Communication avoiding algorithms for LU and QR factorizations

Laura Grigori

Alpines

INRIA Paris - LJLL, Sorbonne Université

October 2019
Plan

• Motivation
• Communication complexity of linear algebra operations
• Communication avoiding for dense linear algebra
  • LU, QR, Rank Revealing QR factorizations
  • Progressively implemented in ScaLAPACK, LAPACK
  • Algorithms for multicore processors
• Conclusions
Motivation - the communication wall

• Runtime of an algorithm is the sum of:
  • #flops x time_per_flop
  • #wordsMoved / bandwidth
  • #messages x latency

• Time to move data >> time per flop
  • Gap steadily and exponentially growing over time
Motivation - the communication wall

- Runtime of an algorithm is the sum of:
  - \( \# \text{flops} \times \text{time \_per \_flop} \)
  - \( \# \text{words \_moved} / \text{bandwidth} \)
  - \( \# \text{messages} \times \text{latency} \)

- Time to move data >> time per flop
  - Gap steadily and exponentially growing over time

<table>
<thead>
<tr>
<th>Annual improvements</th>
<th>Time/flop</th>
<th>Bandwidth</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Network</td>
<td>26%</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>DRAM</td>
<td>23%</td>
<td>5%</td>
</tr>
</tbody>
</table>

- Performance of an application is less than 10% of the peak performance

“We are going to hit the memory wall, unless something basic changes”
[W. Wulf, S. McKee, 95]

Adapted from J. Demmel
Compelling numbers

DRAM latency:
• DDR2 (2007) ~ 120 ns 1x
• DDR4 (2014) ~ 45 ns 2.6x in 7 yrs
• Stacked memory ~ similar to DDR4

Time/flop
• 2008 Intel Nehalem 3.2GHz×4 cores (51.2 GFlops/socket) 1x
• 2017 Intel Skylake XP 2.1GHz×28 cores (1.8 TFlops/socket) 35x in 9 yrs

Network latency
• Interconnect (example one machine today): 0.25µs to 3.7µs MPI latency

Source: G. Bosilca (UTK), S. Knepper (Intel), J. Shalf (LBL)
Selected past work on reducing communication

• Only few examples shown, many references available

A. Tuning
  • Overlap communication and computation, at most a factor of 2 speedup

B. Ghosting
  • Standard approach in *explicit methods*
  • Store redundantly data from neighboring processors for future computations

Example of a parabolic PDE

\[ u_t = \alpha \Delta u \]

with a finite difference,
the solution at a grid point is:

\[ u_{i,j+1} = u(x_i, t_{j+1}) \]
\[ = f(u_{i-1,j}, u_{ij}, u_{i+1,j}) \]
Selected past work on reducing communication

C. Same operation, different schedule of the computation

*Block algorithms for dense linear algebra*
- Barron and Swinnerton-Dyer, 1960
  - LU factorization used to solve a system with 31 equations - first subroutine written for EDSAC 2
  - Block LU factorization used to solve a system with 100 equations using an auxiliary magnetic-tape
  - The basis of the algorithm used in LAPACK

*Cache oblivious algorithms*
- recursive Cholesky, LU, QR
  (Gustavson ‘97, Toledo ‘97, Elmroth and Gustavson ‘98, Frens and Wise ’03, Ahmed and Pingali ‘00)
Selected past work on reducing communication

D. Same algebraic framework, different numerical algorithm

More opportunities for reducing communication, may affect stability

*Dense LU-like factorization* (Barron and Swinnerton-Dyer, 60)

- LU-like factorization based on pairwise pivoting and its block version
  \[ PA = L_1 L_2 \ldots L_n U \]
- With small modifications, minimizes communication between two levels of fast-slow memory
- Stable for small matrices, unstable for nowadays matrices
Communication in CMB data analysis

- **Map-making problem**
  - Find the best map $x$ from observations $d$, scanning strategy $A$, and noise $N^{-1}$
  - Solve generalized least squares problem involving sparse matrices of size $10^{12}$-by-$10^7$

- **Spherical harmonic transform (SHT)**
  - Synthesize a sky image from its harmonic representation
    - Computation over rows of a 2D object (summation of spherical harmonics)
    - Communication to transpose the 2D object
    - Computation over columns of the 2D object (FFTs)

---

Map making, with R. Stompor, M. Szydlarski.
Results obtained on Hopper, Cray XE6, NERSC.

SHT, with R. Stompor, M. Szydlarski.
Simulation on a petascale computer.
Motivation

• The communication problem needs to be taken into account higher in the computing stack

• A paradigm shift in the way the numerical algorithms are devised is required

• Communication avoiding algorithms - a novel perspective for numerical linear algebra
  • Minimize volume of communication
  • Minimize number of messages
  • Minimize over multiple levels of memory/parallelism
  • Allow redundant computations (preferably as a low order term)
Evolution of numerical libraries

**LINPACK (70’s)**
- vector operations, use BLAS1
- HPL benchmark based on Linpack LU factorization

**LAPACK (80’s)**
- Block versions of the algorithms used in LINPACK
- Uses BLAS3

**ScaLAPACK (90’s)**
- Targets distributed memories
- 2D block cyclic distribution of data
- PBLAS based on message passing

