TYPESHIELD: Practical Defense Against Code Reuse Attacks using Binary Type Information

tba.

Abstract

We propose, TYPESHIELD, a binary runtime forward-edge and backward-edge protection tool which instruments program executables at load time. TYPESHIELD enforces a novel runtime control-flow integrity (CFI) policy based on function parameter type and count in order to overcome the limitations of available approaches and to efficiently verify dynamic object dispatches and function returns during runtime. To enhance practical applicability, TYPESHIELD can be automatically and easily used in conjunction with legacy applications or where source code is missing to harden binaries. We evaluated TYPESHIELD on highly relevant open source programs and the SPEC CPU2006 benchmark and were able to efficiently and with low performance overhead protect these applications from forward-edge and backward-edge corruptions. Finally, in a direct comparison with the state-of-the-art tools, TYPESHIELD achieves higher caller/-callee matching (*i.e.,* precision), while maintaining low runtime overhead and a calltarget set per callsite reduction gain of up to 41% compared to state-of-the-art tools.

1. Introduction

The C++ programming language offers object-oriented programming (OOP) concepts which are highly relevant during development of large, complex and efficient software systems, in particular, when runtime performance and reliability are primary objectives. A key OOP concept is polymorphism. This concept is based on C++ virtual functions. These functions enable late binding and allow programmers to overwrite a virtual function of the base-class with their own implementations. In order to implement virtual functions, the compiler needs to generate a virtual table metadata structures for all virtual functions and provide to each instance (object) of such a class a (virtual) pointer (its value is computed during runtime) to the aforementioned table. While this approach represents a main source of program indirection (*i.e.,* forward-edges), the basic compiler implementation provides unfortunately no security assurances (*i.e.,* Clang-CFI [7] virtual call protection is not used).

While the reasons for unwanted outcomes can be highly diverse, our work is primarily motivated by the presence of at least one exploitable memory corruption (*e.g.,* buffer overflow, etc.), which can enable the execution of sophisticated Code-Reuse Attacks (CRAs) such as the advanced COOP attack [30] and its extensions [4, 13, 22, 23]. A necessary ingredient for this class of attacks is the ability to corrupt a virtual object pointer in order to call gadgets by using a list of fake objects. To address such object dispatch corruptions and in general any type of indirect control flow transfer violation, Control-Flow Integrity (CFI) [8, 9] was originally developed to secure indirect control flow transfers by adding runtime checks before forward-edges and backward-edges. Unfortunately, COOP and its brethren bypass most deployed CFI-based enforcement policies, since these attacks do not exploit indirect backward-edges (*i.e.,* function returns), but rather exploit the forward indirect control flow transfers (*i.e.,* object dispatches) imprecision which cannot be statically precisely determined upfront as alias analysis is undecidable [28] in program binaries.

More promising *source code* tools such as: SafeDispatch [19], ShrinkWrap [18], VTI [12], and IFCC/VTV [32] rely on source-code availability which limits their applicability (*i.e.,* proprietary libraries). In contrast, *binary*-based tools typically protect only the forward-edge based on a CFI policy which assumes that a shadow stack [21] (bypassed in [6]) technique for protecting the backward edges is in place. Examples include binCFI [36, 37], vfGuard [27], vTint [35], VCI [15], Marx [26] and TypeArmor [34].

Unfortunately, VCI and Marx can be used to enforce only a CFI-based policy on the forward-edges based on a approximated class hierarchy (*i.e.,* no root class determined and the edges between the classes are not oriented) using several heuristics and assumptions. Further, these tools assume that a shadow stack protection policy is in place. TypeArmor enforces a forward-edge policy which does take into account only the number of parameter provided and consumed without imposing any constraint on their types. Thus, these forward-edge protection tools are too permissive and for this reason we seek in this paper for a more fined-grained protection technique which makes no assumptions on the presence of a shadow stack to protect the backward edges.

In this paper, we present TYPESHIELD, the first fine-grained CFI-complete (forward and backward edges) open source runtime binary-level protection tool. TYPESHIELD

backward-edgepolicy is based on the observation that backward-edges of a program can be precisely protected if there is a precise forward-edge mapping between caller and callees in first place determined. Due to the fact that TYPESHIELD significantly reduces the number valid forward-edges then previous work [34] we are able to build a precise backward-edge policy which represents an effective alternative along shadow stack based techniques. Thus there is no need to assume that other backward-edge protection mechanism (*i.e.,* shadow stack) is in place as most of the only forward-edge protection tools do. TYPESHIELD does not rely on RTTI data (*i.e.,* metadata emitted by the compiler, most of the time stripped) or particular compiler flags, and is applicable to industrial software. TYPESHIELD takes the binary of a program as input and it automatically instruments it in order to detect illegitimate indirect control flow transfers during runtime. In order to achieve this, TYPESHIELD analyzes x86-64 program binaries by carefully analyzing function parameter register wideness (parameter type) and the provided and consumed number of function parameters. Based on the used ABI, TYPESHIELD is consequently able to track up to 6 function arguments for the Itanium C++ ABI [2] x86-64 calling convention. The Itanium ABI caller callee calling convention essentially means that every called function will return at the next address located after the callsite which was used in first place to call this function. This means that there is a one to one mapping between each caller and callee contained in the program. However, we stress that the presented technique is usable for the ARM ABI [5] and Microsoft's C++ ABI [3] as well which are similar to the Itanium ABI w.r.t. the caller callee calling convention.

More precisely, the analysis performed by TYPESHIELD: (1) uses for each function parameter its register wideness (*i.e.,* ABI dependent) in order to map calltargets per callsites, (2) uses an address taken (AT) analysis for all calltargets, (3) compares individually parameters of callsites and calltargets in order to check if an indirect call transfer is acceptable or not, and (4) based on the provided forward-edge caller-callee mapping it builds a mapping back from each callee to the legitimate addresses located next to each caller, thereby providing a more strict callsite per calltarget compared to other state-of-the-art tools and a fine-grained shadow stack alternative for backward edges. TYPESHIELD uses automatically inferred parameter types which are used to build a more precise approximation of both the callee parameter types and callsite signatures.

TYPESHIELD's analysis is based on a use-def callees analysis to derive the function prototypes, and a liveness analysis at indirect callsites to approximate callsite signatures. This efficiently leads to a more precise control flow graph (CFG) of the binary program in question, which can be used also by other systems in order to gain a more precise CFG on which to enforce other types of CFI-related policies. These analysis results are used to determine a mapping between all callsites and legitimate calltarget sets. Further, this mapping is used in a backward analysis for determining the set of legitimate returns addresses for each function return determined by the each calltarget. Note that we consider each calltarget to be the start of function.

TYPESHIELD incorporates an improved forward-edge protection policy which is based on the insight that if the binary adheres to the standard calling convention (*i.e.,* Itanium ABI) for indirect calls, undefined arguments at the callsite are not used by any callee by design and that based on the passed function parameter types can be approximated by their corresponding register wideness. TYPESHIELD uses a forward-edge based propagation analysis used to determine a set of possible return addresses for calltargets (*i.e.,* function returns) which approximates the caller callee function calling convention in a fine-grained way. This policy is based on the observation that in case a fine-grained forward-edge policy can be between callers and callees determined than this mapping can be backwards reflected in order to construct a fine grained policy from callees to legitimate callers. Our backward-edge policy represents a fine-grained Safe Stack [21] (recently bypassed [6]) alternative. This attack shown that in general the protection offered by shadow stacks is questionable (at least four attack vectors) since it is relatively easy for a motivated attacker to disclose the shadow stack and bypass it.

We implemented TYPESHIELD using DynInst [11], which is a binary rewriting framework that allows program binary instrumentation during loading or runtime. We evaluated TYPESHIELD with several highly relevant open source programs and the SPEC CPU2006 benchmark and show that that our forward-edge policy is more precise than state-of-the-art and our backward-edge policy is a precise alternative to shadow stacks.

In summary, we make the following contributions:

- We designed a novel fine-grained CFI technique for protecting forward and backward edges against code reuse attacks.
- We implemented, TYPESHIELD, a prototype which enforces the aforementioned technique in stripped program binaries. TYPESHIELD can serve as platform for developing other binary based protection mechanisms.
- We conduct a thorough set of evaluative experiments in which we show that TYPESHIELD is more precise and effective than other state-f-the-art tools. Further, we show that our tool has a higher calltarget set reduction per callsite, thus further reducing the attack surface.

2. Overview an Threat Model

In §2.1 we present the main steps performed by TYPESHIELD in order to harden a program binary while in §2.2 we introduce the threat model used in this paper.

