Finding Missed Compiler Optimizations by Differential Testing (preprint, v1.2)

Gergö Barany
Inria Paris
France
gergo.barany@inria.fr

Abstract

Randomized differential testing of compilers has had great success in finding compiler crashes and silent miscompilations. In this paper we investigate whether we can use similar techniques to improve the quality of the generated code: Can we compare the code generated by different compilers to find optimizations performed by one but missed by another?

We have developed a set of tools for running such tests. We compile C code generated by standard random program generators and use a custom binary analysis tool to compare the output programs. Depending on the optimization of interest, the tool can be configured to compare features such as the number of total instructions, multiply or divide instructions, function calls, stack accesses, and more. A standard test case reduction tool produces minimal examples once an interesting difference has been found.

We have used our tools to compare the code generated by GCC, Clang, and CompCert. We have found previously unreported missing arithmetic optimizations in all three compilers, as well as individual cases of unnecessary register spilling, missed opportunities for register coalescing, dead stores, redundant computations, and missing instruction selection patterns.

Keywords optimization, differential testing, randomized testing

1 Motivation

Over the last few years, randomized differential testing has become a standard method for testing programming tools. Hundreds of compiler bugs, both compiler crashes and silent generation of incorrect code, have been found by random testing in many mature compilers for various programming languages [Eide and Regehr 2008; Lidbury et al. 2015; Midtgaard et al. 2017; Yang et al. 2011]. The method can be similarly effective for finding crashes and correctness bugs in static analyzers [Cuoq et al. 2012].

In this approach, randomly generated programs satisfying certain correctness properties are compiled with different compilers, executed, and their results compared. Different results for well-defined programs imply a bug in at least one of the compilers. Such interesting input programs can be reduced by a test case minimizer to often just a few lines of code that provoke the bug.

```
int fn1(int p, int q) {
  return q + (p % 6) / 9;
}
```

Figure 1. Missed optimization example: `(p % 6) / 9` is always zero. Clang used to be unable to eliminate this computation.

Given the success of this method in finding bugs, we wanted to test whether it can also be used to find missed optimizations: Is it possible to build a system that finds minimal examples of missed optimizations by compiling random C programs, comparing the generated code, and minimizing the input program?

The answer is yes. Figure 1 shows a small example of an arithmetic optimization missed by Clang found by our tool (with small manual modifications). The value of the expression `(p % 6)` is always in the interval $[-5, 5]$. Dividing any number in this interval by 9, with truncation toward 0 as prescribed by the C standard, gives a value of 0 for any value of `p`. GCC therefore generates code to simply copy `q` to the return register and return. Until a recent fix motivated by our findings, Clang did not perform this arithmetic simplification and generated code to evaluate the modulo and divide operations.

In this paper we describe the following contributions:

- a differential testing methodology for finding missed optimizations in C compilers;
- a set of configurable tools for automating the method; and
- the experimental evaluation of the method and tools, with examples of previously unreported missed optimizations of various kinds in all three C compilers we tested.

In addition to the examples throughout this paper, an evolving list of previously unreported missed optimizations we found is available online at https://github.com/gergo-/missed-optimizations. We have filed reports

Unpublished preprint, 2017
2017.

or patches against all compilers tested, and several issues have been fixed.

2 Random differential compiler testing

When searching for miscompilations, randomized differential compiler testing proceeds by generating random programs, compiling them with different compilers, executing them, and comparing the results. In order to make these results comparable, the random programs must fulfill certain criteria: They must be complete programs that avoid invoking any undefined behaviors in the target language, as well as (for C and similar languages) implementation-defined features such as the order of evaluation of expressions.

To compare results, Csmith [Yang et al. 2011] generates programs that compute a checksum over the final values of all global variables and prints it to the standard output before exiting; Orange3 [Nagai et al. 2014] precomputes expected results during generation and adds corresponding conditional print statements to the program. In either case, the textual output of the test program compiled with different compilers can easily be compared by a driver script. Any difference indicates a miscompilation bug in at least one of the compilers under test. Other random generators such as CCG [Balestrat 2016] or ldrgen [Barany 2017] do not generate complete programs that are meant to be executed; they can still be useful for finding compiler crashes.

The generators mentioned above all target the C programming language, but similar tools exist for other languages, such as CLsmith for OpenCL [Lidbury et al. 2015], jsfunfuzz for JavaScript [Ruderman 2015], or efftester for OCaml [Midtgaard et al. 2017].

Having found the symptoms of a compiler bug, an important final phase is the reduction of the input program to a minimal example showing the bug. Automatically generated programs typically contain hundreds or thousands of lines of code. Finding the cause of a bug in such a large test case is tedious and often unrealistic. Most bugs also don't need large triggers: In a large-scale study of reported compiler bugs, Sun et al. [2016] found that test cases for compiler bugs 'are typically small, with 80 % having fewer than 45 lines of code'.

To help identify the actual bug, test case reducers simplify the input program by removing parts or by other transformations such as replacing variables by constants. C-Reduce [Regehr et al. 2012] is a generic reducer for C family programs that is run with a user-provided 'interestingness test'. This test is a program that returns 0 if the current version of the program is still 'interesting' (i. e., shows the buggy behavior one is investigating) and some other value otherwise. For finding miscompilations, the test will typically compile the program with the compilers under test and see if the resulting binaries still show different behavior; for finding compiler crashes, the test will check if the compiler still crashes. Starting from an initial program, C-Reduce explores a search space of incrementally reduced programs that are interesting according to the test provided. It stops when it has reached a minimal program that it cannot reduce further.

