Abstract—Security and privacy-sensitive smartphone applications use trusted execution environments (TEEs) to protect sensitive operations from malicious code. By design, TEEs have privileged access to the entire system but expose little to no insight into their inner workings. Moreover, real-world TEEs enforce strict format and protocol interactions when communicating with trusted applications (TAs), which prohibits effective automated testing.

TEEzz is the first TEE-aware fuzzing framework capable of effectively fuzzing TAs in situ on production smartphones, i.e., the TA runs in the encrypted and protected TEE and the fuzzer may only observe interactions with the TA but has no control over the TA’s code or data. Unlike traditional fuzzing techniques, which monitor the execution of a program being fuzzed and view its memory after a crash, TEEzz only requires a limited view of the target. TEEzz overcomes key limitations of TEE fuzzing (e.g., lack of visibility into the executed TAs, proprietary exchange formats, and value dependencies of interactions) by automatically attempting to infer the field types and message dependencies of the TA API through its interactions, designing state- and type-aware fuzzing mutators, and creating an in situ, on-device fuzzer.

Due to the limited availability of systematic fuzzing research for TAs on commercial-off-the-shelf (COTS) Android devices, we extensively examine existing solutions, explore their limitations, and demonstrate how TEEzz improves the state-of-the-art. First, we show that general-purpose kernel driver fuzzers are ineffective for fuzzing TAs. Then, we establish a baseline for fuzzing TAs using a ground-truth experiment. We show that TEEzz outperforms other blackbox fuzzers, can improve greybox approaches (if TAs source code is available), and even outperforms greybox approaches for stateful targets. We found 13 previously unknown bugs in the latest versions of OPTEE TAs in total, out of which TEEzz is the only fuzzer to trigger three. We also ran TEEzz on popular phones and found 40 unique bugs for which one CVE was assigned so far.

Index Terms—Fuzzing, Android, TEE, ARM TrustZone

I. INTRODUCTION

Smartphones operate on private user data and perform sensitive functionality, e.g., financial transactions [31], user authentication [76], or handling digital rights management (DRM) protected media [30]. To defend against various application- and kernel-level exploits [21], [70], [84] applications leverage TEEs [24] (e.g., ARM TrustZone (TZ) [7]) as an additional hardware-based defense. TEEs enforce the integrity and confidentiality of their applications. Partially due to recent research that demonstrated the usefulness of TEE applications [25], [56], [42], [50], [87], [89], called TAs, their number, as well as their complexity, is steadily increasing, leading to more TA-based vulnerabilities [15], [16], [34]. Unlike regular applications, where the vulnerability affects only the application, a vulnerability in a TA compromises the security of the entire system [88], potentially even the secure boot process [66].

While the security of these TAs is foundational to the security of the device, performing effective testing (e.g., fuzzing) remains an open challenge. Smartphones ship with the trusted OS (tOS) and numerous pre-installed TAs, prohibiting the normal world (e.g., Android) from inspecting their code at runtime. TA interactions are stateful and use complex proprietary message formats [39]. The entities in the secure world (TEE and TAs) are often encrypted and get decrypted in secure memory at runtime, prohibiting the use of static analysis-based vulnerability detection techniques. Dynamic analysis, i.e., fuzzing, is an effective alternative.

There are two principled approaches for fuzzing TAs: rehosting through emulation or on-device instrumentation.

Rehosting the TEE in an emulated environment overcomes the inaccessibility of the TEE’s internal state. PartEmu [39] rehosts Samsung’s proprietary TEE software stacks. They rehost the tOS and its TAs, to an emulated system-on-a-chip (SoC), gaining unrestricted access to the TEE’s internal state. Limitations to this approach are (1) the reverse engineering and implementation effort for emulated software and hardware components, (2) the inaccuracy of these implementations, (3) the lack of public data sheets, and (4) industry involvement leading to non-disclosure agreements for existing solutions. Especially the last limitation deserves further emphasis. PartEmu is the only existing rehosting solution targeting multiple TEEs. The prototype validates the feasibility of rehosting proprietary software stacks deployed on Samsung devices and is not publicly available.

The second approach, on-device fuzzing, mitigates these limitations and inaccuracies of emulation approaches. However, it lacks access to the TEE’s internal state and must fall back to blackbox fuzzing techniques. Unlike typical fuzzing techniques, which can analyze the binary, system memory, and executed instructions, an on-device TEE fuzzer must infer bugs from a far more limited view of the execution. Interactions with TAs happen through a vendor-provided interface (e.g., a driver [1], [3], [74]) in the rich OS (rOS), which ultimately generates an secure monitor call (SMC) to communicate with the secure world. The only observable execution effects are returned data (e.g., return values) and the status of the TA.

The gateway to the TEE is usually a TEE-specific kernel driver. While kernel fuzzers like Syzkaller [32] or DI-FUZE [20] are capable of fuzzing kernel drivers, these tools
fail to capture the message semantics required to interact with TAs, rendering them ineffective when targeting TEEs. Thus, novel techniques are required to find and fix bugs in these security-critical applications.

We present TEEzz, a fuzzing framework for TAs running on commercial smartphones. TEEzz targets three popular TEE implementations: the Qualcomm Secure Execution Environment (QSEE) [62], used on many phones including the popular Nexus and Pixel series; TrustedCore (TC) [79], found on Huawei devices; and the Open Portable Trusted Execution Environment (OPTEE) [51], the de-facto reference implementation for TZ-based TEEs. The analysis first identifies the TAs within the TEE and then manually triggers interactions with them. During these interactions, TEEzz records the data passed both into and out of the TEE to automatically reconstruct the message format and complex message and value dependencies. Lastly, this message format, along with the dependencies of the interaction (i.e., generating a cryptographic key before it is used for encryption), are fed into our fuzzer. The fuzzer explores the TEE while continuously checking for liveness and monitoring for crashes.

TEEzz necessitates diverse contributions. First, the complex and proprietary data structures of TAs require fuzzed inputs to be well-formed, or else the parsing logic in the tOS will reject them. Thus, we designed TEEzz as a mutation-based fuzzer that operates on type- and state-aware seeds generated from legitimate interactions with TAs. To infer the necessary knowledge of the API, we design an inference mechanism that maps high-level abstractions to low-level messages used to communicate with the TEE. TEEzz automatically generates memory introspection logic for each parameter type of the exposed interface and then abstracts the interaction protocol from the recorded traces. At runtime, we dynamically instrument this interface, parse the values corresponding to each type on-the-fly from memory, and save the type-aware token sequence to disk. The observed type- and state-aware interactions become the specification for efficient mutation.

Second, we automatically generate type-aware mutators for the enriched seeds. We convert the type definitions used by TA-facing interfaces into type-aware mutator plugins for TEEzz’s mutator engine. While fuzzing, TEEzz leverages these type-specific mutators to manipulate input tokens.

Third, many TAs are stateful, and value dependencies between invocations need to be resolved, i.e., a value returned from one invocation must be used as an input for a future invocation. Leveraging the previously recorded type-enriched interaction sequences, which include ingoing and outgoing data, TEEzz employs a novel value dependency inference technique to add state-awareness to its seeds.

Finally, TEEzz is the first end-to-end solution capable of continuously fuzzing TAs on COTS Android devices. Although we use known techniques such as dynamic binary instrumentation-based introspection (DBII) and semantic reconstruction, the novelty of TEEzz stems in solving technical challenges (Section IV) to apply these techniques to a restricted environment of a COTS device with no direct access to secure world entities. In addition, TEEzz features an extensible type-aware mutation engine, a state-aware fuzzing paradigm that considers entire interaction sequences as seeds and resolves interaction dependencies during runtime. Further, it supports state reset mechanisms to deterministically build up TA state to facilitate the reproduction of crashes.

