

Finding Failure Causes through Automated Testing

Holger Cleve
Universität Passau
Lehrstuhl Software-Systeme
Innstraße 33
94032 Passau, Germany
+49 851 509-3094
cleve@fmi.uni-passau.de

Andreas Zeller
Universität Passau
Lehrstuhl Software-Systeme
Innstraße 33
94032 Passau, Germany
+49 851 509-3095
zeller@acm.org

Abstract

A program fails. Under which circumstances does this failure occur? One single algorithm, the *delta debugging* algorithm, suffices to determine these *failure-inducing circumstances*. Delta debugging tests a program systematically and automatically to isolate failure-inducing circumstances such as the program input, changes to the program code, or executed statements.

Keywords: *Testing and debugging, debugging aids, combinatorial testing, execution tracing*

1 Debugging by Testing

Debugging falls into three phases: reproducing a failure, finding the root cause of the failure, and correcting the error such that the failure no longer occurs. While failure reproduction and correction are important issues, it is the second phase, finding the root cause, which is the most significant. Early studies have shown that finding the root cause accounts for 95% of the whole debugging effort [6].

The common definition of a cause is *some preceding event without which the effect would not have occurred*. This implies that any claim for causality can only be verified by experimentation. To prove that an event is the cause, the effect must no longer occur if the event is missing.

In the context of debugging, the “effect” is a failure, and the “event” is some circumstance of the program execution. To prove that some circumstance has caused a failure, one must remove it in a way that the failure no longer occurs. This is typically done as the very last debugging step: After correcting the error, one re-tests the program, verifies that the failure is gone, and thus proves that the original error was indeed the failure cause.

In order to prove causality, there can be no substitute for testing—not even the most sophisticated analysis of the original program run. In other words: *Analysis can show the absence of causality, but only testing can show its presence.*

In this paper, we give an overview on *delta debugging*—an automated debugging method that relies on systematic testing to prove and isolate failure causes—circumstances such as the program input, changes to the program code, or executed statements. Basically, delta debugging sets up subsets of the original circumstances, and tests these configurations whether the failure still occurs. Eventually, delta debugging returns a subset of circumstances where every single circumstance is relevant for producing the failure.

This paper unifies our previous work on delta debugging [3, 8] by showing that a single algorithm suffices. For the first time, the isolation of failure-inducing statements is discussed. After basic definitions and a discussion of the algorithm, we show how to apply delta debugging to identify failure-inducing changes, failure-inducing program input, and failure-inducing statements. We close with discussions of related and future work.

2 Configurations and Tests

Let us begin with some basic definitions. First of all, what are the “circumstances” of a failure? Roughly speaking, a failure circumstance is anything that might influence the existence of the failure. Without any further knowledge, this is anything that may influence the program’s execution: its environment, its input, its code. All these are circumstances of a failure.

We call a set of circumstances a *scenario*. Obviously, the root cause of a problem is somewhere within this scenario. To isolate the root cause, we must separate the chaff from the wheat, or irrelevant from relevant failure circumstances.

Let us now have a specific failing scenario to investigate. Normally, we also know a *working scenario* under which the failure does *not* occur. Let us assume we have some *gradual transition* from the working scenario to the failing scenario—for instance, by adding or altering circumstances. The idea then is to systematically test scenarios along this transition in order to isolate failure-inducing circumstances

and to use the test outcome to generate new hypotheses.

Formally, we view a *failure-inducing scenario* C as the result of applying a number of *changes* $\Delta_1, \Delta_2, \dots, \Delta_n$ to some working scenario. This way, we have a gradual *transition* from the working scenario (= no changes applied) to C (= all changes applied).

We can describe any scenario between the working scenario and C as a *configuration of changes*:

Definition 1 (Scenario) Let $C = \{\Delta_1, \Delta_2, \dots, \Delta_n\}$ be a set of changes. A change set $c \subseteq C$ is called a *scenario*. \square

A scenario is constructed by applying changes to the working scenario:

Definition 2 (Working scenario) An empty scenario $c = \emptyset$ is called the *working scenario*. \square

We do not impose any constraints on how changes may be combined; in particular, we do not assume that changes are ordered. Thus, in the worst case, there are 2^n possible scenarios for n changes.

