

Intro to Scalasca/Score-P/CUBE

Brian Wylie (Jülich Supercomputing Centre)

HORIZON-EUROHPC-JU-2023-COE

EuroHPC Joint Undertaking

1 January 2024– 31 December 2026

Grant Agreement No 101143931

Scalasca

DOI 10.5281/zenodo.7440854

- Collection of trace-based performance tools
  - Specifically designed for large-scale systems
  - Features automatic trace analyzer providing wait-state, critical-path & delay analysis
  - Supports MPI, OpenMP, POSIX threads, and hybrid MPI+OpenMP/Pthreads
  - Uses Score-P instrumentation & measurement infrastructure and CUBE analysis report infrastructure
- Available under 3-clause BSD open-source license

- Documentation & sources:
  - https://www.scalasca.org
- Contact:
  - mailto: scalasca@fz-juelich.de

trace tools scalasca

EuroHPC Joint Undertaking

Score-P

DOI 10.5281/zenodo.10822140

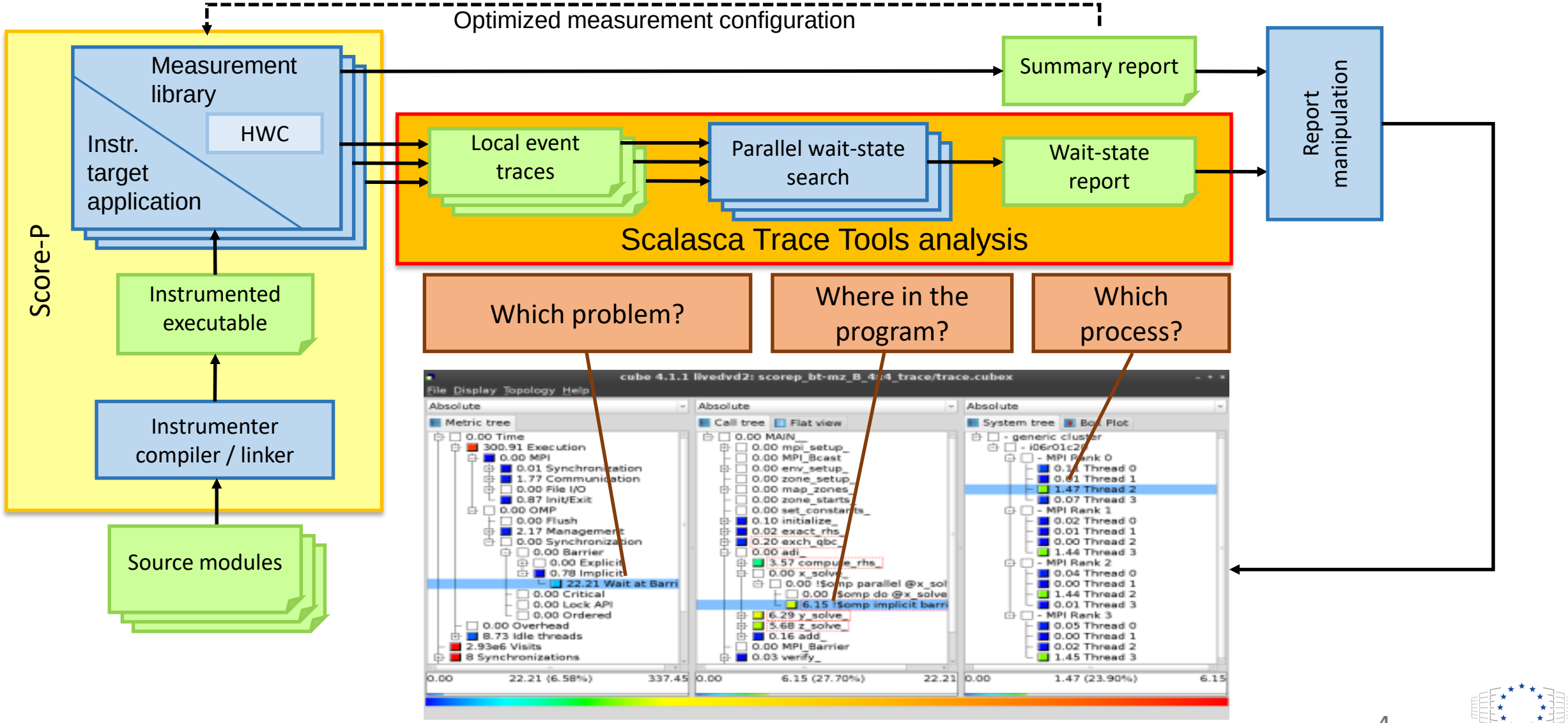
- Infrastructure for instrumentation and performance measurements
- Instrumented application can be used to produce several results:
  - Call-path profiling: CUBE4 data format used for data exchange
  - Event-based tracing: OTF2 data format used for data exchange

- Supported parallel paradigms:
  - Multi-process: MPI, SHMEM
  - Thread-parallel: OpenMP, Pthreads
  - Accelerator-based: CUDA, HIP, OpenCL, OpenACC, Kokkos

- Open Source; portable and scalable to all major HPC systems
- Initial project funded by BMBF
- Further developed in multiple third-party funded projects

- Documentation & sources: https://www.score-p.org

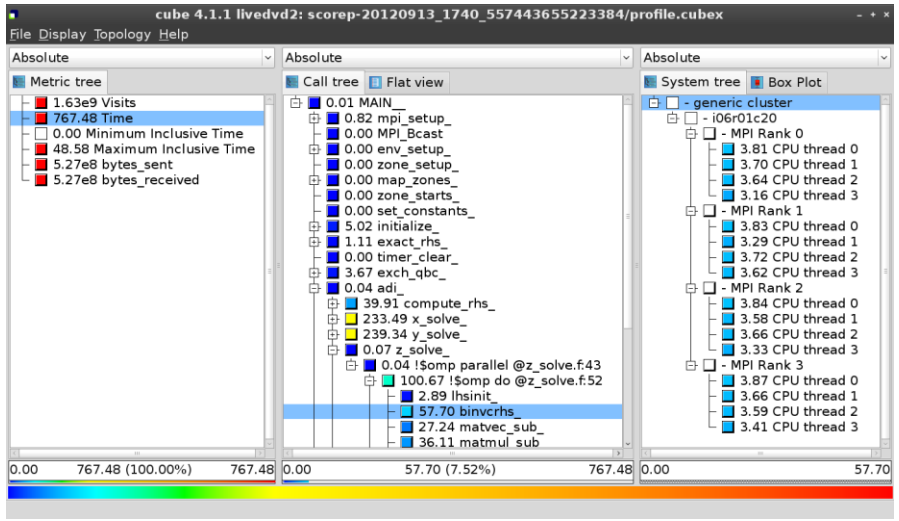
Scalasca workflow



Cube

DOI 10.5281/zenodo.8345207

- Parallel program analysis report exploration tools
  - Libraries for Cube report reading & writing
  - Algebra utilities for report processing
  - GUI for interactive analysis exploration
- Available under 3-clause BSD open-source license
- Documentation & sources:
  - http://www.scalasca.org
- User guide also part of installation:
  - <prefix>/share/doc/CubeGuide.pdf
- Contact:
  - mailto: scalasca@fz-juelich.de



