Symbolic execution for security researchers

Arnau Gàmez i Montolio

Navaja Negra Conference XI - October 6, 2023 - Albacete

About

Arnau Gàmez i Montolio

Hacker, Reverse Engineer & Mathematician

Occupation

- Senior Expert Engineer, Security @ Activision
- Founder, Researcher & Trainer @ Fura Labs
- PhD @ City, University of London

Contact

arnaugamez.com

🔥 Slides are here!

Preliminaries

Expectations

- Aims to be pretty introductory
- Demystify symbolic execution for a non-specialized audience
- Focus on understanding ideas rather than specific tooling
    - Apply it to your own areas of interest
    - Make it easy to use *any* tools (and understand what you are doing)
    - Make it easy to contribute to/write your own (open source) tools

Calculator

Concrete calculations

$$\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = 4 - 2 \cdot 3 = -2$$

Calculator

Concrete calculations

$$\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = 4 - 2 \cdot 3 = -2$$

Computer Algebra System (CAS)

Symbolic calculations and expression manipulation

$$\begin{vmatrix} 1 & 2 \\ a & 4 \end{vmatrix} = 4 - 2a = 2(2 - a)$$

Intermediate representation / language (IR/IL)

Language of an abstract machine designed to aid in the analysis of computer programs:

- Compilation: common ground for architecture independent processing
- Decompilation (binary analysis): lifting from ASM to canonical *higher level* representation
- Transpiling: source to source compilation

What is symbolic execution?

Roughly speaking, just a **computer algebra system** for:

- Programming languages: C, C++, Java, Rust...
- Assembly languages: x86, x86-64, ARM64, MIPS, RISC-V...
- Intermediate languages: LLVM-IR, SMT-LIB, r2 ESIL, IDA Microcode, $YOUR_OWN...

More specifically, symbolic execution is a **program analysis technique**:

- Represent inputs as *symbolic* variables instead of *concrete* values (normal execution or emulation)
- Derive constraints that encode control-flow and data-flow with respect to these symbolic variables

Use these constraints to reason about and extract information from the program

```
int foo(int x, int y) {
    x = y - 3*x;
    if (x < y) {
        return 2*x - x^y;
    }
    else {
        return 3*y + x|y;
    }
}
```

```c
int foo(int x, int y) {
    x = y - 3*x;
    if (x < y) {
        return 2*x - x^y;
    }
    else {
        return 3*y + x|y;
    }
}
```

Concrete execution (or emulation)

$$x = 1337, y = 7331$$

$$x = 7331 - 3 \times 1337 = 3320$$

$$(3320 < 7331)$$

$$2 \times 3320 - 1337 \oplus 7331 = 2068$$

$$\hookrightarrow 2068$$

```
int foo(int x, int y) {
    x = y - 3*x;
    if (x < y) {
        return 2*x - x^y;
    }
    else {
        return 3*y + x|y;
    }
}
```

Concrete execution (or emulation)

$$x = 1337, y = 7331$$
$$x = 7331 - 3 \times 1337 = 3320$$

$$(3320 < 7331)$$
$$2 \times 3320 - 1337 \oplus 7331 = 2068$$
$$\hookrightarrow 2068$$

Symbolic execution

$$x = \mathbf{x}, y = \mathbf{y}$$
$$\mathbf{x} = \mathbf{y} - 3\mathbf{x}$$

```c
int foo(int x, int y) {
    x = y - 3*x;
    if (x < y) {
        return 2*x - x^y;
    }
    else {
        return 3*y + x|y;
    }
}
```

Concrete execution (or emulation)

$$x = 1337, y = 7331$$
$$x = 7331 - 3 \times 1337 = 3320$$

$$(3320 < 7331)$$
$$2 \times 3320 - 1337 \oplus 7331 = 2068$$
$$\hookrightarrow 2068$$

Symbolic execution

$$x = \mathbf{x}, y = \mathbf{y}$$
$$\mathbf{x} = \mathbf{y} - 3\mathbf{x}$$

$$(\mathbf{y} - 3\mathbf{x} < \mathbf{y})$$
$$\hookrightarrow 2(\mathbf{y} - 3\mathbf{x}) - (\mathbf{y} - 3\mathbf{x}) \oplus \mathbf{y}$$

```c
int foo(int x, int y) {
    x = y - 3*x;
    if (x < y) {
        return 2*x - x^y;
    }
    else {
        return 3*y + x|y;
    }
}
```

Concrete execution (or emulation)

$x = 1337, y = 7331$

$x = 7331 - 3 \times 1337 = 3320$

$(3320 < 7331)$

$2 \times 3320 - 1337 \oplus 7331 = 2068$

$\hookrightarrow 2068$

Symbolic execution

$x = \mathbf{x}, y = \mathbf{y}$

$\mathbf{x} = \mathbf{y} - 3\mathbf{x}$

$(\mathbf{y} - 3\mathbf{x} < \mathbf{y})$

$\hookrightarrow 2(\mathbf{y} - 3\mathbf{x}) - (\mathbf{y} - 3\mathbf{x}) \oplus \mathbf{y}$

$(\mathbf{y} - 3\mathbf{x} \geq \mathbf{y})$

$\hookrightarrow 3(\mathbf{y} - 3\mathbf{x}) - (\mathbf{y} - 3\mathbf{x}) \vee \mathbf{y}$

But how does it *actually* work?

1. Define two data structures:
    - **path_constraint**: conditions required to reach current instruction
    - **state_map**: symbolic mapping for the variables (registers, memory locations)
2. Extract the semantics of each statement (instruction)
3. Update these two data structures to account for the effects of the *executed* statement (instruction)
4. If there is control-flow branching, *fork* these structures to keep track of different execution paths

The **state_map** represents *data-flow* updates, i.e. the (computational) process though which a variable ends up holding a certain value at a given point in the program execution.

The **state_map** represents *data-flow* updates, i.e. the (computational) process though which a variable ends up holding a certain value at a given point in the program execution.

The **path_constraint** represents *control-flow* tracking, i.e. the set of constraints (conditions) on the variables that need to be satisfied for the execution to reach a given point in the program.

Visual example

```
_start:
    mov rax, 123  <=0=
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> rax
                 rbx -> rbx
                 rdi -> rdi
                 rsi -> rsi
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi  <=0=
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
```

```
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> 123
                 rbx -> rbx
                 rdi -> rdi
                 rsi -> rsi
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi  <=0=
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> (123 + rsi)
                 rbx -> rbx
                 rdi -> rdi
                 rsi -> rsi
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2      <=0=
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi

    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> ((123 + rsi) ^ rdi)
                 rbx -> rbx
                 rdi -> rdi
                 rsi -> rsi
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx   <=0=
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> ((123 + rsi) ^ rdi)
                 rbx -> 2
                 rdi -> rdi
                 rsi -> rsi
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3    <=0=
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> (((123 + rsi) ^ rdi) + 2)
                 rbx -> 2
                 rdi -> rdi
                 rsi -> rsi
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax  <=0=
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> (((123 + rsi) ^ rdi) + 2)
                 rbx -> 2
                 rdi -> 3
                 rsi -> rsi
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx  <=0=
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> (((123 + rsi) ^ rdi) + 2)
                 rbx -> 2
                 rdi -> 3
                 rsi -> (((123 + rsi) ^ rdi) + 2)
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi  <=0=
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> ((((123 + rsi) ^ rdi) + 2) + 2)
                 rbx -> 2
                 rdi -> 3
                 rsi -> (((123 + rsi) ^ rdi) + 2)
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7     <=0=
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> (((((123 + rsi) ^ rdi) + 2) + 2) ^ 3)
                 rbx -> 2
                 rdi -> 3
                 rsi -> (((123 + rsi) ^ rdi) + 2)
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx  <=0=
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> (((((123 + rsi) ^ rdi) + 2) + 2) ^ 3)
                 rbx -> 7
                 rdi -> 3
                 rsi -> (((123 + rsi) ^ rdi) + 2)
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336 <=0=
    add rax, rdi
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> ((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7)
                 rbx -> 7
                 rdi -> 3
                 rsi -> (((123 + rsi) ^ rdi) + 2)
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi  <=0=
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> ((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7)
                 rbx -> 7
                 rdi -> 1336
                 rsi -> (((123 + rsi) ^ rdi) + 2)
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337 <=0=
    jnz bad

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> (((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7) + 1336)
                 rbx -> 7
                 rdi -> 1336
                 rsi -> (((123 + rsi) ^ rdi) + 2)
                 zf  -> zf
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337
    jnz bad        <=0=

good:
    xor rdi, rdi
    jmp exit

bad:
    mov rdi, 1

exit:
    mov rax, 60
    syscall
```

```
path_constraint  true
state_map        rax -> (((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7) + 1336)
                 rbx -> 7
                 rdi -> 1336
                 rsi -> (((123 + rsi) ^ rdi) + 2)
                 zf  -> (((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7) + 1336)
                        == 1337 ? 1 : 0
```

```
_start:
    mov rax, 123
    add rax, rsi
    xor rax, rdi
    mov rbx, 2
    add rax, rbx
    mov rdi, 3
    mov rsi, rax
    add rax, rbx
    xor rax, rdi
    mov rbx, 7
    and rax, rbx
    mov rdi, 1336
    add rax, rdi
    cmp rax, 1337
    jnz bad

good:
    xor rdi, rdi  <=1=
    jmp exit

bad:
    mov rdi, 1    <=2=

exit:
    mov rax, 60
    syscall
```

```
path_constraint  (((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7) + 1336) == 1337
state_map        ...
                 zf  -> 1

path_constraint  (((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7) + 1336) != 1337
state_map        ...
                 zf  -> 0
```

How do we *reason* about this information?

How do we *reason* about this information?

With an SMT solver

How do we *reason* about this information?

With an SMT solver

Mostly

SMT solver

SMT solver

Satisfiability Modulo Theories

- **S**atisfiability (SAT): determine if a (boolean) formula can be satisfied (can be true)
- **M**odulo: take into account (not only boolean formulas but also)...
- **T**heories: ...integer numbers, real numbers, floating point, **bit vectors**, and more

SMT solver

Satisfiability Modulo Theories

- **S**atisfiability (SAT): determine if a (boolean) formula can be satisfied (can be true)
- **M**odulo: take into account (not only boolean formulas but also)...
- **T**heories: ...integer numbers, real numbers, floating point, **bit vectors**, and more

From a very practical standpoint: a *magic black-box* that can only answer a very simple question.

Question

Given some variables of some type, and some constraints on these variables:

- Is there any variable assignment that makes the set of constraints satisfiable, i.e. such that (all) the constraints hold true?

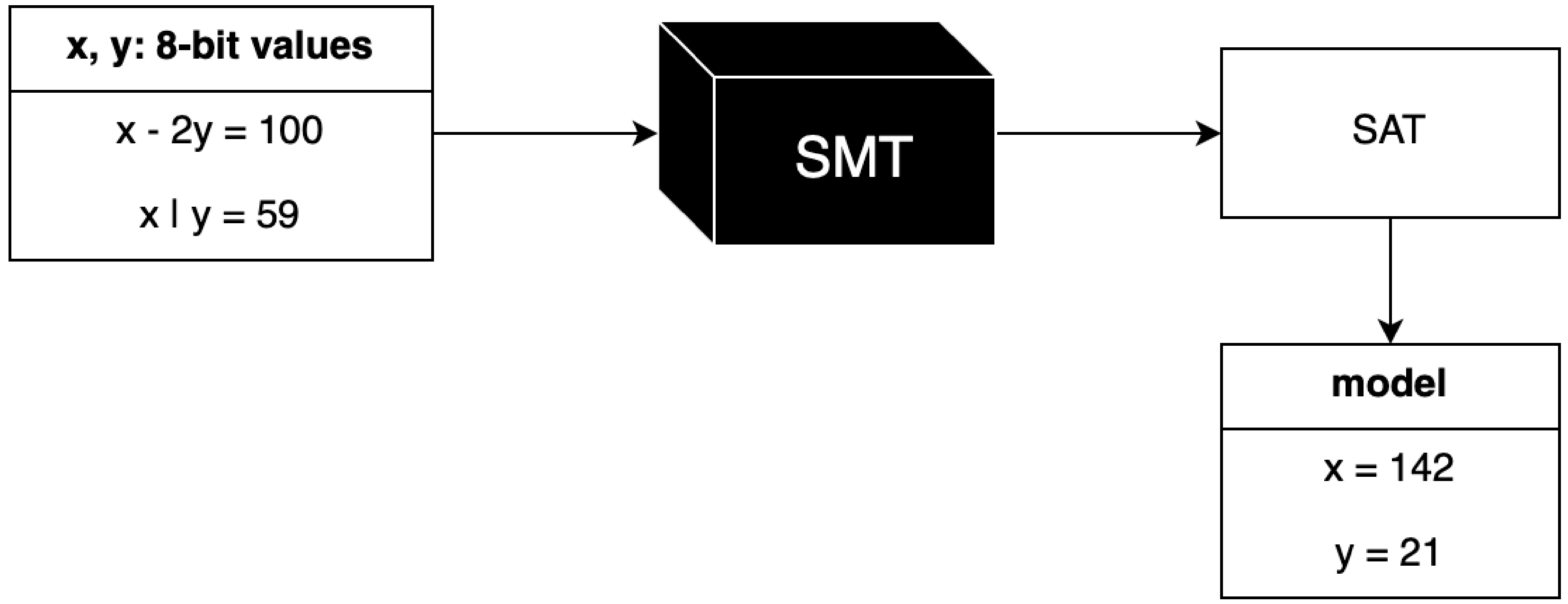
Question

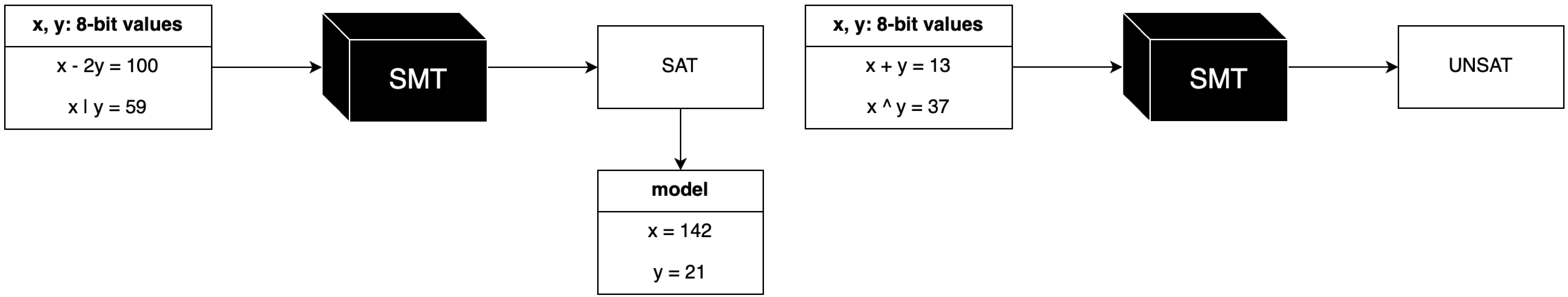
Given some variables of some type, and some constraints on these variables:

- Is there any variable assignment that makes the set of constraints satisfiable, i.e. such that (all) the constraints hold true?

