A model of speculative evaluation

CRAIG DISSELKOEN, University of California San Diego and Mozilla Research Internship
RADHA JAGADEESAN, DePaul University
ALAN JEFFREY, Mozilla Research
JAMES RIELY, DePaul University

1 INTRODUCTION

This paper studies information flow caused by speculation mechanisms in hardware and software.

Information flow from high (hi) to low (lo) provides a formal foundation for end-to-end security. Informally, a program is secure if there is no observable dependency of lo-observables on hi-inputs. The precise formalization of this intutive idea has been the topic of extensive research (e.g., see [14] for a detailed survey till 2006), and can be generalized to account for a variety of language features and observables such as non-determinisim [16], concurrency [15], reactivity [12], and probability [7]. The static and dynamic enforcement of these definitions in general purpose languages [11] has also been studied extensively and has influenced language design and implementation.

A key parameter in the definitions cited above is the notion of the *observational power* of the attacker model. Whereas the classical input-output behavior is often an adequate foundation, it has long been known that side-channels that leak information arise from other observables such as execution time and power consumption.

The focus of this paper is the development of a compositional semantic model of executions of programs to explicate side channel attacks that arise from speculation. In particular, we model conditionals such that it is possible for the behaviour of a conditional ($\texttt{if}\ (M)\ \{\, C \,\}\ \texttt{else}\ \{\, D \,\}$) to depend on the behaviour of D even when the condition M is true.

Our study addresses several sources of speculation in the concurrent execution of programs on modern microprocessors.

- Pipelined microprocessors use predictive schemes to utilize resources more efficiently and improve performance. Informally, these schemes use the prior history to predict the path of a program, e.g. by guessing memory dependencies, or the outcome of conditionals. Execution proceeds along the predicted path until it is possible to validate the prediction; the execution is committed if the prediction is correct; otherwise, the execution is rolled back. The speculation does not affect the observable input-output behavior of the program, thus ensuring correctness with respect to the usual intended semantics of the program. However, since there is clear timing differences between the cases where the prediction is and is not successful, it raises the concern that the predictive mechanisms themselves could cause information-leaks via timing side-channels; a fear that is realized with devastating impact by the Spectre family of attacks [8].

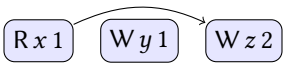
Authors' addresses: Craig Disselkoen, University of California San Diego, Mozilla Research Internship, cdisselk@cs.ucsd.edu; Radha Jagadeesan, DePaul University, rjagadeesan@cs.depaul.edu; Alan Jeffrey, Mozilla Research, ajeffrey@mozilla.com; James Riely, DePaul University, jriely@cs.depaul.edu.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2018 Copyright held by the owner/author(s).
XXXX-XXXX/2018/7-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

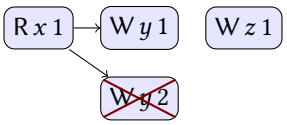
- Several modern microprocessors [4] support transactions to aid in the design of correct and efficient concurrent programs. Transactions are executed optimistically; they are committed if there are no conflicts, and aborted otherwise. All memory effects of an aborted transaction are rolled back; so there is no way for a concurrent observer to detect an aborted transaction. However, the thread of the aborted transaction, via abort-handler code, gets notified of the cancellation of the transaction.
  Thus, when a transaction of a thread aborts, the thread can deduce plausible information about specific memory accesses of a concurrently executing transaction. In combination with the techniques used in Spectre, this potential has been exploited recently to accelerate and scale up the efficacy of the attacker in the Spectre family of attacks [5].
- Compiler optimizations may depend on both branches of a conditional, for example one that replaces $(\texttt{if}\,(M)\,\{\,C\,\}\,\texttt{else}\,\{\,C\,\})$ by C. In particular, $(\texttt{if}\,(r)\,\{\,x:=1\,\}\,\texttt{else}\,\{\,x:=1\,\})$ can be optimized to $(x:=1)$, which does not have a control dependency from r to the assignment to x. In contrast, the program $(\texttt{if}\,(r)\,\{\,x:=1\,\}\,\texttt{else}\,\{\,x:=2\,\})$ cannot be optimized, so the control dependency cannot be removed.
  This would be fine if there were no program constructs which could observe control dependencies, but unfortunately relaxed memory models such as JMM [10], C++ [3] or LLVM [18] allow for such observations. This means there is the possibility for information flows caused by optimizing compilers, which we investigate in this paper.
  Whereas information flows caused by speculation in hardware, or by transactional memory are known [**? ?**], these attacks on compiler optimizations, and the relaxed memory models that justify them, is new. In this paper we provide both a formal model for such attacks, and experimental evidence about their practicality

Information flow attacks motivate the main technical development of this paper. This line of research was initiated by Zhang et al. [17]. Whereas they explore static annotations to address sidechannels in the context of hardware description languages, we explore a model of programs that captures enough detail to reveal and analyze the presence of side channels revealed by speculative execution.

The model is based on *partially ordered multisets* [13], whose labels are given by read and write actions. These can be visualized as a graph where the edges indicate dependencies, for example $(r:=x;\ y:=1;\ z:=r+1)$ has an execution modelled by the pomset:



The edge from $(\mathrm{R}\,x\,1)$ to $(\mathrm{W}\,z\,2)$ indicates a data dependency. The novel aspect of the model is that events have *preconditions* which may be false (and are visualized by crossing out the event). These are used in giving the semantics of conditionals, for example $(\texttt{if}\,(x)\,\{\,y:=1;\ z:=1\,\}\,\texttt{else}\,\{\,y:=2;\ z:=1\,\})$ has an execution:



Th edges from $(\mathrm{R}\,x\,1)$ to $(\mathrm{W}\,y\,1)$ and $(\mathrm{W}\,y\,2)$ indicate control dependencies. The presence of a crossed out $(\mathrm{W}\,y\,2)$ indicates a failed speculation.

The novel contributions of this paper are:

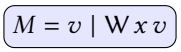
- a model of program execution that includes speculation (§2),

- examples showing how the model can be applied, including information flow attacks on hardware, optimizing compilers, and transactional memory (§3), and
- experimental evidence about how practical it is to mount the new class of attacks (§4).

The information flow attacks against optimizing compilers were developed as a direct result of building the formal model.

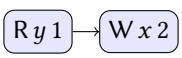
2 MODEL

The model used in this paper is one of sets of pomsets with event labels of the form $(\phi \mid a)$, where ϕ is the event's precondition (such as $M = v$) and a is the event's action (such as $\mathrm{W}\, x\, v$). For example the semantics of the program $(x := M)$ includes the case where M is v, which is written to x, and is captured by the one-event pomset:

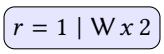


For this reason, the model is parameterized by a logic, used to express the preconditions of actions. We make few requirements of this logic, save that it has includes equalities between expressions, is closed under substitution, and supports a notion of implication (and in particular a notion of when a formula is a tautology).

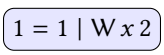
The semantics is defined compositionally. For example, the program $(x := y + 1)$ is shorthand for $(r := y;\ x := r + 1)$, which contains the pomset:



This pomset is build compositonally. First, $[\![x := r + 1]\!]$ contains the pomset:

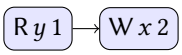


Next, we perform the substitution of r with 1 in every precondition, to get that $[\![x := r + 1]\!][1/r]$ contains the pomset:

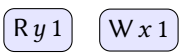


and since $(1 = 1)$ is a tautology, we elide it:

This substitution is performed in defining $[\![r := y;\ x := r + 1]\!]$, which contains the pomset:



as required. There is an ordering $(\mathrm{R}\, y\, 1) < (\mathrm{W}\, x\, 2)$ because the precondition $(r = 1)$ depends on r. If the preconditon was independent of r then there would be no ordering, for example $[\![r := y;\ x := r + 1 - r]\!]$ contains the pomset:



since the precondition $(r + 1 - r = 1)$ is independent of r.

The main novelty of our semantics, since it is designed to model speculative evaluation, is in modelling conditionals. In most sets-of-pomsets semantics, a pomset in $[\![\texttt{if}\ (M)\ \{\, C \,\}\ \texttt{else}\ \{\, D \,\}]\!]$ would either be given by a pomset in $[\![C]\!]$ or a pomset in $[\![D]\!]$. To model speculative evaluation,

we need to allow a pomset in $[\![\texttt{if}\,(M)\,\{\,C\,\}\,\texttt{else}\,\{\,D\,\}]\!]$ to be given by both a pomset in $[\![C]\!]$ *and* a pomset in $[\![D]\!]$. For example, $[\![\texttt{if}\,(M)\,\{\,x:=1\,\}\,\texttt{else}\,\{\,x:=2\,\}]\!]$ contains the pomset:

$$\boxed{M \neq 0 \mid \mathrm{W}\,x\,1} \quad \boxed{M = 0 \mid \mathrm{W}\,x\,2}$$

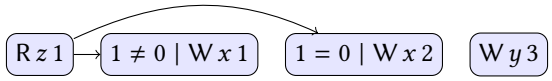
that is we have recorded behaviour from both branches of execution. Moreover, an action which is performed on both sides of the conditional can be merged, producing only one event in the resulting pomset. The precondition of the merged event is the disjunction of the preconditions of the original events. For example $[\![\texttt{if}\,(M)\,\{\,x:=1;\,y:=3\,\}\,\texttt{else}\,\{\,x:=2;\,y:=3\,\}]\!]$ contains the pomset:

$$\boxed{M \neq 0 \mid \mathrm{W}\,x\,1} \quad \boxed{M = 0 \mid \mathrm{W}\,x\,2} \quad \boxed{(M \neq 0) \vee (M = 0) \mid \mathrm{W}\,y\,3}$$

