Perspectives on Parallel Numerical Linear Algebra (PPNLA)
18 July 2012
School of Mathematics, The University of Manchester

# *Algorithmic and Software Challenges when Moving Towards Exascale (Not Your Father's Math Library)*

Jack Dongarra
University of Tennessee
Oak Ridge National Laboratory
University of Manchester

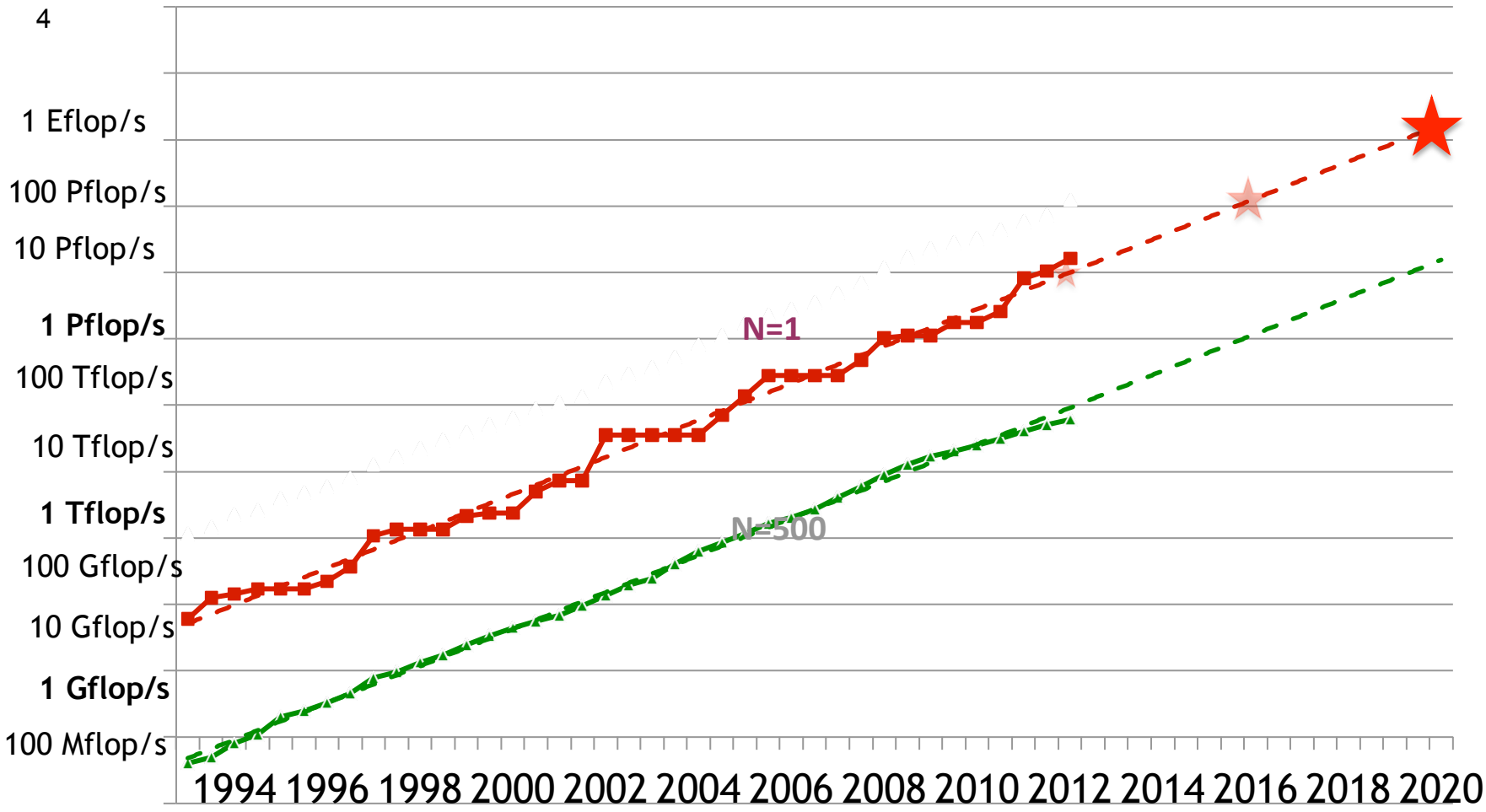
# June 2012: The TOP10

| Rank | Site | Computer | Country | Cores | Rmax [Pflops] | % of Peak | Power [MW] | MFlops /Watt |
|---|---|---|---|---|---|---|---|---|
| 1 | DOE / NNSA L Livermore Nat Lab | Sequoia, BlueGene/Q (16c) + custom | USA | 1,572,864 | 16.3 | 81 | 8.6 | 1895 |
| 2 | RIKEN Advanced Inst for Comp Sci | K computer Fujitsu SPARC64 VIIIfx (8c) + custom | Japan | 705,024 | 10.5 | 93 | 12.7 | 830 |
| 3 | DOE / OS Argonne Nat Lab | Mira, BlueGene/Q (16c) + custom | USA | 786,432 | 8.16 | 81 | 3.95 | 2069 |
| 4 | Leibniz Rechenzentrum | SuperMUC, Intel (8c) + IB | Germany | 147,456 | 2.90 | 90* | 3.52 | 823 |
| 5 | Nat. SuperComputer Center in Tianjin | Tianhe-1A, NUDT Intel (6c) + Nvidia GPU (14c) + custom | China | 186,368 | 2.57 | 55 | 4.04 | 636 |
| 6 | DOE / OS Oak Ridge Nat Lab | Jaguar, Cray AMD (16c) + custom | USA | 298,592 | 1.94 | 74 | 5.14 | 377 |
| 7 | CINECA | Fermi, BlueGene/Q (16c) + custom | Italy | 163,840 | 1.73 | 82 | .821 | 2099 |
| 8 | Forschungszentrum Juelich (FZJ) | JuQUEEN, BlueGene/Q (16c) + custom | Germany | 131,072 | 1.38 | 82 | .657 | 2099 |
| 9 | Commissariat a l'Energie Atomique (CEA) | Curie, Bull Intel (8c) + IB | France | 77,184 | 1.36 | 82 | 2.25 | 604 |
| 10 | Nat. Supercomputer Center in Shenzhen | Nebulea, Dawning Intel (6) + Nvidia GPU (14c) + IB | China | 120,640 | 1.27 | 43 | 2.58 | 493 |
| 500 | Energy Comp | IBM Cluster, Intel + IB | Italy | 4096 | .061 | 93* | | |

# Top in UK, 25 Systems

| Rank | Name | Computer | Site | Cores | Rmax |
|---|---|---|---|---|---|
| 13 | Blue Joule | BlueGene/Q, Power BQC 16C | STFC - Daresbury Lab | 114688 | 1207844 |
| 20 | DiRAC | BlueGene/Q, Power BQC 16C | University of Edinburgh | 98304 | 1035295 |
| 32 | HECToR | Cray XE6, Opteron 6276 16C | University of Edinburgh | 90112 | 660243 |
| 34 | | Power 775, POWER7 8C 3.84 GHz | ECMWF | 24576 | 548996 |
| 35 | | Power 775, POWER7 8C 3.83GHz | ECMWF | 24576 | 548996 |
