Abstract

Compilers for ML and Haskell use intermediate languages that incorporate deeply-embedded assumptions about order of evaluation and side effects. We propose an intermediate language into which one can compile both ML and Haskell, thereby facilitating the sharing of ideas and infrastructure, and supporting language developments that move each language in the direction of the other. Achieving this goal without compromising the ability to compile as good code as a more direct route turned out to be much more subtle than we expected. We address this challenge using monads and unpointed types, identify two alternative language designs, and explore the choices they embody.

1 Introduction

Functional programmers are typically split into two camps: the strict (or call-by-value) camp, and the lazy (or call-by-need) camp. As the discipline has matured, though, each camp has come more and more to recognise the merits of the other, and to recognise the huge areas of common interest. It is hard, these days, to find anyone who believes that business is never useful, or that strictness is always bad. While there are still pervasive stylistic differences between strict and lazy programming, it is now often possible to adopt lazy evaluation at particular places in a strict language (Okasaki [1996]), or strict evaluation at particular points in a lazy one (for example, Haskell’s strictness annotations (Peterson et al. [1997])).

This rapprochement has not yet, however, propagated to our implementations. The insides of an ML compiler look pervasively different to those of a Haskell compiler. Notably, sequencing and support for side effects and exceptions are usually implicit in an ML compiler’s intermediate language (IL), but explicit (where they occur) in a Haskell compiler [Launchbury & Peyton Jones [1995]]. On the other hand, thunk formation and forcing are implicit in a Haskell compiler’s intermediate language, but explicit in an ML compiler. These pervasive differences make it impossible to share code, and hard to share results and analyses, between the two styles.

To say that “support for side effects is implicit in an ML compiler’s IL” (for example) is not to say that an ML compiler will take no notice of side effects; on the contrary, an ML compiler might well perform a global analysis that identifies pure sub-expressions (though in practice few do). However, one might wonder whether the analysis would discover all the pure sub-expressions in a Haskell program translated into the IL. In the same way, if an ML program were translated into a Haskell compiler’s IL, the latter might not discover all the occasions in which a function argument was guaranteed to be already evaluated. This thought motivates the following question: could we design a common compiler intermediate language (IL) that would serve equally well for both strict and lazy languages? The purpose of this paper is to explore the design space for just such a language.

We restrict our attention to higher order, polymorphically typed intermediate languages. There is considerable interest at the moment in type-directed compilation for polymorphic languages, in which type information is maintained accurately right through compilation and even on to run time (Harper & Morrisett [1995]; Shao & Appel [1995]; Tarditi et al. [1996]). Hence we focus on higher order, statically typed source languages, represented in this paper by ML (Milner [1991]) and Haskell (Peterson et al. [1997]).

At first we expected the design to be relatively straightforward, but we discovered that it was not. In particular, making sure that the IL has good operational properties for both strict and lazy languages turns out to be rather subtle. Identifying these subtleties is the main contribution of the paper:

- We employ monads to express and delimit state, input/output, and exceptions (Section 3). Using monads in this way is now well known to theorists (Moggi [1991]) and to language designers (Launchbury & Peyton Jones [1995]; Peyton Jones & Wadler [1993]; Wadler [1992a]), but, with one exception1, no compiler that we know has monads built into its intermediate language.

- We employ unpointed types to express the idea that an expression cannot diverge (Section 3.1). We show that the straightforward use of unpointed types does not lead to a good implementation (Section 3.6). This leads us to explore two distinct language designs. The first, \( \mathcal{L}_1 \), is mathematically simple, but cannot be compiled well (Section 3). An alternative design, \( \mathcal{L}_2 \), adds operational significance to unpointed types, by guaranteeing that a variable of unpointed type is evaluated (Section 4); this means \( \mathcal{L}_2 \) can be compiled well, but weakens its theory.

- We identify an interaction between unpointed types, polymorphism, and recursion in \( \mathcal{L}_1 \) (Section 3.5). Interestingly, the problem turns out to be more easily solved in \( \mathcal{L}_2 \) than \( \mathcal{L}_1 \) (Section 4.2).

