



ChEESE

Work Package 2:
Exascale technical challenges in flagship codes

Project kick-off meeting Barcelona,
19-20 January 2023

Piero Lanucara
CINECA



Project funded by EuroHPC under the grant agreement No 101093038.





Objectives

- | | |
|---|--|
| 1 | Address technical challenges (potentially at exascale) in the 11 flagship codes |
| 2 | Exploit actual accelerators , in particular those targeting EuroHPC pre-exascale machines |
| 3 | Audit the codes and improve single node, multi-node, and I/O performances at scale. |
| 4 | Continuously monitor the code performance efficiency during the project |



Tasks

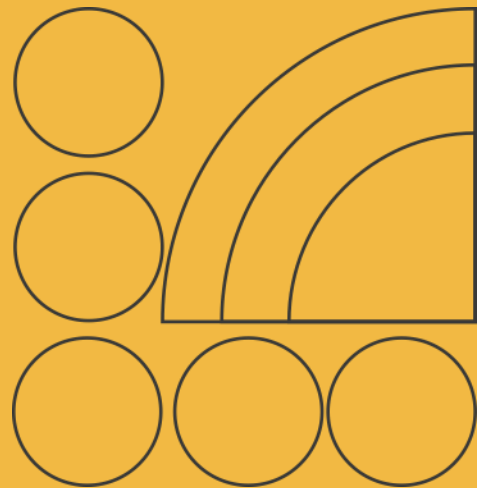
Task	Task name	Start	End	Lead
Task 2.1	Code audits	3 (27)	15 (36)	HLRS
Task 2.2	Increasing the performance portability of flagship codes for different accelerators	1	48	CIN
Task 2.3	Optimisation for exascale	6	48	CIN
Task 2.4	Continuous Performance Efficiency Monitoring	18	48	BSC

Effort (PMs)													
CIN	CSIC	BSC	INGV	HLRS	TUM	LMU	UMA	CNRS	IPGP	BULL	CSC	UM	SU
67	12	18	18	18	22	14	12	28	1	14	8	8	7

Deliverables

D	Deliverable name	Lead	Type	Level	M	Date
2.1	First code audit report	HLRS	R	PU	12	Dec 2023
2.2	First report on improving code performance at exascale	CIN	R	PU	24	Dec 2024
2.3	Second code audit report	HLRS	R	PU	30	Jun 2025
2.4	Final report on improving code performance at exascale and Best Practice Guide	CIN	R	PU	48	Dec 2026

Flagship codes



Initial status

In terms of (strong) scalability and parallel performance, we define the following criterion for the "code level":

- **Level 0 codes.** Codes able to perform full capability runs or capacity workloads on exascale machines ("future tier-0 systems" of the upcoming exascale era). Good scalability (>90% efficiency) proven in more than 10000 nodes (> 10^5 - 10^6 cores). No tests at this scale exist yet.
- **Level 1 codes.** Codes able to perform full capability runs or capacity workloads on current pre-exascale machines ("future tier-1 systems" of the upcoming exascale era). Good scalability (>90% efficiency) proven in more than 1000 nodes (> 10^4 - 10^5 cores). Up to 5 ChEESE flagship codes are considered to be already at or close to level 1 (SeisSol, SPECfem3D, xSHELLS, LaMEM and pTatin3D).
- **Level 2 codes.** Codes able to perform full capability runs or capacity workloads on current petascale machines ("future tier-2 systems" of the upcoming exascale era). Good scalability (>90% efficiency) proven in more than 100 nodes (> 10^3 - 10^4 cores). Up to 6 ChEESE flagship codes are considered to be already at or close to level 2 (ExaHyPE, Tandem, HySEA, FALL3D, OpenPDAC and Elmer/Ice).

