A Direct-Style Effect Notation for Sequential and Parallel Programs

David Richter
Technische Universität Darmstadt, Germany

Timon Böhler
Technische Universität Darmstadt, Germany

Pascal Weisenburger
Universität St. Gallen, Switzerland

Mira Mezini
Technische Universität Darmstadt, Germany

hessian.AI, Darmstadt, Germany

Abstract

Modeling sequential and parallel composition of effectful computations has been investigated in a variety of languages for a long time. In particular, the popular do-notation provides a lightweight effect embedding for any instance of a monad. Idiom bracket notation, on the other hand, provides an embedding for applicatives. First, while monads force effects to be executed sequentially, ignoring potential for parallelism, applicatives do not support sequential effects. Composing sequential with parallel effects remains an open problem. This is even more of an issue as real programs consist of a combination of both sequential and parallel segments. Second, common notations do not support invoking effects in direct-style, instead forcing a rigid structure upon the code.

In this paper, we propose a mixed applicative/monadic notation that retains parallelism where possible, but allows sequentiality where necessary. We leverage a direct-style notation where sequentiality or parallelism is derived from the structure of the code. We provide a mechanisation of our effectful language in Coq and prove that our compilation approach retains the parallelism of the source program.

2012 ACM Subject Classification Software and its engineering → Domain specific languages; Software and its engineering → Concurrent programming structures; Software and its engineering → Parallel programming languages

Keywords and phrases do-notation, parallelism, concurrency, effects

Digital Object Identifier 10.4230/LIPIcs.ECOOP.2023.25

Supplementary Material Software (ECOOP 2023 Artifact Evaluation approved artifact):
https://doi.org/10.4230/DARTS.9.2.17
Software: https://github.com/stg-tud/parseq-notation
archived at: suh1:dir:b77c1bb5b7f47b79976c4134d1c1b143c2a7b8704

Funding David Richter: German Federal Ministry of Education and Research iBlockchain project (BMBF No. 16KIS0902).
Timon Böhler: Hessian Ministry of Higher Education, Research, Science and the Arts via the project 3rd Wave of AI (3AI).
Pascal Weisenburger: The University of St. Gallen (IPF, No. 1031569); Swiss National Science Foundation (SNSF, No. 200429).
Mira Mezini: Hessian Ministry of Higher Education, Research, Science and the Arts via the project 3rd Wave of AI (3AI); BMBF and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE; German Federal Ministry of Education and Research iBlockchain project (BMBF No. 16KIS0902).
Programming language designers often select a few common effects (state, IO, network) and bake them into the language. It is, however, impossible to predict what effects developers will need in the future (as was the case with integrated queries [32, 30, 4], reactive programming [31], asynchronous programming [51, 5], multitier programming [36, 43], differentiable programming [23], etc). Thus, we argue that language designs should be equipped with support for developer-implementable effects.

Modeling effectful computations has long been the subject of investigation in the context of various languages. As a result, there exist different abstractions and notations with different properties. A prominent abstraction for modeling effectful computations are monads (e.g., known from Haskell) and the do-notation that emerged from monadic comprehensions [55] providing a lightweight way to embed monads into programs. But monads force effects to be executed sequentially, ignoring potential for parallelism. Notations for parallelism, such as idiom bracket notation for applicatives [29], on the other hand, do not support sequential effects. Yet, programs are rarely only parallel or only sequential; thus it is desirable to support both sequential and parallel composition of effectful operations.

To the best of our knowledge, there are no approaches that provide such support. The ApplicativeDo approach by Marlow et al. [27, 28] attempts to retrofit parallelism into the do-notation, i.e., with ApplicativeDo, developers write code using the do-notation and an optimising compiler tries to infer which computations are parallelizable. Yet, in the general case, it is not possible to decide statically whether two monadic operations are actually parallelizable or whether the result of one operation depends on the execution of the other. Hence, there is a danger that the compiler either incorrectly decides that two operations can be executed in parallel, which can lead to race conditions, or conservatively decides to not parallelize operations that could actually be parallelized, reducing the potential for improved performance. To counteract race conditions, the ApplicativeDo approach requires developers to adhere to specific coding conventions such as only using expressions which are either all read-only or write-only [27].

Another weak point of Haskell’s do-notation (and thus also of the approach by Marlow et al.) is that it enforces a specific structure upon the code with strict adherence to one effect per line, which does not allow effects in arbitrary places. The do-notations in Idris [21] and Lean [25, 52] are less restrictive and support direct-style effect usage. Scala supports both structuring effectful code in do-notation via for-comprehensions and for in some cases in direct-style via async-await [44]; but, both are based on monads, thus force sequential execution. Although async-await was explicitly designed for concurrency, developers must be careful to start parallel execution before accessing their result to preserve parallelism.

In this paper, we propose a direct-style notation that enables sequential and parallel composition of effectful operations (using monads and applicatives, respectively) without forcing a specific structure of the code. We present a one-pass translation from direct-style to monadic effect combinators. Instead of trying to infer the potential for parallelism on top of sequential programs, our approach preserves parallelism that is inherent in the structure of the code thanks to direct style. This makes it easier to reason about the correctness of the proposed translation process and we present a correctness proof and its mechanization in Coq [9]. We conceptualize the preservation of parallelism as the span of a term (the length of the longest path of effectful operations) and the work of a term (the sum of all effectful operations therein). Our translation is span-preserving leveraging applicatives and monads. In contrast, notations based on monads alone are not span-preserving, as they have to chain all effectful operations into a single sequence.
Our compilation has an elegant description as a set of equations forming a structurally recursive function over the syntax, whose equations are the well-known monadic and applicative laws and free theorems. Implementations for do-notation are essentially compilers for an effectful language. They can produce efficient code by avoiding administrative redexes and generating proper tail calls. Including this optimisation in standard effectful languages modeled by monads can be seen as performing partial evaluation of the code via the semantics, extended by the monad laws. In our case, we target mixed monadic and applicative code. Therefore, our optimised translation also combines the use of the monadic laws with the use of applicative laws.

Contributions. In summary, this paper makes the following contributions:
- We present the first mixed applicative/monadic direct-style effect notation.
- We formalize an optimised one-pass translation from direct-style to effect combinators.
- We prove that our translation preserves typability, semantics, and parallelism.
- We mechanize the proof using parametric higher order abstract syntax.
- We implement the proposed translation in the Scala programming language using Scala macros, which enables us to stay close to the formal development.

Structure. The remainder of the paper is structured as follows. Section 2 provides code examples and an intuitive overview of our approach. Section 3 formally defines the proposed translation and provides a proof that it preserves typability, semantics, and parallelism, which is mechanized using parametric higher order abstract syntax. Section 4 presents the implementation in Scala. Section 5 surveys related work. Section 6 concludes the paper and presents ideas for future work.

2 Overview

In this section, we (a) briefly discuss the difference between monadic, applicative, and mixed notations by examples in Scala, and (b) informally present our mixed direct notation and its implementation by translation to effect combinators.

