Optimization of EC-Earth 3.2 Model
K. Serradella, M. Castrillo and M. Acosta*

*Barcelona Supercomputing Center
C/Jordi Girona 29, 08034-Barcelona, Spain

Abstract

The increase in capability of Earth System Models (ESMs) is strongly linked to the amount of computing power, given that the spatial resolution used for global climate experimentation is a limiting factor to correctly reproduce climate mean state and variability. However, higher spatial resolutions require new and High Performance Computing (HPC) platforms, where the improvement of the computational efficiency of the ESMs will be mandatory. In this context, porting a new ultra-high resolution configuration into a new and more powerful HPC cluster is a challenging task, involving a technical expertise to deploy and improve the computational performance of such a novel configuration. In this paper, we focus on the work done in the context of a PRACE Preparatory Access Project, aiming to optimize the T1279-ORCA12 configuration of the EC-Earth 3.2 coupled climate model. In this case, all runs have been performed in the MareNostrum IV supercomputer of the Barcelona Supercomputing Center.

Introduction

In the last decade, our understanding of climate change has increased [1], as society’s needs grow for advice and policy. However, whilst there is a general consensus that climate change is an ongoing phenomenon, there remain uncertainties [2], for example, on the levels of greenhouse gas emissions and aerosols likely to be emitted, or perhaps even more significantly, on the degree of warming and the likely impacts[3]. Increasing the capability and comprehensiveness of ‘whole Earth system’ models (ESMs), in order to represent with ever-increasing realism and detail new scenarios for our future climate, is the only way to reduce these latter uncertainties [4].

However, the increase in capability of ESMs is strongly linked to the amount of computing power and data storage capacity available. Previous works [5] have shown the benefits of increasing the horizontal resolution of a global ESM for seasonal climate prediction, and therefore there is a vital need for more high-performance computing in order to predict the future evolution of climate [6].

Most applications targeting machines with huge computational power require some degree of rewriting to expose more parallelism, and many of those face severe strong-scaling challenges if they are effectively to achieve this improvement in computational performance, as it is demanded by their scientific goals. There is an ongoing need of support for software maintenance, and tools to manage and optimize workflows all across the infrastructure. As a result, a priority for this work is to continue the development of these ESMs and their infrastructure to be able to run ultra-high-resolution experiments (10-20 km range) focused on new climate processes.

* Corresponding author. E-mail address: kim.serradell@bsc.es; miguel.castrillo@bsc.es; mario.acosta@bsc.es
One of these ESMs, which urgently needs to improve its capability at the higher resolutions, is EC-Earth. EC-Earth is the European global climate community model [7], based on IFS, the world-leading weather forecast model of ECMWF (European Centre of Medium Range Weather Forecast) in its seasonal prediction configuration, along with NEMO, a state-of-the-art modelling framework for oceanographic forecasting and climate studies, which is developed by the NEMO European Consortium [8]. EC-Earth is one of the models chosen by most of the meteorological centers and research institutes across Europe. Among other purposes, this model is used to provide information to address the regional impacts of climate change and determine appropriate adaptation and mitigation measures on a more regional basis.

The Barcelona Supercomputing Center (BSC) has developed a coupled version of EC-Earth 3.2 at a groundbreaking resolution. In the atmosphere the horizontal domain is based on a spectral truncation of the atmospheric model (IFS) at T1279 (approx. 15 km globally, i.e. the highest resolution we can use with the standard IFS - higher resolutions would require e.g. nonhydrostatic parameterizations) together with 91 vertical levels. The ocean component (NEMO) is run on the so called ORCA12 tripolar (cartesian) grid at a horizontal resolution of about 1/12° (approximately 16 km), with 75 vertical levels, whose thickness increases from 1m below surface up to 500m in the deep ocean.

One of the main goals of this work is to achieve a good scalability of the EC-Earth model running on this new T1279-ORCA12 ultra-high resolution.

In particular, the goal of the performance activities in this preparatory access were mainly adapting the new configuration to be more parallel, scalable and robust. These were performed through three consecutive steps:

1. Deployment of EC-Earth and subsequent steps needed to run the new ultra-high configuration on a novel HPC platform.
2. Localization of some possible bottlenecks of EC-Earth when this configuration is used and provide possible solutions for optimization.
3. Study the scalability of EC-Earth in order to evaluate the ideal number of processes for the particular configuration.

Deployment of an ESM in a state-of-the-art HPC facility

This optimization work was carried out in MareNostrum IV, the new HPC cluster in BSC. Therefore, a prior step consisting in deploying and testing the model on the new environment had to be performed. This task had to be done before being able to start the scalability tests.

MareNostrum IV was deployed during the second quarter of 2017, replacing MareNostrum III and is equipped with the new Intel Scalable Processors and the new Intel Skylake microarchitecture.

EC-Earth is an Earth System model made up of different components, each one having different dependencies, so the first step for deploying the model in any new machine is to ensure all the needed libraries and utilities are installed and operating. This work was done in coordination with the Support team of the BSC.

