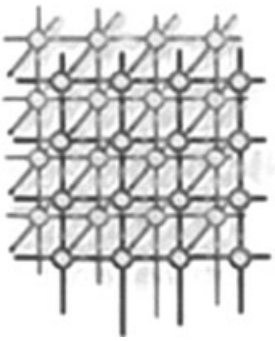
CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2010; **22**:15–44
Published online 11 August 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cpe.1467



Scheduling dense linear algebra operations on multicore processors

Jakub Kurzak[1, *, †], Hatem Ltaief[1], Jack Dongarra[1, 2, 3] and Rosa M. Badia[4]

[1]*Department of Electrical Engineering and Computer Science, University of Tennessee, TN, U.S.A.*
[2]*Computer Science and Mathematics Division, Oak Ridge National Laboratory, TN, U.S.A.*
[3]*School of Mathematics and School of Computer Science, University of Manchester, Manchester, U.K.*
[4]*Barcelona Supercomputing Center—Centro Nacional de Supercomputación, Barcelona, Spain*

SUMMARY

State-of-the-art dense linear algebra software, such as the LAPACK and ScaLAPACK libraries, suffers performance losses on multicore processors due to their inability to fully exploit thread-level parallelism. At the same time, the coarse–grain dataflow model gains popularity as a paradigm for programming multicore architectures. This work looks at implementing classic dense linear algebra workloads, the Cholesky factorization, the QR factorization and the LU factorization, using dynamic data-driven execution. Two emerging approaches to implementing coarse–grain dataflow are examined, the model of nested parallelism, represented by the Cilk framework, and the model of parallelism expressed through an arbitrary Direct Acyclic Graph, represented by the SMP Superscalar framework. Performance and coding effort are analyzed and compared against code manually parallelized at the thread level. Copyright © 2009 John Wiley & Sons, Ltd.

Received 6 February 2009; Revised 28 April 2009; Accepted 7 June 2009

KEY WORDS: task graph; scheduling; multicore; linear algebra; factorization; Cholesky; LU; QR; direct acyclic graph; dynamic scheduling; matrix factorization

*Correspondence to: Jakub Kurzak, Department of Electrical Engineering and Computer Science, University of Tennessee, TN, U.S.A.
†E-mail: kurzak@eecs.utk.edu

1. INTRODUCTION AND MOTIVATION

The current trend in the semiconductor industry to double the number of execution units on a single die is commonly referred to as *the multicore discontinuity*. This term reflects the fact that the existing software is inadequate for the new architectures and the existing code base will be incapable of delivering increased performance, possibly not even capable of sustaining the current performance.

This problem has already been observed with state-of-the-art dense linear algebra libraries, LAPACK [1] and ScaLAPACK [2], which deliver a small fraction of the peak performance on current multicore processors and multi-socket systems of multicore processors, mostly following *Symmetric Multi-Processor* (SMP) architecture.

The problem is twofold. Achieving good performance on emerging chip designs is a serious problem, calling for new algorithms and data structures. Reimplementing the existing code base using a new programming paradigm is another major challenge, specifically in the area of high performance scientific computing, where the level of required skills makes the programmers a scarce resource and millions of lines of code are in question.

The main contribution of this paper is a critical look at a representative set of emerging parallel programming frameworks for multicore processors through implementations of classic dense linear algebra workloads and, specifically, exposing the weakness of models relying on nested parallelism.

2. BACKGROUND

In large-scale scientific computing, targeting distributed memory systems, the recent push towards the PetaFlop barrier caused a renewed interest in *Partitioned Global Address Space* (PGAS) languages, such as *Co-Array Fortran* (CAF) [3], *Unified Parallel C* (UPC) [4] or Titanium [5], as well as the emergence of new languages, such as Chapel (Cray) [6], Fortress (Sun) [7] and X-10 (IBM) [8], sponsored through the DARPA's *High Productivity Computing Systems* (HPCS) program.

In more mainstream, server and desktop computing, targeting mainly shared memory systems, the well-known *dataflow model* is rapidly gaining popularity, where the computation is viewed as a *Direct Acyclic Graph* (DAG), with nodes representing computational tasks and edges representing data dependencies among them. The coarse–grain dataflow model is the main principle behind emerging multicore programming environments such as $Cilk/Cilk++$ [9], Intel® *Threading Building Blocks* (TBB) [10,11], Tasking in OpenMP 3.0 [12–15] and *SMP Superscalar* (SMPSs) [16].

All these frameworks rely on a very small set of extensions to common imperative programming languages such as C/C++ and Fortran and involve a relatively simple compilation stage and potentially much more complex runtime system.

The following sections provide a brief overview of these frameworks, as well as an overview of a rudimentary scheduler implemented using POSIX threads, which will serve as a baseline for performance comparisons.

Since tasking facilities available in Threading Building Blocks and OpenMP 3.0 closely resemble the ones provided by Cilk, Cilk is chosen as a representative framework for all three (also due to the reason that it is available in open source).

DOI: 10.1002/cpe

2.1. Cilk

Cilk was developed at the MIT Laboratory for Computer Science starting in 1994 [9]. Cilk is an extension of the C language with a handful of keywords (*cilk, spawn, sync, inlet, abort*) aimed at providing general-purpose programming language designed for multithreaded parallel programming. When the Cilk keywords are removed from Cilk source code, the result is a valid C program, called the *serial elision* (or C elision) of the full Cilk program. The Cilk environment employs a source-to-source compiler, which compiles Cilk code to C code, a standard C compiler, and a runtime system linked with the object code to provide an executable.

The main principle of Cilk is that the programmer is responsible for exposing parallelism by identifying functions free of side effects (e.g. access to global variables causing race conditions), which can be treated as independent tasks and executed in parallel. Such functions are annotated with the *cilk* keyword and invoked with the *spawn* keyword. The *sync* keyword is used to indicate that execution of the current procedure cannot proceed until all previously spawned procedures have completed and returned their results to the parent.

Distribution of work to multiple processors is handled by the runtime system. Cilk scheduler uses the policy called *work-stealing* to schedule execution of tasks to multiple processors. At runtime, each processor fetches tasks from the top of its own stack—in *First In First Out* (FIFO) order. However, when a processor runs out of tasks, it picks another processor at random and 'steals' tasks from the bottom of its stack—in *Last In First Out* (LIFO) order. This way the task graph is consumed in a *depth-first* order, until a processor runs out of tasks, in which case it steals tasks from other processors in a *breadth-first* order.

Cilk also provides the mechanism of locks. The use of locks can, however, easily lead to a deadlock. 'Even if the user can guarantee that his program is deadlock free, Cilk may still deadlock on the user's code because of some additional scheduling constraints imposed by Cilk's scheduler' [17]. In particular locks cannot be used to enforce parent–child dependencies between tasks.

Cilk is very well suited for expressing algorithms that easily render themselves to recursive formulation, e.g. *divide-and-conquer* algorithms. Since stack is the main structure for controlling parallelism, the model allows for straightforward implementations on shared memory multiprocessor systems (e.g. multicore/SMP systems). The simplicity of the model provides for the execution of parallel code with virtually no overhead from scheduling.

2.2. OpenMP

OpenMP was born in the 1990 s to bring a standard to the different directive languages defined by several vendors. It has been recently extended (version 3.0) to provide construct similar to those offered by Cilk. The new OpenMP directives allow the programmer to identify units of independent work (tasks), leaving the scheduling decisions to the runtime system. The main difference between Cilk and OpenMP 3.0 is that the latter can combine both types of parallelism, worksharing and tasks.

2.3. Intel® threading building blocks

Intel® Threading Building Blocks is a runtime-based parallel programming model for C++. Similar to OpenMPI 3.0 and Cilk, it is a runtime-based system, which emphasizes data parallelism, through

DOI: 10.1002/cpe

constructs like *parallel for*, but also provides constructs similar to Cilk for expressing nested parallelism.

2.4. SMPSs

SMPSs [16] is a parallel programming framework developed at the Barcelona Supercomputer Center (Centro Nacional de Supercomputación), part of the STAR Superscalar family, which also includes Grid Supercalar and Cell Superscalar [18,19]. While Grid Superscalar and Cell Superscalar address parallel software development for Grid environments and the Cell processor, respectively, SMP Superscalar is aimed at 'standard' (x86 and like) multicore processors and SMP systems.

The principles of SMP Superscalar are similar to the ones of Cilk. Similar to Cilk, the programmer is responsible for identifying parallel tasks, which have to be side effect-free (atomic) functions. Additionally, the programmer needs to specify the directionality of each parameter (input, output, inout). If the size of a parameter is missing in the C declaration (e.g. the parameter is passed by pointer), the programmer also needs to specify the size of the memory region affected by the function. Unlike Cilk, however, the programmer is not responsible for exposing the structure of the task graph. The task graph is built automatically, based on the information of task parameters and their directionality.

Similar to Cilk, the programming environment consists of a source-to-source compiler and a supporting runtime library. The compiler translates C code with pragma annotations to standard C99 code with calls to the supporting runtime library and compiles it using the platform native compiler.

At runtime the main thread creates worker threads, as many as necessary to fully utilize the system, and starts constructing the task graph (populating its ready list). Each worker thread maintains its own ready list and populates it while executing tasks. A thread consumes tasks from its own ready list in LIFO order. If that list is empty, the thread consumes tasks from the main ready list in FIFO order, and if that list is empty, the thread steals tasks from the ready lists of other threads in FIFO order.

The SMPSs scheduler attempts to exploit locality by scheduling dependent tasks to the same thread, such that output data is reused immediately. In addition, in order to reduce dependencies, SMPSs runtime is capable of renaming data, leaving only the true dependencies, which is the same technique used by superscalar processors [20] and optimizing compilers [21].

The main difference between Cilk and SMPSs is that, while the former allows mainly for expression of nested parallelism, the latter handles computation expressed as an arbitrary DAG. In addition, while Cilk requires the programmer to create the DAG by means of the spawn keyword, SMPSs creates the DAG automatically. Construction of the DAG does, however, introduce overhead, which is virtually inexistent in the Cilk environment.

2.5. Static pipeline

The *static pipeline* scheduling presented here was originally implemented for dense matrix factorizations on the CELL processor [22,23]. This technique is extremely simple and yet provides good locality of reference and load balance for regular computation, like dense matrix operations.

DOI: 10.1002/cpe

In this approach each task is uniquely identified by the $\{m, n, k\}$ triple, which determines the type of operation and the location of tiles operated upon. Each core traverses its task space by applying a simple formula to the $\{m, n, k\}$ triple, which takes into account the *id* of the core and the total number of cores in the system.

Task dependencies are tracked by a global progress table, where one element describes the progress of computation for one tile of the input matrix. Each core looks up the table before executing each task to check for dependencies and stalls if dependencies are not satisfied. Each core updates the progress table after completion of each task. Access to the table does not require mutual exclusion (using, e.g. mutexes). The table is declared as *volatile*. Update is implemented by writing to an element. Dependency stall is implemented by *busy-waiting* on an element.

