Automated Programmable Control and Parameterization of Compiler Optimizations

Qing Yi (qingyi@cs.utsa.edu)
University of Texas at San Antonio

Abstract—We present a framework which effectively combines programmable control by developers, advanced optimization by compilers, and flexible parameterization of optimizations to achieve portable high performance. We have extended ROSE, a C/C++/Fortran source-to-source optimizing compiler, to automatically analyze scientific applications and discover optimization opportunities. Instead of directly generating optimized code, our optimizer produces parameterized scripts in POET, an interpreted program transformation language, so that developers can freely modify the optimization decisions by the compiler and add their own domain-specific optimizations if necessary. The auto-generated POET scripts support extra optimizations beyond those available in the ROSE optimizer. Additionally, all the optimizations are parameterized at an extremely fine granularity, so the scripts can be ported together with their input code and automatically tuned for different architectures. Our results show that this approach is highly effective, and the code optimized by the auto-generated POET scripts can significantly outperform those optimized using the ROSE optimizer alone.

I. INTRODUCTION

Compiler optimization is critical for scientific applications to automatically achieve high performance on modern computers. However, due to the excessive complexity, compilers have lagged behind in precisely modeling the behaviors of applications running on diverse architectures. While computational specialists can adopt low-level programming in C or assembly to directly manage machine resources and can parameterize their algorithm implementations to accommodate architectural diversity [21], [2], [20], [7], [15], this approach is extremely error prone and time consuming, and compiler technology needs to be developed to automate the process.

We present a framework, shown in Fig. 1, to reach a balance between direct control of resources by programmers and automation of optimization by compilers. This framework starts with a specialized source-to-source optimizing compiler (the ROSE analysis engine) which interacts with developers to automatically discover applicable optimizations and then produces output in a transformation scripting language, POET [25]. Developers can modify the auto-generated POET scripts as well as writing new POET transformations to directly control the optimization of applications. The POET output can then be ported together with an annotated input program to different machines, where an empirical transformation engine can dynamically interpret the POET scripts with different optimization configurations until satisfactory performance is achieved for the input code.

This research is funded by NSF through award CCF0747357 and CCF-0833203, and DOE through award DE-SC0001770

Fig. 1. The optimization workflow

The merit of our overall approach lies in its unique integration of programmable control by developers, automated optimization by compilers, and empirical tuning of the optimization space by search engines. It permits different levels of automation and programmer intervention, from fully-automated tuning to semi-automated optimization to fully programmable control. The POET language was initially designed by Yi et al. [25] and has been used to achieve extremely high performance for performance critical kernels [26]. This paper focuses on extending a source-to-source compiler, the ROSE loop optimizer [23], to automatically generate high quality POET scripts, so that developers do not need to manually manage hardware resources using POET (as done by [26]).

Our extended ROSE optimizer can produce POET scripts to support 7 optimizations: OpenMP parallelization, loop blocking, array copying, loop unroll & jam, scalar replacement, strength reduction, and loop unrolling. The auto-generated POET transformations are extensively parameterized so that each optimization can be turned on or off independently for each relevant array or code region, and arbitrary integers can be given as the blocking or unrolling factor for each loop being transformed. The granularity of external control is far beyond those supported by existing iterative compilation frameworks [9], [16], [1], [14], [13]. Independent search engines can be substituted with ease, and developers can easily interfere by modifying the auto-generated POET scripts.
III. THE OVERALL APPROACH

The key challenge of our approach is using POET to explicitly manage interactions between different transformations, as illustrated by Fig. 2 using two transformations, loop blocking and unroll-&-jam. After applying blocking, the three nested loops (nest1, nest3, and nest2) in Fig. 2(a) become 6 nested loops in (b). Subsequent transformations, e.g., array copying, may be applied to the outer 3 loops in (b). However, some other transformations, e.g., unroll-&-jam and scalar replacement, need to be applied to the inner 3 loops, as demonstrated by Fig. 2(c). The matter becomes increasingly more complex as scalar replacement may be applied after array copying to the same array, and strength reduction may be applied to simplify array references. Conventional compilers overcome such difficulties by separating optimizations into multiple phases, where each phase re-analyzes the output of an earlier phase before performing additional transformations. In contrast, our auto-generated POET scripts do not perform any program analysis in between the extensively parameterized transformations. Specifically, the auto-generated POET scripts aim to explicitly model the interactions between parameterized transformations and dynamically trace these interactions so that different transformations can be applied one after another in a well-coordinated fashion.

This paper presents techniques to extensively parameterize and dynamically coordinate a number of complex optimizations in the absence of any sophisticated program analysis in between the transformations. As a result, an optimizing compiler needs to analyze an input code only once, and multi-phase optimizations can be automatically coordinated and collectively parameterized to support programmable control and flexible empirical tuning. Our contributions include:

- We present a framework where optimization analyses and program transformations are separated into independent components (i.e., the ROSE optimizer and the POET engine), so that compiler optimizations can be managed by developers with programmable control, and fine-grained parameterization can be explored by independent search engines for portable performance tuning.
- We show that by carefully modeling the interactions of different transformations, a long sequence of dynamically reconfigurable optimizations can be effectively applied one after another in a well-coordinated fashion without resorting to any intermediate program analysis.
- We study several linear algebra routines and show that the new approach can produce significantly better optimized code than using the ROSE optimizer alone.

The rest of this paper is organized as follows. Section II introduces our overall framework. Section III presents algorithms for automatically generating parameterized POET scripts. Section IV presents experimental results. Sections V and VI present related work and conclusions.

