Automatic Detection of Speculative Execution Combinations

Xaver Fabian
Cispa Helmholtz Center for Information Security
Saarbrücken, Germany
xaver.fabian@cispa.de

Marco Guarnieri
IMDEA Software Institute
Madrid, Spain
marco.guarnieri@imdea.org

Marco Patrignani
University of Trento
Trento, Italy
marco.patrignani@unitn.it

ABSTRACT

Modern processors employ different speculation mechanisms to speculate over different kinds of instructions. Attackers can exploit these mechanisms simultaneously in order to trigger leaks of speculatively-accessed data. Thus, sound reasoning about such speculative leaks requires accounting for all potential speculation mechanisms. Unfortunately, existing formal models only support reasoning about fixed, hard-coded speculation mechanisms, with no simple support to extend said reasoning to new mechanisms.

In this paper we develop a framework for reasoning about composed speculative semantics that capture speculation due to different mechanisms and implement it as part of the SPECTCTOR verification tool. We implement novel semantics for speculating over store and return instructions and combine them with the semantics for speculating over branch instructions. Our framework yields speculative semantics for speculating over any combination of these instructions that are secure by construction, i.e., we obtain these security guarantees for free. The implementation of our novel semantics in SPECTCTOR let us verify programs that are vulnerable to Spectre v1, Spectre v4, and Spectre v5 vulnerabilities as well as new snippets that are only vulnerable to their compositions.

CCS CONCEPTS

• Security and privacy → Formal security models; Systems security.

KEYWORDS

Spectre; Speculative Execution; Speculative information flows; Speculative non-interference; Combinations of speculative semantics

ACM Reference Format:


1 INTRODUCTION

Speculative execution avoids pipeline stalls by predicting intermediate results and by speculatively executing instructions based on such predictions. When a prediction turns out to be incorrect, the processor squashes the speculative instructions, thereby rolling back their effect on the architectural state. Speculative instructions, however, leave footprints in microarchitectural components (like caches) that persist even after speculative execution terminates. As shown by Spectre [24], attackers can exploit these side effects to leak information about speculatively accessed data.

Modern general-purpose processors have different speculation mechanisms (branch predictors, memory disambiguators, etc.) that are used to speculate over different kinds of instructions: conditional branching [24], indirect jumps [24], store and load operations [23], and return instructions [25]. While well-known attacks target only individual speculation mechanism (e.g., Spectre-PHT [24] targets branch predictors), some speculative leaks only arise due to the interaction of multiple mechanisms.

Listing 1: Speculative leak arising from speculation over branch and store instructions combined.

1 x = 0;
2 p = &secret;
3 p = &public;
4 if (x != 0)
5 temp &= A[*p];

For example, the code in Listing 1 can speculatively leak the value of $secret in Line 5 whenever (1) the memory write to $p in Line 3 is predicted to have a different address than the memory read $*p on Line 5, and (2) the branch instruction on Line 4 is mispredicted as taken. This leak, therefore, arises from the combination of two speculation mechanisms: branch prediction and memory disambiguation prediction. Hence, leaks like the one in Listing 1 are missed by sound analyses for speculative leaks that consider speculation over only one of these speculation mechanisms.

Sound reasoning about speculative leaks requires accounting for all potential speculation mechanisms. However, existing formal models (also called speculative semantics) support multiple speculation mechanisms poorly. Some of them support only fixed speculation mechanisms: branch prediction [20, 21, 38–40] and (in addition) memory disambiguation prediction [9, 14, 32]. Furthermore, the different speculation mechanisms are hard-coded into the formal semantics [9, 14, 21]. Extending these semantics with new speculation mechanisms (e.g., speculation over return addresses or value prediction) requires changes to the formal model and to any security proof relying on it. This is not a scalable approach for developing comprehensive formal models and analyses for speculative leaks.

In this paper we develop a framework for composing speculative semantics that capture speculation due to different mechanisms and implement it as part of the SPECTCTOR verification tool. The combination yields a single operational semantics that can be used
to reason about leaks involving all speculation mechanisms of the components (as in Listing 1). Our framework lets us define the speculative semantics of each mechanism independently, which leads to simpler formalisation. Additionally, the security of the composed semantics is derived automatically from the security of its sub-parts, maximising proof reuse. Finally, the composed semantics can be easily implemented in SPECTCTOR, which can be used to verify the absence of leaks like those in Listing 1.

Concretely, this paper makes the following contributions:
- It introduces $\mathcal{L}_\mathit{S}$ and $\mathcal{L}_\mathit{R}$, two novel semantics for speculation over store and return instructions (Section 3).
- It defines the framework for composing different speculative semantics and formalises its key properties: if the individual semantics fulfil some (expected) security conditions (which we prove for all the semantics we combine), then the composed semantics is also secure (Section 4).
- It instantiates the framework with $\mathcal{L}_\mathit{S}$, $\mathcal{L}_\mathit{R}$, and $\mathcal{L}_\mathit{B}$, the semantics for speculation over branch instructions from [21], creating all the possible compositions ($\mathcal{L}_{\mathit{S}+\mathit{R}}, \mathcal{L}_{\mathit{B}+\mathit{S}}, \mathcal{L}_{\mathit{B}+\mathit{R}}, \mathcal{L}_{\mathit{B}+\mathit{S}+\mathit{R}}$) and proving their security (Section 5). All these semantics are mechanised in Coq, and we write $\phi$ to indicate when traces are calculated mechanically.
- It extends the SPECTCTOR verification tool with all these semantics and validates this extension on both existing benchmarks (for speculation on store and return instructions) as well as on new snippets (for combined speculation) that we define (Section 6).

The rest of the paper first presents background notions, such as the security notion we rely on, and the formal language we extend with the novel speculative semantics (Section 2) and then related work (Section 8) and conclusions (Section 9).

Additional material: Full details of the semantics and proofs can be found in the technical report available at [18]. The extended version of SPECTCTOR is available at [16], whereas the mechanisation of our speculative semantics in Coq are available at [17].

2 BACKGROUND: $\mu$ASM, SPECULATIVE SEMANTICS AND SECURITY DEFINITION

This section first describes the attacker model and the security definition we consider (Section 2.1). Then, it presents the syntax (Section 2.2) and the semantics (Section 2.3) of $\mu$ASM, a simple assembly-style language, followed by $\mathcal{L}_\mathit{B}$, the semantics for speculation over branch instructions (Section 2.4). Most of the notions that we overview next are taken from Guarnieri et al. [21].

2.1 Attacker Model and Security Definition

We adopt a commonly-used attacker model [3, 9, 14, 19–21, 31, 38], a passive attacker observing the execution of a program through events $\tau$. These events, which we call observations, model timing leaks through cache and control flow while abstracting away low-level microarchitectural details.

$$
\text{Obs} ::= \text{load } n \mid \text{store } n \mid \text{pc } n \mid \text{call } f \mid \text{ret } n \mid \text{start}_x n \mid \text{rlbx}_x n \mid \tau ::= \epsilon \mid \text{Obs} \mid \text{start}_x n \mid \text{rlbx}_x n
$$

The $\text{store } n$ and $\text{load } n$ events denote read and write accesses to memory location $n$, so they model cache leakage. In contrast, $\text{pc } n$, $\text{call } f$, and $\text{ret } n$ events record the control-flow of the program. The $\text{start}_x n$ and $\text{rlbx}_x n$ observations denote the start and the finish of a speculative transaction [21] (with identifier $n$) produced by the speculative semantics $x$ (we use $x$ and $y$ to range over the speculative semantics we define later).

An observation $\tau$ is either an event $\text{Obs}$ or the empty observation $\epsilon$. Traces $\mathcal{T}$ are sequences of observations; we indicate sequences of elements $[\epsilon_1; \cdots; \epsilon_n]$ as $\tilde{e}$, and adding an element $e$ to $\tilde{e}$ as $\tilde{e} \cdot e$.

The non-speculative projection $\mid_{\mathit{ns}}$ of a trace $\mathcal{T}$ deletes all speculative observations by removing all sub-traces enclosed between $\text{start}_x n$ and $\text{rlbx}_x n$. The remaining trace, then, captures all non-speculative observations.

Speculative Non-Interference: With this trace model we can define the security property we use in this paper: Speculative Non-Interference (SNI) [21]. Intuitively, SNI requires that programs do not leak more information under the speculative semantics than under the non-speculative semantics.

SNI is parametric in a policy $\phi$, which describes public/low information for the program, and in the used speculative semantics $x$, which models how the program executes. Following Guarnieri et al. [21], a policy $\phi$ consists of a list of public registers and public memory locations. Two configurations $\sigma^1, \sigma^2$ are called low-equivalent for a policy $\phi$, written $\sigma^1 \equiv_{\phi} \sigma^2$, if they agree on all register and memory locations in $\phi$. The speculative semantics $x$ defines how (speculative) traces describing the program behaviour are generated. We indicate that program $p$ generates trace $\mathcal{T}$ from state $s$ with semantic $x$ as $\text{Beh}_x^A(p, s) = \mathcal{T}$. We formalise multiple speculative semantics in later sections, each one instantiating $\text{Beh}_x^A(p, s)$.

