Flaky Test Detection in Android via Event Order Exploration

Zhen Dong
National University of Singapore, Singapore
zhen.dong@comp.nus.edu.sg

Xiao Liang Yu
National University of Singapore, Singapore
xiaoly@comp.nus.edu.sg

Abhishek Tiwari
National University of Singapore, Singapore
dcsabhi@nus.edu.sg

Abhik Roychoudhury
National University of Singapore, Singapore
abhik@comp.nus.edu.sg

ABSTRACT
Validation of Android apps via testing is difficult owing to the presence of flaky tests. Due to non-deterministic execution environments, a sequence of events (a test) may lead to success or failure in unpredictable ways. In this work, we present an approach and tool FlakeScanner for detecting flaky tests through exploration of event orders. Our key observation is that for a test in a mobile app, there is a testing framework thread which creates the test events, a main User-Interface (UI) thread processing these events, and there may be several other background threads running asynchronously. For any event $e$ whose execution involves potential non-determinism, we localize the earliest (latest) event after (before) which $e$ must happen. We then efficiently explore the schedules between the upper/lower bound events while grouping events within a single statement, to find whether the test outcome is flaky. We also create a suite of subject programs called FlakyAppRepo (containing 33 widely-used Android projects) to study flaky tests in Android apps. Our experiments on the subject-suite FlakyAppRepo show FlakeScanner detected 45 out of 52 known flaky tests as well as 245 previously unknown flaky tests among 1444 tests.

CCS CONCEPTS
• Software and its engineering → Software testing and debugging

KEYWORDS
flaky tests, non-determinism, concurrency, event order

1 INTRODUCTION
Regression testing aims to discover the adverse effects of the recently added code changes. Ideally, a test failure during regression testing should reliably signal issues with recent code modifications. However, some test failures are not due to the recent updates but due to flaky tests. Recent research [6, 17, 30] establishes flaky tests as tests that non-deterministically pass or fail when running on the same code version. Unfortunately, the presence of flaky tests significantly harms developers’ productivity [1, 25, 30].

Android is a reactive system, and its non-deterministic execution environment often causes concurrency-related flaky tests. In Android, a GUI test usually simulates user interactions to exercise the app’s functionality. However, due to non-determinism, a test could generate such interactions at “incorrect” timings. For example, a test to download an image and then open it without considering the internet’s speed could show non-deterministic behavior. A potential reason for such concurrency issues is the lack of synchronization among threads, e.g., in the failing run, a background thread would still be downloading the image while the testing thread tries to open it without synchronization with the background thread. Recent studies [34, 42] confirm that such synchronization issues lead to a significant number of flaky tests in Android apps. Thorgeir et al. [42] studied 77 flakiness-related commits in Android projects and discovered that 36% failures are due to synchronization issues between testing thread and app under test, e.g., a test accesses data in the app before the data is available. Similarly, Romano et al.’s study [34] shows that 33% flaky test failures are caused by threads’ synchronization issues (tests interacting with the UI elements before the elements are fully loaded).

Challenges. Exposing a concurrency-related flaky test in a test suite is challenging as such flakiness is only observed when events get executed in a particular order. In Android, a test’s event execution order may show non-determinism due to the non-deterministic execution environment. Some of such (irregular) event orders may cause a test to fail/pass occasionally. Consequently, detecting such a situation would require exploring such event orders deterministically; unfortunately, this poses a set of challenges for the existing flaky test detection approaches, which we discuss below.

• RERUN. A typical way to detect a flaky test is to rerun (expressed as RERUN) it multiple times. If it passes in some runs and fails in others, the test is marked as flaky. However, this approach can struggle to detect concurrency-related flaky tests because the target’s execution environment may not introduce the needed non-determinism. Besides, it may require too many runs to witness the flakiness.

• Noise Based. Several approaches [4, 38, 40, 48] run tests with environmental perturbations (e.g., changing CPU load or test execution orders) in the hope of observing unexpected behaviors.
Despite being simple and easy to be adopted in practice, adding noise does not guarantee to explore different execution orders. Consequently, they fail to detect many concurrency related flaky tests, as observed during our experiments.

- **Event Race.** The concurrency-related flakiness may be caused due to a synchronization problem on the testing thread. Besides, a concurrency bug on the app under test could also cause a non-deterministic test failure. Thus, in principle, approaches identifying concurrency bugs should also detect a subset of such flaky tests. Recent works [8, 18, 19, 24, 32] leverage program analysis techniques to analyze possible races in Android apps. For instance, EventRacer [8] records the execution trace of a test and statically analyzes it to find potential races. However, such approaches are prone to multiple false positives, which may create more problems for the developers during the Continuous Integration (CI) process.

**Our Approach.** We introduce a lightweight technique that automatically detects a concurrency flaky test in Android apps within few test runs. Our technique explores the possible event execution orders in a test by scheduling a non-deterministic (async) event such that each test run explores a different event execution order. Besides, we identify and explore event orders in which the test flakiness manifests such that a flaky test can be detected in a few test runs.

**Insight.** Possible event execution orders are introduced by non-deterministic execution of async events. Android apps use a concurrent event-driven model, in which only the main (UI) thread can access GUI and processes user events. To keep GUI responsive, the UI thread offloads a long-running task (e.g., internet accessing) to a background thread. Once the task is finished, the background thread sends an event (called async event) to the UI thread to perform a GUI update. In a test run, the testing thread and background threads send events to the UI thread simultaneously, resulting in races among the events; the test may pass for some event orders and fail in others. For the example mentioned earlier, the testing thread simulates the “open the image” event, and the background thread downloads the image and sends an update event after completing the download. A race may occur between these two events. If the internet is fast, the update event reaches the UI thread before the “open the image” event, and the test passes. Otherwise, the “open the image” event is processed first, and the test fails.