Cube GUI options

- Run ***remote*** *(often convenient)*
  - start X server (e.g., Xming) locally or use mobaXterm or VNC
  - connect with X forwarding enabled
    - **-Y** may be faster but is insecure!
  - load cube module and start cube remotely

- Install & run ***local***
  - install Cube GUI locally on desktop
    - binary packages available for macOS & Windows and externally provided by OpenHPC and various Linux distributions
    - source package available for Linux, requires Qt
      - configure/build/install manually or use your favourite framework (e.g. Spack or EasyBuild)
  - copy .cubex file (or entire scorep directory) to desktop from remote system
    OR locally mount remote filesystem
  - start cube locally

https://www.**scalasca.org**/scalasca/software/cube-4.x/download.html

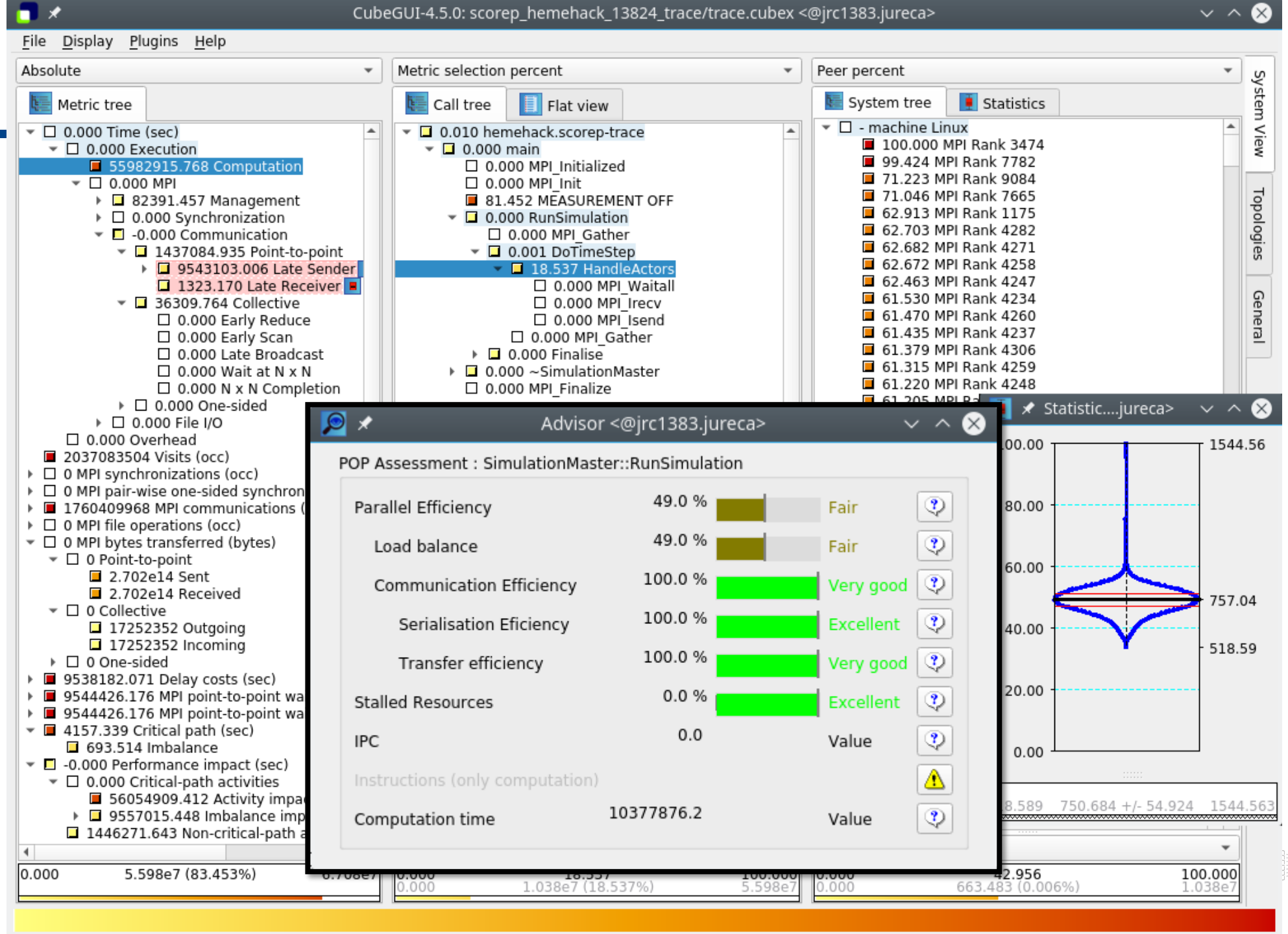
mailto: **scalasca@fz-juelich.de**

EuroHPC
Joint Undertaking

cube scalasca

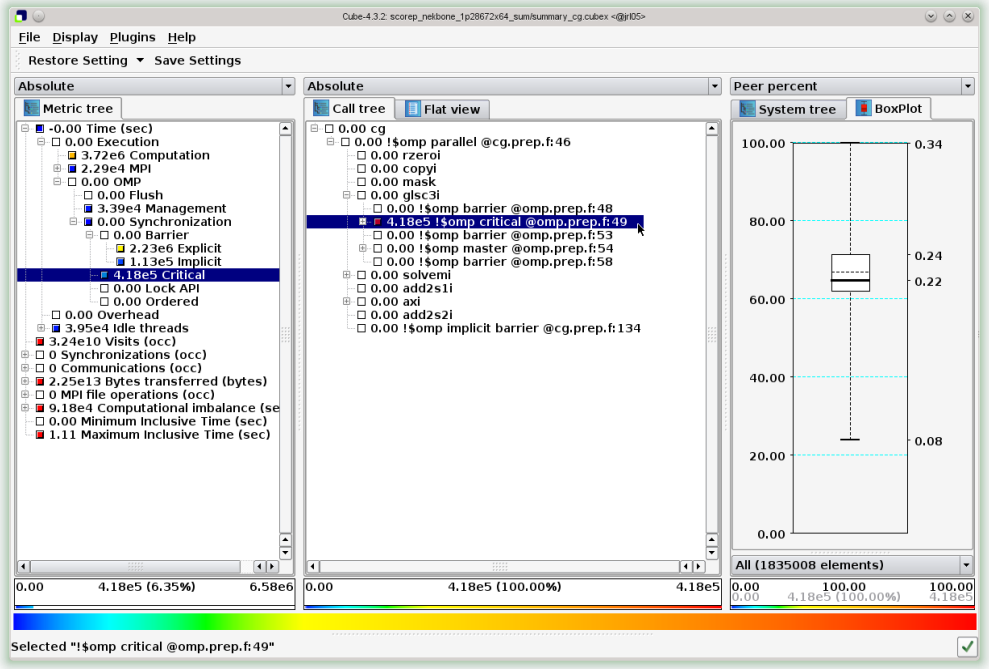
Advisor assessment

Assessment of execution efficiency factors using POP model

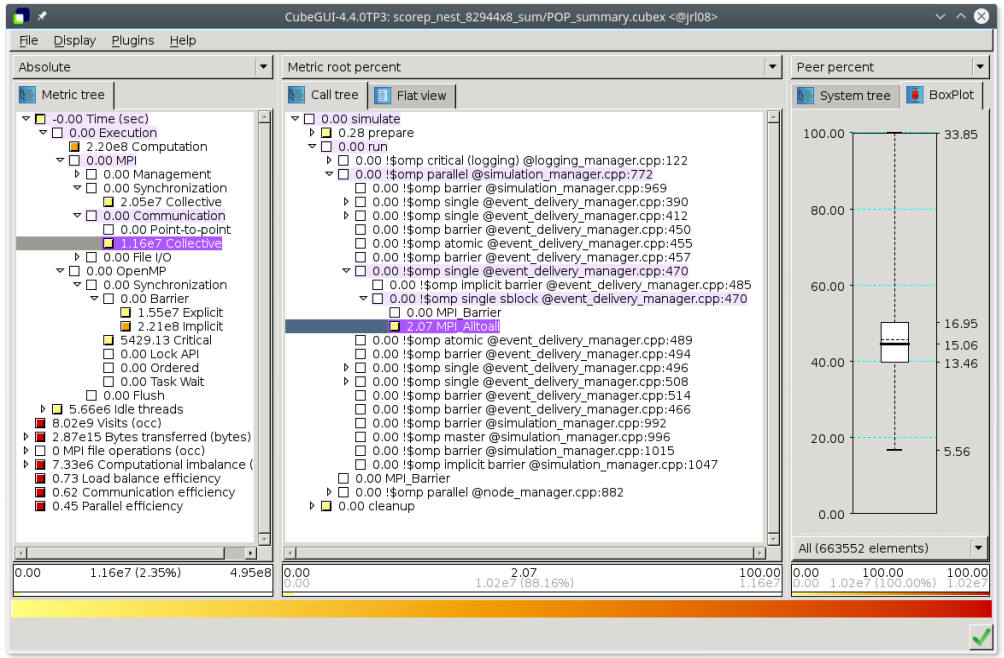


Scalasca Exascale Readiness

- Largest experiment
  - Application: Nekbone
  - System: *JUQUEEN* IBM BG/Q
  - 28,672 x 64 = 1,835,008 threads



- Largest experiment by user
  - Application: NEST
  - System: *K* computer
  - 82,944 x 8 = 663,552 threads



