Geosci. Model Dev. Discuss., 8, C4128–C4131, 2016
www.geosci-model-dev-discuss.net/8/C4128/2016/

Geoscientific Model Development Discussions

Interactive comment on "Performance and results of the high-resolution biogeochemical model PELAGOS025 within NEMO" *by* I. Epicoco et al.

A. Porter (Referee)

andrew.porter@stfc.ac.uk

Received and published: 7 March 2016

General Comments ————————-

This paper presents a first analysis of the computational performance of a high-resolution biogeochemical model. The model itself is constructed by coupling the NEMO oceanographic model and the BFM biogeochemical model. Such coupled models are very important in developing our understanding of the Earth System and therefore their ability to make use of the computer architectures appearing on the road to Exascale is of great interest. This paper analyses the performance of the model on two very different architectures (an IBM BlueGene and an Intel-based cluster) and highlights the areas of the model that require attention (optimisation or re-engineering). I

found the paper interesting and informative. It provides a good discussion of the features of the two models and is clear about the model configuration and experiments performed. The paper clearly highlights the issues in balancing the demands of what are two quite dissimilar models and suggests some possible strategies in order to improve performance.

Specific Comments ————————-

The Introduction is good (sections 1 and 2). I think a bit more detail on the mechanics of the coupling scheme would be useful (accepting that full detail is available in the references). Am I correct in understanding that the BFM and NEMO models exchange data every time-step? Do separate processes execute the BFM and NEMO components or does every PE do both components for its subdomain? Please clarify.

I like the description of the Performance analysis. However, I don't agree with the assertion on lines 252-261 that Figures 2, 3 and 4 show that the MPI communication time is decreasing - those plots show only metrics derived from overall execution time. In fact, it would be nice to see an analysis of the time spent in MPI communication vs the time spent in the rest of the code. Having said that, I agree with the argument that the MPI communication is decreasing - I'd just like to see some evidence. Does the BFM component contribute to the MPI communications or are they only required to support the stencils in the NEMO component?

In section 4 (Code Profiling) it would be nice to have more detail on the % of run-time used by each of the significant routines. Are the entries in Table 3 ordered (e.g. by % of run-time)? A discussion of the differences (if any) of the profiles from the two machines would also be interesting and may highlight opportunities for optimisation. This is a point you touch on later but it would be good to see more detail here.

I like the discussion of memory structures in section 5. It is worth emphasising here that NEMO is designed for vector processors and is therefore good for the increasingly wide SIMD instructions that are appearing. How well does BFM SIMD vectorise? Since

BFM has zero-dimensional state at each ocean point, presumably it would only need a 2D data structure rather than the 3d one of NEMO (i.e. it has no need of depth information) in order to map onto the topology required by the numerical domain? Would this structure benefit the optimisation of the numerical components of BFM? e.g. remove indirection, promote SIMD vectorisation etc? How much does the load imbalance present in BFM (because of non-ocean points) impact the model as a whole?

I think Section 6 is the weakest in the paper and am unsure of its value since really the key issue is of load balancing performance and thus the ocean points and that is covered in Section 5. Memory consumption is just one symptom of this. However, the data in Figure 11 are intriguing - it seems to indicate that PEs fall into one of three categories and the variation in memory requirements between these is substantial, even for PEs that have very similar numbers of ocean points. I think this deserves further investigation.

Technical Corrections

295: describe choice of ordering of entries in the table. Were the key routines the same on the two architectures? 305: add the core count information to the caption of Figure 5. 315: suggest replace "are unaffected by scaling" with "do not scale at all". 316 replace "the results got by" with "the results obtained by" 327 is that "theoretical" peak performance? What is the theoretical peak performance of a single core? 338 replace "routines" with "routine" 345 I'm not sure that 'random' is right - there must be some deterministic relationship. 348 "With this architecture, there are more routines with a speedup value far..."

Section 5.2: at the risk of self-promotion there is this report http://purl.org/net/epubs/work/63488 which discusses changing the domain-decomposition strategy of NEMO in order to load-balance the number of ocean points between processors.

473: replace "sensible" with "sensitive" (I think "sensible" is strictly correct but "sen-

sitive" is more readily understandable). 528-530: this is true if the BFM and NEMO models couple every time-step. A comparison of PEs with large and small memory footprints would be interesting.

Conclusions, $\sim$640. Although a smarter domain decomposition might help at lower process counts, I think Figure 8 shows that by the time you get to 1024 PEs the difference is negligible. Is that right?

Notes on Tables and Figures.

As already mentioned, Table 3 does not specify how the routines are ordered. Are they in order of significance (in terms of time taken)? Table 4: is the "total elapsed time" fully inclusive of start-up costs etc.? If so, some justification that they are negligible is required. Figures 2, 3 and 4: could be improved by more closely restricting the range of the x-axis to match the range of the data. The data being plotted is not continuous and therefore there need to be points as well as lines on the plots. Figures 5 and 6: I think these plots would be much better as bar charts. The shadows on the points made me think there were two data values at each ordinand at first. As mentioned earlier, please specify the PE counts used in the captions of these figures. Figure 9 came out as grey scale for me. If it really is grey scale I suggest colour would really help pick out the differences. Possibly a histogram of number of the number of ocean points on each PE would be useful. Figure 11 - do you have any explanation for the $\sim$trimodal behaviour seen in this plot? e.g. at very low numbers of Ocean Points some PEs have 0.5 GB allocated while others have 0.7 GB - nearly 50% more. There are yet others with values $\sim$mid-way between these two extremes. Figure 13 - again data points as well as lines are required on the plot and it would be nice to limit the range of the x-axis a bit more.

Interactive comment on Geosci. Model Dev. Discuss., 8, 10585, 2015.