Progetto di ricerca

Titolo: Tecniche per la gestione di guasti in sistemi HPC

Tutor: Prof. A. Ciampolini

Motivazioni

I moderni supercomputer, o sistemi per High Performance Computing (HPC) sono composti da centinaia di migliaia di nodi di computazione contenuti in CPU tradizionali oppure in GP-GPU. La loro finalità è spesso legata a computazioni particolarmente onerose in ambito scientifico o industriale, che richiederebbero tempi troppo lunghi se eseguite su un hardware comune. L'architettura di questi supercomputer differisce notevolmente da quella di computer tradizionale, ed anche le problematiche che possono insorgere durante la computazione presentano spesso una natura diversa, generalmente più complessa.

Un esempio particolarmente significativo è rappresentato dalla frequenza degli errori hardware.

Infatti, sebbene il tempo medio tra l'insorgere di due guasti successivi su un computer tradizionale sia progressivamente aumentato nell'ultima decade, raggiungendo l'ordine delle centinaia di anni, la concentrazione di unità di elaborazione dei sistemi HPC è talmente elevata che in questi sistemi di elaborazione è stato spesso osservato un decremento del tempo medio tra due guasti, dell'ordine di qualche ora.

Per questo motivo il problema dell'identificazione e della gestione dei guasti in ambito HPC ha assunto un'importanza sempre maggiore negli ultimi anni. Alcune soluzioni si dedicano ai cosiddetti *hard-faults*, ovvero guasti che interessano uno o più nodi di elaborazione, determinandone il completo arresto e la conseguente perdita dei dati. Altri errori, denominati *soft-faults*, sono invece causati da eventi esterni (come, ad esempio, alterazioni termiche o radiazioni) il cui risultato non è l'arresto del nodo, ma l'alterazione del valore di qualche bit. Queste piccole variazioni del valore di dati elaborati dal nodo interessato dal guasto, dipendentemente dal tipo di computazione in atto, possono essere successivamente propagate e amplificate, determinando errori significativi sui risultati dell'elaborazione.

Obiettivi

La ricerca sarà incentrata sullo studio di metodologie e tecniche per il trattamento di guasti nei sistemi HPC.

In particolare, l'analisi partirà dai risultati fin qui ottenuti nell'ambito della soluzione di sistemi lineari con la parallelizzazione e modifica volta alla tolleranza ad *hard-faults* di un solutore denominato Metodo delle Interdizioni (IMe)[1]. A partire da tali risultati, verranno studiati metodi per migliorare l'efficienza delle tecniche già proposte e diminuire l'overhead imputabile a computazione ed a scambio di messaggi introdotto dal sistema di tolleranza ai guasti. Si estenderà inoltre la ricerca all'ambito dei *soft-faults* e si valuterà la possibilità di applicare tecniche analoghe nella soluzione di altri problemi di algebra lineare tipicamente richiesti in ambito scientifico, quali il calcolo degli autovalori o il prodotto tra matrici.

La ricerca esplorerà inoltre la possibilità di applicare tecniche per la rilevazione, la gestione e il recupero di guasti a livello di sistema, aumentando la generalità, la flessibilità e la trasparenza della soluzione proposta.

Gli obiettivi della attività possono essere sintetizzati come segue:

1) Analisi e studio dello stato dell'arte nell'ambito della soluzione di sistemi lineari con tolleranza a guasti di tipo hard-fault.

2) Individuazione di strategie per aumentare l'efficienza delle soluzioni esistenti e sviluppo di soluzioni che integrino la tolleranza a guasti di tipo soft-fault
3) Estensione della gestione dei guasti ad altri ambiti tramite l'applicazione di tecniche a livello di algoritmo (i.e., adattamento dello stato dell'arte ad altre computazioni di algebra lineare) oppure a livello di sistema.

Attività

Pianificazione delle attività

Si prevede uno svolgimento di 12 mesi per il programma complessivo.
I primi 3 mesi saranno dedicati ad una fase di analisi ed approfondimento dei modelli esistenti in letteratura, e dei risultati raggiunti finora. Durante tale periodo il candidato dovrà acquisire competenze eterogenee, legate agli algoritmi esistenti per la soluzione di sistemi lineari e alle tecniche allo stato dell'arte per la gestione dei guasti.
Seguiranno 4 mesi dedicati al miglioramento delle soluzioni algorithm-based esistenti (con particolare riferimento al Metodo delle Interdizioni),e all'estensione di tali approcci con la gestione dei soft faults.
Gli ultimi 5 mesi saranno dedicati al progetto di una libreria per la gestione trasparente dei guasti a livello di algoritmo o a livello di sistema.

