ABSTRACT

Recursion leads to automatic variable blocking for dense linear-algebra algorithms. The recursive way of programming algorithms eliminates using BLAS level 2 during the factorization steps. For this and other reasons recursion usually speeds up the algorithms. The Cholesky factorization algorithm for positive definite matrices and LU factorization for general matrices are formulated. Different storage data formats and recursive BLAS are explained in this paper. Performance graphs of packed and recursive Cholesky algorithms are presented.

1. INTRODUCTION

This work is a continuation of the work of Gustavson and Toledo described in [5; 9]. These papers describe the application of recursion to the numerical dense linear algebra algorithms. Recursion leads to automatic variable blocking for the dense linear-algebra algorithms. This leads to modifications of the LAPACK [1] algorithms. LAPACK’s level-2 version routines are transformed into level-3 codes by using recursion.
Fortran 90 allows recursion (see [8]). The programs are very concise and the recursion part is automatic as it is handled by the compiler. The intermediate subroutines obey the Fortran 90 standard too (see [3]).

Section 2 shows the recursive Cholesky factorization algorithm. Section 3 formulates the recursive algorithm of Gaussian elimination without pivoting and LU factorization with partial pivoting. Section 4 explains two recursive BLAS: RTRSM and RSYRK.

2. CHOLESKY FACTORIZATION

We would like to compute the solution to a system of linear equations $AX = B$, where $A$ is real symmetric or complex Hermitian and, in either case, positive definite matrix, $X$ and $B$ are rectangular matrices or vectors. The Cholesky decomposition can be used to factor $A$, $A = LL^T$ or $A = U^TU$, where $U$ is an upper triangular matrix and $L$ is a lower triangular ($L = U^T$). The factored form of $A$ is then used to solve the system of equations $AX = B$.

A recursive algorithm of Cholesky factorization is described in detail in [10; 5]. Here we give the final recursive algorithms for the lower triangular and upper triangular cases. We assume that $A$ is $n$ by $n$.

**Recursive Algorithm 2.1.** Cholesky recursive algorithm if lower triangular part of $A$ is given (cholesky):

- Do recursion
  - if $n > 1$ then
    - $L_{11} := \text{cholesky of } A_{11}$
    - $L_{21}L_{11}^T = A_{21} \rightarrow \text{RTRSM}$
    - $A_{22} := A_{22} - L_{21}L_{21}^T \rightarrow \text{RSYRK}$
    - $L_{22} := \text{cholesky of } A_{22}$
  - otherwise
    - $L := \sqrt{A}$
- End recursion

**Recursive Algorithm 2.2.** Cholesky recursive algorithm if upper triangular part of $A$ is given (cholesky):

- Do recursion
  - if $n > 1$ then
    - $U_{11} := \text{cholesky of } A_{11}$
    - $U_{11}^TU_{12} = A_{12} \rightarrow \text{RTRSM}$
    - $A_{22} := A_{22} - U_{11}^TU_{12} \rightarrow \text{RSYRK}$
    - $U_{22} := \text{cholesky of } A_{22}$
  - otherwise
    - $U := \sqrt{A}$
- End recursion

The matrices $A_{11}$, $A_{12}$, $A_{21}$, $A_{22}$, $L_{11}$, $L_{21}$, $L_{22}$, $U_{11}$, $U_{12}$ and $U_{22}$ are submatrices of $A$, $L$ and $U$ respectively.
The sizes of the submatrices are: for $A_{11}$, $L_{11}$ and $U_{11}$ is $h \times h$, for $A_{21}$ and $L_{21}$ is $(n-h) \times h$, for $A_{12}$ and $U_{12}$ is $h \times (n-h)$, and for $A_{22}$, $L_{22}$ and $U_{22}$ is $(n-h) \times (n-h)$, where $h = n/2$. Matrices $L_{11}$, $L_{22}$, $U_{11}$ and $U_{22}$ are lower and upper triangular respectively. Matrices $A_{11}$, $A_{12}$, $A_{21}$, $A_{22}$, $L_{21}$ and $U_{12}$ are rectangular.

The RTRSM and RSYRK are recursive BLAS of _TRSM and _SYRK respectively. _TRSM solves a triangular system of equations. _SYRK performs the symmetric rank k operations (see 2.2).

