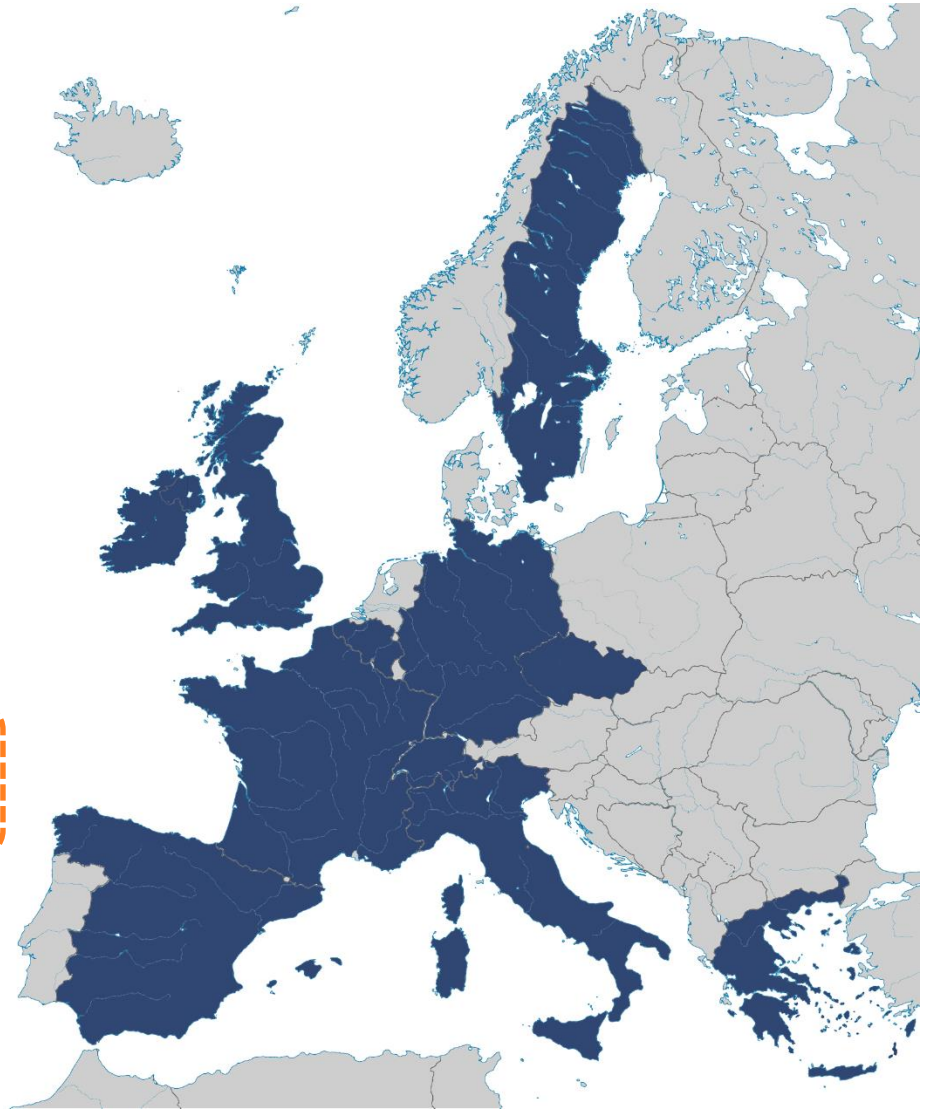
SEA Projects towards modular Exascale systems: advances in SW, storage systems and interconnects

HiPEAC 2024 Industrial Session

Hans-Christian Hoppe (Jülich Supercomputing Centre)

The SEA Projects (April 2021 – March 2024)

SEA Projects



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

cea

CEITEC
Central European Institute of Technology
BRNO | CZECH REPUBLIC

TECHNISCHE UNIVERSITÄT DARMSTADT

ECMWF

ETH zürich

EVIDEN
an atos business

EXAPSYS

exact

EXTOLL
latency matters.

