Some Notes about BLAS

(slides are obtained from: https://sites.google.com/lbl.gov/cs267-spr2018/)
Basic Linear Algebra Subroutines (BLAS)

• Industry standard interface (evolving)
  • www.netlib.org/blas, www.netlib.org/blas/blast-forum

• Vendors, others supply optimized implementations

• History
  • BLAS1 (1970s): 15 different operations
    • vector operations: dot product, saxpy (y=α*x+y), root-sum-squared, etc
    • m=2*n, f=2*n, q = f/m = computational intensity ~1 or less
  • BLAS2 (mid 1980s): 25 different operations
    • matrix-vector operations: matrix vector multiply, etc
    • m=n^2, f=2*n^2, q~2, less overhead
    • somewhat faster than BLAS1
  • BLAS3 (late 1980s): 9 different operations
    • matrix-matrix operations: matrix matrix multiply, etc
    • m <= 3n^2, f=O(n^3), so q=f/m can possibly be as large as n, so BLAS3 is potentially much faster than BLAS2

• Good algorithms use BLAS3 when possible (LAPACK & ScaLAPACK)
  • See www.netlib.org/{lapack,scalapack}
  • More later in the course
BLAS speeds on an IBM RS6000/590

Peak speed = 266 Mflops

Matrices start in DRAM Memory

BLAS 3 (n-by-n matrix matrix multiply) vs BLAS 2 (n-by-n matrix vector multiply) vs BLAS 1 (saxpy of n vectors)
Dense Linear Algebra: BLAS2 vs. BLAS3

- BLAS2 and BLAS3 have very different computational intensity, and therefore different performance

Data source: Jack Dongarra

CS267 - Lecture 3
Measuring Performance — (Run: me)
Measuring Performance — (Flops/Cycle)

Image and paper by G. Ofenbeck, R. Steinman, V. Caparrós Cabezas, D. Spampinato, M. Püschel