An efficient distributed randomized solver with application to large dense linear systems

Dulceneia Becker and
George Bosilca and
Anthony Danalis and
Jack Dongarra
Innovative Computing Laboratory
University of Tennessee, Knoxville, USA
Email: [bosilca,adanalis,dbecker7,dongarra]@eecs.utk.edu

Marc Baboulin
Inria Saclay and University Paris-Sud
Orsay, France
Email: marc.baboulin@inria.fr

***Abstract*—Randomized algorithms are gaining ground in high performance computing applications as they have the potential to outperform deterministic methods, while still providing accurate results. In this paper, we propose a randomized algorithm for distributed multicore architectures to efficiently solve large dense symmetric indefinite linear systems that are encountered, for instance, in parameter estimation problems or electromagnetism simulations. This solver combines an efficient implementation of a multiplicative preconditioning based on recursive random matrices, with a runtime (DAGuE) that automatically adjusts data structures, data mappings, and the scheduling as systems scale up. Both the solver and the supporting runtime environment are innovative. To our knowledge, this is the first distributed solver for large dense symmetric indefinite systems, and the randomization approach associated with this solver has never been used in public domain software for such systems. The underlying runtime framework allows seamless data mapping and task scheduling, mapping its capabilities to the underlying hardware features of heterogeneous distributed architectures. We show that the performance of our software is similar to that of symmetric definite systems, but requires only half the execution time and half the amount of data storage of a general dense solver.**

Keywords: randomized algorithms, distributed linear algebra solvers, symmetric indefinite systems, LDL^T factorization, DAGuE runtime

I. Introduction

The last several years saw the development of randomized algorithms in high performance computing applications. This increased interest is motivated by the fact that the resulting algorithms are able to outperform deterministic methods while still providing very accurate results (see e.g. random sampling algorithms that can be applied to least squares solutions or low-rank matrix approximation [1]). In addition to being easier to analyze, the main advantage of such algorithms is that they can lead to much faster solution by performing a smaller number of floating-point operations (e.g. [2]), or by involving less communication (e.g. [3]). As a result, they potentially allow domain scientists to address larger simulations.

However, to be of full interest for applications, these randomized algorithms must be able to exploit the computing capabilities of current highly distributed parallel systems, which can commonly achieve performance of more than one Tflop/s per node. Since randomized algorithms are supposed to be useful for very large problems, the main challenge for them is to exploit efficiently these distributed computing units and their associated memories. As a matter of fact, large-scale linear algebra solvers from standard parallel distributed libraries like ScaLAPACK [4] often suffer from expensive inter-node communication costs. Another important requirement is to be able to schedule such algorithms dynamically.

The advent of multicore processors has undermined the dominance of the SPMD programming style, reviving interest in more flexible approaches such as dataflow approaches. Indeed, several projects [5], [6], [7], [8], [9], mostly in the field of numerical linear algebra, revived the use of DAGs, as an approach to tackle the challenges of harnessing the power of multicore and hybrid platforms. In [10], an implementation of a tiled algorithm based on dynamic scheduling for the LU factorization on top of UPC is proposed. [11] uses a static scheduling for the Cholesky factorization on top of MPI to evaluate the impact of data representation structures. All these projects propose ad-hoc solutions to the challenging problem of harnessing all the computing capabilities available on today's heterogeneous platforms, solutions that do not expose enough flexibility to be generalized outside their original algorithmic design space. In the DAGuE project [12] we address this problem in a novel way. Using a layered runtime, we decouple the algorithm itself from the data distribution, as the algorithm is entirely expressed as flows of data, and from the underlying hardware, allowing the developer to focus solely on the algorithmic level without constraints

regarding current hardware trends. The depiction of the algorithm uses a concise symbolic representation (similar to Parameterized Task Graph (*PTG*) proposed in [13]), requiring minimal memory for the DAG representation and providing extreme flexibility to quickly follow the flows of data starting from any task in the DAG, without having to unroll the entire DAG. As a result, the DAG unrolls on-demand, and each participant never evaluates parts of the DAG pertaining to tasks executing on other resources, thereby sparing memory and compute cycles. Additionally, the symbolic representation enables the runtime to dynamically discover the communications required to satisfy remote dependencies, on the fly, without a centralized coordination. Finally, the runtime provides a heterogeneous environment where tasks can be executed on hybrid resources based on their availability.

We illustrate in this paper how randomized algorithms for linear system solutions can exploit distributed parallel systems in order to address large-scale problems. In particular, we will show how our approach automatically adjusts data structures, mapping, and scheduling as systems scale up. The application we consider in this paper is the solution of large dense symmetric indefinite systems. Even though dense symmetric indefinite matrices are less common than sparse ones, they do arise in practice in parameter estimation problems, when considering the augmented system approach [14, p. 77], and in constrained optimization problems [15, p. 230].

For instance, the symmetric indefinite linear system

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix} \tag{1}$$

arises from the optimization problem

$$\min \frac{1}{2} \|r - b\|_2{}^2 \text{ subject to } A^T b = 0 \tag{2}$$

where x is the vector of Lagrange multipliers, $r = b - Ax$ is the residual, and I is the identity matrix.

Another class of applications is related to simulations in electromagnetics, when the 3-D Maxwell equations are solved in the frequency domain. The goal is to solve the Radar Cross Section (RCS) problem that computes the response of an object to the wave. In that case, the discretization by the Boundary Element Method (BEM) yields large dense symmetric complex systems which are non Hermitian [16].

Many of the existing methods do not exploit the symmetry of the system, which results in extra computation and extra storage. The problem size commonly encountered for these simulations is a few hundreds of thousands. When the matrix does not fit into the core memories of the computer, this may require the use of out-of-core methods that use disk space as auxiliary memory, which inhibits performance (note that an option for very large size is iterative/multipole methods, for which the matrix has not to be stored). Our in-core solver will be an alternative to these existing methods by minimizing the number of arithmetical operations and data storage.

Symmetric indefinite systems are classically solved using a LDL$^{\mathrm{T}}$ factorization

$$PAP^T = LDL^T \tag{3}$$

where P is a permutation matrix, A is a symmetric square matrix, L is unit lower triangular, and D is block-diagonal, with blocks of size 1×1 or 2×2.

