# Structure-aware fuzzing for Clang and LLVM with libprotobuf-mutator

Kostya Serebryany, Vitaly Buka, Matt Morehouse; Google October 2017

# Agenda

- Fuzzing
- Fuzzing Clang/LLVM
- Fuzzing Clang/LLVM better (structure-aware)
  - o Ilvm-isel-fuzzer
  - clang-proto-fuzzer

## Testing vs Fuzzing

```
// Test // Fuzz
MyApi(Input1); while (true)
MyApi(Input2); MyApi(
MyApi(Input3); Fuzzer.GenerateInput());
```

## Types of fuzzing engines

- Coverage-guided
  - libFuzzer
  - o AFL
- Generation-based
  - o Csmith
- Symbolic execution
  - KLEE

• ...

## Coverage-guided fuzzing

- Acquire the initial corpus of inputs for your API
- while (true)
  - Randomly mutate one input
  - Feed the new input to your API
  - new code coverage => add the input to the corpus

#### libFuzzer

```
bool FuzzMe(const uint8 t *Data, size t DataSize) { // fuzz me.cc
  return DataSize >= 3 &&
      Data[0] == 'F' &&
      Data[1] == 'U' &&
      Data[2] == 'Z' &&
      Data[3] == 'Z'; // :-<
extern "C" int LLVMFuzzerTestOneInput(const uint8 t *Data, size t Size) {
  FuzzMe(Data, Size);
  return 0;
% clang -g -fsanitize=address, fuzzer fuzz me.cc && ./a.out
```

## Simple Fuzzers in LLVM

- clang-format-fuzzer
- clang-fuzzer
- Ilvm-dwarfdump-fuzzer
- Ilvm-as-fuzzer
- Ilvm-mc-assemble-fuzzer
- Ilvm-mc-disassemble-fuzzer
- Ilvm-demangle-fuzzer (Ilvm) & cxa\_demangle\_fuzzer (Iibcxxabi)
- ...

#### OSS-Fuzz + LLVM

- https://github.com/google/oss-fuzz
  - Continuous automated fuzzing for OSS projects
  - Usenix Security 2017
- TL;DR: fuzzers in, bug reports out
- LLVM: <a href="https://github.com/google/oss-fuzz/tree/master/projects/llvm/">https://github.com/google/oss-fuzz/tree/master/projects/llvm/</a>

# cxa\_demangle\_fuzzer

extern "C"

```
int LLVMFuzzerTestOneInput(
   const uint8_t *data, size_t size) {
  char *str = new char[size+1];
  memcpy(str, data, size);
  str[size] = 0;
  free( cxa demangle(str, 0, 0, 0));
  delete [] str;
  return 0;
```

```
Ilvm libcxxabi: Timeout in Ilvm libcxxabi cxa demangle fuzzer ClusterFuzz Reproducible
Ilvm libcxxabi: ASSERT: k0 <= k1 && "parse_type() mutated the name stack" ClusterFuzz Reproducible
Ilvm libcxxabi: ASSERT: k0 <= k1 && "parse type() mutated the name stack" ClusterFuzz Reproducible
Ilvm libcxxabi: Out-of-memory in Ilvm libcxxabi cxa demangle fuzzer ClusterFuzz Reproducible
Ilvm libcxxabi: Timeout in Ilvm libcxxabi cxa demangle fuzzer ClusterFuzz Reproducible
Ilvm libcxxabi: Timeout in Ilvm libcxxabi cxa demangle fuzzer ClusterFuzz Reproducible
Ilvm libcxxabi: Out-of-memory in Ilvm libcxxabi cxa demangle fuzzer ClusterFuzz Reproducible
Ilvm libcxxabi: Out-of-memory in Ilvm libcxxabi cxa demTraceback ClusterFuzz Reproducible
Ilvm libcxxabi: Use-of-uninitialized-value in cxxabiv1::parse nested name ClusterFuzz Reproducible
Ilvm libcxxabi: Null-dereference READ in cxxabiv1::parse nested name ClusterFuzz Reproducible
llvm libcxxabi: Sanitizer CHECK failure in ((IsAligned(reinterpret cast<uptr>(p), page size ))) != (0) (0, 0) ClusterFuzz
Ilvm libcxxabi: Bad-free in std:: 1:: vector base<std:: 1::vector< cxxabiv1::Node*, cxxabiv1::short al Cluster</p>
Ilvm libcxxabi: ASSERT: FromPosition <= names.size() ClusterFuzz Reproducible</p>
Ilvm libcxxabi: Use-of-uninitialized-value in cxxabiv1::Node::hasRHSComponent ClusterFuzz Reproducible
Ilvm libcxxabi: Null-dereference READ in cxxabiv1::Node::hasRHSComponent ClusterFuzz Reproducible
Ilvm libcxxabi: Bad-free in std:: 1:: vector base<std:: 1::vector< cxxabiv1::Node*, cxxabiv1::short al Cluster</p>
```