| 43 | | Power 775, POWER7 8C 3.836GHz | UK Met Office | 18432 | 411747 |
| 51 | | Power 775, POWER7 8C 3.84 GHz | UK Met Office | 15360 | 343123 |
| 93 | Darwin | Dell PowerEdge | Cambridge University | 9728 | 183379 |
| 114 | Blue Wonder | iDataPlex DX360M4, Xeon E5-2670 | STFC - Daresbury Lab | 8192 | 158696 |
| 134 | | iDataPlex DX360M4, Xeon E5-2670 | Durham University | 6720 | 130180 |
| 143 | | Power 775, POWER7 8C 3.836GHz | UK Met Office | 5120 | 124756 |
| 144 | Blackthorn | Bullx B500 Cluster, Xeon X56xx | Atomic Weapons Est | 12936 | 124600 |
| 153 | | Power 575, p6 4.7 GHz, Infiniband | ECMWF | 8320 | 115900 |
| 154 | | Power 575, p6 4.7 GHz, Infiniband | ECMWF | 8320 | 115900 |
| 155 | | Cluster Platform 3000 BL280c G6 | Government | 19536 | 115380 |
| 159 | EMERALD | Cluster Platform SL390s G7, NVIDIA | STFC - RAL | 6960 | 114400 |
| 203 | | xSeries iDataPlex, Xeon E5645 6C | U of Southampton | 11088 | 94692 |
| 237 | | BladeCenter HS22 Cluster, Xeon | Financial Institution (P) | 15744 | 88718 |
| 291 | N8 | Raackable Cluster, Xeon E5-2670 8C | Leeds N8 | 5088 | 81170 |
| 292 | | iDataPlex DX360M3, Xeon X5650 6C | Financial Institution (P) | 14400 | 81144 |
| 350 | | BladeCenter HS22 Cluster (WM), | Classified | 13356 | 75261 |
| 396 | | xSeries x3650M3, Xeon X56xx 2.93 | Bank (J) | 12312 | 69378 |
| 397 | | xSeries x3650M3, Xeon X56xx 2.93 | Bank (J) | 12312 | 69378 |
| 405 | | Cluster Platform 3000 BL460c G7 | IT Service Provider | 7968 | 68638 |
| 440 | | Cluster Platform 4000 BL685c G7 | IT Service Provider | 14556 | 65825 |

# Performance Development in Top500



# Major Changes to Algorithms/Software

- **Must rethink the design of our algorithms and software**
  - **Manycore and Hybrid architectures (58) are disruptive technology**
    - **Similar to what happened with cluster computing and message passing**
  - **Rethink and rewrite the applications, algorithms, and software**
  - **Data movement is expensive**
  - **Flops are cheap**

# The High Cost of Data Movement

- Flop/s or percentage of peak flop/s become much less relevant

**Approximate power costs (in picoJoules)**

| | 2011 |
|---|---|
| DP FMADD flop | 100 pJ |
| DP DRAM read | 4800 pJ |
| Local Interconnect | 7500 pJ |
| Cross System | 9000 pJ |