2.1 Monadic, Applicative, Mixed and Direct Style Notations

Functors, Applicatives and Monads. A functor (in functional programming) for \( F \): \( \text{Type} \rightarrow \text{Type} \) is a method \texttt{map} that turns a function on values into a function on values wrapped in the functor. Intuitively, a value of type \( F A \) represents an effectful computation of type \( A \). An applicative for \( F \) is a functor and a method \texttt{pure} to wrap a value into the functor, and a method \texttt{ap} (which we occasionally also write \( f \circ x \) instead of \( \text{ap} f x \)) to apply an effectful computation returning a function, to an effectful computation returning an argument. A monad for \( F \) is an applicative for \( F \) and a method \texttt{bind}, which runs an effectful computation and feeds the resulting value to another effectful computation. Below we show the mathematical description and an encoding in Scala via traits:

```
trait Functor[F[_]]:
  def map(f: A => B, a: F[A]): F[B]

trait Applicative[F[_]] extends Functor[F]:
  def pure(a: A): F[A]
  def ap(f: F[A => B], a: F[A]): F[B]

trait Monad[F[_]] extends Applicative[F]:
  def bind(f: A => F[B], a: F[A]): F[B]
```

Scala
For convenience, we will write $\text{pure}(x)$, $x.\text{bind}(f)$ and $f.\text{ap}(x)$, so we define them in Scala as an extension methods for every object which has a corresponding instance, and $\text{pure}(x)$ as a top-level function. Note that we swapped arguments for bind as a method $x.\text{bind}(f)$ with regard to its type as a function $\text{bind}(f)(x)$.

**Monadic Notation.** To illustrate monadic notations, consider the two lines of code below (left side) that use a for-comprehension for $\ldots$ yield $\ldots$. The programs execute two effectful statements $\text{fetchX}$ and $\text{fetchY}$ and bind the result in variables $x$ and $y$, respectively, to be combined by a function call to $f$. The for-comprehension (monadic notation) is desugared into explicit use of monadic $\text{bind}$ (right side).

\[
\begin{align*}
\text{for } x & \leftarrow \text{fetchX}; y \leftarrow \text{fetchY} \text{ yield } f(x)(y) \\
\text{for } y & \leftarrow \text{fetchY}; x \leftarrow \text{fetchX} \text{ yield } f(x)(y)
\end{align*}
\]

\[
\begin{align*}
\text{Scala} \quad & \text{fetchX.bind}(x \Rightarrow \text{fetchY.bind}(y \Rightarrow \text{pure}(f(x)(y)))) \\
\text{Scala} \quad & \text{fetchY.bind}(y \Rightarrow \text{fetchX.bind}(x \Rightarrow \text{pure}(f(x)(y))))
\end{align*}
\]

**Applicative Notation.** In the programs above, $x$ does not depend on $y$ and vice versa. If all statements in the program of an applicative are independent from each other – e.g., none of the variables that are introduced in the for-comprehension are used in the for part, but only after the yield, which has access to all variables introduced above – we can interpret the for-comprehension as an applicative notation instead of monadic notation. Then, the program below on the left side would be translated into the program below on the right side, using applicative $\text{ap}$ to encode actual parallelism, where $\text{ap}$ is parallel execution followed by function application. In the example program, $\text{pure}(f)$, $x$ and $y$ are executed in parallel, followed by the function application of the result of $\text{pure}(f)$ to the result of $x$ and the result of $y$:

\[
\begin{align*}
\text{for } x & \leftarrow \text{fetchX}; y \leftarrow \text{fetchY} \text{ yield } f(x)(y) \\
\text{Scala} \quad & \text{pure}(f).\text{ap}(\text{fetchX}).\text{ap}(\text{fetchY})
\end{align*}
\]

**Mixed Notation.** To illustrate where these notations for effectful computations fall short, consider the following larger program that fetches four resources from the Internet. The program first fetches a resource $\text{urlXX}$, which contains another url $\text{urlX}$, and then fetches a resource from $\text{urlX}$ and stores it in $x$. The program further fetches a resource $\text{urlYY}$, which contains another url $\text{urlY}$, and then fetches the resource from $\text{urlY}$ and stores it in $y$. Finally, the program concatenates $x$ and $y$:

\[
\begin{align*}
\text{val } \text{urlXX} & = "https://example.org/configx " \\
\text{val } \text{urlYY} & = "https://example.org/configy " \\
\text{for } \text{urlX} & \leftarrow \text{fetch}(\text{urlXX}) \\
\text{x} & \leftarrow \text{fetch}(\text{urlX}) \\
\text{urlY} & \leftarrow \text{fetch}(\text{urlYY}) \\
\text{y} & \leftarrow \text{fetch}(\text{urlY}) \\
\text{yield } x & ++ y
\end{align*}
\]

Observe that $\text{urlXX}$ needs to be fetched before $\text{urlX}$ can be fetched and $\text{urlYY}$ needs to be fetched before $\text{urlY}$ can be fetched. But fetching $\text{urlXX}$ and $\text{urlYY}$ is independent from each other; and so is $\text{urlX}$ and $\text{urlY}$ (as illustrated in the diagram above). Thus, the example contains

---

1 For-comprehensions for $\ldots$ yield $\ldots$ are Scala’s equivalent of Haskell’s do-notation do $\ldots$ return $\ldots$. The main difference besides Scala’s and Haskell’s monadic notation is that every for-comprehension must end with a yield. Yet, this does not reduce expressive power, as any do-block without a final return can be expressed with an additional binding, e.g., do $\ldots$ e can be represented by for $\ldots$ tmp $\leftarrow$ e yield tmp.
both parallel and sequential elements. However, if implemented via monadic notation, the program will run sequentially, fetching \texttt{urlXX}, then \texttt{urlX}, then \texttt{urlYY}, then \texttt{urlY}. The applicative notation, on the other hand, is not even possible because it would require all effects to be independent from each other.

**Direct-Style Mixed Notation.** In our direct-style notation, we combine the syntactic form \texttt{for ... yield ...} into a single instruction \texttt{purify}. Now, the program can be written to look like the following snippet, which reads Concatenate \texttt{(1) the result of fetching the value pointed to by a URL by fetching urlXX with (2) the result of fetching the value pointed to by a URL by fetching urlYY}:

\begin{verbatim}
  {
    val x = a.
    val y = b.
    val z = c.
    e
  }
  a.bind { x ⇒
    b.bind { y ⇒
      c.map { z ⇒
        e }
    }
  }
\end{verbatim}

The \texttt{purify} operation introduces an operation of type \texttt{↓\textbackslash{}F[X] ⇒ X} used like \texttt{... \downarrow\textcdot \downarrow}. In direct style, potential for parallelism is implicitly defined by the structure of the code. In particular, the function arguments of \texttt{concat \texttt{++}} naturally have no dependency on each other, and can therefore be executed in parallel. The corresponding program with explicit mixed monadic/applicative combinators is:

\begin{verbatim}
  pure(x ⇒ y ⇒ x ++ y).ap( fetch(urlXX).bind(fetch) ).ap( fetch(urlYY).bind(fetch) )
\end{verbatim}

The approach we propose in this paper exploits the information encoded in our direct-style notation to define a compositional (e.g., structurally recursive) and provably correct compiler that transforms such direct-style programs into semantically equivalent mixed applicative/monadic programs.

