ARC++: Effective Typestate and Lifetime Dependency Analysis

Xusheng Xiao¹,² xxiao2@ncsu.edu
Gogul Balakrishnan¹ bgogul@gmail.com
Naoto Maeda³ n-maeda@bp.jp.ne.com
Aarti Gupta¹ agupta@nec-labs.com
Deepak Chhetri¹ deepak.chhetri@nectechnologies.in
¹NEC Labs America, USA ²North Carolina State University, USA ³NEC Corporation, Japan
⁴NEC Technologies India Ltd., India ⁵Google, Inc., USA

ABSTRACT

The ever-increasing reliance of today’s society on software requires scalable and precise techniques for checking the correctness, reliability, and robustness of software. Object-oriented languages have been used extensively to build large-scale systems, including Java and C++. While many scalable static analysis approaches for C and Java have been proposed, there has been comparatively little work on the static analysis of C++ programs. In this paper, we provide an abstract representation to model C++ objects, containers, references, raw pointers, and smart pointers. Further, we present a new analysis called lifetime dependency analysis, which allows us to precisely track the complex lifetime semantics of temporary objects in C++. Finally, we propose an implementation of our techniques and present promising results on a large variety of open-source software.

Categories and Subject Descriptors
D.2.4 [Software Engineering]: Software/Program Verification; D.2.5 [Software Engineering]: Testing and Debugging—Debugging aids

General Terms
Verification, Reliability

Keywords
Bug finding, C++, abstract representation, typestate analysis, static analysis

1. INTRODUCTION

The ever-increasing reliance of today’s society on software requires scalable and precise techniques for checking the correctness, reliability and robustness of software. Recently, automatic approaches, such as automated test generation [12, 20, 31] and static software verification [7, 22, 19, 23], have been used to detect bugs in software and improve software quality. This paper targets the static analysis of object-oriented programs written in C++ using light-weight abstract interpretation techniques by looking for common programming mistakes that are specific to C++.

In the past decade, there has been strong interest in developing practical techniques to find bug patterns [21, 32, 34]. Bug patterns are code idioms that generally lead to various types of software issues, such as run-time failures, performance issues, etc. Such code idioms are targeted because they address an often observed misunderstanding of the language semantics or run-time behavior [21]. Recently, it was also shown that bug patterns can be used to detect non-functional bugs, such as performance bugs [24].

Additionally, in systems built using object-oriented languages, such as C++ or Java, object protocols or typestates [10, 16, 39] have been found to be a useful abstraction for developing and maintaining large software projects. They have also been effectively used to capture errors arising from misuse of various operations on objects. Typestates refer to an abstraction of a concrete object state, and they may change due to the operations performed on the object [39]. For example, a file object $f$ may only be read after being properly opened. Typestates are usually modeled as finite state automata, and can be used to detect bugs by searching for violations of usage constraints [8]. In this paper, we represent bug patterns as typestate automata and seek to develop an effective analysis that can support typestate checking in C++ programs.

Motivation: Static Analysis of C++.

Software development teams have shifted their development from C to object-oriented languages, including C++ and Java. The benefits of using an object-oriented language include reusability, better maintainability, encapsulation, and the use of inheritance. In particular, C++ is often chosen due to its ability to interact with legacy C-based systems, including system-level C libraries. Thus, development in C++ often necessitates a mixed programming style combining object-oriented constructs with lower-level C code. Whereas a large volume of work on verification has focused on C or Java, there has been comparatively little work on the verification of C++ programs.

Handling C++ programs using static analysis is non-trivial because of the complexities of the C++ semantics. One particular issue in handling C++ compared to other object-oriented programming languages such as Java is the com-
void test(auto_ptr<int> p);  
int main() {  
    auto_ptr<int> p(new int);  // initializes p  
test(p);  // transfer ownership from p to formal parameter  
return *p;  // segmentation fault  
}  

Figure 1: Example C++ program using auto_ptr.

Complexity of allowed class hierarchies. C++ allows multiple inheritance, where a class may inherit from more than one class. Furthermore, multiple inheritance complicates the semantics of otherwise simple operations such as casts and field accesses. Therefore, techniques developed for Java are not readily applicable to C++ programs. It is important to emphasize that multiple inheritance is used quite frequently by developers even in large C++ projects. We developed the C++CHROME object representation to address some of this challenge for software model checking in our earlier work [11].

Exceptions. Another difference between Java and C++ semantics is the handling of exceptions. The exception specification and checking mechanism in Java is more strongly enforced than in C++. Indeed, in C++, exception specifications other than an indication that a function does not throw any exception, are discouraged due to negative performance impact and the possibility that they are inaccurate (in which case an overloaded function std::unexpected() is called). A major difference from Java is that, in C++, when an exception is passed to another scope or leaves a functional scope, all stack allocated objects along the way to the appropriate exception handler need to be deallocated using a process called stack unwinding. During stack unwinding, object destructors are called, which may generate further exceptions, leading to a scenario with multiple live exceptions. We address the issue of exceptions by building an inter-procedural exception control flow graph (ICCFG) [13].

Standard libraries. Another challenge is the use of standard libraries, such as STL and BOOST. These libraries introduce containers (such as vector and stack) and new types of pointers (such as iterators and smart pointers), and heavily use templates or operator overloading to define new operations around these new container and pointers. For example, it = vector.begin() initializes the iterator it (a new kind of pointer) to point to the first item of a vector.

Consider the C++ program shown in Fig. 1. The program has two functions, test and main. The function test accepts an auto pointer ptr, and performs certain operations using the memory managed by ptr (not shown here). An auto pointer is a wrapper class of a pointer to an object, that ensures that the pointed to object is destroyed when control leaves the auto pointer scope. The function main allocates memory for an auto pointer p and makes a call to test (line 4). After the call to test finishes, the program modifies the value in the memory managed by p (line 5).

Although this simple program seems semantically correct, its execution results in a segmentation fault at line 5. The reason is that when an assignment operation or a copy construction takes place between two auto_ptr objects, ownership of the memory is transferred, which means that the auto_ptr object losing ownership is reset to null. Since the parameter ptr of test is declared as pass-by-value, a copy constructor is invoked to initialize ptr using p at line 4. Thus, after line 4, p becomes null, and an error occurs when the program tries to dereference p at line 5.

void test(auto_ptr<int> p);  
int main() {  
    int *tmp1;  // Declartions only, i.e.  
    auto_ptr<int> p, tmp2;  // not C++ style semantics  
tmp1 = new int;  // start of Line 3  
Use(tmp1);  // argument in constructor  
Create(p,tmp1);  // constructor for p  
Use(Addr(p));  // start of Line 4: calls copy constructor  
Create(tmp2,Addr(p));  // copy constructor for argument  
Destroy(p);  // ownership transferred to formal parameter  
Call(test,tmp2);  // function call  
Use(p);  // Line 5: Use—after—destroy (error)  
}  

Figure 2: ARC++ representation for Fig. 1.

