TYPESHIELD: Precise Protection of Forward Indirect Calls in Binaries

Your N. Here
Your Institution

Second Name
Second Institution

Abstract

High security, high performance and high availability applications such as the Firefox and Chrome web browsers are implemented in C/C++ for modularity, performance and compatibility to name just a few reasons. Virtual functions, which facilitate late binding, are a key ingredient in facilitating run-time polymorphism in C++ because it allows and object to use general (its own) or specific functions (inherited) contained in the class hierarchy. However, because of the specific implementation of late binding, which performs no verification in order to check where an indirect call site (virtual object dispatch through virtual pointers (*vptrs*)) is allow to call inside the class hierarchy, this opens a large attack surface which was successfully exploited by the COOP attack. Since manipulation (changing or inserting new *vptrs*) violates the programmer initial pointer semantics and allows an attacker to redirect the control flow of the program as he desires, *vptrs* corruption has serious security consequences similar to those of other data-only corruption vulnerabilities. Despite the alarmingly high number of *vptr* corruption vulnerabilities, the *vptr* corruption problem has not been sufficiently addressed by the researchers.

In this paper, we present *TypeShild*, a run-time *vptr* corruption detection tool. It is based on executable instrumentation at load time and uses a novel run-time type and function parameter counter technique in order to overcome the limitations of current approaches and efficiently verify dynamic dispatching during run-time. In particular, *TypeShild* can be automatically and easily used in conjunction with legacy applications or where source code is missing. It achieves higher caller/caller matching (precision) and with reasonable run-time overhead. We have applied *TypeShild* to real life software such as web servers and FTP servers and were able to efficiently and with low performance overhead protect this applications from *vptr* corruptions vulnerabilities. Our evaluation shows that our target reduction schema achieves an additional reduction of the possible call targets per call-site of up to 20% with an overall reduction of about 9% when comparing to other parameter count based approaches.

1 Introduction

Motivation. Control-Flow Integrity (CFI) [4, 5] is one of the most used techniques to secure program execution flows against advanced Code-Reuse Attacks (CRAs). Advanced CRAs such as the recently published COOP [40] and its extensions [17] or the attacks described by the Control Flow Bending paper [14] are able to bypass most traditional CFI solutions, as they focus on indirect call-sites, which are not as easy to decide at compile time.

Problem. This is a problem for applications written in C++, as one of its principle is inheritance and virtual functions. The concept of virtual functions allows the programmer to overwrite a virtual function of the base-class with his own implementation. While this allows for much more flexible code, this flexibility is the reason COOP actually works. The problem is that in order to implement virtual functions, the compiler needs to generate a table of all virtual functions for each class containing them and provide each instantiation of such a class with a pointer to said table. COOP now leverages a memory corruption to inject their own object with a fake virtual pointer, which basically gives him control over the whole program, while the control flow still looks genuine, as no code was replaced.

Current solutions. There exist several source code based solutions that either insert run-time checks during the compilation of the program like SafeDispatch [21], ShrinkWrap [20] or IFCC/VTV [42], which is the solution it is based on. Others modify and reorder the contents of the virtual table as their main aspect like the paper by Bounov et al. [10]. While the recently published Redactor++ [17] implements a combination of those ideas.

While this might seem that only C++ is vulnerable, while C is safe, this notion is wrong, as the Control Flow Bending paper [14] proposes attacks on nginx leveraging global function pointers, which are used to provide configurable behavior.

As previously mentioned, there exist many solutions when one tries to tackle this problem while access to the application in question is provided. However, when we are faced with proprietary third party binaries, which are provided as is and without the actual source code, the number of tools that can protect against COOP or similar attacks is rather low.

Lack. TypeArmor [44] is such a tool that implements a fine grained forward edge CFI solution for binaries. It calculates invariants for call-targets and indirect call-sites based on the number of parameters they use by leveraging static analysis of the binary, which then is patched to enforce those invariants during run-time. However, as of today we are not able to access the source code of TypeArmor, which is why we implement our own approximation of the tool.

The main shortcoming of TypeArmor is that even with high precision in the classification of call-targets and call-sites, one cannot exclude calltargets with lower parameter number from call-sites, for one due compatibility and also due to variadic functions, which are a special case in themselves. This basically means that when a call-site prepares 6 parameters, it is able to call all address taken functions.

We implemented TYPESHIELD to show a possible remedy of this problem by introducing parameter types into the classification of call-sites and call-targets.

Our Idea. In this paper, we present TYPESHIELD, a runtime illegitimate forward calls detection tool that can be seamlessly integrated with large scale applications such as web servers. It takes the binary of an program as input and it can automatically instrument the binary in order to detect illegitimate indirect calls at runtime.

Goal of This Paper. The aim of this paper is twofold. First, implement our own indirect forward edge classification schema in order to fix some of the shortcomings of previous state-of-the-art binary based approaches used to mitigate indirect forward edge based advanced code reuse attacks such as COOP [40]. Second, we compared our tool with TypeArmor [44].

Contributions. In summary, we make the following contributions:

- We designed and implemented a call-site and call-target classification schema that is based on the wideness of parameters alone. We implemented configurable reaching and liveness analysis algorithms that operate on the full set of general purpose integer registers of a x86-64 CPU and evaluated various path merge operators. Although the basic idea of our aproach to rely only on the wideness of a type is rather simple, we still achieved a reduction of up to 20% less call-sites per call-target with an overall of about 9% when compared to our implementation of a parameter count based matching schema.
- We implemented an approximation of the matching schema employed by TypeArmor proposed by [44], because we had no access to their sour-cecode and could achieve similar results regarding parameter matching, partially verifying their results.
- give one more here.

Outline. The rest of this paper is organized as follows. § 2 explains forbidden forward indirect calls issues and their security implications. § 3 contains a high level overview of TYPESHIELD. § 4 describes the theory used and decisions made during the design of TYPESHIELD. § 5 briefly presents the implementation details of TYPESHIELD. § 6 evaluates several properties of TYPESHIELD and § 7 surveys related work, respectively. § 9 highlights future research venues while § 8 contains the discussion, respectively. Finally, § 10 concludes this paper.

2 C++ Forbidden Forward Calls Exposed

Polymorphism in C++. Polymorphism along inheritance and encapsulation are the most used modern object-oriented concepts in C++. Polymorphism in C++ allows to access different types of objects through a common base class. A pointer of the type of the base object can be used to point to object(s) which are derived from the base class. In C++ there are several types of polymorphism: *a)* compile-time (or static, usually is implemented with templates), *b)* runtime (dynamic, is implemented with inheritance and virtual functions), *c)* ad-hoc (e.g., if the range of actual types that can be used is finite and the combinations must be individually specified prior to use), and *d)* parametric (e.g., if code is written without mention of any specific type and thus can be used transparently with any number of new types it is called parametric polymorphism). The first two are implemented through early and late binding, respectively. In C++, overloading concepts fall under the category of *c)* and Virtual functions; templates or parametric classes fall under the category of pure polymorphism. C++ provides polymorphism through: *i)* virtual functions, *ii)* function name overloading, and *iii)* operator overloading. In this paper, we will be concerned with dynamic polymorphism—based on virtual functions (10.3 and 11.5 in ISO/IEC N3690 [22])—because these can be exploited to call: *x)* illegitimate vTable entries not/contained in the class hierarchy by varying or not the number of parameters and types, *y)* legitimate vTable entries not/contained in the class hierarchy by varying or not the number of parameters and types, *z)* fake vTables entries not contained in the class hierarchy by varying or not the number of parameters and types. By legitimate and illegitimate vTable entries we mean those vTable entries which for a single indirect call site lie in the vTable hierarchy. More precisely, a vTable entry is legitimate for a call site if from the call site to the vTable containing the entry there is an inheritance path (see [20]). Virtual functions have several uses and issues associated, but for the scope of this paper we will look at the indirect call sites which are exploiter by calling illegitimate vTable entries (functions) with varying number and type of parameters, *x)*. More precisely, *1)* load-time enforcement: as calling each indirect call site (callee) requires a fix number of parameters which are passed each time the caller is calling, we enforce a fine-grained CFI policy by statically determining the number and types of all function parameter that belong to an indirect call site. *2)* run-time verification: as checking during run-time legitimate from illegitimate indirect caller/callee pairs requires parameter type (along parameter number), we check during run-rime before each indirect call-site if the caller matches to the callee based on the previously

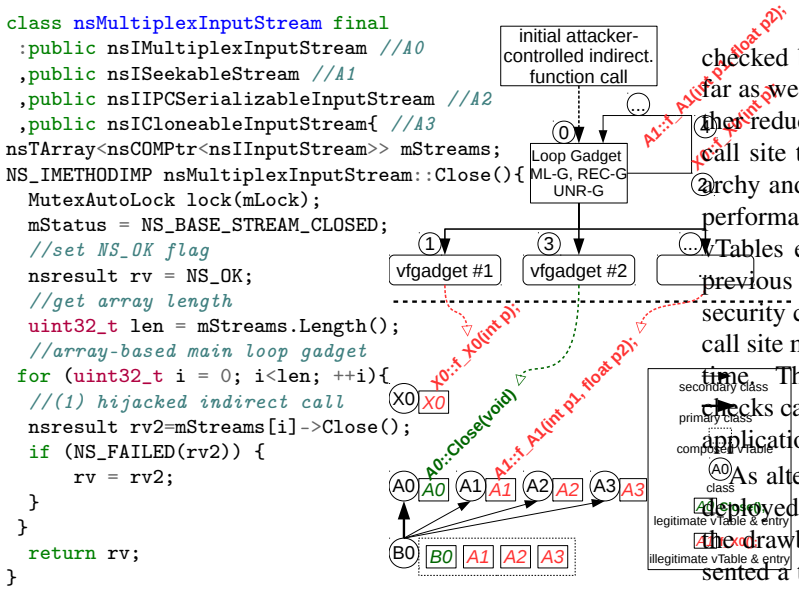

Figure 1: Code example used to illustrate how a COOP loop gadget works.

added checks.

Figure 1 depicts a C++ code example where it is illustrated how a COOP loop gadget (ML-G, REC-G, UNR-G, see [17]) works. (1) can be exploited in several ways, see *x), y) and z)*. The indirect call site (line 17) can be exploited to call by passing a varying number of parameters and types on each object contained in the array a different vTable entry contained in the: *1)* class hierarchy (overall, whole program), *2)* class hierarchy (partial, only legitimate for this call site), *3)* vTable hierarchy (overall, whole program), *4)* vTable hierarchy (partial, only legitimate for this call site), *5)* vTable hierarchy and/or class hierarchy (partial, only legitimate for this call site), and *6)* vTable hierarchy and/or class hierarchy (overall, whole program). There is no language semantics—such as cast checks—in C++ for vCall sites dispatch checking and as consequence the loop gadget indicated in Figure 1 can basically call all around in the class and vTable hierarchy by not being constrained by any build in check during runtime. The attacker corrupts an indirect function call, ①, next she invokes gadgets, ①, ③, through the calls, ②, ④, contained in the loop. As it can be observed in in Figure 1 she can invoke from the same call site legitimate functions residing in the vTable inheritance path (this type of information is usually very hard to recuperate from executables) for this particular call site, indicated with green color vTable entries. However, a real COOP attack invokes illegitimate vTable entries residing in the whole initial program hierarchy (or the extended one) with less or no relationship to the initial call site, indicated with red color vTable entries.

Checking Indirect Forward-Edge Calls in Practice. As far as we know, there is only the IFCC/VTV [42] tools (up to 8.7% performance overhead) deployed in practice which can be used to check legitimate from illegitimate indirect forward-edge calls during compile time. vPointers are checked based on the class hierarchy. ShrinkWrap [20] (as far as we know not deployed in practice) is a tool which further reduces the legitimate vTbles ranges for a given indirect call site through precise analysis of the program class hierarchy and vTable hierarchy. Evaluation results show similar performance overhead but more precision w.r.t. to legitimate vTables entries per call site. We noticed by analyzing the previous research results that the overhead incurred by these security checks can be very high due to the fact that for each call site many range checks have to be performed during runtime. Therefore, despite its security benefit these types of checks can not be applied in our opinion to high performance applications.

As alternative, there are other highly promising tools (not deployed in practice) that can be used to mitigate some of the drawbacks of the previous tools. Bounov et al [10] presented a tool (≈ 1% runtime overhead) for indirect forward-edge call site checking based on vTable layout interleaving. The tool has better performance than VTV and better precision w.r.t. allowed vTables per indirect call site. Its precision (selecting legitimate vTables for each call site) compared to ShrinkWrap is lower since it does not consider vTable inheritance paths. vTrust [46] (average run-time overhead 2.2%) enforces two layers of defense (virtual function type enforcement and vTable pointer sanitization) against vTable corruption, injection and reuse. TypeArmor [44] (≤ than 3 % runtime overhead) enforces an CFI policy based on runtime checking of caller/caller pairs based on function parameter count matching (coarse grained, parameter types and more than six parameters can be used as well). Important to notice is that there are no C++ language semantics which can be used to enforce type and parameter count matching for indirect call/callee pairs, this could be addresses with specifically intended language constructs in future.

Security Implications of Forbidden Indirect Calls. The C++ language standard (12.7 [22]) does not specify what happens when calling different vTable entries from an indirect call site. The standard says that we have have a virtual function related undefined behavior when: "a virtual function call uses an explicit class member access and the object expression refers to the complete object of x or one of that object's base class subobjects but not x or one of its base class subobjects". As undefined behavior is not a clearly defined concept we argue that in order to be able to deal with undefined behavior or unspecified behavior related to virtual function calls one needs to know how these language dependent concepts are implemented inside the used compilers.

Forbidden forward-edge indirect calls are the result of a vPointer corruption. A vPointer corruption is not a vulnerability but rather a capability which can be the result of a spatial or temporal memory corruption through: (1) bad-casting [26] of C++ objects, (2) buffer overflow in a buffer adjacent to a C++ object or a use-after-free condition [40]. A vPointer corruption can be exploited in several ways. A manipulated vPointer can be exploited by pointing it in any existing or added program vTable entry or into a fake vTable which was added by an attacker. For example in case a

vPointer was corrupted than the attacker could highjack the control flow of the program and start a COOP attack [40].

vPointer corruptions are a real security threat which can be exploited if there is a memory corruption (e.g. buffer overflow) which is adjacent to the C++ object or a use-after-free condition. As a consequence each corruption which can reach an object (e.g. bad object casts) is a potential exploit vector for a vPointer corruption. Interestingly to notice in this context is that through: (1) memory layout analysis (through highly configurable compiler tool chains) of source code based locations which are highly prone to memory corruptions such as declarations and uses of buffers, integers or pointer deallocations one can obtain the internal machine code layout representation. (2) analysis of a code corruption which is adjacent (based on (1)) to a C++ object based on application class hierarchy, the vTble hierarchy and each location in source code where an object is declared and used (e.g., modern compiler tool chains can spill out this information for free), one can derive an analysis which can determine—up to a certain extent—if a memory corruption can influence (is adjacent) to a C++ object.

