Embracing modular supercomputers with hybrid workflows
The SPECFEM3D success story

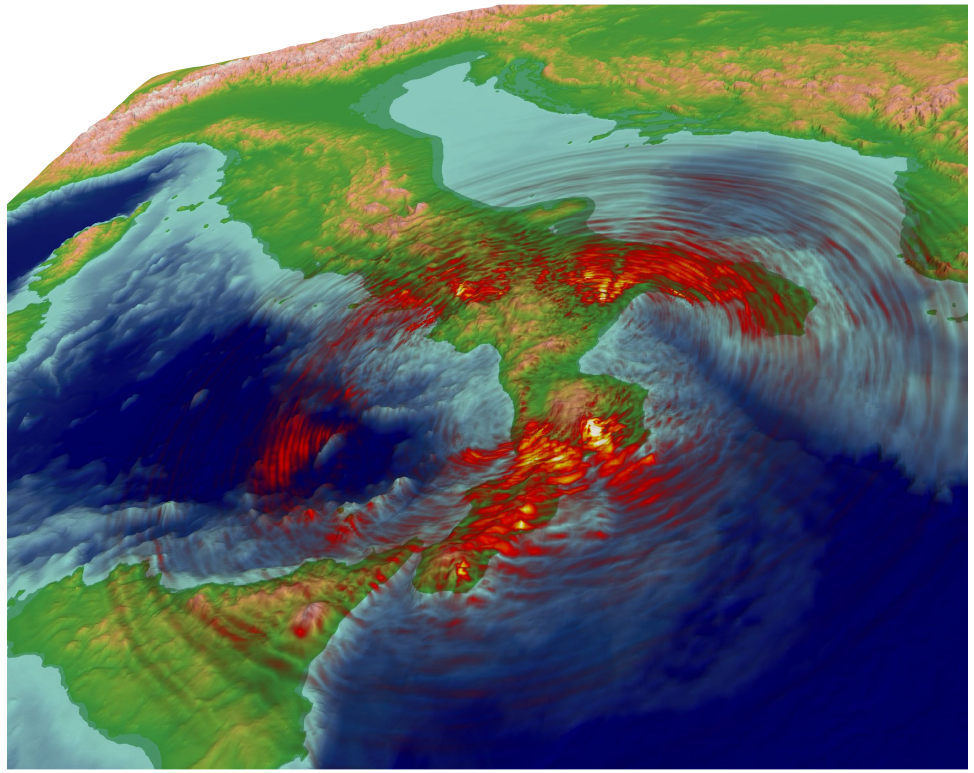
Iacopo Colonnelli, CINI-UniTO <iacopo.colonnelli@unito.it>
Emanuele Casarotti, INGV <emanuele.casarotti@ingv.it>
Piero Lanucara, CINECA <p.lanucara@cineca.it>

EUPEX

EuroHPC Joint Undertaking

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101033975. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Germany, Italy, Greece, United Kingdom, Czech Republic, Croatia.

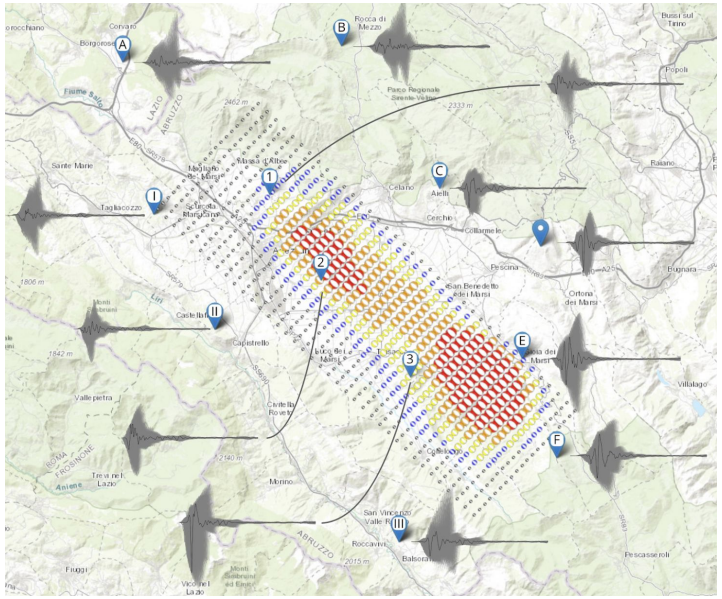
SEISMOLOGY workflow

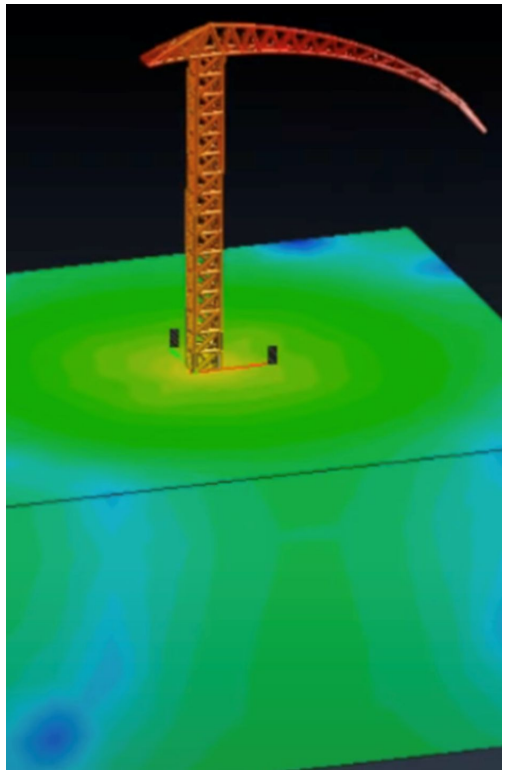


Integrated seismological/engineering **probabilistic** workflow

regional scale	local scale
ground motion acceleration scenarios	impact on strategic buildings

Coupling applications with high-performance parallel streams.