No.	Code	Initial Level	Area	Lead
1	SeisSol	1	Computational seismology Tsunami modeling	LMU
2	SPECfem3D	1	Computational seismology	CNRS
3	ExaHyPE	2	Computational seismology Tsunami modeling	TUM
4	Tandem	2	Computational seismology	LMU
5	xSHELLS	1	Magnetohydrodynamics	CNRS
6	HySEA	2	Tsunami modeling	UMA
7	FALL3D	2	Volcanology	CSIC
8	OpenPDAC	2	Volcanology	INGV
9	LaMEM	1	Geodynamics	UM
10	pTatin3D	1	Geodynamics	SU
11	Elmer/Ice	2	Glacier hazards	CSC

Initial status

In terms of (strong) scalability and parallel performance, we define the following criterion for the "code level":

- **Level 0 codes.** Codes able to perform full capability runs or capacity workloads on exascale machines ("future tier-0 systems" of the upcoming exascale era). Good scalability (>90% efficiency) proven in more than 10000 nodes (> 10^5 - 10^6 cores). No tests at this scale exist yet.
- **Level 1 codes.** Codes able to perform full capability runs or capacity workloads on current pre-exascale machines ("future tier-1 systems" of the upcoming exascale era). Good scalability (>90% efficiency) proven in more than 1000 nodes (> 10^4 - 10^5 cores). Up to 5 ChEESE flagship codes are considered to be already at or close to level 1 (SeisSol, SPECfem3D, xSHELLS, LaMEM and pTatin3D).
- **Level 2 codes.** Codes able to perform full capability runs or capacity workloads on current petascale machines ("future tier-2 systems" of the upcoming exascale era). Good scalability (>90% efficiency) proven in more than 100 nodes (> 10^3 - 10^4 cores). Up to 6 ChEESE flagship codes are considered to be already at or close to level 2 (ExaHyPE, Tandem, HySEA, FALL3D, OpenPDAC and Elmer/Ice).

No.	Code	Initial Level	Area	Lead
1	SeisSol	1	Computational seismology Tsunami modeling	LMU
2	SPECfem3D	1	Computational seismology	CNRS
3	ExaHyPE	2	Computational seismology Tsunami modeling	TUM
4	Tandem	2	Computational seismology	LMU
5	xSHELLS	1	Magnetohydrodynamics	CNRS
6	HySEA	2	Tsunami modeling	UMA
7	FALL3D	2	Volcanology	CSIC
8	OpenPDAC	2	Volcanology	INGV
9	LaMEM	1	Geodynamics	UM
10	pTatin3D	1	Geodynamics	SU
11	Elmer/Ice	2	Glacier hazards	CSC

Initial status

In terms of (strong) scalability and parallel performance, we define the following criterion for the "code level":

- **Level 0 codes.** Codes able to perform full capability runs or capacity workloads on exascale machines ("future tier-0 systems" of the upcoming exascale era). Good scalability (>90% efficiency) proven in more than 10000 nodes (> 10^5 - 10^6 cores). No tests at this scale exist yet.
- **Level 1 codes.** Codes able to perform full capability runs or capacity workloads on current pre-exascale machines ("future tier-1 systems" of the upcoming exascale era). Good scalability (>90% efficiency) proven in more than 1000 nodes (> 10^4 - 10^5 cores). Up to 5 ChEESE flagship codes are considered to be already at or close to level 1 (SeisSol, SPECfem3D, xSHELLS, LaMEM and pTatin3D).
- **Level 2 codes.** Codes able to perform full capability runs or capacity workloads on current petascale machines ("future tier-2 systems" of the upcoming exascale era). Good scalability (>90% efficiency) proven in more than 100 nodes (> 10^3 - 10^4 cores). Up to 6 ChEESE flagship codes are considered to be already at or close to level 2 (ExaHyPE, Tandem, HySEA, FALL3D, OpenPDAC and Elmer/Ice).

