2000

Performance Enhancement by Memory Reduction

Yonghong Song
Rong Xu
Cheng Wang
Zhiyuan Li
Purdue University, li@cs.purdue.edu

Report Number:
00-016
Performance Enhancement by Memory Reduction *†

Yonghong Song  Rong Xu  Cheng Wang  Zhiyuan Li
Department of Computer Sciences
Purdue University
West Lafayette, IN 47907
{songyh,xur,wangc,li}@cs.purdue.edu

ABSTRACT
In this paper, we propose a technique to reduce the virtual memory required to store program data. Specifically, we present an optimal algorithm to combine loop shifting, loop fusion and array contraction to reduce the temporary array storage required to execute a collection of loops. Memory reduction is formulated as a network flow problem, which is solved by the proposed algorithm in polynomial time. When applied to 20 benchmark programs on two platforms, our technique reduces the memory requirement, counting both the data and the code, by 50% on average. The transformed programs gain a speedup of 1.57 on average, due to the reduced working set and, consequently, the improved data locality. In the best case, a maximum speedup of 41.3 is achieved for one of the benchmark programs.

1. INTRODUCTION
Compiler techniques, such as tiling [29, 30], to exploit temporal data locality within a single loop nest have been studied extensively. However, how to effectively exploit temporal locality between different loop nests remains unclear. This state of the art makes it important to seek locality enhancement techniques beyond tiling.

In this paper, we approach the locality issue by reducing the virtual memory required to store program data. In particular, we seek opportunities to contract the number of dimensions of arrays. For example, a two-dimensional array may be contracted to a single dimension, or a whole array may be contracted to a scalar.

A significant potential benefit, among others, of such a reduction in data size is the increased reuse of the archived data due to the reduced working set.

Consider an extremely simple example (Example 1 in Figure 1(a)), where array A is assumed dead after loop L2. After right-shifting loop L2 by one iteration (Figure 1(b)), L1 and L2 can be fused (Figure 1(c)). Array A can then be contracted to two scalars, a1 and a2, as Figure 1(d) shows. (As a positive side-effect, temporal locality of array E is also improved.) The aggressive fusion proposed here also improves temporal data locality between different loop nests.

For a collection of loops defined later in this paper, we formulate the memory reduction problem as a network flow problem, which is optimally solvable in polynomial time. Additional loop transformations, such as loop interchange and circular loop skewing [30], are used to create opportunities for aggressive fusion.

We have implemented our memory reduction technique in our research compiler. We apply our technique to 20 benchmark programs on two platforms in the experiments. On average, the memory requirement for these benchmarks is reduced by 50%, counting both the code and the data, using the arithmetic mean. The transformed programs have an average speedup of 1.57 (using the geometric mean). A speedup of 41.3 is achieved for one of the benchmarks.

In the rest of this paper, we will present some preliminaries in Section 2. We formulate the network flow problem and prove its complexity in Section 3. We present controlled fusion and discuss enabling techniques in Section 4. Section 5 provides the experimental results, followed by related work and conclusion.

2. PRELIMINARIES
2.1 Program Model
We consider a collection of loop nests, \( L_1, L_2, \ldots, L_m \), \( m \geq 1 \), in their lexical order, as shown in Figure 2(a). The label \( L_i \) denotes a perfect nest of loops with indices \( L_{i,1}, \ldots, L_{i,n}, n \geq 1 \), starting from the outmost loop. (In Example 1, i.e. Figure 1(a), we have \( m = 2 \) and \( n = 1 \).) Loop \( L_{i,j} \) has the lower bound \( l_{i,j} \) and the upper bound

*Technical report CSD-TR-00-0016, Department of Computer Sciences, Purdue University, November, 2000.
†This work is sponsored in part by National Science Foundation through grants OCI-9975309 and MIP-9910379, by Indiana 21st Century Fund, by Purdue Research Foundation, and by a donation from Sun Microsystems, Inc.
The array regions referenced in the given collection of loops are divided into three classes:

- **An input array region** is upwardly exposed to the beginning of \( L_1 \).
- **An output array region** is live after \( L_m \).
- **A local array region** does not intersect with any input or output array regions.

By utilizing the existing dependence analysis, region analysis and live analysis techniques [4, 11, 12, 18], we can compute input, output and local array regions efficiently. Note that input and output regions can overlap with each other. In Example 1 (Figure 1(a)), \( E[0 : N] \) is both the input array region and the output array region, and \( L_1 : N \) is the local array region. Figure 1(b) shows a more complex example (Example 2), which resembles one of the well-known Livermore loops. In Example 2, where \( m = 4 \) and \( n = 2 \), each declared array is of dimension \([J+N+1, K+N+1] \). \( ZP, ZR, ZQ, ZZ, ZA[1 \ldots K], ZB[2 : JN, K]+1 \) and \( ZB[2 : JN, K]+1 \) are input array regions. \( ZP, ZR, ZQ, ZZ \) are output array regions. \( ZA[2,JN,2:KN] \) and \( ZB[2 : JN, 2 : KN] \) are local array regions.

Figure 2(b) shows the code form after loop shifting but before loop fusion, where \( p' \) represents the shifting factor for loop \( L_{i,j} \). In the rest of this paper, we assume that loops \( L_i \) are coalesced into single level loops \([30, 26] \) after loop shifting but before loop fusion, and Figure 2(c) shows the code form after loop shifting. The loops are coalesced to ease code generation for general cases. However, in most common cases, loop coalescing is unnecessary [26]. Figure 2(d) shows the code form after loop fusion without loop coalescing applied. Array contraction will then be applied to the code shown in either Figure 2(d) or in Figure.
2.2 Loop Dependence Graph

We extend the definitions of the traditional dependence distance vector and dependence graph \[14\] to a collection of loops as follows.

**Definition 1.** Given a collection of loop nests, \( L_1, \ldots, L_m \), as in Figure 2(a), if a data dependence exists from iteration \((i_1, i_2, \ldots, i_n)\) of loop \( L_1 \) to iteration \((j_1, j_2, \ldots, j_n)\) of loop \( L_2 \), we say the distance vector is \((j_1-i_1, j_2-i_2, \ldots, j_n-i_n)\) for this dependence.

**Definition 2.** Given a collection of loop nests, \( L_1, L_2, \ldots, L_m \), a loop dependence graph (LDG) is a directed graph \( G = (V, E) \) such that each node in \( V \) represents a loop nest \( L_i \), \( 1 \leq i \leq m \). (We denote \( V = \{L_1, L_2, \ldots, L_m\} \).) Each directed edge, \( e = L_i \rightarrow L_j \), in \( E \) represents a data dependence (flow, anti- or output dependence) from \( L_i \) to \( L_j \). The edge \( e \) is annotated by a distance vector \( d(e) \).

For each dependence edge \( e \), if its distance vector is not constant, we replace it with a set of edges as follows. Let \( \overrightarrow{d(e)} = (d_1, d_2, \ldots, d_s) \) be the lexicographically minimum distance in \( S \). Then \( d_1 \leq d_2 \leq \cdots \leq d_s \). For any vector \( d_i \) in \( S_i \), there exists no other vector in \( S \) which is smaller than \( d_i \). We replace the original edge \( e \) with \( (|S_1|+1) \) edges, annotated by \( d_0 \) and \( d_i \) \( (d_i \in S_i, 1 \leq i \leq |S_1|) \) respectively.

Figure 3 shows the dependence graph for the example in Figure 2(a), without showing the array regions. As an example, the flow dependence from \( L_1 \) to \( L_3 \) with \( d(e) = (0, 0) \) is due to array region \( ZA(2:2, JN:2:KN) \). In Figure 3(b), where multiple dependences of the same type (flow, anti- or output) exist from one node to another, we use one arc to represent them all in the figure. All associated distance vectors are then marked on this single arc.