II. THE OVERALL APPROACH

The key to our approach is adapting an existing optimizing compiler to alternatively generate POET scripts which dynamically coordinate a long sequence of parameterized program transformations without requiring any intermediate program analysis. Fig. 3 shows the POET script auto-generated by our ROSE optimizer after analyzing the code in Fig. 2(a), which includes some POET annotations automatically inserted by the optimizer to tag the code regions being optimized. Tab. I and Fig. 5 complement Fig. 3 with additional details.

A. The POET Language

POET is a scripting language [25] designed for applying parameterized source-to-source program transformations to code in arbitrary programming languages such as C, C++,

```c
void dgemm_test(const int M, const int N, const int K, const double alpha, const double *A, const int lda, const double *B, const int ldb, const double beta, double *C, const int ldc) {
    int i, j, l, i_bk, j_bk, l_bk;
    for (l_bk=0; l_bk<K; l_bk+=16)
        for (i_bk=0; i_bk<M; i_bk+=16)
            for (j_bk=0; j_bk<N; j_bk+=16)
                if (0==l)
                    for (i=0; i<MIN(M-i_bk,16); i+=1)
                        for (j=0; j<MIN(N-j_bk,16); j+=1)
                            for (l=0; l<MIN(K-l_bk,16); l+=1)
                                if (1+j<MIN(N-j_bk,16))
                                    for (i=0; i<MIN(M-i_bk,16); i+=1)
                                        for (j=0; j<MIN(N-j_bk,16); j+=1)
                                            for (l=0; l<MIN(K-l_bk,16); l+=1)
                                                for (i_bk=0; i_bk<M; i_bk+=16)
                                                    for (j_bk=0; j_bk<N; j_bk+=16)
                                                        for (l_bk=0; l_bk<K; l_bk+=16)
                                                            C[j * ldc + i] = beta * C[j * ldc + i] + alpha * A[i * lda + i_bk+j_bk*ldc+i];
}
```

(b) After applying loop blocking by factors of (16, 16, 16)

```c
void dgemm_test(const int M, const int N, const int K, const double alpha, const double *A, const int lda, const double *B, const int ldb, const double beta, double *C, const int ldc) {
    int i, j, l, i_bk, j_bk, l_bk;
    for (l_bk=0; l_bk<K; l_bk+=16)
        for (i_bk=0; i_bk<M; i_bk+=16)
            for (j_bk=0; j_bk<N; j_bk+=16)
                if (0==l)
                    for (i=0; i<MIN(M-i_bk,16); i+=1)
                        for (j=0; j<MIN(N-j_bk,16); j+=1)
                            for (l=0; l<MIN(K-l_bk,16); l+=1)
                                if (1+j<MIN(N-j_bk,16))
                                    for (i=0; i<MIN(M-i_bk,16); i+=1)
                                        for (j=0; j<MIN(N-j_bk,16); j+=1)
                                            for (l=0; l<MIN(K-l_bk,16); l+=1)
                                                for (i_bk=0; i_bk<M; i_bk+=16)
                                                    for (j_bk=0; j_bk<N; j_bk+=16)
                                                        for (l_bk=0; l_bk<K; l_bk+=16)
                                                            C[i+i_bk+j_bk*ldc+j*ldc] = beta*C[i+i_bk+j_bk*ldc+j*ldc] + alpha*A[i+i_bk+j_bk*ldc+i];
}
```

(c) After applying loop unroll-&-jam by factors of (2, 1)

Fig. 2. Transforming a matrix multiply routine using POET transformations without requiring any intermediate program analysis. Fig. 3 shows the POET script auto-generated by our ROSE optimizer after analyzing the code in Fig. 2(a), which includes some POET annotations automatically inserted by the optimizer to tag the code regions being optimized. Tab. I and Fig. 5 complement Fig. 3 with additional details.
Global declarations and commands

- `<input syntax="from=test" test=""> Parse file `f` using syntax descriptions defined in file `s`, then evaluate result in `test`.
- `<output syntax="from=usef /> Parse the result of `f` using syntax descriptions defined in file `s`.
- `<trace a,...,m/> Declare a list of related trace handles a1,...,am.
- `<parameter p type=default value=v>` Declares a command-line parameter `p` which has type `t`, default value `v`, and its meaning is defined in string `d`.
- `<eval s1,...,sm/> Evaluate statements s1,...,sm.

**POET expressions: constructing and operating on different types of values**

\[
\begin{align*}
& r_1, r_2, \ldots, r_n \quad \text{A list of n elements c_1, c_2, \ldots, c_n} \\
& \text{HEAD(x)} \quad \text{The first element of a list in variable x} \\
& \text{TAIL()} \quad \text{The rest of the list after the first element} \\
& \text{append(c,d)} \quad \text{Value in data field d of object c} \\
& \text{f[x]} \quad \text{Invoke a routine f using parameters x1,...,xm, with p1,...,pm as default values for optional parameters p0} \\
& \text{REPLACE(c1,c2,e)} \quad \text{Replace all occurrences of c1 with c2 in AST e} \\
& \text{REPLACE(\{(c1),c2\},(\{(c3),c4\},c5))} \quad \text{Traverse AST in pre-order to locate and replace each c1 by c2 in AST e} \\
\end{align*}
\]

**Selected routines from the POET opt library**

- `<ParallelizeLoop[...]>` Parallelize loop nest `x` via OpenMP using `d` threads; treat variables in `c` as private; modify `b` to include OMP header file; use trace handle `a` to save modifications to `o`.
- `<blockLoops[...]>` Block loops that are outside `n` and inside `x` using factors of `c`; the loops are non-perfectly nested at location `b`; use trace handle `d` to save the inner tided loop nest; trace modifications to expression `e`.
- `<TreeTraversal[...]>` Traverse AST in order to perform actions on `o`.

