NEMO performance on pre-Exascale processors : NEC SX-Aurora TSUBASA and Fujitsu PRIMEHPC FX700

Eric Maisonnave
CECI, UMR CERFACS/CNRS No5318, France

WN/CMGC/22/22

Abstract

This note complements a preliminary work , in which the performance of the NEMO ocean-only model was measured on NEC SX-Aurora TSUBASA (20B) and favorably compared to the current scalar platforms. In this document, we resume the study on the same platform and extend it to the sea-ice (SI3) and bio-geo-chemistry (TOP-PISCES) modules of the framework. In a second step, NEMO is also ported and tested on Fujitsu PRIMEHPC FX700, a scalar machine with SIMD vector capability, built from cell phone processors (arm). The NEMO porting on these two new platforms appeared to be easy, but getting optimal performance required more efforts. Interesting performance at low development cost seems to be more affordable on the NEC vector machine. Our result does not allow to determine the level of efficiency we could reach on Fujitsu. The hardware dependency of the comprehensive rewriting needed is another concern on this platform. Beside the never ending race to electricity wasting, we think that there is room for energy efficient and scientifically justified ocean simulations that must be performed on the right equipment.

Table of Contents

In our previous study [1], good vector performance of the NEMO ocean-only model were achieved with a light porting and optimisation effort. The ocean model [2], initially built in the 90's for vector hardware, was still efficient despite 30 years of modifications. The main limitation found, but less sensible at high resolution, was the vector length bounding to the inner loop length (i.e. number of grid points in longitude) that reduces the possibility of 2D spatial decomposition of the grid in MPI subdomains, which leads to the increase of useless land-only sub-domains. At the opposite, with a limited amount of resources, the better efficiency of our code on the vector platform was obvious.

We propose here to extend this first result by porting and check the vectorisation of the two other modules of the NEMO framework (sea-ice and bio-geo-chemistry). These two modules, more recently added, may not follow the vector compliant coding rules that drove the ocean part building.

In this preliminary work, we also explored the possibility to exploit the full computing node (vector engine for computing + x86 host for disk access) by using the host for disk writing, reserving the vector engine part for what it is made for (computations). This time, we propose to use the official XIOS output server to be able to deliver at the end a standard version of the NEMO code.

Finally, to have a larger picture of what the new technologies are currently proposing, the performance of our ocean model is also tested on a scalar A64FX arm v8.2-A processor. This processor equips the top 500 leader Fugaku Fujitsu PRIMEHPC supercomputer[1] and seems to allow geophysics code to be used with reasonably good performance [3].

Again, we repeat that this study is not a comprehensive machine benchmarking with NEMO, but a simple attempt to evaluate how the vector potential of our code has evolved.

1- NEC SX-Aurora TSUBASA

1.1 - SI3 and TOP-PISCES performance

For this study, we used one computing node (8 Vector engines -VE-, 64 vector cores) of the vector partition of machine (`nesh`) owned by the Christian Albrecht University at Kiel[2]. The same NEMO 4.0.4 version that was previously used for ocean only testing is taken as reference to start our study. Its include the modifications already implemented to compile the ocean code on NEC SX-Aurora TSUBASA. This allows to setup (compiling and running) two ocean/sea-ice and ocean/bio-geo-chemistry configurations without any further modification. Our ocean code also includes the enhancements necessary to be vectorised at

1 `https://www.top500.org/system/179807/`

2 *`nesh`* is an hybrid computing system, based on a scalar NEC HPC Linux Cluster including GPUs and a NEC SX-Aurora TSUBASA vector system (8 NEC SX vector nodes, 10B first generation technology)

its best. The realistic BENCH-1 [4] configuration is preferred, to simplify the input file setup and to avoid any input/output perturbation during the measurements.

We first measure the vector performance of the ocean only[3] on one node of our machine. Its speed is equal to 380 SYPD. The whole optimised code spends 99% of its time in vector sectors and shows an average vector length (AVL) of 175. It approximately fits the number obtained by B. Lodej in our previous study (315 SYPD, 95% and AVL=177). To get these numbers, the code is compiled as previously with a simple `-O2` optimisation. At the opposite of our first study, because this option leads to out of bound errors at runtime on the CAU machine, the `key_vectopt_loop` pre-processing option is switched off in our configuration. For other details about measurement strategy and details see [1].