To determine whether a scenario fails, we assume a *testing function* with three standard outcomes:

Definition 3 (Test) The function $test : 2^C \rightarrow \{\mathbf{X}, \checkmark, ?\}$ determines for a scenario whether some given failure occurs (\mathbf{X}) or not (\checkmark) or whether the test is unresolved ($?$). \square

In practice, $test$ would construct the scenario by applying the given changes to the working scenario, execute the scenario and return the outcome.

Let us now model our initial setting. We have some *working scenario* that works fine and some scenario that fails:

Axiom 4 (Scenarios) $test(\emptyset) = \checkmark$ (“working scenario”) and $test(C) = \mathbf{X}$ (“failing scenario”) hold. \square

Our goal is now to *minimize* the failing scenario C —that is, making it as similar as possible to the working scenario. A scenario c being “minimal” means that no subset of c causes the test to fail. Formally:

Definition 5 (Minimal scenario) A scenario $c \subseteq C$ is *minimal* if $\forall c' \subset c (test(c') \neq \mathbf{X})$ holds. \square

This is what we want: to minimize a scenario such that *all circumstances are relevant in producing the failure*—removing any change causes the failure to disappear.

3 Minimality of Scenarios

How can one actually determine a minimal scenarios? Here comes bad news. Let there be some scenario c consisting of $|c|$ changes to the minimal scenario. Relying on $test$ alone to

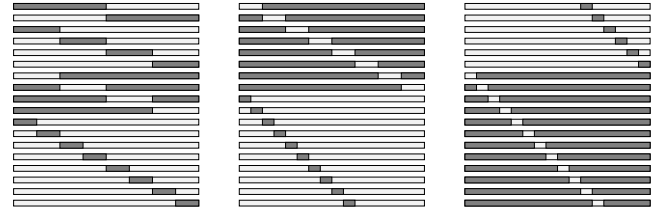


Figure 2: Tests carried out by *dadmin*

determine minimality requires testing all $2^{|c|} - 1$ true subsets of c , which obviously has exponential complexity.

What we can determine, however, is an *approximation*—for instance, a scenario where still every part on its own is significant in producing the failure, but we do not check whether removing several parts at once might make the scenario even smaller. Formally, we define this property as *1-minimality*, where n -minimality is defined as:

Definition 6 (n -minimality) A scenario $c \subseteq C$ is *n -minimal* if $\forall c' \subset c (|c| - |c'| \leq n \Rightarrow (test(c') \neq \mathbf{X}))$ holds. \square

If c is $|c|$ -minimal, then c is minimal in the sense of Definition 5.

Definition 6 gives a first idea of what we should be aiming at. However, given a scenario with, say, 100,000 changes, we cannot simply undo each individual change in order to minimize it. Thus, we need an effective algorithm to reduce our scenario efficiently.

An example of such a minimization algorithm is the *minimizing delta debugging algorithm* *dadmin*, shown in Figure 1 on the following page and discussed in [3]. *dadmin* is based on the idea of *divide and conquer*. Initially, it partitions the set of changes into two subsets c_1 and c_2 and tests each of them: if any test fails, the search is continued with this reduced subset.

If no test fails, *dadmin* *increases the granularity* by *doubling* the number of subsets. It then tests each subset and each *complement*; if the test fails, the search is continued with this reduced subset. The process is continued until each subset has a size of 1 and no further reduction is possible.

The whole process is illustrated in Figure 2. Here, we assume that every test outcome is unresolved. We see how *dadmin* first partitions the whole set of changes (a rectangle) into two subsets (gray areas), then into four, then into eight, sixteen, etc., testing each subset as well as its complement.

The *dadmin* algorithm guarantees that the returned set is 1-minimal; that is, no single change that can be removed such that the test still fails. In the last stages of Figure 2, we see how this guarantee is achieved: *dadmin* returns only when all complements have passed the test, and this means that each remaining change has been removed at least once.

