Testing Noninterference, Quickly

Cătălin Hrițcu1 John Hughes2 Benjamin C. Pierce1 Antal Spector-Zabuzky1 Dimitrios Vytiniotis3
Arthur Azevedo de Amorim1 Leonidas Lampropoulos1

1University of Pennsylvania 2Chalmers University 3Microsoft Research

Abstract
Information-flow control mechanisms are difficult to design and labor intensive to prove correct. To reduce the time wasted on proof attempts doomed to fail due to broken definitions, we advocate modern random testing techniques for finding counterexamples during the design process. We show how to use QuickCheck, a property-based random-testing tool, to guide the design of a simple information-flow abstract machine. We find that both sophisticated strategies for generating well-distributed random programs and readily falsifiable formulations of noninterference properties are critically important. We propose several approaches and evaluate their effectiveness on a collection of injected bugs of varying subtlety. We also present an effective technique for shrinking counterexamples to minimal, easily comprehensible ones. Taken together, our best methods enable us to quickly and automatically generate simple counterexamples for all these bugs.

Categories and Subject Descriptors D.2.5 [Testing and Debugging]: Testing tools (e.g., data generators, coverage testing); D.4.6 [Security and Protection]: Information flow controls

General Terms Security, Languages, Design

Keywords random testing; security; design; dynamic information-flow control; noninterference; abstract machine; QuickCheck

1. Introduction
Secure information-flow control (IFC) is nearly impossible to achieve by careful design alone. The mechanisms involved are intricate and easy to get wrong: static type systems must impose numerous constraints that interact with other typing rules in subtle ways, while dynamic mechanisms must appropriately propagate taints and raise security exceptions when necessary. This intricacy makes it hard to be confident in the correctness of such mechanisms without detailed proofs; however, carrying out these proofs while labor intensive to prove correct. To reduce the time wasted on proof attempts doomed to fail due to broken definitions, we advocate modern random testing techniques for finding counterexamples during the design process. We show how to use QuickCheck, a property-based random-testing tool, to guide the design of a simple information-flow abstract machine. We find that both sophisticated strategies for generating well-distributed random programs and readily falsifiable formulations of noninterference properties are critically important. We propose several approaches and evaluate their effectiveness on a collection of injected bugs of varying subtlety. We also present an effective technique for shrinking counterexamples to minimal, easily comprehensible ones. Taken together, our best methods enable us to quickly and automatically generate simple counterexamples for all these bugs.

To answer this question, we take as a case study the task of extending a simple abstract stack-and-pointer machine to track dynamic information flow and enforce termination-insensitive noninterference [23]. Although our machine is simple, this exercise is both nontrivial and novel. While simpler notions of dynamic taint tracking are well studied for both high- and low-level languages, it has only recently been shown [1, 24] that dynamic checks are capable of soundly enforcing strong security properties. Moreover, sound dynamic IFC has been studied only in the context of lambda-calculi [1, 16, 25] and While programs [24]; the unstructured control flow of a low-level machine poses additional challenges. (Testing of static IFC mechanisms is left as future work.)

We show how QuickCheck [9], a popular property-based testing tool, can be used to formulate and test noninterference properties of our abstract machine, quickly find a variety of missing-taint and missing-exception bugs, and incrementally guide the design of a correct version of the machine. One significant challenge is that both the strategy for generating random programs and the precise formulation of the noninterference property have a dramatic impact on the time required to discover bugs; we benchmark several variations of each to identify the most effective choices. In particular, we observe that checking the unwinding conditions [14] of our noninterference property can be much more effective than directly testing the original property.

Our results should be of interest both to researchers in language-based security, who can now add random testing to their tools for debugging subtle enforcement mechanisms; and to the random-testing community, where our techniques for generating and shrinking random programs may be useful for checking other properties of abstract machines. Our primary contributions are: (1) a demonstration of the effectiveness of random testing for discovering counterexamples to noninterference in a low-level information-flow machine; (2) a range of program generation strategies for finding such counterexamples; (3) an empirical comparison of how effective combinations of these strategies and formulations of noninterference are in finding counterexamples; and (4) an effective methodology for shrinking large counterexamples to smaller, more readable ones. Our information-flow abstract machine, while simple, is also novel, and may be a useful artifact for further research.

2. Basic IFC
We begin by introducing the core of our abstract machine. In §5 we will extend this simple core with control flow (jumps and procedure calls), but the presence of pointers already raises opportunities for some subtle mistakes in information-flow control.

We write \(|\) for the empty list and \(x : s\) for the list whose first element is \(x\) and whose tail is \(s\); we also write \([x_0, x_1, \ldots, x_n]\) for the list \(x_0 : x_1 : \cdots : x_n : |\). If \(l\) is a list and \(0 \leq j < |l|\), then \(l(j)\) selects the \(j^{th}\) element of \(l\) and \(l(j \mapsto x)\) produces the list that is like \(l\) except that the \(j^{th}\) element is replaced by \(x\).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICFP ’13, September 25–27, 2013, Boston, MA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2326-0/13/09 . . . $15.00.
http://dx.doi.org/10.1145/2500365.2500574
**Bare machine** Our basic machine (without information-flow labels) has seven instructions:

\[
\text{Instr} ::= \text{Push } x \mid \text{Pop } \mid \text{Load } \mid \text{Store } \mid \text{Add } \mid \text{Noop } \mid \text{Halt }
\]

The \( x \) argument to \text{Push} is an integer (an immediate constant).

A machine state \( S \) is a 4-tuple consisting of a program counter \( pc \) (an integer), a stack \( s \) (a list of integers), a memory \( m \) (another list of integers), and an instruction memory \( i \) (a list of instructions), written \([pc\,s\,m\,i]\). Often, \( i \) will be fixed in some context and we will write just \([pc\,s\,m]\) for the varying parts of the machine state.

The single-step reduction relation on machine states, written \( S \Rightarrow S' \), is straightforward to define; we elide it here, for brevity. (It is included in a longer version of the paper, available from [http://www.crash-safe.org/node/24](http://www.crash-safe.org/node/24).) This relation is a partial function: it is deterministic, but some machine states don’t step to anything. Such a stuck machine state is said to be \text{halted} if \( \upsilon(pc) = \text{Halt} \) and failed in all other cases (e.g., if the machine is trying to execute an \text{Add} with an empty stack, or if the \( pc \) points outside the bounds of the instruction memory). We write \( \Rightarrow^* \) for the reflexive, transitive closure of \( \Rightarrow \). When \( S \Rightarrow^* S' \) and \( S' \) is a halted state, we write \( S \downarrow \downarrow S' \).

**Machine with labeled data** In a (fine-grained) dynamic IFC system [1, 16, 24, 25] security levels (called labels) are attached to runtime values and propagated during execution, enforcing the constraint that information derived from secret data does not leak to untrusted processes or to the public network. Each value is protected by an individual IFC label representing a security level (e.g., secret or public). We now add labeled data to our simple stack machine.

To illustrate the design methodology we hope to support, we first introduce in the initial state; because of the form of our initial states, such values can appear only in instruction memories. By secret inputs we mean values labeled \( \top \) in the initial state; secret outputs labeled \( \bot \) in the state.

**Noninterference (EENI)** We define what it means for this basic IFC machine to be “secure” using a standard notion of termination-insensitive noninterference [1, 16, 23]; we call it \textit{end-to-end noninterfering} (or EENI) to distinguish it from the stronger notions we will introduce in §6. The main idea of EENI is to directly encode the intuition that secret inputs should not influence public outputs. By secret inputs we mean values labeled \( \top \) in the initial state; because of the form of our initial states, such values can appear only in instruction memories. By secret outputs we mean values labeled \( \bot \) in a halted state. More precisely, EENI states that for any two executions starting from initial states that are indistinguishable to a low observer (or just indistinguishable) and ending in halted states \( H_1 \) and \( H_2 \), the final states \( H_1 \) and \( H_2 \) are also indistinguishable. Intuitively, two states are indistinguishable if they differ only in values labeled \( \top \). To make this formal, we define an equivalence relation on states compositionally from equivalence relations over their components.

**2.1 Definition:** Two values \( x_1 \) and \( x_2 \) are said to be indistinguishable, written \( x_1 \equiv x_2 \), if either \( x_1 = x_2 \) or \( x_1 = \top \) or else \( x_1 = x_2 \) and \( L_1 = L_2 = \bot \). Two instructions \( i_1 \) and \( i_2 \) are indistinguishable if they are the same instruction, or if \( i_1 = \text{Push } v_1 \) and \( i_2 = \text{Push } v_2 \), and \( v_1 \equiv v_2 \). Two lists (memories, stacks, or instruction memories) \( l_1 \) and \( l_2 \) are indistinguishable if they have the same length and \( l_1(x) \equiv l_2(x) \) for all \( x \) such that \( 0 \leq x < |l_1| \).

For machine states we have a choice as to how much of the state we want to consider observable; we choose (somewhat arbitrarily) that the observer can only see the data and instruction memories, but not the stack or the \( pc \). (Other choices would give the observer either somewhat more power—e.g., we could make the stack and \( pc \) observable—or somewhat less—e.g., we could restrict the observer to some designated region of “I/O memory,” or extend the architecture with I/O instructions and only observe the traces of inputs and outputs.)

**2.2 Definition:** Machine states \( S_1 = [pc_1\,s_1\,m_1\,i_1] \) and \( S_2 = [pc_2\,s_2\,m_2\,i_2] \) are indistinguishable with respect to memories, written \( S_1 \approx_{\text{mem}} S_2 \), if \( m_1 \equiv m_2 \) and \( i_1 \equiv i_2 \).

**2.3 Definition:** A machine semantics is \textit{end-to-end noninterfering} with respect to some sets of states \text{Start} and \text{End} and an indistinguishability relation \( \approx \), written \( \text{EENI}_{\text{Start}, \text{End}, \approx} \), if for any \( S_1, S_2 \in \text{Start} \) and \( H_1, H_2 \in \text{End} \) such that \( S_1 \approx S_2 \) and such that \( S_1 \Downarrow H_1 \) and \( S_2 \Downarrow H_2 \), we have \( H_1 \approx H_2 \).