**TABLE I**

<table>
<thead>
<tr>
<th>Routines from the POET opt library</th>
<th>Description</th>
</tr>
</thead>
</table>
| `<ParallelizeLoop[...]>`           | Parallelize loop nest `x` via OpenMP using `d` threads; treat variables in `c` as private; modify `b` to include OMP header file; use trace handle `a` to save modifications to `o`.
| `<blockLoops[...]>`                | Block loops that are outside `n` and inside `x` using factors of `c`; the loops are non-perfectly nested at location `b`; use trace handle `d` to save the inner tided loop nest; trace modifications to expression `e`.
| `<TreeTraversal[...]>`             | Traverse AST in order to perform actions on `o`.

**Fig. 3** Auto-generated POET scripts for Fig. 2(a)

1. Include `opt.pi`
2. `<trace target/>`
3. `<input target="syntaxциальненексимдекфронтоукоде" from="(dgemm.c)"/>`

**FORTRAN.** For example, line 3 of Fig. 3 parses the matrix multiplication code in file `Cfront.code` and then stores the resulting AST to a global variable named `target`. The output command at line 47 serves to unpars the optimized AST to standard output. The inclusion of file `opt.pi` at line 1 ensures that the POET `opt` library, which supports a large collection of compiler transformations, can be invoked by the given script. POET provides strong programming support for flexibly combining a long sequence of heavily parameterized program transformations. It uses a special collection of global variables called `trace handles` which can be embedded inside an input code to automatically keep track of various code fragments as they go through different transformations. In particular, these trace handles can be embedded inside parameters of a POET routine invocation so that the routine can dynamically modify these handles without knowing their names. For example, lines 4-13 of Fig. 3 declare and initialize all the trace handles that will be used to accommodate interactions between different POET transformations, discussed in Section III-C. In Fig. 3, the 11 optimizations that will be later applied to transform the input code are defined at lines 25-35 using the `DELAY` operator. All transformations invoke routines from the `opt` library to operate on the trace handles declared at lines 4-13, and their configurations are extensively controlled by the command-line parameters declared at lines 14-24.

**TABLE I**

<table>
<thead>
<tr>
<th>POET expressions: constructing and operating on different types of values</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>r_1, r_2, \ldots, r_n</code> A list of n elements <code>c_1, c_2, \ldots, c_n</code></td>
</tr>
<tr>
<td><code>\text{HEAD(x)}</code> The first element of a list in variable <code>x</code></td>
</tr>
<tr>
<td><code>\text{TAIL()}</code> The rest of the list after the first element</td>
</tr>
<tr>
<td><code>\text{append(c,d)}</code> Value in data field <code>d</code> of object <code>c</code></td>
</tr>
<tr>
<td><code>f[x]</code> Invoke a routine <code>f</code> using parameters <code>x_1, \ldots, x_m</code>, with <code>p_1, \ldots, p_m</code> as default values for optional parameters <code>p_0</code></td>
</tr>
<tr>
<td><code>\text{REPLACE(c1,c2,e)}</code> Replace all occurrences of <code>c_1</code> with <code>c_2</code> in AST <code>e</code></td>
</tr>
<tr>
<td><code>\text{REPLACE(\{(c1),c2\},(\{(c3),c4\},c5))}</code> Traverse AST in pre-order to locate and replace each <code>c_1</code> by <code>c_2</code> in AST <code>e</code></td>
</tr>
</tbody>
</table>

**Selected routines from the POET opt library**

- `ParallelizeLoop[...]>` Parallelize loop nest `x` via OpenMP using `d` threads; treat variables in `c` as private; modify `b` to include OMP header file; use trace handle `a` to save modifications to `o`.
- `blockLoops[...]>` Block loops that are outside `n` and inside `x` using factors of `c`; the loops are non-perfectly nested at location `b`; use trace handle `d` to save the inner tided loop nest; trace modifications to expression `e`.
- `TreeTraversal[...]>` Traverse AST in order to perform actions on `o`.

**TABLE I**

<table>
<thead>
<tr>
<th>Selected POET operations used in Figs. 3 and 5</th>
</tr>
</thead>
</table>
| `\text{ParallelizeLoop[...]>}` Parallelize loop nest `x` via OpenMP using `d` threads; treat variables in `c` as private; modify `b` to include OMP header file; use trace handle `a` to save modifications to `o`.
| `blockLoops[...]>` Block loops that are outside `n` and inside `x` using factors of `c`; the loops are non-perfectly nested at location `b`; use trace handle `d` to save the inner tided loop nest; trace modifications to expression `e`.
| `TreeTraversal[...]>` Traverse AST in order to perform actions on `o`.

**Fig. 3** Auto-generated POET scripts for Fig. 2(a)
Approach entails the following technical challenges. In an explicit fashion than previously possible. This non-conventional developers and empirical search engines in a much more ex-

integrating POET as an alternative output language of existing regions to be transformed later, an example of which is shown modification to input code by ROSE is the tagging of code advanced analysis to make optimization decisions. The only gram transformations to POET, where ROSE merely performs The adapted optimizer has essentially delegated the actual pro-

binary focus on managing a different set of architect-

Conventional compil-

Interactions between optimizations. Since AST transfor-

mations within ROSE have been redirected to a new POET scripting interface, each optimization now assumes that it directly operates on the original input code. However, since program transformations are actually applied one after another by the auto-generated POET script, the input to each transformation could be dramatically different from the original input code. This discrepancy could result in incorrect program transformations.

Phase Ordering of Optimizations. Conventional compi-

eres apply optimizations in multiple phases, with each phase focusing on managing a different set of architectural features. However, since the actual program transformations are now postponed and delegated to the auto-generated POET script, our ROSE optimizer can no longer separate these optimizations into different phases.