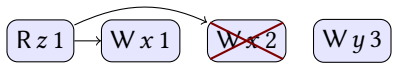
and since $(M \neq 0) \vee (M = 0)$ is a tautology, this is:

$$\boxed{M \neq 0 \mid \mathrm{W}\,x\,1} \quad \boxed{M = 0 \mid \mathrm{W}\,x\,2} \quad \boxed{\mathrm{W}\,y\,3}$$

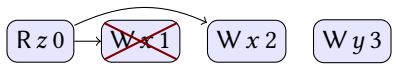
Combining this model of conditionals with the model of memory using substitutions gives that $[\![\texttt{if}\,(z)\,\{\,x:=1;\,y:=3\,\}\,\texttt{else}\,\{\,x:=2;\,y:=3\,\}]\!]$ contains the pomset:



and since $(1 \neq 0)$ is a tautology and $(1 = 0)$ is unsatisfiable, this is:



Similarly, $[\![\texttt{if}\,(z)\,\{\,x:=1;\,y:=3\,\}\,\texttt{else}\,\{\,x:=2;\,y:=3\,\}]\!]$ contains the pomset:



Note that this semantics captures control dependencies such as $(\mathrm{R}\,z\,0) < (\mathrm{W}\,x\,1)$, independencies such as $(\mathrm{R}\,z\,0) \not< (\mathrm{W}\,y\,3)$, and failed speculations such as the crossed out $(\mathrm{W}\,x\,1)$.

In summary, the features we need of the underlying data model are:

- *actions*, which may read or write to memory locations, and
- *preconditions*, which form a logic closed under substitution.

from which we can define the operations used in defining the semantics of programs, which include:

- *prefixing* $(\phi \mid a) \to \mathcal{P}$, which adds an event with precondition ϕ and action a to pomsets in $\mathcal{P}$,
- *substitution* $\mathcal{P}[M/x]$, which performs a substitution on every precondition in $\mathcal{P}$, and
- *join* $\mathcal{P}_1 \parallel \mathcal{P}_2$, which unions pomsets from $\mathcal{P}_1$ and $\mathcal{P}_2$, allowing events to be merged.

We make data models precise in §2.1, define pomsets in §2.2, and operations on sets of pomsets in §2.3, which are used to give a compositional semantics for a simple imperative language in §**??**.

2.1 Data models

A *data model* consists of:

- a set of *memory locations* $\mathcal{X}$, ranged over by x and y,
- a set of *registers* $\mathcal{R}$, ranged over by r and s,
- a set of *values* $\mathcal{V}$, ranged over by v and w,
- a set of *expressions* $\mathcal{E}$, ranged over by M and N,
- a set of *logical formulae* Φ, ranged over by ϕ and ψ, and

- a set of *actions* $\mathcal{A}$, ranged over by a and b,

such that:

- values include at least the constants 0 and 1,
- expressions include at least registers and values,
- expressions are closed under substitutions of the form $M[N/r]$,
- formulae include at least true, false, and equalities of the form $(M = N)$ and $(x = N)$,
- formulae are closed under negation, conjunction, disjunction,
- formulae are closed under substitutions of the form $\phi[x/r]$ or $\phi[N/x]$,
- there is a relation $\vDash$ between formulae, and
- there are partial functions R and W : $(\mathcal{A} \times \mathcal{X}) \to \mathcal{V}$,

We shall say *a reads v from x* whenever $\mathsf{R}(a, x) = v$, and *a writes v to x* whenever $\mathsf{W}(a, x) = v$. We shall say ϕ *implies* ψ whenever $\phi \vDash \psi$, ϕ is a *tautology* whenever true $\vDash \phi$, ϕ is *unsatisfiable* whenever $\phi \vDash$ false, and ϕ is *independent of x* whenever $\phi \vDash \phi[v/x] \vDash \phi$ for every v. In examples, the actions are of the form $(\mathsf{R}\, x\, v)$, which reads v from x, and $(\mathsf{W}\, x\, v)$, which writes v to x. Going forward, we assume a given data model, though some examples in §3 make use of particular $\mathcal{A}$.

2.2 3-valued pomsets

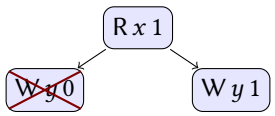
Recall the definition of a pomset from [**?**]:

Definition 2.1. A *pomset* $(E, \leq, \lambda)$ with alphabet Σ is a partial order $(E, \leq)$ together with a function $\lambda : E \to \Sigma$.

Going forward, we fix the alphabet $\Sigma = (\Phi \times \mathcal{A})$. We will write $(\phi \mid a)$ for the pair (ϕ, a), elide ϕ when ϕ is a tautology, and write ~~ϕ~~ when ϕ is unsatisfiable. We lift terminology from logical formulae and actions to events, for example if $\lambda(e) = (\phi \mid a)$ and $\lambda(d) = (\psi \mid b)$ then we say:

- e implies d whenever ϕ implies ψ,
- e is a tautology whenever ϕ is a tautlogy,
- e is unsatisfiable whenever ϕ is unsatisfiable,
- e is independent of x whenever ϕ is independent of x,
- e writes v to x whenever a writes v to x, and
- e reads v from x whenever a reads v from x.

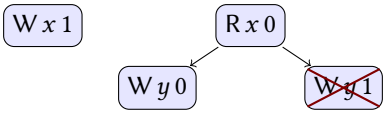
We visualize a pomset as a graph where the nodes are drawn from E, each node e is labelled with $\lambda(e)$, and an edge $d \to e$ corresponds to an ordering $d \leq e$. For example:



is a visualization of the pomset where:

$$0 \leq 1 \quad 0 \leq 2 \quad \lambda(0) = (\mathsf{true}, \mathsf{R}\, x\, 1) \quad \lambda(1) = (\mathsf{false}, \mathsf{W}\, y\, 0) \quad \lambda(2) = (\mathsf{true}, \mathsf{W}\, y\, 1)$$

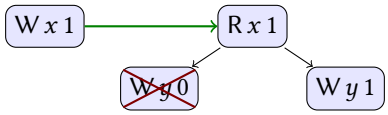
We are building a compositional semantics of shared memory concurrency, which means we require a notion of when a read has a matching write. This is a property we require of closed programs, but *not* of open programs. For example a program whose semantics includes:



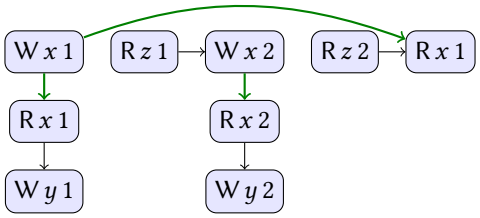
may be put in parallel with another program which writes 0 to x. If the program is closed wrt x is scoped, though, such an execution cannot exist, so we need each read of x to have a matching write. This is captured by defining when *e reads x from d* [2]. A preliminary definition (which as we shall see, needs strengthened) is:

- $d < e$,
- if e is satisfiable, then d is a tautology,
- d writes v to x, and e reads v from x, and
- there is no $d < c < e$ such that c writes to x.

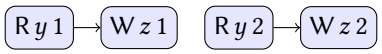
In diagrams, for readability we often highlight the reads-from edges, for example:



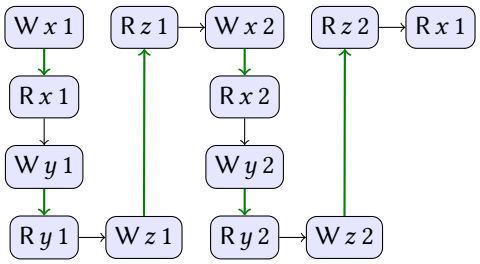
Unfortunately by itself, this is not enough. The problem is the final clause saying that there does not exist an *x-blocking* event c between d and e. Unfortunately, concurrency can turn events that were not x-blockers into an x-blocker, *even if the new thread does not mention x*. To see this, consider a program with execution:



If this is placed in parallel with:



then a possible excution is:



and now the (W x 2) event is an x-blocker, so (R x 1) cannot read from (W x 1).

There are a number of ways this can be addressed, for example in models such as [**?**] the reads-from relation is taken as a primitive. In this paper, we propose *3-valued posets* as a solution. These are posets in which as well as positive statements ($d < e$) (interpreted as e depends on d) we also have negative statements ($d \not< e$) (interpreted as e cannot depend on d).

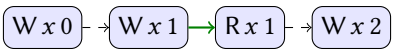
Definition 2.2. A *3-valued poset* $(E, \leq, \not<)$ is a poset $(E, \leq)$ together with $\not< \subseteq (E \times E)$ such that:

- if $d \leq e$ then $e \not< d$,
- if $d \leq e$ and $d \not< e$ then $d = e$,

- if $c \geq d \not< e$ or $c \not< d \geq e$ then $c \not< e$.

Definition 2.3. A *3-valued pomset* $(E, \leq, \not<, \lambda)$ is a 3-valued poset $(E, \leq, \not<)$ and a pomset $(E, \leq, \lambda)$.

In diagrams, we visualize $(e \not< d)$ as a dashed arrow from d to e (note the flip of direction). We will refer to edges introduced by $(d < e)$ as *strong edges* and those introduced by $(e \not< d)$ as *weak edges*, for example:



Structures similar to 3-valued posets have come up in many guises, for example rough sets [**?**], fuzzy sets [**?**] or ultrametrics over $\{0, {}^1/_2, 1\}$. They correspond to axioms A1–A3 of Lamport's *system executions* [**?**]. They are the notion of poset given by interpreting $d \leq e$ in a 3-valued logic [**?**].

