Lecture Notes on
Control Flow Safety

Matt Fredrikson

Carnegie Mellon University
Lecture 8

1 Introduction & Recap

In the last lecture we talked about enforcing a more granular type of memory safety policy to ensure that parts of our program don't read or write portions they aren't supposed to. This was motivated by our hypothetical career as an app developer who wants to monetize with advertising, and is thus compelled by Vladimir's discount ad shop to run untrusted rendering code within our program:

$$\texttt{if}(\textit{display ads}) \, \alpha \, \texttt{else} \, \textit{continue without ads}$$

We discussed sandboxing policies where a region of memory is designated for the untrusted α to "play" in, such as the upper portion of memory at addresses 8-15 in the diagram below.
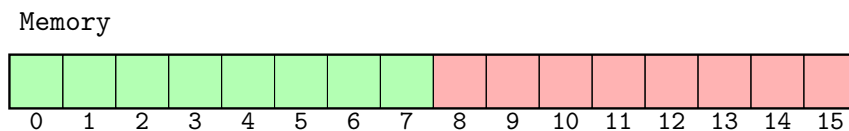
As long as we can enforce this policy, and we are careful about writing our program to save and restore variable state, then we can ensure that whatever the sandbox does will not affect the rest of our program's execution.

We discussed an approach called *software fault isolation* [SMB$^+$10, YSD$^+$09] (SFI) for properly isolating the malicious or buggy effects of α from the rest of our program. SFI works by inlining enforcement directly into α, changing its behavior so that it can't violate the sandbox policy and if it attempts to do so then it still won't have any effect on the rest of our execution. SFI is a very practical technique, and has been used effectively in real applications to isolate untrusted code execution from browsers, operating

systems, and other critical applications. In the next lab, you will implement a prototype SFI policy for your server.

Today we will look at a related technique called *control flow integrity* [ABEL09], which ensures that the attacker cannot influence the control flow of a program to diverge from a pre-defined control flow policy. But in order for this defense to have any purpose, we need to introduce indirect control flow commands into our language, bringing it closer yet to the features that real platforms in need of rigorous security defenses have in practice. We will then generalize the safety enforcement techniques discussed so far, introducing a flexible and practical method for enforcing safety policies provided as *security automata* [Sch00].

2 Indirect control flow

So far the programs that we have considered have a particularly nice property. Namely, once the programmer decides which commands are in the program and how they are sequenced together with compositions, conditionals, and while loops, then all of the possible sequences of commands that the program will ever execute are fixed once and for all. There is no way for a user to provide data that could cause some of the commands to be skipped over or added, and as long as the program is executed faithfully to the semantics, the control flow will be as the programmer envisioned when the program was written.

Programs executed on "real" machines do not enjoy this property, thanks to indirect transfers of control flow. An indirect control flow transfer is commonly implemented using a function pointer in high-level languages, or a `jmp` instruction with a pointer operand. We'll add this functionality to our language by considering a program command of the form:

$$\mathtt{if}(Q)\,\mathtt{jump}\,e \tag{1}$$

The command in (1) first tests whether a formula Q is true in the current state. If it is, then control transfers to the instruction indexed by the term e. Otherwise, control proceeds to the next instruction.

But this doesn't make much since yet, because we haven't discussed indexing of instructions. Programs in the simple imperative language are themselves just commands, which can be built from other commands by connecting them with composition, conditional, and looping constructs. We will now change the language so that rather than having structured high-level commands like $\mathtt{if}(Q)\,\alpha\,\mathtt{else}\,\beta$ and $\mathtt{while}(Q)\,\alpha$, we will assume that programs are sequences of unstructured "atomic" commands. So the commands are defined by the syntax:

$$\alpha \ ::= \ x := e \mid \mathtt{Mem}(e) := \tilde{e} \mid \mathtt{assert}(Q) \mid \mathtt{if}(Q)\,\mathtt{jump}\,e$$

Then a program Π is a finite sequence of commands,

$$\Pi = (\alpha_0, \alpha_1, \ldots, \alpha_n) \tag{2}$$

We will write $\Pi(i)$, where $0 \leq i \leq n$, to refer to the command α_i in program Π. If i is negative, or $n < i$, then $\Pi(i)$ is undefined.