2.3 Assumptions

We make the following three assumptions in order to simplify our formulation in Section 3.

**Assumption 1.** The loop trip counts for perfect nests \( L_i \) and \( L_j \) are equal at the same corresponding loop level \( h \), \( 1 \leq h \leq n \). This can also be stated as \( u_{i,h} - i_{i,h} + 1 = u_{j,h} - i_{j,h} + 1, 1 \leq i, j \leq m, 1 \leq h \leq n \).

To enforce Assumption 1, one could either partition the iteration spaces of certain loops into equal pieces, or apply loop peeling.

Throughout this paper, we use \( \beta^{(h)} \) to denote the loop trip count of loop \( L_h \) at level \( h \), which is constant or symbolically constant w.r.t. the program segment under consideration. Denote \( \beta = (\beta^{(1)}, \ldots, \beta^{(n)}) \). We let \( \sigma^{(1)} = 1 \) and \( \sigma^{(n)} = \sigma^{(n+1)} = 1, 1 \leq h \leq n - 1 \). Let \( \sigma = (\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(n)}) \). In this paper, we also denote \( \sigma \) as the number of static write references due to local array regions \(^3\) in loop \( L_i \). We arbitrarily assign each static write reference in \( L_i \) a number \( 1 \leq k \leq \sigma_i \), in order to distinguish them. Take loops in Figure 3(b) as an example, we have \( \beta = (KN:1, JN:1), \sigma = (KN-1, 1), \sigma_1 = 1 \) and \( \sigma_2 = 0 \).

\(^{3}\)In the rest of this paper, the term of "a static write reference" means "a static write reference due to local array regions".
We make the following assumption about the dependence distance vectors.

**Assumption 2.** The sum of the absolute values of all dependence distances at loop level \( h \) in loop dependence graph \( G = (V, E) \) should be less than one-fourth of the trip count of a loop at level \( h \). This assumption can also be stated as \( \sum_{e \in E} |dv(e)| < \frac{1}{4}T \) for all \( e \in E \) annotated with the dependence distance vector \( dv(e) \).

Assumption 2 is reasonable because for most programs, the constant dependence distances are generally very small. If non-constant dependence distances exist, the techniques discussed in Section 4.2, such as loop interchange and circular loop skewing, may be utilized to reduce such dependence distances.

**Assumption 3.** For each static write reference \( r \), each instance of \( r \) writes to a distinct memory location. No IF-statement guards the statement which contains the reference \( r \). Different static write references write to different portions of main memory.

If a static write reference does not write to a distinct memory location in each loop iteration, we apply scalar or array expansion to this reference [30]. Later on, our technique should minimize the total size of the local array regions. In case of IF statements, we assume both branches will be taken. In [26], we discussed the case where the regions written by two different static write references are the same or overlap with each other.

### 2.4 LDG Simplification

The loop dependence graph can be simplified by keeping only dependence edges necessary for memory reduction. The simplification process is based on the following three claims.

**Claim 1.** Any dependence from \( L_i \) to itself is automatically preserved after loop shifting, loop coalescing and loop fusion. This is because we are not reordering the computation within any loop \( L_i \).

**Claim 2.** Among all dependence edges from \( L_i \) to \( L_j \), \( i \neq j \), suppose that the edge \( e \) has the lexicographically minimum dependence distance vector. After loop shifting and coalescing, if the dependence distance associated with \( e \) is nonnegative, it is legal to fuse loops \( L_i \) and \( L_j \). This is because after loop shifting and coalescing, the dependence distances for all other dependence edges remain equal to or greater than that for the edge \( e \) and thus remain nonnegative. In other words, no fusion-preventing dependencies exist. We will prove this claim in Section 3 through Lemma 3.

**Claim 3.** The amount of memory needed to carry a computation is determined by the lexicographically maximum flow-dependence distance vectors which are due to local array regions. We will discuss this claim further in Section 2.5.

During the simplification, we classify all edges into two classes: \( L \)-edges and \( M \)-edges. The \( L \)-edges are used to determine the legality of loop fusion. The \( M \)-edges will determine the minimum memory requirement. All \( M \)-edges are flow dependence edges. But an \( L \)-edge could be a flow, an anti- or an output dependence edge. It is possible that one edge is both an \( L \)-edge and an \( M \)-edge. The simplification process is as follows.

- Based on the claims 1 and 3, for each combination of the node \( L_i \) and the static write reference \( r \) in \( L_i \) where \( \tau_i > 0 \), among all dependence edges from \( L_i \) to itself due to \( r \), we keep only the one whose flow dependence distance vector is lexicographically maximum. This edge is an \( M \)-edge.
- Based on the claims 1 and 3, for each node \( L_i \) such that \( \tau_i = 0 \), we remove all dependence edges from \( L_i \) to itself.
- Based on the claims 2 and 3, for each node \( L_i \) where \( \tau_i > 0 \), among all dependence edges from \( L_i \) to \( L_j \) \( (j \neq i) \), we keep only one dependence edge for legality such that its dependence distance vector is lexicographically minimum. This edge is an \( L \)-edge. For any static write reference \( r \) in \( L_i \), among all dependence edges from \( L_i \) to \( L_j \) \( (j \neq i) \) due to \( r \), we keep only one flow dependence edge whose distance vector is lexicographically maximum. This edge is an \( M \)-edge.
- Based on the claims 2 and 3, for each node \( L_i \) where \( \tau_i = 0 \), among all dependence edges from \( L_i \) to \( L_j \) \( (j \neq i) \), we keep only the dependence edge whose dependence distance vector is lexicographically minimum. This edge is an \( L \)-edge.

The above process simplifies the program formulation and makes graph traversal faster. Figure 4(c) shows the loop dependence graph after simplification of Figure 4(b). In Figure 4(c), we do not mark the classes of the dependence edges. As an example, the dependence edge from \( L_1 \) to \( L_2 \) marked with \((0,0)\) is an \( L \)-edge, and the one marked with \((0,1)\) is an \( M \)-edge. The latter edge is associated with the static write reference \( ZA(J,K) \).

#### 2.5 Reference Windows

Loop shifting is applied before loop fusion in order to honor all the dependences. We associate one integer vector \( \bar{p}(L_i) \) with each loop nest \( L_i \) in the loop dependence graph. Denote \( \bar{p}(L_i) = (p^1(L_i), \ldots, p^n(L_i)) \) where \( p^j(L_i) \) is the shifting factor of \( L_i \) at loop level \( k \) (Figure 2(b)). For each dependence edge \( < L_u, L_v > \) with the distance vector \( dv \), the new distance vector is \( \bar{p}(L_u) + dv - \bar{p}(L_v) \). Our memory minimization problem, therefore, reduces to the problem of determining the shifting factor, \( p^j(L_i) \), for each Loop \( L_i,j \), such that the total temporary array storage required is minimized after all loops are coalesced and legally fused.

In [9], Gannon et al. use a reference window to quantify the minimum cache footprint required by a dependence with a loop-invariant distance. We shall use the same concept to quantify the minimum temporary storage to satisfy a flow.
Definition 3. (from [9]) The reference window, \( W(\pi X) \), for a dependence \( \pi X : S_1 \to S_2 \) on a variable \( X \) at time \( t \), is defined as the set of all elements of \( X \) that are referenced by \( S_1 \) at or before \( t \) and will also be referenced (according to the dependence) by \( S_2 \) after \( t \).

In Figure 1(a), the reference window due to the flow dependence (from \( L_1 \) to \( L_2 \) due to array \( A \)) at the beginning of each loop \( L_2 \) iteration is \( \{ A(I), A(I + 1), \ldots, A(N) \} \). Its reference window size ranges from 1 to \( N \). In Figure 1(c), the reference window due to the flow dependence (caused by array \( A \)) at the beginning of each loop iteration is \( \{ A(I - 1) \} \). Its reference window size is 1.

