*International Journal of Parallel Programming, Vol. 28, No. 5, 2000*

# Generation of Efficient Nested Loops from Polyhedra

Fabien Quilleré,[1] Sanjay Rajopadhye,[1] and Doran Wilde[2]

*Received August 1999; revised March 2000*

Automatic parallelization in the polyhedral model is based on affine transformations from an original computation domain (iteration space) to a target space-time domain, often with a different transformation for each variable. Code generation is an often ignored step in this process that has a significant impact on the quality of the final code. It involves making a trade-off between code size and control code simplification/optimization. Previous methods of doing code generation are based on loop splitting, however they have nonoptimal behavior when working on parameterized programs. We present a general parameterized method for code generation based on dual representation of polyhedra. Our algorithm uses a simple recursion on the dimensions of the domains, and enables fine control over the tradeoff between code size and control overhead.

**KEY WORDS:** Automatic parallelization; polyhedral model; nested loops; code generation.

## 1. INTRODUCTION

The polyhedral model is a formalism for reasoning about parallel computations, using *Systems of Affine Recurrence Equations* defined over polyhedral shaped domains. Its use in automatic parallelization of nested loops goes back to the work of Kuck,[(1)] who showed that the domain of nested loops with affine lower and upper bounds can be described in terms of a polyhedron, and the seminal work of Karp *et al.*[(2)] on scheduling systems of uniform recurrence equations. The model draws on a number of results

[1] Irisa, Campus de Beaulieu, F-35042 Rennes Cedex, Rennes, France.
[2] Brigham Young University, Provo, Utah.

from data flow analysis of static loop programs,(3) scheduling, using compact representations, of indexed computations (essentially uniform or affine recurrence equations) in the context of parallelization(2, 4) or systolic array synthesis.(5, 6) Independently of the polyhedral model, many authors have developed linear-algebraic formulations for a large number of loop transformations,(7–9) which consist of reindexing the source index domain using affine transformations and then scanning the new domain, generally in a different order.

There has been considerable work on finding good transformations to optimize various measures of performance such as time, as in the work on scheduling(4, 10) communication volume(11, 12) synchronizations, etc. However, code generation, the back-end of automatic parallelization, has received relatively less attention, even though it has a significant impact on the quality of the resulting code.

In this paper, we address the problem of generating imperative code for a set of statements defined over polyhedral context domains. The context domain of a statement is a finite union of (disjoint) polyhedral sets of index vectors, for which the statement has to be evaluated. The main idea is that since each domain is a union of disjoint polyhedra, it can be individually scanned by using a set of perfectly nested loops, which are then arranged in an appropriate textual order. This uses and generalizes the techniques for the polyhedron scanning problem, namely finding a set of nested do loops which visit each integral point of a polyhedron in the lexicographic order. This problem was first studied by Ancourt and Irigoin(13) and subsequently by many authors.

The loop body is the code implementing the evaluation of the statements. Generating code for statements (the loop body) is relatively straight forward and is not the focus of this paper. Instead, we focus on generating the loops and control code.

The code generation problem is more difficult when the domains are not disjoint. Furthermore, in the most general case, the domain of *each* statement may be a finite union of polyhedra. Typically, this necessitates imperfectly nested loops that enumerate all the index vectors of the domain.

We present an algorithm to scan a union of polyhedral domains in the lexicographic order of their indices. The method can deal with affine-by-statement and even piecewise-affine mappings, since domains can be separated into multiple sub-domains, one for each mapping.

The remainder of this paper is organized as follows. In Section 2, we first present notations and background material, including details of some of the routines we use for manipulating polyhedra. These routines are drawn from a library, Polylib, developed at IRISA.(14) Polylib is freely available under Gnu license.[3] Next, we discuss the problem of scanning a single polyhedron and recall the results of Le Verge *et al.*,(15) which exploit the dual representation used in Polylib. In Section 4, we formulate and discuss the issues involved in scanning a union of (not necessarily disjoint) polyhedra. In Section 5 we present our algorithm, first as an overview using an illustrative example and then in more detail. An important subproblem that arises is that of "sorting" polyhedra, which is also discussed there. In Section 6, we develop a number of extensions in notably to (i) tradeoffs between code-size and efficiency; (ii) nonunimodular mapping; and (iii) generation of (data) parallel code. Finally, we discuss related work and present our conclusions.

## 2. NOTATION AND BACKGROUND

We now introduce the basic mathematical notation and also the tools from the Polylib library that we need to manipulate polyhedral domains. Polylib was first developed at IRISA,(14) and later extended by a number of authors, notably at Brigham Young University and Strasbourg.

### 2.1. Basic Notation

A statement $S$ defined over a polyhedral domain $\mathscr{D}$ is called a domain-qualified statement, and we denote it $\mathscr{D} :: S$. A system of recurrence equations can always be rewritten as a set of domain-qualified statements, as shown in Fig. 1.

In our notation, upper case indices are program parameters and are not explicitly declared in the indices of a domain. As explained later, this is only a matter of notation. In the polyhedral model, the symbolic program parameters are represented by dimensions added to the polyhedral domain of each statement. In the example of Fig. 1, the domain of statement S1 may be viewed either as a two-dimensional parameterized domain or equivalently, as a single four-dimensional polyhedron.

In this representation, the set of constraints that can be expressed on the parameters is the standard one of the polyhedral model, i.e., a set of linear constraints. The constraints on the symbolic parameters can be specified as a polyhedron, whose dimension is the number of parameters.

We now review the concept of polyhedra to help in understanding the polyhedron scanning problem. A polyhedron is defined by $\mathscr{D} = \{x \mid Ax \geqslant b\}$ where $A$ and $b$ are a constant matrix and vector respectively. A parameterized polyhedron is a family $\mathscr{D}(p)$ of polyhedra, one per

[3] See http://www.irisa.fr/cosi/Polylib.

(a)

```
{i,j | 1<=i<=N; 1<=j<=M} :: S1;
{i,j | i=j; 3<=j<=N} :: S2;
```

(b)

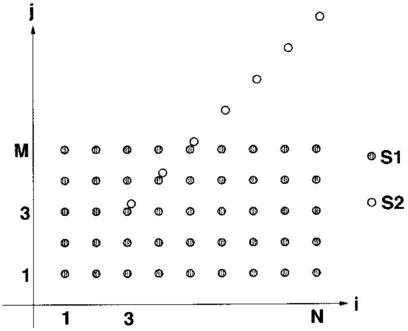


Fig. 1. (a) Systems of statements qualified by domains; (b) domains for the case where $4 \leqslant M \leqslant N$.

instance of the parameters $p$, and may be described by replacing vector $b$ with an affine combination of the parameters:

$$\mathscr{D}(p) = \{x \mid Ax \geqslant Bp + b\}$$

where $A$ and $B$ are constant matrices and $b$ is a constant vector. This parameterized polyhedron can be rewritten in the form of a nonparameterized polyhedron in the combined data and parameter index space as:

$$\mathscr{D}(p) = \left\{ x \;\middle|\; (A \quad -B) \begin{pmatrix} x \\ p \end{pmatrix} \geqslant b \right\}$$

$$\mathscr{D}' = \left\{ \begin{pmatrix} x \\ p \end{pmatrix} \;\middle|\; A' \begin{pmatrix} x \\ p \end{pmatrix} \geqslant b \right\}$$

A polyhedron may be equivalently specified either with a set of constraints (inequalities) or by its generators (vertices and rays). This description was first proposed by Motzkin *et al.*[16] Polylib is based on this double description, and uses the Chernikova algorithm[17] to transform one to the other. It turns out to be very useful for many of the operations. However, keeping the dual representation is only feasible for small dimensional domains since a $d$-polyhedron with $n$ constraints might have as many as $d^{\lfloor n/2 \rfloor}$ rays/vertices, in the worst case. We rely on the fact that computational domains tend to be relatively small polyhedra (in terms of the dimension and number of constraints) due to the fact that loops are not deeply nested. We have not experienced significant problems, related to the space inefficiency.

