Access Normalization: Loop Restructuring for NUMA Computers

WEI LI and KESHAV PINGALI
Cornell University

In scalable parallel machines, processors can make local memory accesses much faster than they can make remote memory accesses. Additionally, when a number of remote accesses must be made, it is usually more efficient to use block transfers of data rather than to use many small messages. To run well on such machines, software must exploit these features. We believe it is too onerous for a programmer to do this by hand, so we have been exploring the use of restructuring compiler technology for this purpose. In this article, we start with a language like HPF-Fortran with user-specified data distribution and develop a systematic loop transformation strategy called *access normalization* that restructures loop nests to exploit locality and block transfers. We demonstrate the power of our techniques using routines from the BLAS (Basic Linear Algebra Subprograms) library. An important feature of our approach is that we model loop transformations using *invertible* matrices and integer lattice theory.

Categories and Subject Descriptors: C.1.2 [**Processor Architectures**]: Multiple Data Stream Architectures (Multiprocessors)—*Multiple-instruction-stream multiple-data-stream processors (MIMD)*, *NUMA*, D.1.3 [**Programming Techniques**]: Concurrent Programming—*parallel programming*; D.3.4 [**Programming Languages**]: Processors—*compilers*; *optimization*; *code generation*

General Terms: Algorithms, Experimentation, Languages, Performance

Additional Key Words and Phrases: Communication, data locality, loop transformation, nonsingular loop transformation, nonuniform memory access machines, parallelizing compilers

1. INTRODUCTION

Scalable parallel machines are usually organized as networks of processor-memory pairs in which a processor can access local data much faster than it can access remote data. For example, in the BBN Butterfly, accesses to local memory take 0.6 microseconds while accesses to remote memory take about 6.6 microseconds [BBN 1989]. Distributed-memory machines like the Intel iPSC/i860 have even greater nonuniformity in memory access times because access to remote data must be orchestrated through the exchange of mes-

This research was supported by an NSF Presidential Young Investigator award CCR-8958543, NSF grant CCR-9008526, ONR grant N00014-93-1-0103, and by grants from the Hewlett-Packard Corporation and the Cornell Theory Center.
Author's address: Department of Computer Science, Cornell University, Ithaca, NY 14853.

sages. If nonlocal accesses are on the critical path through a program, making these accesses local through proper data management will speed up program execution.

A second feature of such architectures is that block transfer of data between processors is more efficient than sending this data using many small messages. Data transfer between processors can be viewed as a pipeline with a large setup time compared to the time per stage. For example, on the Intel iPSC/i860, it takes 70 microseconds to start up communication, but it takes only 1 microsecond to transfer a double-precision floating-point number between nearest neighbors once the communication has been set up. Therefore, when a number of data items must be sent from one processor to another, it is preferable to use a single long message to amortize startup time.

Contention in the network has the effect of increasing the expected latency of nonlocal references; therefore, data management to avoid nonlocal references has the added benefit of reducing contention, thereby improving performance. Interestingly, some analytical studies show that long messages can increase the latency of nonlocal accesses [Agarwal 1991]. This is an argument against long messages, but on current machines, this effect seems to be of secondary importance compared to the benefits to amortizing startup time, as we show in Section 8.

For the software writer, the implication of these features of nonuniform memory access (NUMA)[1] machines is that programs must not only exploit parallelism but must also manage data to eliminate nonlocal references wherever possible; where nonlocal references are necessary, they should be grouped together for block transfers. We believe that it is too onerous for the programmer to accomplish this by hand, so we have been exploring the use of restructuring compilers for this purpose. Existing compiler technology is oriented mostly toward *uniform* memory access machines in which the only concern is exploitation of parallelism. Parallel code is generated by distributing iterations of the outermost loop in a loop nest among the processors, with synchronization instructions being inserted to take care of dependences carried by this loop. To reduce synchronization, transformations like loop interchange are carried out to move parallel loops outermost wherever possible [Allen and Kennedy 1987; Banerjee 1988; Midkiff and Padua 1987; Wolf and Lam 1991b]. This approach does not perform any data management, so it is not suitable for generating good code on NUMA architectures.

An alternative approach, implemented by the Id Nouveau [Rogers and Pingali 1989] and Fortran-D systems [Hiranandani et al. 1991], among others, is to give the programmer control over how data structures are distributed across the processors. The compiler uses this *data decomposition* information to determine how to assign work to processors. One simple way to do this is to use the so-called *ownership* rule—a processor executes an assignment statement if the left-hand-side variable of the statement is

[1] We use this term in a broad sense to include distributed-memory machines.

mapped to the local memory of that processor. A processor executes a loop iteration if it has any work to do in the body for that iteration. Although this strategy takes data mappings into account, it can generate inefficient code, in which all processors execute all iterations "looking for work to do" if the structure of the loop nest does not match the data distribution [Zima and Chapman 1990]. In many of these cases, loop restructuring can improve code quality, but no general approach to loop transformation has been available in this context [Hiranandani et al. 1991].

In this article, we present a systematic approach to loop restructuring for parallel machines with a memory hierarchy. As in the ownership approach, our starting point is a language like HPF-Fortran with user-specified data decomposition. However, rather than use this information directly to generate code, we use the data distribution information to drive *access normalization*, which transforms loop nests so that code can be generated by distributing iterations of the outermost loop among the processors without compromising locality. The structure of inner loops is chosen so that data can be transferred using block transfers wherever possible.

Our work makes two contributions:

—We describe a new loop transformation strategy called *access normalization* that is useful for compiling programs for parallel machines with nonuniform memory access. It has applications in other areas such as the generation of vector code.

—Our loop transformations are expressed in the framework of *invertible* matrices and integer lattice theory, which is an important generalization of existing frameworks that use unimodular matrices.

The rest of the article is organized as follows. In Section 2, we discuss a simple example that gives an overview of our compiling strategy. We also introduce the *data access matrix*, which plays a key role in the development. In Section 3, we discuss the framework of *invertible* matrices as a foundation for loop transformations. For some programs, the data access matrix is invertible and can be used directly to transform the loop nest, as we show in Section 4. In general, however, this matrix may not be invertible, and the techniques of Section 5 must be used to produce an invertible matrix for the transformation that respects program dependences; this is done in Section 6. In Section 7, we discuss how code can be generated after loops have been restructured according to our methods. We present experimental results in Section 8 that demonstrate that our methods work well on programs of practical interest such as routines from the BLAS (Basic Linear Algebra Subroutines) library [Coleman and van Loan 1988]. Finally, we discuss related work in Section 9.