---

1Personal communication, Nick Benton, Persimmon IT Ltd. 1997.
None of these ingredients are new. Our contribution is to explore the interactions of mixing them together. We emerge with the core of a practical IL that has something to offer both the strict and lazy community in isolation, as well as offering them a common framework. Our long-term goal is to establish an intermediate language that will enable the two communities to share both ideas (analyses, transformations) and systems (optimisers, code generators, run-time systems, profilers, etc) more effectively than hitherto.

2 The ground rules

We seek an intermediate language (IL) with the following properties:

- It must be possible to translate both (core) ML and Haskell into the IL. Extensions that add laziness to ML, or strictness to Haskell, should be readily incorporated. We make no attempt to treat ML’s module system, though that would be a desirable extension.

- In order to accommodate ML and Haskell the IL’s type system must support polymorphism. This ground rule turns out to have very significant, and rather unfortunate, impact upon our language designs (Section 3.5), but it seems quite essential. Nearly all existing compilers generate polymorphic target code, and although researchers have experimented with compiling away polymorphism by type specialisation (Jones [1994]; Tolmach & Oliva [1997]), problems with separate compilation and potential code explosion remain unresolved.

- The IL should be explicitly typed (Harper & Mitchell [1993]). We have in mind a variant of System F (Girard [1990]), with its explicit type abstractions and applications. The expressiveness of System F really is required. For example, there are several reasons for wanting polymorphic arguments to functions: the translation of Haskell type class constructors “dictionaries” with polymorphic components; we would like to be able to simulate modules using records (Jones [1996]); rank-2 polymorphism is required to express encapsulated state (Launchbury & Peyton Jones [1995]); and data-structure fusion (Gill, Launchbury & Peyton Jones [1993]).

- The IL programs can readily be type-checked, but there is no requirement that one could infer types from a type-erased IL program.

- The IL should have a single well-defined semantics. On the face of it, compilers for both strict and lazy languages already use a common language, namely the lambda calculus. But this similarity is only at the level of syntax; the semantics of the two calculi differ considerably. In particular, the code generator from a strict-language compiler would be completely unusable in a lazy-language compiler, and vice versa. Our goal is to have a single, neutral, semantics and hence a single optimiser and code generator.

- ML (or Haskell) programs thus compiled should be as efficient as those compiled by a good ML (resp. Haskell) compiler. In other words, compiling through the common IL should not impose any unavoidable efficiency penalty, either by way of loss of transformations (especially when starting from Haskell) or by way of a less efficient basic evaluation model (especially when starting from ML). Indeed, our hope is that we may ultimately be able to generate better code through this new route.

3 $L_1$, a totally explicit language

It is clear that the IL must be explicit about things that are implicit in “traditional” compiler ILs. Where are these implicit aspects of a “traditional” IL currently made explicit? Answer: in the denotational semantics of the IL. For example, the denotational semantics of a call-by-value lambda calculus looks something like this\(^3\)

$$\mathcal{E}[e_1, e_2]_\rho = (\mathcal{E}[e_1]_\rho) \ b, \ \text{if } a = b \perp,$$

$$\perp, \ \text{if } a = \perp$$

Here, the two cases in the right-hand side deal with the possible non-termination of the argument. What is implicit in the IL – the evaluation of the argument. in this case – becomes explicit in the semantics. An obvious suggestion is therefore to make the IL reflect the denotational semantics of the source language directly, so that everything is explicit in the IL, and nothing remains to be explicated by the semantics. This is our first design, $L_1$.

Figure 1 gives the syntax and type rules for $L_1$. We note the following features:

- As a compromise in the interest of brevity all our formal material describes only a simply-typed calculus, although supporting polymorphism is one of our ground rules. The extensions to add polymorphism, complete with explicit type abstractions and applications in the term language, are fairly standard (Harper & Mitchell [1993]; Peyton Jones [1996]; Tarditi et al. [1996]). However polymorphism adds some extra complications (Section 3.5.3.6).

- We omit recursive data types, constructors, and case expressions for the sake of simplicity, being content with pairs and selectors.

- let is simply very convenient syntactic sugar. It is not there to introduce polymorphism, even in the polymorphic extension of the language; explicit typing removes this motivation for let.