We evaluated TEEzz’s performance in terms of coverage and bug-finding capabilities in a ground-truth experiment. For this purpose, we extended the OPTEE platform with (1) permanently shared memory between client applications (CAs) and TAs, (2) TA instrumentation to populate the coverage bitmap, (3) TA instrumentation to collect program counters during post-processing, and (4) support for TA constructors to initialize the instrumentation. Due to the non-availability of related fuzzers, we truthfully replicate the state-of-the-art based on AFL++ and compare TEEzz against three TA-aware AFL++ variants. Our results show that TEEzz finds bugs that are unreachable for existing fuzzers. In fact, TEEzz was the only fuzzer capable of finding three previously unknown bugs in OPTEE TAs. Further, we tested TEEzz on 18 TAs covering four popular Google and Huawei phones. TEEzz successfully generated enriched seeds, inferred interaction dependencies, and fuzzed each TA. Across these proprietary targets, TEEzz successfully found 40 unique bugs that we responsibly reported to the corresponding vendors. One CVE (CVE-2019-10561) was assigned so far, and we await further replies. Some of these crashes force the phone into a factory reset—wiping all user data—to resume normal functionality. In contrast, others allowed us to extract protected cryptographic keys from the TEE, a stepping stone to launch brute-force attacks against a device’s disk encryption.

In summary, our contributions are as follows:

- We developed TEEzz, available at [GitHub](https://github.com/HexHive/teezz-fuzzer), the first end-to-end automated fuzzing framework capable of fuzzing TAs on commercially available smartphones;
- an automated, dynamic-analysis-based technique for inferring field types of messages, as well as their dependencies, to facilitate type-aware fuzzing of stateful TAs;
- type and state-aware fuzzing mutators that leverage the message and dependency information inferred from analyzing the interactions with the TAs; and
- a thorough evaluation of TEEzz against other state-of-the-art fuzzing techniques on production TAs.

II. Background

In this section, we present the relevant background for ARM TrustZone and its usage on the Android platform. Then, we cover the fuzzing concepts necessary for our research. These concepts include type-awareness and state-awareness.

a) TEEs on the Android Platform: The vast majority of smartphones uses modern ARMv8-based SoCs that leverage TZ to provide a hardware-supported TEE for security-critical functionality. Figure 1 shows the privilege levels (i.e., exception levels) of ARMv8 TrustZone and the location of the complex feature-rich code (e.g., Android) in the “normal” world,
and the sensitive code (e.g., the management of authentication keys or DRM media) in the “secure” world. Software in each of these worlds is referred to as either untrusted in case of the normal world (i.e., the rOS on N-EL1 and its CAs on N-EL0), or trusted in case of the secure world (i.e., the tOS on S-EL1 and its TAs on S-EL0). The secure monitor on the highest privilege level (EL3) is responsible for switching a central processing unit (CPU) core from one world to the other; therefore, both OSes have to call into the secure monitor to interact with the respective other world.

The security-critical features in TEEs are usually implemented through dynamically loadable TAs. On Android, the interaction with TAs is primarily initiated by system services that expose their functionality to regular apps via the Android Framework application program interface (API) (i.e., Binder-based inter-process communication (IPC)). Figure 1 depicts a simplified but typical communication paradigm between CAs and TAs on the Android platform.

An Android app uses system services to access hardware features (e.g., microphones, cameras, and sensors). Many of these system services leverage the capabilities provided by a TEE in the background (e.g., keystroke, gatekeeper, and fingerprint). Thus, a TEE is treated as a hardware device in Android’s architecture. To have common abstractions across vendors and their hardware-specific implementations, Android defines hardware abstraction layers (HALs) [33] for all system services dealing with hardware. Regarding the TEE, each system service uses a proprietary CA library (CAlib), which exposes the common HAL interface and implements the TA-specific marshaling of the request. Additionally, this CAlib usually invokes a further proprietary TEE library that implements abstractions for interacting with the vendor-specific TEE driver (1) in Figure 1.

The TEE driver, usually invoked using an ioctl[20] system call, is responsible for forwarding the request to the tOS (we use the terms driver interface and ioctl interface interchangeably). Since the rOS and tOS have different virtual address spaces (i.e., they use separate page tables), the rOS needs to provide the request in a coordinated shared memory region before it calls the privileged SMC instruction (2) in Figure 1 to switch worlds. Once the message is received by the tOS, the tOS forwards the message to the corresponding TA (3) in Figure 1.

b) Type-Awareness: Naïve mutation primitives consist of random modifications, appending data, and removing data. Other mutations may depend on the type of the data. For example, an enum type with a set of possible values could be mutated to ensure that it will have one of the possible values with high probability. The mutations aim to reduce the range of interesting values for a mutation to trim the search space for inputs overall. We refer to this principle’s extension from primitive to complex data types as type-awareness.

Automated type-aware fuzzers such as DIFUZE[20] and HFL[44] require source code for the generation of their fuzzing templates. They thus cannot be effectively applied to TAs as they are available only in binary form and often encrypted (e.g., Huawei).

c) State-Awareness: Many applications and libraries have internal finite state machines. For instance, a file cannot be read from or written to without first opening it. This dependency is enforced programmatically by requiring that a valid file descriptor be passed to the read and write function. This descriptor can be obtained by opening a file through open. We say that these functions have a value dependency and a consumer of this API needs to be state-aware [35], i.e., use functions in an expected order and resolve value dependencies, e.g., open before read and correctly passing the file descriptor.

Fuzzers that have an automated approach to state-awareness either require source code (FuzzGen [40] and HFL [44]), or fine-grained logs (IMF [38] and MoonShine [59]), both of which are hard or rather impossible to get on COTS devices with a proprietary TEE.

III. INTERACTING WITH TAs

One of the fundamental requirements to fuzz TAs is the ability to execute them by providing interesting inputs. This section discusses different ways to interact with TAs along with their tradeoffs, thereby providing the rationale for choices made in TEEzz.

A. Executing TAs

TAs require the corresponding TEE, whose execution in turn requires the specific hardware (see Section II). The two straightforward choices are executing the TAs on-device or rehosting the TAs [37], i.e., executing a hardware-dependent program in an emulated environment.
a) On-device execution: Executing a TA directly on the original (or at least compatible) hardware along with the corresponding TEE is the most precise way to capture the runtime behavior of TAs. In cases like Huawei, where the TEE firmware is encrypted, on-device execution is the only way to execute corresponding TAs.

No Introspection: For the majority of available devices, the original equipment manufacturer (OEM) restricts modifications to the TEE. This prevents the analyst from introspecting the TA on the device as it requires modifying or instrumenting the TEE. TEE software exploits or hardware attacks to gain code execution within the TEE are out of scope. These options would enable more powerful introspection capabilities but are challenging to carry out and not universally available.

b) Rehosting: These techniques enable the execution of a hardware-dependent program, such as a TEE, in an emulated environment. Samsung’s not-publicly-available PartEmu [39] rehosts proprietary TEE software stacks, consisting of the iOS and TAs, to an emulated environment based on QEMU [10] and Panda [23].

Incompleteness: Existing approaches suffer from inaccuracy of emulated hardware and software. Chip designs and corresponding data sheets are often unavailable publicly, and accurately emulating hardware without this information requires non-trivial reverse-engineering. Although PartEmu emulates peripherals such as the TrustZone address-space controllers (TZASC) and protection controllers (TZPC), chip manufacturers add custom components that are difficult to emulate. For example, Qualcomm chips have a proprietary eXtended Protection Unit “XPU” enforcing dynamic memory protection of shared buffers between the secure world and the normal world, making it an interesting fuzzing target. Furthermore, PartEmu does not model certain proprietary and necessary peripherals like the fingerprint sensor, ARM CryptoCell [91], or face identification hardware. This lack of perfect emulation results in a lack of fuzzing accuracy causing false positives and negatives. For an emulator, practically feasible emulation often involves sacrificing some hardware accuracy.