No.	Code	Initial Level	Area	Lead
1	SeisSol	1	Computational seismology Tsunami modeling	LMU
2	SPECfem3D	1	Computational seismology	CNRS
3	ExaHyPE	2	Computational seismology Tsunami modeling	TUM
4	Tandem	2	Computational seismology	LMU
5	xSHELLS	1	Magnetohydrodynamics	CNRS
6	HySEA	2	Tsunami modeling	UMA
7	FALL3D	2	Volcanology	CSIC
8	OpenPDAC	2	Volcanology	INGV
9	LaMEM	1	Geodynamics	UM
10	pTatin3D	1	Geodynamics	SU
11	Elmer/Ice	2	Glacier hazards	CSC

Initial status

In terms of generic HPC main code features:

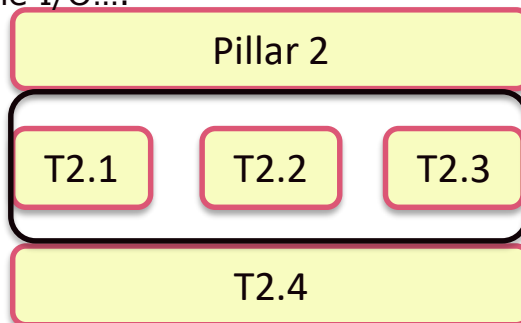
- **MPI or MPI+X** where X is equal to:
 - **OpenMP** (mostly) for CPUs (with some exception)
 - **CUDA, OpenACC (but also HIP)** for accelerators
- C, C++, Fortran (also Python) as selected programming languages

No.	Code	Initial Level	Area	Lead
1	SeisSol	1	Computational seismology Tsunami modeling	LMU
2	SPECFEM3D	1	Computational seismology	CNRS
3	ExaHyPE	2	Computational seismology Tsunami modeling	TUM
4	Tandem	2	Computational seismology	LMU
5	xSHELLS	1	Magnetohydrodynamics	CNRS
6	HySEA	2	Tsunami modeling	UMA
7	FALL3D	2	Volcanology	CSIC
8	OpenPDAC	2	Volcanology	INGV
9	LaMEM	1	Geodynamics	UM
10	pTatin3D	1	Geodynamics	SU
11	Elmer/Ice	2	Glacier hazards	CSC

Preparation strategy (Pillar 2)

The transversal WP2 will deal with the technical challenges (including porting, optimization, audit and performance monitoring) devoted to the preparation of the flagship codes to Exascale. Those challenges will consider node heterogeneity and different levels of parallelism on complex hardware architectures and will require a development and optimization methodology to guarantee a full utilization of the HPC resources available. Two developments Tasks:

- ❑ T2.1 Code audits
- ❑ T2.2 “accelerator arena”, with a clear focus on EuroHPC hardware (Lumi, Leonardo, MN5) and associated GPUs (AMD, NVIDIA but also Intel GPUs...)
- ❑ T2.3 “generic” audit-driven development and tuning activities addressing known or unexpected code issues, starting from the single-node level and passing through the multi-node scaling, the I/O....



Adequacy of flagship codes to
EuroHPC heterogeneous
hardware



Preparation strategy (Pillar 2). Task T2.1

Task T2.1: Code audits

❑ **Lead: HLRS**

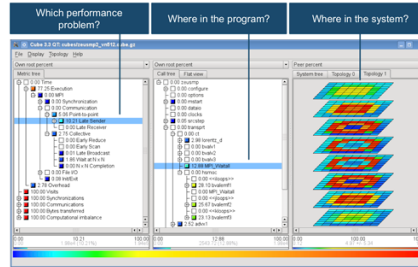
❑ **Partners: CSIC, INGV, CIN, TUM, LMU, UMA, CNRS, IPGP, CSC, UM, SU**

Goals:

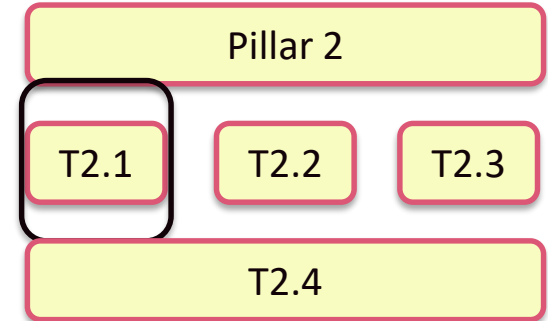
- ❑ Set performance baseline for optimization work and progress monitoring
- ❑ Identify performance bottlenecks

