ChEESE and POP: a Story of Success and Fruitful Interaction

30th POP Webinar
11th November 2024

Arnau Folch (CSIC, Spain)
Piero Lanucara (CINECA, Italy)

Many thanks to Vadim Montellier (CNRS), Jose Gracia (HLRS), Brian Wylie (JSC), Marta García (BSC) and the rest of people form ChEESE WP2

Project funded by EuroHPC under the grant agreement No 101093038.

Outline

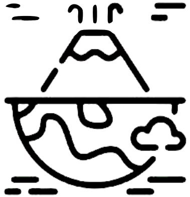
01 ChEESE-2P overview

02 The ChEESE-2P WP2 strategy

03 SPECFEM3D showcase

ChEESE : The CoE in geophysics (and geohazards)

ChEESE

	Earthquakes	Volcanoes	Tsunamis	MHD	Geodynamics	Glaciers
ChEESE-1P (2019-2022)						
ChEESE-2P (2023-2026)						

ChEESE in a nutshell

ChEESE covers 3 approaches to exascale

1

Capability computing

Solve problems that traditionally have been parameterised because are unaffordable with current hardware

2

Capacity computing

Solve ensembles of single problems affordable with petascale-range machines but that can aggregate into an exascale workflow (e.g. data inversion, model data assimilation, uncertainty quantification, etc)

3

Urgent computing

Solve capability/capacity problems under strict time constrains in terms of time-to-solution (emergency situations)

ChEESE in a nutshell

ChEESE sustains on 3 pillars

11

Flagship codes

Open source community codes on seismology, tsunamis, volcanoes, geodynamics, magneto-hydrodynamics, and glaciers

9

Pilot Demonstrators

Workflows that address underpinning capability/capacity/ UC exascale computational challenges

15

Simulation Cases and services

Enabling of services on socially-relevant aspects of geohazards like urgent computing for disaster response, early warning and hazard assessment

ChEESE Pillar 1: the 11 flagship codes

Domain	No	Code	Accelerated	Mini-app	Licence	PDs
Computational Seismology	1	SeisSol	CUDA, SYCL	yes	BDS3	PD1,PD4
	2	SPECFEM3D	CUDA, HIP	yes	GNU GPL v3	PD1,PD2
	3	ExaHyPE	on-going	no	Modified BSD3	PD1,PD4
	4	Tandem	on-going	yes	BSD3-clause new	PD1
MHD	5	xSHELLS	CUDA	yes	CeCILL	PD7
Tsunami modelling	6	HySEA	CUDA	yes	Core under CCA	PD3,PD4
Volcanology	7	FALL3D	OpenACC	yes	GNU GPL v3	PD5
	8	OpenPDAC	on-going	no	GNU GPL v3	PD6
Geodynamics	9	LaMEM	on-going	no	GNU GPL v3	PD8
	10	pTatin3D	CUDA	yes	GNU GPL v3	PD8
Glacier modelling	11	Elmer/ICE	on-going	no	GNU GPL v2-v3	PD9

ChEESE Pillar 1: the 11 flagship codes (wide range of typologies)

Code	Wave propagation	CFD (multiphysics)	CFD (geodynamics)	Shallow waters	Transport
SeisSol	✓				
SPECFEM3D	✓				
ExaHyPE	✓				
Tandem			✓		
xSHELLS		✓			
HySEA				✓	
FALL3D					✓
OpenPDAC		✓			
LaMEM			✓		
pTatin3D			✓		

ChEESE Pillar 1: the 11 flagship codes

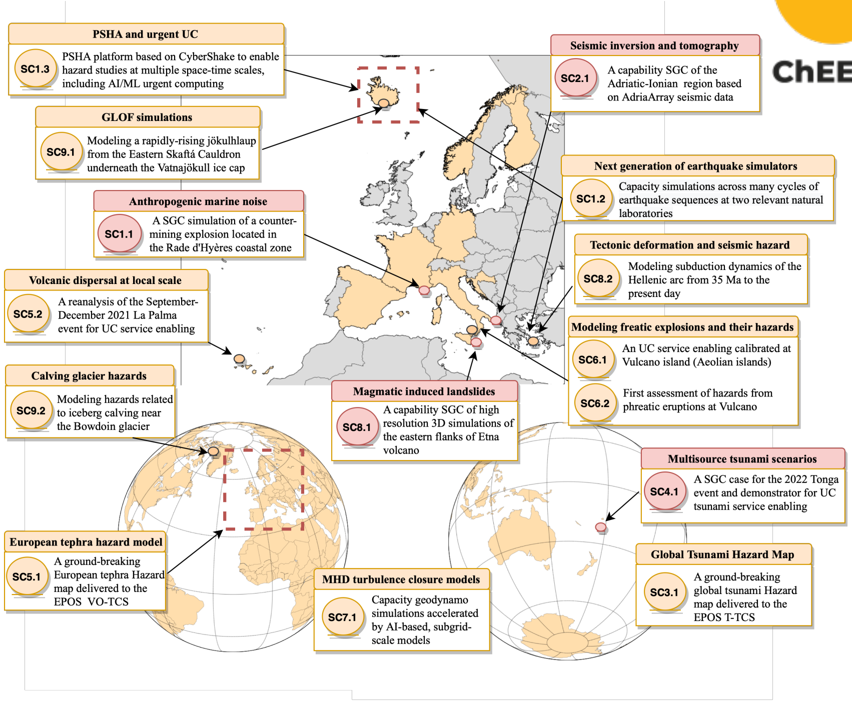
ChEESE

WP	WP Title
1	Project Management
2	Exascale technical challenges in flagship codes
3	Co-design with European HPC vendors and technologies
4	Simulation ecosystem
5	Farming Pilot Demonstrators
6	Simulation cases and service enabling

