Compas'2023 : Parallélisme/ Architecture / Système
LISTIC/USMB - Annecy France, 4-7 juillet 2023

Conception d'un sous-système mémoire pour traitement de données creuses

Valentin Isaac--Chassande *

Univ. Grenoble Alpes, CEA, List, F-38000 Grenoble, France
valentin.isaacchassande@cea.fr

Résumé
Démarrée en octobre 2022, cette thèse a pour objectif de *concevoir un sous-système mémoire pour traitement de données creuses*, c'est-à-dire pour traiter des structures de données possédant de nombreux éléments nuls. Ce sujet s'inscrit dans l'intérêt croissant porté aux solutions matérielles et plus particulièrement aux accélérateurs spécialisés, pour améliorer les performances de calcul [8]. Par ailleurs, de nombreux problèmes de calcul haute performance dépendent de la résolution d'opérations de multiplication sur de très larges matrices creuses [1, 4, 6, 9, 14]. De telles matrices sont souvent compressées via des formats spécifiques pour réduire l'empreinte mémoire et éviter les opérations inutiles (multiplications par zéro), au détriment de l'efficacité de l'utilisation de la bande passante mémoire qui subit les coûts des accès indirectes induit par les formats creux [3, 5, 10]. Ainsi, l'objectif de la thèse est de proposer une solution matérielle spécialisée améliorant l'efficacité de la bande passante mémoire pour du calcul hautes performances, avec des matrices creuses de grandes dimensions.
Une étude approfondie de l'état de l'art a permis, entre autre, de rassembler et d'étudier différents accélérateurs existants ayant le potentiel d'améliorer significativement les calculs avec données creuses [2, 7, 11–13, 15–17]. Il est récurent que les auteurs utilisent des notations et bases de comparaison différentes pour avancer leurs solutions. En effet, chacune d'entre elles utilisent des méthodes, technologies, formats creux et algorithmes qui diffèrent en partie. Remarquant ainsi un manque d'analyse comparative, les premiers mois de recherches de la thèse s'inscrivent donc dans l'objectif de classifier ces accélérateurs en rassemblant des caractéristiques communes, pour pouvoir donner aux architectes une base de comparaison ainsi que les outils nécessaires pour concevoir des architectures efficaces dans le domaine des matrices creuses. Dans le cadre de la thèse, cet objectif permet également de se positionner pour proposer une nouvelle solution matérielle.
Pour la suite du doctorat, l'objectif est de proposer des améliorations à des solutions existantes à intégrer dans un nouvel accélérateur, en prenant en considération les besoins en bande passante et les contraintes de densité et dimensions des données. En particulier, un premier axe de travail sera porté sur la conception d'une micro-architecture de type mémoire cache avec accumulations de données in situ.

Mots-clés : Calcul Haute Performance (HPC), Accélérateurs, Bande-passante mémoire, Matrices creuses.

---

*. Le texte a été relu par Adrian Evans, Frédéric Rousseau et Yves Durand

Bibliographie

1. Azad (A.), Buluç (A.) et Gilbert (J.). – Parallel triangle counting and enumeration using matrix algebra. – In *2015 IEEE International Parallel and Distributed Processing Symposium Workshop*, pp. 804–811, May 2015.
2. Barredo (A.), Beard (J. C.) et Moretó (M.). – Poster : Spidre : Accelerating sparse memory access patterns. – In *2019 28th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, pp. 483–484, New York, NY, USA, 2019. Institute of Electrical and Electronics Engineers.
3. Cong (J.), Fang (Z.), Lo (M.), Wang (H.), Xu (J.) et Zhang (S.). – Understanding performance differences of fpgas and gpus. – In *2018 IEEE 26th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, pp. 93–96, April 2018.
4. Gilbert (J. R.), Reinhardt (S.) et Shah (V. B.). – High-performance graph algorithms from parallel sparse matrices. – In Kågström (B.), Elmroth (E.), Dongarra (J.) et Waśniewski (J.) (édité par), *Applied Parallel Computing. State of the Art in Scientific Computing*, pp. 260–269, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.
5. Goumas (G.), Kourtis (K.), Anastopoulos (N.), Karakasis (V.) et Koziris (N.). – Understanding the performance of sparse matrix-vector multiplication. – In *16th Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP 2008)*, pp. 283–292, Feb 2008.
6. Han (S.), Mao (H.) et Dally (W. J.). – Deep compression : Compressing deep neural networks with pruning, trained quantization and huffman coding, 2016.
7. Hegde (K.), Asghari-Moghaddam (H.), Pellauer (M.), Crago (N.), Jaleel (A.), Solomonik (E.), Emer (J.) et Fletcher (C. W.). – Extensor : An accelerator for sparse tensor algebra. – In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, MICRO '52*, MICRO '52, p. 319–333, New York, NY, USA, 2019. Association for Computing Machinery.
8. Hennessy (J.) et Patterson (D.). – A new golden age for computer architecture : Domain-specific hardware/software co-design, enhanced security, open instruction sets, and agile chip development. – In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, pp. 27–29, June 2018.
9. Itoh (S.), Ordejón (P.) et Martin (R. M.). – Order-n tight-binding molecular dynamics on parallel computers. *Computer Physics Communications*, vol. 88, n2, 1995, pp. 173–185.
10. Matam (K.), Krishna Bharadwaj Indarapu (S. R.) et Kothapalli (K.). – Sparse matrix-matrix multiplication on modern architectures. – In *2012 19th International Conference on High Performance Computing*, pp. 1–10, Dec 2012.
11. Mukkara (A.), Beckmann (N.) et Sanchez (D.). – Phi : Architectural support for synchronization- and bandwidth-efficient commutative scatter updates. – In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, MICRO '52*, MICRO '52, p. 1009–1022, New York, NY, USA, 2019. Association for Computing Machinery.
12. Pal (S.), Beaumont (J.), Park (D.-H.), Amarnath (A.), Feng (S.), Chakrabarti (C.), Kim (H.-S.), Blaauw (D.), Mudge (T.) et Dreslinski (R.). – Outerspace : An outer product based sparse matrix multiplication accelerator. – In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 724–736, New York, NY, USA, 2018. Institute of Electrical and Electronics Engineers.
13. Srikanth (S.), Jain (A.), Conte (T. M.), Debenedictis (E. P.) et Cook (J.). – Sortcache : Intelligent cache management for accelerating sparse data workloads. *ACM Trans. Archit. Code Optim.*, vol. 18, n4, sep 2021.
14. Yamazaki (I.) et Li (X. S.). – On techniques to improve robustness and scalability of a pa-

rallel hybrid linear solver. – In Palma (J. M. L. M.), Daydé (M.), Marques (O.) et Lopes (J. C.) (édité par), *High Performance Computing for Computational Science – VECPAR 2010*, pp. 421–434, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

15. Zhang (C.), Bremer (M.), Chan (C.), Shalf (J.) et Guo (X.). – Asa : Accelerating sparse accumulation in column-wise spgemm. *ACM Trans. Archit. Code Optim.*, vol. 19, n4, sep 2022.
16. Zhang (G.), Attaluri (N.), Emer (J. S.) et Sanchez (D.). – Gamma : Leveraging gustavson's algorithm to accelerate sparse matrix multiplication. – In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, *ASPLOS '21*, ASPLOS '21, p. 687–701, New York, NY, USA, 2021. Association for Computing Machinery.
17. Zhang (Z.), Wang (H.), Han (S.) et Dally (W. J.). – Sparch : Efficient architecture for sparse matrix multiplication. *CoRR*, vol. abs/2002.08947, 2020.