- letrec introduces recursion. Though we only give it one binding here, our intention is that it should accommodate multiple bindings. We use it rather than a constant fix because the latter requires heavy encoding for mutual recursion that is not reflected in an implementation. We discuss recursion in detail in Section 3.5, including the unspecified side condition mentioned in the rule.

- Following Moggi [1991], we express “computational effects” — such as non-termination, assignment, exceptions, and input/output — in monadic form. The type $M \tau$ is the type of $M$-computations returning a value of type $\tau$, where $M$ is drawn from a fixed family of monads. The syntactic forms let$_M$ and ret$_M$ are

\(^3\)We use the following standard notation. If $T$ is a complete partial order (CPO), then the CPO $T_\perp$, pronounced “$T$ lifted”, is defined thus: $T_\perp = \{a_\perp \mid a \in T\} \cup \{\perp\}$, with the obvious ordering.
The semantics of the two monads is given by their bind and unit functions. From the semantics one can prove that both $\beta$ and $\eta$ are valid with respect to the semantics, and that monadic expressions admit a number of standard transformations, given in Figure 3.

Figure 2 gives the semantics of $\mathcal{L}_1$. The semantic function $T$ gives the meaning of types. If it looks somewhat boring, that is the point! The function arrow in $\mathcal{L}_1$ is interpreted by function arrow in the underlying category of complete partial orders ($\mathit{CPO}$). Product is interpreted by (categorical i.e. un-lifted) product, and integers are interpreted by the integers. (If $\mathcal{L}_1$ were expanded to have sum types, they would be interpreted by (categorical, separated) sums.) Lastly, each monad is specified by an interpretation. The monad of lifting is interpreted by lifting, while a state transformer is interpreted by a function from the current “state” to a result and the new state. The “state” is a finite mapping from location identifiers (modeled by the natural numbers, $\mathcal{N}$) to their contents.

The semantic function $E$ gives the meaning of expressions. Again, many of its equations are rather dull: application is interpreted by application in the underlying category, lambda abstraction by functional abstraction, and so on. The semantics of the two monads is given by their bind and unit functions. From the semantics one can prove that both $\beta$ and $\eta$ are valid with respect to the semantics, and that monadic expressions admit a number of standard transformations, given in Figure 3.

Figure 2: Semantics of $\mathcal{L}_1$
3.1 Termination and non-termination

As we have mentioned, the interpretation of a type in $\mathcal{L}_1$ is a complete partial order (CPO). However, the interpretation of a type is not necessarily a pointed CPO; that is, the CPO does not necessarily contain a bottom element. For example, the data type of integers, $\text{Int}$, is interpreted by the unpointed CPO of integers, $\mathcal{Z}$. That is, if an expression has type $\text{Int}$, then it denotes an integer, and cannot denote a non-terminating computation. How, then, do we express the type of possibly-terminating integer-valued computations? As we have seen, $\mathcal{L}_1$ has an explicit type constructor for each monadic (i.e., computation) type, of which lifting is one. To express the type of a possibly-terminating integer we use the lifting monad. A possibly-diverging integer-valued expression therefore has type $\text{Lift Int}$.

So $\mathcal{L}_1$’s type system can distinguish surely-terminating expressions from possibly-diverging ones. The main reason for making this distinction in the type system is so that we can express the idea that a function takes an evaluated argument. The $\mathcal{L}_1$ lambda abstraction $\lambda x:\text{Int}.e$ expresses that $x$ cannot possibly be $\bot$, and so is a suitable translation of a lambda abstraction from a call-by-value language. On the other hand $\lambda x:\text{Lift Int}.e$ expresses that $x$ might perhaps be $\bot$, which fits a call-by-name or call-by-need language.

A second motivation for distinguishing pointed types from unpointed ones is that some useful program transformations that are not valid in general, hold unconditionally when one has more control over pointedness. Several researchers have explored languages that employ a distinction between pointed and unpointed types (Howard [1996]; Launchbury & Paterson [1996]), and others have explored pure languages without pointed types altogether (Cockett & Fukushima [1992]; Hagino [1987]; Turner [1995]). The presence of unpointed types has consequences for recursion, as we discuss in Section 3.5.