Think of this language as a simplified version of assembly language. Memory update commands $\text{Mem}(e) := \tilde{e}$ correspond to store instructions (i.e., `mov` into a memory cell), memory dereference terms $\text{Mem}(e)$ correspond to memory fetch instructions (i.e., `mov` from a memory cell to a register), and $\text{if}(Q)$ jump e to conditional `jmp` instructions. We don't have an explicit stack or notion of procedure to worry about, but if we did then `halt` commands would correspond to `ret` instructions.

Semantics. Recall that program states ω are composed of a variable map ω_V and memory ω_M. Now that programs Π are composed of indexed commands, and can transfer control to any command in Π, states will also need to track a program counter ω_i that determines which command executes next. The program counter will range from $i \in 0$ to n, denoting that the command Π_i executes next.

Definition 1 (Operational semantics of programs). The small-step transition relation $\leadsto_\Pi$ of program Π composed of commands $\alpha_0, \ldots, \alpha_n$ in state $\omega = (\omega_i, \omega_V, \omega_M)$ is given by the following cases:

$$
(\omega_i, \omega_V, \omega_M) \leadsto_\Pi
\begin{cases}
(\omega_i + 1, \omega_V\{x \mapsto \omega[\![e]\!]\}, \omega_M) & \textit{if} \quad \Pi_i = x := e \text{ and } \omega[\![e]\!] \text{ is defined} \\
(\omega_i + 1, \omega_V, \omega_M\{\omega[\![e]\!] \mapsto \omega[\![\tilde{e}]\!]\}) & \textit{if} \quad \Pi_i = \text{Mem}(e) := \tilde{e} \text{ and} \\
& \qquad \omega[\![\tilde{e}]\!] \text{ is defined and } 0 \leq e < U \\
(\omega_i + 1, \omega_V, \omega_M) & \textit{if} \quad \Pi_i = \text{assert}(Q) \text{ and } \omega \models Q \\
(\omega[\![e]\!], \omega_V, \omega_M) & \textit{if} \quad \Pi_i = \text{if}(Q) \text{ jump } e \text{ and} \\
& \qquad 0 \leq \omega[\![e]\!] \leq n \text{ and } \omega \models Q \\
(\omega_i + 1, \omega_V, \omega_M) & \textit{if} \quad \Pi_i = \text{if}(Q) \text{ jump } e \text{ and} \\
& \qquad 0 \leq \omega[\![e]\!] \leq n \text{ and } \omega \models \neg Q \\
\Lambda & \textit{if} \quad \text{otherwise and } 0 \leq \omega_i \leq n
\end{cases}
$$

Having defined the small-step transition semantics, we can define the trace semantics as all sequences of states $\omega_1, \omega_2, \ldots$ that either terminate, diverge (i.e. terminate in no state and run forever), or abort by terminating in $\omega = \Lambda$.

Definition 2 (Trace semantics). Given a program Π composed of commands $\alpha_0, \ldots, \alpha_n$, the trace semantics $[\![\Pi]\!]$ is the set of traces obtainable by repeated application of the small-step relation $\leadsto_\Pi$.

$$[\![\Pi]\!] = \{(\omega_0, \omega_1, \ldots) \ : \ \omega_i \leadsto_\Pi \omega_{i+1} \text{ for all indices } 0 \leq i \text{ of the trace}\}$$

The definitions of terminating, diverging, and aborting traces are the same as they were in previous definitions of the trace semantics.

Example.   Consider the following program to illustrate how the operational and trace semantics work.

$$
\begin{aligned}
&\texttt{0:}\quad \texttt{assert}(0 \leq x)\\
&\texttt{1:}\quad \texttt{Mem}(0) := \texttt{Mem}(0) + 1\\
&\texttt{2:}\quad x := x - 1\\
&\texttt{3:}\quad \texttt{if}(0 \leq x)\,\texttt{jump}\,1
\end{aligned}
$$

Suppose that we begin in the following state:

$$
\begin{aligned}
\omega_i &= 0\\
\omega_V(x) &= 2 \text{ and all other variables map to } 0\\
\omega_M(i) &= 0 \text{ for all } 0 \leq i < U
\end{aligned}
$$