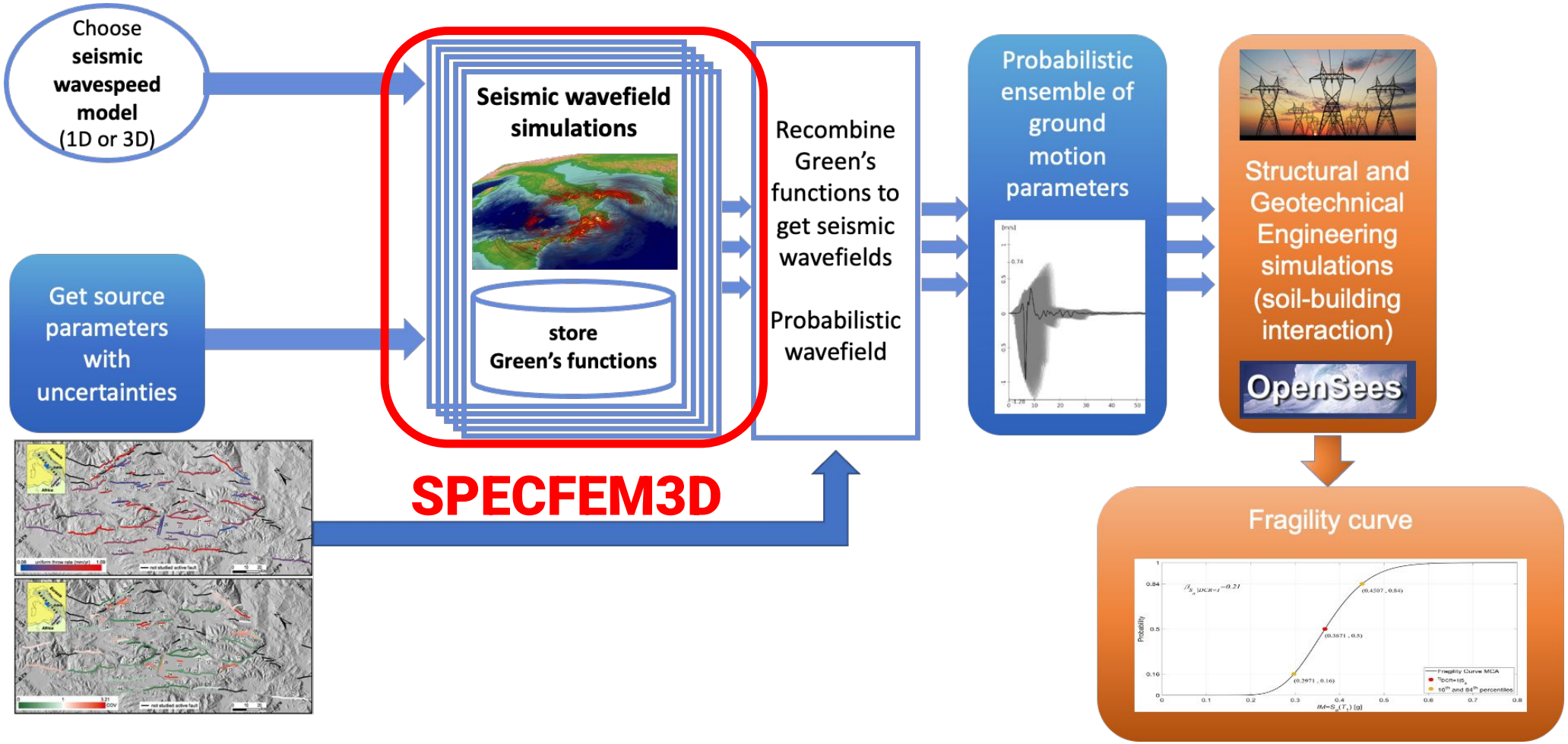




EUPEX

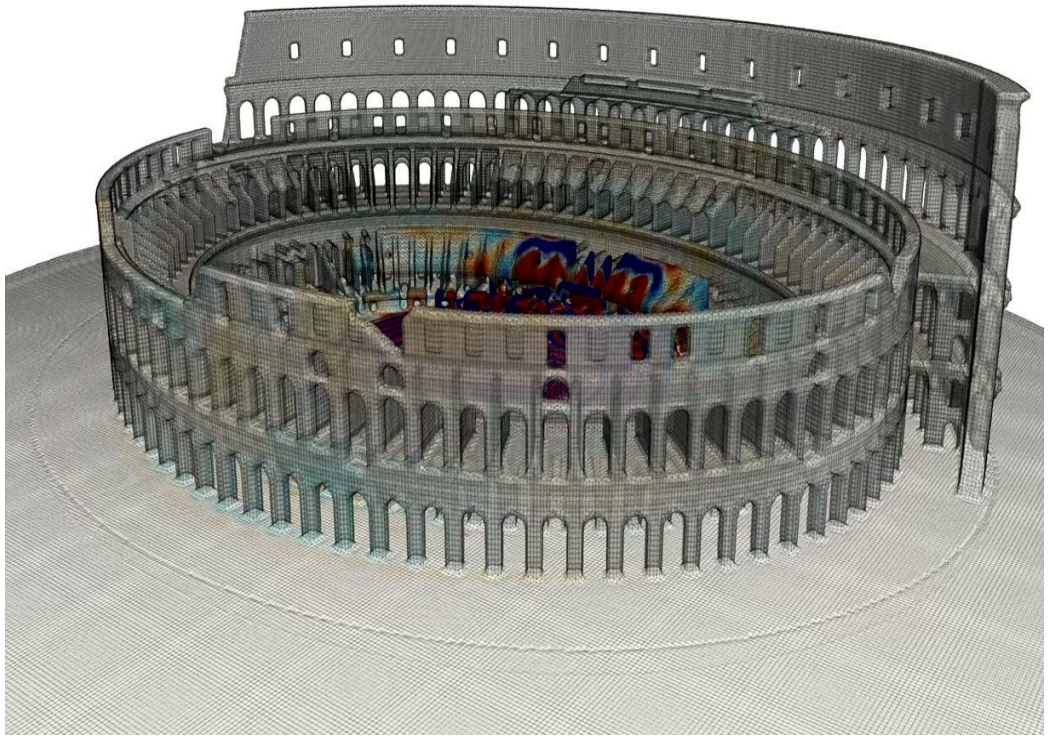
SEISMOLOGY WORKFLOW

From geology to probabilistic seismic engineering analysis



SPECFEM3D

- Exploits Spectral-Element Method's **accuracy** and **computational efficiency**
- Can model seismic waves propagating in sedimentary basins or any other regional geological model
- Simulations of **acoustic, elastic, coupled acoustic/elastic, poroelastic** or seismic wave propagation in any type of conforming mesh of **hexahedra** (structured or not)
- Implementation of **free or absorbing surfaces (PML), attenuation, anisotropy, structural heterogeneity** (topography, tomography), point or finite **kinematic/dynamic** sources
- Forward, **adjoint**, noise cross-correlation simulations
- **Finite frequency sensitivity kernels** for elastic/anelastic and isotropic/anisotropic media



EUPEX

ChEESE-2P CoE

17 **Beneficiary Organisations**
From 8 different countries

1 **Affiliated Entity**
IPGP (affiliated to CNRS)

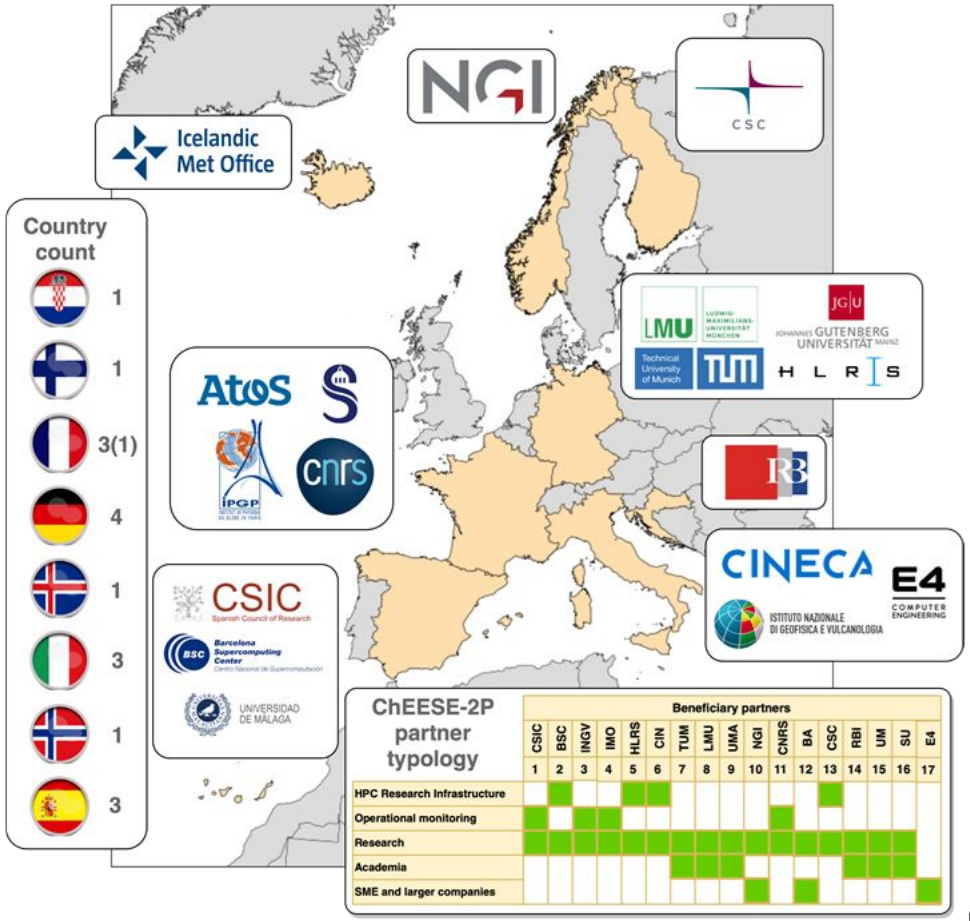
4 **HPC tier-0 Centers**
BSC, CIN, HLRS, CSC

