Frugal computations with NEMO latest versions on a NEC vector platform

Eric Maisonnave
CECI, UMR CERFACS/CNRS No5318, France

WN/CMGC/23/17

Abstract

This note complements three other preliminary works, in which the performance of the NEMO ocean-only model was measured on NEC SX-Aurora TSUBASA (20B) and favorably compared with similar results on contemporary scalar platforms. In this document, we resume the study on similar vector engines, recently deployed on a higher number of nodes at the French CRIANN supercomputing mesocenter. The version upgrade (trunk, including RK3 time stepping) of our ocean model code, while conserving its vector properties, was easy. A direct energy consumption measurement reveals, for the most favourable but popular NEMO configuration (ORCA025), the x8 advantage of the vector platform compared to its Intel SkyLake competitor (x6 to IceLake). Beside the never ending race to electricity overconsumption and developers labour waste on esoteric hardware, this study shows that there is room for energy efficient and scientifically justified ocean simulations when performed on the suitable equipment

Table of Contents

This working note follows three other reports related to the NEMO [1] ocean model performance on NEC SX-Aurora TSUBASA platforms: an initial vector potential evaluation of the ocean core engine [2], an extension to the sea-ice and bio-geo-chemistry modules [3] and a full use, in coupled mode, of the associated scalar host resources for I/O [4].

1- Code porting and upgrade

1.1- Porting

For this study, we used the nine computing nodes of the vector partition of `boreale`, owned by the "Centre Régional Informatique et d'Applications Numériques de Normandie" - CRIANN (Rouen, France) [1]. To facilitate the porting, the NEMO 4.2 version is used in a first step. Its include the modifications regularly prescribed to compile the ocean code on NEC SX-Aurora TSUBASA, particularly those made by Janna Abalichin (BSH) and Jens-Olaf Beismann (NEC Germany) to use the model on the DWD (German Weather Service) supercomputer. Since both CRIANN and DWD hardware and software are slightly the same, the setup (compiling and running) was done without any further modification.

As usual for performance measurement, the realistic BENCH [5,6] configuration (ocean only) is preferred, because its simplifies the input file setup and speeds up the experiment (simplified choice of model parameters, short duration runs). The level of 99% of the time spent on vector sectors is immediately reached without modification. We notice that the average vector length (AVL), which measures the capacity of our program to benefit from the full vector register length (256) is closer to this limit with the BENCH025 global ¼ degree grid (240) than with the BENCH1 global 1 degree grid (170). For that reason, we decide to focus on this BENCH025 intermediate horizontal resolution. As previously, and to be able to keep this level of AVL, we also need to precise by namelist the 2D horizontal decomposition of the MPI sub-domain (practically a 1D decomposition, since the longitude dimension decomposition X is set to 1). This disables the automatic computation of the best decomposition, which has several consequences, from which:

- the impossibility, at this resolution, to remove land-only processors,
- the difficulty to minimise communication and load imbalance.

In particular, the fixed Y decomposition forces the algorithm to give a rather small Y dimension to the last sub-domain, which introduces a large load imbalance and downgrades the scalability on more than 4 vector nodes. For that reason, on 8 nodes, a 2x256 decomposition was preferred to the 1x512.

1 *boreale* is the recent CRIANN vector system with 9 NEC SX-Aurora TSUBASA nodes ("Vector Hosts" in NEC terminology). Each node includes two Intel Xeon Ice Lake 6326 bi sockets of 16 cores each (host) and 8 vector engines VE (20B) of 8 cores each

A performance comparison can be seen on Figure 1. The same code is tested on a bi-socket Intel Xeon Gold 6140 (SkyLake) with 18x2 cores (2.3 Ghz).

