www.bsc.es

Barcelona, 1st July 2015

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

BSC Performance tools suite: study cases on improving the efficiency of the EC-EARTH model components

Miguel Castrillo, Oriol Tintó, Kim Serradell
Francisco J. Doblas-Reyes, Jesús Labarta

Efficiency in Earth science models

- Efficiency is especially **critical** for Earth science models
- Simulations use a **huge amount** of computational resources
- Future simulations will need **many more resources**
  - Computational time
  - Storage and postprocess
  - Software to simplify the usage of the model

Energy efficiency

- Energy efficiency

$$EE = \frac{Performance}{Power\ consumed}$$

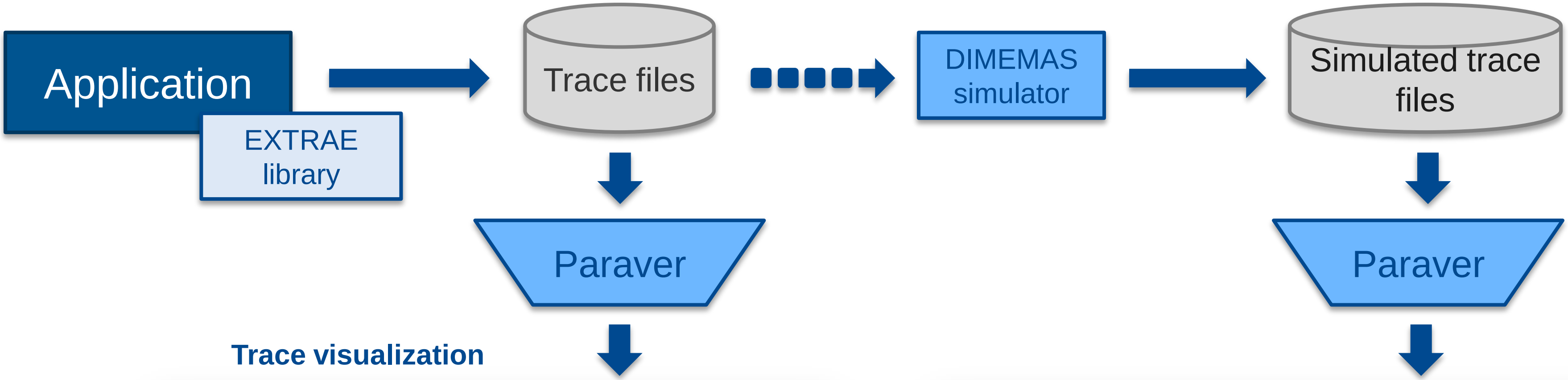
Increase performance

Reduce power

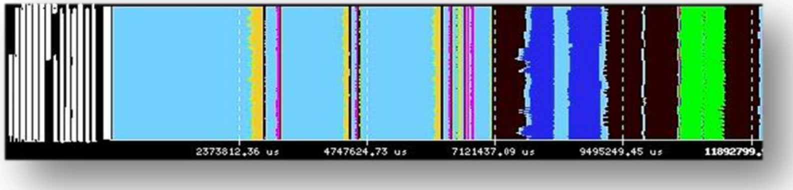
- Performance loss caused by:
  - Bad programming
  - Load imbalance
  - Synchronization
  - Resource contention
  - …

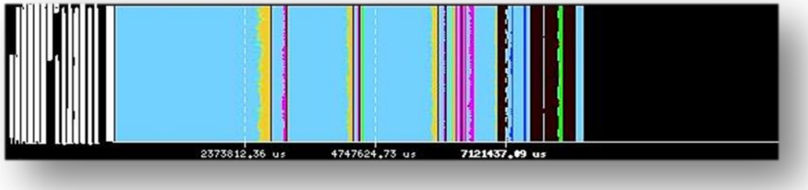
BSC Computer Sciences Performance Tools

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



Trace visualization





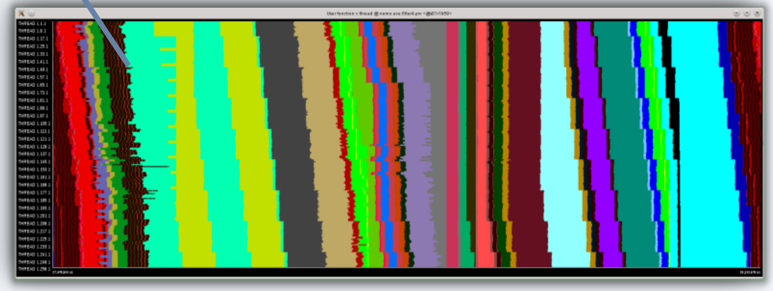
DIMEMAS generated trace. Target = ideal machine

