

Enabling grid-driven supercomputing for oceanographic applications – theory and deployment of hybrid OpenMP+MPI parallel model for the OPATM-BFM application

Alexey Cheptsov

e-mail: cheptsov@hls.de

Service Management and Business Processes department,
High Performance Computing Center Stuttgart (HLRS)
University of Stuttgart
70550 Stuttgart
Germany

Shortly about the author

Dr. Alexey Cheptsov joined HLRS after attaining a Ph.D. degree at the Research Institute of Simulation Problems in Power Engineering, National Academy of Science of Ukraine. After successful completing the Int.EU.Grid project, he currently works at the Service Management and Business Processes group as a research scientist, focusing on parallel and distributed computing of large-scale scientific applications. His scientific background includes computer science and progressive information technology, in particular simulation support of complex dynamic systems and parallel and high-performance simulation technology. He is currently involved in the DORII project, where he is leading the research activity, and LarKC EU projects.

Research object

The object of the project investigation was the OPATM-BFM application [1]. OPATM-BFM is a MPI-parallel physical-biogeochemical simulation model developed in the frame of the MERSEA project [2] and practically applied for short-term forecasts of key biogeochemical variables (e.g. chlorophyll, salinity and other) for a wide range of coastal areas, among others Mediterranean Sea and used in the DORII project [3].

The model solves the transport-reaction equations (1) for the generic biogeochemical concentration c_i based on the advection-diffusion processes:

$$\frac{\partial c_i}{\partial t} + v \cdot \nabla c_i = w_i \frac{\partial c_i}{\partial z} + k_h \nabla_h c_i + \frac{\partial}{\partial z} \left[k_z \frac{\partial c_i}{\partial z} \right] + R_{bio}(c_i, c_1 \dots c_N, T, I \dots) \quad (1)$$

where v is the current velocity, w_i is the sinking velocity, k_h and k_z the eddy diffusivity constants and R_{bio} is the biogeochemical reactor that depends, in general, on the other concentrations and on temperature T , short-wave radiation I and other physical variables.

The complexity of the OPATM-BFM model consists in the high number of prognostic variables to be integrated, dimension of analysed ecosystems and usability for the long-term forecasts. In this context, the OPATM-BFM application poses some challenging scenarios for the efficient usage of modern HPC systems the application is running on.

The current realization

The core of the OPATM-BFM makes a 3D Ocean General Circulation modelling System [4] up, coupled off-line with the BFM chemical reactor (with the $1/8^\circ$ horizontal resolution and 72 vertical levels), developed at the Instituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS) of Trieste, Italy. The adoption is done programmatically through the hard-coding. That rules the possibility to use newer versions of the third-party software (mainly with regard to the OPA part) practically out.

OPATM-BFM is parallelized using the domain decomposition over longitudinal elements (Figure 1), that enables using massive-parallel computing resources for the application execution. The number of domains corresponds to the number of computing nodes, the application is running on. The consistence of the computation by the parallelization and domain decomposition is ensured by the inter-domain communication pattern that enables passing data cells resided on the domain bounds needed for the computation inside domains.