To detect this bug, a static analysis requires a model of the implicit copy constructor and its side-effect on p. To model the high-level semantics of such libraries, we propose an abstract representation for C++ statements called Arc++ (Abstract Representation for C++). Fig. 2 shows the ARC++ representations for the program shown in Fig. 1.

Temporary objects. One of the peculiarities of the C++ semantics is related to the notion of lifetimes of temporary objects. In C++, temporary objects are often created by the compiler, causing performance and correctness issues that are often hard to find and understand. Temporary objects are unnamed objects created on the stack by the compiler. They are used during reference initialization and during evaluation of expressions including standard type conversions, argument passing, function returns, and evaluation of the throw expression. Performance bottlenecks can arise due to the unnecessary creation and destruction of such temporary objects. Correctness issues can arise due to the complex lifetime semantics of temporary objects often leading to accesses of previously freed/deleted memory.

Generally, the correctness issues related to object lifetimes are hidden during testing due to the fact that stale uses of object storage often occur shortly after destruction of the object. Nevertheless, in an actual deployed production environment, such short-term stale uses can cause hard to find runtime errors and memory corruption, leading to memory faults. Furthermore, such memory corruption can also be potentially exploited by malicious users.

Low-level memory management. Furthermore, C++ needs to deal with legacy C semantics, APIs and programs, with low-level memory management issues that are not of concern for Java. One such issue is interaction with the semantics of temporary objects. In a prior effort, we have used model checking for finding interaction bugs involving STL strings and C-style strings [6]. A very short example is shown in Fig. 3. The call to s.substr() returns a temporary string object. The call to c_str() on that temporary object returns an internal pointer into the temporary string object, which represents a properly terminated C string. However, the temporary string object is destroyed after its lifetime, which is right after the call to c_str() in this case. Thus, the call to strlen in the following line is performed on a C string that has just been deallocated. To deal with such issues, this paper introduces a notion of lifetime dependency that captures the dependency of the validity of one object on the validity of another object, and we use a light-weight static analysis instead of model checking, which is more expensive.
Our approach.

This paper presents an abstract representation called Arc++ for modeling C++ objects as well as containers and smart pointers introduced by standard libraries. The main goal of Arc++ is to make object creation, destruction, usage, lifetime, and pointer operations explicit in the abstract model. The Arc++ representation provides the basis for a novel static analysis that we call lifetime dependency analysis to support typestate checking for C++ objects. Arc++ is also used for specifying common bug patterns in the form of typestate automata on these objects. We have developed an automatic framework to generate Arc++ representations from C++ programs and use our novel static analysis to perform typestate checking for many bug patterns.

Abstract representation. Each statement in C++ is mapped to a list of abstract instructions in Arc++. For example, Fig. 2 shows the Arc++ representation for the program in Fig. 1. Arc++ captures the ownership transfer of auto_ptr’s copy constructor and generates a Destroy statement for p, enabling static program analysis to detect the bug. Note that besides object creation and destruction, Arc++ also handles C++ references and pointer operations of containers and smart pointers, which have different semantics with respect to pointer aliasing (such as it = vector.begin()). Moreover, we utilize the results of the exceptional flow analysis [33], enabling developers to compose bug patterns related to exceptions.

Lifetime dependency analysis. The lifetime semantics of objects in C++ is quite complicated and can easily lead to correctness or performance issues. One such issue is due to the interaction of low-level memory management with the lifetimes of temporary objects. We use Arc++ to highlight the lifetimes of various objects. However, as we will show for the above example, many times the analysis of such C++ programs can be analyzed precisely by considering a new static analysis that we call lifetime dependency analysis.

To handle the lifetime semantics of C++, we consider lifetime dependencies between objects such that the destruction of one of the objects immediately implies the destruction of other dependent objects. For example, for the program shown in Fig. 3, we would introduce a lifetime edge for the unnamed temporary object returned from the substr() method call and the C string pointer str. At the time of the temporary object destruction, which occurs after the call to c_str(), we follow lifetime dependency edges to other implicitly destructed objects or access paths.

Contributions.

- We present an abstract representation (Arc++) that models C++ objects as well as containers, and smart pointers introduced by standard libraries. Arc++ makes object creation, destruction, usage, lifetime, and pointer operations explicit in the abstract model, providing a basis for static analysis on C++ programs.
- We propose a notion of “lifetime dependency” to track implied destructions between objects. We show that this provides an effective high-level abstraction for many types of issues involving temporary objects and internal buffers. This is used in a novel flow-sensitive, context-sensitive, and path-insensitive static analysis that supports typestate checking for C++ objects.
- We propose a framework that (i) automatically generates Arc++ representations from C++ programs, and (ii) performs typestate checking to detect bugs that are specified as typestate automata over Arc++ representations.
- We present five bug patterns using Arc++ and experimental results for checking them on 20 open-source C++ programs of more than 2M LOC. The results show that our framework is very fast and detected 23 bugs and 124 bad programming practices.

2. Arc++ Abstraction

This section describes the details of our abstract representation called Arc++. Each statement of a C++ program is mapped to a list of abstract instructions in Arc++ automatically. These abstract instructions can be directly used for bug pattern specification and checking. Fig. 4 presents a subset of Arc++.

Expressions. Arc++ expressions represent program variables, field accesses, cast operations, return values of function calls, and functional computations without side effects or control flow. The parameters e and type refer to the expressions and types in a C++ program. A Variable expression represents program variables of primitive or user-defined types. A Pointer expression represents C pointers, C++ references, iterators of containers, and the various pointer classes (e.g., shared_ptr) in the standard library. Note that an iterator is conceptually treated as a pointer because an iterator may point to some element of the associated container. A Container expression is associated with an lvalue or a rvalue expression that represent containers. A container is either a C++ array or one of Vector, List, Stack, Map, and Set container types in the standard C++ library. The type argument in a Container expression represents the type of elements stored in the container. A Call expression represents the return value of either a global function call or a member function call. Besides the type of the return value, a Call expression also provides the information about the receiver object and an argument list.