PARAVER trace analysis

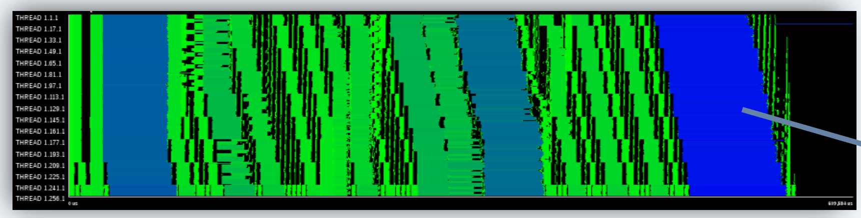
Serial efficiency

Each color represents a function

Functions instrumentation

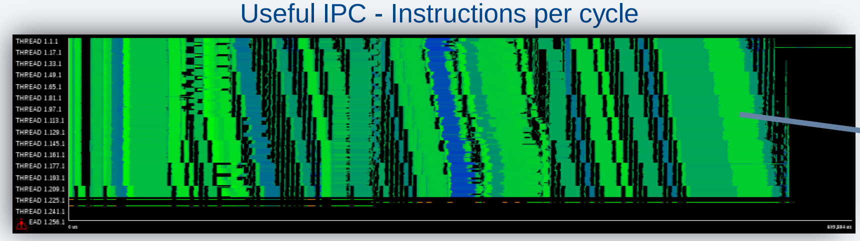


Useful duration



Darker color = Longer duration

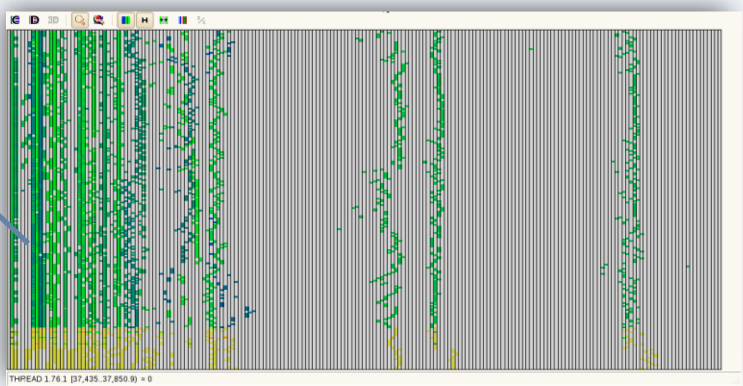
Useful IPC - Instructions per cycle



Lighter color = Less IPC

Horizontal axis -> Time component

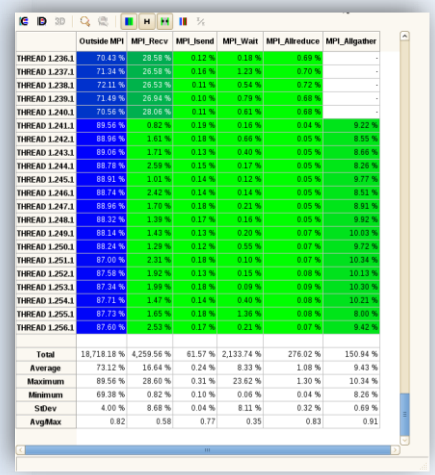
Correlation between two functions



THREAD 1.76.1 [37,435..37,850.9) = 0

Darker color = Higher values

Parallel efficiency

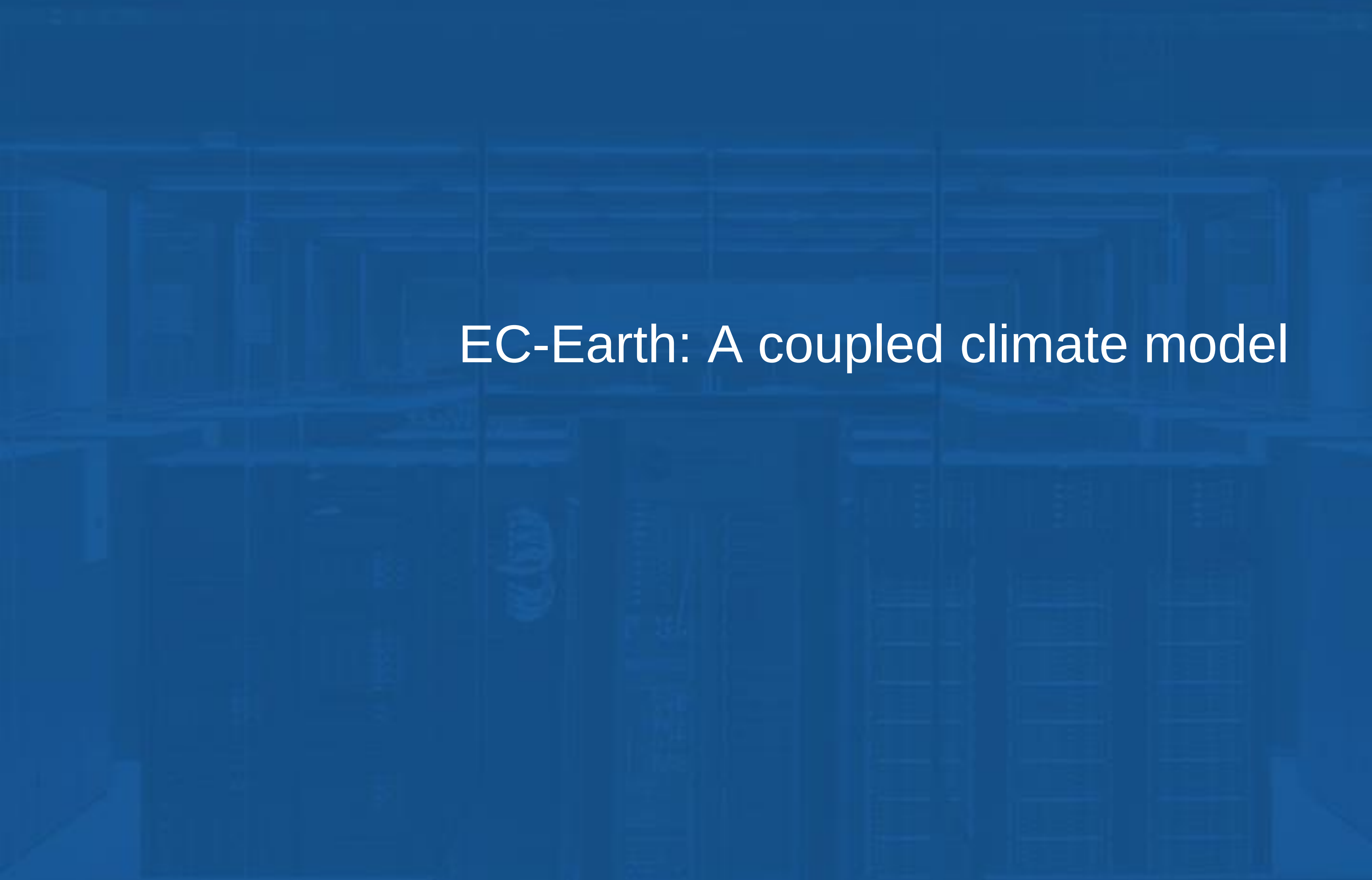


	Outside MPI	MPI_Recv	MPI_Isend	MPI_Wait	MPI_Allreduce	MPI_Allgather
THREAD 1.236.1	70.43 %	28.58 %	0.12 %	0.18 %	0.69 %	-
THREAD 1.237.1	71.34 %	26.58 %	0.16 %	1.23 %	0.70 %	-
THREAD 1.238.1	72.11 %	26.53 %	0.11 %	0.54 %	0.72 %	-
THREAD 1.239.1	71.49 %	26.94 %	0.10 %	0.79 %	0.68 %	-
THREAD 1.240.1	70.56 %	28.06 %	0.11 %	0.61 %	0.68 %	-
THREAD 1.241.1	89.56 %	0.82 %	0.19 %	0.16 %	0.04 %	9.22 %
THREAD 1.242.1	88.96 %	1.61 %	0.18 %	0.66 %	0.05 %	8.55 %
THREAD 1.243.1	89.06 %	1.71 %	0.13 %	0.40 %	0.05 %	8.66 %
THREAD 1.244.1	88.78 %	2.59 %	0.15 %	0.17 %	0.05 %	8.26 %
THREAD 1.245.1	88.91 %	1.01 %	0.14 %	0.12 %	0.05 %	9.77 %
THREAD 1.246.1	88.74 %	2.42 %	0.14 %	0.14 %	0.05 %	8.51 %
THREAD 1.247.1	88.96 %	1.70 %	0.18 %	0.21 %	0.05 %	8.91 %
THREAD 1.248.1	88.32 %	1.39 %	0.17 %	0.16 %	0.05 %	9.92 %
THREAD 1.249.1	88.14 %	1.43 %	0.13 %	0.20 %	0.07 %	10.03 %
THREAD 1.250.1	88.24 %	1.29 %	0.12 %	0.55 %	0.07 %	9.72 %
THREAD 1.251.1	87.00 %	2.31 %	0.18 %	0.10 %	0.07 %	10.34 %
THREAD 1.252.1	87.58 %	1.92 %	0.13 %	0.15 %	0.08 %	10.13 %
THREAD 1.253.1	87.34 %	1.99 %	0.18 %	0.09 %	0.09 %	10.30 %
THREAD 1.254.1	87.71 %	1.47 %	0.14 %	0.40 %	0.08 %	10.21 %
THREAD 1.255.1	87.73 %	1.65 %	0.18 %	1.36 %	0.08 %	8.00 %