3.2 Stateful computations

In a similar way, we use the $\text{ST}$ monad to express in the type system the distinction between pure and stateful computations. For example, an expression of type $\text{Lift Int}$ denotes a pure (side-effect free), albeit possibly-divergent, computation; on the other hand, and expression of type $\text{ST Int}$ denotes a computation that might diverge, or might perform some side effects on a global state and deliver an integer. Further monads can readily be added to model exceptions, or continuations, or input/output.

This use of monads is well known. Moggi pioneered the idea of using monads to encapsulate computations (Moggi [1991]; Wadler [1992a]). The lazy functional programming

\[
\begin{align*}
\text{(M1)} & \quad \text{let}_M x \leftarrow \text{ret}_M e \text{ in } b = \text{let } x \mapsto e \text{ in } b \\
\text{(M2)} & \quad \text{let}_M x \leftarrow (\text{let}_M y \leftarrow e_1 \text{ in } e_2) \text{ in } b = \text{let } y \mapsto e_1 \text{ in } (\text{let}_M x \leftarrow e_2 \text{ in } b) \quad y \notin fV(b) \\
\text{(M3)} & \quad \text{let}_M x \leftarrow (\text{let}_M y \leftarrow e_1 \text{ in } e_2) \text{ in } b = \text{let } y \mapsto e_1 \text{ in } (\text{let}_M x \leftarrow e_2 \text{ in } b) \quad y \notin fV(b) \\
\text{(M4)} & \quad \text{let}_M x \leftarrow (\text{letrec } y \leftarrow e_1 \text{ in } e_2) \text{ in } b = \text{letrec } y \mapsto e_1 \text{ in } (\text{let}_M x \leftarrow e_2 \text{ in } b) \quad y \notin fV(b) \\
\text{(M5)} & \quad \text{let}_M x \leftarrow e \text{ in } \text{ret}_M x = e \\
\text{(M6)} & \quad \text{let } x \mapsto e \text{ in } \text{ret}_M b = \text{ret}_M (\text{let } x \mapsto e \text{ in } b)
\end{align*}
\]

Figure 3: Monad transformations

<table>
<thead>
<tr>
<th>Types $S, T$</th>
<th>$::=\text{Int} \mid \emptyset \mid S \times T \mid S \rightarrow T \mid \text{Ref S}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haskell only</td>
<td>$\text{ST } S$</td>
</tr>
<tr>
<td>Terms $M, N$</td>
<td>$::= x \mid i \mid M \mid N \mid \lambda x:T.M \mid M + N$</td>
</tr>
<tr>
<td>Haskell only</td>
<td>$\text{let } x:T \leftarrow M \text{ in } N$</td>
</tr>
<tr>
<td>Haskell only</td>
<td>$\text{let } M \mid \text{pair } M \mid \text{fst } M \mid \text{snd } M$</td>
</tr>
<tr>
<td>Haskell only</td>
<td>$\text{new } M \mid \text{rd } M \mid \text{wr } M \mid N$</td>
</tr>
<tr>
<td>Haskell only</td>
<td>$\text{let}S_T x:T \leftarrow M \text{ in } N$</td>
</tr>
</tbody>
</table>

Figure 4: Syntax of $S$

community has been using monads very effectively to isolate and encapsulate stateful computations and input/output within pure, lazy programs (Launchbury & Peyton Jones [1995]; Peyton Jones, Gordon & Finne [1996]; Peyton Jones & Wadler [1993]; Wadler [1992b]). Nevertheless, there are surprisingly subtle design choices to make, as we discuss in Section 3.4.

3.3 Translating ML and Haskell into $\mathcal{L}_1$

Before discussing its design any further, we first emphasise $\mathcal{L}_1$’s role as a target for both strict, stateful, and pure, lazy languages by giving translations from both into $\mathcal{L}_1$. Figure 4 gives the syntax of a tiny generic source language, $S$. We regard $S$ as a prototype for either ML or Haskell, by giving it a strict or lazy interpretation respectively. In either case, $S$ is assumed to have been explicitly annotated with type information by a type inference pass.