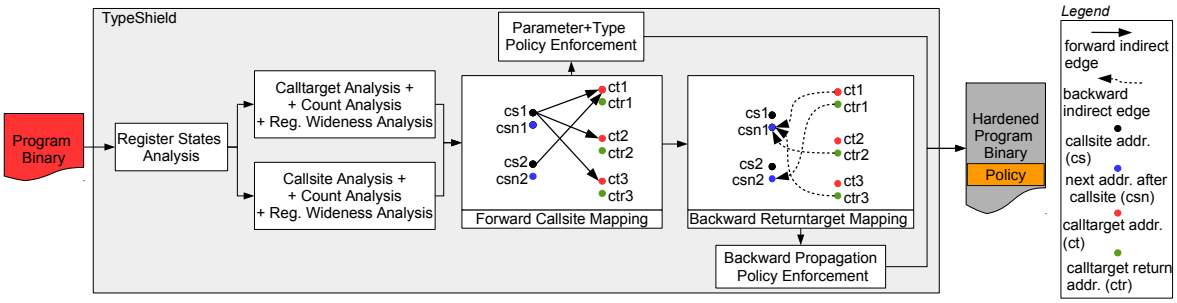

Figure 1: Overview of the main steps performed by TYPESHIELD when hardening a program binary.

2.1 Approach Overview

Figure 1 depicts the overview our approach. From right to left the program binary is analyzed (see left hand side in Figure 1) by TYPESHIELD and the calltargets and callsite analysis are performed for determining how many parameters are provided, how many are consumed and their register wideness. After this step labels are inserted at each previously identified callsite and at each calltarget. The enforced policy is schematically represented by the black highlighted dots (addresses) in Figure 1 which are allowed to call only legitimate red highlighted dots (addresses). Next for each function return address the address set determined by each address located after each legitimate (is allowed to call the function) callsite is collected. This information is obtained by using the previously determined callsite forward-edge mapping to derive a function return backward map containing function returns as key and return targets as values. In this way TYPESHIELD has for each function return site a set of legitimate addresses where the function return site is allowed to transfer the program control flow. Finally, range or compare checks are inserted before each function return site. This checks are used to check during runtime if the address where the function return wants to jump to is contained in the legitimate set for each particular return site. This is represented in Figure 1 by green highlighted dots (addresses) that are allowed to call only legitimate blue highlighted dots (addresses). Finally, the result is a hardened program binary (see right hand side in Figure 1).

2.2 Threat Model

We align our threat model with the same basic assumptions as described in [34]. More precisely, we assume a resourceful attacker that has read and write access to the data sections of the attacked program 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. Further, we assume that the attacker has the ability to execute a memory corruption to hijack the program control flow. Finally, the analyzed program binary is not hand-crafted and the compiler which was used to generate the binary adheres to one of the standard calling conventions mentioned in §1.

3. Design

In this section, we present in **??** we present our function parameter count based policy, in §3.1, we present the details of our type policy, and in §3.2 we introduce the definitions for our instructions analysis based on register states, in §3.3 we present the design of our calltarget analysis, while in §3.4 we depict the design of our callsite analysis[1]. Finally, in §3.6.1 we present our forward-edge policy instrumentation strategy, and in §3.5 we highlight our function backward-edgeanalysis and policy instrumentation strategy.

3.1 Parameter Register Wideness Based Policy

We use the register width of the function parameter in order to infer the type information. As previously mentioned, there are 4 types of reading and writing accesses. Therefore, our set of possible types for parameters is $\texttt{TYPE} = \{64, 32, 16, 8, 0\}$; where 0 models the absence of a parameter. Since Itanium C++ ABI specifies 6 registers (*i.e.,* `rdi`, `rsi`, `rdx`, `rcx`, `r8`, and `r9`) as parameter passing registers during function calls, we classify our callsites and calltargets into $\texttt{TYPE}^6$. Similar to our count policy, we allow overestimations of callsites and underestimations of calltargets, on the parameter types as well. Therefore, for a callsite cs it is possible to call a calltarget ct, only if for each parameter of ct the corresponding parameter of cs is not smaller w.r.t. the register width. This results in a finer-grained policy which is further restricting the possible set of calltargets for each callsite.

Further, we built a function parameter count-based policy similar to [34]. Calltargets are classified based on the number of parameters that these provide and callsites are classified by the number of parameters that these require. Further, we consider the generation of high precision measurements for such classification with binaries as the only source of information rather difficult. Therefore, over-estimations of parameter count for callsites and underestimations of the parameter count for calltargets is deemed acceptable. This classification is based on the general purpose registers that the

[1] Callsites detection in the binary is based on the capabilities of DynInst.

call convention of the current ABI—in this case the Itanium C++ ABI [2]—designates as parameter registers. Furthermore, we do not consider floating point registers or multi-integer registers for simplicity reasons. The *count* policy is based on allowing any callsite cs, which provides c_{cs} parameters, to call any calltarget 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 calltargets for the lower callsites, the restriction capability drops rather rapidly when reaching higher parameter counts, with callsites that use 6 or more parameters being able to call all possible calltargets. This is more precisely expressed as $\forall\ cs_1, cs_2;\ c_{cs_1} \leq c_{cs_2} \rightarrow \| \{ct \in \mathcal{F} \mid c_{ct} \leq c_{cs_1}\}\| \leq \| \{ct \in \mathcal{F} | c_{ct} \leq c_{cs_2}\}\|$.

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 callsites, which can access the same set of calltargets, a route which we will explore within this work.

3.2 Analysis of Register States

Our register state analysis is register state based, another alternative would be to do symbol-based data-flow analysis which we will leave as future work. In order for the reader to understand our analysis we will first give some definitions. The set `INSTR` describes all possible instructions that can occur within the executable section of a program binary. In our case, this is based on the x86-64 instruction set. An instruction $i \in$ `INSTR` can non-exclusively perform two kinds of operations on any number of existing registers. Note that there are registers that can directly access the higher 8-bit of the lower 16-bit. For our purpose, we register this access as a 16-bit access. (1) Read n-bit from the register with $n \in \{64, 32, 16, 8\}$, and (2) Write n-bit to the register with $n \in \{64, 32, 16, 8\}$.

Next, we describe the possible change within one register as $\delta \in \Delta$ with $\Delta = \{w64, w32, w16, w8, 0\} \times \{r64, r32, r16, r8, 0\}$. Note that 0 represents the absence of either a write or read access and $(0, 0)$ represents 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.,* $r64$ implies $r32$). Note that we exclude 0, as it means the absence of any access. Intanium C++ ABI specifies 16 general purpose integer registers. Therefore, we represent the change occurring at the processor level as $\delta_p \in \Delta^{16}$. In our analysis, we calculate this change for each instruction $i \in$ `INSTR` via the function $decode$: `INSTR` $\mapsto \Delta^{16}$.

3.3 Calltarget Analysis

Our calltarget analysis classifies calltargets according to the parameters they expect. Underestimations are allowed, however, overestimations are not permitted. For this purpose, we employ a customizable modified liveness analysis algorithm, which iterates over address-taken[2] functions with the goal of analyzing register state information in order to determine if these registers are used for arguments passing.

3.3.1 Liveness Analysis

A variable is alive before the execution of an instruction, if at least one of the originating paths performs a read access before any write access on that variable. If applied to a function, this calculates the variables that need to be alive at the beginning, as these are its parameters.

Algorithm 1: Basic block liveness analysis.

Input : The basic block to be analyzed - block : `INSTR`*
Output: The liveness state - $\mathcal{S}^{\mathcal{L}}$

```
1  Function analyze(block : INSTR*) : S^L is
2      state = Bl        ▷ Initialize the state with first block
3      foreach inst ∈ block do
4          state' = analyze_instr(inst)   ▷ Calc. changes
5          state = merge_h(state, state')      ▷ Merge changes
6      end
7      states = ∅                          ▷ Set of succ. states
8      blocks = successor(block)             ▷ Get succ. blocks
9      foreach block' ∈ blocks do
10         state' = analyze(block')        ▷ Analyze succ. block
11         states = states ∪ {state'}        ▷ Add succ. states
12     end
13     state' = merge_h (states)             ▷ Merge succ. states
14     return merge_v(state, state')       ▷ Merge to final state
15 end
```

Algorithm 1 is based on the liveness analysis algorithm presented in [20], which basically is a depth-first traversal of basic blocks. For customization, we rely on the implementation of several functions which we will present next. $\mathcal{S}^{\mathcal{L}}$ is the set of possible register states which depends on the specific implementations of the following operations.

