International PostExascale

Workshop Series

InPEx 2025 workshop – April 15-17, Japan

InPEx working groups results and achievements since the Sitges (Spain) InPEx 2024 workshop

Co-design/Co-development of Community-driven set of Motifs-based proxy/mini-apps and software components streamlined with performance analysis tools and methodologies

Masaaki Kondo (Riken), Anshou Dubey (ANL), Jean-Pierre Vilotte (CNRS)

InPEx

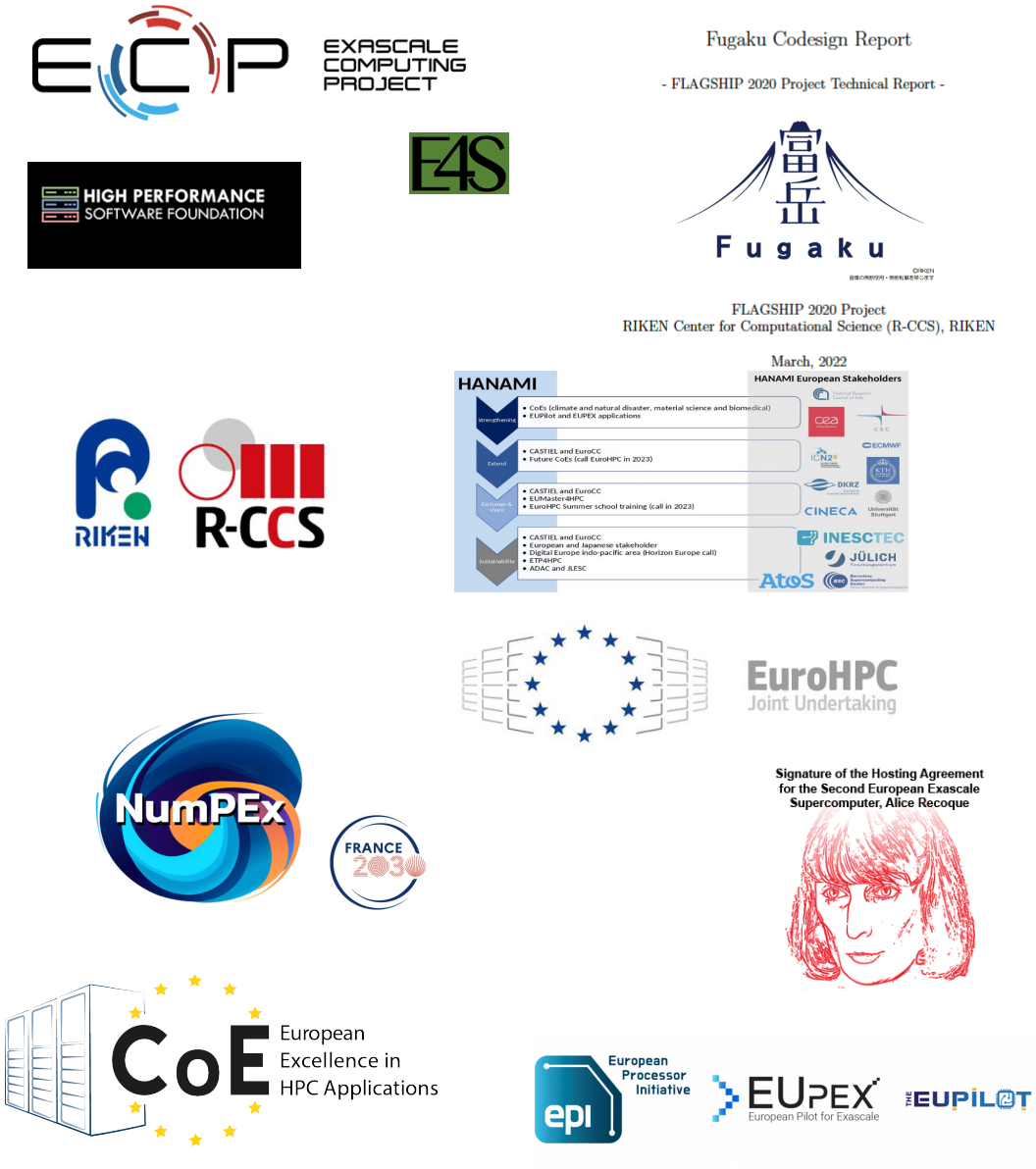
International Post-Exascale Initiative

Context: Productive (Post-) exascale systems

- US Post Exascale Computing Projects
- Fugaku & Fugaku nEXT co-design projects
- Euro-HPC JU Initiatives, ETP4HPC
- French NumPEx project & 2026 Alice Recoque Exascale System
- High performance Software Foundation (HPSF)