## 2.2. Projection of Polyhedra

The library function *DomainAddRays* joins a set of lines, rays, or points to a domain and produces the resulting domain with all redundancies eliminated. It is used in this paper to eliminate (or project out) the inner loop indices from the bound expressions of outer loops. This is illustrated in Fig. 2 where $x_i$ is an outer loop variable and $x_j$ is an inner loop variable. To compute the loop bounds for the $x_i$-loop as a function of parameters and outer loop variables, the inner loop variables $x_j$, for $j > i$ must be removed from the domain. This is done by projecting the domain in the direction of the inner loop variables onto the $x_i$-axis, giving *lb* and *ub*, the lower and upper bounds of loop variable $x_i$, respectively. The resulting polyhedron, as shown in Fig. 2, is a cylinder, i.e., has no constraints on the inner loop variables.

Polylib achieves this with its dual representation, by just adding lines $\{l_{i+1}, l_{i+2},...\}$ in the direction of all of the inner loop variables $\{x_{i+1}, x_{i+2},...\}$ to the vertex-ray representation of the polyhedron, and a call to the Chernikova algorithm.

## 2.3. Simplification of Polyhedra in Context

The simplification of a polyhedral domain $\mathscr{D}$ within a *context* domain $\mathscr{C}$ is defined as follows:

Given domains $\mathscr{D}$ and $\mathscr{C}$, then Simplify $(\mathscr{D}, \mathscr{C}) = \mathscr{D}'$, when

1. $\mathscr{D}' \cap \mathscr{C} = \mathscr{D} \cap \mathscr{C}$, $\mathscr{D}' \supseteq \mathscr{D}$, and
2. $D'$ is maximal, in the sense that there does not exist a domain $\mathscr{D}'' \supset \mathscr{D}'$ such that $\mathscr{P}'' \cap \mathscr{C} = \mathscr{D} \cap \mathscr{C}$.

The simplify function finds the largest domain set (i.e., the domain with the smallest list of constraints) that, when intersected with the context $\mathscr{C}$ is equal to $\mathscr{D} \cap \mathscr{C}$. The simplify operation is done by computing the intersection $\mathscr{D} \cap \mathscr{C}$ and while doing so, recording which constraints of $\mathscr{D}$ are

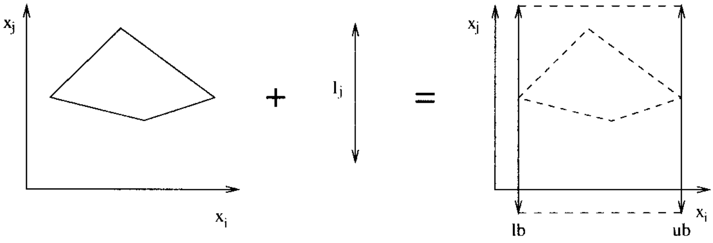


Fig. 2. Adding a line to project out an index.

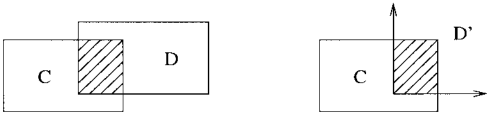


Fig. 3. Domain simplification: $\mathscr{D}' = \text{Simplify}(\mathscr{D}, \mathscr{C})$.

"redundant" with the result of the intersection. The result $\mathscr{D}'$ of the simplify operation is then the domain $\mathscr{D}$ with the "redundant" constraints removed. A simple example is shown in Fig. 3, where domain $\mathscr{D}$ is simplified (resulting in domain $\mathscr{D}'$) by eliminating the two constraints that are redundant with context domain $\mathscr{C}$.

Polyhedral simplification will turn out to be extremely useful later when we seek to eliminate redundant guards and loop bounds tests.

## 3. SCANNING A SINGLE POLYHEDRON

In this section we describe a method to generate perfectly nested loops to scan a single polyhedron.[15] Unlike the standard methods based on Fourier–Motzkin elimination,[13] it exploits the dual representation of polyhedra, as implemented in Polylib.

We formulate the *polyhedron scanning problem* as follows: given a parameterized polyhedral domain $\mathscr{D}(p)$ in terms of a parameter vector $p$ and a set of $k$ constraints: $\mathscr{D}(p) = \{x \mid Ax \geqslant Bp + b\}$, produce a set of loop bound expressions $lb_1, ub_1, ..., lb_n, ub_n$ such that the loop nest:

```
$lb_1 \leqslant x_1 \leqslant ub_1$
  $lb_2 \leqslant x_2 \leqslant ub_2$
    $\ddots$
      $lb_n \leqslant x_n \leqslant ub_n$
        loop body
```

will visit once and only once all integer points in the domain $\mathscr{D}(p)$ in lexicographic order of the elements of $x = (x_1, ..., x_n)$. For a loop variable $x_i$, we use the term *outer loops* to refer to loops which enclose the $x_i$-loop, that is, the loops over index variables $x_j$, for $j < i$. We use *inner loops* to refer to the loops contained in the $x_i$-loop, that is, the loops over indices $x_j$, for $j > i$. In the code to be produced, loop bounds should be the minimum or maximum of affine functions of outer loop variables and parameters.

The problem of finding loop bounds is related to integer linear programming and shares its complexity. Fortunately, these problems tend to be relatively small (in terms of the dimension and number of constraints) due to the fact that loops are not deeply nested, and exact solutions for typical problems can be found in reasonable time.

## 3.1. Basis of the Method

The algorithm for scanning a single polyhedron resembles the standard method Fourier–Motzkin (FM) elimination,[(13)] in that it projects the polyhedron in the canonical direction of inner loop variables in order to eliminate all constraints on them. Thus bounds on a loop are found independent of inner loop indices. However, the FM method considers all pairs of constraints when finding the bound on the projection. In Fig. 4, for example, lower bounds *lb2* and *lb* and upper bounds *ub2* and *ub* are all considered by the FM method. In general, given $n$ constraints, as many as $n^2/4$ loop bounds could be generated in eliminating a single variable, and the total number grows exponentially with the number of variables. Most of these bounds turn out to be redundant (nontight) and must be eliminated. The elimination of these redundant loop bounds is a significant problem when using the FM method.

Our algorithm only considers pairs of constraints that are *adjacent*, and therefore never generates any redundant bounds (like *lb2* and *ub2* in the example). Thus the redundant bound elimination phase of the FM method is replaced with an adjacency test on pairs of constraints in the proposed method. This test is accomplished very efficiently assuming the rays and vertices of the polyhedron are known. The algorithm maintains both the constraint and the ray/vertex representation of a polyhedron allowing us to employ the adjacency test.

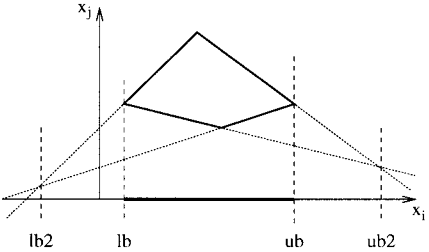


Fig. 4. Projection using Fourier–Motzkin elimination.

## 3.2. Loop Nest Synthesis

We use a recursive algorithm which separates (or factors) a specified domain $\mathscr{D}$ into an intersection of the initial context domain $\mathscr{C}_0$ and a sequence of loop domains $\mathscr{D}_1, \mathscr{D}_2, \ldots$, such that:

1. $\mathscr{D} = \mathscr{C}_0 \cap \mathscr{D}_1 \cap \cdots \cap \mathscr{D}_n$.
2. Each loop domain $\mathscr{D}_i$ has no constraint involving inner loop variables $x_{i+1}, x_{i+2}, \ldots$, i.e., in these dimensions it is a cylinder. Thus, it is completely specified (in terms of constraints) by its projection, $\mathscr{L}_i$, on the outermost $i$ dimensions. We sometimes abuse the notation and use $\mathscr{D}$ and $\mathscr{L}$ interchangeably.
3. The linear constraints defining each loop domain $\mathscr{D}_i$ are not redundant with any constraint established by any outer loop (i.e., within $\mathscr{C}_{i-1} = \mathscr{C}_0 \cap \mathscr{D}_1 \cap \cdots \cap \mathscr{D}_{i-1}$, which we call the *context* of this loop).

Our algorithm "splits" the constraints of $\mathscr{D}$ onto $n$ separate, but not necessarily disjoint (remember that they have the same number of dimensions) polyhedra. The constraints defining each polyhedron $\mathscr{D}_i$ specify the lower and upper bounds of loop index $i$.

At each step $i$ of the recursion, with a context $\mathscr{C}_{i-1}$, we generate one additional loop domain $\mathscr{D}_i$ by

- projecting out the inner loop variables $x_j$, for $j > i$, from domain $\mathscr{D}$;
- simplifying the resulting domain $\mathscr{D}_i$ in the context $\mathscr{C}_{i-1}$; and
- determining the context $\mathscr{C}_i = \mathscr{C}_{i-1} \cap \mathscr{D}_i$ for the next step of the recursion.