We strengthen the definition of reads-from to require not just that no blocker exists, but that any candidate blocker must either have $d \not< c$ or $c \not< e$. This ensures that any further concurrency cannot turn a non-blocker into a blocker.

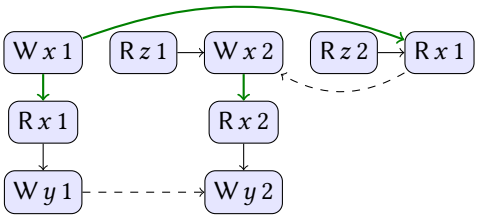
Definition 2.4. In a 3-valued pomset, *e can read x from d* whenever:

- $d < e$,
- if e is satisfiable, then d is a tautology,
- d writes v to x, and e reads v from x, and
- if c writes to x then either $d \not< c$ or $c \not< e$.

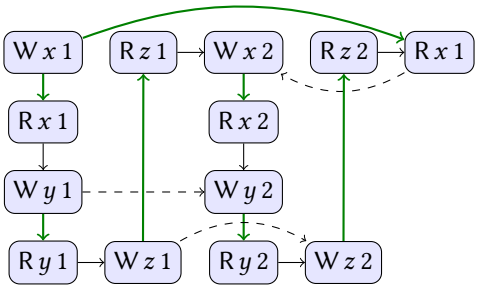
Definition 2.5. A 3-valued pomset is x-closed if for every $e \in E$:

- e is independent of x, and
- if e reads from x, then there is a d such that e can read x fom d.

In our previous example, in order for $(\mathrm{R}\, x\, 1)$ to read from $(\mathrm{W}\, x\, 1)$, we either need $(\mathrm{W}\, x\, 1) \not< (\mathrm{W}\, x\, 2)$ or $(\mathrm{W}\, x\, 2) \not< (\mathrm{W}\, x\, 1)$, for example:



then putting this in parallel as before results in:



but this is *not* a valid 3-valued pomset, since $(\mathrm{W}\, x\, 2) < (\mathrm{R}\, x\, 1)$ but also $(\mathrm{W}\, x\, 2) \not< (\mathrm{R}\, x\, 1)$, which is a contradiction.

2.3 Sets of 3-valued pomsets

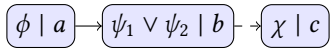
Our model of programs is going to be sets of rf-pomsets. In this section we define the operations on pomsets which are used in giving the semantics. These operations such as prefixing, parallel composition, restriction and so on, which are familiar from models of concurrency such as [**? ?**], but adapted to the setting of speculative evaluation.

Definition 2.6. Let $P_0 \in (\mathcal{P}_1 \parallel \mathcal{P}_2)$ whenever there are $P_1 \in \mathcal{P}_1$ and $P_2 \in \mathcal{P}_2$ such that:

- $E_0 = E_1 \cup E_2$,
- if $e \leq_1 d$ or $e \leq_2 d$ then $e \leq_0 d$,
- if $e \not<_1 d$ or $e \not<_2 d$ then $e \not<_0 d$,
- if $\lambda_0(e) = (\phi_0 \mid a)$ then either:
  - $\lambda_1(e) = (\phi_1 \mid a)$ and $\lambda_2(e) = (\phi_2 \mid a)$ and ϕ_0 implies $\phi_1 \vee \phi_2$,
  - $\lambda_1(e) = (\phi_1 \mid a)$ and $e \notin E_2$ and ϕ_0 implies ϕ_1, or
  - $\lambda_2(e) = (\phi_2 \mid a)$ and $e \notin E_1$ and ϕ_0 implies ϕ_2.

We use $\mathcal{P}_1 \parallel \mathcal{P}_2$ in defining the semantics of conditionals and concurrency. It contains the union of pomsets from $\mathcal{P}_1$ and $\mathcal{P}_2$, allowing overlap as long as they agree on actions. For example, if $\mathcal{P}_1$ and $\mathcal{P}_2$ contain:

then $\mathcal{P}_1 \parallel \mathcal{P}_2$ contains:

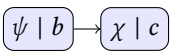


Definition 2.7. Let $(\phi \mid a) \rightarrow \mathcal{P}$ be the set $\mathcal{P}'$ where $P' \in \mathcal{P}'$ whenever there is $P \in \mathcal{P}$ such that:

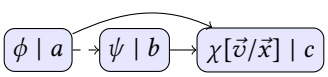
- $E' = E \cup \{c\}$,
- if $d \leq e$ then $d \leq' e$,
- if $d \not< e$ then $d \not<' e$,
- $\lambda'(c) = (\psi, a)$, where ψ implies ϕ, and
- if $\lambda(e) = (\psi \mid b)$ then:
  - $\lambda'(e) = (\psi' \mid b)$,
  - ψ' implies $\psi[\vec{v}/\vec{x}]$,
  - if ψ is dependent on any $\vec{x}$, then $c \leq' e$, and
  - if b writes to $\vec{y}$, then $c \not>' e$.

where a reads $\vec{v}$ from $\vec{x}$ and writes $\vec{w}$ to $\vec{y}$.

Prefixing is used to define the semantics of reads and writes, and adds a new event c with action a and a precondition which implies ϕ. It performs a substitution on the preconditions of any subsequent events. The tricky part of the definition is the last two clauses, which places requirements on $c \leq e$ and $c \not> e$. In normal prefixing, these would always be required, but we only require it when either the precondition of e depends on one of the variables read by a, or if c and e write to the same variables. For example, if $\mathcal{P}$ contains:



where a reads $\vec{v}$ from $\vec{x}$ and writes $\vec{w}$ to $\vec{y}$, b writes to $\vec{y}$, and ψ is independent of $\vec{x}$, then $(\phi \mid a) \rightarrow \mathcal{P}$ contains:



Definition 2.8. Let $\mathcal{P}[M/x]$ be the set $\mathcal{P}'$ where $P' \in \mathcal{P}'$ whenever there is $P \in \mathcal{P}$ such that:

- $E' = E$,
- if $d \leq e$ then $d \leq' e$, and
- if $d \not\leq e$ then $d \not\leq' e$, and
- if $\lambda(e) = (\psi \mid a)$ then $\lambda'(e) = (\psi[M/x] \mid a)$.

and similarly for $\mathcal{P}[x/r]$.

Definition 2.9. Let $(\phi \mid \mathcal{P})$ be the subset of $\mathcal{P}$ such that $P \in \mathcal{P}$ whenever:

- if $\lambda(e) = (\psi \mid a)$ then ϕ implies ψ.

Definition 2.10. Let $(vx\,.\,\mathcal{P})$ be the subset of $\mathcal{P}$ such that $P \in \mathcal{P}$ whenever P is x-closed.

We can use the operations defined above to give the semantics of a simple concurrent imperative programming language.

Definition 2.11.

$$\begin{aligned}
[\![\texttt{skip}]\!] &= \{\emptyset\} \\
[\![x := M; C]\!] &= \bigcup_v (M = v \mid \mathrm{W}\,x\,v) \to [\![C]\!][M/x] \\
[\![r := x; C]\!] &= [\![C]\!][x/r] \cup \bigcup_v (\mathrm{R}\,x\,v) \to [\![C]\!][x/r] \\
[\![\texttt{if}\,(M)\,\{C\}\,\texttt{else}\,\{D\}]\!] &= (M \neq 0 \mid [\![C]\!]) \parallel (M = 0 \mid [\![D]\!]) \\
[\![C \,||\, D]\!] &= [\![C]\!] \parallel [\![D]\!] \\
[\![\texttt{var}\,x; C]\!] &= vx\,.\,[\![C]\!]
\end{aligned}$$

A write generates a write event that may be visible to other threads. A read may see a thread-local value, or it may generate a read event that must be justified by another thread. In the latter case, occurrences of r are replaced with x (rather than v) to ensure that dependencies are tracked properly.

We have completed the formal definition of our model of speculative evaluation, and now turn to examples of this model in use.

3 EXAMPLES

3.1 Sequential memory accesses

In the semantics of memory, there are two very different ways memory can be accessed: sequentially or concurrently. These are modelled differently, since hardware and compilers give very different guarantees about their behaviour. In this section, we discuss the sequential semantics, and leave the concurrent semantics to §3.2.

Consider the program `(x := 0; y := x+1;)`. One execution of this program is where the write to y uses the sequential value of x, which is 0:

$$\boxed{\mathrm{W}\,x\,0} \quad \boxed{\mathrm{W}\,y\,1}$$

To see how this execution is modelled, we first expand out the syntax sugar to get the program `(x := 0; r := x; y := r+1; skip)`. Now $[\![\texttt{skip}]\!]$ is just $\{\emptyset\}$, and $[\![y := r + 1; \texttt{skip}]\!]$ includes:

$$(r + 1 = 1 \mid \mathrm{W}\,y\,1) \to [\![\texttt{skip}]\!][1/y]$$

which contains the pomset:

$$\boxed{r + 1 = 1 \mid \mathrm{W}\,y\,1}$$

expressing that this program can write 1 to y, as long as the precondition $(r + 1 = 1)$ is satisfied. Now $[\![r := x;\ y := r + 1;\ \texttt{skip}]\!]$ has two cases, the sequential case (which does not introduce a read action) and the concurrent case (which does). For the moment, we are interested in the sequential case, which is:

$$[\![y := r + 1;\ \texttt{skip}]\!][x/r]$$

which contains the pomset:

$$\boxed{x + 1 = 1 \mid \mathrm{W}\, y\, 1}$$

In this pomset, the precondition is $(x + 1 = 1)$, which specifies a property of the thread-local value of x. Finally $[\![x := 0;\ r := x;\ y := r + 1;\ \texttt{skip}]\!]$ includes:

$$(0 = 0 \mid \mathrm{W}\, x\, 0) \to [\![r := x;\ y := r + 1;\ \texttt{skip}]\!][0/x]$$

which contains the pomset:

$$\boxed{0 = 0 \mid \mathrm{W}\, x\, 0} \quad \boxed{0 + 1 = 1 \mid \mathrm{W}\, y\, 1}$$

all of whose preconditions are tautologies, so this has the expected behaviour:

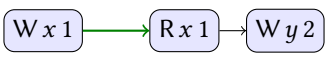
$$\boxed{\mathrm{W}\, x\, 0} \quad \boxed{\mathrm{W}\, y\, 1}$$

Note that there is no requirement of order between $(\mathrm{W}\, x\, 0)$ and $(\mathrm{W}\, y\, 1)$.

