Under-Constrained Symbolic Execution: Correctness Checking for Real Code

David A. Ramos and Dawson Engler, Stanford University

https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/ramos
Under-Constrained Symbolic Execution: Correctness Checking for Real Code

David A. Ramos
ramos@cs.stanford.edu

Dawson Engler
engler@csl.stanford.edu

Stanford University

Abstract
Software bugs are a well-known source of security vulnerabilities. One technique for finding bugs, symbolic execution, considers all possible inputs to a program but suffers from scalability limitations. This paper uses a variant, under-constrained symbolic execution, that improves scalability by directly checking individual functions, rather than whole programs. We present UC-KLEE, a novel, scalable framework for checking C/C++ systems code, along with two use cases. First, we use UC-KLEE to check whether patches introduce crashes. We check over 800 patches from BIND and OpenSSL and find 12 bugs, including two OpenSSL denial-of-service vulnerabilities. We also verify (with caveats) that 115 patches do not introduce crashes. Second, we use UC-KLEE as a generalized checking framework and implement checkers to find memory leaks, uninitialized data, and unsafe user input. We evaluate the checkers on over 20,000 functions from BIND, OpenSSL, and the Linux kernel, find 67 bugs, and verify that hundreds of functions are leak free and that thousands of functions do not access uninitialized data.

1 Introduction
Software bugs pervade every level of the modern software stack, degrading both stability and security. Current practice attempts to address this challenge through a variety of techniques, including code reviews, higher-level programming languages, testing, and static analysis. While these practices prevent many bugs from being released to the public, significant gaps remain.

One technique, testing, is a useful sanity check for code correctness, but it typically exercises only a small number of execution paths, each with a single set of input values. Consequently, it misses bugs that are only triggered by other inputs.

Another broad technique, static analysis, is effective at discovering many classes of bugs. However, static analysis generally uses abstraction to improve scalability and cannot reason precisely about program values and pointer relationships. Consequently, static tools often miss deep bugs that depend on specific input values.

One promising technique that addresses the limitations of both testing and static analysis is symbolic execution [4, 5, 40]. A symbolic execution tool conceptually explores all possible execution paths through a program in a bit-precise manner and considers all possible input values. Along each path, the tool determines whether any combination of inputs could cause the program to crash. If so, it reports an error to the developer, along with a concrete set of inputs that will trigger the bug. Unfortunately, symbolic execution suffers from the well-known path explosion problem since the number of distinct execution paths through a program is often exponential in the number of if-statements or, in the worst case, infinite. Consequently, while symbolic execution often examines orders of magnitude more paths than traditional testing, it typically fails to exhaust all interesting paths. In particular, it often fails to reach code deep within a program due to complexities earlier in the program. Even when the tool succeeds in reaching deep code, it considers only the input values satisfying the few paths that manage to reach this code.

An alternative to whole-program symbolic execution is under-constrained symbolic execution [18, 42, 43], which directly executes an arbitrary function within the program, effectively skipping the costly path prefix from main to this function. This approach reduces the number and length of execution paths that must be explored. In addition, it allows library and OS kernel code without a main function to be checked easily and thoroughly.

This paper presents UC-KLEE, a scalable framework implementing under-constrained symbolic execution for C/C++ systems code without requiring a manual specification or even a single testcase. We apply this framework to two important use cases. First, we use it to check whether patches to a function introduce new bugs, which may or may not pose security vulnerabilities. Ironically, patches intended to fix bugs or eliminate security vulnerabilities are a frequent source of them. In many cases,
UC-KLEE can verify (up to a given input bound and with standard caveats) that a patch does not introduce new crashes to a function, a guarantee not possible with existing techniques.

Second, we use UC-KLEE as a general code checking framework upon which specific checkers can be implemented. We describe three example checkers we implemented to find memory leaks, uses of uninitialized data, and unsanitized uses of user input, all of which may pose security vulnerabilities. Additional checkers may be added to our framework to detect a wide variety of bugs along symbolic, bit-precise execution paths through functions deep within a program. If UC-KLEE exhaustively checks all execution paths through a function, then it has effectively verified (with caveats) that the function passes the check (e.g., no leaks).

We evaluated these use cases on large, mature, and security-critical code. We validated over 800 patches from BIND [3] and OpenSSL [36] and found 12 bugs, including two OpenSSL denial-of-service vulnerabilities [12, 16]. UC-KLEE verified that 115 patches did not introduce new crashes, and it checked thousands of paths and achieved high coverage even on patches for which it did not exhaust all execution paths.

We applied our three built-in checkers to over 20,000 functions from BIND, OpenSSL, and the Linux kernel and discovered 67 new bugs, several of which appear to be remotely exploitable. Many of these were latent bugs that had been missed by years of debugging effort. UC-KLEE also exhaustively verified (with caveats) that 771 functions from BIND and OpenSSL that allocate heap memory do not cause memory leaks, and that 4,088 functions do not access uninitialized data.

The remainder of this paper is structured as follows: § 2 presents an overview of under-constrained symbolic execution; § 3 and § 4 discuss using UC-KLEE for validating patches and generalized checking, respectively; § 5 describes implementation tricks; § 6 discusses related work; and § 7 concludes.

2 Overview

This paper builds upon our earlier work on UC-KLEE [43], an extension to the KLEE symbolic virtual machine [5] designed to support equivalence verification and under-constrained symbolic inputs. Our tool checks C/C++ code compiled as bitcode (intermediate representation) by the LLVM compiler [29]. As in KLEE, it performs bit-accurate symbolic execution of the LLVM bitcode, and it executes any functions called by the code. Unlike KLEE, UC-KLEE begins executing code at an arbitrary function chosen by the user, rather than main.

With caveats (described in § 2.2), UC-KLEE provides verification guarantees on a per-path basis. If it exhausts all execution paths, then it has verified that a function has the checked property (e.g., that a patch does not introduce any crashes or that the function does not leak memory) up to the given input size.

Directly invoking functions within a program presents new challenges. Traditional symbolic execution tools generate input values that represent external input sources (e.g., command-line arguments, files, etc.). In most cases, a correct program should reject invalid external inputs rather than crash. By contrast, individual functions typically have preconditions imposed on their inputs. For example, a function may require that pointer arguments be non-null. Because UC-KLEE directly executes functions without requiring their preconditions to be specified by the user, the inputs it considers may be a superset (over-approximation) of the legal values handled by the function. Consequently, we denote UC-KLEE’s symbolic inputs as under-constrained to reflect that they are missing preconditions (constraints).

