mlirSynth: Automatic, Retargetable Program Raising in Multi-Level IR using Program Synthesis

Citation for published version:

Digital Object Identifier (DOI):
10.1109/PACT58117.2023.00012

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
2023 32nd International Conference on Parallel Architectures and Compilation Techniques (PACT)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abstract—MLIR is an emerging compiler infrastructure for modern hardware, but existing programs cannot take advantage of MLIR’s high-performance compilation if they are described in lower-level general purpose languages. Consequently, to avoid programs needing to be rewritten manually, this has led to efforts to automatically raise lower-level to higher-level dialects in MLIR. However, current methods rely on manually-defined raising rules, which limit their applicability and make them challenging to maintain as MLIR dialects evolve.

We present mlirSynth—a novel approach which translates programs from lower-level MLIR dialects to high-level ones without manually defined rules. Instead, it uses available dialect definitions to construct a program space and searches it effectively using type constraints and equivalences. We demonstrate its effectiveness by raising C programs to two distinct high-level MLIR dialects, which enables us to use existing high-level dialect specific compilation flows. On Polybench, we show a greater coverage than previous approaches, resulting in geomean speedups of 2.5x (Intel) and 3.4x (AMD) over state-of-the-art compilation flows. mlirSynth also enables retargetability to domain-specific accelerators, resulting in a geomean speedup of 21.6x on a TPU.

I. INTRODUCTION

The end of Dennard scaling has led, in recent years, to the development of a diverse range of specialized hardware. Examples include tensor cores (TPU [34], NVIDIA [39]) and AI specialized accelerators [47]. Such hardware holds the promise of efficient performance, but at the cost of increased programming complexity.

A popular approach to overcoming this challenge is the use of domain-specific programming languages (DSLs), such as Halide [46], TensorFlow [2] and PyTorch [45]. These languages allow programmers to easily specify the essential structure of a problem without concern for low-level details. Crucially, this separation of concerns enables domain-specific compilers [54], [16] to efficiently map programs down to a wide range of idiosyncratic accelerators.

The need for existing code to harness the power of domain specific compilation has been well recognized by the compiler community with the development of MLIR [38]. MLIR is a new extensible representation within LLVM that captures high-level representations of programs. Once programs are expressed in the appropriate MLIR dialect, vendors can develop and exploit an efficient compilation path to their platform.

Need for Raising IRs.. This, however, presents a new challenge: how to translate existing code, currently represented in a low-level intermediate representation (IR), into a higher dialect so as to leverage domain-specific compilation. Furthermore, given the proliferation of MLIR dialects, any compiler technique that attempts to translate from a low to high dialect faces an additional challenge: how to adapt to a world of ever-changing high-level targets.

Previous Approaches.. This lifting of abstraction from a low to a high level is called program lifting [35] or raising [2], [56]. There is a large body of work in this area that has recently received increased interest [41], [9], [44], [8], [22], [18], [33]. In [35], lifting is applied to legacy FORTRAN code to generate high-performance Halide programs. This is achieved by representing both source and target languages in a common internal language and deploying an off-the-shelf program synthesis tool [51], [53], and proving the synthesized target code is equivalent to the original via loop invariants. This was later applied to C++ [6] and expanded into an LLVM-based framework [1]. While powerful, such an approach requires the user to manually define the semantics of all operations in the target language semantics in the common internal language. As the number of target high-level languages diversifies, this is not a scalable approach. There has been adjacent work in replacing code with library calls [23], [40]. Such approaches, however, are also fundamentally non-scalable as they focus on a fixed API rather than the open-ended nature of DSLs and their IRs.

Multi Level Tactics (MLT) [15] more recently, directly addressed this issue by showing raising to a high-level MLIR dialect enabled significant performance improvement. However, their approach requires domain-specific raising rules to be implemented by the compiler writer which are dialect specific. As we show in section VI, they are restricted in the number of programs they can tackle and rules need to be rewritten for each source and target dialect. Ideally, we would like a generic scheme that is robust, raising a large number of programs and is able to target new MLIR dialects without any compiler writer intervention.

Our Approach.. This paper presents mlirSynth, which automatically raises MLIR dialects from low to high-level without any hardwired compiler transformation or raising rules. Instead, mlirSynth automatically uses the available dialect definitions (within MLIR’s TableGen [38]) to construct a program space
II. Motivating Example

To illustrate the benefits of raising, consider the program in Figure 1 in the box labeled C Program. This loop nest implements the doitgen computation, and was taken from the Polybench benchmark suite. 