3 **Private Companies**
BA, NGI, E4

4 **Operational Monitoring**
CSIC, INGV, IMO, CNRS

6 **Academia**
TUM, LMU, UMA, RBI, UM, SU

ChEESE-2P Consortium Composition



SPECFEM3D in ChEESE-2P

- SPECFEM3D targeted to ChEESE-2P Scientific Challenges
- Extreme computational requirements (billions of DoFs)
- ChEESE-2P objectives (CPU & GPU investigation):
  - Prepare for pre-Exascale and Exascale systems
  - Improve scalability, GPU efficiency, and portability
  - Support heterogeneous architectures (NVIDIA, AMD)

CPU: Mini-apps

- Available mini-apps (different physical models):
- Acoustic isotropic
- Elastic isotropic
- Elastic isotropic (already vectorized)
- Elastic isotropic with attenuation
- Elastic anisotropic
- Software environment:
  - GNU 11.5.0, Intel oneAPI 2024
  - OpenMPI 5.0.7 / Intel MPI 2021
  - compiler flags targeting AVX2 and Neoverse-V2 for x86 and ARM respectively

CPU: Characterization & Optimization

- Mini-apps are predominantly memory bound
- Initial lack of compiler auto-vectorization
- ChEESE-2P optimizations:
  - !$OMP SIMD directives
  - Loop splitting and reordering
- Moderate speed-up up to 1.6×
- Vectorization put higher pressure on memory bandwidth

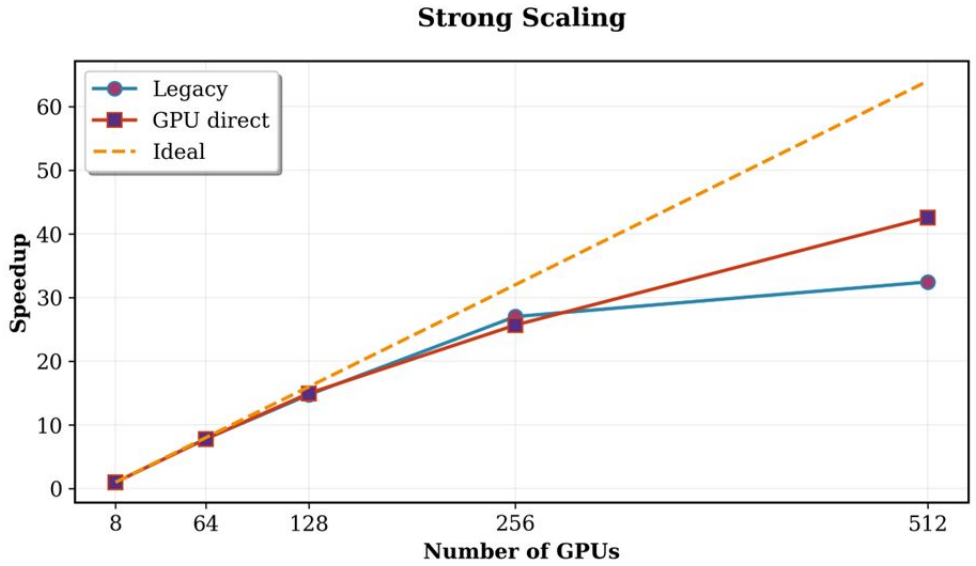
CPU: Architecture Insights & Conclusions

- HBM study on Intel Sapphire Rapids:
  - Speed-up from 4% up to 44%
- SVE analysis on ARM (NVIDIA Grace):
  - Optimized mini-apps fully vectorized

- Conclusions:
  - Mini-apps enable performance portability
  - Vectorization is crucial for performances on ARM CPUs
  - HBM benefits memory-bound kernels

GPU: Code Modernization and Scalability

- Use of GPU-Aware MPI (OpenMPI 4.1.6)
- Direct GPU-to-GPU communication
- Removal of host-mediated data transfers
- Strong scaling up to 512 GPUs
- Clear performance gains over legacy MPI



GPU: Methodology and Tools

- Performance analysis with Score-P and Scalasca
- POP efficiency model (GPU-focused)
- Good GPU utilization and load balance
- Mini-app strategy:
  - Extraction of core SPECFEM3D kernels
  - Rapid testing and tuning on new architectures

	Problem size	1024x1024	1024x1024	1024x1024	1024x1024	1024x1024	1024x1024
	MPI|GPU ranks	**8**	**32**	**64**	**128**	**256**	**512**
v1	Wall time [s]	**2001.806**		**255.948**	**135.478**	**73.846**	**61.400**
GPU	Global scaling efficiency	0.995		0.973	0.919	0.843	0.507
	- Computation time scaling	*1.000*		0.987	0.972	0.950	0.937
	- Parallel efficiency	0.995		0.986	0.945	0.887	0.541
	- - Load balance efficiency	1.000		0.998	0.996	0.994	0.989
	- - Orchestration efficiency	0.995		0.988	0.948	0.892	0.547
v2	Wall time [s]	**1996.055**	**505.209**	**259.019**	**133.395**	**77.756**	**46.871**
GPU	Global scaling efficiency	0.997	0.985	0.960	0.932	0.800	0.664
	- Computation time scaling	*1.000*	0.997	0.989	0.976	0.957	0.952
	- Parallel efficiency	0.997	0.988	0.971	0.956	0.836	0.697
	- - Load balance efficiency	1.000	0.999	0.998	0.991	0.990	0.992
	- - Orchestration efficiency	0.997	0.989	0.973	0.964	0.845	0.703