**PLASMA (2008): new algorithms**
- Targets many-core
- Block data layout
- Low granularity, high asynchronicity

Project developed by U Tennessee Knoxville, UC Berkeley, other collaborators.
Source: inspired from J. Dongarra, UTK, J. Langou, CU Denver
Evolution of numerical libraries

• Did we need new algorithms?
  • Results on two-socket, quad-core Intel Xeon EMT64 machine, 2.5 GHz per core
  • LU factorization of an m-by-n matrix, m=10^5 and n varies from 10 to 1000

LU using communication avoiding approach
LU from PLASMA using operations on tiles
LU from Intel MKL using lapack block operations
Lapack will not get even 1/2 of this performance
LU using vector operations
Communication Complexity of Dense Linear Algebra

- Matrix multiply, using $2n^3$ flops (sequential or parallel)
  - Lower bound on Bandwidth = $\Omega \left( \frac{\text{#flops}}{M^{1/2}} \right)$
  - Lower bound on Latency = $\Omega \left( \frac{\text{#flops}}{M^{3/2}} \right)$

- Same lower bounds apply to LU using reduction
  - Demmel, LG, Hoemmen, Langou 2008

$$\begin{pmatrix} I & -B \\ A & I \\ I & I \end{pmatrix} = \begin{pmatrix} I & I \\ A & I \\ I & AB \end{pmatrix} \begin{pmatrix} I & -B \\ I & I \end{pmatrix}$$

- And to almost all direct linear algebra [Ballard, Demmel, Holtz, Schwartz, 09]
Lower bounds for linear algebra

- Computation modelled as an n-by-n-by-n set of lattice points
  \((i,j,k)\) represents the operation \(c(i,j) += f_{ij} \cdot (g_{ijk} \cdot (a(i,k) \cdot b(k,j)))\)
- The computation is divided in \(S\) phases
- Each phase contains exactly \(M\) (the fast memory size) load and store instructions
- Determine how many flops the algorithm can compute in each phase, by applying discrete Loomis-Whitney inequality:

\[
W^2 \leq N_A N_B N_C
\]

Algorithms in direct linear algebra:

\[
\begin{align*}
for & \ i,j,k = 1:n \\
c(i,j) &= f_{ij}(g_{ijk}(a(i,k),b(k,j))) \\
endfor
\end{align*}
\]

- set of points in \(R^3\), represent \(w\) arithmetics
- orthogonal projections of the points onto coordinate planes \(N_A, N_B, N_C\) represent values of \(A, B, C\)
Lower bounds for matrix multiplication (contd)

- Discrete Loomis-Whitney inequality:

\[ w^2 \leq N_A N_B N_C \]

- Since there are at most \( 2M \) elements of A, B, C in a phase, the bound is:

\[ w \leq 2\sqrt{2} M^{3/2} \]

- The number of phases \( S \) is \#flops/w, and hence the lower bound on communication is:

\[
\# \text{messages}(S) \geq \frac{\# \text{flops}}{w} = \Omega\left(\frac{\# \text{flops}}{M^{3/2}}\right)
\]

\[
\# \text{loads/stores} \geq \Omega\left(\frac{\# \text{flops}}{M^{1/2}}\right)
\]
Matrix distributions

1) 1D Column Blocked Layout
2) 1D Column Cyclic Layout

3) 1D Column Block Cyclic Layout

4) Row versions of the previous layouts

5) 2D Row and Column Blocked Layout
6) 2D Row and Column Block Cyclic Layout

Generalizes others

Source slide: J. Demmel
MatMul with 2D Layout

- Consider processors in 2D grid (physical or logical)
- Processors can communicate with 4 nearest neighbors
  - Broadcast along rows and columns

\[
\begin{array}{ccc}
  p(0,0) & p(0,1) & p(0,2) \\
  p(1,0) & p(1,1) & p(1,2) \\
  p(2,0) & p(2,1) & p(2,2) \\
\end{array}
\]

\[
\begin{array}{ccc}
  p(0,0) & p(0,1) & p(0,2) \\
  p(1,0) & p(1,1) & p(1,2) \\
  p(2,0) & p(2,1) & p(2,2) \\
\end{array}
\]

\[
= \begin{array}{ccc}
  p(0,0) & p(0,1) & p(0,2) \\
  p(1,0) & p(1,1) & p(1,2) \\
  p(2,0) & p(2,1) & p(2,2) \\
\end{array}
\]

\[
* \begin{array}{ccc}
  p(0,0) & p(0,1) & p(0,2) \\
  p(1,0) & p(1,1) & p(1,2) \\
  p(2,0) & p(2,1) & p(2,2) \\
\end{array}
\]

- Assume \( p \) processors form square \( s \times s \) grid, \( s = p^{1/2} \)

Source slide: J. Demmel
Cannon’s Algorithm

... \( C(i,j) = C(i,j) + \sum_k A(i,k)B(k,j) \)
... assume \( s = \sqrt{p} \) is an integer

forall \( i = 0 \) to \( s-1 \) … “skew” \( A \)
left-circular-shift row \( i \) of \( A \) by \( i \)
... so that \( A(i,j) \) overwritten by \( A(i,(j+i)\mod s) \)

forall \( i = 0 \) to \( s-1 \) … “skew” \( B \)
up-circular-shift column \( i \) of \( B \) by \( i \)
... so that \( B(i,j) \) overwritten by \( B((i+j)\mod s), j) \)

for \( k = 0 \) to \( s-1 \) … sequential

forall \( i = 0 \) to \( s-1 \) and \( j = 0 \) to \( s-1 \) … all processors in parallel
\( C(i,j) = C(i,j) + A(i,j)B(i,j) \)
left-circular-shift each row of \( A \) by 1
up-circular-shift each column of \( B \) by 1

