POP CoE
SEM46 Audit (POP2_AR_033)

Judit Gimenez (judit@bsc.es) , BSC - Nov 2019

EU H2020 Centre of Excellence (CoE)

1 December 2018 – 30 November 2021

Grant Agreement No 824080

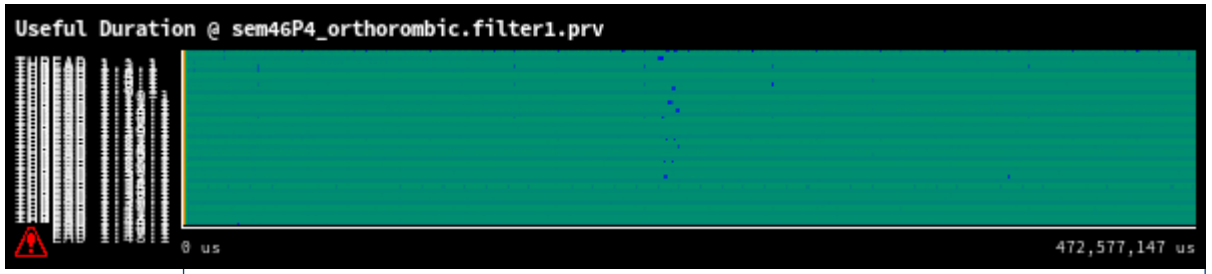
Background

- Applicant: Romain Brossier, Univ. Grenoble Alpes (Core developer)
- Name of the code: SEM46
- Scientific/technical area: Earth and atmospheric sciences
- Programming: Fortran; MPI

- Input case: 2 different inputs from production runs. Each input with 2 configurations (M1 – forward 1 shot, M2 – FWI all waves 1 shot)
- Scale: 48 cores for the small case, 96 and 192 cores for the big case
- Platform: BSC MareNostrum4 (2 x Intel Xeon Platinum 8160 24C at 2.1 GHz per node)
- Initial set-up: 9000 iterations. Reduced to 180 iterations after check
- POP user collected the performance data

Application structure

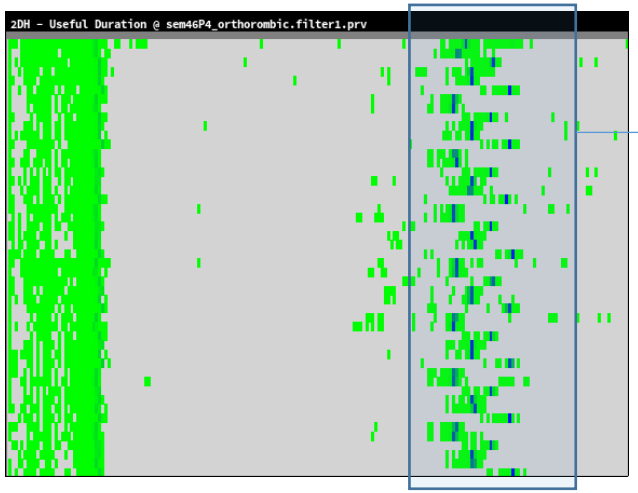
- First analysis with the small input case, 48 MPI, 9000 iterations
  - Phases clearly identified in the tracefile: initialization – iterative computation



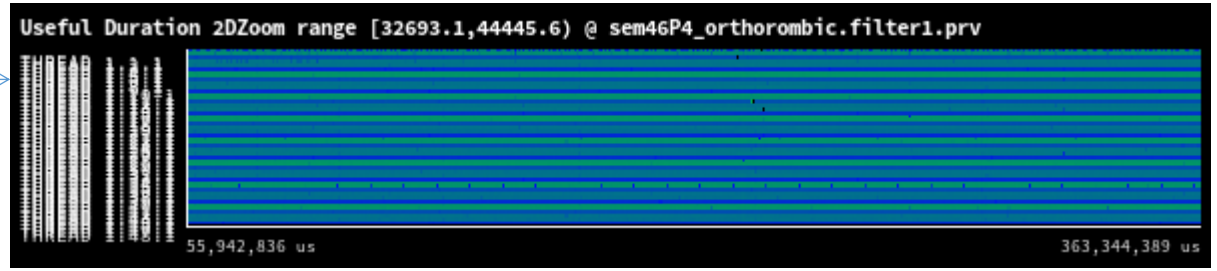
Initialization

Relevant iterative phase

A quick look on this tracefile verified there is no time correlation (similar unbalance in all the iterations) → Can focus the analysis in a smaller number of iterations



