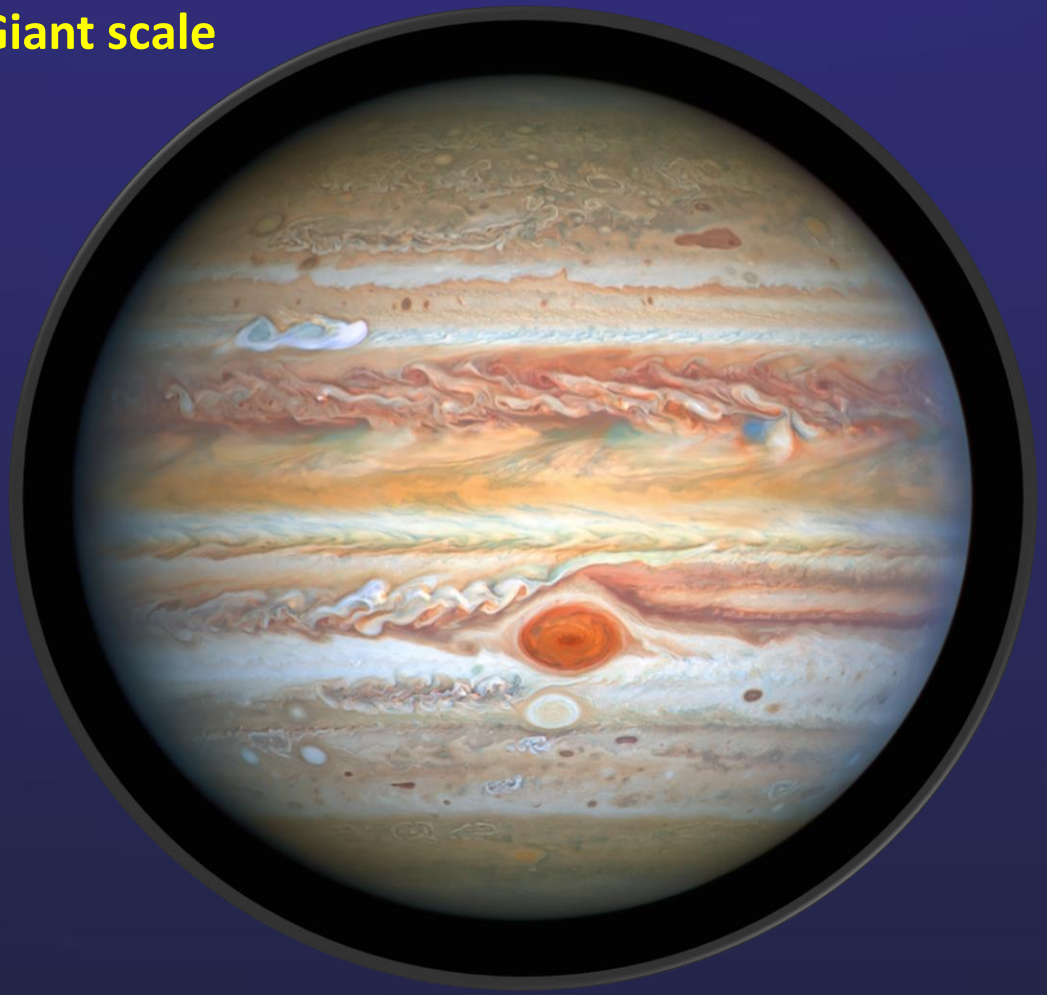
EuroHPC DEVELOPMENT PROJECT(S)

Scalasca / Score-P (exa-)scalable parallel performance tools

2024/10/22 I BRIAN WYLIE [b.wylie@fz-juelich.de]

EuroHPC Joint Undertaking
HORIZON-EUROHPC-JU-2023-COE
Grant Agreement No 101143931
JÜLICH Forschungszentrum
JÜLICH SUPERCOMPUTING CENTRE

