Compiler Blockability of Dense Matrix Factorizations

STEVE CARR
Michigan Technological University
and
R. B. LEHOUCQ
Argonne National Laboratory

The goal of the LAPACK project is to provide efficient and portable software for dense numerical linear algebra computations. By recasting many of the fundamental dense matrix computations in terms of calls to an efficient implementation of the BLAS (Basic Linear Algebra Subprograms), the LAPACK project has, in large part, achieved its goal. Unfortunately, the efficient implementation of the BLAS results often in machine-specific code that is not portable across multiple architectures without a significant loss in performance or a significant effort to reoptimize them. This article examines whether most of the hand optimizations performed on matrix factorization codes are unnecessary because they can (and should) be performed by the compiler. We believe that it is better for the programmer to express algorithms in a machine-independent form and allow the compiler to handle the machine-dependent details. This gives the algorithms portability across architectures and removes the error-prone, expensive, and tedious process of hand optimization. Although there currently exist no production compilers that can perform all the loop transformations discussed in this article, a description of current research in compiler technology is provided that will prove beneficial to the numerical linear algebra community. We show that the Cholesky and optimized automatically by a compiler to be as efficient as the same hand-optimized version found in LAPACK. We also show that the QR factorization may be optimized by the compiler to perform comparably with the hand-optimized LAPACK version on modest matrix sizes. Our approach allows us to conclude that with the advent of the compiler optimizations discussed in this article, matrix factorizations may be efficiently implemented in a BLAS-less form.

Categories and Subject Descriptors: D.3.4 [**Programming Languages**]: Processors—*compilers*; *optimization*; F.2.1 [**Analysis of Algorithms and Problem Complexity**]: Numerical Algorithms and Problems—*computations on matrices*; G.1.3 [**Numerical Analysis**]: Numerical Linear Algebra—*linear systems*; G.4 [**Mathematics of Computing**]: Mathematical Soft-

This research was supported by NSF grants CCR-9120008 and CCR-9409341. R. B. Lehoucq was also supported by the U.S. Department of Energy contracts DE-FG0f-91ER25103 and W-31–109–Eng-38.
Authors' addresses: S. Carr, Department of Computer Science, Michigan Technological University, Houghton, MI 49931; email: carr@cs.mtu.edu; R. B. Lehoucq, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439; email: lehoucq@mcs.anl.gov; http://www.mcs.anl.gov/home/lehoucq/index.html.

ware—*efficiency*; *portability*

General Terms: Languages, Performance

Additional Key Words and Phrases: BLAS, cache optimization, Cholesky decomposition, LAPACK, LU decomposition, QR decomposition

1. INTRODUCTION

The processing power of microprocessors and supercomputers has increased dramatically and continues to do so. At the same time, the demand on the memory system of a computer is to increase dramatically in size. Due to cost restrictions, typical workstations cannot use memory chips that have the latency and bandwidth required by today's processors. Instead, main memory is constructed of cheaper and slower technology, and the resulting delays may be up to hundreds of cycles for a single memory access.

To alleviate the memory speed problem, machine architects construct a hierarchy of memory where the highest level (registers) is the smallest and fastest and where each lower level is larger but slower. The bottom of the hierarchy for our purposes is main memory. Typically, one or two levels of cache memory fall between registers and main memory. The cache memory is faster than main memory, but is often a fraction of the size. The cache memory serves as a buffer for the most recently accessed data of a program (the working set). The cache becomes ineffective when the working set of a program is larger than its size.

The three factorizations considered in this article—the LU, Cholesky, and QR—are among the most frequently used by numerical linear algebra and its applications. The first two are used for solving linear systems of equations, while the last is typically used in linear least-squares problems. For square matrices of order n, all three factorizations involve on the order of n^3 floating-point operations for data that need n^2 memory locations. With the advent of vector and parallel supercomputers, the efficiency of the factorizations were seen to depend dramatically upon the algorithmic form chosen for the implementation [Dongarra et al. 1984; Gallivan et al. 1990; Ortega 1988]. These studies concluded that managing the memory hierarchy is the single most important factor governing the efficiency of the software implementation computing the factorization.

The motivation of the LAPACK [Anderson et al. 1995] was to recast the algorithms in the EISPACK [Smith et al. 1976] and LINPACK [Dongarra et al. 1979] software libraries with *block* ones. A block form of an algorithm restructures the algorithm in terms of matrix operations that attempt to minimize the amount of data moved within the memory hierarchy while keeping the arithmetic units of the machine occupied. LAPACK blocks many dense matrix algorithms by restructuring them to use the level 2 and 3 BLAS [Dongarra et al. 1988; 1990]. The motivation for the Basic Linear Algebra Subprograms (BLAS) [Lawson et al. 1979] was to provide a set of

commonly used vector operations so that the programmer could invoke the subprograms instead of writing the code directly. The level 2 and 3 BLAS followed with matrix-vector and matrix-matrix operations, respectively, that are often necessary for high efficiency across a broad range of high-performance computers. The higher-level BLAS better utilize the underlying memory hierarchy. As with the level 1 BLAS, responsibility for optimizing the higher-level BLAS was left to the machine vendor or another interested party.

This article investigates whether a compiler has the ability to block matrix factorizations. Although the compiler transformation techniques may be applied directly to the BLAS, it is interesting to draw a comparison by applying them directly to the factorizations. The benefit is the possibility of BLAS-less linear algebra package that is nearly as efficient as LAPACK. For example, in Lehoucq [1992], it was demonstrated that on some computers the best optimized set of BLAS was available.

We deem an algorithm *blockable* if a compiler can automatically derive the most efficient block algorithm (for our study, the one found in LAPACK) from its corresponding machine-independent point algorithm. In particular, we show that LU and Cholesky factorizations are blockable algorithms. Unfortunately, the QR factorization with Householder transformations is not blockable. However, we show an alternative block algorithm for QR that can be derived using the same compiler methods as those used for LU and Cholesky factorizations.

This article has yielded two major results. The first, which is detailed in another paper [Carr and Kennedy 1992], reveals that the hand loop unrolling performed when optimizing the level 2 and 3 BLAS [Dongarra et al. 1988; 1990] is often unnecessary. While the BLAS are useful, the hand optimization that is required to obtain good performance on a particular architecture may be left to the compiler. Experiments show that, in most cases, the compiler can automatically unroll loops as effectively as hand optimization. The second result, which we discuss in this article, reveals that it is possible to block matrix factorizations automatically. Our results show that the block algorithms derived by the compiler are competitive with those of LAPACK [Anderson et al. 1995]. For modest-sized matrices (on the order of 200 or less), the compiler-derived variants are often superior.

We begin our presentation with a review of background material related to compiler optimization. Then, we describe a study of the application of compiler analysis to derive the three block algorithms in LAPACK considered above from their corresponding point algorithms. We present an experiment comparing the performance of hand-optimized LAPACK with the compiler-derived algorithms attained using our techniques. We also briefly discuss other related approaches. Finally, we summarize our results and provide and draw some general conclusions.

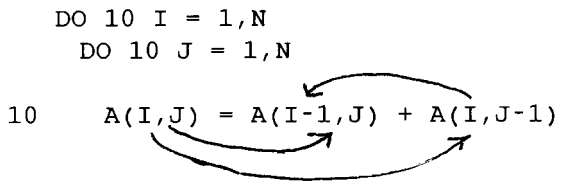

Fig. 1. Loop-carried dependencies.

2. BACKGROUND

The transformations that we use to create the block versions of matrix factorizations from their corresponding point versions are well known in the mathematical software community [Dongarra et al. 1991]. This section introduces the fundamental tools that the compiler needs to perform the same transformations automatically. The compiler optimizes point versions of matrix factorizations through analysis of array access patterns rather than through linear algebra.