2. OVERVIEW OF NUMA COMPILATION

In this section, we give an overview of our compilation strategy for NUMA architectures. We also introduce a key data structure called the *data access matrix*.

2.1 NUMA Compilation

Our compiler accepts programs written in Fortran-77 extended with data distribution declarations that specify how arrays are to be distributed across the local memories of the machine. We support most of the data distributions commonly used by programmers of NUMA machines, such as *wrapped* and *blocked* column and row distributions. In a wrapped column distribution, the columns of an array are distributed in a round-robin manner to the processors: if P is the number of processors, then processor 0 gets columns 0, P, $2P$ and so on, while processor 1 gets columns 1, $P + 1$, $2P + 1$, etc. Most of the examples in this article use a *wrapped column distribution*. Blocked column distribution is similar, except that a processor gets a contiguous set of columns. We also support so-called 2D blocks in which rectangular *subblocks* of the array are distributed to the processors [Hiranandani et al. 1991], but for lack of space, we will not consider them any further.

Data distributions can be specified precisely using a distribution function.

Definition 2.1 A *distribution function* is a function from array indices to integers between 0 and P-1, where P is the number of processors in the machine. An array dimension is a *distribution dimension*, if the dimension is used in the distribution function for the array.

For example, the distribution function for the wrapped column distribution of a 2D array is $W_2(i, j) = j \bmod P$, and the second dimension of the array is a distribution dimension.

To understand the need for loop restructuring, consider the program in Figure 1a, which is a simplified version of the SYR2K code discussed in Section 8. Assume that both A and B have a wrapped column distribution. Distributing iterations of the outer loop among the processors (Figure 1b) results in processor p executing iterations p, $p + P$, etc. Consider accesses to elements of array B. Each iteration of the outer loop makes $N_2(b - b/\mathrm{P})$ nonlocal accesses, and the total number of nonlocal accesses is $N_1 N_2 b(1 - 1/P)$.

The ownership rule uses data decomposition information to generate code. A processor is involved in the execution of an iteration (i, j, k) if it owns any of the elements referenced in the body of the loop in that iteration. Therefore,

```
for i = 0, N1 − 1
  for j = i, i + b − 1
    for k = 0, N2 − 1
      B[i,j − i] = B[i,j − i] + A[i,j + k]
```
(a)

```
for i = p, N1 − 1, step P
  for j = i, i + b − 1
    for k = 0, N2 − 1
      B[i,j − i] = B[i,j − i] + A[i,j + k]
```
(b)

```
for u = 0, b − 1
  for v = u, u + N1 + N2 − 2
    for w = 0, N1 − 1
      B[w,u] = B[w,u] + A[w,v]
```
(c)

```
for u = p, b − 1, step P
  for v = u, u + N1 + N2 − 2
    read A[ * ,v];
    for w = 0, N1 − 1
      B[w,u] = B[w,u] + A[w,v]
```
(d)

Fig. 1. Transformation and code generation for a simple example.

processor p has work to do in iteration (i, j, k) if $(j - i) \bmod P = p$ (it must update an element of B) or if $(j + k) \bmod P = p$ (it must send an element of A to whichever processor is updating B in that iteration). This is accomplished by placing these conditional tests in front of the statement and having all the processors execute all iterations "looking for work to do" [Callahan and Kennedy 1988; Rogers and Pingali 1989]. In simple programs, these conditional tests can be optimized away, but in general they must be executed at run-time, which is inefficient. Moreover, in our program, the code cannot make use of block transfers of elements of A since the elements of A referenced during one iteration of the j loop are mapped to different processors.

Now, consider the program of Figure 1c. This program computes the same function as Figure 1a, but if we distribute the outermost loop among the processors as before (Figure 1d), there are no nonlocal accesses to B. There are nonlocal accesses to A, but these can be performed using block transfers since the subscript in the distribution dimension of A is invariant in the innermost loop. The loop transformations described in this article transform the program of Figure 1a to that of Figure 1c. Given the transformed program, the code generation techniques described in Section 7 generate the parallel code shown in Figure 1d.

2.2 Data Access Matrix

Since the transformations are driven by the data access patterns, it is convenient to define a data structure to represent array subscripts in a loop nest in a convenient way. This data structure is called the *data access matrix*. It is used by our loop-restructuring system as the starting point for determining what transformations to apply to the loop nest. For the loop nest in Figure 1a, the data access matrix is

$$\begin{pmatrix} -1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

This matrix represents the subscripts in the sense that the product of the data access matrix with the column vector $[i, j, k]^T$ yields a column vector in which each element is a subscript from the program. For our example, this product is the column vector $[j - i, j + k, i]^T$ which corresponds to the three subscripts of the program. Constants in a subscript are omitted from the corresponding entry in the data access matrix.

The order in which these subscripts are represented in the data access matrix is important and corresponds to an estimate of their relative importance for achieving good performance. A reasonable heuristic is to give highest importance to subscripts in the distribution dimension(s) of arrays; in our example, the subscripts $j - i$ and $j + k$ dominate the subscript i since they occur in the distribution dimensions of arrays B and A. Notice that $j - i$ occurs twice, but $j + k$ occurs only once. Therefore, we let $j - i$ dominate $j + k$. This yields the data access matrix shown above.

The technical development in the rest of the article is independent of how subscripts were ordered to obtain the data access matrix. Additionally, a subscript that is "overly complex" for any reason (such as a nonlinear function of loop indices) may be omitted from the data access matrix without affecting correctness.

3. LOOP TRANSFORMATIONS AND INVERTIBLE MATRICES

In this section, we show how invertible matrices can be used to model the loop transformations of interest in the NUMA context. Consider a simple loop nest:

```
for i = 1, 3
  for j = 1, 3
    A[2i + 4j, i + 5j] = j;
```

It is to be restructured to the form

```
for u = 6, 18 step 2
  for v = u / 2 + max(3[(u − 6) / 4], 3),
          u / 2 + min(3[(u − 2) / 4], 9)
          step 3
    A[u, v] = (2v − u) / 6;
```

To determine how to perform the transformation, consider the iteration spaces of the two loops, shown in Figure 2. Since the bodies of both loops have the same statement, we must ensure that the work done in any iteration of the original loop nest is done in exactly one iteration of the new loop nest. Therefore, we must construct a one-to-one mapping from the old iteration space to the new one. Moreover, every iteration of the new loop nest must correspond to some point in the old iteration space, so the mapping must be an onto mapping. In other words, we must construct an invertible mapping between the two iteration spaces. One such mapping can be described concisely by the following set of equations, written in matrix form:

$$\begin{pmatrix} 2 & 4 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} i \\ j \end{pmatrix} = \begin{pmatrix} u \\ v \end{pmatrix}.$$

The mapping can be represented using an invertible integer matrix because it is a *linear*, *integral*, *invertible mapping* between the two iteration spaces.