We take \( \text{EENI}_{\text{Init}, \text{Halted}, \approx_{\text{mem}}} \) as our baseline security property; i.e., we often consider executions starting in initial states and ending in halted states, and we use indistinguishability with respect to memories. The EENI definition above is, however, more general, and we will consider other instantiations of it later.

**Information-flow rules** Our next task is to enrich the rules for the step function to take information-flow labels into account. For most of the rules, there are multiple plausible ways to do this, and some opportunities for subtle mistakes even with these few instructions. To illustrate the design methodology we hope to support, we first propose a naive set of rules and then use QuickCheck-generated counterexamples to identify and help repair mistakes until no more can be found.

\[
\begin{align*}
\text{NOOP} & \\
\{pc\,s\,m\} & \Rightarrow \{pc+1\,s\,m\} \\
\text{PUSH} & \\
\{pc\,s\,m\} & \Rightarrow \{pc+1\,v:s\,m\} \\
\text{POP} & \\
\{pc\,v:s\,m\} & \Rightarrow \{pc+1\,s\,m\} \\
\text{LOAD} & \\
\{pc\,x@L\,s\,m\} & \Rightarrow \{pc+1\,m(x):s\,m\} \\
\text{STORE} & \\
\{pc\,x@L\,v:s\,m\} & \Rightarrow \{pc+1\,s,m\{x \mapsto v\}\} \\
\text{ADD} & \\
\{pc\,x@L\,y@L\,s\,m\} & \Rightarrow \{pc+1\,(x+y):@\,s\,m\}
\end{align*}
\]

The NOOP rule is the same as in the unlabeled machine. In the PUSH and POP rules, we simply change the relevant integers to be labeled values; luckily, this obvious adaptation happens to be correct. But now our luck runs out: the simple changes that we’ve
made in the other rules will all turn out to be wrong. (We include a star in the names of incorrect rules to indicate this. The rule \texttt{STORE*AB} actually contains two bugs, which we refer to as \texttt{A} and \texttt{B}; we will discuss them separately later.) Fortunately, QuickCheck can rapidly pinpoint the problems, as we will see.

Figure 1 shows the first counterexample that QuickCheck gives us when we present it with the step function defined by the six rules above and ask it to invalidate the EENI property. (The \texttt{EFGX} source for all the figures was generated automatically by our QuickCheck testing infrastructure.) The first line of the figure is the counterexample itself: a pair of four-instruction programs, differing only in the constant argument of the second \texttt{Push}. The first program pushes \texttt{0@T}, while the second pushes \texttt{1@T}, and these two values are indistinguishable. We display the two programs, and the other parts of the two machine states, in a “merged” format. Pieces of data that are the same between the two machines are written just once; at any place where the two machines differ, the value of the first machine is written above the value of the second machine, separated by a horizontal line. The rest of the figure shows what happens when we run this program. On the first step, the \texttt{pc} starts out at 0; the memory, which has two locations, starts out as \texttt{[0@\bot,0@\bot]}; the stack starts out empty; and the next instruction to be executed (\texttt{i(pc)}) is \texttt{Push 1@\bot}. On the next step, this value has been pushed on the stack and the next instruction is either \texttt{Push 0@T} or \texttt{Push 1@T}; one or the other of these values is pushed on the stack. On the next, we \texttt{Store} the second stack element (\texttt{1@\bot}) into the location pointed to by the first (either \texttt{0@T} or \texttt{1@T}), so that now the memory contains \texttt{1@\bot} in either location 0 or location 1 (the other location remains unchanged, and contains \texttt{0@\bot}). At this point, both machines halt. This pair of execution sequences shows that EENI fails: in the initial state, the two programs are indistinguishable to a low observer (their only difference is labeled \texttt{T}), but in the final states the memories contain different values at the same location, both of which are labeled \texttt{\bot}.

Thinking about this counterexample, it soon becomes apparent what went wrong with the \texttt{Store} instruction: since pointers labeled \texttt{T} are allowed to vary between the two runs, it is not safe to store a low value through a high pointer. One simple but draconian fix is simply to stop the machine if it tries to perform such a store (i.e., we could add the side-condition \texttt{L_s = \bot} to the rule). A more permissive option is to allow the store to take place, but require it to taint the stored value with the label on the pointer:

\[
i(pc) = \text{Store} \quad \frac{pc \ x@L_s : y@L_y : s \ m \Rightarrow pc+1 \ s \ \{x \mapsto y@(L_x \cup L_y)\}}{(\text{STORE*B})}
\]

Unfortunately, QuickCheck’s next counterexample (Figure 2) shows that this rule is still not quite good enough. This counterexample is quite similar to the first one, but it illustrates a more subtle point: our definition of noninterference allows the observer to distinguish between final memory states that differ only in their labels.\footnote{See the first clause of Definition 2.1. One might imagine that this could be fixed easily by changing the definition so that labels are not observable—i.e., \texttt{x@\bot} \approx \texttt{x@T} for any \texttt{x}. Sadly, this is known not to work \cite{22}. (QuickCheck can also find a counterexample; see \S A).}

Since the \texttt{STORE*B} rule taints the label of the stored value with the label of the pointer, the fact that the \texttt{Store} changes different locations is visible in the fact that a label changes from \texttt{\bot} to \texttt{T} on a different memory location in each run. To avoid this issue, we adopt the “no sensitive upgrades” rule \cite{1,28}, which demands that the label on the current contents of a memory location being stored into are above the label of the pointer used for the store—we get, i.e., it is illegal to overwrite a low value via a high pointer (and trying to do so terminates the machine). Adding this side condition brings us to a correct version of the \texttt{STORE} rule:

\[
i(pc) = \text{Store} \quad \frac{pc \ x@L_x : y@L_y : s \ m \Rightarrow pc+1 \ s \ \{x \mapsto y@(L_x \cup L_y)\}}{(\text{STORE})}
\]

The next counterexample found by QuickCheck (Figure 3) points out a straightforward problem in the \texttt{ADD} rule: adding \texttt{0@\bot} to \texttt{0@T} yields \texttt{0@\bot}. (We elide the detailed execution trace for this example and most of the ones that we will see later, for brevity. They can be found in the long version.) The problem is that the taints on the arguments to \texttt{Add} are not propagated to its result. The \texttt{Store} is needed in order to make the difference observable. The easy (and standard) fix is to use the join of the argument labels as the label of the result:

\[
i(pc) = \text{Add} \quad \frac{pc \ x@L_x : y@L_y : s \ m \Rightarrow pc+1 \ s \ \{x \mapsto y@(L_x \cup L_y)\}}{(\text{ADD})}
\]

The final counterexample (Figure 4) alerts us to the fact that the \texttt{LOAD} rule contains a similar mistake to the original \texttt{STORE*AB} rule: loading a low value through a high pointer should taint the loaded value. The program in Figure 4 is a little longer than the one in Figure 1 because it needs to do a little work at the beginning to set up a memory state containing two different low values. It then pushes a high address pointing to one or the other of those
cells onto the stack; loads (different, low addresses) through that
pointer; and finally stores \texttt{0}@\perp to the resulting address in memory
and halts. In this case, we can make the same change to \texttt{LOAD*}
as we did to \texttt{STORE*AB}: we taint the loaded value with the join
of its label and the address’s label. This time (unlike the case of
\texttt{Store}, where the fact that we were changing the memory gave us
additional opportunities for bugs), this change gives us the correct
rule for \texttt{Load},
\[\texttt{i(pc)} = \texttt{Load} \quad \frac{x \at \texttt{L}_x : s \cdot m}{\texttt{pc+1} \at m(x) \at \texttt{L}_x : s \cdot m} \quad \text{(LOAD)}\]
and QuickCheck is unable to find any further counterexamples.

**More bugs** The original IFC version of the step rules illustrate
one set of mistakes that we might plausibly have made, but there
are more possible mistakes:
\[\texttt{i(pc)} = \texttt{Push} \at x \at \texttt{L} \quad \frac{\texttt{pc}s@m}{\texttt{pc+1}@x@\perp s@m} \quad \text{(PUSH*)}\]
\[\texttt{i(pc)} = \texttt{Store} \quad \frac{x \at \texttt{L}_x : y \at \texttt{L}_y : s \cdot m}{\texttt{pc+1}@m(x \mapsto y@\perp)} \quad \text{(STORE*C)}\]

Although it is unlikely that we’d write these rather silly rules by
accident, it is worth including them in our experiments because
they can be invalidated by short counterexamples and thus provide
useful data points for less effective testing strategies.

We will also gather statistics for a partially fixed but still wrong
rule for \texttt{Store}, in which the no-sensitive-upgrades check is performed
but the result is not properly tainted:
\[\texttt{i(pc)} = \texttt{Store} \quad \frac{x \at \texttt{L}_x : v : s \cdot m}{\texttt{pc+1}@m(x \mapsto v)} \quad \text{(STORE*A)}\]

3. QuickCheck

We test noninterference using QuickCheck [9], a tool that tests
properties expressed in Haskell. Often, QuickCheck is used to test
properties that should hold for all inhabitants of a certain type.
QuickCheck repeatedly generates random values of the desired
type, instantiates the property with them, and checks it directly
by evaluating it to a boolean. This process continues until either a
counterexample is found or a specified timeout is reached. Quick-
Check supplies default test data generators for many standard types.
Additionally, the user can supply custom generators for their own
types. In order to test EENI, for example, we needed to define cus-
tom generators for values, instructions, and machine states (each of
which depends on the previous generator: machine states contain
instructions, some of which contain values). The effectiveness of
testing (i.e., mean time to discover bugs) depends on the sophisti-
cation of these generators, a topic we explore in detail in §4.

QuickCheck properties may also be guarded by \texttt{preconditions};
EENI is an example of why this is necessary, as it only applies to
pairs of indistinguishable initial machine states that both success-
fully execute to halted states. Testing a property with a precondi-
tion proceeds similarly: a sequence of random values are generated
and tested, up to a user-specified maximum. The difference is that
if there is a precondition, it is instantiated with the random value
first. If the precondition does not hold, this random value is sum-
mmarily discarded. If the precondition does hold, then the rest of
the property is checked just as before. Although preconditions are very
useful, too high a proportion of discards can lead to slow testing or
a badly skewed distribution of test cases (since some kinds of test
case may be discarded much more often than others). To help di-
agnose such problems, QuickCheck can collect statistics about the
tests it tried.