The Orange3 random generator includes its own specialized reduction mode. Both of these reducers work well in practice. They often produce very small examples, at times with further opportunities for manual cleanup.

```
char fn2(float p) {
  if (isfinite(p) &&
      CHAR_MIN <= p && p <= CHAR_MAX) {
    return (char) p;
  }
  return 0;
}
```

(a) Example input for useless spilling by GCC. Guards added manually for well-defined behavior on all inputs.

```
vcvt.u32.f32 s15, s0   ; float -> unsigned int
sub sp, sp, #8         ; allocate stack frame
vstr.32 s15, [sp, #4]  ; spill float reg
ldrb r0, [sp, #4]      ; reload to int reg
add sp, sp, #8         ; free stack frame
```

(b) Annotated ARM code generated by GCC for the `float` to `char` conversion. Spilling is useless, the last four instructions could be replaced by a single direct copy (`vmov r0, s15`).

Figure 2. Example of unnecessary spill code generation.

3 Searching for missed optimizations

Our goal is to exploit this existing differential testing infrastructure in novel ways to find missed optimizations.

Consider the case of inefficient register allocation, resulting in more spill code than is strictly needed for a given function. The amount of spill code can be computed for different compiled programs and can thus form the basis of an interestingness test. If we can start from a program on which some compiler generates useless spills, a reducer can then produce a corresponding program that is minimal but still shows a difference in spilling between compilers.

Figure 2 shows an example program found by our tool, then manually extended with guards against undefined behavior on overflow. The figure also shows part of the ARM assembly code[1] generated for this program by GCC. The conversion between floating-point and integer types is performed by the `vcvt` instruction, which puts

[1] In all our ARM examples, we generate code for the ARMv7-A instruction set with VFPv3-D16 hardware floating-point extensions.

its result into a floating-point register. This value must be copied to the integer return register `r0`. It can then be zero-extended from 8 bits to implement the truncation to `char`, which is unsigned on the target machine. GCC models the copy and zero extension by spilling an integer word to the stack and reloading a byte, whereas Clang and CompCert simply generate a register-to-register copy instruction. (CompCert then zero-extends. Clang does not, as the guards ensure that the value is already in the correct range.) We have reported this issue; a GCC developer found that the spill is chosen because it is not marked as having higher cost than a zero-extension pattern with a direct copy.

Our hypothesis was that this process can (a) be seeded efficiently with randomly generated programs, (b) be generalized to many kinds of missed optimizations, and (c) reduce programs to reasonably small examples that show clearly useless code that should be optimized away. In other differential testing approaches, compilers are used as 'correctness oracles' for each other. In contrast, we use different compilers' outputs as *optimization oracles*.

Our method thus works as follows:

1. generate a random C program
2. compile the program with different compilers to obtain different binaries
3. if binaries show an 'interesting' difference:
   - reduce to minimal program showing the same 'interesting' difference

We use randomly generated programs instead of open source software or known benchmark suites because random generators provide an unlimited supply of programs with characteristics of our choice: With or without floating-point operations, loops, branches, etc., and with tunable parameters such as the number of variables or maximum size of basic blocks. In particular, we found that loops are problematic in our current approach (see Section 5). Random generators allow us to find interesting missed optimizations in loop-free generated code without having to artificially limit our search to the subset of loop-free functions in existing software.

We also knew from previous experience that the output of a test case reducer often looks very 'artificial', and it is known that the test case reduction process can 'jump' from one compiler bug to another.[2] These observations mean that starting reduction from a real benchmark function doesn't guarantee that the missed optimization in the final, reduced result is actually present in the original code: All we know is that it is interesting according to the same criterion. We therefore opted for the flexibility of randomly generated input programs instead of using a fixed benchmark set.

[2] https://blog.regehr.org/archives/1284

```
int fn3(double c, int *p, int *q) {
  int i = (int) c;
  *p = i;
  *q = i;
  return i;
}
```

(a) Program causing Clang to generate redundant code.

```
vcvt.s32.f64    s2, d0  ; convert double->int
vstr    s2, [r0]        ; store to *p
vcvt.s32.f64    s2, d0  ; convert again
vcvt.s32.f64    s0, d0  ; convert yet again
vmov    r0, s0          ; copy to return reg
vstr    s2, [r1]        ; store to *q
```

(b) Annotated ARM code generated by Clang. The type conversion is performed three times; one time would suffice.

Figure 3. Redundant repeated type conversions introduced by Clang for a single conversion in the source.

For generating programs, we have experimented with the Csmith [Yang et al. 2011] and ldrgen [Barany 2017] tools. By default, Csmith generates a complete application with a `main` function and typically several other functions. We use command line flags to request generation of a single function and suppress generation of `main`. We also disable generation of global variables because we have found that different compilers generate different code for loading the addresses of globals; these inessential differences introduced unnecessary complications when comparing binaries.

Both generators were useful for finding missed optimizations. Csmith covers a larger subset of the C language than ldrgen, allowing us to find some issues that could not be found using ldrgen. On the other hand, Csmith often produces large amounts of dead code, leading to trivial binary code generated even for complex-looking input source code. In contrast, ldrgen tries to avoid generating dead code, and binaries compiled from ldrgen-generated code are typically larger and more complex than those compiled from Csmith-generated code of similar size.

Figure 3 shows one example of a missed optimization found using a Csmith-generated seed program. This function receives pointers as arguments and contains assignments to their target. Clang used to duplicate the type conversion for each use of the variable `i`. We have reported this issue, and a partial fix eliminating the conversions but leaving some unnecessary copies has been added to Clang. This missed optimization could not have been found using only ldrgen because its current version never generates assignments through pointers.

Our system for interestingness tests on binaries is described in the following section. In the spilling example of Figure 2, the interestingness test searched for different numbers of loads through the stack pointer. Other tests consider features such as function calls or certain classes of arithmetic instructions.

For reducing to minimal programs showing missed optimizations, we use C-Reduce. C-Reduce has many different passes that implement different minimizing transformations. We use the default pass set, except for disabling a pass that transforms local variables into global ones. The reason is the same as above: Accesses to globals introduce uninteresting differences in binaries.

