Isolating Cause-Effect Chains from Computer Programs

Andreas Zeller, *Member, IEEE Computer Society*

Abstract— **Consider the execution of a failing program as a sequence of program states. Each state induces the following state, up to the failure. Which variables and values of a program state are relevant for the failure? We show how the *Delta Debugging* algorithm isolates the relevant variables and values by systematically narrowing the state difference between a passing run and a failing run—by assessing the outcome of altered executions to determine wether a change in the program state makes a difference in the test outcome. Applying Delta Debugging to multiple states of the program automatically reveals the *cause-effect chain* of the failure—that is, the variables and values that caused the failure.**

In a case study, our prototype implementation successfully isolated the cause-effect chain for a failure of the GNU C compiler: "Initially, the C program to be compiled contained an addition of 1.0; this caused an addition operator in the intermediate RTL representation; this caused a cycle in the RTL tree—and this caused the compiler to crash." The prototype implementation is available as a public Web service, where anyone can submit programs to be debugged automatically.

Index Terms—automated debugging, program analysis, testing tools, combinatorial testing, diagnostics, tracing.

I. Introduction

PROGRAM debugging is commonly understood as the process of identifying and correcting errors in the program code. Debugging is a difficult task, because normally, errors can only be detected indirectly by the failures they cause. Now, let us assume we have some program test that fails. How did this failure come to be?

Traditionally, approaches to facilitate debugging have relied on static or dynamic program analysis to detect anomalies or dependencies in the source code and thus narrowed the set of potential erroneous code. In this paper, we propose a novel and very different approach. Rather than focusing on the source code as potential error cause, we concentrate on *program states* as they occur during program execution—especially, on the *difference* between the program states of a run where the failure in question occurs, and the states of a run where the failure does not occur.

Using *automated testing*, we systematically narrow these initial differences down to a small set of variables: "The failure occurs if and only if variable x has the value y (instead of y')". That is, $x = y$ is a *cause* for the failure: if x is altered to y', the failure no longer occurs. If we narrow down the relevant state differences at multiple locations in the program, we automatically obtain a *cause-effect-chain* listing the consecutive relevant state differences—from the input to the failure: "Initially, variable v_1 was x_1, thus variable v_2 became x_2, thus variable v_3 became x_3 ... and thus the program failed."

State differences are not only causes of failures, but also *effects* of the program code. By increasing the granularity of the cause-effect chain, one can interactively isolate the moment where the program state changed from "intended" to "faulty". This moment in time is when the piece of code was executed that caused the faulty state (and thus the failure)—that is, "the error" in the program to be examined.

Our approach also differs from program analysis in that it is *purely experimental*: It only requires the ability to execute an automated test and to access and alter program states. Knowledge or analysis of the program code is not required, although hints on dependencies and anomalies can effectively guide the experimental narrowing process.

Our prototype implementation is available as a public Web service where anyone can submit programs to be debugged automatically. The Web service only requires the executable in question and the invocations of the differing runs; it then determines a cause-effect chain for the entire run. Such a facility could easily be integrated within an automated test suite, automatically generating a precise diagnosis whenever a new failure occurs.

This paper is organized as follows. Section II recapitulates how to isolate failure-inducing circumstances automatically, using a GCC failure as example. Section III shows how to access program states and to isolate their difference, obtaining a cause-effect chain for the GCC failure. Section IV shows how to narrow down failure-inducing program code—that is, "the error" in GCC. Section V discusses further case studies, and Section VI discusses the current limits of our approach. In Section VIII, we discuss related work and Section IX closes with conclusion and consequences.

II. Isolating Relevant Input

In this Section, we recapitulate our previous work on isolating failure-inducing input [1]. As an ongoing example, consider the *fail.c* program in Figure 1. This program is interesting in one

```
double mult(double z[], int n)
{
  int i, j;

  i = 0;
  for (j = 0; j < n; j++) {
    i = i + j + 1;
    z[i] = z[i] * (z[0] + 1.0);
  }
  return z[n];
}
```

Fig. 1
THE *fail.c* PROGRAM THAT CRASHES GCC.

Andreas Zeller is with Universität des Saarlandes, Lehrstuhl für Softwaretechnik, Postfach 15 11 50, 66041 Saarbrücken, Germany. E-mail: zeller@computer.org.

#	GCC input	*test*
1	double **mult**(…) { int *i*, *j*; *i* = 0; for (…) { … } … }	✘
2	double **mult**(…) { int *i*, *j*; *i* = 0; for (…) { … } … }	✔
3	double **mult**(…) { int *i*, *j*; *i* = 0; for (…) { … } … }	✔
4	double **mult**(…) { int *i*, *j*; *i* = 0; for (…) { … } … }	✔
5	double **mult**(…) { int *i*, *j*; *i* = 0; for (…) { … } … }	✘
6	double **mult**(…) { int *i*, *j*; *i* = 0; for (…) { … } … }	✔
⋮	⋮	⋮
18	… *z*[*i*] = *z*[*i*] ∗ (*z*[0] + 1.0); …	✘
19	… *z*[*i*] = *z*[*i*] ∗ (*z*[0] + 1.0); …	✔
20	… *z*[*i*] = *z*[*i*] ∗ (*z*[0] + 1.0); …	**?**
21	… *z*[*i*] = *z*[*i*] ∗ (*z*[0] + 1.0); …	**?**

TABLE I
ISOLATING FAILURE-INDUCING GCC INPUT

aspect: It causes the GNU C compiler (GCC) to crash—at least, when using version 2.95.2 on Intel-Linux with optimization enabled:

```
$ gcc -O fail.c
gcc: Internal compiler error:
   program cc1 got fatal signal 11
$ _
```

If we say "*fail.c* causes GCC to crash", what do we actually mean? Generally, the *cause* of any event is a preceding event without which the event in question (the *effect*) would not have occurred. Indeed, if we remove the contents of *fail.c* from the input—that is, we compile an empty file—, GCC works fine. These two experiments (the failing and the passing run) actually prove that *fail.c* is a cause of the failure.

A. A "Trial and Error" Process

In practice, we typically want a more *precise* cause than just the contents of *fail.c*—that is, a smaller difference between the failing and the passing experiment. For instance, we could try to isolate the *smallest possible difference* between two GCC inputs. This can be done using a simple "trial and error" method, as illustrated in Table I.

In Step 1, we see the entire input *fail.c*, causing the GCC failure ("✘"). In Step 2, the contents of *fail.c* have been deleted; this empty input compiles fine ("✔"). In Step 3, we take away only the *mult* body. This input also compiles fine. Thus, we have narrowed the failure-inducing difference to the *mult* body: "Something within the *mult* body causes GCC to crash". In Steps 4 and 5, we have narrowed the cause to the *for* loop.

Continuing this "trial and error" method, we eventually narrow down the cause to the character set "+1.0" in Steps 18 and 19. This difference "+1.0" is minimal, as it can not be further reduced: Removing either "+" or "1.0" would result in a GCC syntax error (Steps 20 and 21)—a third outcome besides failure and success, denoted here as "**?**". So, we have isolated "+1.0" is a minimal difference or precise cause for the GCC failure—GCC fails if and only if "+1.0" is present in *fail.c*.

B. Delta Debugging

The interesting thing about the "trial and error" process to isolate failure-inducing input is that it can be *automated*—all one needs is a means to alter the input, and an automated test to assess the effects of the input. In fact, Table I does not show the narrowing process as conducted by a human, but the execution of the *Delta Debugging* algorithm, an automatic experimental method to isolate failure causes [1].

Delta Debugging requires two program runs $r_✘$ and $r_✔$—one run $r_✘$ where the failure occurs, and one run $r_✔$ where the failure does not occur. The *difference* between these two runs is denoted as δ; the difference can be *applied* to $r_✔$ to produce $r_✘$, or $\delta(r_✔) = r_✘$. Formally, δ is a failure cause—the failure occurs if and only if δ is applied. The aim of Delta Debugging, though, is to produce a cause that is as precise as possible. We thus decompose the original difference into a number of *atomic* differences $\delta = \delta_1 \circ \delta_2 \circ \cdots \circ \delta_n$.

Let us illustrate these sets in our GCC example. $r_✔$ is the GCC run on the empty input, and $r_✘$ is the run on the failure-inducing input *fail.c*. We model the difference δ between $r_✘$ and $r_✔$ as a set of atomic deltas δ_i, where each δ_i inserts the i-th C token of *fail.c* into the input. We further assume the existence of a *testing function* $test$ that takes a set of atomic differences, applies the differences to $r_✔$, and returns the test outcome—✘ if the test fails (i.e. the expected failure occurs), ✔ if the test passes (the failure does *not* occur), and **?** in case the outcome is *unresolved*—such as a non-expected failure.

Let us define $c_✔ = \emptyset$ and $c_✘ = \{\delta_1, \delta_2, \ldots, \delta_n\}$ as sets of atomic differences. By definition, $test(c_✔) = ✔$ holds (because *no* changes are applied to $r_✔$); $test(c_✘) = ✘$ holds, too (because *all* changes are applied to $r_✔$, changing it to $r_✘$).

In our GCC example, $test$ constructs the input from the given changes and checks whether the failure occurs. $test(c_✔)$ applies no changes to the empty input and runs GCC; the failure does not occur. $test(c_✘)$ inserts all characters of *fail.c* into the empty input, effectively changing the input to *fail.c*, and runs GCC; the failure would occur. The $test$ function returns ✘ if and only if the original failure occurs, ✔ if the program exits normally, and **?** in all other cases.