Outcomes

- SAT: there is a variable assignment that makes all the constraints hold true.
    - It will actually find a model, which is a particular solution (a concrete variable assignment)
- UNSAT: there is NO variable assignment that makes all the constraints hold true.
- UNKNOWN: unable to answer the question (usually due to a time-out)


x, y: 8-bit values
x - 2y = 100
x | y = 59
SMT
SAT
model
x = 142
y = 21


x, y: 8-bit values
x - 2y = 100
x | y = 59
SMT
SAT
model
x = 142
y = 21
x, y: 8-bit values
x + y = 13
x ^ y = 37
SMT
UNSAT

Symbolic execution + SMT solver

Symbolic execution + SMT solver

Some basic ideas

Data-flow analysis

Data-flow analysis

- Embed *compiler optimization* techniques into the `state_map` population process:
  - Constant propagation: by construction
  - Constant folding: evaluate intermediate expressions on constant values
  - Reaching definitions: calculate at a given point the set of definitions that reach it
  - Liveness analysis: calculate at a given point the *live* variables (may be read before updated)

1. The symbolic execution engine is used to extract the formula of the return value of a function with respect to its inputs parameters: check its value in the `state_map`

1. The symbolic execution engine is used to extract the formula of the return value of a function with respect to its inputs parameters: check its value in the `state_map`

2. The formula is fed into the SMT solver

1. The symbolic execution engine is used to extract the formula of the return value of a function with respect to its inputs parameters: check its value in the `state_map`

2. The formula is fed into the SMT solver

3. The SMT can:
    - Attempt to simplify the formula to get a nicer representation
    - Craft inputs value that will make the formula evaluate to a desired output (i.e. inputs that will make the function return a desired value)

Control-flow analysis

Control-flow analysis

1. The symbolic execution engine is used to extract the formulae (constraints) for a given path branching to happen: check its `path_constraint`

Control-flow analysis

1. The symbolic execution engine is used to extract the formulae (constraints) for a given path branching to happen: check its `path_constraint`

2. The constraints are fed into the SMT solver

Control-flow analysis

1. The symbolic execution engine is used to extract the formulae (constraints) for a given path branching to happen: check its `path_constraint`

2. The constraints are fed into the SMT solver

3. The SMT solver can prove the feasibility of the constraints, meaning the path is reachable
    - If it is, retrieve a model for it, i.e. input values that will make the program execution to reach it
    - If it is not, we have detected an obfuscating opaque predicate and can ignore/patch it away

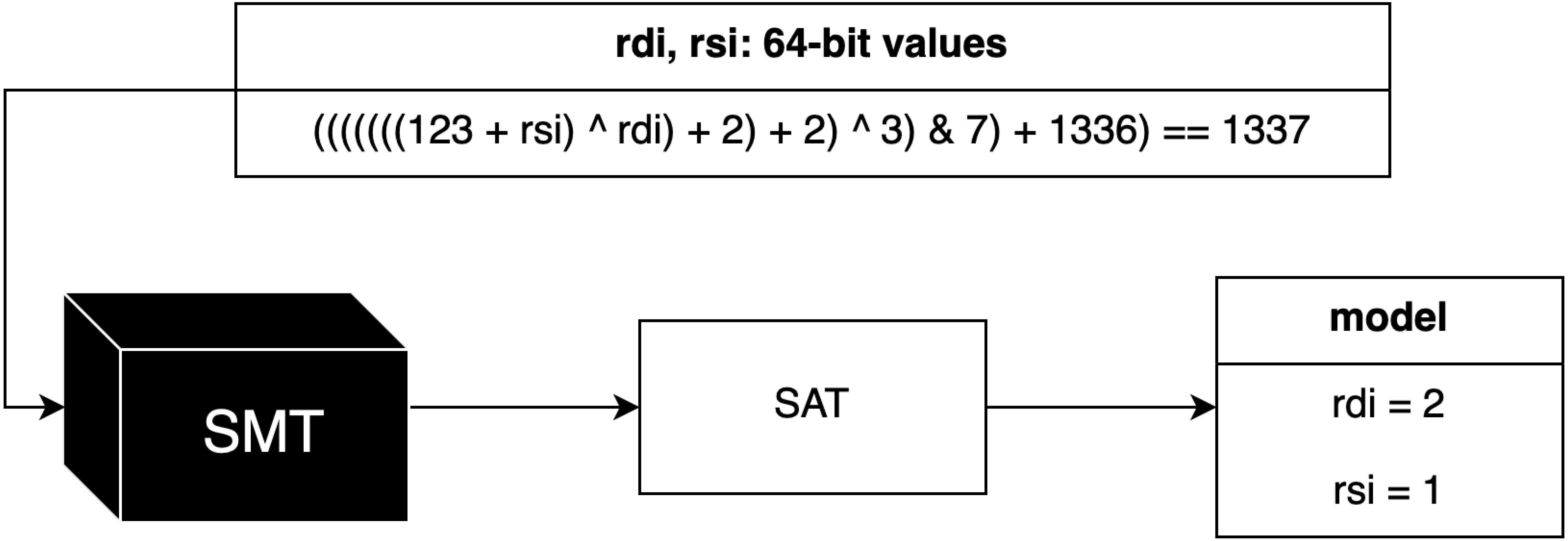
Example

```
path_constraint  (((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7) + 1336) == 1337
```

```
path_constraint  (((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7) + 1336) == 1337
```

Given 64-bit variables `rdi` and `rsi`:

- Is there any variable assignment (for `rdi` and `rsi`) that makes the `path_constraint` satisfiable?


rdi, rsi: 64-bit values
(((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7) + 1336) == 1337
SMT
SAT
model
rdi = 2
rsi = 1

```python
import z3

rdi, rsi = z3.BitVecs('rdi rsi', 64)
path_constraint = (((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7) + 1336) == 1337

solver = z3.Solver()
solver.add(path_constraint)

if solver.check() == z3.sat:
    print(solver.model())
```

```python
import z3

rdi, rsi = z3.BitVecs('rdi rsi', 64)
path_constraint = ((((((((123 + rsi) ^ rdi) + 2) + 2) ^ 3) & 7) + 1336) == 1337

solver = z3.Solver()
solver.add(path_constraint)

if solver.check() == z3.sat:
    print(solver.model())
```

```
[rdi = 2, rsi = 1]
```

Tooling

Tooling

Welcome to the jungle

Implementation technology

- **Interpreter based**: Miasm, Triton, Angr, Maat, radius2
- **Instrumentation based**: QSYM
- **Compiler based**: KLEE, SymCC, SymQEMU

Implementation technology

- **Interpreter based**: Miasm, Triton, Angr, Maat, radius2
- **Instrumentation based**: QSYM
- **Compiler based**: KLEE, SymCC, SymQEMU

Target

- **Binary**: Miasm, Triton, Angr, Maat, radius2, QSYM, SymQEMU
- **Source code**: KLEE, SymCC

Implementation technology

- **Interpreter based**: Miasm, Triton, Angr, Maat, radius2
- **Instrumentation based**: QSYM
- **Compiler based**: KLEE, SymCC, SymQEMU

Target

- **Binary**: Miasm, Triton, Angr, Maat, radius2, QSYM, SymQEMU
- **Source code**: KLEE, SymCC

Focus

- **Analysis**: Miasm, Triton, Maat
- **Automagic**: Angr, radius2
- **Test generation**: QSYM, KLEE, SymCC, SymQEMU

Practical applications

Practical applications

An appetizer

Analysis of complex code

Detect (and patch) opaque predicates

Opaque predicates

A conditional statement P whose truth value is known a priori.

Opaque predicates

A conditional statement P whose truth value is known a priori.

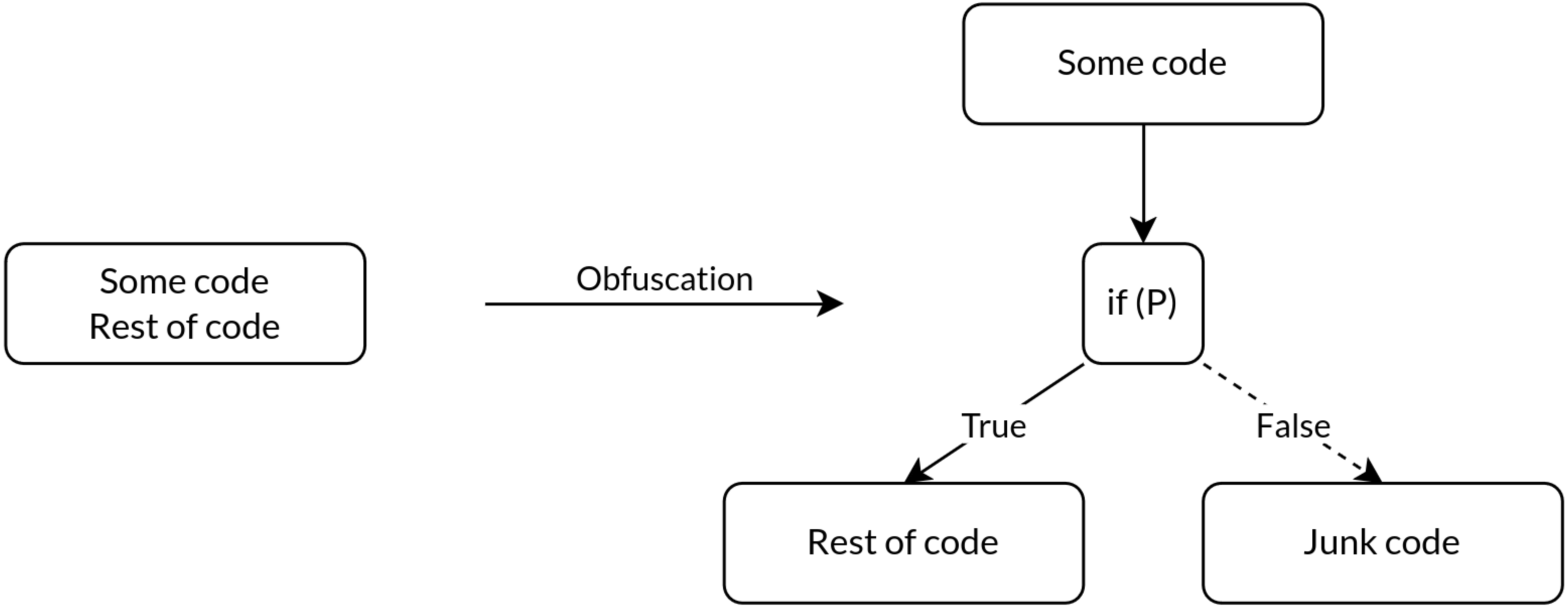
$x^2 \geq 0$ is always true.