The use of invertible matrices to model loop transformations is a generalization of the *unimodular* approach which is used to model loop interchange, skewing, and reversal [Banerjee 1990; Wolf and Lam 1991b]. Invertible matrices include unimodular matrices as a special case and permit us to model *loop scaling*, as well. An example of this transformation, which replaces a loop index with an integer multiple of the loop index, is shown below

```
for i = 1, 3          for u = 2, 6, 2
  A[2*i] = i            A[u] = u / 2
```

(a) original code (b) loop scaling

Loop scaling may introduce integer divisions, as is shown in the example, but these operations can be strength reduced and replaced with additions.

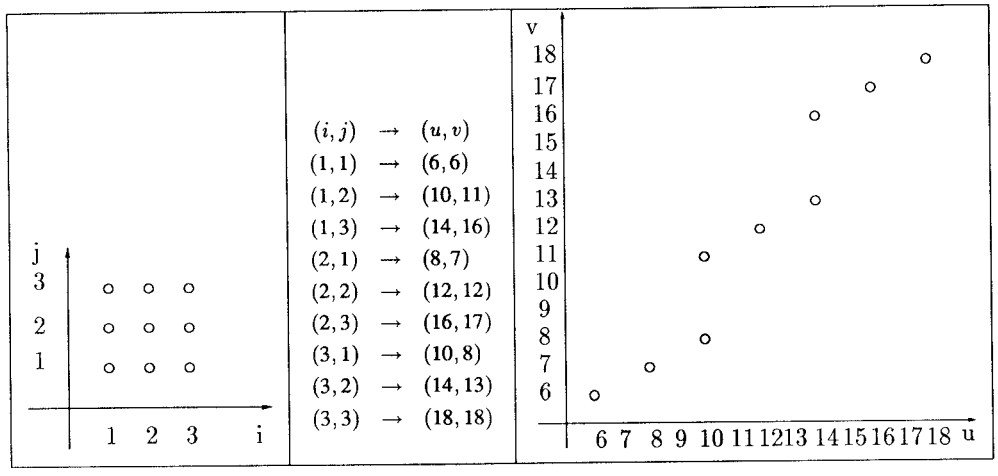


Fig. 2. Mapping between iteration spaces.

Like skewing or reversal, loop scaling is not particularly interesting in isolation, but combined with the other transformations, it lets us do wholesale loop restructing for NUMA architectures.

The algorithm for generating a restructured program starting from a loop nest and an invertible mapping is given in Li and Pingali [1992b]. This algorithm is nontrivial since the new loop nest must traverse points in the new iteration space in lexicographic order, and the starting point, ending point, and step size of a loop in the restructured loop nest can depend on only the loop indices of outer loops (for instance, the values for the outermost loop must be loop constant). It is not immediately obvious that this can be done for any inverticle matrix T. Fortunately, the iteration space of the loop nest forms what is called an *integer lattice*; by applying some results from integer lattice theory, we can easily construct the required loop nest.

Since invertible matrices are closed under matrix product, it follows that any sequence of these loop transformations (permutation, reversal, skewing, and scaling) can also be modeled as an invertible matrix. This means that the problem of performing the right sequence of loop transformations now reduces to that of finding an appropriate invertible matrix that models the desired sequence of transformations. We show how to do this next.

4. INVERTIBLE DATA ACCESS MATRICES

In this section, we consider the simple case where the data access matrix is invertible. Consider the program of Figure 1 again. The data access matrix for the program is X.

$$X = \begin{pmatrix} -1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

It is easy to verify that X is invertible; the result of transforming the source program using X as the transformation matrix was shown in Figure 1c.

Consider what happens when code is generated for the new loop nest by distribution iterations of the outermost loop among the processors in a round-robin manner. Since the outermost loop index is also the subscript of the distribution dimension of array B, all references to B will be purely local. We cannot accomplish this for both A and B simultaneously since the subscripts in the distribution dimensions of A and B are different; therefore, there will be nonlocal accesses to A. However, since the subscript in the distribution dimension of the reference to A was placed second in the data access matrix, this subscript in the new loop nest corresponds to the second loop index, and we can perform block transfers from accesses to A; as was shown in Figure 1d.

For future reference, we define the following notion.

Definition 4.1 Given an array reference, an array subscript is *normal* with respect to loop i if it is equal to the loop index variable i.

In this example, the data access matrix yielded the transformation without any complications. This is not the case in general. First, the data access matrix may not be invertible. We handle this case in Section 5. Second, the transformation suggested by the data access matrix may violate one or more data dependences. We take care of this problem in Section 6. In both cases, the goal is to produce an invertible matrix that retains as many rows of the data access matrix as possible.

5. NONINVERTIBLE DATA ACCESS MATRICES

In general, the data access matrix is not invertible, so it cannot be used directly to transform the loop nest. The techniques in this section convert such a matrix into an invertible matrix that retains as many rows (subscripts) of the data access matrix as possible. This is done in two stages—first, we eliminate linearly dependent rows from the data access matrix using algorithm *BasisMatrix*, and second, we pad this reduced matrix with additional rows using algorithm *Padding*, to get a matrix that is invertible. The details of these algorithms can be found in the associated technical report; here we outline what these algorithms do.

5.1 Basis Matrix

It is easy to design an inefficient algorithm that takes a data access matrix and selects as many linearly independent rows as possible; we simply go down the rows of the matrix in sequence, discarding a row if it is linearly dependent on the rows before it, and keeping it otherwise. It is important to traverse the rows in sequence since it ensures that less important rows are discarded in favor of more important ones. For future reference, let us call the resulting matrix the *basis matrix* corresponding to the data access matrix.

Definition 5.1 The *basis matrix* of a data access matrix A is the first row basis of A.

```
Input:  An m × n data access matrix A.
Output: An m × m permutation matrix and the rank of A.
Algorithm BasisMatrix(A) : (PermMatrix, Rank)
begin
  P = I, where I is the m × m identity matrix.
  done = false;
  i = 1;
  While not done do
    /*Consider the submatrix A[i : m, i : n]* /
    Search for the first j ≥ i such that A[j, i : n] ≠ 0⃗;
    If no such j exists Then
      done = true;
    Else
      If j ≠ i then
        Exchange A[i, 1 : n] with A[j, 1 : n]
        Exchange P[i, 1 : m] with P[j, 1 : m]
      End-If
      Apply the elementary column operations to make
        A[i, i] nonzero and A[i, i + 1 : n] zero.
      i = i + 1;
    End-If
  End-While
  return (P, i − 1);
end
```

Fig. 3. Computing a basis matrix.