- `merge_v` : $\mathcal{S}^{\mathcal{L}} \times \mathcal{S}^{\mathcal{L}} \mapsto \mathcal{S}^{\mathcal{L}}$, (merge vertically block states) describes how to merge the current state with the following state change.
- `merge_h` : $\mathcal{P}(\mathcal{S}^{\mathcal{L}}) \mapsto \mathcal{S}^{\mathcal{L}}$, (merge horizontally block states) describes how to merge a set of states resulting from several paths.
- `analyze_instr` : `INSTR` $\mapsto \mathcal{S}^{\mathcal{L}}$, (analyze instruction) calculates the state change that occurs due to the given instruction.
- `succ` : `INSTR`* $\mapsto \mathcal{P}$(`INSTR`*), (successor of a basic block) calculates the successors of the given block.

In our specific case, the function `analyze_instr` needs to also to handle non-jump and non-fall-through successors, as these are not handled by DynInst. Essentially, there are three relevant cases. First, if the current instruction is an indirect call or a direct call and the analysis algorithm is set not to follow calls, then our analysis will return a state where

[2] A program function is defined to have its address taken if there is at least one binary instruction which loads the function entry point into memory. Note that by definition, indirect calls can only target AT functions.

all registers are considered to be written before read. Second, if the current instruction is a direct call and the analysis algorithm is set not to follow calls, then we start an analysis of the target function an return its result. 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. Finally, in any other case, we simply return the decoded state. This leaves us with the two undefined merge functions and the undefined liveness state $\mathcal{S}^{\mathcal{L}}$.

3.3.2 Required Parameter Wideness

For our type policy, we need a finer representation of the state of one register as follows. (1) W represents write before read access, (2) $r8, r16, r32, r64$ represents read before write access with 8-, 16-, 32-, 64-bit width, and (3) C represents the absence of access. This gives us the following $S^{\mathcal{L}} = \{C, r8, r16, r32, r64, W\}$ register state which translates to the register super state $\mathcal{S}^{\mathcal{L}} = (S^{\mathcal{L}})^{16}$. As there could be 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. To permit this, we allow our merge operations to also return the value RW, which represents the existence of both read and write access and then can use W with the functionality of an end marker. Therefore, our vertical merge operator conceptually intersects all read accesses along a path until the first write occurs $merge\_v^i$. In any other case, it behaves like the previously mentioned vertical merge function. Our horizontal merge $merge\_h$ function is a pairwise combination of the given set of states, which are then combined with an union-like operator with W preceding WR and WR preceding R and R preceding C. Unless one side is W, read accesses are combined in such a way that always the higher one is selected.

3.3.3 Required Parameter Count

For our count policy, we need a coarse representation of the state of one register, thus we use the following representation. (1) W represents write before read access, (2) R represents read before write access, and (3) C represents the absence of access. Further, this gives us the $S^{\mathcal{L}} = \{C, R, W\}$ as register state, which translates to the register super state $\mathcal{S}^{\mathcal{L}} = (S^{\mathcal{L}})^{16}$. We implement `merge_v` in such a way that a state within a superstate is only updated if the corresponding register was not accessed, as represented by C. Our reasoning is that the first access is the relevant one in order to determine read before write. Our horizontal *merge(*$merge\_h$*)* function is a simple pairwise combination of the given set of states, which are then combined with an union like operator with W preceding R and R preceding C. The index of highest parameter register based on the used call convention that has the state R considered to be the number of parameters a function at least requires to be prepared by a callsite.

3.3.4 Void/Non-Void Calltarget

In order to determine if a calltarget is a void or non-void return function TYPESHIELD traverses backwards the basic blocks from the return instruction of the function an looks for the `RAX` register. In case there is a write operation on the `RAX` register then TYPESHIELD infers that the function return is non-void and thus provides a pointer value back.

3.4 Callsite Analysis

Our callsite analysis classifies callsites according to the parameters they provide. Overestimations are allowed, however, underestimations are not permitted. For this purpose we employ a customizable modified reaching definition algorithm, which we will show first.

3.4.1 Reaching Definitions

An assignment to a variable is a reaching definition after the execution of a set of instruction if that variable still exists in at least one possible execution path. If applied to a callsite, this calculates the values that are provided by this callsite to the function it then invokes.

Algorithm 2: Basic block reaching definition analysis.

Input : The basic block to be analyzed - *block* : `INSTR`*
Output: The reaching definition state - $\mathcal{S}^{\mathcal{R}}$

Function *analyze(block : INSTR*) :* $\mathcal{S}^{\mathcal{R}}$ **is**
  $state$ = BI ▷ Initialize the state with first block
  foreach *inst* ∈ *reversed(block)* **do**
    $state' = analyze\_instr(inst)$ ▷ Calculate changes
    $state = merge\_v(state, state')$ ▷ Merge changes
  end
  $states = \emptyset$ ▷ Set of predecessor states
  $blocks = pred(block)$ ▷ Get predecessors blocks
  foreach *block' ∈ blocks* **do**
    $state' = analyze(block')$ ▷ Analyze pred. block
    $states = states \cup \{state'\}$ ▷ Add pred. states
  end
  $state' = merge\_h(states)$ ▷ Merge predecessors states
  return $merge\_v(state, state')$ ▷ Merge to final state
end

Algorithm 2 is based on the reaching definition analysis presented in [20], which basically is a reverse depth-first traversal of basic blocks of a program. For customization, we rely on the implementation of several functions. $\mathcal{S}^{\mathcal{R}}$ is the set of possible register states which depends on the specific reaching definition implementation of the following operations.

- `merge_v` : $\mathcal{S}^{\mathcal{R}} \times \mathcal{S}^{\mathcal{R}} \mapsto \mathcal{S}^{\mathcal{R}}$, (merge vertically block states) describes how to merge the current state with the following state change.
- `merge_h` : $\mathcal{P}(\mathcal{S}^{\mathcal{R}}) \mapsto \mathcal{S}^{\mathcal{R}}$, (merge horizontally block states) describes how to merge a set of states resulting from several paths.
- `analyze_instr` : `INSTR` $\mapsto \mathcal{S}^{\mathcal{R}}$, (analyze instruction) calculates the state change that occurs due to the given instruction.

• `pred` : $\texttt{INSTR}^* \mapsto \mathcal{P}(\texttt{INSTR}^*)$, (predecessor of a basic block) calculates the predecessors of the given block.

In our specific case, the function `analyze_instr` does not need to handle normal predecessors, as DynInst will resolve those for us. However there are several instructions that have to be handled as depicted in the following situations. (1) If the current instruction is an indirect call or a direct call and the analysis algorithm is set not to follow calls, then return a state where all registers are considered trashed. (2) If the instruction is a direct call and the analysis algorithm is set to follow calls, then we start an analysis of the target function. (3) In all other cases we simply return the decoded state. This leaves us with the two merge functions and the undefined reaching definitions state $\mathcal{S}^{\mathcal{R}}$.

Previous work [20] provides a reaching definition analysis on blocks, which we use to arrive at the algorithm depicted in Algorithm 2 to compute the liveness state at the start of a basic block. We apply the reaching analysis at each indirect callsite directly before each call instruction.

This algorithm relies on various functions that can be used to configure its behavior. We 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 $analyze\_instr$. Note, that the function $pred$, which retrieves all possible predecessors of a block is provided by the DynInst instrumentation framework.

The $analyze\_instr$ function calculates the effect of an instruction and is the core of the analyze function (see Algorithm 2). 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: (1) If the instruction is an indirect call or a direct call but we chose not to follow calls, then return a state where all trashed 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, and (3) 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 $\mathcal{S}^{\mathcal{L}}$ into parameters up to the following subsections.

3.4.2 Provided Parameter Width

In order to implement our type policy, we use a finer representation of the states of one register, thus we consider: (1) T represents a trashed register, (2) $s8, s16, s32, s64S$ represents a set register with 8-, 16-, 32-, 64-bit width, and (3) U represents an untouched register. This gives us the following $S^{\mathcal{L}} = \{T, s64, s32, s16, s8, U\}$ register state which translates to the register super state $\mathcal{S}^{\mathcal{R}} = (S^{\mathcal{R}})^{16}$.

However, 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 $merge\_h$ function is a simple pairwise combination of the given set of states, which are then combined with an union like operator with T preceding S and S preceding U. Note, that when both states are set, we pick the higher one.