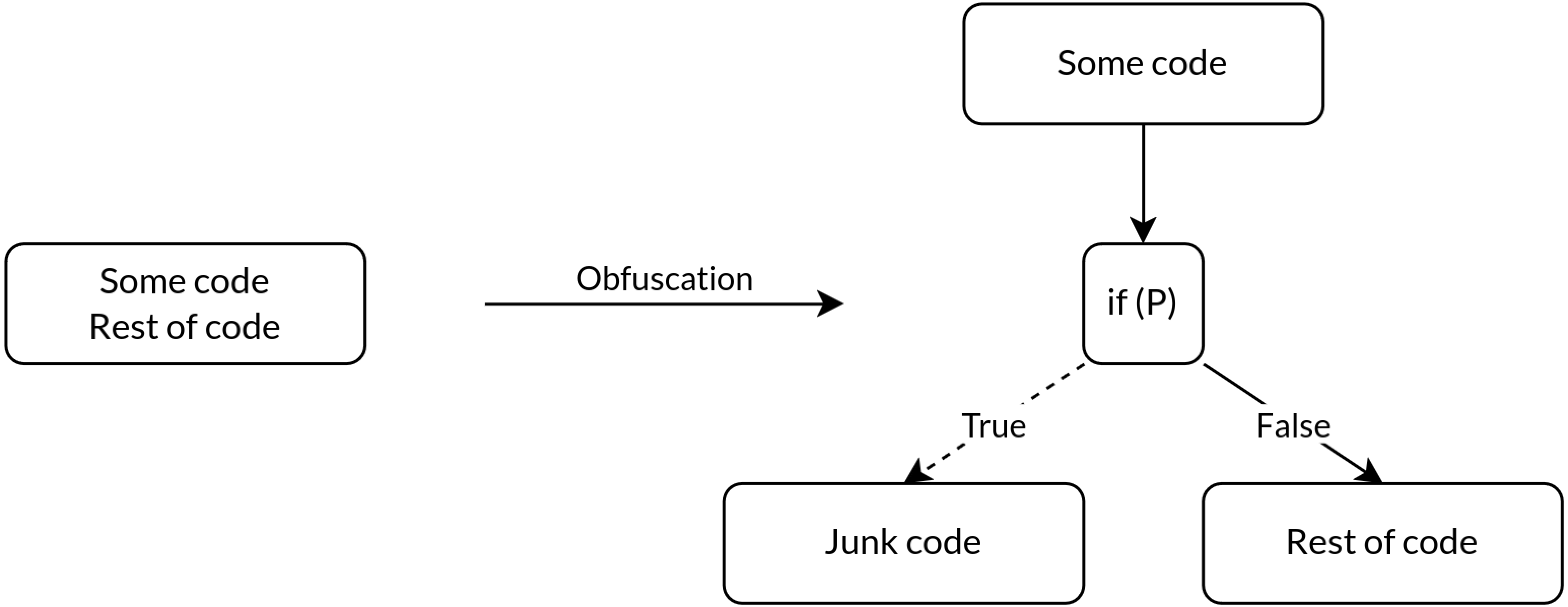
Opaque predicates

A conditional statement P whose truth value is known a priori.

$x^2 \geq 0$ is always true.

$7y^2 - 1 = x^2$ is always false.

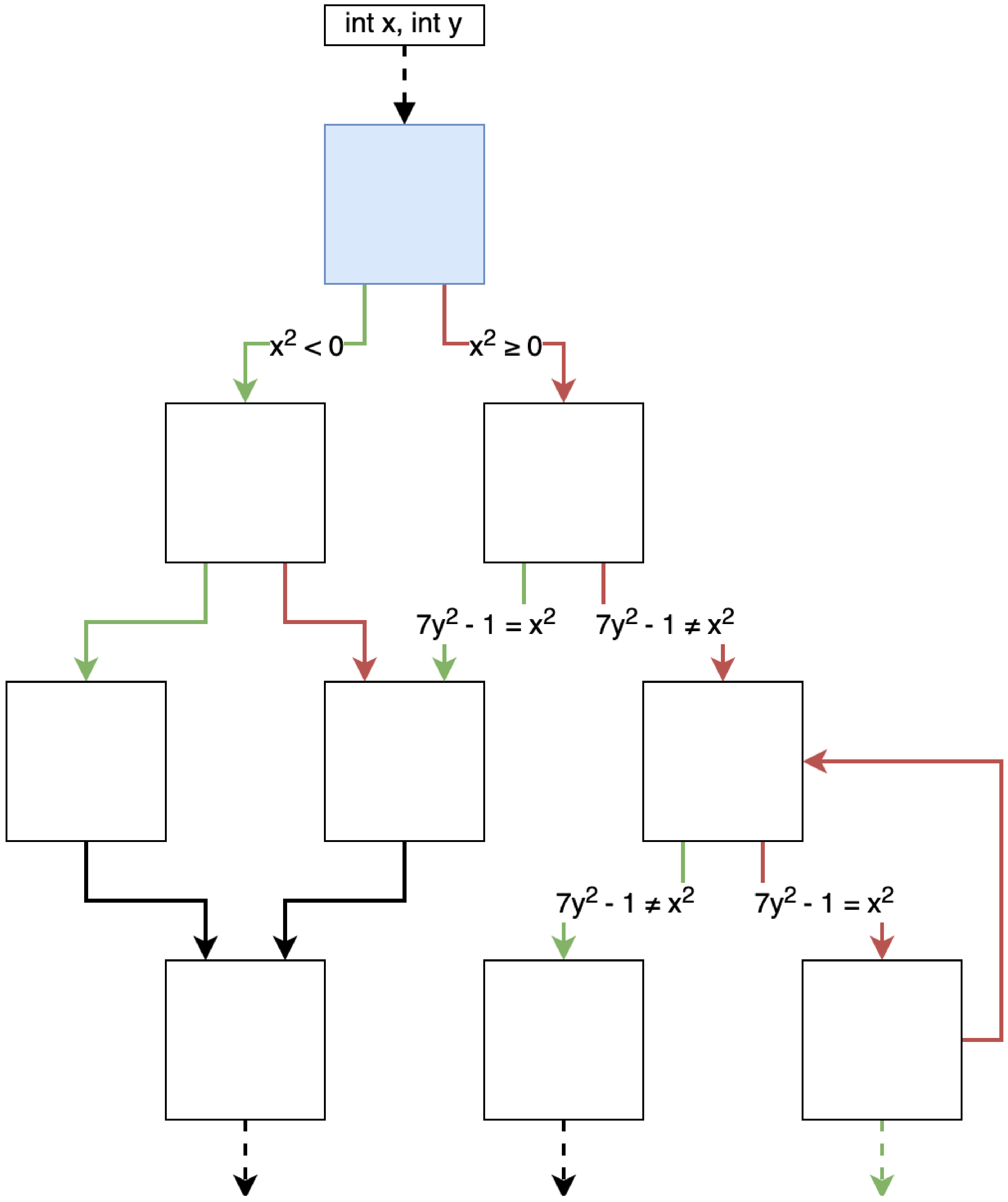

Some code
Rest of code
Obfuscation
Some code
if (P)
True
False
Rest of code
Junk code


Some code
Rest of code
Obfuscation
Some code
if (P)
True
False
Junk code
Rest of code

Detect (and patch) opaque predicates

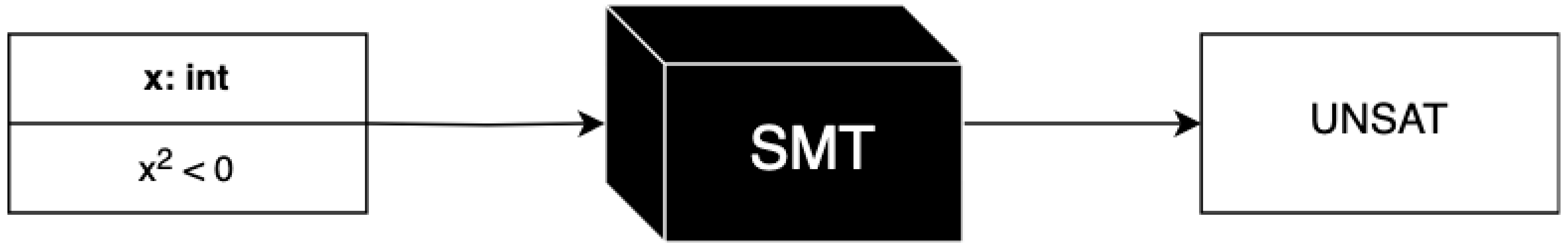
- Symbolically execute a basic block
- Extract the branching constraints
- Check if the constraints are either always true (or false)
- Patch it to continue execution at the only possible branch and remove (`NOP`) the unreachable branch

