PASC19-MiniSymposia Session VI

Friday 14.06

Computational Performance Evaluation for Hardware and Software Alternatives to increase the HPC efficiency of Earth System Models

Organizers: Mario Acosta (Barcelona Supercomputing Center) and Tim Whitcomb (U.S. Naval Research Laboratory)

Presentations: 10:30 - 11:00	New Methodologies for Computational Performance Evaluation of Climate and Weather Models
10.50 - 11.00	Authors: Mario Acosta, Miguel Castrillo, Stella Paronuzzi, Kim Serradell, Oriol Tinto, Xavier Yepes
11:00 - 11:30	Improving Ocean Model Computational Performance by using Mixed-Precision Approaches
	Authors: Oriol Tinto, Mario C. Acosta, Miguel Castrillo, Kim Serradell, Francisco Doblas-Reyes
11:30 - 12:00	Reduced Numerical Precision in Atmosphere Models and Computational Performance Evaluation
	Authors: Peter Dueben, Mario Acosta
12:00 - 12:30	Evaluating Commercial Distributed Computing for Numerical Weather Prediction
	Authors: Tim Whitcomb, Daniel Arevalo



Barcelona Supercomputing Center Centro Nacional de Supercomputación



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

New Methodologies for Computational Performance Evaluation of Climate and Weather Models

Mario C. Acosta, Oriol Tintó, Miguel Castrillo, Stella Paronuzzi, Xavier Yepes and Kim Serradell

14/06/2019

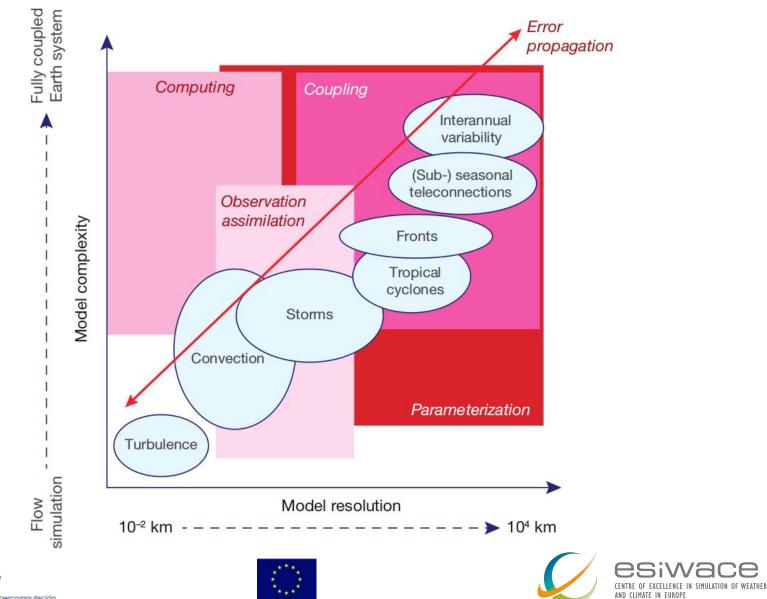
PASC19 Zurich - Minisymposia 06





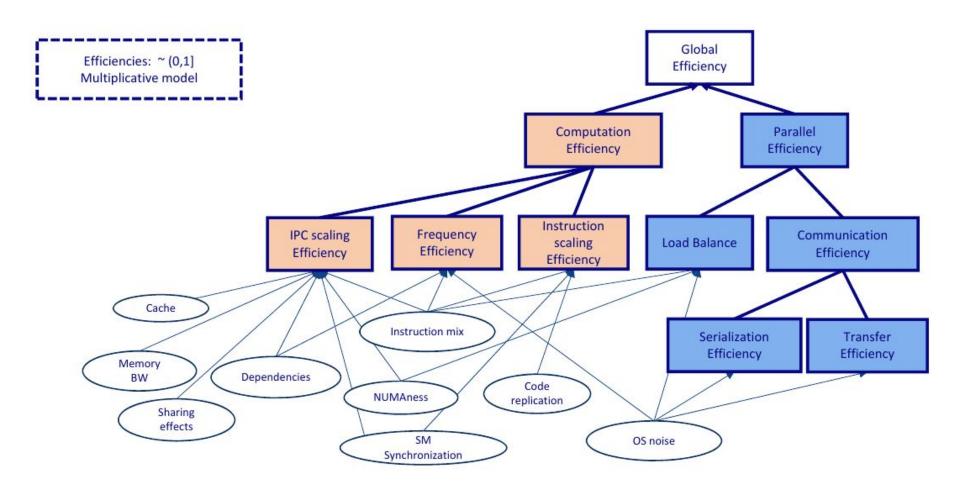






Barcelona Supercomputing Center Centro Nacional de Supercomputación











- The necessary refactoring of numerical codes is given a lot of attention and is stirring a number of discussions.
 - Computational performance analysis and new optimizations are needed for actual numerical models.
 - Study new algorithms for the new generation of high performance platforms (path to exascale).
- Several European institutions and projects working together on the same direction (ESCAPE2, ESiWACE2, IS-ENES3, ETP4HPC...)



AND CLIMATE IN EUROP

- Area of study
- Deployment efficiency
- Benchmarking
- Profiling analysis
- Validation







• Area of study

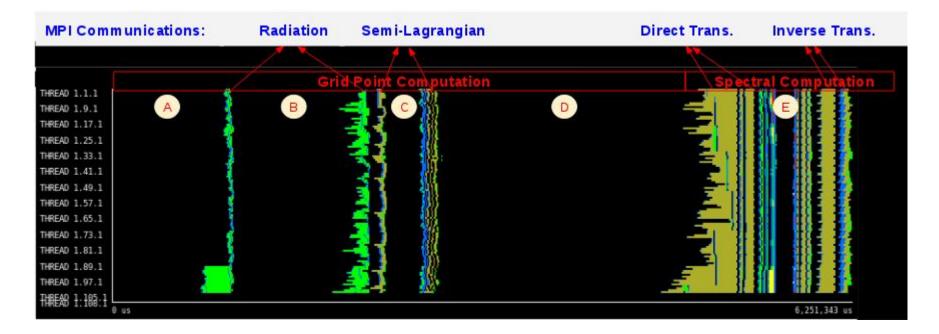
- Configuration used (Operational, New algorithms, Global, Parallelization paradigm...)
- Components activated and cyclic patterns
 - IO, ICE, Radiation, MPI, OpenMP
- Area of study
 - 1 complete time step
- Deployment efficiency
- Benchmarking
- Profiling analysis
- Validation







- Area of study (IFS)
 - 24 hours of simulation, T511L137 on CCA (ECMWF)
 - Selected 1 time step: 104 MPI processes + 4 IO (No OpenMP)
 - Metrics collected for large areas of computation automatically