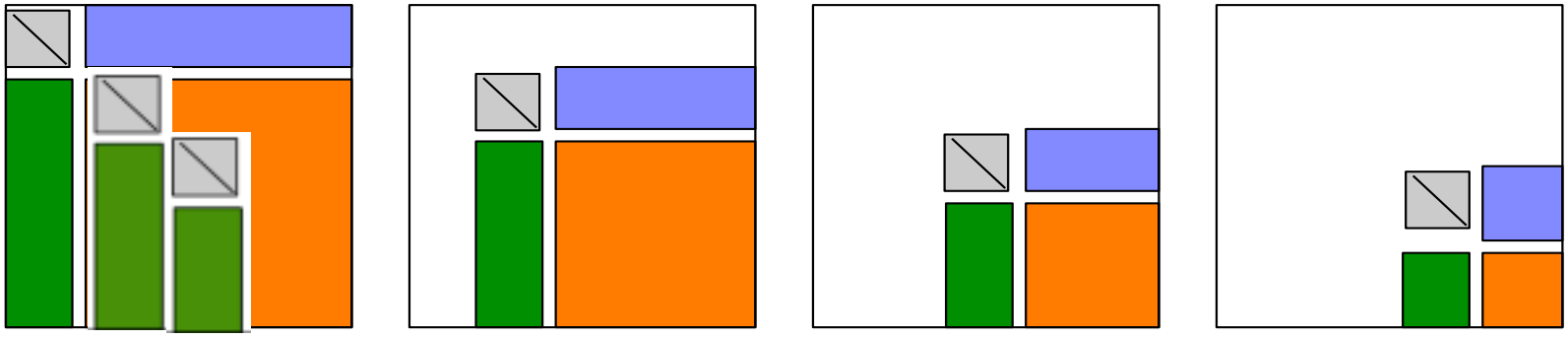
Source: John Shalf, LBNL

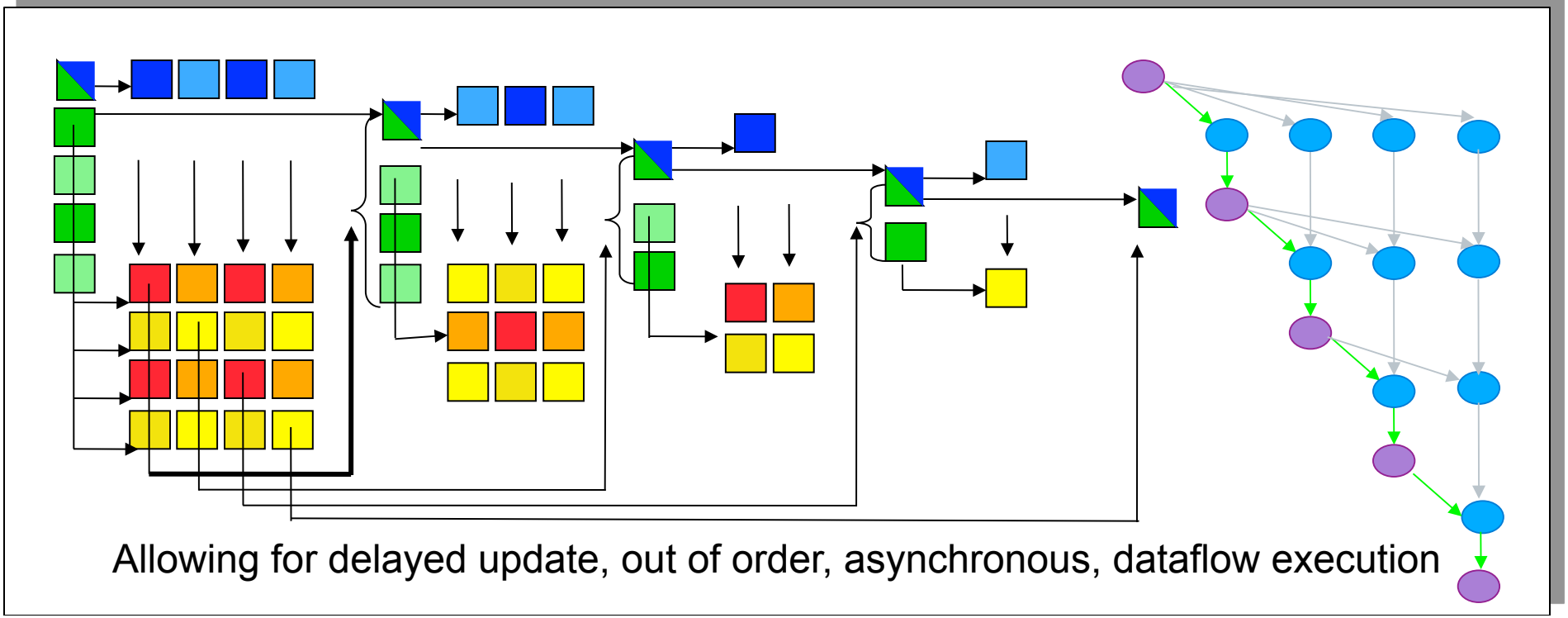
- Algorithms & Software: minimize data movement; perform more work per unit data movement.