Visual example



Path constraint T branch

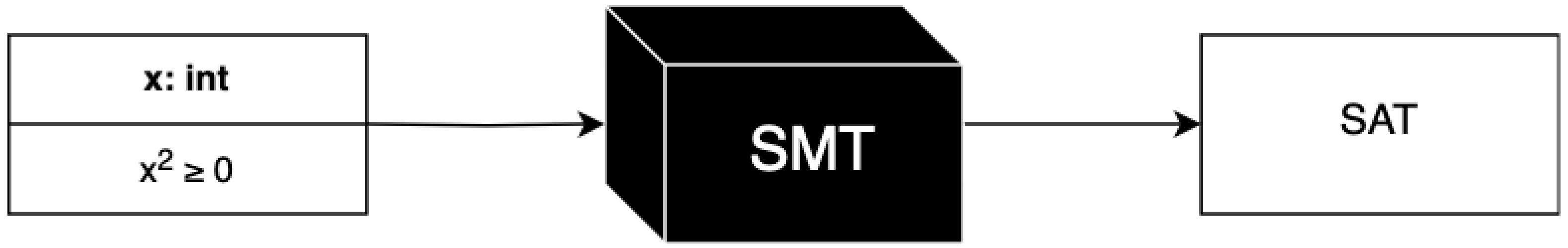
$$x^2 < 0$$



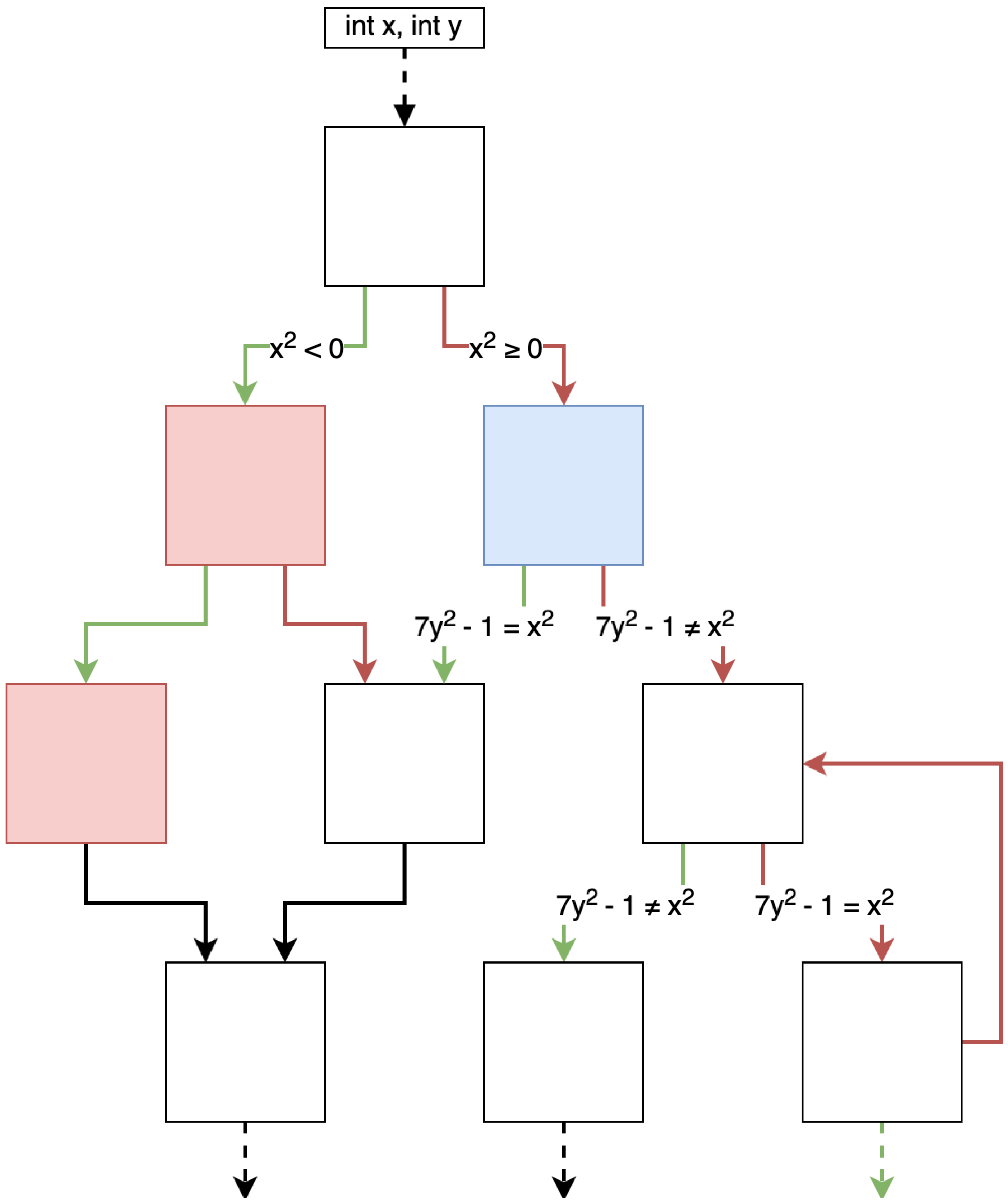
❌

Path constraint F branch

$$x^2 \geq 0$$

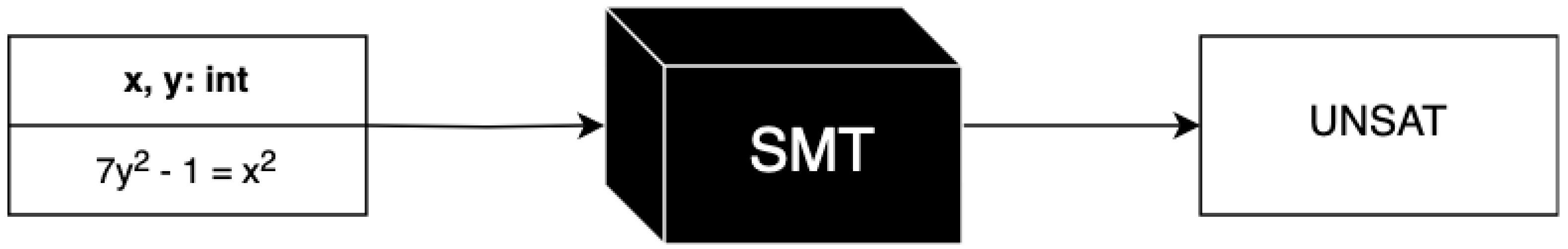


✅



Path constraint T branch

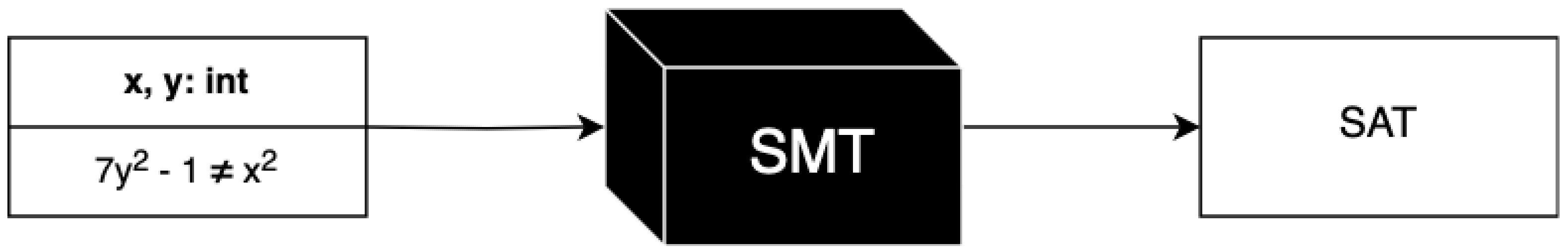
$$7y^2 - 1 = x^2$$



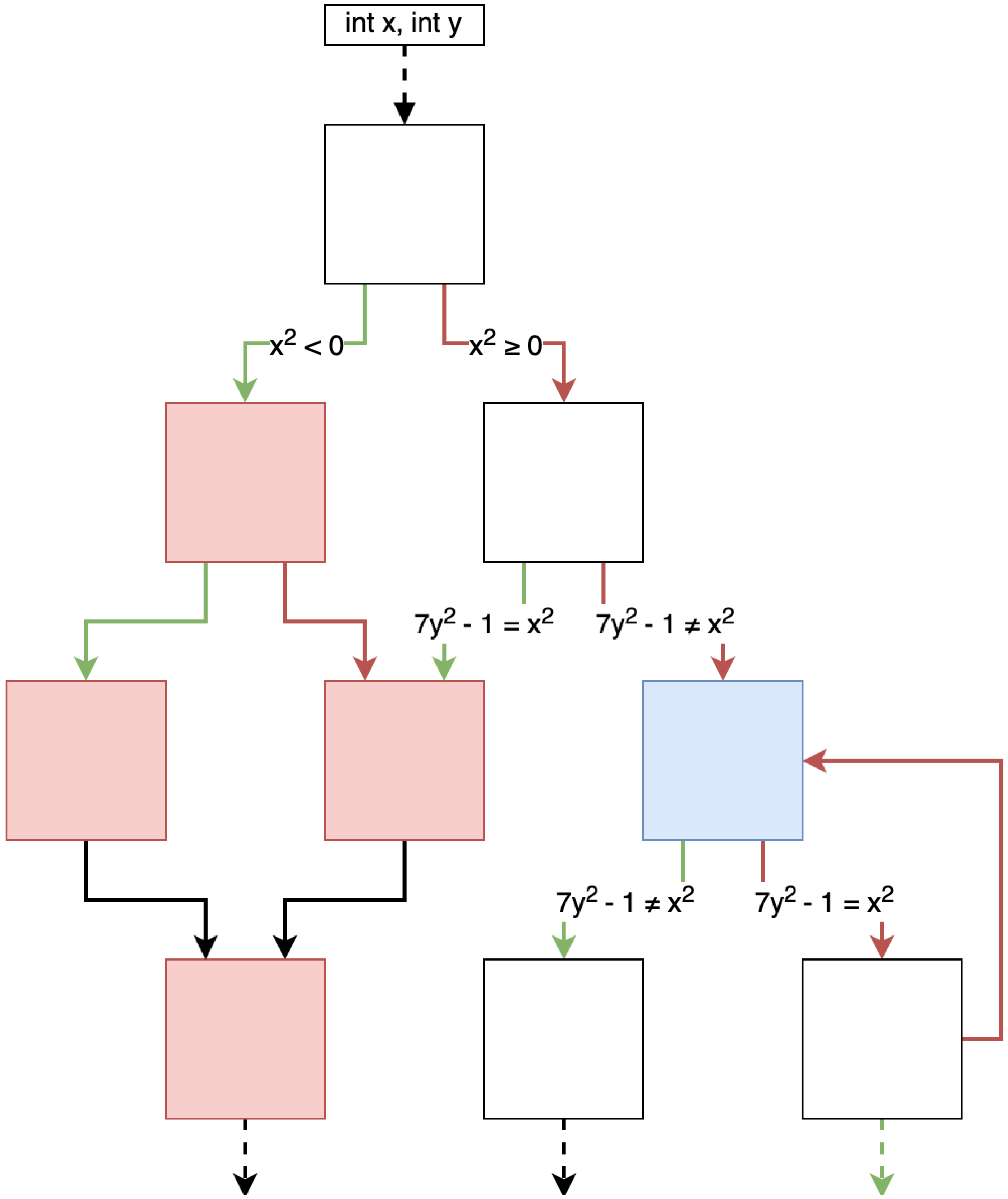
❌

Path constraint F branch

$$7y^2 - 1 \neq x^2$$

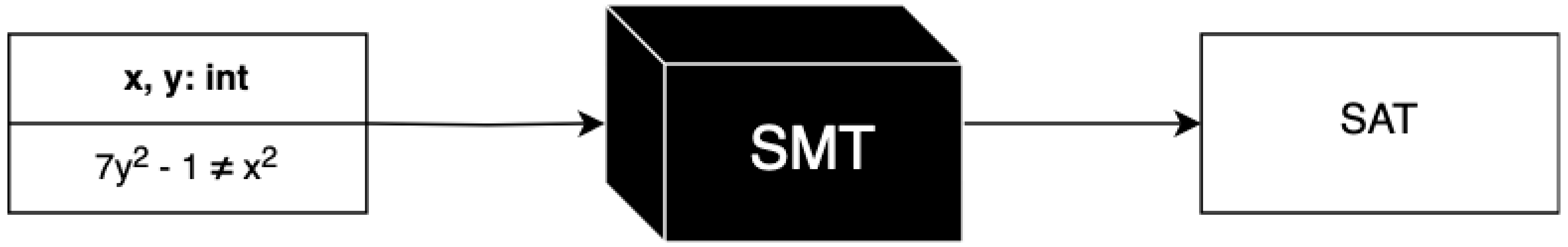


✅



Path constraint T branch

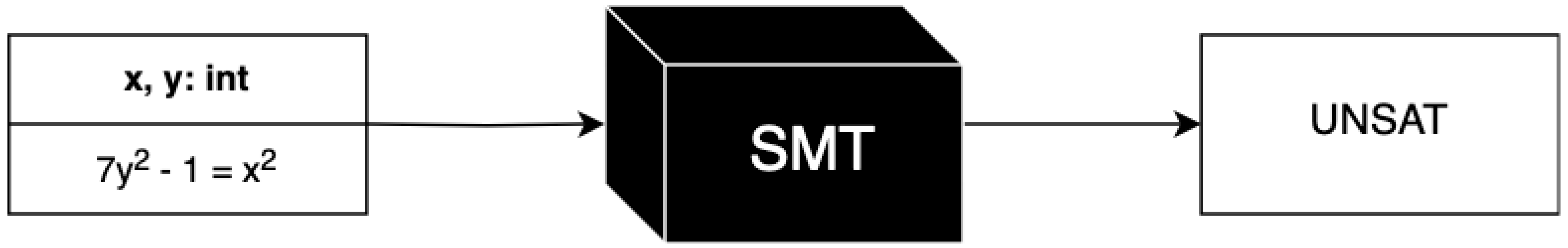
$$7y^2 - 1 \neq x^2$$



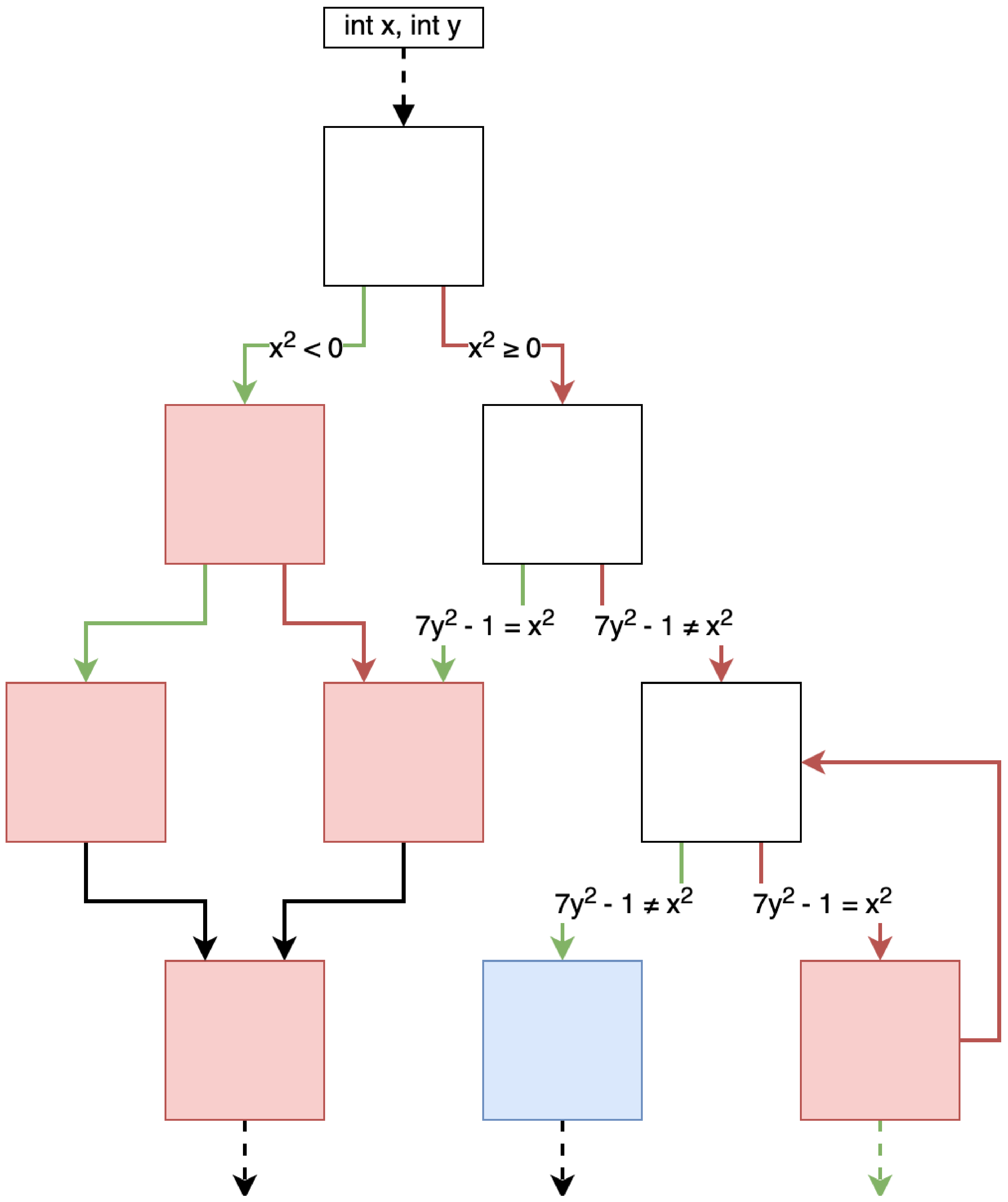
✅

Path constraint F branch

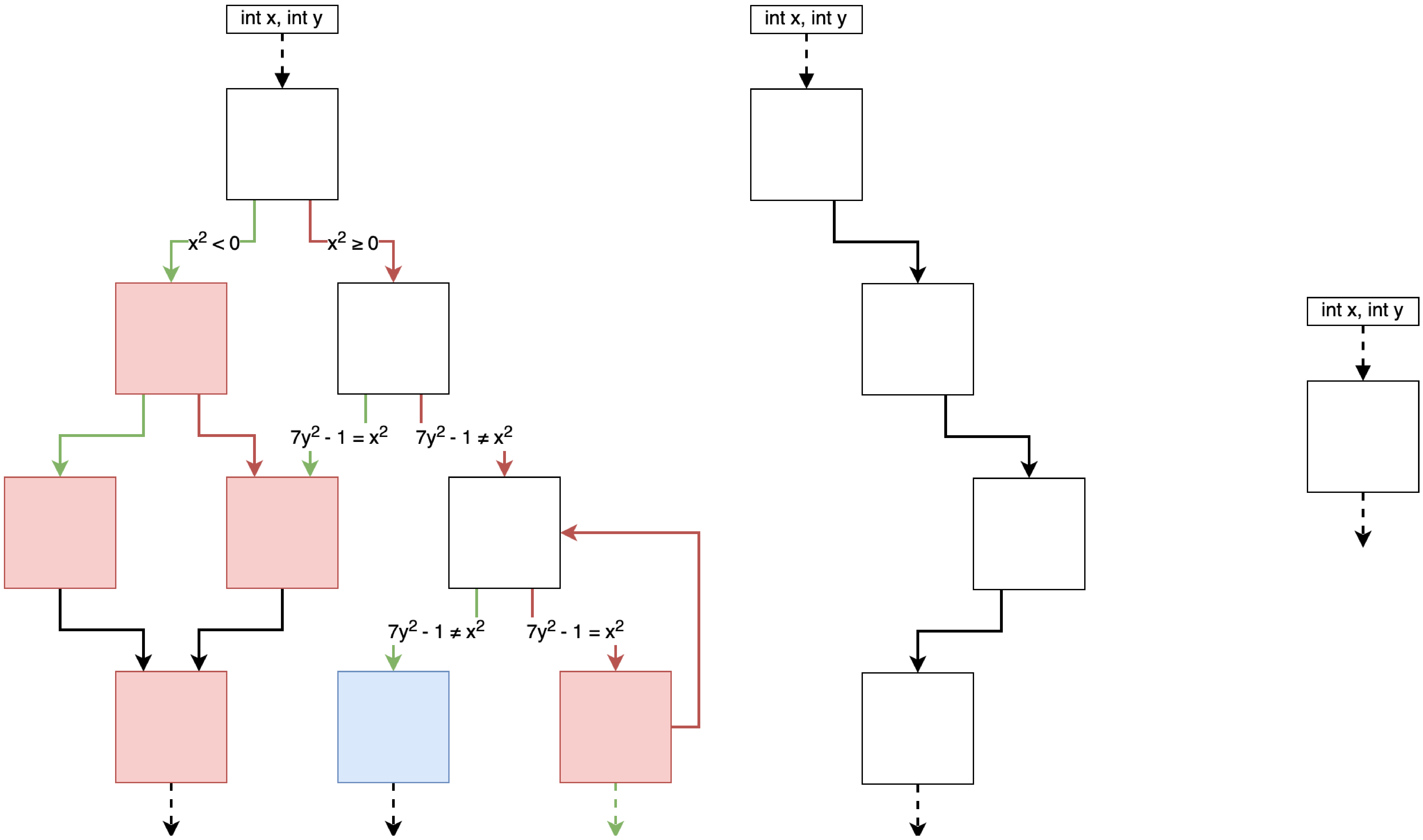
$$7y^2 - 1 = x^2$$



❌



int x, int y
x2 < 0
x2 ≥ 0
7y2 - 1 = x2
7y2 - 1 ≠ x2
7y2 - 1 ≠ x2
7y2 - 1 = x2



int x, int y
$x^2 < 0$
$x^2 \geq 0$
$7y^2 - 1 = x^2$
$7y^2 - 1 \neq x^2$
$7y^2 - 1 \neq x^2$
$7y^2 - 1 = x^2$
int x, int y
int x, int y

Example

Example

