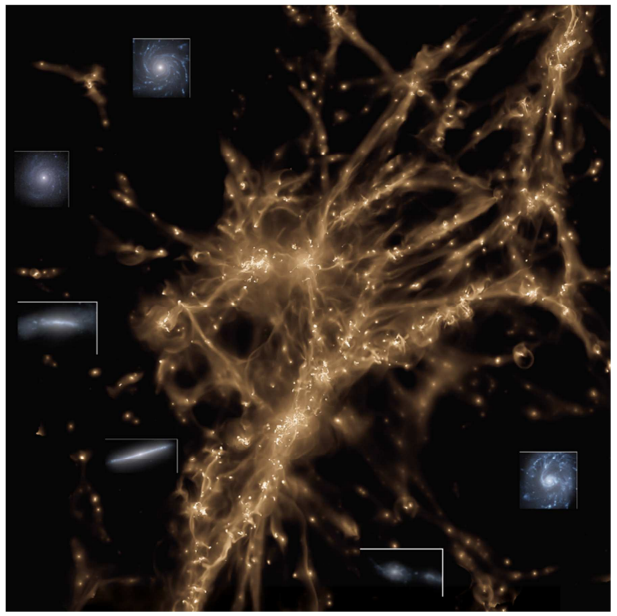
→ Ready to be benchmarked on DEEP cluster (2nd half of the project)

Validated by use cases

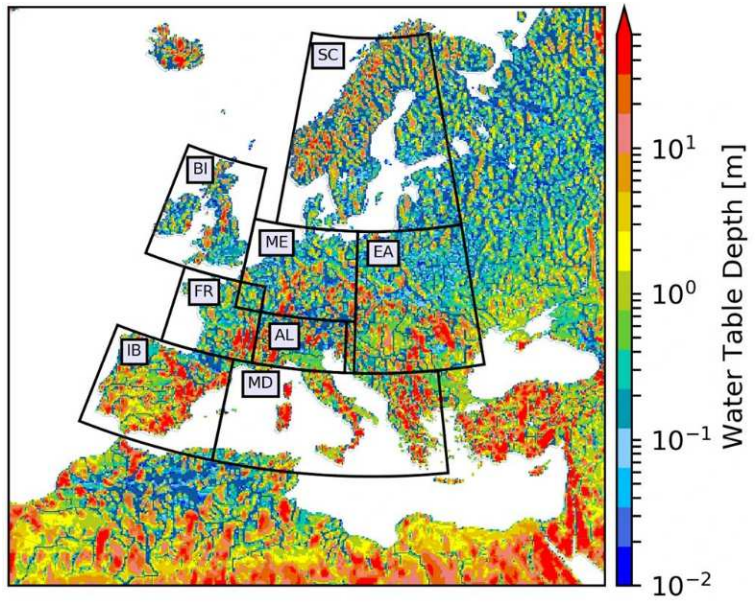
The Applications



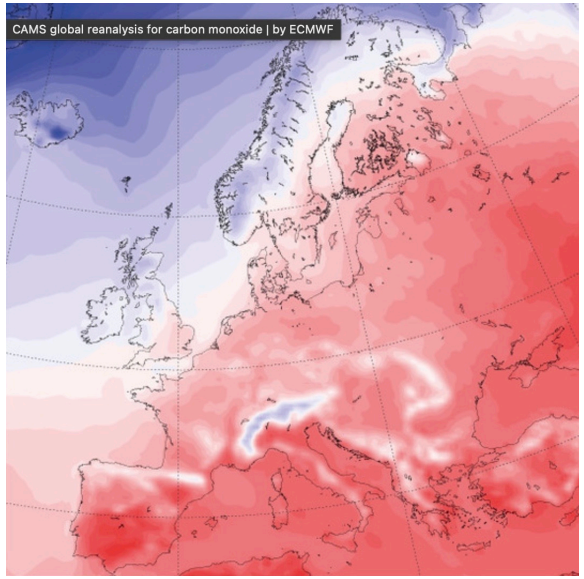
Lattice QCD (Particle Physics)



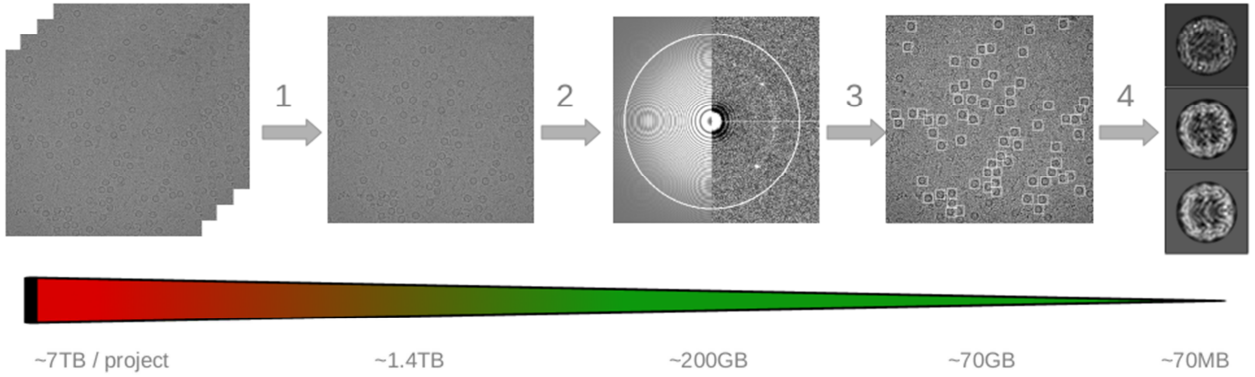
RAMSES (Astrophysical Systems)



Terrestrial Systems Modeling (TSMP)



Weather Forecasting Workflows



Electron Microscopy Imaging

IO-SEA

Questions?