**LLVM IR.** As it is written in C, the loop nest is represented within the LLVM compiler by the standard SSA IR form shown in the box labeled LLVM IR. If we apply Polly [31], a polyhedral optimizing compiler to this IR, it is able to automatically generate parallel and cache efficient code. In this case, it is able to achieve a 1.9x speedup over the default -O3 pass on an Intel i7 platform as shown in the performance results plot.

**Affine.** Although Polly delivers a significant speedup, if the LLVM IR could be rewritten in an alternative MLIR dialect, then we can potentially achieve greater performance. This is the motivation behind the Affine dialect in MLIR, which captures high level polyhedral information, such as linear array access, convex iteration space and static control-flow. Polygeist [42] is a tool takes in C code and produces Affine IR for appropriate loop nests. They then apply the Pluto [14] polyhedral cache and parallelism optimizer to this Affine IR, which results in a 3.2x speedup. While the performance achieved is greater than Polly, the Affine dialect also acts as a convenient starting point for lifting to higher level dialects and is the source IR for the MLT compiler [15].

**Linalg.** Consider the Linalg IR version of the program in the box labeled Linalg in Figure 1. It is semantically equivalent to the Affine version, but is in a form that the MLIR compiler can generate more efficient code from. Rather than the Affine IR polyhedral representation of the program, Linalg describes

---

2CPU: Intel i7-8700k (6 cores / 12 threads), TPU: Google TPU v3 (8 cores).
the given code as three high level operators: a core matrix multiplication operation, surrounded by two tensor reshaping operations. As matrix multiplication and reshaping are often highly tuned on different hardware targets, it allows the MLIR compiler to exploit kernel libraries and generate a highly efficient implementation. The performance results plot shows that if we were able to lift code to this dialect, we were able to achieve an even greater speedup of 8.3x.

A. Raising IR

While MLT tries to automatically raise the Affine code to the LinAlg form, it, unfortunately, fails as its hand-coded matching rules do not consider this IR pattern. If additional patterns were added, then it would achieve a higher speedup.

HLO. Our approach is not limited to one target dialect of MLIR. It is able to raise to the higher level HLO dialect. Figure 1 further shows an implementation of the same computation in HLO (box labeled HLO). Rather than the three operations of LinAlg, the computation is expressed using a single high-level operation. This allows XLA [48], a compiler for HLO, greater flexibility in its implementation. If we consider the i7 platform, XLA able to achieve a speedup of over 100x relative to LLVM -03, justifying such a dialect. One benefit of this representation is that it can leverage platform specific compilers. If we change the hardware target to a TPU, the XLA compiler delivers a speedup of over 806x.

Summary. Higher-level representations in MLIR allow greater performance than lower-level ones and LLVM IR, as they allow compilation with high-performance domain-specific compilers. To enable such compilers for programs written in lower-level programming languages such as C, a raising technique is required. With the ongoing emergence of new dialects, we need a flexible raising technique to automatically leverage high-performance compilation flows.

III. SYSTEM OVERVIEW

In this section, we briefly introduce the notion of dialects within MLIR before describing the mlirSynth design which raises low-level dialects to higher ones.

A. MLIR

MLIR is an infrastructure for developing domain-specific compilers. To aid this, MLIR provides reusable building blocks and shared tools that allow us to define domain-specific languages and their compilation pipelines. The key concept that enables this is a dialect.

Dialects define sets of operations, types, and attributes. There are many dialects currently deployed (35 in the MLIR repository) and, crucially for automatic synthesis, each of these is defined by a structured TableGen description which contains the typed operands, attributes, and regions for each operation.

Figure 2 shows a small subset of MLIR’s dialects, relevant to this paper and the associated compilation flows. The lowest level IR we consider is LLVM IR, the default format for languages such as C/C++/FORTRAN. From this, the LLVM compiler generates code for all supported platforms. Many higher dialect compilers can progressively lower their dialect to LLVM IR.

Different compilation flows exploit high level dialect information to generate efficient implementations. For instance, Polygeist leverages the polyhedral representation available in Affine IR [42]. While some programming languages / DSLs such as TensorFlow can benefit from the XLA compiler due to its representation as HLO, this is not available to languages such as C. If we can raise code using MLT or mlirSynth to higher MLIR levels, then we can leverage the pre-existing compilation flows for performance. While a lowering path is always provided, raising is considerably more challenging.

B. mlirSynth design

mlirSynth operates in a three stage process that takes in a user program in the Affine IR dialect plus a description of the target dialect and outputs a raised program in the new dialect, as shown in Figure 3. The central idea is that we apply classic program synthesis techniques to lift a dialect to a higher one and then lower both the original and raised program down to the same representation and check they are equivalent. We use smallest-program first enumeration, discarding candidate programs based on type information and guided by heuristics where available. The approach is comprised of 3 stages:
**Pre-processing.** Initially, we apply pre-processing steps to simplify the program before enumerative synthesis is applied. Specifically, we use polyhedral analysis to distribute the original loop structure into smaller ones that can be synthesized independently and reduce the size of any array/matrix-like data structures in the code.