**2.2 Discussion**

We discuss the similarities and differences between different notations. For illustration, consider that monadic code in the \texttt{for/yield} notation can be easily refactored into direct style, roughly by replacing \texttt{←} with \texttt{=} \texttt{↓}. In turn, monadic code is compiled into explicit use of monadic operators by calling \texttt{bind} on each value and \texttt{call}ing \texttt{map} on the last:

\begin{verbatim}
  for x ← a
    y ← b
    z ← c
  yield e

  val x = a.
  val y = b.
  val z = c.
  e
\end{verbatim}

**Scaling.** A benefit of direct-style code is that it “scales” better for larger programs in the sense that it integrates well with common language constructs. In particular, direct style composes better with branching. Consider the following versions of the same program. On the left, the program is written in monadic notation, implemented in pure Scala, which requires us to leave and enter monadic notation a second time. The program fetches the time of the last change and the last caching of a certain request. If the resource has been updated since the last caching, we count the cache miss for statistics, request the resource freshly,
pass it to the cache and return it. Otherwise, we count the cache hit for statistics, and return
the answer from the cache. Now, compare the program on the left with a clearer direct-style
representation on the right, which is implemented using our approach (code on the right).

Sub-notations. Direct-style notation subsumes three different explicit effect notations
(Table 1). First, if a \texttt{purify} expression contains exactly one down arrow \(\downarrow\) as a mark for effect
execution (Row 1), then the notation translates to solely using the Functor interface. This
case corresponds to a standard map operation.

Second, if a \texttt{purify} expression contains multiple such marks, which are “parallel” with
guard to each other (Row 2) – i.e., when they are side-by-side inside different arguments to a
function – then the expression translates to solely using the Applicative interface. Crucially,
in this case, different effect executions cannot depend on each other. We call these effect
executions “parallel” as contrasted with “sequential” code, where a statement can depend on
the previous one. Such parallel composition enables parallel execution of effects at run time.

Third, if the expression makes use of nested marks (Row 3) – or equivalently of marks
which make use of previously bound variables which contained marks (Row 4) – then the
expression translates to using the full Monad interface, which models sequential code.

Direct-style enables to define parallelism naturally by structuring code such that the
execution of effects are independent, which gives rise to parallel execution of code where this
is possible and using sequential execution where necessary.

2.3 From Laws to a Rewrite System

We refresh the laws of functors, applicatives, and monads and give an intuition how they can
be used to optimise effectful programs. Then we state a completion of the laws into a rewrite
system. We use the symbol \(\circ\) for function composition as in \(f \circ g\), and use the symbol \((\circ)\) as
the name of function composition, when not used as an infix operator, i.e., \((f g) v := f (g v)\).

\begin{tabular}{|c|c|c|}
\hline
\textbf{Scheme} & \textbf{Description} & \textbf{Typeclass} \\
\hline
\texttt{purify{ ... \downarrow \ ... \}} & one mark & Functor map \\
\texttt{purify{ ... \downarrow \ ... \downarrow \ ... \}} & multiple marks & Applicative ap \\
\texttt{purify{ ... \downarrow \ ... \downarrow \ ... \}} & nested marks & Monad bind \\
\texttt{purify{ ... \downarrow \ ... \downarrow \ ... \}} & consecutive marks & Monad bind \\
\hline
\end{tabular}

\textbf{Table 1} Subnotations.
Laws. Typically, the coherence laws are formulated as follows [29, 26].

For Functors, \( \text{map} \) preserves identity and composition, i.e., applying the identity function to an effectful computation is the same as not doing anything; and applying two function in sequence to an effectful value is the same as applying the composite once.

\[
\begin{align*}
\text{identity} : \text{map id v} &= \text{id v} \\
\text{composition} : \text{map (f \circ g) v} &= \text{map f (map g v)}
\end{align*}
\]

For Applicatives, \( \text{pure} \) creates an effect-free, i.e., pure value from a value. The homomorphism law states that, applying a pure function to a pure argument is pure. The identity law states that, if the function is just pure, then there is nothing to do. The interchange law states that, if the argument is pure, then we can swap the pure argument with the effectful function.

\[
\begin{align*}
\text{homomorphism} : \text{pure f \circ pure v} &= \text{pure (f v)} \\
\text{identity} : \text{pure id \circ v} &= v \\
\text{interchange} : f \circ \text{pure v} &= \text{pure (\lambda f', f' v) \circ f} \\
\text{composition} : u \circ (v \circ w) &= \text{pure (\circ) \circ u \circ v \circ w}
\end{align*}
\]

For Monads, the first and second laws state that executing a pure computation amounts to not having to execute any effect at all, allowing us to eliminate the \( \text{bind} \). The third law states that \( \text{bind} \) is associative, i.e., applying two effectful functions in sequence is the same as applying the effectful composite of the two functions once.

\[
\begin{align*}
\text{leftunit} : \text{bind f (pure v)} &= f v \\
\text{rightunit} : \text{bind pure v} &= v \\
\text{associativity} : \text{bind g (bind f v)} &= \text{bind (bind g \circ f) v}
\end{align*}
\]

Free theorems. The laws are phrased as an equational theory – to create a compiler from the laws, we need to rephrase them as a terminating rewrite system. To do so, we first complete our set of equations with the following free theorems. Free theorems hold in programming languages with parametric polymorphism by parametricity for free [54, 53], therefore they are often not stated specifically in the laws, because there is no additional effort required to make them hold. On the other hand, as we are interested in making use of laws for optimisation purposes, we are allowed to make use of the free theorems as well.

Consider the function \( \text{pure} : \forall A, A \rightarrow F A \). Because it must work over all \( A \) it cannot change or create new elements of type \( A \), but only duplicate or forget values of type \( A \). Therefore, it does not matter whether we apply a function \( g \) to change the \( A \)s into \( B \)s before or after applying \( \text{pure} \). On the left we apply \( f \) on the argument of \( \text{pure} \), on the right we apply \( f \) on the result:

\[
\text{free_pure} : \text{map f (pure v)} = \text{pure (f v)}
\]

Similarly, consider the function \( \text{ap} : \forall A B, F (A \rightarrow B) \rightarrow F A \rightarrow F B \) and \( \text{bind} : \forall A B, (A \rightarrow F B) \rightarrow (F A \rightarrow F B) \). On the left we apply \( f \) on the argument of \( \text{ap} \) and \( \text{bind} \), on the right we apply \( f \) on the result, where \( (f o) \) stands for \( \lambda g, f \circ g \):

\[
\begin{align*}
\text{free_ap} : \text{ap (map (f o) g) v} &= \text{map f (ap g v)} \\
\text{free_bind} : \text{bind (map f o g) v} &= \text{map f (bind g v)}
\end{align*}
\]
We instantiate \( g \) with \( \text{pure} \ id \) in the applicative case and with \( \text{pure} \) in the monadic case, then we can extend the equation chain to the left by the free theorem of \( \text{pure} \ (\text{map} f \circ \text{pure}) \), and to the right by the identity applicative law \((\text{ap} \ (\text{pure} \ id)) \ v = v\) respectively the left-unit monad law \((\text{bind} \ \text{pure} \ v = v)\):

\[
\begin{align*}
*\text{free}_\text{ap} & : \text{ap} \ (\text{pure} \ f) \ v = \text{ap} \ (\text{map} \ f \ (\text{pure} \ id)) \ v = \text{map} \ f \ v \\
*\text{free}_\text{bind} & : \text{bind} \ (\text{pure} \circ f) \ v = \text{bind} \ (\text{map} \ f \circ \text{pure}) \ v = \text{map} \ f \ (\text{bind} \ \text{pure} \ v) = \text{map} \ f \ v
\end{align*}
\]