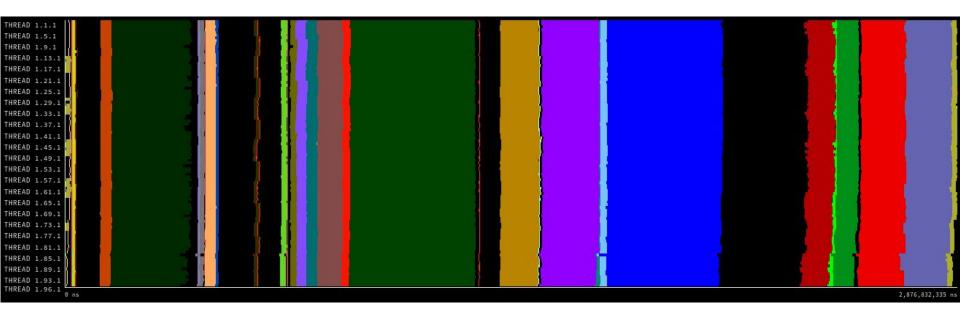








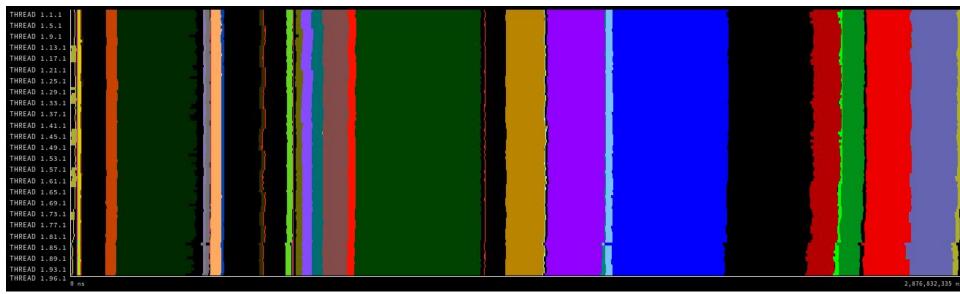
- Area of study (NEMO)
 - 1 day of simulation, ORCA025L91 on MN4 (BSC)
 - Selected the fastest time step automatically
 - 1 time step: 72 MPI processes (No IO, No OpenMP, No SI3)
 - Metrics collected for User functions manually

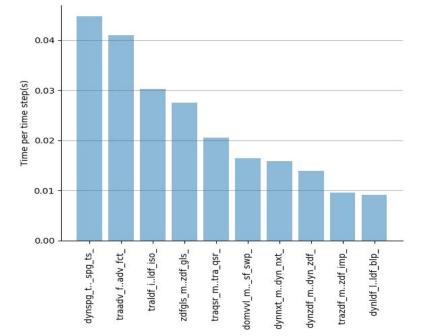














BSC Barcelona Supercomputing Center Centro Nacional de Supercompulación

- Area of study
- Deployment efficiency
 - Compilation flags
 - Comparing fp options (fast, precise, strict...) and optimization options (OX, vectorization, approximations...)
 - Checking external libraries compilation
 - Debug flags (-g, Optimization reports, -f-instrument-functions...)
- Benchmarking
- Profiling analysis
- Validation







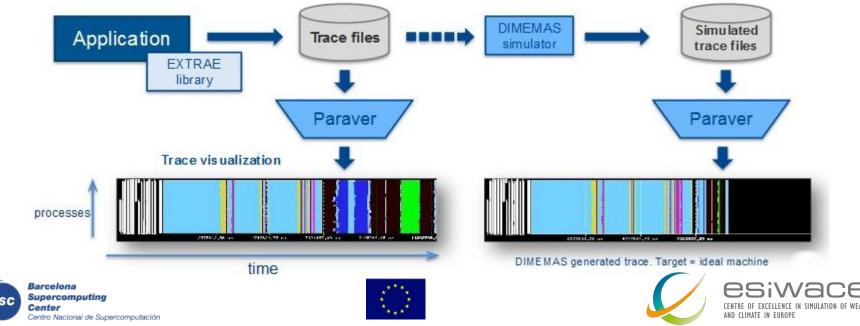
- Area of study
- Deployment efficiency
- Benchmarking
 - Basic Tests to collect Hardware metrics
 - Communications (Latency, Bandwidth, CPU, Parallel Efficiency...)
 - Weak and Strong scaling (MPI, OpenMP, Block processing and Hybrid sets)
 - Comparing optimizations (Double VS Single Precision...)
 - Extrae metrics collection and trace production
- Profiling analysis
- Validation



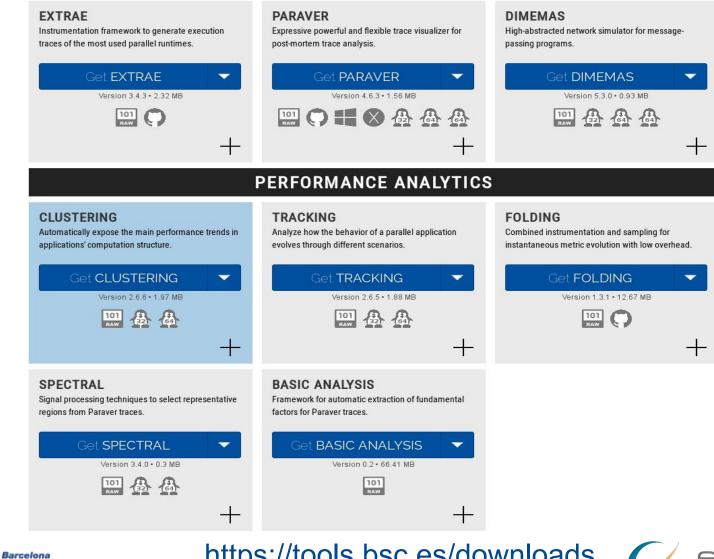




- Since 1991
- Based on traces
- Open Source → http://www.bsc.es/paraver
- Extrae: Package that generates Paraver-trace files for a post-morten analysis
- Paraver: Trace visualization and analysis browser
- Dimemas: Message passing simulator
- Include traces manipulation: Filter, cut traces...



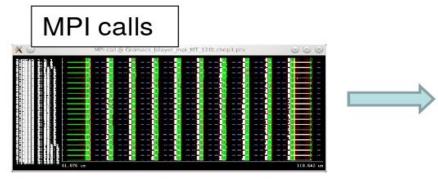
CORE TOOLS



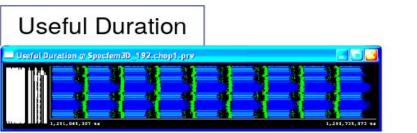
Supercomputing Center Centro Nacional de Supercomputación https://tools.bsc.es/downloads



From timelines to tables



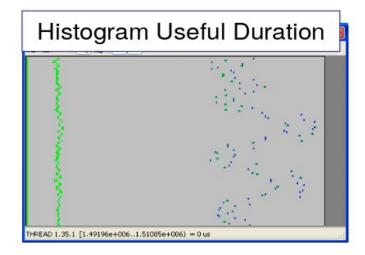
	call	s pi	One	9	ilayer_mpi_MT	_120t chop1 prv			9
	Outside MPI	MPI Send	MPI_Recv	MPI Isend	MPI Irecv	MPI Waitall	MPI_Bcast	MPI Reduce	MPL Allo
THREAD 1.113.1	67.6081 %	0.0682.%	9.9182 %	2.5777 %	1.7698 %	5.1676 %	0.5934 %	0.1465 %	
THREAD 1.114.1	42.8434 %			1.1947 %	1.0400 %	7.7056 %		-	
THREAD 1.115.1	68.6127 %	0.0707 %	9.6223.%	2.2589 %	2,0177.95	3.9825 %	0.5249 %	0.0297 %	31
THREAD 1.116.1	74,6039.%	0.0531 %	9.6084 %	2.8813 %	2 5593 %	2.9296 W	0.5095 %	0.0483 %	100
THREAD 1.117.1	74,3733 %	0:0691 %	9.7012 %	2.8517 %	2.5240 %	X 🚽 202	. MPI call profile @	Gramacs_bilayer_mpi_	NT_12014
THREAD 1.118.1	72,7770 %	0.0545 %	9.5489 %	2.8489 %	2 5353 %	C D 3		H ++ %	
THREAD 1.119.1	66,7994 %	0.0682 %	10.0674 %	2.4206 %	1.9741 %				
THREAD 1.120.1	43.7224 %)	20.5273 %	1.1912.56	1.0175 %				
(
Total	8,012,4546 %	7.3174 %	1.370.5276 %	288,6168 %	253.0137 %	54			
Average	66.7705 %	0.0690 %	11.4211 %	2.4051 %	2.1084 %				
Maximum	75.6821%	0.4390 %	21.2505 %	2.9706 %	2.6369 %				
Minimum	40.5200 %	0.0129 %	8.8583 %	1.1489 %	1.0077%				
StDev	11.3685 %	0.0474 %	4.0613 %	0.5984 %	0.5406 %				
Avg/Max	0.8822	0.1572	0.5374	0.8095	D.7996				







