Next-Generation High Performance Computing for the Environmental Sciences: a Roadmap

Outcome Summary of ReCoVER Workshop

"Next Generation HPC Architectures for Studying Environmental Variability"

Held at University of Exeter, 20-22nd March 2017

9th October 2017

This document is an outcome summary of the two day workshop "Next Generation HPC Architectures for Studying Environmental Variability" held at the University of Exeter on 20-22nd March 2017, organized by Beth Wingate, Lawrence Mitchell and Chris Budd with the assistance of Peter Ashwin and Emily Paremain. Details are on: https://sites.google.com/site/recoverlwec/activities-and-events/hpc-workshop

The workshop was supported by the EPSRC grant EP/M008495/1, *Research in Changes of Variability and Environmental Risk* (ReCoVER, http://www.recoverlwec.org).

Introduction

At the time of writing, next-generation high performance computer (HPC) architectures are expected to have 100-200 million way parallelism in the next 3-5 years. The purpose of the meeting was to explore and reimagine the scientific landscape to take advantage of this for the environmental sciences in general and climate/weather/earth system modelling in particular.

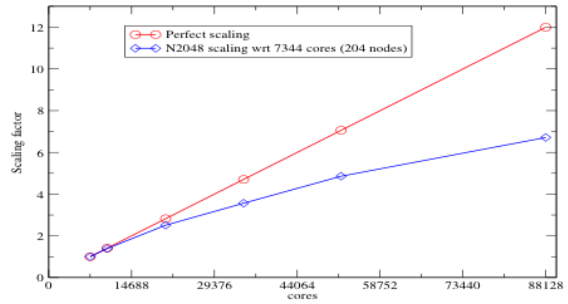

Figure 1: UK Met Office paper titled "Viability of 2020-era NWP resolutions with the ENDGame dynamical core" by Andy Malcom and Paul Selwood. We quote: "The fact that the core model has been able to run on approaching 90, 0000 cores without reaching scaling limits is very encouraging. However, the best runtime achieved at this scale is still 40% over that achieved by current resolutions and resources. This suggests that an N2048 model is unlikely to be achievable operationally with today's model on today's computer architectures, no matter how large a system is used."

The workshop was called to address a fundamental challenge: most parallelization strategies used for state-of-the-art simulation models will **not** be able to utilize this increase in power - and in particular translate increases in power to increases in grid resolution - because of the following issues:

- Increases in grid resolution require corresponding decreases in time-step (this is due to the mathematical structure of the equations). Consequently, more time-steps are required, the computational cost of which cannot be amortised by parallelism as the mathematical methods are not parallel in time.
- The approach of the strong scaling limit. For a fixed problem size (i.e. number of degrees of freedom) increasing the number of processors produces diminishing returns as the balance between computation and communication changes until the latter dominates and reducing simulation run-time is no longer possible.

Therefore, simulation models will have to be retooled to improve both numerical methods and compute/software infrastructure to make use of more parallelism. As an example, we note Figure 1 illustrates the drop off in effectiveness of parallelization that we currently face. This is a problem not only for earth system simulations, but for many applications in physics and scientific computing.

Executive Summary

We present a roadmap of how to make best use of next-generation HPC in the UK for environmental modelling (in particular climate, weather, and earth system modelling). We identify three outcomes of the discussion at the workshop that should help to address these challenges:

1) UK researchers in industry and academia are in a good position to lead an international conversation about developing HPC in environmental sciences at this key moment in time.

2) Much of this knowledge and expertise lies across traditional boundaries, skillsets and organizations, and we recommend a number of actions can be taken to enable and promote these developments.

3) We identify some critical skills gaps and training gaps in developing a workforce with the right combination of skills to prepare for the future challenges that HPC computing in this area will face. This will bring additional benefits to the use of next-generation HPC for a wide variety of scientific computing challenges.

Our suggestions align with the goals of the EPSRC 2016-2020 Delivery Plan that highlights computing as an important facet for 3 out of its 4 priorities: namely productivity, connectedness, and resilience. It also ties in well with goals of the other research councils, in particular NERC's vision as outlined in 'The Business of the Environment' and STFC's goals recommended in 'Computing Strategic Review 2015', and is core to the development of the science at the Met Office and ECMWF. We suggest that it is important that the research councils and a wide range of institutions should work together across traditional boundaries and is an opportunity for the RCUK.

We outline two recommendations to enable the developments:

Recommendation 1. Develop activities in a number of areas as a way to meet the challenges:

1) Exploit domain specific information to augment computational performance, by developing tools that automatically provide architecture specific optimisations. These tools can be considered as 'fit for purpose' intermediate steps for compiling codes on different architectures.
2) Develop new numerical algorithms and use new modes of numerical analysis (eg performance modelling) for increasing concurrency or moving beyond the timestep constraints for new and current algorithms.
3) Create new statistical and machine learning methods to take advantage of the Era of Ensemble and Data Science.