In fact, these two equation share a common right-hand side, and thus we can combine them to get a connection between applicative \( \text{ap} \) and monadic \( \text{bind} \):

\[
\text{ap}_\text{bind} : \text{ap} \ (\text{pure} \ f) \ v = \text{bind} \ (\text{pure} \circ f) \ v
\]

Completion. We can use the free theorems to complete the functor, applicative and monad laws into a more suitable form. In particular, we replace the identity law of the applicative with their generalization derived above. Similarly, the right hand side of the interchange law contained the left hand side of the identity law, therefore we simplify it by composition with the identity law. Further, observe that the homomorphism law becomes superfluous, as it can be constructed by applying the identity law (or equivalently by the interchange law) followed by the free theorem of \( \text{pure} \); however we will still make use of it in swapped direction, such that reading the laws from left to right, it does not overlap with the other applicative laws. The complete set of equations is now:

\[
\begin{align*}
\text{identity} & : \text{map} \ \text{id} \ v = v \\
\text{composition} & : \text{map} \ f \ (\text{map} \ g \ v) = \text{map} \ (f \circ g) \ v \\
\text{homomorphism} & : \text{pure} \ (f \ v) = \text{pure} \ f \circ \text{pure} \ v & \text{—— swapped} \\
\text{identity} & : \text{pure} \ f \circ v = \text{bind} \ (\lambda v', \text{pure} \ (f \ v')) \ v & \text{—— generalised by ap_bind} \\
\text{interchange} & : f \circ \text{pure} \ v = \text{bind} \ (\lambda f', \text{pure} \ (f' \ v)) \ f & \text{—— combined with identity} \\
\text{composition} & : u \circ (v \circ w) = \text{map} \ (\circ) \ u \circ v \circ w & \text{—— combined with identity} \\
\text{leftunit} & : \text{bind} \ f \ (\text{pure} \ v) = f \ v \\
\text{rightunit} & : \text{bind} \ \text{pure} \ v = v \\
\text{associativity} & : \text{bind} \ g \ (\text{bind} \ f \ v) = \text{bind} \ (\text{bind} \ g \circ f) \ v
\end{align*}
\]

Looking at the equations above, we see that the identity and interchange law show that a non-pure argument to \( \text{ap} \) on the left and on the right can each be represented by a \( \text{bind} \), so one might think both laws can be unified by a single law, using two \( \text{binds} \) like \( f \circ x = \text{bind} \ (\lambda \ f', \text{bind} \ (\lambda x, \text{pure} (f \ x)) \ x) \). However, it is not valid to assume this equation. Actually, there are at least two possible trivial instances of an applicative for any monad, the left-to-right applicative above, but also the right-to-left applicative: \( f \circ x = \text{bind} \ (\lambda x, \text{bind} \ (\lambda f, \text{pure} (f \ x)) \ f) \ x \). There is no reason to prefer one over the other, and, in general, the assumption of either of these equations is too strong; committing to one such equation would allow elimination of all \( \text{ap}s \) into \( \text{binds} \), and thus implies full sequentiality. To support parallelism, we have to make neither assumption and only rely on the equations derived from the applicative laws.

2.4 Translation

We present a rewrite system based on the laws, and prove its terminating by phrasing it as a structurally recursive function.
We distinguish between a source language and a target language below. The source language consists of term formers for variables, function application, and the direct-style effect execution \( \downarrow \) represented as Each. The target language consists of term formers for variables, function application, and effect combinators Pure, Bind and Ap; and parallelism is explicitly structured by those combinators.

\[
\begin{align*}
\text{(Src)} & \quad e, f ::= \text{Var } x \mid \text{App } f e \mid \text{Each } e \\
\text{(Tgt)} & \quad g, h ::= \text{Var } x \mid \text{App } g h \mid \text{Pure } g \mid \text{Ap } g h \mid \text{Bind } g h
\end{align*}
\]

The source language uses direct style. In programs written in the source language, parallelism is implicitly defined by the structure of the code. In particular, function arguments naturally have no dependency on each other, and can therefore be executed in parallel. Our compiler translates direct style into monadic and applicative combinators. The essence of our compilation strategy is to use the monadic and applicative laws directly as the actual transformation rules.

**Basic Translation.** The translation starts with the PURE expression, which is implemented as a structurally recursive function over syntax, expanding the direct-style use of effects \( \downarrow \) into the effect operation Bind, while variables are wrapped in a PURE, and function application is translated to applicative Ap, realising that function arguments can be executed in parallel.

\[
\begin{align*}
\text{PURE: Src } & \rightarrow \text{ Tgt } \\
\text{PURE } (\text{Var } x) & = \text{Pure } (\text{Var } x) \\
\text{PURE } (\text{Each } e) & = \text{Bind } (\text{PURE } e) \text{id} \\
\text{PURE } (\text{App } f e) & = \text{Ap } (\text{PURE } f) (\text{PURE } e)
\end{align*}
\]

**Optimising Translation.** If we only cared about a correct translation from the direct-style notation to the pure calculus with explicit combinators, then the translation we discussed so far is sufficient. Yet, we consider an optimising translation (Figure 1), where instead of term using the constructors Bind and Ap (capitalized) directly, we use the smart constructors AP and BIND (all capitals) instead. Both the constructor and the smart constructor of a term do construct terms that are semantically indistinguishable, i.e., \( \text{AP } f x \equiv \text{Ap } f x \). Smart constructors, however, internalize the optimisation by reducing to a simpler term if possible. The translation PURE we have described earlier can be seen as such a smart constructor for the Pure term constructor. It also preserves the semantics, i.e., \( \text{PURE } x \equiv \text{Pure } x \).

The only difference between the basic and the optimised translation, is that the optimised smart constructor PURE calls to the smart constructor AP instead of using the term constructor Ap directly, which can lead to further optimisations. In this way, we can leverage smart constructors to integrate the translation with an optimisation into a one-pass optimised translation. For the optimisation, the smart constructors apply the monadic and applicative laws, only in the other direction than the translation, i.e., bubbling up Pure in a structurally recursive way through the term, and thereby removing superfluous effect combinators in the generated code.

In particular, \( \text{AP} \) (Figure 1) will reduce the applicative application of a pure function to a pure argument back into the pure function application with only the result wrapped into Pure (which is simply the reverse rule of the homomorphism law we used above). Similarly, if either side of AP is pure, there are no two effects to be executed in parallel but just a single effect. Hence, we can reduce the term to a single monadic bind. Finally, if neither argument to AP is marked as pure, then we simply return the actual term former Ap and retain the
A Direct-Style Effect Notation for Sequential and Parallel Programs

\[
\begin{align*}
    \text{(Src)} & \quad e, f ::= \text{Var } x \mid \text{App } f \ e \mid \text{Each } e \\
    \text{(Tgt)} & \quad g, h ::= \text{Var } x \mid \text{App } g \ h \mid \text{Pure } e \mid \text{Ap } g \ h \mid \text{Bind } g \ h \\
\end{align*}
\]

\[
\begin{align*}
    \text{PURE:}& \quad \text{Src} \rightarrow \text{Tgt} \\
    \text{AP:}& \quad \text{Tgt} \rightarrow \text{Tgt} \\
    \text{BIND:}& \quad \text{Tgt} \rightarrow \text{Tgt} \\
\end{align*}
\]