This example demonstrates how preconditions capture the sequential semantics of memory. In an execution containing an event with label $(\phi \mid a)$, one way the precondition ϕ can be discharged is by a write $x := M$, which performs a substitution $[M/x]$. This is a variant of the usual Hoare semantics for assignment, where if C has preconditon ϕ then $x := M;\ C$ has precondition $\phi[M/x]$.

3.2 Concurrent memory accesses

We now turn to the case of concurrent accesses to memory. Consider a concurrent version of the program from §3.1: `(x := 1; || y := x+1;)`. One execution of this program is where the write to y performs a concurrent read of x:



To see how this execution is modelled, we first expand out the syntax sugar to get the program `(x := 1; skip || r := x; y := r+1; skip)`. As before, $[\![y := r + 1;\ \texttt{skip}]\!]$ includes:

$$(r + 1 = 2 \mid \mathrm{W}\, y\, 2) \to [\![\texttt{skip}]\!][2/y]$$

which contains the pomset:

$$\boxed{r + 1 = 2 \mid \mathrm{W}\, y\, 2}$$

Now $[\![r := x;\ y := r + 1;\ \texttt{skip}]\!]$ has two cases, the sequential case (which does not introduce a read action) and the concurrent case (which does). We are now interested in the concurrent case, which includes:

$$(\mathrm{R}\, x\, 1) \to [\![y := r + 1;\ \texttt{skip}]\!][x/r]$$

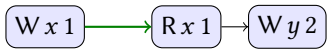
which contains the pomset:

$$\boxed{\mathrm{R}\, x\, 1} \to \boxed{\mathrm{W}\, y\, 2}$$

Note that $(\mathrm{R}\, x\, 1)$ reads 1 from x, and while $(x + 1 = 2)[1/x]$ is a tautlogy $(x + 1 = 2)$ is *not* independent of x and so there is an ordering $(\mathrm{R}\, x\, 1) < (\mathrm{W}\, y\, 2)$ modelling the data dependency of the write of y on the read of x.

Now, $[\![x := 1;\ \texttt{skip}]\!]$ includes the pomset:

and so $[\![x := 1;\ \texttt{skip}\ ||\ r := x;\ y := r + 1;\ \texttt{skip}]\!]$ includes:



as expected, including an reads-from dependency between the concurrent write of x and the matching read.

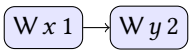
This example demonstrates how read and write events capture the concurrent semantics of memory. In an execution containing an event with label $(\mathrm{R}\, x\, v)$, if the execution is x-closed, then there must be an event it reads from, for example one labelled $(\mathrm{W}\, x\, v)$.

3.3 Independent writes

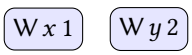
Consider an example with two independent writes `(x := 1; y := 2;)`. This has the semantics which includes:

$$(1 = 1 \mid \mathrm{W}\, x\, 1) \rightarrow (2 = 2 \mid \mathrm{W}\, y\, 2) \rightarrow \{\emptyset\}$$

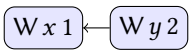
One of the executions this contains is:



but it also contains:



and:

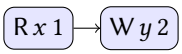


since there is no requirement that $(\mathrm{W}\, x\, 1) \leq (\mathrm{W}\, y\, 2)$.

Thus, the semantics of `(x := 1; y := 2;)` is the same as the semantics of `(y := 2; x := 1;)`.

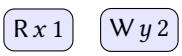
3.4 Independent reads and writes

Whereas write prefixing introduces weak dependencies on events which write the the same location, read prefixing introduces strong dependencies on preconditions which depend on the location being read. For example in §3.2 we saw that the program `(r := x; y := r+1;)` includes the pomset:



but since $(x + 1 = 2)$ depends on x, we have the requirement that $(\mathrm{R}\, x\, 1) \leq (\mathrm{W}\, y\, 2)$.

This is in contrast to the program `(r := x; y := r+2-r;)`. Since $(x + 2 - x = 2)$ is independent of x (at least for integer arithmetic) this contains:



and we can show the semantics of `(r := x; y := r+2-r;)` is the same as the semantics of `(y := 2; r := x;)`.

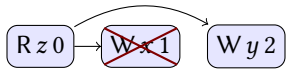
Note this this example shows that we are not just dealing with a syntactic notion of dependency, which is common in hardware models of memory. In syntactic dependency, since r occurs free in $(y := r + 2 - r)$, there would be a dependency between $(r := x)$ and $(y := r + 2 - r)$. In contrast, this model is based on logical implication, which can be interpreted semantically.

3.5 Control dependencies

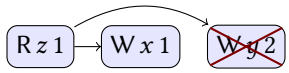
Conditionals introduce control dependencies, for example consider the program:

```
r := z; if (r) { x := 1; } else { y := 2; }
```

This includes executions in which the false branch is taken:



and ones where the true branch is taken:



In both cases, we record the actions in the branch that was not taken. This is a novel feature of this model, and is intended to capture speculative evaluation. In §3.7 we will show how this model captures Spectre-like information flow attacks, once the attacker is provided with the ability to observe such speculations.

To see how these executions are modelled, consider the semantics of $[\![x := 1;\ \texttt{skip}]\!]$, which contains any pomset of the form:

$$\phi \mid \mathrm{W}\, x\, 1$$

in particular it contains:

$$r \neq 0 \mid \mathrm{W}\, x\, 1$$

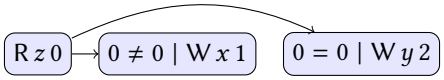
Similarly $[\![y := 2;\ \texttt{skip}]\!]$ contains:

$$r = 0 \mid \mathrm{W}\, y\, 2$$

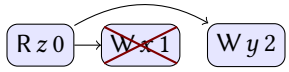
and so $[\![\texttt{if}\,(r)\,\{\,x := 1;\ \texttt{skip}\,\}\,\texttt{else}\,\{\,y := 2;\ \texttt{skip}\,\}]\!]$ contains:

$$r \neq 0 \mid \mathrm{W}\, x\, 1 \qquad r = 0 \mid \mathrm{W}\, y\, 2$$

Now, the semantics of concurrent read performs substitutions, for example:



which gives the required pomset:



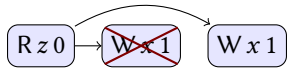
Note that the precondition $r = 0$ is dependent on r, and so there is a dependency $(\mathrm{R}\, z\, 0) < (\mathrm{W}\, y\, 2)$, modelling the control dependency introduced by the conditional.

3.6 Control independencies

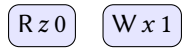
In most models of control dependencies, the dependency relation is syntactic, based on whether the action occurs inside syntactically inside a conditional. In contrast, the notion in this model is semantic: if an action can occur on both sides of a conditional, there is no control dependency. Consider a variant of the example from §3.5:

```
r := z; if (r) { x := 1; } else { x := 1; }
```

This has the expected execution in which the control dependencies exist:



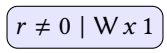
but it also has an execution in which the two writes of 1 to x are merged, resulting in no dependency:



To see how this arises in the model, consider the definition of $[\![\texttt{if}\ (r)\ \{\ x\texttt{:=}1\texttt{;}\ \texttt{skip}\ \}\ \texttt{else}\ \{\ x\texttt{:=}1\texttt{;}\ \texttt{skip}\ \}]\!]$:

$$\mathcal{P}_1 \sqcup \mathcal{P}_2 \quad \text{where} \quad \mathcal{P}_1 = (r \neq 0 \mid [\![x\texttt{:=}1\texttt{;}\ \texttt{skip}]\!]) \quad \text{and} \quad \mathcal{P}_2 = (r = 0 \mid [\![x\texttt{:=}1\texttt{;}\ \texttt{skip}]\!])$$

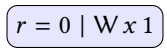
Now, one pomset in $\mathcal{P}_1$ is:



that is P_1 where:

$$E_1 = \{e\} \quad \lambda_1(e) = (r \neq 0, \mathrm{W}\,x\,1)$$

and similarly, one pomset in $\mathcal{P}_2$ is:



that is P_2 where:

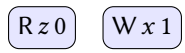
$$E_2 = \{e\} \quad \lambda_2(e) = (r = 0, \mathrm{W}\,x\,1)$$

Crucuially, in the definition of $\mathcal{P}_1 \sqcup \mathcal{P}_2$ there is *no* requirement that E_1 and E_2 are disjoint, and in this case they overlap at e. As a result, one pomset in $\mathcal{P}_1 \sqcup \mathcal{P}_2$ is P_0 where:

$$E_0 = \{e\} \quad \lambda_0(e) = (r \neq 0 \vee r = 0, \mathrm{W}\,x\,1)$$

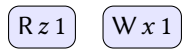
that is:

Note that this pomset has no precondition dependent on r, since $(r \neq 0 \vee r = 0)$ does not depend on r, which is why we end up with an execution without a control dependency:

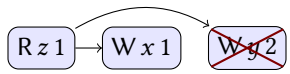


This semantics captures compiler optimizations which may, for example merge code executed on both branches of a conditional, or hoist constant assignments out of loops.