Finally, we notice that by building tools based on this two concepts (i.e., (1) and (2)) attackers (e.g., used to find new vulnerabilities) and for defenders which can harden the source code with checks only at the places which are most exposed to such vulnerabilities (i.e., we name this targeted security hardening).

Real COOP Attack Example. The given example depicted in Figure 2 is a proof of concept exploit extracted from [40] and used in order to perform a COOP attack on the Firefox browser. A buffer overflow bug was used in order to call into existing vTable entries by using the a main loop gadget. The attack concludes with opening of an Unix shell. A real-world bug, CVE-2014-3176, was exploited by Crane et al. [17] in order to perform another COOP attack on the Chromium browser. The details of the second attack are far to complex (i.e., involves not properly handled interaction of extensions, IPC, the sync API, and Google V8) and for this reason we briefly present the first documented COOP exploit on a Linux machine.

The C++ class `nsMultiplexInputStream` contains a main loop gadget inside the function `nsMultiplexInputStream::Close(void)` which is performing an indirect calls by dispatching indirect calls on the objects contained in the array. The objects contained in the array during normal execution are of type `nsInputStream` and each of the objects will call the `Close(void)` function in order to close each of the previously opened streams. In order to perform the COOP attack the attacker crafts a C++ program containing a array buffer holding six fake objects. Fake objects can call inside (and outside) the initial class and vTable hierarchies with no constraints. During the attack a buffer is created in order to hold the fake objects. The crafted buffer will be called in stead of the real code in order to call different functions available in the program code. For example the attacker calls a function contained in the class `xpcAccessibleGeneric` which is not in the class

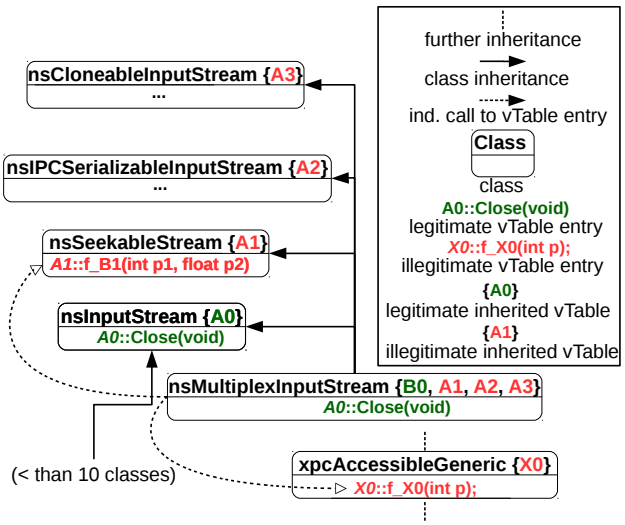

Figure 2: Class inheritance hierarchy of the classes involved in the COOP attack against the Firefox browser. Red letters indicate forbidden vTble entries and green letters indicate allowed vTable entries for the given indirect call site contained in the main loop gadget.

hierarchy or vTable hierarchy of the initially intended type of objects used inside the array. Moreover, the header file of this class (`xpcAccessibleGeneric`) is not included in the class `nsMultiplexInputStream`. In total six fake objects are used to call into functions residing in not related class hierarchies with varying number of parameters and return types. The final goal of this attack is to prepare the program memory such that a Unix shell can be opened at the end of this attack.

This example illustrates why detecting vPointer corruptions is not trivial for real-world applications. As depicted in Figure 2 the class `nsInputStream` has 11 classes which inherit directly or indirectly from this class. The classes `nsSeekableStream`, `nsIPCSerializableInputStream` and `nsClonea-bleInputStream` provide additional inherited vTables which represent illegitimate call targets for the initial `nsInputStream` objects and legitimate call targets for the six fake objects which were added during the attack. Furthermore, declaration and usage of the objects can be wide spread in the source code. This makes detection of the object types (base class), range of vTables (longest vTable inheritance path for a particular call site) and parameter types of the vTable entries (functions) in which it is allowed to call a trivial task for source code (current research work is mostly concerned with performance issues) applications but a hard task in our opinion when one wants to apply similar security policies (e.g. which rely on parameter types of vTable entries) to executables.

3 TypeShield Overview

Adversary Model and Assumptions. We largely use the same threat model and the same basic assumptions as de-

scribed in the TypeArmor paper [44], meaning that our attacker has read and write access to the data sections of the attacked binary. We also assume that the protected binary does not contain self modifying code, handcrafted assembly or any kind of obfuscation. We also consider pages to be either writable or executable but not both at the same time. Furthermore we assume that our attacker has the ability to execute a memory corruption to hijack the programs control flow. As our schema targets only forward control flow, namely indirect function calls, we assume that a solution for backward CFI edges is in place.

Invariants for Targets and Callsites. Advanced code reuse attacks attempt to change the calltargets that are invoked within indirect call-sites, standard CFI solutions cannot defend against this and TypeArmor proposed the approach of creating two sets of invariants.

1. Indirect call-sites provide a number of parameters (possibly overestimated compared to source)

2. Call-targets require a minimum number of parameters (possibly underestimated compared to source)

The main idea is now that a call-site might not call any function in the binary but only call-targets that do not require more parameters than provided by the call-site itself. To achieve this classification of call-targets and call-sites, TypeArmor proposed to use a modified version of forward liveness analysis for call-targets and backward reaching definitions analysis for call-sites.

TYPESHIELD Impact on COOP. The problem with relying solely on the parameter count is that a call-site being classified as using 6 or more parameters can use basically all address taken functions within the binary. This is however counterproductive and we attempt a possible solution, by extending the classification schema to the single parameters themselves:

1. Indirect call-sites provides a maximum wideness of value to each parameter (possibly overestimated compared to source)

2. Call-targets require a minimum wideness of value for each parameter(possibly underestimated compared to source)

Basically the principle stays the same, but instead of just requiring that the call-site parameter count is not lower than the call-target parameter count we now require the same also for the wideness of each parameter.

While there are still occurrences where call-sites may target all call-targets, we split the buckets up into smaller ones, as shown in Figure 3. For example in the parameter-count oriented schema a call-site classified as (32,32) would be able to call functions classified as (64,0), however in the parameter wideness oriented schema that is not possible.

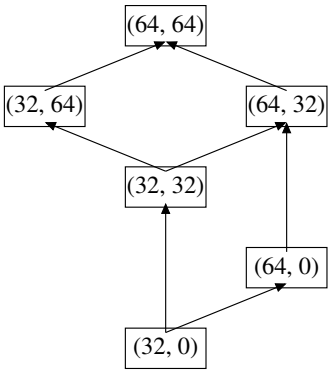

Figure 3: Example for the wideness based schema when only using a parameter wideness of 64, 32 and 0 bits.

4 Design

In this section, we cover the design of TYPESHIELD. We first present the details of the *count* policy in § 4.1—as introduced by [44]—and the new *type* policy in § 4.2. Then we describe general theory needed to transform set-based analysis to register based ones in § 4.3. We follow this up by presenting the theory needed implement the analysis for call-targets in § 4.4 and call-sites in § 4.5 for each policy. Finally, in § 4.6 we introduce a version of address taken analysis based on [49] to restrict the number of available call-targets even more.

4.1 *Count* Policy

What we call the *count* policy is essentially the policy introduced by TypeArmor [44]. The basic idea revolves around classifying call-targets by the number of parameters they provide and call-sites by the number of parameters they require. The schema to match those is that we have call-targets requiring parameters and the call-sites providing parameters as depicted in Figure 4.

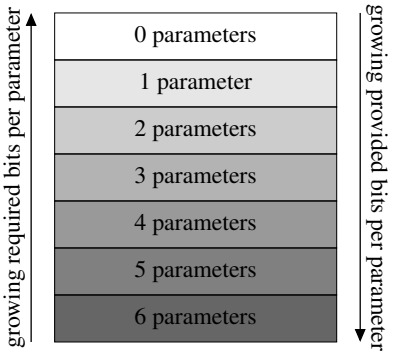

Figure 4: *Count* policy classification schema for call-sites and call-targets.

Furthermore, generating 100% precise measurements for such classification with binaries as the only source of information is rather difficult. Therefore over-estimations of parameter count for call-sites and underestimations of the pa-

rameter count for call-targets is deemed acceptable. This classification is based on the general purpose registers that the call convention of the current ABI—in this case the SystemV ABI—designates as parameter registers. Furthermore, we completely ignore floating point registers or multi-integer registers. The core of the *count* policy is now to allow any call-site *cs*, which provides c_{cs} parameters, to call any call-target *ct*, which requires c_{ct} parameters, iff $c_{ct} \leq c_{cs}$ holds. However, the main problem is that while there is a significant restriction of call-targets for the lower call-sites, the restriction capability drops rather rapidly when reaching higher parameter counts, with call-sites that use 6 or more parameters being able to call all possible call-targets: $\forall cs_1, cs_2.c_{cs_1} \leq c_{cs_2} \Longrightarrow \|\{ct \in \mathscr{F} | c_{ct} \leq c_{cs_1}\}\| \leq \|\{ct \in \mathscr{F} | c_{ct} \leq c_{cs_2}\}\|$ One possible remedy would be the ability to introduce an upper bound for the classification deviation of parameter counts, however as of now, this does not seem feasible with current technology. Another possibility would be the overall reduction of call-sites, which can access the same set of call-targets, a route we will explore within this work.

4.2 *Type* Policy

What we call the *type* policy is the idea of not only relying on the parameter count but also on the type of a parameter. However due to complexity reasons, we are restricting ourselves to the general purpose registers, which the SystemV ABI designates as parameter registers. Furthermore we are not inferring the actual type of the data but the wideness of the data stored in the register. The schema again is that we have call-targets requiring wideness and the call-site providing wideness as depicted in Figure 5.

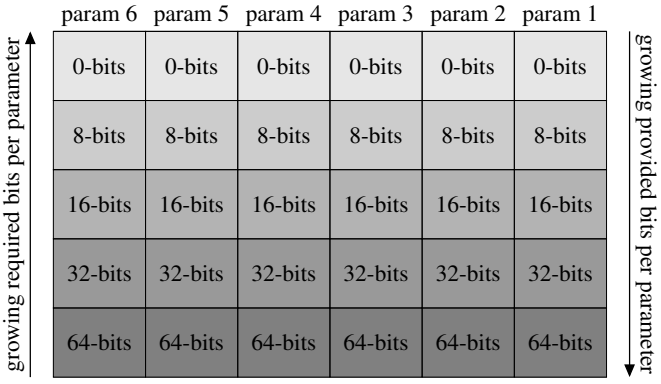

Figure 5: *Type* policy schema for call-sites and call-targets.

We are currently interested in x86-64 binaries, the registers we are looking at are 64-bit registers that can be accessed in four different ways: *1)* the whole 64-bits of the register, meaning a wideness of 64, *2)* the lower 32-bits of the register, meaning a wideness of 32, *3)* the lower 16-bits of the register, meaning a wideness of 16, and *4)* the lower 8-bits of the register, meaning a wideness of 8.

Four of those registers can also directly access the higher 8-bits of the lower 16-bits of the register. For our purpose we register this access as a 16-bit access. Based on this information, we can assign a register one of 5 possible types $\mathscr{T} = \{64, 32, 16, 8, 0\}$. We also included the type 0 to model the absence of data within a register. Similar to the *count* policy, we allow overestimation of types in call-sites and underestimation of types in call-targets. However, the matching idea is different, because as can we depict in Figure 5, the type of a call-target and a call-site no longer depends solely on its parameter count, each call-site and call-target has its type from the set of $\mathscr{T}^6$, with the following comparison operator: $u \leq_{type} v :\Longleftrightarrow \forall_{i=0}^{5} u_i \leq v_i$, with $u, v \in \mathscr{T}^6$ Again we allow any call-site *cs* call any call-target *ct*, when it fulfills the requirement $ct \leq cs$. The way we represent this is by letting the type for a call-target parameter progress from 64-bit to 0-bit—if a call-target requires a 32-bit value in its 1st parameter, it also should accept a 64-bit value from its call-site—and similarly we let the type for a call-site progress from 0-bit to 64-bit - If a call-site provides a 32-bit value in its 1st parameter it also provides a 16-bit, 8-bit and 0-bit to a call-target. Now the advantage of the *type* policy in comparison to the *count* policy is that while our type comparison implies the count comparison, the other direction does not hold. Meaning, just having an equal or lesser number of parameters than a call-site, does no longer allow a call-target being called there, thus restricting the number of call-targets per call-site even further. A function that requires 64-bit in its first parameter, and 0-bit in all other parameters, would have been callable by a call-site providing 8-bit in its first and second parameter when using the *count* policy, however in the *type* policy this is no longer possible.

4.3 Instruction Analysis

Usually data-flow analysis algorithms are based on set of variable or sets of definitions, which both are basically unbounded. However, we are analyzing the state of registers, which are baked into hardware and therefore their number is given, thus requiring us to adapt the data-flow theory to work on tuples.

The set $\mathscr{I}$ describes all possible instructions that can occur within the executable section of a binary. (in our case this is based on the instruction set for x86-64 processors)

An instruction $i \in \mathscr{I}$ can non-exclusively perform two kinds of operations on any number of existing registers:

1. Read *n*-bits from the register with $n \in \{64, 32, 16, 8\}$.
2. Write *n*-bits to the register with $n \in \{64, 32, 16, 8\}$.

Thus we describe the possible change that occurs in one register with the set $S = \{w64, w32, w16, w8, 0\} \times \{r64, r32, r16, r8, 0\}$. Note that 0 signals the absence of either a write or read access and $(0,0)$ signals the absence of both. Furthermore *wn* or *rn* with $n \in \{64, 32, 16, 8\}$ implies all *wm* or *rm* with $m \in \{64, 32, 16, 8\}$ and $m < n$ (e.g. *r*64 implies *r*32). Note that we exclude 0, as it means the absence of any access.

SystemV ABI specifies 16 general purpose integer registers, thus for our purpose we represent the change occurring at the processor level as $\mathscr{S} = S^{16}$.

At last we declare a function, which calculates the change occurring in the processor state, when executing an instruction from $\mathscr{I}$: $decode : \mathscr{I} \mapsto \mathscr{S}$

However, we do not go into detail how this function actually calculates this sate, because we rely on external libraries to perform this task. Implementing this function our self is out of scope due to the lengthy work required, as the x86-64 instruction set is quite large.

4.4 Call-target Analysis

For either *count* or *type* policy to work, we need to arrive at an underestimation of the required parameters by any function existing within the targeted binary. We will employ a modified version of liveness analysis that tracks registers instead of variables to generate the needed underestimation. As our algorithm will be customizable, we look at the required merge functions to implement *count* and *type* policy. Furthermore we need to eliminate the passing of variadic parameter lists from variadic functions, as this might cause our analysis to overestimate the required parameters.