Method (**from DoW**):

- ❑ Two performance audit campaigns (M3-M15, M27-M36)
- ❑ Actual performance audit and analysis done by POP3
- ❑ Target EuroHPC systems such as Leonardo, Lumi, ...
- ❑ Main effort is organize POP campaigns and prepare codes for analysis



Task T2.1 outcomes will (partly) feed Task T2.2 and T2.3 and will contribute to Deliverables D2.1 and D2.3



Adequacy of flagship codes to
EuroHPC heterogeneous
hardware



Preparation strategy (Pillar 2). Task T2.2

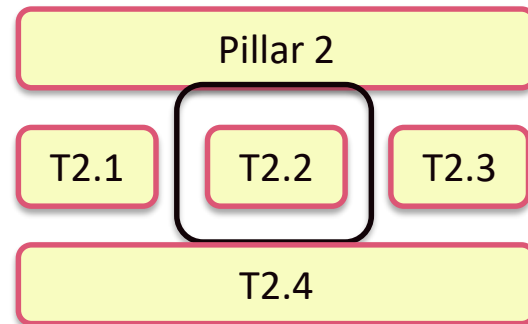
Task T2.2: Increasing the performance portability of flagship codes for different accelerators.

❑ **Lead: CINECA**

❑ **Partners: CSIC, INGV, TUM, LMU, UMA, CNRS, IPGP, CSC, UM, SU**

Task T2.2 will contribute to increase the “performance portability” of flagship codes in the “accelerator arena” (focus on EuroHPC heterogeneous hardware) using a

- ❑ clear and rigorous methodology based on available literature (including a suitable metric for that target)
- ❑ performance portable software tools and libraries (including OpenMP, OpenACC, SYCL, Kokkos....)
- ❑ new development team (lead by CINECA), acting on flagship codes



OpenMP

SYCL



kokkos

Adequacy of flagship codes to
EuroHPC heterogeneous
hardware

Task T2.2 will (partly) depends on the output of Task T2.1 and will contribute to Deliverables D2.2 and D2.4



Preparation strategy (Pillar 2). Task T2.3

Task T2.3: Optimisation for Exascale.

❑ **Lead: CINECA**

❑ **Partners: CSIC, BSC, INGV, TUM, LMU, UMA, BULL, UM, SU**

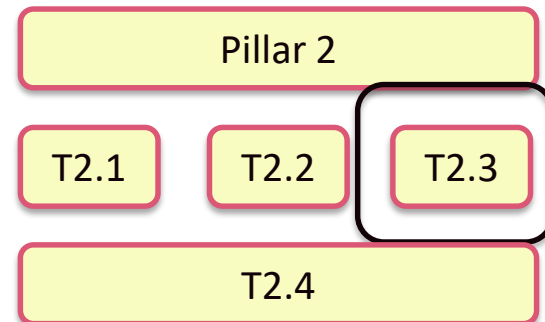
Task T2.3 will guarantee technical support to each flagship code to enhance its performance and scalability.

This task will 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

Task T2.3 will depends on the output of Task T2.1 and will contribute to Deliverables D2.2 and D2.4



Adequacy of flagship codes to EuroHPC heterogeneous hardware



Monitoring performances: TALP in ChEERE-2P

Task T2.4: Continuous Performance Efficiency Monitoring

❑ Lead: BSC

❑ Partners: CSIC, INGV, HLRS, TUM, LMU, UMA, CNRS, IPGP, UM

01

A tool for codes

Integrate TALP with the different codes.

To allow a continuous performance monitoring.

