

Evaluation strategies for monadic computations

Tomas Petricek

Computer Laboratory
University of Cambridge
United Kingdom

tomas.petricek@cl.cam.ac.uk

Monads have become a powerful tool for structuring effectful computations in functional programming, because they make the order of effects explicit. When translating pure code to a monadic version, we need to specify evaluation order explicitly. Two standard translations give *call-by-value* and *call-by-name* semantics. The resulting programs have different structure and types, which makes revisiting the choice difficult.

In this paper, we translate pure code to monadic using an additional operation *malias* that abstracts out the evaluation strategy. The *malias* operation is based on *computational comonads*; we use a categorical framework to specify the laws that are required to hold about the operation.

For any monad, we show implementations of *malias* that give *call-by-value* and *call-by-name* semantics. Although we do not give *call-by-need* semantics for *all* monads, we show how to turn certain monads into an extended monad with *call-by-need* semantics, which partly answers an open question. Moreover, using our unified translation, it is possible to change the evaluation strategy of functional code translated to the monadic form without changing its structure or types.

1 Introduction

Purely functional languages use lazy evaluation (also called *call-by-need*) to allow elegant programming with infinite data structures and to guarantee that a program will not evaluate a diverging term unless it is needed to obtain the final result. However, reasoning about lazy evaluation is difficult thus it is not suitable for programming with effects.

An elegant way to embed effectful computations in lazy functional languages, introduced by Moggi [16] and Wadler [26], is to use monads. Monads embed effects in a purely functional setting and explicitly specify the evaluation order of monadic (effectful) operations.

Wadler [26] gives two ways of translating pure programs to a corresponding monadic version. One approach leads to a *call-by-value* semantics, where effects of function arguments are performed before calling a function. However, if an argument has an effect and terminates the program, this may not be appropriate if the function can successfully complete without using the argument. The second approach gives a *call-by-name* semantics, where effects are performed only if the argument is actually used. However, this approach is also not always suitable, because an effect may be performed repeatedly. Wadler leaves an open question whether there is a translation that would correspond to *call-by-need* semantics, where effects are performed only when the result is needed, but at most once.

The main contribution of this paper is an alternative translation of functional code to a monadic form, parameterized by an operation *malias*. The translation has the following properties:

- A single translation gives monadic code with either call-by-name or call-by-value semantics, depending on the definition of *malias* (Section 2). When used in languages such as Haskell, it is possible to write code that is parameterized by the evaluation strategy (Section 4.1).

- The translation can be used to construct monads that provide the *call-by-need* semantics (Section 4.2), which partly answers the open question posed by Wadler. Furthermore, for some monads, it is possible to use *parallel call-by-need* semantics, where arguments are evaluated in parallel with the body of a function (Section 4.3).
- The *malias* operation has solid foundations in category theory. It arises from augmenting a *monad* structure with a *computational semi-comonad* based on the same functor (Section 3). We use this theory to define laws that should be obeyed by *malias* implementations (Section 2.2).

This paper was inspired by work on *joinads* [19], which introduced the *malias* operation for a similar purpose. However, operations with the same type and similar laws appear several times in the literature. We return to joinads in Section 4.4 and review other related work in Section 5.

1.1 Translating to monadic code

We first demonstrate the two standard options for translating purely functional code to monadic form. Consider the following two functions that use *pureLookupInput* to read some configuration property. Assuming the configuration is already loaded in memory, we can write the following pure computation¹:

```
chooseSize :: Int → Int → Int
chooseSize new legacy =
  if new > 0 then new else legacy

resultSize :: Int
resultSize =
  chooseSize (pureLookupInput "new_size")
             (pureLookupInput "legacy_size")
```

The *resultSize* function reads two different configuration keys and chooses one of them using *chooseSize*. When using a language with lazy evaluation, the call *pureLookupInput "legacy_size"* is performed only when the value of "new_size" is less than or equal to zero.

To modify the function to actually read configuration from a file as opposed to performing in-memory lookup, we now use *lookupInput* which returns *IO Int* instead of the *pureLookupInput* function. Then we need to modify the two above functions. There are two mechanical ways that give different semantics.

Call-by-value. In the first style, we call *lookupInput* and then apply monadic bind on the resulting computation. This reads both of the configuration values before calling the *chooseSize* function, and so arguments are fully evaluated before the body of a function as in the *call-by-value* evaluation strategy:

```
chooseSizecbv :: Int → Int → IO Int
chooseSizecbv new legacy =
  return (if new > 0 then new else legacy)

resultSizecbv :: IO Int
resultSizecbv = do
  new ← lookupInputcbv "new_size"
  legacy ← lookupInputcbv "legacy_size"
  chooseSizecbv new legacy
```

¹Examples are written in Haskell and can be found at: <http://www.cl.cam.ac.uk/~tp322/papers/malias.html>

In this version of the translation, a function of type $A \rightarrow B$ is turned into a function $A \rightarrow M B$. For example, the $chooseSize_{cbv}$ function takes integers as parameters and returns a computation that returns an integer and may perform some effects. When calling a function in this setting, the arguments may not be fully evaluated (the functional part is still lazy), but the effects associated with obtaining the value of the argument happen before the function call.

For example, if the call $lookupInput_{cbv}$ "new_size" read a file and then returned 1024, but the operation $lookupInput_{cbv}$ "legacy_size" caused the program to crash because a specified key was not present in a configuration file, then the entire program would crash.

Call-by-name. In the second style, we pass unevaluated computations as arguments to functions. This means we call $lookupInput$ to create an effectful computation that will read the input, but the computation is then passed to $chooseSize$, which may not need to evaluate it:

```

chooseSize_cbn :: IO Int → IO Int → IO Int
chooseSize_cbn new legacy = do
  newVal ← new
  if newVal > 0 then new else legacy

resultSize_cbn :: IO Int
resultSize_cbn =
  chooseSize_cbn (lookupInput_cbn "new_size")
                 (lookupInput_cbn "legacy_size")

```

The translation turns a function of type $A \rightarrow B$ into a function $M A \rightarrow M B$. This means that the $chooseSize_{cbn}$ function takes a computation that performs the I/O effect and reads information from the configuration file, as opposed to taking a value whose effects were already performed.

Following the mechanical translation, $chooseSize_{cbn}$ returns a monadic computation that evaluates the first argument and then behaves either as *new* or as *legacy*, depending on the obtained value. When the resulting computation is executed, the computation which reads the value of the "new_size" key may be executed repeatedly. First, inside the $chooseSize_{cbn}$ function and then repeatedly when the result of this function is evaluated. In this particular example, we can easily change the code to perform the effect just once, but this is not generally possible for computations obtained by the *call-by-name* translation.

2 Abstracting evaluation strategy

The translations demonstrated in the previous section have two major problems. Firstly, it is not easy to switch between the two – when we introduce effects using monads, we need to decide to use one or the other style and changing between them later on involves rewriting of the program and changing types. Secondly, even in the *IO* monad, we cannot easily implement a *call-by-need* strategy that would perform effects only when a value is needed, but at most once.

2.1 Translation using aliasing

To solve these problems, we propose an alternative translation. We require a monad m with an additional operation *alias* that abstracts out the evaluation strategy and has a type $m a \rightarrow m (m a)$. The term *aliasing* refers to the fact that some part of effects may be performed once and their results shared in multiple monadic computations. The translation of the previous example using *alias* looks as follows:

```

chooseSize :: IO Int → IO Int → IO Int
chooseSize new legacy = do
  newVal ← new
  if newVal > 0 then new else legacy
resultSize :: IO Int
resultSize = do
  new ← malias (lookupInput "new_size")
  legacy ← malias (lookupInput "legacy_size")
  chooseSize new legacy

```

The types of functions and access to function parameters are translated in the same way as in the *call-by-name* translation. The *chooseSize* function returns a computation *IO Int* and its parameters also become computations of type *IO Int*. When using the value of a parameter, the computation is evaluated using monadic bind (e.g. the line *newVal ← new* in *chooseSize*).

The computations passed as arguments are not the original computations as in the *call-by-name* translation. The translation inserts a call to *malias* for every function argument (and also for every let-bound value, which can be encoded using lambda abstraction and application). The computation returned by *malias* has a type $m (m a)$, which makes it possible to perform the effects at two different call sites:

- When simulating the *call-by-value* strategy, all effects are performed when binding the outer monadic computation before a function call.
- When simulating the *call-by-name* strategy, all effects are performed when binding the inner monadic computation, when the value is actually needed.

These two strategies can be implemented by two simple definitions of *malias*. However, by delegating the implementation of *malias* to the monad, we make it possible to implement more advanced strategies as well. We discuss some of them later in Section 4. We keep the translation informal until Section 2.3 and discuss the *malias* operation in more detail first.