While this technique allows previously-unreachable code to be deeply checked, the missing preconditions may cause false positives (spurious errors) to be reported to the user. UC-KLEE provides both automated heuristics and an interface for users to manually silence these errors by lazily specifying input preconditions using simple C code. In our experience, even simple annotations may silence a large number of spurious errors (see § 3.2.5) and this effort is orders of magnitude less work than eagerly providing a full specification for each function.

2.1 Lazy initialization

UC-KLEE automatically generates a function’s symbolic inputs using lazy initialization [26, 46], which avoids the need for users to manually construct inputs, even for complex, pointer-rich data structures. We illustrate lazy initialization by explaining how UC-KLEE executes the example function listSum in Figure 1(a), which sums the entries in a linked list. Figure 1(b) summarizes the three execution paths we explore. For clarity, we elide error checks that UC-KLEE normally performs at memory accesses, division/remainder operations, and assertions.

UC-KLEE first creates an under-constrained symbolic value to represent the sole argument n. Although n is a pointer, it begins in the unbound state, not yet pointing to any object. UC-KLEE then passes this symbolic argument to listSum and executes as follows:

Line 7 The local variable sum is assigned a concrete value; no special action is taken.

Line 8 The code checks whether the symbolic variable n is non-null. At this point, UC-KLEE forks execution and considers both cases. We first consider the false path where n = null (Path A). We then return to the true path where n ≠ null (Path B). On Path A, UC-KLEE adds n = null as a path constraint and skips the loop.

Line 12 Path A returns 0 and terminates.
We now consider Path B.

**Line 8** UC-KLEE adds the constraint $n \neq \text{null}$ and enters the loop.

**Line 9** The code dereferences the pointer $n$ for the first time on Path B. Because $n$ is unbound, UC-KLEE allocates a new block of memory, denoted node1, to satisfy the dereference and adds the constraint $n = \&\text{node1}$ to bind the pointer $n$ to this object. At this point, $n$ is no longer unbound, so subsequent dereferencess of that pointer will resolve to node1 rather than trigger additional allocations. The (symbolic) contents of node1 are marked as unbound, allowing future dereferences of pointers in this object to trigger allocations. This recursive process is the key to lazy initialization. Next, $\text{sum} = 0$ is incremented by the symbolic value node1.val.

**Line 10** $n$ is set to the value node1.next. Path B then returns to the loop header.

**Line 8** The code tests whether $n$ (set to node1.next) is non-null. UC-KLEE forks execution and considers both cases. We first consider node1.next = null, which we still refer to as Path B. We will then return to the true path where node1.next $\neq$ null (Path C). On Path B, node1.next = null is added as a path constraint and execution exits the loop.

**Line 12** Path B returns node1.val and terminates.

We now consider Path C.

**Line 8** UC-KLEE adds node1.next $\neq$ null as a path constraint, and Path C enters the loop.

**Line 9** Path C dereferences the unbound symbolic pointer node1.next, which triggers allocation of a new object node2. This step illustrates the unbounded nature of many loops. To prevent UC-KLEE from allocating an unbounded number of objects as input, the tool accepts a command-line option to limit the depth of an input-derived data structure ($k$-bounding [17]). When a path attempts to exceed this limit, our tool silently terminates it. For this example, assume a depth limit of two, which causes UC-KLEE to terminate Path D (not shown) at line 9 during the next loop iteration.

**Line 10** $n$ is set to the value node2.next.

**Line 8** UC-KLEE forks execution and adds the path constraint node2.next = null to Path C.

**Line 12** Path C returns node1.val + node2.val and exits.

This example illustrates a simple but powerful recursive technique to automatically synthesize data structures from under-constrained symbolic input. Figure 2 shows an actual data structure our tool generated as input for one of the BIND bugs we discovered (Figure 5). The edges between each object are labeled with the field names contained in the function’s debug information and included in UC-KLEE’s error report.

### 2.2 Limitations

Because we build on our earlier version of UC-KLEE, we inherit its limitations [43]. The more important examples are as follows. The tool tests compiled code on a specific platform and does not consider other build configurations. It does not handle assembly (see §4 for how we skip inline assembly), nor symbolic floating point operations. In addition, there is an explicit assumption that input-derived pointers reference unique objects (no aliasing, and no cyclical data structures), and the tool assigns distinct concrete addresses to allocated objects.

When checking whether patches introduce bugs, UC-KLEE aims to detect crashing bugs and does not look for performance bugs, differences in system call arguments, or concurrency errors. We can only check patches that do not add, remove, or reorder fields in data structures or change the type signatures of patched functions. We plan to extend UC-KLEE to support such patches by implementing a type map that supplies identical inputs to each version of a function in a “field aware” manner.
ever, our current system does not support this, and we excluded such patches from our experiments.

3 Patch checking

To check whether a patch introduces new crashing bugs, UC-KLEE symbolically executes two compiled versions of a function: \( P \), the unpatched version, and \( P' \), the patched version. If it finds any execution paths along which \( P' \) crashes but \( P \) does not (when given the same symbolic inputs), it reports a potential bug in the patch.

Recall that due to missing input preconditions, we cannot simply assume that all crashes are bugs. Instead, UC-KLEE looks for paths that exhibit differing crash behavior between \( P \) and \( P' \), which usually share an identical set of preconditions. Even if UC-KLEE does not know these preconditions, in practice, real code tends to show error equivalence \cite{error_equivalence}, meaning that \( P \) and \( P' \) both crash (or neither crashes) on illegal inputs. For example, if a precondition requires a pointer to be non-null and both versions dereference the pointer, then \( P \) and \( P' \) will both crash when fed a null pointer as an argument.

In prior work, UC-KLEE \cite{error_equivalence} verified the equivalence of small library routines, both in terms of crashes and outputs. While detecting differences in functionality may point to interesting bugs, these discrepancies are typically meaningful only to developers of the checked code. Because this paper evaluates our framework on large, complex systems developed by third parties, we limit our discussion to crashes, which objectively point to bugs.

To check patches, UC-KLEE automatically generates a test harness that sets up the under-constrained inputs and invokes \( P \) and \( P' \). Figure 3 shows a representative test harness. Lines 2–3 create an under-constrained input \( n \). Line 4 calls \( \text{fooB}(P) \). Note that UC-KLEE invokes \( P' \) before \( P \) to facilitate path pruning (§ 3.1). Line 5 discards any writes performed by \( \text{fooB} \) but preserves the path constraints so that \( \text{fooA}(P) \) will see the same initial memory contents and follow the corresponding path. Line 6 invokes \( \text{fooA} \).