This guarantee comes at a price. In the worst case (every test fails but the last complement), *dadmin* requires up to $n^2 +$

Minimizing Delta Debugging Algorithm

The *minimizing delta debugging algorithm* $ddmin(c)$ is

$$ddmin(c) = ddmin_2(c, 2) \quad \text{where}$$

$$ddmin_2(c, n) = \begin{cases} ddmin_2(c_i, 2) & \text{if } test(c_i) = \mathbf{X} \text{ for some } i \text{ ("reduce to subset")} \\ ddmin_2(\bar{c}_i, \max(n-1, 2)) & \text{else if } test(\bar{c}_i) = \mathbf{X} \text{ for some } i \text{ ("reduce to complement")} \\ ddmin_2(c, \min(|c|, 2n)) & \text{else if } n < |c| \text{ ("increase granularity")} \\ c & \text{otherwise ("done").} \end{cases}$$

where $c_1, \dots, c_n \subseteq c$ such that $\bigcup c_i = c$, all c_i are pairwise disjoint, $\forall c_i (|c_i| \approx |c|/n)$, as well as $\bar{c}_i = c - c_i$.

The recursion invariant (and thus precondition) for $ddmin_2$ is $test(c) = \mathbf{X} \wedge n \leq |c|$.

Figure 1: Minimizing delta debugging algorithm

$3n$ tests for n changes [3]. However, in the best case (every test fails), $ddmin$ requires only $\log_2 n$ tests.

The performance of $ddmin$ can be dramatically improved if we know that once a set of changes passes the test, then every subset passes the test as well—the so-called *monotony*. If we know that a set of changes is monotone, we need not test sets whose supersets have already passed the test: $test$ can simply return \checkmark without actually carrying out the test.

This optimization makes $ddmin$ linear at worst—but only if there are no unresolved test outcomes. The number of tests required by $ddmin$ largely depends on our ability to group the scenario transitions such that we avoid unresolved test outcomes and increase our chances to get failing tests.

4 Finding Failure-Inducing Changes

In our first case study [8], we applied delta debugging to a common, yet simple regression case. The situation is as follows: There is some old version of the program (“yesterday”), which works, and a new version (“today”) which does not. The goal is to identify the changes to the program code which induce the failure.

This situation fits nicely into the scenario model of Section 2. The “yesterday” version is the working scenario; the changes are textual changes to the program code; the “today” version is the failing scenario where all changes are applied. Using the scenario minimization algorithm, we should be able to minimize the set of applied changes and thus identify a small set of failure-inducing changes.¹

Our results with delta debugging were quite promising. First, if the dependency between changes was known (from a version history, for instance), then delta debugging became quite trivial: If every change Δ_i depended on earlier

¹In this first case study, we actually used a variant of $ddmin$, called dd^+ [8]. For the examples in this Section, dd^+ has exactly the same performance as $ddmin$ under monotony; i.e. $test(c)$ returns \checkmark without actually testing c if a superset of c has already passed the test. $ddmin$ does not guarantee 1-minimality, which is why we recommend $ddmin$ (possibly with the monotony optimization), as a general replacement for dd^+ .

changes $\Delta_1, \dots, \Delta_{i-1}$, then every scenario which would not fulfill these constraints could be rejected straight away. Under these conditions, only configurations with a full set of changes $\Delta_1, \dots, \Delta_{i-1}$ actually had to be tested— $ddmin$ degraded into a simple *binary search* with logarithmic complexity.

But even when the dependency between changes was not known, or when a single logical change was still too large, $ddmin$ performed satisfactorily. In one case study, we had a single 178,000-line change to the GNU debugger (GDB); this change was broken down into 8721 textual changes in the GDB source, with any two textual changes separated by a context of at least two unchanged lines. The problem was, of course, that applying any subset of these 8721 changes did not have many chances to result in something consistent.