Then consulting the operational semantics, we see that $\Pi_0 = \texttt{assert}(0 \leq x)$ and $\omega \models 0 \leq x$, so the next state is $(\omega_i + 1, \omega_V, \omega_M)$.

$$
(0, \omega_V, \omega_M) \rightsquigarrow_\Pi
$$
$$
(1, \omega_V, \omega_M)
$$

Now $\Pi_1 = \texttt{Mem}(0) := \texttt{Mem}(0) + 1$ and $\texttt{Mem}(0) = 0$. So the next state is $(2, \omega_V, \omega_M\{0 \mapsto 1\})$.

$$
(0, \omega_V, \omega_M) \rightsquigarrow_\Pi
$$
$$
(1, \omega_V, \omega_M) \rightsquigarrow_\Pi
$$
$$
(2, \omega_V, \omega_M\{0 \mapsto 1\})
$$

Now $\Pi_2 = x := x - 1$ and the semantics tell us to update ω_V at x.

$$
(0, \omega_V, \omega_M) \rightsquigarrow_\Pi
$$
$$
(1, \omega_V, \omega_M) \rightsquigarrow_\Pi
$$
$$
(2, \omega_V, \omega_M\{0 \mapsto 1\}) \rightsquigarrow_\Pi
$$
$$
(3, \omega_V\{x \mapsto 0\}, \omega_M\{0 \mapsto 1\})
$$

We now get to the jump because $\Pi_3 = \texttt{if}(0 \leq x)\,\texttt{jump}\,1$. The number of instructions $n = 3$, so the next state has program counter 1. We continue in this way, until we reach the conditional jump again. At this point $\omega_V(x) = -1$, and so the program counter increments to 4.

$$
(0, \omega_V, \omega_M) \rightsquigarrow_\Pi
$$
$$
(1, \omega_V, \omega_M) \rightsquigarrow_\Pi
$$
$$
(2, \omega_V, \omega_M\{0 \mapsto 1\}) \rightsquigarrow_\Pi
$$
$$
(3, \omega_V\{x \mapsto 0\}, \omega_M\{0 \mapsto 1\}) \rightsquigarrow_\Pi
$$
$$
(1, \omega_V\{x \mapsto 0\}, \omega_M\{0 \mapsto 1\}) \rightsquigarrow_\Pi
$$
$$
(2, \omega_V\{x \mapsto 0\}, \omega_M\{0 \mapsto 1\}\{0 \mapsto 2\}) \rightsquigarrow_\Pi
$$
$$
(3, \omega_V\{x \mapsto 0\}\{x \mapsto -1\}, \omega_M\{0 \mapsto 1\}\{0 \mapsto 2\}) \rightsquigarrow_\Pi
$$
$$
(4, \omega_V\{x \mapsto 0\}\{x \mapsto -1\}, \omega_M\{0 \mapsto 1\}\{0 \mapsto 2\})
$$

Now the program counter is outside the instruction bounds in Π. The operational semantics does not define a next state, so the computation effectively terminates.

3 Control Flow Integrity

Let's return to our problem with untrusted advertising code. Now that the language α is written in can make indirect jumps, what could go wrong? Assuming that we are using SFI to enforce a sandboxing policy, there is still no way for α to read or write memory outside the sandbox. Is this true? Consider the following situation, where we can assume that SFI has been applied to the untrusted α starting at command 20.

$$
\begin{array}{rl}
& \vdots \\
10: & z := \texttt{Mem}(x) \\
11: & \texttt{if}(i \geq 0)\,\texttt{jump}\,y \\
& \vdots \\
20: & i := 0 \\
21: & x := \textit{attacker's desired address} \\
\alpha \Big\{ \quad 22: & y := 24 \\
23: & \texttt{if}(0 = 0)\,\texttt{jump}\,10 \\
24: & \textit{copy memory contents from } z \\
& \vdots
\end{array}
$$

Here, our original program (not the untrusted α) dereferences memory and makes use of indirect control flow transfer. More specifically,

1. At command 10, it dereferences memory on the variable x and stores the result into z. Because this command is not in the untrusted portion α, it was not rewritten with SFI and can readily access memory outside the sandbox.