Abstract instructions. Arc++ abstract instructions represent operations on expressions. These include object operations (Create, Destroy or Use), lifetime operations, and others. A Create instruction represents a call to the constructor of a class or a memory allocation for a pointer (such as new/malloc). A Destroy instruction represents a call to the destructor of a class, a memory deallocation (such as delete/free), or the reset on pointers (e.g., set pointer to null). Note that the implicit calls to copy constructors, destructors, and pointer resets (such as auto_ptr in Fig. 1) are made explicit and the corresponding abstract instructions are generated.

Use operations. A Use instruction represents how the value of an object is accessed. A Value operation represents that the underlying expression is a simple variable, a Dereference operation signifies that the underlying expression is a pointer dereference, a Field operation represents an access to a field of an object, and an Index operation represents that the underlying expression is an index into a container object, such as an array or a vector.
Expressions:
SPtrType ::= AutoPtr | UniquePtr | SharedPtr
PtrType ::= CPtr | Reference | Iterator | SPtrType
ContainerType ::= Array | Vector | List | Stack | Map | Set
Expr ::= Variable(e, type)
| Pointer(PtrType, Expr, type)
| Container(e, ContainerType, type)
| Call(e, type, Expr option, Expr list)

Figure 4: ARC++ Abstraction (Abstract Representation for C++). The terminal symbols e and type represents the expression and types in the C++ program, respectively.

Lifetime operations. The operation link(e₁, e₂) links the lifetime of e₁ with the operations of e₂. After the link operation, an invalidateLinked(e₂) operation on e₂ causes the invalidation of e₁. For example, the link(it, v) operation can be used to represent the fact that after the statement it = v.begin(), the lifetime of iterator it is linked with the lifetime of the vector v. Similarly, invalidateLinked(v) can be used to represent the fact that v.push_back() invalidates the iterators that are linked with v.

Dependency labels for member functions. To generalize the bug detector for objects of user-defined classes, we automatically attach a dependency label to a member function that returns a pointer or a reference to the receiver object’s internal buffer. For ease of presentation, we only consider fields of receiver objects as potential candidates.

Def. 2.1. Given a member function foo returning a pointer or reference, the call p = r.foo() should be treated as an abstract operation that makes p dependent on the receiver r, if the memory accessible from a returned pointer or reference is destroyed whenever r is destroyed or cleared.

Based on the above definition, if a field f of an object satisfies any of the following conditions, our technique attaches a dependency label to a member function that returns a pointer or a reference and satisfies one of these conditions:
1. If the member function returns a field f pointing to dynamically allocated memory whose lifetime is kept by the object (i.e., deleted in the object’s destructor).
2. If the member function returns the address or reference of a field f, where the type of f is either an array or a user-defined class.

Invalidate operations for member functions. If a class C has a method that returns an internal buffer, then for a Destroy(o) operation associated with an object o of class C, the invalidateLinked(o) operation is also generated. We also generalize the invalidateLinked operation to other user-defined methods. If a method foo of a class reallocates any of the internal buffers that are returned by other methods of the class, then the method call a.foo() also causes the invalidateLinked(a) to be generated.

3. TYPESTATES AND BUG PATTERNS

A bug pattern is specified as a finite state automaton over observable program operations. An observable operation is either a C++ statement or an ARC++ abstract instruction that changes the typestate corresponding to a bug pattern.

Def. 3.1 (Bug Pattern). A bug pattern P is a tuple P = (Q, Σ, τ, init, error), where Q is set of automaton states, Objects o ∈ O
Identifiers id ∈ Identifiers
Labels l ∈ Labels
Fields fld ∈ Fields
Operations op ∈ Operations
Program pgm ::= proc
Procedure proc ::= id(id){stmt}
Access path ap ::= id | *ap | ap.fld
Statement stmt ::= ap := ap | ap.op(âp) | call id(âp)
| ap := &o | goto l₁ l₂ | l : stmt

Σ is the set of observable operations in a program, τ ∈ (Q × Σ × Q) is the transition function mapping a state and an operation to a successor state, init ∈ Q is the unique initial state, and error ∈ Q is the unique error state.

Our implementation described in operates on C++ programs annotated with ARC++. However, to illustrate our approach, we use the Mini language shown in Fig. We assume a set of identifiers Identifiers, a set of fields Fields, a (possibly infinite) set of dynamically allocated objects O, and a set of observable operations Operations over which the bug patterns are specified. A Mini language program pgm consists of a list of procedures proc, and each proc has a unique identifier id ∈ Identifiers, a list of arguments and statements. In a Mini program, all objects are accessed through an access path, which is either an identifier, a memory dereference, or a field access. In the case of cyclic data structures, we limit the length of access paths by a constant. A Mini language statement stmt is either an assignment from one access path to another, an observable operation, a call, initialization of an access path with the address of an object, non-deterministic goto, or a labeled statement. Let AP be the set of all access paths in the program, and Proc be the set of all procedures in a Mini program. We assume that transitive assignments induced by pointer dereferences and field accesses are present explicitly.

Def. 3.2 (CFG). A procedure p ∈ Proc in the Mini language program is represented by a control-flow graph (CFG), which is a directed graph Gp = (Np, Ep), where Np is a set of CFG nodes and Ep ⊆ Np × Np is a set of CFG edges. The set of nodes Np consists of a unique start node sp ∈ Np, a unique exit node ep ∈ Np, and nodes associated with each statement in the program.

Let N = ⋃p∈Proc Np be the set of all nodes in the program, E = ⋃p∈Proc Ep be the set of all edges in the program,
4. ACCESS-PATH CLUSTERS ABSTRACT DOMAIN

Since it is computationally intractable to compute the set of all concrete states precisely for a MINI program, we use abstract interpretation [14] to efficiently analyze programs. Abstract interpretation requires an abstract domain \( D \). An abstract domain \( D \) is a lattice \((D, \sqcup, \sqcap)\) along with an abstraction map \( \alpha : 2^S \rightarrow D \) and a concretization map \( \gamma : D \rightarrow 2^S \).

One of the main challenges in defining an abstract domain for programs with pointers lies in incorporating aliasing relationships into the representation of the abstract object so that strong updates can be performed for common cases. A strong update replaces the current abstract value with a new abstract value, and is performed when the abstract interpreter can determine that the abstract state being updated corresponds to a unique concrete state. On the other hand, a weak update only appends the new abstract value to the current abstract value. A large number of weak updates can adversely affect the precision resulting in false positives.

