Minimizing Communication in Linear Algebra

Grey Ballard, Mark Hoemmen

{ballard,mhoemmen}@cs.berkeley.edu

Summary

- New communication lower bounds for (nearly) all dense or sparse, sequential or parallel, direct linear algebra problems
- New algorithms that attain lower bounds (sequential and parallel)
- Measured and modeled speedups, not just asymptotics
- Open problems in dense and sparse linear algebra

Motivation

"Communication" means

- Parallel: Data movement between processors
- Sequential: Data movement between levels of memory hierarchy
- # words (inverse bandwidth) and # messages (latency)

Communication matters because:

- Much slower than flops, and getting *exponentially slower* over time
- Moving data much more *energy-intensive* than computing on it

Lower Bounds

Dense Matrix Multiplication

Lower bound on:	Lower bound
# words	$\Omega\left(\text{\# flops} / (\text{local/fast memory size})^{1/2}\right)$
# messages	$\Omega\left(\text{\# flops} / (\text{local/fast memory size})^{3/2}\right)$

- Results due to Hong-Kung [HK81], Irony/Tishkin/Toledo [ITT04]
- Attained by block algorithm (sequential) and Cannon's algorithm (parallel)

Extensions to (nearly) all direct problems

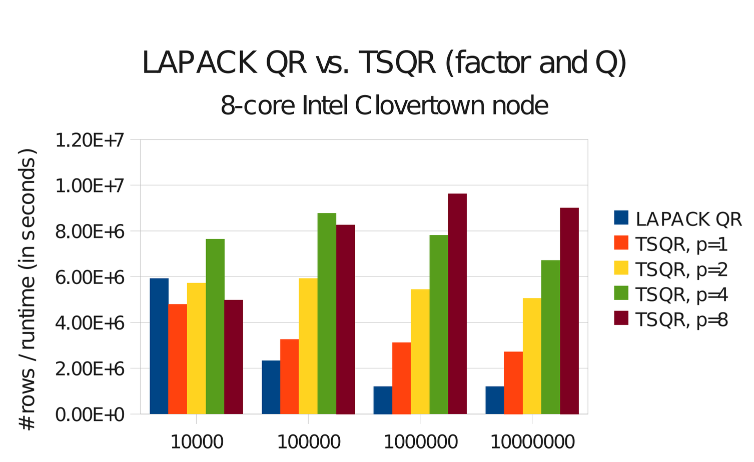
- Theorem: same lower bounds hold for LU, Cholesky, QR, EV/SVD problems
  - Sequential or parallel, dense or sparse
  - See [BDHS09] for details and proof
- Existing library routines not both bandwidth and latency optimal
  - ScaLAPACK: only Cholesky is optimal; LAPACK: Cholesky bandwidth only

New Algorithms

Communication-Avoiding QR (CAQR)

- Factor panel with "Tall Skinny QR" (TSQR): block reduction with QR as operator
- Measured speedup of parallel TSQR: up to $6.7\times$ on 16 processors of a Pentium III cluster
- Modeled speedup of parallel CAQR: up to $9.7\times$ on an IBM Power5 system
- See [DGHL08] for details, models, and more performance results
- Standalone TSQR useful for iterative methods (orthogonalize basis vectors)

TSQR performance results



- Single node of 8-core Intel Clovertown (we have cluster and out-of-core versions too)
- Includes factorization and assembling explicit Q factor
- Best number of threads for LAPACK QR (MKL and stock LAPACK): 1
- Even better measured and modeled speedups on clusters

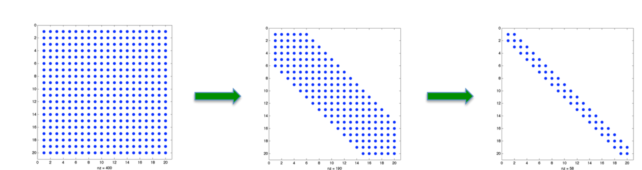
Communication-Avoiding LU (CALU)