FORTH

Fraunhofer ITWM

NUI Galway
OÉ Gaillimh

INFN
Istituto Nazionale di Fisica Nucleare

JÜLICH
Forschungszentrum

KATHOLIEKE UNIVERSITEIT LEUVEN

KTH
VETENSKAP OCH KONST

Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

JGU
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

UCLM

Technical University of Munich

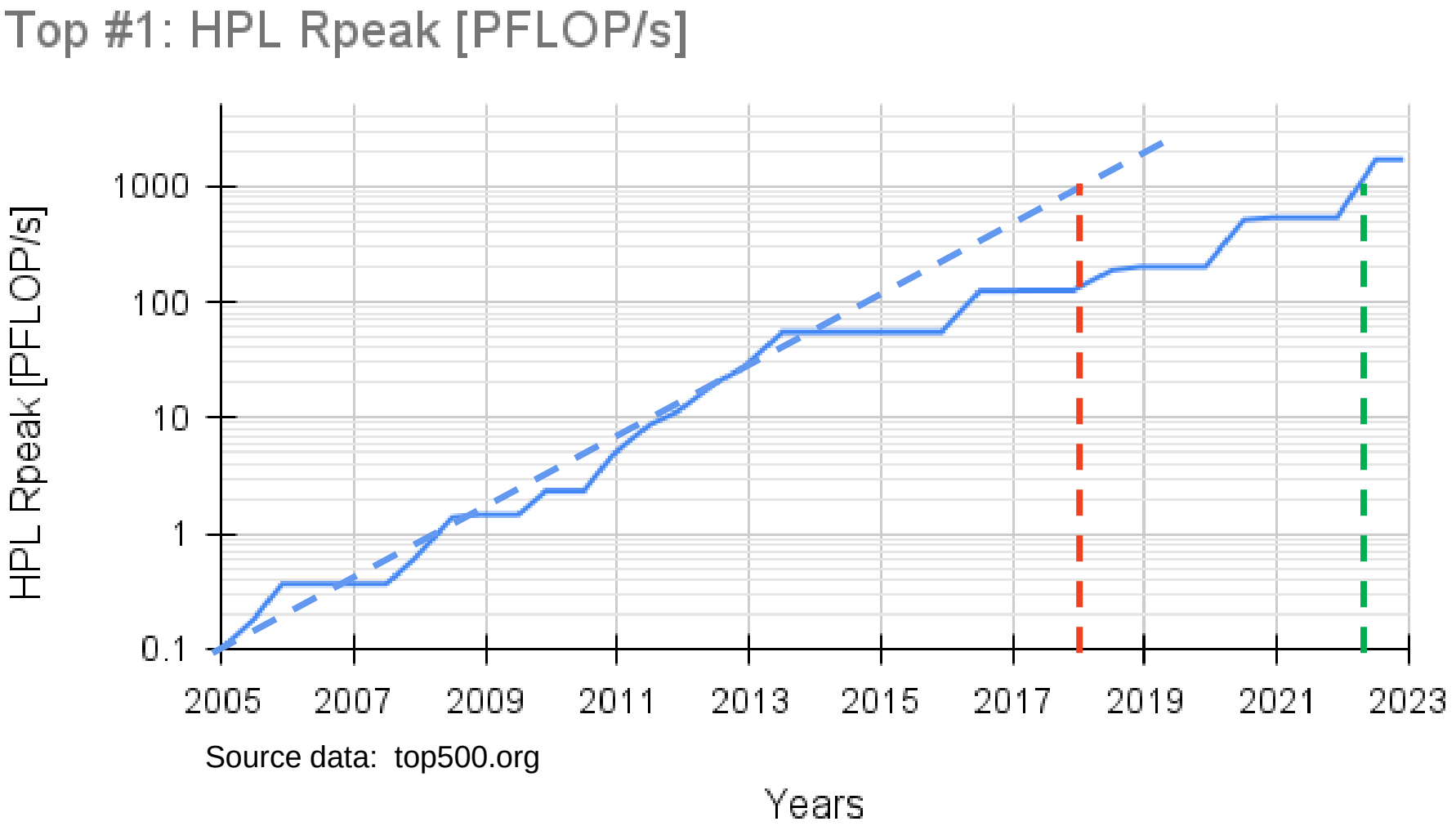
ParTec
MODULAR SUPERCOMPUTING

SEAGATE

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Industry partners

HPC Peak Performance Evolution



- **1997:** First **1TFlop/s** computer: (*ASCI Red/9152)*
- **2008:** First **1 PFlop/s** computer: (*Roadrunner*)
- So…. First **1 EFlop/s** computer: **2018 !!**
  - Well… not really
- It took 4 years longer…. **2022** for *Frontier* to appear

Exascale Challenges

Application parallelism

- Applications must support billions of individual threads
- Lower-scaling applications / parts of applications must not run on a full Exascale system

DEEP-SEA

Truly scalable systems

- Huge numbers of devices need to exchange data with each other
- Collective communication operations are “slowing down” due to larger system sizes
- Network contention and reliability become worries

RED-SEA

Energy efficiency

- Accelerators clearly beat CPUs for many (most?) codes
- System heterogeneity is a must
- Yet – portable accelerator programming is hard