The use of a global progress table is a potential scalability bottleneck. It does not pose a problem, however, on small-scale multicore/SMP systems for small to medium matrix sizes. Many alternatives are possible. (Replicated progress tables were used on the CELL processor [22,23].)

As further discussed in Sections 4.3 and 5.3, this technique allows for pipelined execution of factorizations steps, which provides similar benefits to dynamic scheduling, namely, execution of the inefficient Level 2 BLAS operations in parallel with the efficient Level 3 BLAS operations.

The main disadvantage of the technique is potentially suboptimal scheduling, i.e. stalling in situations where work is available. Another obvious weakness of the static schedule is that it cannot accommodate dynamic operations, e.g. *divide-and-conquer* algorithms.

3. RELATED WORK

Dynamic data-driven scheduling is an old concept and has been applied to dense linear operations for decades on various hardware systems. The earliest reference, that the authors are aware of, is the paper by Lord *et al*. [24]. A little later dynamic scheduling of LU and Cholesky factorizations was reported by Agarwal and Gustavson [25,26]. Throughout the years dynamic scheduling of dense linear algebra operations has been used in numerous vendor library implementations such as ESSL, MKL and ACML (numerous references are available on the Web). In the recent years, the authors of this work have been investigating these ideas within the framework *Parallel Linear Algebra Software for Multicore Architectures* (PLASMA) at the University of Tennessee [27–30].

The most important issue in performance optimization of orthogonal transformations is aggregation of transformations leading to efficient use of the memory system. The idea was first demonstrated by Dongarra *et al*. [31], later by Bischof and van Loan [32], and yet later by Schreiber and van Loan [33], resulting in the compact *WY* technique for accumulating Householder reflectors.

Elmroth and Gustavson [34–36] generalized this work to produce high-performance recursive QR factorization. In this work, the problem of reducing the amount of extra floating point operations was addressed by the introduction of *mini blocking/register blocking*, referred to as *inner blocking* in this paper. Serial implementation was presented as well as parallel implementation with dynamic scheduling of tasks on SMPs.

One of the early references discussing methods for updating matrix factorizations is the paper by Gill *et al.* [37]. Berry *et al.* successfully applied the idea of using orthogonal transformations to annihilate matrix elements by tiles, in order to achieve a highly parallel distributed memory implementation of matrix reduction to the block upper-Hessenberg form [38].

DOI: 10.1002/cpe

It is crucial to note that the technique of processing the matrix by square blocks only provides the performance in tandem with data organization by square blocks, a fact initially observed by Gustavson [39,40] and recently investigated in depth by Gustavson *et al.* [41]. The layout is referred to as *Square Block* (SB) format by Gustavson *et al.* and as *Block Data Layout* (BDL) in this work. The paper by Elmroth *et al.* [42] gives an excellent introduction to many of the important issues concerning deep memory hierarchies and the use of recursion and hybrid data structures and also contains a section on the QR factorization.

Seminal work leading to the *tile QR* algorithm presented here was done by Elmroth and Gustavson [34–36]. Gunter and van de Geijn presented an 'out-of-core' (out-of-memory) implementation [43], Buttari *et al.* an implementation for 'standard' (x86 and alike) multicore processors [29,30] and Kurzak *et al.* an implementation for the CELL processor [23]. The LU algorithm used here was originally devised by Quintana-Ortí and van de Geijn for 'out-of-core' (out-of-memory) execution [44].

4. CHOLESKY FACTORIZATION

The Cholesky factorization (or Cholesky decomposition) is mainly used for the numerical solution of linear equations $Ax=b$, where A is symmetric and positive definite. Such systems arise often in physics applications, where A is positive definite due to the nature of the modeled physical phenomenon. This happens frequently in numerical solutions of partial differential equations.

The Cholesky factorization of an $n\times n$ real symmetric positive-definite matrix A has the form

$$A=LL^{\mathrm{T}}$$

where L is an $n\times n$ real lower triangular matrix with positive diagonal elements. In LAPACK the double precision algorithm is implemented by the DPOTRF routine. A single step of the algorithm is implemented by a sequence of calls to the LAPACK and BLAS routines: DSYRK, DPOTF2, DGEMM, DTRSM. Owing to the symmetry, the matrix can be factorized either as an upper triangular matrix or as a lower triangular matrix. Here the lower triangular case is considered.

The algorithm can be expressed using either the top-looking version, the left-looking version of the right-looking version, the first being the most *lazy* algorithm (depth-first exploration of the task graph) and the last being the most *aggressive* algorithm (breadth-first exploration of the task graph). The left-looking variant is used here, with the exception of Cilk implementations, which favor the most aggressive right-looking variant.

Mathematically, the tile Cholesky algorithm is identical to the block Cholesky algorithm implemented in LAPACK. Operations on relatively large submatrices (blocks) are replaced with operations on relatively small submatrices (tiles). In addition, the call to LAPACK DPOTF2 routine is replaced with a call to the DPOTRF routine and multiple calls to the DTRSM routine. The algorithm relies on four basic operations implemented by four computational kernels (Figure 1).

DSYRK: The kernel applies updates to a diagonal (lower triangular) tile T of the input matrix, resulting from factorization of the tiles A to the left of it. The operation is a symmetric rank-k update.

DOI: 10.1002/cpe

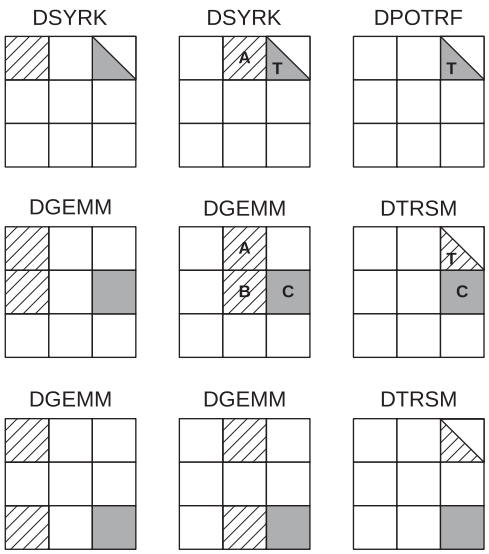


Figure 1. Tile operations in the tile Cholesky factorization. The sequence is left-to-right and top-down. Hatching indicates input data, shade of gray indicates in/out data.

```
FOR k = 0..TILES-1
  FOR n = 0..k-1
    A[k][k] ← DSYRK(A[k][n], A[k][k])
  A[k][k] ← DPOTRF(A[k][k])
  FOR m = k+1..TILES-1
    FOR n = 0..k-1
      A[m][k] ← DGEMM(A[k][n], A[m][n], A[m][k])
    A[m][k] ← DTRSM(A[k][k], A[m][k])
```

Figure 2. Pseudocode of the tile Cholesky factorization (left-looking version).

DPOTRF: The kernel performs the Cholesky factorization of a diagonal (lower triangular) tile T of the input matrix and overrides it with the final elements of the output matrix.

DGEMM: The operation applies updates to an off-diagonal tile C of the input matrix, resulting from factorization of the tiles to the left of it. The operation is a matrix multiplication.

DTRSM: The operation applies an update to an off-diagonal tile C of the input matrix, resulting from factorization of the diagonal tile above it and overrides it with the final elements of the output matrix. The operation is a triangular solve.

Figure 2 shows the generic pseudocode of the left-looking Cholesky factorization. Figure 3 shows the task graph of the tile Cholesky factorization of a 5×5 tiles matrix, produced automatically from the pseudocode using the DOT language. Although the code is as simple as four loops with three levels of nesting, the task graph is far from intuitive, even for a tiny size.

DOI: 10.1002/cpe

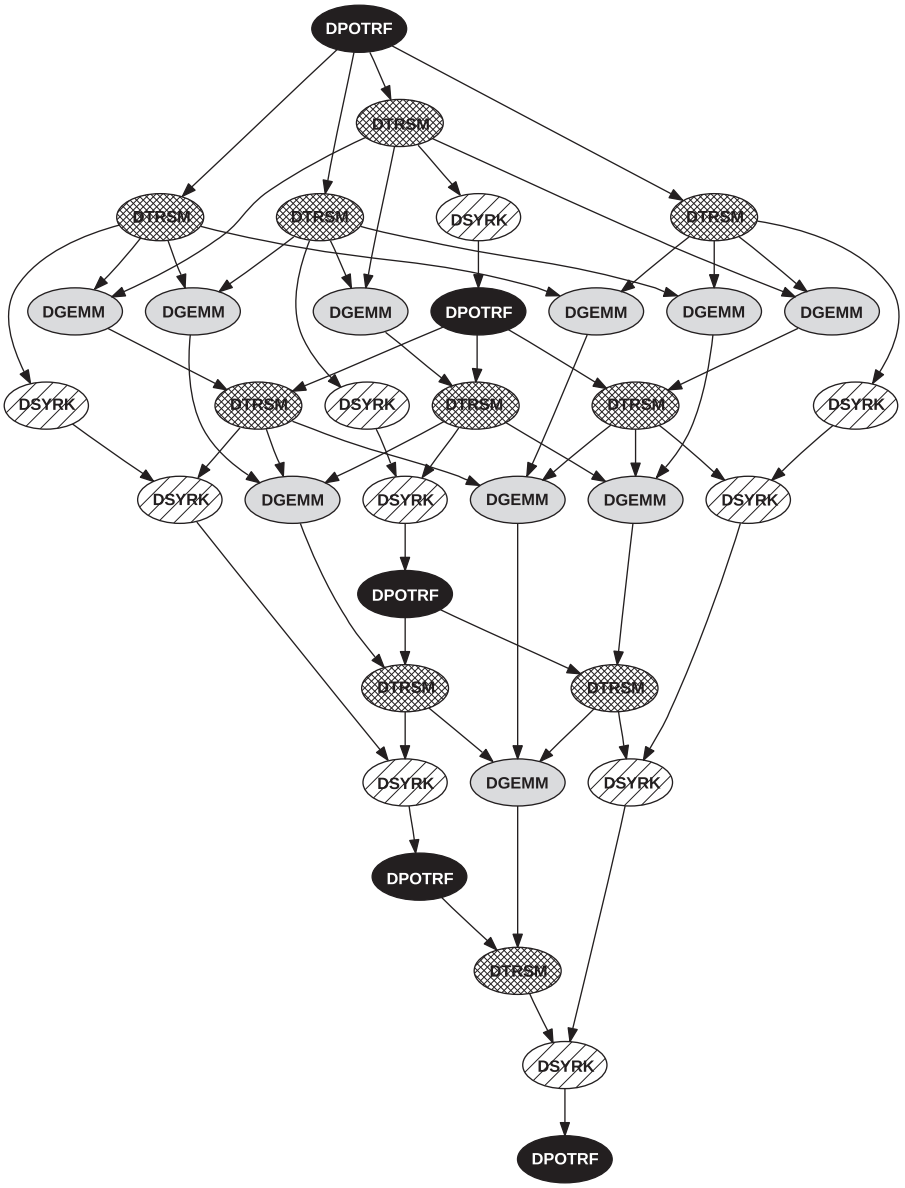


Figure 3. Task graph of the tile Cholesky factorization (5×5 tiles).

