1 Automatic generation of software

We describe a new paradigm in software development, which we refer to as software automation. This can be viewed as part of a natural evolution that has transformed systems research. The main ideas have their roots in programming languages and compilers.

We illustrate different areas of software development where this approach has proved essential. In particular, in the area of scientific software development, software automation has revolutionized an area with very demanding requirements. We give extensive examples.

Software automation has two components:

- (correct) translation from a high-level problem description to low-level executable code
- optimization of generated code.

As such, this describes the standard paradigm of a compiler for a programming language. What is different is the allowance for

- a hierarchical, or multifaceted, approach to the problem.

The typical programming language+compiler approach involves just two levels: the source language and the compiler output. In software automation, multiple levels exist, either hierarchically or in some more complex form of interaction.

The multifaceted approach of software automation facilitates leveraging inherent abstract representations of each segment of a software system. It can result in both better optimization and increased expressiveness in the problem description. In some cases, true optimization of generated code for a given segment can be performed, as opposed to just applying ad hoc transformations. It can both reduce cost and improve correctness of generated code.
1.1 Evolution in systems research

In the past, systems research was focused in major core areas, such as

- programming languages and compilers,
- databases, and
- operating systems.

Two things have happened to transform systems research. New areas have emerged, such as embedded, pervasive and ubiquitous computing. But as well, traditional areas have expanded: data mining has augmented database research; networking and distributed computing have expanded operating systems research. We argue that the traditional model of programming languages and compilers is also expanding into a new model, which we call software automation. As with other areas of systems, this area is more closely aligned with applications than the traditional base discipline.

1.2 High-productivity computing

There is a welcome shift in emphasis from a narrow focus on high-performance computing to the need for high productivity computing environments. In some contexts, this means parallel computing [45] involving the continuing evolution of computer architectures. This is still the dominant issue for problems where the scale of the computation is the limiting factor. But high productivity also means that software systems allow rapid advances in areas where new natural phenomena are incorporated in models. Rapid prototyping systems are often used in such a context, but these are typically low-performance systems (such as Matlab). There is a need to have systems that allow both rapid development of new models as well as rapid deployment on novel architectures. We indicate how software automation allows these twin objectives to be approached simultaneously.
2 Efficiency–expressiveness trade-off

There is a fundamental tension between expressiveness and efficiency. Ideally the naive user could code in a high-level specification language and have it translated into efficient machine code. So far, this remains only a dream in most areas of software development. The basic problem with a single-language approach is illustrated in Figure 1.

Some causes of the efficiency–expressiveness conflict can be identified. Typical compilers perform transformations, not optimization. Optimization at compile time, even if based on realistic performance models, would be wildly expensive, so only heuristics are applied in practice. Actual performance is often data dependent. True optimization would have to involve evaluating performance, and this has been done in certain contexts [13, 15].

Many problems have a hierarchical description, and transformations at one level of description may not be appropriate for another. The full efficiency is only obtained by performing the best transformations appropriate for each level. Current language support for hierarchical abstraction is limited. You can define the abstractions that you need, but you cannot specify how to compile it.

Many large codes are themselves multidisciplinary, and the best transformations are often domain dependent. Compiler optimizations cannot be chosen to fit the different parts of an application. One must live with what the compiler writer gives you.

2.1 Multi-lingual approaches

One solution that is often used in practice is multi-lingual. Different parts of the system are developed in appropriate languages, and they are merged in an ad hoc fashion. In a high-level (e.g., functional) language, new functionality can be introduced that allows efficient computation at a low level, as indicated in Figure 2.

One common example of this is the efficiency gain when one ‘vectorizes’ Matlab code. Analysa [2] combined efficiency and expressiveness by using a functional programming lan-
guage (AlScheme, an extension of Scheme providing among other things efficient linear algebra) as a scripting language which linked with C, C++ and Fortran code (for efficiency).

There are unfortunately serious limits to linking multiple languages informally. At the moment, it lacks formal description, requires development (and maintenance) of ad hoc interface. Different memory models (allocation, garbage collection) can interact adversely at run time. Our experience gained from the development of Analysa [2] shows that this can lead to difficult debugging problems.

2.2 Telescoping languages

Ken Kennedy proposed the use of ‘telescoping languages’ [22] in which fully optimized low-level code would be included as high-level operations in an extended language. This may be viewed as the reverse direction to what is done with the multi-lingual approach, as depicted in Figure 3.

The Broadway compiler [18] supports domain-specific compiler optimizations. It provides compiler support for a wide range of domains and in the context of existing programming languages using a technique called ‘library-level optimization’ which recognizes and exploits the domain-specific semantics of software libraries.

3 Hierarchy in software development

Some problems possess a natural hierarchical structure that allows a segmentation of the code generation process. For example, we might have the following levels.

- Model description [application domain]
- Algorithm discovery [mathematical description]
- Algorithm implementation [programming language]
In the typical programming language approach, the first three levels are all expressed in a single language. The compiler makes the transition from the top levels to the bottom level. However, there is potential for automation to be performed at each level. Interaction between levels can also be tested to optimize performance.

