AUTOMATIC INFERENCE AND EFFECTIVE APPLICATION OF TEMPORAL SPECIFICATIONS

A Dissertation

Presented to

the Faculty of the School of Engineering and Applied Science

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in Computer Science

by

Jinlin Yang

May 2007

Approval Sheet

This dissertation is submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

Jinlin Yang

This dissertation has been read and approved by the examining Committee:

David E. Evans, Advisor

Manuvir Das

Jack W. Davidson

Joanne B. Dugan

John C. Knight

Mary Lou Soffa, Chair

Accepted for the School of Engineering and Applied Science:

James H. Aylor, Dean, School of Engineering and Applied Science

May 2007

Abstract

Software specifications are the foundation of many software development activities including maintenance, testing, and verification. However, specifications are rarely available for real systems. This dissertation describes a dynamic analysis technique to automatically infer program specifications. We focus on temporal specifications, an important category of specifications that constrain the order of occurrence of events. Our technique generates execution traces of the program by running a set of test cases and then analyzes the properties satisfied by traces using inference techniques and a set of predefined patterns. Our approach makes three contributions over previous work, focusing on enabling effective dynamic inference on large programs under realistic conditions. First, our inference algorithm scales well to large execution traces that have millions of events and thousands of distinct events. Second, our statistical inference algorithm can successfully infer property specifications from imperfect traces collected from buggy programs or using inadequate instrumentation tools. Third, our heuristics can effectively reduce the large number of properties inferred to a manageable set of mostly interesting properties. We implemented our dynamic analysis technique in a prototype tool called Perracotta. To evaluate the usefulness of dynamically inferred temporal properties, we applied Perracotta to aid program understanding, verification, and differencing. Results include inferring a 24-state finite state machine from the JBoss transaction manager that is consistent with the J2EE specification, inferring interesting API rules for the Windows kernel, and detecting a previously unknown deadlock bug in Windows by checking the inferred properties with the ESP verifier.

Dissertation Advisor: David E. Evans

Title: Associate Professor, Computer Science

Acknowledgments

I am very lucky to have David Evans as my advisor. Without his advice, encouragement, and patience, I would not have been able to complete my PhD study. David taught me everything from how to select research problem to conduct experiments, from how to write papers and proposals to give talks, and from how to be a rigorous researcher to be a mentor. I am very grateful for all the opportunities I had of attending conferences and meeting department visitors, which not only allowed me to get feedback of my work but also to start building my own professional network. I also want to thank him for his support, effort, and understanding in my job hunting process.

I want to thank my dissertation committee members for all their feedback and advice about my research. I want to thank Manuvir Das for being a wonderful mentor of my internship at Microsoft and an enthusiast of my research. It is always exciting to discuss my project with Manuvir. I want to thank Jack Davidson for his kindness and support of graduate students. I want to thank Joanne Dugan for her interest in my research and asking thought-provoking questions about my work. I enjoyed learning fault trees from Joanne who is the best expert on this subject. I want to thank John Knight for all his encouragement and support during my five years at UVa. Attending John's class in dependable computing systems introduced me to research in building reliable systems. I want to thank May Lou Soffa for giving me many detailed comments about my work, dissertation, and job talk. I appreciate her thoughtful comments about my writings and presentations. In addition, I want to thank Manuvir Das, John Knight, and Mary Lou Soffa for their support in my job hunting.

The University of Virginia provided a stimulating environment for my PhD study. I want to thank Sudhanva Gurumurthi, Kim Hazelwood, Greg Humphreys, and Westley Weimer for sharing their experience and helping me when I was looking for job. Westley Weimer's Programming

Language class is one of the best computer science courses I have ever taken. He is always willing to give helpful advice to his students. I thank Westley for all the time and effort he spent on helping me during my job searching. I thank Shukang Zhou for being a wonderful friend in research and also social life. I am grateful for being included in the PLEASE reading group and certainly have enjoyed it very much. I thank Chengdu Huang for helping me when I had trouble in experiments. I thank members of our research group, Benjamin Cox, Karsten Nohl, Nathanael Paul, Jeffrey Shirley, Ana Nora Sovarel, and Joel Winstead, for being very helpful and encouraging friends. I thank John Pfaltz for his interest and advice for my research.

Interning at the Center for Software Excellence at Microsoft was a wonderful experience. I thank everybody in the Program Analysis Group and the Test Effectiveness Group for their friendliness and willingness to help. In particular, I thank Deepali Bhardwaj and Thirumalesh Bhat for supporting my work. I thank Stephen Adams, Jason Yang, and Zhe Yang for helping me with running ESP. I thank Lei Zhang and other anonymous Windows developers for their enthusiasm in my project. I thank Brian Hackett, Aquinas Hobor, Sumant Kowshik, Henning K. Rohde, Frances Spalding, Daniel Wang and Zhe Yang for being wonderful friends during my internship.

I thank Zhong Xiu for being a supportive friend since college for all his advice and friendship.

Lastly, but not least, I thank my parents (Ying Wang and Lijiang Yang), grandma (Aimei Du), and wife (Thao Doan) for all their support and love. I appreciate that my wife proofread my dissertation with very short notice.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

Software is pervasive. From ATMs to cell phones, from payroll to health insurance, and from cars to airplanes, nearly every aspect of a person's life depends on software. An average car today has about 35 million lines of code that runs in about 30 microprocessors communicating on a high-bandwidth network [Duvall05]. The amount of software installed in cars will continue increasing. IBM estimates that, by 2010, 90% of the new innovations packed into cars will come from software that controls everything from a car's headlights to its brake system [Duvall05].

Researchers have claimed, for many years, that the use of software specifications can improve many software development activities. A formal specification documents important properties and hence is useful in understanding programs [Hoar69, Pnueli77]. Formal specifications can be used to automatically generate test inputs [Dick93, Memon01, Coppit05]. Program verification needs a formal specification that defines the correct behaviors of a program [Hoar69, Pnueli77, Gries81]. Other uses include refining a specification into a correct program [Abrial96] and protecting a programmer from making changes that violate important invariants [Ernst01].

Despite these benefits, formal specifications are nearly always absent in real systems [Holloway96, Knight97, Lamsw00]. To help realize the benefits of formal specifications, this work develops techniques for automatically inferring formal specifications and investigates the uses of inferred specifications in software development.

This work focuses on temporal properties. Temporal properties constrain the order of a program's states [Pnueli77, Kröger87]. For example, acquiring a lock should always be followed by releasing the lock. Temporal properties are important in many types of systems such as network protocols. Satisfying temporal properties is crucial for establishing a program's correctness proper-

ties.

1.1 Earlier Attempts to Specification Inference

Several researchers have recognized the unavailability of specifications as an important problem and studied specification inference and its application to software development [Ernst01, Flan01, Ammons02, Whaley02, Henkel04, Alur05, Weimer05]. A *specification inference technique* automatically discovers a formal specification of a target program by analyzing program artifacts. A specification inference tool is often called a *specification miner* [Ammons02], *specification synthesizer* [Alur05], or *specification prospector* [Mande05]. A program's source code or execution traces are the most common artifacts used by specification miners, though it is also possible to analyze other artifacts including design documents, bug reports, and revision history [Cook98, Livs05].

A static inference technique analyzes source code. In addition to the advantage of not needing to execute the target program, a static inference technique examines all execution paths of a program and therefore can infer precise specifications of a program. In contrast, a dynamic inference technique analyzes execution traces and only sees execution paths present in the traces.

Hence, a dynamic inference technique might infer specifications that are stronger than the program itself when the executions do not cover all possible scenarios. However, features such as pointers, branches, loops, threads, inheritance, and polymorphic interfaces, almost ubiquitous in modern programming languages, can be expensive to analyze precisely from source code. A static inference technique must balance precision and efficiency. An efficient static inference technique often produces specifications that are too general to be useful. A static inference technique with high precision usually fails to scale to large programs. On the other hand, an execution trace can have precise information about pointers, branches, threads, and the other features of programming executions.

We take the dynamic inference approach in our work for several reasons. The temporal specifications we aim to infer typically involve objects and threads that are difficult to analyze statically. In addition, our goal is to be able to apply our techniques to large real systems that typically have complex control flow structures which pose a scalability challenge for static techniques.

Previous dynamic specification inference techniques have shown promising results in many areas, including bug detection [Hangal02, Nimm02, Pytlik03, Livs05], test case selection [Gupta03, Hard03, Xie06], and program steering [Lin04]. However, all of the results to date have been on relatively small execution traces. For example, the largest execution traces analyzed by Daikon have only hundreds of variables, whereas a large system usually has thousands of variables [Ernst01, Perk04].

Scaling dynamic inference to handle large real systems involves several important challenges. The inference technique must effectively deal with imperfect execution traces. An *imperfect execution trace* is a trace that contains event sequences that violate a property specification that is necessary for the correctness of a system. Suppose we want to learn the temporal specification of a type of lock. If our target program neglects to release this type of lock during some executions (due to bugs), running this program would produce an imperfect trace that fails to exhibit the rule for correctly using the lock. A dynamic inference technique needs to effectively deal with such imperfect execution traces otherwise it would risk missing important specifications in the inference results. For example, the Strauss specification miner requires human guidance to tune an imperfect trace so that it can discover important specifications that would otherwise be missing [Ammons02]. Daikon requires 100% satisfaction of a pattern [Ernst01], which might exclude important specifications if the execution trace is imperfect. Although we aim to handle imperfect traces, we assume that our traces are mostly correct - our target program should exhibit the desirable behavior most of the time.

Our inference technique must be able to select interesting specifications. An *interesting speci-*

fication is a specification for which developers are likely to make mistakes and violation of which would produce bad consequences. For example, we consider specifications about using critical system resources such as locks and transactions to be interesting. Such properties are important because violating them can have serious consequences such as causing system crashes [Ball01, Das02] and opening security vulnerabilities [Chen02]. Selecting interesting specifications is important because for a large program thousands of properties may be inferred, only a small fraction of which are interesting. We present several techniques that can effectively increase the percentage of interesting properties in the results.

In summary, we address several issues that prevent earlier dynamic inference techniques from being applied effectively to large programs:

1. The inference algorithms scale poorly with the size of the program and the execution traces.

2. Previous dynamic inference techniques do not work well in situations where perfect traces are not available.

3. A significant portion of the inferred properties are uninteresting. For small programs, it is feasible to manually select the interesting properties; for large programs, property selection must be mostly automated.

1.2 Contributions

The thesis of this dissertation is that ***dynamic inference techniques can automatically produce temporal specifications for large programs that are useful for a variety of software development tasks including program understanding, verification, and differencing.***

This dissertation describes dynamic inference techniques for automatically inferring temporal specifications and experimentally evaluates our techniques on real systems, as well as several different applications of the inferred properties. Our dynamic inference techniques address the three

issues of earlier inference work described in the previous section. In particular we make the following contributions in the area of scaling dynamic inference techniques:

1. We develop a scalable inference algorithm that can analyze large execution traces.

2. We create a statistical inference algorithm that can deal with imperfect execution traces.

3. We develop two heuristics for eliminating uninteresting properties. These heuristics increase the percentage of interesting properties in the inference results and are crucial for the approach to be useful in practice.

4. We present a chaining method for constructing large finite state automata out of a set of smaller ones. This method is useful for presenting a large number of inferred properties in a more readable way.

In order to evaluate our approach, we built a prototype tool called Perracotta and applied it to several real systems including Microsoft Windows and the JBoss Application Server. The results demonstrate that the dynamic analysis technique is useful in several different software development activities. We make the following experimental contributions:

5. We show that the dynamic analysis technique can abstract important temporal behaviors of a complex system. Hence, the inferred properties can help programmers understand the temporal behaviors of real systems.

6. We demonstrate that checking the inferred properties with verification tools can find application-specific bugs in real systems.

7. We show that the technique can aid in program differencing by discovering important differences among multiple versions of real systems.

1.3 Overview

Chapter 2 describes the dynamic inference techniques including the scalable inference algorithm (Sections 2.4 and 2.5), the heuristics for selecting properties (Section 2.7), and the chaining method (Section 2.8).

Chapter 3 presents the inference results on real systems with a focus on helping program understanding. Chapter 4 describes other uses of the inferred properties. Section 4.1 describes combining Perracotta with two program verification tools. Section 4.2 presents the experiments of using Perracotta in program differencing.

Chapter 5 evaluates our inference approach based on the experimental results. Chapter 6 surveys related work. Chapter 7 concludes with a discussion about the contributions, limitations, and future work.

Chapter 2
Specification Inference[1]

This chapter presents our dynamic temporal specification inference approach. We use a simple Producer-Consumer program as our running example (Section 2.1). To infer temporal specifications for a program, our inference approach follows several steps as shown in Figure 2.1. An instrumentor instruments the program to monitor information of interest (Section 2.2). Then, the instrumented program is executed against a set of test cases to produce execution traces (Section 2.3). Next, the inference engine matches the traces against a set of predefined property templates (Section 2.4). A post-processor selects the interesting properties out of the matched properties using several heuristics (Section 2.7). The chaining method aims to condense the inferred properties so that users can better comprehend them (Section 2.8). Section 2.9 describes Perracotta, a prototype implementation of our inference approach.

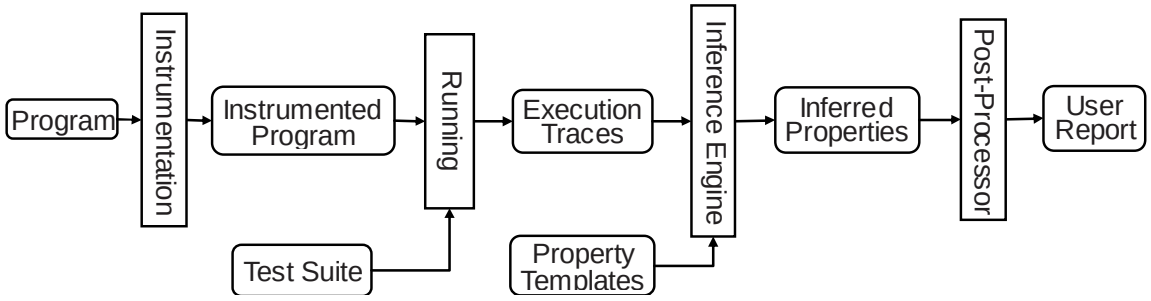


Figure 2.1: Overview of our approach.

2.1 A Running Example: Producer-Consumer

The Java program in Figure 2.2 implements a simplified version of the Producer-Consumer problem. The Producer class and the Consumer class implement a Producer and a Consumer respectively. Only one Producer and one Consumer object exist at any time. The Producer interacts

[1] This chapter is partly based on [Yang04a] and [Yang06].

```
class Buffer {
   int queue = -1;
   public synchronized int take() {
      int value;
      while (queue < 0)
         try { wait(); } catch(InterruptedException ex) {}
      value = queue;
      queue = -1;
      notifyAll();
      return value;
   }
   public synchronized void add(int x) {
   while (queue != -1)
         try { wait(); } catch(InterruptedException ex) {}
      queue = x;
      notifyAll();
   }
   public synchronized void stop() {
      while (queue != -1)
         try { wait (); } catch (InterruptedException ex) {}
      queue = 0; notifyAll ();
   }
}

class Heap { static Buffer buf; }

class Producer {
   static public void main(String[] args) {
      Heap.buf = new Buffer();
      (new Consumer()).start();
      for(int i = 1; i < Integer.valueOf(args[0]).intValue(); i++)
         Heap.buf.add(i);
      Heap.buf.stop();
   }
}

class Consumer extends Thread {
   public void run () {
      int tmp = -1;
      while ((tmp = Heap.buf.take ()) != 0)
         System.err.println ("Result: " + tmp);
   }
}
```

Figure 2.2: A Java implementation of the simplified Producer-Consumer problem.

with the Consumer through a global Buffer object, buf, which is a static member of the Heap class. At any time, a Buffer object can only hold one integer element, queue, whose value can be retrieved through the take method and updated through either the add method or the stop method. The buffer is empty when the value of queue is -1. The Producer inserts a new integer into buf by calling its add method and the Consumer removes an integer from buf by calling its take method. The Producer iteratively inserts integers from 1 to n (as designated by program arguments) into buf, while the Consumer takes those numbers from buf and prints them out. After the Producer calls the stop method that writes 0 to buf, the Consumer reads 0, exits the run loop, and terminates.

All three methods of the Buffer class, take, add, and stop, are declared with the Java synchronized keyword to ensure mutual exclusion among multiple threads that access a same Buffer object. We implement synchronization among multiple threads using the commonly used Java wait-notify idiom.

This program exhibits two interesting properties: (1) inserting an integer to buf alternates with removing an integer from buf. Hence, the Producer cannot overwrite a new element before the Consumer retrieves it. Furthermore, the Consumer cannot take an element from an empty buffer. (2) once the Producer calls the stop method, the Consumer must eventually stop.

2.2 Instrumentation

A program execution has a large amount of information: values of object fields, thread contexts, values of arguments, branches taken, exceptions raised, etc. Some information represents an object's state (e.g., object fields), whereas other information captures control flow (e.g., branches taken, exceptions raised).

Ideally we would want to record every detail of a program's execution. This is impractical for several reasons. Instrumenting everything without affecting a program's normal behavior is

```
Enter:Producer.main():[main]
Enter:Buffer.add():[main]
Enter:Consumer.run():[Thread-1]
Exit:Buffer.add():[main]
Enter:Buffer.take():[Thread-1]
Enter:Buffer.add():[main]
Exit:Buffer.take():[Thread-1]
Exit:Buffer.add():[main]
Enter:Buffer.add():[main]
Enter:Buffer.take():[Thread-1]
Exit:Buffer.take():[Thread-1]
Exit:Buffer.add():[main]
Enter:Buffer.take():[Thread-1]
Enter:Buffer.add():[main]
Exit:Buffer.take():[Thread-1]
Exit:Buffer.add():[main]
Enter:Buffer.take():[Thread-1]
Enter:Buffer.stop():[main]
Exit:Buffer.take():[Thread-1]
Exit:Buffer.stop():[main]
Enter:Buffer.take():[Thread-1]
Exit:Producer.main():[main]
Exit:Buffer.take():[Thread-1]
Exit:Consumer.run():[Thread-1]
```

Figure 2.3: A trace of running the Producer-Consumer program.

Each line corresponds to a single event that represents the entrance or exit of a method. We omit the method's signature, argument values, and return values to simplify presentation. For example, Enter:Producer.main():[main] indicates that the main thread enters the main method in the Producer class.

difficult. For example, in a real-time system, the overhead introduced by the instrumentation code might prevent a process from meeting its deadline. Furthermore, collecting all information does not scale to long-running systems as the size of the data becomes too large to be efficiently stored and processed.

Our instrumentor instruments a program at the method level and records the thread contexts, argument values, and return values (Section 2.9.1). Our instrumentor instruments the entrance and exit events of all the methods in the Producer-Consumer program. For example, executing the instrumented Producer-Consumer program with 5 as its input produces the execution trace shown in Figure 2.3. To simplify presentation, we do not include the argument values and return values in this trace.

Each line in the trace corresponds to a single event. An event starts with either Enter or Exit corresponding to the entrance and exit event respectively. The middle part of an event is a method's name. The last part of an event includes the thread context information and, optionally, argument values. For example, Enter:Producer.main():[main] indicates that the main thread enters the main method in the Producer class.

2.3 Running

The executions affect the results of any dynamic analysis. Our goal is to develop a dynamic inference technique that works with readily available or easily produced execution traces. Hence, we run a target system's regression test suite if one is available. Furthermore, when generating inputs is necessary, we either randomly select inputs from a program's input domain [Duran84, Weyu80] or exhaustively generate all inputs within certain bounds [Boya02, Khurshid02, Coppit05].

For the example, we run the Producer-Consumer program with 100 randomly selected integers between 1 and 10000.

2.4 Inference Engine

This section describes the inference engine. First, we introduce the predefined property templates in Section 2.4.1. Section 2.4.2 describes our inference algorithm that scales to large execution traces. Section 2.4.3 presents an algorithm for inferring the strictest pattern.

2.4.1 Property Templates

A property template abstracts a set of concrete properties. Our property templates have two parameters that can be substituted with values to generate concrete properties. Property templates determine the properties that can be inferred. It is essential that these templates capture properties users care about.

Dwyer et al. developed a temporal property pattern library after surveying hundreds of tempo-

Table 2.1: Temporal property templates.

For example, the regular expression of the MultiEffect pattern is $(PSS^*)^*$. Hence, PSS is a string that satisfies the MultiEffect template, whereas PPS or SPS do not satisfy the template.

Name	Regular Expression	Satisfying Example	Violating Examples
Response	*S*(PP*SS*)**	*SPPSS*	*SPPSSP, PSP*
Alternating	*(PS)**	*PSPS*	*PSS, PPS, SPS*
MultiEffect	*(PSS*)**	*PSS*	*PPS, SPS*
MultiCause	*(PP*S)**	*PPS*	*PSS, SPS*
EffectFirst	*S*(PS)**	*SPS*	*PSS, PPS*
CauseFirst	*(PP*SS*)**	*PPSS*	*SPSS, SPPS*
OneCause	*S*(PSS*)**	*SPSS*	*PPSS, SPPS*
OneEffect	*S*(PP*S)**	*SPPS*	*PPSS, SPSS*

ral property specifications checked by program verification tools [Dwyer99]. One pattern in their library is the Response pattern, which constrains the cause-effect relationship between two events P and S so that P's occurrence must be followed by S's occurrence. We use regular expressions to represent these patterns. For example, $[\neg P]^*(P[\neg S]^*S[\neg P]^*)^*$ is the regular expression for the Response pattern. After removing all events other than P and S, the Response pattern can be simplified as $S^*(PP^*SS^*)^*$. The Response pattern does not constrain the number of P events, the number of S events, or whether the S event can occur before the P event. As a result, knowing that two events satisfy the Response pattern does not give us precise information about their relationship.

We introduce seven property patterns based on the Response pattern. Table 2.1 shows the seven new patterns, their representation as a regular expression, and example strings that satisfy or violate the patterns. For example, the MultiEffect pattern is $(PSS^*)^*$; PSS is a string that satisfies the MultiEffect pattern. Furthermore, the MultiEffect pattern only allows one P event to occur between two S events and also requires that the first P event to occur before the first S event. Hence, PPS and SPS do not satisfy the MultiEffect pattern. Another pattern is Alternating that requires a strictly alternating relationship between two events. Its regular expression is $(PS)^*$. For example, $PSPS$

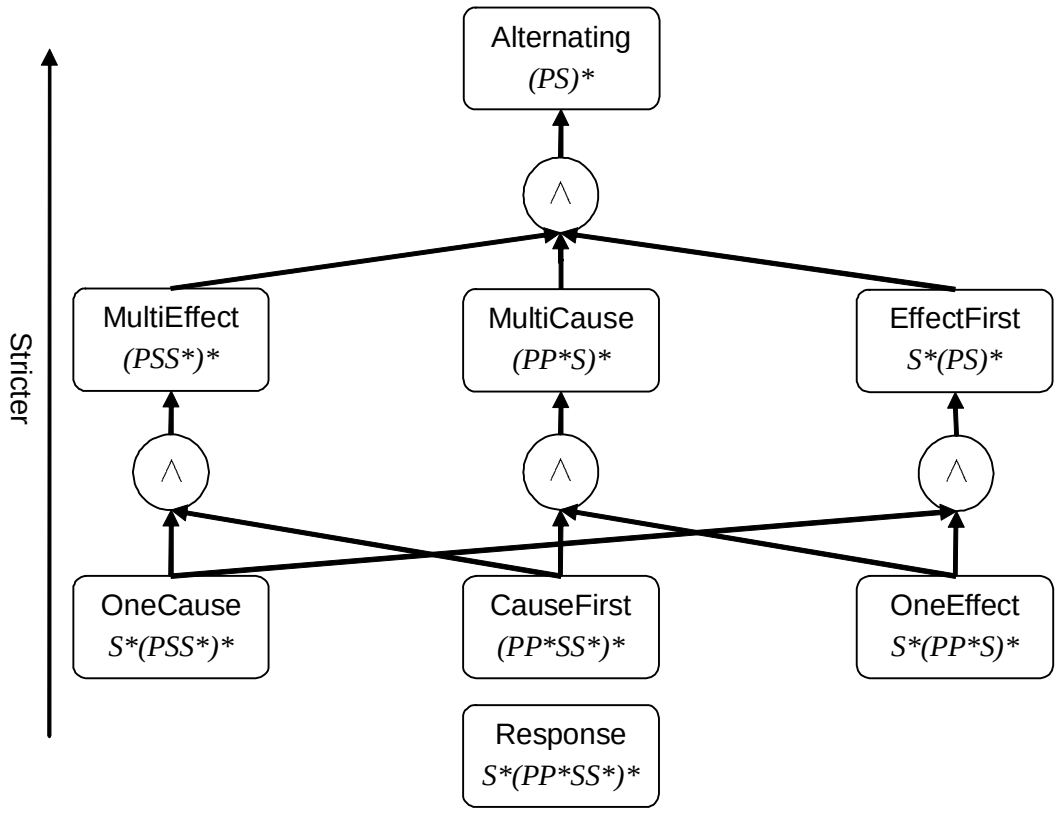


Figure 2.4: Partial order of property templates.

Each box shows a property template and its regular expression representation. A pattern A is stricter than another pattern B if $L(A) \subset L(B)$, where $L(A)$ means all the strings accepted by A. The eight patterns form a partial order in terms of their strictness. For example, Alternating is the strictest pattern among them. In addition, these patterns have an internal logical relationship as illustrated by the logical $\wedge$ operators among them. For example, a string satisfies the MultiEffect pattern if and only if it satisfies the OneCause and CauseFirst patterns.

satisfies the Alternating template, whereas PSS, PPS, and SPS all violate the Alternating pattern. We also use $P \rightarrow S$ to indicate that P and S satisfy the Alternating pattern.