The algorithm described informally above is simple, but it is expensive to keep checking rows for independence. A more efficient algorithm is obtained by using a variation of computing the Hermite normal form [Li and Pingali 1992a]. A detailed understanding of this algorithm is not important for reading the rest of the article, so we give an informal description of what it does. Given a data access matrix, algorithm *BasisMatrix* in Figure 3 returns a permutation matrix P and the rank d of the data access matrix (the number of linearly independent rows). The first d rows of the permutation matrix P tell us which rows of the data access matrix are in the basis matrix. The following example should make this clear.

Consider the data access matrix

$$X = \begin{pmatrix} 1 & 1 & -1 & 0 \\ 2 & 2 & -2 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

This data access matrix can arise from the following program:

```
for i = ...
  for j = ...
    for k = ...
      for l = ...
        R[i + j − k, 2i + 2j + 2k, k − l] = ...
```

```
Input:  An m × n basis matrix B.
Output: An (n − m) × n padding matrix H.
Algorithm Padding(B) : PadMatrix
begin
  H = I, where I is an n × n identity matrix.
  For i = 1, m do
    /* Consider the submatrix B[i:m, i:n]* /
    apply the elementary column operations to make
      B[i, i] nonzero and B[i,i + 1 : n] zero.
    If columns i and j have been exchanged Then
      exchange rows i and j of H
    End-If
  End-For
  return (H[m + 1 : n, 1 : n]);
end
```

Fig. 4. Computing a padding matrix.

Algorithm *BasisMatrix*(X) returns the permutation matrix

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

and rank $d = 2$. The first two rows of the permutation matrix tell us which rows of A form a linearly independent basis: the position of the nonzero entry in these rows of P indicates which row of A is in the basis. In this example, the first and third rows form the basis matrix

$$B = \begin{pmatrix} 1 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

The significance of this in terms of transformations is that only the first and third subscripts can be normalized. This is reasonable because the subscript $2i + 2j - 2k$ is just a multiple of the subscript $i + j - k$.

5.2 Padding Matrix

To extend the basis matrix to an invertible matrix, we need to add additional mutually indepedent rows which are also independent of the rows of the basis matrix. There is some flexibility in the choice of the padding matrix, and we will use this flexiblity to our advantage in the next section when we discuss dependences. Algorithm *Padding* in Figure 4 constructs one possible padding matrix as follows. It is well known that for a full row rank matrix, there exist m columns that are linearly independent. We simply need to pad these columns with $\vec{0}$ and the rest of the columns with columns from the $(n - m) \times (n - m)$ identify matrix I. For the above program, since the first column and the third column are linearly independent, the padding matrix is

$$H = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

The mapping between the old and new iteration spaces is

$$\begin{pmatrix} 1 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} i \\ j \\ k \\ l \end{pmatrix} = \begin{pmatrix} u \\ v \\ w \\ z \end{pmatrix}.$$

In the transformed program shown below, the reference becomes $R[u, 2u, v]$, and second index is not normalized.

```
for u = ...
  for v = ...
    for w = ...
      for z = ...
        R[u, 2u, v]
```

6. DATA DEPENDENCES

The result of Section 5 shows that a basis matrix can always be padded to yield an integer matrix. However, there is no guarantee that the transformation corresponding to this final matrix is legal, because this transformation may violate data dependences.

There are three kinds of data dependence between statements. A dataflow dependence exists when one statement assigns to a variable read by another statement. A data antidependence exists when one statement reads a variable that is reassigned later by another statement. A data output dependence exists when one statement assigns to a variable that is reassigned later by another statement. A data dependence in the loop nest can be represented by a *distance vector* or *direction vector*. For example, the distance vector

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

tells us that the dependence is between successive iterations of the innermost loop. A dependence vector has the property that its leading nonzero is always positive; a legal transformation must preserve this property for each dependence, since the source of the dependence must be executed before its destination. More information on data dependences and techniques of dependence analysis can be found in [Banerjee 1988].

To understand the legality of transformations, consider

$$A = \begin{pmatrix} -1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$$

a basis matrix, and

$$D_A = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

the dependence matrix. Each column of the dependence matrix represents the distance vector of a dependence in the loop nest. In our example, there is just one dependence. If T is an invertible matrix representing a loop transforma-

tion, it is easily shown that TD is the dependence matrix of the restructured loop nest; therefore, the leading nonzero element in each column of TD must be positive. By looking at the product

$$AD_A = \begin{pmatrix} 0 \\ -1 \end{pmatrix},$$

we can see at once that A cannot be padded to give us a transformation that respects data dependences. The intuition is that the first two rows of A determine the two outermost loops of the transformed loop nest. In the original program, the dependence was carried by the innermost loop, but in the new program, the dependence is "carried" by the second loop. Unfortunately, the negative value of the second dimension of AD_A means that the source of the dependence will be executed after the sink. Clearly, there is nothing we can do in the *inner* loops that would remedy this situation, so it is impossible to pad A to yield a legal transformation.

To get around this problem, we proceed in two steps. We start with the basis matrix and use algorithm *LegalBasis* to produce a new basis matrix that does not violate dependences. Then, we pad this matrix using algorithm *LegalInvt* to yield the final transformation. In this article, we discuss only the case when dependences are represented by distances; it is straightforward to extend these results to dependence *directions* [Li and Pingali 1992a].