EUPEX

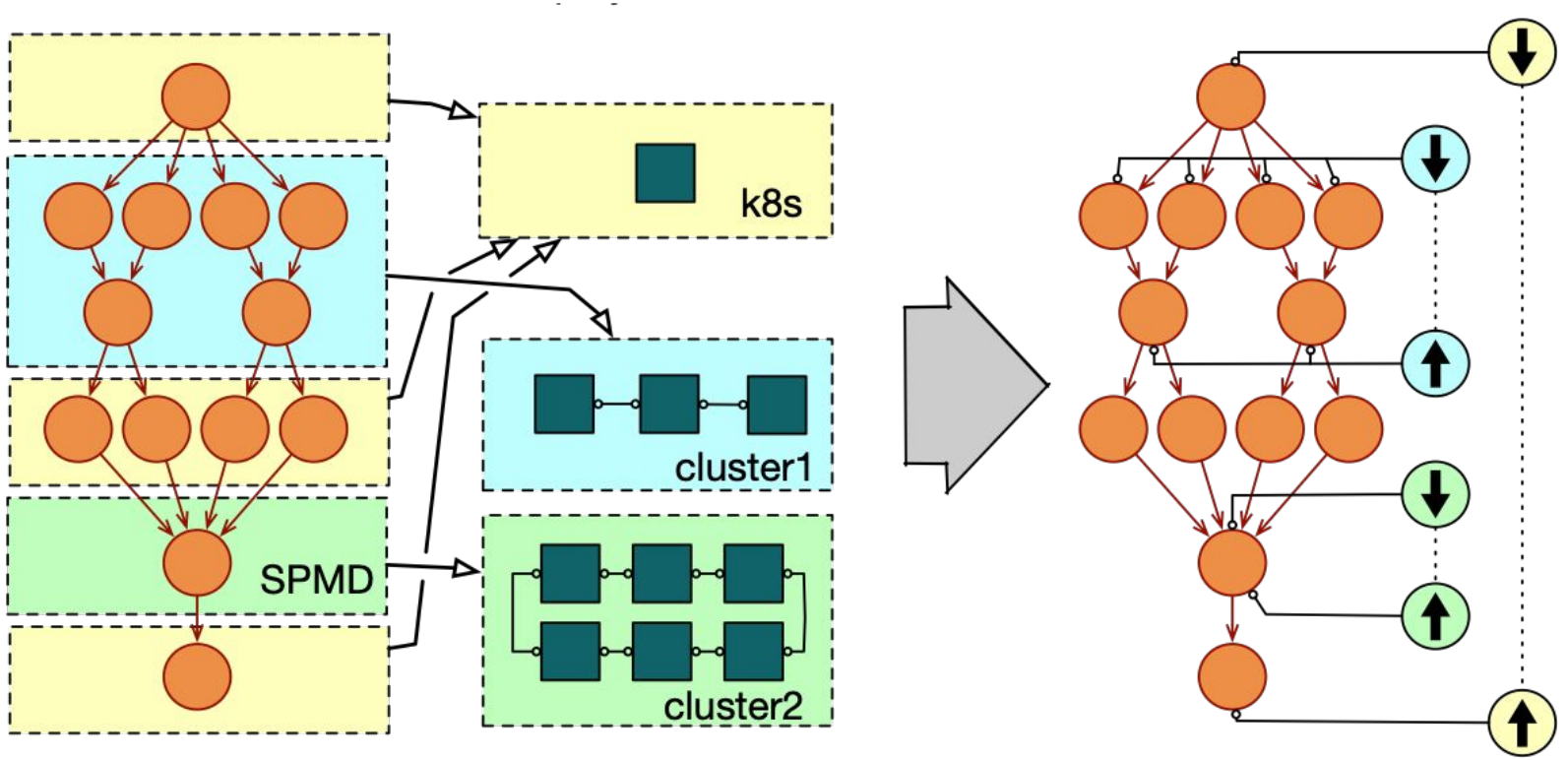
GPU: Performance Portability

- Improved scalability and efficiency on GPU systems
- Demonstrated performance portability across:
  - NVIDIA A100
  - NVIDIA GH200
  - AMD MI250X
- SPECFEM3D is now portable, maintainable, and Exascale-ready

	Leonardo (Custom A100)		Thea (GH200)		Adastra (MI250X)		PP metric
Version	Time Loop (s)	App. Eff.	Time Loop (s)	App. Eff.	Time Loop (s)	App. Eff.	
CUDA	8.56	100%	3.85	100%	N.A.	N.A.	0
HIP	9.82	87%	4.41	87%	4.5	100%	0.91

Hybrid scientific workflows and StreamFlow

A **hybrid workflow** is a workflow whose steps can span **multiple**, **heterogeneous**, and **independent** computing infrastructures

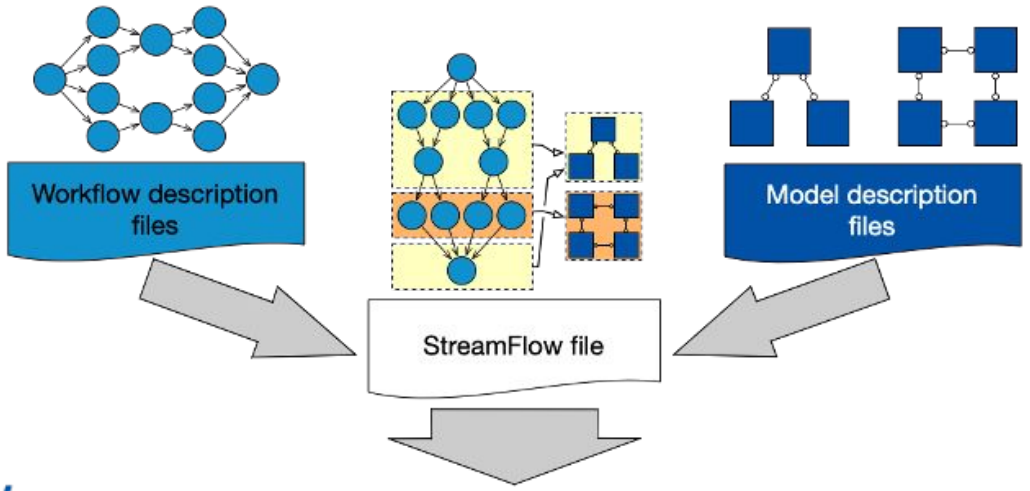


Hybrid scientific workflows and StreamFlow

COMMON WORKFLOW LANGUAGE

jupyter

StreamFlow
Towards Cloud-HPC Continuum



docker

flux

slurm
workload manager

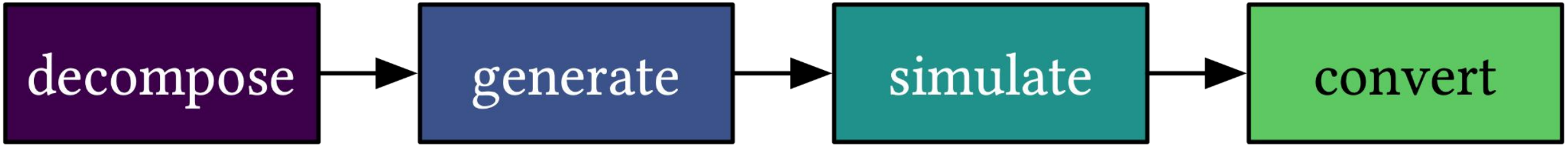
HyperQueue

IBM Spectrum LSF

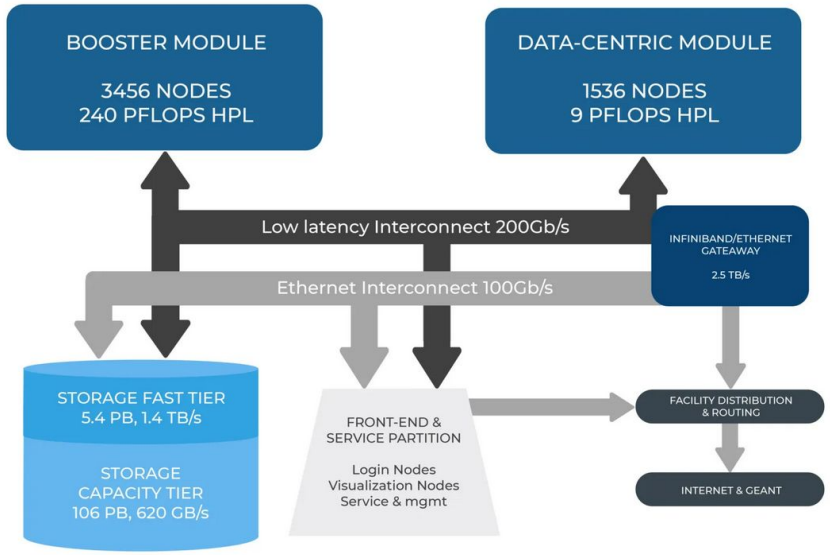
I. Colonnelli, B. Cantalupo, I. Merelli and M. Aldinucci, "StreamFlow: cross-breeding cloud with HPC," in *IEEE Transactions on Emerging Topics in Computing*, vol. 9, iss. 4, p. 1723-1737, 2021. doi: 10.1109/TETC.2020.3019202

EUPEX

The SPECFEM3D workflow: single execution



- Written in **Common Workflow Language (CWL)** for portability
- Orchestrated with **StreamFlow** (one of the EUPEX WP5 technologies) for scalability
- Executed on the **CINECA@Leonardo** modular HPC facility for performance
- Publicly available on **GitHub** for reproducibility and FAIRness