The entire process sketched above is tied together by a shell script that simply calls a generator, the compilers, and the reducer. All the power in the process comes from these tools. Reduced programs are checked by a human to see if they are trivial, duplicates, or really interesting missed optimizations.

4 Finding optimization differences

As we use standard random program generators and a standard test case reduction tool, the only optimization-specific part of our toolchain is the interestingness test. This interestingness test compares the code generated by different compilers.

There are several possible ways of comparing code: We might use dynamic approaches measuring execution time [Chen and Regehr 2010] or recording dynamic traces of executed instructions [Moseley et al. 2009a,b]. For this work we decided to try a simpler, static approach based on analyzing the binary without executing it. This is an approach that composes better with test case reduction: As we reduce a program, we would also have to reduce test inputs along with it if we wanted to execute it for dynamic analysis. In contrast, for static analysis we do not have to worry about inputs at all as we never execute the code. In addition, timing the program under test is inherently non-deterministic, and any small perturbations could confuse the reducer.

For these reasons we decided to try a simple, static approach. We implemented it in a tool we tentatively call `optdiff`. Our goal was to build a tool that could be reconfigured for different kinds of missed optimization tests, in particular for identifying different amounts of spill code. As spill code in loops is more expensive than outside, we wanted to be able to estimate execution frequencies of instructions. Finally, we wanted the tool to be retargetable to at least the ARM and x86-64 architectures.

We use the angr binary analysis framework[3] [Shoshitaishvili et al. 2016] as the basis for our tool. This framework collects various tools useful for cross-platform binary analysis, such as an executable loader, a library of architecture descriptions, and predefined program analyses. Its API is presented as a Python library.

Our `optdiff` tool is a Python program that uses angr's binary loader to load code from the binaries to be compared. We then use the algorithm of Wei et al. [2007] to construct a loop nesting tree for the input program and use that as a basis for basic block frequency estimation. We estimate frequencies by assigning the function's entry block the frequency 1, then propagating frequencies such that (a) every block's frequency is distributed equally to its successors at the same loop nesting level, and (b) loop entry blocks get assigned a frequency of 8 times their predecessors outside the loop (i. e., we assume that every loop iterates 8 times).

`optdiff` then uses functions we call *checkers* to compute a *score* for each binary. The score is a number expressing the prevalence of the feature of interest (e. g., the amount of spill code) in the input program.

A checker is a function that computes a local score for each instruction in the input program. The total score for the program is the sum of the individual instruction scores weighted by the estimated block frequencies. Thus for a checker c and an input function f consisting of blocks b with frequencies (weights) w_b containing instructions i, the total score s is given by:

$$s = \sum_{b \in f} w_b \cdot \sum_{i \in b} c(i)$$

Checkers are small Python functions annotated with the `@checker` decorator. They return a number or a boolean with `True` treated as 1, `False` treated as 0. Figure 4 shows the source code of two example checkers predefined in `optdiff`: the trivial checker for counting instructions, and the ARM-specific part of the checker for the number of memory loads. As ARM has load instructions for doublewords and a flexible load-multiple instruction, this checker counts the number of registers written by the instruction.

Note that checkers are purely local: they look at each instruction in isolation, without information about the surrounding code, or any way of propagating information to other instructions. This is for simplicity, although it makes some interesting things impossible. In particular, this framework cannot be used for cycle-accurate performance estimation, which would necessitate propagation of information about the states of pipelines and caches between instructions.

[3] http://angr.io/

```
@checker
def instructions(arch, instr):
    """Number of instructions."""
    return 1
```

(a) The trivial checker for counting instructions.

```
@checker
def loads(arch, instr):
    """Number of memory loads."""
    op = instr.insn.mnemonic
    if is_arm(arch):
        if op == 'ldrd':
            return 2
        elif re.match('ldm.*', op):
            return len(instr.insn.operands)-1
        return bool(re.match('v?ldr.*', op)
    ... # other architectures
```

(b) The checker for counting memory accesses (excerpt).

Figure 4. Two example checkers from `optdiff`.

The `optdiff` tool provides command line flags for customizing its operation. It exits with a status indicating 'interesting' if the scores for the input binaries differ. It also provides a flag for a minimum absolute difference to be considered interesting. In other cases, we are not just interested in whether the scores s_1 and s_2 for the two binaries differ, but want only to consider cases where a certain compiler is 'better' than another one. For this, `optdiff` provides flags specifying that only the case $s_1 < s_2$, or only the case $s_1 > s_2$, should be considered interesting.

At the time of writing, `optdiff` comprises 573 physical lines of Python code, of which 246 lines compute the loop nesting tree and estimate basic block frequencies, 171 lines are checker definitions, and the remaining 156 lines parse command line arguments, load binaries, call the other components to compute scores, and exit with an appropriate result. There are currently 14 checkers, some of which are architecture-specific or not yet implemented for all architectures. We have checkers for the total number of instructions; memory loads and stores from the stack or from any address; register copies; additive or multiplicative integer arithmetic operations; additive, multiplicative, or arbitrary floating-point arithmetic operations; x86-64 packed (SIMD) instructions; and function calls.

The x86-64 packed instructions are an interesting case of a checker where a lower score does not necessarily indicate a 'better' program. Indeed, a higher number of such instructions may mean that one compiler succeeded in vectorizing a loop but another didn't. We have not found such cases so far, but the search did result in an interesting issue shown in Figure 5. This function always returns either 0 if `N` ≤ 0, or `p1[N-1]` otherwise. Thus the loop can be replaced by a simple branch, as only the last iteration matters. Clang is able to perform this optimization, but GCC is not. Instead, on x86-64, it generates a complex unrolled loop.

```
int N;
double fn5(double *p1) {
  int i = 0;
  double v = 0;
  while (i < N) {
    v = p1[i];
    i++;
  }
  return v;
}
```

Figure 5. Example loop extensively optimized by GCC on x86-64. Clang simply replaces the loop by a branch.