Duration of the main computation



Different colors between processes maintained along time

Structured pattern → groups of 6 processes

Focus of Analysis (FOA)

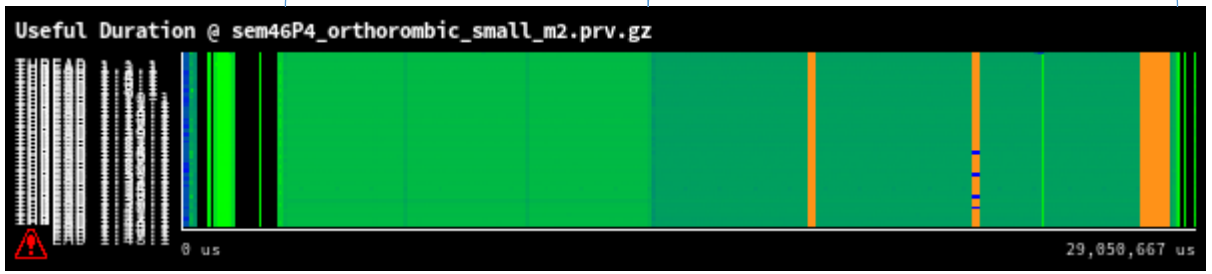
- Selected Focus of Analysis

M1

Initialization FOA

First phase Second phase

M2



Initialization FOA

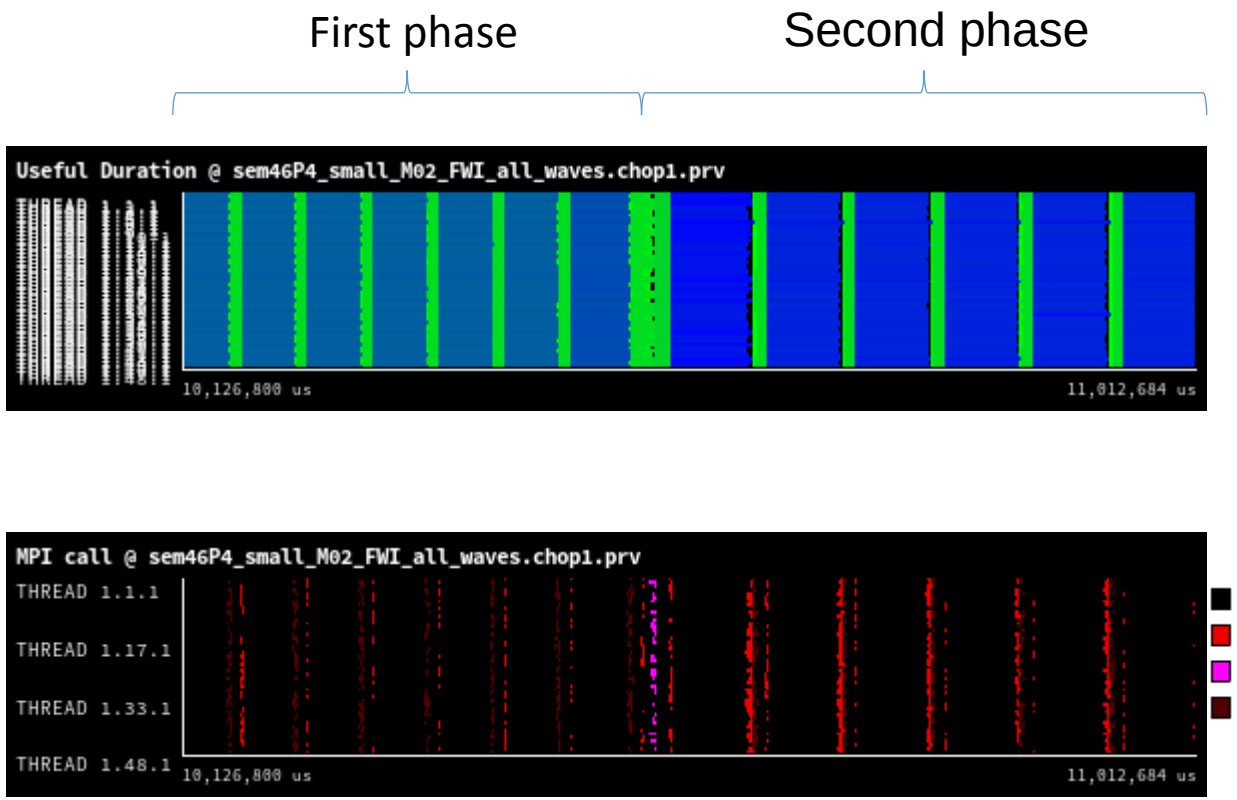
The traces were obtained twice because the first set showed perturbations in one of the cores

The selected focus of analysis is the 180 iterations as the M2 configuration has two phases.

