Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-230-RC2, 2017

Geoscientific Model Development Discussions

Interactive comment on "Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4888 GPUs with COSMO 5.0" *by* Oliver Fuhrer et al.

R.W. Ford (Referee)

rupert.ford@stfc.ac.uk

Received and published: 22 December 2017

This paper is well written and the authors are world leaders in their ability to run climate models on GPU's (due to their DSL approach) - they are the only ones currently doing so for a production model.

The main theme of this paper is to examine how well the DSL version of the COSMO model runs on the largest GPU supercomputer in Europe and compare this to what will be required for 1km production runs.

The paper makes a number of important contributions

Printer-friendly version

Discussion paper

1) it demonstrates that production climate models can run on large GPU machines (this has not been done before) and presents performance results 2) it shows the crossover between GPU and CPU performance for lower resolution runs 3) it shows what can be run on todays supercomputers (certain AMIP experiments could run at 2km resolution) and how much more performance will be required to run simulations at 1k resolution (which is considered to be one of the major challenges when running on next generation supercomputers as it allows convection to be resolved rather than being parameterised which results in much improved simulations) 4) it compares their performance (SYPD) and energy use with earlier work and presents improved results thereby laying down a performance baseline that others can compare their models to 5) it introduces and discusses a new performance metric 6) in the appendix it outlines a new approach for minimising data movement in a CDAG allowing complex models such as COSMO to be analysed and optimised.

I have a couple of general suggestions

1) the simulation is "near global" as COSMO is a regional model with a regular structured 3d grid. Global climate models do not scale if they implement regular structured 3d grids, both numerically and computationally, due to the "pole problem". Therefore scalable global models tend to use quasi-uniform grids with potential associated performance and algorithmic issues. It would be good (at least for me) if you could outline how these approaches relate and therefore give a feel of what global models can learn from this work. I guess P8 L10-15 would be the appropriate place. I looked at the reference at it did not give me this information.

2) I think the Memory Usage Efficiency metric is interesting but it raises many questions. As such I think it is a little early to suggest its use as a metric by the community. I would like to see a separate paper (or papers) just on this metric, with a number of case studies, comparisons etc. before its use is advocated. I think it would be worth toning down the claims for the metric in this paper and limit it to saying that it was useful for this analysis and that it will be further analysed in future work.

For the authors information, specific questions on MUE relating to point 2 that I thought when reading section 4 are below. I am not expecting you to answer them all; as said above I think it would be better to say that future work will study the metric in more detail.

a) The statement that algorithmic optimisation has to focus on minimising data movement is not the whole story. It is certainly the main factor for memory bound applications, which this application is, but other applications may have a higher ratio of operations to memory accesses and other opimisations may also be important e.g. improving load balance. Could you please modify this section slightly, perhaps just to say that "algorithmic optimization is having to increasingly focus". I note that the appendix does relate to memory bound climate models but the reader is not likely to know this.

b) It is not immediately clear whether internode comms is optimised here. I presume not but it should be mentioned somewhere.

c) Can MUE be used on memory bound codes that do not use STELLA and therefore do not have a CDAG?

d) is the metric expected to be useful for CPU's as well given their complex memory hierarchies?

e) If I understand correctly then you determine a single achievable bandwidth value based on micro-benchmarks. However, a program might have different phases with different access patterns potentially leading to different achievable bandwidths e.g. ECMWF's IFS.

f) how does MUE relate to roofline (and other metrics)

g) is MUE really two metrics (ratios) rather than one?

I also have a number of minor suggestions for changes mostly to do with improving the readability of the paper, which are given below

P1 Title: 4’888 -> 4,888

P1 L1: biases, is through -> biases is by

P1 L8: propose the new -> propose a new

P1 L15: Currently available -> Current

P2 L6: must thus be -> must be

P2 L7: to kilometer-scale resolution -> to the kilometer scale

P2 L7: allow to explicitly represent -> allow the explicit representation of

P2 L9: around this -> at this

P2 L11: US -> United States

P2 L14: appalling -> significant

P2 L15: Indeed, increasing -> Increasing

P2 L15: factor 25ˆ3 = 15000 -> factor of 25ˆ3 = 15,000

P2 L17: of the simulation -> of a simulation

P2 L18: what is a hydrometeros?

