Exploring the use of mixed precision in NEMO

Oriol Tintó Prims[#1], Miguel Castrillo[#2]

[#]*Earth Sciences Department,Barcelona Supercomputing Center, Barcelona (Spain)*

[1]oriol.tinto@bsc.es, [2]miguel.castrillo@bsc.es

***Keywords*— Mixed precision, Ocean model, Climate Sciences**

EXTENDED ABSTRACT

It has been a widely extended practice in scientific computing to use 64-bit to represent data without even considering which level of precision is really needed. In many applications, 32-bit should provide enough accuracy, and in other cases 64-bit is not enough. In climate science, the inherent difficulties collecting data imply a considerable level of uncertainty, which suggest that the general use of 64-bit to represent the data may be a waste of resources, while on the other hand, some specific algorithms could benefit from an increment of the precision used. These factors suggest that in the future more attention has to be paid to the precision used in scientific software, to use the resources wisely and also avoid losing accuracy. In this work we question whether the precision used in the oceanic model NEMO is necessary and sufficient, and the potential benefits of adjusting this precision.

A. The model

The Nucleus for European Modelling of the Ocean (NEMO) is a state-of-the-art modelling framework for oceanographic research, operational oceanography, seasonal forecast and climate studies, which is used by a large community: about 100 projects and 1000 registered users around the word. It is controlled and maintained by an European consortium, made up by CNRS and Mercator-Ocean from France, NERC and Met Office from the United Kingdom and CMCC and INGV from Italy.

Several millions of computer hours are invested each year in simulations involving NEMO, resources valued in tenths, or even hundreds, of thousands of euros.

The core of NEMO is the OPA module that solves the Navier-Stokes equations from regional to global scales, using Euler first-order discretization methods on a three-dimensional (3D) grid. The model was parallelized and is able to be executed in both shared and distributed memory environments, using MPI to overcome data dependencies.

B. Is the use of 64-bit data justified in NEMO?

While it can be affirmed that high levels of precision could be used only on sensitive calculations, the generalized lack of background in numerical analysis provoques that it often has been much easier and cheaper to employ high-precision arithmetic than to take care of the necessary precision [1]. However, having the cost of scientific computations grown in several orders of magnitude, the possibility to employ mixed precision algorithms has won a lot of interest. These algorithms can provide performance benefits while maintaining the accuracy [2][3].(

Moreover, while it is clear that reducing the overall precision of a model will unequivocally increase the arithmetic and rounding errors, whether this errors are important or not has to be evaluated. In a computational model, arithmetic and rounding operations are not the only sources of error, faas truncation and data uncertainty can also harm the accuracy of the results. If the errors coming from the other sources are important, spending resources in increasing the precision may not be worth.

This is the case in climate science, where the inherent difficulties to collect the data and the chaotic nature of the simulated systems suggest that minimizing arithmetic errors may not benefit the accuracy of the results, and therefore reducing the precision from the default 64-bit to 32-bit may not be as harmful as to renounce to the performance benefits that it can provide.

C. Precision and computational performance

There is a direct relation between the precision used and the cost of a computation. Most modern processors have vectorial instructions for 32-bit and 64-bit data, performing the first the double of operations per cycle than the second. Moreover, the gap between processor and memory speeds implies that the time to bring the data to the CPU can be as long as the time to compute the operations itself in the CPU and in some cases becomes the computational performance bottleneck. So finally in both CPU bound and memory bandwidth bound computations, the expected difference between using 32-bit and 64-bit representation is about a factor of 2.

D. Preliminary approach for NEMO

To explore how the use of mixed precision can benefit the computational performance of this model the first approach is to move the whole model from 64 to 32-bit wherever it is possible.

Although at the first stages of the model development the precision was intended to be easily selected by only changing one single parameter, later developments did not respect that approach and several changes had to be done to change the precision. As a consequence, some issues had to be solved before obtaining a compiling version of the code.

Nevertheless, this preliminary version of the model in 32-bit crashed in several places, but luckily enough, when the changes cause the model to crash, it was easy to track where the floating point exceptions are occurring. With this trial-and-error approach, we determined that the module corresponding to sea-ice was especially problematic and that

could be convenient to keep this part of the model in 64-bit due to the higher sensitivity to the precision used. After solving the issues derived from having variables of different precisions in the same code, it was possible to obtain a running version able to produce results.

