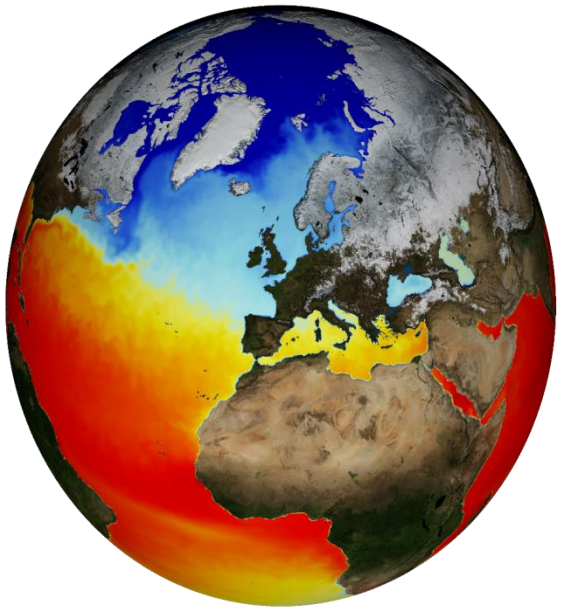
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Performance tools for climate models optimization

Miguel Castrillo, Oriol Tintó-Prims, Harald Servat, Germán Llort, Kim Serradell, Oriol Mula-Valls, Francisco Doblas-Reyes

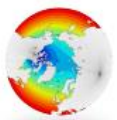
Efficiency in Earth Science models

- Especially critical in Earth science models.
- Simulations use a **huge amount of computational resources**.
- Future simulations will need much more resources.
- NEMO: **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.
- Almost 170.000 lines of FORTRAN 90 code
- Parallelization based on spatial domain decomposition through MPI.
- Mostly small stencil element calculations



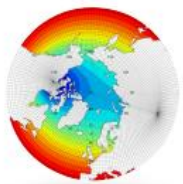
ORCA 2
550 MB of memory
8 CPU hours
10 Gigabytes of output (daily)

0,84M points



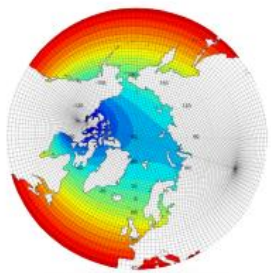
ORCA ¼
47 Gigabytes of memory
3500 CPU hours
120 Gigabytes of output (daily)

67,72M points



ORCA 1/12
414 Gigabytes of memory
90 000 CPU hours
1 Terabyte of output (daily)

991M points

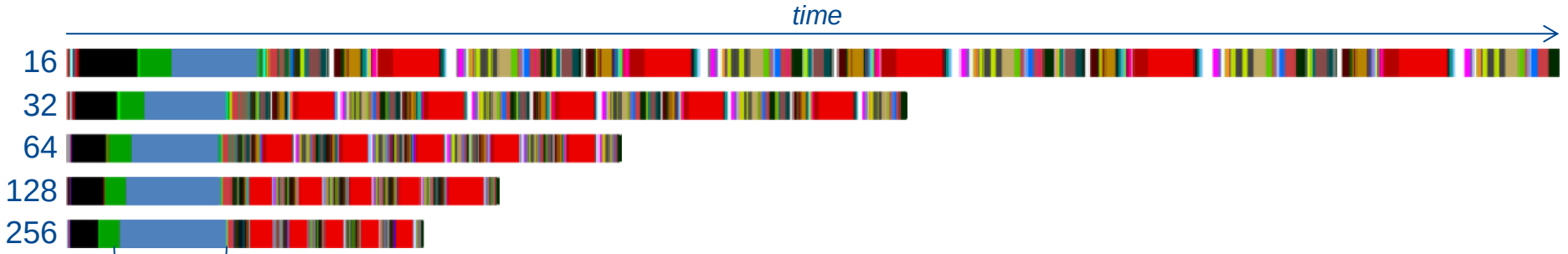


ORCA 1/36
> 1 Terabytes of memory
~4 000 000 CPU hours
> 5 Terabytes of output (daily)