As an example of hierarchy, we consider the problem of signal processing and the solution provided by the Spiral project [41]. The hierarchical levels in this case are as follows.

- Model description [Discrete Fourier Transform: DFT]
- Algorithm discovery [Cooley-Tukey FFT]
- Algorithm implementation [FFTW]
- Executable code [BLAS, ATLAS, OSKI]

Automation has been performed at each level. For example, the Cooley-Tukey FFT algorithm can be derived automatically from an abstract definition of the discrete Fourier transform (DFT) and its mathematical properties. But optimizations can be carried out at each level. There is no need to hand-code for specific architectures.

Hierarchy is also common in other areas of scientific computing. Models are often built from simpler models:

\[
\text{Laplace} \implies \text{Stokes} \implies \text{Navier-Stokes} \implies \text{Grade-2 fluids}
\]

Moreover, there are natural abstractions to describe simulations of natural phenomena. Equations are the language of science and engineering:

\[
E = mc^2
\]
\[
F = ma
\]
\[
-\nu \Delta u + \text{curl}(u - a \Delta u) \times u + \nabla p = f
\]

The natural language of an application area can be exploited both to automate the generation of code and to provide an environment to carry out optimizations.

Not all applications have simple hierarchical abstractions. In general, more complex relations will occur. We explore one of these in Section 4.

4 Solving PDE’s: the FEM

Optimization of code for solving differential equations has been studied widely [31, 32, 47, 48]. Many of these approaches have been based on the finite element method (FEM). There are four distinct areas of finite element codes: function spaces, domain geometry/mesh, differential equation, and equation solution. These are not hierarchical, but rather interact in a more complex way. Each area has its own natural language and its own optimizations. But they also interact in ways that may require a type of inter-procedural analysis to obtain ideal performance.
4.1 FIAT and SyFi
The automation of code related to function spaces can be done in different ways. The principal task is to generate integrals of combinations of derivatives of finite element basis functions. There is often the need to interact with both the mesh and the PDE description. The evaluation of finite element basis functions, and related inner-product data has been automated by FIAT [24, 25] and by SyFi [38].

4.2 Finite element form compilers
One common representation for PDE’s is the language of variations forms. The variational forms must be compiled into code that is compatible with both the function space and the mesh.

The FEniCS Form Compiler (FEC) was introduced in [27] and further developed in [28]. AnalySia [2] also included a form compiler.

4.3 FErari
In [26], efficient evaluation of finite element matrices was addressed. This paper posed a complex optimization problem that was further studied in [29]. In [30], a different type of optimization, based on the geometric properties of certain tensors, was explored.

The paper [23] examines the effect of using complexity-reducing relations to generate optimized code for the evaluation of finite element variational forms. The optimizations are implemented in a prototype code named FErari. The authors demonstrate that by invoking FErari as an optimizing backend to the form compiler FFC, they obtain reduced local operation counts by as much as a factor 7.9 and speedups for the assembly of the global sparse matrix by as much as a factor 2.8.

4.4 Mesh generation
All discretization methods (finite element, finite difference, finite volume, spectral element, boundary element) require a mesh. This is an industry unto itself, so we do not try to survey it extensively, but give a few examples.

Triangle [46] is a program for efficiently generating 2D triangulations and Voronoi diagrams. Features include user-specified constraints on angles and triangle areas, user-specified holes and concavities, and the economical use of exact arithmetic to improve robustness. The paper [46] discusses many of the key implementation decisions, including the choice of triangulation algorithms and datastructures, the steps taken to create and refine a mesh, a number of issues that arise in Ruppert’s algorithm, and the use of exact arithmetic.

NETGEN [44] is an advancing front 2D/3D-Mesh generator based on abstract rules. The process of tetrahedral mesh generation is broken up into four steps: special point calculation, edge following, surface meshing, and volume mesh generation. Several techniques of mesh optimization are tested for quality.

The Bank-Holst adaptive meshing paradigm [3] is an efficient approach for parallel adaptive meshing of elliptic partial differential equations. It is designed to keep communication...
costs low and to take advantage of existing sequential adaptive software. The ‘isosurface stuffing’ algorithm [35] fills an isosurface with a tetrahedral mesh whose dihedral angles are bounded. It is fast and robust as it generates tetrahedra from a small number of precomputed stencils. It is the first algorithm that rigorously guarantees the suitability of tetrahedra for finite element methods in domains whose shapes are substantially more challenging than boxes. If the isosurface is a smooth 2-manifold with bounded curvature, and the tetrahedra are sufficiently small, then the boundary of the mesh is guaranteed to be a geometrically and topologically accurate approximation of the isosurface.