XTunnel @ APT28: ac3e087e43be67bdc674747c665b46c2

Example

XTunnel @ APT28: ac3e087e43be67bdc674747c665b46c2

Based on: https://github.com/mrphrazer/r2con2020_deobfuscation/blob/master/remove_opaque.py by Tim Blazytko (aka mrphrazer)

Demo

-

Malware deobfuscation through opaque predicates removal

Fuzzing

Increase code coverage

Code coverage

Measure of the degree to which the code of a program is executed when a set of inputs is run.

- Subroutines called
- Basic blocks reached
- Statements executed

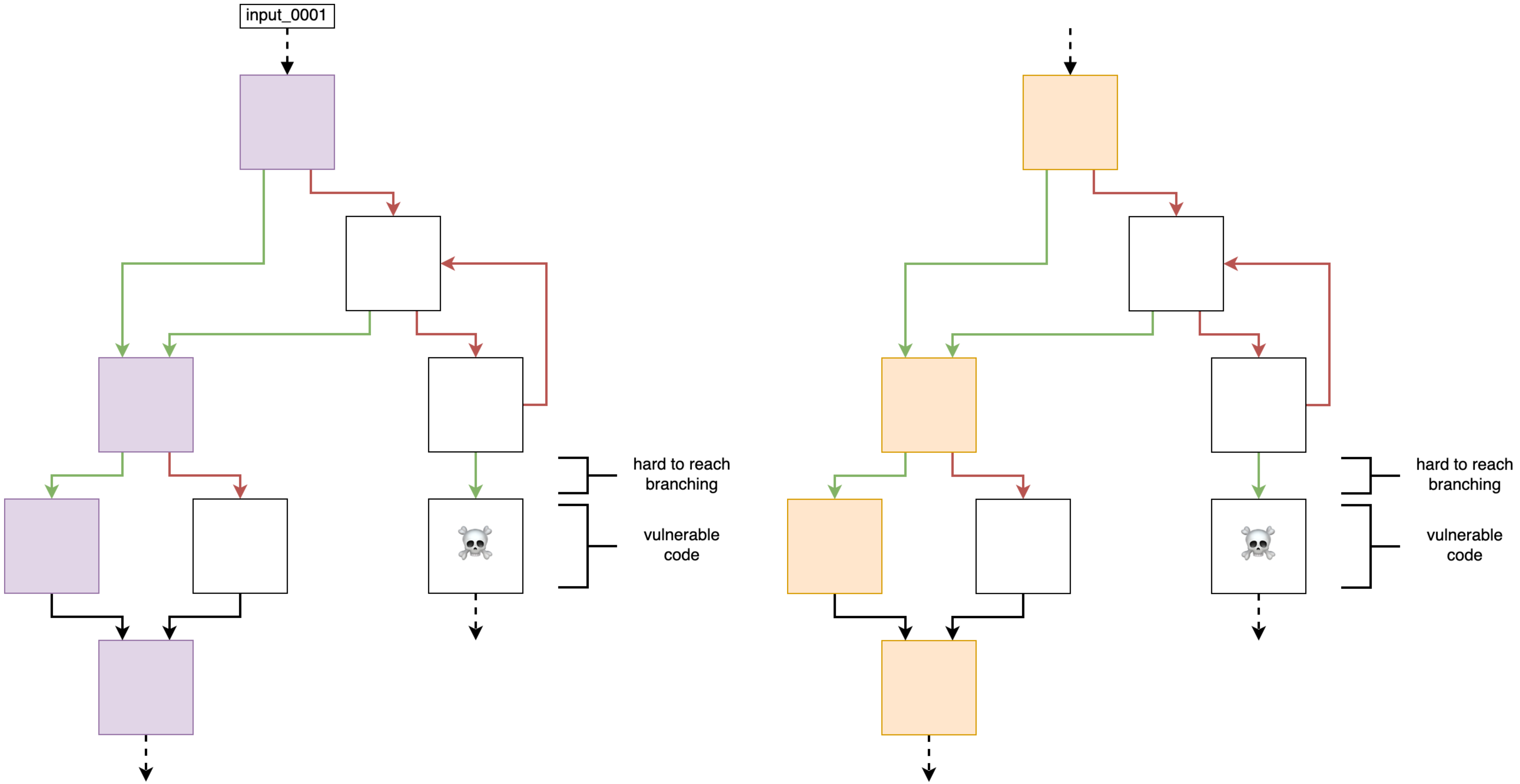
Higher code coverage → higher chance of hitting *interesting* (vulnerable) code

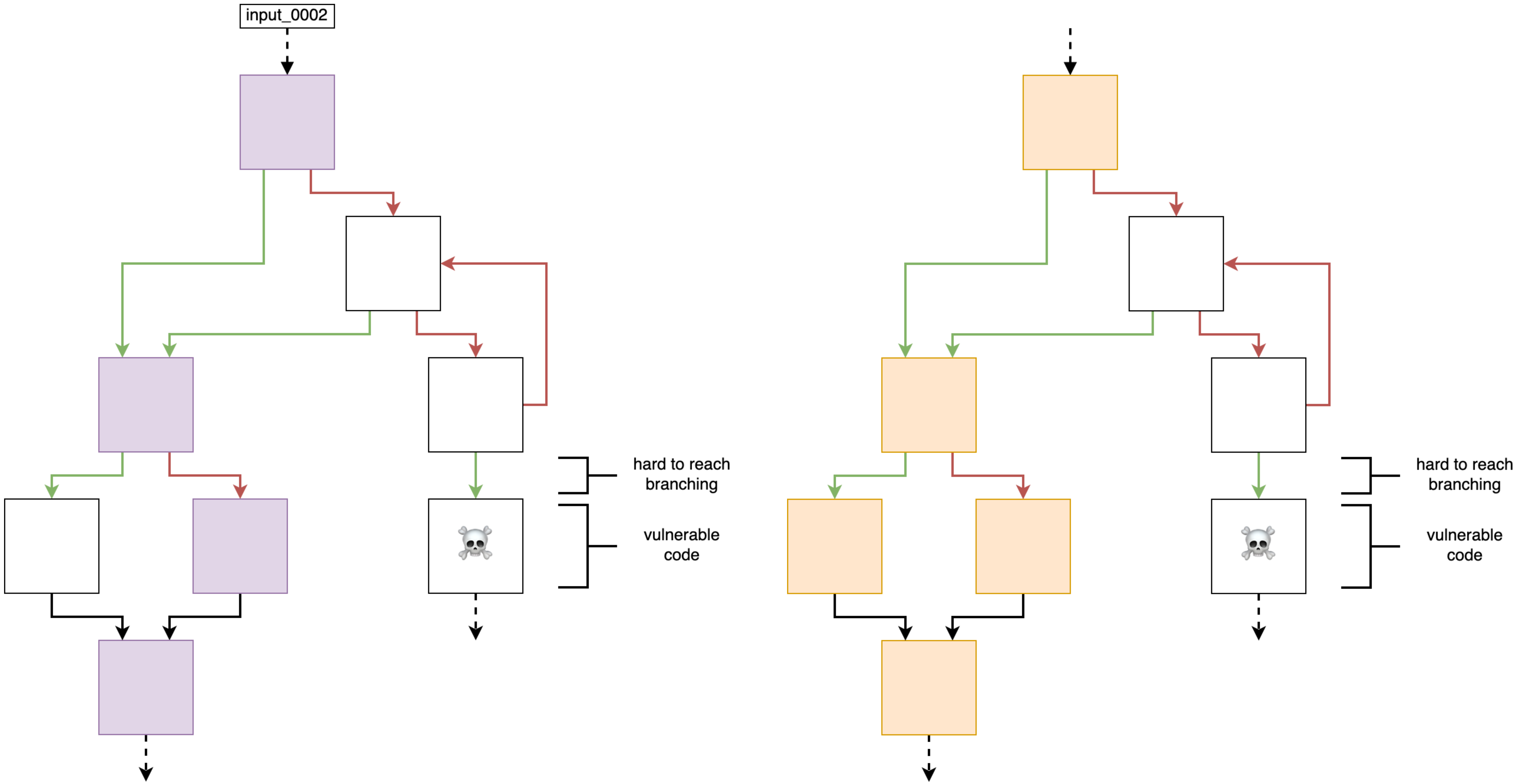
Increase code coverage

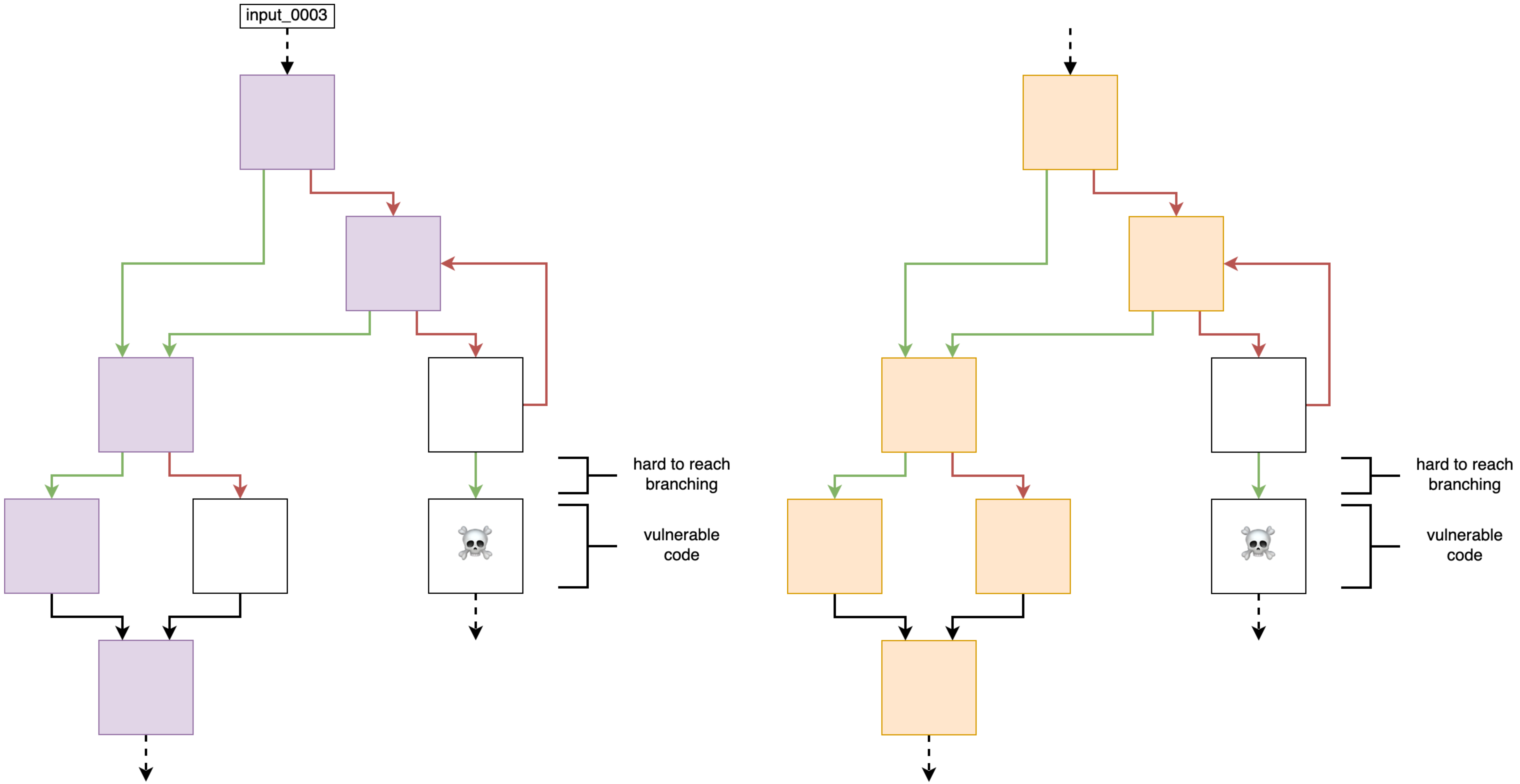
1. Start fuzzing your target with an initial seed/corpus
2. Use symbolic execution to generate inputs that trigger non-explored paths
3. Feed these new inputs into the fuzzer
4. Repeat