The Alternating pattern is *stricter* than the MultiEffect pattern because all strings that satisfy the Alternating pattern must also satisfy the MultiEffect pattern but not vice versa. Formally, we say a pattern A is stricter than another pattern B if $L(A) \subset L(B)$, where $L(A)$ is the set of strings accepted by A. The seven new patterns form a partial order in terms of their strictness as illustrated in Figure 2.4. In particular, OneCause, CauseFirst, and OneEffect are the three primitive patterns that are the least strict among the seven. Above them come MultiEffect, MultiCause, and EffectFirst. Alternating is the strictest pattern among the seven patterns. In addition, these patterns

have internal logic relationship as illustrated by the logical $\wedge$ operators among them. For example, a string satisfies the MultiEffect pattern if and only if it satisfies the OneCause and CauseFirst patterns.

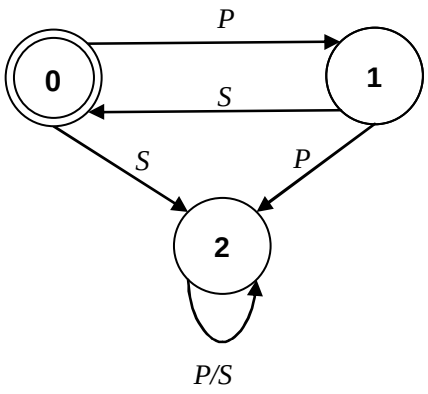
A string can satisfy more than one of the seven patterns. For example, PSS satisfies the MultiEffect, CauseFirst, and OneCause patterns. We want to infer the strictest pattern a string can satisfy. For example, the strictest pattern PSS satisfies is the MultiEffect pattern. We can derive the strictest pattern a string satisfies by exploring the logical relationship among the patterns. Our algorithm first determines which of the three primitive patterns a string satisfies and then deduces the strictest pattern. For example, if a string satisfies all three primitive patterns, then the strictest pattern it satisfies is Alternating.

In addition to the patterns with two parameters, we also derive two Alternating patterns that have three parameters. The two-effect-alternating pattern, $P \rightarrow S \mid T$, allows P to alternate with either S or T. The two-cause-alternating pattern, $P \mid S \rightarrow T$, requires either P or S to alternate with T. These patterns correspond to properties of real programs. For example, a file, after being successfully opened, can be either read or written, and finally be closed.

Section 2.4.2 presents a scalable algorithm for deciding whether a string satisfies a pattern. Next, Section 2.4.3 extends the algorithm to infer the strictest pattern a string satisfies.

2.4.2 Pattern Matching Algorithm

All the patterns described in the previous section have only two or three parameters (e.g., P, S, and T). We present our pattern matching algorithm using the Alternating template as an example, although the essential ideas of the algorithm also work for other templates. Given a trace with N distinct events and L events total, we want to infer which pairs of events can satisfy the Alternating pattern. For example, a hypothetical trace $ABCACBDC$ has four distinct events: A, B, C, and D. Hence, there are 16 ways to instantiate the Alternating template: $(AA)^*$, $(AB)^*$, $(AC)^*$, $(AD)^*$,



(a) Finite state machine

Event \ Current State	0	1	2
P	1	2	2
S	2	0	2

(b) Table

Figure 2.5: Representing the Alternating template in different forms.
State 0 is the initial state and the accepting state (not shown in the table).

..., $(DC)^*$, $(DD)^*$. We want to determine which of the 16 instantiations the string $ABCACBDC$ satisfies. The only Alternating property that string satisfies is $A \rightarrow B$.

A naïve algorithm would check the string against all 16 instantiations of the property one by one. However, this algorithm does not scale when the number of distinct events becomes large. For N distinct events, there are N^2 instantiations of a pattern with two parameters. Checking the string against each instantiation needs to traverse the string once. Hence, the time complexity of this naïve algorithm is in $\Theta(N^2L)$.

Next, we introduce a more efficient inference algorithm with time complexity in $\Theta(NL)$ and space complexity in $\Theta(N^2)$.

The algorithm encodes a property template as a table. Figure 2.5(a) shows a finite state machine representation of the Alternating template. State 0 is both the initial state and the accepting state. State 2 is the error state. We can encode the transitions in this FSM as the table shown in Figure 2.5(b). The column header is the current state of the FSM. The row header is the current event in the trace. Given the current state and the current event, our algorithm decides the next state by looking up in the table. For example, if the current state is state 0 and the event is P, the next state is state 1.

```
1     void Infer( RandomAccessFile tracefile )
2
3         byte[ ][ ] ALTERNATING = { {1, 2}, {2, 0}, {2, 2} };
4         // A mapping between an event to its index
5         Hashtable event2index = new Hashtable();
6         // The number of distinct events
7         int n = 0;
8         String current_event = null;
9
10        // First pass: create event table
11        while( ( current_event = tracefile.readLine() ) != null )
12            if( !event2index.contains( current_event ) )
13                event2index.add( current_event, n++ );
14
15        // Create a table for keeping track of the states
16        // state[ i ][ j ] records the current state of the "Event_i → Event_j" template
17        byte state[ ][ ] = new byte[ n ][ n ];
18        for( int i = 0; i < n; i++ )
19            for( int j = 0; j < n; j++)
20                if( i == j )
21                    state[ i ][ j ] = 2;
22                else
23                    state[ i ][ j ] = 0;
24
25        // Second pass
26        // restart from the beginning of the trace file
27        tracefile.seek( 0 );
28        while( ( current_event = tracefile.readLine() ) != null )
29            k = event2index.get( current_event );
30            for( int i = 0; i < n; i++ )
31                // Update the state when current_event is the P event
32                state[ k ][ i ] = ALTERNATING[ state[ k ][ i ] ][ 0 ];
33                // Update the state when current_event is the S event
34                state[ i ][ k ] = ALTERNATING[ state[ i ][ k ] ][ 1 ];
35
36        // Check final state
37        for( int i = 0; i < n; i++ )
38            for( int j = 0; j < n; j++ )
39                // Is the state in an accepting state?
40                if( state[ i ][ j ] == 0 )
41                    System.out.println( i + "→" + j );
```

Figure 2.6: The inference algorithm for the Alternating pattern.

Figure 2.6 shows our inference algorithm. It scans a trace twice. In the first pass, it identifies all the distinct events and creates a mapping between the event names and integer index numbers (lines 3-13). After scanning the whole trace, the algorithm creates state, an $N \times N$ array, for keeping track of the states of the instantiations of the Alternating pattern (line 17). The value of state[i][j] corresponds to the FSM state of the instantiation in which P is $Event_i$ and S is $Event_j$. The elements of this array except the diagonal ones are initialized to 0 since all FSMs start in the initial state (lines 18-23). The diagonal elements are initialized to 2 (i.e., the error state) because the two events cannot be equal (line 21). In the second pass, our algorithm rescans the execution trace (line 27). When it reads an event from the trace, it updates the state array (lines 28-34). Here the key observation is that an event, $Event_k$, could be either the P event or the S event. If $Event_k$ is the P event, our algorithm updates the k-th row of the state array (line 32). If $Event_k$ is the S event, our algorithm updates the k-th column of the state array (line 34). Our algorithm updates the state by looking up the pre-encoded tables of the FSMs (Figure 2.5).

After scanning the trace twice, if state[i][j] is in an accepting state, our algorithm outputs $Event_i \rightarrow Event_j$ as a satisfied Alternating property (lines 37-41).

This algorithm has time complexity in $\Theta(NL)$ and space complexity in $\Theta(N^2)$. In the loop from line 11 to 13, it scans the trace once. For each new event, the algorithm updates the event2index mapping, which is a constant-time computation. Therefore, the time complexity of the loop is in $\Theta(L)$. The loop from line 18 to 23 has time complexity in $\Theta(N^2)$. In the loop from line 28 to 34, the algorithm updates one row and one column of the state array for each event in the trace. Hence, the time complexity of the loop is in $\Theta(NL)$. Finally, the loop from line 37 to 41 has time complexity in $\Theta(N^2)$. As a result, the total time complexity of our algorithm is in $\Theta(L + N^2 + NL + N^2)$. Because $L \geqslant N$, the time complexity of the algorithm is in $\Theta(NL)$. The algorithm requires creating an $N \times N$ array. Therefore, its space complexity is in $\Theta(N^2)$.

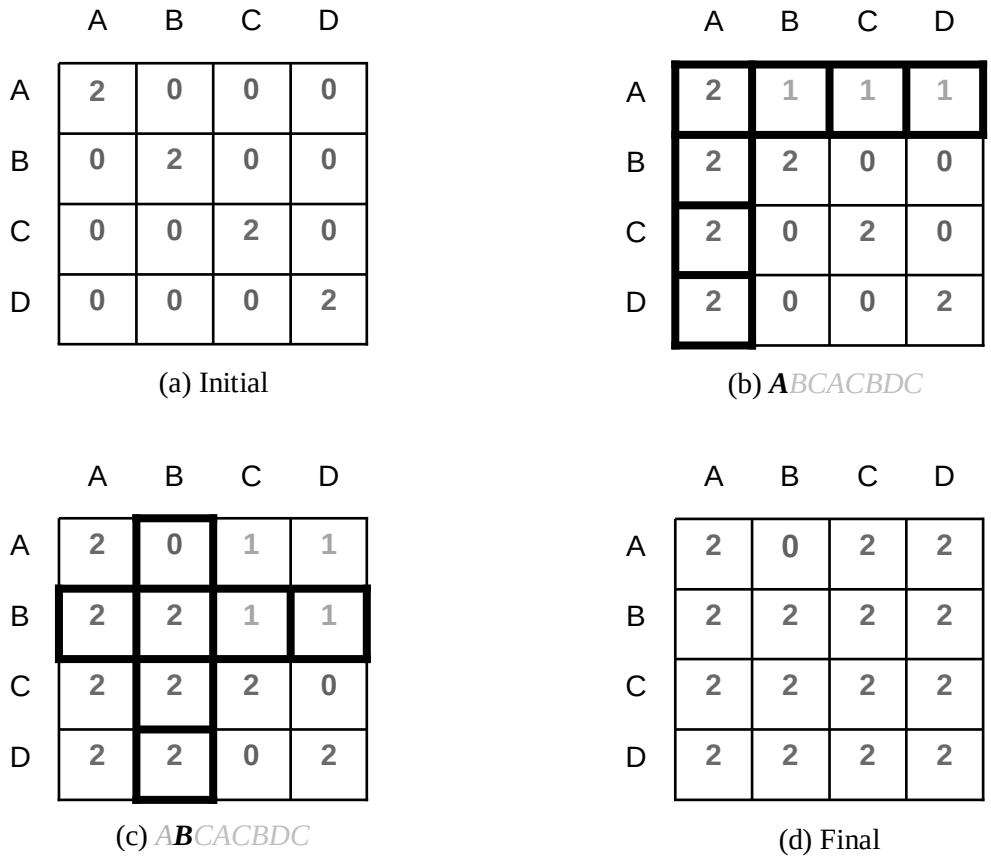


Figure 2.7: Inferring Alternating properties from a hypothetical trace $ABCACBDC$.

Enter:Buffer.stop():[main] → Exit:Consumer.run():[Thread-1]
Exit:Buffer.stop():[main] → Exit:Consumer.run():[Thread-1]

Enter:Buffer.add():[main] → Exit:Buffer.add():[main]
Enter:Consumer.run():[Thread-1] → Exit:Consumer.run():[Thread-1]
Enter:Producer.main():[main] → Exit:Producer.main():[main]
Enter:Buffer.stop():[main] → Exit:Buffer.stop():[main]
Enter:Buffer.take():[Thread-1] → Exit:Buffer.take():[Thread-1]

Enter:Producer.main():[main] → Enter:Consumer.run():[Thread-1]
Enter:Producer.main():[main] → Enter:Buffer.stop():[main]
Enter:Producer.main():[main] → Exit:Buffer.stop():[main]
Enter:Producer.main():[main] → Exit:Consumer.run():[Thread-1]
Enter:Consumer.run():[Thread-1] → Enter:Buffer.stop():[main]
Enter:Consumer.run():[Thread-1] → Exit:Buffer.stop():[main]
Enter:Consumer.run():[Thread-1] → Exit:Producer.main():[main]
Enter:Buffer.stop():[main] → Exit:Producer.main():[main]

Figure 2.8: Alternating properties inferred from the trace in Figure 2.3.

Suppose we have a hypothetical trace $ABCACBDC$ with four distinct events: A, B, C, and D. Figure 2.7 shows how our inference algorithm infers which of the 16 instantiations of the Alternating template the trace satisfies. Initially, the algorithm creates a 4×4 array (Figure 2.7 (a)). Every cell of this array stores the current state of the corresponding pair of events. Our algorithm initializes all except the diagonal cells to 0 – the initial state of the Alternating template. It sets the diagonal cells to the error state (2 in this case) because the template does not allow the two events to be equal. Next, the algorithm scans the first event, A (Figure 2.7 (b)). Because A can be either the P event or the S event in the Alternating template, our algorithm updates the first row of the array in which A is the P event and the first column of the array in which A is the S event. For example, for the cell (A, B), where A is the P event and B is the S event, the algorithm updates the cell's value from 0 to 1 because A is the P event. After reading the next event from the trace, B, our algorithm updates the second row and second column in the array (Figure 2.7 (c)). The state array after scanning the whole trace is shown in Figure 2.7 (d). Only the cell corresponding to $A \rightarrow B$ is in the accepting state 0. As a result, the algorithm outputs that $A \rightarrow B$ satisfies the Alternating template.

The Producer-Consumer trace in Figure 2.3 (see page 10) has 10 distinct events. Hence, there

are 100 candidate Alternating properties. The inference algorithm determines that the trace satisfies the 17 Alternating properties shown in Figure 2.8. The two properties in boldface represent the property that whenever the Producer sends out the stop signal, the Consumer will stop execution eventually. The next five properties are uninteresting because they correspond to the trivial fact that entering and exiting a method always alternate. The remaining properties reflect the static call graph and are not very interesting either (e.g., Producer.main() calls Consumer.run()). Section 2.7 describes the post-processing component that selects interesting properties, eliminates redundant and uninteresting properties, and better presents the results.

2.4.3 Inferring the Strictest Pattern

Section 2.4.1 introduced a hierarchy of seven property templates whose strictness forms a partial order as shown in Figure 2.4 (see page 13). Given a string, we would like to know the strictest template it can satisfy. To achieve this, we first determine whether the string satisfies the three primitive templates: OneCause, CauseFirst, and OneEffect. Then we deduce the strictest pattern the string satisfies using the logical relationship among the patterns. We extend the algorithm for a single pattern (Figure 2.6) to simultaneously track the three primitive patterns: OneCause, CauseFirst, and OneEffect. After the algorithm finishes inferring the primitive patterns, it collates the state array for each of the three primitive patterns and deduces the strictest pattern.

2.5 Approximate Inference

The algorithm in Figure 2.6 only infers properties that are completely satisfied by the trace. For example, P and S satisfy the Alternating template in $PSPSPSPSPSPSPSPSPSPS$ but not in $PSPSPSPSPSPSPSPSPSPP$ because the last P does not have a corresponding S. It is, however, apparent that P and S satisfy the Alternating template for most parts of the second trace. This 100% satisfaction requirement is a big limitation of the original algorithm when applied to

```
1       int i=0;
2       while(true){
3           lock.acquire();
4           // do something here
5           ……
6           if(++i>10)
7               break;
8           lock.release();
9       }
```

Figure 2.9: A hypothetical program that forgets to release a lock.

traces from real systems. This algorithm would miss many interesting properties because the traces are imperfect.

2.5.1 Imperfect Traces

An *imperfect trace* is a trace that contains event sequences that violate a property specification that is necessary for the correctness of a system.

Bugs in a program are the most insurmountable reason for imperfect execution traces. The hypothetical buggy program in Figure 2.9 illustrates this. Suppose the code on line 5 has no side effect on either i or $lock$. The while loop exits when i becomes greater than 10. During each loop except the last one, the loop body acquires a lock, does some work, and releases the lock. On the last loop, however, the program does not release the lock. As a result, executing the program would produce a trace $PSPSPSPSPSPSPSPSPSP$, where P represents lock.acquire and S represents lock.release. Although this trace does not satisfy the $P \rightarrow S$ property, it is clear that $P \rightarrow S$ is the dominant pattern in the trace.

In addition to buggy programs, sampling can also cause imperfect traces. Monitoring the complete execution of long running programs such as operating systems is impractical. In practice, users often sample partial execution by recording either the complete run-time data for a short period or randomly selected data for the whole execution. In either case, the trace does not capture the whole execution. For example, the instrumentor samples the acquisition and release of locks. As a

result, the execution trace might miss some lock acquisition or release events, thus it will not satisfy the needed Alternating properties. Alternatively, the instrumentor might only record the first 10000 function calls when executing the program. Consequently, the trace might miss some lock release events that fall out of the first 10000 function calls.

Finally, instrumentation tools have some limitations on the data they can capture. For example, an instrumentor might not be able to record argument values and return values. So the execution trace can miss information such as the identity of a lock. When there are multiple instances of a lock, the identity serves as a way to differentiate operations on the locks. If a trace does not have the identity of a lock or other ways to distinguish different locks, all calls to acquire or release different locks will appear to be the same events and hence violate the Alternating property.

2.5.2 Detecting the Dominant Behavior

To deal with the imperfect traces, we adapt the algorithm from Figure 2.6 (page 16). The new algorithm decides what fraction of an execution trace satisfies a property template instead of just determining satisfaction of a property template. The new algorithm partitions the original trace into subtraces, decides whether each subtrace satisfies a pattern, and computes the fraction of the subtraces that satisfy a pattern.

In general, we can define a subtrace using a regular expression. We call this regular expression the *monitor template* and its finite state machine the *monitor FSM*. We call the finite state machine of a property template (e.g., Alternating) the *property FSM*. We use $P_{property\ template}$ to represent the percentage of a trace that satisfies a property template. For example, $P_{Alternating}$ indicates the percentage of the subtraces that satisfy the Alternating template.

One intuitive definition of a subtrace is P^+S^+. In addition, the leading S events form a subtrace and so do the trailing P events. For example, we can partition $PSPSPSPSPSPSPSPSP$ into 10 subtraces $PS, PS, PS, PS, PS, PS, PS, PS, PS, P$. The first nine subtraces satisfy the

```
1   while (not at the end of the trace)
2       read the next event from the trace
3       update the property FSMs
4       update the monitor FSM
5       if (the monitor FSM is in an accepting state)
6           Increase the counter of the monitor FSM by one
7           for each property FSM
8               if (the property FSM is in an accepting state)
9                   increase the counter of the property FSM by one
10              reset the property FSM to its start state
```

Figure 2.10: The approximate inference algorithm.

$P \rightarrow S$ property, whereas the last one does not. Therefore, 90% of the subtraces satisfy $P \rightarrow S$ and $P_{Alternating}$ is 0.90.

The new algorithm tracks one monitor FSM and one or more property FSMs. As with the original algorithm, it scans the execution trace twice. The two algorithms differ in how they update the states during the second scan. Figure 2.10 shows the pseudo-code of the new algorithm. When the new algorithm reads an event (line 2), it first updates all the property FSMs (line 3) in the same way as the original algorithm (lines 25-29 in Figure 2.6 on page 16). Then the new algorithm updates the state of the monitor FSM (line 4). If the monitor FSM reaches one of its accepting states (line 5), this indicates the end of a subtrace. The new algorithm increases the counter of the monitor FSM by one (line 6) and then checks whether the property FSMs reach their accepting states (line 7-8). If a property FSM is in an accepting state, this subtrace satisfies the property FSM and hence the new algorithm increases the property FSM's counter by one (line 9). Finally, the new algorithm resets the property FSMs to their starting states (line 10) before analyzing the next subtrace.

After scanning the trace, the new algorithm examines the counters of the property FSMs and outputs the satisfaction percentage.

This new algorithm requires one more $N \times N$ array to store the counters of the property FSMs and another $N \times N$ array to track the state of the monitor FSM. Hence, the space complexity of the new algorithm is still in $\Theta(N^2)$, same as the original algorithm. In addition to updating the

property FSMs' states, the new algorithm updates the monitor FSM, increases the counters, and resets the property FSMs. However, these additional steps only increase the time complexity by constant factors. Therefore, the new algorithm also has the time complexity in $\Theta(NL)$, same as the original algorithm.

2.6 Handling Context Information

A major advantage of dynamic analysis over static analysis is the ready availability of precise context information including threads, objects, argument values, and return values. This section presents our techniques for using context information to infer more precise properties.

We use three general approaches: *context-neutral*, *context-sensitive*, and *context-slicing*. The context-neutral approach treats two events with same static signature but different context information as the same event, whereas the context-sensitive approach considers them as two distinct events.

For example, consider the example trace in Figure 2.11. The context-neutral approach sees two distinct events (lock.acq and lock.rel), but the context-sensitive approach sees four events (lock1.acq, lock1.rel, lock2.acq, and lock2.rel). Context-neutral analysis does not infer lock.acq $\rightarrow$ lock.rel. On the other hand, context-sensitive analysis infers six Alternating properties, only two of which are useful (shown in boldface in Figure 2.11). Neither context-sensitive nor context-neutral analysis infers that lock.acq and lock.rel alternate for a same lock object. To infer this property, we need to generalize the results of the context-sensitive analysis by slicing the original trace into separate traces based on object identity. We call this the *context-slicing* approach. Context-slicing produces two traces from which our inference algorithm infers lock.acq $\rightarrow$ lock.rel. In addition to slicing on object identities, context-slicing can also slice thread identities, argument values, and return values.

The results of context-sensitive analysis are the most complete, but are not useful without gen-

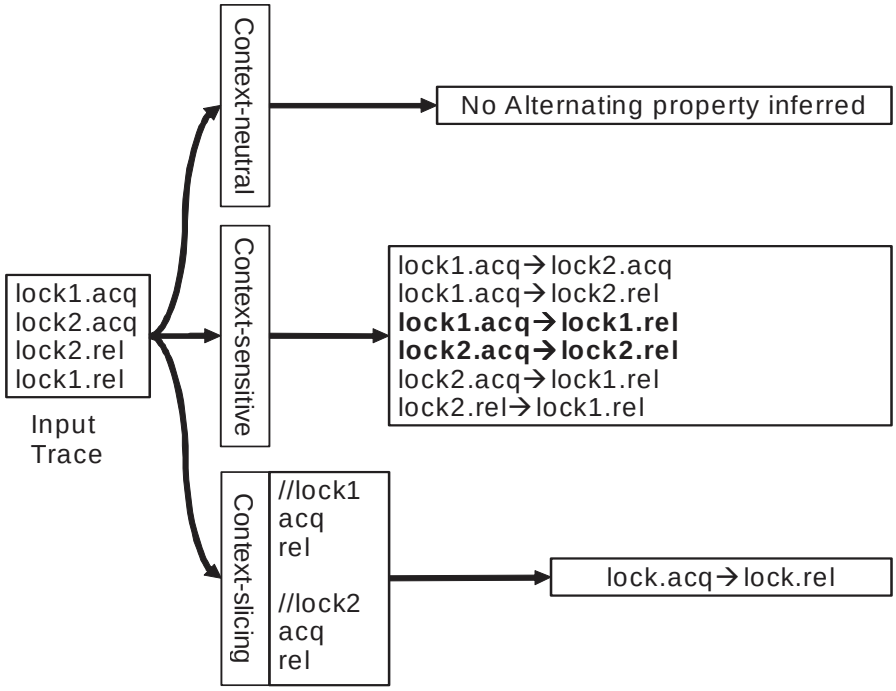


Figure 2.11: Context handling techniques.

Context-neutral does not differentiate between *lock1.acq/rel* and *lock2.acq/rel* and does not infer any Alternating property. *Context-sensitive* differentiates between the methods of *lock1* and *lock2* and infers six Alternating properties. However, only two of the six properties shown in boldface correspond to the property that acquiring a lock should alternate with releasing a lock. *Context-slicing* slices the original trace by the identity of the lock and produces two subtraces. Hence, *context-slicing* infers lock.acq → lock.rel.

eralization. Context-slicing is a simple way to generalize the results of context-sensitive analysis. A limitation of context-slicing is that it cannot detect properties that involve more than one context. For example, if slicing is done by threads, an Alternating pattern between P in one thread and S in another thread would not be detected.

2.7 Property Selection

When processing a large trace that has many distinct events, our inference technique typically infers thousands of properties, which are too many to be effectively used in practice. Hence, one big challenge is to select a small fraction of interesting properties. An *interesting* property is a property for which developers are likely to make mistakes and violation of which would produce bad consequences. For example, we consider properties of critical system resources such as locks and transactions to be interesting. Such properties are important because violating them can have serious consequences such as causing system crashes [Ball01, Das02] and opening security vulnerability [Chen02]. Next, Sections 2.7.1 and 2.7.2 introduce two heuristics for selecting interesting properties.

In addition to selecting interesting properties, it is also important to present the results so that users can easily understand them. Section 2.8 presents a chaining method for presenting the inferred properties.

2.7.1 Static Call Graph Based Heuristic

This subsection introduces a heuristic for identifying interesting properties based on a program's static call graph [Grov01]. The key observation is that a property is more likely to be interesting when the two events it involves are not reachable in the static call graph. Figure 2.12(a) illustrates this idea. Suppose our inference technique infers two Alternating properties: KeSetTimer → KeSetTimerEx and ExAcquireFastMutexUnsafe → ExReleaseFastMutexUnsafe. In the first property, Ke-

```
void KeSetTimer( ) {
   KeSetTimerEx( );
}
```

```
void X( ) {
   …
   ExAcquireFastMutexUnsafe(&m);
   …
   ExReleaseFastMutexUnsafe(&m);
   …
}
```

(a) A concrete example

```
A( ) {
   …
   B ( );
   …
}
```

```
X( ) {
   …
   C ( );
   …
   D ( );
   …
}
```

(b) Abstract form

Figure 2.12: Two scenarios of static call graph.

SetTimer is a wrapper of KeSetTimerEx. Therefore, whenever KeSetTimer is called, KeSetTimerEx must also be called. In the second property, ExAcquireFastMutexUnsafe and ExReleaseFastMutex-Unsafe do not call each other and therefore, their executions are asynchronous.

Figure 2.12(b) shows an abstract form of the two scenarios. $A \rightarrow B$ and $C \rightarrow D$ are Alternating properties. The second property is usually more interesting than the first one for two reasons. The first property represents a trivial relationship that can be easily discovered by constructing a static call graph. On the other hand, the second property captures two events that do not have obvious static relationship. In addition, the second property captures a protocol between two functions, which developers are more likely to forget. The developers are obligated to call the pair of functions together. In contrast, the first property represents a delegation relationship between the two functions, which does not suggest any obligation on the developers.