WP	Code preparation activities	
5	Model physics	New code physics, couplings and forcing terms (T5.1)
2	Code performance (audit driven)	Code audit(s) and related POP metrics (T2.1)
		GPU porting and fine tuning (T2.2)
		Single heterogeneous node performance (T2.3)
		Multi-node performance (T2.3)
		Algorithmic improvements (T2.3)
		Resilience and fault tolerance (T2.3)
		IO performance (T2.3)
3	co-desing	Co-design with mini-apps (T3.1, T3.2 and T3.3)

ChEESE Pillar 2: the 9 Pilot Demonstrators (PDs)

No	PD name	Area	Initial TRL	Target TRL
1	Extreme-scale modeling of seismic hazards	CS	5	7
2	Joint seismic inversion and tomography		4	6
3	Global tsunami hazard and uncertainty quantification	T	6	8-9
4	Complex multi-source tsunami modeling		3	5
5	Ensemble-based volcanic dispersal at multiple scales	V	5	7-8
6	Multiphase 3D volcanic explosion modeling		4	7
7	The Earth's dynamo model	MHD	3	5
8	Geodynamics to geohazards	GD	3	5
9	Glacial outburst floods	GL	2	4



PDs will materialize in **15 Simulation Cases** (SCs) to produce:

- Open datasets
- Urgent computing service enabling
- TCS-TSU and TCS-VO in EPOS

15 Simulation cases (4 Scientific Grand Challenges)

PD	No	SC name	Type	SGC	Expected outcome
PD1	SC1.1	Anthropogenic noise and seismicity, wave propagation at higher frequencies in Rade d'Hyères	**capability**	**yes**	open datasets
	SC1.2	Next generation of earthquake simulators for the Hellenic arc and Iceland	capacity	-	earthquake catalogs
	SC1.3	Physics-based PSHA and urgent UQ for Iceland	capacity	-	**UC service enabling**
PD2	SC2.1	Seismic tomography of the Adriatic-Ionian region	**capability**	**yes**	open datasets
PD3	SC3.1	The Global Tsunami Hazard Map	capacity	-	**TCS-TSU service in EPOS**
PD4	SC4.1	Multi-source local high-resolution scenario with emphasis on the 2022 Tonga case	**capability**	**yes**	**UC service enabling**
PD5	SC5.1	Towards the European tephra hazard map	capacity	-	**TCS-VO service in EPOS**
	SC5.2	Volcanic dispersal at local scale: a reanalysis of the 2021 La Palma case	capacity	-	**UC service enabling**
PD6	SC6.1	Urgent high-resolution, 3D multiphase flow simulation of phreatic eruptions at Vulcano	**capability**	-	**UC service enabling**
	SC6.2	Long-term probabilistic hazard maps for phreatic eruptions at Vulcano island	capacity	-	open datasets
PD7	SC7.1	MHD turbulence closure models in planetary cores using AI	capacity	-	open datasets
PD8	SC8.1	Magmatic systems and induced landslides: the Etna case	**capability**	**yes**	open datasets
	SC8.2	3D evolution of tectonic deformation as driver of seismic hazard in the Hellenic arc	**capability**	-	open datasets
PD9	SC9.1	Eastern Skaftá cauldron GLOF simulation, Iceland	capacity	-	open datasets
	SC9.2	Calving glacier hazard in Greenland	capacity	-	open datasets

ChEESE Pillar 3: enabling of 3 typologies of services

1

Urgent Computing (UC) service enabling at EuroHPC systems (emergency access mode)

Collaboration with IUB members (trials) and EuroHPC Infrastructure Advisory Group (INFRAG) for UC service certification, deployment, and access

PD1	SC1.3	Physics-based PSHA and urgent UQ
PD4	SC4.1	Multi-source high-resolution tsunamigenic scenarios
PD5	SC5.2	Volcanic dispersal at multiple scales
PD6	SC6.1	Multiphase flow simulation of phreatic eruptions

2

Services integrated in EPOS Tematic Core Services (TCS)

TCS-TSU and TCS-VO

PD3	SC3.1	Global Tsunami Hazard Map
PD5	SC5.1	The European tephra hazard map

3

CI/CD in EuroHPC systems

Collaboration with CASTIEL2

EOSC SYNERGY → ChEESE

- Software Quality Assurance as a Service (SQAaaS)

ChEESE and POP

	ChEESE-1P and POP2	ChEESE-2P and POP3
Audit campaigns	• 2 campaigns (10 codes)	• 2 campaigns (11 codes)
Audit purpose	• Identify code bottlenecks	• Identify code bottlenecks • Code performance baseline • Code performance monitoring based on POP metrics (including TALP)
Audit tools	• Score-p (level 1 codes) • Extrae (level 2 and codes)	• Several (systems dependent)
Audit focus	• CPU (POP hybrid metrics were still under development at that time)	• CPU and GPU
Other		• Second Level Services (SLS) for "Correctness-check", "Energy-efficiency study", and "Advisory study”.