To summarize: Nearly all tests were unresolved. $ddmin$ had a chance to succeed only after the subsets and their complements had become sufficiently close to \emptyset and C , respectively. As shown in Figure 3, $ddmin$ required 470 tests to isolate the single failure-inducing change. Each test involved change application, smart reconstruction of GDB, and running the test, which took an average time of 190 seconds.²

After grouping changes by location criteria (i.e. common files and directories), $ddmin$ required 289 tests. After applying syntactic criteria (grouping of changes according to functions), $ddmin$ required only 97 tests, or about 4 hours.

²All times were measured on a Linux PC with a 500 MHz Pentium III processor.

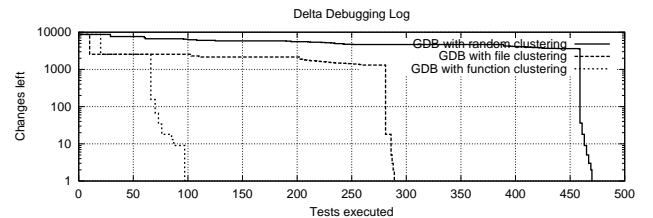


Figure 3: Simplifying failure-inducing GDB changes

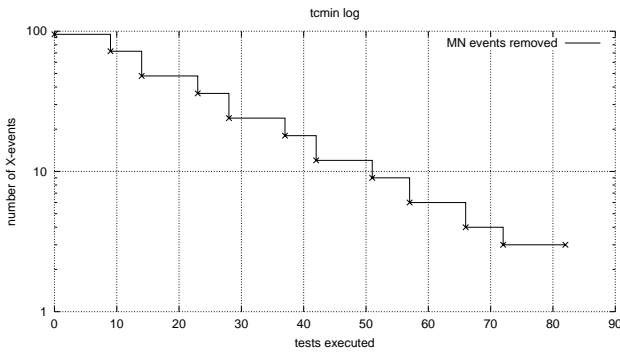


Figure 4: Simplifying failure-inducing Mozilla input

5 Simplifying Test Cases

In a second case study [3], we used delta debugging to simplify *failing test cases*—by minimizing the differences between the failing input and a trivial (empty) input.

Having a minimal test case is an important issue for debugging. Not only does it show the relevant failure circumstances; simplifying test cases also helps to identify and eliminate duplicate bug reports. It turned out that *dadmin* works just as well to minimize failure-inducing input than to minimize failure-inducing changes.

Here comes another case study. *Mozilla*, Netscape’s new web browser project [5], is still in beta status and thus shows numerous bugs. We hunted a bug in which printing a certain WWW page caused the browser to fail. This scenario gave us two things to look for:

- Which parts of the WWW page are relevant in reproducing the failure?
- Which user actions are relevant in reproducing the failure?

Using delta debugging, we could successfully simplify both the sequence of user actions as well as the WWW page. It turned out that a number of user actions (such as changing the printer settings) were not relevant in reproducing the failure. As shown in Figure 4, our prototype minimized 95 user actions into 3 after 39 test runs (or 25 minutes): In order to crash Mozilla, it suffices to press *Alt+P* and press and release the mouse button on *Print*.³ Likewise, in 57 test runs, our prototype reduced the WWW page from 896 lines to a single line:

```
<SELECT NAME="priority" MULTIPLE SIZE=7>
```

Automatic test case minimization was also applied to the GNU C compiler as well as various UNIX utilities, all with promising results. All one needs is an input, a sequence of user actions, an observable failure—and a little time to let the computer do the minimization.

³Note that releasing the *P* key is not required.

6 Reducing Execution Traces

As a further example, we show how delta debugging can be helpful in identifying *failure-inducing events* during the execution of the program. We assume an *execution trace* as a sequence of *program states* during execution, starting with a working (empty) state and ending in a failing state. During a program run, statements are executed that access and alter the program state. The goal is now to identify those events (i.e. statement executions) that were relevant in producing the failure and to eliminate those that were not.

