Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-257-RC2, 2020
© Author(s) 2020. This work is distributed under
the Creative Commons Attribution 4.0 License.

Geoscientific Model Development Discussions

Interactive comment on "Parallel I/O in FMS and MOM5" *by* Rui Yang et al.

Michael Kuhn (Referee)

michael.kuhn@informatik.uni-hamburg.de

Received and published: 2 March 2020

The authors present a detailed study of implementing parallel I/O using NetCDF in the Modular Ocean Model version 5 via the Flexible Modelling System. Even though the implementation is quite specific to MOM5, the paper can serve as a useful experience for developers aiming to implement parallel I/O within other scientific software packages. Overall, I believe the paper is worth publishing, especially since I/O aspects are often neglected. There are still some points for improvement, though.

Specific comments:

- Lines 36-38: Where is the number of 350 MB/s for disk throughput coming from? The HDDs I know about typically max out at roughly 200 MB/s. While I understand the point you are trying to make with these sentences, I believe some more details would make them easier to follow. How long does a one-year simulation typically take? Is writing

out one terabyte of data even relevant in this case?

- Lines 87-96: Please elaborate why you have selected NetCDF for your parallelization efforts. There are also other approaches such as SIONlib or ADIOS. While NetCDF probably makes the most sense for geoscientific applications, this should at least be discussed briefly.

- Lines 191-195: Have you considered the alignment of chunks? We have shown in "A Best Practice Analysis of HDF5 and NetCDF-4 Using Lustre (Bartz, Chasapis, Kuhn, Nerge, Ludwig)" that chunk alignment can have very significant impact on parallel I/O performance. Sadly, NetCDF did not (and apparently still does not) expose this functionality while HDF5 does. It is therefore necessary to patch NetCDF to enable HDF5's chunk alignment. Missing alignment could be the cause of contention you describe when increasing the number of I/O PEs per I/O domain.

- Lines 295-299: See previous comment, this could also be caused by missing alignment.

- Lines 451-453: The serial I/O versions with 720 PEs ran for 6 hours while the ones with 1440 PEs were killed after 5 hours. Did the 720 PE version run on a different partition? If so, is it still possible to compare the two?

- Lines 508-512: Why did you develop your own I/O profiling tool? There are existing options such as Score-P or Darshan. Please state why the existing tools did not meet your requirements.

- Line 526: I gave the GitHub repository a quick look but could only find the source code. According to GMD's code and data policy, the data must also be provided. You have also not mentioned in the paper which commit you were using to perform the model runs.

Technical corrections:

- Line 28: The acronym OS has been introduced before in line 23 and does not need

to be repeated here.

- Lines 62-70: Since you talk about "single file I/O" in the paragraph before, it might be worth mentioning explicitly that one file is created per I/O domain in this case.

- Line 73: "A typical 0.25° global simulations ..." - It should be "simulation".

- Line 183: "... in Table 3, ..." - This should be "Table 2".

- Line 227: "... of the I/O parameters in Table 3." - Should be "Table 2".

- Line 238: "... grids are disturbed over ..." - This should probably be "distributed".

- Line 375: "... in the charts below for each library." - This should rather reference the figures directly since they are placed in the appendix.

- Line 429: "... in Figure 14." - Figure 14 seems to be rather blurry while the others are fine. Please provide a high-resolution version if possible.

- Lines 581-622: Are the reported values averages? If so, you should mention this somewhere and also give deviations. Figure 14 already includes them but the others do not.

- Lines 625-639: Bright orange is hard to read on white, so it might make sense to change the color for the profiling graphs.

- Line 665: "Number of Output File" - This should be "Files".

- Lines 680-685: To better assess the scaling behavior, please also mention the number of nodes in addition to the number of PEs.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-257, 2019.