One issue is to be able to partition a given mesh for parallel calculation. The paper [39] describes an efficient approach to partitioning unstructured meshes that occur naturally in the finite element and finite difference methods. The approach makes use of the underlying geometric structure of a given mesh and finds a provably good partition in random O(n) time. It applies to meshes in both two and three dimensions. The new method has applications in efficient sequential and parallel algorithms for large-scale problems in scientific computing. This is an overview paper written with emphasis on the algorithmic aspects of the approach. Many detailed proofs can be found in companion papers.

Another issue is to maintain shape regularity as meshes are subdivided. The papers [40, 49] present efficient techniques for constructing simplicial meshes in which each simplex is small enough, according to an application specific error test, and the number of distinctly-shaped simplices in the mesh is small, depending only on the dimension of the problem. The techniques include the decomposition of simplices of a certain kind into smaller similar simplices, the refinement of a hierarchy of simplices to enforce an element size continuity condition, and the processing of such a refined hierarchy to produce a simplicial, tree-structured mesh.

## 5 Other scientific computing areas

Due to the mathematical nature of scientific simulation and analysis, it is natural that software automation can be applied pervasively. In section 3, we mentioned automation of signal processing code. We briefly mention other areas where automation has already reached a high level of development.

### 5.1 Quantum chemistry

Quantum chemistry has made major advances recently through automation. This has allowed more complex theories to be applied and has also provided a way to optimize code in novel ways. See [1, 4, 5, 42]. Also [6, 10, 11, 12, 16, 19, 34, 36, 37, 43].

We give some numbers to indicate the scale at which automation operates in the NWChem code. Table 1 lists some indicators of the automatic code generation for computation of dipole polarizability tensor single element evaluation. There are six terms per tensor evaluation, but only two unique pieces of code to compute them. Thus, the data is listed as m+n for each of the two pieces.

The ‘coupled-cluster order’ $n$ is a measure of the complexity (and accuracy) of the theory being used. The number of diagrams indicates the level of complexity of the interactions.
Table 1: Some indicators of the size of the automatically generated code for NWChem computation of dipole polarizability tensor single element evaluation. Note, there are six terms per tensor evaluation, but only two unique pieces of code. Thus, the \textit{m+n} is for each of the two pieces.

<table>
<thead>
<tr>
<th>Coupled-cluster order</th>
<th>number of diagrams</th>
<th>subroutines</th>
<th>F77 lines of code</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>12+80</td>
<td>29+269</td>
<td>3395+34302</td>
</tr>
<tr>
<td>3</td>
<td>21+266</td>
<td>42+807</td>
<td>6113+122463</td>
</tr>
<tr>
<td>4</td>
<td>23+560</td>
<td>40+1529</td>
<td>7053+313116</td>
</tr>
</tbody>
</table>

The number of indices in the largest tensors in the interaction terms in the computation is \(2n\). Thus for \(n = 4\) there are tensors of order eight, i.e., arrays with eight indices. Tensor contraction is a major part of the computation, and the automatically generated code optimizes these contractions.

5.2 Dense linear algebra

Dense linear algebra has also made major advances in terms of automation of code generation and optimization [7, 8, 9].

6 Program analysis and transformation

There are automatic tools that extract information from existing codes. Two areas are performance analysis and sensitivity analysis (a.k.a., differentiation).

6.1 Automating performance analysis

The paper [20] presents a framework for parallel performance data mining and knowledge discovery. The PerfExplorer framework is part of the authors' ongoing research into automatic parallel performance analysis. PerfExplorer addresses the need to manage large-scale data complexity using techniques such as clustering and dimensionality reduction, and the need to perform automated discovery of relevant data relationships using comparative and correlation analysis techniques. The intended uses of the framework include, but are not limited to, benchmarking, procurement evaluation, modeling, prediction and application optimization.

6.2 Automatic Differentiation

Algorithmic, or automatic, differentiation (AD) is concerned with the accurate and efficient evaluation of derivatives for functions defined by computer programs [17]. No truncation errors are incurred, and the resulting numerical derivative values can be used for all scientific computations that are based on linear, quadratic, or even higher order approximations to nonlinear scalar or vector functions. In particular, AD has been applied to optimization,
parameter identification, equation solving, the numerical integration of differential equations, and combinations thereof. Apart from quantifying sensitivities numerically, AD techniques can also provide structural information, e.g., sparsity pattern and generic rank of Jacobian matrices.

7 Software automation in other domains

Software automation is by no means limited to scientific computing. It can be applied in any area in which there is a suitable abstraction.

7.1 Compilers

Compilers can be defined formally since they translate one language into another in a well-defined way. This formalism provides a natural abstraction for automation. Not surprisingly, automatic generation of code has been performed in compiler design [14, 21].

7.2 Algebra of compiler optimization

In [33], the authors (according to their abstract) “use Kleene algebra with tests to verify a wide assortment of common compiler optimizations, including dead code elimination, common subexpression elimination, copy propagation, loop hoisting, induction variable elimination, instruction scheduling, algebraic simplification, loop unrolling, elimination of redundant instructions, array bounds check elimination, and introduction of sentinels. In each of these cases, we give a formal equational proof of the correctness of the optimizing transformation.”

References


12