Although in many scenarios this work may be done in an automatic or semi-automatic way by using deployment frameworks [9], it is not rare to experience problems between some libraries and architectures, especially when they are being installed in a brand new HPC. In our case, the biggest problem came from a writing problem in the NEMO model, only reproducible when using an instance of the netCDF library linked to PnetCDF, which enables parallel I/O for CDF-1, 2, and 5 formats. Therefore, we use a version of netCDF not linked to PnetCDF.
Being the main novelties in MareNostrum IV the Skylake processors and the OmniPath network, most of the tests were aimed at assessing the efficiency and resiliency of the new network fabric (OPA), as also the exploitation by the model of the new AVX-512 instructions, available in the new Intel Scalable Processors.

Regarding the advanced vectorization instructions, previous tests with the standalone version of the NEMO model showed up to x10 speedup when enabling the 512-bit instructions, decreasing with the number of cores and the ratio between computation and communication. Independently of the performance benefits, testing the AVX-512 instructions in the EC-Earth model revealed numerical issues in some routines of IFS when the O2 or O3 optimization flags were also being enabled. By using the ARM DDT debugger, some strange values in the problematic regions of the code were found, which seemed to be due to bad automatic optimization of some loops. This hypothesis was verified when it was proved that using the !DIR$ NOVECTOR compiler directive in that loops the problem was not being reproduced. However, the AVX-512 flag is not used neither in our following tests nor in our operational runs until it is demonstrated that the mentioned fix is safe in terms of model results.

Working with new OPA network has been a challenge at the time of coming up with a stable and reliable configuration. At several stages we have experienced different kind of communication issues, from model crashes to random stops that hampered our operation. In collaboration with the Support team we tried several networking environment variables to work on network internals, sometimes to change the communication fabric at runtime, other times to use a different version of the PSM2 library (Performance Scaled Messaging by Intel) or change its configuration.

| export I_MPI_ADJUST_BCAST=3 |
| export PSM2_MTU=8196 |
| export PSM2_MEMORY=large |
| export PSM2_MQ_RNDV_HFI_THRESH=1 |
| export I_MPI_DEBUG=5 |
| export LD_LIBRARY_PATH="/apps/PSM2/10.3-37/usr/lib64:$LD_LIBRARY_PATH" |
| export I_MPI_FABRIC=tmi |

Figure 1: Intel MPI environment variables being used to run EC-Earth production runs in MareNostrum IV.

XIOS, the NEMO asynchronous I/O server, makes an intensive use of MPI polling mode (busy wait). At the same time NEMO also performs a lot of communications. Our hypothesis is that using these components with a high number of processors increases the chances of producing errors in the underlying OPA layer, which seems to be not suitable for this kind of communication.

At the end of this study, our operational high-resolution runs have 0.86 ratio of success when using ~1.5K processors, and ~0.98 ratio of success when using ~0.5K processing units.

At this point, once the most stable versions and compilation options for the different library dependencies were found, and additionally, the appropriate MPI environment variables for running EC-Earth in MareNostrum IV were identified, it was time to ensure that the model was using a balanced configuration, as well as an optimal domain decomposition, for each case.

**Optimizing ESMs performance**

A very important factor for the performance of parallel computational models in general, and of ESMs in particular, is the horizontal domain decomposition, due to its tight relation with the amount of computation (workload distribution, redundant computation), and communications (ratio of inter-node communications) among the parallel resources used. Previous optimizations have been developed at BSC and applied to the NEMO model [10][11].

For the NEMO model this is a critical issue, given that this model by default performs the same computations in land-only subdomains than over the ocean ones and then discard the results using a
mask. In order to find an optimal domain decomposition for this model, the ELPiN tool, recently developed by our team, was included in the workflow of the EC-Earth model and reported to EC-Earth consortium. In NEMO the way to enable land-subdomain removal is really cumbersome, because the user has to know in advance the exact number of ocean subdomains contained in a given decomposition of the spatial domain. ELPiN (Exclude Land Processes in NEMO) performs this calculation and at the same time finds the best amount of resources and horizontal subdivision (which determines the border/surface relation) by considering the overlapping inherent to the usage of ghost cells.

In order to include ELPiN in the EC-Earth community repository, its configuration had to be adapted to the EC-Earth standards, which makes use of a semi-automatic compilation and set-up system. At the same time, it was improved the procedure to require minimal input, and to be able to run from the model running directory, automatically detecting the configuration used and the bathymetry files. With the aim to save resources, the tool was also adapted to save performed calculations to cache, in order to avoid extra computation in future runs.

Speaking about the I/O side, also some work was done to find a right configuration for the XIOS I/O server, used by NEMO to write output data. Indeed, XIOS has a high memory consumption profile and increasing the resolution forced us to run some tests to determine the right number of I/O processors in order to avoid memory crashes. For the T1279-ORCA12 configuration it was necessary to use 192 XIOS servers in 4 nodes.