The out-of-the-box results are poor, which makes mandatory the checking of the vectorisation, as already done for the ocean code. NEMO is compiled and linked with the `-report-all` option. On diagnostics files, vectorisation is detailed for each line of the code and can be optimised through directives. Notice that the total amount of lines that we checked in the code is limited by the `namelist` parameters chosen (the parameters of the official release). A more comprehensive work would be mandatory to ensure the NEC vector compatibility in all configurations.

For sea-ice, three problems are preventing the full vectorisation of the code:

1. The recurrent use of the `SIGN` FORTRAN instruction. In several time consuming routines (`dyn_rhg_evp`, `dyn_adv_pra`, `limthd`, `icethd_dh`), the instruction is replaced by the corresponding `IF` condition,
2. The partial definition of an array (`IF` without `ELSE`) in `dyn_rhg_evp`,
3. The dependency, in a computation of a `n` index variable, to the `n+i` index variable (`icethd_zdf_bl99`). This remains unsolved (and takes approximately 14% of the whole simulation) but the algorithm seems to be also problematic on scalar machine. A general rewriting is expected before proposing a fully vector suitable solution.

In total, we increase the AVL (ocean+sea-ice) from 155 to 160, the time spent on vector sectors from 88 to 97% and the simulation speed from 59 to 161 SYPD.

In TOP-PISCES, in addition to the `SIGN` issue already mentioned (`traadv_mus`, `trc_sink2`), we found inner loops not done on the longitude index but rather on tracers (`p4zpoc`) or vertical levels (`trc_sink2`). On `p4zpoc`, it was possible to re-organise the different loops and ensure long enough vector length by using `ji` index in the inner loop. The same rewriting was impossible for `trc_sink2`, which stays for that reason at 96% of vectorisation and AVL=82. We found that various less time consuming routines would benefit for an index re-ordering. Again, this problem also needs to be addressed for routines not called in our parametrisation. This index re-ordering issue seems more problematic in the BGC code than in the sea-ice part.

3 ORCA1 global grid, defined on 362x332x75 points

In total, we increase the AVL (ocean+BGC) from 81 to 170, the time spent on vector sectors from 77 to 96% and the simulation speed from 4 to 45 SYPD.

Considering these gains, for both sea-ice and BGC parts, it seems clear that the vectorisation of much of the NEMO routines is mandatory to take the most of the machine. This implies that (i) the whole NEMO code has to be tested with all parametrisations before being able to certify the full model compliance with NEC SX-Aurora TSUBASA machine and (ii) a re-ordering of loop indexes must be done to increase the AVL of some routines (particularly in TOP-PISCES) that are currently bottlenecks for performance.

1.2 – XIOS porting

Despite our efforts, it was not possible to include XIOS in our coupled configuration. The main reason of this limitation was the unavailability of the external libraries (with appropriate version) required by XIOS (`boost`, `netCDF`, `hdf5`) on both VE and VH processors. Without support of the CAU help desk, it was impossible to install the full set of required libraries. This strongly emphasises the lack of portability of the XIOS IO suite and prevents further attempts to install the full ECHAM or OpenIFS / NEMO coupled model on the Kiel vector machine. However, a new try will be given at DWD, where more libraries seems to be available.

2. Fujitsu PRIMEHPC FX700

A short allocation is provided by a French computing center for public research (TGCC) on a special partition of their top leading machine [4]. This partition includes 80 nodes of arm FX64 processors[5]. This scalar architecture implement the Scalable Vector Extension (SVE), an SIMD extension that allows vectorisation on up to 8 double precision real. The bandwidth per core is rather large (21Gb/s per core) but to get the most of it, the particular memory access technology supposes a fine tuning of the threads mapping, that NEMO cannot provide due to the lack of OpenMP parallelism. However, we propose to test the model in its standard MPI configuration, assuming that a further implementation of OpenMP could give even better performance (on this topic, see [5]) . The same NEMO 4.0.4 version than on NEC SX-Aurora TSUBASA is used, without any IO or additional module (no sea-ice, no BGC).

This is not the first time that NEMO performance was evaluated on an arm processor [6] but this previous attempt, on a Cavium ThunderX2, was led with the Intel compiler instead of the native arm compiler, not available at that time. Performance just similar to the Intel Broadwell reference was observed. Similarly, a recent test on Fujitsu A64FX preferred the GNU compiler [7] and gave deceptive results.

4 `https://www.top500.org/system/179700/`

5 A64FX armv8.2-A SVE @1.8Ghz, 48 cores per node

In this study, we propose to use the native Fujitsu compiler[6] with basic optimisation options[7]. No code modification was necessary for porting, except the adding of the `key_nosignedzero` pre-processing option.

