WORKSHOP

"Preparing Applications and European Users to Efficiently Exploit Future ARM-based Exascale Machines"

January 26 - 28, 2026

Kraków, Poland

Software Ecosystem

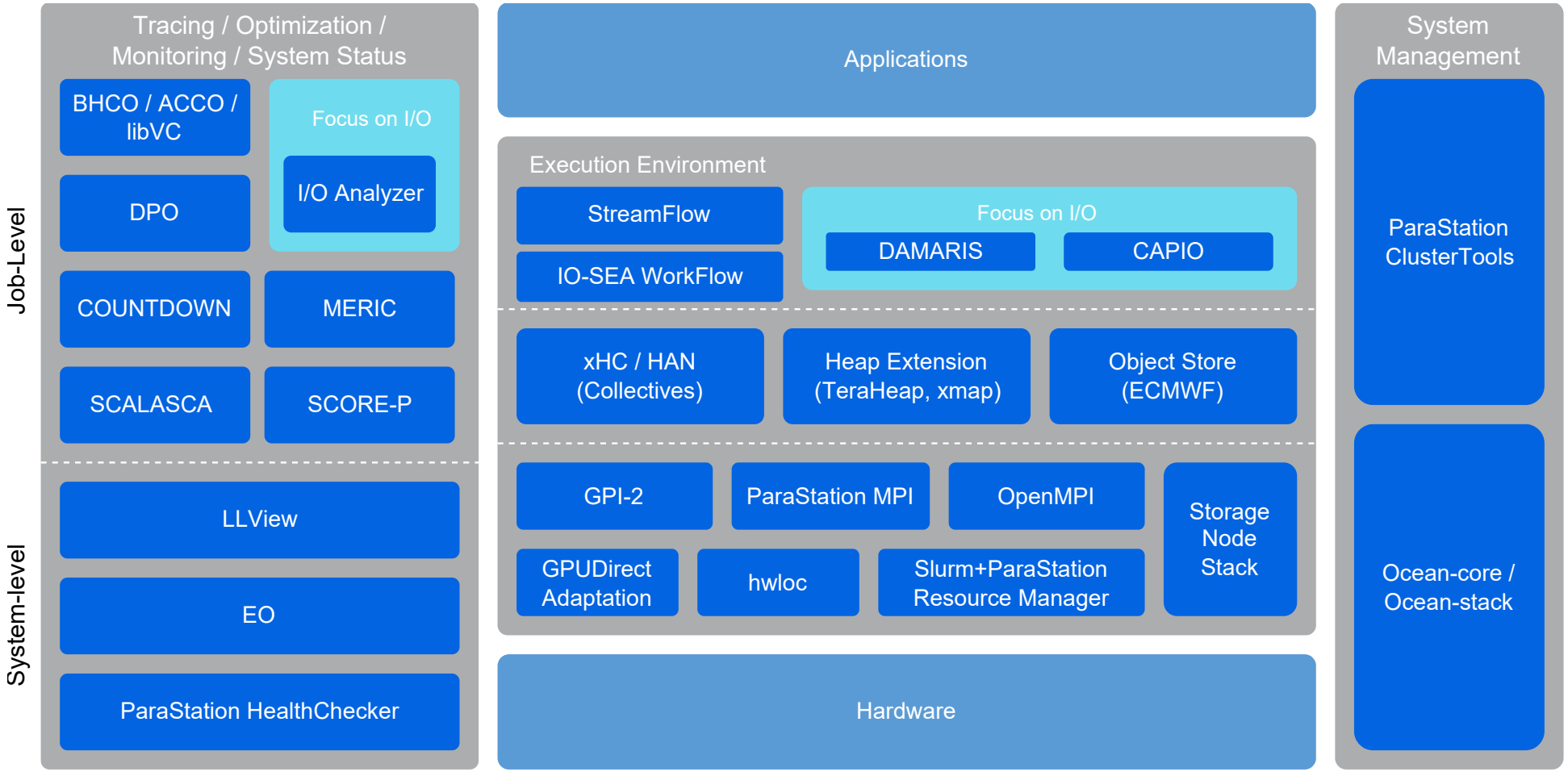
Thomas Moschny

moschny@par-tec.com

EUPEX WP5 Software Stack Objectives/Tasks

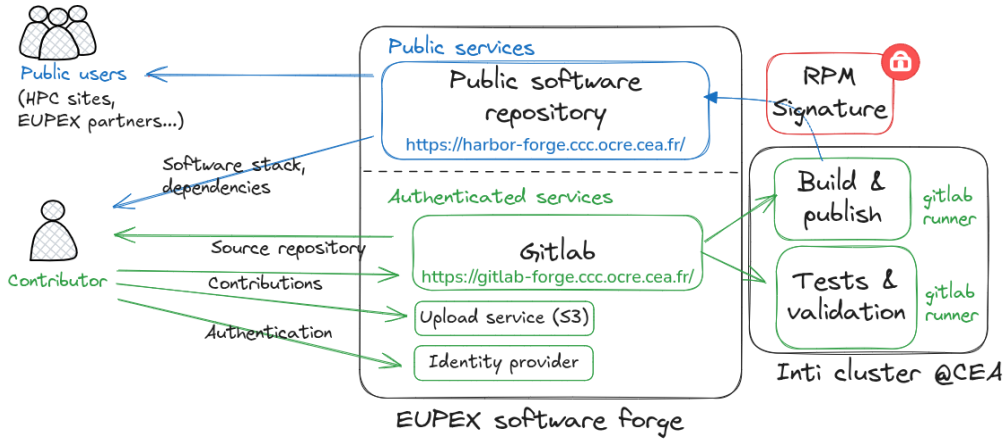
- European **Management Software Stack** for large-scale MSA systems and future modular architectures
- Integration of different components forming the **Execution Environment** for system-level, module-level, and node-level programming of modular architectures
- Tools to monitor, profile, and semi-automatically improve **Energy Efficiency**, leveraging HW power management knobs
- **Storage, I/O capabilities**: Provide fast and scalable data access stack for workflows and applications, leveraging work from other projects

The EUPEX Software Stack



Management Software (Deployment, Monitoring, Provisioning …)

- **Ocean**
  - Ocean-core: RPM-based distribution with packaged tools and SW updates for HPC cluster management
  - Ocean-stack: defines core services architecture, integration of 3rd party services and operating procedures
  - Version 9.4 released (Alma 9.4 MOFED 5.4 Lustre 2.15, full aarch64 support)
  - OcenForge: allows partners to contribute their packages