5 Dealing with undefined behavior

One particular feature of our system is the absence of checks for undefined behavior in the programs we generate or reduce. Such checks are standard in other randomized differential approaches that look for miscompilations [McKeeman 1998; Nagai et al. 2014; Yang et al. 2011] since undefined programs may be compiled differently by different compilers, without this being an indication of a miscompilation. For this reason, random generators like Csmith or Orange3 exclude certain undefined behaviors at generation time (for example, Csmith uses flow-sensitive pointer analysis) or generate code to guard against undefined behavior at runtime by using a library of safe wrapper functions for arithmetic operations. For our search for missed optimizations, hiding most arithmetic operations behind a function call would be inconvenient: It would prevent almost all arithmetic simplifications that we may want to test for.

As another difference to our approach, differential bug-finding tools detect differences in compilation by executing the generated programs. For this, they include concrete inputs for their code, typically as initialized global variables. In contrast, we never execute our programs, only compare the generated binary code. We therefore do not need input data. Inputs are function parameters with unknown values provided by a hypothetical caller. Some values for these parameters may lead to undefined behavior. For example, in Figure 5, the pointer `p1` may be `NULL` or otherwise invalid; in Figure 3, certain values of the `double` parameter may overflow the `int` it is converted to.

This is normal for C programs; compilers expect programmers to call such functions with valid arguments and typically do not give guarantees if such implicit preconditions are violated. On the other hand, they do not go out of their way to pessimize the cases where there *are* such violations; they just generate code as if such cases never occurred. Therefore, pragmatically, we expected no problems in practice with comparing programs in which certain input values may cause undefined behavior.

All in all, our experiments confirmed that this was usually the case for undefined behavior during expression evaluation that was dependent on input values. On the other hand, we did find cases where the undefined behavior was unconditional and independent of input data. As a representative example, we did sometimes observe reduced functions such as the following:

```
int fn(int a) {
  int x = 0;
  int b = a / x;
  return b;
}
```

Neither GCC nor Clang warn about the unconditional division by zero in this case, although Clang recognizes it and 'optimizes' it to a function that returns immediately. GCC does not do this, thus the generated binaries differ, and this program is considered 'interesting' by some of our checkers. However, as the function does not admit any well-defined execution, we do not consider this an interesting optimization.

We have experimented with including a static analyzer in the interestingness test to exclude the cases where every execution path leads to undefined behavior. We ran the Eva abstract interpreter [Blazy et al. 2017], part of the Frama-C program analysis platform [Kirchner et al. 2015]. This analyzer is effective at flagging unconditionally undefined cases. However, we found that this only pushes the reduction process in the direction of slightly more obscure versions of essentially the same undefined operations.

Avoiding such uninteresting, partially undefined cases in practice is thus an open problem. For now, we use the following workarounds:

- These cases are relatively frequent when counting all instructions, but rarer with more specialized checkers. We accept that they occur from time to time, as reduced cases must be manually inspected and classified anyway.
- Of the many undefined operations on C, in practice we mostly had problems with division and pointer dereferences. We can configure our program generators to avoid such operations.

As Figure 1 shows, we are still able to find some interesting missed optimizations involving divisions.

Loops are a more problematic case. With our instruction count checker, C-Reduce likes to reduce functions containing loops to nonsensical code like the following:

```
void fn(char p3, int p5) {
  double v;
  while (p3)
    v = !p5;
  (int)v & 1;
}
```

Entering an infinite loop without externally visible side effects (such as I/O) is undefined in C, and different compilers 'optimize' this program differently. This, too, sends our reduction process down uninteresting paths. We have briefly experimented with the CPAchecker platform [Beyer and Keremoglu 2011] in the hope that its termination analysis would catch such cases of functions that do not terminate for all inputs. Unfortunately, we found that it returned 'unknown' in too many cases to be practical for our purposes.

We therefore leave the case of loops as an open problem for future work. Despite the single interesting find of the example in Figure 5, the false positive rate due to nonterminating functions was too high for meaningful use. In most of our experiments we have therefore disabled the generation of loops.

6 Experimental evaluation

We have evaluated our method of finding missed optimizations in a series of unstructured experiments performed occasionally over the course of about five months, incrementally testing and evolving our `optdiff` tool and our methods in response to our observations. We used up-to-date development versions of all compilers at optimization level `-O3` and with `-fomit-frame-pointer` (where supported) to make spill code easier to identify.

6.1 Development of the project

We used the Csmith and ldrgen program generators in our experiments. The two generators are complementary: Csmith's larger supported fragment of the C language allowed us to find the issues in Figures 3 and 6, while arithmetic optimizations are more quickly found using ldrgen. The library of `optdiff` checkers began with a simple checker for counting instructions and a very simple load checker that gradually evolved into the more sophisticated form shown in Figure 4. Other checkers were added as we observed patterns in generated programs and wanted to focus on more specific features.

As a side-effect of various issues encountered over the course of the project, we have submitted patches or bug reports to the Csmith, angr, and ldrgen developers.

```
struct S0 {
  int f0;
  int f1;
  int f2;
  int f3;
};

int fn6(struct S0 p) {
  return p.f0;
}
```

Figure 6. The function returns the structure's first field, which on ARM is passed in register `r0`. This is also the return register; the function could return immediately. GCC first generates useless code to spill the structure to the stack, then reload the first field.

The missed optimizations we collected were thus found with a mix of various versions of different tools, reflecting the exploratory nature of this work. We varied the command line flags of generators and checker definitions to search for issues of various kinds.

