SparseKaffe: High-performance, auto-tuned, energy-aware algorithms for sparse direct methods on modern heterogeneous architectures

Meng Tang, Mohammed Aref Gadou, Tania Banerjee and Sanjay Ranka
Department of Computer and Information Science and Engineering, University of Florida

UF UNIVERSITY of FLORIDA

IN COLLABORATION WITH UNIVERSITY OF TENNESSEE AND TEXAS A&M

BACKGROUND

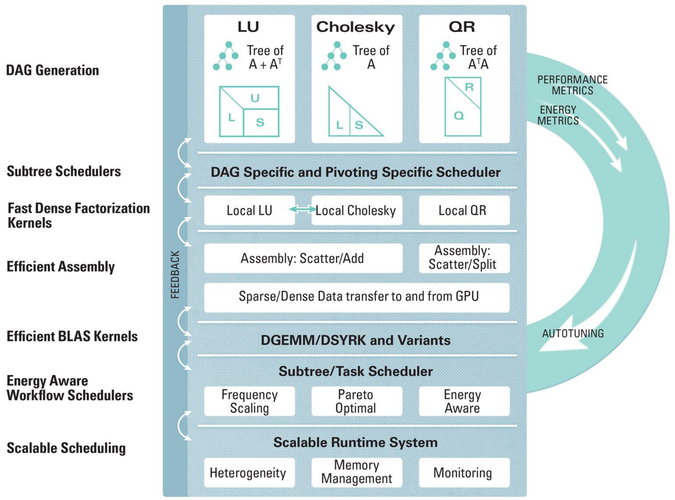
- **Direct methods (QR, Cholesky, and LU factorization)** can be used to find solutions in many numerical algebra applications, including sparse linear systems, sparse linear least squares, and eigenvalue problems; consequently they form the backbone of a broad spectrum of large scale applications.
- **Hybrid Multicore Processors (HMPs) -** multicore processors with one or more coprocessors, such as GPUs or Xeon Phi cores—are poised to dominate the landscape of the next generation of computing, from desktops to extreme scale systems.

OBJECTIVES

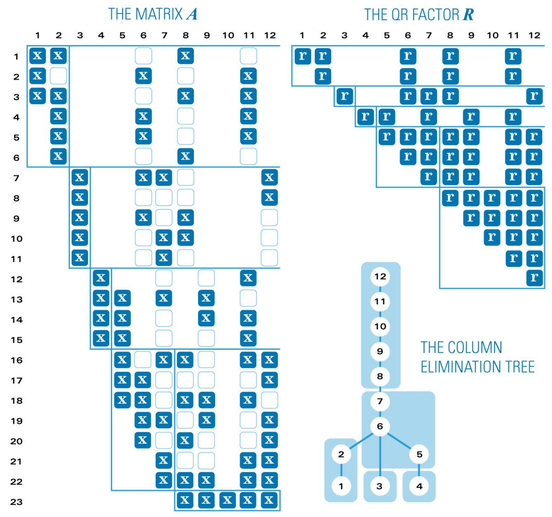
- **Build a common and auto-tunable framework for parallel sparse multifrontal algorithms:** In each of these methods, the computation has a nonuniform and hierarchical workflow, where each node of the workflow is a factorization of a dense submatrix. The edges represent an irregular data movement, where the results from a child node are assembled into the frontal matrix of the parent.
- **Develop novel scheduling algorithms for nonuniform hierarchical workflow**: At or near the leaves of the tree, a small number of threads in a single thread-block on a GPU, or a single core of a Xeon Phi coprocessor, will work on their own frontal matrices. Further up the tree, multiple thread-blocks, GPUs, Xeon Phi's, and/or CPUs will collaborate to factorize a frontal matrix. Still higher up, multiple computational units will cooperate to factor the largest frontal matrices in a distributed-memory fashion.
- **Develop novel algorithms for optimizing the energy requirements**: We will develop scheduling algorithms that allow for exploitation of features such as Dynamic Voltage and Frequency Scheduling (DVFS), core disabling, clock gating, and reconfiguring caches to reduce energy requirements.