Again, this nicely fits into our scenario model of Section 2. The “working” scenario is no execution at all. The “failing” scenario is the result of the state changes $C = \{\Delta_1, \dots, \Delta_n\}$ induced by the executed statements. Applying delta debugging to minimize C means to isolate exactly those state changes that were relevant in producing the final state—very much like well-known dynamic slicing techniques [1, 4, 7], but relying on *partial execution* rather than analysis of control and data flow and thus showing real causality instead of potential causality.

As an example, consider the following PERL program. It reads in two numbers a and b and computes the sum $sum = \sum_{i=a}^b i$ as well as the product $mul = \prod_{i=a}^b i$:

```
1 $sum = 0;
2 $mul = 1;
3 print "a? "; $a = <>;
4 print "b? "; $b = <>;
5 while ($a ≤ $b) {
6   $sum = $sum + $a;
7   $mul = $mul * $a;
8   $a = $a + 1;
9 }
10 print "sum = ", $sum, "\n";
11 print "mul = ", $mul, "\n";
```

Here is an example run of `sample.pl`:

```
$ perl ./sample.pl
a? 0
b? 5
sum = 15
mul = 0
$ -
```

For this run, we wanted to determine the events which have influenced the final program output. We set up a prototype called STRIPE⁴ which applies the *dadmin* algorithm on execution traces.

In a first step, STRIPE determines the execution trace—that is, the values of the program counter during execution. Then, STRIPE runs the *dadmin* algorithm on the execution trace. To omit a statement S from execution, STRIPE uses a conventional interactive debugger to insert a breakpoint at S . The breakpoint causes the program to interrupt whenever S is reached; an associated breakpoint command causes the debugger to resume execution behind S .

⁴STRIPE = “Systematic Trace Reduction by Iterative Partial Execution”