For most of the project, we focused on comparing GCC and Clang, mostly generating code for ARM due to familiarity with its assembly language. While we are particularly interested in register allocation and spilling, it turned out that searching for differences in spill reloads often finds issues that are not directly related to register allocation. In particular, we found several arithmetic simplification issues while trying to find differences in register allocation. A common pattern was a function with more arguments than fit in argument registers, having to receive some arguments on the stack. If a computation on one of these stack arguments is redundant and is optimized away by one compiler but not by the other, this will show up in the compiled code as a difference in the number of reads from the stack. As this is a roundabout way of searching for arithmetic optimization opportunities, we implemented more direct checkers. In contrast, when searching for actual differences in register allocation, we typically tell the random code generator to limit the number of arguments according to the target platform's argument registers. Even so, we keep finding more spills due to differences in arithmetic optimizations (as less optimized code uses more registers) than due to real differences between register allocators.

6.2 Results

In the early stages of the project, we reported some issues we found interesting to the compilers' bug trackers. As these are not correctness bugs, they are treated with low priority and mostly remained unfixed. We therefore stopped submitting them to avoid spamming bug trackers and instead decided to aggregate all in a public list. After we published this preliminary list online, there was some renewed interest by LLVM developers, and some previously reported issues were fixed.

It is difficult to count missed optimizations as we cannot be sure which seemingly related issues are really caused by the same piece of code. However, to the best of our knowledge, we have identified the following numbers of previously unreported, distinct missed optimization issues that we would consider worth changing and assume to be reasonably simple.

GCC We have identified 19 issues in GCC, for which we have filed five reports so far, including the examples in Figures 2, 6, and 7. One of our reports contained two issues that we now suspect to have separate underlying causes and that we now count separately. All of our reports were confirmed by developers, and one issue has been fixed in GCC's development version.

Clang We have identified six issues in Clang/LLVM, of which two were reported and fixed by the developers (Figures 1 and 3), and a third one was fixed by a patch we submitted.

CompCert We have not tested CompCert extensively but have identified nine separate missed optimizations. One was fixed by the CompCert developers, two were fixed by patches we submitted.

6.3 Classification of the issues found

The issues we identified fall into several categories. Many are peephole optimizations of arithmetic or bitwise operations, such as the examples in Figures 1 and 7. As with most peephole optimizations, these may appear trivial in isolation and unlikely to appear in hand-written code. However, they may be enabled by other optimizations, notably by function inlining followed by constant propagation; much of the power of inlining comes from the fact that it exposes opportunities for specializing the inlined code.

Other issues concern register allocation and spilling, as in Figure 2. We also found similar code (not shown here for lack of space) causing GCC to spill although enough registers are available, and cases where it generates dead spill code (i.e., it stores values to the stack but never reloads them).

The example in Figure 6 also contains dead stack stores generated by GCC, but it is due to particular handling of argument registers, not due to general computations. We found a similar (previously known) case where Clang also generates dead stack stores into structures passed by value.

Some other issues we found (mostly in GCC) include bad instruction scheduling of load-immediate instructions, increasing register pressure and leading to unnecessary spills; missed tail-call optimization for compiler intrinsics; and the much too aggressive loop optimization of Figure 5.[4]

6.4 Causes of the issues found

We have also tried to identify the underlying reasons for the missed optimizations we found. Our analysis is based on the patches we developed ourselves, the patches developed by compiler maintainers, code inspection, and comments by maintainers on the issues we reported. We have not been able to classify all issues.

6.4.1 Forgotten/faulty rules

Pattern matching rules for instruction selection, arithmetic simplifications, and in various program analyses are important parts of every optimizing compiler. With large numbers of rules and possible interactions between them, forgetting rules or getting the associated costs and priorities wrong is an unsurprising source of mistakes.

Of the examples in this paper, the ones in Figures 1, 2, and 3 fall into this category. The first one is part of a group that were considered missing rules in LLVM's instruction simplifier by LLVM developers. The second is confirmed by a GCC developer to be due to a missing cost annotation in a match rule. The third was due to redundant instruction selection patterns that, in the words of an LLVM developer, could 'confuse the cost logic' in the selector, leading to code duplication. Another LLVM example we have found and submitted a patch for was a missing rule for selecting the ARM `vnmla` 'floating-point multiply accumulate with negation' instruction for expressions of the form `-(a * b) - c`; the selector only had a pattern for the symmetric case of `-c - (a * b)`.

Five of the issues we identified in CompCert also fall into this category: a missing constant-folding rule for the modulo operator; a missing instruction selection rule for ARM's `movw` move-immediate instruction; missing constant folding rules for the ARM `mla` integer multiply-add instruction; and missing reassociation rules for multiplications by constants of the form `c1 * x * c2` to `x * (c1 * c2)`. Our patches for the `movw` and `mla` issues have been merged.

6.4.2 Phase ordering

Phase ordering issues are a well-known and much researched problem in compiler construction. Some of the problems we found are due to such ordering problems. One example is shown in Figure 7. The branch condition

```
int fn7_1(int p1) {
  int a = 6;
  int b = (p1 / 12 == a);
  return b;
}
```

(a) The division and comparison could be optimized to a subtraction and a comparison, but GCC failed to do this.

```
int fn7_2(int p1) {
  int b = (p1 / 12 == 6);
  return b;
}
```

(b) GCC was able to optimize this equivalent variant.

Figure 7. Example of a confirmed phase-ordering problem in GCC. The rule for simplifying the division and comparison used to be run before constant propagation.

in either case is equivalent to $72 \leq \texttt{p1} < 84$ and can be optimized to a subtraction and two signed comparisons (or a single unsigned comparison) instead of the expensive division. GCC managed to do this in the second case, but not the first one. We reported this issue as a baffling case of seemingly failed constant propagation. It turned out that the matching rule for this optimization was only run before constant propagation, but not after. In response to our report, GCC developers have moved this rule from the early matching phase to a later one.