3.4.3 Provided Parameter Count

For implementing our count policy, we use a coarse representation of the state of one register, thus we use the following representation. (1) T represents a trashed register, (2) S represents a set register (written to), and (3) U represents an untouched register. This gives us the following $S^{\mathcal{L}} = \{T, S, U\}$ register state which translates to the register super state $\mathcal{S}^{\mathcal{R}} = (S^{\mathcal{R}})^{16}$.

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, our vertical merge function $merge\_v$ behaves as follows. In case the first given state is U, than the return value is the second state and in all other cases it will return the first state.

Our horizontal merge $merge\_h$ function is a pairwise combination of the given set of states, which are then combined with an union like operator with T preceding S and S preceding U.

The index of the highest parameter register based on the used call convention that has the state S is considered to be the number of parameters a callsite prepares at most.

3.4.4 Void/Non-Void Callsite

In order to determine if a callsite is a void or non-void return function TYPESHIELD looks at the callsite if there is an read before write on the `RAX` register. In case there is a read before write operation on the `RAX` register then TYPESHIELD infers that the callsite is non-void and thus expects a pointer to be provided when the called function returns.

3.5 Backward-Edge Analysis

In order to protect the backward edges of our previously determined calltargets for each callsite we designed an analysis which can determine possible legitimate return target addresses.

Algorithm 3 depicts how the forward mapping between callsites and calltargets is used to determine the backward address set for each return address contained in each address taken function. The $fMap$ is obtained after running the callsite and calltarget analysis (see §3.3 and §3.4). These mapping contains for each callsite the legal calltargets where the forward-edge indirect control flow transfer is allowed to jump to. This mapping is reflected back by construction a second mapping between the return address of each function for which we have the start address and a return target address set.

Algorithm 3: Calltarget return set analysis.

```
Input  : Forward edge callsite to calltargets map - fMap
Output: Backward edge to return addresses map - rMap
1  Function backwardAddressMapping(fMap) : rMap is
       ▷ visit all detected callsites in the binary
2      foreach callsite ∈ fMap do
           ▷ get calltargets for callsite address key
3          calltargetSet = getCalltargetSet(callsite, fMap)
           ▷ calltarget is the function start address
           ▷ visit all calltargets of a callsite
4          foreach calltarget ∈ calltargetSet do
               ▷ get the next address after the callsite
5              rTarget = getNextAddress(callsiteKey)
               ▷ find the address of function return
6              rAddress =
               getReturnOfCalltarget(calltarget)
               ▷ rAddress is map key; rTarget is value
7              rMap = rMap ∪
               rMap add (rAddress, rTarget)
8          end
9      end
       ▷ return the backward-edgeaddresses mappings
10     return rMap
11 end
```

The return target address set for a function return is determined by getting the next address after each callsite address which is allowed to make the forward-edge control flow transfer (*i.e.,* recall the caller callee calling convention). The $rMap$ is obtained by visiting each function return address and assigning to it the address next to the callsite which was used in order to transfer the control flow to the function in first place. At the end of the analysis all callsites and all function returns have been visited and a set for each function return address of backward-edgeaddresses will be obtained. Note that the function boundary address (*i.e.,* retn) was detected by a linear basic block search from the beginning of the function (calltarget) until the first return instruction was encountered. We are aware that other promising approaches for recuperating function boundaries (*e.g.,* [10]) exist, and plan to experiment with them in future work.

3.6 Binary Instrumentation

3.6.1 Forward-Edge Policy Enforcement

The result of the forward callsite and calltarget analysis is a mapping between the allowed calltargets for each callsite. In order to enforce this mapping during runtime each callsite and calltarget contained in the previous mapping are instrumented inside the binary program with two labels and a callsite located CFI-based checking mechanism. At each callsite the number of provided parameters are encoded as a series of six bits. At the calltarget the label contains six bits denoting how many parameters the calltarget expects. Additionally, at the callsite six bits encode which register wideness types each of the provided parameters have while at the calltarget another six bits are used to encode the types of the parameters expected. Further, at the callsite another bit is used to define if the function is expecting a `void` return type or not. All this information are written in labels before each callsite and calltarget. During runtime before each callsite these labels are compared by performing a xor operation between the bits contained in the previously mentioned labels. In case the xor operation returns false than the transfer is allowed else the program execution is terminated.

3.6.2 Backward-Edge Policy Enforcement

The previously determined $rMap$ in Algorithm 3 will be used to insert a check before each function (calltarget) return present in the $rMap$. We propose three modes of operation based on three types of checks which can be inserted before each function return instruction. Depending on the specific needs one of the following modes of operation can be selected. Depending on the runtime time needed to perform each backward-edge indirect transfer we define the following three modes of operation.

Super fast mode. Based on the $rMap$, for each AT function return the minimum and the maximum address out of the return set for a particular $rAdress$ return address will be determined. Next, these two values will be used to insert a range check having as left and right boundaries these two values. Before the return instruction of the function is executed the value of the function return is compared against these two values previously mentioned. In case the check fails than the program will be terminated else the indirect control flow transfer will be allowed. Note that this check has insignificant runtime overhead but on the other side it could contain not legitimate return addresses depending on the entropy of the $rAddresse$s. In short, this means that as far as the min and max addresses are from each other the more leeway the attacker will have.

Fast mode. Based on the $rMap$, before each AT function return a randomly generated label (*i.e.,* the value 7232943 loaded trough one level of indirection) value will be inserted. The same label will be inserted before each legitimate (*i.e.,* based on the forward-edge policy) target address (next address after a legitimate callsite) of the function return. In this way a function return will be allowed to jump to only the instruction which follows next to the address of the callsites which are allowed to call the calltarget which contains this particular function return. For callsites which are allowed to call the the calltarget mentioned and another calltarget than in this cases TypeShield will perform a search in order to detect if the callsite has already a label attached to the next address after the callsite. In this case the label will be reused. In this situation two callsites share their labels. The solution to this is to use single labels for each function return address. In this case multiple labels have to be stored for each address following a legitimate callsite. Further, addresses located after a callsite that are not allowed to call a particular calltarget will get another randomly generated label. In this way calltarget return labels are grouped together based on the $rMap$. This mode allows at least (additionally

the callsites which are allowed to call more than one call-target are added) the same number of function return sites as the forward-edge policy enforces for each callsite and it is runtime efficient since label checking is based on a single compare check.

Slow mode. Based on the $rMap$, before each AT function return a series of comparison checks are inserted in the binary. Before the return instruction of the function a series of comparison checks between the appropriate addresses stored in $rMap$ and the address where the function wants to return are performed. In case one of the check fails than the program will be terminated. The total number of comparison checks added is equal to the size of return address set which contains $rTarget$ values. Note that these types of checks are precise since only legitimate addresses are allowed but on the other side the runtime overhead is higher than in the case of the fast path because the number of checks is in general higher.

4. Implementation

We implemented TYPESHIELD using the DynInst [11] (v.9.2.0) instrumentation framework. In total, we implemented TYPESHIELD in 5501 lines of code (LOC) of C++ code. We currently restricted our analysis and instrumentation to x86-64 bit elf binaries using the Itanium C++ ABI call convention. We focused on the Itanium C++ ABI 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 support other ABIs as well. We developed the main part of our binary analysis 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, which can be customized with relative ease, as we allow for arbitrary path merging functions. Next, we implemented a Clang/LLVM (v.4.0.0, trunk 283889) back-end pass (416 LOC) 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 contained in 3239 LOC of Python code.

5. Evaluation

We evaluated TYPESHIELD by instrumenting various open source applications and conducting a thorough analysis. Our test sample includes the two FTP servers *Vsftpd* (v.1.1.0, C code), Pure-ftpd (v.1.0.36, C code) and *Proftpd* (v.1.3.3, C code), web server Lighttpd (v.1.4.28, C code); the two database server applications *Postgresql* (v.9.0.10, C code) and *Mysql* (v.5.1.65, C++ code), the memory cache application *Memcached* (v.1.4.20, C code), the *Node.js* application server (v.0.12.5, C++ code). We selected these applications in order to allow for comparison with [34]. In our evaluation we addressed the following research questions (RQs).

