[Project timeline: April 2021 – March 2024]

IO-SEA

IO Software for Exascale Architecture

Sai Narasimhamurthy (ParTec)

Dissemination and Exploitation co-lead, IO-SEA

Presented at the ALL SEA Workshop, LRZ, 16th Jan’2024

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.

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

- **Data Scalability:**
  Massive increase of the stored data and metadata
- **System Scalability:**
  Increase of the number of clients to storage systems
- **CPU/GPU evolution:**
  I/O and storage not keeping up with the rapid progress in heterogeneity and parallelism
- **Data Placement:**
  Manage data locality and movements across multiple tiers
- **Data Heterogeneity:**
  Different workloads and different types of resources

<u>Consequence:</u>
Currently used I/O paradigms will **not scale** to Exascale and beyond.

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

DEEP-SEA

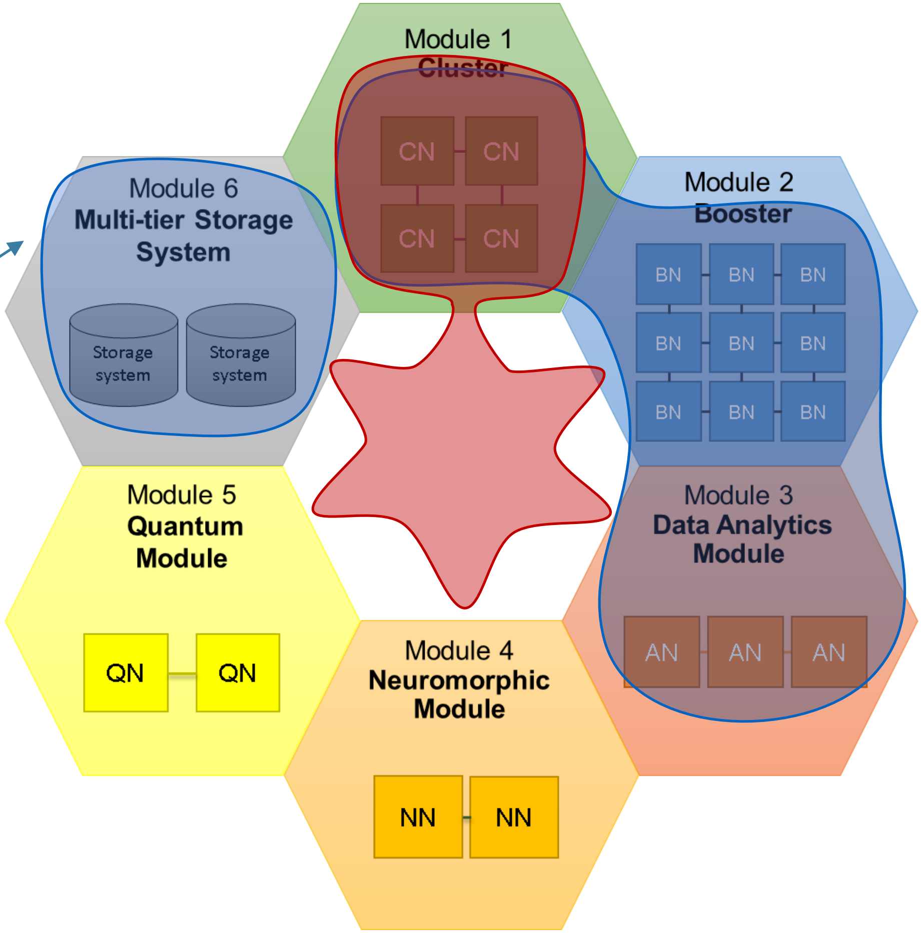
Software stack for Exascale heterogeneity