2.1 Dependence

As in vectorizing and parallelizing compilers, *dependence* is a critical compiler tool for performing transformations to improve the memory performance of loops. Dependence is necessary for determining the legality of compiler transformations to create blocked versions of matrix factorizations by giving a partial order on the statements within a loop nest.

A dependence exists between two statements if there exists a control flow path from the first statement to the second and if both statements reference the same memory location [Kuck 1978].

—If the first statement writes to the location, and the second reads from it, there is a *true dependence*, also called a *flow dependence*.

—If the first statement reads from the location, and the second writes to it, there is an *antidependence*.

—If both statements write to the location, there is an *output dependence*.

—If both statements read from the location, there is an *input dependence*.

A dependence is *carried* by a loop if the references at the source and sink (beginning and end) of the dependence are on different iterations of the loop and if the dependence is not carried by an outer loop [Allen and Kennedy 1987]. In the loop below, there is a true dependence from `A(I,J)` to `A(I-1,J)` carried by the `I`-loop, a true dependence from `A(I,J)` to `A(I,J-1)` carried by the `J`-loop and an input dependence from `A(I,J-1)` to `A(I-1,J)` carried by the `I`-loop (see Figure 1).

To enhance the dependence information, *section* analysis can be used to describe the portion of an array that is accessed by a particular reference or set of references [Callahan and Kennedy 1987; Havlak and Kennedy 1991].

```
      DO 10 I = 1,M
        DO 10 J = 1,10
10        A(J,I) = …
```

Fig. 2. Section analysis.

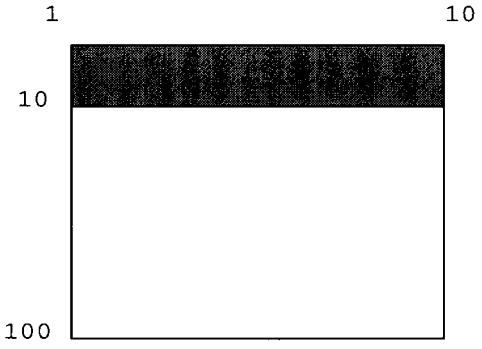

Fig. 3. Section of A.

Sections describe common substructures of arrays such as elements, rows, columns, and diagonals. As an example of section analysis consider the loop in Figure 2.

If A were declared to be 100×100, the section of A accessed in the loop would be that shown in the shaded portion of Figure 3.

Matrix factorization codes require us to enhance basic dependence information because only a portion of the matrix is involved in the block update. The compiler uses section analysis to reveal that portion of the matrix that can be block updated. Section 3 discusses this in detail.

3. AUTOMATIC BLOCKING OF DENSE MATRIX FACTORIZATIONS

In this section, we show how to derive the block algorithms for the LU and the Cholesky factorizations using current compiler technology and section analysis to enhance dependence information. We also show that the QR factorization with Householder transformations is not blockable. However, we present a performance-competitive version of the derivable by the compiler.

3.1 LU Factorization

The LU decomposition factors a nonsingular matrix A into the product of two matrices, L and U, such that $A = LU$ [Golub and Van Loan 1996]. L is a unit lower triangular matrix, and U is an upper triangular matrix. This factorization can be obtained by multiplying the matrix A by a series of elementary lower triangular matrices, $M_{n-1} \cdots M_1$ and pivot matrices

```
      DO 10 K = 1,N-1
        DO 20 I = K+1,N
20        A(I,K) = A(I,K) / A(K,K)
        DO 10 J = K+1,N
          DO 10 I = K+1,N
10          A(I,J) = A(I,J) - A(I,K) * A(K,J)
```

Fig. 4. Right-looking LU factorization.

$P_{n-1} \cdots P_1$, where $L^{-1} = M_{n-1}P_{n-1} \cdots M_1P_1$ and $U = L^{-1}A$. (The pivot matrices are used to make the LU factorization a numerically stable process.)

We first examine the blockability of LU factorization. Since pivoting creates its own difficulties, we first show how to block LU factorization without pivoting. We then show how to handle pivoting.

3.1.1 *No Pivoting*. Consider the algorithm for LU factorization (Figure 4).

This point algorithm is referred to as an unblocked right-looking [Dongarra et al. 1991] algorithm. It exhibits poor cache performance on large matrices. To transform the point algorithm to the block algorithm, the compiler must perform *strip-mine-and-interchange* on the K-loop [Lam et al. 1991; Wolf and Lam 1991]. This transformation is used to create the block update of A. To apply this transformation, we first *strip* the K-loop into fixed-size sections (this size is dependent upon the target architecture's cache characteristics and is beyond the scope of this article [Coleman and McKinley 1995; Lam et al. 1991]), as shown in Figure 5.

Here KS is the machine-dependent strip size that is related to the cache size. To complete the transformation, the KK-loop must be distributed around the loop that surrounds statement 20 and around the loop nest that surrounds statement 10 before being interchanged to the innermost position of the loop surrounding statement 10 [Wolfe 1986]. This distribution yields the algorithm shown in Figure 6.

Unfortunately the loop is no longer correct. This loop scales a number of values before it updates them. Dependence analysis allows the compiler to detect and avoid this change in semantics by recognizing the dependence cycle between A(I,KK) in statement 20 and A(I,J) in statement 10 carried by the KK-loop.

Using basic dependence analysis only, it appears that the compiler would be prevented from blocking LU factorization due to the cycle. However, enhancing dependence analysis with section information reveals that the cycle only exists for a portion of the data accessed in both statements. Figure 7 shows the sections of the array A accessed for the entire execution of the KK-loop. The section accessed by A(I,KK) in statement 20 is a subset of the section accessed by A(I,J) in statement 10.

```
      DO 10 K = 1,N-1,KS
        DO 10 KK = K,MIN(K+KS-1,N-1)
          DO 20 I = KK+1,N
20          A(I,KK) = A(I,KK) / A(KK,KK)
          DO 10 J = KK+1,N
            DO 10 I = KK+1,N
10            A(I,J) = A(I,J) - A(I,KK) * A(KK,J)
```

Fig. 5. Stripping the K-loop.

```
      DO 10 K = 1,N-1,KS
        DO 20 KK = K,MIN(K+KS-1,N-1)
          DO 20 I = KK+1,N
20          A(I,KK) = A(I,KK) / A(KK,KK)
        DO 10 KK = K,MIN(K+KS-1,N-1)
          DO 10 J = KK+1,N
            DO 10 I = KK+1,N
10            A(I,J) = A(I,J) - A(I,KK) * A(KK,J)
```

Fig. 6. Distributing the KK–loop.

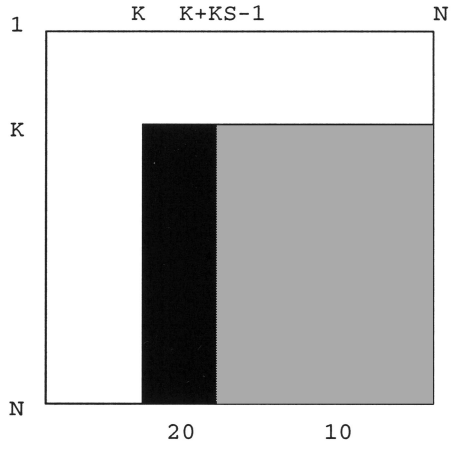

Fig. 7. Sections of A in LU factorization.