When a test fails, the failing test case is often large, containing
many irrelevant details. QuickCheck then tries to \texttt{shrink} the test
case, by searching for a similar but smaller test case that also fails.
To do this, it greedily explores a number of “shrinking candidates”;
modifications of the original failing test case that are “smaller” in
some sense. The property is tested for each of these, and as soon
as a failure is found, that candidate becomes the starting point for
a new shrinking search (and the other candidates are discarded).
Eventually this process terminates in a failing test case that is
\textit{locally minimal}: none of its shrinking candidates fails. This failing
case is then reported to the user. It is often very much smaller than
the original randomly generated test case, and it is thus easy to use it
to diagnose the failure because it (hopefully) contains no irrelevant
details. Just like generation strategies, shrinking strategies are type
dependent; they are defined by QuickCheck for standard types,
and by the user for other types. We discuss the custom shrinking
strategies we use for machine states in §7.

4. Test Generation Strategies

We are ready now to begin exploring ways to generate potential
counterexamples. At the outset, we need to address one fundament-
ally issue. Noninterference properties quantify over a \textit{pair} of indis-
nuishable starting states: \(\forall S_1, S_2 \in \text{Start}. S_1 \approx S_2 \implies \ldots\)
This is a very strong precondition, which is unlikely to be satisfied
for independently generated states. Instead, we generate \textit{indistin-
guishable} pairs of states together. The first state is generated ran-
domly using one of the techniques described later in this section.
The second is obtained by randomly varying the “high parts” of
the first. We refer to the second state as the \textit{variation} of the first. The
resulting pair thus satisfies indistinguishability by construction. Note
that we have not compromised completeness: by generating a ran-
dom state and randomly varying we still guarantee that it is pos-
sible to generate all pairs of indistinguishable states. Naturally,
the resulting distributions will depend on the specifics of the genera-
tion and variation methods used, as we shall see.

Since EENI considers only executions that start at initial states,
we only need to randomly generate the contents of the instruction
memory (the program that the machine executes) together with the
\textit{size} of the data memory (in initial states, the contents of the
memory are fixed and the stack is guaranteed to be empty).

Figure 5 offers an empirical comparison of all the generation
strategies described in this section. For a given test-generation
strategy, we inject bugs one at a time into the machine definition
and measure the time spent on average until that bug is found (\textit{mean
time to failure}, or \textit{MTTF}). Tests were run one at a time on seven
identical machines, each with 4 × 2.4 GHz Intel processors and
11.7 GB of RAM; they were running Fedora 16 and GHC 7.4.2,
and using QuickCheck 2.5.1.1. We run each test for 5 minutes
(300 seconds) or until 4000 counterexamples are found, whichever
comes first.

\textbf{Naive instruction generation} The simplest way to generate
programs is by choosing a sequence of instructions \textit{independently} and
\textit{uniformly}. We generate individual instructions by selecting an
instruction type uniformly (i.e., \texttt{Noop, Push, etc.}) and then filling in
its fields using QuickCheck’s built-in generators. Labels are also
chosen uniformly. We then build the instruction memory by sam-
pling a number (currently a random number between 20 and 50) of
instructions from this generator.

The first column of Figure 5 shows how this strategy performs
on the bugs from §2. Disappointingly, but perhaps not too surpris-
One obvious weakness is that the discard rate is quite high, indicating that one or both machines often fail to reach a halted state. By asking QuickCheck to collect statistics on the execution traces of test cases (Figure 6), we can also see a second problem: the average execution length is only 0.46 steps! Such short runs are not useful for finding counterexamples to EENI (at a minimum, any counterexample must include a valid address), and similarly for other instructions that use stack elements. The results are shown in the third column of Figure 5. With sequence generation we can now find all bugs, faster than before. Programs run for slightly longer (3.87). As expected, stack underflows are less common than before (Figure 8) and out-of-range addresses are now the second biggest reason for termination.

Generating useful instruction sequences more often  To further reduce stack underflows we can generate sequences of instructions that make sense together. For instance, instead of generating single Store instructions, we can additionally generate sequences of the form [Push a, Store] (where a is a valid address), and similarly for other instructions that use stack elements. The results are shown in the second column of Figure 5. Although this strategy still performs badly for the Load and Store instructions, there is a significant improvement on both discard rates and the MTTF. Run length is also better, averaging 2.71 steps. As Figure 7 shows, executing Halt is now the main reason for termination, with stack underflows and out-of-range accesses close behind.

Generation by execution  We can go even further. In addition to weighted distributions, sequences, and smart integers, we try to generate instructions that do not cause a crash. In general (for more interesting machines) deciding whether an arbitrary instruction sequence causes a crash is undecidable. In particular we cannot know in advance all possible states in which an instruction will be executed. We can only make a guess—a very accurate one for this simple machine. This leads us to the following generation by execu-
Average execution steps: 11.60 (generated) / 11.26 (variation). 95% halt / halt, 3% halt / load or store out of range, and 1% halt / sensitive upgrade.

Figure 10. Execution statistics for generation by execution, broken down for the variations.

\[ i = \begin{cases} \text{Push} \frac{2}{5} \oplus \top, \text{Jump}, \text{Push} 1 \oplus, \text{Push} 0 \oplus, \text{Store}, \text{Halt} \end{cases} \]

Figure 11. Counterexample to Jump^AB: A textbook example of an implicit flow

\[ \text{tion} \text{ strategy: We generate a single instruction or a small sequence of instructions, as before, except that now we restrict generation to instructions that do not cause the machine to crash in the current state. When we find one, we execute it to reach a new state and then repeat the process to generate further instructions. We continue until we have generated a reasonably sized instruction stream (currently, randomly chosen between 20–50 instructions). We discuss how this idea generalizes to machines with nontrivial control flow in \S 5.} \]

As we generate more instructions, we make sure to increase the probability of generating a Halt instruction, to reduce the chances of the machine running off the end of the instruction stream. As a result, (i) we maintain low discard ratios for EENI since we increase the probability that executions finish with a Halt instruction, and (ii) we avoid extremely long executions whose long time to generate and run could be more fruitful used for other test cases.

The MTTF (last column of Figure 5) is now significantly lower than in any previous generation method, although this strategy runs fewer tests per second than the previous ones (because both test case generation and execution take longer). Figure 10 shows that 94% of the pairs both successfully halt, which is in line with the very low discard rate of Figure 5, and that programs run for much longer. Happily, varying a machine that successfully halts has a high probability of generating a machine that also halts.

5. Control Flow

Up to this point, we’ve seen how varying the program generation strategy can make orders-of-magnitude difference in the speed at which counterexamples are found for a very simple—almost trivial—information-flow machine. Now we are ready to make the machine more interesting and see how these techniques perform on the new bugs we’ve already seen. In this section, we add Jump, Call, and Return instructions—and, with them, the possibility that information can leak via the program’s control flow.

Jumps, implicit flows, and the \textit{pc} label

We first add a new Jump instruction that takes the first element from the stack and sets the \textit{pc} to that address:

\[ i(\text{pc}) = \text{Jump} \quad \text{(Jump^{AB})} \]

(The jump target may be an invalid address. In this case, the machine will be stuck on the next instruction.)

Note that this rule simply ignores the label on the jump target on the stack. This is unsound, and QuickCheck easily finds the counterexample in Figure 11—a textbook case of an implicit flow [23]. A secret is used as the target of a jump, which causes the instructions that are executed afterwards to differ between the two machines; one of the machines halts immediately, whereas the other

\[ i = \begin{cases} \text{Push} 1 \oplus, \text{Push} \frac{4}{6} \oplus \top, \text{Jump}, \text{Halt}, \text{Push} 0 \oplus, \text{Store}, \text{Push} 3 \oplus, \text{Jump} \end{cases} \]

Figure 12. Counterexample to Jump^B: Jump should not lower the \textit{pc} label

one does a Store to a low location and only then halts, causing the final memories to be distinguishable. The standard way to prevent implicit flows is to label the \textit{pc}—i.e., to make it a value, not a bare integer. Initial states have \textit{pc} = 0 \oplus, and after a jump to a secret address the label of the \textit{pc} becomes \top.

The JUMP rule is still wrong, however, since it not only raises the \textit{pc} label when jumping to a high address but also lowers it when jumping to a low address. The counterexample in Figure 12 illustrates that the latter behavior is problematic. The fix is to label the \textit{pc} after a jump with the \textit{join} of the current \textit{pc} label and the label of the target address.

\[ i(\text{pc}) = \text{Jump} \quad \text{(JUMP)} \]

With this rule for jumps QuickCheck no longer finds any counterexamples. Some readers may find this odd: In order to fully address implicit flows, it is usually necessary to modify the rules for memory stores to handle the case where the \textit{pc} is labeled high [1, 22]. The current machine doesn’t require this, but the reason is subtle: here, the \textit{pc} can go from \top to \bot when we jump to a secret address, but it never goes from \top to \bot! It doesn’t matter what the machine does when the \textit{pc} is high, because none of its actions will ever be observable—all high machine states are indistinguishable.
To make things more interesting, we need to enrich the machine with some mechanism that allows the \( pc \) to safely return to \( \perp \) after it has become \( \top \). One way to achieve this is to add \textit{Call} and \textit{Return} instructions, a task we turn to next.

### Restoring the \( pc \) label with calls and returns

IFC systems (both static and dynamic) generally rely on control flow merge points (i.e., post-dominators of the branch point in the control flow graph where the control was tainted by a secret) to detect when the influence of secrets on control flow is no longer relevant and the \( pc \) label can safely be restored. Control flow merge points are, however, much more evident for structured control features such as conditionals than they are for jumps. Moreover, since we are doing purely dynamic IFC we cannot distinguish between safe uses of jumps and unsafe ones (e.g., the one in Figure 12). So we keep jumps as they are (only raising the \( pc \) label) and add support for structured programming and restoring the \( pc \) label in the form of \textit{Call} and \textit{Return} instructions, which are of course useful in their own right.