The constants $\text{pair}, \text{fst}, \text{snd}$ have the same (obvious) $S$ types in both interpretations. The constants $\text{new}, \text{rd}, \text{wr}$ create, read, and write a mutable variable. Unlike $\text{pair}$, their types differ in the two interpretations. As Figure 4 shows. In the lazy interpretation their types explicitly involve the source-language $\text{ST}$ monad, and $S$ also includes

\[
\begin{align*}
\text{"ML" constants} & \\
\text{new} & \equiv \text{\forall } a. a \rightarrow \text{Ref } a \\
\text{rd} & \equiv \text{\forall } a. \text{Ref } a \rightarrow a \\
\text{wr} & \equiv \text{\forall } a. \text{Ref } a \rightarrow a 
\end{align*}
\]

\[
\begin{align*}
\text{"Haskell" constants} & \\
\text{new} & \equiv \text{\forall } a. a \rightarrow \text{ST } (\text{Ref } a) \\
\text{rd} & \equiv \text{\forall } a. \text{Ref } a \rightarrow \text{ST } a \\
\text{wr} & \equiv \text{\forall } a. \text{Ref } a \rightarrow a \rightarrow \text{ST } ()
\end{align*}
\]
The translation given here introduces quite a few "administrative
One consequence of our decision to allow a type to be modeled by an unpointed CPO is that we have to take care with recursion. The rule (REC) in Figure 1 suggests that a \( \text{letrec } x : \tau = e_1 : \tau \text{ in } e_2 : \rho \) can be constructed at any type \( \tau \), but that is not so. Consider \( \text{letrec } x : \text{Int} = \ldots x \ldots \text{ in } \ldots \). Such a recursive definition is plainly nonsense, because \( \text{Int} \) is an unpointed type and has no bottom element; so there might be no solution, or many solutions, to the recursive definition. We can only do recursion over \textit{pointed} CPOs!

How, then, can we make sense of recursion? One solution is to link recursion to the \( \text{Lift} \) monad, since \( \text{Lift} \) adds a bottom to its argument domain:

\[
\Gamma, x : \tau \vdash e_1 : \tau \quad \Gamma, x : \tau \vdash e_2 : \rho \\
\Gamma \vdash \text{letrec } x : \tau = e_1 \text{ in } e_2 : \rho
\]

This solution is not very satisfactory. For a start, it cannot type:

\[
\text{letrec } f = \lambda x. \ldots \text{ in } \ldots
\]

because the type of a lambda abstraction has the form \( \tau \rightarrow \rho \), not \( \text{Lift} \tau \), and lifting all functions raises the spectre of having to force the definition on each recursive call. Nor can it type recursive definitions of \( \text{ST} \) computations. Furthermore, this loss of expressiveness is completely unnecessary, since a function type whose result type is pointed is itself pointed; and any \( \text{ST} \) computation is pointed. The right solution is to fix (REC) by adding a side condition that \( \tau \) must be pointed:

\[
\Gamma, x : \tau \vdash e_1 : \tau \\
\Gamma, x : \tau \vdash e_2 : \rho \\
\Gamma \vdash \tau \text{ pointed}
\]

\[
\Gamma \vdash \text{letrec } x : \tau = e_1 \text{ in } e_2 : \rho
\]

Figure 6 gives rules for determining when a type is pointed. Unfortunately, the extension to a polymorphic type system is problematic: \textit{is the type a pointed or not?} There are three possible choices:

- We could decide that type variables can only range over pointed types. This is precisely the restriction proposed by Peyton Jones & Launchbury [1991], but it is unacceptable in our IL because we expect (the translations of) most ML data types to be unpointed. For example, an ordinary, non-recursive polymorphic function such as the identity function could not be applied to both \( 3 \) and \( \text{ret}_{\text{LST}} 3 \), because one has a lifted type and one does not.