- **ParaStation Modulo**: Comprehensive software suite for Modular Supercomputing: additional administrative modules
- **LLView**: Graphical system monitoring tool w/ node telemetry
- **Ticketing system** and information hub for projects' systems



https://ocean.eupex.eu/

https://par-tec.com/hpc/

https://www.fz-juelich.de/jsc/llview
https://github.com/FZJ-JSC/LLview

Execution Environment (Programming Models)

- **ParaStation MPI**: MPI execution environment for large-scale systems w/ unique support for modular systems
  - Combines management system and MPI runtime
  - Support for PMIx5; scalable startup using RRCom (scalable daemon2daemon comm)
  - Supports efficient device-to-device communication via Atos BXI*

ParaStation MPI

https://par-tec.com/hpc/

- **Atos Open MPI**: well-established MPI implementation
  - Provides native Atos BXI* support
  - Optimizes collective operations using hierarchical communication patterns adapted to system topology (using Forth's HAN (inter-node) and xHC (intra-node) interfaces)

FORTH Institute of Computer Science

Open MPI

https://www.open-mpi.org/

* https://eviden.com/solutions/high-performance-computing/hpc-solutions/bullsequana-exascale-interconnect-bxi/

EUPEX

Execution Environment (Performance Tuning & Orchestration)