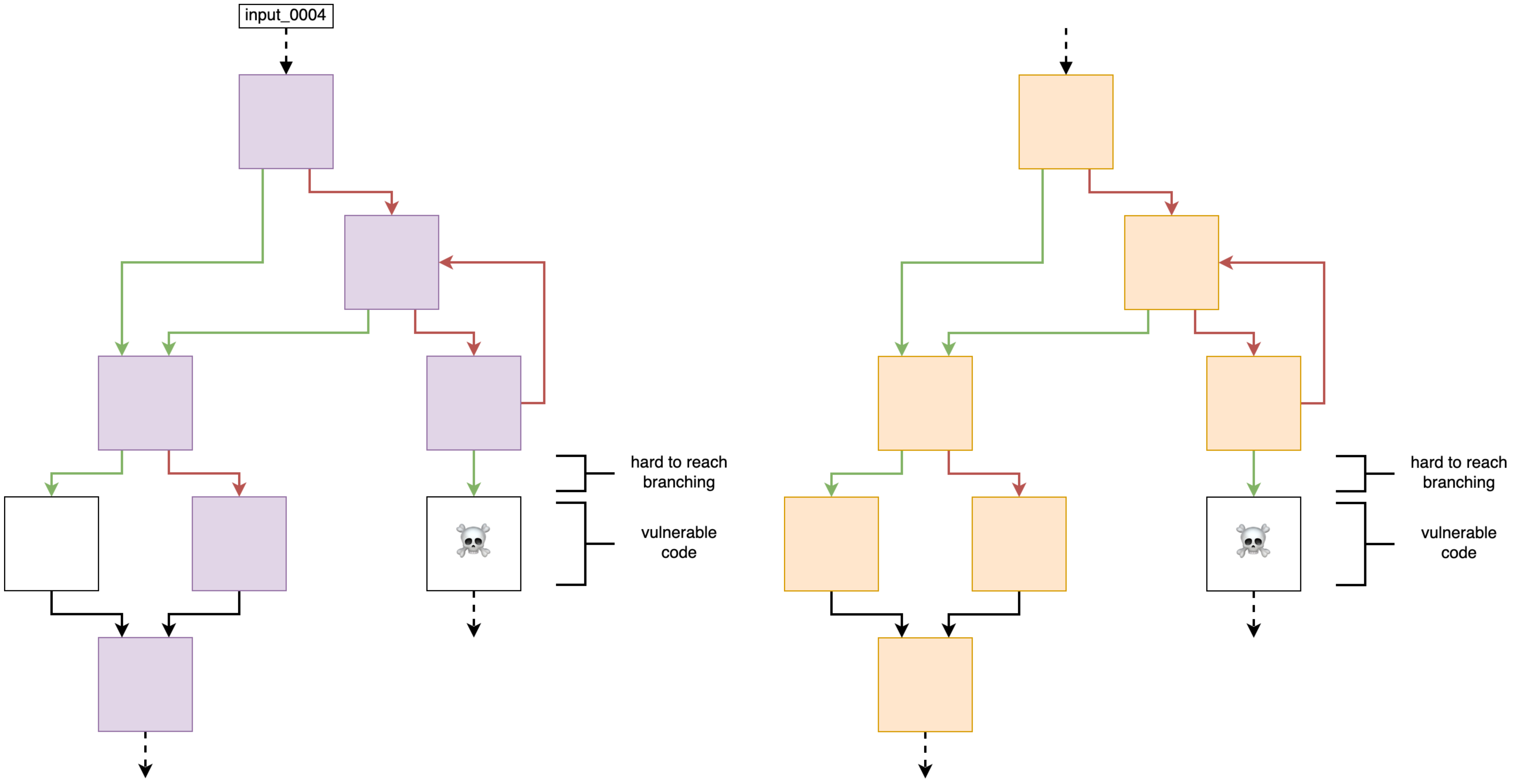
Visual example

Start fuzzing your target with an initial seed/corpus

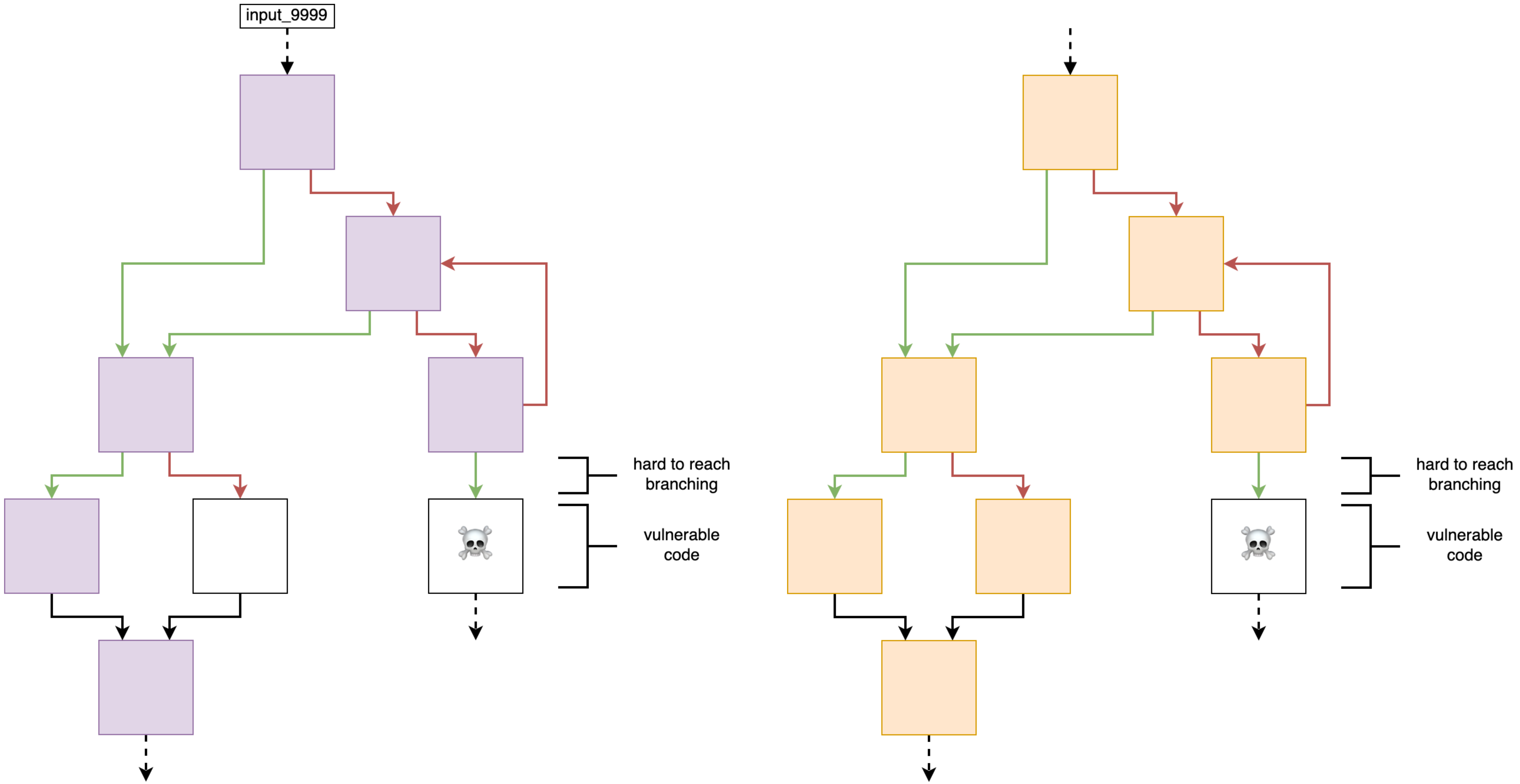

input_0001
hard to reach branching
vulnerable code
hard to reach branching
vulnerable code


input_0002
hard to reach branching
vulnerable code
hard to reach branching
vulnerable code

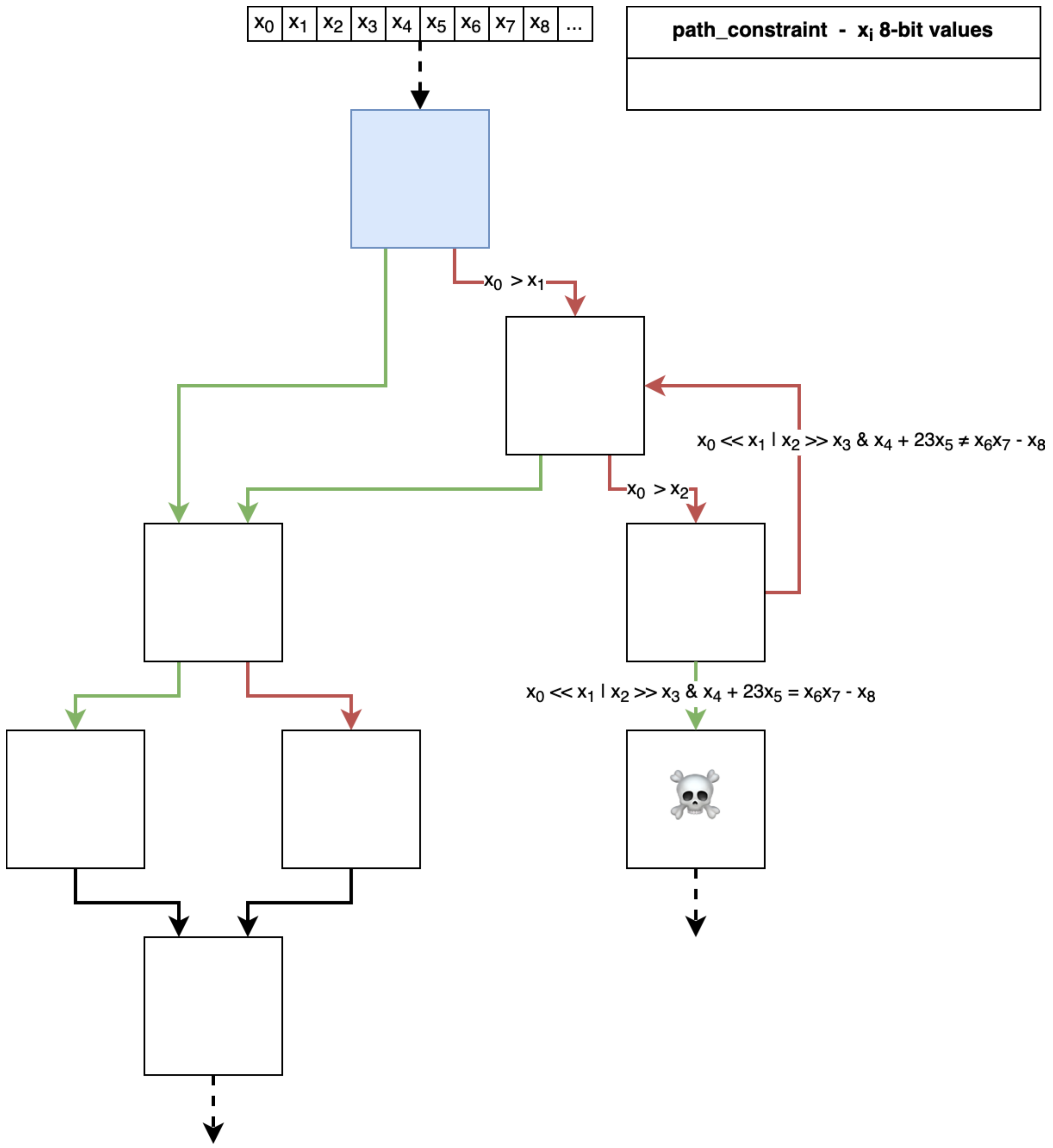

input_0003
hard to reach
branching
vulnerable
code
hard to reach
branching
vulnerable
code


input_0004
hard to reach branching
vulnerable code
hard to reach branching
vulnerable code