If a path through \( \text{fooB} \) crashes, UC-KLEE unwinds the stack and resumes execution at line 5. If \( \text{fooA} \) also crashes on this path, then the two functions are crash equivalent and no error is reported. However, if \( \text{fooA} \) returns from line 6 without crashing, we report an error to the user as a possible bug in \( \text{fooB} \). For this use case, we do not report errors in which \( \text{fooA}(P) \) crashes but \( \text{fooB}(P') \) does not, which suggest bugs fixed by a patch.

3.1 Path pruning

UC-KLEE employs several path pruning techniques to target errors and avoid uninteresting paths. The underlying UC-KLEE system includes a static cross-checker that walks over the LLVM \cite{llvm} control flow graph, conservatively marking regions of basic blocks that differ between the original function \( P \) and the patched function \( P' \). This algorithm is fairly straightforward, and we elide details for brevity. UC-KLEE soundly prunes paths that:

1. have never executed a “differing” basic block, and
2. cannot reach a differing basic block from their current program counter and call stack.

The second condition uses an inter-procedural reachability analysis from the baseline UC-KLEE system. Paths meeting both of these criteria are safe to prune because they will execute identical instruction sequences.

In addition, UC-KLEE introduces pruning techniques aimed specifically at detecting errors introduced by a patch. As our system executes \( P' \) (see Figure 3), it prunes paths that either:

1. return from \( P' \) without triggering an error, or
2. trigger an error without reaching differing blocks.

In the first case, we are only concerned with errors introduced by the patch. In the second case, \( P \) and \( P' \) would both trigger the error.

Error uniquing. Our system aggressively uniques errors by associating each path executing \( P \) with the program counter (PC) of the error that occurred in \( P' \). Once our system executes a non-error path that returns from \( P \) (and reports the error in \( P' \)), it prunes all current and future paths that hit the same error (PC and type) in \( P' \). In practice, this enabled our system to prune thousands of redundant error paths.

3.2 Evaluation

We evaluated UC-KLEE on hundreds of patches from BIND and OpenSSL, two widely-used, security critical systems. Each codebase contains about 400,000 lines of C code, making them reasonable measures of UC-KLEE’s scalability and robustness. For this experiment, we used a maximum symbolic object size of 25,000 bytes and a maximum symbolic data structure depth of 9 objects.

3.2.1 Patch selection and code modifications

We tried to avoid selection bias by using two complete sets of patches from the git repositories for recent stable branches: BIND 9.9 from 1/2013 to 3/2014 and OpenSSL 1.0.1 from 1/2012 to 4/2014. Many of the patches we encountered modified more than one function; this section uses patch to refer to changes to a single function, and commit to refer to a complete changeseet.

We excluded all patches that: only changed copyright information, had build errors, modified build infrastructure only, removed dead functions only, applied only to disabled features (e.g., win32), patched only BIND contrib features, only touched regression/unit tests, or used variadic functions. We also eliminated all patches
<table>
<thead>
<tr>
<th>Codebase</th>
<th>Function</th>
<th>Type</th>
<th>Cause</th>
<th>New</th>
<th>Vulnerability</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIND</td>
<td>receive_secure_db</td>
<td>assert fail</td>
<td>double lock acquisition</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>BIND</td>
<td>save_nsec3param</td>
<td>assert fail</td>
<td>uninitialized struct</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>BIND</td>
<td>configure_zone_acl</td>
<td>assert fail</td>
<td>inconsistent null argument handling</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>BIND</td>
<td>isc_idetach</td>
<td>assert fail</td>
<td>input parsing logic</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>OpenSSL</td>
<td>PROXY_PROPP2_BIND</td>
<td>uninitialized pointer dereference</td>
<td>inconsistent null check</td>
<td>✓</td>
<td>CVE-2015-0292</td>
</tr>
<tr>
<td>OpenSSL</td>
<td>dtls1_process_record</td>
<td>uninitialized pointer dereference</td>
<td>unchecked return value</td>
<td>✓</td>
<td>CVE-2014-0198</td>
</tr>
<tr>
<td>OpenSSL</td>
<td>dtls1_final_finish_mac</td>
<td>null pointer dereference</td>
<td>call err side effect after null check</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>OpenSSL</td>
<td>do_ssl3_write</td>
<td>null pointer dereference</td>
<td>unchecked return value</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>OpenSSL</td>
<td>PKCS7_datadecode</td>
<td>out-of-bounds array access</td>
<td>negative count passed to memcp y</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>OpenSSL</td>
<td>EVP_DecodeUpdate</td>
<td>use-after-free</td>
<td>improper error handling</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>OpenSSL</td>
<td>dtls1_buffer_record</td>
<td>use-after-free</td>
<td>improper error handling</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>OpenSSL</td>
<td>pkey_db</td>
<td>use-after-free</td>
<td>improper error handling</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: Summary of bugs UC-KLEE reported while checking patches. New indicates that the bug was previously unknown.

that yielded identical code after compiler optimizations. Because of tool limitations, we excluded patches that changed input datatypes (§2.2). Finally, to avoid inflating our verification numbers, we excluded three BIND commits that patched 200-300 functions each by changing a pervasive linked-list macro and/or replacing all uses of memcp y with memmove. Neither of these changes introduced any errors and, given their near-trivial modifications, shed little additional light on our tool’s effectiveness. This yielded 487 patches from BIND and 324 patches from OpenSSL, both from 177 distinct commits to BIND and OpenSSL (purely by coincidence).

We compiled patched and unpatched versions of the codebase for each revision using an LLVM 2.7 toolchain. We then ran UC-KLEE over each patch for one hour. Each run was allocated a single Intel Xeon E5645 2.4GHz core and 4GB of memory on a compute cluster running 64-bit Fedora Linux 14. For these runs, we configured UC-KLEE to target crashes only in patched routines or routines they call. While this approach allows UC-KLEE to focus on the most likely source of errors, it does not detect bugs caused by the outputs of a function, which may trigger crashes elsewhere in the system (e.g., if the function unexpectedly returns null). UC-KLEE can report such differences, but we elide that feature in this paper.

**Code modifications.** In BIND and OpenSSL, we canonicalized several macros that introduced spurious code differences such as the __LINE__, VERSION, SRCID, DATE, and OPENSSL_VERSION_NUMBER macros. To support function-call annotations (§3.2.5) in BIND, we converted four preprocessor macros to function calls.

For BIND, we disabled expensive assertion-logging code and much of its debug malloc functionality, which UC-KLEE already provided. For OpenSSL, we added a new build target that disabled reference counting and address alignment. The reference counting caused many false positives; UC-KLEE reported double free errors due to unknown preconditions on an object’s reference count.

