Computer Physics Communications 221 (2017) 160–173

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Exploiting the chaotic behaviour of atmospheric models with reconfigurable architectures

Francis P. Russell [a],*, Peter D. Düben [b], Xinyu Niu [a], Wayne Luk [a], T.N. Palmer [b]

[a] Department of Computing, Imperial College London, United Kingdom
[b] Atmospheric Oceanic and Planetary Physics, University of Oxford, United Kingdom

ARTICLE INFO

Article history:
Received 14 March 2017
Received in revised form 10 August 2017
Accepted 11 August 2017
Available online 19 August 2017

Keywords:
FPGA
Chaotic system
Precision reduction
Weather modelling

ABSTRACT

Reconfigurable architectures are becoming mainstream: Amazon, Microsoft and IBM are supporting such architectures in their data centres. The computationally intensive nature of atmospheric modelling is an attractive target for hardware acceleration using reconfigurable computing. Performance of hardware designs can be improved through the use of reduced-precision arithmetic, but maintaining appropriate accuracy is essential. We explore reduced-precision optimisation for simulating chaotic systems, targeting atmospheric modelling, in which even minor changes in arithmetic behaviour will cause simulations to diverge quickly. The possibility of equally valid simulations having differing outcomes means that standard techniques for comparing numerical accuracy are inappropriate. We use the Hellinger distance to compare statistical behaviour between reduced-precision CPU implementations to guide reconfigurable designs of a chaotic system, then analyse accuracy, performance and power efficiency of the resulting implementations. Our results show that with only a limited loss in accuracy corresponding to less than 10% uncertainty in input parameters, the throughput and energy efficiency of a single-precision chaotic system implemented on a Xilinx Virtex-6 SX475T Field Programmable Gate Array (FPGA) can be more than doubled.

1. Introduction

Climate and weather prediction are computationally intensive; even with high-performance computing resources, it is typically impossible to resolve important convective cloud systems in global models [1]. Numerical models of weather and climate show significant model error due to limited resolution and complexity, necessitating a need for even more resource-intensive models. Performance and power requirements for running such models with hard time constraints, for example in operational weather forecasts, have led to the exploration of hardware accelerators such as GPUs and FPGAs to obtain greater throughput and power efficiency [2,3].

With reconfigurable architectures becoming more prominent and increasingly accessible for running accelerated computation, we choose to investigate how designs may be customised to take advantage of the characteristics of climate simulations. A common approach to enhance throughput of hardware designs is to reduce precision of calculations so that additional hardware resources can be employed to increase parallelism. Excessive precision reduction however, reduces calculation quality and usefulness, so a trade-off must be made between performance and accuracy.

Lorenz showed that weather forecasting involves the prediction of a chaotic system [4]. This implies an exponential growth of errors in any perturbation of the model, such as a slight change in initial conditions; significant divergence between a CPU and hardware implementation is expected simply due to implementation differences, and is not in itself an indicator of error. Diagnostics that rely on solution convergence between implementations for validation are no longer appropriate.

Weather and climate models are known to show significant limitations due to limited resolution or insufficient complexity in the representation of the Earth system. Short-term forecasts will be perturbed by uncertainty in initial conditions due to the nature of measurements and data assimilation. If the model setup is changed, e.g. by a change of software/hardware implementation, model parameters or model forcing (such as CO_2 concentration), it is almost impossible to predict the response of the model due to the chaotic nature of the system and numerical simulation is necessary to identify the impact.

A reduction in precision will reduce computational cost and resource usage but will also alter arithmetic behaviour and influence model simulations. However, given the model limitations that are outlined above, one may consider a reduction in precision

* Corresponding author.
E-mail address: francis.russell02@imperial.ac.uk (F.P. Russell).

http://dx.doi.org/10.1016/j.cpc.2017.08.011

appropriate so long as the behaviour change introduced is insignificant compared to those introduced by other factors. The resources saved through precision reduction may then facilitate an increase of design throughput or permit more complex designs.

To this end, we investigate the reduction of precision in chaotic systems using the Lorenz 1996 model (a.k.a. Lorenz 1995 and referred to hereafter as Lorenz '96), designed by Lorenz [5] to study interactions of atmospheric processes with non-linear, chaotic dynamics. Our analysis has applicability beyond this model — scale interactions within the Lorenz '96 model resemble scale interactive behaviour of various parts of numerical atmosphere models that is typically difficult to capture in idealised systems. Similar scale interactions are also important for turbulent energy cascades relevant to most applications in CFD (computational fluid dynamics).

The Lorenz '96 model has been used in extensive studies in the literature to investigate new methods for data assimilation, the propagation and representation of model error and improved numerical algorithms involving dynamical systems [6–10].