Our representation for an abstract object state is inspired by the work on aliasing-aware typestate verification developed for Java by Fink et al. [15]. Fink et al. use a notion of access paths to objects together with focus operations that split heap regions to selectively allow strong updates on certain abstract objects using these access paths. Fink et al. consider must and must-not access paths, such that successor abstract objects may be split up to gain a certain level of path-sensitivitiy [18]. We briefly present an adaption of their domain for our purposes, and in [14] describe how we extend it to deal with objects whose types are inter-dependent (such as LifeOp in [12]).

To define an abstract object, we assume that results of a pointer analysis such as Steensgaard’s [38] or Andersen’s [4] are available, and it supports the following operations:

- **mayPtsTo** \( (a, n) \) is the set of concrete objects pointed to by the access path \( a \) at node \( n \).
- **alias** \( (a_1, a_2, n) \) is the set of true if \( a_1 \) and \( a_2 \) are possible aliases at node \( n \), and \( alias(a_1, a_2) \) is the flow-insensitive version.

We define several concepts necessary for the abstraction.

**Def. 4.1 (Alias Cluster).** Let \( AP \) be the set of all access paths in the program. An alias cluster \( c \subseteq AP \) is a set of access paths such that \( \forall a_1, a_2 \in c, alias(a_1, a_2) \). An alias cluster \( c \) represents the set of objects pointed-to by the access paths in the cluster:

\[
[c] = \{ o \in O \mid \exists a \in c, o \in mayPtsTo(a) \}
\]

The universe of alias clusters \( C \subseteq 2^{AP} \) is a partitioning of the access paths \( AP \) into alias clusters such that all of the following hold: (1) \( \forall c \in C, c \) is an alias cluster, (2) \( \cup c \subseteq (c) = AP \), and (3) \( \forall c_1, c_2 \in C, c_1 \cap c_2 = \emptyset \Rightarrow [c_1] \cap [c_2] = \emptyset \).

In other words, \( C \) is a set of alias clusters such that each alias cluster \( c \subseteq C \) represents a distinct set of concrete objects.

**Def. 4.2 (Abstract Object).** An abstract object \( o^a \) is a tuple \((c, \mu, \nu)\), where \( c \in C \) is an alias cluster and \( \mu, \nu \subseteq AP \) is a set of access paths. An abstract object \( o^a \) characterizes a subset of concrete objects represented by the alias cluster \( c \). Specifically, it represents the set of objects that are definitely pointed to by the access paths in \( \mu \) (must set), but are not pointed to by the access paths in \( \nu \) (must-not set).

**Example 4.1.** Fig. 6 shows three different abstract objects. Let \( c = \{a_1, a_2, a_3\} \) be an alias cluster, \( [c] \) be the corresponding set of abstract objects (denoted by the three circles in Fig. 6). The must and must-not sets in an abstract object partition the objects of an alias cluster as shown in Fig. 6.

The abstract object \( o^a \) is a tuple \((c, \mu, \nu)\), where \( c \in C \) is an alias cluster and \( \mu, \nu \subseteq AP \) is a set of access paths. An abstract object \( o^a \) characterizes a subset of concrete objects represented by the alias cluster \( c \). Specifically, it represents the set of objects that are definitely pointed to by the access paths in \( \mu \) (must set), but are not pointed to by the access paths in \( \nu \) (must-not set).

We define exclusive accessibility to cover cases where strong updates can be performed.

**Def. 4.3 (Accessibility).** An access path \( a \) has access to abstract object \( o^a \) if \( a \triangleright o^a \), denoted as \( a \triangleright o^a \), if one of the following conditions hold: (1) \( a \in \mu \), or (2) \( a \notin \mu \land a \notin \nu \). Otherwise, we say \( a \) has no access to \( o^a \), denoted as \( a \not\triangleright o^a \). Intuitively, \( a \triangleright o^a \), if the set of objects accessible by a overlap with the set of objects represented by \( o^a \), i.e., \([c, \{a\}, \emptyset]\) and \([o^a]\) are not empty.

If \( a \in \mu \), the set of objects represented by \( o^a \) is entirely contained within the set of objects represented by \((c, \{a\}, \emptyset)\). In such cases, we say that \( a \) has exclusive access to \( o^a \), denoted \( a \triangleright o^a \).

**Example 4.2.** In Fig. 6, \( a_1 \triangleright (c, \{a_1\}, \{a_2, a_3\}) \), \( a_3 \not\triangleright (c, \{a_1\}, \{a_2, a_3\}) \). Finally, \( a_2 \triangleright (c, \{a_1\}, \emptyset) \) because the set of objects pointed to by \( a_2 \) may overlap with the set of objects pointed to by \( a_1 \).

The notion of accessibility allows us to determine when a strong or weak update can be performed on an abstract object \( o^a \). Consider an automaton operation \( op() \) at a node \( n \) encountered during the analysis. The abstract interpreter has to update the state of abstract objects accessible through a according to the state transitions of \( op() \). If \( a \triangleright o^a \) at \( n \),

\[
N_s = U_{p \in Proc[p]} S_p \text{ be the set of start nodes in the program, and } N_e = U_{p \in Proc[p]} E_p \text{ be the set of exit nodes in the program. Let } main \in Proc \text{ be the entry function. Finally, for each procedure } p \in Proc, \text{ let } Caller_p \subseteq N \text{ be the set of call nodes that call } p. \text{ Similarly, } N \in Proc[n] \text{ represents the procedure to which the node } n \text{ belongs.}
\]
then the analysis can perform a strong update, because \( o^2 \) only represents objects accessible through \( a \). On the other hand, if \( a \triangleright o^2 \) then a weak update has to be performed because it may represent objects not accessible through \( a \) as well. Furthermore, if \( a \not\triangleright o^2 \) the state of \( o^2 \) is unchanged.

To facilitate strong updates, an abstract object can be refined by using the focus operator, \( \text{focus}((c, \mu, \nu), a) = \left\{ \langle c, \mu \cup \{ a \}, \nu \rangle \mid a \in c \land a \not\in \nu \land \mu \right\} \) if \( a \in c \), \( \nu \not\in a \), and \( \mu \) otherwise.

The focus operation splits a given abstract object \( o^2 \) into \( o^2_1 \) containing objects definitely accessible by \( a \) and \( o^2_2 \) containing objects that are not accessible by \( a \). Therefore, an operation that is performed through the access path \( a \) should update the state of \( o^2_2 \) and leave the state of \( o^2_1 \) unchanged.