Since the recurrence exists for only a portion of the iteration space of the loop surrounding statement 10, we can split the J-loop into two loops—one loop iterating over the portion of A where the dependence cycle exists and one loop iterating over the portion of A where the cycle does not exist—using a transformation called *index-set splitting* [Wolfe 1987]. J can be split at the point J = K + KA – 1 to create the two loops as shown in Figure 8.

```
   DO 10 K = 1,N-1,KS
     DO 10 KK = K,MIN(K+KS-1,N-1)
       DO 20 I = KK+1,N
20       A(I,KK) = A(I,KK) / A(KK,KK)
       DO 30 J = KK+1,MIN(K+KS-1,N)
         DO 30 I = KK+1,N
30         A(I,J) = A(I,J) - A(I,KK) *A(KK,J)
       DO 10 J = K+KS,N
         DO 10 I = KK+1,N
10         A(I,J) = A(I,J) - A(I,KK) * A(KK,J)
```

Fig. 8. Index-set splitting.

```
   DO 10 K = 1,N-1,KS
     DO 30 KK = K,MIN(K+KS-1,N-1)
       DO 20 I = KK+1,N
20       A(I,KK) = AI9,KK) / A(KK,KK)
       DO 30 J = KK+1,MIN(K+KS-1,N)
         DO 30 I = KK+1,N
30         A(I,J) = A(I,J) - A(I,KK) * A(KK,J)
     DO 10 KK = K,MIN(K+KS-1,N)
       DO 10 J = K+KS,N
         DO 10 I = KK+1,N
10         A(I,J) = A(I,J) - A(I,KK) * A(KK,J)
```

Fig. 9. Strip-mine-and-interchange.

Now the dependence cycle exists between statements 20 and 30, and statement 10 is no longer in the cycle. Strip-mine-and-interchange can be continued by distributing the KK-loop around the two new loops as shown in Figure 9.

To finish strip-mine-and-interchange, we need to move the KK-loop to the innermost position in the nest surrounding statement 10. However, the lower bound of the I-loop contains a reference to KK. This creates a *triangular* iteration space as shown in Figure 10. To interchange the KK- and I- loops, the intersection of the line I=KK+1 with the iteration space at the point (K,K+1) must be handled. Therefore, interchanging the loops requires the KK-loop to iterate over a trapezoidal region with an upper bound of I-1 until $I - 1 > K + KS - 1$ (see Wolfe [1987] and Carr and Kennedy [1992] for more details on transforming nonrectangular loop nests). This gives the loop nest shown in Figure 11.

At this point, a right-looking [Dongarra et al. 1991] block algorithm has been obtained. Therefore, statement 10 is a matrix-matrix multiply that can be further optimized depending upon the architecture. For superscalar architectures whose performance is bound by cache, outer loop unrolling on

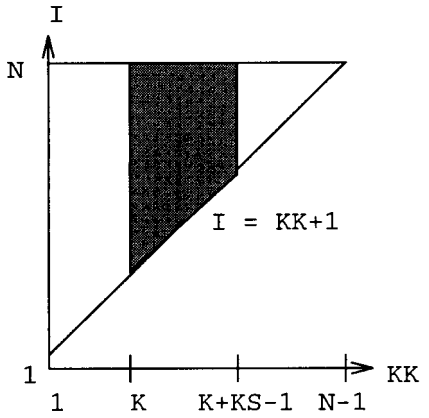

Fig. 10. Iteration spce of the LU factorization.

```
   DO 10 K = 1,N-1,KS
     DO 30 KK = K,MIN(K+KS-1,N-1)
       DO 20 I = KK+1,N
20       A(I,KK) = A(I,KK) / A(KK,KK)
       DO 30 J = KK+1,MIN(K+KS-1,N)
         DO 30 I = KK+1,N
30         A(I,J) = A(I,J) - A(I,KK) * A(KK,J)
     DO 10 J = K+KS,N
       DO 10 I = K+1,N
         DO 10 KK = K,MIN(I-1,MIN(K+KS-1,N-1))
10         A(I,J) = A(I,J) - A(I,KK) * A(KK,J)
```

Fig. 11. Right-looking block LU decomposition.

nonrectangular loops can be applied to the J- and I-loops to further improve performance [Carr and Kennedy 1992; 1994]. For vector architectures, a different loop optimization strategy may be more beneficial [Allen and Kennedy 1987].

Many of the transformations that we have used to obtain the block version of LU factorization are well known in the compiler community and exist in many commercial compilers (e.g., HP, DEC, and SGI). One of the contributions of this article to compiler research is to show how the addition of section analysis allows a compiler to block matrix factorizations. We remark that none of the aforementioned compilers uses section analysis for this purpose.

3.1.2 *Adding Partial Pivoting*. Although the compiler can discover the potential for blocking in LU decomposition without pivoting using index-set

```
        DO 10 K = 1,N-1
C
C …   pick pivot --- IMAX
C
          DO 30 J = 1,N
            TAU = A,(K,J)
     25     A(K,J) = A(IMAX,J)
     30     A(IMAX,J) = TAU
          DO 20 I = K+1,N
     20     A(I,K) = A(I,K) / A(K,K)
          DO 10 J = K+1,N
            DO 10 I = K+1,N
     10       A(I,J) = A(I,J) - A(I,K) * A(K,J)
```

Fig. 12. LU decomposition with partial pivoting.

```
      DO 10 KK = K,K+KS-1
        DO 30 J = 1,N
          TAU = A(KK,J)
   25     A(KK,J) = A(IMAX,J)
   30     A(IMAX,J) = TAU
        DO 10 J = KK+KS,N
          DO 10 I = KK+1,N
   10       A(I,J) = A(I,J) - A(I,KK) * A(KK,J)
```

Fig. 13. LU decomposition with partial pivoting.

splitting and section analysis, the same cannot be said when partial pivoting is added (see Figure 12 for LU decomposition with partial pivoting). In the partial-pivoting algorithm, a new recurrence exists that does not fit the form handled by index-set splitting. Consider sections of code, shown in Figure 13, after applying index-set splitting to the algorithm in Figure 12.

The reference to `A(IMAX,J)` in statement `25` and the reference to `A(I,J)` in statement `10` access the same sections. Distributing the `KK`-loop around both `J`-loops would convert the true dependence from `A(I,J)` to `A(IMAX,J)` into an antidependence in the reverse direction. The rules for the preservation of data dependence prohibit the reversing of a dependence direction. This would seem to preclude the existence of a block analogue similar to the nonpivoting case. However, a block algorithm that ignores the preventing recurrence and distributes the `KK`-loop can still be mathematically derived [Dongarra et al. 1991].

Consider the following. If

$$M_1 = \begin{pmatrix} 1 & 0 \\ -m_1 & I \end{pmatrix}, P_2 = \begin{pmatrix} 1 & 0 \\ 0 & \hat{P}_2 \end{pmatrix}$$

then

$$P_2 M_1 = \begin{pmatrix} 1 & 0 \\ -\hat{P}_2 m_1 & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \hat{P}_2 \end{pmatrix} \equiv \hat{M}_1 P_2 \tag{1}$$

This result shows that we can postpone the application of the eliminator M_1 until after the application of the permutation matrix P_2 if we also permute the rows of the eliminator. Extending Eq. (1) to the entire formulation, we have

$$U = M_{n-1}\hat{M}_{n-2}\hat{M}_{n-3}\cdots\hat{M}_1 P_{n-1}P_{n-2}P_{n-3}\cdots P_1 A = MPA.$$

In the implementation of the block algorithm, P_i cannot be computed until step i of the point algorithm. P_i only depends upon the first i columns of A, allowing the computation of k P_i and $\hat{M}_i$, where k is the blocking factor, and then the block application of the $\hat{M}_i$ [Dongarra et al. 1991].

To install the above result into the compiler, we examine its implications from a data dependence viewpoint. In the point version, each row interchange is followed by a whole-column update in which each row element is updated independently. In the block version, multiple row interchanges may occur before a particular column is updated. The same computations (column updates) are performed in both the point and block versions, but these computations may occur in different locations (rows) of the array. The key concept for the compiler to understand is that row interchanges and whole-column updates are commutative operations. Data dependence alone is not sufficient to understand this. A data dependence relation maps values to memory locations. It reveals the sequence of values that pass through a particular location. In the block version of LU decomposition, the sequence of values that pass through a location is different from the point version, although the final values are identical. Unless the compiler understands that row interchanges and column updates commute, LU decomposition with partial pivoting is not blockable.

