EUROHPC SUMMIT 2024

ANTWERP

UNLEASHING THE POWER OF EUROPEAN HPC AND QUANTUM COMPUTING

Co-Designing the Future of European HPC: Eco-friendly Technologies, Systems and Software

AGENDA

19.03.2024

ADMIRE malleable data solutions for HPC

DEEP-SEA

RED-SEA

textarossa

EUPEX European Pilot for Exascale

IO-SEA

eProcessor

TIME-X

THE EUPILOT

EuroHPC Summit 2024, Antwerp, 18 – 20, March 2024

Agenda & Speakers

Topic Title	Contributing Projects Involved	Speakers	Time
Introduction	N/A	Rene Chatwell [EuroHPC]	3 minutes
Co-Design in Europe: A Summary	All	Hans-Christian [ParTec/FZJ]	7 Minutes
Technology Area: European Processor Ecosystem	eProcessor, EUPEX The European PILOT	Carlos Puchol [BSC]	10 Minutes
Technology Area: Dynamic Resource Management	ADMIRE DEEP-SEA IO-SEA Regale Time-X	Jesus Carretero [UC3M]	10 Minutes
Technology Area: European System Architecture Advancements	DEEP-SEA EUPEX IO-SEA RED-SEA TEXTAROSSA	Hans-Christian [ParTec/FZJ]	10 Minutes
Technology Area: I/O, Storage and Data Management	ADMIRE IO-SEA	Maike Gilliot [CEA]	10 Minutes
Panel Discussion [38 Minutes] - Speakers will be in the panel, project representatives are also in the audience to answer questions if needed			
Conclusions [2 Minutes] - Moderator			

Project Representatives

Project	Representatives
ADMIRE	Jesus Carretero [UC3M]
DEEP-SEA	Hans-Christian Hoppe [ParTec/FZJ]
eProcessor	Carlos Puchol [BSC]
EUPEX	Jean-Robert Bacou [Eviden]
IO-SEA	Sai Narasimhamurthy [ParTec]
RED-SEA	Pascale Bernier-Bruna [Eviden]
Regale	Andry Razafinjatovo [RYAX]
TEXTAROSSA	Daniele Gregori [E4]
The European PILOT	Carlos Puchol [BSC]
Time-X	Jesus Carretero [UC3M]

EUROHPC SUMMIT 2024

ANTWERP

UNLEASHING THE POWER OF EUROPEAN HPC AND QUANTUM COMPUTING

Co-Design in Europe: A Summary

Hans-Christian Hoppe, ParTec AG & Forschungszentrum Jülich

ADMIRE
malleable data solutions for HPC

DEEP-SEA

RED SEA

textarossa

EUPEX
European Pilot for Exascale

IO-SEA

eProcessor

TIME-X

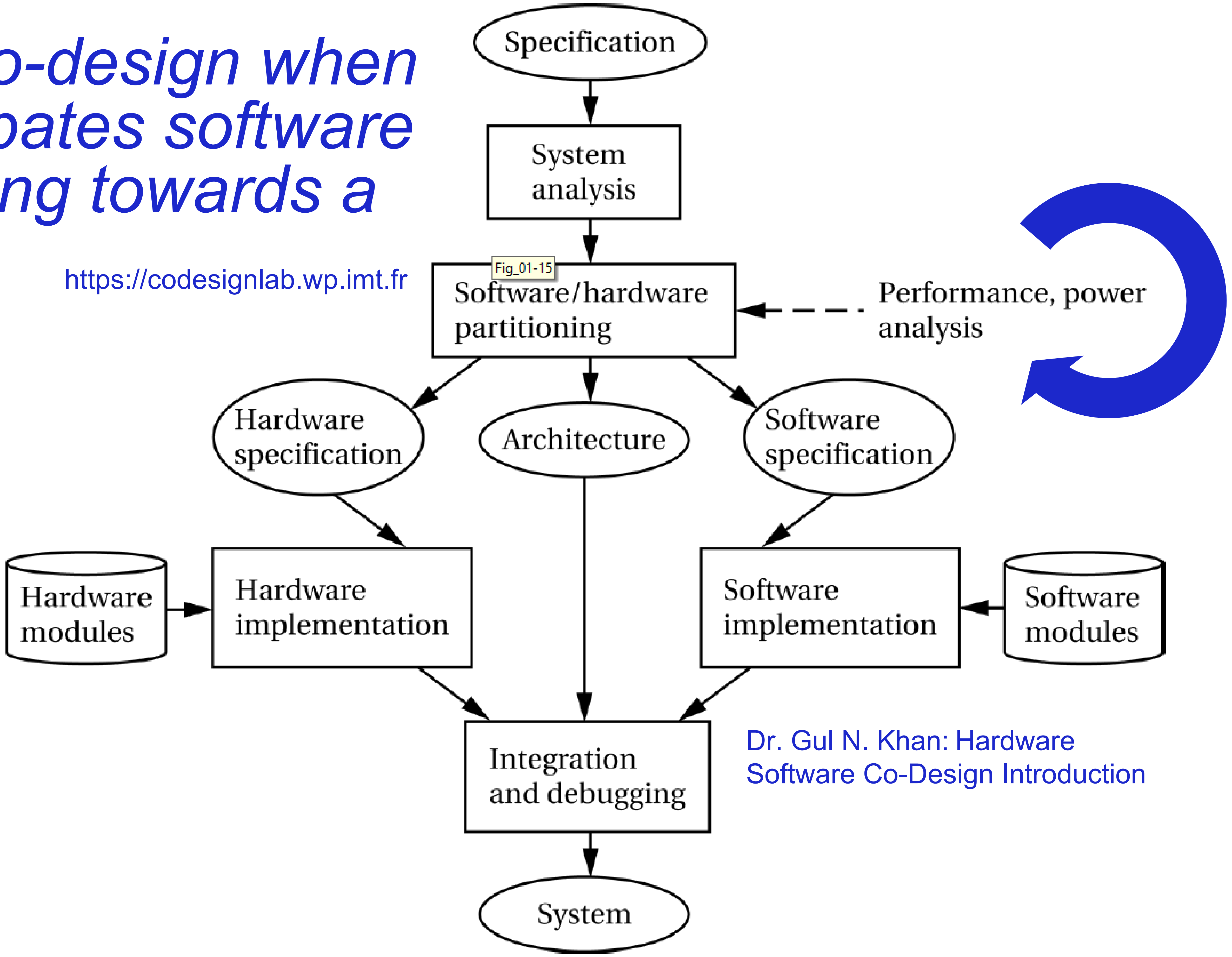
THE EUPILOT

What is Co-Design?

“Electronic and computer scientists speak of co-design when they want to describe the way hardware anticipates software and software adapts to hardware, both evolving towards a better integration”

https://codesignlab.wp.imt.fr

- Common practice in embedded computing
  - Producing an *integrated* HW/SW system with prescribed behavior from known HW & SW design kits and optimize cost, time to market, performance, form factor, energy, …
  - Complete HW/SW system is specified in SW and can be simulated



Dr. Gul N. Khan: Hardware Software Co-Design Introduction

Co-Design & High Performance Computing

HPC HW *and* SW are both highly complex and not amenable to be quickly changes/refactored

Co-Design is a sequential flow

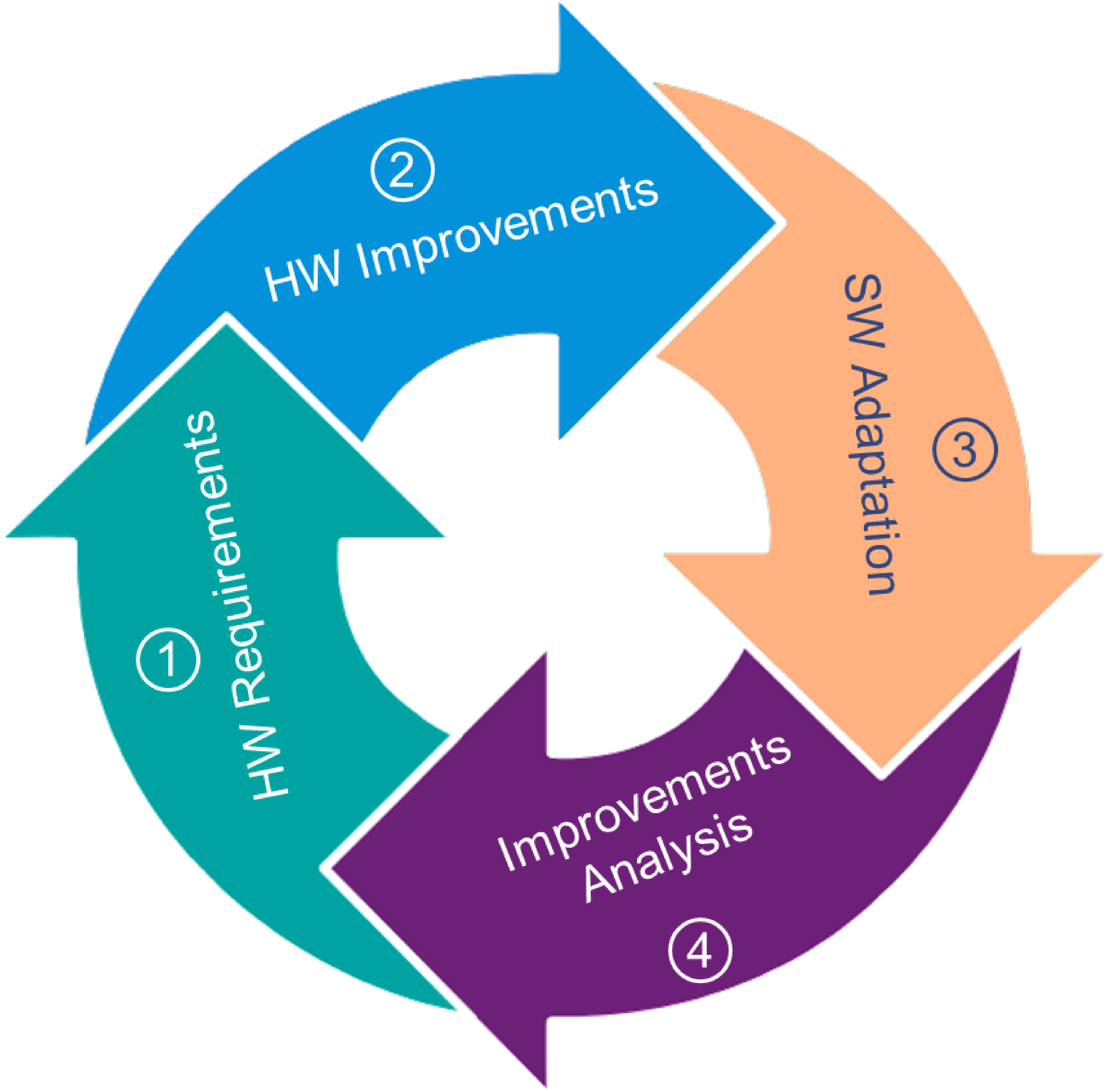
- Pipelining could increase throughput
- Careful definition of APIs to overlap steps 2 & 3

Problematic steps

- Step 2 – HW takes (lots of) time & (lots of) money
- Simulators/emulators can accelerate steps 2 & 3
- Step 4 – Adapting representative workloads takes time

How many iterations

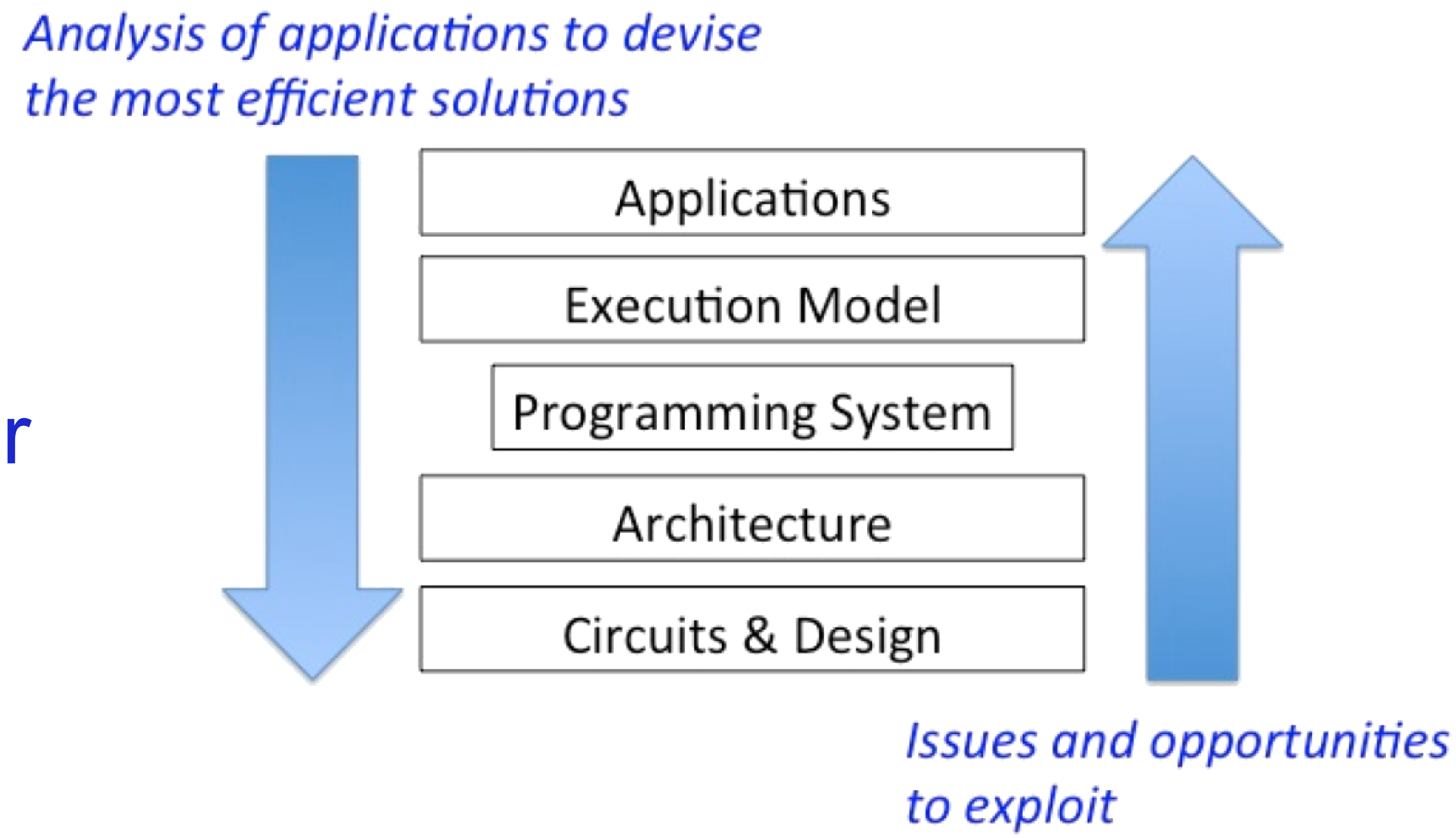
- Do we need
- Can we afford?



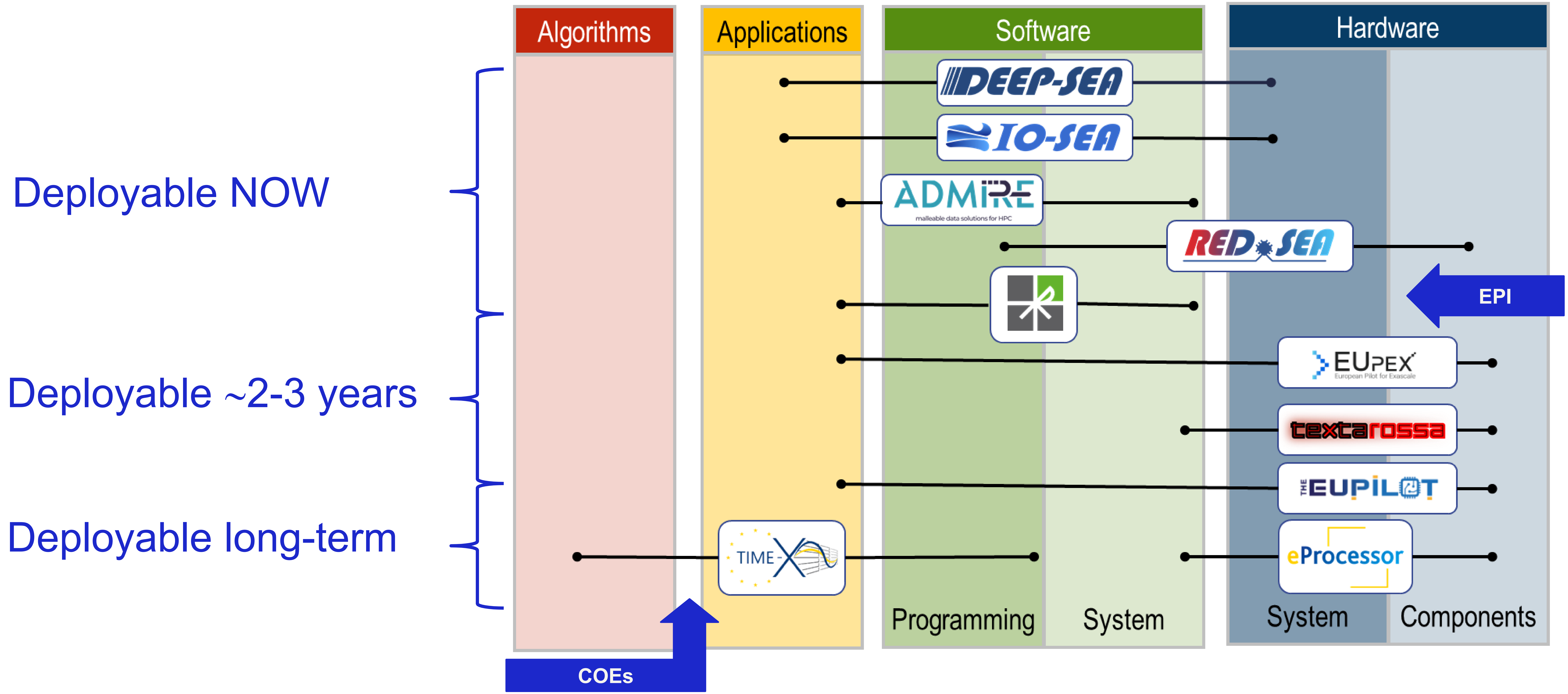
There's More than just HW and SW

HPC co-design involves many layers

- HW/applications co-design
- I/O co-design: storage systems and I/O middleware or applications
- HW/HW co-design: multiple HW layers/devices, like node & accelerators
- SW/SW co-design: multiple SW layers, like compiler/application co-design
- Co-design involving (actual or modelled) workload mixes, such as Scheduler/workload co-design


Analysis of applications to devise the most efficient solutions
Applications
Execution Model
Programming System
Architecture
Circuits & Design
Issues and opportunities to exploit

EuroHPC JU Projects Cover the Spectrum



Co-Design in HPC – European Experience

Some notable successes

- NextGenIO, SAGE-1 & 2, ADMIRE, IO-SEA: advanced storage architectures & APIs, proof points for novel memory/storage technology
- DEEP, DEEP-ER, DEEP-EST: validated flexible, modular architecture for heterogeneous systems
- DEEP-SEA, IO-SEA, RED-SEA, EUPEX: integrate and evolve common European HPC SW stack
- REGALE: prototype architecture for integrating monitoring & resource management

Challenges encountered

- Timescale for co-design iterations longer than typical project terms
- Reliance on externally developed technology which failed in the market
- Costs and timescale for developing ready-to-deploy hardware
- Convincing end-users, application developers and HPC centre operators to take up results

HPC and AI Co-Design – Looking Ahead

Coordination between funding programs and collaboration between projects

- Exploration of ideas and follow-through with “best” approach
- Application CoEs and technology projects must closely interact

Reconciliation of co-design timeframes and project terms

- Chains of projects vs. long-term funding lines

Open hardware presents a great opportunity

- Steep rise in costs for silicon design, manufacturing and validation is challenging

Balance between HPC and AI technology requirements to be found

- Example: replacing 64-bit floating point by smaller data types?