The decomposition of $\mathscr{D}$ into a sequence of loop domains $\mathscr{D}_i$ yields a loop nest where each loop of depth $i$ is specified by the polyhedron $\mathscr{D}_i$. Note that this polyhedron has constraints involving only outer loop variables $x_j$, $j \leqslant i$. In the rest of this paper, we write a loop nest as $\mathscr{L}_1 :: \mathscr{L}_2 :: \cdots :: \mathscr{L}_n :: S$.

For example, consider the following polyhedral domain $\mathscr{D}$ (Fig. 5), in a program parameterized by $\mathsf{N}$ and $\mathsf{M}$:

$$\{\mathsf{i, j \mid j <= i <= N; j >= 0; i >= M; i + j <= N + M; M >= 2; M < N}\}$$

Let us decompose $\mathscr{D}$ into a loop nest, i.e., compute $\mathscr{D}_1$ and $\mathscr{D}_2$ such that

$$\mathscr{D} = \mathscr{C}_0 \cap \mathscr{D}_1 \cap \mathscr{D}_2$$

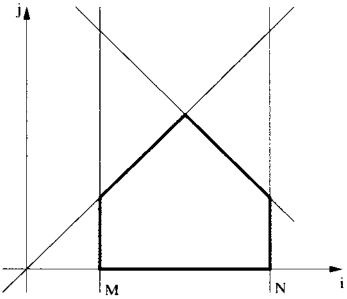


Fig. 5. Domain $\{\mathsf{i, j \mid j <= i <= N; j >= 0; i >= M; i+j <= N+M; M >= 2; M < N}\}$.

The context $\mathscr{C}_0$ is the polyhedron constraining the program parameters, i.e., the projection of $\mathscr{D}$ on the N, M dimensions:

$$\mathscr{C}_0 = \{\mathsf{N, M \mid M >= 2; M < N}\}$$

We start with the *i*-loop. The context for this step is $\mathscr{C}_0$. We first project out the inner dimension *j*, and this yields the following polyhedron:

$$\{\mathsf{i, j \mid M <= i <= N; M >= 2; M < N}\}$$

After simplification in the context of $\mathscr{C}_0$, it yields a polyhedron giving the bounds of the outer loop:

$$\mathscr{L}_1 = \{\mathsf{i \mid M <= i <= N}\}$$

We now generate the next loop, in the context $\mathscr{C}_1 = \mathscr{C}_0 \cap \mathscr{D}_1$. This level of the recursion does not require any projection, since there are no inner loop variables. The polyhedron $\mathscr{D}$ is just simplified in the context $\mathscr{C}_1$, yielding the polyhedron:

$$\mathscr{L}_2 = \{\mathsf{i, j \mid 0 <= j <= i; j <= M + N - i}\}$$

Finally, we obtain the following loop nest:

```
$M \leqslant i \leqslant N$
    $0 \leqslant j \leqslant \min(i, (M+N-i))$
        loop body
```

## 4. THE MULTIPLE POLYHEDRA SCANNING PROBLEM

In the previous section, we discussed loop nest synthesis for a statement defined in the context of a single polyhedron. We now address the problem of generating code for multiple statements, each defined over a different polyhedron (or a finite union of polyhedra). We assume that the definition domains are independent of each other, i.e., they may, or may not, be all disjoint.

In general, data dependencies between statements imply constraints on the scanning order. Thus, we consider the problem of scanning the union of all the domains under the constraint of a given scanning order, formulated as follows:

- a *schedule* assigns a logical execution time to each index vector of each domain. We assume that this time is represented by a $k$-dimensional time vector ($k$ is the same for all domains), and is an affine function $\Lambda$ of the index vector;
- the scanning order must respect the lexicographic order of the logical execution times. We define the lexicographic order $\prec$ as follows: let $x$ and $y$ be two $n$-dimensional vectors; we define

$$x \prec y \equiv \exists p, 1 \leqslant p < n, (x_{(1..p)} = y_{(1..p)}) \wedge (x_{p+1} < y_{p+1})$$

It follows that domains cannot be scanned independently. For example, consider the domains $\mathscr{D}_1 = \{\mathsf{i} \mid \mathsf{1} <= \mathsf{i} <= \mathsf{10}\}$ and $\mathscr{D}_2 = \{\mathsf{i} \mid \mathsf{4} <= \mathsf{i} <= \mathsf{12}\}$, with respective schedules $\Lambda_1(i) = i + 1$ and $\Lambda_2(i) = 2i$. The statements associated with points 1 and 10 of $\mathscr{D}_1$ have to be executed at logical time 2 and 11, while the statements associated with points 4 and 8 of $\mathscr{D}_2$ have to be executed at times 8 and 16. Since $1 < 8 < 10 < 16$, we cannot first completely scan $\mathscr{D}_1$, and then scan $\mathscr{D}_2$, and neither can we scan $\mathscr{D}_2$ followed by $\mathscr{D}_1$.

Hence, it is not possible to generate separate loops to scan both $\mathscr{D}_1$ and $\mathscr{D}_2$. These loops must be partially merged, so that the statements are executed in the order given by the schedule.

The simplest solution uses a perfectly nested loop to scan a convex superset (such as the bounding box) of this union. However, since the domain scanned by this loop is also a superset of *each* statement's domain, we cannot unconditionally execute the statements within the loop body. Rather, each statement $S_i$ must be guarded by conditions which test that the current loop index vector belongs to $\mathscr{D}_i$.

(a)

```
for (i=1; i<=N; ++i){
  for (j=1; j<=N; ++j){
    if (j<=M)
      S1;
    if (i==j && i>=3)
      S2;
  }
}
```

(b)

```
for (i=1; i<=N; ++i) {
  for (j=1; j<=min ((i+3N+M-4)/4,
                    i+M-1, N);
       ++j) {
    if (j<=M)
      S1;
    if (i==j && i>=3)
      S2;
  }
 }
}
```

Fig. 6. Code scanning a superset of the union of domains of Fig. 1: (a) bounding box; and (b) convex closure.

The superset of the union scanned by the perfect loop may simply be the bounding box of the domain (Fig. 6a, as done by the LooPo code generator[18]), or the convex closure of the domain (Fig. 6b as done by the Omega code generator[19]).

Solutions based on perfect loop nests yield compact but inefficient code because of the following limitations.

- Empty iterations: a perfectly nested loop scans a convex polyhedron. Since a domain that is a union of polyhedra may not be convex, some iterations of the loop may not execute any statements. It is often especially inefficient when a domain has an extreme aspect ratio. For example in Fig. 6a, only about M out of every N points are usefully visited. Depending on the parameter values (say if $M \ll N$) and on the complexity of the loop body, this could have a significant overhead.
- Control overhead: each loop iteration must test the guards of all guarded statements, causing a significant control overhead.
- Finally, finite unions of finite parameterized polyhedra do not always admit finite convex supersets. Consider for example the following parameterized domains: $\{i \mid 1 <= i <= N\}$ and $\{i \mid 1 <= i <= M\}$; the smallest convex super-set of these domains is the infinite polyhedron $\{i \mid 1 <= i\}$ (note that $\{i \mid 1 <= i <= \max(M, N)\}$ is not a polyhedron).

To avoid this overhead, we can separate a union of polyhedra into several distinct regions that can each be scanned using imperfectly nested loops. The resulting code is more efficient, since:

- Imperfectly nested loops can scan nonconvex regions, avoiding empty iterations;
- It may be possible to choose these loops such that some statement guards become always true or always false; when a guard is always true, the guard may be removed; when a guard is always false, the entire statement can be removed. This optimization increases the ratio of the number of executed statements to the number of tested guards, and thus reduces the control overhead. This can be carried through to its logical limit to yield code with no guards.

However, the efficiency is obtained at the expense of code size for two reasons. First, a perfect loop is replaced by an imperfectly nested loop, which scans a disjoint union of polyhedra. Second, a statement domain may be divided into several disjoint polyhedra, where each polyhedron is scanned by a different loop. In this case, the statement code has to be duplicated in each of these loops. Figure 7 illustrates this for the example of Fig. 1.

```
for (i = 1; i <= 2; ++i) {
  for (j = 1; j <= M; ++j) {
    S1;
  }
}
for (i = 3; i <= N; ++i) {
  for (j = 1; j <= min((-1 + i),M); ++j) {
    S1;
  }
  if (i<=M) {
    j = i;
    S1;
    S2;
  }
  if (i>=M+1) {
    j = i;
    S2;
  }
  for (j = (1 + i); j <= M; ++j) {
    S1;
  }
}
```

Fig. 7. Imperfectly nested loops scanning the union of domains of Fig. 1.

