A COMPILER-BASED ONLINE ADAPTIVE OPTIMIZER

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
KAVON FAR FARVARDIN

CHICAGO, ILLINOIS
DECEMBER 2020
To my cats Jiji and Pumpkin.
Table of Contents

List of Figures vii
List of Tables ix
Acknowledgments x
Abstract xi

I Introduction 1

1 Motivation 2
1.1 Adaptation ...................................................... 4
1.2 Trial and Error .................................................. 5
1.3 Goals ............................................................ 7

2 Background 8
2.1 Terminology .................................................... 9
2.2 Profile-guided Compilation ................................. 10
2.3 Dynamic Compilation ....................................... 11
2.4 Automatic Tuning .............................................. 12
2.5 Finding Balance ................................................. 14

3 Thesis 16

4 Related Work 18
4.1 By Similarity .................................................... 18
4.1.1 Active Harmony ........................................... 18
4.1.2 Kistler’s Optimizer ....................................... 22
4.1.3 Jikes RVM .................................................. 23
4.1.4 ADAPT ....................................................... 24
4.1.5 ADORE ....................................................... 25
4.1.6 Suda’s Bayesian Online Autotuner .................... 26
4.1.7 PEAK ........................................................ 27
4.1.8 PetaBricks .................................................. 27
List of Figures

1.1 The affect of quadrupling the inlining threshold in the IBM J9 JVM compiler for the 101 hottest methods across 20 Java benchmark programs; from Lau et al. [2006]. ................................................................. 3

4.1 Client-side setup code for a parallelized online Active Harmony + CHiLL tuning session [Chen, 2019]. The tuned parameter names are known by the server to refer to loop tiling and unrolling for each loop nest. ............... 20
4.2 The main loop of an online Active Harmony + CHiLL application that uses the code-server to search for code variants of a naive matrix multiplication implementation for optimally performing configurations [Chen, 2019]. ............................. 21

5.1 The client-server separation used by HALO. ........................................ 34
5.2 The entry-point of patchable function, in an unpatched state. ................. 38
5.3 A patchable function that was dynamically redirected. .......................... 38
5.4 The function redirection routine. ....................................................... 39
5.5 An overview of Halo Server’s major structures and the flow of information. Solid arrows point to information consumers and dotted arrows indicate a dependence. .......................................................... 40
5.6 An example of two JSON-formatted knob specifications used by HALO server. .................................................................................. 41
5.7 A call-graph versus a calling-context tree for the same program. .......... 43
5.8 Computing the total IPC of the tuning section \( \{B, A, E\} \). .................. 45
5.9 An ancestor chain of the CCT used to choose a tuning section root. ........ 47
5.10 A state machine for a once-compiled tuning section (i.e., the jit-once manager). ........................................................................ 49

6.1 The state machine for the Adaptive tuning manager. ............................ 52
6.2 An overview of the calculation to determine how to amortize a bakeoff’s overhead during the Payback state. Figure is not to scale. ............. 59
6.3 Probability of selecting library \( i \) in the Retry state. ............................ 61

7.1 An example decision tree from the spectralnorm benchmark. ............... 71

8.1 Results for the workstation machine. Higher speedups are better. ......... 76
8.2 Results for the desktop machine. Higher speedups are better. .............. 77
8.3 Comparing the Call and IPC metrics on the workstation. Higher speedups are better.
List of Tables

7.1 Settings tuned by HALO for each tuning section. All options are integers. The Default column indicates LLVM’s default setting for the given option. 63
Acknowledgments

First and foremost, I would like to thank my advisor John H. Reppy for his ample support, kindness, and advice over the past six years. Thanks also to the rest of my dissertation committee, Ravi Chugh, Hal Finkel, and Sanjay Krishnan, for their time and feedback. I am also grateful for my summer internships with Hal Finkel and Simon Peyton Jones that greatly expanded my horizons.

This dissertation would not exist without the social and intellectual support from my friends, family, and colleagues: Amna, Andrew M., Aritra, Brian, Charisee, Connor, Cyrus, Elnaz, The Happy Hour Gang, The Ionizers, Jean, Joe, Justina, Kartik, Lamont, The Ross Street Crew, Saeid, Sheba, Suhail, Will, and numerous others. In ways both subtle and overt, many of them have helped me grow as a person, persevere through difficult times, or both; I am incredibly lucky to have had them in my life.
Abstract

The primary reason for performing compiler optimizations before running the program is that they are “free” in the sense that the analysis and transformations involved in applying the optimizations do not contribute any overhead to program execution. For optimizations that are inherently or provably beneficial, there is no reason to perform them dynamically at runtime (or online). Other optimizations are situational or speculative in the sense that their profitability depends on certain assumptions that cannot be fully determined statically, since they depend on dynamic program behavior or machine characteristics.

Today, the main use of online adaptive optimization has been for language implementations that would otherwise have high baseline overheads. This overhead is normally due to the use of an interpreter or a large degree of dynamically-determined behavior (e.g., dynamic types). The optimizations applied in online adaptive systems are those that, if the profiling data is a good predictor of future behavior, will certainly be profitable according to the optimization’s cost model. But, there are many compiler optimizations whose cost models are unreliable!

In this dissertation, I describe an adaptive optimization system that uses a search-based approach, i.e., it uses a trial-and-error search over the available compiler optimization flags to automatically tune them for the program with respect to its dynamic environment. This LLVM-based system is unique in that it is fully online: all profiling and adaptation happens while the program is running in production, thanks to newly developed techniques to manage the cost of tuning. The system delivers nearly a $2 \times$ speed-up for some C/C++ benchmarks relative to the best optimizations available in a production-grade compiler.
Part I

Introduction
Chapter 1

Motivation

The primary reason for performing compiler optimizations before running the program is that they are “free” in the sense that the analysis and transformations involved in applying the optimizations do not contribute any overhead to program execution. For optimizations that are inherently or provably beneficial, there is no reason to perform them dynamically at runtime (or online). Other optimizations are situational or speculative in the sense that their profitability depends on certain assumptions that cannot be fully determined statically, since they depend on dynamic program behavior or machine characteristics.

For example, unrolling a loop is situational because it is only profitable under the assumption that the loop has a large iteration bound. This assumption helps ensure that the overhead of additional code added by the optimization, which decides between entering either the unrolled or remainder loops, will be paid-off. Evidence to support such assumptions can sometimes be inferred statically, but profile data, which provides information about the dynamic behavior and environment of the program, is often what is relied upon when estimating whether a situational optimization will yield a profit.

As an additional example, let us consider expansive function inlining, i.e., the inlining of a function that has more than one call-site. Inlining can be viewed as a speculative optimization because it specializes the body of the function to a particular call-site. This specialization provides later intraprocedural optimization passes with information about how the arguments at that call-site will be used. The downside is that too much inlining can also degrade performance, because it increases the size of the program and causes ill effects during execution such as instruction cache misses. Compiler-writers often navigate this inlining trade-off by setting a simple threshold on the size of the caller or callee when considering to inline a particular call-site. But, a simple metric such as function
size does not guarantee an optimal decision. For example, Lau et al. [2006] experimented with quadrupling the default inlining threshold, which limits the size of the callee to be inlined, in the production-grade IBM J9 compiler. Figure 1.1 shows that the inlining trade-off is not optimally reasoned about with just a function size threshold. Raising the size threshold significantly benefits a handful of functions but significantly degrades another handful, with little change for the rest. Is there a better way to navigate the trade-offs when deciding whether to inline a call-site?

Cost Models All compiler optimizations rely on a cost model to account for performance trade-offs. A cost model predicts the net benefit of the transformation on the program’s performance, signaling whether the transformation is likely to be profitable or not. These cost models may be simple or not explicitly stated, e.g., the implicit model behind eliminating useless code is that executing less code will reduce execution time.

For optimizations where an explicit cost model is needed to accurately estimate the trade-offs, such as for inlining, the models are often parameterized to rely on abstract threshold values, profile data, or both. The function-size threshold for inlining and the large loop-bound for unrolling are examples of simple but explicit cost models. Profile data helps by providing concrete evidence to support or refute assumptions, such as “this loop’s bound is large,” because it reduces uncertainty. The trouble is that profile data can only be collected after compilation by executing the program.
1.1 Adaptation

A compiler is a program translator and does not execute the program like an interpreter does. Compilers that are run prior to program execution, or ahead-of-time (AOT), cannot access profile data without additional help from the programmer. The programmer must set-up a build system that compiles the program once and performs a profiling run on a test workload. Then, the profiling data is fed back into the compiler again to recompile the program. This feedback loop of profile data is at the heart of adaptive optimization because it allows the compiler to adapt its optimizations based on information it is not cognizant of statically.

In the absence of profile data to guide cost models, compilers make either pessimistic or optimistic decisions, such as disabling the optimization or using heuristics to make assumptions for the typical program. Kennedy and Allen [2002, Chapter 5.11] discuss an example of this phenomenon for loop vectorization, where an optimistic assumption is used when faced with statically unknown loop bounds during cost modeling:

When loop bounds are symbolic, trading off loop lengths versus strides and other parameters is very difficult. With no additional input from the programmer, compilers must generally assume that all loops with unknown bounds are long enough for efficient vector execution.

Pessimism and optimism are two general approaches to dealing with risk, which in our case takes the form of the compiler possibly yielding a program with poor performance. As an alternative, the adaptive approach is to get a clearer picture of that risk through experimentation, in the form of actually running the program, and then basing the optimization’s decisions on the information gained. The fundamental downside of using purely static optimization instead of adaptive optimization is that less information known about the program’s runtime behavior or environment to make smart decisions [Adve et al., 1997].

Online Adaptive Optimization   Adaptive systems that operate online (i.e., during program execution) can take advantage of up-to-the-second profile information and quickly react to changes in a program’s workload or environment. The key advantage of performing online adaptive optimization (OAO) is that the optimizer can directly obtain fresh, runtime-only information instead of making pessimistic optimization decisions.

A number of implementations of programming languages such as JAVASCRIPT, JAVA, and SELF [Ungar and Smith, 1987] have been very successful in applying online adaptive
optimization [Gal et al., 2009; Hölzle, 1994; Kotzmann et al., 2008]. These language implementations employ online adaptive optimization in the form of a profiler that gathers and analyzes data for a dynamic, or just-in-time (JIT), compiler in order to make better optimization decisions. Online adaptive optimization is not just an idea within the ivory tower of academia: it is a crucial technique for language implementations used by everyday people [Arnold et al., 2005; Aycock, 2003]. For example, all major web browsers, such as Firefox, Safari, and Chromium, rely on online adaptive optimization in their runtime system to accelerate their JavaScript interpreter [Gal et al., 2009; Google; Pizlo, 2014, 2016]. The runtime systems in these web browsers are so efficient at optimizing JavaScript that they have been repackaged and are widely used to run server and desktop applications too [OpenJS Foundation, a,b].

1.2 Trial and Error

Today, the main use of online adaptive optimization (OAO) has been for language implementations that would otherwise have high baseline overheads. This overhead is normally due to the use of an interpreter or a large degree of dynamically-determined behavior (e.g., dynamic types). The optimizations applied in online adaptive systems are those that, if the profiling data is a good predictor of future behavior, will certainly be profitable according to the optimization’s cost model.

Let us consider two examples of how the profile data is used in existing OAO systems. First, to reduce the overhead of running an interpreted program, profile data is used to balance the cost of running the JIT compiler with the benefits of patching-in a native version of the most commonly executed parts of code [Plezbert and Cytron, 1997]. As long as the program continues to execute those regions that were compiled to native code, the cost of running the compiler will quickly pay for itself. Second, compiler optimizations are forced to pessimize about how the value can be used when faced with values whose types can vary dynamically. So, profile data about particular program values is used to determine whether the type of a value appears to be stable; if so, the code is customized based on that assumption [Calder et al., 1997; Chambers et al., 1989]. If the type of the value changes in the future, then the optimization may be undone. Otherwise, the type-based customizations are at least as efficient as the previous version of the code and usually significantly better. In both examples of optimizations commonly used in today’s OAO systems, the cost models for these optimizations are accurate as long as the profile data is an accurate summary of future behavior.
Search Methods  What if a compiler optimization’s cost model is unreliable?  As discussed earlier, the trade-offs of function inlining are complex and difficult to model, even with the availability of profile data. Instead, compiler-writers commonly use simple heuristics and abstract threshold values, which are hand-tuned to perform well on a fixed set of benchmark programs. Thus, the cost model may not reliably produce optimal answers, which prevents programs from performing better due to the model’s overly optimistic or pessimistic assumptions.

Utilizing search is one way to overcome an unreliable or incomplete cost model. Empirical autotuning (a portmanteau for “automatic tuning”) is a form of adaptive optimization that follows a search-based ideology. For example, instead of cost-modeling, we can simply experiment with all combinations of optimization decisions, profiling the program’s overall performance, and pick the the combination with the best rewards based on the data. Thus, instead of feeding back profile data for the purposes of informing the optimization’s cost model, the data is used by an autotuner as an indicator of the quality of one combination of decisions.

High-performance computing applications, such as scientific simulations on supercomputers, are the main users of autotuning today [Balaprakash et al., 2018]. An example of an empirical autotuning system for C programs is Orio [Hartono et al., 2009], which tunes loop optimizations such as unrolling, interchange, and tiling. When applied to computational kernels like sparse matrix computations and a sequence of linear algebra operations, Orio was able to find configurations for loop optimizations that match or drastically improve performance relative to both production-grade compilers and hand-tuning by expert programmers. The magnitude of improvements for some kernels tuned by Orio ranged from 26% to 277%! Thus, search methods can find significant performance improvements that are missed by the typical cost models used by compiler optimizations.

Performing a search avoids the tension between optimism and pessimism because it eliminates the risk caused by uncertainty through experimentation. But, searching creates another avenue of risk: failing to find a worthwhile performance gain for the time and machine resources spent. Since the space of possible optimization decisions can grow exponentially, autotuning can quickly become intractable, regardless of the resources available or the search technique used. In some instances, the time necessary to perform the search can add days to the deployment of software to users [Hoste et al., 2010]. If the autotuning process relies on a test workload that does not truly represent an end-user’s workload, which can be expected for large-scale applications [McFarling, 2003], the program’s performance may end up worse than if tuning were not performed at all!
1.3 Goals

Empirical autotuning offers an opportunity to harness the full power of compiler optimizations, but suffers from a number of practical drawbacks:

- Developers have to create a representative input for the software to accurately simulate its end-user’s usage, and perform tuning on the hardware that end-users are expected to use.
- The build and release process for the software must be augmented with search and final-compilation phases.
- Once the software is deployed, the tuning remains fixed even as the hardware and end-user’s usage evolves over time.

The goal of this dissertation is to investigate whether online autotuning for compiler optimizations can overcome these drawbacks, while still improving the overall performance of the tuned software. Online autotuning delays the process of tuning until after the software has been deployed in production to end-users. Specifically, experimentation happens live on the end-user’s system; no test inputs are required. The use of search is what distinguishes online autotuning from more traditional language runtime systems that feature just-in-time compilation.

Online autotuning is not a new idea (Section 4.1), but it suffers from a number of unique challenges that have prevented broader application. Central to these challenges is the balancing of search overheads with the performance gains. For example, the search may discover badly-optimized code in a failed experiment. How do we minimize the impact of this experiment on the program while it is running in production? On the other hand, if the search discover better-optimized code, what should be done afterwards? Specifically, for how long should the search pause to exploit the better version of the code, e.g., to repay the accrued debts of the search?

Roadmap  [[ describe structure and organization of the dissertation. ]]
Chapter 2

Background

Many experts throughout the years have highlighted the need for adaptive optimization to fully access the hardware’s peak performance. Griswold et al. [1996] believe that “a small, stable programming language with abstraction features that support the development of self-tuning, optimizing, easily adaptable, integrable layered systems” is needed for the next millennium. Smith [2000] advocated for the use of online adaptive optimization with concrete examples where an online approach could better optimize programs. Smith also believes that the largest barrier to adaptive optimization is people’s aversion to mutating code, because of their fear of it silently introducing bugs. Since then, dynamic code generation has become a widely-accepted technique. When envisioning the next 50 years of compiler research, Hall et al. [2009] paraphrased a private communication with researcher David Kuck who discussed the importance of adaptive optimization:

Compiler fundamentals are well understood now, but where to apply what optimization has become increasingly difficult over the past few decades. Compilers today are set to operate with a fixed strategy (such as on a single function in a particular data context) but have trouble shifting gears when different code is encountered in a global context (such as in any whole application).