A program $p$ satisfies SNI (Definition 1) for a speculative semantics $x$ if any pair of low-equivalent initial configurations $\sigma^1$ and $\sigma^2$ that generate the same observations without speculative events also generate the same observations with speculative events too.

Definition 1 (SNI). Program $p$ satisfies SNI (denoted $p \vdash_x \text{SNI}$) if for all $\sigma^1, \sigma^2$, if $\sigma^1 \equiv_{\phi} \sigma^2$ and $\text{Beh}_x^A(p, \sigma^1) \mid_{\mathit{ns}} = \text{Beh}_x^A(p, \sigma^2) \mid_{\mathit{ns}}$, then $\text{Beh}_x^A(p, \sigma^1) = \text{Beh}_x^A(p, \sigma^2)$.

2.2 $\mu$ASM

(Expressions) $e ::= n \mid x \mid \text{or } e \mid e_1 \otimes e_2$

(Instructions) $i ::= \text{skip } \mid x \leftarrow e \mid \text{load } x, e \mid \text{store } x, e \mid \text{jmp } e \mid \text{beqz } x, l \mid x \leftarrow e \mid \text{sbarr } \mid \text{call } f \mid \text{ret}$

$\mu$ASM is an assembly-like language whose syntax is presented above. Programs $p$ in $\mu$ASM are sequences of mappings from natural numbers $n$ (i.e., the instruction address) to instructions $i$ or $\bot$. Instructions include skipping, register assignments, loads, stores, indirect jumps, conditional branches, conditional assignments, speculation barriers, calls, and returns. Instructions can refer to expressions, which are constructed by combining registers and values with unary and binary operators. Registers come from the set $\text{Regs}$, containing register identifiers and designated registers $\text{pc}$ and $\text{sp}$ modelling the program counter and stack pointer respectively, while values come from the set $\text{Vals}$, which includes natural numbers and $\bot$. 


In the following, we use instruction keywords to denote the set of all instructions of a given type. For instance, beqz is the set of all branch instructions, i.e., \( \text{beqz} = \{ \text{beqz} \, x, l \mid x \in \text{Regs} \land l \in \text{Vals} \} \).

### 2.3 Non-speculative Semantics of \( \mu \text{ASM} \)

\( \mu \text{ASM} \) has a small-step operational non-speculative semantics \( \rightarrow \) that describes how programs execute without speculative execution. The judgment for this semantics is \( (p, \sigma) \xrightarrow{\text{Rule Beqz-Sat}} (p, \sigma') \) and it reads: “A program state \( (p, \sigma) \) steps to a new program state \( (p, \sigma') \) producing observation \( \tau \)”. Program states \( (p, \sigma) \) consist of the program \( p \) and the configuration \( \sigma \). The program \( p \) is used to look up the current instruction, whereas the configuration \( \sigma = (m, a) \) is used to read from/write to the memory \( m \) and register file \( a \). Memories map addresses (which are natural numbers) to values, whereas register files map register identifiers to values.

Most of the rules of the semantics are standard and thus omitted; we present selected rules below (see [21] for all rules). The rules rely on the evaluation of expressions (indicated as \( \llbracket e \rrbracket(a) = n \) where expression \( e \) is evaluated to value \( n \) under register file \( a \). In the rules, \( a[x \mapsto n] \), where \( x \in \text{Regs} \cup \text{In} \) and \( n \in \text{Vals} \) denotes the update of a map (memory or registers), whereas \( a(x) \) denotes reading from a map. Finally, \( \sigma(x) \), where \( x \in \text{Regs} \) and \( \sigma = (m, a) \), denotes \( a(x) \).

#### 2.3.1 Symbolic semantics

Following [21], we introduce a symbolic non-speculative semantics \( \rightarrow^S \) that is at the basis of SPECTATOR’s analysis. This symbolic semantics differs from \( \rightarrow \) in two key ways: (1) concrete configurations \( \sigma \) are replaced with symbolic configurations \( \sigma^S \), and (2) path condition constraints are generated in the standard way and they are encoded as part of the symbolic trace \( \tau \). Given a symbolic trace \( \tau \), \( \mu(\tau) \) denotes the set of all concrete traces that can be obtained by concretizing \( \tau \) with values consistent with \( \tau \)’s path condition. The symbolic non-speculative behavior \( \text{Beh}^S_{\text{NS}}(p) \) of a program \( p \) consists of all symbolic traces derived by applying \( \rightarrow^S \), and \( \mu(\text{Beh}^S_{\text{NS}}(p)) \) is the set of all concrete traces derived from \( p \)’s symbolic traces. As proved by Guarnieri et al. [21], \( \text{Beh}^S_{\text{NS}}(p) = \mu(\text{Beh}^S_{\text{NS}}(p)) \).

#### 2.4 \( \Ld_B \): Speculating Over Branch Instructions

To model and reason about the effects of speculation over branch instructions, Guarnieri et al. [21] propose three related semantics: an always-mispredict semantics (Section 2.4.1), an oracle semantics (Section 2.4.2), and a symbolic semantics (Section 2.4.3). The always-mispredict semantics, our main focus, is a safe overapproximation of the oracle semantics, which explicitly models the behavior of the branch predictor using a prediction oracle. Finally, the symbolic semantics, which is used in the SPECTATOR program analysis tool, is the symbolic version of the always-mispredict semantics. We summarize the properties of these semantics in Section 2.4.4. With a slight abuse of notation, \( \Ld_B \) indicates both the three speculative semantics, and the AM one alone (since it is the most relevant one).

#### 2.4.1 Always-mispredict (AM) Semantics

At every branch instruction, the always-mispredict semantics first speculatively executes the wrong branch for a fixed number of steps and then continues with the correct one. As a result, this semantics is deterministic and agnostic to implementation details of the branch predictor [21].

The state \( \Sigma_B \) of the AM semantics is a stack of speculative instances \( \Phi_B \) where reductions happen only on top of the stack. Whenever we start speculating, a new instance is pushed on top of the stack (Rule B:AM-branch). The instance is then popped when speculation ends (Rule B:AM-Rollback). Each instance \( \Phi_B \) contains the program \( p \), a counter \( ctr \) that uniquely identifies the speculation instance, a configuration \( \sigma \), and the remaining speculation window \( n \) describing the number of instructions that can still be executed speculatively (or \( 1 \), when no speculation is happening). Throughout the paper, we fix the maximal speculation window, i.e., the maximum number of speculative instructions, to a global constant \( \omega \).

\[ \Sigma_B := \Ld_B \quad \Phi_B := (p, ctr, \sigma, n) \]

This judgement for the AM semantics is: \( \Sigma_B \xrightarrow{\text{B:AM-branch}} \Sigma_B^' \)

\[ p(\sigma(pc)) \xrightarrow{\text{beqz \: x, l}} (p, \sigma', j) \]

\[ \sigma' = \sigma[pc \mapsto l] \quad \tau = \tau \quad \text{start}_{\Phi_B} \quad \text{ctr} \cdot \text{pc} \cdot l \]

\[ l' = \sigma(pc) + 1 \quad \text{if} \quad \sigma'(pc) = l \]

\[ l' = \sigma'(pc) - 1 \quad \text{if} \quad \sigma'(pc) \neq l \]

\[ \tau \xrightarrow{\text{AM-NoSpec}} (p, ctr, \sigma', n) \]

\[ (p, ctr, \sigma, n + 1) \xrightarrow{\text{AM-NoSpec}} (p, ctr, \sigma', n) \]

\[ \tau \xrightarrow{\text{AM-Branch}} (p, ctr, \sigma', n) \]

\[ \tau \xrightarrow{\text{AM-Rollback}} (p, ctr, \sigma', n) \]
When speculation ends, Rule $\text{B}$ will instantiate $\text{Z}$ for all semantics in this paper. In the definition we indicate that a properties in a single definition (Definition 2), which we will prove.

2.4.4 Properties of symbolic semantics look like those of the AM one, and the behaviour $\sigma\Sigma$ are that (1) concrete states $[21]$. The only two differences w.r.t. the AM semantics $\text{B}$ are replaced with symbolic states $\sigma\Sigma$ instead of concrete configurations $\sigma$, and (2) the semantics uses the symbolic non-speculative semantic instead of the concrete one. The rules of the symbolic semantics look like those of the AM one, and the behaviour $\text{Beh}_B^S(p)$ of a program $p$ is defined as for the AM semantics.

2.4.2 Oracle Semantics. The oracle semantics explicitly models the branch predictor using an oracle $O$, that relies on the branching history $h$ of the program $p$ to predict branch outcomes.

Here, we quickly summarize the key differences with the AM semantics; see [21] for the full definition. First, speculative instances are extended to track the branching history $h$, which records the outcomes of prior branch instructions. Second, when executing a beqz instruction, the oracle predicts the branch outcome (based on the branching history $h$) and a new speculative instance is pushed on top of the stack. Finally, whenever the speculation window of an instance anywhere on the stack reaches 0, the execution needs to be rolled back or committed. Rolling back deletes all the instances above the rolled back instance, whereas committing updates the configuration, the counter and the branching history $h$ of the instance below and the committed instance is deleted.