IO-SEA

IO Software stack for the Exascale

RED-SEA

Network solutions for Exascale systems

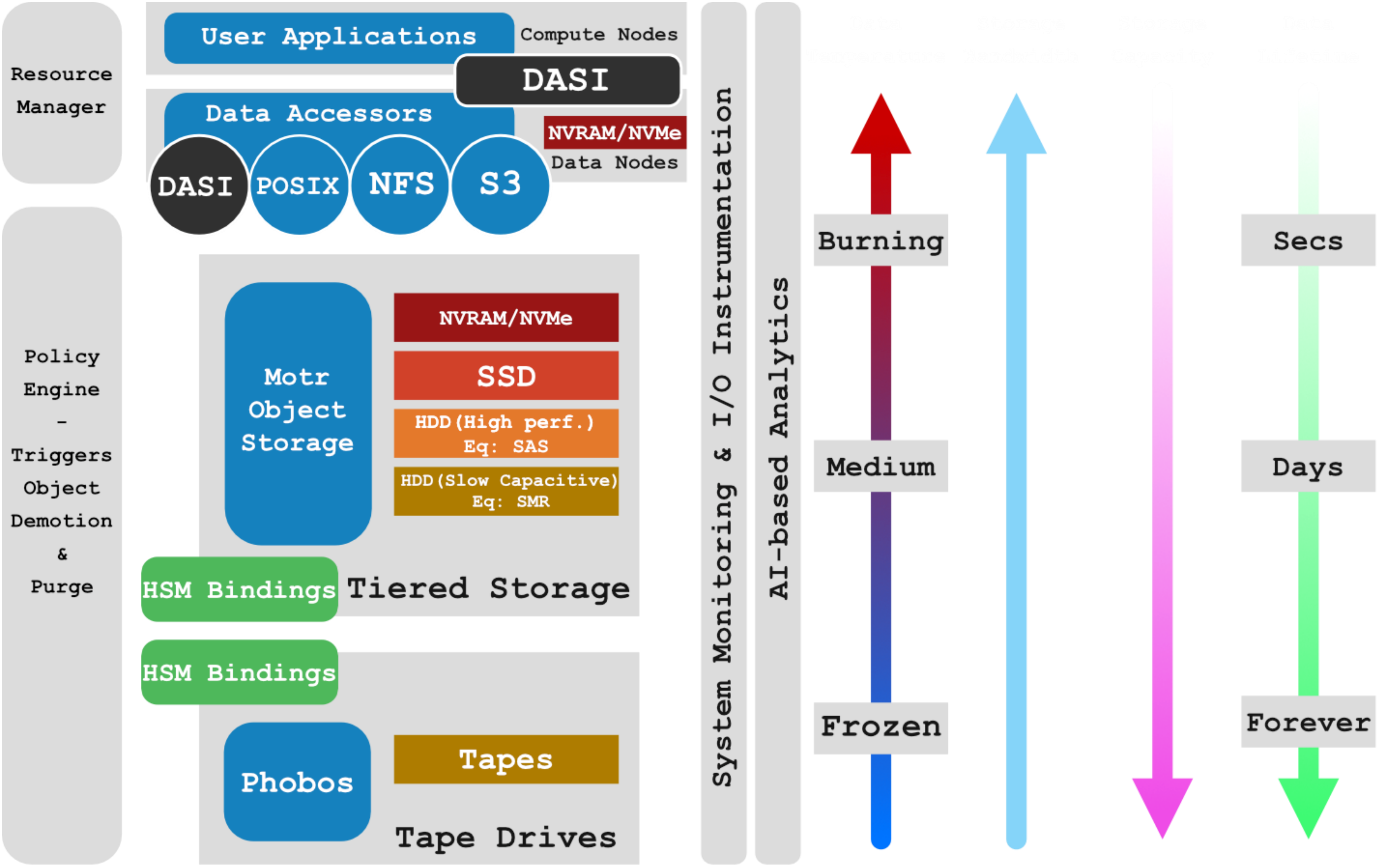


The areas explored by IO-SEA

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