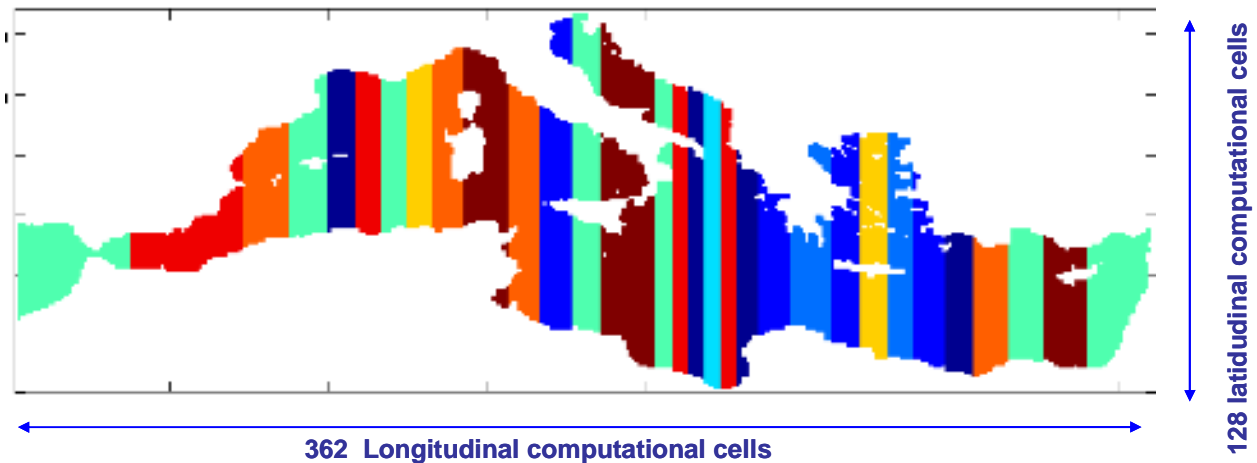


Figure 1. Domain decomposition used in the OPATM-BFM.

The current realization of the communication pattern is purely based on the Message-Passing Interface (MPI) and implemented by means of single point-to-point operations. Cumulatively, more as 250000 messages are transmitted on each step of the numerical solution in the current realization. The analysis duration for a standard short-term forecast (17 days) on a generic XEON-cluster is approximately 8 hours. The size of the standard input data set is 1.5Gb. For the standard use case, the simulation model produces 80Gb of the output data.

Research's objectives

Due to the parallel implementation (by means of MPI), the OPATM-BFM simulation model has been applied for a whole basin of the Mediterranean Sea. Being the core of the ecological forecasting system developed in framework of the Italian Group of Operational Oceanography, OPATM-BFM can be safely used for large-scale assessments, such as: estimation of the carrying capacity of the Mediterranean Basin (a valuable information for ecosystem-based approaches to fisheries management), provision of habitat suitability indicators for risk assessment of a non-Mediterranean species invasion (such as algae), regional ecosystem responses to climate change, scenario analyses, design of observational research cruises and activities and many others.

In light of promising results that has been achieved using an IBM SP5 machine of CINECA [5] (PowerPC processors) for the short-term forecast, the model is expected to produce additional output products and information at different time scales (the long-term forecast for up to 15 years of analysis is a challenging scenario of the planned OPATM-BFM usage), as well as for a wider set of climatic scenario analyses (multi-decadal period of integration etc.). With regard to this, obtaining maximal performance (in terms of the execution time) and scalability (in terms of speed-up gained due to running on the increasing number of computing nodes) is mandatory for OPATM-BFM practical usability for tasks of the real complexity. Moreover, the application poses great challenge for different HPC architectures with regard to both optimal utilization of resources for performing the identified complicated tasks of environmental simulation and development of algorithms enabling such an efficient usage.

The computation time can be reduced by scaling the number of computing nodes the application is running on. This requires reconfiguration of the domain decomposition, corresponding to the number of available computing nodes, ensuring the optimal load balancing of high-performance resources' nodes. For the current longitudinal configuration, the decomposition only for 16, 32 and 64 domains is foreseen by the OPATM-BFM realization. Moreover, the performance effect gained due to scaling the

number of computing nodes is not optimal (as the preliminary analysis had revealed, the performance speed-up by scaling from 32 to 64 nodes is less as 50%). This proves the necessity of the qualitative improvement of the application communication pattern. That involves a detail analysis of the communication pattern, resulting in performance shortcomings and bottlenecks identification. According to the analysis results, proposals for improving the application performance and resources utilization (load-balancing) are to be identified. The realization of those proposals should also allow the OPATM-BFM to overcome the most considerable performance limitation – scaling only up to 64 computing nodes, due to the specific of the realized domain decomposition.

In the frame of the proposed HPC-Europa project on “Enabling grid-driven supercomputing for oceanographic applications – theory and deployment of hybrid OpenMP+MPI parallel model for the OPATM-BFM application”, apart from a number of improvements of the current “pure MPI” realization of the application communication and I/O pattern, the combination of memory parallelization on the node interconnect with shared memory parallelization inside the node will be implemented. This will result in a hybrid OpenMP+MPI model for the OPATM-BFM, ensuring that the full potential of HPC is enabled.

The research should also support the OPATM-BFM developers in further understanding the communication patterns, improving the computing nodes’ load balance for different use cases, defining bottlenecks more precisely as well as working out solutions for resolving the shortcomings and maximizing the application performance and scalability. Moreover, the results obtained for the OPATM-BFM will be important for other scientific parallel applications that implement similar communication pattern.