To resolve these issues, a key observation is that since no compiler optimization is allowed to alter the semantics of its input program, each optimization safe for the original input typically remains safe even after other optimizations have been applied. However, since the input code may have been modified by other optimizations, the auto-generated POET script must precisely model such interactions to ensure correctness of transformation. Further, optimizations at different phases often share the same program analysis. For example, if a loop nest should be blocked for cache reuse, it is safe and typically profitable to apply unroll&jam to the blocked loops to further promote register reuse. We therefore can use a single optimization analysis to drive related optimizations at different phases. Sections III discusses details of these solutions.

B. The ROSE Loop Optimizer

We have built our analysis engine in Fig. 1 by adapting the ROSE loop optimizer [24] to automatically produce POET scripts as output without affecting how it works otherwise. The adapted optimizer has essentially delegated the actual program transformations to POET, where ROSE merely performs advanced analysis to make optimization decisions. The only modification to input code by ROSE is the tagging of code regions to be transformed later, an example of which is shown in Fig. 2(a). Our framework demonstrates the practicality of integrating POET as an alternative output language of existing optimizing compilers, so that compilers can interact with developers and empirical search engines in a much more explicit fashion than previously possible. This non-conventional approach entails the following technical challenges:

- Interactions between optimizations. Since AST transformations within ROSE have been redirected to a new POET scripting interface, each optimization now assumes that it directly operates on the original input code. However, since program transformations are actually applied one after another by the auto-generated POET script, the input to each transformation could be dramatically different from the original input code. This discrepancy could result in incorrect program transformations.

- Phase Ordering of Optimizations. Conventional compilers apply optimizations in multiple phases, with each phase focusing on managing a different set of architectural features. However, since the actual program transformations are now postponed and delegated to the auto-generated POET script, our ROSE optimizer can no longer separate these optimizations into different phases.

To resolve these issues, a key observation is that since no compiler optimization is allowed to alter the semantics of its input program, each optimization safe for the original input typically remains safe even after other optimizations have been applied. However, since the input code may have been modified by other optimizations, the auto-generated POET script must precisely model such interactions to ensure correctness of transformation. Further, optimizations at different phases often share the same program analysis. For example, if a loop nest should be blocked for cache reuse, it is safe and typically profitable to apply unroll&jam to the blocked loops to further promote register reuse. We therefore can use a single optimization analysis to drive related optimizations at different phases. Sections III discusses details of these solutions.

C. Programmable Control And Portable Tuning

After our adapted ROSE optimizer automatically producing a POET script, the script can be modified by a developer if necessary to change the ordering of transformations or to integrate additional domain-specific optimizations. The final script together with the tagged input source code can then be ported to a variety of different machines and empirically tuned. The POET interpreter and library, shown in Fig. 1, are lightweight and easily portable to different architectures. Compared to existing iterative compilation frameworks, our infrastructure clearly offers better modularity, flexibility and portability, as compiler optimizations are completely opened up for programmable control by developers, the optimizing compiler does not need to reside on the same machine that the user application is optimized for, all the optimizations applied by a compiler are made explicitly available for developers to modify if necessary, and an explicit well-defined parameterization space can be tuned using arbitrary independent search engines to find satisfactory optimization configurations. Every component is independent of the others through clearly defined interfaces. Researchers specialized in different areas can therefore easily collaborate with each other in building a collective high-performance computing infrastructure.

We have structured the auto-generated POET scripts so that it is trivially easy to disable a specific transformation or swap the ordering of different optimizations, for example, by simply modifying lines 36-46 of Fig. 3. However, it is more challenging to add new domain-specific optimizations into an existing POET script, due to the complexity of maintaining correct tracing of different transformations (which has been managed automatically by our ROSE optimizer in the auto-generated POET script). Our ongoing work includes adding GUI support for composing and modifying POET scripts so that developers can easily add domain-specific optimizations without explicitly maintaining the relevant trace handles.

III. Automatically Generating POET Scripts

We have adapted three existing optimizations within the ROSE loop optimizer: loop blocking, array copying, and loop unrolling, and have added analysis support for loop parallelization to automatically generate POET transformation scripts. The adaptation process includes the following steps.

1) Separate optimization analyses from transformations, so that the latter can be alternatively redirected to POET.
2) Generate a sequence of parameterized POET transformations based on results of the optimization analyses.
3) Monitor interactions between different POET transformations to ensure correctness of transformation.

The translation mapping between ROSE optimization analyses and the sequence of POET transformations is shown in Tab. II. The following addresses each of the above steps in detail.

A. Separating Optimization Analysis and Transformation

Most compiler optimizations can be naturally separated into an analysis and a transformation phase. The program analyses that we have adapted from ROSE include transitive loop
### TABLE II

<table>
<thead>
<tr>
<th>ROSE opt anal.</th>
<th>POET transformations</th>
</tr>
</thead>
<tbody>
<tr>
<td>thread level</td>
<td>cache level</td>
</tr>
<tr>
<td>reg. level</td>
<td>CPU level</td>
</tr>
<tr>
<td>loop parallel</td>
<td></td>
</tr>
<tr>
<td>OMP parallel.</td>
<td>loop block.</td>
</tr>
<tr>
<td></td>
<td>unroll&amp;jam</td>
</tr>
<tr>
<td>array copying</td>
<td>array copy.</td>
</tr>
<tr>
<td></td>
<td>scalar repl.</td>
</tr>
<tr>
<td></td>
<td>+ sit. reduct.</td>
</tr>
<tr>
<td></td>
<td>+ sit. reduct.</td>
</tr>
</tbody>
</table>

