Simplification and Run-time Resolution of Data Dependence Constraints for Loop Transformations
Diogo Sampaio, Louis-Noël Pouchet, Fabrice Rastello

To cite this version:

HAL Id: hal-01653819
https://inria.hal.science/hal-01653819
Submitted on 5 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Simplification and Runtime Resolution of Data Dependence Constraints for Loop Transformations

Diogo N. Sampaio  
Univ. Grenoble Alpes, Inria, CNRS,  
LIG F-38000  
Grenoble, France  
diogo.nunes-sampaio@inria.fr

Louis-Noël Pouchet  
Colorado State University  
Fort Collins, USA  
pouchet@cse.ohio-state.edu

Fabrice Rastello  
Univ. Grenoble Alpes, Inria, CNRS,  
LIG F-38000  
Grenoble, France  
fabrice.rastello@inria.fr

ABSTRACT
Loop transformations such as tiling, parallelization or vectorization are essential tools in the quest for high-performance program execution. Precise data dependence analysis is required to determine whether the compiler can apply a transformation or not. In particular, current static analyses typically fail to provide precise enough dependence information when the code contains indirect memory accesses or polynomial subscript functions to index arrays. This leads to considering superfluous may-dependences between instructions that prevent many loop transformations to be applied.

In this work we present a new hybrid (static/dynamic) framework that allows to overcome several limitations of purely static dependence analyses: For a given loop transformation, we statically generate a test to be evaluated at runtime. This test allows to determine whether the transformation is valid, and if so triggers the execution of the transformed code, falling back to the original code otherwise. Such test, originally constructed as a loop-based code with $O(n^2)$ iterations ($n$ being the number of iterations of the original loop-nest), are reduced to a loop-free test of $O(1)$ complexity thanks to a new quantifier elimination scheme that we introduce in this paper. The precision and low overhead of our method is demonstrated over 25 kernels.

CCS CONCEPTS
• Computing methodologies → Symbolic and algebraic algorithms; • Software and its engineering → Compilers;

KEYWORDS
Hybrid Dependence Analysis, Loop optimization, May-alias, Polynomials, Polyhedral Transformation, Quantifier Elimination

ACKNOWLEDGMENTS
This work was funded in part by the U.S. National Science Foundation, award 1524127 and by CAPES - Brazil. The authors are also very grateful to Alain Ketterlin, who developed ASTER.

1 INTRODUCTION
Loop optimizations span from vectorization, scalar promotion, loop invariant code motion, or software pipelining to loop fusion, skewing, tiling [59, Ch.9], or loop parallelization. The polyhedral compilation framework [6, 18] subsumes most of these loop transformations in a single, unified formalism that abstracts in an algebraic way the iteration space of multidimensional loops and the dependencies between instructions. Polyhedral program transformation has reached a sufficient level of maturity and software robustness to now be integrated in production compilers, as exemplified with LLVM (Polly [23]), GCC (Graphite [57]), or IBM XL/C [29]. But to date, these optimizers are effective only in restricted cases and are severely challenged by polynomial array subscript functions and may-aliasing issues, for instance.

Generally speaking, integrating advanced loop transformation systems in production compilers involves solving problems inherent to the compiler intermediate representation (IR) lowering: array linearization, type erasure and pointers with possibly multiple indirections makes the task of optimizing loops hard. The main reason is that most loop transformations require precise data dependence analysis that is expressed between loop iterations. When employing generic IRs (e.g., LLVM IR [40] or GIMPLE [21]), semantic key information, such as multidimensional array references, are removed (if not already absent at source level), significantly decreases the effectiveness of static analyses.

Static dataflow analysis [13, 17] based on polyhedral abstraction [6, 18] succeeds in computing precise dependencies for array subscript references that use affine access functions of the loop indices [23, 46]. But, whenever their is no known relationship between the base-pointers, or whenever the access functions are not affine, existing loop-based data dependence analyses fail to compute exact dependences. In those very common scenarios, the compiler will conservatively choose not to implement any loop transformation. This is disheartening given the numerous demonstrations of how polyhedral optimizers can produce high-performance code [6, 10, 22, 32, 36, 39, 52] by exposing parallelism and data locality.

Vectorization is an example of loop optimization that suffers from poor alias information: as most current architectures support SIMD instructions (e.g. AVX), enabling aggressive loop vectorization [35] in production compilers became a critical concern. In many cases
production compilers where not able to implement SIMD vectorization. The reason was the presence of spurious dependencies due to conservative assumptions by the compiler about potentially overlapping memory regions being referenced (i.e., aliasing). The practical observation of the inefficiency of static alias analysis led to the development of a specific disambiguation technique for vectorization [34, 43], to enable efficient SIMDization. As opposed to speculative pointer disambiguation, this approach uses code cloning and versioning and determines at runtime which code version can be safely executed.

In a similar spirit, we develop a may-dependence violation checking technique that allows to safely apply loop transformations even in the presence of may-aliasing arrays with polynomial accesses functions. This combination of static analyses and runtime checks allows to assert the correctness of the transformation with regards to (runtime) data dependencies, falling back to the non-transformed code if the runtime validity test fails.

Two instructions are (data-)dependent if they both access the same memory cell and at least one of them is a write. To preserve the semantics, a transformation must preserve the relative execution order of dependent instructions: the producer shall be executed prior to the consumer. One can create a simple runtime check at compile-time (i.e., before the production of the target code), and (b) if so whether they are executed in the same order as in the original program. The purpose of our separation of concerns is to enable efficient SIMDization. As opposed to auto-vectorization [34], to enable efficient SIMDization, we develop a may-dependence violation check for a multiple of 4):

• A new quantifier elimination method for integer-valued polynomial constraints.
• A method to generate simple runtime tests for a variety of critical loop transformations, such as tiling, parallelization, vectorization, etc. to ensure the legality of these transformations.
• An extensive evaluation demonstrating that our elimination scheme is precise and our approach has negligible runtime overhead, thus enabling optimizations.

Sec. 2 shows practical examples motivating the need for our approach, and Sec. 3 provides the theoretical foundations on the program representation we use. Sec. 4 presents the expression of runtime checks for numerous key program transformations, while Sec. 5 presents the variable elimination techniques for polynomial constraints. Experimental results are presented in Sec. 6, and related work is discussed in Sec. 7 before concluding.

2 MAY-DEPEND ILLUSTRATING CASES

In this section we go through a few motivating examples for which state-of-the-art static analyses are typically not able to disambiguate some dependencies, and thus prevent compilers from implementing aggressive program optimizations. Our proposed runtime dependence test, presented in Sec. 4 and beyond, can handle all these cases even for complex loop transformations.