We can now see the counterintuitive behavior of conditionals in the presence of control dependencies. There are programs such as `(r := z; if (r) { x := 1; } else { x := 1; })` with executions in which $(\mathrm{W}\,x\,1)$ is independent of $(\mathrm{R}\,z\,1)$:



while programs such as `(r := z; if (r) { x := 1; } else { y := 2; })` only have executions in which $(\mathrm{W}\,x\,1)$ is dependent on $(\mathrm{R}\,z\,1)$:



so these programs have different dependency relations, depending on conditional branches that were not taken. In §3.9 we shall see that this has security implications, since relaxed memory can observe dependency. The attack is similar to Spectre, so we shall take a detour to see how Spectre can be modeled in this setting.

3.7 Spectre

We give a simplified model of Spectre attacks, ignoring the details of timing. In this model, we extend programs with the ability to tell whether a memory location has been touched (in practice this is implemented using timing attacks on the cache). For example, we can write a SPECTRE program as:

```
var a;
if (canRead(SECRET)) { a[SECRET] := 1; }
else if (touched a[0]) { x := 0; }
else if (touched a[1]) { x := 1; }
```

This is a low-security program, which is attempting to discover the value of a high-security variable SECRET. The low-security program is allowed to attempt to escalate its privileges by checking that it is allowed to read a high-security variable:

```
if (canRead(x)) { ... code allowed to read x ... }
else { ... fallback code ... }
```

In this case, the canRead(SECRET) is false, so the fallback code is executed. Unfortunately, the escalated code is speculatively evaluated, which allows information to leak by testing for which memory locations have been touched.

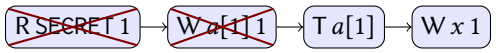
We model the touched test by introducing a new action $(\mathsf{T}\,x)$ and defining:

$$[\![\,\mathtt{if\ touched}\,x\,\{\,C\,\}\,\mathtt{else}\,\{\,D\,\}\,]\!] \quad = \quad ((\mathsf{T}\,x) \to [\![C]\!]) \cup [\![D]\!]$$

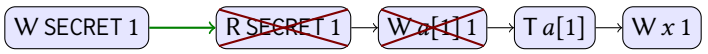
The additional requirement we need to add for x-closure is:

- if $\lambda(e) = (\phi \mid \mathsf{T}\,x)$ then there is $d < e$ where d reads or writes x.

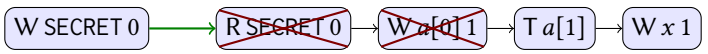
Note that there is no requirement that d be satisfiable, and indeed SPECTRE has the execution:



Putting this in parallel with a high-security write to SECRET gives:



but due the requirement of a-closure we do *not* have:



Thus, the attacker has managed to leak the value of a high-security location to a low-security one.

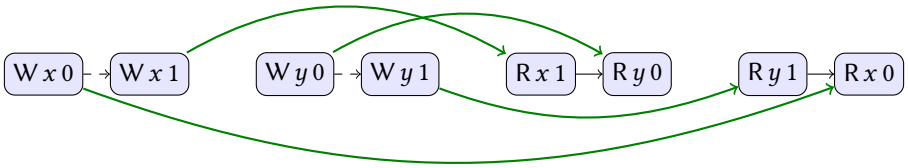
This shows how our model of speculative evaluation can express (very abstract, untimed) Spectre attacks.

3.8 Relaxed memory

In §3.9 we present an information flow attack on relaxed memory, similar to Spectre in that it relies on speculative evalutation. Unlike Spectre it does not depend on timing attacks, but instead is based on the sensitivity of relaxed memory to data dependencies. For this reason, we present a simple model of relaxed memory, which is strong enough to capture this attack. The model includes concurrent memory accesses, which can introduce concurrent reads-from. Since we are allowing events to be partially ordered, this gives a simple model of relaxed memory, for example an independent read independent write (IRIW) example is:

```
x := 0; x := x+1; || y := 0; y := y+1; || if (x) { r := y; } || if (y) { s := x; }
```

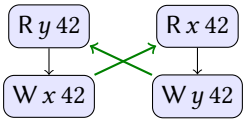
which includes the execution:



This model does not introduce thin-air reads (TAR), for example the TAR pit program is:

```
x := y; || y := x;
```

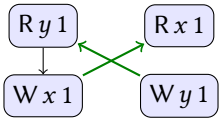
but an attempt to produce a value from thin air fails, for the usual reason of producing a cycle in $\leq$:



This cycle can be broken if one of the writes does not depend on the read, for example:

```
x := y; || r := x; y := r+1-r;
```

has execution:



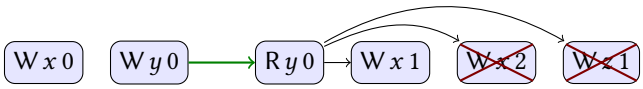
Note that $(\mathrm{R}\,x\,1) \not\leq (\mathrm{W}\,y\,1)$, so this does not introduce a cycle.

3.9 Information flow attacks on relaxed memory

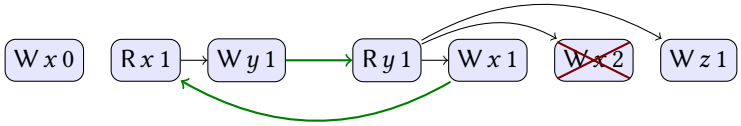
Consider an attacker program, again using security checks to try to learn a SECRET. Whereas SPECTRE uses hardware capabilities, which have to be modeled by adding extra capabilities to the language, this new attacker works by exploiting relaxed memory which can result in unexpected information flows. The attacker program is:

```
(
  x := 0; y := x;
) || (
  if (y == 0) { x := 1; }
  else if (canRead(SECRET)) { x := SECRET; }
  else { x := 1; z := 1; }
)
```

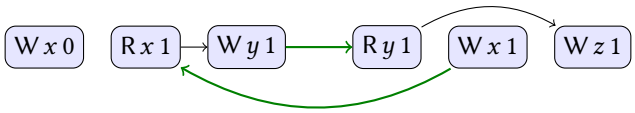
In the case where SECRET is 2, this has many executions, one of which is:



but there are no executions which exhibit (W z 1), since any attempt to do so produces a cycle:



In the case where SECRET is 1, there is an execution:



Note that in this case, there is no dependency from (R y 1) to (W x 1), which is what makes this execution possible. Thus, if the attacker sees an execution with (W z 1), they can conclude that SECRET is 1, which is an information flow attack.

This attack is not just an artifact of the model, since the same behavior can be exhibited by compiler optimizations. Consider the program fragment:

```
if (y == 0) { x := 1; }
else if (canRead(SECRET)) { x := SECRET; }
else { x := 1; z := 1; }
```

Now, in the case where SECRET is a constant 1, the compiler can inline it:

```
if (y == 0) { x := 1; }
else if (canRead(SECRET)) { x := 1; }
else { x := 1; z := 1; }
```

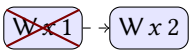
and lift the assignment to x out of the `if` statement:

```
x := 1;
if (y == 0) { }
else if (canRead(SECRET)) { }
else { z := 1; }
```

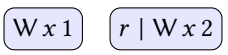
After these optimizations, a sequentially consistent execution exhibits (W z 1). We discuss the practicality of this attack further in §4.

3.10 Dead store elimination

A common compiler optimization is *dead store elimination*, in which writes are omitted if they will be overwritten by a subsequent write later in the same thread. For example, in the program `(x := 1; x := 2)`, the first write to x can be eliminated, since the second is guaranteed:



but in the program `(x := 1; if (r) { x := 2; })` the first write cannot be eliminated, since the second might not happen:



A simple model of dead store elimination changes the semantics of write to only introduce a write event if there is no subsequent guaranteed write of the same variable. We can do this by giving a looser interpretation of prefixing: in the definition of $(\phi \mid a) \to \mathcal{P}$, replace the requirement:

- $\lambda'(0) = (\psi, a)$, where ψ implies ϕ,

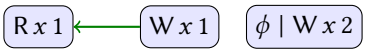
by:

- either:
  - $\lambda'(0) = (\psi, a)$, where ψ implies ϕ, or
  - $\lambda(0) = (\psi, b)$, where ϕ implies ψ, and every location a writes to is also written to by b.

This simple model includes the examples above. Note that if dead store elimination is *always* performed, then there is an information flow attack similar to the one in §3.9. Consider the program:

```
(
  r := x;
) || (
  x := 1;
  if (canRead(SECRET)) { if (SECRET) { x := 2; } }
  else { x := 2; }
)
```

In the case that SECRET is 0, there is an execution:



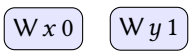
where ϕ is (¬canRead(SECRET)), which is not a tautology, and so the (W x 1) event is not omitted. In the case that SECRET is not 0, the matching execution is:

Now the (W x 2) event is a guaranteed write, so the (W x 1) is omitted. In the case that the attacker can rely on dead store elimination taking place, this is an information flow: if the attacker observes x to be 1, then they know SECRET is 0. We return to this attack in §4.

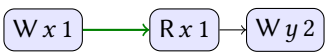
3.11 Thread inlining

One property one could ask of a model of shared memory is thread inlining: any execution of $[\![P; Q]\!]$ is an execution of $[\![P \mid\mid Q]\!]$. This is *not* a goal of our model, and indeed is not satisfied, due to the different semantics of concurrent and sequential memory accesses. We demonstrate this by considering an example from the Java Memory Model [**?**], which shows that Java does not satisfy thread inlining either.