2. At command 11, the program tests $i \geq 0$, and if the test holds then jumps to whatever command is currently in y.

3. The untrusted code sets up the program state: *(i)* the indirect jump at 11 will occur by setting $i := 0$ on line 20; *(ii)* the indirect jump at line 11 will return control to α by setting $y := 24$; *(iii)* setting $x := \ldots$ at 21 so that the address read at line 10 will be whatever the attacker wants, presumably outside the sandbox bounds.

4. The command at 23 then executes an indirect jump on a trivial test, targeting 10 so that the attacker's choice of memory is read and control returns to α after the indirect jump at 11.

To summarize, the attacker identifies a sequence of commands in the trusted portion of the program, and sets things up in a way so that unauthorized memory is copied into a variable that the attacker can later access once control is returned to the untrusted code.

This should remind you of a *return-oriented programming* (ROP) attacks [Sha07] that you learned about in 15-213. If we assume that the attacker knows the text of our program, then it is possible for them to identify "gadgets" in *code that we wrote* to do their bidding. But this crucially relies on the ability to change control flow using indirection so that commands are executed in the order needed by the attacker to carry out their goals.

3.1 Coarse-grained safety

How can we prevent this? One idea is to use the same approach as we did for SFI. In that case, we designed a sandbox between memory address s_l and s_h, and then rewrote the indices in all memory operations to ensure that accesses stayed within those bounds. Perhaps we can do a very similar thing here, by assigning a "code sandbox" between commands at $\mathtt{pc}_l$ and $\mathtt{pc}_h$. Then we can rewrite indirect jump commands to ensure that their target always lies within these bounds.

$$\text{Rewrite all } \mathtt{if}(Q)\,\mathtt{jump}\,e \text{ commands as } \mathtt{if}(Q)\,\mathtt{jump}\,(e\,\&\,\mathtt{pc}_h)\mathbin{|}\mathtt{pc}_l \tag{3}$$

This is a form of *control flow integrity* (CFI) [ABEL09], a technique for enforcing a broad class of safety properties that place limits on the allowed control flow paths in a program. As long as we choose $\mathtt{pc}_l$ and $\mathtt{pc}_h$ to satisfy similar conditions as those used in Theorem 1 from the previous lecture, i.e.

$$0 \le \mathtt{pc}_l \le (e\,\&\,\mathtt{pc}_h)\mathbin{|}\mathtt{pc}_l \le \mathtt{pc}_h \le n \tag{4}$$

then we can ensure that the untrusted code will never jump out of its sandbox. This seems to address our concerns with the advertising scenario, when everything untrusted resides in a well-specified region of code known in advance.

4 Finer-grained control-flow safety

But what if this isn't the case? Suppose that we want to enforce other invariants on untrusted code, such as that they do not modify a protected variable under certain conditions. So for example if x is negative, then we want to jump over any assignment to x. We make the following replacements, among others:

$$\mathtt{i}{:}\quad x := e \qquad \text{becomes} \qquad \begin{array}{ll}\mathtt{i}{:} & \mathtt{if}(x < 0)\,\mathtt{jump}\,i+2 \\ \mathtt{i{+}1}{:} & x := e\end{array}$$

Now if we use the coarse-grained CFI policy from before, can we actually enforce the policy using this approach? It would seem not, at least as long as the attacker knows that this is how we will attempt to do so. The problem arises because of the fact that according to the coarse-grained CFI policy, any address in the untrusted code is an allowed target of a jump. So it is perfectly acceptable (according to the coarse-grained policy) for the attacker to jump directly past the inlined check.

To see this more concretely, consider the following proof-of-concept attack code. Suppose that the attacker wants to set x to 0 regardless of what its value is before the

untrusted code executes. This obviously violates the policy, and to accomplish it the attacker will provide a program that takes the inline enforcement code into account. So after providing the code on the left,

$$
\begin{array}{ll}
\texttt{i-1:} & \texttt{if}(0 = 0)\,\texttt{jump}\,i + 1 \\
\texttt{i:} & x := 0
\end{array}
\qquad \text{becomes} \qquad
\begin{array}{ll}
\texttt{i-1:} & \texttt{if}(0 = 0)\,\texttt{jump}\,(i + 1 \,\&\, s_h) \mid s_l \\
\texttt{i:} & \texttt{if}(x < 0)\,\texttt{jump}\,i + 2 \\
\texttt{i+1:} & x := 0
\end{array}
$$

In short, the attacker sets up the program to jump directly over the enforcement code.