Next, we extend Definition 3 for a set of flow dependences as follows.

Definition 4. Given flow dependence edges \( e_1, e_2, \ldots, e_s \), suppose their reference windows at time \( t \) are \( W_1, W_2, \ldots, W_s \) respectively. We define the reference window of \( \{ e_1, e_2, \ldots, e_s \} \) at time \( t \) as \( \bigcup_{j=1}^{s} W_j \).

Since the reference window characterizes the minimum memory required to carry a computation, the problem of minimizing the memory required for the given collection of loop nests is equivalent to the problem of choosing loop shifting factors such that the loops can be legally coalesced and fused and that, after fusion, the reference window size of all flow dependences due to local array regions is minimized. Given a collection of loop nests which can be legally fused, we need to predict the reference window after loop coalescing and fusion.

Definition 5. For any loop node \( L_i \) (in an LDG) which writes to local array regions \( R \), suppose iteration \( (j_1, \ldots, j_n) \) becomes iteration \( (j_1', \ldots, j_n') \) after loop coalescing and fusion. We define the predicted reference window of \( L_i \) in iteration \( (j_1, \ldots, j_n) \) as the reference window of all flow dependences due to \( R \) in the beginning of iteration \( j \) of the coalesced and fused loop. Suppose the predicted reference window with iteration \( (j_1', \ldots, j_n') \) has the largest size of those due to \( R \). We define it as the predicted reference window size of the entire loop \( L_i \) due to \( R \). We define the predicted reference window due to a static write reference \( r \) in \( L_i \) as the predicted reference window of \( L_i \) due to the array regions written by \( r \). (For convenience, if \( L_i \) writes to nonlocal regions only, we define its predicted reference window to be empty.)

Based on Definition 5, we have the following lemma:

Lemma 1. The predicted reference window size for the \( k \)th static write reference \( r \) in \( L_i \) must be no smaller than the predicted reference window size for any flow dependence due to \( r \).

**Proof:** This is because the predicted reference window size for any flow dependence should be no smaller than the minimum required memory size to carry the computation for that dependence. The predicted reference window size for the \( k \)th static write reference \( r \) in \( L_i \) should be no smaller than the memory size to carry the computation for all flow dependences due to \( r \). □

Theorem 1. Minimizing memory requirement is equivalent to minimizing the predicted reference window size for all flow dependences due to local array regions.

**Proof.** By Definition 5 and Lemma 1. □

In this paper, \( \langle \rangle \) denotes the inner product of \( u \) and \( v \). Given a dependence \( \pi \) with the distance vector \( dv = (d_1, d_2, \ldots, d_n) \) after loop shifting, \( \langle dv \rangle \) is the distance dependence for \( \pi \) after loop coalescing but before loop fusion, which we also call the coalesced dependence distance. Due to Assumption 3, \( \langle dv \rangle \) also represents the predicted reference window size of \( \pi \) both in the coalesced iteration space and in the original iteration space.

**Lemma 2.** Loop fusion is legal if and only if all coalesced dependence distances are nonnegative.

**Proof.** This is to preserve all the original dependences. □

We now use loop node \( L_3 \) in Figure 4(c) to illustrate how to compute the size of the predicted reference window for one particular static write reference. In this example, the predicted reference window size of \( L_2 \) due to the static write reference \( ZB(J, K) \) is the same as the predicted reference window size of \( L_2 \). There exist two dependence edges from \( L_2 \) to \( L_3 \), one L-edge and one M-edge, with distance vectors \((-1, 0) \) and \((0, 0) \). There also exist two dependence edges from \( L_2 \) to \( L_4 \), one L-edge and one M-edge. Let

\[
\begin{align*}
\mathbf{d}v_1 &= \mathbf{p}(L_3) + (-1, 0) - \mathbf{p}(L_2) \geq 0 \\
\mathbf{d}v_2 &= \mathbf{p}(L_3) + (-1, 0) - \mathbf{p}(L_2) \geq 0 \\
\mathbf{d}v_3 &= \mathbf{p}(L_4) + (0, 0) - \mathbf{p}(L_2) \geq 0 \\
\mathbf{d}v_4 &= \mathbf{p}(L_4) + (0, 0) - \mathbf{p}(L_2) \geq 0
\end{align*}
\]

Note that loop shifting and coalescing is always legal. To make loop fusion legal, the following constraint is enforced:

\[ \langle dv \rangle \geq 0, 1 \leq i \leq 4 \]

The constraint (5) guarantees that the coalesced dependence distance is nonnegative for all dependences after loop shifting and coalescing but before loop fusion.

\[ \langle dv \rangle \] represents the predicted reference window size for the flow dependence from \( L_2 \) to \( L_3 \), and \[ \langle dv \rangle \] for the predicted
reference window size for the flow dependence from $L_2$ to $L_4$. The size of the predicted reference window of $L_0$ can be computed by taking the greater one of the above two reference window sizes, i.e., $\max(\delta d_{v3}, \delta d_{v4})$, according to Lemma 1.

Next, we formulate the objective function for memory reduction to minimize the size of local array regions.

3. OBJECTIVE FUNCTION

In this section, we first formulate a graph-based system to minimize the predicted reference window size, thus minimizing the total memory requirement. We then transform our problem to a network flow problem, which is solvable in polynomial time.

Given a loop dependence graph $G$, the objective function to minimize the size of the predicted reference window for all loop nests can be formulated as follows. ($e =< L_i, L_j >$ is an edge in $G$)

$$\min(\Sigma_{v=1}^{m}, \Sigma_{k=1}^{r} \delta M_{i,k}) \tag{6}$$

subject to

$$\delta(\vec{p}(L_j) + \delta v(e) - \vec{p}(L_i)) \geq 0, \forall L$-edge $e \tag{7}$$

$$\delta M_{i,k} \geq \delta(\vec{p}(L_j) + \delta v(e) - \vec{p}(L_i)), \forall M$-edge $e, 1 \leq k \leq \tau_i \tag{8}$$

We call the above defined system as Problem 1. In (8), $\delta M_{i,k}$ represents the predicted reference window size for the local array regions due to the $k$th static write reference in $L_i$.

Constraint (7) says that the coalesced dependence distance must be nonnegative for all $L$-edges after loop coalescing but before loop fusion. Constraint (8) says that the predicted reference window size, $\delta M_{i,k}$, must be no smaller than the predicted reference window size for every $M$-edge originated from $L_i$ and due to the $k$th static write reference in $L_i$.

Combining the constraint (7) and Assumption 2, the following lemma says that the coalesced dependence distance is also nonnegative for all $M$-edges.

**Lemma 3.** If the constraint (7) holds, $\delta(\vec{p}(L_j) + \delta v(e) - \vec{p}(L_i)) \geq 0$ holds for all $M$-edges $e =< L_i, L_j >$ in $G$.

**Proof.** If $i = j$, we have $\delta(\vec{p}(L_j) + \delta v(e) - \vec{p}(L_i)) = \delta v(e)$. If $\delta v(e) = 0$, then $\delta v(e) = 0$ holds. Otherwise, assume that the first non-zero component of $\delta v(e)$ is the $k$th component. Based on Assumption 2, we have $\delta v(e) \geq \delta(d(1, \ldots, 0, 1, \ldots, -\frac{1}{2} q^{(k+1)} + 1, \ldots, -\frac{1}{2} q^{(m+1)}) > 0$.

For an $M$-edge $e_2 =< L_i, L_j >, i \neq j$, there must exist an $L$-edge $e_1 =< L_i, L_j >$. The constraint (7) guarantees that $\delta(\vec{p}(L_j) + \delta v(e_2) - \vec{p}(L_i)) \geq 0$ holds. We have $\delta(\vec{p}(L_j) + \delta v(e_2) - \vec{p}(L_i)) = \delta(\vec{p}(L_j) + \delta v(e_1) - \vec{p}(L_i)) + \delta v(e_2) - \delta v(e_1)) \geq \delta v(e_2) - \delta v(e_1))$.