4.1. Cilk implementation

Figure 4 presents implementation of Cholesky factorization in Cilk. The basic building blocks are the functions performing the tile operations. *dsyrk()*, *dtrsm()* and *dgemm()* are implemented

DOI: 10.1002/cpe

```
cilk void dsyrk(double *A, double *T);
cilk void dpotrf(double *T);
cilk void dgemm(double *A, double *B, double *C);
cilk void dtrsm(double *T, double *C);


for (k = 0; k < TILES; k++) {

   spawn dpotrf(A[k][k]);
   sync;

   for (m = k+1; m < TILES; m++)
      spawn dtrsm(A[k][k], A[m][k]);
   sync;

   for (m = k+1; m < TILES; m++) {
      for (n = k+1; n < m; n++)
         spawn dgemm(A[k][n], A[m][n], A[m][k]);
      spawn dsyrk(A[k][n], A[k][k]);
   }
   sync;
}
```

Figure 4. Cilk implementation of the tile Cholesky factorization with 2D work assignment (right-looking version).

by calls to a single BLAS routine. *dpotrf()* is implemented by a call to the LAPACK DPOTRF routine. The functions are declared using the *cilk* keyword and then invoked using the *spawn* keyword.

The input matrix is stored using the format referred to in the literature as *Square Block* (SB) format or *Block Data Layout* (BDL). The latter name will be used here. In this arrangement, each function parameter is a pointer to a continuous block of memory, which greatly increases the cache performance and virtually eliminates cache conflicts between different operations.

For implementation in Cilk the right-looking variant was chosen, where factorization of each panel is followed by an update to all the remaining submatrix. The code on Figure 4 presents a version, referred here as *Cilk 2D*, where task scheduling is not constrained by data reuse considerations (there are no provisions for reuse of data between different tasks).

Each step of the factorization involves:

- factorization of the diagonal tile—spawning of the *dpotrf()* task followed by a *sync*,
- applying triangular solves to the tiles below the diagonal tile—spawning of the *dtrsm()* tasks in parallel followed by a *sync*,
- updating the tiles to the right of the panel—spawning of the *dsyrk()* and *dgemm()* tasks in parallel followed by a *sync*.

It is not possible to further improve parallelism by pipelining the steps of the factorization. Nevertheless, most of the work can proceed in parallel and only the *dpotrf()* task has to be executed sequentially.

Since the disregard for data reuse between tasks may adversely affect the algorithm's performance, it is necessary to consider an implementation facilitating data reuse. One possible approach

DOI: 10.1002/cpe

```
void dsyrk(double *A, double *T);
void dpotrf(double *T);
void dgemm(double *A, double *B, double *C);
void dtrsm(double *T, double *C);

cilk void cholesky_panel(int k)
{
   int m;

   dpotrf(A[k][k]);

   for (m = k+1; m < TILES; m++)
      dtrsm(A[k][k], A[m][k]);
}

cilk void cholesky_update(int n, int k)
{
   int m;

   dsyrk(A[k][n], A[k][k]);

   for (m = n+1; m < TILES; m++)
      spawn dgemm(A[k][n], A[m][n], A[m][k]);

   if (n == k+1)
      spawn cholesky_panel(k+1);

}


spawn cholesky_panel(0);
sync;

for (k = 0; k < TILES; k++) {
   for (n = k+1; n < TILES; n++)
      spawn cholesky_update(n, k);
   sync;
}
```

Figure 5. Cilk implementation of the tile Cholesky factorization with 1D work assignment (right-looking version).

is processing of the tiles of the input matrix by columns. In this case, however, work is being dispatched in relatively big batches and load imbalance in each step of the factorization will affect the performance. A traditional remedy to this problem is the technique of *lookahead*, where update of step N is applied in parallel with panel factorization of step $N+1$. Figure 5 shows such implementation, referred here as *Cilk 1D*.

First, panel 0 is factorized, followed by a *sync*. Then updates to all the remaining columns are issued in parallel. Immediately after updating the first column, next panel factorization is spawned. The code synchronizes at each step, but panels are always overlapped with updates. This approach implements one-level lookahead (lookahead of depth one). Implementing more levels of lookahead would further complicate the code, while giving little hope for significant performance improvement.

DOI: 10.1002/cpe

```
#pragma css task input(A[NB][NB]) inout(T[NB][NB])
void dsyrk(double *A, double *T);

#pragma css task inout(T[NB][NB])
void dpotrf(double *T);

#pragma css task input(A[NB][NB], B[NB][NB]) inout(C[NB][NB])
void dgemm(double *A, double *B, double *C);

#pragma css task input(T[NB][NB]) inout(B[NB][NB])
void dtrsm(double *T, double *C);


#pragma css start
for (k = 0; k < TILES; k++) {

   for (n = 0; n < k; n++)
      dsyrk(A[k][n], A[k][k]);
   dpotrf(A[k][k]);

   for (m = k+1; m < TILES; m++) {
      for (n = 0; n < k; n++)
         dgemm(A[k][n], A[m][n], A[m][k]);
      dtrsm(A[k][k], A[m][k]);
   )
)
#pragma css finish
```

Figure 6. SMPSs implementation of the tile Cholesky factorization (left-looking version).

4.2. SMPSs implementation

Figure 6 shows implementation using SMPSs. The functions implementing parallel tasks are designated with *#pragma ccs task* annotations defining directionality of the parameters (input, output, inout). The parallel section of the code is designated with *#pragma ccs start* and *#pragma ccs finish* annotations. Inside the parallel section the algorithm is implemented using the canonical representation of four loops with three levels of nesting, which closely matches the pseudocode definition of Figure 2.

The SMPSs runtime system schedules tasks based on dependencies and attempts to maximize data reuse by following the parent–child links in the task graph when possible.

4.3. Static pipeline implementation

As already mentioned in Section 2.5 the *static pipeline* implementation is a hand-written code using POSIX threads and primitive synchronization mechanisms (*volatile* progress table and busy-waiting). Figure 7 shows the implementation.

The code implements the left-looking version of the factorization, where work is distributed by rows of tiles and steps of the factorization are pipelined. The first core that runs out of work in step N proceeds to factorization of the panel in step $N+1$, following cores proceed to update in step $N+1$, then to panel in step $N+2$ and so on (Figure 8).

DOI: 10.1002/cpe

```
void dsyrk(double *A, double *T);
void dpotrf(double *T);
void dgemm(double *A, double *B, double *C);
void dtrsm(double *T, double *C);


k = 0; m = my_core_id;
while (m >= TILES) {
   k++; m = m-TILES+k;
} n = 0;

while (k < TILES && m < TILES) {
   next_n = n; next_m = m; next_k = k;

   next_n++;
   if (next_n > next_k) {
      next_m += cores_num;
      while (next_m >= TILES && next_k < TILES) {
         next_k++; next_m = next_m-TILES+next_k;
      } next_n = 0;
   }

   if (m == k) {
      if (n == k) {
         dpotrf(A[k][k]);
         core_progress[k][k] = 1;
      }
      else {
         while(core_progress[k][n] != 1);
         dsyrk(A[k][n], A[k][k]);
      }
   }
   else {
      if (n == k) {
         while(core_progress[k][k] != 1);
         dtrsm(A[k][k], A[m][k]);
         core_progress[m][k] = 1;
      }
      else {
         while(core_progress[k][n] != 1);
         while(core_progress[m][n] != 1);
         dgemm(A[k][n], A[m][n], A[m][k]);
      }
   }
   n = next_n; m = next_m; k = next_k;
}
```

Figure 7. Static pipeline implementation of the tile Cholesky factorization (left-looking version).

The code can be viewed as a parallel implementation of the Cholesky factorization with one-dimensional partitioning of work and lookahead, where lookahead of varying depth is implemented by processors that run out of work.

5. QR FACTORIZATION

The QR factorization (or QR decomposition) offers a numerically stable way of solving underdetermined and overdetermined systems of linear equations (least-squares problems) and is also the basis for the *QR algorithm* for solving the eigenvalue problem.

DOI: 10.1002/cpe

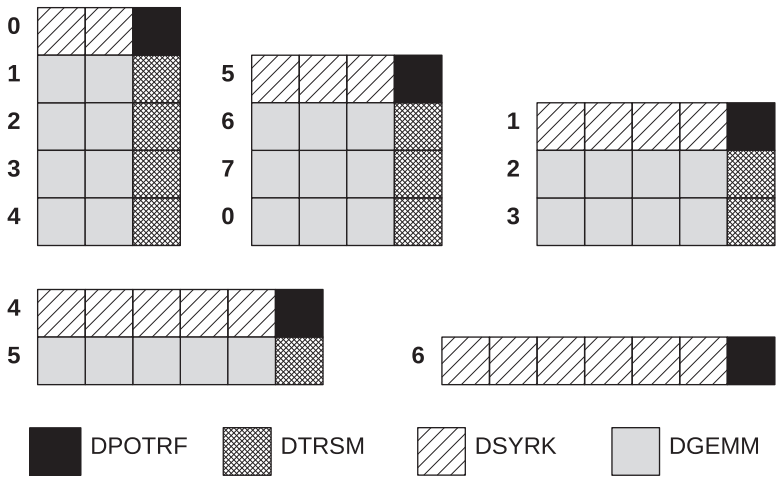


Figure 8. Work assignment in the static pipeline implementation of the tile Cholesky factorization. Consecutive steps of the factorization are shown (the sequence is left-to-right and top-down). Numbers signify the cores responsible for processing corresponding rows of the matrix.

The QR factorization of an $m \times n$ real matrix A has the form

$$A = QR$$

where Q is an $m \times m$ real orthogonal matrix and R is an $m \times n$ real upper triangular matrix. The traditional algorithm for QR factorization applies a series of elementary Householder matrices of the general form

$$H = I - \tau v v^{\mathrm{T}}$$

where v is a column reflector and τ is a scaling factor. In the block form of the algorithm a product of nb elementary Householder matrices is represented in the form

$$H_1 H_2 \ldots H_{nb} = I - VTV^{\mathrm{T}}$$

where V is an $N \times nb$ real matrix those columns are the individual vectors v, and T is an $nb \times nb$ real upper triangular matrix [32,33]. In LAPACK the double precision algorithm is implemented by the DGEQRF routine.

