Orca: Differential Bug Localization in Large-Scale Services

Ranjita Bhagwan  Rahul Kumar  Chandra Sekhar Maddila  Adithya Abraham Philip  
Microsoft Research India

Abstract
Today, we depend on numerous large-scale services for basic operations such as email. These services are complex and extremely dynamic as developers continuously commit code and introduce new features, fixes and, consequently, new bugs. Hundreds of commits may enter deployment simultaneously. Therefore one of the most time-critical, yet complex tasks towards mitigating service disruption is to localize the bug to the right commit.

This paper presents the concept of differential bug localization that uses a combination of differential code analysis and software provenance tracking to effectively pin-point buggy commits. We have built Orca, a customized code search-engine that implements differential bug localization.

On-Call Engineers (OCEs) of Orion, a large enterprise email and collaboration service, use Orca to localize bugs to the appropriate buggy commits. Our evaluation shows that Orca correctly localizes 77% of bugs for which it has been used. We also show that it causes a 3x reduction in the work done by the OCE.

1 Introduction
Orion is a large enterprise email and collaboration service that supports several millions of users, runs across hundreds of thousands of machines, and serves millions of requests per second. Thousands of developers contribute code to it at the rate of hundreds of commits per day. Dozens of new builds are deployed every week. Software bugs are bound to be common in such a complex and dynamic environment. It is critical to detect and promptly localize such bugs since service disruptions lead to customer dissatisfaction and significantly lower revenues [21].

When a service disruption happens because of a software bug, the first-step towards mitigating its effect is to localize the responsible bug to the right commit. We call this commit-level bug localization. This is a non-trivial task since the intense pace of development demands that multiple commits be aggregated into a single deployment. In addition, commit-level bug localization needs to happen as quickly as possible so that buggy commits can be reverted promptly thereby restoring service health. About half of all Orion’s service disruptions are caused by software bugs.

Unfortunately, bug localization in large services such as Orion is a cumbersome, time-consuming, and error-prone task. The On-Call Engineers (OCEs) are the first points-of-contact when a disruption occurs, and they are responsible for bug localization. Though knowledgeable, on-call engineers can hardly be expected to have complete and in-depth understanding of all recent commits. Moreover, bugs that emerge after deployment are complex and often non-deterministic. And yet, very few tools exist to enable OCEs to perform this critical task.

Our goal is to build a tool that will help OCEs correctly and swiftly localize a bug to the buggy commit. Over a period of eight months, we studied post-deployment bugs, their symptoms, the buggy commits that caused them, and the current approach to bug localization that Orion’s OCEs follow. We made four key observations.

1) Bug localization is time-critical, bug fixing is not. When a bug disrupts a service, the OCE’s task is to keep the service disruption time to a minimum. She finds the buggy commit as fast as possible and reverts it rather than wait for the concerned developer to fix the bug. The reason is that, depending on the complexity of the bug, the developer may take a long time to fix it. Therefore to keep disruption to a minimum, it is better to revert the buggy commit first and introduce the fix at a later time. Thus, fast commit-level bug localization is critical.

2) Rich monitoring infrastructure exists but is insufficient because of uncaptured dependencies. Since service disruptions are a major concern, developers have created thousands of active probes that periodically monitor service-components or API calls and raise alerts if they fail. Despite this, bug localization is a challenge because a probe to a component may fail not because of any change to the component itself, but because of a change to another component that depends upon it. For instance, a server-side probe failed with an exception Type RecipientId not supported because a developer made a commit to client-side code that added support for the data type RecipientId without adding support on the server-side. To make matters worse, as the service evolves fast, new dependencies emerge at a rapid
rate and no tool can completely capture all of them.

3) Symptom descriptions and their causes tend to have similar terms. We have found that when a probe detects a bug, a similarity often exists between terms in the unhealthy probe name or exception text that it generates and the source-code change that caused the bug. In the example mentioned in the previous paragraph, the term recipient occurs in both symptom (the exception text) and cause (added support to the data type on the client). We also see this similarity in some customer complaints as well which predominantly use natural language. For instance, a customer recently complained that “Email ID suggestions for people I know is not working.”. The cause for this was an incorrect change to a function named PeopleSuggest.

4) Bugs may appear well after the buggy commit is deployed. We observed that while the symptoms of a bug appear in a current build, the cause may be a commit deployed in a much older build. These are particularly challenging for the OCE to localize because they have to investigate, in the worst-case, all commits in the current build before moving on to investigate an older build. This can significantly lengthen service disruptions.

Keeping these observations in mind, we design a novel search technique that we call differential bug localization. Using descriptions of the bug as a query, we detect changes to the abstract syntax tree in the source-code and search only these changes for text-based similarity. We call this differential code analysis. To find offending commits in older builds, we introduce a construct called the build provenance graph that captures dependencies between builds. We designed Orca, a custom code search-engine that leverages differential bug localization to provide a ranked list of “suspect” commits.

The Orion service has integrated Orca into its alerting and monitoring processes. This paper describes Orca and makes the following contributions:

- We provide a study of post-deployment bugs found in the Orion service and their characteristics (Section 3).

- We introduce differential bug localization, which uses two constructs: differential code analysis of the abstract syntax tree, and the build provenance graph. In addition, we use a prediction of commit risk to call out riskier commits in the list of potential root-causes (Section 4).

- We have designed Orca, a tool that Orion’s OCEs are actively using to localize bugs (Section 5).

- We provide an evaluation of Orca for bugs found in the Orion service (Section 6).

To the best of our knowledge, ours is the first study of a bug localization tool deployed on a large-scale enterprise service. We have evaluated Orca on 48 post-deployment bugs found in Orion since October 2017. We show that Orca correctly localizes 37 out of 48 bugs for a recall of 77%. In 30 of the 37 cases, the correct commit was ranked in the top 5 records shown by our UI (Section 6). We also show that Orca causes a 3× reduction in the work done by the OCE.