\[
\begin{align*}
    \text{PURE } (\text{Var } x) & = \text{PURE } (\text{Var } x) & & \text{--- indistinguishable} \\
    \text{PURE } (\text{Each } e) & = \text{BIND } \text{id} \ (\text{PURE } e) & & \text{--- effect translation} \\
    \text{PURE } (\text{App } f \ e) & = \text{AP } (\text{PURE } f) \ (\text{PURE } e) & & \text{--- homomorphism law} \\
    \text{AP } (\text{Pure } f) \ e & = \text{BIND } (\lambda x, \text{Pure } (\text{App } f \ x)) \ e & & \text{--- identity law} \\
    \text{AP } f \ (\text{Pure } e) & = \text{BIND } (\lambda x, \text{Pure } (\text{App } x \ e)) \ f & & \text{--- interchange law} \\
    \text{AP } f \ e & = \text{App } f \ e & & \text{--- indistinguishable} \\
    \text{BIND } g \ (\text{Bind } f \ e) & = \text{BIND } (\text{Bind } g \circ \text{App } f) \ e & & \text{--- associativity law} \\
    \text{BIND } f \ (\text{Pure } e) & = \text{App } f \ e & & \text{--- left unit law} \\
    \text{BIND } \text{Pure } e & = e & & \text{--- right unit law} \\
\end{align*}
\]

\[\text{Figure 1 Optimised Translation.}\]

parallelism. The optimisation rules that apply to \text{BIND} (Figure 1) are similar. If either of its arguments is marked as pure, we can avoid performing effects at all. If we have nested binds, we can apply the associativity rule to generate a chain of binds.

Overall, seven of the ten equations above come from our generalized laws; two hold by semantic \textit{indistinguishability}, and one is the basis for our \textit{effect translation}, namely the translation of the imperative \texttt{←} to an explicit \texttt{bind}.

In the following section, we extend the language, formalize the language and the translation using a Coq mechanisation, and prove correctness.

3 Mechanisation

We define the source language that features our effect notation and a translation to a target language which is the subset of the source language that does not include the effect notation. We prove that our translation preserves typability, semantics, and parallel execution, which we measure through the program’s span and work. We have mechanized our language and proofs in Coq.

3.1 Definitions

We use (parametric) higher-order abstract syntax (PHOAS) [40, 8], which enables us to reuse the binders of the host language as binders of the guest language. PHOAS avoids the need to define first-order syntax, an operational semantics and capture-avoiding substitution, thereby removing intricate lemmas regarding substitution and hundreds of lines of code from the mechanisation, bringing the proof more in line with a more legible pen-and-paper formalisation.

Further, we use intrinsically typed terms [11, 3, 1, 2], and a type-theoretic semantics [17]. Using intrinsically typed terms together with dependent pattern matching allows us to define total evaluation (in contrast to using untyped terms or simple pattern matching where we could just define partial evaluation). The reason is that such an approach only needs to consider well-formed terms that \texttt{don’t go wrong} [33].
Listing 1 Lawful Monad.

Class Monad F := {
  map {A B}: (A → B) → F A → F B;
  pure {A}: A → F A;
  ap {A B}: F (A → B) → F A → F B;
  bind {A B}: (A → F B) → F A → F B;
}.

Class LawfulMonad F := {
  monad => Monad F;
 idl {A B} (f: A → F B) {x}: bind f (pure x) = f x;
  idr {A B} (f: A → B) {x}: bind (pure ◦ f) x = map f x;
  asc {A B C} (f: A → F B) (g: B → F C) {x}: bind g (bind f x) = bind (bind g ◦ f) x;
  apl {A B} (f: A → B) {x}: ap (pure f) x = map (λ x', f x') x;
  apr {A B} (f: F (A → B)) {x}: ap f (pure x) = map (λ f', f' x) f;
  aplr {A B} (f: A → B) {x}: map f (pure x) = pure (f x);
  map_map {A B C} (g: A → B) {x}: map f (map g x) = map (λ x, f (g x)) x;
}.

The common strategy behind all these approaches is to carve out a subset of the host language, that is the language we want to define (the guest language), and then reusing all the power of the host language to define the guest language, avoiding having to reimplement tedious implementation details: The guest types simply mirror the host types, the guest terms mirror the host terms, and the evaluation function maps guest terms to host terms.

Lawful monads. For brevity, we do not define Functor, Applicative and Monad separately. We define a class Monad and a class LawfulMonad (Listing 1). Monad contains the functions map, pure, ap, and bind. LawfulMonad extends Monad and further contains idl, idr, asc, apl, apr, aplr, and map_map, corresponding to the left and right unit law, and the associativity law of the monad, and the identity and interchange law of the applicative, the free theorem of pure, and the composition law of the functor.

Static semantics. From Coq, we use use units (tt: Unit), products ((a,b): A×B), functions ((λa, b): A→B). Mirroring the data types of the host language, we define the types for unit (T), sums (s ∨ t), products (s ∧ t), functions (s → t) and effects (M) in the guest language (Listing 2a). We define a data type ef to label terms with, as belonging to the source language src, the target language tgt, or the either language com (common) (Listing 2c). Label denotation EF m: ef (Type Type) assigns each functor in the host language. Concretely, the target and common label is assigned the identity effect functor (e.g. no effect), and the source language is assigned the effect m given as an argument.

We define a data type tm Γ B t for the syntax of our guest language (Listing 2d). The terms are parametrized by a type denotation Γ, a language label B and a type t. The common term formers are abstraction Lam e, application App e f and variables Var v to represent functions; unit Unt e, tuple Prd (e, f) and projections Fst e and Snd e to represent products. The source language has an additional term former Each e, which represents the direct-style effect application ↓ from above. The target language has additional term formers Pure e, Ap e, Map f e, and Join e representing the effect combinators.
The term former $\text{Lam}$ binds variables. In PHOAS, guest-level bindings are represented using the host language's bindings. This is why this construct takes as an argument a function which binds a variable, represented as a value of type $\Gamma t$.

**Example.** In our encoding, the terms of the guest language can be written similarly to the terms of the host language where each term former of the host language is wrapped by a term former of the guest language.

For example, the identity function $(\lambda x, x)$ can be encoded in the guest language as $\text{Lam}(\lambda x, \text{Var } x)$. A term $(\lambda x y, \text{add } x y) a b$ – an eta-expanded addition function applied to some arguments – can be expressed as $\text{Lam}(\lambda x, (\text{Lam}(\lambda y, \text{add}' (\text{App}(\text{Var } y)' (\text{App}(\text{Var } x))'))')' \text{App}' (\text{Var } a)' \text{App}' (\text{Var } b)$, writing application $x' \text{App}' y$ infix for convenience. Constructing a term of unit type $\text{tt}$ is written in the guest language as $\text{Unt } \text{tt}$. Similarly, projection on a pair $(a,b)$ is written in the guest language as $\text{Fst } (\text{Prd } (a,b))$.

**Dynamic semantics.** Next, we define the dynamic semantics corresponding to the static semantics. The static semantics has three parts: the types, the language labels and the terms. Therefore, the dynamic semantics also defines three parts.

The denotation of a type $\text{EVAL } m: \text{ty}$ (Listing 2b) maps each guest type to its corresponding host type, and is parametrized by a type constructor $m$, corresponding to the monad we evaluate in.