Giant scale



10^{18}

=

1000000
000000
000000
000000

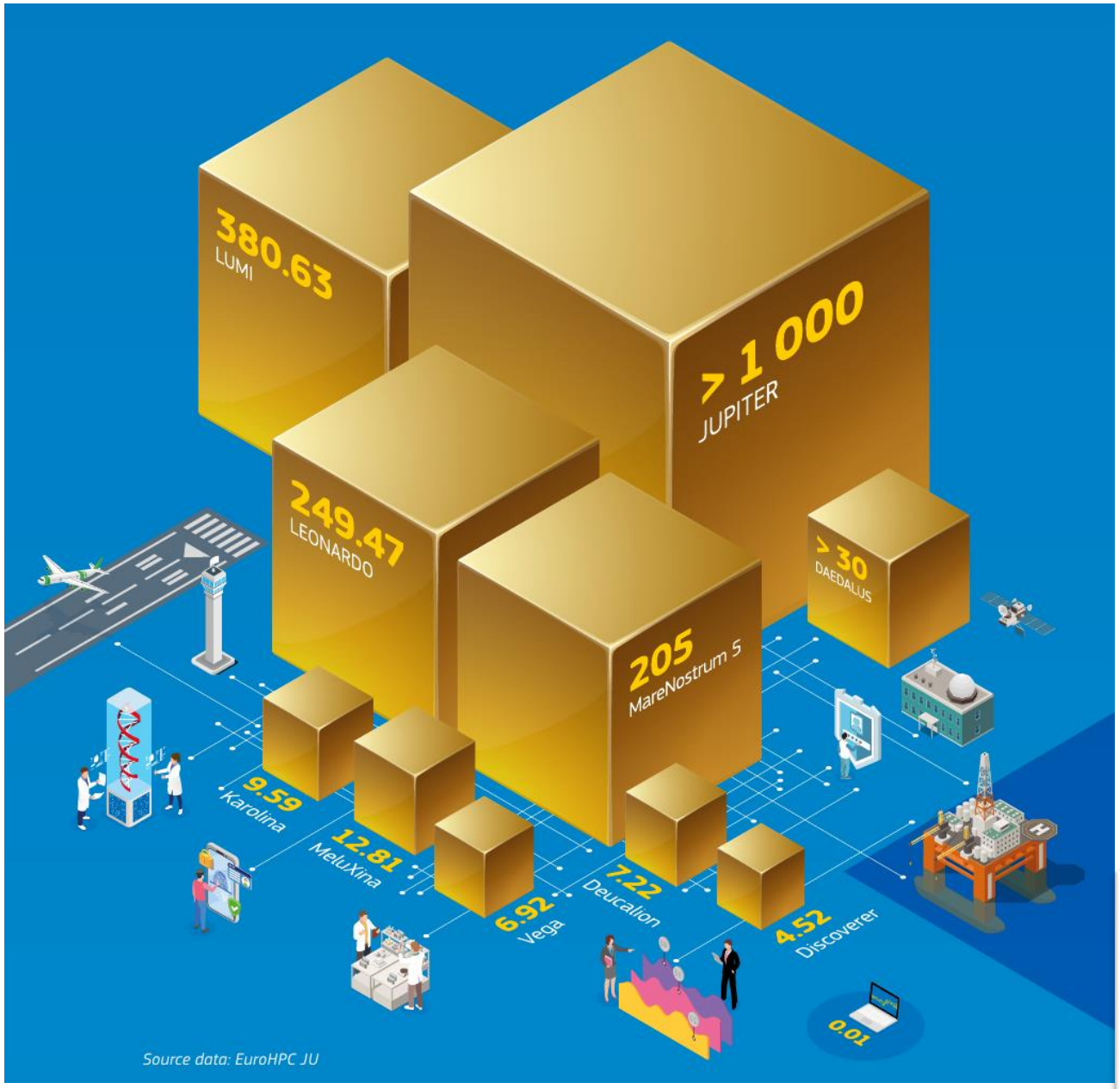
fz-juelich.de/jupiter | #exa_jupiter

Lightning speed

Disclaimer: content provided by AI.

EuroHPC Joint Undertaking
Federal Ministry of Education and Research
Ministry of Culture and Science of the State of North Rhine-Westphalia
GCS Gauss Centre for Supercomputing

EUROHPC COMPUTERS & CENTRES OF EXCELLENCE



bioexcel
Centre of Excellence for Computational Biomolecular Research

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

ChEESE

EXCELLERAT

CEEC
Centre of Excellence in Exascale CFD

MAX

Performance Optimisation and Productivity

MICROCARD

dealII-X

SPACE

E-CoE

PLASMA PEPSC

HiDALGO

Multiscale

CoE European Excellence in HPC Applications

CASTIEL 2

JÜLICH Forschungszentrum

OUTLINE

- Jülich Supercomputing Centre open-source scalable parallel performance tools
  - ***Scalasca***: scalable performance analysis of large-scale parallel applications
  - ***Score-P***: community-developed instrumentation & measurement infrastructure
- Performance Optimisation & Productivity Centre of Excellence (POP CoE)
  - Assessments for HPC application domain CoEs
    - *SPECFEM3D* on Leonardo-B [ChEESE CoE]
    - *Tandem* on LUMI-C [ChEESE CoE]
    - *ESPResSo* on Vega-C [MultiXscale CoE]
    - *ecTrans_dwarf* on Karolina-G [ESiWACE CoE]
  - Training: CASTIEL2/EuroCC Training Sprint on Karolina (CPUs & GPUs)