THREAD 1.256.1	87.60 %	2.53 %	0.17 %	0.21 %	0.07 %	9.42 %
Total	18,718.18 %	4,259.56 %	61.57 %	2,133.74 %	276.02 %	150.94 %
Average	73.12 %	16.64 %	0.24 %	8.33 %	1.08 %	9.43 %
Maximum	89.56 %	28.60 %	0.31 %	23.62 %	1.30 %	10.34 %
Minimum	69.38 %	0.82 %	0.10 %	0.06 %	0.04 %	8.26 %
StDev	4.00 %	8.68 %	0.04 %	8.11 %	0.32 %	0.69 %
Avg/Max	0.82	0.58	0.77	0.35	0.83	0.91

MPI stats reflect the percentage of time invested in computation for each thread.

Total stats give the communication efficiency and the load balance

* Examples from NEMO 3.4–LIM2 / ORCA025



EC-Earth: A coupled climate model

EC-Earth

- Earth System Model
- Reliable in-house predictions of global climate change
- Part of a Europe-wide consortium
- Being used in **large** European **projects**
  - EMBRACE
  - EUPORIAS
  - IS-ENES
  - SPECS
  - PRIMAVERA
- 3.1 version → IFS + NEMO-LIM + OASIS

EC EARTH

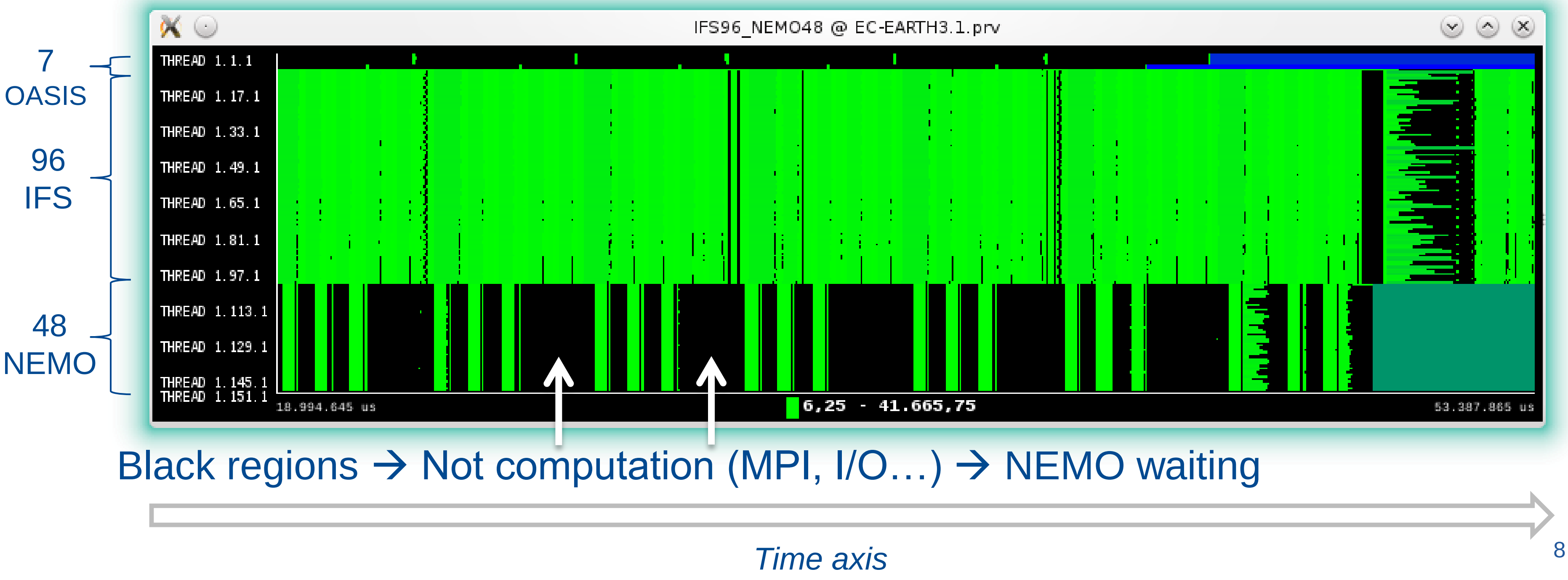
PRIMAVERA Project

- Preparing the next CMIP6 high-resolution simulations called HiResMIP
  - Target resolutions are T511-ORCA025 and T1279-ORCA12
- 19 European groups involved
- No experience in analyzing efficiency on these resolutions

	Single climate experiment (10 members, 60 start dates, 10 years simulated)		
Resolution	Grid Size	Output Size	Computation Time
T511-ORCA025	Atmos 40km - Ocean 25km	720 Tb	132.0 x 10^6
T1279-ORCA12	Atmos 16km - Ocean 9km	1,4 Pb	NA

An EC-Earth Paraver trace

- Motivation: Finding a good configuration to **optimize** the resources usage
- IFS T255L91-ORCA1L46
- Configuration widely used in **production**
  - Using 7 cores for OASIS, 96 for IFS and 48 for NEMO
- 1 day simulation traces
- Traces generated in burst mode (only computational regions > 100us)
- Paraver view → Useful duration (displays duration of computational bursts)