This paper contributes and presents:

- the hardware architecture of a two-scale Lorenz '96 simulation using Runge–Kutta time-stepping;
- a demonstration of how it is possible to make trade-offs between precision and throughput for a system where numerical divergence is expected;
- an analysis of the impact of varying the precision of variables at different scales in the Lorenz '96 implementation using error metrics appropriate to chaotic systems (the Hellinger distance);
- performance, precision and power consumption comparisons of reduced and single-precision FPGA implementations.

This paper is an extended version of work presented at the 23rd IEEE International Symposium on Field-Programmable Custom Computing Machines in 2015 (FCCM'15) [11]. Additional contributions and content include:

- a more detailed presentation of the chaotic properties of Lorenz '96 and the effect of arithmetic precision and parameter changes on the evolution of the system (Section 4.1);
- a detailed overview of the methodology we used to determine the value representations we chose to use for hardware builds (Section 4.3);
- the application of the same methodology with fixed-point representations, and the presentation of power, precision and performance results for the resulting builds (Section 4.7 & 6);
- a more detailed performance analysis of hardware builds, including calculations of arithmetic throughput and bandwidth utilisation (Section 6.2).

2. Field Programmable Gate Arrays as specialised accelerators

Field Programmable Gate Arrays (FPGAs) are integrated circuits whose architecture can be reconfigured. FPGAs are achieving increasing visibility as an accelerator architecture, present in Microsoft, Amazon and IBM data centres — a major benefit being a high compute-to-power ratio. Microsoft's Project Catapult has deployed FPGA accelerators into Microsoft data-centre servers where a performance improvement of 95% in search engine scoring was obtained relative to a software implementation running on 12-core Intel Sandy Bridge CPUs but with a maximum power overhead of only 22.7 W [12]. A revised accelerator architecture is now being deployed at hyperscale in Microsoft's production datacentres worldwide [13].

Alternative accelerator architectures include GPUs (Graphics Processing Units) and the Intel Xeon Phi. Performance and power comparisons typically show highly application-dependent results with arithmetic intensity and regularity of data access being important factors [14,15]. Given the significant power requirement for executing climate models, our work primarily focuses on power efficiency rather than the computational throughput of a given device. High-performance GPUs typically have large power requirements [16], but can be power-efficient when high computational throughput is achieved e.g. for operations such as matrix-multiply [15]. However, finite difference computations like those used in many weather models are typically bandwidth bound and may only achieve a fraction of peak performance. In situations where near peak performance of a GPU cannot be attained, FPGAs can provide a significant advantage over GPUs in terms of energy efficiency [16].

FPGAs typically contain reconfigurable logic blocks, random access memory elements (Block RAM) totalling tens of MBs and DSP (digital signal processing) elements which provide efficient implementation of various numerical primitives. Through configuration of the programmable logic, arbitrary designs can be placed on the chip provided that resource, routing and timing constraints can be satisfied.

Conventional CPU cores contain pipelines designed for the execution of general purpose calculations where operations in a pipeline are dependent on an instruction stream that is unknown until execution of a program. Although it is possible to materialise similar architectures on FPGAs, most efficient utilisation occurs when the architecture chosen is customised to the computation being performed.

In an architecture designed for a specific problem, the pipeline will only involve steps required for the calculation. Since the pipeline is intended solely to execute a specific calculation, there is typically no need for a fetch–decode–execute cycle or techniques such as branch prediction or out-of-order execution which are frequently employed in general purpose architectures. Obtaining effective performance from an FPGA thus depends on the creation of an architecture that can perform a given computation efficiently rather than mapping it to a pipeline chosen in advance.

Optimised FPGA architectures typically make extensive use of pipeline parallelism, and for compute intensive workloads, may perform thousands of operations per cycle. This extensive parallelism enables FPGAs to achieve significant throughput despite having lower clock frequencies (typically hundreds of MHz) than conventional CPU architectures. Key to achieving efficiency is also the choice of number representation — real numbers need not be represented as IEEE floating-point values and using reduced-precision floating or fixed-point representations can significantly reduce resources, making it possible to place higher throughput designs on a device. We exploit the ability for FPGAs to use custom number representations in this work, by choosing representations that possess only the precision necessary to perform calculations of interest. In contrast, a conventional CPU is only capable of performing calculations efficiently with representations chosen during its design.

FPGAs have previously been applied to Limited Area Models such as BOLAM [17] and the global shallow water equations [18] using finite difference schemes though such work has considered short-term simulations that show only limited propagation of model errors due to the chaotic dynamics of the atmosphere. In this work, we consider the influence of reduced precision on a long-term diagnostic for a system with strong chaotic behaviour. Furthermore, the Lorenz '96 model allows us to choose the level of precision reduction to apply at different scales, which we also explore. Previous explorations have been limited to software-based simulations which could only be run at much smaller scale [19,20].