- Factor panel once with "Tall Skinny LU" (like a block reduction) to choose pivots
- Swap pivot rows to top and factor *again* without pivoting – $O(n^2)$ extra flops
- Measured speedup of parallel TSLU: up to $5.58\times$ on Cray XT4
- Measured speedup of parallel CALU (size $10^4 \times 10^4$): $1.31\times$ on Cray XT4
- See [DGX08] for details, models, and more performance results

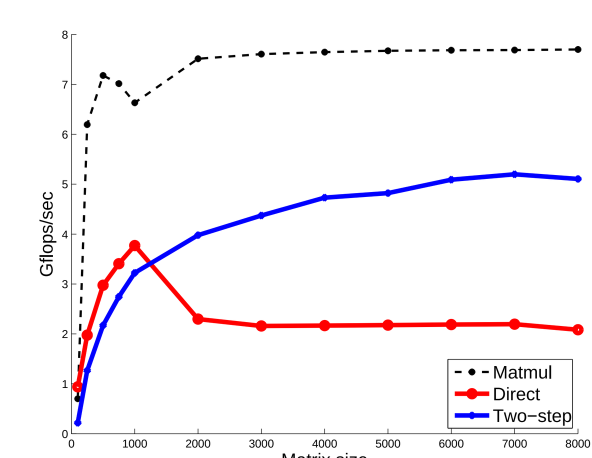
Current Work

Eigenvalue/SVD Problems

- Successive Band Reduction (SBR)
  - uses two-step reduction to tridiagonal rather than one-step



  - pays off when only eigen/singular values are required
  - costs constant factor more flops when vectors are required



- First step (full to banded) can be done in optimal way
  - ∗ using optimal QR factorization and BLAS 3 kernels
- Second step (banded to tridiagonal) is lower order term
- MKL driver routines do not yet take advantage of two-step approach

- Randomized divide-and-conquer approach
  - no reductions, uses randomized rank-revealing QR factorization
  - communication-optimal in asymptotic sense
  - costs (larger) constant factor more flops
- More flops → pay-off in future

Sparse Cholesky on 5-pt Stencil Matrix

- Gilbert ('73) proved lower bounds for sparse Cholesky factorization
  - $\Omega(n^{3/2})$ flops, $\Omega(n \log n)$ fill-in
- Communication lower bound is then $\Omega\left(\max\left\{\frac{n^{3/2}}{\sqrt{M}}, n \log n\right\}\right)$
- Nested dissection attains computation/fill-in lower bound
  - New variant attains communication lower bound (we think)

Credits

- Research supported by Microsoft (Award #024263) and Intel (Award #024894) funding and by matching funding by U.C. Discovery (Award #DIG07-10227).
- Joint work with J. Demmel, L. Grigori, O. Holtz, J. Langou, M. Mohiyuddin, O. Schwartz, H. Xiang

References

[BDHS09] G. Ballard, J. Demmel, O. Holtz, and O. Schwartz. A general communication lower bound for linear algebra, 2009. Submitted to FOCS.

[DGHL08] J. Demmel, L. Grigori, M. Hoemmen, and J. Langou. Communication-optimal parallel and sequential QR and LU factorizations. Technical Report EECS-2008-89, University of California Berkeley EECS, August 2008. Submitted to SIAM J. Sci. Comp.

[DGX08] J. Demmel, L. Grigori, and H. Xiang. Communication-avoiding Gaussian elimination. In *Supercomputing 2008*, 2008.

[HK81] Jia-Wei Hong and H. T. Kung. I/O complexity: The red-blue pebble game. In *STOC '81: Proceedings of the thirteenth annual ACM symposium on theory of computing*, pages 326–333, New York, NY, USA, 1981. ACM.

[ITT04] D. Irony, S. Toledo, and A. Tiskin. Communication lower bounds for distributed-memory matrix multiplication. *J. Parallel Distrib. Comput.*, 64(9):1017–1026, 2004.