2.7.2 Naming Similarity Heuristic

The second heuristic exploits the naming conventions used in many real systems. For example, Microsoft uses the Hungarian Notation [Simo]. The key observation behind the naming similarity

heuristic is that a property is more likely to be interesting if the events it involves have similar names. For example, KeAcquireInStackQueuedSpinLock and KeReleaseInStackQueuedSpinLock only differ by one word and clearly appear to represent a desirable locking discipline relationship.

The naming similarity heuristic partitions the two events' names into words and then computes a similarity score of the two names. For a property $A \rightarrow B$, where A has w_A words, B has w_B words, and A and B have w words in common, this heuristic computes the similarity score of A and B as

$$Similarity_{AB} = \frac{2w}{w_A + w_B}$$

We distinguish words by capital letters or underscores. For example, KeAcquireInStackQueued-SpinLock has seven words, and so does KeReleaseInStackQueuedSpinLock (i.e., Ke, Acquire/Release, In, Stack, Queued, Spin, and Lock). Therefore, $w_A = w_B = 7$. There are six common words and so $w = 6$. As a result, the similarity score of these two names is 85.7%.

The naming similarity heuristic is especially effective for selecting properties that tend to involve events with similar function names. For example, locking disciplines and resource allocation/deletion protocols usually have a pair of functions with similar names. Although this heuristic does eliminate some interesting properties that involve events with very different names, our goal is to increase the density of interesting properties in the results. The experimental results (Section 3.4) demonstrate that the naming heuristic is effective for realizing this goal.

2.8 Chaining Method

The chaining method presents the inferred Alternating properties in a condensed form so that users can gain a better picture of how a system works. Section 2.8.1 introduces the *property graph* that is a graph representation of a set of Alternating properties, proves several important properties of the property graph, and defines the *chaining problem*. Section 2.8.2 proves that the chaining

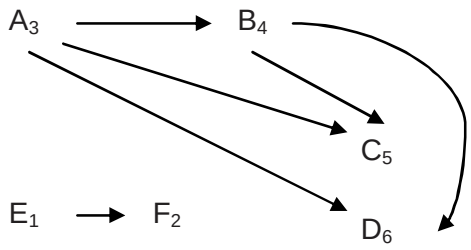


Figure 2.13: Alternating Chains.

problem is in NP-complete. Section 2.8.3 describes a brute force algorithm for solving the chaining problem. As explained later, this algorithm performs well in our experiments due to the low density of the property graphs.

2.8.1 Property Graph

We map a set of Alternating properties to a directed graph $G =< V, E >$, where V is a set of nodes and E is a set of edges. Each distinct event corresponds to a node in V. Therefore, $|V| = N$. An edge from node i to node j corresponds to an Alternating property $Event_i \rightarrow Event_j$. We call G the *property graph*. Figure 2.13 shows a property graph involving six events and six Alternating properties (the subscripts correspond to the topological numbers explained later in this subsection).

If all the Alternating properties in a property graph have $p_{AL} = 1.0$, the property graph is a directed acyclic graph (DAG). In other words, for any i and j, if there is a path from node i to node j, there does not exist a path from node j to node i. To prove this, we first prove the following lemma: if there is a path from node i to node j, the first $Event_i$ must occur before the first $Event_j$ in the trace.

We prove the lemma by induction on l, the length of a path from node i to node j. The base case is $l = 1$. In this case there is an edge from node i to node j. According to the definition of the property graph, $Event_i \rightarrow Event_j$ is true. Therefore, according to the definition of the Alternating property, the first $Event_i$ must occur before the first $Event_j$ in the trace. For $k > 1$, assume that

the lemma holds for all l where $1 \leqslant l < k$. Next, we prove the lemma when $l = k$. Suppose there is a path from node i to node j and this path's length is k. If the next to last node on this path is node p, there is a path from node i to node p and this path's length is $k - 1$. According to the induction hypothesis, the first $Event_i$ must occur before the first $Event_p$ in the trace. In addition, there is an edge from node p to node j. According to the induction hypothesis, the first $Event_p$ must occur before the first $Event_j$ in the trace. As a result, the first $Event_i$ must occur before the first $Event_j$ in the trace.

Using the above lemma, we prove by contradiction that a property graph is a DAG. If a property graph G is not a DAG, then G contains a cycle. There must exist two nodes, i and j, such that there exists a path P from i to j and a path P' from j to i. According to the above lemma, the existence of P implies that the first $Event_i$ must occur before the first $Event_j$ in the trace. In addition, the existence of P' implies the first $Event_j$ must occur before the first $Event_i$ in the trace, which is a contradiction. Hence, a property graph is a DAG.

When a property graph includes Alternating properties with $p_{AL} < 1.0$, it might still be a DAG. As explained later, our chaining algorithm first checks if a property graph (with properties whose $p_{AL} < 1.0$) has any cycles. Our algorithm only performs the chaining operation when the property graph is a DAG. The rest of this section assumes a property graph is a DAG.

A *topological number* of a node in a DAG is an integer such that if there exists an edge from node i to node j, then the topological number of node i is less than the topological number of node j [Cormen01]. Because a property graph is a DAG, we can sort the nodes based on their topological numbers. The subscript of each node in Figure 2.13 indicates its topological number.

An *alternating chain* is a subgraph $G' =< V', E' >$ of a property graph $G =< V, E >$, where $V' \subseteq V$ and $E' = \{(i, j) \mid i, j \in V' \text{ and } (i, j) \in E\}$, such that if the topological number of node i is less than the topological number of node j, then $(i, j) \in E'$. By definition, an edge in a property

graph is a trivial alternating chain. For example, in Figure 2.13, the subgraph consisting of A and B is a trivial alternating chain. In addition, the subgraph consisting of A, B, and C is an alternating chain, while the subgraph consisting of A, B, C, and D is not an alternating chain.

A maximal alternating chain is an alternating chain $G' =< V', E' >$ of a property graph $G =< V, E >$ such that $\forall i \in V - V'$, $G'' =< V'', E'' >$ is not an alternating chain, where $V'' = V' \cup \{i\}$ and $E'' = E' \cup \{(i,j) \mid j \in V' \text{ and } (i,j) \in E\} \cup \{(j,i) \mid j \in V' \text{ and } (j,i) \in E\}$. For example, in Figure 2.13, the subgraph consisting of A, B, and C and the subgraph consisting of E and F are maximal alternating chains.

The *chaining problem* is the problem that, given a property graph G and an integer k, determine whether there exists an alternating chain of k nodes in G. The *chain enumeration problem* is the problem that, given a property graph G, identify all the maximal alternating chains of G.

2.8.2 Chaining is in NP-Complete

In this section, we prove that the chaining problem is an NP-Complete problem.

Given any subgraph of a property graph, we can easily verify in polynomial time whether the subgraph is an alternating chain. Hence, the chaining problem is in NP.

We prove the NP-hardness of the chaining problem by reducing the clique problem to the chaining problem. The *clique problem* in graph theory states "for an undirected graph G and an integer k, does G have a complete subgraph that has k nodes" [Karp72]. The clique problem is a well-known NP-Complete problem. Next we construct a polynomial-time reduction from the clique problem to the chaining problem.

Given an undirected graph G, we can convert all of its undirected edges to directed edges in polynomial time. We call the resulting directed graph G'. The conversion works by following the standard black-gray-white color depth-first-search (DFS) algorithm [Cormen01]. Figure 2.14 shows the transformation algorithm. Initially, all the nodes are marked as white (lines 7-8). Next,

```
1     int total;                  // the number of nodes
2     int undirected[ ][ ];   // total x total, an undirected graph with no self-loop
3     int directed[ ][ ];       // total x total
4     int color[ ];               // total
5
6     void undirected2directed ( )
7        for ( int i = 0; i < total; i++ )
8           color[ i ] = WHITE;
9        for ( int i = 0; i < total; i++ )
10          if (color [ i ] == WHITE)
11             DFS ( i );
12
13    void DFS ( int i )
14       color [ i ] = GRAY;
15
16       for ( int j = 0; j < total; j++ )
17          if ( undirected [ i ][ j ] == CONNECTED )
18             switch ( color [ j ] )
19                case WHITE:
20                   directed [ i ][ j ] = CONNECTED;
21                   DFS ( j );
22                   break;
23                case GRAY:
24                   directed [ j ][ i ] = CONNECTED;
25                   break;
26                case BLACK:
27                   break;
28
29       color [ i ] = BLACK;
```

Figure 2.14: Algorithm for transforming an undirected graph to a DAG.

the algorithm checks each node and performs DFS for white nodes (lines 9-11). DFS is a recursive function that takes a node i to be explored as its argument (line 13). The node that is currently being explored is marked as gray (line 14). Next, all the adjacent nodes of i are explored (lines 16-17). If the adjacent node j is a white node, the algorithm converts the corresponding edge to a directed edge from i to j and continues to explore j (lines 19-22); if j is a gray node, the algorithm converts the corresponding edge to a directed edge from j to i (lines 23-25); if j is a black node, the algorithm does not do anything (lines 26-27). After all the adjacent nodes of i have been explored, node i is marked as black (line 29). The time complexity of the conversion algorithm is polynomial because DFS is in $\Theta(|V| + |E|)$.

Now we prove by contradiction that the resulting graph G' is a DAG. Suppose there is a cycle, $i, ..., j, i$, in G'. Without loss of generality, let us assume that i is the first node in the cycle that is explored during the conversion process. In other words, when i changes color from white to gray, the other nodes on the cycle are still white. According to the algorithm, there are two ways the edge from j to i can be created. First, when the adjacent nodes of j are explored (hence, j is gray), i (one of j's adjacent nodes) is white (lines 19-22). This case is impossible due to our assumption that i is the first node on the cycle that has gray color. Second, when the adjacent nodes of i are explored, j is gray too (lines 23-25). So the algorithm creates an edge from j to i (line 24). This also contradicts our assumption that j is white when i is gray. Therefore, it is impossible to have an edge from j to i in G', which contradicts the last edge on the cycle.

Given an instance of the clique problem (determining whether an undirected graph G has a complete subgraph with k nodes), the above transformation converts the clique problem to an instance of the chaining problem (determining whether the resulted DAG G' has an alternating chain with k nodes). Next we show that if we have a solution to the chaining problem, we can use it to solve the clique problem. We use $[i, j]$ to represent a directed edge from i to j and (i, j) to represent an undirected edge between i and j. Suppose our algorithm determines that G' has an alternating chain, $X' =< V_{X'}, E_{X'} >$ with k nodes. The counterpart of X' in G is $X =< V_X, E_X >$, where $V_X = V_{X'}$ and $E_X = \{(i, j) | i, j \in V_X, [i, j] \text{ or } [j, i] \in E_{X'}\}$. According to the definition of Alternating Chain, $\forall i, j \in V_{X'}$, $[i, j]$ or $[j, i] \in E_{X'}$. Therefore, $\forall i, j \in V_X$, $(i, j) \in E_X$. Hence, X is a clique with k nodes in G.

As a result, if we have a solution to the chaining problem, we can use it to solve the clique problem. Since the clique problem is in NP-hard, the chaining problem is also in NP-hard.

We already proved the chaining problem is in NP, so the chaining problem is in NP-Complete.

2.8.3 The Chaining Algorithm

Assume $P \neq NP$, no algorithm can have a better worst-case performance than a brute force algorithm. The performance of a brute force algorithm is highly dependent on the structure of the property graph. For a directed graph with n nodes and m edges, we compute its edge density as $2m/n^2$. Intuitively, a brute force algorithm generally performs better on a sparse property graph than a dense property graph because there are fewer edges (and so much fewer combinations of edges) to try in a sparse property graph. In our experiments, all of our property graphs are very sparse with densities around 10%. Next we present a brute force algorithm whose worst-case performance is exponential. The running time scales with the edge density, which we have found to be low in practice. We have applied our chaining algorithm to analyze a property graph with 91 nodes and 490 edges in less than one minute.

Given a property graph G, we can convert all its directed edges to be undirected. We call the resulted graph $G_{undirected}$. If a subgraph C of G is an alternating chain, then the corresponding subgraph $C_{undirected}$ of $G_{undirected}$ must be a clique, and vice versa. Finding all the maximal alternating chains in G, therefore, can be solved by finding all the maximal cliques in $G_{undirected}$.

Our chaining algorithm first identifies all the connected components in $G_{undirected}$ using a depth-first-search [Cormen01]. Then the algorithm identifies the maximal cliques in each connected component. To convert a maximal clique to an alternating chain, we output the clique in topological order. Figure 2.15 shows the chaining algorithm.

The chaining algorithm is a work-list algorithm. The algorithm first creates a worklist (line 5). The worklist stores the cliques that will be examined. Each clique is represented as a set of nodes. The worklist is a set of sets of nodes, initially containing one set for each edge in the graph consisting of the endpoints of that edge (line 6). The algorithm removes the first clique, clq, from the worklist (line 8). For each node of the current connected components, the algorithm first tests if it is in clq

```
1    for each cc in G_undirected
2        chaining ( cc );
3
4    chaining ( cc )
5        Vector worklist = new Vector();
6        add all the edges of cc to worklist;
7        while ( worklist is not empty )
8            clq = worklist.remove ( 0 );
9            boolean cannotExpand = true;
10           for each node i of cc
11               if ( i is in clq )
12                   continue;
13               if ( there is an edge between i and each node of clq )
14                   newclq = clq U { i };
15                   cannotExpand = false;
16                   for each element x of worklist
17                       if ( x is a subset of newclq )
18                           remove x from worklist;
19                   insert newclq to the beginning of worklist;
20                   break;
21           if ( cannotExpand )
22               print out clq in topological order;
```

Figure 2.15: The chaining algorithm.

(lines 11-12). For a node i that is not in clq, the algorithm tests if there is an edge between i and each node of clq (line 13). If so, the algorithm expands clq to newclq (line 14), removes all the cliques in worklist that are subset of newclq (lines 15-18), and inserts newclq to the beginning of worklist (line 19). The boolean variable, cannotExpand, indicates whether clq can be expanded (lines 9 and 15). If cannotExpand is true after processing clq, the algorithm outputs clq according to the topological order in the directed property graph (lines 21-22).

Next we analyze the complexity of our chaining algorithm. Suppose a connected component, cc, has n nodes and m maximal cliques. Suppose the ith maximal clique has n_i nodes. In order to construct the maximal cliques, the while loop (lines 8-22) executes $\sum_{i=1}^{m} n_i$ times because the clique grows one node in each loop. The for loop (lines 11-20) executes n_i times. During each for loop, the algorithm needs to check up to n_i edges (line 13). In addition, when a larger clique, newclq, is formed, the algorithm needs to remove redundant cliques from the worklist (lines 16-18).

Chain #1
Enter:Buffer.add():[main] →
Exit:Buffer.add():[main]

Chain #2
Enter:Buffer.take():[Thread-1] →
Exit:Buffer.take():[Thread-1]

Chain #3
Enter:Producer.main():[main] →
Enter:Consumer.run():[Thread-1] →
Enter:Buffer.stop():[main] →
Exit:Buffer.stop():[main] →
Exit:Producer.main():[main] →
Exit:Consumer.run():[Thread-1]

Figure 2.16: Alternating chains for the Producer-Consumer program.

A clique, x, is redundant if it is a subset of newclq. When newclq has p nodes, in worst-case, the number of redundant cliques is in $\Theta(C_{n_i}^{p-1})$. Therefore, the for loop is in $\Theta(n_i(n_i + C_{n_i}^{p-1}))$. To construct the ith maximal clique, the algorithm's complexity is in $\Theta(\sum_{p=1}^{n_i}(n_i(n_i + C_{n_i}^{p-1})))$ = $\Theta(n_i 2^{n_i} + n_i^3)$. Therefore, the worst-case complexity of our chaining algorithm is still exponential. The complexity of the chaining algorithm varies by the density of the property graph. In practice, the density of the property graph tends to be small (typically around 10% in our experiments). Therefore, the performance of our algorithm is acceptable in practice.

Figure 2.16 shows the three chains constructed for the Producer-Consumer program. The longest chain (#3) has six events and represents an important property: after the Producer calls the stop method, the Consumer eventually stops. Each of the other two chains has only two events and is uninteresting because these Alternating properties correspond to the trivial fact that entering a method alternates with exiting the method.

2.9 Perracotta

This section describes the prototype implementation of our inference approach. We adapted a Java instrumentation tool (Section 2.9.1) and implemented the inference engine in a prototype tool

called Perracotta (Section 2.9.2). In addition to implementing the inference algorithm, Perracotta also implements the two heuristics, the chaining method, and the context-handling techniques.

2.9.1 Instrumentation

To instrument Java programs, we adapted the Java Runtime Analysis Toolkit (JRat) [JRat]. JRat has two important components: an instrumentor and a runtime system. The instrumentor component uses the Byte Code Engineering Library (BCEL) to parse and insert hooks into Java bytecode [BCEL]. When an instrumented Java application executes, the hooks generate events for method entrances, method exits, and exceptions. The JRat runtime system processes the events by delegating them to one or more handlers. Different handlers process the events in different ways. JRat provides an event handler Service Provider Interface (SPI). Users can develop their own handlers by implementing this SPI and can configure the handlers that the JRat runtime system uses by either setting an environment variable or supplying a configuration file. We developed an event handler for collecting method execution traces. For each Java method, our handler records its entrance, exit, signature, arguments, return value, and any exceptions generated. If an argument is of a primitive type, the handler outputs its value. If an argument is of an object type, the handler outputs its hashcode as its object ID.

For small C programs, we manually instrument the source code to monitor function calls. For Windows kernel, we use a Vulcan-based instrumentor [Sriva01]. This instrumentor works on x86 binaries and can monitor function calls and thread information. However, this instrumentor cannot monitor argument values and return values.

2.9.2 Inference Engine

Perracotta implements the inference engine and post-processing components. Perracotta is implemented using 12000 lines of Java code. Perracotta can be run in two modes: the strictest pattern mode for inferring the strictest pattern and the approximate mode for inferring properties whose

satisfaction ratio is greater than a threshold between 0.0 and 1.0. In addition, it also has an interface that accepts user-specified templates and determines the satisfaction ratios. To implement the static call graph based heuristic for Java, Perracotta has a module for computing the static call graph. For C/C++, it accepts static call graph as a textual file. In addition, Perracotta can eliminate uninteresting properties based on the naming similarity based heuristic. Furthermore, Perracotta implements the chaining method for the Alternating properties as presented in Section 2.8.3.

Perracotta also provides several utility programs that slice traces as described in Section 2.6. The *ThreadSlicer* slices the trace into subtraces based on the thread identities. Similarly, the *ObjectSlicer* slices the trace into subtraces based on the object identities.

Chapter 3
Inference Experiments[1]

This chapter presents experiments applying Perracotta to several real programs. We evaluate the usefulness of the inferred properties in revealing important information of a program. Chapter 4 investigates other uses of the inferred properties including program verification (Section 4.1) and program differencing (Section 4.2).

Static analysis is still very limited in achieving both scalability and precision. On the other hand, dynamic analysis can access precise runtime information such as branches taken, threads, and pointers, which is beneficial for understanding a complex program. Complex systems are hard to understand in that it is challenging to recognize delocalized plans [Leto86]. *Delocalized plans* are "programming plans realized by lines scattered in different parts of the program." [Leto86] Programmers often only rely on local information when making changes to a system, which fails to reveal the interaction between specific pieces of code and other pieces of code or data some "distance" away [Leto86, Corb89]. We hypothesize that our dynamic temporal specification inference technique can effectively detect certain delocalized plans.

In order to test the hypothesis and also evaluate the scalability, accuracy, and efforts of using our inference technique, we conducted several case studies on a wide range of programs. Table 3.1 summarizes the characteristics of these six testbeds. These programs are from a diverse range of application domains, are written in Java, C, or C++, and range from small prototypes with several hundred lines of code to large products with millions of lines of code. The programs are:

1. *Producer-consumer*, a Java implementation of the classic synchronization problem. We use Producer-Consumer as a running example in Chapter 2.

[1] This chapter is partly based on [Yang04a] and [Yang06].

Table 3.1: Characteristics of Testbeds.

Name	Category	Language	Size	Maturity	Temporal Specifications Available?	Additional Experiments
Producer-Consumer	Classic synchronization problem	Java	59	Prototype	Yes, in English	
Bus Simulator	Student multi-threaded program	C/C++	259	Prototype	Yes, in English	Program differencing
Daisy	Unix-like file system	Java	2K	Prototype	Limited	Program verification
OpenSSL (Handshake Protocol)	Network protocol	C	32K (418)[1]	Production	Yes, SSL specification in English and FSM	Program differencing
JBoss (Transaction Management)	Network middleware	Java	1M (7K)[2]	Production	Yes, JTA specification in English and FSM	
Windows Vista Kernel APIs	OS kernel	C/C++	50M	Production	Limited, MSDN and MS internal document including SLAM	Program verification

1. In OpenSSL version 0.9.7d, the implementation of the SSL specification (i.e., **.c* and **.h* files in the *ssl* directory) has thirty-two thousand lines and the implementation of the SSL handshake protocol on the server's side (i.e., the *ssl3_accept* function in the *s3_srvr.c* file) has 418 lines.

2. In JBoss version 4.0.2, all the **.java* files have one million lines. The transaction management module (i.e., all the **.java* files belonging to the *org.jboss.tm* package) has seven thousand lines.

2. *Bus Simulator*, a collection of student submissions for an assignment of implementing a multi-threaded C program in a graduate course taught at the University of Virginia.

3. *Daisy*, a prototype implementation of a Unix-like file system in Java. Daisy was developed as a common testbed for different program verification tools [Daisy04].

4. *OpenSSL*, a C implementation of the Secure Sockets Layer (SSL) specification [SSL, OpenS]. Our experiment focuses on its handshake protocol.

5. *JBoss*, a Java application server conforming to the J2EE specification [JBoss, J2EE]. Our experiment focuses on its transaction management module.

6. *Microsoft Windows kernel APIs*, a set of about 1800 functions written in C or C++. These functions are the foundation of the Windows operating system. Our experiment is on Windows Vista.

Producer-consumer, Bus Simulator, and Daisy are small prototype programs. They are good for proof-of-concept experiments. The other three programs are complex and widely used. They are valuable for understanding the strength and weakness of our technique when it is applied to real systems.

Our testbeds differ in the quality of specification available. As a well studied problem, Producer-Consumer has several well-known temporal properties that have been described in both English and formal logic. The Bus Simulator comes with instructor's English description of a list of properties a valid implementation ought to have. Daisy does not have any temporal specification except a list of properties in English provided by one of its developers. OpenSSL has an extensive specification, the SSL specification [SSL, OpenS]. The SSL specification includes a detailed description as well as a finite state automaton of the handshake protocol. The JBoss application server implements the J2EE specification [JBoss, J2EE]. In particular, its transaction management module implements

the Java Transaction API (JTA) specification [JTA], which has an extensive English description and an object interaction diagram illustrating the temporal behavior of the components participating in a transaction. For the Windows kernel APIs, some of their temporal rules are documented either publicly in the MSDN library or privately in some internal documents. These documented rules are, however, by no means complete as our experiment discovered many undocumented important rules.

We include two systems with extensive specifications (OpenSSL and JBoss) so that we can compare the inferred properties against the existing specifications. These existing specifications serve as a guideline of what properties are important and interesting, without which it would be much more difficult to tell whether the inference approach produces useful results.

Next, Sections 3.1 to 3.4 present the inference results for all the testbeds except OpenSSL and Bus Simulator. We evaluate the usefulness of the inferred properties in program understanding. In addition, we also evaluate the scalability, accuracy, and efforts of using our techniques. Section 3.5 summarizes the experiments with discussion of lessons learned. We defer the presentation of the experiments on OpenSSL and Bus Simulator to Section 4.2 because the focus of these experiments is on evaluating the usefulness of the inferred properties in program differencing.

3.1 Producer-Consumer

We adapted our Producer-Consumer implementation (see Figure 2.2 on page 8) from the Bandera distribution [Corb00a, Corb00b]. Our implementation satisfies two desirable properties. First, inserting an element in a Buffer object alternates with removing an element from the Buffer object. Notice that two methods of the Buffer class can insert elements: Buffer.add and Buffer.stop. Therefore, we express this property as Enter:Buffer.add | Enter:Buffer.stop → Enter:Buffer.take.

Another property is that once the Producer calls the Buffer.stop method, the Consumer must eventually stop. We express this property as Enter:Buffer.stop → Exit:Consumer.run.

We used JRat to monitor 10 distinct events corresponding to the entrance and exit of the five methods: Buffer.add, Buffer.stop, Buffer.take, Consumer.run, and Producer.main. Running the instrumented Producer-Consumer program with 100 randomly generated integers between 1 and 10000 produced 100 traces with about two million events. In the approximate mode, Perracotta analyzed the traces in 40 seconds on a machine running Windows XP Professional with one 3GHz CPU and 1GB RAM.

Table 3.2 shows the 23 inferred properties whose satisfaction ratio is greater than zero (i.e., $p_{AL} > 0$). If both $P \rightarrow S$ and $S \rightarrow P$ have $p_{AL} > 0$, we include only the one that has a greater p_{AL}. We sort the properties by their p_{AL} first and then by their names. P represents the Producer thread and C represents the Consumer thread. For example, Enter:Buffer.stop():P indicates the invocation of the Buffer.stop() method in the Producer thread, and Exit:Consumer.run():C means the completion of the Consumer.run() method in the Consumer thread.

Property 3 captures the expected property that once the Producer calls the Buffer.stop method (i.e., Enter:Buffer.stop():P), the Consumer must eventually stop (i.e., Exit:Consumer.run():C). These two events satisfy the Alternating pattern in all the traces ($p_{AL} = 1$). Property 20 is closely related to property 3 except the first event is Exit:Buffer.stop. Notice that these two events only approximately satisfy the Alternating template ($p_{AL} = 0.97$), which means, in some traces, these two events do not alternate. Inspecting the code shows that a thread switch might happen right after Buffer.stop calls notifyALL. Hence, the Consumer thread may stop before the Buffer.stop method returns.