To support these instructions, we need some way of representing stack frames. We choose a straightforward representation, in which each stack element can now be either a value (as before) or a return address, marked \( R \), recording the \( pc \) (including its label) from which the corresponding \textit{Call} was made. We also extend the indistinguishability relation on stack elements so that return addresses are only equivalent to other return addresses and \( R(x_1@L_1) \approx R(x_2@L_2) \) if either \( L_1 = L_2 = \top \) or else \( x_1 = x_2 \) and \( L_1 = L_2 = \perp \) (this is the same as for values).

We also need a way to pass arguments to and return results from a called procedure. For this, we annotate the \textit{Call} and \textit{Return} instructions with an integer indicating how many stack values should be passed or returned (0 or 1 in the case of \textit{Return}). Formally, \( Call \) \( n \) expects an address \( x@L \) followed by \( n \) values on the stack. It sets the \( pc \) to \( x \), labels this new \( pc \) by the join of \( L \), and the current \( pc \) label (as we did for \textit{Jump}—we’re eliding the step of getting this bit wrong at first and letting QuickCheck find a counterexample), and adds the return address frame to the stack under the \( n \) arguments.

\[
\begin{align*}
i(pc) &= \text{Call } n \quad L \equiv L_x \lor L_{pc} \\
&\quad \left\{ \begin{array}{l}
x_{pc}@L_{pc} \quad x@L : v_1 : \ldots : v_n : s \quad m \\
x@L : v_1 : \ldots : v_n @ R(x_{pc}+1@L_{pc}) : s \quad m
\end{array} \right. \quad (\text{CALL*B})
\end{align*}
\]

\textit{Return} \( n' \) traverses the stack until it finds the first return address and jumps to it. Moreover it restores the \( pc \) label to the label stored in that \( R \) entry, and preserves the first \( n' \) elements on the stack as return values, discarding all other elements in this stack frame.

\[
\begin{align*}
i(pc) &= \text{Return } n' \quad n' \in \{0, 1\} \quad k \geq n' \\
&\quad \left\{ \begin{array}{l}
pc v_1 : \ldots : v_n : R(x@L_x) : s \quad m \\
x@L_x : v_1 : \ldots : v_{n'} @ s \quad m
\end{array} \right. \quad (\text{RETURN*AB})
\end{align*}
\]

Finally, we observe that we cannot expect the current EENI instantiation to hold for this changed machine, since now one machine can halt in a high state while the other can continue, return to a low state, and only then halt. Since we cannot equate high and low states (see §A), we need to change the EENI instance we use to EENI\textit{Init.Halted}↓\textit{Low}, \( \text{Init.Halted}↓\textit{Low} \), where \( \text{Low} \) denotes the set of states with \( pc = \perp \). Thus, we only consider executions that end in a low halting state.

After these changes, we can turn QuickCheck loose and start finding more bugs. The first one, listed in Figure 13, is essentially another instance of the implicit flow bug, which is not surprising given the discussion at the end of the previous subsection. We need to change the rule for \textit{Store} so that the new memory contents are tainted with the current \( pc \) label. This eliminates the current counterexample; QuickCheck then finds a very similar one in which the labels of values in the memories differ between the two machines. The usual way to prevent this problem is to extend the no-sensitive-upgrades check so that low-labeled data cannot be overwritten in a high context \([1, 28]\). This leads to the correct rule for stores:

\[
\begin{align*}
i(pc) &= \text{Store } L_{pc} \lor L_x \subseteq L_m(x) \\
&\quad \left\{ \begin{array}{l}
\text{pc } v@L_x : y@L_y : s \quad m \Rightarrow \\
p@1 + s \quad m\{ x := v @ (L_x \lor L_y \lor L_{pc}) \}
\end{array} \right. \quad (\text{STORE})
\end{align*}
\]

The next counterexample found by QuickCheck (Figure 14) shows that returning values from a high context to a low one is unsound if we do not label those values as secrets. To fix this, we taint all the returned values with the pre-return \( pc \) label.

\[
\begin{align*}
i(pc) &= \text{Return } n' \quad n' \in \{0, 1\} \quad k \geq n' \\
&\quad \left\{ \begin{array}{l}
\text{pc } v@L_x : y@L_y : s \quad m \Rightarrow \\
x@L_x v_1 @ L_{pc} : \ldots : v_{n'} @ L_{pc} : s \quad m
\end{array} \right. \quad (\text{RETURN*B})
\end{align*}
\]

The next counterexample, listed in Figure 15, shows (maybe somewhat surprisingly) that it is unsound to specify the number of results to return in the \textit{Return} instruction, because then the number of results returned may depend on secret flows of control. To restore soundness, we need to pre-declare at each \textit{Call} whether the corresponding \textit{Return} will return a value—i.e., the \textit{Call} instruction should be annotated with two integers, one for parameters and the other for results; accordingly, each stack frame should include not only a return address but also a number of return values. These changes lead us to the correct rules:

\[
\begin{align*}
i(pc) &= \text{Call } n \quad n' \in \{0, 1\} \quad L \equiv L_x \lor L_{pc} \\
&\quad \left\{ \begin{array}{l}
x_{pc}@L_{pc} x@L_x : v_1 : \ldots : v_n : s \quad m \\
x@L : v_1 : \ldots : v_n @ R(x_{pc}+1@L_{pc}) : s \quad m
\end{array} \right. \quad (\text{CALL})
\end{align*}
\]

\[
\begin{align*}
i(pc) &= \text{Return } n' \quad n' \in \{0, 1\} \\
&\quad \left\{ \begin{array}{l}
\text{pc } v_1 : \ldots : v_{n'} : R(x, n')@L_x : s \quad m \Rightarrow \\
x@L_x v_1 @ L_{pc} : \ldots : v_{n'} @ L_{pc} : s \quad m
\end{array} \right. \quad (\text{RETURN})
\end{align*}
\]

Figure 13. Counterexample to old \textit{STORE} rule in the presence of \textit{Call} and \textit{Return}: Raising \( pc \) label is not enough to prevent implicit flows. Once we have a mechanism (like \textit{Return}) for restoring the \( pc \) label, we need to be more careful about stores in high contexts.

Figure 14. Counterexample to \textit{RETURN*AB}: \textit{Return} needs to taint the returned values.

Figure 15. Counterexample to \textit{CALL*B} and \textit{RETURN*B}: It is unsound to choose how many results to return on \textit{Return}. 
i = [\textit{Push }5@\perp, \textit{Call }0 \, 1, \textit{Push }0@\perp, \textit{Store, Halt}, \textit{Push }0@\perp, \textit{Push }\frac{8}{9} @\top, \textit{Call }0 \, 0, \textit{Pop, Push }0@\perp, \textit{Return}]

\textbf{Figure 16.} Counterexample to \textit{Pop}\textsuperscript{*}: It is unsound not to protect the call stack.

The final counterexample found by QuickCheck is quite a bit longer (see Figure 16). It shows that we cannot allow instructions like \textit{Pop} to remove addresses from the stack, as does the following broken rule (we use \(v \rightarrow e\) to denote an arbitrary stack entry):

\[
i (pc) = \text{Pop} \quad \text{(Pop\textsuperscript{*})}
\]

To protect the call frames on the stack, we change this rule to only pop values (all the other rules can already only operate on values).

\[
i (pc) = \text{Pop} \quad \text{(Pop)}
\]

\textbf{Generation by execution and control flow} Generation by execution is still applicable in the presence of interesting control flow but we have to make small modifications to the original algorithm. We still generate a single instruction or sequence that does not crash, as before, and we execute to compute a new state. However, unlike before, while executing this newly generated sequence of instructions, it is possible to “land” in a position in the instruction stream where we have already generated an instruction. (e.g. via a backward jump): If this happens then we keep executing the already generated instructions. If the machine stops (or we reach a loop-avoiding cutoff) then we stop the process and return the so-far generated instruction stream. If there are no more instructions to execute then we go on to generate more instructions. There is one more possibility though: the machine may \textit{crashes} while executing an already generated instruction. To address this issue, we make sure that we never generate an instruction that causes the machine to crash in a number of steps. We refer to this number of steps as the \textit{lookahead} parameter and in our experiments we use a lookahead of just 2 steps. If we cannot generate such an instruction, we retry with a smaller lookahead, until we succeed.

Since it now becomes possible to generate instruction streams that cause the machine to crash in some number of steps, one might be worried about the discard ratio for EENI. However, the ever increasing probability of generating a \textit{Halt} (discussed in §4) counterbalances this issue.

\textbf{Finding the bugs} We experimentally evaluated the effectiveness of testing for this new version of the machine, by adding the bugs discussed in this section to the ones applicable for the previous machine. The results of generation by execution with lookahead for this machine are shown in the first column of Figure 17. As we can see, all old bugs are still found relatively fast. It takes slightly longer to find them when compared to the previous machine, but this is to be expected: when we extend the machine, we are also increasing the state space to be explored. The new control-flow-specific bugs are all found, with the exception of Pop\textsuperscript{*}. Discard rates are much higher compared to generation by execution in Figure 5, for two reasons. First, control flow can cause loops, so we discard machines that run for more than 50 steps without halting. Second, as described previously, generation by execution in the presence of control flow is much less accurate.

\begin{figure}
\centering
\begin{tabular}{|c|c|c|}
\hline
\textit{Push} & \textit{Pop} & \textit{Store} \\
\hline
\end{tabular}
\caption{Counterexample to \textit{Push} and \textit{Store}}
\end{figure}

\begin{table}
\centering
\begin{tabular}{|c|c|c|c|}
\hline
\textit{Store} & \textit{Push} & \textit{Pop} \\
\hline
\end{tabular}
\caption{Counterexample to \textit{Store} and \textit{Push}}
\end{table}

\section{Strengthening the Tested Property}

The last few counterexamples in §5 are fairly long and quite difficult for QuickCheck to find, even with the best test-generation strategy. In this section we explore a different approach: strengthening the \textit{property} we are testing so that counterexamples become shorter and easier to find. Figure 17 summarizes the variants of non-interference that we consider and how they affect test performance.