Kuck also said, “The best hope for the future is adaptive compilation that exploits equivalence classes based on ‘codelet’ syntax and optimization potential. This is a deep research topic, and details are only beginning to emerge. Success could lead to dramatic performance gains and compiler simplifications while embracing new language and architecture demands.”

After discussing some important terminology, this chapter will explore a few general
forms of adaptive optimization, at a high-level, to help provide necessary context. Then Chapter 4 takes a deeper look at the most closely related prior works and how they differ from this work.

2.1 Terminology

There is no consensus on the terminology surrounding adaptive optimization, but we will try to form a coherent universe of definitions. Smith [2000] defines *feedback-directed optimization* in the most general sense: “any technique that alters a program based on information gathered at run time.” Then, Smith focuses specifically on feedback-directed optimization that improves the performance of programs. Others such as Hansen [1974] opted to use the term “adaptive” instead of “feedback-directed,” but use the term in a way consistent with Smith. According to Cooper et al. [2002], the term *adaptive compilation* refers to the specialization of the compiler’s optimization pipeline to particular programs, workloads, machines, or any combination of those, to minimize some objective function; but their definition does not require profile data. In contrast, *adaptive recompilation* refers to a technique that uses runtime information to dynamically select heavily-used code and recompiles (and optimizes) them for better performance [Hölzle, 1994]. Throughout this dissertation, *adaptive optimization* has the same meaning as *feedback-directed optimization*, despite the fact some works consider adaptation without runtime information.

The typical goal of adaptive optimization is to minimize metrics such as cache misses, energy consumption, or the execution time of frequently executed, or *hot*, code regions. These metrics may be directly or indirectly influenced by a set of configurable parameters for which values can chosen arbitrarily without affecting the correctness of the program.\(^1\) I refer to such parameters as *knobs* because “parameter” is an overused term.

Much like its physical counterpart, a knob is defined in terms of a finite space of configurations, or options, and a state which indicates exactly one element from the configuration space that is currently-selected at any point in time. A collection of knobs can be thought of as a single knob whose configuration space is the cross-product of the spaces of each individual knob in the collection.

The kinds of knobs that one can imagine influencing a program are numerous. For example, an algorithm may offer a knob with two options, selecting either a linked-list or an array, that changes the underlying data structure used by the implementation. Changing

\(^1\)Work in approximate computing feature knobs that trade program correctness for performance [Sidiroglou-Douskos et al., 2011]. Such perilous knobs will not be considered here.
this knob only affects the performance of the program, not its result. While the inputs or
data processed by the program affects the program’s performance, they are not normally
considered knobs unless if the values can be instantiated arbitrarily among its allowed
options.

The focus of this work is on the adaptive optimization of knobs that control the code
optimization process within a compiler. Widely-used, production-grade compilers such
as GCC have over a hundred such knobs available for experts to tweak, controlling vari-
ous parts of the compilation pipeline [Fursin et al., 2014]. Ordinary programmers are
typically familiar with just a single knob that controls the compiler’s general optimiza-
tion level: an \(-\text{On}\) command-line flag, where \(n \in [0, 3]\) and a larger \(n\) tells the compiler to
spend more time to aggressively optimize. But, ignoring the other knobs available within
a compiler can leave significant performance gains untapped. For example, using just
unbiased random selection over the knobs in GCC, Fursin et al. [2014] found 79 unique
combinations of flags that yielded better performance over \(-\text{O3}\), in some cases up to \(3.75\times\)
 faster, after 5000 search iterations. For the remainder of this chapter, we will take a tour
of some of the major works within the realm of compiler-based adaptive optimization.

2.2 Profile-guided Compilation

One of the earliest advocates of the use of a program profile data for compilers is by Knuth
[1971], who conducted an eye-opening study of real-world FORTRAN programs. Knuth
recognized the disconnect between what compiler writers thought was important to op-
timize and the kind of code that programmers wrote in practice. Program profiles of two
forms were analyzed: static and dynamic.

A static profile is information about the program that is obtained through static analy-
sis, i.e., analyzing the program’s source code without executing it. Knuth was interested
in the style of code written in FORTRAN, such as how often GOTO statements appear in
the code. These statistics were intended to inform compiler writers about where to focus
their attention, based on how popular various language constructs are in practice. While
static analysis is limited because it does not execute the program, there are a number of
analytical techniques and effective heuristics that infer dynamic information [Ball and
Larus, 1993; Cousot and Cousot, 1977; Patterson, 1995; Wagner et al., 1994].

In contrast to a static profile, a dynamic profile is information gathered through the ex-
ecution of the program about its runtime characteristics. These characteristics are typically
statistics about the program’s source code, such as how often a function or statement
is executed. There are two overall strategies for obtaining dynamic profile data. One is sampling-based profiling that periodically interrupts a program thread to examine its state, such as the next instruction to be executed or the callers of the current function. The other is instrumentation-based profiling that relies on the injection of code in places such as function entry-points and loop-backedges to record whenever those points in the program are reached. Throughout this dissertation, discussion of an unqualified “profile” refers to the dynamic kind.

Dynamic profile data can inform both humans when writing their programs and compilers when optimizing them. Techniques where a compiler uses profile data that consists of information that indicates where the program spends its time to inform the optimizations applied is called profile-guided compilation (PGC) [Smith, 2000]. Profile-guided compilation has found its footing in many widely-used compilers such as GCC and CLANG [Chen et al., 2016; Hubička, 2005]. The kinds of profile data used in PGC typically includes statistics about how frequently a control-flow edge or path within a function is executed [Ball and Larus, 1994; Ball et al., 1998]. These profiles serve many uses, such as to optimize code layout for cache locality [Pettis and Hansen, 1990], focus compilation-time on the important parts of the program [Whaley, 2001], inform cost models for speculative optimizations [Bodík et al., 1998; Gupta et al., 1997; Hazelwood and Conte, 2000], and even for instruction selection [Krishnaswamy and Gupta, 2002].

2.3 Dynamic Compilation

Dynamic compilation is any technique that uses a compiler during program execution, and is often called just-in-time (JIT) compilation. Early JIT systems used to implement virtual machines (VMs) were compile-only in the sense that VM code compiled to native code and then executed when control-flow first reaches it. Plezbert and Cytron [1997] observed that it is sometimes faster to interpret the code instead, pioneering a profile-driven technique for mixed-mode execution that interleaves interpretation and native execution of the program. This adaptive technique is now widely found in language runtime systems for JAVA and JAVASCRIPT [Gal et al., 2009; Hookway and Herdeg, 1997; Kotzmann et al., 2008; Paleczny et al., 2001]. The CoreCLR language runtime system for C# (and other languages) uses a mixed ahead-of-time and just-in-time compilation strategy rather than interpretation [Strehovský, 2019]. For a more complete overview of early JIT compilation techniques, see Aycock [2003].

The LLVM-based ClangJIT project by Finkel et al. [2019] aims to bring JIT-compilation
to C++. through annotations on template definitions. This essentially turns C++ into a sort of multi-staged language, because template instantiation can be selectively delayed until runtime and specialized on dynamic values or types [Veldhuizen, 2000]. Depending on the availability of in-language profiling facilities, a multi-staged language could offer the ability to implement adaptive optimization as a library.

**Code Multi-versioning** In essence, code multi-versioning is a solution to the classic code selection problem, where the performance of an algorithm or piece of code has a strong input and/or environment dependence [Rice, 1976]. Compilers featuring ahead-of-time multi-versioning suffer from a combinatorial explosion of code versions that need to be generated and stored in the binary, though recent work has used profiling to help mitigate the problem [Rodriguez et al., 2016; Zhou et al., 2014].

Dynamic code multi-versioning is most frequently realized in runtime systems using a tiered JIT optimization approach. A tiered JIT consists of several levels of optimizations that are reserved for different levels of code importance, as determined by profiling for code hotness [Arnold et al., 2000]. These tiers help balance the cost of compilation with the potential code improvements seen through more aggressive optimization, since compilation happens concurrently with program execution on the same machine. Gu and Verbrugge [2008] were able to use dynamic tracking of program phases to help predict whether a method is worth recompiling at a higher optimization level, since the time spent at a lower level is lost performance. JIT-based runtime systems also version and dynamically select code based on other information about the code, such as the most common kinds of arguments to a function [Hölzle et al., 1991].

### 2.4 Automatic Tuning

A type of adaptive optimization called autotuning (short for “automatic tuning”) is the process of identifying an optimal knob configuration within its space of possible choices. At its core, autotuning is an optimization problem where the goal is to find an optimal knob configuration within a huge space of possible configurations. At a high-level, we can view this problem of autotuning more formally as an instance of black-box optimization (BBO). The “black-box” is $f$, the unknown cost function (or fitness function) that takes as input a configuration $\theta$ drawn from a large space of configurations $\Theta$ and out-
puts the configuration’s cost $c$. The goal is to find (or “select”) $\theta' \in \Theta$ such that $f(\theta')$ is minimal. What makes this goal challenging is that the cost function is assumed to be expensive to evaluate, so exhaustive search methods that guarantee global optimality are not feasible. In autotuning, the cost function may involve the process of recompiling the application according to the configuration, followed by a number of executions under a sample workload to obtain its average running time, which is output as the cost. There are three major strategies for tackling the selection problem for autotuning, which I will describe in the following paragraphs. The space of existing work in automatic tuning is huge; for a good overview see Balaprakash et al. [2018] and Ashouri et al. [2018].

Model-free Selection A wide variety of heuristic search methods have been developed to navigate arbitrary configuration spaces efficiently [Blum and Roli, 2003]. A model-free strategy focuses on sampling the configuration space by choosing configurations and executing $f$ to determine their quality. Two major types of meta-heuristics for model-free selection are trajectory methods that focus on the neighborhood of their best known state (e.g., hill climbing and simulated annealing [Bertsimas and Tsitsiklis, 1993]) and evolutionary strategies that maintain a population of high-fitness states for recombination over a series of generations (e.g., genetic algorithms, neuro-evolution [Stanley and Miikkulainen, 2002]). For example, Cooper et al. [1999] used genetic algorithms to search the space of compiler flags to minimize the executable’s size for embedded systems. Knijnenburg et al. [2003] applied several heuristics, such as simulated annealing and random search, to tune loop unrolling and tiling factors for a number of CPU architectures. All of these search-methods obtain convergence by assuming the cost function is deterministic and remains fixed throughout the search process. This assumption does not necessarily hold for online automatic tuning.

Model-based Selection An autotuner that uses model-based selection relies on analytical models or machine-learning to learn how to select an optimal configuration. Thus, model-based strategies avoid employing search and directly produce a tuned configuration based on the characteristics of the program or machine. For example, Cavazos et al. [2007] collected data to train a model that learns how to predict good compiler optimization flags based on the state of performance counters. This pre-trained model is then used by their system to select compiler optimizations specific to the input program. The input

---

2In the case of multi-objective optimization, $c$ may be a vector or there may be multiple cost functions [Durillo and Fahringer, 2014].
program is run a few times to collect the performance counter data, which is input to the model to select one good configuration. Similarly, Milepost GCC [Fursin et al., 2011] uses a model trained to predict good compiler optimization flags based on features of the input program, e.g., the number of basic blocks or number of assignment instructions. A number of additional systems are summarized by Wang and O’Boyle [2018].

**Hybrid Selection** The hybrid selection strategy combines search-based strategies with an analytical or machine learning model. In the typical formulation, it is a strategy that relies on a model to help focus a search over the configuration space. For example, Balaprakash et al. [2013a] train a model to learn about the cost function during a search to progressively learn how to filter out future poor-performing or less-useful configurations. Some formulations train the model prior to using it to focus the search [Agakov et al., 2006]. In Active Harmony, the sensitivity of each knob is modeled based on configurations already tested to better-focus the search [Chung and Hollingsworth, 2004]. Pouchet et al. [2008] leverage an analytical model to focus the search over configurations of high-level loop optimizations. As a final example, OpenTuner dynamically switches between multiple heuristics during the search, using a simple adaptive model that learns which heuristic has been most successful recently [Ansel et al., 2014].

### 2.5 Finding Balance

The major challenge faced by all online optimization systems is reaching a *break-even point*, which is the point at which the costs accrued reoptimize is paid-off by using the code version that is better than the original [Diniz et al., 1997; Kistler, 1999]. The optimizer may not reach a break-even point for a code region if: (1) the program’s use of the optimized code region ends before the accumulated debt is paid off, or (2) the optimizer is unable to produce code that is better than the baseline.

The first scenario is primarily mitigated through dynamic profiling and a prediction model that determines whether the code region is likely to remain heavily-used in the future. For example, in early versions of the influential HotSpot adaptive optimizer, a code region is predicted to be hot if instrumentation-based performance counters exceed thresholds that were determined heuristically [Paleczny et al., 2001].

To ensure profitability under the more challenging second scenario, it is important for an online adaptive optimizer to focus on high-impact optimizations for hot code regions, while keeping the system’s operating overhead low [Kistler, 1999]. An online system
will always introduce some sort of overhead, because replacing or modifying a piece of code at runtime has a cost. In fact, an online system may replace the code multiple times, either for profiling purposes or to experiment with differently-optimized versions of the code during a search. Furthermore, these new versions of code may have much worse performance than what it replaced, adding onto the costs. Thus, knowing when to stop reoptimizing is important. Vuduc et al. [2004] developed a statistical technique that estimates the probability of finding a better configuration, based on the results of an empirical trial-and-error search.
Chapter 3

Thesis

The specific goal of this dissertation is to investigate the following high-level question: can an automatic, online, search-based, adaptive recompilation system be effective in improving the performance of programs? The key factors that distinguish this question from existing work is the use of search-based methods in an automatic, online system. The remainder of this section expands on what is meant by “effective” and “performance” under the context of the three broad barriers facing the wider use of search-based adaptive optimization, as described by Basu et al. [2013]: usability, generality, and managing overheads.

Usability There are two groups of software developers for whom usability of an adaptive optimizer matters: average developers and domain experts. A domain expert possesses the knowledge to profile and manually optimize their performance-critical system. Thus, the major usability concern for an expert user of adaptive optimization is its ability to deliver performance improvement that matches or exceeds the expert’s ability to perform manual optimization. On the other hand, an average application developer is interested in using adaptive optimization to meet their performance needs, but do not know precisely how to achieve them. An easy-to-use, portable system is a bigger usability concern for average users than its performance gains. This research aims to evaluate the usability of adaptive optimizers for average developers.

Generality Basu et al. highlights two generality concerns: the customization available to expert users across different problem domains to guide the system, and the compatibility across different programming languages, operating systems, and hardware. Manual customization of the optimization process by an expert, i.e., programmer-directed adaptive optimization, may yield better results sooner than a fully automatic system. This type
of customization is not required for an online system to be considered effective, for two reasons.

First, in an online system, optimization is meant to occur during useful program execution, so the rate at which the optimizer can converge is constrained by the program’s execution behavior and the overheads added by the online system’s infrastructure. Thus, the value of customization is less clear in an online system, unless it is used like an offline optimizer, where the program is repeatedly performing the exact same work in a test environment. Second, customization primarily benefits domain experts, but my focus is on usability by average developers. This work’s effectiveness, in terms of generality, will be evaluated through the lens of compatibility, leaving customization to future work.

**Performance and Managing Overheads** Any adaptive optimization system will have some overhead, whether it be the time it takes the optimizer to converge on an optimal solution, or the slowdown due to instrumentation and sampling for dynamic profiling. If the overheads outweigh the performance improvements, the system is either ineffective at managing overheads or finding performance gains.

Suppose a user has a program and one computing resource with which they would like to try search-based adaptive optimization, e.g., autotuning. Applying a traditional autotuner, such as OpenTuner [Ansel et al., 2014], consists of creating and a running tuning phase prior to obtaining and deploying the resulting optimized version for real-world use. Because application-specific tuning is performed during a phase prior to real-world use, the focus of most offline systems is to minimize the tuning time by reducing of the number of evaluations required to converge on a good result. While the total time to execute this process depends on a number of other factors, such as the running time of each program test run and the user’s patience, times on the order of hours to days are not unusual (e.g., Section 4.2.6). Fortunately for offline systems, the cost of tuning is completely separate from the use of its resulting optimized configuration.

