

July 2017

Implications of Reduced-Precision Computations in HPC: Performance, Energy and Error

Stefano Cherubin^a Giovanni Agosta^a Imane Lasri^b
Erven Rohou^b and Olivier Sentieys^b

^a *DEIB - Politecnico di Milano, name.surname@polimi.it*

^b *INRIA, name.surname@inria.fr*

Abstract. Error-tolerating applications are increasingly common in the emerging field of real-time HPC. Proposals have been made at the hardware level to take advantage of inherent perceptual limitations, redundant data, or reduced precision input [15], as well as to reduce system costs or improve power efficiency [14]. At the same time, works on floating-point to fixed-point conversion tools [6] allow us to trade-off the algorithm exactness for a more efficient implementation. In this work we aim at leveraging existing, HPC-oriented hardware architectures, while including in the precision tuning an adaptive selection of floating and fixed point arithmetic.

Our proposed solution takes advantage of the application domain knowledge of the programmer by involving them in the first step of the interaction chain. We rely on annotations written by the programmer on the input file to know which variables should be converted to fixed-point. The second stage replaces the floating point variables in the kernel with fixed-point equivalents. It also adds to the original source code the utility functions to perform data type conversions from floating-point to fixed-point, and vice versa. The output of the second stage is a new version of the kernel source code which exploits fixed-point computation instead of floating point computation.

We discuss the effect of our solution in terms of time-to-solutions, error and energy-to-solution.

Keywords. Approximate Computing, Fixed Point, Compilers

1. Introduction

High Performance Computing (HPC) has been traditionally the domain of grand scientific challenges and industrial sector such as oil & gas or finance, where investments are large enough to support massive computing infrastructures. Nowadays HPC is recognized as a powerful technology to increase the competitiveness of nations and their industrial sectors, including small scale but high-tech businesses – *to compete, you must compute* has become an ubiquitous slogan [2].

The current roadmap for HPC systems aims at reaching the Exascale level (10^{18} FLOPS) within the 2023 – 24 timeframe – with a $\times 1000$ improvement over