Given $c_✔$, $c_✘$, and $test$, Delta Debugging now isolates two sets $c'_✔$ and $c'_✘$ with $c_✔ \subseteq c'_✔ \subseteq c'_✘ \subseteq c_✘$, $test(c'_✔) = ✔$, and $test(c'_✘) = ✘$. Furthermore, the set difference $\Delta = c'_✘ - c'_✔$ is *1-minimal*—that is, no single $\delta_i \in \Delta$ can be removed from $c'_✘$ to make the test pass or added to $c'_✔$ to make the test fail. Hence, Δ is a precise cause for the failure.

Applied to the GCC input, Delta Debugging executes exactly the tests as illustrated in Table I. Delta Debugging first splits the input in two parts[1]. Compiling the header alone works fine (Step 3), and adding the initialization of i and j to the input (Step 4) does not yet make a difference. However, adding the "for" loop (Step 5) makes GCC crash. At this stage, $c'_✔$ is set up as shown in Step 4, $c'_✘$ is set up as in Step 5, and their difference $\Delta = c'_✘ - c'_✔$ is exactly the "for" loop—in other words, the "for" loop is a more precise cause of the failure.

Continuing the narrowing process eventually leads to $c'_✘$ as shown in Step 18 and $c'_✔$ as shown in Step 19: The remaining difference is exactly the addition of "+1.0". Steps 20 and 21 verify that each of these two remaining tokens is actually relevant for the failure. So, the remaining 1-minimal difference "+1.0" is what Delta Debugging returns—after only 21 tests, or

[1] In this example, we assume a "smart" splitting function that splits input at C delimiters like parentheses, braces, or semicolons.

Let $+$ be the set of all differences (in input or state) between program runs. Let $test : 2^+ \to \{\text{✘}, \text{✔}, \text{?}\}$ be a testing function that determines for a configuration $c \subseteq +$ whether some given failure occurs (✘) or not (✔) or whether the test is unresolved (?).

Now, let $c_{\text{✔}}$ and $c_{\text{✘}}$ be configurations with $c_{\text{✔}} \subseteq c_{\text{✘}} \subseteq +$ such that $test(c_{\text{✔}}) = \text{✔} \wedge test(c_{\text{✘}}) = \text{✘}$. $c_{\text{✔}}$ is the "passing" configuration (typically, $c_{\text{✔}} = \emptyset$ holds) and $c_{\text{✘}}$ is the "failing" configuration.

The *Delta Debugging algorithm* $dd(c_{\text{✔}}, c_{\text{✘}})$ isolates the failure-inducing difference between $c_{\text{✔}}$ and $c_{\text{✘}}$. It returns a pair $(c'_{\text{✔}}, c'_{\text{✘}}) = dd(c_{\text{✔}}, c_{\text{✘}})$ such that $c_{\text{✔}} \subseteq c'_{\text{✔}} \subseteq c'_{\text{✘}} \subseteq c_{\text{✘}}$, $test(c'_{\text{✔}}) = \text{✔}$, and $test(c'_{\text{✘}}) = \text{✘}$ hold and $c'_{\text{✘}} - c'_{\text{✔}}$ is *1-minimal*—that is, no single circumstance of $c'_{\text{✘}}$ can be removed from $c'_{\text{✘}}$ to make the failure disappear or added to $c'_{\text{✔}}$ to make the failure occur.

The dd algorithm is defined as $dd(c_{\text{✔}}, c_{\text{✘}}) = dd_2(c_{\text{✔}}, c_{\text{✘}}, 2)$ with

$$dd_2(c'_{\text{✔}}, c'_{\text{✘}}, n) = \begin{cases} dd_2(c'_{\text{✔}}, c'_{\text{✔}} \cup \Delta_i, 2) & \text{if } \exists i \in \{1, \ldots, n\} \cdot test(c'_{\text{✔}} \cup \Delta_i) = \text{✘} \\ dd_2(c'_{\text{✘}} - \Delta_i, c'_{\text{✘}}, 2) & \text{if } \exists i \in \{1, \ldots, n\} \cdot test(c'_{\text{✘}} - \Delta_i) = \text{✔} \\ dd_2\big(c'_{\text{✔}} \cup \Delta_i, c'_{\text{✘}}, \max(n-1, 2)\big) & \text{else if } \exists i \in \{1, \ldots, n\} \cdot test(c'_{\text{✔}} \cup \Delta_i) = \text{✔} \\ dd_2\big(c'_{\text{✔}}, c'_{\text{✘}} - \Delta_i, \max(n-1, 2)\big) & \text{else if } \exists i \in \{1, \ldots, n\} \cdot test(c'_{\text{✘}} - \Delta_i) = \text{✘} \\ dd_2\big(c'_{\text{✔}}, c'_{\text{✘}}, \min(2n, |\Delta|)\big) & \text{else if } n < |\Delta| \\ (c'_{\text{✔}}, c'_{\text{✘}}) & \text{otherwise} \end{cases}$$

where $\Delta = c'_{\text{✘}} - c'_{\text{✔}} = \Delta_1 \cup \Delta_2 \cup \cdots \cup \Delta_n$ with all Δ_i pairwise disjoint, and $\forall \Delta_i \cdot |\Delta_i| \approx (|\Delta| / n)$ holds.
The recursion invariant for dd_2 is $test(c'_{\text{✔}}) = \text{✔} \wedge test(c'_{\text{✘}}) = \text{✘} \wedge n \leq |\Delta|$.

Fig. 2

THE DELTA DEBUGGING ALGORITHM IN A NUTSHELL. THE FUNCTION dd ISOLATES THE FAILURE-INDUCING DIFFERENCE BETWEEN TWO SETS $c_{\text{✔}}$ AND $c_{\text{✘}}$. FOR A FULL DESCRIPTION OF THE ALGORITHM AND ITS PROPERTIES, SEE [1].

roughly 2 seconds.[2] Let us assume that an automated test already exists (for instance, as part of the GCC test suite); also, let us assume we have a simple scanner to decompose the input (a 10-minute programming assignment). Then, finding the cause in any GCC input comes at virtually no cost, compared to the manual editing and testing of *fail.c*.

The actual algorithm is summarized in Figure 2. The number of required tests grows with the number of unresolved test outcomes. In the worst case, nearly all outcomes are unresolved; then, the number of tests is $t = |c_{\text{✘}}|^2 + 3|c_{\text{✘}}|$. However, this worst case never occurs in practice, because in case of unresolved outcomes, the Delta Debugging algorithm has been designed to try runs more similar to $c_{\text{✔}}$ and $c_{\text{✘}}$. The central assumption is that the closer we are to the original runs, the lesser are the chances of unresolved test outcomes—an assumption backed by a number of case studies [1]. In the best case, we have no unresolved test outcomes; all tests are either passing or failing. Then, the number of tests t is limited by $t \leq \log_2\big(|c_{\text{✘}}|\big)$—basically, we obtain a binary search.

III. ISOLATING RELEVANT STATES

Let us reconsider the isolated cause "+1.0". Although removing "+1.0" from *fail.c* makes the GCC failure disappear, this is not the way to fix the error once and for all; we rather want to fix the GCC code instead. Unfortunately, processing such arithmetic operations is scattered all over the compiler code. Nonetheless, during the compilation process, "+1.0" eventually induces a faulty GCC state which manifests itself as a failure. How does "+1.0" eventually cause the failure? And how do we get to the involved program code?

The basic idea of this paper is illustrated in Figure 3 on the following page, which depicts a program execution as a series of *program states*—that is, variables and their values. On the left hand side, the program processes some input. Only a part of the input is relevant for the failure—that is, a difference like "+1.0" between an input that is failure-inducing and an input that is not. This difference in the input causes a difference in later program states, up to the difference in the final state that determines whether there is a failure or not.

The problem, though, is, that even a minimized difference in the input may well become a large difference in the program state. Yet, only some of these differences are relevant for the failure. So, we *apply Delta Debugging on program states* in order to isolate the variables and values that are relevant for the failure; these isolated variables constitute the *cause-effect chain* that leads from the root cause to the failure. This is the contribution of this paper: a fully automatic means to narrow down program states and program runs to the very small fraction that is actually relevant for the given failure.

A. Accessing and Comparing States

Our requirements are easy to satisfy; all we need is an ordinary debugger tool that allows us to retrieve and alter variables and their values. Let us initiate two GCC runs: a run $r_{\text{✘}}$ on *fail.c* and a run $r_{\text{✔}}$ on *pass.c*, where *fail.c* and *pass.c* differ only by "+1.0". Using the debugger, we interrupt both runs at the same location L; then, we retrieve each GCC state as a set of (*variable*, *value*) pairs. As a mental experiment, let us assume that all variables have identical values in $r_{\text{✔}}$ and $r_{\text{✘}}$, except for the four variables in Table II.

Obviously, this difference in the program state is the difference which eventually causes the failure: If we set the differing variables in $r_{\text{✔}}$ to the values found in $r_{\text{✘}}$ and resume execution, then GCC should behave as in $r_{\text{✘}}$ and fail.