Source slide: J. Demmel
Cannon’s Matrix Multiplication

Cannon’s Matrix Multiplication Algorithm

\[
C(1,2) = A(1,0) \times B(0,2) + A(1,1) \times B(1,2) + A(1,2) \times B(2,2)
\]

Source slide: J. Demmel
Cost of Cannon’s Algorithm

forall  i=0 to s-1  ... recall s = sqrt(p)
left-circular-shift row i of A by i  ... cost ≤ s*(α + β*n^2/p)
forall  i=0 to s-1
up-circular-shift column i of B by i  ... cost ≤ s*(α + β*n^2/p)
for k=0 to s-1
forall  i=0 to s-1 and j=0 to s-1
C(i,j) = C(i,j) + A(i,j)*B(i,j)  ... cost = 2*(n/s)^3 = 2*n^3/p^{3/2}
left-circular-shift each row of A by 1  ... cost = α + β*n^2/p
up-circular-shift each column of B by 1  ... cost = α + β*n^2/p

° Total Time = 2*n^3/p + 4*s*α + 4*β*n^2/s  - Optimal!
° Parallel Efficiency = 2*n^3 / (p * Total Time)
  = 1/( 1 + α * 2*(s/n)^3 + β * 2*(s/n) )
  = 1/(1 + O(sqrt(p)/n))
° Grows to 1 as n/s = n/sqrt(p) = sqrt(data per processor) grows

Source slide: J. Demmel
## Sequential algorithms and communication bounds

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Minimizing #words (not #messages)</th>
<th>Minimizing #words and #messages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cholesky</strong></td>
<td>LAPACK</td>
<td>[Gustavson, 97] [Ahmed, Pingali, 00]</td>
</tr>
<tr>
<td><strong>LU</strong></td>
<td>LAPACK (few cases) [Toledo,97], [Gustavson, 97] both use partial pivoting</td>
<td>[LG, Demmel, Xiang, 08] [Khabou, Demmel, LG, Gu, 12] uses tournament pivoting</td>
</tr>
<tr>
<td><strong>QR</strong></td>
<td>LAPACK (few cases) [Elmroth,Gustavson,98]</td>
<td>[Frens, Wise, 03], 3x flops [Demmel, LG, Hoemmen, Langou, 08] [Ballard et al, 14]</td>
</tr>
<tr>
<td><strong>RRQR</strong></td>
<td></td>
<td>[Demmel, LG, Gu, Xiang 11] uses tournament pivoting, 3x flops</td>
</tr>
</tbody>
</table>

- Only several references shown for block algorithms (LAPACK), cache-oblivious algorithms and communication avoiding algorithms
- CA algorithms exist also for SVD and eigenvalue computation
2D Parallel algorithms and communication bounds

- If memory per processor = $n^2 / P$, the lower bounds become
  \[ \#\text{words\_moved} \geq \Omega \left( \frac{n^2}{P^{1/2}} \right), \quad \#\text{messages} \geq \Omega \left( P^{1/2} \right) \]

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Minimizing #words (not #messages)</th>
<th>Minimizing #words and #messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesky</td>
<td>ScaLAPACK</td>
<td>ScaLAPACK</td>
</tr>
<tr>
<td>LU</td>
<td>ScaLAPACK</td>
<td>[LG, Demmel, Xiang, 08]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Khabou, Demmel, LG, Gu, 12]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>uses tournament pivoting</td>
</tr>
<tr>
<td>QR</td>
<td>ScaLAPACK</td>
<td>[Demmel, LG, Hoemmen, Langou, 08]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Ballard et al, 14]</td>
</tr>
<tr>
<td>RRQR</td>
<td>ScaLAPACK</td>
<td>[Demmel, LG, Gu, Xiang 13]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>uses tournament pivoting, 3x flops</td>
</tr>
</tbody>
</table>

- Only several references shown, block algorithms (ScaLAPACK) and communication avoiding algorithms
- CA algorithms exist also for SVD and eigenvalue computation
LU factorization (as in ScaLAPACK pdgetrf)

LU factorization on a $P = P_r \times P_c$ grid of processors

For $ib = 1$ to $n-1$ step $b$

$A^{(ib)} = A(ib:n, ib:n)$

(1) Compute panel factorization
   - find pivot in each column, swap rows

(2) Apply all row permutations
   - broadcast pivot information along the rows
   - swap rows at left and right

(3) Compute block row of U
   - broadcast right diagonal block of L of current panel

(4) Update trailing matrix
   - broadcast right block column of L
   - broadcast down block row of U

$O(n \log_2 P_r)$

$O(n / b(\log_2 P_c + \log_2 P_r))$

$O(n / b \log_2 P_c)$

$O(n / b(\log_2 P_c + \log_2 P_r))$
General scheme for QR factorization by Householder transformations

• Apply orthogonal transformations to annihilate subdiagonal entries

\[ H_n H_{n-1} \ldots H_2 H_1 A = R \rightarrow A = \left( H_n H_{n-1} \ldots H_2 H_1 \right)^T R \]

\[ Q = H_1 H_2 \ldots H_n \]