As before, the behaviour $\text{Beh}_B^O(p)$ of a program $p$ under the oracle semantics is the set of all its traces generated from an initial state until termination using the reflexive-transitive closure of $\text{Z}_B$.

3.1 $\mathcal{L}_S$: Speculation on Store Instructions

Modern processors write stores to main memory asynchronously to reduce delays caused by the memory subsystem. For this, processors employ a Store Queue where not-yet-committed store instructions are stored before being permanently written to memory. When executing a load instruction, the processor first inspects the store queue for a matching memory address. If there is a match, the value is retrieved from the store queue (called store-to-load forwarding), and otherwise the memory request is issued to the memory subsystem. To speed up computation, processors employ memory disambiguation predictors to predict if memory addresses of loads and stores match. Since the prediction can be incorrect, processors may speculatively bypass a store instruction in the store queue leading to a load instruction retrieving a stale value.

Example 1 (Store Speculation Vulnerability). Consider the example in Listing 2:

```
Listing 2: Code vulnerable to store speculation.
1. p = &secret;
2. p = &public;
3. temp = B[*p * 512];
```
Whenever the speculative window reaches 0. The key difference
store value from the stale
Ze. A transaction by pushing a new speculative instance on top of the
counter without updating the memory and starts a new speculative
instance from the stack.

This section first introduces the extended trace model required
to talk about speculation over store instructions (Section 3.1.1).
Next, it presents the speculative AM semantics (Section 3.1.2) and
the corresponding oracle semantics (Section 3.1.3) and symbolic
semantics (Section 3.1.4). This semantics is a secure speculative
semantics (Theorem 2).

**Theorem 2** ($\mathcal{L}_S$ is SSS). $\vdash \mathcal{L}_S$ SSS

3.1.1 Extended Trace Model. We extend the trace model $\text{Obs}$ with
starts $n$ and rlb $n$ observations to mark start and end of a spec-
ulative transaction $n$ started by a store bypass. Furthermore, we add a
bypass $n$ observation denoting that the store instruction at pro-
gram counter $n$ was speculated bypassed.

$$\text{Obs}_S := \text{Obs} \cup \text{starts} n \cup \text{rlb} n \cup \text{bypass} n$$

3.1.2 Speculative Semantics. The overall structure of the $\mathcal{L}_S$ se-
manics is similar to that of $\mathcal{L}_B$: speculative execution is modeled
using a stack of speculative states, instructions that do not start
speculative transactions are executed by delegating back to the non-
speculative semantics, and speculative transactions are rolled back
whenever the speculative window reaches 0. The key difference
between $\mathcal{L}_S$ and $\mathcal{L}_B$ is the differing source of speculation:
beq instructions for $\mathcal{L}_B$ and store instructions for $\mathcal{L}_S$.

The states used in $\mathcal{L}_S$ are similar to those of $\mathcal{L}_B$:

- **Spec. States** $\Sigma_S := \Phi_S$
- **Spec. Instance** $\Phi_S := \langle p, ctr, \sigma, n \rangle$

Judgement $\Sigma_S \vdash_{\mathcal{L}_S} \Sigma_S'$ describes how $\Sigma_S$ steps to $\Sigma_S'$
emitting observation $r$. As in $\mathcal{L}_B$, reductions only happen on top of
the stack.

$$(\text{AM-Store})\hspace{1cm} p(\sigma(\text{pc})) = \text{store } x, e \hspace{1cm} \langle p, \sigma \rangle \rightarrow \langle p, \sigma' \rangle\hspace{1cm} j = \min(\omega, n)\hspace{1cm} \sigma'' = \sigma[pc \mapsto \sigma(\text{pc}) + 1]$$

$$(\text{AM-NoSpec})\hspace{1cm} p(\sigma(\text{pc})) \not\in \text{store} \cup \Sigma_S \hspace{1cm} \langle p, \sigma \rangle \rightarrow \langle p, \sigma' \rangle\hspace{1cm} (p, ctr, \sigma, n + 1) \vdash_{\mathcal{L}_B} (p, ctr, \sigma', n)$$

To model the effect of bypassing a store instruction, Rule $S$:AM-
Store bypasses the store instruction by increasing the program
counter without updating the memory and starts a new speculative
transaction by pushing a new speculative instance on top of the
state. A load instruction loading from the same memory location
as the bypassed store instruction, therefore, retrieves a stale value.

Similarly to $\mathcal{L}_B$, all instructions that are not store instructions
(and are not in $\Sigma_S$) are handled by delegating back to the non-
speculative semantics (Rule $S$:AM-NoSpec) and when the specu-
ation window reaches 0, a roll back occurs that pops the topmost
speculative instance from the stack.

The set $\text{Beh}_{\mathcal{L}_S}^S(p)$ contains all traces generated from an initial
state until termination using the reflexive-transitive closure of $\mathcal{L}_S$.

3.1.3 Oracle Semantics. Instead of bypassing every store instruc-
tion, the oracle semantics employs an oracle $O$ that decides if the
store instruction should be speculatively bypassed or not. As before,
the behaviour $\text{Beh}_{\mathcal{L}_B}^S(p)$ of a program $p$ is the set of all traces starting
from an initial state until termination using the reflexive-transitive
closure of the oracle semantics.

3.1.4 Symbolic Semantics. Similarly to $\mathcal{L}_B$, the symbolic specu-
tative semantics $\mathcal{L}_S^S$ requires two changes w.r.t. the always-mispred-
icate one: concrete configurations $\sigma$ and the non-speculative seman-
tics are replaced by symbolic configurations $\sigma^S$ and the symbolic
non-speculative semantics respectively. The behaviour $\text{Beh}_{\mathcal{L}_S}^S(p)$ of
a program $p$ is the set of all its symbolic traces.

3.2 $\mathcal{L}_R$: Speculation on Return Instructions

The return-stack-buffer (RSB) is a small stack used by the CPU to
save return addresses upon call instructions. These saved return
addresses are speculatively used when the function returns, be-
cause accessing the RSB is faster than looking up the return address
on the stack (stored in main memory). This works well because
return addresses rarely change during function execution. However,
mispredictions can be exploited by an attacker.

**Example 2** (Return Speculation Vulnerability). Consider the ex-
ample in Listing 3 and recall that register $sp$ is used to find return
addresses saved on the stack.

**Listing 3: A program exploiting RSB speculation.**

```
1 Manip_Stack:
2 \text{sp} \leftarrow \text{sp} + 8
3 \text{ret}
4 \text{Speculate:}
5 \text{call Manip_Stack}
6 \text{load eax, secret}
7 \text{load edx, eax}
8 \text{ret}
9 \text{Main:}
10 \text{call Speculate}
11 \text{skip}
```

Each function call pushes a return address on the stack and decre-
ments the $sp$ register. After reaching the function $\text{Manip_Stack}$, the
$sp$ register is incremented (line 2). Thus, $sp$ points to the previous
return address on the stack (i.e., line 11), and the non-speculative
execution continues in $\text{Main}$ and terminates. However, the return
address of the call in line 5 is line 6 and it is on top of the RSB. Thus,
the CPU speculatively executes lines 6–7 and leaks the secret.

This section describes the AM semantics (Section 3.2.1), the
oracle semantics (Section 3.2.2), and the symbolic semantics (Sec-
tion 3.2.3). Then, it discusses formalising different implementations
of the RSB in the CPU (Section 3.2.4). This semantics is a secure
speculative semantics (Theorem 3).

**Theorem 3** ($\mathcal{L}_R$ is SSS). $\vdash \mathcal{L}_R$ SSS
3.2.1 Speculative Semantics. Unlike before, the state of \( \mathcal{L}_R \) contains a model of the RSB which is used to retrieve return addresses instead of relying on the stack. Thus, speculative instances of \( \mathcal{L}_R \) are extended with an additional entry \( R \) for tracking the RSB, whose size is limited by a global constant \( R_{size} \) denoting the maximal RSB size. A speculative instance \( \Phi_R \) now consists of the program \( p \), the counter \( ctr \), the configuration \( \sigma \), the speculation window \( \omega \) and the RSB \( R \). As before, a state \( \Sigma_R \) is a stack of speculative instances \( \Phi_R \).