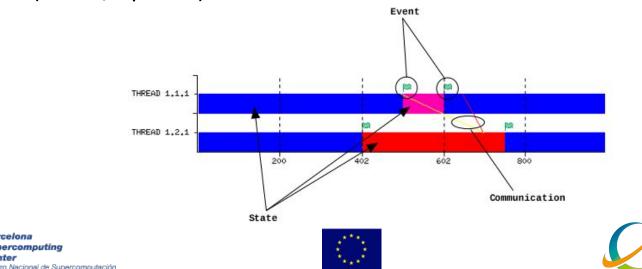








- Paraver traces: made up from records (timestamp + event or activity) of three different kind:
 - State records: intervals of thread status, i.e, waiting in a barrier (either MPI or OpenMP), waiting for a message, computing...
 - Event records: punctual event occurred in a given timestamp, as entry & exit points of user functions, MPI routines, OpenMP parallel regions...
 - Communication records: relationship between two objects, as communication between two processes (MPI), task movement among threads (OpenMP/OmpSs) or memory transfers (CUDA/OpenCL).

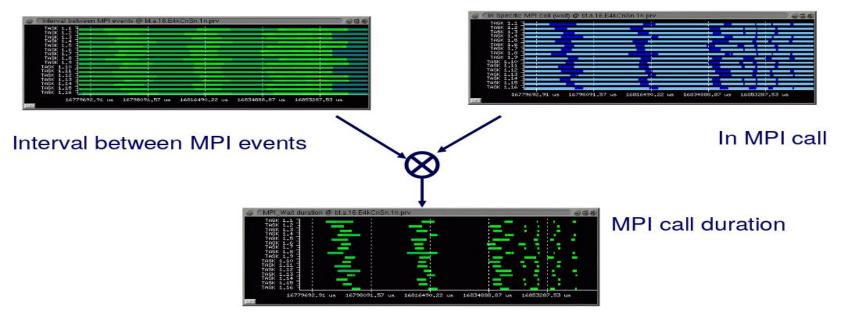




larcelona Supercomputing

Semantic Funcionality

- Derived windows
 - Point wise operation
 - $S = \alpha * S^a < op > \beta * S^b$
 - <op>:+,-,*,/,...









- Area of study
- Deployment efficiency
- Benchmarking
- Profiling analysis
 - MPI and OpenMP profile summary and Basic Analysis Tool
 - PAPI counters
 - MPI and OpenMP evaluation in detail
 - Clustering and Tracking Tools
 - Sampling and Folding Tools
 - Connection to the code
 - Dimemas Tool
- Validation







MPI Profile Summary

Parallel and Communication efficiency, Global load balance \rightarrow less than 85%?

IFS

Parallel Efficiency

							11 0						
IC ID 30) 🔾 🍭 I	н	1 II X	Σ									
	Outside MPI	MPI_Send	MPI_Recv	MPI_lsend	MPI_Irecv	MPI_Wait	MPI_Barrier	MPI_Alltoallv	MPI_Gatherv	MPI_Comm_rank	MPI_Comm_size	MPI_Bsend	MPI_Waitany
Total	66,578.44 %	1.71 %	773.76 %	646.21 %	239.35 %	12,362.37 %	806.93 %	10,757.31 %	35.56 %	2.49 %	448.23 %	0.81 %	7,746.82 %
Average	66.31 %	0.00 %	0.77 %	0.64 %	0.24 %	12.31 %	0.80 %	10.71 %	0.04 %	0.00 %	0.45 %	0.81 %	7.72 %
Maximum	72.93 %	0.01 %	2.98 %	1.60 %	0.80 %	18.56 %	1.84 %	25.06 %	1.12 %	0.01 %	1.88 %	0.81 %	19.25 %
Minimum	57.05 %	8.00 %	0.01 %	0.08 %	0.07 %	3.11 %	0.00 %	5.25 %	0.00 %	0.00 %	0.16 %	0.81 %	0.31 %
StDev	2.03 %	0.00 %	0.57 %	0.36 %	0.06 %	2.52 %	0.41 %	3.57 %	0.12 %	0.00 %	0.10 %	0 %	3.18 %
Avg/Max	0.91	0.31	0.26	0.40	0.30	0.66	0.44	0.43	0.03	0.34	0.24	1	0.40
								· · · · · · · · · · · · · · · · · · ·		· · · · · · · · · · · · · · · · · · ·			
MPI Send													

Communication Efficiency

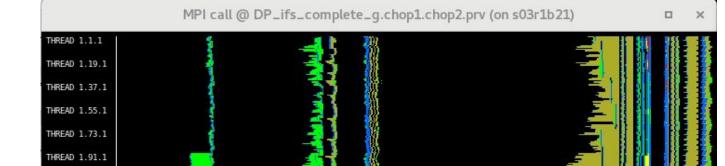
Global Load Balance







IFS



IG ID 30 Q 🗮 II H H II 🕂 Σ ½

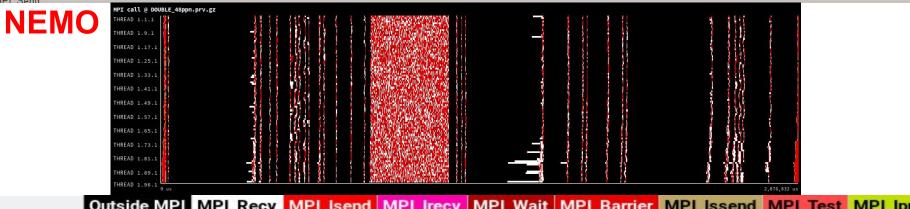
THREAD 1.108.1 0 us

	Outside MPI	MPI_Send	MPI_Recv	MPI_lsend	MPI_Irecv	MPI_Wait	MPI_Barrier	MPI_Alltoallv	MPI_Gatherv	MPI_Comm_rank	MPI_Comm_size	MPI_Bsend	MPI_Waitany
Total	66,578.44 %	1.71 %	773.76 %	646.21 %	239.35 %	12,362.37 %	806.93 %	10,757.31 %	35.56 %	2.49 %	448.23 %	0.81 %	7,746.82 %
Average	66.31 %	0.00 %	0.77 %	0.64 %	0.24 %	12.31 %	0.80 %	10.71 %	0.04 %	0.00 %	0.45 %	0.81 %	7.72 %
Maximum	72.93 %	0.01 %	2.98 %	1.60 %	0.80 %	18.56 %	1.84 %	25.06 %	1.12 %	0.01 %	1.88 %	0.81 %	19.25 %
Minimum	57.05 %	0.00 %	0.01 %	0.08 %	0.07 %	3.11 %	0.00 %	5.25 %	0.00 %	0.00 %	0.16 %	0.81 %	0.31 %
StDev	2.03 %	0.00 %	0.57 %	0.36 %	0.06 %	2.52 %	0.41 %	3.57 %	0.12 %	0.00 %	0.10 %	0 %	3.18 %
Avg/Max	0.91	0.31	0.26	0.40	0.30	0.66	0.44	0.43	0.03	0.34	0.24	1	0.40