**Event Order Exploration.** Leveraging these insights, our approach explores possible event execution orders by scheduling an async event in a test run. Our approach first identifies the schedule space for each async event with dynamic analysis and schedules the async event in its schedule space to avoid infeasible orders. To manifest a flaky test failure quickly, our approach schedules events at positions where the test is more likely to fail. Specifically, it schedules an async event at a statement boundary position. The statement boundary position is between the last event that a test statement triggers and the first event that the next statement triggers. We note that the tests are often flaky due to missing an appropriate synchronization operation between two test statements, e.g., the synchronizing operation after clicking a download button. Thus we are more likely to trigger a test failure if an async event is executed after executing certain test statements.

**Instrumentation-free Tool.** We implement our approach into a tool called FlakeScanner, which leverages the debug mode supported in the Android framework to perform dynamic analysis based event scheduling. Thus FlakeScanner requires no instrumentation in either Android apps or the Android framework and works on both Android emulators and devices. According to the study [41], code instrumentation often disrupts test executions and prevents the manifestation of flaky test failures. Moreover, FlakeScanner supports multiple widely-used testing frameworks with which developers write tests such as Espresso [2] or Robotium [13].

**Experiment.** We evaluated FlakeScanner on 33 widely-used Android projects. Our experiments show that FlakeScanner successfully detected 45 out of 52 known flaky tests. On average, it detected a flaky test within three test runs. FlakeScanner outperformed the recently published flaky test detection technique Shaker [40] and the baseline tool RERUN in terms of both the number of detected flaky tests and average execution time. FlakeScanner also detected 245 flaky tests that were previously unknown. Out of these 245 unknown flaky tests, we reported 20 to developers; 13 out of these 20 have been confirmed and addressed by developers.

**Contributions.** Our contributions can be summarized as follows.

- We present an event scheduling approach that explores different event execution orders for each test run to detect concurrency-related flaky tests in Android apps. To avoid exploring all possible event orders, our approach adopts heuristics to identify and explore event orders in which the test flakiness is likely to occur.
- We implement our approach into an instrumentation-free tool that works on Android emulators and physical devices while supporting widely-used Android testing frameworks.
- We curate a suite of subject programs containing 33 widely-used Android projects with developer tests, called FlakyAppRepo for evaluating flaky test detection techniques. To facilitate future research on flaky tests, we make our prototype FlakeScanner and subject-suite FlakyAppRepo available at https://github.com/AndroidFlakyTest.

2 BACKGROUND

![Android concurrency model & testing framework.](image-url)
After obtaining the location data, the async thread updates the result to the UI thread, and then the UI thread updates this result to CaptureLocationActivity. Due to the lack of synchronization between the testing thread and async thread, the async thread might send the location data before or after the testing thread validates it. If the validation occurs before the location data is received, the test fails, else it passes.

Detecting concurrency-related flaky tests is non-trivial as such failures manifest when the events are executed in a specific order. The traditional approach RERUN blindly executes the test many times in the hope of witnessing the flaky test failures. However, this approach becomes ineffective when the environment under which the test runs does not produce the required non-determinism. For example, RERUN failed to witness the flaky test failure in CaptureLocationActivityTest during our experiments. Another approach is to add noise in the execution environment to increase the likelihood of observing such errors. However, such approaches do not proactively explore different event execution orders and are prone to miss crucial event orderings. In our evaluations, the recent (noise-based) related work Shaker [40] failed to detect many confirmed flaky tests, including CaptureLocationActivityTest. We also evaluated EventRacer, a dynamic analysis-based technique for detecting event races in Android apps. However, such approaches are prone to report many races. For example, in our evaluation, EventRacer reported over 200 data races for CaptureLocationActivityTest. Unfortunately, none of these reported data races from EventRacer involve events originated from the testing thread, and hence do not demonstrate the specific flakiness we are illustrating in this section. Besides, EventRacer does not validate whether reported races can cause the test failure.

4 OVERVIEW OF OUR APPROACH

This section describes the main components of our framework and lists the principles that make event exploration suitable for detecting concurrency-related flaky tests. First, the category of flaky tests should be well defined. It is acceptable to give up on all classes of flaky tests and focus on one; concurrency-related flaky tests. Second, using only one test run, we explore (all) possible scheduling space of an async event. Third, we execute these new schedules to observe the flaky behavior, i.e., there exist at least two executions for which a test depicts different outcome (pass/fail).

4.1 Basic Concepts

Before diving into details of our approach, we define a miniature domain-specific syntax extended from [20] for an Android test.

<table>
<thead>
<tr>
<th>Test</th>
<th>$T$</th>
<th>$S$</th>
<th>$\rightarrow$</th>
<th>Assertions</th>
<th>$\text{SYNC}$</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>$e$</td>
<td>$M_{\text{UI thread}}$</td>
<td>$M_{\text{background thread}}$</td>
<td>$M_{\text{OS}}$</td>
<td>$\text{SYNC}(T)$</td>
<td></td>
</tr>
</tbody>
</table>

A test $T$ is composed of a series of program statements $S$. For concurrency-related flaky tests, we are interested in exploring the scheduling space of events, and thus for ease of representation, we categorize a test statement into four parts; statements posting an event, assertion statements, testing framework-specific synchronizing statements, and all other Android-specific statements. An event, denoted as $e$, is a message object created in the UI thread.
Given a test statement \( M \), let \( \mathcal{M} \) be the set of all possible sequences of events created by reordering an async event \( e_{async} \) in \( M \) from the position \( i \) to a new position \( j \). Then, \( \mathcal{M} \) will be infeasible if \( \exists e_k \in \mathcal{M} \) after the position \( j \) and \( e_{async} \) depends on \( e_k \).