• Householder transformation

\[ H_i = I - \tau_i h_i h_i^T \]

• Orthogonal factor Q can be represented implicitly by \( mxn \) lower trapezoidal matrix \( Y = (h_1 h_2 \ldots h_n) \)

\[ Q = I - YTY^T \]
Blocked QR factorization by Householder transformations

Block Householder QR iterates over block of columns and
1. Computes panel factorization
2. Updates trailing matrix
Block QR factorization

\[ A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = Q_1 \begin{pmatrix} R_{11} & R_{12} \\ \hline & A_{22} \end{pmatrix} \]

Block QR algebra:

1. Compute panel factorization:
   \[
   \begin{pmatrix} A_{11} \\ A_{12} \end{pmatrix} = Q_1 \begin{pmatrix} R_{11} \end{pmatrix}, \quad Q_1 = H_1 H_2 \ldots H_b
   \]

2. Compute the compact representation:
   \[
   Q_1 = I - Y_1 T_1 Y_1^T
   \]

3. Update the trailing matrix:
   \[
   \begin{pmatrix} I - Y_1 T_1^T Y_1^T \end{pmatrix} \begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix} = \begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix} - Y_1 T_1^T Y_1^T \begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix} = \begin{pmatrix} R_{12} \\ A_{22} \end{pmatrix}
   \]

4. The algorithm continues recursively on the trailing matrix.
TSQR: QR factorization of a tall skinny matrix using Householder transformations

- QR decomposition of $m \times b$ matrix $W$, $m \gg b$
  - $P$ processors, block row layout

- Classic Parallel Algorithm
  - Compute Householder vector for each column
  - Number of messages $\propto b \log P$

- Communication Avoiding Algorithm
  - Reduction operation, with QR as operator
  - Number of messages $\propto \log P$

$$ W = \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{bmatrix} \rightarrow \begin{bmatrix} R_{00} \\ R_{10} \\ R_{20} \\ R_{30} \end{bmatrix} \rightarrow R_{01} \rightarrow R_{02} \rightarrow R_{11} \rightarrow R_{12} $$

J. Demmel, LG, M. Hoemmen, J. Langou, 08
Parallel TSQR

References: Golub, Plemmons, Sameh 88, Pothen, Raghavan, 89, Da Cunha, Becker, Patterson, 02
Algebra of TSQR

Parallel:

\[ W = \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{bmatrix} \rightarrow \begin{bmatrix} R_{00} \\ R_{10} \\ R_{20} \\ R_{30} \end{bmatrix} \rightarrow \begin{bmatrix} R_{01} \\ R_{11} \end{bmatrix} \rightarrow R_{02} \]

\[
\begin{pmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{pmatrix} = \begin{pmatrix} Q_{00} R_{00} \\ Q_{10} R_{10} \\ Q_{20} R_{20} \\ Q_{30} R_{30} \end{pmatrix} = \begin{pmatrix} Q_{00} \\ Q_{10} \\ Q_{20} \\ Q_{30} \end{pmatrix} \begin{pmatrix} R_{00} \\ R_{10} \\ R_{20} \\ R_{30} \end{pmatrix}
\]

\[
\begin{pmatrix} R_{00} \\ R_{10} \\ R_{20} \\ R_{30} \end{pmatrix} = \begin{pmatrix} Q_{01} R_{01} \\ Q_{11} R_{11} \end{pmatrix} = \begin{pmatrix} Q_{01} \\ Q_{11} \end{pmatrix} \begin{pmatrix} R_{01} \\ R_{11} \end{pmatrix} = Q_{02} R_{02}
\]

Q is represented implicitly as a product

Output: \{Q_{00}, Q_{10}, Q_{00}, Q_{20}, Q_{30}, Q_{01}, Q_{11}, Q_{02}, R_{02}\}
Flexibility of TSQR and CAQR algorithms

Parallel: \[ w = \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{bmatrix} \rightarrow R_{00} \rightarrow R_{01} \rightarrow R_{02} \]

Sequential: \[ w = \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{bmatrix} \rightarrow R_{00} \rightarrow R_{01} \rightarrow R_{02} \rightarrow R_{03} \]

Dual Core: \[ w = \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{bmatrix} \rightarrow R_{00} \rightarrow R_{01} \rightarrow R_{11} \rightarrow R_{02} \rightarrow R_{03} \]

Reduction tree will depend on the underlying architecture, could be chosen dynamically.
Algebra of TSQR

Parallel: \[ W = \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{bmatrix} \rightarrow \begin{bmatrix} R_{00} \\ R_{10} \\ R_{20} \\ R_{30} \end{bmatrix} \rightarrow \begin{bmatrix} R_{01} \\ R_{11} \end{bmatrix} \rightarrow R_{02} \]
QR for General Matrices

• Cost of CAQR vs ScALAPACK’s PDGEQRF
  • n x n matrix on P^{1/2} x P^{1/2} processor grid, block size b
  • Flops: \((4/3)n^3/P + (3/4)n^2b \log P/P^{1/2}\) vs \((4/3)n^3/P\)
  • Bandwidth: \((3/4)n^2 \log P/P^{1/2}\) vs same
  • Latency: \(2.5 n \log P / b\) vs \(1.5 n \log P\)