MPI_Alltoallv

6,251,343 us

MPI Send



	Outside MPI	MPI_Recv	MPI_lsend	MPI_lrecv	MPI_Wait	MPI_Barrier	MPI_Issend	MPI_Test	MPI_lprobe
Total	9,047.79 %	434.54 %	38.68 %	0.03 %	38.73 %	40.14 %	0.06 %	0.01 %	0.01 %
Average	94.25 %	4.53 %	0.40 %	0.00 %	0.40 %	0.42 %	0.00 %	0.00 %	0.00 %
Maximum	95.86 %	13.65 %	0.54 %	0.00 %	1.04 %	1.39 %	0.00 %	0.00 %	0.00 %
Minimum	84.07 %	2.77 %	0.28 %	0.00 %	0.12 %	0.11 %	0.00 %	0.00 %	0.00 %
StDev	1.94 %	1.74 %	0.06 %	0.00 %	0.21 %	0.32 %	0.00 %	0.00 %	0.00 %
Avg/Max	0.98	0.33	0.75	0.38	0.39	0.30	0.59	0.24	0.33

Basic Analysis

Overview of the collected raw data:

I	10	08 108
Runtime (us)	110741508.7	76 71238767.9
Runtime (ideal)	105675625.0	68396939.23
Useful duration (average)	88427932.0	57382830.24
Useful duration (maximum)	94410288.	.2 61484222.58
Useful duration (total)	9196504931.	.3 5967814345.21
Useful duration (ideal, max)	94410288.	.2 61484222.58
Useful instructions (total)	2679842251571	14 23201423473963
Useful cycles (total)	2198500033287	74 14299301515415
Parallel efficiency	79.85%	80.55%
Overview of the computed model	108	108
	=================	
Load balance	93.66%	93.33%
Communication efficiency	85.25%	86.31%
Serialization efficiency	89.34%	89.89%
Transfer efficiency	95.43%	96.01%
Computation scalability	100.00%	154.10%
Global efficiency	79.85%	124.13%
	l I	
IPC scalability	100.00%	133.11%
IPC scalability Instruction scalability	100.00% 100.00%	133.11% 115.50%
IPC scalability Instruction scalability Frequency scalability		
Instruction scalability	100.00%	115.50%
Instruction scalability Frequency scalability	100.00% 100.00%	115.50% 100.23%





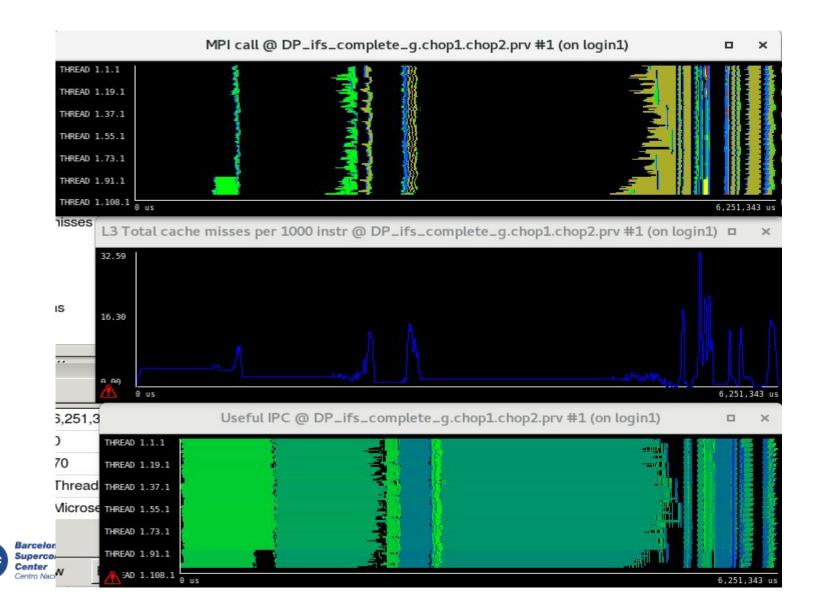
- PAPI counters collected during the execution
- Some of them are based on other native PAPI counters and derived from the base metrics

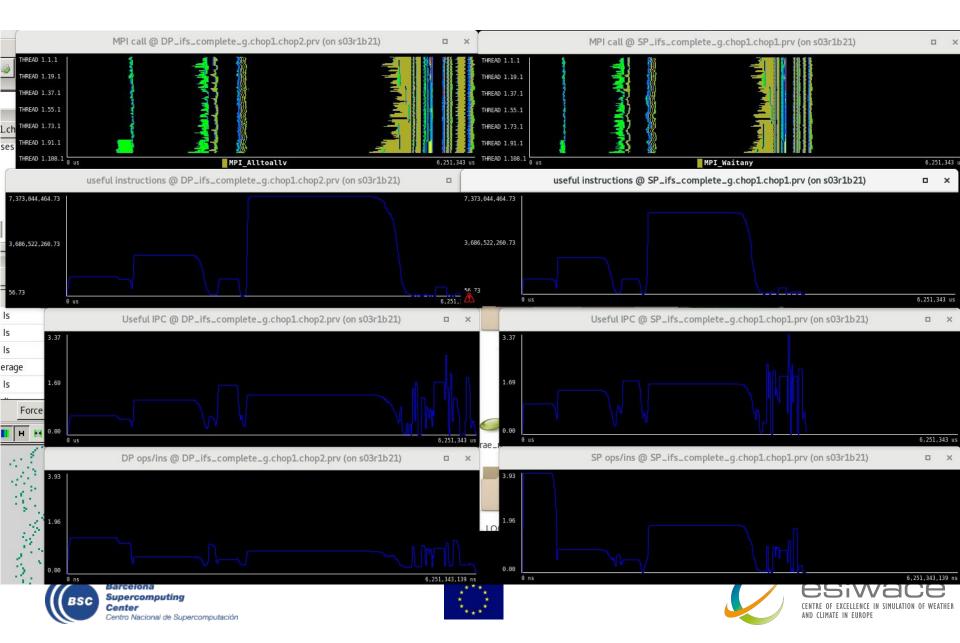
	Derived
Instructions	
Cycles	
Useful Duration	Х
Useful Instructions	Х
Useful IPC	Х
Loads	
Stores	
L3/L2/L1_Total_Misses	
L3/L2/L1_MISS_RATIO	Х
FP_OPS	
FP_TOT_INS	
INS_VEC	Х