We have designed Orca for usability and ease-of-adoption. While this paper concentrates on Orion’s deployment of the tool, Orca has been deployed on five other services within our enterprise. Our techniques are generic and extend well to other large services. To make it easy for OCEs to use the tool interactively, we have optimized its performance through multiple caching and preprocessing techniques. Our user-interface provides results with an average run-time of 5.9 seconds per query.

2 Related Work

The Programming Languages, Software Engineering and Systems communities have extensively studied bug detection, bug localization and debugging. While Orca takes inspiration from some of this prior work, it targets a fundamentally different application space, i.e. large-scale service deployments. Also, Orca is meant to be used by on-call engineers, not developers.

A bug localization tool for such services needs to be fast: the query response time should be at most a few seconds since OCEs will use the UI interactively. It should be general: the techniques should support code in different languages and should not need an OCE or developer to provide specification. It should be non-intrusive: we should not require any changes to the service’s existing coding and deployment practices. Finally, it should be adaptive: it should work in an extremely dynamic and changing environment. We now describe prior work in the area and explain why it does not satisfy some or all of these requirements.

IR-based Bug localization techniques [3, 25, 35, 36, 38, 42, 49] use a given bug-report to search, based on textual similarity, for similar bug reports in the past. For each match, they localize the bug at the file-level, not at commit-level, to files that have been changed to address similar bug-reports. Wang et al. [38] present a structured and detailed study of the various techniques that are used for information-retrieval based bug localization. They use similar search techniques over five major concepts: version history, bug reports, stack traces, and reporter information.

While these techniques are fast, general and non-intrusive, they assume a stationary system, i.e. they as-
Moreover, a bug in a large deployment may not be reproducing, this may require a large number of testing cycles. Given the last good commit, it uses an algorithm based on instrumented programs. Hence, it is intrusive.

Dynamic Analysis techniques are the most commonly used and widely studied approaches for detecting bugs and issues in software. Testing and automated fuzz testing techniques [6, 18] provide an effective method to automatically generate test-cases that produce random inputs for the underlying software. Although these techniques are useful, they are complementary to our approach. Testing is never complete and does not provide guarantees about the correctness of the software. As a result, in spite of comprehensive testing, bugs still emerge regularly in service deployments, as we have noticed with Orion.

To find post-deployment bugs, previous work has presented statistical techniques [7, 9, 26] to automatically isolate bugs in programs by tracking and analyzing successful and failed runs of the program. While these techniques hold promise, they are intrusive: they require fine-grained instrumentation and a large number of program traces from the service deployment. Given the stringent performance needs and dynamism in services, we do not have the luxury of utilizing such techniques.

Delta Debugging techniques [12, 45] help automate the problem of debugging by providing the debugger or programmer information about the state of the program in passing and failing runs. The possible search space of the root cause is systematically pruned by using information from the various runs and by creating new executions. The ideas in delta debugging rely heavily on program slicing [1, 39]. Given its requirements, delta debugging tends to require a large number of tests and the data from instrumented programs. Hence, it is intrusive.

The GIT bisect command [17] is similar in flavor. Given the last good commit, it uses an algorithm based on binary search to go through subsequent commits, repeatedly test the code, and ultimately localize the problem to the buggy commit. However, like delta debugging, this may require a large number of testing cycles. Moreover, a bug in a large deployment may not be reproducible by simple testing.

Static analysis entails analyzing software without executing the programs in question. The analysis may be performed at the level of the source-code itself, or at the level of the object code that is generated for execution (byte code or binary). Analysis is performed by automated tools that tend to be rooted in some formal method such as model checking [11], symbolic execution [29], and abstract interpretation [13]. Although such techniques have shown significant promise in the past, performing such analysis on a large scale for services has proven to be very slow and intractable. Performing program analysis and verification at smaller scales for individual components is the current limit of such techniques.

Tools such as Semmle [14] provides a unified framework that implements various program analysis techniques and correctness checks. In our experience, Semmle has proven to be somewhat useful for simple situations, but it lacks generality.

Differential static analysis techniques such as SymDiff [24] are immediately relevant to the problem discussed in this paper. But differential analysis techniques are usually property driven; two versions of the program are analyzed with respect to a specific correctness property. For example, the analysis may be performed for asserting differences in the new version w.r.t. null pointer dereferences. We believe this approach too lacks generality since it is not feasible to enumerate all such properties of code in large dynamic services with multiple dependent components. Yet, we do draw inspiration from this work to build Orca’s AST-based differential analysis.

Tools such as Gumtree [15] use differences in ASTs to derive accurate edit scripts. Their techniques need to be fine-grained and therefore use heavyweight algorithms that implement isomorphism detection. Orca uses differences in ASTs with a different goal: for a changed file, we use the AST difference to enumerate all changed entities, such as changed variables, methods, classes, namespaces, etc. So we use faster, more coarse-grained heuristics than Gumtree, making our techniques much more performant.

Log Enhancement techniques [44, 47] improve log-based debugging by making logs richer and more targeted towards diagnosability. We believe such work is complementary to our approach and Orca can gain significantly with such techniques.

3 Overview

In this Section, we first describe the system development lifecycle of the Orion enterprise email service and the role of the OCE. We next describe the characteristics
of post-deployment bugs in Orion and motivate the approaches we adopt in Orca. We provide an overview of Orca and its goals, and finally state Orca’s scope.

3.1 System Development Lifecycle

Figure 1 shows Orion’s system development lifecycle (SDLC) and Table 1 holds a summary definition of each term we use for the reader’s convenience. Multiple developers commit code, where a commit varies in complexity from a small tweak to a single file to changes to hundreds, even thousands of files. These commits are reviewed by one or more reviewers that the developer chooses. After multiple iterations with the reviewers, the commits are tested using unit, integration, and functional tests. Periodically, the administrator creates a new build by combining a set of commits. A build is a unit of deployment of the service and may contain just one, or hundreds of commits.