Ilvm libcxxabi: Negative-size-param in cxxabiv1::NodeArray cxxabiv1::Db::makeNodeArray<std:: 1:: wrap if</p>

Ilvm\_libcxxabi: Stack-buffer-overflow in std::\_\_1::basic\_string<char, std::\_\_1::char\_traits<char>, \_\_cxxabiv1::malloc

Ilvm\_libcxxabi: Negative-size-param in std::\_\_1::char\_traits<char>::move ClusterFuzz Reproducible
Ilvm\_libcxxabi: Out-of-memory in Ilvm\_libcxxabi\_cxa\_demangle\_fuzzer ClusterFuzz Reproducible
Ilvm\_libcxxabi: Timeout in Ilvm\_libcxxabi\_cxa\_demangle\_fuzzer ClusterFuzz Reproducible

Ilvm\_libcxxabi: Negative-size-param in std::\_\_1::char\_traits<char>::copy ClusterFuzz Reproducible
Ilvm libcxxabi: Out-of-memory in Ilvm libcxxabi cxa demangle fuzzer ClusterFuzz Reproducible

Ilvm libcxxabi: ASSERT: FromPosition <= names.size() ClusterFuzz Reproducible</p>

Ilvm\_libcxxabi: Out-of-memory in Ilvm\_libcxxabi\_cxa\_demangle\_fuzzer ClusterFuzz Reproducible
Ilvm\_libcxxabi: Heap-buffer-overflow in \_\_cxxabiv1::parse\_encoding ClusterFuzz Reproducible
Ilvm\_libcxxabi: Heap-buffer-overflow in \_\_cxxabiv1::parse\_new\_expr ClusterFuzz Reproducible
Ilvm\_libcxxabi: Timeout in Ilvm\_libcxxabi cxa\_demangle\_fuzzer ClusterFuzz Unreproducible

# clang-format-fuzzer

```
extern "C" int LLVMFuzzerTestOneInput(uint8 t *data, size t size) {
  // FIXME: fuzz more things: different styles, different style features.
  std::string s((const char *)data, size);
  auto Style = getGoogleStyle(clang::format::FormatStyle::LK Cpp);
  Style.ColumnLimit = 60;
  auto Replaces = reformat(Style, s, clang::tooling::Range(0, s.size()));
  auto Result = applyAllReplacements(s, Replaces);
  // Output must be checked, as otherwise we crash.
  if (!Result) {}
                                 Ilvm: Out-of-memory in Ilvm clang-format-fuzzer ClusterFuzz Reproducible
  return 0;
                                 Ilvm: ASSERT: getClient() && "DiagnosticClient not set!" ClusterFuzz Reproducible
                                 Ilvm: Stack-overflow in clang::format::AnnotatingParser::parseAngle ClusterFuzz Reproducible
                                 Ilvm: Null-dereference READ in clang::format::AnnotatingParser::consumeToken ClusterFuzz Reproducible
                                 Ilvm: ASSERT: IndentPrefix.startswith("//") ClusterFuzz Reproducible
                                 Ilvm: Timeout in Ilvm clang-format-fuzzer ClusterFuzz Unreproducible
                                 ASSERT: Changes[i - 1].OriginalWhitespaceRange.getBegin() != C.OriginalWhitespaceRange.g ClusterFuzz Reproducit
```

# Ilvm-dwarfdump-fuzzer

```
extern "C"
int LLVMFuzzerTestOneInput(uint8 t *data, size t size) {
  std::unique ptr<MemoryBuffer> Buff = MemoryBuffer::getMemBuffer(
      StringRef((const char *)data, size), "", false);
  Expected<std::unique ptr<ObjectFile>> ObjOrErr =
      ObjectFile::createObjectFile(Buff->getMemBufferRef());
  if (auto E = ObjOrErr.takeError()) {
    consumeError(std::move(E));
    return 0:
  ObjectFile &Obj = *ObjOrErr.get();
  std::unique ptr<DIContext> DICtx = DWARFContext::create(Obj);
  DIDumpOptions opts;
  opts.DumpType = DIDT All;
  DICtx->dump(nulls(), opts);
  return 0:
```

```
Ilvm: ASSERT: result <= UINT32_MAX ClusterFuzz Reproducible</p>
Ilvm: Heap-buffer-overflow in Ilvm::object::WasmObjectFile::pars
```