The lack of thread inlining is related to the different dependency relations introduced by sequential and concurrent access. Recall from §3.1 that the program `(x := 0; y := x+1;)` has execution:



but that `(x := 1; || y := x+1;)` has:

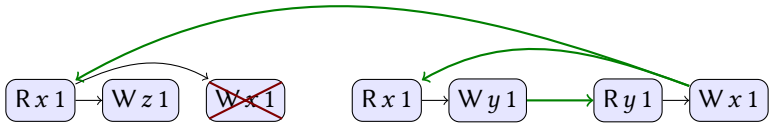


That is in the sequential case there is no depenency from the write of x to the write of y, but in the concurrent case there is such a dependecy.

This can be used to construct a counter-example to thread inlining, based on [**?** , Ex 11]:

```
x := 0; if (x == 1) { z := 1; } else { x := 1; } || y := x; || x := y;
```

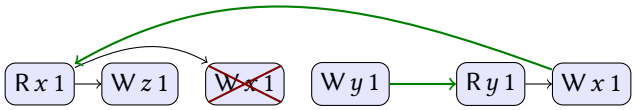
This has no execution containing $(\mathrm{W}\, z\, 1)$. Any attempt to build such an execution results in a cycle:



Inlining the thread `(y := x)` gives [**?** , Ex 12]:

```
x := 0; if (x == 1) { z := 1; } else { x := 1; } y := x; || x := y;
```

with execution:



To see why this execution exists, consider the program fragment:

```
if (x == 1) { z := 1; } else { x := 1; } y := x;
```

Removing the syntax sugar, this is:

```
r1 := x; if (r1 == 1) {
  z := 1; r2 := x; y := r2; skip
} else {
  x := 1; r3 := x; y := r3; skip
}
```

Now, $[\![z := 1; r_2 := x; y := r_2; \texttt{skip}]\!]$ includes pomset:

$$r_1 = 1 \mid \mathrm{W}\, z\, 1 \qquad r_1 = x = 1 \mid \mathrm{W}\, y\, 1$$

and $[\![x := 1; r_3 := x; y := r_3; \texttt{skip}]\!]$ includes pomset:

$$r_1 \neq 1 \mid \mathrm{W}\, x\, 1 \qquad r_1 \neq 1 \mid \mathrm{W}\, y\, 1$$

so $[\![\texttt{if}\,(r_1 = 1)\,\{\, z := 1; r_2 := x; y := r_2; \texttt{skip}\,\}\,\texttt{else}\,\{\, x := 1; r_3 := x; y := r_3; \texttt{skip}\,\}]\!]$ includes:

$$r_1 = 1 \mid \mathrm{W}\, z\, 1 \qquad r_1 \neq 1 \mid \mathrm{W}\, x\, 1 \qquad (r_1 = x = 1) \vee (r_1 \neq 1) \mid \mathrm{W}\, y\, 1$$

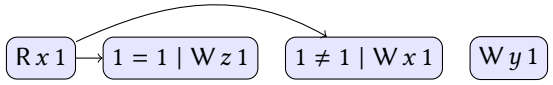
which means $[\![\texttt{if}\,(r_1 = 1)\,\{\, z := 1; r_2 := x; y := r_2; \texttt{skip}\,\}\,\texttt{else}\,\{\, x := 1; r_3 := x; y := r_3; \texttt{skip}\,\}]\!][x/r_1]$ includes:

$$x = 1 \mid \mathrm{W}\, z\, 1 \qquad x \neq 1 \mid \mathrm{W}\, x\, 1 \qquad (x = x = 1) \vee (x \neq 1)) \mid \mathrm{W}\, y\, 1$$

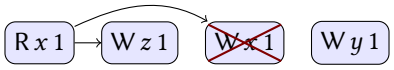
Now $(x = x = 1) \vee (x \neq 1)$ is a tautology, so this is just:

$$x = 1 \mid \mathrm{W}\, z\, 1 \qquad x \neq 1 \mid \mathrm{W}\, x\, 1 \qquad \mathrm{W}\, y\, 1$$

and so $[\![r_1 := x; \texttt{if}\,(r_1 = 1)\,\{\, z := 1; r_2 := x; y := r_2; \texttt{skip}\,\}\,\texttt{else}\,\{\, x := 1; r_3 := x; y := r_3; \texttt{skip}\,\}]\!]$ includes:



which simplifies to:



as required. The rest of the example is straightforward, and shows that our semantics agrees with the JMM in not supporting thread inlining.

3.12 Word tearing

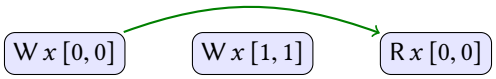
{Remove this section, since it's not needed for transactions?}

In §3.14, we shall be considering transactional memory, and in §3.14 show that we can model a simplified version of an information flow attack on transactions. In order to model transactions, we need to consider actions that can write many memory locations at once, since this is part of the semantics of commitment. To lead up to this, we first consider a simpler scenario of many-location writes and reads, which is word tearing.

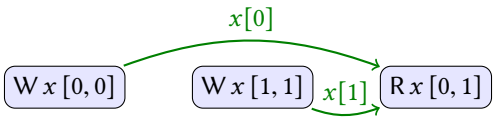
In word tearing, a program contains a write instruction with data larger than the hardware word size, for example copying a byte array, or assigning a 64-bit float on a 32-bit architecture. For example, consider the program:

```
(x := [0, 0];) || (x := [1, 1];) || (r := x;)
```

This has executions in which the read of x only reads from one of the writes, for example:



but also has executions in which the read of x reads from both writes, for example:



Word tearing can occur, for example, in Java extended floating point [**?**], LLVM 64-bit instructions on 32-bit hardware [**?**], or in JavaScript SharedArrayBuffers [**?**].

3.13 Release/acquire synchronization

In relaxed memory models, synchronization actions act as memory fences: that is, they are a barrier to reordering memory accesses. In this section, we present a simple model of release/acquire fencing. In §3.14, we show that this can be scaled up to a model of transactional memory.

We assume there are sets Rel and Acq $\subseteq \mathcal{A}$. We say that a is a *release action* if $a \in$ Rel and a is an *acquire action* if $a \in$ Rel. In a pomset, a release event is one labelled with a release action, and an aquire event is one labelled by an aquire action. The semantics of fences are given by adding an extra constraint to the definition of $(\phi \mid a) \to \mathcal{P}$ (recalling that c is the a-labelled event being introduced):

- $c \leq e$ whenever c is an aquire event or e is a release event.

This constraint ensures that events are ordered before a release and after an acquire.

In examples, we will use releasing writes and acquiring reads:

- (Rel x v), a release action that writes v to x, and
- (Acq x v), an acquire action that reads v from x.

The semantics of programs with releasing write and acquiring read are the same as for regular write and read and, but with $\mathsf{Rel}\,x\,v$ replacing $\mathsf{W}\,x\,v$ and $\mathsf{Acq}\,x\,v$ replacing $\mathsf{R}\,x\,v$.

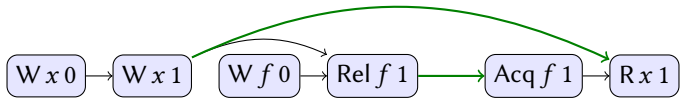
$$[\![\mathsf{rel}\,x := M;\,C]\!] \;=\; \bigcup_v (M = v \mid \mathsf{Rel}\,x\,v) \to [\![C]\!][M/x]$$

$$[\![\mathsf{acq}\,r := x;\,C]\!] \;=\; [\![C]\!][x/r] \cup \bigcup_v (\mathsf{true} \mid \mathsf{Acq}\,x\,v) \to [\![C]\!][x/r]$$

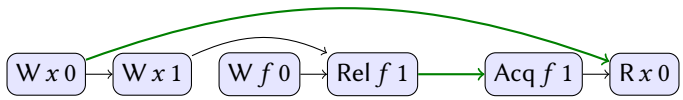
For example, consider the program:

```
(x := 0; f := 0; x:= 1; rel f := 1;) || (acq r := f; s := x;)
```

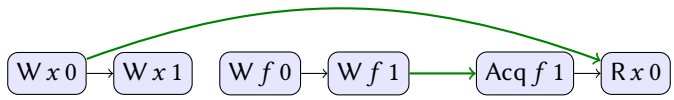
This has an execution:



but *not*:



since $(\mathsf{W}\,x\,0) < (\mathsf{W}\,x\,1) < (\mathsf{R}\,x\,0)$, so this pomset does not satisfy the requirements to be an rf-pomset. If we replace the release with a plain write, then the outcome $(\mathsf{Acq}\,f\,1)$ and $(\mathsf{R}\,x\,0)$ is possible:



since no order is required between $(\mathsf{W}\,x\,1)$ and $(\mathsf{W}\,f\,1)$. Symmetrically, if we replace the acquire of the original program with a plain read, then the outcome $(\mathsf{R}\,f\,1)$ and $(\mathsf{R}\,x\,0)$ is possible.

3.14 Transactions

We present a simple model of serializable transactions. The action $(\mathsf{B}\,v) \in \mathsf{Acq}$ represents the begin of a transaction with id v and $(\mathsf{C}\,v) \in \mathsf{Rel}$ represents the corresponding commit, with semantics:

$$[\![r := \mathtt{begin}\,v;\,D]\!] \;=\; (\mathsf{true} \mid \mathsf{B}\,v) \to [\![C]\!][v/r]$$

$$[\![r := \mathtt{commit}\,v;\,D]\!] \;=\; (\mathsf{true} \mid \mathsf{C}\,v) \to [\![D]\!][1/r] \cup (\mathsf{false} \mid \mathsf{C}\,v) \to [\![D]\!][0/r]$$

At top level, we require that pomsets be *serializable*, as defined below.