# Critical Issues at Peta & Exascale for Algorithm and Software Design

- Synchronization-reducing algorithms
  - Break Fork-Join model
- Communication-reducing algorithms
  - Use methods which have lower bound on communication
  - Follows from the 1981 Hong & Kung paper on minimum communication for matrix multiply
- Mixed precision methods
  - 2x speed of ops and 2x speed for data movement
- Autotuning
  - Today's machines are too complicated, build "smarts" into software to adapt to the hardware
- Fault resilient algorithms
  - Implement algorithms that can recover from failures/bit flips
- Reproducibility of results
  - Today we can't guarantee this. We understand the issues, but some of our "colleagues" have a hard time with this.

# Synchronization (in LAPACK LU)

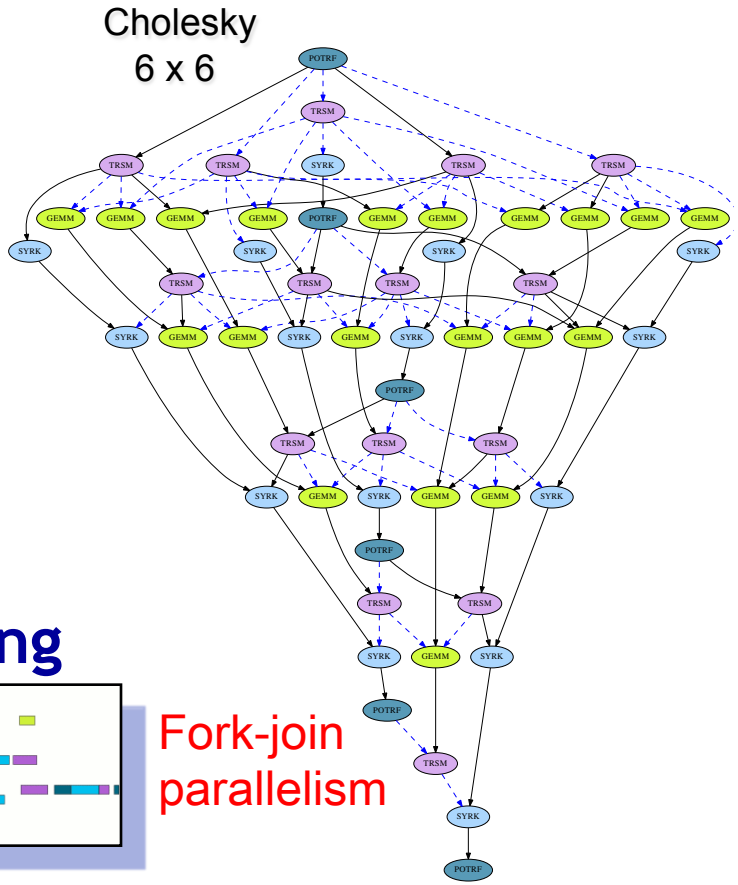


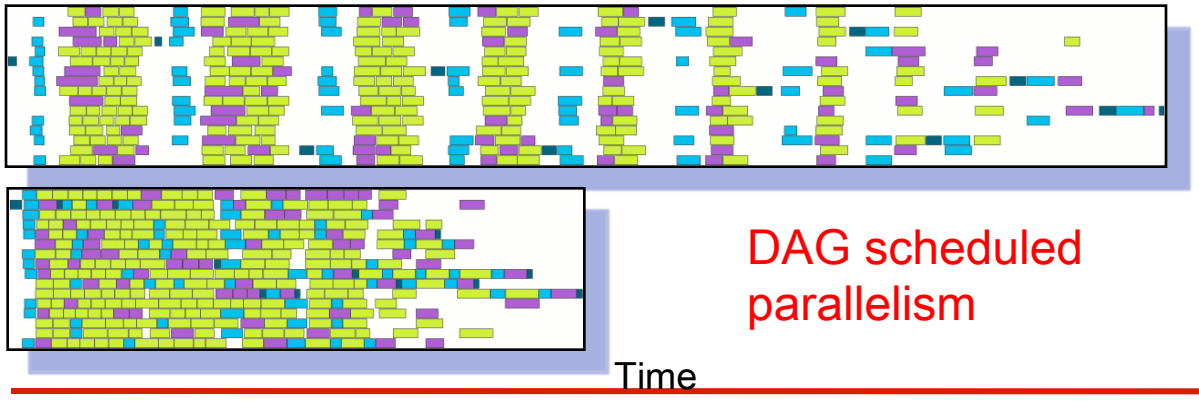



Allowing for delayed update, out of order, asynchronous, dataflow execution

# PLASMA/MAGMA: Parallel Linear Algebra s/w for Multicore/Hybrid Architectures

- **Objectives**
  - High utilization of each core
  - Scaling to large number of cores
  - Synchronization reducing algorithms
- **Methodology**
  - Dynamic DAG scheduling (QUARK)
  - Explicit parallelism
  - Implicit communication
  - Fine granularity / block data layout
- **Arbitrary DAG with dynamic scheduling**