**Enumerative synthesis.** The key lifting is done by a classic enumerative synthesis technique. We generate a grammar (box Grammar Generator) for the synthesis process from the TableGen for the target dialect (Target Dialects), automatically generate an input-output specification (IO Spec Generator) for the target program, and then use a bottom-up synthesis process (Enumerator) to search the space of possible programs. This space is large, so we guide the enumerative search with a number of heuristics, based on polyhedral analysis and observational equivalence. Lifted candidates that satisfy the input-output specification (Checker) can be proved to be equivalent to the original code using bounded model checking (Validator).

**Post-processing.** Once we have a successful candidate, it is inlined back into the program and all data structure sizes are restored. It is now in the appropriate high level dialect and can be mapped to hardware by an appropriate compiler toolchain.

The following section gives details of the enumerative synthesis stage.

**IV. SYNTHESIS**

We base our synthesis procedure on classic enumerative synthesis algorithms from the literature [12]. The core synthesis loop, shown in Algorithm 1, is two phases: enumerating candidate programs and then checking the candidate against a specification. We thus need two inputs: a grammar that defines the space of possible programs to enumerate and a specification for the synthesized program. In the following section, we describe how we automatically generate grammars and specifications for our synthesis algorithm, the exact details of the enumeration process and the heuristics we use.

**A. Grammar generation**

In MLIR, dialects are defined in the declarative TableGen language. This language provides a structured way to define new dialects, including specifying the number and types of operands, attributes and regions that are required per operation. Given a target MLIR dialect, we use a custom TableGen extraction tool to generate a simple recursive context-free grammar for our enumerative synthesis process. Specifically, we generate an initial grammar $G$ that contains a single non-terminal for each type in the MLIR dialect, and a production rule for each operator.

This grammar does not precisely capture the syntactic restrictions of the MLIR dialect (since MLIR dialects are in general not context-free languages), but we are able to use inbuilt MLIR syntactic checks to discard any invalid programs later on in the enumerative loop. The grammar is, however, specific enough to rule out the majority of poorly typed programs. This automatic grammar generation simplifies retargeting our technique to new dialects.

**B. Specification generation**

Given a grammar $G$ and a reference function $f$ that we wish to lift, which takes input $x$, our goal is to synthesize a function $f'$, such that $\forall x. f(x) = f'(x)$ and $f'$ and $f'$ is in $L(G)$, the language defined by the grammar.

Checking a function satisfies the full equivalence specification above, for arbitrarily large input data structures, is in general undecidable and a significant challenge to state-of-the-art verification tools. Consequently, we apply a multi-staged synthesis process using two approximations of the specification. In the first specification, $\phi_{\delta-eq}$, we minify the input data structures, reducing $x$, a potentially very large data structure, to $x_{min}$, a small, bounded size data structure. We then assert that the relative error between the outputs of the functions is smaller than a small $\delta$. This $\delta$ accounts for the deviations introduced by compiler optimizations due to non-associativity of floating-point arithmetic. The second specification checks the observational equivalence of the functions, i.e., it checks for equivalent behavior on a finite set of $n$ inputs. We denote this $\phi_{obs_n}$ when it checks behavior on $n$ inputs and $n$ is small (i.e., less than 10) and $\phi_{obs_N}$ when it checks behavior on $N$ inputs, where $N$ is large (i.e., $\geq 10$). Formally, the specifications hold

![Fig. 3. Design of mlirSynth, showing its processing pipeline and component interaction.](image-url)
Algorithm 1 Core synthesis algorithm

```
function SYNTHESIZE(f, G)
    C ← initCandidates(f)
    I_n ← genRandomInputs(f, n)
    operations ← pickOperations(f, G)
    while true do
        f' ← enumerate(C, I_n, operations, f)
        I_N ← genRandomInputs(f, N)
        if specCheck(I_n, f, f') then
            return f'
        else
            I_n ← genRandomInputs(f, n)
    
for a given candidate f' under the following conditions:

\[ \phi_{\delta-eq} \Leftrightarrow \forall x_{min}, \text{obs}(f(x_{min}) - f'(x_{min}))/f(x_{min}) < \delta \]
\[ \phi_{\text{obs}_n} \Leftrightarrow \forall i \in I_n, f(i) = f'(i) \]
    where \( I_n \) is a small finite set of \( n \) inputs

\[ \phi_{\text{obs}_N} \Leftrightarrow \forall i \in I_N, f(i) = f'(i) \]
    where \( I_N \) is a large finite set of \( N \) inputs

We automatically generate \( \phi_{\text{obs}_n} \) and \( \phi_{\text{obs}_N} \) by randomly generating sets of inputs. We then check for observational equivalence by compiling and executing both \( f \) and \( f' \) on the inputs, shown in Algorithm 3. The data for \( \phi_{\text{obs}_n} \) is initially sampled from the range \([-10, 10]\) to work around the effect of numerical instabilities and to avoid cases where it is likely that the synthesized code will crash.

For speed, the core enumerative algorithm, Algorithm 2 checks candidates against \( \phi_{\text{obs}_n} \). Initially, \( I_n \) contains only a single input example, although this input is a high-dimensional tensor or matrix. The outer loop checks candidates against \( \phi_{\text{obs}_N} \). We use CBMC [36], a bounded model checker, to check \( \phi_{\delta-eq} \) post-synthesis.

Algorithm 2 Enumeration