**OPT ANALYSIS**

- Optimization analysis: block. - blocking; str. reduct. - strength reduction

**TABLE III**

<table>
<thead>
<tr>
<th>Mapping between ROSE and POET Optimizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>GenPOETScript</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Algorithm for generating POET transformations**

1. Create a new POET script which starts with the inclusion of the *opt* library and a command to parse the input code and save the parsing result to a trace handle named target, as illustrated by lines 1-3 of Fig. 3.
2. Declare and initialize trace handles required to explicitly coordinate different POET transformations, as illustrated by lines 4-13 of Fig. 3. These trace handles are defined in Tab. III and explained in detail in Section III-C.
3. Declare command-line parameters to parameterize and optionally turn off each POET transformation, identified by each pair of individual ROSE optimization and POET transformation phases. These parameter declarations are illustrated by lines 14-24 of Fig. 3.
4. Independently generate each relevant POET transformation using the *DELAY* operator, illustrated by lines 25-35 of Fig. 3. For each optimization phase and saved ROSE optimization, the algorithm determines whether a POET translation is applicable and appends a new transformation to the POET script if necessary.
5. Invoke the delayed POET transformations and output optimization result, illustrated by lines 36-46 of Fig. 3.

**C. Interactions Between POET Transformations**

A key challenge in our approach is that after the ROSE optimizer is finished, the auto-generated POET transformations are no longer driven by any sophisticated program analysis. While the ROSE optimizer assumes all transformations are applied to the original input code, each POET transformation must operate on the result of previous transformations. Note each POET transformation is extensively parameterized and can be...
TABLE III
INTERACTIONS BETWEEN TRANSFORMATIONS IN Fig. 3

<table>
<thead>
<tr>
<th>Variable</th>
<th>Code fragment to trace</th>
<th>Used by</th>
<th>Modified by</th>
</tr>
</thead>
<tbody>
<tr>
<td>target</td>
<td>file the entire input code</td>
<td>opt. input</td>
<td>all opt.</td>
</tr>
<tr>
<td>nest1</td>
<td>outermost loop to transform</td>
<td>opt. input</td>
<td>all opt.</td>
</tr>
<tr>
<td>nest2</td>
<td>innermost loop to transform</td>
<td>opt. input</td>
<td>all opt.</td>
</tr>
<tr>
<td>nest3</td>
<td>middle loop to transform</td>
<td>opt. input</td>
<td>all opt.</td>
</tr>
<tr>
<td>incl_files</td>
<td>new include-file declarations</td>
<td>opt. output</td>
<td>par_nested1</td>
</tr>
<tr>
<td>decl_nested1</td>
<td>new variable declarations</td>
<td>opt. output</td>
<td>all opt.</td>
</tr>
<tr>
<td>top_nested1</td>
<td>code equiv. to the original nested</td>
<td>opt. output</td>
<td>all opt.</td>
</tr>
<tr>
<td>private_nested1</td>
<td>private vars for parallel nested</td>
<td>par_nested1</td>
<td>all opt.</td>
</tr>
<tr>
<td>cleanup_nested1</td>
<td>cleanup scope for nested</td>
<td>cleanup_nested1</td>
<td>par_nested1</td>
</tr>
<tr>
<td>top_nested1</td>
<td>input for unroll&amp;jam nested</td>
<td>unroll&amp;jam_nested1</td>
<td>block_nested1</td>
</tr>
<tr>
<td>nest1_C_dim</td>
<td>size of each dimension of array C accessed in nested</td>
<td>copy(C_nest, scalar_C_nest, par(C_nest, _block, C_block, scalar_C_block))</td>
<td>nested1</td>
</tr>
<tr>
<td>nest1_C</td>
<td>how to access array C using loop index variables in nested</td>
<td>copy(C_nest, par(C_nest, _block, C_block, scalar_C_block))</td>
<td>nested1</td>
</tr>
<tr>
<td>nest1_A_dim</td>
<td>size of each dimension of array A accessed in nested</td>
<td>copy(A_nest, scalar_A_nest, par(A_nest, _block, A_block, scalar_A_block))</td>
<td>nested1</td>
</tr>
<tr>
<td>nest1_A</td>
<td>how to access array A using loop index variables in nested</td>
<td>copy(A_nest, par(A_nest, _block, A_block, scalar_A_block))</td>
<td>nested1</td>
</tr>
<tr>
<td>nest1_B_dim</td>
<td>size of each dimension of array B accessed in nested</td>
<td>copy(B_nest, _block, scalar_B_nest, C_nest)</td>
<td>nested1</td>
</tr>
<tr>
<td>nest1_B</td>
<td>how to access array B using loop index variables in nested</td>
<td>copy(B_nest, scalar_B_nest, par(B_nest, _block, B_block, scalar_B_block))</td>
<td>nested1</td>
</tr>
</tbody>
</table>

opt.: optimization;

optional turned off entirely. Interactions between different transformations therefore must be dynamically traced to ensure correctness of program transformations.

Tab. III shows the collection of POET trace handles used to dynamically manage interactions between different transformations in Fig. 3. These trace handles are essentially place holders which can be embedded inside complex data structures (e.g., the AST of an input code) so that they can be automatically modified by the POET opt library even though the library routines cannot access them directly via their names. For example, when the opt library invokes the REPLACE operator in Tab. I to modify an AST e, the operator will automatically modify all the trace handles embedded inside e to contain the corresponding transformed code fragments. The trace handles in Tab. III are declared at lines 4-13 of Fig. 3 and can be categorized as follows.

