Communication Lower Bounds for Programs that Reference Arrays

James Demmel, UC Berkeley

October 24, 2012

Communication, i.e. moving data, whether it is between levels of a memory hierarchy or between parallel processors on a network, can greatly dominate the cost of arithmetic, so algorithms that minimize communication can run much faster (and use less energy) than algorithms that do not. Motivated by this, attainable communication lower bounds have been established for a variety of algorithms including linear algebra. The lower bound approach used initially by Irony/Tiskin/Toledo for $O(n^3)$ matrix multiplication, and later by Ballard/Demmel/Holtz/Schwartz for many other linear algebra algorithms, depended on a geometric result by Loomis and Whitney: This result bounded the volume of a 3D set (representing multiplications 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 the Loomis/Whitney result, we generalize this lower bound approach to a much larger class of algorithms, essentially including any algorithm that accesses arrays indexed by arbitrary linear combinations of the loop variables. In other words, the algorithm can do arbitrary operations on any number of variables like $A(i,j-2 \cdot i, 3 \cdot i-4 \cdot k+7 \cdot l,...)$. We also discuss when optimal algorithms exist that attain the lower bounds; this has already led to new asymptotically faster algorithms for several problems including linear algebra, n-body, and tensor contractions.

This is joint work with Michael Christ, Nicholas Knight, Thomas Scanlon and Katherine Yelick.