**Spec. States** \( \Sigma_R := \Phi_R \)  
**Spec. Instance** \( \Phi_R := (p, ctr, \sigma, R, n) \)

As before, in \( \Sigma_R \overset{\tau}{\not\rightarrow} \Sigma_R \) reductions happen on the top of the stack.

\[
\begin{align*}
\sigma & = \langle m, a \rangle \\
\sigma' & = \sigma[pc \mapsto \alpha(\text{sp}) + 1] \\
\sigma'' & = \sigma[pc \mapsto \alpha(\text{sp}) + 8] \\
\tau & = \tau \cdot \text{start}_R \cdot \text{ret} \cdot l \\
\langle p, c \rangle & \in \mathbb{R} \\
\langle p, c, \sigma, R, n + 1 \rangle & \not\rightarrow \langle p, c, \sigma', R', n \rangle \cdot \langle p, c + 1, \sigma'', R', j \rangle \\
\end{align*}
\]

(RAM-Ret-Spec)

\[
\begin{align*}
p(\sigma(pc)) & = \text{ret} \\
\tau(p, \sigma, R, n + 1) & \not\rightarrow \langle p, c, \sigma, R, n \rangle \cdot \langle p, c + 1, \alpha(pc + 1) \rangle \\
\end{align*}
\]

(RAM-Ret-Empty)

During call instructions (Rule RAM-Call), the return address is pushed on top of the RSB (if there is space available) and during ret instructions, the return address stored on the RSB is used if the entry on top of the RSB is different from the one stored on the stack (Rule RAM-Ret-Spec). Then, the rule creates a new speculative instance that uses the return address from the RSB. Note that speculation only happens when the return address from the RSB differs from the one on the stack (stored in \( m(\alpha(\text{sp})) \)).

Here, we overview how our semantics behaves with empty and full RSB; full formalisation is available in the technical report [18]. Whenever the RSB is empty, executing a ret instruction does not cause speculation and we return to the address pointed by sp. In contrast, whenever the RSB is full, executing a call instruction does not add entries to the RSB, i.e., we model an acyclic RSB.

The behaviour \( \text{Beh}_R(\sigma) \) is the set of all traces generated from an initial state until termination using \( \not\rightarrow_R \).

3.2.2 Oracle Semantics. Unlike before, the oracle cannot decide the outcome of the ret instruction, because the CPU always uses the return address stored in the RSB (if there is one) and it does not speculate otherwise [9]. The only thing the oracle decides here is the size of the speculation window \( \omega \).

3.2.3 Symbolic Semantics. Just as before, the symbolic speculative semantics \( \mathcal{S}_R \) replaces concrete configurations and the non-symbolic semantics with the symbolic non-speculative semantics respectively. We remark that the program counter \( pc \) is always concrete in the symbolic non-speculative semantics [21]. As a result, the RSB only contains concrete values (and return addresses). The behaviour \( \text{Beh}_R(\sigma) \) of a program \( p \) is the set of all traces starting from an initial state until termination using the reflexive-transitive closure of the symbolic semantics.

1We follow the way AMD processors handle this kind of speculation [27].

3.2.4 Different Behaviours of Empty and Full RSBs. Modern CPUs use different RSB implementations that differ in how they handle underflows and overflows, i.e., when the RSB is empty or full [27]. For example, cyclic RSB implementations overwrite old entries when the RSB is full. Alternatively, CPUs can fallback to other predictors (like the indirect branch predictor) to predict return addresses whenever the RSB is empty.

In our model, the RSB is not cyclic and there is no speculation when the RSB is empty (Rule R:AM-Ret-Empty).

\[
\begin{align*}
p(\sigma(pc)) & = \text{ret} \\
\tau(p, \sigma, R, n + 1) & \not\rightarrow \langle p, c, \sigma', R', 0, n \rangle \\
\end{align*}
\]

We remark that extending \( \not\rightarrow_R \) to support different RSB implementations can be done with minimal effort.

4 A FRAMEWORK FOR COMPOSING SPECULATIVE SEMANTICS

The presented speculative semantics allow us to verify programs for violations of SNI but they do not capture the vulnerability in Listing 1, as the traces of Example 3 show.

**Example 3 (SNI for Listing 1, \( \phi \)).** The traces generated are:

\[
\begin{align*}
\tau^1_1 & = \tau^2_1 := \text{store } p \cdot \text{store } p \cdot \text{start}_B \cdot \text{load } p \cdot \text{load } A + \text{public} \cdot \text{rlb}_B \cdot \text{pc } 9 \\
\tau^1_2 & = \tau^2_2 := \ldots \cdot \text{store } p \cdot \text{start}_S \cdot \text{bypass } 1 \cdot \text{pc } \perp \cdot \text{rlb}_S \cdot \text{pc } \perp \\
\end{align*}
\]

The program in Listing 1 seems secure since there is no secret value leaked in the speculative transaction; thus the program satisfies SNI for \( \not\rightarrow_R \) and \( \not\rightarrow_s \) in isolation. However, this program satisfies speculatively leaks when considering speculation over beqz and store instructions, but we need our combined semantics to detect this vulnerability; see Section 5.3.

The vulnerability only appears when the branch predictor (Section 2.4.1) and the memory disambiguator (Section 3.1.2) are used together. Intuitively, we know that CPUs use all the speculation mechanisms described here (and many others as well) at the same time. Thus, we should not only focus on these different speculation mechanisms in isolation but we need to look at their combinations as well. That is, we need a way to compose the different semantics into new semantics that can reason about these “combined” leaks.

This section presents a novel, general framework for composing two speculative semantics \( x \) and \( y \), each one capturing the effects of a single speculation mechanism, to allow for speculation from both mechanisms \( x \) and \( y \). The semantics \( x \) and \( y \) are also called the source semantics of the composition. Next, we first introduce the new composed semantics, which consists of an always-mispredict semantics, an oracle semantics, and a symbolic semantics (Section 4.1). Then, we present the notion of well-formed composition which we use to study the properties of composed semantics (Section 4.2).

**New Notation.** The states \( \Sigma_{xy} \), instances \( \Phi_{xy} \), and the trace model \( OBS_{xy} \) are defined as the union of the source parts. Furthermore, we define a projection function \( \tau_{xy} \) and two projections \( \tau_{xxy} \) and \( \tau_{xyy} \).
\[ \begin{align*}
\delta_{xy}^{\cup} & \text{ that return the first and second projection of the pair from } \delta_{xy}. \\
& \text{These functions are lifted to states by applying them pointwise:}
\end{align*} \]

\[ \begin{align*}
\delta_{xy}^{\cup} : \Phi_{xy} \mapsto (\delta_{x} \cup \delta_{y}) \\
\delta_{xy}^{\cap} : \Phi_{xy} \mapsto (\delta_{x} \cap \delta_{y}) \\
\delta_{xy}^{\setminus} : \Phi_{xy} \mapsto (\delta_{x} \setminus \delta_{y}) \\
\delta_{xy}^{\setminus} : \Phi_{xy} \mapsto (\delta_{y} \setminus \delta_{x}) \\
\end{align*} \]

\[ \begin{align*}
\delta_{xy}^{\cup} \Phi_{x} \cup \Phi_{y} & \quad \Sigma_{x} \cup \Sigma_{y} \\
\delta_{xy}^{\cap} \Phi_{x} \cap \Phi_{y} & \quad \Sigma_{x} \cap \Sigma_{y} \\
\delta_{xy}^{\setminus} \Phi_{x} \setminus \Phi_{y} & \quad \Sigma_{x} \setminus \Sigma_{y} \\
\delta_{xy}^{\setminus} \Phi_{x} \setminus \Phi_{y} & \quad \Sigma_{y} \setminus \Sigma_{x} \\
\end{align*} \]

For example, the \( \Phi_{x+y} \) state resulting from the union of \( \Phi_{x} \) and \( \Phi_{y} \) states (from Section 3.1.2 and Section 3.2.1 respectively) is \( \langle p, c, \sigma, R, n \rangle \), as it contains all common elements (the program \( p \), the counter \( c \), the state \( \sigma \), and the speculation count \( n \)) plus the return stack buffer \( R \) from \( \Phi_{R} \) only. Taking the \( \delta_{x+y}^{\cup} \) of a \( \Phi_{x+y} \) state returns the \( \Phi_{x+y} \) subpart, i.e., all but the return stack buffer.

We overload \( \delta_{x}^{\cup} \) and \( \delta_{x}^{\cap} \) to also work on traces \( \tau \). The projection \( \delta_{x}^{\cup} \) deletes all speculative transactions (marked by \( \text{start}_y \) \text{id} and \( \text{rlb}_y \) \text{id}) not generated by the source semantics \( x \). The definition of \( \delta_{x}^{\cap} \) is similar by replacing \( x \) with \( y \):

\[ \begin{align*}
\delta_{x}^{\cup} \tau & = \varepsilon \\
\delta_{x}^{\cap} \tau & = (\tau \cdot (\varepsilon)) \delta_{x}^{\cup} \\
\delta_{x}^{\cup} \langle p, \sigma, \tau, (\text{start}_y \text{id}, \ldots, \text{rlb}_y \text{id}) \rangle & = \tau \delta_{x}^{\cap} \tau
\end{align*} \]

We indicate source semantics for \( x \) and \( y \) as \( \delta_{x} \) and \( \delta_{y} \) respectively and use \( \delta_{xy} \) to indicate the composed semantics.

