INMIO high resolution global ocean model as a benchmark for different supercomputers *

K.V. Ushakov[1,3], R.A. Ibrayev[2,1,3], M.N. Kaurkin[1,3]

Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow, Russia[1],
Marchuk Institute of Numerical Mathematics of RAS, Moscow, Russia[2],
Marine Hydrophysical Institute of RAS, Sevastopol, Russia[3]

The purpose of this work is to demonstrate the cross-platform and speed estimation of the global ocean dynamics model INMIO (open-source, GPLv2, `http://model.ocean.ru`) [1], developed by the authors on various supercomputers. For this, the INMIO model was implemented in a high (eddy-resolving) resolution mode, which is currently being successfully used in numerical experiments to study the global eddy meridional ocean heat transport. Performance tests were carried out on three Russian supercomputers MVS-10Q (JSCC RAS), Cray XC-LC (Roshydromet), Sugon (MHI RAS) and utilized up to 1500 processor cores. All tested supercomputers have processors of the same architecture and similar clock frequency (Intel Xeon E5, the possible performance difference is about 10%), but different inter-node connects: Omnipath, Aries and Infiniband FDR, respectively. So, all three supercomputers use different interconnects operating at speeds from 56 Gb/s (Infiniband FDR) to 120 Gb/s (Cray Aries). Therefore, from the analysis of the obtained performance results, it is possible to make conclusions which of the interconnects is better suited for the task of high-resolution ocean modeling.

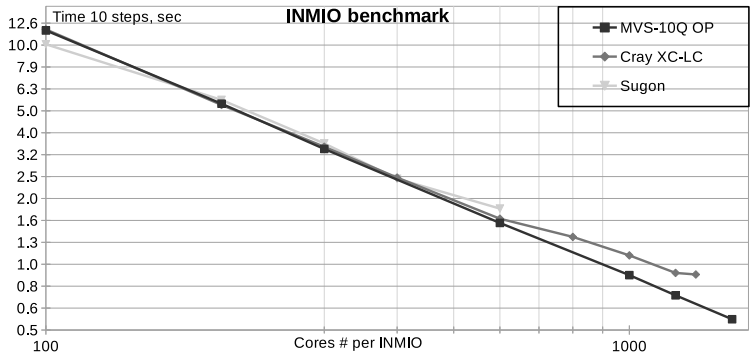

Figure 1. Dependence of the execution time of 10 steps of the *INMIO benchmark* on the number of processor cores on various supercomputers. Logarithmic scale.

The INMIO model was tuned to a fixed configuration: World Ocean spatial resolution of 0.1 degrees and 49 vertical levels. The average time of 10 model steps was measured. Each step, in addition to the required local (in each processor subregion) geophysical calculations, is accompanied by updating the required fields in the boundary cells using CMF [2] functions based on persistent communication queries (`MPI_SEND_INIT`, `MPI_RECV_INIT` and then `MPI_STARTALL`, `MPI_WAITALL` procedures), which allows reducing the overhead costs of communication between processes and communication controller. In this configuration, the boundary band around the perimeter of each processor subregion for exchanges has a width of one grid node, but can be increased, which is sometimes required for more complex difference schemes. The results of numerical experiments are summarized in Fig. 1 which shows the execution time of 10 model steps vs. the number of cores used.

MPI profiling. To assess the effectiveness of the parallel implementation of the INMIO

* The research was supported by the Russian Science Foundation (project no. 17-77-30001) and performed at the Federal State Budget Scientific Institution Marine Hydrophysical Institute of RAS.

model, we have to profile it using `mpiP` library [3]. It uses statistical sampling to record profiling data, thus no code changes are required (only compiled with the -g flag and linked with the `mpiP`). The `mpiP` performance report contains the following sections: the percentage of time each rank is spending in MPI calls; MPI call sites in the code; the top call sites that spend the most time in MPI and send the most data; MPI call site statistics (number of times called, average/max/min time spent, and percentage of time; average/max/min/total bytes sent).

By analyzing these reports we can draw a number of conclusions. There is no hard disbalance among the cores, they all use MPI by 50-60%. At the same time, with an increase of the number of cores, this ratio does not change. This means that, from an architectural point of view, the model is implemented correctly and an increase in the number of used cores does not lead to a sharp increase in the number of exchanges, which is confirmed by the previous experiments.

From the point of view of communications, the `o_d_brtr_sweq_sngl` (shallow water equations) and `o_td_scalar_advection_fctz` (Zalesak tracer transport scheme) subroutines are very costly. Since they require addition exchanges of data in the boundary cells, the data is needed immediately after the exchange. In total, communications in these subprograms occupy about 31% and 11% of the entire program work time. It is worth noting that improving interconnect performance does not solve this problem, which is proven experimentally on Omnipath, since it is caused by the heterogeneity of the computational load between the processor subdomains (for example, due to the presence of land cells). This is supported by the fact that the top of `Aggregate Times` consists of calling `MPI_WAITALL` operations. But the amount of sent/recv data is relatively small (about 4 GB per model day integration, see `Aggregate Sent Message Size` section of the report). Also, significant costs (9.55% of the entire program) are required by the operation `MPI_ALLREDUCE` when calculating the average level throughout the ocean in order to prevent its drift.

Conclusion. In this paper, an attempt was made to use a direct numerical experiment to understand the efficiency of the model on various clusters and how important high-performance interconnect is for model scalability. Experiments have shown that even an IB FDR is sufficient to provide near linear scalability (at least up to 600 cores). In this case, a small difference in the performance should be attributed to the difference in the frequency of the processors. This results are confirmed by profiling with `mpiP`. This kind of research is important to conduct when a software product (INMIO model) becomes distributed as *open-source* and its users can understand what performance they can rely on with their equipment. Similar studies have been published for the HYCOM [4] and POP [5] models of ocean dynamics used worldwide.

References

1. Ibrayev, R.A., Khabeev, R.N. and Ushakov, K.V. 2012. Eddy-resolving 1/10° model of the World Ocean. Izvestiya, Atmospheric and Oceanic Physics. 48(1), pp. 37-46.

2. Kalmykov, V.V., Ibrayev, R.A., Kaurkin, M.N. and Ushakov, K.V. 2018. Compact Modeling Framework v3.0 for high-resolution global ocean-ice-atmosphere models. Geoscientific Model Development. 11. pp. 3983-3997.

3. mpiP: Lightweight, Scalable MPI Profiling V.3.4.1 , `http://mpip.sourceforge.net`

4. Barker, Kevin J. and Darren J. Kerbyson. 2005. A Performance Model and Scalability Analysis of the HYCOM Ocean Simulation Application. IASTED PDCS. pp. 650-658.

5. Jones, P.W., Worley, P.H., Yoshida, Y., White, J.B. and Levesque, J. 2005. Practical performance portability in the parallel ocean program (POP), Concurrency and Computation: Practice and Experience, 17(10), pp. 1317-1327.