Task 4
Precision auto-tuning and verified computing

Projet Interflop

juin 2023

Etat d'avancement

- Résolution de systèmes linéaires par raffinement itératif sur GPUs ✓
  - Couplage avec raffinement itératif de l'algorithme de factorisation LU sur GPU Tensor Cores de Lopez & Mary (Florent Lopez & Theo Mary. Mixed Precision LU Factorization on GPU Tensor Cores: Reducing Data Movement and Memory Footprint, Int. J. High Perform. Comput. Appl., 2023).
  - Gain d'un facteur 2 en temps et stockage par rapport à l'état de l'art
  - Résultats présentés à SIAM PP 2022
- Calcul fiable de racines de polynômes en précision arbitraire ✓
  - Version stochastique des itérations de Newton, du PGCD polynomial, de la division euclidienne polynomiale
  - S. Graillat, F. Jézéquel, E. Queiros Martins, M. Spyropoulos, Computing multiple roots of polynomials in stochastic arithmetic with Newton method and approximate GCD, SYNASC 2021. `http://hal.archives-ouvertes.fr/hal-03274453`

Etat d'avancement

- Reproductibilité de l'algo BiCGStab ✓
  - Xiaojun Lei, Tongxiang Gu, Stef Graillat, Xiaowen Xu & Jing Meng. Comparison of Reproducible Parallel Preconditioned BiCGSTAB Algorithm Based on ExBLAS and ReproBLAS, HPC Asia'23.
  - Roman Iakymchuk, Stef Graillat & José Ignacio Aliaga. General framework for deriving reproducible Krylov subspace algorithms: A BiCGStab case study, PPAM 2022.
- Survey sur l'arrondi stochastique ✓
  - Matteo Croci, Massimiliano Fasi, Nicholas Higham, Theo Mary & Mantas Mikaitis. Stochastic Rounding: Implementation, Error Analysis, and Applications, Roy. Soc. Open Sci., 9(3), 1–25 (2022).
- Survey sur les algos d'algébre linéaire en précision mixte ✓
  - Nicholas Higham & Theo Mary. Mixed precision algorithms in numerical linear algebra, Acta Numerica, 31:347–414 (2022).

Etat d'avancement

- Produit matrice-vecteur creux en précision adaptative et application aux solveurs de Krylov (réalisé/perspectives)
  - En fonction des éléments de la matrice, sommes partielles évaluées dans différentes précisions
  - Par rapport à la précision uniforme, speed-up (jusqu'à $7\times$) et gain en stockage (jusqu'à $36\times$) pour une précision comparable
  - Pour des précisions cibles raisonnables, le produit matrice-vecteur creux adaptatif n'affecte pas la converge des méthodes de Krylov testées (GMRES, BiCGstab, CG)
  - Jusqu'à 7 précisions testées : bfloat16, fp24, fp32, fp40, fp48, fp56, fp64
    conversions non optimisées (optimisation à venir, en collaboration avec RIKEN : séjour de Roméo Molina à RIKEN cet été)
  - S. Graillat, F. Jézéquel, T. Mary, R. Molina, Adaptive precision matrix-vector product and its application to Krylov solvers, 2022. `https://hal.science/hal-03561193`

Etat d'avancement

- "Instrumentizer" : modification de sources C/C++ pour les implémentations en précision mixte (réalisé/perspectives)
  - Auparavant :
    - utilisation dans PROMISE d'un parser "maison" pour identifier les variables et leurs types, les types de retour des fonctions.
    - cadnaizer: script PERL
  - Travail réalisé :
    - Utilisation du parser C de Clang
    - Extraction à partir de l'AST (Abstract Syntax Tree) des variables et de leurs types
    - Génération d'un code avec des types modifiés
    - Utilisation de cet intrumentizer dans PROMISE (autotuning de précision)
      ⇒ Améliorations de PROMISE :
      utilisation de la précision mixte dans des portions de code
    - Nouveau cadnaizer
  - Perspective :
    - prendre en compte d'autres outils (FLDLib ?)

Etat d'avancement

- Optimisation de la précision dans les réseaux de neurones via PROMISE (réalisé/en cours)
  - Modèles Keras/Pytorch → codes C++ en précision mixte
  - Résultats présentés pour 4 réseaux de neurones dont MNIST et CIFAR : configurations de types obtenues, temps pris par l'autotuning

    Q. Ferro, S. Graillat, T. Hilaire, F. Jézéquel, B. Lewandowski, Neural Network Precision Tuning Using Stochastic Arithmetic, NSV'22.
    `https://hal.archives-ouvertes.fr/hal-03682645`
  - Mesure du gain en mémoire et en temps (sur les codes vectorisés)
- Parallélisation de la recherche de configurations dans PROMISE ✓
- Autotuning de codes HPC avec PROMISE ✓
  - codes MPI
  - codes OpenMP
- Comparaisons avec les travaux de Y. Fakhreddine (Perpignan) sur l'autotuning de calculs itératifs fondé sur LLVM

Etat d'avancement

- Combinaison de NSAN/INSANE et de l'arithmétique stochastique (en cours)
  - Utilisation de la "shadow memory" pour implanter l'arithmétique stochastique
  - Echanges avec Mathys Jam sur MCASync
- Introduction des FPInt (en cours)
  - Combinaison de FP-ANR et intervalles : précision représentée par le bit à 1 le plus à droite dans un seul flottant
  - Algorithme de conversion FPInt vers intervalle

Perspectives à plus long terme

- Etude de la différentiation automatique pour la validation numérique et l'auto-tuning de précision
  Pour chaque variable, dérivée du calcul $\approx$ conditionnement
  conditionnement et précision de travail $\rightarrow$ précision de la variable

- Proposition et implantation d'algorithmes d'auto-tuning en précision arbitraire
  - Exploration de la différentiation automatique pour accélérer la recherche des configurations possibles

- Conversion automatique en précision mixte des noyaux d'algèbre linéaire au sein d'outils d'auto-tuning de précision

Petites annonces

Offre

PostDoc de 2 ans (PEPR NumPEx) :
Precision auto-tuning and numerical validation of high performance simulations
voir `http://www.lip6.fr/Fabienne.Jezequel`

Demande

Recherche d'appli à la suite de NSV'20 (F. Jézéquel, S. Graillat, D. Mukunoki, T. Imamura, R. Iakymchuk, Can we avoid rounding-error estimation in HPC codes and still get trustworthy results?, NSV'20)

En résumé :

- Données perturbées (par ex issues de calculs)
- Remplacement de CADNA par 3 exécutions
- Perfs ☺ pour les multiplications de matrices

⇒ Recherche d'application : code/algo avec multiplications de matrices ?