Variable Liveness Analysis Theory. A variable is alive before the execution of an instruction, if at least one of the originating paths contains a read access before the variable is written to again. We employ liveness analysis, because we are looking for the parameters a function requires. This essentially requires read before write access, however global variables usually would also fall into this category, however these would not reside within parameter registers at the start of a function.

Khedker et al. [23] defines live variable analysis on blocks in the following manner:

$$In_n := (Out_n - Kill_n) \cup Gen_n \tag{1a}$$

$$Out_n := \begin{cases} Bl & \text{n is end block} \\ \bigcup_{s \in succ(n)} In_s & \text{otherwise} \end{cases} \tag{1b}$$

Bl is the default state at the end of a path of execution and in our case reaching that state would mean that a variable has never been used (neither written nor read). The set $Kill_n$ describes all variables that are no longer live after the block *n*, meaning that a variable occurring within this set has been written to. The set Gen_n describes all variables that are alive due to the block *n*, meaning that a variable occurring within this set has been read before it was written to.

However, we cannot use variable liveness analysis as is, because the analysis is based on potentially unbound variable sets, while we are restricted to a finite number of registers and states. We also require an underestimation of live variables an not an overestimation as provided by standard liveness analysis. Furthermore we have to define how to interpret the changes occurring within one block based on the the change caused by its instructions. Considering this, we arrive at algorithm 6 to compute the liveness state at the start of a basic block.

```
Function analyze(block : BasicBlock) : 𝒮^ℒ is
    state = Bl
    foreach inst ∈ block do
        state' = analyze_instr(inst)
        state = merge_v(state, state')
    end
    states = { }
    blocks = succ(block)
    foreach block' ∈ blocks do
        state' = analyse(block')
        states = states ∪ { state' }
    end
    state' = merge_h (states)
    return merge_v(state, state')
end
```

Figure 6: Algorithm to analyze the liveness of a Basic Block.

This algorithm relies on various functions that can be used to configure its behavior. We need to define the function *merge_v*, which describes how to compound the state change of the current instruction and the current state, the function *merge_h*, which describes how to merge the states of several paths, the instruction analysis function *analyze_instr*. The function *succ*, which retrieves all possible successors of a block won't be implemented by us, because we rely on the DynInst instrumentation framework to achieve this.

$$merge\_v : \mathscr{S}^{\mathscr{L}} \times \mathscr{S}^{\mathscr{L}} \mapsto \mathscr{S}^{\mathscr{L}} \tag{2a}$$

$$merge\_h : \mathscr{P}(\mathscr{S}^{\mathscr{L}}) \mapsto \mathscr{S}^{\mathscr{L}} \tag{2b}$$

$$analyze\_instr : \mathscr{I} \mapsto \mathscr{S}^{\mathscr{L}} \tag{2c}$$

$$succ : \mathscr{I} \mapsto \mathscr{P}(\mathscr{I}) \tag{2d}$$

As the *analyze_instr* function calculates the effect of an instruction and is the heart of the analyze function. It will also handle non jump and non fall-through successors, as these are not handled by DynInst in our case. We essentially have four cases that we handle:

1. if the instruction is an indirect call or a direct call but we chose not follow calls, then return a state where all registers are considered written.
2. if the instruction is a direct call and we chose to follow calls, then we spawn a new analysis and return its result.
3. if the instruction is a constant write (e.g. xor of two registers) then we remove the read portion before we return the decoded state.
4. in all other cases we simply return the decoded state

This leaves us with the two merge functions remaining undefined and we will leave the implementation of these and the

interpretation of the liveness state $\mathscr{S}^{\mathscr{L}}$ into parameters up to the following subsections.

Required Parameter Count. To implement the *count* policy, we only need a coarse representation of the state of one register, thus we are interested in the following three different exclusive informations:

1. Was the register written to before its value could be read ?
   We represent this with the state W.
2. Was the register read from before its value was overwritten ?
   We represent this with the state R.
3. Did neither read nor write access occur for the register ?
   We represent this with the state C.

This gives us the following register state $S^{\mathscr{L}} = \{C, R, W\}$ which translates to the register superstate $\mathscr{S}^{\mathscr{L}} = (S^{\mathscr{L}})^{16}$. Now, we assume that unless the instructions we are looking at does discard the value it is reading (`xor rax rax` would be such an instruction that we call const_write) that reading does precede the writing withing one instruction. Furthermore we are only interested in the first occurrence of a R or W within one path, as following reads or writes do not give us more information. Therefore, we can define our vertical merge function in the following way:

$$merge\_v^r(cur, delta) = \begin{cases} delta & cur = C \\ cur & otherwise \end{cases} \tag{3}$$

$$merge\_v(cur, delta) = (s'_0, ...s'_1 5) \text{ with } s'_j = merge\_v^r(cur_j, delta_j) \tag{4}$$

Our horizontal merge function is a simple pairwise combination of the given set of states:

$$merge\_h(\{s\}) = s \tag{5}$$

$$merge\_h(\{s\} \cup s') = s \circ merge\_h(s') \tag{6}$$

We have three viable possibilities for our combination operator $\circ$, depicted in Table 1, which all give priority to W:

$\bigsqcap^{\mathscr{L}}$ is what we call the destructive combination operator, as it returns W on any mismatch.

$\bigcap^{\mathscr{L}}$ is what we call the intersection operator, as it returns C, when combining C and R, similar to an intersection.

$\bigcup^{\mathscr{L}}$ is what we call the union operator, as it returns R, when combining C and R similar to a union.

We apply the liveness analysis for each function with the entry block of the function as start and the return blocks as end and after an analysis run for a function, the index of highest parameter register based on the used call convention that has the state R is considered to be the number of parameters a function at least requires to be prepared by a call-site.

Required Parameter Wideness. To implement the *type* policy, we need a finer representation of the state of one register, thus we are interested in the following three different not neccessarily exlusive informations:

$\bigsqcap^{\mathscr{L}}$	C	R	W
C	C	W	W
R	W	R	W
W	W	W	W

$\bigcap^{\mathscr{L}}$	C	R	W
C	C	C	W
R	C	R	W
W	W	W	W

$\bigcup^{\mathscr{L}}$	C	R	W
C	C	R	W
R	R	R	W
W	W	W	W

Table 1: Different mappings for combining two liveness state values in horizontal matching for the *count* policy.

1. Was the register written to before its value could be read ?
   We represent this with the state W.
2. How much was read from the register before its value was overwritten?
   We represent this with the states $\{r8, r16, r32, r64\}$ using R as a placeholder for arbitrary reads.
3. Did neither read nor write access occur for the register ?
   We represent this with the state C.

This gives us the following register state $S^{\mathscr{L}} = \{C, r8, r16, r32, r64, W\}$ which translates to the register superstate $\mathscr{S}^{\mathscr{L}} = (S^{\mathscr{L}})^{16}$. Now, we assume that unless the instructions we are looking at does discard the value it is reading (`xor rax rax` would be such an instruction that we call const_write) that reading does precede the writing withing one instruction.

As there could happen more than one read of a register before it is written, we might be interested in more than just the first occurrence of a write or read on a path. We arrive therefore at three possible vertical merge functions:

- The same vertical merge operator as used in the *count* policy, which only gives us the first non C state ($merge\_v^r$).
- A vertical merge operator that conceptually intersects all read accesses along a path until the first write occurs ($merge\_v^i$).
- A vertical merge operator that conceptually calculates the union of all read accesses along a path until the first write occurs ($merge\_v^u$).

Our horizontal merge function is a simple pairwise combination of the given set of states:

$$merge\_h(\{s\}) = s \tag{7}$$

$$merge\_h(\{s\} \cup s') = s \circ merge\_h(s') \tag{8}$$

The results of our experiments with the implementation of call-target classification gave presented us with essentially one possible candidate that we can base our horizontal merge function on, namely the union operator with an analysis function that follows into direct calls. The basic schema of the merging is depicted in 2 and it essentially behaves as if it was the union operator (when both states are set, the higher one is chosen). However, we have to account for W being used as an end marker, which is why we added mapping for RW, which is essentially that.

$\bigcup^{\mathscr{L}}$	C	R	W	RW
C	C	R	W	RW
R	R	$R^{\cup}$	W	$R^{\cup}W$
W	W	W	W	W
RW	RW	$R^{\cup}W$	W	RW

Table 2: The union mapping operator for liveness in the *type* policy.

```
00000000004222f0 <make_cmd>:
  4222f0:       push   %r15
  4222f2:       push   %r14
  4222f4:       push   %rbx
  4222f5:       sub    $0xd0,%rsp
  4222fc:       mov    %esi,%r15d
  4222ff:       mov    %rdi,%r14
  422302:       test   %al,%al
  422304:       je     42233d <make_cmd+0x4d>
  422306:       movaps %xmm0,0x50(%rsp)
  42230b:       movaps %xmm1,0x60(%rsp)
  422310:       movaps %xmm2,0x70(%rsp)
  422315:       movaps %xmm3,0x80(%rsp)
  42231d:       movaps %xmm4,0x90(%rsp)
  422325:       movaps %xmm5,0xa0(%rsp)
  42232d:       movaps %xmm6,0xb0(%rsp)
  422335:       movaps %xmm7,0xc0(%rsp)
  42233d:       mov    %r9,0x48(%rsp)
  422342:       mov    %r8,0x40(%rsp)
  422347:       mov    %rcx,0x38(%rsp)
  42234c:       mov    %rdx,0x30(%rsp)
  422351:       mov    $0x50,%esi
  422356:       mov    %r14,%rdi
  422359:       callq  409430 <pcalloc>
```

Figure 7: ASM code of the `make_cmd` function with optimize level O2, which has a variadic parameter list.

Variadic Functions. Variadic functions are special functions in C/C++ that have a basic set of parameters, which they always require and a variadic set of parameters, which as the name suggests may vary. A prominent example of this would be the *printf* function, which is used to output text to *stdout*.

The problem with these functions is that to allow for easier processing of parameters usually all potential variadic parameters are moved into a contiguous block of memory, as can be seen in the assembly in Figure 7. Our analysis interprets that as a read access on all parameters and we arrive at a problematic overestimation.

Our solution to this problem is to find these spurious reads and ignore them. A compiler will implement this type of operation very similar foll all cases, thus we can achieve this using the following steps:

- Look for what we call the xmm-passthrough block, which entierly consist of moving the values of registers `xmm0` to `xmm7` into contiguous memory (in our case basic block [`0x422306`, `0x42233d` [ ).
- Look at the predecessor of the xmm-passthrough block, which we call the entry block (in our case basic block [`0x4222f0`, `0x4222f2` [ ). Check if the successors of the entry block consist of the xmm-passthrough block and the successor of the xmm-passthrough block, which we call the param-passthrough block (in our case basic block [`0x42233d`, `0x42235e` [ ).
- Look at the param-passthrough block and set all instructions that move the value of a parameter register into memory to be ignored (in our case the instructions `0x42233d`, `0x422342`, `0x422347` and `0x42234c`).

4.5 Call-site Analysis

For either *count* or *type* policy to work, we need to arrive at an overestimation of the provided parameters by any indirect call-site existing within the targeted binary. We will employ a modified version of reaching analysis that tracks registers instead of variables to generate the needed overestimation. As our algorithm will be customizable, we look at the required merge functions to implement *count* and *type* policy.

Reaching Definitions Theory. An assignment of a value to a variable is a reaching definition at the end of a block n, if that definition is present within at least one path from start to the end of the block n without being overwritten by another value assignment to the same variable. We employ reaching definitions analysis, because we are looking for the parameters a call-site provides. This essentially requires the last known set of definitions that reach the actual call instruction within the parameter registers.

The book [23] defines reaching definition analysis on blocks in the following manner:

$$In_n := \begin{cases} Bl & \text{n is start block} \\ \bigcup\limits_{p \in pred(n)} Out_p & \text{otherwise} \end{cases} \tag{9a}$$

$$Out_n := (In_n - Kill_n) \cup Gen_n \tag{9b}$$

Bl is the default state at the start of a path of execution and in our case reaching that state would mean that we do not know whether a value has been provided for the variable and therefore we assume that one has been provided, reaching an overestimation. The set $Kill_n$ describes all definitions that are removed within this block, meaning that the value of a variable has been overwritten. The set Gen_n describes the new definitions that have been provided by the block n, meaning that the value of a variable has been assigned. Considering this, we can assume that $Gen_n \subseteq Kill_n$, as we can always create new definitions, but not simply remove definitions without assigning a new value to the variable.

However, we cannot use reaching definition analysis as is, because the analysis is again based on potentially unbound variable sets, while we are restricted to a finite number of registers and states. This time however the analysis provides us with an overestimation, we however want to get a result as close as possible so we again want to customize merge functions. Furthermore we have to define how to interpret the

Function *analyze(block : BasicBlock) :* $\mathscr{S}^{\mathscr{R}}$ **is**
    state = Bl
    foreach *inst* $\in$ *reversed(block)* **do**
        state' = analyze_instr(inst)
        state = merge_v(state, state')
    end
    states = {}
    blocks = pred(block)
    foreach *block'* $\in$ *blocks* **do**
        state' = analyse(block')
        states = states $\cup$ { state' }
    end
    state' = merge_h (states)
    return merge_v(state, state')
end

Figure 8: Algorithm to analyse the reaching definitions of a Basic Block.

changes occuring withing one block based on the the change caused by its instructions. Considering this, w, we arrive at algorithm 8 to compute the liveness state at the start of a basic block.

This algorithm relies on various functions that can be used to configure its behavior. We need to define the function *merge_v*, which describes how to compound the state change of the current instruction and the current state, the function *merge_h*, which describes how to merge the states of several paths, the instruction analysis function *analyze_instr*. The function *pred*, which retrieves all possible predecessors of a block won't be implemented by us, because we rely on the DynInst instrumentation framework to achieve this.

$$merge\_v : \mathscr{S}^{\mathscr{R}} \times \mathscr{S}^{\mathscr{R}} \mapsto \mathscr{S}^{\mathscr{L}} \tag{10a}$$

$$merge\_h : \mathscr{P}(\mathscr{S}^{\mathscr{R}}) \mapsto \mathscr{S}^{\mathscr{R}} \tag{10b}$$

$$analyze\_instr : \mathscr{I} \mapsto \mathscr{S}^{\mathscr{R}} \tag{10c}$$

$$pred : \mathscr{I} \mapsto \mathscr{P}(\mathscr{I}) \tag{10d}$$

As the *analyze_instr* function calculates the effect of an instruction and is the heart of the analyze function. It will also handle non jump and non fall-through successors, as these are not handled by DynInst in our case. We essentially have three cases that we handle:

- if the instruction is an indirect call or a direct call but we chose not follow calls, then return a state where all trashed are considered written.
- if the instruction is a direct call and we chose to follow calls, then we spawn a new analysis and return its result.
- in all other cases we simply return the decoded state.

This leaves us with the two merge functions remaining undefined and we will leave the implementation of these and the interpretation of the liveness state $\mathscr{S}^{\mathscr{L}}$ into parameters up to the following subsections.