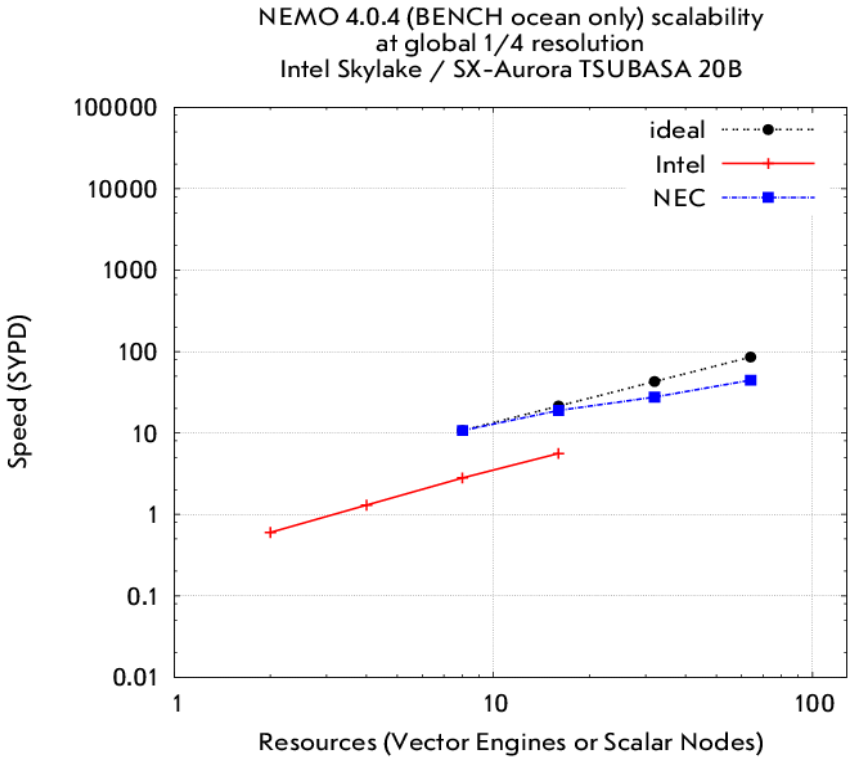

Figure 1: Comparison scalar/vector performance of NEMO v4.0.4 (1 NEC vector engine of 8 cores -8 MPI subdomains- is compared with 1 Intel node of 2 sockets of 18 cores -36 MPI subdomains-)

A rough x4 gain is noticed at low parallelism. This gain is certainly decreasing at scale, possibly nullified at node (o) 100. Whether the MPI library speed or the NEMO communication routines (`lbc_lnk`) vectorisation is the bottleneck remains unclear, despite our effort to vectorise a maximum of instructions in these routines. It seems cleat that NEMO exhibits a better scalability on scalar nodes (where smaller sub-domains take a better benefit of the fast cache accesses) than on vector accelerators. This result is coherent with recent NEMO porting on other devices like GPU.

In addition, better scalar performance of the same code is noticed on more recent hardware (Ice Lake, x1.4 compared to Sky Lake) which reduces the vector advantage, at low parallelism, to x3.

1.2- Main branch

The up to date distribution of the NEMO code (*gitlab* forge [2]) is installed on the vector machine. During our experiment (December 2022), the main branch downloaded included the

2 https://forge.nemo-ocean.eu/nemo

4.2 modifications, and 6 months of additional works. A simple merge of the existing vector related modifications is applied to the code. On Figure 2, one could notice the small additional cost, probably due to an increase of the model computations, given that vector performances are unchanged compared to the 4.0.4.

This result suggests that, at least for the ocean only BENCH configuration, vectorisation is broadly conserved during the current NEMO developments. In particular, the impact of the modifications related to the MPI sub-domain halo extension seems to be minimal.

1.3 – New RK3 time-stepping scheme

This result is consolidated when the new third order Runge-Kutta (RK3) time stepping scheme [7] is activated in our BENCH configuration. This option slightly changes the actual routine tree of our program, which necessarily needs new vectorisation adjustments on two new routines[3].