6.1 Generating a Legal Basis

Algorithm *LegalBasis*, shown in Figure 5, takes a basis matrix and checks each row against the dependences. For example, consider the product of the first row and D_A. This gives us a row vector in which entries can be positive, zero, or negative. If an entry is positive, it means that the corresponding dependence will be carried by the new outermost loop. Therefore, the structure of the inner loops does not matter as far as this dependence is concerned, and we may delete it from the D_A matrix for the rest of the algorithm. If the entry is zero, then the dependence will not be carried by the potential outermost loop, so we leave the dependence in the D_A matrix. However, if we have a negative entry, the dependence is "carried" by the potential outer loop, but the order of the iterations is wrong, Notice that if all of the entries of the row vector are 0 or negative (intuitively, for all dependences, the potential outer loop either does not carry the dependence, or the source of the dependence is executed after the sink), we can simply reverse the direction of the loop. Problems arise only if some entries are positive and others negative—in that case, we cannot keep that row of the basis matrix, and we delete it from the basis matrix. For the above example, *LegalBasis* (A) generates the basis

$$A_1 = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}.$$

6.2 Legal Padding Matrix

To pad a legal basis matrix, we need to satisfy two constraints. First, any row added must be linearly independent of other rows, so that the final matrix is

```
Input:  An m × n basis matrix B
        and a dependence matrix D.
Output: A legal basis Matrix.
Algorithm LegalBasis(B,D) : BasisMatrix
begin
  Let B_i be the ith row of B
    and d_i be the ith column of D.
  For i = 1, m
    f^T = B_i D
    If each element of f is non-negative then
      D = D − d_j, where f[j] > 0
    Elseif each element of f is non-positive then
      B_i = (−1) B_i;
      D = D − d_j, where f[j] < 0
    Else
      B = B − B_i;
    End-If
  End-For
  return B;
end
```

Fig. 5. Computing a legal basis matrix.

invertible. Second, the row must not violate dependence constraints. Once a new row has been added during padding, all dependences carried by the loop corresponding to this row may be dropped from consideration when filling in the rest of the matrix. When there are no further dependences to be satisfied, we can apply algorithm *Padding* of section 5.2 to complete the generation of a legal, invertible matrix.

As an example, consider the basis matrix $B = (-1\ \ 1\ \ 0)$ which is legal with respect to the dependence matrix

$$D = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

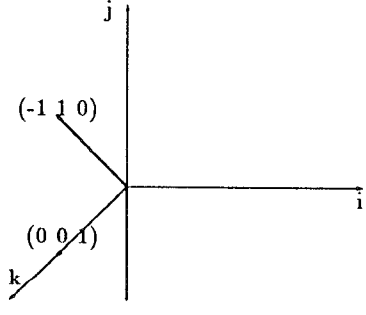
The first dependence is carried by the new outermost loop represented by the first row of B and can be dropped from consideration for the rest of the procedure. The inner product of the first row with the second dependence is 0, meaning that this dependence is not carried by the new outermost loop; therefore, it must be taken into account when padding the matrix. To pad B, we need to find a row whose inner product with the second dependence vector is nonnegative. In the geometric sense, the angle between the two vectors must be less than or equal to 90 degrees. Thus, the general problem can be stated succinctly as that of finding a vector that is linearly independent of the existing row vectors in the basis matrix and within 90 degrees of each dependence vector.

It is not immediately clear that such a vector exists; fortunately, algorithm *LegalInvt* in Figure 6 gives a positive answer by computing such a vector

```
Input:  An m × n legal basis matrix B
        and a dependence matrix D.
Output: An n × n legal invertible matrix T.
Algorithm LegalInvt(B,D) : Matrix
begin
  /* Let B_i be row i of B, and d_i be column i of D* /
  For i = 1, m
    f^T = B_i D
    D = D − d_j, where f[j] > 0
  End-For
  r = m + 1;
  While D is not empty do
    Z^T = the basis matrix of D^T;
    find the first e_k that is not orthogonal to D;
    x = cZ(Z^T Z)^-1 Z^T e_k;
      where c is a positive integer that makes x
        an integer vector.
    f^T = x^T D; /*f[j] ≥ 0* /
    D = D − d_j, where f[j] > 0
    B_r = x^T;
    r = r + 1;
  End-While
  H = Padding(B);
  return(append(B,H));
end
```

Fig. 6. Computing a legal invertible matrix.

using a standard result about projections. This vector can be written as $x = cZ(Z^TZ)^{-1}Z^Te_k$ for some positive scaling integer c that makes all of the entries integers, where $e_i^T = [0, 0, \ldots, 1, \ldots, 0]$, with the 1 in the ith position, and Z is a column basis from D.



For our example, the remaining dependence to be satisfied is e_3. The new row vector for the padding is $x = e_3$. Since the dependence is carried by the loop corresponding to this new row vector, we can drop the dependence from consideration now. The dependence matrix is empty at this point. The new legal basis matrix is

$$B_1 = \begin{pmatrix} -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then we can use the algorithm *Padding* to produce an invertible matrix. The final matrix

$$T = \begin{pmatrix} -1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

is a linear, invertible matrix, and the corresponding transformation satisfies all of the dependences.

The correctness of Algorithm *LegalInvt* follows from the following theorem.

THEOREM 6.1 *The invertible matrix returned by algorithm LegalInvt is consistent with program dependences.*

PROOF. Notice that the dependence vectors that need to be satisfied are orthogonal to the row vectors of the basis matrix. If we can find a vector from the subspace spanned by the dependence vectors, then this vector must be orthogonal to the basis rows therefore linearly independent of the existing row vectors. The invariants of the while-loop are that $AD = 0$; the rows of the A are linear independent; and for every column d_i of D, $e_k^T d_i \geq 0$. Let $d_i = Zy$ for some y since Z is a basis of the columns of D. Since $x = cZ(Z^TZ)^{-1}Z^Te_k$ and $AZ = 0$, $Ax = 0$. Since $e_k^T d_i \geq 0$, we conclude that $x^T d_i \geq 0$. After each step, the rank of the column space of D decreases at least by one; so the size of D is decreasing, and the algorithm will terminate. □

As a final remark, we note that the choice of the padding matrix in this article is quite arbitrary. For a machine in which processors have a first-level cache, there is the obvious possibility of selecting the padding to improve cache performance by incorporating results on blocking of nested loops [Gannon et al. 1988; Schreiber and Dongorra 1990]. We leave this for future work.

6.3 Direction Vectors

Direction vectors provide a conservative approximation when the distance of dependences cannot be detected at compile-time. They can be represented by signs "<", ">", "=", and "*". "<" means that the distance is positive, ">" negative, and "*" unknown. The leading nonzero of a direction vector is positive.

The algorithm in this article can be expended to handle direction vectors. For lack of space, details can be found in the associated technical report [Li and Pingali 1992a].

7. NUMA CODE GENERATION

Once the program has been transformed by access normalization, we must generate the code that will run on each processor. We generate the same code for each processor, but this code is parameterized by the processor number so that each processor does only the work for which it is responsible.