2.1. Full and Packed Storage Data Format

The Cholesky factorization algorithm can be programmed either in "full storage" or "packed storage". For example the LAPACK subroutine POTRF works on full storage, while the routine PPTRF is programmed for packed storage. Here we are interested only in full storage and packed storage holding data that represents dense symmetric positive definite matrices. We will compare our recursive algorithms to the LAPACK POTRF and PPTRF subroutines.

The POTRF subroutine uses the Cholesky algorithm in full storage. A storage for the full array $A$ must be declared even if only $n \times (n+1)/2$ elements of array $A$ are needed and, hence $n \times (n-1)/2$ elements are not touched. The PPTRF subroutine uses the Cholesky algorithm on packed storage. It only needs $n \times (n+1)/2$ memory words. Moreover the POTRF subroutine works fast while the PPTRF subroutine works slow. Why? The routine POTRF is constructed with BLAS level 3, while the PPTRF uses the BLAS level 2. These LAPACK data structures are illustrated by the figs. 1 and 2 respectively.

<table>
<thead>
<tr>
<th>Lower triangular case</th>
<th>Upper triangular case</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A = \begin{pmatrix} A_{11} &amp; A_{12} \ A_{21} &amp; A_{22} \end{pmatrix}$</td>
<td>$U = \begin{pmatrix} U_{11} &amp; U_{12} \ U_{21} &amp; U_{22} \end{pmatrix}$</td>
</tr>
</tbody>
</table>

Figure 1. The mapping of 7× 7 matrix for the LAPACK Cholesky Algorithm using the full storage.
2.2. The Recursive Storage Data Format

We introduce a new storage data format, the recursive storage data format, using the recursive algorithms 2.1 and 2.2. Like packed data format this recursive storage data format requires \( n \times (n + 1)/2 \) storage for the upper or lower part of the matrix. The recursive storage data format is illustrated in fig. 3. A buffer of the size \( p \times (p - 1)/2 \), where \( p = \lceil n/2 \rceil \) (integer division), is needed to convert from the LAPACK packed storage data format to the recursive packed storage data format and back. No buffer is needed if data is given in recursive format.

We can apply the BLAS level 3 using the recursive packed storage data data format. The performance of the recursive Cholesky algorithm with the recursive packed data format reaches the performance of the LAPACK POTRF algorithm. A graph with the performance between different packed storages is presented in fig. 4.

The graph in fig. 4 presents seven curves. From the bottom: The first two curves represent LAPACK POTRF performance results for upper and lower case respectively. The third curve represents IBM ESSL performance results.

![Lower triangular case](image1)

**Figure 2.** The mapping of \( 7 \times 7 \) matrix for the LAPACK Cholesky Algorithm using the packed storage.

![Upper triangular case](image2)

**Figure 3.** The mapping of \( 7 \times 7 \) matrix for the LAPACK Cholesky Algorithm using the packed recursive storage.

<table>
<thead>
<tr>
<th></th>
<th>Lower triangular case</th>
<th>Upper triangular case</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1 2 4 7 11 16 22</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>3 5 8 12 17 23</td>
</tr>
<tr>
<td>3</td>
<td>9 14</td>
<td>6 9 13 18 24</td>
</tr>
<tr>
<td>4</td>
<td>10 15 19</td>
<td>10 14 19 25</td>
</tr>
<tr>
<td>5</td>
<td>11 16 20 23</td>
<td>15 20 26</td>
</tr>
<tr>
<td>6</td>
<td>12 17 21 24 26</td>
<td>21 27</td>
</tr>
<tr>
<td>7</td>
<td>13 18 22 25 27 28</td>
<td>28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Lower triangular case</th>
<th>Upper triangular case</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1 2 4 7 10 13 16</td>
</tr>
<tr>
<td>2</td>
<td>3 5 6</td>
<td>3 5 8 11 14 17</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>6 9 12 15 18</td>
</tr>
<tr>
<td>4</td>
<td>7 11 15 19</td>
<td>19 20 22 25</td>
</tr>
<tr>
<td>5</td>
<td>8 12 16 20 21</td>
<td>21 23 26</td>
</tr>
<tr>
<td>6</td>
<td>9 13 17 22 24 26</td>
<td>24 27</td>
</tr>
<tr>
<td>7</td>
<td>10 14 18 23 25 27 28</td>
<td>28</td>
</tr>
</tbody>
</table>
of the lower case. The last four curves give the performance of variants of the recursive algorithm. The conversion time from LAPACK packed data format to the recursive packed data format and back is included here. The fourth and fifth curves give performance of the $L$ and $U$ variants without any compiler directives. The sixth and seventh curves give performance of the $L$ and $U$ variants using compiler parallelizing directives. The SMP parallel ESSL DGEMM was used by the last five algorithms.