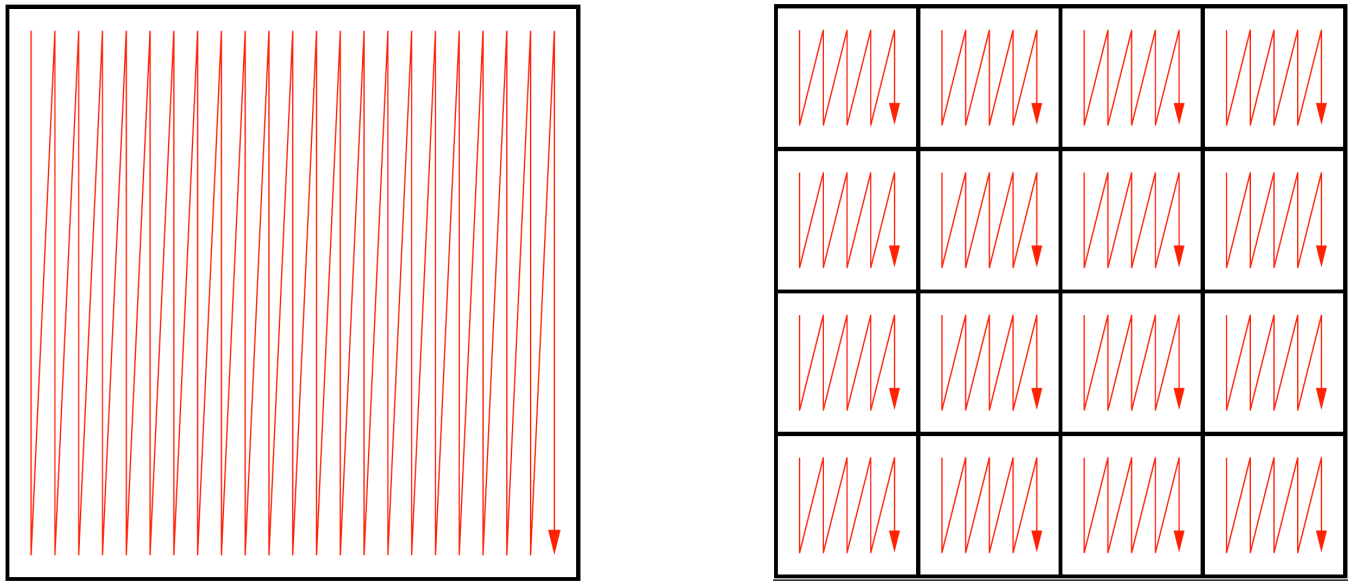


# What's New in PLASMA

| FUNCTIONALITY | COVERAGE |
|---|---|
| Linear Systems of Equations | Cholesky,<br>**LU (partial pivoting, Recursive Parallel Panel)**<br>**LDL (Randomization Pivoting)** |
| Explicit Matrix Inversion | **Cholesky (inter-procedural pipelining)**<br>**LU (partial pivoting, Recursive Parallel Panel, inter-procedural pipelining)** |
| Least Squares | **QR & LQ (Communication Reducing)** |
| Mixed Precision Iterative Refinement | Cholesky, LU (linear systems); QR, LQ (least squares) |
| Symmetric Eigenvalue Problem | **Eigenvalues and eigenvectors (Communication Reducing, two-stage)** |
| Symmetric Generalized Eigenvalue Problem | **Generalized eigenvalues and eigenvectors (Communication Reducing, two-stage)** |
| Singular Value Problem | **Singular values and singular vectors (Communication Reducing, two-stage)** |
| Level 3 Tile BLAS | GEMM, HEMM, HER2K, HERK, SYMM, SYR2K, SYRK, TRMM, TRSM |
| In-Place Layout Translation | **CM, RM, CCRB, CRRB, RCRB, RRRB (parallel)** |
| Norm calculation | 1, ∞, Frobenius |

# Data Layout is Critical for Performance



- Tile data layout where each data tile is contiguous in memory
- Decomposed into several fine-grained tasks, which better fit the memory of the small core caches

# Recursive LU Panel Factorization

- Lower order term $O(n^2)$
- Poses a problem in the parallel setting
- Avoids tuning parameters

