BDEC workshop

Weather and Climate Modelling: ready for exascale?

P.L. Vidale, H. Weller and B.N. Lawrence, NCAS, University of Reading, UK

Weather and climate models are amongst the most compute-intensive tools used in geoscience. These models simulate a range of earth system processes by solving sets of coupled hyperbolic/parabolic partial differential equations on domains that range from individual country to global scales. They have been around for many decades and have undergone several transformations to exploit every emerging technology. After an entire generation of models developed to run on vector supercomputers, the latest drive has been towards massive parallelism. The trend in W&C modelling has always been to increase numerical resolution (decrease the "mesh size") and increase model complexity (add more processes) to make the fullest possible use of available HPC. More recently, increases in CPU capacity have also been spent on exploiting data assimilation and on running multiply perturbed simulations ("ensembles"), to better initialise and sample the possible solution trajectories in the real physical systems.

At the World Modelling Summit for Climate Prediction[1] (ECMWF, 2008) a number of weather and climate modellers gathered to talk about future technology and modelling trends and to assess whether our codes are ready for future architectures. A number of requirements were discussed, based on predictions of what would be available, and are summarised in Table 1. The predictions were accurate enough, in that, as predicted, by 2012 many global NWP codes have pushed resolution to nearly 10km and a select few global climate models are operating at 25-50km resolution. However, most codes still struggle to scale on O(10^5) cores and beyond, validating the 2008 conclusion that "core counts above O(10^4) are unprecedented for weather or climate codes, so the last 3 columns require getting 3 orders of magnitude in scalable parallelization".

Table 1: affordable GCM mesh size (km) as a function of available computing capability, as predicted in 2008

	Earth Simulator 2002-2009		PRACE-HERMIT 2012-		
Peak Rate:	**10 TFLOPS**	**100 TFLOPS**	**1 PFLOPS**	**10 PFLOPS**	**100 PFLOPS**
Cores	1,400 (2005)	12,000 (2007)	80-100,000 (2009)	300-800,000 (2011)	6,000,000? (20xx?)
Global NWP[0]: 5-10 days/hr	18 - 29	8.5 - 14	4.0 - 6.3	1.8 - 2.9	0.85 - 1.4
Seasonal[1]: 50-100 days/day	17 - 28	8.0 - 13	3.7 - 5.9	1.7 - 2.8	0.80 - 1.3
Decadal[1]: 5-10 yrs/day	57 - 91	27 - 42	12 - 20	5.7 - 9.1	2.7 - 4.2
Climate Change[2]: 20-50 yrs/day	120 - 200	57 - 91	27 - 42	12 - 20	5.7 - 9.1

teraFLOPS = 10^{12} (trillion) floating point operations per second
petaFLOPS = 10^{15} (quadrillion) floating point operations per second
exaFLOPS = 10^{18} (quintillion) floating point operations per second

Range: Assumed efficiency of 10-40%
0 - Atmospheric General Circulation Model (AGCM; 100 vertical levels)
1 - Coupled Ocean-Atmosphere-Land Model (CGCM; ~ 2X AGCM)
2 - Earth System Model (with biogeochemical cycles) (ESM; ~ 2X CGCM)

Thanks to Jim Abeles (IBM)

[1] http://www.nature.com/news/2008/080514/full/453268a.html

Increasing the resolution, by for example halving the mesh size, can increase the computational cost by nearly a factor 10, enabling strong parallelism, due to the increased number of points and the concurrent need for increasingly shorter time steps. Longer time steps can be used with more sophisticated numerical schemes, such as the variants of semi-implicit and semi-Lagrangian schemes widely used in numerical weather and climate prediction models. However, a bottleneck in semi-implicit schemes is the need for a three-dimensional elliptic solve at each model time step, which requires global data communication; this increasingly amounts to a significant proportion of the model runtime. Codes developed for vector supercomputers, with low core numbers, assumed that global data communications between nodes would be few and rare; at high core numbers, and with massive domain decomposition (mostly in the horizontal direction), global communications have become so intense as to compromise scalability. Future models, on very large core counts, may revert to explicit time schemes.

