Implementation and computing performance of NEMO ORCA-Km configuration

*E. Maisonnave, C. Lévy *, S. Masson **

* Sorbonne Universités -CNRS-IRD-MNHN,
LOCEAN Laboratory, Paris, France

TR/CMGC/19/18

Abstract

Popularity of one kilometre scale global geophysical models is continuously growing in Europe. In this study, we increase the resolution of the NEMO ocean model, to better identify technical limitations of the present implementation and deduce, from the performance measured on one of our top supercomputers, the hardware characteristics needed to routinely use a global ocean one kilometre configuration. 24,000 Intel Broadwell cores of the Météo-France supercomputer are required to measure the performance of a simplified version of NEMO global kilometre scale. From them, assuming a perfect scaling up to 9,000,000 (nine million) computing cores, we infer that we could target a maximum speed of approximately 2 SYPD with the ORCA-Km global model. This configuration needs, at least, 4 (four) orders of magnitude more CPU power and energy than one of the most demanding coupled model of the community (CNRM-CM6-HR). It is premature to start considering the possibility of any production run based on the global one kilometre scale NEMO model, because no present day machine can fulfil the ORCA-Km memory/bandwidth/computing power requirements. And without a major breakthrough in ocean modelling science paradigms (to strongly reduce the amount of calculations and time to solution required), a simultaneous revolution affecting microprocessor industry (to reduce supercomputer energy consumption) and a code rewriting to fit the unconventional requirements of such energy-efficient chips, it is, from some point of view, not advisable to do so.

Table of Contents

Since the description of the advantages that climate modelling and computing industry could take from such configuration [1], popularity of one-kilometre scale global geophysical models is continuously growing in Europe [2,3]. We propose to increase the resolution of the NEMO ocean model [4], already available for tests at 1/36° resolution, to better identify technical limitations of the present implementation and deduce, from the performance measured on one of our top supercomputers, the hardware characteristics needed to routinely use a global ocean at one-kilometre horizontal resolution.

1- Test case suite configuration

NEMO is a framework that includes several modules (OPA ocean, LIM sea-ice, TOP-PISCES biogeochemistry, XIOS I/O server …). Modules compose configurations adapted to various needs. We propose to focus our study on the framework core: the ocean routines. Due to its pivotal position in the framework, any enhancements of computing performance related to this module will benefit to all NEMO configurations.

Since sea-ice routines can be linked separately from the main NEMO executable [5], and considering that the modifications needed to increase sea-ice or ocean module computing performance strongly differ, we prefer to study separately the analysis of the 1 km resolution increase impact in the ocean and in the sea-ice components. In addition, the ongoing development of a new version of the sea-ice model forbids to lead its study first, and suggests to take benefit of the rewriting to reconsider the algorithmic of the code also for this purpose.

The configuration we chose includes TOP and XIOS. This ensures that the benchmark results will be easily related to full complexity configurations.

Among several spatial discretisations (horizontal grids) available, GYRE appears to be the one that could, at the same time, facilitate the porting on the various platforms and the sharing of a common configuration between the several laboratories involved in our project (ATOS, BSC, CMCC), along with keeping most of the characteristic of the realistic ORCA configurations:

- With its flat bottom and rectangular boundaries, GYRE does not need input files to define its bathymetry and its forcing conditions. The whole initial and external constrains can be defined via a simple FORTRAN namelist, which insures that all partners can investigate the very same problem without risk of input file differences or local code modifications induced by the different nature of the partner platforms.
- Its resolution can be simply changed by namelist. This facilitates the definition of a set of configurations from 1° to 1 Km, which helps to investigate the model weak scaling.
- Actually, the resolution increase is mimicked by a surface increase of the rectangular pool. Parameters of physics and dynamics of the 1° resolution model are the same during the whole experiment. This avoids any numerical instabilities usually linked to a realistic resolution increase and prevents to change the time step length (kept to 3600 s). We emphasise that the result correctness is not an aim of this study, considering the perfect computing performance similarity between a realistic 1km resolution model and our benchmark configuration.