In-place block xMMADD. We consider a multi-dimensional data space (e.g. a 2D-matrix, an image, etc.) and a simple kernel that performs a block-computation (e.g., addition of two sub-matrices) that can possibly be in-place Computations restricted to a sub-block (or “window”) of a multi-dimensional array can be found in numerous function libraries. Linearized arrays (see below) can also occur in such codes (e.g. OpenCV [31]).

In this example, there is no restriction on A nor B, and actually they could be equal. In this context, there is a may-dependence between all read accesses and all write accesses. Output dependencies between accesses to A are the only ones that can be disambiguated stably. Indeed, in the case where A = B, depending on the respective values of ai, aj, bi, and bj, any overlap is possible. Hybrid alias analysis as developed in most production compilers for auto-vectorization [34] could disambiguate the innermost loop as follows 1 (considering vectors of length 4 and K a multiple of 4):

1Note that accesses are shown linearized here as they would appear in the compiler intermediate representation

production of the program representation we use. Sec. 4 presents the expression of runtime checks for numerous key program transformations, while Sec. 5 presents the variable elimination techniques for polynomial constraints. Experimental results are presented in Sec. 6, and related work is discussed in Sec. 7 before concluding.

2 MAY-DEPEND ILLUSTRATING CASES

In this section we go through a few motivating examples for which state-of-the-art static analyses are typically not able to disambiguate some dependencies, and thus prevent compilers from implementing aggressive program optimizations. Our proposed runtime dependence test, presented in Sec. 4 and beyond, can handle all these cases even for complex loop transformations.

In-place block xMMADD. We consider a multi-dimensional data space (e.g. a 2D-matrix, an image, etc.) and a simple kernel that performs a block-computation (e.g., addition of two sub-matrices) that can possibly be in-place Computations restricted to a sub-block (or “window”) of a multi-dimensional array can be found in numerous function libraries. Linearized arrays (see below) can also occur in such codes (e.g. OpenCV [31]).

In this example, there is no restriction on A nor B, and actually they could be equal. In this context, there is a may-dependence between all read accesses and all write accesses. Output dependencies between accesses to A are the only ones that can be disambiguated stably. Indeed, in the case where A = B, depending on the respective values of ai, aj, bi, and bj, any overlap is possible. Hybrid alias analysis as developed in most production compilers for auto-vectorization [34] could disambiguate the innermost loop as follows 1 (considering vectors of length 4 and K a multiple of 4):

1Note that accesses are shown linearized here as they would appear in the compiler intermediate representation
The idea of this technique is to check the overlapping of accessed areas. As an example the first equality expresses the overlapping of intervals \(A[i:i+1][j:j+1] \subseteq A[i:i+1][j:j+1] \cup A[i:i+1][j:j+1]\). Apart from the potential complexity of the runtime test in the cited technique (due to the absence of algebraic simplifications), no existing technique can disambiguate the block regions necessary for register tiling. The main reason is the multi-dimensionality of data structures.

**FW with Triangular Matrix.** Here we consider a symmetric matrix (representing for example a tensor with symmetries) stored in a compact way. \(T[i,j]\) is represented as \(T[1*(i-1)/2+j]\) if \(i \geq j\) and \(T[j*(j-1)/2+i]\) otherwise. Such matrix representation (named “packed”) is used by most high-performance linear algebra libraries such as LAPACK [37], or MKL [30] for triangular, symmetrical, or Hermitian forms. Boost [7] also uses packed representation for banded matrices. The example used in later Sec. 3 which corresponds to one part of the computation of Floyd-Warshall with symmetries uses this representation. It turns out that such code would highly benefit from tiling [60], as the data reuse available has similarities with matrix-matrix multiplication. However, none of the existing dependence analyses (even hybrid ones) can disambiguate this code. The reason is the presence of polynomials in access functions. Obviously delinearization techniques that are based on pattern matching cannot infer the 2D structure of memory accesses [25]. Any compiler will fail at tiling this loop nest.

**ADI with Linearized Array.** Here we consider some computation on linearized arrays. An access to cell \((j, i)\) of array \(T[H][H]\) is written as \(T[H][H][j+i]\). Multi-dimensional array linearization is widely present in high-performance legacy code. One of the reason is that languages such as C99 or Java (e.g. Colt [14], EJML [16]) implement multi-dimensional arrays as pointers of pointers (even for fixed-size arrays in Java) which turns out to be less efficient than the equivalent linearized version. Also, even if some programming languages support packed multi-dimensional arrays, some compiler intermediate representations (e.g. LLVM-IR [38]) will not support it and linearization is performed early in the compilation stages. As an example we consider a simple 2D loop nest that interleaves two sweeps, one on each direction.

For the same reason than for the previous example (FW), as linearization introduces polynomials, static dependence analysis will fail in disambiguating the memory accesses. Also, depending on the actual values of \(H\) and \(n\) at runtime, loop interchange might not be valid. Without code cloning, the compiler has to be conservative regarding the presence of data dependencies and cannot fully optimize the loop nest.

### 3 THEORETICAL FOUNDATIONS

To illustrate our notations and terminology, consider the example code below. We consider nested loops so that any statement is uniquely identified by the iteration vector \((k, i, j)\) here.

```plaintext
for (k = 2; k < n; k++) // j < k
  for (i = 1; i < k; i++)
    for (j = 0; j < k; j++)
      d[i*(i-1)/2 + j] = d[i*(i-1)/2 + j] + d[k*(k-1)/2 + j];
for (k = 1; k < n-1; k++) // j < k
  for (i = k + 1; i < n; i++)
    for (j = 0; j < k; j++)
      d[i*(i-1)/2 + j] = d[i*(i-1)/2 + j] + d[k*(k-1)/2 + j];
```

Without loss of generality, we use a canonical representation of the nested loop [18] where to each iteration vector is associated an unique statement instance. As an example the previous example is equivalently modeled as follows:

```plaintext
for (t = 0; t < 2; t++)
  if (t = 0)
    for (k = 2; k < n; k++) // j < k
      for (i = 1; i < k; i++)
        for (j = 0; j < i; j++)
          S1: ...
  else if (t = 1)
    for (k = 1; k < n-1; k++) // j < k
      for (i = k + 1; i < n; i++)
        for (j = 0; j < k; j++)
          S2: ...
```

In this form any statement is indexed by a vector of possibly larger size than the original loop depth \((t, k, i, j)\) here. The iteration domain of a nested loop is a geometric representation of all possible values of loop indices. It can be an over-approximation of it [3]. In our example, the iteration domain for statement S1 is \(D_{S1} = \{(t, k, i, j) : t = 0 \land 0 \leq j \leq i < k < n\}\). In this paper we restrict the iteration domain to be an union of convex polyhedra, in other words it can be represented as a disjunctive normal form of affine inequalities.