EuroHPC Joint Undertaking



POP3_AR_002

SPECFEM3D (MPI+CUDA) on Leonardo-B performance assessment report

Brian Wylie, Jülich Supercomputing Centre, Jan 2024

HORIZON-EUROHPC-JU-2023-COE

EuroHPC Joint Undertaking

1 January 2024– 31 December 2026

Grant Agreement No 101143931

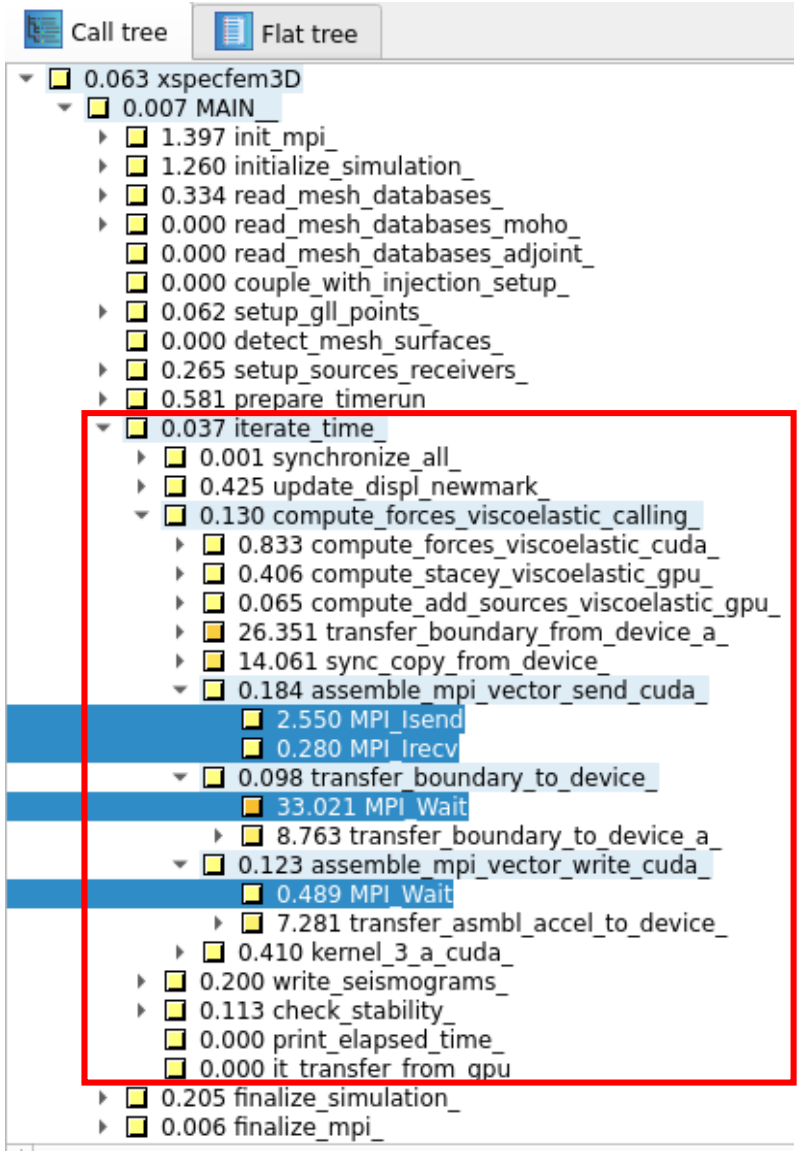
SPECFEM3D (Leonardo-B)

- SPECFEM3D
  - Software package for simulation of seismic wave propagation based on the spectral-element method
  - Assessment for HPC CoE for Exascale in Solid Earth (ChEESE)
  - Version 4.0.0 (release)
- Fortran90 (and some C) parallelized with MPI & CUDA: one MPI process per GPU
  - Intel oneAPI 2023.0.0 compilers and 2021.7.1 MPI libraries (not GPU-Aware)
- Testcase: 1 source in elastic domain; 4 seismic receiver stations
  - 48000 solver timesteps with intermediate writing disabled
  - weak scaling (22x 128x128 = 360,448 elastic elements per rank)
  - strong scaling (22x 1024x1024 = 23,068,672 elastic elements total)
- Executed on *Leonardo-Booster* Atos Bull Sequana XH21355 (CINECA)
  - 2345 compute nodes with 32-core Intel Xeon Platinum 8358 ('IceLake') CPUs @ 2.6 GHz and quad Nvidia A100 ('Ampere') GPUs [64GB]
  - Nvidia Mellanox HDR DragonFly++ interconnection network
- Measurements with Scalasca/2.6.1 using Score-P/8.3
- Focus of analysis (FOA): xspecfem3D/*iterate_time*