Fortunately, a compiler can be equipped to understand that operations on whole columns are commutable with row permutations. To upgrade the compiler, one would have to install pattern matching to recognize both the row permutations and whole-column updates to prove that the recurrence involving statements `10` and `25` of the index-set split code could be ignored. Forms of pattern matching are already done in commercially available compilers. Vectorizing compilers pattern match for specialized computations such as searching vectors for particular conditions [Levine et al. 1991]. Other preprocessors pattern match to recognize matrix multiplication and, in turn, output a predetermined solution that is optimal for a

```
      DO 10 K = 1,N-1,KS
        DO 10 KK = K,MIN(K+KS-1,N-1)
          A(KK,KK) = SQRT( A(KK,KK) )
          DO 20 I = KK+1,N
20          A(I,KK) = A(I,KK) / A(KK,KK)
          DO 10 J = KK+1,N
            DO 10 I = J,N
10            A(I,J) = A(I,J) - A(I,KK) * A(J,KK)
```

Fig. 14. Strip-mined Cholesky factorization.

particular machine. So, it is reasonable to believe that pivoting can be recognized and implemented in commercial compilers if its importance is emphasized.

3.2 Cholesky Factorization

When the matrix A is symmetric and positive definite, the factorization may be written as

$$A = LU = LD(D^{-1}U) = LD^{1/2}D^{1/2}L^T \equiv \hat{L}\hat{L}^T,$$

where $\hat{L} = LD^{1/2}$ and where D is the diagonal matrix consisting of the main diagonal of U. The decomposition of A into the product of a triangular matrix and its transpose is called the Cholesky factorization. Thus, we need only work with the lower triangular half of A, and essentially the same dependence analysis that applies to the LU factorization without pivoting may be used. Note that with respect to floating-point computation, the Cholesky factorization only differs from LU in two regards. The first is that there are n square roots for Cholesky, and the second is that only the lower half of the matrix needs to be updated.

The strip-mined version of the Cholesky factorization is shown in Figure 14.

As is the case with LU factorization, there is a recurrence between `A(I,J)` in statement `10` and `A(I,KK)` in statement `20` carried by the `KK`-loop. The data access patterns in Cholesky factorization are identical to LU factorization (see Figure 7); index-set splitting can be applied to the `J`-loop at `K+KS-1` to allow the `KK`-loop to be distributed, achieving the LAPACK block algorithm.

3.3 QR Factorization

In this section, we examine the blockability of the QR factorization. First, we show that the algorithm from LAPACK is not blockable. Then, we give an alternate algorithm that is blockable.

3.3.1 *LAPACK Version*. The LAPACK point algorithm for computing the QR factorization consists of forming the sequence $A_{k+1} = V_k A_k$ for $k = 1, \ldots, n-1$. The initial matrix $A_1 = A$ has m rows and n columns, where for this article we assume $m \geq n$. The elementary reflectors $V_k = I - \tau_k v_k v_k^T$ update A_k so that the first k columns of A_{k+1} form an upper triangular matrix. The update is accomplished by performing the matrix-vector multiplication $w_k = A^T v_k$ followed by the rank-one update $A_{k+1} = A_k - \tau_k v_k w_k^T$. Efficiency of the implementation of the level 2 BLAS subroutines determines the rate at which the factorization is computed. For a more detailed discussion of the QR factorization see Golub and Van Loan [1996].

The LAPACK block QR factorization is an attempt to recast the algorithm in terms of calls to level 3 BLAS [Dongarra et al. 1991]. If the level 3 BLAS are hand-tuned for a particular architecture, the block QR algorithm may perform significantly better than the point version on large matrix sizes (those that cause the working set to be much larger than the cache size).

Unfortunately, the block QR algorithm in LAPACK is not automatically derivable by a compiler. The block application of a number of elementary reflectors involves both computation and storage that do not exist in the original point algorithm [Dongarra et al. 1991]. To block a number of eliminators together, the following is computed:

$$\begin{aligned} Q &= (I - \tau_1 v_1 v_1^T)(I - \tau_2 v_2 v_2^T) \cdots (I - \tau_{n-1} v_{n-1} v_{n-1}^T) \\ &= I - VTV^T. \end{aligned}$$

The compiler cannot derive $I - VTV^T$ from the original point algorithm using dependence information. To illustrate, consider a block of two elementary reflectors

$$\begin{aligned} Q &= (I - \tau_1 v_1 v_1^T)(I - \tau_2 v_2 v_2^T), \\ &= I - (v_1 v_2) \begin{pmatrix} \tau_1 & \tau_1 \tau_2 (v_1^T v_2) \\ 0 & \tau_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix}. \end{aligned}$$

The computation of the matrix

$$T = \begin{pmatrix} \tau_i & \tau_1 \tau_2 (v_1^T v_2) \\ 0 & \tau_2 \end{pmatrix}$$

is not part of the original algorithm. Hence, the LAPACK version of block QR factorization is a different algorithm from the point version, rather than just a reshaping of the point algorithm for better performance. The compiler can reshape algorithms, but it cannot derive new algorithms with data dependence information. In this case, the compiler would need to understand linear algebra to derive the block algorithm.

```
                DO II = 1, N, IB

                  DO I = II, MINO(II+IB-1,N)
*
*                   Generate elementary reflector V_i.
*
                    DO J = I+1, M
                      A(J,I) = A(J,I)/(A(I,I)-BETA)
                    END DO
*
*                   Update A(i:m,i+1:n) with V_i
*
                    DO J = I+1, N

                      T1 = ZERO
                      DO K = 1, M
                        T1 = T1 + A(K,J)*A(K,I)
                      ENDDO

                      DO K = I, M
                        A(K,J) = A(K,J) - TAU(I)*T1*A(K,I)
                      ENDDO

                    ENDDO
                  ENDDO
                ENDDO
```

Fig. 15. Strip-mined point QR decomposition.

In the next section, a compiler-derivable block algorithm for QR factorization is presented. This algorithm gives comparable performance to the LAPACK version on small matrices while retaining machine independence.

3.3.2 *Compiler-Derivable QR Factorization*. Consider the application of j matrices V_k to A_k,

$$A_{k+j} = (I - \tau_{k+j-1} v_{k+j-1} v_{k+j-1}^T) \cdots (I - \tau_{k+1} v_{k+1} v_{k+1}^T)(I - \tau_k v_k v_k^T) A_k.$$

The compiler-derivable algorithm, henceforth called *cd*-QR, only forms columns k through $k + j - 1$ of A_{k+j} and then updates the remainder of matrix with the j elementary reflectors. The final update of the trailing $n - k - j$ columns is "rich" in floating-point operations that the compiler organizes to best suit the underlying hardware. Code optimization techniques such as strip-mine-and-interchange and unroll-and-jam are left to the compiler. The derived algorithm depends upon the *compiler* for

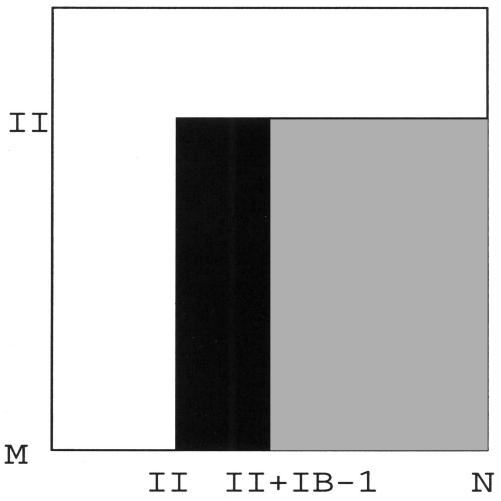

Fig. 16. Regions of A accessed by QR decomposition.

efficiency, in contrast to the LAPACK algorithm which depends on hand optimization of the BLAS

Cd-QR can be obtained from the point algorithm for QR decomposition using array section analysis. For reference, segments of the code for the point algorithm after strip mining of the outer loop are shown in Figure 15. To complete the transformation of the code in Figure 15 to obtain *cd*-QR, the `I`-loop must be distributed around the loop that surrounds the computation of V_i and around the update before being interchanged with the `J`-loop. However, there is a recurrence between the definition and use of `A(K,J)` within the update section and the definition and use of `A(J,I)` in the computation of V_i. The recurrence is carried by the `I`-loop and appears to prevent distribution.