• Close to optimal (modulo log P factors)
  • Assume: O(n^2/P) memory/processor, O(n^3) algorithm,
  • Choose b near \(n / P^{1/2}\) (its upper bound)
  • Bandwidth lower bound:
    \(\Omega(n^2 /P^{1/2})\) – just log(P) smaller
  • Latency lower bound:
    \(\Omega(P^{1/2})\) – just polylog(P) smaller
Performance of TSQR vs Sca/LAPACK

- Parallel
  - Intel Xeon (two socket, quad core machine), 2010
    - Up to 5.3x speedup (8 cores, $10^5 \times 200$)
  - Pentium III cluster, Dolphin Interconnect, MPICH, 2008
    - Up to 6.7x speedup (16 procs, 100K x 200)
  - BlueGene/L, 2008
    - Up to 4x speedup (32 procs, 1M x 50)
  - Tesla C 2050 / Fermi (Anderson et al)
    - Up to 13x (110,592 x 100)
  - Grid – 4x on 4 cities vs 1 city (Dongarra, Langou et al)
  - QR computed locally using recursive algorithm (Elmroth-Gustavson) – enabled by TSQR

- Results from many papers, for some see [Demmel, LG, Hoemmen, Langou, SISC 12], [Donfack, LG, IPDPS 10].
Modeled Speedups of CAQR vs ScaLAPACK

Petascale machine with 8192 procs, each at 500 GFlops/s, a bandwidth of 4 GB/s.

\[ \gamma = 2 \cdot 10^{-12} \text{s}, \alpha = 10^{-5} \text{s}, \beta = 2 \cdot 10^{-9} \text{s/word}. \]
Impact

• TSQR/CAQR implemented in
  • Intel Data analytics library
  • GNU Scientific Library
  • ScaLAPACK
  • Spark for data mining

• CALU implemented in
  • Cray’s libsci
  • To be implemented in lapack/scapalack
Algebra of TSQR

Parallel: \[ W = \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{bmatrix} \rightarrow R_{00} \rightarrow R_{01} \rightarrow R_{02} \]

\[ \rightarrow R_{10} \rightarrow R_{11} \rightarrow R_{12} \]

TSQR-HR

CAQR
Reconstruct Householder vectors from TSQR

The QR factorization using Householder vectors

\[ W = QR = (I - YTY_1^T)R \]

can be re-written as an LU factorization

\[ W - R = Y(-TY_1^T)R \]
\[ Q - I = Y(-TY_1^T) \]
Reconstruct Householder vectors TSQR-HR

1. Perform TSQR
2. Form Q explicitly (tall-skinny orthonormal factor)
3. Perform LU decomposition: $Q - I = LU$

4. Set $Y = L$
5. Set $T = -U \ Y_1^{-T}$

$$I - YTYY^T = I - \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}T\begin{bmatrix} Y_1^T \\ Y_2^T \end{bmatrix}$$
Strong scaling

Hopper: Cray XE6 (NERSC) – 2 x 12-core AMD Magny-Cours (2.1 GHz)
Edison: Cray CX30 (NERSC) – 2 x 12-core Intel Ivy Bridge (2.4 GHz)
Effective flop rate, computed by dividing $2mn^2 - 2n^3/3$ by measured runtime
Ballard, Demmel, LG, Jacquelin, Knight, Nguyen, and Solomonik, 2015.
The LU factorization of a tall skinny matrix

First try the obvious generalization of TSQR.

\[
W = \begin{pmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{pmatrix} = \begin{pmatrix} \Pi_{00} & \Pi_{10} & \Pi_{20} & \Pi_{30} \end{pmatrix} \begin{pmatrix} L_{00} \\ L_{10} \\ L_{20} \\ L_{30} \end{pmatrix} \begin{pmatrix} U_{00} \\ U_{10} \\ U_{20} \\ U_{30} \end{pmatrix}
\]

\[
\begin{pmatrix} U_{00} \\ U_{10} \\ U_{20} \\ U_{30} \end{pmatrix} = \begin{pmatrix} \Pi_{01} & \Pi_{11} \end{pmatrix} \begin{pmatrix} L_{01} & L_{11} \end{pmatrix} \begin{pmatrix} U_{01} \\ U_{11} \end{pmatrix} = \prod_{\Pi_2} L_{02} U_{02}
\]
Obvious generalization of TSQR to LU

• Block parallel pivoting:
  • uses a binary tree and is optimal in the parallel case

\[ W = \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{bmatrix} \rightarrow U_{00} \rightarrow U_{01} \rightarrow U_{02} \]

• Block pairwise pivoting:
  • uses a flat tree and is optimal in the sequential case
  • introduced by Barron and Swinnerton-Dyer, 1960: block LU factorization used to solve a system with 100 equations on EDSAC 2 computer using an auxiliary magnetic-tape
  • used in PLASMA for multicore architectures and FLAME for out-of-core algorithms and for multicore architectures

\[ W = \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{bmatrix} \rightarrow U_{00} \rightarrow U_{01} \rightarrow U_{02} \rightarrow U_{03} \]
Stability of the LU factorization

• The backward stability of the LU factorization of a matrix $A$ of size $n$-by-$n$

$$
\left\| \hat{L} \cdot \hat{U} \right\|_\infty \leq (1 + 2(n^2 - n)g_w )\|A\|_\infty
$$

depends on the growth factor

$$
g_w = \frac{\max_{i,j,k} |a_{ij}^k|}{\max_{i,j} |a_{ij}|}
$$

where $a_{ij}^k$ are the values at the $k$-th step.