Can also be integrated with CI/CD platforms

02

Extending TALP

Extend TALP to address the needs of ChEERE-2P codes:

- ... to collect PAPI HWC counters
- ... with OpenMP hybrid metrics
- ... with GPU metrics

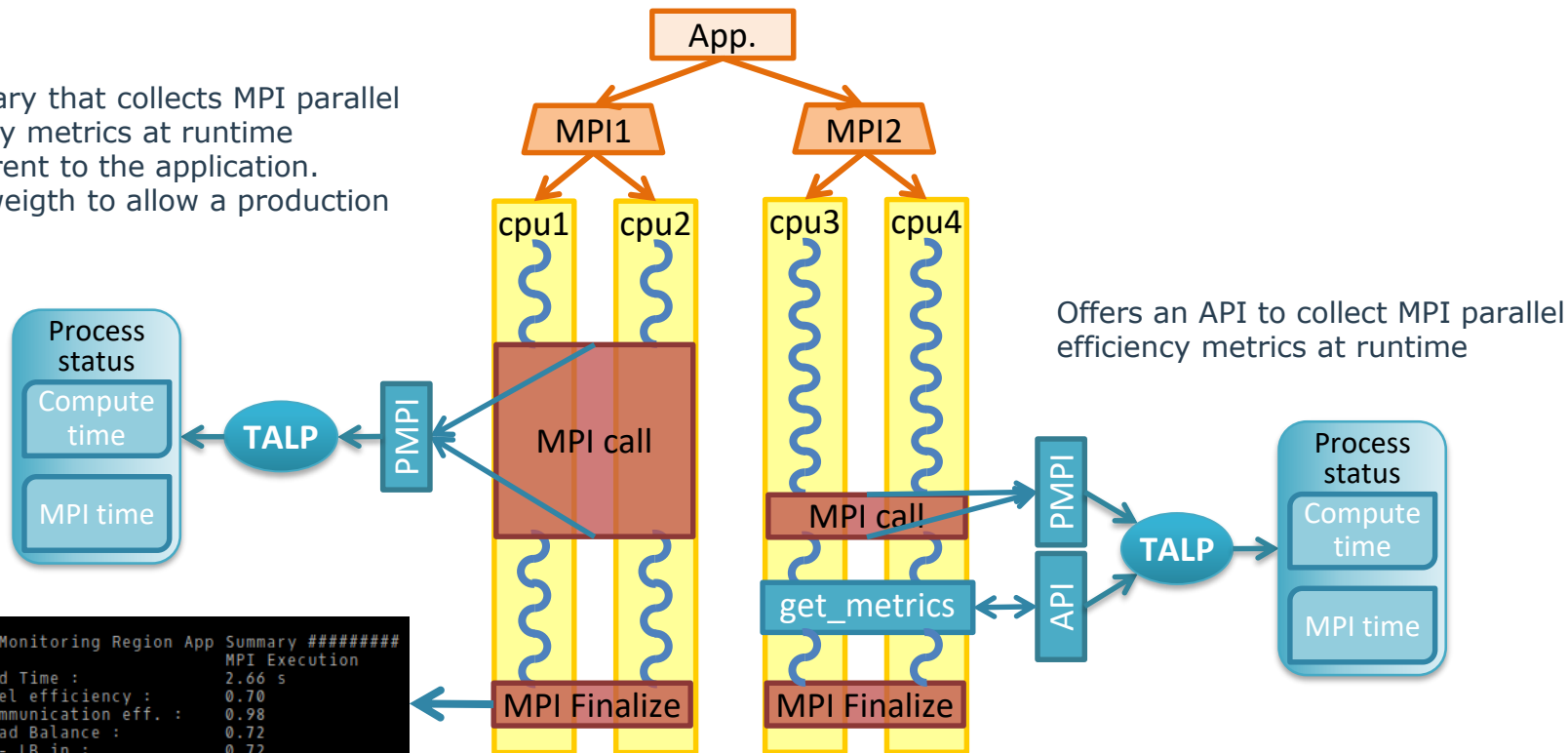


TALP

Is a library that collects MPI parallel efficiency metrics at runtime transparent to the application.
Is lightweight to allow a production use.

```
##### 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
```