Implementing call-by-name. To implement the *call-by-name* strategy, the *malias* operation needs to return the computation specified as an argument inside the monad. In the type $m (m a)$, the outer *m* will not carry any effects and the inner *m* will be the same as the original computation:

```

malias :: m a → m (m a)
malias m = return m

```

From the monad laws (see Figure 1), we know that applying monadic bind to a computation created from a value using *return* is equivalent to just passing the value to the rest of the computation. This means that the additional binding in the translation does not have any effect and the resulting program behaves as the *call-by-name* strategy. A complete proof can be found in Appendix A.

Implementing call-by-value. Implementing the *call-by-value* strategy is similarly simple. In the returned computation of type $m (m a)$, the computation corresponding to the outer *m* needs to perform all the effects. The computation corresponding to the inner *m* will be a computation that simply returns the previously computed value without performing any effects:

```

malias :: m a → m (m a)
malias m = m >>= (return ∘ return)

```

Functor with *unit* and *join*:

$$\begin{aligned} \text{unit} &:: a \rightarrow m a \\ \text{map} &:: m a \rightarrow (a \rightarrow b) \rightarrow m b \\ \text{join} &:: m (m a) \rightarrow m a \end{aligned}$$

Join laws (*unit* and *map* laws omitted):

$$\begin{aligned} \text{join} \circ \text{map join} &= \text{join} \circ \text{join} \\ \text{join} \circ \text{map unit} &= \text{id} = \text{join} \circ \text{unit} \\ \text{join} \circ \text{map (map } f) &= \text{map } f \circ \text{join} \end{aligned}$$

Definition using *bind* ($\gg\equiv$) and *return*:

$$\begin{aligned} \text{return} &:: a \rightarrow m a \\ \gg\equiv &:: m a \rightarrow (a \rightarrow m b) \rightarrow m b \end{aligned}$$

Monad laws about *bind* and *return*:

$$\begin{aligned} \text{return } a \gg\equiv f &\equiv f a \\ m \gg\equiv \text{return} &\equiv m \\ (m \gg\equiv f) \gg\equiv g &\equiv m \gg\equiv (\lambda x \rightarrow f x \gg\equiv g) \end{aligned}$$

Figure 1: Two equivalent ways of defining monads with monad laws

In Haskell, the second line could be written as $\text{liftM return } m$. The liftM operation represents the functor associated with the monad. This means that binding on the returned computation performs all the effects, obtains a value v and returns a computation $\text{return } v$.

When calling a function that takes an argument of type $m a$, the argument passed to it using this implementation of *malias* will always be constructed using the *return* operation. Hence the resulting behaviour is equivalent to the original *call-by-value* translation. Detailed proof can be found in Appendix A.

2.2 The malias operation laws

In order to define a reasonable evaluation strategy, we require the *malias* operation to obey a number of laws. The laws follow from the theoretical background that is discussed in Section 3, namely from the fact that *malias* is the *cojoin* operation of a computational semi-comonad.

The laws that relate *malias* to the monad are easier to write in terms of *join*, *map* and *unit* than using the formulation based on $\gg\equiv$ and *return*. For completeness, the two equivalent definitions of monads with the monad laws are shown in Figure 1. Although we do not show it, one can be easily defined in terms of the other. The required laws for *malias* are the following:

$$\begin{aligned} \text{map (map } f) \circ \text{malias} &= \text{malias} \circ (\text{map } f) && (\textit{naturalit}) \\ \text{map malias} \circ \text{malias} &= \text{malias} \circ \text{malias} && (\textit{associativity}) \\ \text{malias} \circ \text{unit} &= \text{unit} \circ \text{unit} && (\textit{computationality}) \\ \text{join} \circ \text{malias} &= \text{id} && (\textit{identity}) \end{aligned}$$

The first two laws follow from the fact that *malias* is a *cojoin* operation of a comonad. The *naturalit* law specifies that applying a function to a value inside a computation is the same as applying the function to a value inside an aliased computation. The *associativity* law specifies that aliasing an aliased computation is the same as aliasing a computation produced by an aliased computation.

The *computationality* law is derived from the fact that the comonad defining *malias* is a *computational comonad* with *unit* as one of the components. The law specifies that aliasing of a pure computation creates a pure computation. Finally, the *identity* law relates *malias* with the monadic structure, by requiring that *join* is a *left inverse* of *malias*. Intuitively, it specifies that aliasing a computation of type $m a$ and then joining the result returns the original computation.

All four laws hold for the two implementations of *malias* presented in the previous section. We prove that the laws hold for any monad using the standard monad laws. The proofs can be found in Appendix B. We discuss the intuition behind the laws in Section 2.4 and describe their categorical foundations in Section 3. The next section formally presents the translation algorithm.

$$\begin{aligned}
\llbracket x \rrbracket_{cbn} &= x \\
\llbracket \lambda x. e \rrbracket_{cbn} &= \mathit{unit} (\lambda x. \llbracket e \rrbracket_{cbn}) \\
\llbracket e_1 e_2 \rrbracket_{cbn} &= \mathit{bind} \llbracket e_1 \rrbracket_{cbn} (\lambda f. f \llbracket e_2 \rrbracket_{cbn}) \\
\llbracket \mathbf{let} x = e_1 \mathbf{in} e_2 \rrbracket_{cbn} &= (\lambda x. \llbracket e_2 \rrbracket_{cbn}) \llbracket e_1 \rrbracket_{cbn}
\end{aligned}$$

Figure 2: Wadler’s call-by-name translation of λ calculus.

$$\begin{aligned}
\llbracket x \rrbracket_{cbv} &= \mathit{unit} x \\
\llbracket \lambda x. e \rrbracket_{cbv} &= \mathit{unit} (\lambda x. \llbracket e \rrbracket_{cbv}) \\
\llbracket e_1 e_2 \rrbracket_{cbv} &= \mathit{bind} \llbracket e_1 \rrbracket_{cbv} (\lambda f. \mathit{bind} \llbracket e_2 \rrbracket_{cbv} (\lambda x. f x)) \\
\llbracket \mathbf{let} x = e_1 \mathbf{in} e_2 \rrbracket_{cbv} &= \mathit{bind} \llbracket e_1 \rrbracket_{cbv} (\lambda x. \llbracket e_2 \rrbracket_{cbv})
\end{aligned}$$

Figure 3: Wadler’s call-by-value translation of λ calculus.

2.3 Lambda calculus translation

The *call-by-name* and *call-by-value* translations given in Section 1.1 were first formally introduced by Wadler [26]. In this section, we present a similar formal definition of our translation based on the *malias* operation. For our source language, we use a simply-typed λ calculus with let-binding:

$$\begin{aligned}
e \in \mathit{Expr} \quad e &::= x \mid \lambda x. e \mid e_1 e_2 \mid \mathbf{let} x = e_1 \mathbf{in} e_2 \\
\tau \in \mathit{Type} \quad \tau &::= \alpha \mid \tau_1 \rightarrow \tau_2
\end{aligned}$$

The target language of the translation is identical but for one exception – it adds a type scheme $M \tau$ representing monadic computations. The *call-by-name* and *call-by-value* translation of the lambda calculus are shown in Figure 2 and Figure 3, respectively. In the translation, we write \gg as *bind*. The translation of types and typing judgements are omitted for simplicity and can be found in the original paper [26].

Our translation, called *call-by-alias*, is presented below. The translation has similar structure to Wadler’s *call-by-name* translation, but it inserts *malias* operation in the last two cases:

$$\begin{aligned}
\llbracket \alpha \rrbracket_{cba} &= \alpha \\
\llbracket \tau_1 \rightarrow \tau_2 \rrbracket_{cba} &= M \llbracket \tau_1 \rrbracket_{cba} \rightarrow M \llbracket \tau_2 \rrbracket_{cba} \\
\llbracket x \rrbracket_{cba} &= x \\
\llbracket \lambda x. e \rrbracket_{cba} &= \mathit{unit} (\lambda x. \llbracket e \rrbracket_{cba}) \\
\llbracket e_1 e_2 \rrbracket_{cba} &= \mathit{bind} \llbracket e_1 \rrbracket_{cba} (\lambda f. \mathit{bind} (\mathit{malias} \llbracket e_2 \rrbracket_{cba}) f) \\
\llbracket \mathbf{let} x = e_1 \mathbf{in} e_2 \rrbracket_{cba} &= \mathit{bind} (\mathit{malias} \llbracket e_1 \rrbracket_{cba}) (\lambda x. \llbracket e_2 \rrbracket_{cba})
\end{aligned}$$

$$\llbracket x_1 : \tau_1, \dots, x_n : \tau_n \vdash e : \tau \rrbracket_{cba} = x_1 : M \llbracket \tau_1 \rrbracket_{cba}, \dots, x_n : M \llbracket \tau_n \rrbracket_{cba} \vdash \llbracket e \rrbracket_{cba} : M \llbracket \tau \rrbracket_{cba}$$

The translation turns user-defined variables of type τ into variables of type $M \tau$. A variable access x is translated to a variable access, which now represents a computation (that may have effects). A lambda expression is turned into a lambda expression wrapped in a pure monadic computation.