B. Input Injection

Due to restrictions placed on the secure world (i.e., the inability to modify or debug code), inputs to TAs must necessarily originate from the normal world. Within the normal world, fuzzing inputs can be injected into the system (with the intent of reaching TAs) at several locations, each having accompanying tradeoffs:

Client Application (CA): Requests to TAs typically originate from a CA. Thus, directly exercising the CA generates valid requests to the corresponding TAs. However, the input to the CA has little or no control over the values that are ultimately sent to the TAs. As the vendor-specific proprietary TEE client library (see Figure 1) sanitizes the forwarded input, this interface is ineffective for fuzzing.

Driver Interface (1) in Figure 1: There are already techniques, specifically Syzkaller [32] and DIFUZE [20], that can provide fuzz inputs at the driver interface. However, they fail to generate valid requests through the SMC interface (2) in Figure 1 since they are neither aware of the input format accepted by TAs nor their internal finite state machines.

SMC Interface (2) in Figure 1: While it is straightforward to fuzz at the SMC call interface, the generated inputs are likely to fail to reach the TA since they most probably do not adhere to the TA-specific protocol or data format.

This multi-interface invocation of a TA presents a tension between the high-fidelity, low-control interface of the CA and the low-fidelity, raw interface of the SMC call. As fuzzed input injection moves farther away from the TEE, the amount of control over the fuzzed data decreases while the likelihood of the data ultimately reaching the targeted TA increases. For example, the SMC interface permits complete control over the data but requires an accurate reconstruction of the entire communication protocol. On the contrary, the CA provides limited ability to modify the handled data, as numerous checks and mutations are performed, but creates valid packets for the TA.

C. Interaction with TAs in TEEzz

To prevent inaccuracies introduced by software and hardware emulation, we interact by executing TAs on the device. We handle the lack of introspection by using a coarse-grained feedback mechanism, as explained in Section VI to determine the state of TAs.

We inject fuzzed inputs through the driver interface, as it gives sufficient control over the data sent to a TA and makes our input injection mechanism independent of the SMC calling convention (as determined by the SoC). However, as mentioned before, generating valid requests to a TA through the driver interface requires us to know the input format expected by the TA, along with the status of its internal finite state machine. We will discuss these challenges and how TEEzz handles them in the next section.

IV. CHALLENGES AND TEEZZ’S APPROACH

Effectively sending arbitrary but well-formatted data to a TA through the driver interface requires (i) Type Awareness, an understanding of the structure and type of data accepted by the TA; and (ii) API State-Awareness, an understanding of the interdependencies between the data sent to a TA across multiple requests. For instance, a sign request to the keymaster TA has to contain information about the key to be used, which has been generated or imported by a preceding request.

A. Challenge #1: Structure Recovery (Type-Awareness)

The data expected by TAs through the TEE driver interface (1) in Figure 1 is proprietary and specific to the TA. But, the structured and typed data from the upper layers (i.e., the CAlib) enter the driver interface (1) in a serialized and untyped format.
To effectively send input to a TA, the format of the expected input and the marshaling method expected by the driver interface must be known. We cannot analyze the entities in the secure world (right half of Figure 1) as the firmware is mostly encrypted with device keys. Even in the normal world, recovering the structure and type of data expected by a TA has the following two sub-challenges:

**Java Native Interface (JNI):** As shown in Figure 1 in the normal world, the data from the app to the TA goes through various layers before reaching the driver interface. Each layer marshals the incoming data in a specific format before sending it to the next layer. We need to track the data flow to understand which parts of the data sent to the driver interface are needed by the TA. However, transitions between these layers are complex, such as the app calling from a Java context into a native library using JNI and the inter-process communication with the system service using Binder. But, precise static tracking of information flow across the JNI layer and process boundaries is still an open research problem.

**Type and structure recovery on binaries:** Few of these layers involved in the data flow from the app to the driver interface contain potentially closed-source, stripped, and obfuscated binary code, e.g., CAlib. These libraries contain essential information about the structure and type of the data accepted by TAs. Recovering this information requires precise structure and type inference at the binary level, which is also an open research problem.

**TEEzz’s Approach:** TEEzz recovers formats expected by TAs using a novel combination of dynamic binary instrumentation (DBI), multi-interface message recording, and semantic deduction (i.e., automatically bridging the semantic gap between the low-level interface and high-level API). As we will explain in Section V-C, TEEzz exploits the availability of types from the CAlib layer interfacing the proprietary vendor code to reconstruct the format carried within the messages crossing the driver interface.

### B. Challenge #2: API Statefulness (State-Awareness)

As mentioned before, interactions with TAs are stateful. Existing techniques, such as PartEmu, reset (or restart) the state of TAs after each request and hence do not account for the accumulation of state and fail to explore the state machine of TAs.

An effective fuzzer must account for the TEE implementation-specific state machine for loading TAs and establishing sessions as depicted in Figure 1. Moreover, it also has to account for the TA-specific state machine. Any fuzzing technique must adhere to the proper TA-specific protocol to go beyond shallow bugs.

**TEEzz’s Approach:** As mentioned before, we use DBI to record the messages exchanged with TAs. In addition, we also preserve the order of these recorded messages. Using this ordered list of messages and their corresponding data, TEEzz attempts to infer the API model of TAs for a given message sequence (i.e., the appropriate fields and invocation order) by analyzing data dependencies between messages (as explained in Section V-C).

## V. System Design

TEEzz is designed as a pluggable framework to ensure its portability and extensibility as existing TEEs continue to evolve, and new TEE vendors enter the landscape. For each platform, TEEzz must be initialized to ensure effective fuzzing. Figure 2 provides a high-level overview of the workflow of our system. The first step is to identify the CAlib interacting with TAs that are installed within a given TEE, obtain their interfaces, and create CAlib consumers, if none exist, to trigger interactions (Section V-A). Next, TEEzz automatically generates DBI recorders and type-aware mutators from these CAlib interfaces (Section V-B). Then, TEEzz leverages these recorders to carry out multi-interface interaction capturing. Having the same interaction recordings on a semantically-rich high-level interface and a semantically-poor low-level interface serves us to propagate types to the lower level where we have more flexibility to inject mutated inputs for fuzzing. Further, based on the recordings of entire interaction sequences, TEEzz attempts to infer value dependencies between interactions, resulting in type- and state-aware TA interactions that can be used as seeds by the fuzzing engine (Section V-C). Finally, TEEzz employs these enriched seeds and the previously generated CA-specific mutators to fuzz TAs on COTS Android devices (Section V-D).

### A. CA Identification

CAs are built to interact with TAs, and they are part of the normal world execution context that we fully control. TEEzz is based on the idea of extracting the knowledge of CAs about the protocol and message formats needed to interact with TAs. We chose a dynamic approach to capture this knowledge by recording interactions (Section V-C) and therefore need a way to trigger these interactions.

From our observation, we can find four usage scenarios for CAlibs on COTS Android devices:

- **Android Open Source Project (AOSP) System Service.** These are services usually present on all Android devices, and Google specifies the open source CAlib interface. The well-known keystored, gatekeeperd, fingerprintd, and mediaserver are examples for these services.

- **Vendor System Service.** These are vendor-specific services, and the CAlib interface is not publicly available. CAs for secure storage or anti theft features are examples for these services.

- **Unused CA.** Some vendors deploy CAlibs to their devices that are not used by any component. In this case, the corresponding TAs are present and fully functional. They are just not used.

- **Non-Existen CA.** This scenario applies to the situation where a TA exists without having a corresponding CA deployed on the device. Since TEEzz’s approach requires CAs, this scenario is out of scope.
TEEzz identifies CAs based on our observation that all vendors use a single TEE client library that serves as an abstraction to interact with the TEE driver (e.g., libQSEEComAPI.so on Qualcomm, libteec.so on Huawei, libmcclient.so on Samsung). By generating the dependency tree using a static analysis that recursively traverses all dependent objects, we identify all CAs and, if applicable, their corresponding services.

We obtain the CAlib interface for all CAlibs used in AOSP system services from the AOSP repository. For CAlibs used in vendor system services and unused CAlibs an analyst has to manually spend the effort to obtain interface definitions and trigger interactions with TAs.