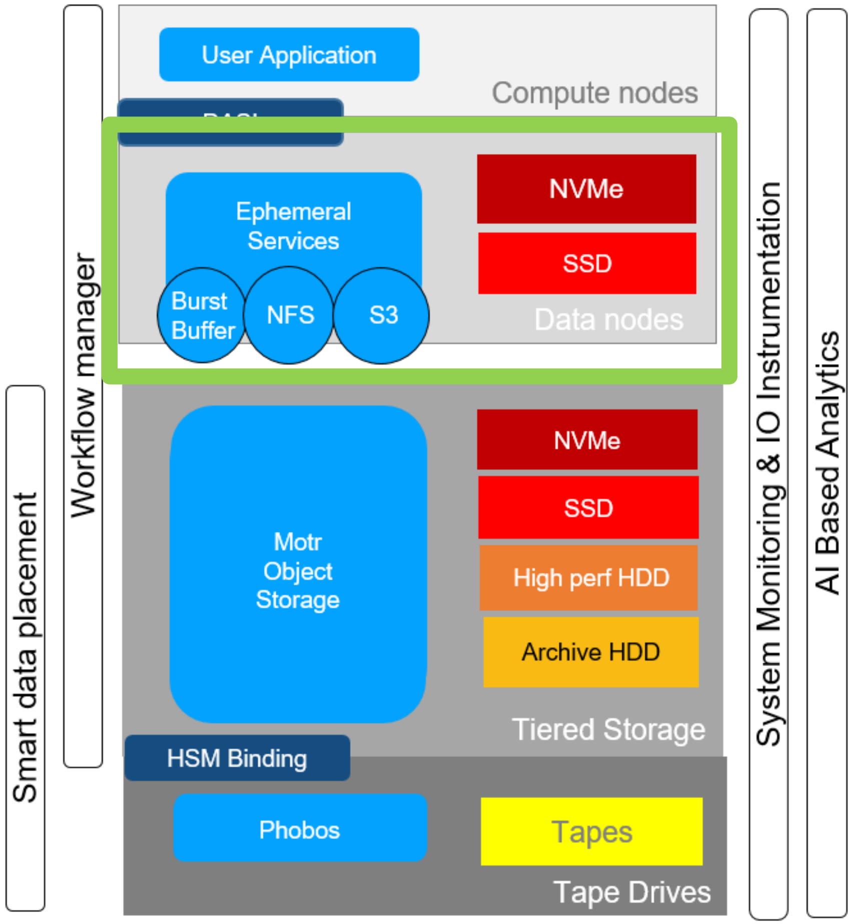
IO-SEA

Ephemeral Data Access Environment

The data nodes

- 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
- Data nodes sit at the interface between the