### 4.1 Combined Speculative Semantics

The combined semantics delegates back to the source semantics of \( x \) and \( y \) to model the effects of both speculation mechanisms (modeled by \( x \) and \( y \)). This is captured in the two core rules below:

\[ \begin{align*}
\Phi_{xy} \delta_{xy}^{\cup} \tau & \quad \delta_{xy}^{\cap} \delta_{xy}^{\cup} \\
\Phi_{xy} \delta_{xy}^{\cap} \delta_{xy}^{\cup} & \quad (\text{AM-x-step}) \\
\Phi_{xy} \delta_{xy}^{\cap} \delta_{xy}^{\cup} & \quad (\text{AM-y-step}) \\
\end{align*} \]

The combined semantics does a step by either delegating back to the \( x \) source semantics (Rule AM-x-step) or to the \( y \) one (Rule AM-y-step). The rules rely on metaparameter \( Z_{xy} \), which is a pair of two metaparameters \( Z_{x} := (Z_{x}, Z_{y}) \) — one for \( x \) and one for \( y \). We overload the projections \( \delta_{xy}^{\cup} \) and \( \delta_{xy}^{\cap} \) to extract the corresponding metaparameter from \( Z_{xy} \), e.g., \( \delta_{x+y}^{\cup} = Z_{x} \).

The role of \( Z \) is to centralize making the composed semantics work as expected. It restricts how the combined semantics delegates execution to the components to ensure that the correct rule is applied.

With \( Z = (\emptyset, \emptyset) \), consider the execution of the \( \text{beqz} \) instruction in Line 4 in Listing 1. The combined semantics \( \delta_{xy}^{\cup} \) can use Rule AM-x-step to delegate back to \( \delta_{x} \) for \( \text{beqz} \) instructions, creating a new speculative transaction (Rule B:AM-branch). However, \( \delta_{xy}^{\cup} \) can also use Rule AM-y-step, because \( \text{beqz} \) instructions are also handled by \( \delta_{y} \). Unfortunately, this does not start speculation, which happens only on \( \text{store} \) instructions (Rule S:AM-NoSpec).

Intuitively, \( \delta_{xy}^{\cup} \) should delegate back to \( \delta_{y} \), so Rule AM-y-step should not be applicable. This can be obtained by instantiating \( \delta_{xy}^{\cup} = (\text{store}, \text{beqz}) \), so that its projections are \( \delta_{y} = \text{store} \) and \( \delta_{z} = \text{beqz} \). Now, \( \delta_{xy}^{\cup} \) can only apply Rule AM-x-step on the \( \text{beqz} \) of Line 4, because \( \delta_{z} \) ensures that \( \delta_{y} \) cannot execute \( \text{beqz} \) instructions, as depicted in the full rule for \( \delta_{xy}^{\uparrow} \) below (where we indicate the instructions derived from \( Z_{y} = \text{beqz} \) in blue):

\[ \begin{align*}
\delta_{xy}^{\uparrow} & \quad (\text{AM-NoSpec}) \\
\delta_{xy}^{\uparrow} & \quad (p, \sigma) \rightarrow (p, \sigma') \quad \text{(with } \text{beqz} \text{ in blue)}
\end{align*} \]

Having clarified the intuition behind the semantics, we can define the behaviour \( \delta_{xy}^A \) as the set of all traces generated from initial states until termination using \( \delta_{xy}^A \).

#### 4.1.1 Oracle Combination

Oracle Combination. Instead of using one oracle, the combination uses a pair of oracles, one from each source semantics. When delegating back to either source, the correct oracle of the source is handed over to the source semantics.

#### 4.1.2 Symbolic Combination

Instead of using the AM semantics for delegation, the combined symbolic semantics \( \delta_{xy}^S \) uses the symbolic source semantics for delegation. Furthermore, the new notation (union, projections) is lifted to the symbolic combination to create the symbolic states \( \Sigma_{xy}^S \). The behaviour \( \delta_{xy}^S(p) \) of program \( p \) is the set of all traces generated using the symbolic semantics.

### 4.2 Properties of Composition

We now illustrate the benefits of our composition framework. For this, we first introduce a notion of well-formed composition (Section 4.2.1), which intuitively tells when a combined semantics “makes sense”. Then, we show that for well-formed compositions, if the source semantics are SSS, so is the combined semantics (Section 4.2.2). Since we proved this property for any well-formed composition in our framework, all (well-formed) compositions we present in Section 5 are SSS for free. This proof reuse and extensibility is our framework’s key advantage over having ad-hoc semantics combining multiple speculation mechanisms, which requires one to manually prove the SSS results we instead obtain for free.

#### 4.2.1 Well-formed Compositions

The well-formedness conditions for the composition ensures that the delegation is done properly (Definition 3), they are the minimal set of assumptions that let us derive SSS of the combined semantics for free:

**Definition 3 (Well-formed composition).** A composition \( \delta_{xy} \) of two speculative semantics \( \delta_{x} \) and \( \delta_{y} \) is well-formed, written \( \delta_{xy} : \text{WFC} \), if:

1. \( (\text{Confluence}) \) \text{ Whenever } \Sigma_{xy} \delta_{xy}^A \Sigma_{xy}^A \text{ and } \Sigma_{xy} \delta_{xy}^B \Sigma_{xy}^B, \text{ then } \Sigma_{xy} = \Sigma_{xy}^A \text{.} \]

2. \( (\text{Projection preservation}) \) For all \( p \), \( \delta_{xy}^A(p) = \delta_{xy}^B(p) \) and \( \delta_{xy}^A(p) = \delta_{xy}^B(p) \text{.} \]

3. \( (\text{Relation preservation}) \) If \( \Sigma_{xy} \approx_{xy} X_{xy} \text{ and } \Sigma_{xy} \delta_{xy}^A X_{xy}^A \text{, then } X_{xy} \approx_{xy} X_{xy}^A \text{.} \]

4. \( (\text{Symbolic preservation}) \) If \( \Sigma_{xy} \delta_{xy}^A S_{xy} \Sigma_{xy}^A \text{ and } \delta_{xy}^A(p) = \delta_{xy}^A(p) \text{, then } S_{xy} \approx_{xy} X_{xy} \text{, s.t. } \Sigma_{xy} \delta_{xy}^A X_{xy}^A \text{ and } \mu(S_{xy}^A) = \mu(S_{xy}) \text{.} \]

Next, we explain the well-formedness conditions:

- **Confluence (point 1)** ensures that the non-determinism of the combined semantics (non-deterministically delegates back to
its sources) is not harmful. Consider the assignment in Line 1 in Listing 1. \(\mathcal{Q}_{B+S}\) can delegate to either \(\mathcal{Q}_B\) or \(\mathcal{Q}_S\) to reduce the assignment. If the combined semantics is confluent, then it does not matter which source rule executes the assignment in Line 1 in Listing 1, the semantics reaches the same state either way.

- Projection preservation (point 2) ensures that the combined semantics is not hiding or forgetting traces of its sources. Any observable emitted by a source semantics must be propagated to the combined one, this is also the reason why \(\text{Obs}_{xy}\) is defined as the union of the source \(\text{Obs}\).

- To explain relation preservation (point 3), we need to mention a technical detail: the state relation (denoted \(\equiv_{xy}\) and defined in our technical report) between the AM states \((\Sigma_{xy})\) and the Oracle ones \((X_{xy})\). Intuitively, two states are related if they are the same or if one is waiting on a speculation of the other to end. Then, point (3) ensures that whenever we start from related states \((\Sigma_{xy} \equiv_{xy} X_{xy})\) and we do one or more steps of the AM composed semantics \((\Sigma_{xy} ? \bigcup_{xy} ? \Sigma_{xy})\), then we can always find a related state \((\Sigma'_{xy} \equiv_{xy} X'_{xy})\) that is reachable by performing one or more steps of the composed oracle semantics \((X_{xy} ? \bigcup_{xy} ? X_{xy})\). This fact is used when proving that SNI of a program under the composed AM semantics implies SNI under the composed oracle semantics (point 1 of Definition 2). Thus, it is not important for the AM and the Oracle semantics to produce the same traces, just that the two AM traces and the two Oracle traces are pairwise equivalent – which follows from the state relation.

- Finally, symbolic preservation (point 4) ensures that any step of the always-mispredict composed semantics corresponds to the concretization of a step of the symbolic composed semantics (and vice versa\(^7\)). Note that proving symbolic preservation is almost trivial whenever both source semantics enjoy the same property (like our semantics \(\mathcal{Q}_B, \mathcal{Q}_S, \) and \(\mathcal{Q}_R\)).

4.2.2 SSS preservation. The key result of our framework is that well-formed compositions whose sources are secure speculative semantics (SSS) are also SSS (Theorem 4). Note that our proof of Theorem 4, available in the companion technical report [18], holds for any well-formed composition in our framework and, therefore, it applies for free to all the compositions in Section 5.

**Theorem 4 (\(\mathcal{Q}_{xy}\) is SSS).** If\(\mathcal{Q}_x\) SSS and\(\mathcal{Q}_y\) SSS and\(\mathcal{Q}_{xy} : WFC\) then\(\mathcal{Q}_{xy}\) SSS.