The automatic extraction of these interface definitions and automatic unit-test generation are open research problems. For example, FUDGE [8] and FuzzGen [41] deal with this problem. Both approaches rely on source code and the existence of library consumers. These prerequisites are not given regarding the problem of fuzzing TAs on COTS devices. Thus, we consider the automation of this step as an orthogonal research problem and opted for a manual approach. From our experience, the CA layer is usually designed for interaction, meaning that the affected libraries have exported symbols. Furthermore, we did not encounter any obfuscation techniques. For vendor system services, we can trigger interactions and have a dynamic component for the interface analysis.

B. CA Interface Processing

Given a CA interface definition in C or C++, TEEzz automatically generates DBII recorders and type-aware mutators from the abstract syntax tree (AST) representation of the interface.

a) DBII Recorder Generation: Memory introspection is a well-known technique used in the context of virtual machines. Virtual machine introspection describes the process of a host that reads and parses raw bytes of a guest to reconstruct the meaning of values. For TEEzz, we leverage this technique to record each incoming and outgoing parameter of the CAlib interface according to its type using DBI. For some predefined data types, this technique might be known from the commonly known strace tool. strace can parse and print information about parameter types of system calls while tracing. The parsing logic for these types is hard-coded and limited to a few widely used complex types.

TEEzz, similar to strace, is capable of parsing primitive types, like char, short, and int. In contrast to strace, TEEzz can automatically generate the parsing logic for complex types from the definition of this type. Consider the function `func()` in Figure 3. It accepts the parameter `s`
of type struct S*. TEEzz generates an onenter and 
an onleave hook for each function of a given interface. 
These hooks, triggered on function entry (onenter) and exit 
(onleave), contain the parsing logic to retrieve the complete 
parameter from memory according to its type. For the struct 
S* type, this logic would dereference the pointer passed as the 
first parameter and read three chunks of memory: four bytes 
for int a, four bytes for enum E e, and 64 bytes for char b[64].

Our approach misses to accurately figure out the size of 
picticular objects, e.g., void* and the size of buffers pointed to 
by an int* parameter. In these cases, we use specific 
heuristics similar to the prior work [53].

For interaction recording, the onenter hook is responsible 
for recording all incoming parameters, and the onleave hook 
is responsible for recording all outgoing parameters and the 
function’s return value.

b) Type-Aware Mutator Generation: From the CAlib 
interface, we know which types are passed to the lower layers 
and eventually to the TA. The intuition for the generation of 
type-aware mutators is that during fuzzing, we want to create 
semi-valid inputs, and the knowledge of types helps to reduce 
mutations that likely lead to invalid inputs.

TEEzz converts the type definitions known from the CAlib 
interface into type-aware mutators. Given the token sequence 
of an enriched seed, TEEzz can choose from several type- 
specific mutations. In Figure 3 we illustrate a mutator for an 
enum. For example, the constants of the enumeration enum 
E { x, y, z }; are encoded using four bytes in C/C++. A 
a naive mutator would flip random bits of these four bytes 
and disregard the fact that the type already indicates that only 
x=0, y=1, and z=2 are valid values. This mutation usually 
leads to many wasted cycles because the parsing component 
immediately rejects the input. Using the mutators generated by 
TEEzz, mutations on this enumeration take the value space of 
these four bytes into account and assign the values indicated 
by the type with a higher probability.

C. Seed Recording

In this stage of TEEzz, we obtain type- and state-aware 
interaction sequences for each TA. We capture TA interactions 
at two interfaces, the semantically-rich and low-control CA 
interface and the semantically-poor and high-control TEE driver 
interface. This multi-interface interaction recording allows us 
to propagate the types observed at the CA interface to the TEE 
driver interface. This lower-level interface is more appropriate 
for fuzzing because we have more control over the inputs 
eventually passed to TAs and can bypass the sanitization 
logic of the CA layer. The seed preprocessing is finalized 
by inferring the inter-interaction value dependencies. We can 
identify outgoing data utilized as input in a later interaction 
using the interaction recordings. The availability of types 
facilitates this inference.

The recording of seeds is a TA-specific process, meaning 
that each TA has its command handlers, expected parameters, 
and interaction sequences. The corresponding CA knows about 
these specific internals. Given a component that exercises the 
CA interface (i.e., an AOSP system service, a vendor system 
service, or a manually crafted CA driver), we can capture valid 
interaction sequences that can be used as seeds for fuzzing.

a) Multi-Interface Recording: Interaction recording can 
be carried out on several levels of abstraction. The CA layer 
exposes a semantically rich interface containing API calls 
that map directly to TA-implemented command handlers. Unfortunately, the CA layer also implements sanitization logic 
that reject invalid inputs early before they even reach the TEE.

The TEE driver layer exposes an interface that is primarily 
designed as a transport layer to pass serialized interactions 
back and forth between the rOS and the tOS. This interface 
allows for arbitrary manipulations of the inputs passed to TAs 
but does not give us any high-level semantics about types or 
dependencies between interactions.

Based on this observation, we decided for a multi-interface 
interaction recording to get the best of both worlds. Using DBI, 
we record input and output messages at the CA interface and 
the driver interface to correlate fields and type information 
later. Once a variety of messages have been recorded (e.g., 
every exposed function is exercised), TEEzz automatically 
generates type- and state-aware TA interaction seeds.

b) Type Propagation: TEEzz leverages the types and pa-
rameters for each function, which are automatically extracted 
from the interface definition (i.e., C- and C++-header files), to 
map the high-level semantic information to the raw values that 
were recorded in memory. Consider Figure 4, which shows the 
data recorded for the parameter P and K at a representative 
CAlib interface method, f1(). By also recording the input 
and output buffers at the low-level driver interface, TEEzz 
identifies matching subsequences and similarly maps the high-
level semantic information. A linear scan yields the offsets 
of recorded CAlib parameters within the data sent to (and from) 
the driver interface. For example, the four byte size_t field 
len is easily observed at the beginning of the driver input 
buffer.

In addition to the CAlib parameter identification, TEEzz 
applies further structure reconstruction heuristics to also iden-
tify length fields, offsets, and strings (e.g., off_t the 
0x10 offset to the uint8_t[] k at the beginning of the buffer). 
Each pair of recordings at the CAlib and driver layer results 
in a model for an enriched seed that TEEzz later uses for 
type-aware fuzzing.

c) Interaction Dependency Inference: TEEzz is also ca-
pable of stateful fuzzing. The function, f1(), in Figure 4 
has an output parameter Kout of type keyblob_t**. Note 
that in order for the subsequent function, f2(), to succeed, 
this return value, Kout, which is an output parameter to f1(), 
must be correctly generated and passed as the input parameter. 
To meet this stateful requirement, TEEzz tracks multiple calls 
and automatically identifies output parameters of a call that 
are used as input parameters in subsequent calls. With the 
raw buffer from the driver recorder and the recorded type-
annotated parameters from the CAlib interface, TEEzz can 
infer the structure used at the driver interface. This type of
interaction is quite common in real-world TEE interactions. For example, to perform cryptographic operations using a key inside the TEE, the key must be generated first, with a known reference value.

Furthermore, we replay a recorded sequence multiple times to eliminate false positives and disable the resolution of individual value dependencies. If the dependent call still succeeds, we remove the value dependency.

Identifying these value dependencies is crucial to be later able to fuzz stateful APIs. Overall, performing type-aware and stateful fuzzing of TAs is an essential contribution of TEEzz.

D. Target Fuzzing

The actual fuzzing is carried out by a host component, called TEEzzH, and a target component, called TEEzzT. TEEzzH selects the call sequences to be fuzzed according to the API model and performs type-aware mutation based on the types inferred in the type-aware model generation step. Then, it sends the mutated call sequence to TEEzzT and evaluates the response. TEEzzT is a proxy that forwards the inputs to the TEE driver and returns its responses.

VI. IMPLEMENTATION

In this section, we present the implementation details of TEEzz.