To our knowledge, there is no parallel distributed solver in public domain libraries (e.g ScaLAPACK) for solving dense symmetric indefinite systems using factorization (3). This is why, in many situations, large dense symmetric indefinite systems are solved via LU factorization, resulting in an algorithm that requires twice as much, both arithmetical operations and storage than LDL$^{\mathrm{T}}$. Moreover, as mentioned in [14], the pivoting strategy adopted in LAPACK [17] (based on the Bunch-Kaufman algorithm [18]) does not generally give a stable method for solving the problem (2), since the perturbations introduced by roundoff do not respect the structure of the system (1).

The approach we propose in this paper avoids pivoting thanks to a randomization technique described in [19]. The main advantage of randomizing here is that we can avoid the communication overhead due to pivoting (this communication cost can represent up to 40% of the global factorization time depending on the architecture, and on the problem size [3]). The resulting distributed solver enables us to address the large dense simulations mentioned before. To our knowledge, this is the first parallel distributed solver in public domain software for dense symmetric indefinite systems. It is combined with an innovative runtime system that allows seamless data mapping and task scheduling on heterogeneous distributed architectures.

II. A DISTRIBUTED RANDOMIZED ALGORITHM

In Sections II-A and II-B, we briefly present the LDL$^{\mathrm{T}}$ factorization and the symmetric random butterfly transformation (SRBT). In Section II-C, we describe the kernels required by both methods, and in Section II-D we explain how these kernels are implemented using the *DAGuE* approach.

A. LDL$^{\mathrm{T}}$ Factorization

When no pivoting is applied (*i.e.* $P = I$), the LDL$^{\mathrm{T}}$ factorization given by Equation (3) can be written as

$$A = LDL^T \tag{4}$$

For this particular case, D is diagonal (instead of block-diagonal), and L remains unit lower triangular[1].

In this work, a tiled LDL$^{\mathrm{T}}$ factorization is used [19]. The tiled algorithm starts by decomposing A in $NT \times NT$ tiles (blocks), such as

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1,NT} \\ A_{21} & A_{22} & \dots & A_{2,NT} \\ \vdots & \vdots & \ddots & \vdots \\ A_{NT,1} & A_{NT,2} & \dots & A_{NT,NT} \end{bmatrix}_{N \times N} \quad (5)$$

where each A_{ij} is a tile of size $NB \times NB$. The same decomposition can be applied to L and D. Upon this decomposition and using the principle of the Schur complement, a series of tasks can be generated to calculate each L_{ij} and D_{ii}. These tasks can be executed out of order, as long as dependencies are observed, rendering parallelism.

The decomposition into tiles allows the computation to be performed on small blocks of data that fit into cache. This leads to the need of a reorganization of the data layout. The so-called tile layout reorders data in such a way that all data of a single block is contiguous in memory. The tasks can either be statically scheduled to take advantage of cache locality and reuse or be dynamically scheduled based on dependencies among data and computational resources available.

The tiled algorithm for the LDL$^{\mathrm{T}}$ factorization is based on the following operations:

`xSYTRF`: This LAPACK based subroutine is used to perform the LDL$^{\mathrm{T}}$ factorization of a symmetric tile A_{kk} of size $NB \times NB$ producing a unit triangular tile L_{kk} and a diagonal tile D_{kk}.
Using the notation *input* $\longrightarrow$ *output*, the call `xSYTRF`(A_{kk}, L_{kk}, D_{kk}) will perform

$$A_{kk} \longrightarrow L_{kk}, D_{kk} = \text{LDL}^{\text{T}}\ (A_{kk})$$

`xSYTRF2`: This subroutine first calls `xSYTRF` to perform the factorization of A_{kk} and then multiplies L_{kk} by D_{kk}. The call `xSYTRF2`(A_{kk}, L_{kk}, D_{kk}, W_{kk}) will perform

$$A_{kk} \longrightarrow L_{kk}, D_{kk} = \text{LDL}^{\text{T}}\ (A_{kk}),$$
$$W_{kk} = L_{kk} D_{kk}$$

`xTRSM`: This BLAS subroutine is used to apply the transformation computed by xSYTRF2 to an A_{ik} tile by means of a triangular system solve. The call `xTRSM`(W_{kk}, A_{ik}) performs

$$W_{kk}, A_{ik} \longrightarrow L_{ik} = A_{ik} W_{kk}^{-T}$$

`xSYDRK`: This subroutine is used to update the tiles A_{kk} in the trailing submatrix by means of a matrix-matrix multiply. It differs from `xGEMDM` by taking advantage of the symmetry of A_{kk} and by using only the lower triangular part of A and L. The call `xSYDRK`(A_{kk}, L_{ki}, D_{ii}) performs

$$A_{kk}, L_{ki}, D_{ii} \longrightarrow A_{kk} = A_{ik} - L_{ki} D_{ii} L_{ki}^{T}$$

`xGEMDM`: This subroutine is used to update the tiles A_{ij} for $i \neq j$ in the trailing submatrix by means of a matrix-matrix multiply. The call `xGEMDM`(A_{ij}, L_{ik}, L_{jk}, D_{kk}) performs

$$A_{ij}, L_{ik}, L_{jk}, D_{kk} \longrightarrow A_{ij} = A_{ij} - L_{ik} D_{kk} L_{jk}^{T}$$

Given a symmetric matrix A of size $N \times N$, NT as the number of tiles, such as in Equation (5), and making the assumption that $N = NT \times NB$ (for simplicity), where $NB \times NB$ is the size of each tile A_{ij}, then the tiled LDL$^{\mathrm{T}}$ algorithm can be described as in Algorithm 1.

Algorithm 1 Tile LDL$^{\mathrm{T}}$ Factorization

```
1:  for k = 1 to NT do
2:    xSYTRF2(A_kk, L_kk, D_kk, W_kk)
3:    for i = k + 1 to NT do
4:      xTRSM(W_kk, A_ik)
5:    end for
6:    for i = k + 1 to NT do
7:      xSYDRK(A_kk, L_ki, D_ii)
8:      for j = k + 1 to i − 1 do
9:        xGEMDM(A_ij, L_ik, L_jk, D_kk)
10:     end for
11:   end for
12: end for
```

B. Randomizing symmetric indefinite systems

Let us recall here the main definitions and results related to the randomization approach that is used for symmetric indefinite systems. The randomization of the matrix is based on a technique described in [21] and [3] for general systems and applied in [19] for symmetric indefinite systems. The procedure to solve $Ax = b$, where A is symmetric, using a random transformation and the LDL$^{\mathrm{T}}$ factorization is:

1) Compute $A_r = U^T A U$, with U a random matrix,
2) Factorize $A_r = LDL^T$ (without pivoting),
3) Solve $A_r y = U^T b$ and compute $x = Uy$.