INTELLECTUAL MERIT

- **Multiobjective Optimization:** Requires the development of novel and innovative algorithms for scheduling, energy minimization, and memory management.
- **Autotuning for Performance and Energy:** This requires the development of building blocks to support different types of cores, each calling for novel user-guided autotuning algorithms that exploit different hardware characteristics in terms of energy, performance, and parallelism.
- **Building a Common Software Framework**: This requires a careful consideration of the common and different characteristics of these methods.

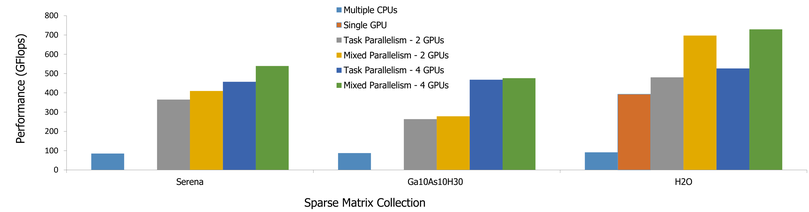


Overview of SparseKaffe Project

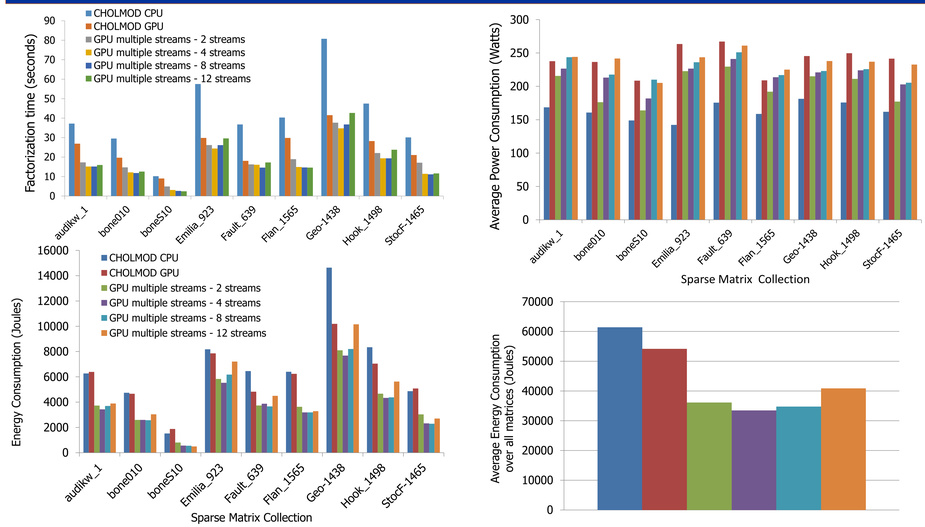


A sparse matrix and its column elimination tree. Each x is a nonzero in A, each dot is an entry that will become nonzero as the matrix is factorized, and each r is a nonzero in R. Each node of the tree is a column of A or row of R, grouped together when adjacent rows of R have the same nonzero pattern.

QR FACTORIZATION



CHOLESKY FACTORIZATION



CONCLUSIONS

- Using our multi-GPU version of QR factorization, we obtained up to 3 times improvement in performance using four GPUs compared to a single GPU.
- Our multi-stream enhancement to CHOLMOD version 4.5.4, improved both performance and energy consumption. In particular, performance improvement was up to 95% using 4 parallel streams.
- We are working on enhancements to CHOLMOD version 4.6.0. Early results show improvement in performance when two GPU streams are used compared to the original version.

PUBLICATIONS

- Meng Tang, Gadou and Sanjay Ranka, "A Multithreaded Algorithm for Sparse Cholesky Factorization", ICCS 2017 and Elsevier Procedia Computer Science Journal.
- Mohamed Gadou, Timothy A. Davis and Sanjay Ranka, "Sparse QR Factorization on heterogeneous systems with multiple GPUs". submitted to JPDC.