Recommendation 2. Coordinate these lines of activity so they can move forward together.

Note that the benefits of taking these steps will be not only to Environmental science but more generally to parallel modelling of continuum systems, for example in biomedical technologies or materials science.

The HPC Roadmap

There were two main themes of discussion at the workshop:

1. What novel numerical algorithms might be developed (and applied to environmental modelling) that can more fully exploit emerging architectures. What challenges do these pose to numerical analysts and those using scientific computing? For example there was discussion of:
    a. Potential disruptive algorithms (eg that break out of time-step constraints).
    b. New paradigms and modes of scientific computing, such as using data science in a new way.
    c. New ways to optimize the use of computer resources (eg variable precision, performance modelling).
2. How do we make immediate best use of these new emerging architectures For example, there was discussion of:
    a. Developing the science of ensembles.
    b. Bringing statistics/machine learning methods closer to environmental simulations.
    c. New science questions that can be addressed that will have significant impact on UK and international science.

We make two recommendations as a result of the discussions.

Recommendation 1: Three key activities

1) ***Create fit for purpose tools and software stack*** (time scale: 5 years).

    Key challenges for effectively using new computer architectures are the rapid evolution and disparate nature of programming models. To maintain single source science code that can be optimized for different architectures requires an extra stage in compilation that can automatically target, for example, different programming models such as OpenMP or OpenACC. To make progress this HPC Roadmap calls for research and development of:

    a) Computing work-flow schedulers that can take programs in C, Fortran, etc, and map them onto specified hardware.
    b) Domain specific languages that can be used to capture application-specific knowledge/structure in the simulation code, which can be used by special-purpose compilers to produce optimised implementations for a variety of HPC platforms.
    c) Development of communication and functionality 'standards', 'units' and 'tests', comparable to standards that exist for telecommunications processes.

d) Develop models that help give better understanding of the costs of decreasing the time-to-solution and optimization of computations in terms of energy use.
e) Mathematical maps of performance models that map directly to different architecture types and include memory-bandwidth, hierarchical memory, interconnect, I/O. Mathematical analysis of these to deduce limitations and opportunities.
f) Connection with people studying algorithms and mathematical structure.

2) ***Create new numerical algorithms and numerical analysis for increasing concurrency or moving beyond current time-step limitations for new and current algorithms.*** (time-scale: 10 years)
Since at least the 1990s there have been enormous efforts in parallelization strategies for breaking spatial domains (e.g. rivers, tidal plains, the Earth) into subdomains. This strategy, called *domain decomposition*, has been successful at reducing time to solution but with models typically only decomposing in the horizontal dimension, other forms of parallelism are required. One of the primary reasons standing in the way of reduced time-to-solution is the well-known time-step limitation. To go beyond current methods and algorithms the HPC Roadmap calls for research and development of:

   a) New algorithms with potential to reduce the time-to-solution (e.g., exponential integrators, multistage, time-parallel, parallel-across-the-step, parallel-across-the-method, methods that go beyond the time-step limitations, probabilistic numerics, etc.)
   b) New mathematical analysis of the governing equations to find low frequencies that can be time-stepped with larger time-steps. This should collaborate with numerical analysis.
   c) Numerical analysis of new and current methods that lead to new insights into extending the time-step.
   d) Collaboration with Activities 1, 2 and domain scientists.

3) ***Create new statistical methods and data analytics to take advantage of the Era of Ensemble science and Data Analytics*** (time scale: 5 years)
One possible way to take immediate advantage of new computer architectures is to run many instantiations of current models at the same time. This would give scientists access to ensemble information that has never before been possible and bring on a new era of statistical or ensemble science. Under such conditions much will be learned about the notion of accuracy and physics models will be improved. There may be previously unanswered science questions due to uncertainties that can finally be addressed. The HPC Roadmap calls for research and development of:

   a) Collaboration with domain scientists to develop new science questions and new statistical methods to answer them using ensembles.
   b) Improving and advancing statistical models and theories.

c) HPC models with uncertainty quantification
d) Using ensembles to improve physical models.
e) Machine learning to improve data intensive and compute intensive simulations
f) Using statistics and machine learning to intelligently design and build ensembles.
g) Exploring the role of machine learning and data analytics. Are there new questions that can be answered, either for science or society?
h) Can we use machine learning with ensembles to improve the algorithms or find low-frequency information?
i) Any improvements need to be shared with Activity 1 and 2.

Recommendation 2: Promoting collaboration

Although all three of the activities can make progress by working alone, consciously connecting them could move the impact-time forward.

This is illustrated in the HPC Roadmap diagram, at right. The dashed lines indicate the three activities relative to their time-scales of impact on learning to use, and using, next generation architectures.