## 5. AN ALGORITHM TO SCAN UNIONS OF POLYHEDRA

As described earlier, the simple method for scanning unions of polyhedra is to scan a convex superset using guards. The inefficiencies are then resolved by post-processing. Our solution is more holistic. It is recursive, similar to the algorithm for scanning a single polyhedron. This enables us to precisely control the tradeoff between efficiency and code size. We first present the outline of the algorithm using an illustrative example, then formally develop the algorithm, and finally describe an important subproblem, namely that of sorting polyhedra.

In this section, we assume that all loops are sequential, and defer to Section 6 the extension to parallel code generation and also the size-speed tradeoff. Thus the number of time dimensions $k$ is equal to the dimension $n$ of all polyhedra. This necessitates that all polyhedra have the same number of dimensions, but this does not introduce any loss of generality: any polyhedron of dimension $n'$ can be extended to a polyhedron of dimension $n > n'$.

### 5.1. Outline of the Algorithm and Illustration

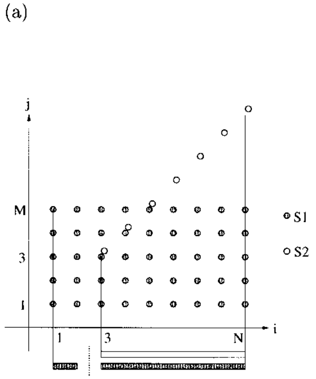
Our algorithm recursively decomposes the union of polyhedra into imperfectly nested loops, starting from outermost loops to innermost loops. At each step, we seek to solve the following problem: given a context (i.e., a polyhedron in $\mathbb{Z}^{(d-1)}$ containing the outer loops and system parameters), and a union of polyhedra in $\mathbb{Z}^n$, $n \geqslant d$, generate a loop that scans this union in lexicographic order.

We generate each additional level of loops by:

1. Projecting the polyhedra onto the outermost $d$ dimensions;
2. Separating these projections into disjoint polyhedra;
3. Recursively generating loop nests that scan each of these;
4. Sorting these loops so that their textual order respects the lexicographic order.

There are two subtle details that we will need to address. First, the idea of "sorting" polyhedra (in step 4), necessitates an order relation and an associated algorithm. Second, we need to ensure that the "separation" of polyhedra (in step 2) must be such that the resulting polyhedra can be so sorted.

Before giving the details, we illustrate the algorithm with an example. Let us start with the program of Fig. 1, and generate the first level of loops.



(b)

```
{i | 1<=i<=2} ::
  {i,j | 1<=i<=2; 1<=j<=M} ::
    S1;
{i | 3<=i<=N} ::
  {i,j | 3<=i<=N; 1<=j<=M} ::
    S1;
  {i,j | 3<=i<=N;  i=j} ::
    S2;
```

Fig. 8. Projections of domains on first dimension and separation into disjoint loops: (a) disjoint polyhedra; and (b) sorted loops.

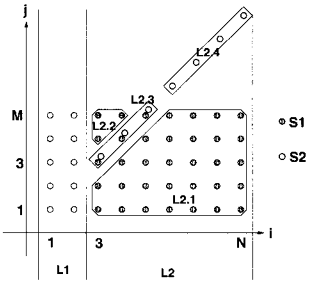
Each domain is projected onto the first dimension and the parameters, i.e., $\{$i, N, M$\}$. These projections are then separated into three disjoint regions: one region containing only S1, one region containing both S1 and S2, and one region containing only S2. For our example, the last region is empty, as shown in Fig. 8a. Furthermore, note that some regions (those that involve the difference of polyhedra) could be unions of polyhedra.

The two remaining regions represent two loops scanning different pieces of dimension i, parameterized by N and M. At this point, only the first level polyhedra ($\{\mathsf{i} \mid \mathsf{1} <= \mathsf{i} <= \mathsf{2}\}$ and $\{\mathsf{i} \mid \mathsf{3} <= \mathsf{i} <= \mathsf{N}\}$) can be interpreted as loops. The other (two-dimensional) polyhedra act as guards on the statements. Note that in Fig. 8, these guards are partially redundant with their context. The redundant constraints are eliminated later.

We will recursively generate separate pieces of code to scan, respectively, regions 1 (containing S1 alone) and 2 (containing both S1 and S2). Then we will, textually place the former *before* the latter. This order respects the lexicographic schedule. Finding such a textual order is trivial for the first dimension, but becomes more complicated for subsequent levels. In the rest of this example, whenever such an order is needed, we will give a valid order without discussing how to obtain it.

Now, we generate the next level of loops in the context of the first level loops. The first i-loop (labeled L1 in Fig. 9a) contains only one statement, and thus the generation of its inner loop is a perfect loop generation problem. Next, consider the second *i*-loop, L2. It contains two guarded statements. In the next level of transformation, these domains are first separated into four disjoint polyhedra: two of them (namely L2.1 and L2.2) containing S1 alone, one (L2.3) containing both S1 and S2, and

(a)



(b)

```
L1      {i | 1<=i<=2} ::
L1.1      {i,j | 1<=j<=M} ::
            S1;
L2      {i | 3<=i<=N} ::
L2.1      {i,j | 1<=j<=i-1; j<=M} ::
            S1;
L2.3      {i,j | i=j; j<=M} ::
            S1;
            S2;
L2.4      {i,j | i=j; M+1<=j} ::
            S2;
L2.2      {i,j | i+1<=j<=M} ::
            S1;
```

Fig. 9. Second level: projection and separation of the second loop nest (L2). (a) disjoint polyhedra; and (b) sorted loops.

one (L2.4) containing S2 alone. Then, we sort these polyhedra, such that the following constraints are respected.

- Loop (L2.1) precedes loops (L2.2), (L2.3), and (L2.4).
- Loop (L2.3) precedes loop (L2.2).

There exist multiple valid textual orderings of these four loops, and we propose one of them in Fig. 9b. When we discuss loop sorting later in Section 5.3 we will describe other solutions for this example. Also note that there is no constraint on the relative position of loops (L2.2) and (L2.4), since their respective contexts ($\{i \mid i <= M - 1\}$ and $\{i \mid 1 >= M + 1\}$) are distinct. In other words, for any value of the parameter M and the outer loop index i, at least one of these loops is empty; it follows that any textual order of these loops will execute correctly.

Finally, we generate code by pretty printing the sorted nested loops, yielding the code we saw previously in Fig. 7.

## 5.2. Formal Presentation of the Algorithm

Recall that the algorithm is recursive, starting from the outer to the inner loop indices. At each step of the recursion, given a context, i.e., a polyhedron which defines the outer loops' bounds and the parameter constraints, we generate an additional set of loops by (i) projecting the statement domains onto the context and the current loop variable; (ii) separating these projections into distinct polyhedra; (iii) recursively generating nested loops; and (iv) sorting them to determine their textual order.

The following algorithm generates one level of the loops at depth $d$, given a union of domain-qualified statements $P_i :: S_i$, and a context $\mathscr{C}$.

1. Replace each polyhedron $P_i$ by its intersection with the context $\mathscr{C}$. This ensures that $P_i$ defines exactly the domain of statement $S_i$ in the context of the current nested loop.

2. Compute the projection $L_i$ of each polyhedron $P_i$ onto the outermost $d$ loop variables and the parameters, i.e., eliminate the inner variables. Each statement $S_i$ is now qualified by its validity domain $P_i$ and the projection of this domain $L_i$, written as $L_i :: P_i :: S_i$. Note that we do not simplify $P_i$ in the context of $L_i$ but retain it unchanged. This notation is thus redundant, since $P_i$ is the domain of $S_i$. As explained later, the constraints of $P_i$ will be simplified in the context at the last step of the recursion, and all remaining constraints—i.e., those that are not loop bounds—will remain as guards on $S_i$.

