REQUEST FOR A SPECIAL PROJECT 2014–2016

MEMBER STATE: United Kingdom

Principal Investigator[1]: Tim Palmer

Affiliation: Oxford University

Address:
Department of Physics
Atmospheric, Oceanic and Planetary Physics
Clarendon Lab.
Parks Road
Oxford OX1 3PU

E-mail: T.N.Palmer@atm.ox.ac.uk

Other researchers:

Peter Düben, Hugh McNamara, Andreas Döring, Antje Weisheimer

Project Title:

The use of imprecise arithmetic to increase resolution in atmospheric models

If this is a continuation of an existing project, please state the computer project account assigned previously.	**SP** __________________	
Starting year: (Each project will have a well defined duration, up to a maximum of 3 years, agreed at the beginning of the project.)	2014	
Would you accept support for 1 year only, if necessary?	YES ☒	NO ☐

Computer resources required for 2014-2016: (The maximum project duration is 3 years, therefore a continuation project cannot request resources for 2016.)	**2014**	**2015**	**2016**
High Performance Computing Facility (units)	5.000.000	10.000.000	15.000.000
Data storage capacity (total archive volume) (gigabytes)	6.000	10.000	15.000

An electronic copy of this form ***must be sent*** *via e-mail to:* *special_projects@ecmwf.int*

Electronic copy of the form sent on (please specify date):
12.06.2013

Continue overleaf

---

[1] The Principal Investigator will act as contact person for this Special Project and, in particular, will be asked to register the project, provide an annual progress report of the project's activities, etc.

Principal Investigator: Tim Palmer

Project Title: The use of imprecise arithmetic to increase resolution in atmospheric models

Extended abstract

It is expected that Special Projects requesting large amounts of computing resources (500,000 SBU or more) should provide a more detailed abstract/project description (3-5 pages) including a scientific plan, a justification of the computer resources requested and the technical characteristics of the code to be used. The Scientific Advisory Committee and the Technical Advisory Committee review the scientific and technical aspects of each Special Project application. The review process takes into account the resources available, the quality of the scientific and technical proposals, the use of ECMWF software and data infrastructure, and their relevance to ECMWF's objectives. - Descriptions of all accepted projects will be published on the ECMWF website.

1. Aim and structure of this proposal

The aim of this project is to study the limits and prospects of the use of imprecise arithmetic in a global atmosphere model (IFS). In particular, we want to investigate the potential use of so-called stochastic processors and reduced precision floating point arithmetic and the propagation of uncertainties, such as hardware errors or random forcings, in a model for the global atmosphere. The use of stochastic processors and reduced precision floating point arithmetic has the potential to reduce the computational cost significantly, due to a reduced energy demand and/or an increase in performance. A reduction of computing costs allows higher computing power within the same budget for computing facilities. Higher computing power allows higher resolution and the resolution in state-of-the-art atmospheric simulations is still far from being adequate (*Shukla et al., 2010*). If the results of this project show that it is possible to use imprecise arithmetic to perform large parts of the computation of an atmosphere model this project might lead to a fundamental change in the future computing hardware used for weather and climate modelling.

Section 2 will provide background information on imprecise arithmetic and give a short overview of the ongoing research. Section 3 will outline the structure of the project. Section 4 provides information on the technical requirements and characteristics.

2. Imprecise arithmetic in weather and climate models

While the resolution of today's weather and climate models is limited by the available computing power, it is likely that the exponential growth of computational power that is predicted by Moorse law will not continue towards exascale computing. Two of the major challenges to be overcome in the development of exascale high performance computers are the energy demand and the error resilience.

There are two new approaches in hardware development that exploit a weakening of the constraints to work with bit reproducible architecture and double precision floating point arithmetic and might allow a significant increase in computational power: the use of mixed precision arithmetic and stochastic processors.

While present computing architectures are typically restricted to the use of either single or double precision for floating points, a mixed precision architecture would allow to allocate a suitable precision for each floating point number. Such an approach has the potential to reduce the number of bits that are needed to represent the physical fields in weather and climate simulations significantly. This reduction would reduce the numerical costs, since the amount of floating point operations is reduced and more data would fit into memory and cache, and would furthermore reduce the amount of storage capacity needed.