For online systems, the training phase is continuous and interleaves with the usage of its results. This makes overhead management a central point of evaluation because of the need to pay-off the overhead of exploration, i.e., breaking-even. Because of this difference, the performance of an online optimization system is harder to evaluate, especially relative to an offline system. The ability to adapt to unforeseen workloads or environments is the unique advantage that an online system offers over an offline one. Thus, the other point of evaluation is the performance improvements seen through latent, on-demand tuning that is specific to the workload or machine.
Chapter 4

Related Work

Adaptive optimization is the general technique of utilizing information obtained during the runtime of the program to modify the program for performance improvement (Section 2.1). Due to the enormous volume of existing work that satisfies such a broad definition, it is not feasible to provide a survey of the area in this dissertation. Instead, the focus of this chapter is on works most relevant to the thesis (Chapter 3). No attempt is made to categorize the related works using an objective or rigid classification system; intuition is used instead.

4.1 By Similarity

This section contains a detailed overview of prior work subjectively believed to be most close to addressing the thesis. Distinctions between the objectives of this work and the prior work will be emphasized as needed.

4.1.1 Active Harmony

The Active Harmony project has gone through a number of revisions and is the most closely related work. In Hollingsworth and Keleher [1999]; Keleher et al. [1999] Harmony was positioned as a global resource management system for the coming distributed-computing boom. The goal was to support online reconfiguration and adaptation of resources to optimize along trade-offs, such as throughput versus latency, for long-running applications like databases. The architecture of Harmony used a client-server model, where the server manages the simple greedy search and adaptation options that are described in a domain-specific language. Clients use the Harmony C API to synchronize...
their configuration with the server.

Later on, the Harmony project became more focused on general automatic tuning for application libraries and parameters in Tăpuş et al. [2002]. The major idea was that many libraries or algorithms implement the same functionality, but some versions are more appropriate for certain situations than others. Notable earlier work by Whaley and Dongarra [1998] in the Automatically Tuned Linear Algebra Software (ATLAS) employed this idea. ATLAS is effectively an “active library” [Veldhuizen and Gannon, 1998] that automatically generates, tests and uses matrix-multiply kernels tailored for the particular CPU it is running on.

In Tăpuş et al. [2002], Harmony would generate an API based on a specification that wraps one or more implementations of that interface. Clients would access the libraries through the wrapper API, which would dynamically monitor the performance of the underlying library implementation and tune both the choice of library and the chosen library’s tuning parameters on-the-fly. The BBO search process for an optimal configuration uses a custom variant of the Nelder-Mead simplex method [Nelder and Mead, 1965] that does not assume the function being minimized is defined or continuous. Chung and Hollingsworth [2004] later greatly improved this search technique by analyzing prior configurations and focused the search by determining which parameters are the most important through the use of parameter sensitivity testing.

**Offline Harmony with CHiLL**  Tiwari et al. [2009a] combined Active Harmony with the CHiLL polyhedral loop transformation system [Chen et al., 2008] to autotune compiler optimizations in an offline setting. The tuning search process was made parallel using the Parallel Rank Ordering algorithm Tabatabaee et al. [2005]; Tiwari et al. [2009b], where multiple clients running the same application and performing the same work connect to the server and experiment with different configurations in parallel as directed by the central server. Essentially, one search for a single optimal configuration is parallelized across multiple machines. They were able to improve the performance of computational kernels by 1.4x–3.6x over the Intel compiler. Tiwari et al. [2011] further evaluated their offline CHiLL-based autotuner from 2009 on a full application called SMG2000. They were able to improve SMG2000’s overall performance by 27% through the tuning of the main kernel, which saw a 2.37x speed-up.

**Online Harmony with CHiLL**  Tiwari and Hollingsworth [2011] describe their extensions to Active Harmony from 2009 to allow for online autotuning with dynamic code
generation and loading. The key new feature is the ability to configure one or more code servers that handle compilation requests using a standalone compiler, with CHiLL used in their evaluation, to produce shared libraries with the new code variant. A system similar to Online Harmony called AARTS was proposed by Teodoro and Sussman [2011].

```c
// MPI Initialization and Harmony API initialization are omitted.
if (masterClient) {
    hdef_t* hdef = ah_def_alloc(); // Create Harmony tuning search.
    ah_def_name(hdef, "gemm"); // server knows the "gemm" kernel.
    ah_def_strategy(hdef, "pro.so"); // request PRO search algorithm
    ah_def_layers(hdef, "codegen.so"); // request codegen

    // initialize name, min, max, and step-size for each parameter.
    // here "Tx" means "tile loop x" and "Ux" means "unroll loop x"
    ah_def_int(hdef, "TI", 2, 500, 2, NULL);
    ah_def_int(hdef, "TJ", 2, 500, 2, NULL);
    ah_def_int(hdef, "TK", 2, 500, 2, NULL);
    ah_def_int(hdef, "UI", 1, 8, 1, NULL);
    ah_def_int(hdef, "UJ", 1, 8, 1, NULL);
    htask = ah_start(hdesc, hdef);
    ah_def_free(hdef);
} else { /* Other nodes join master client’s session */ }
```

Figure 4.1: Client-side setup code for a parallelized online Active Harmony + CHiLL tuning session [Chen, 2019]. The tuned parameter names are known by the server to refer to loop tiling and unrolling for each loop nest.

In order to use Online Harmony for tuning with code generation, the user must first extract the code they would like to tune into a separate library that can be compiled with a standalone tool. The library source code along with the tuning configuration is placed on the server and given a unique name. Clients who connect refer to a library through the session name when initializing the tuning session (Figure 4.1). When the clients begin to execute the tuning loop specified by the user, experimental code is sent by the server and dynamically loaded using `dlopen` and `dlsym` at the point where a new configuration is fetched through the Harmony C API (Figure 4.2). This code’s performance is measured using a testing workload and reported back to the server.

It is important to recognize that Active Harmony is effectively a traditional autotuning system that is accessible through a C API. That is, the user of Harmony must manually synchronize the clients and setup their own tuning loop, which runs for a fixed period of time or until convergence, to test a new configuration on each iteration. The user
must perform their own profiling, both to identify a function that is worth tuning and to evaluate each configuration sent by the server.

Harmony server also does not explicitly tackle the problem of finding balance, i.e., managing the exploration versus exploitation trade-off of online tuning to minimize overheads. The only case where this problem is addressed is when the server is not ready with a new configuration. In this case, an Active Harmony client simply uses the existing configuration instead of blocking to wait for the server.

```c
for (i = 1; i < SEARCH_MAX; ++i) {
    // Retrieve a new point to test from the tuning session.
    // This call modifies tuned variables and may call dlopen, etc.
    fetch_configuration();

    // Execute and measure the client application kernel.
    memset(C, 0, sizeof(C)); // clear output matrix
    time_start = timer();
    code_so(500, A, B, C); // compute C=A*B for the 500x500 matrices
    time_end = timer();

    // Report our performance result to the Harmony server.
    perf = calculate_performance(time_end - time_start);
    ah_report(htask, &perf);

    if (! harmonized) {
        harmonized = check_convergence();
        if (harmonized) {
            // Harmony server has converged. One final fetch
            // to load the harmonized values and disconnect.
            fetch_configuration();
            ah_leave(htask);
            break;
        }
    }
}
```

Figure 4.2: The main loop of an online Active Harmony + CHiLL application that uses the code-server to search for code variants of a naive matrix multiplication implementation for optimally performing configurations [Chen, 2019].
4.1.2 Kistler’s Optimizer

Kistler [1999]’s dissertation\(^1\) describes a general, extensible architecture for online program optimization built on Oberon System 3 for the Macintosh [Gutknecht, 1994; Wirth and Gutknecht, 1989]. Kistler’s system consists of five modular components: a code-generating loader, continuous profiler, manager, optimizer, and replacer. The manager is a low-priority thread that periodically consults the profiling data being gathered, looking for changes in behavior. If the recent profiles of a function are considered different enough according to a similarity metric, the manager requests reoptimization of the function.

If the manager determines that the estimated speedup is not worth the cost of performing reoptimization, then no change is made to that function. Otherwise, a fixed-order sequence of optimization components are applied to the function. Each optimization component consists of a set of one or more optimization passes that perform some compiler optimization (e.g., loop-invariant code motion) plus a profitability analysis that considers code features and the profiling data before deciding to apply the optimization.

Adaptive optimization is exhibited in Kistler’s system from its use of dynamically toggled compiler optimizations based on continuously monitored program behavior. For example, an optimization component may look to see if a certain profiling counter exceeds some threshold in order to be deemed profitable. On the other hand, if an optimization component was applied in a previous version of the function, and new profiling data suggests that the function’s performance became negative or did not change, then the component is marked as not-profitable in a database. Since the database’s information is aged during execution and discarded after the application exits, an optimization component that was disabled may be reenabled, or a component that was skipped may be applied in the future. Thus, there is an air of search-based automatic tuning in Kistler’s system.

It is important to recognize that the optimization components in Kistler’s system toggle themselves independently based on their own cost models, without a central system monitoring the overall performance. If all optimization components were completely independent and isolated from each other, this would not be a problem. I believe this is the reason why the results were not as appetizing as one might hope. Kistler’s evaluation of the system does not discuss total system performance gain or loss for a given benchmark. This is because even in an ideally optimized case, most benchmarks only saw a roughly 5% performance gain, except for one outlier which improved by 125%. An evaluation

of adaptively toggling optimizations was not presented presumably because it was not effective.

In contrast, the newly developed profile-guided optimizations in the dissertation, i.e., object layout and instruction scheduling were effective. Adaptive object layout yielded an average 24% performance improvement, and instruction trace scheduling a 4% improvement, over the baseline of an ideally optimized program.

Additionally, Kistler analyzes break-even points, which are points at which the adaptively optimized program saves enough execution time to overcome the cost of optimizing it (Section 2.5).

4.1.3 Jikes RVM

The Jikes RVM\textsuperscript{2} is an adaptive Java virtual machine featuring JIT compilation [Arnold et al., 2000, 2002]. Jikes RVM periodically considers reoptimizing the hottest methods and uses a simple cost-benefit model based on execution time estimates to decide whether the compilation cost is worth the investment.

The key innovations of Jikes are the low-overhead profiling techniques and continuous recompilation used to drive a five online adaptive optimizations. First, profiling is used to decide whether to promote the code’s optimization level in order to balance compilation costs with code quality. Second, a dynamic call graph is continually updated based on profiling data to identify hot call edges and perform adaptive inlining. Third, code within a method is laid out to prioritize locality based on profiling data. Fourth, loop unrolling is also adapted by either doubling the compiler’s heuristic unrolling threshold, or halving it, based on whether profiling determines the loop is hot or cold respectively. Finally, path profiles are used to focus an optimistic transformation called merge splitting, where the goal is to eliminate control-flow merges within a method via code duplication to aid later optimization and analysis passes such as redundancy elimination [Arnold et al., 2002; Bodík et al., 1998; Chambers and Ungar, 1991]. A major a distinguishing factor is Jikes’s use of profile-guided heuristics to determine what optimizations to dynamically apply to the code, in contrast to the use of empirical search.

Jikes’s dynamic profile-guided inlining yielded improvements of 11% on average and a peak of 73% on their benchmark suite for start-up performance. Instrumentation to generate profile data were added at most 1–2% overhead. Overall, Jikes improved peak steady-state performance by 4.3% on average and up to 16.9% in one case. The top two

\textsuperscript{2}Originally called the Jalapeño Adaptive Optimization System.
most impactful optimizations were profile-guided code layout and merge splitting, which aided redundancy elimination in removing method accessor guards introduced by inlining.

4.1.4 ADAPT

Voss and Eigemann [2001] describe ADAPT, which is an online adaptive optimization system for FORTRAN 77. What sets ADAPT apart from prior systems, such as JVMs featuring JIT compilation like HotSpot and Jikes (Section 4.1.3), is ADAPT’s use of an iterative search process to find a good optimization configuration for simple loop nests. ADAPT identifies a hot loop that executes long enough and begins experimenting with differently optimized versions of that code. In essence, this dissertation is similar to ADAPT in that they both use search-based online tuning, though there are a number of important differences between the two.

ADAPT’s search process is primarily driven by scripts written in a domain-specific language “AL”. These scripts implement heuristics written by compiler developers that decide what to tune and how the tuning should be done for a given loop. For example, their AL script for loop unrolling is a program implementing a linear search of unrolling factors of at most 10 that evenly divide the loop’s bound. While these AL scripts are similar to offline tuning with a shell script, the AL language for these scripts features high-level constructs to simplify the compiler autotuning task. These constructs include the ability to specify a parameter space, constrain the tuning by the results of compiler analyses, and re-run the tuning if the user determines the best version has become stale.

One of the main shortcomings of relying on these scripts is that it offloads the most challenging aspects of online autotuning onto the users: effective search strategies and managing exploration and exploitation. All of the AL scripts used to evaluate ADAPT experiment with at most 10 configurations using linear search, which is a trivial autotuning task. Because the parameter space is incredibly small, the explore-exploit problem also did not need to be addressed.

The overheads of ADAPT are claimed to be at most 5%, but when using their do-nothing AL script to evaluate overheads, some programs actually ran a few percent faster than when not using ADAPT, with an average 1.6% improvement on Linux. The authors state the reason for this is that ADAPT always performs “inter-procedural constant propagation and applies some simplifications that may lead to improved performance” when initially compiling the program and runtime system components, so the true overheads
are unknown.

The performance benefits of using ADAPT for autotuning were notable: average improvements of 35% on the backend flag selection problem and 18% on loop unrolling relative to their baseline on five SPEC2000 floating-point benchmarks. It is important to note that each of these benchmark programs ran for 8 to 70 minutes (average of 21.8 minutes), presumably to give the system time to experiment, but the authors did not specify how they determined the amount of work to be performed for each benchmark.

4.1.5 ADORE

Lu et al. [2004] describe a binary optimizer called ADORE (Adaptive Object code RE-optimization) for speculative online adaptive optimization. ADORE is a uses performance monitoring, sampling-based profiling, and phase-change detection to deliver dynamically optimized cache behavior that improves performance by 3%–106% while incurring 1%–2% overhead on average for the SPEC 2000 benchmark programs considered. Intuitively, a program phase roughly corresponds to contiguous period of time of consistent or “similar” program behavior within that period [Hind et al., 2003].

ADORE is shared library that is linked into a single-threaded executable that is specially compiled to allow for code mutation, namely, the compiler reserves some free registers in the generated code for dynamic modification by ADORE. The dynamic optimization of ADORE is driven by the fixed rate at which sampling data is delivered to ADORE from the CPU’s performance monitoring units (PMUs), which was empirically chosen to be 400,000 cycles/sample. The sampling data from the PMU contains information about data cache misses and recently-taken branches. Later work by Weifeng Zhang et al. [2005] proposes additional hardware extensions that would enhance ADORE with event-driven optimization and better profiling to reduce overheads.

As the program is executing, ADORE periodically mutates hot code to redirect control-flow to a dynamically optimized code “trace,” or piece of code that has a single entry-point but multiple exists. ADORE uses PMU sampling information for two optimizations within a code trace. The first and primary optimization determines which load instruction is the most costly based on its data reference pattern in order to schedule an appropriate data-cache prefetching directive. The second utilizes the recently-taken branch data to build a path profile that is used to layout code within the trace for better instruction-cache locality. Once the code trace is in place, the ADORE system hibernates until a high-level phase change is detected, where it will then reoptimize the code based on the
new sampling data.

There are two aspects of ADORE that put it outside of the class of search-based automatic tuners. First, its process of “searching” a configuration space of knobs is passive, being driven by the phase-changes of the program. Additionally, once a phase-change occurs, cost models are used to determine how and where a prefetching directive should be added, instead of using experimentation. No active effort is made to undo poorly-chosen prefetch directives, leaving it up to the phase detector to adjust the optimizations once the program’s behavior changes.

4.1.6 Suda’s Bayesian Online Autotuner

Suda [2007, 2010] tackles the high-level problems of online autotuning through the lens of a formal, abstract model of the online autotuning process. Empirical autotuning is seen as having two distinct kinds of executions of the program under a given configuration. There is the trial execution, which corresponds to an evaluation of an exploratory configuration (often with sample inputs), and the practice execution that exploits the best-discovered execution. Suda remarks that offline tuning is distinct from online tuning in that it performs all exploration before any exploitation. Suda employs a Bayesian approach to manage the case of online autotuning, i.e., where one must strike a careful balance of exploration and exploitation.

Suda’s proposed Bayesian approach is based on cost models, which can be either analytical or trained via machine learning, that accurately describe the cost function to be minimized. The key limitations of the Bayesian approach are the reliance on a fixed, predetermined number of experiments during the tuning process, and strong dependence on the estimation of the accuracy of cost model. In addition, assumptions are made about the stability and deterministic behavior of the trial execution’s reported cost, namely, that the costs are normally distributed around some mean when the configuration remains fixed.