Builds are deployed in stages onto rings. A ring is a pre-determined set of machines on which the same build runs. The build is first deployed onto the smallest ring, or “Ring 0”, consisting of a few thousand machines. When it is considered safe, the build progresses through multiple rings such as Ring 1 and Ring 2 until it is finally deployed world-wide. The idea of this staged deployment is to find bugs early in the life-cycle.

Once the build is deployed to a ring, several tools monitor it. A tool may use passive or active monitoring techniques, either analyzing logs or sending periodic probes to a component. It uses anomaly detection techniques to raise an alert that the OCE receives.

If only the machines running a specific build raise alerts, the OCE concludes that the build is buggy and she begins bug localization. Roughly half of all alerts in Orion are caused by bugs. Consequently, bugs are a significant reason for service disruption. If the alerts are not confined to a particular build, the bug is likely due to other reasons such as faulty networks or hardware misconfiguration. Root-causing infrastructure issues is not our focus as several tools already exist for this purpose [43, 46].

To localize the bug to a commit, the OCE picks the commit that she feels is most likely to cause an issue and contacts the developer who created it. If the developer responds in the affirmative that her commit may indeed have caused the bug, the commit is immediately reverted and the service is restored to a healthy state. Note that the developer does not necessarily debug or fix the bug before responding to the OCE. If the developer says that their commit is not responsible for the bug, then the OCE picks the next most likely commit, and repeats the process until the service becomes healthy. Compared to a novice, an experienced OCE with domain-knowledge may pick the correct commits more promptly and therefore restore the service much faster. Orca removes this dependency on experience and domain-knowledge by codifying it in its search algorithms.

3.2 Post-Deployment Bugs

Over a period of eight months, we analyzed various post-deployment bugs and the buggy source-code that caused them. Table 2 outlines a few characteristic issues. The table shows the type of alert, it provides an overview of the symptom, and a description of the root-cause. It also shows the number of commits (and the number of files) that an OCE has to consider while performing bug attribution which, in some cases is more than 200.

In general, we have found that bugs fall predominantly into one of the following categories:

- Bugs specific to certain environments. Often, a component starts failing because files implementing that component have a bug (Bugs 2, 5). Usually, the failures happen only for a specific type of client such as web-based clients, or in a specific region such as Japan. Tests do not catch the bug
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commit</td>
<td>Set of file changes made by one developer.</td>
</tr>
<tr>
<td>Review</td>
<td>Recommendations made by one or more developers for a commit.</td>
</tr>
<tr>
<td>Build</td>
<td>Unit of deployment for the service consisting of one or more commits.</td>
</tr>
<tr>
<td>Ring</td>
<td>Set of machines onto which a build is deployed.</td>
</tr>
<tr>
<td>Probe</td>
<td>Periodic checks to functions/APIs to ensure they are working as expected.</td>
</tr>
<tr>
<td>Alert</td>
<td>An email- or web-based notification that warns the OCE of a problem.</td>
</tr>
</tbody>
</table>

Table 1: Terms used in the SDLC description and their definitions

not all configurations, clients and environments are tested.

- **Bugs due to uncaptured dependencies.** Dependencies can be of various types. In Bug 10, a server-side implementation is modified without appropriately modifying the client-side code. This happens because developers often overlook such dependencies as no compile-time tool captures them completely. Another example of an uncaptured dependency is Bug 4. A commit modifies a certain library, but unbeknownst to the developer, another component depends on certain features in the older version of the library and stops working correctly.

- **Bugs that introduce performance overheads.** Several probes track performance issues. For instance, in Bug 8, a code addition that was not thread-safe caused CPU overload that slowed down the service. The bug emerges only when a large number of users use the service. Hence it is not caught in testing.

- **Bugs in the user-interface.** A UI Feature starts misbehaving, so a customer complains. An example of this is Bug 1.

### 3.3 Orca Overview

Studying these bugs and observing the OCEs gave us valuable insights. We state these insights, and describe how Orca’s design is influenced by them.

*Often, the same meaningful terms occur both in the symptom and the cause.* Table 2 captures this under the “Term Similarity” column. Some matched terms are proper nouns such as the component name (Bug 3) or global data types (Bug 7). They can also be commonly used terms such as protocols (*Imap* in Bug 9) or the function performed (*suggest* in Bug 1 and *migrate* in Bug 6). Given the term similarity between symptom and cause, we designed Orca as a custom-designed search engine, using the symptom as the query-text, and giving us a ranked list of commits as results. Orca searches for the symptomatic terms in names of the modified code by performing differential code analysis on the abstract syntax tree. We describe this procedure in Section 4.1.

Testing and anomaly detection algorithms do not always find a bug immediately. A bug may start surfacing in a new build despite being introduced through a commit to a much older build. Bug 3 in Table 2 is an example. We introduce a build provenance graph to allow Orca to expand its search to older builds from which the current build has been derived. We describe this in Section 4.2.

*Builds may have hundreds of commits, so manually attributing bugs can be a long-drawn task.* For instance, Bug 2 appears in a build that had 201 commits. Bug 3 appears in a build with 160 commits but the root-cause was in the previous build which had 41 commits. The OCE is faced with the uphill task of analyzing, in these cases, up to 200 commits before discovering the buggy commit. OCEs often work at odd hours and are constantly pressed for time. Orca therefore ranks commits based on a *prediction of commit risk*. Orca uses machine-learning and several features such as developer experience, code hot-spots and commit type to make this prediction. We describe this in Step 4 of Section 4.3.

To facilitate its use, we have built an Orca user interface and leverage caching and parallelism to ensure an interactive experience for the OCE. We describe our optimizations in Section 5.