RQ1: How **precise** is TYPESHIELD? (§5.1)
RQ2: How **effective** is TYPESHIELD? (§5.2)
RQ3: What **overhead** imposes TYPESHIELD? (§5.3)
RQ4: What **binary blow-up** has TYPESHIELD? (§5.4)
RQ5: What **security level** offers TYPESHIELD? (§5.5)
RQ6: Which **attacks** mitigates TYPESHIELD? (§5.6)
RQ7: Is TYPESHIELD **effective** against COOP? (§5.7)
RQ8: Are other tools **better** than TYPESHIELD? (§5.8)
RQ9: Is TYPESHIELD **better** than ShadowStack? (§5.9)

Comparison Method. We used TYPESHIELD to analyze each program binary individually. Next TYPESHIELD was used to harden each binary with forward and backward checks. The data generated during analysis and binary hardening was written into external files for later processing. Finally, the previous obtained data was extracted with our Python based framework and inserted into spreedsheet files in order to be able to better compare the obtained results with other existing tools.

Experimental Setup. Our used setup consisted in a VirtualBox (version 5.0.26r) instance, in which we ran a Kubuntu 16.04 LTS (Linux Kernel version 4.4.0). We had access to 3GB of RAM and 4 out of 8 provided hardware threads (Intel i7-4170HQ @ 2.50 GHz).

5.1 Precision (RQ1)

In order to measure the precision of TYPESHIELD, we need to compare the classification of callsites and calltargets as provided by our tool with some ground truth data. We generated the ground truth data by compiling our test targets using a custom back-end Clang/LLVM compiler (v.4.0.0 trunk 283889) MachineFunction pass inside the x86-64-Bit code generation implementation of LLVM. During compilation, we essentially collect three data points for each callsite and calltarget as follows. (1) the point of origination, which is either the name of the calltarget or the name of the function the callsite resides in, (2) the return type that is either expected by the callsite or provided by the calltarget, and (3) the parameter list that is provided by the callsite or expected by the calltarget, which discards the variadic argument list.

5.1.1 Quality and Applicability of Ground Truth

Table 1 depicts the results obtained w.r.t. the investigation of callrgets comparability and the callsites compatibility. We assessed the applicability of our collected ground truth, by assessing the structural compatibility of our two data sets. Table 1 shows three data points w.r.t. calltargets for the optimization level `-O2`: (1) Number of comparable calltargets that are found in both datasets, (2) Clang miss: Number of calltargets that are found by TYPESHIELD, but not by our Clang/LLVM pass, and (3) TypeShield miss: Number of calltargets that are found by our Clang/LLVM pass, but not by TYPESHIELD. Both columns (Clang miss and Type-

-O2	Calltargets			Callsites		
Target	match	Clang miss	TypeShield miss	match	Clang miss	TypeShield miss
Proftpd	1202	0 (0%)	1 (0.08%)	157	0 (0)	0 (0.08)
Pure-ftpd	276	1 (0.36%)	0 (0%)	8	2 (20)	5 (0)
Vsftpd	419	0 (0%)	0 (0%)	14	0 (0)	0 (0)
Lighttpd	420	0 (0%)	0 (0%)	66	0 (0)	0 (0)
MySQL	9952	9 (0.09%)	7 (0.07%)	8002	477 (5.62)	52 (0.07)
Postgressql	7079	9 (0.12%)	0 (0%)	635	80 (11.18)	40 (0)
Memcached	248	0 (0%)	0 (0%)	48	0 (0)	0 (0)
Node.js	20337	926 (4.35%)	23 (0.11%)	10502	584 (5.26)	261 (0.11)
geomean	1460.87	4.07 (0.60%)	1.89 (0.40%)	203.77	9.04 (3.00)	6.37 (0.40)

Table 1: The quality of structural matching provided by our automated verify and test environment, regarding callsites and calltargets 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 TypeShield denotes elements not found in the data-set of TYPESHIELD.

Shield miss) show a relatively low number of encountered misses. Therefore, we can state that our structural matching between ground truth and TYPESHIELDs callsites is almost perfect.

Calltargets. The obvious choice for structural comparison regarding calltargets is their name, as these are functions. First, we have to remove internal functions from our datasets like the `_init` or `_fini` functions, which are of no relevance for this investigation. 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 defining 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 return type, its name-space and the types of its parameter list. Therefore, we need to reverse this process and then compare the fully typed names.

The problematic column is the Clang miss column, as these values might indicate problems with TYPESHIELD. These numbers are relatively low (below 1%) with only Node.js shows a noticeable higher value than the rest. 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 state that our structural matching between ground truth and TYPESHIELDs calltargets is very good.

Callsites. While our structural matching of calltargets is rather simple, matching callsites is more complex. Our tool can provide accurate addressing of callsites within the binary. However, Clang/LLVM does not have such capabilities in its intermediate representation (IR). Furthermore, the IR is not the final representation within the compiler, as the IR is transformed into a machine-based representation (MR), which is again optimized. Although, we can read information regarding parameters from the IR, it is not possible with the MR. Therefore, we extract that data directly after the conversion from IR to MR and read the data at the end of the compilation. To not unnecessarily pollute our dataset, we only considered calltargets, which have been found in both datasets.

5.1.2 Type Based Classification Precision

02	Calltargets			Callsites		
Target	#cs gt	perfect args	perfect return	#ct gt	perfect args	perfect return
Proftpd	1009	835 (82.75%)	861 (85.33%)	157	125 (79.61%)	113 (71.97%)
Pure-Ftpd	128	101 (78.9%)	54 (42.18%)	8	4 (50%)	8 (100%)
Vsftpd	315	256 (81.26%)	179 (56.82%)	14	14 (100%)	14 (100%)
Lighttpd	289	253 (87.54%)	244 (84.42%)	66	48 (72.72%)	57 (86.36%)
MySqld	9728	6141 (63.12%)	7684 (78.98%)	8002	4477 (55.94%)	6449 (80.59%)
Postgressql	6873	5730 (83.36%)	4952 (72.05%)	635	455 (71.65%)	573 (90.23%)
Memcached	133	110 (82.7%)	70 (52.63%)	48	43 (89.58%)	48 (100%)
Node.js	20069	15161 (75.54%)	13911 (69.31%)	10502	4757 (45.29%)	8841 (84.18%)
geomean	1097.06	867.43 (79.06%)	723.70 (65.96%)	203.77	139.08 (68.25%)	180.59 (88.62%)

Table 2: *type* based policy classification of callsites.

Table 2 depicts the number and ratio of perfect classifications and the number and ratio of problematic classifications, which in the case of calltargets refers to overestimations and in case of callsites refers to underestimations. We used the `-O2` optimization level, when comparing to the ground truth obtained by our Clang/LLVM pass. The #cs gt and #ct gt labels mean total number of callsites and calltarget based on the ground truth, respectively. The label perfect args denotes all occurrences when our result and the ground truth perfectly match regarding the required/provided arguments. The label perfect return denotes this for return values.

Calltargets. For the first experiment we used the union combination operator with an $analyze$ function that follow into occurring direct calls and a vertical merge and that intersects all reads until the first write. The results indicate a rate of perfect calltargets classification is over 79% while for the returns it is over 65%.

Callsites. For the second experiment we used the union combination operator with an $analyze$ function that does not follow into occurring direct calls while relying on a backward inter-procedural analysis. The results indicate a rate of perfect classification of over 68% while for the returns it is over 88%.

5.2 Effectiveness (RQ2)

Table 3 depicts the the average number of calltargets per callsite, the standard deviation σ and the median. We evaluated the effectiveness of TYPESHIELD by leveraging the results of several experiment runs. First, we established a baseline using the data collected from our Clang/LLVM pass. These are the theoretical limits of our implementation which can be reached for both the count and the type schema. Second, we evaluated the effectiveness of our count policy. Third, we evaluated the effectiveness of our type policy.