Actually, some characteristics of the ORCA configurations are not present in GYRE's, in particular:

- coast lines and varying depth of ocean bottom
- consequently, continent-only sub-domains resulting from an horizontal MPP decomposition cannot be eliminated
- forcing files reading (an analytical forcing is applied to GYRE surface)
- periodic conditions (East-West and Northern Polar regions)

However, for a given resolution, the total amount of computations and communications is almost the same for GYRE and ORCA. In particular:

- sub-time step value, that modifies the cost of the surface pressure gradient trend calculation routine[1], only depends of the maximum depth (in metre) and not of the bathymetry shape nor vertical level number. For historical reasons (vector machines), calculations are performed similarly on non-masked and masked grid points, where calculations with constant values are not even altered by any compiler optimisation. This also prevents any influence of the bathymetry on the calculations speed.
- even though there is no eliminated sub-domain in the domain decomposition mesh of GYRE, the communication pattern is supposed to be rather similar to ORCA. Boundary sub-domain can be assimilated to sub-domain with no communication on one or two sides.
- forcing files are scarcely read in the ORCA time loop part of the simulation (the one that is mainly investigated here) and, with an efficient I/O system, this should not influence significantly computing performance. We also mention that this problem does not need to be addressed for ocean-atmosphere simulations.
- East-West periodic conditions are not included in the standard GYRE configuration. This could be done but should not strongly change the communication pattern but can be responsible of a lack of scalability. The separated treatment of communications on Northern Polar fold was recently revisited. This supposed positive effect could be questioned at kilometre scale but is excluded from the current study.

The vertical resolution is kept constant (31 levels) during the whole study. An extrapolation to the appropriate vertical resolution (75 levels if horizontal resolution > 1 degree) can be evaluated at current vertical resolution. At the opposite, to strictly estimate the cost of special numerical schemes better suited for high resolution, it is mandatory to change physics parametrisation and dynamics schemes. This will be done (in a second step) by a simple change in namelist.

For this study, we will mainly use GYRE-12 (global 1/12 degrees, 4500x3000 grid points, 31 levels) and GYRE-Km (global 1Km, 36000x24000 grid points, 31 levels).

1 `dynspg_ts`, time splitting loop including halo communications

2- Design of experiment

2.1- Initial design

The NEMO[2] routines were compiled to create a standard GYRE "GYRE_PISCES" executable. Among several platforms available, the research dedicated Météo-France machine[3] is chosen. Consequently, routines are compiled with the INTEL compiler and standard computing optimisations (-02 -xAVX)[4].

