Energy consumption of the NEMO ocean model measured with the Energy Scope tool

Eric Maisonnave & Isabelle d'Ast

WN/CMGC/21/88

Abstract

The Energy Scope tool, developed at INRIA Bordeaux, is used on an Intel Skylake cluster to measure the energy consumption at scale of the NEMO ocean model.

This work establishes that the NEMO model does not differ significantly from other applications regarding energy consumption, and that this consumption is weakly dependant on model resolution or parallel decomposition, conferring on scalability the quality of single criteria for energy efficiency. More analysis are required to identify possible energy bottleneck in our model.

Table of Contents

The energy consumption of ICT infrastructure is a rising issue in our society, since today its direct annual carbon footprint is approximately of 1400 Mt of CO2 equivalents [1]. The HPC research activities does not contribute significantly to these emissions [2], but some scientists put the stress on the particular leading role of Earth Science toward sustainability or, more simply mention "a personal level of being consistent with why they are a scientist in the first place" [3]. These questions partially encounter the concerns of the silicon industry on the path to energy efficient platforms that could match the Exascale constraints [4].

Since partners of the NEMO consortium already started to evolve towards more sustainable behaviours [5], we propose to contribute to the energy consumption evaluation of the NEMO ocean model [6], one of the main components of the CNRM-CM6 [7] climate model currently in use in our laboratory. In Berthoud at al., the carbon footprint of one core.hour of computing is evaluated for an average usage of the machine (Intel Skylake Gold 6130 CPU @ 2.10GHz), regarding CPU, memory, network and disk access. The power consumption is equal to 16,5 W per core, taking PUE (Power Usage Effectiveness) into account. We are convinced that a first evaluation of this kind, but focused on the energy/carbon footprint of our models and not on the machine, is necessary to better understand their capacity to efficiently take benefit of the computing resources, in comparison to other models. In a second step, it may be possible to identify the key actions needed to enhance this capacity.

Experimental setting

Several techniques are available to evaluate the consumption of a software. A CERFACS' production machine[1] is targeted to host the test, which forbids the use of a direct measurement of the electric consumption on one single processor independently of the rest of the machine.

The chosen NEMO configuration relies on the 4.0.1 version, revision 10984. The BENCH testing configuration [8], made to facilitate a benchmarking exercise while keeping the computing behaviour of a production model, is including the SI3 sea ice but no bio-geo-chemistry module.

We choose an indirect measurement, via hardware counters, as proposed by the Energy Scope tool [9]. This software developed at INRIA Bordeaux, is designed to measure the energy consumption through a large number of resources, which fits with the parallel nature of our codes. An Python based acquisition package is launched by a code wrapper. It runs on the cluster during the simulation and provides measures at approximately one second frequency.

In order to be able to access to hardware counters, the `msr-tools` package should be

1 Cluster LENOVO "`kraken`" with 185 bi-socket nodes, 18 cores Intel Xeon Gold 6140 (Skylake), 2.3 Ghz, 96 Gb memory, OmniPath interconnect 100 Gb/sec

installed on the compute nodes and the read access to Message Service Routine (MSR) device has to be given[2].

Our experimental setup aims to quantify how parallelism (strong scaling) and resolution (weak scaling) affect energy consumption. In that perspective, (i) several measurements are performed on a variable number of nodes, ranging from 1 to 128 and (ii) two BENCH configurations are used, based on ORCA1 (1 degree horizontal resolution) and ORCA025 (¼ degree horizontal resolution) discretisation and their corresponding namelist parameters.

Measurements

A previous study [10] has noticed the strong discrepancy in NEMO energy consumption during checkpoint-restart operations, in comparison to the compute phases. Unfortunately, the checkpoint-restart only occurs at the beginning and the end of large simulations, and the time spent in these operations, even if increasing with horizontal resolution, can be largely neglected in comparison to the compute phase. Thanks to the tag option provided by the Energy Scope toolkit, we bound our measurements to the code time loop, excluding not only the checkpoint-restart operations but the whole initialisation and termination phases. To avoid any interference caused by disk access, unreproducible by nature, data or log output are also switched off.

