HiCMA: Hierarchical Computations on Manycore Architectures

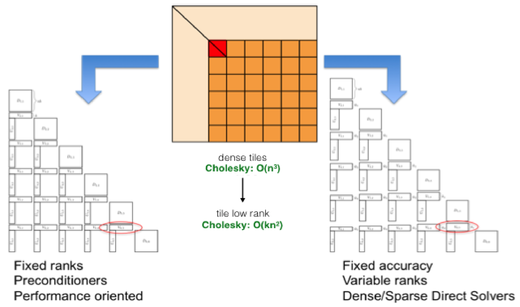
H. Ltaief, K. Akbudak, A. Charara, A. Mikhalev, S. Abdulah, and David Keyes
Extreme Computing Research Center
King Abdullah University of Science and Technology

Motivations

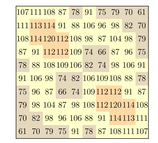
The Hierarchical Computations on Manycore Architectures (HiCMA) library aims to redesign existing dense linear algebra libraries to exploit the data sparsity of the matrix operator. Data sparse matrices arise in many scientific problems (e.g., in statistics-based weather forecasting, seismic imaging, and materials science applications) and are characterized by low-rank off-diagonal tile structure. Numerical low-rank approximations have demonstrated attractive theoretical bounds, both in memory footprint and arithmetic complexity. The core idea of HiCMA is to develop fast linear algebra computations operating on the underlying tile low-rank data format, while satisfying a specified numerical accuracy and leveraging performance from massively parallel hardware architectures.

Tile Low-Rank Algorithms

- Double Precision Floating-Point Arithmetic
- Matrix-Matrix Multiplication
- Cholesky Factorization/Solve
- Task-based Programming Models
- Support for StarPU Dynamic Runtime Systems
- Shared [2] and Distributed-Memory [4] Environments
- Testing Suite and Examples
- Available for download at: `https://github.com/ecrc/hicma`

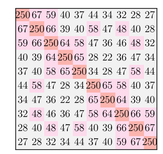


Matrix Kernels

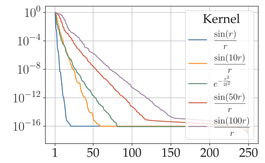


Synthetic λ=50.

Synthetic λ=100.



Real-world acc=10^{-8}.



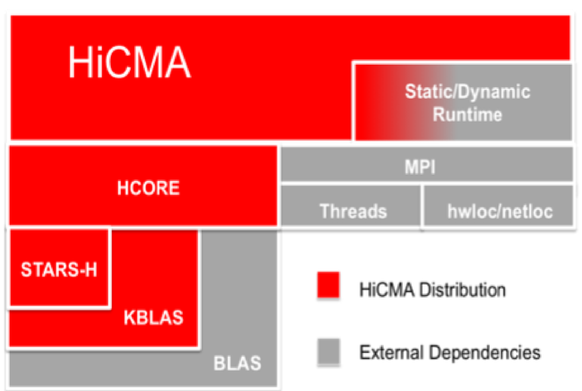
Singular values.

HiCMA 101

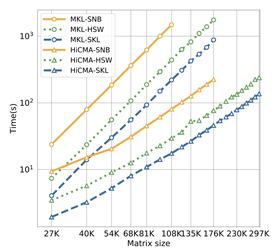
HiCMA's Methodology

1. Compress the large dense matrix using the off-diagonal tile low-rank approximation [1, 2].
2. Redesign the numerical algorithm so that it may operate on the compressed data structures.
3. Rely on the task-based programming model for fine-grained computations [3].
4. Employ a dynamic runtime system to ensure asynchronous executions, to maintain load balancing and to maximize hardware occupancy.
5. Ensure a separation of concerns by abstracting the hardware complexity from users.

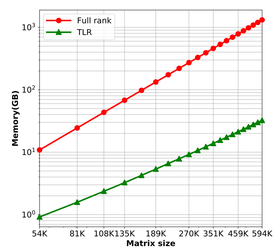
HiCMA's Software Stack



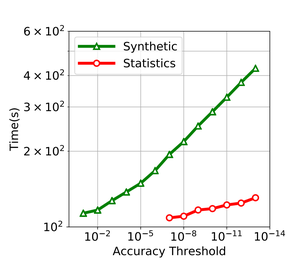
Performance Results



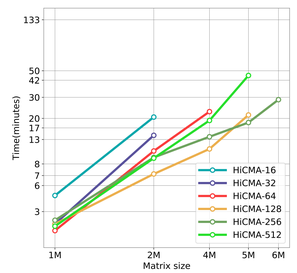
Synthetic ($\lambda = 100$).