One of the downsides Suda discovered is that if the cost model is deemed to be very accurate by the Bayesian method, then virtually no exploration will occur, but if the model is determined to be too inaccurate, then it effectively performs all exploration up-front before doing any exploitation. The proposed solution adds the assumption that the cost function is linear. Then, a fixed number of exploration steps $n_{init}$ is chosen to be at least equal to the number of degrees-of-freedom in the linear model. For execution step $i > n_{init}$, the proposed system decides to exploit the best configuration if $i \geq c \log i$, for
some arbitrary $c$, otherwise it will explore a new configuration. This type of decision-making process for whether to explore or exploit is effectively a non-random version of approaches used in multi-armed bandit models.

### 4.1.7 PEAK

Pan and Eigenmann [2008] describe PEAK, an automatic compiler-optimization tuning system that only needs partial-program executions rather than full executions to evaluate configurations. The advantage over whole-program executions is that the time to tune a particular function is greatly reduced, since a function will be called many times during a single program execution. PEAK is designed to perform autotuning using training data in a tuning stage in order to produce a final production-ready version of code, thus it is classified as an offline tuning system.

Nevertheless, from the viewpoint of the implementation techniques used in its pre-tuning and tuning stages, PEAK is similar to an online autotuner. Prior to tuning, PEAK analyzes profiling data to choose a worthwhile function to tune [Pan and Eigenmann, 2006]. During tuning, the tuned function (and its callees) is dynamically switched-out with a differently-optimized versions to evaluate multiple configurations within one execution of the program. PEAK uses a number of practical methods for comparing the performance of these different versions, since raw execution times from two arbitrary time slices of a program’s execution may not have equivalent workloads for the function [Pan and Eigenmann, 2004]. Both of these techniques are solutions to problems faced by online autotuning frameworks, though PEAK only utilized them to make offline autotuning more efficient.

### 4.1.8 PetaBricks

Ansel et al. [2009] developed PetaBricks, an implicitly parallel language that incorporates runtime-tunable values. Tuning is performed for language-level program concepts such as parallelization techniques and algorithm choice, rather than compiler optimizations. SiblingRivalry, PetaBricks’s technique for online performance auditing, partitions resources and races two variants in real-time (through process forking) to determine which configuration is better [Ansel et al., 2012]. PetaBricks’s search procedure uses evolutionary techniques and a multi-armed bandit model to choose the mutation operator to apply to the current best configuration to yield the next one Ansel et al. [2011, 2014]; Fialho et al. [2010]; Maturana et al. [2009].
4.2 By Philosophy

This section contains an overview of a number of works that are spiritually similar to the goals of the proposed research, but are not sufficiently adjacent to the thesis. By spiritually, I mean that these works recognize the importance of leveraging dynamic profile data to improve program performance. The specific works chosen for inclusion in this section either intersect with the thesis in some ways, or are particularly interesting to consider.

4.2.1 Adaptive Fortran

Hansen [1974] created Adaptive Fortran (AF), one of the earliest known works in online adaptive optimization. Motivated by the findings of Knuth [1971] and others, profile information is used to determine when and where extra time should be spent dynamically optimizing the program. Thus, AF primarily reduces the cost of static compilation, optimization, and native code loading time by delaying these tasks so they occur on-demand at runtime. One of the challenges faced by AF was in controlling the rate at which optimizations were applied to the program and sometimes kicks in too early in order to pay off the cost. For example, their heuristic for determining whether to further optimize code, predetermined execution frequency thresholds, worked well for some benchmarks but different thresholds were needed for others.

4.2.2 Dynamic Feedback

Diniz et al. [1997] developed a technique they call “dynamic feedback” for adapting programs to their runtime environment by dynamically selecting among a small number of predetermined versions of code. These versions implement different optimization policies, such as whether two critical sections in the code should be merged or more finely broken down in order to reduce synchronization overhead. What sets this work apart from pure code multi-versioning (Section 2.3) is the continuous adaptation loop that switches between two fixed-time phases. The performance sampling phase explores the effectiveness of each code version under the current process’s environment by testing out each version empirically for a short time. The production phase then exploits the best version determined in the sampling phase to amortize the cost of exploration. Under a number of assumptions, the authors provide an analysis of the worst-case performance bounds of dynamic feedback, which can be used to pick a time for the production phase that guarantees a minimum level of performance.
4.2.3 Dynamo

One of the major influential works in adaptive optimization is the Dynamo project [Bala et al., 2000], an online optimizer for native machine-code binaries. The project’s main vision is that application end-users can adaptively optimize programs without needing access to source code. Dynamo readapts to new program behaviors after reaching a steady state through periodic flushing of their code cache. Good results were achieved on SPEC95’s benchmark suite relative to default optimization with a static compiler: up to 22% improvement in some cases and 9% on average through partial procedure inlining and improved code block layout. But, if the static compiler utilizes profile data to produce the optimized binary (i.e., profile-guided compilation), Dynamo is unable to improve application performance. Recent work by Panchenko et al. [2019] that adaptively optimizes machine-code binaries, primarily through improved code layout, has been able to significantly outperform the the capabilities of static compilation.

4.2.4 CoCo

Childers et al. [2003] proposed the Continuous Compiler (CoCo) framework and performed preliminary experiments with a simulator to gauge the effectiveness of CoCo in adapting loop optimizations. While the philosophy of CoCo is the same as in this work, CoCo fully relies on expert-defined analytical models to predict the impact of an optimization, which are detailed by Zhao et al. [2002]. Thus, hand-crafted models are used instead of any trial-and-error for adaptation. Follow on work by Zhao et al. [2005] extended their analytical models to predict the effectiveness of scalar optimizations, though the work was implemented using the Machine SUIF compiler [Smith and Holloway, 2002].

4.2.5 MATE

Morajko et al. [2007] created the Monitoring, Analysis and Tuning Environment (MATE) to perform dynamic automatic tuning for C/C++ applications. Their particular focus was on programs running on a cluster or supercomputer. For each application, MATE requires the developer to provide information about what can be tuned, where to measure for performance changes, and a performance model. Their performance model is a set of rules that signal if performance has deteriorated, improved, or has reached optimality. Their evaluation of MATE first focused on the scenario of tuning a communication library (similar to MPI [Gropp et al., 1999]) within a space of 8 total options using exhaustive search,
with a baseline running time of 12.2 minutes. The second scenario tuned one floating-point value within \((0, 1]\) that controls the division of work in a parallel application under a variable workload to demonstrate its ability to adapt over the default.

MATE was sometimes able to improve the performance of an application. For example in the first scenario, when tuning a certain binary option by itself, MATE slowed down the overall execution of the program by 5.1%. However, when tuning all three binary options the execution time improved by 27.7%, though it is unclear whether the cause was a lucky ordering of the configuration space searched.

4.2.6 AOS

Hoste et al. [2010]'s Adaptive Optimization System (AOS) is built on top of the Jikes RVM (Section 4.1.3) and is similar to this work in terms of combining automatic tuning and dynamic compilation. The key difference is that AOS is an automatic tuner for a JIT compiler, not one performs online adaptive optimization using a JIT compiler. That is, in a tuning phase prior to the use of the JIT compiler in production, AOS searches for an optimal set of optimizations to be applied to JIT-compiled methods at each optimization tier (e.g., -O0, -O1, -O2). Thus, while AOS itself is an offline tuning system, the object being tuned is an online adaptive optimizer.

Notably, AOS's tuning phase is broken into two stages. The first stage is a search over a compiler optimization configuration space under a scenario where activity such as garbage-collection and concurrent runtime compilation are excluded (e.g., the heap size is set to be very large). The evolutionary search narrows down a space of \(2^{33}\) configurations down to 8 Pareto-optimal configurations across the dimensions of execution-time and compile-time. The second phase then uses another evolutionary search to find an optimal mapping from configurations to optimization tiers, giving rise to 92 possible assignments to explore. This reduced search is performed in a more realistic environment where the previously-excluded overheads are now present. The total time needed to tune the Jikes RVM with AOS across 16 standard Java benchmark programs on a single machine take 550 hours for the first stage and 1,320 hours for the second stage, i.e., a total of 1,870 machine hours or nearly 78 days. Since this offline tuning can be parallelized, they were able to complete the tuning in only 75 hours (3 days) of actual time, which means on the order of 25 machines were needed. Hoste et al. performed an extensive evaluation of AOS and were generally successful in showing that autotuning of a JIT compiler can match manual tuning.
4.2.7 Testarossa

Sanchez et al. [2011] describe an experimental augmentation of the IBM Testarossa JVM that utilizes a support-vector machine learner to produce function-specific compiler optimization plans, which toggle the optimizations to be applied to the JIT-compiled function. The model is trained ahead-of-time to recognize code features of functions and produce a specially-tuned optimization plan. A similar approach was used in Milepost GCC [Fursin et al., 2011] where machine learning was used to predict optimal application-specific plans, but this work focuses on function-specific plans.

Data to train the model is generated through iterated compilation while training programs execute. Both random search and a type of simulated annealing are used to explore the compilation-plan space during data collection. Their experiments showed that their machine-learning predicted optimization plans under-perform in steady-state performance when compared to Testarossa’s default optimization plans. But, compilation overhead (and thus JVM warm-up time) were reduced with the plans produced by the model.

Nuzman et al. [2013] extended the IBM Testarossa JVM with support for C/C++. Their main contribution was to show that adaptive optimization through JIT compilation can be implemented in such a way that the overheads are low enough for use by languages that are normally compiled to machine code. Nuzman et al. provide a detailed overview of their runtime system’s infrastructure and how they avoid introducing overhead. Overall, they were able to demonstrate a 7% total average performance improvement through dynamic optimization of SPECint2006. This work is exploratory in the sense that their dynamic optimizer under-utilizes the information available to it: they only optimize the code based on dynamically-profiled branch probabilities to improve code layout. Nevertheless, it shows that there is room for future work in this space.
Part II

Halo: Wholly Adaptive LLVM Optimizer
Chapter 5

System Overview

The main goal of HALO\(^1\) is to leverage a novel combination of techniques to overcome the challenges of online adaptive optimization and investigate the thesis (Chapter 3). HALO optimizes programs represented in LLVM’s intermediate representation (IR), which is a C-like static single-assignment IR that can be efficiently serialized as “bitcode.”

The LLVM compiler infrastructure is used by major compilers for C, C++, RUST, SWIFT, and other languages to target a variety of hardware architectures such as x86-64 and ARM. Lattner and Adve [2004] originally positioned LLVM as a compiler-based lifelong program optimization framework based on its generality and modularity. LLVM is designed around a common intermediate representation (IR) that is a typed, static single-assignment representation of the program with constructs that mirror a high-level assembly language with unstructured control-flow. In many respects, LLVM IR is similar to a verbose, normalized C language with support for features like exceptions. Thus, prior work in adaptive optimization for C-like languages are should be applicable at the level of LLVM IR with sufficient metadata from the compiler front-end.

Since HALO itself operates solely on a compiler IR, we extend the LLVM-based CLANG compiler for C/C++ with support for generating executables that utilize HALO for adaptive optimization. Users of the system do not need to make any changes to their C or C++ program’s source code: simply add the `-fhalo` flag when compiling with CLANG to produce a HALO executable.\(^2\)

The system is based on a client-server model (Figure 5.1) to isolate the activity of the optimizer from the running application to prevent interference that degrades application performance. A *client* is a running executable that utilizes HALO’s monitor subsystem,

---

\(^1\) An acronym for Wholly Adaptive LLVM Optimizer.

\(^2\) Currently, only Linux systems that use ELF object files are supported.
which is controlled by a server for performance enhancement. After establishing a connection, the most common types of messages sent by the client are perf events samples, with the server tuning each client by sending re-optimized versions of the client’s code. The connection is terminated whenever the client process’s main function returns, which invokes the monitor’s destructor. For the remainder of this chapter, I describe in more detail the three major components that make up the HALO optimization system.

5.1 Clang

The HALO system can optimize unmodified C and C++ programs through an augmented CLANG compiler. By simply providing the -fhalo flag, the CLANG compiler produces executables that can be modified during runtime by the HALO monitor (Section 5.2), even though the monitor does not export a public API. Instead, the monitor has requirements on the format of the executable, which CLANG ensures in the following ways.

First, a snapshot of the unoptimized LLVM bitcode corresponding to the application’s source code is taken to be saved as data in a special section of the executable file. This “fat-binary” approach has been shown to be a low-overhead way of providing dynamic optimization for statically-compiled languages [Nuzman et al., 2013]. Next, early in CLANG’s optimization pipeline, we analyze the program to determine which functions will be made patchable. Functions that are only ever called once, such as those in the .text.startup section or the main function are not patchable. All other functions are marked as patchable unless if they have fewer than 100 LLVM IR instructions, have no loops, and is a leaf function. Finally, the halomon library is linked into the final executable. This library contains a single, static global class definition whose constructor will be run prior to main. The constructor spawns a thread for the monitor and the destructor is run...
when the executable exits to cleanly shutdown the monitor.

5.2 Halo Monitor

The HALO Monitor, also called halomon, is a static C++ library that is linked into every HALO-enabled executable by CLANG (Section 5.1). When the monitor is initially launched (in its own thread), it tries to connect to the HALO server via a TCP/IP connection to receive requests (Figure 5.1). The types of requests sent by the server involve tasks such as profiling, code patching, and dynamic linking.

5.2.1 Instrumentation-based Profiling

The HALO monitor offers a simple instrumentation-based profiling method to provide an estimate of the frequency of calls to a patched function. The instrumentation consists of a call-counter that is incremented each time the client program calls a patched function, specifically, during the function redirection (Section 5.2.3). The monitor thread then records a timestamp every time it takes a snapshot of the latest counts before sending both pieces of information to the server. Based on the difference in time between timestamps and the number of calls that elapsed during that time, a metric of “number of calls per fixed time unit” is made available as a metric for measuring the performance of a tuning section.

One of the unique aspects of this call-rate metric is that it captures the performance of the program with respect to its usage of a function, since the time between calls includes both the time spent inside and outside of the function. A similar and more traditional metric would be the amount of time elapsed to complete a call, which has been used in prior work to measure the performance of the function with respect to its usage by the program [Lau et al., 2006]. The downside of time-per-call is that it requires more instrumentation: on both function entry and exit. To handle recursive calls efficiently, a timestamp could be pushed onto the program’s call-stack instead of maintaining a separate stack of timestamps. Nevertheless, both metrics are still fallible in the sense that they do not control for workload variations in the program.
5.2.2 Sampling-based Profiling

Modern CPUs offer performance-monitoring units (PMUs) that can be sampled to provide useful profile information with low overhead, at the cost of accuracy and consistency [Chen et al., 2013, 2016; Weaver, 2015]. The Linux kernel provides access to the PMU via the `perf.events` API [Corbet, 2009, 2011; Weaver, 2016], which is the primary source from which halommon obtains reports about the performance of the process to the server. To make use of the `perf.events` data, halommon first provides the server with information about the correspondence between code addresses and functions in the process, i.e., the code map. The code map is built using two sources of information. The first piece is the name, size, and offset of every function symbol in the process by reading the executable’s `.text` section. The second piece is the starting address of the `.text` section in the process’s memory map, which can be obtained from the operating system [noa].

Linux `perf.events` provides samples on a periodic basis from the PMUs after a number of CPU events (or a length of time) has been observed. A poorly chosen period length can cause the sampling to become “synchronized” with the code being executed, which yields biased samples [Chen et al., 2013]. We use the event “number of instructions retired” and set the period length to a large prime number (e.g. 15,485,867). This way, the number of instructions within a loop body is less likely to be a divisor of the period length. The size of the period directly influences the overhead of `perf.events` sampling, and we hand-picked the length period to ensure that the overhead is less than one percent. Through `perf.events`, the halommon library is configured to save the following information from each sample, where a sample’s information corresponds to the state of the PMUs when the event was triggered.

- `ip` — A pointer to the instruction that triggered the event. This is often not exact due to sampling skid, which is the number of instructions that executed during the time it took for the kernel to respond to the CPU’s event notification to stop and sample the PMUs.

- `thread_id` — The thread ID (according to the operating system) that triggered the event. Each sample’s information is specific to the particular thread.

- `time` — A nanosecond timestamp that indicates when the event occurred, relative to some fixed point in time.

- `call_context` — A stack of return addresses that are currently on the call stack.
When the event is triggered, the kernel will walk the thread’s call stack (by following frame-pointers) to a bounded depth to obtain this information.