Definition 3.1. We say that event c *matches* b if $\lambda(c) = (\mathsf{C}\,v)$ and $\lambda(b) = (\mathsf{B}\,v)$, for some v. We say that a begin event *aborts* every matching commit is unsatisfiable. We say that b *begins* e if $b < e$ is a tautologous begin and there is no intervening tautologous matching commit; in this case e *belongs to* b. We say that c *commits* e if $c > e$ is a tautologous commit and there is no intervening tautologous matching begin. A pomset is *serializable* if:

(1) no two begins have the same id,
(2) every commit follows the matching begin,
(3) $<$ totally orders tautological begins and commits,
(4) if b begins e, but not d, and $d < e$ then $d < b$,
(5) if c ends e, but not d, and $e < d$ then $c < d$,
(6) if e and d belong to b and read the same location, then both read the same value,

(7) if e belongs to b, then e implies some matching c that ends e, and
(8) if b aborts then there must be some unsatisfiable matching commit c and events $b < e < c$ and $d < e$ such that both e and d touch the same location.

In discussion, we identify transactions by their unique begin event. A transaction that does not abort is *successful*. Conditions 1-5 ensure serializability. Conditions 4-6 also ensure strong isolation for non-transactional events [6]. Condition 7 ensures that all events in aborted transactions are unsatisfiable. Condition requires that aborts only occur due to caching conflicts—this is similar to the treatment of the touch operation in §3.7.

4 EXPERIMENTS

One theme of this paper is that optimizations not typically part of formal abstractions can result in information-flow leaks. This is typified by the Spectre attack, which leverages speculative execution, a hardware optimization. Sections 3.9 and 3.10 presented other attacks along this same theme, which leverage relaxed memory models and dead store elimination respectively. In particular, the latter attack (and, to a degree, the former attack), result not from hardware optimizations, but from common *compiler* optimizations. These attacks also, unlike Spectre, do not rely on timing side channels, or indeed timers of any kind, bypassing many common Spectre mitigations [**?**]. Here we demonstrate the efficacy of each of these attacks against modern compilers and hardware, including both the `clang` and `gcc` compilers.

All of our experiments are performed on a **{describe machine}** with `clang` version **{clang version}** and `gcc` version **{gcc version}**.

4.1 Relaxed memory attack

4.2 Dead store elimination attack

In this section we return to the attack in Section 3.10 based on dead store elimination.

As in Section 4.1, we assume that there is a `SECRET` which an attacker wishes to learn. For instance, `SECRET` may be a cryptographic key hardcoded into the application. This `SECRET` is known to the compiler at compile time, but may not be accessed except behind a security check. We assume that the security check always evaluates to `false` at runtime for the attacker, but that the attacker is allowed to write arbitrary code subject to the above restrictions. Despite the attacker's apparent inability to access `SECRET`, we show that the attacker can learn its value using the idea in Section 3.10. This attack was tested and works on both `clang` and `gcc`.

First, we start from the simple form of the attack presented in Section 3.10, and extend it to leak a secret consisting of an arbitrary number `N` of bits. To do this, we simply compile `N` copies of the function above, each performing a boolean test on a single bit of the secret. The function used for reading the `k`th bit is as follows (for `N <= 64`):

```
(
  r := x;
) || (
  x := 1;
  if (canRead(SECRET)) {
    if (SECRET & (1 << k)) { x := 2; }
  } else {
    x := 2;
  }
)
```

Then, we test each function in turn, each time noting the value of `r` observed by the second thread. The extension to the general case (with truly arbitrary `N`) is straightforward; `SECRET` becomes an array of 64-bit values, and we use `k / 64` and `1 << (k & 63)` as the array index and bitmask respectively.

We make three additional tweaks to improve the reliability so that the attacker can confidently infer the value of `SECRET` based on the observed values of `r`. First, we insert additional time-consuming computation immediately following the `x := 1` operation. This lengthens the timing window in which `x` has the value 1, increasing the likelihood that the other thread will be able to observe `x == 1` (unless the `x := 1` write was eliminated, of course). Inserting this computation can be done without interfering with the dead store elimination process itself, so that the compiler will continue to eliminate the `x := 1` write if and only if the appropriate bit of `SECRET` was 1. For `gcc`, we have a fair amount of freedom with the time-consuming computation (for instance, we can use an arbitrarily long loop), but with `clang`, the computation must be branch-free, and furthermore not consist of too many instructions. This is because `clang`'s dead store elimination pass operates only within basic blocks, and uses a heuristic to stop scanning the basic block early if it is too large. Nonetheless, we find that even with these restrictions, we are able to construct a reliable and fast attack against both `clang` and `gcc`.

Second, rather than simply observing `x` with `r := x` in the 'listening' thread, we continuously load `x` in a loop until a nonzero value is observed – i.e., we perform

```
do {
  r := x;
} while(r == 0);
```

This remedies the case where `r := x` could observe a value of `x` from 'before' either of the two possible writes performed by the other thread.

Finally, we redundantly execute the entire attack several times, noting the final value of `r` (the first observed nonzero value of `x`) in each case. We note that if *any* of the redundant runs produces `r == 1` for a particular bit position, we can be certain that the corresponding bit of `SECRET` *must* be 0, as it implies that the `x := 1` write was not eliminated in that particular function. On the other hand, the more runs that observe `r == 2` in a particular bit position despite our other reliability-increasing measures taken above, the more certain we can be that the `x := 1` write was eliminated in that function, and the appropriate bit of `SECRET` is 1.

Our implementation has two important "knobs" which trade off reliability vs. performance. First, we have the length of time which the writing thread attempts to "stall" immediately after the `x := 1` write. Second, we have the number of entire redundant runs of the attack that are performed before the attacker reaches her conclusion. Increased reliability can be achieved by adjusting either of these knobs, and they each have (different) effects on the overall performance of the attack. After exploring the parameter space, we found that 3 redundant runs is sufficient to provide near-100% accuracy while allowing us to maximize the speed of the attack. Specifically, on our machine, our attack on `gcc` reaches speeds of **{exact gcc leak speed}** bits leaked per second (**{exact gcc raw leak speed}** 'raw' bits leaked per second, that is, before error correction) with **{exact gcc accuracy}**, while our attack on `clang` reaches speeds of **{exact clang leak speed}** bits leaked per second (**{exact clang raw leak speed}** 'raw' bits leaked per second) with **{exact clang accuracy}**. In particular, this means our attack can leak a 2048-bit cryptographic key in under **{exact speed}** ms, on either `gcc` or `clang`, with probability **{exact probability}** that there are exactly zero bit errors in the leaked key, or probability **{exact probability}** that there is at most one bit error in the leaked key.

5 LOGIC

We work with a variant of PLTL [1] interpreted over pomsets. The atoms of our logic are write and read events, and we use strictly in the past modal operators in addition to the usual boolean connectives.

$$\begin{array}{rcl} \phi & ::= & \mathsf{R}\,x\,v \mid \mathsf{W}\,x\,v \\ & & \phi \wedge \psi \mid \neg\phi \\ & & \Diamond^{-1}\phi \mid \Box^{-1}\phi \end{array}$$

As usual, we write $\phi \vee \psi$ for $\neg(\neg\phi \wedge \neg\psi)$, and $\phi \Rightarrow \psi$ for $\neg\phi \vee \psi$.

Definition 5.1 (Satisfaction). Given an rf-pomset $P = (E, \leq, \lambda)$ with alphabet Σ, and $e \in E$, define:

$$\begin{array}{lcl} P, e & \models & \mathsf{R}\,x\,v, \text{ if } \lambda(e) = (\mathsf{true}, \mathsf{R}\,x\,v) \\ P, e & \models & \mathsf{W}\,x\,v, \text{ if } \lambda(e) = (\mathsf{true}, \mathsf{W}\,x\,v) \\ P, e & \models & \phi \wedge \psi, \text{ if } P, e \models \phi \text{ and } P, e \models \psi \\ P, e & \models & \neg\phi\,, \text{ if } P, e \not\models \phi \\ P, e & \models & \Diamond^{-1}\phi, \text{if } (\exists d \leq e, d \neq e)\ P, d \models \phi \\ P, e & \models & \Box^{-1}\phi, \text{if } (\forall d \leq e, d \neq e)\ P, d \models \phi \end{array}$$

Define $P \models \phi$ if $(\forall e \in E)\ P, e \models \phi$.
For a set of rf-pomsets $\mathcal{P}$, define $\mathcal{P} \models \phi$ if $(\forall P \in \mathcal{P})\ P \models \phi$.

The definition of satisfaction of formulas by pomsets validates the rule: if $P \models \phi$ then $P \models \phi \wedge \Box^{-1}\phi$.

While the boolean connectives are interpreted as expected, the strict past operators explore the past of the current event as determined by the pomset ordering. Their interpretation does not include the current instant. Thus, while they are connected by a DeMorgan law $\Box^{-1}\phi = \neg\Diamond^{-1}(\neg\phi)$ they do *n*ot satisfy the rule $\Box^{-1}\phi \Rightarrow \Diamond^{-1}\phi$. However, they do satisfy:

Coinduction. $(\phi \Rightarrow \Diamond^{-1}\phi) \Rightarrow \neg\phi$
Induction. $(\Box^{-1}\phi \Rightarrow \phi) \Rightarrow \phi$

We next state a composition result in the style of Abadi and Lamport [1]. Our statement and proof is intentionally less general to simplify the presentation. We view the composition result as capturing key aspects of no-ThinAirRead, as will become clearer in the examples below.

The statement of the theorem requires us to incorporate environment assumptions in the satisfaction relation.