ChEESE

EuroHPC Joint Undertaking

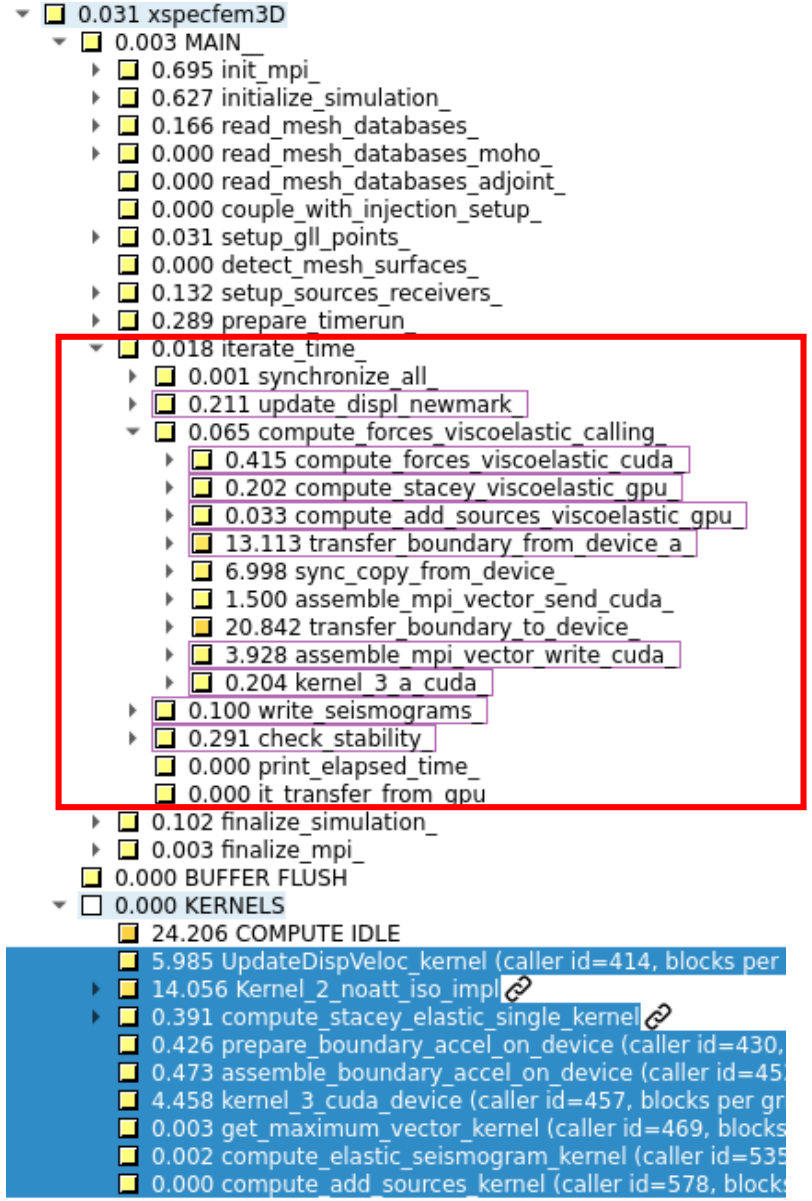
Execution call-tree & Focus of Analysis



- Structure
  - setup/initialise (amortised in full run)
    - read (w/o MPI), MPI_Bcast, MPI_Reduce, etc
  - solver (*iterate_time*)
    - 48000 timesteps
      - non-blocking point-to-point communication for boundary exchange with 2D neighbours
      - data transfer to/from associated GPU device and corresponding stream synchronization
    - summary output every 500 steps
      - collective MPI_Reduce
  - *write_seismograms* executed only once
- Focus of Analysis selected for assessment: *iterate_time*