Here a derivative of the block algorithm is used called the *tile QR* factorization. The ideas behind the tile QR factorization are very well known. The tile QR factorization was initially developed to produce a high-performance 'out-of-memory' implementation (typically referred to as 'out-of-core') [43] and, more recently, to produce high-performance implementation on 'standard' (x86 and alike) multicore processors [29,30] and on the CELL processor [23].

The algorithm is based on the idea of annihilating matrix elements by square tiles instead of rectangular panels (block columns). The algorithm produces the same R factor as the classic algorithm, e.g. the implementation in the LAPACK library (elements may differ in sign). However,

DOI: 10.1002/cpe

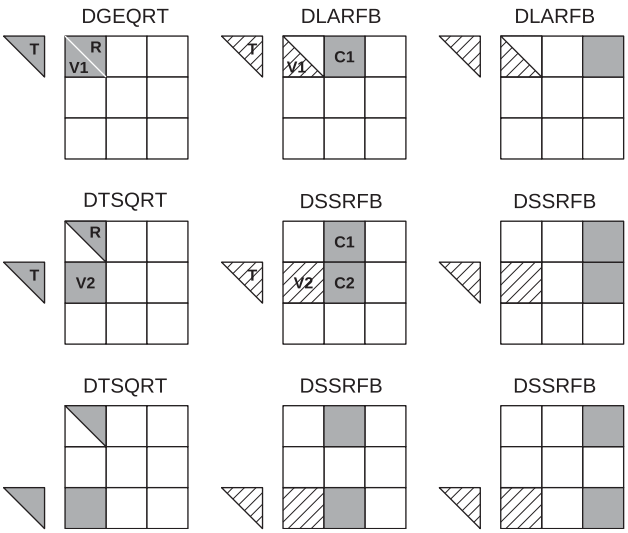


Figure 9. Tile operations in the tile QR factorization. The sequence is left-to-right and top-down. Hatching indicates input data, shade of gray indicates in/out data.

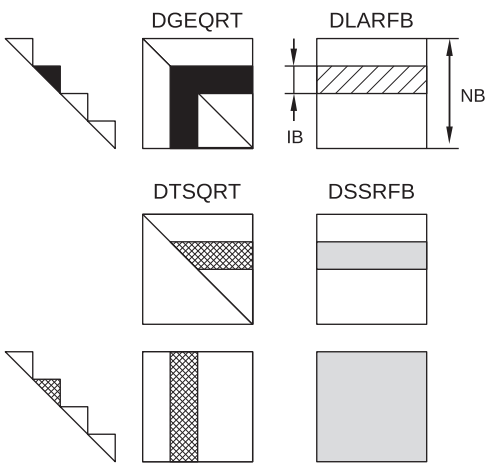


Figure 10. Inner blocking in the tile QR factorization.

a different set of Householder reflectors is produced and a different procedure is required to build the Q matrix. Whether the Q matrix is actually needed depends on the application. The tile QR algorithm relies on four basic operations implemented by four computational kernels (Figure 9).

DGEQRT: The kernel performs the QR factorization of a diagonal tile of the input matrix and produces an upper triangular matrix R and a unit lower triangular matrix V containing the

DOI: 10.1002/cpe

Householder reflectors. The kernel also produces the upper triangular matrix T as defined by the compact *WY* technique for accumulating Householder reflectors [32,33]. The R factor overrides the upper triangular portion of the input and the reflectors override the lower triangular portion of the input. The T matrix is stored separately.

DTSQRT: The kernel performs the QR factorization of a matrix built by coupling the R factor, produced by DGEQRT or a previous call to DTSQRT, with a tile below the diagonal tile. The kernel produces an updated R factor, a square matrix V containing the Householder reflectors and the matrix T resulting from accumulating the reflectors V. The new R factor overrides the old R factor. The block of reflectors overrides the square tile of the input matrix. The T matrix is stored separately.

DLARFB: The kernel applies the reflectors calculated by DGEQRT to a tile to the right of the diagonal tile, using the reflectors V along with the matrix T.

DSSRFB: The kernel applies the reflectors calculated by DTSQRT to two tiles to the right of the tiles factorized by DTSQRT, using the reflectors V and the matrix T produced by DTSQRT.

Naive implementation, where the full T matrix is built, results in 25% more floating point operations than the standard algorithm. In order to minimize this overhead, the idea of *inner-blocking* is used, where the T matrix has sparse (block-diagonal) structure (Figure 10) [34–36].

Figure 11 shows the pseudocode of the tile QR factorization. Figure 12 shows the task graph of the tile QR factorization for a matrix of 5×5 tiles. Orders of magnitude larger matrices are used in practice. This example only serves the purpose of showing the complexity of the task graph, which is noticeably higher than that of the Cholesky factorization.

5.1. Cilk implementation

The task graph of the tile QR factorization has a much denser net of dependencies than the Cholesky factorization. Unlike for Cholesky the tasks factorizing the panel are not independent and have to be serialized and the tasks applying the update have to follow the same order. The order can be arbitrary. Here top-down order is used.

Figure 13 shows the first Cilk implementation, referred to as *Cilk 2D*, which already requires the use of lookahead to achieve performance. The basic building blocks are the functions performing the tile operations. Unlike for Cholesky, none of them is a simple call to BLAS or LAPACK. Owing

```
FOR k = 0..TILES-1
  A[k][k], T[k][k] ← DGRQRT(A[k][k])
  FOR m = k+1..TILES-1
    A[k][k], A[m][k], T[m][k] ← DTSQRT(A[k][k], A[m][k], T[m][k])
  FOR n = k+1..TILES-1
    A[k][n] ← DLARFB(A[k][k], T[k][k], A[k][n])
    FOR m = k+1..TILES-1
      A[k][n], A[m][n] ← DSSRFB(A[m][k], T[m][k], A[k][n], A[m][n])
```

Figure 11. Pseudocode of the tile QR factorization.

DOI: 10.1002/cpe

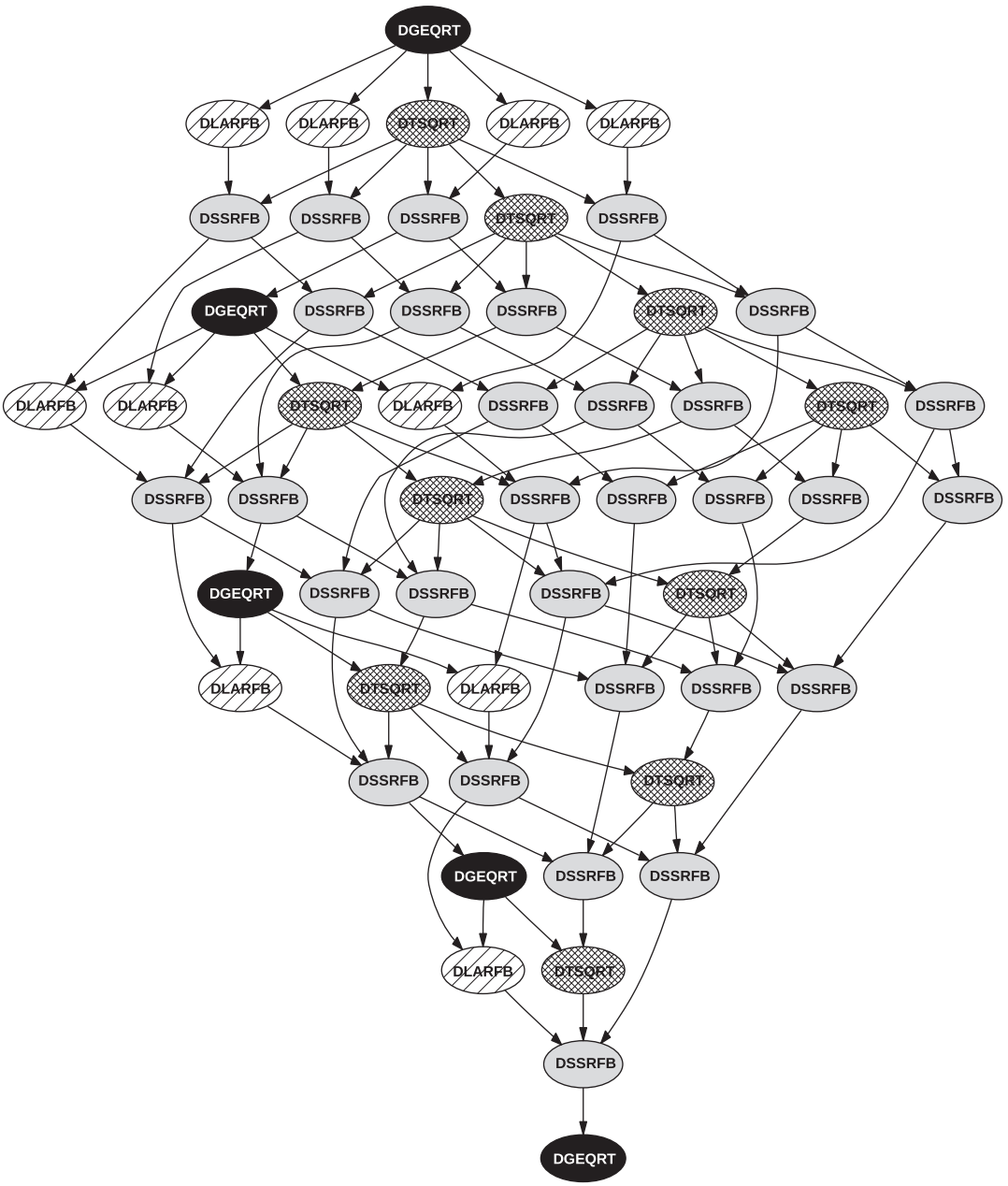


Figure 12. Task graph of the tile QR factorization (matrix of size 5×5 tiles).

DOI: 10.1002/cpe

```
cilk void dgeqrt(double *RV1, double *T);
cilk void dtsqrt(double *R, double *V2, double *T);
cilk void dlarfb(double *V1, double *T, double *C1);
void dssrfb(double *V2, double *T, double *C1, double *C2);

cilk void dssrfb_(int m, int n, int k)
{
   dssrfb(A[m][k], T[m][k], A[k][n], A[m][n]);

   if (m == TILES-1 && n == k+1 && k+1 < TILES)
      spawn dgeqrt(A[k+1][k+1], T[k+1][k+1]);

   if (n == k+1 && m+1 < TILES)
      spawn dtsqrt(A[k][k], A[m+1][k], T[m+1][k]);
}


spawn dgeqrt(A[0][0], T[0][0]);
sync;

for (k = 0; k < TILES; k++) {

   for (n = k+1; n < TILES; n++)
      spawn dlarfb(A[k][k], T[k][k], A[k][n]);

   if (k+1 < TILES)
      spawn dtsqrt(A[k][k], A[k+1][k], T[k+1][k]);
   sync;

   for (m = k+1; m < TILES; m++) {
      for (n = k+1; n < TILES; n++)
         spawn dssrfb_(m, n, k);
      sync;
   }
}
```

Figure 13. Cilk implementation of the tile QR factorization with 2D work assignment and lookahead.

to the use of inner-blocking the kernels consist of loop nests containing a number of BLAS and LAPACK calls (currently coded in FORTRAN 77).