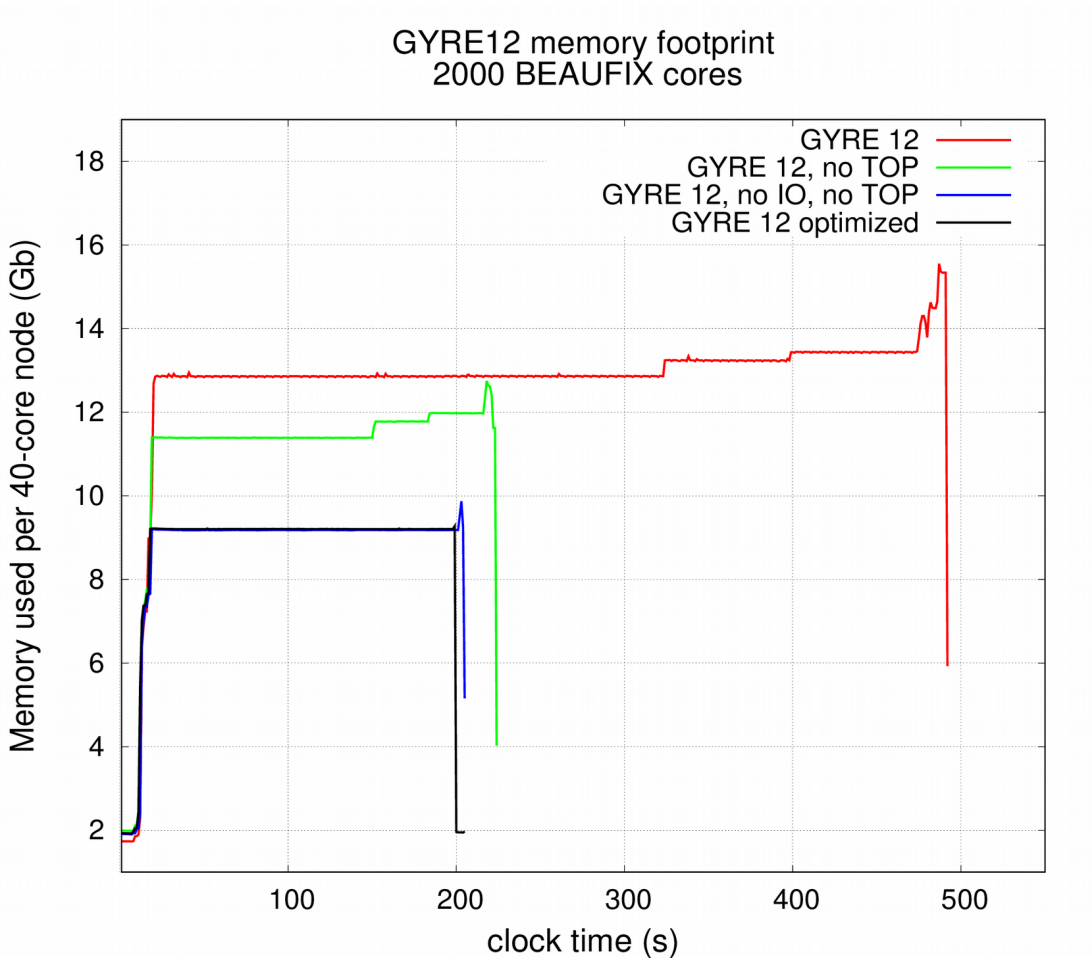

Figure 1: Memory trace of GYRE-12 on the peak memory consuming node of the `beaufix2` Météo-France supercomputer. Ratio of time to solution between GYRE-12 with or without TOP/IO is about 500s/200s

2.2- Operational constraints

2 `dev_merge_2017` branch, revision 9466
3 "`beaufix2`", `https://www.top500.org/system/178962`
4 `-O3` optimisation doesn't change significantly the performance. Intel Fortran compiler version is 16.0.1 20151021.

High end configuration usually requires to reserve a substantial part of the machine resources. The first attempts to perform short simulations with GYRE-Km, even without the XIOS external I/O server, revealed that the memory consumption easily overshot the node capacity (61 Gb), even using the maximum of 300 nodes (12,000 cores) available for standard users. Moreover, since the machine is not able to cope with "out of memory" error on a node (the operating system have to be restarted after such error), the administration strictly forbids to generate such error and an a priori estimation of the necessary amount of memory was required.

A memory requirement analysis of GYRE-12 was performed with the "`collectl`[5]" tool and the results extrapolated to the GYRE-Km configuration. Figure 1 shows the maximum memory use (RSS), as seen by the LINUX OS (red line), during the simulation (x-axis = elapsed time). This first set-up already excludes the memory consuming XIOS server and includes an improvement in the GYRE initialisation phase (the unnecessary allocation of a full global array for bathymetry by the MPI master process).

An extrapolation of this result revealed that more than the half of the machine would have been required to safely conduct a GYRE-Km simulation. Consequently, we had to modify our initial design of experiment to strongly reduce the memory consumption of our benchmark configuration.

2.3- Adapted design

In a first step, the passive tracer advection (TOP) was switched off, resulting a large save in computing time but smaller memory gain (Figure 1, green line). The switch out of the whole I/O system (external server and internal client/diagnostics) was required to reach satisfactory levels (blue line). Eventually, an additional restriction (no restart writing) is applied to avoid the final peak (black curve).

The new GYRE-Km "light" configuration (L-GYRE-Km) was ready to be used on 600 nodes (24,000 cores). Figure 2 shows the memory trace of two simulations with two different time lengths (red and orange lines), in comparison with the previous L-GYRE-12 test (green line). The small extra memory available per node (10 Gb) prevents to test the model on less nodes. Considering the number of nodes reasonably available per user, it was not possible either to test the model on more than 600 nodes. A larger machine, including nodes with larger memory capacity, would be required to investigate L-GYRE-Km scalability.