Provided Parameter Count. To implement the *count* policy, we only need a coarse representation of the state of one register, thus we are interested in the following three different exclusive informations:

$\sqcap^{\mathscr{R}}$	U	S	T
U	U	T	T
S	T	S	T
T	T	T	T

$\cap^{\mathscr{R}}$	U	S	T
U	U	U	T
S	U	S	T
T	T	T	T

$\cup^{\mathscr{R}}$	U	S	T
U	U	S	T
S	S	S	T
T	T	T	T

$\sqcup^{\mathscr{R}}$	U	S	T
U	U	S	T
S	S	S	S
T	T	S	T

Table 3: Different mappings for combining two reaching state values in horizontal matching for the *count* policy.

- Was the register value trashed ?
  We represent this with the state T.
- Was the register written to ?
  We represent this with the state S.
- Was the register neither trashed nor written to ?
  We represent this with the state U.

This gives us the following register state $S^{\mathscr{L}} = \{T, S, U\}$ which translates to the register superstate $\mathscr{S}^{\mathscr{R}} = (S^{\mathscr{R}})^{16}$. We are only interested in the first occurrence of a S or T within one path, as following reads or writes do not give us more information. Therefore, we can define our vertical merge function in the following way:

$$merge\_v^r(cur, delta) = \begin{cases} delta & cur = U \\ cur & otherwise \end{cases} \tag{11}$$

$$merge\_v(cur, delta) = (s'_0, ...s'_1 5) \text{ with } s'_j = merge\_v^r(cur_j, delta_j) \tag{12}$$

Our horizontal merge function is a simple pairwise combination of the given set of states:

$$merge\_h(\{s\}) = s \tag{13}$$

$$merge\_h(\{s\} \cup s') = s \circ merge\_h(s') \tag{14}$$

We have four viable possibilities for our combination operator $\circ$, depicted in table 3, which all (except one) give priority to T:

$\sqcap^{\mathscr{R}}$ is what we call the destructive combination operator, as it returns T on any mismatch.

$\cap^{\mathscr{R}}$ is what we call the intersection operator, as it returns U, when combining U and S, similar to an intersection.

$\cup^{\mathscr{R}}$ is what we call the union operator, as it returns S, when combining U and S similar to a union.

$\sqcup^{\mathscr{R}}$ is what we call the true union operator, as it gives S precedence over everything and returns T or U only when both sides are T or U being more inclusive than a union.

Provided Parameter Wideness. To implement the *type* policy, we need a finer representation of the state of one register, thus we are interested in the following three informations:

- Was the register value trashed ?
  We represent this with the state T.

- Was the register written to and how much ?
  We represent this with the states $\{s64, s32, s16, s8\}$ using S as a placeholder for arbitrary writes.
- Was the register neither trashed nor written to ?
  We represent this with the state U.

This gives us the following register state $S^{\mathscr{L}} = \{T, s64, s32, s16, s8, U\}$ which translates to the register superstate $\mathscr{S}^{\mathscr{R}} = (S^{\mathscr{R}})^{16}$. Again, we are only interested in the first occurrence of a state that is not U in a path, as following reads or writes do not give us more information.

Therefore we can use the same vertical merge function as for the *count* policy, which is essentially a pass-through until the first non U state.

Our horizontal merge function is again a simple pairwise combination of the given set of states:

$$merge\_h(\{s\}) = s \tag{15}$$
$$merge\_h(\{s\} \cup s') = s \circ merge\_h(s') \tag{16}$$

However, we have different possibilities regarding the merge operator. Experiments with our implementations for call-site classification in the *count* policy have given us the following results:

- The best candidate to minimize the problematic matches is the union operator without following direct calls.
- The best candidate to maximize precision is the intersection operator with following direct calls.

We therefore arrive at three viable possibilities for our combination operator $\circ$, depicted in table 4, which all (except one) give priority to T:

$\bigcap^{\mathscr{R}}$ is what we call the intersection operator, as it returns U, when combining U and S, similar to an intersection furthermore we also calculate the intersection of states when both states are set (the lower of the two is returned).

$\bigsqcap^{\mathscr{R}}$ is what we call the half intersection operator, as it returns U, when combining U and S, similar to an intersection but we calculate the union of states when both states are set (the higher of the two is returned).

$\bigcup^{\mathscr{R}}$ is what we call the union operator, as it returns S, when combining U and S similar to a union furthermore we calculate the union of states when both states are set (the higher of the two is returned).

Initial experiments with this implementation showed two problems regarding provided wideness detection. Parameter lists with "holes" and address wideness underestimation.

Parameter Lists with *Holes*. This refers to parameter lists that show one or more `void` parameters between start to the last actual parameter. These are not existant in actual code but our analysis has the possibility of generating them through the merge operations. An example would be the following: A parameter list of $(64, 0, 64, 0, 0, 0)$ is concluded, although the actual parameter list might be $(64, 32, 64, 0, 0, 0)$. While the trailing 0es are what we expect, the 0 at the second parameter position will cause trouble, because it is an underestimation at the single parameter level, which we need to avoid. Our solution is to simply scan our reaching analysis result for these holes and replace them with the wideness 64, causing a (possible) overestimation.

$\bigcap^{\mathscr{R}}$	U	S	T
U	U	U	T
S	U	$S^{\cap}$	T
T	T	T	T

$\bigsqcap^{\mathscr{R}}$	U	S	T
U	U	U	T
S	U	$S^{\cup}$	T
T	T	T	T

$\bigcup^{\mathscr{R}}$	U	S	T
U	U	S	T
S	S	$S^{\cup}$	T
T	T	T	T

Table 4: Different mappings for combining two reaching state values in horizontal matching for the *type* policy.

Address Wideness Unterestimation. refers to the problem that while in the call-site a constant value of 32-bit is written to a register, however the call-target uses the whole 64-bit register. This can occur when pointers are passed from the call-site to the call-target. Specifically this happens when pointers to memory inside the ".bss", ".data" or ".rodata" section of the binary are passed. Our solution is to enhance our instruction analysis to watch out for constant writes. In case a 32-bit constant value write is detected, we check if the value is an address within the ".bss", ".data" or ".rodata" section of the binary. If this is the case, we simply return a write access of 64-bits instead of 32-bits. (This is not problematic, because we are looking for an overestimation of parameter wideness) It should be noted that the same problem can arise when a constant write causes the value 0 to be written to a 32-bit register. We use the same solution and set the wideness to 64-bits instead of 32-bits.

4.6 Address Taken Analysis

As of now, we use the maximum available set of call-targets—the set of all function entry basic blocks—as input for our algorithm. To restrict the number of call-targets per call-site even further, we explored the possibility of incorporating an address taken analysis into our application. We base our theory on the paper by Zhang et al. [49], which introduced various types of taken addresses. An address is considered to be taken, when it is loaded into memory or a register.

Address Taken Targets. Based on the notions of [49], which classified taken addresses into several types of indirect control flow targets, we only chose Code Pointer Constants (CK) and discarded the others:

- Code Pointer Constants (CK) are addresses that are calculated during the compilation of the binary and point within the possible range of addresses in the current module or to instruction boundaries. We are however only interested in addresses that directly point to an entry basic block of a function, as these are the only valid

targets for any call-site.

- Computed code pointers (CC) are the result of simple pointer arithmetic, however these are only used for intra-procedural jumps. We rely on DynInst to resolve those and only focus on indirect call-sites, therefore these are of no interest to us.

- Exception handling addresses (EH) are used to handle exceptions within C++ functions and are modeled as jumps within the function. These are therefore within the normal control flow that we rely on DynInst to resolve for us.

- Exported function addresses (ES) are essentially functions that point outside of our current module (usually to dynamically linked libraries) and are implemented as jumps, which are of no concern to us, because our analysis is only concerned about the current object.

- Return addresses (RA), which are the addresses next to a call instruction, are also of no interest to us, because we only implement forward control flow integrity.

Binary Analysis. Our approach of identifying taken addresses consists of two steps: First, we iterate over the raw binary content of data sections. Second, we iterate over all functions within the disassembled binary. We rely on DynInst to provide us with the boundaries of the sections inside the binary and in case of shared libraries with the needed translation to current memory addresses:

- We look at three different data sections of the binary, which could possibly contain taken addresses: the .data, .rodata and .dynsym sections. As [49] proposed, we slide a four byte window over the data within those sections and look for addresses that point to function entry blocks. However, we are looking at x64 binaries therefore we additionally use an eight byte window. In case of shared libraries, we need to let DynInst translate the raw address, we extracted, so we can perform the function check.

- We specifically look for instructions that load a constant value into a register or memory, and again check whether the address points to the entry block of a function.

5 Implementation

We implemented TYPESHIELD as a module pass for the *di-opt* environment pass provided by the DynInst [8] instrumentation framework (v. 9.2.0). However, converting the pass to a standalone executable is also possible, as we do not rely on an extended set of DynInst features except for the pass abstraction.

We currently restricted our analysis and instrumentation to x86-64 bit elf binaries using the SystemV call convention, because the DynInst library does not yet support the Windows platform. However, there is currently work going on in order to allow DynInst to work with Windows binaries as well. We focused on the SystemV call convention as most C/C++ compilers on Linux implement this ABI, however we encapsulated most ABI dependent behavior, so it should be possible to implement other ABIs with relative ease. Therefore, we deem it possible to implement TYPESHIELD for the Windows platform in the near future, as we do not use any other platform-dependent API's.

We developed the core part of our pass in an instruction analyzer, which relies on the DynamoRIO [1] library (v. 6.6.1) to decode single instructions and provide access to its information. The analyzer is then used to implement our version of the reaching and liveness analysis (similar to PathArmor [44]), which can be customized with relative ease, as we allow for arbitrary path merging functions. However, we implemented the three basic versions as follows: destructive, intersection and union. In order to accomplish this we patched the DynInst library in order to allow for local annotation of call-targets with arbitrary information, leveraging its relocation schema, which relies on the basic block abstraction.

We implemented a Clang/LLVM (v. 4.0.0, trunk 283889) pass used for collecting ground truth data in order to measure the quality and performance of our tool. The ground truth data is then used to verify the output of our tool for several test targets. This is accomplished with the help of our python based evaluation and test environment.

In total we implemented TYPESHIELD in 5123 source code lines (SLoC) of C++ code, our Clang/LLVM pass in 200 SLoC of C++ code and our test environment in 2674 SLoC of Python code.

6 Evaluation

We evaluated our TYPESHIELD by instrumenting various open source applications and analyzing the result. We used the two ftp server applications vsftpd (version 1.1.0) and proftpd (version 1.3.3), the two http server applications postgresql (version 9.0.10) and mysql (5.1.65), the memory cache application memcached (version 1.4.20) and the node.js server application node (version 0.12.5). We chose these applications, which are a subset of the applications also used by the TypeAmor [44] to allow for later comparison. In our evaluation of the two modes of TYPESHIELD, we are trying to answer the following questions:

- **RQ1:** How precise is TYPESHIELD in recovering parameter count and type information for call-sites and call-targets from a given binary?

- **RQ2:** How effective is TYPESHIELD in restricting the possible number of call-targets per call-site?

- **RQ3:** What is the performance overhead introduced by TYPESHIELD?

add a binary patch that does not crash none of the programs from SPEC2006.

need a table with all the results for each of the SPEC2006 programs and a bar diagram

- **RQ4:** What is the instrumentation overhead introduced by TYPESHIELD? Here we measure how much the binaries increased in size after the instrumentation was added to the binaries.

  Measure the size (in bytes) of the SPEC2006 testes in RQ3 before and after adding all the patches

- **RQ5:** What level of security does TYPESHIELD offer? We look at our implementation conceptually and assess qualitatievly whether our implementation can interfere with various classes of attacks.

  see the TypeAmor paper w.r.t. security analysis in the evaluation, a CDF figure is here required.

- **RQ6:** Need to be defined see down?

  RQ6: need to define first this question.

Comparison Method. As we do not have access to the source code of TypeAmor, we implemented two modes in TYPESHIELD. The first mode of our tool is an approximate implementation of what we understand is the *count* policy implemented by TypeAmor. The second mode is our implementation of the *type* policy on top of our implementation of the *count* policy.

Experimental Setup. We setup our environment within a VirtualBox (version 5.0.26r) instance, which runs Kubuntu 16.04 LTS (Linux Kernel version 4.4.0) and has access to 3GB of RAM and 4 of 8 provided hardware threads (Intel i7-4170HQ @ 2.50 GHz).

6.1 RQ1: Precision of TYPESHIELD

In this section we need just one or two Table similar to what TypeArmor contains, first we need to define the fields which make most sense.

To measure the precision of TYPESHIELD, we need to compare the classification of call-sites and call-targets as is given by our tool to some sort of ground truth for our test targets. We generate this ground truth by compiling our test targets using a custom compiled Clang/LLVM compiler (version 4.0.0 trunk 283889) with a MachineFunction pass inside the x86 code generation implementation of LLVM. We essentially collect three data points for each call-site/call-target from our LLVM-pass:

- The point of origination, which is either the name of the call-target or the name of the function the call-site resides in.
- The return type that is either expected by the call-site or provided by the call-target.
- The parameter list that is provided by the call-site or expected by the call-target, which discards the variadic argument list.

However, before we can proceed to measure the quality and precision of TYPESHIELD's classification of call-targets and call-sites using our ground truth, we need to evaluate the quality and applicability of the ground truth, we collected.

6.1.1 Quality and Applicability of Ground Truth

To assess the applicability of our collected ground truth, we essentially need to assess the structural compatibility of our two datasets. First, we take a look at the comparability of call-targets, which is quite high throughout optimization levels. Second, we take a look at the compatibility of call-sites, which is qualitatively low in the higher optimization levels, while five of our test targets start with 0% mismatch in O0, mysql stays throughout all levels at a constant mismatch rate of around 18%, and the others between 2% and 17%.

Call-targets. The obvious choice for structural comparison regarding call-targets is their name, as these are simply functions. First, we have to however remove internal functions from our data-sets like the `_init` or `_fini` functions, which are of no consequence for us. Furthermore, while C functions can simply be matched by their name as they are unique through the binary, the same cannot be said about the language C++. One of the key differences between C and C++ is function overloading, which allows to define several functions with the same name, as long as they differ in namespace or parameter type. As LLVM does not know about either concept, the Clang compiler needs to generate unique names. The method used for unique name generation is called mangling and composes the actual name of the function, its the return type, its name-space and the types of its parameter list. We therefore need to reverse this process, which is called demangling and then compare the fully typed names. The table 5 shows three data points regarding call-targets for optimization levels O0, O1, O2 and O3:

- The number of comparable call-targets that are found in both datasets
- The number of call-targets that are found by TYPESHIELD but not by our Clang/LLVM pass, named Clang miss
- The number of call-targets that are found by our Clang/LLVM pass but not by TYPESHIELD, named tool miss

The problematic column is the Clang miss column, as these might indicate problems with TYPESHIELD. These numbers are relatively low (below 1%) throughout optimization levels, with only node showing a significant higher value than the rest of around 1.6%. The column labeled tool miss lists higher numbers, however these are of no real concern to us, as our ground truth pass possibly collects more data: All source

files used during the compilation of our test-targets are incorporated into our ground truth. The compilation might generate more than one binary and therefore not necessary all source files are used for our test-target.