Figure 16 shows the sections of the array `A(:,:)` accessed for the entire execution of the `I`-loop. If the sections accessed by `A(J,I)` and `A(K,J)` are examined, a legal partial distribution of the `I`-loop is revealed (note the similarity to the LU and Cholesky factorization). The section accessed by `A(J,I)` (the black region) is a subset of the section accessed by `A(K,J)` (both the black and gray regions), and the index-set of `J` can be split at the point `J = -I + IB - 1` to create a new loop that executes over the iteration space where the memory locations accessed by `A(K,J)` are disjoint from those accessed by `A(J,I)`. The new loop that iterates over the disjoint region can be further optimized by the compiler, depending upon the target architecture.

3.3.3 *A Comparison of the Two QR Factorizations.* The algorithm *cd*-QR does not exhibit as much cache reuse as the LAPACK version on large matrices. The reason is that the LAPACK algorithm is able to take advantage of the level 3 BLAS routine `DGEMM`, which can be highly opti-

Table I. Summary of the Compiler Transformations Necessary to Block Matrix Factorizations

1 Dependence Analysis (Section 2.1 [Goff et al. 1991; Kuck 1978])
2 Array Section Analysis (Section 2.1 [Callahan and Kennedy 1987; Havlak and Kennedy 1991])
3 Strip-Mine-and-Interchange (Section 3.1 [Wolfe 1987; Wolfe and Lam 1991])
4 Unroll-and-Jam (Section 3.1 [Carr and Kennedy 1994])
5 Index-Set Splitting (Section 3.1)
6 Loop Distribution (Section 3.1 [Wolfe 1986])
7 Handling of Nonrectangular Iteration Spaces (Section 3.1 [Carr and Kennedy 1992; Wolfe 1987])
8 Automatic Block-Size Selection [Coleman and McKinley 1995; Lam et al. 1991]
9 Pattern Matching for Pivoting (Section 3.1.2)

mized. *Cd*-QR uses operations that are closer to the level 2 BLAS and that have worse cache reuse characteristics. Therefore, we would expect the LAPACK algorithm to perform better on larger matrices, as it could possibly take advantage of a highly tuned matrix-matrix multiply kernel.

3.4 Summary of Transformations

In summary, Table I lists the analyses and transformations that must be used by a compiler to block matrix factorizations. Items 1 and 2 are discussed in Section 2. Items 3 through 7 are discussed in Section 3.1. Item 8 is discussed in the compiler literature [Coleman and McKinley 1995; Lam et al. 1991]. Item 9 is discussed in Section 3.1.2. Many commercial compilers (e.g., IBM [Sarkar 1997], HP, DEC, and SGI) contain items 1, 3, 4, 5, 6, 7, and 8. However, it should be noted that items 2 and 9 are not likely to be found in any of today's commercial compilers.

4. EXPERIMENT

We measured the performance of each block factorization algorithm on four different architectures: the IBM POWER 2 model 590, the HP model 712/80, the DEC Alpha 21164, and the SGI model Indigo2 with a MIPS R4400. Table II summarizes the characteristics of each machine. These architectures were chosen because they are representative of the typical high-performance workstation.

On all the machines, we used the vendor's optimized BLAS. For example, on the IBM Power2 and SGI Indigo2, we linked with the libraries `-lessl` (Engineering and Scientific Subroutine Library [IBM 1994]) and `-lblas`, respectively.

Our compiler-optimized versions were obtained by hand using the algorithms in the literature. The reason that this process could not be fully automated is because of a current deficiency in the dependence analyzer of our tool [Callahan et al. 1987; Carr 1992]. Table III lists the Fortran compiler and the flags used to compile our factorizations.

In Tables IV–IX, performance is reported in double-precision megaflops. The number of floating-point operations for the LU, QR, and Cholesky

Table II. Machine Characteristics

Machine	Clock Speed (MHz)	Peak MFLOPS	Cache Size (KB)	Associativity	Line Size (Bytes)
IBM POWER2	66.5	264	256	4	256
HP 712	80	80	256	1	32
DEC Alpha	250	500	8	1	32
SGI Indigo2	200	100	16	1	32

Table III. Fortran Compiler and Switches

Machine	Compiler	Flags
IBM POWER2	xlf AIX v1.3.0.24	-O3
HP 712	f77 v9.16	-O
DEC Alpha	f77 v3.8	-O5
SGI Indigo2	f77 v5.3	-O3 -mips2

Table IV. LU Performance on IBM and HP

	IBM POWER2					HP 712				
Size	LABlk	LAMf	CBlk	CMf	Speedup	LABlk	LAMf	CBlk	CMf	Speedup
25×25	1,16,32,64	21	8,16	44	2.10	1	21	8	21	1.00
50×50	32	48	8,16	74	1.54	1	33	8	28	0.83
75×75	16	81	16	95	1.17	1	26	8	31	1.17
100×100	16	106	16	112	1.06	1	25	8	31	1.23
150×150	16	132	16	132	1.00	64	21	16	31	1.49
200×200	32	143	16	138	0.965	64	20	16	33	1.63
300×300	32	157	32	147	0.936	32	18	32	36	2.03
500×500	64	166	32	161	0.970	32	17	32	40	2.28

Table V. LU Performance on DEC and SGI

	DEC Alpha					SGI Indigo				
Size	LABlk	LAMf	CBlk	CMf	Speedup	LABlk	LAMf	CBlk	CMf	Speedup
25×25	1	43	8	53	1.25	8	20	8	21	1.05
50×50	8	74	8	78	1.05	8	34	8	28	0.824
75×75	16	96	8	96	1.00	8	34	8	29	0.853
100×100	16	116	8	110	0.95	8	37	8	29	0.784
150×150	32	138	8	113	0.82	8	39	8	28	0.718
200×200	32	156	8	124	0.79	8	40	16	29	0.725
300×300	32	181	16	132	0.73	8	41	16	30	0.732
500×500	32	212	8	148	0.70	38	32	32	29	0.763

factorizations are $2/3n^3$, $2mn^2 - 2/3n^3$, and $1/3n^3$, respectively, where m and n are the number of rows and columns, respectively. We used the LAPACK subroutines `dgetrf`, `dgeqrf`, and `dpotrf` for the LU, QR, and Cholesky factorizations, respectively. Each factorization routine is run

with block sizes of 1, 2, 4, 8, 16, 24, 32, 48, and 64.[1] In each table, the columns should be interpreted as follows:

—**LABlk**: The best blocking factor for the algorithm.

—**LAMf**: The best megaflop rate for the algorithm (corresponding to LABlk).

—**CBlk**: The best blocking factor for the compiler-derived algorithm.

—**CMf**: The best megaflop rate for the compiler-derived algorithm (corresponding to CBlk).

In order to explicitly set the block size for the LAPACK factorizations, we have modified the LAPACK integer function `ILAENV` to include a common block.