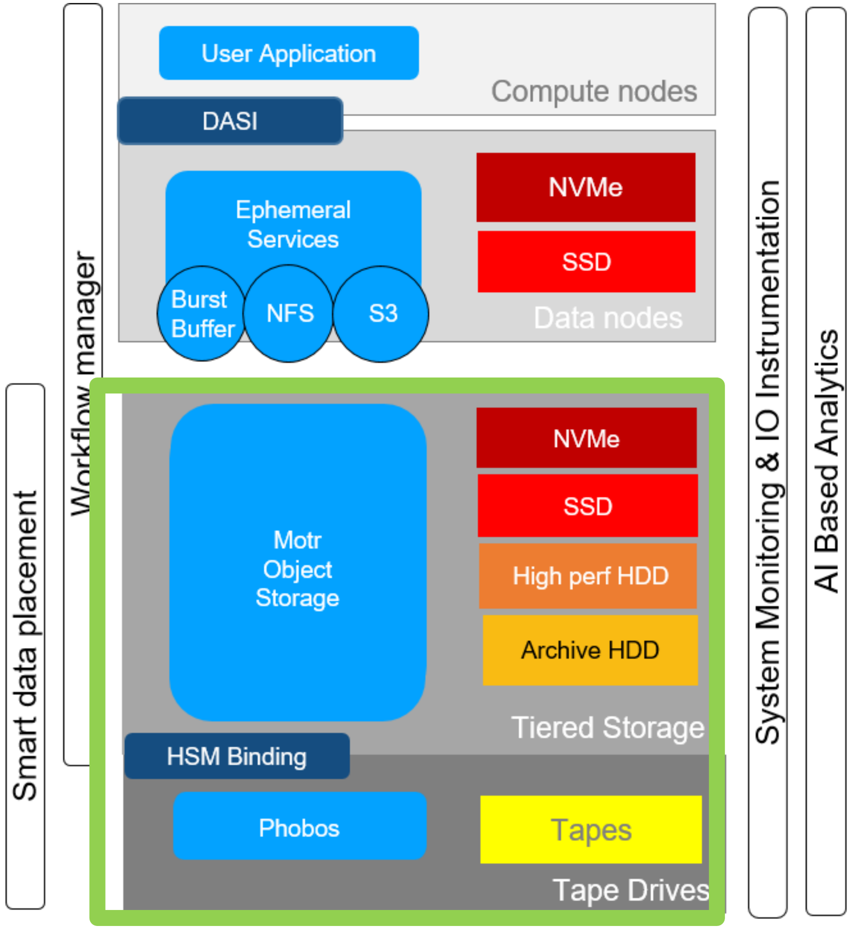
Modules



HSM Features

The right data at the right place

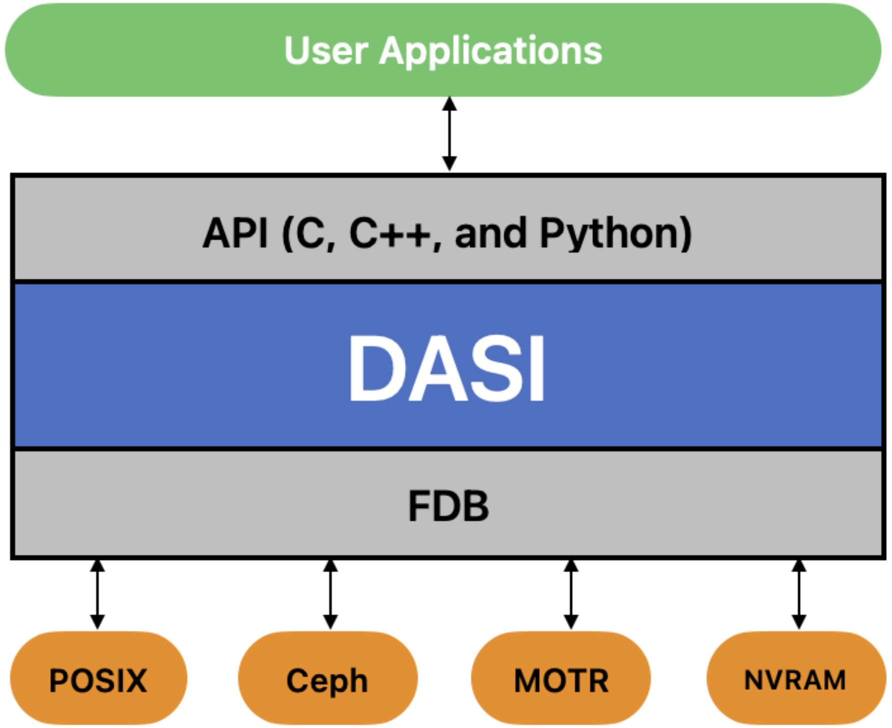
- HSM Mechanism for managing data movements between multiple tiers of Persistent Storage tiers, such as:
    - NVMe
    - SSD
    - Disk
    - Tape

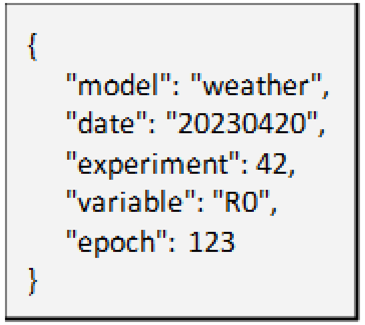


Application Interfaces and DASI:

Where users meet their data

- DASI provides an abstraction for scientific data handling
  - granting access to the underlying complex storage mechanisms
  - is simple for application developers to adopt and understand.
- The key feature of DASI is that it provides a “semantic interface” for data,
  - data is indexed and uniquely identified by set of scientifically-meaningful metadata keys.