*There are thousands of probes in the system, and probe failures and exceptions are continuously logged.* Therefore there is rich data on what symptoms, or potential queries to Orca look like. This allows us to track frequency of terms that appear in the queries and use the *Inverse Query Frequency (IQF)* rather than the Inverse Document Frequency (IDF) in our search rankings. We explain this further in in Section 4.3.

### 3.4 Orca Scope

In this section, we elaborate what Orca does *not* aim to solve. This is primarily because existing techniques already address these issues.

Orca does not target issues caused by faults in the infrastructure. Several techniques [22, 43] exist to do this. Orca does not solve the anomaly detection problem directly as several techniques already exist for this [4, 10].
But we do recognize that anomaly detection algorithms are imperfect. Orca’s build provenance graph helps find bugs even when the anomaly detection algorithm detects a bug well after a commit introduces the bug. 

Orca does not handle bugs where the query does not have any context-specific information. Notable examples are performance issues, where the symptoms are, simply, out-of-memory exceptions or CPU-above-threshold exceptions. There exist other techniques [10, 48] in the literature which can debug such issues and therefore, can be used in combination with Orca. We discuss this further in Section 7.

### 4 Design

In this Section, we describe Orca’s differential bug localization constructs in more detail. The input query to Orca is a symptom of the bug. This could be the name of the anomalous probe, an exception message, a stack-trace, or the words of a customer-complaint. The idea is to search for this query through changes in code or configurations that could have caused this bug. Thus the “documents” that the tool searches are properties that changed with a commit, such as names of files that were added, removed or modified, commit and review comments, modified code and modified configuration parameters. Orca’s output is a ranked list of commits, with the most likely buggy commit displayed first.

First, we describe two novel constructs that we use for search-space pruning and search-space expansion respectively: differential code analysis, and the build provenance graph. Next, we describe the machine-learning based model of root-cause commits. Finally we provide a detailed description of the search algorithm.
4.1 Differential Code Analysis

Orca can search the entire code file to perform bug localization. But this approach can find false matches and drop Orca’s precision, especially since we have found that every commit changes, on average, only about 20 lines of source-code per-file (including file additions), whereas the entire file can consist of hundreds of lines of code.

Another simple approach, on the other end of the spectrum, would be to search only the file names for matches, and not look into the code at all. While this is partially effective, our evaluation in Section 6 shows that this does not give us satisfactory recall.

We therefore employ a middle-path – differential code analysis – within Orca. Prior work has studied differential code analysis, albeit mostly at the semantic level [23]. It identifies relevant pieces of the code change that can potentially cause different behavior in the new version relative to the old version, but with reference to a specified property, such as null dereference, memory consistency, etc. Such techniques rely on computationally expensive techniques such as differential symbolic execution [34] and regression verification [19].

We do not go with the semantic approach because (a) it is difficult to determine the full set of properties to capture all bugs in large-scale services, and (b) we would like to avoid the performance overhead of traditional static analyzers for differential analysis.

On the other hand, we could go with a complete lexical analysis on the differences, i.e., we can match terms in the query with terms in the difference, without any syntactic or semantic understanding. This approach will be relatively lightweight. However, it will miss several root-causes because very often, terms that match, such as protocol names, are parts of higher-level structures such as the method names and classes that have been modified. Consider Bug 6 in Table 2, for instance. the term migrat appears in the name of the function that has been changed, but not in text that has changed. Our techniques therefore need to identify syntactic constructs, such as methods and classes, that have changed.

Therefore, rather than going with a semantic or lexical analysis, we perform a syntactic analysis. We use the abstract syntax trees (AST) of the old and new version of the program to discover relevant parts of the source that have been added, removed, and modified.

Our analysis finds differences in entities of the following types: class, method, reference, condition, and loop. We create a “difference set” $D$ of two ASTs, $A_{old}$ and $A_{new}$, in the following way. Say $e_i$ is the old version of an entity, and $e_j$ the new version. Say $t$ is the type of the entity. Then,

$$D = \begin{cases} d_{added} = \forall e \in A_{new} \mid e \not\in A_{old} \\ d_{removed} = \forall e \in A_{old} \mid e \not\in A_{new} \\ d_{changed} = \forall e_{new} \in A_{new} \mid type(e_{new}) = t \\
\quad \land e_{new} \in A_{old} \land e_{new} \not= e_{old} \\
\quad \land type(e_{old}) = t \\
\quad \land e_{diff} = e_{old} \Delta e_{new} \\
\quad \land e_{new} \in A_{new} \\
\quad \land type(e_{new}) = t \land e_{old} \in A_{old} \\
\quad \land e_{new} \not= e_{old} \end{cases}$$

Thus, the difference set $D$ captures entities that have been added, removed, or changed. For all entities in $D$ that have been changed, we also capture the differences ($d_{diff}$) between the two versions of the entity using a heuristic. For instance, say our heuristic detects that two lines of a function $F$ have been changed. In addition to $D$ containing the name of the function $F$, $D$ also includes $d_{diff}$, which contains the entire text of the two changed lines: both the old version, and the new version.

We would like to point out that our syntactic approach to differential code analysis is not sound: we may detect changes even though there are none. For instance, consider the case where a function name changes completely, but the body remains unchanged. Our algorithm will treat this as a completely new function. While this could cause a precision drop in Orca since we are including more textual differences than actually exist, the algorithm does ensure that all changes are captured.

4.2 Build Provenance Graph

In Section 3.2, we have shown that a buggy commit to an older build may show symptoms only in a subsequent build. This could be because an inaccurate anomaly detection algorithm detects an anomaly too late. It may also be that a subtle bug manifests only in certain environments or only when a large number of users hit the service.

To accommodate this scenario, we expand our search to include previous builds that the symptomatic build is “derived” from. We maintain a build provenance graph (BPG) that captures dependencies between various builds along the axes of time and ring ID. Figure 2 shows an example fragment of a build provenance graph.