scalasca

DOI 10.5281/zenodo.7440854

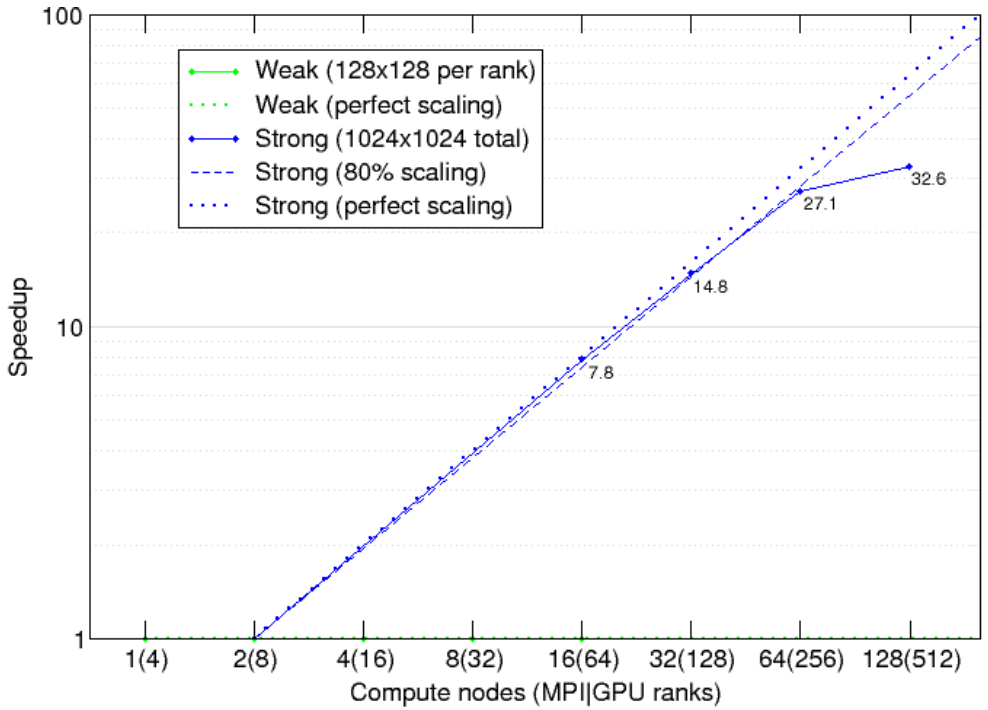
Developed to support scalable performance analysis of large-scale parallel applications

- available under open-source license from www.**scalasca.org**
- offers flexible runtime summarization/profiling and event tracing
- based on ***Score-P*** instrumentation & measurement infrastructure and ***CUBE*** analysis report utilities & explorer GUI
- MPI + OpenMP, extended to support Pthreads and other threading paradigms plus accelerator kernel offload with OpenACC, OpenCL, CUDA, HIP, etc.
- support for large-scale HPC computer systems and clusters
  - used with up to 1.75 M threads (or 1.28 M processes) on JUQUEEN BG/Q

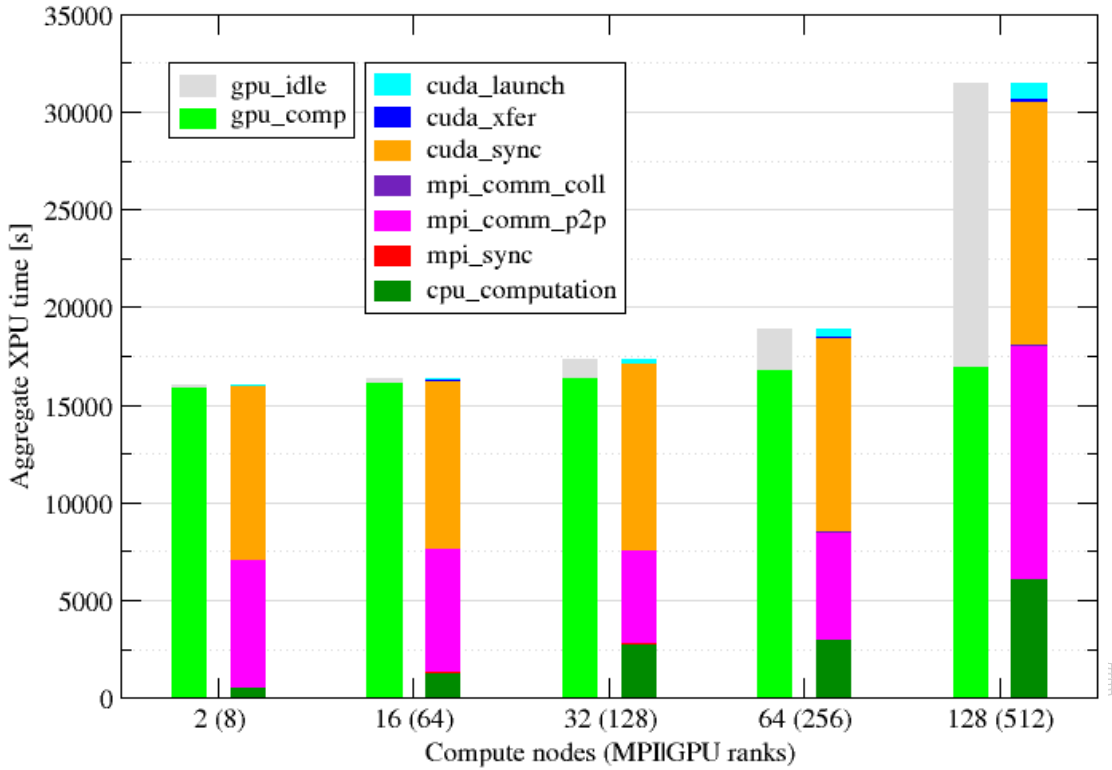
scalasca & Score-P

- Centre of Excellence in HPC Applications' Performance Optimisation & Productivity
- JSC 'flagship' codes to be deployed on (all) EuroHPC supercomputer systems
  - M12 goal of up-to-date deployments on 4 systems (both CPU & GPU partitions)
    - on target for publicly-installed modules to be accessible by 2024/12
  - ***Score-P***/8.4 available on Karolina & Leonardo*
    - older installations available on MeluXina, MN5, Vega* (and LUMI-C)
  - ***Scalasca***/2.6.1 available on Karolina, Leonardo*, MeluXina, Vega* (and MN5-GPP)
- Engaged with CASTIEL2 CI/CD task
  - following progress of EESSI & GitLab/Jacamar prototype solutions
  - build & install recipes in IT4I EasyConfigs repository