\textbf{Making entire low states observable} Every counterexample that we’ve seen involves pushing an address, executing a \textit{Store} instruction, and halting. These steps are all necessary because of the choice we made in §2 to ignore the stack when defining indistinguishability on machine states. A counterexample that leaks a secret onto the stack must continue by storing it into memory; similarly, a counterexample that leaks a secret into the \textit{pc} must execute \textit{Store} at least twice. This suggests that we can get shorter counterexamples by redefining indistinguishability as follows:

\textbf{6.1 Definition:} Machine states \( S_1 = [pc_1 \, s_1 \, m_1 \, i_1] \) and \( S_2 = [pc_2 \, s_2 \, m_2 \, i_2] \) are indistinguishable with respect to entire low states, written \( S_1 \approx_{\text{low}} S_2 \), if either \( L_{pc_1} = L_{pc_2} = 1 \) or else \( L_{pc_1} = L_{pc_2} = 0, m_1 \approx m_2, i_1 \approx i_2, s_1 \approx s_2, \) and \( pc_1 \approx pc_2 \).

We now strengthen \textit{EENIInit.Halted\textsuperscript{Low}, \approx_{\text{mem}}} the property we have been testing, to \textit{EENIInit.Halted\textsuperscript{Low}, \approx_{\text{low}}}; this is stronger because \( \approx_{\text{mem}} \) and \( \approx_{\text{low}} \) agree on initial states, while for halted states \( \approx_{\text{low}} \subset \approx_{\text{mem}} \). Indeed, for this stronger property, QuickCheck finds bugs faster (compare the first two columns of Figure 17).

\textbf{Quasi-initial states} Many counterexamples begin by pushing values onto the stack and storing values into memory. This is necessary because each test starts with an empty stack and low, zeroed memory. We can make counterexamples easier to find by allowing the two machines to start with arbitrary (indistinguishable)

\footnote{Detailed profiling revealed that 18% of the pairs of machines both loop, and loopy machines push the average number of execution steps to 22.}
The function \( \text{trace} \) computes the (possibly infinite) list of states obtained by executing our machine starting from state \( S \). This is a function because our machine is deterministic.

6.2 Definition: \( \text{trace} \) is infinite

While, in practice, we test LLNI over finite prefixes of traces, the definition below is also valid for potentially infinite traces.

6.3 Definition: A machine semantics is \textit{low-lockstep noninterference} with respect to the indistinguishability relation \( \approx \) (written \( \text{LLNI}_\approx \)) if, for any quasi-initial states \( S_1 \) and \( S_2 \) with \( S_1 \approx S_2 \), we have \( \text{trace} \ S_1 \approx^* \text{trace} \ S_2 \), where \( \approx^* \) is defined coinductively by the following rules:

\[
S_1, S_2 \in \text{Low} \quad S_1 \approx S_2 \quad t_1 \approx^* t_2 \quad \Rightarrow \quad (S_1 : t_1) \approx^* (S_2 : t_2) \quad \text{(LOW LOCKSTEP)}
\]

\[
S_1 \notin \text{Low} \quad t_1 \approx^* t_2 \quad \Rightarrow \quad (S_1 : t_1) \approx^* t_2 \quad \text{(HIGH FILTER)}
\]

\[
\text{Halted} \quad S_1 \approx^* \quad \Rightarrow \quad S_1 \approx^* \quad \text{(LOCKSTEP END)}
\]

\[
S_1 \notin \text{Low} \quad (S_1) \approx^* t_2 \quad \Rightarrow \quad S_1 \approx^* \quad \text{(HIGH END)}
\]

\[
S_1 \in \text{Low} \quad S_1 \notin \text{Halted} \quad S_1 \approx S_2 \quad \Rightarrow \quad S_1 \approx^* (S_2 : t_2) \quad \text{(LOW ERROR END)}
\]

\[
t_1 \approx^* t_2 \quad t_2 \approx^* t_1 \quad \Rightarrow \quad \text{symmetry} \quad \text{(SYMMETRY)}
\]

The rule \text{LOW LOCKSTEP} requires low states in the two traces to be pointwise indistinguishable, while \text{HIGH FILTER} (together with \text{SYMMETRY}) simply filters out high states from either trace. The remaining rules are about termination: because we are working with termination-insensitive noninterference, we allow one of the traces to continue (maybe forever) even if the other has terminated in a state that is not low (\text{HIGH END}) or not halted (\text{LOW ERROR END}).

### Table 1: Experiments for the Control Flow Machine. MTTF given in milliseconds.

<table>
<thead>
<tr>
<th>Tested property</th>
<th>EENI</th>
<th>EENI</th>
<th>EENI</th>
<th>LLNI</th>
<th>SSNI</th>
<th>SSNI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting property</td>
<td>( \approx_{\text{mem}} )</td>
<td>( \approx_{\text{mem}} )</td>
<td>( \approx_{\text{mem}} )</td>
<td>( \approx_{\text{low}} )</td>
<td>( \approx_{\text{low}} )</td>
<td>( \approx_{\text{low}} )</td>
</tr>
<tr>
<td>Generation strategy</td>
<td>BYEXEC</td>
<td>BYEXEC</td>
<td>BYEXEC</td>
<td>BYEXEC</td>
<td>BYEXEC</td>
<td>BYEXEC</td>
</tr>
<tr>
<td>ADD*</td>
<td>37.07</td>
<td>2.38</td>
<td>1.38</td>
<td>0.36</td>
<td>0.24</td>
<td>0.11</td>
</tr>
<tr>
<td>PUSH*</td>
<td>0.22</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>1.06</td>
<td>0.06</td>
</tr>
<tr>
<td>LOAD*</td>
<td>155.07</td>
<td>37.50</td>
<td>5.73</td>
<td>1.14</td>
<td>3.25</td>
<td>0.61</td>
</tr>
<tr>
<td>STORE*A</td>
<td>20018.67</td>
<td>18658.56</td>
<td>124.78</td>
<td>84.08</td>
<td>289.63</td>
<td>16.32</td>
</tr>
<tr>
<td>STORE*B</td>
<td>13.02</td>
<td>12.87</td>
<td>16.10</td>
<td>5.25</td>
<td>31.11</td>
<td>0.33</td>
</tr>
<tr>
<td>STORE*C</td>
<td>0.35</td>
<td>0.34</td>
<td>0.33</td>
<td>0.08</td>
<td>0.73</td>
<td>0.03</td>
</tr>
<tr>
<td>JUMP*A</td>
<td>48.84</td>
<td>7.58</td>
<td>5.26</td>
<td>0.08</td>
<td>1.45</td>
<td>0.09</td>
</tr>
<tr>
<td>JUMP*B</td>
<td>2421.99</td>
<td>158.36</td>
<td>104.62</td>
<td>2.80</td>
<td>16.88</td>
<td>0.49</td>
</tr>
<tr>
<td>STORE*D</td>
<td>13289.39</td>
<td>12295.65</td>
<td>873.79</td>
<td>232.19</td>
<td>8.77</td>
<td>1.13</td>
</tr>
<tr>
<td>STORE*E</td>
<td>1047.56</td>
<td>1129.48</td>
<td>717.72</td>
<td>177.75</td>
<td>2.26</td>
<td>0.29</td>
</tr>
<tr>
<td>CALL*A</td>
<td>3919.08</td>
<td>174.66</td>
<td>118.15</td>
<td>5.97</td>
<td>31.71</td>
<td>0.62</td>
</tr>
<tr>
<td>RETURN*A</td>
<td>12804.51</td>
<td>4698.17</td>
<td>1490.80</td>
<td>337.74</td>
<td>1110.09</td>
<td>3.10</td>
</tr>
<tr>
<td>CALL<em>B+RETURN</em>B</td>
<td>69081.50</td>
<td>6940.67</td>
<td>1811.66</td>
<td>396.37</td>
<td>1194.30</td>
<td>4.56</td>
</tr>
</tbody>
</table>

### Figure 17

<table>
<thead>
<tr>
<th>MTTF arithmetic mean</th>
<th>6847.81</th>
<th>1526.75</th>
<th>219.46</th>
<th>194.44</th>
<th>2.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTTF geometric mean</td>
<td>135.76</td>
<td>46.48</td>
<td>7.69</td>
<td>12.87</td>
<td>0.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Average tests / second</th>
<th>2795</th>
<th>2797</th>
<th>2391</th>
<th>1224</th>
<th>8490</th>
<th>18407</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average discard rate</td>
<td>65%</td>
<td>65%</td>
<td>69%</td>
<td>0%</td>
<td>40%</td>
<td>9%</td>
</tr>
</tbody>
</table>
Additionally, we allow the two traces to terminate simultaneously (LOCKSTEP END). We implement these rules in Haskell as a recursive predicate over lazy lists.

In general, LLNI implies EENI, but not vice versa. However, the correct version of our machine does satisfy LLNI, and we have not observed any cases where QuickChecking a buggy machine with LLNI finds a bug that is not also a bug with regard to EENI. Testing LLNI instead of EENI leads to significant improvement in the bug detection rate for all bugs, as the results in the fourth column of Figure 17 show. In these experiments no generated machine states are discarded, since LLNI applies to both successful (halting) executions and failing or infinite executions. The generation strategies described in §4 also apply to LLNI without much change; also, as for EENI, generation by execution (with lookahead of 2 steps) performs better than the more basic strategies, so we don’t consider those for LLNI.

SSNI: Single-step noninterference Until now, we have focused on using sophisticated (and potentially slow) techniques for generating long-running machine states, and then checking equivalence for low halting states (EENI) or at every low step (LLNI). An alternative is to define a stronger property that talks about all possible single steps of execution starting from two indistinguishable states.

Proofs of noninterference usually go by induction on a pair of execution traces; to preserve the corresponding invariant, the proof needs to consider how each execution step affects the indistinguishability relation. This gives rise to properties known as “unwinding conditions” [14]; the corresponding conditions for our machine form a property we call single-step noninterference (SSNI).

