Reduced numerical precision in weather and climate models

Peter Düben, Stephen Jeffress, Tim Palmer,
J Schlachter, Parishkrati, S Yenugula, J Augustine, C Enz, K Palem, F Russell, X Niu, W Luk

University of Oxford, EPFL, IITM, Rice University, Imperial College

Why should we use reduced numerical precision in weather and climate predictions?

- Numerical models are crucial for reliable forecasts of future weather and climate.
- The quality of these forecasts dependents strongly on the resolution and complexity of the numerical models used.
- Resolution is limited by the computational power of state-of-the-art super computers.

Why should we use reduced numerical precision in weather and climate predictions?

- Numerical models are crucial for reliable forecasts of future weather and climate.
- The quality of these forecasts dependents strongly on the resolution and complexity of the numerical models used.
- Resolution is limited by the computational power of state-of-the-art super computers.

- The free lunch is over: We can not assume that the steady increase in resolution and computational power will continue.

What is reduced precision hardware?

My definition: Reduced precision hardware is using a level of numerical precision which is smaller than double precision.

- Reduced precision hardware allows a reduction of power consumption and/or an increase in performance and therefore a reduction of computational cost.
- This would allow simulations at higher resolution and possibly more accurate forecasts.

What is reduced precision hardware?

My definition: Reduced precision hardware is using a level of numerical precision which is smaller than double precision.

- Reduced precision hardware allows a reduction of power consumption and/or an increase in performance and therefore a reduction of computational cost.
- This would allow simulations at higher resolution and possibly more accurate forecasts.

It turns out that there are plenty of different approaches in hardware development that study the trade between precision and performance!

What is reduced precision hardware?

My definition: Reduced precision hardware is using a level of numerical precision which is smaller than double precision.

- Reduced precision hardware allows a reduction of power consumption and/or an increase in performance and therefore a reduction of computational cost.
- This would allow simulations at higher resolution and possibly more accurate forecasts.

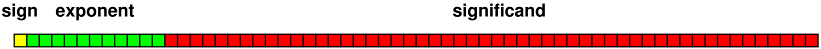
It turns out that there are plenty of different approaches in hardware development that study the trade between precision and performance!

Easiest way: double $\rightarrow$ single ($\rightarrow$ half).

Three approaches to imprecise processing

Stochastic processor

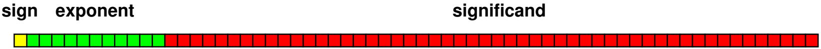
- If we reduce the applied voltage or the wall clock time beyond a certain level, we will get hardware errors, but we will save power.
- The error rate of a stochastic processor can be reduced massively, if the architecture is changed.



Three approaches to imprecise processing

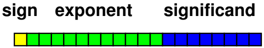
Stochastic processor

- If we reduce the applied voltage or the wall clock time beyond a certain level, we will get hardware errors, but we will save power.
- The error rate of a stochastic processor can be reduced massively, if the architecture is changed.



Pruning

Parts of the CPU that are hardly used or do not have a strong influence on significant bits are removed.



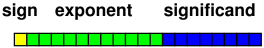
Three approaches to imprecise processing

Stochastic processor

- If we reduce the applied voltage or the wall clock time beyond a certain level, we will get hardware errors, but we will save power.
- The error rate of a stochastic processor can be reduced massively, if the architecture is changed.

Pruning

Parts of the CPU that are hardly used or do not have a strong influence on significant bits are removed.



Field Programmable Gate Array (FPGA)

- FPGAs are integrated circuits that can be configured by the user.
- Numerical precision can be customised to the application.

A scale-selective approach

Spectral models use spherical harmonics as basis functions to represent physical fields. They allow to treat different scales at different levels of precision.

A scale-selective approach

Spectral models use spherical harmonics as basis functions to represent physical fields. They allow to treat different scales at different levels of precision.

We can push the small scales harder than the large scales.

A scale-selective approach

Spectral models use spherical harmonics as basis functions to represent physical fields. They allow to treat different scales at different levels of precision.

We can push the small scales harder than the large scales.

This is intuitive due to the high inherent uncertainty in small scale dynamics (parametrisation, viscosity, data-assimilation).

A scale-selective approach

Spectral models use spherical harmonics as basis functions to represent physical fields. They allow to treat different scales at different levels of precision.

We can push the small scales harder than the large scales.

