Communication Optimal Algorithms for Linear Algebra

Talk by J. Demmel and L. Grigori

In this talk we present COALA, an INRIA associated team that focuses on the design and implementation of numerical algorithms for today's large supercomputers formed by thousands of multicore processors, possibly with accelerators. We focus on operations that are at the heart of many scientific applications as solving linear systems of equations or least squares problems. The algorithms belong to a new class referred to as communication avoiding that provably minimize communication, where communication means the data transferred between levels of memory hierarchy or between processors in a parallel computer. This research is motivated by studies showing that communication costs can already exceed arithmetic costs by orders of magnitude, and the gap is growing exponentially over time. This represents one of main challenges today in high performance computing.

We first present results showing that lower bounds on communication for operations in linear algebra can be determined by extending known communication lower bounds for dense matrix multiplication. We then show that most of the existant algorithms, as for example implemented in LAPACK and ScaLAPACK libraries do asymptotically more communication than the lower bounds require.

Second, we discuss new direct factorization algorithms for dense and sparse matrices that provably minimize communication. In the dense case we will discuss LU, QR and rank-revealing QR (RRQR) factorizations. Both LU and RRQR require new, numerically stable pivoting schemes. We show large speedups on multicore and clusters of multicore machines compared to conventional algorithms, in the LAPACK, ScaLAPACK, MKL, and ESSL libraries. In the sparse case, for matrices arising from discretizations on 2D and 3D regular grids, we present communication-optimal sequential and parallel sparse Cholesky algorithms.