We start by observing that LLNI implies that, if two low states are indistinguishable and each takes a step to another low state, then the resulting states are also indistinguishable. However, this alone is not a strong enough invariant to guarantee the indistinguishability of whole traces. In particular, if the two machines perform a Return from a high state to a low state, we would need to conclude that the two low states are equivalent without knowing anything about the original high states. This indicates that, for SSNI, we can no longer consider all high states indistinguishable. The indistinguishability relation on high states needs to be strong enough to ensure that when both machines return to low states, those low states are also indistinguishable. Moreover, we need to ensure that if one of the machines takes a step from a high state to another high state, then the old and new high states are equivalent. The following definition captures all these constraints formally; we write \( S^L \) for a machine state whose pc label is \( L \).

6.4 Definition: A machine semantics is single-step noninterfering with respect to the indistinguishability relation \( \approx \) (written \( \text{SSNI}_{\approx} \)) if the following four conditions are all satisfied:

1. For all low states \( S_1^L \) and \( S_2^L \), if \( S_1^L \approx S_2^L \), \( S_1^H \Rightarrow S_1 \), and \( S_2^H \Rightarrow S_2 \), then \( S_1 \approx S_2 \);  
2. For all high states \( S^H \) with \( S^H \Rightarrow S \), we have \( S^H \approx S \);  
3. For all high states \( S_1^H \) and \( S_2^H \), if \( S_1^H \approx S_2^H \), \( S_1^L \Rightarrow S_1 \), \( S_2^L \Rightarrow S_2 \), and \( S_1^L \) and \( S_2^L \) are low, then \( S_1^L \approx S_2^L \);  
4. For all low states \( S_1^L \) and \( S_2^L \), if \( S_1^L \approx S_2^L \) and \( S_1^L \) is halted, then \( S_1^L \) is stuck.

Note that SSNI talks about completely arbitrary states, not just (quasi-)initial ones.

The definition of SSNI is parametric in the indistinguishability relation used, and it can take some work to find the right relation. As discussed above, \( \approx_{\text{full}} \) is too weak and QuickCheck can easily find counterexamples to condition 3, e.g., by choosing two indistinguishable machine states with \( \approx_{\text{full}} \Rightarrow \text{Return} \), \( \text{pc} = 0@\leftarrow \tau \), and \( s = [\text{Pop}, \text{pc} = 0@\leftarrow \tau]; \) after a single step the two machines have distinguishable \( pc = 0@\bot \) and \( 1@\bot \), respectively. On the other hand, treating high states exactly like low states in the indistinguishability relation is too strong. In this case QuickCheck finds counterexamples to condition 2, e.g., a single machine state with \( i = [\text{Pop}, \text{pc} = 0@\leftarrow \tau] \), and \( s = [0@\bot; \text{steps to a state with } \approx = 7] \), which would not be considered indistinguishable. These counterexamples show that indistinguishable high states can have different \( pc \)s and can have completely different stack frames at the top of the stack. So all we can require for two high states to be equivalent is that their memories and instruction memories agree and that the parts of the stacks below the topmost low return address are equivalent. This is strong enough to ensure condition 3.

6.5 Definition: Machine states \( S_1 = [pc_1 | s_1 | m_1 | i_1] \) and \( S_2 = [pc_2 | s_2 | m_2 | i_2] \) are indistinguishable with respect to whole machine states, written \( S_1 \approx_{\text{full}} S_2 \), if \( m_1 \approx m_2 \), \( i_1 \approx i_2 \), \( L_{pc_1} = L_{pc_2} \), and additionally

- if \( L_{pc_1} = \bot \) then \( s_1 \approx s_2 \) and \( pc_1 \approx pc_2 \), and
- if \( L_{pc_1} = \top \) then \( \text{cropStack} s_1 \approx \text{cropStack} s_2 \).

The \( \text{cropStack} \) helper function takes a stack and removes elements from the top until it reaches the first low return address (or until all elements are removed).

The fifth column of Figure 17 shows that, even with arbitrary starting states generated completely naively, \( \text{SSNI}_{\approx_{\text{full}}} \) performs very well. If we tweak the weights a bit and additionally observe that since we only execute the generated machine for only one step, we can begin with very small states (e.g., the instruction memory can be of size 2), then we can find all bugs very quickly. As the last column of Figure 17 illustrates, each bug is found in under 20 milliseconds. (This last optimization is a bit risky, since we need to make sure that these very small states are still large enough to exercise all bugs we might have—e.g., an instruction memory of size 1 is not enough to exhibit the \( \text{CALL} + \text{B} + \text{RETURN} + \text{B} \) bug using SSNI.) Compared to other properties, QuickCheck executes many more tests per second with SSNI for both generation strategies.

Discussion In this section we have seen that strengthening the noninterference property is a very effective way of improving the effectiveness of random testing our IFP machine. It is not without costs, though. Changing the security property required some expertise and, in the case of LLNI and SSNI, manual proofs showing that the new property implies EENI, the baseline security property. In the case of SSNI we used additional invariants of our machine (captured by \( \approx_{\text{full}} \)) and finding these invariants would probably constitute the most creative part of doing a full security proof. While we could use the counterexamples provided by QuickCheck to guide our search for the right invariants, we expect that for more realistic machines the process of interpreting the counterexamples and manually refining the invariants will be significantly harder than for our very simple machine.

The potential users of our techniques will have to choose a point in the continuum between testing and proving that best matches the characteristics of their practical application. At one end, we present ways of testing the original EENI property without changing it in any way, by putting all the smarts in clever generation strategies. At the other end, one can imagine using random testing just as the first step towards a full proof of a stronger property such as SSNI. For our simple machine, Azevedo de Amorim et al. [2] did in fact prove recently in Coq that SSNI holds, and did not find any bugs that had escaped our testing. Moreover, we proved in Coq that for any deterministic machine and for any indistinguishability relation that is an equivalence, SSNI implies LLNI and LLNI implies EENI.
7. Shrinking Strategies

The counterexamples presented in this paper are not the initial randomly generated tests; they are the result of QuickCheck shrinking these to minimal examples. For example, randomly generated counterexamples to EENI for the PUSH* bug usually consist of 20–40 instructions; the minimal counterexample uses just four. In this section we describe the shrinking strategies we used.

**Shrinking labeled values, instructions, and stack elements** By default, QuickCheck already implements a shrinking strategy for integers. For labels, we shrink $\mathbb{T} \rightarrow \bot$, because we prefer to see counterexamples in which labels are only $\mathbb{T}$ if this is essential to the failure. Values are shrunk by shrinking either the label or the contents. (If we need to shrink both the label and the contents, then this is achieved in two separate shrinking steps.)

We allow any instruction to shrink to $\texttt{Noop}$, which preserves a counterexample if the instruction was unnecessary; or to $\texttt{Halt}$, which preserves a counterexample if the bug had already manifested by the time that instruction was reached. To avoid an infinite shrinking loop, we do not shrink $\texttt{Noop}$ at all, while $\texttt{Halt}$ can only to $\texttt{Noop}$. Instructions of the form $\texttt{Push a}$ are also shrunk by shrinking $a$. Finally, instructions of the form $\texttt{Call a r}$ are also shrunk by shrinking $a$ or $r$, or by being replaced with $\texttt{Jump}$.

For quasi-initial or arbitrary states the stack contains a mixture of values and return addresses, which are shrunk pointwise.

**Machine states** Machine states contain an instruction memory, a data memory, a stack, and the initial pc, the first three of which can be shrunk. We allow any element of the memories or the stack to be shrunk by the methods above; additionally, shrinking may remove elements from any of the three.

We allow shrinking to remove arbitrary elements of the data memory or the stack, but in the case of the data memory we first try to remove the last value. This is because removing other elements changes all subsequent memory addresses, likely invalidating the counterexample and thus rendering the shrinking step unsuccessful.

In the case of the instruction memory, we only try to remove $\texttt{Noop}$ instructions, since removing other instructions is likely to change the stack or the control flow fairly drastically, and thus risks invalidating a counterexample. Other instructions can still be removed in two stages, by first shrinking them to a $\texttt{Noop}$.

**Variations** One difficulty that arises when shrinking noninterference counterexamples is that the test cases must be pairs of indistinguishable machines. Shrinking each machine state independently will most likely yield distinguishable pairs, which are invalid test cases, since they fail to satisfy the precondition of the property we are testing. In order to shrink effectively, we need to shrink both states of a variation simultaneously, and in the same way.

For instance, if we shrink one machine state by deleting a $\texttt{Noop}$ in the middle of its instruction memory, then we must delete the same instruction in the corresponding variation. Similarly, if a particular element gets shrunk in a memory location, then the same location should be shrunk in the other state of the variation, and only in ways that produce indistinguishable states. We have implemented all of the shrinking strategies described above as operations on pairs of indistinguishable states, and ensured that they generate only shrinking candidates that are also indistinguishable.

When we use the full state equivalence $\approx_{\text{full}}$, we can shrink stacks slightly differently: we only need to synchronize shrinking steps on the low parts of the stacks. Since the equivalence relation ignores the high half of the stacks, we are free to shrink those parts of the two states independently, provided that high return addresses don’t get transformed into low ones.

**Optimizing shrinking** We applied a number of optimizations to make the shrinking process faster and more effective. One way we sped up shrinking was by turning on QuickCheck’s “smart shrinking,” which optimizes the order in which shrinking candidates are tried. If a counterexample $a$ can be shrunk to any $b_i$, but the first $k$ of these are not counterexamples, then it is likely that the first $k$ shrinking candidates for $b_{k+1}$ will not be counterexamples either, because $a$ and $b_{k+1}$ are likely to be similar in structure and so to have similar lists of shrinking candidates. Smart shrinking just changes the order in which these candidates are tried: it defers the first $k$ shrinking candidates for $b_{k+1}$ until after more likely ones have been tried. This sped up shrinking dramatically in our tests.

We also observed that many reported counterexamples contained $\texttt{Noop}$ instructions—in some cases many of them—even though we implemented $\texttt{Noop}$ removal as a shrinking step. On examining these counterexamples, we discovered that they could not be shrunk because removing a $\texttt{Noop}$ changes the addresses of subsequent instructions, at least one of which was the target of a $\texttt{Jump}$ or $\texttt{Call}$ instruction. To preserve the behavior of the counterexample, we needed to remove the $\texttt{Noop}$ instruction and adjust the target of a control transfer in the same shrinking step. Since control transfer targets are taken off the stack, and such values can be generated during the test in many different ways, we simply allowed $\texttt{Noop}$ removal to be combined with any other shrinking step—which might, for example, decrement any value on the initial stack, or any value stored in the initial memory, or any constant in a $\texttt{Push}$ instruction. This combined shrinking step was much more effective in removing unnecessary $\texttt{Noops}$.