4.2.1 Construction

We now describe how we construct the build provenance graph. The BPG captures how builds are created and promoted in the different rings. Every build is represented as $B_{r,v}^i$, where $i$ is a the build identifier, $r$ is the ring identifier, and $v$ is the version of the build within a ring.
Say a build spends a certain amount of time in Ring 0. If it is stable and shows no unhealthy behavior for a while, it is considered for promotion to the next ring. Build $B^n_{0,0}$ is one such build. It forks off build $B^n_{1,0}$ which runs in the next ring, Ring 1. Since $B^n_{1,0}$ is directly derived from $B^n_{0,0}$, any bugs that emerge in $B^n_{1,0}$ could potentially be due to commits originally made to $B^n_{0,0}$. Thus, we introduce an inter-ring edge between the two builds. Similarly, once a build is considered stable in Ring 1, it is forked off to Ring 2. This introduces the inter-ring edge between $B^n_{1,0}$ and $B^n_{2,0}$. For a given build identifier $n$, only one inter-ring edge can exist between two consecutive rings.

Meanwhile, developers make fresh commits within each ring too, thereby creating the next build versions within the same ring. So in Figure 2, $B^n_{0,1}$ is derived from $B^n_{0,0}$, $B^n_{1,1}$ from $B^n_{1,0}$, and so on. We call the edges between these builds intra-ring edges. Several such intra-ring edges may exist in all rings though most are in Ring 0 which is the most dynamic and experimental ring. At any given time, a vertical line drawn through the graph yields the active build within each ring at that time. For instance, at time $t_1$, the active builds are $B^n_{0,3}$, $B^n_{1,1}$, and $B^n_{2,0}$. For ease of explanation, we assume that at any given time there is only one active build in every ring though in reality there could be many.

We now explain a third edge-type called the back-port edge, shown in Figure 2 by dotted arrows. A critical bug that goes undetected may, with time, propagate to a large number of builds across all rings. Say at time $t_0$, such a bug causes an alert in build $B^n_{1,1}$. Say at time $t_1$, using the process described in Section 4.2.2, we localize the bug to a commit made to the earlier build $B^n_{0,0}$. The bug is fixed through a new commit $c$ to the code of $B^n_{1,1}$, and this generates $B^n_{1,2}$, with an intra-ring edge between them.

We can see that since the bug originated in $B^n_{0,0}$, it also exists in the active builds within Ring 0 and Ring 2 that have been derived from it, i.e. $B^n_{0,3}$ and $B^n_{2,0}$. Consequently we apply the commit $c$, or we back-port it, to $B^n_{0,3}$ and $B^n_{2,0}$. This creates new builds $B^n_{0,4}$ in Ring 0 and $B^n_{2,1}$ in Ring 2. Thus we add two back-port edges with the label $c$ from $B^n_{1,2}$ to $B^n_{0,4}$, and $B^n_{1,2}$ to $B^n_{2,1}$.

Finally, we describe how the build identifier $n$ gets incremented in the build provenance graph. All builds across all rings that have the build identifier $n$ are derived originally from $B^n_{0,0}$. Thus $B^n_{0,0}$ is called an origin build. With time, several new commits are applied in Ring 0. To ensure that these commits are fully deployed across all rings, a subsequent build from Ring 0 is chosen to be the next origin build. In the figure, this is $B^n_{0,4}$, which we rename as $B^n_{0,n+1}$, or the next origin build. All subsequent builds are now derived from this new origin build.

It can be seen that, barring backport edges, every node in the build provenance graph has only one incoming edge. This can be either an inter-ring or an intra-ring edge.

### 4.2.2 Traversal

We now describe how Orca uses the build provenance graph to expand its search-space and find potential buggy commits in older builds. Given a symptomatic build $B^n_{p,q}$, the purpose of the traversal is to find a list of candidate commits for the search.

We observe that Ring 0 is the most experimental of all rings. Builds in Ring 0 see a large number of significant commits. Consequently, our intuition is that, to localize a bug that appears in $B^n_{p,q}$, we should search all builds back to the origin build $B^0_{0,0}$, which is in Ring 0. Thus using inter-ring and intra-ring edges, we backtrack from $B^n_{p,q}$ to the origin build $B^n_{0,0}$. In addition, we also include all back-ported commits to every build on the same path. Since every build has only one incoming inter-ring or intra-ring edge, there is only one such path from $B^n_{p,q}$ to $B^n_{0,0}$. The candidate list of commits to search will include all commits made to the builds on this path, and the back-ported commits on the same path.

We now explain this through an example with Figure 2. Say an alert is raised in $B^n_{2,1}$. Backtracking from $B^n_{2,1}$ to $B^n_{0,0}$ yields the set of commits $\{C(B^n_{2,1}), C(B^n_{2,0}), C(B^n_{1,0}), C(B^n_{0,0})\}$, where $C(B_{i,j})$ is the set of commits that were made to build $B_{i,j}$. To this, we add $c$, which is a backported commit from $B^n_{1,2}$ to $B^n_{2,1}$, thereby giving us the final list of commits to search in. That is,

$$\Gamma = \{C(B^n_{2,1}), C(B^n_{2,0}), C(B^n_{1,0}), C(B^n_{0,0}), c\}$$

### 4.3 Algorithm

In this Section, we describe Orca’s search algorithm which uses differential code analysis and the build provenance graph. The Orca search algorithm consists of four

\[\text{Algorithm}\]
steps: 1) Query pre-processing where we perform tokenization, stemming and stop-word removal, 2) build graph traversal described in Section 4.2.2, 3) token-matching in code changes using Differential Code Analysis and 4) ranking and visualization of results. Our system runs differential code analysis and constructs the build provenance graph in the background periodically so that these tasks do not slow down the query response time.

