Résumé de thèse Roméo Molina

« Précision mixte pour le calcul haute performance, application aux mesures de radiation gamma de faible énergie»

Dans le cadre de cette thèse, nous nous intéressons aux formats de précision numérique, au contrôle de la validité de résultats numériques et aux opportunités qu'ils offrent dans le cadre du calcul haute performance.
Plus précisément, nous nous intéressons à la précision mixte, qui consiste à mêler plusieurs formats de précision dans un même code pour tirer partie à la fois des gains de performances apportés par les précisions faibles et de validité et de la stabilité des précisions élevées. Nous étudions ici deux approches visant à introduire de la précision mixte. D'une part le tuning de précision qui consiste à développer des outils d'aide à la décision permettant d'introduire des précisions plus faibles que celles d'origine, dans un code existant, tout en effectuant un contrôle sur la validité des résultats ainsi obtenus. D'autre part, le développement d'algorithmes d'algèbre linéaire intrinsèquement mixtes en termes de précision pouvant ensuite être utilisés comme des briques de bases pour des applications diverses.

Dans ce manuscrit, nous nous intéressons à une application en particulier : la recherche en physique nucléaire, c'est-à-dire l'étude des particules et des interactions qui régissent cette échelle.
Cet objectif est souvent atteint par l'étude de valeurs extrêmes, en particulier sur des noyaux très instables à l'aide de détecteurs à haute résolution.
AGATA est une collaboration européenne visant à mettre au point un détecteur de rayons gamma au Germanium de Haute Pureté. Celui-ci s'appuie sur deux nouvelles technologies :
la segmentation électrique des cristaux de Germanium et la reconstruction du parcours complet d'un rayon dans le détecteur. Pour cela, une étape d'analyse de la forme des traces mesurées
dans chaque segments avec ceux d'une base de donnée préalablement calculée ou obtenue par calibration du cristal permet d'identifier les points d'interaction et leurs énergies associées.
La quantité de données mesurée implique un traitement en direct mais cette étape doit aussi être réalisée avec précision car une résolution de 5mm est requise. Nous avons donc cherché à accélérer cette étape en réduisant le volume des données en utilisant des formats de précisions réduite. Afin de vérifier le maintien de la validité des résultats obtenus, nous nous sommes appuyés sur l'arithmétique stochastique mais aussi sur une évaluation du nombre de points identifiés pareillement par les différentes méthodes. Nous avons ainsi mis en évidence que l'exécution de l'algorithme d'origine pouvait se faire sans perte de qualité en FP16 plutôt qu'en FP32. Nous avons également effectué une réécriture de l'algorithme pour l'adapter à l'architecture GPU, montrant des résultats positifs et incitant à choisir ce type de matériel pour effectuer cette étape.

Nous avons également réalisé un travail sur le produit matrice-vecteur creux en développant une version en précision mixte de cet algorithme. Celui-ci s'appuie sur une analyse rigoureuse qui permet de répartir les éléments en buckets et de les calculer dans une précision inversement proportionnelle à leur magnitude. Cet algorithme permet de garantir une précision cible mais aussi d'utiliser plusieurs formats de précision qu'ils soient natifs ou émulés.
Cet algorithme permet d'obtenir des gains très importants en mémoire et en temps d'exécution mais se trouvait limité dans son utilisation des formats de précision non standards du fait de leur absence d'implémentation hardware.
Nous avons donc décidé d'intégrer des accesseurs optimisés au sein du produit matrice-vecteur adaptatif et développé de nouveaux formats utilisant un exposant réduit tirant partie de la faible variation de magnitude au sein de chaque bucket.