The random matrix U is chosen among a particular class of matrices called *recursive butterfly matrices* and the resulting transformation is referred to as **Symmetric Random Butterfly Transformation** (SRBT).

We recall that a butterfly matrix is defined as any n-by-n matrix of the form:

$$B = \frac{1}{\sqrt{2}} \begin{pmatrix} R & S \\ R & -S \end{pmatrix} \quad (6)$$

[1]See [20] for more details about the LDL$^{\mathrm{T}}$ factorization.

where $n \geq 2$ and R and S are random diagonal and nonsingular $n/2$-by-$n/2$ matrices.

A recursive butterfly matrix U of size n and depth d is a product of the form

$$U = U_d \times \cdots \times U_1, \tag{7}$$

where U_k ($1 \leq k \leq d$) is a block diagonal matrix expressed as

$$U_k = \begin{pmatrix} B_1 & & \\ & \ddots & \\ & & B_{2^{k-1}} \end{pmatrix} \tag{8}$$

each B_i being a butterfly matrix of size $n/2^{k-1}$. In particular U_1 is a butterfly as defined in Formula (6). Note that this definition requires that n is a multiple of 2^d which can always be obtained by "augmenting" the matrix A with additional 1's on the diagonal.

We generate the random diagonal values used in the butterflies as $e^{\rho/10}$, where ρ is randomly chosen in $[-\frac{1}{2}, \frac{1}{2}]$. This choice is suggested and justified in [21] by the fact that the determinant of a butterfly has an expected value 1. Then the random values r_i used in generating butterflies are such that

$$e^{-1/20} \leq r_i \leq e^{1/20}.$$

Using these random values, it is shown in [22] that the 2-norm condition number of the randomized matrix A_r verifies

$$\mathrm{cond}_2(A_r) \ \leq 1.2214^d \mathrm{cond}_2(A) \ , \tag{9}$$

and thus is kept almost unchanged by the randomization. We recall also that the LDL$^{\mathrm{T}}$ algorithm without pivoting is potentially unstable [23, p. 214], due to a possibly large growth factor. We can find in [21] explanations about how recursive butterfly transformations modify the growth factor of the original matrix A. To ameliorate this potential instability, we systematically add in our method a few steps of iterative refinement in the working precision as indicated in [23, p. 232].

A butterfly matrix and a recursive butterfly matrix can be stored in a packed storage using a vector and a matrix, respectively. The computational cost of the randomized matrix depends on the order of the matrix to be transformed, n, and on the number of recursion levels, d. Given that

$$A_r = U^T A U = \prod_{i=1}^{d} U_i^T \ A \ \prod_{i=d}^{1} U_i, \tag{10}$$

for each recursion level k, $U_k^T Q U_k$ must be computed as a block matrix of the form

$$\begin{pmatrix} B_1^T Q_{11} B_1 & \cdots & B_1^T Q_{p1}^T B_p \\ \vdots & \ddots & \vdots \\ B_p^T Q_{p1} B_1 & \cdots & B_p^T Q_{pp} B_p \end{pmatrix} \tag{11}$$

where $p = 2^{k-1}$ and Q is a partial random transformation of A (levels d to $k+1$) given by

$$Q = \prod_{i=k+1}^{d} U_i^T \ A \ \prod_{i=d}^{k+1} U_i$$

Equation (11) requires two computational kernels:

1) symmetric $B^T C B$ with C symmetric, and
2) general $B^T C B'$.

Each matrix expressed in (11) requires p symmetric kernels and $p(p-1)/2$ general (nonsymmetric) kernels operating on matrices of size n/p. Therefore, the number of operations involved in randomizing A by an SRBT of depth d is

$$\begin{aligned} C(n,d) &\simeq \sum_{k=1}^{d} \left(p \cdot 2(n/p)^2 + p(p-1)/2 \cdot 4(n/p)^2 \right) \\ &= 2dn^2 \end{aligned}$$

We will consider a number of recursions d such that $d < \log_2 n \ll n$. Numerical tests described in [19] and performed on a collection of matrices from the Higham's Matrix Computation Toolbox [23] have shown that, in practice, $d = 2$ enables us to achieve satisfying accuracy. Similarly to the product of a recursive butterfly by a matrix, the product of a recursive butterfly by a vector does not require the explicit formation of the recursive butterfly since the computational kernel will be a product of a butterfly by a vector, which involves $\mathcal{O}(n)$ operations. Then the computation of $U^T b$ and Uy can be performed in $\mathcal{O}(dn)$ flops and, for small values of d, can be neglected compared to the $\mathcal{O}(n^3)$ cost of the factorization.

C. Randomization kernels

As described in Section II-B, the matrix A of Equation (3) can be transformed by U of Equation (7) such as

$$Ax = b \ \equiv \ \underbrace{U^T A U}_{A_r} \ \underbrace{U^{-1} x}_{y} \ = \ \underbrace{U^T b}_{c} \tag{12}$$

For simplicity, n (order of matrices U and A) is supposed to be a multiple of 2^d hereafter (if not the system is augmented with additional 1's on the diagonal).

In order to transform the system $Ax = b$, two kernels are required:

$$A_r = U^T A U \tag{13}$$

$$c = U^T b \tag{14}$$

After solving $A_r y = c$, x is obtained by

$$x = Uy \tag{15}$$

Equations (14) and (15) can be taken as particular cases of Equation (13), where the U matrix on the right side of A is an identity matrix. The latter requires the multiplication of B on both sides, while the former ones only require multiplication of B to the left. This means that the data dependencies described in what follows is similar but simpler for Equations (14) and (15). Since the implementation uses the same principle for all three operations, only $U^T AU$ is detailed.