ChEESE-2P and POP

- 10 codes have been analysed for baseline scaling behaviour and POP metrics for the Focus of Analysis (FoA)

Code	Programing model	Version	System	Audit tool
SeisSol	MPI + SYCL + CUDA	1.1.3	Leo-B	Extrae; SeisSol self-profiling
SPECFEM3D	MPI + CUDA	devel 4.0.0	Leo-B	Scalasca/2.6.1; Score-P/8.3
ExaHyPE	MPI + OpenMP target		Leo-B	Score-P/8.3
Tandem	MPI	1.1.0	LUMI-C	Score-P/8.3
xSHELLS	MPI + CUDA		Vega-G	
HySEA	MPI + CUDA	4.3.0	Leo-B	Extrae
FALL3D	MPI + OpenACC	8.2.1	Leo-B	Extrae and mpiP
OpenPDAC	MPI	1.0.0	LUMI-C	DLB/TALP and mpiP
LaMEM	MPI		LUMI-C	Extrae
pTatin3D	MPI + HIP		LUMI-G	Extrae
ELMER/ICE	MPI	fdd6c58f1	LUMI-C	Extrae

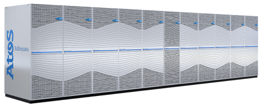
- **Issues encountered**:
  - Heterogeneity of the performance analysis tools across EuroHPC systems
  - Need for dedicated project access resources (no audit access mode)

Outline

Bridging the gap between ChEESE-2P Flagship codes and the European HPC ecosystem

Area	No	Code	Lead	ChEESE -1P
CS	1	SeisSol	LMU/TUM	yes
	2	SPECFEM3D	CNRS	yes
	3	ExaHyPE	TUM	yes
	4	Tandem	LMU	no
MHD	5	xSHELLS	CNRS	yes
T	6	HySEA	UMA	yes
V	7	FALL3D	CSIC	yes
	8	OpenPDAC	INGV	no
GD	9	LaMEM	UM	no
	10	pTatin3D	SU	no
GL	11	Elmer/ICE	CSC	no

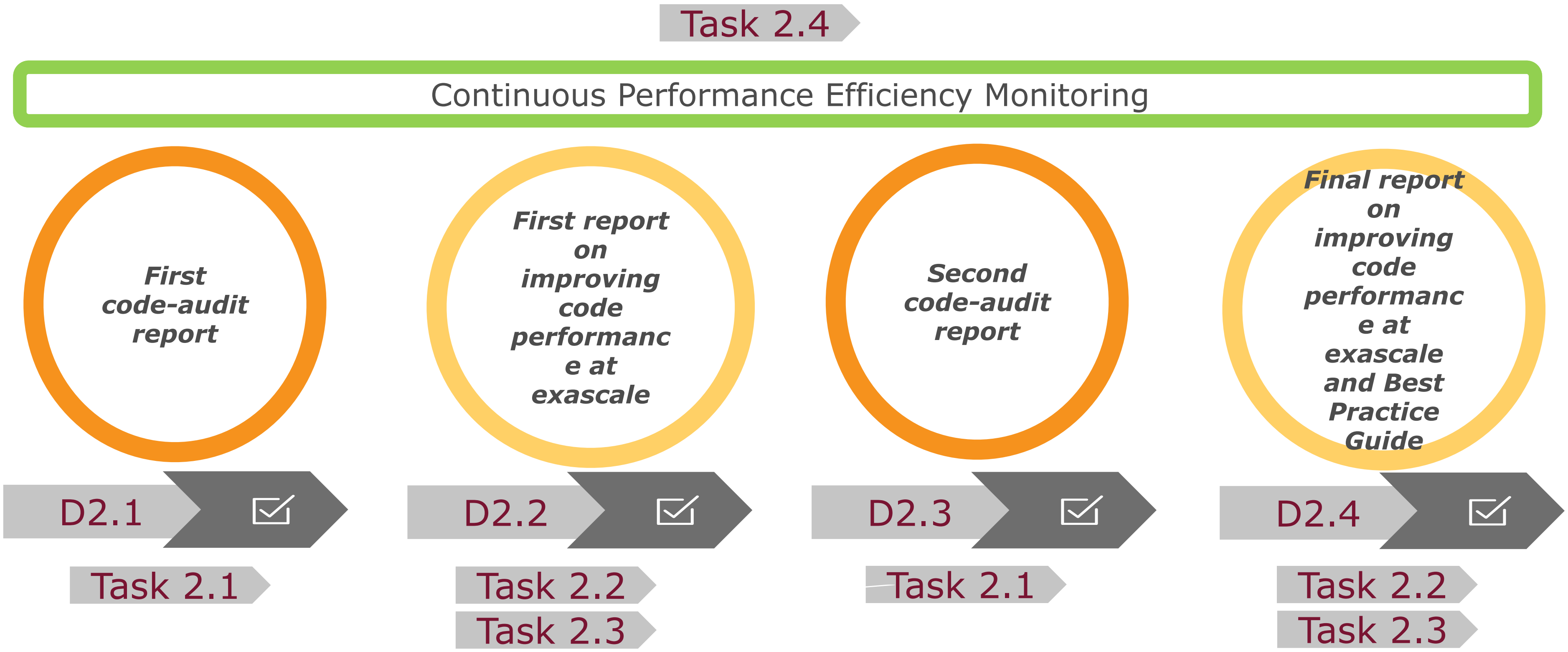
Tier-0 (pre-exascale)			TOP500
LUMI	Finland	~500 PFlop/s peak	5
Leonardo	Italy	~300 PFlop/s peak	6
MN5 ACC	Spain	~250 PFlop/s peak	8