Intuitively, \( \text{depends} \) specifies the \textit{Happens-before} relation. In practice, avoiding infeasible event orders would require computing the dependencies among events in the test execution. However, computing event dependencies is challenging for multiple reasons. First, a test can trigger many events (more than 500 events for some cases). Computing dependencies among them can be complex as one event’s execution may depend on multiple other events. Second, event dependencies are not specified in the events but handed over to the components that respond to the events. These components may belong to Android frameworks or third-party libraries, making event dependency analysis difficult. We propose a dynamic analysis to explore various event execution orders to address this challenge. Given an async event \( e \) in a test run, the analysis identifies the scheduling space \( \langle e_{lo}, e_{up} \rangle \), where \( e_{lo} \) is lower bound event and \( e_{up} \) is upper bound event, i.e., \( e \) cannot be executed earlier than \( e_{lo} \) or later than \( e_{up} \) for all execution environments. Consequently, we can explore event execution orders by scheduling the async event \( e \) at various positions between \( e_{lo} \) and \( e_{up} \).

### 4.2 Identifying the Scheduling Space

The successful realization of the scheduling space requires noticing two crucial points. First, the scheduling space needs to reflect some level of determinism, i.e., there need to be some non-async events. Second, each async event’s scheduling space should be well defined.

**Observation 1.** For a statement \( s \) in a test \( T \), the first event it generates will not be an async event, i.e., its order will not change for all runs.

Android is a reactive system, and tests in an Android app invoke the functionality of Android’s components (e.g., an activity) by simulating user events. For example, a test would emulate a button click to invoke some functionality provided by an activity. The concept of background threads in Android is trivial, and Android’s components utilize them to offload long-running tasks in the background. Notably, a background thread is always created by an Android component (via UI Thread). Thus, the interaction from a test (via a statement) to an Android component would always follow a sequence where the test would invoke the component, and then the component may invoke a background thread. Considering such event chaining, the first event of a test statement will always belong to \( M_{UiThread} \) or \( M_{OS} \). Based on this observation, we define anchor events as:

**Definition 4.3 (Anchor Events).** Let \( s \in S \) be a program statement in a test \( T \), and it generates \( n \) events in the following order: \( \langle e_1, ..., e_n \rangle \). Then, we define \( e_1 \) as the anchor event for \( s \).

To localize anchor events, we execute a test, statement by statement; we record all events triggered by each statement, and build
a map between them. According to this map, we identify anchor events for the test. Defining the lower bound of an async event is now straightforward. Intuitively, for a given statement, the lower bound of an async event would always be the anchor event of this statement, as the anchor event will always happen-before async event (Observation 1).

Localizing the upper bound event of an async event $e_{async}$ is more involved and requires identifying events that depend on $e_{async}$, i.e., events that occur only after $e_{async}$. Such event dependence is maintained by thread synchronization operations. For instance, the testing framework Espresso uses `onView()` operation to synchronize the testing thread and app under test. The testing thread waits until specific threads or resources are idle to ensure that a widget specified by `onView()` shows up on the screen. Formally, we define the upper bound of an async event as:

**Definition 4.4 (Upper Bound Event).** Let $T$ be a test containing a set of statements $S$. We define $e_i$ as the upper bound event for an async event $e_{async}$, where $e_j$ is the first event that cannot happen-before $e_{async}$.

We propose a what-if analysis to localize the upper bound event of $e_i$. Specifically, after the test is launched, we hook the async thread that posts $e_i$ at runtime and suspend it, keeping other threads free. Meanwhile, we monitor the testing thread to check at which statement the testing thread stops and waits for the suspended thread to be completed. Suppose that the testing thread stops at a statement $s$ on executing the event $e_j$. We then deem operations in $s$ to depend on $e_i$, and these operations will not be executed until $e_i$ is processed. Thus, the anchor event $e_j$ triggered by $s$ is upper bound event of $e_i$. The idea behind the analysis is what-if it takes forever to complete the long-running task that corresponds $e_j$; operations in a test that depend on $e_i$ will not be executed due to thread synchronization, and those that do not depend on $e_i$ will be executed.

### 4.3 Scheduling Events

Next, we explore the scheduling strategies for an async event $e_{async}$. Given the lower and upper bound events of $e_{async}$, the anchor events that lie between this interval are identified. Then, $e_{async}$ is scheduled before each of these anchor events. To manifest flaky test failures as soon as possible, we prioritize positions closer to the upper bound event (maximizing the runtime of the async event), i.e., we first schedule $e_{async}$ just before its upper bound and then move towards its lower bound.

For two reasons, we only explore positions before anchor events in the scheduling space of an async event instead of all possible positions. First, for non-anchor events in the scheduling space, their execution orders may change from run to run. Using them as hooks for event scheduling may introduce infeasible orders. Second, anchor events are boundary events of test statements. Scheduling an async event before them is likely to trigger a flaky test failure.

### 5 METHODOLOGY

Figure 2 shows the workflow of our approach. For a given test, our framework performs a concrete execution to trace all the generated events. Next, a map between the statements in the test and their corresponding events is created. The test is executed multiple times to compute possible schedules for async events. Consequently, a set of event orders that might occur in execution environments are created. Finally, we explore these possible event execution orders; if a test failure is detected during the exploration, the test is identified as a flaky test (since we have already seen passing runs of the test).

#### 5.1 Event Tracing and Mapping

Event tracing is often used in the dynamic analysis of Android apps. It can be achieved by simply logging events that are generated at runtime. However, such techniques cannot fulfill our task. Event information (e.g., event id) produced in logs is dynamically generated by Android runtime and changes in each run. Our approach requires an event identifier to identify an event uniquely across different test runs. Async events that are identified during event tracing need to be hooked and scheduled in runs that are performed for event order exploration.