- **ACCO**: Tuning of Open MPI performance via environment variables
  - Explores the parameter space of several key parameters
  - Extended to support the tuning of AI workloads
  - Study the use of AI for ACCO tuning capabilities
- **BHCO**: Toolbox for accelerating hybrid MPI/OpenMP applications
  - Enables efficient MPI communication in OpenMP computing sections
  - Adapts number of OpenMP threads of MPI processes on a node according to computational load
- **StreamFlow**: workflow management for hybrid cloud/HPC infrastructures
  - Uses Common Workflow Language (CWL) & YAML/JSON descriptions
  - Supports local processes, containers, SSH nodes, Kubernetes (Helm) & batch systems (Slurm, PBS, …)
  - Improvements: Iterative workflows; high-throughput workflows; container support; SPECFEM3D+ integration with CAPIO library

Atos

StreamFlow
Towards Cloud-HPC Continuum

https://streamflow.di.unito.it/

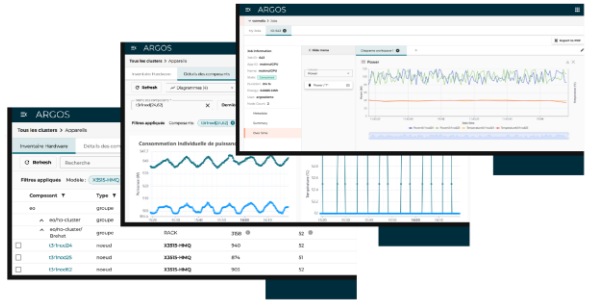
Execution Environment (Low-level Hardware Interfaces)

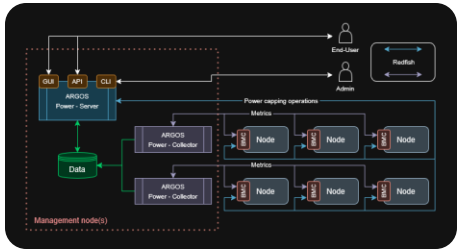
- **xmap** & **TeraHeap**: alleviate DRAM capacity limitations
  - Extends application heaps transparently on fast storage devices for managed (TeraHeap) and unmanaged (xmap) languages
  - Added support for asynchronous switching between huge and regular pages (Linux) ;
  - Support for memory reclamation for the slow, high-capacity memory tier
- PGAS Programming Model (**GASPI/GPI**) & GPUDirect RDMA extensions
  - Design and implement PGAS/MPI benchmark suite
  - Porting of Portals4 for GASPI standard – reference implementation with GPI-2
  - GPUDirect RDMA extensions for multiple, distributed GPUs w/ BXI
- Hardware Locality (**hwloc**): expose hardware topology and memory hierarchies to upper layers
  - Includes firmware (ACPI tables) and kernels (sysfs) information

https://www.open-mpi.org/projects/hwloc/

Performance and Energy Efficiency Tools (System Level & Node Level)

- Power Component of **ARGOS** (formerly Energy Optimizer (B)**EO**): monitor cluster infrastructure & jobs
  - OOB HW Monitoring: Power, energy, temperature: 1Hz time series
  - Manage constraints (power caps)
  - Per-job energy accounting
  - Working with WP3 on ecTrans & dyablo to characterize the impact of power capping and GPU frequency scaling
- Low-level components to access to prototype board/blade sensors (energy/power/temperature)
- Collect and forward sensor data via **ParaStation Modulo** and **PMIx**





Atos

E4 COMPUTER ENGINEERING

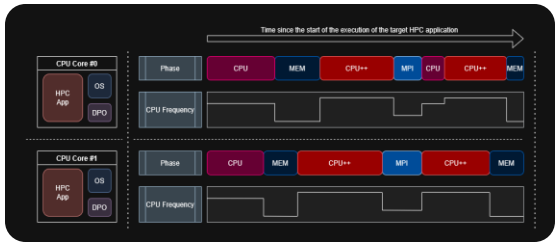
ParaStation MODULO
https://par-tec.com/hpc/

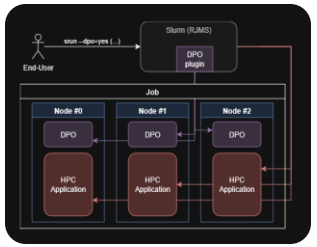
PMIx 10^18
https://pmix.github.io/

EUPEX

Performance and Energy Efficiency Tools (Application/Job Level)