A trivial phase ordering case we found in CompCert concerns the conditions of useless branching statements like `if (c) {} else {}`. The corresponding useless jump instructions are cleaned up by one of the backend passes, but this pass only runs after dead code elimination. The code evaluating the condition `c` was therefore left in the program. This has been fixed based on our report.

6.4.3 Unimplemented optimizations

Some optimizations may be missing from compilers simply because developers have not thought of implementing them, or have lacked the resources to do so. We collect some such cases because we find them interesting as opportunities to learn from other compilers' behavior. Clearly, these should not be considered mistakes of the same kind as forgotten cases or faultily implemented optimizations such as those discussed above.

We found two such cases, concerning optimization of floating-point expressions in Clang but not GCC: Clang can promote some floating-point computations to integers where the result is known to be an exact integer, such as in `i = i * 10.0` where `i` is a 32-bit `int`. Similarly, it can simplify `x + 0`, where the `double` variable `x` was initialized from an `int` and thus cannot be a negative zero, which would prohibit simplification to `x`.

[4] See these examples and more at https://github.com/gergo-/missed-optimizations.

6.4.4 Other/uncategorized issues

We have not been able to conclusively categorize some other issues, such as those of Figures 5 and 6. We suspect that all of these cases fall into the general category of faulty rules, either triggering unwanted code transformations or failing to trigger a presumably existing beneficial transformation.

Another possible source of missed optimizations is simply bad luck due to the fact that many computationally hard problems, in particular related to register allocation, must be solved using heuristics. We do not have any examples that we can definitely blame on weak or unlucky heuristics, although Figure 8 may be an example.

In this code, a register is spilled by CompCert but not by other compilers. Investigating the program, we were surprised to find that the statement `v = 4;` is dead code (its result is never used), and it does not correspond to any instructions in the machine code (the constant 4 never appears), yet it was not removed by C-Reduce. We found that the dead statement is necessary for the spill code interestingness test, and removing it manually makes the spill of register `r4` go away. This is strange as dead code should never influence register allocation at all. CompCert's developers also expressed puzzlement at this case: While CompCert in general is not expected to be as powerful as other compilers, its register allocator based on iterated register coalescing [George and Appel 1996] is meant to be competitive. In addition to the spill, some `mov` instructions should be coalesced, but CompCert's heuristics miss this. (Finally, the `mla` instruction computes $0 \cdot \texttt{r0} + 0$, but CompCert missed the opportunity to constant fold. We submitted a patch for this latter issue, which was accepted.)

6.5 Case study: finding regressions

The compilers compared by our method do not have to be completely different, they can also be different versions or different optimization levels of the same compiler. Such a comparison can be useful for finding regressions in optimizations [Iwatsuji et al. 2016].

To see whether our system was able to do this as well, we tested two development versions of GCC just before and after the fix for the issue in Figure 7. Searching for cases where the new version produced more instructions than the old, in 8 hours we generated 8097 programs, of which 16 matched the interestingness test and were reduced to minimal examples. (Generating and testing programs is fast; the bulk of the time is in reduction, which commonly takes on the order of 15 minutes to an hour.) We found a case that could be considered a regression: In code like

```
a = 4;
if (x / a) ...
```

```
int fn8(int p1) {
  int a, b, c, d, e, v, f;
  a = 0;
  b = c = 0;
  d = e = p1;
  v = 4;
  f = e * d | a * p1 + b;
  return f;
}
```

(a) Input source code containing a dead statement `v = 4;`.

```
  str r4, [sp, #8]      ; spill
  mov r4, #0
  mov r12, #0
  mov r1, r0            ; could be coalesced
  mov r2, r1            ; could be coalesced
  mul r3, r2, r1
  mla r2, r4, r0, r12   ; compute 0 * r0 + 0
  orr r0, r3, r2        ; compute r3 | 0
  ldr r4, [sp, #8]      ; reload
```

(b) The core of the code generated by CompCert. The spill of `r4` is caused by dead code, and coalescing and constant folding are missed.

Figure 8. Example function exposing several missed optimizations in CompCert.

where the divisor is a power of 2, the division and comparison against 0 were previously implemented using a single shift that updated the condition code register. After the change, this is no longer treated specially but compiled to two instructions (an add and a compare).

6.6 Example of practical use

To illustrate how compiler developers might use our tools, we ran a representative experiment, configuring `optdiff` to look for differences of at least 10 instructions in ARM code generated by GCC and LLVM. Over an 8-hour run on a laptop with an Intel Core i7 CPU at 2.60 GHz, running C-Reduce with up to 4 threads, we generated 38 programs. Of these 34 were found interesting by `optdiff` and reduced (average reduction time: 14 minutes). The reduced programs have on the order of 5 to 15 lines of C code and compile to about 2 to 30 instructions.

A majority of the reduced programs display duplicates of issues we know: in 10 cases, the lack of general value range analysis in LLVM, causing problems like the one in Figure 1, and 8 cases where GCC prefers to emit sequences of adds and shifts for multiplications by constants instead of a single multiply instruction. A minority are 'false positives' where the difference between the generated code isn't actually interesting. Some of these (7 cases) are due to a current limitation our basic

block frequency estimator, which is misled by functions containing predicated return instructions. Some others (7 cases) are simplifications of very specific forms of expressions or exploitation of signed integer overflow that we assume compiler developers would not find interesting.

Finally, among the 34 reduced cases, we identified two new issues that we consider missed optimizations of interest in GCC: One case where its value range analysis appears to fail, and one case of missed reassociation and canceling in integer computations of a form similar to $(a + x) - (b + x)$ that should be simplified to $a - b$.

With some experience, inspection of the reduced programs is very quick: We spend less than a minute per uninteresting case. Overall, compiler developers looking for missed optimizations could run the generation/reduction process overnight and quickly sort through the results in the morning.