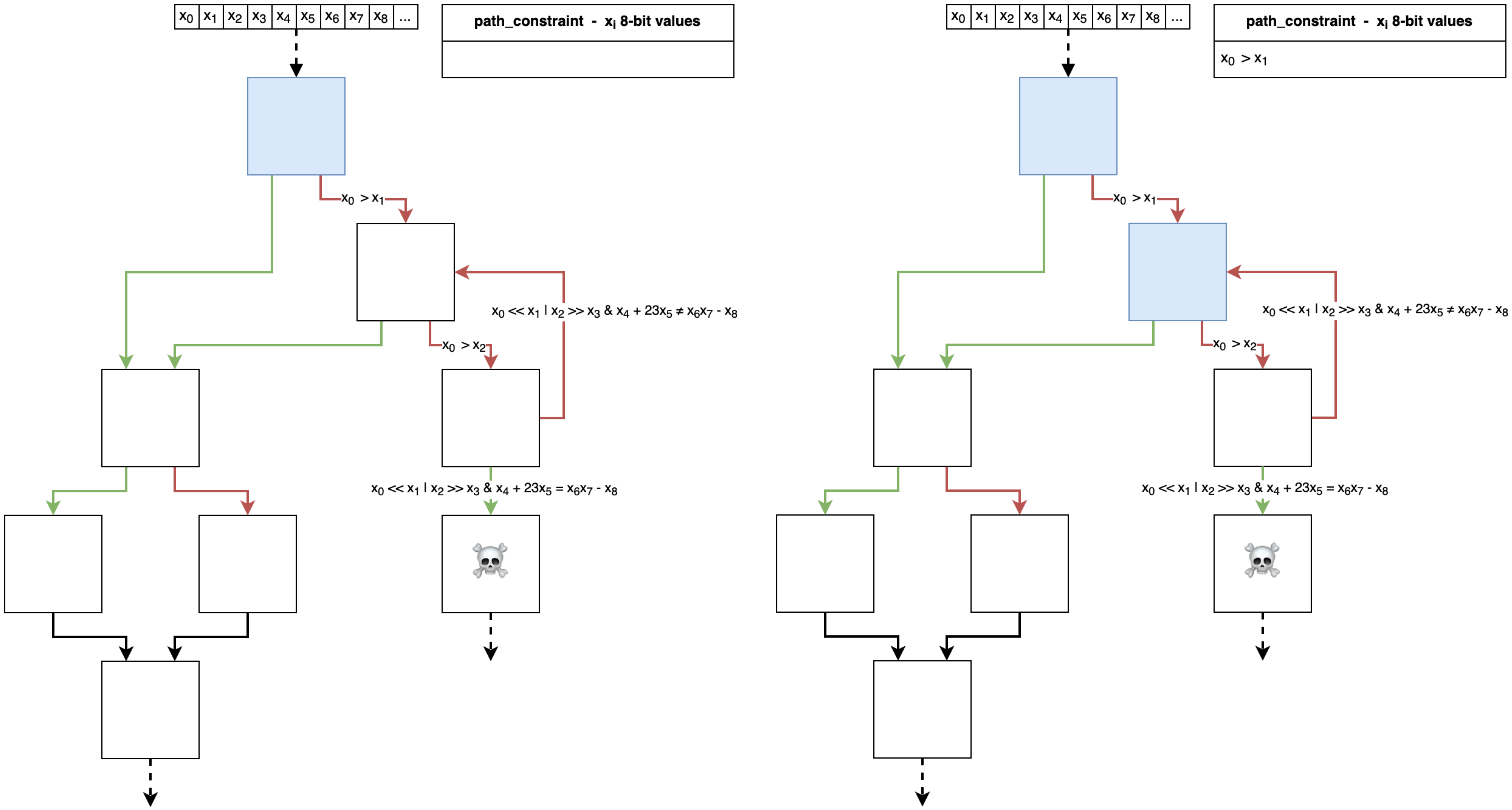
•••


input_9999
hard to reach branching
vulnerable code
hard to reach branching
vulnerable code

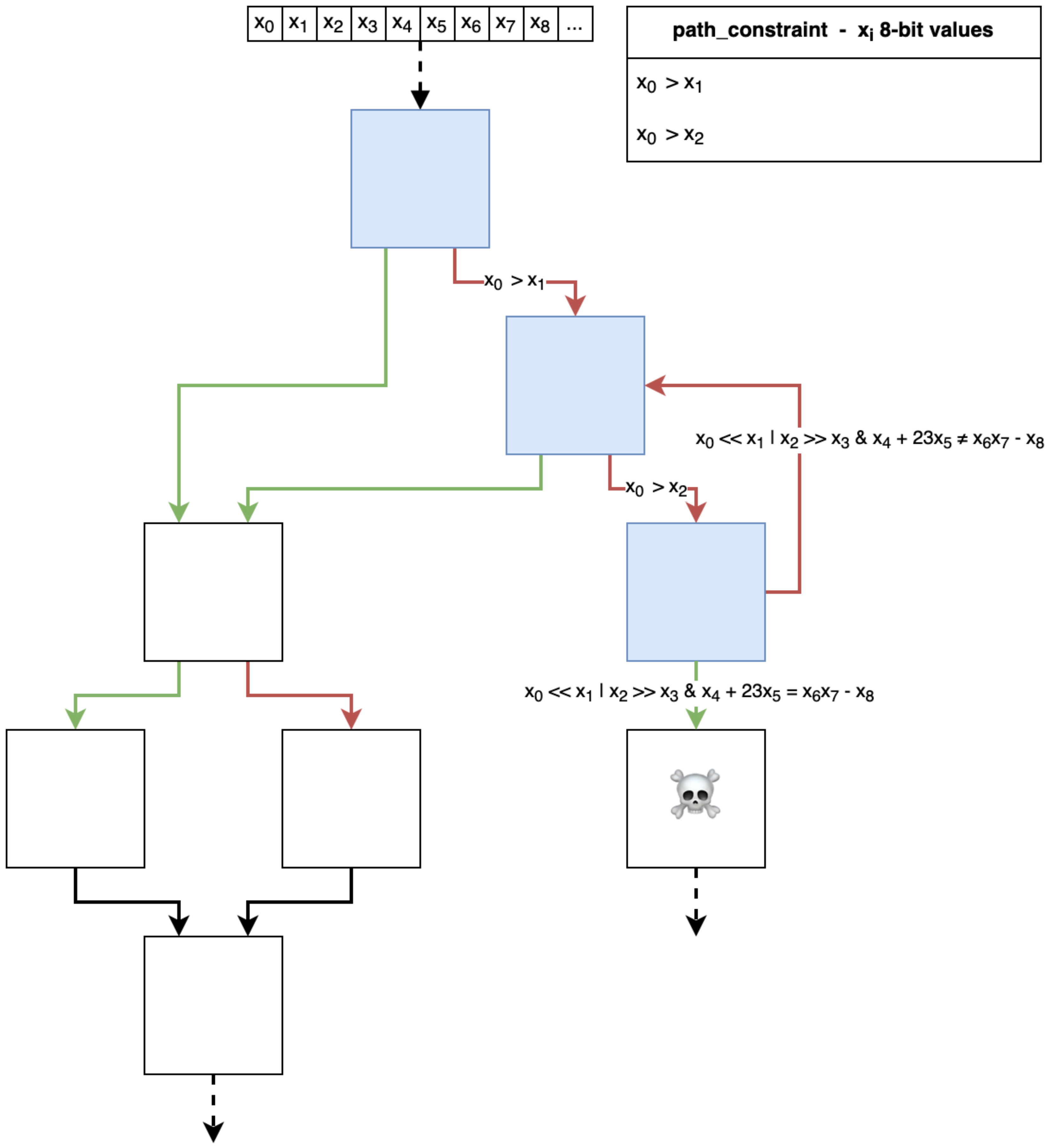
Use symbolic execution to generate inputs that trigger non-explored paths



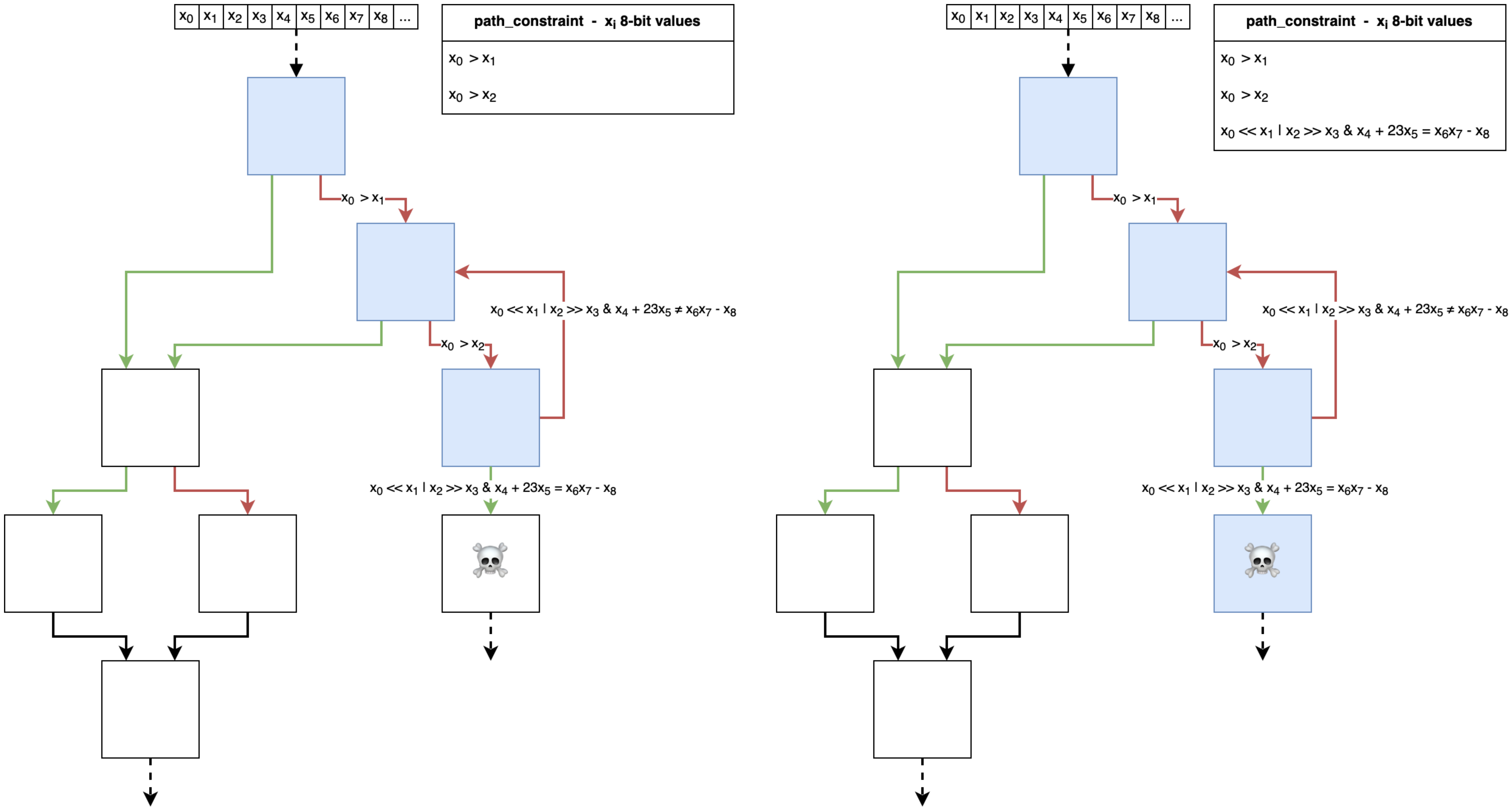
x0 x1 x2 x3 x4 x5 x6 x7 x8 ...
path_constraint - xi 8-bit values
x0 > x1
x0 << x1 | x2 >> x3 & x4 + 23x5 ≠ x6x7 - x8
x0 > x2
x0 << x1 | x2 >> x3 & x4 + 23x5 = x6x7 - x8



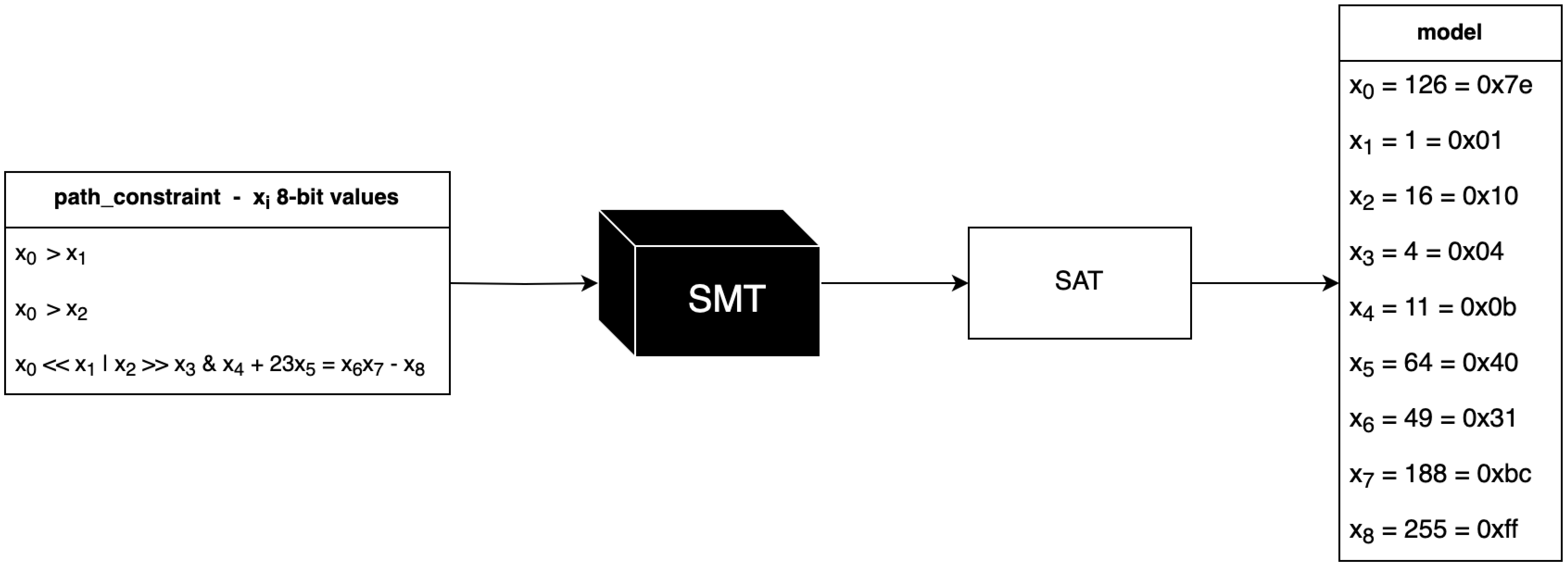
x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 ...
path_constraint - x_i 8-bit values
$x_0 > x_1$
$x_0 << x_1 \mid x_2 >> x_3 \;\&\; x_4 + 23x_5 \neq x_6x_7 - x_8$
$x_0 > x_2$
$x_0 << x_1 \mid x_2 >> x_3 \;\&\; x_4 + 23x_5 = x_6x_7 - x_8$
x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 ...
path_constraint - x_i 8-bit values
$x_0 > x_1$
$x_0 > x_1$
$x_0 << x_1 \mid x_2 >> x_3 \;\&\; x_4 + 23x_5 \neq x_6x_7 - x_8$
$x_0 > x_2$
$x_0 << x_1 \mid x_2 >> x_3 \;\&\; x_4 + 23x_5 = x_6x_7 - x_8$