A. Hook-based Requests Recording

TEEzz is able to automatically inject hooks using dynamic binary instrumentation (DBI) by leveraging Frida [38], a popular and stable DBI framework. By specifying a function to be hooked (e.g., the ioctl-wraper function within libc or functions of the CAlib), Frida allows for the injection of logic at the very beginning (onenter) and the very end (onleave) of a function. This technique allows TEEzz to record input and output parameters, as well as the return value, without corrupting the hooked function’s logic.

Frida expects the recording logic executed by these hooks to be specified in JavaScript. The complexity of this logic is directly dependent on the complexity of the data structures to be recorded. For the TEE driver interface, the recording logic is relatively simple because the data structures are flat and consist primarily of length fields and their corresponding buffers (i.e., uint8_t[]). Since it is a one-time effort, we manually implemented simple cases like this for each targeted TEE driver interface.

However, consider the parameter p of type param_set_t* from our request recording illustration in Figure 4. The recording logic for this parameter needs to traverse a nested struct, and also account for the runtime value of len that indicates how many elements (not bytes) of type param_t are referenced by p. In reality, the structure of the CAlib interface parameter types is often complex. For complex cases like these, TEEzz implements a DBII recorder generator capable of generating complex introspection logic for Frida hooks based on an interface definition. This generator is written in Python and uses Python’s libclang bindings to parse the header files describing the CAlib interfaces. After the parsing step, TEEzz emits Frida-hooks in JavaScript for each CAlib function containing proper introspection logic for parameters according to their types.

By using the AST, TEEzz’s DBII recorder generator produces hooks that can traverse each parameter node of the AST down to its leaf nodes (primitive types) potentially following pointers through memory. We record the corresponding value from memory for each leaf node and annotate it with its respective type. A simplified example of this process is illus-

---

Fig. 4. An indicative function call sequence of a crypto API as exposed by the keystore system service. Our recorders (green) log the incoming (left) and outgoing (right) parameters of the CAlib and the driver interfaces (yellow). After recording, we match the typed data logged by the CAlib recorder with the raw buffers from the driver recorder using our type recovery (purple) to generate fuzzing enriched seeds for fuzzing. Beyond that, we recognize K being an output parameter of f1() and an input parameter of f2(), thus, accounting for the TA internal stateful API.
treated in Figure 3 where the leaf nodes of func’s parameter
struct S* s are recorded.

In cases where explicit information about the relation of parameters or structure members is unavailable, TEEzz uses
heuristics. For example, given a struct that encodes a buffer’s
length (i.e., one parameter being a buffer and the other
one being its size), it may not be explicit from the header
definitions to see if the current parameter name is a substring
of the neighboring parameter suffixed with a size indicator (e.g.,
buf and buf_{len,sz,length,size}). If so, the size
parameter’s value is read from memory and used to record the
buffer. This heuristic is implemented conservatively to prevent
reading random memory content; we do not record the buffer
if our heuristic does not succeed.

B. Type Recovery

To perform type recovery of the buffers recorded at the
ioctl interface, TEEzz first matches parameters from the
CAlib interface with byte sequences of the recorded buffers at the ioctl interface.

The implementation of the matching algorithm originates
from the intuitive way of implementing serializers. Given a
nested data structure, it is common to apply a depth-first
traversal and store the data of neighboring AST leaf nodes
next to each other. Thus, our matching algorithm traverses the
recorded leaf node values of the high-level interface using a
depth-first approach and tries to find identical values within
the byte sequences captured from the low-level interface. If we
encounter collisions, we apply a greedy strategy and prioritize
the match with a higher number of bytes.

In the second stage, TEEzz applies heuristics to identify
offset, size, and constant fields. TEEzz assumes that the size
and offset fields will be four or eight bytes depending on the
bitness of the targeted TA. For offset fields, TEEzz scans for
any offset-sized byte sequence that points to data from the
beginning of the buffer (alignments of 4, 8, or 16 bytes are
considered). When an offset candidate points to the beginning
of an already identified type from the previous step, it is a
good indicator that our heuristic identified the offset correctly.
For size fields, TEEzz scans for any size-sized byte sequence
that matches the size of an already identified type from the
previous step. Size fields are not only identified in terms of
their number of bytes (i.e., a buffer length) but also by the
number of elements in a list (i.e., structures stored in an array).
If a value is identified that does not change across all observed
recordings, we consider it a constant.

C. API Model Inference

Our recorded interaction sequences keep the chronological
order of calls to the ioctl interface of the TEE driver to
ensure that our fuzzed inputs satisfy the state and protocol
requirements of the underlying TA. Furthermore, TEEzz identifies
value dependencies between outputs and inputs. For
example, one interaction produces a value consumed by a
later interaction. To identify these dependencies, we scan
through the typed fields of the output for each interaction and
search for a matching typed field within the inputs of later
interactions. Our current heuristic creates a value dependency
if the type and value of a field in the output and input match.
This approach allows TEEzz to satisfy dependencies from a
request to a prior response, indicating which bytes from a
given response need to be replayed verbatim in future requests,
permitting the fuzzing of stateful TA APIs.

To eliminate false positives, we replay each sequence and
systematically remove value dependencies while comparing
the output behavior to the output of the original recording. If
we can remove the dependency and achieve the same behavior,
we find a false positive and do not consider this dependency
for fuzzing.

D. Fuzzing

TEEzz’s fuzzer is implemented in two parts: a program
running on a machine connected to the phone that generates
inputs and logs results, called TEEzzH, and a stub that runs
on the phone and works as a proxy to pass the fuzzed inputs
to the appropriate interface, called TEEzzT.

TEEzzH orchestrates TEEzzT to load/unload TAs and
open/close sessions which is necessary to reach the core logic
of any TA (see Figure 4). Given an opened session, TEEzzT
accepts inputs that it forwards to the TA and sends the output
(the TA’s response) back to TEEzzH. TEEzzT uses the
TEE-specific API to interface the TEE driver, establish shared
memory with the TEE, and send commands to TAs.

TEEzzH observes the output behavior of the TA and initiates
reboots via Android Debug Bridge (ADB) if necessary. TAs
potentially accumulate state over time which is why we reboot
the device after a configurable number n (n = 500, by default)
of inputs sent to the target to reset the TA’s state.

TEEzz’s mutations are composed of type-aware mutations,
bit-flips, and well-known mutations. For the type-aware mu-
tators, TEEzz compiles all types used in the CAlib header
files into Google’s Protobuf interface description format. This
translation is again based on traversing the AST of all type
definitions leveraging Python and libclang. Given all avail-
able types specified in the Protobuf format, we use Google’s
Protobuf compiler for Python to access the AST types from our
mutators. An example for a mutator for an enum declaration
is illustrated in Figure 3. These type-aware mutators can be
combined with the bit-flip and well-known mutators. The
well-known mutations include edge cases for various width
signed and unsigned integers, null-byte insertion for strings,
increment and decrement operations, and the population with
random data.

After mutating an input, TEEzzH resolves value depen-
dencies by propagating values from prior TA outputs to the current
input. Figure 5 illustrates these value dependencies for the
keymaster TA. We store these dependencies in a DAG-like
data structure to resolve them while fuzzing.
TA crashes come in two flavors. TEEzz recognizes crashes that immediately reboot the phone and uses return codes from the TEE to evaluate target-specific crash conditions. When a potential crash is detected, a message is propagated to the runner, which persists the entire sequence of mutants. All the recorded seeds and mutants can later be checked for reproducibility using TEEzz.

VII. EVALUATION

In this section, we perform a comprehensive evaluation of TEEzz and show the effectiveness of each of our techniques. First, we conduct ground-truth experiments exploring different fuzzers’ capabilities to achieve coverage within TAs based on OPTEE (Sec VII-A). Second, we examine TEEzz’s type-recovery and value-dependency identification capabilities (Section VII-B). Third, we evaluate TEEzz and its various techniques on COTS TAs (Sec VII-C). As a bonus, we include our unsuccessful attempt to use Kernel driver fuzzers to fuzz TAs in Section A and a comparative evaluation with the closed-source PartEmu [39] system in Section B of our appendix.