**Example 4.3.** Before applying an automaton operation \( a.\text{op}() \) on an abstract object \( o^2 \), the focus operator can be used to split \( o^2 \) into \( o^2_1 \) and \( o^2_2 \) such that \( a \triangleright o^2_1 \) and \( a \not\triangleright o^2_2 \). Therefore, we can perform a strong update on the abstract object \( o^2_1 \) and leave the state of \( o^2_2 \) unchanged.

For scalable verification, access paths in the must and must-not sets may be discarded, thus merging various abstract objects potentially leading to imprecise (but still sound) verification results. A blur operation discards access paths [18].

The abstract domain \( D = C \rightarrow 2^c \) is a map associating an abstract object with a set of automaton states. The \( \cup \) and \( \sqsubseteq \) on abstract objects and the set of automaton states can be extended pointwise in a straightforward manner to \( D \). \( (D, \sqsubseteq, \sqsubseteq) \) forms a lattice.

A flow-sensitive abstract map \( \eta^s : N \rightarrow D \) associates each CFG node \( n \in N \) with an abstract state map \( s^2 \in D \). Given an inductive map \( \eta^i \) with respect to a bug pattern \( \mathcal{P} \), the Mini language program is safe for the bug pattern \( \mathcal{P} \), if and only if \( \forall m \in N, a \in \mathcal{AP} : \text{error} \not\in \eta^i[m][m] \). To compute the fixpoint, we use the IFDS tabulation solver [37]. Except for the differences discussed in [3] the abstract transfer functions are similar to Fink et al. [18].

After the fixpoint is computed, if none of the nodes in the program have an error state, then the program does not have the given bug pattern. Otherwise, the program may have an execution matching the path.

## 5. LIFETIME DEPENDENCY ANALYSIS

Consider an access path \( a \in \mathcal{AP} \) and its alias cluster \( c \in \mathcal{C} \). Assume that an operation \( a.\text{op}() \) only affects objects accessible through \( a \) or its alias cluster \( c \). However, there are bug patterns where an operation \( a.\text{op}() \) also affects the typestates of objects in other alias clusters. Consider the iterator pattern shown in Fig. [1] that describes a pattern for iterator objects. Note that state \( \text{Init}(v) \) depends upon the container object \( v \) to which the iterator \( i \) is initialized. Thus, \( \text{Init}(v) \) in the automaton represents several concrete states, one for each possible container object. When the iterator is

```
1. i1 = v1.begin(); // [v1] ~ i1
2. i2 = i1;
3. i1 = v2.begin(); // [v1] ~ [i1, i2]
4. v1.push_back(); // [v1] ~ [i1, i2], [v2] ~ [i1]
```

**Figure 8:** Dependence information for an example.

- The dependency information can be incorporated into the typestate analysis described in [3] and [4]. In addition to the abstract-state maps, the analysis also maintains the set \( D \subseteq \Omega^s \times \Omega^t \) of dependency edges between objects at every node in the CFG. If \( (o^2_1, o^2_2) \in D \) then \( o^2_1 \sim o^2_2 \).

- Consider an automaton operation such as \( b = a.\text{op}() \) that makes the state of the abstract objects accessible through \( b \) to be dependent on the abstract objects accessible through \( a \). For such operations, the set of dependency edges \( D \) is updated as follows:

\[
D := D \cup \{ o^2_1 \sim o^2_2 \mid (a \triangleright o^2_1) \lor (b \triangleright o^2_2) \}
\]

First, the existing dependency information for objects for which either \( a \) or \( b \) has exclusive access is removed and the new dependency information is added between objects for which either \( a \) or \( b \) has exclusive access. By removing the existing dependency relations for abstract objects with exclusive access, the operation performs a strong update. For all other statements, whenever a new abstract object \( o^2 \) is added through a focus(\( a^2, a \)) or blur(\( a^2, a \)) operator the dependency information of \( o^2 \) is copied over to \( o^2 \).

**Example 5.1.** Consider line 3 in Fig. [8] where a new dependency edge between \( v2 \) and \( i1 \) is added. During the lifetime dependency analysis, the edge \( v1 \sim i1 \) is removed and the edge \( v2 \sim i1 \) is added.

**Uses of Lifetime Dependency Analysis.** We are interested in discovering the stale use of dependent objects. Thus, we consider lifetime dependencies: An object \( o_2 \) is lifetime dependent on another object \( o_1 \) if an operation on \( o_1 \) or its modification or destruction directly implies the destruction of object \( o_2 \). These types of lifetime dependencies
and especially interesting in the context of a language that relies on low-level memory management.

Example 5.2. Consider the stale buffer access shown in Fig. 3. We illustrate the lifetime dependency analysis in Fig. 3. On the left we show the source code simplifications, which introduce explicit calls to all implicit function calls and temporary object creation as needed. For example, the original statement `str = s.substr(i,n).c_str();` is desugared into three statements including the copy-construction of a temporary variable called `tmp` and finally its destruction after the call to `c_str()` is completed.

On the right we show the state of the two relevant objects and pointers of interest, namely the C++ string object `tmp` and the C string `str`. We assume that the C++ string `s` is valid throughout. Assume that for C++ strings, we have typestates `GoodStr` to represent a properly initialized string, and `UninitStr` to denote an uninitialized or deallocated string. We have similar states for C strings, which are denoted `GoodCStr` and `UninitCStr`.

We show the typestates for the two objects/access path clusters of interest here at the beginning of the sequence of statements. The interesting step is after the call to `c_str()`, where a lifetime dependency edge (LTDE) is added denoting the dependency of the abstract object (`GoodStr`, `str`) on the abstract object (`GoodCStr`, `tmp`). Then, the destruction of `tmp` in the following step not only destroys `tmp`, but also `str` via dependence analysis. Thus, we can easily identify the stale use of `str` in the following line.

Internal-buffer pattern refers to bugs, where a deallocated object’s internal buffer is accessed. The iterator pattern shown in Fig. 3 is an internal-buffer pattern, because an iterator is a pointer to an internal buffer of some container. When the container is destroyed, its internal buffer is deallocated and the iterator becomes stale.

6. IMPLEMENTATION

We have implemented a framework (shown in Fig. 11) that automatically generates Arc++ and uses it for typestate checking to find bugs in C++ programs. A given C++ program is initially translated by our frontend based on EDG [17] into a simplified version of C++ called CILPP [41]. CILPP is similar to CIL [50] with suitable additions for C++ constructs such as inheritance and exceptions. The exceptions are handled by the IECFG module [52] which generates an inter-procedural exceptional control flow graph that is similar to the CFG defined in Defn. 22 with additional edges related to exceptional flows. The new module Arc++ Gen. then automatically generates the Arc++ abstract representation, where each abstract instruction is associated with the control flow information of the IECFGs. With bug patterns provided by users, the framework automatically performs typestate checking on the Arc++ representations and reports detected bugs.