```
function ENUMERATE(C, I_n, operations, f)
    while true do
        for op in operations do
            ops ← filterTypes(C, op)
            attr ← genAttrs(op)
            regs ← genRegions(op)
            for f' in cartesianProduct(ops, attr, regs) do
                if not staticCheck(f') then continue
                if observationallyUnique(C, f') then
                    C ← C ∪ f'
                if specCheck(I_n, f, f') then
                    return f'
        
C. Bottom-Up Enumeration

The core enumeration is a bottom-up synthesis algorithm inspired by [7]. The enumeration combines previous candidates with each other to generate more complex ones until a candidate matching the specification is found. An example of this is shown in Fig. 4.

Initialisation. We start by creating a candidate set \( C \) of valid (i.e., well-formed) candidates that each produce a computation-ally distinct value from the other candidates. We initialize this set with candidates that return the arguments of the reference function \( f \), as shown in the left-most Candidate Set box in Figure 4, as well as simple constants in the shape and data type of the arguments and results of the function. We will use this active candidate set as the base of our enumeration.

Enumeration. The synthesis loop enumerates through the set of operations in the grammar. For each operation, we first identify sets of possible operands, attributes and regions. We do this according to the operation signature in the grammar.

We populate the set of possible operands with all expressions in the candidate set of the correct type, highlighted yellow in each iteration in Fig 4. For operators with \( y \) operands of type \( \tau \), we add each active candidate of type \( \tau \) to the set of possible operands \( y \)-times to allow operators to have two or more identical operands.

For the attributes, which need to be known statically, we generate a large number of them, depending on their type, once the operator is selected. For regions, which contain groups of operations with arguments, we generate simple ones that perform binary mathematical operations on the function arguments (specifically, the operations addition, subtraction, multiplication and division). Region generation is not shown in Fig. 4, but these are generated in each iteration after the operator is selected.

We generate a set of all possible candidates by taking the Cartesian product of sets of operands, attributes and regions. Candidate Checking. Each candidate in the set is validated using a series of static checks, ordered by their complexity, such that the cheapest checks are performed first, and the expensive checks are performed last. We use MLIR’s type and shape inference system and built-in verification method chain to perform these checks.

Equivalence pruning and validating the candidate. If the static checks succeeded so far, we use MLIR’s execution engine to just-in-time compile the candidate. We then check \( \phi_{\text{obs}_n} \) by executing the candidate program \( f' \) on the set of inputs and comparing the output value with the output value produced by the reference function. If \( \phi_{\text{obs}_n} \) is satisfied, we send the candidate to the outer loop check.

We also check if the candidate is observationally unique on the inputs used by \( \text{specObs} \), that is, there does not exist a
Fig. 4. Synthesis example showing the enumeration of the candidate set at each iteration. In each iteration, an operator is selected, and the active set (highlighted in yellow) is chosen based on the types of the operands, and a set of attributes and regions is generated. We enumerate through the cartesian product until a correct candidate is found. If no correct candidate is found, any observationally unique candidates are added to the candidate set and we start a new iteration with a new operator.

candidate in the candidate set \( C \) that behaves the same as \( f' \) on all inputs. In other words, if \( f' \) is observationally unique, the following formula is valid \( \forall f \in C. \forall i \in I, f'(i) = f(i) \). If this is the case, we add \( f' \) to the candidate set \( C \). If not, we discard the candidate.

This central enumeration process is repeated for each operation, until either a program matching the specification \( \phi_{obs} \) is found, or a timeout expires. Once a candidate satisfies \( \phi_{obs} \), it is then checked against \( \phi_{obs,N} \). If a candidate fails this check, we can restart the synthesis loop with a new set of random inputs for \( \phi_{obs} \).