By the definition of $L$-edges and $M$-edges, we have $\delta v(e_2) - \delta v(e_1) \geq 0$. Similar to the proof for the case of $i = j$ in the above, we can prove that $\delta(\delta v(e_2) - \delta v(e_1)) \geq 0$ holds. $\square$

From the proof of Lemma 3, we can also see that for any dependence $\sigma$ which is eliminated during our simplification process in Section 2.4, its coalesced dependence distance is also nonnegative, given that the constraint (7) holds. Hence, the coalesced dependence distances for all the original dependences (before simplification in Section 2.4) are nonnegative, after loop shifting and coalescing but before loop fusion. Loop fusion is legal according to Lemma 2.

In Section 2.4, we know that for any flow dependence edge $e_3$ from $L_i$ to $L_j$ due to the static write reference $\tau$ which is eliminated during the simplification process, there must exist an $M$-edge $e_4$ from $L_i$ to $L_j$ due to $\tau$. From the proof of Lemma 3, $\delta(\vec{p}(L_j) + \delta v(e_4) - \vec{p}(L_i)) \geq \delta(\vec{p}(L_j) + \delta v(e_3) - \vec{p}(L_i))$ holds. Hence, the constraint (8) computes the predicted reference window size, $\delta M_{i,k}$, over all flow dependences originated from $L_i$ due to the $k$th static write reference in the unsimplified loop dependence graph (see Section 2.2). According to Lemma 1, the constraint (8) correctly computes the predicted reference window size, $\delta M_{i,k}$.

### 3.1 Transforming the Original Problem

We define a new problem, Problem 2, by adding the following two constraints to Problem 1. ($e =< L_i, L_j >$ is an edge in $G$)

$$\vec{p}(L_j) + \delta v(e) - \vec{p}(L_i) \geq 0, \forall L$-edge $e \tag{9}$$

$$M_{i,k} \geq \vec{p}(L_j) + \delta v(e) - \vec{p}(L_i), \forall M$-edge $e, 1 \leq k \leq \tau_i \tag{10}$$

In the following, we show that given an optimal solution for Problem 1, we can construct an optimal solution for Problem 2 with the same value for the objective function (6), and vice versa.

**Lemma 4.** Given any optimal solution for Problem 1, we can construct an optimal solution for Problem 2, with the same value for the objective function (6).

**Proof.** The search space of Problem 2 is a subset of that of Problem 1. Given an LDG $G$, the optimal objective function value (6) for Problem 2 must be equal to or greater than that for Problem 1. Given any optimal solution for Problem 1, we find the shifting factor ($\vec{p}$) and $M_{i,k}$ values, for Problem 2 as follows.

1. Initially let $\vec{p}$ and $M_{i,k}$ values from Problem 1 be the solution for Problem 2. In the following steps, we will adjust these values so that all the constraints for Problem 2 are satisfied and the value for the objective function (6) is not changed.

2. If all $\vec{p}$ values satisfy the constraint (9), go to step 4. Otherwise, go to step 3.
3. This step finds \(\Pi\) values which satisfy the constraint (9).

Following the topological order of nodes in \(G\), find the first node \(L_i\) such that there exists an \(L\)-edge \(e = (L_i, L_j)\) where the constraint (9) is not satisfied. (Here we ignore self cycles since they must represent \(M\)-edges in \(G\).) Suppose \(\delta' = p(L_i) + \delta v(e) - p(L_j) = (0, \ldots, 0, c_1, \ldots)\) where \(c_1 < 0\) is the \(s\)th and the first nonzero component of \(\delta'\). Let \(\delta = (0, \ldots, 0, -c_1, c_1\delta^{(s+1)}, 0, \ldots)\) where the only two nonzero components are the \(s\)th and the \((s+1)\)th. Change \(p(L_j)\) by \(\delta p(L_j)\) to \(p(L_j) + \delta\). Because of \(\delta \delta = 0\), the new \(\Pi\) values, including \(\delta p(L_j)\), satisfy the constraints (7) and (8). The value for the objective function (6) is also not changed.

If \(\delta p(L_j) + \delta v(e) - p(L_i)\) is still lexicographically negative, we can repeat the above process. Such a process will terminate within at most \(n\) times since otherwise the constraint (7) would not hold for the optimal solution of Problem 1.

Note that the node \(L_i\) is selected based on the topological order and the shifting factor \(\delta p(L_j)\) is increased compared to its original value. For any \(L\)-edge with the destination node \(L_j\), if the constraint (9) holds before updating \(p(L_j)\), it still holds after the update. Such a property will guarantee our process to terminate.

Go to step 2.

1. This step finds \(\Pi_{i,k}\) values which satisfy the constraint (10).

Given 1 \(\leq i \leq m\) and 1 \(\leq k \leq r_1\), find the \(\Pi_{i,k}\) value which satisfies the constraint (10) such that the constraint (10) becomes equal for at least one edge.

If \(\Pi_{i,k}\) achieved above satisfies the constraint (8), we are done. Otherwise, we increase the \(r_i+k\) component of the \(\Pi_{i,k}\) value until the constraint (8) holds and becomes equal for at least one edge.

Find all \(\Pi_{i,k}\) values. The value for the objective function (6) is not changed.

With such \(\Pi\) and \(\Pi_{i,k}\) values, the value for the objective function (6) for Problem 2 is the same as that for Problem 1. Hence, we get an optimal solution for Problem 2 with the same value for the objective function (6). \(\square\)

**THEOREM 2.** Any optimal solution for Problem 2 is also an optimal solution for Problem 1.

**Proof.** Given any optimal solution of Problem 2, we take its \(\Pi\) and \(\Pi_{i,k}\) values as the solution for Problem 1. Such \(\Pi\) and \(\Pi_{i,k}\) values satisfy the constraints (7)-(8), and the value for the objective function (6) for Problem 1 is the same as that for Problem 2. Such a solution must be optimal for Problem 1. Otherwise, we can construct from Problem 1 another solution of Problem 2 which has lower value for the objective function (6), according to Lemma 4. This contradicts to the optimality of the original solution for Problem 2. \(\square\)

---

**Figure 5:** The transformed graph \((G_1)\) for Figure 4(c)

Based on Theorem 2, given an optimal solution for Problem 2, we immediately have an optimal solution for Problem 1. In the rest of this section, we try to solve Problem 2 instead.

By expanding the vectors in Problem 2, an integer programming (IP) problem results. General solutions for IP problems, however, do not take the LDG graphical characteristics into account. Instead of solving the IP problem, we transform it into a network flow problem, as discussed in the rest of this section.

3.2 Transforming Problem 2

Given a loop dependence graph \(G\), we generate another graph \(G_1 = (V_1, E_1)\) as follows.

- For any node \(L_i \in G\), create a corresponding node \(\hat{L}_i\) in \(G_1\).
- For any node \(L_i \notin G\), if \(L_i\) has an outgoing \(M\)-edge, let the weight of \(\hat{L}_i\) be \(u(\hat{L}_i) = -r_1 \hat{\sigma}\). For each static write reference \(r_k\) (1 \(\leq k \leq r_1\)) in \(L_i\), create another node \(\hat{L}_i^{(k)}\) in \(G_1\), which is called the sink of \(\hat{L}_i\) due to \(r_k\). Let the weight of \(\hat{L}_i^{(k)}\) be \(u(\hat{L}_i^{(k)}) = \hat{\sigma}\).
- For any node \(L_i \in G\) which does not have an outgoing \(M\)-edge, let the weight of \(\hat{L}_i\) be \(\hat{\sigma}\).
- For any \(M\)-edge \(< L_i, L_j >\) in \(G\) due to the static write reference \(r_k\), suppose its distance vector \(\delta v\). Add an edge \(< \hat{L}_i, \hat{L}_j^{(k)} >\) to \(G_1\) with the distance vector \(-\delta v\).
- For any \(L\)-edge \(< L_i, L_j >\) in \(G\), suppose its distance vector \(\delta v\). Add an edge \(< \hat{L}_i, \hat{L}_j >\) to \(G_1\) with the distance vector \(\delta v\).

