Geophysical Research Abstracts
Vol. 14, EGU2012-12983, 2012
EGU General Assembly 2012

Towards petascaling of the NEMO ocean model

J. Donners (1), N. Audiffren (2), and J.-M. Molines (3)

(1) SARA, Amsterdam, The Netherlands, (2) CINES, Montpellier, France, (3) LEGI, Grenoble, France

PRACE, the Partnership for Advanced Computing in Europe, offers acces to the largest high-performance computing systems in Europe. These systems follow the trend of increasing numbers of nodes, each with an increasing number of cores. To utilize these computing systems, it is necessary to use a model that is parallellized and has a good scalability. This poster describes different efforts to improve the scalability of the NEMO ocean model.

Most importantly, the problem size needs to be chosen adequately: it should contain enough computations to keep thousands of cores busy, but foremostly it has to be scientifically relevant. The global, 1/12degree, NEMO ocean model configuration, developed by the Mercator team, is used for operational ocean forecasting. Therefore, PRACE selected this model for the PRACE Benchmarking suite.

However, an increased problem size alone was not enough to efficiently use these petascale systems. Different optimizations were required to reach the necessary performance. Scientifically, the model should simulate one year within a wallclock day. Technically, the application needs to scale up to a minimum number of cores. For example, to utilize the fastest system in Europe, the new Curie system in France, the lower limit is 2048 cores.

Scalability can be increased by minimizing the time needed for communication between cores. This has been done in two ways.

Firstly, advanced parameters of the MPI-communication library were optimized. The improvement consists in:
1. using RDMA for eager messages (NEMO messages size are below the eager size limit) conjugated with adequate openib flags.
2. tuning for openMPI for collective communication through the btl_coll_tuned_dynamic_rules flag.
Overall, the improvement is 33%.

Secondly, NEMO uses a tri-polar and staggered grid, which involves a complicated fold across the north-pole. Communication along this fold involves collective gather and scatter operations which create a bottleneck at a single core, so these were replaced by a small series of send-receive messages. This increases performance by 35% when using 2048 cores.

The distribution of tasks is of primary order. We face to multi-exchange starting at the same time on every rank. If communicating ranks are on the same node but should also communicate with external processes a contention is observed on the Global queue and the QPI on the node. We get an improvement of 15% on 3584 cores when the processes only communicate through the InfiniBand network.

These optimizations are especially advantageous at scale, where the fraction of computation versus communication decreases rapidly.