Occasionally, we observed shrunk counterexamples containing two or more unnecessary $\texttt{Noops}$, but where removing just one $\texttt{Noop}$ led to a non-counterexample. We therefore used QuickCheck’s double shrinking, which allows a counterexample to shrink in two steps to another counterexample, even if the intermediate value is not a counterexample. With this technique, QuickCheck could remove all unnecessary $\texttt{Noops}$, albeit at a cost in shrinking time.

We also observed that some reported test cases contained unnecessary sequences of instructions, which could be elided together, but not one by one. We added a shrinking step that can replace any two instructions by $\texttt{Noops}$ simultaneously (and thus, thanks to double shrinking, up to four), which solved this problem.

With this combination of methods, almost all counterexamples we found shrink to minimal examples, from which no instruction, stack element, or memory element could be removed without invalidating the counterexample.

8. Related Work

Generating random inputs for testing is a large research area, but the particular sub-area of testing language implementations by generating random programs is less well studied. PLT Redex [18] is a domain-specific language for defining operational semantics within PLT Scheme, which includes a property-based random testing framework inspired by QuickCheck. This framework uses a formalized language definition to automatically generate simple test cases. To generate better test cases, however, Klein et al. find that the generation strategy needs to be tuned for the particular language; this agrees with our observation that fine-tuned strategies are required to obtain the best results. They argue that the effort required to find bugs using PLT Redex is less than the effort required for a formal proof of correctness, and that random testing is sometimes viable in cases where full proof seems infeasible.

CSmith [27] is a C compiler testing tool that generates random C programs, avoiding ones whose behavior is undefined by the C99 standard. When generating programs, CSmith does not attempt to model the current state of the machine; instead, it chooses program fragments that are correct with respect to some static safety analysis (including type-, pointer-, array-, and initializer-safety, etc.). We found that modeling the actual state of our (much simpler) machine
to check that generated programs were hopefully well-formed, as
in our generation by execution strategy, made our test-case genera-
tion far more effective at finding noninterference bugs. In order
to get smaller counterexamples, Regehr et al. present C-Reduce [21],
a tool for reducing test-case C programs such as those produced
by CSmith. They note that conventional shrinking methods usually
introduce test cases with undefined behavior; thus, they put a great
deal of effort and domain specific knowledge into shrinking well-
defined programs only to programs that remain well-defined. To do
this, they use a variety of search techniques to find better reduc-
tion steps and to couple smaller ones together. Our use of Quick-
Check’s double shrinking is similar to their simultaneous reduc-
tions, although we observed no need in our setting for more sophis-
ticated searching methods than the greedy one that is guaranteed
to produce a local minimum. Regehr et al.’s work on reduction is
partly based on Zeller and Hildebrandt’s formalization of the delta
debugging algorithm admn [29], a non-domain-specific algorithm
for simplifying and isolating failure-inducing program inputs with
an extension of binary search. In our work, as in Regehr et al.’s,
domain-specific knowledge is crucial for successful shrinking.

Another relevant example of testing programs by generating
random input is Randoop [19], which generates random sequences
of calls to Java APIs. Noting that many generated sequences crash
after only a few calls, before any interesting bugs are discovered,
Randoop performs feedback directed random testing, in which pre-
viously found sequences that did not crash are randomly extended.
This enables Randoop to generate tests that run much longer be-
fore crashing, which are much more effective at revealing bugs.
Our generation by execution strategy is similar in spirit, and like-
wise results in a substantial improvement in bug detection rates.

A state-machine modeling library for (an Erlang version of)
QuickCheck has been developed by Quviq [17]. It generates se-
quencies of API calls to a stateful system satisfying preconditions
formulated in terms of a model of the system state, associating a
(model) state transition function with each API call. API call gen-
erators also use the model state to avoid generating calls whose pre-
conditions cannot be satisfied. Our generation-by-execution strat-
ey works in a similar way for straightforward code.

A powerful and widely used approach to testing is symbolic
execution—in particular, concolic testing and related dynamic
symbolic execution techniques [8]. The idea is to mix symbolic
and concrete execution in order to achieve higher code coverage.
The choice of which concrete executions to generate is guided by
a constraint solver and path conditions obtained from the symbolic
executions. Originating with DART [13] and PathCrawler [26], a
variety of tools and methods have appeared. We wondered whether
dynamic symbolic execution could be used instead of random test-
ing for finding noninterference bugs. As a first step, we imple-
mented a simulator for a version of our abstract machine in C and
tested it with KLEE [7], a state-of-the-art symbolic execution tool.
Using KLEE out of the box and without any expert knowledge in
the area, we attempted to invalidate various assertions of noninter-
ference. Unfortunately, we were only able to find a counterexample
for PUSH*, the simplest possible bug, in addition to a few imple-
mentation errors (e.g., out-of-bound pointers for invalid machine
configurations). The main problem seems to be that the state space
we need to explore is too large, so we don’t cover enough of it to
reach the particular IFC-violating configurations.

In interactive theorem provers, automatically generating coun-
terexamples for false conjectures can prevent wasting time and
error on proof attempts doomed to fail. Dybjer [10] propose a
QuickCheck-like tool for the Agda/Alfa proof assistant. Berghofer
and Nipkow [3] proposed a QuickCheck-like tool for Isabelle/HOL.
This was recently extended by Bulwahn [5] to also support exhaust-
ive and narrowing-based symbolic testing. Moreover, Bulwahn’s
tool uses Horn clause data flow analysis to automatically devise
generators that only produce data that satisfies the precondition of
the tested conjecture [6]. Eastlund [11] implemented DoubleCheck,
an adaption of QuickCheck for ACL2. Chamari et al. [20] later
proposed a more advanced counterexample finding tool for ACL2s,
which uses the full power of the theorem prover and libraries to
simplify conjectures so that they are easier to falsify. While all these
tools are general and only require the statement of the conjecture
to be in a special form (e.g., executable specification), so they could
in principle be applied to test noninterference, our experience with
QuickCheck suggests that for the best results one has to incorpo-
rate domain knowledge about the machine and the property being
tested. We hope to compare our work against these tools in the fu-
ture and provide experimental evidence for this intuition.

On the dynamic IFC side Birgisson et al. [4] have a good
overview of related work. Our correct rule for Store is called the
no-sensitive-upgrades policy in the literature and was first proposed
by Zdancewic [28] and later adapted to the purely dynamic IFC setting
by Austin and Flanagan [1]. Hedin and Sabelfeld [15] improve the
precision of the no-sensitive-upgrades policy by explicit upgrade
annotations, which raise the level of a location before branching
on secrets. They apply their technique to a core calculus of JavaScript
that includes objects, higher-order functions, exceptions, and, dy-
amic code evaluation. Birgisson et al. [4] show that random test-
ing with QuickCheck can be used to infer upgrade instructions in
this setting. The main idea is that whenever a random test causes
the program to be stopped by the IFC monitor because it attempts a
sensitive upgrade, the program can be rewritten by introducing an
upgrade annotation that prevents the upgrade from being deemed
sensitive on the next run of the program.

9. Conclusions and Outlook

We have shown how random testing can be used to discover
counterexamples to noninterference in a simple information-flow
machine and how to shrink counterexamples discovered in this way
to simpler, more comprehensible ones. The techniques we present
bring many orders of magnitude improvement in the rate at which
bugs are found, and for the hardest-to-find bugs (to EENI) the min-
imal counterexamples are 10-15 instructions long—well beyond
the scope of naive exhaustive testing. Even if we ultimately care
about full security proofs [2], using random testing should greatly
speed the initial design process and allow us to concentrate more
of our energy on proving things that are correct or nearly correct.

What crucially remains to be seen is whether our results will
scale up to more realistic settings. We are actively pursuing this
question in the context of CRASH/SAFE, an ambitious hardware–
software co-design effort underway at Penn, Harvard, Northeast-
er, and BAE Systems, with IFC as a key feature at all levels, from
hardware to application code. The design involves a number of ab-
stract machines, all much more complex than the stack machine we
have studied here. We hope to use random testing both for checking
noninterference properties of individual abstract machines and for
checking that the code running on lower-level abstract machines
correctly implements the higher-level abstractions.

Just how well do our methods need to perform to find bugs ef-
effectively in these machines? The true answer is anybody’s guess,
but for a very rough estimate we might guess that there will be around
10× as many instructions on a realSAFE machine, that each in-
struction might be on average 3× more complex, and that several
cross-cutting features will induce additional complexity factors—
e.g., “public labels” [16] (2×), dynamic memory allocation
(2×), a “fat pointer” memory model (2×), a more complex lattice
of labels (2×), and dynamic principal generation (2×). Combin-
ing these, we could be finding bugs more than 1000× more slowly.
If this calculation is in the right ballpark, then EENI is nowhere

near fast enough: even for our stack machine, it can take several
minutes to find some bugs. Between LLNI and SSNI, on the other
hand, there is a tradeoff (discussed in §6) between the overhead of
finding good invariants for SSNI and the increased bug-finding rate
once this is done. Both approaches seem potentially useful (and
potentially fast enough), perhaps at different points in the design
process. In particular, checking SSNI may help find invariants that
will also be needed for a formal proof.

Preliminary experiments on testing a more complex register ma-
chine with some of the extra features enumerated above show that
most of our techniques scale directly to the new setting with only
minor modifications. For this more complex machine, EENI is (as
expected) impractical for finding non-trivial IFC bugs. LLNI vari-
ants, on the other hand, uncovered a plethora of (truly accidental)
bugs in our initial implementation, and also aided in the discov-
er of increasingly strong invariants. So far we have been unable
to come up with a correct SSNI invariant; however, we have been
using the testing framework to hone in on a correct formulation.