Challenges:

- Co-design/Co-development of Motifs-based logical collections of software components (libraries, frameworks, tools)
- Motif-based proxy/mini-apps as multi-layered framework with a high-level parametrisable abstraction interface for improving CSE software development methodologies, tuning and optimisation
- Increasing CSE application workload complexity: multi-physics/multi-scale applications, AI-coupled HPC/HPDA workflows,
- Profiling tools with shared methodologies and benchmarking specifications
- Improved sustainable application development methodologies & performance portability



Reminder of some break-out sessions outputs

- Co-design and co-development in rapidly changing HPC world
  - AI-driven HPC hardware and architecture evolution (high probability)
  - Possible customisation (networks, accelerators) provided by a limited number of vendors (low probability)
  - Specialised hardware and architecture (very low probability, narrow market)
- Increasing application workload and complexity
  - multi-physics/multi scale (AI enabled) coupling,
  - AI-driven and AI-coupled HPC/HPDA workflows: coupling programming models, move and map work and data layout to compute and memory targets
- Increasing rate of (AI-driven) change (GPUs/Accelerators, specialisation, heterogeneity) in exascale .
  - AI-driven low-precision accelerators: challenging algorithmic development (mixed precision)
  - Multiple memory spaces and complexity: memory access model (unified memory, gpu-direct, transfer)
  - Expansive data movements: minimise data transfers, efficient I/Os, ML-based in-situ data analytics
  - Programming them directly with traditional HPC languages (C++ / Fortran) is challenging.
  - Domain-specific programming models and frameworks (RAJA, Kokkos, PyTorch, TensorFlow, JAX) and performance portability
- From Motif-based proxy/mini-apps to multi-layered framework with a parametrisable high-level abstraction interface for tuning and optimisation
  - Combine dynamically different Motifs-based components (libraries, tools) and data structures/layout
  - Experiment different programming and execution models, hardware abstraction frameworks
  - Allow more efficient cross-layer optimisation, proper tuning and performance evaluation methodologies

Reminder of some break-out session outputs

- Lively discussions about proxy/mini-apps goals
  - Improved software development methodologies: experiment with different algorithms, libraries, data structures/layout, programming and execution models
  - making benchmarking easier

- Too many insufficiently documented and curated proxy/mini-apps
  - Takes effort to evaluate usefulness, collect information: needs clear standardised documentation, maintenance and more parametrisable inputs
  - Instruct apps developers to adopt standards, write document about how to setup/prepare input
  - Capture cross-cutting Motif-based execution (interaction, coupling) in application kernels and AI-coupled HPC/HPDA workflows
  - Selected out of test-suites, because they cannot be left out of sync with the code
  - Defined objectives of the proxy/mini-app: extracting the computation, communication and coupling behaviour of an app vs. making benchmarking easier
  - Share catalog of proxy/mini-apps that meet a minimum quality requirements, together with shared results information in standardised way across the InPEx community

Proposed Action plans

Set working groups to better define:

- How to move from Motif-based proxy/mini-apps to multi-layered frameworks with a parametrisable abstraction interface for more sustainable CSE software co-development methodologies , performance portability and performance evaluation
- How to build and share a catalog of well-documented proxy/mini-apps that meet a minimum quality requirements, and share results information in a standardised way
- How to coordinate and leverage existing initiatives in InPEx
- What resources to sustain it through cross-functional international collaborations (e.g. HPSF)
- How to foster adoption by the computational science and engineering community

Achievements of the working group

- Today no clear collective roadmap, initiatives and achievements ***across*** InPEx since June 2024
- Many individual partners and national initiatives and achievements (e.g. Riken, DOE, NumPEx, EuroHPC)
- Strong emphasis everywhere in increasing AI-driven and AI-coupled HPC/HPDA CSE application workflows
- Link to the new WGs on sharing AI-centric benchmarks and hybrid HPC/HPDA workflows, and on Generative AI for science
- Still need to set-up a momentum and collaborative tools to address the identified action plan

? More specific comments from: Masaaki Kondo and Jen Dome (R-CCS), Anshu Dubey (ANL)