This is intuitive due to the high inherent uncertainty in small scale dynamics (parametrisation, viscosity, data-assimilation).

The smallest scales are the most expensive once.

Our vision....

A global atmosphere and/or ocean model which is using just the right level of precision and reducing numerical precision with scales.

Large scales: Double precision,   Small scales: Half precision

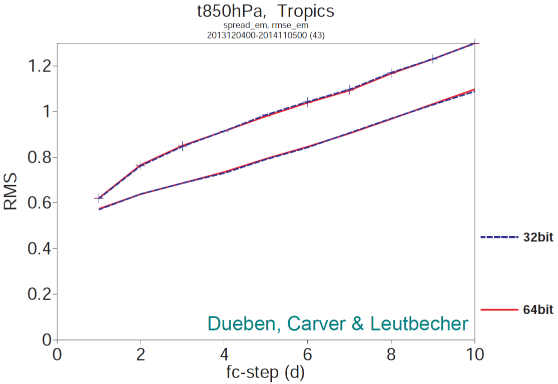
Our vision....

A global atmosphere and/or ocean model which is using just the right level of precision and reducing numerical precision with scales.

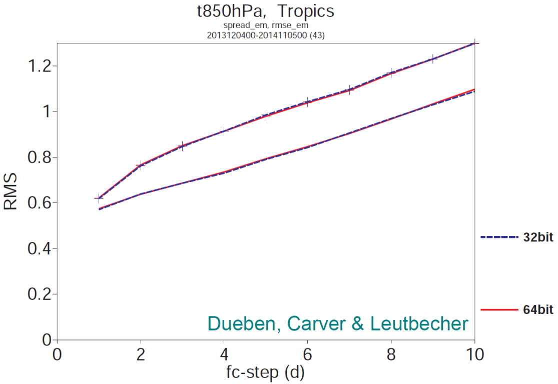
Large scales: Double precision,   Small scales: Half precision

How will rounding errors affect the solution?

How to approach full-blown earth system models? OpenIFS in single precision



How to approach full-blown earth system models? OpenIFS in single precision



- Approximately one third speed-up.
- Less computing nodes are needed due to reduced memory requirements.

Reduced precision in an atmosphere model

- We calculate weather forecasts with a spectral dynamical core (IGCM) in a "Held-Suarez world" and compare results against a high resolution truth.
- Floating point precision for the significand is reduced to 8 or 10 bits instead of 52 bits for double precision using an emulator.
- In the reduced precision setup, only 2% of the computational cost of the control simulation is calculated in double precision.
- Scale separation turned out to be really important.

What are the savings?

- In cooperation with Rice University (USA) and EPFL (Switzerland) we derive hardware setups of the floating point unit, memory and cache that show comparable error pattern.
- We analyse the possible savings and trace the application to obtain an estimate for the power consumption on the exact and the reduced precision hardware.

What are the savings?

- In cooperation with Rice University (USA) and EPFL (Switzerland) we derive hardware setups of the floating point unit, memory and cache that show comparable error pattern.
- We analyse the possible savings and trace the application to obtain an estimate for the power consumption on the exact and the reduced precision hardware.

Resolution	Precision FP significand	Normalised Energy Demand	Forecast error day 2
235 km	52	1.0	2.3
315 km	52	0.47	4.5
235 km	**10**	0.32	2.3
235 km	**8**	0.29	2.5

Forecast error: Mean error in geopotential height.
See Düben et al., DATE, 2015 for more details.

What are the savings?

- In cooperation with Rice University (USA) and EPFL (Switzerland) we derive hardware setups of the floating point unit, memory and cache that show comparable error pattern.
- We analyse the possible savings and trace the application to obtain an estimate for the power consumption on the exact and the reduced precision hardware.

Resolution	Precision FP significand	Normalised Energy Demand	Forecast error day 2
235 km	52	1.0	2.3
315 km	52	0.47	4.5
235 km	**10**	0.32	2.3
235 km	**8**	0.29	2.5

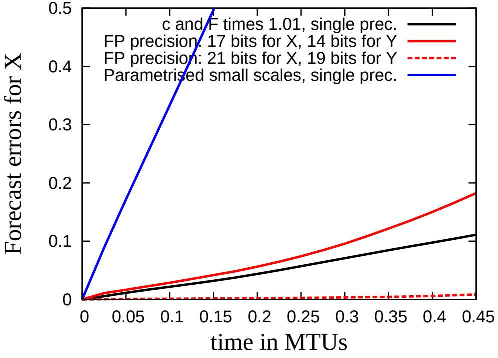
Forecast error: Mean error in geopotential height.
See Düben et al., DATE, 2015 for more details.