The factorization proceeds in the following steps:

- Initially the first diagonal tile is factorized—spawning of the *dgeqrt()* task followed by a *sync*. Then the main loop follows with the remaining steps.
- Tiles to the right of the diagonal tile are updated in parallel with factorization of the tile immediately below the diagonal tile—spawning of the *dlarfb()* tasks and the *dtsqrt()* task followed by a *sync*.
- Updates are applied to the tiles right from the panel—spawning of the *dssrfb()* tasks by rows of tiles (*sync* following each row). The last *dssrfb()* task in a row spawns the *dtsqrt()* task in the next row. The last *dssrfb()* task in the last row spawns the *dgeqrt()* task in the next step of the factorization.

DOI: 10.1002/cpe

```
void dgeqrt(double *RV1, double *T);
void dtsqrt(double *R, double *V2, double *T);
void dlarfb(double *V1, double *T, double *C1);
void dssrfb(double *V2, double *T, double *C1, double *C2);

cilk void qr_panel(int k)
{
   int m;

   dgeqrt(A[k][k], T[k][k]);

   for (m = k+1; m < TILES; m++)
      dtsqrt(A[k][k], A[m][k], T[m][k]);
}

cilk void qr_update(int n, int k)
{
   int m;

   dlarfb(A[k][k], T[k][k], A[k][n]);

   for (m = k+1; m < TILES; m++)
      dssrfb(A[m][k], T[m][k], A[k][n], A[m][n]);

   if (n == k+1)
      spawn qr_panel(k+1);
}


spawn qr_panel(0);
sync;

for (k = 0; k < TILES; k++) {
   for (n = k+1; n < TILES; n++)
      spawn qr_update(n, k);
   sync;
}
```

Figure 14. Cilk implementation of the tile QR factorization with 1D work assignment and lookahead.

Although lookahead is used and factorization of the panel is, to some extent, overlapped with applying the update, tasks are being dispatched in smaller batches, which severely limits opportunities for scheduling.

The second possibility is to process the tiles of the input matrix by columns, the same as was done for Cholesky. Actually, it is much more natural to do it in the case of QR, where work within a column has to be serialized. Load imbalance comes into picture again and lookahead is the remedy. Figure 14 shows the implementation, referred to as *Cilk 1D*.

The implementation follows closely the *Cilk 1D* version of Cholesky. First, panel 0 is factorized, followed by a *sync*. Then updates to all the remaining columns are issued in parallel. Immediately after updating the first column, next panel factorization is spawned. The code synchronizes at each step, but panels are always overlapped with updates. This approach implements one-level lookahead (lookahead of depth one). Implementing more levels of lookahead would further complicate the code.

DOI: 10.1002/cpe

```
#pragma css task \
   inout(RV1[NB][NB]) output(T[NB][NB])
void dgeqrt(double *RV1, double *T);

#pragma css task \
   inout(R[NB][NB], V2[NB][NB]) output(T[NB][NB])
void dtsqrt(double *R, double *V2, double *T);

#pragma css task \
   input(V1[NB][NB], T[NB][NB]) inout(C1[NB][NB])
void dlarfb(double *V1, double *T, double *C1);

#pragma css task \
   input(V2[NB][NB], T[NB][NB]) inout(C1[NB][NB], C2[NB][NB])
void dssrfb(double *V2, double *T, double *C1, double *C2);


#pragma css start
for (k = 0; k < TILES; k++) {

   dgeqrt(A[k][k], T[k][k]);

   for (m = k+1; m < TILES; m++)
      dtsqrt(A[k][k], A[m][k], T[m][k]);

   for (n = k+1; n < TILES; n++) {
      dlarfb(A[k][k], T[k][k], A[k][n]);
      for (m = k+1; m < TILES; m++)
         dssrfb(A[m][k], T[m][k], A[k][n], A[m][n]);
   }
}
#pragma css finish
```

Figure 15. SMPSs implementation of the tile QR factorization.

5.2. SMPSs implementation

Figure 15 shows implementation using SMPSs, which follows closely the one for Cholesky. The functions implementing parallel tasks are designated with *#pragma ccs task* annotations defining directionality of the parameters (input, output, inout). The parallel section of the code is designated with *#pragma ccs start* and *#pragma ccs finish* annotations. Inside the parallel section the algorithm is implemented using the canonical representation of four loops with three levels of nesting, which closely matches the pseudocode definition of Figure 11.

The SMPSs runtime system schedules tasks based on dependencies and attempts to maximize data reuse by following the parent–child links in the task graph when possible.

There is a caveat here, however. *V1* is an input parameter of task *dlarfb()*. It is also an inout parameter of task *dtsqrt()*. However, *dlarfb()* only reads the lower triangular portion of the tile, while *dtsqrt()* only updates the upper triangular portion of the tile. Since in both cases the tile is passed to the functions by the pointer to the upper left corner of the tile, SMPSs sees a false dependency. As a result, the execution of the *dlarfb()* tasks in a given step will be stalled until all the *dtsqrt()* tasks complete, despite the fact that both types of tasks can be scheduled in parallel as soon as the *dgeqrt()* task completes. Figure 16 shows conceptually the change that needs to be done.

DOI: 10.1002/cpe

```
#pragma css task \
   inout(RV1[NB][NB]) output(T[NB][NB])
void dgeqrt(double *RV1, double *T);

#pragma css task \
   inout(R[◥], V2[NB][NB]) output(T[NB][NB])
void dtsqrt(double *R, double *V2, double *T);

#pragma css task \
   input(V1[◣], T[NB][NB]) inout(C1[NB][NB])
void dlarfb(double *V1, double *T, double *C1);

#pragma css task \
   input(V2[NB][NB], T[NB][NB]) inout(C1[NB][NB], C2[NB][NB])
void dssrfb(double *V2, double *T, double *C1, double *C2);


#pragma css start
for (k = 0; k < TILES; k++) {

   dgeqrt(A[k][k], T[k][k]);

   for (m = k+1; m < TILES; m++)
      dtsqrt(A[k][k]◥, A[m][k], T[m][k]);

   for (n = k+1; n < TILES; n++) {
      dlarfb(A[k][k]◣, T[k][k], A[k][n]);
      for (m = k+1; m < TILES; m++)
         dssrfb(A[m][k], T[m][k], A[k][n], A[m][n]);
   }
}
#pragma css finish
```

Figure 16. SMPSs implementation of the tile QR factorization with improved dependency resolution for diagonal tiles.

Currently, SMPSs is not capable of recognizing accesses to triangular matrices. There are however multiple ways to enforce the correct behavior. The simplest method, in this case, is to drop dependency check on the *V1* parameter of the *dlarfb()* function by declaring it as *volatile**. Correct dependency will be enforced between the *dgeqrt()* task and the *dlarfb()* tasks through the *T* parameter. This implementation is further referred to as *SMPSs**.

5.3. Static pipeline implementation

The *static pipeline* implementation for QR is very close to the one for Cholesky. As already mentioned in Section 2.5 the *static pipeline* implementation is a hand-written code using POSIX threads and primitive synchronization mechanisms (*volatile* progress table and busy-waiting). Figure 17 shows the implementation.

The code implements the right-looking version of the factorization, where work is distributed by columns of tiles and steps of the factorization are pipelined. The first core that runs out of work in step N proceeds to factorization of the panel in step $N+1$, following cores proceed to update in step $N+1$, then to panel in step $N+2$ and so on (Figure 18).

DOI: 10.1002/cpe

```
void dgeqrt(double *RV1, double *T);
void dtsqrt(double *R, double *V2, double *T);
void dlarfb(double *V1, double *T, double *C1);
void dssrfb(double *V2, double *T, double *C1, double *C2);


k = 0; n = my_core_id;
while (n >= TILES) {
   k++; n = n-TILES+k;
} m = k;

while (k < TILES && n < TILES) {
   next_n = n; next_m = m; next_k = k;

   next_m++;
   if (next_m == TILES) {
      next_n += cores_num;
      while (next_n >= TILES && next_k < TILES) {
         next_k++; next_n = next_n-TILES+next_k;
      } next_m = next_k;
   }

   if (n == k) {
      if (m == k) {
         while(progress[k][k] != k-1);
         dgeqrt(A[k][k], T[k][k]);
         progress[k][k] = k;
      }
      else{
         while(progress[m][k] != k-1);
         dtsqrt(A[k][k], A[m][k], T[m][k]);
         progress[m][k] = k;
      }
   }
   else {
      if (m == k) {
         while(progress[k][k] != k);
         while(progress[k][n] != k-1);
         dlarfb(A[k][k], T[k][k], A[k][n]);
      }
      else {
         while(progress[m][k] != k);
         while(progress[m][n] != k-1);
         dssrfb(A[m][k], T[m][k], A[k][n], A[m][n]);
         progress[m][n] = k;
      }
   }
   n = next_n; m = next_m; k = next_k;
}
```

Figure 17. Static pipeline implementation of the tile QR factorization.

The code can be viewed as a parallel implementation of the tile QR factorization with one-dimensional partitioning of work and lookahead, where lookahead of varying depth is implemented by processors that run out of work.

DOI: 10.1002/cpe

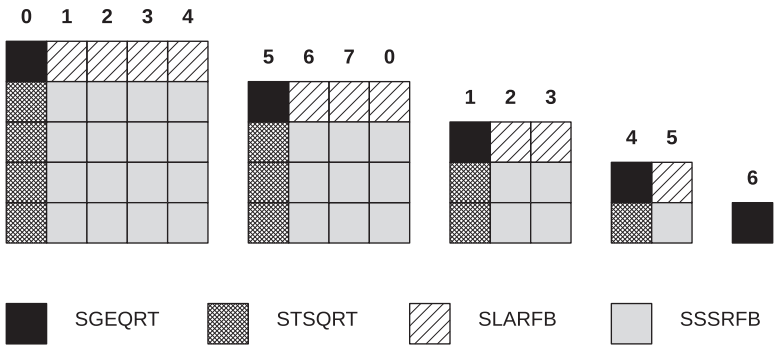


Figure 18. Work assignment in the static pipeline implementation of the tile QR factorization. Consecutive steps of the factorization are shown (the sequence is left-to-right). Numbers signify the cores responsible for processing corresponding rows of the matrix.