For the original graph in Figure 4(c), Figure 5 shows the transformed graph.

We associate a vector \(\delta\) to each node in \(G_1\) as follows.

- For each node \(\hat{L}_i \in G_1\), \(\delta_i = \hat{p}(\hat{L}_i)\).
- For each node \(\hat{L}_i^{(k)} \in G_1\), \(\delta_i = M_{i,k} + \hat{p}(\hat{L}_i^{(k)})\).

The new system, which we call Problem 3, is defined as follows. \(e = (u_v, v_j)\) is an edge in \(G_1\) annotated by \(\delta_{\text{e}}\).
subject to
\[ \sigma(q_j - q_i + d_k) \geq 0, \forall e \]  
(13)

Theorem 3. Problem 3 is equivalent to Problem 2.

Proof. We have
\[ \sum_{i=1}^{V_1} w(v_i) q_i \]
(11)

For each edge \( e = < Li, Lj > \) in \( G_1 \), the inequality (12) is equivalent to
\[ p(L_i) - p(L_j) + v(e_1) \geq \bar{0}, \]
(14)
where \( e_1 \) is an L-edge in \( G \) from \( L_i \) to \( L_j \). Inequality (14) is equivalent to (9), hence inequality (12) is equivalent to (9).

For each edge \( e = < Li, Lj > \) in \( G_1 \), the inequality (12) is equivalent to
\[ \bar{M}_{i,k} + \bar{p}(L_i) - \bar{p}(L_j) - d v(e_1) \geq \bar{0}, \]
(15)
where \( e_1 \) is an M-edge in \( G \) from \( L_i \) to \( L_j \) due to the kth static write reference in \( L_i \). Inequality (15) is equivalent to (10), hence inequality (12) is equivalent to (10).

Similarly, it is easy to show that the constraints (7) and (8) are equivalent to constraint (13).

Note that one edge in \( G \) could be both an L-edge and an M-edge, which corresponds to two edges in \( G_1 \). Assumption 2 can derive the following inequality for the transformed graph \( G_1 \):
\[ \sum_{k=1}^{E_1} |d v(e_k)| < \frac{1}{2} \bar{d}, \]
(16)
where \( e_k \in E_1 \) is annotated with the dependence distance vector \( d v(e_k) \).

If we consider the vector as the basic computation unit, Problem 3 is a nonlinear system, due to the constraint (13). The same as Problem 2, such a nonlinear system can be solved by linearizing the vector representation so that the original problem becomes an integer programming problem, which in its general form, is NP-complete. In the next, however, we show that we can achieve an optimal solution in polynomial time for Problem 3 by utilizing the network flow property.

### 3.3 Optimality Conditions

We develop optimality conditions to solve Problem 3. We utilize the network flow property. A network flow consists of a set of vectors such that each vector \( f(e_i) \) corresponds to each edge \( e_i \in E_1 \) and for each node \( v_i \in V_1 \), the sum of flow values from all the in-edges should be equal to \( w(v_i) \) plus the sum of flow values from all the out-edges. That is,
\[ \sum_{k=1}^{V_1} w(v_i) q_i = \sum_{k=1}^{E_1} f(e_k) (q_j - q_i) \]
(17)
where \( e_k = < v_i, v_j > \) represents an in-edge of \( v_i \) and \( e_k = < v_i, v_j > \) represents an out-edge of \( v_i \).

Lemma 5. Given \( G_1 = (V_1, E_1) \), there exists at least one legal network flow.

Proof. Find a spanning tree \( T \) of \( G_1 \). Assign the flow value to be \( \bar{0} \) for all the edges not in \( T \). Hence, if we can find a legal network flow for \( T \), the same flow assignment is also legal for \( G_1 \).

We assign flow value to the edges in \( T \) in reverse topological order. Since the total weight of the nodes in \( T \) is equal to \( \bar{0} \), a legal network flow exists for \( T \).

Based on equation (17), given a legal network flow, we have
\[ \sum_{i=1}^{V_1} w(v_i) q_i = \sum_{k=1}^{E_1} f(e_k) (q_j - q_i) \]
(18)
where \( e_k = < v_i, v_j > \in E_1 \).

For any node \( v \in V_1 \), we have \( w(v) = c \bar{d} \), where \( c = -r_i, 0 \) or 1. For our network flow algorithm, we abstract out the factor \( \bar{d} \) from \( w(v) \) such that \( w(v) \) is represented by \( c \) only. Such an abstraction will give each flow value the form \( f(e_k) = c_k \bar{d} \), where \( c_k \) is an integer constant.

Suppose \( f(e_k) \geq \bar{0} \) for the edge \( e_k \in E_1 \), which is equivalent to \( c_k \geq 0 \). With the constraint (13), we have
\[ f(e_k) (q_j - q_i) + d_k = \bar{c}_k (q_j - q_i) + \bar{d}_k \geq 0. \]
(19)

Hence, we have
\[ f(e_k) (q_j - q_i) \geq -f(e_k) \bar{d}_k. \]
(20)

Therefore, with the equation (18), if \( f(e_k) \geq \bar{0} \), we have
\[ \sum_{i=1}^{V_1} w(v_i) q_i \geq -\sum_{k=1}^{E_1} f(e_k) \bar{d}_k. \]
(21)

Collectively, we have the optimality conditions stated as the following theorem such that if they hold, the inequality (21) becomes the equality and the optimality is achieved for Problem 3.

Theorem 4. If the following three conditions hold,
1. Constraints (12) and (19) are satisfied, and

2. A legal network flow \( f(e_k) = c_k \sigma \) exists such that \( c_k \geq 0 \) for \( 1 \leq k \leq |E_1| \), and

3. \( \Sigma_{k=1}^{|E_1|} w(u_i) \delta_i = -\Sigma_{k=1}^{|E_1|} f(e_k) \delta_k \) holds, i.e., inequality (20) becomes an equality.

Problem 3 achieves an optimal solution \( -\Sigma_{k=1}^{|E_1|} f(e_k) \delta_k \).

**Proof.** Obvious from the above discussion. \( \square \)

### 3.4 Solving Problem 3

Here, let us consider each vector \( w(u_i), \delta_i \) and \( \delta_k \) as a single computation unit. Based on the duality theory [24, 2], Problem 3, excluding the constraint (13), is equivalent to

\[
\max |E_1| (-f(e_k) \delta_k)
\]

subject to

\[
\Sigma_{k=1}^{|E_1|} w(u_i) \delta_i = w(u_i) + \Sigma_{k=1}^{|E_1|} f(e_k) \delta_k, \quad 1 \leq i \leq |V_i|.
\]

\[
f(e_i) \geq 0, \quad 1 \leq i \leq |E_1|.
\]

The constraint (13) is mandatory for the equivalence between Problem 3 and its dual problem, following the development of optimality conditions in Section 3.3 [1]. The constraint (23) in the dual system precisely defines a flow property, where each edge \( e_i \) is associated with a flow vector \( f(e_i) \). We define Problem 4 as the system by (11)-(12) and (22)-(24). Similar to \( w(u_i) \), the vector \( f(e_i) \) is represented by \( c_k \) where \( f(e_k) = c_k \sigma \). Although apparently the search space of Problem 4 encloses that of Problem 3, Problem 4 has correct solutions only within the search space defined by Problem 3.