The general technique for partitioning the iteration space of the loop nest among the processors is called *tiling*. Here, we will restrict ourselves to the

for $i = 1, u$	for $i = \lceil \frac{l-p}{P} \rceil * P + p,$ $u,$ step P	for $i = max(l, p * S),$ $min(u, (p+1) * S - 1)$
(a) unit step	(b) task p for wrapped distribution	(c) task p for blocked distribution
for $i = l, u,$ step s	for $i = l + n_0 * s,$ $u,$ step $P / (P,s) * s$	for $i = max(l, p * S),$ $min(u, (p+1) * S - 1),$ step s
(d) non-unit step	(e) task p for wrapped distribution	(f) task p for blocked distribution

Fig. 7 Distributed loops among processors

special case of wrapped and blocked distributions introduced in Section 2. For these distributions, it is sufficient to distribute the iterations of the outermost loop of the transformed loop nest among the processors. Consider the first row of the transformation matrix: one of the following cases must be true.

—The row was present in the data access matrix, so it corresponds to a subscript in the original program, and this subscript is in a distribution dimension.

—The row was present in the data access matrix, but it is not a distribution dimension.

—The row was introduced by padding.

In cases (ii) and (iii), access normalization cannot exploit locality, and we generate code simply by assigning iterations to processors in a round-robin manner. This code can still exploit block transfers. For case (i), an iteration should be executed by a processor if the corresponding data element is mapped to its local memory.

First, consider the case when the step size is 1. For a wrapped distribution, processor p owns the *data segments* p, $p + P$, $p + 2P, \ldots$ etc., where a data segment is a column in the wrapped column distribution or a row in the wrapped row distribution. Since the iterations that access the data segments on processor p are assigned to processor p, it is easy to verify that the iterations executed by processors p are the ones shown in Figure 7b. The lower bound $\lceil (1-p)/(P) \rceil * P + p$ is the first iteration between l and u that belongs to process p.

When the step size is not 1 (Figure 7d), we must solve a linear congruence for the wrapped distribution. Assume that the step size is positive, since the solution can be easily extended to handle the case when the step size is negative. The iterations can be represented by $i = l + n * s$ where n is a parameter with integer values. The iterations that belong to process p are these satisfying the equation $l + n * s = p \pmod{P}$. Using results from number theory, we know that when the g.c.d. of s and P (written as (s, P)) divides $(l - p)$, there is an infinite number of solutions in the form of $n = n_0 + t * P/(s, P)$ for some integer solution $0 \leq n_0 < P/(s, P)$ and integer free variable t. However, only certain t's are solutions for iterations within

the loop bounds. Since $l \le i \le u$ and $i = l + (n_0 + t * P/(s, P)) * s$, the range of t is $\lceil(-n_0)/(P/(s, P))\rceil = 0 \le t \le \lfloor(u - l - n_0 * s)/(P/(s, P) * s)\rfloor$. Therefore the loop for processor p is as shown in Figure 7e.

Given this assignment of iterations to processors, we must generate synchronization instructions to take care of dependences carried by the outermost loop and insert block transfers wherever possible. Inserting block transfers is similar to message vectorization in distributed-memory machines or block-invalidates for software cache-coherent schemes. These steps are routine [Gerndt 1989; Midkiff and Padua 1987; Rogers 1990] and are omitted from this article.

8. EMPIRICAL RESULTS AND PERFORMANCE ANALYSIS

In this section, we report the performance of our techniques on routines from the BLAS library. The target machine is a BBN Butterfly GP-1000. On this machine, a processor can access its local memory in about 0.6 microseconds, but a nonlocal access takes about 6.6 microseconds even in the absence of contention in the network. For block transfers, the startup time is about 8 microseconds, and after that, a byte is transferred every 0.31 microseconds [BBN 1989]. Our compiler takes as input Fortran-77 programs with data distribution information, and it generates C code for each processor; this node program is compiled into native code using the Green Hills C compiler (Release 1.8.4). The C compiler performs only conventional code optimizations, so our experimental results are not skewed by any restructuring performed by this compiler. We will use pseudocode in discussing examples.

For the GEMM code, our techniques are successful in eliminating nonlocal accesses significantly, so block transfers contribute just a small amount to overall performance. In the SYR2K code, the reduction of nonlocal accesses is less significant, so block transfers of nonlocal data are important for good performance.

8.1 GEMM

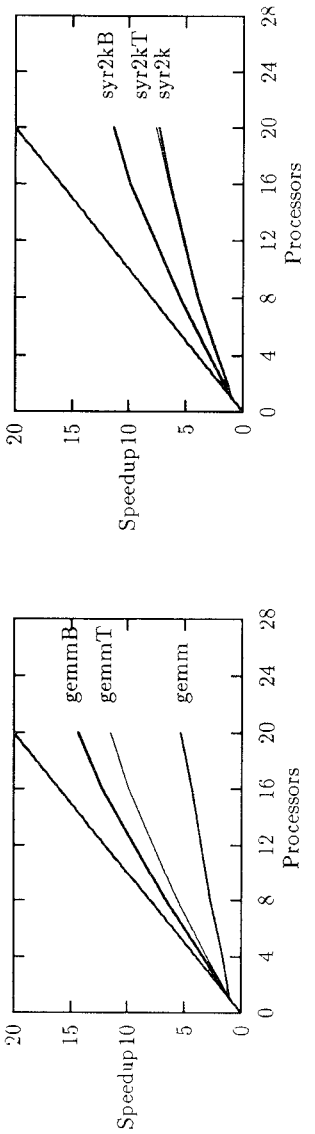
General matrix multiplication (GEMM) is one of the central subroutines in BLAS.

```
for i = 1, N
  for j = 1, N
    for k = 1, N
      C[i, j] = C[i, j] + A[i, k] * B[k, j]
```

All arrays are of size 400 by 400 and are distributed in wrapped-column manner. By distributing the outermost loop among the processors without doing any transformations, we obtain the graph labeled *gemm* in Figure 8a.

The data access matrix is

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$



(a) GEMM (b) banded SYR2K

Fig 8. Speedups of GEMM and banded SYR2K

and the dependence matrix is

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

The invertible matrix for the transformation is,

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

The transformed loop nest yields the following parallel code with the performance labeled *gemmB* (Figure 8a). The curve labeled *gemmT* is the speedup without block transfers.

```
for u = p, N, step P
  for v = 1, N
    read A[*, v];
    for w = 1, N
      C[w, u] = C[w, u] + A[w, v] * B[v, u]
```

After access normalization, accesses to C and B are local, but there are nonlocal accesses to A. Since three out of four data structure accesses in each iteration have become local, the effect of block transfers is relatively small.