6. LU FACTORIZATION

The LU factorization (or LU decomposition) with partial row pivoting of an $m \times n$ real matrix A has the form

$$A = PLU$$

where L is an $m \times n$ real unit lower triangular matrix, U is an $n \times n$ real upper triangular matrix and P is a permutation matrix. In the block formulation of the algorithm, factorization of nb columns (the panel) is followed by the update of the remaining part of the matrix (the trailing submatrix) [45,46]. In LAPACK the double precision algorithm is implemented by the DGETRF routine. A single step of the algorithm is implemented by a sequence of calls to the following LAPACK and BLAS routines: DGETF2, DLASWP, DTRSM, DGEMM, where DGETF2 implements the panel factorization and the other routines implement the update.

Here a derivative of the block algorithm is used called the *tile LU* factorization. Similar to the tile QR algorithm, the tile LU factorization originated as an 'out-of-memory' ('out-of-core') algorithm [44] and was recently rediscovered for the multicore architectures [30]. No implementation on the CELL processor has been reported so far.

Again, the main idea here is the one of annihilating matrix elements by square tiles instead of rectangular panels. The algorithm produces different U and L factors than the block algorithm (e.g. the one implemented in the LAPACK library) and produces a different pivoting pattern, which is farther discussed in more detail. The tile LU algorithm relies on four basic operations implemented by four computational kernels (Figure 19).

DGETRF: The kernel performs the LU factorization of a diagonal tile of the input matrix and produces an upper triangular matrix U, a unit lower triangular matrix L and a vector of pivot indexes P. The U and L factors override the input and the pivot vector is stored separately.

DTSTRF: The kernel performs the LU factorization of a matrix build by coupling the U factor, produced by DGETRF or a previous call to DTSTRF, with a tile below the diagonal tile. The kernel produces an updated U factor and a square matrix L containing the coefficients corresponding

DOI: 10.1002/cpe

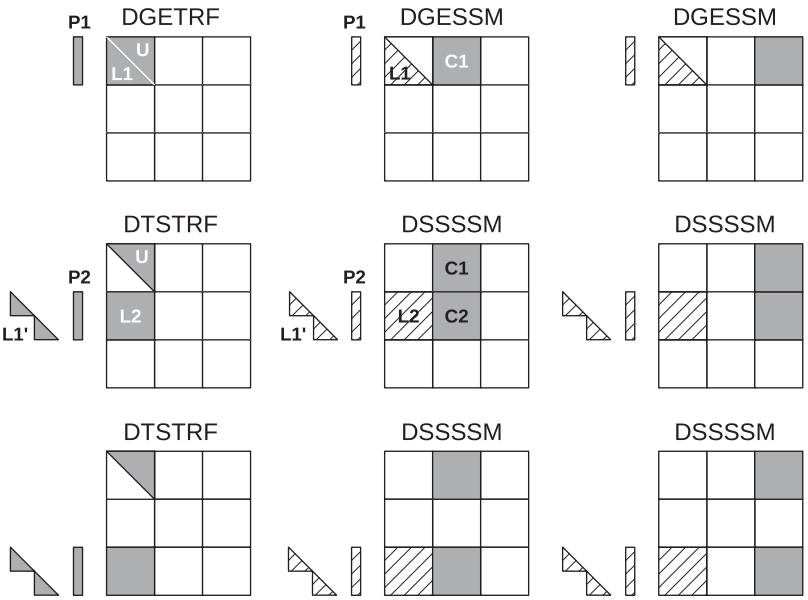


Figure 19. Tile operations in the tile LU factorization with inner blocking. The sequence is left-to-right and top-down. Hatching indicates input data, shade of gray indicates in/out data.

to the off-diagonal tile. The new U factor overrides the old U factor. The new L factor overrides corresponding off-diagonal tile. New pivot vector P is created and stored separately. Owing to pivoting, the lower triangular part of the diagonal tile is scrambled and also needs to be stored separately as L'.

DGESSM: The kernel applies the transformations produced by the DGETRF kernel to a tile to the right of the diagonal tile, using the L factor and the pivot vector P.

DSSSSM: The kernel applies the transformations produced by the DTSTRF kernel to the tiles to the right of the tiles factorized by DTSTRF, using the L' factor and the pivot vector P.

One topic that requires further explanation is the issue of pivoting. Since in the tile algorithm only two tiles of the panel are factorized at a time, pivoting only takes place within two tiles at a time, a scheme that could be described as *block-pairwise pivoting*. Clearly, such pivoting is not equivalent to the 'standard' *partial row pivoting* in the block algorithm (e.g. LAPACK). A different pivoting pattern is produced, and also, since pivoting is limited in scope, the procedure results in a less numerically stable algorithm. The numerical stability of the tile algorithm is not discussed here. As of today the authors are not aware of an exhaustive study of the topic.

As already mentioned earlier, due to pivoting, the lower triangular part of the diagonal block gets scrambled in consecutive steps of panel factorization. Each time this happens, the tiles to the right need to be updated, which introduces extra floating point operations, not accounted for in the standard formula for LU factorization. This is a similar situation to tile QR factorization, where the extra operations are caused by the accumulation of the Householder reflectors. For LU the

DOI: 10.1002/cpe

impact is yet bigger, resulting in 50% more operations for a naive implementation. The problem is remedied in the exact same way as for the tile QR factorization, by using the idea of inner blocking (Figure 19).

Another issue that comes into play is the concept of *LAPACK-style* pivoting versus *LINPACK-style* pivoting. In the former case, factorization of the panel is followed by row swaps both to the right of the panel and to the left of the panel. When using the factorization to solve the system, first permutations are applied to the entire right-hand side vector, and then straightforward lower triangular solve is applied to perform the forward substitution. In the latter case, factorization of the panel is followed by row swaps only to the right of the panel (only to the trailing submatrix). As a result, in the forward substitution phase of solving the system, applications of pivoting and Gauss transforms are interleaved.

The tile algorithm combines LAPACK pivoting within the panel, to achieve high performance for the kernels on a cache-based system, and LINPACK pivoting between the steps of the factorization, to facilitate flexible scheduling of tile operations. The combination of the two pivoting techniques is explained in great detail by Quintana-Ortí and van de Geijn [44].

6.1. Parallel implementation

The tile LU factorization is represented by a DAG of the exact same structure as the one for QR factorization. In other words, the tile LU factorization is identical, in terms of parallel scheduling, to the tile QR factorization. For that reason, the parallel implementations of the tile LU factorization are virtually identical to the parallel implementation of the tile QR factorization and all the facts presented in Section 5 hold here. In the codes, in Figures 11, 13–16, the DGEQRT operation is replaced by the DGETRF operation, DLARFB operation by DGESSM operation, DTSQRT by DTSTRF and DSSRFB by DSSSSM.

7. RESULTS AND DISCUSSION

Results were collected on a 2.4 GHz quad-socket quad-core (16 cores total) Intel Tigerton system running Linux kernel 2.6.18. Cilk and SMPSs codes were built using Cilk 5.4.6, SMPSs 2.0 and GCC 4.1.2. Static pipeline codes were built using ICC 10.1. Kernels coded in FORTRAN were compiled using IFORT 10.1. All codes were linked with MKL 10.0.1. Random input matrices were used (diagonally dominant for Cholesky factorization). Block Data Layout was used in all cases. Memory was allocated using huge TLB pages of size 2 MB.

It is the author's intention to use vendor compilers and libraries as much as possible. However, problems were encountered when compiling source distributions of Cilk and SMPSs using ICC, and GCC was used instead. This, however, has virtually no impact on the performance, since in all cases, the performance critical functions are performed by BLAS provided in MKL.

Figure 20 shows execution traces of all the implementations of Cholesky factorization. The figure shows a small run (9×9 tiles, 1080×1080 elements) on a small number of cores (four). The goal here is to clearly illustrate differences in scheduling by the different approaches.

DOI: 10.1002/cpe

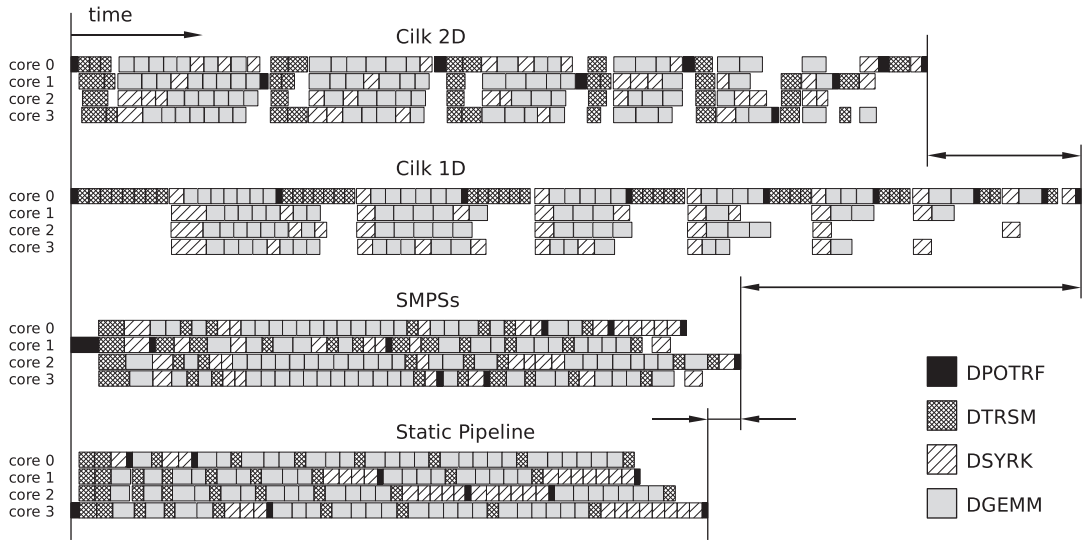


Figure 20. Execution traces of the tile Cholesky factorization in double precision on four cores of a 2.4 GHz Intel Tigerton system. Matrix size $N=1080$, tile size $nb=120$, total number of tasks $=140$.

The *Cilk 1D* implementation performs the worst. The 1D partitioning of work causes a disastrous load imbalance in each step of the factorization. Despite the lookahead, panel execution is very poorly overlapped with the update, in part due to the triangular shape of the updated submatrix and quickly diminishing amount of work in the update phase.

The *Cilk 2D* implementation performs much better by scheduling the *dtrsm()* operations in the panel in parallel. In addition, scheduling the *dsyrk()* and *dgemm()* tasks in the update phase without constraints minimizes the load imbalance. The only serial task, *dpotrf()*, does not cause disastrous performance losses.

Far better is the SMPSs implementation, where tasks are continuously scheduled without gaps until the very end of the factorization, where small stalls occur. Data reuse is clearly visible through clusters of *dsyrk()* tasks. Yet better is the *static pipeline* schedule, where no dependency stalls occur at all and data reuse is exploited to the fullest.