- **Input trace handles**, e.g., target, nest1, nest2, and nest3, which are used to tag and dynamically trace various fragments of the original input code.
- **Output trace handles**, e.g., decl_nested1, incl_files, and top_nested1, which are used to save new code fragments created by various POET transformations.
- **Configuration trace handles**, e.g., those declared at lines 9-13 of Fig. 3, which are used to properly configure various POET transformations.

Only the input and output handles need to be embedded inside the input code so that they can be automatically modified by all POET transformations, as shown by lines 4-8 of Fig. 3. The configuration handles are modified by explicitly passing them as side-effect parameters of the opt library routines. The following explains how each POET transformation modifies these trace handles to dynamically coordinate with others.

![Fig. 5. Details of missing transformations in Fig. 3](image-url)
Loop blocking, illustrated by lines 12-22 of Fig. 5. This transformation uses a local variable \texttt{bdim} to set up how to block each loop and then block all three loops accordingly by invoking the \texttt{BlockLoops} routine (lines 16-21). The invocation of \texttt{BlockLoops} at lines 19-21 uses the \texttt{nonPerfect} parameter to specify that \texttt{nest2} is not perfectly nested inside \texttt{nest1} (see Fig. 2(a)) and therefore code sinking is necessary to block the loops. Two trace handles, \texttt{decl\_nest1} and \texttt{private\_nest1}, are modified to include a new loop index variable for each blocked loop. Additionally, the unroll\&jam trace handle \texttt{uj\_nest} is modified as a side-effect of \texttt{BlockLoops} to contain the inner tiled loops after blocking.

Array copying and strength reduction, which are illustrated by lines 23-47 of Fig. 5 and are enabled only when loop blocking has already been applied to \texttt{nest1}. If array copying is desired, the transformation uses a local variable \texttt{cpDim} to set up how to copy each array dimension, invokes the \texttt{CopyRepl} routine to copy array elements accessed by each block into a continuous region (a new array \texttt{C\_buf}), and then invokes the \texttt{FiniteDiff} routine to reduce the cost of accessing elements from the copied buffer via strength reduction (lines 29-39). If strength reduction is desired without array copying (lines 42-46), the \texttt{FiniteDiff} routine is invoked directly to simplify element accesses of the original array \texttt{C} after the surrounding loops have been blocked.

Two trace handles, \texttt{decl\_nest1} and \texttt{private\_nest1}, are modified to include the new loop index variables created to enumerate array elements being copied (lines 25-26). Additionally, as side-effects of invoking \texttt{CopyRepl} and \texttt{FiniteDiff}, \texttt{nest1\_C} is modified to contain an equivalent element access of the new array \texttt{C\_buf}, and both \texttt{decl\_nest1} and \texttt{private\_nest1} are again modified to contain the new variables created to facilitate the transformations. Finally, the dimensionality of the array is modified accordingly at line 40.

Loop unroll\&jam, which is illustrated by lines 48-57 of Fig. 5 and is the first register-level optimization in our current collection of auto-generated POET transformations. If loop blocking has already been applied (i.e., \texttt{uj\_nest1}!=\texttt{nest1}), all register level optimizations should operate on the inner tile (traced by \texttt{uj\_nest1}) resulted from the blocking transformation, so lines 49-54 of Fig. 5 adjust all the input trace handles (i.e., \texttt{nest1, nest3, and nest2}) to go inside \texttt{uj\_nest1}. Then, the actual transformation is applied if desired by invoking the \texttt{UnrollJam} routine at line 56.

Scalar replacement and strength reduction, which are similar to the array copying and strength reduction transformations illustrated by lines 23-47 of Fig. 5. The main differences are that scalar replacement does not depend on loop blocking, there is no \texttt{cpBlock} parameter to the invocation of \texttt{CopyRepl}, and it does not need additional loop index variables to enumerate array elements (i.e., lines 25-26 is no longer necessary). Further, \texttt{scalar} is set to 1 when invoking \texttt{CopyRepl}, and \texttt{cpDim, fdDim}, and copying locations are set differently to reflect the needs of register-level optimizations.

Loop Unrolling, which is illustrated by line 34 of Fig. 3 and does not require any additional tracing support besides the input/output trace handles used by all optimizations.

Cleaning up transformed code, which is invoked at line 35 of Fig. 3 and is the last POET transformation applied. It searches the input code region (\texttt{clnup\_nest1}) for special tags inserted by loop blocking and unroll\&jam and applies loop splitting to remove the if-conditional inserted inside loops by previous transformations.

D. Correctness, Robustness, And Generality

Our approach essentially uses the ROSE loop optimizer to determine potential optimizations applicable to an input code and then uses POET to dynamically compose the transformations one after another. The correctness of the POET transformations relies on the following conditions:

- The POET \texttt{opt} library is implemented correctly and modifies all the input and output trace handles as expected.
- After each transformation, all the input trace handles (e.g., \texttt{nest1, nest2, and nest3} in Fig. 2(a)) carry the same dependence constraints as their original code fragments.
- All the interactions between different transformations are explicitly modeled via the collection of configuration trace handles, and each transformation properly updates all the affected configuration trace handles.

In summary, as no optimization can modify the dependence constraints of the input code (guaranteed by the ROSE optimizer), and we keep all transformation configurations up-to-date by modifying the affected trace handles after each transformation, all the POET optimizations are guaranteed to be safe, unless given an invalid optimization configuration, where the POET \texttt{opt} library will exit with errors or simply skip the transformation. For example, if a search engine decides to invoke the POET script in Fig. 3 by blocking \texttt{nest1} by factors of (16 16 16) and then unrolling the innermost loop by 32, an error will be reported. Note that all POET transformations in Fig. 5 use the input, output, and configuration trace handles as parameters instead of directly using values generated by the ROSE optimizer, so they always operate on the most up-to-date values. For example, whether or not \texttt{par\_nest1} is applied, loop blocking and array copying are always applied to the sequential loop nest evaluated by each thread.