• $g_w \leq 2^{n-1}$, attained for Wilkinson matrix

but in practice it is on the order of $n^{2/3} - n^{1/2}$

• Two reasons considered to be important for the average case stability [Trefethen and Schreiber, 90] :

- the multipliers in $L$ are small,

- the correction introduced at each elimination step is of rank 1.
Block parallel pivoting

- Unstable for large number of processors \( P \)
- When \( P = \) number rows, it corresponds to parallel pivoting, known to be unstable (Trefethen and Schreiber, 90)
Block pairwise pivoting

- Results shown for random matrices
- Will become unstable for large matrices

\[ W = \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{bmatrix} \rightarrow U_{00} \rightarrow U_{01} \rightarrow U_{02} \rightarrow U_{03} \]
Tournament pivoting - the overall idea

- At each iteration of a block algorithm

\[ A = \begin{pmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ A_{21} & A_{22} \end{pmatrix} \]

\[ b \quad n - b \]

, where \[ W = \begin{pmatrix} A_{11} \\ A_{21} \end{pmatrix} \]

- Preprocess \( W \) to find at low communication cost good pivots for the LU factorization of \( W \), return a permutation matrix \( P \).
- Permute the pivots to top, ie compute \( PA \).
- Compute LU with no pivoting of \( W \), update trailing matrix.

\[ PA = \begin{pmatrix} L_{11} & 0 \\ L_{21} & I_{n-b} \end{pmatrix} \begin{pmatrix} U_{11} & U_{12} \\ 0 & A_{22} - L_{21}U_{12} \end{pmatrix} \]
Tournament pivoting for a tall skinny matrix

1) Compute GEPP factorization of each $W_i$, find permutation $\Pi_0$

$$W = \begin{pmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \end{pmatrix} = \begin{pmatrix} \Pi_{00}L_{00}U_{00} \\ \Pi_{10}L_{10}U_{10} \\ \Pi_{20}L_{20}U_{20} \\ \Pi_{30}L_{30}U_{30} \end{pmatrix}$$

Pick b pivot rows, form $A_{00}$

Same for $A_{10}$

Same for $A_{20}$

Same for $A_{30}$

2) Perform $\log_2(P)$ times GEPP factorizations of $2b$-by-$b$ rows, find permutations $\Pi_1, \Pi_2$

$$\begin{pmatrix} A_{00} \\ A_{10} \\ A_{20} \\ A_{30} \end{pmatrix} = \begin{pmatrix} \Pi_{01}L_{01}U_{01} \\ \Pi_{11}L_{11}U_{11} \end{pmatrix}$$

Pick b pivot rows, form $A_{01}$

Same for $A_{11}$

$$\begin{pmatrix} A_{01} \\ A_{11} \end{pmatrix} = \Pi_2L_{02}U_{02}$$

3) Compute LU factorization with no pivoting of the permuted matrix:

$$\Pi_2^T\Pi_1^T\Pi_0^TW = LU$$
Tournament pivoting

\[
\begin{align*}
P_0 & \quad \begin{pmatrix} W_0 \\ \Pi_0^T W_0 \\ \Pi_0^T W_0 \end{pmatrix} = \Pi_0 L_0 U_0 \\
\begin{pmatrix} 2 & 4 \\ 0 & 1 \\ 2 & 0 \\ 1 & 2 \end{pmatrix} & \quad \begin{pmatrix} 2 & 4 \\ 2 & 0 \\ 4 & 1 \\ 2 & 0 \end{pmatrix} & \quad \begin{pmatrix} 4 & 1 \\ 2 & 4 \\ 4 & 2 \\ 1 & 4 \end{pmatrix} = \Pi_0 L_0 U_0 \\
\end{align*}
\]

Good pivots for factorizing \( W \)

\[
\begin{align*}
P_1 & \quad \begin{pmatrix} W_1 \\ \Pi_1^T W_1 \\ \Pi_1^T W_1 \end{pmatrix} = \Pi_1 L_1 U_1 \\
\begin{pmatrix} 2 & 0 \\ 0 & 0 \\ 4 & 1 \\ 1 & 0 \end{pmatrix} & \quad \begin{pmatrix} 4 & 1 \\ 2 & 0 \end{pmatrix} & \quad \begin{pmatrix} 4 & 1 \\ 2 & 4 \\ 4 & 2 \\ 1 & 4 \end{pmatrix} = \Pi_1 L_1 U_1 \\
\end{align*}
\]

\[
\begin{align*}
P_2 & \quad \begin{pmatrix} W_2 \\ \Pi_2^T W_2 \\ \Pi_2^T W_2 \end{pmatrix} = \Pi_2 L_2 U_2 \\
\begin{pmatrix} 0 & 1 \\ 1 & 4 \\ 0 & 0 \\ 0 & 2 \end{pmatrix} & \quad \begin{pmatrix} 1 & 4 \\ 0 & 2 \\ 4 & 2 \\ 0 & 2 \end{pmatrix} & \quad \begin{pmatrix} 4 & 2 \\ 1 & 4 \end{pmatrix} = \Pi_2 L_2 U_2 \\
\end{align*}
\]

\[
\begin{align*}
P_3 & \quad \begin{pmatrix} W_3 \\ \Pi_3^T W_3 \\ \Pi_3^T W_3 \end{pmatrix} = \Pi_3 L_3 U_3 \\
\begin{pmatrix} 2 & 1 \\ 0 & 2 \\ 1 & 0 \\ 4 & 2 \end{pmatrix} & \quad \begin{pmatrix} 4 & 2 \\ 0 & 2 \end{pmatrix} \quad \text{time} \quad \end{align*}
\]
Growth factor for binary tree based CALU