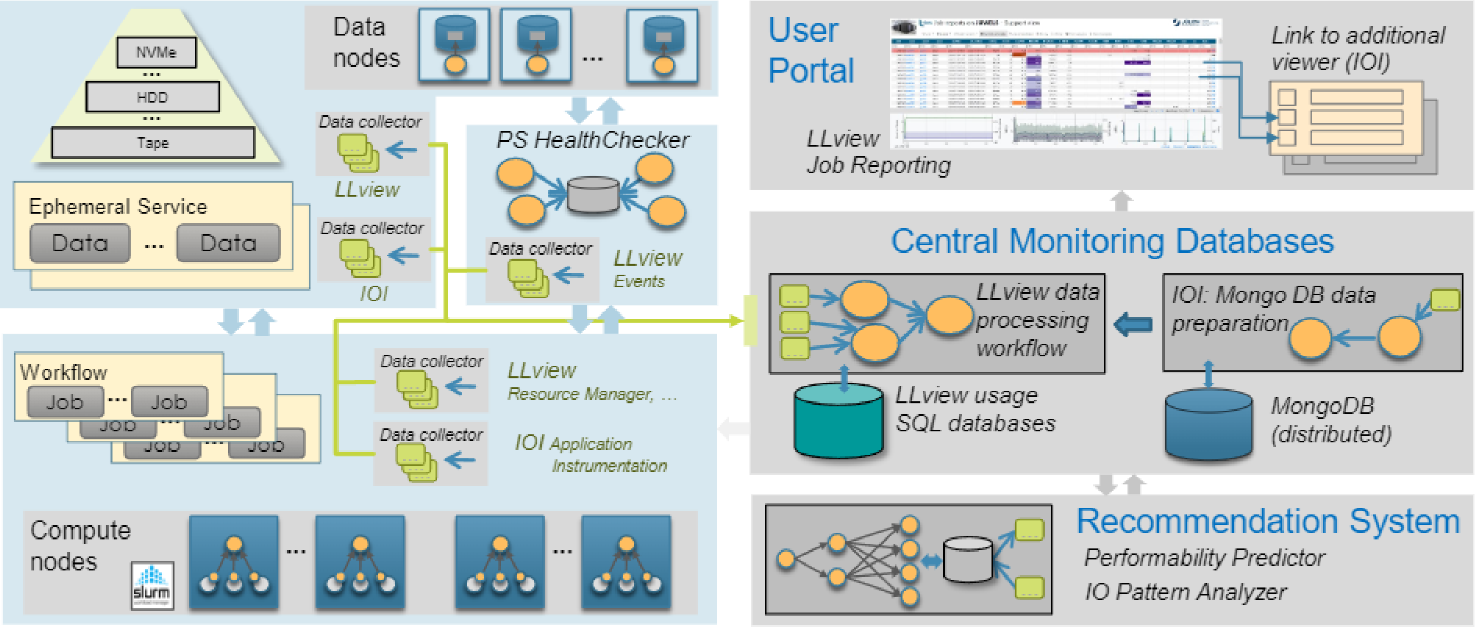
```
{
    "model": "weather",
    "date": "20230420",
    "experiment": 42,
    "variable": "R0",
    "epoch": 123
}
```