- Dynamic Power Optimizer (formerly (B)**DPO**) of **ARGOS**: improves application energy efficiency
  - Detects and classifies application phases (memory, compute, communication, …)
  - Scales CPU voltage & frequency to achieve good efficiency
  - Working with WP3 on ecTrans & dyablo to characterize the impact of power capping and GPU frequency scaling
  - Integration with MERIC and demonstration with WP3





- **COUNTDOWN**: reduces power consumption during MPI wait times
  - Intercepts MPI communication calls & detects wait phases
  - Scales down CPU voltage & frequency accordingly

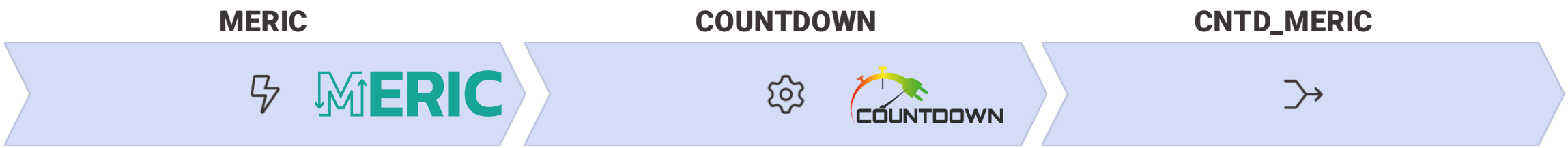
https://github.com/EEESlab/countdown

- **MERIC**: dynamically tunes CPU and accelerator settings for energy efficiency
  - Optimises core & uncore voltage & frequency, number of OpenMP threads
  - MERICext library providing access to energy measurement in a unified way (Intel/AMD RAPL, A64FX, OCC, Nvidia NVML, AMD ROCm, HDEEM, HWMON, and DiG)

https://code.it4i.cz/vys0053/meric

Performance and Energy Efficiency Tools (Application/Job Level)

- Leverage synergies: MERIC for HW interface, COUNTDOWN for the interface with the application
  - Real-time tracking of energy consumption on ARM-based power monitoring system, during MPI-based applications
  - Provides insights into total energy used, average power draw, MPI communication overheads, memory usage, and workload behavior
  - Reduces energy consumption with negligible overhead by slowing down the core's frequencies during the MPI communication phase



Fine-grained energy monitoring capabilities

Run-time optimisations for MPI applications

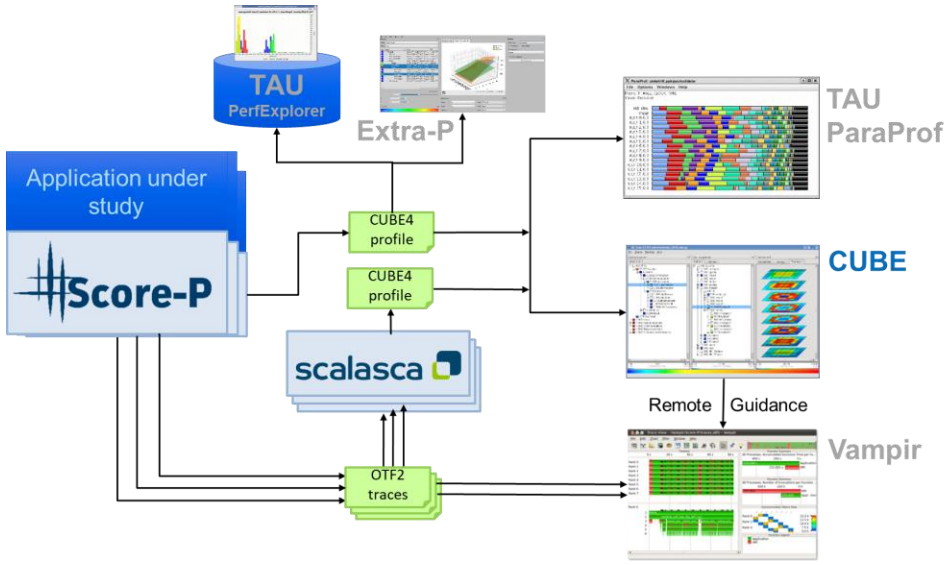
Unified framework for energy-performance analysis