### 3.2.2 Bugs found

From the patches we tested, UC-KLEE uncovered three previously unknown bugs in BIND and eight bugs in OpenSSL, six of which were previously unknown. These bugs are summarized in Figure 4.

Figure 5 shows a representative double-lock bug in BIND found by cross-checking. The patch moved the LOCK ZONE earlier in the function (line 1), causing existing error handling code that jumped to failure (line 9) to bypass the UNLOCK ZONE (line 12). In this case, the subsequent call to dns_zone_idetach (line 14) reacquires the already-held lock, which triggers an assertion failure. This bug was one of several we found that involved infrequently-executed error handling code. Worse, BIND often hides goto failure statements inside a CHECK macro, which was responsible for a bug we discovered in the save_nsec3param function (not shown). We reported the bugs to the BIND developers, who promptly confirmed and fixed them. These examples demonstrate a key benefit of UC-KLEE: it explores non-obvious execution paths that would likely be missed by a human developer, either because the code is obfuscated or an error condition is overlooked.

UC-KLEE is not limited to finding new bugs introduced by the patches; it can also find old bugs in patched code. We added a new mode where UC-KLEE flags errors that occur in both P and P’ if the error must occur for all input values following that execution path (must-fail error described in §3.2.5). This approach allowed us to find one new bug in BIND and four in OpenSSL. It also re-confirmed a number of bugs found by cross-checking above. This mode could be used to find bugs in functions that have not been patched, but we did not use it for that purpose in this paper.

Figure 6 shows a representative must-fail bug, a previously unknown null pointer dereference (denial-of-service) vulnerability we discovered in OpenSSL’s...
If the pointer is null, explicitly checking whether a developer attempted to prevent this bug by explicitly CVE-2014-0198 [12] being issued. In this case, works correctly when an alert is pending a developer may write tests to make sure this function illustrates the weaknesses of regression testing. While 3.2.3 Patches verified

In addition to finding new bugs, UC-KLEE exhaustively verified all execution paths for 67 (13.8%) of the patches in BIND, and 48 (14.8%) of the patches in OpenSSL. Our system effectively verified that, up to the given input bound and with the usual caveats, these patches did not introduce any new crashes. This strong result is not possible with imprecise static analysis or testing.

The median instruction coverage (§ 3.2.4) for the exhaustively verified patches was 90.6% for BIND and 100% for OpenSSL, suggesting that these patches were thoroughly tested. Only six of the patches in BIND and one in OpenSSL achieved very low (0-2%) coverage. We determined that UC-KLEE achieved low coverage on these patches due to dead code (2 patches); an insuffi-

Figure 6: OpenSSL null pointer bug in do_ssl3_write.

do_ssl3_write function that led to security advisory CVE-2014-0198 [12] being issued. In this case, a developer attempted to prevent this bug by explicitly checking whether wb->buf is null (line 1). If the pointer is null, ssl3_setup_write_buffers allocates a new buffer (line 2). On line 6, the code then handles any pending alerts [20] by calling ssl_dispatch_alert (line 8). This call has the subtle side effect of freeing the write buffer when the common SSL_MODE_RELEASE_BUFFERS flag is set. After freeing the buffer, wb->buf is set to null (not shown), triggering a null pointer dereference on line 15.

This bug would be hard to find with other approaches. The write buffer is freed by a chain of function calls that includes a recursive call to do_ssl3_write, which one maintainer described as “sneaky” [44]. In contrast to static techniques that could not reason precisely about the recursion, UC-KLEE proved that under the circumstances when both an alert is pending and the release flag is set, a null pointer dereference will occur. This example also illustrates the weaknesses of regression testing. While a developer may write tests to make sure this function works correctly when an alert is pending or when the release flag is set, it is unlikely that a test would exercise these conditions simultaneously. Perhaps as a direct consequence, this vulnerability was nearly six years old.

3.2.4 Patches partially verified

This section measures how thoroughly we check non-terminating patches using two metrics: (1) instruction coverage, and (2) number of execution paths completed.

We conservatively measure instruction coverage by counting the number of instructions that differ in $P'$ from $P$ and then computing the percentage of these instructions that UC-KLEE executes at least once. Figure 7 plots the instruction coverage. The median coverage was 81.1% for BIND, 86.9% for OpenSSL. Efficient symbolic input bound (2 patches); comparisons between input pointers (we assume no aliasing, 1 patch); symbolic malloc size (1 patch); and a trivial stub function that was optimized away (1 patch).

Figure 7: Coverage of patched instructions: 100% coverage for 98 BIND patches (20.1%) and 124 OpenSSL patches (38.3%). Median was 81.1% for BIND, 86.9% for OpenSSL.

Figure 8 plots the number of completed execution paths for each patch we did not exhaustively verify (§ 3.2.3) that hit at least one patched instruction. These graphs exclude 31 patches for BIND and 32 patches for OpenSSL for which our system crashed during the one hour execution window. The crashes were primarily due to bugs in our tool and memory exhaustion/blowup caused by symbolically executing cryptographic ciphers.

For the remaining patches, UC-KLEE completed a median of 5,828 distinct paths per patch for BIND and 1,412 for OpenSSL. At the upper end, 154 patches for BIND (39.6%) and 79 for OpenSSL (32.4%) completed over 10,000 distinct execution paths. At the bottom end, 58 patches for BIND (14.9%) and 46 for OpenSSL (18.9%) completed zero execution paths. In many cases, UC-KLEE achieved high coverage on these patches but neither detected errors nor ran the non-error paths to com-
false positives we encountered bugs from standard practice of running the patch on a few tests. 

The independence between the paths explored (greater is preferable). On the other hand, they grossly under-count the number of distinct concrete values each symbolic path reasons about simultaneously. One would generally expect that exercising 1,000 or more paths through a patch, where each path simultaneously tests all feasible values, represents a dramatic step beyond the current standard practice of running the patch on a few tests.

3.2.5 False positives

This section describes our experience in separating true bugs from false positives, which were due to missing input preconditions. The false positives we encountered were largely due to three types of missing preconditions:

1. Data structure invariants, which apply to all instances of a data structure (e.g., a parent node in a binary search tree has a greater value than its left child).
2. State machine invariants, which determine the sequence of allowed values and the variable assignments that may exist simultaneously (e.g., a counter increases monotonically).
3. API invariants, which determine the legal inputs to API entry points (e.g., a caller must not pass a null pointer as an argument).