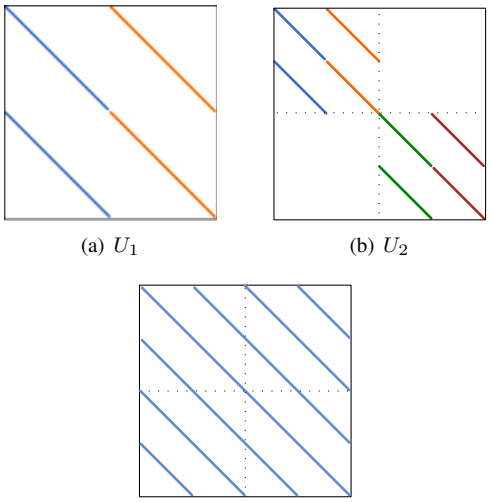
The recursive matrix U of Equation (13), as defined by Equation (10), can be expanded as

$$A_r = U_1^T \times U_2^T \times \cdots \times U_d^T \times A \times U_d \times \cdots \times U_2 \times U_1.$$

Note that U_i is a sparse matrix, with sparsity pattern as shown in Figure 1, and that the matrix product of U_i results in a different sparsity pattern, which depends on the number of levels of recursion. To avoid storing the product of U_i and to maintain the symmetry, the computation can be performed by recursively computing

$$A_r^{(i-1)} = U_i^T A^{(i)} U_i$$

where $A^{(d)} = A$ and $A_r = A_r^{(0)}$.



(a) U_1 (b) U_2

(c) $U_2 \times U_1$

Fig. 1. Sparsity pattern of matrix U.

It can be observed that, for each level, $U_i^T A^{(i)} U_i$ can be written as blocks given by $B_i^T A_{ij} B_j$. For instance, for the second level

$$\begin{aligned} U_2^T A^{(2)} U_2 &= \begin{bmatrix} B_1^T & \\ & B_2^T \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_1 & \\ & B_2 \end{bmatrix} \\ &= \begin{bmatrix} B_1^T A_{11} B_1 & B_1^T A_{12} B_2 \\ B_2^T A_{21} B_1 & B_2^T A_{22} B_2 \end{bmatrix} \end{aligned}$$

Hence, the so-called *core kernel* of a random butterfly transformation is given by

$$B_i^T A_{ij} B_j \tag{16}$$

where A_{ij} is a block of A and B_* is a random butterfly matrix, both of size $m \times m$. The block A_{ij} can either be symmetric (diagonal block, *i.e.* $i = j$) or non-symmetric (off-diagonal block, *i.e.* $i \neq j$).

Recalling that B has a well defined structure

$$B = \begin{bmatrix} R & S \\ R & -S \end{bmatrix} \text{ and } B^T = \begin{bmatrix} R & R \\ S & -S \end{bmatrix}$$

where R and S are diagonal matrices, and given that A_{ij} is divided into four submatrices of same size, such as

$$A_{ij} = \begin{bmatrix} TL & TR \\ BL & BR \end{bmatrix}$$

and that

$$\begin{aligned} W_{TL} &= (TL + BL) + (TR + BR), \\ W_{BL} &= (TL - BL) + (TR - BR), \\ W_{TR} &= (TL + BL) - (TR + BR), \\ W_{BR} &= (TL - BL) - (TR - BR). \end{aligned}$$

Equation (16) can be written as

$$B_i^T A_{ij} B_j = \begin{bmatrix} R \cdot W_{TL} \cdot R & R \cdot W_{TR} \cdot S \\ S \cdot W_{BL} \cdot R & S \cdot W_{BR} \cdot S \end{bmatrix}.$$

Note that only the signs differ in calculating each W_*. Hence, all four cases can be generalized as

$$W = (TL \circ BL) \circ (TR \circ BR). \tag{17}$$

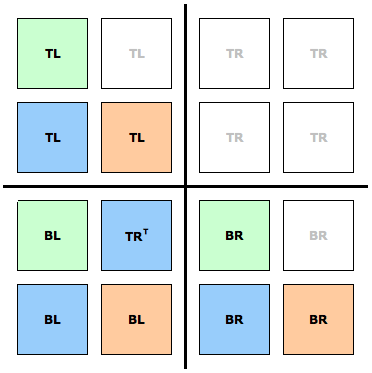
Equation (17) shows that each matrix W_* depends on all four submatrices of A_{ij}. More specifically, any given element of W depends on four elements of A. Therefore, the data dependencies among elements could be depicted as:

$$\underbrace{\left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 2 & 3 \\ 2 & 4 & 5 & 2 & 4 & 5 \\ 3 & 5 & 6 & 3 & 5 & 6 \\ \hline 1 & 2 & 3 & 1 & 2 & 3 \\ 2 & 4 & 5 & 2 & 4 & 5 \\ 3 & 5 & 6 & 3 & 5 & 6 \end{array}\right]}_{Symmetric} \quad \underbrace{\left[\begin{array}{ccc|ccc} 1 & 4 & 7 & 1 & 4 & 7 \\ 2 & 5 & 8 & 2 & 5 & 8 \\ 3 & 6 & 9 & 3 & 6 & 9 \\ \hline 1 & 4 & 7 & 1 & 4 & 7 \\ 2 & 5 & 8 & 2 & 5 & 8 \\ 3 & 6 & 9 & 3 & 6 & 9 \end{array}\right]}_{General}$$

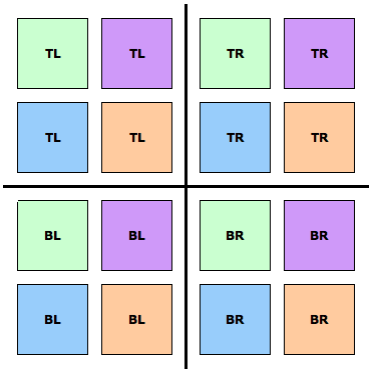
where same numbers means these elements depend on each other. For the symmetric case, the strictly upper triangular part is not calculated, and for this work, not stored either. Hence, elements above the diagonal ($i < j$) must be computed as their transpose, *i.e.* A_{ij} is read and written as A_{ji}.

This can be extended to blocks, or tiles, to comply with the LDL$^{\mathrm{T}}$ algorithm presented in Section II-A. If each number above is taken as a tile (or block), the data

dependencies can be sketched as in Figure 2. The actual computation can be done in different ways; details of the approach adopted in this work are given in Section II-D.



(a) Symmetric kernel



(b) General kernel

Fig. 2. SRBT core kernel ($B_i^T A_{ij} B_j$), data dependencies among tiles of A_{ij}, given by matching colors.