**Event Identification.** We identify an event based on interactions between the event and app under test at runtime. Two events triggered in different test runs are considered identical if: (1) they are triggered by the same test statement; (2) they are processed by the same sequence of methods at runtime. For instance, a `pressDown` event is associated with an identifier constructed using the line number of the statement that triggers the event and signatures of a sequence of methods that process it. This practice of event identification comes from our investigation of the Android framework.

**Tracing and Mapping.** Algorithm 1 outlines the procedure of event tracing and mapping. First, it launches the app under test and takes control of the Android runtime with a module called `ARThandler`. Given a test $T$ consisting of a series of statements, ARThandler runs the test $T$ in the testing thread and executes statements one by one (via the Android debugger). When one statement is executed, ARThandler monitors the event queue of the UI thread and hooks injected events. For each event, ARThandler records the tuple $(isAsync, sq_m)$ where $isAsync$ denotes whether it is an async event and $sq_m$ denotes the signatures of a sequence of methods that have processed the event. This tuple and the line number of the statement being executed form the event’s identifier and get stored in a list (Line 11-14). As stated before, a statement in the test might perform long-running tasks, which are executed in async threads. When these tasks are completed is non-deterministic, and the async event might be posted after a long time. To not miss async events that are triggered by a single statement, we keep hooking events until two criteria are satisfied: (a) there are no new events and (b) the event queue of the UI thread is empty, which often indicates the system is not running tasks. This practice is also used in the
To compute possible event execution orders, we perform a list of steps to analyze the event dependency. For each async event in the event trace, we start by hooking the async event and then explore event orders during the event trace. This process takes the generated list of statements as input. For each statement, we identify the event dependencies and determine which event can be executed before a statement is reached. Once statement $m$ is reached, we suspend the testing thread and free the suspended thread to post async events.

### 5.2 Identifying Event Schedule Space

To compute possible event execution orders, we perform a what-if dynamic analysis to resolve event dependencies caused by thread synchronization between apps and testing frameworks. Algorithm 2 shows the procedure of resolving event dependencies. It takes the event trace $List$ generated in the previous step as input. For each async event $e_{async}$ in $List$, the algorithm launches the test and starts to hook event $e_{async}$ (Line 4-8). Once hooked, the algorithm suspends the thread that posts $e_{async}$ such that $e_{async}$ cannot be posted (Line 9). Meanwhile, it keeps checking the status of the testing thread (Line 10-14). If the testing thread’s status is WAITING, it considers the testing thread is performing thread synchronization with threads in the app and waiting for $e_{async}$ to be executed. Thus, we consider the statement $s$ that is being executed in the testing thread attempts to trigger an event (e.g., $e_j$) which depends on $e_{async}$. Therefore, the schedule space of $e_{async}$ is bounded by $e_j$, i.e., the first event that is triggered by $s$. So statement $s$ is identified as the upper bound of schedule space of async event $e_{async}$. Statement $s$ is recorded and set as the upper bound of event $e_{async}$ and $e_{async}$ is restored to $List$. Finally, schedule spaces for all async events in $List$ are recorded.

### 5.3 Scheduling Events

Schedule space of each async event in the event trace $List$ is identified in the previous steps. Next, we explore event orders during test execution. An async event $e_{async}$ can be simply represented by a triplet $(id, n, m)$, where $n$ and $m$ are bounds of the scheduling space of $e_{async}$. Specifically, $n$ is the index of the statement in the test that triggers $e_{async}$, and $m$ is the index of the statement that triggers the upper bound event of $e_{async}$.

Similar to scheduling space identification, we can schedule $e_{async}$ by operating threads. We first hook event $e_{async}$ after the test is launched and suspend the thread that posts $e_{async}$. Then, we free the testing thread and monitor whether the statement being executed is statement $m$. Once statement $m$ is reached, we suspend the testing thread and free the suspended thread to post $e_{async}$. After the async thread is terminated or idle, i.e., event $e_{async}$ has been posted, we free the testing thread. In such a way, event $e_{async}$ can be executed before statement $m$. In next test run, we schedule $e_{async}$ to be executed prior to statement $m - 1$, until all statements between $n$ and $m$ are explored. This procedure is repeated for each async event in $List$.

### 6 IMPLEMENTATION

Our system is implemented in Scala and runs on a computer that connects a physical Android device or an emulator. Unlike existing techniques, it requires no instrumentation on apps or the Android framework and can be adapted to different versions of Android.

**Taking Control of Android Runtime.** The Android framework supports running an app in the debug mode, under which the Android runtime can be fully controlled. Specifically, we connect the Android runtime via Android Debug Bridge (ADB) and use the Android debugger to execute the app under test. With the debugger’s help, we can perform execution step by step and monitor the app state during thread manipulation.

**Hooking Events.** Android adopts the event-driven model and manages events using an event queue. Each event implements an interface method called enqueueMessage() and Android puts
With the Android debugger, we can execute a testing thread step by step and observe execution at the statement level by inserting a breakpoint at each statement in the test. It also allows us to obtain the current stack frames of a running thread that are used for event identification. Operating Threads. We leverage the Android debugger to retrieve threads running in the app under test, including the testing thread and UI thread. The Android debugger provides commands to remotely operate threads, e.g., inspecting thread status and suspending a selected thread. It also allows us to obtain the current stack frames of a running thread that are used for event identification. With the Android debugger, we can execute a testing thread step by step and observe execution at the statement level by inserting a breakpoint at each statement in the test.

7 EVALUATION

We empirically evaluate FlakeScanner’s effectiveness in detecting flaky tests in the test suites of large-scale Android projects.

7.1 Subject Apps

Our technique is designed for a test that runs on the Android framework platform and checks whether it is flaky. To evaluate our technique, we need Android projects that contain such tests. Basically, there are two types of tests in Android projects: instrumented tests that run on a physical device or Android emulator and local unit tests that run on local Java virtual machines. Thus, we need Android projects that contain instrumented tests. Unfortunately, most Android projects in existing benchmarks such as AndroTest [11], industrial app benchmark [44], and data set that is used for flaky test empirical study [42] have no instrumented tests.