More than 1000 PFlop/s peak aggregated computing power…

…but with a variety of different (accelerated) architectures

WP2 methodology



T2.1 Code audits (including D2.1 statistics)

- ❑ **Goal: set performance baseline for optimization work and progress monitoring. Identify performance bottlenecks.**
- ❑ **Method: two performance audit campaigns (M3-M15, M27-M36)**
- ❑ **Actual performance analysis done in collaboration with POP3**
- ❑ **Target EuroHPC systems such as Leonardo, LUMI, …**

POP metrics*	final	preliminary
#	7	4

POP discussion	observations only	recommendations	none
#	2	4	4

The same scientific use cases to be used for both assessments

System	Leo-G	LUMI-C	LUMI-G
#	5	4	1

Scaling*	weak	strong
#	3	8

max MPI ranks	512	1024	8192
#	2	1	1

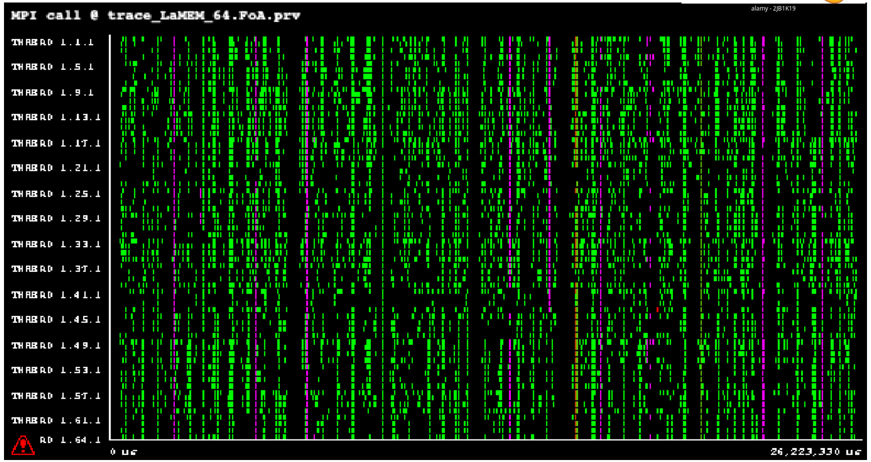
max GPUs*	16	32	64	256	512
#	1	1	2	2	1

- ❑ **Access to EuroHPC systems challenging and complicated**
- ❑ **Availability of performance analysis tools on EuroHPC systems**

T2.1 Code audits (including some D2.1 Highlights)

- ❑ **Useful to have a clear idea on what is exactly the bottleneck in the performance and where is located**

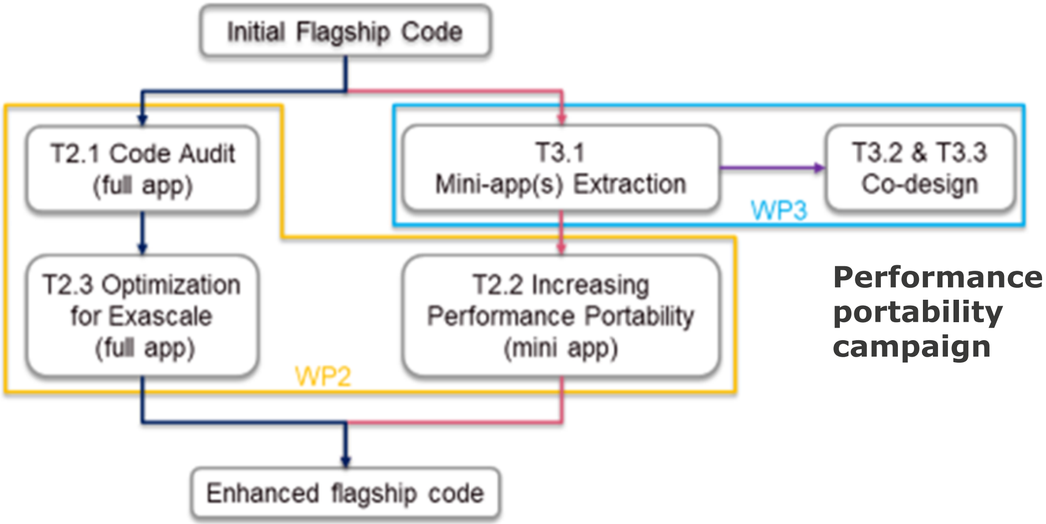
- ❑ **Useful to have a quantitative measure of the performance obtained in a scaling campaign and the particular metric(s) responsible for the loss of performance**