Properties 16 to 19 are closely related. They only approximately satisfy the Alternating template (i.e., $p_{AL} < 1$). They indicate that Buffer.add and Buffer.take do not strictly alternate because the last Buffer.take event corresponds to a Buffer.stop event, although they alternate in most of the subtraces (i.e., $p_{AL} > 0.97$).

Perracotta supports the three-parameter Alternating template $P \rightarrow S \mid T$ that means P alter-

Table 3.2: Inferred Producer-Consumer properties with $p_{AL} > 0$.

No	P_{AL}	$Event_1$ → $Event_2$	$Event_1$ Frequency	$Event_2$ Frequency
1	1	Enter_Buffer.add():P → Exit_Buffer.add():P	486781	486781
2	1	Enter_Buffer.stop():P → Exit_Buffer.stop():P	100	100
3	1	Enter_Buffer.stop():P → Exit_Consumer.run():C	100	100
4	1	Enter_Buffer.stop():P → Exit_Producer.main():P	100	100
5	1	Enter_Buffer.take():C → Exit_Buffer.take():C	486881	486881
6	1	Enter_Consumer.run():C → Enter_Buffer.stop():P	100	100
7	1	Enter_Consumer.run():C → Exit_Buffer.stop():P	100	100
8	1	Enter_Consumer.run():C → Exit_Consumer.run():C	100	100
9	1	Enter_Consumer.run():C → Exit_Producer.main():P	100	100
10	1	Exit_Buffer.stop():P → Exit_Producer.main():P	100	100
11	1	Enter_Producer.main():P → Enter_Buffer.stop():P	100	100
12	1	Enter_Producer.main():P → Enter_Consumer.run():C	100	100
13	1	Enter_Producer.main():P → Exit_Buffer.stop():P	100	100
14	1	Enter_Producer.main():P → Exit_Consumer.run():C	100	100
15	1	Enter_Producer.main():P → Exit_Producer.main():P	100	100
16	0.98	Exit_Buffer.add():P → Exit_Buffer.take():C	486781	486881
17	0.98	Enter_Buffer.add():P → Exit_Buffer.take():C	486781	486881
18	0.98	Enter_Buffer.take():C → Exit_Buffer.add():P	486881	486781
19	0.97	Enter_Buffer.add():P → Enter_Buffer.take():C	486781	486881
20	0.97	Exit_Buffer.stop():P → Exit_Consumer.run():C	100	100
21	0.97	Exit_Producer.main():P → Exit_Consumer.run():C	100	100
22	0.48	Enter_Buffer.add():P → Enter_Consumer.run():C	486781	100
23	0.04	Exit_Buffer.add():P → Enter_Consumer.run():C	486781	100

nates with either S or T (the approximate inference algorithm is not implemented for three-event templates). Perracotta infers the Enter:Buffer.add | Enter:Buffer.stop → Enter:Buffer.take property as discussed at the beginning of this section.

The other properties are uninteresting. For example, entering and exiting a non-recursive method always alternate. Therefore, properties 1, 2, 5, 8, and 15 are trivial. Properties 22 and 23 have very low satisfaction ratios and therefore would have been eliminated with a higher acceptance threshold for p_{AL}. The remaining properties correspond to the call-chain of the Producer-Consumer program. The chaining method converts these properties into Alternating Chains.

In summary, Perracotta infers several important properties including the two properties that the program is expected to have. These properties are not apparent from inspecting the code because they capture the interaction between two threads. Therefore, the inferred properties can help programmers gain insight into the program.

3.2 Daisy

Daisy is a prototype Unix-like file system implemented in 2000 lines of Java code [Daisy04]. Daisy's architecture has four layers as shown in Figure 3.1. At the bottom, Daisy emulates the hard drive using a RandomAccessFile (RAF) object. Above it, the disk layer abstracts the hard drive into byte streams. The next layer abstracts the byte streams into blocks. The top layer provides an interface for files and directories.

We used JRat to monitor the invocation of all Daisy methods except those inherited from the Object class (e.g., toString). We created a wrapper for the Java RandomAccessFile class so that JRat can monitor its methods. JRat recorded a method's signature, thread, this object, and arguments.

To execute Daisy, we adapted the test harness in the Daisy distribution. Our test harness, DaisyTest, takes four parameters: F, the number of files to be created initially, T, the number

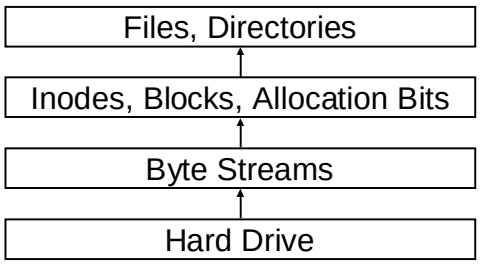


Figure 3.1: Daisy's System Architecture.

Daisy emulates the hard drive through a Java RandomAccessFile object.

of threads to be created, N, the number of iterations each thread executes, and R, the seed for the random number generator. The main thread of DaisyTest first creates F files and T threads. Next, each child thread makes a sequence of N calls to randomly selected methods of the DaisyDir class (one of read, write, set_attr, or get_attr). These methods are invoked with arguments randomly selected within the valid range.

3.2.1 Inference Results

We ran DaisyTest with *F = 5*, *T = 5*, *M = 15*, and *R = 0*. This execution produced a trace of about 70000 events. We used Perracotta to slice the original trace by threads and obtained six subtraces (five for the child threads and one for the main thread). We ran Perracotta in approximate mode with 0.70 as the acceptance threshold for p_{AL}. Our analysis only considered the 40 distinct events that occurred more than 10 times in the trace. Perracotta inferred 70 properties, 52 of which had a satisfaction ratio less than one. We applied Perracotta's chaining method to infer nine Alternating Chains.

The six shortest chains, with length from one to three events, are uninteresting because they correspond to wrapper functions. The other three chains also contain uninteresting edges due to wrapper functions. Next we applied Perracotta's call graph based heuristic to eliminate these wrapper properties and got eight properties.

Several properties provide insight into the temporal behaviors of Daisy. For example, Daisy-

Disk.readAllocBit → DaisyLock.relb ($p_{AL} = 0.97$) indicates that reading the allocation bit of a block (DaisyDisk.readAllocBit) often alternates with releasing the lock on the block (DaisyLock.relb). Because these two methods are not eliminated by the call graph based heuristic, they represent an interesting pair of asynchronous operations. These two methods do not necessarily alternate because DaisyLock.relb is called in several places (e.g., Daisy.read and Daisy.write) where DaisyDisk.readAllocBit is not called. Another interesting property is LockManager.acq → LockManager.rel ($p_{AL} = 0.86$) that captures an important locking relationship. The satisfaction ratio of this property is less than 1.0 because the traces are not sliced by objects.

Next, we used Perracotta to slice the this object and a method's first argument. Perracotta inferred two properties with $p_{AL} = 1.0$: Mutex.acq → Mutex.rel and LockManager.acq → LockManager.rel. Slicing on other arguments did not lead Perracotta to infer more properties. Object slicing missed some useful properties that involve more than one object such as LockManager.acq → Mutex.rel.

We also ran Perracotta with the two-effect-alternating and the two-cause-alternating patterns (Section 2.4.1). Perracotta inferred an important property: RAF.seek → RAF.readByte | RAF.writeByte. We found a race condition in Daisy that violates this property using the Java PathFinder model checker (Section 4.1.1). This race condition bug was also detected [Daisy04] by Willem Visser using Java PathFinder [Viss03] and Klaus Haveland using Java Explorer [Have04].

In summary, the inferred properties represent interesting temporal behaviors of Daisy. Several of the inferred properties such as DaisyDisk.readAllocBit → DaisyLock.relb involve more than one class, which indicate delocalized plans that would be useful to aid programmers in understanding the system.

3.3 JBoss Application Server

An *application server (AS)* is a piece of middleware that provides important services such as transactions, security, and caching for running web applications [J2EE]. A web application built upon an application server reuses these well-tested services and components. This makes developing web applications easier because programmers can focus on implementing the important business logic without having to worry about how to implement services such as transaction management.

A *Java application server* is a class of application server that typically runs on top of a Java virtual machine [J2EE]. The J2EE specification defines the interface between a web application and a Java application server [J2EE]. Popular Java application servers include Sun's Java System Application Server [SunAS], IBM's WebSphere [WebSphere], BEA's WebLogic Server [WebLogic], and Red Hat's JBoss Application Server [JBoss]. JBoss is the only one of these that is open-source. JBoss was originally developed by JBoss Inc. (acquired by the Red Hat Inc in April 2006) and is currently one of the most widely used Java application servers on the market [JBoss].

We are particularly interested in the APIs of the transaction management service because a transaction occurs in multiple stages with certain temporal ordering constraints. The Java Transaction API (JTA) specification defines the interfaces between a transaction manager and the other participants in a distributed transaction system: the application, the resource manager, and the application server [JTA]. The JTA specification has an object interaction diagram [Fowler03] as an illustration of how an application server may handle a transactional connection request from an application (note that the diagram is just one typical scenario, but not a specification) [JTA]. Next, we describe this typical scenario. An application server starts a transaction by first calling the begin method of the transaction manager (TM). Next the AS tries to get a transactional resource from the resource adapter (RA). The AS calls the enlistResource method to declare its ownership of a resource. Then the application does its work. To finish a transaction, the AS calls the delistResource method to

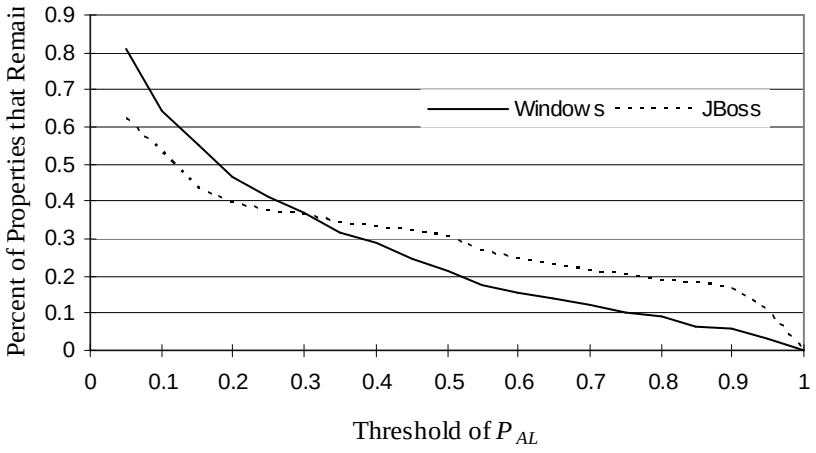


Figure 3.2: Inferred properties versus the acceptance threshold for p_{AL}.

release its ownership of the corresponding resource and then commits the transaction. The transaction commission follows a two-stage committing protocol that first prepares and then commits the transaction.

3.3.1 Inference Results

We obtained the source code of the JBoss Application Server version 4.0.2 from *www.jboss.org* (the latest one at the time of our experiments). The source code distribution included about 4000 fully automated regression test cases. We used JRat to instrument all method invocations of the transaction management module (i.e., all classes in the org.jboss.tm package) and ran the regression test suite. After dropping events that occurred less than 10 times, the execution trace contained 2.5 million events with 91 distinct events (we only monitored the entrance events). Perracotta analyzed the trace in 80 seconds on a machine with one 3GHz CPU, 1GB RAM, and Windows XP Professional.

Figure 3.2 shows the percentage of all instantiations of the Alternating pattern with p_{AL} greater than an acceptance threshold that increases from 0 to 1. The other line is for the Windows experiments described in Section 3.4. We arbitrarily picked 0.9 as the acceptance threshold to select properties. The initial result had 490 properties, too many to inspect manually. Next, the chaining

method converted the properties to 17 chains. We applied the chaining method before applying other heuristics because other heuristics might prevent a long chain from being formed. Then we pruned the results by applying the static call graph based heuristic, which reduced the number of chains down to 15 as shown in Table 3.3.

3.3.2 Comparison with JTA Specification

The longest chain (Figure 3.3) has 24 events including not only the public methods declared in the JTA specification but also private methods internal to the implementation of JBoss. After omitting the private methods, we obtain a shorter chain (Figure 3.4). The *TxManager* and *TransactionImpl* classes implement the JTA *TransactionManager* and *Transaction* interfaces respectively. This chain is almost identical to the object interaction diagram in the Java Transaction API (JTA) specification except that Perracotta does not infer the alternating relationship between enlistResource and delistResource. This is because whenever enlistResource is called, either delistResource or commitResources must be called. Therefore, a resource does not have to be delisted. As shown in Figure 3.4, Perracotta incorrectly infers enlistResource $\rightarrow$ commitResources as it is the dominant behavior in the trace.

The longest chain reveals more than just how the public APIs interact. It also provides insight into the internal implementation such as starting and committing a transaction, which would be useful for new developers to understand JBoss.

In summary, Perracotta successfully infers a complex finite state machine that is consistent with the JTA specification. This demonstrates that Perracotta can help programmers understand a real legacy system. Suppose there is no specification available for the transaction management module, it would have been a great challenge for programmers to discover its temporal behaviors by hand because these properties cross the boundary of many modules and represent a non-trivial delocalized plan. These properties would also be difficult for static analysis to infer because JBoss

Table 3.3: The JBoss AS TM Alternating Chains.

No	JBoss Transaction Management Module Alternating Chains
1	org.jboss.tm.TransactionImpl.lock org.jboss.tm.TransactionImpl.unlock
2	org.jboss.tm.TxManager.setRollbackOnly org.jboss.tm.TxManager.rollback
3	org.jboss.tm.TxManager.setTransactionTimeout org.jboss.tm.TxManager.suspend
4	org.jboss.tm.TxUtils.isActive org.jboss.tm.TxManager.commit
5	org.jboss.tm.usertx.server.UserTransactionSessionImpl.getInstance org.jboss.tm.usertx.server.UserTransactionSessionImpl.getTransactionManager
6	org.jboss.tm.GlobalId.computeHash org.jboss.tm.XidFactory.extractLocalIdFrom
7	org.jboss.tm.XidImpl.getLocalId org.jboss.tm.TransactionImpl.getGlobalId
8	org.jboss.tm.TransactionLocal.storeValue org.jboss.tm.TxManager.storeValue
9	org.jboss.tm.XidImpl.getFormatId org.jboss.tm.XidImpl.getGlobalTransactionId org.jboss.tm.XidImpl.getBranchQualifier
10	org.jboss.tm.TransactionImpl.checkWork org.jboss.tm.TransactionImpl.rollbackResources org.jboss.tm.TxManager.incRollbackCount
11	org.jboss.tm.TransactionLocal.getValue org.jboss.tm.TxManager.getValue
12	org.jboss.tm.TransactionPropagationContextUtil.getTPCFactoryClientSide org.jboss.tm.TxManager.getTransactionPropagationContext org.jboss.tm.TxManager.getTransaction
13	org.jboss.tm.TransactionLocal.initialValue org.jboss.tm.TransactionLocal.set org.jboss.tm.TransactionLocal.containsValue org.jboss.tm.TxManager.containsValue
14	org.jboss.tm.TransactionImpl.associateCurrentThread org.jboss.tm.TransactionImpl.commit org.jboss.tm.TxManager.resume org.jboss.tm.TxManager.suspend
15	See Figure 4.5

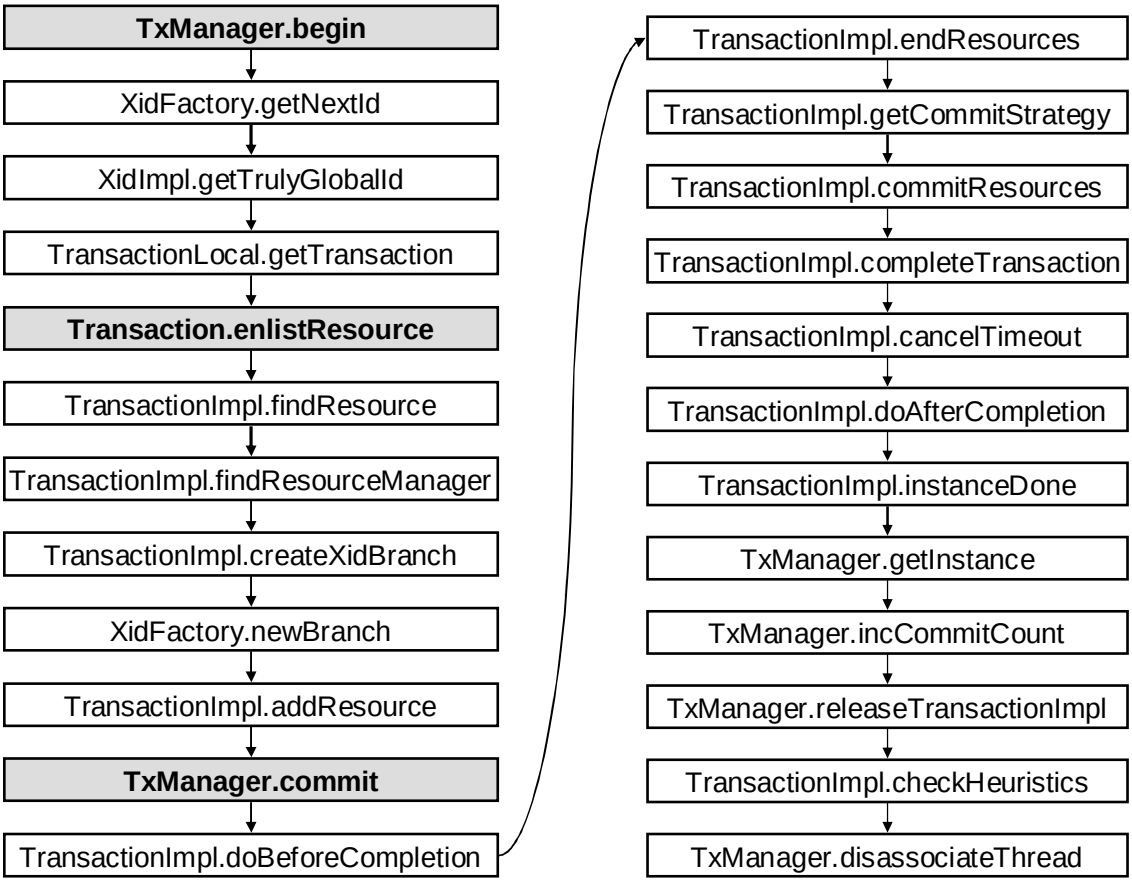


Figure 3.3: Alternating Chain for the JBoss AS TM module.

Grayed events are public methods.

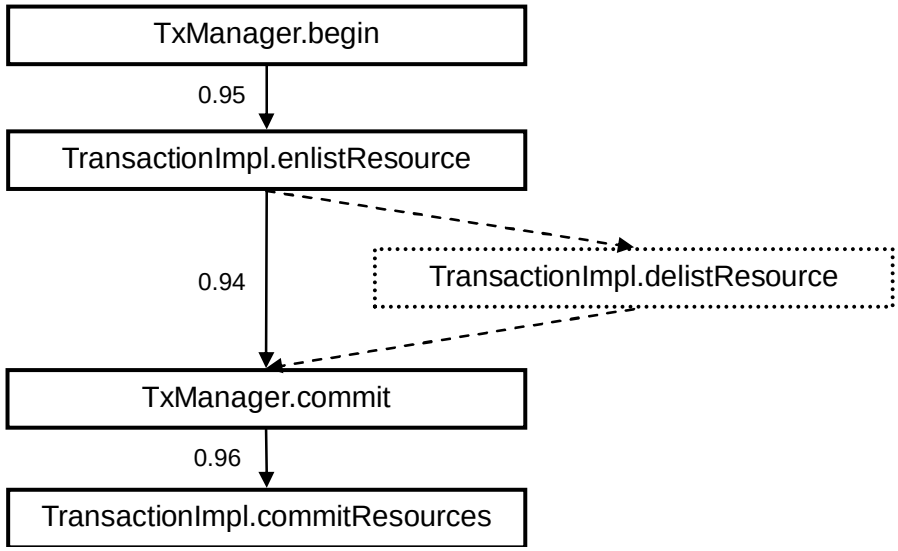


Figure 3.4: Alternating Chain for the public APIs of the JBoss AS TM module.

extensively uses polymorphic interfaces, pointers, and exception handlers. Precisely handling all these features is still beyond the state-of-the-art of static analysis. In addition, the static call graph based heuristic and the chaining method are helpful for reducing the number of properties and presenting the results. On the other hand, the naming similarity based heuristic is not very useful in the experiment since being a Java program, JBoss has few methods that implement basic resource management and locking disciplines.

3.4 Windows

In this experiment, we applied Perracotta to infer API rules for the latest (as of summer 2005) kernel (*ntoskrnl.exe*) and core components (*hal.dll* and *ntdll.dll*) of Windows Vista. Perracotta not only inferred many documented properties, but also found several important but undocumented properties. We checked 10 arbitrarily selected properties using the ESP verifier, which found a serious deadlock bug in the NTFS file system (see Section 4.1.2).

3.4.1 Inference Results

We obtained 17 execution traces from a developer in the Windows core development team. This developer instrumented APIs of the Windows kernel and core components and collected these traces by running some typical Windows applications (e.g., Windows MediaPlayer, Windows MovieMaker). He collected these traces mainly for performance tuning and debugging. We did not have any control of generating execution traces because these traces had already been generated before we started our experiments. In particular, the execution traces included the thread context information, but did not include the values of function arguments.

The lengths of the traces ranged from about 300,000 to 750,000 events, for about 5.8 million total events. The number of distinct events in each trace varied from around 30 to 1300. On average, each execution trace had about 500 distinct events. Perracotta analyzed all traces in 14 minutes on

Table 3.4: Impact of selection heuristics.

Prop	Name Similarity (>0.5)		Call Graph Only				Both	
	Prop	Reduction	Unreachable	Unknown	Total	Reduction	Prop	Reduction
7611	185	97.6%	3280	3326	6606	23.5%	142	98.13%

a machine running Windows XP Professional with one 3GHz CPU and 1GB RAM. As with JBoss, we set the acceptance threshold for p_{AL} to 0.90. Perracotta inferred 7611 properties, too many to manually inspect. We randomly selected 200 inferred properties and found that only 2 of them are interesting. So we applied the call graph and naming similarity heuristics to select the interesting properties. We used the static call graph of ntoskrnl.exe generated by ESP [Das02]. 142 properties remained after using both selection heuristics.

Table 3.4 summarizes the impact of the two heuristics. The naming similarity based heuristic alone reduces the number of properties from 7611 to 185, which is a 97.6% reduction. Although the static call graph based heuristic has a smaller reduction rate than the naming similarity, it is still very helpful for reducing the number of properties as indicated by the 23.5% reduction.

We manually inspected the 142 properties and identified 56 interesting ones, which was 40% of the 142 properties. The properties we deemed interesting are relevant to either resource allocation/deallocation or locking discipline. The heuristics increased the density of interesting properties and therefore were effective. Table 3.5 shows 20 sample properties. Appendix A lists all 56 properties. The approximation algorithm is essential for detecting useful properties such as ObpCreateHandle → ObpCloseHandle and ExCreateHandle → ExDestroyHandle that otherwise would be missing.

We compared the 56 inferred properties against those checked by the Static Driver Verifier (SDV), and found that Perracotta inferred four of the 16 sequencing properties that the SDV checked [SDV]. For example, KeAcquireQueuedSpinLock → KeReleaseQueuedSpinLock is one of the four

Table 3.5: Selected properties inferred for Windows.

Properties in bold are neither documented anywhere in MSDN nor checked by SDV.

P_{AL}	Property
1.0	ExAcquireFastMutex→ExReleaseFastMutex
1.0	**IoAcquireVpbSpinLock→IoReleaseVpbSpinLock**
1.0	**ExAcquireRundownProtectionCacheAwareEx→ ExReleaseRundownProtectionCacheAwareEx**
1.0	KefAcquireSpinLockAtDpcLevel→ KefReleaseSpinLockFromDpcLevel
1.0	KeAcquireQueuedSpinLock→KeReleaseQueuedSpinLock
1.0	KfAcquireSpinLock->KfReleaseSpinLock
1.0	KiAcquireSpinLock->KiReleaseSpinLock
1.0	MmSecureVirtualMemory→MmUnsecureVirtualMemory
1.0	**ObpAllocateObjectNameBuffer→ObpFreeObjectNameBuffer**
1.0	SeLockSubjectContext→SeUnlockSubjectContext
0.993	ObpCreateHandle→ObpCloseHandle
0.988	**GreLockDisplay→GreUnlockDisplay**
0.985	RtlActivateActivationContextUnsafeFast→ RtlDeactivateActivationContextUnsafeFast
0.982	KeAcquireInStackQueuedSpinLock→KeReleaseInStackQueuedSpinLock
0.977	SeCreateAccessState→SeDeleteAccessState
0.972	IoAllocateIrp→IoFreeIrp
0.961	**CmpLockRegistry→CmpUnlockRegistry**
0.959	**ObAssignSecurity→ObDeassignSecurity**
0.954	ExCreateHandle→ExDestroyHandle
0.954	**ExpAllocateHandleTableEntry→ExpFreeHandleTableEntry**

properties. Perracotta missed seven properties such as KeAcquireSpinLock → KeReleaseSpinLock that the SDV checked because our execution traces did not include those events.

Perracotta missed the other five properties that the SDV checked because the property templates could not express them. For example, our property templates cannot represent the property that an event only happens once. Therefore, Perracotta cannot infer the property that IoInitializeTimer is called only once.

Perracotta also inferred two important properties KiAcquireSpinLock → KiReleaseSpinLock and KfAcquireSpinLock → KfReleaseSpinLock that the SDV did not check because these functions were internal to Windows and therefore were invisible to the device driver developers.

The Windows experimental results are encouraging. The inferred properties capture critical rules which Windows developers are expected to follow when using the Windows kernel and other core components. Violating these rules could result in system crashes and would be extremely difficult to diagnose. Furthermore, many of the inferred properties are not properly documented, which reflects the sad fact that specifications are rarely available. In our personal communication with several Windows developers, they express the need for a tool to help them learn the important rules of Windows such as the ones Perracotta infers. For example, two summer interns in the Windows group were assigned a task of writing a program using a type of queue in Windows kernel. Unfortunately they could not find any documentation about this queue. As a result, they had to manually look at programs to distill the rules about how to use the queue, which was very time-consuming. Although the effectiveness of using Perracotta in development remains to be investigated, our results so far provide evidence that Perracotta can assist developers in gaining insight into critical properties of systems as complex as Windows.