Therefore, we built the first subject-suite FlakyAppRepo in which each Android project contains instrumented tests that are written by developers. As shown in Table 1, FlakyAppRepo contains 33 Android projects (majority of them are well-known, such as WordPress) and over 5000 instrumented tests from developers. We collected Android projects containing instrumented tests as follows: (1) searching popular open-source Android projects such as Firefox and manually checking their repositories to select projects with instrumented tests (intuitively, popular projects are well maintained and more likely have developer tests); (2) searching keywords such as flak, flakiness, or intermit on the Github and manually checking searched repositories to select Android projects containing instrumented tests. Furthermore, we explored these keywords in the commit history as well to include already fixed flaky tests. A few searched projects have 0 star on the Github. To achieve diversity, we did not exclude these projects and kept them in the subject-suite.

7.2 Research Questions

Our evaluation aims to address the following research questions.

**RQ1** Can FlakeScanner detect known flaky tests that are reported by developers?

**RQ2** How does FlakeScanner compare with existing techniques in terms of number of detected flaky tests?

**RQ3** Can FlakeScanner be used to discover flaky tests in Android projects that were previously unknown?

7.3 Experiment Setup

To answer the research questions, we conduct three empirical studies on test cases that are written by developers in real world Android app projects.

7.3.1 Study 1.

We first evaluate FlakeScanner’s effectiveness in flaky test detection by running it on known flaky tests in Android projects and checking how many of them are marked as flaky tests by FlakeScanner.
Table 3: Dynamic analysis based tools for detecting flaky tests or concurrency issues in Android apps.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Categories</th>
<th>Instrumentation</th>
<th>Year</th>
<th>Reason for not selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>RERUN</td>
<td>Flaky tests</td>
<td>No</td>
<td>2016</td>
<td></td>
</tr>
<tr>
<td>Shaker [40]</td>
<td>Flaky tests</td>
<td>No</td>
<td>2020</td>
<td></td>
</tr>
<tr>
<td>APEChecker [16]</td>
<td>Async bugs</td>
<td>No</td>
<td>2018</td>
<td>Publicly unavailable</td>
</tr>
<tr>
<td>ERVA [20]</td>
<td>Event race</td>
<td>Yes</td>
<td>2016</td>
<td>Publicly unavailable</td>
</tr>
</tbody>
</table>

Data Set. We selected 52 known flaky tests from FlakyAppRepo that are caused by concurrency or synchronization issues for Study 1. In Android projects, flaky tests reported by developers are marked with a dedicated annotation @FlakyTest such that flaky tests can be automatically filtered out during test execution. We identified 269 known flaky tests in FlakyAppRepo using @FlakyTest annotation. We selected known tests from them in the following manner.

Concurrency. We select concurrency related flaky tests by manually analyzing (1) details about reported flaky tests such as the reason of why a test is flaky, which can be collected from detail element of @FlakyTest annotation; (2) commit messages when flaky tests were introduced or fixed in a subsequent version; (3) whether there exist statements in tests that are commonly involved in concurrency or synchronization execution, e.g., operations of Runnable or AsyncTask objects.

Passing. We ran each selected test in our execution environment and ensured it passes in the initial run, avoiding failures due to environment setup.

Reproducing. We reproduce the flaky test failure for each selected test by analyzing its commit messages and detailed description in the repository including root causes and failing scenarios. For cases lacking the description, we contacted developers to seek more insights on the failure reproduction. A test is selected if its failures is reproduced.

In the end, 52 known flaky tests were selected for the evaluation. Meanwhile, we also analyzed root causes of these flaky test failures and classified them into four categories which are shown in Table 2.

7.3.2 Study 2.

To answer RQ2, we evaluate FlakeScanner and existing tools on the data set used in Study 1. We reviewed most recent works on flaky tests or event race detection for Android apps and summarize them in Table 3. We selected the following approaches for comparison.

- **RERUN.** In practice, developers often run a test many times to check whether the test is flaky. Similarly, we run a test 100 times and check whether the flaky test failure manifests during execution. We call the approach 100RUN and take it as our baseline.
- **Shaker [40]** is the most recently reported state-of-the-art technique for detecting concurrency-related flaky tests in Android apps. **Shaker** attempts to manifest a flaky test failure by adding noises in the execution environment to affect event execution order, e.g., changing CPU workload.
- **EventRacer [8]** is a state-of-the-art dynamic technique of detecting event races in Android apps, which infers races by analyzing runtime traces produced by an instrumented Android framework.

7.3.3 Study 3.

To answer RQ3, we run FlakeScanner on tests in FlakyAppRepo without @FlakyTest annotation, i.e., tests that are not reported as flaky tests. We first run these tests in our execution environment and exclude tests that fail, then feed passing tests to FlakeScanner and report tests that are labeled as flaky tests by FlakeScanner.

Effect of debug mode. FlakeScanner runs apps in the debug mode. The debug mode environment may affect flaky test detection by impacting the Android event dispatching mechanism or increasing execution workload. According to the official Android documentation, the difference between normal execution mode and debug mode is that the Android virtual machine (VM) loads an additional Android Runtime Tooling Interface (ARTTI) plugin that exposes runtime internals (e.g., variable values). The ARTTI plugin runs as a VM-level component and it is transparent to event dispatching that occurs in the app’s main thread, i.e., running apps in debug mode does not impact the mechanism of event dispatching.