- Random matrices from a normal distribution
- Same behaviour for all matrices in our test, and $|L| \leq 4.2$
Stability of CALU (experimental results)

- Results show $\|PA-LU\|/\|A\|$, normwise and componentwise backward errors, for random matrices and special ones
  - See [LG, Demmel, Xiang, SIMAX 2011] for details
  - BCALU denotes binary tree based CALU and FCALU denotes flat tree based CALU
Our “proof of stability” for CALU

• CALU as stable as GEPP in following sense:
  In exact arithmetic, CALU process on a matrix $A$ is equivalent to GEPP process on a larger matrix $G$ whose entries are blocks of $A$ and zeros.

• Example of one step of tournament pivoting:

\[
A = \begin{pmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22} \\
A_{31} & A_{32}
\end{pmatrix}
\]

\[
G = \begin{pmatrix}
\overline{A}_{11} & \overline{A}_{12} \\
A_{21} & A_{21} \\
A_{31} & -A_{31} & A_{32}
\end{pmatrix}
\]

\[
\text{tournament pivoting:}
\begin{align*}
A_{11} & \rightarrow A_{11} \\
A_{21} & \rightarrow A_{21}
\end{align*}
\]

• Proof possible by using original rows of $A$ during tournament pivoting (not the computed rows of $U$).
Outline of the proof of stability for CALU

- Consider \( A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{pmatrix} \), and the result of TSLU as \( \begin{pmatrix} A_{11} \\ A_{21} \\ A_{31} \end{pmatrix} \rightarrow A_{11} \rightarrow A_{11} \)

- After the factorization of first panel by CALU, \( A_{s32} \) (the Schur complement of \( A_{32} \)) is not bounded as in GEPP,

\[
\begin{pmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{pmatrix} = \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \\ \bar{A}_{31} & \bar{A}_{32} \end{pmatrix} = \begin{pmatrix} \bar{L}_{11} & \bar{U}_{11} \\ \bar{L}_{21} & \bar{I} \end{pmatrix} \begin{pmatrix} \bar{U}_{11} & \bar{U}_{12} \\ \bar{A}_{22} \\ \bar{A}_{s32} \end{pmatrix}
\]

- but \( A_{s32} \) can be obtained by GEPP on larger matrix \( G \) formed from blocks of \( A \)

\[
G = \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{21} \\ -A_{31} & A_{32} \end{pmatrix} = \begin{pmatrix} \bar{L}_{11} & \bar{L}_{11}^{-1} \bar{U}_{11} \\ \bar{L}_{21} & \bar{L}_{21}^{-1} \bar{U}_{11} \bar{U}_{12} \\ \bar{L}_{31} & \bar{I} \end{pmatrix} \begin{pmatrix} \bar{U}_{11} & \bar{U}_{12} \\ \bar{U}_{21} & -L_{21}^{-1} A_{21} \bar{U}_{11} \bar{U}_{12} \\ \bar{A}_{s32} \end{pmatrix}
\]

- GEPP on \( G \) does not permute and

\[
L_{31} L_{21}^{-1} A_{21} \bar{U}_{11}^{-1} \bar{U}_{12} + A_{s32} = L_{31} U_{21} \bar{U}_{11}^{-1} \bar{U}_{12} + A_{s32} = A_{31} \bar{U}_{11}^{-1} \bar{U}_{12} + A_{s32} = \bar{L}_{31} \bar{U}_{12} + A_{s32} = A_{32}
\]
LU factorization and low rank matrices

• For low rank matrices, the factorization of $A_1$ computed as following might not be stable

  Compute $PA=LU$ by using GEPP
  $L(k+1:end,k) = A(k+1:end,k)/A(k,k)$
  Permute the matrix $A_1=PA$
  Compute $LU$ with no pivoting $A_1=L_1U_1$
  $L(k+1:end,k) = L(k+1:end,k)*(1/A(k,k))$

• Example $A = \text{randn}(6,3)*\text{randn}(3,5)$, $\max(\text{abs}(L)) = 1$, $\max(\text{abs}(L_1)) = 10^{15}$

After 4 steps of factorization of $PA$ we obtain:

$$PA_4 = \begin{pmatrix} 1.0000 \\ 0.1729 & 1.0000 \\ 0.6061 & 0.8608 & 1.0000 \\ 0.5776 & 0.0543 & 0.3264 & 1.0000 \\ 0.4789 & -0.2877 & -0.1545 & 2.3333 \\ -0.3264 & -0.7514 & -0.4597 & 1.7778 & 8.3e-17 \end{pmatrix} \cdot \begin{pmatrix} 4.4766 & 3.0163 & -4.7390 & 4.2180 & -0.8164 \\ -1.5439 & -0.4703 & 2.9267 & 1.0925 \\ 1.6149 & 2.3623 & 0.3167 \\ 9.9e-16 & 1.6e-16 & 1 \end{pmatrix}$$

