Addressing Covert Termination and Timing Channels in Concurrent Information Flow Systems

Deian Stefan1  Alejandro Russo2  Pablo Buiras2  Amit Levy1  John C. Mitchell1  David Mazières1
(1) Stanford University, Stanford, CA, USA  (2) Chalmers University of Technology, Gothenburg, Sweden

Abstract
When termination of a program is observable by an adversary, confidential information may be leaked by terminating accordingly. While this termination covert channel has limited bandwidth for sequential programs, it is a more dangerous source of information leakage in concurrent settings. We address concurrent termination and timing channels by presenting a dynamic information-flow control system that mitigates and eliminates these channels while allowing termination and timing to depend on secret values. Intuitively, we leverage concurrency by placing such potentially sensitive actions in separate threads. While termination and timing of these threads may expose secret values, our system requires any thread observing these properties to raise its information-flow label accordingly, preventing leaks to lower-labeled contexts. We implement this approach in a Haskell library and demonstrate its applicability by building a web server that uses information-flow control to restrict untrusted web applications.

Categories and Subject Descriptors D.1.3 [Programming Techniques]: Concurrent Programming; D.6.5 [Security and Protection]: Information flow controls

General Terms Security, Languages, Design

Keywords Monad, Library, Timing channels

1. Introduction
Covert channels arise when programming language features are misused to leak information [30]. For example, when termination of a program is observable to an adversary, a program may intentionally or accidentally communicate a confidential bit by terminating according to the value of that bit. While this termination covert channel has limited bandwidth for sequential programs, it is a significant source of information leakage in concurrent settings. Similar issues arise with covert timing channels, which are potentially widespread because so many programs involve loops or recursive functions. These channels, based on either internal observation by portions of the system or external observation, are also effective in concurrent settings.

We present an information-flow system, in a form of an execution monitor, that mitigates and eliminates termination and timing channels in concurrent systems, while allowing timing and termination of loops and recursion to depend on secret values. Because the significance of these covert channels depends on concurrency, we fight fire with fire by leveraging concurrency to mitigate these channels: we place potentially non-terminating actions, or actions whose timing may depend on secret values, in separate threads. In our system, each thread has an associated current label that keeps track of the sensitivity of the data it has observed and restricts the locations to which the thread can write. Hence, while termination and timing of these threads may expose secret values, our system requires any thread observing these properties to raise its information-flow label accordingly. This prevents lower security threads from observing confidential information. We implement this approach in a Haskell library and demonstrate its applicability by building a web server that applies information-flow control to untrusted web applications. One advantage of Haskell is that the Haskell type system prevents code from circumventing our dynamic information-flow tracking. Although we do not address underlying hardware issues such as cache timing, our language-level methods can be combined with hardware-level mechanisms as needed to provide comprehensive defenses against covert channels.

Termination covert channel Askarov et al. [2] show that, for sequential programs with outputs, leakage using the termination covert channel requires exponential time in the size of the secret. Moreover, if secrets are uniformly distributed, the attacker’s advantage (after observing a polynomial amount of output) is negligible in comparison with the size of the secret. Because of this relatively low risk, accepted sequential information-flow tools, such as Jif [38] and FlowCaml [48], are only designed to address termination-insensitive noninterference. In a concurrent setting, however, the termination covert channel may be exploited more significantly [18]. We therefore focus on termination covert channels in concurrent programs and present an extension to our Haskell LIO library [52], which provides dynamic tracking of labeled values. By providing labeled explicit futures with 1Fork and 1Wait, our extension removes the termination covert channel from sequential and concurrent programs while allowing loops whose termination conditions depend on secret information.

Internal timing channel Multi-threaded programs can leak information through an internal timing covert channel [56] when the order of public events depends on secret data. Generally, an internal timing attack can be carried out whenever a race to acquire a shared public resource may be affected by secrets. We close this covert channel by decoupling the execution of computations that produce public events from computations that manipulate secret data. Using 1Fork and 1Wait, a computation depending on secret data proceeds in a new thread; LIO asserts that the number of instructions executed by threads producing public events does not depend on secrets. Therefore, a possible race to a shared public resource does not depend on the secret, eliminating internal timing leaks.

Permission to make digital or hard copies of all or part 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. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICFP’12, September 9-15, 2012, Copenhagen, Denmark.
Copyright © 2012 ACM … $10.00
2. Background

We build on the Haskell dynamic information flow control library called LIO \cite{52}. This section describes LIO and some of its relevant background.

External timing channel

External timing covert channels, which involve externally measuring the time used to complete operations that may depend on secret information, have been used in practice to leak information \cite{6, 15} and break cryptosystems \cite{17, 28, 57}. While several existing mechanisms mitigate external timing channels \cite{1, 4, 19}, these covert channels are not addressed by conventional information-flow tools and, in fact, most of the previous techniques for language-based information-flow control appear to have limited application. Our contribution to external timing channels is to bring the mitigation techniques from the OS community into the language-based security setting. Generalizing previous work \cite{3}, Zhang et al. \cite{60} propose a black-box mitigation technique that we adapt to a language-based security setting. In this approach, the source of observable events is wrapped by a timing mitigator that delays output events so that they contain only a bounded amount of information. We take advantage of Haskell’s ability to identify computations that produce outputs, and implement the mitigator as part of our information flow control library. Leveraging monad transformers \cite{33}, we show how to modularly extend LIO and other Haskell libraries that perform side-effects, to provide a suitable form of Zhang et al.’s mitigator.

In summary, the main contributions of this paper are:

- We present a dynamic information flow control (IFC) system that eliminates the termination and internal timing covert channels, while mitigating external timing. While these covert channels have been addressed differently in static IFC approaches, this is the first implementation of a language-based dynamic IFC system for concurrency that does not rely on cooperative scheduling. Our system provides support for threads, lightweight synchronization primitives, and allows loops and branches to depend on sensitive values.
- We eliminate termination and internal-timing covert channels using concurrency, with potentially sensitive actions run in separate threads. This is implemented in a Haskell library that uses labeled concurrency primitives\(^1\).
- We provide language-based support for resource-usage mitigation using monad transformers. We use this method to implement the black-box external timing mitigation approach of Zhang et al.; the method is also applicable to other covert channels, such as storage.
- We evaluate the language implementation by building a simple server-side web application framework. In this framework, untrusted applications have access to a persistent key-value store. Moreover, requests to apps may be from malicious clients colluding with the application in order to learn sensitive information. We show several potential leaks through timing and termination and show how our library is used to address them.

Section 2 provides background on information flow, Haskell, and the LIO monad. We describe how to address the termination covert channel in Section 3, the internal timing covert channel in Section 4, and the external timing channel in Section 5. Labeled communication primitives are detailed in Section 6. A formalization of the library is given in Section 7, with security guarantees detailed in Section 8. The implementation and experimental evaluation are presented in Section 9. Related work is described in Section 10. We conclude in Section 11.

2.1 Information flow control

The goal of IFC is to track and control the propagation of information. In an IFC system, every bit has an associated label. Labels form a lattice \cite{9} governed by a partial order \(\subseteq\) pronounced “can flow to.” The value of a bit labeled \(L_{out}\) can depend on a bit labeled \(L_{in}\) only if \(L_{in} \subseteq L_{out}\).

In a floating-label system, such as LIO, every execution context has a label that can rise to accommodate reading more sensitive data. For a computation \(P\) labeled \(L_{P}\) to observe an object labeled \(L_{O}\), \(P\)'s label must rise to the least upper bound or \(\cap\) of the two labels, written \(L_{P} \cup L_{O}\). \(P\)'s label effectively “floats above” the labels of all objects it observes. Furthermore, systems frequently associate a clearance with each execution context that bounds its label.

Specific label formats depend on the application and are not the focus of this work. Instead, we will focus on a very simple two-point lattice with labels Low and High, where Low \(\subseteq\) High and High \(\not\subseteq\) Low. We, however, note that our implementation is polymorphic in the label type and any label format that implements a few basic relations (e.g., \(\subseteq\), join \(\cup\), and meet \(\cap\)) can be used when building applications. Additionally, the LIO library supports privileges which are used to implement decentralized information flow control as originally presented in \cite{37}; though we do not discuss privileges in this paper, our implementation also provides privileged-aware versions of the combinators described in later sections.