E. Preliminary results

With this early-stage mixed precision version of the code, it is possible to evaluate the potential impact of the changes in the computational performance. To do so we compared the model throughput of the 64-bit and mixed-precision versions in three different use cases: a low-weight low resolution configuration in a work-station and a high resolution in a supercomputer. The results are exposed in Table 1, Table 2 and Figure 1.

TABLE I
MODEL THROUGHPUT IN A WORKSTATION

Precision	Throughput (Steps per Second)
Double	1.31
Mixed	2.06

Table 1 Model throughput of a low resolution ORCA2 simulation in an i7-4790 CPU @ 3.60GHz workstation using 4 of 8 cores.

TABLE II
MODEL THROUGHPUT IN MARENOSTURM III

Number of cores	Throughput (Steps per Second)	
	Double	Mixed
256	0.75	1.09
512	1.43	1.92
1024	2.39	3.15
2048	3.14	3.87

Table 2 Model throughput of a High Resolution ORCA025 configuration in Marenostrum III supercomputer at Barcelona Supercomputing Center..

It can be seen that in the workstation case (Table 1), the difference in throughput is 57% faster in mixed precision. In the high resolution case, with 256 cores the mixed precision version is 45% faster, while in the 2048 core case it is only 23%. This is explained because in low core-counts the computation has much more weight and in large core-counts the communication becomes more important.

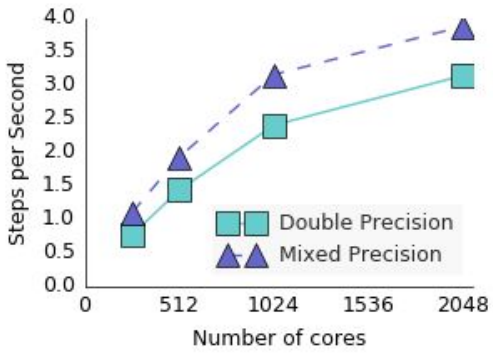

Fig. 1 Model throughput for different number of cores with the double and the mixed precision codes with the high resolution ORCA025 configuration in the Marenostrum III supercomputer..

F. Issues

Although the model can run and produce results, these present significant differences compared to the double precision outputs. It is out of the scope of this work to evaluate these differences, but for sure it will be necessary for further studies.

G. Conclusions

The cost of the resources invested in simulations including NEMO can be estimated to be of the order of hundreds of thousands of euros. Adjusting the models to avoid the overuse of double precision representation generally has been proven to be a solution to reduce considerably the cost of the simulations. This exploratory work proved that using single instead of double precision speeded up the model a 59% in a workstation as also achieved an important improvement in high resolution simulations in Marenostrum III supercomputer, and therefore suggests that this must be studied further.

H. Future work

This study is only the necessary starting point to develop a mixed precision version of the code that uses the necessary precision in each part of the model, allowing an important reduction of the computational cost of the simulations and increasing the quality of the results. To do so it is necessary to develop methods for analyzing the algorithms and their response to the used precision. This kind of work has been done in other models and can be orientative.

References

[1] Bailey, D.h. "High-Precision Floating-Point Arithmetic in Scientific Computation." Computing in Science and Engineering 7.3 (2005): 54-61.

[2] Baboulin, Marc, Alfredo Buttari, Jack Dongarra, Jakub Kurzak, Julie Langou, Julien Langou, Piotr Luszczek, and Stanimire Tomov. "Accelerating scientific computations with mixed precision algorithms." Computer Physics Communications 180.12 (2009): 2526-533.

[3] Leyffer, Sven , Stefan M. Wild, Mike Fagan, Marc Snir, Krishna Palem, Kazutomo Yoshii, Hal Finkel. “Doing Moore with Less -- Leapfrogging Moore's Law with Inexactness for Supercomputing”, 2016

Author biography

Oriol Tintó Prims received his degree in physics from the Universitat Autònoma de Barcelona, his Masters degree in Modelling for Science and Engineering from the Universitat Autònoma de Barcelona and is currently doing his PhD in the Earth Sciences department from the Barcelona Supercomputing Center in collaboration

with the Computer Architecture and Operating Systems department from the Universitat Autònoma de Barcelona.