CCF Transactions on High Performance Computing (2023) 5:99–101
https://doi.org/10.1007/s42514-023-00157-9

EDITORIAL

Editorial for the special issue on architecture, algorithms and applications of high performance sparse matrix computations

Weifeng Liu[1] · Guangming Tan[2] · Xiaowen Xu[3]

Published online: 19 June 2023

Sparse matrix operations are widely used in computational science and engineering applications such as quantum chemistry and finite element analysis, as well as modern machine learning scenarios such as social network and compressed deep neural networks. The University of California, Berkeley in the famous article 'A View of the Parallel Computing Landscape', Asanovic et al. (2009) listed sparse matrix computations as one of the most important parallel computing patterns. In recent decades, how to use massively parallel computing platforms for highly scalable, highly performant, and highly practical sparse matrix computations has been a challenging open problem.

We have eight invited papers selected for this special issue based on a peer-review procedure, which cover several different aspects that related to architecture, algorithms and applications of high performance sparse matrix computations mentioned above.

The first part of the special issue focuses on exploring new architectural and compilation techniques for matrix computations. The two papers propose a fast matrix multiplication architecture on field programmable gate array (FPGA), and several compilation optimizations to sparse tensor algebra, respectively.

- In the first paper, Bessant et al. (2023) propose a parallel multiplication architecture using Strassen and UrdhvaTiryagbhyam multiplier, which involves design of efficient parallel matrix multiplication with flexible implementation of FPGA devices. The architecture incorporates scheduling of blocks, operations on processing elements, block size determination, parallelization and double buffering for storage of matrix elements.
- In the second paper, Zhang et al. (2023) present a strategy called Sgap considering segment group and atomic parallelism, which can resolve two challenges about how to elevate the flexible reduction semantics to sparse tensor algebra compilation: (1) there are wasted parallelism by adopting static synchronization granularity, and (2) static reduction strategy limits optimization space exploration. They use GPU-accelerated sparse matrix-dense matrix multiplication (SpMM) as a use case to demonstrate the effectiveness of segment group in reduction semantics elevation, and achieve obvious speedups over existing work.

The second part of the special issue focuses on the optimization techniques in sparse algorithms. The three papers propose new optimization strategies for eigenvalue problem, sparse triangular solve, and sparse approximate inverse preconditioning, respectively.

- In the first paper, Li et al. (2022) propose a novel parallel structured divide-and-conquer (DC) algorithm for symmetric banded eigenvalue problems, denoted by PBSDC, which computes the eigenpairs directly without tridiagonalization. They compare the work with PBDC and ELPA through numerous experiments on Tianhe-2 supercomputer. For matrices with many deflations and/or small bandwidths, PBSDC can be obviously faster than the tridiagonalization-based DC implemented in LAPACK and ELPA.

✉ Weifeng Liu
weifeng.liu@cup.edu.cn

Guangming Tan
tgm@ict.ac.cn

Xiaowen Xu
xwxu@iapcm.ac.cn

[1] Super Scientific Software Laboratory, China University of Petroleum-Beijing, Beijing 102249, China

[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

[3] Institute of Applied Physics and Computational Mathematics, Beijing 100088, China

- In the second paper, Lu and Liu (2023) propose a tiled algorithm called TileSpTRSV for optimizing SpTRSV on GPUs through exploiting 2D spatial structure of sparse matrices. They design two algorithm implementations, i.e., TileSpTRSV level-set and TileSpTRSV sync-free, on top of level-set and sync-free schemes, respectively. By testing a group of representative matrices, their experimental results show excellent performance compared to cuSPARSE, Sync-free and Recblock algorithms.
- In the third paper, Gao et al. (2023) present a new heuristic sparse approximate inverse (SPAI) preconditioning algorithm on GPUs, called HeuriSPAI. HeuriSPAI fuses the advantages of static and dynamic SPAI preconditioning algorithms, and alleviates the drawback of the existing dynamic SPAI preconditioning methods not suitable for large matrices. The experimental results show that HeuriSPAI outperforms the popular preconditioning algorithms in three public libraries, i.e., cuSPARSE, MAGAMA and ViennaCL, as well as a recent parallel static SPAI preconditioning algorithm.

The third part of the special issue focuses on applications of sparse matrix computations. The first two studies propose software packages for solving large scale eigenvalue problems and sparse linear equations, respectively. The third one proposes an improved multistage preregulator for compositional reservoir simulation.