Definition 5.2. Let $\phi \in \mathtt{PLTL}$. Then, define:

$$\mathtt{Models}(\phi) = \{\mathcal{P} \mid \mathcal{P} \models \phi\}$$

$\mathtt{Models}(\phi)$ are the models of the formula ϕ, i.e. the pomsets that satisfy the formula. We say that ϕ is prefix closed if $\mathtt{Models}(\phi)$ is prefix-closed[2].

Let $\phi \in \mathtt{PLTL}$ be prefix-closed. Let $\psi \in \mathtt{PLTL}$. Then define $\phi, \mathcal{P} \models \psi$ if $\mathtt{Models}(\phi) \parallel \mathcal{P} \models \psi$.

We are now ready to state the composition theorem. In the vocabulary of Abadi and Lamport [1], we are in the special case of invariants without environment assumptions.

[1] Our presentation differs from standard presentations of past LTLLichtenstein et al. [9] in two ways. First, we eschew the previous instant operator, to account for the current setting of partial orders; second, we consider strictly past versions of the "once" and "always in the past" operators, by not evaluating the formula at the current instant.
[2] P_1 is a prefix of P if the carrier set of P_1 is a downwards closed subset of P; i.e if $e \in P_1$ and $d \leq_P e$, then d also in P_1.

Lemma 5.3 (Composition). *Let $\phi \in \mathsf{PLTL}$ be prefix-closed. Let $\mathcal{P}_1, \mathcal{P}_2$ be augmentation-closed*[3]. *Then:*

$$\frac{\phi, \mathcal{P}_1 \models \phi \quad \phi, \mathcal{P}_2 \models \phi}{\mathcal{P}_1 \parallel \mathcal{P}_2 \models \phi}$$

Sketch. We will show that all prefixes in the prefix closures of $\mathcal{P}_1 \parallel \mathcal{P}_2$ satisfy the required property. Proof proceeds by induction on prefixes of $P \in \mathcal{P}_1 \parallel \mathcal{P}_2$.

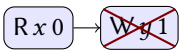
The case for empty prefix follows from assumption that ϕ is prefix closed.

For the inductive case, consider P in the prefix closure of $\mathcal{P}_1 \parallel \mathcal{P}_2$, i.e. $P = P_1 \parallel P_2$ where $P_i \in \mathcal{P}_i$. Since $\mathcal{P}_1$ and $\mathcal{P}_2$ are augmentation closed, we can assume that the restriction of P to the events of P_i coincides with P_i, for $i = 1, 2$.

Consider a prefix (say P') got by deleting a maximal element, say e, of P. There are two cases depending on whether e comes from P_1 or P_2. In the case when e comes from P_1, since P_2 is a prefix of P' and $P' \models \phi$ by induction hypothesis, we deduce that $P_2 \models \phi$. Thus, $P_2 \in \mathtt{Models}(\phi)$. Since $P_1 \in \mathcal{P}_1$, assumption $\phi, \mathcal{P}_1 \models \phi$, we deduce that $P_1 \parallel P_2 \models \phi$. □

6 CONCLUSIONS AND FUTURE WORK

One oddity of the model is that $[\![r := x;\ y := r;\ \mathtt{skip}]\!]$ includes:



where the write action guessed its value incorrectly, and therefore has precondition $0 = 1$. This form of speculative execution does not appear to be used in practice. In order to disallow it, one could change the semantics of `skip` to introduce a tick action denoting successful completion of the thread and only consider executions in which the precondition of every tick action is satisfiable. We leave the elaboration of this idea as future work.

{Comment on the following:}

- coherence = per location total order on $\not<$
- Validation of write removal requires some tricks to ensure that thread does not rf its own write
- Definition of rf can use $\not<$ in first clause, rather than $\leq$. We chose the stronger definition because it makes some of the examples simpler: in particular, the example motivating rf needs a volatile in the other thread if you don't have rf-implies-hb. Old text: The notion rf-pomset is sufficient to capture hardware models and release/acquire access in C++, where reads-from implies happens-before [2]. To model C++ relaxed access, it would be necessary to use a more general notion of rf-pomset, where $(d, x, e) \in \mathsf{RF}$ does not necessarily imply $d < e$, instead requiring that $(< \cup \mathsf{RF})$ be acyclic.
- The design space for transactions is very rich [6]. We have only presented one option. we pun between abort and false commit
- causality test cases 1, 6, 8, 9, 18, 20 (12?) require that the logic make assertions about the domain of variables

REFERENCES

[1] Martín Abadi and Leslie Lamport. 1993. Composing Specifications. *ACM Trans. Program. Lang. Syst.* 15, 1 (Jan. 1993), 73–132. https://doi.org/10.1145/151646.151649

[2] Jade Alglave, Luc Maranget, and Michael Tautschnig. 2014. Herding Cats: Modelling, Simulation, Testing, and Data Mining for Weak Memory. *ACM Trans. Program. Lang. Syst.* 36, 2, Article 7 (July 2014), 74 pages. https://doi.org/10.1145/2627752

[3] P_1 is a augmentation of P if their carrier sets are the same and if $d \leq_P e$, then $d \leq e$ also in P_1.

[3] Hans-J. Boehm and Sarita V. Adve. 2008. Foundations of the C++ Concurrency Memory Model. In *Proceedings of the 29th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '08)*. ACM, New York, NY, USA, 68–78. https://doi.org/10.1145/1375581.1375591

[4] Nathan Chong, Tyler Sorensen, and John Wickerson. 2018. The semantics of transactions and weak memory in x86, Power, ARM, and C++. In *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2018, Philadelphia, PA, USA, June 18-22, 2018*. 211–225. https://doi.org/10.1145/3192366.3192373

[5] Craig Disselkoen, David Kohlbrenner, Leo Porter, and Dean M. Tullsen. 2017. Prime+Abort: A Timer-Free High-Precision L3 Cache Attack using Intel TSX. In *26th USENIX Security Symposium, USENIX Security 2017, Vancouver, BC, Canada, August 16-18, 2017.*, Engin Kirda and Thomas Ristenpart (Eds.). USENIX Association, 51–67. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/disselkoen

[6] Brijesh Dongol, Radha Jagadeesan, and James Riely. 2018. Transactions in relaxed memory architectures. *PACMPL* 2, POPL (2018), 18:1–18:29. https://doi.org/10.1145/3158106

[7] James W. Gray, III. 1992. Toward a Mathematical Foundation for Information Flow Security. *J. Comput. Secur.* 1, 3-4 (May 1992), 255–294. http://dl.acm.org/citation.cfm?id=2699806.2699811

[8] Paul Kocher, Daniel Genkin, Daniel Gruss, Werner Haas, Mike Hamburg, Moritz Lipp, Stefan Mangard, Thomas Prescher, Michael Schwarz, and Yuval Yarom. 2018. Spectre Attacks: Exploiting Speculative Execution. *CoRR* abs/1801.01203 (2018). arXiv:1801.01203 http://arxiv.org/abs/1801.01203

[9] Orna Lichtenstein, Amir Pnueli, and Lenore D. Zuck. 1985. The Glory of the Past. In *Proceedings of the Conference on Logic of Programs*. Springer-Verlag, London, UK, UK, 196–218. http://dl.acm.org/citation.cfm?id=648065.747612

[10] Jeremy Manson, William Pugh, and Sarita V. Adve. 2005. The Java Memory Model. *SIGPLAN Not.* 40, 1 (Jan. 2005), 378–391. https://doi.org/10.1145/1047659.1040336

[11] Andrew C. Myers. 1999. JFlow: practical mostly-static information flow control. In *26th ACM Symp. on Principles of Programming Languages (POPL)*. 228âĂŞ241. http://www.cs.cornell.edu/andru/papers/popl99/popl99.pdf

[12] Kevin R. O'Neill, Michael R. Clarkson, and Stephen Chong. 2006. Information-Flow Security for Interactive Programs. In *Proceedings of the 19th IEEE Workshop on Computer Security Foundations (CSFW '06)*. IEEE Computer Society, Washington, DC, USA, 190–201. https://doi.org/10.1109/CSFW.2006.16

[13] Gordon Plotkin and Vaughan Pratt. 1997. Teams Can See Pomsets (Preliminary Version). In *Proceedings of the DIMACS Workshop on Partial Order Methods in Verification (POMIV '96)*. AMS Press, Inc., New York, NY, USA, 117–128. http://dl.acm.org/citation.cfm?id=266557.266600

[14] A. Sabelfeld and A. C. Myers. 2006. Language-based Information-flow Security. *IEEE J.Sel. A. Commun.* 21, 1 (Sept. 2006), 5–19. https://doi.org/10.1109/JSAC.2002.806121

[15] Geoffrey Smith and Dennis Volpano. 1998. Secure Information Flow in a Multi-threaded Imperative Language. In *Proceedings of the 25th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '98)*. ACM, New York, NY, USA, 355–364. https://doi.org/10.1145/268946.268975

[16] J. Todd Wittbold and Dale M. Johnson. 1990. Information Flow in Nondeterministic Systems. In *IEEE Symposium on Security and Privacy*.

[17] Danfeng Zhang, Aslan Askarov, and Andrew C. Myers. 2012. Language-based Control and Mitigation of Timing Channels. *SIGPLAN Not.* 47, 6 (June 2012), 99–110. https://doi.org/10.1145/2345156.2254078

[18] Jianzhou Zhao, Santosh Nagarakatte, Milo M. K. Martin, and Steve Zdancewic. 2012. Formalizing the LLVM intermediate representation for verified program transformations. In *Proceedings of the 39th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL 2012, Philadelphia, Pennsylvania, USA, January 22-28, 2012*, John Field and Michael Hicks (Eds.). ACM, 427–440. https://doi.org/10.1145/2103656.2103709