The same good results were obtained on other parallel supercomputers, for example on Compaq α DS-20 and SGI Origin 2000.

The forward and back substitutions perform better than LAPACK too. The recursive RTRSM is used here.

3. LU FACTORIZATION

We would like to compute the solution to a system of linear equations $AX = B$, where $A$ is a real or complex matrix, and $X$ and $B$ are rectangular matrices or vectors. Gaussian elimination with row interchanges is used to factor $A$ as $LU = PA$, where $P$ is a permutation matrix, $L$ is a unit lower triangular matrix, and $U$ is an upper triangular matrix. The factored form of $A$ is then used to solve the system of equations $AX = B$.

The recursive algorithm of the Gauss LU factorization is described in detail in [2; 5]. We give two recursive algorithms here. They are listed in figs. 5 and 6.

The matrices $A_1, A_2, A_3, A_{12}, A_{22}, L_1, L_{11}, L_{21}, L_{22}, U_1, U_3, U_{12}$ and $U_{22}$ are submatrices of $A, L$ and $U$ respectively, $a_{11} \in A_1$. 

Figure 4. The performance graphs between different packed storage data formats of the Cholesky factorization algorithm run on the 4-way IBM PowerPC 604e computer, using the double precision arithmetic.
Recursive Algorithm 3.1. Recursive LU factorization without pivoting (rgausslu):

Do recursion
- if min(m, n) > 1 then
  - \((L_1, U_1) = \text{rgausslu of } A_1\)
  - \(L_1U_1 = A_{12} \rightarrow \text{RTRSM}\)
  - \(\hat{A}_{22} = A_{22} - L_{21}U_{12} \rightarrow \_\text{GEMM}\)
  - \((L_{22}, U_{22}) = \text{rgausslu of } \hat{A}_{22}\)
- otherwise
  - \(L_1 := A_1/a_{11} \text{ and } U_1 = a_{11}\)
End recursion
- if \(n > m\) then
  - \(LU_3 = A_3 \rightarrow \text{RTRSM}\)

Figure 5.

Recursive Algorithm 3.2. Recursive LU=PA factorization with partial pivoting (rgausslu):

Do recursion
- if min(m, n) > 1 then
  - \((P_1, L_1, U_1) = \text{rgausslu of } A_1\)
  - Forward pivot \(A_2\) by \(P \rightarrow \_\text{LASWP}\)
  - \(L_1U_1 = A_{12} \rightarrow \text{RTRSM}\)
  - \(\hat{A}_{22} := A_{22} - L_{21}U_{12} \rightarrow \_\text{GEMM}\)
  - \((P_2, L_{22}, U_{22}) = \text{rgausslu of } \hat{A}_{22}\)
  - Back pivot \(A_1\) by \(P_2 \rightarrow \_\text{LASWP}\)
  - \(P = P_2P_1\)
- otherwise
  - pivot \(A_1\)
  - \(L_1 := A_1/a_{11} \text{ and } U_1 := a_{11}\)
End recursion
- if \(n > m\) then
  - Forward pivot \(A_3\) by \(P \rightarrow \_\text{LASWP}\)
  - \(LU_3 = A_3 \rightarrow \text{RTRSM}\)

where \(P_1\) and \(P_2\) are permutation matrices.

Figure 6.
4. RECURSIVE BLAS AND THEIR PARALLELISM

Two recursive BLAS (Basic Linear Algebra Subprograms, see [12]) subroutines are used in our recursive Cholesky and recursive LU algorithms: RTRSM and RSYRK. These two routines will be explained below.

4.1. RTRSM

RTRSM is a recursive formulation of TRSM, where _ is a precision and arithmetic indicator: S, D, C or Z. TRSM subroutine solves one of the matrix equation

\[ AX = \alpha B, A^T X = \alpha B, X A = \alpha B, \text{or} X A^T = \alpha B, \]

where \( \alpha \) is a scalar. \( X \) and \( B \) are \((m \times n)\) rectangular matrices. \( A \) is a unit, or non-unit, upper or lower \((m \times m)\) triangular matrix. The matrix \( X \) is overwritten on \( B \). We have 16 different triangular equations because \( A \) and \( A^T \) can be either upper or lower triangular matrices, and the diagonal is normal or unit.

