Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-301-AC1, 2018

Geoscientific Model Development Discussions

Interactive comment on "Simulation of the Performance and Scalability of MPI Communications of Atmospheric Models running on Exascale Supercomputers" *by* Yongjun Zheng and Philippe Marguinaud

Yongjun Zheng and Philippe Marguinaud

yongjun.zheng@meteo.fr

Received and published: 19 March 2018

Dear reviewer,

Thank you very much for your comments. The followings are our responses. Please also find the revised manuscript in the supplement to this comment.

Best regards,

Yongjun ZHENG and Philippe MARGUINAUD

Reviewer #1:

The authors present work of simulated scaling analysis for different communication algorithms commonly used in atmospheric models using a skeleton codes and a simulation package to examine the scaling performance on possible future supercomputers. This represents significant new information on how these algorithms may perform and is likely to be of interest to the community. The methods used are well described and appear robust.

Thank you very much for your careful comments.

Some of the assumptions made about future architectures, in effect single CPU core nodes, are unlikely to be entirely valid. Whilst these are made the entirely reasonable purpose of make the simulations tractable, they may weaken some of the conclusions. For example, almost all CPU base supercomputer are multi-core and multi-socket nodes which then have significant network hierarchy. Moreover, many of the largest machines in the top 500 list have non-CPU architectures such as GPUs and Xeon Phi. These have more complex hierarchies and are unlikely to, or even cannot be, programmed with a single MPI rank bound to single "core". Whilst the authors don't hide this, this is not discussed in the conclusions.

The main purpose of this study is to analyse the performance and scalability of communications over an interconnect network between nodes. Thus, single CPU core per node is adopted; because this not only makes the simulations tractable, but also eliminates the intra-node communications, which in turn makes it easy to draw robust conclusions for the inter-node communications without the complicated hierarchical

network. But we totally agree with the reviewer that the architectures of current and futures supercomputers are multi-core and multi-socket nodes, even non-CPU architectures; intra-node communications significantly distinguish from inter-node communications. For example, some MPI implementations implement the intra-node communication using the shared memory communication mechanism for multi-core and multi-socket nodes, or using proprietary inter-processor networks and API for non-CPU architectures. However, an MPI rank can be bound to any core for multi-core and multi-socket nodes; and an MPI rank can be bound to any processor/co-processor for MIC architectures such as Xeon Phi; with CUDA-aware MPI, an MPI rank can be bound to a CPU core but can communicated with GPUs for GPU architectures. Because A multi-core node behaves more or less like a more powerful single core node when the OpenMP is used for the intra-node parallelization, the assumption of an MPI rank bound to a single core should apply to the complex hierarchical system. We have added a discussion for more complex hierarchical architectures in the conclusions. Please refer to the conclusions section in the revised manuscript.

Most of the results are presented in the form of graphs. Unfortunately, they are simply too small and it not possible to read the legends, axis labels etc. This makes it difficult to judge the quality of the results and the inferences drawn. These should be reproduced to appear much larger.

Thank you very much for pointing out the regibility of some figures. We have reproduced most the figures so that they are legible, especially for the legends and axis labels.

Moreover, it would appear (although hard to be sure) that some of the plots have num-

ber of processors as the x-axis. This is a discrete variable and so line graphs should not be used, a bar chart may be appropriate. Whilst it may be common practice to present scaling data in this way, it is still wrong. This paper has the potential to become an interesting and significant work, but not in its current form.

Fig.5, Fig.9a-c, and Fig.10a-b have number of processes as the x-axis which is a discrete variable. In the revised manuscript, we have added one statement (lines 405-408) about the discrete values adopted in this study. We have tried to change the line plots to a bar chart, but it is not as clear as a line to demonstrate the trend of communications times, which varies as the number of processes increases. But we changed the lines to the lines with markers which indicates the number of processes, and added explanations in the captions of the figures. Thank you again for your careful comments.

Once some revisions have been made it should be review again. In particular, there are three changes which are necessary.

i) The plots must be made bigger so they are legible

We have changed the Fig.3, Fig.4, especially, Fig.5, Fig.8, Fig.9, and Fig.10 so that they are legible now. Please refer to the revised manuscript.

ii) Plots against discrete variables shouldn't be line graphs
As mentioned above, we have changed the Fig.5, Fig.9a-c, and Fig.10a-b. Please refer to the revised manuscript.

iii) The authors should comment on and discuss what conclusions can be drawn from simulations of single core nodes for more complex node architectures and the consequent differences to communication patterns.
As mentioned above, the binding of an MPI rank is possible for non-CPU architectures; thus, the conclusions for inter-node communications could be generalized to more complex node architectures. As we already discussed, the intra-node communications significantly distinguish from inter-node communications, and multiple MPI processes per node in complex node architectures may result in congestion in the network interface controller for inter-node communication. The congestion can be mitigated even eliminated if more network interface controllers per node or a network interface controller with multi-ports (such as a mini-switch) in a node. From this point of view, our conclusion should still be valid for this complex hierarchical architectures, but the scalability might be affected. We agreed with the reviewer that a discussion should be included and we have added a statement about this. Please refer to the conclusions section in the revised manuscript.

Please also note the supplement to this comment:
https://www.geosci-model-dev-discuss.net/gmd-2017-301/gmd-2017-301-AC1-supplement.pdf

---

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-301, 2018.