Considering this, we can safely state that our structural matching between ground truth and TYPESHIELD regarding call-targets is nearly perfect (above 98%)

O2	call-targets			call-sites		
Target	match	Clang miss	tool miss	match	Clang miss	tool miss
proftpd	1015	0 (0.0%)	15 (1.45%)	155	0 (0.0%)	0 (0.0%)
vsftpd	318	0 (0.0%)	0 (0.0%)	14	0 (0.0%)	0 (0.0%)
lighttpd	290	0 (0.0%)	311 (51.74%)	66	0 (0.0%)	0 (0.0%)
nginx	921	0 (0.0%)	0 (0.0%)	266	0 (0.0%)	0 (0.0%)
mysqld	9742	13 (0.13%)	3690 (27.47%)	7923	24 (0.3%)	25 (0.31%)
postgres	6930	1 (0.01%)	1512 (17.91%)	687	1 (0.14%)	0 (0.0%)
memcached	133	0 (0.0%)	91 (40.62%)	48	1 (2.04%)	0 (0.0%)
node	20638	339 (1.61%)	620 (2.91%)	10965	29 (0.26%)	26 (0.23%)

Table 5: Table shows the quality of structural matching provided by our automated verify and test environment, regarding call-sites and call-targets when compiling with optimization level O2. The label Clang miss denotes elements not found in the data-set of the Clang/LLVM pass. The label tool miss denotes elements not found in the data-set of TYPESHIELD.

Call-sites. While our structural matching of call-targets is rather simple, we have not so much luck regarding call-sites. While our tool can provide accurate addressing of call-sites within the binary, Clang/LLVM does not have such capabilities in its intermediate representation. Therefore we assume that the ordering of call-sites stays roughly the same within one function and that we exclude all functions, which report a different amount of call-sites in both datasets. The table 5 shows three data points regarding call-sites for optimization levels O0, O1, O2 and O3:

- The number of comparable call-sites that are found in both datasets.
- The number of call-sites that are discarded due to mismatch from the dataset of TYPESHIELD, named Clang miss.
- The number of call-sites that are discarded due to mismatch from the dataset of our Clang/LLVM pass, named tool miss.

Second, we look at call-sites and this is more problematic, as Clang/LLVM does not have a notion of instruction address in its IR, therefore we assume the ordering in a function is the same in both data-sets and when the call-site count is not the same for dyninst and Clang/padyn, we discard it. The are several reasons for mismatch: One is the tailcall optimization, which means that a call instructions at the end of a function are converted into jump instructions. Another one is call-site merging, which happens when a call to a function exists several times within a function and the compiler can merge the paths to this function. Furthermore we already eliminated multiple compilations of the same source file during one test-target compilation (this would have skewed the results in the case of memcached).

Normally up to 20% mismatch would not be that much of a problem, however this percentage is only based on the number of call-sites per function and we cannot really give any guarantees that the call-sites in both data-sets are the same. (Although for O0 there is quite a high possibility due to the absence of nearly all optimizations, however mysqld and node remain problematic).

6.1.2 Precision Call-target Classification (*count*)

We are going to present the experiments and values to find the best possible combination operator for the call-target analysis in the *count* policy.

Experiment Setup. To choose the best possible combination operator for the call-target analysis in implementing the *count* policy, we generated data for all six possible versions of liveness analysis.

- Destructive combination operator with an *analyze* function that does not follow into occurring direct calls see Table 6 for results.
- Destructive combination operator with an *analyze* function that follows into occurring direct calls see Table 6 for results.
- Intersection combination operator with an *analyze* function that does not follow into occurring direct calls see Table 6 for results.
- Intersection combination operator with an *analyze* function that follows into occurring direct calls see Table 6 for results.
- Union combination operator with an *analyze* function that does not follow into occurring direct calls see Table 7 for results.
- Union combination operator with an *analyze* function that follows into occurring direct calls see Table 7 for results.

For each possible version we measured two data points per testtarget, the number and ratio of perfect classifications and the number and ratio of problematic classifications, which in this case refers to overestimations. **Results.** The results of destructive combination and intersection combination are the same. Regardless of operator choice the problem rate is extremely low (under 0.1%). Overall there is a slight improvement when following calls in all result sets. The union operator (geometric mean for O2: 82.24%) is slightly more precise than the destructive/intersection operator (geometric mean for O2: 79.93%). Which presents us the union combination operator as the best possible option here.

O2		not following calls		following calls	
Target	#	perfect	problem	perfect	problem
proftpd	1015	874(86.1%)	0(0.0%)	883(86.99%)	0(0.0%)
vsftpd	318	232(72.95%)	0(0.0%)	248(77.98%)	0(0.0%)
lighttpd	290	251(86.55%)	0(0.0%)	253(87.24%)	0(0.0%)
nginx	921	691(75.02%)	0(0.0%)	695(75.46%)	0(0.0%)
mysqld	9742	6512(66.84%)	1(0.01%)	6567(67.4%)	1(0.01%)
postgres	6930	5752(83.0%)	0(0.0%)	5878(84.81%)	0(0.0%)
memcached	133	117(87.96%)	0(0.0%)	117(87.96%)	0(0.0%)
node	20638	15124(73.28%)	0(0.0%)	15315(74.2%)	0(0.0%)

Table 6: The results for call-target analysis using the destructive/intersection combination operator for the *count* policy throughout different optimizations.

O2		not following calls		following calls	
Target	#	perfect	problem	perfect	problem
proftpd	1015	880(86.69%)	0(0.0%)	920(90.64%)	0(0.0%)
vsftpd	318	233(73.27%)	0(0.0%)	249(78.3%)	0(0.0%)
lighttpd	290	266(91.72%)	0(0.0%)	268(92.41%)	0(0.0%)
nginx	921	720(78.17%)	0(0.0%)	724(78.61%)	0(0.0%)
mysqld	9742	6684(68.61%)	1(0.01%)	6759(69.38%)	1(0.01%)
postgres	6930	5865(84.63%)	0(0.0%)	6001(86.59%)	0(0.0%)
memcached	133	117(87.96%)	0(0.0%)	117(87.96%)	0(0.0%)
node	20638	15626(75.71%)	0(0.0%)	15863(76.86%)	0(0.0%)

Table 7: The results for call-target analysis using the union combination operator for the *count* policy throughout different optimizations.

6.1.3 Precision Call-site Classification (*count*)

We are going to present two series of experiments and values to find the best possible combination operator for the call-site analysis in the *count* policy.

6.1.4 Without inter-procedural Analysis

Experiment Setup. To choose the best possible combination operator for the call-site analysis in implementing the *count* policy without a backward inter-procedural analysis, we generated data for all eight possible versions of reaching definition analysis.

- Destructive/intersection combination operator with an *analyze* function that does not follow into occurring direct calls see Table 8 for results.
- Destructive/intersection combination operator with an *analyze* function that follows into occurring direct calls see Table 8 for results.
- Union combination operator with an *analyze* function that does not follow into occurring direct calls see Table 9 for results.
- Union combination operator with an *analyze* function that follows into occurring direct calls see Table 9 for results.
- Union combination operator with an *analyze* function that does not follow into occurring direct calls see Table 10 for results.
- Union combination operator with an *analyze* function that follows into occurring direct calls see Table 10 for results.

For each possible version we measured two data points per test-target, the number and ratio of perfect classifications and the number and ratio of problematic classifications, which in this case refers to underestimations.

Results. The results of destructive combination and intersection combination are the same. The true union operator is overall inferior to the union operator. The union operator exhibits the lowest error rate when not following calls and is therefore designated a candidate for the safe version of the call-site combination operator. The destructive/intersection operator exhibits the highest precision (geometric mean for O2: 80.11%) and is therefore designate a candidate for the precision version of the call-site combination operator (we chose the version that follows functions).

O2		not following calls		following calls	
Target	#	perfect	problem	perfect	problem
proftpd	112	100(89.28%)	4(3.57%)	100(89.28%)	4(3.57%)
vsftpd	14	14(100.0%)	0(0.0%)	14(100.0%)	0(0.0%)
lighttpd	48	34(70.83%)	8(16.66%)	34(70.83%)	8(16.66%)
nginx	234	181(77.35%)	27(11.53%)	181(77.35%)	27(11.53%)
mysqld	6671	4758(71.32%)	673(10.08%)	4758(71.32%)	673(10.08%)
postgres	565	484(85.66%)	30(5.3%)	484(85.66%)	30(5.3%)
memcached	47	39(82.97%)	0(0.0%)	39(82.97%)	0(0.0%)
node	8599	5886(68.44%)	489(5.68%)	5886(68.44%)	489(5.68%)

Table 8: The results for call-site analysis using the destructive/intersection combination operator for the *count* without inter-procedural analysis policy throughout different optimizations.

O2		not following calls		following calls	
Target	#	perfect	problem	perfect	problem
proftpd	112	56(50.0%)	0(0.0%)	23(20.53%)	1(0.89%)
vsftpd	14	4(28.57%)	0(0.0%)	1(7.14%)	0(0.0%)
lighttpd	48	2(4.16%)	0(0.0%)	0(0.0%)	4(8.33%)
nginx	234	66(28.2%)	4(1.7%)	56(23.93%)	5(2.13%)
mysqld	6671	1921(28.79%)	112(1.67%)	2044(30.64%)	199(2.98%)
postgres	565	179(31.68%)	0(0.0%)	165(29.2%)	4(0.7%)
memcached	47	30(63.82%)	0(0.0%)	29(61.7%)	0(0.0%)
node	8599	2181(25.36%)	101(1.17%)	2280(26.51%)	143(1.66%)

Table 9: The results for call-site analysis using the union combination operator for the *count* without inter-procedural analysis policy throughout different optimizations.

O2		not following calls		following calls	
Target	#	perfect	problem	perfect	problem
proftpd	112	53(47.32%)	0(0.0%)	20(17.85%)	1(0.89%)
vsftpd	14	4(28.57%)	0(0.0%)	1(7.14%)	0(0.0%)
lighttpd	48	2(4.16%)	0(0.0%)	0(0.0%)	4(8.33%)
nginx	234	71(30.34%)	6(2.56%)	49(20.94%)	5(2.13%)
mysqld	6671	1907(28.58%)	141(2.11%)	1934(28.99%)	199(2.98%)
postgres	565	183(32.38%)	0(0.0%)	141(24.95%)	4(0.7%)
memcached	47	29(61.7%)	0(0.0%)	26(55.31%)	0(0.0%)
node	8599	2330(27.09%)	137(1.59%)	2154(25.04%)	143(1.66%)

Table 10: The results for call-site analysis using the true union combination operator for the *count* without inter-procedural analysis policy throughout different optimizations.

6.1.5 With inter-procedural Analysis

Experiment Setup. To choose the best possible combination operator for the call-site analysis in implementing the *count* policy with a backward inter-procedural analysis, we generated data for all eight possible versions of reaching definition analysis.

- Destructive/Intersection combination operator with an *analyze* function that does not follow into occurring direct calls see Table 11 for results.
- Destructive/Intersection combination operator with an *analyze* function that follows into occurring direct calls see Table 11 for results.
- Union combination operator with an *analyze* function that does not follow into occurring direct calls see Table 12 for results.
- Union combination operator with an *analyze* function that follows into occurring direct calls see Table 12 for results.
- Union combination operator with an *analyze* function that does not follow into occurring direct calls see Table 13 for results.
- Union combination operator with an *analyze* function that follows into occurring direct calls see Table 13 for results.

For each possible version we measured two data points per testtarget, the number and ratio of perfect classifications and the number and ratio of problematic classifications, which in this case refers to underestimations.

Results. The results are essentially the same as in the experiment series without inter-procedural analysis, but with slightly higher precision. The results of destructive combination and intersection combination are the same. The true union operator is overall inferior to the union operator. The union operator exhibts the lowest error rate when not following calls and is therefore designated the safe version of the call-site combination operator, as it is superior in precision (geometric mean for O2: 35.79%) to the non inter-procedural version (geometric mean for O2: 26.55%). The destructive/intersection operator exhibts the highest precision (geometric mean for O2: 87.85%) and is therefore designated the best operator for the precision version of the call-site combination operator (we chose the version that follows functions).

O2		not following calls		following calls	
Target	#	perfect	problem	perfect	problem
proftpd	112	100(89.28%)	4(3.57%)	100(89.28%)	4(3.57%)
vsftpd	14	14(100.0%)	0(0.0%)	14(100.0%)	0(0.0%)
lighttpd	48	34(70.83%)	8(16.66%)	34(70.83%)	8(16.66%)
nginx	234	181(77.35%)	28(11.96%)	181(77.35%)	28(11.96%)
mysqld	6671	4789(71.78%)	673(10.08%)	4789(71.78%)	673(10.08%)
postgres	565	489(86.54%)	30(5.3%)	489(86.54%)	30(5.3%)
memcached	47	44(93.61%)	0(0.0%)	44(93.61%)	0(0.0%)
node	8599	5947(69.15%)	493(5.73%)	5947(69.15%)	493(5.73%)

Table 11: The results for call-site analysis using the destructive/intersection combination operator for the *count* with inter-procedural analysis policy throughout different optimizations.

O2		not following calls		following calls	
Target	#	perfect	problem	perfect	problem
proftpd	112	56(50.0%)	0(0.0%)	53(47.32%)	2(1.78%)
vsftpd	14	4(28.57%)	0(0.0%)	8(57.14%)	0(0.0%)
lighttpd	48	14(29.16%)	0(0.0%)	13(27.08%)	6(12.5%)
nginx	234	81(34.61%)	6(2.56%)	141(60.25%)	13(5.55%)
mysqld	6671	2079(31.16%)	141(2.11%)	2775(41.59%)	353(5.29%)
postgres	565	189(33.45%)	0(0.0%)	315(55.75%)	7(1.23%)
memcached	47	30(63.82%)	0(0.0%)	38(80.85%)	0(0.0%)
node	8599	2418(28.11%)	137(1.59%)	3473(40.38%)	256(2.97%)

Table 12: The results for call-site analysis using the union combination operator for the *count* with inter-procedural analysis policy throughout different optimizations.

O2		not following calls		following calls	
Target	#	perfect	problem	perfect	problem
proftpd	112	53(47.32%)	0(0.0%)	53(47.32%)	2(1.78%)
vsftpd	14	4(28.57%)	0(0.0%)	8(57.14%)	0(0.0%)
lighttpd	48	2(4.16%)	0(0.0%)	13(27.08%)	6(12.5%)
nginx	234	71(30.34%)	6(2.56%)	141(60.25%)	13(5.55%)
mysqld	6671	1907(28.58%)	141(2.11%)	2775(41.59%)	353(5.29%)
postgres	565	183(32.38%)	0(0.0%)	315(55.75%)	7(1.23%)
memcached	47	29(61.7%)	0(0.0%)	38(80.85%)	0(0.0%)
node	8599	2330(27.09%)	137(1.59%)	3457(40.2%)	256(2.97%)

Table 13: The results for call-site analysis using the true union combination operator for the *count* with inter-procedural analysis policy throughout different optimizations.