Black regions → Not computation (MPI, I/O…) → NEMO waiting

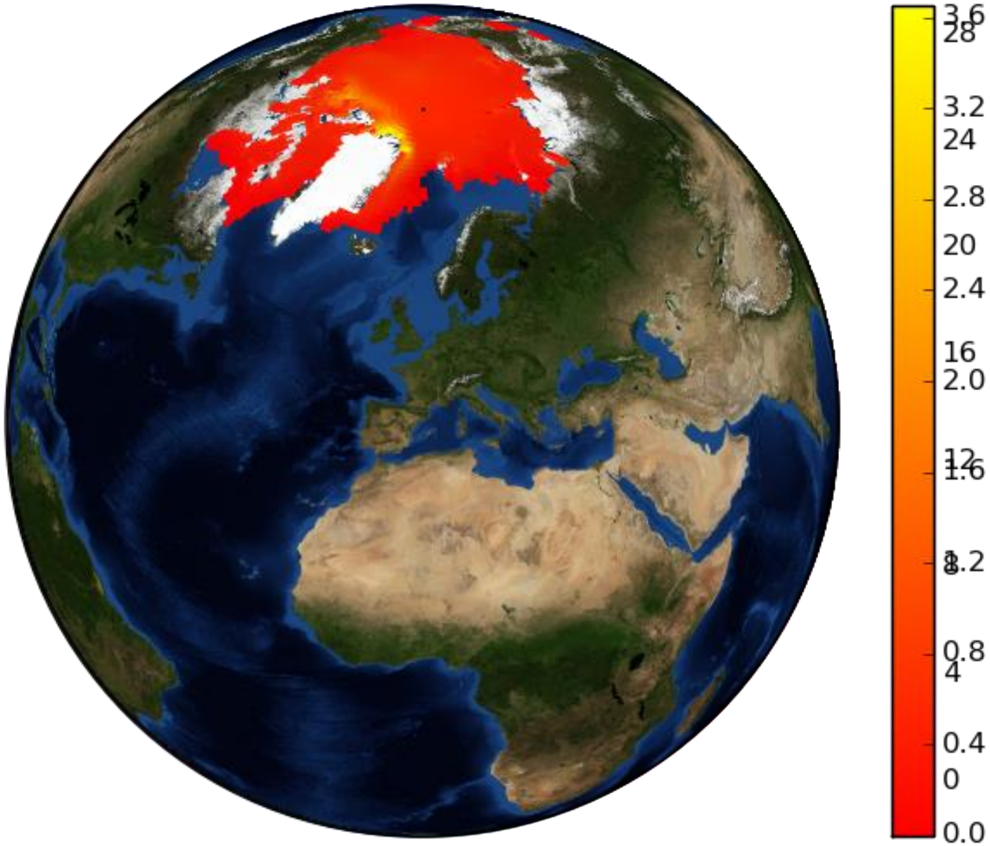
NEMO: An ocean model

NEMO ocean model

Nucleus for European Modeling of the Ocean (NEMO) is a **state-of-the-art** global **ocean model**

It is used in oceanographic research, operational oceanography, seasonal forecast and climate studies

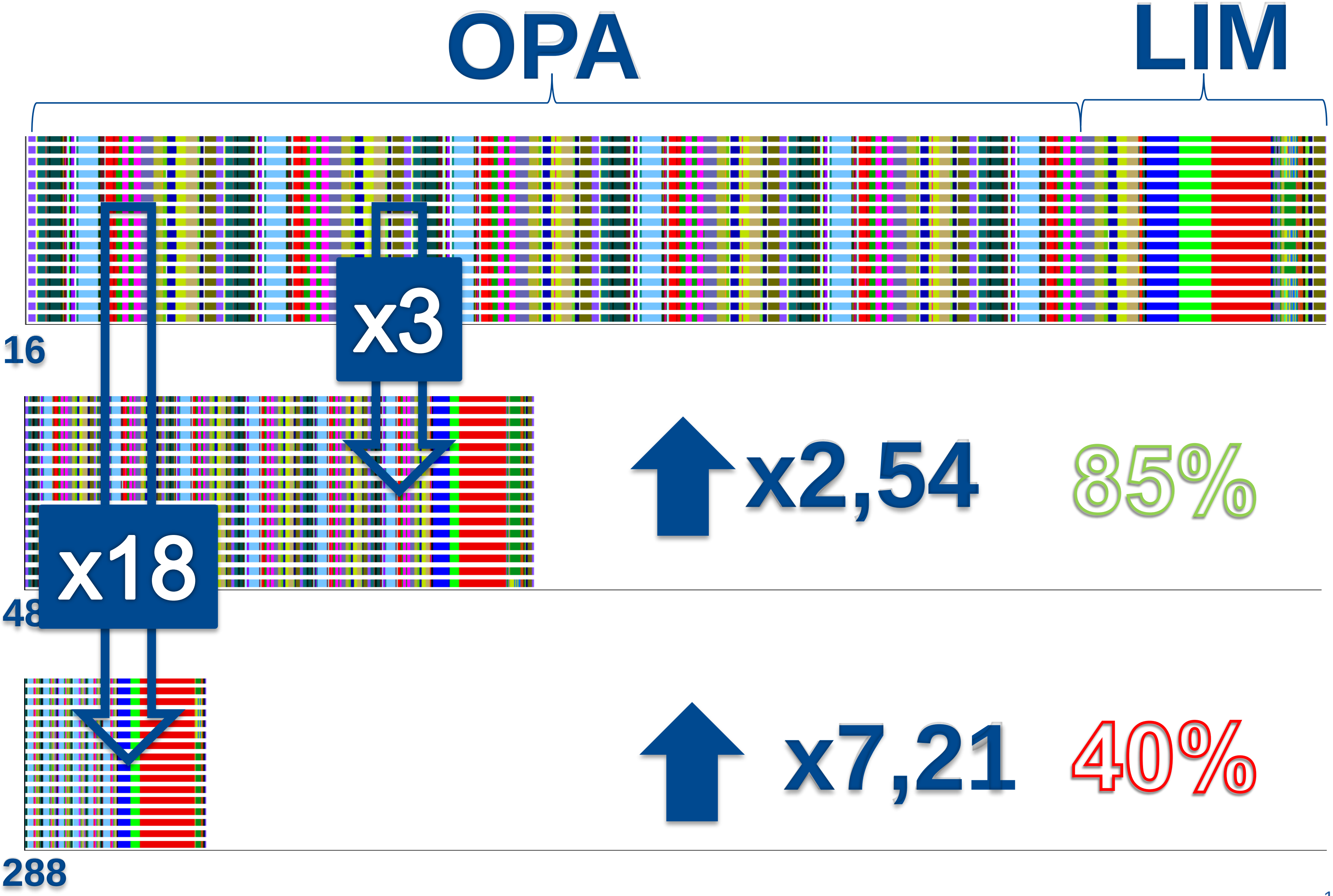
Includes several **sub-models**. Many of them can work in standalone version , many others need to be coupled