Ilvm: Heap-buffer-overflow in Ilvm::StringMapImpl::LookupBucke

Ilvm: Heap-buffer-overflow in Ilvm::identify\_magic ClusterFuzz Rep

Ilvm: ASSERT: sizeof(Elf\_Ehdr) <= Buf.size() && "Invalid buffer"
Ilvm: Out-of-memory in Ilvm\_Ilvm-dwarfdump-fuzzer ClusterFuzz R</pre>

Ilvm: Heap-buffer-overflow in readInitExpr ClusterFuzz Reproducible

Ilvm: Heap-buffer-overflow in checkDylibCommand ClusterFuzz Re

Ilvm: Crash in Ilvm::DataExtractor::getUnsigned ClusterFuzz Reproducible
Ilvm: Abrt in Ilvm::report\_bad\_alloc\_error ClusterFuzz Reproducible

IIvm: ASSERT: result <= INT32\_MAX && result >= INT32\_MIN (

Ilvm: Heap-buffer-overflow in Ilvm::DataExtractor::getU32 Cluster
Ilvm: Heap-buffer-overflow in Ilvm::raw\_svector\_ostream::write\_

Ilvm: Abrt in Ilvm::Ilvm unreachable internal ClusterFuzz Reproducil

Ilvm: Heap-buffer-overflow in Ilvm::DataExtractor::getCStr Cluster
Ilvm: Heap-buffer-overflow in Ilvm::DataExtractor::getU32 Clusterl

Ilvm: Heap-buffer-overflow in Ilvm::DataExtractor::getUnsigned (
Ilvm: Abrt in Ilvm::Ilvm\_unreachable\_internal ClusterFuzz Reproducit

Ilvm: Heap-buffer-overflow in Ilvm::DataExtractor::getCStr Cluster
Ilvm: Heap-buffer-overflow in Ilvm::StringMapImpl::FindKey Cluster

llvm/llvm-dwarfdump-fuzzer: Heap-buffer-overflow in llvm::identi

## clang-fuzzer

```
void clang_fuzzer::HandleCXX(const std::string &S,
                             const std::vector<const char *> &ExtraArgs) {
 llvm::InitializeAllTargets();
 llvm::InitializeAllTargetMCs();
 llvm::InitializeAllAsmPrinters();
 llvm::InitializeAllAsmParsers();
 llvm::opt::ArgStringList CC1Args;
 CC1Args.push_back("-cc1");
 for (auto &A : ExtraArgs)
   CC1Args.push back(A);
 CC1Args.push_back("./test.cc");
 llvm::IntrusiveRefCntPtr<FileManager> Files(
     new FileManager(FileSystemOptions()));
 IgnoringDiagConsumer Diags;
 IntrusiveRefCntPtr<DiagnosticOptions> DiagOpts = new DiagnosticOptions();
 DiagnosticsEngine Diagnostics(
     IntrusiveRefCntPtr<clang::DiagnosticIDs>(new DiagnosticIDs()), &*DiagOpts,
     &Diags, false);
 std::unique_ptr<clang::CompilerInvocation> Invocation(
     tooling::newInvocation(&Diagnostics, CC1Args));
 std::unique_ptr<llvm::MemoryBuffer> Input =
     11vm::MemoryBuffer::getMemBuffer(S);
 Invocation->getPreprocessorOpts().addRemappedFile("./test.cc",
                                                    Input.release());
 std::unique_ptr<tooling::ToolAction> action(
     tooling::newFrontendActionFactory<clang::EmitObjAction>());
 std::shared ptr<PCHContainerOperations> PCHContainerOps =
     std::make_shared<PCHContainerOperations>();
  action->runInvocation(std::move(Invocation), Files.get(), PCHContainerOps,
                        &Diags);
```

Ilvm: ASSERT: DelayedTypos.empty() && "Uncorrected typos!" ClusterFuzz Reproducible

Ilvm: ASSERT: ParmVarDeclBits.ScopeDepthOrObjCQuals == scopeDepth && "truncation!" ClusterFuzz Reprodu