8.2 SYR2K

When remote accesses are necessary due to the problem structure, it is beneficial to use block data transfers to amortize the cost of the startup time. Consider the rank $2k$ update SYR2K from BLAS [Coleman and Van Loan 1988]. The subroutine computes $C = \alpha A^T B + \alpha B^T A + C$. Suppose A and B are banded matrices with band width b, then C is symmetric and banded with band width $2b - 1$. The banded matrices A, B are stored in $n \times 2b - 1$ arrays A_b, B_b such that the elements $A[i, j]$, $B[i, j]$ are in $A_b[i, j - i + b - 1]$ and $B_b[i, j - i + b - 1]$. C is symmetric so only the upper triangular matrix is stored in an $n \times (2b - 1)$ array C_b such that $C[i, j]$ is in $C_b[i, j - i]$. The program is shown below.

```
for i = 1, N
  for j = i, min(i + 2b - 2, N)
    for k = max(i - b + 1, j - b + 1, 1),
        min(i + b - 1, j + b - 1, N)
      C_b[i, j - i + 1] = C_b[i, j - i + 1]
        + αA_b[k, i - k + b] * B_b[k, j - k + b]
        + αA_b[k, j - k + b] * B_b[k, i - k + b]
```

Assume that we are given a wrapped-column mapping for each array. The data access matrix is

$$\begin{pmatrix} -1 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \end{pmatrix}.$$

If we apply algorithm *BasisMatrix*, we get a base matrix B consisting of the first three rows. However, the dependence matrix is $[0, 0, 1]^T$. The legal base mapping is

$$B_{legal} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

which is B with the second row negated. This matrix is invertible. Using B_{legal} as the transformation matrix and parallelizing the new nest, we get the following parallel code.

```
for u = p, 2b − 2, step P
  for v = 1 − b, b − u
    read A_b[*, −u − v + b], read A_b[*, −v + b];
    read B_b[*, −v + b]; read B_b[*, −u − v + b];
      for w = max(1, u + v), min(N, N + v)
        C_b[−u − v + w + 1, u] = C_b[−u − v + w + 1, u]
          +αA_b[w, −u − v + b] * B_b[w, −v + b]
          +αA_b[w, −v + b] * B_b[w, −u − v + b]
```

The experimental results are shown in Figure 8b. The problem size is 500 with a band size of 200. Block transfers are relatively important in this example, since there are many nonlocal accesses left in the transformed code.

A simple performance model explaining these results can be found in the associated technical report.

9. SUMMARY AND RELATED WORK

This article is a contribution to the state of the art of compiling programs in languages like HPF-Fortran that permit user-defined data decomposition for parallel machines with a memory hierarchy. This compilation problem is the goal of a number of projects including Fortran-D, Id Nouveau, Superb, and Crystal [Callahan and Kennedy 1988; Hiranandani et al. 1991; Koelbel and Mehrotra 1991; Li and Chen 1989; Mirchandaney et al. 1988; Tseng 1989; Rogers and Pingali 1989; Zima and Chapman 1990].

The emphasis in these projects has been on code generation mechanisms (such as the ownership rule discussed in Section 2) and on recognizing and exploiting special patterns of computation and communication, such as reductions. Although it is well known that loop restructuring before code generation can improve performance, no systematic loop-restructuring mechanism using a general loop transformation framework has been available.

We have attempted to exploit locality by matching code to the data distribution across the machine. This is a *static* notion of locality and must be differentiated from the *dynamic* locality that must be exploited on parallel machines with coherent caches [Kendal Square Research 1991]. On such machines, the key to high performance is *data reuse*, and the code must be restructured to allow reuse of cached data wherever possible. Restructuring techniques for doing this have been explored by Wolf and Lam [1991a]. Their approach is complementary to the one described here. It is likely that scalable parallel architectures will be organized as networks of processor-memory

pairs in which processors have an on-chip cache and perhaps a second-level cache between the processor and its local memory. The techniques in this article can be used to partition work and data among the processors; techniques to enhance data reuse can be used to optimize uniprocessor cache performance.

Our use of matrix techniques generalizes the unimodular matrix approach [Banerjee 1990; Wolf and Lam 1991b]. Unimodular matrices were used by Kumar et al. [1991] to eliminate outermost loop-carrier dependence in generating code for distributed-memory machines. In our work, we use invertible matrices which include unimodular matrices as a special case. This lets us model loop scaling as well, which is important in the NUMA context. In general, it is easier to work with invertible matrices since there are fewer constraints to be satisfied in generating invertible matrices, as opposed to unimodular matrices. There are a number of other loop transformations like distribution, jamming, and alignment that are useful in generating code for parallel machines [Wolf 1989]. It would be useful to extend the matrix framework to incorporate these transformations. Related work on loop transformations can be found [Allen and Kennedy 1987; Ancourt and Irigoin 1991; Gannon et al. 1988; Irigoin and Triolet 1988; Lamport 1974; Lu 1991; Padua and Wolfe 1986; Porterfield 1989; Schreiber and Dongarra 1990; Whitfield and Soffa 1991; Wolf and Lam 1991b; Wolfe 1989].

The data access matrix is a new concept introduced in this article, and access normalization is useful in other contexts such as code generation for vector machines. On many vector machines such as the CRAY-1 and CRAY-2, vector loads and stores must have constant stride. Even on machines such as the Fujitsu FACOM that support scatter-and-gather operations, it is more efficient to use constant-stride accesses wherever possible since address generation for vector elements is faster. The techniques in this article can be used to accomplish this [Li and Pingali 1992a].

We require the programmer to specify data distributions. Automatic deduction of this information for special programs has been investigated by Balasundaram et al. [1990], by Gannon et al. [1988] on CEDAR-like architectures, by Hudak and Abraham [1990] for sequentially iterated parallel loops, by Knobe et al. [1990] for SIMD machines, by Li and Chen [1989] for index domain alignment and by Ramanujam and Sadayappan [1991] who find communication-free partitioning of arrays in fully parallel loops. These efforts focus on deducing good data distributions for particular kinds of programs such as fully parallel loops, and no general solution to this problem is known. We speculate that it might be possible to start with the dependence matrix and use our techniques in reverse, so to speak, to determine what a good data distribution should be. The main difficulty in doing this is to ensure that the resulting parallel code is load balanced.

ACKNOWLEDGMENTS

We thank Radha Jagadeesan and Danny Ralph for useful discussion, Richard Huff for proofreading the draft, and Mark Charney, Richard Johnson, Mayan

Moudgill, and Paul Stodghill for comments. The anonymous referees from both *ASPLOS '92* and *TOCS* have provided us with many helpful comments. In particular, we are grateful to Mary Lou Soffa for shepherding the *ASPLOS* version of this article.