3.5 Discussion

We have presented the experimental results of applying Perracotta to several systems. Our results demonstrate that Perracotta is a useful technique for inferring interesting temporal specifications.

Perracotta's inference algorithm scales very well to execution traces of real systems. For example, on a machine with one 3GHz CPU, 1GB RAM, and running Windows XP Professional, it only took Perracotta 14 minutes to analyze the Windows traces that had 5.8 million events and 500 distinct events on average.

Perracotta's approximate inference algorithm effectively handles imperfect traces. In all the experiments, Perracotta inferred interesting properties whose p_{AL} was below 1.0. These properties involve operations on key system resources such as locks and file handles.

Perracotta's heuristics for selecting interesting properties effectively increase the density of interesting properties. After applying heuristics, the manual effort required to select properties was reduced to allow efficient analysis of results even for systems as complex and large as Windows and JBoss.

Furthermore, our experimental results strongly support that Perracotta can be useful for program understanding. Perracotta discovered interesting temporal properties for all the testbeds. For the Windows kernel, Perracotta inferred 56 interesting properties, many of which were undocumented. For JBoss, Perracotta inferred a 24-event finite state machine that was consistent with the JTA specification. Many of the JBoss properties represent delocalized plans as the events cross multiple "distant" modules of the target systems. Discovering these delocalized plans is valuable because they can be hard to discover by manual inspection and violating them when modifying the system could introduce serious errors.

Chapter 4
Using Inferred Properties[1]

This chapter presents experiments using the inferred properties. Section 4.1 describes our experiments using a verification tool to check the inferred properties. Section 4.2 describes our experiments using the inferred properties to identify differences among multiple versions of a program.

4.1 Program Verification

Program verification techniques try to decide whether or not a program conforms to a specification. Although verifying any non-trivial property is undecidable [Rice53], recently researchers have achieved much practical success in verifying important generic safety properties that are independent from programs [Bush00]. Verifying application-specific properties also shows great promise [Evans96, Ball01, Das02, Viss03]. For example, Microsoft uses the Static Driver Verifier (also known as the SLAM project) to check device drivers against a list of driver related rules. The adoption of such tools, however, is limited by the unavailability of property specifications.

Chapter 3 demonstrates inferring certain application-specific rules is practical. For example, Perracotta infers the protocols for using Windows kernel by examining execution traces of existing programs that use Windows kernel. We can apply Perracotta to automatically infer important temporal properties and then feed these properties to a verification tool. Figure 4.1 illustrates this process.

We present two case studies on the Daisy file system and the Microsoft Windows kernel respectively. Chapter 3 presented the inference results. Next, we describe the results of feeding the inferred properties to a static verification tool. In particular, we used the ESP verifier to check the Windows kernel properties and the Java PathFinder model checker to verify the Daisy properties

[1] This chapter is partly based on [Yang04b, Yang06].

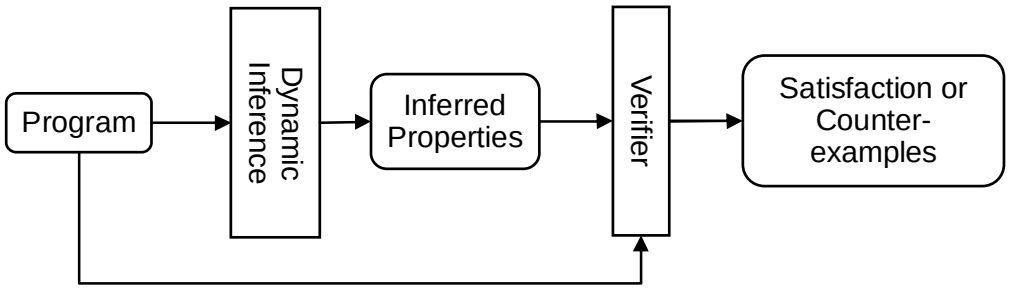


Figure 4.1: Use inferred properties in program verification.

[Das02, Viss03].

4.1.1 Daisy

Section 3.2 introduced Daisy, a prototype Unix-like file system [Daisy04]. Table 4.1 recaps the 22 inferred properties. We applied Java PathFinder (JPF) to verify these inferred properties [Viss03]. The last column in Table 4.1 shows the verification results.

Java PathFinder is an explicit-state model checker for Java programs [Viss03]. It checks deadlocks, race conditions, unhandled exceptions, and user-specified assertions. It can scale to a program of about 10000 lines. Upon finding a violation of a property, it produces an execution path illustrating the problem. It has been used to find complex concurrent bugs in real systems [Viss03].

Java PathFinder Setup

We developed a Java class, Alt, for checking the Alternating template (Figure 4.2). As presented in Figure 2.5 (page 15), the Alt class encodes the Alternating pattern as a four by two array, rule (line 2). The field, currState, keeps track of the current state (line 3) and is initialized to the initial state 0 (line 4-6). The update method updates the FSM's state (line 9). We encode a safety property as an assertion on the FSM's state, which says the current state cannot be an error state (line 10). Our instrumentor instruments Daisy so that it calls the update method whenever the P or S event occurs. Our instrumentor also inserts a call to the checkExitState method (line 13-15) to ensure that the current state is in an accepting state before the program terminates (to catch bugs such as when

Table 4.1: Results of checking Daisy properties with Java PathFinder.

No	Causing Event (*P*)	Effect Event (*S* or *S*|*T*)	JPF Found Violation?
1	Daisy.alloc()	DaisyDisk.writeAllocBit(blockno, ...)	
2	Daisy.creat()	Daisy.alloc()	
3	Daisy.creat()	Daisy.ialloc()	
4	Daisy.creat()	DaisyDisk.writeAllocBit(blockno, ...)	
5	Daisy.get_attr(inodeno, ...)	Daisy.get_attr(inode, ...)	
6	Daisy.ialloc()	Daisy.alloc()	
7	Daisy.ialloc()	DaisyDisk.writeAllocBit(blockno, ...)	
8	Daisy.iget(inodeno)	DaisyDisk.readi(inodeno, inode)	✓
9	Daisy.iget(inodeno)	DaisyLock.acqi(inodeno)	✓
10	Daisy.iget(inodeno)	DaisyLock.reli(inodeno)	✓
11	Daisy.read(inodeno, ...)	Daisy.read(inode, ...)	✓
12	Daisy.write(inodeno, ...)	Daisy.write(inode, ...)	✓
13	DaisyDir.writeLong(inodeno, ...)	Utility.longToBytes(...)	
14	DaisyDisk.readi(inodeno, inode)	DaisyLock.reli(inodeno)	
15	DaisyLock.acqb(blockno)	DaisyLock.relb(blockno)	
16	DaisyLock.acqi(inodeno)	DaisyDisk.readi(inodeno, inode)	✓
17	DaisyLock.acqi(inodeno)	DaisyLock.reli(inodeno)	✓
18	LockManager.acq(lockno)	LockManager.rel(lockno)	✓
19	Mutex.acq()	Mutex.rel()	✓
20	Petal.read(location, ...)	RAF.length()	
21	Petal.write(location, ...)	RAF.writeByte(...)	
22	RAF.seek(location)	RAF.readByte() | RAF.writeByte(...)	✓

```
1   public class Alt {
2       private final static byte[ ][ ] rule = { { 1, 2 }, { 2, 0 }, { 2, 2 } };
3       private byte currState;
4       public Alt() {
5           currState = 0;
6       }
7       public synchronized void update(int event) {
8           Verify.beginAtomic();
9           currState = rule[currState][event];
10          assert (currState != 2);
11          Verify.endAtomic();
12      }
13      public synchronized void checkExitState() {
14          assert (currState == 0);
15      }
```

Figure 4.2: The Java code for monitoring Alternating properties.

a lock is not released.

JPF did not support many Java native classes such as RandomAccessFile (RAF). We created an array to emulate RAF. We also created a simpler test harness, DaisyTestSimple, which only creates one file and two threads. Each thread either reads from or writes to the created file once. We used JPF's Verify.random in place of Java's random number generator so that JPF would automatically explore all possible results of the random number generator.

In our preliminary experiments, JPF did not finish analyzing several properties within 24 hours. JPF allows users to indicate a sequence of statements as an atomic segment by enclosing the statements between Verify.beginAtomic and Verify.endAtomic. This significantly reduces the number of states JPF has to check. To improve performance, we enclosed the initialization code of DaisyTestSimple and the monitoring code in atomic segments (line 8-11 in Figure 4.2).

Verification Results

When JPF finds a counterexample, it might be a bug. For example, consider the RAF.seek $\rightarrow$ (RAF.read | RAF.write) property (number 22 in Table 4.1), which says whenever RAF.seek is called, RAF.read or RAF.write must be called before the next invocation of RAF.seek. JPF detected

a violation of this property, where RAF.seek was called twice without a call to either RAF.read or RAF.write in between. Diagnosing this problem revealed a race condition in Daisy that was detected by several other verification tools [Daisy04]. After one thread moves the file pointer to location A, another thread starts executing and moves the file pointer to location B. If the first thread is scheduled to execute again, it would write to an incorrect position.

A counterexample can also result from a faulty property inferred from inadequate execution. For example, Perracotta inferred DaisyLock.acqi(inodeno) → DaisyLock.reli(inodeno) (number 17 in Table 4.1). These two methods operate on the lock of the inode whose inode number equals inodeno. Although this property appears to be valid, JPF found a counterexample. Inspecting the code revealed a subtle and interesting aspect of Daisy. DaisyLock.acqi(inodeno) calls LockManager.acq(lockno) that calls Mutex.acq() corresponding to the Mutex object that has the lockno. Similarly, DaisyLock.reli(inodeno) calls LockManager.rel(lockno) that calls Mutex.rel() corresponding to the Mutex object that has the lockno. Therefore, as long as the implementation of Mutex guarantees synchronized access to an inode, an upper level class (e.g., DaisyLock) does not have to ensure synchronization. JPF detected counterexamples to properties 8 to 12 and 16 to 19 in Table 4.1 for a similar reason. Although such counterexamples do not reveal bugs, they provide insight into some important yet subtle properties of Daisy.

Furthermore, the counterexample of Mutex.acq() → Mutex.rel() (number 19 in Table 4.1) revealed a limitation of our inference technique. Figure 4.3 shows the implementation of the two methods. Recall that our instrumentor monitors the entrance of a method. Hence, the Mutex.acq() event corresponds to entry of the acq() method of the Mutex class (line 4). Similarly, Mutex.rel() event corresponds to entry of the rel() method of the Mutex class (line 14). Therefore, Mutex.acq() does not have to alternate with Mutex.rel() because Mutex.acq() does not correspond to the start of the critical section (line 12).

```
1   class Mutex {
2         boolean locked;
3         ……
4         synchronized void acq() {
5            while (locked) {
6                  try {
7                        this.wait();
8                  } catch (Exception e) {
9                        System.out.println(e);
10                 }
11           }
12           locked = true;
13        }
14        synchronized void rel() {
15           locked = false;
16           this.notify();
17        }
18 }
```

Figure 4.3: The Mutex class in Daisy.

4.1.2 Windows

Section 3.4 presented the inference results of Windows. The inferred properties include interesting rules about using the Windows kernel that are neither documented by the Static Driver Verifier nor the MSDN [Ball01]. Next, we present experiments checking selected properties using the ESP verifier [Das02].

ESP is a validation tool for typestate properties [Stro86]. Typestates are more expressive than ordinary types: for an object created during program execution, its ordinary type is invariant through the lifetime of the object, but its typestate may be updated by certain operations. ESP allows a user to write a custom specification encoded in a finite state machine to describe typestate transitions. Based on the specification, ESP instruments the target program with the state-changing events. It then employs an inter-procedural data-flow analysis algorithm to compute the typestate behavior at every program point [Reps95]. ESP handles inter-procedural analysis through the use of partial transfer functions via function summaries.

From the 56 inferred properties (see Appendix A), we randomly selected ten locking properties

```
1     void PushNewInfo ( struct1 s1, struct2 s2 ) {
2         …
3         acquire ( s1.mutex);
4         …
5         if ( s2.flag )
6             GetData ( s1, FALSE );
7         …
8     }
9
10    void GetData ( struct1 s1, boolean locked ) {
11        …
12        if ( !locked ) {
13            acquire ( s1.mutex );
14        …
15    }
```

Figure 4.4: The NTFS bug in Windows Vista.

and checked them using ESP. We converted the inferred properties to the specification language of ESP. ESP found one previously unknown deadlock bug in the NTFS file system. The property is a typical locking discipline property: acquiring a specific type of kernel Mutex must be followed by releasing the same Mutex. Figure 4.4 shows a snippet of the buggy code (the function names have been changed at Microsoft's request to conceal proprietary information). Whenever GetData is called from PushNewInfo, the Mutex will be acquired twice (line 3 and 13), causing the system to deadlock. ESP clearly showed this execution path. To fix it, the second argument on line 6 should be changed to TRUE. The Windows development team confirmed that this was a real bug and subsequently fixed the problem.

For all 10 properties, ESP produced false positives. The number of false positives was large for some properties. For example, ESP found 600 counterexamples of one property. We manually examined several counterexamples that all turned out to be false positives due to known limitations of ESP. Manually examining all counterexamples would take too much time and might not be worthwhile as previous experience indicated the false positive rate could be high. Instead, we wrote a simple Perl script to recognize the syntactical pattern of known false positives, ran the script

on the counterexamples, and only manually examined those counterexamples that the script did not recognize. This process helped us eliminate those cases that were highly likely false positives. For example, the script matched all 600 counterexamples mentioned earlier with known false positive patterns. Therefore, we did not further examine any of them.

One common type of false positive we observed was caused by the imprecision of pointer analysis. For example, ESP does not precisely analyze the target of a function pointer and therefore might consider infeasible paths. About half of the 600 false positives were caused by the imprecision of function pointer analysis. Another type of false positive we observed was caused by the imprecision in modeling non-linear arithmetic operators. ESP uses a theorem prover to decide whether a branch should be taken. Because its theorem prover cannot precisely model the effect of non-linear arithmetic operators such as shift, ESP might explore infeasible paths.

Diagnosing the counterexamples of a property took at least a day of human effort. Some types of false positives are very difficult to diagnose. For example, one counterexample caused by the imprecise modeling of the shift operator took four people including two ESP developers and one Windows developer eight hours to diagnose. Initially, three of the four people believed it was a bug in Windows until the fourth person found it to be a false positive after careful inspection of the code.

Even though diagnosis of the counterexamples was very time-consuming, it increased our confidence in the correctness of the Windows code. In addition, some of the false positives motivated the ESP developers to enhance ESP to reduce the false positives and to make diagnosis of counterexamples easier.

4.1.3 Discussion

Both case studies demonstrate that the inferred properties are useful for bug detection. For example, ESP found a serious deadlock bug in Windows code and JPF found a race condition in Daisy. Hence, feeding inferred properties to a verification tool is a promising way to detect

application-specific defects. However, checking inferred properties for bug detection is not yet close to cost-effective because current program verification tools are not at the point where this can be done in a fully automated way. A programmer has to manually inspect the counterexamples to determine whether a reported violation is a real defect. Analyzing the ESP results for each checked property consumed at least a day of human effort.

Nevertheless, checking the inferred properties has several other benefits. Diagnosing the counterexamples lead users to closely examine the target system and therefore increases their confidence in its correctness. Diagnosing the counterexamples often provides insight into a complex system's behavior. For example, the counterexamples JPF found for the Daisy locking properties reveal an interesting design decision that crosscuts multiple components. Furthermore, as demonstrated in the ESP experiments, identifying false positives can also provide helpful feedbacks for the developers of a verification tool to improve the tool to reduce the false positives and to ease diagnosis of counterexamples.

4.2 Program Differencing

Inferring the differences between two versions of a program is an important problem in program evolution. The majority of previous work focuses on static approaches, ranging from simple syntactic differencing [Mill85] to more complex static approaches that consider program semantics using dependences [Horw90, Horw94, Jack94, Bink01]. Precisely determining the semantic differences between two programs is, in general, undecidable. Therefore, static techniques usually produce an approximation of the precise semantic differences, which might be very imprecise.

We hypothesize that Perracotta can infer useful properties for differentiating programs. In particular, comparing the inferred properties can increase users' confidence that desirable temporal properties are preserved by new modifications. Furthermore, inconsistencies among the inferred

properties can reveal interesting facts such as bug fixes or program enhancement.

This section presents experiments applying Perracotta to two families of programs: student implementations of a multi-threaded programming assignment in a graduate software systems course, and archived versions of OpenSSL [OpenS]. Because all programs in each case implement the same informal specification, any differences in the inferred temporal properties are likely to be interesting. At the time of these experiments, we had not yet developed the approximate inference algorithm. Hence, the acceptance threshold for p_{AL} for all the properties reported in this section is 1.0. We use Perracotta to infer the strictest properties for each version of a program. We call the inferred properties the *signature* of each version. We compare the signatures of multiple versions of a program.

4.2.1 Tour Bus Simulator

The first experiment was on submissions to an assignment in a graduate software systems course taught at the University of Virginia in fall 2003. The assignment was a multi-threaded program simulating the operation of city bus with an informal specification, paraphrased below:

Write a program that takes three inputs: n, the number of passengers, C, the maximum number of passengers the bus can hold (C must be $\leq n$), and T, the number of trips the bus takes, and simulates a tour bus transporting passengers around town. The passengers repeatedly wait to take a tour of town in the bus, which can hold a maximum of C passengers. The bus waits until it has a full load of passengers, and then drives around town. After finishing a trip, each passenger gets off the bus and wanders around before returning to the bus for another trip. The bus makes up to T trips in a day and then stops.

The assignment also specified the format of input and output. Figure 4.5 shows the output of a typical execution when $n = 2$, $C = 1$, and $T = 1$.

Because the outputs corresponded to events of interest to us, we did not need to instrument the

```
Bus waiting for trip 1
Passenger 0 gets in
Bus drives around Charlottesville
Passenger 0 gets off
Bus waiting for trip 2
Passenger 1 gets in
Bus drives around Charlottesville
Passenger 1 gets off
Bus stops for the day
```

Figure 4.5: Sample output of Bus Simulator with $n = 2$, $C = 1$, and $T = 1$.

programs. Instead, we mapped the output logs directly to event sequences. In the mapping, we considered these five events:

1. wait (*Bus waiting for trip n*)
2. drives (*Bus drives around Charlottesville*)
3. stops (*Bus stops for the day*)
4. gets in (*A passenger gets in*)
5. gets off (*A passenger gets off*)

Note that the numbers of the trip and passenger were ignored in our event mapping, which reduced the number of distinct events.

A correct solution must satisfy several temporal properties including:

1. The bus always rides with exactly C passengers.
2. No passenger will jump off or on the bus while it is running.
3. No passenger will have another trip before getting off the bus.
4. All passengers get off the bus before passengers for the next trip begin getting on.

We analyzed eight different submissions, all of which were previously evaluated as correct by a grader.

Table 4.2: Bus Simulator Properties.

P	*S*	Property in Correct Versions	Property in Faulty Version
wait	drives	Alternating	MultiEffect
wait	gets off	MultiEffect	CauseFirst
drives	gets off	MultiEffect	CauseFirst
wait	gets in	MultiEffect	MultiEffect
gets in	drives	MultiCause	MultiCause
gets in	gets off	CauseFirst	CauseFirst
drives	stops	MultiCause	N/A
gets in	stops	MultiCause	N/A
wait	stops	MultiCause	N/A
gets off	stops	MultiCause	N/A

Inference Results

We executed each solution 100 times with randomly generated parameters ($20 \leq C \leq 40$, $C + 1 \leq n \leq 2C$, and $1 \leq T \leq 10$). We used Perracotta to infer the strictest pattern the execution traces satisfy (see Section 2.4.1 and 2.4.3). Perracotta inferred the same set of temporal properties for seven out of the eight submissions. Table 4.2 summarizes the results.

Perracotta inferred the Alternating pattern, wait $\rightarrow$ drives, for seven of the programs. In the other program, the strictest pattern inferred for wait and drives was MultiEffect. Recall that the regular expression of the MultiEffect pattern is $(PSS^*)^*$. Hence, the result indicated multiple drives events corresponded to one wait event. This led us to find the bug shown in Figure 4.6. At the end of function go_for_drive, the bus thread releases the lock (line 12). This effectively allows the passenger threads to compete for the lock (line 2) and to possibly "get in" the bus before the bus starts waiting for passengers. In most cases, the bus thread can successfully obtain the lock (line 2) before it has been filled to capacity (i.e., num_riders $<$ capacity on line 3), so it can generate the wait event (line 4). However, the bus can be already full when the bus thread obtains the lock (i.e., num_riders $\geqslant$ capacity on line 3), in which case it does not produce the wait event. In such situations, wait and drives do not alternate. One way to fix this bug would be to use a conditional variable to

```
1   void go_for_drive() {
2       pthread_mutex_lock (&mutex[mutex_lock]);
3       if (num_riders < capacity) {
4           printf ("Bus waiting for trip %d\n", num_trips);
5           pthread_cond_wait (&cond[cond_shuttle_full], &mutex[mutex_lock]);
6       }
7       printf ("Bus drives around Charlottesville\n");
8       sleep (3);
9       pthread_cond_broadcast (&cond[cond_ride_over]);
10      num_riders = 0;
11      num_trips--;
12      pthread_mutex_unlock (&mutex[mutex_lock]);
13  }
```

Figure 4.6: A synchronization bug in one bus implementation.

synchronize the bus thread and the passenger threads and make sure the bus thread generates the wait event before it broadcasts that condition.

Another difference concerned the property between drives and gets off. In seven of the implementations, drives and gets off satisfied MultiEffect, but in the other implementation the strictest property satisfied is CauseFirst ($(PP^*SS^*)^*$), which meant it was possible for the bus to drive around Charlottesville more than once without allowing passengers to get off between these trips. This turned out to be another bug of missing synchronization between the bus thread and the passenger threads. As shown in Figure 4.6, the bus thread broadcasts that the ride is over to all passenger threads after driving around the city (line 9). Then it should wait for all the passengers to get off before starting the next trip. If the bus thread runs before any passengers depart, it will still be full and will begin the next trip. The third difference in the property between wait and gets off was caused by the same bug.

4.2.2 OpenSSL

Our second experiment considered six versions of OpenSSL [OpenS], a widely used open source implementation of the Secure Sockets Layer (SSL) specification. The SSL protocol provides secure communication over TCP/UDP using public key cryptography [SSL]. We focused on the *handshake*

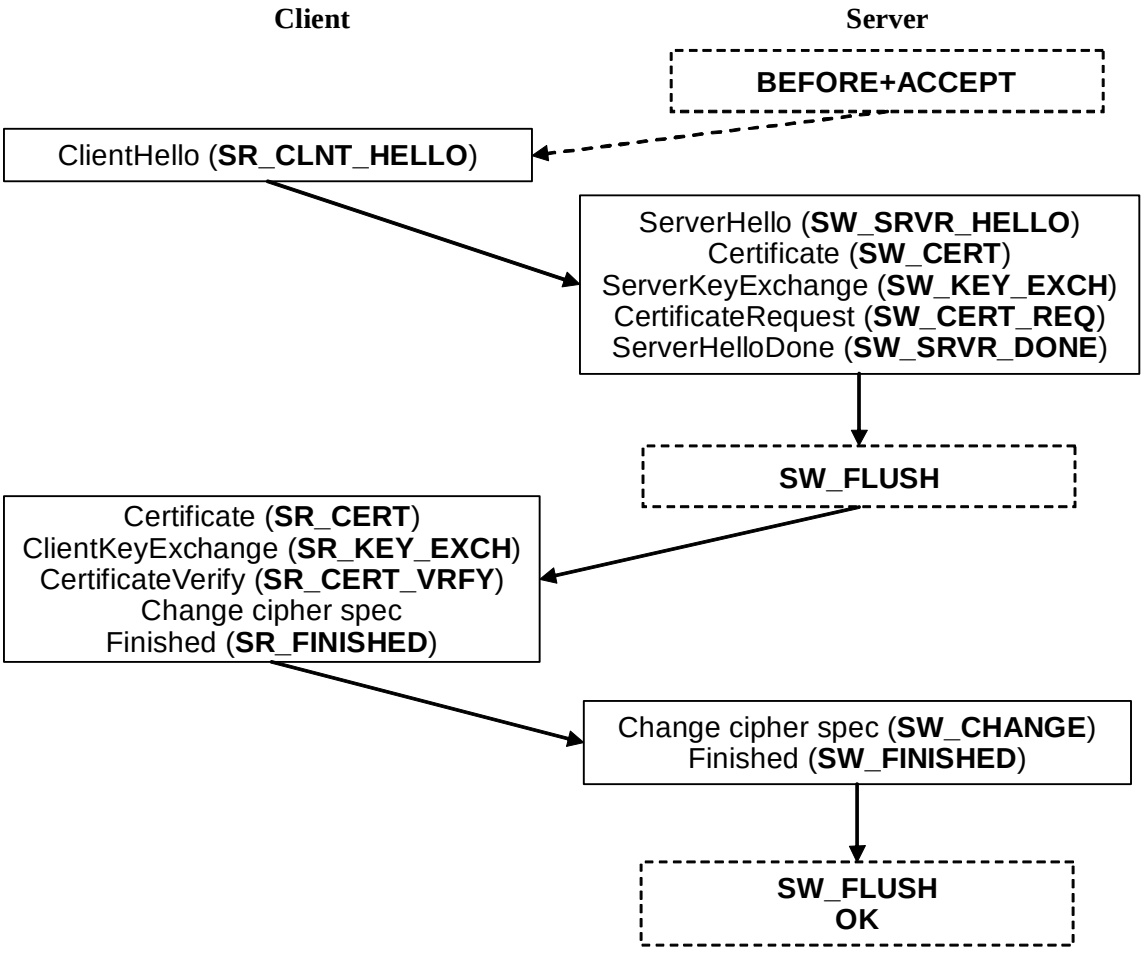


Figure 4.7: SSL handshake protocol states.

protocol that performs authentication and establishes important cryptographic parameters before data transmission starts.