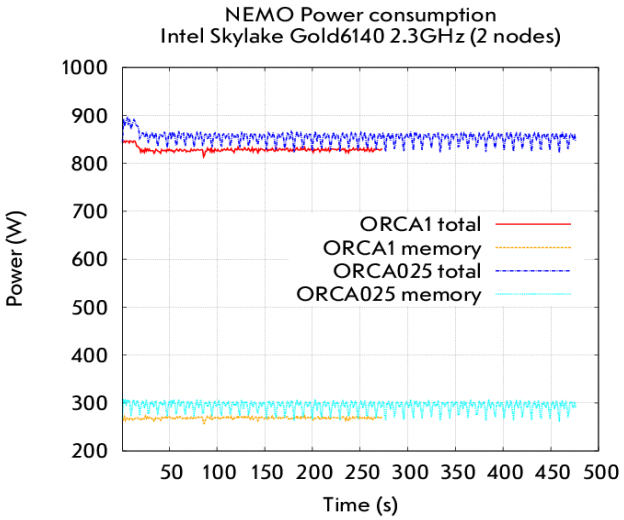

Figure 1: Power dissipation of two Skylake nodes, during resp. 1000 and 100 time steps of NEMO simulations, using resp. ORCA1 and ORCA25 configurations

As shown in Figure 1, the total consumption (CPU+Memory) along the simulation remains constant, although small drops, already noticed by Ferrero, can be seen in the ORCA025 profile. That is probably the relative slowness of the ORCA025 configuration compared to

2 The following installation steps are followed by the cluster administrator :

```
yum -y  install msr-tools
chmod a+r /dev/cpu/*/msr
setcap cap_sys_rawio=ep /usr/sbin/rdmsr
```

ORCA1 that allows to visualise these drops at this particular resolution. These drops are also present in the consumption part attributed to the memory access, which lead to the hypothesis that the line drops point out the intra time step heterogeneity of memory access. Unfortunately, limits on time sampling frequency prevent a finer per subroutine analysis and any further algorithm enhancements. That said, the detected variations are small compared to the consumption raw value and probably give evidence of poor chances of enhancement. We also noticed that only one third of the consumption can be attributed to the memory access. This ratio, as the raw consumption, is approximately the same for other parallel applications tested during this exercise, such as AVBP [11], which suggests that the NEMO code is not particularly memory bound, as usually stated in the available literature, nor more energy efficient than others.

It looks like the CPU energy consumption does not really depend on the kind of operation performed (except maybe disk access, measurable at simulation beginning and end). At the opposite, slight changes can be noticed in memory access energy consumption. We take benefit of the possible parametrisation of our Lenovo cluster by changing the bus speed mode[3], enabling the adjustment of this speed to the application needs. This leads to (i) an increase of the memory access consumption, probably over solicited at these speeds, (ii) a decrease of the CPU consumption, probably less solicited, (iii) a small reduction of the overall dissipated power at the beginning of the simulation, but an increase to the standard value after a few dozens of second and (iv) a total energy to solution less favourable (about +10%). If our model (neither AVBP), on one node, does not take benefit of this change in the cluster setting, the experiment proves, at least, that the consumption ratio between CPU and memory can be modulated and its addition possibly enhanced on other platforms.

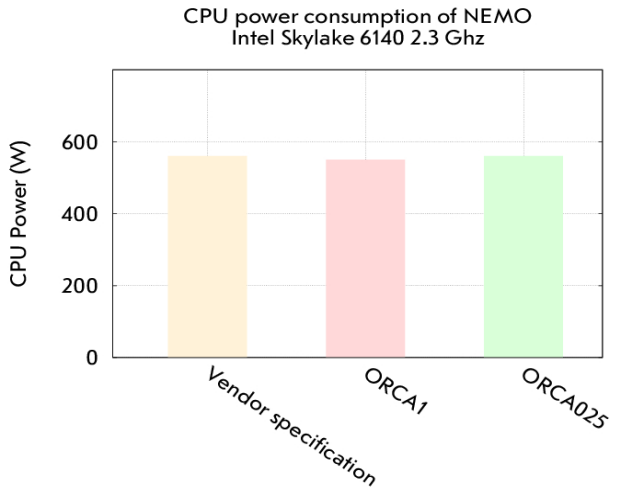

Figure 2: Average power dissipation of NEMO ORCA1 and ORCA025 configurations on two Intel Skylake nodes, compared with vendor specifications

On Figure 2, we also compare the average CPU power measurement (excluding the

3 LENOVO Firmware / BIOS / Microcode Settings (Operating Mode) : "Efficiency -Favour Power", provides the best features for reducing power and increasing performance in applications where maximum bus speeds are not critical

consumption of memory) of our two ORCA configurations, to conclude that, even though a small increase in consumption can be seen with ORCA025, maybe because more intense computations occur with larger sub-domain dimension arrays, the resolution does not affect significantly the per node consumption. Of course, the total energy to solution is different, but can be roughly estimated as the cube of the resolution ratio, assuming that the number of grid points only differs in the horizontal dimensions and that the time step varies in the same proportion than the horizontal resolution (CFL condition). We also noticed that the measurements are close to the TDP (Thermal Dissipation Power) specification provided by the vendor.