~2000M points

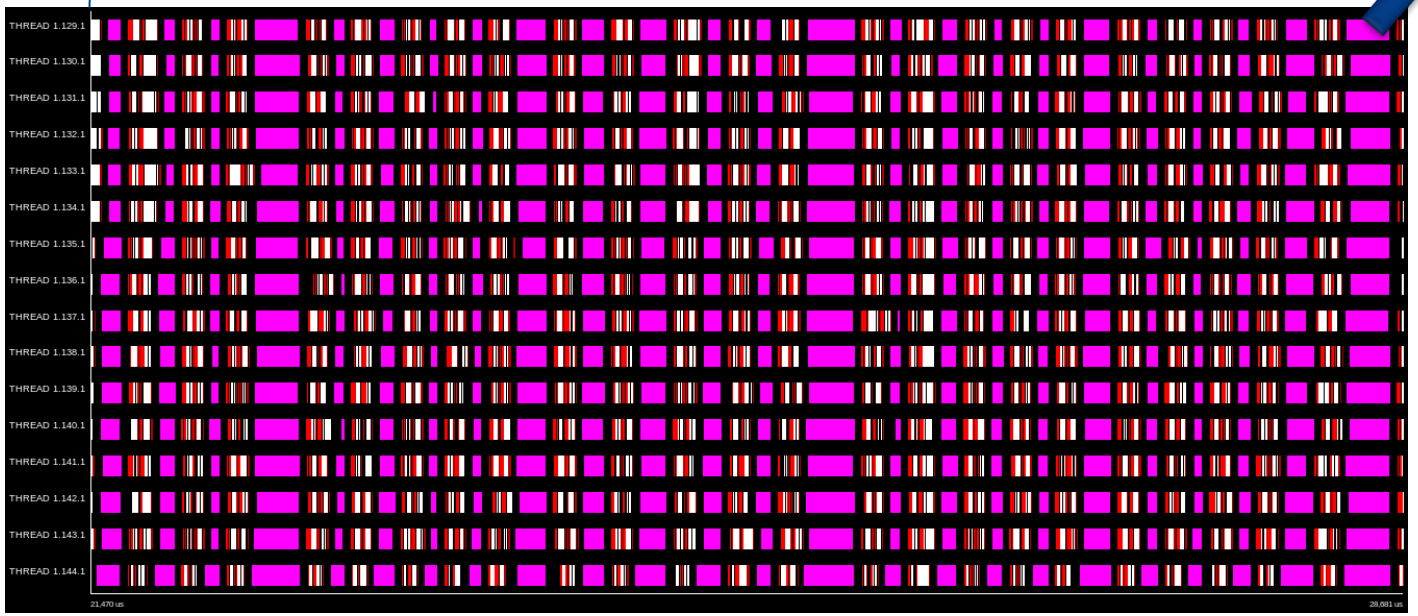
Methodology: PARAVER trace analysis

Model scalability overview: Routines analysis → Each color represents one different routine



LIMHDF (Sea-ice horizontal diffusion) → Bad scalability & big portion of time spent

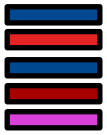
LIMHDF communications



Global Communication **every** loop **iteration**

60% of time

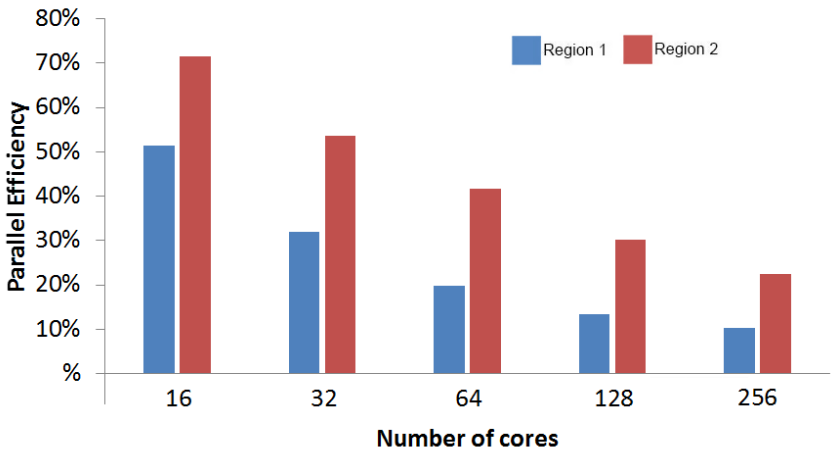
Only **20%** time invested in **computation.**