The POET language can express arbitrary source-to-source program transformations, including those that eliminate code (e.g., strength reduction) or modify control flow (e.g., loop blocking). The robustness of our approach is determined by its ability to always trace the corresponding modified code after each transformation. The tracing may be broken when given illegal combinations of optimization configurations or if the result of some transformation is not traceable. As a result some optimizations may have to be excluded from our approach if they render some input trace handles no longer traceable. The trace handles we use are based on the current collection of transformations supported by our ROSE optimizer. Some optimization configurations, e.g., the index variables \texttt{(j and i)} of array references at lines 27-28 of Fig. 5, are not traced as they are not modified by our current POET transformations.
As new optimizations are included, additional trace handles may need to be modeled.

A key technical contribution of this paper is a practical approach where compiler optimizations can be cleanly separated into two phases: program analysis and transformation, where the transformation phase does not require any intermediate analysis. This clean separation of concerns allows programmable control to be given to developers so that they can conveniently intervene and modify how their programs will be optimized. Our approach is dramatically different from how conventional compiler optimizations are implemented. As a result, it requires the tracing support of the POET scripting language. Since we use an optimizing compiler to automatically generate POET scripts, we have released HPC library (e.g., ATLAS [21]) developers from having to manually compose code optimizations from scratch and allows them to better utilize optimization capabilities within compilers.

IV. EXPERIMENTAL RESULTS

This paper focuses on adapting an existing source-to-source optimizing compiler to automatically generate extensively parameterized transformation scripts in POET so that developers can freely modify and extend optimization decisions by the compiler, and optimization configurations can be empirically determined via performance tuning on a wide variety of different architectures. To confirm that the auto-generated POET scripts can indeed achieve portable high performance across different architectures, we have selected several linear algebra routines and have used our ROSE optimizer both to directly generate optimized code and to alternatively produce POET scripts together with an annotated but un-optimized source code. We then compare the best performance achieved by the auto-generated POET scripts with that achieved using the ROSE optimizer only. Additionally, when applicable, we compare the best performance we achieved against that achieved by the code generator of ATLAS [21], which is well-known for its high performance matrix computation kernels.

A. Experimental Design

We have used our adapted ROSE optimizer to optimize four linear algebra routines, matrix-matrix multiplication (dgemm), matrix-vector multiplication (dgemv), vector-vector multiplication (dger) and LU factorization with partial pivoting (dgetrf). The source code of dgemm is shown in Fig. 2(a). The other routines are written similarly. These routines are chosen because they are known to benefit from the collection of optimizations that we support. We chose the dgetrf routine to demonstrate that through code sinking combined with a later cleanup step, the auto-generated POET scripts are capable of blocking arbitrarily nested loops (which is the case with dgetrf) despite the lack of program analysis support. For all benchmarks, the best performance achieved by the auto-generated POET scripts is compared against that achieved by using the ROSE optimizer alone. For dgemm, dgemv, and dger, the performance is additionally compared with that achieved by the code generator of ATLAS release 3.9.4.

We evaluated all benchmarks on two multi-core machines: a quad-core machine running Linux with two dual-core 3 GHz AMD Opteron Processors (each with 128KB L1 cache), and an eight-core machine running MacOS with two quad-core 3 GHz Intel processors (each with 32KB L1 cache). All the optimized code are compiled with -O2 option using gcc 4.2.4 on the AMD machine and gcc 4.4.4 on the Intel machine. We did not use -O3 to prevent gcc from applying aggressive loop optimizations to our already heavily optimized code. To discover the best performance achieved by each approach, we used an optimization-specific search engine implemented using Perl [17]. The search algorithm assumes domain-specific knowledge about each optimization being tuned and therefore can avoid being stuck at local minima which is a known problem for other general-purpose generic search algorithms. Our search engine has been used to tune both the auto-generated POET scripts and the adapted ROSE optimizer but has not been used in tuning the ATLAS generated code (which was tuned by the ATLAS search engine). We present the best performance found for all cases. Note that how to efficiently explore the immensely complex multi-dimensional optimization space is beyond the scope of this paper. Our framework is independent of specific search algorithms and can easily collaborate with other existing search engines in the literature[13], [19], [20], [27], [4].

B. Performance of Optimized Code

Figs. 6 and 7 show the best performance of the four linear algebra routines when optimized using five different approaches: auto-POET-small and auto-POET-large, the best performance achieved by auto-generated POET scripts using small and large input sizes (100*100 vs. 1000*1000 randomly initialized matrices) respectively; auto-ROSE-small and auto-ROSE-large, the best performance achieved directly by the ROSE loop optimizer without going through POET, using small 100*100 matrices and large 1000*1000 matrices respectively; and ATLAS, the best performance achieved by ATLAS release 3.9.4. The small and large input sizes are used to measure both the in-cache and out-of-cache/multi-threading performance of each optimized code. The ATLAS result is the out-of-cache single-threaded performance of its kernels.

From Figs. 6 and 7, auto-POET has performed significantly better than auto-ROSE for almost all cases. In particular, it outperformed auto-ROSE by factors of 2-10 for dger using large matrices on the 8-core Intel machine, dgemv using large matrices on both machines, and dgemm using both small and large matrices. It performed better than auto-ROSE by 25-50% for dgemv and dger using small matrices. For dgetrf, auto-POET performed similarly as auto-ROSE.