SSL Handshake Protocol

Figure 4.7 illustrates the handshake protocol (derived from the SSL specification) [SSL]. The three boxes with dashed outlines contain internal events introduced by the OpenSSL implementation but not specified in the SSL specification. The remaining boxes contain sequences of events corresponding to messages defined by the SSL handshake protocol. We gave each server event a more descriptive name and showed the original server event in the parentheses.

The handshake protocol begins when the server receives a ClientHello message from a client.

BEFORE+ACCEPT → SR_CLNT_HELLO → SW_SRVR_HELLO → SW_CERT → SW_KEY_EXCH → SW_CERT_REQ → SW_SRVR_DONE → SR_CERT → SR_KEY_EXCH → SR_CERT_VRFY → SR_FINISHED_SW_CHANGE → SW_FINISHED → OK

Figure 4.8: A server event trace of normal handshake process.

Then the server sends out five messages consecutively (ServerHello, Certificate, ServerKeyExchange, CertificateRequest, and ServerHelloDone). Next, the server enters the SR_CERT state in which it tries to read certificate from the client (whether or not the client sends its certificate depends on the server's certificate request message). Then the server reads four consecutive messages from the client (Certificate, ClientKeyExchange, CertificateVerify, and Finished). If no error occurs, the server sends out its ChangeCipherSpec message and wraps up the handshake by sending its Finished message.

As shown in the dash lined boxes in Figure 4.7, the server implements several additional internal states which are also monitored. First, the server always initializes its state to BEFORE+ACCEPT at the beginning of the handshake. After sending each batch of messages, the server flushes the socket by entering the SW_FLUSH state. In addition, OK is another internal state indicating that the server cleans things up and is ready for data transmission. A typical event trace is shown in Figure 4.8.

In the OpenSSL server implementation, the handshake process is encapsulated in the function, ssl3_accept. This method implements the protocol state machine as an infinite loop that checks the current protocol state, sends or receives messages, and advances the state accordingly. We manually instrumented this function to monitor the 15 events shown in Figure 4.7.

Running OpenSSL

We adapted the OpenSSL client to generate OpenSSL server traces. We were particularly interested in analyzing the server's behavior when the client does not follow the protocol correctly, since this is often a source of errors. Therefore, we modified the OpenSSL client so that it transitioned to

some randomly selected state different from the original one with 5% probability.

Our test harness was based on a simple OpenSSL-based implementation of the HTTPS protocol: wclient and wserver (version 20020110) originally developed by Eric Rescoria [Resc01]. We added one more command-line option to wclient so that users could seed the random number generator with a specific integer to enable reproducibility of the experiments. We modified wserver so that it only accepted one connection and exited after that. We also added handler functions to print out SIGSEGV and SIGPIPE signals.

We selected six versions of OpenSSL: 0.9.6, 0.9.7, 0.9.7a, 0.9.7b, 0.9.7c, and 0.9.7d. For each version, we executed both wclient and wserver in tandem 1000 times on two machines, which produced 1000 server traces. We used the default keys and certificates supplied in the example program's package and the default choice of cryptographic algorithm.

Inference Results

We first applied Perracotta to all traces. Then we partitioned the traces into four groups: 1) correct client (i.e., the client neither changes to any unintended state nor generates segmentation faults); 2) faulty client (i.e., the client changes its state to some unintended one at least once) but no errors generated in traces; 3) segmentation fault; 4) faulty client generating errors other than segmentation faults. We applied Perracotta to each group of traces.

All traces

Table 4.3 highlights three key differences among the inferred Alternating properties.

In the first row, Perracotta inferred SW_CERT → SW_KEY_EXCH, for all versions except 0.9.6. Investigating the traces revealed that the server sometimes crashed after entering SW_CERT state with a SIGPIPE signal. We found this was caused by a documented critical bug in earlier versions of OpenSSL [OpenS].

In the second row, Perracotta inferred SR_KEY_EXCH → SR_CERT_VRFY, for all versions up

Table 4.3: Alternating properties satisfied by six versions of OpenSSL.

	0.9.6	0.9.7	0.9.7a	0.9.7b	0.9.7c	0.9.7d
SW_CERT→SW_KEY_EXCH		✓	✓	✓	✓	✓
SR_KEY_EXCH→SR_CERT_VRFY	✓	✓	✓	✓		
SW_SRVR_DONE→SR_CERT		✓				

to 0.9.7b but not for 0.9.7c and 0.9.7d. This was caused by a change to version 0.9.7c in order to make the implementation conform to the SSL 3.0 specification (documented in the change log of version 0.9.7c [OpenS]). Starting from version 0.9.7c, a server did not process any certificate message it received from a client if it did not previously request authentication of that client. The condition of whether the server requested client authentication was recorded in a variable, which could be set using a command line option. The server checked this variable to determine its next state after entering the receiving certificate state (SR_CERT). If the server did not require client authentication, it directly advanced to receiving key exchange state (SR_KEY_EXCH). Otherwise, it first read and examined the client's certificate before changing to that state. Our faulty client may send its certificate even if the server did not request one. A server earlier than 0.9.7c noticed this as an error and stopped the handshake immediately: the SR_KEY_EXCH state was not entered at all. In contrast, a server of version 0.9.7c or 0.9.7d ignored the client's certificate, continued to change its state to SR_KEY_EXCH, and then stopped the handshake because of the wrong type of message the client sends (it expected the ClientKeyExchange message, but got the Certificate message).

In the third row, Perracotta inferred SW_SRVR_DONE → SR_CERT, only for version 0.9.7. Using the log messages we identified the cause to be a race condition present in all versions. When the client changed to a false state and got some unexpected message from the server, it tried to send an alert message to the server to stop the handshake process. After that, the client disconnected the socket with the server. If the client disconnected the socket during the server sending messages, the

BEFORE+ACCEPT → SR_CLNT_HELLO → SW_SRVR_HELLO → SW_CERT → SW_KEY_EXCH → SW_CERT_REQ → SW_SRVR_DONE → SR_CERT → SR_KEY_EXCH → SR_CERT_VRFY → SR_FINISHED_SW_CHANGE → SW_FINISHED → OK

Figure 4.9: Inferred alternating chains for correct OpenSSL clients.

server would get a sending error message. The server had not and would not get the alert message from the client now because the socket had been disconnected. The server would only be able to get the alert message if it had already finished sending messages to client and entered a receiving state. In the experiment, this receiving state was SR_CERT. If the server was able to enter this state before the client sent the alert message, this event and an alert message would both be printed out at the server. Therefore, there was no guarantee that an alert message would be received after sending. Perracotta discovered this important design decision that was not documented in the specification.

Correct clients

Perracotta inferred an Alternating Chain as shown in Figure 4.9 from the server traces in which the client behaves correctly (i.e., the 5% probability of switching to a random state is never selected). All versions agree on the same Alternating Chain. This result is desirable because the pattern in Figure 4.9 conforms exactly to the SSL specification as shown in Figure 4.7 (page 72). This demonstrates that the server implementation of the handshake protocol conforms to the specification as OpenSSL evolves.

Faulty clients without errors generated

Next, we considered the set of traces corresponding to faulty clients that, surprisingly, did not generate any error event on either the server or client. For all six versions, Perracotta inferred the two Alternating Chains shown in Figure 4.10. These two chains closely follow the normal handshake. However, there are a few key distinctions between the patterns in Figure 4.9 and Figure 4.10.

Two of the Alternating properties that are present in Figure 4.9 do not appear in Figure 4.10.

SR_CLNT_HELLO → SW_SRVR_HELLO → SW_CERT → SW_KEY_EXCH → SW_CERT_REQ → SW_SRVR_DONE → SR_CERT

BEFORE+ACCEPT → SR_KEY_EXCH → SR_CERT_VRFY → SR_FINISHED → SW_CHANGE → SW_FINISHED → OK

Figure 4.10: Inferred alternating chains for non-error faulty OpenSSL clients.

Instead, those event pairs satisfy weaker patterns. SR_CERT and SR_KEY_EXCH satisfy the MultiCause pattern, and BEFORE+ACCEPT and SR_CLNT_HELLO satisfy the MultiEffect pattern. Figure 4.11(a) shows a trace that violates the expected Alternating properties. The key distinction between this trace and the traces produced using correctly behaving clients is that the eight-event sequence appears twice (shown in boldface in Figure 4.11(a)). Figure 4.11(b) shows the corresponding client events. The faulty client falsely changes its state to renegotiate (an internal state in the implementation of the client, not shown in Figure 4.7 since it is not part of the normal handshake process) instead of CW_CERT after reading the five messages from server (ServerHello, Certificate, ServerKeyExchange, CertificateRequest, and ServerHelloDone). Then, the client starts the handshake again by sending the ClientHello message, which causes the server to repeat the hello stage of the handshake again. If a client always changes its state to renegotiate after receiving the ServerHelloDone message, the server and the client will enter an infinite loop.

Suspecting this could be exploited in a denial of service (DOS) attack, we contacted the OpenSSL developers. They argued that it did not indicate a serious DOS vulnerability because the server looped infinitely only when an ill-behaved client kept sending renegotiation requests. This was similar to too many clients attempting to connect to a server, which was a scenario that could not really be prevented at the server. Although this was not a real vulnerability, it did reveal an interesting aspect of OpenSSL, which was not documented in the SSL specification and was not obvious from code inspection.

BEFORE+ACCEPT, OK+ACCEPT,
SR_CLNT_HELLO, SW_SRVR_HELLO, SW_CERT, SW_KEY_EXCH,
SW_CERT_REQ, SW_SRVR_DONE, SW_FLUSH,SR_CERT,
SR_CLNT_HELLO, SW_SRVR_HELLO, SW_CERT, SW_KEY_EXCH,
SW_CERT_REQ, SW_SRVR_DONE, SW_FLUSH,SR_CERT,
SR_KEY_EXCH, SR_CERT_VRFY, SR_FINISHED, SW_CHANGE, SW_FINISHED,
SW_FLUSH, OK

(a) Server trace

BEFORE+CONNECT, OK+CONNECT, CW_CLNT_HELLO,
CR_SRVR_HELLO,CR_CERT, CR_KEY_EXCH, CR_CERT_REQ,CR_SRVR_DONE,
RENEGOTIATE,
BEFORE, CONNECT, BEFORE+CONNECT, OK+CONNECT, CW_CLNT_HELLO,
CR_SRVR_HELLO, CR_CERT, CR_KEY_EXCH, CR_CERT_REQ, CR_SRVR_DONE,
CW_KEY_EXCH, CW_CHANGE, CW_FINISHED, CW_FLUSH, CR_FINISHED, OK

(b) Client trace

Figure 4.11: Traces generated with a faulty OpenSSL client.

Segmentation faults

There were three traces that included segmentation faults in all versions prior to 0.9.7d. These traces were from the faulty client that sent a change_cipher_spec instead of the normal client hello message at the beginning of the handshake process. We examined the change log for version 0.9.7d and found that this was due to a critical update [OpenSec], where an assignment to a null-pointer in the do_change_cipher_spec function caused the server to crash. Although this finding did not result from comparing the inferred temporal properties, it shows that using randomly behaving clients to test a server is powerful enough to uncover important problems.

Faulty clients with other errors

For all versions, Perracotta inferred the same temporal properties for traces within this category. Although we did not detect any interesting problems through such traces, this demonstrated that the server handled misbehaving clients consistently with respect to the properties Perracotta inferred.

4.2.3 Discussion

We have presented experiments using the properties that Perracotta infers to differentiate programs. The experimental results support our hypothesis that the differences among the inferred properties are useful for revealing important changes with respect to a program's temporal behaviors. For example, in the Bus Simulator experiments, the differences among the inferred properties of the eight programs led us to detect two bugs in one program. In the OpenSSL experiments, the differences among the inferred properties led us to detect a documented improvement to later versions, a documented bug in earlier versions, and an undocumented race condition in all versions.

We found that the hierarchy of property templates was useful in diagnosing differences among the inferred properties. The strictest template that two events satisfy in each program provided an important clue to localize the cause of a difference among the inferred properties. For example, in the Bus Simulator experiments, wait and gets off satisfied the MultiEffect template in seven programs, but the strictest template these two events satisfied was CauseFirst. Based on the knowledge of the two templates, we hypothesized that the cause of the difference was in the code between the wait event and gets off event, which reduced the scope of our diagnosis.

Chapter 5
Evaluation

Based on the experimental results presented in Chapters 3 and 4, we evaluate how well our approach and the Perracotta implementation address the three challenges (scalability, dealing with imperfect traces, and selecting interesting properties) described in Section 1.1. In addition, we also assess the effort required to use our approach and how useful the inferred properties are. We discuss more fundamental limitations of our approach later in Section 7.2.

5.1 Scalability

To construct a specification for a small program, many different techniques including manual inspection, static inference, and dynamic inference would serve the purpose well. However, many of these previous techniques do not scale well to large complex programs. Therefore, it is essential for a specification inference technique to be able to work on large systems.

For inferring a specification from execution traces using pre-defined templates, the time complexity of the inference depends on several factors: the number of distinct events in the trace (N), the length of the trace (L), and the complexity of the templates. The complexity of a template can be measured by how many parameters a template has (P). In general, given a trace, the more parameters a template has and the more distinct events the trace has, the more complex inferring properties is. In particular, if P is not known before-hand (i.e., arbitrary model inference), the problem is NP-hard even for inferring properties as simple as regular expressions [Gold78] (see Section 6.1).

In this dissertation, we take a two-step approach to address the scalability challenge. Our approach first infers properties using templates with only two parameters (except for the two-cause and two-effect templates that have three parameters) and then constructs more complex properties

by combining the simple properties. Next, we assess each step based on our experimental results.

For the simple templates used in our approach, our inference algorithm is very scalable. In particular, for templates with only two parameters, our algorithm's time complexity is in $\Theta(NL)$. In our experiments, it only took Perracotta 14 minutes to analyze a Windows trace with 5.8 million events and 500 distinct events on a machine with one 3GHz CPU and 1GB RAM. The largest trace that we had ever attempted to analyze had around 1500 distinct events and 100 million events. It took Perracotta 10 hours to analyze this trace on a machine with 3GHz CPU and 4GB RAM.

The second step uses the chaining method to combine Alternating properties. We have proved, in Section 2.8.2, that the chaining problem is in NP-complete. Therefore, assuming $P \neq NP$, there is no better algorithm than the brute-force algorithm for solving the chaining problem. In our experiments, the property graphs were always very sparse and therefore, we were able to apply the chaining algorithm to some fairly large property graph and obtained useful results. For example, our JBoss trace had 91 distinct events. Initially, Perracotta inferred 490 properties. The chaining method constructed a 24-event Alternating Chain from the 490 properties in less than one minute on the same machine mentioned above.

Scaling the inference of properties of arbitrary forms from large execution traces with many distinct events is very difficult and might not always be possible. However, we have shown that, by breaking down the inference into two steps and focusing the first step on very simple templates, our inference algorithm can process traces whose sizes are typical in real applications in reasonable time.

5.2 Dealing with Imperfect Traces

An *imperfect execution trace* is a trace that contains event sequences that violate a property specification that is necessary for the correctness of a system. Imperfect execution traces are com-

mon in real systems because of bugs in the target systems, imperfect tracing tools, and sampling. At the early stage of this dissertation work, our inference algorithm required 100% satisfaction. Our tool was able to infer useful properties from small programs that we knew were mostly correct. However, when we attempted to apply the algorithm to a large JBoss trace, it did not infer any useful properties. In retrospect, this result is not surprising at all as almost 90% of the 490 properties in our JBoss experiment have $p_{AL} < 1.0$. Similarly, about 50% of the 56 Windows Alternating properties have $p_{AL} < 1.0$. It is clear that a specification inference technique must be able to deal with imperfect traces in order to be useful on real systems. Other researchers' results share a same observation as ours. For example, the Strauss specification miner requires human guidance to tune an imperfect trace so that it can discover important specifications that would otherwise be missing [Ammons02].

Quantifying the effectiveness of our approximate inference algorithm is beyond the scope of this dissertation. However, our experiment results do provide strong evidence that such an approach is feasible for inferring useful properties that would otherwise be missing. For example, in all the four systems to which the approximate inference algorithm was applied, a significant portion of the selected interesting properties have $p_{AL} < 1.0$.

Under the assumption that a well-tested program should be correct most of the time, our approximate inference algorithm is able to automatically tolerate occasional imperfection in a trace. So it can be applied even when perfect tracing tools or immaculate test programs are not available. This works because most widely adopted systems have been through non-trivial testing, which ensures their main paths are mostly correct. The latent bugs that escape from testing typically reside on infrequent paths, which do not represent the program's dominant behavior.

Limitations. We still do not have a way to measure the effectiveness of the approximate inference algorithm. Therefore, we do not know what types of bugs the approximate inference algorithm

can tolerate and cannot predict how well the approximate inference algorithm can tolerate a bug. In addition, the approximate inference is limited to the Alternating template, how to generalize it to other templates still needs to be investigated. Furthermore, the approximate inference algorithm only uses P^+S^+ to partition a trace into subtraces. This definition of subtrace ignores the length and the location of the part that is imperfect. Lastly, we still do not know what the best way to choose the acceptance threshold is.

5.3 Selecting Interesting Properties

An *interesting specification* is a specification for which developers are likely to make mistakes and violation of which would produce bad consequences. For example, we consider specifications about using critical system resources such as locks and transactions to be interesting. Such properties are important because violating them can have serious consequences such as leading to system crashes [Ball01, Das02] and opening security vulnerabilities [Chen02]. Selecting interesting specifications is important because for a large program thousands of properties may be inferred, only a small fraction of which are interesting.

The key metric for measuring the effectiveness of a heuristic for selecting interesting properties is how much the heuristic increases the density of interesting properties. Let us use X to represent the set of initially inferred properties and X' to represent the set of properties after being filtered by a heuristic. Let us use I to represent the set of interesting properties in X and I' the ones in X'. The density of interesting properties before applying the heuristic is $D = |I|\backslash|X|$. Similarly, the density of interesting properties after applying the heuristic is $D' = |I'|\backslash|X'|$.

When $|X|$ is large, it is time-consuming and sometime infeasible to determine precisely the elements of I. Therefore, direct measurement of D might be infeasible. However, we can estimate D through sampling. For example, we can randomly select $\hat{X} \subset X$ and determine the set of

interesting properties $\hat{I}$ in $\hat{X}$. Then, we can estimate D using $\hat{D} = |\hat{I}|\backslash|\hat{X}|$. For example, Perracotta initially inferred 7611 properties for the Windows trace. The number of properties was too large to allow manual inspection. Therefore, we randomly selected 200 properties and found only 2 of them to be interesting properties. Hence, we estimated the density of interesting properties to be 1%.

Overall the static call graph based heuristic and the naming similarity heuristic are very effective for increasing the density of interesting properties. For example, in the Windows experiment, after applying both heuristics, D' became 39.43%, which is almost 40 times of D. The chaining method greatly simplifies the presentation of the properties, even though it does not eliminate any properties. For example, in the JBoss experiments, Perracotta initially inferred 490 properties. The chaining method merged these properties into 17 Alternating Chains. After applying the static call graph based heuristic to the 17 chains, Perracotta produced 15 Alternating Chains that contain 45 total Alternating properties. Next, we summarize the two heuristics and the chaining method.

5.3.1 Static Call Graph Based Heuristic

The static call graph based heuristic explores the call graph of a target system. Our experimental results showed that it was effective for eliminating mostly uninteresting properties in real systems written in either Java or C. This heuristic works because it captures the fact that events in interesting properties involving resource management usually do not have calling relationship.

The limitation of this heuristic is that the static call graph is usually not precise enough when there are function pointers or polymorphic interfaces. Therefore, the number of properties might still be large after applying this heuristic.

5.3.2 Naming Similarity Based Heuristic

The naming similarity based heuristic was very effective in selecting interesting properties in our Windows experiments as described in Section 3.4. It works because most of Windows APIs are related to locking or resource management. However, we did not find this heuristic to be very

useful in the other experiments, where only a small portion of the events we monitored are related to resource management.

5.3.3 Chaining Method

The chaining method simplifies the presentation of properties by grouping correlated Alternating properties into clusters. As our experimental results on JBoss demonstrated, this method groups related properties together into alternating chains and therefore can help users gain better insight into a system's behavior. For example, the 24-state FSM Perracotta inferred for the JBoss transaction management module illustrated the process of a typical transaction.

As for limitations, our chaining algorithm only works for Alternating properties. In practice, the property graphs we encountered were sparse, hence our chaining algorithm scaled well. However, if a property graph is large and dense, our chaining algorithm would not scale well because the chaining problem is in NP-Complete.

5.4 Versatility

A useful specification inference tool should be able to infer many of the useful specifications. This section assesses how well our work and the Perracotta implementation encompass the range of useful temporal specifications. We first describe the three different factors that affect the expressiveness of a template-based specification inference technique. Then, we consider each of the three factors separately with regard to our Perracotta implementation.

The set of properties a template-based specification inference technique can infer are mostly determined by the templates. For example, if all templates are in regular expressions, it is impossible to infer a property that can only be expressed using context-free languages. In addition, a template also designates how many parameters it has. Besides templates, the choice of events also affects the properties that can be inferred. For example, if the events only comprise function calls, it

is impossible to infer properties dealing with loop invariants. The last factor involves how the monitored events are processed. In particular, we consider the context information in a trace such as thread identities and argument values. If the inference algorithm slices the trace by thread identities, it would be impossible to infer properties that cross thread boundaries.

Next we evaluate how our design decisions for each of these three factors impact the properties that can be inferred.

5.4.1 Property Templates

Perracotta currently includes seven two-event templates. In addition to the seven built-in templates, Perracotta accepts any user-specified two-event regular expression template. Therefore, Perracotta is capable of inferring any property that can be expressed as a two-event regular expression. The space encompasses many useful properties that represent causal relationship between two events. For example, locking discipline and resource management properties are both important causal properties.

Among the seven two-event templates, the Alternating property was used most extensively in our experiments. We did not apply the other templates in the Windows and JBoss experiments because Perracotta produced too many properties in our initial tryout. Although we had success applying the other templates to differencing a few small programs, the results were still insufficient to claim that the other six templates and their hierarchy are useful on large real systems.

In addition to two-event properties, Perracotta is also capable of synthesizing complex properties by combining two-event properties, though the chaining method currently only works on the Alternating template. The Alternating Chain extends the concept of a causal relationship between two events to the one among multiple events. As our experiments demonstrated, Alternating Chains captured useful multi-event causal relationship in a few real systems. For example, the handshake protocol in OpenSSL was expressed as an Alternating Chain, so was the transaction processing logic

of JBoss.

Perracotta also incorporates two three-event regular expression templates. We did not attempt to develop more three-event templates. We also did not use the two three-event templates in all our experiments because the three-event templates did not scale well when the number of distinct events became large. Hence, the limited experimental results we obtained on these templates are still insufficient to draw any strong conclusion.

Overall, the main goal of this dissertation is to explore whether it is possible to scale specification inference to real systems. Therefore, this dissertation does not intend to develop a comprehensive library of templates. There are several types of useful properties that Perracotta cannot infer. First, any properties that cannot be expressed as regular expressions are clearly beyond the capability of Perracotta. Furthermore, it is unclear how to extend Perracotta to handle properties that cannot be expressed as regular expressions. Second, properties that involve three or more events remain a challenge for Perracotta's scalability as directly matching traces against templates with three or more events does not scale well when the number of distinct events becomes large. Encouraged by our experimental results with the chaining method, we believe inferring complex properties should be carried out by combining simpler properties. However, generalizing the chaining method is beyond the scope of this dissertation.

5.4.2 Choice of Events

The choice of events to be monitored determines what behaviors of a program can be analyzed. We chose to monitor the function call entrance events in most of our experiments for two reasons. First, many useful specifications are defined at the level of APIs largely due to the proliferation of procedural programming and object-oriented programming. Therefore, it is important to study how to infer API-level specifications. Second, function-level specifications are easy to understand. Programmers are trained to write programs by constructing functions. It is easy for a programmer

to reason about the relationship among functions. Therefore, a tool for inferring function-level specifications is likely to be useful for programmers.

In our experiments, we used Perracotta to infer many interesting API properties for real large systems. We also manually instrumented some inner-function events such as case statements. We only conducted experiments of inner-function events on small programs because it required manual modification. Therefore, we still don't have sufficient evidence to claim such events are feasible for large systems. Exploring different types of events is beyond the scope of this dissertation.

There are several limitations of only monitoring function call events. First, function calls might not be accurate enough to capture a relevant event. This was revealed in the Daisy experiment. Although the Mutex.acq event seemed to represent the acquiring of a mutex, it was the entrance of the critical section that indeed indicated the acquiring of a mutex. Similarly Mutex.rel did not precisely correspond to the releasing of a mutex. Therefore, Mutex.acq $\rightarrow$ Mutex.rel did not accurately represent the causal relationship between the two events and therefore was invalidated by Java PathFinder. Second, data-oriented specifications such as pre/post-conditions cannot be represented by function call events alone. To infer the correlation between a function's pre-condition and post-condition requires instrumenting the variables before and after the function calls.

5.4.3 Context-Handling Techniques

Different types of events are usually analyzed in different ways. For example, a function call event is interpreted differently from a function call's argument values. When a trace comprises different types of events, special techniques are needed to coordinate the different ways of analyzing these events. In particular, Perracotta instruments function calls along with thread identities and argument values. As discussed in Section 2.6, Perracotta has three ways of combining the analysis of the context information and the function call events. For the same trace, each of the three ways produces a different set of properties. Therefore, it is clear that different approaches of combining

the analysis of different types of events can affect the properties to be inferred.

Perracotta sliced all the traces based on thread identities. We found thread-slicing very useful in our experiments (Daisy, JBoss, and Windows). This helped infer interesting thread-local properties. A side-effect of thread-slicing is that it also reduces the number of distinct events and therefore allows efficient analysis of the traces. One limitation of thread-slicing is that it misses any interesting properties that cross thread boundaries.

Due to a limitation of our instrumentor, we were only able to instrument object information on a small program (Daisy). Although we found object-slicing useful for inferring properties for Daisy, the results were insufficient to evaluate the advantages and disadvantages of the three object-handling techniques.