-O2	AT	*count**			*count*			*type**			*type*		
Target		limit (mean ± σ)		median	limit (mean ± σ)		median	limit (mean ± σ)		median	limit (mean ± σ)		median
ProFTPD	396	330.31 ± 48.07		343.0	334.5 ± 51.26		311.0	310.58 ± 60.33		323.0	337.41 ± 54.09		336.0
Pure-FTPD	13	5.5 ± 4.82		6.5	9.87 ± 4.32		13.0	4.37 ± 4.92		2.0	8.12 ± 4.11		7.0
Vsftpd	10	7.14 ± 1.81		6.0	7.85 ± 1.39		7.0	5.42 ± 0.95		6.0	6.42 ± 0.96		7.0
Lighttpd	63	27.75 ± 10.73		24.0	41.19 ± 13.22		41.0	25.1 ± 8.98		24.0	41.42 ± 14.29		38.0
MySQL	5896	2804.69 ± 1064.83		2725.0	4281.71 ± 1267.78		4403.0	2043.58 ± 1091.05		1564.0	3617.51 ± 1390.09		3792.0
Postgressql	2504	1964.83 ± 618.28		2124.0	1990.59 ± 574.53		2122.0	1747.22 ± 727.08		2004.0	1624.07 ± 707.58		1786.0
Memcached	14	11.91 ± 2.84		14.0	12.0 ± 1.38		13.0	9.97 ± 1.45		11.0	10.25 ± 0.77		10.0
Node.js	7230	3406.07 ± 1666.9		2705.0	5306.05 ± 1694.73		5429.0	2270.28 ± 1720.32		1707.0	4229.22 ± 2038.64		3864.0
geomean	216.61	129.77 ± 43.99		127.62	166.09 ± 40.28		171.97	105.13 ± 38.68		92.74	144.06 ± 38.38		141.82

Table 3: Results for allowed callsites per calltarget for several programs compiled with Clang using optimization level `-O2`. Note that the basic restriction to address taken only calltargets (see column AT) is present for each other series. The label $count^*$ denotes the best possible reduction using our *count* policy based on the ground truth collected by our Clang/LLVM pass, while $count$ denotes the results of our implementation of the *count* policy derived from the binaries. The same applies to $type*$ and $type$ regarding the *type* policy. A lower number of calltargets per callsite indicates better results. Note that our *type* policy is superior to the *count* policy, as it allows for a stronger reduction of allowed calltargets. We consider this a good result which further improves the state-of-the-art. Finally, we provide the median and the pair of mean and standard deviation to allow for a better comparison with other state-of-the-art tools.

5.2.1 Theoretical Limits

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

Experiment Setup. Based on the type information collected by our Clang/LLVM pass, we conducted two experiment series. We derived the available number of calltargets for each callsite based on the collected ground truth applying the count and type schemes.

Results. (1) The theoretical limit of the *count** schema has a geometric mean of 129 possible calltargets, which is around 11% of the geometric mean of the total available calltargets (1097, see Table 2), and (2) The theoretical limit of the *type** schema has a geometric mean of 105 possible calltargets, which is 9.5% of the geometric mean of the total available calltargets (1097, see Table 2). When compared, the theoretical limit of the *type** policy allows about 19% less available calltargets in the geomean with Clang `-O2` than the limit of the *count** policy (*i.e.,* 105 vs. 129).

5.2.2 Calltarget Reduction per Callsite

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

Results. (1) The *count* schema has a geometric mean of 166 possible calltargets, which is around 15% of the geometric mean of total available calltargets (1097, see Table 2). This is around 28% more than the theoretical limit of available calltargets per callsite, see *count**, and (2) The *type* schema has a geometric mean of 144 possible calltargets, which is around 13% of the geometric mean of total available calltargets (1097, see Table 2). This is around 37% more than the theoretical limit of available calltargets per callsite, see *type**. Our implementation of the *type* policy allows around 21% less available calltargets in the geomean with Clang `-O2` than our implementation of the *count* policy and further a total reduction of more than 87% (141 vs. 1097) w.r.t. to total number of AT calltargets available after our *count* and *type* policies were applied.

5.3 Runtime Overhead (RQ3)

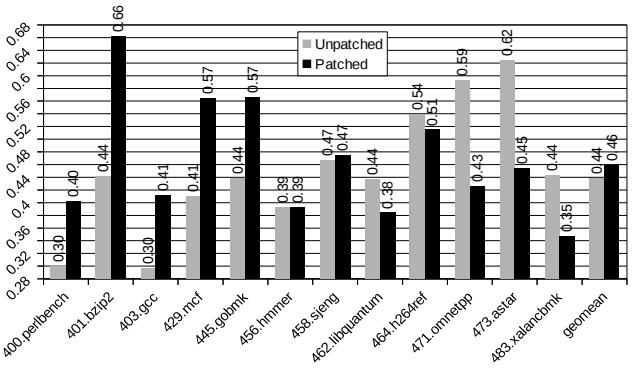

Figure 2: SPEC CPU2006 Benchmark Results.

Figure 2 depicts the runtime overhead obtained by applying TYPESHIELD (forward-edge policy (parameter count and type mode) and backward-edge policy (fast mode)) to several SPEC CPU2006 benchmarks. Out of the used programs: xalanckbmk, astar, and omnetpp are C++ program while the rest are pure C programs. Unpatched means the original vanilla programs while patched means the programs with the CFI checks inserted. After the programs were instrumented we measured the runtime overhead. It is important to notice that none of the instrumented programs crashed and all executed as intended by the benchmark. The obtained runtime overhead is around 3% (geomean) when instrumenting the binary with DynInst. One reason for the performance

drop includes cache misses introduced by jumping between the old and the new executable section of the binary generated by duplicating and patching. This is necessary, because when outside of the compiler, it is nearly impossible to relocate indirect control flow. Therefore, every time an indirect control flow occurs, one jumps into the old executable section and from there back to the new executable section. Moreover, this is also dependent on the actual structure of the target, as it depends on the number of indirect control flow operations per time unit. Another reason for the slightly higher (yer acceptable) performance overhead is due to our runtime policy which is more complex than that of other state-of-the-art tools. However, the runtime overhead of TYPESHIELD (3%) is comparable with other forward-edge protection tools such as: TypeArmor (3%), VCI [15] (7.79% overall and 10.49% on only the SPEC CPU2006 programs), vfGuard [27] (10% - 18.7%), T-VIP [17] (0.6% - 103%), SafeDispatch [19] (2% - 30%), and VTV/IFCC [32] (8% - 19.2%). Finally, this results qualify TYPESHIELD as a highly practical tool.

5.4 Instrumentation Overhead (RQ4)

The instrumentation overhead (*i.e.,* binary blow-up) or the change in size due to patching is mostly due to the method DynInst uses to patch binaries. Essentially, the executable part of the binary is duplicated and extended with the check we insert. The usual ratio we encountered in our experiments is around 40% to 60% with Postgres having an increase of 150% in binary size. One cannot reduce that value significantly, because of the nature of code relocation after loosing the information which a compiler has. Especially indirect control flow changes are very hard to relocate. Therefore, instead each important basic block in the old code contains a jump instruction to the new position of the basic block. Finally, this results should not represent an issue for memory resourceful systems on which these applications typically run.

5.5 Security Analysis (RQ5)

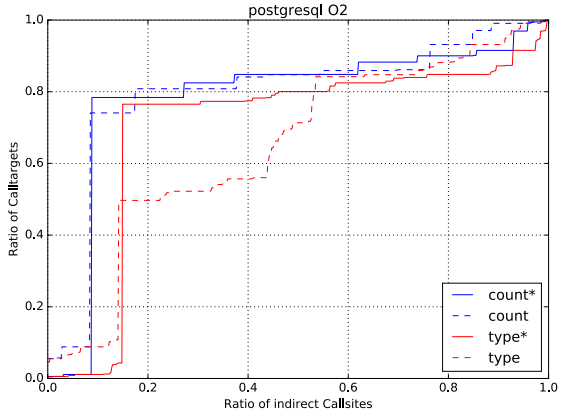

Figure 3: CDF for PostgreSQL compiled with Clang `-O2`.

Figure 3 depicts the cumulative distribution function (CDF) of the PostgreSQL program which was compiled with the Clang `-O2` flag. We selected this program randomly from our sample programs. The CDFs depict the number of legal callsite targets and the difference between the type and the count policies. While the count policies have only a few changes, the number of changes that can be seen within the type policies are vastly higher. The reason for this is fairly straightforward: the number of buckets (*i.e.,* this is the number of equivalence classes) that are used to classify the callsites and calltargets is simply higher. While type policies mostly perform better than the count policies, there are still parts within the type plot that are above the count plot, the reason for that is also relatively simple: the maximum number of calltargets a callsite can access has been reduced. Therefore, a lower number of calltargets is a higher percentage than before. However, Figure 3 depicts clearly that the *count** and *type** have higher values as *count* and *type*, respectively. This further, confirms our assumptions w.r.t. these used metrics. Finally, note that the results dependent on the particular internal structure of the hardened programs.

5.6 Mitigation of Advanced CRAs (RQ6)

Exploit	Stopped	Remark
COOP ML-G [30]		
IE 32 bit	×	Out of scope
IE 1 64-bit	✓(FP)	Argcount mismatch
IE 2 64-bit	✓(FP)	Argcount mismatch
Firefox	✓(FP)	Argcount mismatch
COOP ML-REC [13]		
Chrome	✓(FP)	Void target where non-void was expected
Control Jujutsu [16]		
Apache	✓(FP)	Target function not AT
Nginx	✓(FP)	Void target where non-void was expected
All Backward edge violating attacks	✓(BP)	(1)[a] or (2)[b] or (3)[c]

[a] Jump to an address $\notin$ in the $max - min$ address range.
[b] Jump to an address $\neq$ to one of the legitimate addresses.
[c] Jump to an address label $\neq$ with the calltarget return label.