Stochastic processors sacrifice bit reproducibility since they show faulty calculations, but they offer a significant reduction of the power consumption. Traditional processor design uses rather large guarding bands to prevent natural fluctuations from impacting on the results of calculations and therefore to ensure bit-reproducibility. Guarding against fluctuations that can for example be caused by thermal noise and cosmic ray impacts requires that hardware be run at high voltages. A deliberated reduction of these guarding bands can lead to a significant reduction of the energy demand of processors, but it will allow faults to happen within the calculations of the CPU. While the work on stochastic processors is at an early prototype stage using simplified architectures, results suggest that power consumption could be reduced, on average, by anything from around 12 – 20 % (*Kahng et al. 2010, Sartori et al. 2011*) at low fault rates (1 – 2 %) up to about 90 % (at a fault rate of 10 % (*Lingamneni et al. 2012*). Further cost reductions can be expected in manufacturing, verification and testing.

The use of double precision, bit-reproducible arithmetic might be over-engineering in large parts of a model for the atmosphere or the ocean. Especially near the truncation scale at which the use of parametrisation schemes to represent sub-grid-scale processes and the use of diffusion to obtain a closure of the turbulent cascade does not allow an exact representation of the physical processes, anyway.

In previous studies, we developed software tools that emulate the use of stochastic processors and reduced precision floating point arithmetic in a numerical model. Tests were performed with toy models for atmospheric dynamics and a spectral dynamical core (IGCM) for simulating the global atmosphere. The results with the toy models could show that inexact calculations can be superior to those where smaller scales are parametrised. The results with the dynamical core suggest that neither the use of stochastic processor, nor of reduced precision floating point arithmetic do substantially affect the mean behaviour of simulations of the Held-Suarez test case, provided that the calculation of the dynamics of spherical harmonics with large wave lengths is performed on exact hardware. The calculation of the dynamics of spherical harmonics with short wave lengths and of the non-linear tendencies in grid point space appears to be fairly robust to the use of stochastic processors and reduced precision (*Düben et al., 2013*). This suggests that the use of imprecise arithmetic could reduce computation and power costs in an atmosphere model without adversely affecting the quality of the simulations.

If we can reduce the precision of the arithmetic that is used to compute an atmosphere model in a scale-selective fashion, we may be able to develop much higher resolution models, for a given budget for computing facilities. Consistent with this, a recent paper showed that a decrease in precision on the software level, by using an inexact fast Legendre Transform, can lead to a decrease of the computational cost without a strong degrading of the quality of the model simulations. The increase in performance allows simulations with the IFS with higher resolution (up to T7999) (*Wedi et al. 2013*). This project shall investigate if a similar trade-off is possible on a hardware level.

The amount of computing resources for which we apply in this special project would provide the computing power that we need to investigate the use of stochastic processors and imprecise floating point arithmetic in a global atmosphere model (IFS) with full physics at more realistic resolution for weather and climate simulations. If the project shows that it is possible to use imprecise arithmetic in large parts of the computation of an atmosphere model the results might have a strong impact on future hardware architecture used for modelling weather and climate.

3. Structure of the project

The project would investigate three lines of research:

1. Imprecise floating point arithmetic

In a first step, we want to investigate if it is possible to replace the double precision floating point numbers by single precision floating point numbers in large parts of the IFS. We expect that the use of single precision already leads to a significant speed-up and we already started to build up a setup for the OpenIFS that uses mainly single precision arithmetic. In a second step, we want to investigate the minimal precision which is needed in specific, cost-intensive parts of the model. Such a study would allow an estimate of the performance increase that would be possible with processors that are able to use mixed precision. These processors are not produced yet, but it is likely that they will play an important role in future HPC systems. As a by-product, the knowledge we gain on the necessary floating point precision in IFS could be useful to reduce the amount of data storage – when not all of the output data is stored in double precision – or to provide useful information in the development of imprecise but efficient software techniques, such as the Fast Legendre Transform. The second step would involve the use of the same emulator which is used for stochastic processors, with different rules to induce hardware faults.

2. Stochastic processors

Stochastic processors allow hardware-induced faults in calculations and sacrifice bit-reproducibility in exchange for improvements in performance and/or a reduction in the energy demand. We already developed and tested software tools to emulate hardware faults of stochastic processors in a numerical simulation. We want to use this software to emulate the use of stochastic processors for certain parts of the IFS, for example in the calculation of the dynamics of spherical harmonics with high wave numbers, or to calculate the physical part and parametrisation schemes of the IFS. Towards the use of real stochastic hardware, we collaborate with the research group of Prof. Krishna Palem at Rice University who is working on the hardware development of stochastic processors. This collaboration will allow us to emulate realistic hardware and to make good estimates for possible energy savings.