- branches — An array of information about the most recently retired branch instructions. Modern PMUs can be configured to maintain a ring-buffer, called the last-branch record, that records metadata about branch instructions: its source address, destination address, and whether the branch was mispredicted [Chen et al., 2013; Kleen, 2016]. The size and types of branches recorded in the array are hardware dependent. For example, on Intel’s Haswell, we can specify that only call and return instructions be recorded, but on Intel Ivybridge, all branches are included.

5.2.3 Code Patching

The halomon library uses live code patching to dynamically redirect control-flow in the running process to new code sent by the optimization server. The only points at which code can be redirected are specially-compiled LLVM IR functions that were marked for patching. Dynamic code replacement is relatively straightforward for code that is interpreted by the runtime system Bala et al. [2000]; Gal et al. [2009]. Equivalent mechanisms for native code executing directly on the hardware have also been previously developed for C/C++ Nuzman et al. [2013]. The specific mechanism used to perform patching in HALO is based on the XRay instrumentation infrastructure in LLVM.

Berris et al. [2016] developed XRay as a lightweight function tracing system for LLVM. XRay was designed to facilitate the instrumentation of large-scale systems, such as those used at Google, to track down failures or performance regressions in languages like C or C++. A key goal of XRay is that it does not add any noticeable overhead when instrumentation is disabled, since it is meant to be used for programs compiled to run in production. XRay is fundamentally quite simple: functions marked for XRay instrumentation are compiled such that at the beginning of the function each function’s entry-point and all of its exits, a short-jump followed by empty space (filled with no-op instructions) is inserted (Figure 5.2).

The space between the jump and the beginning of the function allows another thread to redirect control-flow to an XRay instrumentation routine by overwriting the first few instructions of the function. The instrumentation routine receives an ID that identifies the function that was redirected (Figure 5.3).

HALO extends XRay’s patching mechanism with an additional “instrumentation routine” that actually performs a full redirection of control-flow from the original function,
Figure 5.2: The entry-point of patchable function, in an unpatched state.

```assembly
__tunedFunction:
    jmp     funcBody
    nopw   0x200(%rax,%rax,1) ; multi-byte no-op
funcBody:
    ; ... body of function ...
```

Figure 5.3: A patchable function that was dynamically redirected.

```assembly
__tunedFunction:
    mov     $<function id>, %r10d
    call    __functionRedirection
funcBody:
    ; ... body of function ...
```

to a dynamically generated version instead. The redirection routine reads a pointer to the dynamically loaded function pointer from an array (Figure 5.4). The array is used so that the monitor can cheaply change the redirection without performing an atomic write to the memory pages containing the original function’s entry-point.

A mechanism such as on-stack replacement (OSR) would be needed in order to patch in code snippets within a function, such as a new loop body [D’Elia and Demetrescu, 2016]. OSR is a mechanism for dynamically redirecting control-flow at a certain point within a function to another equivalent point in different version of that function. The advantage of OSR over function-call replacement is that we can more quickly optimize a function’s very long-running loops, instead of just the loop’s callees, since this type of function is invoked infrequently but dominates the running time. Mosaner et al. [2019] recently developed simple techniques for performing OSR in LLVM via loop extraction and were able to improve warm-up time without significantly diminishing performance. But, Fink and Feng Qian [2003] found that OSR in an online adaptive optimization system primarily benefits debugging or optimizing pathological code. For performance improvements, they suggest investing effort in the actual code optimizations being adapted instead of the granularity at which they can be done. Thus, HALO replaces code at the function-level granularity, instead of using OSR.
5.2.4 Dynamic Linking

Each object file sent from the server to a client represents a portion of the program that was recompiled with a new tuning configuration. These object files contain one exposed function symbol that represents the entry-point into that version’s code. The client uses standard dlopen-style dynamic linking to load the object files as a dynamic library, or “dylib” in the process. To facilitate this linking, the original executable (which itself is just an object file) is compiled such that any global symbols, such as variables and other patchable functions, are exposed for dlopen to access. This way, when linking a new object file into the process, references to mutable globals (e.g. static local variables in C functions) refer to the version established when the process first launched, instead of creating a dylib-private version.

Currently, HALO does not try to garbage collect any dylibs that might be unused, since it is quite difficult to know when the code is truly dead without further coordination with the running process. One possible solution is to have the monitor fork a child process that invokes Linux’s ptrace, or a “process trace” on its parent, providing the child with debugger-like control to pause all threads and inspect their state. The child can then walk each thread’s stack and look for any values that might be addresses pointing to dylibs, so

---

```assembly
__functionRedirection:
  movq  _ZN6__xray17XRayRedirectTableE(%rip), %r11

; set r11 to the offset of this function's table entry
shlq  $4, %r10      ; r10 contains the function id
addq  %r10, %r11

movq  (%r11), %r10  ; load the function pointer into r10
incq  8(%r11)       ; increment the call counter

; if the function pointer is zero, go back to original
  testq %r10, %r10
  je noRedirect

popq  %r11          ; adjust stack to return to the function's caller
jmq   *%r10         ; call the dynamically-loaded function

noRedirect:
  retq
```

Figure 5.4: The function redirection routine.
that the parent can then free the unreferenced code later. But, this is not a good solution if any of the functions in the dylib have their address taken as a value, since the heap would need to be scanned too.

5.3 Halo Server

The HALO server performs the reoptimization of the code being executed by connected client processes. The major structures and overall flow of information within the HALO server are illustrated in Figure 5.5. Client processes are grouped together into client groups based on the compatibility of their embedded bitcode and CPU architecture, both of which are sent by clients during registration with the server. The optimization server drives the adaptive optimization of client groups independently, based on profile data periodically
Figure 5.6: An example of two JSON-formatted knob specifications used by HALO server.

sent by the clients of each group. Within each group are one or more tuning sections, which are subsets of the code running on the client that have been selected for reoptimization. Specifically, a tuning section is a pair consisting of a patchable “root” function and a set of functions reachable from that root, according to the program’s call-graph. The root function is unique among all active tuning sections in the group, so the root function distinguishes a tuning section.

The server uses a JSON-formatted file that specifies all of the server’s settings and defines the space of options to be searched during tuning. Figure 5.6 provides an example of specifications for two tunable parameters, or knobs, that control HALO server’s compiler. I will refer to the entire space of compiler options available to be tuned as a configuration space, which is equivalent to the cross-product of the sets containing all possible settings for each knob specified in the file. Additionally, a single element in the configuration space is called a configuration, which describes a particular setting of every knob (Section 2).

The values for the name and kind fields of JSON-formatted knob specifications (Figure 5.6) must be already known to the server, because they must be baked into specific parts of the server’s compiler. The other fields are free for users to configure in order to constrain or broaden the search space. Because all integer-based knobs are required to form a contiguous range of values, these knobs offer a scale field specifying the prior scaling that was applied to the given range. In Figure 5.6, the inline-threshold-default knob has a $[0,30]$ range with 1/100 scaling, so the unscaled values form the set of integers $\{0,100,200,\ldots\}$. In other words, the compiler will unscale the value by applying the scale factor’s inverse to determine its true value. Scaling allows the user keep the tuner’s search space small when probing a large space of possible values for a compiler option.
The “Search Method” of each tuning section consists of infrastructure to generate configurations and create parallel compilation jobs (Chapter 7). One a compilation job completes an object file that is ready to be sent to clients is produced. The object file, along with other metadata such as the configurations that produce that object file, are kept together and referred to as a library of the tuning section. This name is consistent with the idea that once an object file is sent to a client, it is loaded dynamically like an ordinary library (Section 5.2.4).

5.3.1 Calling-Context Tree

Since each client group can have multiple clients providing sampling-based profile data, HALO server uses a calling-context tree (CCT) to combine and manage these samples in each group’s profiler. Ammons et al. [1997] originally proposed the CCT as an alternative to the more commonly used call-graph to manage profile data. A call-graph (Figure 5.7a) is a graph that represents the control-flow of the program at a function-level. So, each vertex in a call-graph represents a function and an edge $A \rightarrow B$ represents the relation “$A$ contains a call to $B$.”

One of the downsides of using a call-graph to store performance metrics is that its structure does not keep track of the calling-context of the performance metric. A calling context is a dynamic sequence of functions that have been called at a given point during execution, i.e., a stack of callers for the current function. Consider two performance samples $s_1$ and $s_2$ measured while in function $B$, where the context for $s_1$ is $A \rightarrow B$ and for $s_2$ is $A \rightarrow C \rightarrow B$. When using the call-graph in Figure 5.7a to store this information, we are forced to aggregate these two samples in the one vertex for $B$, which has two unique callers who may use $B$ in vastly different ways.

The calling-context tree (Figure 5.7b) offers finer granularity for tracking samples, at the cost of more space. Specifically, the samples $s_1$ and $s_2$ end up at distinct vertices for $B$ in a CCT. Insertion of a sample into a CCT happens by walking down the tree (rooted at $A$) while following the calling-context associated with the sample. The CCT is built on-demand as contexts are observed, so nodes are created as-needed during a walk of the calling context associated with a sample. A calling-context may contain repeated functions due to recursion, so the CCT is not strictly a tree. Nevertheless, the only types of non-tree edges a CCT will contain are back-edges [Ammons et al., 1997], which are marked with a dashed arrow in Figure 5.7b.
Processing a Sample’s Calling-Context The profile data generated by halomon is well-suited, in theory, for a calling-context tree because the samples contain call_context metadata (Section 5.2.2). But in practice, the metadata is sometimes broken or incomplete for at least two reasons.

The first reason is that sampling happens asynchronously with the running program, so the call stack may be in an indeterminate state at the point where the sample was collected, resulting in a broken calling-context. The second reason is owed to tail-call optimization, which is a compiler optimization to reduce call-stack allocation in control-flow paths that end by returning the value of a function call. The optimization pops the stack-frame of the caller before making a call to a function; this way the callee will return directly to the caller’s caller.

For example, suppose we have a path in the call-graph $F \rightarrow G \rightarrow H$, but $G$ will simply return the value returned by $H$. Then, with tail-call optimization the value returned by $H$ will, dynamically, go directly back to $F$. So, when a \texttt{perf_events} sample is taken in $H$, the calling-context will be \ldots $\rightarrow F \rightarrow H$. According to the program’s call-graph, $F$ does not call $H$ directly, so the calling-context is incomplete and we become stuck at $F$ during the walk to insert the sample.

To resolve a missing function during a calling-context walk, we first search for a shortest path in the CCT from the current, contextually-sensitive node $F$ to a node corresponding to function $H$. If no such path in the CCT exists, then we check the call-graph to see if there is only one path from $F$ to $H$ and create fresh CCT nodes for those intermediate functions.
**Recording Performance Metrics**  After identifying the contextually appropriate CCT node by walking the calling-context, we update the node’s hotness and instructions-per-cycle (IPC) according to the details within the sample. Hotness is a metric that provides an abstract measure of where time is being spent in the program. A CCT node’s hotness is incremented whenever a calling-context walk ends at that node, i.e., the ip field of the sample is within the function (Section 5.2.2). To compute an IPC, we compute the nanoseconds that have elapsed since the last sample was observed at the CCT node (based on each sample’s time field) and divide the sample period (i.e., the fixed number of instructions between samples) by that time difference. For simplicity, the IPC measure remains scaled for a 1 gigahertz clock-rate instead of trying to account for frequency scaling. We use a standard incremental update formula [Sutton and Barto, 1998] to maintain estimated averages for the performance metrics at each CCT node:

\[
NewEstimate \leftarrow OldEstimate + StepSize \times (Observation - OldEstimate)
\]  (5.1)

The \texttt{StepSize} \in (0, 1] is a parameter that controls how much weight is placed on newer observations. For the hotness metric, we only use (5.1) to decay the hotness value by recording a zero-temperature observation at every CCT node at the start of every time-step for the group. Additionally, performance metrics are recorded in each CCT node in two ways: generally (across all libraries) and separately for each library, depending on the library to which the sample belongs. Recording metrics per-library enables more accurate performance comparisons between libraries.

**Leveraging the Last Branch Record**  Samples sent by haomon contain metadata about the most recently executed branch instructions (Section 5.2.2). After walking the sample’s calling-context to reach the context-sensitive node for the sampled function, we walk the last-branch record backwards, from newest to oldest branch, going from each branch’s target to its source. Each time we identify a call, i.e., a branch who’s target is to the beginning of a function, we move up the CCT from the current node to its ancestor (moving backwards in time) and increment the call-hotness of the edge we followed upwards. If the branch goes across two different functions, then a return occurred, so we move down the CCT from the current node to the function that was returned from, creating a new CCT node as-needed. The remaining kind of branches are those that stay within the same function, such as for loops.
\( H_i = \text{sum of hotness in sub-tree } i \)
\[
f_i(V) = \frac{V_{hot}}{H_i} \times (V_{ipc})^{-1}
\]
\[
IPC_1 = (f_1(B1) + f_1(E1) + f_1(A))^{-1}
\]
\[
\approx 4.35
\]
\[
IPC_2 = (f_2(B2) + f_2(E2) + f_2(A))^{-1}
\]
\[
\approx 5.02
\]
\[
H_{tot} = H_1 + H_2
\]
\[
g_i = \frac{H_i}{H_{tot}} \times (IPC_i)^{-1}
\]
\[
IPC_{tot} = (g_1 + g_2)^{-1}
\]
\[
\approx 4.38
\]

(b) Computing \( IPC_{tot} \) for the tuning section using repeated, weighted harmonic means.

Figure 5.8: Computing the total IPC of the tuning section \( \{B, A, E\} \).

In all three cases, the CCT node we end at after processing each branch has its hotness boosted and its IPC updated (if not already updated for the sample). As with the calling-context data, the branches metadata is sometimes incomplete or broken. In such cases where is it not obvious how to continue walking the records, we simply stop early.

**Computing a Tuning Section’s IPC** One of the primary reasons for gathering profiling data is to facilitate performance comparisons. Because context-sensitive versions of a function in the CCT are assigned their own average IPC, and each tuning section can consist of more than one function, a total IPC for a group of functions is needed to make performance comparisons. Specifically, due to the nature of how code patching works in halomon (Section 5.2.3), the relevant nodes in the CCT are all those that are in the tuning section and are reachable from any CCT node representing the root function of the tuning section. To deal with cycles in the CCT, we exclude a node if it represents a function that has already been included during the depth-first search to identify reachable nodes.

Consider the example CCT in Figure 5.8a that is used to manage the profiling data for tuning section \( \{B, A, E\} \), where \( B \) is the root of the tuning section. In Figure 5.8a, CCT
nodes representing a function in the tuning section are shaded with gray (root-functions also use a square) and display their hotness and IPC averages. There are two sub-trees corresponding to the tuning section: \{B_1, A, E_1\} and \{B_2, A, E_2\}. To compute the total IPC for this tuning section (IPC\textsubscript{tot}), we first compute IPC\textsubscript{1} and IPC\textsubscript{2} for each sub-tree, respectively, using a hotness-weighted harmonic mean of the IPCs for each function (Figure 5.8b). An IPC is a rate-of-time measure where the length of time spent in each function is not equal, so we use a harmonic mean to provide a more accurate average IPC than an arithmetic mean [Ferger, 1931]. Using a hotness-weighting further emphasizes the functions that should be the focus of the average measure during a particular time-step, since only the hotness (and not the IPC) decays as time passes. Once we have the IPCs for every sub-tree, to compute IPC\textsubscript{tot} we again use a hotness-weighted harmonic mean, but of the sub-tree IPCs for the tuning section.

5.3.2 Tuning Section Selection

One of the challenges of automatic procedure-level optimization is the selection of procedures that are worth optimizing. For traditional JIT compilation systems, usually a single function or small code fragment is chosen at for reoptimization, with priority given to code contributing to the most execution time in order to pay-off the cost of reoptimization.

For adaptive optimization systems like HALO (Chapter 4), the code where most of the execution time is spent is just as important as in ordinary JIT compilation systems. If the focus of HALO’s tuning effort is not spent on the most-executed parts of code, overheads would take longer to pay-off through performance gains for the tuned code. Because the code may frequently be replaced by HALO with differently-optimized versions, whether for performance comparisons or other adaptation, the frequency at which the function is invoked is uniquely important. Additionally, the tuning section may need to span multiple functions to be effectively tuned, unlike traditional JIT systems which choose the code fragment specifically because there are known opportunities for improving it.