To save power a reduction in precision is much more efficient than a reduction in resolution!

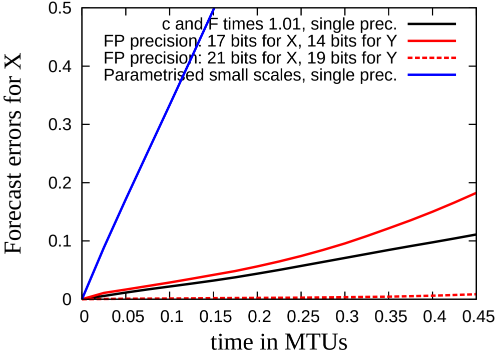
A toy model for atmospheric dynamics on FPGAs

- We implemented the Lorenz ’96 model on an FPGA in cooperation with Xinyu Niu, Francis Russel and Wayne Luk from Imperial College.
- We scale the size of Lorenz ’96 to the size of a high performance application (up to more than 100 million degrees-of-freedom).
- We compare results with reduced precision against results with perturbed parameters (by 1 %) or parametrised small scales.

Lorenz ’96 on an FPGA: Weather



Lorenz ’96 on an FPGA: Weather



Changes in weather type forecasts are comparably small when precision is reduced.

Lorenz ’96 on an FPGA: Climate

Precision	Hellinger distance
c and F times 1.01, single prec.	0.0054
Parametrised small scales, single prec.	0.1137
FP precision, 17 bits for X, 14 bits for Y	0.0079
FP precision, 21 bits for X, 19 bits for Y	0.0029

The Hellinger distance describes the difference between two PDFs.

Lorenz ’96 on an FPGA: Climate

Precision	Hellinger distance
c and F times 1.01, single prec.	0.0054
Parametrised small scales, single prec.	0.1137
FP precision, 17 bits for X, 14 bits for Y	0.0079
FP precision, 21 bits for X, 19 bits for Y	0.0029

The Hellinger distance describes the difference between two PDFs.

Changes in climate type forecasts are comparably small when precision is reduced.

Lorenz ’96 on an FPGA: Speed and Power

Hardware	Speed	Energy efficiency
CPU, 12 cores, single precision	1.0	1.0
CPU, 12 cores, double precision	0.5	-
FPGA, single precision	2.8	10.4
FPGA, 17 bits for X, 14 bits for Y	6.9	23.9
FPGA, 21 bits for X, 19 bits for Y	5.4	18.9

Lorenz '96 on an FPGA: Speed and Power

Hardware	Speed	Energy efficiency
CPU, 12 cores, single precision	1.0	1.0
CPU, 12 cores, double precision	0.5	-
FPGA, single precision	2.8	10.4
FPGA, 17 bits for X, 14 bits for Y	6.9	23.9
FPGA, 21 bits for X, 19 bits for Y	5.4	18.9

We get significant savings in energy consumption and a significant increase in performance if we use FPGAs with reduced precision.

Conclusions

- Double precision as default is overcautious in earth system modelling.
- There are several different ways to trade precision against performance in hardware development (Stochastic processors, pruned FPUs, FPGAs, half precision,...).
- Reduced precision hardware allows significant savings. Freed resources can improve forecast quality.
- To save power, a reduction of precision is more efficient compared to a reduction in resolution.

References

PD Düben, J Joven, A Lingamneni, H McNamara, G De Micheli, KV Palem, TN Palmer, Phil. Trans. A, 2014

PD Düben, TN Palmer, H McNamara, JCP, 2014

TN Palmer, PD Düben, H McNamara, Phil. Trans. A, 2014

PD Düben, TN Palmer, Mon. Weath. Rev., 2014

PD Düben, J Schlachter, Parishkrati, S Yenugula, J Augustine, C Enz, K Palem and TN Palmer, DATE, 2015

F Russell, PD Düben, X Niu, W Luk, TN Palmer, FCCM, 2015

PD Düben, S Jeffress, TN Palmer, EMIT, 2015

PD Düben, SI Dolaptchiev, submitted to TCFD

PD Düben, F Russel, X Niu, W Luk, TN Palmer, submitted to JAMES