As a corollary of Theorem 4, we obtain that the security of well-formed compositions is related to the security of their components (Theorem 5). In particular, whenever a program is insecure w.r.t. one of the components, then it is insecure w.r.t. the composed semantics. Dually, if a program is secure w.r.t. the composed semantics, then it is secure w.r.t. the single components. Note, however, that there are programs that are secure for the single components but insecure w.r.t. the composed semantics like Listing 1.

**Theorem 5 (Combined SNI Preservation).** If\(\mathcal{Q}_{xy} : WFC\) and \(p \not\equiv_{xy} \text{SNI}\) or \(p \not\equiv_{xy} \text{SNI}\), then \(p \not\equiv_{xy} \text{SNI}\).

If\(\mathcal{Q}_{xy} : WFC\) and \(p \equiv_{xy} \text{SNI}\), then \(p \equiv_{x} \text{SNI}\) and \(p \equiv_{y} \text{SNI}\).

These results have an immediate practical impact on Spectector: (1) Spectector’s security analysis relies on the (symbolic) speculative semantics, (2) the source semantics \(\mathcal{Q}_B, \mathcal{Q}_S, \) and \(\mathcal{Q}_R\) are SSS, (3) well-formed compositions are also SSS, and (4) the composition of \(\mathcal{Q}_B, \mathcal{Q}_S, \) and \(\mathcal{Q}_R\) are well-formed. So, the Spectector security analysis equipped with any combination of the \(\mathcal{Q}_B, \mathcal{Q}_S, \) and \(\mathcal{Q}_R\) produces sound results, i.e., whenever the tool proves that a program is leak-free then the program satisfies SNI. So, the next section describes all the compositions and proves they are well-formed (this implies that they are SSS thanks to Theorem 4), whereas the section thereafter describes their implementation in Spectector.

5 INSTANTIATING OUR FRAMEWORK

This section describes all combinations of the speculative semantics \(\mathcal{Q}_B, \mathcal{Q}_S, \) and \(\mathcal{Q}_R\) (Section 5.1), \(\mathcal{Q}_{B+S}\) (Section 5.2), \(\mathcal{Q}_{B+R}\) (Section 5.3), and \(\mathcal{Q}_{B+S+R}\) (Section 5.4). For each one, we overview the combined AM semantics using examples (whereas we computed using our Coq executable compiled semantics) and we prove that the combined semantics is well-formed, i.e., it satisfies Definition 3. In the following, we describe in detail how the \(\mathcal{Q}_{B+S+R}\) semantics can be instantiated as part of our framework; the other combinations can be instantiated similarly and we only provide a higher-level description. Full details and well-formedness proofs are available in the companion technical report [18].

5.1 \(\mathcal{Q}_{S+R}\) Composition

To combine semantics using our framework, we need to define the states, observations, and metaparameter \(\mathcal{Z}_{S+R}\) for the composed semantics \(\mathcal{Q}_{S+R}\). The combined state \(\Sigma_{S+R}\) is the union of the states \(\Sigma_S\) and \(\Sigma_R\); thus it contains the RSB \(\mathcal{R}\) as well.

**Spec. States** \(\Sigma_{S+R} := \Sigma_S + \Sigma_R\)  **Spec. Instance** \(\Phi_{S+R} := \langle p, ctr, \sigma, r, n \rangle\)

The union \(\text{Obs}_{S+R}\) of the trace models \(\text{Obs}_S\) and \(\text{Obs}_R\) is defined as:

\(\text{Obs}_{S+R} := \text{start}_s n | \text{start}_r n | \text{rlb}_s n | \text{rlb}_r n | \text{bypass} n | ...\)

To define the metaparameter \(\mathcal{Z}_{S+R}\), we need to identify, for each component semantics, the instructions that are related with speculative execution. For \(\mathcal{Q}_S\), the only instruction associated with speculative execution is \text{store}, since the semantics can only speculatively bypass stores. For \(\mathcal{Q}_R\), even though the semantics speculates only over \text{ret} instructions, \text{call} instructions also affect speculative execution since \(\mathcal{Q}_R\) pushes return addresses onto the \(\mathcal{R}\) when executing \text{calls}. Therefore, we set the metaparameter \(\mathcal{Z}_{S+R}\) to \(\langle \text{call} \cup \text{ret}, \text{store} \rangle\).

This ensures that in \(\mathcal{Q}_{S+R}\), \text{store} instructions are only executed by delegating back to \(\mathcal{Q}_{S+R}\) whereas \text{call} and \text{ret} instructions are only executed by delegating back to \(\mathcal{Q}_S\) and \(\mathcal{Q}_R\) and \(\mathcal{Q}_{S+R}\).

**Theorem 6 (\(\mathcal{Q}_{S+R}\) is well-formed).** \(\vdash \mathcal{Q}_{S+R} : WFC\)

Listing 4 presents a program that contains a leak that can be detected only by \(\mathcal{Q}_{S+R}\) but not by its components \(\mathcal{Q}_S\) and \(\mathcal{Q}_R\).

**Listing 4: \(\mathcal{Q}_{S+R}\) example**

```plaintext
Manip stack:
sp <= sp + 8
```
In Listing 4, execution starts on Line 12 by calling the function `Speculate` and it continues at Line 5. Next, the function `Manip_Stack` is called and the stack pointer `sp` is incremented (Line 2). This modifies the return address of the function `Manip_Stack` to now point to Line 13 (the return address of the call to `Speculate`). Under $\mathcal{D}_{R}$, mispredicting the return address of `Manip_Stack` using the RSB leads to continuing the execution at Line 6. However, the `store` instructions in Line 7 overwrites the secret value stored in Line 6. Then, the `load` instruction is executed and the secret is leaked, as shown in the traces below. Since $\sigma_1$ is a high value, there are low-equivalent configurations $\sigma_1^1, \sigma_1^2$ that differ in the value of secret. Thus, there are two traces $\mathcal{T}_1$ that differ in the observation `load secret` (highlighted in gray). Hence, the program is not secure under the combined semantics, i.e., Listing 4 $\not\models_{R+S}$ SNI.

$$\mathcal{T}_1^2 \neq \mathcal{T}_1^1 \Downarrow \text{call Speculate} \cdots \text{start}_B \cdots \text{start}_R \cdots \text{bypass} \cdot \text{load p} \cdot \text{load secret} \cdots$$

The relation between the source semantics and their composition is visualised in Figure 1, which shows the insecure programs (with respect to SNI) detected under the different semantics. The combined semantics encompasses all vulnerable programs of $\mathcal{D}_S$ and $\mathcal{D}_R$ and additional programs like Listing 4. These additional programs are the reason why the semantics $\mathcal{D}_{S+R}$ is "stronger than the sum of its parts" $\mathcal{D}_S$ and $\mathcal{D}_R$.

**5.2 $\mathcal{D}_{B+R}$ Composition**

In this combination, the instructions that influence speculative execution are `call` and `ret` ($\mathcal{D}_{B}$) and `beqz` ($\mathcal{D}_{R}$). Thus, we set $\mathcal{D}_{B+R} = (\text{call} \cup \text{ret}, \text{beqz})$ to account for this and to allow speculation from both sources.

Theorem 7 states that $\mathcal{D}_{B+R}$ is well-formed. This allows us to derive "for free" that $\mathcal{D}_{B+R}$ is SSS by applying Theorem 4.

**Theorem 7 ($\mathcal{D}_{B+R}$ is well-formed).** $\vdash \mathcal{D}_{B+R} : \text{WFC}$

**Listing 5: $\mathcal{D}_{B+R}$ example**

```
Manip_Stack:
  sp <- sp +
ret
Speculate:
  call Manip_Stack
  x <- 0
  beqz x, L2
  load eax, secret
L2:
  ret
Main:
  call Speculate
skip
```

Listing 5 presents a leak that can be detected only under $\mathcal{D}_{B+R}$. The execution proceeds similarly to Listing 4 until the `ret` instruction in Line 3 is reached. Under $\mathcal{D}_{B}$, mispredicting the return address leads to function `Manip_Stack` returning to Line 6. However, the `beqz` instructions in Line 7 jumps to Line 10 (since $x$ is 0) and speculation ends without leaking. Under $\mathcal{D}_{R}$, the `beqz` instruction in Line 7 is never executed and the function `Manip_Stack` returns to Line 13 without leaking. Hence, Listing 5 is secure (i.e., it satisfies SNI) when considering $\mathcal{D}_{B}$ and $\mathcal{D}_{R}$ in isolation.

Under the combined semantics $\mathcal{D}_{B+R}$, function `Manip_Stack` returns to Line 6 and the `beqz` instruction is then mispredicted. This leads to executing the load instructions in Line 8, which leaks secret information. The resulting traces $\mathcal{T}_1$ are given below, where we highlight the secret-dependent observations. Given the length of the trace, we present only the most relevant parts, i.e., that both kinds of speculations need to have started for the leak to appear.

$$\tau_1^2 \neq \tau_1^1 \Downarrow \text{call Speculate} \cdots \text{start}_R \cdots \text{start}_B \cdots \text{load secret} \cdots$$

Again, the two traces differ in the observation in the grey box and we have Listing 5 $\not\models_{B+R}$ SNI.