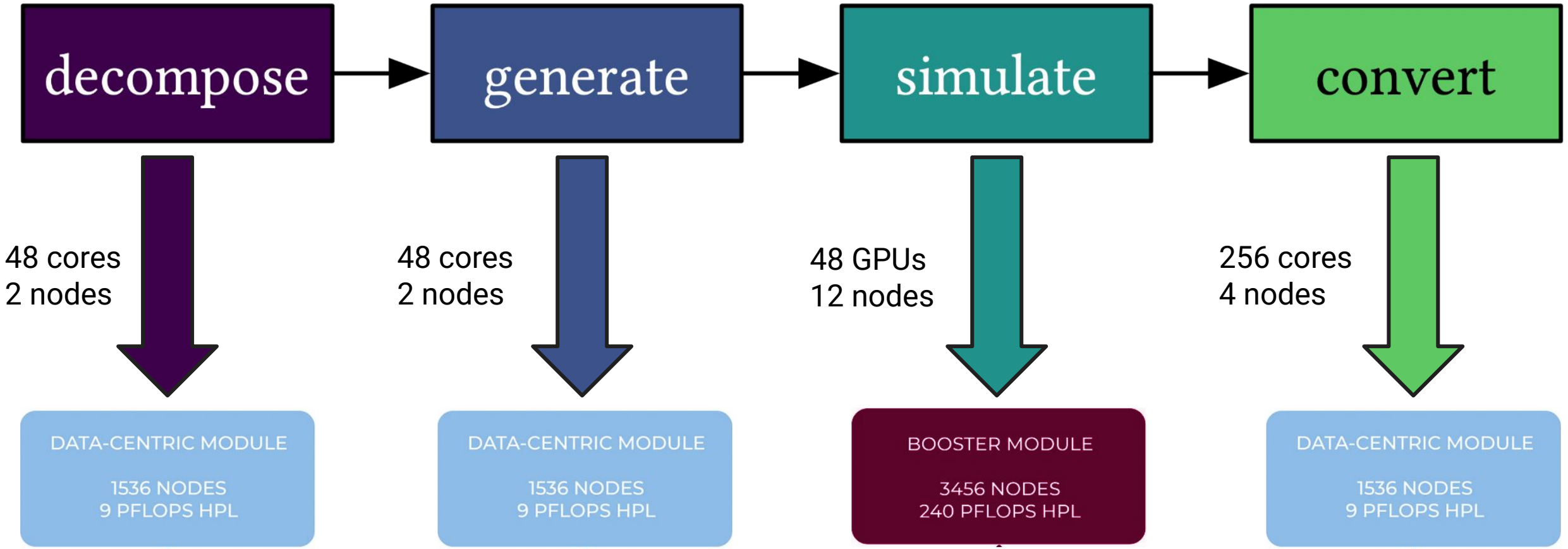
LEONARDO
CINECA

Main features

- 1536 CPU-based nodes (172032 cores)
- 3456 GPU-based nodes (13824 GPUs, 110592 cores)
- 155 Racks (16 CPU, 116 GPU, 12 I/O, 1 System)
- Power Requirements
  - HPL: ~ 8.0 MW
  - Operational: ~ 6.0 MW

