Guest editorial: Special issue on parallel matrix algorithms and applications (PMAA'16)

Emmanuel Agullo, Peter Arbenz, Luc Giraud, Olaf Schenk

Guest editorial

Special issue on Parallel Matrix Algorithms and Applications (PMAA'16)

This special issue of Parallel Computing contains nine articles, selected after peer reviewing, from invited and contributed presentations made at the 8th International Workshop on Parallel Matrix Algorithms and Applications (PMAA'16), that took place at the Université of Bordeaux, France, from July 6-8, 2016. The workshop attracted around 120 participants from all continents, 25% were PhD students and around 10% from industry.

The workshop was co-chaired by Emmanuel Agullo, Peter Arbenz, Luc Giraud, and Olaf Schenk. The members of the program committee were : P. D'Ambra, H. Avron , M. Bader, W. Bangerth, A. Basermann, C. Bekas, P. Bientinesi, R. Blaheta, M. Bollhoefer, E. Boman, G. Bosilca, A. Buttari, X.-C. Cai, H. Calandra, O. Coulaud, E. Chow, E. Darve, Z. Drmac, N. Emad, M. Faverge, E. Gallopoulos, D. Goeddeke, J. Gondahmed Samehzio, D. Gordon, L. Grigori, A. Guermouche, I. Gutheil, G. Houzeaux, T. Huckle, M. Knepley, R. Krause, J. Langou, H. Ltaief, K. Meerbergen, Y. U. Meier, E. Ng, G. Oksa, S. Petiton, P. Ramet, J. Roman, J. E. Roman, C. Petra, T. Sakurai, D. B. Szyld, D. Trystram, M. Tuma, B. Uçar, M. Vajtersic, W. Vanroose, P. Vasconcelos, R. Wyrzykowski. A total of twelve high quality submissions were received. In this special issue nine eventually accepted papers appear.

The nine papers address diverse aspects of linear algebra and high performance computing

1. Jack Dongarra, Mark Gates, Stanimire Tomov address accelerating the SVD two stage reduction and divide-and-conquer using GPUs. The increasing gap between memory bandwidth and computation speed motivates the choice of algorithms to take full advantage of today's high performance computers. For dense matrices, the classic algorithm for the SVD uses a one-stage reduction to bidiagonal form, which is limited in performance by the memory bandwidth. To overcome this limitation, a two-stage reduction to bidiagonal has been gaining popularity. As accelerators, such as GPUs and co-processors, are becoming increasingly widespread in high-performance computing, the authors present an accelerated SVD employing a two-stage reduction to bidiagonal as well as a parallelized and accelerated divide- and-conquer algorithm to solve the subsequent bidiagonal SVD. The new implementation provides a significant speedup compared to existing multi-core and GPU-based SVD implementations.