The denotation of a term with regard to the previously defined type denotation $\text{eval ...}$ : $\text{tm } (\text{EVAL } m) \text{B } m (\text{EVAL } m \text{ t})$ (Listing 2d) interprets the terms in a specific monad depending on which language the terms are labeled from. More concretely it takes a term of type $t$ and of label $B$ to be evaluated in monad $m$, and returns a value of the denotation of the type $\text{EVAL } m t$ wrapped in the denotation of the label $\text{EF } m B$. The evaluation for terms from the source language implicitly have effects and can therefore only be interpreted in a monadic interpreter. For the common and the target language, we define an evaluation as simply the mapping of guest term formers to their corresponding host expressions, while mapping variables to variables.

The decision of which monad to use is governed by the label denotation $\text{EF } m: \text{ef}$ (Listing 2c) mapping the target and the common language (whose terms do not have implicitly any effects) to the identity effect, e.g., no effect, while the src language is mapped to the effect $M$.

Note the way we defined the common, source and target terms, we can relable common terms into source or target terms, e.g., into any other language label $\text{relabel: } T \Gamma \text{com } t - T \Gamma \text{e } t$.

**Example.** Consider the evaluation of the following term, which constructs and then destructs a pair of units, which is equal to unit: $\text{eval } \text{com } (\text{Fst } (\text{Prd } (\text{Unt } \text{tt}, \text{Unt } \text{tt}))) = \text{tt}$.

**Translation.** The compilation from the target into the source language is performed by the smart constructor $\text{PURE}$, i.e., we compile from an effectful language into a pure language that uses monadic effect combinators. We formally define $\text{PURE}$ (Listing 3) that performs both the action of a normal pure, e.g., wraps the argument into an additional effect $\text{tm } \Gamma \text{ src } t - \text{tm } \Gamma \text{ tgt } (M \text{ t})$, and additionally performs a translation from terms form the source language with effect application $\text{Each}$ to terms of the target language using combinators $\text{Pure}$, $\text{Map}$, $\text{Ap}$ and $\text{Bind}$. This translation makes use of the smart constructors $\text{AP}$ and $\text{JOIN}$, that perform optimisations.
## Listing 2 Syntax and semantics.

### (a) Types.

**Inductive**

<table>
<thead>
<tr>
<th>ty: Type ::=</th>
</tr>
</thead>
<tbody>
<tr>
<td>T: ty</td>
</tr>
<tr>
<td>V: ty, ty, ty</td>
</tr>
<tr>
<td>∧: ty, ty, ty</td>
</tr>
<tr>
<td>→: ty, ty, ty</td>
</tr>
<tr>
<td>M: ty, ty</td>
</tr>
</tbody>
</table>

### (b) Type denotation.

**Equations**

<table>
<thead>
<tr>
<th>EVAL (m: Type → Type): ty → Type ::=</th>
</tr>
</thead>
<tbody>
<tr>
<td>m, T ⇒ Unit</td>
</tr>
<tr>
<td>m, s ∨ t ⇒ EVAL m s + EVAL m t</td>
</tr>
<tr>
<td>m, s ∧ t ⇒ EVAL m s × EVAL m t</td>
</tr>
<tr>
<td>m, s → t ⇒ EVAL m s → EVAL m t</td>
</tr>
<tr>
<td>m, M t ⇒ m (EVAL m t).</td>
</tr>
</tbody>
</table>

### (c) Labels and denotation.

**Inductive**

<table>
<thead>
<tr>
<th>ef := src</th>
<th>tgt</th>
<th>com.</th>
</tr>
</thead>
</table>

**Equations**

<table>
<thead>
<tr>
<th>EF m: ef → Type → Type ::=</th>
</tr>
</thead>
<tbody>
<tr>
<td>m, src, t ⇒ m t</td>
</tr>
<tr>
<td>m, com, t ⇒ t</td>
</tr>
<tr>
<td>m, tgt, t ⇒ t</td>
</tr>
</tbody>
</table>

### (d) Term and their denotation.

**Inductive**

<table>
<thead>
<tr>
<th>tm {Γ: ty → Type}: ef → ty → Type ::=</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var {B t}: Γ t ⇒ tm B t</td>
</tr>
<tr>
<td>Unt {B}: Unit ⇒ tm B T</td>
</tr>
<tr>
<td>Prd {B s t}: tm B s × tm B t ⇒ tm B (s ∧ t)</td>
</tr>
<tr>
<td>Fet {B s t}: tm B (s ∧ t) ⇒ tm B s</td>
</tr>
<tr>
<td>Snd {B s t}: tm B (s ∧ t) ⇒ tm B t</td>
</tr>
<tr>
<td>App {B s t}: tm B (s → t) ⇒ tm B (s → t)</td>
</tr>
<tr>
<td>Lam {B s t}: (Γ' s → tm com t) ⇒ tm B (s → t)</td>
</tr>
</tbody>
</table>

**Equations**

<table>
<thead>
<tr>
<th>eval {t m} {M:Monad m} B: tm (EVAL m) B t - EF m B (EVAL m t) ::=</th>
</tr>
</thead>
<tbody>
<tr>
<td>src , Var i ⇒ M.(pure) i</td>
</tr>
<tr>
<td>src , Lam k ⇒ M.(pure) (eval o k)</td>
</tr>
<tr>
<td>src , Unt tt ⇒ M.(pure) tt</td>
</tr>
<tr>
<td>src , Fst e ⇒ M.(map) (λ e', e'.1) (eval e)</td>
</tr>
<tr>
<td>src , Snd e ⇒ M.(map) (λ e', e'.2) (eval e)</td>
</tr>
<tr>
<td>src , App e f ⇒ M.(ap) (eval e) (eval f)</td>
</tr>
<tr>
<td>src , Prd (e, f) ⇒ M.(ap) (M.(map) (λ a', b', (a', b')) (eval e)) (eval f)</td>
</tr>
<tr>
<td>src , Each e ⇒ M.(bind) id (eval e)</td>
</tr>
</tbody>
</table>

| src , Var i ⇒ i                                              | (* com or tgt *) |
| src , Lam k ⇒ eval o k                                       |
| src , Fst e ⇒ (eval e).1                                     |
| src , Snd e ⇒ (eval e).2                                     |
| src , App e f ⇒ (eval e) (eval f)                            |
| src , Prd (e, f) ⇒ (eval e) (eval f)                         |
| src , Unt tt ⇒ tt                                            |
| tgt , Map f e ⇒ M.(map) (eval f) (eval e)                    | (* only tgt *) |
| tgt , Ap f e ⇒ M.(ap) (eval f) (eval e)                      |
| tgt , Pure e ⇒ M.(pure) (eval e)                             |
| tgt , Join e ⇒ M.(bind) id (eval e).                         | Coq |
A Direct-Style Effect Notation for Sequential and Parallel Programs

Listing 3 Translation.

Notation "f`AP`e" := (AP f e) (at level 20).

Equations PURE { Γ } x (e: tm{ Γ } src x): tm{ Γ } tgt (M x) :=
  | Var i ⇒ Pure (Var i)
  | Unt tt ⇒ Pure (Unt tt)
  | Lam j ⇒ Pure (Lam j)
  | Fst e ⇒ Pure (Λ e', Fst (Var e')) `AP` PURE e
  | Snd e ⇒ Pure (Λ e', Snd (Var e')) `AP` PURE e
  | Prd (e, f) ⇒ Pure (Λ e' f', Prd (Var e', Var f')) `AP` PURE e `AP` PURE f
  | App e f ⇒ PURE e `AP` PURE f
  | Each e ⇒ JOIN (PURE e).