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

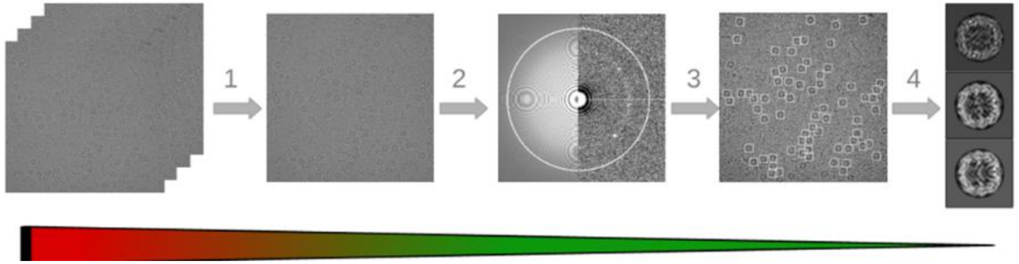
- Gathering knowledge about infrastructure resources to make efficient scheduling decisions
  - AI algorithms will complement scheduling decisions made by users
- I/O & instrumentation tools will be adapted to each protocol (S3, NFS, POSIX, etc)

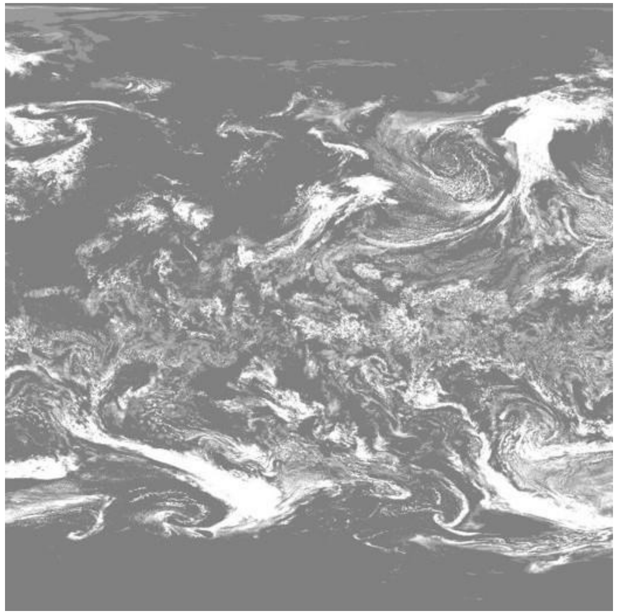


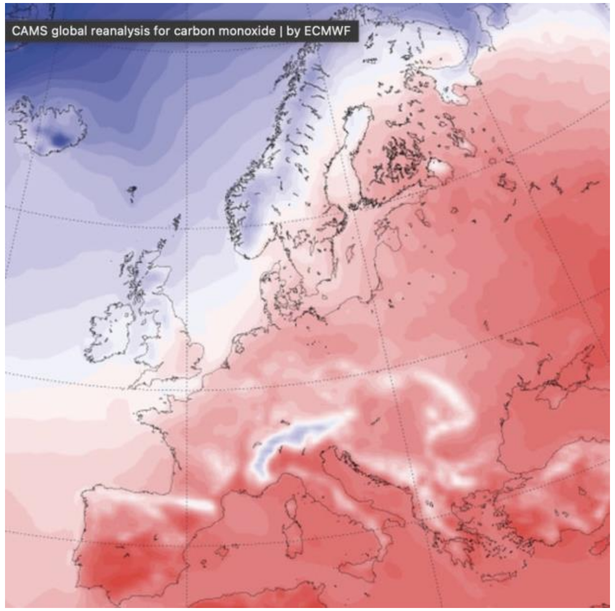
IO-SEA Applications

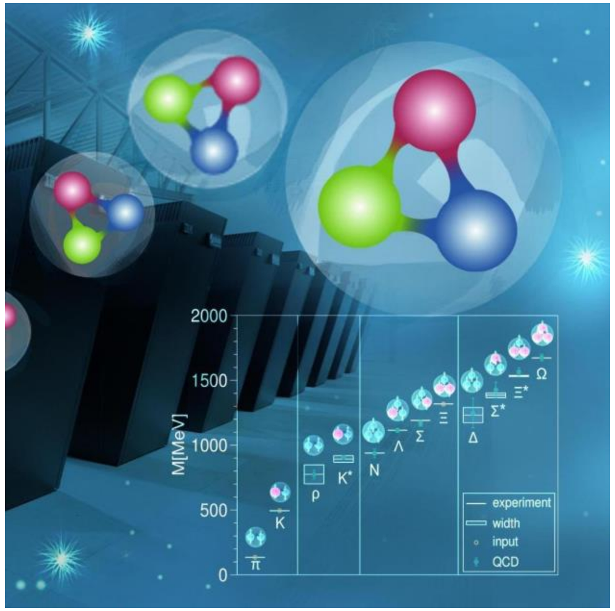
Co-design

- Data Intensive applications employing the IO-SEA Stack
  - TSMP: Multi-Physics Regional Earth Systems Model
  - ECMWF Weather Forecasting workflow
  - Cryo-electron microscopy imaging
  - Lattice QCD









Data Organisation in IO-SEA

In a nutshell

- **Objects** are the fundamental data units
- **DASI** provides scientifically meaningful views on this data for use cases
- Objects are grouped into semantically meaningful **Datasets**
- **Namespaces** impose organization on objects within Data Sets
- **Ephemeral services** are spawned to create and work with namespaces
- Data sets are managed in the hierarchy, moved and archived through **Smart Data Placement**