Sea Surface Temperature

Sea Ice Thickness

NEMO model scalability



OPA

LIM

16

x3

48

x18

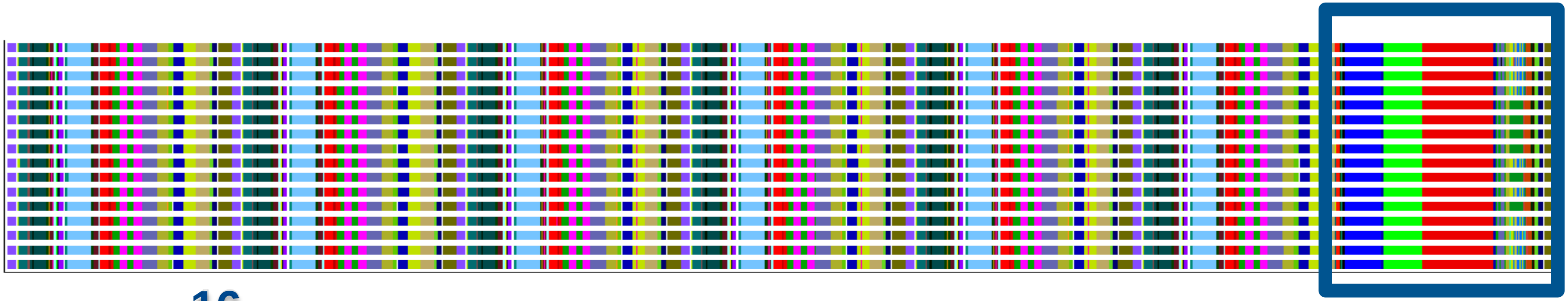
288

x2,54 85%

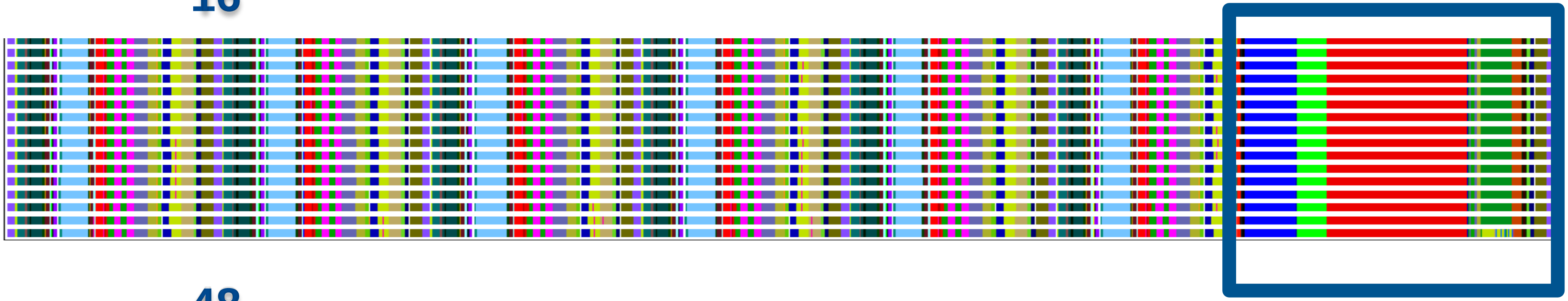
x7,21 40%

*Timelines have the same duration

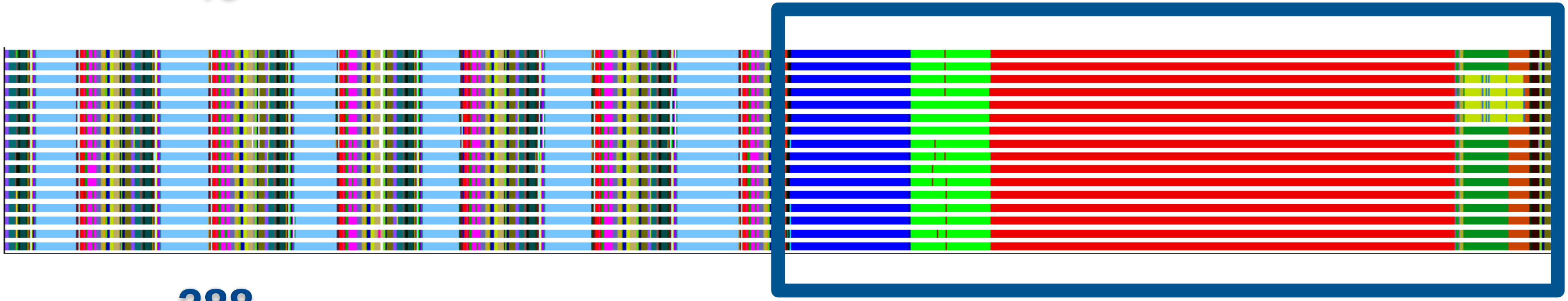
LIM model as a bottleneck