Need to anticipate future use cases and market drivers

- Critical to sustain further technology development

SPONSORED BY THE

Federal Ministry
of Education
and Research

MINISTERIO
DE CIENCIA
E INNOVACIÓN

Financiado por
la Unión Europea
NextGenerationEU

Plan de Recuperación,
Transformación y
Resiliencia

AGENCIA
ESTATAL DE
INVESTIGACIÓN

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ
ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΕΠΕΝΔΥΣΕΩΝ

ΓΓΕΚ
ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ
ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ

MŠMT
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

bpifrance

REPUBLIC OF CROATIA
Ministry of Science and
Education

fwo

Ministero dello
sviluppo economico

FCT
Fundação para a Ciência e a Tecnologia

euresearch
Swiss guide to European
research and innovation

The Research Council
of Norway

Swedish
Research
Council

EuroHPC
Joint Undertaking

These projects have received funding from the EuroHPC JU, under Grant Agreements **955606, 955701, 956748,955776, 955811, 956560, 956702, 956831,101033975, 101034126**.

The EuroHPC Joint Undertaking (JU) receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, Croatia, Cyprus, Czech Republic, France, Germany, Greece, Ireland, Italy, Luxembourg, Norway, Poland, Portugal, Spain, Sweden, Switzerland, and Turkey.

THANK YOU

EUROHPC SUMMIT 2024

ANTWERP

UNLEASHING THE POWER OF EUROPEAN HPC AND QUANTUM COMPUTING

European System Architecture Advancements

Hans-Christian Hoppe, ParTec AG & Forschungszentrum Jülich

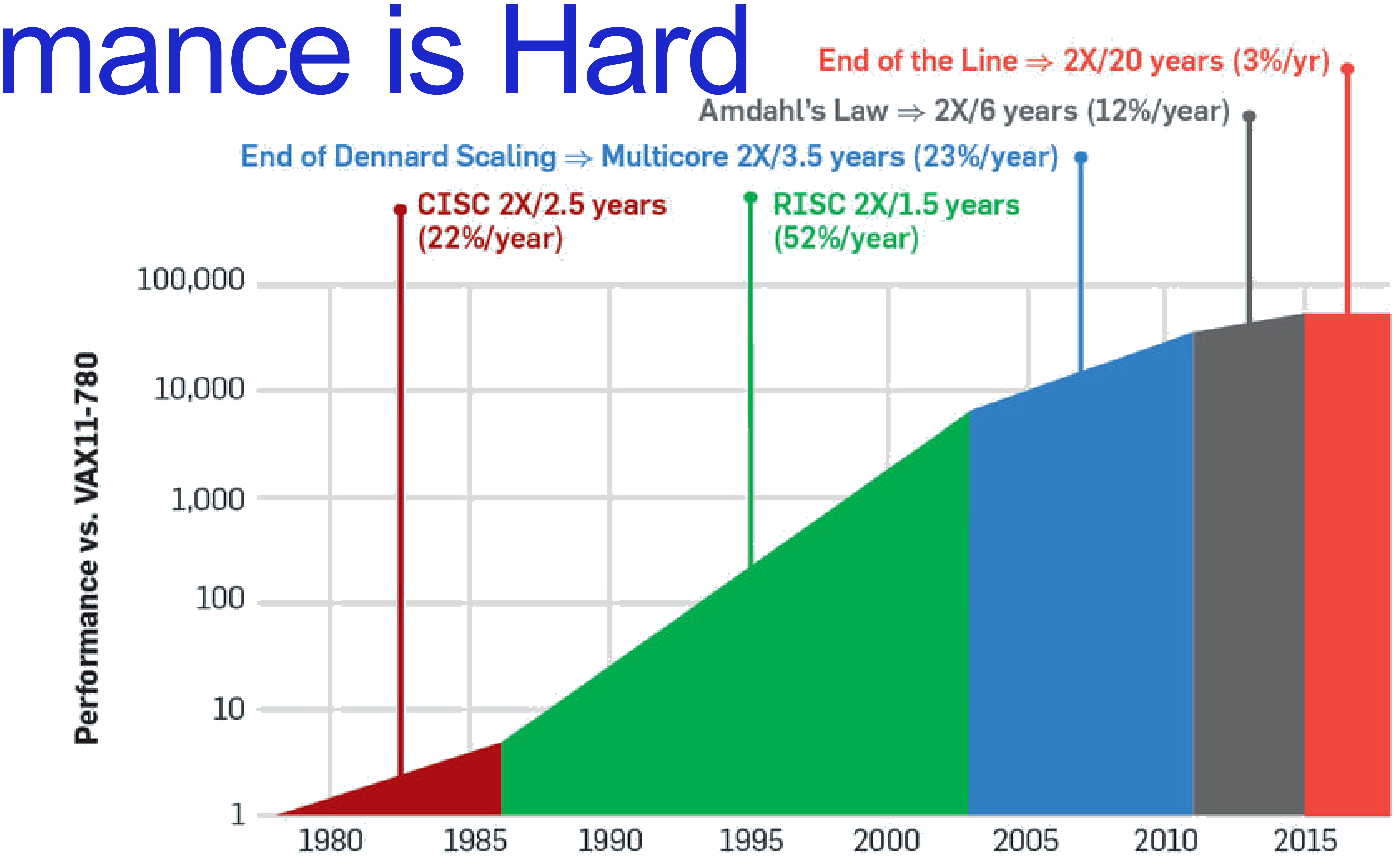
DEEP-SEA EUPEX European Pilot for Exascale IO-SEA RED-SEA textarossa

Reaching Exascale Levels of Performance is Hard

”Free lunch” no longer served by Moore’s Law

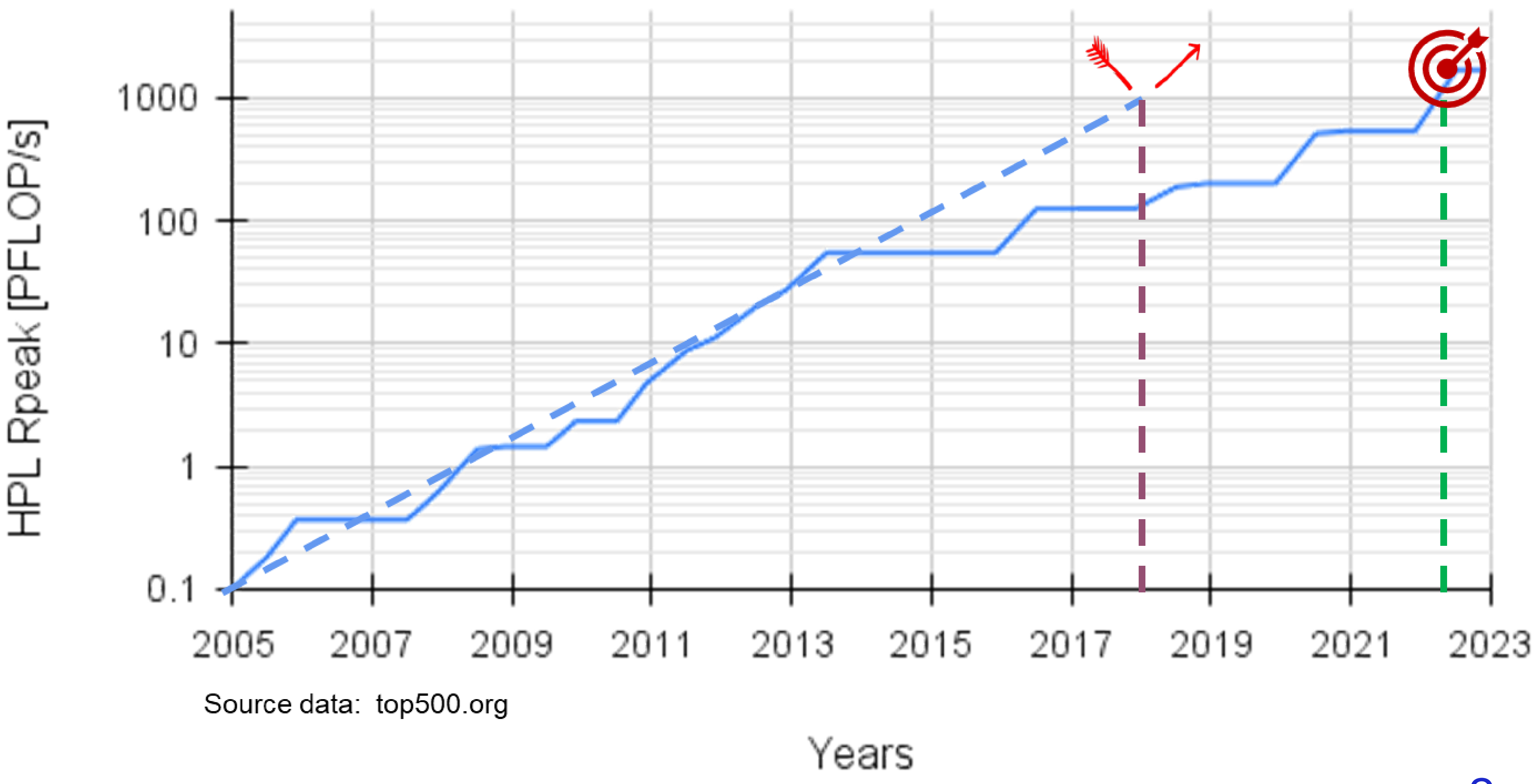
Challenges include

- Highest application parallelism
  - Algorithms and portable parallel programming
- Truly scalable systems
  - (Much) faster interconnect fabrics
- Highest energy efficiency
  - Accelerators and heterogeneous systems
  - Efficient cooling
- Memory and Storage
  - Close performance gap to compute engines
- Diversity of applications requirements
  - Deployed systems must support a huge set of applications with very different system requirements



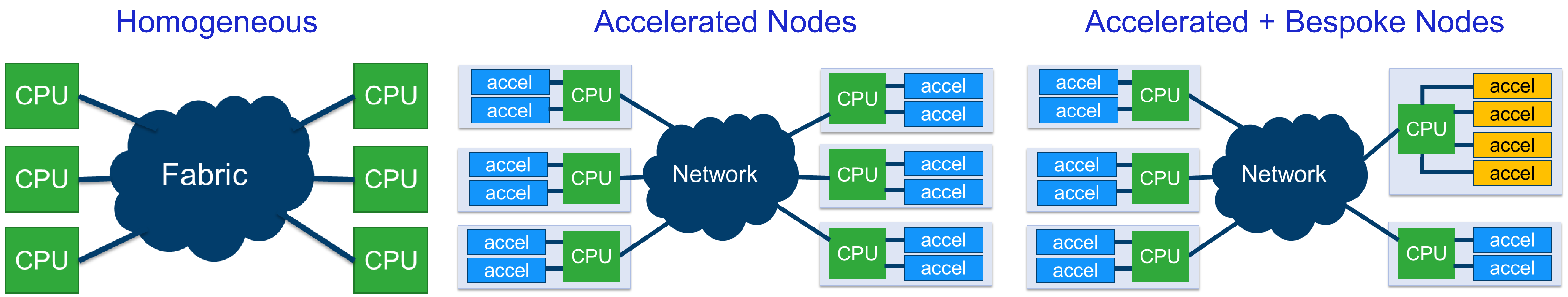
https://www.researchgate.net/publication/343096513_Energy_Efficient_Computing_Systems_Architectures_Abstractions_and_Modeling_to_Techniques_and_Standards

Top #1: HPL Rpeak [PFLOP/s]

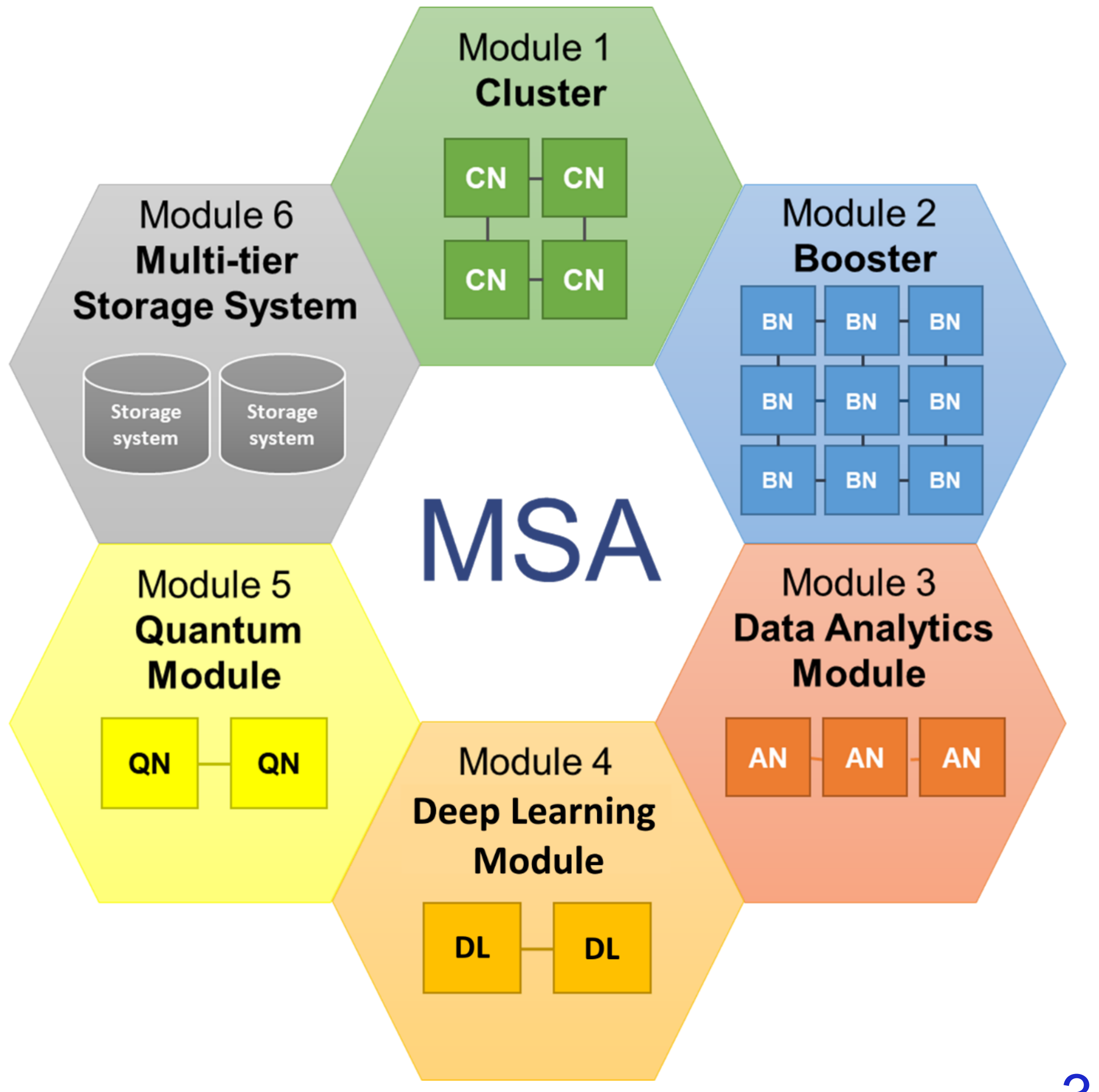


Source data: top500.org

Making Heterogenous Systems Flexible



- Accelerated nodes fix ratio of CPUs vs. accelerators, complicate sharing resources across nodes
  - Many applications will not use all resources, and capital & energy is wasted
- Really want to aggregate different compute, storage and network resources according to dynamic demand
- Adding „bespoke nodes“ for special tasks does not provide this flexibility



Modular Supercomputing Architecture

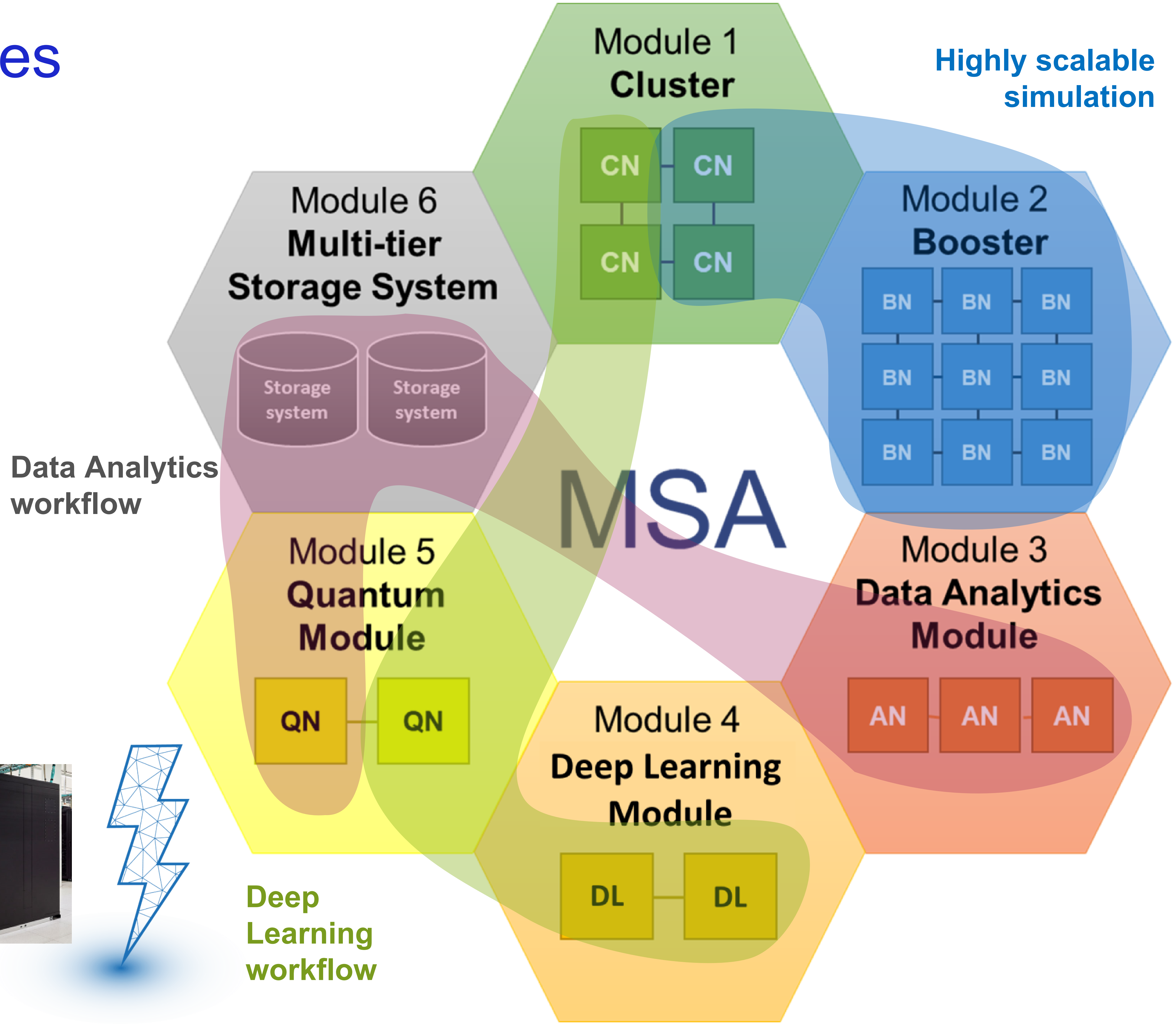
The MSA achieves composability of resources