The two interesting cases are application and let-binding. When translating function application, we bind on the computation representing the function. We want to call the function f with an aliased computation as an argument. This is achieved by passing the translated argument to *malias* and then applying *bind* again. The translation of let-binding is similar, but slightly simpler, because it does not need to use *bind* to obtain a function.

The definition of $\llbracket - \rrbracket_{cba}$ includes a well-typedness-preserving translation of typing judgements. The details of the proof can be found in Appendix C. In the translation, let-binding is equivalent to $(\lambda x.e_2) e_1$, but we include it to aid the intuition and to simplify motivating examples in the next section.

2.4 The meaning of *malias* laws

Having provided the translation, we can discuss the intuition behind the *malias* laws and what they imply about the translated code. The *naturality* law specifies that *malias* is a natural transformation. Although we do not give a formal proof, we argue that the law follows from the parametricity of the *malias* type signature and can be obtained using the method described by Voigtländer [25].

Effect conservation. The meaning of *associativity* and *identity* can be informally demonstrated by treating effects as units of information. Given a computation type involving a number of occurrences of m , we say that there are *effects* (or *information*) associated with each occurrence of m .

The *identity* law specifies that $join \circ malias$ does not lose effects – given a computation of type $m a$, the *malias* operation constructs a computation of type $m (m a)$. This is done by splitting the effects of the computation between two monadic computations. Requiring that applying *join* to the new computation returns the original computation means that all the effects of the original computation are preserved, because *join* combines the effects of the two computations.

The *associativity* law specifies how the effects should be split. When applying *malias* to an aliased computation of type $m (m a)$, we can apply the operation to the inner or the outer m . Underlining second aliasing, the following two computations should be equal: $\underline{m} (m (m a)) = m (\underline{m} (m a))$. The law forbids implementations that split the effects in an asymmetric way. For instance, if $m a$ represents a step-wise computation that can be executed by a certain number of steps, *malias* cannot execute the first half of steps and return a computation that evaluates the other half. The proportions associated with individual m values would be $1/4, 1/4$ and $1/2$ on the left and $1/2, 1/4$ and $1/4$ on the right-hand side.

Semantics-preserving transformations. The *computationality* and *identity* (again) laws also specify that certain semantics-preserving transformations on the original source code correspond to equivalent terms in the code translated using the *call-by-alias* transformation. The source transformations corresponding to *computationality* and *identity*, respectively, are the following:

$$\begin{aligned} \mathbf{let } f = \lambda x.e_1 \mathbf{ in } e_2 &\equiv e_2[f \leftarrow \lambda x.e_1] \\ \mathbf{let } x = e \mathbf{ in } x &\equiv e \end{aligned}$$

The construction of the lambda function in the first equation is the only place where the translation inserts *unit*. The *computationality* law can be applied when a lambda abstraction appears in a position where *malias* is inserted. Thanks to the first monad law (Figure 1), the value assigned to f in the translation is *unit* ($\lambda x.e_1$), which is equivalent to the translation of a term after the substitution. In the second equation, the left-hand side is translated as $bind (malias e) id$, which is equivalent to $join (malias e)$, so the equation directly corresponds to the *identity* law.

Discussion of completeness. The above discussion, together with the theoretical foundations introduced in the next section, supports the claim that our laws are necessary. We do not argue that our set of laws is complete – for instance, we might want to specify that aliasing of an already aliased computation has no effect, which is difficult to express in an equational form.

However, the definition of completeness, in this context, is elusive. One possible approach that we plan to investigate in future work is to expand the set of semantics-preserving transformations that should hold, regardless of an evaluation strategy.

3 Computational semi-bimonads

In this section, we formally describe the structure that underlies a monad having a *malias* operation as described in the previous section. Since the *malias* operation corresponds to an operation of a comonad associated with a monad, we first review the definitions of a monad and a comonad. Monads are well-known structures in functional programming. Comonads are dual structures to monads that are less widespread, but they have also been used in functional programming and semantics (Section 5):

Definition 1. A monad over a category \mathcal{C} is a triple (T, η, μ) where $T : \mathcal{C} \rightarrow \mathcal{C}$ is a functor, $\eta : I_{\mathcal{C}} \rightarrow T$ is a natural transformation from the identity functor to T , and $\mu : T^2 \rightarrow T$ is a natural transformation, such that the following associativity and identity conditions hold, for every object A :

- $\mu_A \circ T \mu_A = \mu_A \circ \mu_{TA}$
- $\mu_A \circ \eta_{TA} = \text{id}_{TA} = \mu_A \circ T \eta_A$

Definition 2. A comonad over a category \mathcal{C} is a triple (T, ε, δ) where $T : \mathcal{C} \rightarrow \mathcal{C}$ is a functor, $\varepsilon : T \rightarrow I_{\mathcal{C}}$ is a natural transformation from T to the identity functor, and $\delta : T \rightarrow T^2$ is a natural transformation from T to T^2 , such that the following associativity and identity conditions hold, for every object A :

- $T \delta_A \circ \delta_A = \delta_{TA} \circ \delta_A$
- $\varepsilon_{TA} \circ \delta_A = \text{id}_{TA} = T \varepsilon_A \circ \delta_A$

In functional programming terms, the natural transformation η corresponds to *unit* $:: a \rightarrow m a$ and the natural transformation μ corresponds to *join* $:: m (m a) \rightarrow m a$. A comonad is a dual structure to a monad – the natural transformation ε corresponds to an operation *counit* $:: m a \rightarrow a$ and δ corresponds to *cojoin* $:: m a \rightarrow m (m a)$. An equivalent formulation of comonads in functional programming uses an operation *cobind* $:: m a \rightarrow (m a \rightarrow b) \rightarrow m b$, which is dual to $\gg=$ of monads.

A simple example of a comonad is the product comonad. The type $m a$ stores the value of a and some additional state S , meaning that $TA = A \times S$. The ε (or *counit*) operation extracts the value A ignoring the additional state. The δ (or *cojoin*) operation duplicates the state. In functional programming, the product comonad is equivalent to the reader monad $TA = S \rightarrow A$.

In this paper, we use a special variant of comonads. Computational comonads, introduced by Brookes and Geva [5], have an additional operation γ together with laws specifying its properties:

Definition 3. A computational comonad over a category \mathcal{C} is a quadruple $(T, \varepsilon, \delta, \gamma)$ where (T, ε, δ) is a comonad over \mathcal{C} and $\gamma : I_{\mathcal{C}} \rightarrow T$ is a natural transformation such that, for every object A ,

- $\varepsilon_A \circ \gamma_A = \text{id}_A$
- $\delta_A \circ \gamma_A = \gamma_{TA} \circ \gamma_A$.

A *computational comonad* has an additional operation γ which has the same type as the η operation of a monad, that is $a \rightarrow m a$. In the work on computational comonads, the transformation γ turns an extensional specification into an intensional specification without additional computational information.

In our work, we do not need the natural transformation corresponding to $\text{counit} :: m a \rightarrow a$. We define a computational *semi-comonad*, which is a computational comonad without the natural transformation ε and without the associated laws. The remaining structure is preserved:

Definition 4. A computational semi-comonad over a category \mathcal{C} is a triple (T, δ, γ) where $T : \mathcal{C} \rightarrow \mathcal{C}$ is a functor, $\delta : T \rightarrow T^2$ is a natural transformation from T to T^2 and $\gamma : I_{\mathcal{C}} \rightarrow T$ is a natural transformation from the identity functor to T , such that the following associativity and computability conditions hold, for every object A :

- $T \delta_A \circ \delta_A = \delta_{TA} \circ \delta_A$
- $\delta_A \circ \gamma_A = \gamma_{TA} \circ \gamma_A$.

Finally, to define a structure that models our monadic computations with the *malias* operation, we combine the definition of a monad and computational semi-comonad. We require that the two structures share the functor T and that the natural transformation $\eta : I_{\mathcal{C}} \rightarrow T$ of a monad coincides with the natural transformation $\gamma : I_{\mathcal{C}} \rightarrow T$ of a computational comonad.

Definition 5. A computational semi-bimonad over a category \mathcal{C} is a quadruple (T, η, μ, δ) where (T, η, μ) is a monad over a category \mathcal{C} and (T, δ, η) is a computational semi-comonad over \mathcal{C} , such that the following additional condition holds, for every object A :

- $\mu_A \circ \delta_A = \text{id}_{TA}$

The definition of computational semi-bimonad relates the monadic and comonadic parts of the structure using an additional law. Given an object A , the law specifies that taking TA to T^2A using the natural transformation δ_A of a comonad and then back to TA using the natural transformation μ_A is identity.