The first phase of M2 configuration is very similar to M1

Focus of Analysis (FOA)

- Zooming into a small area of M2 where we can see the iterations in both phases



With the small input the granularity of the main computations in M1 and M2 phase 1 are around 45 ms while M2 phase 2 goes up to 63 ms. The values grow to 90 ms and 123 ms on the big case → the two input cases cannot be used to analyse scaling

The communications are done using MPI_Sendrecv() and the iterations are synchronized with MPI_Barrier().

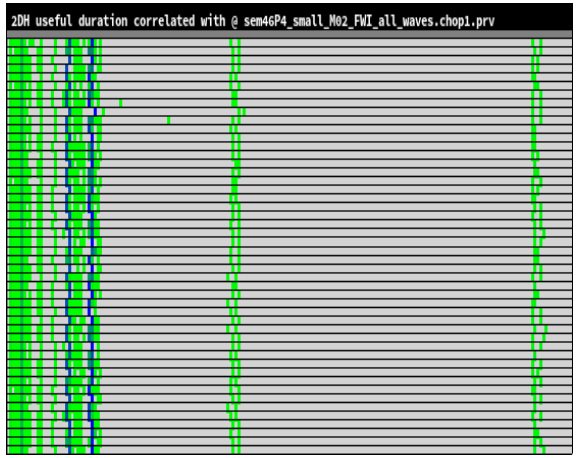
Efficiency model analysis

	small		big	
	M1 - 48	M2 - 48	M1 - 96	M2 - 192
Parallel efficiency	96.28	94.97	96.39	94.48
Load Balance	98.08	97.16	98.74	97.82
Communication eff.	98.16	97.14	97.62	96.59
Serialization	98.95	98.62	98.21	97.27
Transfer	99.20	99.11	99.39	99.30
Average IPC	0.89	0.89	0.91	0.88
Average frequency (GHz)	2.09	2.09	2.09	2.09

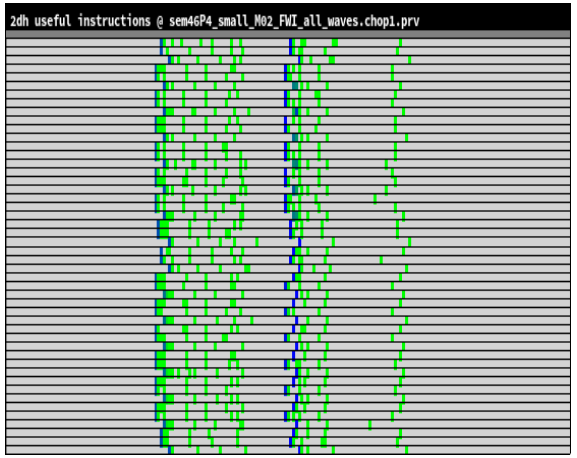
- Efficiencies lower than 80% indicate space for improvement. Lower than 60% there is a clear need for improvement. Good IPC in MareNostrum use to be around 1.2-1.5
- The efficiencies are very good (>= 94.48). Both input cases report lightly worst efficiency for M2 → the degradation may be related with phase 2. Seems to be correlated with load balance and serialization.
- The IPC is a little bit low and should be the target for further analysis/optimizations. The user confirmed there is vectorization of some inner loops with few iterations
- The rest of the report would focus on M2 as M1 is equivalent to M2 phase 1.