A schedule is expressed as a function (say \(T_{S1}\) for statement S1) from \(D\) to \(D' \subseteq \mathbb{Z}^2\) a vector space of dimension \(s \geq 1\) (\(\mathbb{Z}\) represents the integers). \(T_{S1}(0, k, i, j) = T_{S1}(0, k', i', j')\) means that \(S_{1,0, k, i, j}\) and \(S_{1,0, k', i', j'}\) are executed in parallel. \(T_{S1}(0, k, i, j) <_{lex} T_{S1}(0, k', i', j')\) means that \(S_{1,0, k, i, j}\) is executed before \(S_{1,0, k', i', j'}\). A loop transformation is expressed as a schedule, and most composition of loop transformations can be represented as affine schedules [6, 18, 22]. For example, interchanging loops \(i\) and \(j\) for the loop nest encapsulating S1 would lead to \(T'_{S1}(0, k, i, j) = (0, k, j, i)\). The canonical form allows to express the schedule of the original code as the identity function from \(D\) to \(D\).

A data dependence between two statement instances expresses an ordering (represented as \(\angle_d\)) constraint of their respective execution. A data dependence is due to read and write accesses to the same memory location. Here \(S_{1,0, k, i, j}\) writes to location \(W_t(0, k, i, j) = T + i(i-1)/2 + j\) and reads to locations \(R_2(0, k, i, j) = T + i(i-1)/2 + j, R_3(0, k, i, j) = T + k(k-1)/2 + i, R_4(0, k, i, j) = T + k(k-1)/2 + j\). There is a read-after-write (flow) dependence between \(S_{1,0, k, i, j}\) and \(S_{1,0, k', i', j'}\) from \(W_t\) to \(R_2\) as soon as: (1) \((0, k, i, j) \in D\) strictly precedes \((0, k', i', j') \in D\) in the original schedule; which is written as \((0, k, i, j) \angle_{lex} (0, k', i', j')\) (\(\angle_{lex}\) represents a lexicographic comparison); (2) read/write locations overlap; which is written (for atomic data) as \(W_t(0, k, i, j) = R_2(0, k', i', j')\). As an example, \(S_{1,0, 3, 2, 1} <_d S_{1,0, 4, 2, 1}\) as \(W_t(0, 3, 2, 1) = R_2(0, 4, 2, 1) = T + 4\). A read-after-write is a flow dependence; A write-after-read is an output dependence; A write-after-read is an anti-dependence.

\[a_1, \ldots, a_n <_{lex} (b_1, \ldots, b_m) \iff \text{there exists an integer } 1 \leq i \leq \min(n, m) \text{ s.t. } (a_1, \ldots, a_i) = (b_1, \ldots, b_i) \text{ and } a_i < b_i.\]
For any schedule to be valid it must preserve all such dependences. As an example, for the previously considered loop interchange $T_{S1}(0, k, i, j) = (0, k, j, i)$, the dependence $S10_{\text{orig. sched.}} \leq_d S10_{\text{orig. sched.}}$ is preserved as $T_{S1}(0, 3, 2, 1) = (0, 3, 1, 2) \prec_{\text{lex}} T_{S1}(0, 4, 2, 1) = (0, 4, 1, 2)$. The loop interchange is valid as soon as all dependences are preserved.

4 Dependence Validity Tests

Our approach is based on checking at runtime whether the optimized or the original code should be executed. The transformation is correct (with regard to data dependencies) as soon as the transformed schedule does not contain any violated dependence. The following system of equations, where a dependence is represented as a point $(v, v')$ in a multi-dimensional space, defines the exact set of violated dependencies. Let us call this set, the violated set. Here, existential quantifiers are used to express non-emptiness of this set. The contribution of this work is to eliminate those quantifiers.

$$\bigvee_{(A, A') \in mD} \exists (v, v') \in \mathbb{Z}^{2d} \text{ s.t.}$$

$$(\text{exist. quantifier})$$

$$(\text{domain})$$

$$(\text{orig. sched.})$$

$$(\text{same access loc.})$$

$$(\text{violation})$$

where: (1) $d$ is the dimension of the loop nest in canonical form; (2) $A$ and $A'$ are access functions (e.g. $W_i$ and $R_2$ in the example of Section 3), one of them necessarily being a write access (to generate a may-dependence); (3) $D_A$ (resp. $D_{A'}$) is the set of points in the domain $A$ (resp. $A'$) that is accessed (access $S1_{\text{orig. sched.}}$ for $W_i$ and $R_2$ in our example); (4) $mD$ represents all possible pairs of such functions that could lead to a may-dependence (a dependence the compiler cannot disambiguate statically); (5) $T_A$ (resp. $T_{A'}$) represents the schedule of the statement including access $A$ (resp. $A'$) after transformation ($T_{S1}$ in our example).

Observe that emptiness (overall condition with quantifiers that evaluate to false) of the violated set is a sufficient condition for the transformation to be valid. So it is safe to over-approximate it (it is obviously not to under-approximate it), but if too loose it may lead to conservatively decide not to apply an even safe transformation.

In that sense, precision (tightness of the over-approximation) of the quantifier elimination process is important to make the overall approach of hybrid analysis effective. We now illustrate this system of equations through different loop transformations and applications.

Interchange transformation can be used to expose parallelism and locality. Take ADI as an example and the may-dependence $(A, A') = (R_2, W_1)$, where $A(i, j) = R_2(i, j) = -H + Hj + i$ and $A'(i, j) = W_1(i, j) = Hj + i$. Consider loop interchange as the desired transformation, i.e. $T(i, j) = (j, i)$. The system of equations writes as follow:

$$\exists (i, j, i', j') \in \mathbb{Z}^4 \text{ s.t.}$$

$$(\text{domain})$$

$$(\text{orig. sched.})$$

$$(\text{same access loc.})$$

$$(\text{violation})$$

$$(i, j) \prec_{\text{lex}} (i', j')$$

writes as $i < i'$ and thus (orig. sched.) \& (violation) simplifies to: $i < i'$ \& $j < j'$.

Skewing such as $T(i, j) = (i, i + j)$ can be used to enable interchange and tiling. Combined with an interchange, (violation) writes as $(i + j, i) \prec_{\text{lex}} (i' + j', i')$.

Parallelization is valid if there are no loop-carried dependencies. To illustrate this point we consider ADI with the above skewing $(i, j) = (i + j, j)$ that exposes parallelism along the innermost loop. Expressing that we want this loop to be parallel simply consists in folding it, getting the schedule $T(i, j) = (i + j)$. In that case the system of inequalities (for may-dependence $(R_2, W_1)$) simply writes as follows:

$$\exists (i, j, i', j') \in \mathbb{Z}^4 \text{ s.t.}$$

$$(\text{domain})$$

$$(\text{domain})$$

$$(\text{orig. sched.})$$

$$(\text{same access loc.})$$

$$(\text{violation})$$

$$(i, j) \prec_{\text{lex}} (i', j')$$

as $(i + j, i) \prec_{\text{lex}} (i' + j', i')$.