Figure 21 shows execution traces of all the implementations of QR factorization. The same as for Cholesky, the figure shows a small run (7×7 tiles, 1008×1008 elements) on a small number of cores (four). Once again, the goal here is to clearly illustrate the differences in scheduling by the different approaches. Traces for the tile LU factorization for a similar size problem are virtually identical to the QR traces and are not shown here. The following discussion applies equally to the tile QR and to the tile LU factorization.

The situation looks a bit different for the tile QR and LU factorizations compared with the tile Cholesky factorization. The fine-grain *Cilk 2D* implementation performs poorest, which is mostly due to the dispatch of work in small batches. Although the tasks of panel factorization (*dgeqrt()*, *dtsqrt()*) are overlapped with the tasks of the update (*dlarfb()*, *dssrfb()*), synchronization after each row, and related load imbalance, contribute a big number of gaps in the trace.

DOI: 10.1002/cpe

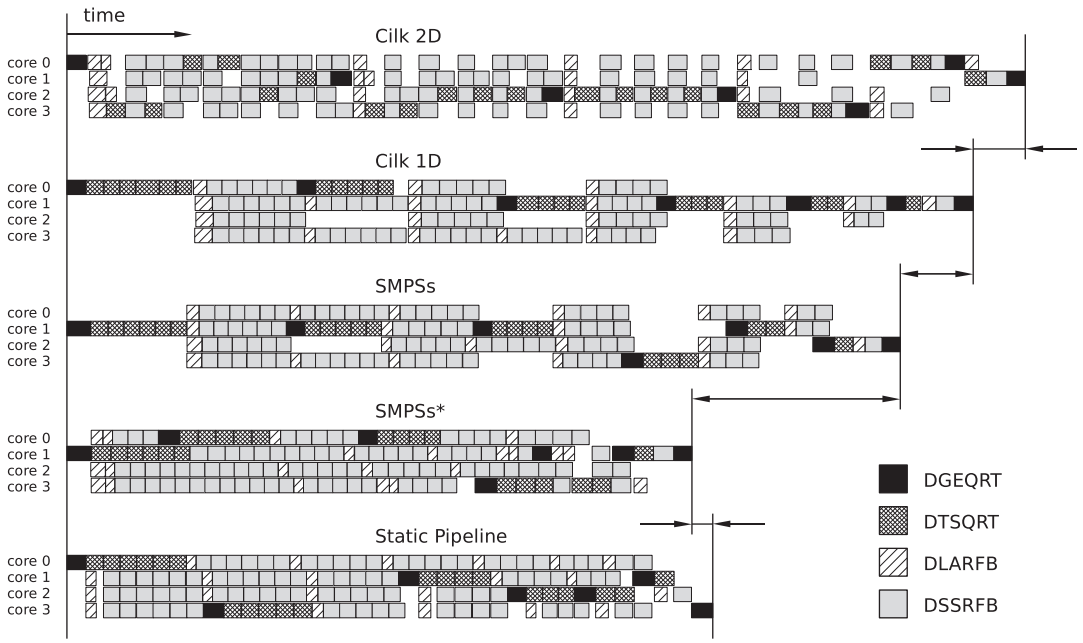


Figure 21. Execution traces of the tile QR factorization in double precision on four cores of a 2.4 GHz Intel Tigerton system. Matrix size N=1008, tile size nb=144, inner block size IB=48, total number of tasks =140.

The *Cilk 1D* version performs better. Although the number of gaps is still significant, mostly due to 1D partitioning and related load imbalance, overall this implementation loses less time due to dependency stalls.

Interestingly, the initial *SMPSs* implementation produces almost an identical schedule to the *Cilk 1D* version. One difference is the better schedule at the end of the factorization. The overall performance difference is small.

The *SMPSs** implementation delivers a big jump in the performance, due to dramatic improvement in the schedule. Here the *static pipeline* schedule is actually marginally worse than *SMPSs* due to a few more dependency stalls. More flexible scheduling of *SMPSs* provides for a better schedule at the end of the factorization. This advantage diminishes on larger number of cores, where the overheads of dynamic scheduling puts the performance of the *SMPSs* implementation slightly behind the one of the *static pipeline* implementation.

Figure 22 shows the performance for the Cholesky factorization, where Cilk implementations provide mediocre performance, SMPSs provides much better performance and *static pipeline* provides the performance clearly superior to other implementations.

Figure 23 shows the performance for the QR factorization. The situation is a little different here. The performance of Cilk implementations is still the poorest and the performance of the *static pipeline* is still superior. However, the performance of the initial *SMPSs* implementation is only marginally better than *Cilk 1D*, while the performance of the improved SMPSs* implementation is

DOI: 10.1002/cpe

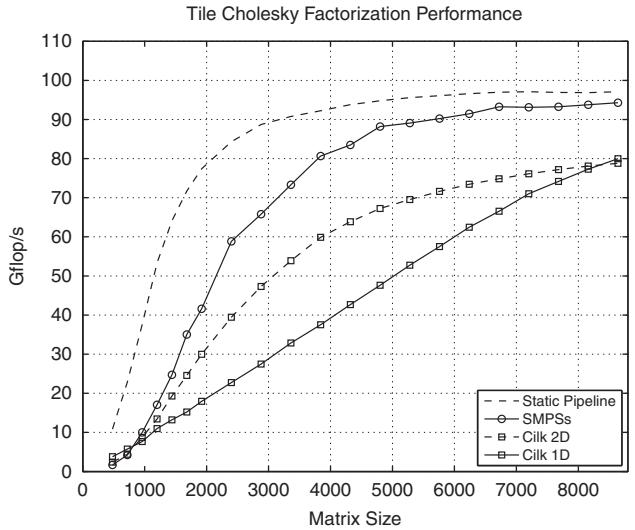


Figure 22. The performance of the tile Cholesky factorization in double precision on a 2.4 GHz quad-socket quad-core (16 cores total) Intel Tigerton system. Tile size nb=120.

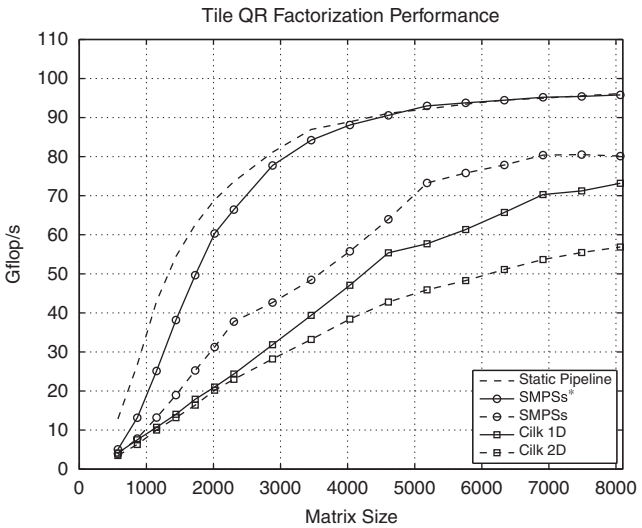


Figure 23. The performance of the tile QR factorization in double precision on a 2.4 GHz quad-socket quad-core (16 cores total) Intel Tigerton system. Tile size $nb = 144$, inner block size $IB = 48$.

DOI: 10.1002/cpe

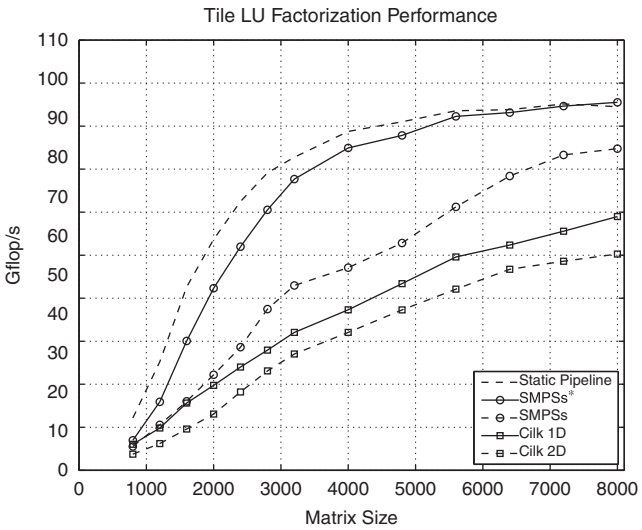


Figure 24. The performance of the tile LU factorization in double precision on a 2.4 GHz quad-socket quad-core (16 cores total) Intel Tigerton system. Tile size nb=200, inner block size IB=40.

only marginally worse than *static pipeline*. The same conclusions apply to the tile LU factorization (Figure 24).

Relatively better performance of SMPSs for the QR and LU factorizations versus the Cholesky factorization can be explained by the fact that the LU factorization is two times more expensive and the QR factorization is four times more expensive, in terms of floating point operations. This diminishes the impact of various overheads for smaller size problems.

8. CONCLUSIONS

In this work, the suitability of emerging multicore programming frameworks was analyzed for implementing modern formulations of classic dense linear algebra algorithms, the tile Cholesky, the tile QR and the tile LU factorizations. These workloads are represented by large task graphs with compute-intensive tasks interconnected with a very dense and complex net of dependencies.

For the workloads under investigation, the conducted experiments show clear advantage of the model, where automatic parallelization is based on construction of *arbitrary* DAGs. SMPSs provides much higher level of automation than Cilk and similar frameworks, requiring only minimal programmer's intervention and basically leaving the programmer oblivious to any aspects of parallelization. At the same time it delivers superior performance through more flexible scheduling of operations.

SMPSs still loses to hand-written code for very regular compute-intensive workloads investigated here. The gap is likely to decrease, however, with improved runtime implementations. Ultimately, it may have to be accepted as the price for automation.

DOI: 10.1002/cpe

Parallel programming based on the idea of representing the computation as a task graph and dynamic data-driven execution of tasks shows clear advantages for multicore processors and multi-socket shared-memory systems of such processors. One of the most interesting questions is the applicability of the model to large-scale distributed-memory systems.