3- Computing performance

3.1- Restrictions effect

5 `http://collectl.sourceforge.net/`

Starting from the 3 short tests performed on the Météo-France supercomputer, we can deduce several quantities that help to guess the characteristics of the future machine that would be able to host production run leaded with a GYRE or ORCA-Km model (without sea-ice).

Due to the current hardware limits, two main restrictions were applied to our initial test-bed configuration: the removal of TOP routines and the disabling of any output or diagnostics performed by XIOS client or server.

XIOS I/O server high capability is already validated with high-resolution configurations [6]. However, special care must be taken in allocating the right amount of memory to each server. This issue is particularly bothering in our machine with our top end configuration, because memory limit cannot be reached without severe disturbances for all the machine users. Further analysis of an "XIOS-Km" configuration is then required to rightly estimate the total amount of memory (and disk space) requested by this resolution. However, the XIOS server capacity to perform writing on disk at the same time than calculations allows the assumption that the related restriction we made with L-GYRE-12 should not lead to significant slow down, independently of the extra resources needed to include more I/O server in the configuration.

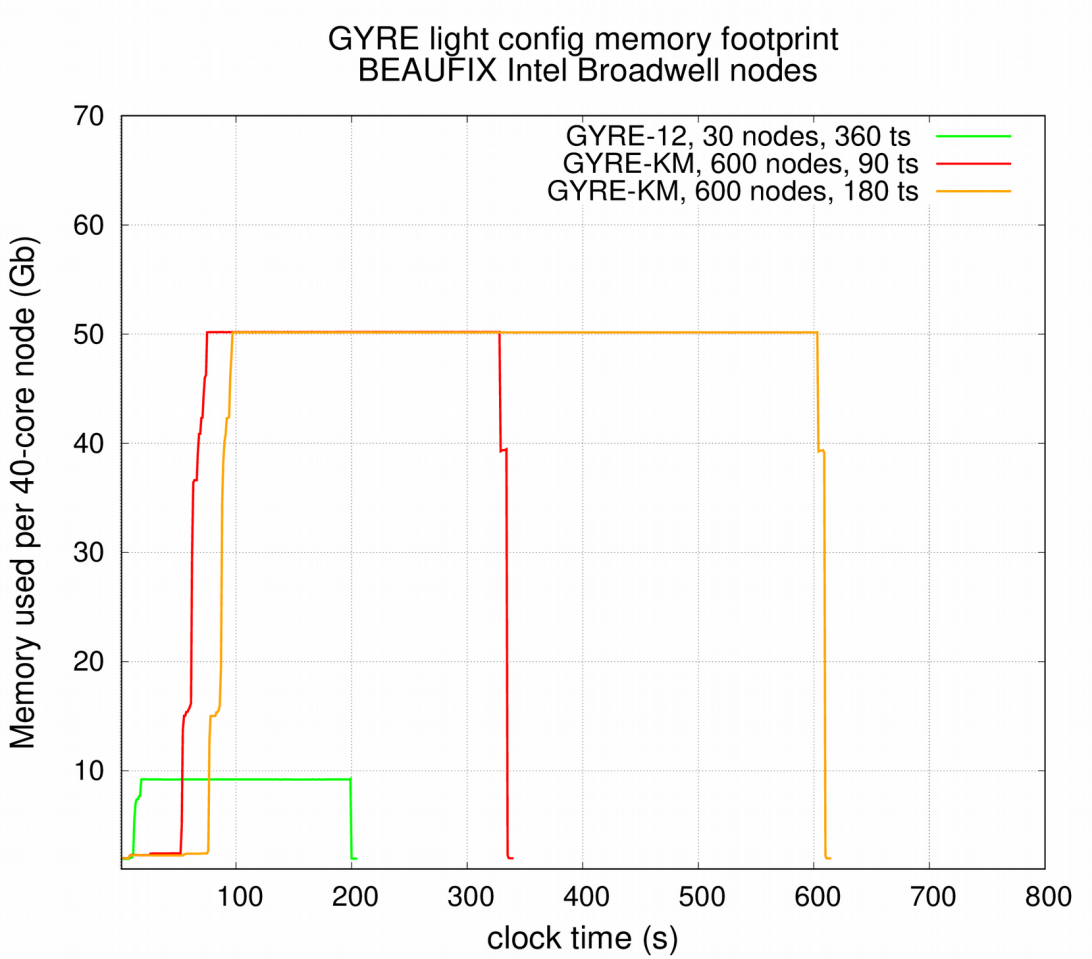

Figure 2: Memory trace of L-GYRE-12 and L-GYRE-Km on the peak memory consuming node of the `beaufix2` Météo-France supercomputer

The effect of the other restrictions (TOP and XIOS client costs) can be easily estimated with GYRE-12 and extrapolated to L-GYRE-Km with a simple ratio of 500/200 in computing time

(see Figure 1, and assuming perfect scalability of XIOS and TOP for both resolution, within the resource range considered in this study).

3.2- Measurements

The total computing time of L-GYRE-Km (initialisation phase and time loop) estimated on 90 and 360 time steps, and memory trace (`collectl` tool) are the only direct measurement that were possible to lead on our machine. Further instrumentations, that are mandatory i.e. to estimate MPI routine cost (with ATOS xPMPI library or INTEL ITAC), were not possible with such high number of computing nodes. To perform 90 timesteps of L-GYRE-Km, 253 seconds were required.

3.3- Extrapolations

Various quantities can be extrapolated from this raw result. We will limit our projections to speed, computing cost and energy consumption. Definitions of these quantities are given in [7].