Equations AP { Γ } s t (f: tm{ Γ } tgt (M (s → t))) (e: tm{ Γ } tgt (M s)): tm{ Γ } tgt (M t) :=
  | Pure f, Pure e ⇒ Pure (App f e)
  | Pure f, e ⇒ Map (Lam (λ x, App f (Var x))) e
  | f, Pure e ⇒ Map (Lam (λ x, App (Var x) e)) f
  | f, e ⇒ Ap f e.

Equations JOIN { Γ } t (e: tm{ Γ } tgt (M (M t))): tm{ Γ } tgt (M t) :=
  | Pure e ⇒ to e
  | e ⇒ Join e.

The Var, App and Each cases were discussed in Section 2.4: The direct-style use of effects Each is expanded into effect operation Bind, while variables are wrapped in PURE, and function application is translated to applicative Ap. The lambda and empty terms describe values and are simply wrapped into a pure as well.

In the case of projections and the case of tuples, we follow the general pattern of the homomorphism law, e.g., we map both the function (projection, tuple) into a Pure and we wrap all arguments in a PURE, and we apply them applicatively.

Example. Assume our language contains an effectful operation fetch. Then, translating the term e := Prd (Each (fetch "foo"), Each (fetch "bar")) yields PURE e = Pure (Lam (λ e', (Lam (λ f', Prd (Var e', Var f')))) `AP` (fetch "foo") `AP` (fetch "bar").

Span and Work. We define span and work (Listing 4), which we use to express the degree of parallelism. Span is the length of the longest chain of unhandled effectful operations, i.e., the longer the path, the more operations need to run sequentially. Hence, a shorter span for the same number of operations means a higher amount of parallelism. Work is the sum of all unhandled effectful operations. Just like evaluation interprets the value of a term, span and work are interpretations to a numeric value of a term.

As our syntax is defined from types, label and terms, we define these new interpretations as a type denotation, an effect denotation and a term denotation as well. The effect denotation for span and work is the identity function, and the type denotation is the constant function mapping all guest types to the type of natural numbers (SPAN, WORK).

More formally, we define the span of an expression to be zero for variables and values, such as empty and lambda, and for pure expressions. The span of Join and direct-style effect application Each is one more (successor S) than the span of their argument. For assertion and projection (access to first and second component), the span is simply the span of its
argument, while the span of a tuple is the maximum of its left or right branch. The span for function application, applicative application and mapping is the maximum of the span of its arguments as well, plus the span of the execution of the specified function on the argument. However, we defined our static semantics such that direct-style effect application cannot be performed under a lambda, therefore the span of the execution of any function is zero. Analogously, we define the work of an expression to be zero for variables, values, and pure expressions. Similar to the span, the work of \texttt{Join} and direct-style effect application \texttt{Each} is one more (successor $S$) than the work of their argument. The work of assertion and projection is the work of its argument. Other than span (which takes the maximum), the work of a tuple is the sum of both arguments. The work for function application, applicative application and mapping is the sum of the work of its arguments, plus the work of the execution of the specified function on the argument, which is zero, because lambdas cannot contain \texttt{Join} or \texttt{Each}.

\textbf{Example.} Assume our language contains an effectful operation \texttt{fetch}. We calculate the span and work of a term in the source language $e := \text{Prd} (\text{Each} (\text{fetch} "foo"), \text{Each} (\text{fetch} "bar"))$ as follows: $\text{span } e = 1$ and $\text{work } e = 2$. This expresses the fact that the two effects can be performed in parallel. The corresponding target language term is $e' := \text{Pure} (\text{Lam} (\lambda e', (\text{Lam} (\lambda f', \text{Prd} (\text{Var } e', \text{Var } f'))) ) \cdot \text{AP} (\text{fetch} "foo") \cdot \text{AP}' (\text{fetch} "bar")$. We get the same results for this term: $\text{span } e' = 1$ and $\text{work } e' = 2$.

\section{3.2 Proof}

Our translation should only change the encoding from direct-style to effect combinators, while the semantics, typability and parallelism of the term should be preserved. We prove that our translation preserves typability, semantics, span and work. Intuitively, the theorems hold, because our translation performed by \texttt{PURE}, \texttt{AP}, and \texttt{JOIN} are the functor, monad and applicative laws.

\textbf{Theorem 3.1 (PURE preserves types).} The translation function takes a well-typed term and produces a well-typed term, i.e., $\forall t. \text{tm } \Gamma \text{src } t \Rightarrow \text{tm } \Gamma \text{tgt } (M t)$

\textbf{Proof.} Using intrinsically-typed representation of terms, the well-typedness of the translated term is guaranteed by the fact that the definition of the translation function \texttt{PURE} is itself well-typed in Coq. ▶
We now consider the preservation of semantics. First, we show that the semantics of the smart constructors is equal to that of the normal constructors, so that they merely represent optimizations of those.

**Lemma 3.2** (AP respects semantics). \( \forall f \ e, \ \text{eval} \ tgt \ (AP \ f \ e) = \text{eval} \ tgt \ (Ap \ f \ e) \)

**Lemma 3.3** (JOIN respects semantics). \( \forall f \ e, \ \text{eval} \ tgt \ (JOIN \ e) = \text{eval} \ tgt \ (Join \ e) \)

**Proof.** By case distinction on the term structure of the arguments, using the functor, monad and applicative laws.

Next, we see that embedding the pure \texttt{com} sublanguage in the target language preserves the semantics:

**Lemma 3.4** (relabel preserves semantics). \( \forall e, \ \text{eval} \ tgt \ (relabel \ e) = \text{eval} \ com \ e \)

**Proof.** By induction on the structure of \( e \).

From this, we can deduce that the \texttt{PURE} transformation preserves the semantics of the source program.

**Theorem 3.5** (PURE preserves semantics). For all lawful monads \( M \) to be evaluated in,

\[ \forall e, \ \text{eval} \ tgt \ (PURE \ e) = \text{eval} \ src \ e \]

**Proof.** By induction on the structure of \( e \), using Lemmas 3.2–3.4.

We now want to show that \texttt{PURE} preserves the work and span of the program. This is similar to semantics preservation, except that the functions we consider map to a monoid (the natural numbers with addition and maximum, respectively) rather than a monad.

We show that AP and JOIN do not increase the span and work of a term, compared to the normal constructors.

**Lemma 3.6** (AP respects span and work).

\[ \forall f \ e, \ \text{span} \ tgt \ (AP \ f \ e) \leq \text{span} \ tgt \ (Ap \ f \ e) \] and \[ \text{work} \ tgt \ (AP \ f \ e) \leq \text{work} \ tgt \ (Ap \ f \ e) \]

**Lemma 3.7** (JOIN respects span and work).

\[ \forall e, \ \text{span} \ tgt \ (JOIN \ e) \leq \text{span} \ tgt \ (Join \ e) \] and \[ \text{work} \ tgt \ (JOIN \ e) \leq \text{work} \ tgt \ (Join \ e) \]

**Proof.** By case distinction on the term structure of the arguments, using the monoid laws.

The pure terms in the \texttt{com} sublanguage are effect-free; therefore, their span and work is equal to 0.