DEEP-SEA

IO-SEA

RED-SEA

Memory and storage

- Ever growing gap between compute throughput and memory bandwidth
- New technologies like HBM suffer from capacity limitations & high energy consumption

DEEP-SEA

IO-SEA

Workload diversity

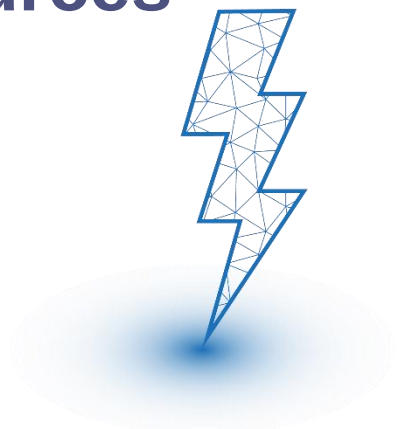
- Exascale centers must run a wide variety of HPC, AI and data analytics workloads with highest energy efficiency

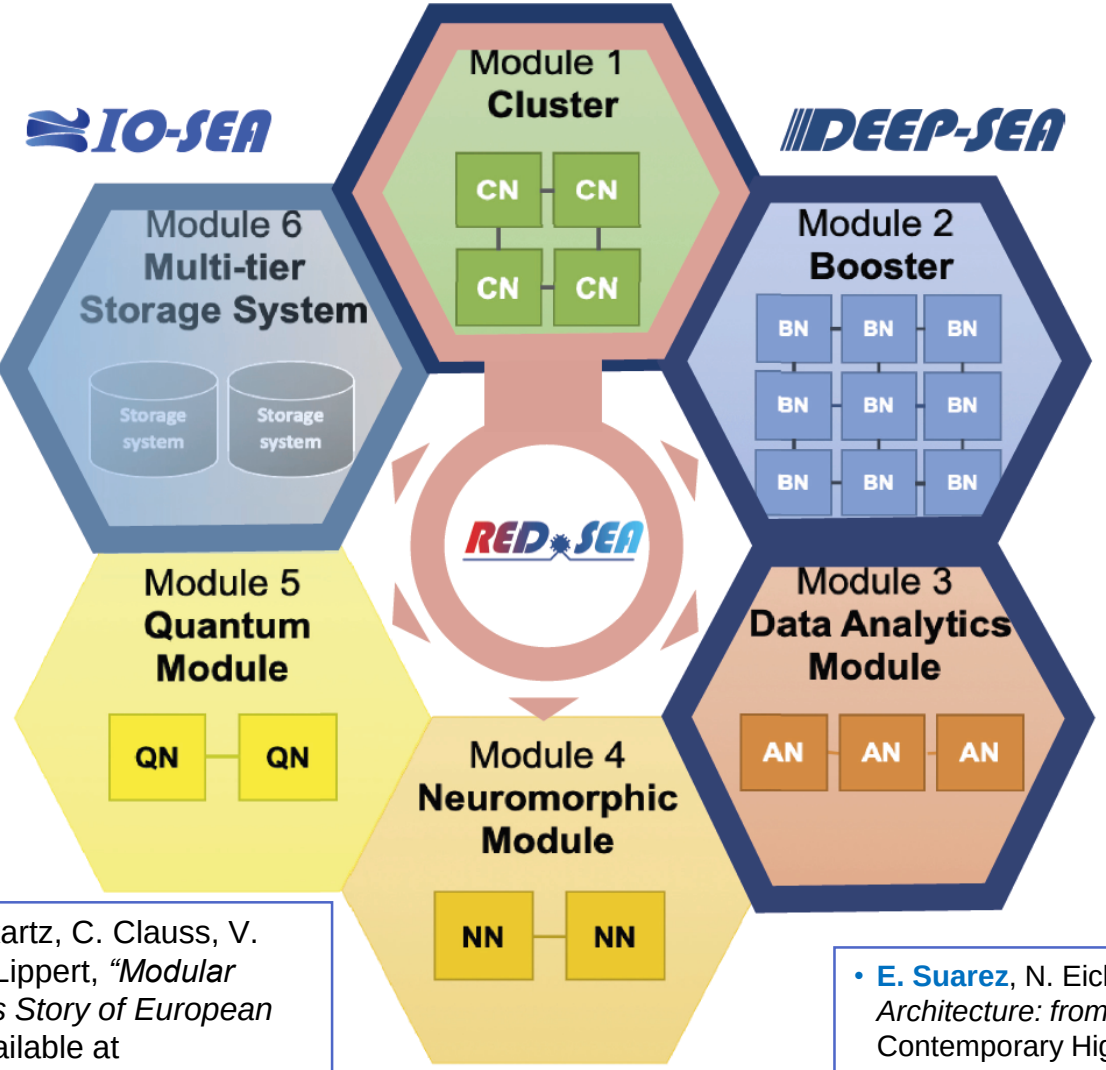
DEEP-SEA

Modular Supercomputing Architecture

SEA Projects