The performance of dgemm achieved by auto-POET is 20%-40% slower than that achieved by ATLAS. When using small matrices but is about 2-4 times better than that by ATLAS when using large matrices. For dgemv and dger, auto-POET has significantly outperformed ATLAS code generator for both small and large input sizes.
C. Interpretation Of Results

Since our benchmarks are individual routines containing single loop nests, we can use compilation flags to control how auto-ROSE optimizes each loop nest. Therefore while the auto-generated POET scripts support more comprehensive parameterization of the combined array copying and scalar replacement optimization, both auto-POET and auto-ROSE share a similar level of parameterization for loop blocking and unrolling. Auto-POET performs three additional optimizations: loop parallelization, unroll-and-jam, and strength reduction, than auto-ROSE. Since both approaches use the same empirical search algorithm, the better performance achieved by auto-POET is mostly due to the extra optimizations, which enable a much larger tuning space, instead of better strategies to explore the common portion of the tuning spaces. Note that the granularity of control via auto-ROSE becomes much coarser when optimizing larger routines with multiple loop nests. Iterative compilation can support extensive parameterization of optimizations internally. However, the internal optimizations cannot be controlled by outside developers or independent search engines.

When using small (100*100) matrices, neither cache blocking nor loop parallelization is necessary for any of our benchmarks. Consequently the best optimized code from both auto-POET and auto-ROSE use only a single thread, and auto-POET performs significantly better than auto-ROSE due to better parameterization of array copying/scalar replacement and the extra loop unroll&jam and strength reduction optimizations it applies to improve the register- and CPU-level performance of the benchmarks. Because matrix multiplication has a much higher computation/memory access ratio than the other routines, when using a single thread, the register- and CPU-level optimizations are critically important for the overall performance of the benchmark. ATLAS features a larger collection of such optimizations, e.g., memory prefetching and better cleanup code, than those currently available using our framework. As a result it achieves about 25-40% better performance than that by auto-POET.

When using large (1000*1000) matrices, auto-POET outperforms both auto-ROSE and ATLAS because of the extra OpenMP loop parallelization optimization. Here the performance improvements on the Intel machine are much higher than those on the AMD, as the Intel machine has 8-cores while the AMD has only 4 cores.

V. RELATED WORK

The initial design of the POET language was published by Yi et al. [25]. Yi and Whaley demonstrated that by manually writing POET scripts to optimize several linear algebra kernels, they can achieve performance comparable to that achieved by manually written assembly in ATLAS [26]. While designed to be easy to use, using POET to manually compose a long sequence of code optimizations remains a challenging and error-prone task for developers. This paper focuses on using a source-to-source optimizing compiler to automatically produce parameterized POET scripts. The goal is to enable seamless integration of the domain-specific knowledge possessed by computational specialists and the automated program analysis and optimization capabilities by compilers.

Several existing projects also focused on facilitating effective communication between optimizing compilers and developers to support better tuning of applications. The work by Hall et al. [8] allows developers to provide a sequence of loop transformation Recipes to guide transformations performed by an optimizing compiler. The X language [6] uses C/C++ pragma to guide the application of a pre-defined collection of compiler optimizations. Instead of asking developers to guide transformations applied by a compiler, we adapt an optimizing compiler to output its optimization transformations to be perused by developers. The degree of parameterization in our auto-generated POET scripts is much more extensive than that supported by existing other approaches.

Empirical tuning of performance has been successfully adopted by many popular scientific libraries, including ATLAS [21], PHiPAC [2], OSKI [20], FTFW [7], SPIRAL [15], among others, which use specialized kernel generators to parameterize and orchestrate differently optimized code. More recent research on iterative compilation has empirically modified the configurations of general-purpose compiler optimizations based on performance feedbacks [9], [16], [1], [14], [13]. None of these existing compilers support programmable control of optimizations by developers. Our framework aims to overcome this weakness by providing better means to support the integration of domain-specific knowledge and general-purpose compiler optimizations. Our auto-generated POET scripts can be easily integrated with existing search techniques [13], [19], [20], [27], [4] to automatically find desirable optimization configurations.

Existing compiler research has developed a large collection of optimization techniques for improving the performance of scientific applications [10], [11], [22], [3], [5]. Our approach...
can be used to similarly extend these optimizations to generate parameterized POET scripts for portable performance tuning and programmable intervention by developers.

A focus of this research is to develop effective techniques that allow collective parameterization of advanced compiler optimizations. Previous research has studied a number of compiler optimizations which have naturally parameterized configurations, including loop blocking, unrolling [9], [14], [18], software pipelining [12], and loop fusion [16]. The work by Cohen, et al. [5] used the polyhedral model to parameterize the composition of loop transformations applicable to a code fragment. Our work is different from the work by Cohel, et al. in that we parameterize the configuration of each individual transformation instead of parameterizing the overall combined optimization space. Our research focuses on composing these parameterized transformations in a well-coordinated fashion without intermediate program analysis support.

VI. CONCLUSIONS AND FUTURE WORK

We have presented a new optimization framework to support the programmable control and extensive parameterization of compiler optimizations so that developers can freely modify and extend optimization decisions by compilers, and optimization configurations can be empirically tuned on a wide variety of different architectures to achieve portable high performance. We have demonstrated the practicality of this framework by automatically generating programmable transformation scripts from a source-to-source optimizing compiler, and have shown that significantly better performance can be achieved by the more flexible and portable tuning framework than using the optimizing compiler alone.

Our approach opens up compiler optimizations to be controlled both by developers and independent search engines. It exposes an explicit well-defined parameterization space which is much bigger than the set of optimization flags supported by conventional compilers and can be tuned using arbitrary independent search engines. Our ongoing research is working on efficient exploration of the optimization space to enable better tuning of application performance.

REFERENCES