**5.3 $\mathcal{D}_{B+S}$ Composition**

In this combination, speculation happens on `beqz` instructions ($\mathcal{D}_{S}$) and on `store` instructions ($\mathcal{D}_{B}$). Thus, we set $\mathcal{D}_{B+S} = (\text{store, beqz})$. Therefore, in the combined semantics $\mathcal{D}_{B+S}$, `beqz` instructions are only executed by delegating back to $\mathcal{D}_{B}$ and `store` instructions are only executed by delegating back to $\mathcal{D}_{S}$. This semantics is also a well-formed composition (Theorem 8) and SSS.

**Theorem 8 ($\mathcal{D}_{B+S}$ is well-formed).** $\vdash \mathcal{D}_{B+S} : \text{WFC}$
Listing 1 from Section 1 contains a leak that can only be detected by \( \mathcal{B}_{\text{RSB}} \) but not by its components. The traces associated with the code (\( \triangleright \)) are given below, where secret-dependent observations are highlighted in gray:

\[
\tau_{\text{RSB}} \neq \tau_{\text{RSB}} \quad \text{def} = \cdots \text{start}_1 \cdot \text{bypass} 1 \cdots \cdot \text{start}_B 2 \cdot \text{pc} 5 \\
\cdot \text{load } p \cdot \text{load } A + \text{secret} \cdot \text{rlb}_B 2 \cdot \text{rlb}_S 1 \cdots
\]

The program is not secure under \( \mathcal{B}_{\text{RSB}} \), i.e., Listing 1 \( \not \vdash_{\text{RSB}} \) SNI.

5.4 \( \mathcal{B}_{\text{RSB}} \) Composition

We conclude this section by combining all three semantics \( \mathcal{B}_{\text{RSB}} \), \( \mathcal{B}_S \), and \( \mathcal{B}_R \). Our framework (Section 4) allows only combining a pair of source semantics into a combined one. For simplicity, we present \( \mathcal{B}_{\text{RSB}} \) as a direct combination of the three source semantics (technically, we obtain \( \mathcal{B}_{\text{RSB}} \) by combining \( \mathcal{B}_{\text{RSB}} \) with \( \mathcal{B}_R \)). The metaparameter \( Z_{\text{RSB}} \) (which we represent as a triple of values) is \( \text{call} \cup \text{ret} \cup \text{store} \), \( \text{call} \cup \text{ret} \cup \text{beqz} \), and \( \text{beqz} \cup \text{store} \). As a result, the combined semantics \( \mathcal{B}_{\text{RSB}} \) can only delegate to the corresponding speculative semantics for the appropriate speculation sources.

As before, \( \mathcal{B}_{\text{RSB}} \) is a well-formed composition (Theorem 9) and we get that \( \mathcal{B}_{\text{RSB}} \) is SSS by applying Theorem 4.

**Theorem 9 (\( \mathcal{B}_{\text{RSB}} \) is well-formed).** \( \vdash \mathcal{B}_{\text{RSB}} : \text{WFC} \)

Listing 6: \( \mathcal{B}_{\text{RSB}} \) example

```plaintext
-Manip_Stack:
sp <- sp + 8
ret
Speculate:
call Manip_Stack
x <- 0
beqz x, L2
load eax, p
load edi, eax
L2:
ret
Main:
store secret, p
store pub, p
call Speculate
```

Listing 6 depicts a leaky program that can be detected only under \( \mathcal{B}_{\text{RSB}} \), since the program satisfies SNI under \( \mathcal{B}_S \), \( \mathcal{B}_R \), and \( \mathcal{B}_{\text{RSB}} \). Under \( \mathcal{B}_{\text{RSB}} \), the store instruction in Line 14 is bypassed. Therefore, when returning from Manip_Stack, the program mispredicts the return address and speculatively returns to Line 6. Here, the beqz instruction in Line 7 is mispredicted and the load instructions are executed, which now leaks the secret value.

The resulting traces (\( \triangleright \)) are given below:

\[
\tau_{\text{RSB}} \neq \tau_{\text{RSB}} \quad \text{def} = \cdots \text{start}_1 \cdot \text{bypass} 1 \cdot \text{call Speculate} \\
\cdot \text{start}_B 2 \cdot \text{ret} 6 \cdot \text{start}_B 3 \cdot \text{pc} 8 \cdot \text{load } p \\
\cdot \text{load } \text{secret} \cdot \text{rlb}_B 3 \cdot \text{rlb}_S 2 \cdot \text{rlb}_S 1 \cdots
\]

Thus, the program is not secure, i.e., Listing 6 \( \not \vdash_{\text{RSB}} \) SNI.

6 IMPLEMENTATION AND EVALUATION

This section describes how our combined semantics can be used to detect leaks introduced by the interaction of multiple speculation mechanisms. For this, we extended Spectector, a symbolic analysis tool for speculative leaks against \( \mathcal{B}_{\text{RSB}} \), with the semantics for \( \mathcal{B}_S \) and \( \mathcal{B}_R \) and for all the combinations from Section 5 (Section 6.1). Using Spectector, we analyze a corpus of 49 microbenchmarks containing speculative leaks generated by different speculation mechanisms (Section 6.2). With these experiments, we aim to show that (1) our \( \mathcal{B}_S \) and \( \mathcal{B}_R \) speculative semantics can correctly identify speculative leaks associated with speculation over store-bypasses and return instructions, and (2) our combined semantics can detect novel leaks that are otherwise undetectable when considering single speculation mechanisms in isolation.

6.1 Implementation

We implemented all our semantics (the symbolic versions of \( \mathcal{B}_S \) and \( \mathcal{B}_R \) plus all compositions from Section 5) as an extension of Spectector [21]. The implementation of compositions closely follows the structure of our framework. As in Section 5, selecting one of the composed semantics in Spectector sets the metaparameter \( Z \), which is used to delegate back to the correct individual semantics. Spectector then uses symbolic execution together with self-composition [6] and an SMT solver to check for SNI against \( \mathcal{B}_S \). Due to this setup, we inherit all limitations of Spectector’s speculative analysis, e.g., path explosion due to symbolic execution and limitations in the translation from x86 to \( \mu \text{Asm} \). We refer to [21] for an in-depth discussion of such limitations.

6.2 Experiments

**Benchmarks:** We analyze 49 snippets of code containing leaks resulting from speculation over branch, store/load, and ret instructions (and their combinations):

- **Spectre-STL:** 13 programs are variants of the Spectre-STL vulnerability. They exploit speculation over memory disambiguation, and they have been used as benchmarks in prior work [14, 32]. For each program, we also analyze a patched version where a manually inserted lfence instruction stops speculation over store-bypasses and prevents the leak.
- **Spectre-RSB:** 5 programs are variants of the Spectre-RSB vulnerability. They exploit speculation over return instructions, and they are obtained from the safeside [1] and transientfail [8] projects\(^3\). For each program, we also analyze manually patched versions obtained by (1) inserting lfences after call instructions (i.e., at the instruction address where ret speculatively returns), and (2) using the modified retpoline defense proposed in [27, Section 6.1].
- **Spectre-Comb:** 4 programs contain leaks that arise from combining speculation mechanisms. These are the programs depicted in listing 1, listing 5, listing 4, and listing 6 and discussed in Section 5. For each program, we also analyze a manually patched version where 1fence instructions prevent the speculative leaks.

\(^3\)Of the three Spectre-RSB examples from safeside [1], we analyze the only one that works against an acyclic RSB like the one supported by \( \mathcal{B}_S \). Programs ca.ip, ca.eop, and ca.ip from transientfail [8] rely on concurrent execution. Since Spectector does not support concurrency, we hardcode the worst-case interleaving in terms of speculative leakage in our benchmark.
We also analyzed programs manually patched with 1fence statements ("listing 1 Fence", "listing 5 Fence", "listing 4 Fence", and "listing 6 Fence" in Figure 3a). As before, Spectector successfully proved the security of patched programs. Even for leaks that arise from multiple speculation mechanisms, it is often sufficient to insert a single 1fence to secure the entire program, e.g., an 1fence after the beqz instruction in Listing 5 is enough to make the program SNI with respect to $\mathcal{D}_{R+4+S+R}$.

Figure 3b reports the average execution time (for 1000 executions) of Spectector's analysis for the Spectre-Comb programs under the different semantics. We highlight the following findings:

- For the programs patched with 1fence, Spectector's execution time under a combined semantics is larger than Spectector's execution times under the corresponding source semantics. This follows from the combined semantics exploring (a) everything explored by the source semantics as well as (b) additional statements resulting from extra interactions between the source semantics.

Note that the placement of 1fences influence execution time. For instance, the execution time for "listing 5 Fence" under $\mathcal{D}_R$ and $\mathcal{D}_R$.
(a) Results of the analysis. For each program, o denotes that SPECTECTOR finds a violation of SNI whereas • denotes that SPECTECTOR proves the program secure under the corresponding semantics. Next to each program, we report if it is Secure or Insecure w.r.t. $\Delta_{B+4+R}$.