REFERENCES

AGARWAL, A. 1991. Limits on interconnection network performance. *IEEE Trans. Parall. Distrib. Syst. 2*, 4, 398–412

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

ALLEN, F., COCKE, J., AND KENNEDY, K. 1981. *Reduction of Operator Strength.* Prentice-Hall, Englewood Cliffs, N J , 79–101.

ANCOURT, C., AND IRIGOIN, F. 1991. Scanning polyhedra with DO loops in *The 3rd ACM Symposium on Principles and Practice of Parallel Programming.* ACM, New York, 39–50.

BALASUNDARAM, V., FOX, G., KENNEDY, K., AND KREMER, U. 1991. An iteractive environment for data partitioning and distribution. In *Proceedings of the 5th Distributed Memory Computing Conference.* IEEE, New York.

BANERJEE, U. 1990. Unimodular transformations of double loops. In *Proceedings of the Workshop on Advances in Languages and Compilers for Parallel Processing.* 192–219

BANERJEE, U. 1988. *Dependence Analysis for Supercomputing.* Kluwer Academic, New York.

BBN ADVANCED COMPUTERS INC. 1989. *Butterfly GP1000 Switch Tutorial.* BBN, Cambridge, Mass

CALLAHAN D., AND KENNEDY, K. 1988. Compiling programs for distributed memory multiprocessors. *J. Supercomput. 2*, 2 (Oct.).

COLEMAN, T., AND VAN LOAN C. 1988. *Handbook for Matrix Computations* SIAM, Philadelphia.

GANNON, D., JALBY, W., AND GALLIVAN, K. 1988. Strategies for cache and local memory management by global program transformations *J. Parall. Distrib. Comput. 5*, 587–616.

GERNDT, H. M. 1989. Automatic parallelization for distributed-memory multiprocessing systems. Ph.D. thesis, Bonn Univ., Bonn, Germany.

HIRANANDANI, S., KENNEDY, K., AND TSENG, C 1991. Compiler optimizations for FORTRAN-D on MIMD distributed-memory machines. Tech. Rep. TR91-156, Rice Univ.

HUDAK, D., AND ABRAHAM, S. 1991 Compiler techniques for data parititioning of sequentially iterated parallel loops. In *Proceedings of the ACM International Conference on Supercomputing.* ACM, New York.

IRIGOIN, F., AND TRIOLET, R. 1988. Supernode partitioning. In *Proceedings of the 15th Annual ACM Symposium on Principles of Programming Languages.* ACM, New York

KENDALL SQUARE RESEARCH CORPORATION 1991. *Parallel Programming Manual.* Waltham, Mass.

KNOBE, K., LUKAS, J., AND STEELE, G. 1990. Data optimization: Allocation of arrays to reduce communication on SIMD machines. *J. Parall Distrib. Comput. 8*, (Feb.), 102–118

KOELBEL AND MEHROTRA, P. 1991. Compiling global namespace parallel loops for distributed execution. *IEEE Trans. Parall. Distrib. Syst. 2*, (Oct.).

KUMAR, K. G., KULKARNI, D., AND BASU, A. 1991 Generalized unimodular loop transformtions for distributed memory multiprocessors. Tech Rep FG-TR-014, Center for Development of Advanced Computing, Bangalore, India.

LAMPORT, L. 1994. The parallel execution of do loops. *Commun. ACM 27*, 2 (Feb.) 83–93.

LI, J., AND CHEN, M. 1989. Index domain alignment: Minimizing cost of cross-referencing between distributed arrays Tech. Rep., Yale Univ.

LI, W., AND PINGALI, K. 1992a. Access normalization: Loop restructuring for NUMA compilers. Tech. Rep. 92-1278, Department of Computer Science, Cornell Univ., Ithaca, N Y.

LI, W. AND PINGALI, K. 1992b. A singular loop transformation framework based on non-singular matrices. Tech. Rep. 92-1294, Dept. of Computer Science, Cornell Univ., Ithaca, N.Y.

LU, L. 1991. A unifed framework for systematic loop transformations. In *The 3rd ACM SIGPLAN Symposium on Principles and Practice of Parallel programming.* ACM, New York 28–38.

MIDKIFF, S. P., AND D. A. PADUA. 1987. Compiler algorithms for synchronization. *IEEE Trans. Comput. C-36*, (Dec.), 1485–1495.

MIRCHANDANEY, R., SALTZ, J., SMITH, R., NICOL, D., AND CROWLEY, K. 1988. Principles of runtime support for parallel processors. In *Proceedings of the 2nd International Conference on Supercomputing.*

PADUA, D., AND WOLFE, M. 1986. Advanced compiler optimizations for supercomputers. *Commun. ACM 29*, 12 (Dec.), 1184–1201.

PORTERFIELD, A. 1989. Software methods for improvement of cache performance on supercomputer applications. Ph.D. thesis, Rice Univ.

RAMANUJAM, J., AND SADAYAPPAN, P. 1991. Compile-time techniques for data distribution in distributed memory machines. *IEEE Trans. Parall. Distrib. Syst. 2* (Oct.).

ROGERS, A. 1990. Compiling for locality of reference. Ph.D. thesis, Cornell Univ., Ithaca, N.Y.

ROGERS, A., AND PINGALI, K. 1989. Process decomposition through locality of reference. In *Proceedings of the 1989 SIGPLAN Conference on Programming Language Design and Implementation.* ACM, New York.

SCHREIBER, R., AND DONGARRA, J. 1990. Automatic blocking of nested loops. Tech. Rep. 90.38, NASA RIACS.

TSENG, P. 1989. A parallelizing compiler for distributed memory parallel computers. Ph.D. thesis, Carnegie-Mellon Univ., Pittsburgh, Penn.

WHITFIELD, D., AND SOFFA, M. L. 1991. Automatic generation of global optimizers. In *Proceedings of the SIGPLAN '91 Conference on Programming Language Design and Implementation. SIGPLAN Not.* (June).

WOLF M., AND LAM, M. 1991a. A data locality optimizing algorithm. In *Proceedings ACM SIGPLAN '91 Conference on Programming Language Design and Implementation.* ACM, New York, 30–44.

WOLF, M., AND LAM, M. 1991b. A loop transformation theory and an algorithm to maximize parallelism. *IEEE Trans. Parall. Distrib. Syst. 13*, 4 (Oct.).

WOLFE, M. 1989. *Optimizing Supercompilers for Supercomputers.* Pitman Publishing, London.

ZIMA, H., AND CHAPMAN, B. 1990. *Supercompilers for Parallel and Vector Computers.* ACM Press Frontier Series, New York.

Received September 1992; revised June 1993; accepted June 1993