ESCAPE 2

Workshop on fault tolerant algorithms and resilient approaches for exascale computing

Dipartimento di Matematica

Politecnico di Milano

Building 14 ‘La Nave’

January 23 2019

Local organizers

Tommaso Benacchio tommaso.benacchio@polimi.it

Luca Bonaventura luca.bonaventura@polimi.it

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 800987

Funded by the European Union

Co-ordinated by ECMWF

The FET-HPC H2020 project ESCAPE2 is the prosecution of the just completed ESCAPE project. In this new initiative, resilient computing techniques and fault tolerant algorithms will also be investigated for applications to numerical weather prediction modelling on exascale computers. The goal of this workshop is to present recent results by some of the leading researchers in this area and to foster discussion with project participants, with the aim of identifying the most promising strategies to build the exascale numerical weather prediction codes whose development is at the core of the ESCAPE2 project.

Workshop programme

January 23

Dipartimento di Matematica, Politecnico di Milano, Sala di Consiglio

7th floor of Building 14 'La Nave', entrance from Via Giuseppe Ponzio 31-33

13.45 Introductory remarks

14.00 *Luc Giraud* (MOX Series Seminar)

15.00 *Mirco Altenbernd*

15.45 Coffee break

16.30 *Chris Cantwell*

17.15 *Keita Teranishi*

18.00 *Peter Düben*

18.45 Closing remarks

Abstracts

Luc Giraud, INRIA

Dealing with unreliable computing platforms at extreme scale

The advent of extreme scale computing platforms will require the use of parallel resources at an unprecedented scale. On the technological side, the continuous shrinking of transistor geometry and the increasing complexity of these devices affect dramatically their sensitivity to natural radiation, leading to a high rate of hardware faults and thus diminishing their reliability. Handling fully these faults at the computer system level may have a prohibitive computational and energetic cost. High performance computing applications that aim at exploiting all these resources will thus need to be resilient. In this talk, we will first give an overview of the current trends towards exascale. We will discuss the new challenges to face in terms of platform reliability and associated variety of possible faults. We will then discuss some of the solutions that have been proposed to tackle these errors before discussing in more detail some contributions in sparse numerical linear algebra. First, in the context of computing node crashes, we will discuss possible remedies in the framework of the solution of linear systems. Second, we will discuss a somehow more challenging problem related to silent transient soft-errors produced by natural radiation, consisting in a bit-flip in a memory cell producing unexpected results at the application level. In that context, we will consider the conjugate gradient (CG) method, that is the most widely used iterative scheme for the solution of large sparse systems of linear equations when the matrix is symmetric and positive definite. We will investigate through extensive numerical experiments the sensitivity of CG to bit-flips and further discuss possible numerical criteria to detect the occurrence of such faults. These research activities have been conducted in collaboration with many colleagues including E. Agullo (Inria), S. Cools (University of Antwerpen), E. Fatih-Yetkin (Kadir Has University), P. Salas (CERFACS), W. Vanroose (University of Antwerpen) and M. Zounon (NAG).

Mirco Altenbernd, University of Stuttgart

Fault-tolerance for linear solvers with a focus on multigrid

There is broad consensus that future leadership-class machines will exhibit a substantially reduced mean-time-between-failure (MTBF). Any simulation run will be compromised without inclusion of resilience techniques into the underlying software stack and system. Especially the rising number of cores, which is expected on the way to exascale computing, has a great impact. Multigrid methods are optimal solvers for diffusion-like problems and widely used as preconditioners or even standalone solvers. We introduce an algorithm-based fault-tolerance scheme to detect and repair soft transient faults (SDC, bitflips). By applying the full approximation scheme (FAS) variant of multigrid to linear systems, we use invariants that enable fault detection and correction, and ultimately lead to a black-box protection of the smoothing stage. We only employ readily available quantities and thus have a minimal overhead in the fault-free case. Furthermore, using multigrid gives the opportunity to use the underlying hierarchy to create compressed checkpoints for a recovery in a node-loss scenario. In addition, we examine the advantages of different lossy compression techniques. For the implementation, we have developed, in cooperation with the University of Muenster, a high-level C++ approach to manage local errors, asynchrony and faults in MPI applications, which integrates seamlessly with the upcoming MPI-ULFM standard. The above mentioned research activities have been conducted in collaboration with Dominik Goeddeke (University of Stuttgart).

Chris Cantwell, Imperial College

Exascale resilience strategies for transient solvers

Time-dependent partial differential equations (PDEs) arise in a wide range of application areas, for example in fluid dynamics. The high-fidelity resolution of complex flows often requires large-scale computational resources and is one of the drivers towards exascale computing. However, to maintain the usefulness and efficiency of the computational tools used to solve these problems, they need to be made more tolerant of the frequent hardware failures anticipated to occur on future exascale systems. Conventional resilience techniques use disk-based check-pointing methods, which have been shown not to scale to large numbers of cores. Recovery typically involves the complete restart of the application. In this talk, I will present our latest efforts to address this challenge. We combine the proposed user-level failure mitigation (ULFM) extensions to MPI and remote in-memory check-pointing in a minimally intrusive way, in order to augment existing transient PDE solvers with scalable fault tolerance capabilities. Our resilience approach improves forward-path efficiency over conventional techniques, by avoiding the parallel file system completely, and allows one or more concurrently failed ranks to be rebuilt with spare ranks on-the-fly and independently of other non-failed processes. I will describe the algorithms and their performance characteristics, and illustrate their application through examples using the Nektar++ spectral/hp element framework.

Keita Teranishi, Sandia National Laboratories

Local Failure Local Recovery: Toward Scalable Resilient Parallel Programing Model

With growing scale and complexity of computational systems, HPC applications are increasingly susceptible to a wide variety of hardware and software faults. Accordingly, applications are ill-equipped to deal with the full spectrum of possible faults and often their response, particularly in synchronous programming models, is disproportionate to fault rate. Alternatively, Local Failure Local Recovery (LFLR), is based on the notion that a fault recovery that is localized around their occurrence is more scalable and efficient than a bulk response characterized by the traditional checkpoint/restart. LFLR is more amenable with an asynchronous programming model as opposed to synchronous ones. In this study, we review the existing resilient parallel programming models and then demonstrate the efficiency and scalability of our resilient programming model for the traditional message passing and emerging asynchronous many task programming models.

Peter Düben, ECMWF

A hands-on approach to secure weather and climate models against hardware faults

Enabling Earth System models to run efficiently on future supercomputers is a serious challenge for model development. One of the major threats for weather and climate predictions on future HPC architectures is the presence of hardware faults that will frequently hit large applications in exascale supercomputing. In this talk, we present a simple hands-on approach to secure the dynamical core of weather and climate models against hardware faults using a backup system that stores coarse resolution copies of prognostic variables. Frequent checks of the model fields on the backup grid allow the detection of severe hardware faults. Prognostic variables that are changed by hardware faults on the model grid can be restored from the backup grid to continue model simulations with no significant delay. To justify the approach, we perform model simulations with a C-grid shallow water model in the presence of frequent hardware faults.