We can use Delta Debugging to narrow down the cause. Now, the deltas δ_i become *differences between variable values*: applying a δ_i in $r_{\text{✔}}$ means setting the i-th differing variable in $r_{\text{✔}}$ to the value found in $r_{\text{✘}}$. The $test$ function executes $r_{\text{✔}}$, interrupts

[2] All times were measured on a LINUX PC with a 500 MHz Pentium III processor.

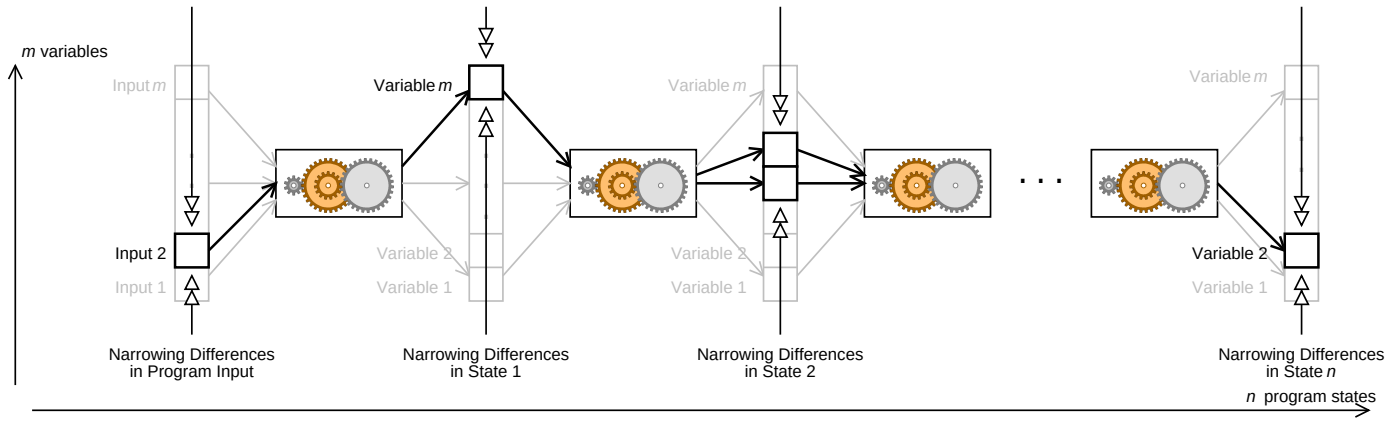


Fig. 3

NARROWING A CAUSE-EFFECT CHAIN. IN EACH STATE, OUT OF *m* VARIABLES, ONLY FEW ARE RELEVANT FOR THE FAILURE. THESE CAN BE ISOLATED BY NARROWING THE STATE DIFFERENCE BETWEEN A WORKING RUN AND A FAILING RUN.

	δ_1	δ_2	δ_3	δ_4	
	reg_rtx_no	*cur_insn_uid*	*first_loop_store_insn*	*last_linenum*	*test*
$r_{✘}$	32	74	0x81fc4e4	15	✘
$r_{✔}$	31	70	0x81fc4a0	14	✔

TABLE II
DIFFERING VARIABLES IN TWO GCC RUNS

#	*reg_rtx_no*	*cur_insn_uid*	*first_loop_store_insn*	*last_linenum*	*test*
1	32	74	0x81fc4a0	14	✔
2	32	74	0x81fc4e4	14	?
3	32	74	0x81fc4a0	15	✔

TABLE III
ISOLATING FAILURE-INDUCING VARIABLES

execution at L, applies the given deltas, resumes execution and determines the test outcome.

As an example, consider δ_1 and δ_2 from Table II. To test δ_1 and δ_2 means to execute GCC on *pass.c* (run $r_{✔}$), to interrupt it at L, to set *reg_rtx_no* to 32 and *cur_insn_uid* to 74, and to resume execution. If we actually do that, it turns out that GCC runs just fine—$test$ returns ✔ and we have narrowed the failure cause by these two differences.

Unfortunately, things are not so simple. If we continue the narrowing process using Delta Debugging, we end up in trouble, as shown in Table III. Step 1 is the application of δ_1 and δ_2, as discussed before—everything fine so far. In Step 2, though, we would apply the change δ_3, setting the pointer variable *first_loop_store_insn* to the address found in $r_{✘}$. This would cause an immediate core dump of the compiler—not really surprising, considering that an address from $r_{✘}$ probably has little meaning in $r_{✔}$.

In Step 3, we can exclude *last_linenum* as a cause and thus effectively isolate *first_loop_store_insn* as remaining failure-inducing difference, so we can easily see that our process is feasible. However, our state model is insufficient: We must also take *derived* variables into account—that is, all memory locations being pointed to or otherwise accessible from the base variables.

B. Memory Graphs

To fetch the *entire* state, we capture the state of a program as a *memory graph* [2]. A memory graph contains all values and all variables of a program, but represents operations like variable access, pointer dereferencing, struct member access, or array element access by edges.

As an example, consider Figure 4 on page 6, depicting a subgraph of the GCC memory graph. The immediate descendants of the ⟨*Root*⟩ vertex are the base variables of the program. The variable *first_loop_store_insn*, for instance, is found by following the leftmost edge from ⟨*Root*⟩. The dereferenced value is found following the edge labeled **()*[3]—a record with three members *value*, *fld[0].rtx*, and *fld[1].rtx*. The latter points to another record which is also referenced by the *link* variable.

Memory graphs are obtained by querying the base variables of a program and by systematically unfolding all data structures encountered; if two values share the same type and address, they are merged to a single vertex. (More details on memory graphs, including formal definitions and extraction methods, are available [2].) Memory graphs give us access to the entire state of a program and thus avoid problems due to incomplete comparison of program states. They also abstract from concrete memory addresses and thus allow for comparing and altering pointer values appropriately.

However, memory graphs also indicate another problem: The set of variables itself may *differ* in the two states to be compared. As an example, consider Figure 5 on the next page. In the upper left corner, you can see two memory graphs $G_{✔}$ and $G_{✘}$, obtained from the two runs $r_{✔}$ and $r_{✘}$. As a human, you can quickly see that, to change $G_{✔}$ into $G_{✘}$, one must insert element 15 into

[3]Each variable name is constructed from the incoming edge, where the placeholder () stands for the name of the parent.

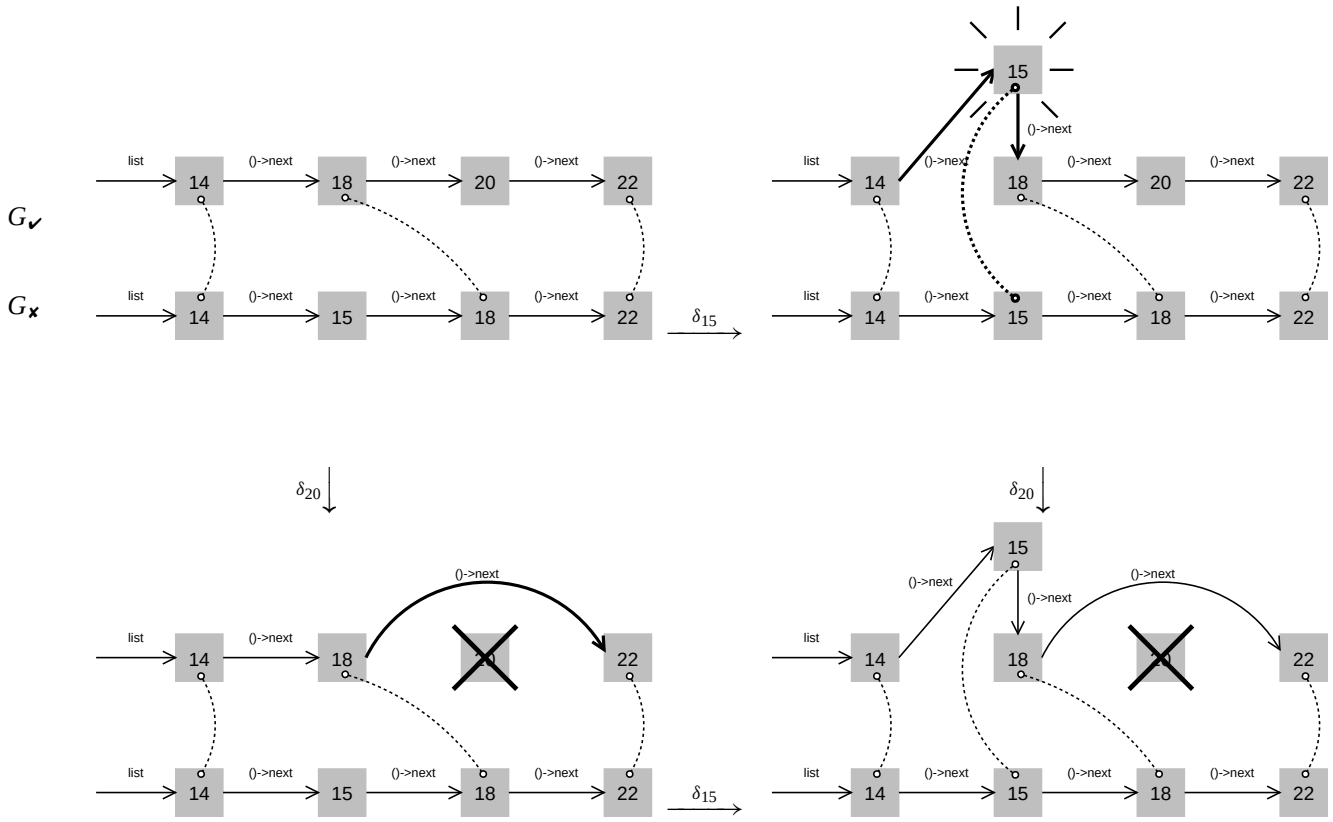


Fig. 5

DETERMINING STRUCTURAL DIFFERENCES BETWEEN MEMORY GRAPHS. ANY NODE NOT CONTAINED IN THE COMMON SUBGRAPH (DOTTED LINES) IS EITHER INSERTED OR DELETED (TOP LEFT). APPLYING δ_{15} ON $r_{✔}$ CREATES THE LIST ELEMENT 15, APPLYING δ_{20} DELETES LIST ELEMENT 20. APPLYING BOTH DELTAS (BOTTOM RIGHT) TRANSFORMS $r_{✔}$ TO $(\delta_{15} \circ \delta_{20})(r_{✔}) = r_{✘}$.

the list and delete element 20. To detect this automatically for arbitrary data structures, one must compute a *common subgraph* of $G_{✔}$ and $G_{✘}$: Any vertex that is not in the common subgraph of $G_{✔}$ and $G_{✘}$ has either been inserted or deleted. Large common subgraphs can be computed effectively by *parallel traversal* along matching edges [2].