All the benchmarks were run when the computer systems were free of other computationally intensive jobs. All the benchmarks were typically run two or more times. The differences in time were within 5%.

4.1 LU Factorization

Tables IV and V show the performance of the compiler-derived version of LU factorization with pivoting versus the LAPACK version.

The IBM POWER2 results (Table IV) show that as the size of the matrix increases to 100, the compiler-derived algorithm's edge over LAPACK diminishes. And for the remaining matrix sizes, the compiler-derived algorithm stays within 7% of the LAPACK one. Clearly, the Fortran compiler on the IBM Power2 is able to nearly achieve the performance of the hand-optimized BLAS available in the ESSL library for the block matrix factorizations.

For the HP 712, Table IV indicates an unexpected trend. The compiler-derived version performs better on all matrix sizes except 50 by 50, with dramatic improvements as the matrix size increases. This indicates that the hand-optimized `DGEMM` does not efficiently use the cache. We have optimized for cache performance in our compiler-derived algorithm. This is evident when the size of the matrices exceeds the size of the cache.

The significant performance degradation for the 50-by-50 case is interesting. For a matrix this small, cache performance is not a factor. We believe the performance difference comes from the way code is generated. For superscalar architectures like the HP, a code generation scheme called software pipelining is used to generate highly parallel code [Lam 1988; Rau et al. 1992]. However, software pipelining requires a lot of registers to be successful. In our code, we performed unroll-and-jam to improve cache performance. However, unroll-and-jam can significantly increase register pressure and cause software pipelining to fail [Carr et al. 1996]. On our version of LU decomposition, the HP compiler diagnostics reveal that

[1]Although the compiler can effectively choose blocking factors automatically, we do not have an implementation of the available algorithms [Coleman and McKinley 1995; Lam et al. 1991].

software pipelining failed on the main computational loop due to high register pressure. Given that the hand-optimized version is highly software pipelined, the result would be a highly parallel hand-optimized loop and a not-as-parallel compiler-derived loop. At matrix size 25 by 25, there are not enough loop iterations to expose the difference. At matrix size 50 by 50, the difference is significant. At matrix sizes 75 by 75 and greater, cache performance becomes a factor. At this time, there are no known compiler algorithms that deal with the trade-offs between unroll-and-jam and software pipelining. This is an important area of future research.

For the DEC Alpha, Table V shows that our algorithm performs as well as or better than the LAPACK version on matrices of order 100 or less. After size 100 by 100, the second-level cache on the Alpha, which is 96K, begins to overflow. Our compiler-derived version is not blocked for multiple levels of cache, while the LAPACK version is blocked for two levels of cache [Kamath et al. 1994]. Thus, the compiler-derived algorithm suffers many more cache misses in the level 2 cache than the LAPACK version. It is possible for the compiler to perform the extra blocking for multiple levels of cache, but we know of no compiler that currently does this. Additionally, the BLAS algorithm utilized the following architectural features that we do not [Kamath et al. 1994]:

—the use of temporary arrays to eliminate conflicts in the level 1 direct-mapped cache and the translation lookaside buffer [Coleman and McKinley 1995; Lam et al. 1991] and

—the use of the memory-prefetch feature on the Alpha to hide latency between cache and memory.

Although each of these optimizations could be done in the DEC product compiler, they are not. Each optimization would give additional performance to our algorithm. Using a temporary buffer may provide a small improvement, but prefetching can provide a significant performance improvement because the latency to main memory is on the order of 50 cycles. Prefetches cannot be issued in the source code, so we were unable to try this optimization.

The results on the SGI (Table V) are roughly similar to those for the DEC Alpha. It is difficult for us to determine exactly why our performance is lower on smaller matrices, because we have no diagnostic tools. It could again be software pipelining or some architectural feature of which we are not aware. We do note that the code generated by the SGI compiler is worse than expected. Additionally, the two-level cache comes into play on the larger matrices.

Comparing the results on the IBM POWER2 and the multilevel cache hierarchy systems (DEC and SGI), shows that our compiler-derived versions are very effective for a single-level cache. It is evident that more work needs to be done in optimizing the update portion of the factorizations to obtain the same relative performance as a single-level cache system on a multilevel cache system.

Table VI. Cholesky Performance on IBM and HP

	IBM POWER2					HP 712				
Size	LABlk	LAMf	CBlk	CMf	Speedup	LABlk	LAMf	CBlk	CMf	Speedup
25×25	1,32,64	24	4	45	1.86	1	10	8	21	2.00
50×50	32	53	16	81	1.53	1	42	8	38	0.67
75×75	32	81	8	114	1.41	1	37	8	31	0.83
100×100	32	105	8	132	1.26	1	33	8	33	1.00
150×150	32	136	16	151	1.11	1	32	16	34	1.05
200×200	32	154	8	158	1.03	1	33	16	36	1.11
300×300	32	185	32	170	0.920	1	23	16	39	1.72
500×500	64	205	32	191	0.932	32	17	16	43	2.56

Table VII. Cholesky Performance on DEC and SGI

	DEC Alpha					SGI Indigo2				
Size	LABlk	LAMf	CBlk	CMf	Speedup	LABlk	LAMf	CBlk	CMf	Speedup
25×25	1	36	4	53	1.50	1	19	4	23	1.21
50×50	1	71	4	107	1.50	1	31	4	32	1.03
75×75	1	94	4	117	1.25	1	33	4	34	1.03
100×100	1	104	4	131	1.27	8	33	4	34	1.03
150×150	1	113	4	141	1.24	16	36	4	34	0.94
200×200	1	116	4	145	1.25	16	38	4	34	0.895
300×300	64	134	4	146	1.09	16	40	4	34	0.850
500×500	64	162	4	149	0.92	16	40	4	32	0.800

4.2 Cholesky Factorization

Tables VI and VII show the performance of the compiler-derived version of Cholesky factorization versus the LAPACK version.

The IBM POWER2 results (Table VI) show that as the size of the matrix increases to 200, the compiler-derived algorithm's edge over the LAPACK diminishes. And for the remaining matrix sizes, the compiler-derived algorithm stays within 8% of the LAPACK one. As was the case for the LU factorization, the compiler version performs very well. Only for the large matrix sizes does the highly tuned BLAS used by the LAPACK factorization cause LAPACK to be faster. Table VI shows a slightly irregular pattern for the block size used by the compiler-derived algorithm. We remark that for matrix sizes 50 through 200, the MFLOP rate for the two block sizes 8 and 16 were nearly equivalent.

On the HP (Table VI), we observe the same pattern as we did for LU factorization. When cache performance is critical, we outperform the LAPACK version. When cache performance is not critical, the LAPACK version gives better results, except when the matrix is small. Our algorithm performed much better on the 25-by-25 matrix most likely due to the high overhead associated with software pipelining on short loops. Since Cholesky factorization has fewer operations than LU factorization in the update portion of the code, we would expect a high overhead associated with small matrices. Also, the effects of cache are not seen until larger

Table VIII. QR Performance on IBM and HP

	IBM POWER2					HP 712				
Size	LABlk	LAMf	CBlk	CMf	Speedup	LABlk	LAMf	CBlk	CMf	Speedup
25×25	32,64	30	8	52	1.73	1	21	1	21	1.00
50×50	64	60	8	84	1.4	1	37	2	28	0.75
75×75	1	81	4	97	1.20	1	38	4	29	0.76
100×100	8	101	4	108	1.07	1	38	4	30	0.80
150×150	8	127	2	116	0.913	1	28	8	29	1.07
200×200	8	144	8,16	118	0.819	64	25	16	31	1.23
300×300	16	164	16	121	0.738	32	25	32	31	1.25
500×500	16,32	183	32	123	0.676	32	23	32	31	1.38

matrix sizes (compared to LU factorization). This is again due to the smaller update portion of the factorization.