6.1.6 Precision Call-target Classification (*type*)

We are going to present a series of experiments and values to find the best possible combination operator for the call-target analysis in the *type* policy.

Experiment Setup. To choose the best possible combination operator for the call-target analysis in implementing the *type* policy, we conducted three experiments based on the data of the Precision call-target Classification(*count*) experiment and the proposed implementations for the *type* policy:

exp1 union combination operator with an *analyze* function that does follow into occurring direct calls and a vertical merge that only accepts the first change see Table 14 for results.

exp2 union combination operator with an *analyze* function that does follow into occurring direct calls and a vertical merge that unions all reads until the first write see Table 14 for results.

exp3 union combination operator with an *analyze* function that does follow into occurring direct calls and a vertical merge that intersects all reads until the first write see Table 15 for results.

For each possible version we measured two data points per test-target, the number and ratio of perfect classifications and the number and ratio of problematic classifications, which in this case refers to overestimations.

Result. The series exp2 shows the lowest problem rate of the three series. Regarding precision the values are as follows relatively equal:

- The series exp1 exhibits a geometric mean of 72.42% for precision in O2.
- The series exp2 exhibits a geometric mean of 72.25% for precision in O2.
- The series exp3 exhibits a geometric mean of 72.52% for precision in O2.

Due to exp2 exhibiting the lowest error value, we designate this setup as the best setup for analyzing the call-targets in the *type* policy.

O2		exp1		exp2	
Target	#	perfect	problem	perfect	problem
proftpd	1015	855(84.23%)	11(1.08%)	852(83.94%)	11(1.08%)
vsftpd	318	230(72.32%)	3(0.94%)	230(72.32%)	3(0.94%)
lighttpd	290	239(82.41%)	6(2.06%)	240(82.75%)	3(1.03%)
nginx	921	599(65.03%)	0(0.0%)	594(64.49%)	0(0.0%)
mysqld	9742	5712(58.63%)	327(3.35%)	5705(58.56%)	316(3.24%)
postgres	6930	5291(76.34%)	585(8.44%)	5291(76.34%)	572(8.25%)
memcached	133	104(78.19%)	10(7.51%)	103(77.44%)	10(7.51%)
node	20638	13673(66.25%)	479(2.32%)	13671(66.24%)	438(2.12%)

Table 14: The results for call-target analysis for exp1 and exp2 of the *type* policy throughout different optimizations.

O2		exp3	
Target	#	perfect	problem
proftpd	1015	853(84.03%)	13(1.28%)
vsftpd	318	230(72.32%)	3(0.94%)
lighttpd	290	240(82.75%)	7(2.41%)
nginx	921	599(65.03%)	0(0.0%)
mysqld	9742	5714(58.65%)	338(3.46%)
postgres	6930	5280(76.19%)	600(8.65%)
memcached	133	105(78.94%)	10(7.51%)
node	20638	13693(66.34%)	530(2.56%)

Table 15: The results for call-target analysis for exp3 of the *type* policy throughout different optimizations.

6.1.7 Precision Call-site Classification (*type*)

We are going to present a series of experiments and values to find the best possible combination operator for the call-site analysis in the *type* policy.

Experiment Setup. To choose the best possible combination operator for the call-site analysis in implementing the *type* policy, we conducted three experiments based on the data of the two Precision call-site Classification(*count*) experiments and the proposed implementations for the *type* policy:

exp1 intersection combination operator that intersects when both paths are set with an *analyze* function that does follow into occurring direct calls with backward inter-procedural analysis see Table 16 for results.

exp2 intersection combination operator that unions when both paths are set with an *analyze* function that does follow into occurring direct calls with backward inter-procedural analysis see Table 16 for results.

exp3 union combination operator with an *analyze* function that does not follow into occurring direct calls with backward inter-procedural analysis see Table 17 for results.

For each possible version we measured two data points per testtarget, the number and ratio of perfect classifications and the number and ratio of problematic classifications, which in this case refers to underestimations.

Result. The series exp3 easily holds the lowest rate of problematic classification and therefore we designate it the setup for a safe implementation of call-site analysis for the *type* policy. The results for the series exp1 and the series exp2 are the same exhibiting a geometric mean of 57.30% for precision in O2. Therefore it does not matter which of the two setups we choose for the precision implementation of call-site analysis for the *type* policy.

O0		exp1		exp2	
Target	#	perfect	problem	perfect	problem
proftpd	112	100(89.28%)	4(3.57%)	100(89.28%)	4(3.57%)
vsftpd	14	14(100.0%)	0(0.0%)	14(100.0%)	0(0.0%)
lighttpd	48	33(68.75%)	9(18.75%)	33(68.75%)	9(18.75%)
nginx	234	151(64.52%)	59(25.21%)	151(64.52%)	59(25.21%)
mysqld	6671	4189(62.79%)	918(13.76%)	4187(62.76%)	918(13.76%)
postgres	565	443(78.4%)	31(5.48%)	442(78.23%)	31(5.48%)
memcached	47	5(10.63%)	10(21.27%)	5(10.63%)	10(21.27%)
node	8599	4829(56.15%)	1227(14.26%)	4828(56.14%)	1227(14.26%)

Table 16: The results for call-site analysis for exp1 and exp2 of the *type* policy throughout different optimizations.

O2		exp3	
Target	#	perfect	problem
proftpd	112	56(50.0%)	0(0.0%)
vsftpd	14	4(28.57%)	0(0.0%)
lighttpd	48	14(29.16%)	1(2.08%)
nginx	234	57(24.35%)	38(16.23%)
mysqld	6671	1602(24.01%)	418(6.26%)
postgres	565	154(27.25%)	1(0.17%)
memcached	47	0(0.0%)	10(21.27%)
node	8599	1659(19.29%)	876(10.18%)

Table 17: The results for enhanced call-site analysis for exp3 of the *type* policy throughout different optimizations.

6.2 RQ2: Effectiveness of TYPESHIELD

In this section we need just one or two Table similar to what TypeArmor contains, first we need to define the fields which make most sense.

We are now going to evaluate the effectiveness of TYPESHIELD leveraging the result of several experiment runs: First we are going to establish a baseline using the data collected from our Clang/LLVM pass, which are the theoretical limits our implementation can reach. Second we are going to evaluate the effectiveness of our *count* policy and third we are going to evaluate the effectiveness of our *type* policy. At last we are going to look at the effect our address taken analysis had on the results.

6.2.1 Theoretical Limits.

We explore the theoretical limits regarding the effectiveness of the *count* and *type* policies by relying on the collected ground truth data, which is essentially assume perfect classification.

Experiment Setup Based on the type information collected by our Clang/LLVM pass, we conducted two experiment series:

- We derived the *count* schema using the ground truth and calculated the available number of call-targets for each call-site, see table 18 for results.
- We derived the *type* schema using the ground truth and calculated the available number of call-targets for each call-site, see table 18 for results.

For each series we collected three data points per test target, the average number of call-targets per call-site, the standard deviation σ and the median.

Results.

- The theoretical limit of the *count* schema has an overall geometric mean of 959 possible call-targets, which is 53.75% of the geometric mean of total available call-targets.
- The theoretical limit of the *count* schema has an overall geometric mean of 823 possible call-targets, which is 46.10% of the geometric mean of total available call-targets.

When compared, the theoretical limit of the *type* policy allows about 15% less available call-targets in the geomean in O2 than the limit of the *count* policy.

O0	AT	*count**			*type**		
Target		limit (mean $\pm$ σ)		median	limit (mean $\pm$ σ)		median
proftpd	1171	927.71	$\pm$ 203.99	941.0	867.43	$\pm$ 206.78	941.0
vsftpd	391	300.0	$\pm$ 73.0	300.0	159.0	$\pm$ 68.0	159.0
lighttpd	354	216.33	$\pm$ 77.55	154.0	198.41	$\pm$ 68.89	154.0
nginx	1098	611.94	$\pm$ 294.31	636.0	611.94	$\pm$ 294.31	636.0
mysqld	13020	8421.79	$\pm$ 2658.56	8994.0	7812.74	$\pm$ 2377.9	8994.0
postgres	9273	6549.07	$\pm$ 1720.76	6868.0	6039.57	$\pm$ 1942.2	6868.0
memcached	233	211.59	$\pm$ 23.55	229.0	154.42	$\pm$ 19.28	140.0
node	20638	12841.45	$\pm$ 4441.18	14062.0	10977.88	$\pm$ 4143.54	14062.0

Table 18: The results of comparing theoretical limits for the different restriction policies throughout different optimizations.

6.2.2 TYPESHIELD implementation of the *count* policy

We explore the effectiveness of our safe and precise versions for the *count* policy implementation.

Experiment Setup. We setup our two experiment series based on our previous evaluations regarding the classification precision for the *count* policy.

- For the safe version, we chose the union combination operator, with *analyse* that follows into occurring direct calls to calculate the call-target invariant. For the call-site invariant, we chose the union operator that does not follow into occurring direct calls with backwards inter-procedural analysis. See table 19 for results.
- For the precise version, we chose the union combination operator, with *analyse* that follows into occurring direct calls to calculate the call-target invariant. For the call-site invariant, we chose the intersection operator that follows into occurring direct calls with backwards inter-procedural analysis. See table 19 for results.

For each series we collected three data points per test target, the average number of call-targets per call-site, the standard deviation σ and the median.

Results.

- The average number of available targets provided by the safe implementation of the *count* schema has an overall geometric mean of 1283 possible call-targets, which is 71.87% of the geometric mean of total available call-targets. This is 33.78% more than the theoretical limit of available call-targets per call-site.
- The average number of available targets provided by the precise implementation of the *count* schema has an overall geometric mean of 1030 possible call-targets, which is 57.70% of the geometric mean of total available call-targets. This is 7.40% more than the theoretical limit of available call-targets per call-site.

When compared, the precise implementation of of the *count* policy allows about 20% less available call-targets in the geomean in O2 than the safe implementation of the *count* policy.

O2	AT	*count* safe		*count* prec	
Target		limit (mean $\pm$ σ)	median	limit (mean $\pm$ σ)	median
proftpd	1015	954.41 $\pm$ 27.08	961.0	862.14 $\pm$ 158.04	935.0
vsftpd	318	298.14 $\pm$ 29.45	306.0	224.57 $\pm$ 51.49	192.0
lighttpd	290	270.64 $\pm$ 23.16	286.0	192.18 $\pm$ 96.06	227.0
nginx	921	801.91 $\pm$ 210.01	913.0	559.46 $\pm$ 262.55	913.0
mysqld	9742	8560.9 $\pm$ 1323.59	8550.0	6889.73 $\pm$ 2289.89	9742.0
postgres	6930	6074.52 $\pm$ 525.69	6012.0	4800.62 $\pm$ 1718.52	5262.0
memcached	133	126.08 $\pm$ 5.04	129.0	120.91 $\pm$ 12.58	129.0
node	20638	18221.13 $\pm$ 2479.72	14996.0	15325.62 $\pm$ 4407.39	14996.0

Table 19: The results of comparing *count* safe and precision implementation throughout different optimizations.

6.2.3 TYPESHIELD implementation of the *type* policy

We explore the effectiveness of our safe and precise versions for the *type* policy implementation.

Experiment Setup. We setup our two experiment series based on our previous evaluations regarding the classification precision for the *type* policy.

- For the safe version, we chose the union combination operator with an *analyze* function that does follow into occurring direct calls and a vertical merge that unions all reads until the first write to calculate the call-target invariant. For the call-site invariant, we chose the union combination operator with an *analyze* function that does not follow into occurring direct calls with backward inter-procedural analysis. See table 20 for results.
- For the precise version, we chose union combination operator with an *analyze* function that does follow into occurring direct calls and a vertical merge that unions all reads until the first write to calculate the call-target invariant. For the call-site invariant, we chose the intersection combination operator that intersects when both paths are set with an *analyze* function that does follow into occurring direct calls with backward inter-procedural analysis. See table 20 for results.

For each series we collected three data points per test target, the average number of call-targets per call-site, the standard deviation σ and the median.

Results.

- The average number of available targets provided by the safe implementation of the *type* schema has an overall geometric mean of 1144 possible call-targets, which is 64.08% of the geometric mean of total available call-targets. This is 39.00% more than the theoretical limit of available call-targets per call-site.
- The average number of available targets provided by the precise implementation of the *type* schema has an overall geometric mean of 907 possible call-targets, which is 50.81% of the geometric mean of total available call-targets. This is 10.20% more than the theoretical limit of available call-targets per call-site

When compared, the precise implementation of of the *count* policy allows about 21% less available call-targets in the geomean in O2 than the safe implementation of the *count* policy.

O2 Target	AT	*type* safe limit (mean ± σ)			median	*type* prec limit (mean ± σ)			median
proftpd	1015	895.17	±	122.6	935.0	816.77	±	176.59	935.0
vsftpd	318	246.14	±	81.88	306.0	172.85	±	30.26	192.0
lighttpd	290	249.08	±	53.1	152.0	174.18	±	93.65	148.0
nginx	921	732.73	±	238.77	618.0	503.72	±	234.88	618.0
mysqld	9742	8082.58	±	1528.73	9742.0	6505.05	±	2143.58	2155.0
postgres	6930	5678.06	±	1157.81	6012.0	4443.7	±	1781.19	5262.0
memcached	133	96.23	±	14.3	90.0	92.55	±	12.8	90.0
node	20638	16558.62	±	3566.3	17944.0	13928.35	±	4307.4	14996.0

Table 20: The results of comparing *type* safe and precision implementation throughout different optimizations.

6.2.4 Effect of our AddressTaken Analysis

We are providing experiment results and an evaluation regarding the impact of our address taken analysis on the theoretical limits and our various policy implementations, namely the safe and precise versions of the *count* and *type* policies.

Experiment Setup. We conducted the same three experiments as before but with the initial set of call-targets filtered by our implementation of address taken analysis:

- We setup the same experiment regarding theoretical limits as described in subsection 6.2.1 but with restricting the possible call-targets to only address taken functions. The results are presented in table 21.
- We setup the same experiment regarding the *count* policy as described in subsection 6.2.1 but with restricting the possible call-targets to only address taken functions. The results are presented in table **??**.
- We setup the same experiment regarding the *type* policy as described as in subsection 6.2.1 but with restricting the possible call-targets to only address taken functions. The results are presented in table **??**.

For each series we collected three data points per test target, the average number of call-targets per call-site, the standard deviation σ and the median.

Results. First of all we observed an overall reduction of the geometric mean of overall available call-targets to 64% before our policies were applied. Notable outliers are memcached, which had a 90% reduction of available call-targets and vsftpd, which even achieved a 97% reduction in available call-targets.