- Cost-effective scaling
- Effective resource-sharing
- Match workload diversity

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)

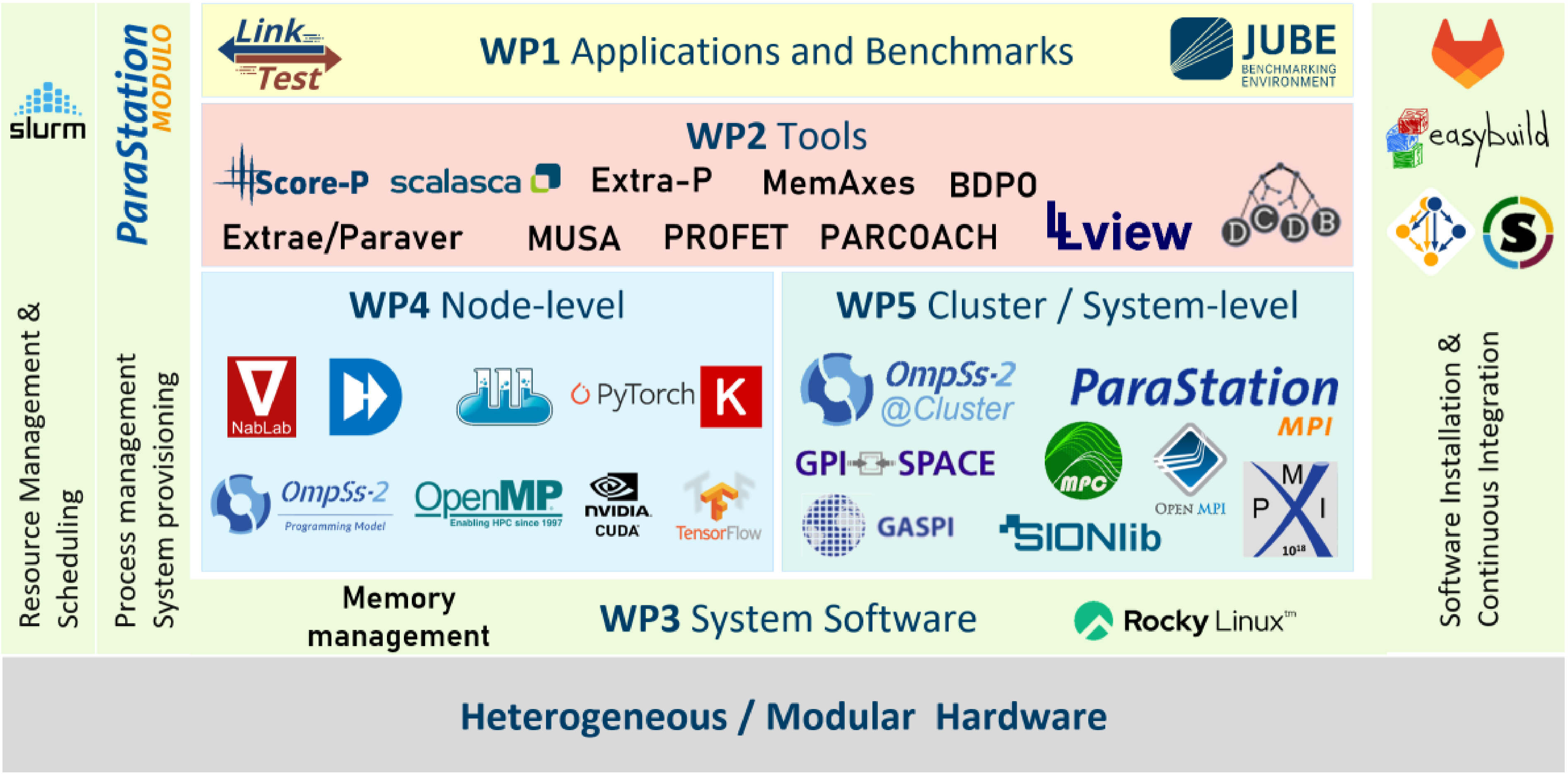
DEEP-SEA EUPEX European Pilot for Exascale IO-SEA RED-SEA

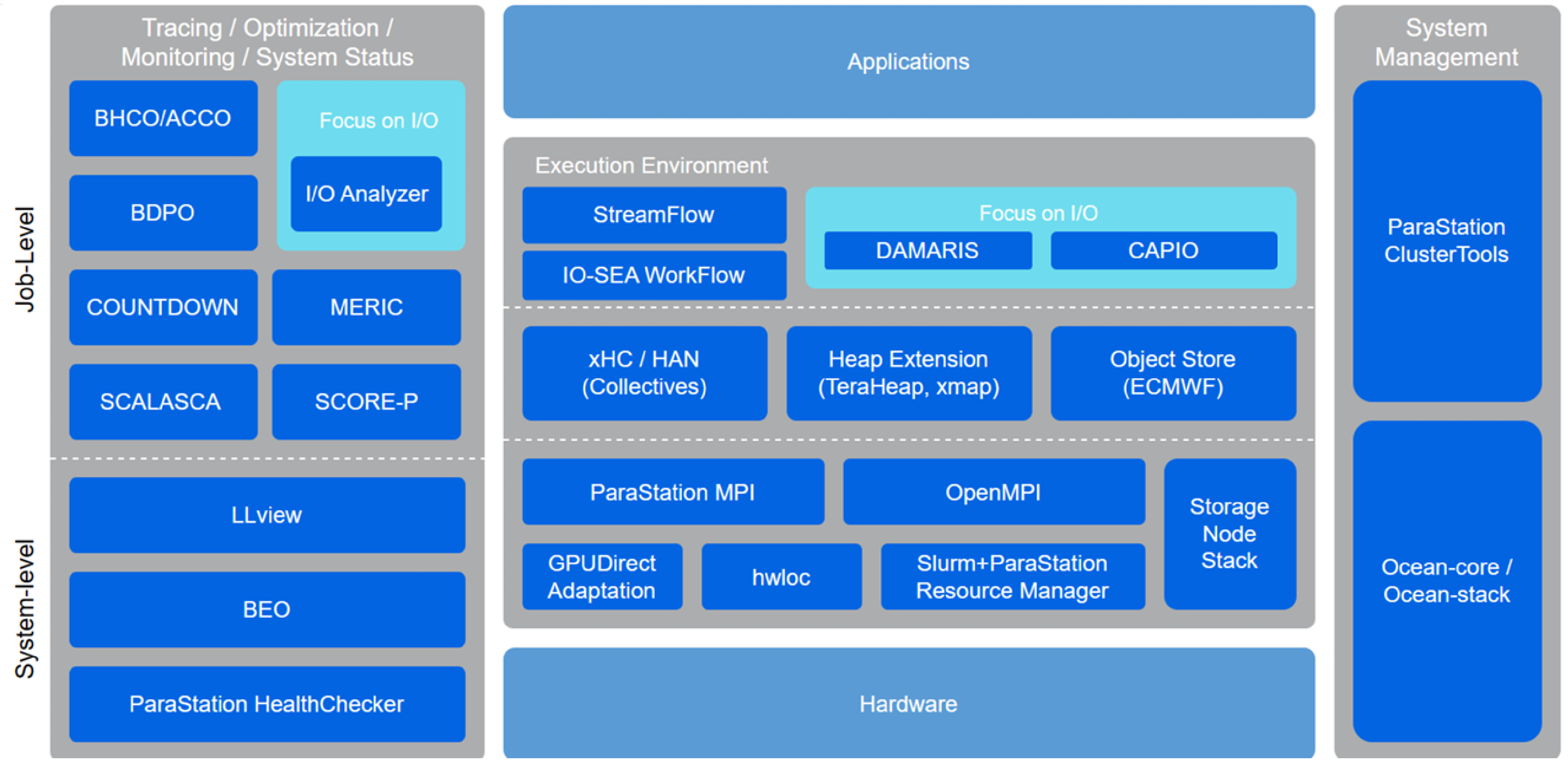


DEEP-SEA & EUPEX SW Stacks

Cross-project collaboration towards a common, integrated European SW stack for HPC and AI

- DEEP-SEA provides a full set of programming models, APIs and tools for programming and operating heterogeneous systems
- EUPEX adds support for workflows and additional I/O components from IO-SEA and ports to the SiPearl Rhea CPU
- RED-SEA contributes optimized communication libraries for BXI
- Most SW components freely available, early access to EUPEX platform



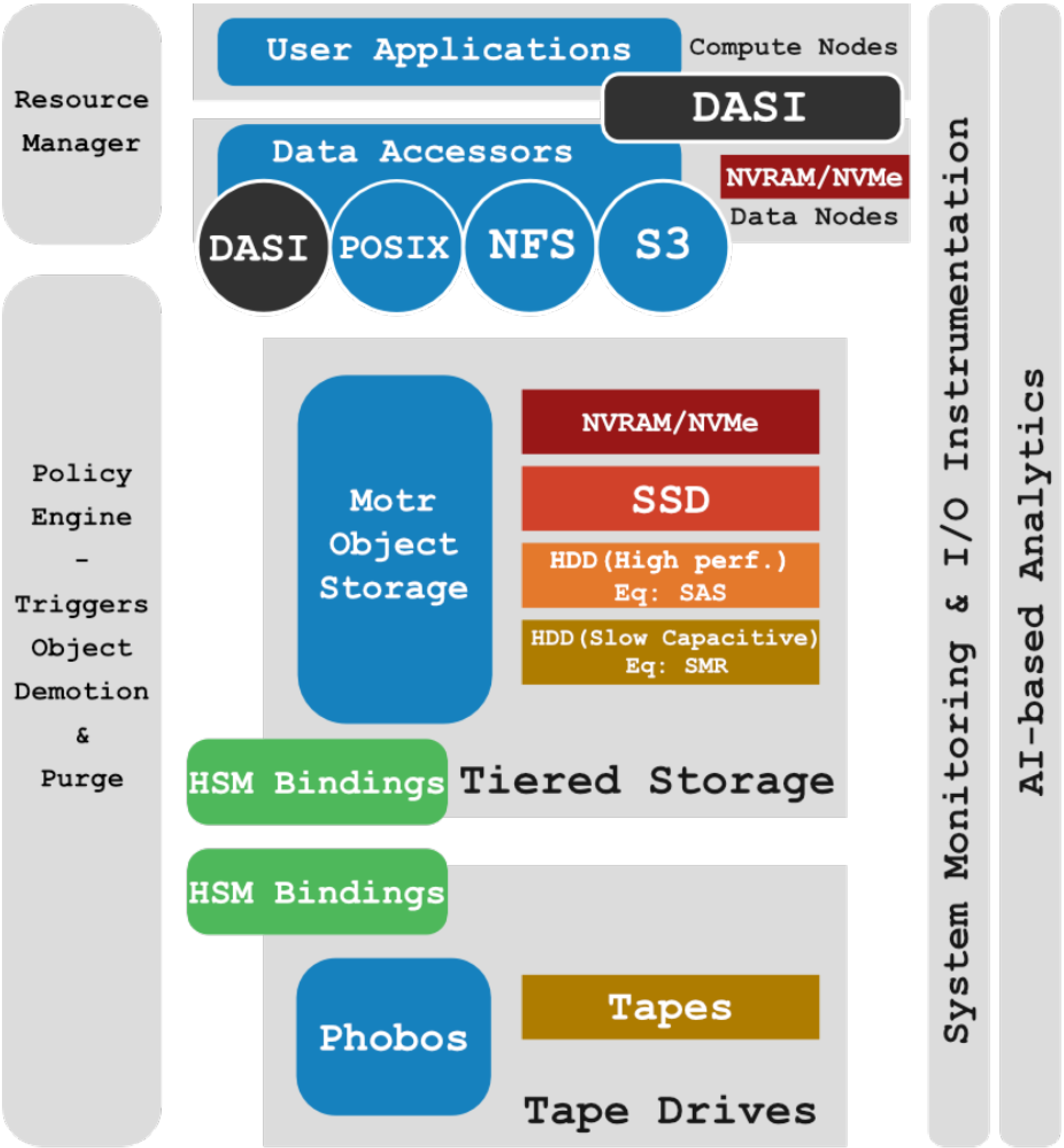


DEEP-SEA EUPEX European Pilot for Exascale IO-SEA RED-SEA

IO-SEA Storage Architecture & Services

IO-SEA adds an integrated storage system to the MSA

- Implemented as storage modules within an MSA system
- Covering the full range from ad-hoc storage on SSDs/NVM to archive storage on tape drives
- DASI semantic interface to scientific data
- Ephemeral storage services support applications on demand
- Fully integrated with MSA resource management, scheduling and operational management



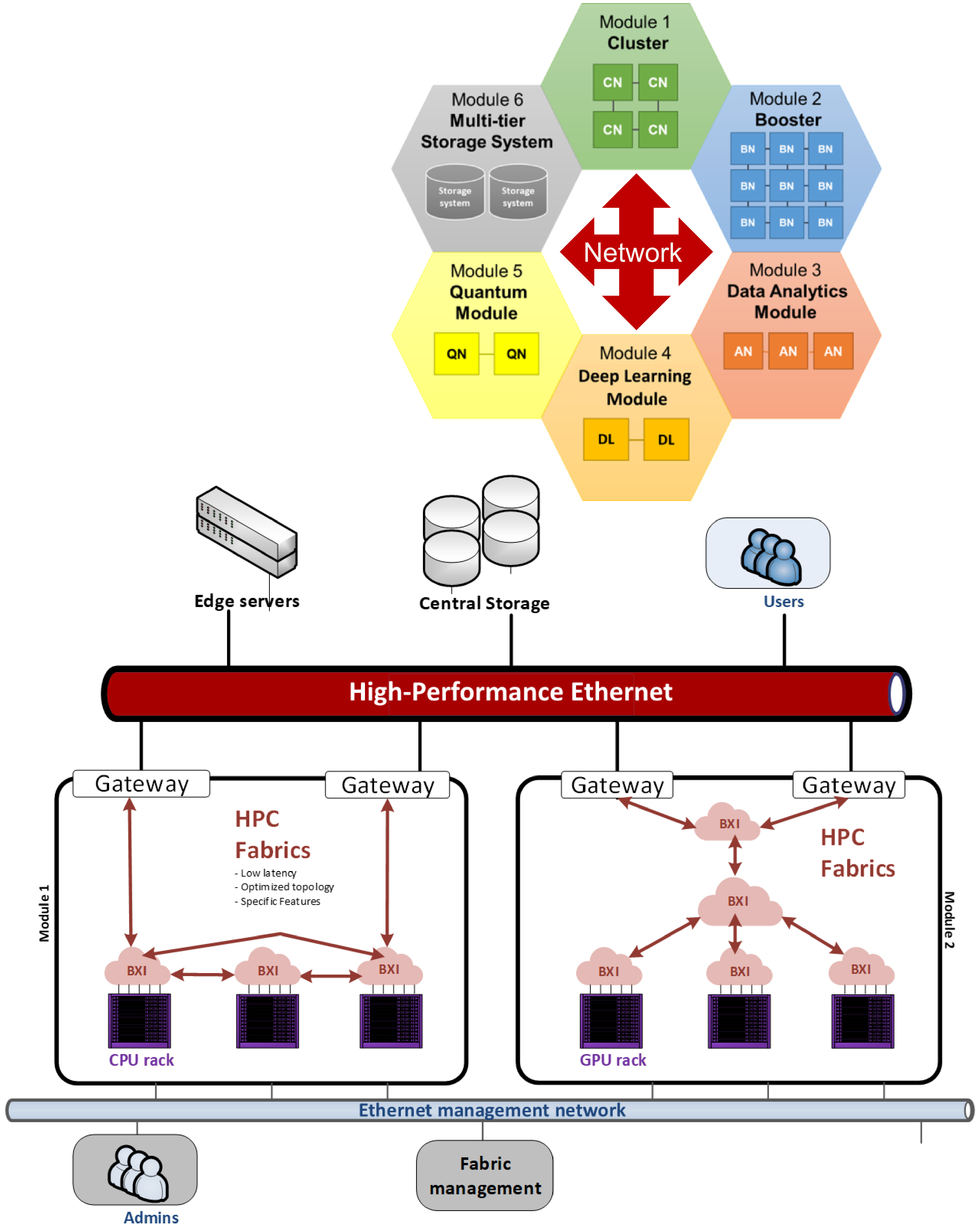
RED-SEA Interconnect

High performance Ethernet as *federation network* with state-of-the-art low latency RDMA communication

Bull eXascale Interconnect (BXI) as *HPC fabric* built from BXI NICs and BXI switches

Zero-copy protocols and on-NIC message matching enable superior performance for message passing, RDMA and partitioned global address spaces

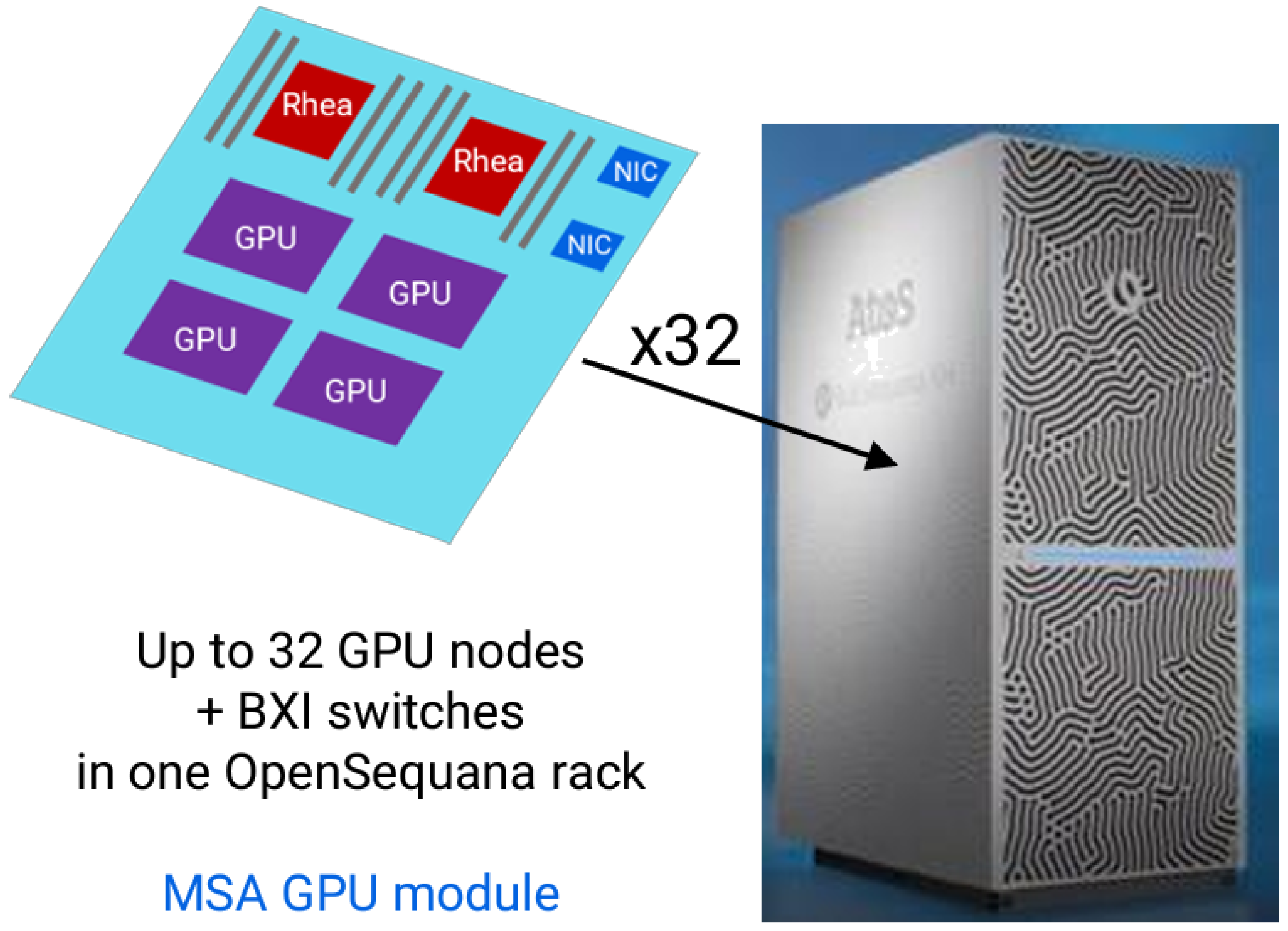
- Fully satisfies requirements of HPC, AI and data analytics