**Step 1:** The search queries are symptoms of the problem, consisting of probe names, exception texts, log messages, etc. We first tokenize the terms in these symptoms by using a custom-built code tokenizer. This tokenizer uses heuristics that we have built specifically for code and log messages, such as splitting large complex strings along Camel-cased or Pascal-cased fragments. We also create n-gram based tokens since we have found that bigrams, such as `ImapTransfer` and `mailboxSync`, capture important information.

Next, we filter out irrelevant words also called stopwords [27] from these symptoms. Previous work has shown that logs have a lot of inherent structure [40]. For instance, all exception names have the suffix `Exception` and almost all log messages have a timestamp. Unlike conventional search-engines, even before we built Orca, we had access to about 8 million alerts consisting of probe names, exceptions and log messages from Orion’s log store. We therefore perform stop-word removal on these to weed out commonly used or irrelevant terms such as `Exception` or timestamps. This step gives us a list of relevant “tokens” in the symptom. For each token $t$, we also maintain an Inverse Query Frequency (IQF) value [41] that we call $t_{IQF}$, obtained by analyzing Orion’s logs. $t_{IQF}$ is calculated as (No. of queries/No. of queries in which token $t$ appears). A high value of $t_{IQF}$ implies that the token $t$ is more important.

**Step 2:** We traverse the build provenance graph to find all builds related to the symptomatic build. From each build we discover, we enumerate all the commits that caused bugs in deployment are labeled “risky” while those that did not, we labeled “safe”. We have put in considerable effort into engineering the features for this task. The features that we input to the learner roughly fall into four categories:

- **Developer Experience.** Developers who are new to the organization and to the code-base tend to create more post-deployment bugs. Hence, we associate several experience-related features with each commit.
- **Code ownership.** We found that certain files that were mostly changed by a single developer caused fewer bugs than files that were constantly touched by several developers. Hence, for each commit, we use features to capture whether it touched files with very few owners or many owners.
- **Code hotspots.** Certain code-paths, when touched, tend to cause more post-deployment bugs than others. Some of our features capture this fact.
- **Commit complexity.** Several features, such as file types changed, number of lines changed, and number of reviewer comments capture the complexity of the commit.

Thus, for commits that have the same relevance score based on the terms match, we use the commit risk as a
secondary sort key to obtain a rank-order for Orca’s output.

5.1 Data Loaders

Since Orca requires information about various parts of the system – source-code, builds, deployment information and alerts – a significant part of our implementation are data loaders for these different types of data. Figure 3 shows an architectural overview of the implementation. At the heart of Orca is a standard SQL database. This database is populated by data loaders at a predefined frequency. We now describe the different data loaders we use.

5.2 Background Analyses

As our data loaders periodically load new data into the SQL database, we periodically initiate differential code analysis, build provenance graph construction, IQF calculation, and the commit risk prediction used in Section 4.3. If needed, the frequency of an analysis can be changed to make it more/less frequent. For example, the IQF calculation runs once a week, while commit risk prediction runs once every day. The other two processes run once every hour. Finally, it should be noted that all analyses that have been developed and deployed within this system are agnostic of the data source. The SQL database schema is normalized; thus, providing the same interface to all analyzers irrespective of the data source.

5.3 API Implementation

We now describe the Orca API and its implementation. Each Orca request is processed in real-time by the core Orca engine, and results are returned in JSON format. Clients decide on the relevant parts of the return result and how to display them. We make use of a Redis Cache [8] for improving our lookup times associated with data that is static. This includes data about the

5.4 Apostle

As a demonstration of the Orca platform we have created a web interface that allows users to search for relevant code commits. Users can specify the commit range, such as the last 4 days, and the search results are ranked based on the Orca score. The interface allows users to drill down into the source code of the commit and view the changes made.

5.5 Private APIs

We have also implemented private APIs for clients to access the Orca data in a more granular manner. These APIs allow clients to query the database based on specific criteria, such as the commit author, the commit message, or the file paths. This provides clients with more control over the data they access.

5.6 Open Source

Orca is open source and is available on GitHub [9]. The source code is under the Apache 2.0 license, and contributions are welcome.

5.7 Conclusion

In this paper, we have described the design and implementation of Orca, a system for searching and searching for relevant code commits. Orca uses a combination of data loaders, background analyses, and API implementation to provide a powerful tool for developers. We have demonstrated the usefulness of Orca through a web interface, and we encourage other developers to use Orca in their projects.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. 1562596. We would like to thank the anonymous reviewers for their helpful comments and suggestions.
builds, files, source-code, difference sets, and the build provenance graph. Internally, the Orca system records all requests made and any feedback provided, which we use for learning and improvement.

All our services, servers, databases, and caches are implemented and operated using Azure as both a Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) provider.

### 5.4 Usage

Orca can be used in two ways: OCEs can use it to make ad-hoc queries interactively, or an alerting infrastructure can query Orca to get a list of suspect commits for an alert and include it in the alert itself.

For the OCEs, we have built a web-based UI that allows them to enter the details of their query and view the results. Figure 4 shows a screenshot of the UI with sensitive information removed. We have also built a PowerShell `cmdlet` with which the OCE can interact through a command-line interface.

To integrate Orca with alerting infrastructure, we have built an API that the alerting system can query directly. Currently, this mechanism is being used by multiple groups that are generating alerts within the Orion group. The web-based UI and `cmdlet` tend to be used by OCEs when new information such as log text or exception text has been discovered after the original alerts were generated.

Today, Orca has been deployed on multiple code-bases for six large-scale services within our enterprise. The combination of data loaders used in each code-base is slightly different and unique, but fully operational and functional.

### 6 Evaluation

In this section, we provide results we obtain by evaluating Orca. First, we evaluate Orca’s result quality, i.e. how often it attributes a bug correctly to the right commit. Next, we evaluate how much effort an OCE saves by using Orca. Third, we evaluate the performance of Orca and the savings we get by using a Redis-based cache.