The other major bottleneck is memory bandwidth: most simulations are limited by the time taken to get data from memory, rather than doing the calculations. This can be ameliorated by using higher-order accuracy, so that more calculations are done on the data loaded from memory. It is also important not to access the same data too many times from memory; once some data are loaded, they should be used for everything that needs them, i.e., the data at one point are needed to calculate gradients at all the surrounding points. This can be achieved by ordering data along space filling curves so that data stay in cache until no surrounding points need them any more. This can dramatically improve cache hits and speed up simulations. However, the semi-Lagrangian method relies on accessing data at distant departure points, which are unlikely to be in cache. The semi-Lagrangian method is therefore not suited to modern computer architectures.

Similarly, just handling the data input and output at higher resolution is becoming a bottleneck, with significant fractions of time in any given simulation job being spent on initialisation and data output (checkpoints – start dumps – still require, at times, O(1hr) to be written to disk). For instance, in the case of the 2010 version of the UK Unified Model at 25km resolution, IO was already starting to limit scalability at 1500 cores. The use of "IO server" technology alongside hybrid parallelism, assigned a large fraction of global communications to a number of specialised nodes, which made it possible to improve scalability by nearly a factor of 10. When the entire scientific workflow (including post-processing analysis) is factored in, data handling issues can dominate the time to solution (from problem conception to result); these data issues are exacerbated by increased complexity, data assimilation, and large ensembles.

The next generation of models, such as those being developed in the UK Next Generation Weather and Climate Prediction[2] (NGWCP) project, aimed at massively increasing scalability, will exploit new dynamical cores with more efficient grid topologies and numerical schemes, also reducing global data communication, even at the cost of requiring much shorter time steps. To run such models operationally, the new solvers will need to scale to $O(10^5\text{-}10^6)$ processor cores on modern computer architectures. However, even if progress with such models is slow, data problems are expected to come sooner: a modern climate model ensemble can output O(200GB) per simulated month, which, at about 2 model years per wall clock day, could result in about 5 TB per wall clock-day (the PRACE-UPSCALE[3] project sustained an output of about 2 TB/day over approximately 200 days in 2012). It is not unreasonable to assume that over the next decade some applications of such models will lead to output increases of nearly four orders of magnitude[4] (since much of that increase will not be constrained by limitations of scalability, as it comes from increasing complexity and ensemble size). This would lead to outputs of O(10-100PB) per wall-clock day. While not all such data will need to be stored, at the very least, much of it will have to be analysed in concurrent

[2] http://www.nerc.ac.uk/research/programmes/ngwcp/background.asp

[3] http://www.prace-ri.eu/PRACE-system-equips-science?lang=en

[4] For example, see The Scientific Case for High Performance Computing in Europe, Guest et al, 2012, available at http://goo.gl/nIHC0 (as downloaded Apr 24, 2013).

and/or post-processing mode, yielding the requirement for significantly increased co-located compute. Wherever this is done, it is likely that such data will need to be compared with simulations (and observations) stored elsewhere: data migration and storage will be a major problem.

Even when migration and storage are solved (e.g. by dedicated analysis facilities with high performance parallel disk and light paths to supercomputing, such as JASMIN in the UK), the problem of analysis remains. The analysis of weather and climate datasets with sizes of 10-100TB is still in its infancy: while parallel IO (e.g. in HDF5 and NetCDF4) helps, most advanced analysis involves the extraction of “features” (phenomena like intense storms) that require custom-made codes, most of which are not parallelised. As a result, while a typical climate modelling experiment can be completed in ensemble mode within one year on a petascale supercomputer, analysis will lag from 6 months to several years. Accordingly, the community needs to invest significantly in tools for improving productive parallelisation in such environments.

In summary, while there are significant problems with exploiting massive parallelisation in weather and climate computing – which are being addressed – weak scalability, coupled with increases in ensemble size and model complexity, will yield similar, if not greater problems, associated with data handling and storage technologies. The community is much less ready for these.