- The theoretical available targets were overall reduced to 25.96% in the *count* policy case (geometric mean of 249 in O2) and to 27.27% in the *type* policy case (geometric mean of 224 in O2). The difference between their geometric means shrank from about 15% to about 10%.
- The average targets provided by the safe and the precise implementation of the *count* policy were reduced to 25.33% (geometric mean of 325 in O2) and 25.92% (geometric mean of 267) respectively. The difference between their geometric means shrank from about 20% to about 18%.
- The average targets provided by the safe and the precise implementation of the *count* policy were reduced to 26.13% (geometric mean of 299 in O2) and 26.79% (geometric mean of 243) respectively. The difference between their geometric means shrank from about 20% to about 18%.

When comparing the precision focused implementations of our *type* policy and our *count*, we observe that the difference between them shrank to about 9% when comparing their geometric means.

Overall we were able to reduce the number of available call-targets per call-site from 1785 to 243 using our precision focused implementation of the *type* policy, which is an overall reduction to 13.61%.

6.3 RQ3: Runtime Performance Overhead

In this section we need one or two Table similar to what TypeArmor contains, first we need to define the fields which make most sense.

Here we measure how much performance overhead the instrumentation incurs. Here we measure with the same SPEC2006 programs that was used in the TypeArmor paper. spec 2006.

add a binary patch that does not crash none of the programs from SPEC2006.

need a table with all the results for each of the SPEC2006 programs and a bar diagram

6.4 RQ4: Instrumentation Overhead

here we need a bar chart, see TypeArmor paper.

Here we measure how much the binaries increased in size after the instrumentation was added to the binaries.

Measure the size (in bytes) of the SPEC2006 testes in RQ3 before and after adding all the patches

6.5 RQ5: Security Level of TypeShield

here we need a a CDF figure, see the TypeAmor paper w.r.t. security analysis in the evaluation. The buckets have to be defined first, reason about if they make sense!

We look at our implementation conceptually and assess qualitatievly whether our implementation can interfere with various classes of attacks.

see the TypeAmor paper w.r.t. security analysis in the evaluation, a CDF figure is here required.

O2	AT	*count**			*count*			*type**			*type*		
Target		limit (mean ± σ)		median	limit (mean ± σ)		median	limit (mean ± σ)		median	limit (mean ± σ)		median
proftpd	390	348.55 ± 59.5		369.0	366.16 ± 48.39		382.0	333.11 ± 72.18		369.0	356.23 ± 59.79		382.0
vsftpd	10	7.14 ± 1.8		6.0	7.14 ± 1.8		6.0	5.42 ± 0.9		6.0	5.42 ± 0.9		6.0
lighttpd	59	36.87 ± 15.13		47.0	39.31 ± 21.2		50.0	34.55 ± 14.03		33.0	36.39 ± 20.72		36.0
nginx	543	313.87 ± 150.2		266.0	347.76 ± 152.94		543.0	313.87 ± 150.2		266.0	316.36 ± 135.22		362.0
mysqld	5879	4103.38 ± 1054.55		3167.0	4615.36 ± 1165.18		5879.0	3882.72 ± 937.79		3167.0	4427.54 ± 1097.93		1885.0
postgres	2490	2056.06 ± 670.46		2284.0	2083.21 ± 640.18		2332.0	1885.18 ± 812.27		2284.0	1954.39 ± 722.62		2332.0
memcached	14	12.27 ± 2.35		14.0	13.12 ± 1.33		14.0	10.27 ± 0.96		11.0	11.53 ± 1.31		12.0
node	7528	5068.77 ± 1547.25		5522.0	5995.79 ± 1419.43		5940.0	4363.56 ± 1497.97		5522.0	5500.11 ± 1397.84		5940.0

Table 21: The results of comparing our implementation results with the theoretical limits for the different restriction policies combined with an an address taken analysis for optimization level O2.

6.6 RQ6: Add another evaluation dimention we did not think off, maybe "RQ6: TypeShield Deployments" (if it is easy or hard to deploy our tool in comparison with TypeArmor.)

need to define first the question.

7 Related Work

Type-Inference on Executables. Recovering variable types from executable programs is very hard in general for several reasons. First, the quality of the disasembly can very much from used framework to another. TYPESHIELD is based on DynInst and the quality of the execuatble disasembly fits our needs. For a more comprehensive review on the capabilities of DynInst and other tools we advice the reader to have a look at [6]. Second, alias analysis in binares is undecidable in theory and intractable in practice [32]. There are several most promising tools such as: Rewards [28], BAP [11], SmartDec [19], and Divine [7]. These tools try with more or less success to recover type information from binary programs with different goals. Typical goals are: *i)* full program reconstruction (binary to code convertion, reversing), *ii)* checking for buffer overflows, *iii)* integer overflows and other types of memory corruptions. For a more exhaustive review of such tools we advice the reader to have a look at the review of Caballero et al. [13]. Intresting to notice is that the code from only a few of this tools is available.

While smartde seemed promising due to its simple type lattice that we wanted to leverage for our classification schema. Its integration into our DynInst based environment was not successful mostly for time constraints, as it was deemed to time consuming to extract the whole machinery and implemnt an interface to the DynInst disassembler. Therefore we finally implemented our own version of type analysis and only focused on the wideness of the types, resulting in a simpler lattice than we initially wanted.

Mitigation of Code-Reuse Attacks. In the last couple of years researchers have provided many versions of new Code Reuse Attacks (CRAs). These new attacks were possible since DEP [30] and ASLR [39] were successfully bypassed mostly based on Return Oriented Programming (ROP) [12, 24, 41] on one hand and on the other hand due to the discovery of new exploitable hardware and software primitives.

ROP started to present itself in the last couple of years in many faceted ways such as: Jump Oriented Programming (JOP) [9, 16, 18] which uses jumps in order to divert the control flow to the next gadget and Call Oriented Programming (COP) [15] which uses calls in order to chain gadgets together. CRAs have many manifestations and it is out of

scope of this work to list them all.

On one hand, CRAs can be mitigated in general in the following ways: *(i)* binary instrumentation, *(ii)* source code recompilation and *(iii)* runtime application monitoring. On the other hand, there is a plethora of tools and techniques which try to enforce CFI based primitives in executables, source code and during runtime. Next we briefly present the solution landscape together with the approaches and the techniques on which these are based: *(a)* fine-grained CFI with hardware support, PathArmor [43], *(*(b)) coarse-grained CFI used for binary instrumentation, CCFIR [48], *(c)* coarse-grained CFI based on binary loader, CFCI [50] *(d)* fine-grained code randomization, O-CFI [31], *(e)* cryptografy with hardware support, CCFI [29], *(f)* ROP stack pivoting, PBlocker [38], *(g)* canary based protection, DynaGuard [36], *(h)* runtime and hardware support based on a combination of LBR, PMU and BTS registers CFIGuard [45], and *(i)* source code recompilation with CFI and/or randomization enforcement against JIT-ROP attacks, MCFI [33], RockJIT [34] and PiCFI [35].

The above list is not exhaustive and new protection techniques can be obtained by combining available techniques or by using newly available hardware features or software exploits. However, none of the above techniques and tools can mitigate against COOP attacks.

Mitigation of Forward-Edge based Attacks. Recursive-COOP [17], COOP [40] and Subversive-C [27]. are advanced CRAs since these attacks can not be addressed: *i)* with shadow stacks techniques (i.e., do not violate the caller/calle convention), *ii)* coarse-grained Control-Flow Integrity (CFI) [4, 5] techniques are useless against these attacks, *iii)* hardware based approaches such as Intel CET [3] can not mitigate this attack for the same reason as in *i)*, and *iv)* with OS-based approaches such as Windows Control Flow Guard [2] since the precomputed CFG does not contain edges for indirect call sites which are explicitly exploited during the COOP attack. However, the following tools can protect against COOP attacks:

Source code based. Indirect call site targets are checked based on vTable integrity. Different types of CFI policies are used such as in the following tools: SafeDispatch [21], IFCC/VTV [42] LLVM and GCC compiler. Additionally, the Redactor++ [17] uses randomization vTrust [46] checks call target function signatures, CPI [25] uses a memory safety technique in order to protect against the COOP attack.

There are several source code based tools which can successfully protect against the COOP attack. Such tools are: ShrinkWrap [20], IFCC/VTV [42], SafeDispatch [21], vTrust [46], Readactor++ [17], CPI [25] and the tool presented by Bounov et al. [10]. These tools profit from high precision since they have access to the full semantic context of the program though the scope of the compiler on which they are based. Because of this reason these tools target mostly other types of security problems than binary-based tools address. For example some last advancec in compile based protection against code reuse attacks address mainly performance issues. Currently, most of the above presented tools are only forward edge enforcers of fine-grained CFI policies with an overhead from 1% up to 15%.

We are aware that there is still a long research path to go until binary based techniques can recuperate program based semantic information from executable with the same precision as compiler based tools. These path could be even endless since compilers are optimized for speed and are designed to remove as much as possbile semantic information from an executable in order to make the program run as fast as possible. In light of this fact, TYPESHIELD is another attempt to recuperate just the needed semantic information (types and number of function parameters from indirect call sites) in order to be able to enforce a precise and with low overhead primitive against COOP attacks.

Rather than claiming that the invariants offered by TYPESHIELD are suffiecient to mitigate all versions of the COOP attack we take a more conservative path by claiming that TYPESHIELD further raises the bar w.r.t. what is possible when defending against COOP attacks on the binary level.

Binary based. vTable protection is addressed through binary instrumentation in tools such as: vfGuard [37], vTint [47]. However, none of these tools can help to mitigate against COOP. The only binary based tool which we are aware of that can mitigate protect against COOP is TypeArmor [44]. TypeArmor uses a fine-grained CFI policy based on caller (only indirect call sites)/callee matching which consists in checking during runtime if the number of provided and needed parameters match.

TYPESHIELD is most similar to TypeArmor [44] since we also enforce strong binary-level invariants on the number of function paramters. TYPESHIELD similarly to TypeArmor targets exclusive protection against advanced exploitation techniques which can bypass fine-grained CFI schemes and VTable protections at the binary level.

However, TYPESHIELD offers a better restriction of call targets to call sites, since we not only restrict based on the number of parameters but also on the wideness of their types. This results in much smaller buckets that in turn can only target a smaller subset of all address taken functions. However, we rely for that on the variety of parameter types and when there is none, we will degrade into a parameter count policy.

Runtime based. "There is something available out there but I can not use it" *Anonymous*. Long story short conclusion: There are several promising runtime-based line of defenses against advanced CRAs but none of them can successfully protect against the COOP attack.

IntelCET [3] is based on, `ENDBRANCH`, a new CPU instruction which can be used to enforce an efficient shadow stack mechanism. The shadow stack can be used to check during program execution if caller/return pairs match. Since the COOP attack reuses whole functions as gadgets and does not violate the caller/return convention than the new feature provided by intel is useless in the face of this attack. Nevertheless other highly notorious CRAs may not be possible after this feature will be implemented main stream in OSs and compilers.

Windows Control Flow Guard [2] is based on a user-space and kernel-space components which by working closely to-

gether can enforce an efficient fine-grained CFI policy based on a precomputed CFG. These new feature available in Windows 10 can considerably rise the bar for future attacks but in our opinion advanced CRAs such as COOP are still possible due the typical characteristics of COOP.

PathArmor [43] is yet another tool which is based on a precomputed CFG and on the LBR register which can give a string of 16 up to 32 pairs of from/to addressed of different types of indirect instructions such as `call`, `ret`, and `jump`. Because of the sporadic query of the LBR register (only during invocation of certain function calls) and because of the sheer amount of data which passes through the LBR register this approach has in our opinion a fair potential to catch different types of CRAs but we think that against COOP this tool can not be used. First, because of the fact that the precomputed CFG does not contain edges for all possible indirect call sites which are accessed during runtime and second, the LBR buffer can be easily triked by adding legitimate indirect call sites during the COOP attack.

8 Discussion

Comparison with TypeArmor. We are looking at two sets of results. First of all, we compare the overall precision of our implementation of the COUNT policy with the results from TypeArmor to set the perspective for the precision of our TYPE policy. We cannot compare data regarding overestimations of calltargets or underestimations of callsites, as TypeArmor did not provide sufficient data. The second point of comparison is the reduction of calltargets per callsite, however, this comparison is rather crude, as we most surely do not have the same measuring environment and not sufficient data to infer its quality.

Precision of Classification. TypeArmor reports a geometric mean of 83.26% for the perfect classification of calltargets regarding parameter count in optimization level O2, which compares rather well to our result of 82.24%. Regarding the perfect classification of callsites we report a geometric mean of 81.6% perfect classification regarding parameter count, while TypeArmor reprots a geometric mean of 79.19%. Howevver we also have a geometric mean of about 7% regarding underestimations in the callsite classification with an upper bound of 16%, while TypeArmor reports that it does not incur underestimations in their callsites. Now, for our type based classification we incur the cost for two error sources. First, the error from the parameter count classification, which we base our type analysis on and second for the type analysis itself. The numbers for the perfect classification of calltargets regarding parameter types we report a 72.25% geometric mean of perfect classification, which is 87.85% of our precision regarding parameter counts. However we report a geometric mean of 57.36% for perfect classification of callsites, which altough seemingly low, is still 69.74% of our precsion regarding parameter counts.

Reduction of Available Calltargets While our count based precision focused implementation achieves a reduction in the same ballpark as TypeArmour regarding our test targets, lets us believe that our implementation of their classification schema is a sufficient approximation to compare against. However, we cannot safely compare those numbers, as the information regarding their test environment are rather sparse and the only data available is the median, which in our opinion does discard valuable information from the actual result set. This is the main reason we implemented an approximation, because we needed more metrics to compare TYPESHIELD and TypeArmor regarding calltargets. Using average and sigma, we can report that our precision focused type based classification can reduce the number of calltargets, by up to 20% more than parameter number based classification with an overall reduction of about 9%.

TypeArmor Discrepancies. As we have no access to source code of TypeArmor, we habe implemented an approximation of TypeArmor. Using this approximation we found some discrepancies between the data that we collected and data that was presented. A minor discrepancy between our results and the results of TypeArmor is that, while they basically implemented what we call a destructive merge operator for the liveness analysis. However, our data suggestes that this operator is marginally inferior to the union pathmerge operator, when we compared them in our implementation. A major concern is the classification of calltargets, while we were able to reduce the number of overerstimations of calltargets regarding parameter counts to essentially 0, the number of underestimations of calltarget did stay at a geometric mean of 7%. This error rate is rather large when compared to the reported 0% underestimation of TypeArmor, however we are not entirely sure what has caused this discrepancy. A possibility is the differing test environments, or a bug within our implementation that we are not aware of, or simply reaching defintions analysis alone is not the best possible algorithm for this particular problem.

Improving TYPESHIELD. To improve our type analysis, we see atleast two possibilities. Incorporating refined dataflow analysis and expanding the scope to also include memory. The main point of improvement is not the precision but for now more importantly the reduction of underestimations in the callsite analysis.