A. State-of-the-Art Ground-Truth Comparison

To the best of our knowledge, PartEmu [39] is the only fuzzer that targets TAs of COTS devices. Unfortunately, neither the PartEmu prototype nor the datasets used for its evaluation are publicly available.

Due to the lack of availability, a first-order comparison with PartEmu is impossible. Therefore, we chose to reimplement PartEmu’s fuzzing approach as truthfully as possible and evaluate it on the available software.

Instrumented+noseed mode. This mode uses an instrumented target TA and follows the typical AFL fuzzing model using no seeds. This mode mimics the AFL PartEmu module and resembles the experiment carried out by Harrison et al. [39]. We consider this configuration to be our baseline.

Instrumented mode. This mode uses the same setup as in the previous configuration, but initializes the fuzzer with the seeds obtained through TEEzz’s seed recording.

Multi-interaction mode. This mode goes one step further than the AFL PartEmu module due to its ability to fire multiple inputs against a target TA and therefore allows for building up state. We also initialize this configuration with seeds obtained through TEEzz.

Our dataset consists of four TAs available for OPTEE: keymaster, gatekeeper, secure storage, and acipher. For keymaster and gatekeeper, we recorded the seeds from the vendor test suite (VTS) binaries available from the AOSP. For secure storage and acipher, we recorded the seeds from the CA command-line executables which are part of the OPTEE project. We use the keymaster and gatekeeper TAs from tag 3.8.0 to leverage the known vulnerabilities from the publicly available security advisory. The two important metrics we experimentally evaluate are each fuzzer’s coverage and capability to find bugs.

For the coverage experiment, we run each fuzzer ten times for 24 hours in a dockerized Ubuntu 20.04 setup on a machine featuring a Xeon Gold 5218 and 64GB of RAM. The coverage results from this experiment are shown in Figure 6.

TEEzz achieves a higher initial coverage compared to the other fuzzers for keymaster and gatekeeper due to its ability to resolve value dependencies between interactions. The API of keymaster contains calls that only succeed when a transitive value dependency across multiple calls is resolved. This property is also indicated by TEEzz’s coverage, which is the only fuzzer that reaches 52% coverage in the best case. Compared to the second best coverage of 25% in the multi-interaction mode configuration, TEEzz covers more than double the code in this target.

For gatekeeper, the seeds provided by TEEzz significantly support the exploration of the instrumented and multi-interaction modes. The configuration without proper seeding, as carried out in PartEmu’s evaluation [39] reaches a maximum coverage of 12% after 24h while the multi-interaction mode already starts at 38% and finishes at 58%.

From our experience, the keymaster and gatekeeper TAs of our dataset are similar to TAs found on production devices in terms of their complexity and interaction patterns. acipher and secure storage are example TAs of the OPTEE project and would likely not be used on production phones in their current implementation. For example, OPTEE’s secure storage TA differs significantly from the secure storage TA found on Huawei devices in terms of its missing session management to hold a session-specific state. Furthermore, acipher only implements two static cryptographic operations that do not require complex input formats, as seen for TAs on COTS devices. Hence, it is not surprising that the instrumented mode, even without proper
seeded, can significantly increase coverage and cover 69% of acipher and 74% of secure storage, respectively.

We found 13 previously unknown bugs in our OPTEE TA dataset. The AFL-based fuzzers combined could detect ten, and TEEzz was the only fuzzer that detected all 13 of these bugs. The three additional bugs found by our system required the target TA to build up state before they could be triggered. Thus TEEzz’s state-aware seeds facilitate finding state-dependent bugs.

In summary, TEEzz’s capability to resolve value dependencies across multiple interactions significantly improves the capabilities of modern fuzzers. Given that our coverage experiment featured coverage-guided fuzzer configurations that are commonly unavailable, whereas TEEzz is a blackbox fuzzer, it is a notable result that our fuzzer outperforms the competitors in realistic scenarios.

B. Accuracy of Type Recovery and State Awareness

In this section, we present an evaluation of TEEzz’s type recovery and state awareness. For the presented experiments, we leverage the Google VTS for the keymaster subsystem and execute it on a HiKey620 development board featuring an Android/OPTEE deployment setup. The vast set of test cases included in the Google VTS allows us to measure TEEzz’s type-recovery capabilities on many interactions and provides us with the ground truth for the behavior of sequences of interactions with the target. We use the latter to examine the accuracy of identified value dependencies (i.e., output values required as input values by later interactions).

a) Type Recovery: In TEEzz’s seed recording stage, we record all inputs and outputs propagating through the high-level and low-level interface. Due to the DBII recorders targeting the high-level interface, we obtain the AST for every single parameter passed to a function and have the run-time values (associated with the AST’s leaf nodes) readily available. This process allows us to map these leaf nodes of the high-level interface to the untyped byte sequences recorded from the low-level interface to recover types.

After executing and recording the entire Google VTS consisting of 102 sequences and 1,874 interactions with the keymaster, we map the input and output AST leaf nodes of the high-level interface to the untyped buffers of the low-level interfaces.

For the input, we were able to uniquely map 8,059 out of 20,676 AST leaf nodes. Our heuristics discarded 9,090 leaf nodes since they consisted of all-zero byte sequences and are indistinguishable from padding in buffers. We could not map 3,527 (17%) leaf nodes.

For the output, we uniquely mapped 14,821 leaf nodes out of 33,614 AST leaf nodes in total. 13,686 nodes were discarded because of our all-zero byte sequence heuristic, and 5,107 (15%) nodes could not be mapped automatically.

Besides uniquely mapped AST nodes, the coverage of the low-level buffers with types is an interesting accuracy metric. Table II shows the type coverage of low-level input and output buffers grouped by their high-level function and indicates the typical size of these buffers. For the eleven functions implemented by the keymaster TA, we can almost perfectly recover the types of inputs provided to six functions (97-100% recovery), perform well on another four functions (65-79% recovery), and only miss some of the types for one function (~30% recovery).

b) State Awareness: Our seed recording stage (Section 4C) also tries to infer the value dependencies of interactions occurring in a recorded interaction sequence.

To assess the accuracy of identified value dependencies, we leverage Google’s VTS for the keymaster TA and, in contrast to the other experiments, filter the tests for interactions that result in successful status codes to facilitate the creation of the ground truth for value dependencies. The keymaster target is representative due to the availability of many test cases and its stateful API. Figure 5 illustrates a typical value dependency graph of this target. In total, we obtain 95 sequences consisting of 728 interactions. Replaying all sequences, we manually mark dependent interactions and monitor if these interactions return successfully. If they return with an error code, we did not manage to identify and resolve the correct value dependencies.

Out of the 728 interactions in our dataset, we manually identified 565 as dependent on prior interactions’ output values. In comparison to this time-consuming manual approach, TEEzz automatically resolved 457 (81%) correctly and only failed to resolve 108 (19%).

Regarding false positives, we replay a recorded sequence several times and disable the resolution of individual value dependencies. If the dependent call succeeds, we remove the value dependency and effectively eliminate false positives.

C. Fuzzing COTS TAs

Using TEEzz, we fuzzed 18 TAs that can be found on COTS Android devices. The detailed TAs are listed in Table III.