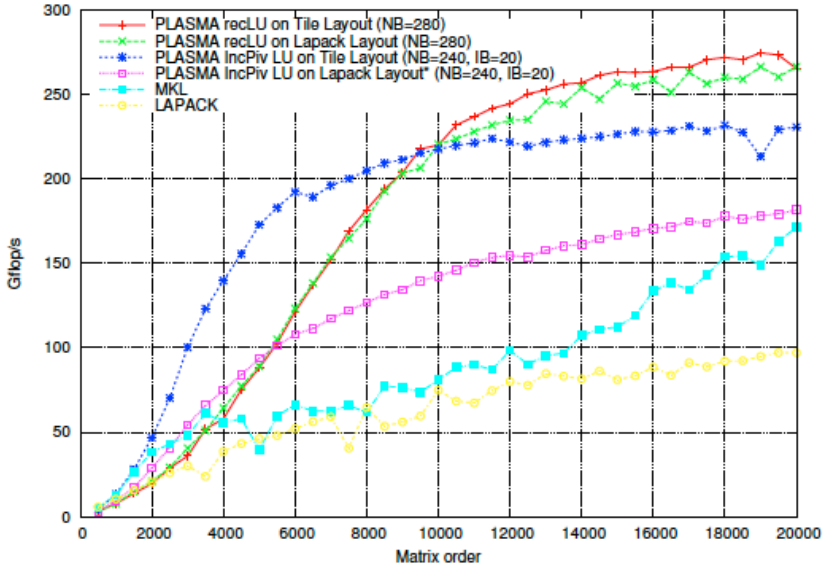
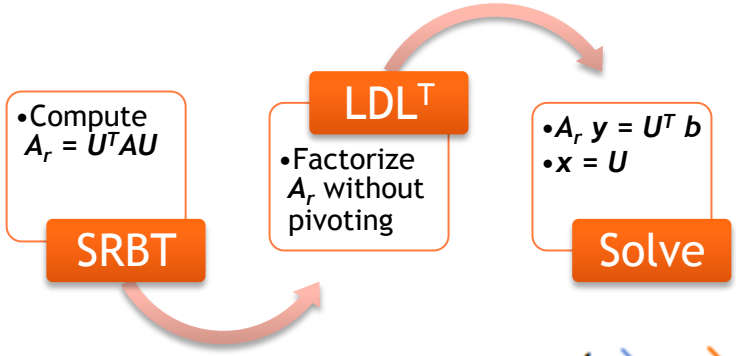



Figure 10. Performances of DGETRF on *MagnyCour-48*.

```
function xGETRFR(M, N, column) {
  if N == 1 {                                  single column, recursion stops
    idx = split_IxAMAX(...)                    compute local maximum of modulus
    gidx = combine_IxAMAX(idx)                 combine local results
    split_xSCAL(...)                           scale local data
  } else {
    xGETRFR(M, N/2, column)                    recursive call to factor left half
    xLASWP(...)                                pivoting forward
    split_xTRSM(...)                           triangular solve
    split_xGEMM(...)                           Schur's complement
    xGETRFR(M, N-N/2, column+N/2)              recursive call to factor right half
    xLASWP(...)                                pivoting backward
```

# Randomize Instead of Pivoting

- *A* is symmetric indefinite. Given the factorization $A = LDL^T$, where L is unit lower triangular and D is diagonal
- Solve $Ax = b$ by solving successively
  $Lz = b, \quad Dy = z, \quad L^T x = y$
- Not stable
  - To ensure stability usually pivoting is used such as
    $PAP^T = LDL^T$, where P is a permutation matrix
  - Pivoting complicated and expensive
- Avoid pivoting using Random Butterfly Transformations (RBT)
- Apply iterative refinement to solution
  - If non-convergence call LU on symmetric matrix
- Performance similar to Cholesky




$$\frac{1}{\sqrt{2}} \begin{pmatrix} R & S \\ R & -S \end{pmatrix} \longrightarrow$$

R and S are random diagonal matrices

| Matrix | Cond A | NP | PP | SRBT (IR) |
|---|---|---|---|---|
| condex | $10^2$ | $10^{-15}$ | $10^{-15}$ | $10^{-15}$ (0) |
| fiedler | $10^5$ | – | $10^{-15}$ | $10^{-15}$ (0) |
| orthog | $10^0$ | $10^{-1}$ | $10^{-14}$ | $10^{-16}$ (1) |
| randcorr | $10^3$ | $10^{-16}$ | $10^{-16}$ | $10^{-16}$ (0) |
| augment | $10^4$ | $10^{-15}$ | $10^{-15}$ | $10^{-16}$ (1) |
| prolate | $10^{18}$ | $10^{-15}$ | $10^{-16}$ | $10^{-15}$ (0) |
| toeppd | $10^7$ | $10^{-16}$ | $10^{-16}$ | $10^{-16}$ (0) |
| ris | $10^0$ | – | $10^{-15}$ | $10^{-1}$ (10) |
| $\lvert i-j \rvert$ | $10^5$ | $10^{-15}$ | $10^{-15}$ | $10^{-14}$ (0) |
| max(i,j) | $10^6$ | $10^{-14}$ | $10^{-15}$ | $10^{-14}$ (0) |
| Hadamard | $10^0$ | $10^0$ | $10^0$ | $10^{-15}$ (0) |
| rand0 | $10^5$ | $10^{-12}$ | $10^{-14}$ | $10^{-15}$ (1) |
| rand1 | $10^5$ | – | $10^{-13}$ | $10^{-15}$ (1) |
| rand2 | $10^5$ | – | $10^{-14}$ | $10^{-15}$ (1) |
| rand3 | $10^4$ | $10^{-13}$ | $10^{-14}$ | $10^{-15}$ (1) |