EUPEX & TEXTAROSSA Node Architecture & Cooling

EUPEX

- Compute module node using SiPearl (EPI) Rhea CPUs and NVIDIA GPUs as accelerators
- Eviden BXI V3 (Enhanced by RED-SEA) as module interconnect
- Fully integrated with Eviden BullSequana XH3000 platform

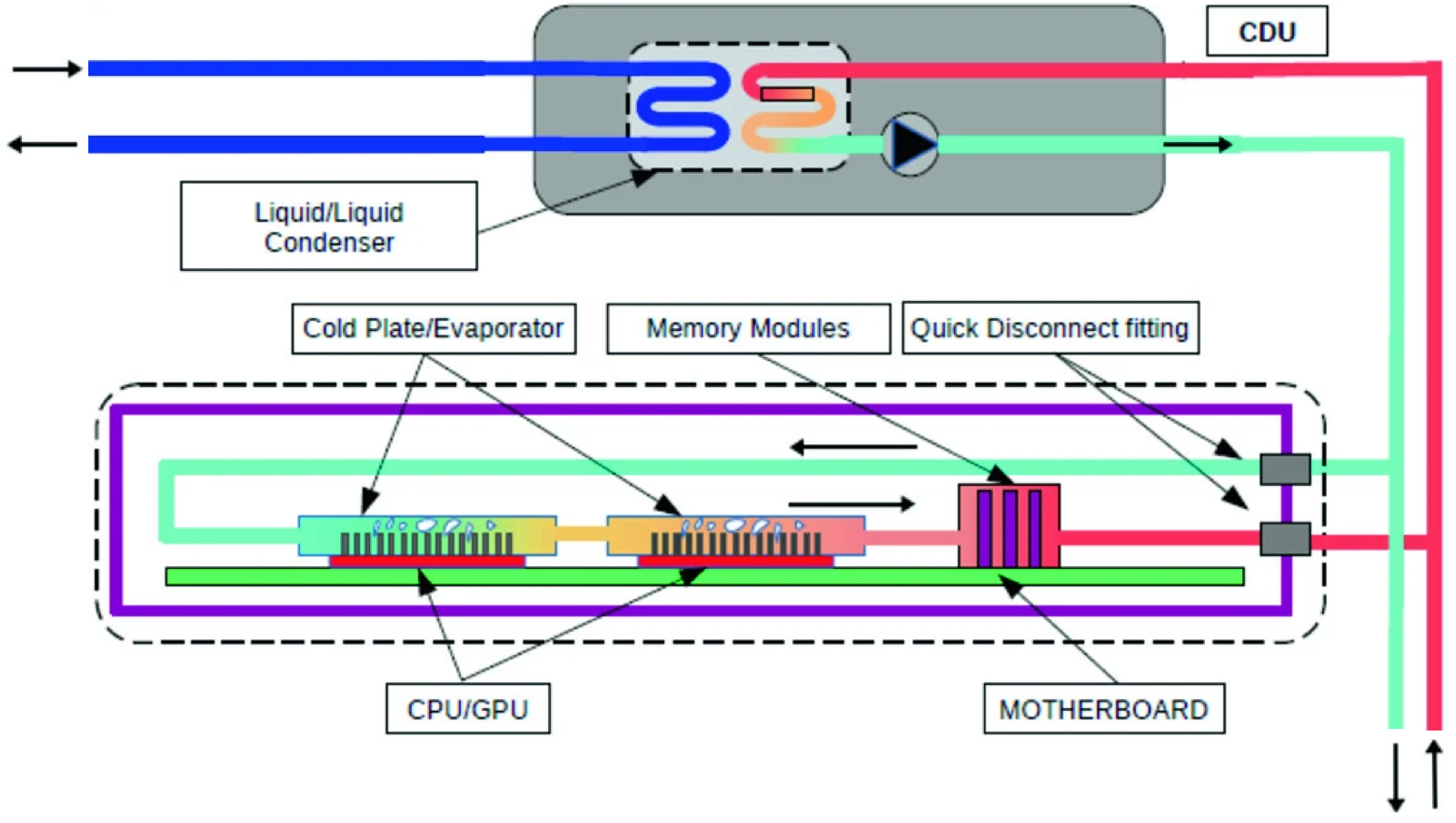


Up to 32 GPU nodes
+ BXI switches
in one OpenSequana rack

MSA GPU module

TEXTAROSSA

- Innovative multi-level thermal management and two-phase cooling
- Covers node and rack level
- Fully integrated with energy optimizing resource management



EUPEX European Pilot for Exascale

textarossa

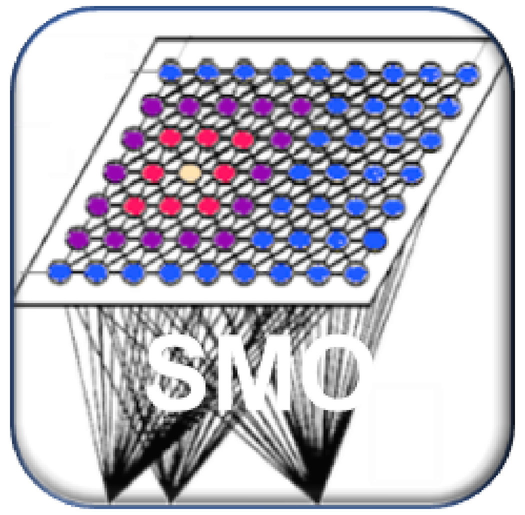
Co-Design Applications

Astrophysics

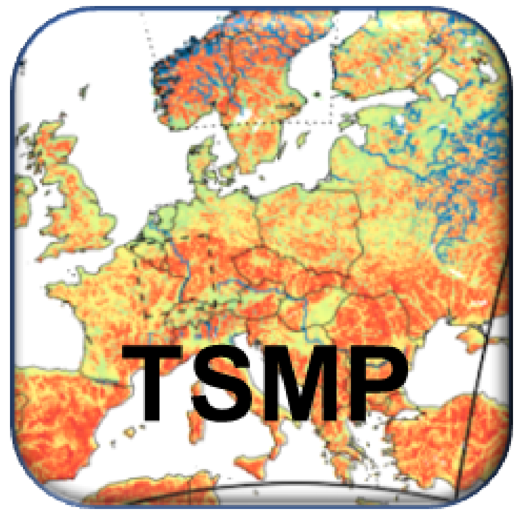
Atomistic Simulations

Brain Simulation

CFD



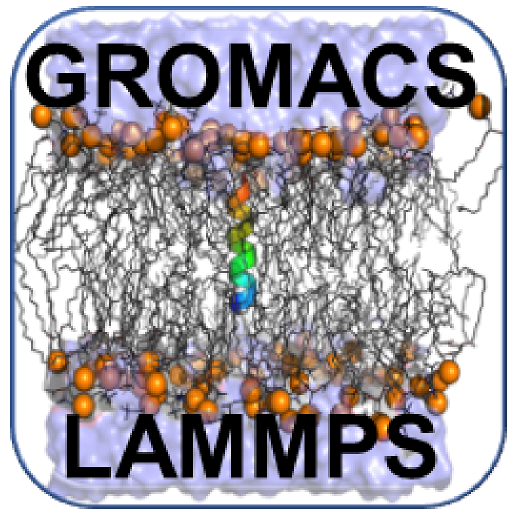
Deep Learning



Earth System Modeling

Drug Design

Finite Elements



Molecular Dynamics

Neutron Transport

Precision Agriculture

QCD

Remote Sensing

Seismic Imaging

Space Weather

Weather Climate

SPONSORED BY THE
Federal Ministry
of Education
and Research

MINISTERIO
DE CIENCIA
E INNOVACIÓN

Financiado por
la Unión Europea
NextGenerationEU

Plan de Recuperación,
Transformación y
Resiliencia

AGENCIA
ESTATAL DE
INVESTIGACIÓN

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ
ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΕΠΕΝΔΥΣΕΩΝ
ΓΓΕΚ
ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ
ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ

MŠMT
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

bpifrance

fwo

Ministero dello
sviluppo economico

Swedish
Research
Council

euresearch
Swiss guide to European
research and innovation

REPUBLIC OF CROATIA
Ministry of Science and
Education

EuroHPC
Joint Undertaking

These projects have received funding from the EuroHPC JU, under Grant Agreements **955606, 955776, 955811, 956831, 101033975**.

The EuroHPC Joint Undertaking (JU) receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, Croatia, Czech Republic, France, Germany, Greece, Ireland, Italy, Poland, Spain, Sweden, and Switzerland.

THANK YOU

I/O, Storage and Data Management

IO Storage and Data Management

- Today HPC systems with up to 1 Eflops (10^{18} flops)
  - How to optimise use of resources such as storage and network ? → They have now become the bottleneck
  - How to optimise data placement ? → Too expensive (energy-wise)

- These questions are here to stay:
  - Expected data volumes for applications such as DestinE or SKA: 1 PB/day
  - Merging AI/ML with HPC: Evolving I/O dynamics and many different workloads

Two projects for looking into these questions:

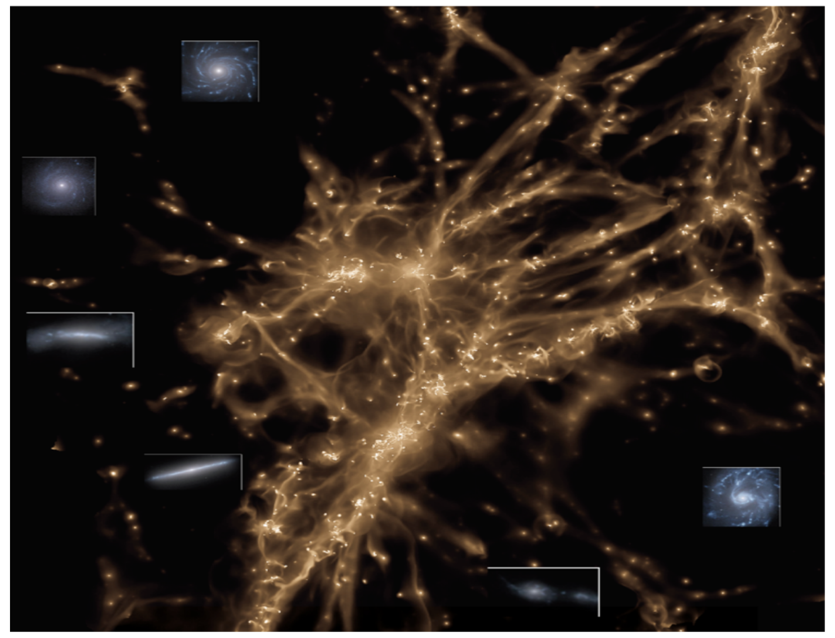
IO-SEA

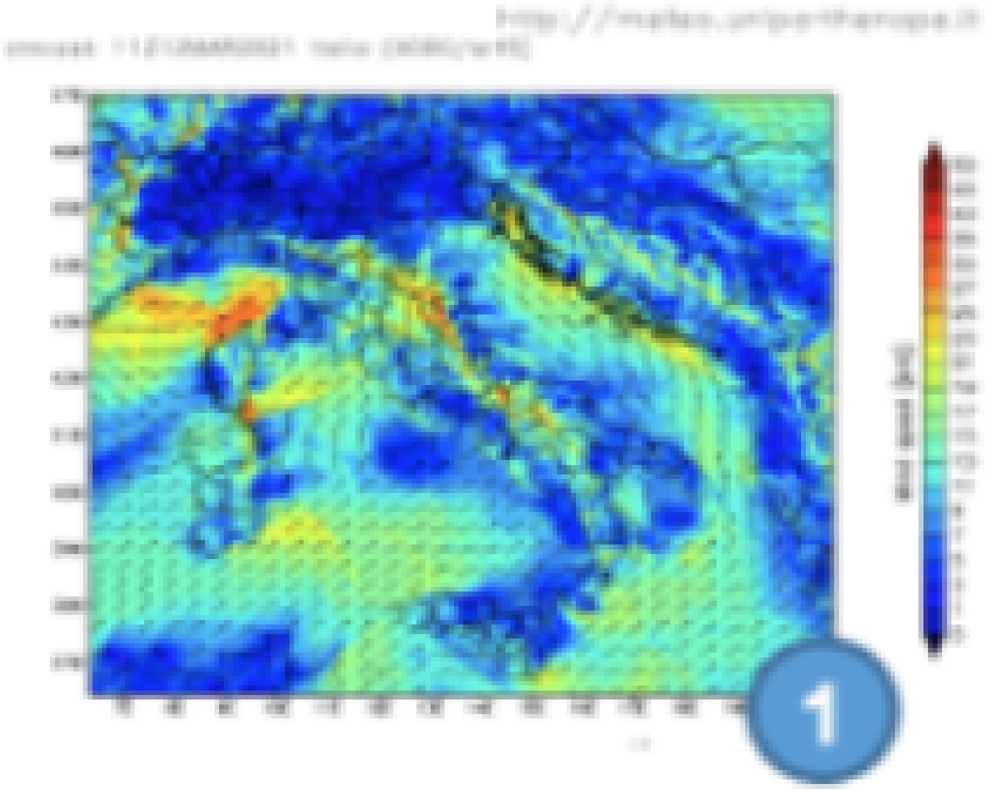
ADMIRE

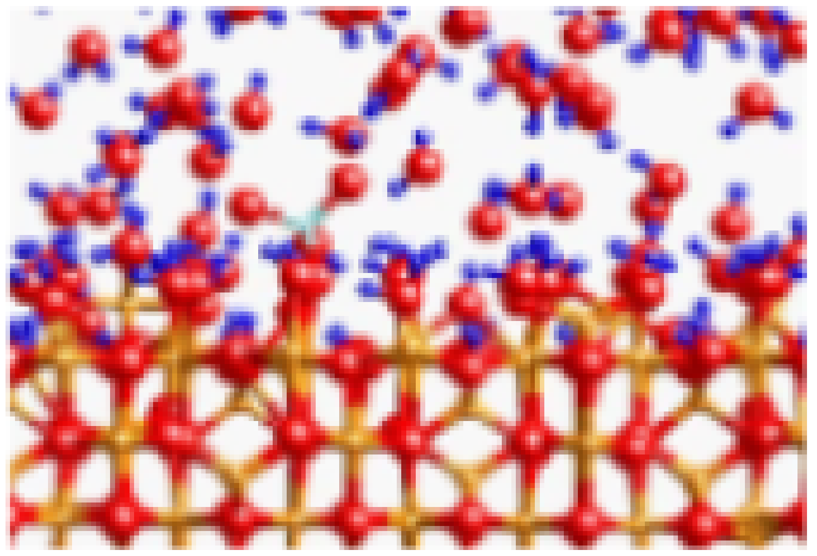
malleable data solutions for HPC

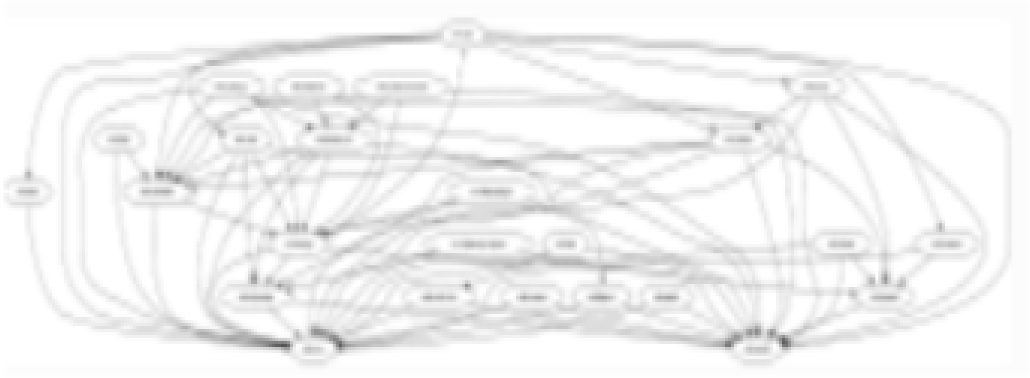
Understanding the applications

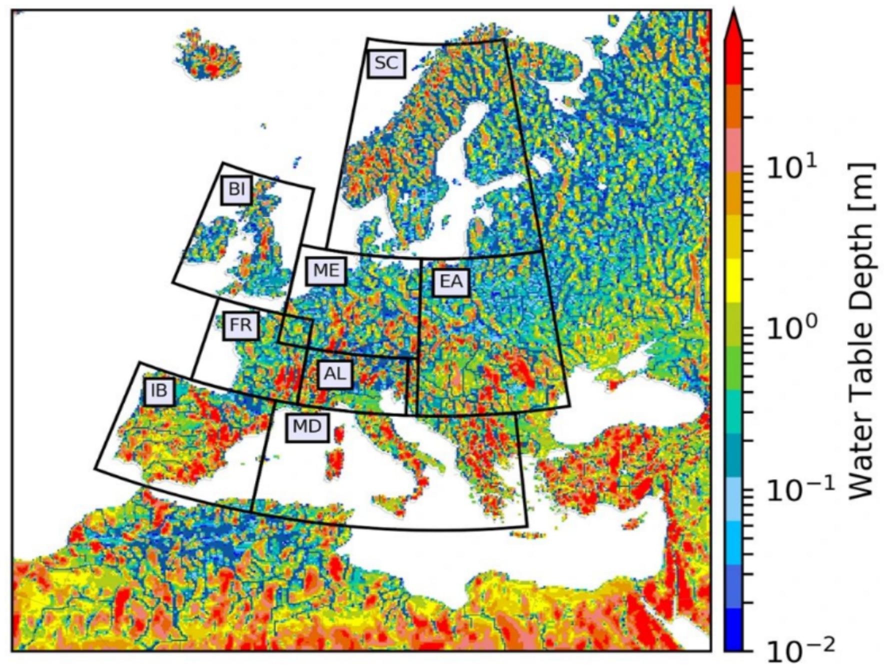
- Weather and Climate
- Molecular Dynamics
- Astrophysics
- CryoEM
- Earth modelling
- Lattice Quantum-Chromodynamics
- Software Heritage Management