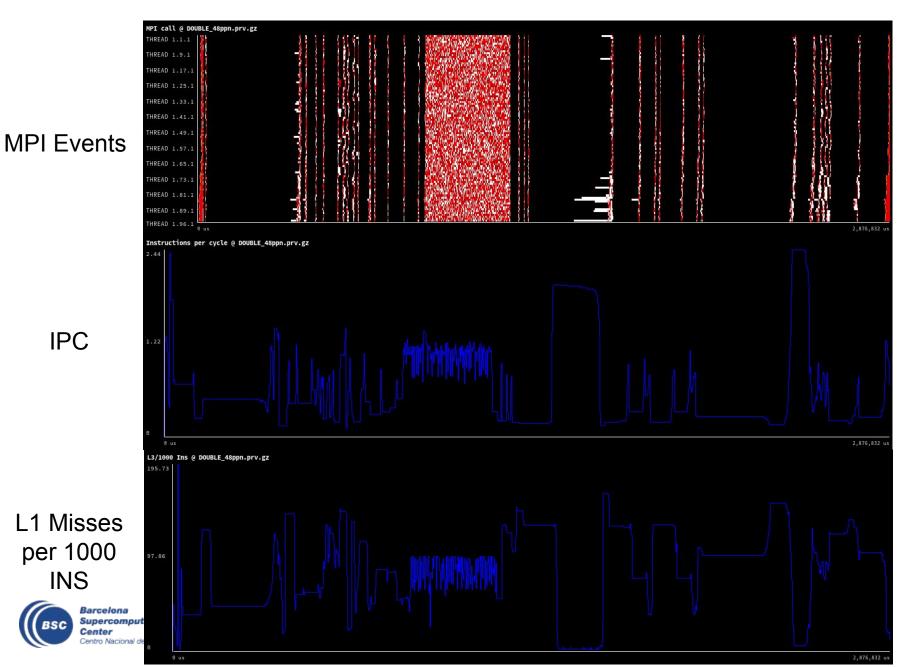




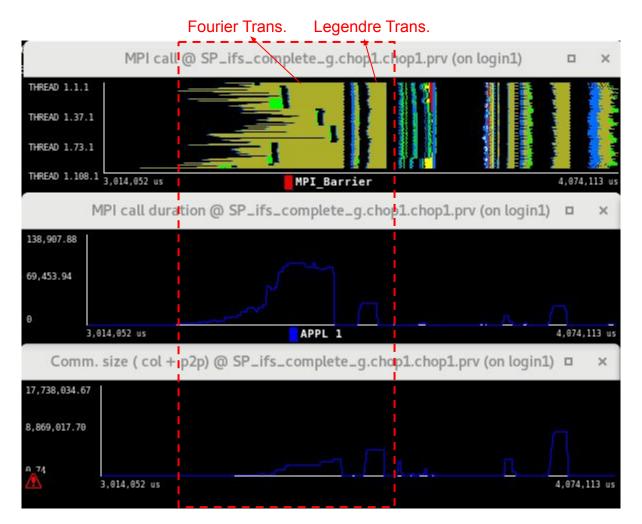








MPI evaluation



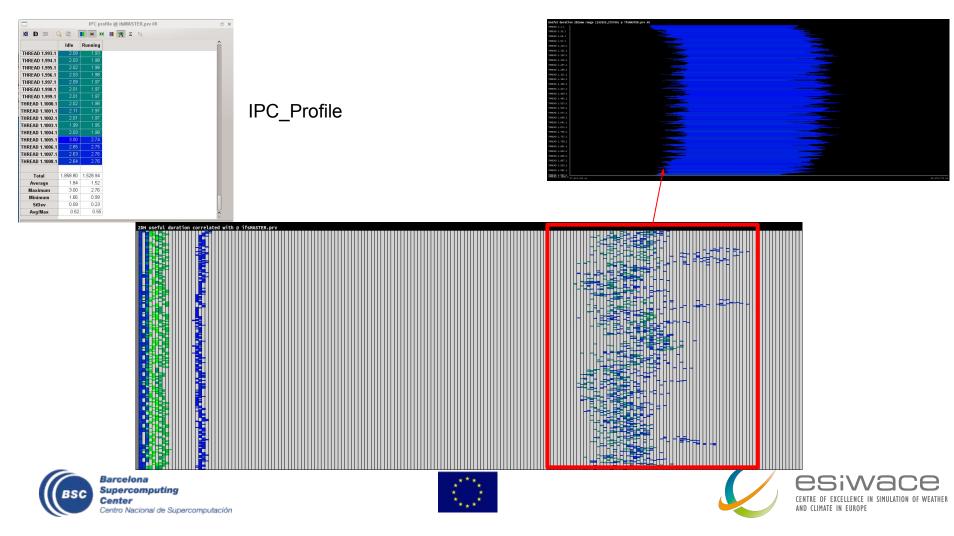






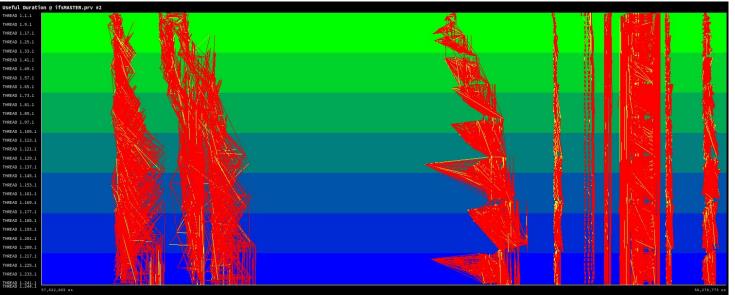
MPI evaluation

- IPC less than 1 for calculation areas?
- Are there load imbalance regions?



MPI evaluation

• Are MPI communications efficient according to the map affinity?



Affinity per node



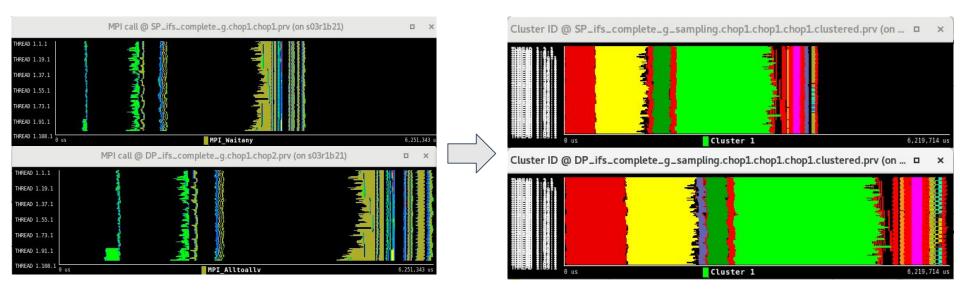
Barcelona Supercomputing Center Centro Nacional de Supercomputación





Clustering Tool

Applying Clustering for an automatic profiling analysis



- Characterizes computing bursts that are similar and groups them into clusters
- Allows to study the behavior of the clusters separately, identify patterns, etc.



Conter entro Nacional de Supercomputación





Tracking Tool

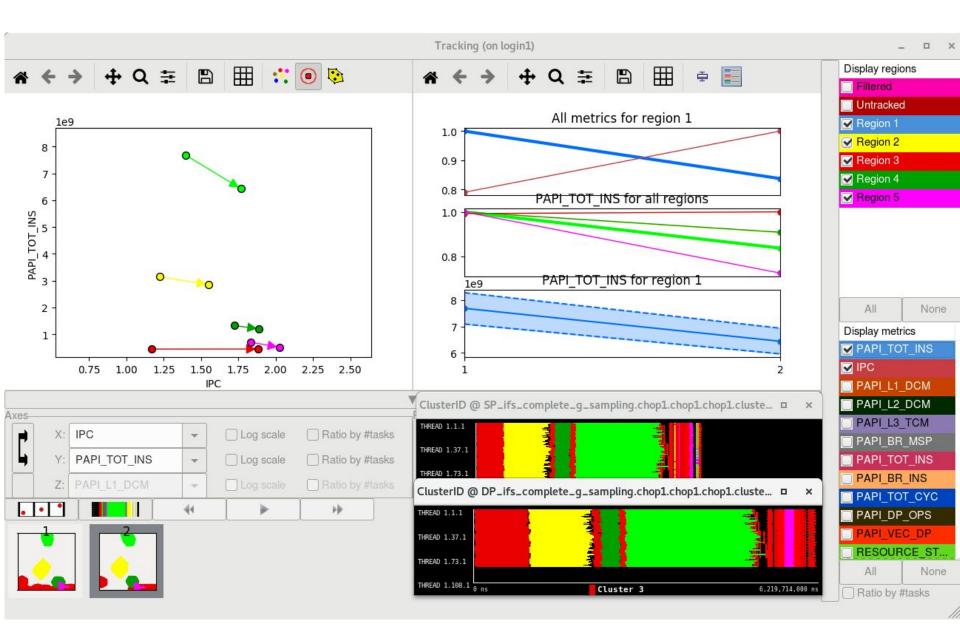
- A friendly way to quantify and visualize the evolution of the clusters among several traces
- The tool has 2 parts
 - Recognition algorithm of "who-is-who", based on heuristics
 - A visualization GUI
- Examples analyzing multiple traces
 - Scaling number of MPI/OpenMP resources (64 128 256...)
 - Testing different microarchitecture features
 - Changing the problem size
 - Trying different compiler optimizations