2.2 Haskell

We chose Haskell because its abstractions allow IFC to be provided by a library \cite{31}. Building a library is far simpler than developing a programming language from scratch (or heavily modifying a compiler). Moreover, a library offers backwards compatibility with a large body of existing Haskell code.

From a security point of view, Haskell’s most distinctive feature is a clear separation of pure computations from those with side-effects. Any computation with side-effects must have a type encapsulated by the monad \texttt{IO}. The main idea behind the LIO library is that untrusted actions must be specified with a new \texttt{Labeled I/O} monad, \texttt{LIO}, instead of \texttt{IO}. Using \texttt{LIO} ensures that all computations obey information flow control flow.

2.3 The LIO monad

LIO dynamically enforces IFC, but without the features described in this paper, provides only termination-insensitive IFC \cite{2} for sequential programs. At a high level, LIO provides the \texttt{LIO} monad intended to be used in place of \texttt{IO}. The library furthermore contains a collection of \texttt{LIO} actions, many of them similar to \texttt{IO} actions from standard Haskell libraries, except that the \texttt{LIO} versions contain label checks that enforce IFC. For instance, \texttt{LIO} provides file operations that look like those of the standard library, except that they confine the application to a dedicated portion of the file system where they store a label along with each file.

The \texttt{LIO} monad keeps a current label, \texttt{Lcut}, that is effectively a ceiling over the labels of all data that the current computation may depend on. \texttt{LIO} also maintains a current clearance, \texttt{Cut}, which specifies an upper bound on permissible values of \texttt{Lcut}.

\texttt{LIO} does not individually label definitions and bindings. Rather, all symbols in scope are identically labeled with \texttt{Lcut}. The only way to observe or modify differently labeled data is to execute actions that internally access privileged symbols. Such actions are responsible for appropriately validating and adjusting the current label. As an example, the \texttt{LIO} file-reading function \texttt{readFile}, when executed on a file labeled \texttt{LF}, first checks that \texttt{LF} \(\subseteq\) \texttt{Cut}, throwing an exception if not. If the check succeeds, the function raises \texttt{Lcut}.

\footnote{The library implementations discussed in this paper can be found at \url{http://www.scs.stanford.edu/~deian/concurrent_lio}}
to $L_{\text{out}} \sqcup L_F$ before returning the file content. The LIO file-writing function, `writeFile`, throws an exception if $L_{\text{out}} \neq L_F$.

As previously mentioned, allowing experimentation with different label formats, LIO actions are parameterized by the label type. For instance, simplifying slightly:

```haskell
readFile :: (Label l) -> FilePath -> LIO 1 String
```

To be more precise, it is really $(LIO l)$ that is a replacement for the $10$ monad, where $1$ can be any label type. The context $(Label 1)$ in `readFile`'s type signature restricts $1$ to types that are instances of the Label typeclass, which abstracts the label specifics behind the basic methods $\in$, $\cup$, and $\sqcap$.

### 2.4 Labeled values

Since LIO protects all nameable values with $L_{\text{out}}$, we need a way to manipulate differently-labeled data without monotonically increasing $L_{\text{out}}$. For this purpose, LIO provides explicit references to labeled, immutable data through a polymorphic data type called Labeled. A locally accessible symbol (at $L_{\text{out}}$) can name, say, a Labeled $\text{Int}$ (for some label type $l$), which contains an Int protected by a different label.

Several functions allow creating and using Labeled values:

- `label :: (Label l) -> l -> LIO l (Labeled l a)`
  - Given label $l$, construct a Labeled value $l a$ with label $l$.

- `unlabel :: (Label l) -> Labeled l a -> LIO l a`
  - If $l$ is a labeled value $l a$ with label $l$, `unlabel` $l a$ returns $l$.

- `toLabeled :: (Label l) -> 1 -> LIO 1 a -> LIO l 1 (Labeled l a)`
  - The dual of `unlabel`: given a label $l$ and an action $a$ that would raise $L_{\text{out}}$ to $L'_{\text{out}}$ where $L'_{\text{out}} \subseteq l \subseteq C_{\text{out}}$, `toLabeled` $l a$ executes $a$ without raising $L_{\text{out}}$, and instead encapsulates $a$'s result in a Labeled value protected by label $l$.

- `labelOf :: (Label l) -> Labeled l a -> l`
  - Returns the label of a Labeled value.

As an example, we show an LIO action that adds two Labeled Ints:

```haskell
addLIO 1A 1B = do a <- unlabel 1A b <- unlabel 1B return (a + b)
```

If the inputs' labels are $L_A$ and $L_B$, this action raises $L_{\text{out}}$ to $L_{\text{out}} \sqcup L_A \sqcup L_B$ and returns the sum of the values. To avoid raising the current label, and instead return a Labeled Int, `addLIO` can be wrapped by `toLabeled`:

```haskell
addLIO 1A 1B = toLabeled (labelOf 1A \u2192 labelOf 1B) (addLIO 1A 1B)
```

### Implicit flows in LIO

We note that in an imperative language with labeled variables, dynamic labels can lead to implicit flows [10].

The canonical example is as follows:

```haskell
public :: Int
if (secret) then return (public)
else return (public + 1)
```

To avoid directly leaking the `secret` bit into `public`, one should track the label of the program counter and determine that execution of the assignment `public := 1` depends on `secret`, and raise `public`'s label when assigning `public := 1`. However, since the assignment executes conditionally depending on `secret`, now `public`'s `label` leaks the `secret` bit. LIO does not suffer from implicit flows. When branching on a `secret`, $L_{\text{out}}$ becomes `High` and therefore no public events are possible.

### 3. The termination covert channel

As mentioned in the introduction, information-flow control results and techniques for sequential settings do not naturally generalize to concurrent settings. In this section we highlight that the sequential LIO library is, like many IFC systems, susceptible to leaks due to termination and show that a naive, but typical, extension that adds concurrency drastically amplifies this leak.

We present a modification to the LIO library that eliminates the termination covert channel from both sequential and concurrent programs; our solution allows for flexible programming patterns, even writing loops whose termination condition depends on secret data.

#### Sequential setting

As described by [2], a brute-force attack, taking exponential time in the size (# of bits) of the secret, is the most effective way to exploit the termination channel in a sequential program. Listing 1 shows an implementation of such attack. Function `bruteForce` takes three arguments: a public string (helper) message, a public upper-bound on the secret, and the secret Int of type Labeled Int. Given the three arguments `bruteForce` returns an LIO action which when executed returns unit, but produces intermediate side-effects. Specifically, `bruteForce` writes to a Low labeled channel using `outputLow` while $L_{\text{out}}$ is Low. We assume that the attack is executed with initial $L_{\text{out}} = \text{Low}$, and secret Int labeled `High`.

The attack consists of iterating (variable $i$) over the domain of the secret ($\text{forM}_{\[0..n\]}$), producing a publicly-observable output at every iteration until the secret is guessed. On every iteration $L_{\text{out}}$ is raised to the label of the secret within a `toLabeled` block. As described in Section 2.4, the current label outside the `toLabeled` block remains unaffected, and so the computation can continue producing publicly-observable outputs if the computation within the `toLabeled` block does not diverge. This is the case unless guess $i$ is equal to the secret, at which point the computation diverges (when $(s = i)$) and no additional publicly-observable outputs are produced. The leak due to termination is obvious: when the attacker, observing the Low labeled output channel, no longer receives any data, the value of the secret can be inferred given the previous outputs. For instance, to leak a 16-bit `secret`, we execute `bruteForce "secret" 65536 secret`. Assuming the value of the secret is 3, executing the action produces the outputs "secret = 0", "secret = 1", and "secret = 2" before diverging. The assumption here, and the rest of the paper, is that the code is untrusted (e.g., provided by the attacker) and therefore an observer that knows the implementation of `bruteForce` can directly infer that the value of the secret is 3. Observe that the code producing public outputs (`outputLow (mag + * a + show i)` does not inspect secret data, which makes it difficult to avoid termination leaks by simply tracking the flow of labeled data inside programs.