EUPEX

The SPECFEM3D workflow: single execution



The SPECFEM3D workflow: single execution

DATA-CENTRIC MODULE

1536 NODES
9 PFLOPS HPL

BOOSTER MODULE

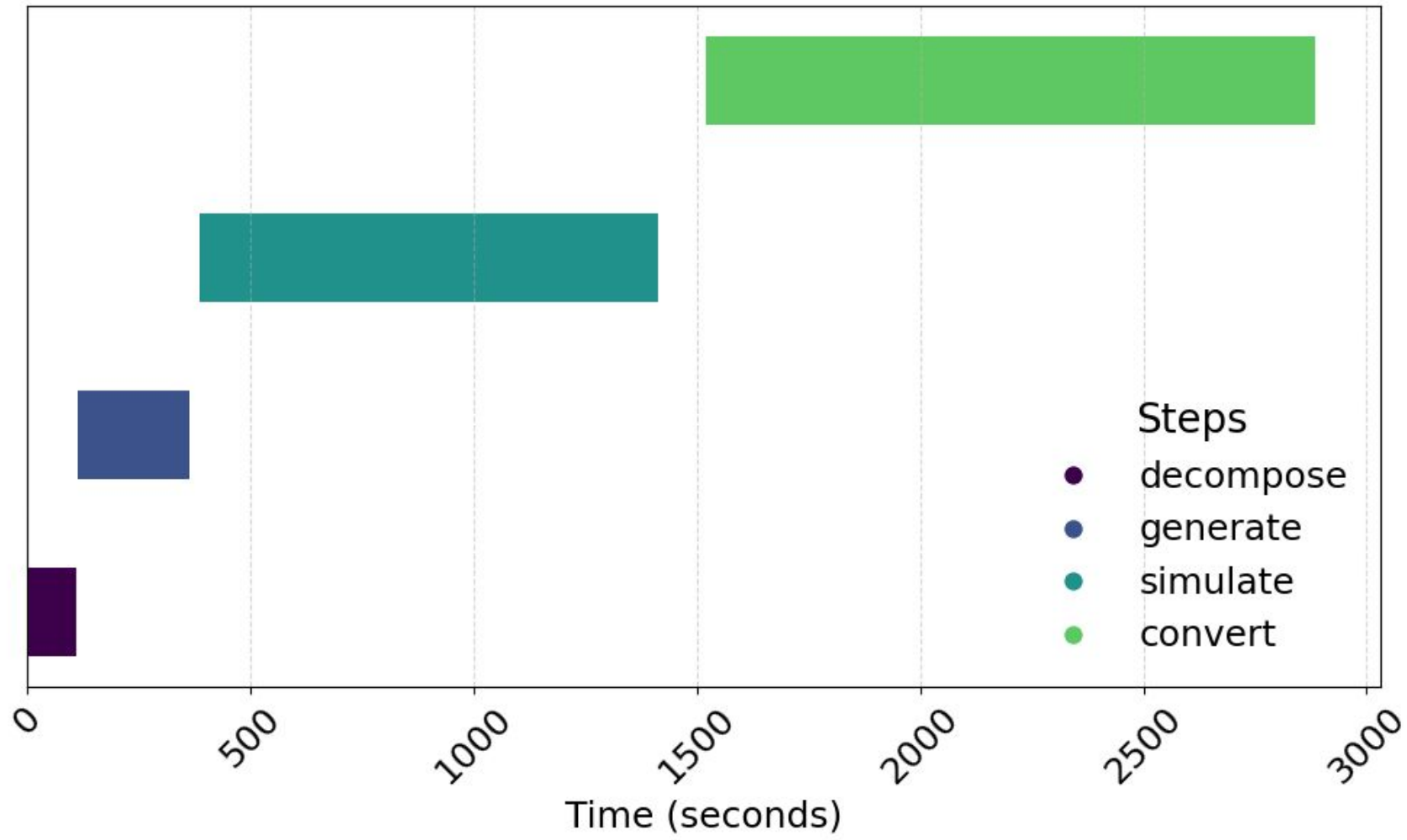
3456 NODES
240 PFLOPS HPL

DATA-CENTRIC MODULE

1536 NODES
9 PFLOPS HPL

DATA-CENTRIC MODULE

1536 NODES
9 PFLOPS HPL



EUPEX

The SPECFEM3D workflow: single execution

DATA-CENTRIC MODULE

1536 NODES
9 PFLOPS HPL

BOOSTER MODULE

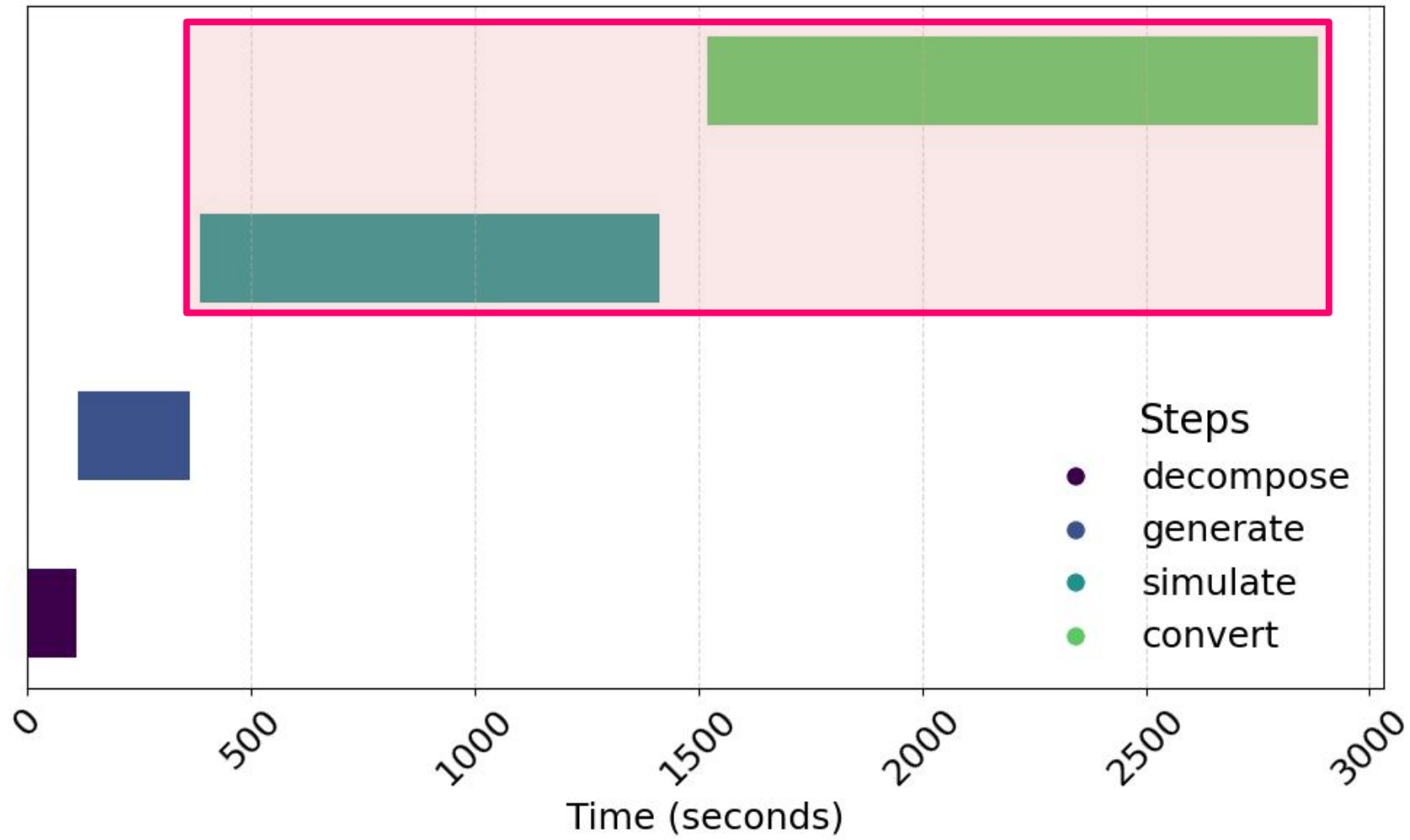
3456 NODES
240 PFLOPS HPL

DATA-CENTRIC MODULE

1536 NODES
9 PFLOPS HPL

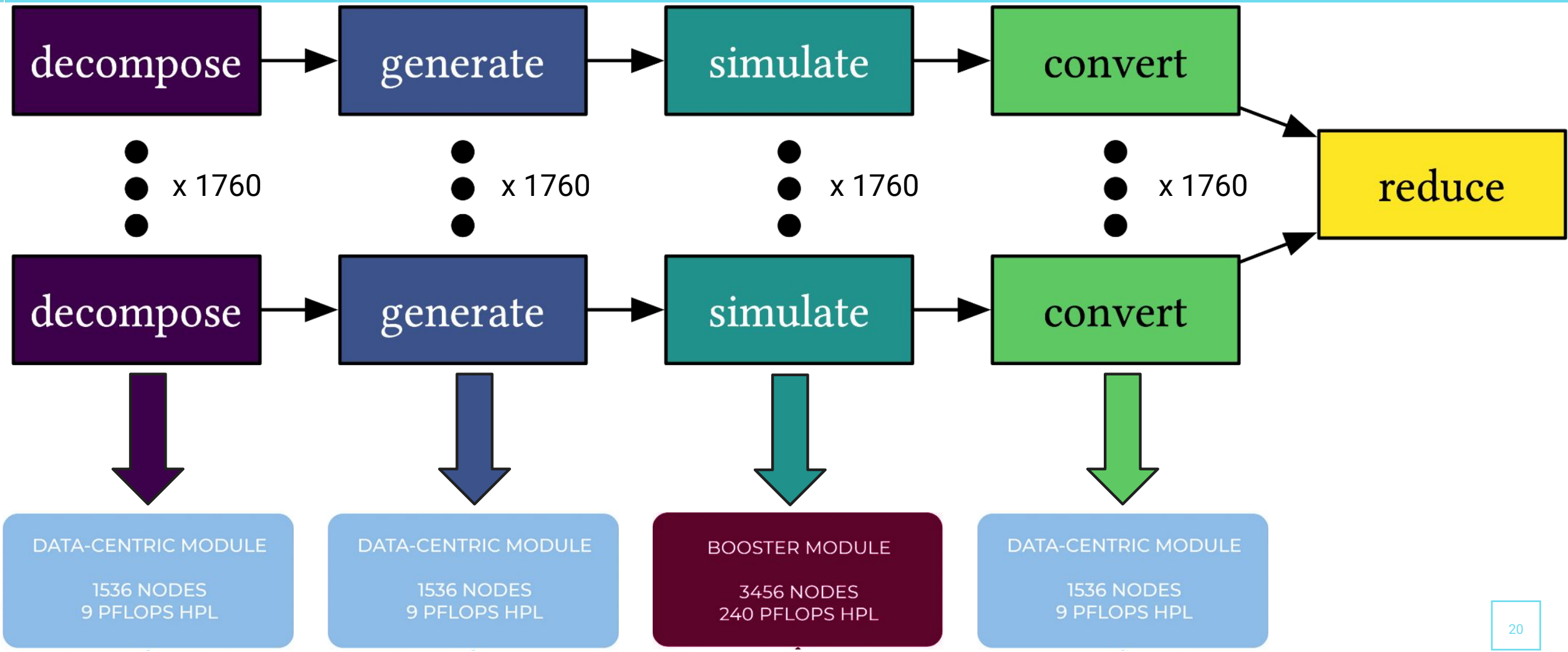
DATA-CENTRIC MODULE

1536 NODES
9 PFLOPS HPL



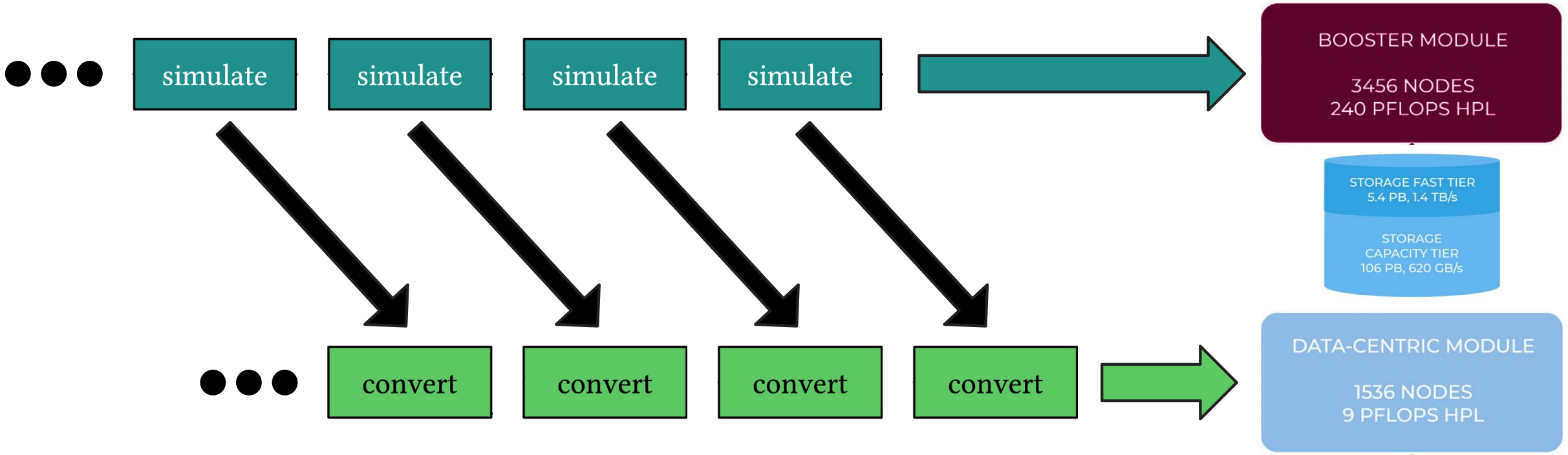
EUPEX

The SPECFEM3D workflow: multiple execution



The SPECFEM3D workflow: multiple execution

Avoid concurrent access to the same machine, which slows down scheduling



The SPECFEM3D workflow: multiple execution