Load Balance (M2)

- Focus on small M2 as it is the execution with lower load balance

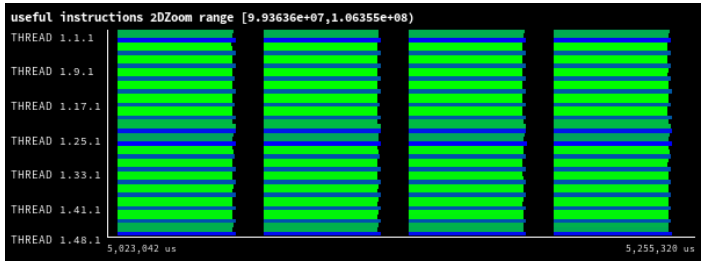


duration

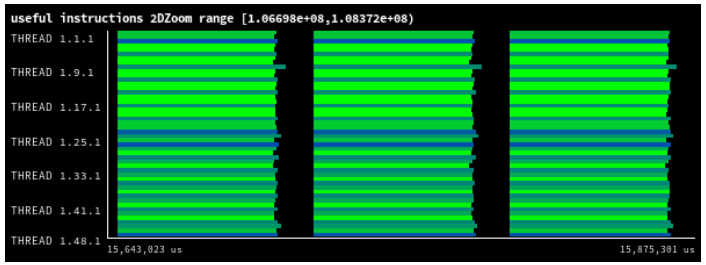


instructions

First phase

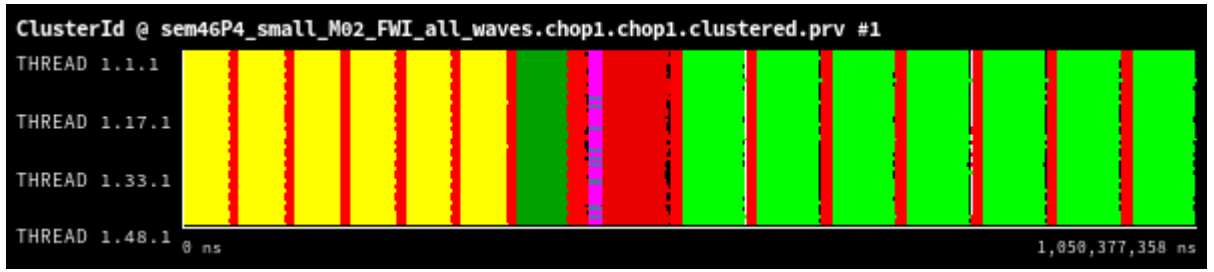


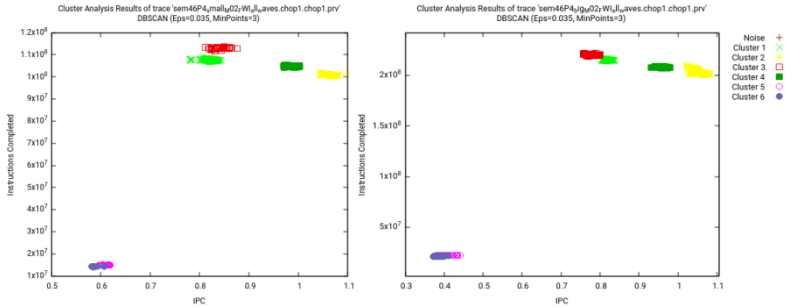
Second phase

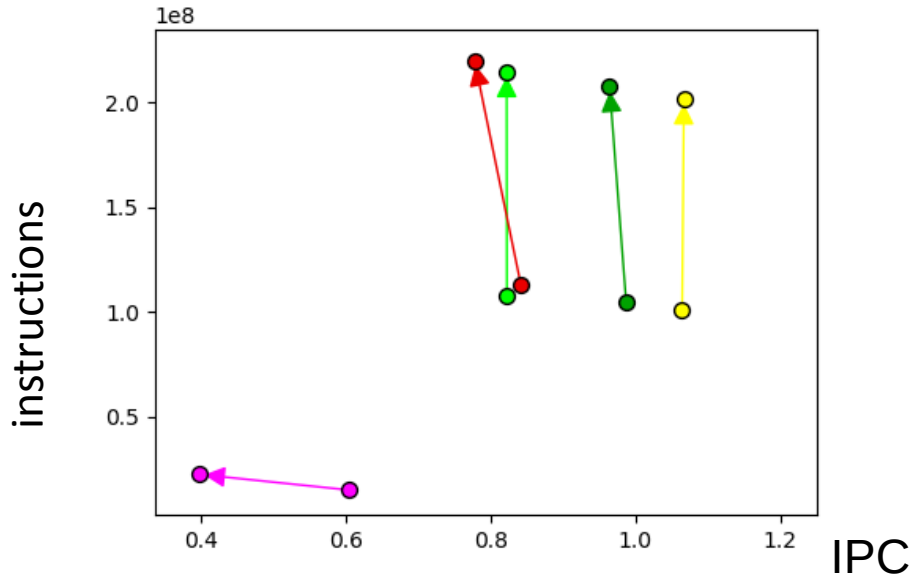


- The duration shows a structured pattern of unbalance (left) that it is correlated with instructions unbalance (right).
- Same structured unbalance appears in both phases (bottom timelines) →domain decomposition?
- In the analysed executions there is no need to improve it (lower unbalance is 97.16)