In Figure 5, we have determined the largest common subgraph, drawn using dotted lines as a *matching* between $G_{✔}$ and $G_{✘}$.[4] It is plain to see that element 15 in $G_{✘}$ has no match in $G_{✔}$; likewise, element 20 in $G_{✔}$ has no match in $G_{✘}$.

For our purposes, we translate these differences into atomic deltas that create or delete new variables—one delta for each non-matched variable. In this example, we obtain a delta δ_{15} that creates the list element 15 and a delta δ_{20} that deletes list element 20. Both deltas can be applied independently (upper right and lower left). Altogether, we thus obtain deltas that change variable values, as sketched in Section III-A, as well as deltas that alter data structures.

[4]An edge is part of the matching (= the common subgraph) if its vertices match; there is no such edge in this example.

C. Isolating the GCC Cause-Effect Chain

Let us now put all these building blocks together. We have built a prototype called HOWCOME that relies on the GNU debugger (GDB) to extract the program state; each δ_i is associated with appropriate GDB commands that alter the state.

From Section II, we already know the failure-inducing difference in the input, namely the token sequence "+1.0", which is present in *fail.c*, but not in *pass.c*. HOWCOME's *test* function is also set up as discussed in Section II.

At which locations do we compare executions? For technical reasons, we require *comparable* states—since we cannot alter the set of local variables, the current program counter and the backtrace of the two locations to be compared must be identical. From the standpoint of causality, though, any location during execution is as causal as any other.

HOWCOME thus starts with a sample of three events, occurring in both the passing run and the failing run:

1. After the program start (in our case, when GCC's subprocess *cc1* reaches the function *main*)
2. In the middle of the program run (when *cc1* reaches the function *combine_instructions*)
3. Shortly before the failure (when *cc1* reaches the function *if_then_else_cond* for the 95th time—a call that never returns)

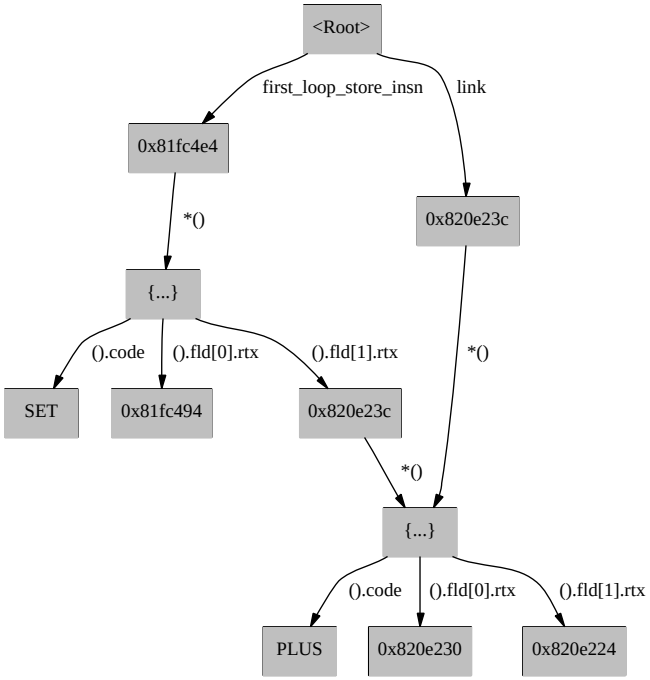


Fig. 4
A SIMPLE MEMORY GRAPH. POINTERS REFERENCE RECORDS, EACH REFERENCING ITS MEMBERS.

C.1 At *main*

HOWCOME starts by capturing the two program states of $r_{\checkmark}$ and $r_{\times}$ in *main*. Both graphs $G_{\checkmark}$ and $G_{\times}$ have 27139 vertices and 27159 edges (Figure 6); to squeeze them through the GDB command-line bottleneck requires 15 minutes each.

After 12 seconds, HOWCOME determines that exactly one vertex is different in $G_{\checkmark}$ and $G_{\times}$—namely *argv*[2], which is `"fail.i"` in $r_{\times}$ and `"pass.i"` in $r_{\checkmark}$. These are the names of the preprocessed source files as passed to *cc1* by the GCC compiler driver. This difference is minimal, so we do not need a Delta Debugging run to narrow it further.

C.2 At *combine_instructions*

As *combine_instructions* is reached, GCC has already generated the intermediate code (called RTL for "register transfer list") which is now optimized. HOWCOME quickly captures the graphs $G_{\checkmark}$ with 42991 vertices and 44290 edges as well as $G_{\times}$ with 43147 vertices and 44460 edges. The common subgraph of $G_{\checkmark}$ and $G_{\times}$ has 42637 vertices; thus, we have 871 vertices that have been added in $G_{\times}$ or deleted in $G_{\checkmark}$.

The deltas for these 871 vertices are now subject to Delta Debugging, which begins by setting 436 GCC variables in the passing run to the values from the failing run ($G_{\times}$). This obviously is a rather insane thing to do, and GCC immediately aborts with an error message complaining about inconsistent state. Changing the other half of variables does not help either. After these two unresolved outcomes, Delta Debugging increases granularity and alters only 218 variables. After a few unsuccessful attempts (with various uncommon GCC messages), this number of

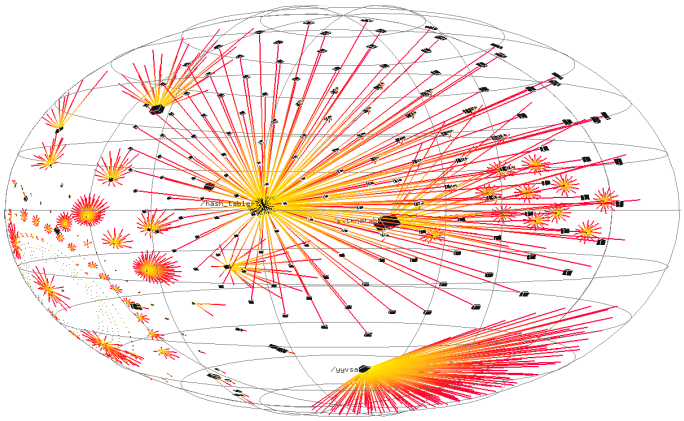


Fig. 6
THE GCC $G_{\checkmark}$ MEMORY GRAPH

altered variables is small enough to make GCC pass (Figure 7 on the facing page). Eventually, after only 44 tests, HOWCOME has narrowed the failure-inducing difference to one single vertex, created with the GDB commands

*set variable $m9 = (struct rtx_def *)malloc(12)*
set variable $m9→code = PLUS
set variable $m9→mode = DFmode
set variable $m9→jump = 0
set variable $m9→fld[0].rtx = loop_mems[0].mem
set variable $m9→fld[1].rtx = $m10
set variable first_loop_store_insn→fld[1].rtx→
fld[1].rtx→fld[3].rtx→fld[1].rtx = $m9

That is, the failure-inducing difference is now the insertion of a node in the RTL tree containing a `PLUS` operator—the proven effect of the initial change "+1.0" from *pass.c* to *fail.c*. Each of the tests required about 20 to 27 seconds of HOWCOME time, and 1 second of GCC time.

C.3 At *if_then_else_cond*

At this last event, HOWCOME captured the graphs $G_{\checkmark}$ with 47071 vertices and 48473 edges as well as $G_{\times}$ with 47313 vertices and 48744 edges. The common subgraph of $G_{\checkmark}$ and $G_{\times}$ has 46605 vertices; 1224 vertices have been either added in $G_{\times}$ or deleted in $G_{\checkmark}$.

Again, HOWCOME runs Delta Debugging on the deltas of the 1224 differing vertices (Figure 8 on the next page). As every second test fails, the difference narrows quickly. After 15 tests, HOWCOME has isolated a minimal failure-inducing difference—a single pointer adjustment, created with the GDB command

set variable link→fld[0].rtx→fld[0].rtx = link

This final difference is the difference that causes GCC to fail: It creates a cycle in the RTL tree—the pointer *link* → *fld*[0].*rtx* → *fld*[0].*rtx* points back to *link*! The RTL tree is no longer a tree, and this causes endless recursion in the function *if_then_else_cond*, eventually crashing *cc1*.