Pan and Eigenmann [2006] recognized this problem of tuning section selection, which is the challenge of automatically selecting a subset of the program to be tuned based on profiling data. In their system called PEAK, a call-graph is built using profiling data from a complete execution of the program on a test input. The edges of the call-graph are annotated with call-frequency and execution time. Using a two-step algorithm based on maximal edge-cuts, PEAK selects tuning sections to maximize execution-time coverage.
while ensuring that each tuning section’s root is called often enough to be dynamically tuned.

In contrast, HALO uses a calling-context tree (Section 5.3.1) and performs online tuning section selection. First, the hottest CCT node that is not already part of another tuning section is used as the starting-point in the search for a tuning root. Starting from the current node, we look at the node’s contextually-sensitive parent node\(^3\) to determine whether the current node should be chosen as a tuning root. The reason for inspecting the parent of the current node is that its immediate parent must be repeatedly making calls to the node in order for halomon’s redirection of the function to take hold. Just like PEAK, the goal is to avoid the situation where the tuning section spends a long time executing inside of the tuning section without the root being called often enough.

Figure 5.9 contains an example of an ancestor chain that is climbed to determine a tuning section’s root. Suppose the node \(D\) (which represents function \(D\) in a specific calling context) is the hottest node in the CCT. When no tuning root has been chosen yet, the current node must simply be patchable (Section 5.1) to become chosen as the tuning root. In this example, node \(D\) is not patchable as indicated by the dashed outline for the node. Thus, we climb up the ancestor chain until we reach the first node that is patchable, \(C\) in our example, and initially choose that node as the tuning root.

To expand the scope of the resulting tuning section, we greedily try to find a better tuning root higher up the ancestor chain, based on profiling data and static code features. Using the node names in Figure 5.9, I will explain how this scope expansion works. We consider the parent \(B\) of the currently-chosen tuning root \(C\) as a candidate tuning root if it is a patchable function. Then, we inspect the grandparent \(A\) of the currently-chosen tuning root \(C\), \(i.e.,\) the parent of the candidate tuning root. At least one of these safety conditions must be satisfied for the candidate \(B\) to become the new chosen tuning root:

1. The grandparent node \(A\) has a hotness that is above a small threshold.

2. The call-edge from the grandparent \(A\) to the parent \(B\) has a non-zero call-frequency.

\(^3\)Ignoring back-edges, which introduce a false parent.
3. One of the ancestors of the candidate $B$ calls its child only from call-sites that are within natural loops.

If the candidate is chosen as the new tuning root, the greedy expansion continues. In our running example, $B$ would be chosen as the final tuning section root because $A$ satisfies the third condition above, as indicated by the allLp metadata.

All three of these safety conditions are designed to ensure that the parent of the final tuning root calls the root function often enough such that tuning can proceed. The third condition above leverages static analysis of the program to infer safety when profiling data is lacking. A lack of profiling data for the grandparent can happen, for example, when the grandparent contains a simple loop that calls the parent function that takes milliseconds to complete for each call. Because the CCT is built using sampling-based profiling, the grandparent’s repeated calls to the moderately long-running parent will not be detected in the profile data. The call-counts provided by instrumentation-based profiling in halomon (Section 5.2.1) could fill this hole in the profiling data, but widespread instrumentation would be costly. Specifically, call-counts are only available for functions that are patched, so all functions would need to be patched (with no redirection) to gather call-counts. Thus, we opted to use static analysis instead of introducing overhead when selecting a tuning root.

Finally, once a tuning root has been selected, our prototype of HALO creates a tuning section that consists of all functions that are reachable from the root, according to the program’s call-graph. Using all reachable functions has the benefit of maximizing coverage of the most-used parts of the program, but this strategy may include irrelevant, less-frequently executed code that bloats the tuning section. During tuning, each library starts with a copy of all the code in the tuning section before compilation. Thus, the all-reachable strategy is not feasible for larger real-world programs, because the server will require more compilation-time, and large object files can bloat client processes with too much cold code. We leave the specific strategy for reducing the size of the tuning section while balancing program coverage to future work.

### 5.3.3 Tuning Section Managers

Each tuning section’s manager is a state machine that is driven by a timer that triggers a step every 100 milliseconds. Because the manager is driven by data from the profiler, some waiting in the manager is unavoidable. Specifically, the profile data is being collected from live client processes, which send their data piecemeal on a periodic basis (on
the order of tens of milliseconds between messages). To maintain uniformity with how other states operate, a poll-and-wait system (where the wait time is equal to the time between steps) was chosen instead of a blocking manager when there is insufficient new profile data. For example, in some states where profile data is not needed, the manager may want to give the client time to exploit the best-optimized version of code. In other states, the manager is waiting for one of many parallel compilation jobs to finish, so poll-and-wait is a simple solution that works for all of these scenarios.

To give a flavor for these state machines, let us consider a simple strategy for optimizing a tuning section that mirrors a standard JIT compiler: recompile the tuning section only once, at the highest default optimization level. This \textit{jit-once strategy} is represented in HALO with the tuning section manager in Figure 5.10. Vertices in the graph represent possible machine states and edges represent actions that transition the machine from state to state. Some transition must be taken by the manager on every step, as driven by the timer. The transitions available depend on the manager’s current state and the decision procedure for that state. For the jit-once state machine, we omit the decision decision procedure for the \textbf{Start} and \textbf{Deployed} states, since only one transition is available once in those states. The \textbf{Compiling} state has a simple decision procedure (Algorithm 5.1).

```
if any compilation job is ready then
    // send the library to all clients
    transitionTo Deployed
else
    transitionTo Compiling
end
```

\textbf{Algorithm 5.1:} The decision procedure for the Compiling state of the jit-once manager.

Because clients can connect or disconnect from the group at any time, at the beginning of all decision procedures, we check the currently-connected clients and synchronize them as-needed. For example, if a client has joined the group late, then that client’s version of the tuning section’s root has not been patched yet. So, that specific client’s state is
synchronized with the tuning section’s manager by re-deploying that library (i.e., sending the object file and a patching command), before invoking the decision procedure for the current state.
Chapter 6

Adaptive Recompilation

The Adaptive tuning manager is the core piece of the HALO server that performs search-based adaptive recompilation. Each tuning section is assigned its own independent instance of the manager so that its decisions are specific to the code being optimized. As with all tuning managers (Section 5.3.3), the Adaptive manager is described by a state machine, illustrated in Figure 6.1, that is driven forward at discrete time-steps.

6.1 Finding Balance

Central to the manager is the Decide state (Figure 6.1) that can choose among three actions: (1) experiment with an unseen knob configuration, (2) retry an experiment using an existing knob configuration, or (3) pause to exploit the best library found so far. All of these actions play an important part of an online search within a non-stationary search space of knob configurations. Good decision-making in the manager is learned through the results, or rewards, of the actions taken. The Adaptive tuning manager’s Decide state helps regulate the overhead of search and adapts the optimization of the tuning section over time.

The multi-armed bandit (MAB) is a learning problem where, at each at discrete time-step, one must select among $k$ actions that give a reward chosen from an unknown probability distribution [Sutton and Barto, 1998]. A “one-armed bandit” is another term for a slot machine that is used for gambling [Oxford English Dictionary]. The objective is to maximize the total rewards as one “gambles” by choosing among multiple actions (or slot machine levers) that may give low or negative rewards in the short-term.

In HALO, the Adaptive tuning manager’s Decide state is viewed as an instance of
the multi-armed bandit problem, where the actions correspond to transitions out of the **Decide** state (Figure 6.1). During any transition back to the **Decide** state, a reward is assigned to the outcome of the last action taken by the tuning manager\(^1\). Specifically, the rewards correspond to whether the outcome is good or not, with higher rewards given to actions that led to a positive outcome.

Solutions to the MAB problem focus on techniques that estimate the value of each available action through trial-and-error, using the estimates to make better decisions over time. In classic formulations of the MAB problem, the probability distributions for each lever are stationary, \textit{i.e.}, the reward probabilities do not change over time. Thus, the value of each action may be estimated to be equal to the average of the total rewards received over time. In HALO, the rewards can change over time because the program’s workload may fluctuate. For our non-stationary MAB problem, we use a weighted average that emphasizes the rewards of actions tried recently.

Suppose we have a map \( E : \text{Action} \rightarrow \text{Expectation} \) that maintains the expected value of taking an action. We define a function to update our expectation mapping \( E \) based the reward \( r \) for an action \( a \) as follows:

\[
\text{update}(E, a, r) = E \pm \{ a \mapsto \text{increment}(E, a, r) \}
\]

\[
\text{increment}(E, a, r) = E[a] + \gamma (r - E[a])
\]

where \( \gamma \in (0, 1] \) is a constant step size, or learning rate, parameter (chosen empirically to be 0.1) and the \( \pm \) operator overwrites a key-value pair in the map. The update

\(^1\)The state machine has persistent memory of the most recent transition out of the **Decide** state.
rule in 6.3, which is the same as rule 5.1 from Section 5.3.1, incrementally calculates an exponential recency-weighted average reward for the action [Sutton and Barto, 1998].

**Decision-making** Knowing the expected rewards for each action serves as a guide for future decisions. Whenever it comes time to make a decision, we could simply pick the action with the highest reward, i.e., make the greedy choice. The problem with greedy decision-making is that it can suffer from tunnel vision. More precisely, a full-greedy strategy results in poor decision-making, because of inherent inaccuracies in our estimate of the average reward of an action [Sutton and Barto, 1998]. This is especially an issue for non-stationary multi-armed bandit problems because knowledge about the expected rewards becomes out-of-date, so some exploration is required. One common alternatives is the $\epsilon$-greedy strategy that picks the action with highest expected reward with probability $1 - \epsilon$ (breaking ties arbitrarily). Typically $\epsilon$ is small (e.g., 0.05 or 0.15) so that we are still mostly greedy, but will also explore other actions; we empirically chose a fixed $\epsilon = 0.1$ for HALO. When the greedy action is not selected, with happens with with probability $\epsilon$, we explore by choosing among all actions uniformly at random.

**Pausing** Tying everything together, let us consider in detail how the HALO server decides to and performs a pause. Pausing allows the each client the utilize the best-known library for the tuning section for a fixed period of time, without any additional overhead due to sampling-based profiling or other activity. During a step in the Decide state (Figure 6.1), suppose we are about to choose an action (i.e., the next state to transition to) with respect to the following expected rewards:

\[
E[\text{Experiment}] = -0.28 \\
E[\text{Retry}] = -0.047 \\
E[\text{Pause}] = 0
\]  

(6.2)

After picking a real number uniformly at random within $[0, 1]$, suppose that number is less than $1 - \epsilon$. This implies the greedy strategy of choosing the action with the largest expected reward, i.e., that we should pause. After transitioning to the Pause state, the tuning manager stays in that state for a fixed number of steps. During that time, the server only communicates with the clients when they first connect to ensure that they have the correct library patched into the process. Then after pausing for long enough, we
transition from \textbf{Pause} to \textbf{Decide} while carrying a reward of 0. Section 6.2 explains why the \textbf{Pause} action’s reward is 0 and discusses the rewards given after other actions. Before choosing the next action in the \textbf{Decide} step, the action-value map is overwritten:

$$E = \text{update}(E, \text{Pause}, 0)$$

$$= E \pm \{ \text{Pause} \mapsto E[\text{Pause}] + \gamma(0 - E[\text{Pause}]) \} \quad (6.3)$$

using the update function from Equation 6.3. Finally, we have completed one cycle and are now ready to start with the next decision in the \textbf{Decide} state.

## 6.2 Bakeoffs

Accurately comparing the quality of two different libraries\textsuperscript{2} is a major challenge in an online setting because we do not have control over what the program does, only how it is done. Specifically, an online system cannot re-run two versions of the code under the exact-same program state [Lau et al., 2006; Pan and Eigenmann, 2004]. It is difficult to re-run a new version of code for comparison with a prior version because the code may depend on the state of global memory, modify its arguments, perform other observable effects such as IO, or any combination of those. In particular, the program may be applying the function to different inputs on each call. Thus, if one were to record and compute the mean running-time of individual function invocations, the variance of the mean is significantly attributable to the function simply doing more work, \textit{i.e.}, the variance is dependent on the function’s inputs. This makes it difficult to make confident conclusions based on the mean running-time alone.

Lau et al. [2006] recognized this problem and used an analysis of the variance of running-times to compute the probability that the difference between the means is significant. The running-times are collected in a live contest between two code versions called a \textit{bakeoff}. A bakeoff is a contest that determines which version of code is better at the particular time the contest is held. Bakeoffs proceed by randomly alternating the version of a function that the program uses between the two contestants while recording how well each performs.

In HALO, a bakeoff begins after entering the \textbf{Bakeoff} state in Figure 6.1. The decision

\textsuperscript{2}Recall that a \textit{library} is a version of the tuning section optimized and compiled to an object file according to a specific configuration of the knobs being tuned (Section 5.3).
if first invocation of the bakeoff decision procedure then
\[
\begin{align*}
\text{CLEAR} & \text{OBSERVATIONS(} \text{deployedLib} \text{)} \\
\text{CLEAR} & \text{OBSERVATIONS(} \text{otherLib} \text{)}
\end{align*}
\]
end

\[
\begin{align*}
deployedObs & \leftarrow \text{GET} \text{OBSERVATIONS(} \text{deployedLib} \text{)} \\
otherObs & \leftarrow \text{GET} \text{OBSERVATIONS(} \text{otherLib} \text{)} \\
\text{comparisonResult} & \leftarrow \text{COMPARE} \text{WITH} \text{STUDENTS} \text{T}(\text{deployedObs}, \text{otherObs})
\end{align*}
\]

switch \text{comparisonResult} do
\[
\begin{align*}
\text{case GreaterThan do } & \text{ transitionTo Payback} \\
\text{case LessThan do} & \\
& \begin{align*}
\text{swap} & \text{ deployedLib with otherLib} \\
\text{transitionTo} & \text{ Payback}
\end{align*} \\
\text{case NoAnswer do} & \\
& \begin{align*}
& \text{if deployedObs has no fresh observations then} \\
& \begin{align*}
& \text{transitionTo Bakeoff}
\end{align*}
\end{align*} \\
& \begin{align*}
& \text{StepsUntilSwap} \leftarrow \text{StepsUntilSwap} - 1 \\
& \text{if StepsUntilSwap} = 0 \text{ then} \\
& \begin{align*}
& \text{if Swaps} \geq \text{MAX then} \text{ the bakeoff has timed out} \\
& \begin{align*}
& \text{deploy bestLib} \\
& \text{transitionTo Payback}
\end{align*}
\end{align*} \\
& \text{swap deployedLib with otherLib} \\
& \text{Swaps} \leftarrow \text{Swaps} + 1 \\
& \text{reset StepsUntilSwap}
\end{align*}
\end{align*}
\]
transitionTo Bakeoff
end
end

Algorithm 6.1: The decision procedure for the Bakeoff state of the Adaptive tuning manager.
procedure for the Bakeoff state is given in Algorithm 6.1. During the transition into the Bakeoff from either Retry or Compile, a candidate library is specified to compete with the current-best library that is being used by all clients. Thus, initially the deployedLib is set to the best library and the otherLib is the candidate. The bakeoff proceeds by assuming all current and future clients during the contest are homogeneous. In the decision procedure, we swap the library used by the clients after a fixed number of time-steps by sending patching commands (and the object file if needed). A time step is not counted during the bakeoff if the deployedLib has insufficient fresh profile data, which is new data attributed to that library that has come in from a client since the last check.

6.2.1 Contest Rules

To determine whether one library is better than another, HALO uses an enhanced version of Lau et al.’s statistical analysis and measurement technique. Instead of measuring running-times, HALO relies on indirect quality metrics to rate different versions of a tuning section. Indirect metrics are used to avoid overhead as much as possible. During development the highest-overhead method, sampling-based profiling, incurred less than 5% overhead when enabled. The two quality metrics available are (1) the average number of instructions retired per CPU cycle (also referred to as an IPC), and (2) the average number of calls to the tuning section’s root function within a fixed time period (also referred to as the call frequency). Under both metrics, we consider a higher value to mean better quality, but sometimes better quality does not result in better performance due to the indirect nature of measurement.