Specifying Bug Patterns. In our framework, the automata for bug patterns are specified in an ML-like language using instructions from Arc++. Fig. 10 shows the bug pattern for detecting invalid use of iterators. First the set of states along with initial and error states are defined. The state transitions for the automaton are provided by the transfer function. In transfer, Arc++ instructions are used to identify the C++ statement of interest and an appropriate action is performed. updateState is an API that describes the state transitions for the typestate analysis.

Given a program and a bug pattern, our framework performs typestate analysis on the program to check the bug pattern. The analysis reports any error transition to the user. Note that the iterator pattern does not specify the dependencies between iterators and containers. This is taken care of during the Arc++ generation, and the lifetime operations `link` and `invalidateLinked` (described in 2) are added to the appropriate C++ statements. For example, `link(it, v)` is generated for `it = v.begin()` and `invalidateLinked(v)` is generated for `v.push_back()`. The typestate analysis automatically takes these operations into account.

7. EVALUATION

To evaluate the effectiveness of the presented techniques, we have implemented some common bug patterns using our Arc++ framework and conducted evaluations on some micro-benchmarks (see [4]) and several open-source C++ projects. We seek to answer the following research questions:

- **RQ1**: Can the bug patterns based on typestates and syntactic checking detect real bugs?
- **RQ2**: Can the bug patterns based on our lifetime dependency analysis detect real bugs?
- **RQ3**: Can our analysis scale to open-source C++ programs of large sizes?
Subjects. We have evaluated the techniques on microbenchmarks and selected modules of some open-source benchmarks. The micro-benchmarks consist of 16 C++ programs that were handcrafted based on bugs that are encountered in real programs and good coding practice guidelines, such as Effective C++ [28] and Dark Side of C++ [40]. These small programs misuse STL libraries and can be used to test the detection capabilities of bug finding tools.

The open-source benchmarks consist of 20 open-source C++ projects. The lines of code (LOC) and the number of files of the projects are shown in Tab. 1. These projects are sampled from various domains, such as games (e.g., GNU Motti and GNUChess), parsers (e.g., TinyXML and AtomicParsley), simulators (e.g., Gama and Stella), and analysis tools (e.g., Cppcheck).

Evaluation Setup. For sake of brevity, we present 5 bug patterns that were checked using our Arc++ framework. Some of these bug patterns capture bad programming practices, and the other bug patterns capture invalid memory accesses, resource leaks and misuses of certain APIs:

- **Auto_Ptr Pattern.** This pattern identifies invalid usage of `auto_ptr` as described in [3].
- **Lock Pattern.** This pattern identifies invalid usage of pthread locks, such as double locking and potential deadlocks on exceptions.
- **Overloaded-Operator Pattern.** This pattern identifies bad programming practices of overloading the assignment operator `=`.
- **Copy-Constructor Pattern.** This pattern identifies the bad programming practice where a class contains fields that point to dynamically allocated memory, but does not contain a user-defined copy constructor.
- **Dependency Patterns.** Dependency patterns include the iterator pattern and the internal-buffer pattern of `c_str()` described in [5].

RQ1: Bug Pattern Analysis.

To address RQ1, we measure the number of bugs detected by the following bug patterns: `auto_ptr`, `pthread` locks, and overloading assignment operator (`=`). Tab. 2 shows the results of RQ1. Column “Pattern” shows the name of the bug pattern. Columns “Micro” and “OSS” show the number of bugs (excluding duplicates) and the number of false positives in brackets found in the benchmarks. We next show some example bugs detected by these patterns and discuss the reasons for the false positives.

**auto_ptr pattern.** Fig. 12 shows a bad practice detected by the `auto_ptr` pattern in the project RCSS Server. The class `Listener` accepts an `auto_ptr` `sender`, and calls the base class constructor by passing the `auto_ptr` `sender` by value. Due to the ownership transfer, any subsequent usage of `sender` in the class `Listener` is invalid. In this case, it is error-prone to extend class `Listener` if `sender` is used in any member function of `Listener`. Although we have not found any invalid usage of `sender` in the class `Listener` is invalid. In this case, it is error-prone to extend class `Listener` if `sender` is used in any member function of `Listener`.