2. David E. Keyes, George Turkiyyah, Hatem Ltaief, Wajih Halim Boukaram present batched QR and SVD algorithms on GPUs with applications in Hierarchical matrix compression. The paper describes high performance implementations of the QR and the SVD of a batch of small matrices hosted on the GPU with applications in the compression of hierarchical matrices. In particular the authors detail multiple kernels based on the level of the GPU memory hierarchy in which the matrices can reside and show substantial speedups against streamed CUSolver SVDs.
3. Akira Imakura, Tetsuya Sakurai consider a complex moment-based parallel nonlinear eigensolver using the block communication-avoiding Arnoldi procedure. Complex moment-based parallel eigensolvers have been actively studied owing to their high parallel efficiency. In this paper, the authors propose a block SS–CAA method, which is a complex moment-based parallel nonlinear eigensolver that makes use of the block communication-avoiding Arnoldi procedure. Numerical experiments indicate that the proposed method has higher performance compared with traditional complex moment-based nonlinear eigensolvers.
4. Chao Chen, Eric F Darve, Erik G Boman, Hadi Pouransari, Sivasankaran Rajamanickam introduce a distributed-memory hierarchical solver for general sparse linear systems. The proposed fully algebraic algorithm exploits the low-rank structure of fill-in blocks during the Gaussian elimination process. The parallel implementation is based on data decomposition such that only local communication is needed for updating boundary information on every processor. The new hierarchical solver achieves an average speedup of 45 compared to its dense counterpart in factorizing three two-million-sized test problems on 256 cores of a supercomputer.
5. David Keyes, George Turkiyyah, Gustavo Chavez, Hatem Ltaief, Stefano Zampini discuss a distributed-memory fast solver for structured linear systems based on accelerated cyclic reduction. More particularly they describe a fast direct solver for rank-compressible block tridiagonal linear systems arising from the discretization of 3D elliptic operators. They provide a baseline for performance and applicability by comparing with the multifrontal method where hierarchical semi-separable matrices are used for compressing the fronts, and with algebraic multigrid. The proposed algorithm exhibits good strong and weak scaling in a distributed context and, as with any direct solver, is advantageous for problems that require the solution of multiple right-hand sides.
6. Chao Yang, Lin Lin, Mathias Jacquelin discuss a new technique named PSelInv - a distributed memory parallel algorithm for selected inversion for non-symmetric matrices. This paper generalizes the parallel selected inversion algorithm called PSelInv to sparse non-symmetric matrices. The PSelInv method computes selected elements of A^{-1}. The selection is confined by the sparsity pattern of the matrix A^T. PSelInv involves a large number of collective data communication activities within different processor groups of various sizes. In order to minimize idle time

and improve load balancing, tree-based asynchronous communication is used to coordinate all such collective communication. Numerical results demonstrate that PSelInv can scale efficiently to 6'400 cores for a variety of matrices.

7. George Karypis, Jongsoo Park, Shaden Smith present HPC formulations of optimization algorithms for tensor completion. Tensor completion is a powerful tool used to estimate or recover missing values in multi-way data. Tensor completion is most often accomplished via low-rank sparse tensor factorization, a computationally expensive non-convex optimization problem. In this work, the authors study three optimization algorithms that have been successfully applied to tensor completion. They evaluate their parallel formulations on a variety of real datasets on a modern supercomputer and demonstrate speedups through 16'384 cores. These improvements reduce time-to-solution from hours to seconds.
8. Alejandro Lamas DavinÌa, Jose E. Roman discuss MPI-CUDA parallel linear solvers for block-tridiagonal matrices in the context of SLEPc's eigensolvers for the computation of a few eigenpairs of a generalized eigenvalue problem $Au = \lambda Bu$. In this kind of computation, it is often necessary to solve a linear system of equations in each eigensolver iteration. The linear solve can be performed through an LU factorization or using preconditioned Krylov methods. In this work, the authors aim at comparing different direct linear solvers that can exploit the block-tridiagonal structure. A parallel implementation based on MPI is developed in the context of the SLEPc library. The use of GPU devices to accelerate local computations shows to be competitive for large block sizes.
9. Cristiano I. Malossi, Alessandro Curioni, Costas Bekas, Efstratios Gallopoulos, Vasileios Kalantzis, Yousef Saad present a scalable iterative dense linear system solver for multiple right-hand sides in data analytics. They describe a Parallel-Projection Block Conjugate-Gradient, a distributed iterative solver for the solution of dense and symmetric positive-definite linear systems with multiple right-hand sides with a particular emphasis on linear systems in uncertainty quantification. Numerical experiments on massively parallel architectures illustrate the performance of the proposed scheme in terms of efficiency and convergence rate on a 4 rack BG/Q with up to 65'536 cores using dense matrices of order as high as half a million unknowns and 800 right-hand sides

The editors of this special issue of Parallel Computing would like to thank all authors who submitted papers and the numerous reviewers whose expert advice was essential in the selection process.

Emmanuel Agullo, Inria, Bordeaux Sud-Ouest, France

Peter Arbenz , ETH Zürich, Computer Science Department, Zürich, Switzerland

Luc Giraud, Inria, Bordeaux Sud-Ouest, France

Olaf Schenk, Università della Svizzera italiana, Institute of Computational Science, Lugano, Switzerland