We will introduce the recursive formulation only for one case \( AX = \alpha B \), where \( A \) is lower triangular. The other cases will be similar.

The matrices \( A, B, \) and \( X \) can be partitioned into smaller submatrices, thus

\[
\begin{pmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{pmatrix}
\begin{pmatrix}
X_{11} & X_{12} \\
X_{21} & X_{22}
\end{pmatrix}
= \alpha
\begin{pmatrix}
B_{11} & B_{12} \\
B_{21} & B_{22}
\end{pmatrix}
\]

The matrices \( A_{11} = A(1 : h,1 : h), A_{21} = A(h+1 : m,1 : h), A_{22} = A(h+1 : m,h+1 : m), B_{11} = B(1 : h,1 : p), B_{12} = B(1 : h,p+1 : n), B_{21} = B(h+1 : m,1 : p), B_{22} = B(h+1 : m,p+1 : n), X_{11} = X(1 : h,1 : p), X_{12} = X(1 : h,p+1 : n), X_{21} = X(h+1 : m,1 : p) \) and \( X_{22} = X(h+1 : m,p+1 : n) \) are submatrices of \( A, B \) and \( X \) respectively.

Multiplying the matrix \( A \) by \( X \) gives:

\[
\begin{pmatrix}
A_{11}X_{11} & A_{11}X_{12} \\
A_{21}X_{11} + A_{22}X_{21} & A_{21}X_{12} + A_{22}X_{22}
\end{pmatrix}
= \begin{pmatrix}
\alpha B_{11} & \alpha B_{12} \\
\alpha B_{21} & \alpha B_{22}
\end{pmatrix}
\]

We have got two independent groups of triangular systems:

\[
\begin{align*}
A_{11}X_{11} &= \alpha B_{11} \\
A_{22}X_{21} &= \alpha B_{21} - A_{21}X_{11}
\end{align*}
\]

\[
\begin{align*}
A_{11}X_{12} &= \alpha B_{12} \\
A_{22}X_{22} &= \alpha B_{22} - A_{21}X_{12}
\end{align*}
\]

We could do a double recursion on \( m \) and \( n \); i.e. on \( h \) and \( p \). However, we do not do the recursion on \( p \). This results in the following algorithm:

**Recursive Algorithm 4.1.** Recursive algorithm for the \( AX = B \) operation (one group only), where \( A \) is a lower triangular matrix (rtrsm):

*Do recursion*
if \( m > 1 \) then
- \( A_{11}X_1 = \alpha B_1 \) \( \rightarrow \) RTRSM
- \( B_2 := \alpha B_1 - A_{21}X_1 \) \( \rightarrow \) GEMM
- \( A_{22}X_2 = \alpha B_2 \) \( \rightarrow \) RTRSM
otherwise
- \( a_{11}X_1 = \alpha B_1 \)

\[ End \ \text{recursion} \]

### 4.2. RSYRK

RSYRK is a recursive formulation of SYRK, where \( _\_ \) is a precision and arithmetic indicator: S, D, C or Z. SYRK performs one of the symmetric rank k operations:

\[
C := \alpha A A^\top + \beta C \quad \text{or} \quad C := \alpha A^\top A + \beta C,
\]

where \( \alpha \) and \( \beta \) are scalars. \( A \) is a rectangular matrix \((m \times n)\). \( C \) is a square symmetric matrix.

We will introduce the recursive formulation only for one of the four cases of SYRK:

\[
C := \alpha A A^\top + \beta C.
\]

The matrices \( A \) and \( C \) can be partitioned into smaller submatrices:

\[
\begin{pmatrix}
C_{11} & C_{12} \\
C_{21} & C_{22}
\end{pmatrix} = \beta
\begin{pmatrix}
C_{11} & C_{21} \\
C_{21} & C_{22}
\end{pmatrix} + \alpha
\begin{pmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{pmatrix}
\begin{pmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{pmatrix}^\top
\]

The matrices \( A_{11} = A(1: h, 1: p) \), \( A_{12} = A(1: h, p + 1: n) \), \( A_{21} = A(h + 1: m, 1: p) \), \( A_{22} = A(h + 1: m, p + 1: n) \), \( C_{11} = C(1: h, 1: h) \), \( C_{21} = C(h + 1: m, 1: h) \), \( C_{22} = C(h + 1: m, h + 1: m) \) are submatrices of \( A \) and \( C \), respectively. The recursion could be done again on two variables, \( h \) and \( p \), but we do recursion on \( h \) only.