- In the first paper, Li et al. (2023) introduce several strategies to improve the efficiency and scalability of the generalized conjugate gradient algorithm and build a package GCGE for solving large scale eigenvalue problems. This method is the combination of damping idea, subspace projection method and inverse power algorithm with dynamic shifts. Numerical results are provided to demonstrate the efficiency, stability and scalability of this work for computing many eigenpairs of large symmetric matrices arising from applications.
- In the second paper, Li et al. (2023) use error-free transformation technology and mixed-precision ideas to construct a reliable parallel numerical algorithm framework based on HYPRE, which solves large-scale sparse linear equations to improve accuracy and accelerate numerical calculations. Experimental results demonstrate that XHYPRE has higher reliability and effectiveness over HYPRE, and reduces the number of iterations.
- In the third paper, Zhao et al. (2023) develop an efficient multistage preconditioner for the fully implicit compositional flow simulation. The method employs an adaptive setup phase to improve the parallel efficiency on GPUs. Furthermore, a multicolor Gauss-Seidel algorithm is applied in the algebraic multigrid methods for the pressure part. Numerical results demonstrate that the proposed method achieves good speedups while yielding the same convergence behavior.

We would like to take this chance to thank all the authors and the reviewers for their splendid contribution to this special issue of CCF THPC.

References

Asanovic, K., Bodik, R., Demmel, J., Keaveny, T., Keutzer, K., Kubiatowicz, J., Morgan, N., Patterson, D., Sen, K., Wawrzynek, J., Wessel, D., Yelick, K.: A view of the parallel computing landscape. Commun. ACM **52**(10), 56–67 (2009). https://doi.org/10.1145/1562764.1562783

Bessant, Y.R.A., Jency, J.G., Sagayam, K.M., Jone, A.A.A., Pandey, D., Pandey, B.K.: Improved parallel matrix multiplication using Strassen and Urdhvatiryagbhyam method. CCF Trans. High Perform. Comput. (2023). https://doi.org/10.1007/s42514-023-00149-9

Gao, J., Chu, X., Wang, Y.: HeuriSPAI: a heuristic sparse approximate inverse preconditioning algorithm on GPU. CCF Trans. High Perform. Comput. (2023). https://doi.org/10.1007/s42514-023-00142-2

Li, S., Liao, X., Lu, Y., Roman, J.E., Yue, X.: A parallel structured banded DC algorithm for symmetric eigenvalue problems. CCF Trans. High Perform. Comput. (2022). https://doi.org/10.1007/s42514-022-00117-9

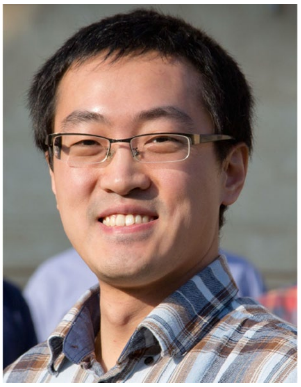
Li, Y., Wang, Z., Xie, H.: GCGE: a package for solving large scale eigenvalue problems by parallel block damping inverse power method. CCF Trans. High Perform. Comput. (2023). https://doi.org/10.1007/s42514-023-00135-1

Li, C., Graillat, S., Quan, Z., Gu, T.-X., Jiang, H., Li, K.: XHYPRE: a reliable parallel numerical algorithm library for solving large-scale sparse linear equations. CCF Trans. High Perform. Comput. (2023). https://doi.org/10.1007/s42514-023-00141-3

Lu, Z., Liu, W.: "TileSpTRSV: a tiled algorithm for parallel sparse triangular solve on GPUs. CCF Trans. High Perform. Comput. (2023)

Zhang, G., Zhao, Y., Tao, Y., Yu, Z., Dai, G., Huang, S., Wen, Y., Petoumenos, P., Wang, Y.: Sgap: towards efficient sparse tensor algebra compilation for GPU. CCF Trans. High Perform. Comput. (2023). https://doi.org/10.1007/s42514-023-00140-4

Zhao, L., Li, S., Zhang, C.-S., Feng, C., Shu, S.: An improved multistage preconditioner on GPUs for compositional reservoir simulation. CCF Trans. High Perform. Comput. (2023). https://doi.org/10.1007/s42514-023-00136-0



Weifeng Liu is a Professor from Super Scientific Software Laboratory, China University of Petroleum-Beijing. His main research interests are numerical linear algebra and parallel computing, with a particular focus on data structures, parallel algorithms and software for sparse matrices. He has served as a TPC members of SC, ICS, IPDPS, ICPP, etc. His research work has been published in major international conferences and journals such as SC, PPoPP, ASPLOS, DAC, ICS, IPDPS, TPDS, JPDC, etc.

Guangming Tan is a professor from the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include parallel programming and algorithm, domain-specific architecture, and bioinformatics. He has served as associate editor for IEEE TPDS and PC members for SC, PPoPP, ICS. He has published papers including conference/ journals like SC, PLDI, PPoPP and IEEE/ ACM Transactions.

Xiaowen Xu is a Professor of Institute of Applied Physics and Computational Mathematics (IAPCM). He is the director of High Performance Computing Center at IAPCM. He got his B.S degree from Xiangtan University in 2002, and his PhD degree in computational mathematics from Chinese Academy of Engineering Physics in 2007. His research interests include high performance numerical algorithm & software in scientific and engineering fields, parallel programming framework for large-scale numerical simulations. He is member of CCF and member of SIAM and CSIAM.