Based on the property of duality, Problem 4 achieves an optimal solution if and only if

- The constraints (13), (23) and (24) hold, and
- The objective function values for (11) and (22) are equal, i.e., \( \Sigma_{i=1}^{|V_i|} w(u_i) \delta_i = -\Sigma_{k=1}^{|E_1|} f(e_k) \delta_k \) holds.

If we can prove that the constraint (13) holds for the optimal solution of Problem 4, such a solution must also be optimal for Problem 3, according to Theorem 4.

There exist plenty of algorithms to solve Problem 4 [1, 2]. Although those algorithms are targeted to the scalar system (the vector length equals to 1), some of them can be directly adapted to our system by vector summation, subtraction and comparison operations. In [2], the authors present a network simplex algorithm, which can be directly utilized to solve our system. The algorithmic complexity, however, is exponential in the worst case in terms of the number of nodes and edges in \( G_1 \). In [1], the authors present several graph-based polynomial-time algorithms, for example, successive shortest path algorithm with the complexity \( O(|V_1|^3) \), double alternating algorithm with the complexity \( O(|V_1||E_1| \log |V_1|) \), and so on. From [1], the current fastest polynomial-time algorithm for solving network flow problem is enhanced capacity scaling algorithm with the complexity \( O((|E_1| \log |V_1|)(|E_1| + \log |V_1|)) \). For these algorithms, we have the following lemma.

**Lemma 6.** For any optimal solution of \( \delta_i \) in Problem 4, there exists a spanning tree \( T \) in \( G_1 \) such that each edge \( e = < u_i, v_j > \) in \( T \) satisfies \( \delta_i - \delta_j + \delta_k = 0 \).

**Proof.** This is true due to the foundation of the simplex method [2]. \( \square \)

Let \( T \) be the spanning tree in Lemma 6. If we fix any \( \delta_i \) to be \( 0 \), all \( \delta_i, 1 \leq i \leq |V_i| \), can be determined uniquely. With such uniquely-determined \( \delta_i \), we have

\[
|\delta_i| \leq 1 |E_1| |\delta_k| , \quad 1 \leq i, j \leq |V_i|.
\]

For any \( e = < u_i, v_j > \in E_1 \) with annotation \( \delta_k \), with the inequality (25), we have

\[
|\delta_i - \delta_j| \leq |\delta_i| + |\delta_j| \leq 2 |E_1| |\delta_k| .
\]

For the inequality (26), based on the inequality (16), we have

\[
|\delta_i - \delta_j| < \beta, e = < u_i, v_j > \in E_1 \text{ is annotated with } \delta_k.
\]

**Lemma 7.** \( \delta_i - \delta_j + \delta_k \geq 0 \), where \( e = < u_i, v_j > \in E_1 \) is annotated with \( \delta_k \).

**Proof.** If \( \delta_i - \delta_j + \delta_k = 0 \), then \( \delta_i - \delta_j + \delta_k \geq 0 \) holds.

Otherwise, assume the first non-zero component is the \( h \)th for \( \delta_i - \delta_j + \delta_k \). Then, \( q^{(s)}_i - q^{(s)}_j + q^{(s)}_k = 0, 1 \leq s \leq h - 1, \) and \( q^{(h)}_i - q^{(h)}_j + q^{(h)}_k > 0 \).

With the constraint (27), we have

\[
\delta_i - \delta_j + \delta_k \geq \sigma(0, \ldots, 0, q^{(h)}_i - q^{(h)}_j + q^{(h)}_k, -q^{(h+1)}_i + q^{(h+1)}_j + q^{(h+1)}_k, \ldots, -q^{(h)}_i + q^{(h)}_j) + \sigma^{(n)}
\]

\[
= \sigma^{(h)} q^{(h)}_i - q^{(h)}_j + q^{(h)}_k - \sigma^{(h+1)} q^{(h+1)}_i + q^{(h+1)}_j + \sigma^{(h+1)} q^{(h+1)}_k - \ldots - \sigma^{(h)} q^{(h)}_i + q^{(h)}_j + \sigma^{(n)}
\]

\[
= \sigma^{(h)} q^{(h)}_i - q^{(h)}_j + q^{(h)}_k - \sigma^{(h)} q^{(h)}_i + q^{(h)}_j + \sigma^{(n)}
\]

\[
> 0 \quad \square
\]

Hence, Inequality (16) guarantees that the constraint (13) always holds when the optimality of Problem 4 is achieved. The optimal solution for Problem 4 is also an optimal solution for Problem 3.
It may also be important to avoid fusing attoo many loop

levels if loops are shifted. This is because, after loop shifting, fusing too many loop levels can potentially increase the number of operations due to the IF-statements added in the loop body or due to the effect of loop peeling. Coalescing, if applied, may also introduce more subscript computation overhead. Although all such costs tend to be less significant than the costs of cache misses and register spills, we carefully control the fusion of innermost loops. If the rate of increased operations after fusion exceeds a certain threshold, we only fuse the outer loops.

4.2 Enabling Loop Transformations

We use several well-known loop transformations to enable effective fusion. Long backward data-dependence distances make loop fusion ineffectivememor y reduction. Such long distances are sometimes due to incompatible loops [27] which can be corrected by loop interchange. Long backward distances may also be due to circular data dependences which can be corrected by circular loop skewing [27]. Furthermore, our technique applies loop distribution to a node, \( L_i \), if the dependence distance vectors originated from \( L_i \) are different from each other. In this case, distributing the loop may allow different shifting factors for the distributed loops, potentially yielding a more favorable result.

4.3 Tiling vs. Reduction

Suppose the collection of loops in Figure 2(a) are embedded in another loop, \( T \), such that the memory reference foot-
To tile the T loop nest, certain arrays may be duplicated [27]. Let $W_1$ represent the array footprint size of the loop body after the array duplication phase (in the number of data elements) but before tiling. ($W_2$ may then be greater than $W_1$.) The average cache miss penalty for each T-iteration after tiling will depend on the number of inner-loop levels which are tiled. Based on a detailed calculation [26], we derive the average miss penalty per T-iteration under two-level tiling as

$$p_1 S_1 W_2 \frac{C_{b1}}{B_1} + p_2 S_2 W_2 \frac{C_{b2}}{B_2}$$  \hspace{1cm} (30)$$

where $S_1$ and $S_2$ represent the skew factors of tiling and $(B_1, B_2)$ represent the tile size.

The average miss penalty per T-iteration under one-level tiling is estimated as

$$p_1 W_2 \frac{C_{b1}}{C_{b2}} + p_2 S_1 W_2 \frac{C_{b2}}{B_1}.$$  \hspace{1cm} (31)$$

Which transformation to choose is then determined by a comparison of the estimated cache miss penalties. Our experimental results will show that these simple cost models work quite well.

5. EXPERIMENTAL RESULTS

We have implemented our memory reduction technique in a research compiler, Panorama [12]. We implemented a network flow algorithm, successive shortest path algorithm [1]. The loop dependence graphs in our experiments are relatively simple. The successive shortest path algorithm takes less than 0.06 seconds for each of all the benchmarks. To measure its effectiveness, we tested our memory reduction technique on 20 benchmarks on a SUN Ultra II uniprocessor workstation and on a MIPS R10K processor within an SCI Origin 2000 multiprocessor. The Ultra II processor has a 16KB directly-mapped L1 data cache with a 16-byte cache line, and it has a 2MB directly-mapped unified L2 cache with a 64-byte cache line. The cache miss penalty is 6 machine cycles for the L1 data cache and 45 machine cycles for the L2 cache. The MIPS R10K has a 32KB 2-way set-associative L1 data cache with a 32-byte cache line, and it has a 4MB 2-way set-associative unified L2 cache with a 128-byte cache line. The cache miss penalty is 9 machine cycles for the L1 data cache and 68 machine cycles for the L2 cache.