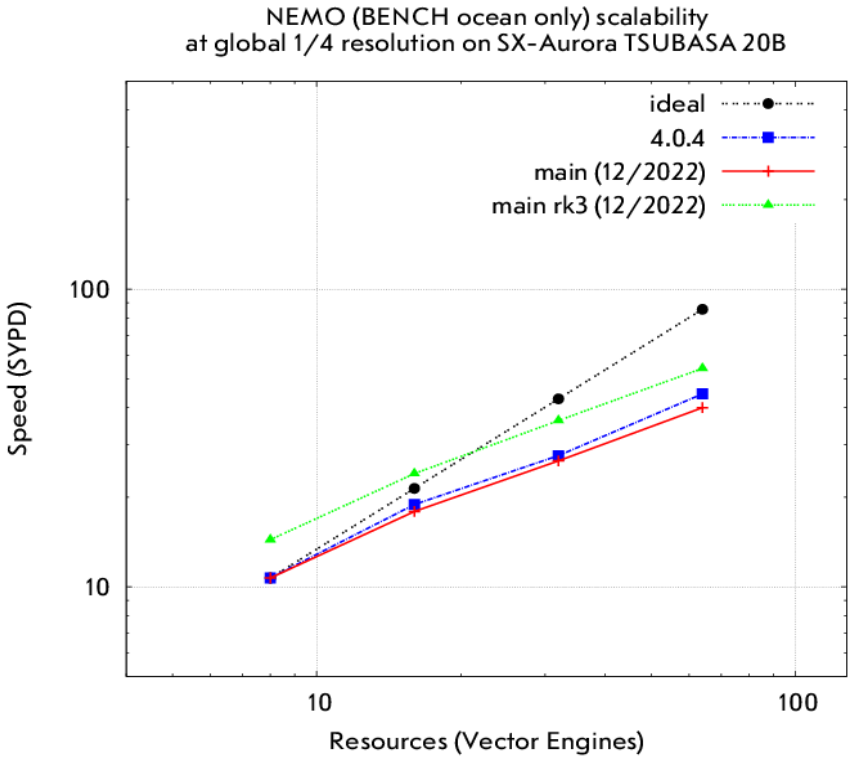

Figure 2: Scalability comparison of NEMO versions

A performance enhancement is observed on Figure 2. As expected, RK3 required more computations and communications, but allows the division by two of our time step (taken into account on Fig. 2). However, a deepening of the previous drawbacks also pops up with this version. Using more than 16 vector engines, the load imbalance increases and a new global communication routine significantly contributes to limit the scaling. An improvement is observed by increasing the Y dimension decomposition from 1 to 2 (taken into account on Fig.

3 `stprk3_stg.F90` and `stp2d.F90`, in addition to small bug correction in `isfstp.F90`

2), but vectorisation decreases for some routines, e.g. during surface pressure gradient computations.

However, it is possible to conclude that the code vectorisation is kept at a good level despite significant code upgrade such as the implementation of the RK3 time stepping.

2. Energy consumption

The energy consumption of the 4.0.1 version of the code was previously measured with the specific Energy Scope[4] toolkit [8]. It appears that the energy efficiency is mainly related to the parallel efficiency, assuming that scalability is its main driver. At its best possible efficiency (8 nodes), the BENCH025 configuration on height Intel Sky Lake nodes was dissipating 3,400W during the execution of the core model time loop (excluding initialisation and finalisation phases).

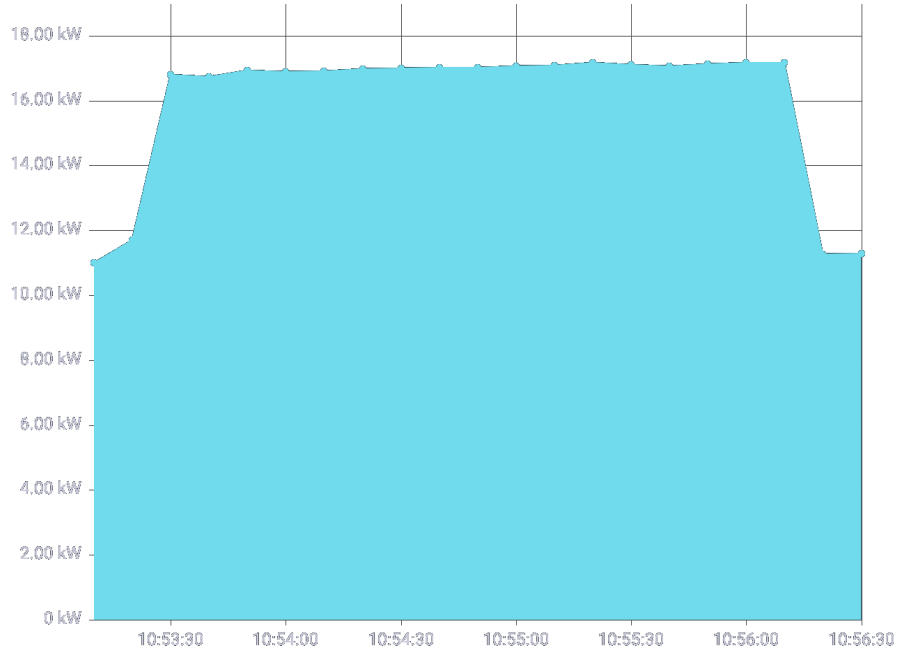

Figure 3: `boreale` total power consumption (kW) during a BENCH025 simulation, measured from its Power Distribution Unit

On the whole `boreale` machine (nine vector nodes), the total consumption was measured by its administrator directly from the Power Distribution Unit associated to the machine racks (via Simple Network Management Protocol), with a period of ten seconds. A trace of this measurement is presented in Figure 3. Starting from the value of 11 kW (idle processors), the consumption rapidly reaches 16.5 kW. As previously observed on other platforms, the machine heated by the running computations seems to slowly increases its consumption, to plateau at 17 kW after a few minutes. Extrapolated on one vector node (eight vector engines), the consumption (including the power needed to feed the idle structure) would be equal to 1,900

4 `https://sed-bso.gitlabpages.inria.fr/datacenter/energy_scope.html`

W, about the half of what eight Intel SkyLake nodes require to run the same configuration at about ¼ of the vector machine speed (ratio 1/8 for energy consumption).