To refine the dataflow analysis, we propse the actual tracking of data values and simple operations, as these can be used to better differentiate the actual wideness stored within the current register. The highest gain, we see here would be the establishment of upper and lower bounds regarding values within the register, which would allow for more sophisticated callsite and calltarget invariants. Essentially we would have to resort to symbolic execution or some other sort of precise abstract interpretation.

Expanding the scope to also include memory, is another possible way of improving the type analysis, as it would allow us to distinguish normal 32 or 64 bit values and pointer addresses. Although we already have a limited approach of that in our reaching implementation, we still see room for improvement, as we only check whether a value is within one of three binary sections or 0.

Limitations of TYPESHIELD. First of all, we are limited

by the capabilities of the DynInst Instrumentation Environment, the main problem, we are facing here is that non returning functions like exit are not detected reliably in some cases, which is why we were not able to test the Pure-FTP server, as it heavily relies on these functions. The problem is that those non returning functions usually appear as a second branch within a function that occurs after the normal control flow, causing basic blocks from the following function to be attributed to the current function. This results in a malformed control flow graph and erroneous attribution of callsites and problematic misclassifications for both calltargets and callsites.

Another limitation of TYPESHIELD is it reliance on variety within the binary, in particular we rely on the fact that functions use more than only 64bit values or pointers within their parameter list. Should this scenario occur, our analysis has nothing to work with and essentially degrades into a parameter count based implementation. Thankfully this occurrence is quite rare, as we experienced within our experiments. When working based on source level information, we could not detect a difference between our TYPE and a COUNT policies. However when leveraging our tool, we were able to detect differences, which reinforces the fact, that we do not rely on declaration of parameters but usage of those.

9 Future Work

Structural matching capability. Improving the structural matching capability is in our opinion the most important further venue of research, as we need a reliable way to match a ground truth against the resulting binary. This is important, because it is a prerequisite to the ability to generate reliable measurements and reduces the current uncertainy (we rely on the number of calltargets per callsite to match callsites and furthermore assume that the order within ground truth and binary is the same).

Better callsite analysis. Finding a better suited callsite analysis would present itself as another important possibility, as we still have a relatively high—up to16%—number of underestimated callsites. However, this venue should only be attempted after significant improvements to the structural matching of callsites.

Better patching schema. Devising a patching schema that is based on Dyninst functionality, which allows annotation of calltargets so they can hold at least 4 bytes of arbitrary data. This is required to hold the type data that we generate using our classification. Keeping the runtime overhead of said patching schema low should be the second goal of this venue after satisfying stability.

Expanding to return values. Expanding our schema to return values is another viable venue of further work, as we were not able to reliably reduce the number of problematic classification regarding the return values of functions to managable levels. Should one attempt this, it should be noted that the responsibilities of callsites and calltargets are reversed in this case: The callsite requires return value wideness, while the calltarget needs to provide it.

Using pointer/memory analysis. Introducing pointer/memory analysis to distinguish simple 32/64bit values and actual addresses to even further restrict the possible number of calltargets per callsite. This would require more precise dataflow analysis, as in calculating value possibilities for registers at each instruction.

10 Conclusion

The family of forward indirect call based attacks which can manifest due to a series of factors such as a memory corruption, binary layout leackage and presence of useful gadgets in sufficiently large executables is a serious security threat. We have developed TYPESHIELD, a runtime based fine-grained CFI enforcing tool which can precissely filter legitimate from illegitimate indirect forward calls in binaries. It uses a novel run-time type checking technique based on function parameter type checking and parameter counting in order to efficiently filter-out legitimate and illegitimate forward edges. TYPESHIELD provides a more precise analysis than existing approaches with a comparable performance overhead. We have implemented TYPESHIELD and applied it to real software such as: web servers, and FTP servers. We demonstrated through extensive experiments and comparisons with related software that TYPESHIELD has higher precision and comparable performance overhead than the existing state-of-the-art tools. To date, we were able to provide a more precise technique than parameter count based techniques by reducing the average target count of up to 20%. This results in a more precise analysis and a considerably reduced attack surface.

Todo list

References

[1] DynamoRIO. http://dynamorio.org/home.html.

[2] Windows Control Flow Guard. http://msdn.microsoft.com/en-us/library/windows/desktop/mt637065(v=vs.85).aspx.

[3] Intel Control-flow Enforcement Technology (CET). http://blogs.intel.com/evangelists/2016/06/09/intel-release-new-technology-specifications-protect-rop-attacks/.

[4] ABADI, M., BUDIU, M., ERLINGSSON, Ú., AND LIGATTI, J. Control Flow Integrity. In *Proceedings of the 12th ACM Conference on Computer and Communications Security (CCS),* (2005).

[5] ABADI, M., BUDIU, M., ERLINGSSON, Ú., AND LIGATTI, J. Control Flow Integrity Principles, Implementations, and Applications. In *ACM Transactions on Information and System Security (TISSEC),* (2009).

[6] ANDRIESSE, D., CHEN, X., VEEN, V. V. D., SLOWINSKA, A., AND BOS, H. An In-Depth Analysis of Disassembly on Full-Scale x86/x64 Binaries. In *Proceedings of the USENIX Conference on Security (USENIX SEC),* (2016).

[7] BALAKRISHNAN, G., AND REPS, T. DIVINE: Discovering Variables in Executables. In *International Conference on Verification, Model Checking, and Abstract Interpretation (VMCAI),* (2007).

[8] BERNAT, A. R., AND MILLER, B. P. Anywhere, Any-Time Binary Instrumentation. In *Proceedings of the 10th ACM SIGPLAN-SIGSOFT workshop on Program analysis for software tools, (PASTE),* (2011).

[9] BLETSCH, T., JIANG, X., FREEH, V. W., AND LIANG, Z. Jump-Oriented Programming: A New Class of Code-Reuse Attack. In *ACM Symposium on Information, Computer and Communications Security (ASIACCS),* (2011).

[10] BOUNOV, D., GÖKHAN KICI, R., AND LERNER, S. Protecting C++ Dynamic Dispatch Through VTable Interleaving. In *Symposium on aravindNetwork and Distributed System Security (NDSS),* (2016).

[11] BRUMLEY, D., JAGER, I., AVGERINOS, T., AND SCHWARTZ, E. J. BAP: A Binary Analysis Platform. In *Proceedings of Computer AidedVTint: Protecting Virtual Function Tab Verification (CAV),* (2011).

[12] BUCHANAN, E., ROEMER, R., SHACHAM, H., AND SAVAGE, S. When Good Instructions Go Bad: Generalizing Return-oriented Programming to RISC. In *ACM Conference on Computer and Communications Security (CCS),* (2008).

[13] CABALLERO, J., AND LIN, Z. Type Inference on Executables. In *ACM Computing Surveys (CSUR),* (2016).

[14] CARLINI, N., BARRESI, A., PAYER, M., WAGNER, D., AND GROSS, T. R. Control-Flow Bending: On the Effectiveness of Control-Flow Integrity. In *Proceedings of the USENIX conference on Security (USENIX SEC),* (2015).

[15] CARLINI, N., AND WAGNER, D. ROP is still dangerous: Breaking Modern Defenses. In *Proceedings of the USENIX conference on Security (USENIX SEC),* (2014).

[16] CHECKOWAY, S., DAVI, L., DMITRIENKO, A., SADEGHI, A.-R., SHACHAM, H., AND WINANDY, M. Return-oriented Programming Without Returns. In *ACM Conference on Computer and Communications Security (CCS),* (2010).

[17] CRANE, S., VOLCKAERT, S., SCHUSTER, F., LIEBCHEN, C., LARSEN, P., DAVI, L., SADEGHI, A.-R., HOLZ, T., DE SUTTER, B., AND FRANZ, M. It's a TRaP: Table Randomization and Protection against Function-Reuse Attacks. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS),* (2015).

[18] DAVI, L., DMITRIENKO, A., SADEGHI, A.-R., AND WINANDY, M. Return-Oriented Programming without Returns on ARM. In *Technical report, Technical Report HGI-TR-2010-002, Ruhr-University Bochum,* (2010).

[19] FOKIN, A., DEREVENETS, Y., CHERNOV, A., AND TROSHINA, K. SmartDec: Approaching C++ decompilation. In *Working Conference on Reverse Engineering (WCRE),* (2011).

[20] HALLER, I., GOKTAS, E., ATHANASOPOULOS, E., PORTOKALIDIS, G., AND BOS, H. ShrinkWrap: VTable Protection Without Loose Ends. In *Annual Computer Security Applications Conference (ACSAC),* (2015).

[21] JANG, D., TATLOCK, T., AND LERNER, S. SafeDispatch: Securing C++ Virtual Calls from Memory Corruption Attacks. In *Symposium on Network and Distributed System Security (NDSS),* (2014).

[22] JTC1/SC22WG21, I. ISO/IEC 14882:2013 Programming Language C++ (N3690). https://isocpp.org/files/papers/N3690.pdf.

[23] KHEDKER, U., SANYAL, A., AND SATHE, B. *Data flow analysis: Theory and Practice*. CRC Press, 2009.

[24] KORNAU, T. Return-Oriented Programming for the ARM Architecture. http://www.zynamics.com/downloads/kornau-tim--diplomarbeit--rop.pdf.

[25] KUZNETSOV, V., SZEKERES, L., PAYER, M., CANDEA, G., SEKAR, R., AND SONG, D. Code-Pointer Integrity. In *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation (OSDI),* (2014).

[26] LEE, B., SONG, C., KIM, T., AND LEE, W. Type Casting Verification: Stopping an Emerging Attack Vector. In *Proceedings of the USENIX Conference on Security (USENIX SEC),* (2015).

[27] LETTNER, J., KOLLENDA, B., HOMESCU, A., LARSEN, P., SCHUSTER, F., DAVI, L., SADEGHI, A.-R., HOLZ, T., AND FRANZ, M. Subversive-C: Abusing and Protecting Dynamic Message Dispatch. In *USENIX Annual Technical Conference (USENIX ATC),* (2016).

[28] LIN, Z., ZHANG, X., AND XU, D. Automatic Reverse Engineering of Data Structures from Binary Execution. In *Symposium on Network and Distributed System Security (NDSS),* (2010).

[29] MASHTIZADEH, A. J., BITTAU, A., BONEH, D., AND MAZIÈRES, D. CCFI: Cryptograhically Enforced Control Flow Integrity. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS),* (2015).

[30] MICROSOFT, C. T. F. I. M. W. X. S. P. . https://technet.microsoft.com/en-us/library/bb457151.aspx.

[31] MOHAN, V., LARSEN, P., BRUNTHALER, S., HAMLEN, K. W., AND FRANZ, M. Opaque Control-Flow Integrity. In *Symposium on Network and Distributed System Security (NDSS),* (2015).

[32] MYCROFT, A. Lecture Notes. https://www.cl.cam.ac.uk/~am21/papers/sas07slides.pdf.

[33] NIU, B., AND TAN, G. Modular Control-Flow VTint: Protecting Virtual Function TabIntegrity. In *ACM Conferece on Programming Language Design and Implementation (PLDI),* (2014).

[34] NIU, B., AND TAN, G. RockJIT: Securing Just-In-Time Compilation Using Modular Control-Flow Inegrity. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS),* (2014).

[35] NIU, B., AND TAN, G. Per-Input Control-Flow Integrity. In *Proceedings the ACM Conference on Computer and Communications Security (CCS),* (2015).

[36] PETSIOS, T., KEMERLIS, V. P., POLYCHRONAKIS, M., AND KEROMYTIS, A. D. DynaGuard: Armoring Canary-based Protections against Brute-force Attacks. In *Annual Computer Security Applications Conference (ACSAC),* (2015).

[37] PRAKASH, A., HU, X., AND YIN, H. Strict Protection for Virtual Function Calls in COTS C++ Binaries. In *Symposium on Network and Distributed System Security (NDSS),* (2015).

[38] PRAKASH, A., AND YIN, H. Defeating ROP Through Denial of Stack Pivot. In *Annual Computer Security Applications Conference (ACSAC),* (2015).

[39] RANDOMIZATION, P. T. A. S. L. `https://pax.grsecurity.net/docs/aslr.txt`.

[40] SCHUSTER, F., TENDYCK, T., LIEBCHEN, C., DAVI, L., SADEGHI, A.-R., AND HOLZ, T. Counterfeit Object-Oriented Programming. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P),* (2015).

[41] SHACHAM, H. The Geometry of Innocent Flesh on the Bone: Return-into-Libc without Function Calls (On the x86). In *ACM Conference on Computer and Communications Security (CCS),* (2007).

[42] TICE, C., ROEDER, T., COLLINGBOURNE, P., CHECKOWAY, S., ERLINGSSON, Ú., LOZANO, L., AND PIKE, G. Enforcing Forward-Edge Control-Flow Integrity in GCC and LLVM. In *Proceedings of the USENIX conference on Security (USENIX SEC),* (2014).

[43] VEEN, V. V. D., ANDRIESSE, D., GÖKTAS, E., GRAS, B., SAMBUC, L., SLOWINSKA, A., BOS, H., AND GIUFFRIDA, C. Practical Context-Sensiticve CFI. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS),* (2015).

[44] VEEN, V. V. D., GOKTAS, E., CONTAG, M., PAWLOWSKI, A., CHEN, X., RAWAT, S., BOS, H., HOLZ, T., ATHANASOPOULOS, E., AND GIUFFRIDA, C. A Tough call: Mitigating Advanced Code-Reuse Attacks At The Binary Level. In *Proceedings oaravindf the IEEE Symposium on Security and Privacy (S&P),* (2016).

[45] YUAN, P., ZENG, Q., AND DING, X. Hardware-Assisted Fine-Grained Code-Reuse Attack Detection. In *Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses (RAID),* (2015).

[46] ZHANG, C., CARR, S. A., LI, T., DING, Y., SONG, C., PAYER, M., AND SONG, D. VTrust: Regaining Trust on Virtual Calls. In *Symposium on Network and Distributed System Security (NDSS),* (2016).

[47] ZHANG, C., SONG, C., ZHIJIE, K. C., CHEN, Z., AND SONG, D. VTint: Protecting Virtual Function TablesÍntegrity. In *Proceedings of the Symposium on Network and Distributed System Security (NDSS),* (2015).

[48] ZHANG, C., WEI, T., CHEN, Z., DUAN, L., SZEKERES, L., MCCAMANT, S., SONG, D., AND ZOU, W. Practical Control Flow Integrity & Randomization for Binary Executables. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P),* (2013).

[49] ZHANG, M., AND SEKAR, R. Control Flow Integrity for COTS Binaries. In *Proceedings of the USENIX conference on Security (USENIX SEC),* (2013).

[50] ZHANG, M., AND SEKAR, R. Control Flow and Code Integrity for COTS binaries: An Effective Defense Against Real-ROP Attcks. In *Annual Computer Security Applications Conference (ACSAC),* (2015).