**Lemma 3.8** (com is effect-free). \( \forall e, \ \text{span} \ com \ e = 0 \) and \( \text{work} \ com \ e = 0 \)

**Proof.** By induction on the term structure of \( e \).

Embedding pure terms into the target language produces a term that does not perform any effects, either.

**Lemma 3.9** (relabeled terms remain effect-free).

\[ \forall e, \ \text{span} \ tgt \ (relabel \ e) = 0 \] and \[ \text{work} \ tgt \ (relabel \ e) = 0 \]

**Proof.** By induction on the term structure of \( e \).

We can then show that the translation \texttt{PURE} does not increase the span or work of the source program, thereby demonstrating that it is parallelism-preserving.
Theorem 3.10 (PURE preserves span and work).
\[ \forall e, \quad \text{span tgt (PURE e)} \leq \text{span src e} \quad \text{and} \quad \text{work tgt (PURE e)} \leq \text{work src e} \]

Proof. By induction on the term structure of \( e \), using the monoid laws of addition and maximum as well as Lemmas 3.6–3.9. ◀

4 Implementation

In this section, we describe the differences and similarities between the mechanisation in Coq and the implementation in Scala built on a macro-based AST transformation.

Structural Recursion. We keep our implementation in a general-purpose language as close to the formal model of our core calculus as possible. To this end, our implementation follows the formal translation as a structurally recursive function over the terms where possible. We use Scala macros to get access to code as AST data type, similar to the \( \text{tm} \) data type in the formalization.

Type-preserving Compilation. In Scala, we process the untyped AST for fine-grained detailed manipulations. Knowing that the translation is typability-preserving by our Coq proof, increases confidence in the implementation.

Exhaustiveness Checks. A difference between the Coq and the Scala implementation is that \texttt{Bind}, \texttt{Ap} and \texttt{Pure} are not syntax forms in Scala but represented by variable and function application in the embedding. Still, we can treat them as syntax forms to construct and destruct by defining custom patterns for pattern matching. Further, Scala macros do not define the Scala syntax as an algebraic data type (to hide compiler internals), and therefore do not offer exhaustiveness checks. Yet, by the fact that the Coq implementation is total, the Scala implementation can be expected to be as well.

Custom Effects. In the formalization, we have only a single effect, while, in the implementation, we allow every use of the notation to be instantiated with a different effect, based on the type of the expression. Our macro inspects the expression’s type and, based on this type, picks the corresponding generated combinators \texttt{bind/ap/pure} of the respective effect.

Arity. In Scala, functions may take multiple arguments. Generally, we can model functions taking multiple arguments as functions taking a single argument of a tuple with multiple fields with appropriate currying and uncurrying. Functions with multiple arguments make the compiler no longer structurally recursive over terms because, besides terms, the compiler additionally needs to mutually recurse over the list of arguments, which would complicate the proof, but is necessary for our implementation.

5 Related work

Do-Notation. Do-notations have been popular for studying a variety of styles for writing effectful code: Wadler extends list-comprehension syntax [55] to monadic comprehensions, from which modern do-notation sprung, and McBride introduced applicatives and idiom brackets as a notation for applicatives [29]. To the best of our knowledge, the only support for mixed sequential and parallel programming was introduced as a Haskell extension [27, 28] to optimise do-notation into mixed monadic/applicative operations (ApplicativeDo). In contrast, our notation preserves the parallelism inherent in the structure of the program, thereby allowing sequentiality where necessary and giving parallelism where possible.
Implementations. Besides theory, implementations for effectful guest language notations are a popular endeavor, for example: In Scala, we can find projects to support effectful programs through compiler plugins such as coroutines \[47\], Scala async \[44\], Monadless \[7\], Effectfull \[10\], Scala Workflow \[49\], Scala Context Workflow \[22\], Scala Computation Expressions \[46\], Dsl.scala \[57\], Dotty CPS \[50\]. In other languages we have: F# computation expressions \[39\], In particular proof-assistants and dependently typed languages have an interest for good support of notations for guest languages, which we can see in Idris’ \[6, 21\] Lean’s \[52, 25\], and Kind’s \[24\] notation. None of them support parallelism.

Further, the following approaches are similar to ApplicativeDo: OCaml’s monadic and applicative let \[37\], Scala avocADO \[45\], and Scala parallel-for \[48\]. But these do not support direct-style effect usage, and do not preserve parallelism.

CPS Translations. In general, effects are implemented by translating to other already known effects. In particular, all effects can be represented by the continuation effect \[14\], and thus, by translating to continuation passing style (CPS) \[42, 15\]. However, naive CPS translations introduce so-called administrative redexes, e.g., expressions containing subexpressions which do not need to be evaluated at run-time, but can already be optimised by a partial evaluation pass at compile-time. Eventually, Danvy and Nielsen \[12\] optimised the CPS-translation into a first-order, one-pass, compositional translation.

Their trick for achieving an optimal result in one pass is to build optimisations into the definitions of their translation functions. We use a similar approach in our translation through the definition of smart constructors which simplify terms using monad and applicative laws when called.

Host supporting effects. Because effects can be implemented by translation to equally or more powerful effects, besides giving a denotational semantics modelling a compiler, there is another approach – that we did not follow – by forwarding effects to the host language as well. Then, compile-time translations like ours can be avoided and effects can be implemented in languages as a library, given the host languages has sufficient powerful effects.

Filinski \[14\] studied the implementation of effects in languages with delimited continuations (e.g. shift and reset). In such an impure language it is possible to implement so-called monadic reflection – a function taking an effectful function and returning a “pure” function. This is of course only possible by exploiting the impurity of the host language to implement the effect using delimited continuation. Later, Forster \[16\] studied the translations between monadic reflection, effect handlers and delimited control. The approach to extend the underlying virtual machine by support for delimited continuations, which are sufficiently efficient for then implementing effects as normal libraries is followed by: the JVM proposal for delimited continuations \[41\], the Haskell proposal for continuation marks \[18\] and multicore-ocaml \[35\].

We are looking for a more general solution for compiling a language, that works independent of whether the runtime already supports delimited continuations or not.

Formalisation Techniques. To focus on the interesting parts of our formalisation, we used modern techniques to define features of the guest language in terms of features of the host language: In particular, we use parametric higher order abstract syntax (PHOAS) \[40, 19, 8\] to inherit binders and capture-avoiding substitution from the host language, and intrinsically-typed syntax \[11, 3, 1, 2\] to inherit type checking. The choice of PHOAS implies a limitation of our work, namely that we can formally only prove theorems about closed terms. Yet, this is a common restriction and lifting it is subject to future work.
6 Conclusion

Existing notations for composing effectful computations fall short on providing both sequential and parallel composition of effects at the same time. In this paper, we proposed a notation for mixed sequential/parallel code. Our notation allows direct-style effects, a feature that enables the sequentiality or parallelism of the effects to be determined by the structure of the code. We proved that our compilation preserves the parallelism of the source program and mechanized the proof in Coq.

An interesting next step for this line of research on direct-style notations for effects is to investigate how to cover more programming language features such as loops and branches, to integrate effects more seamlessly into the language. Besides monad and applicative functors, other effect functors, such as selectives [34, 56], comonads [38], and the theory behind effectful recursion [13] and generalizations such as arrows [20, 26] are promising possibilities.

References

A Direct-Style Effect Notation for Sequential and Parallel Programs