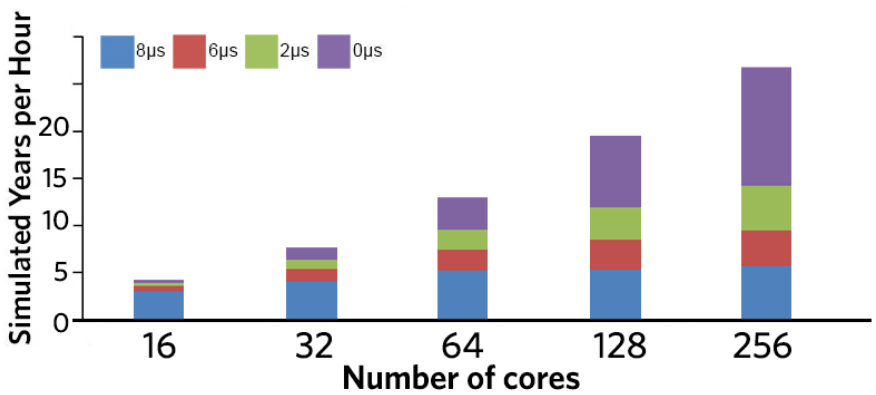


Outside MPI
MPI Isend
MPI Recv
MPI Wait
MPI Allgather

Cause of efficiency loss

- Cluster techniques used to identify different computational trends in the regions of interest → Decrease of computational efficiency due to code duplication.
- However, communications are the main performance problem. Even in the 16-core case parallel efficiency is really bad.
- The figure at the right shows how sensitive the model is to network latency.
- Communications efficiency drops much more faster than computational.

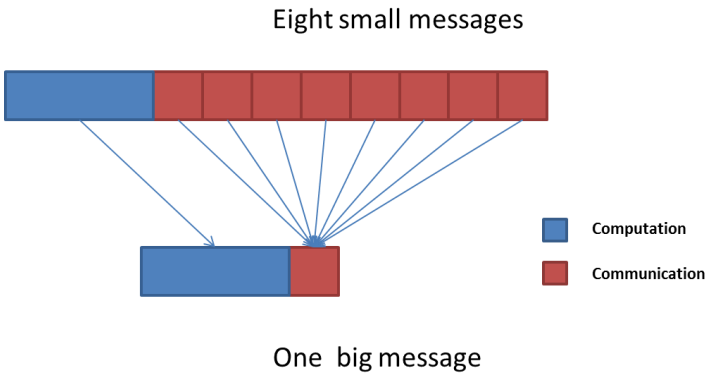




Optimizations

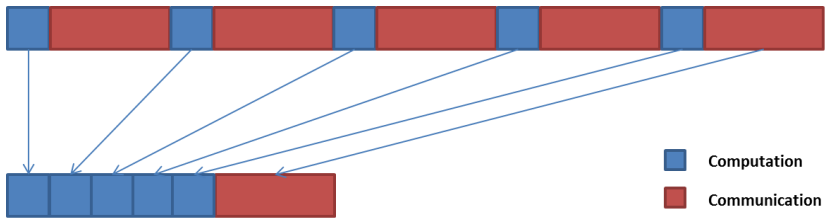
MPI message packing

Taking in account that NEMO is really sensitive to latency, messages aggregation is the best way to reduce the time invested in communications. Therefore, consecutive messages have been packed wherever the computational dependencies allow to do so.



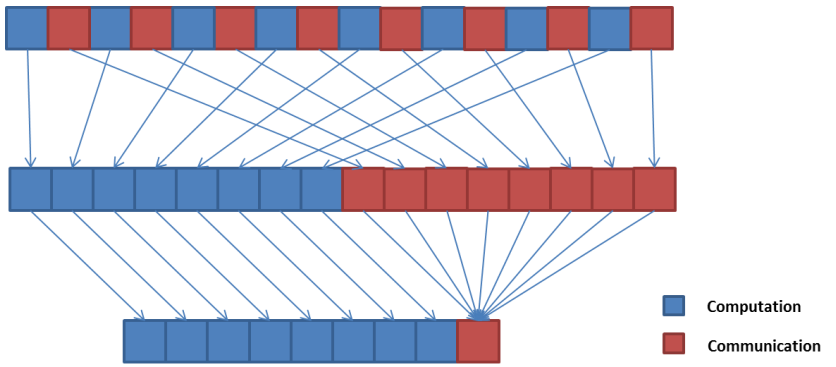
Convergence check reduction

Some routines use collective communications to perform a convergence check in iterative solvers. The cost of this verifications is really high, reaching a 66% of the time. Wherever the model allowed it, we reduced the frequency of this verifications in order to increase parallel efficiency.