Since its deployment with Orion in October 2017, the Orca API has been invoked 4400 times to debug issues within the Orion service. Unfortunately, there is no central location where OCEs retrospectively log information about buggy commits. To evaluate how well Orca localizes a bug, we not only need the complete symptom of the bug, such as error messages or exceptions, but also the root-cause commit. Consequently, we began a manual process towards gaining this information for as many bugs as we could. We interviewed multiple OCEs and manually analyzed source-code, bug-reports and email-threads.

By performing this exercise, we collected complete information for 48 of these bugs. These bugs vary greatly in characteristics. While some were inadvertently introduced by a single line-change, others were caused by complex dependencies between components.

#### 6.1 Result Quality

To measure the quality of Orca’s results, we interviewed several OCEs about how they would quantify result quality. Based on these interviews we determined that it is important that we find the buggy commit in as many cases as possible. This is captured by the Recall, i.e. the fraction of bugs where we found the buggy commit in any position in our results. The OCEs also told us that they linearly scan the list of commits that Orca provides, hence the closer the correct commit is to the top, the more time they save. To capture this, we use the Mean Recip-
Table 3: We use recall to evaluate applying differential code analysis alone (DCA), differential code analysis with build provenance graph (DCA+BPG), and differential code analysis with build provenance graph and commit risk (ALL). Numbers in parentheses are the number of bugs correctly localized. We evaluate aggregating by MAX, SUM, and AVG.

<table>
<thead>
<tr>
<th>Agg.fn.</th>
<th>No. of results = 5</th>
<th>No. of results = 10</th>
<th>No. of results = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DCA</td>
<td>DCA+BPG</td>
<td>ALL</td>
</tr>
<tr>
<td>MAX</td>
<td>0.65(31)</td>
<td>0.60(29)</td>
<td>0.63(30)</td>
</tr>
<tr>
<td>SUM</td>
<td>0.52(25)</td>
<td>0.52(25)</td>
<td>0.60(29)</td>
</tr>
<tr>
<td>AVG</td>
<td>0.60(29)</td>
<td>0.46(22)</td>
<td>0.58(28)</td>
</tr>
<tr>
<td></td>
<td>0.69(33)</td>
<td>0.77(37)</td>
<td>0.77(37)</td>
</tr>
<tr>
<td></td>
<td>0.71(34)</td>
<td>0.67(32)</td>
<td>0.77(37)</td>
</tr>
<tr>
<td></td>
<td>0.67(32)</td>
<td>0.73(35)</td>
<td>0.77(37)</td>
</tr>
<tr>
<td>DCA+BPG</td>
<td>0.70(33)</td>
<td>0.77(37)</td>
<td>0.77(37)</td>
</tr>
<tr>
<td>ALL</td>
<td>0.69(33)</td>
<td>0.77(37)</td>
<td>0.77(37)</td>
</tr>
</tbody>
</table>

Table 4: We use MRR to evaluate applying differential code analysis alone (DCA), differential code analysis with build provenance graph (DCA+BPG), and differential code analysis with build provenance graph and commit risk (ALL). We also evaluate aggregating by MAX, SUM, and AVG.

<table>
<thead>
<tr>
<th>Agg.fn.</th>
<th>No. of results = 5</th>
<th>No. of results = 10</th>
<th>No. of results = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DCA</td>
<td>DCA+BPG</td>
<td>ALL</td>
</tr>
<tr>
<td>MAX</td>
<td>0.41</td>
<td>0.38</td>
<td>0.39</td>
</tr>
<tr>
<td>SUM</td>
<td>0.42</td>
<td>0.36</td>
<td>0.40</td>
</tr>
<tr>
<td>AVG</td>
<td>0.36</td>
<td>0.25</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>0.44</td>
<td>0.38</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>0.44</td>
<td>0.38</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>0.37</td>
<td>0.29</td>
<td>0.41</td>
</tr>
<tr>
<td>DCA+BPG</td>
<td>0.44</td>
<td>0.38</td>
<td>0.42</td>
</tr>
<tr>
<td>ALL</td>
<td>0.44</td>
<td>0.38</td>
<td>0.42</td>
</tr>
</tbody>
</table>

MRR is the most suitable metric described how we need to aggregate file-level relevance value into one value at the commit-level.

- **How many results should we show in the Orca UI?** When asked, the OCEs mentioned that they would not want to see beyond 10 results for each query. Hence, we wanted to evaluate the trade-off between the number of results shown and the recall and MRR.

We now answer each of these questions in order.

**The build provenance graph adds 8% to the recall.**

Observe the data in bold in Table 3. DCA alone localizes 33 bugs for a recall of 0.69. Adding the build provenance graph helps us localize 4 more bugs correctly thereby increasing our recall to 0.77. While at first glance, this may appear to be a small increase, OCEs find it significantly more difficult to attribute bugs that occur in older builds, and therefore the value of finding these 4 bugs eases the OCE’s workload considerably. We show this quantitatively in Section 6.2.

**Adding commit risk-based ranking improves MRR by 11%.** Observing the data in bold in Table 4, one can see that adding the build provenance graph reduced our MRR from 0.44 to 0.38. This happens because the build provenance graph increases our search-space significantly, and this increases the number of false-positive matches between terms and commits. Here, adding our secondary rank order based on commit risk improved our results by restoring the MRR value to 0.42.

The MAX and SUM aggregation functions per-
form better than AVG. While the buggy code-changes match some very high-relevance tokens, several lower-relevance tokens match these code-changes too. Hence taking the average value across all matches dilutes the high-value token matches, therefore reducing both recall and MRR. Such a dilution does not happen if we use MAX or SUM. We choose MAX in our implementation.

Showing 10 results seems a good trade-off between result quality and UI succinctness. We evaluated results while setting the number of results shown as 5, 10 and 20. We find that with 10 results we achieve close to our best recall and MRR values.