D. Implementation details

We use the *DAGuE* [12] runtime system to implement the SRBT transformation and LDL^{T} factorization, because DAGuE enables the algorithm developer to ignore all system details that do not relate to the algorithm that is being implemented. DAGuE is a data-flow engine that takes as input:

1) a set of kernels that will operate on the user data, in the form of compiled functions inside a library or object files,
2) an algebraic description of the data-flow edges between the tasks that will execute the kernels, in a DAGuE specific format that we refer to as the JDF (for Job Data Flow).

Given this input, *DAGuE* will automatically handle the communication, task placement, memory affinity, load balancing, job stealing, and all other aspects of high performance parallel programming that are critical for achieving good performance, but constitute a nuisance to an algorithm developer.

Implementing SRBT randomization using *DAGuE* :

In *DAGuE*'s data-flow view of the world, the unit of data is not a single element, but a tile. A tile can be a block of contiguous memory, or more generally a block of memory containing data separated by strides in any way that an MPI datatype can describe. In our implementation, we assume that the input matrix, given to us by the calling application, is organized according to the format used in the PLASMA [24] library. That is, tiles of contiguous data are stored in a column major order. As discussed in section II-C the data dependencies of the SRBT are those dictated by Equation (17) and exhibit a symmetry. For SRBT at depth 1, the symmetry is about the horizontal and vertical lines that pass through the center of the matrix. For SRBT at depth 2, each quarter of the matrix is independent, and the symmetry is about the center of each quarter. Clearly, if the tile size is such that the number of tiles is not divisible by 2^d, then the symmetries will cause data dependencies that will cross the tile boundaries.

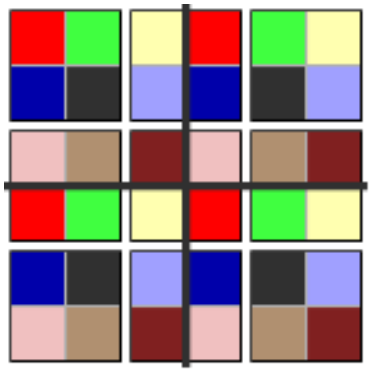


Fig. 4. Tile count (3x3) non divisible by 2^1

As an example, Figure 4 depicts the case where a matrix is divided in 3x3 tiles. Equation (17) demands that each element of each quarter of the matrix is processed with the corresponding elements of the other three quarters of the matrix. However, here, the number of tiles is not divisible by the number of blocks for a butterfly level 1, and therefore processing any tile of the top left quarter requires elements from the other quarters that belong to multiple tiles. The largest blocks of memory that do not cross tile boundaries and have matching blocks are those depicted with the different colors in Figure 4. Clearly, for higher levels of the SRBT transformation, the situation can become even more complicated, as is depicted in Figure 5, where the matrix is organized in 5x5 tiles which is neither divisible by 2^1, nor 2^2.

While choosing a tile size that creates this situation will unequivocally lead to performance degradation (since there will be four times as many messages of

```
Diag(i)
  /* Execution space */
  i = 0 .. mt/2-1

  : A(i,i)

  /* Atl: A from Top Left */
  RW  Atl <- A Reader_TL(i,i) [ inline_c %{ return seg2type(descA, i, i); %}]
          -> A Writer_TL(i,i) [ inline_c %{ return seg2type(descA, i, i); %}]

  /* Abl: A from Bottom Left */
  RW  Abl <- A Reader_BL(i,i) [ inline_c %{ return seg2type(descA, i, i); %}]
          -> A Writer_BL(i,i) [ inline_c %{ return seg2type(descA, i, i); %}]

  /* Abr: A from Bottom Right */
  RW  Abr <- A Reader_BR(i,i) [ inline_c %{ return seg2type(descA, i, i); %}]
          -> A Writer_BR(i,i) [ inline_c %{ return seg2type(descA, i, i); %}]

BODY
...
END
```

Fig. 3. Part of the *DAGuE* input file, JDF, containing the description of the *Diag* task

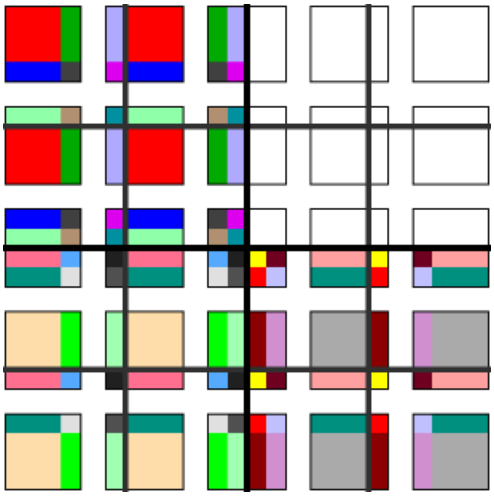


Fig. 5. Tile count (5x5) non divisible by 2^2, or 2^1

significantly smaller size), our implementation of the SRBT transformation supports arbitrary tile sizes and number of tiles. There are two ways to implement this with *DAGuE*. The first is to transfer, for every tile, all tiles that contain matching data, perform the core kernels on the consolidated data, and then transfer the results back to the original location of the tiles. The second way is to logically subdivide the tiles into segments that do not cross tile boundaries (as depicted in Figures 4 and 5 with different colors), and operate on this logical segment space. We chose to implement the latter approach, as it does not perform unnecessary data transfers, since it does not transfter the parts of the tiles that will not be used. We made this choice because the computation performed by SRBT is relatively light weight ($O(n^2)$) and thus it would be difficult to hide the additional overhead of the unnecessary communication.

Besides the performance concerns, there is a correctness concern. As discussed in section II-C, regardless of the size of a contiguous block, all four W_{TL}, W_{BL}, W_{TR}, and W_{BL} require the same data (TL,BL,TR and BR) and also overwrite this data. Clearly, the computation of all four W_* must be performed before any original data is overwritten, using temporary memory, in a double buffering fashion. *DAGuE* provides mechanisms for handling the allocation and deallocation of temporary buffers of size and shape that match a given MPI datatype, simplifying the process.