3.1 Revisiting the laws

The laws of computational semi-bimonad as defined in the previous section are exactly the laws of our monad equipped with the *malias* operation. In this section, we briefly review the laws and present the category theoretic version of all the laws demonstrated in Section 2.2. We require four laws in addition to the standard monad laws (which are omitted in the summary below). A diagrammatic demonstration is shown in Figure 4. For all objects A and B of \mathcal{C} and for all $f : A \rightarrow B$ in \mathcal{C} :

$$\begin{aligned}
 T^2 f \circ \delta_A &= \delta_B \circ T f && (\text{naturalness}) \\
 T \delta_A \circ \delta_A &= \delta_{TA} \circ \delta_A && (\text{associativity}) \\
 \delta_A \circ \eta_A &= \eta_{TA} \circ \eta_A && (\text{computationality}) \\
 \mu_A \circ \delta_A &= \text{id}_{TA} && (\text{identity})
 \end{aligned}$$

The *naturalness* law follows from the fact that δ is a natural transformation and so we did not state it explicitly in Definition 5. However, it is one of the laws that are translated to the functional programming interpretation. The *associativity* law is a law of comonad – the other law in Definition 2 does not apply in our scenario, because we only work with *semi-comonad* that does not have natural transformation ε (*counit*). The *computationality* law is a law of a computational comonad and finally, the *identity* law is the additional law of *computational semi-bimonads*.

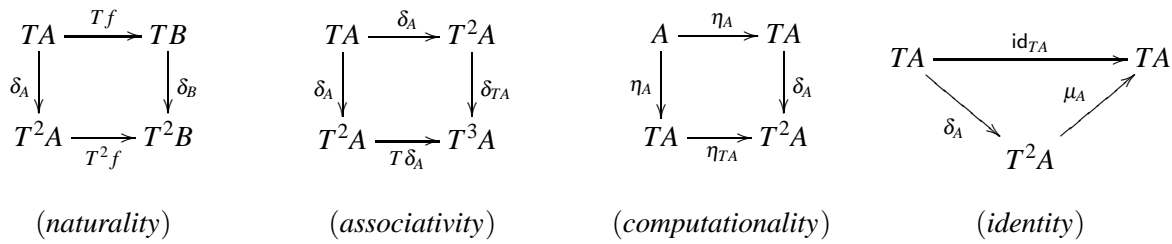


Figure 4: Diagrammatic representation of the four additional properties of semi-bimonads

4 Abstracting evaluation strategy in practice

In this section, we present several practical uses of the *malias* operation. We start by showing how to write monadic code that is parameterized over the evaluation strategy and then consider expressing *call-by-need* in this framework. Then we also briefly consider *parallel call-by-need* and the relation between *malias* and *joinads* and the **docase** notation [19].

4.1 Parameterization by evaluation strategy

One of the motivations of this work is that the standard monadic translations for *call-by-name* and *call-by-value* produce code with different structure. Section 2.3 gave a translation that can be used with both of the evaluation strategies just by changing the definition of the *malias* operation. In this section, we make one more step – we show how to write code parameterized by evaluation strategy.

We define a *monad transformer* [13] that takes a monad and turns it into a monad with *malias* that implements a specific evaluation strategy. Our example can then be implemented using functions that are polymorphic over the monad transformer. We continue using the previous example based on the *IO* monad, but the transformer can operate on any monad.

As a first step, we define a type class named *MonadAlias* that extends *Monad* with the *malias* operation. To keep the code simple, we do not include comments documenting the laws:

```
class Monad m => MonadAlias m where
  malias :: m a -> m (m a)
```

Next, we define two new types that represent monadic computations using the *call-by-name* and *call-by-value* evaluation strategy. The two types are wrappers that make it possible to implement two different instances of *MonadAlias* for any underlying monadic computation *m a*:

```
newtype CbV m a = CbV { runCbV :: m a }
newtype CbN m a = CbN { runCbN :: m a }
```

The snippet defines types *CbV* and *CbN* that represent two evaluation strategies. Figure 5 shows the implementation of three type classes for these two types. The implementation of the *Monad* type class is the same for both types, because it simply uses *return* and $\gg=$ operations of the underlying monad. The implementation of *MonadTrans* wraps a monadic computation *m a* into a type *CbV m a* and *CbN m a*, respectively. Finally, the instances of the *MonadAlias* type class associate the two implementations of *malias* (from Section 2.1) with the two data types.

```

instance Monad m  $\Rightarrow$  Monad (CbV m) where
  return v = CbV (return v)
  (CbV a) >>= f = CbV (a >>= (runCbV  $\circ$  f))
instance Monad m  $\Rightarrow$  Monad (CbN m) where
  return v = CbN (return v)
  (CbN a) >>= f = CbN (a >>= (runCbN  $\circ$  f))

instance MonadTrans CbV where lift = CbV
instance MonadTrans CbN where lift = CbN

instance Monad m  $\Rightarrow$  MonadAlias (CbV m) where
  alias m = m >>= (return  $\circ$  return)
instance Monad m  $\Rightarrow$  MonadAlias (CbN m) where
  alias m = return m

```

Figure 5: Instances of *Monad*, *MonadTrans* and *MonadAlias* for evaluation strategies

Example. Using the previous definitions, we can now rewrite the example from Section 1.1 using generic functions that can be executed using both *runCbV* and *runCbN*. Instead of implementing *alias* for a specific monad such as *IO a*, we use a monad transformer *t* that lifts the monadic computation to either *CbV IO a* or to *CbN IO a*. This means that all functions will have constraints *MonadTrans t*, specifying that *t* is a monad transformer, and *MonadAlias (t m)*, specifying that the computation implements the *alias* operation.

In Haskell, this can be succinctly written using *constraint kinds* [3], that make it possible to define a single constraint *EvalStrategy t m* that combines both of the conditions²:

```

type EvalStrategy t m = (MonadTrans t, MonadAlias (t m))

```

Despite the use of the **type** keyword, the identifier *EvalStrategy* actually has a kind *Constraint*, which means that it can be used to specify assumptions about types in a function signature. In our example, the constraint *EvalStrategy t IO* denotes monadic computations based on the *IO* monad that also provide an implementation of *MonadAlias*:

```

chooseSize :: EvalStrategy t IO  $\Rightarrow$  t IO Int  $\rightarrow$  t IO Int  $\rightarrow$  t IO Int
chooseSize new legacy = do
  newVal  $\leftarrow$  new
  if newVal > 0 then new else legacy
resultSize :: EvalStrategy t IO  $\Rightarrow$  t IO Int
resultSize = do
  new  $\leftarrow$  alias $ lift (lookupInput "new_size")
  legacy  $\leftarrow$  alias $ lift (lookupInput "legacy_size")
  chooseSize new legacy

```

Compared to the previous version of the example, the only significant change is in the type signature of the two functions. Instead of passing computations of type *IO a*, they now work with computations *t IO a*

²Constraint kinds are available in GHC 7.4 and generalize constraint families proposed by Orchard and Schrijvers [18]

with the constraint $EvalStrategy\ t\ IO$. The result of $lookupInput$ is still $IO\ Int$ and so it needs to be lifted into a monadic computation $t\ IO\ Int$ using the $lift$ function.

The return type of the $resultSize$ computation is parameterized over the evaluation strategy t . This means that we can call it in two different ways. Writing $runCbN\ resultSize$ executes the function using *call-by-name* and writing $runCbV\ resultSize$ executes it using *call-by-value*. However, it is also possible to implement a monad transformer for *call-by-need*.

4.2 Implementing call-by-need strategy

In his paper introducing the *call-by-name* and *call-by-value* translations for monads, Wadler noted: “It remains an open question whether there is a translation scheme that corresponds to call-by-need as opposed to call-by-name” [26]. We do not fully answer this question, however we hope to contribute to the answer. In particular, we show how to use the mechanism described so far to turn certain monads into an extended versions of such monads that provide the *call-by-need* behaviour.

In the absence of effects, the *call-by-need* strategy is equivalent to the *call-by-name* strategy, with the only difference in that performance characteristics. In the *call-by-need* (or *lazy*) strategy, a computation passed as an argument is evaluated at most once and the result is cached afterwards.

The caching of results needs to be done in the *malias* operation. This cannot be done for *any* monad, but we can define a monad transformer similar to the ones presented in the previous section. In particular, we use Svenningsson’s package [22], which defines a transformer version of the ST monad [11]. As documented in the package description, the transformer should be applied only to monads that yield a single result. Combining lazy evaluation with non-determinism is a more complex topic that has been explored by Fischer et al. [6].

```
newtype CbL s m a = CbL { unCbL :: STT s m a }
```

Unlike CbV and CbN , the CbL type is not a simple wrapper that contains a computation of type $m\ a$. Instead, it contains a computation augmented with some additional state. The state is used for caching the values of evaluated computations. The type $STT\ s\ m\ a$ represents a computation $m\ a$ with an additional local state “tagged” with a type variable s . The use of a local state instead of e.g. IO means that the monadic computation can be safely evaluated even as part of purely functional code. The tags are used merely to guarantee that state associated with one STT computation does not leak to other parts of the program.