5.1 Benchmarks and Memory Reduction

Table 1 lists the benchmarks used in our experiments, their descriptions and their input parameters. In the table, "m/n" represents the number of loops in the loop sequence (m) and the maximum loop nesting level (n). Note that the array size and the iteration counts are chosen arbitrarily for L1L14, L1L18 and Jacobi. To differentiate two versions of swain in SPEC95 and SPEC2000, we call the SPEC95 version as swain and the SPEC2000 version as swain2. swain2 is almost identical to swain except for its larger data size. For combustion, we change the array size (N1 and N2) from 1 to 10, so the execution time will last for several seconds. Programs climate, laplace-JB, laplace-ge and all the Pardue set problems are from an HPF benchmark suite at Rice University [20].
Table 1: Test programs

<table>
<thead>
<tr>
<th>Benchmark Name</th>
<th>Description</th>
<th>Input Parameters</th>
<th>m/n</th>
</tr>
</thead>
<tbody>
<tr>
<td>LL14</td>
<td>Livermore Loop No. 14</td>
<td>N = 1000, ITMAX = 60000</td>
<td>3/1</td>
</tr>
<tr>
<td>LL18</td>
<td>Livermore Loop No. 18</td>
<td>N = 400, ITMAX = 100</td>
<td>2/2</td>
</tr>
<tr>
<td>Jacobi</td>
<td>Jacobi Kernel w/o convergence test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tomcatv</td>
<td>A mesh generation program from SPEC95fp</td>
<td>reference input</td>
<td>5/1</td>
</tr>
<tr>
<td>swim1</td>
<td>A weather prediction program from SPEC95fp</td>
<td>reference input</td>
<td>2/2</td>
</tr>
<tr>
<td>swim2</td>
<td>A weather prediction program from SPEC95fp</td>
<td>reference input</td>
<td>2/2</td>
</tr>
<tr>
<td>hydro2d</td>
<td>An astrophysical program from SPEC95fp</td>
<td>reference input</td>
<td>10/2</td>
</tr>
<tr>
<td>lucas</td>
<td>A prominence test from SPEC95fp</td>
<td>reference input</td>
<td>3/1</td>
</tr>
</tbody>
</table>
| mg             | A multigrid solver from NPB2.3-serial benchmark                             | Clasp 

For each of the benchmarks, we examine three versions of the code, i.e. the original one, the one after loop fusion but before array contraction, and the one after array contraction. For all versions of the benchmarks, we use the native Fortran compilers to produce the machine codes. On the Ultra II, we follow the recommendations from SUN's optimizing compiler group and use the following optimization flags. For the original tomcatv code, we use "-fast -xchip=ultra2 -xarch=v8plusa -xpad=local:23". For all versions of swim1 and swim2, we use "-fast -xchip=ultra2 -xarch=v8plusa -xpad=common:15". For all versions of combustion, we simply use "-fast" because it produces better-performing codes than using other flags. For all other codes, we use the flag...
"-fast -xchip=ultra2 -xarch=v8plusa -fsimple=2". When we compare the best results of different versions, we switch on and off prefetching (i.e. the "-xprefetch" flag) and pick the better result for each version. On the R10K, we simply use the optimization flag "-O3" except with the following adjustments. We switch off prefetching for laplace-jb, software pipelining for laplace-gs and loop unrolling for purdue-03. For aviml and avin2, the native compiler fails to insert prefetch instructions in the innermost loop body after memory reduction. We manually insert prefetch instructions into the three key innermost loop bodies, following exactly the same prefetching patterns used by the native compiler for the original codes.

Figure 9 compares the code sizes and the data sizes of the original and the transformed codes on the Ultra II. The data size shown for each original program is normalized to 100. The actual data size varies greatly for different benchmarks, which are listed in the table associated with the figure. Similarly, Figure 10 compares the data sizes and the code sizes on the R10K. For mg and climate, the memory requirement differs little before and after the program transformation. For svrnl and bin2, the native compiler fails to insert prefetch instructions in the innermost loop body after memory reduction. We manually insert prefetch instructions into the three key innermost loop bodies, following exactly the same prefetching patterns used by the native compiler for the original codes.

5.2 Performance

Figure 11 compares the normalized execution time on the Ultra II, where "Mid" represents the execution time of the codes after loop fusion but before array contraction, and "Final" represents the execution time of the codes after array contraction. Similarly, Figure 12 compares the normalized execution time on the R10K. The geometric mean of speedup after memory reduction is 1.57 for all benchmarks on both machines. Specifically, the geometric mean is 1.57 on the Ultra II alone, and it is 1.41 on the R10K. For several small purdue benchmarks, the reduction rate is almost 100%.

5.3 Memory Reference Statistics

To further understand the effect of memory reduction on the performance, we examined the cache behavior of different versions of the tested benchmarks. We measured the reference count (dynamic load/store instructions), the miss count of the L1 data cache, and the miss count of the L2 unified cache on both machines. We use the perfex package on the MIPS R10K and the perfmon package on the Ultra II to get the cache statistics. Figures 13 and 14 compare such statistics on the Ultra II, where the total reference counts in the original codes are normalized to 100. Similarly, Figures 15 and 16 compare the statistics on the R10K.

We see three programs actually get slowed down slightly after memory reduction. The execution time of both purdue-13 and laplace-gs on the Ultra II is increased by 2%. The execution time of purdue-08 on the R10K is increased by 1%. Both purdue-08 and purdue-13 make several math library function calls which have dominated the execution time. For laplace-gs, loop peeling is applied which may reduce the effectiveness of scalar replacement, and increase the number of total memory references.

Gao et al. propose to perform array contraction enabled by loop fusion only [10]. With their technique, the geometric mean of speedup after array contraction is 1.30 for all other combinations of benchmarks and machines.
both machines, is reduced by 21.1% after memory reduction. Specifically, the reduction rate is 20.0% on the Ultra II alone, and it is 22.3% on the R10K alone. However, in a few cases, the total reference counts get increased instead. We examined the assembly codes and found a number of reasons:

1. The fused loop body contains more scalar references in a single iteration than before fusion. This increases the register pressure and sometimes causes more register spilling.

2. The native compilers can perform scalar replacement [3] for references to noncontracted arrays. The fused loop body may prevent such scalar replacement for two reasons:
   - If register pressure is high in a certain loop, the native compiler may choose not to perform scalar replacement.
   - After loop fusion, the array dataflow may become more complex, which then may defeat the native compiler in its attempt to perform scalar replacement.

3. Loop peeling may decrease the effectiveness of scalar replacement since fewer loop iterations benefit from it.

Despite the possibility of increased memory reference counts in a few cases due to the above reasons, Figures 13 to 16 show that cache misses are generally reduced by memory reduction. The total number of cache misses, counting all benchmarks on both machines, is reduced by 55.0% after memory reduction. Specifically, the reduction rate is 28.6% on the Ultra II alone, and it is 63.8% on the R10K alone. The total number of L1 data cache misses, counting all benchmarks on both machines, is reduced by 57.3% after memory reduction. Specifically, the reduction rate is 27.5% on the Ultra II alone, and it is 63.0% on the R10K alone. The improved
Figure 17: Performance of the original programs w/ and w/o prefetching on the Ultra II

Figure 18: Performance of the transformed programs w/ and w/o prefetching on the Ultra II

Figure 19: Performance of the original programs w/ and w/o prefetching and software pipelining on the R10K

Figure 20: Performance of the original programs w/ and w/o prefetching and software pipelining on the R10K (cont.)

Figure 21: Performance of the transformed programs w/ and w/o prefetching and software pipelining on the R10K

5.4 Interaction with Other Compiler Optimizations

In this subsection, we examine how our memory reduction technique affects prefetching, software pipelining, register allocation and unroll-and-jam. The issue of concern is whether the memory reduction makes other compiler optimizations suffer. A performance comparison with loop tiling is also presented.