Figure 3 depicts a snippet of the JDF file that *DAGuE* will use as input to guide the execution of the different tasks and the necessary communication between them. As is implied by its name, task *Diag* will process the segments that lie on the diagonal of the matrix. The execution space of the task goes up only to the middle of the matrix, since, due to the translation symmetry we mentioned earlier, the four quarters of the matrix need to be processed together to generate the result. For this reason, the *Diag* task communicates with the Reader_* tasks to receive the corresponding segments from the three quarters of the matrix, top left (TL), bottom left (BL) and bottom right (BR) and sends back the resulting segments to the Writer_* tasks. Clearly, since the matrix is symmetric and this task is processing diagonal blocks, there is no need to fetch the top right block, since it contains the same data as the bottom left

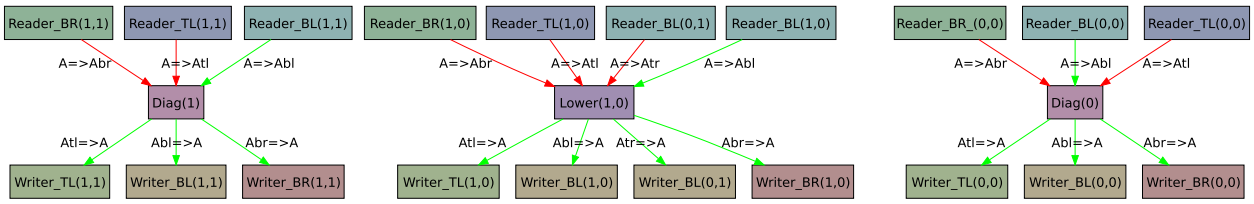


Fig. 6. SRBT DAG for 4x4 tile matrix

block, transposed.

In addition to specifying which tasks constitute communication peers, the JDF specifies the type of data to be exchanged. This can be done with a predetermined, static type, or – as is the case here – using C code that will compute the appropriate type, at runtime, based on parameters of the task, such as "`i`" that identify the segment that is being processed. In this example we make call to a C function called `seg2type()`.

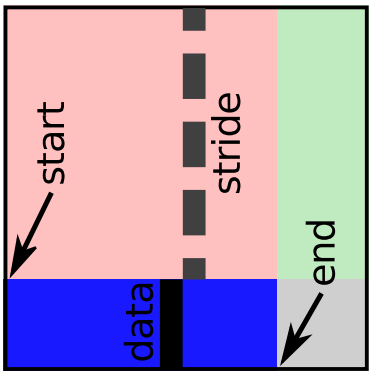


Fig. 7. Segment data inside a tile

The reason for calculating the datatype at runtime becomes clearer if we consider an example from Figure 5. In particular, let us look at the top left tile and consider that we are processing the (blue colored) segment that is in the bottom left of the tile and is short and wide. Figure 7 shows a magnified version of this tile (with fainter colors for readability). Because of the data dependencies of SRBT, the data of the blue segment have to be considered independently of the rest of the tile. However, as is shown in Figure 7, the segment itself is not stored contiguously in memory, but it is stored within a tile. Nevertheless, given the start and end points of the segment, the length of a contiguous block of data, and the stride between blocks of data, we can process the data of the segment in a systematic way and transfer them over the network without wasting bandwidth. The need for dynamic discovery of the type arises because, as can be derived by the multitude of segment shapes and sizes in Figure 5, each segment can potentially have a different type. Since the types depend on the matrix size and the size (or count) of tiles in the matrix, it is impossible to assign those types statically.

Figure 6 shows the Direct Acyclic Graph (DAG) that describes the execution of a level 1 SRBT for a matrix with 4x4 tiles. We can see that the three tasks (*Diag(0)*, *Diag(1)* and *Lower(1,0)*) that process the top left quarter of the matrix do not depend on each other and can therefore run independently. Also, the DAG clearly shows the communication edges between the processing tasks and the corresponding *Reader* and *Writer* tasks that will fetch the data from the other quarters of the matrix. Interestingly, the DAG of Lower(1,0) shows that data is read from the bottom left quarter twice (*Reader_BL(1,0)* and *Reader_BL(0,1)*). This is because the matrix is symmetric and lower triangular, so all data that would belong to the top right quarter are instead found at the transpose location in the bottom left quarter.

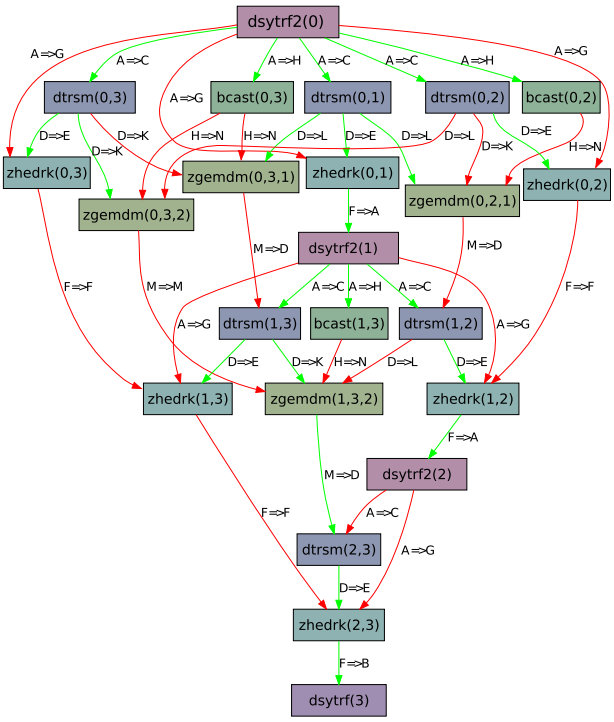
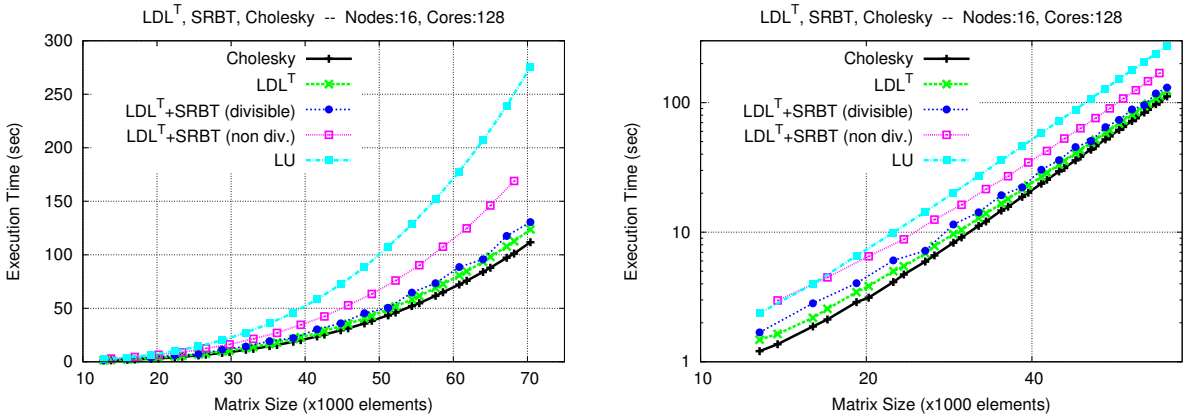