P3 L2: the large legacy simulation -> their large existing

P3 L3: HPC -> High Performance Computing (HPC)

P3 L4: a concern over peak -> peak

P3 L4: towards a concern over improving -> towards improving

P3 L6: 1000 -> 1,000

P3 L7: most of this -> much of its

P3 L9: I’m not familiar with the term "time-compression"

P3 L11: Use the footnote on the second reference to AMIP here.

P3 L12: 1979-2014 driven -> 1979-2014, driven

P3 L13: the framework of -> the context of ???

P3 L14: larger than about -> greater or equal to

P3 L14: 0.3, with which the -> 0.3. At such a rate the

P3 L23: flop/s -> floating point operations per second (flop/s)

P3 L32: put the AMIP footnote on the first reference to AMIP (line 11)

P3 L32: simluations -> simulations

P4 L12: please explain the acronym GPU, as this is the first use (apart from the abstract)

P4 L13: please add a footnote explaining Tflop/s

P4 L14: please add a footnote explaining Pflop/s

P4 L22 1500 km -> 1,500 km

P4 L32: grid spacing. -> grid spacing, respectively.

P5 L1: Such a large -> In their case, such a large

P5 L13: and reducing -> and additionally would require reducing

P5 L14: introduce IFS

P5 L19: If ECMWF used 120s then please explain how you conclude 40-60s.

P5 L24: Unlike with -> In contrast to

P5 L26: Please provide a footnote explaining weak scaling.

P5 L26: applies for the -> applies to the

P6 L8/9: remove additionally

P6 L10: is -> are

P7 L2: vertical a -> vertical, a

P7 L3: allows prescribing -> supports an

P7 L4: neighbor -> neighboring

P7 L6: solvers -> solves

P7 L8: encompasses -> includes

P7 L9: algorithms -> algorithm ???

P7 L11: To enable COSMO -> To enable the running of COSMO

P7 L16: Additionally the -> Additionally, the

P7 L22: 5000 -> 5,000

P7 L23: remove "available"

P7 L27: writing this -> writing, this

P7 L27: 1431 -> 1,431

P7 L30: explain that GB is GBytes, where G is 10ˆ9, as it has not been introduced

P7 L31 fabric based on Aries technology in -> fabric (based on Aries technology) in

P7 L31 please reference, or explain Aries unless it is explained in the Alverson reference.

P8 L3: The total energy -> The total energy (which includes the interconnect)

P8 L14: by then winner -> by the winner

P9 L1: regime accompanied -> regime, accompanied

P9 L4-5: On this balanced initial state, a large-scale local Gaussian perturbation is applied, which -> A large-scale local Gaussian perturbation is then applied to this balanced initial state which

P9 L8: 36000 x 16001 -> 36,000 x 16,001

P9 L9: In zonal -> In the zonal

P9 L14: bracket should start before Jablonowski, not 2006

P10 L12: In the meantime, -> However,

P10 L12-13: changed entirely. -> changed.

P10 L13: are cheap. -> are now relatively cheap.

P10 L13: 1000 -> 1,000

P10 L16: movements -> movement

P10 L22: would -> could ... I don't think it is guaranteed to deceive.

P10 L31: You first introduce the concept of data transfers here. Could you please define what you mean.

P11 L2: stencils of these computations in -> stencils in

P11 L4: to perform only necessary transactions at -> and perform them at the

P11 L14: represents the -> represent the

P11 L16: estimates -> estimate

P11 L18: model assess the -> model determines the

P11 L20: former one -> former

P11 L21: application, the -> application and the

P11 L24: ist he -> is the

P12 L2: The GPU version of COSMO was so far only scaled up to 1000 -> Until this study the GPU version of COSMO had only been scaled up to 1,000

P12 L3: 4932 -> 4,932

P12 L5: Please explain that adding cloud microphysics is not expected to affect strong or weak scalability.