- **Workflow Manager** specifies appropriate services on behalf of use cases

Questions?

sai.narasimhamurthy@par-tec.com

[Backup Slides]

Test Infrastructure

The IT4I platform and the DEEP testcluster

The software is developed using two different infrastructures:

- 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

The project’s main outcomes at this stage

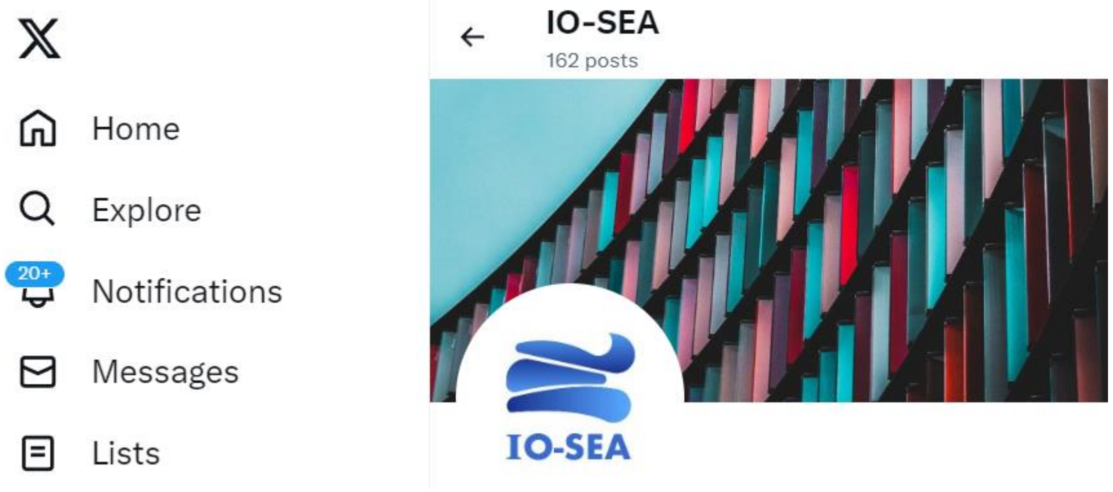
- Workflow manager exposing Ephemeral services deployed on DEEP system
  - Will include HSM API (Hestia) & DASI
- Workflow / data nodes and ephemeral services also available for test on IT4I infrastructure
- The different monitoring tools are integrated and deployed on DEEP cluster

} Validated by use cases

→ More Tests/demos on DEEP system ongoing

→ Work on feeding the outcomes of IO-SEA into EUPEX

Resources



Twitter: @iosea_eu

Youtube channel

Web: https://iosea-project.eu

Collaboration with other projects within EuroHPC-19-1

- With our SEA-friends
  - All projects rely on the **MSA architecture**
  - DEEP system usage for IO-SEA
  - Joint use cases

- 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 be used and deployed on the EUPEX pilot