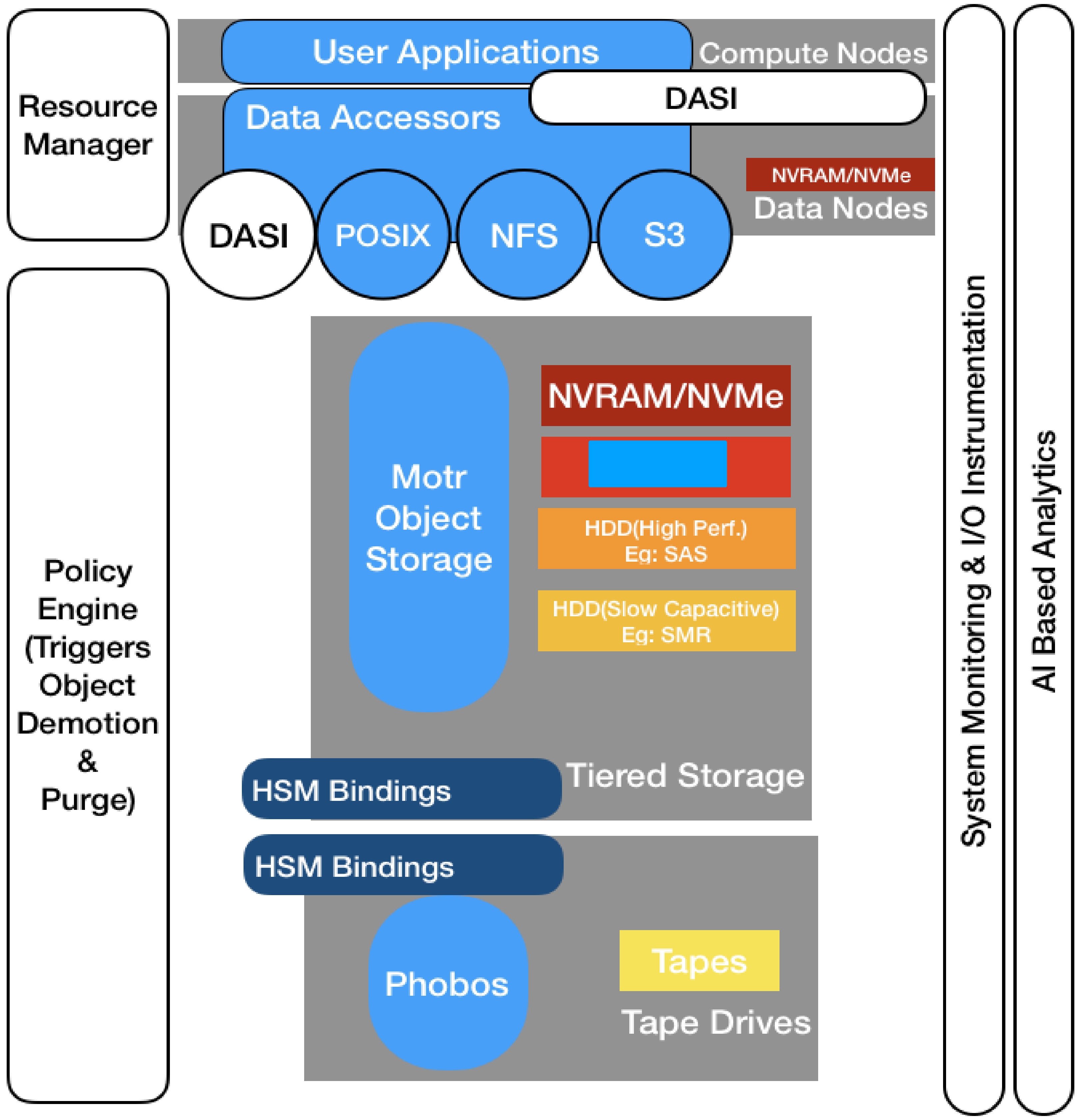






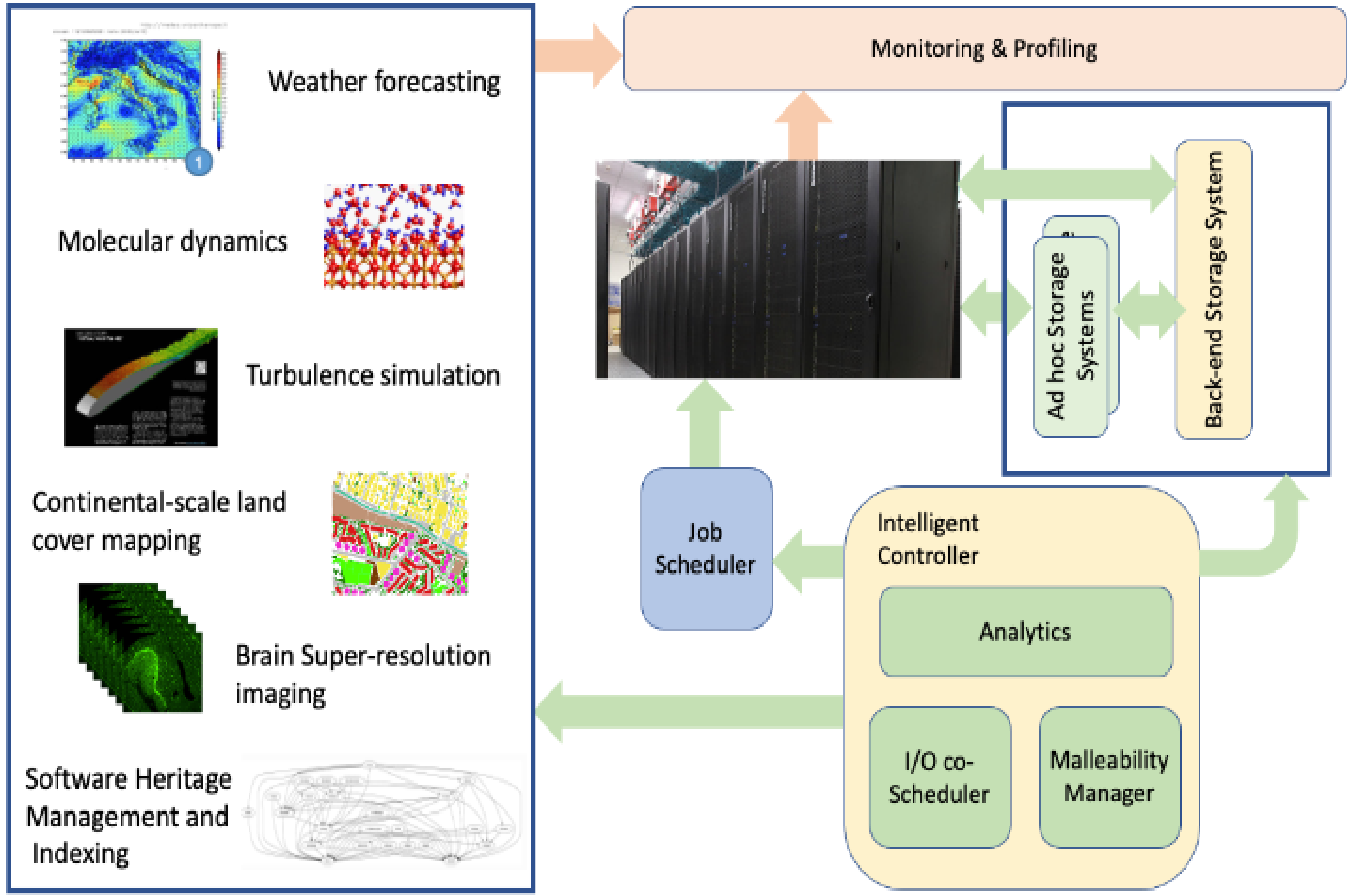
IO-Software for Exacale Architectures (IO-SEA)

- Design and development of a novel management and storage platform
  - Usage of Object Stores
  - Hierarchical storage management (HSM)
  - On-demand/Ephemeral provisioning of storage services & scheduling
- Co-design with next generation I/O intensive HPC oriented applications
  - Development of new flexible application Interface ("DASI")
  - Tools for monitoring and instrumentation allowing the application end users to understand the I/O behaviour of their applications



Adaptive multi-tier data management (ADMIRE)

- Flat storage hierarchies no longer satisfies performance requirements of data-driven HPC applications
- Design an active and dynamic I/O stack, steered by:
  - Intelligent global coordination
  - Computational and I/O malleability
  - Storage resource scheduling across the I/O stack
- Transparent usage of high-performance ephemeral distributed file systems that are deployed ad-hoc
  - Transparent integration through POSIX interface
  - Transparent control and data staging
  - Integration into Lustre HSM
  - Elasticity features on runtime for ad-hoc file systems
- Co-design with several I/O intensive applications covering typical HPC applications (e.g., CFD) as well as ML/AI and workflows



Need to understand the applications (from an I/O point of view)

- Understand the storage behavior of HPC applications
  - Metadata operations, I/O patterns
  - Impact on I/O requirements for various workloads
- Challenges
  - All applications may not run in the same environment
  - General cluster dependencies and hardware availability
  - Instrumentation generally difficult

Our approach :

➢ build a profiling and tracing platform which runs in all environments

➢ Create an analysis tool for visualizing gathered statistics and traces

The I/O Traces Initiative

Rich data collection:
Context-aware I/O tracing
for comprehensive analysis

Cross-disciplinary collaboration:
Breaking down silos between
scientific fields

Empowering developers:
Access to in-depth I/O data
to enhance application tuning

AI/ML I/O patterns:
Addressing the unique needs
of modern HPC workloads

FAIR I/O Traces

Findable	**A**ccessible	**I**nteroperable	**R**eusable
Joint metadata scheme makes data sets Findable within and across archives	Archive will build on European initiatives like OpenAIRE and Zenodo to ensure long-time Accessibility	Traces will follow (de-facto) standards like the Open Trace Format and Darshan file format to enable Interoperability	Reusability will be ensured by open data formats and by automatic analysis inside web-environment

And… what is in for your ?

If you are an application developer

- Learn about your applications
- Compare your workload, performance and insights
- Cite your workload
- Execution time
- Datasets

If you are a storage developer

- Understanding I/O requirement
- Insights on I/O bottlenecks
- Research on real application workload

Conclusion

- Smart I/O and data management is crucial to become more energy-efficient
- Smart I/O and data management is crucial for exploiting exascale machines fully

→ Need to better understand our applications with their I/O habits

- The I/O Trace Initiative: A hub for I/O trace collection and analysis: https://hpcioanalysis.zdv.uni-mainz.de
  - Community engagement: Share, find, and use I/O traces for collaborative advancement
  - Advanced features: Submission, archiving, searching, and visualization tools
  - Commitment to FAIR Principles: Enhancing I/O trace accessibility and utility
  - On-going dedication: Platform evolution to support HPC and AI/ML communities

Your contributions wanted !

bitly

SPONSORED BY THE
Federal Ministry
of Education
and Research

MINISTERIO
DE CIENCIA
E INNOVACIÓN

Financiado por
la Unión Europea
NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

AGENCIA
ESTATAL DE
INVESTIGACIÓN

bpifrance

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ
ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΕΠΕΝΔΥΣΕΩΝ
ΓΓΕΚ
ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ
ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ

MŠMT
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

fwo

Ministero dello
sviluppo economico

Swedish
Research
Council

FCT
Fundação para a Ciência e a Tecnologia

The Research Council
of Norway

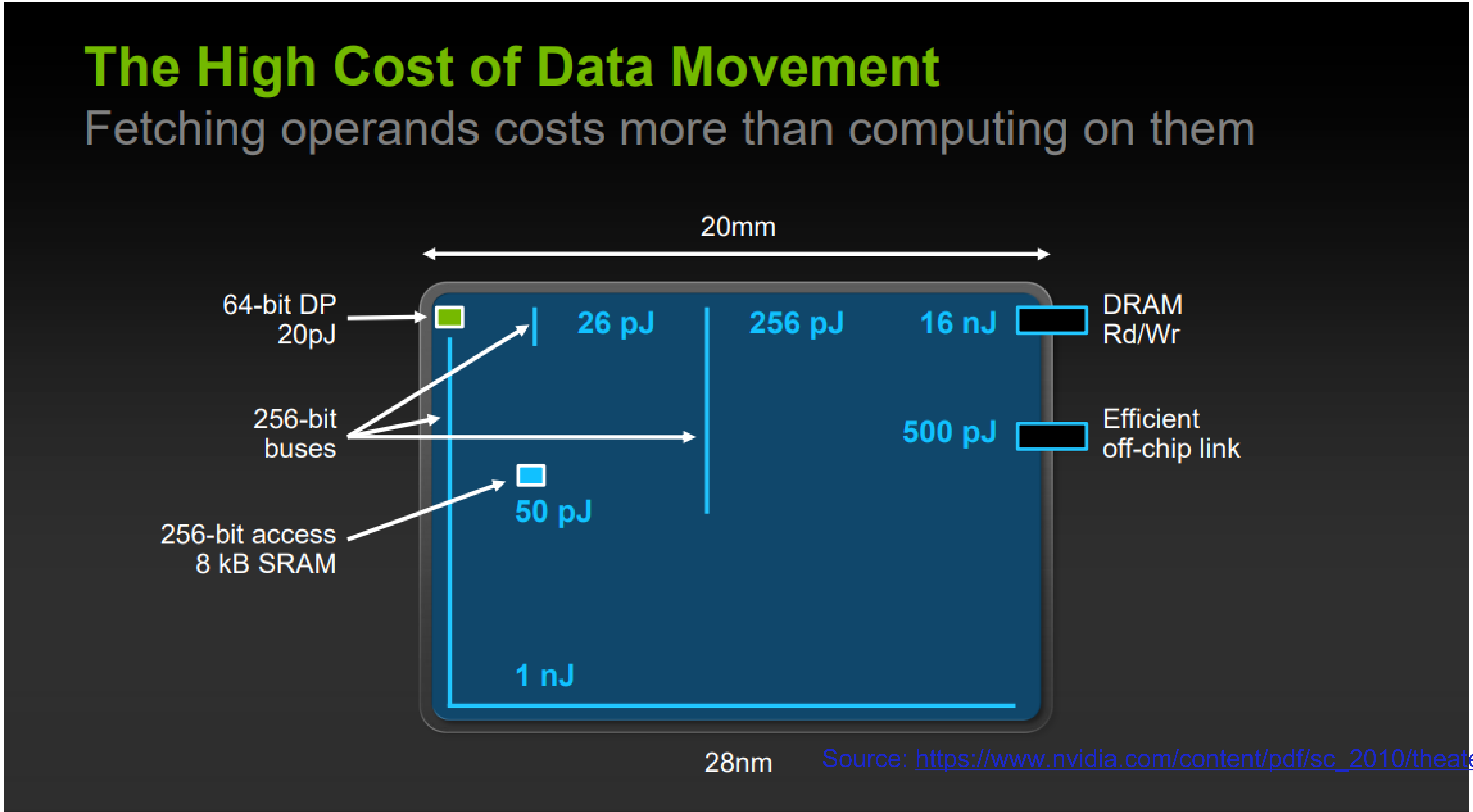
euresearch
Swiss guide to European
research and innovation

These projects have received funding from the EuroHPC JU, under Grant Agreements **956748, 955606, 955811, 955776, 956831, 956560, 101034126, 956702, 101033975**

The EuroHPC Joint Undertaking (JU) receives support from the European Union’s Horizon 2020 research and innovation programme and Belgium, Cyprus, Czech Republic, France, Germany, Greece, Ireland, Italy, Luxembourg, Norway, Poland, Portugal, Spain, Sweden, United Kingdom, Switzerland, and Turkey.

THANK YOU

Smart data placement urgently needed



The High Cost of Data Movement

Fetching operands costs more than computing on them

20mm

64-bit DP
20pJ

26 pJ

256 pJ

16 nJ

DRAM
Rd/Wr

256-bit
buses

500 pJ

Efficient
off-chip link

50 pJ

256-bit access
8 kB SRAM

1 nJ

28nm

Source: https://www.nvidia.com/content/pdf/sc_2010/theater/dally_sc10.pdf

ANTWERP

Co-Designing the Future of European HPC:

European Processor Ecosystem

19.03.2024

ADMIRE
malleable data solutions for HPC

DEEP-SEA

RED-SEA

textarossa

EUPEX
European Pilot for Exascale

IO-SEA

eProcessor

TIME-X

THE EUPILOT

UNLEASHING THE POWER OF EUROPEAN HPC AND QUANTUM COMPUTING

EuroHPC Summit 2024, Antwerp, 18 – 20, March 2024

European Processor Ecosystem: Summary

Objectives of this technology area

- Develop European know-how on the design and development of processors
- Foster Europe's technological sovereignty

- Projects highlighted in this presentation today
    - **eProcessor:** an open-source RISC-V OoO CPU with a full-stack European ecosystem, extendable & energy efficient architecture
    - **EUPEX:** academic and commercial project to co-design a European modular exascale-ready pilot system. Leverage contributions from EPI project (processors)
    - **EUPILOT:** open-source and open standards pre-exascale system based on accelerators for HPC and AI/ML with European components. Leverage contributions from EPI project (accelerators)

eProcessor

EuroHPC
Joint Undertaking

The eProcessor project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No. 956702. The JU receives support from the European Union’s Horizon 2020 research and innovation programme and Spain, Sweden, Greece, Italy, France, and Germany.

eProcessor

High performance 64-bit RISC-V Processor

- 2-way Out-of-Order Core
- Single core & multi-core
- Cache coherent implementation
- Adaptive caches/scratchpad memories
- HW/SW fault tolerance

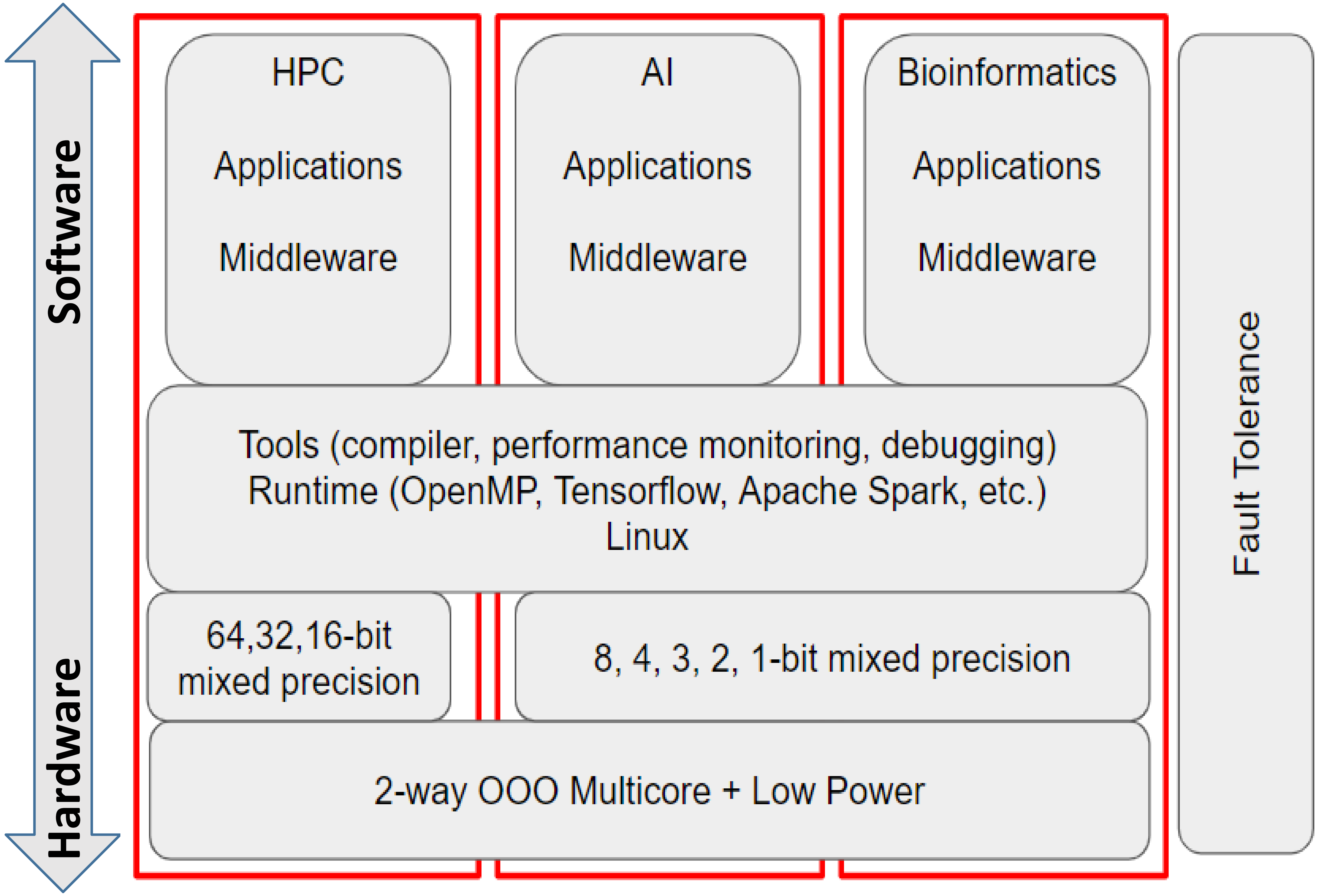
On chip Vector + AI + Bioinformatics accelerator co-processor, reduced & mixed-precision