Composable heterogeneous resources





DEEP-SEA

Software stack and programming model for Exascale heterogeneity

IO-SEA

I/O Software stack for Exascale

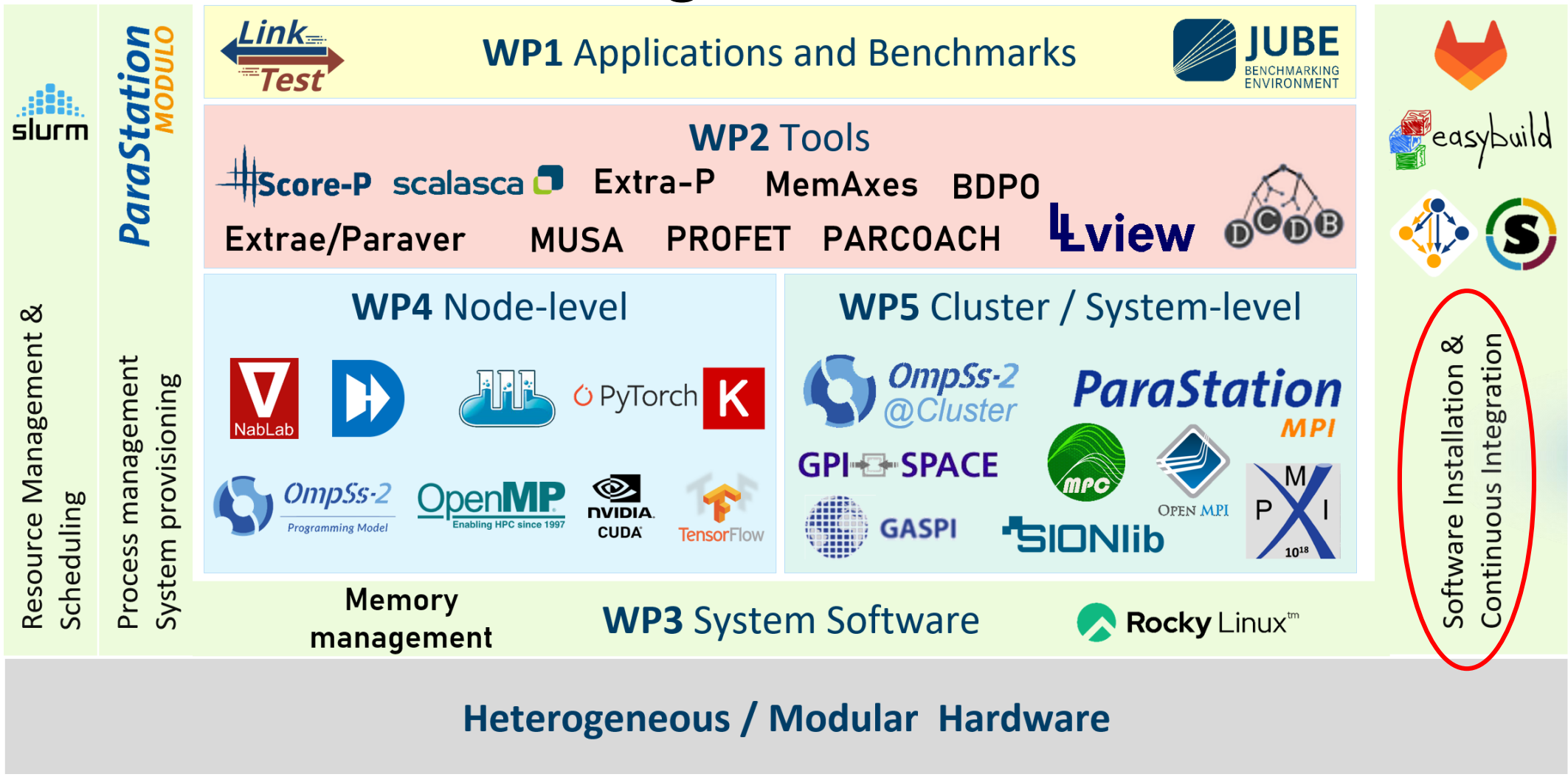
RED-SEA

Network solutions for Exascale systems

- **E. Suarez,** N. Eicker, T. Moschny, S. Pickartz, C. Clauss, V. Plugaru, A. Herten, Kristel Michielsen, T. Lippert, *"Modular Supercomputing Architecture – A Success Story of European R&D",* ETP4HPC White Paper. (2022) Available at https://www.etp4hpc.eu/white-papers.html#msa.