We define GYRE-Km to be the model actually used for measurement, plus TOP and output diagnostics routines (5/2 times slower, see above), changing time step length to match the kilometre spatial scale (36s instead of 3600s, 100 times slower) and the vertical levels also corresponding to the kilometre spatial scale (75 instead of 31).

To extrapolate GYRE-Km speed from our measurement, we use an even more simplified configuration of NEMO (BENCH). This configuration has the same basic geometry and bathymetry but does not include forcing and is free of any control diagnostics (MPI collective communications). Experiments are leaded with the same ORCA12-related namelists, modified to increase the number of vertical levels (from 75 to 31) and to set ORCA1 physics/dynamics. The differences of performance is a pure effect of these two changes (time step is the same for the 3 experiments). A MPI decomposition of 2520 is chosen to keep our measurement to near perfect parallel efficiency conditions, that we are supposed to found in all our study. From results shown in Table 1 and previous extrapolations, we guess that GYRE-Km will be 2.46 x 100 x 500/200 slower than L-GYRE-Km.

Namelist	Standard	k=31	ORCA1
Speed (SYPD)	0.52	1.28	0.625
speedup		*2.46*	*1.20*

Table 1: Compared speed and speedup (italic) of BENCH-12 model (2520 MPI subdomains) based experiment, with standard namelist, namelist with vertical level number equal to 31 instead of 75, and with physics/dynamics of ORCA-1 instead of ORCA-12 (but same time step)

We call ORCA-Km the configuration GYRE-Km with realistic parametrisation at that scale (see Appendix 1 to see parameter changes between ORCA-1 and ORCA-2 namelists). Relying on

the comparison made with BENCH and shown in Figure 3, speed must be divided by a 1.20 ratio.

We call "scalable" ORCA-Km (S-ORCA-Km) the configuration ORCA-Km that would have a perfect scaling up to 10x10 subdomain size. Considering that our measurement is done on 24.000 cores with a 100x380 subdomain size, speed should be multiplied by 380 (and the core number multiplied by the same number is equal to 9,120,000). It must be stated here that the scalability limit of GYRE-Km cannot be measured with our supercomputers. It can be guessed from the scalability limits of BENCH-1, BENCH-025 (reached with `irene`[6] supercomputer [7]). It is a strong assumption that future architectures bandwidth and MPI library functionalities will allow such performance with ORCA-Km.