However, running the ARTTI plugin and accessing runtime internals in the execution could increase the workload of the VM. Due to the workload change, a test can manifest the flaky test failure. To rule out such cases, for each test in our study, we pre-execute it to ensure the test passes in the debug mode environment. To perform end-to-end comparison in Study 2, we run 100RUN and Shaker in the debug mode environment as well.

Execution Environment. We conducted experiments on a physical machine with 64 GB RAM and a 56 cores Intel(R) Xeon(R) E5-2660 v4 CPU, running a 64-bit Ubuntu 16.04 operating system. Each execution instance runs in a Docker container to minimize the potential interference between running instances. App under test runs on an Android 9 (x86) emulator. One execution instance is for one test case for which the Android emulator is initialized to a fresh state at the beginning to provide a clean testing environment.

7.4 RQ1: Efficacy

Table 4 shows results on each known flaky test for study 1. The first column indicates known test IDs, the second column shows app names and testing frameworks used in apps, and the third column indicates test method names. Column “#Event” indicates the number of events observed by FlakeScanner during detection. “#Run” indicates the number of test runs in the event order exploration. Column “Time” reports the time that is used to detect a flaky test. Column “Succ” indicates whether the test is identified as a flaky test by FlakeScanner. “-” indicates that a test is unable to be executed due to library compatibility issues. Firefox-Lite marked with “+” (rows 39-40) is a previous version (commit:465739510e). Note that some test names in the table are shortened for readability.

Results. FlakeScanner successfully detected 45 out of 52 known flaky tests that are from 10 Android projects (including a different version of Firefox Lite). On average, FlakeScanner detected a flaky test in 101 seconds. The maximum detection time is for test 20 from app AntennaPod (691 seconds). The minimum detection time is for
Table 4: Results on known flaky tests by FlakeScanner, Shaker and 100Run.

<table>
<thead>
<tr>
<th>Id</th>
<th>App: Framework</th>
<th>Method name</th>
<th>FlakeScanner</th>
<th>Shaker</th>
<th>100Run</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Events</td>
<td>Run</td>
<td>Time(s)</td>
</tr>
<tr>
<td>1</td>
<td>Surveyor: Espresso</td>
<td>capture</td>
<td>58</td>
<td>2</td>
<td>48</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>twoQuestions</td>
<td>104</td>
<td>2</td>
<td>48</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>multimedia</td>
<td>233</td>
<td>8</td>
<td>542</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>contactDetails</td>
<td>104</td>
<td>2</td>
<td>48</td>
</tr>
<tr>
<td>5</td>
<td>Youtube: JUnit4</td>
<td>testEncipheredVideo</td>
<td>4</td>
<td>2</td>
<td>24</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>testUnembeddable</td>
<td>7</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>testAgeRestrictVideo</td>
<td>5</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>testUsualVideo</td>
<td>5</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>MyExpenses: Espresso</td>
<td>testScenarioForBug5b..</td>
<td>861</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>editCommandKeeps..</td>
<td>101</td>
<td>2</td>
<td>54</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>cloneCommandIncreases..</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>changeOfFractionDigits..</td>
<td>991</td>
<td>2</td>
<td>170</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>changeOfFractionDigitsWith..</td>
<td>991</td>
<td>2</td>
<td>174</td>
</tr>
<tr>
<td>14</td>
<td>Firefox Lite: Espresso</td>
<td>saveImageThenDelete..</td>
<td>627</td>
<td>5</td>
<td>352</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>dismissMenu</td>
<td>165</td>
<td>2</td>
<td>76</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>turnOnTurboMode..</td>
<td>89</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>changeDisplayLang</td>
<td>189</td>
<td>3</td>
<td>72</td>
</tr>
<tr>
<td>18</td>
<td>AntennaPod: Robotium</td>
<td>testGoToPreferences</td>
<td>107</td>
<td>3</td>
<td>56</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>testClickNavDrawer</td>
<td>132</td>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>PlaybackSonicTest#.On..</td>
<td>177</td>
<td>10</td>
<td>691</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>PlaybackSonicTest#.Off..</td>
<td>176</td>
<td>6</td>
<td>495</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>PlaybackTest#.Off.Episodes</td>
<td>175</td>
<td>2</td>
<td>142</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>PlaybackTest#.On.Episodes</td>
<td>162</td>
<td>2</td>
<td>140</td>
</tr>
<tr>
<td>24</td>
<td>FlexBox: Espresso</td>
<td>testScrollToPosition_row</td>
<td>219</td>
<td>6</td>
<td>306</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>testAddViewHolders..</td>
<td>46</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>26</td>
<td></td>
<td>testChangeAttributes..</td>
<td>23</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td>testMinHeight..</td>
<td>58</td>
<td>7</td>
<td>112</td>
</tr>
<tr>
<td>28</td>
<td></td>
<td>testJustifyContent..</td>
<td>58</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td>testJustifyContent_center</td>
<td>58</td>
<td>9</td>
<td>158</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>testFlexWrap..</td>
<td>55</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>31</td>
<td></td>
<td>testFirstViewGone..</td>
<td>58</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>32</td>
<td></td>
<td>testChangeOrder..</td>
<td>26</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>33</td>
<td></td>
<td>testAlignItems..</td>
<td>58</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>34</td>
<td></td>
<td>testAlignContent..</td>
<td>58</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>35</td>
<td></td>
<td>testAlignContent..</td>
<td>58</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>36</td>
<td></td>
<td>testAlignContent.. Padding</td>
<td>58</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>37</td>
<td></td>
<td>testFlexLines..row</td>
<td>74</td>
<td>3</td>
<td>32</td>
</tr>
<tr>
<td>38</td>
<td></td>
<td>testFlexLines..column</td>
<td>69</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>39</td>
<td>Firefox Lite(★): Espresso</td>
<td>browsingWebsite..</td>
<td>642</td>
<td>2</td>
<td>94</td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>saveImageThenDelete..</td>
<td>661</td>
<td>3</td>
<td>96</td>
</tr>
<tr>
<td>41</td>
<td>BackPack: JUnit4</td>
<td>test_with_description</td>
<td>41</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>42</td>
<td></td>
<td>test_with_title</td>
<td>40</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>43</td>
<td></td>
<td>test_bottom_sheet_style</td>
<td>32</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>44</td>
<td></td>
<td>test_alert_style</td>
<td>32</td>
<td>4</td>
<td>31</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td>screenshotTestDialog..</td>
<td>86</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>46</td>
<td></td>
<td>screenshotTestDialog</td>
<td>86</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>47</td>
<td></td>
<td>test_with_buttons</td>
<td>42</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>48</td>
<td>Barista: JUnit4</td>
<td>overflowMenuClick_byTitle</td>
<td>130</td>
<td>3</td>
<td>82</td>
</tr>
<tr>
<td>49</td>
<td></td>
<td>openOverflowMenu__Option</td>
<td>58</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>overflowMenuClick_byId</td>
<td>127</td>
<td>4</td>
<td>82</td>
</tr>
<tr>
<td>51</td>
<td>Kaspresso: JUnit4</td>
<td>CommonFlakyTest#test</td>
<td>93</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>52</td>
<td></td>
<td>UiCommonFlakyTest#test</td>
<td>94</td>
<td>5</td>
<td>295</td>
</tr>
<tr>
<td>Avg/Sum</td>
<td></td>
<td></td>
<td>169</td>
<td>3</td>
<td>101</td>
</tr>
</tbody>
</table>
test 41, 42, and 47 from app BackPack (11 seconds). \textit{FlakeScanner} detected a flaky test within 3 test runs on average. The maximum number of test runs is for test 20 from AntennaPod (10 runs). To understand why \textit{FlakeScanner} successfully detected failures within a few test runs, we inspected source code of projects. The major reason is that tests heavily use synchronization operations. For instance, test 2 from Surveyor contains 21 test statements, 15 of which use the synchronization operation provided by Espresso [2] \texttt{onView()}. When \texttt{onView()} is called in the test statement, the testing thread waits until background threads complete tasks. As designed, \textit{FlakeScanner} does not perform event scheduling for those cases since async events are bounded in the execution of a test statement by the synchronization operation. Then \textit{FlakeScanner} can focus on scheduling events for statements that lack synchronization operations. Moreover, \textit{FlakeScanner} prioritizes exploring positions that are closer to the upper bound event, which likely triggers flaky test failures (Section 4.3). Thus, \textit{FlakeScanner} could detect failures within a few test runs. The results also show \textit{FlakeScanner} is a practical tool. It successfully detected flaky test failures for test 12 and 13 from MyExpenses, for which 991 events were generated in the execution. Meanwhile, \textit{FlakeScanner} worked on Android projects that adopt different testing frameworks such as Espresso [2] and Robotium [3].