Prefetching and Software Pipelining

On the R10K, we compared the performance impact of prefetching and software pipelining on both the original codes and the transformed codes. On the Ultra II, we compared the performance impact of prefetching only, since we cannot specifically switch off software pipelining alone for the native compiler.

Figures 17 and 18 show the normalized execution time with and without prefetching, on the Ultra II, for the original programs and the transformed programs respectively. Prefetching affects the performance little for the transformed codes except tomcatv. Figures 19 and 20 show the normalized cache performance seems to often have a bigger impact on execution time than the total reference counts.
Figure 22: Performance of the transformed programs w/ and w/o prefetching and software pipelining on the R10K (cont.)

Figure 23: Performance of the code with (unroll-and-jam, scalar replacement) on the Ultra II

Figure 24: Performance of the code with (unroll-and-jam, scalar replacement) on the R10K

execution time for the original programs, with and without prefetching and software pipelining, on the R10K. Figures 21 and 22 show the normalized execution time for the transformed programs. Software pipelining and prefetching improves the performance for the transformed codes in most cases. One exception is that \texttt{laplace-jb} with prefetching, where prefetching makes performance worse by 49.4%. A close look at cache statistics with \texttt{perfex} shows that prefetching increases the L1 cache miss count by 50% compared with the code without prefetching. Another exception is that \texttt{laplace-gs} with software pipelining, where software pipelining makes performance worse by 20.7%. Based on the results from \texttt{perfex} software pipelining generates 59% more floating point instructions than without software pipelining (336M vs. 337M).

\textbf{Register Allocation}

As stated earlier in this section, loop fusion may potentially increase register pressure and thus may potentially reduce register reuse. Figures 13 and 14 show that, in 5 of the 20 codes transformed for the Ultra II, slightly more memory references are issued than the original codes. Loop fusion seems to have degraded register reuse somewhat in those codes. However, we should point out that, except in three cases, (\texttt{laplace-gs} on the Ultra II and \texttt{swim1} and \texttt{swim2} on the R10K), all those transformed codes in question actually run faster than their original codes. Nonetheless, it is useful to examine the register matter further.

One interesting question is whether the seemingly degraded register utilization is truly due to the increased register pressure. Alternatively, it might be due to the native compiler's inability to properly perform scalar replacement and unroll-and-jam on the fused loop body. (These two techniques are important to good register allocation.) To find the answer, we manually applied unroll-and-jam and scalar replacement to the codes of concern. We experimented with unroll factors from 1 to 4 (a factor of 1 meaning no unrolling), and we applied scalar replacement where possible. We then picked the best results. Figures 23 and 24 show the results on the Ultra II and on the R10K respectively, where "Org" stands for the original code, "Org-Unroll" for the original code with unroll-and-jam plus scalar replacement manually applied, "Trans" stands for the transformed code and "Trans-Unroll" for the transformed code with unroll-and-jam plus scalar replacement applied. From these figures, we conclude that loop fusion indeed increases register pressure somewhat, as unroll-and-jam and scalar replacement applied manually do not seem to make much difference, before or after memory reduction.

\textbf{Compare with Tiling}

As stated in Section 4.3, for certain loop sequences, both tiling and memory reduction may be applied profitably. In our benchmarks, we have \texttt{LL18}, \texttt{Jacobi}, \texttt{tomcatv}, \texttt{swim1} and \texttt{swim2} which can be tiled profitably. Table 2 compares the performance between memory reduction and tiling.

Using the cost estimation in Section 4.3, our research compiler chooses 2-level tiling for \texttt{Jacobi} on both machines. It chooses 1-level tiling for \texttt{LL18}, \texttt{swim1} and \texttt{swim2} and chooses
memory reduction for tomtom, also on both machines. This turns out to be correct in 9 out of the 10 cases. The exception is LL18 on the RIOK. The tiled assembly code of LL18 on the RIOK shows that the loop index variables of the tile-controlling loop and the time-step loop (i.e. the T loop) are spilled heavily, thus introducing significantly more load/store instructions than the code with memory reduction.

6. RELATED WORK
The work by Fraboulet et al. is the closest to our memory reduction technique [8]. Given a perfectly-nested loop, they use loop alignment to adjust the iteration space for individual statements such that the total buffer size can be minimized. Unlike ours, they only formulate the optimization problem for the 1-D case as a network flow problem, in a form different from ours. For multi-dimensional case, they apply 1-D formulation loop level by loop level. They do not present any experimental results, and they do not consider the effect of memory reduction on cache behavior and execution speed.

Callahan et al. present unroll-and-jam and scalar replacement techniques to replace array references with scalar variables to improve register allocation [3]. However, they only consider the innermost loop in a perfect loop nest. They do not consider loop fusion, neither do they consider array partial contraction. Gao and Sarkar present the collective loop fusion [10]. They perform loop fusion to replace arrays with scalars, but they do not consider partial array contraction. They do not perform loop shifting, therefore they cannot fuse loops with fusion-preventing dependences. Sarkar and Gao perform loop permutation and loop reversal to enable collective loop fusion [23]. These enabling techniques can also be used in our framework.

Lam et al. reduce memory usage for highly-specialized multi-dimensional integral problems where array subscripts are loop index variables [18]. Their program model does not allow fusion-preventing dependences. Lewis et al. propose to apply loop fusion and array contraction directly in array statement level for those array languages such as F90 [16]. The same result can be achieved if the array statements are transformed into various loops and loop fusion and array contraction are then applied in scalar level. They do not consider loop shifting in their formulation. Strunt et al. consider the minimum working set which permits tiling for loops with regular stencil of dependences [28]. Their method applies to perfectly-nested loops only. In [6], Ding indicates the potential of combining loop fusion and array contraction through an example. However, he does not apply loop shifting and does not provide formal algorithms and evaluations. Gannon et al. introduce the concept of reference window, using it to estimate the cache hit rate and to guide program optimization for a software-controlled cache [9]. They do not address the memory reduction problem.

There exist a lot of work related with loop fusion. To name a few, Kennedy and McKinley prove maximizing data locality by loop fusion is NP-hard [13]. They provide two polynomial-time heuristics. Singhal and McKinley present parameterized loop fusion to improve parallelism and cache locality [25]. They do not perform memory reduction or loop shifting. Megiddo and Sarkar use mixed integer programming to optimize weighted loop fusion for parallel programs [19]. Recently, Darte analyzes the complexity of loop fusions [8] and claims that the problem of maximum fusion of parallel loops with constant dependence distances is NP-complete when combined with loop shifting. None of these works address the issue of minimizing memory requirement for a collection of loops and their techniques are very different from ours. Manjikian and Abdelrahman present shift-and-peel [17]. They shift the loops in order to enable fusion. However, they do not consider array contraction.

7. CONCLUSION
In this paper, we present a locality enhancement technique, memory reduction, which is a combination of loop shifting, loop fusion and array contraction. We reduce the memory reduction problem to a network flow problem, which is solved optimally. (The current fastest algorithm has the complexity \(O(|E|\log|V|(|E| + \log|V|))\) where \(G = (V,E)\) is the loop dependence graph.) We propose controlled fusion to prevent excessive register spilling and cache misses which may be caused by excessive loop fusion. We develop a simple memory cost model for memory reduction. For a loop nest where both tiling and memory reduction can apply, the scheme having the smaller cost is chosen. Experimental results so far show that our technique can reduce the memory requirement significantly. At the same time, it speeds up program execution by a factor of 1.57 on average. Furthermore, the memory reduction does not seem to create difficulties for a number of other back-end compiler optimizations. We also believe that memory reduction by itself is vitally important to computers which are severely memory-constrained and to applications which are extremely memory-demanding.

8. REFERENCES