In terms of the partitioning we have three independent formulas:

\[
\begin{align*}
C_{11} &= \beta C_{11} + \alpha A_{11}A_{11}^\top + \alpha A_{12}A_{12}^\top, \\
C_{21} &= \beta C_{21} + \alpha A_{21}A_{11}^\top + \alpha A_{22}A_{12}^\top, \\
C_{22} &= \beta C_{22} + \alpha A_{21}A_{21}^\top + \alpha A_{22}A_{22}^\top.
\end{align*}
\]

These three computations can be computed in parallel. We now formulate a recursive algorithm as follows:

**Recursive Algorithm 4.2.** Recursive algorithm for the \( C := \alpha A A^\top + \beta C \) symmetric rank k operations (rsyrk):

\[ Do \ \text{recursion} \]
• if \( m \geq 1 \) then
  Perform computation \( C_{11} \):
  \[
  \hat{C}_{11} := \beta C_{11} + \alpha A_{11} A_{11}^T \rightarrow \text{RSYRK}
  \]
  \[
  C_{11} := \hat{C}_{11} + \alpha A_{12} A_{12}^T \rightarrow \text{RSYRK}
  \]
  Perform computation \( C_{21} \):
  \[
  \hat{C}_{21} := \beta C_{21} + \alpha A_{21} A_{21}^T \rightarrow \_\text{GEMM}
  \]
  \[
  C_{21} := \hat{C}_{21} + \alpha A_{22} A_{22}^T \rightarrow \_\text{GEMM}
  \]
  Perform computation \( C_{22} \):
  \[
  \hat{C}_{22} := \beta C_{22} + \alpha A_{21} A_{21}^T \rightarrow \text{RSYRK}
  \]
  \[
  C_{22} := \hat{C}_{22} + \alpha A_{22} A_{22}^T \rightarrow \text{RSYRK}
  \]
  End recursion

4.3. A fast _GEMM algorithm

The _GEMM is a precision and arithmetic indicator: \( S, D, C \) or \( Z \). _GEMM subroutine does the following operations

\[
C := \alpha AB + \beta C, \quad C := \alpha AB^T + \beta C, \quad C := \alpha A^T B + \beta C,
\]

where \( \alpha \) and \( \beta \) are scalars. \( A, B \) and \( C \) are rectangular matrices. \( A^T, B^T, A^C \) and \( B^C \) are transpose and conjugate matrices respectively.

The GEMM operation is very well documented and explained in [6; 12]. We can see that work is done by _GEMM for both our BLAS RTRSM 4.1 and RSYRK 4.2. The speed of our computation depends very much from the speed of a good _GEMM. Good _GEMM implementations are usually developed by computer manufacturers. The model implementation of _GEMM can be obtained from netlib [12]; it works correctly but slowly. However, an excellent set of high performance BLAS, called _GEMM based BLAS was developed by Bo Kågström at the University of Umeå in Sweden, see for example [7]. A key idea behind _GEMM based BLAS was to cast all BLAS algorithms in terms of the simple BLAS _GEMM. Recently, the Innovative Computing Laboratory at University of Tennessee in Knoxville developed a system called ATLAS (see 4.3.1) which can produce a fast _GEMM program.

4.3.1. ATLAS

Automatically Tuned Linear Algebra Software (ATLAS) [11]. ATLAS is an approach for the automatic generation and optimization of numerical software for processors with deep memory hierarchies and pipelined functional units. The production of such software for machines ranging from desktop workstations to embedded processors can be a tedious and time consuming task. ATLAS has been designed to automate much of this process. So, having a fast
_GEMM means our RTRSM and RSYRK routines will be fast. The ATLAS GEMM is often better than _GEMM developed by the computer manufacture. What is important, the ATLAS software is available to everybody, free of charge. Every personal computer can have a good GEMM.

REFERENCES


LAWRA – REKURSYVINIAI TIESINĖS ALGEBROS ALGORITMAI

B. ANDERSEN, F. GUSTAVSON, A. KARAIVANOV, J. WASNIEWSKI, P. YALAMOV