\textit{FlakeScanner} successfully detected 45 out of 52 known flaky tests in 10 Android projects. On average, it detected a flaky test within 3 test runs.

7.5 RQ2: Comparison with Existing Techniques

As shown in Table 4, \textit{FlakeScanner} outperforms Shaker and 100RUN in terms of both the number of detected flaky tests and the average execution time. Out of the 52 known flaky tests, \textit{FlakeScanner} detected the most flaky tests (45) and is followed by Shaker (15) and 100RUN (8). For execution time, \textit{FlakeScanner} detected a flaky test in 101 seconds on average, which is less than 861 seconds of Shaker and 497 seconds of 100RUN. For most of tests, 100RUN could not detect a flaky test failure in the first few runs and kept executing them until reaching 100 times, which took a longer time.

Regarding the overlap between flaky tests detected by each tool, \textit{FlakeScanner} detected all the flaky tests detected by Shaker and 100RUN. But Shaker and 100RUN failed to detect the other flaky tests that were detected by \textit{FlakeScanner}. The better results from \textit{FlakeScanner} can be explained as follows: \textit{FlakeScanner} can identify synchronization operations in the test execution and focus on scheduling events for statements that lack synchronization operations. Unexpected event execution orders that cause flaky test failures are more likely to explored by \textit{FlakeScanner}.

We also evaluated EventRacer since it uses dynamic analysis to infer event race for Android apps. EventRacer reported many possible races for each test run. On average, it reported 1237 races for a test. This is because EventRacer focuses on detecting races in Android apps and does not analyze races in which testing frameworks are involved. Furthermore, EventRacer infers races by analyzing recorded traces and cannot validate whether the reported races can cause flaky test failures.

7.6 RQ3: Real-world Flaky Test Detection

We ran 1444 passing tests from the 33 Android projects in FlakyAppRepo which are not annotated as flaky tests (these tests may or may not be flaky). Out of these 33 projects, \textit{FlakeScanner} detected at least one flaky test for 19 projects, and reported 245 flaky tests in total. To validate previously unknown flaky tests that \textit{FlakeScanner} detected, we randomly selected 20 out of the detected 245 flaky tests and reported them to developers. For each selected test, we manually reproduced the failure that \textit{FlakeScanner} witnessed during detection and generated a detailed root-cause-analysis report, and submitted the report on the Github. At the time of writing the paper, we got responses on 15 test cases. Out of the 15 tests, 13 were confirmed as flaky tests and addressed by developers. For the other two tests, developers replied that the reported failures were not encountered yet or not reproduced at their end, without giving us further explanation. Our experience with flaky test reporting shows so far that developers are more interested in identifying which tests are flaky. Once a flaky test is detected, they appear to be more prone to removing them, rather than investigating why it is flaky.

\textit{FlakeScanner} detected 245 previously unknown flaky tests in 19 widely-used Android projects. Out of the reported 20 unknown flaky tests, 13 were confirmed and addressed by developers.