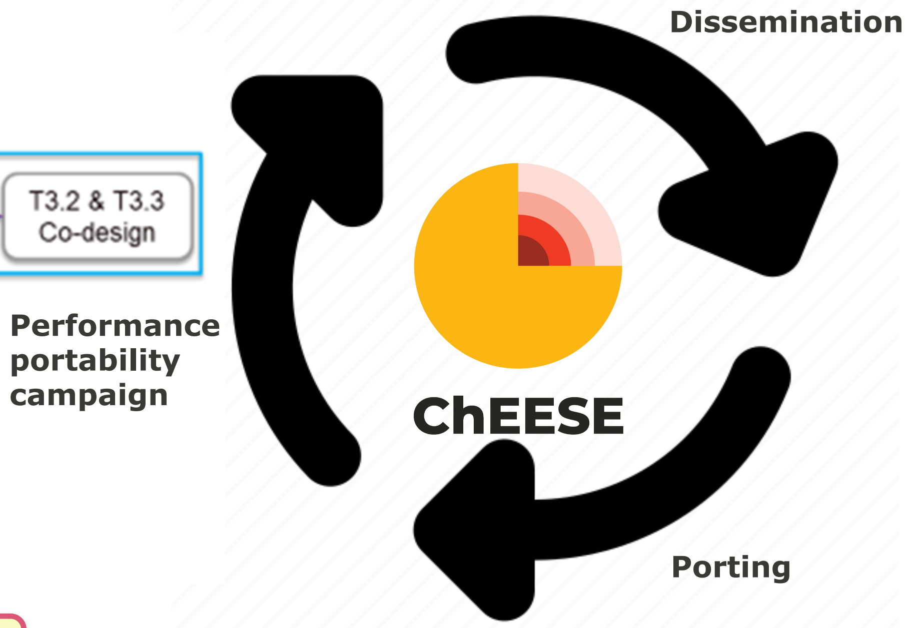


	1	4	8	16	24	32	40	48	56	64
-- Global Parallel efficiency	47.43	47.19	46.80	46.27	45.73	44.48	43.61	43.30	42.34	42.04
-- Global Load balance	50.00	49.91	49.97	49.91	49.92	49.83	49.59	49.64	49.59	49.40
-- GPUs Global efficiency	94.85	94.08	92.94	91.03	89.33	85.99	83.83	82.77	80.17	79.00
-- GPUs Parallel efficiency	94.85	94.33	93.53	92.39	91.23	88.67	86.87	86.18	84.19	83.52
-- GPUs Load Balance	100.00	99.79	99.86	99.65	99.59	99.33	98.77	98.80	98.62	98.15
-- GPUs Communication efficiency	94.85	94.53	93.66	92.71	91.61	89.27	87.95	87.23	85.37	85.09
-- GPUs Computation scalability	100.00	99.73	99.37	98.53	97.92	96.98	96.50	96.05	95.22	94.59

Percentage(%): 0, 20, 40, 60, 80, 100

T2.2 Increasing the performance portability of flagship codes for different accelerators: the methodology





D2.2 M24

D2.4

T2.2 Increasing the performance portability of flagship codes for different accelerators: dissemination

❑ Online events on performance portabilility tools and techniques:

1. *13/06/23 webinar: Porting to GPU a Fortran code using intrinsic parallelism*
2. *28/06/23 webinar: OpenCL GPU Programming for HPC Applications*
3. *07/07/23 webinar: Kokkos*
4. *21/12/23 webinar: Device accelerated solvers with PETSc*

❑ Other webinars will be planned (date to be fixed). So far:

1. *GPU Performance portability for directive-based approach*
2. *MPI GPU-aware communications: some figures and examples from Leonardo*
3. *SYCL programming*
4. *…*

T2.2 Increasing the performance portability of flagship codes for different accelerators: porting

❑ **Goal: use different strategies for increasing the performance portability in flagship codes (or available mini-apps):**

1. *Directive based tools (OpenMP offload/OpenACC)*
2. *Intrinsic language parallelism (stdpar for C/C++, DO CONCURRENT Fortran, …)*
3. *Performance Portable programming model (Kokkos, Raja, SYCL, …)*
4. *Performance Portable Libraries (PETSc, …)*

❑ **Implementation**

1. *WP2 Workplans (by code owners)*
2. *Creation of a development team acting on flagship codes (mini-apps)*
3. *Porting campaign: your mileage may vary (see https://x-dev.pages.jsc.fz-juelich.de/models/ for details)*

OpenMP

kokkos

Fortran
DO CONCURRENT

SYCL™

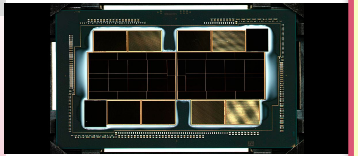
T2.2 Increasing the performance portability of flagship codes for different accelerators: performance portability campaign

- ❑ **Goal: make an “ad-hoc” campaign for effectively measure the performance portability in flagship codes (or available mini-apps). Ingredients:**