Ilvm: ASSERT: CurPtr[-1] == '<' && CurPtr[0] == '#' && "Not a placeholder!" ClusterFuzz Reproducible

Ilvm: ASSERT: !isTokenSpecial() && "Should consume special tokens with Consume\*Token" ClusterFuzz Reprodu

Ilvm: Stack-buffer-overflow in clang::Lexer::SkipLineComment ClusterFuzz Reproducible

Ilvm: ASSERT: Access != AS none && "Access specifier is AS none inside a record dec!" ClusterFuzz Reproducit

Ilvm: ASSERT: CachedTokens[CachedLexPos-1].getLastLoc() == Tok.getAnnotationEndLoc() && "The a Cluster of the Country of the C

Ilvm: ASSERT: lisNull() && "Cannot retrieve a NULL type pointer" ClusterFuzz Reproducible

Ilvm: ASSERT: ResultKind != Found || Decls.size() == 1 ClusterFuzz Reproducible

Ilvm: ASSERT: Tok.is(tok::eof) && Tok.getEofData() == AttrEnd.getEofData() ClusterFuzz Reproducible

Ilvm: ASSERT: Access == AS private || Access == AS protected ClusterFuzz Reproducible

Ilvm: ASSERT: RHS.U.VAL != 0 && "Divide by zero?" ClusterFuzz Reproducible

Ilvm: ASSERT: RHS.U.VAL != 0 && "Divide by zero?" ClusterFuzz Reproducible

Ilvm: ASSERT: !CodeSynthesisContexts.empty() ClusterFuzz Reproducible

Ilvm: ASSERT: Result.isInvalid() && "C++ binary operator overloading is missing candidates!" ClusterFuzz Reproduction

Ilvm: Abrt in Ilvm::Ilvm\_unreachable\_internal ClusterFuzz Reproducible

Ilvm: Direct-leak in clang::Parser::ParseParameterDeclarationClause ClusterFuzz Reproducible

Ilvm: ASSERT: !Prev.isAmbiguous() && "Cannot have an ambiguity in previous-declaration lookup" ClusterFuzz

Ilvm: ASSERT: DD && "queried property of class with no definition" ClusterFuzz Reproducible

Ilvm: ASSERT: getContainingDC(DC) == CurContext && "The next DeclContext should be lexically c ClusterFuz

llvm: ASSERT: (OtherT->isIntegerType() && ConstantT->isIntegerType()) && "comparison with non- ClusterFuzz

IIvm: ASSERT: Ancestor->getEntity() == CurContext && "ancestor context mismatch" ClusterFuzz Reproducible

llvm: ASSERT: CodeDC && !CodeDC->isFileContext() && "statement expr not in code context" ClusterFuzz Repr

Ilvm: ASSERT: BT->isInteger() ClusterFuzz Reproducible

Ilvm/clang-fuzzer: ASSERT: BitWidth && "bitwidth too small" ClusterFuzz Reproducible

# libFuzzer's default (generic) mutations

- Bit flip
- Byte swap
- Insert magic values
- Remove byte sequences
- ...

## clang-fuzzer (using generic mutations)

<u>heap-buffer-overflow in clang::Lexer::SkipLineComment on a 4-byte input</u>

//\\

<u>use-after-free or Assertion `Tok.is(tok::eof) && Tok.getEofData()</u> == AttrEnd.getEofDat a()'.

infinite CPU and RAM consumption on a 62-byte input

cFjass��F:�{�F\*NFF(;F\*�FF=F(JFF=F:FFF.FFF-FFF')





#### Parser



#### **Optimizer**



## Problem with generic mutations

Some APIs consume highly structured data

Generic mutations create invalid data that doesn't parse

#### Structure-aware mutations

- Specialized solution for a given input type
- Parse one input, reject if doesn't parse
- Mutate the AST and/or the leaf nodes in memory

#### Ilvm-isel-fuzzer: structure-aware LLVM IR fuzzer

- Justin Bogner "Adventures in Fuzzing Instruction Selection" <u>Euro LLVM '17</u>
- libFuzzer + Custom Mutator:
  - Parse LLVM IR
  - Mutate IR in memory (llvm/FuzzMutate/IRMutator.h)
  - Feed the mutation to an LLVM pass