5.5 Effort Required

This section assesses the usability of our approach and in particular Perracotta. Whether it is easy to use a technique is subjective. However, we can consider the efforts required to use an approach. In our context, inferring temporal specifications of a real system requires instrumenting the target system, executing the instrumented system to collect traces, and analyzing the inference results. Next, we discuss each part separately.

5.5.1 Instrumentation

Our instrumentor monitored function calls in four of our testbeds, Producer-Consumer, Daisy, JBoss, and Windows. Very little effort was required to pre-select the functions to monitor even for systems as complex as JBoss and Windows. In particular, we identified key components, the transaction management module of JBoss and the kernel APIs for Windows, and monitored all functions of these components. Selecting these components only required high-level knowledge of the target system and was fairly straightforward. We did not select the functions within the

key components because that would require more detailed knowledge about the components. We believe that our choice of events is realistic for what a new user of a system would be able to make. Choosing all the functions of a key component to monitor sometimes could still result in a large number of events and hence a large number of properties. Our experimental results showed that our selection heuristics and chaining method were very effective for pruning the inferred properties. As a result, choosing key components is a practical way to pre-select events to monitor.

For OpenSSL and Bus Simulator, the events were not function calls. In particular, we manually instrumented a switch-case statement in OpenSSL. This choice was fairly straightforward because this switch-case statement corresponded to the finite state machine of the SSL handshake protocol. We derived a list of pre-defined events from the instructions of the Bus Simulator assignment. Our experimental results demonstrated that such non-function-call events also have many interesting temporal properties. Although selecting non-function-call events would require more knowledge of the target system, doing so seems to be feasible for large systems.

Above all, the cost of selecting events to instrument is generally tolerable. The efforts involved in selecting events are often paid off as the knowledge gained through this process can be reused when analyzing the inference results.

5.5.2 Collecting Traces

This section qualitatively evaluates the effort needed to run our testbeds. We classify our testbeds into three categories based on what needs to be done to run them.

The first type of system has a complete testing infrastructure including test harness, test suite, and test execution script. Such systems are usually large-scale infrastructural systems that are widely used and therefore, heavily tested to ensure their quality. For example, the JBoss application server has a fully automated extensive test suite. Running such systems only requires negligible effort.

The second category of system has a test harness for running the system with different parame-

ters but does not have a test suite and test execution script. For example, the Bus Simulator program clearly defines a user interface that takes three integers. For such systems, we treat the target program as a black box and develop a test script that generates parameters for running the program. Users only need to understand the program's input interface. After that, writing the scripts is mostly mechanical. The amount of effort involved depends on the complexity of the program's input interface. In our experience, the effort involved was relatively small.

The last class of system only exposes a set of APIs. Such systems usually are prototypes and do not have a large user base. Running such systems therefore requires creating or adapting a program that uses these functions. For example, Daisy only exposes several functions for operations on files and directories. We had to develop a test harness to execute these functions. Although creating a test harness can be time-consuming, the test harness can be reused in a future analysis.

In summary, the effort required to collect adequate traces depends on how mature and widely adopted the program is. Running widely used complex systems such as JBoss and Windows usually requires little more than running existing testing infrastructure.

We did not investigate the relationship between the executions and the inferred properties. This is an important issue but is beyond the scope of this dissertation. In addition, for the third class of system that does not have any testing infrastructure, building a test harness that covers most of its execution scenarios can be challenging and requires detailed knowledge of the target system.

5.5.3 Analysis of Results

Since the inferred specifications are derived from execution traces, they can be invalid. It is often necessary to examine the initially inferred specifications and filter out obviously invalid specifications before applying them to other tasks such as verification and differencing program behaviors. This process, however, does not intend to prove the validity of a specification as this is generally impossible for non-trivial programs. This section considers the efforts involved in manually validating

and enhancing the inferred specifications.

In our experience, we validated the inferred properties through either looking up the documentation or reading the source code. When looking up the documentation, we typically searched for the event names. For example, we looked up the MSDN documentation and Microsoft internal documentation when analyzing the inferred properties. When we found the relevant document, we compared the description against the inferred properties. Any disparency would suggest the inferred specification might be invalid or the program be buggy. The comparison was straightforward for the Alternating properties since they were simple. An undocumented property does not, however, mean it is invalid. When we could not find the inferred specification anywhere in the documentations, we read the source code. This typically involved reading the relevant functions and the code that uses these functions. We tried to identify any obvious execution path that violated a specification.

When we identified an invalid property, we first tried to adjust it so that it became consistent with the documentation or the source code. For example, comparing the 24-event Alternating Chain with the JTA specification revealed that enlistResource $\rightarrow$ delistResource was missing in the inferred chain. Reading the relevant code that uses these two functions helped us discover that an enlisted resource can either be delisted or committed (without delist). Hence, we modified the original specification to enlistResource $\rightarrow$ delistResource|commitResource. An invalid property was eliminated when it was impossible to adjust it.

This manual validation process is usually very time consuming if a user is not familiar with the target system. Therefore, this process is unlikely to scale when there are too many inferred specifications. However, the efforts spent on investigating an inferred specification are usually paid off because the developer often gains more insight into how the target system works.

Chapter 6
Related Work

This chapter surveys work related to Perracotta. Section 6.1 describes the grammar inference problem and its complexity results, which provides the theoretical context of the specification inference problem. Section 6.2 classifies other inference work based on its underlying techniques and describes the representative work in each category. Finally, Section 6.3 presents previous work on using the inferred specifications.

6.1 Grammar Inference

A grammar G describes a language L if and only if $L(G)$ (the language generated by G) equals to L. The grammar inference problem can be informally stated as: "given some sample sentences in a language (positive samples), and perhaps some sentences specifically not in the language (negative samples), infer a grammar that describes the language." [Cook98] Inferring temporal specifications from a program's execution traces is a concrete example of the grammar inference problem, where the specifications inferred comprise the grammar and the execution traces are the sample sentences. Gold developed the first theoretical framework for analyzing the grammar inference problem [Gold67]. Gold proved that it is NP-hard to infer a deterministic finite-state automaton (DFA) with a minimum number of states from positive and negative sample sentences [Gold78]. In addition, Gold showed that it is impossible to infer such a DFA given only positive samples because an algorithm has no way to determine whether it is overgeneralizing. Cook et al. surveys the grammar inference problem, its theoretical complexity, and several practical inference techniques [Cook98].

Gold's results indicate that it is very difficult to infer the exact grammar (e.g., a DFA with minimal number of states) even for relatively simple languages such as regular languages. This

explains why earlier specification inference approaches to extract a complete finite-state machine do not scale well.

6.2 Property Inference

We classify other property inference work into two main categories: *template-based inference techniques* that have a set of pre-defined templates and try to match either execution traces or static program paths against these templates [Ernst00, Hangal02, Engler01, Weimer05, Flan01, Hack06, Prat06] and *arbitrary model inference techniques* that do not have a set of pre-defined templates and can discover arbitrary models that execution traces or static program paths satisfy [Reiss00, Whaley02, Ammons02, Cook04, Livs05, Fost02]. The major difference between a template-based inference technique and an arbitrary model inference technique is that a template-based technique tries to detect whether and how a program fits in a specific patterns, whereas an arbitrary model inference technique tries to create the best model that fit the target program or its execution traces.

We can further classify techniques in each of the two main categories into two subcategories: machine learning based techniques that use statistical machine learning to infer patterns [Ernst00, Hangal02, Engler01, Weimer05, Reiss00, Whaley02, Ammons02, Cook04, Livs05] and dataflow analysis based techniques that infer properties by either symbolically executing a program [Hack06, Prat06, Fost02] or use a verification tools as an oracle [Flan01]. Following we compare our work with the work in each category.

6.2.1 Template-based inference

Template-based inference techniques have a set of pre-defined templates that can capture pre-condition, post-condition, and invariants [Ernst01, Flan01, Hangal02], temporal constraints [Engler01, Weimer05], or very specific program features such as locks [Prat06] or buffers [Hack06].

Daikon is a tool that automatically infers likely program invariants from a program's execution

traces [Ernst00, Ernst01, Perk04]. The technique uses a set of pre-defined invariant patterns that are matched against the values observed in the traces; invariants that are violated are dropped. Invariants that survive are ranked in order of confidence. Only invariants that are above a specified confidence threshold are reported. The original Daikon inference algorithm is limited by its scalability because the algorithm is cubic in the number of variables monitored [Ernst01]. Several new optimizations recently developed by Perkins and Ernst greatly improve Daikon's performance, but performance remains an impediment to applying Daikon to large-scale programs [Perk04]. This dissertation is distinct from Daikon in that our approach infers specifications of the temporal behaviors of a system, which Daikon does not infer. Our technique also scales much better to large-scale real systems such as Windows. In addition, Daikon requires 100% satisfaction of templates and therefore is less robust to imperfect traces than our approximate inference algorithm.

Diduce is a run-time anomaly detection tool that is based on invariant inference [Hangal02]. Diduce infers data invariants as a program executes and generates alarms whenever the execution violates the current invariants. The invariants Diduce infers are bit field patterns (e.g., the last bit of a byte at certain address is always 0) and therefore are useful for detecting memory usage errors at runtime. Our work is different from Diduce in that our work focuses on API level invariants. In addition, the properties Perracotta infers can be used in several other purposes than run-time monitoring.

Engler et al. proposed a method for extracting properties by statically examining source code based on a set of pre-defined templates [Engler01]. Like our approach, their technique scales well by targeting simple property templates. Our chaining method can build more complex properties that their technique does not infer. To deal with the imprecision of static analysis, they use statistics to prioritize their results. Similarly, our approximate inference algorithm is also statistical. Furthermore, they use a set of specific names to reduce the number of candidate events [Engler01]. In order

to infer a property, Engler's technique requires an event (e.g., a function foo) occurs frequent enough in the code base (e.g., foo is called at many different places). In contrast, our dynamic technique can still infer a property as long as an event occurs frequent enough in the trace. The limitation of Engler's work is that it often produces too many false positives. To lower the rate of false positives, Weimer et al. improved Engler's work by developing an approach that examines a program's exception handling routines [Weimer05]. The intuition is that programs often make mistakes on exceptional paths, even when they are mostly correct on the normal paths. On one hand, Weimer's approach mainly focuses on local properties, while our technique can identify relationships among events that are far removed from each other in program text. On the other hand, Weimer's work and our work complement each other because it is usually very hard to generate test inputs for exception handling routines.

Houdini is an annotation inference tool for ESC [Flan01, Flan02]. The annotations are in the standard Hoare-style logic that includes pre-condition, post-condition, and invariants [Hoar69]. Houdini first generates a set of candidate annotations and then feeds these candidates to ESC. ESC either proves or refutes a property. The main problem of Houdini is that it does not scale well because the underlying ESC is slow.

Other static inference techniques focus on certain program features and are very effective. LockSmith is a tool for automatically inferring locking discipline annotations that are later used to check for deadlocks and race conditions [Prat06]. SALInfer is a tool for inferring annotations for function arguments that are related to buffer usages [Hack06]. SALInfer has been successfully used to help annotate the whole Windows code-base. Both LockSmith and SALInfer use dataflow analysis. Compared to LockSmith, Perracotta infers similar type of property using a regular expression inference from execution traces. Compared to SALInfer, our work focuses on a different class of properties that SALInfer does not infer. Furthermore, our underlying technique is based on regular

expression matching instead of dataflow analysis.

6.2.2 Arbitrary model inference

Arbitrary model inference techniques try to discover a model that fits the target program or its execution traces. Most of the related work in this category analyzes program execution traces [Reiss00, Whaley02, Ammons02, Cook04]. Other artifacts analyzed include revision history [Livs05] and source code [Fost02].

Ammons et al. used an off-the-shelf probabilistic finite state automaton learner to mine temporal and data-dependence specifications for APIs or abstract data structures from dynamic traces [Ammons02, Ammons03]. In addition to handle traces containing bugs (i.e., imperfect traces as defined in our work), their approach required non-trivial human guidance. In contrast, our techniques can automatically tolerate imperfect traces without guidance. Their machine learning algorithm has a high computational cost, whereas our algorithm scales better to larger traces than theirs.

Whaley et al. proposed a static and a dynamic approach for inferring what protocols clients of a Java class must follow [Whaley02]. The protocols their approach found are mainly typestate properties that involve one component and are small. In contrast, our approach is able to discover useful properties involving more than one component. In addition, our chaining method is able to construct large finite state machines efficiently.

Cook et al. developed a statistical dynamic analysis for extracting thread synchronization models from a program's execution traces [Cook04]. Our work differs from theirs in that our approach focuses on detecting API rules and assumes the trace already has the thread information.

Reiss et al. developed a technique to compact large volumes of execution traces [Reiss00]. Their tool uses the sequencing properties on individual objects, while Perracotta detects rules across multiple objects.

DynaMine extracts error patterns from a system's CVS revision histories and dynamically val-

idates inferred patterns [Livs05]. This approach is complementary to our work in that examining a CVS history is a way to select events to monitor at run-time. Their mining algorithm has to filter out a fixed set of frequent events to scale to large scenarios, which is not as general as our heuristics. The patterns their approach infers tend to focus only on methods within a class, whereas our approach can infer properties involving more than one class.

6.3 Use of Inferred Specifications

This section presents related work on using the inferred specifications to improve a variety of software development activities including defect detection [Engler01, Flan01, Hangal02, Whaley02, Nimm02, Fost02, Livs05, Weimer05, Hack06, Prat06, Chilim06, Csall05, Csall06], test case generation [Hard03, Xie06, Csall06], program evolution [Ernst01], program understanding [Mande05, Cook04], theorem proving [Win04], bug localization [Libl05], and data structure repairing at run-time [Demsky06].

6.3.1 Defect Detection

Defect detection is the application to which inferred specifications are most widely applied. We can further divide the related work into several groups: using a separate verification tool [Flan01, Whaley02, Nimm02, Livs05, Hack06], statistical defect detection [Engler01, Hangal02, Weimer05, Chilim06], context-free language reachability analysis [Fost02, Prat06], combining static analysis and test generation [Csall05, Csall06].

The work in the first group simply feeds the inferred properties to a separate program verification tool [Flan01, Whaley02, Nimm02, Livs05, Hack06]. This is similar as our work on checking the inferred properties using Java PathFinder and ESP. The benefits of such an approach are that inferred specifications can be easily checked by different verification tools.

The work using statistical defect detection integrates specification inference tightly with defect

detection [Engler01, Hangal02, Weimer05, Chilim06]. Engler et al. pioneered work in statistical defect detection [Engler01]. Their technique examined a program's source code to observe common behavioral patterns (called *belief*) and detected violations of the common behavioral patterns based on statistics. Weimer et al. tried to reduce the false positive rate by focusing the analysis on exception handling routines [Weimer05]. The Diduce tool used a similar idea as Engler's work except that Diduce detected abnormal program behavioral at runtime [Hangal02]. Chilimbi et al. was inspired by Diduce and aimed to detect abnormal memory usage patterns [Chilim06]. Statistical defect detection works well in practice because real world programs, even though not perfectly correct, behave correctly most of the time and therefore only expose abnormal behaviors occasionally. Its drawback is that the defects that can be detected are hard-coded and therefore extending existing tool to check new defects requires significant efforts. Our approximate inference algorithm bears a similar spirit as the statistical defect detection work. Instead of detecting bugs, our approximate inference algorithm tries to use a statistical approach to tolerate imperfect program behaviors.

Foster's PhD dissertation on type qualifiers introduced the idea of using constraint-based context-free language reachability analysis to infer type annotations and detect bugs [Fost02]. More recent work applied this idea to infer correlation between data and locks [Prat06]. The inferred correlations were checked for race conditions. This work also tightly combined the inference of specifications (type qualifiers or data-lock correlations) with the detection of their violation.

Recent work by Csallner and Smaragdakis opened an interesting direction of combining dynamic inference, static analysis, and automated test generation for bug detection [Csall05, Csall06]. Their work used dynamic inference to infer intended program behaviors so that users did not have to provide the specifications, static verification tool to detect violations of inferred invariants, and test generation to check the validity of the errors reported by the static analysis. One advantage of this approach compared to a purely static analysis was that the test generation eliminated the false

positives in the results. Compared to this work, our work of checking the inferred specifications still produced too many false positives to be practical.

6.3.2 Other Uses

Inferred specifications were used to augment and minimize test suite [Hard03]. The idea was to use inferred properties as a test coverage criterion. Test cases were selected until the set of inferred properties stabilized. Harder's work, however, did not automatically generate new test cases. Xie and Notkin used inferred properties to automatically generate unit-test cases [Xie06]. In our future work, we plan to investigate how to use the inferred temporal specification to select and generate test cases.

Ernst et al. studied using the invariants inferred by Daikon to aid programmers in modifying programs [Ernst01]. In their experiments, programmers were supplied with the inferred invariants, as they added a new feature to an existing program. The programmers found the inferred invariants useful. Our work on using inferred invariants in differencing programs was inspired by Ernst's work. Compared to Ernst's work, our work focused more on automatically detecting the semantic difference of multiple versions of a program, whereas Ernst's work aimed to evaluate whether the inferred specification could be useful for programmers.

Naturally, inferred specifications can be used to help programmers gain more insights into the target program. Cook et al. developed a technique for inferring thread interaction models from execution traces [Cook04]. Their technique used a statistical approach. Mandelin et al. developed a technique for discovering the sequence of APIs that were needed to accomplish some tasks [Mande05]. They called their tool *specification prospector* and the inferred API sequence *jungloid*. Our work on inferring Alternating Chains bore a similar motivation as Mandelin's work.

Another representative use of inferred specifications is in guiding theorem prover to verify distributed programs [Win04].

Liblit et al. showed the inferred specifications can be used to debug a program [Libl05]. Their work monitored a set of program predicates and compared the observed properties of a successful execution against the ones of a failed execution. Their technique used a statistical approach to eliminate irrelevant predicates and rank inferred predicates. One big advantage of this work compared to previous work on bug localization is that it can effectively handle programs that have more than one bug. We also plan to investigate whether the temporal specifications inferred by Perracotta can be useful in debugging.

Much work has been on using inferred specifications to detect bugs, Demsky et al. showed that the inferred specification can also be used to let a program continue its execution in face of data structure corruption [Demsky06]. Their technique used Daikon to infer invariants for data structures and automatically detected and restored corrupted data structures to behave normally with regard to the invariants.

Chapter 7
Conclusion

Software specifications are the foundation of many software development activities including maintenance, testing, and verification. We have presented Perracotta, a tool that automatically infers temporal specifications of programs by analyzing test execution traces. Perfect, fully-automatic specification inference for industrial programs remains an elusive goal, well beyond the state-of-the-art. We have shown, however, that by targeting simple properties that can be efficiently discovered and by using approximation inference techniques along with heuristics for pruning the set of inferred properties, it is possible to obtain useful results even for programs as large and complex as JBoss and Windows.

The experimental results of applying Perracotta to a diverse range of real systems provide strong evidence that our approach is useful in aiding several different software development tasks: program understanding, program differencing, and program verification. In particular, the experimental results demonstrate that our approach is able to automatically identify important temporal properties, identify interesting behavioral differences among multiple versions of programs, and help find bugs.

7.1 Contributions

Next, we summarize the seven contributions this dissertation makes with highlights of relevant experimental results.

1. We developed a scalable inference algorithm that can efficiently analyze large execution traces. Targeted towards simple regular expression templates, the running time of this algorithm is linear in the length of the execution trace and the number of distinct events in the trace. This scalability enables the approach to be successfully applied to several real soft-

ware systems. For example, Perracotta analyzed 17 Windows kernel execution traces with 5.8 million events and 500 distinct events on average in only 14 minutes on a typical desktop machine.

2. We developed a statistical inference algorithm that can effectively deal with imperfect execution traces. Our experimental results show that imperfect traces are common in practice. The approximate inference algorithm is essential for discovering those properties that are not 100% satisfied in the traces. For example, almost half of the 56 interesting Windows properties have a satisfaction rate less than 1.0.

3. We developed the call graph and naming similarity heuristics for selecting interesting properties. These two heuristics effectively increase the density of interesting properties and are crucial for the approach to be of any practical use. For example, in the Windows experiment, Perracotta initially inferred more than 7000 properties of which only about 1% was interesting. Applying these two heuristics reduced the number of properties to 142, over one third of which were interesting.

4. We developed a chaining method for constructing large finite state automata out of a set of smaller ones. This method is useful for presenting a large number of inferred properties in a more readable way. We formally presented the chaining problem and proved its NP-Completeness. In the JBoss experiment, the chaining method produced a 24-event Alternating Chain for the JBoss transaction management module, which was consistent with the J2EE specification.

5. We showed that our dynamic analysis technique can help programmers understand the temporal behaviors of real systems. Our technique inferred temporal properties that involved multiple classes in the JBoss transaction module. The inferred properties can help developers

understand the design of the transaction module of JBoss. Additional key results included 56 interesting rules for Windows APIs, most of which were undocumented. These inferred Windows properties can help developers understand rules they need to follow when using the Windows APIs.

6. We showed that applying verification tools to check the inferred properties can help detect application-specific defects. In particular, we detected one serious deadlock bug in the NTFS file system in Windows Vista using the ESP verifier and a race condition in the Daisy file system using the Java PathFinder model checker. In addition, even though other violations detected by JPF did not lead to bugs, they revealed subtle behavioral differences across the layers of Daisy.

7. We demonstrated that our dynamic analysis technique can aid in program evolution by discovering important temporal behavioral differences among multiple versions of a real system. For the Bus simulator student assignment, the differences among inferred temporal properties helped us find two previously unknown bugs in one submission. In the OpenSSL experiment, comparing the inferred properties across six versions revealed previously detected bugs, identified improvements, and provided evidence that the desirable properties were preserved as the system evolved.

7.2 Limitations

Chapter 5 summarizes limitations with regard to our particular implementation of our dynamic analysis approach. This section discusses several fundamental limitations of our approach.

Our approach, being a dynamic analysis, shares the limitations of any dynamic analysis. In particular, most real systems have an infinite number of execution paths. It is impossible to execute such a system on all of its paths. Dynamic analysis only examines a subset of all paths of a target

program and might produce results that are false for some paths. The properties inferred by our technique might be false and therefore require manual or machine validation before being used as specifications. In addition, dynamic analysis needs to instrument a target program to observe its behaviors. This instrumentation can affect the normal behavior of a target program. The extra computation introduced by the instrumentation might cause a thread in a real-time system to miss its deadline. The extra memory used by the instrumentation might affect cache locality. Hence, dynamic analysis is impractical to analyze a program's properties if these properties can be affected by the instrumentation.

Our approach uses a set of pre-defined property templates. This limits the properties that can be inferred to those that can be expressed using these templates. Even though we can introduce a new template to express new properties, a new template will also introduce many uninteresting properties that require new heuristics to filter. Developing a good heuristic for distinguishing interesting properties from uninteresting ones requires much effort and may not always be possible. Without a good heuristic, inferred properties can be useless if the density of interesting properties is very low. Furthermore, if a new template is complex (i.e., templates with many parameters), it might be very inefficient to infer properties that satisfy it.

To reduce the effort required to analyze the inference results, our approach relies heavily on the effectiveness of heuristics for selecting interesting properties. The effort required to analyze the remaining properties after applying the heuristics sometimes can still be quite high even for users familiar with the target system. In addition, heuristics, no matter how good they are, can mis-classify interesting properties as uninteresting ones. Therefore, interesting properties may be missing in the final results.

One of our assumptions is that a target program is well tested and exhibits desirable specifications most of the time. To tolerate imperfection in the traces, our technique uses a very simple

statistical approach. Even though our experimental results show that this assumption is valid for typical real systems, our approach can hardly be applicable to systems that do not satisfy this assumption.

7.3 Future Work

There are many exciting opportunities to extend our approach to infer new types of properties and to new applications.

Our current templates have only two or three parameters and most work uses only the Alternating template. In future work, we plan to develop techniques for inferring properties involving more events. The events our current implementation instruments are function entrance and exit events. We want to instrument other events including branches, function pre-conditions, and post-conditions. We are also interested in investigating how to infer properties that combine temporal properties with data properties. As an example, consider the property that the value of variable x should remain same between acquiring and releasing lock l. In this property, between acquiring and releasing lock l is a temporal property, whereas the value of variable x should remain same is a data invariant.

We can improve the precision of our approximate inference algorithm. The current algorithm considers a subtrace as P^+S^+. This definition of subtraces ignores the length of a subtrace. Let P^n represent P repeats n times. Our current approximate algorithm gives $P^{100}S^{100}$ and $PPSS$ same weight. Therefore, p_{AL}'s for $PSP^{100}S^{100}$ and $PSPPSS$ are both 50%. The subtrace in $PSP^{100}S^{100}$ that does not satisfy the Alternating pattern is, however, much longer than the corresponding subtrace in $PSPPSS$. We can improve the computation of p_{AL} to include the length of the subtraces, which would give the second trace a higher rank than the first one. Such an improvement more precisely characterizes the relationship between two events in a trace than our current algorithm does.

Currently we only infer properties from a fixed set of test cases. We still don't understand how the inferred properties change as we vary the test cases, when the inferred properties stabilize as we add new test cases, and what size of test suite is required to produce a stable set of properties. Such knowledge can guide users in selecting test cases for inferring properties.

For handling context information such as objects and threads, we rely on context slicing that is very effective for inferring typestate properties. However, it might miss properties that cross modules. In the future, we plan to investigate other techniques for handling context information.

We only qualitatively compare different specification inference techniques in this dissertation. We would like to conduct quantitative studies to compare their strengths and weaknesses in the future.

There are also plenty of opportunities to explore new applications of our inference approach to software development activities. We plan to study how to use inferred properties to help generate and select test cases. We are interested in exploring ways to validate the inferred properties at runtime. Furthermore, we want to study how to use the inferred properties to help localize faults in a program. Lastly, we are interested in studying whether we could use the inferred properties to improve static analysis.

7.4 Summary

This dissertation has presented a dynamic analysis approach for inferring interesting temporal properties. Through experiments on several real systems, we have shown that our approach is scalable, effective, and useful in aiding a variety of software development activities.

Bibliography

[Abrial96] J. R. Abrial. *The B-Book: Assigning Programs to Meanings*. Cambridge University Press. October 1996.