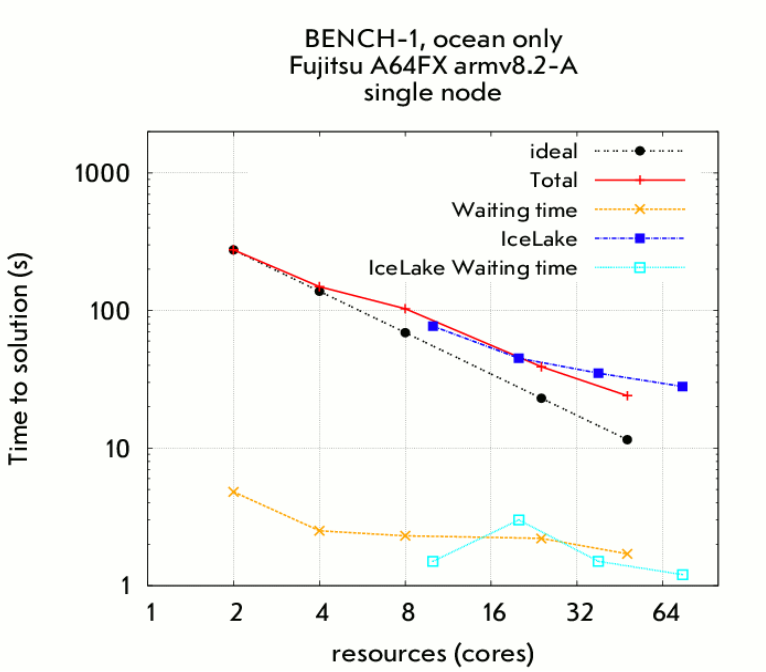

Figure 1: BENCH-1 scalability and communication time on Fujitsu A64FX armv8.2-A and Intel Ice Lake Xeon Gold 6140 @2.30GHz as reference

We present in Fig. 1 the total time to solution (red line) to complete 100 iterations of the main time loop, for several MPI decompositions in a single node. One resource (core), allocated on a single node, is attributed to each subdomain. We call waiting time (orange line) the time spent by NEMO waiting for the horizontal boundary conditions (halos), communicated from their neighbors subdomains. For comparison, the performance of the same model on one Intel Ice Lake node is plotted in blue. The vector `-xCORE-AVX512` option is activated on this machine. Both time to solution and waiting time (mainly load imbalance and time spend in MPI communications) are the same, which does not show a significant improvement but validates the good behavior of our model on the Fujitsu-arm processor, without many porting effort.

To go further, we try to use more than one computing node. Unfortunately, it was not possible to get reasonable performance or even to start simulation on more than 2 nodes of the machine. This can be explained by the relatively unconventional parallel strategy we deploy (full MPI), known to be inefficient on this kind of hardware. In addition, the interconnect network is not the native "Tofu" Fujitsu but the Infiniband EDR (100Go) rather similar to the Infiniband HDR already in use on the TGCC supercomputer. This could also contribute to significantly downgrade our performance. However, we plotted in Fig 2. the performance

6 `frt 4.6.1`

7 `-Free -Kfast -CcdRR8 -zcfc=target_sve,` with `-Kfast` equivalent to `-O3 -Komitfp,mfunc,eval,fp_relaxed,fz,fp_contract,ilfunc,simd_packed_promotion`

measured on two Fujitsu nodes, and compared it with the performance measured on the new Intel Ice Lake platform installed at CERFACS, which again are in the same range.

3. Perspectives

The NEMO porting on these two new platforms appeared to be easy, but getting optimal performance required more efforts. On NEC vector engine, in addition to the already mentioned vector length related limitation for an optimal MPI decomposition, our tests showed additional bottlenecks in sea-ice and biogeochemistry modules. On Fujitsu, an optimum use of the memory hierarchy still need to be done, but requires a specific NEMO implementation like tiling or OpenMP parallelism.

Interesting performance at low development cost seems to be more affordable on the NEC vector machine. Our result does not allow to determine the level of efficiency we could reach on Fujitsu. The hardware dependency of the comprehensive rewriting needed is another concern on this platform. On a slightly similar architecture [8], a very different strategy (parallelism on the vertical levels, asynchronous communications) was recently deployed, with a totally different impact on the code.

