Matrix Computations & Scientific Computing Seminar

Organizer: James Demmel & Ming Gu

Wednesday, 11:00AM–12:00Noon, 380 Soda

Oct. 2 **James Demmel**, UC Berkeley

Communication Lower Bounds and Optimal Algorithms for Programs that Reference Arrays

Our goal is to minimize communication, i.e. moving data, since it increasingly dominates the cost of arithmetic in algorithms. Motivated by this, attainable communication lower bounds have been established by many authors for a variety of algorithms including matrix computations.

The lower bound approach used initially by Irony, Tiskin and Toledo for $O(n^3)$ matrix multiplication, and later by Ballard *et al* for many other linear algebra algorithms, depends on a geometric result by Loomis and Whitney: this result bounds the volume of a 3D set (representing multiply-adds done in the inner loop of the algorithm) using the product of the areas of certain 2D projections of this set (representing the matrix entries available locally, i.e., without communication).

Using a recent generalization of Loomis' and Whitney's result, we generalize this lower bound approach to a much larger class of algorithms, that may have arbitrary numbers of loops and arrays with arbitrary dimensions, as long as the index expressions are affine combinations of loop variables. In other words, the algorithm can do arbitrary operations on any number of variables like $A(i_1, i_2, i_2 - 2 \times i_1, 3 - 4 \times i_3 + 7 \times i_4, \cdots)$. Moreover, the result applies to recursive programs, irregular iteration spaces, sparse matrices, and other data structures as long as the computation can be logically mapped to loops and indexed data structure accesses.

We also discuss when optimal algorithms exist that attain the lower bounds; this leads to new asymptotically faster algorithms for several problems.