3. Perspectives

From the three NEMO studies we led from 2020 onward, it seems possible to conclude that:

- Compared to the huge and short-lived instrumentation required on GPU[5], a light (for ocean and sea-ice module) or simple (for bio-geo-chemistry module) code update is necessary to get a decent vectorisation and good performance on the NEC SX-Aurora TSUBASA vector devices.
- The medium ¼ degree resolution better fits vectorisation requirements, which also forbids land-only sub-domain removal and limits the multi-device scalability. These combined effects make that the optimal use of NEMO on NEC vector devices is reached for a medium resolution configuration on a small number of vector cards.
- Under this experimental setup[6], NEMO simulations are significantly faster and outstandingly more energy efficient than on its scalar competitors.

These considerations should encourage the NEMO community to maintain a vector compatible code. On that purpose, an additional work is required to instrument both SI3 sea-ice and TOP-PISCES bio-geo-chemistry modules, where vectorisation is not inherited from legacy routines.

Despite these limitations, we keep thinking that the current NEC solution is one of our best current choice for quick NEMO computations, particularly if the configuration resolution allows the use of small size platforms, where interconnection speed or array vector lengths do not limit the high processor throughput. The Japanese vendor roadmap seems to ensure the availability of a vector solution for at least several years [7]. In addition, the underlying hybrid technology (host+device) gives further possibilities of an efficient use of the whole card, if it can hosts modular systems like coupled models [8].

At the age of new energy restrictions, particularly in Europe, isn't it the right moment to drop out the race to computing power, less and less uncertain and more and more time consuming for developers, and to adapt the community equipment to the community needs, and only to the community needs ? At least, we encourage our colleagues to take benefit of the current vector machine availability wherever it is already possible.

5 See `PSyclone` optimisation of the NEMO code : `https://github.com/stfc/PSyclone`

6 This setup is not particularly exotic : ORCA025 is broadly adopted in the community, e.g. studies of climate change, seasonal forecasts, long term ocean variability, etc

7 The newly released C401-8 processor is supposed to bring "2.5 times the computing performance and twice the power efficiency of previous models", Tokyo, October 7, 2022 - NEC Corporation

8 See in [2,4] how an IO server can take benefit of the CPU host while the efficient NEMO vector computations are performed on the vector device

Acknowledgments

Thanks to Laurent Gatineau (NEC HPC Europe), Marie-Sophie Cabot, Sébastien Vigneron, Patrick Bousquet-Melou (CRIANN), Isabelle d'Ast & Nicolas Monnier (CERFACS) for having made possible the access to and the use of the CRIANN supercomputer, to Jens-Olaf Beismann (NEC-Germany), Janna Abalichin (BSH) & Vera Maurer (DWD) for having provided the NEC updated version of the NEMO code. This work was granted access to the HPC resources of CRIANN super-computing center during the acceptance testing period of the NEC SX-Aurora TSUBASA machine (MesoNET national project). The author wishes to acknowledge Thomas Williams and Colin Kelley for the development of the Gnuplot program, which analysis and graphics are displayed in this report.

References

[1] Madec, G. & NEMO System Team, 2019: "NEMO ocean engine", Scientific Notes of Climate Modelling Center (27) – ISSN 1288-1619, Institut Pierre-Simon Laplace (IPSL)
[2] Maisonnave, E., 2021: NEMO performance optimisation on NEC SX-Aurora TSUBASA , Working Note, **WN/CMGC/21/37**, CECI, UMR CERFACS/CNRS No5318, France
[3] Maisonnave, E., 2022: NEMO performance on pre-Exascale processors : NEC SX-Aurora TSUBASA and Fujitsu PRIMEHPC FX700 , Working Note, **WN/CMGC/22/22**, CECI, UMR CERFACS/CNRS No5318, France
[4] Maisonnave, E., 2022: OASIS Dedicated Support, 6th annual summary , Technical Report, **TR/CMGC/22/139**, CECI, UMR CERFACS/CNRS No5318, France
[5] 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
[6] Irrmann, G., Masson, S., Maisonnave, E., Guibert, D., & Raffin, E., 2022: Improving ocean modeling software NEMO 4.0 benchmarking and communication efficiency , *Geosci. Model Dev.* , **15**, 1567–1582, doi:10.5194/gmd-15-1567-2022
[7] Téchené, S., Madec, G., Chanut, J., Coward, A., & Storkey, D., 2022: Gain of efficiency with a new time scheme in NEMO : Runge Kutta 3rd order, EGU General Assembly 2022, Vienna, Austria, 23–27 May 2022, EGU22-13426, https://doi.org/10.5194/egusphere-egu22-13426
[8] Maisonnave, E., & d'Ast, I., 2021: Energy consumption of the NEMO ocean model measured with the Energy Scope tool , Working Note, **WN/CMGC/21/88**, CECI, UMR CERFACS/CNRS No5318, France