To double check the number validity, we used two other tools to measure the average power :

- IPMI (Intelligent Platform Management Interface) `dcmi` (Data Center Manageability Interface) commands, available on our LENOVO server, which captures real-time power with sensors
- the `likwid` software[4] that provides measurements of the current energy consumption through the RAPL (Running Average Power Limit) interface to hardware counters

The `dcmi` command gives the same values as Energy Scope, while `likwid` underestimates the power consumption : for example, the ORCA1 test on a single node shows a total energy consumption of 315W per node for `likwid`, instead of 360W for Energy Scope and the `dcmi` command.

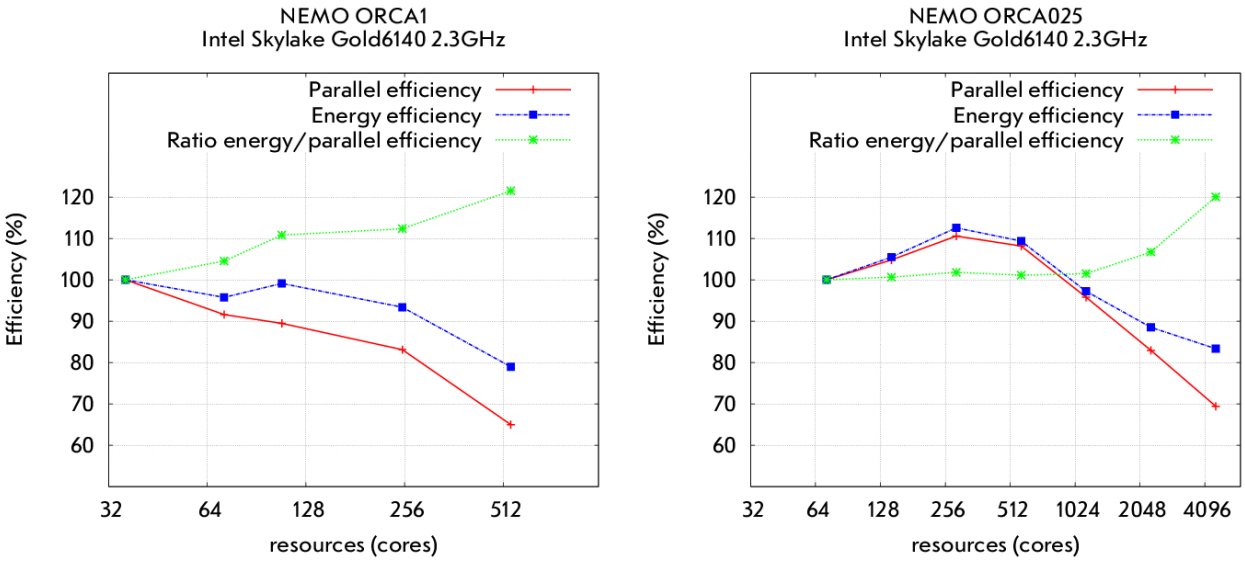

Figure 3: NEMO parallel and energy efficiency, depending on parallelism and model resolution (left: ORCA1, right: ORCA025). The ratio of the two efficiencies is plotted in green.

The figures 3 show how energy consumption can be affected by the parallel decomposition. For the two horizontal resolutions, the model speed and total energy consumption are measured (time loop only) and compared to the measurement with the minimum number of node possible. This number is equal to 1 for ORCA1 and 2 for ORCA025 (the high resolution configuration requires more memory and only fits on two nodes). The blue lines show the parallel efficiency of the model. As expected, an hyper-scaling is observed on the first part of the ORCA025 curb, due to the high memory requirement of each sub-domain there: their size

4 https://github.com/RRZE-HPC/likwid

is equal to 242x203 on one node, and 22x21 on 128 nodes.

What we call energy efficiency is similarly the ratio between the consumption on the minimum number of node and the consumption at a given parallel decomposition. This curb roughly follows the previous one, but shows significantly better efficiency at scale (which can be better visualised on the ratio energy/parallel efficiency in green)[5]. We can assume that the time spend in MPI communications, larger at scale, is the main contributor to this difference, considering that these growing contribution is less energy prone than standard computations. One can notice that the extra consumption in link with a more intensive usage of the interconnect network is not taken into account in our measurement.

We can conclude that the raw values of energy consumption per node is not strongly dependant from the parallel decomposition and mainly varies following the parallel efficiency. The optimum value is reached on the minimum number of node (ORCA1) or at a slightly higher decomposition with configuration showing hyper-scaling (ORCA025).