Reordering

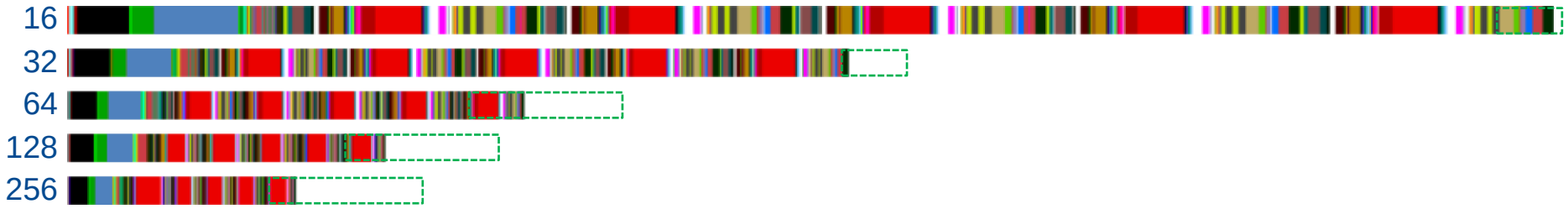
In order to apply the message packing optimization to as many routines as it was possible, it was necessary to rearrange some computation and communication regions, taking into account the dependencies between them , to reduce the number of messages. This way it was possible to compute (and communicate) up to 41 variables at the same time, resulting in a dramatic reduction of the granularity.

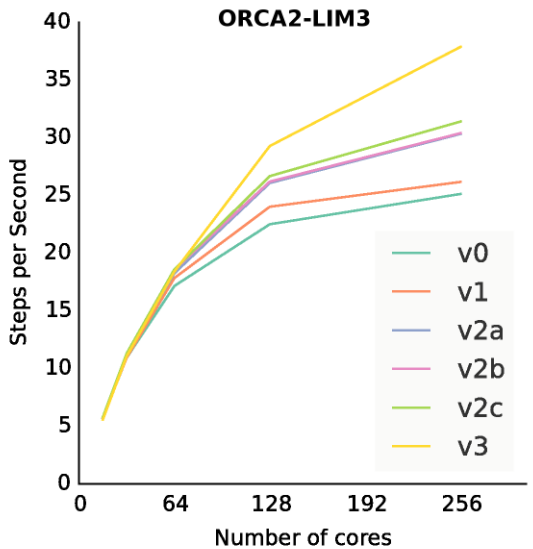


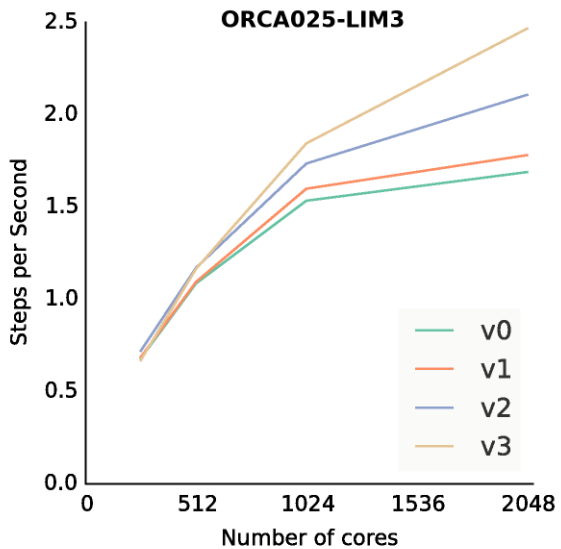
Results

Original code

Optimized code







V0 → Original
V1 → Message packing
V2 → Conv. Check reduction
V3 → Reordering

Work done & future plans

- What has already been done? 2014-2015 (Highlights)
  - Collaboration with the NEMO dev. team, now NEMO HPC Group
  - 2 optimization branches created, both already merged to 3.6 stable and trunk (June 2015, Apr 2016)
  - Technical memorandum (Autumn 2015) and poster at SC15
  - Collaboration @ NEMO merging party ‘16
- Ongoing work
  - Paper: “Finding, analyzing and optimizing MPI communication bottlenecks in Earth System models”
  - Create benchmark configs. → ORCA025-LIM3 (BSC) ready
  - Testing OPA-SAS to decouple ocean and sea-ice.
  - Testing removing land-only subdomains
  - ORCA025 (HiRes) tests
- Future plans
  - New algorithmic approaches (higher order equations)
  - Overlap communication and computation (tasks? Would need OpenMP)

www.bsc.es

Thank you!

For further information please contact
name@bsc.es