Figure 9 illustrates a representative example of a false positive from BIND, which was caused by a missing data structure invariant. The **isc_region_t** type consists of a buffer and a length, but UC-KLEE has no knowledge that the two are related. The code selects a valid buffer length at line 14, the shorter of the two buffers. At line 17, the code calls **memcmp** and supplies this length. Inside **memcmp**, UC-KLEE reported hundreds of false positives involving out-of-bounds memory reads. These errors occurred on false paths where the buffer pointed to by the **base** field was smaller than the associated **length** field.

**Manual annotations.** UC-KLEE supports two types of manual annotations: (1) data type annotations, and (2) function call annotations. Both are written in C and compiled with LLVM. UC-KLEE invokes data type annotations at the end of a path, prior to emitting an error. These are associated with named data types and specify invariants on symbolic inputs of that type (inferred from debug information when available). For the example above, we added the following simple annotation for the **isc_region_t** data type:

```
INARIANT((r->length < OBJECT_SIZE(r->base)));
```

The **INARIANT** macro requires that the condition hold. If it is infeasible (cannot be true) on the current path, UC-KLEE emits an error report with a flag indicating that the annotations have been violated. We use this flag to filter out uninteresting error reports. This one simple annotation allowed us to filter 623 errors, which represented about 7.5% of all the errors UC-KLEE reported for BIND.

Function call annotations are used to run specific code immediately prior to calling a function. For example, we wrote a function call annotation for BIND that runs before each call to **isc_mutex_lock**, with the same arguments:

```
void annot_isc_mutex_lock(isc_mutex_t *mp) {
  EXPECT(*mp == 0);
}
```


**Table 1: Heuristic Effects**

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>Reports</th>
<th>Patches</th>
<th>Reports</th>
<th>Patches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tot.</td>
<td>Bugs</td>
<td>Tot.</td>
<td>Bugs</td>
</tr>
<tr>
<td>Manual annotations</td>
<td>1419</td>
<td>3</td>
<td>1378</td>
<td>153</td>
</tr>
<tr>
<td>must-fail</td>
<td>44</td>
<td>8</td>
<td>1378</td>
<td>153</td>
</tr>
<tr>
<td>concrete-fail</td>
<td>26*</td>
<td>2</td>
<td>878</td>
<td>110</td>
</tr>
<tr>
<td>belief-fail</td>
<td>35*</td>
<td>3</td>
<td>1053</td>
<td>127</td>
</tr>
<tr>
<td>excluding inputs</td>
<td>30*</td>
<td>3</td>
<td>852</td>
<td>102</td>
</tr>
<tr>
<td>True bugs</td>
<td>3*</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

(a) BIND (487 patches, 4 distinct bugs)

**Figure 10:** Effects of heuristics on false positives. *Tot.* indicates the total number of reports, of which *Bugs* are true errors; *Patches* indicates the number of patches that reported at least one error. *P only* refers to errors that occurred only in function *P*; *P and P* indicates that we reviewed all the reports manually.

**Table 2: C annotation macros.**

<table>
<thead>
<tr>
<th>Macro</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>INVARIANT(condition)</td>
<td>Add condition as a path constraint; kill path if infeasible.</td>
</tr>
<tr>
<td>EXPECT(condition)</td>
<td>Add condition as a path constraint if feasible; otherwise, ignore.</td>
</tr>
<tr>
<td>IMPLIES(a, b)</td>
<td>Logical implication: a → b.</td>
</tr>
<tr>
<td>HELDS(a)</td>
<td>Returns true if condition a must hold.</td>
</tr>
<tr>
<td>MAY_HOLD(a)</td>
<td>Returns true if condition a may hold.</td>
</tr>
<tr>
<td>STRIDE(e)</td>
<td>Forces e to be evaluated; prevents compiler from optimizing it away.</td>
</tr>
<tr>
<td>VALID_POINTER(ptr)</td>
<td>Returns true if ptr is valid; false otherwise.</td>
</tr>
<tr>
<td>OBJECT_SIZE(ptr)</td>
<td>Returns the size of the object pointed to by ptr; kills path if pointer is invalid.</td>
</tr>
</tbody>
</table>

**Figure 11:** The manual annotations suppressed 42% of the reports for BIND but only 9.6% for OpenSSL. We attribute this difference to the greater effort we expended writing manual annotations for BIND, for which the automated heuristics were less effective without the annotations.

**Automated heuristics.** We tried numerous heuristics to reduce false reports. UC-KLEE augments each error report with a list of the heuristics that apply. The *must-fail* heuristic identifies errors that must occur for all input values following that execution path, since these are often true errors [18]. For example, assertion failures are must-fail when the condition must be false.

A variation on the must-fail heuristic is the belief-fail heuristic, which uses a form of belief analysis [19]. The intuition behind this heuristic is that if a function *contradicts* itself, it likely has a bug. For example, if the code checks that a pointer is null and then dereferences the pointer, it has a bug, regardless of any input preconditions. On the other hand, a function is generally agnostic to the assumptions made by the functions it calls. For example, if `strcmp` checks whether two strings have the same address, the caller does not acquire this belief, even if the path constraints now indicate that the two addresses match. Following this intuition, the belief-fail heuristic identifies errors that occur for all input values satisfying the belief set, which is the set of constraints (i.e., branch conditions) added within the current function or inherited from its caller, but not its callees. We track belief sets for each stack frame.
A second variation on must-fail is **concrete-fail**, which indicates that an assertion failure or memory error was triggered by a concrete (non-symbolic) condition or pointer, respectively. In practice, this heuristic and belief-fail were the most effective.

These heuristics reduced the total number of reports to a small enough number that we were able to inspect them all manually. While only 8.6% of the belief-fail errors for BIND and 20% of those for OpenSSL were true bugs, the total number of these errors (60) was manageable relative to the number of patches we tested (811). In total, the annotations and belief-fail heuristic eliminated 98.6% of false positives for BIND and 98.2% for OpenSSL.

A subset of the belief-fail errors were caused by reading past the end of an input buffer, and none of these were true bugs. Instead, they were due to paths reaching the input bound we specified. In many cases, our system would emit these errors for any input bound because they involved unbounded loops (e.g., `strlen`). The **excluding inputs** row in Figure 10 describes the subset of belief-fail errors not related to input buffers. This additional filter produced a small enough set of $P$ and $P'$ errors for OpenSSL that we were able to manually inspect them, discovering a number of additional bugs. We note that the **true errors** listed in Figure 10 constitute 12 distinct bugs; some bugs showed up in multiple error reports.