At finalization prints a report with the MPI parallel metrics

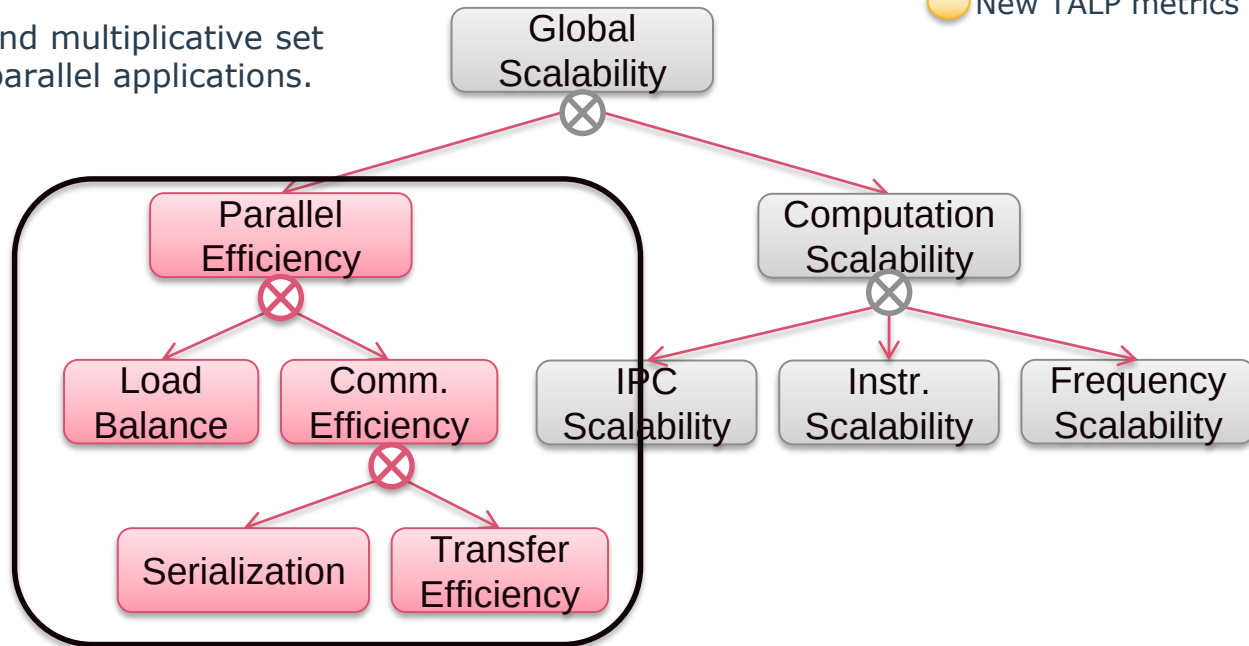


Lopez, V., Ramirez Miranda, G., & Garcia-Gasulla, M. (2021). TALP: A Lightweight Tool to Unveil Parallel Efficiency of Large-Scale Executions. In Proceedings of the 2021 on PERMAVOST (pp. 3-10).

POP efficiency metrics

Are a hierarchical and multiplicative set of metrics for MPI parallel applications.

- POP metrics collected by TALP
- POP metrics NOT collected by TALP
- New TALP metrics

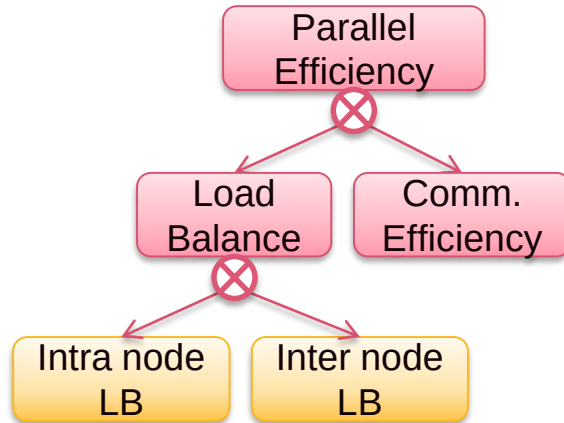


- The metrics are represented as a percentage that corresponds to the amount of time lost due to that inefficiency. I.e. A Communication efficiency of 70% means that 30% of the time in that execution is lost due to communication.
- Scalability metrics are computed with respect to a base case. I.e. in the case of just one sample/trace they are 100%



