Thesis abstract Roméo Molina

« "Mixed precision for High Performance Computing, application to low energy gamma radiation measurements"»

In this thesis, we focus on numerical precision formats, on the control of numerical accuracy of the results and on the opportunities they offer in the context of High Performance Computing. More specifically, we are interested in mixed precision, which consists in mixing several precision formats in the same code to take advantage of both the performance gains of low precision and the validity and stability of high precision. Here, we examine two approaches to introduce mixed precision. On the one hand, precision tuning, which consists in developing decision-support tools to introduce lower precisions than the original ones, into an existing code, while checking the validity of the results thus obtained. On the other hand, the development of linear algebra algorithms that are intrinsically mixed in terms of precision, which can then be used as building blocks for various applications. In this manuscript, we focus on one application in particular: nuclear physics research, i.e. the study of particles and the interactions that govern this scale. This is often achieved by studying extreme values, particularly on highly unstable nuclei, using high-resolution detectors. AGATA is a European collaboration to develop a High-Purity Germanium gamma-ray detector. It is based on the reconstruction of the complete path of a gamma-ray in the detector. To do this, there is a Pulse-Shape Analysis (PSA) that consists in comparing the traces measured in each segment with those of a database previously calculated or obtained by calibrating the crystal, to identify the interaction points and their associated energies. The quantity of data implies an on-line processing, but this step must also be carried out accurately, as a resolution of 5mm is required. We therefore sought to speed up this step by reducing the volume of data, using formats of reduced precision. To check the validity of the results obtained, we used stochastic arithmetic and evaluated the number of points identified by the different methods. I n this way, we demonstrated that the original algorithm could be run without loss of quality when using FP16 rather than FP32. We also adapted the algorithm to GPU architecture, showing positive results and encouraging the choice of this kind of hardware for the PSA. We also worked on the Sparse Matrix-Vector product (SpMV), developing a mixed-precision version of this algorithm. This is based on a rigorous analysis that divides elements into buckets and computes them with a precision inversely proportional to their magnitude. This algorithm not only guarantees target accuracy, but also allows the use of several precision formats, both native and emulated. This algorithm delivers significant gains in memory and execution time, but was limited in its use of non-standard precision formats due to their lack of hardware implementation. We therefore decided to integrate optimized accessors within the adaptive matrix-vector product and developed new formats using a reduced exponent taking advantage of the small variation in magnitude within each bucket.