3. Separate all projections $L_i$ into disjoint polyhedra: for example, with two loops $L_i :: P_i :: S_i$ and $L_j :: P_j :: S_j$, when $L_i \cap L_j$ is nonempty, we generate three regions, containing respectively, $S_1$ alone, $S_2$ alone, and both $S_1$ and $S_2$

   - $(L_i - L_j) :: P_i :: S_i$
   - $(L_j - L_i) :: P_j :: S_j$
   - $(L_i \cap L_j) :: (P_i :: S_i, P_j :: S_j)$.

   Note that the first two regions may be unions of polyhedra (in this case, we consider each polyhedron as a region, and duplicate the associated statements). This was illustrated in Fig. 9, where the region containing S1 alone is the union of L2.1 and L2.2. We repeat the separation until the $L$ are all distinct. Note that this step may associate several statements to a loop $L$.

   The separation of more than two loops is a straightforward extension and explained later.

4. Recurse for each loop $L :: (P_r :: S_r, P_s :: S_s)$, with context $\mathscr{C} \cap L$.

5. Sort the loops, as described in Sect. 0.0, to determine their textual order.

6. Finally, simplify each loop $L_i$ in the context of $\mathscr{C}$; this step eliminates redundant loop bounds.

A couple of points merit a detailed explanation.

**Deferred Simplification of Guards.** Note that we simplify loop guards only in step 6, the very end of the recursion, rather than in step 2. We illustrate, with the following example, how this gives us the possibility to optimize the code. Consider two statements, each one associated to a two-dimensional domain:

```
{i, j, N | 1 <= i <= 10; 1 <= j <= i} :: S1
{i, j , N | 1 <= i <= 10; 11 <= j <= N} :: S2
```

The first level of recursion generates the following program:

```
{i, N | 1 <= i <= 10} ::
  {i, j, N | 1 <= i <= 10; 1 <= j <= i} :: S1
  {i, j, N | 1 <= i <= 10; 11 <= j <= N} :: S2
```

Now suppose that we simplify, right now, the constraint on i in each j-loop, yielding

```
{i, N | 1 <= i <= 10} ::
  {i, j, N | 1 <= j <= i} :: S1
  {i, j, N | 11 <= j <= N} :: S2
```

Now, the information that $i \leqslant 10$ is "lost", and the recursive call is forced to to consider the case where $i \geqslant 11$, and thus generate three loops. However, with the knowledge of the context, it turns out that the two loops are distinct and that a third loop would be empty for any value of i.

**Separation of Multiple Loops.** In step 3, we use the following algorithm which, given a set of polyhedra $\mathscr{P} = \{P_i\}$, computes a set of disjoint polyhedra $\mathscr{Q} = \{Q_i\}$ such that $\bigcup_i P_i = \bigcup_j Q_j$. $\mathscr{Q}$ is initially empty. Each step considers a polyhedron $P_i \in \mathscr{P}$ and computes a new $\mathscr{Q}$ as follows.

- For all $Q_j \in \mathscr{Q}$, add $Q_j - P_i$, associated with the statements of $Q_j$ alone, to (the next round of) $\mathscr{Q}$.

- For all $Q_j \in \mathcal{Q}$, add $Q_j \cap P_i$ associated with the statements of $Q_j$ and $P_i$, to (the next round of) $\mathcal{Q}$.
- Finally add $P_i - (\bigcup_i Q_i)$, associated to the statement of $P_i$ alone.

Our method separates the set of statement domains into disjoint polyhedra without considering any order between these polyhedra. Indeed, the algorithm described earlier manipulates nonordered sets of polyhedra, and only ensures that the result is a set of disjoint polyhedra. Note that this is an essential point of our method. We show in next section that, given a set of disjoint polyhedra, we can generate and textually order loops such that these loops respect the lexicographic order.

## 5.3. Sorting Parameterized Disjoint Polyhedra

In this section, we address the problem of sorting disjoint polyhedral domains. The loop separation step described before does not give any information on how to order the set of loops it yields. This separation is useful only if it is always possible to find a textual order for these loops. In this section, we show that such an order always exists, and can be computed. We first define an order between polyhedra. Then we use this order to textually order loops.

The key idea of sorting two loops (or polyhedra) answers the following question: is it possible to textually place the code that scans one of them before the code that scans the other, without violating the lexicographic order? Such a textual order makes sense only if the two polyhedra share a common context, and hence this order is only a partial order. Moreover, a necessary condition is that the polyhedra are all disjoint. Using two distinct loops, it is not possible to scan in the lexicographic order two polyhedra that intersect.

A loop enclosed in a nested loop of depth $l$, in a program with $p$ symbolic parameters, is specified by a one-dimensional polyhedron parameterized by a $l+p$ vector. This vector defines the context of the loop. A correct textual order for a set of loops has to be valid for any value of the context, i.e., for any value of the program parameters and the outer loop variables.

We first define a partial order on disjoint parameterized polyhedra. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two one-dimensional polyhedra parameterized by $(l+p)$-vectors. We denote by $\mathcal{C}_1$ and $\mathcal{C}_2$, the respective $(l+p$-dimensional domains of parameters of $\mathcal{D}_1$ and $\mathcal{D}_2$. In other words, $\mathcal{C}_1$ and $\mathcal{C}_2$ are the respective projections of $\mathcal{D}_1$ and $\mathcal{D}_2$ on the parameter space. We define the order, noted $\lhd$, as follows: $\mathcal{D}_1 \lhd \mathcal{D}_2$ iff:

1. $\mathscr{D}_1 \cap \mathscr{D}_2 = \varnothing$
2. $\mathscr{C}_1 \cap \mathscr{C}_2 \neq \varnothing$
3. $\forall q \in \mathscr{C}_1 \cap \mathscr{C}_2$, $\forall x \in \mathscr{D}_1(q)$, $y \in \mathscr{D}_2(q)$, $x < y$, i.e., for any value of the context, each point of $\mathscr{D}_1$ precedes all the points of $\mathscr{D}_2$.

As stated before, the first condition implies that this partial order can be applied only to sets of disjoint polyhedra. Indeed, the condition is always ensured by the loop separation algorithm described before.

When the second condition is not verified, then $\mathscr{D}_1$ and $\mathscr{D}_2$ are not comparable. This means that any textual order of the loops which scan these polyhedra will respect the lexicographic order.

When $\mathscr{C}_1$ and $\mathscr{C}_2$ intersect, then $\mathscr{D}_1$ and $\mathscr{D}_2$ are comparable, and exactly one of the following relations is true:

- $\mathscr{D}_1 \lhd \mathscr{D}_2$
- $\mathscr{D}_2 \lhd \mathscr{D}_1$.

Determining whether $\mathscr{D}_1 \lhd \mathscr{D}_2$ or $\mathscr{D}_2 \lhd \mathscr{D}_1$ can be done using basic polyhedral operations. For better understanding, we illustrate the case where $\mathscr{D}_2 \lhd \mathscr{D}_1$. We give the algorithm later. Consider the two polyhedra $\mathscr{D}_1$ and $\mathscr{D}_2$ of Fig. 10. They are distinct, thus there exists at least one hyperplane $H$ which separates the combined context-loop space into two subspaces $S_1$ and $S_2$, containing respectively $\mathscr{D}_1$ and $\mathscr{D}_2$. Moreover, $H$ is not parallel to the loop dimension (were it so, $H$ would be orthogonal to the context, and thus $\mathscr{C}_1$ and $\mathscr{C}_2$ would be disjoint, a contradiction). Thus, $H$ is such that:

$$\forall q \in \mathbb{Z}^{(l+p)}, \forall x, y \in \mathbb{Z} \text{ s.t. } \begin{pmatrix} q \\ x \end{pmatrix} \in S^1 \quad \text{and} \quad \begin{pmatrix} q \\ y \end{pmatrix} \in S_2, \quad \text{then } x < y$$

In particular, this also holds when $\binom{q}{x} \in \mathscr{D}_1$ and $\binom{q}{y} \in \mathscr{D}_2$.

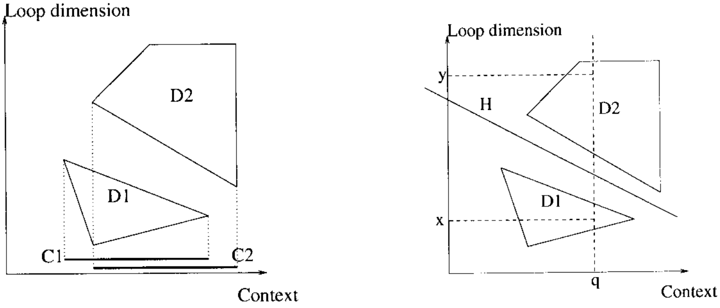


Fig. 10. Precedence relation between two loops: $\mathscr{D}_1 \lhd \mathscr{D}_2$ in the context.