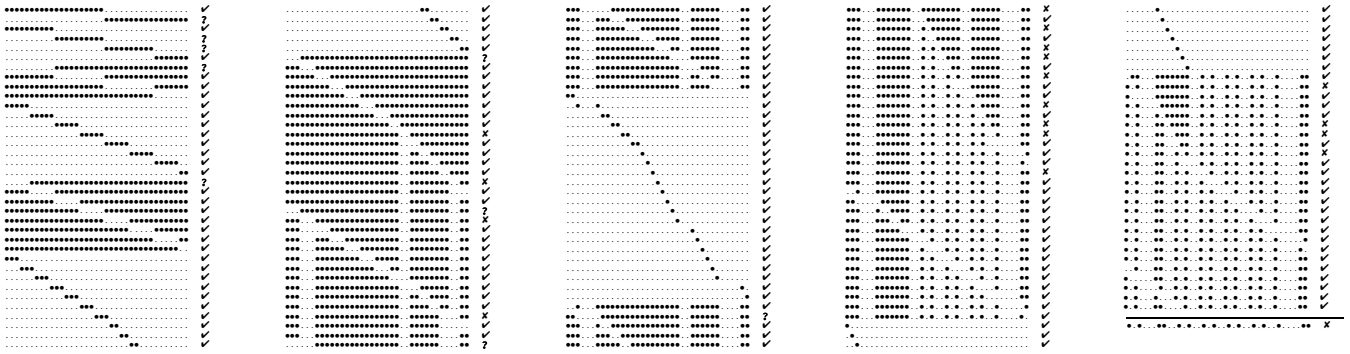


Figure 5: STRIPE run on `sample.pl` with program output as test criterion

Event	Original trace	Reduced trace	... w.r.t. $\$sum$... w.r.t. $\$mul$
1 ₁	$\$sum = 0;$	$\$sum = 0;$	$\$sum = 0;$	
2 ₂	$\$mul = 1;$			
3 ₃	<code>print "a? "; \$a = ();</code>	<code>print "a? "; \$a = ();</code>	<code>print "a? "; \$a = ();</code>	<code>print "a? "; \$a = ();</code>
4 ₄	<code>print "b? "; \$b = ();</code>			
5 ₅	while ($\$a \leq \b) {			
6 ₆	$\$sum = \$sum + \$a;$			
7 ₇	$\$mul = \$mul * \$a;$	$\$mul = \$mul * \$a;$		$\$mul = \$mul * \$a$
8 ₈	$\$a = \$a + 1;$	$\$a = \$a + 1;$	$\$a = \$a + 1;$	
9 ₉	}			
10 ₁₀	while ($\$a \leq \b) {			
6 ₁₁	$\$sum = \$sum + \$a;$	$\$sum = \$sum + \$a;$	$\$sum = \$sum + \$a;$	
7 ₁₂	$\$mul = \$mul * \$a;$			
8 ₁₃	$\$a = \$a + 1;$	$\$a = \$a + 1;$	$\$a = \$a + 1;$	
9 ₁₄	}			
15 ₁₅	while ($\$a \leq \b) {			
16 ₁₆	$\$sum = \$sum + \$a;$	$\$sum = \$sum + \$a;$	$\$sum = \$sum + \$a;$	
17 ₁₇	$\$mul = \$mul * \$a;$			
18 ₁₈	$\$a = \$a + 1;$	$\$a = \$a + 1;$	$\$a = \$a + 1;$	
19 ₁₉	}			
20 ₂₀	while ($\$a \leq \b) {			
21 ₂₁	$\$sum = \$sum + \$a;$	$\$sum = \$sum + \$a;$	$\$sum = \$sum + \$a;$	
22 ₂₂	$\$mul = \$mul * \$a;$			
23 ₂₃	$\$a = \$a + 1;$	$\$a = \$a + 1;$	$\$a = \$a + 1;$	
24 ₂₄	}			
25 ₂₅	while ($\$a \leq \b) {			
26 ₂₆	$\$sum = \$sum + \$a;$	$\$sum = \$sum + \$a;$	$\$sum = \$sum + \$a;$	
27 ₂₇	$\$mul = \$mul * \$a;$			
28 ₂₈	$\$a = \$a + 1;$	$\$a = \$a + 1;$	$\$a = \$a + 1;$	
29 ₂₉	}			
30 ₃₀	while ($\$a \leq \b) {			
31 ₃₁	$\$sum = \$sum + \$a;$	$\$sum = \$sum + \$a;$	$\$sum = \$sum + \$a;$	
32 ₃₂	$\$mul = \$mul * \$a;$			
33 ₃₃	$\$a = \$a + 1;$			
34 ₃₄	}			
35 ₃₅	while ($\$a \leq \b) {			
36 ₁₀₃₆	<code>print "sum = ", $\\$sum$, "\n";</code>	<code>print "sum = ", $\\$sum$, "\n";</code>	<code>print "sum = ", $\\$sum$, "\n";</code>	
37 ₁₁₃₇	<code>print "mul = ", $\\$mul$, "\n";</code>	<code>print "mul = ", $\\$mul$, "\n";</code>		<code>print "mul = ", $\\$mul$, "\n";</code>

Figure 6: Execution traces of `sample.pl`

In our example, the *test* function would return

- ✘ whenever the expected behavior was actually reproduced,
- ✔ when anything else was produced, and
- ? if the program did not produce anything, i.e. the program crashed or hung.

Figure 5 on the preceding page shows the STRIPE run. STRIPE required a total of 176 tests (i.e. partial executions) to reduce the execution trace. Each bullet • stands for an applied Δ_i , or executed statement. We see that most tests either result in ✔ or ?. Only after 50 tests do we see a ✘ test outcome—the expected output was produced and the trace can be reduced to the shown set of statements. The last line shows the final 1-minimal trace. The actual result is shown in Figure 6. All events are listed in the form *line number*_{time}.

- The original trace is shown in the first column; we see how the variables *\$a* and *\$b* are read in (events 3₃ and 4₄ in Figure 5) and how *\$sum* and *\$mul* are computed and printed.
- The reduced trace in the second column shows that several of these statements are actually irrelevant for computing the output—executing the original trace and the reduced trace has the same effect. Thus, we find that the initialization of *\$mul* (1₁) is irrelevant, since whatever it is initialized to, *\$mul* will always be zero.⁵
- In the third column, we have run STRIPE with a *test* function that compares only the *\$sum* output. This means that the few statements related to *\$mul* can be eliminated as well.
- In the fourth and last column, we have run STRIPE with a *test* function that compares only the *\$mul* output. Only the initial assignment and the final output remain.