First, we prepared all devices by rooting them. Then, we obtained the dependency graph to the respective TEE client library (e.g., libteec.so and libQSEEComAPI.so). Based on each dependency graph, we decided for a subset of CAs to fuzz their corresponding TAs. We excluded TAs that require human interaction during their usage (i.e., TAs related

TABLE I

<table>
<thead>
<tr>
<th>Function</th>
<th>Avg Cov Input</th>
<th>Avg Cov Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>addRngEntropy</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>abort</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>deleteKey</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>getKeyCharacteristics</td>
<td>99.75%</td>
<td>58.06%</td>
</tr>
<tr>
<td>exportKey</td>
<td>98.90%</td>
<td>3.99%</td>
</tr>
<tr>
<td>begin</td>
<td>97.48%</td>
<td>30.07%</td>
</tr>
<tr>
<td>finish</td>
<td>79.18%</td>
<td>41.67%</td>
</tr>
<tr>
<td>importKey</td>
<td>71.83%</td>
<td>78.24%</td>
</tr>
<tr>
<td>update</td>
<td>69.17%</td>
<td>28.57%</td>
</tr>
<tr>
<td>attestKey</td>
<td>65.43%</td>
<td>0.00%</td>
</tr>
<tr>
<td>generateKey</td>
<td>30.28%</td>
<td>65.42%</td>
</tr>
</tbody>
</table>
to fingerprint authentication or face identification) since there is no appropriate way to fuzz them.

For each non-AOSP CA, an analyst reverse engineered the CA-layer interface and wrote a small CA driver program that correctly uses the interface. On average, the analyst spent five hours recovering the interface and writing a CA driver. Given that the analyst will realistically spend many more hours on triaging crashes anyway, we argue that writing these drivers is a reasonable engineering trade-off and a great practice to familiarize with the interface.

Having a driver and the CA interface definitions for all of our targeted TAs, we first generated the DBII recorders and type-aware mutators. Then, we installed TEEzz’s multi-interaction interaction recorders and triggered the TA interaction either via the user interface of the phone or our manually developed CA drivers. Based on the interaction recordings, we propagated the types to the low-level high-control TEE driver interface recordings and performed the dependency inference. Finally, we were able to fuzz each of the TAs listed in Table II.

We ran TEEzz for 24 hours for each TA and analyzed all found crashes regarding reproducibility and uniqueness. In total, TEEzz detected 1,541 crashes of which we could reproduce 1,387. We deduplicated the crashes on QSEE manually and used the /dev/hisi_teeelog device on Huawei devices to obtain stack traces of crashing TAs. In total, we found 40 unique crashing inputs. We reported these bugs to the corresponding vendors and already got CVE-2019-10561 assigned.

We reset the target’s state during this experiment by rebooting the device after \( n = 500 \) interactions. On average, this mechanism resulted in 1,517, 1,073, 956, and 756 resets during our 24h experiment for the P9 Lite, P20 Lite, Nexus 5X, and Pixel 2XL devices, respectively.

Fuzzing the keymaster and gatekeeper TAs on the Nexus 5X resulted in unstable ADB connections and corrupted data partitions. These states did not occur on the other devices and are problematic for continuous fuzzing because the host component of TEEzz cannot access the device anymore. We had to perform 130 factory resets and 20 hard resets during this experiment for the keymaster and gatekeeper TA, respectively. To automatically handle these cases and allow for continuous fuzzing, TEEzz is capable of booting the phone into recovery mode and restoring a functional system. Additionally, our fuzzer can carry out hard resets using a phone case equipped with a servo motor that pushes the phone’s power button until it reboots.

TEEzz found 40 unique bugs in TAs deployed across 4 different COTS Android phones. This result provides evidence that TEEzz is capable of effectively finding bugs in TAs deployed on production devices.

### VIII. LIMITATIONS

Although TEEzz provides an effective way to fuzz test TAs, it suffers from several limitations.

**Unlocked device:** TEEzz requires complete control of the rOS for which the bootloader needs to be unlocked. Vendors who do not allow unlocking their bootloaders are challenging targets for TEEzz.

---

**TABLE II**

**RESULTS OF FUZZING COTS TAs USING TEEZZ. WE FUZZED EACH TA ON EACH DEVICE FOR 24H. ♦ – ENCRYPTED TA**

<table>
<thead>
<tr>
<th>TEE Device (OS Vers.) TA</th>
<th>CA</th>
<th>Req/Sec</th>
<th>#Crashes</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC P9 Lite (6.0) keymaster</td>
<td>AOSP</td>
<td>9.6</td>
<td>681</td>
</tr>
<tr>
<td>TC P9 Lite (6.0) gatekeeper</td>
<td>AOSP</td>
<td>11.1</td>
<td>645</td>
</tr>
<tr>
<td>TC P9 Lite (6.0) secure storage</td>
<td>Vendor</td>
<td>10.7</td>
<td>0</td>
</tr>
<tr>
<td>TC P9 Lite (6.0) rpmbkey</td>
<td>Vendor</td>
<td>7.3</td>
<td>0</td>
</tr>
<tr>
<td>TC P9 Lite (6.0) antitheft</td>
<td>Vendor</td>
<td>7.1</td>
<td>0</td>
</tr>
<tr>
<td>TC P9 Lite (6.0) hwsign</td>
<td>Vendor</td>
<td>7.1</td>
<td>0</td>
</tr>
<tr>
<td>TC P20 Lite (8.0) keymaster</td>
<td>♦ AOSP</td>
<td>12.7</td>
<td>10</td>
</tr>
<tr>
<td>TC P20 Lite (8.0) gatekeeper</td>
<td>♦ AOSP</td>
<td>6.4</td>
<td>0</td>
</tr>
<tr>
<td>TC P20 Lite (8.0) secure storage</td>
<td>♦ Vendor</td>
<td>6.5</td>
<td>0</td>
</tr>
<tr>
<td>TC P20 Lite (8.0) rpmbkey</td>
<td>♦ Vendor</td>
<td>5.3</td>
<td>0</td>
</tr>
<tr>
<td>TC P20 Lite (8.0) antitheft</td>
<td>♦ Vendor</td>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>TC P20 Lite (8.0) hwsign</td>
<td>♦ Vendor</td>
<td>5.8</td>
<td>0</td>
</tr>
<tr>
<td>QSEE Nexus 5X (7.1.2) keymaster</td>
<td>♦ AOSP</td>
<td>3.8</td>
<td>55</td>
</tr>
<tr>
<td>QSEE Nexus 5X (7.1.2) gatekeeper</td>
<td>♦ AOSP</td>
<td>6.0</td>
<td>0</td>
</tr>
<tr>
<td>QSEE Nexus 5X (7.1.2) widevine</td>
<td>♦ AOSP</td>
<td>7.2</td>
<td>80</td>
</tr>
<tr>
<td>QSEE Pixel 2 XL (9.0) keymaster</td>
<td>♦ AOSP</td>
<td>5.0</td>
<td>0</td>
</tr>
<tr>
<td>QSEE Pixel 2 XL (9.0) gatekeeper</td>
<td>♦ AOSP</td>
<td>3.3</td>
<td>0</td>
</tr>
<tr>
<td>QSEE Pixel 2 XL (9.0) widevine</td>
<td>♦ AOSP</td>
<td>5.3</td>
<td>70</td>
</tr>
</tbody>
</table>
Availability of CAIBs: In the worst-case scenario, some one-time manual effort is required per TA to obtain its CAIB’s interface definition and implement test cases that trigger interactions with the TA. We argue that synthesizing test cases is a problem at the frontier of science and has barely been solved in situations with source code available (see Fudge [8] and FuzzGen [41]), and is thus an orthogonal problem. With TEEzz we will release and document the tooling that allowed us to fuzz each TA listed in Table [I].

Extending to other TEEs: Our current prototype can fuzz TAs on three popular platforms. While the concepts generally apply to other TEEs, TEEzz needs adaptation to run on those platforms. We plan to support a further target in the future: Samsung’s TEE called TEEGRIS [68]. As far as we know, TEEGRIS is similar to TC and OPTEE, and porting TEEzz to this platform should be straightforward.

IX. RELATED WORK

Several researchers have studied and exploited vulnerabilities in TZ-based TEEs [46], [67], [72], including a class of flaws, called BOOMERANG [52]. Furthermore, there have been various side-channel attacks on TZ [14], [48], [78], [82], [90]. The static analysis techniques employed by the works mentioned above are difficult to generalize for vulnerability detection in TAs, as the structure and implementation of TAs depend on their tOS [4], [55], [75], [80]. Regardless, vendors started to encrypt TAs [85], making it difficult to retrieve their binaries, which renders static analysis techniques inapplicable.