Off-chip coherent CNN FPGA accelerator

eProcessor Co-Design Approach

Software/hardware co-design for improved application performance and energy efficiency

- HPC
- HPDA (AI/ML/DL)
- Bioinformatics


Software
Hardware
HPC
Applications
Middleware
AI
Applications
Middleware
Bioinformatics
Applications
Middleware
Tools (compiler, performance monitoring, debugging)
Runtime (OpenMP, Tensorflow, Apache Spark, etc.)
Linux
64,32,16-bit mixed precision
8, 4, 3, 2, 1-bit mixed precision
2-way OOO Multicore + Low Power
Fault Tolerance

eProcessor Application Use Cases

Targeted Applications	SW/HW Co-Designed Optimizations
• HPC domain ◦ NAS benchmark suite	• Double precision floating point vector instructions • Scratchpad memories • HW/SW dead block management
• Bioinformatics domain ◦ Smith-Waterman-Gotoh ◦ Banded Smith-Waterman-Gotoh ◦ WaveFront Alignment (WFA) ◦ FM-index	• New ad-hoc vector instructions for bioinformatics • Vector instructions with narrow integer data types
• AI domain ◦ DeepHealth Toolkit ◦ Smart Mirror ◦ Surveillance Border Control	• Offloading to CNN FPGA off-chip accelerator • Offloading to on-chip systolic array • Reduced precision floating point vector instructions

eProcessor Future Plans

Wider variety of use cases

- The eProcessor hardware is programmable: it could be leveraged in other application domains
  - eProcessor focuses in Deep Learning (DL) AI use cases, but Large Lange Models (LLMs) could benefit from the systolic array and the mixed precision functional units
  - Many application domains could benefit from the cache-coherent off-chip FPGA accelerator

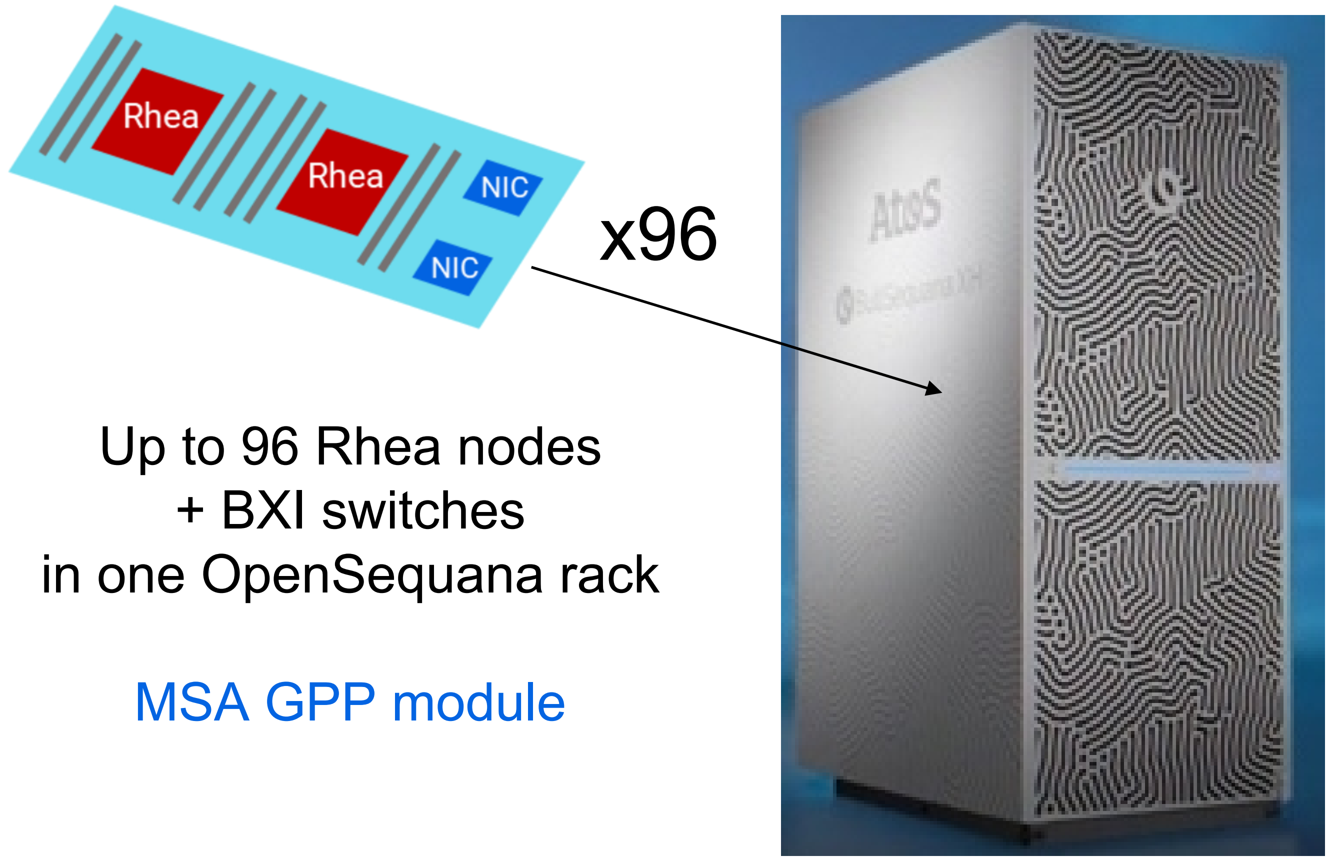
Future plans and exploitation

- Ongoing efforts in defining detailed exploitation plans and licensing schemes
  - Hardware IP blocks open for non-commercial use
- Currently, the eProcessor co-design approach focuses mostly on application and hardware
  - More involvement from runtime systems, libraries and compilers are of interest

EUPEX Objectives

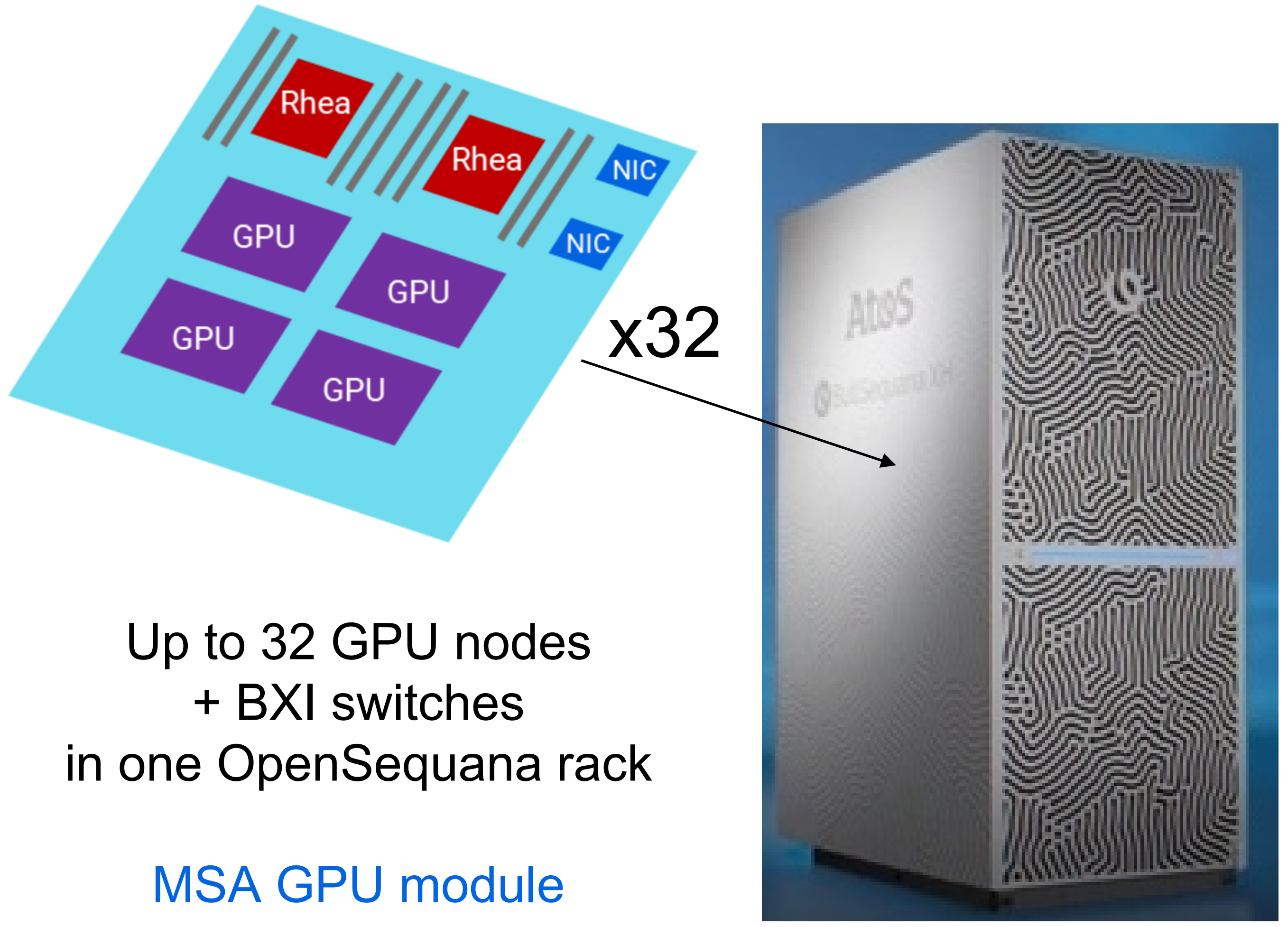
Co-design	Co-design a modular Exascale-pilot system
Deploy	Build and deploy a pilot hardware and software platform integrating European technology
Demonstrate	Demonstrate the readiness and the scalability of the pilot technology in general and the MSA in particular, towards Exascale
Applications	Prepare applications and European users to efficiently exploit the future Exascale machines

The EUPEX Pilot System



Up to 96 Rhea nodes
+ BXI switches
in one OpenSequana rack

MSA GPP module



Up to 32 GPU nodes
+ BXI switches
in one OpenSequana rack

MSA GPU module

- **Modular**
  - OpenSequana-compliant hardware platform
  - matching HPC software ecosystem implementing the Modular Supercomputing Architecture
  - to integrate and manage efficiently a variety of hardware modules and to handle heterogeneous workflows
- **Large enough to be a proof of concept**
  - for a modular architecture relying on European technologies, and in particular on EPI technology
  - to demonstrate the Exascale readiness of the applications selected for co-design
- **Production-grade**
  - technical choices guided by the maturity of the European solutions available

Key Application Domains

Explored by EUPEX for co-design and benchmarking

Climatology, meteorology

- ECMWF, CybeleTech, Atos

Engineering

- IT4I, CINECA

Biology and health

- CINECA, CINI, CEA

Astrophysics

- FORTH, INAF, CEA

Seismology

- INAF, CINECA, CINI, GENCI

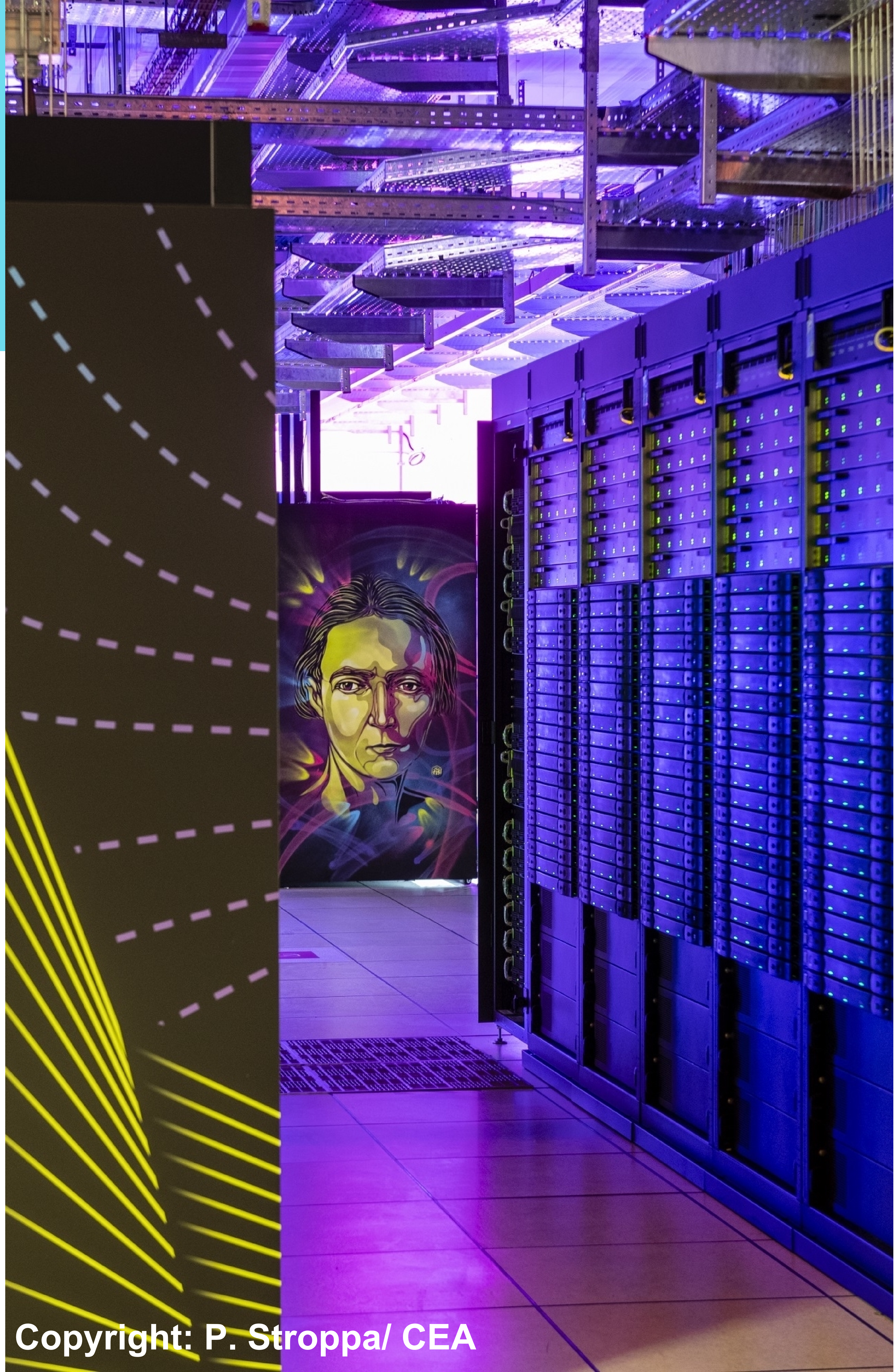
Remote sensing analysis

- FZJ

EUPEX Early Access Programme (EAP)

- EUPEX will open its target Pilot system to interested organisations
- As this Pilot system is only planned for Summer 2025
  - **Phase 1: Access to main Alpha system (CEA Irene A64FX partition) Includes access to planned EUPEX software stack**
  - Phase 2: Access to Pilot system once ready
- **Phase 1 now open** for
  - EuroHPC Centres of Excellence (CoEs)
  - Selected EuroHPC research projects (on request)
- Contact us!
  - https://eupex.eu/early-access-program/#getting-started-with-the-eap

EUPEX



Copyright: P. Stroppa/ CEA

EUPILOT Objectives

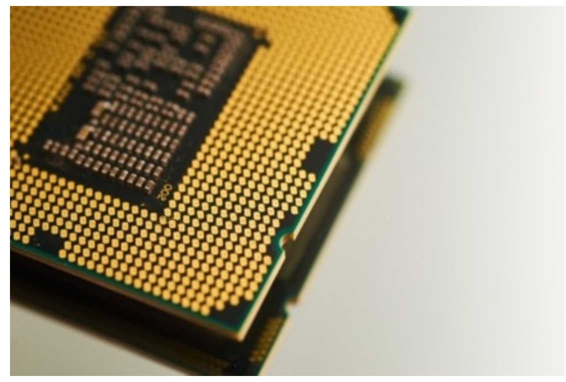
➔ **Demonstrate a European pre-exascale accelerator platform**

- Design, validate, build and deploy a fully-integrated accelerator platform
- Maximize use of European technology & assets
- Stimulate European collaboration, enable future exascale systems
- SW/HW co-design for improved performance and energy efficiency
- Further extend open source into hardware for HPC
- Leverage open source and the RISC-V ISA