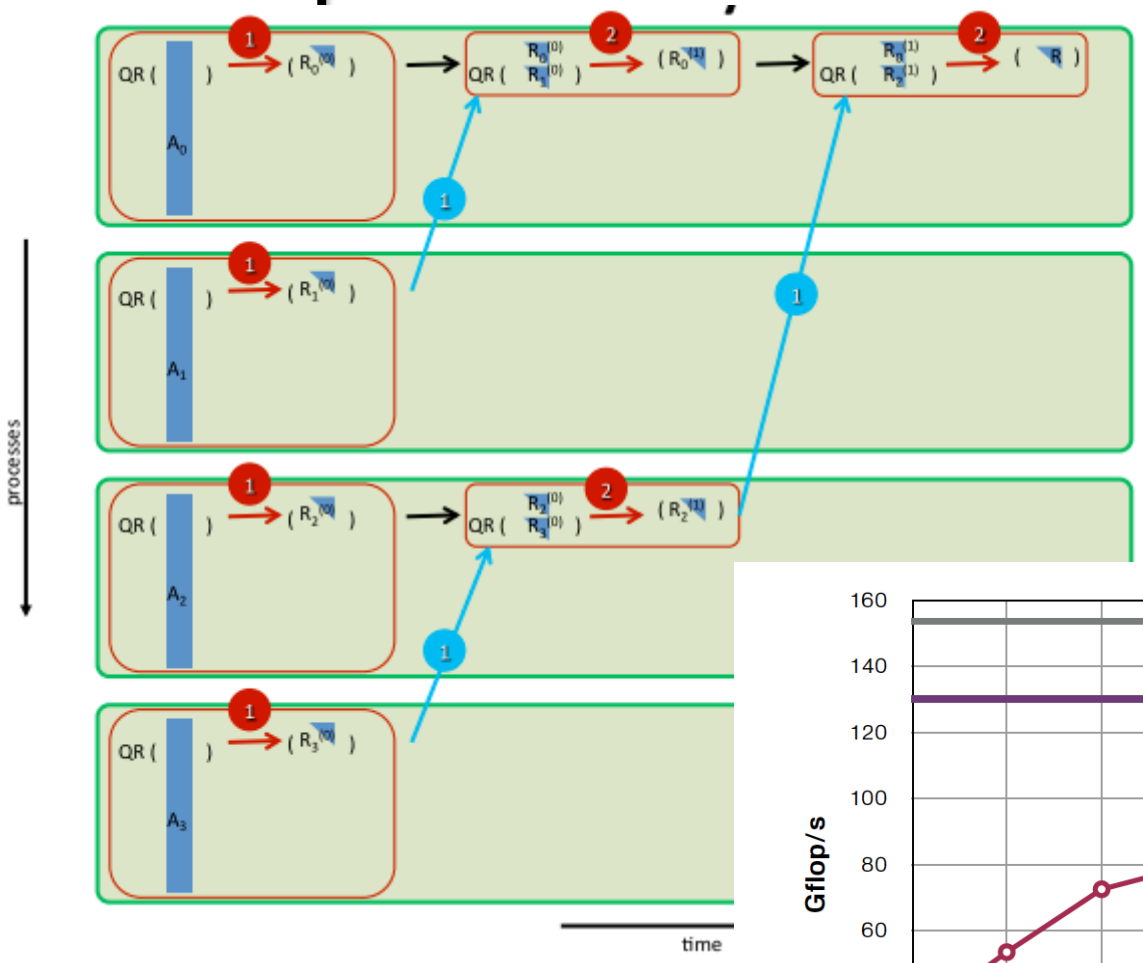
# Same Idea for LU

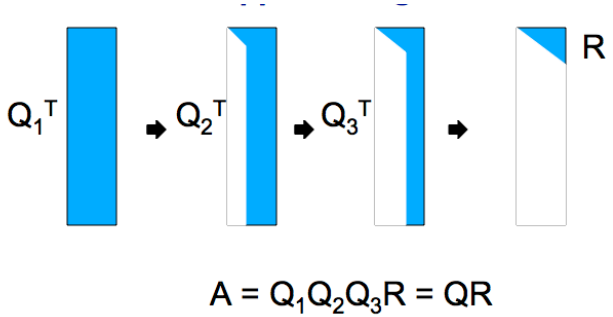
Comparison of errors on linear system solution using PRBT and other solvers

(Higham's collection of matrices - size 1024) $\max_i \frac{|Ax-b|_i}{(|A|\cdot|x|+|b|)_i}$