16

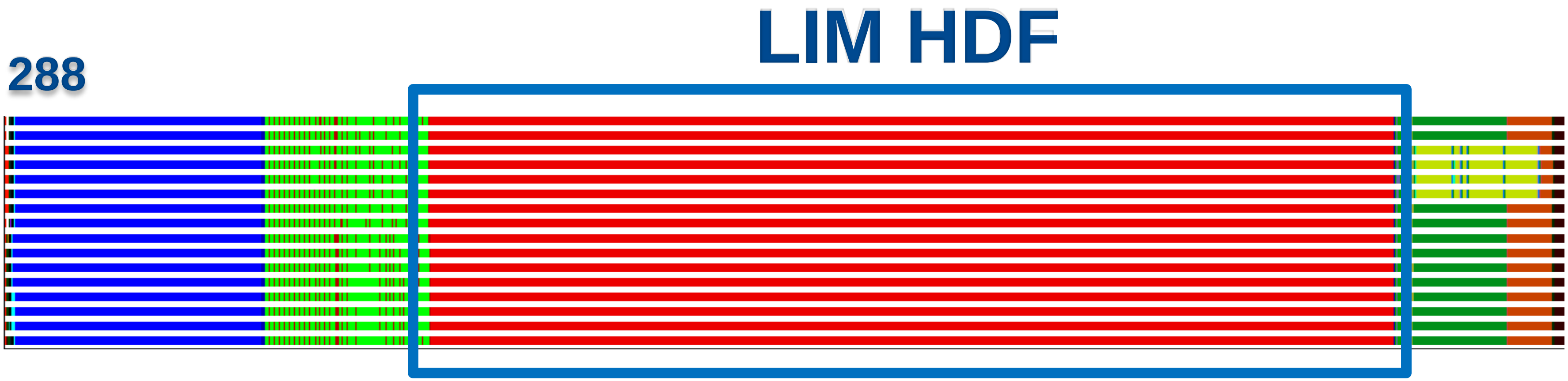


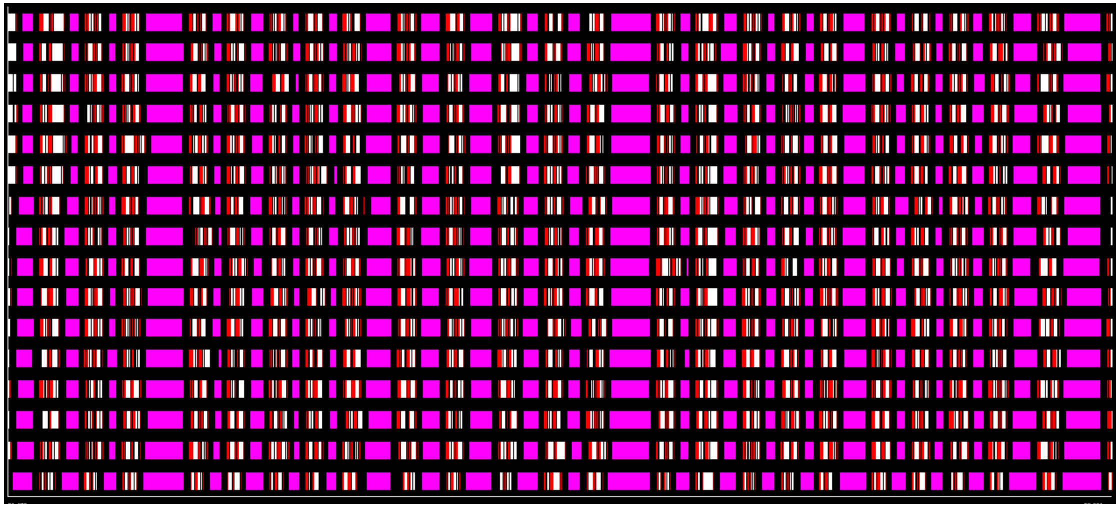
48



288

Sea ice horizontal diffusion





🟦 Outside MPI
🟥 MPI Isend
🟦 MPI Recv
🟥 MPI Wait
🟪 MPI All_gather

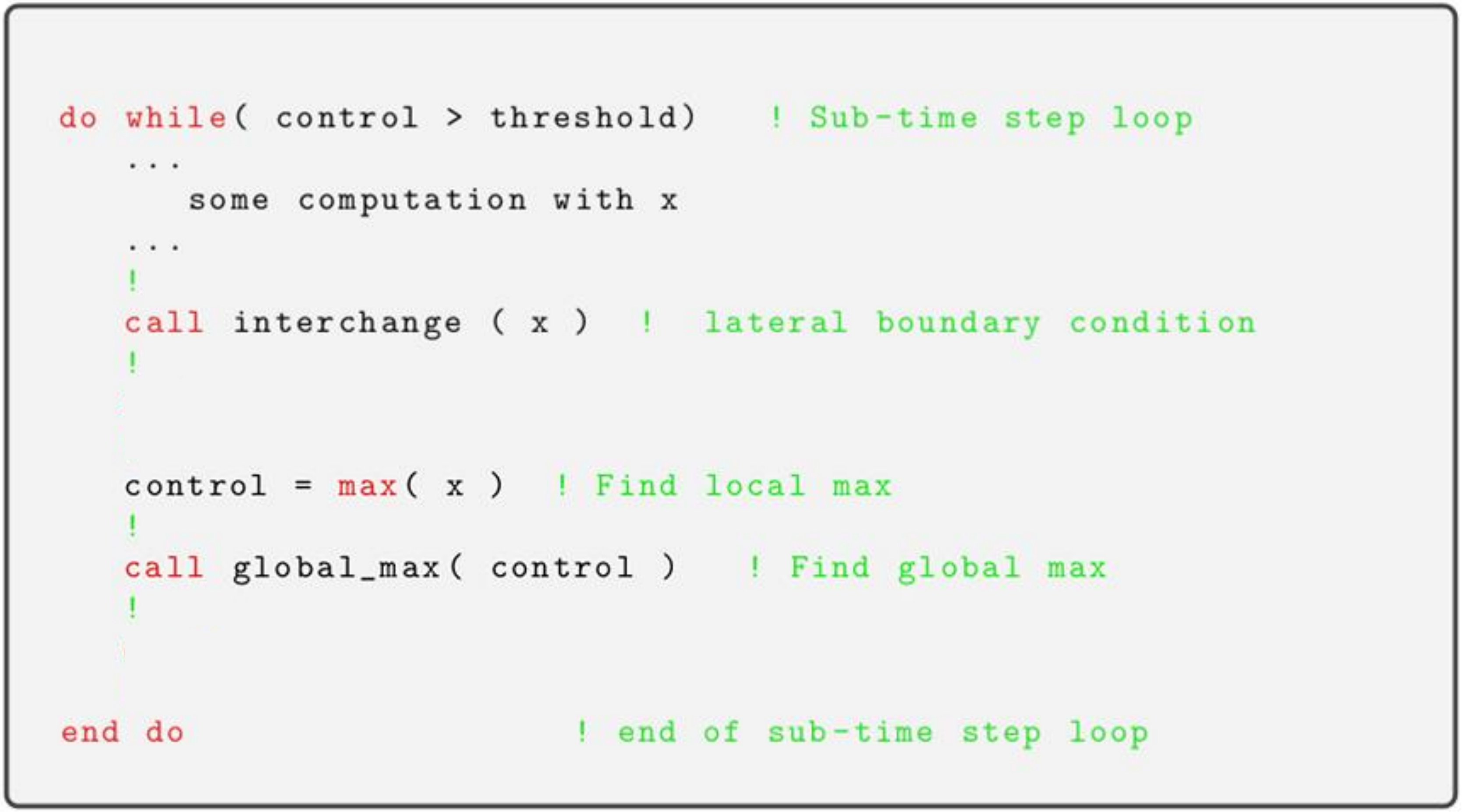
Only **20%** of the time invested on **computation**