Implementing the $Monad$ and $MonadTrans$ instances follows exactly the same pattern as instances of other transformers given in Figure 5. The interesting work is done in the *malias* function of $MonadAlias$:

```
instance Monad m => MonadAlias (CbL s m) where
  malias (CbL marg) = CbL $ do
    r ← newSTRef Nothing
    return (CbL $ do
      rv ← readSTRef r
      case rv of
        Nothing → marg >>= λv → writeSTRef r (Just v) >> return v
        Just v → return v)
```

The *malias* operation takes a computation of type $m\ a$ and returns a computation $m\ (m\ a)$. The monad transformer wraps the underlying monad inside STT , so the type of the computation returned by *malias* is equivalent to a type $STT\ s\ m\ (STT\ s\ m\ a)$.

The fact that both outer and inner *STT* share the same tag *s* means that they operate on a shared state. The outer computation allocates a new reference and the inner computation can use it to access and store the result computed previously. The allocation is done using the *newSTRef* function, which creates a reference initialized to *Nothing*. In the returned (inner) computation, we first read the state using *readSTRef*. If the value was computed previously, then it is simply returned. If not, the computation evaluates *marg*, stores the result in a reference cell and then returns the obtained value.

Discussion. After implementing *runCbL* function (which can be done easily using *runST*), the example in Section 4.1 can be executed using the *CbL* type. With the *call-by-need* semantics, the program finally behaves as desired: if the value of the "new_size" key is greater than zero, then it reads it only once, without reading the value of the "legacy_size" key. The value of "legacy_size" key is accessed only if the value of the "new_size" key is less than zero.

Showing that the above definition corresponds to *call-by-need* formally is beyond the scope of this paper. However, the use of *STT* transformer adds a shared state that keeps the evaluated values, which closely corresponds to Launchbury's environment-based semantics of lazy evaluation [9].

We do not formally prove that the *malias* laws hold for the above implementation, but we give an informal argument. The *naturality* law follows from parametricity. Computations considered in the *associativity* law are of type $m (m (m a))$; both sides of the equation create a computation where the two outer *m* computations allocate a new reference (where the outer points to the inner) and the single innermost *m* actually triggers the computation. The *computationality* law holds, because aliasing of a unit computation cannot introduce any effects. Finally, the left-hand side of *identity* creates a computation that allocates a new reference that is encapsulated in the returned computation and cannot be accessed from elsewhere, so no sharing is added when compared with the right-hand side.

4.3 Parallel call-by-need strategy

In this section, we consider yet another evaluation strategy that can be implemented using our scheme. The *parallel call-by-need* strategy [2] is similar to *call-by-need*, but it may optimistically start evaluating a computation sooner, in parallel with the main program. When carefully tuned, the evaluation strategy may result in a better performance on multi-core CPUs.

We present a simple implementation of the *malias* operation based on the monad for deterministic parallelism by Marlow et al. [14]. By translating purely functional code to a monadic version using our translation and the *Par* monad, we get a program that attempts to evaluate arguments of every function call in parallel. In practice, this may introduce too much overhead, but it demonstrate that *parallel call-by-need* strategy also fits with our general framework.

Unlike the previous two sections, we do not define a monad transformer that can embody any monadic computation. For example, performing IO operations in parallel might introduce non-determinism. Instead, we implement *malias* operation directly for computations of type *Par a*:

```
instance MonadAlias Par where
  malias m = spawn m >>= return ◦ get
```

The implementation is surprisingly simple. The function *spawn* creates a computation *Par (IVar a)* that starts the work in background and returns a mutable variable (I-structure) that will contain the result when the computation completes. The inner computation that is returned by *malias* calls a function *get :: IVar a → Par a* that waits until *IVar* has been assigned a value and then returns it.

Using the above implementation of *malias*, we can now translate purely functional code to a monadic version that uses *parallel call-by-need* instead of the previous standard evaluation strategies (that do not introduce any parallelism). For example, consider a naive Fibonacci function:

$$\begin{aligned} fibSeq\ n \mid n \leq 1 &= n \\ fibSeq\ n &= fibSeq\ (n - 1) + fibSeq\ (n - 2) \end{aligned}$$

The second case calls the $+$ operator with two computations as arguments. The translated version passes these computations to *malias*, which starts executing them in parallel. The monadic version of the $+$ operator then waits until both computations complete. If we translate only the second case and manually add a case that calls sequential version of the function for inputs smaller than 30, we get the following:

$$\begin{aligned} fibPar\ n \mid n < 30 &= return\ (fibSeq\ n) \\ fibPar\ n &= \mathbf{do} \\ &\quad n1 \leftarrow malias\ \$\ fibPar\ (n - 1) \\ &\quad n2 \leftarrow malias\ \$\ fibPar\ (n - 2) \\ &\quad liftM2\ (+)\ n1\ n2 \end{aligned}$$

Aside from the first line, the code directly follows the general translation mechanism described earlier. Arguments of a function are turned to monadic computations and passed to *malias*. The inner computations obtained using monadic bind are then passed to a translated function.

Thanks to the manual optimization that calls *fibSeq* for smaller inputs, the function runs nearly twice as fast on a dual-core machine³. Parallel programming is one of the first areas where we found the *malias* operation useful. We first considered it as part of *joinads*, which are discussed in the next section.

Discussion. Showing that the above implementation actually implements *parallel call-by-need* could be done by relating our implementation of *malias* to the multi-thread transitions of [2]. Informally, each computation that may be shared is added to the work queue (using *spawn*) when it occurs on the right-hand side of a let binding, or as an argument in function application. The queued work evaluates in parallel with the main program and the *get* function implements sharing, so the semantics is *lazy*.

As previously, the *naturality* law holds thanks to parametricity. To consider other laws, we need a formal model that captures the time needed to evaluate computations. We assume that evaluating a computation created by *unit* takes no time, but all other computations take non-zero time. Moreover, all spawned computations start executing immediately (i.e. the number of equally fast threads is unlimited).

In the *associativity* law, the left-hand side returns a computation that spawns the actual work and then spawns a computation that waits for its completion. The right-hand side returns a computation that schedules a computation, which then schedules the actual work. In both cases, the actual work is started immediately when the outer *m* computation is evaluated, and so they are equivalent. The *computationality* law holds, because a *unit* computation evaluates in no time. Finally, the left-hand side of *identity* returns a computation that spawns the work and then waits for its completion, which is semantically equivalent to just running the computation.

4.4 Simplifying Joinads

Joinads [20, 19] were designed to simplify programming with certain kinds of monadic computations. Many monads, especially from the area of concurrent or parallel programming provide additional oper-

³When called with input 37 on a Core 2 Duo CPU, the sequential version runs in 8.9s and the parallel version in 5.1s.

ations for composing monadic computations. Joinads identify three most common extensions: *parallel composition*, *(non-deterministic) choice* and *aliasing*.

Joinads are abstract computations that form a monad and provide the three additional operations mentioned above. The work on joinads also introduces a syntactic extension for Haskell and F# that makes it easier to work with these classes of computations. For example, the following snippet uses the *Par* monad to implement a function that tests whether a predicate holds for all leafs of a tree in parallel:

```

all :: (a → Bool) → Tree a → Par Bool
all p (Leaf v)           = return (p v)
all p (Node left right) =
  docase (all p left, all p right) of
    (False, ?) → return False
    (?, False) → return False
    (allL, allR) → return (allL ∧ allR)

```

The **docase** notation intentionally resembles pattern matching and has similar semantics as well. The first two cases use the special pattern *?* to denote that the value of one of the computations does not have to be available in order to continue. When one of the sub-branches returns *False*, we know that the overall result is *False* and so we return immediately. Finally, the last clause matches if neither of the two previous are matched. It can only match after both sub-trees are processed.

Similarly to **do** notation, the **docase** syntax is desugared into uses of the joinad operations. The choice between clauses is translated using the *choice* operator. If a clause requires the result of multiple computations (such as the last one), the computations are combined using *parallel composition*.

If a computation, passed as an argument, is accessed from multiple clauses, then it should be evaluated only once and the clauses should only access *aliased* computation. This motivation is similar to the one described in this article. Indeed, joinads use a variant of the *malias* operation and insert it automatically for all arguments of **docase**. This is very similar to how the translation presented in this paper uses *malias*. If aliasing was not done automatically behind the scenes, we would have to write:

```

all p (Leaf v)           = return (p v)
all p (Node left right) = do
  l ← malias (all p left)
  r ← malias (all p right)
  docase (l, r) of
    (False, ?) → return False
    (?, False) → return False
    (allL, allR) → return (allL ∧ allR)

```

One of the limitations of the original design of joinads is that there is no one-to-one correspondence between the **docase** notation and what can be expressed directly using the joinad operations. This is partly due to the automatic aliasing of arguments, which inserts *malias* only at very specific locations. We believe that integrating a *call-by-alias* translation, described in this article, in a programming language could resolve this situation. It would also separate the two concerns – composition of computations using *choice* and *parallel composition* (done by joinads) and automatic aliasing of computations that allows sharing of results as in *call-by-need* or *parallel call-by-need*.