7 Related work

To our knowledge, there has not been much work on direct, (semi-)automatic comparison of the quality of the code generated by different compilers.

7.1 Assembly-level missed optimization search

We are aware of a single tool that uses a similar approach to ours for finding missed optimizations, presented in a short paper by Iwatsuji et al. [2016]. This system generates random programs using the Orange3 generator, compiles them with different compilers, and counts instructions in the generated assembly code. If the generated codes are sufficiently different, Orange3's built-in test case reducer produces a minimal example. The authors compared GCC and Clang as well as two different versions of GCC against each other; they found differences between the compilers as well as regressions in the newer version of GCC. Several issues were reported by the authors and fixed by the compilers' developers.

While this system is broadly similar to ours, its design choices are different in almost all respects: Generated programs are straight-line code without branches. Binaries are compared by pure (static) instruction count, not by more specific features such as spill reloads or certain classes of arithmetic as in our checkers. This limits the tool's usefulness for finding certain kinds of missed optimizations. Its model of program scores is also more complex than ours, but it is not clear to us whether this complexity has benefits.

Another difference to our approach is that the Orange3 system always produces well-defined, complete programs that can be run with inputs provided in global variables. In contrast, we focus on individual functions without predefined inputs. As we discussed in Section 5, undefined behavior is a double-edged sword, but admitting certain cases of potentially undefined functions may allow our tools to find more missed optimizations.

7.2 AST-based missed optimization search

A different system related to the one discussed above was presented by Hashimoto and Ishiura [2016]. It also uses the Orange3 program generation system and its reducer, but finds missed optimizations with an interesting twist: It generates random unoptimized C programs, then optimizes these programs itself on the abstract syntax tree (AST) level. The unoptimized and optimized ASTs are both unparsed to C source files, compiled, and the generated assembly codes compared by counting instructions. One interesting aspect of this work is that one does not need to compare two different C compilers; program generation system itself serves as the optimization oracle. The interestingness test is a refinement of the model of Iwatsuji et al. [2016]. Although the AST-level optimization was restricted to constant propagation and folding, the authors found several missed optimizations in both GCC and Clang.

7.3 Superoptimization

Superoptimization is a technique originally introduced by Massalin [1987] for finding the smallest code sequence equivalent to a given piece of input code. In the original formulation, a superoptimizer enumerates all assembly code sequences up to a given length and checks them for equivalence with the target function. Exhaustive enumeration is primitive, while the equivalence check must be engineered to be as efficient as possible. Modern superoptimizers can be stochastic, symbolic (based on SAT or SMT solvers), or combine several of these techniques [Buchwald 2015; Phothilimthana et al. 2016].

Superoptimizers are interesting in the context of this work because they reveal possibly useful previously missed optimizations in the output of compilers, although using a very different strategy from ours.

7.4 Dynamic missed optimization search

Dynamic analysis is another approach for finding missed optimizations. In the work of Chen and Regehr [2010], benchmark functions from open source applications were instrumented to log input values, then extracted from the application, compiled with different compilers, and ran in isolation on realistic inputs. These runs were carefully timed and used to compare optimizations across compilers cycle-accurately. The authors were able to identify very detailed architecture-specific differences in optimizers, such as Clang generating an instruction with a 16-bit immediate operand where apparently 8- and 32-bit immediates are to be preferred on x86-64. Several

of the issues found in this work were quickly fixed in Clang and GCC.

Chainsaw [Moseley et al. 2009a] and OptiScope [Moseley et al. 2009b] are two tools that record dynamic execution traces of programs compiled with different compilers, then correlate those traces to find differences in optimization. They found missed optimizations in compilers including LLVM and GCC, but we do not know if the issues found were communicated to compiler developers and subsequently fixed.

8 Conclusions

We have presented a methodology and supporting tools for finding missed optimizations in C compilers. Using standard tools where possible, we generate random C functions, compile them using different compilers, compare the generated binaries statically, and reduce any interesting cases to a minimal example showing the same interesting property.

The interestingness of programs is determined using a new, customizable tool that compares binaries according to various criteria that may be relevant to performance. Examples of currently implemented criteria are the number of instructions, the number of arithmetic instructions of certain types, types of memory accesses such as loads from the stack indicating register spilling, or function calls. Our binary checker framework is easily extensible with new criteria for missed optimizations.

We have found missed optimizations in all three C compilers we tested: GCC, Clang, and CompCert. Several of the issues we reported to bug trackers have been fixed by developers or by patches submitted by us. The missed optimizations we found fall into different categories such as unnecessary register spilling, missed arithmetic optimizations, redundant computations, and missing instruction selection patterns.

In the future we plan to investigate further the treatment of possibly nonterminating programs, interprocedural optimizations, new optimization checkers, and combinations of checkers.

References

Antoine Balestrat. 2016. CCG: A random C code generator. Source code repository. (2016). https://github.com/Mrktn/ccg

Gergö Barany. 2017. Liveness-Driven Random Program Generation. In *Pre-proceedings of LOPSTR 2017*. https://arxiv.org/abs/1709.04421

Dirk Beyer and M. Erkan Keremoglu. 2011. CPAchecker: A Tool for Configurable Software Verification. In *Computer Aided Verification - 23rd International Conference, CAV*. 184–190. https://doi.org/10.1007/978-3-642-22110-1_16

Sandrine Blazy, David Bühler, and Boris Yakobowski. 2017. Structuring Abstract Interpreters Through State and Value Abstractions. In *Verification, Model Checking, and Abstract Interpretation (VMCAI)*. 112–130. https://doi.org/10.1007/978-3-319-52234-0_7