[Andr04] T. Andrews, S. Qadeer, J. Rehof, S.K. Rajamani, and Y. Xie. Zing: Exploiting Program Structure for Model Checking Concurrent Software. *International Conference on Concurrency Theory*, August/September 2004.

[Alur05] R. Alur, P. Cerny, P. Madhusudan, and W. Nam. Synthesis of Interface Specifications for Java Classes. *Symposium on Principles of Programming Languages*, January 2005.

[Ammons02] G. Ammons, R. Bodik, and J. R. Larus. Mining Specifications. *Symposium on Principles of Programming Languages*, January 2002.

[Ammons03] G. Ammons, D. Mandelin, R. Bodik, and J. R. Larus. Debugging Temporal Specifications with Concept Analysis. *Conference on Programming Language Design and Implementation*, June 2003.

[Ball01] T. Ball and S. K. Rajamani. Automatically Validating Temporal Safety Properties of Interfaces. *International SPIN Workshop on Model Checking of Software*, May 2001.

[BCEL] The Byte Code Engineering Library. http://jakarta.apache.org/bcel/

[Benn95] K. Bennett, T. Bull, E. Younger, and Z. Luo. Bylands: Reverse Engineering Safety-Critical Systems. *International Conference on Software Maintenance*, October 1995.

[Beizer90] B. Beizer. *Software Testing Techniques, Second Edition*. Van Nostrand Rheinold, New York, 1990.

[Bhan06] S. Bhansali, W. Chen, S. de Jong, A. Edwards, R. Murray, M. Drinic, D. Mihocka, and J. Chau. Framework for Instruction-Level Tracing and Analysis of Program Executions. *International Conference on Virtual Execution Environments*, June 2006.

[Bink01] D. Binkley, R. Capellini, L. Raszewski, and C. Smith. An Implementation of and Experiment with Semantic Differencing. *International Conference on Software Maintenance*, November 2001.

[Bowe93] J. Bowen, P. Breuer, and K. Lano. Formal Specifications in Software Maintenance: from Code to Z++ and Back Again. *Information and Software Technology*, November/December 1993.

[Boya02] C. Boyapati, S. Khurshid, and D. Marinov. Korat: Automated Testing Based on Java Predicates. *International Symposium on Software Testing and Analysis*, July 2002.

[Breu91] P. T. Breuer and K. Lano. Creating Specifications from Code: Reverse-Engineering Techniques. *Journal of Software Maintenance: Research and Practice*, Volume 3, 1991.

[Bush00] W. R. Bush, J. D. Pincus, and D. J. Sielaff. A Static Analyzer for Finding Dynamic Programming Errors. *Software - Practice and Experience*, Volume 20, 2000.

[Chak03] S. Chaki, E. Clarke, A. Groce, S. Jha, and H. Veith. Modular Verification of Software Components in C. *International Conference on Software Engineering*, May 2003.

[Chelf06] B. Chelf. Measuring Software Quality: A Study of Open Source Software. Coverity Inc. May 2006.

[Chen02] H. Chen and D. Wagner. MOPS: an Infrastructure for Examining Security Properties of Software. *Conference on Computer and Communications Security*, November 2002.

[Chen06] W. Chen, S. Bhansali, T. Chilimbi, X. Gao, and W. Chuang. Profile-Guided Proactive Garbage Collection for Locality Optimization. *Conference on Programming Language Design and Implementation*, June 2006.

[Chilim06] T. M. Chilimbi and V. Ganapathy. HeapMD: Identifying Heap-Based Bugs Using Anomaly Detection. In *Conference on Architectural Support for Programming Languages and Operating Systems*, October 2006.

[Clar00] E. M. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT Press, 2000.

[Cook98] J. E. Cook and A. Wolf. Discovering Models of Software Processes from Event-Based Data. *ACM Transactions on Software Engineering and Methodology*, Volume 7, 1998.

[Cook04] J. E. Cook, Z. Du, C. Liu, and A. L. Wolf. Discovering Models of Behavior for Concurrent Workflows. *Computers in Industry*, April 2004.

[Coppit05] D. Coppit, J. Yang, S. Khurshid, W. Le, and K. Sullivan. Software Assurance by Bounded Exhaustive Testing. *IEEE Transactions on Software Engineering*, Volume 31, April 2005.

[Corb00a] J. Corbett, M. Dwyer, J. Hatcliff, S. Laubach, C. Pasareanu, Robby, and H. Zheng. Bandera: Extracting Finite-State Models from Java Source Code. *International Conference on Software Engineering*, June 2000.

[Corb00b] J. Corbett, M. Dwyer, J. Hatcliff, and Robby. A Language Framework for Expressing Checkable Properties of Dynamic Software. *International SPIN Workshop on Model Checking of Software*, August 2000.

[Corb89] T. A. Corbi. Program Understanding: Challenge for the 1990s. *IBM Systems Journal*. Volume 28, 1989.

[Cormen01] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, Second Edition*. MIT Press and McGraw-Hill, 2001.

[Csall05] C. Csallner and Y. Smaragdakis. Check 'n' Crash: Combining Static Checking and Testing. In *International Conference on Software Engineering*, May 2005.

[Csall06] C. Csallner and Y. Smaragdakis. DSD-Crasher: A Hybrid Analysis Tool for Bug Finding. In *International Symposium on Software Testing and Analysis*, July 2006.

[Daisy04] Joint CAV/ISSTA Special Event on Specification, Verification, and Testing of Concurrent Software, July 2004.

[Das02] M. Das, S. Lerner, and M. Seigle. ESP: Path-Sensitive Program Verification in Polynomial Time. *Conference on Programming Language Design and Implementation*, June 2002.

[Davi95] A. M. Davis. *201 Principles of Software Development*. McGraw-Hill, 1995.

[Demsky06] B. Demsky, M. D. Ernst, P. J. Guo, S. McCamant, J. H. Perkins, and M. Rinard. Inference and Enforcement of Data Structure Consistency Specifications. In *International Symposium on Software Testing and Analysis*, July 2006.

[Detl96] D. L. Detlefs. An Overview of the Extended Static Checking System. *Workshop on Formal Methods in Software Practice*, January 1996.

[Dick93] J. Dick and A. Faivre. Automating the Generation and Sequencing of Test Cases from Model-Based Specifications. *International Symposium of Formal Methods Europe on Industrial-Strength Formal Methods*, April 1993.

[Dijk72] E. W. Dijkstra. Notes on Structured Programming. *Structured Programming*, Academic Press, 1972.

[Dor04] N. Dor, S. Adams, M. Das, and Z. Yang. Software Validation via Scalable Path-Sensitive Value Flow Analysis. *International Symposium on Software Testing and Analysis*, July 2004.

[Duca05] S. Ducasse and M. Lanza. The Class Blueprint: Visually Supporting the Understanding of Classes. *IEEE Transactions on Software Engineering*, Volume 31, January 2005.

[Duran84] J. W. Duran and S. C. Ntafos. An Evaluation of Random Testing. *IEEE Transactions on Software Engineering*, July 1984.

[Duvall05] M. Duvall. Software Bugs Threaten Toyota Hybrids. *The Baseline Magazine*. August 4, 2005. http://www.baselinemag.com/article2/0,1540,1843934,00.asp

[Dwyer99] M. Dwyer, G. Avrunin, and J. Corbett. Patterns in Property Specifications for Finite-State Verification. *International Conference on Software Engineering*, May 1999.

[Engler01] D. Engler, D. Y. Chen, S. Hallem, A. Chou, and B. Chelf. Bugs as Deviant Behavior: a General Approach to Inferring Errors in Systems Code. *Symposium on Operating Systems Principles*, October 2001.

[Ernst00] M. D. Ernst. *Dynamically Discovering Likely Program Invariants*. Ph.D. dissertation, University of Washington Department of Computer Science and Engineering, August 2000.

[Ernst01] M. Ernst, J. Cockrell, W. Griswold, and D. Notkin. Dynamically Discovering Likely Program Invariants to Support Program Evolution. *IEEE Transactions on Software Engineering*, February 2001.

[Evans96] D. Evans. Static Detection of Dynamic Memory Errors. *Conference on Programming Language Design and Implementation*, May 1996.

[Field03] J. Field, D. Goyal, G. Ramalingam, and E. Yahav. Typestate Verification: Abstraction Techniques and Complexity Results. *International Symposium on Static Analysis*, June 2003.

[Fink06] S. J. Fink, E. Yahav, N. Dor, G. Ramalingam, and E. Geay. Effective Typestate Verification in the Presence of Aliasing. *International Symposium on Software Testing and Analysis*, July 2006.

[Flan01] C. Flanagan, R. Joshi, K. Rustan, and M. Leino. Annotation Inference for Modular Checkers. *Information Processing Letters*, February 2001.

[Flan02] C. Flanagan, K. R. M. Leino, M. Lillibridge, G. Nelson, J. B. Saxe, and R. Stata. Extended Static Checking for Java. *Conference on Programming Language Design and Implementation*, June 2002.

[Fost02] J. S. Foster. *Type Qualifiers: Lightweight Specifications to Improve Software Quality*. Ph.D. thesis. University of California, Berkeley, December 2002.

[Fowler03] M. Fowler. *UML Distilled: A Brief Guide to the Standard Object Modeling Language, Third Edition*. Addison-Wesley Professional, September 2003.

[Gode97] P. Godefroid. Model Checking for Programming Languages Using VeriSoft. *Symposium on Principles of Programming Languages*, January 1997.

[Godfrey00] M. W. Godfrey and Q. Tu. Evolution in Open Source Software: A Case Study. *International Conference on Software Maintenance*, October 2000.

[Gold67] E. Gold. Language Identification in the Limit. *Information and Control*, Volume 10, 1967.

[Gold78] E. Gold. Complexity of Automatic Identification from Given Data. *Information and Control*, Volume 37, 1978.

[Gries81] D. Gries. *The Science of Programming*. Springer Verlag, New York, 1981.

[Grov01] D. Grove and C. Chambers. A Framework for Call Graph Construction Algorithms. *ACM Transactions on Programming Languages and Systems*. Volume 23, November 2001.

[Gupta03] N. Gupta. Generating Test Data for Dynamically Discovering Likely Program Invariants. *Workshop on Dynamic Analysis*, May 2003.

[Hack06] B. Hackett, M. Das, D. Wang, and Z. Yang. Modular Checking for Buffer Overflows in the Large. *International Conference on Software Engineering*, May 2006.

[Hangal02] S. Hangal and M. S. Lam. Tracking Down Software Bugs Using Automatic Anomaly Detection. *International Conference on Software Engineering*, May 2002.

[Hard03] M. Harder, J. Mellen, and M. D. Ernst. Improving Test Suites via Operational Abstraction. *International Conference on Software Engineering*, May 2003.

[Have04] K. Havelund and G. Rosu. An Overview of the Runtime Verification Tool Java PathExplorer. *Formal Methods in System Design*, Volume 24, 2004.

[Henz03] T. A. Henzinger, R. Jhala, R. Majumdar, and S. Qadeer. Thread-Modular Abstraction Refinement. *International Conference on Computer-Aided Verification*, July 2003.

[Henkel04] J. Henkel. Discovering and Debugging Algebraic Specifications for Java Classes. PhD Dissertation, University of Colorado at Boulder, May 2004.

[Hoar69] C. A. R. Hoare. An Axiomatic Basis for Computer Programming. *Communications of the ACM*, October 1969.

[Hoar03] C. A. R. Hoare. The Verifying Compiler: a Grand Challenge for Computing Research. *Journal of the ACM*, Volume 50, January 2003.

[Holloway96] C. M. Holloway and R. W. Butler. Impediments to Industrial Use of Formal Methods. *IEEE Computer*, April 1996.

[Holz97] G. J. Holzmann. The Model Checker Spin. In *IEEE Transactions on Software Engineering*, Volume 23, May 1997.

[Holz02] G. J. Holzmann. The Logic of Bugs. *Symposium on Foundations of Software Engineering*, November 2002.

[Horw90] S. Horwitz. Identifying the Semantic and Textual Differences between Two Versions of a Program. *Conference on Programming Language Design and Implementation*, June 1990.

[Horw94] S. Horwitz and T. Reps. The Use of Program Dependence Graphs in Software Engineering. *International Conference on Software Engineering*, May 1994.

[J2EE] J2EE. *http://java.sun.com/j2ee*

[Jack94] D. Jackson and D. Ladd. Semantic Diff: a Tool for Summarizing the Effects of Modifications. *International Conference on Software Maintenance*, October 1994.

[JBoss] JBoss. *http://www.jboss.org*

[JRat] JRat. *http://jrat.sourceforge.net*

[JTA] Java Transaction API specification. *http://java.sun.com/products/jta*

[Karp72] R. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*. 1972.

[Khurshid02] S. Khurshid and D. Marinov. TestEra: A Novel Framework for Automated Testing of Java Programs. *Automated Software Engineering Journal*, December 2002.

[Knight97] J. C. Knight, C. L. DeJong, M. S. Gibble, and L. G. Nakano. Why are Formal Methods Not Used More Widely? *NASA Langley Formal Methods Workshop*, September 1997.

[Kröger87] F. Kröger. *Temporal Logic of Programs*. Springer-Verlag New York, 1987.

[Lamsw00] A. Lamsweerde. Formal Specification: a Roadmap. In *International Conference on Software Engineering*, May 2000.

[Larus04] J. R. Larus, T. Ball, M. Das, R. DeLine, M. Fahndrich, J. Pincus, S. K. Rajamani, and R. Venkatapathy. Righting Software. *IEEE Software*, May/June 2004.

[Leto86] S. Letovsky and E. Soloway. Delocalized Plans and Program Comprehension. *IEEE Software*, May 1986.

[Li04] Z. Li, S. Lu, S. Myagmar, and Y. Zhou. CP-Miner: a Tool for Finding Copy-Paste and Related Bugs in Operating System Code. *Symposium on Operating System Design and Implementation*, December 2004.

[Libl05] B. Liblit, M. Naik, A. X. Zheng, A. Aiken, and M. I. Jordan. Scalable Statistical Bug Isolation. *Conference on Programming Language Design and Implementation*, June 2005.

[Lin04] L. Lin and M. D. Ernst. Improving Adaptability via Program Steering. *International Symposium on Software Testing and Analysis*, July 2004.

[Lindholm99] T. Lindholm and F. Yellin. *The Java Virtual Machine Specification, Second Edition*. Addison-Wesley Professional, April 1999.

[Liu05] C. Liu, X. Yan, L. Fei, J. Han, and S. P. Midkiff. SOBER: Statistical Model-Based Bug Localization. *Symposium on the Foundations of Software Engineering*, September 2005.

[Livs05] B. Livshits and T. Zimmermann. DynaMine: Finding Common Error Patterns by Mining Software Revision Histories. *Symposium on the Foundations of Software Engineering*, September 2005.

[Mande05] D. Mandelin, L. Xu, R. Bodik, and D. Kimelman. Mining Jungloids: Helping to Navigate the API Jungle. *Conference on Programming Language Design and Implementation*, June 2005.

[Mari01] D. Marinov, and S. Khurshid. TestEra: A Novel Framework for Automated Testing of Java Programs. *International Conference on Automated Software Engineering*, November 2001.

[Memon01] A. M. Memon, M. E. Pollack, and M. L. Soffa. Hierarchical GUI Test Case Generation Using Automated Planning. *IEEE Transactions on Software Engineering*. Volume 27, February 2001.

[Mill85] W. Miller and E. W. Myers. A File Comparison Program. *Software - Practice and Experience*, Volume 15, 1985.

[Mitch03] N. Mitchell and G. Sevitsky. LeakBot: An Automated and Lightweight Tool for Diagnosing Memory Leaks in Large Java Applications. *European Conference on Object-Oriented Programming*, July 2003.

[Musu04] M. Musuvathi and D. Engler. Model-Checking Large Network Protocol Implementations. *Conference on Network System Design and Implementation*, March 2004.

[Meyer97] B. Meyer. *Object-Oriented Software Construction, Second Edition*. Prentice Hall, 1997.

[Neam05] I. Neamtiu, J. S. Foster, and M. Hicks. Understanding Source Code Evolution Using Abstract Syntax Tree Matching. *International Workshop on Mining Software Repositories*, May 2005.

[Nimm02] J. W. Nimmer and M. D. Ernst. Invariant Inference for Static Checking: an Empirical Evaluation. *Symposium on the Foundations of Software Engineering*, November 2002.

[OpenS] OpenSSL. *http://www.openssl.org*

[OpenSec] OpenSSL security advisory, 17 March 2004. http://www.openssl.org/news/secadv_20040317.txt

[Owic82] S. Owicki and L. Lamport. Proving Liveness Properties of Concurrent Programs. *ACM Transactions on Programming Languages and Systems*, July 1982.

[Parnas72] D. L. Parnas. On the Criteria to be Used in Decomposing System into Modules. *Communications of the ACM*, Volume 15, December 1972.

[Perk04] J. Perkins and M. Ernst. Efficient Incremental Algorithms for Dynamic Detection of Likely Invariants. *International Symposium on Foundations of Software Engineering*, November 2004.

[Prat06] P. Pratikakis, J. S. Foster, and M. Hicks. Context-sensitive Correlation Analysis for Detecting Races. In ACM Conference on Programming Language Design and Implementation, June 2006.

[Pnueli77] A. Pnueli. The Temporal Logic of Programs. *Symposium on Foundations of Computer Science*, October/November 1977.

[Pytlik03] B. Pytlik, M. Renieris, S. Krishnamurthi, and S. Reiss. Automated Fault Localization Using Potential Invariants. *International Symposium on Automated and Analysis-Driven Debugging*. September 2003.

[Reiss00] S. P. Reiss and M. Renieris. Encoding Program Executions. *International Conference on Software Engineering*, May 2001.

[Reps95] T. Reps, S. Horwitz, and M. Sagiv. Precise Inter-Procedural Dataflow Analysis via Graph Reachability. *Symposium on Principles of Programming Languages*, January 1995.

[Resc01] E. Rescorla. An Introduction to OpenSSL Programming, Part One. http://www.rtfm.com/ openssl-examples/, October 2001.

[Rice53] H. G. Rice. Classes of Recursively Enumerable Sets and Their Decision Problems. Transactions of the American Mathematics Society, Volume 74, 1953.

[Rose05] J. Rose, N. Swamy, and M. Hicks. Dynamic Inference of Polymorphic Lock Types. *Science of Computer Programming*, Volume 58, 2005.

[Somm00] I. Sommerville. *Software Engineering, Sixth Edition*. Addison Wesley, 2000.

[Simo] C. Simonyi. Hungarian notation. MSDN library. *http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnvs600/html/HungaNotat.asp*

[Spivey92] J. M. Spivey. *The Z Notation: A Reference Manual, Second Edition*. Prentice-Hall, 1992.

[SSL] SSL Specification, Third Version. *http://wp.netscape.com/eng/ssl3*

[SDV] Static Driver Verifier: Finding Bugs in Device Drivers at Compile-Time. *Windows Hardware Engineering Conference*, April 2004.

[Sriva01] A. Srivastava, A. Edwards, and H. Vo. Vulcan: Binary Transformation in a Distributed Environment. *Microsoft Research Technical Report*, MSR-TR-2001-50, April 2001.

[Stro86] R. E. Strom and S. Yemini. Typestate: a Programming Language Concept for Enhancing Software Reliability. *IEEE Transactions on Software Engineering*, Volume 12, January 1986.

[SunAS] The Sun Java System Application Server. http://www.sun.com/software/products/appsrvr/index.xml

[Swif03] M. Swift, B. N. Bershad, and H. M. Levy. Improving the Reliability of Commodity Operating Systems. *Symposium on Operating Systems Principles*, October 2003.

[Viss03] W. Visser, K. Havelund, G. Brat, S. Park, and F. Lerda. Model Checking Programs. *Automated Software Engineering Journal*, April 2003.

[WebLogic] The BEA WebLogic Application Server. http://www.beasys.com/products/weblogic/

[WebSphere] The IBM WebSphere Application Server. http://www.ibm.com/websphere/

[Weimer05] W. Weimer and G. Necula. Mining Temporal Specifications for Error Detection. *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, April 2005.

[Weyu80] E. J. Weyuker and T. J. Ostrand. Theories of Program Testing and the Application of Revealing Subdomains. *IEEE Transactions on Software Engineering*, May 1980.

[Whaley02] J. Whaley, M. C. Martin, and M. S. Lam. Automatic Extraction of Object-Oriented Component Interfaces. *International Symposium on Software Testing and Analysis*, July 2002.

[Win04] T. N. Win, M. D. Ernst, S. J. Garland, D. Kirli, and N. Lynch. Using Simulated Execution in Verifying Distributed Algorithms. *Software Tools for Technology Transfer*, July 2004.

[Xie06] T. Xie and D. Notkin. Tool-Assisted Unit-Test Generation and Selection Based on Operational Abstractions. *Automated Software Engineering Journal*, Volume 13, July 2006.

[Yang04a] J. Yang and D. Evans. Dynamically Inferring Temporal Properties. *Workshop on Program Analysis for Software Tools and Engineering*, June 2004.

[Yang04b] J. Yang and D. Evans. Automatically Inferring Temporal Properties for Program Evolution. *International Symposium on Software Reliability Engineering*, November 2004.

[Yang04c] Junfeng Yang, P. Twohey, D. Engler, and M. Musuvathi. Using Model Checking to Find Serious File System Errors. *Symposium on Operating System Design and Implementation*, December 2004.

[Yang05] J. Yang and D. Evans. Automatically Discovering Temporal Properties for Program Verification. Technical Report, Department of Computer Science, University of Virginia, 2005.

[Yang06] J. Yang, D. Evans, D. Bhardwaj, T. Bhat, and M. Das. Terracotta: Mining Temporal API Rules from Imperfect Traces. *International Conference on Software Engineering*, May 2006.

[Zhou04] P. Zhou, W. Liu, F. Long, S. Lu, F. Qin, Y. Zhou, S. Midkiff, and J. Torrellas. AccMon: Automatically Detecting Memory-Related Bugs via Program Counter-Based Invariants. *International Symposium on Micro-architecture*, December 2004.

Appendix A
Inferred Windows Properties

P_{AL}	Property
1.00	ExAcquireFastMutex→ExReleaseFastMutex
1.00	ExAcquireRundownProtectionCacheAwareEx→ ExReleaseRundownProtectionCacheAwareEx
1.00	HMFreeObject→HMAllocObject
1.00	HvpGetCellMapped→HvpReleaseCellMapped
1.00	IoAcquireVpbSpinLock→IoReleaseVpbSpinLock
1.00	KeAcquireQueuedSpinLock→KeReleaseQueuedSpinLock
1.00	KefAcquireSpinLockAtDpcLevel→KefReleaseSpinLockFromDpcLevel
1.00	KeInitThread→KeStartThread
1.00	KeSuspendThread→KeResumeThread
1.00	KfAcquireSpinLock→KfReleaseSpinLock
1.00	KiAcquireSpinLock→KiReleaseSpinLock
1.00	LdrLockLoaderLock→LdrUnlockLoaderLock
1.00	MiMapPageInHyperSpace→MiUnmapPageInHyperSpace
1.00	MiSecureVirtualMemory→MiUnsecureVirtualMemory
1.00	MmSecureVirtualMemory→MmUnsecureVirtualMemory
1.00	NtfsAcquireFileForCcFlush→NtfsReleaseFileForCcFlush
1.00	NtGdiDdGetDC→NtGdiDdReleaseDC
1.00	NtUserBeginPaint→NtUserEndPaint
1.00	ObpAllocateObjectNameBuffer→ObpFreeObjectNameBuffer
1.00	PopAcquirePolicyLock→PopReleasePolicyLock
1.00	RtlAcquirePebLock→RtlReleasePebLock
1.00	RtlAcquireSRWLockExclusive→RtlReleaseSRWLockExclusive
1.00	RtlActivateActivationContext→RtlDeactivateActivationContext
1.00	RtlDeleteTimer→RtlCreateTimer
1.00	RtlLockHeap→RtlUnlockHeap
1.00	RtlpAllocateActivationContextStackFrame→RtlpFreeActivationContextStackFrame
1.00	RtlpAllocateUserBlock→RtlpFreeUserBlock
1.00	RtlpFindFirstActivationContextSection→RtlFindNextActivationContextSection
1.00	RtlValidSid→RtlCopySid
1.00	SeCaptureSid→SeReleaseSid
1.00	SeLockSubjectContext→SeUnlockSubjectContext
1.00	xxxBeginPaint→xxxEndPaint

P_{AL}	Property
0.99	ObpCreateHandle→ObpCloseHandle
0.99	_GetDC→ _ReleaseDC
0.99	RtlpRemoveListLookupEntry→RtlpAddListLookupEntry
0.99	GreLockDisplay→GreUnlockDisplay
0.99	RtlActivateActivationContextUnsafeFast→ RtlDeactivateActivationContextUnsafeFast
0.98	CmpRemoveFromDelayedClose→CmpAddToDelayedClose
0.98	KeAcquireInStackQueuedSpinLock→KeReleaseInStackQueuedSpinLock
0.98	SeCreateAccessState→SeDeleteAccessState
0.98	KeAcquireInStackQueuedSpinLockRaiseToSynch→ KeReleaseInStackQueuedSpinLockFromDpcLevel
0.97	IoAllocateIrp→IoFreeIrp
0.96	CmpLockRegistry→CmpUnlockRegistry
0.96	ObAssignSecurity→ObDeassignSecurity
0.96	VirtualAllocEx→VirtualFreeEx
0.95	ExCreateHandle→ExDestroyHandle
0.95	ExpAllocateHandleTableEntry→ExpFreeHandleTableEntry
0.95	CmpFreeDelayItem→CmpAllocateDelayItem
0.94	ExInitializeResourceLite→ExDeleteResourceLite
0.94	RtlEnterCriticalSection→RtlLeaveCriticalSection
0.93	MiGetPreviousNode→MiGetNextNode
0.92	RtlValidAcl→RtlCreateAcl
0.92	ObFastReferenceObject→ObFastDereferenceObject
0.92	PsChargeProcessPoolQuota→PsReturnSharedPoolQuota
0.91	EtwpInitializeDll→EtwpDeinitializeDll
0.90	IoFreeMdl→IoAllocateMdl