3. Propagation of uncertainty

The influence of uncertainties – introduced for example by random numbers in stochastic parametrisation schemes, rounding errors, or random hardware faults – is very difficult to estimate. An estimation typically involves the use of very large ensembles of simulations, which implies a huge amount of computing hours for full-grown atmosphere models. We want to investigate to which extent it is possible to estimate the evaluation of uncertainties through a numerical programme without calculating a full ensemble. This involves the development of specific propagation rules for uncertainties for the different numerical operations. Within the model such a propagation of uncertainties can be realised by allocating the prognostic quantities as special types that allow to use overloaded operators. If it is possible to introduce such a tool into the IFS, the approach might be able to provide at least a rough estimate on how for example random forcing terms or rounding errors will influence the ensemble spread of numerical simulations, without the need to run a full ensemble.

4. Technical requirements and characteristics

It is the aim of this project to proceed with each of the three lines of research introduced above (1.-3.) to a state that allows simulations of weather forecasts with the uncoupled IFS. The optimistic aim for the first year of this project is to be able to perform ensemble simulations with T159 horizontal resolution and 91 vertical levels with emulated stochastic processors, and emulated reduced precision floating point arithmetic. The ensemble simulations for a forecast of ten days shall be calculated with fifty ensemble members for thirty different forecasting dates, to get sufficient statistics. To calculate such a setup the for reference simulations with stochastic parametrisation would need approximately 60.000 System Billing Units (SBU). Unfortunately, the use of the emulator for stochastic processors and reduced precision floating point arithmetic comes along with a significant computational overhead and will decrease and not increase the speed of the simulations. Therefore, we calculate the estimated costs for the T159 ensemble simulations with the emulator with an additional factor of ten. Testing will be necessary and it is unlikely that we will find the perfect setup already in the first set of ensemble simulation. We will probably need to test several different setups in which the emulator is used in different parts of the model and at different precision levels or fault rates. For these reasons we multiply the cost for the ensemble simulations with emulator with an additional factor of four. The study of the propagation of uncertainties will not involve ensemble simulations and will therefore be much cheaper. We end up with the approximated amount of 5.000.000 SBUs for the first year of the project. In the second and third year, we want to proceed to simulations with higher resolution. The needed computational resources are therefore higher, as well.

We already started with tests in the OpenIFS. This system allows an easy approach due to the reduction of code to a basic level and the easy compiling procedures, but it still incorporates all parts of the IFS that are necessary for a first approach. When tests with OpenIFS are successful, we will proceed to the IFS. The preferred setup of IFS to emulate the use of stochastic processors and imprecise floating point arithmetic and to estimate the propagation of uncertainties would allocate the prognostic variables as specific types that allow to work with overloaded operators. It is very well possible that such a setup of IFS would be useful for a number of different studies beyond the scope of this project, especially for those that deal with ensemble simulations.

References

Düben, P. D., H. McNamara and T. N. Palmer, 2013: The use of imprecise processing to improve accuracy in weather & climate prediction, Journal of Computational Physics, submitted

Kahng A., S. Kang, R. Kumar, J. Sartori, 2010: Slack redistribution for graceful degradation under voltage overscaling, Design Automation Conference (ASP-DAC), 15th Asia and South Pacific, 2010, pp. 825 –831.

Lingamneni A., K. K. Muntimadugu, C. Enz, R. M. Karp, K. V. Palem, C. Piguet, 2012: Algorithmic methodologies for ultra-efficient inexact architectures for sustaining technology scaling, Proceedings of the 9th conference on Computing Frontiers, CF ’12, ACM, New York, NY

J. Sartori, J. Sloan, R. Kumar, 2011: Stochastic computing: Embracing errors in architecture and design of processors and applications, Compilers, Architectures and Synthesis for Embedded Systems (CASES), 2011 Proceedings of the 14th International Conference on

Shukla J., T.N. Palmer, R. Hagedorn, B. Hoskins, J. Kinter, J.Marotzke, M.Miller, J. Slingo, 2010: Toward a new generation of world climate research and computing facilities, Bulletin of the American Meteorological Society **91** (10)

Wedi, P. W., M. Hamrud, G. Mozdzynski, 2013: A fast spherical harmonics transform for global NWP and climate models, Mon. Wea. Rev., accepted