## 4 Generalized checking

In addition to checking patches, UC-KLEE provides an interface for rule-based checkers to be invoked during symbolic path exploration. These checkers are similar to tools built using dynamic instrumentation systems such as Valgrind [34] or Pin [30]. Unlike these frameworks, however, UC-KLEE applies its checkers to all possible paths through a function, not to a single execution path through a program. In addition, UC-KLEE considers all possible input values along each path, allowing it to discover bugs that might be missed when checking a single set of concrete inputs.

Conceptually, our framework is similar to **WOODPECKER** [8], a KLEE-based tool that allows system-specific checkers to run on top of (whole program) symbolic execution. In this paper, however, we focus on generic checkers we implemented for rules that apply to many systems, and we directly invoke these checkers on individual functions deep within each codebase.

UC-KLEE provides a simple interface for implementing checkers by deriving from a provided C++ base class. This interface provides hooks for a checker to intercept memory accesses, arithmetic operations, branches, and several types of errors UC-KLEE detects.

A user invoking UC-KLEE provides a compiled LLVM module and the name of a function to check. We refer to this function as the **top-level** function. Generally, the module has been linked to include all functions that might be called by the top-level function. When UC-KLEE encounters a function call, it executes the called function. When UC-KLEE encounters a call to a function missing from the LLVM module, however, it may optionally skip over the function call rather than terminate the path with an error message. When UC-KLEE skips a function call, it creates a new under-constrained value to represent the function’s return value, but it leaves the function’s arguments unchanged. This approach **under-approximates** the behaviors that the missing function might perform (e.g., writing to its arguments or globals). Consequently, UC-KLEE may miss bugs and cannot provide verification guarantees when functions are missing.

We briefly experimented with an alternative approach in which we overwrote the skipped function’s arguments with new under-constrained values, but this over-approximation caused significant path explosion, mostly involving paths that could not arise in practice.

In addition to missing functions due to scalability limitations, we also encountered inline assembly (Linux kernel only) and unresolved symbolic function pointers. We skipped these two cases in the same manner as missing functions. For all three cases, UC-KLEE provides a hook to allow a checker to detect when a call is being skipped and to take appropriate actions for that checker.

In the remainder of this section, we describe each checker, followed by our experimental results in § 4.4.

### 4.1 Leak checker

Memory leaks can lead to memory exhaustion and pose a serious problem for long-running servers. Frequently, they are exploitable as denial-of-service vulnerabilities [10, 13, 14]. To detect memory leaks (which may or may not be remotely exploitable, depending on their location within a program), we implemented a leak checker on top of UC-KLEE. The leak checker considers a heap object to be leaked if, after returning from the top-level function, the object is not reachable from a root set of pointers. The root set consists of a function’s (symbolic) arguments, its return value, and all global variables. This checker is similar to the leak detection in Purify [23] or Valgrind’s memcheck [34] tool, but it thoroughly checks all paths through a specific function, rather than a single concrete path through a whole program.

When UC-KLEE encounters a missing function, the leak checker finds the set of heap objects that are reachable from each of the function call’s arguments using a precise approach based on pointer referents [42, 43]. It then marks these objects as **possibly escaping**, since the missing function could capture pointers to these objects and prevent them from becoming unreachable. At the end of each execution path, the leak checker removes any possibly escaping objects from the set of leaked objects.
Doing so allows it to report only true memory leaks, at the cost of possibly omitting leaks when functions are missing. However, UC-KLEE may still report false leaks along invalid execution paths due to missing input preconditions. Consider the following code fragment:

```c
1: char* leaker() { 
2:   char *a = (char*) malloc(10); /* not leaked */
3:   char *b = (char*) malloc(10); /* maybe leaked */
4:   char *c = (char*) malloc(10); /* leaked! */
5:   bar(b); /* skipped call to bar */
6:   return a;
7: }
```

When UC-KLEE returns from the function leaker, it inspects the heap and finds three allocated objects: a, b, and c. It then examines the root set of objects. In this example, there are no global variables and leaker has no arguments, so the root set consists only of leaker’s return value. UC-KLEE examines this return value and finds that the pointer a is live (and therefore not leaked). However, neither b nor c is reachable. It then looks at its list of possibly escaping pointers due to the skipped call to bar on line 6, which includes b. UC-KLEE subtracts b from the set of leaked objects and reports back to the user that c has been leaked. While this example is trivial, UC-KLEE discovered 37 non-trivial memory leak bugs in BIND, OpenSSL, and the Linux kernel (§ 4.4).

### 4.2 Uninitialized data checker

Functions that access uninitialized data from the stack or heap exhibit undefined or non-deterministic behavior and are particularly difficult to debug. Additionally, the prior contents of the stack or heap may hold sensitive information, so code that operates on these values may be vulnerable to a loss of confidentiality.

UC-KLEE includes a checker that detects accesses to uninitialized data. When a function allocates stack or heap memory, the checker fills it with special garbage values. The checker then intercepts all loads, binary operations, branches, and pointer dereferences to check whether any of the operands (or the result of a load) contain garbage values. If so, it reports an error to the user.

In practice, loads of uninitialized data are often intentional; they frequently arise within calls to memcpy or when code manipulates bit fields within a C struct. Our evaluation in § 4.4 therefore focuses on branches and dereferences of uninitialized pointers.

When a call to a missing function is skipped, the uninitialized data checker sanitizes the function’s arguments to avoid reporting spurious errors in cases where missing functions write to their arguments.

### 4.3 User input checker

Code that handles untrusted user input is particularly prone to bugs that lead to security vulnerabilities since an attacker can supply any possible input value to exploit the code. Generally, UC-KLEE treats inputs to a function as under-constrained because they may have unknown preconditions. For cases where inputs originate from untrusted sources such as network packets or user-space data passed to the kernel, however, the inputs can be considered fully-constrained. This term indicates that the set of legal input values is known to UC-KLEE; in this case, any possible input value may be supplied. If any value triggers an error in the code, then the error is likely to be exploitable by an attacker, assuming that the execution path is feasible (does not violate other preconditions).

UC-KLEE maintains shadow memory (metadata) associated with each symbolic input that tracks whether each symbolic byte is under-constrained or fully-constrained. UC-KLEE includes a system-configurable user input checker that intercepts all errors and adds an UNSAFE_INPUT flag to errors caused by fully-constrained inputs. For memory access errors, the checker examines the pointer to see if it contains fully-constrained symbolic values. For assertion failures, it examines the assertion condition. For division-by-zero errors, it examines the divisor.