Attività con altri Enti (italiani ed esteri)

L'attività di progetto qui descritta prevede la collaborazione con l'*Ente per le nuove tecnologie, l'energia e l'ambiente (ENEA), ed in particolare col centro di ricerca di Bologna.*

Research project – post doc position 1 year

Title: Fault management techniques in HPC Systems.

Supervisor: Prof. Anna Ciampolini, Ph.D.

Motivations

Modern supercomputers, or High Performance Computing (HPC) systems are composed of hundreds of thousands of computing nodes contained in traditional CPUs or GP-GPUs. Their purpose is often linked to particularly onerous computations in the scientific or industrial field, which would take too long if performed on common hardware. The architecture of these supercomputers differs considerably from that of traditional computers, and the problems that may arise during computation also often have a different, generally more complex nature.

An example is the frequency of hardware errors.

In fact, although the average time between the occurrence of two successive failures on a traditional computer has progressively increased in the last decade, , the number of processing units in HPC systems is so high that in these processing systems a decrease in the average time between two faults, of the order of a few hours, has often been observed.

For this reason, the problem of detecting and managing failures in the HPC field has become more important in recent years. Some solutions focus on the so-called hard-faults, i.e. failures affecting one or more processing nodes causing their complete stop and consequent loss of data. Other errors, called soft-faults, are instead caused by external events (such as, for example, thermal alterations or radiation) whose effect may be the alteration of the value of a few bits. These small changes in the value of data processed by the faulted node, depending on the type of computation in progress, can be propagated and amplified, leading to significant errors on the results of the computation.

Objectives

The research will focus on the study of methods for fault handling in HPC systems.

In particular, the analysis will start from the results obtained so far in the solution of linear systems with the parallelization and modification aimed at hard-faults tolerance of a solver called the Inhibition Method (IMe). Starting from these results, the research activity will investigate on methods to improve the efficiency of the techniques already proposed and decrease the computation/communication overhead introduced by the fault tolerance system.

The research will also study soft-faults handling and to the possibility of applying similar techniques in the solution of other linear algebra problems typically required in the scientific field, such as the calculation of eigenvalues or the product between matrices.

The research will also explore the possibility of applying techniques for the detection, management and recovery of faults at the system level, increasing the generality, flexibility and transparency of the proposed solution.

The objectives of the activity can be summarized as follows:

1) Analysis and study of the state of the art in the context of the solution of linear systems with tolerance to hard-fault type failures.

2) Identification of strategies to increase the efficiency of existing solutions and development of solutions that integrate soft-fault tolerance

3) Extension of fault management to other areas through the application of techniques at the algorithm level (i.e., adaptation of the state of the art to other linear algebra computations) or at the system level.

Research Plan

Activity plan
The position has a duration of 12 months.

The first 3 months will be dedicated to study of the existing models in the literature, and of the results achieved so far. During this period, the candidate will have to acquire heterogeneous skills, dealing with both existing algorithms for the solution of linear systems and state-of-the-art techniques for fault management.
Following 4 months will be devoted to the improvement of existing algorithm-based solutions (with particular reference to IMe), and to the extension of these approaches with the management of soft faults.
The last 5 months will be dedicated to the project of a library for the transparent management of faults at both algorithm or system level.

Collaboration with other organizations:
The project activity described here provides for collaboration with the l'Ente per le nuove tecnologie, l'energia e l'ambiente (ENEA), and in particular with the Bologna research center.

References

[1] Artioli, Marcello; Loreti, Daniela; Ciampolini, Anna, *Fault Tolerant High Performance Solver for Linear Equation Systems*, in: 2019 38th Symposium on Reliable Distributed Systems (SRDS), 2019, pp. 113 - 122 (Proceedings of 38th Symposium on Reliable Distributed Systems (SRDS), Lione, Francia.