Computations analysis (M2)



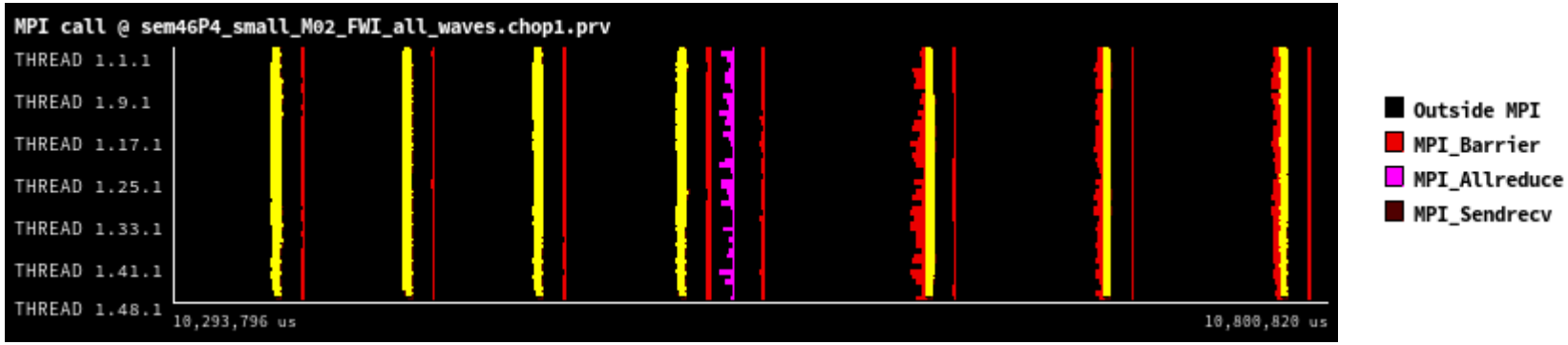




- The same structure is identified in both runs (small and big). The plot in the right show the differences between both inputs. The arrows go from small to big and the increase in the Y dimension reflects the increase on instructions.
- IPC for phase 1 (yellow cluster) seems acceptable while phase 2 (light green) has a lower value → phase 2 should be the first target for optimization
- Cluster pink can be ignored as it represents a very small percentage of time.

MPI analysis (M2)

- Both phases have a similar communication pattern. The only difference is that phase 2 has two calls to MPI_Barrier() per iteration, while phase 1 only has one call.



- As the synchronization is guaranteed by the MPI_Sendrecv() calls, all the calls to MPI_Barrier() can be eliminated in both phases. Nevertheless the impact would be small as the total MPI time is less than 6% in the worst case.
- After eliminating the calls to MPI_Barrier(), the small M1 case was rerun but with a different number of iterations. Very similar results were obtained with differences lower than 1%. Nevertheless the elimination of barriers would improve the execution at very large scales.

Summary of observations

- The audit of SEM46 show very good efficiencies for both input cases (small and big)and the two configurations (M1 and M2). As the typical usage is running multiple configurations on the available resources, there is no need to further stress the set-up running these inputs on a larger scale.
- The analysis identified the weakest factor is a relatively low IPC on the main computations being worst for the second phase of M2. That should be the main target for analysis or optimizations of the code.
- The provided traces did not allowed to do a scalability study because they did not correspond neither to weak nor to strong scaling. The main benefit of a scalability study would be to detect potential code replications when scaling
- A very small unbalance correlated with the work distribution has been detected but the benefits improving it would be very limited.
- The study allowed the user to identify that the calls to MPI_Barrier() were not needed, and despite the limited improvement they represent in the analysed scale, it may make a significant difference when running at very large scales.

Performance Optimisation and Productivity

A Centre of Excellence in HPC

Contact:

https://www.pop-coe.eu

mailto:pop@bsc.es

@POP_HPC

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 676553 and 824080.