### D. Heuristics

Given the search space, heuristics are essential for selecting a set of operations to enumerate. We implement two heuristics: polyhedral model-based and dialect-based heuristics. These heuristics alter the behavior of the function `pickOperations`. Polyhedral model-based heuristics perform a value-based dependence analysis on the reference function to identify reduction dependencies, which are visible as cycles in the polyhedral dependence graph. If such reduction dependencies exist, this heuristic selects reduction operations in the target dialect. For a more detailed discussion of this see [24], [57].

Dialect-based heuristics look at the grammar for the target dialect and the source dialect, and, if an operator is present in the source function, written in the source dialect, it is prioritized in the target dialect. For example, if an add operation exists in the function, this heuristic selects any arithmetic operations that perform an addition in the target dialect. We evaluate the impact of heuristics in Section VI-E.

### E. Translation Validation

We use CBMC [36] to perform a post-synthesis check for the specification \( \phi_{eq} \). CBMC uses symbolic execution to generate a logical formula that is satisfiable if and only if the two functions are not equivalent (or, in path-based mode, multiple logical formulas representing different paths through the program). If the functions are not equivalent, CBMC will generate a counterexample, in the form of a set of inputs for which the outputs of the two functions are not identical. If this were to happen, we could repeat the synthesis process, using the counterexample input generated as one of the input examples in specification \( \phi_{obs} \). If CBMC exceeds a timeout of 1 hour, we substitute this check with extensive testing of \( \phi_{obs,N} \).

### V. Experimental Setup

This section briefly describes the platforms, benchmarks and various compilation flows used to compare against mlirSynth.

#### A. Platforms

All experiments were performed on two multi-core hardware platforms an Intel i7-8700K and an AMD Ryzen 9 3900X with multi-threading enabled. We also evaluated on 1 domain specific accelerator, the Google TPU v3.8.

#### B. Benchmarks

We evaluated all techniques on those benchmarks from Polybench that can be represented in either Linalg and/or HLO. These are shown in Table I. We used Polybench 4.2.1-beta in the large data size and float configuration.

<table>
<thead>
<tr>
<th>Category</th>
<th>Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>datamining</td>
<td>correlation</td>
</tr>
<tr>
<td>kernels</td>
<td>datamining covariance</td>
</tr>
<tr>
<td>datamining</td>
<td>atax</td>
</tr>
<tr>
<td>datamining</td>
<td>2mm</td>
</tr>
<tr>
<td>datamining</td>
<td>3mm</td>
</tr>
<tr>
<td>datamining</td>
<td>doittgen</td>
</tr>
<tr>
<td>datamining</td>
<td>mvt</td>
</tr>
<tr>
<td>datamining</td>
<td>bicg</td>
</tr>
<tr>
<td>datamining</td>
<td>blas</td>
</tr>
<tr>
<td>datamining</td>
<td>syrk</td>
</tr>
<tr>
<td>datamining</td>
<td>syr2k</td>
</tr>
<tr>
<td>datamining</td>
<td>trmm</td>
</tr>
<tr>
<td>datamining</td>
<td>genumer</td>
</tr>
<tr>
<td>datamining</td>
<td>symm</td>
</tr>
<tr>
<td>datamining</td>
<td>gesummv</td>
</tr>
<tr>
<td>datamining</td>
<td>gemm</td>
</tr>
</tbody>
</table>
mlirSynth-Linalg: Our approach, raises Affine to Linalg IR (on tensors) and invokes latest untuned MLIR Linalg compiler
mlirSynth-XLA: Our approach, raises Affine to HLO IR and then invokes the XLA compiler targeting CPUs
mlirSynth-XLA (TPU): As mlirSynth-XLA, targeting TPU

D. Methodology

All experiments were run 10 times with median end-to-end execution time reported. To ensure there was no caching, we ensured a cold start for each experiment, spawning a new process for each run. To further evaluate our raising ability we compare against MLT [15] and KernelFaRer [23], which are state-of-the-art methods for raising to high-level operations from lower-level code.

mlirSynth-Linalg uses the latest default MLIR lowering compiler without any tuning or replacement of named linalg operations with BLAS calls. MLT, however, uses a legacy version of MLIR tuned for performance (e.g. optimized tile sizes). To provide a fair comparison, we retain MLT’s use of this tuned legacy compiler. mlirSynth-Linalg currently targets the exact same subset of Linalg as MLT to allow side-by-side comparison. This, however, restricts the number of programs that can be raised.

While mlirSynth could raise programs to combinations of target dialects, we are limiting the experiments to individual ones to allow a more tractable search. We plan to explore combinations of target dialects in future work.

VI. Evaluation

This section first summarises the performance achieved by each approach before examining coverage. This is followed by a detailed performance comparison and an analysis of mlirSynths compilation time. It concludes with an validity evaluation.