(b) Average execution time for SPECTECTOR’s analysis for the code snippets in the Spectre-Comb benchmark (over 1000 samples) for the relevant individual and composed speculative semantics. The white bar (“non-spec”) represents the analysis time w.r.t. $\mu$Asm non-speculative semantics.

Figure 3: Results of the Spectre-Comb benchmark, where “listing x Fence” is the patched version (using `lfence`) of “listing x”.

$\Delta_{B+4+R}$ is similar because the `lfence` is placed just after the branch instruction of Line 7, thereby stopping $\Delta_{B}$ speculation.

- For most of the unpatched programs, execution time under a combined semantics is again larger than the execution times under the source semantics. This is, however, not always the case. For instance, SPECTECTOR’s execution time for listing 6 and $\Delta_{B+4+R}$ is larger than its execution time for $\Delta_{B+4+R}$. This is due to SPECTECTOR’s terminating early after finding a violation of SNI, which happens under $\Delta_{B+4+R}$ but not under $\Delta_{B+4+R}$ (see also Figure 3a).

7 DISCUSSION

Scope of the models: Lifting the results of the security analysis for our speculative semantics to real-world CPUs is only possible to the extent that these semantics capture the information flows in the target system. Thus, SPECTECTOR’s result may incorrectly classify programs as secure (if our semantics do not capture information flows happening in real-world CPUs) or insecure (if our semantics admit speculations that are impossible on real systems).

Other speculation mechanisms: There are many speculation mechanisms beyond those modeled in $\Delta_{B}$, $\Delta_{S}$, and $\Delta_{R}$:

- Speculation over indirect jumps [24] can be modeled as an always-mispredict semantics (similarly to $\Delta_{B}$) where mispredicted paths can start from any other statement. This, however, makes automated reasoning challenging due to the large number of speculative paths. Mechanisms like Intel Control-Flow-Integrity [34] can improve the situation by restricting potential jump targets.

- CPUs speculate over `ret` instructions in different ways. For instance, there are many different ways of implementing return stack buffers (e.g., cyclic versus acyclic RSBs [27] or RSBs that fall back to indirect branch prediction [41]). Some ARM processors, moreover, use straight-line speculation that allows CPUs to speculatively bypass a `ret` instruction and execute the instructions following it. Both kinds of speculation can be modeled by modifying the Rule R:AM-Ret-Spec rule in $\Delta_{B}$.

- Many proposals for value prediction over different kinds of instructions exist [26, 29, 35]. While naive speculative semantics might have to explore all possible values as prediction, semantics that model specific prediction mechanisms might restrict the set of predicted values (thereby leading to a more tractable analysis).

We expect that most of these mechanisms can be modeled as speculative semantics satisfying our well-formedness conditions. Hence, they could work with our composition framework.

Limitations of composition: Our composition framework has two main limitations:

1. The metaparameter $Z$ is expressed in terms of $\mu$Asm instructions, i.e., the smallest unit of computation in our framework. Since $Z$ restricts how the combined semantics delegates execution to its
sources, this limit the expressiveness of composed semantics. For instance, \(\mathcal{L}_{\text{S}+\text{R}}\) cannot speculate over the \textit{implicit store} writing the return address to the stack that happens as part of call instructions.

(2) Our framework does not support combinations where a single instruction perform speculation-relevant changes in both source semantics. For instance, consider a combination of \(\mathcal{L}_{\text{R}}\) with a semantics modeling straight-line speculation. Here, both semantics start different speculative transactions on executing ret instructions. However, instantiating \(Z\) as \((\emptyset, \emptyset)\), which enables both speculations, violates the confluence well-formedness condition for the composed semantics, whereas setting \(Z = (x, y)\) so that only one of \(x\) and \(y\) is \textit{ret} would only capture one of the two speculation mechanisms.

We leave addressing both limitations as future work.

8 RELATED WORK

Speculative execution attacks: After Spectre [24] has been disclosed to the public in 2018, researchers have identified many other speculative execution attacks [4, 7, 25, 27, 42]. These attacks differ in the exploited speculation sources [23, 25, 27], the covert channels [33, 36, 37] used, or the target platforms [12]. We refer the reader to Canella et al. [8] for a survey of existing attacks.

Security conditions for speculative leaks: Researchers have proposed many program-level properties for security against speculative leaks, which can be classified in three main groups [10]:

1. Non-interference definitions ensure the security of speculative and non-speculative instructions. For instance, speculative constant-time [9] (used also in [3, 14, 38]) extends the constant-time security condition to account also for transient instructions.

2. Relative non-interference definitions [11, 19, 21, 22] ensure that transient instructions do not leak more information than what is leaked by non-transient instructions. For instance, speculative non-interference [21], which we build on, (used also in [20, 31]) restricts the information leaked by speculatively executed instructions (without constraining what can be leaked non-speculatively).

3. Definitions that formalise security as a safety property [31, 32], which may over-approximate definitions from the groups above.

Operational semantics for speculative leaks: In the last few years, there has been a growing interest in developing formal models and principled program analyses for detecting leaks caused by speculatively executed instructions. We refer the reader to [10] for a comprehensive survey on the topic. In the following, we discuss the approaches that are more relevant to our paper.

Our speculative semantics \(\mathcal{L}_{\text{S}}\) and \(\mathcal{L}_{\text{R}}\) capture the effects of transient instructions at a rather high-level, and they are inspired by the always-mispredict \(\mathcal{L}_{\text{B}}\) semantics from [21]. Our \(\mathcal{L}_{\text{S}}\) semantics is also similar to the CT-BPAS speculation contract used by the Revizor testing tool [30]. In contrast, other approaches, which we overview next, explicitly model microarchitectural components like multiple pipeline stages, caches, and branch predictors.

For instance, KLEE-Spectre [39] and SpecuSym [22] consider a semantics that explicitly model the cache, which enable reasoning about the cache content. McIroy et al. [28] go a step further and model a multi-stage pipeline with explicit cache and branch predictor. Their semantics can only model speculation over branch instructions since it lacks store-forwarding or RSB.

Cauligi et al. [9]’s semantics model speculation over branch instructions, store-bypasses, and return instructions. Differently from our semantics, their 3-stage pipeline semantics explicitly models several microarchitectural components like a reorder buffer and an RSB. Their tool detects violations of speculative constant-time induced by speculation over branch instructions and store-bypasses.

Binsec/Haunted [14] detect violations of speculative constant-time due to speculation over store-bypasses and branch instructions. For this, they explicitly model the store buffer, which \(\mathcal{L}_{\text{S}}\) abstracts away. Barthe et al. [5] extend the Jasmin [3] cryptographic verification framework to reason about speculative constant-time and supports speculation over store-bypasses and branch instructions.

While several of these models support multiple speculation mechanisms, these mechanisms are \textit{hard-coded} and no existing approach provides a composition framework or extensible ways of extending the main theoretical results to new mechanisms “for free”. Moreover, while we could have used other semantics as a basis for our framework, this would have resulted in more difficult proofs (since semantics like the one in [9] are significantly more complex than ours).

Axiomatic semantics for speculative leaks: A few approaches formalise the effects of speculatively executed instructions using axiomatic semantics inspired by work on weak memory models. For instance, Colvin and Winter [13] and Disselkoen et al. [15] capture the effects of branch speculation but both lack program analyses.

Ponce de León and Kinder [32] illustrate how one can model leaks resulting from speculation over branch instructions and store-bypasses using the CAT modeling language for memory consistency, and they present a bounded model checking analysis for detecting speculative leaks. Interestingly, they talk about composing several of theirs semantics [32, §IV.F], which should allow them to detect vulnerabilities like Listing 1 (which we detect under \(\mathcal{L}_{\text{B}+\text{S}}\)). However, they do not formally characterize compositions and, therefore, they cannot derive interesting results “for free” about the composed semantics (like we do in Theorem 4). Moreover, even though they state that composability is an advantage of axiomatic models, our framework (and tool implementation) shows that composability can be done with operational semantics as well.

9 CONCLUSION AND FUTURE WORK

This paper presented new speculative semantics for speculation on store and return instructions. It also defined a general framework to reason about the composition of different speculative semantics and instantiated the framework with our new speculative semantics \(\mathcal{L}_{\text{S}}\) and \(\mathcal{L}_{\text{R}}\) and the semantics by Guarnieri et al. [21]. Our framework yields security of the composed semantics (almost) for free, given the security of its parts. All the new semantics have been implemented in the Spectector program analysis tool, which correctly detects all vulnerabilities in existing and novel benchmarks.

Acknowledgments: This work was partially supported by the Madrid regional government under the project S2018/TCS-4339 BLOQUES-CM, by the Spanish Ministry of Science, Innovation, and University under the project RTI2018-102043-B-I00 SCUM, by the Italian Ministry of Education under the Rita Levi Montalcini grant (2019 call), by the German Ministry for Education and Research under the project CISPA-Stanford Center for Cybersecurity (funding number: 16KIS0761), and by a gift from Intel Corporation.
REFERENCES