Performance and Energy Efficiency Tools (Application/Job Level)

- **Score-P**: collects application performance data
  - Supports library interposition and code instrumentation
  - Includes system performance counter data
  - Creates profiling or event data (OTF2 format)
- **Scalasca**: analyses parallel event traces
  - Detects and classifies performance bottlenecks
  - Works with OTF 2 traces
- Prototypical integration with MERIC power measurement (tracing only)

Score-P

https://www.vi-hps.org/projects/score-p/



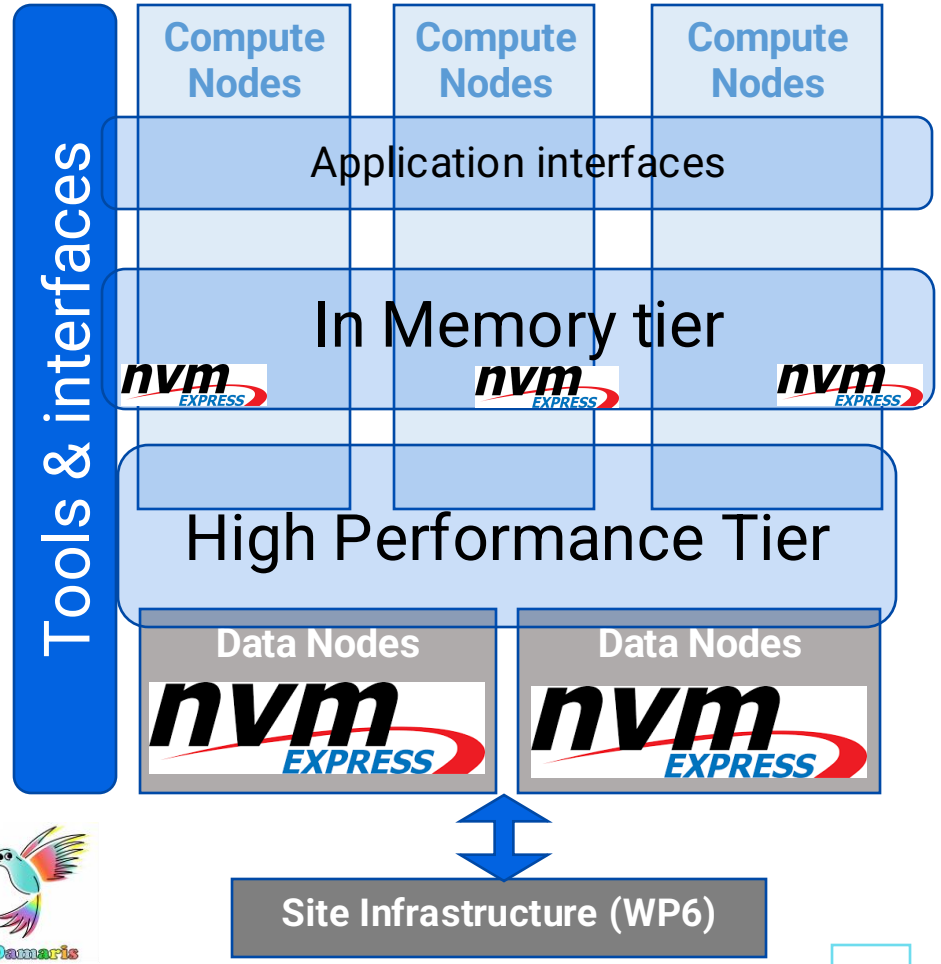
scalasca

https://www.scalasca.org/

EUPEX

Storage and I/O Stack (High-Capacity/High Performance/In-Memory Tiers)