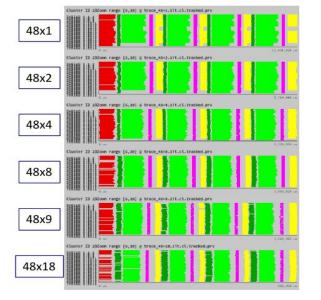






Tracking Tool





109

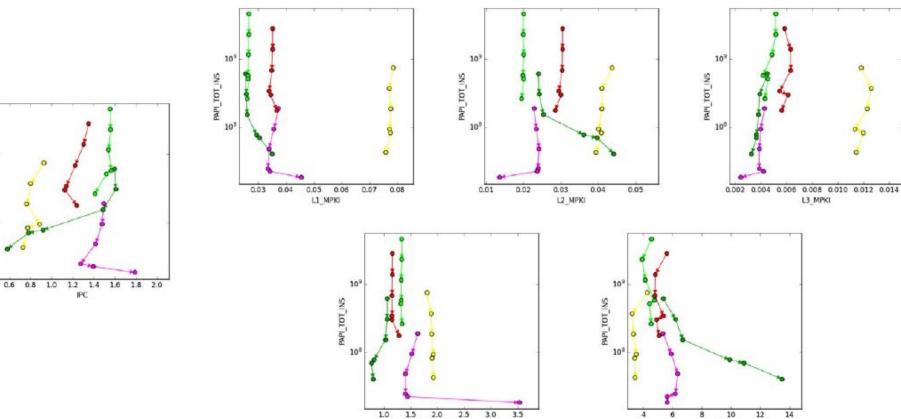
108

PAPI_TOT_INS

Tracking Tool

Tracking IFS MPI+OMP Strong Scaling

1213



L1L2

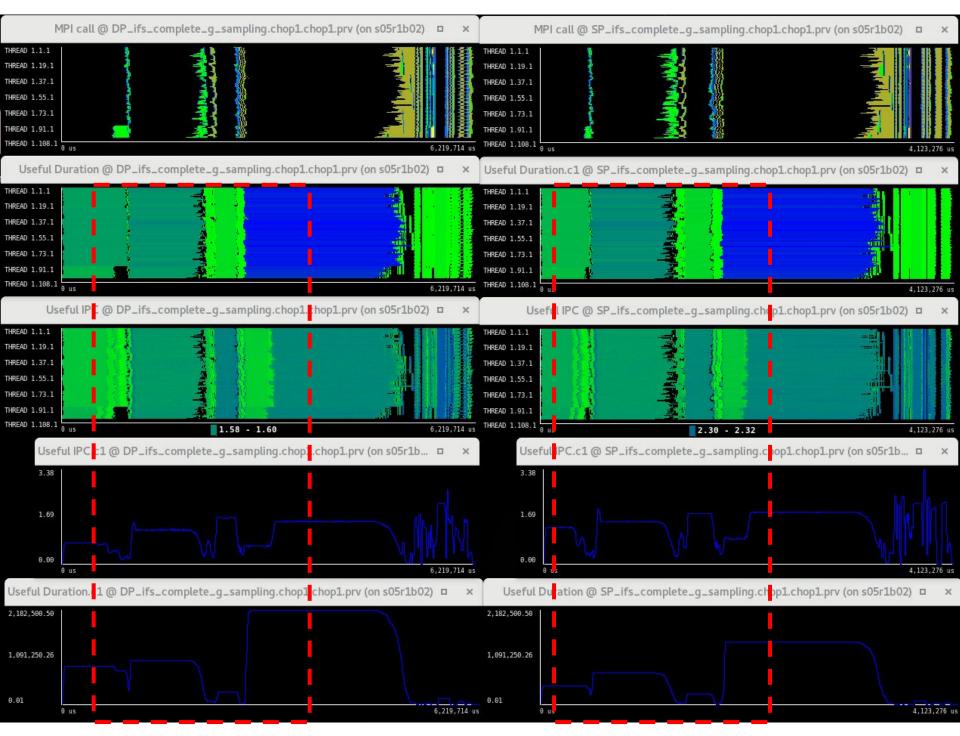
Sampling Tool

- Extrae can be configured to capture performance metrics on a periodic basis using alarm signals and specifying period and variability (10 and 2 respectively for IFS and NEMO tests).
- This means that we will capture samples every 10 ms with a random variability of 2 ms.
- Every sample contains processor performance counters (where every PAPI counter is referred at configured time) and callstack information.









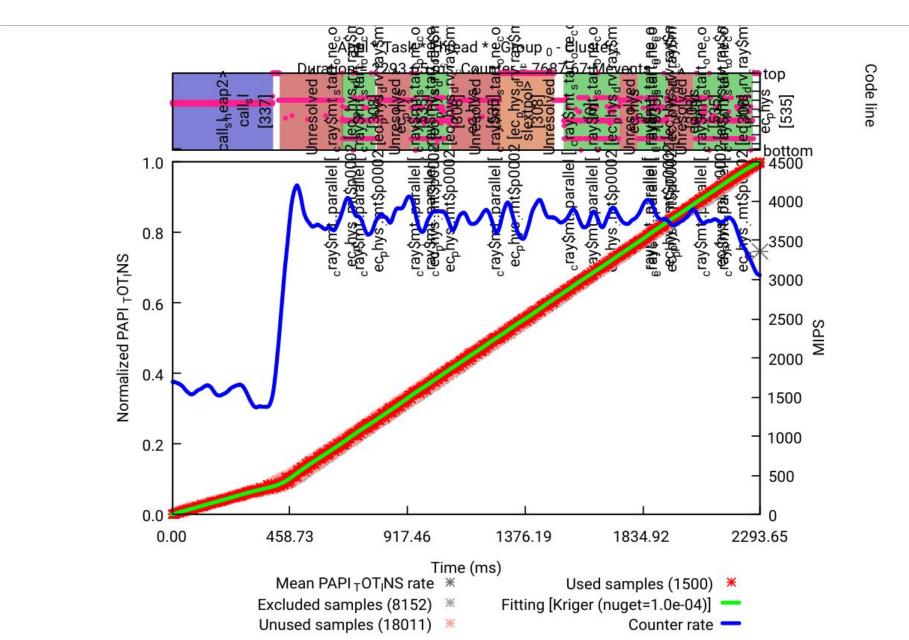
Folding Tool

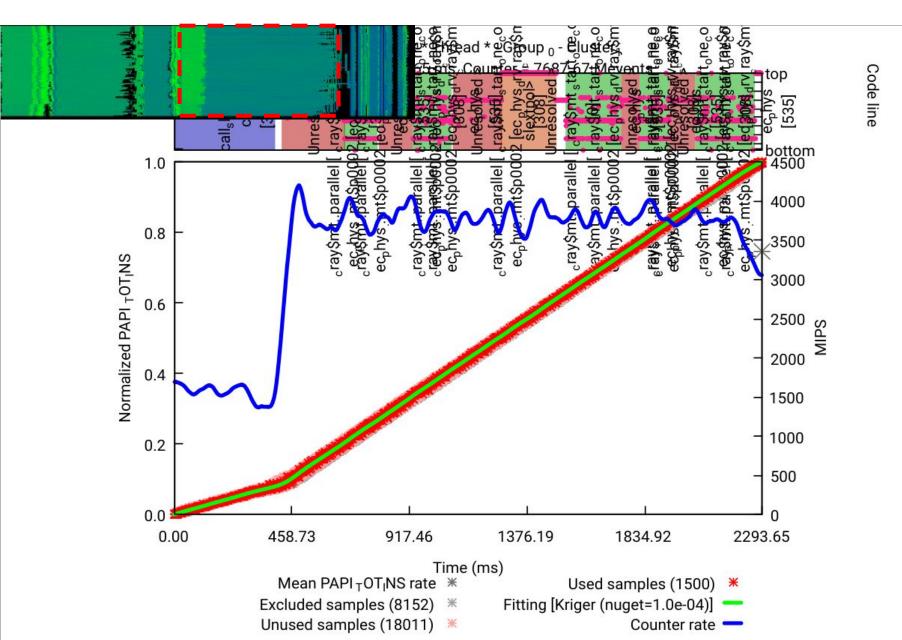
- Combine instrumentation and sampling to provide instantaneous performance metrics, source code and memory references. This mechanism receives a trace-file and generates plots showing the fine evolution of the performance.
- The samples collected are gathered from scattered computing regions into a synthetic region by preserving their relative time within their original region so that the sampled information determines how the performance evolves within the region.
- The performance evolution is connected to source code and memory references at the same time.