Figure 12: A bad practice detected by the `auto_ptr` pattern in RCSS Server.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th># Files</th>
<th>LOC</th>
<th>IECFG</th>
<th>Ptr</th>
<th>Arc++</th>
<th>Lock</th>
<th>Iterator</th>
<th>c_str()</th>
<th>Aptr</th>
<th>Copy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro-Benchmarks</td>
<td>15</td>
<td>317</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
</tr>
<tr>
<td>TinyXML</td>
<td>30</td>
<td>7,511</td>
<td>0.95</td>
<td>0.35</td>
<td>0.24</td>
<td>0.17</td>
<td>0.66</td>
<td>0.17</td>
<td>0.18</td>
<td>0.017</td>
<td>0.005</td>
</tr>
<tr>
<td>motti-3.0.0</td>
<td>8</td>
<td>8,157</td>
<td>0.2</td>
<td>0.16</td>
<td>0.16</td>
<td>0.08</td>
<td>0.06</td>
<td>0.12</td>
<td>0.04</td>
<td>0.002</td>
<td>0.82</td>
</tr>
<tr>
<td>AtomicParsley-0.9.0</td>
<td>12</td>
<td>11,395</td>
<td>0.75</td>
<td>0.61</td>
<td>0.54</td>
<td>0.72</td>
<td>1.14</td>
<td>1.11</td>
<td>0.64</td>
<td>0.016</td>
<td>5.53</td>
</tr>
<tr>
<td>flac-1.2.1</td>
<td>5</td>
<td>25,678</td>
<td>0.47</td>
<td>0.25</td>
<td>0.25</td>
<td>0.28</td>
<td>0.63</td>
<td>0.23</td>
<td>0.26</td>
<td>0.004</td>
<td>4.35</td>
</tr>
<tr>
<td>TUIO_CPP-1.4</td>
<td>45</td>
<td>28,378</td>
<td>0.67</td>
<td>0.24</td>
<td>0.11</td>
<td>0.09</td>
<td>0.46</td>
<td>0.43</td>
<td>0.05</td>
<td>0.271</td>
<td>6.32</td>
</tr>
<tr>
<td>sipp-svn-3.2</td>
<td>73</td>
<td>37,080</td>
<td>6.81</td>
<td>6.77</td>
<td>4.73</td>
<td>8.13</td>
<td>31.29</td>
<td>29.32</td>
<td>18.52</td>
<td>7.37</td>
<td>2,104.40</td>
</tr>
<tr>
<td>gcc-0.2</td>
<td>157</td>
<td>46,013</td>
<td>2.02</td>
<td>0.89</td>
<td>0.69</td>
<td>0.48</td>
<td>1.06</td>
<td>1.25</td>
<td>0.38</td>
<td>0</td>
<td>6.77</td>
</tr>
<tr>
<td>faust-0.9.46</td>
<td>179</td>
<td>53,853</td>
<td>1.14</td>
<td>0.77</td>
<td>0.6</td>
<td>0.55</td>
<td>1.14</td>
<td>1.11</td>
<td>0.64</td>
<td>0.016</td>
<td>5.53</td>
</tr>
<tr>
<td>rcssserver-15.0.1</td>
<td>97</td>
<td>62,303</td>
<td>17.97</td>
<td>15.59</td>
<td>4.57</td>
<td>12.77</td>
<td>21.96</td>
<td>19.47</td>
<td>14.12</td>
<td>0.008</td>
<td>101.32</td>
</tr>
<tr>
<td>code-0.12</td>
<td>184</td>
<td>82,917</td>
<td>7.25</td>
<td>12.69</td>
<td>8.81</td>
<td>11.76</td>
<td>16.19</td>
<td>16.05</td>
<td>10.04</td>
<td>0.005</td>
<td>82.80</td>
</tr>
<tr>
<td>amos-3.1.0-rc1</td>
<td>239</td>
<td>85,891</td>
<td>39</td>
<td>33.42</td>
<td>14.28</td>
<td>31.17</td>
<td>196.35</td>
<td>96.84</td>
<td>28.34</td>
<td>0.093</td>
<td>439.49</td>
</tr>
<tr>
<td>cppcheck-1.5.1</td>
<td>118</td>
<td>88,314</td>
<td>6.67</td>
<td>4.48</td>
<td>2.92</td>
<td>7.03</td>
<td>43.38</td>
<td>8.36</td>
<td>6.22</td>
<td>0</td>
<td>45.35</td>
</tr>
<tr>
<td>p7zip-0.20.1</td>
<td>352</td>
<td>109,261</td>
<td>6.67</td>
<td>4.48</td>
<td>2.92</td>
<td>7.03</td>
<td>43.38</td>
<td>8.36</td>
<td>6.22</td>
<td>0</td>
<td>45.35</td>
</tr>
<tr>
<td>cmake-2.8.6</td>
<td>479</td>
<td>163,277</td>
<td>14.3</td>
<td>15</td>
<td>8.02</td>
<td>21.78</td>
<td>78.29</td>
<td>135.38</td>
<td>20.58</td>
<td>0.01</td>
<td>293.36</td>
</tr>
<tr>
<td>Greenstone</td>
<td>286</td>
<td>263,228</td>
<td>15.18</td>
<td>9.66</td>
<td>5.48</td>
<td>10.86</td>
<td>81.17</td>
<td>15.01</td>
<td>9.99</td>
<td>0.092</td>
<td>148.34</td>
</tr>
<tr>
<td>scummvm-1.2.1</td>
<td>1310</td>
<td>1,275,768</td>
<td>46.58</td>
<td>74.46</td>
<td>53.27</td>
<td>80.51</td>
<td>133.84</td>
<td>127.9</td>
<td>69.3</td>
<td>5.35</td>
<td>591.21</td>
</tr>
<tr>
<td>Total</td>
<td>3,829</td>
<td>2,402,404</td>
<td>178.90</td>
<td>195.70</td>
<td>118.20</td>
<td>206.80</td>
<td>649.40</td>
<td>562.91</td>
<td>185.20</td>
<td>7.37</td>
<td>2,104.40</td>
</tr>
</tbody>
</table>

Table 1: Evaluation subjects and analysis time of bug patterns (in seconds).

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Micro</th>
<th>OSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overload Op</td>
<td>3 (0)</td>
<td>5 (1)</td>
</tr>
<tr>
<td>Auto_ptr</td>
<td>2 (0)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Lock</td>
<td>3 (0)</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Copy Ctr</td>
<td>1 (0)</td>
<td>118 (21)</td>
</tr>
<tr>
<td>Iterator</td>
<td>3 (0)</td>
<td>15 (26)</td>
</tr>
<tr>
<td>c_str()</td>
<td>2 (0)</td>
<td>6 (0)</td>
</tr>
<tr>
<td>Total</td>
<td>15 (0)</td>
<td>147 (56)</td>
</tr>
</tbody>
</table>

Table 2: Number of real bugs found by patterns (false positives in brackets).
execution will go to the exception handler (not shown here) without unlocking the \texttt{objectList}. The exception handler also catches exceptions on paths with \texttt{objectList} not being locked, and thus does not unlock the object. This situation will likely lead to a deadlock after the exception is handled. Detecting these kinds of bugs requires the analysis of exceptional control flows, and few existing bug-finding tools \cite{21, 53, 3} have the capabilities to do that. Our lock pattern also detects double locking bugs for TUIO\_CPP, and has 3 false positives due to a lack of enough path-sensitivity.

**Overloaded-operator and Copy-Constructor patterns.** In the micro-benchmarks, the overloaded-operator pattern detects 2 bugs, where the operator does not return a reference to the receiver object, and another bug where the assignment operator of a base class is not invoked in the assignment operator of a derived class. The copy-structor pattern detects a bug where a field of a class is allocated with dynamic memory, but a copy constructor is missing. We use the typestate analysis to check whether a field is allocated with dynamic memory at the end of the constructor.

In the open-source projects, the overloaded-operator pattern detects 5 bugs with 1 false positive, and the copy-structor pattern detects 57 with 10 false positives. Fig.\ref{fig:overloaded_op_bug} shows the overloaded-operator bugs found by our framework in the project Gama. Struct \texttt{E\_3} does not overload the assignment operator and the default generated one returns a reference to the object itself. Thus, objects of \texttt{E\_3} support chained assignments, such as \texttt{x = y = z}. However, \texttt{E\_3} overloads the compound assignment operators (such as \texttt{+=} and \texttt{-=}), and returns \texttt{void}. This causes an inconsistency in using \texttt{E\_3}, and violates the rule that these overloaded operators should return a reference to the object itself.