➔ **Strengthen European digital autonomy and supply chain**

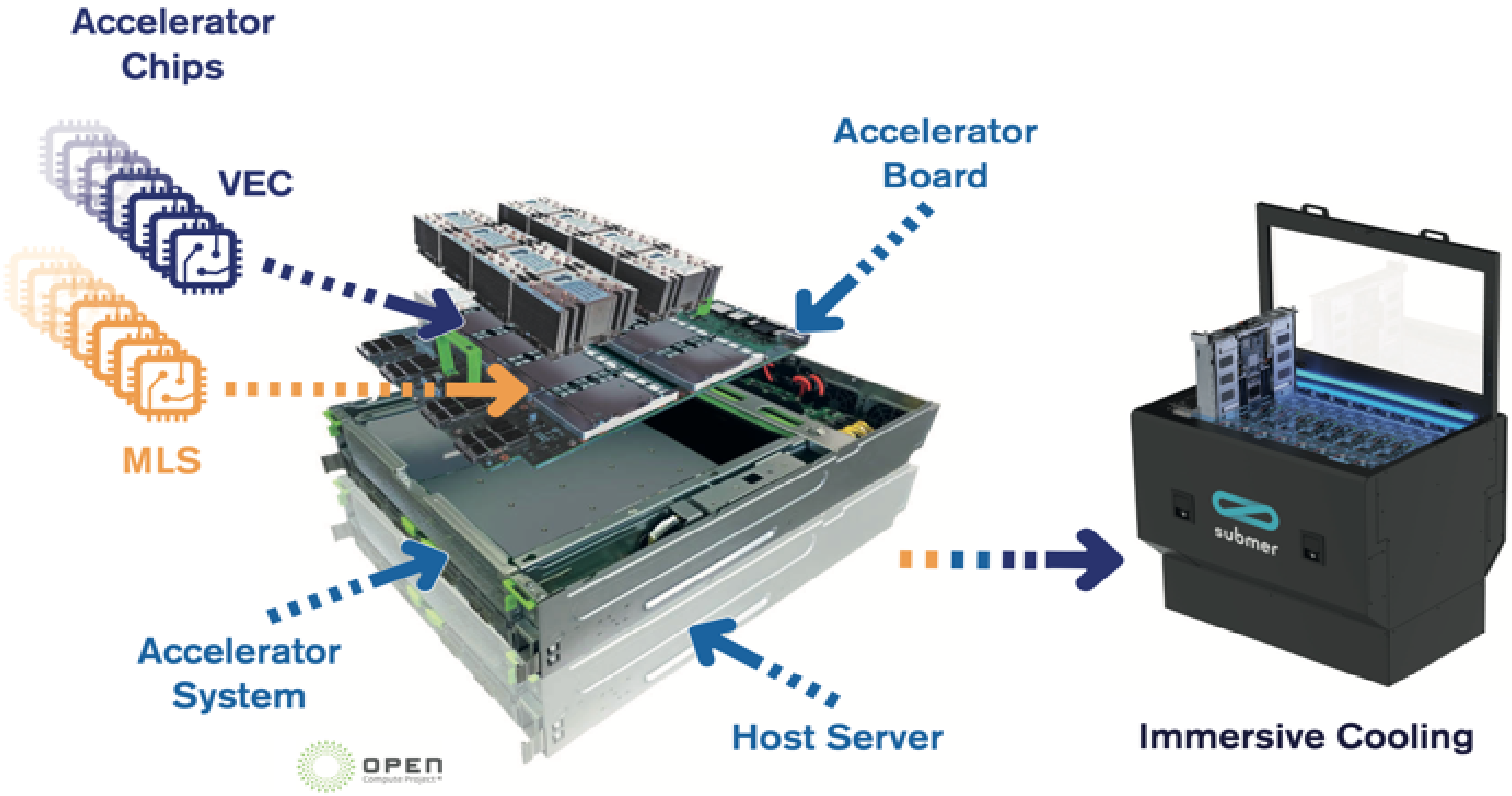
Target: Chips → Deployments

- → **HW: From Chips to Modules to Boards**
- → **SYS: From Boards to Systems to Liquid Immersion Deployments**
- → **SW: From Drivers to OS to Compilers to Frameworks to Apps**

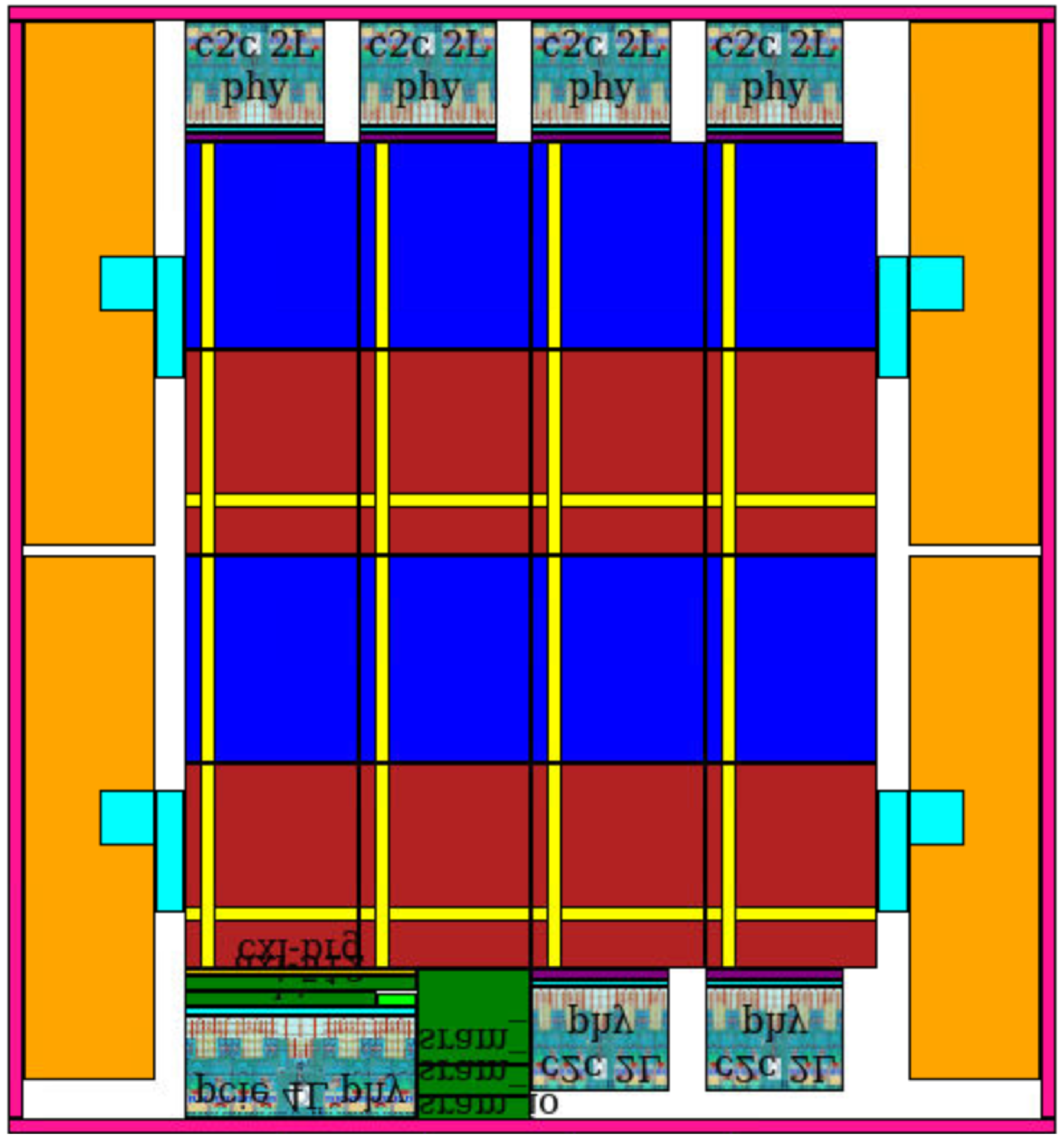


- Leverage accelerator results from EPI project
- Three tape-outs .. from 22nm → 12nm
  - One test chip
  - Two accelerator chips: one VEC & one MLS
- Target HPC (VEC) and HPDA (MLS) applications
  - Port apps/frameworks/libraries to RISC-V
  - Develop toolsets for manual or automatic optimizations
- Deploy OCP-based systems to datacenter with liquid immersion cooling tanks
  - Higher workloads, density & capacity
  - Reduced environmental impact
  - Significantly reduced targets for water usage and PUE from ~1.08 to ~1.03

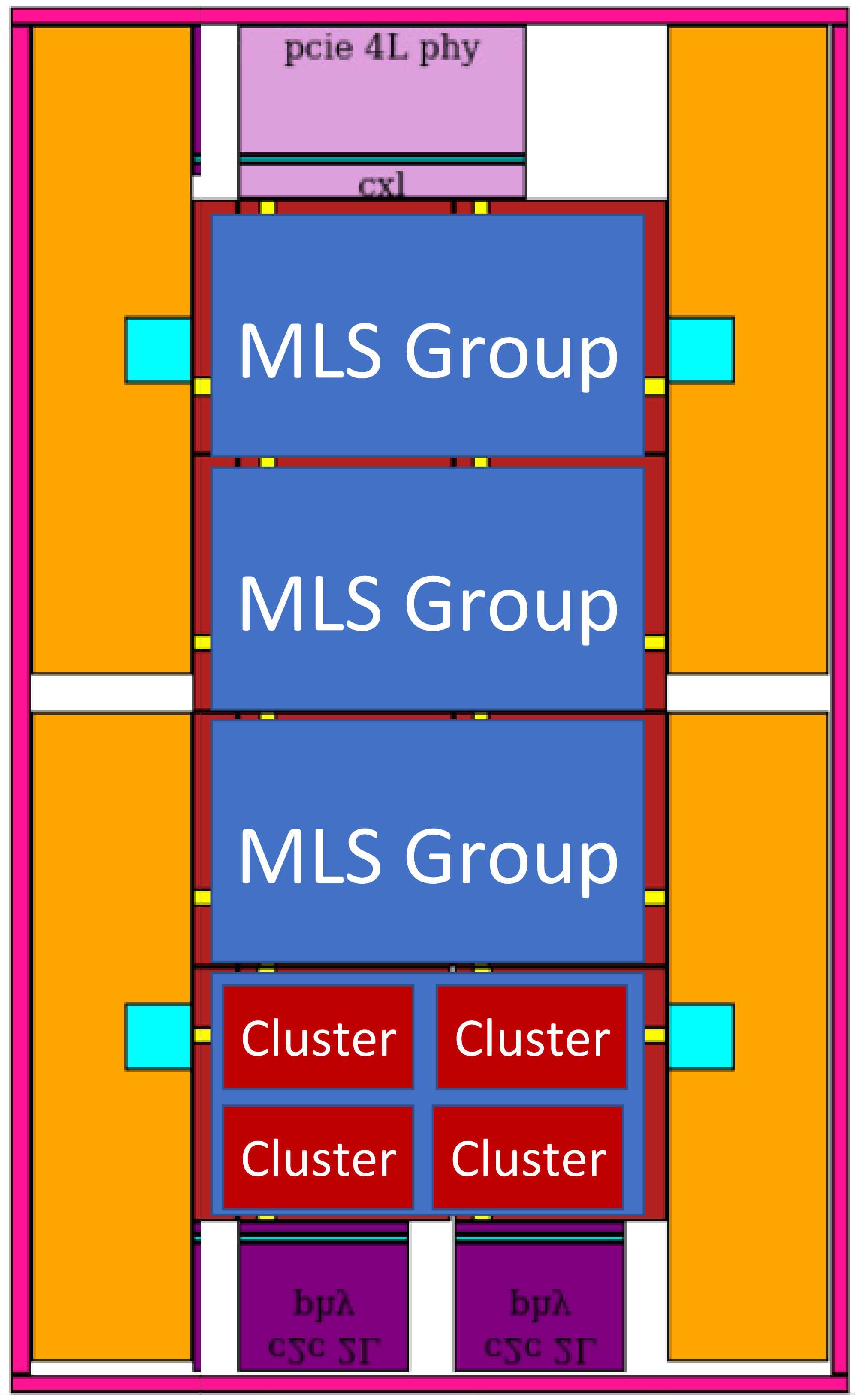
EUPILOT Top Level View



EUPILOT Chip Layouts

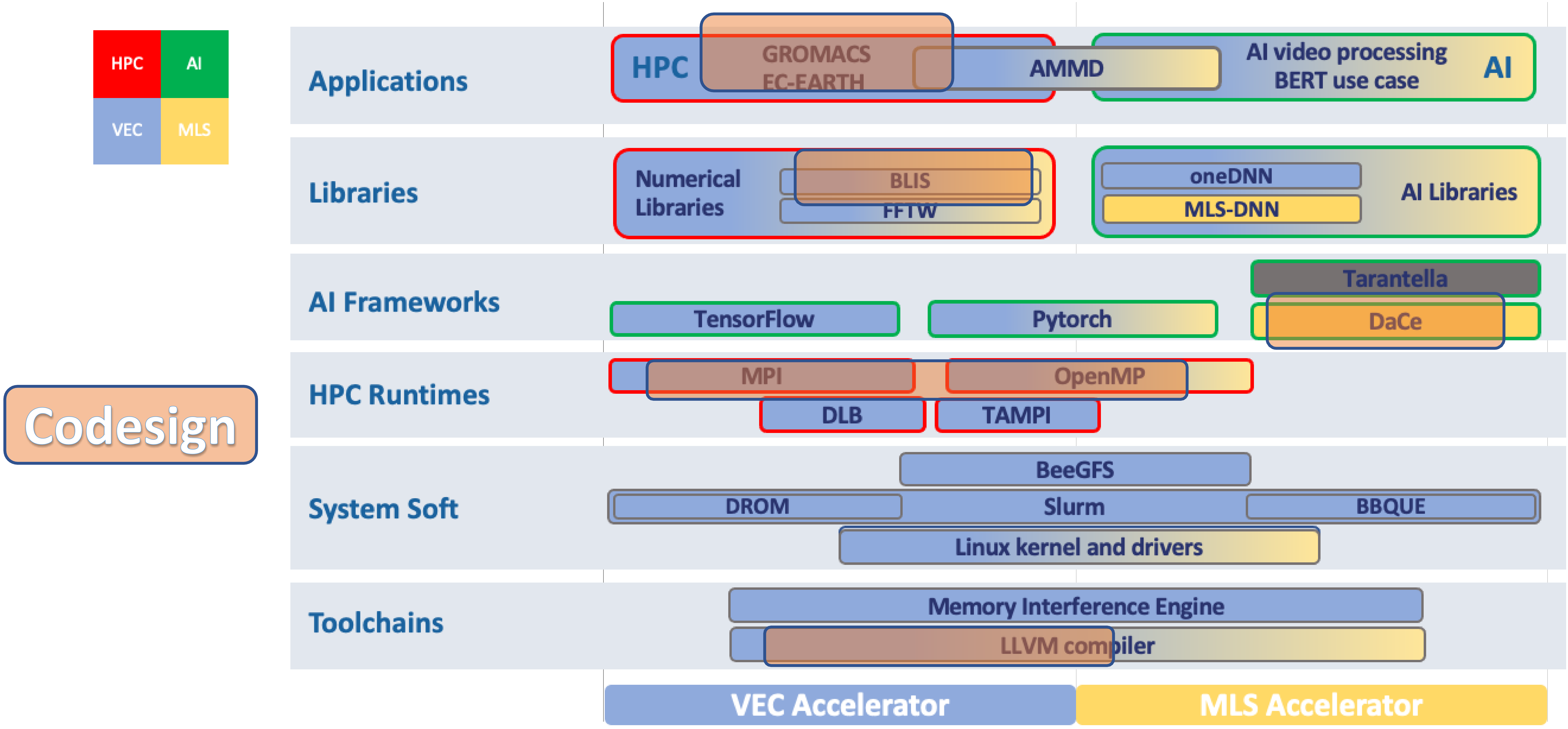


VEC



MLS

Software Stack & Co-design



HPC | AI
VEC | MLS

Codesign

Applications: HPC, GROMACS, EC-EARTH, AMMD, AI video processing, BERT use case, AI

Libraries: Numerical Libraries, BLIS, FFTW, oneDNN, MLS-DNN, AI Libraries

AI Frameworks: TensorFlow, Pytorch, Tarantella, DaCe

HPC Runtimes: MPI, OpenMP, DLB, TAMPI

System Soft: BeeGFS, DROM, Slurm, BBQUE, Linux kernel and drivers

Toolchains: Memory Interference Engine, LLVM compiler

VEC Accelerator | MLS Accelerator

SPONSORED BY THE
Federal Ministry
of Education
and Research

MINISTERIO
DE CIENCIA
E INNOVACIÓN

Financiado por
la Unión Europea
NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

AGENCIA
ESTATAL DE
INVESTIGACIÓN

bpifrance

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ
ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΕΠΕΝΔΥΣΕΩΝ
ΓΓΕΚ
ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ
ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ

MŠMT
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

fwo

Ministero dello
sviluppo economico

Swedish
Research
Council

FCT
Fundação para a Ciência e a Tecnologia

The Research Council
of Norway

euresearch
Swiss guide to European
research and innovation

- These projects have received funding from the EuroHPC JU, under Grant Agreements **956748, 955606, 955811, 955776, 956831, 956560, 101034126, 956702,** 101033975

- The EuroHPC Joint Undertaking (JU) receives support from the European Union’s Horizon 2020 research and innovation programme and Belgium, Cyprus, Czech Republic, France, Germany, Greece, Ireland, Italy, Luxembourg, Norway, Poland, Portugal, Spain, Sweden, United Kingdom, Switzerland, and Turkey.

THANK YOU

ANTWERP

Co-Designing the Future of European HPC: Eco-friendly Technologies, Systems and Software

19.03.2024

UNLEASHING THE POWER OF EUROPEAN HPC AND QUANTUM COMPUTING

Technology Area: Dynamic Resource Management

- HPC community has to cope today with highly complex HPC applications and workflows with new requirements:
  - Scientific workflows that leverage parallelism differently (AI/ML simulations combined with MPI, ensemble runs, ..)
  - Converged computing: Leveraging both HPC and cloud resources to improve turnaround times
  - Hardware overprovisioning: efficient system design under power constraints and support for variable capacity scheduling
  - Resource Heterogeneity: mapping of workload to increase energy efficiency, as fragmentation may lead to underutilization,
- Currently deployed tools (MPI, SLURM, PFS, ..) are very rigid and cannot handle those requirements on runtime.
  - Frameworks and tools are specialized for some solutions, I/O systems are static in most HPC systems, MPI is not readily amenable for elasticity….
- Joint effort to extend them or create new tools to provide dynamicity on resource management, both for computing and I/O resources, that can allow runtime system adaptation to changing application requirements.
  - Projects involved – ADMIRE, DEEP-SEA, IO-SEA, Regale, Time-X

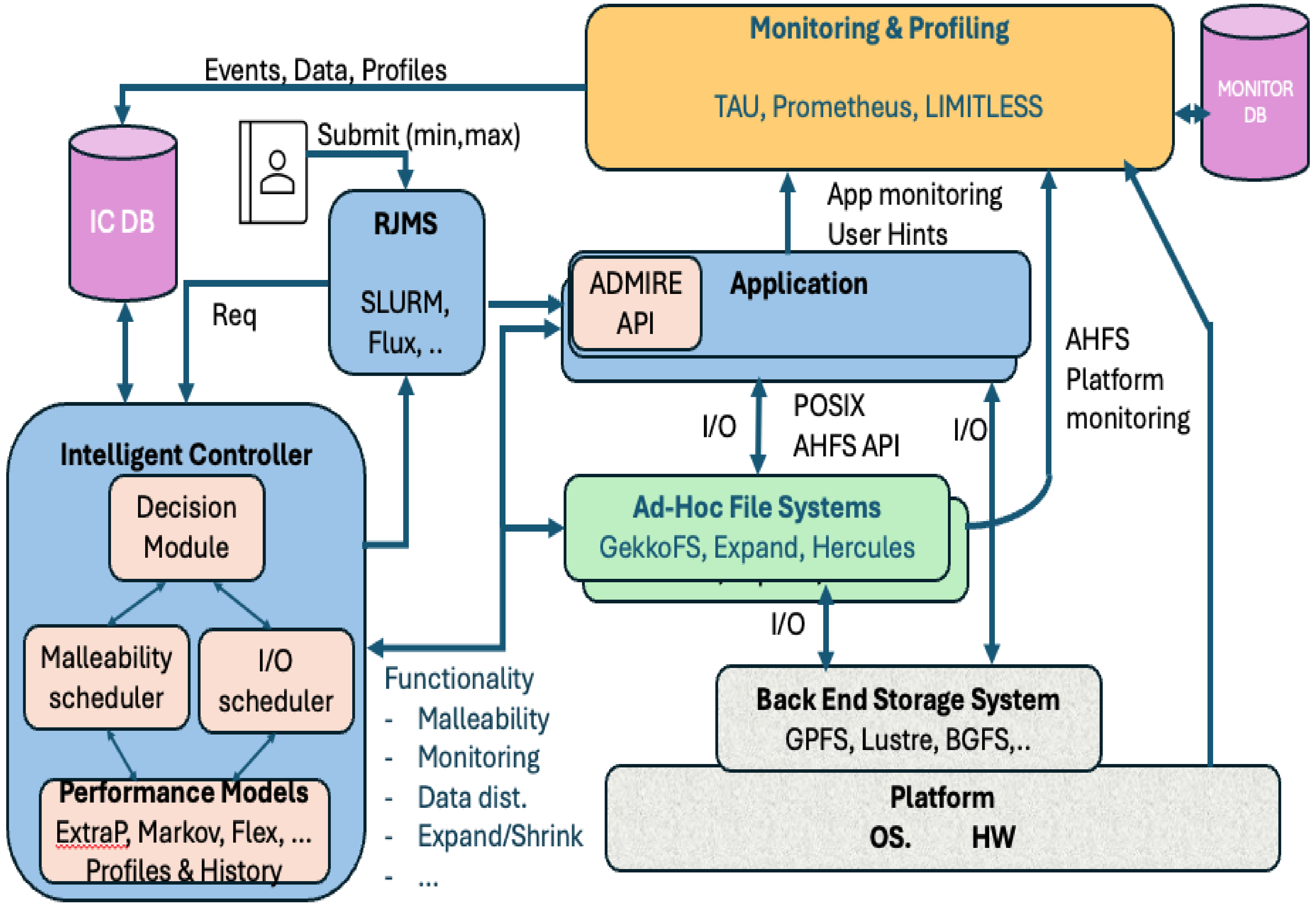
Dynamic Resource Management: ADMIRE Project

ADMIRE provides a framework for DRM that is valid for traditional HPC applications, ML, and workflows.