It is difficult to compare the raw value of "energy to solution" to any other model (which necessarily have ... a totally different solution). The comparison is even impossible with other ocean models such as ROMS [12], since it is complicated to find in literature comparable numbers such as model speed (given in Simulated Years Per Day, SYPD) together with the problem size.

The only possible comparison can be done with the same model on different platforms. For example, a direct consumption measurement of the platform during a recent porting on NEC-SX Aurora TSUBASA revealed that ORCA025 needed 2kW to run at 10 SYPD. At the same speed, the configuration requires about 20 kW on our Intel Skylake machine (or 13 W per core, in the same range than Berthoud et al. 2020). Of course, a larger number of information would be required for a comprehensive comparison, such as PUE or network and disk access consumption. It would also be cleaner to be able to use the same measurement tool on the various machines.

We can conclude from this work that the NEMO model does not differ significantly from other applications regarding energy consumption, and that this consumption is weakly dependant on model resolution or parallel decomposition, conferring on the scalability the quality of single criteria for energy efficiency. More analysis are required to identify possible energy bottleneck in our model.

Acknowledgements: The authors are in debt with Hervé Mathieu (INRIA) for providing the Energy Scope package, for helping them installing and interpreting the results. The authors wish to acknowledge Thomas Williams and Colin Kelley for the development of the Gnuplot program, which analysis and graphics are displayed in this report, in addition to graphics from Matplotlib, a Sponsored Project of NumFOCUS, a 501(c)(3) non profit charity in the United States. This study did not received funding from the European Union's Horizon 2020 research and innovation programme. "I took my power in my hand / And went against the world / 'T was not so much as David had / But I was twice as bold "

5 Notice that speed, not power consumption, is the quantity chosen to calculate efficiencies. It means that what we are comparing here are "energy to solution" (Joule) and not power (Watt). This is why the "energy efficiency" is so strongly linked with "parallel efficiency"

References

[1] Bordage, F., 2019: GreenIT.fr: Environmental footprint of the digital world, https://www.greenit.fr/environmental-footprint-of-the-digital-world/ [accessed 19-07-21]

[2] Berthoud, F., Bzeznik, B., Gibelin, N., Laurens, M., Bonamy, C. et al., 2020: Estimation de l'empreinte carbone d'une heure.coeur de calcul, rapport de recherche, UGA - Université Grenoble Alpes, CNRS, INP Grenoble, INRIA. ffhal-02549565v4f

[3] Rosen, J., 2017: Sustainability: A greener culture, Nature, ***546***, 565–567 https://doi.org/10.1038/nj7659-565a

[4] Heldens, S., Hijma, P., Werkhoven, B. V., Maassen, J., Belloum, A. S., & Van Nieuwpoort, R. V., 2020: The landscape of exascale research: A data-driven literature analysis. ACM Computing Surveys (CSUR), ***53(2)***, 1-43.

[5] Locean-Climactions, et al., 2020: Vers une transition bas carbone au LOCEAN : démarche, mise en place et perspectives, https://www.researchgate.net/publication/348621107_Vers_une_transition_bas_carbone_au_LOCEAN_demarche_mise_en_place_et_perspectives [accessed 26-07-21]

[6] Madec, G. & NEMO System Team, 2019: "NEMO ocean engine", Scientific Notes of Climate Modelling Center (27) – ISSN 1288-1619, Institut Pierre-Simon Laplace (IPSL)

[7] Voldoire et al., 2019. Evaluation of CMIP6 DECK experiments with CNRM-CM6-1, Journal of Advances in Modeling Earth Systems, https://doi.org/10.1029/2019MS001683

[8] Maisonnave, E. & Masson, S., 2019: NEMO 4.0 performance: how to identify and reduce unnecessary communications , Technical Report, TR/CMGC/19/19, CECI, UMR CERFACS/CNRS No5318, France

[9] https://sed-bso.gitlabpages.inria.fr/datacenter/energy_scope.html [accessed 29-07-21]

[10] Ferrero, F. , 2017: Analysis and dynamic optimization of energy consumption on HPC applications based on real-time metrics, Tesi di Laurea Magistrale, Politecnico di Torino

[11] Gourdain, N. et al., 2009: High performance parallel computing of flows in complex geometries : I. methods. Comput. Sci. Disc., 2 :015003

[12] Shchepetkin, A. & McWilliam, J., 2009: Computational kernel algorithms for fine-scale, multi-process, long-term oceanic simulations. In:Handbook of Numerical Analysis, Vol. XIV: Computational Methods for the Ocean and the Atmosphere, P. G. Ciarlet, editor, R. Temam & J. Tribbia, guest eds., Elsevier Science, pp. 121-183