Development of the hybrid OpenMP+MPI model

With regard to the communication pattern, there were some strong arguments in favour of a hybrid model which tend to underline the assumption that OpenMP+MPI realization of the communication pattern should lead to improved parallel efficiency of the OPATM-BFM as compared to current pure MPI implementation [6] . This is especially important from the perspective of the application evolution towards running in high-performance grid environments, where the price/performance sweet spot is settled at a point (European Grid e-Infrastructure, like those set up by EGEE [7] and DEISA [8] projects). Obviously, obtaining high performance and scalability of the application is a necessary condition for the efficient operation of the identified tasks on a standard element of a high-performance Grid – a cluster of multi-core SMP nodes.

The experience for other applications from different problem areas has shown that to the expected benefits of the hybrid parallelization can be referred: enabling additional level of parallelism with inherent performance and scalability optimization, improved load balancing, reduced memory consumption and many others [9] . By the development of the hybrid OpenMP+MPI communication pattern we were guided by the experience of the Earth Science Department of the Barcelona Supercomputing Center that acted as the main hosting organization for the proposed HPC-Europa project.

The high-performance resources used

The investigation has been performed on the supercomputer MareNostrum located at Barcelona Supercomputing Center - Centro Nacional de Supercomputación (The Spanish National Supercomputing Center). MareNostrum is one of the most powerful supercomputer in Europe. The supercomputer consists of 2560 JS21 blade computing nodes, each with 2 dual-core IBM 64-bit PowerPC 970MP processors running at 2.3 GHz for 10240 CPUs in total. For our tests, a total of 512 cores were available.

Software tools used

For application run profiling we used tools from the Valgrind tool suite [10]. The investigation of the message-passing communication pattern was performed using the Paraver tool, developed in the Barcelona Supercomputing Center [11]. The Paraver was also very beneficial for our investigations, as it fully supports hybrid OpenMP+MPI communication models.

Main results and achievements

- We ported the OPATM-BFM application to the Marenostrom, adapting the sources and supplement files to the specific of available compilers, libraries, etc. Corresponding changes have been committed to the native source code repository.
- We identified the tool set for performance analysis of the application. For application run profiling we used tools from the Valgrind tool suite. Benchmarking both inter-domain message-passing and file I/O communication patterns was performed using the Paraver tool. The Paraver was also very beneficial for our investigations, as it fully supports hybrid OpenMP+MPI communication models.
- We identified the main phases of the application execution (initialization, iterative main simulation routine, synchronisation, data flushing and storage) and performed further analysis of the communication pattern with regard to those phases.
- We specified the test use case consisting of only three initial steps in the main simulation routine. This allowed us to avoid profiling periodically repeated parts and minimize the size of the recorded trace data with time stamps of occurred communication events.
- We holistically investigated each of the application phases with regard to the communication and file I/O events for the test use case and measured time characteristics scaling the number of computing nodes.
- Based on the analysis results, we identified main shortcomings of the communication pattern with the most impact on the performance degradation. To those can be referred: passing a big number of small messages between domains (for example, in the advection routine 1264 messages are transmitted on every step of the numerical solution, whereas the size of the message is only 4Kb), serialization of the write-out mechanism by the root process that composes the entire domain from all processes through MPI communication and then stores the complete dataset to a NetCDF file, sequential performing file I/O operations and others.
- We implemented encapsulation of transmitted data to segments. The optimal size of the segment was found for each of the main application routines. This allowed us to decrease the communication time by 50%.
- We implemented parallel data input from NetCDF files (the time reduction was more as 60%), that was a serious bottle by scaling the number of computing nodes.
- We measured the overall impact of proposed improvements for both test and real use cases. Whereas file I/O operations are dominating for the test case, the overall performance improvement due to optimization of the MPI communication becomes significant only for a long-term simulation (816 steps for the real case). The total amount of realized optimizations allowed us to reduce the duration of the application execution for the real case by up to 5% (from initially measured 309 min down to 293 min). Furthermore, the optimization for the increasing number of nodes from 32 up to 64 grew up to 15% (from 213 min down to only 185 min). The application scalability grew accordingly from 145% up to 158%.
- We identified source code's regions where shared memory parallelization inside the node can be implemented.
- We elaborated mechanism how to append the memory parallelization on the node interconnect with shared memory parallelization inside the node for the identified regions.
- We developed the pilot version of the hybrid OpenMP+MPI communication pattern. This allowed us to test the application performance scaling the number of computing nodes up to 512. The expected

outcome of the communication pattern's hybridisation is performance grow up to 10 times.