Sebastian Buchwald. 2015. Optgen: A Generator for Local Optimizations. In *Compiler Construction: 24th International Conference, CC 2015*. 171–189. https://doi.org/10.1007/978-3-662-46663-6_9

Yang Chen and John Regehr. 2010. Comparing Compiler Optimizations. Blog post. (2010). https://blog.regehr.org/archives/320

Pascal Cuoq, Benjamin Monate, Anne Pacalet, Virgile Prevosto, John Regehr, Boris Yakobowski, and Xuejun Yang. 2012. Testing static analyzers with randomly generated programs. In *Proceedings of the 4th NASA Formal Methods Symposium (NFM 2012)*. http://www.cs.utah.edu/~regehr/papers/nfm12.pdf

Eric Eide and John Regehr. 2008. Volatiles Are Miscompiled, and What to Do About It. In *Proceedings of the 8th ACM International Conference on Embedded Software (EMSOFT '08)*. 255–264. https://doi.org/10.1145/1450058.1450093

Lal George and Andrew W. Appel. 1996. Iterated Register Coalescing. In *Proceedings of POPL '96*. 208–218. https://doi.org/10.1145/237721.237777

Atsushi Hashimoto and Nagisa Ishiura. 2016. Detecting Arithmetic Optimization Opportunities for C Compilers by Randomly Generated Equivalent Programs. *IPSJ Transactions on System LSI Design Methodology* 9 (2016), 21–29. https://doi.org/10.2197/ipsjtsldm.9.21

Mitsuyoshi Iwatsuji, Atsushi Hashimoto, and Nagisa Ishiura. 2016. Detecting Missed Arithmetic Optimization in C Compilers by Differential Random Testing. In *The 20th Workshop on Synthesis And System Integration of Mixed Information technologies (SASIMI 2016)*. http://ist.ksc.kwansei.ac.jp/~ishiura/publications/C2016-10a.pdf

Florent Kirchner, Nikolai Kosmatov, Virgile Prevosto, Julien Signoles, and Boris Yakobowski. 2015. Frama-C: A software analysis perspective. *Formal Aspects of Computing* 27, 3 (2015), 573–609. https://doi.org/10.1007/s00165-014-0326-7

Christopher Lidbury, Andrei Lascu, Nathan Chong, and Alastair F. Donaldson. 2015. Many-core Compiler Fuzzing. In *Proceedings of PLDI '15*. 65–76. https://doi.org/10.1145/2737924.2737986

Henry Massalin. 1987. Superoptimizer: A Look at the Smallest Program. In *Proceedings of ASPLOS II*. 122–126. https://doi.org/10.1145/36206.36194

William M. McKeeman. 1998. Differential Testing for Software. *Digital Technical Journal* 10, 1 (1998), 100–107.

Jan Midtgaard, Mathias Nygaard Justesen, Patrick Kasting, Flemming Nielson, and Hanne Riis Nielson. 2017. Effect-driven QuickChecking of Compilers. *Proc. ACM Program. Lang.* 1, ICFP, Article 15 (Aug. 2017). https://doi.org/10.1145/3110259

Tipp Moseley, Dirk Grunwald, and Ramesh Peri. 2009a. Chainsaw: Using Binary Matching for Relative Instruction Mix Comparison. In *PACT 2009*. 125–135. https://doi.org/10.1109/PACT.2009.12

Tipp Moseley, Dirk Grunwald, and Ramesh Peri. 2009b. OptiScope: Performance Accountability for Optimizing Compilers. In *Proceedings of CGO '09*. 254–264. https://doi.org/10.1109/CGO.2009.26

Eriko Nagai, Atsushi Hashimoto, and Nagisa Ishiura. 2014. Reinforcing Random Testing of Arithmetic Optimization of C Compilers by Scaling up Size and Number of Expressions. *IPSJ Transactions on System LSI Design Methodology* 7 (2014), 91–100. https://doi.org/10.2197/ipsjtsldm.7.91

Phitchaya Mangpo Phothilimthana, Aditya Thakur, Rastislav Bodik, and Dinakar Dhurjati. 2016. Scaling Up Superoptimization. In *Proceedings of ASPLOS '16*. 297–310. https://doi.org/10.1145/2872362.2872387

John Regehr, Yang Chen, Pascal Cuoq, Eric Eide, Chucky Ellison, and Xuejun Yang. 2012. Test-case Reduction for C Compiler

Bugs. In *Proceedings of PLDI '12*. 335–346. https://doi.org/10.1145/2254064.2254104

Jesse Ruderman. 2015. jsfunfuzz. Source code repository. (2015). https://github.com/MozillaSecurity/funfuzz/tree/master/js/jsfunfuzz

Yan Shoshitaishvili, Ruoyu Wang, Christopher Salls, Nick Stephens, Mario Polino, Andrew Dutcher, John Grosen, Siji Feng, Christophe Hauser, Christopher Kruegel, and Giovanni Vigna. 2016. SoK: (State of) The Art of War: Offensive Techniques in Binary Analysis. In *IEEE Symposium on Security and Privacy*.

Chengnian Sun, Vu Le, Qirun Zhang, and Zhendong Su. 2016. Toward Understanding Compiler Bugs in GCC and LLVM. In *Proceedings of the 25th International Symposium on Software Testing and Analysis (ISSTA 2016)*. 294–305. https://doi.org/10.1145/2931037.2931074

Tao Wei, Jian Mao, Wei Zou, and Yu Chen. 2007. A New Algorithm for Identifying Loops in Decompilation. In *Proceedings of the 14th International Conference on Static Analysis (SAS'07)*. 170–183. http://dl.acm.org/citation.cfm?id=2391451.2391464

Xuejun Yang, Yang Chen, Eric Eide, and John Regehr. 2011. Finding and Understanding Bugs in C Compilers. In *Proceedings of PLDI '11*. 283–294. https://doi.org/10.1145/1993498.1993532