EuroHPC
Joint Undertaking

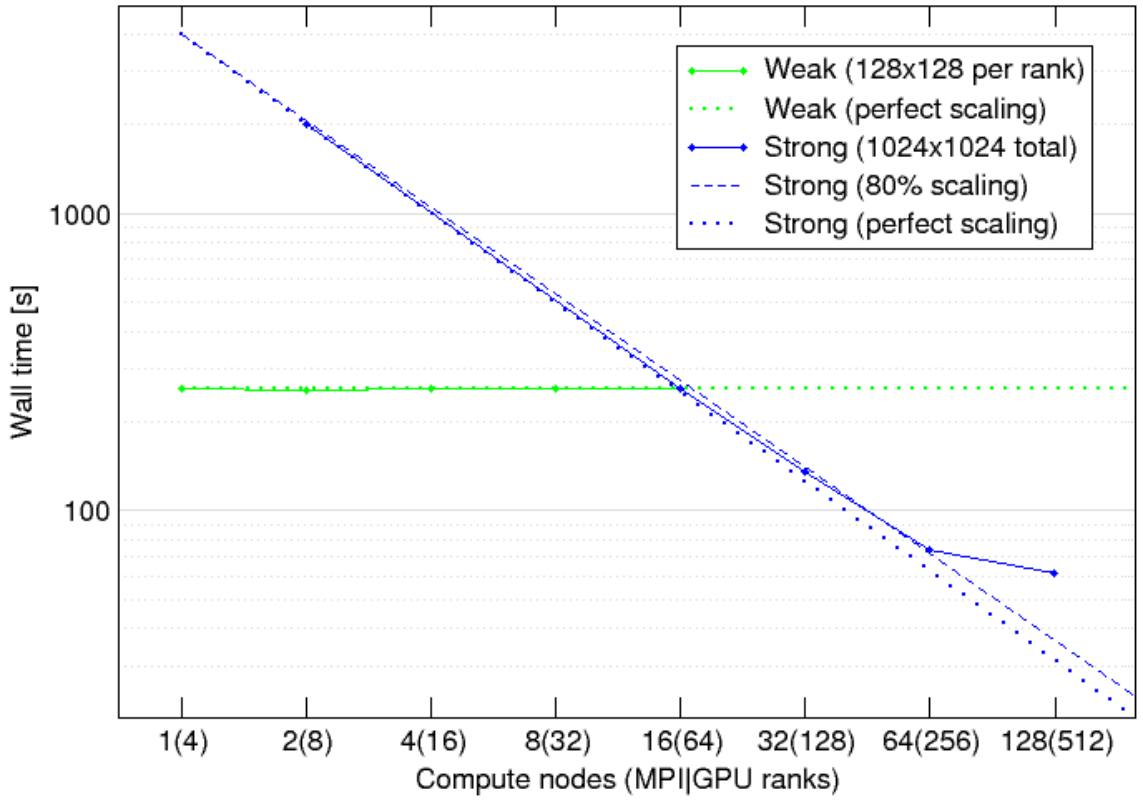
Execution call-tree & kernels

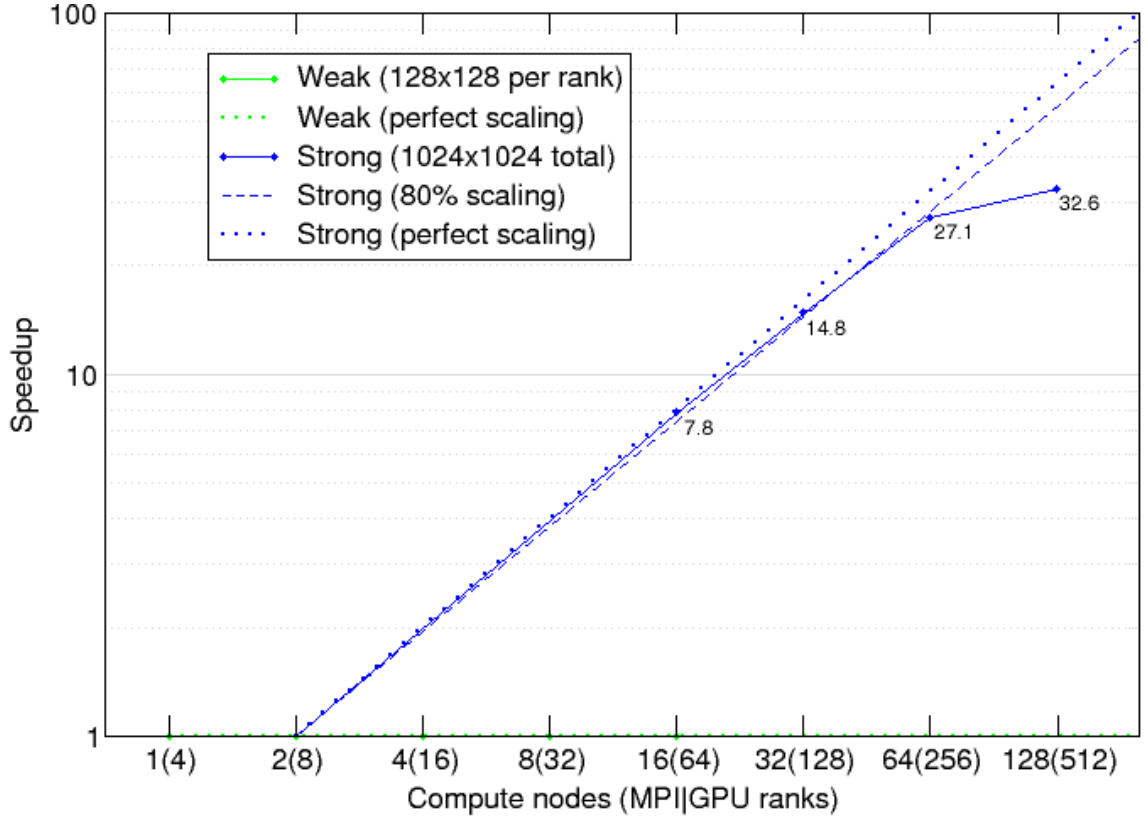


- Structure
  - solver (*iterate_time*)
    - contains all of the CUDA kernel executions
    - 48000 timesteps
      - seven of nine kernels executed by all ranks
      - characteristics oft determined by position in 2D grid
    - *compute_add_sources_kernel* only executed by a single GPU (rank 243 of 512)
    - *compute_elastic_seismogram* only by 4 nearby GPUs (ranks 241, 245, 273, 277 of 512)
- Focus of Analysis selected for assessment: *iterate_time*

EuroHPC
Joint Undertaking

Scaling & speed-up





- xspecfem3d FOA *iterate_time* on Leonardo-Booster
- Excellent weak scaling (expected to continue to higher node counts)
- Very good strong scaling (above 80% of perfect) to around 64 compute nodes (256 GPUs)

EuroHPC
Joint Undertaking

Efficiency model (weak scaling)

	Problem size	256x256	512x256	512x512	1024x512	1024x1024
	MPI|GPU ranks	**4**	**8**	**16**	**32**	**64**
	Wall time [s]	**255.873**	**255.462**	**256.035**	**255.898**	**255.948**
XPU	Global scaling efficiency	0.514	0.516	0.514	0.514	0.514
	- Computation time scaling	*1.000*	0.992	0.977	0.972	0.963
	- Parallel efficiency	0.514	0.520	0.526	0.529	0.534
	- - Load balance efficiency	0.522	0.526	0.533	0.536	0.541
	- - Orchestration efficiency	0.986	0.988	0.988	0.988	0.988
GPU	Global scaling efficiency	0.986	0.987	0.985	0.986	0.986
	- Computation time scaling	*1.000*	1.000	1.000	1.000	1.000
	- Parallel efficiency	0.986	0.987	0.985	0.986	0.986
	- - Load balance efficiency	1.000	0.999	0.997	0.998	0.998
	- - Orchestration efficiency	0.986	0.988	0.988	0.988	0.988

- Orchestration efficiency
  - MPI communication & CUDA management
  - aka Communication efficiency

Key:
1.10
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00

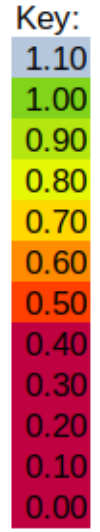
- Excellent GPU weak scaling efficiency
- Very poor CPU efficiency?
- Moderate XPU (GPU+CPU) efficiency?

EuroHPC
Joint Undertaking

Efficiency model (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**
XPU	Global scaling efficiency	0.515	0.504	0.476	0.437	0.263
	- Computation time scaling	*1.000*	0.943	0.861	0.835	0.715
	- Parallel efficiency	0.515	0.534	0.553	0.523	0.367
	- - Load balance efficiency	0.518	0.541	0.583	0.586	0.671
	- - Orchestration efficiency	0.995	0.988	0.948	0.892	0.547
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