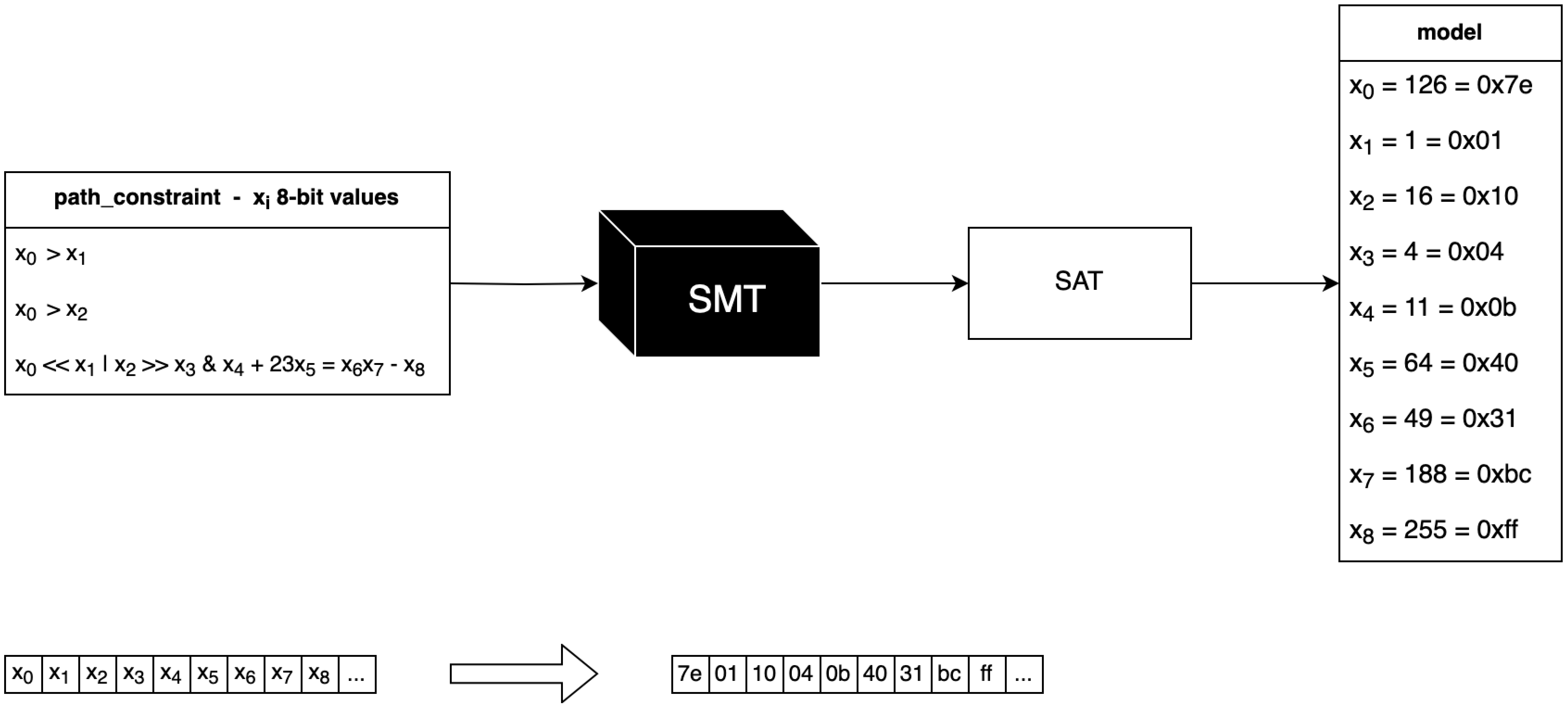


x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 ...
path_constraint - x_i 8-bit values
$x_0 > x_1$
$x_0 > x_2$
$x_0 > x_1$
$x_0 << x_1 \mid x_2 >> x_3 \,\&\, x_4 + 23x_5 \neq x_6x_7 - x_8$
$x_0 > x_2$
$x_0 << x_1 \mid x_2 >> x_3 \,\&\, x_4 + 23x_5 = x_6x_7 - x_8$
☠️



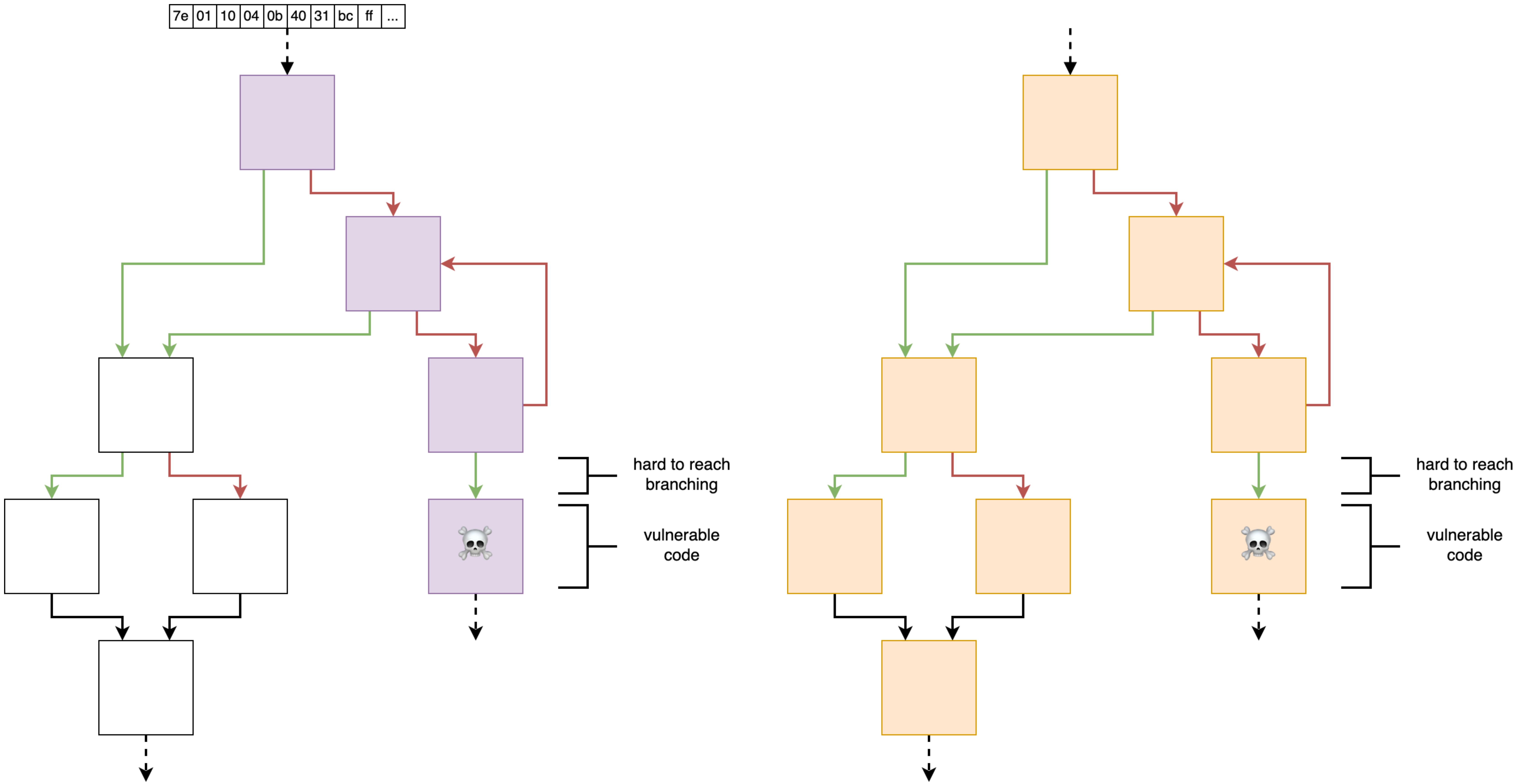
x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 ...
path_constraint - x_i 8-bit values
$x_0 > x_1$
$x_0 > x_2$
$x_0 > x_1$
$x_0 << x_1 \mid x_2 >> x_3 \,\&\, x_4 + 23x_5 \neq x_6x_7 - x_8$
$x_0 > x_2$
$x_0 << x_1 \mid x_2 >> x_3 \,\&\, x_4 + 23x_5 = x_6x_7 - x_8$
x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 ...
path_constraint - x_i 8-bit values
$x_0 > x_1$
$x_0 > x_2$
$x_0 << x_1 \mid x_2 >> x_3 \,\&\, x_4 + 23x_5 = x_6x_7 - x_8$
$x_0 > x_1$
$x_0 << x_1 \mid x_2 >> x_3 \,\&\, x_4 + 23x_5 \neq x_6x_7 - x_8$
$x_0 > x_2$
$x_0 << x_1 \mid x_2 >> x_3 \,\&\, x_4 + 23x_5 = x_6x_7 - x_8$


path_constraint - x_i 8-bit values
$x_0 > x_1$
$x_0 > x_2$
$x_0 << x_1 \mid x_2 >> x_3 \& x_4 + 23x_5 = x_6x_7 - x_8$
SMT
SAT
model
$x_0 = 126 = 0x7e$
$x_1 = 1 = 0x01$
$x_2 = 16 = 0x10$
$x_3 = 4 = 0x04$
$x_4 = 11 = 0x0b$
$x_5 = 64 = 0x40$
$x_6 = 49 = 0x31$
$x_7 = 188 = 0xbc$
$x_8 = 255 = 0xff$



path_constraint - x_i 8-bit values
$x_0 > x_1$
$x_0 > x_2$
$x_0 << x_1 \mid x_2 >> x_3 \ \& \ x_4 + 23x_5 = x_6x_7 - x_8$
SMT
SAT
model
x_0 = 126 = 0x7e
x_1 = 1 = 0x01
x_2 = 16 = 0x10
x_3 = 4 = 0x04
x_4 = 11 = 0x0b
x_5 = 64 = 0x40
x_6 = 49 = 0x31
x_7 = 188 = 0xbc
x_8 = 255 = 0xff
x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 ...
7e 01 10 04 0b 40 31 bc ff ...

Feed these new inputs into the fuzzer


7e 01 10 04 0b 40 31 bc ff ...
hard to reach branching
vulnerable code
hard to reach branching
vulnerable code

Repeat ↺

Example

Example

Trigger a hard-to-reach division by zero

Example

Trigger a hard-to-reach division by zero

Based on: *Fuzzing combined with symbolic execution: a demonstration on SymCC and AFL* by AdaLogics

```c
#include <stdio.h>
#include <stdlib.h>
#include <string.h>

int stuff(char *data, long fsize) {
    for (size_t i = 0; i < fsize; i++) {
        if (data[i] == ('F' ^ i)):
            return i+1;
    }

    if (*(int*)data != 0xfa1afe1):
        return 0;

    return (int)(0x1337/(fsize - 10)); // <== TRIGGER DIV BY 0 HERE
}
```

I made a (dumb) SymCC fork (SymCC++) to make it work with AFL++

https://github.com/arnaugamez/symccpp

Demo

-

Run AFL++ with SymCC++ to reach the division by zero

Limitations

Limitations

And some ideas to overcome them

- Path explosion: the number of control-flow paths grows exponentially ($\rightarrow\ \infty$ for unbounded loops)
    - Manual location of interesting code
    - Concolic (**con**crete + symb**olic**) execution

- Path explosion: the number of control-flow paths grows exponentially ($\rightarrow \infty$ for unbounded loops)
  - Manual location of interesting code
  - Concolic (**con**crete + symb**olic**) execution

- Support for syscalls, standard C library functions, etc.:
  - Same as with any emulator: *hook 'em all*

- Path explosion: the number of control-flow paths grows exponentially ($\rightarrow\ \infty$ for unbounded loops)
  - Manual location of interesting code
  - Concolic (**con**crete + symb**olic**) execution

- Support for syscalls, standard C library functions, etc.:
  - Same as with any emulator: *hook 'em all*

- Limits of SMT solvers (e.g. due to high algebraic complexity through MBA transformations):
  - Program synthesis
  - Math™
  - Imagination

Training

Advanced Software Protection - Attacks and Defense

For Security Researchers and Developers

- Public / Private
- In-person / Remote
- 4 days (flexible)
- Details: https://furalabs.com/trainings

Upcoming

- February 20-23, 2024 @ Ringzer0 - Austin, TX (USA)

Thank you

Q&A