Enforcing the control flow graph. To address this, we can rely on the control flow graph (CFG) of the untrusted code. Recall from lecture 5 that a program's CFG is a graph that encodes all of the possible valid transitions between commands in the program. In this case, we will obtain a control flow graph for the original untrusted program[1], and ensure that the program instrumented with inline safety checks follows the same CFG modulo any checks.

So in this case, the original control flow graph is given on the left of the diagram below. Obviously it is not the case that $0 \neq 0$, so the edge from i-1 to i is never taken. After the invariant instrumentation is inserted, the correct translation of the control flow, preserving the relative edges from the original CFG, is shown on the right. The instrumentation replaced the instruction originally at i with an inline check, and shifted all of the subsequent instructions up by one address. So this moves i to i+1, and i+1 to i+2. To make this clear in the diagram, the nodes and edges corresponding to instrumentation are marked in blue.

---
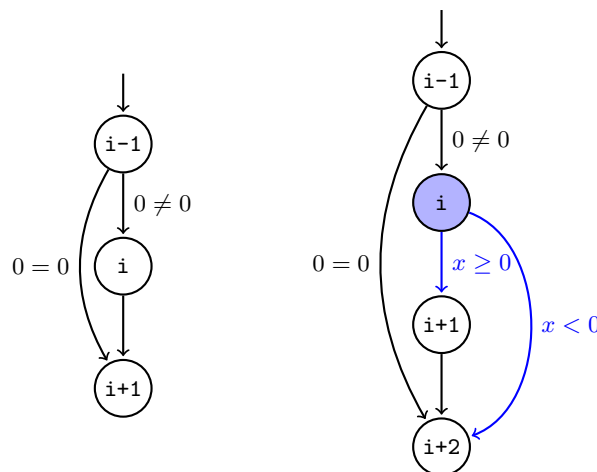
[1]Obtaining the CFG for an arbitrary program with indirect jumps is a difficult problem indeed. It may not always be possible to do so, and we will come back to this in later lectures. For now, we will just assume that we have obtained the correct CFG for α by some unknown means.

Now to correctly enforce the safety policy that the CFG on the right is respected by the code, the original jumps are rewritten accordingly.

$$\begin{aligned} \texttt{i-1:} \quad & \texttt{if}(0 = 0)\,\texttt{jump}\,i + 2 \\ \texttt{i:} \quad & \texttt{if}(x < 0)\,\texttt{jump}\,i + 2 \\ \texttt{i+1:} \quad & x := 0 \end{aligned}$$

By enforcing the original control flow of the program, after taking any added instrumentation into account with dealing with instruction addresses, we prevented the attacker from bypassing our inlined safety policy enforcement. But notice that it didn't really matter what control flow graph we started out with. It could have been arbitrary, perhaps completely different from the actual CFG of the original program, and we could still enforce it by inserting and replacing conditional jumps.

References

[ABEL09] Martín Abadi, Mihai Budiu, Úlfar Erlingsson, and Jay Ligatti. Control-flow integrity: Principles, implementations, and applications. *ACM Transactions on Information and Systems Security*, 13(1):4:1–4:40, November 2009.

[Sch00] Fred B. Schneider. Enforceable security policies. *ACM Transactions on Information Systems Secur.*, 3(1):30–50, February 2000.

[Sha07] Hovav Shacham. The geometry of innocent flesh on the bone: Return-into-libc without function calls (on the x86). In *Proceedings of CCS 2007*, October 2007.

[SMB+10] David Sehr, Robert Muth, Cliff Biffle, Victor Khimenko, Egor Pasko, Karl Schimpf, Bennet Yee, and Brad Chen. Adapting software fault isolation to contemporary cpu architectures. In *Proceedings of the 19th USENIX Conference on Security*, 2010.

[YSD+09] Bennet Yee, David Sehr, Greg Dardyk, Brad Chen, Robert Muth, Tavis Ormandy, Shiki Okasaka, Neha Narula, and Nicholas Fullagar. Native client: A sandbox for portable, untrusted x86 native code. In *IEEE Symposium on Security and Privacy*, 2009.