- **E. Suarez**, N. Eicker, Th. Lippert, "*Modular Supercomputing Architecture: from idea to production*", Chapter 9 in Contemporary High Performance Computing: from Petascale toward Exascale, Volume 3, p 223-251, CRC Press. (2019)

DEEP-SEA Integrated HPC SW Stack



At the heart of the JUPITER system

Public release at https://gitlab.jsc.fz-juelich.de/deep-sea/wp3/software/easybuild-repository-deep-sea

DEEP-SEA Memory Tools

- How much, if any, do the applications need to be modified?
- Which layer manages the memory? When?
- How much can the applications benefit?

Modifying the application

Standard interfaces
OpenMP 5.0, SICM
OpenMP

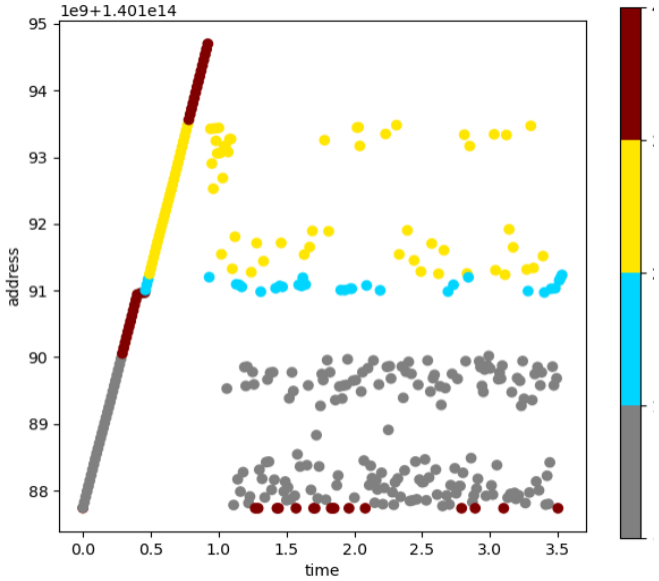
Transparent to the application

Static
(compile time)

ecoHMEM
object level mapping by source-to-source compiler based on profile run

Dynamic
(run time)

SHAMBLES
Run time page-level migration by Operating System



SHAMBLES scatter plot example for sparse kernel

DEEP-SEA Malleability

Usual HPC workload resource reservation (constant # cores or nodes over time)

Actual use of resources varies over time (yellow curve)

Workload is able to use more resources in certain phases (arrow)

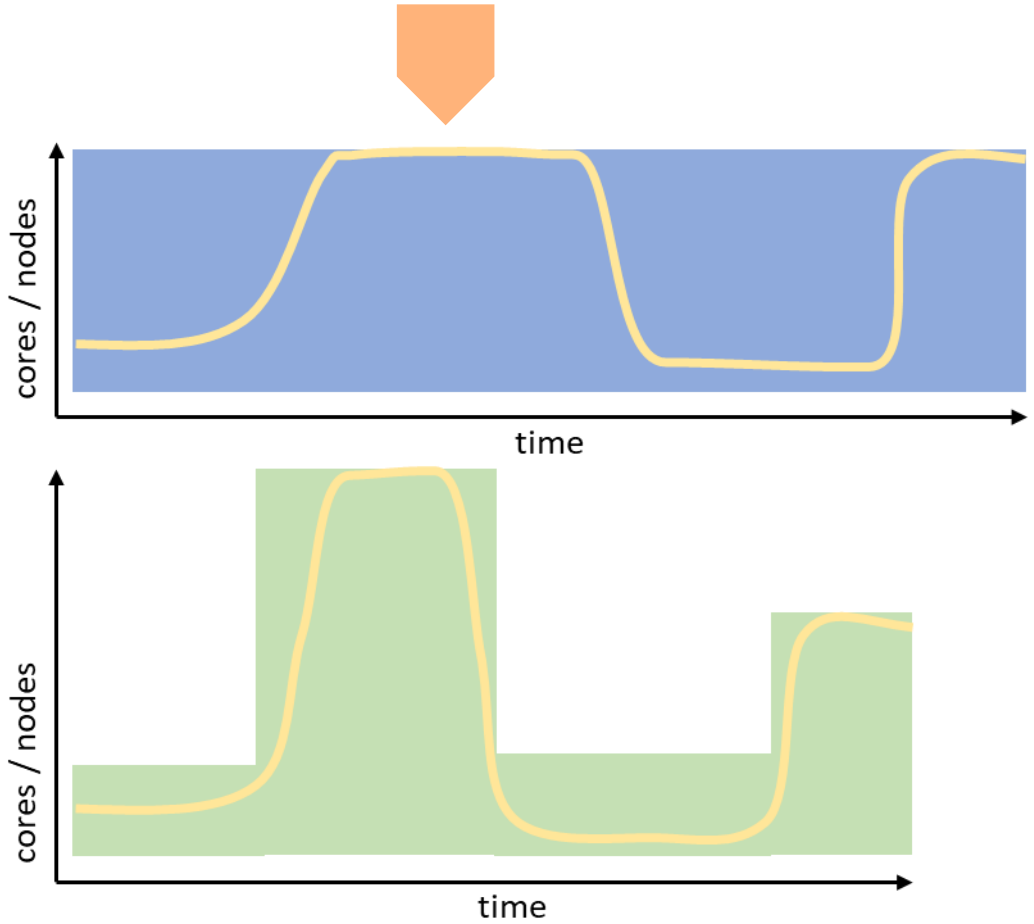
Ideal resource allocation for the workload in green