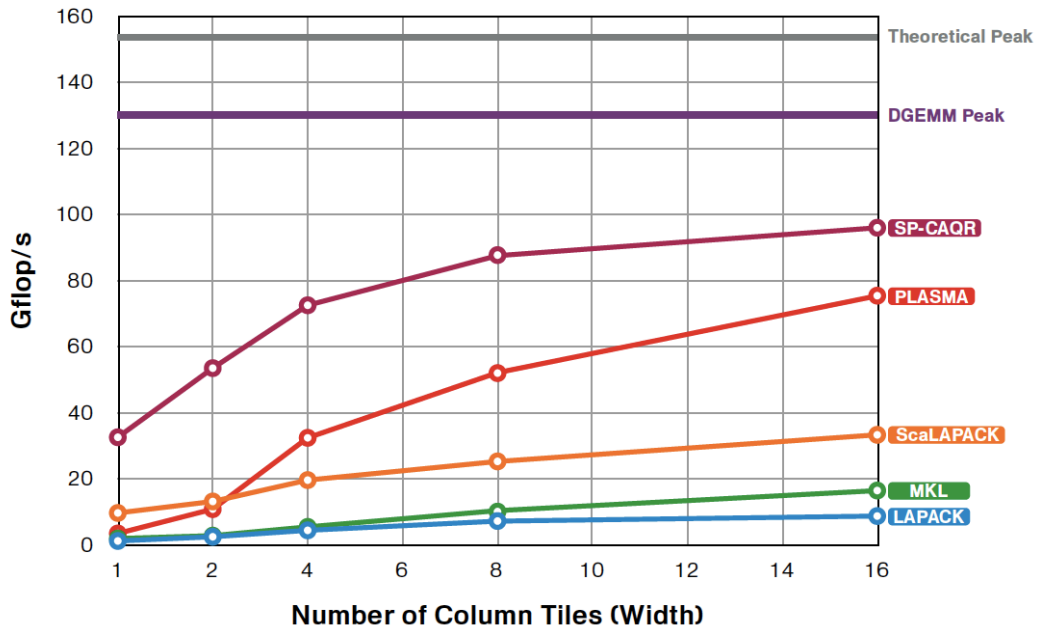
| Matrix | Cond | GENP | GEPP | QR | PRBT | REC | IR |
|---|---|---|---|---|---|---|---|
| augment | $4 \cdot 10^{4}$ | $1.28 \cdot 10^{-14}$ | $2.28 \cdot 10^{-15}$ | $2.99 \cdot 10^{-16}$ | $2.81 \cdot 10^{-16}$ | 1 | 1 |
| gfpp | $5 \cdot 10^{2}$ | $9.01 \cdot 10^{-01}$ | $6.88 \cdot 10^{-01}$ | $1.06 \cdot 10^{-16}$ | $1.27 \cdot 10^{-16}$ | 1 | 1 |
| chebspec | $2 \cdot 10^{14}$ | $1.19 \cdot 10^{-15}$ | $3.29 \cdot 10^{-16}$ | $5.22 \cdot 10^{-15}$ | $3.23 \cdot 10^{-14}$ | 1 | 0 |
| circul | $1 \cdot 10^{3}$ | $1.74 \cdot 10^{-13}$ | $1.66 \cdot 10^{-15}$ | $2.66 \cdot 10^{-15}$ | $2.66 \cdot 10^{-15}$ | 1 | 0 |
| condex | $1 \cdot 10^{2}$ | $7.32 \cdot 10^{-15}$ | $5.98 \cdot 10^{-15}$ | $8.34 \cdot 10^{-15}$ | $6.50 \cdot 10^{-15}$ | 1 | 0 |
| fiedler | $7 \cdot 10^{5}$ | Fail | $2.11 \cdot 10^{-15}$ | $1.54 \cdot 10^{-14}$ | $7.90 \cdot 10^{-15}$ | 1 | 0 |
| Hadamard | $1 \cdot 10^{0}$ | $0 \cdot 10^{0}$ | $0 \cdot 10^{0}$ | $7.58 \cdot 10^{-16}$ | $8.33 \cdot 10^{-15}$ | 1 | 0 |
| normaldata | $3 \cdot 10^{4}$ | $2.03 \cdot 10^{-12}$ | $6.30 \cdot 10^{-15}$ | $2.38 \cdot 10^{-16}$ | $3.30 \cdot 10^{-16}$ | 1 | 1 |
| orthog | $1 \cdot 10^{0}$ | $5.64 \cdot 10^{-01}$ | $4.33 \cdot 10^{-15}$ | $3.70 \cdot 10^{-16}$ | $4.31 \cdot 10^{-16}$ | 2 | 1 |
| randcorr | $3 \cdot 10^{3}$ | $5.12 \cdot 10^{-16}$ | $4.04 \cdot 10^{-16}$ | $5.73 \cdot 10^{-16}$ | $5.92 \cdot 10^{-16}$ | 1 | 0 |
| toeppd | $7 \cdot 10^{5}$ | $2.53 \cdot 10^{-13}$ | $2.60 \cdot 10^{-15}$ | $8.39 \cdot 10^{-15}$ | $5.71 \cdot 10^{-15}$ | 1 | 0 |
| Foster | $5 \cdot 10^{2}$ | $1 \cdot 10^{0}$ | $1 \cdot 10^{0}$ | $1.90 \cdot 10^{-16}$ | $3.30 \cdot 10^{-16}$ | 2 | 1 |
| | | No Pivoting LU | Conventional Pivoting LU | QR Factorization | Randomize No Pivoting LU + Iterative Refinement | | |

# Communication Avoiding QR Example




$A = Q_1 Q_2 Q_3 R = QR$




Quad-socket, quad-core machine Intel Xeon EMT64 E7340 at 2.39 GHz.
Theoretical peak is 153.2 Gflop/s with 16 cores.
Matrix size 51200 by 3200

# Summary

- These are old ideas (today SMPss, StarPU, Charm++, ParalleX, Swarm,…)
- Major Challenges are ahead for extreme computing
  - Power
  - Levels of Parallelism
  - Communication
  - Hybrid
  - Fault Tolerance
  - … and many others not discussed here
- Not just a programming assignment.
- This opens up many new opportunities for applied mathematicians and computer scientists