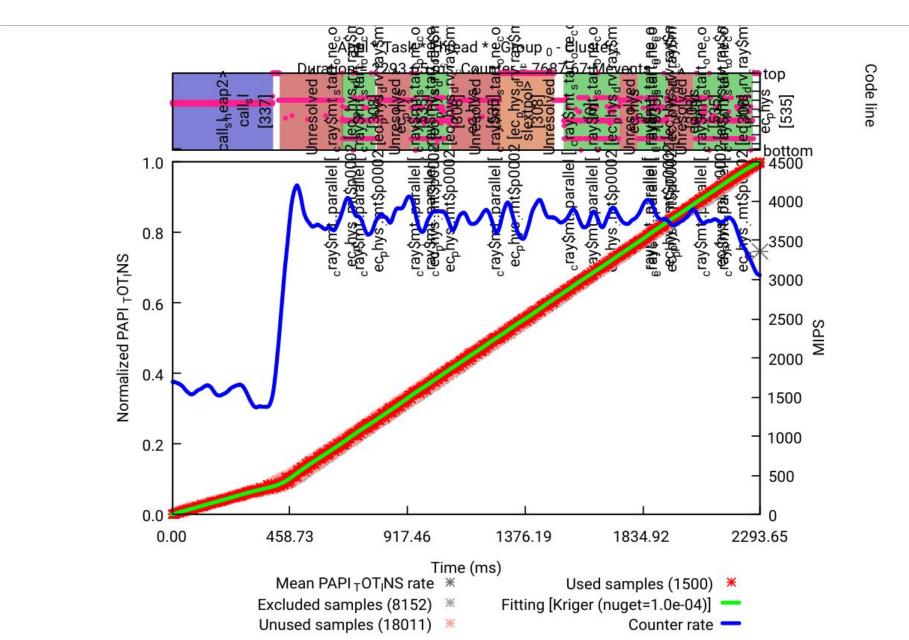










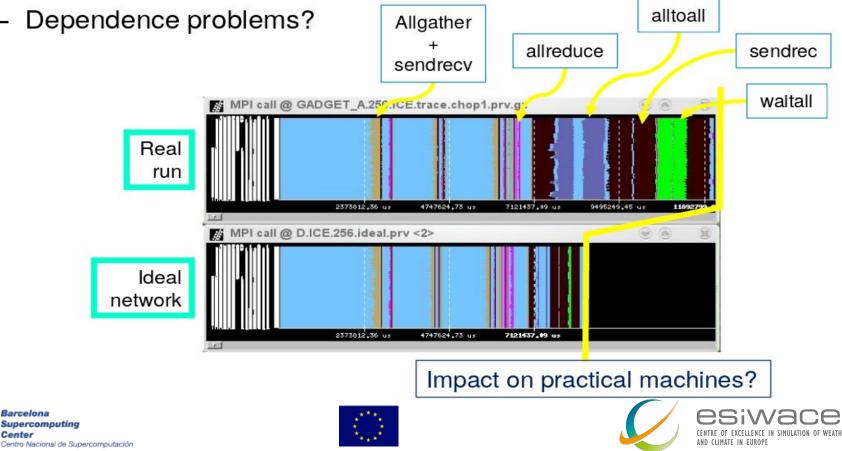




DIMEMAS Tool

The impossible machine: $BW = \infty$, L = 0

- Actually describes/characterizes intrinsic application behavior
 - Load balance problems?



DIMEMAS Tool

Grid point Ideal Network for IFS Semi Grid point calculation Lagrangian execution cmputation (Physics) Actual run MPI call @ fsMASTER.384.1it.prv 843,783 us 1,229,686 us Ideal network call @ D.ideal.ifsMASTER.384.11t.pr 704,067 us 1,089,969 us Imbalance Transfer sensitive Why does not disappear ?

Profiling Methodology

- Area of study
- Deployment efficiency
- Benchmarking
- Profiling analysis
- Validation
 - Reproducibility Test
 - Validation Test

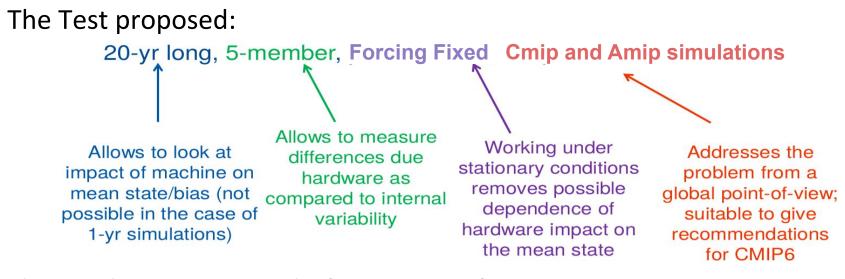




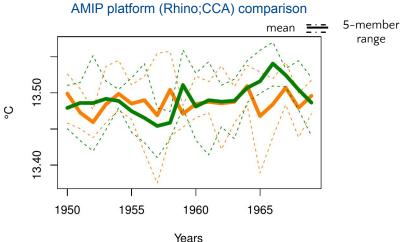


Validation

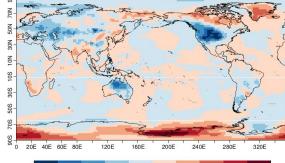
Reproducibility Test: Are your results comparable to the EC-Earth community results?



The results comparing platforms or configurations:

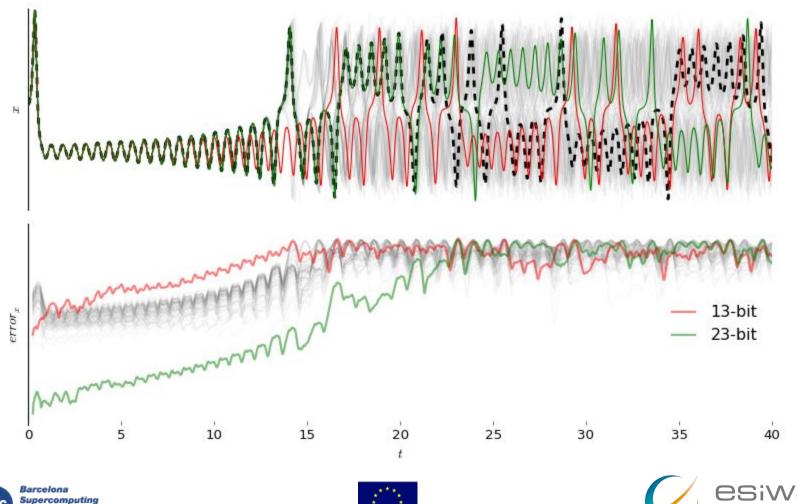


AMIP platform (Rhino;CCA) comparison Kolmogorov-Smirnov differences of two 5-members ensambles