The following algorithm computes the precedence relation between two parameterized polyhedra $\mathscr{D}_1$ and $\mathscr{D}_2$. A precondition is that these polyhedra do not intersect. The result of the algorithm is either "incomparable," $\mathscr{D}_1 \lhd \mathscr{D}_2$ or $\mathscr{D}_2 \lhd \mathscr{D}_1$.

1. Compute the projections $\mathscr{C}_1$ and $\mathscr{C}_2$ of $\mathscr{D}_1$ and $\mathscr{D}_2$ on the context space.
2. Intersect the projections: $\mathscr{I} = \mathscr{C}_1 \cap \mathscr{C}_2$.
3. If the intersection $\mathscr{I}$ is empty, then the polyhedra are not comparable, exit.
4. The polyhedra are comparable, and either $\mathscr{D}_1 \lhd \mathscr{D}_2$, or $\mathscr{D}_2 \lhd \mathscr{D}_1$.
5. Arbitrarily assume that $\mathscr{D}_1 \lhd \mathscr{D}_2$. By definition, $\mathscr{D}_1 \lhd \mathscr{D}_2$ implies that

$$\forall p \in \mathscr{I},\ \forall x_1 \in \mathscr{D}_1(p),\ \forall x_2 \in \mathscr{D}_2(p),\ \exists x \text{ such that } x_1 < x < x_2$$

Let $\mathscr{H}(p) = \{x \mid x_1 < x < x_2, x_1 \in \mathscr{D}_1(p),\ x_2 \in \mathscr{D}_2(p)\}$. To verify that, for all $p$, this polyhedron is not empty, it is sufficient to verify that the polyhedon $\mathscr{P}$, defined as follows, is not empty:

- Select from $\mathscr{D}_1$ the constraints "$x > \cdots$"
- Select from $\mathscr{D}_2$ the constraints "$x < \cdots$"
- Compute the resulting polyhedron $\mathscr{P}$.

$\mathscr{P}$ includes $\mathscr{D}_1$, $\mathscr{D}_2$, and $\mathscr{H}$. If $\mathscr{P}$ is empty, then the assumption was false, i.e., $\mathscr{D}_2 \lhd \mathscr{D}_1$, else it is verified, and thus $\mathscr{D}_1 \lhd \mathscr{D}_2$.

This algorithm is implemented in our polyhedral library. It requires three polyhedral operations (i.e., six calls to the Chernikova algorithm) to decide whether the two polyhedra are comparable, and one more operation to compute the order.

Given this partial order, a method to choose a textual order (out of possibly many solutions) computes a precedence graph of the set of polyhedra, and selects a traversal of this graph. To build this graph, whose nodes are the polyhedra, one has to compare every pair of polyhedra, and set an edge between the corresponding nodes if the polyhedra are comparable. For $x$ polyhedra, building this graph requires $x(x-1)/2$ comparisons. However, this is only an upper bound on the effective number of comparisons, since some comparisons can be avoided by applying transitivity. Note that the worst case is when no pair of polyhedra can be compared—the resulting graph has no edges, thus any textual order is valid, but deciding this requires the maximum number of comparisons.

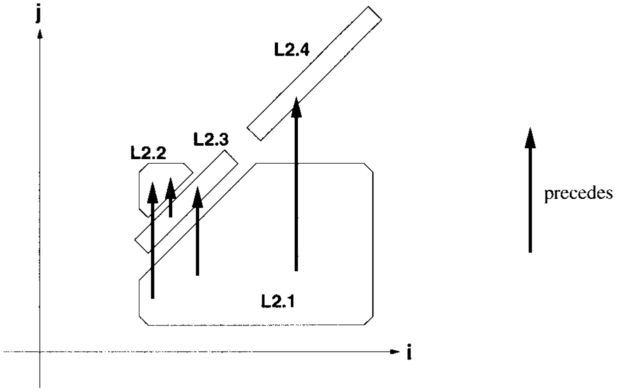


Fig. 11. Partial order between parameterized polyhedra of Fig. 9a.

Finally, the following algorithm selects a valid order:

1. Let $i = 0$;
2. Find a polyhedron which has no predecessor, label it with $i$ and remove it from the graph;
3. Increment $i$ and repeat step 2 until there is no polyhedron left.
4. Each polyhedron is now labeled with a position. The generated loops can be textually placed following the ascending order of these positions.

We illustrate this on our running example. We describe the computation of the set of solutions for sorting the loops in the second step of the recursion (Fig. 9a). Loop L2 contains four distinct $j$-loops, noted L2.1, L2.2, L2.3, and L2.4 (Fig. 11).

- For any value of i and of the parameters, loop L2.1 precedes all the other loops, i.e., its corresponding node is a root of the precedence graph.
- Now ignore loop L2.1. Loops L2.3 and L2.4 have no predecessors, and thus can be placed immediately after L2.1. At this time, we already have two partial solutions:

  —L2.1; L2.3 or

  —L2.1; L2.4.

- Let us consider the first solution. The two remaining loops L2.2 and L2.4 are not comparable, so their textual order is meaningless. In other words,

  —L2.1; L2.3; L2.2; L2.4, and

  —L2.1; L2.3; L2.4; L2.2

  are both valid textual orders.

- Let us now consider the other solution (L2.1; L2.4). The remaining loops L2.2 and L2.3 admit an order, and thus

  —L2.1; L2.4; L2.3; L2.2

  is the third valid order.

## 6. EXTENSIONS

We now address a number of extensions to the main algorithm. These extensions are implemented in our loop generator.

The algorithm described in Section 5 generates imperfectly nested loops, optimized for efficiency. It removes all the guards on the statements and, consequently, ensures that each iteration of a loop will execute all the statements it contains. However, it does not ensure that all loops are visited in a given context. To illustrate this limitation, consider our running example in Fig. 9. Loop L2.2 is empty in the context $i \geqslant M$, and loop L2.3 is also empty when $i \geqslant M+1$. In other words, the code generated for loops L2.2 and L2.3 is dead code for some of the iterations of L2. If $N \ll M$, these two empty loops represent a significant execution overhead.

This is known at compile time, and thus the generated code can be further optimized. This problem is resolved by adding a second separation procedure in the bottom-up phase of the recursion, between steps 5 and 6 of the algorithm. By projecting every loop on the context space, we obtain for each loop, the sub-context where it is not empty. Then we separate these sub-contexts, using the same separation algorithm as for the loops. Finally, we sort the separated contexts. For the loop program of Fig. 9, this extended algorithm generates the loops given in Fig. 12. Note that the overhead is removed at the expense of code size.

### 6.1. Trade-off Between Code Size and Efficiency

At the other extreme, another extension of our algorithm can limit code size. Recall that our method is recursive, and that at each step of the

```
L1     {i | 1<=i<=2} ::
L1.1     {i,j | 1<=j<=M} ::
           S1;
L2     {i | 3<=i<=M} ::
L2.1     {i,j | 1<=j<=i-1; j<=M} ::
           S1;
L2.4     {i,j | i=j; M+1<=j} ::
           S2;
L2.2     {i,j | i+1<=j<=M} ::
           S1;
L3     {i | i=M+1} ::
L3.1     {i,j | 1<=j<=i-1; j<=M} ::
           S1;
L3.3     {i,j | i=j; j<=M} ::
           S1;
           S2;
L4     {i | 3<=i<=M} ::
L4.1     {i,j | 1<=j<=i-1; j<=M} ::
           S1;
L4.4     {i,j | i=j; M+1<=j} ::
           S2;
```

Fig. 12. Second separation of loop L2 into L2, L3, and L4.

recursion we separate the polyhedra into disjoint ones in order to generate imperfectly nested loops. We can always choose to not separate the polyhedra and simply generate loops with guards. We implemented a simple heuristic which considers that the depth of a guard indicates its cost. A similar heuristic is also implemented in the Omega code generator.(19) The heuristic kicks in for all loops whose depth is smaller than the chosen one. This corresponds to generating perfectly nested loops which scan the outermost dimensions, and then generating optimized loops for the inner dimensions.