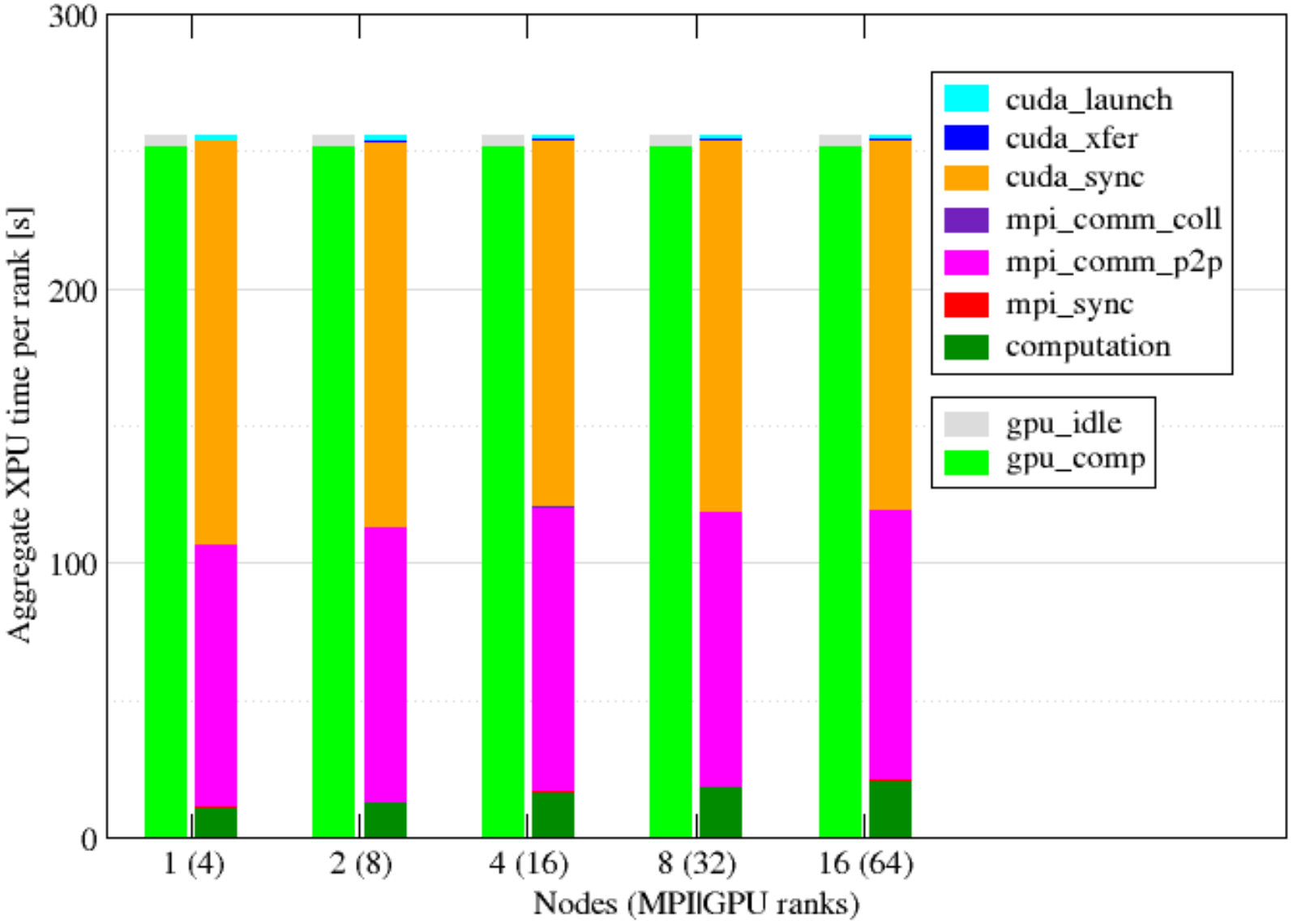
- Orchestration efficiency
  - MPI communication & CUDA management
  - aka Communication efficiency



- Good GPU weak scaling efficiency to 128 GPUs (excellent load balance)
- Very poor CPU efficiency?
- Moderate XPU (GPU+CPU) efficiency?

EuroHPC
Joint Undertaking

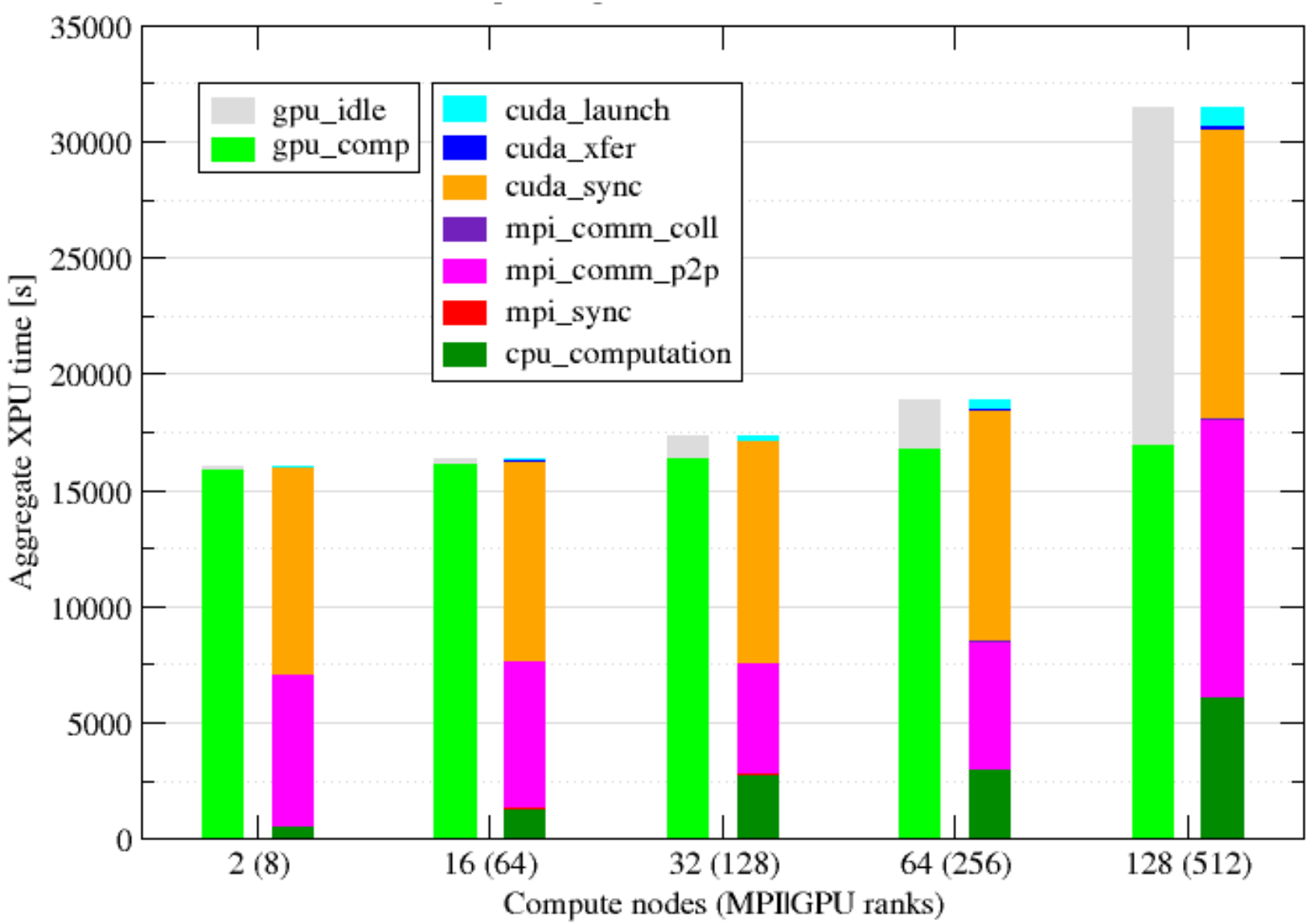
Weak scaling (128x128 per rank)