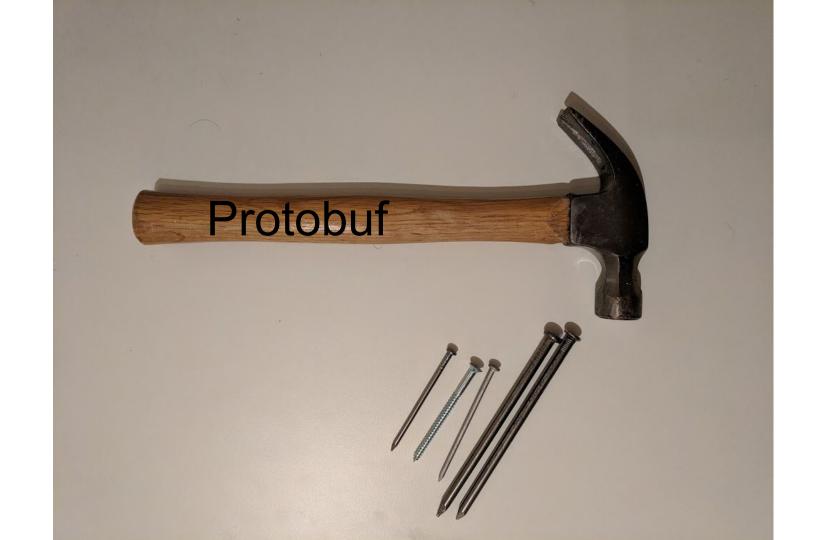
#### Ilvm-isel-fuzzer

```
https://bugs.chromium.org/p/oss-fuzz/issues/detail?id=3628
```

```
LLVM ERROR: VReg has no regclass after selection
source filename = "M"
define void @f() {
BB:
  br label %BB1
BB1:
               : preds = %BB
  %G13 = getelementptr i16*, i16** undef, i1 false
 %A6 = alloca i1
 %A2 = alloca i1*
 %C1 = icmp ult i32 2147483647, 0
  store i1* %A6, i1** %A2
  store i1 %C1, i1* %A6
  store i16** %G13, i16*** undef
  ret void
```

#### https://bugs.chromium.org/p/oss-fuzz/issues/detail?id=3629

```
Assertion `Offset <= INT MAX && "Offset too big to fit
in int."' failed.
source filename = "M"
define void @f() {
BB:
 %A11 = alloca i16
 %A7 = alloca i1, i32 -1
 %L4 = load i1, i1* %A7
 store i16 -32768, i16* %A11
  br label %BB1
BB1:
            ; preds = %BB
 %C5 = icmp eq i1 %L4, %L4
 store i1 %C5, i1* undef
 store i16*** undef, i16**** undef
 ret void
```





# https://github.com/google/protobuf

Protocol Buffers (a.k.a., protobuf) are Google's language-neutral, platform-neutral, extensible mechanism for serializing structured data

```
// Msg.proto
message Msg {
   string str = 1;
   int32 num = 2;
}
```

```
// orig.txt
str: "hello"
num: 42
```

## https://github.com/google/libprotobuf-mutator

Applies a single random mutation to a protobuf message

Valid message in - valid message out

```
// Msg.proto
message Msg {
  string str = 1;
  int32  num = 2;
}
// orig.txt
str: "hello"
num: 42

// mut1.txt
str: "help"
num: 42

// mut2.txt
str: "help"
num: 42
```

## https://github.com/google/libprotobuf-mutator

```
// my_api.cpp
void MyApi(const Msg &input) {
  if (input.str() == "help" && input.num() == 911)
      abort(); // bug
}
```

```
// my_api_fuzzer.cpp
DEFINE_PROTO_FUZZER(const Msg& input) {
   MyApi(input);
}
```

# Fuzz clang/llvm via protobufs

- Define a protobuf type that represent a subset of C++
  - o message Function { ...

```
// tools/clang-fuzzer/cxx proto.proto
message BinaryOp {
 enum Op {
  PLUS = 0;
  MINUS = 1; ...
required Op op = 1;
required Rvalue left = 2;
required Rvalue right = 3;
message Rvalue {
oneof rvalue oneof {
  VarRef varref = 1;
  Const cons = 2;
   BinaryOp binop = 3;
message AssignmentStatement {
required Lvalue lvalue = 1;
required Rvalue rvalue = 2;
} ...
```

## Fuzz clang/llvm via protobufs

 Define a protobuf type that represent a subset of C++

```
o message Function { ...
```