However, the estimate is not realistic. The cost of a guard is the number of times it is potentially evaluated, i.e., the volume of its context polyhedron. Since there are well known techniques for estimating the number of integer points in a parameterized polyhedron using Ehrhart polynomials,(20) we could use an exact function rather then a heuristic. This is the subject of our ongoing work.

## 6.2. Non-unimodular Mappings

The polyhedral model can also deal with non-unimodular mappings. We use a generalization of the change of basis transformation to apply non-unimodular mappings to polyhedra. The intuition behind this is that any nonsingular matrix $M$ admits a unimodular row extension $U$ such that $U$ admits a unimodular column extension. $U$ specifies a one-to-one transformation. If we are given a non-unimodular mapping $M$, we apply a change of basis using $U$. The image of a $n$-dimensional polyhedron $P$ by $U$ is an $m$-dimensional polyhedron $Q$, where $m \geqslant n$. The change of basis by $U$ ensures that $P$ and $Q$ have the same number of integer points, and moreover that the $n$ first dimensions of $P$ correspond to the mapping specified by $M$.

For example, consider the polyhedral domain $\{\mathsf{i} \mid \mathsf{1} <= \mathsf{i} <= \mathsf{N}\}$, where $\mathsf{N}$ is a program parameter, and the transformation $(\mathsf{i} \text{-}> \mathsf{2i} + \mathsf{1})$. This transformation is not unimodular, but the extended transformation $(\mathsf{i} \text{-}> \mathsf{2i} + \mathsf{1}, \mathsf{i})$ is one-to-one (its inverse is $(\mathsf{i}, \mathsf{j} \text{-}> \mathsf{j})$). The image of the domain by this extended transformation is

$$\{\mathsf{i}, \mathsf{j} \mid \mathsf{3} <= \mathsf{i} <= \mathsf{2N} + \mathsf{1}; \mathsf{i} = \mathsf{2j} + \mathsf{1}\}$$

This polyhedron is embedded in two-dimensional space, but it is also constrained by an equality, and thus defines a one-dimensional set of points. In other words, it can be scanned using a single loop, and the stride of this loop can be determined easily, using polyhedral operations.

### 6.3. Data-parallel Code

We now address the generation of code where the outermost $k$ loops are sequential, and the other loops are parallel.

We introduce a method to generate data-parallel code as a generalization of our method of sequential code generation. We extend the algorithm described in previous section to the case where the schedule dimension $k$ is lower than (or equal to) the dimensions of the definition domains.

Given a $k$-dimensional schedule $\Lambda_V$ for a variable $V$, we compute a row extension $T_V$ of $\Lambda_V$ such that $T_V$ admits a unimodular column extension. Hence $T_V$ defines a one-to-one transformation, and can be used to change the basis of $V$. Since the first $k$ rows of $T_V$ are the rows of $\Lambda_V$, the first $k$ dimensions of $V$ in the new basis now carry all the data dependencies. In other words, only the first $k$ depths of loops need to respect the lexicographic order of the $k$ first indices. We generate imperfectly nested loops of depth $k$, using the algorithm described in Section 5. Then, since there are no more data dependencies, we generate perfectly nested loops which scan the remaining dimensions of the definitions domain of each statement, and we annotate these loops such that they can be recognized as parallel loops by the back-end code generator.

## 7. RELATED WORK

The problem of loop synthesis was first addressed by Ancourt and Irigoin.(13) They use Fourier–Motzkin pairwise elimination to produce bounds at each level. The idea is to use the Fourier–Motzkin pairwise elimination algorithm to find the bounds for each loop. The system of constraints is transformed into a triangular system by projecting out, for each loop, the inner loop indices. This algorithm yields a system where each loop index depends only on the outer loop indices and the program parameters. The main drawback of this method is that it creates redundant bound equations, that have to be eliminated afterwards.

Fourier–Motzkin based methods are used by a number of parallelizing compilers, including the Pandore II compiler,(21) SUIF,(22) and the PIPS project.(23)

Li and Pingali(24) extend the previous method to deal with nonunit loop strides, by allowing nonunimodular transformations. Their method is based on the Hermite Normal Form of matrices(25): they decompose the transformation into two parts, a unimodular one, and nonunimodular which preserves the lexicographic order. They generate a nested loop scanning the image of the source index domain by the unimodular part of the transformation. Then, in the loop body, the variables of the loop are replaced by their image by the other part of the transformation. Ramanujam(26, 27) and Xue(28) also presented methods to deal with non unit loop strides. They use the Fourier–Motzkin elimination to compute the loop bounds, and the Hermite decomposition to determine the strides. All these methods generate loops that scan exactly the image of the original domain by the non-unimodular transformation, i.e., a $Z$-polyhedron. Moreover, they are restricted to scanning a single domain. In contrast our method, which deals with multiple polyhedra, uses the Smith Normal Form to compute polyhedra whose projections are the respective $Z$-polyhedra, and generates a loop nest which scans these polyhedra.

Collard *et al.*(29) show how the dual simplex method(25, 30) can be used to generate loop bounds. It has been implemented in Bouclettes, a Fortran parallelizer.(31)

All these methods deal essentially with the problem of scanning a single polyhedron. There has not been much work on the multiple polyhedra scanning problem. To the best of our knowledge, only four other authors have addressed this problem.

Our initial ideas on polyhedra separation and the intuition behind sorting were briefly presented by Quinton *et al.*,(32) and implemented in an early prototype Alpha compiler. However, the sorting step was not completely formalized, and the generator did not correctly handle domains with equalities. A similar method for loop separation was proposed by Albiez.[33]

Collard[34] addresses the generation of loops from multiple polyhedra, with emphasis on the case of loops with incompatible strides. The generated code contains many guards. The paper assumes that there is a single time dimension, but the schedule may not admit a unimodular completion. The author does not therefore face the problem of polyhedron separation and/or sorting, but must confront non-unimodular transformations. Much of the paper is devoted to the (nonunit) strides due to the schedule, and to the removal of redundant guards.

The LooPo project[35] at the University of Passau also explores the parallelization in the polyhedral model and its extension and addresses the problem of code generation. The currently available working generator produces simplistic code that scans the bounding box of the union of all the domains and executes guarded statements. In a recent paper, they also address the problem of imperfectly nested loop generation.[18] Given a set of domain-qualified statements, they seek to recursively merge the polyhedra, in much the same way as we do. It seems however, that they do not exploit the *complete* context information that is available which leads to naive code. They claim that, with unknown program parameters, the number of branches—loop sequences—in the final program is an exponential of the number of unknown loop bounds (exact number of permutation). Their preliminary experiments were discouraging, as evidenced by their statement that, "a simple program of a few lines with three nested loops [...leads to code with an] if-cascade of 40320 branches". We have seen in this paper that generating code for every permutation is not necessary, since parameterized polyhedra admit a partial order.

The Omega system,[19, 36] which is also freely available under the Gnu public license, produces code to scan unions of convex sets, in an extension of the polyhedral model. Although the generated code is very similar to ours, it uses a different, nonrecursive method. It starts from the naive code (a perfect nest of loops which scan the convex closure of the union of domains), and then iteratively tries to remove sources of overhead (guards and min/max operations in loop bounds). However, it generates incorrect code for domains with equalities, and because the generator is no longer maintained, it is difficult to determine whether this is a simple bug or an inherent limitation. The most interesting aspect of their code generator is their technique of post-optimization. As described before, they seek using a bottom-up method to "lift" guards outside loops. Essentially, an affine guard is expressed as an inequality between the current loop index and an affine function $f$ of all the surrounding ones. We can split the loop into two by breaking it at $f$ and, in each of the two parts, we can be sure that the guard is either always true or always false. Hence we can unconditionally suppress the statement in one of the parts and unconditionally execute it in the other. Since the tests are done one guard at a time and one dimension at a time, they experience considerable code explosion. The trade-off between code size and code efficiency is controlled by specifying the depths from which overhead is eliminated.

Finally, let us mention another optimization, which is independent of those presented here. Since loop bounds are minimum or maximum of affine functions of outer loop variables and parameters, the bounds testing may also lead to overheads. Chamski(37) describes a method to eliminate minimum and maximum expressions, by statically computing these expressions and generating imperfect sequences of loops. This is implemented in our code generator as a post-optimization, since it is completely orthogonal to the methods presented here.