For example, a higher call frequency may indicate that the program is performing better, since it is able to call the function more often. But, a sudden rise in call frequency may have occurred simply because the function’s inputs or program workload have changed during the bakeoff. A higher IPC is also an indicator the code is performing better, since it is able to retire more instructions in the same amount of time while executing in the tuning section. The IPC metric is less dependent on the function’s input than a running-time. For example, an IPC can remain steady if a loop’s trip-count depends on an input to the function, while the call frequency or execution time will always be a function of the trip-count. However, an IPC still maintains some dependence on the data being processed by the function. Specifically, each control-flow path within a loop can have different IPCs, and even the same CPU instruction can require more cycles to complete, depending on its inputs and the state of the cache [Fog et al., 2011].
Comparisons  Let us consider COMPAREWITHSTUDENTST from Algorithm 6.1, which is used to compare two libraries and determine if one is better. Suppose we have two libraries of a tuning section, called $A$ and $B$, with corresponding quality measurements $A_1 \ldots A_n$ and $B_1 \ldots B_m$. To apply a classic statistical inference, we first make the assumption that the $A$’s and $B$’s are independent and are drawn from normally-distributed populations with true means $\mu_A$ and $\mu_B$. We proceed with a two-sample $t$ test with null-hypothesis $H_0$ and alternate hypotheses $H_a$ and $H_b$

$$
H_0 : \mu_A - \mu_B = 0 \\
H_a : \mu_A - \mu_B > 0 \\
H_b : \mu_A - \mu_B < 0
$$

(6.4)

To perform the test, we take $H_0$ as our prior assumption, i.e., that the two libraries have the same performance. Based on the quality measurements for $A$ and $B$, we try to reject $H_0$ with strong confidence in favor of one of the two alternatives: $H_a$ supposes that $A$ strictly better than $B$, in terms of quality, whereas $H_b$ supposes $B$ is strictly better.

A $t$ test will give us a confidence estimates for these two alternate hypotheses. The higher the confidence level, the more certain we are that one of the alternatives is the case, i.e., that there is a significant difference in performance between the two versions. With $s_A^2$ representing sample variance of $A$, the test statistic $t$ becomes

$$
t = \frac{\bar{A} - \bar{B}}{\sqrt{\frac{s_A^2}{n} + \frac{s_B^2}{m}}} \tag{6.5}
$$

and the degrees of freedom $v$ estimated from the data is

$$
v = \left\lfloor \frac{\left(\frac{s_A^2}{n} + \frac{s_B^2}{m}\right)^2}{\frac{(s_A^2/n)}{n-1} + \frac{(s_B^2/m)}{m-1}} \right\rfloor \tag{6.6}
$$

We reject $H_0$ in favor of $H_a$ with 95% confidence if $t \geq t_{0.95,v}$ and return $GreaterThan$ from COMPAREWITHSTUDENTST. The threshold value $t_{0.95,v}$ is the corresponding $p$-value threshold for Student’s $t$ distribution [Mandel, 1964, Table II]. Similarly, we reject $H_0$ in favor of $H_b$ if $t \leq -t_{0.95,v}$ and return $LessThan$. If neither or both of those inequalities are true, then we cannot determine with certainty whether $A$ or $B$ is better, so the comparison returns $NoAnswer$. 

57
Lau et al. [2006] makes no specific mention of two-sample hypothesis testing as we have described, but we believe that is essentially what they use. What is described is that is they construct a normal distribution with mean \( |\mu_A - \mu_B| \) and variance \( \frac{s_A^2}{n} + \frac{s_B^2}{m} \) to derive conclusions based on confidence intervals. This is close to what is described earlier, if a two-sample \( z \) test with null-hypothesis \( H_0 : |\mu_A - \mu_B| = 0 \) were used instead. Fundamentally, a \( z \) test either requires a knowing the population variance ahead-of-time, or relying on the law of large numbers and using the sample variance in place of the population variance [Devore and Berk, 2012]. Lau et al. use the latter approach of requiring a large number of samples to construct a normal distribution.

In practice, running-times are not normally distributed, so Lau et al.’s test required roughly 1,000 samples to discern a 10% speed-up. Our use of a \( t \) test instead of a \( z \) test allows us to make inferences with fewer numbers of samples. Specifically, we estimate the degrees-of-freedom \( v \) to use Student’s \( t \) distribution instead of relying on the law of large numbers. It is important to recognize, however, that as the number of samples grows, the \( t \) distribution turns into a normal distribution. Thus, with a large number of samples, Lau et al.’s approach to comparing two means is essentially the same as HALO’s.

6.2.2 Debt Repayment

Following the bakeoff, we transition from the Bakeoff state to the Payback state that helps offset the debt, \( i.e. \), overheads incurred, regardless of the bakeoff’s outcome. The Payback state is designed around the idea of amortizing the cost of performing a bakeoff through a period of exploiting the best known library following the bakeoff with no profiling overhead. During the first step of the Payback state, HALO calculates the number of steps needed to amortize the bakeoff by examining the bakeoff’s history.

Suppose we have a bakeoff history \( H \) for libraries A and B containing \( n \) quality observations, where at each time step \( i \), the deployed library saw fresh profiling data, and thus \( H_i \) is a fresh quality rating for that library. Figure 6.2 contains an example plot of the bakeoff history \( H_0 \ldots H_n \) with the horizontal axis \( t \) representing discretized time and the vertical representing quality. We can see that library B performed better than A, so that is the deployed library upon transitioning to the Payback state. The real performance \( R \) exhibited by the clients during the bakeoff is the area under the curve formed by \( H \), \( i.e. \), performance is the total quality during a period of time:
Figure 6.2: An overview of the calculation to determine how to amortize a bakeoff’s overhead during the Payback state. Figure is not to scale.

\[ R = \int_{0}^{n} H_t \, dt = \sum_{i=1}^{n} H_i \times \frac{H_i + H_{i-1}}{2} \]  \hspace{1cm} (6.7)

The value \( R \) is directly computed as a Riemann sum using the midpoint rule. Had we not performed any switching during the bakeoff and only used \( B \), ideally we would observe constant quality \( I_Q \), the maximum observed quality of \( B \) in \( H \). Knowing this ideal quality, we compute the debt \( D \) incurred during the bakeoff as:

\[ I = I_Q \times n \]
\[ D = |I - R| \]  \hspace{1cm} (6.8)

Since we will pause while using library \( B \) in the Payback state with sampling-based profiling disabled, we estimate the performance of \( B \) during debt repayment to be roughly 5% higher than \( I_Q \). It is through pausing while profiling is disabled that we pay off our debt \( D \). Finally, we compute the total payment \( P \) and the number of payback steps \( p \) as:
\[ J = I_Q \times 1.05 \]
\[ p = \left\lceil \frac{D}{J - I_Q} \right\rceil \]
\[ P = J \times p \geq D \]  

After computing \( p \), the decisions procedure for the Payback state transitions to itself \( p \) times before transitioning to the Decide state.

### 6.3 Exploration

So far, we have considered the decision-making process and how to evaluate a candidate library, but not how the candidate libraries are chosen. In order to adapt, HALO explores the space of knob configurations through two types of library selection decisions made by the Adaptive tuning manager: **Experiment** and **Retry**.

**Experiment**  The first kind of exploration is the decision to transition to the **Experiment** state (Figure 6.1), which generates a completely new library based on a knob configuration that is chosen by the automatic tuning strategy (Chapter 7). After submitting a compile job to the thread pool, the Adaptive tuning manager transitions to **Compile** and waits for the job to complete. During this wait, the client continues to execute using the best-known library, without any profiling enabled. If the resulting library’s object file is identical to an existing library (as determined by its SHA1 hash), then the new configuration is added to the existing library; effectively merging the two libraries. This merging of libraries is not an uncommon occurrence because not all knob changes will result in changes to how the program is compiled. Detecting duplicate libraries and merging them allows HALO server to use its compiler to rapidly search the configuration space without involving the clients. Thus, sometimes we transition back to **Experiment** to consult the automatic tuner for the next configuration to try. To prevent hangs, we limit the number of backward transitions from **Compile**. Once the limit is reached, an exiting library is chosen uniformly at random to be the candidate instead of compiling.

**Retry**  The second kind of exploration is to try one of the existing libraries, excluding the current-best already in use. Retrying a library can be beneficial if the workload or
environment has changed to favor a library that was previously considered unfavorable. To select an existing library, we first rank the libraries in descending order according the average quality following their most recent bakeoff. Then, we perform a random selection process that consists of repeated Bernoulli trials, so that we bias the selection towards better-performing libraries. We begin at the first-ranked library and flip a coin that yields heads with probability \( p \). If the coin yielded heads, the library is chosen to compete in a bakeoff, otherwise we move to the next library and flip the coin again. The probability distribution induced by this selection process follows the geometric distribution with the probability of success \( p = 1/3 \), as illustrated in Figure 6.3.

### 6.4 Rewards

During the transition from **Payback** to **Decide**, the reward assigned to the decision is selected based on the result of the bakeoff. The reward values were manually chosen to fall within the range \([-1, 1]\), with positive values indicating desirable outcomes:

- **Candidate Won Bakeoff** = 1
- **Bakeoff Timed-out** = −0.5
- **Candidate Lost Bakeoff** = −1

The rewards are fixed values because the cost of a bakeoff is amortized (Section 6.2.2). But, the outcome of each bakeoff is still unknown, which is why the decision problem in the **Decide** state is modeled as a multi-armed bandit. The range of rewards are centered around zero because the **Pause** action always yields a zero reward. To encourage early exploration, the initial estimated reward for the **Explore** action is set to 1/3, with the other actions initially estimated to yield a reward of zero.
Chapter 7

Automatic Tuning

Autotuning in HALO is focused on tuning LLVM’s optimization and compilation pipeline for the code contained in each tuning section (Section 5.3). The autotuner’s search method is a heuristic that explores the configuration space, i.e., it selects interesting candidates from the space based on the quality of candidates selected already. Section 7.1 discusses the types of LLVM compiler optimizations that are tuned by HALO for each tuning section. The remaining Sections of this Chapter discuss the search methods used during tuning.

7.1 Compiler Optimization Tuning

Each tuning section is considered to be a fresh instance of a tuning task. The tuning is performed seamlessly while the client processes are executing (online) using dynamic code patching (Section 5.2.3). Non-root functions included in the tuning section are made private to the tuning section’s compilation unit, which is also known as an LLVM module. This way, cross-function specializations or inlining that normally would be impossible can occur during tuning. Global values, especially those which are mutable, are marked as external to the module so that dynamic linking resolves to the correct symbols in the client process. Additionally, when the unoptimized module is initially created for a tuning section, all natural loops within functions are given a module-unique name to facilitate the tuning of optimizations for a particular loop.

Table 7.1 details all of the current knobs tuned by HALO. Because we create a set of knobs for each loop identified in the module, the size of the configuration space depends on the tuning section. All knobs may also be set to a value indicating that the setting is not specified, leaving it up to LLVM to decide or use its default value. The table is
Table 7.1: Settings tuned by HALO for each tuning section. All options are integers. The Default column indicates LLVM’s default setting for the given option.

<table>
<thead>
<tr>
<th>General Knobs (non-speculative)</th>
<th>Options</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Function Attributor Pass</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Use Partial Inliner</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Use Unroll-and-Jam</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Use NewGVN</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Use NewGVN Hoist</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Use GVN Sinking</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Extra Vectorizer Passes</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Experimental Alias Analysis</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Tune for Native CPU</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Use Interprocedural Register Allocation</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Use PBQP Register Allocator</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Optimization Pipeline Level</td>
<td>$i \in [2,3]$</td>
<td>–</td>
</tr>
<tr>
<td>Codegen Optimization Level</td>
<td>$i \in [2,3]$</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>General Knobs</th>
<th>Options</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inlining Threshold</td>
<td>${100i \mid i \in [0,30]}$</td>
<td>2.25</td>
</tr>
<tr>
<td>Jump-threading Threshold</td>
<td>$i \in [0,100]$</td>
<td>6</td>
</tr>
<tr>
<td>SLP Vectorization Cost Threshold</td>
<td>$i \in [-50,50]$</td>
<td>0</td>
</tr>
<tr>
<td>Use Loop Prefetching for Writes</td>
<td>$i \in [0,1]$</td>
<td>0</td>
</tr>
<tr>
<td>Loop Prefetching Distance</td>
<td>$i \in [0,100]$</td>
<td>0</td>
</tr>
<tr>
<td>Loop Versioning % Invariant Threshold</td>
<td>$i \in [0,100]$</td>
<td>disabled</td>
</tr>
<tr>
<td>Loop Interchange Cost Threshold</td>
<td>$i \in [-100,100]$</td>
<td>disabled</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knobs Per Loop in Module</th>
<th>Options</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unroll Count ($i = 0$ disabled)</td>
<td>${2i \mid i \in [0,16]}$</td>
<td>unset</td>
</tr>
<tr>
<td>Disable Runtime Unrolling</td>
<td>$i \in [0,1]$</td>
<td>unset</td>
</tr>
<tr>
<td>Enable Loop Distribution</td>
<td>$i \in [0,1]$</td>
<td>unset</td>
</tr>
<tr>
<td>Vectorization Width ($i = -1$ disabled)</td>
<td>${2i \mid i \in [-1,4]}$</td>
<td>unset</td>
</tr>
<tr>
<td>Interleaving Count ($i = -1$ disabled)</td>
<td>$i \in [-1,4]$</td>
<td>unset</td>
</tr>
</tbody>
</table>
separated into three groups of knobs. The first group are knobs which change an aspect
of the LLVM compiler for the entire module, e.g., “Experimental Alias Analysis” enables
non-default alias analysis infrastructure when optimizing the code. All of these non-
speculative general knobs control an existing aspect of LLVM’s optimizations or code
generation. I refer to this group as non-speculative because it is not clear that the option
exists because it might result in a performance regression. Rather, these non-speculative
knobs exist in LLVM for a few different reasons: (1) the option’s net benefit for application
runtime is not usually worth the extra compile-time required, (2) the option may generate
code that is not fully supported for the commonly-targeted CPU or ABI version, or (3) the
option controls a feature that is a work-in-progress. All of the other knobs are speculative in
the sense that they control an aspect of the compiler’s optimizations that require some cost
modelling or has trade-offs, and thus are more likely to yield a performance regression.
Specifically, speculative knobs do not only exist to save on compile time for common
cases. The remainder of this section details what the speculative knobs tuned by HALO
control and why they were chosen for tuning.

7.1.1 Inlining

The Inlining Threshold knob sets an abstract cost threshold for the LLVM function inliner. Setting this knob to a higher value indicates a higher tolerance for risk, so more aggressive inlining will happen. A number of existing works have found that function inlining is worth tuning [Cavazos and O’Boyle, 2005; Kulkarni et al., 2013; Lau et al., 2006; Waterman, 2006]. For LLVM, a higher threshold value indicates a stronger willingness to perform inlining at a call-site, because the threshold sets the tolerance for a program size increase due to inlining a call-site. But, the threshold cannot be described in a simple way due to the ad-hoc nature of the cost model’s decision procedure. Based on my examination of the source code, there are a number of ad-hoc bonuses and multipliers that scale the estimated costs based on properties of the targeted CPU and the callee. For example, a bonus is applied “if the callee has a single reachable basic block at the given callsite context.”

7.1.2 Jump Threading

In LLVM, jump threading duplicates basic blocks within a function to expose additional optimization opportunities by generating new control-flow paths within a function. Specifically, from the source code:
[Jump threading] looks at blocks that have multiple predecessors and multiple successors. If one or more of the predecessors of the block can be proven to always jump to one of the successors, we forward the edge from the predecessor to the successor by duplicating the contents of this block.

Thus, jump threading is a speculative optimization that achieves an effect similar to merge splitting, which was performed in the context of Jikes RVM (Section 4.1.3). The threshold value tuned by HALO controls the maximum number of instructions allowed in the duplicated basic block, with a value of zero disabling jump threading.

### 7.1.3 SLP Vectorization

First introduced by Larsen and Amarasinghe [2000], SLP Vectorization is an optimization that tries to increase the amount of superword-level parallelism (hence the name “SLP”) by combining groups of independent instructions into vector instructions. It is a type of auto-vectorization that operates on straight-line code rather than loops. HALO tunes a profitability threshold used within LLVM’s cost model controls whether to perform vectorization, with negative values making the optimizer more willing to vectorize.