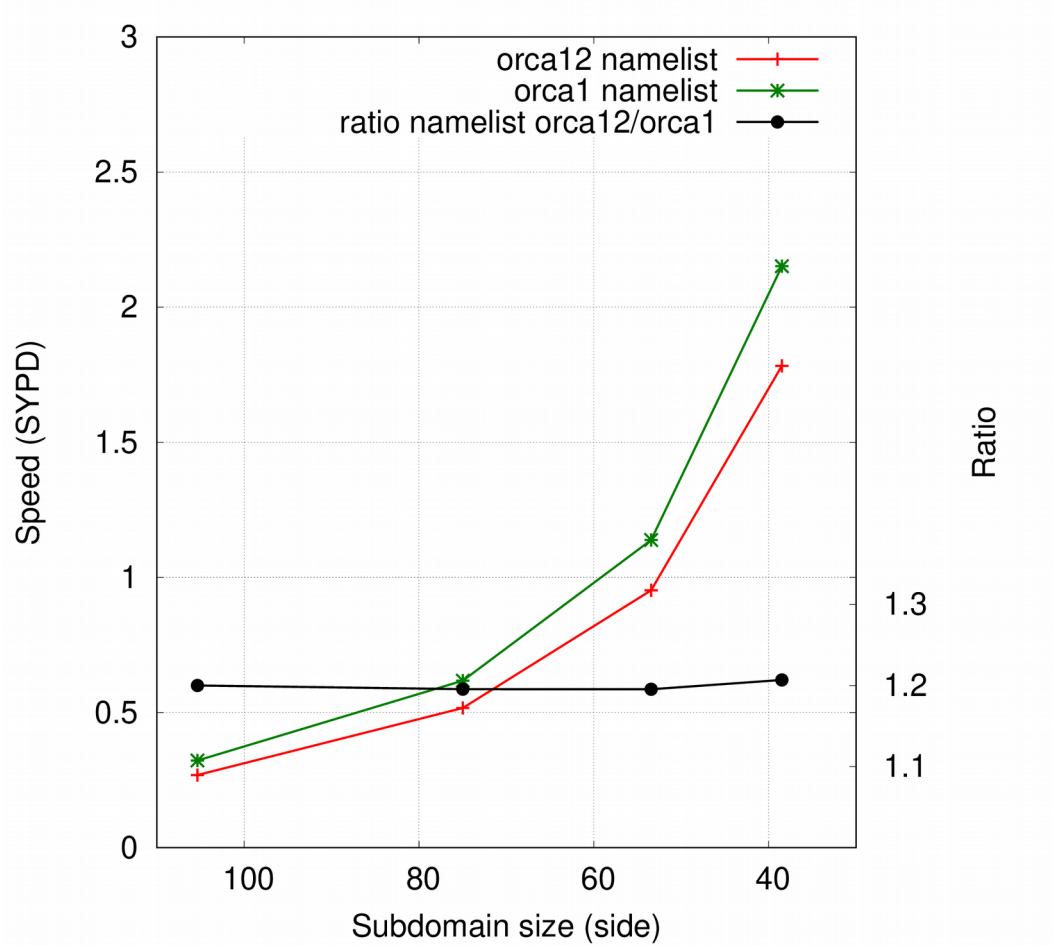

Figure 3: BENCH-12 scalability with ORCA-12 and ORCA-1 physics, ratio ORCA12/ORCA1, measurement on `beaufix2` Météo-France machine

As shown in Table 2, the estimated speed of S-ORCA-Km (more than one simulated year per day), without sea-ice or output, and thanks to extremely favourable assumptions, is acceptable for some short term climate/oceanography studies (seasonal to decadal forecasts). This is true only if we assume the perfect scaling of the model up to 10x10 subdomains and the full availability of a machine (130 times bigger than `beaufix2`) that could provide the required amount of resources. For that reason, under current computing technology conditions, the NEMO model (3.6 version) configured with a set of parametrisations suitable for high resolution, is not able to routinely deliver results globally at kilometre scale. The speed of ORCA-Km, .i.e. the maximum speed that the current ORCA code can reach using one of the

6 https://www.top500.org/system/179411

most powerful supercomputer at the moment, is below any acceptable limit (less than 1 simulated year *per year*).