TALP efficiency metrics

Is a subset of the POP metrics
plus two additional ones



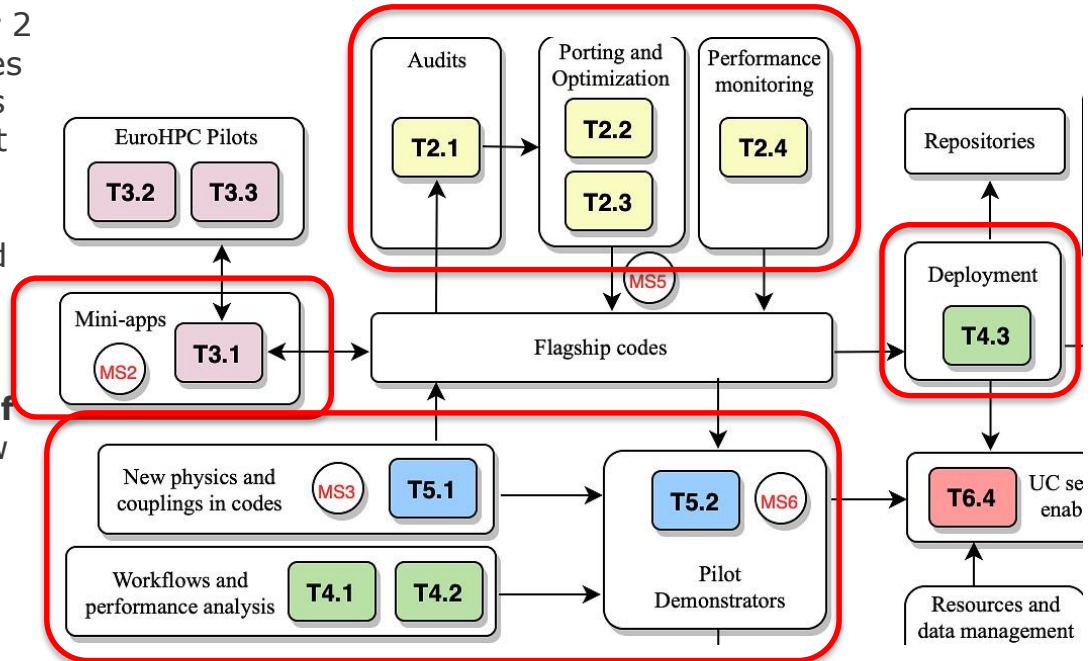
- POP metrics collected by TALP
- New TALP metrics



The complete “Big Picture” and WP2 external dependencies

WP2 will take care of Pillar 2 (adequacy of flagship codes to EuroHPC heterogeneous hardware) and will interact with:

- ❑ **Task T3.1 of WP3** for Mini-apps extraction and preparation
- ❑ **Task T4.3 of WP4** for deployment
- ❑ **Tasks T5.1 and T5.2 of WP5** for addressing new development in flagship codes and Pilot Demonstrators



WP2 will contribute to the fulfillment of milestone MS5 at month 36

Milestone No.	Milestone Name	Lead beneficiary	Means of verification	Due date (month)
MS5	Exascale code preparation	CINECA	Improvements in the flagship codes performance metrics after the second code audit round (D2.3)	36

Wrap-up and & Next Action items

The first WP2 targets are:

- ➡ ☐ Deliverable 2.1 (M12): First code audit report
- ☐ Milestone MS5 (M36): exascale code preparation