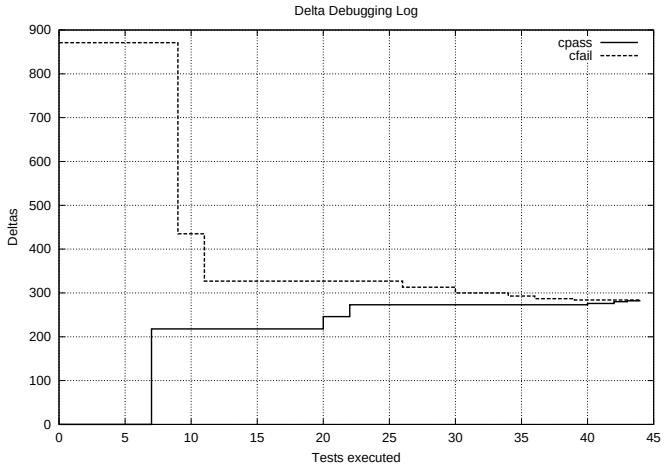


Fig. 7
NARROWING AT *combine_instructions*

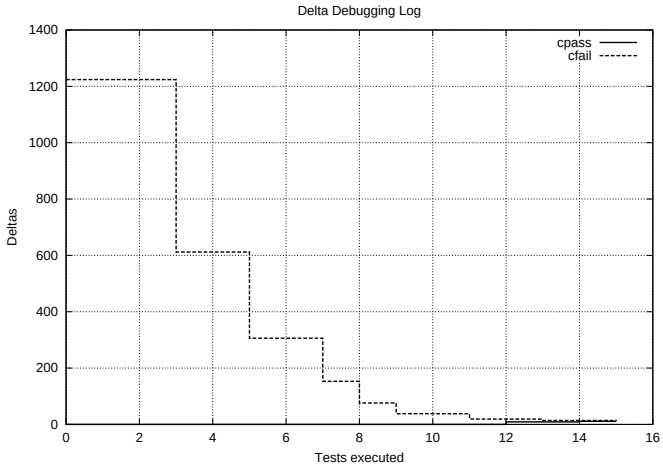


Fig. 8
NARROWING AT *if_then_else_cond*

D. The GCC Cause-Effect Chain

The total cause-effect chain for *cc1*, as reported by HOWCOME, looks like this:

Cause-effect chain for './gcc/cc1'
Arguments are -O fail.i (instead of -O pass.i)
therefore at main, *argv[2] = "fail.i" (instead of "pass.i")*
therefore at combine_instructions,
**first_loop_store_insn→fld[1].rtx→fld[1].rtx→*
fld[3].rtx→fld[1].rtx = ⟨new variable⟩
therefore at if_then_else_cond,
link→fld[0].rtx→fld[0].rtx = link (instead of i1dest)
therefore the run fails.

With this output, the programmer can easily follow the cause-effect chain from the root cause (the passed arguments) via an intermediate effect (a new node in the RTL tree) to the final effect (a cycle in the RTL tree). The whole run was generated automatically; no manual interaction was required. HOWCOME required 6 runs to extract GCC state (each taking 15–20 minutes) and 3 Delta Debugging runs (each taking 8–10 minutes) to isolate the failure-inducing differences.[5]

It should be noted again that the output above is produced in a fully automatic fashion. All the programmer has to specify is the program to be examined as well as the passing and failing invocations of the automated test. Given this information, HOWCOME then automatically produces the cause-effect chain as shown above.

IV. ISOLATING THE ERROR

The ultimate aim of debugging is to break the cause-effect chain such that the failure no longer occurs. Our cause-effect chain for GCC lists some possibilities: One could prevent an input of "+1.0", avoid PLUS operators in RTL or break cycles in the RTL tree. Again, from the standpoint of causality, each of these fixes is equivalent in preventing the failure.

For the programmer, though, these fixes are *not* equivalent—obviously, we need a fix that not only prevents the failure in question, but also prevents similar failures, while preserving the existing functionality. The programmer must thus choose the best place to break the cause-effect chain—a piece of code commonly referred to as "the error". Typically, this piece of code is found by determining the *transition* between an *intended* program state and a *faulty* program state. In the absence of an oracle, we must rely on the programmer to make this distinction: A cause can be determined automatically; the fault is in the eye of the beholder.

Nonetheless, cause-effect chains can be an effective help for the programmer to isolate the transition: All the programmer has to do is to decide whether the isolated state in the failing run is intended or not. In the GCC example, the states at *main* and at *combine_instructions* are intended; the RTL cycle at *if_then_else_cond* obviously is not. So, somewhere between *combine_instructions* and *if_then_else_cond*, the state must have changed from intended ("✔") to faulty ("✘"). We focus on this interval to isolate further differences.

Figure 9 shows the narrowing process. We isolate the failure-inducing state at some point in time between the locations *combine_instructions* and *if_then_else_cond*, namely at the file *combine.c* in line 1758: Here, the *newpat* variable points back to *link*—the cause for the cycle and thus a faulty state. The transition between intended and faulty state must have occurred between *combine_instructions* and line 1758.

Only two more narrowing steps are required: At line 4011, HOWCOME again isolates an additional PLUS node in the RTL tree—an intended effect of the "+1.0" input (not faulty);[6] at line 4271, HOWCOME again finds a failure-inducing RTL cycle (faulty). This isolates the transition down to lines 4013–4019. In this piece of code, executed only in the failing

[5] A non-prototypical implementation could speed up state access by 1–3 magnitudes by bypassing the GDB command line.

[6] Actually, HOWCOME reports this PLUS node as being located at *undobuf.undos→next→next→next→next→next→next→next→next→next*, which indicates that finding the most appropriate denomination for a memory location is an open research issue.

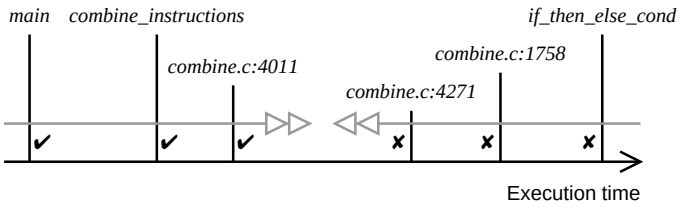


Fig. 9
NARROWING DOWN RELEVANT EVENTS

run, the RTL code (MULT (PLUS *a b*) *c*) is transformed to (PLUS (MULT *a* c_1)(MULT *b* c_2)) where $c = c_1 = c_2$ holds.[7] Unfortunately, c_1 and c_2 are created as *aliases* of *c*, which causes the cycle in the RTL tree! To fix the error, one should make c_2 a true copy of c_1—and this is how the error was fixed in GCC 2.95.3.

Do we really need the programmer to narrow down the point in time where the state becomes faulty? Not necessarily. First, one could simply increase the granularity of the cause-effect chain, and thus present more detailed information. Second (and this is a research topic), one could attempt to isolate *cause transitions* automatically. For instance, the narrowing process as shown above could also have been guided by the fact whether the RTL tree difference PLUS is relevant or not—and would have isolated the very same location. We are currently experimenting with different heuristics (such as code being executed in one run only) to automatically focus on events that are likely to be relevant.

V. CASE STUDIES

A. GNU Programs

Besides GCC, we have applied HOWCOME to some more GNU programs to isolate cause-effect chains (Table IV):

- In the *sample* example from the GNU DDD manual, Delta Debugging quickly isolated a bad *shell_sort* call.[8]
- In the *bison* parser generator, a shift/reduce conflict in the grammar input causes the variable *shift_table* to be altered, which in turn generates a warning.
- In the *diff* file comparison program, printing of differences is controlled by *changes*, whose value is again caused by *files*→ *changed_flag*.
- Invoking the *gdb* debugger with a different debuggee changes 18 variables, but only the change in the variable *arg* is relevant for the actual debuggee selection.

In all cases, the resulting failure-inducing difference contained only one relevant element (“Rel”); the number of tests was at most 42.

B. Siemens Programs

In a second case study, we applied Delta Debugging on the Siemens test suite [4], as summarized in Table V. Each program version (“Version”) was altered by hand to introduce one single new failure [5]. Each version was subject to Delta Debugging between 6 and 191 times (“Runs”); the average number of deltas between program states is shown in “Deltas”. The locations were automatically determined at 50%, 75%, and 90% of the execution coverage; the *test* function was determined automatically from the differing output (see Section VII for details how and why this is done). Overall, Delta Debugging had been run 1995 times.

Event	Edges	Vertices	Deltas	Tests	Rel
sample at *main*	26	26	12	4	1
sample at *shell_sort*	26	26	12	7	1
sample at *sample.c*:37	26	26	12	4	1
cc1 at *main*	27139	27159	1	0	1
cc1 at *combine_instructions*	42991	44290	871	44	1
cc1 at *if_then_else_cond*	47071	48473	1224	15	1
bison at *open_files*	431	432	2	2	1
bison at *initialize_conflicts*	1395	1445	431	42	1
diff at *analyze.c*:966	413	446	109	9	1
diff at *analyze.c*:1006	413	446	99	10	1
gdb at *main.c*:615	32455	33458	1	0	1
gdb at *exec.c*:320	34138	35340	18	7	1

TABLE IV
SUMMARY OF GNU CASE STUDIES

The number of relevant variables in the cause-effect chain never exceeded 2; the average number is shown in the “Relevant” column. The number of tests required to isolate the relevant variables (“Tests”) is pretty much logarithmic with respect to the number of deltas.