SPECFEM3D@Leo-B strong scaling



Problem size	1024x1024	1024x1024	1024x1024	1024x1024	1024x1024
MPI|GPU ranks	**8**	**64**	**128**	**256**	**512**
Wall time [s]	**2001.806**	**255.948**	**135.478**	**73.846**	**61.400**
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



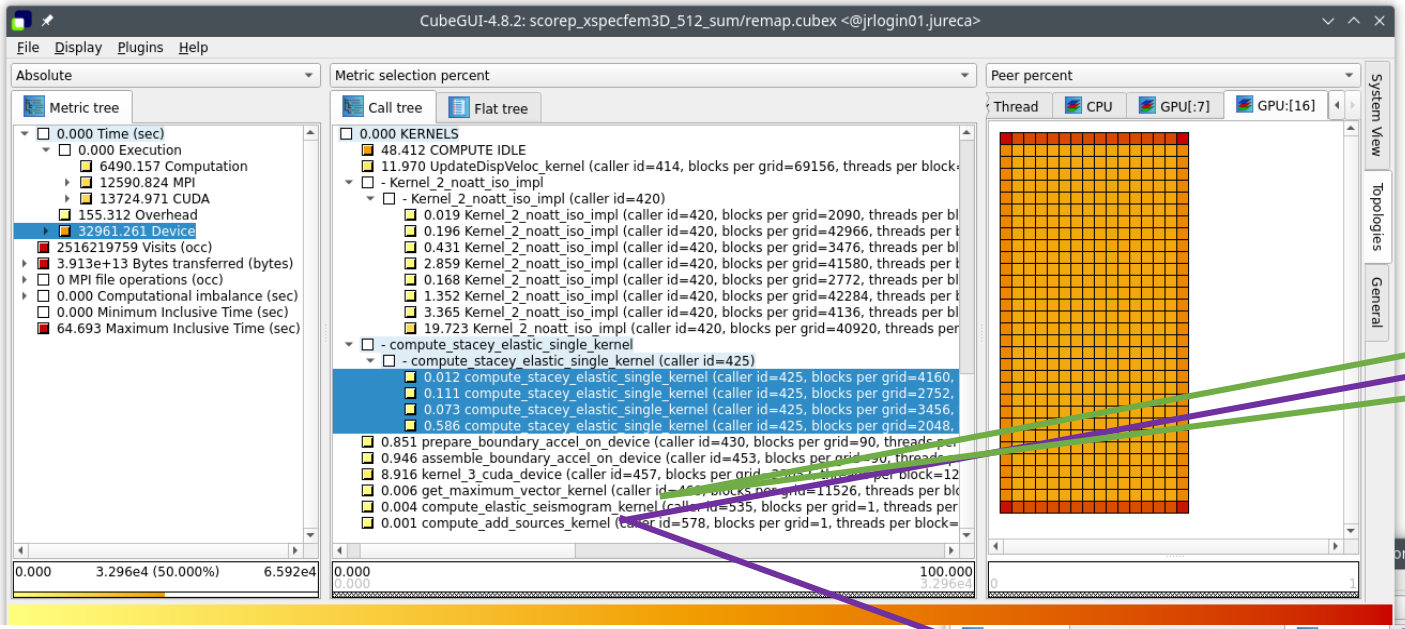
ChEESE

POP3_AR_002

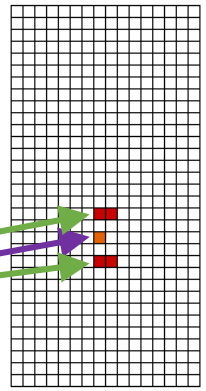
- Fortran90 parallelised with MPI+CUDA (1 rank/GPU)
- *iterate_time* (solver) chosen as Focus of Analysis
- Good strong scaling up to 256 GPUs
- With 512 GPUs no longer able to sufficiently overlap MPI communication with CUDA kernels

DOI 10.5281/zenodo.13643996

SPECFEM3D@Leo-B (512 GPUs)