#### Concurrent setting

Suppose that we (naively) add support for concurrency to LIO using a hypothetical primitive `fork`, which sim-
ply executes a given computation in a new thread. Although we can preserve termination-insensitive non-interference, we can extend the previous brute force attack to leak a secret value in linear, as opposed to exponential, time. In general, adding concurrency primitives in a straight-forward manner makes attacks that leverage the termination covert channel very effective [18]. To illustrate this point, consider the attack shown in Listing 2, which leaks the bit-contents of a secret value in linear time. Given the bit-length $k$ of the secret and the labeled secret, concurrentAttack returns an action which, when executed, extracts each bit of the secret (extractBit $i$ $a$) and spawns a corresponding thread to recover the bit using the sequential brute-force attack of Listing 1 (bruteForce ($i$ $a$)). By collecting the public outputs generated by the different threads (having the form "0-bit $≠$ 0", "3-bit $≠$ 0", "1-bit $≠$ 0", etc.), it is directly possible to recover the secret value. Observe that the divergence of one thread does not affect the termination of other threads and thus does not require observations external to the program, as in the sequential case.

### 3.1 Removing the termination covert channel in LIO

Since LIO has floating labels, a leak to a Low channel due to termination cannot occur after the current label is raised to High, unless the label raise is within an enclosed toLabeled computation. Hence, we can deduce that a piece of LIO code can exploit the termination covert channel only when using toLabeled. The key insight is that toLabeled is the single LIO combinator that effectively allows a piece of code to temporarily raise its current label, perform a computation, and then continue with the original current label. The attack in Listing 1 is a clear example that leverages this property of toLabeled to leak information.

Consider the necessary conditions for eliminating the termination channel of Listing 1: the execution of the publicly-observable outputLow action must not depend on the data or control flow of the secret computation executed within the toLabeled block. Hence, one approach to close the termination covert channel is by decoupling the execution of computations enclosed by toLabeled. To this end, we provide an alternative to toLabeled that executes computations that might raise the current label (as in toLabeled) in a newly-spawned thread. To observe the result (or non-termination) of such a spawned computation, the current label is firstly raised to the label of the (possibly) returned result. In doing so, after observing a secret result (or non-termination) of a spawned computation, actions that produce publicly-observable side-effects can no longer be executed. In this manner, the termination channel is closed.

In Listing 1, the execution of outputLow depends on the termination of the computation enclosed by toLabeled. However, using our proposed approach of spawning a new thread when performing a toLabeled, if the code following the sensitive block wishes to observe whether or not the High computation has terminated, it would first need to raise the current label to High. Thereafter, an outputLow action cannot be executed regardless of the result (or termination) of the toLabeled computation.

Concretely, we close the termination channel by removing the insecure function toLabeled from LIO and, instead, provide the following (termination sensitive) primitives.

```agda
1Fork :: Label _ = 1 → LIO 1 a → LIO 1 (Result 1 a)
1Wait :: Label _ = Result 1 a → LIO 1 a
```

Intuitively, 1Fork can be considered as a concurrent version of toLabeled. 1Fork 1 1o spawns a new thread to perform the computation 1io, whose current label may rise, and whose result is a value labeled with 1. Rather than block, immediately after spawning a new thread, the primitive returns a value of type Result 1 a, which is simply a handler to access the labeled result produced by the spawned computation. Similar to unlabeled, we provide 1Wait.

**Listing 3 Internal timing leak**

```agda
doGuess secret guess cond = do
tolabeled High $ do v ← unlabeled secret 
  when (v = guess) $ loopUntil cond 
outputLow (show guess) 
broadcastCondition cond
```

attack = Labeled 1 Bool → LIO 1 ()
attack secret = do cond ← mkSharedCond
  (- thread 1: -) fork $ doGuess secret True cond
  (- thread 2: -) fork $ doGuess secret False cond
```

which inspects values returned by spawned computations, i.e., values of type Result 1 a. The labeled wait, 1Wait, raises the current label to the label of its argument and then proceeds to inspect it. In principle, rather than forking threads, it would be enough to prove that computations involving secrets terminate, e.g., by writing them in Coq or Agda. However, while this idea works in theory, it is still possible to crash an Agda or Coq program at runtime: for example, with a stack overflow. Generally, abnormal termination due to resource exhaustion exploits the termination channel just as effectively. Forking threads removes the termination channel by design. Although it might seem expensive, forking threads in Haskell is a light-weight operation [25].

We note that adding concurrency to LIO is a major modification which introduces security implications beyond that of handling the termination channel. In the following section, we describe the internal timing covert channel, a channel that is only present in programming languages that have support for concurrency and shared-resources.

### 4. The Internal timing covert channel

Multi-threaded programs, wherein threads share a common (public) resource, can leak sensitive information through the internal timing covert channel [56]. In an internal timing attack, sensitive data can be leaked by affecting the timing behavior of threads, which consequently alters the order of events on a shared public channel.

Listing 3 shows an example of the internal timing attack in LIO with the added fork primitive. Here, action mkSharedCond creates a shared resource that is used as a “condition variable”, while loopUntil waits until the condition is satisfied and broadcastCondition sets the condition to signal all the waiting threads. The shared resource may be defined in terms of language level constructs, such as mutable references: mkSharedCond can create a public Bool reference set initially to False, loopUntil loops until the dereferenced value is True and broadcastCondition assigns True to it. Alternatively, it can be defined in terms of an implicit state, such as the scheduler: mkSharedCond and broadcastCondition can simply return False, while loopUntil delays the running thread for a reasonable amount of time (e.g., by using threadDelay).

Although in isolation both threads are secure (i.e., they satisfy non-interference), by executing them concurrently it is possible to leak information about secret. When executing attack secret, if secret is (labeled) True, thread 2 will output to the public (low) channel after thread 1 regardless of which thread is executed first. In other words, the produced output will be “True”, “False”. The converse holds when secret is False and the program prints out “False”, “True”. Notice that an attacker can infer the value of secret by simply observing the outputs on the public channel: the order of True and False is influenced by the secret data.

Unlike other timing channel attacks, internal timing attacks do not require a powerful attacker that must measure the execution time as to deduce secret information. The interleaving of threads can directly be used to produce leaks! Additionally, we note that
although the example of Listing 3 only leaks a single bit, it is easy to construct an attack that uses a loop to leak the bit-contents of a secret value in linear time of its length. Tsai et al. [54] show such an amplified attack and demonstrate its effectiveness even in settings where little information about the run-time system (e.g., the scheduler) is available.

4.1 Removing the internal timing channel

As indicated by our example, the internal timing covert channel can be exploited when the time to produce public events (e.g., writing data to a public channel) depends on secrets. In other words, leaks due to internal timing occur when there is a race to acquire a public shared resource that may be affected by secret data. To close this channel, we apply the technique used to close the termination covert channel: we decouple the execution of computations that produce public events from computations that manipulate secret data. By using fork and monad transformers, computations dealing with secrets are executed in a new thread. Consequently, any possible race to a shared public resource cannot depend on sensitive data, making leaks due to internal timing infrasible.

5. The external timing covert channel

In a real-world scenario, IFC applications interact with unlabeled, publicly observable, resources. For example, a server-side IFC web application interacts with a browser, which may itself be IFC-unaware, over a public network channel. Consequently, an adversary can take measurements external to the application (e.g., the application response time) from which they may infer information about confidential data computed by the web application. Although our results generalize (e.g., to the storage covert channel), in this section we address the external timing covert channel: an application can leak information over a channel to an observer that precisely measures message-arrival timings.

Most of the language-based IFC techniques that consider external timing channels are limited. Despite the successful use of external timing attacks to leak information in web [6, 15] and cryptographic [17, 28, 57] applications, they remain widely unaddressed by mainstream, practical IFC tools, including JIf [38]. Furthermore, the most techniques that provide IFC in the presence of the external timing channel [1, 4, 19] are overly restrictive, e.g., they do not allow folding over secret data. In this work, we show a modular approach of mitigating the external timing covert channel for Haskell libraries, such as LIO.

5.1 Mitigating the external timing channel

Recently, a predictive black-box mitigation technique for external timing channels has been proposed [3, 60]. The predictive mitigation technique assumes that the attacker has control of the application (computing on secret data) and can measure the time a message is placed on a channel (e.g., when a response is sent to the browser). Treating the application as a black-box event source, a mitigator is interposed between the application and system output.

Internally, the mitigator keeps a schedule describing when outputs are to be produced. For example, the time mitigator might keep a schedule “predicting” that the application will produce an output every 1 ms. If the application delivers events according to the schedule, or at a higher rate, the mitigator will be able to produce an output at every 1 ms interval, according to the schedule, and thus leak no information.

The application may fail to deliver an event to the mitigator on time, and thus render the mitigator’s schedule prediction false. At this point, the mitigator must handle the misprediction by selecting, or “predicting”, a new schedule for the application. In most cases, this corresponds to doubling the application’s quantum. For instance, following a misprediction where the quantum was 1 ms, the application will subsequently be expected to produce an output every 2 ms. It is at the point of switching schedules where an attacker learns information: rather than seeing events spaced at 1 ms intervals, the attacker now observes outputs at 2 ms intervals, indicating that the application violated the predicted behavior (a decision that can be affected by secret data). Askarov et al. [3] show that the amount of information leaked by this slow-doubling mitigator is polylogarithmic in the application runtime.

The aspects of the predictive mitigation technique of [3, 60] that make it particularly attractive to use in LIO are:

- The mitigator can adaptively reduce the quantum, as to increase the throughput of a well-behaved application in a manner that bounds the covert channel bandwidth (though with the leakage factor slightly larger than that of the slow-doubling mitigator);
- The mitigator can leverage public factors to decide a schedule. For example, in a web application setting where responses are mitigated, the arrival of an HTTP request can be used as a “reset” event. This is particularly useful as a quiescent application would otherwise be penalized for not producing an output according to the predicted schedule. Our web application of Section 9 implements this mitigation technique.
- The amount of information leaked is bounded by a combinatorial analysis on the number of attacker observations.

Monadic approach to black-box mitigation

The functionality of different monads, such as I/O and error handling, can be combined in a modular fashion using monad transformers [33]. A monad transformer $t$, when applied to a monad $m$, generates a new, combined monad, $t m$, that shares the behavior of monad $m$ as well as the behavior of the monad encoded in the monad transformer. The modularity of monad transformers comes from the fact that they consider the underlying monad $m$ opaque, i.e., the behavior of the monad transformer $t$ does not depend on the internal structure of $m$. In this light, we adopt Zhang et al.’s system-oriented predictive black-box mitigator to a language-based security setting in the form of a monad transformer.

5.2 Language-based mitigators

We envision the implementation of mitigators that address covert channels other than external timing. For example, we prototype a mitigator for the storage covert channel, which addresses attacks in which the message length is used to encode secret information. Hence, our mitigation monad transformer $\mathrm{MitM} \ a \ s$ is polymorphic in the mitigator-specific state $s$ and quantum type $q$:

$$\text{newtype } \mathrm{MitM} \ a \ s \ q \ m \ a = \mathrm{MitM} \ldots$$

We provide the function $\text{evalMitM}$, which takes an action of type $\mathrm{MitM} \ a \ s \ q \ m \ a$ and returns an action of type $\mathrm{MitM} \ a \ s \ q \ m \ a$, which when executed will mitigate the computation outputs. We note that the value constructor for the mitigation monad must not be exported to untrusted code, which can use it to circumvent the mitigation.

The time-mitigation monad transformer is a special case:

$$\text{type } \mathrm{TimeMitM} = \mathrm{MitM} \mathrm{Timestamp} \mathrm{TimestampDiff}$$

where the internal state $\mathrm{Timestamp}$ is a time stamp, and the quantum $\mathrm{TimestampDiff}$ is a time difference. Superficially, a value of type $\mathrm{TimeMitM} \ m \ a$ is a monadic computation that produces a value of type $a$. Internally, a time measurement is taken whenever an output is to be emitted in the underlying monad $m$, the internal state and quantum are updated to reflect the output event, and the output is delayed if it was produced ahead of the predicted schedule. Consider, for instance, a version of $\text{hPut}$ executing in the time mitigated IO monad, where every handle is mitigated:

$$\text{type } \mathrm{MIO} = \mathrm{TimeMitM} \ IO$$

$$\text{hPut} :: \mathrm{Handle} \rightarrow \mathrm{ByteString} \rightarrow \mathrm{MIO} ()$$
The context MonadConcur m is used to impose the requirement that the underlying monad be an I/O-like monad which allows forging new I/O threads (as to separate the mitigator from the computation being mitigated) and operations on synchronizing variables, MVars [25] (which are internal to the hPut transformer). The mitMitigated function is used to create a mitigated handle given an initial state, quantum, and underlying constructor. The default implementation of mitMitigated creates the mitigator state (internal to the transformer) corresponding to the handle. A simplified version of our openFile operation shows how mitMitigated is used:

```haskell
openFile :: FilePath -> IO Mode -> MIO Handle
openFile f mode = mitMitigated Nothing q $ do
  h <- IO.openFile f mode
  return h
  where q = mkQuant 1000
```

Here, the constructor IO.openFile creates a file handle to the file at path f. This constructor is supplied to mitMitigated, in addition to the "empty" state Nothing, and initial quantum q = 1 ms, which creates the corresponding mitigator and mitMitigated handle (recall Handle is a type alias of TimeMitigated IO.Handle). We note that although the default definition of mitMitigated creates a mitigator per handle, instances may provide a definition that is more coarse-grained (e.g., associate mitigator with all handles of a thread).

Unlike for mkMitigated, each mitigator must define mitigate, which specifies how a computation should be mitigated. The function takes two arguments: the mitigated handle and a computation that produces an output on given the underlying, "raw" handle. Our time mitigator instance

```haskell
instance ... ⇒ Mitigator m TStamp TStampDiff where
  mitigate m h act = ... -- Actual mitigation code
```

provides a definition for mitigate. Using mitigate we define our time mitigated hPut function as:

```haskell
hPut :: Handle → ByteString → MIO ()
hPut h bs = mitigate m (hPut h bs)
```

The mitigate function first retrieves the internal state of the mitigator corresponding to the mitigated handle and forks a new thread (allowing other mitigated actions to be executed). In the new thread, a time measurement t₁ is taken. Then, if the time difference between t₁ and the internal mitigator time stamp t₀ exceeds the quantum q, the new mitigator quantum is set to 2q; otherwise, the computation is delayed for t₁ – t₀ microseconds. Following, the IO action is executed, and the internal timestamp is replaced with the current time. We force operations on the same handle to be sequential and thus follow the latest schedule.

We finally remark that adapting an existing program to have mitigated outputs comes almost for free: a trustworthy programmer needs to define the constructor functions, such as openFile, and output functions, such as hPut, and simply lift all the remaining operations. We provide a definition for the function lift :: Monad m ⇒ m a → MitM s q m a, which lifts a computation in the m monad into the mitigation monad, without performing any actual mitigation. A simple example illustrating this is the definition of hGet which reads a specified number of bytes from a handle:

```haskell
hGet :: Handle → Int → TimeMitigated IO ByteString
hGet = lift ○ IO.hGet ○ mitVal
```

Here, mitVal simply returns the underlying "raw" handle.

6. Synchronization primitives in concurrent LIO

In the presence of concurrency, synchronization is vital. This section introduces an IFC-aware version of MVars, which are well-established synchronization Haskell primitives [25]. As with MVars, LMVars can be used in different manners: as synchronized mutable variables, as channels of depth one, or as building blocks for more complex communication and synchronization primitives.

A value of type LMVar a is a mutable location that is either empty or contains a value of type a labeled with 1. LMVars are associated with the following operations:

```haskell
ewEmptyLMVar :: (Label 1) ⇒ 1 → LMVar 1 a
putLMVar :: (Label 1) ⇒ LMVar 1 a → a → LMVar 1 a
takeLMVar :: (Label 1) ⇒ LMVar 1 a → LMVar 1 a
```

Function newEmptyLMVar takes a label 1 and creates an empty LMVar 1 a for any desired type a. The creation succeeds only if the label 1 is between the current label and clearance of the IIO computation that creates it. Function putLMVar fills an LMVar 1 a with a value of type a if it is empty and blocks otherwise. Dually, takeLMVar empties an LMVar 1 a if it is full and blocks otherwise.

Note that both takeLMVar and putLMVar observe if the LMVar is empty in order to proceed to modify its content. Precisely, takeLMVar and putLMVar perform a read and a write of the mutable location. Consequently, from a security point of view, operations on a given LMVar 1 a are executed only when the label 1 is below or equal to the clearance (i.e., 1 ≤ Lcut due to the read) and above or equal to the current label (i.e., 1 ≥ Lcur due to the write). Moreover, after either operation, Lcut is raised to 1.

In cases where schedule mispredictions are common, it is important to implement the l-grace period policy of [60]. The policy states that when there are more than l mispredictions, the new scheduling should affect all mitigators.
Many communication channels used in practice are similarly bi-directional, i.e., a read produces a write (and vice versa). For instance, reading a file may allow the access time in the inode; writing to a socket may produce an observable error if the connection is closed, etc. As described above, LMVar are bi-directional channels. If we treated them as uni-directional, observe that, a termination leak would be possible: a thread, whose current label is Low, whose current label is High; the High thread can then decide to empt...
Listing 6 Semantics for non-standard expressions.

\[
E ::= \ldots \mid \text{label } E \mid \text{unlabel } E \mid \text{out } E \mid \text{out } l E \\
| \text{1Fork } E \mid \text{newLMVar } E \mid \text{takeLMVar } E \\
| \text{putLMVar } E \mid \text{label10fLMVar } E
\]

\[
\text{(LAB)}
\]

\[
\sigma_1 l \in l \in \sigma. cl\r
\]

\[
(\Sigma, (\sigma, [E[\text{label } l \text{ }])]) \rightarrow (\Sigma, (\sigma, [E[\text{return } (Lb \ l \ e)])])
\]

\[
\text{(UNLAB)}
\]

\[
l' = \sigma_1 l \cup l \quad l' \in \sigma. cl\quad \sigma' = \sigma[1bl \rightarrow l']
\]

\[
(\Sigma, (\sigma, [E[\text{unlabel } (Lb \ l \ e)])]) \rightarrow (\Sigma, (\sigma', [E[\text{return } e)])]
\]

\[
\text{(OUTPUT)}
\]

\[
\sigma_1 l \in l \in \sigma. cl\quad \Sigma' = \Sigma[\alpha_l \rightarrow \alpha_l \rightarrow \text{out }(v)]
\]

\[
(\Sigma, (\sigma, [E[\text{out } l \ e)])]) \rightarrow (\Sigma', (\sigma, [E[\text{return }(\lambda)])])
\]

\[
\text{(LORK)}
\]

\[
\sigma_1 l \in l \in \sigma. cl\quad \Sigma' = \Sigma[\phi \rightarrow \Sigma. \phi[m \rightarrow Lb \ l \ o]]
\]

\[
\delta' = e \rightarrow \delta \rightarrow \lambda x. \text{putLMVar } m x \quad m \text{ fresh}
\]

\[
(\Sigma, (\sigma, [E[\text{1Fork } l \ e)])]) \rightarrow (\Sigma', (\sigma, [E[\text{return } (R \ m)])])
\]

\[
\text{(NLwait)}
\]

\[
\Sigma' = \Sigma[\phi \rightarrow \Sigma. \phi[m \rightarrow Lb \ l \ o]]
\]

\[
(\Sigma, (\sigma, [E[\text{newLMVar } l \ e)])]) \rightarrow (\Sigma', (\sigma, [E[\text{return } m)])]
\]

\[
\text{(TMVar)}
\]

\[
\Sigma. \phi(m) = Lb \ l \ e \quad e \in \sigma\quad \sigma_1 l \in l \in \sigma. cl\quad \sigma' = \sigma[1bl \rightarrow \sigma. 1bl \cup l]
\]

\[
\Sigma' = \Sigma[\phi \rightarrow \Sigma. \phi[m \rightarrow Lb \ l \ o]]
\]

\[
(\Sigma, (\sigma, [E[\text{takeLMVar } m)])]) \rightarrow (\Sigma', (\sigma', [E[\text{return } c)])]
\]

\[
\text{(PLMVAR)}
\]

\[
\Sigma. \phi(m) = Lb \ l \ o \quad \sigma_1 l \in l \in \sigma. cl\quad \sigma' = \sigma[1bl \rightarrow \sigma. 1bl \cup l]
\]

\[
\Sigma' = \Sigma[\phi \rightarrow \Sigma. \phi[m \rightarrow Lb \ l \ o]]
\]

\[
(\Sigma, (\sigma, [E[\text{putLMVar } m \ e)])]) \rightarrow (\Sigma', (\sigma', [E[\text{return } c)])]
\]

\[
\text{(GLABR)}
\]

\[
e = \Sigma. \phi(m)
\]

\[
(\Sigma, (\sigma, [E[\text{label10fLMVar } m \ e)])]) \rightarrow (\Sigma', (\sigma, [E[\text{label10f } e)])]
\]

The relation \((\Sigma, (\sigma, [e])) \rightarrow (\Sigma', (\sigma', [e']))\) represents a single evaluation step from expression \(e\), under the run-time environments \(\Sigma\) and \(\sigma\), to expression \(e'\) and run-time environments \(\Sigma'\) and \(\sigma'\). We define this relation in terms of a structured operational semantics via evaluation contexts [14]. We say that \(e\) reduces to \(e'\) in one step. We write \(\rightarrow^*\) for the reflexive and transitive closure of \(\rightarrow\). Symbol \(\gamma\) ranges over the internal events triggered by expressions. We utilize internal events to communicate between the threads and the scheduler. Listing 6 shows the reduction rules for the core contributions in our library. Rules (LAB) and (UNLAB) impose the same security constrains as for the sequential version of L10 [52]. Rule (LAB) generates a labeled value if and only if the label is between the current label and clearance of the L10 computation. Rule (UNLAB) requires that, when the content of a labeled value is “retrieved” and used in a L10 computation, the current label is raised \((\sigma' = \sigma[1bl \rightarrow l'])\), where \(l' = \sigma_1 l \cup l\), thus capturing the fact that the remaining computation might depend on \(e\). Output channels are treated as deques of events. We use a standard deque-like interface with operations \([e]\) and \((e)\) for front and back insertion (respectively), and we also allow pattern-matching in the rules as a representation of deconstruction operations. Rule (OUTPUT) adds the event \(\text{out}(v)\) to the end of the output channel at security level \(l(\Sigma, [\alpha_l \rightarrow \text{out}(v)])\).

The main contributions of our language are related to the primitives for concurrency and synchronization. Rule (LORK) allows for the creation of a thread and generates the internal event \(\text{fork}(e')\), where \(e'\) is the computation to spawn. The rule allocates a new \(\text{LMVar}\) in order to store the result produced by the spawned thread \((e \rightarrow \delta \rightarrow \lambda x. \text{putLMVar } m x)\). Using that \(\text{LMVar}\), the rule provides a handle to access to the thread’s result \(\text{return } (R \ m)\). Rule (LWAIT) simply uses the \(\text{LMVar}\) for the handle. Rule (TMVAR) describes the creation of a new \(\text{LMVar}\) with a label bounded by the current label and clearance \((\sigma_1 l \in l \in \sigma. cl)\). As mentioned in Section 4, operations on \(\text{LMVar}\) are bi-directional and consequently the rules (TMVAR) and (PLMVAR) require not only that the label of the mentioned \(\text{LMVar}\) be between the current label and current clearance of the thread \((\sigma_1 l \in l \in \sigma. cl)\), but that the current label be raised appropriately. As such, considering the security level \(l\) of a \(\text{LMVar}\), rule (TMVAR) raises the current label \((\sigma' = \sigma[1bl \rightarrow \sigma. 1bl \cup l])\) when emptying \((\Sigma. \phi[m \rightarrow Lb \ l \ o])\) its content \((\Sigma. \phi(m) = Lb \ l \ e)\). Similarly, considering the security level \(l\) of a \(\text{LMVar}\), rule (PLMVAR) raises the current label \((\sigma' = \sigma[1bl \rightarrow \sigma. 1bl \cup l])\) when filling \((\Sigma. \phi[m \rightarrow Lb \ l \ o])\) its content \((\Sigma. \phi(m) = Lb \ l \ o)\). Finally, rule (GLABR) fetches a labeled \(\text{LMVar}\) from the \(\text{LMVar}\) store \((e = \Sigma. \phi(m))\), i.e., a value of the form \(\text{Lb } l'\), and returns its label. To simplify the formalism, insecure programs “get stuck” in their evaluation. In practice, however, an exception is raised to the most outer trusted code, which handles it in an application-specific manner, e.g., in the case of a web server, the trusted code handles such exceptions by not sending a reply to the client. In some other cases, where the user is trusted, it may be desirable to display a notification explaining the source of error.