Loop Invariant Code Motion means that (in terms of dependence the hoisted code does not depend on any computation inside the loop. Consider the statement $S1$ in the example of Section 3. The load $d[k \times (k-1)/2+i]$ can be hoisted outside the inner most loop $j$. Considering may-anti-dependence with the write operation at the same statement $d[i \times (i-1)/2+j]$, we get the following system of inequalities:

$$(\text{domain})$$

$$(\text{domain})$$

$$(\text{orig. sched.})$$

$$(\text{same access loc.})$$

$$(\text{violation})$$

Vectorization. Consider again the ADI case after skewing $(T(i, j) = (i + j, j))$. There exists two ways to test for vectorization. The first and precise way to express the condition to vectorize the new inner-most loop by a vector of size $l$ is to use: $T(i, j) = (i + j, j/l)$. In that condition the system of equations for $(R_2, W_1)$ may-dependence writes:

$$\exists (i, j, i', j') \in \mathbb{Z}^4 \text{ s.t.}$$

$$(\text{domain})$$

$$(\text{domain})$$

$$(\text{orig. sched.})$$

$$(\text{same access loc.})$$

$$(\text{violation})$$

Note that floor (or modulo) can be handled by adding two new variables (say $r$ and $d$): $[i/l] = d \mod 0 \leq r < l \& d + r = j$. Alternatively the (violation) condition (knowing that $i \leq i'$ from the (orig. sched.) condition can also be relaxed into $i' - j < l \& i' \geq j'$.

Note that floor (or modulo) can be handled by adding two new variables (say $r$ and $d$): $[i/l] = d \mod 0 \leq r < l \& d + r = j$. Alternatively the (violation) condition (knowing that $i \leq i'$ from the (orig. sched.) condition can also be relaxed into $i' - j < l \& i' \geq j'$.

Here, the strict inequality in the (orig. schedule) constraint allows to optimize this case.
Loop-Body Scheduling / Software Pipelining. As observed above, the necessary condition for vectorization of size \( l \) to be valid is that all loop-carried dependences must be of distance greater or equal to \( l \). Dependence distance (which actually corresponds to a lower-bound, not to an exact distance) is used by standard loop-body scheduling used for EPIC architectures. For any given dependence distance, the system of inequalities is very similar to the one constructed for vectorization.

Tiling / Permutability. The legality of loop tiling can be assessed by detecting that a set consecutive loops are permutable, and tiling can then be implemented in a general way [6]. Tiling exposes locality and parallelism simultaneously at several levels. Rectangular tiling (possibly after some skimming) is a very powerful loop transformation [60]. Let us consider again the example given in Section 3 and suppose we want to tile the loop nest enclosing \( S1 \). The band \((k,i,j)\) of the domain \( D_{S1} = \{(0,k,i,j), 0 \leq j < i < k < n\}\) is fully permutable if individually \((k,i),(k,j),\) and \((i,j)\) can be interchanged [48]. The conditions can actually be expressed in a more condensed way [2]: all dependencies should be within the positive cone. Restricting to may-dependence variety of complex compositions of loop transformations, including interchange, skewing, fusion, distribution, shifting/retiming, peeling, vectorization, parallelization, etc.

5 QUANTIFIER ELIMINATION

This section describes our elimination scheme that allows to reduce the complexity of the validity test to \( O(1) \). The presence of an existential quantifier means that a loop is needed at runtime to explore possible values. Eliminating a quantifier eliminates the need for the corresponding loop. Geometrically speaking, the quantifier elimination corresponds to perform an orthogonal projection of the violated set along the eliminated dimensions. As already mentioned this set can be safely over-approximated and so is its projection. But the tightest the better, and as we will see further, real-valued relaxation (our initial problem is an integer-valued) usually leads to too loose test.

Our method to perform quantifier elimination is an extension of the Fourier-Motzkin Elimination (FME) method that handles integer multivariate-polynomials. Figure 1 presents an overview of the process. For this section, let \( S(P, V) : \{E_0, E_1, \ldots, E_n\} \geq 0 \land \{E'_0, E'_1, \ldots, E'_m\} = 0 \) be our system of constraints. \( E \) and \( E' \) are expressions over \( (P, V) \). \( V \) represents the set of quantifiers (loop variants / variables) to be eliminated. \( P \) a set of free-variables (loop invariant / constants). Our quantifier elimination will project \( S(P, V) \) into \( S'(P) \).

Schweighofer Tester. One of the most important part of the algorithm relies in the ability to prove, whenever possible, the sign of an expression in a given system of constraints. To this end we use the theorem described by [54] to evaluate expressions’ signs under a system of constraints 3.

Given a system of inequalities \( S : \{E_1, E_2, \ldots, E_n\} \geq 0 \), Schweighofer provides a method to evaluate if the inequality \( I : E \geq 0 \) is implied by \( S \). If \( I \) can be written as the sum of the product of any power of the inequalities of \( S \cup \{1\} \) multiplied by non-negative factors (multivariate-polynomial in \( E_1, \ldots, E_n \) with positive coefficients), then \( S \) implies \( I \). As an example, if the system \( S = \{E_1 = -xy - x \geq 0 \land E_2 = x - 1 \geq 0 \land E_3 = y \geq 0\} \) is true, then \( y \geq 1 \). Indeed, \( y - 1 = 2E_3 + E_2 + E_1 \). Unfortunately, raising a subset \( S' \subset S \) of cardinality \( m \) to degree \( d \), leads to a polynomial (in \( E_1 \)) with a number of terms combinatorial with \( m \) and \( d \). Take as an example the subset \( S' = \{E_1, E_2\} \) we have \( m = 2 \) and \( d = 2 \).

The goal (to prove the positivity of \( E \)) is to find coefficients \( \lambda_i \) of the polynomial \( \lambda_0 + \lambda_1E_1 + \lambda_2E_2 + \lambda_3E_1E_2 - E = 0 \) that are positive. This is done by expanding the polynomial and building a linear program by matching the coefficient of each of its term to 0. In our case, the expanded form \((\lambda_0 + 1) + (\lambda_1 + \lambda_2 + \lambda_12)x - y + (\lambda_1 + \lambda_2)xy - \lambda_12x^2 - \lambda_12xy^2 = 0 \) leads to the linear program \( \{\lambda_i \geq 0 \land \lambda_0 + 1 = 0 \land -\lambda_1 + \lambda_2 + \lambda_12 = 0 \land -1 = 0 \ldots \} \) that has no solution in that case. Our implementation chooses, heuristically, subsets \( S' \subset S \) of bounded cardinality \( m \), and limits the maximum degree \( d \) to which \( S' \) are raised. Solving the LP is heuristically done by using a linear solver where positiveness of the \( \lambda_i \)'s is enforced using a simple heuristic. Whether the precision gain using a LP solver is worth compared to the overhead has to be determined. Observe that the result to the LP possibly gives a more accurate inequality as it allows (in case of success) to prove that \( E \geq 0 \). As discussed further maximizing \( \lambda_0 \) can be used as the objective function of the linear program when looking for tight bounds.