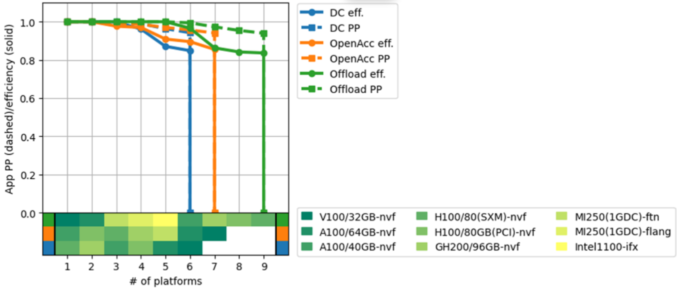
1. *Computational budget on EuroHPC available (accelerated) hardware or even Software Development Vehicles (SDV)*
2. *Accelerated flagship codes (mini-apps) from T2.2 “arena”*
3. *Suitable performance portability metrics from literature to monitor the Performance Portability in different scenarios (see for example [Navigating Performance, Portability, and Productivity | IEEE Journals & Magazine | IEEE Xplore](#))*

IMPLEMENTATION

1. *Performance portability campaign calendarization (by code)*
2. *Performance portability campaign*
3. *Performance portability campaign outcomes (to be used in deliverable D2.2). Lessons learned to flagship codes.*



PPC



T2.3 Optimisation for Exascale

This task guarantee technical support to each flagship code to enhance its performance and scalability.

First set of observations and recommendations from the code audits at M12

D2.1

This task address primarily the following features (not an exhaustive list):

- **single heterogeneous node performance**
- **multi-node performance**
- **algorithmic improvements**
- **resilience and fault tolerance**
- **I/O performance**

Exascale features

Other improvements

Adequacy of flagship codes to EuroHPC heterogeneous hardware

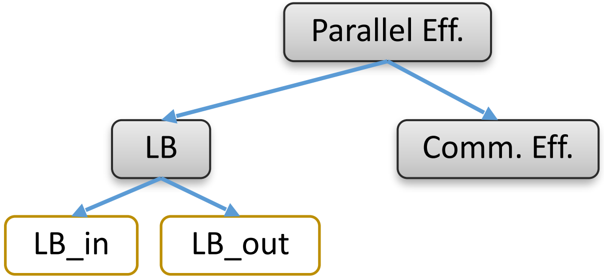
D2.2 M24

D2.4

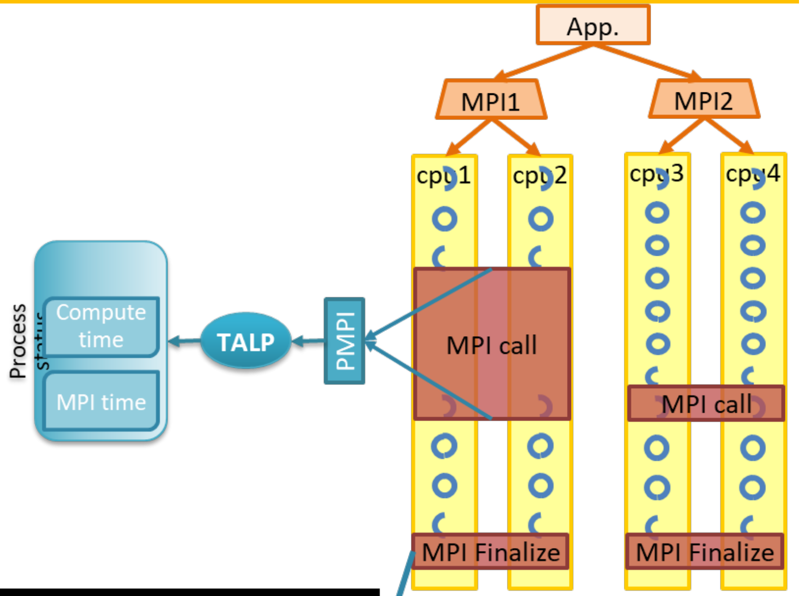
T2.4 Continuous Performance Efficiency Monitoring

- **Goal:** Provide lightweight, straightforward, and easy-to-use mechanism to measure the parallel efficiency of flagship codes at Exascale.
- Based on TALP that gathers POP metrics at runtime, without tracing.

Current implementation support MPI metrics (and something else)



Development of new features in ChEESE-2P (e.g. GPUs support)




```
######### Monitoring Region App Summary #########
### Name:                              MPI Execution
### Elapsed Time :                     2.66 s
### Parallel efficiency :              0.70
###   - Communication eff. :           0.98
###   - Load Balance :                 0.72
###       - LB_in :                    0.72
###       - LB_out:                    1.00
```

Outline

01 ChEESE-2P overview

02 The ChEESE-2P WP2 strategy

03 SPECFEM3D showcase

SPECFEM3D

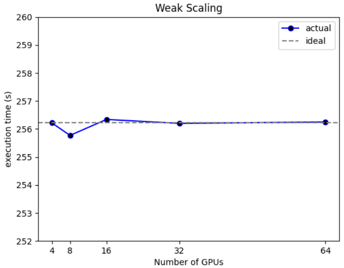
https://specfem.org/

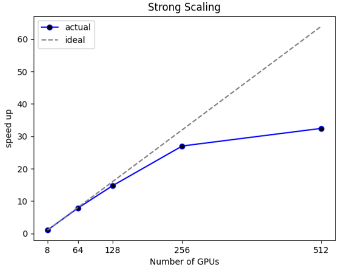
- ***Spectral-element method:***
  - ***CPU, GPU, MPI (domain decomposition)***

- ***Solves linear seismic wave propagation in 3D models :***
  - ***Elastic, viscoelastic, piroelastic, and fluid solid interactions***
  - ***Implements imaging and Full Waveform Inversion for complex models***