Global Communication at **every** loop **iteration → 60%** of the time

NEMO model optimizations

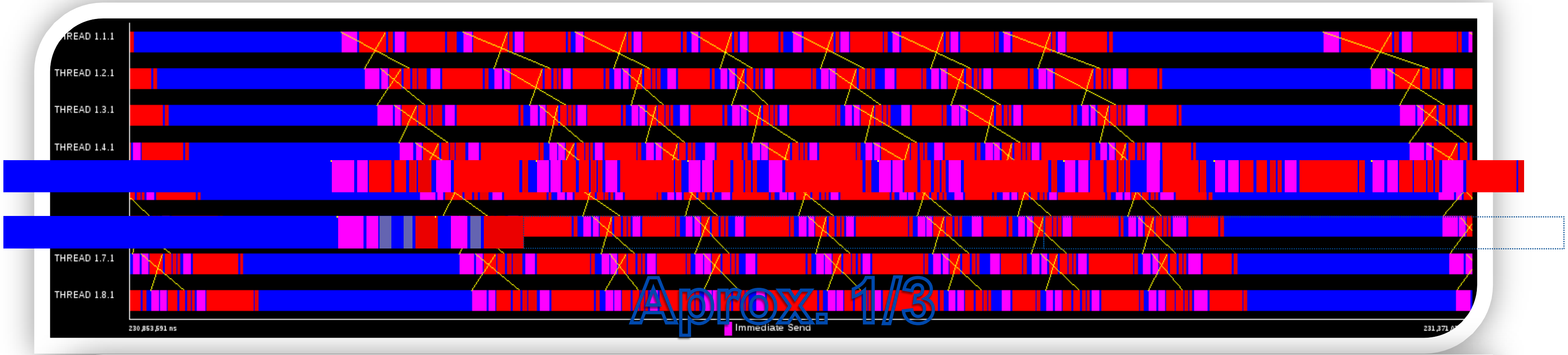
Global communications reduction

- Convergence check that is executed in every loop iteration
- Control structure put to reduce the frequency by a factor of N

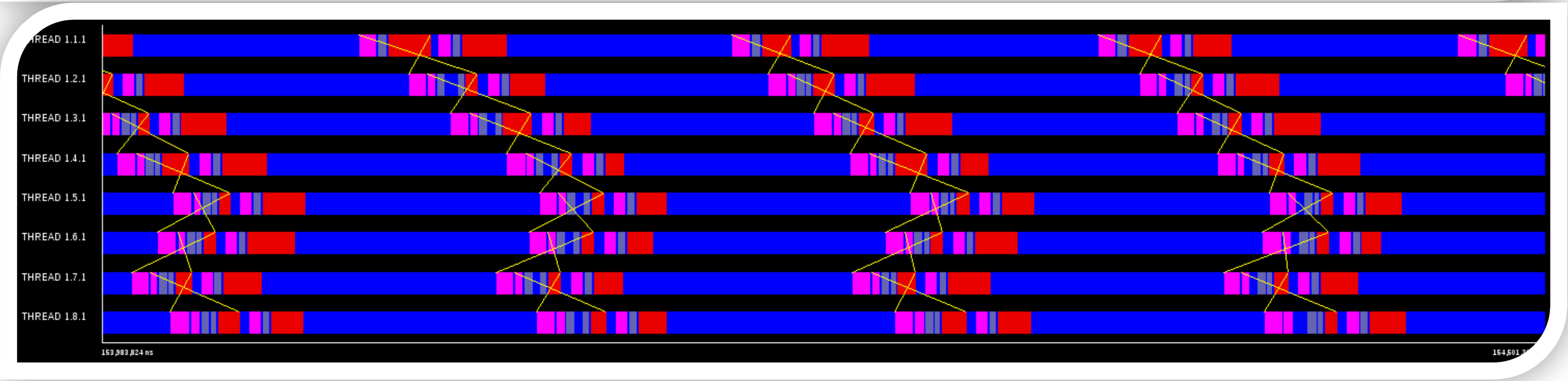


```
do while( control > threshold)    ! Sub-time step loop
   ...
      some computation with x
   ...
   !
   call interchange ( x )  !  lateral boundary condition
   !


   control = max( x )  ! Find local max
   !
   call global_max( control )   ! Find global max
   !


end do                  ! end of sub-time step loop
```

Message packing

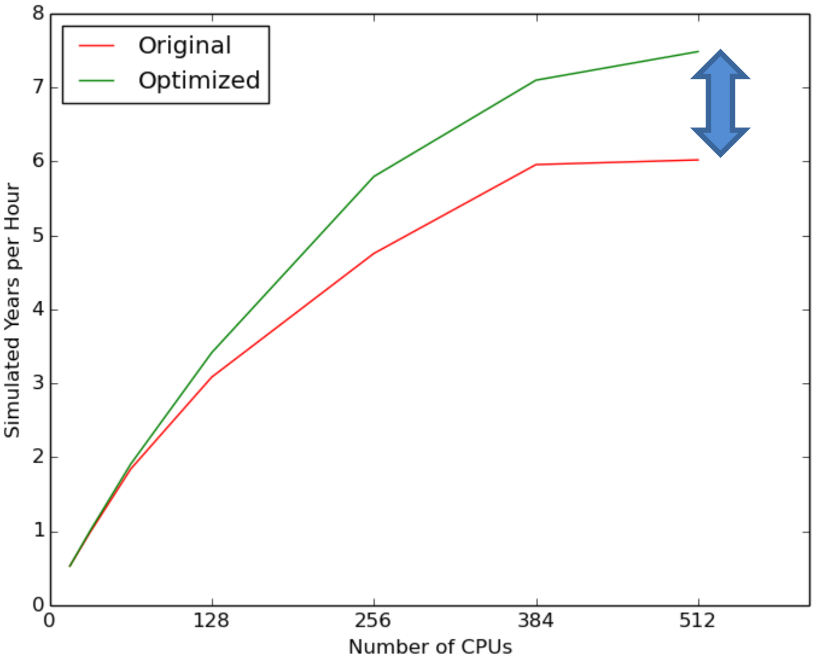


Aprox. 1/3

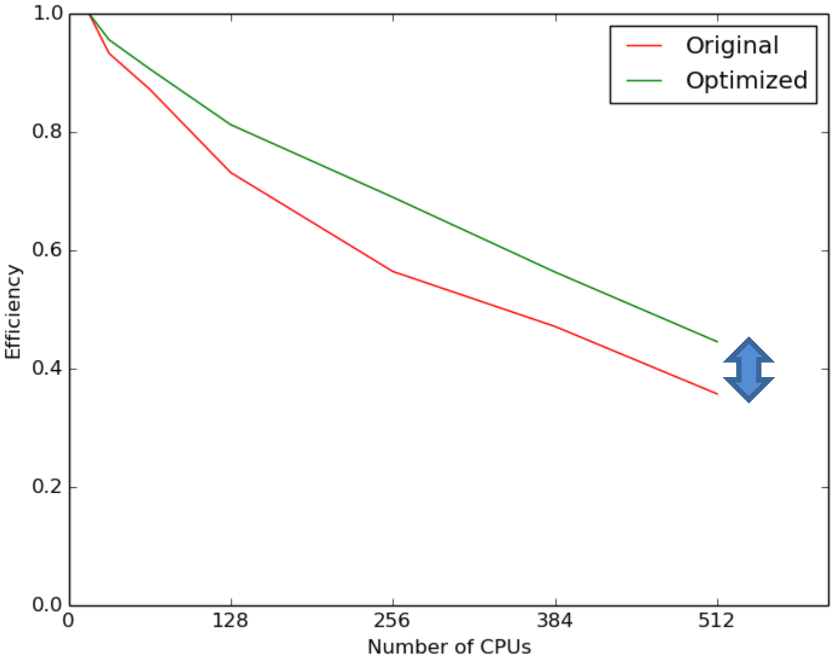


Performance improvement

- Message Packing + Reduction of global communications
- Increase in the model scalability and efficiency



Scalability Improved!