5 Related work

Most of the work that directly influenced this work has been discussed throughout the paper. Most importantly, the question of translating pure code to a monadic version with *call-by-need* semantics was posed by Wadler [26]. To our knowledge, this question has not been answered before, but there is various work that either uses similar structures or considers evaluation strategies from different perspectives.

Monads with aliasing. Numerous monads have independently introduced an operation of type $m a \rightarrow m (m a)$. The *eagerly* combinator of the Orc monad [10] causes computation to run in parallel, effectively implementing the *parallel call-by-need* evaluation strategy. The only law that relates *eagerly* to operations of a monad is too restrictive to allow the *call-by-value* semantics.

A monad for purely functional lazy non-deterministic programming [6] uses a similar combinator *share* to make monadic (non-deterministic) computations lazy. However, the *call-by-value* strategy is inefficient and the *call-by-name* strategy is incorrect, because choosing a different value each time a non-deterministic computation is accessed means that *generate and test* pattern does not work.

The *share* operation is described together with the laws that should hold. The *HNF* law is similar to our *computationality*. However, the *Ignore* law specifies that the *share* operation should not be strict (ruling out our *call-by-value* implementation of *malias*). A related paper [4] discusses where *share* needs to be inserted when translating lazy non-deterministic programs to a monadic form. The results may be directly applicable to make our translation more efficient by inserting *malias* only when required.

Abstract computations and comonads. In this paper, we extended monads with one component of a comonadic structure. Although less widespread than monads, comonads are also useful for capturing abstract computations in functional programming. They have been used for dataflow programming [23], array programming [17], environment passing, and more [8]. In general, comonads can be used to describe context-dependent computations [24], where *cojoin* (natural transformation δ) duplicates the context. In our work, the corresponding operation *malias* splits the context (effects) in a particular way between two computations.

We only considered basic monadic computations, but it would be interesting to see how *malias* interacts with other abstract notions of computations, such as *applicative functors* [15], *arrows* [7] or additive monads (the *MonadPlus* type-class). The monad for lazy non-deterministic programming [6], mentioned earlier, implements *MonadPlus* and may thus provide interesting insights.

Evaluation strategies. One of the key results of this paper is a monadic translation from purely functional code to a monadic version that has the *call-by-need* semantics. We achieve that using the monad transformer [13] for adding state.

In the absence of effects, *call-by-need* is equivalent to *call-by-name*, but it has been described formally as a version of λ -calculus by Ariola and Felleisen [1]. This allows equational reasoning about computations and it could be used to show that our encoding directly corresponds to *call-by-need*, similarly to proofs for other strategies in Appendix A. The semantics has been also described using an environment that models caching [9, 21], which closely corresponds to our actual implementation.

Considering the two basic evaluation strategies, Wadler [27] shows that *call-by-name* is dual to *call-by-value*. We find this curious as the two definitions of *malias* in our work are, in some sense, also dual or symmetric as they associate all effects with the inner or the outer monad of the type $m (m a)$. Furthermore, the duality between *call-by-name* and *call-by-value* can be viewed from a logical perspective

thanks to the Curry-Howard correspondence. We believe that finding a similar logical perspective for our generalized strategy may be an interesting future work.

Finally, the work presented in this work unifies monadic *call-by-name* and *call-by-value*. In a non-monadic setting, a similar goal is achieved by the *call-by-push-value* calculus [12]. The calculus is more fine-grained and strictly separates *values* and *computations*. Using these mechanisms, it is possible to encode both *call-by-name* and *call-by-value*. It may be interesting to consider whether our computations parameterized over evaluation strategy (Section 4.1) could be encoded in *call-by-push-value*.

6 Conclusions

We presented an alternative translation from purely functional code to monadic form. Previously, this required choosing either the *call-by-need* or the *call-by-value* translation and the translated code had different structure and different types in both cases. Our translation abstracts the evaluation strategy into a function *malias* that can be implemented separately providing the required evaluation strategy.

Our translation is not limited to the above two evaluation strategies. Most interestingly, we show that certain monads can be automatically turned into an extended version that supports the *call-by-need* strategy. This answers part of an interesting open problem posed by Wadler [26]. The approach has other interesting applications – it makes it possible to write code that is parameterized by the evaluation strategy and it allows implementing a *parallel call-by-need* strategy for certain monads.

Finally, we presented the theoretical foundations of our approach using a model described in terms of category theory. We extended the monad structure with an additional operation based on *computational comonads*, which were previously used to give intensional semantics of computations. In our setting, the operation specifies the evaluation order. The categorical model specifies laws about *malias* and we proved that the laws hold for *call-by-value* and *call-by-name* strategies.

Acknowledgements. The author is grateful to Sebastian Fischer, Dominic Orchard and Alan Mycroft for inspiring comments and discussion, and to the latter two for proofreading the paper. We are also grateful to anonymous referees for detailed comments and to Paul Blain Levy for shepherding of the paper. The work has been partly supported by EPSRC and through the Cambridge CHESS scheme.

References

- [1] Zena M. Ariola & Matthias Felleisen (1997): *The Call-By-Need Lambda Calculus*. *Journal of Functional Programming* 7, pp. 265–301, doi:10.1017/S0956796897002724.
- [2] Clem Baker-Finch, David J. King & Phil Trinder (2000): *An Operational Semantics for Parallel Lazy Evaluation*. In: *Proceedings of ICFP 2000*, doi:10.1145/351240.351256.
- [3] Max Bolingbroke (2011): *Constraint Kinds for GHC (Unpublished manuscript)*. Available at <http://blog.omega-prime.co.uk/?p=127>.
- [4] B. Braßel, S. Fischer, M. Hanus & F. Reck (2011): *Transforming Functional Logic Programs into Monadic Functional Programs*. In: *Proceedings of WFLP 2010*, LNCS 6559, pp. 30–47, doi:10.1.1.157.4578.
- [5] Stephen Brookes & Shai Geva (1992): *Computational Comonads and Intensional Semantics*. In: *Applications of Categories in Computer Science: Proceedings of the LMS Symposium*, 177, pp. 1–44, doi:10.1.1.45.4952.
- [6] Sebastian Fischer, Oleg Kiselyov & Chung-chieh Shan (2011): *Purely Functional Lazy Non-Deterministic Programming*. *Journal of Functional Programming* 21(4–5), pp. 413–465, doi:10.1145/1631687.1596556.