Malleable applications

- Release resources not required
- Acquire more resources if advantageous

Change in # of nodes do require data redistribution in the workload

DEEP-SEA provides MPI & Slurm prototypes for enabling application-driven (active) malleability



IO-SEA I/O Architecture

Data Access and Storage Interface (DASI) Layer (Language) abstracts the complex storage layer

- Uses semantic description of data – speaking the scientific domain language

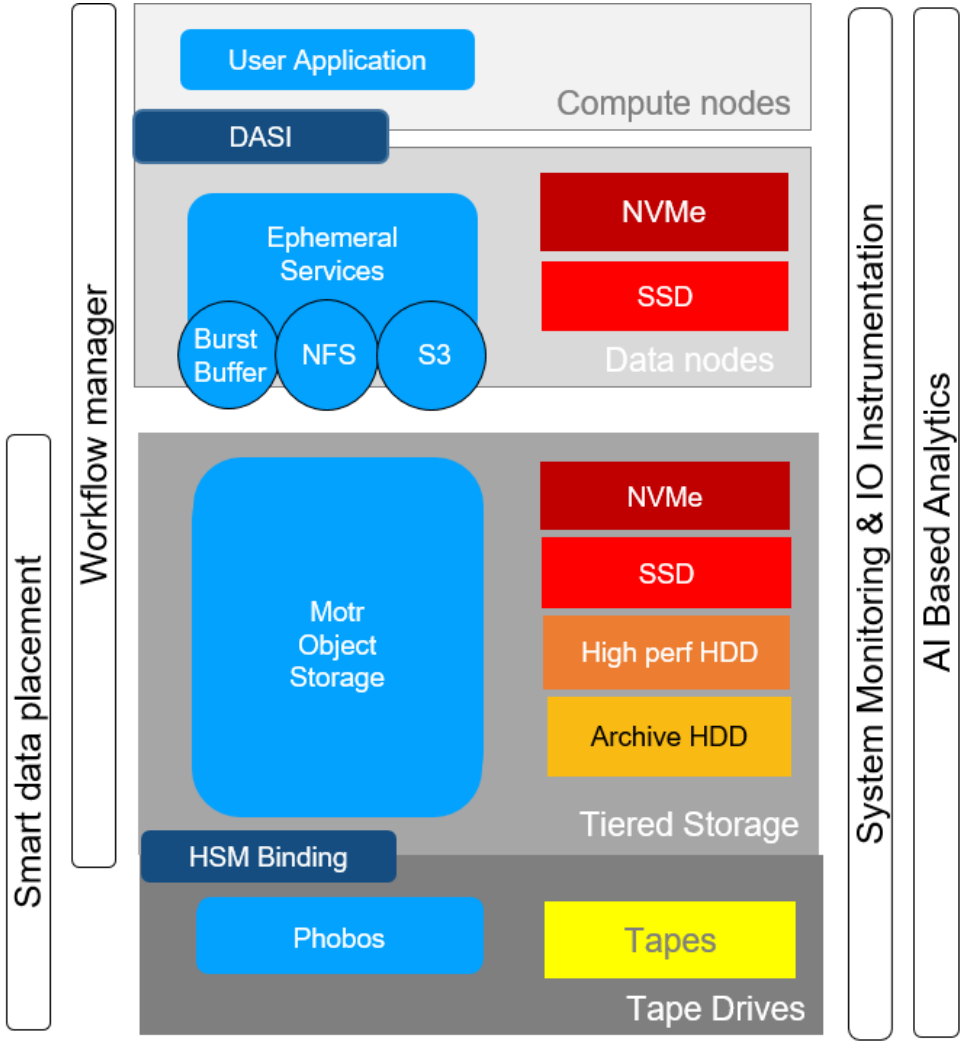
Specialised data access environment for applications and workflows