Table 4: Stopped CRAs, forward-edge policy (FP) and backward-edge policy (BP).

Table 4 depicts several attacks that can be successfully stopped by TYPESHIELD by deploying only the forward-edge or the backward-edge policy. For testing if the COOP attack can be prevented we instrumented the Firefox library (libxul.so) which was used to perform the original COOP attack as presented in the original paper. We observed that due to the forward-edge policy this attack was no longer possible. For testing if backward-edge attacks are possible after applying TYPESHIELD we used several open source ROP attacks which are explicitly violating the control flow of the program through backward-edge violation which are

based on C++ programs. Next we instrumented the binaries of these programs which were used to violate the return edge in order to call different gadgets. Each attack which was using one of the protected function returns was successfully stopped.

In summary, all forward-edge and backward-edge attacks can be successfully mitigated by TYPESHIELD as long these attacks are not aware of the policy in place and thus can not selectively use gadgets which have their start address in the allowed set for the legitimate forward-edge and backward-edge transfers, respectively.

5.7 Effectiveness Against COOP (RQ7)

We investigated the effectiveness of TYPESHIELD against the COOP attack by looking at the number of register arguments which can be used to enable data-flow between gadgets. In order to determine how many arguments remain unprotected after we apply the forward-edge policy of TYPESHIELD we compared the number of parameter overestimation and compare it with the ground truth obtained with the help of an LLVM compiler pass. Next we used some heuristics to determine how many ML-G and REC-G callsites exist for each of the C++ only server applications. Finally, we compared these results with the one obtained by TypeArmor.

		Overestimation					
Program	#cs	0	+1	+2	+3	+4	+5
MySQL (ML-G)	192	184	3	1	0	1	3
Node.js (ML-G)	134	131	1	0	1	0	1
geomean	160	155	1	1	1	1	1
MySql (REC-G)	289	273	10	2	3	0	1
Node.js (REC-G)	72	69	2	0	0	0	1
geomean	144	137	4	1	1	1	1

Table 5: Parameter overestimation for ML-G and REC-G.

Table 5 depicts the results obtained after counting the number of perfectly and overestimation of protected ML-G and REC-G gadgets. As it can be observed we obtained a 96% (184 vs. 192) accuracy (geomean) of perfectly protected ML-G callsites for MySQL while TypeArmor obtaines for the same program an 94% accuracy (geomean). Further, TYPESHIELD obtained a 97% (131 vs. 134) accuracy (geomean) for Node.js while TypeArmor obtained 95% accuracy on the same program. Further, for the REC-G case TYPESHIELD obtained an 94% (273 vs.289) exact argument accuracy for MySQL while TypeArmor had 86%. For Node.js TYPESHIELD obtained an exact parameter matching of 95% (69 vs. 72) while TypeArmor obtained an 96% perfect matching.

Overall TYPESHIELDs forward-edge policy obtained an perfect accuracy of 95% while TypeArmor obtained 92%. While this is not a big difference we point out that the remaining overestimated parameters represent 5% and this do not leave much room for the attacker to perform her attack.

5.8 Forward-Edge Policy vs. Other Tools (RQ8)

Target	IFCC	TypeArmor (CFI+CFC)	AT	TypeShield (count)	TypeShield (type)
Lighttpd	6	47	63	41	38
Memcached	1	14	14	13	10
ProFTPD	3	376	396	311	336
Pure-FTPD	0	4	13	13	7
vsftpd	1	12	10	7	7
PostgreSQL	12	2304	2504	2122	1786
MySQL	150	3698	5896	4403	3792
Node.js	341	4714	7230	5429	3864
geomean	7.6	162.1	216.6	172.0	141.8

Table 6: Calltargets per callsite reduction statistics.

Table 6 depicts a comparison between TYPESHIELD, TypeArmor and IFCC with respect to the count of calltargets per callsites. The values depicted in this table for TypeArmor and IFCC are taken from the original TypeArmor paper. Note that the smaller the geomean numbers are, the better the technique is. AT is a technique which allows calltargets that are address taken. IFCC is a compiler based solution and depicted here as a reference for what is possible when source code is available. TypeArmor and TypeShield on the other hand are binary-based tools. TYPESHIELD reduces the number of calltargets by up to 35% (geomean) when compared to the AT functions, by up to 41% (12 vs. 7) for a single test program and by 13% (geomean) when comparing with TypeArmor, respectively. Finally, TYPESHIELD represents a strong improvement w.r.t. calltarget per callsite reduction in binary programs.

5.9 Comparison with Shadow-Stack (RQ9)

The safe stack implementation of Abadi et al. [8] has the highest security level [24] w.r.t. backward-edge protection. This solution has: (1) a high rumtime overhead ($\geq$ 21%), (2) is not open source, (3) uses a proprietary binary analysis framework (*i.e.,* Vulcan), (4) reuses a restricted number of labels, *i.e.,* each function called from inside a function will get the same label, and (5) the shadow stacks can be disclosed by a motivate attacker. This labels will be stored in all function shadow stacks, see Figure 1 in [8].

For this reason we propose an alternative backward-edge protection solution which is more runtime efficient. In order to show the precision of TYPESHIELD backward-edge protection we will give the average number of legitimate return addresses for each calltarget return address and relate it to the total number of available addresses without any protection.

Program	Total #RA	Total #RATs	Total #RATs/RA	%RATs/RA w.r.t. prog. binary
MySQL	5896	3792	0.64	0.014%
Node.js	7230	3864	0.53	0.011%
geomean	6529	3827	0.58	0.012%

Table 7: Backward-edge policy statistics.

Table 7 depicts the statistics w.r.t. the backward-edge policy legitimate return targets. In Table 7 we use the following abbreviations: total number of return addresses (Total #RA), total (median) number of return address targets (Total #RATs), total (median) number of return addresses targets per return addresses (Total. # RATs/RA), percent of legitimate return address targets per return addresses w.r.t. the total number of addresses in the program binary (% RATs/RA w.r.t. program binary). By applying TYPESHIELD backward-edge policy we obtain a reduction of 0.43 ($1 - 0.58$) ratio (geomean) of total number of return addresses targets per return addresses over total number of return addresses which means that only 43% of the total number of return addresses are actual targets for the function returns. The results indicate a percentage of 0.012% (geomean) of the total addresses in the program binaries are legitimate targets for the function returns. This means that our policy can eliminate 99.98% (100% - 0.012%) of the addresses which an attacker can use for his attack inside the program binary.

6. Related Work

TypeArmor [34] is a binary instrumentation tool that can protect against COOP. TypeArmor uses a fine-grained CFI policy based on caller/callee (but only indirect callsites) matching, which checks during runtime if the number of provided and needed parameters match. TYPESHIELD is related to TypeArmor [34], since we also enforce strong binary-level invariants on the number of function parameters. Further, TYPESHIELD also aims for 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 calltargets to callsites, since we not only restrict based on the number of parameters, but also on the width of their types. This results in much smaller buckets that in turn can only target a smaller subset of all address-taken functions.

VCI [15] is a binary based tool (7.9%) based on DynInst which can protect forward edge indirect control flow violations based on reconstructing a quasi program class hierarchy (*i.e.,* no class root node and the edges are not directed). The authors claim that VCI is 10 times more precise w.r.t. reducing the calltarget set per callsite. In contrast to TYPESHIELD VCI can not protect backward-edge violations and we arguably due to the conservative analysis the VCI could skip some corner situations allowing not legitimate calltargets.

Marx [26] is most similar to VCI and as VCI this tool reconstructs the same type of quasi program class hierarchy. No runtime efficiency numbers were provided in the paper. The authors claim that Marx can recuperate a class hierarchy which is more precise than that of IDAPro. The paper is geared towards first providing a tool which can be used by analyst in order to reverse engineer a binary. The precision of the calltarget set reduction per callsite should be similar to those of VCI but no comparison was compared in the paper. Compared to TYPESHIELD Marx can not protect against backward-edge violations and arguably has in common with VCI several limitations.

