Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-301-AC3, 2018

Geoscientific
Model Development
Discussions
Open Access

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: 11 May 2018

Dear reviewer,

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

Best regards,

Yongjun ZHENG and Philippe MARGUINAUD

Reviewer #2:

The article presents an important aspect often ignored in NWP model development. It studies the impact of network topology, not only for one particular algorithm, but for multiple representative algorithms found in NWP models. It illustrates that the choice of equivalent but different numerical algorithms may well depend on the available network layout. In this case a semi-Lagrangian approach using nearest-neighbour communication for wide halo-exchange is studied. Further, a spectral transform method is studied consisting of large distributed matrix transpositions, and finally a Krylov solver consisting of multiple AllReduce operations. The results are presented in a detailed yet clear manner.

I have attached a edited PDF of the original article containing comments and suggestions. If these are addressed, I am happy to see the article published.

We really appreciate your comments and your efforts to edit the original manuscript.

For clarity, I will report the major comments and questions below (besides being present already in the attached PDF).

1) Throughout the article the term ‘radix" is used. It would be good to formulate a definition of it in this paper’s context.

After searched the term ‘radix" in the manuscript, we found there are mainly three parts that uses the term ‘radix":

1. In Table 1 in page 10, the last column ‘radix" is related to the number of ports of

a switch. We found this radix is never referenced in this paper, so it is removed from the Table 1 to avoid the confusion with the following two usages.

2. In Table 2 in page 14, in the description of the ring-k algorithm for a spectral transposition, the ‘radix k" represents the number of processes to (from) which a process sends (receives) messages. Thus, the ‘radix k" is self-explained.
3. The remainning usages of the term ‘radix" are for the recursive-k algorithm for the allreduce operation. The ‘radix k" represents the number of processes involved in a sub-reduce operation of the resursive-k algorithm. Since this is not obvious, we added the definition of ‘radix k" and made some changes so that the description of the recursive-k algorithm is more accurate. Please refer to the revised manuscript (lines 390-404).

2) Line 135: I recommend following more representative citation instead of "Kuhnlein et al., 2017": Smolarkiewicz et al., 2016: A finite-volume module for simulating global all-scale atmospheric flows, J. Comput. Phys., 314, pp. 287-304, doi:10.1016/j.jcp.2016.03.015

The citation has be changed, please refer to line 136 in the revised manuscript.

3) Line 363: It’s worth noting that this regularity is only possible for structured grids. Even then there are differences between regular and reduced grids. Unstructured grids would not have a preferred x or y sweeping, and communication must be done in a single sweep. Does the following analysis still hold in this case?

Yes, we agree with you that halo exchange for unstructured grids must be done in a single sweep. Two sweep method for a regular grid has the advantage that each process

only exchange messages with his TWO neighbors in the corresponding direction, less processes involved in a communication usually reduce the possibility of congestions; but two sweep method has an overhead time since the second sweep has to wait for the finish of the first sweep. One single sweep method can avoid the overhead time in the two sweep method; but its disadvantage is that each process need to communicate with its EIGHT neighbors simultanuously, this would increase the possibility of congestions. In short, these two methods should have be similar in term of communication times. We adapted the halo exchange skeleton program to the single sweep method, and compared the communication times between the two sweep method and the single sweep method. The result (see Fig. 1) for the halo exchange with a halo of 20 grid points show that the difference between two methods is minor. Thus, we believe the analyses in our paper are also held for unstructured grids.

4) Line 603, whole paragraph: Can MPI tasks be carefully pinned to cores using knowledge of the domain decomposition to reduce congestion?

The domain decomposition, the topology of the interconnet network, and the communication pattern all have an impact on the congestion. The halo exchange usually has a local communication pattern; thus, with the knowledges of the domain decomposition and the underlying topolopy, it is possible to pin MPI tasks to cores so that each process exchanges messages with the near processes to reduce congestion; for example, a regular 2D/3D domain decomposition is mapped to a 2D/3D torus network. The transpositon and allreduce operation are all-to-all communications, it is not easy, if not possible, to map MPI tasks to cores to reduce the congestion.

5) Line 639: "However, the bandwidth of memory limits the performance and scalability of computation for multi-core or many-core systems". This statement seems taken without reasoning. Surely this cannot apply to any algorithm. Could the authors elabo-

rate?

Our intention of using the statement is to elicit the last sentence in the manuscript. Because this paper investigated the communication of atmospheric models using a single-core CPU per node, a singe-core CPU per node is good to assess the communication. But the architectures of current and future supercomputers are multi-core and multi-socket nodes, even non-CPU architectures. Because multi-core or many-core processors share a memory bus, it is possible for a memory-intensive application (such as an atmospheric model) to saturate the memory bus and result in degraded performances of all the computations running on that processor. Thus, our subsequent study will focus on the assessment of computations. We have clarified the last two sentences, please refer to the revised manuscript.

6) Acknowledgements: The Horizon 202 program ESCAPE acknowledgement has more strict rules on how to acknowledge (e.g. mention of EU and program number). I recommend asking the project manager for details.

Thank you very much for pointing out this to us. We have updated the acknowledgement to conform to the rules of the Horizon 2020 program ESCAPE.

In addiction, all the other sugestions presented in your attached PDF have been incorporated into the revised manuscript.

Please also note the supplement to this comment:
https://www.geosci-model-dev-discuss.net/gmd-2017-301/gmd-2017-301-AC3-supplement.pdf

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

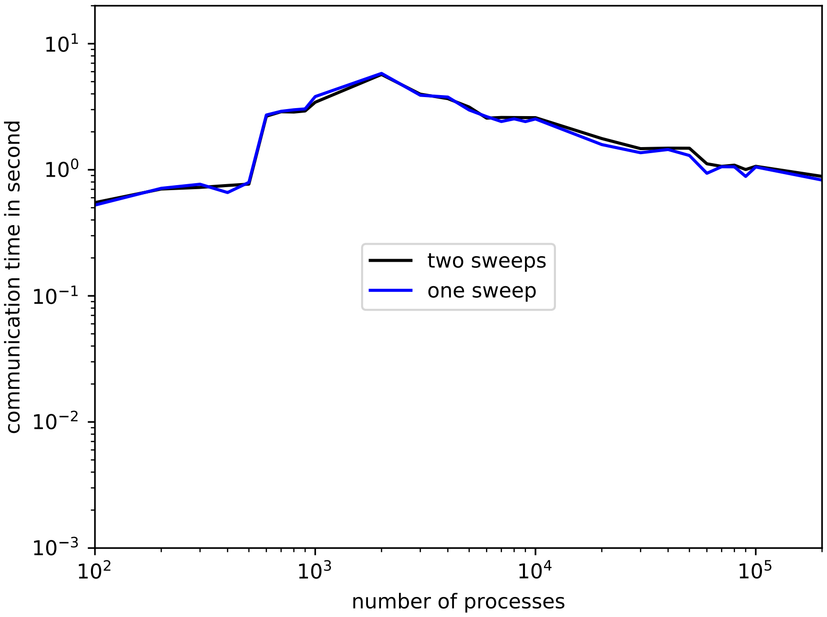

Fig. 1. Communication times of wide halo exchange between the two sweep method and the single sweep method