7.7 Threats to Validity

\textit{External Validity}: Threats to external validity relate to the generalizability of the experimental results. \textit{FlakeScanner} is evaluated so far on 33 Android projects. Our results may not generalise beyond the 33 Android projects to which we have applied \textit{FlakeScanner}. To mitigate this threat, we not only choose Android projects that are popular and well maintained but also include less popular Android projects (i.e., less stars on the Github) which were searched on the Github via keywords.

\textit{Internal Validity}: Threats to internal validity concern factors in our experimental methodology that may affect our results. In Study 1, we note that 52 concurrency or synchronization related known flaky tests are chosen by manually analyzing their related descriptions and commit messages, which might result in selection bias. Similarly, we manually analyze failures detected by each tool under evaluation and validate the results, which might introduce bias as well. To mitigate these risks, two authors of this paper independently performed the manual tasks, and cross-checked each other’s results.

7.8 Data Availability

To facilitate future research on flaky tests, we make our prototype \textit{FlakeScanner} and subject-suite FlakyAppRepo available at link \url{https://github.com/AndroidFlakyTest}
8 RELATED WORK

Flaky Test Detection. Bell et al. [6] use code coverage based differential analysis to identify flaky tests. A test is deemed flaky if it fails during regression testing and its execution does not reach any code that was recently changed by developers. Lam et al. propose iDFlakies [27], a framework for detecting order dependent flaky tests, tests that fail when run in different orders. Dutta et al. [13] develop an approach to detect random number related flaky tests, tests that fail due to difference in the sequence of random numbers generated in different runs. Shi et al. [39] propose an approach to fix order-dependent flaky tests by leveraging passing tests. Shi et al. [37] propose to rerun a test multiple times on each mutant and obtain reliable coverage results such that the effects of flaky tests on mutation testing can be mitigated. In contrast, FlakeScanner detects concurrency or synchronization related flaky tests in Android projects by exploring feasible event execution orders.

Event Race Detection. Another branch of works that are close to ours is event race detection [8, 18–21, 31–33, 35, 36, 45]. Instead of detecting flaky tests, these works leverage dynamic and static analysis to detect event races. For instance, ER Catcher [35], DROIDRACER [32], EventRacer [8], CAVA [19], and nAdroid [18] capture happens-before-relation among events and infer possible event races. In addition, Ozkan et al. [24] propose to detect asynchronous bugs by exploring different execution orders of event handlers in Android apps. SARD [46] leverages happens-before analysis to detect use-after-free issues in Android apps. These techniques have the potential to be applicable for flaky test detection, but face challenges to capture complete and precise happens-before relations when a test is executed by a testing framework. Many false positives can be reported by event-race detectors due to incomplete happen-before relations being compute. In contrast, FlakeScanner performs a system-level dynamic analysis to capture precise event dependencies to avoid such false positives.

Empirical Studies on Flaky Tests. Multiple studies [14, 26, 30, 42] confirm concurrency as the major cause of flaky tests. Luo et al. [30] performed an empirical analysis of flaky tests in 51 open-source projects. They identified Concurrency and Async wait as the most common cause of flaky tests. They pointed out that the majority of these cases arose because they do not wait for external resources. Finally, they described the common fixing strategies the developers use to fix flaky tests. In a separate study, Eck et al. [14] surveyed 21 professional developers to classify 200 flaky tests they fixed. They identified four unreported causes of flaky tests, which are also considered difficult to fix. Thorve et al. [42] conducted an empirical study of flaky tests in Android apps. They searched 1000 projects for the commits related to flakiness and found only 77 relevant commits from 29 projects. They found 36% of commits occurred due to concurrency related issues. Fan et al. [16] proposed a hybrid approach towards manifesting asynchronous bugs in Android apps with fault patterns.

Concurrency Bug Detection. There have been several testing based approaches [10, 12, 22, 23, 28, 47] to identify concurrency related bugs. Maple [47] proposed a coverage-driven approach to expose untested thread interleavings. Letko [28] proposed a combination of testing and dynamic analysis with metaheuristic techniques. Choudhary et al. [10] presented a coverage-guided approach for generating concurrency tests to detect bugs in thread-safe classes. Multiple related works [5, 7, 9, 15, 29, 43] manipulated event orders to control non-determinism in multi-threaded programs. Liu et al. [29] proposed a deterministic multithreading system that replaces pthreads library in C/C++ apps. Emmi et al. [15] proposed a search prioritization strategy to discover concurrency bugs. They add non-determinism to deterministic schedulers by delaying their next-scheduled task. Adamsen et al. [5] presented an automated program repair technique for event race errors in JavaScript. Given a repair policy, they controlled the event handler scheduling in the browser to avoid bad orderings.

9 DISCUSSION

Flaky tests pose a significant problem in validating mobile apps. In this paper, we presented an approach for detecting flaky tests via systematic event order exploration. We introduced FlakeScanner, a tool to detect flaky tests for Android apps. FlakeScanner explores the space of possible execution environments which may cause relevant threads to interleave differently. Due to the lack of a testing benchmark for flaky tests, we created the first subject-suite FlakyAppRepo that is used to study test flakiness. FlakyAppRepo contains 33 widely-used Android apps with around 2.5k stars on average in GitHub. We applied FlakeScanner to tests from FlakyAppRepo. Results show that FlakeScanner not only detected known flaky tests but also reported 245 new flaky tests. We believe that our tool and results hold out promise for tackling flaky tests, which is a significant pain point in the practice of testing.

ACKNOWLEDGEMENTS

This work was partially supported by the National Research Foundation Singapore (National Satellite of Excellence in Trustworthy Software Systems).

REFERENCES


Yongjian Hu, Iulian Neamtiu, and Arash Alavi. 2016. Automatically verifying...