Fig. 8. LDL^T DAG for 4x4 tile matrix



(a) Execution Time (linear)

(b) Execution Time (logarithmic)

Fig. 9. Problem Scaling comparison of LDL$^{\mathrm{T}}$ +SRBT against Cholesky on a $16 \times 8 = 128$ core cluster

Implementing LDL$^{\mathrm{T}}$ using *DAGuE* :

Unlike the SRBT transformation, the LDL$^{\mathrm{T}}$ factorization was rather straight forward to implement in *DAGuE*. In particular, the operation exists as a serial C code in the PLASMA library, similar to the pseudocode shown in Algorithm 1. Because of that, we were able to use the kernels verbatim from the PLASMA library and generate the JDF automatically using the serial code compiler of *DAGuE* (followed by hand-optimizations). The DAG of this factorization on a 4x4 tile matrix is depicted in Figure 8.

III. Performance Results

The performance experiments were performed on a 16 node cluster with Infiniband 20G. Each node features two NUMA Nehalem Xeon E5520 processors clocked at 2.27GHz, with four cores each and 2GB of memory (4GB total per node). The Infiniband NICs are connected through PCI-E and the operating system is Linux 2.6.31.6-2. All experiments presented in this paper were performed using double precision arithmetics.

Figure 9 shows the results of problem scaling when all $16 \times 8 = 128$ cores of the cluster are used. We plot the LDL$^{\mathrm{T}}$ with and without the SRBT transformation and we separate, into two different curves, the cases where the number of tiles is divisible by 2^2 ($d = 2$ since we perform two levels of SRBT) and the cases where it is non-divisible. Also, the figure includes the execution time of Cholesky and LU on a matrix of the same size. These curves help give some context to our measurements. Cholesky can only operate on a subset of the matrices that SRBT+LDL$^{\mathrm{T}}$ can operate on, but it consists of a lower bound in the execution time of SRBT+LDL$^{\mathrm{T}}$, since it performs slightly less operations and communication (because there is no randomization). On the other hand, LU is a general factorization algorithm, but in terms of execution time it provides us with an upper bound, since it performs significantly more operations. We provide both a linear and a logarithmic plot of the data, since the former tends to exaggerate the larger values at the expense of the small ones and vice versa.

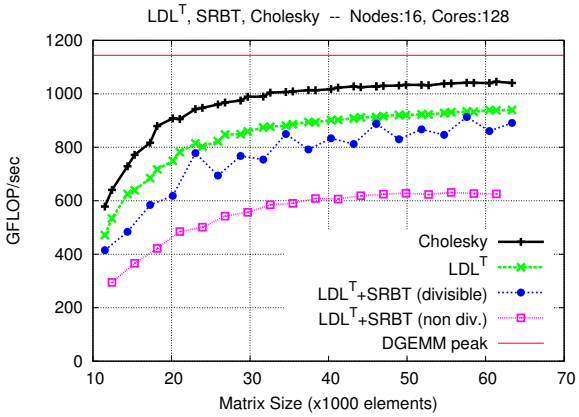


Fig. 10. Problem Scaling (Flops) on a 16x8=128 core cluster

In addition, we show, in Figure 10, the performance of LDL$^{\mathrm{T}}$ +SRBT in Gflops per second. In this figure we have excluded LU, since it algorithmically performs a different number of operations, but we provide the value of the "practical peak". We estimate this peak by measuring the performance of DGEMM (double precission, general, matrix-matrix multiply) operations on the CPU cores without any communication, or synchronization between them. From these graphs we can make the following observations:

1) LDL$^{\mathrm{T}}$, even without the SRBT, is slightly slower than Cholesky because of a) additional synchronization (due to the scaling of the diagonal at the

end of the factorization) and b) the broadcasting of the diagonal elements during the factorization.
2) When the number of tiles is divisible by 2^d, SRBT adds a small overhead, which is especially small when the number of tiles is divisible by 32 (2*NP) since that helps with the process-to-tile mapping.
3) When the number of tiles is not divisible by 2^d, SRBT adds significant overhead.
4) Even in the worst case, SRBT+LDL$^{\mathrm{T}}$ is significantly faster than LU, and the difference in execution time grows with the problem size.

The first conclusion that can be drawn from these observations is that for matrices that are symmetric, but not positive definite, there is a clear advantage in using SRBT+LDL$^{\mathrm{T}}$ instead of LU in terms of performance. Second, it is important, for performance, to have a number of tiles that is divisible by 2^d, i.e., 4, and it is even better if the number of tiles is divisible by the number of nodes that participate in the solution. Since the matrix tile size (and thus the number of tiles in the matrix) is a parameter chosen by the application that calls the *DAGuE* implementation of the algorithm, we provide this insight as a suggestion to the application developers who might wish to use our code.

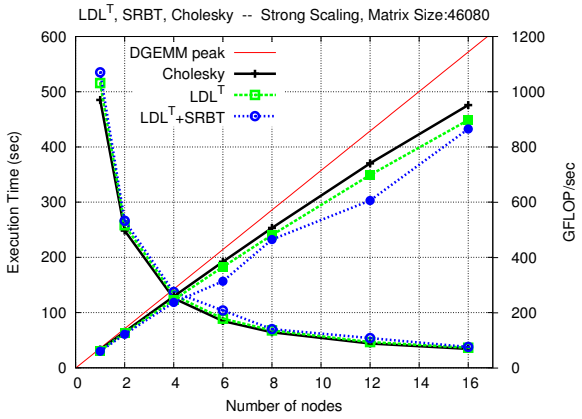


Fig. 11. Strong Scaling using a 46080x46080 matrix

Figure 11 presents strong scaling results for a moderate problem size (N=46K). Both execution time and operations per second are depicted in the graph. Clearly, the algorithm exhibits good performance, similar to that of Cholesky and rather close to the practical peak of the machine. Interestingly, the performance drops for the cases when 6 and 12 nodes were used. The reason for this behavior is that in these cases, the number of nodes does not divide the number of tiles, and thus the (2D block-cyclic) mapping of tiles to nodes creates additional communication.