- Lowers pressure on backend storage system

Leverages NVRAM/NVMe resources available on data nodes

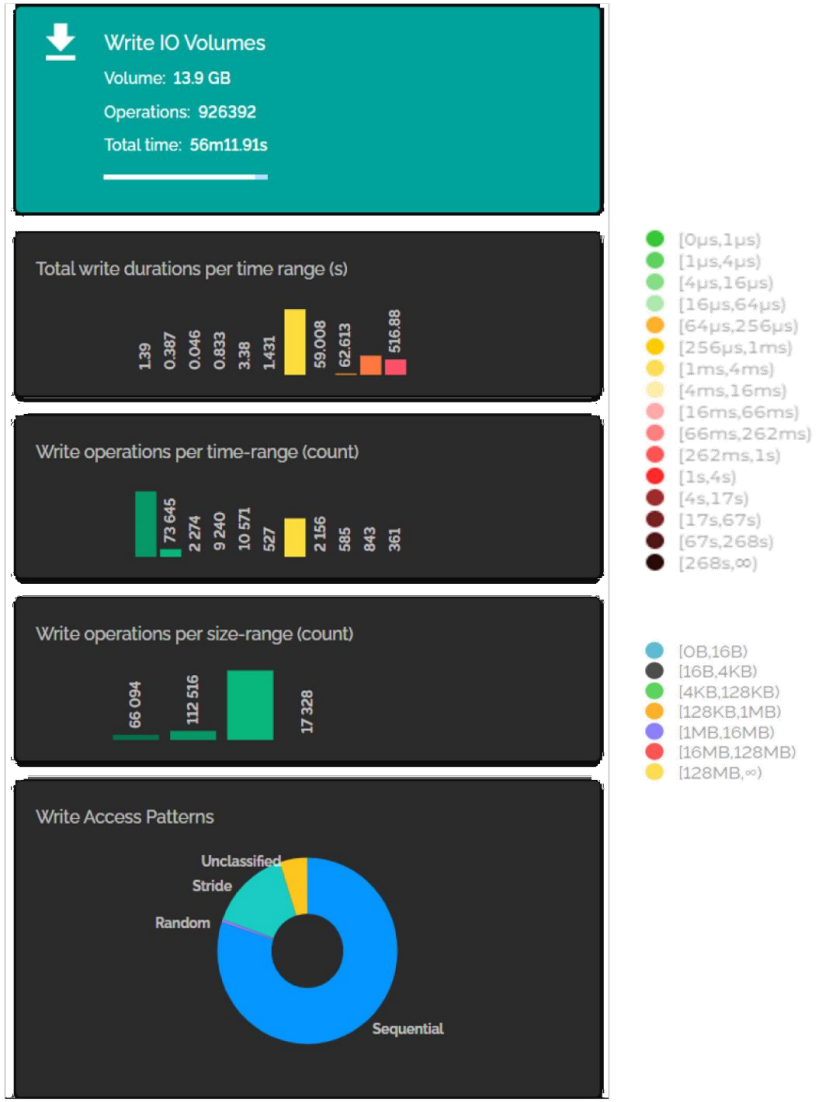
Schedules data accesses on demand through Ephemeral Services (linked to policies)