- [7] John Hughes (1998): *Generalising Monads to Arrows*. *Science of Computer Programming* 37, pp. 67–111, doi:10.1016/S0167-6423(99)00023-4.
- [8] Richard B. Kieburtz (1999): *Codata and Comonads in Haskell (Unpublished manuscript)*.
- [9] John Launchbury (1993): *A Natural Semantics for Lazy Evaluation*. In: *Proceedings of POPL 1993*, pp. 144–154, doi:10.1145/158511.158618.
- [10] John Launchbury & Trevor Elliott (2010): *Concurrent orchestration in Haskell*. In: *Proceedings of Haskell Symposium, Haskell 2010*, pp. 79–90, doi:10.1145/1863523.1863534.
- [11] John Launchbury & Simon Peyton Jones (1995): *State in Haskell*. In: *Proceedings of the LISP and Symbolic Computation Conference, LISP 1995*, doi:10.1007/BF01018827.
- [12] Paul Blain Levy (2004): *Call-By-Push-Value*. Springer. ISBN: 978-1-4020-1730-8.
- [13] Sheng Liang, Paul Hudak & Mark Jones (1995): *Monad Transformers and Modular Interpreters*. In: *Proceedings of POPL 1995*, doi:10.1145/199448.199528.
- [14] Simon Marlow, Ryan Newton & Simon Peyton Jones (2011): *A Monad for Deterministic Parallelism*. In: *Proceedings of the 4th Symposium on Haskell, Haskell 2011*, doi:10.1145/2034675.2034685.
- [15] Conor McBride & Ross Paterson (2008): *Applicative Programming with Effects*. *Journal of Functional Programming* 18, pp. 1–13, doi:10.1017/S0956796807006326.
- [16] Eugenio Moggi (1991): *Notions of Computation and Monads*. *Inf. Comput.* 93, pp. 55–92, doi:10.1016/0890-5401(91)90052-4.
- [17] Dominic Orchard, Max Bolingbroke & Alan Mycroft (2010): *Ygnos: Declarative, Parallel Structured Grid Programming*. In: *Proceedings of DAMP 2010*, pp. 15–24, doi:10.1145/1708046.1708053.
- [18] Dominic A. Orchard & Tom Schrijvers (2010): *Haskell Type Constraints Unleashed*. In: *Proceedings of FLOPS 2010*, pp. 56–71, doi:10.1007/978-3-642-12251-4_6.
- [19] Tomas Petricek, Alan Mycroft & Don Syme: *Extending Monads with Pattern Matching*. In: *Proceedings of Haskell Symposium, Haskell 2011*, doi:10.1145/2096148.2034677.
- [20] Tomas Petricek & Don Syme (2011): *Joinads: A Retargetable Control-Flow Construct for Reactive, Parallel and Concurrent Programming*. In: *Proceedings of PADL 2011*, doi:10.1007/978-3-642-18378-2_17.
- [21] S. Purushothaman & Jill Seaman (1992): *An Adequate Operational Semantics for Sharing in Lazy Evaluation*. In: *Proceedings of the 4th European Symposium on Programming*, doi:10.1007/3-540-55253-7_26.
- [22] Josef Svenningsson (2011): *The STMonadTrans package*. Available at <http://hackage.haskell.org/package/STMonadTrans-0.3.1>.
- [23] Tarmo Uustalu & Varmo Vene (2006): *The Essence of Dataflow Programming*. *Lecture Notes in Computer Science* 4164, pp. 135–167, doi:10.1007/11894100_5.
- [24] Tarmo Uustalu & Varmo Vene (2008): *Comonadic Notions of Computation*. *Electron. Notes Theor. Comput. Sci.* 203, pp. 263–284, doi:10.1016/j.entcs.2008.05.029.
- [25] Janis Voigtländer (2009): *Free Theorems Involving Type Constructor Classes: Functional Pearl*. In: *Proceedings of ICFP 2009, ICFP 209, ACM, New York, NY, USA*, pp. 173–184, doi:10.1145/1596550.1596577.
- [26] Philip Wadler (1990): *Comprehending Monads*. In: *Proceedings of Conference on LISP and Functional Programming, LFP 1990, ACM, New York, NY, USA*, pp. 61–78, doi:10.1145/91556.91592.
- [27] Philip Wadler (2003): *Call-By-Value Is Dual to Call-By-Name*. In: *Proceedings of ICFP, ICFP 2003*, pp. 189–201, doi:10.1145/944705.944723.

$$\begin{aligned}
& \llbracket e_1 e_2 \rrbracket_{cba} \\
&= \text{bind } \llbracket e_1 \rrbracket_{cba} (\lambda f. \text{bind } (\text{malias } \llbracket e_2 \rrbracket_{cba}) f) && \text{(translation)} \\
&= \text{bind } \llbracket e_1 \rrbracket_{cba} (\lambda f. \text{bind } (\text{unit } \llbracket e_2 \rrbracket_{cba}) f) && \text{(definition)} \\
&= \text{bind } \llbracket e_1 \rrbracket_{cba} (\lambda f. f \llbracket e_2 \rrbracket_{cba}) && \text{(left identity)} \\
&= \text{bind } \llbracket e_1 \rrbracket_{cbn} (\lambda f. f \llbracket e_2 \rrbracket_{cbn}) && \text{(induction hypothesis)} \\
&= \llbracket e_1 e_2 \rrbracket_{cbn} && \text{(translation)} \\
\\
& \llbracket \text{let } x = e_1 \text{ in } e_2 \rrbracket_{cba} \\
&= \text{bind } (\text{malias } \llbracket e_1 \rrbracket_{cba}) (\lambda x. \llbracket e_2 \rrbracket_{cba}) && \text{(translation)} \\
&= \text{bind } (\text{unit } \llbracket e_1 \rrbracket_{cba}) (\lambda x. \llbracket e_2 \rrbracket_{cba}) && \text{(definition)} \\
&= (\lambda x. \llbracket e_2 \rrbracket_{cba}) \llbracket e_1 \rrbracket_{cba} && \text{(left identity)} \\
&= (\lambda x. \llbracket e_2 \rrbracket_{cbn}) \llbracket e_1 \rrbracket_{cbn} && \text{(induction hypothesis)} \\
&= \llbracket \text{let } x = e_1 \text{ in } e_2 \rrbracket_{cbn} && \text{(translation)}
\end{aligned}$$

Figure 6: Proving that using an appropriate *malias* is equivalent to *call-by-name*.

A Equivalence proofs

In this section, we show that our translation presented in Section 2.3 can be used to implement standard *call-by-name* and *call-by-value*. We prove that using an appropriate definition of *malias* from Section 2.1 gives the same semantics as the standard translations described by Wadler [26].

Call-by-name. The translation of types is the same for our translation and the *call-by-name* translation. In addition, the rules for translating variable access and lambda functions are also the same. This means that we only need to prove that our translation of let-binding and function application are equivalent. When implementing *call-by-name* using our translation, we use the following definition of *malias*:

$$\text{malias } m = \text{unit } m$$

Using this definition and the *left identity* monad law, we can now show that our *call-by-alias* translation is equivalent to the translation from Figure 2. The Figure 6 shows the equations for function application and let-binding.

Call-by-value. Proving that appropriate definition of *malias* gives a term that corresponds to the one obtained using standard *call-by-value* translation is more difficult. In the *call-by-value* translation, functions are translated to a type $\tau_1 \rightarrow M \tau_2$, while our translation produces functions of type $M \tau_1 \rightarrow M \tau_2$. As a reminder, the definition of *malias* that gives the *call-by-value* behaviour is the following:

$$\text{malias } m = \text{bind } m (\text{unit } \circ \text{unit})$$

To prove that the two translations are equivalent, we show that the following invariant holds: when using the above definition of *malias* and our *call-by-alias* translation, a monadic computation of type $M \tau$ that is assigned to a variable x always has a structure $\text{unit } x_v$ where x_v is a variable of type τ .

The sketch of the proof is shown in Figure 7. We write $e_1 \cong e_2$ to mean that the expression e_1 in the *call-by-alias* translation corresponds to an expression e_2 in *call-by-value* translation. This means that expressions of form $\text{unit } x_v$ translate to values x , variable declarations of x_v (in lambda abstraction and let-binding) translate to declarations of x and all function values of type $M \tau_1 \rightarrow M \tau_2$ become $\tau_1 \rightarrow M \tau_2$.

$$\begin{aligned}
\llbracket x \rrbracket_{cba} &= \mathit{unit} \ x_v && \text{(assumption)} \\
&\cong x && \text{(correspondence)} \\
&= \llbracket x \rrbracket_{cbv} && \text{(translation)} \\
\\
\llbracket e_1 \ e_2 \rrbracket_{cba} &= \mathit{bind} \ \llbracket e_1 \rrbracket_{cba} \ (\lambda f. \mathit{bind} \ (\mathit{malias} \ \llbracket e_2 \rrbracket_{cba}) \ f) && \text{(translation)} \\
&= \mathit{bind} \ \llbracket e_1 \rrbracket_{cba} \ (\lambda f. \mathit{bind} \ (\mathit{bind} \ \llbracket e_2 \rrbracket_{cba} \ (\mathit{unit} \circ \mathit{unit})) \ f) && \text{(definition)} \\
&= \mathit{bind} \ \llbracket e_1 \rrbracket_{cba} \ (\lambda f. \mathit{bind} \ \llbracket e_2 \rrbracket_{cba} \ (\lambda x_v. \mathit{bind} \ (\mathit{unit} \ (\mathit{unit} \ x_v)) \ f)) && \text{(associativity)} \\
&= \mathit{bind} \ \llbracket e_1 \rrbracket_{cba} \ (\lambda f. \mathit{bind} \ \llbracket e_2 \rrbracket_{cba} \ (\lambda x_v. f \ (\mathit{unit} \ x_v))) && \text{(left identity)} \\
&\cong \mathit{bind} \ \llbracket e_1 \rrbracket_{cbv} \ (\lambda f. \mathit{bind} \ \llbracket e_2 \rrbracket_{cbv} \ (\lambda x. f \ x)) && \text{(correspondence)} \\
&= \llbracket e_1 \ e_2 \rrbracket_{cbv} && \text{(translation)} \\
\\
\llbracket \mathit{let} \ x = e_1 \ \mathit{in} \ e_2 \rrbracket_{cba} &= \mathit{bind} \ (\mathit{malias} \ \llbracket e_1 \rrbracket_{cba}) \ (\lambda x. \llbracket e_2 \rrbracket_{cba}) && \text{(translation)} \\
&= \mathit{bind} \ (\mathit{bind} \ \llbracket e_1 \rrbracket_{cba} \ (\mathit{unit} \circ \mathit{unit})) \ (\lambda x. \llbracket e_2 \rrbracket_{cba}) && \text{(definition)} \\
&= \mathit{bind} \ \llbracket e_1 \rrbracket_{cba} \ (\lambda x_v. \mathit{bind} \ (\mathit{unit} \ (\mathit{unit} \ x_v)) \ (\lambda x. \llbracket e_2 \rrbracket_{cba})) && \text{(associativity)} \\
&= \mathit{bind} \ \llbracket e_1 \rrbracket_{cba} \ (\lambda x_v. (\lambda x. \llbracket e_2 \rrbracket_{cba}) \ (\mathit{unit} \ x_v)) && \text{(left identity)} \\
&\cong \mathit{bind} \ \llbracket e_1 \rrbracket_{cbv} \ (\lambda x. \llbracket e_2 \rrbracket_{cbv}) && \text{(correspondence)} \\
&= \llbracket \mathit{let} \ x = e_1 \ \mathit{in} \ e_2 \rrbracket_{cbv} && \text{(translation)}
\end{aligned}$$

Figure 7: Proving that using an appropriate *malias* is equivalent to *call-by-value*.