Note to reviewers: The final version will include a statement on the effectiveness of the deltas—i.e. how well do the isolated causes refer to statements that were altered to induce the error.

C. Lessons Learned

What we found most surprising about these experiments was that one can alter program variables to more or less meaningless values and get away with it. We made the following observations, all used by Delta Debugging:

- First, the altered values are not meaningless; they stem from a consistent state, and it is only a matter of statistics (e.g. which and how many variables are transferred) whether they induce an inconsistent state. The chances for consistency can be increased by grouping variables according to the program structure (which HOWCOME does not do yet).
- Second, the remainder of the program (and the final *test* function) acts as a *filter*: If anything happens that did not happen in the two original runs, the test outcome becomes unresolved, and the next alternative is sought. If variables have been altered and the outcome is still similar to the original two runs, then these variables are obviously irrelevant for the outcome. Precision can be arbitrarily increased by making the *test* function pickier about similarity [1].
- Third, in a program with a good separation of concerns, only a few variables should be responsible for a specific behavior, including failures—and this small number makes Delta Debug-

[7] This application of the distributive law allows for potential optimizations, especially for addresses.

[8] A HOWCOME demonstration program for this example, is available online, including sample Delta Debugging source code [3].

Name	Version	Runs	$\overline{\text{Deltas}}$	$\overline{\text{Tests}}$	$\overline{\text{Relevant}}$
print_tokens	2	55	44.25	9.51	1.31
print_tokens	3	7	52.14	15.00	1.71
print_tokens	4	23	36.22	9.04	1.43
print_tokens	5	2	59.00	16.00	2.00
print_tokens	6	47	49.04	12.81	1.66
print_tokens	7	11	59.00	6.55	1.00
print_tokens2	1	36	12.33	4.86	1.28
print_tokens2	2	22	19.86	6.77	1.27
print_tokens2	3	51	14.41	5.71	1.35
print_tokens2	4	32	27.75	6.03	1.28
print_tokens2	5	54	14.13	5.85	1.44
print_tokens2	6	33	16.42	6.76	1.42
replace	1	17	17.29	7.24	1.29
replace	2	16	21.88	9.12	1.31
replace	3	6	20.00	8.17	1.33
replace	7	39	16.97	4.72	1.08
replace	8	18	20.61	7.89	1.28
replace	9	12	20.17	7.67	1.25
schedule	1	21	46.24	6.29	1.05
schedule	3	20	85.50	9.75	1.10
schedule2	1	29	49.17	8.93	1.14
schedule2	2	56	42.62	7.46	1.16
schedule2	3	92	48.09	6.26	1.09
schedule2	4	35	67.71	6.57	1.11
schedule2	5	179	58.44	6.82	1.08
schedule2	6	191	62.24	5.82	1.02
schedule2	7	132	48.14	6.64	1.03
schedule2	8	187	58.42	6.35	1.01
schedule2	10	164	48.68	7.10	1.13
tcas	38	10	39.70	5.60	1.00
tcas	39	29	38.41	5.59	1.00
tcas	40	18	32.94	5.89	1.00
tcas	41	44	35.64	7.45	1.11
tot_info	1	105	721.34	11.30	1.14
tot_info	4	48	444.62	11.06	1.19
tot_info	5	154	693.35	30.68	1.88

TABLE V
SUMMARY OF SIEMENS CASE STUDIES

ging efficient.

- Fourth and last, program state has a *structure* and can thus easily be decomposed. In contrast, decomposing *input* as sketched in Section II requires the input syntax to be specified manually for each new program. And, of course, isolating relevant states is much more valuable than isolating input alone, since we can actually look at what's going on inside the program.

VI. LIMITS

Let us now discuss some general limits of Delta Debugging, and how we can deal with them.

A. Alternate Runs

Delta Debugging always requires an *alternate run* in order to compare states and narrow down differences. However, "alternate run" does not necessarily translate into "alternate invocation", at least not of the whole program. If one wants to to isolate the cause-effect chain from some *subprogram*, it suffices to have multiple invocations of that subprogram within a single program run. The only requirement is that each invocation can be individually tested for success or failure.

For instance, GCC invokes *combine_instructions* for each compiled procedure. Let us assume we have one single input for *cc1* that first includes the contents of `"pass.i"` and then the contents of `"fail.i"`. We can compare the two invocations of *combine_instructions* just like we compared the two invocations of GCC—and find out how the difference propagates in the cause-effect chain. We expect that many programs contain such *repetitive elements* that can be exploited by Delta Debugging; in fact, *multiple repetitions* such as loop traversals may turn out as a nice source for regular behaviour.

If a program (or subprogram) fails under all conditions, there is no way Delta Debugging (or any other automated process) could make it work. Here, anomaly detection techniques (Section VIII-B) are probably a better choice.

B. State Transfer

In practice, transferring a program state (or parts thereof) from one run to another run can be very tricky. Problems occur with both *internal* and *external* state.

In the C language, the programmer has complete freedom to organize dynamic memory, as long as he keeps track of memory contents. Consequently, there is no general way

- to detect whether a pointer points to an element, an array, or nothing,
- to determine the sizes of dynamically allocated arrays, or
- to find out which member of a union is currently being used.

In all these cases, HOWCOME relies on *heuristics* that query the run-time library for current memory layout, or check for common conventions of organizing memory. These heuristics can be wrong, which may or may not lead to unresolved test outcomes. A better way would be to keep track of all memory contents at run-time such that questions like the ones above can be answered adequately. Fortunately, in languages other than C (notably languages with garbage collection), such heuristics are not required.

A program state may encompass more than just memory, though. As an example, think of the state of a database. If we have a query that works and a query that fails, should we include the entire database in our state, and if so, at which abstraction level?

As another example, think of file handles. We might be able to transfer file handles from one state to another, but should we really do so without considering the state of the file?

One can argue that every external state becomes part of internal memory as soon as it is accessed - databases and files are read and written. But differences may also reside *outside* of the program state—for instance, a file handle may have the same value in $r_{✘}$ and $r_{✔}$, but be tied to a different file. We are working on how to capture such external differences.

C. False Positives

The states generated by Delta Debugging can make the program fail in a number of ways. One must take care not to confound these with the "original" failure.

As an example, consider the following piece of code:

```
bool x := ...;
bool y := x;
⇐ state is accessed here
if (x ≠ y)
  fail();
else if (x ∧ y)
  fail();
fi
```

Let us assume that in a Delta Debugging run, we find the following configurations:

Run	x	y
$c_{✔}$	false	false
$c_{✘}$	true	true
Tested c	false	true

The tested configuration c causes the program to fail (the *fail* function is invoked). However, the failure occurs at a different invocation. Thus, the program fails in a way that is different from the original failure.

A pragmatic way to handle with such false positives is to take the *backtrace* of the failure into account—the stack of functions that were active at the moment of the failure. The $test$ function should return ✘ if and only if the failure occurs at the same location as $r_{✘}$—that is, the program counter and the backtrace of calling functions must be identical. This need not be a program crash; one could also consider the backtrace of an output or exit statement, for instance.

In the example above, the second *fail* invocation results in a different backtrace—$test$ would return **?** rather than ✘.

Besides the backtrace, there are more things one can consider—the execution time, for instance, the set of reached functions, or the final state of the execution. If one requires a higher similarity with the failing run, though, one will also obtain a larger set of causes that is required to produce exactly this similarity.

D. Infeasible Context

Each cause, as reported by Delta Debugging, consists of two configurations (states) $c'_{✘}$ and $c'_{✔}$ such that the difference $\Delta = c'_{✘} - c'_{✔}$ is minimal. This difference Δ between states determines whether the outcome is ✔ or ✘ and thus is an actual failure cause.

However, Δ is a failure cause only in a specific *context*—the configuration $c'_{✔}$—and this context may or may not be related to the original passing or failing runs.

Here's a code example that exhibits this problem:

```
if (a ∧ ¬ b ∧ c) ∨ (c ∧ d))
  fail();
fi
```

Let us assume that after a Delta Debugging run, we find the following configurations:

Run	a	b	c	d
$c_{✔}$	false	false	false	false
$c_{✘}$	true	true	true	true
Found $c'_{✔}$	false	false	true	false
Found $c'_{✘}$	true	false	true	false

The difference in a between $c'_{✔}$ and $c'_{✘}$ is failure-inducing—a is a failure cause, at least in the common context of b, c, and d being false, true, and false, respectively. But altering a alone in $c_{✔}$ or $c_{✘}$ does not change the outcome—one needs to change at least one further variable of the context, too. This is no problem if the common context is feasible, i.e. could be reached within a program run. However, Delta Debugging does not guarantee feasibility.

This example raises several questions. Does it harm if a is reported as a failure cause, although it is only part of a cause? How do we deal with infeasible contexts? How can we relate the isolated difference to the original runs? How likely is it that such interferences occur? Such questions require much further research and experience.