Configuration	L-GYRE-Km	*GYRE-Km*	*ORCA-Km*	*S-ORCA-Km*
Speed (SYPD)	3.51	*0.00561*	*0.00469*	*1.78*

Table 2: Speed in simulated years per day (SYPD) measured ("light" GYRE-Km) and estimated (all others, italic)

It is already possible to predict the computing cost on the hypothetical Exascale machine that could host the S-ORCA-Km model. It is estimated to about 123 million core hours per simulated year (CHSY), 4 (four) orders of magnitude more expensive than the most expensive Météo-France/CERFACS coupled model used during CMIP6 (~ 21,000 CHSY).

Due to the lack of a direct measurement of the energy consumption of the nodes processing the simulation, we can only estimate this quantity, considering the formula given in [8]. We deduce from the computing cost, the energy consumption (excluding computing centre cooling) of the entire `beaufix2` machine $E = 2.15^{12}$ J per month and the aggregate compute hours $A = 5.2^{7}$ core hours per month on this machine:

$$EC = C \times E / A = 1{,}390 \text{ MWh/SY}$$

During one year, the future supercomputer would be able to simulate 650 years. We compare the energy consumed by this simulation during one year, and the energy produced by two European power plants during the same period in Table 3.

	S-ORCA-Km	Fessenheim (F)	Jänschwalde (D)
Energy (TWh)	0.9	8.4[7]	22.0[8]
ratio		11 %	4 %

Table 3: compared yearly energy consumption and production of a global kilometre scale ocean model and two nuclear and coal power stations

Another quantity can be deduced from the total yearly CO_2 emissions of the Jänsschwalde power plant (24.1 Mt): the corresponding CO_2 emission of the 650 year long simulation with S-ORCA-Km (1.0 Mt), equal to the total emissions of the Réunion Island during the same period [9]. Ten experiments of decadal predictions (without sea-ice nor atmosphere model), including each ten starting dates of ten member ensembles, would emit more CO_2 in the atmosphere than the whole Republic of Slovenia in one year.

The future computing scientists of our community would have to face 3 simultaneous challenges to deliver an ORCA-Km model capable to be used routinely (to say nothing about the difficulty of training and attracting those people):

1. ensure a perfect scaling on more than nine million cores (S-ORCA-Km),

7 Data source: EDF yearly communication, 2016

8 Data source: Umweltbundesamt (UBA)

2. improve the speed of the sequential configuration by one order of magnitude to reach the minimum speed required for regular studies and maybe limit the amount of extra calculations / memory accesses of the successive future NEMO versions,
3. make NEMO code compatible with very energy efficient processors (if any) to prevent climate modelling contributing significantly … to climate global change.

As expected [10], an incremental strategy to adapt NEMO to such extreme scales is not suitable. It is premature to start considering the possibility of any production run based on the global one kilometre scale NEMO model, because no present day machine can fulfil the ORCA-Km memory/bandwidth/computing power requirements. And without a major breakthrough in ocean modelling science paradigms (to strongly reduce the amount of calculations and time to solution required), a simultaneous revolution affecting microprocessor industry (to reduce supercomputer energy consumption) and a code rewriting to fit the unconventional requirements of such energy-efficient chips, it is, from some point of view [11], not advisable to do so.

Figure 4: during the exposure time of this picture (1/500s), the Jänschwalde power plant could not produce enough energy to perform 1 time step of S-ORCA-Km but could contribute to the earth global warming emitting 1.6 Kg of CO_2

The authors strongly acknowledge Isabelle d'Ast (CERFACS) for her advices related to Vtune-ITAC, Michel Pottier (Météo-France) for organizing the high-end experiments on `beaufix2`, Cédric Trivino (ATOS) for his help with BULL tools, Romain Bourdallé-Badie (MERCATOR-Océan) for providing the ORCA-12 namelist, Sophie Valcke (CERFACS) for CNRM-CM6-HR computing performance measurements and Silvia Mocavero (CMCC)/Uwe Fladrich (SMHI) for their careful review. This work is part of the ESiWACE project which received funding from the European Union's Horizon 2020 research and innovation program under grant agreement n° 675191.