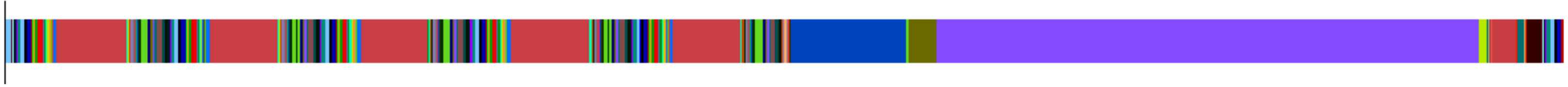


Efficiency Improved!

But it can be even better…

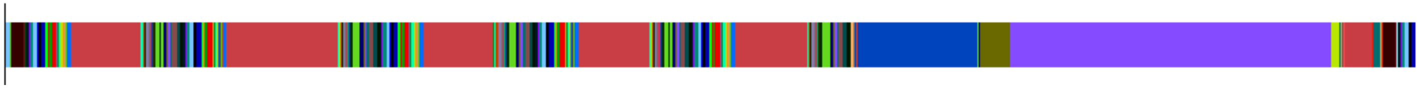
- Even though, in each LIM iteration we have:
  - 41 **lim_hdf** calls
  - More than 1400 collectives and border interchanges
- **lim_hdf** calls are *(almost)* independent → Reorder it to achieve coarser granularity and reduce collectives number by using the message packing

Original



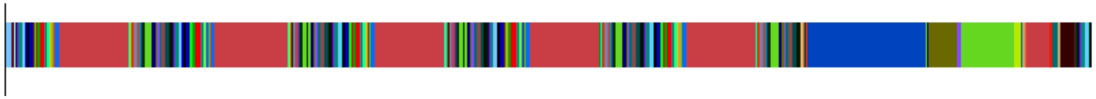
9390 Isend/Recv – 1163 All_Reduce

Message packing + Convergence check



9659 Isend/Recv – 393 All_Reduce

Message packing + Convergence check + lim_hdf multiple



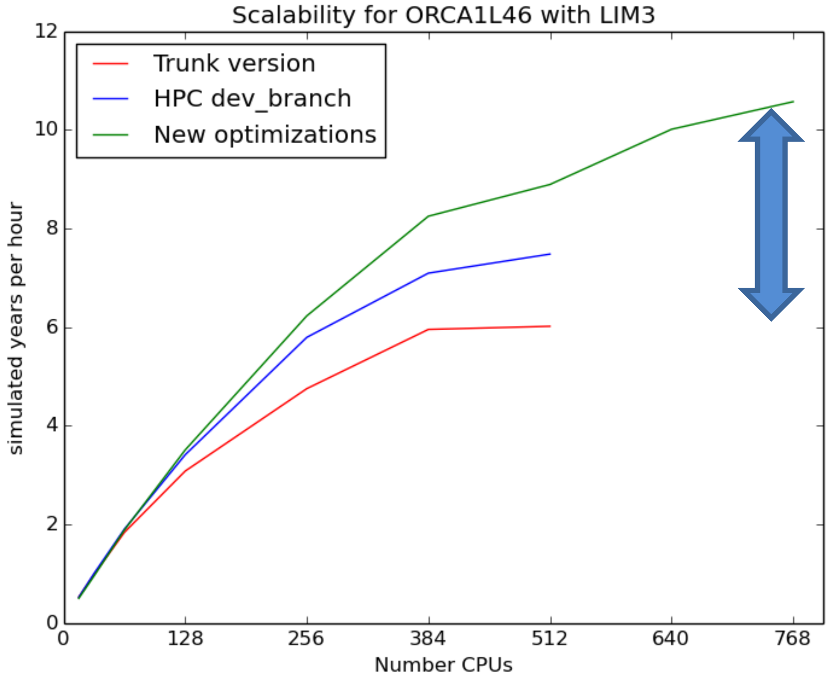
6432 Isend/Recv – **163** All_Reduce

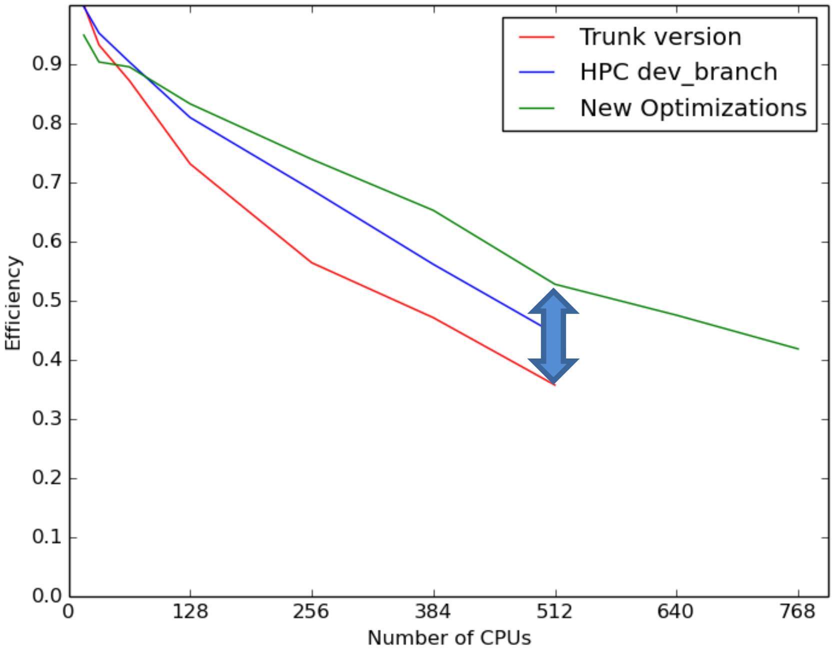
-32%! -86%!

Just reordering one routine!

More performance improvement

- Previous optimizations already included in a new model branch and now are merged into the NEMO 3.6 trunk
- New optimizations increase further scalability





Scalability Improved!

Efficiency Improved!

Conclusions

Conclusions

- **Little changes** in the configuration can significantly improve the performance

- **Trace analysis** can **guide** the **users** in understanding the behavior of the code

- A precise analysis and prediction can generate ideas that **direct** the **restructuring** of the application in the most productive way

- Performance analysis is critical for a **rational usage** of the **resources** www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

Thank you!

For further information please contact
miguel.castrillo@bsc.es