Geoscientific
Model Development
Discussions

Interactive comment on "A new distributed algorithm for routing network generation in model coupling and its evaluation based on C-Coupler2" by Hao Yu et al.

Moritz Hanke (Referee)

hanke@dkrz.de

Received and published: 28 April 2020

Dear Mr. Li Liu,

I am happy to engage in a discussion about the manuscript and hope you do not mind if move the discussion of topics not directly linked to the manuscript (incl. comments specific to YAC) to a personal email.

Regarding the meaning of "collective communication":
The MPI Standard defines this as: "communication that involves a group or groups of processes" [1]. Your manuscript did not explicitly limit the meaning to the subgroup

of collective algorithms like gather or broadcast. If limited your definition of "collective communication" to these operations, then your sorting algorithm has indeed a better complexity. However, the common meaning of "collective communication" also includes operations like reduction, barrier or alltoall.

On the sorting algorithm:
I may have been a little bit hasty regarding my comments concerning the common knowledge about your sorting algorithm. After some research, I could not find any paper describing an algorithm similar to yours. However, I also looked a little bit more into your algorithm. In the manuscript it is described that in each iteration all data of the local table is exchanged with the respective communication partner. This also matches with the actual implementation, if I understood it correctly. In figure 2 however, only a one-directional data transfer is depicted, which is actually sufficient for the algorithm to work correctly. In addition, instead of sending all entries in the local table, you have enough information available to only send the data that is actually required by the receiving process. If you apply these two optimisations, the movement of the table entries between the processes is more or less identical to that of a collective alltoallv, if implemented using Brook's algorithm [2][3].
Therefore, I would interpret your sorting algorithm as a modified alltoallv. And by the definition of "collective communication" in the MPI Standard, your algorithm itself is a collective operation.

Focus of the manuscript:
The title, the abstract and the remaining contents of the manuscript indicate for me a major focus on the introduction of a new algorithm and its performance. In that case, I still stand by previous assessment, that this is not enough contribution to modelling science and I would still not recommend this manuscript for publication.
Yet, I agree with you on your work potentially being beneficial to other software used in modelling science. Your manuscript could focus more on C-Coupler and its performance. After a more detailed analysis of the coupler initialisation, which identifies the

router initialisation as a significant issue, you could provide your algorithm as a solution for this. This might require some more performance data detailing the impact of router initialisation on the overall initialisation of the coupler. And as mentioned in my review, I would prefer a different presentation of the data shown in figure 3. Since, the manuscript also explicitly mentions the algorithm within MCT, you could also add a comparison with its initialisation.

With best regards,
Moritz Hanke

1: MPI: A Message-Passing Interface Standard (Version 3.1) Message Passing Interface Forum (2015) http://www.mpi-forum.org/docs/
2: Brooks, E.D. The butterfly barrier. Int J Parallel Prog 15, 295–307 (1986). https://doi.org/10.1007/BF01407877
3: Thakur, Rajeev. (2005). Optimization of Collective Communication Operations in MPICH. International Journal of High Performance Computing Applications. 19. 49-66. 10.1177/1094342005051521.