Includes HSM support



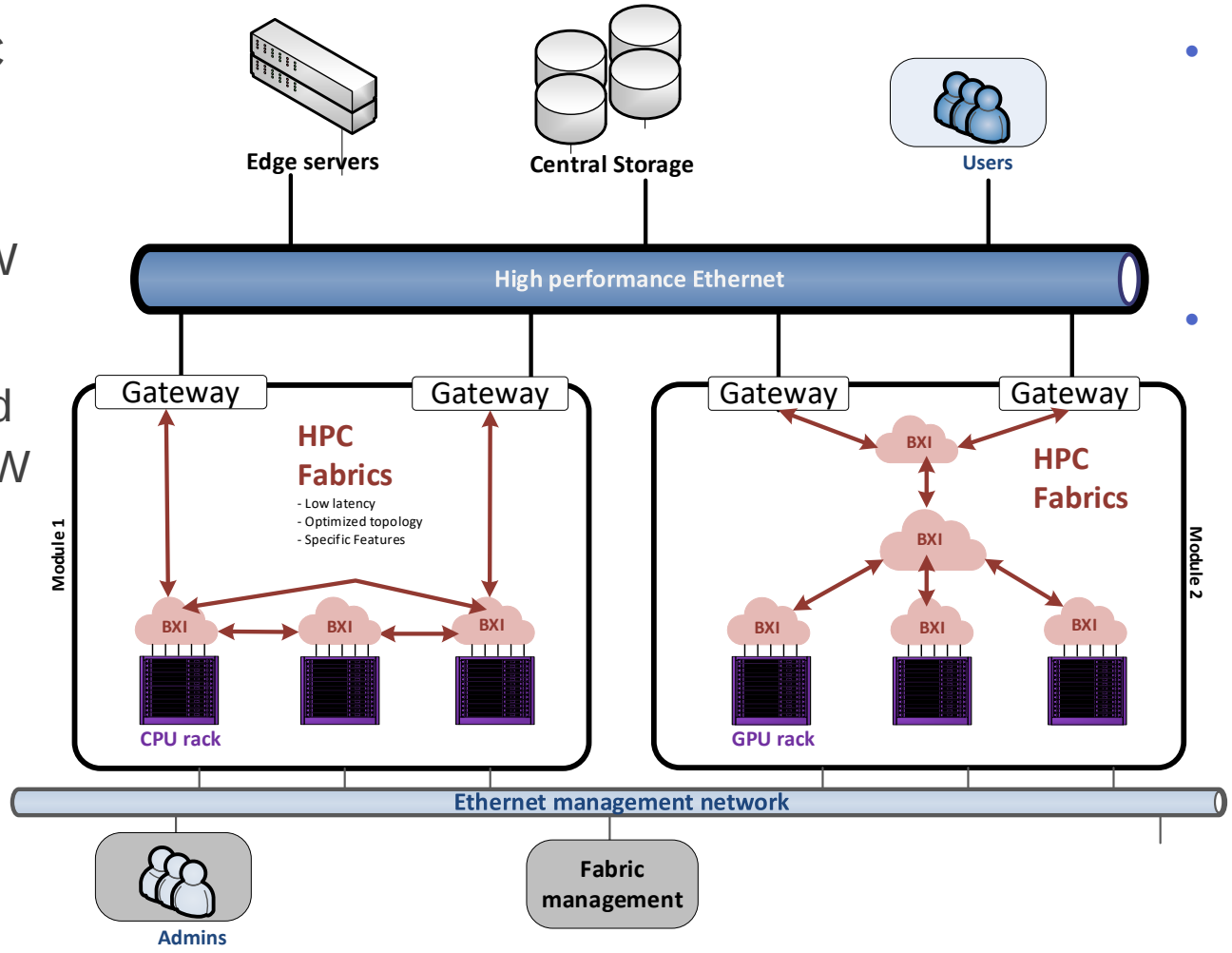
IO-SEA IOI I/O Instrumentation

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



RED-SEA Interconnect Networks

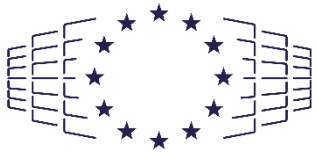
- Atos/Eviden BXI as HPC fabric implementing the Portals 4 interface
- Collective operations HW offload
- Full MSA-aware MPI and network management SW stack
- Load balancing & QoS



- High performance Ethernet as federation network with low latency RDMA communication
- Congestion detection and mitigation

EuroHPC Joint Undertaking

SEA Projects



EuroHPC Joint Undertaking

fwo SPONSORED BY THE

MINISTRY OF EDUCATION, YOUTH AND SPORTS

bpifrance

Federal Ministry of Education and Research

ΓΓΕΚ ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ

MINISTERIO DE CIENCIA E INNOVACIÓN

Financiado por la Unión Europea NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

AGENCIA ESTATAL DE INVESTIGACIÓN

Swedish Research Council

euresearch Swiss guide to European research and innovation

The SEA projects have received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreements n° 955606, 95811, and, 955776 and support from France, the Czech Republic, Germany, Spain, Ireland, Sweden, Switzerland, Italy and Greece.

Thank you

https://sea-projects.eu/
https://deep-projects.eu/
https://iosea-project.eu/
https://redsea-project.eu/