## 8. CONCLUSION

We have addressed the problem of scanning the union of (not necessarily disjoint) polyhedra in lexicographic order their indices. The problem arises in the context of the code-generation phase of parallelizing compilers, especially when one permits different mappings to be applied to the individual statements in the body of a static control loop. Our specific context is that of compiling Alpha, a data-parallel functional language based on systems of affine recurrence equations defined over polyhedral domains. The MMAlpha system for manipulating Alpha programs (written using the Mathematica system) is available under the Gnu public license (see http://www.irisa.fr/cosi/ALPHA). The loop generation module in the Alpha code generator is written in C + + and is also available under the Gnu public license (at http://www.irisa.fr/cosi/ALPHA/codegen).

Our solution is based on a recursive algorithm to generate imperfectly nested loops. The key idea is that at each level of the recursion (responsible for generating a set loops at that level), we use two main operations, namely *separation* of the polyhedra, and *sorting* polyhedra in a given context. Because of a modular separation of the two operations, and because they are applied only in a limited recursive context, we avoid the explosion in the number of comparisons that other methods seem to perform. Furthermore, our algorithm allows fine control of the tradeoff between code size and code efficiency.

We use a unified representation for loops and guards, which ensures that bounds and remaining guards are not redundant, and our algorithm is based on a small set of polyhedral operations drawn from Polylib, a library of polyhedral operations.(14)

We have proposed a solution to the problem of sorting parameterized loops, and shown that any set of disjoint parameterized loops admit a textual order that respects a given lexicographic schedule. We have defined a partial order on parameterized polyhedra, and used this order to sort loops.

In the context of the MMAlpha system, our ongoing work involves the code generation problem from hierarchical Alpha programs, which will involve generating procedure calls and memory allocation issues. In addition, in the context of silicon compilation for dedicated VLSI implementations we are also exploring the problem of automatic generation of interfaces. We are also investigating extensions to the code generation procedure that will allow code to be generated under precise code size constraints.

## ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their useful comments on an earlier draft of this paper.

## REFERENCES

1. D. J. Kuck, *The Structure of Computers and Computations*, Wiley, New York (1978).
2. R. Karp, R. Miller, and S. Winograd, The organization of computations for uniform recurrence equations, *J. Assoc. Computing Machinery* **14**(3):563–590 (July 1967).
3. P. Feautrier, Dataflow analysis of array and scalar references, *IJPP* **20**(1):23–53 (February 1991).
4. P. Feautrier, Some efficient solutions to the affine scheduling problem, Part I, one dimensional time, and Part II, multidimensional time, *IJPP* **21**(5/6):313–348; 389–420 (October 1992).
5. S. Rajopadhye and R. Fujimoto, Synthesizing systolic arrays from recurrence equations, *Parallel Computing* **14**:163–189 (June 1990).
6. P. Quinton and V. Van Dongen, The mapping of linear recurrence equations on regular arrays, *J. VLSI Signal Proc.* **1**(2):95–113 (1989).
7. U. Banerjee, *Loop Transformations for Restructuring Compilers: The Foundations*, Kluwer Academic Publishers, Norwell, Massachusetts (1993).
8. F. Irigoin, Code generation for the hyperplane method and for loop interchange, Technical Report ENSMP-CAI-88-E102/CAI/I, Ecole Nationale Superieure des Mines de Paris (October 1988).
9. M. E. Wolf and M. Lam, Loop transformation theory and an algorithm to maximize parallelism, *IEEE Trans. Parallel Distrib. Syst.* **2**(4):452–471 (October 1991).
10. A. Darte and F. Vivien, Optimal fine and medium grain parallelism detection in polyhedral reduced dependence graphs, Rapport de recherche 96-06, LIP, ENS Lyon (April 1996).

11. A. Darte and Y. Robert, Mapping uniform loop nests onto distributed memory architectures, Rapport de recherche 93-03, LIP (January 1993).
12. P. Feautrier, Toward automatic distribution, *Parallel Proc. Lett.* **4**(3):233–244 (September 1994).
13. C. Ancourt and F. Irigoin, Scanning polyhedra with DO loops, *Third Symp. Principles and Practice of Parallel Programming (PPoPP)*, ACM SIGPLAN, ACM Press, pp. 39–50 (1991).
14. D. Wilde, A library for doing polyhedral operations, Technical Report PI 785, IRISA, France (December 1993).
15. H. Le Verge, V. Van Dongen, and D. Wilde, Loop nest synthesis using the polyhedral library, Technical Report PI 830, IRISA, France (May 1994).
16. T. S. Motzkin, H. Raiffa, G. L. Thompson, and R. M. Thrall, *Contribution to the Theory of Games*, Princeton University Press (1953).
17. N. V. Chernikova, Algorithm for finding a general formula for the no *n*-negative solution of a system of linear inequalities, *U.S.S.R. Computational Math. Math. Phys.* **5**(2):228–233 (1965).
18. M. Griebl, C. Lengauer, and S. Wetzel, Code generation in the polytope model, *Proc. Int'l. Conf. on Parallel Architectures and Compilation Techniques (PACT'98)*, IEEE Computer Society Press, pp. 106–111 (1998).
19. W. Kelly, W. Pugh, and E. Rosser, Code generation for multiple mappings, *Frontiers '95: Fifth Symp. Frontiers of Massively Parallel Computation*, McLean, Virginia (February 1995).
20. P. Clauss and V. Loechner, Parametric analysis of polyhedral iteration spaces, *IEEE Int'l. Conf. on Application Specific Array Processors*, *ASAP'96*, IEEE Computer Society (August 1996).
21. F. Andre, M. Le Fur, Y. Maheo, and J.-L. Pazat, The Pandore data-parallel compiler and its portable runtime, *LNCS* **919**:176–183 (1995).
22. S. P. Amarasinghe, Parallelizing compiler techniques based on linear inequalities, Ph.D. thesis, Stanford University (January 1997).
23. F. Irigoin, P. Jouvelot, and R. Triolet, Semantical interprocedural parallelization: An overview of the PIPS project, *Confe. Proc. Int'l. Conf. Supercomputing*, Cologne, Germany, ACM SIGARCH, pp. 244–251 (June 1991).
24. W. Li and K. Pingali, A singular loop transformation framework based on non-singular matrices, *IJPP* **22**(2):183–205 (April 1994).
25. A. Schrijver, *Theory of Linear and Integer Programming*, A Wiley-Interscience Publication, John Wiley, Chichester, New York (1986).
26. J. Ramanujam, Nonunimodular transformations of nested loops, *Supercomputing '92*, Minneapolis, Minnesota, pp. 214–223 (November 1992).
27. J. Ramanujam, Beyond unimodular transformation, *J. Supercomputing* **9**(4):365–389 (1995).
28. J. Xue, Automating nonunimodular loop transformations for massive parallelism, *Parallel Computing* **20**(5):711–728 (April 1994).
29. J.-F. Collard, T. Risset, and P. Feautrier, Construction of DO loops from systems of affine constraints, *Parallel Proc. Lett.* **5**(3):421–436 (September 1995).
30. P. Feautrier, Parametric integer programming, *Rairo Recherche Opérationnelle* **22**(3): 243–268 (September 1988).
31. P. Boulet, Bouclettes: A Fortran loop parallelizer, *LNCS* **1067**:784–791 (1996).
32. P. Quinton, S. Rajopadhye, and D. Wilde, Deriving imperative code from functional programs, *Seventh Conf. Functional Prog. Lang. and Computer Architecture*, ACM, La Jolla, California, pp. 36–44 (June 1995).

33. O. Albiez, Parcours d'une superposition de domaines, *RenPar'10* (June 1998).
34. J.-F. Collard, Code generation in automatic parallelizers, C. Girault (ed.), *Proc. Int'l. Conf. on Applications in Parallel and Distributed Computing*, IFIP W.G 10.3, Caracas, Venezuela, North Holland, pp. 185–194 (April 1994).
35. M. Griebl and C. Lengauer, The loop parallelizer LooPo, Michael Gerndt (ed.), *Proc. Sixth Workshop on Compilers for Parallel Computers*, *Konferenzen des Forschungszentrums Jülich*, Forschungszentrum Jülich, Vol. 21, pp. 311–320 (1996).
36. W. Kelly and W. Pugh, A framework for unifying reordering transformations, Technical Report CS-TR-3193, Dept. of Computer Science, University of Maryland (April 1993).
37. Z. Chamski, Nested loop sequences: Towards efficient loop structures in automatic parallelization, Technical Report PI 772, IRISA, France (October 1993).