On the DEC (Table VII) we outperform the LAPACK version up until the 500×500 matrix. This is the same pattern as seen in LU factorization except that it takes longer to appear. This is due to the smaller size of the update portion of the factorization.

The results on the SGI show that the compiler-derived version performs better than the LAPACK for matrix sizes up to 100. As the matrix size increases to 500 from 150, the compiler-derived algorithm's performance decreases by 15% compared to that of the LAPACK factorization. We believe that this has to do with the two-level cache hierarchy.

We finally remark that although Tables VI and VII show a similar pattern as Tables IV and V, there are differences. Recall, that as explained in Section 3.2, the Cholesky factorization only has approximately half of the the floating-point operations of LU, because it neglects the strict (above the diagonal) upper triangular portion of the matrix during the update phase. Moreover, there is the computation of the square root of the diagonal element during each of the n iterations.

4.3 QR Factorization

Tables VII and IX show the performance of the compiler-derived version of QR factorization versus the LAPACK version. Since the compiler-derived algorithm for block QR factorization has worse cache performance than the LAPACK algorithm, but $O(n^2)$ less computation, we expect worse performance when the cache performance becomes critical. In plain words, the LAPACK algorithm uses the level 3 BLAS matrix multiply kernel `DGEMM`, but the compiler-derived algorithm can only utilize operations similar to the level 2 BLAS.

On the HP (Table VIII), we see the same pattern as before. However, because the cache performance of our algorithm is not as good as the LAPACK version, we see a much smaller improvement when our algorithm has superior performance. Again, we also see that when the matrix sizes stay within the limits of the cache, LAPACK outperforms our algorithm.

Table IX. QR Performance on DEC and SGI

	DEC Alpha					SGI Indigo2				
Size	LABlk	LAMf	CBlk	CMf	Speedup	LABlk	LAMf	CBlk	CMf	Speedup
25×25	1	50	4	66	1.31	1	15	8	23	1.53
50×50	1	85	2	98	1.15	4	26	8	30	1.15
75×75	1	100	2	107	1.07	8	29	8	29	1.00
100×100	16	114	4	111	0.98	8	34	8	29	0.853
150×150	16	138	8	110	0.79	8	38	8	28	0.737
200×200	16	158	16	115	0.72	8	39	8	27	0.692
300×300	16	180	16	114	0.64	8	40	8	25	0.625
500×500	32	213	16	115	0.54	8	39	8	25	0.641

For the IBM Power2 (Table VIII) and DEC and SGI machines (Table IX), we see the same pattern as on the previous factorizations, except that our degradations are much larger for large matrices. Again, this is due to the inferior cache performance of *cd*-QR. An interesting trend revealed by Table VIII is that the IBM POWER2 has a slightly irregular block size pattern as the matrix size increases. We remark that only for matrix sizes less than or equal to 75, is there interesting behavior. For the first two matrix sizes, the optimal block size is larger than the dimension of the matrix. This implies that no blocking was performed; the level 3 BLAS was not used by the LAPACK algorithm. For the matrix size 75, the rate achieved by the LAPACK algorithm with block size 8 was within 4–6% of the unblocked factorization.

5. RELATED WORK

We briefly review and summarize other investigations parallel to ours. It is evident that there is an active amount of work to remove the substantial hand coding associated with efficient dense linear algebra computations.

5.1 Blocking with a GEMM-Based Approach

Since LAPACK depends upon a set of highly tuned BLAS for efficiency, there remains the practical question of how they should be optimized. As discussed in the introduction, an efficient set of BLAS requires a nontrivial effort in software engineering. See Kägström et al. [1995a] for a discussion on software efforts to provide optimal implementations of the level 3 BLAS.

An approach that is both efficient and practical is the GEMM-based one proposed by Kägström et al. [1995a] in a recent study. Their approach advocates optimizing the general matrix multiply and add kernel `_GEMM` and then rewriting the remainder of the level 3 BLAS in terms of calls to this kernel. The benefit of their approach is that only this kernel needs to be optimized—whether by hand or the compiler. Their thorough analysis highlights the many issues that must be considered when attempting to construct a set of highly tuned BLAS. Moreover, they provide high-quality implementations of the BLAS for general use as well as a performance evaluation benchmark [Kägström et al. 1995b].

We emphasize that our study examines only whether the necessary optimizations may be left to the compiler and whether they should be applied directly to the matrix factorizations themselves. What is beyond the ability of the compiler is that of recasting the level 3 BLAS in terms of calls to _GEMM

5.2 PHiPAC

Another recent approach is the methodology expressed for developing a Portable High-Performance matrix-vector libaries in ANSI C (PHiPAC) [Bilmes et al. 1996]. The project is motivated by many of the reasons as outlined in our introduction. However, there is a major difference in approaches that does not make it a parallel study. As in the GEMM-based approach, they seek to support the BLAS and aim to be more efficient than the vendor-supplied BLAS. However, unlike our study or the GEMM one, PHiPAC assumes that ANSI C is the programming language. Because of various C semantics PHiPAC instead seeks to provide parameterized generators that produce the optimized code. See Bilmes et al. [1996] for a discussion on the inhibitors in C that prevent an optimizing compiler from generating efficient code.

5.3 Autoblocking Matrix Multiplication

Frens and Wise [1997] present an alternative algorithm for matrix-matrix multiply that is based upon a quadtree representation of matrices. Their solution is recursive and suffers from the lack of interprocedural optimization in most commercial compilers. Their results show that when paging becomes a problem on SGI multiprocessor systems, the quadtree algorithm has superior performance to the BLAS 3. On the smaller problems, the quadtree algorithm has inferior performance. In relation to our work, we could not expect the compiler to replace the BLAS 3 with the quadtree approach when appropriate, as it is a change in algorithm rather than a reshaping. In addition, the specialized storage layout used by Frens and Wise calls into question the effect on an entire program.

6. SUMMARY

We have set out to determine whether a compiler can automatically restructure matrix factorizations well enough to avoid the need for hand optimization. To that end, we have examined a collection of implementations from LAPACK. For each of these programs, we determined whether a plausible compiler technology could succeed in obtaining the block version from the point algorithm.

The results of this article are encouraging: we have demonstrated that there exist implementable compiler methods that can automatically block matrix factorization codes to achieve algorithms that are competitive with those of LAPACK. Our results show that for modest-sized matrices on advanced microprocessors, the compiler-derived variants are often superior. These matrix sizes are typical on workstations.

Given that future machine designs are certain to have increasingly complex memory hierarchies, compilers will need to adopt increasingly sophisticated memory-management strategies so that programmers can remain free to concentrate on program logic. Given the potential for performance attainable with automatic techniques, we believe that it is possible for the user to express machine-independent point matrix factorization algorithms without the BLAS and still get good performance if compilers adopt our enhancements to already existing methods.

ACKNOWLEDGMENTS

Ken Kennedy and Richard Hanson provided the original motivation for this work. Ken Kennedy, Keith Cooper, and Danny Sorensen provided financial support for this research when it was begun at Rice University. We also wish to thank Tomas Lofgren and John Pieper of DEC for their help. We also thank Per Ling of the University of Umeå and Ken Stanley of the University of California Berkeley for their help with the benchmarks and discussions.

REFERENCES

ALLEN, R. AND KENNEDY, K. 1987. Automatic translation of FORTRAN programs to vector form. *ACM Trans. Program. Lang. Syst. 9*, 4 (Oct.), 491–542.

ANDERSON, E., BAI, Z., BISCHOF, C. H., DEMMEL, J., DONGARRA, J. J., DU CROZ, J., GREENBAUM, A., HAMMARLING, S., MCKENNEY, A., OSTROUCHOV, S., AND SORENSEN, D. C. 1995. *LAPACK User's Guide*. 2nd ed. Society for Information Management and The Management Information Systems, Minneapolis, MN.