References

[1] Palmer, T., 2014: Climate forecasting: Build high-resolution global climate models, Nature, 515, 338–339, https://doi.org/10.1038/515338a

[2] Fuhrer, O., Chadha, T., Hoefler, T., Kwasniewski, G., Lapillonne, X., Leutwyler, D., Lüthi, D., Osuna, C., Schär, C., Schulthess, T. C., and Vogt, H., 2018 : Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4888 GPUs with COSMO 5.0, Geosci. Model Dev., 11, 1665-1681, https://doi.org/10.5194/gmd-11-1665-2018

[3] Neumann, P., Biercamp, J., Fast, I., Bauer, P., Brueck, M., Mauritsen, T., Klocke, D., 2018 : Implementation of ICON global 1km atmosphere-only demonstrator and performance analysis (D2.9). Zenodo.

[4] Madec, G., 2018 : NEMO ocean engine,Note du Pôle de modélisation, Institut Pierre-Simon Laplace (IPSL), France, No 27, ISSN No 1288-1619

[5] Maisonnave, E. and Masson, S., 2015: Ocean/sea-ice macro task parallelism in NEMO, Technical Report, TR/CMGC/15/54, SUC au CERFACS, URA CERFACS/CNRS No1875, France

[6] Masson, S., Hourdin, C., Benshila, R., Maisonnave, E., Meurdesoif, Y., Mazauric, C., Samson, G., Colas, F., Madec, G., Bourdallé-Badie, R., Valcke, S., Coquart, L., 2012 : Tropical Channel NEMO-OASIS-WRF Coupled simulations at very high resolution , 11.4. 13th WRF Users' Workshop – 25-29 June 2012, Boulder, CO

[7] Maisonnave, E. and Masson, S., 2019: NEMO 4.0 performance : how to identify and reduce unnecessary communications, Technical Report, to be published, SUC au CERFACS, URA CERFACS/CNRS No1875, France

[8] Balaji, V., Maisonnave, E., Zadeh, N., Lawrence, B. N., Biercamp, J., Fladrich, U., Aloisio, G., Benson, R., Caubel, A., Durachta, J., Foujols, M.-A., Lister, G., Mocavero, S., Underwood, S., and Wright, G., 2017: CPMIP: measurements of real computational performance of Earth system models in CMIP6, Geosci. Model Dev., 10, 19–34, https://doi.org/10.5194/gmd-10-19-2017

[9] Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Olivier, J.G.J., Peters, J.A.H.W., Schure, K.M., 2017: Fossil CO2 and GHG emissions of all world countries, EUR 28766 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-79-73207-2, doi:10.2760/709792, JRC107877

[10] Lawrence, B. N., Rezny, M., Budich, R., Bauer, P., Behrens, J., Carter, M., Deconinck, W., Ford, R., Maynard, C., Mullerworth, S., Osuna, C., Porter, A., Serradell, K., Valcke, S., Wedi, N., and Wilson, S., 2018 : Crossing the chasm: how to develop weather and climate models for next generation computers?, Geosci. Model Dev., 11, 1799-1821, https://doi.org/10.5194/gmd-11-1799-2018

[11] IPCC: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp.

Appendix 1: Differences of BENCH-1/BENCH-12 physics (namelists)

BENCH 1	BENCH 12
Advection scheme order nn_fct_h = 2, nn_fct_v = 2	nn_fct_h = 4, nn_fct_v = 4
Lateral diffusion ln_traldf_lap = .true.	ln_traldf_blp = .true., ln_traldf_msc = .true.
Mixed Layer Eddy (MLE) parameterisation ln_mle = .true.	ln_mle = .false.
Eddy induced velocity parameterization ln_ldfeiv = .true., nn_aei_ijk_t = 20	ln_ldfeiv = .false.
Momentum advection ln_dynadv_vec = .true., nn_dynkeg = 1	ln_dynadv_ubs = .true.
Vorticity/ Coriolis nn_een_e3f = 1	nn_een_e3f = 0
ln_dynldf_lap = .true., ln_dynldf_hor = .true. , nn_ahm_ijk_t = 30	ln_dynldf_OFF = .true.