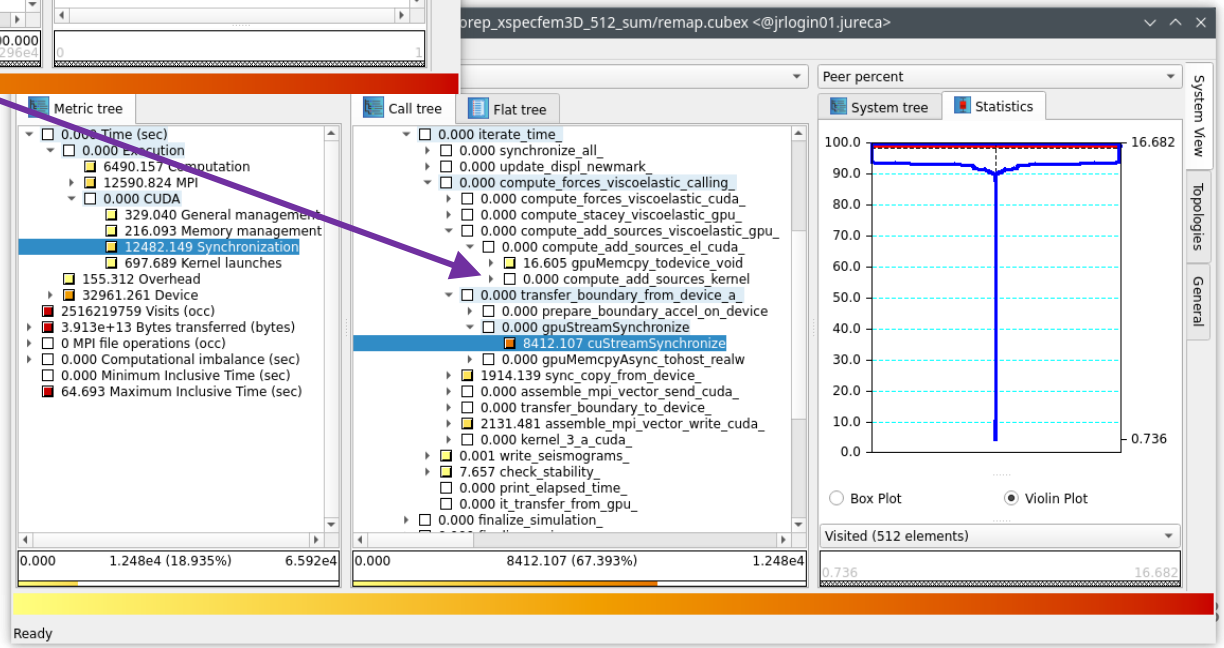


16x32 grid

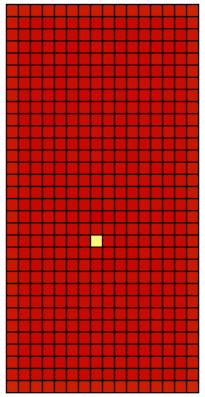


2D domain decomposition on GPUs

- many kernels are well balanced
- some kernels execute much faster for interior compared to edges
- only four GPUs handle seismogram receivers
- only one GPU (#243) executes *compute_add_sources_kernel*



16x32 grid



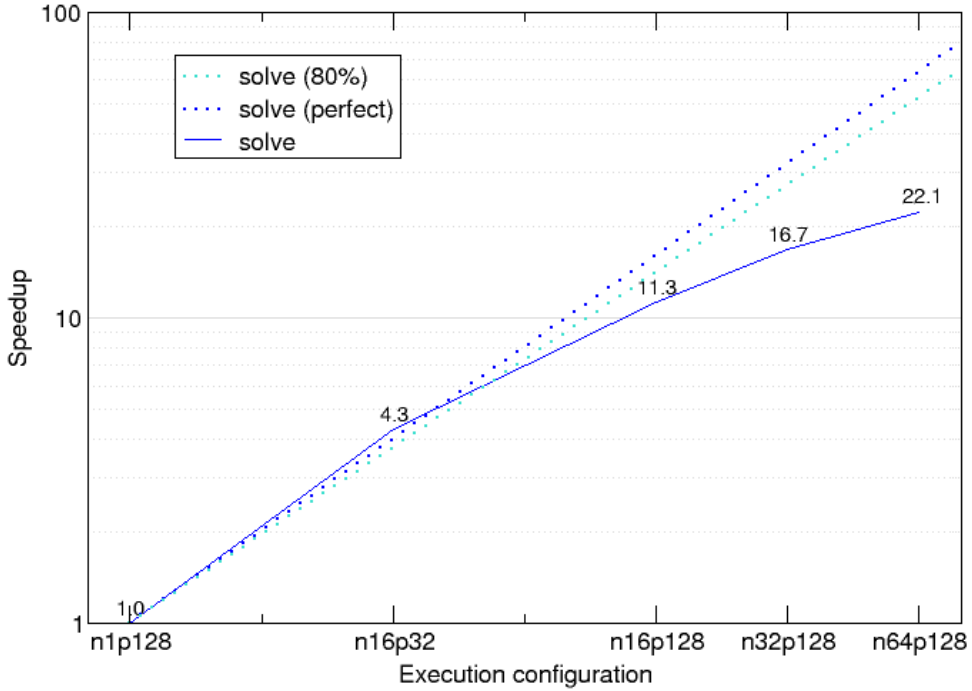
compute_add_sources_kernel executed on single GPU (#243) is rather short, however, results in all other GPUs having very long synchronization times in following *transfer_boundary_from_device_a*