BILMES, J., ASANOVIĆ, K., DEMMEL, J., LAM, D., AND CHIN, C. W. 1996. PHiPAC: A portable, high-performance, ANSI C coding methodology and its application to matrix multiply. LAPACK Working Note 111, University of Tennessee, Knoxville, TN.

CALLAHAN, D. AND KENNEDY, K. 1988. Analysis of interprocedural side effects in a parallel programming environment. In *Proceedings of the 1st International Conference on Supercomputing* (Athens, Greece, June 1987), E. N. Houstis, T. S. Papatheodorou, and C. D. Polychronopoulos, Eds. Lecture Notes in Computer Science, vol. 297. Springer-Verlag New York, Inc., New York, NY, 138–171.

CALLAHAN, K., COOPER, R., KENNEDY, K., AND TORCZON, L. 1987. ParaScope: A parallel programming environment. In *Proceedings of the 1st International Conference on Supercomputing* (Athens, Greece, June 8-12, 1987). ACM Press, New York, NY.

CARR, S. M. 1992. Memory-hierarchy management. Ph.D. thesis, Rice University, Houston, TX.

CARR, S. AND KENNEDY, K. 1992. Compiler blockability of numerical algorithms. In *Proceedings of Supercomputing '92* (Minneapolis, MN, Nov. 16–20, 1992), R. Werner, Ed. IEEE Computer Society Press, Los Alamitos, CA, 114–124.

CARR, S. AND KENNEDY, K. 1994. Improving the ratio of memory operations to floating-point operations in loops. *ACM Trans. Program. Lang. Syst. 16*, 6 (Nov.), 1768–1810.

CARR, S., DING, C., AND SWEANY, P. 1996. Improving software pipelining with unroll-and-jam. In *Proceedings of the 29th Annual Hawaii International Conference on System Sciences* (Maui, Hawaii, Jan. 1996). IEEE Computer Society Press, Los Alamitos, CA.

COLEMAN, S. AND MCKINLEY, K. S. 1995. Tile size selection using cache organization and data layout. *SIGPLAN Not. 30*, 6 (June), 279–290.

DONGARRA, J. J., DU CROZ, J., HAMMARLING, S., AND DUFF, I. 1990. A set of level 3 Basic Linear Algebra Subprograms. *ACM Trans. Math. Softw. 16*, 1 (Mar.), 1–17.

DONGARRA, J. J., DU CROZ, J., HAMMARLING, S., AND HANSON, R. J. 1988. An extended set of FORTRAN Basic Linear Algebra Subprograms. *ACM Trans. Math. Softw. 14*, 1 (Mar.), 1–17.

DONGARRA, J. J., DUFF, I. S., SORENSEN, D. C., AND VAN DER HORST, H. A. 1991. *Solving Linear Systems on Vector and Shared Memory Computers*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

DONGARRA, J. J., GUSTAVSON, F. G., AND KARP, A. 1984. Implementing linear algebra algorithms for dense matrices on a vector pipeline machine. *SIAM Rev. 26*, 1 (Jan.), 91–112.

DONGARRA, J. J., MOLER, C. B., BUNCH, J. R., AND STEWART, G. W. 1979. *LINPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

FRENS, J. AND WISE, D. S. 1997. Auto-blocking matrix multiplication, or tracking BLAS3 performance from source code. *SIGPLAN Not. 32*, 7 (July), 206–216.

GALLIVAN, K. A., PLEMMONS, R. J., AND SAMEH, A. H. 1990. Parallel algorithms for dense linear algebra computations. *SIAM Rev. 32*, 1 (Mar.), 54–135.

GOFF, G., KENNEDY, K., AND TSENG, C.-W. 1991. Practical dependence testing. *SIGPLAN Not. 26*, 6 (June), 15–29.

GOLUB, G. H. AND VAN LOAN, C. F. 1996. *Matrix Computations*. 3rd ed. Johns Hopkins University Press, Baltimore, MD.

HAVLAK, P. AND KENNEDY, K. 1991. An implementation of interprocedural bounded regular section analysis. *IEEE Trans. Parallel Distrib. Syst. 2*, 3 (July), 350–360.

IBM. 1994. *Engineering and Scientific Subroutine Library Version 2 Release 2, Guide and Reference*. IBM Corp., Riverton, NJ.

KÅGSTRÖM, B., LING, P., AND VAN LOAN, C. 1995a. GEMM-based level 3 BLAS: High-performance model implementations and performance evaluation benchmark. Tech. Rep. UMINF-95.18, Dept. of Computing Science, University of Umeå, Umeå, Sweden. Also available as LAPACK Working Note 107.

KÅGSTRÖM, B., LING, P., AND VAN LOAN, C. 1995b. GEMM-based level 3 BLAS: Installation, tuning, and use of the model implementations and the performance evaluation benchmark. Tech. Rep. UMINF 95.19, Dept. of Computing Science, University of Umeå, Umeå, Sweden. Also available as LAPACK Working Note 108.

KAMATH, C., HO, R., AND MANLEY, D. P. 1994. DXML: A high-performance scientific subroutine library. *Digital Tech. J. 6*, 3 (Summer), 44–56.

KUCK, D. 1978. *The Structure of Computers and Computations*. Vol. 1. John Wiley & Sons, Inc., New York, NY.

LAM, M. 1988. Software pipelining: An effective scheduling technique for VLIW machines. In *Proceedings of the ACM SIGPLAN '88 Conference on Programming Language Design and Implementation* (Atlanta, Georgia, June 22-24, 1988), R. L. Wexelblat, Ed. ACM Press, New York, NY, 318–328.

LAM, M. D., ROTHBERG, E. E., AND WOLF, M. E. 1991. The cache performance and optimizations of blocked algorithms. *SIGARCH Comput. Archit. News 19*, 2 (Apr.), 63–74.

LAWSON, C. L., HANSON, R. J., KINCAID, D. R., AND KROGH, F. T. 1979. Basic linear algebra subprograms for Fortran usage. *ACM Trans. Math. Softw. 5*, 3, 308–323.

LEHOUCQ, R. 1992. Implementing efficient and portable dense matrix factorizations. In *Proceedings of the 5th SIAM Conference on Parallel Processing for Scientific Computing*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

LEVINE, D., CALLAHAN, D., AND DONGARRA, J. 1991. A comparative study of automatic vectorizing compilers. *Parallel Comput. 17*, 1223–1244.

ORTEGA, J. M. 1988. *Introduction to Parallel and Vector Solution of Linear Systems*. Plenum Press, New York, NY.

RAU, B. R., LEE, M., TIRUMALAI, P. P., AND SCHLANSKER, M. S. 1992. Register allocation for software pipelined loops. *SIGPLAN Not. 27*, 7 (July), 283–299.

SARKAR, V. 1997. Automatic selection of high order transformations in the IBM XL Fortran compilers. *IBM J. Res. Dev. 41*, 3 (May).

SMITH, B. T., BOYLE, J. M., DONGARRA, J. J., GARBOW, B. S., IKEBE, Y., KLEMA, V. C., AND MOLER, C. B. 1976. *EISPACK Guide*. 2nd ed. Springer Lecture Notes in Computer Science, vol. 6. Springer-Verlag New York, Inc., New York, NY.

WOLFE, M. 1986. Advanced loop interchange. In *Proceedings of the 1986 International Conference on Parallel Processing*.

WOLFE, M. 1987. Iteration space titling for memory hiearchies. In *Proceedings of the 3rd SIAM Conference on Parallel Processing for Scientific Computing*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

WOLF, M. E. AND LAM, M. S. 1991. A data locality optimizing algorithm. *SIGPLAN Not. 26*, 6 (June), 30–44.

Received: August 1995; revised: October 1996 and February 1997; accepted: February 1997