We see how STRIPE effectively reduces execution traces to those events which were actually relevant (or *critical*) in reaching the final state. We thus call this sequence of events a *critical slice*—similar to the critical slices as explored by DeMillo, Pan and Spafford [2], but guaranteeing 1-minimality and with a much better best-case efficiency. Furthermore, delta debugging does not require any program analysis (in the case of PERL programs such as *sample.pl*, this is an almost impossible issue, anyway).

7 Conclusion and Future Work

Delta debugging automates the most time-consuming debugging issue: determining the *relevant problem circumstances*. Relevant circumstances include the program input, changes

⁵Likewise, all later assignments to *\$mul* (7₁₂, 7₁₇, 7₂₂, ...) do not change its value. Finally, all control statements are irrelevant, since the control flow is explicitly stated in the execution trace.

to the program code, or executed statements. All that is required is an automated test.

Delta debugging comes at a price: Although the *dadmin* algorithm guarantees 1-minimality, the worst-case quadratic complexity is a severe penalty for real-world programs—especially considering program runs with billions of executed statements. Consequently, our future work will concentrate on introducing *domain knowledge* into delta debugging. In the domain of code changes, we have seen significant improvements by grouping changes according to files, functions, or static program slices, and rejecting infeasible configurations; we expect similar improvements for program input and program statements.

Our long-term vision is that, to debug a program, all one has to do is to set up an appropriate *test* function. Then, one can let the computer do the debugging, isolating failure circumstances using a combination of program analysis and automated testing. Automatic isolation of failure causes is no longer beyond the state of the art. It is just a question of how much computing power and program analysis you are willing to spend on it.

Acknowledgements. Ralf Hildebrandt, Kerstin Reese, and Gregor Snelling provided valuable comments on earlier revisions of this paper.

Further information on delta debugging is available at

<http://www.fmi.uni-passau.de/st/dd/>.

References

- [1] H. Agrawal and J. R. Horgan. Dynamic program slicing. In *Proceedings of the ACM SIGPLAN 1990 Conference on Programming Language Design and Implementation (PLDI)*, volume 25(6) of *ACM SIGPLAN Notices*, pages 246–256, White Plains, New York, June 1990.
- [2] R. A. DeMillo, H. Pan, and E. H. Spafford. Critical slicing for software fault localization. In S. J. Zeil, editor, *Proc. ISSTA 1996 – International Symposium on Software Testing and Analysis*, volume 21(3) of *ACM Software Engineering Notes*, pages 121–134, San Diego, California, USA, Jan. 1996.
- [3] R. Hildebrandt and A. Zeller. Simplifying failure-inducing input. In *Proc. ISSTA 2000 – International Symposium on Software Testing and Analysis*, Portland, Oregon, Aug. 2000.
- [4] B. Korel and J. Laski. Dynamic slicing of computer programs. *The Journal of Systems and Software*, 13(3):187–195, Nov. 1990.
- [5] Mozilla web site. <http://www.mozilla.org/>.
- [6] G. J. Myers. *The Art of Software Testing*. John Wiley & Sons, Inc., New York, 1979.
- [7] F. Tip. A survey of program slicing techniques. *Journal of Programming Languages*, 3(3):121–189, Sept. 1995.
- [8] A. Zeller. Yesterday, my program worked. Today, it does not. Why? In O. Nierstrasz and M. Lemoine, editors, *Proc. ESEC/FSE’99 – 7th European Software Engineering Conference / 7th ACM SIGSOFT Symposium on the Foundations of Software Engineering*, volume 1687 of *Lecture Notes in Computer Science*, pages 253–267, Toulouse, France, Sept. 1999. Springer-Verlag.