IV. Conclusion

This paper describes an innovative randomized algorithm for solving large dense symmetric indefinite systems on clusters of multicores. Such solvers have immediate application in several domains, such as the Radar Cross Section problems in Electromagnetism. Existing direct methods, generally based on LU solvers, have a computational and storage cost increased by a factor of two compared to the approach proposed in this paper. The randomization step, used in our approach, allows us to eliminate the pivoting completely and therefore reduce the critical path. Instead, a communication pattern with higher parallelism, which takes place prior to the factorization, is used to apply a preconditioner to the original matrix. The cost of this randomization step is significantly lower than the overall cost of pivoting. The randomization and the LDL$^{\mathrm{T}}$ algorithms have been implemented using the *DAGuE* runtime system. Due to the tile data layout used by the proposed algorithm, the runtime is capable of enforcing a strict data locality scheduling strategy, resulting in high cache locality. As a result the performance obtained is similar to the Cholesky solver, and close to the practical peak of the platform.

We have shown that for symmetric indefinite systems, the proposed algorithm, together with an efficient runtime, is a better performing approach than LU, and therefore a well suited replacement for it.

References

[1] M. W. Mahoney, "Randomized algorithms for matrices and data," *Foundations and Trends in Machine Learning*, vol. 3, no. 2, pp. 123–224, 2011.

[2] A. Avron, P. Maymounkov, and S. Toledo, "Blendenpick: Supercharging LAPACK's least-squares solvers," *SIAM Journal on Scientific Computing*, vol. 32, pp. 1217–1236, 2010.

[3] M. Baboulin, J. Dongarra, J. Herrmann, and S. Tomov, "Accelerating linear system solutions using randomization techniques," *ACM Transactions on Mathematical Software (to appear)*, 2011, LAPACK Working Note 246.

[4] L. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. Whaley, *ScaLAPACK User's Guide*. SIAM, 1997.

[5] A. Buttari, J. J. Dongarra, J. Kurzak, J. Langou, P. Luszczek, and S. Tomov, "The impact of multicore on math software," in *Applied Parallel Computing. State of the Art in Scientific Computing, 8th International Workshop, PARA*, ser. Lecture Notes in Computer Science, vol. 4699. Springer, 2006, pp. 1–10.

[6] E. Chan, F. G. Van Zee, P. Bientinesi, E. S. Quintana-Ortí, G. Quintana-Ortí, and R. van de Geijn, "Supermatrix: a multithreaded runtime scheduling system for algorithms-by-blocks," in *PPoPP '08: Proceedings of the 13th ACM SIGPLAN Symposium on Principles and practice of parallel programming*. ACM, 2008, pp. 123–132.

[7] E. Agullo, J. Demmel, J. J. Dongarra, B. Hadri, J. Kurzak, J. Langou, H. Ltaief, P. Luszczek, and S. Tomov, "Numerical linear algebra on emerging architectures: The PLASMA and MAGMA projects," *Journal of Physics: Conference Series*, vol. 180, 2009.

[8] R. Dolbeau, S. Bihan, and F. Bodin, "HMPP: A hybrid multi-core parallel programming environment," in *Workshop on General Purpose Processing on Graphics Processing Units (GPGPU 2007)*, 2007.

[9] C. Augonnet, S. Thibault, R. Namyst, and P.-A. Wacrenier, "StarPU: a unified platform for task scheduling on heterogeneous multicore architectures," *Concurrency and Computation: Practice and Experience*, vol. 23, no. 2, pp. 187–198, 2011.

[10] P. Husbands and K. A. Yelick, "Multi-threading and one-sided communication in parallel LU factorization," in *Proceedings of the ACM/IEEE Conference on High Performance Networking and Computing, SC 2007, November 10-16, 2007, Reno, Nevada, USA*, B. Verastegui, Ed. ACM Press, 2007.

[11] F. G. Gustavson, L. Karlsson, and B. Kågström, "Distributed SBP Cholesky factorization algorithms with near-optimal scheduling," *ACM Trans. Math. Softw.*, vol. 36, no. 2, pp. 1–25, 2009.

[12] G. Bosilca, A. Bouteiller, A. Danalis, T. Herault, P. Lemarinier, and J. J. Dongarra, "DAGuE: A generic distributed DAG engine for high performance computing," *Parallel Computing*, 2011, http://dx.doi.org/10.1016/j.parco.2011.10.003.

[13] M. Cosnard and E. Jeannot, "Automatic Parallelization Techniques Based on Compact DAG Extraction and Symbolic Scheduling," *Parallel Processing Letters*, vol. 11, pp. 151–168, 2001.

[14] Å. Björck, *Numerical Methods for Least Squares Problems*. SIAM, 1996.

[15] Y. Saad, *Iterative Methods for Sparse Linear Systems*. SIAM, 2000, second edition.

[16] J.-C. Nédélec, *Acoustic and electromagnetic equations. Integral representations for harmonic problems*. New-York: Springer-Verlag, 2001, vol. 144.

[17] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK User's Guide*. SIAM, 1999, third edition.

[18] J. R. Bunch and L. Kaufman, "Some stable methods for calculating inertia and solving symmetric linear systems," *Math. Comput.*, vol. 31, pp. 163–179, 1977.

[19] D. Becker, M. Baboulin, and J. Dongarra, "Reducing the amount of pivoting in symmetric indefinite systems," in *9th International Conference on Parallel Processing and Applied Mathematics (PPAM 2011)*, ser. Lecture Notes in Computer Science, R. Wyrzykowski et. al., Ed., vol. 7203. Heidelberg: Springer-Verlag, 2012, pp. 133–142.

[20] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: The Johns Hopkins University Press, 1996.

[21] D. S. Parker, "Random butterfly transformations with applications in computational linear algebra," Computer Science Department, UCLA, Technical Report CSD-950023, 1995.

[22] M. Baboulin, D. Becker, and J. Dongarra, "A parallel tiled solver for dense symmetric indefinite systems on multicore architectures," *Proceedings of IPDPS 2012 (to appear)*, 2012, LAPACK Working Note 261.

[23] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*. SIAM, 2002, second edition.

[24] U. of Tennessee, *PLASMA Users' Guide, Parallel Linear Algebra Software for Multicore Architectures, Version 2.3*, 2010.