Listing 7 shows the formal semantics for threadpools. The relation \(\rightarrow\) represents a single evaluation step for the threadpool, in contrast with \(\rightarrow\) which is only for a single thread. We write \(\rightarrow^*\) for the reflexive and transitive closure of \(\rightarrow\). Configurations are of the form \((\Sigma, t_s)\), where \(\Sigma\) is the global runtime environment and \(t_s\) is a queue of sequential configurations. The front of the queue is the thread that is currently executing. Threads are scheduled in a round-robin fashion. The thread at the front of the queue executes one step, and it is then moved to the back of the queue (rule (STEP)). If this step involves a fork (represented by \(\rightarrow\)), a new
thread is created at the back of the queue (rule (FORK)). The identiﬁer \( t_{\text{new}} \) is bound in rule (FORK), and it stands for the conﬁguration of the newly-forked thread, i.e., \( t_{\text{new}} = (\sigma, e) \) if the parent thread transition had a label \( \text{fork}(e) \). Threads are also moved to the back of the threadpool if they are blocked, e.g., waiting to read a value from an empty LMVar (rule (NO-STEP) deﬁnes \( \rightarrow_\ast \) as the impossibility to make any progress). When a thread ﬁnishes, i.e., it can no longer reduce, the thread is removed from the queue (rule (EXIT)).

Considering IFC for a general scheduler could lead to reﬁnement attacks (e.g., [21, 22, 39, 50, 51, 56]) or the need to severely restrict programs (e.g., [47]). By considering a deterministic scheduler, our approach is more permissive — it ignores fewer programs — and robust against reﬁnement attacks. We remark that it is possible to generalize our work by considering a range of deterministic schedulers (e.g., those of [40]) without drastically changing our proof technique.

8. Security guarantees

In this section, we show that LIO computations satisfy termination-sensitive non-interference. As in [32, 43, 52], we prove this property by using the term erasure technique. The erasure function \( \varepsilon_L \) rewrites data at security levels that the attacker cannot observe into the syntax node \( \bullet \).

Listing 8 deﬁnes the erasure function \( \varepsilon_L \). This function is deﬁned in such a way that \( \varepsilon_L(e) \) contains no information above\(^1 \) level \( L \), i.e., the function \( \varepsilon_L \) replaces all the information more sensitive than \( L \) in \( e \) with a hole (\( \bullet \)). In most of the cases, the erasure function is simply applied homomorphically (e.g., \( \varepsilon_L(\cdot_1 \cdot_2) = \varepsilon_L(\cdot_1) \varepsilon_L(\cdot_2) \)). For threadpools, the erasure function is mapped into all sequential conﬁgurations; all threads with a current label above \( L \) are removed from the pool (filter \( \lambda(\sigma, e).e \neq \bullet \) (map \( \varepsilon_L t_s \)), where \( \varepsilon_L t_s \) denotes syntactic equivalence). The computation performed in a certain sequential conﬁguration is erased if the current label is above \( L \). For runtime environments and stores, we map the erasure function into their components. An output channel is erased into the empty channel (\( e \)) if it is above \( L \), otherwise the individual output events are erased according to \( \varepsilon_L \). A labeled value is erased if the label assigned to it is above \( L \).

Following the deﬁnition of the erasure function, we introduce a new evaluation relation \( \rightarrow_L \) as follows:

\[
\begin{align*}
\varepsilon_L((\Sigma, t_s)) &\rightarrow (\varepsilon_L(\Sigma), \text{filter}(\lambda(\sigma, e).e \neq \bullet) (\text{map } \varepsilon_L t_s)) \\
\varepsilon_L((\sigma, e)) &\rightarrow \begin{cases} 
(\sigma, \bullet) & \text{ if } \varepsilon_L(\sigma) \neq L \\
(\sigma, \varepsilon_L(e)) & \text{ otherwise }
\end{cases} \\
\varepsilon_L(\Sigma) &\rightarrow \Sigma[\phi \mapsto \varepsilon_L(\Sigma[\phi])](\alpha_l \mapsto \varepsilon_L(\alpha_l))_{\text{clabels}} \\
\varepsilon_L(\alpha_l) &\rightarrow \begin{cases} 
\varepsilon & \text{ if } \varepsilon \neq L \\
\text{map } \varepsilon \alpha_l & \text{ otherwise}
\end{cases} \\
\varepsilon_L(\phi) &\rightarrow \begin{cases} 
\{x, \varepsilon_L(\phi(x)) : x \in \text{dom}(\phi)\} & \text{ otherwise}
\end{cases} \\
\varepsilon_L(\text{lb } l e) &\rightarrow \begin{cases} 
\text{lb } l \bullet & \text{ if } \varepsilon \neq L \\
\text{lb } l \varepsilon_L(e) & \text{ otherwise}
\end{cases}
\end{align*}
\]

In the rest of the cases, \( \varepsilon_L \) is homomorphic.

a relation \( \rightarrow_L \) as follows:

\[
(\Sigma, t_s) \rightarrow (\varepsilon_L(\Sigma), t_s')
\]

As usual, we write \( \rightarrow_L^* \) for the reﬂexive and transitive closure of \( \rightarrow_L \). In order to prove non-interference, we will establish a simulation relation between \( \rightarrow_L^* \) and \( \rightarrow_L^L \) through the erasure function: erasing all secret data and then taking evaluation steps in \( \rightarrow_L \) is equivalent to taking steps in \( \rightarrow_L^L \) and then erasing all secret values in the resulting conﬁguration. Note that this relation would not hold if information from some level above \( L \) was being leaked by the program. In the rest of this section, we only consider well-typed terms to ensure there are no stuck conﬁgurations.

For simplicity, we assume that the address space of the memory store is split into different security levels and that allocation is deterministic. Therefore, the address returned when creating an LMVar with label \( L \) depends only on the LMVars with label \( L \) already in the store.

We start by showing that the evaluation relations \( \rightarrow_L \) and \( \rightarrow_L^L \) are deterministic.

**Proposition 1** (Determinacy of \( \rightarrow_L \)). If \( (\Sigma, t_s) \rightarrow_L (\Sigma', t_s') \) and \( (\Sigma, t) \rightarrow_L (\Sigma'', t) \) then \( (\Sigma', t_s') \approx (\Sigma'', t) \).

**Proof.** By induction on expressions and evaluation contexts, showing there is always a unique redex in every step.

**Proposition 2** (Determinacy of \( \rightarrow_L^L \)). If \( (\Sigma, t_s) \rightarrow_L^L (\Sigma', t_s') \) and \( (\Sigma, t) \rightarrow_L^L (\Sigma'', t) \) then \( (\Sigma', t_s') \approx (\Sigma'', t) \).

**Proof.** By induction on expressions and evaluation contexts, showing there is a unique redex in every step and using Lemma 1.

The next lemma establishes a simulation between \( \rightarrow^* \) and \( \rightarrow^L \).

**Lemma 1** (Many-step simulation). If \( (\Sigma, t_s) \rightarrow^* (\Sigma', t_s') \), then \( \varepsilon_L((\Sigma, t_s)) \rightarrow_L^L \varepsilon_L((\Sigma', t_s')) \).

**Proof.** In order to prove this result, we rely on properties of the erasure function, such as the fact that it is idempotent and homomorphic to the application of evaluation contexts and substitution. We show that the result holds by case analysis on the rule used to derive \( (\Sigma, t_s) \rightarrow^* (\Sigma', t_s') \) and considering different cases for threads whose current label is below (or not) level \( L \).

The \( L \)-equivalence relation \( \approx_L \) is an equivalence relation between conﬁgurations (and their parts), deﬁned as the equivalence kernel of the erasure function \( \varepsilon_L : (\Sigma, t_s) \approx_L (\Sigma', t') \) iﬀ \( \varepsilon_L((\Sigma, t_s)) = \varepsilon_L((\Sigma', t_s)) \). If two conﬁgurations are \( L \)-equivalent, they agree on all data below or at level \( L \), i.e., they cannot be distinguished by an attacker at level \( L \). Note that two queues are \( L \)-equivalent iff the threads with current label that ﬂows to \( L \) are pairwise \( L \)-equivalent in the order appearing in the queue.