The overloaded-operator pattern detects 5 bugs while cpcheck \cite{4} detects only 3 of them, because the CHROME object representation used in our framework effectively addresses the challenges posed by inheritance and method overloads. Both our analysis and cpcheck have 1 false positive in RCSS Server. This is caused by the presence of an internal struct \texttt{Holder} in class RCSSCLangLexer. \texttt{Holder} overloads operator \texttt{‘=’} for facilitating the lexical processing of different types of tokens, thus making it inappropriate for operator \texttt{‘=’} to return the reference to the \texttt{Holder} object.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{overloaded_op_bug.png}
\caption{Overloaded operator bug in Gama.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{stale_iterator.png}
\caption{A stale iterator use in Amos-3.1.0.}
\end{figure}

\textbf{RQ3: Scalability Analysis.}
To address RQ3, we measure the analysis time of each bug pattern and the execution time of the modules in our framework. Tab.\ref{tab:scalability} shows the results of RQ3. Columns “IECFG”, “Ptr”, and “Arc++” show the analysis time for constructing IECFGs, point-to-analysis, and generating Arc++ representations, respectively. Columns “Lock”, “Iterator”, “c\_str()”, “Aptr”, “Copy” show the analysis time for the corresponding bug patterns. Column “Total” shows the total time of preprocessing and pattern checking for each project. The analysis of the micro-benchmarks and the Overload-Operator pat-
tern based on syntactic checking is very fast. For most of the projects it takes less than 0.01 second to finish. Thus, we do not show the execution time for them. The copy-constructor pattern could not be checked for projects reactfIVision-1.4, p7zip 9.20.1, and stella-3.5 due to some failures in earlier modules unrelated to the typestate checker. For this reason, the execution time of this pattern for these projects is shown as 0. To reduce false positives, our framework also performs liveness analysis that helps remove invalid aliasing relationships. The execution time of the liveness analysis is similar to pointer analysis and is not shown here.

The results show that for the patterns without dependency analysis (Auto_ptr and Lock), the maximum analysis time is 80.51 s. For the dependency analysis patterns, the maximum analysis time is a bit higher (the Iterator pattern requires 196.35 s), but it is still reasonable. Moreover, the total time required for the pre-processing steps is just 374.60 s, and the total time for the pre-processing and typestate analysis in all the projects is 2104.40 s.

Answer: Our framework completes the analysis of open-source benchmarks of 2M LOCs in a reasonable time.

Threats to Validity.

Threats to external validity. We evaluated our program analysis techniques on a large variety of benchmarks. However, we cannot guarantee that the set of benchmark is representative of all domains. To mitigate this limitation, we strived to include a large range of benchmark applications, with varying application domains and of varying sizes. Moreover, we also evaluated our program analysis techniques on micro-benchmarks that are extracted from real bugs and other bug sources, such as Effective C++.

Threats to internal validity. One internal validity threat is the correctness of our implementation to preprocess and analyze C++. We rely on the correctness of our CILPP framework for C++, which has been thoroughly tested and is in production use for program analysis within NEC in a tool called Varvet. We have also inspected the newly developed code for correctness.

8. RELATED WORK

Our work is quite closely related to pattern-based bug finding tools for Java. FindBugs is a bug pattern detector for Java, which syntactically matches code to suspicious programming practice, in a manner similar to ASTLog. PMD and Jlint perform syntactic checks to analyze Java, and find problems of unused variables, unnecessary object creation, etc. In addition, Jlint detects synchronization problems using data flow analysis. While these tools focus on Java, our bug-finding framework handles more complex C++ semantics, and performs typestate analysis and dependency analysis over Arc++ representations to detect bugs. The state-machine framework of LLVM/Clang provides a possibility for bug pattern checking. However, based on our experience with open-source projects, pointer analysis, object abstraction, and the reasoning about lifetimes and dependencies are critical for effective typestate analysis.

There also exist bug finding approaches for C++. Rose is an open and extensible source-to-source compiler infrastructure, and Quinlan et al. propose approaches to search for bug patterns based on Rose’s interface to the abstract syntax tree (AST). Cppcheck is a static analysis tool for C++ code, which allows users to specify bug patterns using regular expressions. While these two tools detect bugs by searching for patterns in the AST or matching code using regular expressions, our framework further supports typestate-based bug patterns based on typestate analysis. PR-Miner mines implicit programming rules and detects violations for C code. PR-Miner cannot be applied directly on C++ code, since it does not handle the specific challenges brought by C++, such as exceptional control flows shown in Fig. 13. Coccinelle provides a semantic patch language for specifying desired matches and transformations in C code, which allows automatic bug detection based on the specified semantic patches. The semantic patches used by Coccinelle are quite different from our typestate analysis with dependency analysis. To perform a typestate-like analysis of multiple interacting objects, Naeem and Lhoták propose an abstract representation for objects, which is similar to our access-path clusters domain. While they only allow access paths that represent local variables, our domain allows arbitrary access paths, which is required for analyzing C++ programs.

Model checking and theorem proving techniques are also explored to detect bugs. Bandera is a Java verification tool based on model checking and abstraction. ESC/Java performs formal verification of properties of Java source code, and allows developers to add preconditions, post-conditions, and loop invariants in the form of special comments. Blanc et al. propose an operational model of the behavior guaranteed by the STL standard and apply predicate abstraction to a modified C++ program for verification. Blast provides an observer specification language for users to specify temporal safety properties of C programs, which allows specification of type states and checking of type states. Besides typestate analysis, our framework provides a dependency analysis to support bug patterns that require dependencies among several objects. Moreover, these tools target either Java or C, while we address C++. Finally, we also note recent work on theorem proving based approaches for C++ compiler certification in the CompCert project.

9. CONCLUSIONS

In this paper, we address the static analysis of complex C++ programs with an abstract representation and aliasing-aware typestate analysis techniques. We introduced a notion of lifetime dependency to target complex C++ lifetime semantics by proposing a lifetime dependency analysis. Finally, we show that an implementation of our techniques can find many interesting bugs in open-source projects. For the future, we will be extending our implementation on top of EDG and Cilpp to handle programs written using the new C++11 standard. In fact, newly introduced smart pointer types such as std::unique_ptr and other changes and library additions will lead to new bug patterns as well.
10. REFERENCES