- over two-thirds of CUDA synch time and over 30% of total CPU time

DOI 10.5281/zenodo.13643996

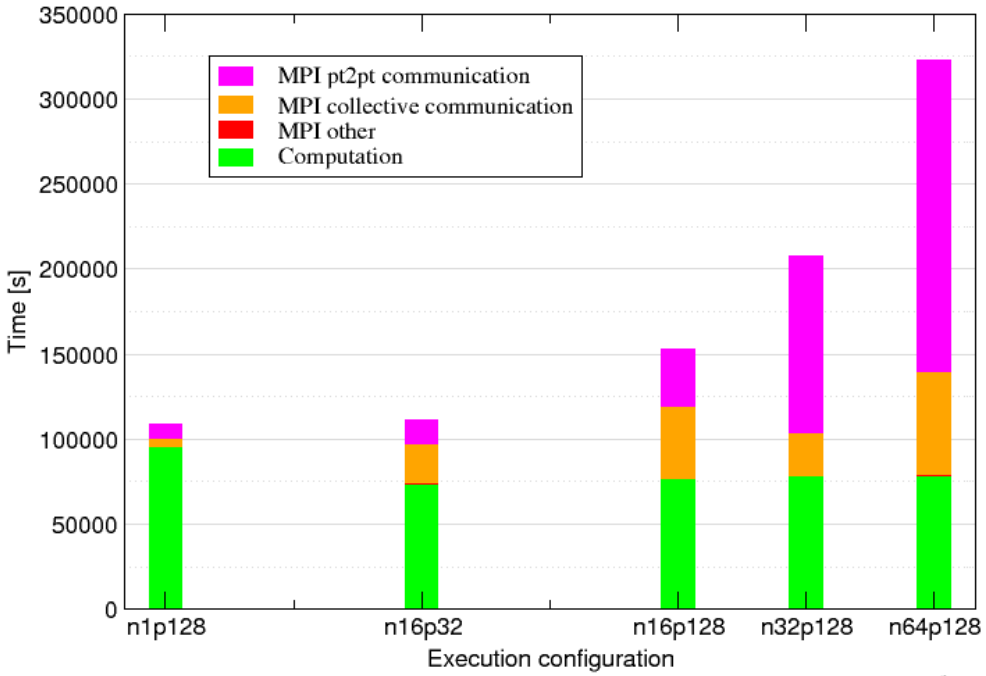
EuroHPC Joint Undertaking

Tandem@LUMI-C strong scaling



MPI ranks	128	512	2048	4096	8192
MPI configuration	n1p128	n16p32	n16p128	n32p128	n64p128
Solve iterations	18	18	19	20	20
Wall-clock time [s]	**850.78**	**216.77**	**74.76**	**50.73**	**39.42**
Computation time [s]	95671.33	74046.30	76451.98	78419.05	78618.84
Global scaling efficiency	0.879	0.862	0.624	0.461	0.296
- Computation time scaling	*1.000*	1.292	1.251	1.221	1.217
- Parallel efficiency	0.879	0.667	0.499	0.377	0.243
- - Load balance efficiency	0.904	0.882	0.735	0.783	0.638
- - Communication efficiency	0.971	0.756	0.679	0.482	0.382