A. Overall Summary

Figure 5 shows the average speedups of the various compilation flows on three platforms across the Polybench benchmarks. Speedups are relative to LLVM-O3. MLT lifting to Linalg achieves a 1.1x speedup on the AMD platform, and 1.2x on Intel. Replacing named operations with BLAS routines however achieves 3.3x improvement on the Intel platform, rising to 4.2x on AMD. Although significant, MLT-BLAS’s performance is limited by the number of kernels it can raise. In fact, Polly is able to achieve greater improvement: 6.4x and 6.1x speedups. mlirSynth-Linalg is able to synthesize a larger number of kernels than MLT, which gives a performance improvement across both CPU platforms. However, MLT-BLAS uses a tuned MLIR legacy compiler and substituted BLAS routines gives increased performance. When lifting to HLO and invoking the XLA compiler, mlirSynth-XLA achieves a geometric mean speedup of 20.8x on the Intel platform, rising to a geometric mean of 20.9x on the AMD. This significant increase is because XLA makes use of vendor-optimized kernel

C. Compilation flows

We evaluated a number of different compilation flows:

LLVM-O3: General-purpose compiler, used as baseline [37]
Polly: Polyhedral compiler, optimizes LLVM IR for caches and parallelism [30]
Polygeist: Polyhedral compiler, takes C to Affine IR, then uses Pluto for cache- and parallelism optimization [42]
MLT: Raises Affine to Linalg IR before invoking a tuned MLIR Linalg compilation flow [15]
MLT-BLAS: As MLT, replacing named Linalg operations with BLAS calls after raising [15]
mlirSynth-Linalg: Our approach, raises Affine to Linalg IR (on tensors) and invokes latest untuned MLIR Linalg compiler
mlirSynth-XLA: Our approach, raises Affine to HLO IR and then invokes the XLA compiler targeting CPUs

This section first summarises the performance achieved by each approach before examining coverage. This is followed by a detailed performance comparison and an analysis of mlirSynths compilation time. It concludes with an validity evaluation.

A. Overall Summary

Figure 5 shows the average speedups of the various compilation flows on three platforms across the Polybench benchmarks. Speedups are relative to LLVM-O3. MLT lifting to Linalg achieves a 1.1x speedup on the AMD platform, and 1.2x on Intel. Replacing named operations with BLAS routines however achieves 3.3x improvement on the Intel platform, rising to 4.2x on AMD. Although significant, MLT-BLAS’s performance is limited by the number of kernels it can raise. In fact, Polly is able to achieve greater improvement: 6.4x and 6.1x speedups. mlirSynth-Linalg is able to synthesize a larger number of kernels than MLT, which gives a performance improvement across both CPU platforms. However, MLT-BLAS uses a tuned MLIR legacy compiler and substituted BLAS routines gives increased performance. When lifting to HLO and invoking the XLA compiler, mlirSynth-XLA achieves a geometric mean speedup of 20.8x on the Intel platform, rising to a geometric mean of 20.9x on the AMD. This significant increase is because XLA makes use of vendor-optimized kernel

D. Methodology

All experiments were run 10 times with median end-to-end execution time reported. To ensure there was no caching, we ensured a cold start for each experiment, spawning a new process for each run. To further evaluate our raising ability we compare against MLT [15] and KernelFaRer [23], which are state-of-the-art methods for raising to high-level operations from lower-level code.

mlirSynth-Linalg uses the latest default MLIR lowering compiler without any tuning or replacement of named linalg operations with BLAS calls. MLT, however, uses a legacy version of MLIR tuned for performance (e.g. optimized tile sizes). To provide a fair comparison, we retain MLT’s use of this tuned legacy compiler. mlirSynth-Linalg currently targets the exact same subset of Linalg as MLT to allow side-by-side comparison. This, however, restricts the number of programs that can be raised.

While mlirSynth could raise programs to combinations of target dialects, we are limiting the experiments to individual ones to allow a more tractable search. We plan to explore combinations of target dialects in future work.
libraries. When targeting the TPU, mlirSynth is able to achieve over 175x improvement, a significant result.

B. Coverage

The performance achieved by raising critically depends on the number of raised programs. Figure 6 shows the percentage of programs raised by different approaches for each of the Polybench categories shown in Table I. KernelFarer [23] is a robust GEMM detector and is able to detect two routines in the BLAS and kernel category, but no others due to its hardwired pattern matching rules. MLT performs better capturing most of the kernels (5 out of 6) and some of the BLAS (3 out of 7) categories. It is unable to capture any of the data mining kernels due to its restricted Linalg coverage. When raising to HLO, mlirSynth is able to raise all candidates except for `trmm`. Its computation cannot be represented in HLO operations and therefore, results in raising to fail.

