Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-114-RC2, 2016

Geoscientific
Model Development
Discussions
Open Access

Interactive comment on “Coarse-grained component concurrency in Earth System modeling” *by* V. Balaji et al.

Anonymous Referee #2

Received and published: 1 July 2016

Coarse-grained component concurrency in Earth System modeling V. Balaji, R. Benson, B. Wyman, and I. Held

General Comments

The paper describes an approach to exploit a previously unexploited level of concurrency from multi-component, coupled Earth System Models (ESMs). The work is motivated by a discussion of the trends in the development of extreme-scale high performance computing where future performance gains require the exploitation of vastly higher amounts of concurrency than previously required.

The motivation is very well argued with respect to the need for new techniques to exploit more concurrency in such models and from the algorithmic perspective. For example, in the specific case of attempting to execute the radiation model concurrently

with the rest of the atmosphere model, the resulting concurrent algorithm is known to be theoretically unstable but, nevertheless, the implementation is found to be stable in practice due to other physical processes.

The implementation of the new concurrent algorithm exploits the shared memory nature of modern multi-core CPUs in order to minimise the impact on data movement between the component models. This is a crucial aspect, since a traditional approach to exploiting model concurrency - running two models on separate sets of processes and exchanging data through a message passing-based model coupler - would lead to excessive data movement between processor cores that would severely impact performance.

The paper successfully demonstrates that there is much more concurrency in ESMs than has previously been exploited and that there are techniques that can exploit this concurrency effectively on the emerging extreme-scale HPC systems. The new algorithm allows the (expensive) radiation model to run more frequently (i.e with a smaller timestep, hence coupling more frequently) while maintaining the overall simulation rate while using more processor cores. This is a very encouraging story on the road to exascale for Earth System Modelling.

Specific Comments

page 2, around line 20: the discussion focuses on performance aspects only. Mention of the implications for energy/power use would also be useful here, as the focus is on extreme-scale systems.

page 3, line 12: "how it is achieved without increasing data movement" should, I think, be changed to "and how it is achieved with minimal impact on data movement". This point is discussed further below [*].

page 5, Figure 1: The role of A_t should be depicted in the top figure (to be consistent with equations 1 and 2), I feel. Also, in many models, the atmosphere is executed on more processors than the ocean (because it scales better). Is this diagram consistent in this respect with the FMS model being described?

Also, the bottom figure in Figure 1 implies that the ocean is executed on fewer processors in the concurrent set up. Is the intention to simply show a deployment utilising the same number of processors in total? If so, that should be made clear in the caption and text.

[*] The following point is my main concern with the paper as it stands.

I believe the paper would benefit by being clearer on the use, and limitations, of shared memory threading to implement the concurrent execution of the radiation and the rest of the atmosphere model.

It is clear that for a given multi-core processor there will be limits on the number of MPI processes per node and the number of threads each model may use within an MPI process without incurring potentially expensive data movement between caches. The example results given are based on two threads for each model. It would be good to make clear the rationale for this choice. Here are some thoughts on this and suggestions for possible changes which might help achieve this.

Currently, the use of threads for executing the radiation in parallel with the rest of the atmosphere is described as not incurring any communication costs. While it is true there will be no MPI communication incurred between cores running the atmosphere and cores running the radiation, I believe there may well be some extra remote data accesses (i.e. cache misses) incurred between cores running the atmosphere and cores running radiation. The magnitude of this effect is, of course, architecture dependent and also depends on the number of threads used for each model and the mapping of the threads to cores.

The results presented are for AMD Interlagos processors with two threads used for the atmosphere model and two threads used for the radiation model (within each MPI

process). The Interlagos processor chip consists of eight 2-core modules. Two threads executing on the same module share an L2 cache. All 2 core modules share a large L3 cache. So, if one atmosphere and one radiation thread share a module, they can share data in the L2 cache (as well as the L3 cache). If two Atmosphere threads share a module and two radiation threads share a module, they will communicate through the L3 cache, which is more expensive in terms of cycles to access. If threads are on separate processor chips, there will be (even more expensive) data movement within the shared memory node.

The cache behaviour in either of the above cases is likely to be different to that of a single thread running first the atmosphere and then the radiation.

If more than two threads were used for each model, some sharing would have to take place via the L3 cache. Total thread numbers are clearly limited by the core count of a shared memory node.

I would suggest to the authors that some clarification of these issues be made. For example:

- in Figures 3 and 4, the images depicting MPI and OpenMP could be re-drawn to illustrate the relationship of threads within MPI in each case. In Figure 3, this would simply show multiple threads in an MPI task and multiple MPI tasks. In Figure 4, MPI tasks could be shown with both atmosphere and radiation threads or with ocean threads. For Figure 4, this might be something like:

[AARR] [AARR]... [O] [O]...

(where [] here represent MPI processes and letters represent threads and their models)

- In Section 3 (perhaps?), a brief description of a multi-core processor (like the Interlagos) could be given along with the implications of thread-to-core mapping. This description would help to explain the benefit of using OpenMP to exploit parallelism between the atmosphere and radiation models (i.e. no MPI communication) and pave the way for a discussion of the potential for sharing data between caches in the specific configurations presented in the results section.

- page 9, line 2: "chosen to offer optimal load balance" could be extended to "chosen to offer optimal load balance and data sharing", for example. [I generally have a concern over the use of the work "optimal", which has a formal sense of "provably best". The word "good" might be better unless the load balance is provably optimal?]

- page 11, line 9: The above arguments are tied up with the statement that "All runs use the optimal processor/thread layout for a given PE count". Some explanation about what this layout is and how it was chosen could be added.

Technical Corrections

page 1, line 4: I suggest changing "based on marginal increases in clock speed" to, for example, "based on, at best, marginal increases in clock speed" since it is likely clock speeds may decrease in future in some systems.

page 1, line 14: Define the acronym CCC here.

page 1, line 15: is a little ambiguous about what is running in parallel ("and all other atmospheric physics components". I would suggest making it clear that there are only two concurrent components (i.e. not all "other atmospheric components" are executed in parallel with each other!

page 2, line 3: perhaps provide a reference to the IPPC assessments.

Section 2:

page 4, line 4: needs a closing bracket after "example".

page 8, line 24: "Individual" should be "individual".

page 9, Figure 4: In this figure, the Land and Ice models are shown as executing concurrently but this is not mentioned in the text. This should be explained (or made

consistent with Figure 3).

page 10, line 9: "that that" should be "that".

page 10, line 11: This sentence would benefit from having a reference added.

page 10, line 13: Expand the acronym GPCP here as a definition.

page 10, lines 15-17: The point here is, I think, that this result is counter intuitive. If that is correct, it would be worth stating.

page 11, line 3: "less expensive as..." should be "less expensive as the...".

page 12, line 4: the figures for processor count and SYPD given in this line are rounded versions of those in Table 1. Those in the previous sentence are not rounded. Please use the precise figures for consistency.

page 13, line 18: It would be worth giving the definition of CCC again here to remind the reader.

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-114, 2016.