In all cases, the checker inspects the fully-constrained inputs responsible for an error and determines whether any path constraints compare the inputs to under-constrained data (originating elsewhere in the program). If so, the checker assumes that the constraints may properly sanitize the input, and it suppresses the error. Otherwise, it emits the error. This approach avoids reporting spurious errors to the user, at the cost of missing errors when inputs are partially (but insufficiently) sanitized.

We designed this checker primarily to find security vulnerabilities similar to the OpenSSL “Heartbleed” vulnerability [1, 11] from 2014, which passed an untrusted and unsanitized length argument to memcpy, triggering a severe loss of confidentiality. In that case, the code never attempted to sanitize the length argument. To test this checker, we ran UC-KLEE on an old version of OpenSSL without the fix for this bug and confirmed that our checker reports the error.

### 4.4 Evaluation

We evaluated UC-KLEE’s checkers on over 20,000 functions from BIND, OpenSSL, and the Linux kernel. For BIND and OpenSSL, we used UC-KLEE to check all functions except those in the codebases’ test directories. We used the same minor code modifications described in § 3.2.1, and we again used a maximum input...
Figure 12: Summary of results from running UC-KLEE checkers on functions from each codebase. Bugs shows the number of distinct true bugs found (67 total). Reports shows the total number of errors reported by UC-KLEE in each category (multiple errors may point to a single bug). False reports the percentage of errors reported that did not appear to be true bugs (i.e., false positives). ¹denotes that we inspected only a handful of errors for that category.

Figure 13: Linux kernel memory leak in RPCSEC GSS protocol implementation used by NFS server-side AUTH GSS.

Figure 14: Functions verified (with caveats) by UC-KLEE.

4.4.1 Leak checker

The leak checker was the most effective. It reported the greatest number of bugs (37 total) and the lowest false positive rate. Interestingly, only three of the 138 leak reports for BIND were spurious errors, a false positive rate of only 2.2%. For OpenSSL, we excluded 269 additional reports involving the library’s obfuscated ASN.1 [25] parsing code, which we could not understand. Of the remaining 272 reports, the checker found five bugs but had a high false positive rate of 90.1%.

For the Linux kernel, we wrote simple C annotations (about 60 lines) to intercept calls to mmalloc, vmalloc, kfree, vfree, and several similar functions, and to forward these to UC-KLEE’s built-in mmalloc and vfree functions. Doing so allowed us to track memory management without the overhead of symbolically executing the kernel’s internal allocators. We then ran UC-KLEE on all functions that directly call these allocation functions.

Our system discovered 23 memory leaks in the Linux kernel. One particularly interesting example (Figure 13) involved the SunRPC layer’s server-side implementation of AUTH GSS authentication for NFS. Each connection triggering an upcall causes 512 bytes allocated at line 10 to be leaked due to a missing kfree that should be present around line 17. Since this leak may be triggered by remote connections, it poses a potential denial-of-service (memory exhaustion) vulnerability. The NFS maintainers accepted our patch to fix the bug.

UC-KLEE found that at least 2909 functions in BIND and at least 3700 functions in OpenSSL (or functions they call) allocate heap memory. As shown in Figure 14, UC-KLEE verified (with caveats) that 388 functions in BIND and 383 in OpenSSL allocate heap memory but do not leak it. Our system also verified that 1776 functions in BIND and 1648 functions in OpenSSL do not allocate heap memory, making them trivially leak-free.

4.4.2 Uninitialized data checker

The uninitialized data checker reported a total of 19 new bugs. One illustrative example, shown in Figure 15, in-
volves OpenSSL’s elliptic curve cryptography. If the call to EC_POINT_new on line 4 fails, the code jumps to line 8, leaving the points array partially uninitialized. Line 13 then passes uninitialized pointers from the array to EC_POINT_free, which dereferences the pointers and passes them to free, potentially corrupting the heap. This is one of many bugs that we found involving infrequently executed error-handling code, a common source of security bugs.

UC-KLEE discovered an interesting bug (Figure 16) in BIND’s UDP port randomization fix for Kaminsky’s cache poisoning attack [9]. To prevent spoofed DNS replies, BIND must use unpredictable source port numbers. The dispatch_createudp function calls the get_udpsocket function at line 9, which selects a pseudorandom number generator (PRNG) at line 18 based on whether we are using a UDP or TCP connection. However, the socktype field isn’t initialized in dispatch_createudp until line 12, meaning that the PRNG selection is based on uninitialized data. While it appears that the resulting port numbers are sufficiently unpredictable despite this bug, this example illustrates UC-KLEE’s ability to find errors with potentially serious security implications.

For the Linux kernel, we checked the union of the functions we used for the leak checker and the user input checker (discussed below) and found 10 bugs.

Due to time limitations, we exhaustively inspected only the most serious category of errors: uninitialized pointers. The checker reported too many uninitialized branches for us to examine completely, but we did inspect a few dozen of these errors in an ad-hoc manner. All three of the bugs from BIND and one bug from the Linux kernel fell into this category. The remaining bugs were uninitialized pointer errors. We did not inspect the error reports for binary operations or load values.

Finally, our system verified (with caveats) that about a third of the functions from BIND (2045) and OpenSSL (2043) do not access uninitialized data. We believe that providing this level of guarantee on such a high percent-

Figure 15: OpenSSL dereference/free of uninitialized pointer in ec_vNAF_precompute_mult function.

Figure 16: BIND non-deterministic PRNG selection bug.
5.1 Object sizing

Recall that when an unbound symbolic pointer is dereferenced, UC-KLEE must allocate memory and bind the pointer to it. One challenge in implementing this functionality is picking a useful object size to allocate. If the size is too small, later accesses to this object may trigger out-of-bounds memory errors. On the other hand, a size that is too large can hide legitimate errors. We handled this tradeoff using two approaches.

5.2 Error reporting

With whole program symbolic execution, symbolic inputs typically represent unstructured strings or byte arrays from command line arguments or file contents. In this case, an error report typically contains a single set of concrete inputs that trigger the error, along with a backtrace. With under-constrained symbolic execution, however, the inputs are often complex, pointer-rich data structures since UC-KLEE directly executes individual functions within a program. In this case, a single set of concrete values is not easily understood by a user, nor can it be used to trivially reproduce the error outside of UC-KLEE because pointer inputs expect memory objects (i.e., stack, heap, and globals) to be located at specific addresses.