Implement a proto => C++ converter

```
// tools/clang-fuzzer/proto-to-cxx/proto to cxx.cpp
std::ostream &operator<<(std::ostream &os,</pre>
                           const BinaryOp &x) {
os << "(" << x.left();
switch (x.op()) {
   case BinaryOp::PLUS: os << "+"; break;</pre>
   case BinaryOp::MINUS: os << "-"; break; ...</pre>
return os << x.right() << ")";</pre>
}
std::ostream &operator<<(std::ostream &os,</pre>
                          const Rvalue &x) {
   if (x.has varref()) return os << x.varref();</pre>
   if (x.has cons()) return os << x.cons();</pre>
   if (x.has_binop()) return os << x.binop();</pre>
   return os << "1";
}
std::ostream &operator<<(std::ostream &os,
             const AssignmentStatement &x) {
return os <<x.lvalue() << "=" << x.rvalue() << ";\n";</pre>
```

## Fuzz clang/llvm via protobufs

- Define a protobuf type that represent a subset of C++
  - o message Function { ...
- Implement a proto => C++ converter
- Implement a fuzz target
  - HandleCXX same as in clang-fuzzer
- Current state: toy prototype

clang hangs in llvm::JumpThreadingPass::ComputeValueKnownInPredecessors

```
void foo(int *a) {
  while ((1 + 1)) {
    while ((a[96] * a[96])) {
      a[0] = (1024);
      while (a[0]) {
        while (a[0]) {
          (void)0;
          while ((a[96] * ((a[96] * a[96]) < 1))) {
            a[96] = (1 + 1);
          a[0] = (a[0] + a[0]);
```





#### Parser



#### **Optimizer**



<u>use-after-poison in Ilvm::SelectionDAG::Combine</u>

```
void foo(int *a) {
 while (1) {
  a[0] = (a[0] + (15134));
  while ((1 / a[6])) {
   (void)0;
  a[0] = (a[0] + (1 + 1));
  a[8] =
     (-2147483648)) *
          a[0]) +
          ((1 + 1) + (0))) -
         a[0]) *
         a[0]) +
          ((1 + 1) + (0))) *
         a[0])) -
        a[0]) *
       - [0]
```





#### Parser



#### **Optimizer**



<u>fatal error: error in backend: Cannot select: t195: i1 = add t192, t194 (in HexagonDAGToDAGISel::Select)</u>

```
void foo(int *a) {
 while ((
   (((a[0] -
     a[0]) *
        a[0]) *
        a[0]) +
           a[0]) *
          a[0]) *
         a[0]) &
         1) -
        1)) &
       1) -
      1) *
      1) *
     a[26])) *
    a[0]) *
    a[0]) +
   a[0])) {
  a[0] = (((a[26] * 1) + a[0]) * 1);
```

Lexer



Parser



Optimizer



#### null deref in llvm::ScalarEvolution::getMulExpr

```
void foo(int *a) {
  while (1) {
    a[60] = ((1 + a[60]) + a[0]);
    while ((a[60] + a[0])) {
        a[0] = (a[0] + 1);
    }
  }
}
```

#### Lexer



#### Parser



#### **Optimizer**



#### Custom IR mutator

- More work (but already done)
- Easier to reuse existing tests as corpus (?)
- No need to introduce one more IR
- Doesn't involve Clang => faster

#### vs proto-mutator

- Full C++ is hard to express in protobuf
- Easy to target a specific subset of C++
- Not LLVM-specific, can apply to other compilers and languages

#### Csmith?

- https://embed.cs.utah.edu/csmith/
- Generation-based, does not use coverage feedback
- Generates valid runnable programs
- Wish: Csmith + libFuzzer + protobuf-mutator == Csmith v2

#### **Problems**

- Bugs are being fixed too slow (if at all)
  - Not suitable for 'starter' projects due to code review latency
- Timeouts
- Clang/LLVM is very slow on small inputs
  - 5-20 inputs per second, w/o hitting timeouts

#### What's next

- clang-proto-fuzzer & Ilvm-isel-fuzzer run on OSS-Fuzz
  - let's observe
- How to contribute to the clang-proto-fuzzer prototype:
  - Try to express other/larger subset of C++ in a protobuf
    - Loop nests for to fuzz polly?
  - Try to make programs runnable (like csmith)
  - Try with other compilers
- How to contribute to fuzzing LLVM in general:
  - Fix <u>crashes</u>, <u>timeouts</u>, <u>and OOMs</u> and/or review the fixes
  - Developing a new feature? Create a dedicated fuzzer & <u>add it to OSS-Fuzz</u>

# Q&A