REFERENCES

1. Anderson E, Bai Z, Bischof C, Blackford LS, Demmel JW, Dongarra JJ, Du Croz J, Greenbaum A, Hammarling S, McKenney A, Sorensen D. *LAPACK Users' Guide*. SIAM: Philadelphia, PA, 1992. Available at: http://www.netlib.org/lapack/lug/ [2 June 2009].
2. Blackford LS, Choi J, Cleary A, D'Azevedo E, Demmel J, Dhillon I, Dongarra JJ, Hammarling S, Henry G, Petitet A, Stanley K, Walker D, Whaley RC. *ScaLAPACK Users' Guide*. SIAM: Philadelphia, PA, 1997. Available at: http://www.netlib.org/scalapack/slug/ [2 June 2009].
3. Co-Array Fortran. Available at: http://www.co-array.org/ [2 June 2009].
4. The Berkeley Unified Parallel C (UPC) project. Available at: http://upc.lbl.gov/ [2 June 2009].
5. Titanium project home page. Available at: http://titanium.cs.berkeley.edu/ [2 June 2009].
6. Cray, Inc. *Chapel Language Specification 0.775*. Available at: http://chapel.cs.washington.edu/spec-0.775.pdf [2 June 2009].
7. Sun Microsystems Inc. *The Fortress Language Specification*, *Version 1.0*, 2008. Available at: http://research.sun.com/projects/plrg/Publications/fortress.1.0.pdf [2 June 2009].
8. Saraswat V, Nystrom N. *Report on the Experimental Language X10*, *Version 1.7*, 2008. Available at: http://dist.codehaus.org/x10/documentation/languagespec/x10-170.pdf [2 June 2009].
9. Blumofe RD, Joerg CF, Kuszmaul BC, Leiserson CE, Randall KH, Zhou Y. Cilk: An efficient multithreaded runtime system. *Principles and Practice of Parallel Programming*, *Proceedings of the Fifth ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, *PPOPP'95*. ACM: Santa Barbara, CA, 19–21 July 1995; 207–216. DOI: 10.1145/209936.209958.
10. Intel Threading Building Blocks. Available at: http://www.threadingbuildingblocks.org/ [2 June 2009].
11. Reinders J. *Intel Threading Building Blocks*: *Outfitting C++ for Multi-core Processor Parallelism*. O'Reilly Media, Inc., 2007. Available at: http://www.amazon.com/exec/obidos/ASIN/0596514808/ ISBN: 0596514808 [2 June 2009].
12. OpenMP Architecture Review Board. *OpenMP Application Program Interface*, *Version 3.0*, 2008. Available at: http://www.openmp.org/mp-documents/spec30.pdf [2 June 2009].
13. The community of OpenMP users, researchers, tool developers and providers. Available at: http://www.compunity.org/ [2 June 2009].
14. Ayguadé E, Copty N, Duran A, Hoeflinger J, Lin Y, Massaioli F, Su E, Unnikrishnan P, Zhang G. A proposal for task parallelism in OpenMP. *A Practical Programming Model for the Multi-Core Era*, *3rd International Workshop on OpenMP*, *IWOMP 2007* (*Lecture Notes in Computer Science*, vol. 4935), Beijing, China. Springer: Berlin, 3–7 June 2007; 1–12. DOI: 10.1007/978-3-540-69303-1_1.
15. Duran A, Perez JM, Ayguadé RM, Badia Labarta J. Extending the OpenMP tasking model to allow dependent tasks. *OpenMP in a New Era of Parallelism*, *4th International Workshop*, *IWOMP 2008* (*Lecture Notes in Computer Science*, vol. 5004), West Lafayette, IN. Springer: Berlin, 12–14 May 2008; 111–122. DOI: 10.1007/978-3-540-79561-2_10.
16. Barcelona Supercomputing Center. *SMP Superscalar* (*SMPSs*) *User's Manual*, *Version 2.0*, 2008. Available at: http://www.bsc.es/media/1002.pdf [2 June 2009].
17. Supercomputing Technologies Group. *Cilk 5.4.6 Reference Manual*, MIT Laboratory for Computer Science, 1998. Available at: http://supertech.csail.mit.edu/cilk/manual-5.4.6.pdf [2 June 2009].
18. Bellens P, Perez JM, Badia RM, Labarta J. CellSs: A programming model for the Cell BE architecture. *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*. Tampa, FL. ACM: New York, 11–17 November 2006; DOI: 10.1145/1188455.1188546.
19. Perez JM, Bellens P, Badia RM, Labarta J. CellSs: Making it easier to program the Cell Broadband Engine processor. *IBM Journal of Research and Development* 2007; **51**(5):593–604. DOI: 10.1147/rd.515.0593.
20. Smith JE, Sohi GS. The microarchitecture of superscalar processors. *Proceedings of the IEEE* 1995; **83**(12):1609–1624.
21. Kuck DJ, Kuhn RH, Padua DA, Leasure B, Wolfe M. Dependence graphs and compiler optimizations. *Proceedings of the 8th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. ACM: Williamsburg, VA, January 1981; 207–218. DOI: 10.1145/209936.209958.
22. Kurzak J, Buttari A, Dongarra JJ. Solving systems of linear equation on the CELL processor using Cholesky factorization. *IEEE Transactions on Parallel and Distributed Systems* 2008; **19**(19):1175–1186. DOI: TPDS.2007.70813.

23. Kurzak J, Dongarra JJ. QR factorization for the CELL processor. *Scientific Programming* 2009; **17**(1–2):31–42. DOI: 10.3233/SPR-2009-0268.
24. Lord RE, Kowalik JS, Kumar SP. Solving linear algebraic equations on an MIMD computer. *Journal of the ACM* 1983; **30**(1):103–117. DOI: 10.1145/322358.322366.
25. Agarwal RC, Gustavson FG. A parallel implementation of matrix multiplication and LU factorization on the IBM 3090. *Proceedings of the IFIP WG 2.5 Working Conference on Aspects of Computation on Asynchronous Parallel Processors*. North-Holland Publishing Company: Stanford, CA, 22–25 August 1988; 217–221. ISBN: 0444873104.
26. Agarwal RC, Gustavson FG. Vector and parallel algorithms for Cholesky factorization on IBM 3090. *Proceedings of the 1989 ACM/IEEE Conference on Supercomputing*, Reno, NV. ACM: New York, 13–17 November 1989; 225–233. DOI: 10.1145/76263.76287.
27. Kurzak J, Dongarra JJ. Implementing linear algebra routines on multi-core processors with pipelining and a look ahead. *Applied Parallel Computing*, *State of the Art in Scientific Computing*, *8th International Workshop*, *PARA 2006* (*Lecture Notes in Computer Science*, vol. 4699), Umeå, Sweden. Springer: Berlin, 18–21 June 2006; 147–156. DOI: 10.1007/978-3-540-75755-9_18.
28. Buttari A, Dongarra JJ, Husbands P, Kurzak J, Yelick K. Multithreading for synchronization tolerance in matrix factorization. *Scientific Discovery Through Advanced Computing*, *SciDAC 2007* (*Journal of Physics*: *Conference Series*, vol. 78:012028), Boston, MA. IOP Publishing: Bristol, U.K., 24–28 June 2007. DOI: 10.1088/1742-6596/78/1/012028.
29. Buttari A, Langou J, Kurzak J, Dongarra JJ. Parallel tiled QR factorization for multicore architectures. *Concurrency and Computation*: *Practice and Experience* 2008; **20**(13):1573–1590. DOI: 10.1002/cpe.1301.
30. Buttari A, Langou J, Kurzak J, Dongarra JJ. A class of parallel tiled linear algebra algorithms for multicore architectures. *Parallel Computing*: *Systems and Applications* 2009; **35**:38–53. DOI: 10.1016/j.parco.2008.10.002
31. Dongarra JJ, Kaufman L, Hammarling S. Squeezing the most out of eigenvalue solvers on high-performance computers. *Linear Algebra and its Applications* 1986; **77**:113–136.
32. Bischof C, van Loan C. The WY representation for products of Householder matrices. *Journal on Scientific and Statistical Computing* 1987; **8**:2–13.
33. Schreiber R, van Loan C. A storage-efficient WY representation for products of Householder transformations. *Journal on Scientific and Statistical Computing*1991; **10**:53–57.
34. Elmroth E, Gustavson FG. Applying recursion to serial and parallel QR factorization leads to better performance. *IBM Journal of Research and Development* 2000; **44**(4):605–624.
35. Elmroth E, Gustavson FG. High-performance library software for QR factorization. *Applied Parallel Computing*, *New Paradigms for HPC in Industry and Academia*, *5th International Workshop*, *PARA 2000* (*Lecture Notes in Computer Science*, vol. 1947), Bergen, Norway. Springer: Berlin, 18–20 2000; 53–63. DOI: 10.1007/3-540-70734-4_9.
36. Elmroth E, Gustavson FG. New serial and parallel recursive QR factorization algorithms for SMP systems. *Applied Parallel Computing*, *Large Scale Scientific and Industrial Problems*, *4th International Workshop*, *PARA'98* (*Lecture Notes in Computer Science*, vol. 1541), Umeå, Sweden. Springer: Berlin, 14–17 June 1998; 120–128. DOI: 10.1007/BFb0095328. Available at: http://dx.doi.org/10.1007/BFb0095328 [2 June 2009].
37. Gill PE, Golub GH, Murray WA, Saunders MA. Methods for modifying matrix factorizations. *Mathematics of Computation* 1974; **28**(126):505–535.
38. Berry MW, Dongarra JJ, Kim Y. LAPACK working note 68: A highly parallel algorithm for the reduction of a nonsymmetric matrix to block upper-Hessenberg form. *Technical Report UT-CS-94-221*, Computer Science Department, University of Tennessee, 1994. Available at: http://www.netlib.org/lapack/lawnspdf/lawn68.pdf [2 June 2009].
39. Gustavson FG. Recursion leads to automatic variable blocking for dense linear-algebra algorithms. *IBM Journal of Research and Development* 1997; **41**(6):737–756. DOI: 10.1147/rd.416.0737.
40. Gustavson FG. New generalized matrix data structures lead to a variety of high-performance algorithms. *Proceedings of the IFIP WG 2.5 Working Conference on Software Architectures for Scientific Computing Applications*. Kluwer Academic Publishers: Ottawa, Canada, 2–4 October 2000; 211–234. ISBN: 0792373391.
41. Gustavson FG, Gunnels JA, Sexton JC. Minimal data copy for dense linear algebra factorization. *Applied Parallel Computing*, *State of the Art in Scientific Computing*, *8th International Workshop*, *PARA 2006 PARA 2006*. (*Lecture Notes in Computer Science*, vol. 4699), Umeå, Sweden. Springer: Berlin, 18–21 June 2006; 540–549. DOI: 10.1007/978-3-540-75755-9_66.
42. Elmroth E, Gustavson FG, Jonsson I, Kågström B. Recursive blocked algorithms and hybrid data structures for dense matrix library software. *SIAM Review* 2004; **46**(1):3–45. DOI: 10.1137/S0036144503428693.
43. Gunter BC, van de Geijn RA. Parallel out-of-core computation and updating the QR factorization. *ACM Transactions on Mathematical Software* 2005; **31**(1):60–78. DOI: 10.1145/1055531.1055534.
44. Quintana-Ortí ES, van de Geijn RA. Updating an LU factorization with pivoting. *ACM Trans. Math. Softw* 2008; **35**(2):11. DOI: 10.1145/1377612.1377615.
45. Dongarra JJ, Duff IS, Sorensen DC, van der Vorst HA. *Numerical Linear Algebra for High-Performance Computers*. SIAM: Philadelphia, 1998. ISBN: 0898714281.
46. Demmel JW. *Applied Numerical Linear Algebra*. SIAM: Philadelphia, 1997. ISBN: 0898713897.