- Excellent weak scaling
- Little GPU idle time
- MPI communication effectively overlapped with GPU kernel computation

EuroHPC
Joint Undertaking

Strong scaling (1024x1024 total)

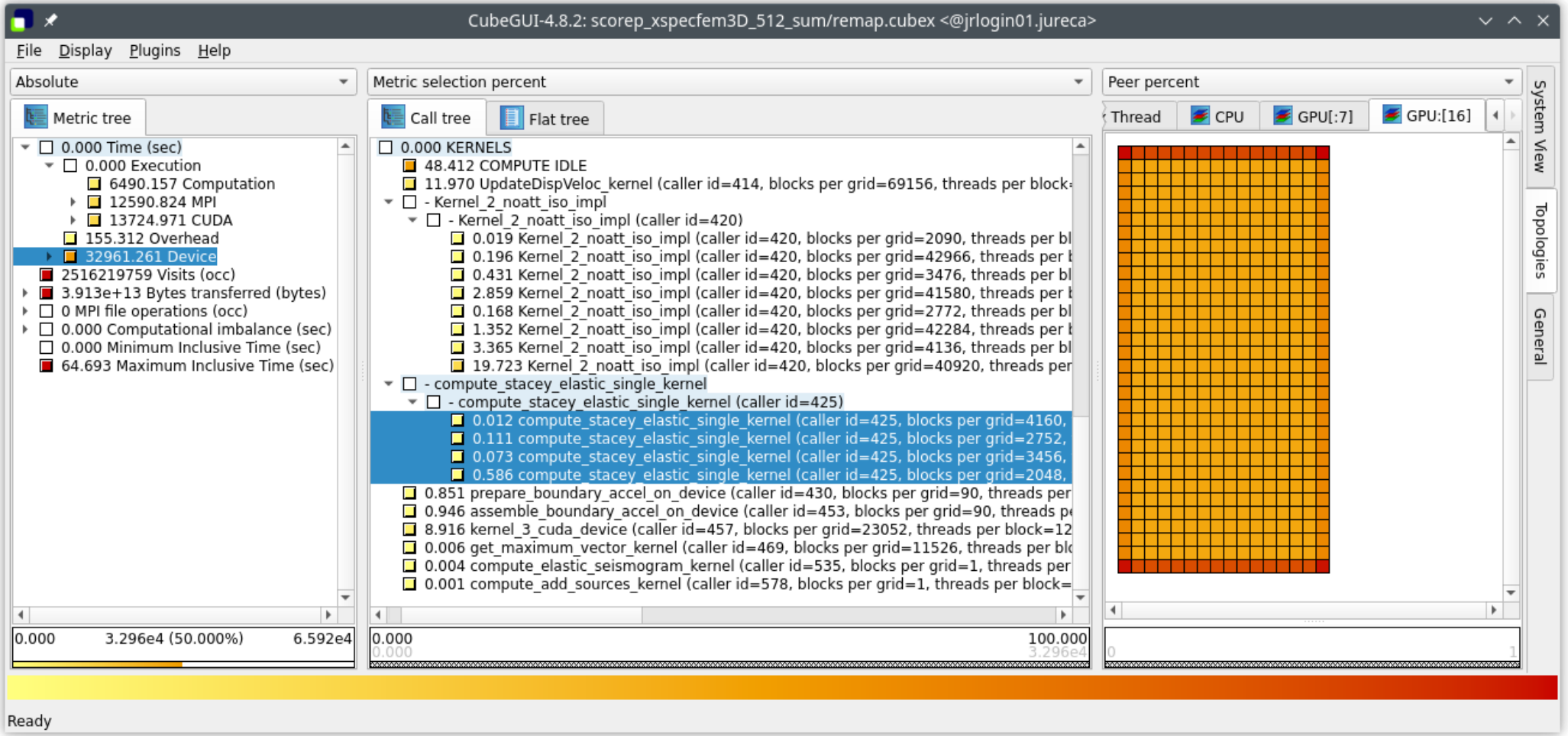


- Good scaling to 256 GPUs (64 nodes)
- GPU computation time slowly grows progressively
- GPU idle time grows for 256 & particularly 512 GPUs
- CPU computation time grows substantially
  - sync_copy_from_device & transfer_boundary_to_device_a
- For 512 GPUs, growing MPI communication no longer fully overlapped with GPU kernel computation

EuroHPC Joint Undertaking

Topological inhomogeneities

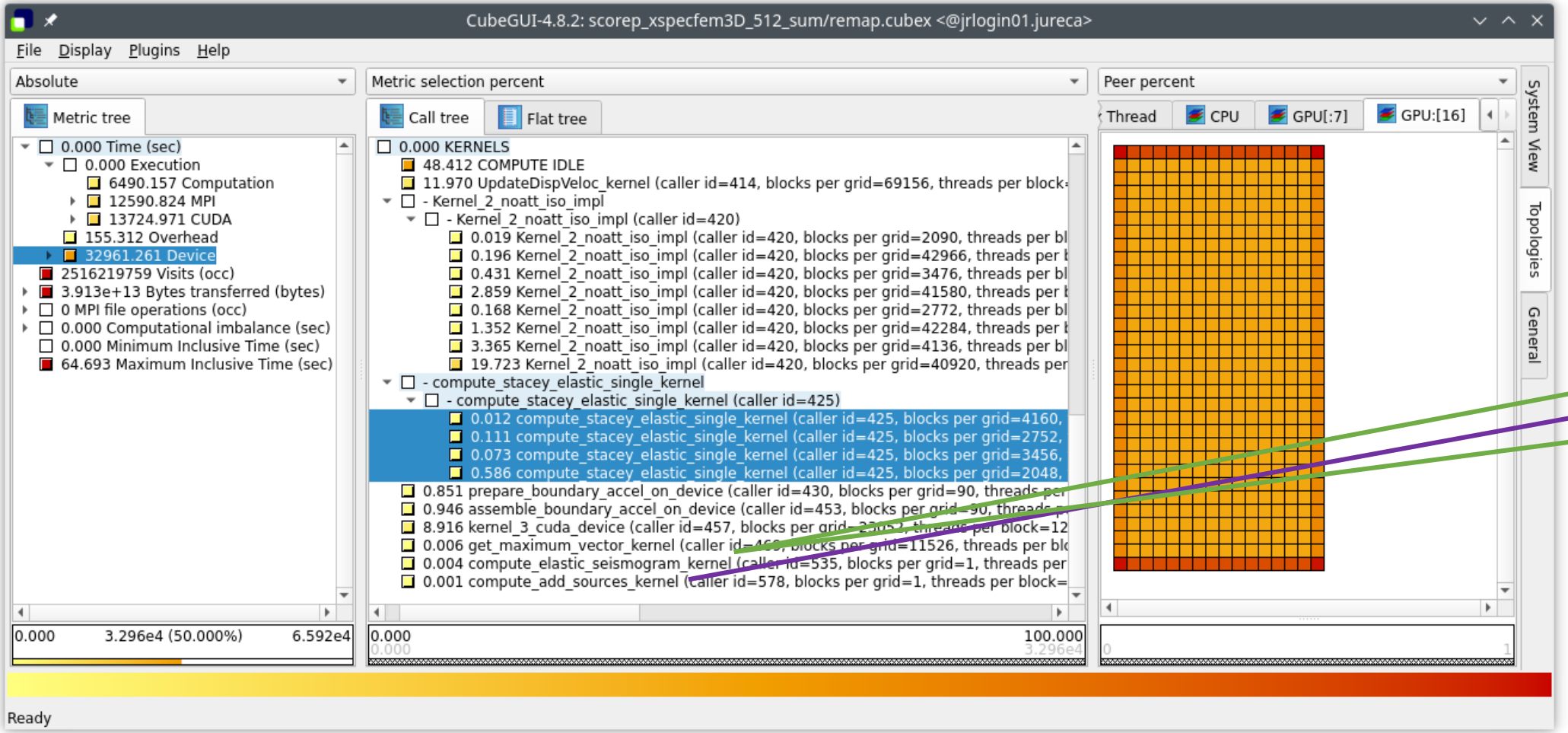
Kernel variant executed (characteristics and corresponding execution time) varies according to position in 2D grid: four corners, upper/lower & left/right edges, interior