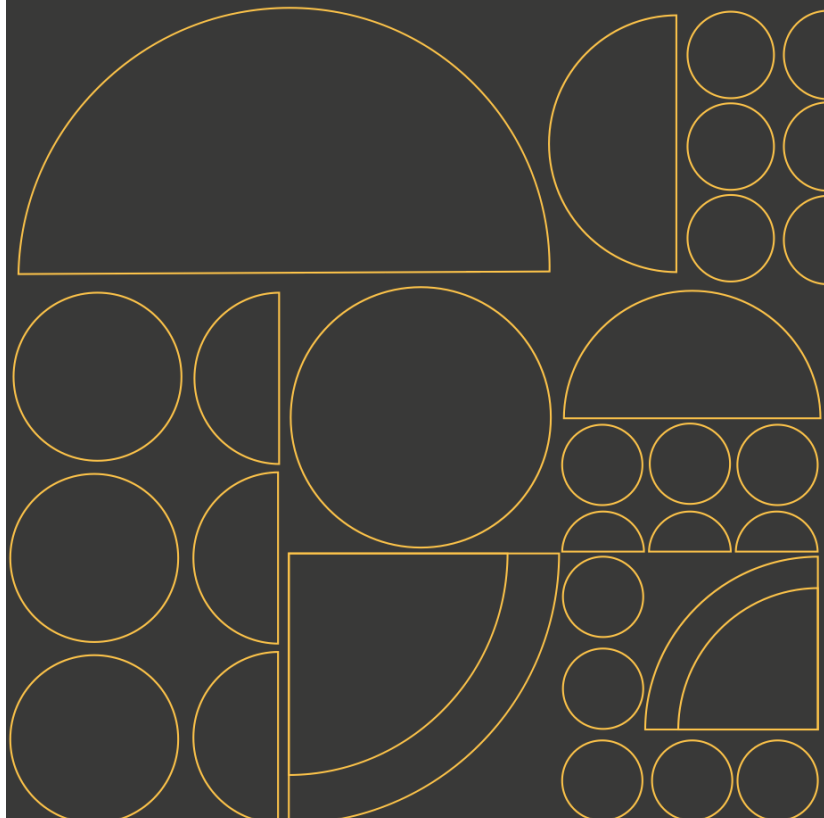
Action items (TBD)

- ☐ **Setting a doodle for the monthly meeting**
- ☐ **Define a presentation roadmap per flagship code (mainly T2.2 related ...)**
- ☐ **Setting up the access to the available hardware for first code audit**
- ☐ **Setting up the WP2 main contact points and responsibilities**
- ☐ **Defining a ChEESA-2P git repository (not only for the WP2?)**
- ☐ **T2.1 “issue” (next slide)**

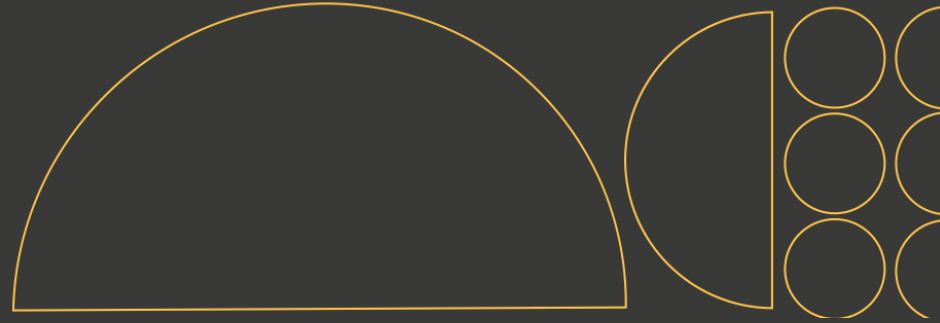


Wrap-up and Next Action items: T2.1 “issue”

- ❑ POP3 has not been funded.
ChEESE will have to do code audits on its own.
- ❑ Good: ChEESE has sufficient expertise in code auditing (BSC, HLRS are in POP)
- ❑ Problem: main effort shifts to collecting performance data and doing analysis
- ❑ Suggestions (TBD):
 - **Continuous effort throughout project rather than 2 distinct campaigns**
 - **More effort for code owners to collect performance data**
 - **Very light-weight initial audit for POP metrics. Code audit only where necessary.**
How to prioritize?



Thank you!



@cheese-coe



<http://cheese2.eu>



@cheese-coe



@cheese_coe@techhub.social