The sign detection of an expression \( E \) is done by evaluating if \( S \) implies \( E \geq 0 \) or \( S \) implies \( -E \geq 0 \). When both are implied, if one of the inequalities can be made strict, then \( S \equiv \text{false} \), otherwise \( E = 0 \). Whenever none of the inequalities can be proven to be implied by any \( S' \), \( E \) is factorized, if possible, and its sub-expressions are recursively evaluated, with particular rules for exponential, multiplication and sum. To avoid replicated sign tests a cache of signs application and sum. To avoid replicated sign tests a cache of signs system changes, the tester must be rebuilt, clearing the cache.

Fourier-Motzkin Lowering (FML). The Fourier-Motzkin elimination (FME) is a well known algorithm for removing variables of linear inequality systems on \( \mathbb{R} \). For a given linear system \( S \), the
Fourier-Motzkin Lowering
Choose Variable
Select variable with lowest
degree as target t

Check Signs
For all inequalities F \cdot t \geq P
Decide sign of F

Eliminate
Combine bounds of t, lowering its
degree.

Cut Sign
Split the system into 3:
Si(\alpha > 0)
Si(\alpha=0)
Si(\alpha < 0)

Simplify
System S1′ implies S2′?
Remove S1′

Prune
System S1′ implies S2′?
Remove S1′

Output:
S1' or S2' or ... Sm'
over Parameters

Figure 1: Quantifier Elimination scheme. Input and output are sets of conjunctive systems of polynomial inequalities. “A” stands for affine hyper-planes.

FME works as follows: (1) Consider the variable t ∈ V and a set of inequalities S on that variable; (2) isolate t so that any inequalities from S can either be written as:

\[ t \leq U_j(P, V \setminus \{t\}) \quad \text{(upper bounds)} \]
\[ t \geq L_j(P, V \setminus \{t\}) \quad \text{(lower bounds)} \]

(3) eliminate t from the system by combining each lower-bound with each upper-bound:

\[ \forall (i,j), \quad L_j(P, V \setminus \{t\}) \leq U_j(P, V \setminus \{t\}) \]

The basic FME is designed for a system made up of linear inequalities on \( \mathbb{R} \). As we will see in Section 7 there exists several extensions. One considers linear inequalities but possibly with parameters as coefficients [27]. Another one considers linear inequalities on \( \mathbb{Z} \). The goal here, is to extend the method to polynomial inequalities on \( \mathbb{Z} \). The main difference with a linear system is that each elimination step (of the Convergence mode – see below) actually consists in lowering the degree of the variable t. The first step, Choose Variable, selects the variable t to eliminate, as the one with lowest degree. The second step, Check Signs has two modes. In the Precision mode, it writes every inequalities as:

\[ L_j(P, V \setminus \{t\}) \leq C_j(P, V) \cdot t \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]

where each \( L_j \) and \( U_j \) have degree in t (say t-degree) strictly lower than \( \alpha \). In both modes, inequalities with non-null coefficients \( C_j \) are normalized to the form

\[ L_j(P, V) \leq C_j(P, V) \cdot t^\alpha \]
elimination. Our normalization extends the one from [51] to non-numerical coefficients. For any inequality in the form \( F(P) - G(P, V) \geq H(P) \), if Schweighofer Tester proves that \(|F(P)| \geq |H(P)|\), then the inequality becomes \( \pm G(P, V) \geq \{\pm 1, 0\} \), depending on the signs of \( F \) and \( H \).

After applying Pre-conditioning, if the system has no more free-variable, the process branches to Simplify phase. If some free-variables remain, but changes made to the system, the entire process is restarted (Schweighofer Tester). If no changes have been performed then it proceeds to the variable elimination stage (Fourier-Motzkin Lowering).

Simplify. Once all quantifiers have been eliminated, a last phase consists in: 1. simplifying each inequality; 2. eliminating redundant ones or proving its contrary. Hence, 1. Function Simplify takes each inequality \( E \geq 0 \) of the system and factorizes it (as much as possible) into \( E = \bigwedge F_i \geq 0 \). Then, any factor \( F_i \) for which the (strict) sign can be proven (using the other inequalities) is removed. Observe that the orientation of the inequality can be potentially reversed depending on the number of negative factors. 2. Once this is done, for every inequality \( I \cdot E \geq 0 \) we test if \( |S \cdot I| \) implies \( I \).

If it does then \( I \) is removed; If it can be proved that \( |S \cdot I| \) implies \( E < 0 \), then a contradiction \( (I \land \neg I) \) is found, and the system is turned into false.

Prune. The last “simplification” step acts over the disjunction formed by all systems \( S'_0(P) \). Using the Schweighofer Tester, if a system \( S'_0 \) implies all inequalities of a system \( S'_1 \), then \( S'_1 \) is removed. All pairs are tested.

The output of this method is a disjunctive normal form of systems over the parameters \( P \). If (only) a single one of these systems evaluates to true, then it is unsafe to use the optimized version of the loop.

6 EXPERIMENTAL RESULTS

Framework implementation. We recall the three steps of our framework. Given a (sequence of) loop transformation(s) that forms the tentatively optimized program, and a representation of the original program as described in Sec. 3, our framework: (1) builds sufficient conditions to prove the validity of the transformation with regard to may-dependences; (2) eliminates loop indices in the constructed validity test; (3) incorporates the simplified test at the entry point of the program region, deciding on branching to the optimized version or fallback to the original (un-optimized) program version.

We have implemented an end-to-end prototype of the framework presented in this paper. Our prototype can handle C code with loops and array access functions containing polynomial expressions. ASTER extracts the program representation. PoCC computes a program transformation for this representation and emits ASTER directives to describe its transformation. ASTER generates the complete validity test (with complexity in the order of \( O(n^2d) \) where \( d \) is the loop depth), which is fed to our implementation of the quantifier elimination strategy. Glue scripts are used to create a binary using versioning, containing the optimized code and the original one (fallback), the one being executed at runtime depending on the result of the evaluation of the simplified test.

6.1 A Simple Example

We start by demonstrating the capacity to generate tight runtime tests using a simple example that corresponds to the code of X-MADD. As input to our toolchain, we provided the original code along with pragmas to specify the transformation for which a validity test is to be generated.