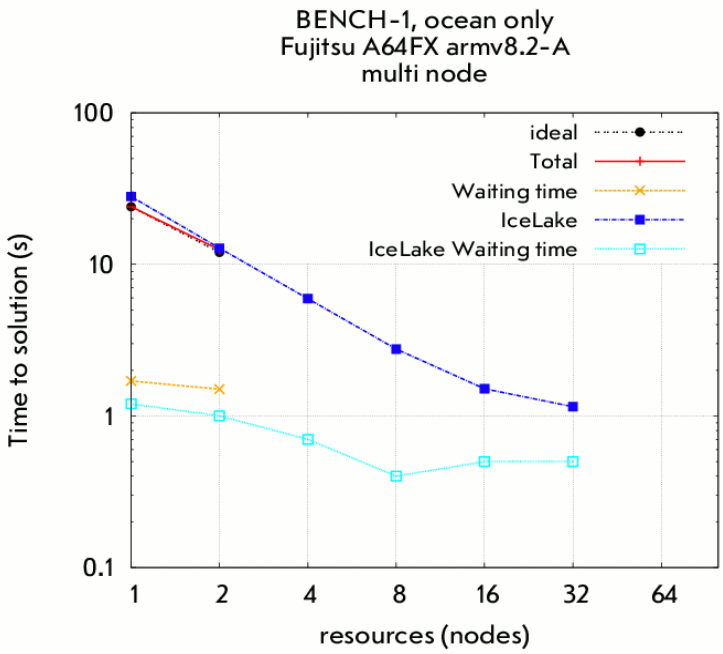

Figure 2: Same than Fig. 1. but on more than one node

To conclude, our knowledge were insufficient to be able to establish the power consumption of our model on these platforms. Unlike on AMD or Intel [9], we lack here of the appropriate measurement software, that would also tell us how vectorisation [10] helps to save energy on

NEC machines.

Despite these limitations, we keep thinking that the NEC machine 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 is not a bottleneck but benefit can be taken from the high processor throughput. Beside the never ending race to electricity wasting, we think that there is room for energy efficient and scientifically justified ocean simulations that must be performed on the right equipment.

Acknowledgments

Thanks to Wonsun Park and Mojib Latif (GEOMAR) for providing the access to the Kiel Christian-Albrecht University (CAU) supercomputer, to Jens-Olaf Beismann (NEC-Germany) and Isabelle d'Ast (CERFACS) for their respective support on the CAU NEC/CERFACS Intel Ice Lake machines, to Vera Maurer (DWD) for providing the NEC compatible libraries for XIOS. This work was granted access to the HPC resources (Fujitsu) of Saclay TGCC super-computing center under the allocation AP010113157 made by GENCI and to the HPC resources (NEC) of the Computing Center of the CAU. 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, in addition to graphics from Matplotlib, a Sponsored Project of NumFOCUS, a 501(c)(3) non profit charity in the United States. This project did not received funding from the European Union's Horizon 2020 research and innovation program.

References

[1] Maisonnave, E., 2021: NEMO performance optimisation on NEC SX-Aurora TSUBASA , Working Note, **WN/CMGC/21/37**, CECI, UMR CERFACS/CNRS No5318, France

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

[3] Dongarra, J., 2020: "Report on the Fujitsu Fugaku system" *University of Tennessee-Knoxville Innovative Computing Laboratory, Tech. Rep. ICLUT-20-06*

[4] 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

[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] d'Ast, I., Maisonnave, E. & Monnier, N., 2019: arm CAVIUM THUNDERX2 benchmarks (in French), CERFACS Internal Report, France

[7] Banchelli, F., Peiro, K., Ramirez-Gargallo, G., Vinyals, J., Vicente, D., Garcia-Gasulla, M., & Mantovani, F. , 2021: Cluster of emerging technology: evaluation of a production HPC system based on A64FX, in *2021 IEEE International Conference on Cluster Computing (CLUSTER)* (pp. 741-750)

[8] Ye, Y., Song, Z., Zhou, S., Liu, Y., Shu, Q., Wang, B., Liu, W., Qiao, F., & Wang, L., 2022: swNEMO_v4.0: an ocean model NEMO for the next generation Sunway supercomputer, *Geosci. Model Dev. Discuss.* [preprint], https://doi.org/10.5194/gmd-2022-33, in review

[9] 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

[10] Guermouche, A. ,& Orgerie, A.-C., 2021: Thermal design power and vectorized instructions behavior. Concurrency and Computation: Practice and Experience, Wiley, In press, pp.1-18. 10.1002/cpe.6261.hal-03185821