C. Detailed CPU performance results

Figure 7 shows a more detailed performance evaluation of different compilation paths relative to LLVM-O3. We show Polly and Polygeist as the polyhedral compilers, MLT, MLT-BLAS, finally mlirSynth raising Linalg and XLA operations.

MLT-BLAS is able to improve on Polly on the Intel platform in 3 cases where large matrix multiplications dominate execution time, `2mm`, `3mm`, `gemm`. On the AMD platform, in addition, it performs well on `atax`. Overall however it performs less well than Polly as it is unable to raise all the benchmarks to Linalg.

As mlirSynth-Linalg is able to synthesize a larger number of kernels than MLT it achieves performance improvement in seven of the benchmarks. In 2 cases, `bicg` and `gemver` MLT performs better due to its tuned legacy MLIR compiler.

Comparing mlirSynth-XLA to MLT-BLAS, we see MLT-BLAS is able to achieve comparable performance to mlirSynth-XLA on matrix-matrix multiply like kernels (`2mm`, `3mm`, `gemm`). This is because XLA uses similar BLAS kernel libraries like MLT, particularly on the Intel platform. On other programs such as `mvt`, mlirSynth-XLA is superior as it synthesizes a more efficient, but computationally equivalent program.

It is clear that raising beyond Linalg to HLO enables significant performance improvement due to the superior XLA
compilation flow. On both the CPU platforms, in all cases, it outperforms the corresponding Linalg implementations. While it is less performant than polyhedral compilers for small problems, on average it significantly outperforms them.

D. TPU results

If we consider the domain-specific TPU accelerator, then HLO enables even greater performance improvement across the benchmarks, as shown in Figure 8. The XLA compiler is able to achieve over 3000x speedups in some cases, performing particularly well on programs containing large matrix operations.

E. Synthesis time

While mlirSynth is able to raise programs to higher-level dialects, it requires search-based synthesis which could be expensive. Figure 9 shows the synthesis time for each benchmark for each dialect and compares it against a naive synthesizer that does not restrict the types and the set of operations.

In 7 out of 13 cases, we are able to raise programs in less than 2 seconds. In the 4 more complicated examples, synthesis time increases to over 2 minutes. The impact of type information and candidate pruning is significant. Compared to the naive algorithm, we are able to reduce synthesis time by an order of magnitude.

Table II provides a more detailed breakdown of the synthesis process for HLO. The synthesis time is correlated to the largest synthesis subproblem size, whereas the largest solved one was 5 high-level operations. While the number of candidates considered varies from 720 to c. 20 million, over 90% of these can be discarded based on type and shape filtering. The remaining candidates are then evaluated, with those that are found to be equivalent to existing candidates eliminated from further consideration. The number discarded this way varies and increases in impact as the number of enumerated candidates increases: from 0% for 3mm up to 60% for symm.

F. Validity

As described in section IV-E, we employ model checking [36] to determine if our raised program is equivalent to the original [17]. In our post-synthesis checks, 4 solutions are proven to be equivalent or δ-equivalent (satisfying φ_{int-δ}). CBMC failed to find any errors within 1 hour when checking φ_{int-δ} for 8 of the remaining HLO solutions, but we were able to prove that these are equivalent when using integers and discounting floating-point arithmetic. That is, the specification φ_{int-δ} was shown to hold, where φ_{int-δ} ⇔ ∀x ∈ X_{int}.f_{int}(x) = f_{int}'(x) and X_{int} is the set of all possible integers and f_{int} and f_{int}' use integer arithmetic throughout. We are unable to verify the remaining two solutions (correlation and