Here, we mark the inner-most loop for parallelization (e.g., for SIMD). This program is fed to ASTER, which produces the complete (with free-variables) system of constraints of the validity test. This system is then passed to our quantifier elimination tool, which after simplification successfully eliminates all existential variables (i.e., loop indices) and ends up generating the following runtime test:

```
if ( | (aj+aia2N2+2aj+aiaN2) || (aj+aia2NK+2aj+aiaN2) || (aj+aiaNK+aj+aiaN2) || (aj+aiaNK+aj+aiaN2) || (aj+aiaNK+aj+aiaN2))
    (aj+aiaNK+aj+aiaN2) || (aj+aiaNK+aj+aiaN2) || (aj+aiaNK+aj+aiaN2) || (aj+aiaNK+aj+aiaN2))
else ( optimized )
```

If at least one of the disjunction evaluates to true then the code branches to the original (non-optimized) version. Note that evaluating this test at runtime is extremely quick in comparison to the naive test produced by ASTER, thanks to the elimination of quantifiers.

For comparison purpose we have also evaluated QEPCAD-B [8] for the simplification process. Here QEPCAD-B determines the system to be false for the \texttt{WAW} dependence, but fails to process the systems for all remaining dependences.

6.2 Evaluation

We now demonstrate the ability for our technique to generate a low-overhead runtime validity test for complex polynomial array index functions that may alias. We also present experimental data about the performance improvements that can be reached by enabling aggressive loop transformations on these programs. We emphasize here that our purpose is not to find the utmost best transformation for each program, but instead to demonstrate the potential impact of implementing complex polyhedral transformations safely, after a complete may-alias and validity runtime check.

Experimental protocol. All experiments were conducted on a 4-core Intel Haswell Core i5-4590 CPU @ 3.30GHz running Fedora 25, with Turbo boost and DVFS/frequency scaling turned off. All benchmarks have been compiled with GCC 6.3.1 with flags `-O3 -fopenmp -march=native -mtune=native` and ICC 17.0.2 with flags `-O3 -xHost -parallel -fopenmp. Each experiment was run 20 times, and we report the average performance. Note that the standard deviation of the execution time never exceeded 2% in any of the experiments we conducted. We use Polybench/C 4.1 with DATA_TYPE_IS_DOUBLE and MEDIUM_DATASET.
Table 1: Results summary. Original, Optimized GF/s and Speed-up of the applications when compiled gcc or icc. If it was
various domains, including linear algebra, image processing, or
of the solved systems. The Transformation performed is reported, as described in Sec. 6.2.

<table>
<thead>
<tr>
<th>Appli. name</th>
<th>gcc - GF/s</th>
<th>Spd</th>
<th>icc - GF/s</th>
<th>Spd</th>
<th>Ex</th>
<th>Op</th>
<th>#P</th>
<th>Naive test - O(n^2)</th>
<th>Simplified test</th>
<th>QEPCAD-B</th>
<th>Trans</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ori</td>
<td>opt</td>
<td>upx</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cmp</td>
<td>#Ineqs</td>
<td>#Sys</td>
<td>Cmp</td>
</tr>
<tr>
<td>2mm</td>
<td>2.0</td>
<td>9.6</td>
<td>4.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3mm</td>
<td>1.4</td>
<td>11.0</td>
<td>8.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>atax</td>
<td>1.8</td>
<td>5.8</td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bicg</td>
<td>1.2</td>
<td>5.8</td>
<td>4.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>choles</td>
<td>4.3</td>
<td>4.3</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correl</td>
<td>1.3</td>
<td>2.2</td>
<td>1.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>covari</td>
<td>1.3</td>
<td>12.1</td>
<td>9.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>doigate</td>
<td>1.8</td>
<td>9.0</td>
<td>5.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FW</td>
<td>4.5</td>
<td>9.2</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gemm</td>
<td>9.3</td>
<td>36.4</td>
<td>3.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gemver</td>
<td>2.6</td>
<td>5.8</td>
<td>2.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gesumm</td>
<td>1.2</td>
<td>0.9</td>
<td>0.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gramsc</td>
<td>1.0</td>
<td>2.9</td>
<td>2.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>heat3D</td>
<td>11.7</td>
<td>26.9</td>
<td>2.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jaco1D</td>
<td>8.5</td>
<td>4.2</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jaco2D</td>
<td>8.0</td>
<td>11.4</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lu</td>
<td>1.3</td>
<td>1.3</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mvt</td>
<td>1.2</td>
<td>4.3</td>
<td>3.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seidel</td>
<td>0.9</td>
<td>0.9</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>symm</td>
<td>2.9</td>
<td>3.6</td>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sy2k</td>
<td>1.6</td>
<td>2.2</td>
<td>1.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>syrk</td>
<td>5.1</td>
<td>25.9</td>
<td>5.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trisol</td>
<td>1.0</td>
<td>1.2</td>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trmm</td>
<td>1.6</td>
<td>1.9</td>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FW-tri</td>
<td>2.2</td>
<td>3.1</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Results summary. Original, Optimized GF/s and Speed-up of the applications when compiled gcc or icc. If it was
various domains, including linear algebra, image processing, or
datataming. PolyBench natively contains affine program regions
using multidimensional arrays. We transformed each such array access into a linearized form, e.g., A[i][j] for an array A[N][N+1] is rewritten as A[i*(N+1)+j]. Note that in practice many compilers,
e.g., LLVM, will linearize the multidimensional array accesses leading
to non-affine access functions, and preventing optimization with
off-the-shelf polyhedral optimizers. In specific cases techniques like optimistic delinearization may be applied [25], yet issues such as
array padding may quickly prevent the technique to be applied
even on PolyBench [25]. We show that in contrast our approach can lead to systematic successful analysis and optimization of linearized PolyBench/C benchmarks.

Note that codes with polynomial (e.g., linearized) array accesses do not arise only from compiler IR lowering. For example Netlib BLAS kernels on symmetric matrices, e.g. in [37], use compacted storage leading to polynomial access functions. We leave the evaluation of such codes for future work, as our current prototype does not support memory address registers, used to compute an array displacement that depends on the loop induction variable.

Candidate optimization generation. We have generated one candidate optimization per benchmark. This optimization is the result of a complex polyhedral transformation implemented by the PoCC compiler [47]. We have extended PoCC to output, in addition to the transformed program, a series of pragmas describing the transformation applied. ASTER reads this information to build the initial validity test as described in Sec. 4, before we perform quantifier elimination.

The optimization candidate was generated using high-level PoCC optimization primitives, such as the PLuTo optimizer [5] which computes a sequence of loop transformations to enable parallelism and data locality enhancements, typically using loop tiling. This transformation, taking the form of a multidimensional schedule for each statement, can be seen as a sequence of high-level loop transformations. We report in Table 1, last column, the transformation as a sequence of fusion, Fission, loop interchange, thread-level parallelism, skewing, tiling, vectorization.

We recall that our objective is not to demonstrate the maximal performance achievable for each benchmark, but only to demonstrate the feasibility and potential impact of using polyhedral transformations on non-affine programs. For example, we have not tuned the tile sizes, or did not perform extensive exploration of the performance of various possible polyhedral optimizations. We mostly followed the data from Park et al. [45] which points to good PoCC optimizations for PolyBench.

