Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-193-AC3, 2016

Geoscientific Model Development Discussions

Interactive comment on "Enabling BOINC in Infrastructure as a Service Cloud Systems" *by* Diego Montes et al.

Diego Montes et al.

kabute@uvigo.es

Received and published: 23 December 2016

We would like to thank the reviewer for the time and the effort to review our work and for the good and constructive comments. About the questions:

Generating data is only one part of the story. Data needs to be stored and analysed. The costs of storage in the cloud can be substantial and the computational require- ments for data analytics are very different to those needed to generate data in the first place, particularly if the data is to be open to a large number of scientists/data analysts

We totally agree with your comments but we would like to note that the impact of data transfer and storage (for us mainly S3) is minimal (and academic institutions can benefit

from waivers) Obviously this is completely dependent on the duration that you wish to store the complete raw results. Also within the scope of the submitted paper we are considering the execution of GCM type climate models and not the analysis of their outputs as the methods chosen are very different within different groups with no single clear method appearing as the community standard as yet. Anyway, we include more specific costs on the final version of the document.

Although I fully accept that this paper describes a proof of concept, the models run in this experiment are, relatively speaking, low resolution. There will be different challenges using higher resolution or more complex models. This could be related to the information shown in Figure 1 for run-time as it is surprising that smaller instances perform better than bigger instances. A better discussion of why this is the case would be very useful for other researchers as it has very clear financial implications.

We agree that there will indeed be different challenges for more complex modes. It is clear that models that require many processors to run will not benefit from the use of smaller instance types. However, the point we are making to highlight the need to accuratly benchmarks on the application used to ensure that you are maximising the value of the computational resources that you are intending to use. The better performance with smaller instance is due to the fact that vCPUs are hyperthreads and in smaller instance types there is greater chance the CPU is running at a lower utilization and our instances can scavenge extra CPU cycles. This information is now included in the text (also see Uhe et al., Utilising Amazon Web Services to provide an on demand urgent computing facility for climateprediction.net, Proceedings of the 2016 IEEE 12th International Conference on e-Science).

Specific points: - Page 2, line 1: "the number of members in each ensemble tends to be small due to computational constrains" The use of computational resources in climate modelling is a balancing act between resolution, complexity and ensemble members as the authors point out a few lines earlier. It is not that the number of ensemble members is small due to computational constrains, it is often a conscious choice made

by researchers. Having a big ensemble is desirable but it only makes sense if your model captures the right processed"

The authors agree with this point and it is clear that the models resolution and the size of the ensemble is completely dependent on the experiments that the submitter is considering to analyse. It is though a common complaint of researchers that they cannot access the scale of resources desirable to create ensembles numbers large enough due to lack of resources or having to share them with other members of the community in larger national scale systems.

Re "only makes sense if your model captures the right processes that you are interested in" It should also be noted that if the model used doesn't capture the climate process necessary to be analysed for the application of interest then the researchers should be questioning why they are using that model in the first place. This is also again out of scope for the study being published here.

Page 6, section 3.2.2 How could the data in S3 be used by other applications beyond climateprediction.net? How would this cope with much bigger datasets of hundreds of terabytes or petabytes?

S3 can store files up to 5TB (as described on https://aws.amazon.com/s3/faqs/) so if datasets are larger than that another solutions should be explored (like a CephFS cluster, which is compatible with S3 via a Gateway API). With S3 share can be simply done by using the built-in web server and access policies. This information is now included in the text.

A few typos have been detected...

Thanks, indicated typos have been fixed.