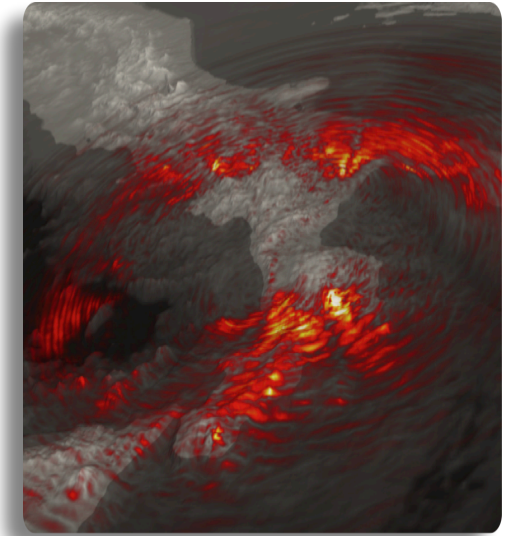
- ***Scalability evaluated on Leonardo booster:***
  - ***Weak scaling : from 4 to 64 GPUs, 360448 mesh elements per GPU***
    excellent weak scaling up to 64 GPUs (and likely beyond)
  - ***Strong scaling :***
    good strong scaling up to 256 GPUs
    overlapping MPI communication with GPU computation

SPECFEM3D has been optimized for various GPU architectures (using primarily CUDA and HIP as GPU programming model).

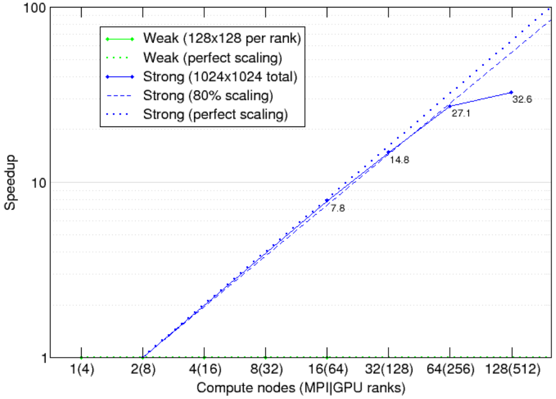




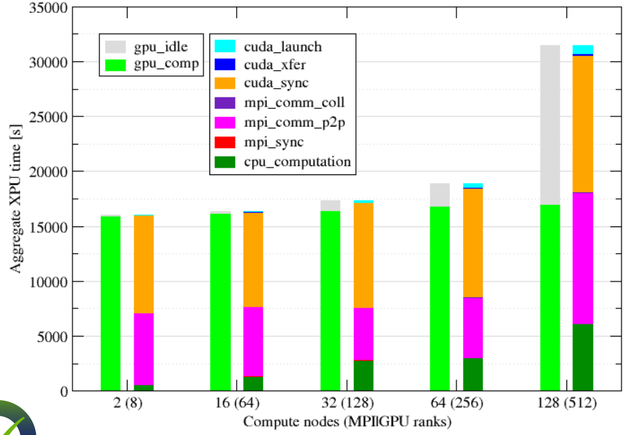
Performance Optimisation
and Productivity



SPECFEM3D



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



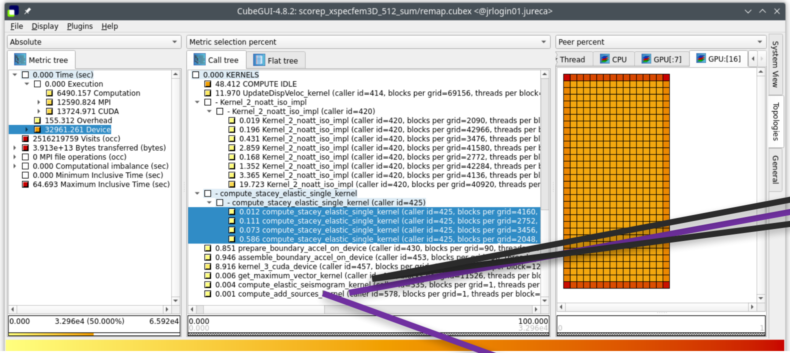
POP3_AR_002

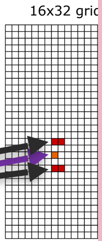
- ***iterate_time (solver) chosen as Focus of Analysis***
- ***Good strong scaling up to 256 GPUs***
- ***With 512 GPUs no longer able to sufficiently overlap MPI communication with CUDA kernels***