### 7.1.4 Loop Prefetching

Software prefetching is an optimization that recognizes memory access patterns within loop bodies and decides whether to insert cache prefetching hints for the CPU to reduce memory stalls. On the x86-64, these hints take the form of a prefetch instruction that takes the form of a load instruction [Intel, 2020]. However, the prefetch instruction is simply a hint to the CPU that the cache-line referenced by the given address should be brought into the first-level cache because it will be used soon. Too many prefetch instructions, prefetches that are too early, or prefetches for data that is only conditionally used, can pollute the first-level cache. On the other hand, a well-timed prefetch can significantly improve performance. Software prefetching has historically been a prime candidate for adaptive optimization [Chilimbi and Hirzel, 2002; Lee et al., 2012; Lu et al., 2004; Saavedra and Daeyeon Park, 1996]. The knob controlling distance is in terms of number of instructions between the prefetch instruction and the first use of data on that cache line is the primary way that LLVM tunes its prefetching pass. A distance of zero disables the use of software prefetching.
7.1.5 Loop Versioning
[[ discuss LLVM’s LICM versioning pass ]]

7.1.6 Loop Interchange
[[ discuss ]]

7.1.7 Loop Unrolling
[[ discuss all loop unrolling knobs ]]

7.1.8 Loop Vectorization
[[ discuss loop vectorization width, interleaving, and distribution knobs ]]

7.2 Random Search

When attempting to search large space of knob configurations, simple random search over the entire space is easy to implement and a good first-strategy to try. Seymour et al. [2008] compared a number of more sophisticated search heuristics in the context of empirical autotuning for compilers, including Nelder-Mead simplex, genetic algorithms, and simulated annealing. They found that random search is the most effective overall strategy. The primary reason, according to Seymour et al., is that the search spaces are not overly complex: “there are many points with performance within 5% of the true maximum.” Knijnenburg et al. [2003] also found that random search can quickly find good performing configurations, which is very desirable for an overhead-sensitive system such as HALO. But, better-customized search strategies, such as the Nelder-Mead simplex algorithm in Active Harmony [Chung and Hollingsworth, 2004; Tiwari et al., 2009a], are more likely to avoid particularly bad configurations than random search for more difficult search spaces.

In HALO, we leverage two flavors of random search: local and global. Global random search is quite simple: for each knob, select among all possible options uniformly at random. Because a knob can be “unset” in HALO, all knobs in Table 7.1 are considered to have an additional option, added to the lower-end of the allowable range, that tells the compiler to use LLVM’s default setting.
Local Random Search  Since a knob can be viewed as a range of integers, they allow us to loosely think of the distance between two knob configurations as some norm that combines the distances between each knob value in each configuration. To perform local random search, we define a perturbing function that accepts a knob configuration $C$ plus a real-valued energy level $e \in [0, 1]$ and returns a configuration that is “nearby” the given configuration.\(^1\) For each knob in the input configuration, the perturbing function draws values from a normal distribution defined over the knob’s integer range, where the mean value is $C[k]$ for knob $k$. To help determine the variance of the distribution with respect to a knob, we define a vibration factor

$$V(knob, energy) = \frac{\left|\text{knob}.max - \text{knob}.min\right| \times energy}{2}$$

so that a higher energy level corresponds to a greater variance. Thus, the perturbing function produces a value nearby $C[k]$ for the new configuration’s setting for $k$ by first drawing $v$ from the normal distribution $\mathcal{N}(C[k], \sigma^2)$, where $\sigma = V(k, e)$, and then truncating $v$ within the knob’s range.

7.3 Surrogate Search

HALO’s primary implementation of automatic tuning uses a hybrid strategy that combines a search heuristic with a model to help filter out bad configurations (Section 2.4). Specifically, HALO’s surrogate search strategy combines random search with supervised learning progressively learn about the configuration space during exploration. A fresh learning model is dynamically trained from scratch for each tuning section. Viewing the automatic tuning problem as a black-box optimization problem to find a maximal input to a quality function $f$, in HALO that $f$ takes a knob configuration $C$ corresponding to some library, and outputs its quality according to its metric. Thus, our goal is to find a high-quality knob configuration for the tuning section. To execute that black-box quality function, we must conduct a bakeoff that is expensive and time consuming. The purpose of the model is to predict the quality of a configuration, based solely on characteristics of the knob configuration. The model serves as a surrogate for conducting a bakeoff, allowing HALO to rapidly explore the configuration space and only attempt a bakeoff for high-quality candidates. This idea is an extension of work by Nelson et al. [2015], who

\(^1\)This notion is borrowed from work in simulated annealing [Bertsimas and Tsitsiklis, 1993], where a point in space is like a particle that is vibrating within its neighborhood according to its energy-level.
used a surrogate to accelerate an exhaustive search of a small configuration space.

7.3.1 Bootstrapping

The importance of any particular knob in predicting the quality of a library is dependent upon the code contained in the tuning section. While other works have used a generic model that is trained to consider code features [Fursin et al., 2011], HALO trains a model specific to each tuning section. Thus, a process for bootstrapping the model is needed to generate data that the supervised model can learn from.

Bootstrapping consists of conducting bakeoffs using configurations selected by some other means, in order to gather some initial data points. The selection of initial configurations can have a dramatic effect on the number of search iterations required to find an optimal configuration [Balaprakash et al., 2013b]. Thus, care is taken to ensure that the initial configurations consist of some hand-picked configurations that are expected to perform well and some that are exotic. Two hand-crafted configurations are defined in HALO to be configurations that are sensible optimization choices as determined by a compiler expert. The first of these configurations unsets all knobs and then sets only the optimization and code generation levels to -O3, along with enabling Tune for Native CPU (Table 7.1). The second hand-crafted configuration is a superset of the first that additionally enables:

- Interprocedural Register Allocation
- The PBQP (Partitioned Boolean Quadratic Problems) Register Allocator
- The Function Attributor Pass
- Experimental Alias Analysis

These knobs were chosen entirely based on my own gut-feeling that these knobs are sensible to have enabled, but are not enabled by default in LLVM. The remaining initial configurations are randomly chosen to provide some variety in the initial batch of training data for the surrogate.

Algorithm 7.1 summarizes the procedure for selecting a configuration while in the Explore state of the Adaptive tuning manager (Figure 6.1). The procedure is written as a coroutine, where yield is similar to a return that saves the state of the loop in order
Input: history — a map from configurations to quality of the last bakeoff. 
Result: a configuration to try in the next bakeoff.

1. expertConfigs ← {...}
2. readyConfigs ← ∅
3. while true do
   4. if readyConfigs ≠ ∅ then
      5. yield readyConfigs.pop()
   6. else generate configurations
      7. if history.size() < minPrior = 5 then we must bootstrap
         8. if expertConfigs ≠ ∅ then
            9. yield expertConfigs.pop()
         10. else
            11. yield GENUNIFORMRANDOMCONFIG()
      12. else
         13. readyConfigs ← readyConfigs ∪ SURROGATESEARCH(history, ...)
      14. end
   15. end
   16. end

Algorithm 7.1: The surrogate-based configuration selection method.

to resume at that point on the next procedure invocation. The function GENUNIFORMRANDOMCONFIG uses the global random search strategy from Section 7.2 to generate well-varied configurations to explore the configuration space.

7.3.2 Generating Configurations

The SURROGATESEARCH function (Algorithm 7.2) generates a set of configurations that are worth experimenting with in a bakeoff. Similar to Nelson et al. [2015], the generation process relies on supervised learning to estimate a function that predicts the quality of a configuration, based on the existing history of configurations and their bakeoff performances. From Algorithm 7.2, the approximated function is modeled as the surrogate $S$ that allows us to quickly filter the configurations in $\mathcal{X}$. By default, the batchSize is set to 10 and controls how often we retrain a new surrogate to predict the quality of another exploreSize configurations, which is on the order of hundreds.

To prevent future overfitting of the model to its own opinions, a small percentage of the configurations returned by SURROGATESEARCH are chosen uniformly at random instead of relying on the surrogate’s quality estimate. This percentage is controlled by batchPct and by default 80% of the configurations returned are the surrogate’s top predictions,
Input: history — a map from config to quality from its last bakeoff  
    bestConfig — best config currently  
    exploreSize — number of configs to try with surrogate  
    explorePct — proportion of global versus local search  
    batchSize — number of configs requested  
    batchPct — proportion of random versus surrogate-chosen configs  

Result: a set of configurations, of size batchSize  

Algorithm 7.2: Implementation of SURROGATE SEARCH that uses a model to find candidate configurations for future bakeoffs.
with the rest randomly chosen. The set of configurations $\mathcal{X}$ explored by the surrogate is made up, by default, of equal-parts global random selection and local random selection that uses the configuration perturbing discussed in Section 7.2.

While supervised learning requires many examples to be effective, each quality rating in HALO is associated with a library that can contain multiple knob configurations. Thus, the merging of equivalent configurations into libraries after compilation (Section 6.3) helps generate additional data to train the surrogate model, since those configurations will be assigned the same quality rating.

The underlying surrogate model is implemented with fairly standard gradient-boosted decision trees [Friedman, 2001, 2002]. The TRAINMODEL procedure used in Algorithm 7.2 employs simple cross-validation with a randomly-selected held-out set for testing, and the learning objective function uses regression with squared loss. Each knob within a configuration corresponds to a feature of the model. The initial prediction of all instances prior to training, i.e., the global bias, is set to the mean of the entire dataset. Decision trees were chosen as the learner of the model for two practical reasons: (1) decision trees are somewhat easier to read in debugging dumps than other learned models, and (2) there are high-quality implementations accessible from C++, which is HALO’s implementation.
language. Figure 7.1 is an example of one decision tree in the trained model that was obtained after 60 executions of the spectralnorm benchmark on the workstation machine from Chapter 8.

At each leaf of the decision tree, val is an indicator of the quality of matching configurations (with larger values meaning better quality), and cover tells us how many samples in the training set matched that node. The tree in Figure 7.1 appears to have overfit, probably due to a small number of examples, because Partial Inlining is unlikely to actually matter for the benchmark’s tuning section. But, HALO found the first configuration that is roughly 1.75× better as a suggestion from SURROGATESEARCH.
Chapter 8

Experimental Results

This chapter provides an evaluation of the capabilities of a prototype of HALO, as described in earlier chapters, across a number of C and C++ benchmark programs and machines. We start with an empirical performance comparison that compares HALO with ahead-of-time optimization and traditional just-in-time optimization (Section 8.2). Then, we take a deeper look at the two quality metrics in HALO in Section 8.3. Finally, in Section 8.4 we consider the overhead of compiling and distributing binaries that are compatible with HALO when the server is not available.

8.1 Experiment Setup

A number of benchmark programs from the LLVM compiler’s test suite were chosen to evaluate HALO’s capabilities. Programs were chosen if they are not heavily reliant on external libraries, because such library code will not be visible to HALO for optimization due to limitations of the prototype. The benchmark programs are as follows:

- **lpbench** — a modernized version of the classic LINPACK benchmark [Dongarra et al., 2003]. Matrices of size 1000 × 1000 are used.
- **matrix** — a synthetic benchmark that multiplies square matrices of size 256 × 256.
- **nbody** — a benchmark that performs a double-precision N-body simulation. It models the orbits of Jovian planets, using a symplectic-integrator [Gouy].
- **perlin** — a program that applies Perlin [1985, 2002]’s noise filter to arbitrary data.
• **spectralnorm** — a synthetic benchmark that computes the spectral norm of a 2000 × 2000 matrix using the power method [Gouy].

• **sphereflake** — a raytracer that utilizes a bounding-volume hierarchy for collision detection. An image with dimensions 1280 × 1280 pixels is rendered for a scene resembling a three-dimensional Koch [1904] snowflake made of reflective spheres.

These programs perform one heavy-duty task so that we can evaluate the most important aspects of HALO within a reasonable amount of time. Specifically, the dominant time spent in these programs is automatically captured within a single tuning section, so that we can explore the the long-term behavior of the HALO prototype. During each trial-execution of these benchmark programs, the work performed by the program is repeated a number of times to emphasize its execution-time overhead. We perform five trials of each program that performs a fixed workload, *i.e.*, repeated execution of the computational kernel.

**Machine Specifications**  Rather than dynamically varying the inputs to the programs, we will evaluate another piece of latent information for adaptive recompilation: hardware. We experimented with three types of machines with the following CPU specifications:

• **workstation** — 2 × Intel Xeon Gold 6142 CPUs each with a 22 MB cache.

• **desktop** — Intel i7-3770 CPU with an 8MB cache.

• **mobile** — Broadcom BCM2837B0, with four ARM Cortex-A53 cores and 512 KB last-level cache; running in 32-bit mode (*armhf*) with fan-cooled heatsinks.

### 8.2 Performance Comparison

In this section, we test the effectiveness of HALO’s ability to optimize programs in production. An experimental trial consists of a single instance of the HALO server, followed by a number of “tuning iterations” that are repeated executions of the compiled benchmark program that utilizes the server. The same instance of the HALO server is used for all tuning iterations of a benchmark program so that tuning process can be picked-up again between program executions. Except for the mobile machine, the server process is
running on the same machine as the client programs. To help avoid interference with the client programs, the server’s parallel compilation thread-pool is limited to two.

We record the execution time of each tuning iteration and plot the progression over time in Figures 8.1 and 8.2. For comparison, we use the running time of each program under the best offline or ahead-of-time optimization methods available to ordinary users, called \textit{aot} in the Figures, as a baseline.

The \textit{aot} method does not use HALO and instead utilizes a two-stage compilation procedure, where the program is first compiled and run with instrumentation to gather profile data. That profiling data is used by the compiler when compiling the program for the second and final time, using the flags \texttt{-fprofile-instr-use=<file> -O3 -mcpu=native}.

The \textit{jit} method utilizes HALO’s “jit once” tuning manager from Section 5.3.3 that compiles the tuning section with options equivalent to \texttt{-O3 -mcpu=native}. This method is designed to simulate a traditional single-stage JIT compiler that does not perform search.

The remaining methods utilize HALO’s adaptive tuning manager with two different performance metrics, instructions-per-cycle (Section 5.3.1) and call frequency (Section 5.2.1). Overall we can see that the IPC metric is not accurate in the majority of these benchmarks. Deeper investigation into why has revealed that it is not an inaccuracy of the measure itself, but instead a disconnect between IPC increases and execution-time decreases. In some instances, the tuner produces code that has a higher IPC only because the code contains many unnecessary cheap instructions, increasing the overall running time of the program. Thus, the remainder of our discussion is focused on HALO under the call-frequency metric.

**Workstation Performance** The \textit{matrix} benchmark sees a large speedup of nearly $1.4 \times$ with HALO compared to the AOT and JIT approaches by the 5th tuning iteration (Figure 8.2b). For \textit{spectralnorm} peak performance is not reached until the 15th iteration, but a huge speed-up of $1.75 \times$ is achieved by halo (Figure 8.1e). For other programs, the performance of HALO with the call metric is in line with the JIT approach, except for \textit{perlin}. In \textit{perlin} we see large spikes of performance regressions that last for one or two iterations at a time, followed by a recovery period.

**Desktop Performance** \texttt{[[ No improvement for matrix, but a small improvement for lp-bench. An even bigger speed-up for spectralnorm: 2\times! ]]}
Figure 8.1: Results for the **workstation** machine. Higher speedups are better.
Figure 8.2: Results for the desktop machine. Higher speedups are better.
Figure 8.3: Comparing the Call and IPC metrics on the workstation. Higher speedups are better.

8.3 Quality Metrics

[[ discuss why the IPC metric does not work well ]]

8.4 Offline Overhead

[[ Compute avg runtime overhead of -O1 compiled programs with and without -fhalo, without running the halo server at all. I’ll also summarize the size overhead of the application. This stresses the overhead of -fhalo the most because it optimizes but does not perform inlining. For all benchmarks on all machines and put them into a table, then discuss findings ]]
Chapter 9

Conclusions

[[ A search-based online adaptive optimization system is feasible and can yield sizable performance improvements for some programs and machines. But, more work is needed to improve the quality metric without introducing additional overhead, because the margins to stay within the same range of AOT optimization are slim. ]]

9.1 Future Work

[[ more ]]

Population Experiments When multiple HALO clients are code-compatible, the tuning effort can become a collaborative process where clients with similar behavior in-the-large, i.e., similar profiles, are tuned together by the server. In particular, randomized controlled trials among the population of similar clients can be used to determine the quality of a code variant during the empirical search process. This helps mitigate the negative impacts of a poor-performing code variant on the productivity of the group of clients as a whole, because we only experiment on a subset.
Abbreviations

AOT  Ahead-of-time
BBO  Black-box Optimization
CCT  Calling-context tree
HALO Wholly Adaptive LLVM Optimizer
JIT  Just-in-time
MAB  Multi-armed Bandit
OAO  Online Adaptive Optimization
VM   Virtual Machine
References


Jason Ansel, Maciej Pacula, Yee Lok Wong, Cy Chan, Marek Olszewski, Una-May O’Reilly, and Saman Amarasinghe. SiblingRivalry: online autotuning through lo-


Michael Paleczny, Christopher Vick, and Cliff Click. The Java HotSpot™ Server Compiler. page 13, April 2001.