Schur complement after 4 elimination steps

After 4 steps of factorization of $A_1$ we obtain:

$$A_1^4 = \begin{pmatrix} 1.0000 \\ 0.1729 & 1.0000 \\ 0.6061 & 0.8608 & 1.0000 \\ 0.5776 & 0.0543 & 0.3264 & 1.0000 \\ 0.4789 & -0.2877 & -0.1545 & 2.3333 & 4.9e-32 \\ -0.3264 & -0.7514 & -0.4597 & 1.7778 & -7.4e-17 \end{pmatrix}$$
LU_PRRP: LU with panel rank revealing pivoting

- Pivots are selected by using strong rank revealing QR on each panel
- The factorization after one panel elimination is written as

\[
PA = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} I_b & \cdot \\ A_{21}A_{11}^{-1} & I_{n-b} \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{22} - A_{21}A_{11}^{-1}A_{12} \end{pmatrix}
\]

\(A_{21}A_{11}^{-1}\) is computed through strong rank revealing QR and \(\max(|A_{21}A_{11}^{-1}|)_{ij} \leq f\)

- LU_PRRP and CALU_PRRP stable for pathological cases (Wilkinson matrix) and matrices from two real applications (Volterra integral equation - Foster, a boundary value problem - Wright) on which GEPP fails.

A. Khabou, J. Demmel, LG, M. Gu, 2012
Growth factor in exact arithmetic

- Matrix of size m-by-n, reduction tree of height $H = \log(P)$.
- (CA)LU_PRPP select pivots using strong rank revealing QR (A. Khabou, J. Demmel, LG, M. Gu, SIMAX 2013)
- “In practice” means observed/expected/conjectured values.

<table>
<thead>
<tr>
<th></th>
<th>CALU</th>
<th>GEPP</th>
<th>CALU_PRPP</th>
<th>LU_PRPP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Upper bound</strong></td>
<td>$2^{n(\log(P)+1)-1}$</td>
<td>$2^{n-1}$</td>
<td>$(1+2b)^{(n/b)\log(P)}$</td>
<td>$(1+2b)^{(n/b)}$</td>
</tr>
<tr>
<td><strong>In practice</strong></td>
<td>$n^{2/3} -- n^{1/2}$</td>
<td>$n^{2/3} -- n^{1/2}$</td>
<td>$(n/b)^{2/3} -- (n/b)^{1/2}$</td>
<td>$(n/b)^{2/3} -- (n/b)^{1/2}$</td>
</tr>
</tbody>
</table>

Better bounds

- For a matrix of size $10^7$-by-$10^7$ (using petabytes of memory)
  
  $n^{1/2} = 10^{3.5}$
CALU – a communication avoiding LU factorization

• Consider a 2D grid of $P$ processors $P_r$-by-$P_c$, using a 2D block cyclic layout with square blocks of size $b$.

For $ib = 1$ to $n-1$ step $b$

$A^{(ib)} = A(ib:n, ib:n)$

(1) Find permutation for current panel using TSLU $O(n/b \log_2 P_r)$

(2) Apply all row permutations ($pdla$swp) $O(n/b(\log_2 P_c + \log_2 P_r))$

- broadcast pivot information along the rows of the grid

(3) Compute panel factorization ($dtrsm$)

(4) Compute block row of $U$ ($pdtrsm$)

- broadcast right diagonal part of $L$ of current panel

(5) Update trailing matrix ($pdgemm$)

- broadcast right block column of $L$
  - broadcast down block row of $U$
LU for General Matrices

• Cost of **CALU** vs **ScaLAPACK**’s **PDGETRF**
  • n x n matrix on P^{1/2} x P^{1/2} processor grid, block size b
  • Flops: \((2/3)n^3/P + (3/2)n^2b / P^{1/2}\) vs \((2/3)n^3/P + n^2b/P^{1/2}\)
  • Bandwidth: \(n^2 \log P/P^{1/2}\) vs same
  • Latency: \(3n \log P / b\) vs \(1.5n \log P + 3.5n \log P / b\)

• Close to optimal (modulo log P factors)
  • Assume: \(O(n^2/P)\) memory/processor, \(O(n^3)\) algorithm,
  • Choose b near \(n / P^{1/2}\) (its upper bound)
  • Bandwidth lower bound:
    \(\Omega(n^2 / P^{1/2})\) – just log(P) smaller
  • Latency lower bound:
    \(\Omega(P^{1/2})\) – just polylog(P) smaller
Performance vs ScaLAPACK

- Parallel TSLU (LU on tall-skinny matrix)
  - IBM Power 5
    - Up to $4.37x$ faster (16 procs, 1M x 150)
  - Cray XT4
    - Up to $5.52x$ faster (8 procs, 1M x 150)

- Parallel CALU (LU on general matrices)
  - Intel Xeon (two socket, quad core)
    - Up to $2.3x$ faster (8 cores, $10^6$ x 500)
  - IBM Power 5
    - Up to $2.29x$ faster (64 procs, 1000 x 1000)
  - Cray XT4
    - Up to $1.81x$ faster (64 procs, 1000 x 1000)

- Details in SC08 (LG, Demmel, Xiang), IPDPS’10 (S. Donfack, LG).