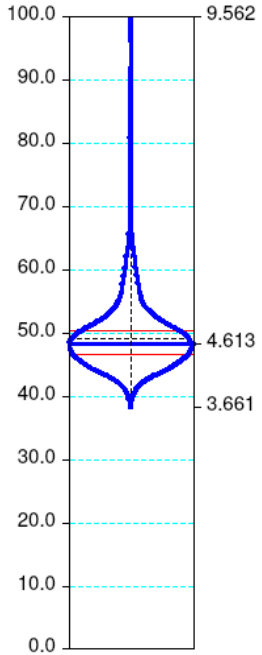
ChEESE

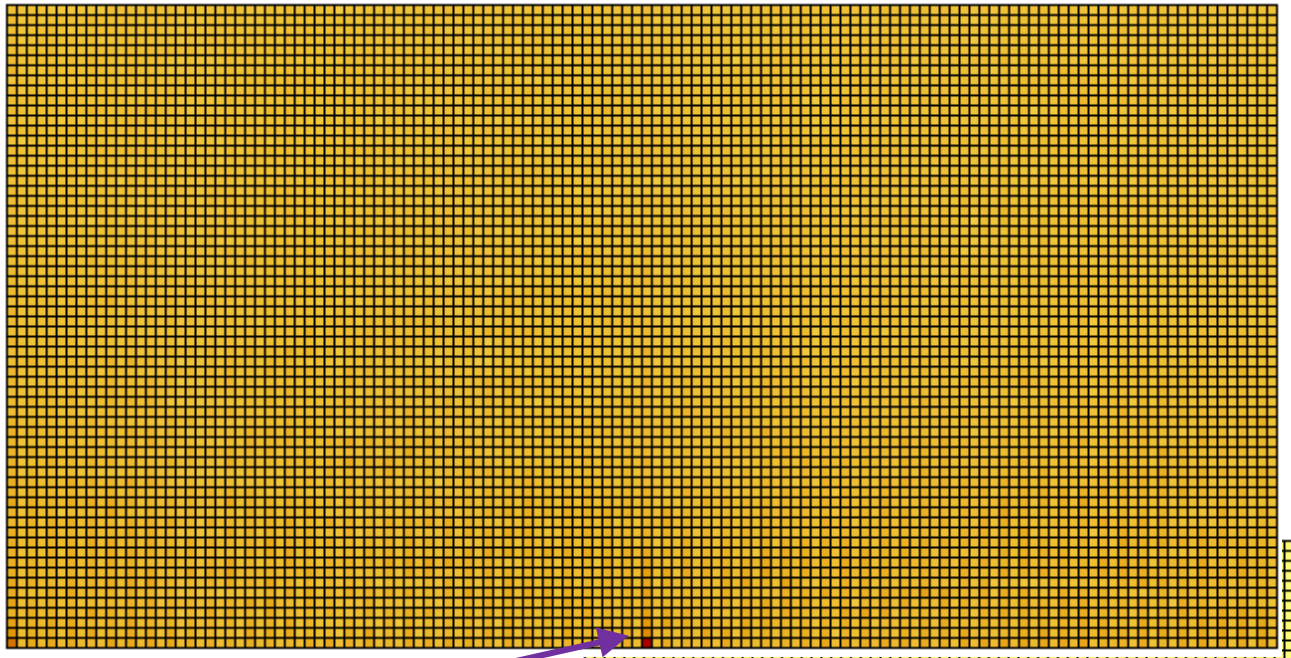


POP3_AR_009

- C++17 with MPI (& CUDA) built upon PETSc
- *static_problem* (incl. solver) chosen as Focus of Analysis
- Efficiency drops below 80% for more than 512 CPUs
- MPI pt2pt communication time grows significantly
- MPI collective communication time varies

Tandem@LUMI-C 8192p MPI usage





Each grid row corresponds to processes within a node

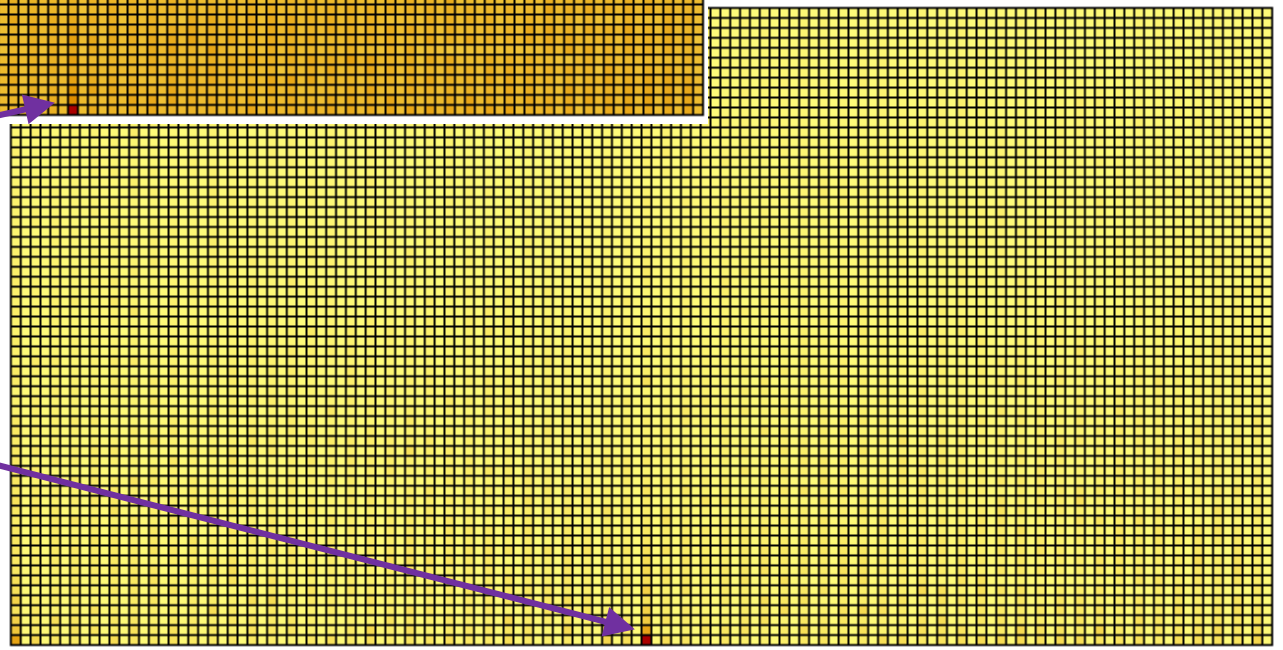
split after n elements: 128

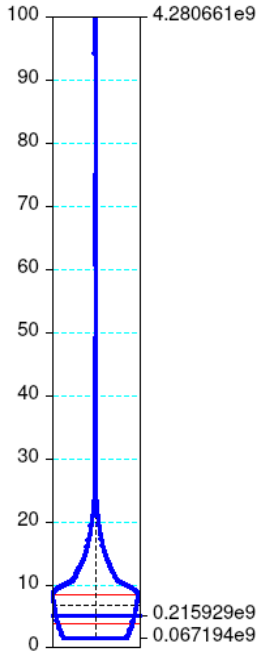
Computation time (above)

- MPI rank 8128 requires notably longer than others - more than twice the mean

MPI P2P bytes transferred (right)