E. Differences vs. Faults

The chosen alternate run determines the causes Delta Debugging can infer: A variable v can be isolated as a cause only if it exists in both runs and if its value differs. However, v may not be faulty at all; it may very well be that v has exactly the intended value. In the GCC example, the PLUS node in the RTL tree was perfectly legal, although it eventually caused the failure.

In the absence of a specification, no automated process can distiguish faulty from intended values. However, it may be helful to further understand how state differences cause a difference in the later computation. For instance, if some variable like the PLUS node causes specific pieces of the code to be executed (which otherwise would not) or specific variables to be accessed (which otherwise would not be), then it may be helpful to report these differences as causes, too.

F. Multiple Causes

Delta Debugging as presented here isolates only one cause from several potential causes—for instance, *fail.c* can be changed in several ways besides removing "+1.0", and so can the induced states. Although Delta Debugging could easily be extended to search for alternative causes—which is the most relevant cause, then, to present to the programmer?

We are currently experimenting with indications for relevance by analyzing *multiple* runs. If a failure cause occurs in several failing runs, it is more relevant than a cause that occurs only in few failing runs—in other words, altering the "frequent" cause fixes more failures than the "non-frequent" cause. Such multiple runs can either be given, or generated from random delta configurations.

G. Efficiency

In our case studies, Delta Debugging was always efficient. This is not necessarily the case, though. Suppose you have a program, which, among other things, verifies a checksum c over a message m. Now let two runs differ by exactly these two items c and m.

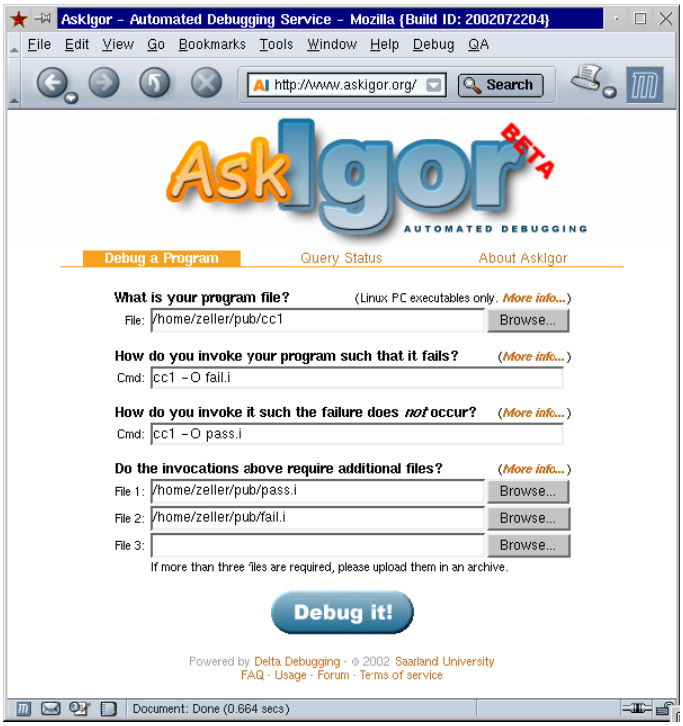


Fig. 10
THE *AskIgor* PUBLIC DEBUGGING SERVER

Any applied delta invalidates the relationship between checksum and message—in fact, you can either have no delta or all of them. Delta Debugging would nonetheless try to apply every single delta individually. Assuming that there is a delta for every single message character, there this requires a test run for each character in the message.

A pragmatic solution would be to process in a *hierarchical* fashion: Try to group all variables according to the object(s) they apply to—for instance, apply all deltas related to m in a group as long as possible, and break the group only when all else fails. This way, even a premature termination of the search process would at least return the differences in c and m, without going into all the details of m.

VII. A DEBUGGING SERVER

To further examine the limits and possible alternatives, we have built a *debugging server* named *AskIgor* [6] where anyone can submit failing programs via the Web to have HOWCOME compute a failure diagnosis. The main design issue in *AskIgor* was to make the use of the service as easy as possible.

From a user's perspective, *AskIgor* works as follows:

1. The user invokes *AskIgor* at the Web address [6] (Figure 10).
2. The user submits
   - the executable to be debugged,
   - an invocation $r_{\text{✘}}$ where the program fails,
   - an invocation $r_{\text{✔}}$ where the program does not fail,
   - (optionally) additional files required to make the program run, such as input or configuration files.
3. After a few minuts, the user obtains a diagnosis on a separate Web page, containing the cause-effect chain of the failure.

The user need not specify a *test* function; *AskIgor* automatically generates a *test* function based on the differing behavior of the two given invocations; the criteria for ✔ or ✘ include the program output and the exit status. The user has the chance to provide her own *test* function by specifying alternate failure symptoms.

By default, *AskIgor* computes a cause-effect chain at three events: when 50%, 75%, and 90% of the functions executed in the failing run have been covered. The rationale behind such a coverage-based metric is that the likelihood of an error is the same across the program code; hence, we choose coverage rather than execution time as a metric base. The level of detail increases at the end of the execution, since these events are more likely to be relevant for the failure.

To narrow down errors interactively, as sketched in Section IV, *AskIgor* provides a "How did this happen" feature. For every event e, the user can select to obtain a cause-effect chain for the interval between e and the previous event. This way, the user can increase the granularity of the cause-effect chain as desired.

On March 10, 2003, a total of 150 programs had been submitted to *AskIgor*; more than two thirds of these programs could be processed successfully. The current summary of submissions is available to the public via the status page [6].

As *AskIgor* requests feedback from its users, we will be able to evaluate the effectiveness and usability of our diagnoses for a large number of real-life case studies. We plan to extend *AskIgor* to accomodate and combine a variety of services for program comprehension, including program slicing and anomaly detection (Figure 11).

In the future, services like *AskIgor* will be part of integrated programming environments. Then, users would not have to submit their programs explicitly; instead, the cause-effect chain would automatically be computed as soon as a test fails. This way, users would not only learn *that* their test has failed, but also *why* it failed—and all without the least bit of interaction.

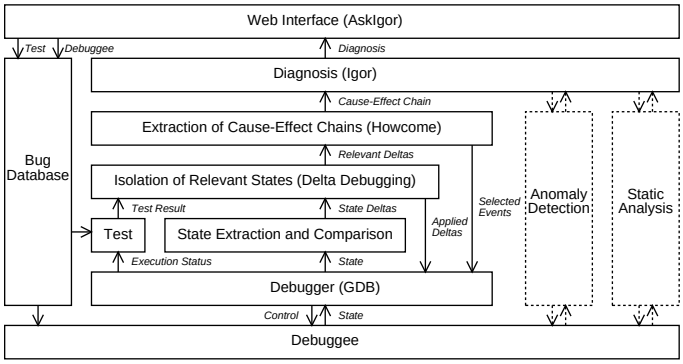


Fig. 11
THE *AskIgor* ARCHITECTURE

VIII. Related Work

A. Program Slicing

Program slicing [7], [8] facilitates debugging by focusing on relevant program fragments. A *slice* for a location p in a program consists of all other locations that could possibly influence p ("all locations that p depends upon"). As an example, consider the code fragment

$$\textbf{if}\ \ p\ \ \textbf{then}\ \ x' := x * y\ \ \textbf{fi}$$

Here, the variable x' is control dependent on p and data dependent on x and y (but *not* on, say, z); the slice of x' would also include earlier dependencies of p, x, and y. The slice allows the programmer to focus on relevant statements; a slice also has the advantage that it is valid for *all possible program runs* and thus needs to be computed only once.

Dynamic slicing [9], [10], [11] is a variant of slicing that takes a *concrete program run* into account. The basic idea is that within a concrete run, one can determine more accurate data dependencies between variables, rather than summarizing them as in static slicing. In the dynamic slice of x', as above, x' is dependent on x, y, and p only if p was found to be true.

In practice, slicing is not yet as useful as would be expected, since each statement is quickly dependent on many other statements. The end result is often a program slice which is not dramatically smaller than the program itself—the program dependencies are too coarse [12]. Also, data and control-flow analysis of real-life programs is non-trivial. For programs with pointers, the necessary points-to analysis makes dependencies even more coarse [13].

The GCC example also illustrates the limits of slicing. In *combine_instructions* (Section III-C.2), we found 871 deltas between the passing and the failing run. Any slicing method, static or dynamic, would have to incorporate these 871 variables in the forward slice of the changed input. Although 871 variables is just 2% of all variables, it is still a large number in absolute terms.

In *cause-effect chains*, p, x, and y are the cause for the value of x' if and only if altering them also changes the value of x', as proven by test runs. If $x = 0$ holds, for instance, p can never be a cause for the value of x', because x' will never alter its value; y cannot be a cause, either. Consequently, cause-effect chains have a far higher precision than static or dynamic slices—returning only 1 out of 871 changes as relevant for the failure.

On the other hand, cause-effect chains require several test runs (which is possibly slower than analysis), apply to a single program run only, and give no hints on the involved statements. The intertwining of program analysis and experimentation promises several mutual advantages.

B. Detecting Anomalies

Dicing [14] determines the *difference* of two program slices. For instance, a dynamic dice could contain all the statements that may have influenced a variable v at some location in a failing run $r_{\text{✘}}$, but not in a passing run $r_{\text{✔}}$. The dice is likely to include the statement relevant for the value of v.

Running several tests at once allows one to establish *relationships* between the executed code and the test outcome. For instance, one could isolate code that was only executed in failing tests [15]. This differential approach would also have isolated the erroneous code in our GCC example.