Benchmarks. We focus our evaluation on the PolyBench/C 4.1 [49] benchmark suite. It contains a variety of numerical kernels from various domains, including linear algebra, image processing, or datataming. PolyBench natively contains affine program regions using multidimensional arrays. We transformed each such array access into a linearized form, e.g., A[i][j] for an array A[N][N+1] is rewritten as A[i*(N+1)+j]. Note that in practice many compilers, e.g., LLVM, will linearize the multidimensional array accesses leading to non-affine access functions, and preventing optimization with off-the-shelf polyhedral optimizers. In specific cases techniques like optimistic delinearization may be applied [25], yet issues such as array padding may quickly prevent the technique to be applied even on PolyBench [25]. We show that in contrast our approach can lead to systematic successful analysis and optimization of linearized PolyBench/C benchmarks.

Note that codes with polynomial (e.g., linearized) array accesses do not arise only from compiler IR lowering. For example Netlib BLAS kernels on symmetric matrices, e.g. in [37], use compacted storage leading to polynomial access functions. We leave the evaluation of such codes for future work, as our current prototype does not support memory address registers, used to compute an array displacement that depends on the loop induction variable.

Candidate optimization generation. We have generated one candidate optimization per benchmark. This optimization is the result of a complex polyhedral transformation implemented by the PoCC compiler [47]. We have extended PoCC to output, in addition to the transformed program, a series of pragmas describing the transformation applied. ASTER reads this information to build the initial validity test as described in Sec. 4, before we perform quantifier elimination.

The optimization candidate was generated using high-level PoCC optimization primitives, such as the PLuTo optimizer [5] which computes a sequence of loop transformations to enable parallelism and data locality enhancements, typically using loop tiling. This transformation, taking the form of a multidimensional schedule for each statement, can be seen as a sequence of high-level loop transformations. We report in Table 1, last column, the transformation as a sequence of fusion, Fission, loop interchange, thread-level parallelism, skewing, tiling, vectorization.

We recall that our objective is not to demonstrate the maximal performance achievable for each benchmark, but only to demonstrate the feasibility and potential impact of using polyhedral transformations on non-affine programs. For example, we have not tuned the tile sizes, or did not perform extensive exploration of the performance of various possible polyhedral optimizations. We mostly followed the data from Park et al. [45] which points to good PoCC optimizations for PolyBench.
Comparison with QEPCAD-B. We compared our approach against QEPCAD-B version b 1.69 [8], a publicly available quantifier elimination implementation that uses Partial Cylindrical Algebraic Decomposition. As most of our systems contain more than eight variables, we allow QEPCAD-B to execute with the maximum amount of memory cells we could use (by using -N1073741822). As, QEPCAD-B turns out to be quite slow, the overall number of systems being close to 2000, we limited the execution time per system to 30 minutes. However, as we will see further, most of the time QEPCAD-B gives up, running out of memory.

Detailed results. Table 1 presents the results on PolyBench/C 4.1, using the medium dataset size where programs typically run in a few milliseconds, to exacerbate any potential effect of executing the runtime test.4 We report for each case statistics on the size of the runtime test, before (naive) and after (simplified) applying our quantifier elimination strategy. Note that in all cases, the simplified test has a complexity of O(1), meaning no loop is occurring in the simplified test. The runtime test computation compared to the total region execution time is negligible. It amounts here to evaluating a single large conditional, with no loop in it. In our experiments, we measured this overhead and confirmed it is marginal, typically well below 0.01% of the program execution time.

We also present results using a variation of PolyBench’s Floyd-Warshall benchmark, where we use a compact (triangular) representation. No existing polyhedral technique, including overapproximation, could generate a tiling-based optimization due to the purely polynomial access. With our proposed scheme, off-the-shelf polyhedral optimizers can be seamlessly applied on programs with polynomial array access functions, leading to potentially high gains as shown in Table 1.

Precision. Elimination techniques could generate correct but very conservative solutions, leading to a condition that always evaluates to true even if the transformation is in fact safe. In our experiments we compare our results to QEPCAD-B. The numerous failures (that is, a value > 0 in the #to and #err columns) illustrate the limitations of using QEPCAD-B in this context. The advantage of our FML scheme is that it allows the use of cutting planes on-demand, allowing to divide the space into different systems that can be processed in parallel and have their resulting systems recombined when all quantifiers have been eliminated without lost of precision.

To evaluate the precision of our technique, we ask for sophisticated transformations to be done by the optimizer, including vectorization, tiling and parallelization, leading to complex systems to be simplified.

As we can see in Table 1, despite the high complexity of the initial systems in our experiments, our scheme leads to a test that turns out to be tight, allowing the use of the optimized code in all cases.5 Our FML scheme successfully processed all systems within the time limit, while QEPCAD-B only successfully produced a final result to jacobid, mvt and trisolv applications. For all remaining tests it timed out or failed (running out of memory) to generate the quantifier free formula.6 To evaluate the precision of QEPCAD-B, we replaced all system that it failed to solve with our (precise) solution, so as to build a complete test. For six benchmarks (bigc, correlation, covariance, FW, gemm, gemver) the test fails to validate the transformation at runtime showing that for each of those benchmarks, at least one of the system generated by QEPCAD-B is not precise enough.

Focusing on the Floyd-Warshall algorithm, for a symmetric distance graph stored in a packed lower triangular matrix we obtain four different loops, depending on the size relations between i, j, k. In this example we described a distinct optimization per loop nest, as the following:

```c
void kernel_floyd_warshall_tri (int n, double *D)
int i, j, k, I, J, K;
#pragma scop
for (k)/\forall k < j < i < n
for parallel (i)
for (j)
51: D[i*(i-1)/2+j] = min (D[i*(i-1)/2+k] + D[(j-1)/2+k]);

for (k) //\forall k < j < i < n
for parallel (i)
for (j)
52: D[i*(i-1)/2+j] = min (D[i*(i-1)/2+k] + D[k*(k-1)/2+j]);

for (k)/\forall k < j < i < n
for parallel (i)
for (j)
53: D[i*(i-1)/2+j] = min (D[(k-1)/2+j] + D[k*(k-1)/2+j]);

for (K, step = 32) //\forall i < j < k < n
for (I, step = 32)
for (J, step = 32)
for (k)
for parallel (j)
54: D[j*(j-1)/2+i] = min (D[k*(k-1)/2+j] + D[j*(j-1)/2+k]);
```

A total of 385 systems were generated and our quantifier elimination evaluated all to false. QEPCAD-B fails to process any of the 385 systems given a 30 min limit per system.