The CFI based implementation of Abadi et al. [9] is the first binary based implementation of a shadow stack. While at first quite promising this implementation suffers from high performance overhead which is around 21% due to the fact that the inserted checks before each function return instruction are not runtime efficient. Further, this tool has high imprecision du to the fact that labels are reused and in this way not legitimate return addresses become legitimate, thus these could be exploited by a skilled attacker.

7. Conclusion

We presented TYPESHIELD the first open source binary hardening tool that can protect the forward and backward edges inside a stripped (no RTTI information) program binary. In our evaluation we evaluated TYPESHIELD with real open source programs and shown that the tool is practical and effective when protecting system binaries. Further our evaluation results indicate that TYPESHIELD can reduce the forward-edge legal call target of up to 41% while providing high precision and low overhead w.r.t. the backward-edge policy.

References

[1] Dynamorio. ˙http://dynamorio.org/home.html.

[2] Itanium c++ abi. ˙https://mentorembedded.github.io/cxx-abi/abi.html.

[3] J. gray. c++: Under the hood. 1994. ˙http://www.openrce.org/articles/files/jangrayhood.pdf.

[4] Bluelotus team, bctf challenge: bypass vtable read-only checks. 2015.

[5] C++ abi for the arm architecture. 2015. ˙http://infocenter.arm.com/help/topic/com.arm.doc.ihi0041e/IHI0041Ecppabi.pdf.

[6] Bypassing clang's safestack for fun and profit. In *Blackhat Europe*, 2016. ˙https://www.blackhat.com/docs/eu-16/materials/eu-16-Goktas-Bypassing-Clangs-SafeStack.pdf.

[7] Clang cfi. 2017. ˙https://clang.llvm.org/docs/ControlFlowIntegrity.html#cfi-strictness.

[8] Martin Abadi, Mihai Budiu, Úlfar Erlingsson, and Jay Ligatti. Control flow integrity. In *Proceedings of the 12th ACM Conference on Computer and Communications Security (CCS)*, 2005.

[9] Martin Abadi, Mihai Budiu, Úlfar Erlingsson, and Jay Ligatti. Control flow integrity principles, implementations, and applications. In *ACM Transactions on Information and System Security (TISSEC)*, 2009.

[10] Dennis Andriesse, Asia Slowinska, and Herbert Bos. Compiler-agnostic function detection in binaries. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2017.

[11] Andrew R. Bernat and Barton P. Miller. Anywhere, anytime binary instrumentation. In *Proceedings of the 10th ACM SIGPLAN-SIGSOFT workshop on Program analysis for software tools, (PASTE)*, 2011.

[12] Dimitar Bounov, Rami Gökhan Kici, and Sorin Lerner. Protecting c++ dynamic dispatch through vtable interleaving. In *Symposium on Network and Distributed System Security (NDSS)*, 2016.

[13] Stephen Crane, Stijn Volckaert, Felix Schuster, Christopher Liebchen, Per Larsen, Lucas Davi, Ahmad-Reza Sadeghi, Thorsten Holz, Bjorn De Sutter, and Michael Franz. 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.

[14] Lucas Davi, M. Hanreich, D. Paul, A.-R Sadeghi, P. Koeberl, D. Sullivan, O. Arias, and Y Jin. Hafix: hardware-assisted flow integrity extension. In *Design Automation Conference (DAC)*, 2015.

[15] Mohamed Elsabagh, Dan Fleck, and Angelos Stavrou. Strict virtual call integrity checking for c ++ binaries. In *ACM Asia Conference on Computer and Communications Security (ASIACCS)*, 2017.

[16] Isaac Evans, Fan Long, Ulziibayar Otgonbaatar, Howard Shrobe, Martin Rinard, Hamed Okhravi, and Stelios Sidiroglou-Douskosr. Control jujutsu: On the weaknesses of fine-grained control flow integrity. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015.

[17] Robert Gawlik and Thorsten Holz. Towards automated integrity protection of c++ virtual function tables in binary programs. In *Annual Computer Security Applications Conference (ACSAC)*, 2014.

[18] Istvan Haller, Enes Goktas, Elias Athanasopoulos, G. Portokalidis, and Herbert Bos. Shrinkwrap: Vtable protection without loose ends. In *Annual Computer Security Applications Conference (ACSAC)*, 2015.

[19] D. Jang, T. Tatlock, and S. Lerner. Safedispatch: Securing c++ virtual calls from memory corruption attacks. In *Symposium on Network and Distributed System Security (NDSS)*, 2014.

[20] Uday Khedker, Amitabha Sanyal, and Bageshri Sathe. *Data flow analysis: Theory and Practice*. CRC Press, 2009.

[21] Volodymyr Kuznetsov, László Szekeres, Mathias Payer, George Candea, R. Sekar, and Dawn Song. Code-pointer integrity. In *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2014.

[22] Bingchen Lan, Yan Li, Hao Sun, Chao Su, Yao Liu, and Qingkai Zeng. Loop-oriented programming: A new code reuse attack to bypass modern defenses. In *IEEE Trustcom/BigDataSE/ISPA*, 2015.

[23] Julian Lettner, Benjamin Kollenda, Andrei Homescu, Per Larsen, Felix Schuster, Lucas Davi, Ahmad-Reza Sadeghi, Thorsten Holz, and Michael Franz. Subversive-c: Abusing and protecting dynamic message dispatch. In *USENIX Annual Technical Conference (USENIX ATC)*, 2016.

[24] Burow Nathan, Scott A. Carr, Joseph Nash, Per Larsen, Michael Franz, Stefan Brunthaler, and Mathias Payer. Control-flow integrity: Precision, security, and performance. In *ACM Computing Surveys (CSUR)*, 2017.

[25] Ben Niu and Gang Tan. Modular control-flow integrity. In *ACM Conferece on Programming Language Design and Implementation (PLDI)*, 2014.

[26] Andre Pawlowski, Moritz Contag, Victor van der Veen, Chris Ouwehand, Thorsten Holz, Herbert Bos, Elias Athanasopoulos, and Cristiano Giuffrida. Marx : Uncovering class hierarchies in c++ programs. In *Symposium on Network and Distributed System Security (NDSS)*, 2017.

[27] Aravind Prakash, Xunchao Hu, and Heng Yin. Strict protection for virtual function calls in cots c++ binaries. In *Symposium on Network and Distributed System Security (NDSS)*, 2015.

[28] G. Ramalingam. The undecidability of aliasing. In *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 1994.

[29] Felix Schuster, Thomas Tendyck, Christopher Liebchen, Lucas Davi, Ahmad-Reza Sadeghi, and Thorsten Holz. Evaluating the effectiveness of current anti-rop defenses. In *Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses (RAID)*, 2014.

[30] Felix Schuster, Thomas Tendyck, Christopher Liebchen, Lucas Davi, Ahmad-Reza Sadeghi, and Thorsten Holz. Counterfeit object-oriented programming. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2015.

[31] Michael Theodorides and David Wagner. Breaking active-set backward-edge cfi. In *International Symposium on Hardware Oriented Security and Trust (HOST)*, 2017.

[32] Caroline Tice, Tom Roeder, Peter Collingbourne, Stephen Checkoway, Úlfar Erlingsson, Luis Lozano, and Geoff Pike. Enforcing forward-edge control-flow integrity in gcc and llvm. In *Proceedings of the USENIX conference on Security (USENIX SEC)*, 2014.

[33] Victor van der Veen, Dennis Andriesse, Enes Göktas, Ben Gras, Lionel Sambuc, Asia Slowinska, Herbert Bos, and Cristiano Giuffrida. Practical context-sensiticve cfi. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015.

[34] Victor van der Veen, Enes Goktas, Moritz Contag, Andre Pawlowski, Xi Chen, Sanjay Rawat, Herbert Bos, Thorsten Holz, Elias Athanasopoulos, and Cristiano Giuffrida. A tough call: Mitigating advanced code-reuse attacks at the binary level. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2016.

[35] Chao Zhang, Chengyu Song, Kevin Chen Zhijie, Zhaofeng Chen, and Dawn Song. vtint: Protecting virtual function tablesíntegrity. In *Proceedings of the Symposium on Network and Distributed System Security (NDSS)*, 2015.

[36] Chao Zhang, Tao Wei, Zhaofeng Chen, Lei Duan, Laszlo Szekeres, Stephen McCamant, Dawn Song, and Wei Zou. Practical control flow integrity & randomization for binary ex-

ecutables. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2013.

[37] Mingwei Zhang and R. Sekar. Control flow integrity for cots binaries. In *Proceedings of the USENIX conference on Security (USENIX SEC)*, 2013.