DATA-CENTRIC MODULE

1536 NODES
9 PFLOPS HPL

BOOSTER MODULE

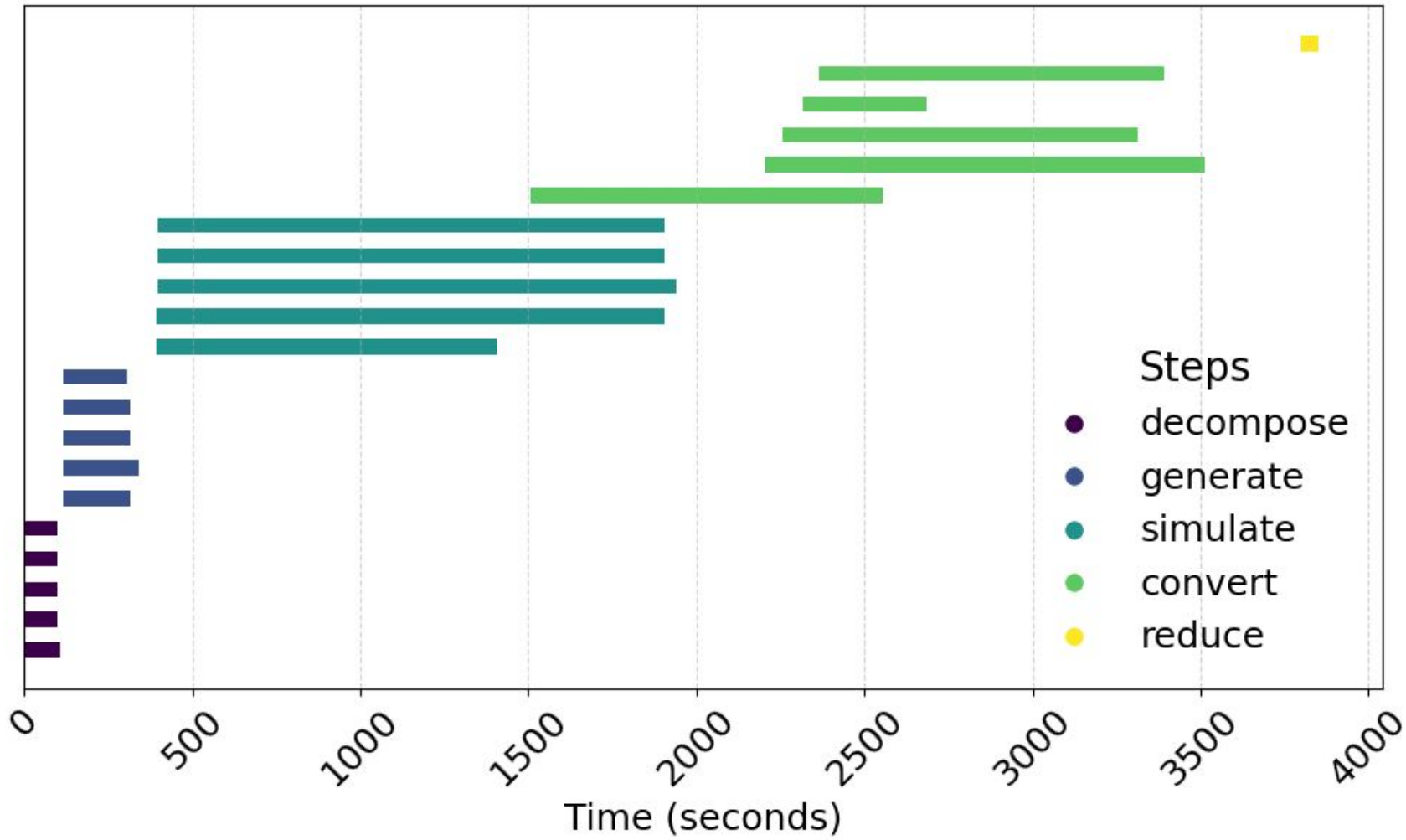
3456 NODES
240 PFLOPS HPL

DATA-CENTRIC MODULE

1536 NODES
9 PFLOPS HPL

DATA-CENTRIC MODULE

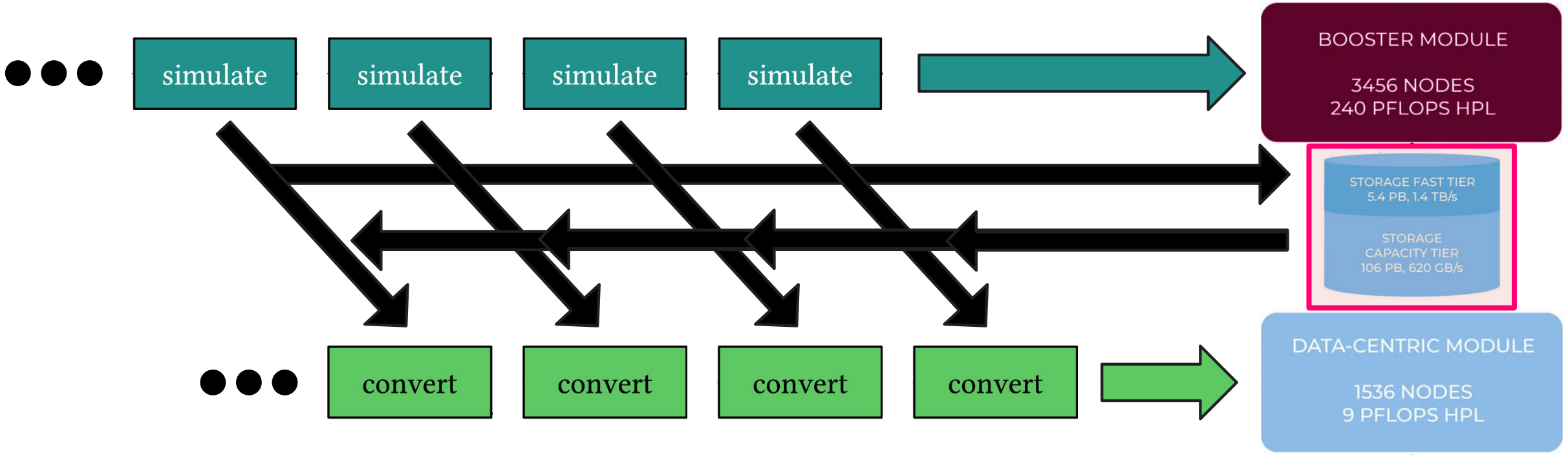
1536 NODES
9 PFLOPS HPL



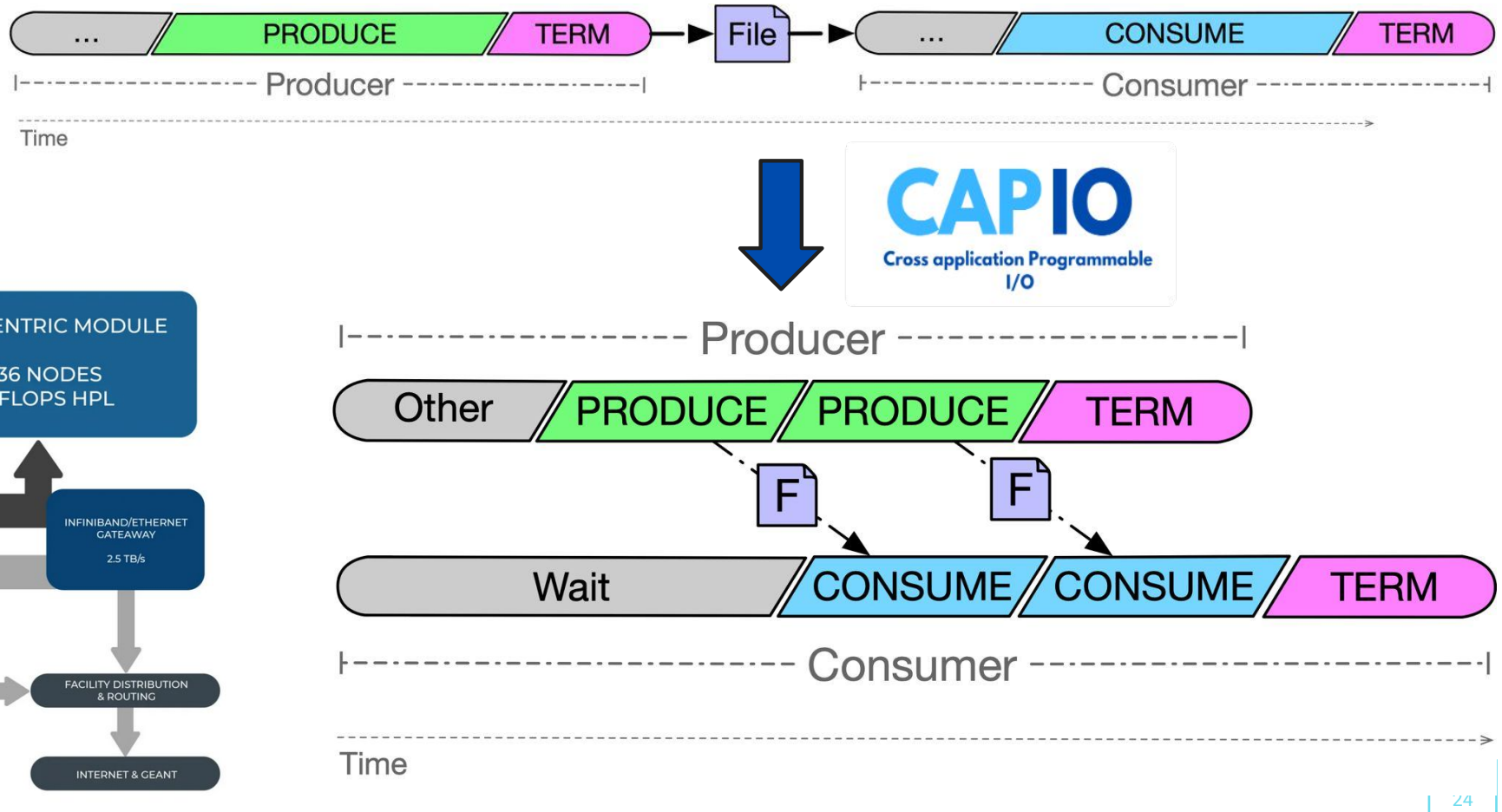
EUPEX

The SPECFEM3D workflow: multiple execution

Still, we have **concurrent access to the shared file system** (LUSTRE) and **no overlap of computation and I/O**



Transparent in-situ workflows and CAPIO

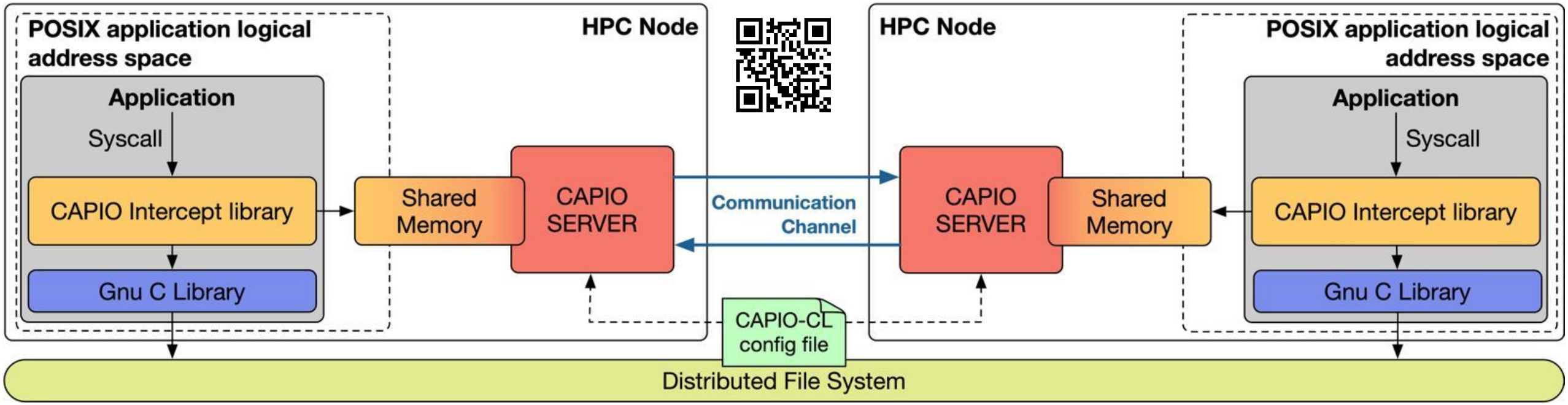


Transparent in-situ workflows and CAPIO

simulate

CAPIO
Cross application Programmable I/O

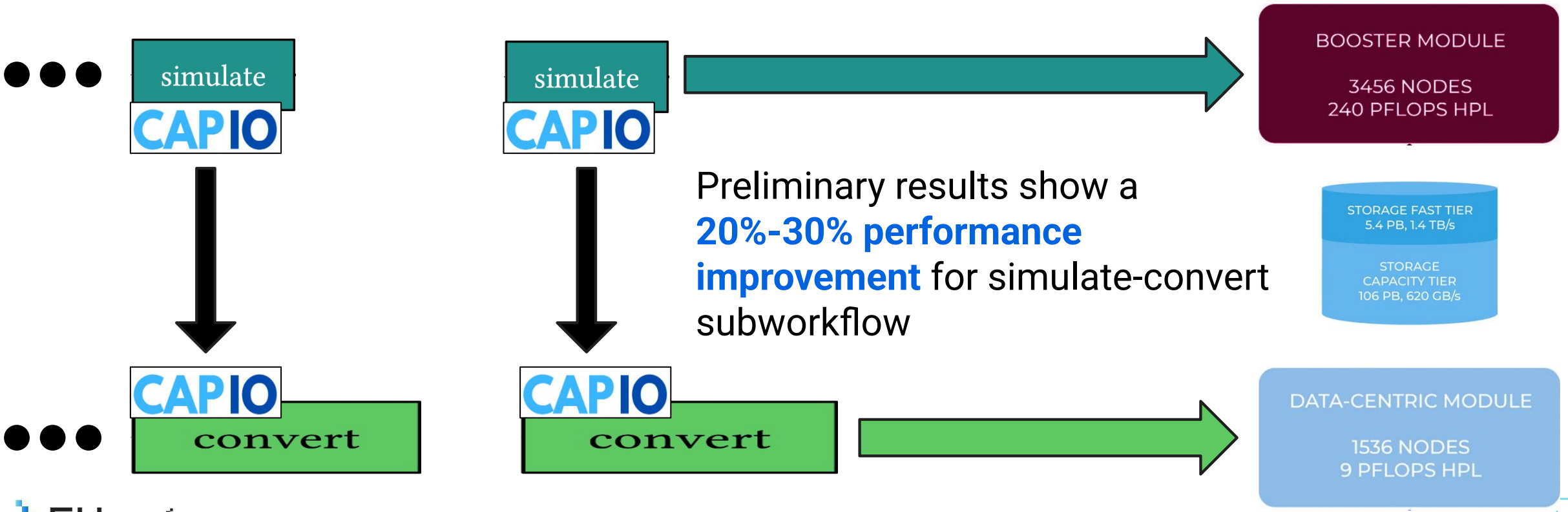
convert



EUPEX

The SPECFEM3D workflow: streaming execution

Co-schedule steps and **reduce I/O overhead** through transparent streaming



Preliminary results show a **20%-30% performance improvement** for simulate-convert subworkflow

EUPEX

Conclusion

- SPECFEM3D is **portable**, **maintainable**, and **Exascale ready** thanks to the support for CPU vectorisation and GPU-accelerated machines
- The hybrid workflow approach scales **beyond the single HPC facility**, and is valid for **any modular supercomputing architecture** (including European HW)
- **Transparent I/O optimisations** overcome the limitations of file-based workflows and shared parallel file systems without touching the codebase (and HW compatibility)
- These three independent optimisation directions enable **efficient urgent computing** at the scale of the whole EuroHPC Federation

Thank you for your attention…

https://eupex.eu

https://www.linkedin.com/company/eupex-pilot

https://bsky.app/profile/eupex-project.bsky.social

…and stay tuned!

EUPEX