Once the model was efficiently performing, last actions were devoted to the optimization of the data workflow. This was done by reducing the data movement in the EC-Earth post-processing stage, another issue that was detected as a possible bottleneck in our study. After the optimization, post-processed data is directly moved from the HPC scratch directory to the archive without staying in any intermediate storage, and using the data transfer nodes for that operations, achieving a 10x speedup in comparison to transfers out of the GPFS file system. As the commands required to use the Marenostrum IV data nodes are non-blocking versions of their counterparts, submitted to a special queue, it was needed to adapt the workflow manager [12] templates to control the evolution of the transfers, as also the correctness of the operations.

Methodology for assessing ESMs scalability

The scalability tests performed within the scope of this project had the objective of evaluating both the strong and weak scaling of the EC-Earth model, with the aim of running operationally the proposed configuration (T1279-ORCA12).

Being EC-Earth a coupled model, strong scalability tests become a two-dimensional problem, in which for every increment in amount of resources, the balance between the two major components (IFS & NEMO) has to be achieved.

To find the more balanced configurations for every given amount of resources, two different but complementary approaches were followed. The first, costlier, tries to find the optimal distribution by assigning the same number of processors to IFS & NEMO first, and then moving resources from one model to the other alternately, identifying the intervals in which the model performance increases, by using a variation of the half-interval search algorithm. The second, much cheaper, starts from one separate scalability test for each one of the components that then is used to determine the best configuration. In our case, the first method could be used to validate the conclusions taken from the second.

For the weak scaling exercise, two different configurations were tested, namely T511-ORCA025 (~30 km) & T1279-ORCA12 (~16 km), the latter having approximately 9x more points that the former.
Results

Figure 2 shows the scalability of EC-Earth for a different number of processes for the two main components (IFS and NEMO), using the T511-ORCA025 configuration. The results are evaluated in Simulation Year(s) Per Day (SYPD), a very popular metric in Earth System Modelling. In this case the model is able to scale up to 3,456, where the efficiency drops off 70%.

Figure 3 shows similar results for the T1279-ORCA12 configuration. In this case, the “Ideal” series (orange color) show perfect speedup from the smaller case, while “Ideal II” (yellow color) starts from the more efficient one. The first configuration is affected by resource contention, leading to a superlinear speedup, that is not the case when we start from the more efficient configuration. Figure 4 and Figure 5 show different visualizations from model results running this configuration.

Figure 2: Strong scaling of EC-Earth 3.2 using a T511-ORCA025 configuration (75 vertical levels, 900 s. timestep) with production output data. The throughput is shown SYPD.
Figure 3: Strong scaling of EC-Earth 3.2 using a T1279-ORCA12 configuration (75 vertical levels, 360 s. timestep) with production output data. The throughput is shown in SYPD.

Figure 4: Left, Global Sea Surface Temperature of the ocean component NEMO. Right, Global Speed Wind at 10m of atmosphere component IFS.
Sharing results with EC-Earth community

EC-Earth is developed in the framework of a European consortium where more than fifteen partners collaborate together in the development of the model. Each institution, based on their interests, develops different modules. The work can be tracked and discussed in a development portal, being a “Technical Working Group” in charge of the coordination of the technical aspects.

All the work performed under this project has been shared with the community. Exchanging this kind of information is really useful for any other institution working with EC-Earth, especially if they are users of similar hardware. As an example, a direct collaboration with CINECA was established to support them in solving issues related to EC-Earth runs in Marconi. Having this HPC a similar architecture compared to MareNostrum IV, many of the tweaks tested and validated in MareNostrum IV were also useful in CINECA. In addition, the development team working in EC-Earth in Royal Netherlands Meteorological Institute (KNMI) has applied some of the developments presented in this paper.

These examples show how the community benefits from a preparatory access, reaches different research institutes and therefore, increases the value of the work done.

Summary

MareNostrum IV is a great and powerful machine but the choice of the brand new OPA network had a strong impact in the deployment of our coupled climate code. For the case of this project, we surely have payed a price for being “early adopters” of that technology. On the other hand, this increased our technical knowledge of the application. As EC-Earth is the current production application, and will stay there for years, this has been a valuable exercise.

This task entailed a lot of testing and debugging, having constant communication with the Support team in BSC Operations department, as also high use of the ARM DDT debugger. Some MPI stability issues also required of using different MPI variables, finally suggested by Operations, that had to be tested to find their reliability.

The effort of including the ELPiN tool in the EC-Earth compilation and running workflow, allowed to share it in the community repositories, so it will be used in upcoming projects. This will help to save
millions of computing hours in the high-resolution simulations done for all the institutions in the consortium.

There is still much to do for this configuration which is a key piece of H2020 projects as PRIMAVERA. The amount of SYPD that can be achieved is still low and should be increased to produce results in a sustained throughput. In this sense, we are currently performing another scalability exercise using a configuration of the model with reduced output, in order to compare it with the results obtained from this project. This will allow us to better understand the impact of the output at high resolution configurations, especially on the IFS model, which does not count with an I/O server in EC-Earth, as NEMO does.

References


Acknowledgements

This work was financially supported by the PRACE project funded in part by the EU’s Horizon 2020 Research and Innovation programme (2014-2020) under grant agreement 730913.