Dynamic invariants [16] can be used to detect anomalous program behavior [17]. During execution, a tool checks the program against a model that is continuously updated; invariant violations can be immediately reported. This approach has several exciting uses; one related to our work is to check a failing run against invariants obtained from a passing run.

As discussed in Section IV, the idea that an automated process could isolate "the" erroneous code automatically in the absence of an oracle can only be based on *heuristics*, and this is what these approaches provide—including the risk of being misleading. Nonetheless, a heuristic can be very good at isolating possible causes; and it can be even more helpful when guiding a trial-and-error approach like Delta Debugging.

C. The Debugging Process

Algorithmic debugging [18] automates the debugging *process*. The idea is to isolate a failure-inducing clause in a PROLOG program by querying systematically whether subclauses hold or not. The query is resolved either manually by the programmer or by an oracle relying on an external specification. This could easily be combined with our approach to narrow down the failure-inducing code as discussed in Section IV: "Is PLUS in the RTL tree correct (y/n)?"

D. Testing for Debugging

Surprisingly, there are very few applications of testing for purposes of debugging or program understanding. Our own contributions [1] as well as inferring relationships between code and tests [15] have already been mentioned.

Specifically related to our GCC case study is the isolation of failure-inducing RTL optimizations in a compiler, using simple binary search over the optimizations applied [19]. An experimental approach comparable to Delta Debugging is *change impact analysis* [20], identifying code changes that are relevant for a failure.

IX. Conclusion and Consequences

Cause-effect chains explain the causes of program failures automatically and effectively. All that is required is an automated test, two comparable program runs and access to the state of an executable program. Although relying on several test runs to prove causality, the isolation of cause-effect chains requires no manual interaction and thus saves valuable developer time.

As the requirements are simple to satisfy, we expect that future automated test environments will come with an automatic isolation of cause-effect chains. Whenever a test fails, the cause-effect chain could be automatically isolated, thus showing the programmer not only *what* has failed, but also *why* it has failed. Although fixing the program is still manual (and creative) work, we expect that the time spent for debugging will be reduced significantly.

All this optimism should be taken with a grain of salt, as there is still much work to do. Our future work will address the limits discussed in Section VI, concentrating on the following topics:

Optimization. As stated in Section III-D, HOWCOME could be running faster by several orders of magnitude by bypassing the GDB bottleneck and re-implementing HOWCOME in a compiled language. Regarding Delta Debugging, we are working on *grouping variables* such that variables related by occurring in the same function or module are changed together, rather than having a random assignment of variables to subsets.

Program analysis. As hinted at in Section VIII, the integration of program analysis could make extracting cause-effect chains much more effective. For instance, variables that cannot influence the failure in any way could be excluded right from the start. Anomaly detection could help to guide the search towards specific variables or specific events.

Greater state. Right now, our method only works on the state that is accessible via the debugger. However, differences may also reside *outside* of the program state—for instance, a file descriptor may have the same value in $r_{✘}$ and $r_{✔}$, but be tied to a different file. We are working on how to capture such external differences.

A discipline of debugging. Notions like causes and effects and approaches like running experiments under changed circumstances can easily be generalized to serve in arbitrary debugging contexts. We are currently compiling a *textbook* [21] that shows how debugging can be conducted as systematically and as all other software engineering disciplines—be it manually or automated.

Overall, we expect that debugging may become as automated as testing—not only detecting *that* a failure occurred, but also *why* it occurred. And since computers were built to relieve humans from boring, monotonous tasks—let's have them do the debugging!

Acknowledgments. The concept of isolating cause-effect chains was born after a thorough discussion with Jens Krinke on the respective strengths and weaknesses of program slicing and Delta Debugging. Tom Zimmermann implemented the initial memory graph extractor and the common subgraph algorithms. Holger Cleve conducted the Siemens experiments. Holger Cleve, Stephan Diehl, Petra Funk, Kerstin Reese, Barbara Ryder, and Tom Zimmermann provided substantial comments on earlier versions of this paper. Many thanks go to the anonymous reviewers for their constructive and helpful comments.

More information on isolation of cause-effect chains, including a HOWCOME demonstration program, is available at the Delta Debugging web site [3].

Figure 4 was generated by AT&T's DOT graph layouter; Figure 6 is a screenshot of Tamara Munzner's H3VIEWER.

Delta Debugging research is funded by Deutsche Forschungsgemeinschaft, grant Sn 11/8-1.

REFERENCES

[1] Andreas Zeller and Ralf Hildebrandt, "Simplifying and isolating failure-inducing input," *IEEE Transactions on Software Engineering*, vol. 28, no. 2, pp. 183–200, Feb. 2002.

[2] Thomas Zimmermann and Andreas Zeller, "Visualizing memory graphs," in *Proc. of the International Dagstuhl Seminar on Software Visualization*, Stephan Diehl, Ed. 2002, vol. 2269 of *LNCS*, pp. 191–204, Springer-Verlag.

[3] "Delta debugging web site," `http://www.st.cs.uni-sb.de/dd/`.

[4] Monica Hutchins, Herb Foster, Tarak Goradia, and Thomas Ostrand, "Experiments of the effectiveness of dataflow- and controlflow-based test adequacy criteria," in *Proceedings of the 16th International Conference on Software Engineering*. 1994, pp. 191–200, IEEE Computer Society Press.

[5] Gregg Rothermel and Mary Jean Harrold, "A safe, efficient regression test selection technique," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 6, no. 2, pp. 173–210, 1997.

[6] "*AskIgor* web site," `http://www.askigor.org/`.

[7] Frank Tip, "A survey of program slicing techniques," *Journal of Programming Languages*, vol. 3, no. 3, pp. 121–189, Sept. 1995.

[8] M. Weiser, "Programmers use slices when debugging," *Communications of the ACM*, vol. 25, no. 7, pp. 446–452, 1982.

[9] Hiralal Agrawal and Joseph R. Horgan, "Dynamic program slicing," in *Proceedings of the ACM SIGPLAN 1990 Conference on Programming Language Design and Implementation (PLDI)*, White Plains, New York, June 1990, vol. 25(6) of *ACM SIGPLAN Notices*, pp. 246–256.

[10] Tibor Gyimóthy, Árpád Beszédes, and István Forgács, "An efficient relevant slicing method for debugging," in *Proc. ESEC/FSE'99 – 7th European Software Engineering Conference / 7th ACM SIGSOFT Symposium on the Foundations of Software Engineering*, Toulouse, France, Sept. 1999, vol. 1687 of *LNCS*, pp. 303–321, Springer-Verlag.

[11] Bogdan Korel and Janusz Laski, "Dynamic slicing of computer programs," *The Journal of Systems and Software*, vol. 13, no. 3, pp. 187–195, Nov. 1990.

[12] Daniel Jackson and Eugene J. Rollins, "A new model of program dependences for reverse engineering," in *Proc. 2nd ACM SIGSOFT symposium on the Foundations of Software Engineering (FSE-2)*, New Orleans, Lousiana, Dec. 1994, pp. 2–10.

[13] Michael Hind and Anthony Pioli, "Which pointer analysis should I use?," in *Proc. ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*, Portland, Oregon, Aug. 2000, pp. 113–123.

[14] J. R. Lyle and M. Weiser, "Automatic program bug location by program slicing," in *2nd International Conference on Computers and Applications*, Peking, 1987, pp. 877–882, IEEE Computer Society Press, Los Alamitos, California.

[15] James A. Jones, Mary Jean Harrold, and John Stasko, "Visualization of test information to assist fault localization," In ICSE 2002 [22].

[16] Michael D. Ernst, Jake Cockrell, William G. Griswold, and David Notkin, "Dynamically discovering likely program invariants to support program evolution," *IEEE Transactions on Software Engineering*, vol. 27, no. 2, pp. 1–25, Feb. 2001.

[17] Sudheendra Hangal and Monica S. Lam, "Tracking down software bugs using automatic anomaly detection," In ICSE 2002 [22].

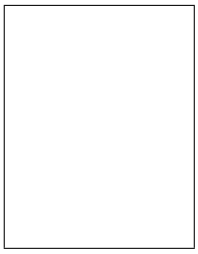
[18] Ehud Y. Shapiro, *Algorithmic Program Debugging*, Ph.D. thesis, MIT Press, 1982, ACM Distinguished Dissertation.

[19] David B. Whalley, "Automatic isolation of compiler errors," *ACM Transactions on Programming Languages and Systems*, vol. 16, no. 5, pp. 1648–1659, 1994.

[20] Barbara G. Ryder and Frank Tip, "Change impact analysis for object-oriented programs," in *Proc. of the Workshop on Program Analysis for Software Tools and Engineering (PASTE)*, Snowbird, Utah, June 2001, pp. 46–53.

[21] Andreas Zeller, *Automated Debugging*, Morgan Kaufmann Publishers, Aug. 2003, To appear (ISBN 1-55860-866-4).

[22] *Proc. International Conference on Software Engineering (ICSE)*, Orlando, Florida, May 2002.



Andreas Zeller received his PhD in computer science from Technische Universität Braunschweig, Germany in 1997. He is currently a professor of computer science at Universität des Saarlandes, Saarbrücken, Germany. His general research interest is program analysis and automated debugging – and how to make these techniques applicable to real programs with real problems. He is a member of the IEEE computer society, the ACM, and the GI.