The next theorem shows the non-interference property. It essentially states that if we take two executions of a program with two \( L \)-equivalent inputs, then for every intermediate step of the computation of the ﬁrst run, there is a corresponding step in the computation of the second run which results in an \( L \)-equivalent conﬁguration. Note that this also includes the termination channel, since \( L \)-equivalence of conﬁgurations requires the same public threads to be terminated. We formulate the theorem in terms of a function since we only consider programs that receive input (represented by the argument to the function) at the beginning of their execution and then produce outputs (represented by the output primitive).

---

\(^1\)We loosely use the word “above” to mean \( \notin \), since labels may not be comparable.
LIO does not impose major challenges since its interface is very simple and MVars make it possible to implement the forking primitives, and types Result and burdens R and forkIO, respectively.

\[
\forall e, e_1, e_2 : \text{Labeled } \tau \rightarrow \text{LIO} \quad \text{(Labeled } \tau \rightarrow \text{LIO)}
\]

\[
\Rightarrow \exists \Sigma_2, \sigma, e_1, e_2 \quad \text{such that } \Sigma_2 = (\sigma, e_2) \quad \text{and } \Sigma_2, \sigma, e_1 = \Sigma_2, \sigma, e_2 = \epsilon \quad \text{for all levels } k, k'\]

\[
\text{and } \Sigma_2, \sigma, e_1 = \Sigma_2, \sigma, e_2 = \epsilon \quad \text{for all levels } k, k'.
\]

\[
\exists \Sigma_2, \sigma, e_1, e_2 \quad \text{such that } \Sigma_2 = (\sigma, e_2) \quad \text{and } \Sigma_2, \sigma, e_1 = \Sigma_2, \sigma, e_2 = \epsilon \quad \text{for all levels } k, k'.
\]

Proof. Since \( e, e_1, e_2 \) are L-equivalent and \( \Sigma_2 \) and \( \Sigma_2 \) are initially empty, the initial configurations \( \langle \Sigma_1, (\sigma, e_1) \rangle \) \( \text{and } \langle \Sigma_2, (\sigma, e_2) \rangle \) must be L-equivalent. This implies that the erased configurations \( e (\Sigma_1, (\sigma, e_1)) \) \( \text{and } e (\Sigma_2, (\sigma, e_2)) \) must be syntactically equivalent. Also, by Lemma 1 (Simulation) we have \( e (\Sigma_1, (\sigma, e_1)) \rightarrow^* (\Sigma_1, e_2) \) \( \text{and } e (\Sigma_2, (\sigma, e_2)) \rightarrow^* (\Sigma_2, e_2) \) \( \text{where } e (\Sigma_1, e_2) = e (\Sigma_2, e_2) \). By Proposition 2 (Determinacy), we can always find a reduction \( e (\Sigma_1, (\sigma, e_2)) \rightarrow^* (\Sigma_1, l) \) \( \text{and } e (\Sigma_2, (\sigma, e_2)) \rightarrow^* (\Sigma_2, l) \) \( \text{where } e (\Sigma_1, l) = e (\Sigma_2, l) \). By Lemma 1 again, we have \( \langle \Sigma_1, (\sigma, e_2) \rangle \rightarrow^* (\Sigma_2, e_2) \) \( \text{and therefore } \langle \Sigma_1, l \rangle \) \( \text{and } \langle \Sigma_2, l \rangle \) are L-equivalent.

\[\Box\]

9. Example Application: Dating Website

We implemented the concurrency primitives discussed in Sections 3, 4, and 6 using Concurrent Haskell [25]. We rely on forkIO and MVar to implement the forking primitives, and types Result and LMVar. Similarly, we implement the time-based mitigator detailed in Section 5, and a small library that mitigates the standard I/O file handle functions. We refer the interested reader to the source code, available at http://www.scs.stanford.edu/~deslam/concurrent_l.io. In this section we evaluate the feasibility of leaking information through timing-based covert channels as well as the effectiveness of LIO in addressing these leaks.

To this end, we built a simple dating website that allows third-party apps to interact with users and read and write to the web server that executes apps written using LIO. In the context of our dating-website, a malicious app term, running on behalf of an (authenticated) user a, can be used to leak information on another (target) user b as follows:

- Adversary a issues a request containing a guess that user b has an interest in. Get /term?target=b&guess=g
- The trusted app container invokes the app term and forwards the request to it.
- The app term then executes the following LIO code:

\[
\text{toLabeled } t \text{ do } v \leftarrow \text{lookupDB } t \quad \text{when } (g =: v) \downarrow \text{return } \text{mhtmlResp200 } \text{Bad guess}
\]

Here, lookupDB t is used to perform a database lookup with key t. If g is present in the database entry, the app will not terminate, otherwise it will respond, denoting the guess was wrong.

We found the termination attack to be very effective. Specifically, we measured the time required to reconstruct a database of 10 users to be 73 seconds.4

If toLabeled is prohibited and forkIO is used instead, the termination attack cannot be mounted. This is because wait first raises the label of the app request handler. An attempt to output a response to the client browser will not succeed since the current label of the handler cannot flow to the label of the client’s browser. (The browser label is used to restrict apps from sending responses that the end-user, in this case a, cannot observe.) It is important to note that errors of this kind are made indistinguishable from non-terminating requests. To accomplish this, our dating site caches label violation errors and converts them to 1.

Internal timing covert channel To carry out an internal timing attack, an app must execute two threads that share a common resource. Concretely, an app can use internal timing to leak information on a target user t as follows:

- Adversary a issues a request containing a guess that t is interested in. Get /internal?target=t&guess=g
- The trusted app container invokes the app internal.
- App internal then executes the following LIO code:

\[
\text{varHigh } \leftarrow \text{fork } \\
\text{toLabeled } t \text{ do } \\
v \leftarrow \text{lookupDB } t \\
\text{when } (g =: v) \quad \text{(sleep 5000)} \\
\text{appendToAppStorage } g \\
\text{varLow } \leftarrow \text{fork } \\
\text{do } \text{sleep 3000} \\
\text{appendToAppStore } -1 \\
\text{wait varHigh} \\
\text{wait varLow} \\
r \leftarrow \text{readFromAppStore} \\
\text{return } \text{mhtmlResp200 } r
\]

The code spawns two threads. The first reads the high value in varHigh, sleeps for 5 seconds if the guess is correct, and then writes the guess to a Low-labeled persistent store5 The second thread simply writes a placeholder (-1) after waiting for 3 seconds. Here, the ordering of the data in the store reveals whether the guess is correct. If the guess is incorrect, the store will read g, -1; if the guess is correct, the store will read -1,g.

We implemented a magnified version of the attack above by sending several requests to the server. The adversary repeatedly sends requests to internal for each user in the system as a guess g. As with the termination channel attack, we found that internal timing attack is feasible. For a database of 10 users we managed to recover all the database entries in 66.92 seconds.

4 All our measurements were conducted on a laptop with an Intel Core i7 2620M (2.7GHz) processor and 8GB of RAM, with GHC 7.4.1.

5 Apps can write to the database on behalf of invoking user, we use the store notion for simplicity.
Our modifications to LIO can be used to address the internal timing attacks described above; replacing `toLabeled` with `lFork` eliminates the internal timing leaks. We observe that by using `lFork`, the time when the app executes `appendToAppStore` cannot be influenced by sensitive data. Hence, replacing `fork` and `wait` by their LIO counterparts renders the attack futile.

**External timing covert channel** We consider a simple external timing attack to our dating website in which the adversary `a` has access to a high-precision timer. An app `external` colluding with `a` can use external timing to leak a target user `t`'s interested-in relationship as follows:

- Authenticated adversary `a` issues requests containing the target user `t`: `GET /external?target=t&guess=g`
- The trusted container invokes `external` with the request.
- App `external` then proceeds to execute the following LIO code:

```haskell
lFork t $ do
  v <- lookupDB t
  when (g == v) (sleep 5000)
  return $ mkHtmlResp200 "done"
```

Given a target `t` and guess `g`, if the `g` is correct the thread sleeps; otherwise it does nothing. In both cases the final response is public. The attacker thus simply measures the response time — recognizing a delay as a correct guess. Despite its simplicity, we also found this attack to be effective. In 33 seconds, we recovered a database of 10 users. To address the leak, we mitigated the app handler, as described in Section 5. Concretely, the response time of an app was mitigated, taking into account the arrival of a request. Although we managed to recover 3 of the 10 user entries in 64 seconds—we found that recovering the remaining user entries was infeasible. The performance of well-behaved apps is unaffected.

10. Related Work

**IFC security libraries** The seminal work by Li and Zdancewic [31] presents an implementation of information-flow security as a Haskell library using arrows. Russo et al. [43] show a similar IFC security library based solely on monads, that library leverages Haskell’s type-system to statically enforce non-interference. Tsai et al. [54] extend [31] by considering side-effects and concurrency. Different from our approach, they provide termination-insensitive non-interference under a cooperative scheduler and no synchronization primitives. Jaskelioff and Russo [24] propose a library that enforces non-interference by executing the program as many times as security levels, known as secure multi-execution [11]. More recently, we propose the use of the LIO monad to track information-flow dynamically [52]. Morgenstern et al. [36] encoded an authorization- and IFC-aware programming language in Agda. Their encoding, however, does not consider computations with side-effects. Devriese and Piessens [12] used monad transformers and parametrized monads to enforce non-interference, both dynamically and statically. None of the above approaches handle the termination covert channel. Moreover, except for [54] they do not consider a concurrent language.

**Internal timing covert channel** In addressing the internal timing covert channel, compared to this work, other language-based approaches sacrifice standard semantics, practical enforcement, permissiveness, and language expressiveness. The works [49–51, 56] rely on an unrealistic primitive `protect(c)` which hides the timing behavior of a command `c`. However, assuming a scenario where it is possible to modify the scheduler, [5, 40] show how an interaction between threads and the scheduler can be used to implement a generalized version of `protect(c)`. To close internal timing leaks, the work in [4] makes the unlikely assumption that rolling back a transaction takes the same time as committing it. In contrast, our `forkLIO` and `waitLIO` are implemented using standard concurrency primitives available in Haskell.

Low-determinism [58] states that public outputs must be deterministic such that no race on public data is possible. This concept inherently makes enforcement mechanisms non-compositional (e.g., two parallel threads that only write to a public channel is considered insecure). A model-checking and type-system approach to enforcing low-determinism have been presented in [23], and [53], respectively. Mantel et al. [34] use synchronization barriers after branching on secret data and before producing public outputs. Different from these enforcement techniques, our library scales to a large number of threads.

With respect to permissiveness, some works do not allow publicly-observable events after branching on secrets. Specifically [7, 8] avoid internal timing leaks by disallowing public events after branching on secret data. They consider a fixed number of threads and no synchronization primitives. Conversely, we allow spawning arbitrary threads that branch (or loop) on secrets while the program continues producing public events. Several approaches consider a restrictive language where dynamic thread creation is not allowed [7, 8, 16, 50, 51, 56, 58]. Russo and Sabelfeld [41] remove internal timing leaks under a cooperative scheduling by manipulating `yield` commands. However, the termination channel is intrinsically present under cooperative scheduling. Closer to our approach, Russo et al. [42] transform sequential programs into concurrent programs that spawn new threads when executing branches and loops on secret values. Although the idea of spawning threads for sensitive computations is similar, we use the approach in a different context. Firstly, Russo et al. apply their technique for a simple sequential While-language, while we consider concurrent programs with synchronization primitives in the context of a practical language. Secondly, and different from our work, their approach does not consider leaks due to termination, i.e., their transformation only guarantees termination-insensitive non-interference. Finally, their transformation approach is conservative in preserving security and, as such, the termination behavior of a transformed program may change. Our proposal, on the other hand, guarantees that the semantics of the program is that which the programmer writes.

**Termination and external covert channels** There are several language-based mechanisms for addressing the termination and external timing channels. Smith and Volpano [51, 55] describe a type-system that removes the termination channel for loops whose conditional depend on secrets. This restriction is also used in [34, 47]. The work by Hedé and Sands [19] avoids the termination and external timing covert channels for sequential Java bytecode by disallowing outputs after branching on secrets. This is similar to our approach; however, we allow the spawning new threads for such sensitive tasks, while the rest of the program can still perform public events. Agat [1] describes a code transformation that removes external timing leaks by padding programs with dummy computations, and avoids the termination channel by disallowing loops on secrets. One drawback of Agat’s transformation is that if there is a conditional on secret data, and only one of the branches is non-terminating the transformed program is non-terminating. Despite this, the approach has been adapted for languages with concurrency [44, 45, 47]. Moreover, the transformation has been rephrased as a unification problem [29] and implemented with transactions [4]. While targeting sequential programs, secure multi-execution [11] removes both the termination and external timing channels. However, the latter is only closed in a special configuration, e.g., if there are as many CPUs (or cores) as security levels. We refer the reader to the systematization of knowledge paper [26] for a more detailed description of possible enforcements for timing- and termination-sensitive non-interference.
Recently, Zhang et al. [61] propose a language-based mitigation approach for a simple While-language extended with a `mitigate` primitive. Their work relies on static annotations to provide information about the underlying hardware. Compared to their work, our functional approach is more general and can be extended to address other covert channels (e.g., storage). However, their attack model is more powerful in considering the effects of hardware, including caches. Nevertheless, we find their work to be complimentary: our system can leverage static annotations and the Xen “no-fill” mode to address attacks relying on underlying hardware.

**Secure operating systems and the termination channel** A number of operating systems have been developed that intentionally left termination channels out of a belief that closing them was intractable, e.g., IX [35]. Another is Asbestos, whose limited their security ambitions to ensuring “that at least two cooperating processes are required to communicate information in violation of a label policy” [13]. In their seminal paper on the decentralized label model [37], which revived the operating system community’s interest in information flow control, Myers and Liskov expressed skepticism the problem could ever be overcome with purely dynamic checks. HiStar [59] avoided hard-coding termination channels into the operating system. In practice, however, privileged software had to implement them anyway explicitly using privileges through untangling gates, because operating-system-level resource management requires knowing when a process has exited. By contrast, we believe that we have found abstractions that are both practical on their own, and sound with respect to non-interference.

**π-calculus and information-flow** Honda et al. [22] present a sophisticated type-system that addresses internal and termination covert channels in π-calculus. They classify channels into two types: *truly linear*, used exactly once, and *non-linear* (non-deterministic), used an arbitrary number of times. The type-system allows public outputs after reading from linear channels but prevents a process from sending public outputs after receiving secret values on a non-linear channel. Without this restriction, a termination leak might occur because data might never arrive on the (non-linear) channel. The typing judgements guarantee that for every sender on a linear channel there is a corresponding receiver. Since it is not possible to have two processes writing to a common public linear channel, leaks due to internal timing are not possible. Our library relies on essentially the same mechanism Honda et al. use to prevent leaks associated with non-linear channels. However, our approach enforces IFC dynamically rather than statically. The two systems are incomparable: we are more permissive in taking the dynamic approach [46], while Honda et al. are more permissive by allowing, in certain situations, outputs on public channels after inspecting secret data. Subsequent work [21] describes a more advanced type-system that utilizes a different classification for channels (namely, linear instead of truly linear, and affine instead of non-linear) but imposes restrictions similar to [22]. Focusing on simplicity, Pottier [39] describes a type-system that disallows public outputs after reading secrets from a channel, similar to the restriction imposed by non-linear channels described above. Our work can be understood as a dynamic analog to [39] and [20]. Kobayashi [27] improves the precision of the type-system described in [21] to allow synchronization locks, or binary semaphores (similar to MVars). Kobayashi eliminates the termination covert channel. Though conceptually the work addresses and eliminates the internal timing covert channel, this is a consequence of the formalism in which a conditional and its branches “reduce” according to a structural relation. For example, the rule \( \text{if } \text{true then } P \text{ else } Q \leq P \) essentially assumes that evaluation of guards takes no time and as a consequence no internal timing attack is possible. We, like the work generalized by Kobayashi, address the internal timing covert channel by considering what we believe is a more appropriate model.

11. Summary

Many information flow control systems allow applications to sequence code with publicly visible side-effects after code that computes over sensitive data. Unfortunately, such sequencing leaks sensitive data through termination channels (which affect whether the public side-effects ever happen), internal timing channels (which affect the order of publicly visible side-effects), and external timing channels (which affect the response time of visible side-effects). Such leaks are far worse in the presence of concurrency, particularly when untrusted code can spawn new threads.

We demonstrate that such sequencing can be avoided by introducing additional concurrency when public values must reference the results of computations over sensitive data. We implemented this idea in an existing Haskell information flow library, LIO. In addition, we show how our library is amenable to mitigating external timing attacks by quantizing the appearance of externally visible side-effects. To evaluate our ideas, we prototyped the core of a dating web site showing that our interfaces are practical and our implementation does indeed mitigate these covert channels.

**Acknowledgments** We thank the anonymous reviewers for insightful comments and bringing several references to our attention. This work was funded by DARPA CRASH under contract #N00010-10-2-0488, by multiple gifts from Google, and by the Swedish research agency VR. D. Stefan is supported by the DoD through the NDEG Fellowship Program.

**References**