Other Recommendations

We list some detailed recommendations arising from the workshop, along with suggestions for some of the actors who need to be involved:

- UC= University researchers in computer science
- UM= University researchers in mathematics
- UP= University researchers in physical/ earth systems modelling
- R= Research Councils
- EI= Environmental Modelling Industry (eg Met Office)
- CI= HPC Industry (eg IBM)

- Need for national collaboration at the Tier 2 computer level, preparing the way for critical scientific runs on National and International-scale architectures. (UC,UM,UP,R,EI,CI)
- Need for international collaboration. Different computer architectures are developing groups of "early adopters". Because the UK will be experimenting with new algorithms and domain specific embedded compilers, we can play an international-leading role in the way machines develop in the future. This is essential for maximum return on investment. (UC,UM,UP,R,EI,CI)
- The UK already has world-leading researchers in this area. In particular, the UK has:
  1) strong and established collaborations between the Met Office and Universities supported by both NERC and EPSRC;

2) world-leading numerical analysis, statistics and mathematics related to planet Earth in the EPSRC remit;
3) EPSRC investment in Research Software Engineers (RSE) and next-generation computer architectures,
4) Strength in mathematics of algorithms and computation (Alan Turing Institute).
5) NERC and EPSRC support for domain specific language. For example FEniCS/Firedrake which already makes the bones of Met Office's next-generational dynamical core available to mathematics and computer science. For finite difference programming models, there is similarly significant expertise in the UK, e.g. OpenSBLI, OPESCI.

There is a great opportunity for the UK to maintain and build on these strengths, especially because other governments, such as USA and China have put significant resource into hardware. (UC,UM,UP,R,EI,CI)

- The subject matter funding for this is split between disciplines and councils. Some of the greatest challenges in terms of developing new algorithms and developing and advancing mathematics as well as developing RSEs modes of scientific computing that will face the HPC challenges. (UC,UM,UP,R)
- There are opportunities for improvements in HPC that will significantly benefit research funded by NERC, BBSRC and STFC but this should benefit from more buy-in from EPSRC-facing researchers. For the environment, another challenge will be for NERC and EPSRC to contribute its expertise to this highly multidisciplinary endeavour. This should be underpinned by development of a wider paradigm of performance modelling of computer algorithms, both theory and application. This is a great (UC,UM,UP,R)
- There is a clear "skills gap" of scientific computing experts who have adequate knowledge of the algorithms and the parallelization and of next generation HPC. This is not only important for the next decade and the available next-generation architectures, but because the HPC industry will continue to change as strategies for computing are examined for application until a new paradigm is reached. UK must participate in this international conversation to meet its goals in almost all of the RCUK. (UC,UM,UP,R)
- There is a need to develop innovative MSc/PhD Programmes in next-generation HPC for scientific computing and environmental modelling, embracing the full range of HPC applications (data intensive to compute intensive). There is a need to develop career paths for those at the interface of these academic disciplines to ensure a healthy development in this area. (UC,UM,UP)

Attendees

The workshop website: https://sites.google.com/site/recoverlwec/activities-and-events/hpc-workshop

Organizers:

Beth Wingate	University of Exeter
Chris Budd	University of Bath
Lawrence Mitchell	Imperial College London

Speakers:

Matthew Chantry	University of Oxford
Katherine Evans	Oakridge National Laboratory, US Department of Energy
Christopher Maynard	Met office
Lawrence Mitchell	Imperial College London
Willem Deconinck	ECMWF
Hans Vandierendonck	Queen's University Belfast
Beth Wingate	University of Exeter

Contributions from:

Peter Challenor	University of Exeter
Paul Dellar	University of Oxford
Richard Everson	University of Exeter
Serge Guillas	UCL
Dejice Jacob	University of Glasgow
Eike Mueller	University of Bath
Jemma Shipton	Imperial College London

Other delegates:

Delegate	University/Organisation
David Acreman	University of Exeter
Peter Ashwin	University of Exeter
Bob Beare	University of Exeter
Jack Betteridge	University of Bath
Tim Dodwell	University of Exeter
Matthew Griffith	University of Bath
Iva Kavcic	Met Office
Edward Keedwell	University of Exeter
Tim Lenton	University of Exeter
Nathan Mayne	University of Exeter
David Moxey	University of Exeter
Esmond G. Ng	Lawrence Berkeley National Laboratory, US Department of Energy
Emily Paremain	University of Exeter
Adam Peddle	University of Exeter
Andy Richards	University of Exeter
Zhe Sha	University of Bristol
Stefan Siegert	University of Exeter
Ben Shipway	Met Office
John Thuburn	University of Exeter

Follow-on funding:

As part of the EPSRC network ReCoVER, participants were offered an opportunity to apply for a feasibility funding project to follow up ideas from the workshop. Any projects funded through this and their outcomes are announced via the ReCoVER website http://recoverlwec.org.