To provide more comprehensible error reports, UC-KLEE addresses.
KLEE emits a path summary for each error. The path summary provides a complete listing of the source code executed along the path, along with the path constraints added by each line of source. The path constraints are expressed in a C-like notation and use the available LLVM debug information to determine the types and names of each field. Below we list example constraints that UC-KLEE included with error reports for BIND (§ 3.2):

```
int x = y / z;
if (x > 10) /* query: y / z > 10 */
...  
```

(a) Eager constraints (standard)

```
int x = lazy_x; /* adds lazy constraint: lazy_x = y / z */
if (x > 10) /* query: lazy_x > 10 */
...  
```

(b) Lazy constraints

Figure 19: Lazy constraint used for integer division operation.

5.3 General KLEE optimizations

We added several scalability improvements to UC-KLEE that apply more broadly to symbolic execution tools. To reduce path explosion in library functions such as strlen, we implemented special versions that avoid forking paths by using symbolic if-then-else constructs. We also introduced scores of rules to simplify symbolic expressions [42]. We elide further details due to space.

5.3.1 Lazy constraints

During our experiments, we faced query timeouts and low coverage for several benchmarks that we traced to symbolic division and remainder operations. The worst cases occurred when an unsigned remainder operation had a symbolic value in the denominator. To address this challenge, we implemented a solution we refer to as lazy constraints. Here, we defer evaluation of expensive queries until we find an error. In the common case where an error does not occur or two functions exhibit crash equivalence along a path, our tool avoids ever issuing potentially expensive queries. When an error is detected, the tool re-checks that the error path is feasible (otherwise the error is invalid).

Figure 19(a) shows a simple example. With eager constraints (the standard approach), the if-statement at line 2 triggers an SMT query involving the symbolic integer division operation \( y / z \). This query may be expensive, depending on the other path constraints imposed on \( y \) and \( z \). To avoid a potential query timeout, UC-KLEE introduces a lazy constraint (Figure 19(b)). On line 1, it replaces the result of the integer division operation with a new, unconstrained symbolic value \( \text{lazy}_x \) and adds the lazy constraint \( \text{lazy}_x = y / z \) to the current path. At line 2, the resulting SMT query is the trivial expression \( \text{lazy}_x > 10 \). Because \( \text{lazy}_x \) is unconstrained, UC-KLEE will take both the true and false branches following the if-statement. One of these branches may violate the constraints imposed on \( y \) and \( z \), so UC-KLEE must check that the lazy constraints are consistent with the full set of path constraints prior to emitting any errors.

6 Related work

This paper builds on prior work in symbolic execution [4], particularly KLEE [5] and our early work on UC-KLEE [43]. Unlike our previous work, which targeted small library routines, this paper targets large systems.
and supports generalized checking.

Other recent work has used symbolic execution to check patches. DiSE [39] performs whole program symbolic execution but prunes paths unaffected by a patch. Differential Symbolic Execution (DSE) [38] and regression verification [21] use abstraction to achieve scalability but may report false differences. By contrast, our approach soundly executes complete paths through each patched function, eliminating this source of false positives. Impact Summaries [2] complement our approach by soundly pruning paths and ignoring constraints unaffected by a patch.

SymDiff [27] provides a scalable solution to check the equivalence of two programs with fixed loop unrolling but relies on imprecise, uninterpreted functions. Differential assertion checking (DAC) [28] is the closest to our work and applies SymDiff to the problem of detecting whether properties that hold in $P$ also hold in $P'$, a generalization of crash equivalence. However, DAC suffers from the imprecisions of SymDiff and reports false differences when function calls are reordered by a patch. Abstract semantic differencing [37] achieves scalability through clever abstraction but, as with SymDiff, suffers additional false positives due to over-approximation.

Recent work has used symbolic execution to generate regression tests exercising the code changed by a patch [41, 31, 32]. While they can achieve high coverage, these approaches use existing regression tests as a starting point and greedily redirect symbolic branch decisions toward a patch, exploring only a small set of execution paths. By contrast, our technique considers all possible intermediate program values as input (with caveats).

Dynamic instrumentation frameworks such as Valgrind [34] and PIN [30] provide a flexible interface for checkers to examine a program’s execution at runtime and flag errors. However, these tools instrument a single execution path running with concrete inputs, making them only as effective as the test that supplies the inputs.

Similar to our use of generalized checking in UC-KLEE is WOODPECKER [8], which uses symbolic execution to check system-specific rules. Unlike UC-KLEE, WOODPECKER applies to whole programs, so we expect it would not scale well to large systems. However, WOODPECKER aggressively prunes execution paths that are redundant with respect to individual checkers, a technique that would be useful in UC-KLEE.

Prior work in memory leak detection has used static analysis [45], dynamic profiling [24], and binary rewriting [23]. Dynamic tools such as Purify [23] and Valgrind [34] detect a variety of memory errors at runtime, including uses of uninitialized data. CCured [33] uses a combination of static analysis and runtime checks to detect pointer errors. Our user input checker relates to prior work in dynamic taint analysis, including TaintCheck [35] and Dytan [7].

7 Conclusions and future work

We have presented UC-KLEE, a novel framework for validating patches and applying checkers to individual C/C++ functions using under-constrained symbolic execution. We evaluated our tool on large-scale systems code from BIND, OpenSSL, and the Linux kernel, and we found a total of 79 bugs, including two OpenSSL denial-of-service vulnerabilities.

One avenue for future work is to employ UC-KLEE as a tool for finding general bugs (e.g., out-of-bounds memory accesses) in a single version of a function, rather than cross-checking two functions or using specialized checkers. Our preliminary experiments have shown that this use case results in a much higher rate of false positives, but we did find a number of interesting bugs, including the OpenSSL denial-of-service attack for which advisory CVE-2015-0291 [15, 22, 42] was issued.

In addition, we hope to further mitigate false positives by using ranking schemes to prioritize error reports, and by inferring invariants to reduce the need for manual annotations. In fact, many of the missing input preconditions can be thought of as consequences of a weak type system in C. We may target higher-level languages in the future, allowing our framework to assume many built-in invariants (e.g., that a length field corresponds to the size of an associated buffer).

Acknowledgements

The authors would like to thank Joseph Greathouse and the anonymous reviewers for their valuable feedback. In addition, the authors thank the LibreSSL developers for their quick responses to our bug reports, along with Evan Hunt and Sue Graves of ISC for granting us access to the BIND git repository before it became public. This work was supported by DARPA under agreements 1190029-276707 and N66001124088, by the United States Air Force Research Laboratory (AFRL) through contract FA8650-10-C-7024, and by a National Science Foundation Graduate Research Fellowship under grant number DGE-0645962. The views expressed in this paper are the authors’ own.

References

[5] Cadar, C., Dunbar, D., and Engler, D. KLEE: Unassisted and automatic generation of high-coverage tests for complex sys-