B Proofs for two implementations

In this section, we prove that the two implementations of *malias* presented in Section 2.1 obey the *malias* laws. We use the formulation of monads based on a functor with additional operations *join* and *unit*. The proof relies on the following laws that hold about *join* and *unit*:

$$\begin{aligned}
\mathit{map} \ (g \circ f) &= (\mathit{map} \ g) \circ (\mathit{map} \ f) && \text{(functor)} \\
\mathit{unit} \circ f &= \mathit{map} \ f \circ \mathit{unit} && \text{(natural unit)} \\
\mathit{join} \circ \mathit{map} \ (\mathit{map} \ f) &= \mathit{map} \ f \circ \mathit{join} && \text{(natural join)} \\
\mathit{join} \circ \mathit{map} \ \mathit{join} &= \mathit{join} \circ \mathit{join} && \text{(assoc join)} \\
\mathit{join} \circ \mathit{unit} &= \mathit{join} \circ \mathit{map} \ \mathit{unit} = \mathit{id} && \text{(identity)}
\end{aligned}$$

The first law follows from the fact that *map* corresponds to a functor. The next two laws hold because *unit* and *join* are both natural transformations. Finally, the last two laws are additional laws that are required to hold about monads (a precise definition can be found in Section 3).

The proofs that the two definitions of *malias* (implementing *call-by-name* and *call-by-value* strategies) are correct can be found in Figure 8 and Figure 9, respectively. The proofs use the above monad laws. We use a definition $\mathit{malias} \equiv \mathit{map} \ \mathit{unit}$, which is equivalent to the definition in Appendix A. The figures include proofs for the four additional *malias* laws as defined in Section 2.2: *naturality*, *associativity*, *computationality* and *identity*.

$$\begin{aligned}
& \mathit{map} (\mathit{map} f) \circ \mathit{malias} \\
&= \mathit{map} (\mathit{map} f) \circ \mathit{unit} && \text{(definition)} \\
&= \mathit{unit} \circ \mathit{map} f && \text{(naturality)} \\
&= \mathit{malias} \circ \mathit{map} f && \text{(definition)} \\
\\
& \mathit{map} \mathit{malias} \circ \mathit{malias} \\
&= \mathit{map} \mathit{unit} \circ \mathit{unit} && \text{(definition)} \\
&= \mathit{unit} \circ \mathit{unit} && \text{(natural unit)} \\
&= \mathit{malias} \circ \mathit{malias} && \text{(definition)} \\
\\
& \mathit{malias} \circ \mathit{unit} \\
&= \mathit{unit} \circ \mathit{unit} && \text{(definition)} \\
\\
& \mathit{join} \circ \mathit{malias} \\
&= \mathit{join} \circ \mathit{unit} && \text{(definition)} \\
&= \mathit{id} && \text{(identity)}
\end{aligned}$$

Figure 8: Call-by-name definition of *malias* obeys the laws

$$\begin{aligned}
& \mathit{map} (\mathit{map} f) \circ \mathit{malias} \\
&= \mathit{map} (\mathit{map} f) \circ \mathit{map} \mathit{unit} && \text{(definition)} \\
&= \mathit{map} (\mathit{map} f \circ \mathit{unit}) && \text{(functor)} \\
&= \mathit{map} (\mathit{unit} \circ f) && \text{(natural unit)} \\
&= \mathit{map} \mathit{unit} \circ \mathit{map} f && \text{(functor)} \\
&= \mathit{malias} \circ \mathit{map} f && \text{(definition)} \\
\\
& \mathit{map} \mathit{malias} \circ \mathit{malias} \\
&= \mathit{map} (\mathit{map} \mathit{unit}) \circ (\mathit{map} \mathit{unit}) && \text{(definition)} \\
&= \mathit{map} (\mathit{map} \mathit{unit} \circ \mathit{unit}) && \text{(functor)} \\
&= \mathit{map} (\mathit{unit} \circ \mathit{unit}) && \text{(natural unit)} \\
&= \mathit{map} \mathit{unit} \circ \mathit{map} \mathit{unit} && \text{(functor)} \\
&= \mathit{malias} \circ \mathit{malias} && \text{(definition)} \\
\\
& \mathit{malias} \circ \mathit{unit} \\
&= \mathit{map} \mathit{unit} \circ \mathit{unit} && \text{(definition)} \\
&= \mathit{unit} \circ \mathit{unit} && \text{(natural unit)} \\
\\
& \mathit{join} \circ \mathit{malias} \\
&= \mathit{join} \circ \mathit{map} \mathit{unit} && \text{(definition)} \\
&= \mathit{id} && \text{(identity)}
\end{aligned}$$

Figure 9: Call-by-value definition of *malias* obeys the laws

C Typing preservation proof

In this section, we show that our translation preserves typing. Given a well-typed term e of type τ , the translated term $\llbracket e \rrbracket_{cba}$ is also well-typed and has a type $\llbracket \tau \rrbracket_{cba}$. In the rest of this section, we write $\llbracket - \rrbracket$ for $\llbracket - \rrbracket_{cba}$. To show that the property holds, we use induction over the typing rules, using the fact that $\llbracket \tau \rrbracket = M \tau'$ for some τ' . The inductive construction of the typing derivation follows the following rules:

$$\begin{array}{c}
\frac{}{\llbracket \Gamma, x : \tau \vdash x : \tau \rrbracket} = \frac{}{\llbracket \Gamma \rrbracket, x : M \llbracket \tau \rrbracket \vdash x : M \llbracket \tau \rrbracket} \\
\frac{\llbracket \Gamma, x : \tau_1 \vdash e : \tau_2 \rrbracket}{\llbracket \Gamma \vdash \lambda x. e : \tau_1 \rightarrow \tau_2 \rrbracket} = \frac{\llbracket \Gamma \rrbracket, x : M \llbracket \tau_1 \rrbracket \vdash \llbracket e \rrbracket : M \llbracket \tau_2 \rrbracket}{\llbracket \Gamma \rrbracket \vdash \mathit{unit} (\lambda x. \llbracket e \rrbracket) : M (M \llbracket \tau_1 \rrbracket \rightarrow M \llbracket \tau_2 \rrbracket)} \\
\frac{\llbracket \Gamma \vdash e_1 : \tau_1 \rightarrow \tau_2 \rrbracket \quad \llbracket \Gamma \vdash e_2 : \tau_1 \rrbracket}{\llbracket \Gamma \vdash e_1 e_2 : \tau_2 \rrbracket} = \frac{\llbracket \Gamma \rrbracket \vdash \llbracket e_1 \rrbracket : M (M \llbracket \tau_1 \rrbracket \rightarrow M \llbracket \tau_2 \rrbracket) \quad \Gamma \vdash \llbracket e_2 \rrbracket : M \llbracket \tau_1 \rrbracket}{\Gamma \vdash \mathit{bind} \llbracket e_1 \rrbracket (\lambda f. \mathit{bind} (\mathit{malias} \llbracket e_2 \rrbracket) f) : M \llbracket \tau_2 \rrbracket} \\
\frac{\llbracket \Gamma \vdash e_1 : \tau_1 \rrbracket \quad \llbracket \Gamma, x : \tau_1 \vdash e_2 : \tau_2 \rrbracket}{\llbracket \Gamma \vdash \mathbf{let} x = e_1 \mathbf{in} e_2 : \tau_2 \rrbracket} = \frac{\llbracket \Gamma \rrbracket \vdash \llbracket e_1 \rrbracket : M \llbracket \tau_1 \rrbracket \quad \llbracket \Gamma \rrbracket, x : M \llbracket \tau_1 \rrbracket \vdash \llbracket e_2 \rrbracket : M \llbracket \tau_2 \rrbracket}{\llbracket \Gamma \rrbracket \vdash \mathit{bind} (\mathit{malias} \llbracket e_1 \rrbracket) (\lambda x. \llbracket e_2 \rrbracket) : M \llbracket \tau_2 \rrbracket}
\end{array}$$