16x32 grid

Topological inhomogeneities

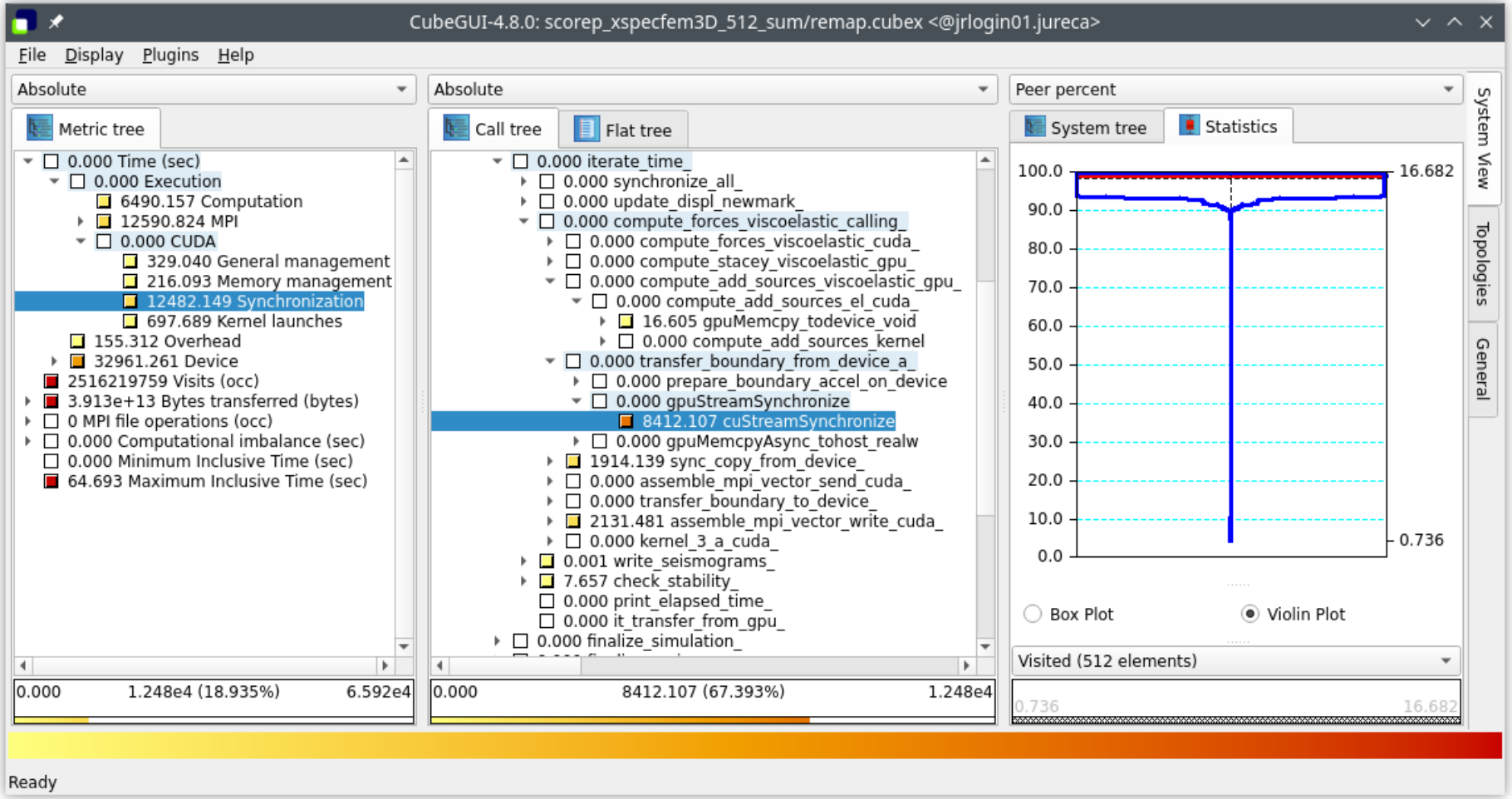
Kernel variant executed (characteristics and corresponding execution time) varies according to position in 2D grid: four corners, upper/lower & left/right edges, interior

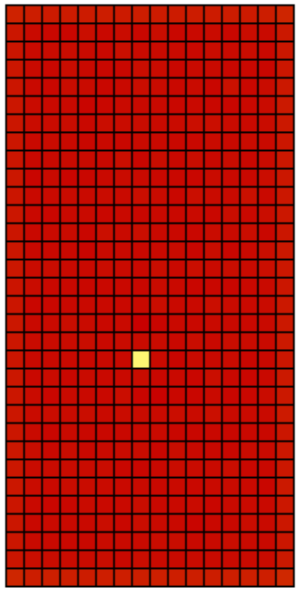


16x32 grid

GPU computation imbalance

36% of CPU execution time is CUDA synchronization, 67% of which is 16s within *transfer_boundary_from_device_a* following *compute_add_sources_kernel* that's only executed by a single GPU (source rank 243 of 512)





Summary of observations

- *iterate_time* (solver) chosen as focus of analysis
  - negligible time for initialization/finalization
- Excellent weak scaling up to 16 nodes (64 GPUs) and likely beyond
  - Computation very well balanced over GPUs; Excellent GPU utilization
  - MPI P2P communication time grows with scale, but effectively overlapped with GPU computation kernels
- Good strong scaling speedup up to 64 nodes (256 GPUs)
  - Computation remains very well balanced over GPUs
  - Orchestration efficiency progressively diminishes
    - *compute_add_sources_kernel* execution by a single GPU seems the main origin
  - MPI P2P communication time grows significantly, becomes no longer fully overlapped with GPU computation kernels

EuroHPC
Joint Undertaking

Performance Optimisation and Productivity 3

A 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 and the Czech Republic.