- rank 8128 sends & receives several times more data than any others



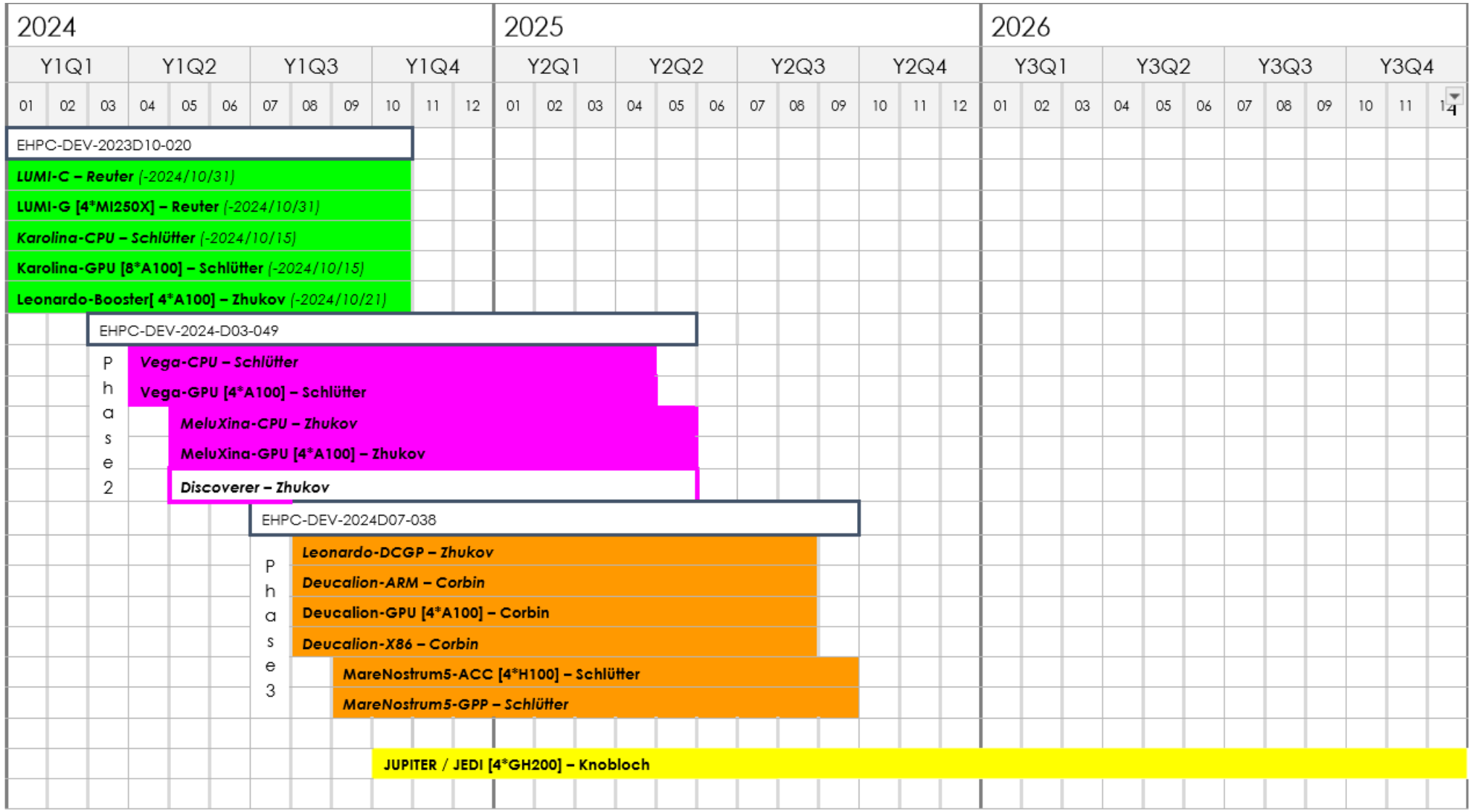


TRAINING COLLABORATIONS

Co-organised events & contributions to third-party events

- Training Sprint with Czechia+5 nearby EuroCC HPC National Competence Centres (NCCs)
  - 3-day hands-on virtual VI-HPS Tuning Workshop using Karolina CPUs & GPUs
    - bring-your-own-code for expert coaching in performance analysis/tuning
    - presentation & demonstration of all POP CoE tools (except energy assessment tools)
- 3 other in-person VI-HPS Tuning Workshops (CALMIP/F, NHR/D, LRZ/D)
- DiRAC/N8 Performance Analysis Workshop Series (DurhamU/UK)
- Archer2 AMD GPU performance analysis workshop (EPCC/UK)
- HPC Spectra Int'l Summer School (R-CCS/J) & EPICURE GPU hackathon (CINECA/I)
  - additional workshop using Fugaku proposed for HANAMI support not supported (yet)
- ISC HPC half-day hands-on tutorial on POP methodology & tools

EHPC-DEV PROJECT ALLOCATIONS



EuroHPC Development Access Calls

- 12 month projects
- no project extensions allowed(?)
- arrangement specific for HPC CoEs?

PLANS

Subject to revision

- Transition to project/allocations of POP CoE as existing allocations expire
- Work with HEs (and EPICURE?) on installations & modules for latest ***Scalasca/Score-P***
- Continue to collaborate with CASTIEL2 on CI/CD prototype(s)
- Follow-on assessments (POP second-level services) for CoE lighthouse codes
- First assessments for additional CoE lighthouse codes
  - neko & nekRS on JUWELS-Booster/JEDI/JUPITER, …
- Co-organise advanced performance analysis/tuning hackathons with EPICURE & HANAMI
- Identify NCC partner(s) for 2025 (and 2026) Training Sprint / VI-HPS Tuning Workshop
  - prioritise 'widening' and 'low R&I performing' states

Performance Optimisation & Productivity

Centre of Excellence in HPC

Contact:

- https://www.pop-coe.eu
- pop@bsc.es
- @POP_HPC
- youtube.com/POPHPC

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101143931.
The JU receives support from the European Union's Horizon Europe research and innovation programme and Spain, Germany, France, Portugal & Czechia.