With 10 results and using MAX, we obtained a recall of 0.77, i.e. we found the root-cause in 37 out of 48 cases. The MRR was 0.42. We also studied the matched terms for these 37 bugs and found that term-similarity serves as a good proxy to capture different types of bugs. Table 2 showed that matched terms fall roughly into four categories: they either match a component name, a function that the component performs such as migrat or suggest, data types, or protocol names such as imap. We found in our case that of the 37 correctly localized cases, in 14 cases the token was a component name, in 17 cases it captured the function being performed, in 4 cases it matched a protocol name, and in the remaining 2 cases, the match was on a data type. Therefore term-similarity is quite versatile: it helps us catch a variety of bugs.

Of the 11 cases that we could not localize, in 1 case the issue was related to performance. In 4 cases, the problem was because a configuration setting changed, and that triggered the use of code that was committed much earlier than our build provenance graph covered. In future work, we therefore plan to include all configuration settings in our differential analysis. In the remaining 6 cases, term similarity just did not capture the high-level semantics of the bug, and static or dynamic analysis may be required.

### 6.2 Reduction of OCE Workload

We now investigate the effect of Orca on reducing the OCE’s workload. There are multiple ways to measure this. One way is to measure the amount of time saved by Orca for the OCE. Another is to determine the decrease in the number of commits that the OCE needs to manually investigate, both with and without Orca. We chose the latter metric because it allows us to quantify the reduction of OCE workload both at the level of every individual bug and as an aggregate. The former metric, i.e. the amount of time saved, can only provide us an aggregate across all alerts. Moreover, unless we shadow OCEs over an extended period of time, it is difficult to accurately quantify the time saved.

Given a bug, an OCE will investigate an average of \( c/2 \) commits to localize it, where \( c \) is the number of commits in the buggy build. If the OCE uses Orca, they need investigate at most \( r \) commits, where \( r \) is the rank of the correct commit in the results that the UI shows. If Orca does not find the correct commit, then apart from the 10 commits that Orca shows, the OCE needs to investigate an additional \((c−10)/2\) commits in expectation.

Figure 5 shows the number of commits investigated with and without Orca, for all 48 bugs that we evaluated, as a cumulative distribution function. Over all 48 bugs, without Orca, the OCE investigates a median of 22.75 commits, and an average of 28.9 commits. With Orca, she investigates a median of just 3.5 commits and an average of only 9.4 commits. Therefore, using Orca causes a \( 6.5 \times \) reduction in median OCE workload and a \( 3 \times (67\%) \) reduction in the OCE’s average workload. For the 37 bugs Orca localizes, this reduction factor in the average is much higher, i.e. \( 9.7 \times \). For the 4 bugs that were caught only because of the build provenance graph, the OCE had to investigate an average of 59.4 commits without Orca, and only 1.25 commits with it. This is a \( 47.5 \times \) improvement. These numbers point out the benefits that the build provenance graph gives us, and the benefits overall of using Orca for commit-level bug localization.

### 6.3 Performance

Finally, we evaluate the performance of Orca. We run Orca on a 32 core, 2GHz Intel Xeon E7-4820 CPU with 64 GB memory. We ran all 48 queries in sequence to obtain these results. First, we evaluated the effect of the Redis cache on Orca’s average query response time. Using
Redis, the average response time reduced from 12.4 seconds to 5.97 seconds, a gain of 51.8%. Next, we varied the degree of parallelism in Orca from 32 to 128 using the C# MAXDOP parameter [37] and noticed a significant effect on average query response time. With parallelism set to 32, the average response time is 30.94 seconds, whereas with parallelism of 128, it is 5.97 seconds. Our evaluation shows that there is a significant potential for parallelizing Orca further, thereby catering to many more queries and providing lower response times. Therefore we can effectively scale Orca out to more services within our enterprise without loss in performance.

![Graph showing cache hits and cache misses over time.](image)

Figure 6: The efficacy of using Redis in Orca’s deployment.

Figure 6 shows the effect of using Redis with Orca’s deployment over two months, starting from August 2018. On average, the number of hits is 2.2 times the number of misses. This shows that users of the Orca API make queries that have locality and therefore benefit greatly from the use of the Redis cache. This is to be expected as successive queries to Orca will be highly likely for the same builds and therefore will access similar difference sets.

7 Discussion

In this section, we first discuss the generalizability of Orca to other services. We then discuss the limitations, and how we intend to address these in the future.

7.1 Generalizing Orca

Fundamentally, in a CI/CD pipeline, to save time and resource-usage, services combine various commits before they build, test and deploy. Performing these procedures after every commit would be prohibitively expensive. Since such an aggregation of commits is absolutely necessary, so is the need for a tool that localizes bugs at the commit-level, such as Orca.

Though we describe Orca in the context of a large service and post-deployment bugs, we believe the techniques we have used also apply generically to many Continuous Integration/Continuous Deployment (CI/CD) pipeline. This is based on our experience with multiple services that Orca is operational on within our organization. Orca needs expressive symptoms as input. Modern-day services use rich monitoring systems enabled by infrastructure such as Nagios [33] and AlertSite [2] which can provide probe-level symptoms of problems.

7.2 Future Work

Two types of bugs that Orca currently cannot localize are performance and configuration issues. Addressing these bugs, therefore, are immediate next-steps.

To deal with performance-related bugs, we plan to incorporate better anomaly detection algorithms, and correlate anomalies with code-changes that could potentially cause them. We believe that the principal idea of Orca, i.e. keyword match, can also be applied to bugs that arise from faulty configuration settings. Therefore, we plan to include configuration-based difference sets into our search engine.

8 Conclusion

In this paper, we described Orca and the differential bug localization algorithm. Orca uses differential code analysis and the build provenance graph to find buggy commits in large-scale services. Orca is deployed with a large email and collaboration platform. We have shown that Orca finds the correct buggy commits in about 77% of bugs that we studied, and causes a $3 \times$ reduction in the work done by the OCE. We have also shown that Orca is efficient, accurate and easy to deploy.

References