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Enumerated</th>
<th>Static filtered</th>
<th>Evaluated</th>
<th>Equiv filtered</th>
<th>Ops (max)</th>
<th>Synth time</th>
<th>Formal Guarantee</th>
</tr>
</thead>
<tbody>
<tr>
<td>2mm</td>
<td>49067</td>
<td>46504</td>
<td>1043</td>
<td>709</td>
<td>7 (3)</td>
<td>0.65s</td>
<td>φ_{int-eq}</td>
</tr>
<tr>
<td>3mm</td>
<td>2484</td>
<td>2409</td>
<td>3</td>
<td>0</td>
<td>3 (1)</td>
<td>0.14s</td>
<td>φ_{int-eq}</td>
</tr>
<tr>
<td>atax</td>
<td>18960</td>
<td>17042</td>
<td>1166</td>
<td>763</td>
<td>8 (3)</td>
<td>0.62s</td>
<td>φ_{δ-eq}</td>
</tr>
<tr>
<td>bicg</td>
<td>18961</td>
<td>17046</td>
<td>1173</td>
<td>771</td>
<td>8 (3)</td>
<td>0.59s</td>
<td>φ_{δ-eq}</td>
</tr>
<tr>
<td>correlation</td>
<td>1420241</td>
<td>1173035</td>
<td>188577</td>
<td>159679</td>
<td>22 (3)</td>
<td>174.11s</td>
<td>φ_{obs_N}</td>
</tr>
<tr>
<td>covariance</td>
<td>382674</td>
<td>374083</td>
<td>5799</td>
<td>2049</td>
<td>6 (3)</td>
<td>4.21s</td>
<td>φ_{δ-eq}</td>
</tr>
<tr>
<td>doigen</td>
<td>9972</td>
<td>9879</td>
<td>71</td>
<td>18</td>
<td>1 (1)</td>
<td>0.16s</td>
<td>φ_{int-eq}</td>
</tr>
<tr>
<td>gemm</td>
<td>607638</td>
<td>572798</td>
<td>13695</td>
<td>6745</td>
<td>4 (3)</td>
<td>7.26s</td>
<td>φ_{δ-eq}</td>
</tr>
<tr>
<td>gesummv</td>
<td>29221</td>
<td>24566</td>
<td>3919</td>
<td>2333</td>
<td>9 (3)</td>
<td>1.37s</td>
<td>φ_{δ-eq}</td>
</tr>
<tr>
<td>mvt</td>
<td>27977</td>
<td>24460</td>
<td>2855</td>
<td>1631</td>
<td>4 (2)</td>
<td>1.09s</td>
<td>φ_{obs_N}</td>
</tr>
<tr>
<td>symm</td>
<td>5353361</td>
<td>4943595</td>
<td>309752</td>
<td>163310</td>
<td>4 (4)</td>
<td>134.85s</td>
<td>φ_{int-eq}</td>
</tr>
<tr>
<td>syrk</td>
<td>20820281</td>
<td>18547932</td>
<td>1467901</td>
<td>1022725</td>
<td>12 (5)</td>
<td>2438.69s</td>
<td>φ_{int-eq}</td>
</tr>
<tr>
<td>syrk</td>
<td>3532229</td>
<td>2954620</td>
<td>433798</td>
<td>297594</td>
<td>7 (5)</td>
<td>467.79s</td>
<td>φ_{int-eq}</td>
</tr>
</tbody>
</table>
API replacement. Replacing matched code/IR to a fixed API call [19] is a limited form of raising. KernelFarer [23] works at the program level and restricts its attention to just GEMM API targets, but is more robust than IDL matching significantly more user code. This robustness is extended further in [55], [40] which uses behavioral equivalence to match code. Such approaches, however, are intrinsically limited as they focus on fixed APIs rather than the open-ended nature of DSLs and their IRs.

Raising with synthesis. Using program synthesis to generate programs from a specification is a long-studied area [25], [50]. Using a low-level program as the specification and a high level-one as the target was tackled in [35]. Here appropriate stencil like loops in FORTRAN are lifted to their equivalent in Halide [46]. This has been extended to a more generic LLVM framework [3] based on a common IR. While this has the potential to allow lifting to multiple targets [5], [4], it requires the compiler writer to provide a compiler and decompiler from each potential source and target into the IR which is not scalable. MLIR-Fuzz [27] offers a fuzzer mlir enumerate which enumerates type-correct MLIR programs bottom-up for any dialect by translating MLIR's TableGen files into the dialect definition language IRDL [26].

Example driven synthesis. The use of input/output examples to synthesize high-level code has been explored in a number of projects [32], [58], [21], [20]. It has been used to generate pytorch or tensor-flow code from tensor inputs [49], [43]. TF-coder [49] uses type- constraints and equivalences to efficiently apply enumerative program search while [43] uses a DeepCoder [11] style predictive model to guide code generation. AutoPandas [13] uses a more powerful graph neural network based model to guide the generation of Panda code. As there is no ground-truth program to lift, just examples, such schemes cannot be directly used for IR raising. Furthermore, both source format and target output are hardwired for each domain.

Polyhedral compilation. The use of polyhedral analysis to drive program optimization [14] has been extensively explored in the compilation community [31]. It has been used for driving systolic code generation [29], memory hierarchy optimization, parallelization and GPU code generation [10] and forms the core for many modern tensor algebra compilers. Polly [30] is able to generate efficient cache optimized and parallel code directly from LLVM IR.

Low-level loop and memory reference representation in LLVM IR can make analysis difficult. This has motivated LLVM IR extensions to facilitate parallelization [52] and motivated Polygeist, a C to Affine IR compiler [42] that uses Pluto [14] to generate cache optimized and parallel code. All of these approaches use polyhedral analysis to lower code, rather than mlirSynth which uses it to raise dialect levels.
REFERENCES