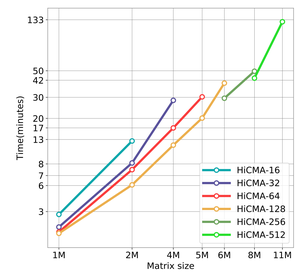
Memory footprint on 1M matrix size.



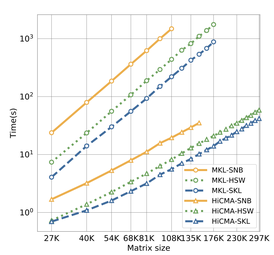
Accuracy impact on performance.



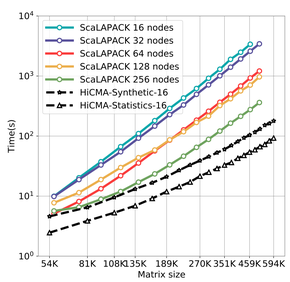
Synthetic, λ=50, Shaheen-2.



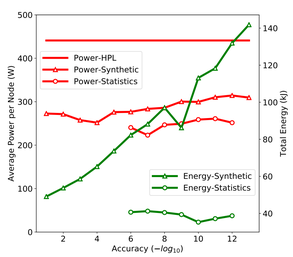
Statistics, Shaheen-2.



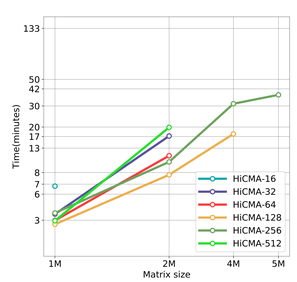
Statistics.



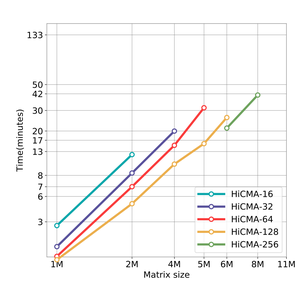
Time to solution.



Energy study.



Synthetic, λ=100, Shaheen-2.



Statistics, Cray-SKL.

Future Research Directions

- LU Factorization/Solve
- Schur Complements
- Preconditioners
- Hardware Accelerators [6]
- Support for Multiple Precisions
- Autotuning: Tile Size, Fixed Accuracy and Fixed Ranks
- Support for OpenMP, PaRSEC and Kokkos
- Support for HODLR, H, HSS and H^2

References

[1] P. Amestoy, C. Ashcraft, O. Boiteau, A. Buttari, J.-Y. L'Excellent, and C. Weisbecker. Improving Multifrontal Methods by Means of Block Low-Rank Representations. *SIAM SISC*, 37(3):A1451–A1474, 2015.

[2] K. Akbudak, H. Ltaief, A. Mikhalev, and D. Keyes. Tile low rank cholesky factorization for climate/weather modeling applications on manycore architectures. In *International Supercomputing Conference*, pages 22–40. Springer, 2017.

[3] E. Agullo, J. Demmel, J. Dongarra, B. Hadri, J. Kurzak, J. Langou, H. Ltaief, P. Luszczek, and S. Tomov. Numerical Linear Algebra on Emerging Architectures: The PLASMA and MAGMA projects. *Journal of Physics: Conference Series*, 180(1):012037, 2009.

[4] K. Akbudak, H. Ltaief, A. Mikhalev, A. Charara, and D. Keyes. Exploiting Data Sparsity for Large-Scale Matrix Computations. In *EuroPar Conference (Submitted), available at http://repository.kaust.edu.sa/kaust/handle/10754/627403*, 2018.

[5] S. Abdulah, H. Ltaief, Y. Sun, M. Genton, and D. Keyes. Exploiting Data Sparsity for Large-Scale Matrix Computations. In *ICPP Conference (Submitted), available at https://arxiv.org/abs/1804.09137*, 2018.

[6] A. Charara, D. Keyes, and H. Ltaief. Exploiting Data Sparsity for Large-Scale Matrix Computations. In *EuroPar Conference (Submitted), available at http://repository.kaust.edu.sa/kaust/handle/10754/627402*, 2018.