We expect that our techniques are flexible enough to be applied
to checking other special classes of programs (i.e., properties
of pairs of related runs)—in particular, the many variants and gen-
eralizations of noninterference. Beyond noninterference properties,
preliminary experiments with checking correspondence between
conce concrete and abstract versions of our current stack machine suggest
that many of our techniques can also be adapted for this purpose.
For example, the generate-by-execution strategy and many of the
shrinking tricks apply just as well to single programs as to pairs of
related programs. This gives us hope that they may be useful for
checking yet further properties of abstract machines.

Acknowledgments We thank the participants in the discussion
at the IFIP WG 2.8 Storulvån meeting that originated this work:
Ulf Norell, Rishiyur S. Nikhil, Michal Palka, Nick Smallbone,
and Meng Wang. We also thank Johannes Borgström, Cristian
Cadar, Delphine Demange, Matthias Felleisen, Robby Findler,
Alex Groce, Casey Klein, Ben Karel, Scott Moore, Michal Palka,
John Regehr, Howard Reubenstein, Alejandros Russo, Deian Stefan,
Greg Sullivan, and Andrew Tolmach, for providing feedback on a
draft, and the members of the CRASH/SAFE team and Manolis Pa-

dakis for fruitful discussions. This material is based upon work

funded under NSF grants CCF-10-C-7090. The work is also partially funded under
10-C-7090. The views expressed are those of the authors and do not

relate to the U.S. Government. The work is also partially funded under

References

flow analysis. In Workshop on Programming Languages and Analysis

D. Pichardie, B. C. Pierce, R. Pollack, and A. Tolmach. A verified
information-flow architecture. Under submission, July 2013.

International Conference on Software Engineering and Formal

of dynamic information-flow tracking by testing. In 17th European

and symbolic testing under one roof. In 2nd International Conference
on Certified Programs and Proofs (CPP), volume 7679 of Lecture Notes
in Computer Science. 2012.

International Conference on Logic for Programming, Artificial
Intelligence, and Reasoning (LPAR), volume 7180 of Lecture Notes in
Computer Science. 2012.

generation of high-coverage tests for complex systems programs. In 6th
USENIX conference on Operating systems design and implementation,
OSDI. 2008.

mann, and W. Visser. Symbolic execution for software testing in prac-
tice: preliminary assessment. In 33rd International Conference on Soft-

testing of Haskell programs. In 5th ACM SIGPLAN International

[10] P. Dybjer, Q. Haiyan, and M. Takeyama. Combining testing and
proving in dependent type theory. In 16th International Conference
on Theorem Proving in Higher Order Logics (TPHOLs), volume 2758


random testing. In ACM SIGPLAN Conference on Programming Lan-
guage Design and Implementation, PLDI. 2005.


[15] D. Hedin and A. Sabelfeld. Information-flow security for a core of
(CSF), CSF. 2012.

All your IFCException are belong to us. In 34th IEEE Symposium on

[17] J. Hughes. QuickCheck testing for fun and profit. In 9th Internationa

Conference on Practical Aspects of Declarative Languages (PADD),
volume 4354 of Lecture Notes in Computer Science. 2007.

[18] C. Klein, J. Clements, C. Dimoulas, C. Eastlund, M. Felleisen,
M. Flatt, J. A. McCarthy, J. Ratzkin, S. Tobin-Hochstadt, and R. B.
Findler. Run your research: On the effectiveness of lightweight mecha-
nization. In Principles of Programming Languages (POPL), 2012.

[19] C. Pacheco and M. D. Ernst. Randoop: feedback-directed random test-
ing for Java. In 22nd ACM SIGPLAN Conference on Object-Oriented

Integrating testing and interactive theorem proving. In ACL2, 2011.

Test-case reduction for C compiler bugs. In 33rd ACM SIGPLAN conference

[22] A. Russo and A. Sabelfeld. Dynamic vs. static flow-sensitive security
analysis. In 23rd Computer Security Foundations Symposium (CSF),
CSF. 2010.

IEEE Journal on Selected Areas in Communications, 21(1):5–19,

[24] A. Sabelfeld and A. Russo. From dynamic to static and back: Riding
the roller coaster of information-flow control research. In Ershov
Memorial Conference. 2009.

information flow control in Haskell. In 4th Symposium on Haskell.
2011.

[26] N. Williams, B. Marre, and P. Mouy. On-the-fly generation of K-
path tests for C functions. In 19th IEEE International Conference on

[27] X. Yang, Y. Chen, E. Eide, and J. Regehr. Finding and understanding
bugs in C compilers. In ACM SIGPLAN Conference on Programming
Language Design and Implementation, PLDI. 2011.

[28] S. A. Zdancewic. Programming Languages for Information Security,
Weakening EENI when adding calls and returns

The counterexample in Figure 20 shows that once we have a way to restore the pc label, we can no longer expect all pairs of halting states to be indistinguishable in EENI. In particular, the counterexample shows, one machine can halt in a high state, while the other can return to low, and only then halt. Since our indistinguishability relation only equates states with the same pc label, these two halting states are distinguishable. The solution we use in §5 is to weaken the EENI instance, by considering only ending states that are both halting and low (i.e., we change to EENI_{init, Halted, Low, \neq_{\text{meas}}}).

Indistinguishability for stack elements when adding calls and returns

In §5 we defined the indistinguishability relation on stack elements so that return addresses are only equivalent to other return addresses and (as for values) R(x1@L1) \approx R(x2@L2) if either L1 = L2 = \top or x1 = x2 and L1 = L2 = \bot. If instead we considered high return addresses and high values to be indistinguishable, QuickCheck would find a counterexample. This counterexample requires quasi-initial states (and \neq_{\text{meas}}) and is listed in Figure 21. The first machine performs only one Return that throws away two elements from the stack and then halts. The second machine returns twice: the first time to the same Return, unwinding the stack and raising the pc; and the second time to the Halt instruction, labeling the return value high in the process. The final states are indistinguishable because the elements on the stack have different labels. As we saw earlier, such a counterexample can be extended to one in which values also differ.

```
\[
\begin{array}{cccc}
\text{pc} & \text{m} & \text{s} & \text{i(pc)} \\
0@\perp & [\bot] & \bot & \text{Push 1@\bot} \\
1@\bot & [\bot] & \bot & \text{Push 0@\top} \\
2@\perp & [\bot, \bot] & \bot & \text{Push 1@\bot} \\
3@\bot & [\bot, 0@\top] & [\bot] & \text{Store} \\
4@\bot & [0@\top, 0@\top] & [\bot] & \text{Push 1@\top} \\
5@\bot & [0@\top, 0@\top] & [1@\top] & \text{Store} \\
6@\bot & [0@\top, 1@\top] & [\bot] & \text{Halt} \\
\end{array}
\]
```

Figure 18. A counterexample showing that it is wrong to make high atoms equivalent to all other atoms.

```
\[
\begin{array}{cccc}
\text{pc} & \text{m} & \text{s} & \text{i(pc)} \\
0@\bot & [0@\bot] & [\bot] & \text{Push 1@\bot} \\
1@\bot & [0@\bot] & [1@\bot] & \text{Push 0@\top} \\
2@\bot & [0@\bot] & [0@\top, 1@\bot] & \text{Push 0@\top} \\
3@\bot & [0@\bot] & [1@\bot] & \text{Store} \\
4@\bot & [0@\top] & [1@\bot] & \text{Push 1@\top} \\
5@\bot & [0@\top] & [1@\bot] & \text{Call 1@0} \\
\end{array}
\]
```

Machine 1 continues...

```
\[
\begin{array}{cccc}
\text{pc} & \text{m} & \text{s} & \text{i(pc)} \\
7@\top & [0@\bot] & [1@\bot, R(6,0)@\bot] & \text{Push 0@\bot} \\
8@\top & [0@\bot] & [0@\bot, 1@\bot, R(6,0)@\bot] & \text{Store} \\
9@\bot & [0@\top] & [R(6,0)@\bot] & \text{Return} \\
6@\bot & [1@\bot] & [\bot] & \text{Halt} \\
\end{array}
\]
```

Machine 2 continues...

```
\[
\begin{array}{cccc}
\text{pc} & \text{m} & \text{s} & \text{i(pc)} \\
7@\top & [0@\bot] & [1@\bot, R(6,0)@\bot] & \text{Return} \\
8@\bot & [0@\bot] & [0@\bot] & \text{Halt} \\
\end{array}
\]
```

Figure 19. A counterexample showing that it is also wrong to make high atoms equivalent to low atoms with the same payload.

```
\[
\begin{array}{cccc}
\text{pc} & \text{m} & \text{s} & \text{i(pc)} \\
0@\bot & [0@\bot] & [\bot] & \text{Push 1@\bot} \\
1@\bot & [0@\bot] & [1@\bot] & \text{Push 0@\top} \\
2@\bot & [0@\bot] & [0@\top, 1@\bot] & \text{Push 0@\top} \\
3@\bot & [0@\bot] & [0@\bot] & \text{Store} \\
4@\bot & [0@\bot] & [1@\bot] & \text{Push 1@\top} \\
5@\bot & [0@\bot] & [1@\bot] & \text{Call 1@0} \\
\end{array}
\]
```

Machine 1 continues...

```
\[
\begin{array}{cccc}
\text{pc} & \text{m} & \text{s} & \text{i(pc)} \\
1@\bot & [\bot] & [\bot] & \text{Push 2@\top} \\
2@\bot & [\bot] & [\bot] & \text{Call 0@0} \\
\end{array}
\]
```

Machine 2 continues...

```
\[
\begin{array}{cccc}
\text{pc} & \text{m} & \text{s} & \text{i(pc)} \\
3@\top & [\bot] & [R(2,0)@\bot] & \text{Return} \\
2@\bot & [\bot] & [\bot] & \text{Halt} \\
\end{array}
\]
```

Figure 20. A counterexample justifying the change to EENI_{init, Halted, Low, \neq_{\text{meas}}}, in §5.

```
\[
i = [\text{Push 1@\bot, Push 0@\top, Push 0@\bot, Store, Push \text{Return}}]
\]
```

```
\[
i = [\text{Push 2@\top, Call 0@0, Halt, Return}]
\]
```

```
\[
i = [\text{Return, Halt}]
\]
```