- Based on results of the performed investigations, we prepared a scientific publication ("Analysis and optimization of performance characteristics of parallel scientific applications").

- We will keep investigating the possibilities of the communication pattern hybridisation, in particular for different architecture platforms, in order to ensure the optimal utilization of the high-performance resources.

Difficulties we faced during the research

As the analysis had revealed, the application communication pattern can be characterized through the very intensive inter-domain element exchange, that takes more as 60% of the execution time. Hence, the performance impact of the proposed hybrid OpenMP+/MPI communication pattern's realization is not as considerable as expected before starting the research. Therefore, additional research on optimization of the already implemented communication pattern is necessary, in order to minimize the total communication time down to at least 20%, whereby the use of the hybrid OpenMP+/MPI communication pattern will be really beneficial.

Project resource utilization

The deviation from the requested CPU hours with actually consumed amount is expected to be not higher as 10%.

Scientific achievements

As the outcome of this HPC-Europa attendance, a common article on "Analysis and optimization of performance characteristics of parallel scientific applications on the Grid (a case study for the OPATM-BFM environmental application)" was published in the proceedings of the conference "Instrumenting the Grid – INGRID'09". The results will be also presented in the scientific seminars of the HLRS, in particular in the HPC-Surgeries organized in frame of the HPC-Europa project.

Acknowledgments

The work has been performed under the HPC-EUROPA2 project (project number: 228398) with the support of the European Commission - Capacities Area - Research Infrastructures.

I am very grateful to the HPC-Europa project and BSC for providing the presented research project with an access to the high-performance facilities of the Marenostrom system and the HPC-Europa project's supporting team, especially to Mrs. Maria Carreras Godal, for the organization of the visit.

I would also like to thank to the Head of the Earth Science Department, Prof. Jose Baldasano for hosting the project, as well as the Head of the Computer Science Department, Mr. Jesus Labarta for valuable discussions with him and his colleagues during the stay in his department.

References

- [1] A. Crise, P. Lazzari, S. Salon, and A. Teruzzi. MERSEA deliverable D11.2.1.3 - Final report on the BFM OGS-OPA Transport module, 21 pp., 2008.
- [2] See the web page of the Marine Environment and Security for the European Area (MERSEA) European Integrated project – <http://www.mersea.eu.org>
- [3] See the web page of the Deployment of the Remote Instrumentation Infrastructure (DORII) project – <http://www.dorii.org>
- [4] Madec G., P. Delecluse, M. Imbard, and C. Lévy, 1998: OPA 8.1 Ocean General Circulation Model reference manual. Note du Pôle de modélisation XX, Institut Pierre-Simon Laplace, France, 91 pp.
- [5] See the description of the IBM SP5 machine on the CINECA's web page – <https://hpc.cineca.it/docs/user-guide-zwiki/SP5UserGuide>

- [6] Georg Hager, Gabriele Jost, and Rolf Rabenseifner: Communication Characteristics and Hybrid MPI/OpenMP Parallel Programming on Clusters of Multi-core SMP Nodes. Proceedings of the Cray Users Group Conference 2009 (CUG 2009), Atlanta, GA, USA, May 4-7, 2009, https://fs.hlr.de/projects/rabenseifner/publ/CUG09_Hager_Jost_Rabenseifner.pdf
- [7] Enabling Grids for E-Science (EGEE) project website, <http://www.eu-egee.org/>
- [8] Distributed European Infrastructure for Supercomputing Applications (DEISA) project website, <http://www.deisa.eu/>
- [9] B. Simo, O. Habala, E. Gatia, L. Hluchy. Leveraging interactivity and MPI for environmental applications. Computing and Informatics, Vol. 27, 2008, 271-284.
- [10] J. Seward, N. Nethercote, J. Weidendorfer and the Valgrind Development Team. Valgrind 3.3 - Advanced Debugging and Profiling for GNU/Linux applications. <http://www.network-theory.co.uk/valgrind/manual/>
- [11] See the description of the Paraver tool at the site of the Barcelona Supercomputing Center - http://www.bsc.es/plantillaA.php?cat_id=485