7 RELATED WORK

Related work reported here concerns: 1. static/dynamic analysis, and speculative execution as techniques (similar to our) to overcome the limitations of static dependence analysis and poor alias information; 2. Array delinearization that recovers multidimensional data-structure sizes by a pattern matching technique; 3. The handling of polynomials and in particular quantifier elimination schemes.

Static/dynamic analysis and speculative execution. Most of the existing work which goal is to expose parallelism and locality on code with poor alias/dependence information exploits speculative execution framework. As opposed to the vast majority that are restricted to instruction level parallelism (e.g. [11, 15, 20, 28]), the work of Jimborean at al. [33] later improved by [41, 55] allows to perform arbitrary complex polyhedral transformations: The idea is to instrument memory accesses during execution; if a linear pattern is observed, then an optimizing transformation is applied, expecting the access pattern to hold. Without guaranties that the memory access prediction would hold, a roll-back to a safe point must be done if a misprediction is detected. The main limitation of this approach is the overhead due to memory instrumentation and backtracking support. Our work based on hybrid analysis allows to avoid speculation as the validity of the transformation is

---

4Note that 6 of the 30 PolyBench/C 4.1 codes are missing, as they are not supported by the version of PoCC we used.
5By design of our experiments, the transformation we apply is always valid, so an accurate-enough validity test should always lead to executing the transformed code.
6The time constraints we did not evaluate heat-3d test with QEPCAD-B.
checked before executing the optimized code in an inspector/executor manner. In the approaches advocated by Venkat et al. [58], the inspector corresponds to loops that actually inspect the data (i.e. the test complexity is not $O(1)$). This approach is motivated as it focuses on indirect indexes through read-only indexing arrays. Hybrid analysis with $O(1)$ runtime checks is used in [1, 27, 34, 51, 53] but it is always restricted to testing non-overlapping of intervals. This is useful in testing if a loop is fully parallel (or could be vectorized) but cannot serve as testing the validity of arbitrary loop transformation (such as loop-interchange). Hence, [1] only tackles the may-aliasing problem while GCC is capable of performing auto-vectorization [34] using aliasing and interval overlap tests. Rus et al. [53], and later Oancea and Rauchwerger [44] use “uniform set representation” along with appropriate inference rules to reason about data dependence, and an associated predicate language for efficient runtime tests that validate loop parallelization.

Array Delinearization. Another topic that is related to our work is array delinearization, where a set of loop bounds and polynomial access functions is used to reconstruct array dimensions and sizes. As far as we know, the oldest attempt at delinearization dates back to the work of Maslov [42], but recent advances have made it applicable to linear access functions with parametric coefficients [24, 25]. Whereas delinearization consists in reconstructing affine array accesses to compute or apply a transformation, our work consists in testing the validity of a given transformation directly on the linearized accesses. However, apart the fact that this approach does not tackle pointer disambiguation, this pragmatic approach success rate is related to its “library” of access patterns. In particular array delinearization in [25] does not handle arbitrary polynomials (e.g., for triangular array accesses), and requires some assumptions on the possible sizes of rectangular arrays.

Handling Polynomials & Quantifier Elimination. Polynomials have long been recognized as an important class of functions for loop analysis. Claus [12] uses Bernstein basis decomposition to maximize a polynomial over a polyhedron. Unfortunately this technique is restricted to real numbers and its tentative use to design a quantifier elimination scheme leads to too imprecise results. Recently, Feautrier [19] has drawn the attention on two theorems on polynomials (see Section 5), and listed several potential applications. Our work can be seen as leveraging these theorems for the case of variable elimination.

The work that appears nearest to ours is Gröbner’s treatment of quantifier elimination [26, 27]. Constraint system splitting is used to structure the final projection as a tree, and no further simplification seems to be applied to the result. Another difference is that Fourier-Motzkin is applied only to polynomials that are linear in loop counters but have parametric coefficients, i.e., it could not be applied to our triangular array example. Several applications of the technique are mentioned, but as far as we know the resulting algorithm has never been applied to data dependence analysis. We use QEPCAD-b [9] to compare our results against cylindrical algebraic approaches, as used in that work. Whenever a solution could be provided (i.e. eliminate all variables leading to a $O(1)$ test) the result was not precise enough, prohibiting the use of the optimized code. More recently, on the work of Suriana [56], an extension to the FME was developed to handle parameter coefficients and used to produce loop boundaries. Limited to systems with a single symbolic coefficient, the author performs normalization as described on [59]. Without a technique to determine the sign of parametric coefficient the scheme always evaluate all possibilities, except when determined in the environment cache.

8 CONCLUSION AND FUTURE WORK
This paper developed a dynamic disambiguation technique to address the issue of dependence analysis and program optimization validity, suitable for many loop transformations such as skewing, tiling, vectorization, interchange, loop invariant code motion, parallelization, etc. We particularly focused on polynomial access functions, which are a challenge for current compiler analyses. Such functions arise typically with low-level IR (which are more and more common to address performance portability) but also in legacy code with pointers (e.g. C), linearized arrays, etc. To make the approach practical, the test to be evaluated at runtime to assess the validity of a transformed program must be as simple as possible. This objective was addressed by developing a powerful new quantifier elimination scheme on integer multivariate-polynomials, as prior works suffer from significant precision issues when dealing with quantifiers on $\mathbb{Z}$.

The quality of the presented quantifier elimination technique is very important to make this approach realistic. In particular, it must be precise enough such that the test succeeds in practical cases, and must lead to negligible overhead. The integer aspect makes the problem very challenging due to rounding errors that must be repeatedly removed. The precision and overhead has been evaluated on a set of 25 benchmarks using complex loop transformations.

Due to the combinatorial aspect of the elimination scheme, our current implementation is clearly not suited for Just-In-Time compilation. Fortunately, even if in theory the number of inequalities can square at every step of the Fourier-Motzin scheme, practice for affine systems many generated terms are redundant. It turns out to be the same for polynomial systems, and the use of the Schweighofer redundancy detection technique allows to make the approach really practical, requiring a few minutes to solve all the systems of our 20+ benchmarks. However, we believe many improvements are possible to decrease the analysis time. One consists in eliminating false systems (when cutting), using lifting to generate over-approximations such as intervals or polyhedra. A second one consists in caching Schweighofer’s redundancy results as it turns out identical systems appear many times in practice. Another one consists in (when available) using profile-based values. Such information could be used for three purposes: first, if profiled values state that a term turns out to be true in practice, as it is always safe to replace a term by true (which means transformation is not possible) it can be eliminated in any conjunctions. Second, interesting cuts mostly appear on actual dependences. Dependences observed by profiling (using shadow memory) can be used to add initial cuts. Lastly, “degraded” but faster techniques that use lifting or Bernstein expansion when it turns out to be accurate enough on the simple parts of the system. One could run degraded schemes first, check if the result is accurate enough with regard to profiled values, and only if it is not accurate, use a more aggressive (and costly) approach.