- Elastic initial deployment of apps and ad-hoc file systems
- Malleability at runtime for processes and AHFS based on application phases' requirements
- Guided by app and intelligent controller (multicriteria schedulers for computing and I/O)
- Holistic monitoring controlling providing profiles, significant events and users' feedback.

Goal: optimizing resources usage while increasing global system throughput by co-scheduling apps and I/O.

- Transparent app co-design through a reduced API
- Energy efficiency is next step



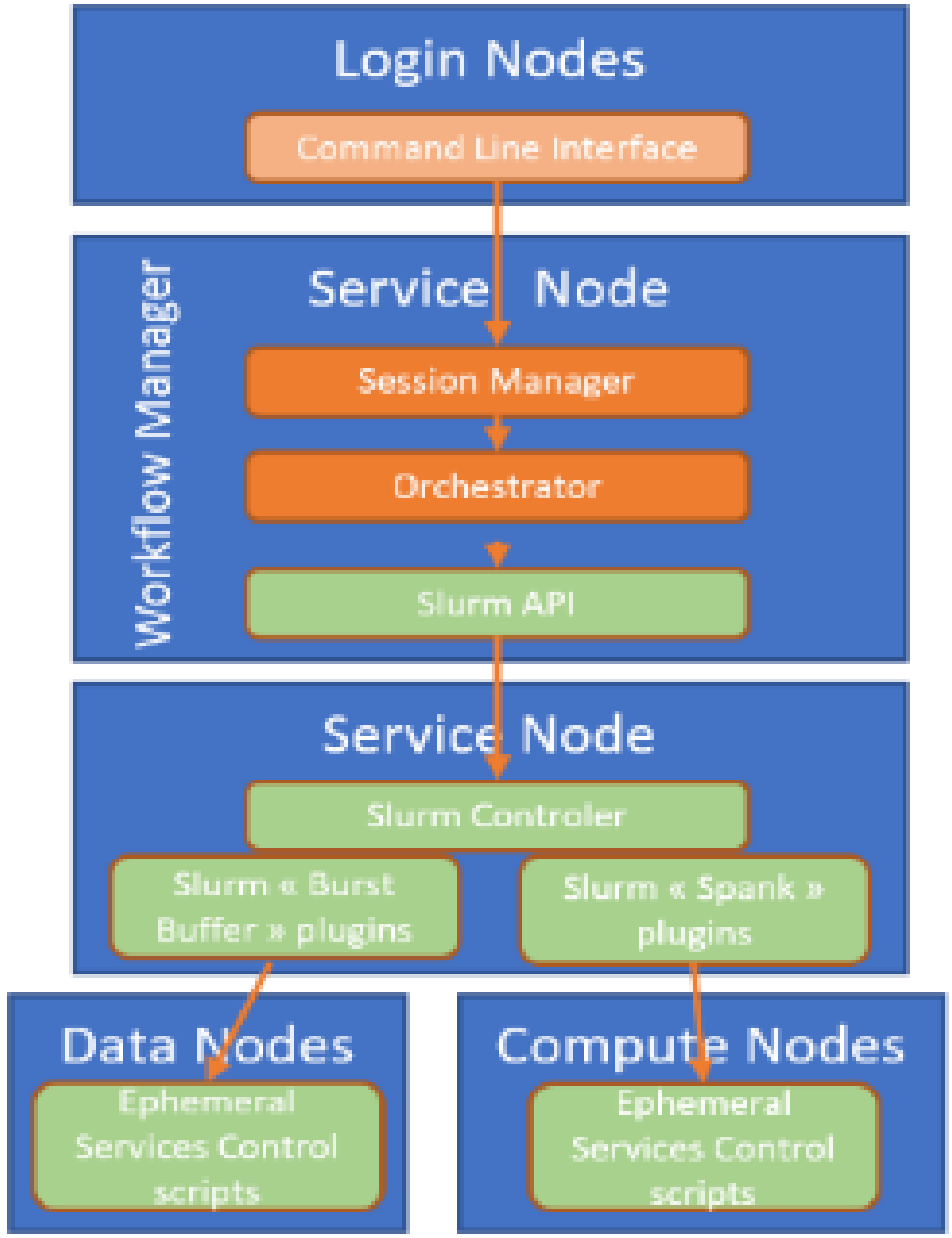
https://admire-eurohpc.eu/

Dynamic Resource Management: DEEP-SEA

- DEEP-SEA proposed and integrated a low-level API with MPI for “Malleability from the Ground Up” based on MPI Sessions and using the PMIx community standard
- Supports applications to request resource changes (“active” malleability) and to react to them (“passive” malleability); data redistribution must be handled by applications
- Integrated and validated with ParaStation MPI (MPICH-based), leading to many clarifications and bug fixes in the PMIX and MPI Session implementations
- Can be used by applications, or serve as a solid “portability” interface for higher-level malleability layers
- DEEP-SEA extended the Slurm job scheduler to handle resource variability during a job (both active and passive malleability)
- Proof of concept implementation using Slurm plugin mechanisms and PMIx

Dynamic Resource Management : I/O-Sea Project

- IO-SEA introduced the paradigm of ephemeral services
- Ephemeral servers run on data nodes and are associated with compute jobs, which are running on compute nodes.
- All IO operations are done through the IO proxy, reducing the pressure on the underlying storage system.
- The IO-Proxy/compute job coupling is performed by applying a scheduling policy, optimized for each use case by a recommendation system,
- This model saves network bandwidth, avoids bottlenecks, and strongly reduces the associated energy consumption.

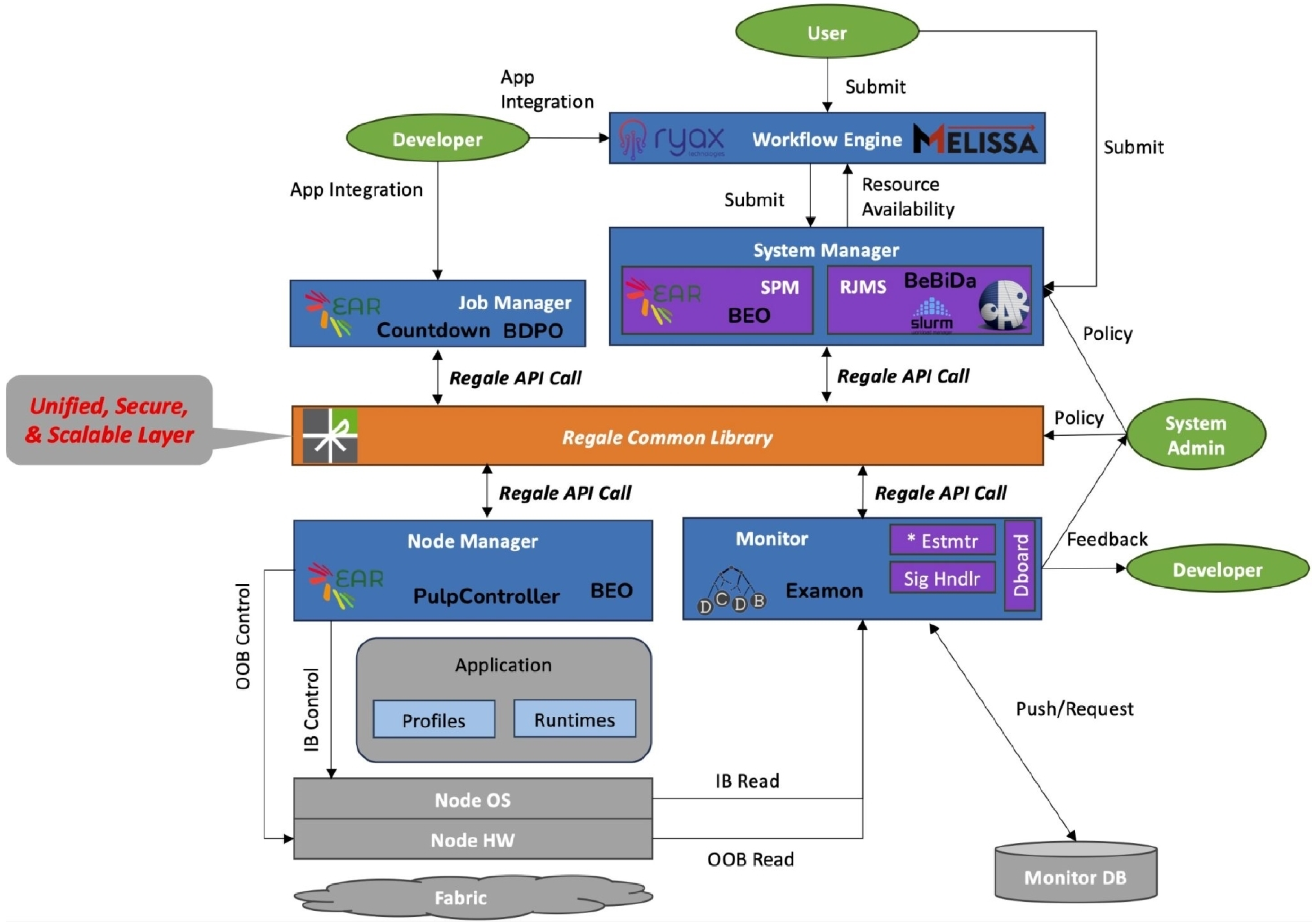


Dynamic Resource Management: REGALE Project

REGALE defines an architecture and implements an instantiation that:

- Switches between different optimization targets (performance, energy, power)
- Adapts to application phases for effective power capping
- Co-schedule applications for increased performance

With a goal to effectively utilize resources in an open, modular, and scalable way



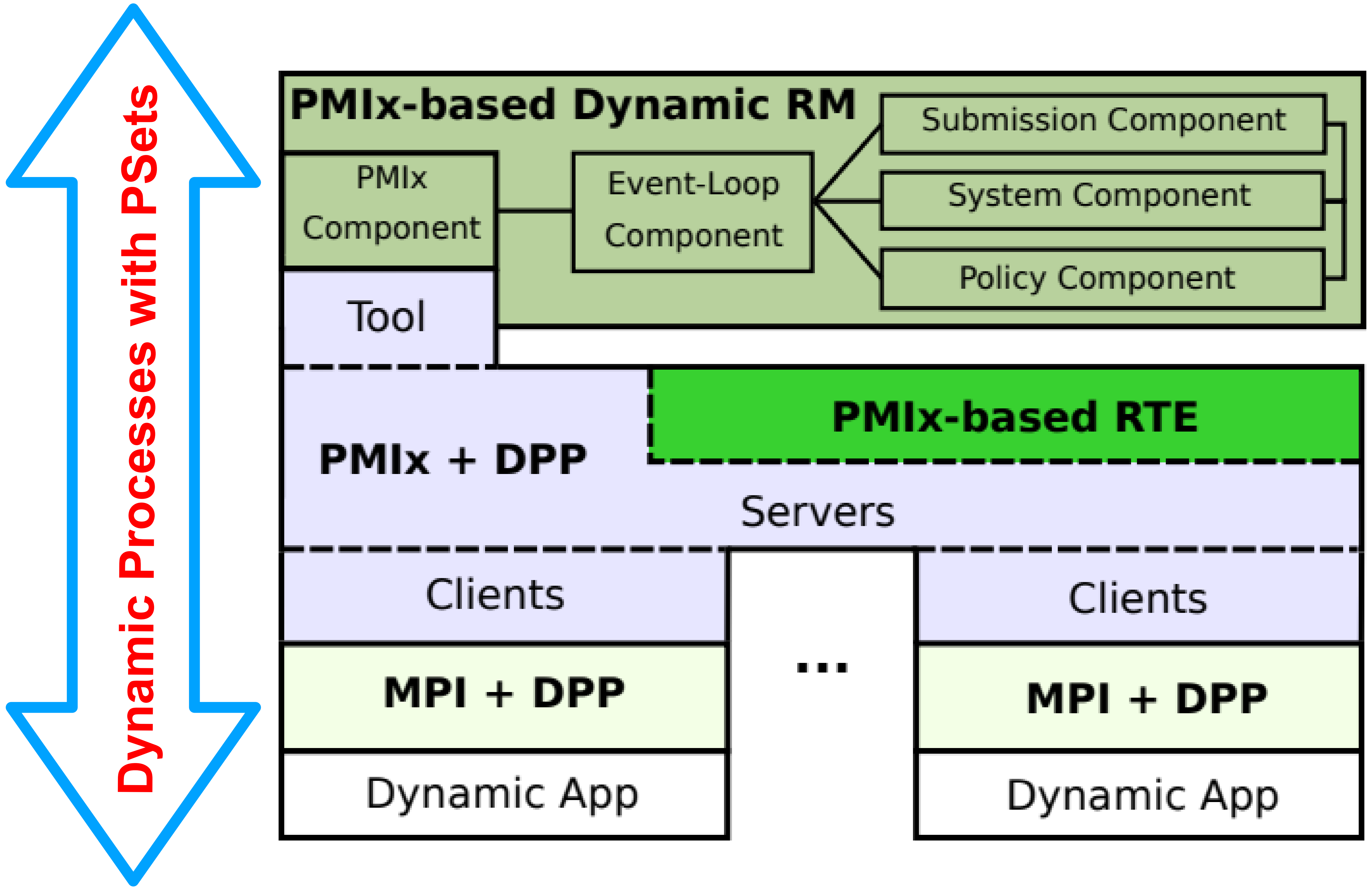
Dynamic Resource Management: Time-X Project

- TIMEX developed a generic programming model for dynamic resources: **Dynamic Processes with PSets (DPP)**

- Flexible **set- and graph-based abstraction** to cover all kinds of application use cases

- **Application/System co-design** for cooperative resource optimization

- **Research-oriented development** to explore new strategies for dynamic applications and scheduling



Dynamic Resource Management: Co-Design Use Cases

- Data-Intensive workflows
  - Monitoring and modelling marine, weather and Air quality (ADMIRE)
  - In-Transit workflow for ubiquitous sensitivity analysis and meta-model Training (REGALE)
  - ECMWF Weather prediction object-based I/O and data management (IO-SEA)
- Machine Learning
  - Continental-scale land cover mapping with scalable and automatic deep learning frameworks. (ADMIRE)
  - Super-resolution imaging using Opera microscopy and SRRF/ImageJ software. (ADMIRE)

- Complex simulations
  - Car-Parrinello molecular dynamic simulation of large molecules and small proteins (ADMIRE)
  - Simulation of large scale turbulent flow (ADMIRE)
  - Industrial scale unsteady adjoint-based Shape optimization of hydraulic turbines (REGALE)
  - Climate simulation (DEEP-SEA)
- HPDA
  - Software Heritage Management & Indexing (ADMIRE)
  - Enterprise Risk Assessment (REGALE)
  - Simulation and data analytics in Earth Science (DEEP-SEA)

Dynamic Resource Management: Exploitable Outcomes

- Prototype of an integrated monitoring environment, including system and apps, that can extract patterns dynamically, and raise signals on thresholds and values defined by the system administrator and users of apps with a usable web interface.
- MPI malleability extensions give application and middleware developers a solid base for using malleability functions; Slurm plugins show a way towards extending Slurm by dynamic resource management.
- Ad-hoc file systems that can be deployed dynamically for each application and adapted dynamically on runtime to cope with changes in I/O phases.
- Application-level DRM tools that can be applied to all kinds of dynamic resource requirements
  - Flex-MPI from ADMIRE, ParaStation MPI from DEEP-SEA, or "Dynamic Processes with PSets" (DPP) from Time-X.
- Architecture to exploit different optimization targets (performance, energy, power) and to co-schedule applications for increased performance and energy efficiency.

SPONSORED BY THE
Federal Ministry
of Education
and Research

MINISTERIO
DE CIENCIA
E INNOVACIÓN

Financiado por
la Unión Europea
NextGenerationEU

Plan de Recuperación,
Transformación y
Resiliencia

AGENCIA
ESTATAL DE
INVESTIGACIÓN

bpifrance

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ
ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΕΠΕΝΔΥΣΕΩΝ
ΓΓΕΚ
ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ
ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ

MŠMT
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

fwo

Ministero dello
sviluppo economico

Swedish
Research
Council

FCT
Fundação para a Ciência e a Tecnologia

The Research Council
of Norway

euresearch
Swiss guide to European
research and innovation

These projects have received funding from the EuroHPC JU, under Grant Agreements **956748, 955606, 955811, 955776, 956831, 956560, 101034126, 956702, 101033975**

The EuroHPC Joint Undertaking (JU) receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, Cyprus, Czech Republic, France, Germany, Greece, Ireland, Italy, Luxembourg, Norway, Poland, Portugal, Spain, Sweden, United Kingdom, Switzerland, and Turkey.

THANK YOU

Pls adapt the slide as per your talk

Dynamic Resource Management: Future Plans

- Determine what kinds of DRM are actually required we need to collect a lot more data.
  - In ADMIRE and I/O-SEA projects we collect I/O traces (I/O traces initiative). We need also Computing behaviour.
- Extending prototypes of holistic system-wide monitoring tools, integrating system, OS, and applications, to extract global knowledge.
- Enhancing performance and energy models with the collected knowledge, to predict short and mid-term needs to optimize resource usage and to increase energy efficiency by steering allocation policies dynamically.
- Together with other collaborators, we see the need for a thorough overhaul of Slurm or for developing a new dynamic resource manager & scheduler together with applications, monitoring, system, and other components. Critical challenge is the buy-in and support by HPC centres.
- Extending/Integrating current prototypes of Flex-MPI, MPI-X, ParaStation MPI and PMIx environments to generate a single environment for application design & development.
- Provide a methodology for code refactoring/porting to the HPC community to help with the transition of applications to DRM.