Performance Optimisation and Productivity

https://zenodo.org/records/13643996

SPECFEM3D



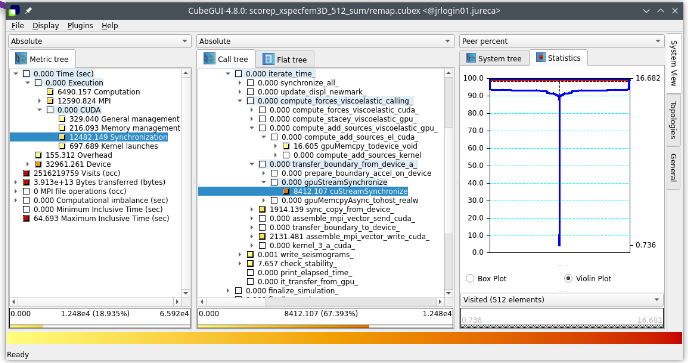


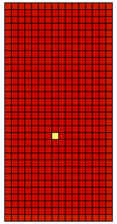
2D domain decomposition on GPUs

- *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*

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**





Performance Optimisation and Productivity

Using SPECFEM3D to address Scientific Challenge

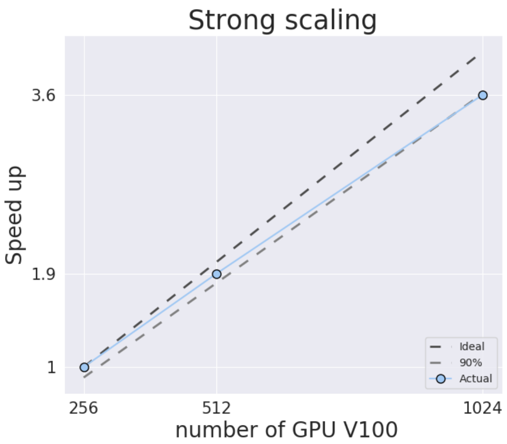
SPECFEM3D used to numerically assess the underwater noise pollution and seismic risks induced by the unexploded ordnance disposal in the Bay of Hyères (France).

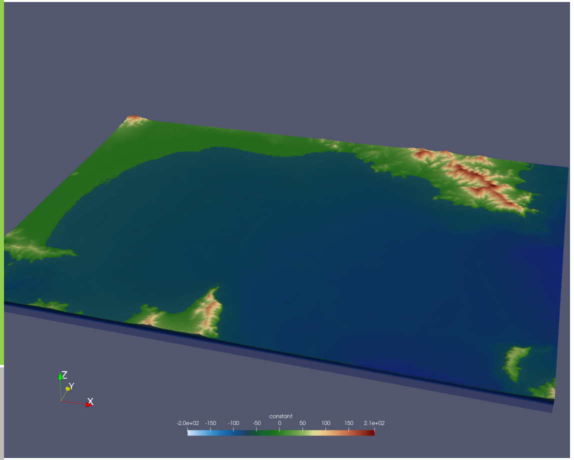
- ***Sismo-acoustic modelization : fluid (acoustic) / solid (viscoelastic)***
- ***Test case : - Coastal zone of the Rade d'Hyère (Southern France)***
- ***Constraints :***

1. ***Computation up to 30 Hz***
2. ***Sedimentary basin → very low shear-wave velocity → small size elements (something like 6m in sediments)***

- ***Size: 27 millions of elements (interpolation polynomials degree 7)***
- ***2 millions of time steps***

SPECFEM3D modelization already performed on Petascale machine (CINECA Marconi 100)





Using SPECFEM3D to address Scientific Challenge

The video shows the vertical particle velocity on sea bottom

- The movie is related to the shaking of this surface while seismic waves propagate. The color is related to the intensity.
- Seismic wave propagation are triggered by an explosion on the middle of the zone and at the sea bottom.
- The goal is to estimate the maximum shaking we can expect in order to anticipate some damages in buildings at the coast (this area is densely populated) upstream the real explosions that are operated by French Navy.

Using SPECFEM3D to address Scientific Challenge

Scientific Challenge planned in ChEESE-2P: increase the accuracy of the simulation

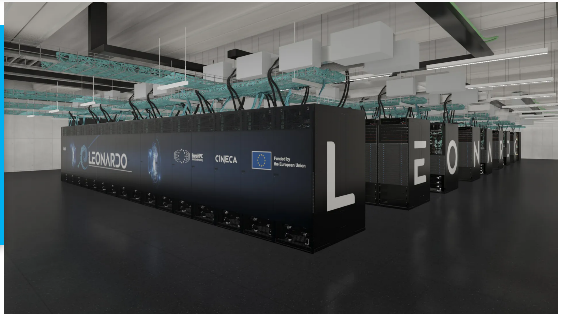
- Passing from 30 to 100 Hz reduce the element size by a factor 3.33
- With 100 Hz, 1 billion of elements and 4 millions time step are needed

Also the total memory needed for the simulation increase by a factor of 74 and the run time modeling a 100 second signal increase of a factor 123

The ultimate goal of the research requires significant computational resources, because of the domain size (about 20x20x5 km), the high upper limit of the explosion frequency range (300 Hz), and the low shear-wave velocity in sediments (200 m/s). This results in large meshes and very small time steps, pushing current supercomputers to their limits.

Exascale needed

Use of the improved performance portable version of SPECFEM3D from ChEESE-2P



Conclusions

- ChEESE and POP have established a very productive collaboration
- Many codes audited during campaigns (10 in ChEESE-1P and 11 in ChEESE-2P)
- ChEESE-2P makes use of audits to:
  - Identify code bottlenecks (audit-driven optimisations)
  - Establish a code performance baseline
  - Continuous code monitoring based on POP metrics (including TALP)
- Additionally, some ChEESE codes have started making use of POP SLS
- Additional takeaways:

❑ POP tools needs to (continue to) improve their functionalities and evolve (with hw/sw)

❑ Metrics (performance, others?) are the key of success

Thank you!

@cheese-coe

http://cheese2.eu

@cheese-coe

@cheese_coe@techhub.social