Commercial TEEs with encrypted TAs are essentially blackbox systems that expose certain functionalities or APIs [27] that are accessible from the rOS through SMC [6] instructions. Fuzzing [13], [22], [25], [28] is a well-known technique to test blackbox systems. There has been significant progress in whitebox [60] and greybox [11] fuzzing regarding performance [54], coverage [12], [49], [64], [71], [73], and bug finding ability [17], [18], [61]. However, these techniques need access to the binary of the program under test. Although, Harrison et al. recently proposed a TOS emulation technique, PartEmu [39], which can be used to fuzz the corresponding TAs, this technique fails when the TOS and TAs are encrypted. Consequently, these techniques cannot be applied directly to fuzz TAs.

A further technique to fuzz blackbox systems is grammar-based fuzzing [29]. This technique uses the grammar of the target program’s input to generate fuzz inputs. Peach [60] is a well-known commercial grade tool for grammar-based fuzzing capable of processing complex inputs. Despite this, getting the input format accepted by TAs is difficult because there is no standard input format. Furthermore, well-known interface recovery techniques [9], [19], [63], [65], [81] cannot be used as they rely on the availability of the program binary.

Almost all commercial TEEs expose an interface in the rOS, usually in the form of a device driver. [1], [6], [74]. Most of these device drivers are open-source and expose high-level formats of the input accepted by the TAs, like raw request and response buffers. [2]. Recently, fuzzing techniques targeting device drivers have been proposed [20], [38], [43], [77]. Specifically, DIFUZE [20] uses static analysis to infer the accepted input format. This format is then used to effectively fuzz the driver, or applied by other device-driver fuzzers, like Syzkaller [32]. As we show in Section A in our appendix, these techniques are ineffective in generating inputs for TAs. In comparison, TEEzz is the first work opportunistically using different techniques to fuzz commercial TAs effectively.

X. CONCLUSIONS

While TEEs in modern smartphones are intended to provide a safe haven for sensitive data and computations, they also pose a major security risk. Privileged code running within the TEE has complete access to every aspect of the smartphone (e.g., cryptographic keys, hardware peripherals, and sensitive user data). Unfortunately, despite this potentially catastrophic security risk, traditional security analyses, such as fuzz testing, are rendered useless due to the limited access to and lack of feedback from production TAs. To address this analysis gap, we present TEEzz, the first TEE-aware fuzzer capable of effectively fuzzing TAs on production smartphones. TEEzz leverages a combination of DBI and stateful replay techniques to ensure that both the protocol and structure expected by the targeted TAs are met, enabling almost all of the fuzzed inputs to be processed. Indeed, TEEzz discovered over 40 unique crashes on QSEE and TC combined, resulting in one CVE so far, and found 13 previously unknown bugs in TAs running on OPTEE, an open-source reference implementation for TZ-based TEE.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their insightful feedback and appreciate the opportunity for a major revision to conduct additional experiments. This project was supported by European Research Council (ERC) grant No. 850868, SNSF grant PCEGP2_186974, DARPA grants R001119S0089-AMP-PP-034 and N6600120C4031, DHS grant FA8750-19-2-0005, and ONR grants N00014-20-1-2632 and N00014-17-1-2897. Any findings are those of the authors and do not necessarily reflect the views of our sponsors.

REFERENCES

use a kernel fuzzer to fuzz the TEE and TAs in particular. Our intuition for kernel fuzzers is that a general-purpose approach to device driver fuzzing is insufficient for the complex inputs and interactions required by the TEE. To validate this hypothesis and get a deeper understanding of its implications, we chose two representative kernel fuzzers, namely Syzkaller [32] and DIFUZE [20], to fuzz TEE drivers on COTS Android devices. We prepared both of these fuzzers according to the available documentations and deployed them on two COTS devices, the Pixel 2 XL (Android 9.0) and the Huawei P20 Lite (Android 8.0). The Pixel 2 XL is based on a Qualcomm chipset and, hence, runs QSEE. The Huawei P20 Lite is based on a HiSilicon chipset and runs TC. For this experiment, we rooted both devices and deployed custom kernels. The modifications to the original kernels are minimal. We instrumented the locations where the Linux kernel executes a secure monitor call to count the number of context switches to the TEE and to count the number of \texttt{ioctl} calls that the fuzzers generated. Additionally, we enabled \texttt{kcov} for the Syzkaller experiment. Since DIFUZE is not a coverage-guided fuzzer, we did not enable this feature.

For Syzkaller, we manually specified the TEE driver’s syscalls and different \texttt{ioctl}-handlers, including the different \texttt{argp} data structures expected by these handlers. Furthermore, Syzkaller’s grammar allowed us to specify stateful APIs. For example, a filedescriptor that is returned from an \texttt{open} syscall can be linked to the filedescriptor consumed by an \texttt{ioctl} syscall.

DIFUZE automatically generates data models to fuzz \texttt{ioctl} calls from the driver’s source code, but, in comparison to syzkaller, it does not support stateful APIs or coverage-guidance.

<table>
<thead>
<tr>
<th>Technique</th>
<th>QSEE (Pixel 2 XL - taimen)</th>
<th>TrustedCore (Huawei P20 Lite)</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{DIFUZE}</td>
<td>$7,370,756$ (85.3)</td>
<td>$10,847,220$ (125.5)</td>
</tr>
<tr>
<td>\texttt{Syzkaller}</td>
<td>$7,006,579$ (81.1)</td>
<td>$2,245,970$ (20.71)</td>
</tr>
</tbody>
</table>

According to our experiment shown in Table III DIFUZE and Syzkaller have a respectable \texttt{ioctl} throughput of up to 85.3 and 125.5 requests per second ($\Sigma_{\text{ioctl}}$ per sec on the Pixel 2XL and the Huawei P20 Lite, respectively).

Contrary to our hypothesis that driver interface fuzzing techniques would not be able to generate any SMC requests (i.e., $\Sigma_{\text{SMC}} = 0$), DIFUZE and Syzkaller were both able to generate those requests by just fuzzing the driver interface. On QSEE, a small fraction (1.53% and 3.28%) of invocations reach the TEE, and on TC, up to a third of the calls reaches the TEE (35.67% and 20.71%).

To better understand this counterintuitive observation, we investigated the specific \texttt{ioctl} requests that yield SMCs. A TEE driver supports several command handlers that are not related to the TA lifecycle. Those handlers include facilities to query the tOS’s version or synchronize the time with the tOS. Not a single generated SMC is triggered by a command handler related to the interaction with a TA. Consequently, the interface to communicate with TAs is not reached at all.

Looking at the design of both kernel fuzzers, it is apparent that they would require extensive adaptations to incorporate a fuzzing harness that supports establishing sessions to TAs. Hence, we conclude that kernel fuzzers are ineffective for fuzzing TAs.

### Appendix B

Since PartEmu [39] is the only fuzzer targeting proprietary TAs, we contacted the authors to support our evaluation of TEEzz because PartEmu is not open to the public. According to the authors, the prototype cannot be made available due to parts of it being under a non-disclosure agreement with Samsung Research, and the individual TAs or firmware images used as a dataset in their evaluation cannot be revealed due to the security-sensitive nature of TZ.

We agreed with them to compare our results of a six-hour fuzzing session of the \texttt{widedvineTA} on the Pixel 2XL platform since this target was a common denominator. PartEmu was able to discover two shallow bugs in the target. One of these bugs was already discovered and reported during the development of TEEzz and the corresponding CVE-2019-10561 is assigned to us. The other bug discovered by PartEmu was not found by TEEzz because PartEmu is not open to the public. According to the authors, PartEmu only discovered shallow bugs and never reached deeper into the target’s logic than two to three function calls deep.

An inherent limitation of PartEmu is that it cannot be utilized for the encrypted TEE firmware images used on the Huawei P20 Lite. In comparison, TEEzz could not only fuzz the TAs on this device but also found a bug in the \texttt{keymaster} TA.