- Integration with High-Capacity Tiers through Lustre GWs
  - Done for SDV hosted @ EVIDEN
  - OCEAN integration currently under way
- High-Performance Tier
  - Data nodes with NVMe as ephemeral storage resources
  - Integrated and optimized KV store (KVS) interfaces with ported IO-SEA stack
- In-Memory Tier
  - Data coupling in workflows via NVM and in-memory devices using a Cross-Application Programmable I/O layer (CAPIO)
  - Orchestrate transfers, non-intrusive in-situ/ in-transit data processing with DAMARIS
  - Initial porting and optimizing of Specfem3d on StreamFlow + CAPIO

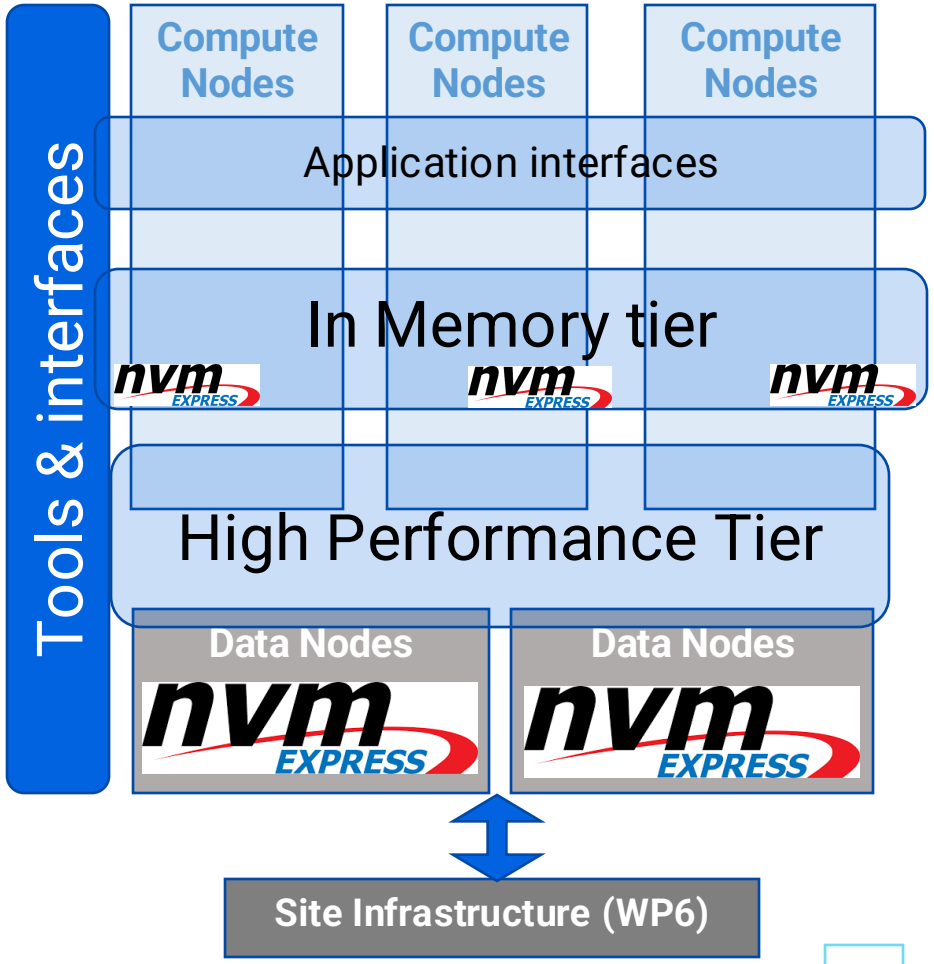


Damaris
https://project.inria.fr/damaris/

Storage and I/O Stack (Application-Specific Interfaces / Tools)

- Application Specific Interfaces
  - Tiered meteorology object store (ECMWF Fields Database https://github.com/ecmwf/fdb)
  - Preliminary runs, validation, and improved integration of FDB layer with KV store
- Tools & Interfaces
  - Integrated KV store with I/O workflow mgmt stack ported from IO-SEA
  - Early prototype integration of Modular and Automated Workflow Analysis for HPC (MAWA-HPC)

IO-SEA
https://iosea-project.eu/



EUPEX