P12 L7: 4888 -> 4,888

P12 L10 next-neighbor -> nearest-neighbor

P13 L1: this is the first time mixed precision for this model is introduced. Some explanation is needed.

P13 Figure 6: Perhaps add a marker showing where 200 x 200 is on the 19km and 3.7km results to help the reader.

P13 L2: 1000 -> 1,000

P13 L6: 1000 -> 1,000

P13 L6: is there any benefit in using OpenMP, either to reduce comms, or for hyper-threading?

P14 L2: 4888 -> 4,888

P14 L3: 12h -> 12 h

P14 L4: 1000 -> 1,000

P14 L9: perhaps recap that 0.2-0.3 sypd is the minimum feasible value to remind the reader before concluding.

P14 L12: In the following, we -> In the remainder of this section we

P14 L12: bracket before Yang, not before 2016

P14 L15: As argued -> As explained

P14 L16: admissible -> feasible

P14 L17: factor 3 to -> factor of 3

P14 L19: factor 1.5 -> factor of 1.5

P14 L20: factor 2.5 -> factor of 2.5

P14 L22: planet, and -> planet and

P14 L24: a factor 2 or more -> at least a factor of 2

P14 L26: cf. -> see

P14 L26: we provide -> we now provide

P14 L27: 4888 -> 4,888

P14 L28 MWh/SY while -> MWh/SY, while

P14 L28: 4888 -> 4,888

P14 L28: comparison the -> comparison, the

P15 L2: imply -> resulting in

P15 L3: 6500 -> 6,500

P15 L5: 2052 -> 2,052

P15 L6: 1059.7 -> 1,059.7

P15 L7: 15371 -> 15,371

P15 L4-9: Please explain this argument in more detail. It is unclear to me what the logic is to get to 5 times more power efficient.

P15 L13: Appendix -> Appendix A

P16 L3: Whilst the results are impressive, claiming 66% as being "close to optimal" is perhaps a little strong.

P16 L4: Because on -> On

P16 L4: unit, we -> unit. Therefore we

P16 L5: This -> This result

P16 L8: This conclusion is less obvious to me as you have made the assumption that the unoptimised version achieves peak bandwidth and that may not be the case.

P16 L10: With the work presented here, we are setting -> The work presented here sets

P16 L10: full-fledged -> fully-fledged

P16 L12: 4888 -> 4,888

P16 L19: 1000 -> 1,000

P16 L19: also add that the overall performance is better on CPU's than GPU's due to the strong scaling

P16 L20: by Jablonowski and Williamson (2006). -> in this paper.

P17 L2: does this count as extreme scale? - I would say not. Perhaps just repeat that it is the most powerful supercomputer in Europe at this time

P17 L2: scale running -> scale, running

P17 L2: 4888 -> 4,888

P17 L13: 4888 -> 4,888

P17 L14: Energy -> The energy

Printer-friendly version

Discussion paper

P17 L22: reach -> reach the

P17 L23: Given -> Given that

P17 L23: was -> is

P17 L25: as goal post -> as a goal post

P17 L26: deployed next -> deployed in the next

P19 L15: former one -> former approach

P19 L16: and not -> and is not

P19 L17: latter one -> latter approach

P19 L20: those -> these

P19 L24: bracket before Hong and remove before 1981

P20 L9: load partition -> load partitions

P20 L22: Register -> Registers

P21 Figure A1 caption except of -> except for

P21 L9: Table ??

P21 L10: Table ??

P22 L12: pre-dominantly -> predominantly

P24 L14: Bony et al. S. -> Bony, S. et al.

P26 L5: Lee et al., V. W. -> Lee, V. W. et al.

P26 L16: Michalakes et al., J. -> Michalakes, J. et al.

P27 L13: Yang et al., C. -> Yang, C. et al.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-230, 2017.