Validation Test

Verifying a non-linear model: a simple example



Center Centro Nacional de Supercomputación



WEATHEI

AND CLIMATE IN EUROPE

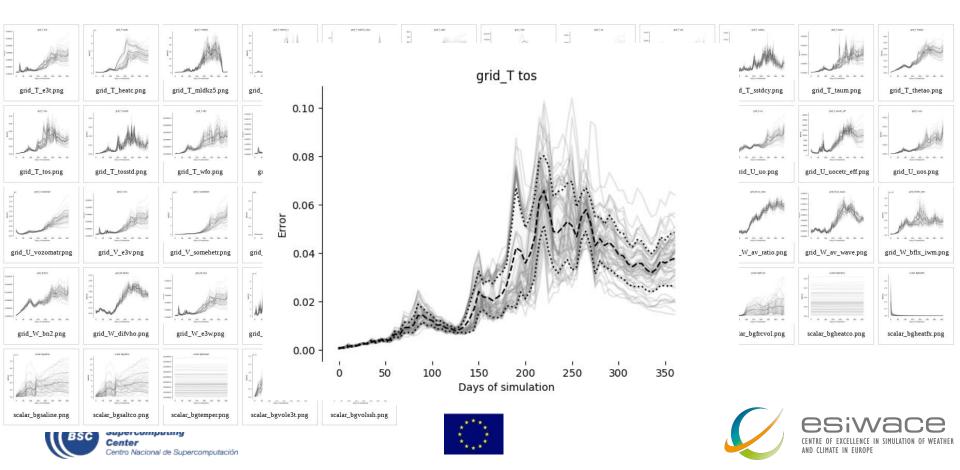
- Initial conditions perturbed with white noise in the 3D temperature field.
- Evaluating 53 output variables.

ntro Nacional de Supercomputación

grid_T_e3t.png	grid_T_heatc.png	grid_T_mldkz5.png	grid_T_mldr10_1.png	grid_T_mldr10_1dcy.png	g grid_T_saltc.png	grid_T_sbt.png	grid_T_so.png	grid_T_sos.png	grid_T_sstdcy.png	grid_T_taum.png	grid_T_thetao.png
grid_T_tos.png	grid_T_tosstd.png	grid_T_wfo.png	grid_T_zos.png	grid_T_zosstd.png	grid_U_e3u.png	grid_U_sozohetrpng	grid_U_sozosatrpng	grid_U_tauuo.png	gid_U_uo.png	grid_U_ucetr_effpng	grid_U_uos.png
gid_U_vozomatrpng	grid_V_e3vpng	grid_V_somehettrpng	grid_V_somesatr.png	grid_V_tauvo.png	grid_V_vo.png	grid_V_vocetr_effpng	grid_V_vomematr.png	grid_V_vospng	grid_W_av_ratio.png	grid_W_av_wave.png	grid_W_bflx_iwm.png
grid_W_bn2.png	grid_W_difvho.png	grid_W_e3wpng	grid_W_emix_iwm.png	grid_W_pcmap_iwm.png	g gid_W_vovematrpng	grid_W_wo.png	scalar_bgfrchfx.png	scalar_bgfrctem.png	scalar_bgfrcvol.png	scalar_bgheatco.png	scalar_bgheatfx.png
scalar_bgsaline.png	scalar_bgsaltco.png	scalar_bgtemper.png	scalar_bgvole3t.png	scalar_bgvolssh.png	٠	•.				esiv	/ace
((BS	Center	anna			- <u>\$</u>	÷.					IN SIMULATION OF WEATHER

AND CLIMATE IN EUROPE

- Initial conditions perturbed with white noise in the 3D temperature field.
- Evaluating 53 output variables.



Example: Compiling with -xHost

grid_T_e3t.png	grid_T_heatc.png	grid_T_mldkz5.png	grid_T_mldrl0_1.png	grid_T_mldr10_1dcy.png	grid_T_saltc.png	grid_T_sbt.png	grid_T_so.png	grid_T_sos.png	grid_T_sstdcy.png	grid_T_taum.png	grid_T_thetao.png
grid_T_tos.png	grid_T_tosstd.png	grid_T_wfo.png	grid_T_zos.png	grid_T_zosstd.png	grid_U_e3u.png	grid_U_sozohetrpng	grid_U_sozoatrpng	grid_U_tauuo.png	grid_U_uo.png	grid_U_uocetr_effpng	grid_U_uos.png
grid_U_vozomatr.png	grid_V_e3v.png	grid_V_somehetr.png	grid_V_somesatrpng	grid_V_tauvo.png	grid_Vvo.png	grid_V_vocetr_effpng	grid_V_vomematr.png	grid_V_vos.png	grid_W_av_ratio.png	grid_W_av_wave.png	grid_W_bflx_iwm.png
grid_W_bn2.png	grid_W_difvho.png	grid_W_e3wpng	grid_W_emix_iwm.png	gid_W_pcmap_iwm.png	grid_W_vovematr.png	grid_W_wo.png	scalar_bgfrchfx.png	sealar_bgfretem.png	scalar_bgfrcvol.png	scalar_bgheatco.png	scalar_bgheatfx.png
				a bi sublighter bi bi b							



scalar_bgsaltco.png

scalar_bgsaline.png

Barcelona Supercomputing Center Centro Nacional de Supercomputación

scalar_bgtemper.png

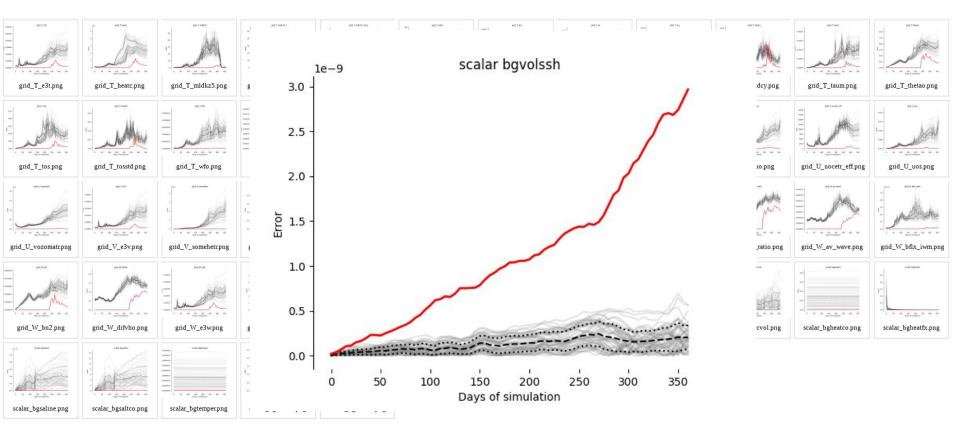
scalar_bgvole3t.png

scalar_bgvolssh.png





Example: Compiling with -xHost





Barcelona Supercomputing Center Centro Nacional de Supercomputación



AND CLIMATE IN

Summary

- The complexity of our climate and weather models requires complexity for the methodologies used to study the computational performance.
- The methodology proposed can be used to find main bottlenecks across platforms, compiler options... for different configurations of the model.
- It can be used to compare computational optimizations (see now Single and Double precision comparison for IFS and NEMO!) and validate the results.
- Profiling analysis include different tools for different purposes:
 - EXTRAE+PARAVER
 - CLUSTERING+TRACKING
 - SAMPLING+FOLDING
 - DIMEMAS









Barcelona Supercomputing Center Centro Nacional de Supercomputación



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

Thank you

The research leading to these results has received funding from the EU H2020 Framework Programme under grant agreement H2020 GA 675191.

The content of this presentation reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

mario.acosta@bsc.es