\footnote{There is a substantial literature on the categorical treatment of recursion (for example, Pitts [1996]), but the discussion of this section focuses on the specific setting of CPOs.}
We could allow type variables to range over all types, but prohibit recursion at a type variable. This would irritatingly reject recursive functions whose result type is a type variable, such as the function \( \text{nth} \) that selects the \( n \)th element from a list:

\[
\text{nth} : \forall \alpha. \text{Int} \to (\text{List} \alpha) \to \alpha
\]

Alternatively, we could employ qualified universal quantification, where type variables at which fixpoints are taken are explicitly qualified:

\[
\text{nth} : \forall \alpha \in \text{Pointed} \text{Int} \to (\text{List} \alpha) \to \alpha
\]

Launchbury & Paterson [1996] elaborate on this idea.

Since the first two choices are untenable, we conclude that adding polymorphism to a language with both recursion and unpointed types requires the use of qualified universal quantification.

3.6 Controlling evaluation in \( \mathcal{L}_1 \)

While \( \mathcal{L}_1 \) seems to be quite suitable from a theoretical point of view, it suffers from a serious practical drawback: \( \mathcal{L}_1 \) is vague about the timing and degree of evaluation. Consider the \( \mathcal{L}_1 \) expression:

\[
\text{let } x : \tau = e \text{ in } f \ x
\]

What code should the code generator produce for such an expression?

- An ML compiler writer would probably expect the code to evaluate the right-hand side of the \( \text{let} \) and then call \( f \) passing the value thus computed. But this eager strategy is incorrect in general if \( e \) diverges, and \( f \) does not evaluate its argument, as a quick glance at Figure 2 will confirm.

- A safe strategy is to build a thunk (suspension) for the right-hand side, bind \( x \) to this thunk, and call \( f \) passing the thunk to it. That is precisely what the code generator for a lazy language would do.

Now suppose that we are compiling code for \( f \), and that \( f \) has type \( \text{Int} \to \text{Int} \). The major motivation for distinguishing \( \text{Int} \) from \( \text{Lift} \text{Int} \) was to allow the compiler to treat values of type \( \text{Int} \) as certainly-evaluated, just as a strict-language compiler would assume (Section 3.1). It is unacceptable for \( f \) to test whether its argument is evaluated; such a choice would guarantee that no ML compiler would use this intermediate language! Alas, the safe strategy for preparing the \( f \)'s argument does indeed pass an unevaluated thunk, so \( f \) must be prepared for this eventuality.

Can we instead use a hybrid strategy?

- A hybrid strategy for compiling \( \text{let} \) expressions might use the type of the bound variable to decide what to do: for types whose values are sure to converge (such as \( \text{Int} \)) it can evaluate the right-hand side eagerly, otherwise it can build a thunk. This strategy works for a simply-typed language but fails (again!) when we introduce polymorphism. What is the code generator to do with a \( \text{let} \) that binds a value of type \( \alpha \)? Either the instantiating type must be passed as an argument, or we must have two versions of the code, one for terminating types and one for possibly-diverging ones.

We regard these complications as a very serious (and far from obvious) objection to using \( \mathcal{L}_1 \) for operational purposes.

3.7 Summary

We expected it to be a routine matter to translate both Haskell and ML into a common language built directly on top of the standard mathematics for programming-language semantics. To our surprise it was not, as Sections 3.5-3.6 describe.

\( \mathcal{L}_1 \) may still be quite useful as a kernel language for reasoning about programs. However, as Section 3.6 has shown, it is unsuitable as a compiler intermediate language. Thus motivated, we now turn our attention to a second design that is more suitable as an IL.

4 \( \mathcal{L}_2 \), a language of partial functions

Our second design starts from the problem we described in Section 3.6. Operationally, it is essential to be able to control exactly when evaluation takes place, so that the recipient of a value knows for sure whether or not it is evaluated.

Since we want to control what evaluation is done when, the obvious thing to do is to make \( \text{let} \) (and, of course, function application) eager. That is, to evaluate \( \text{let } x : \tau = e \text{ in } b \) one evaluates \( e \) binds it to \( x \), and then evaluates \( b \). (We use the operational term “eager”, rather than the semantic term “strict” because the latter does not mean anything if the type of \( e \) has no bottom element.) How, then, are we to translate the \( \text{let}s \) and function applications of a lazy language? There is a standard way to do so, namely by making the construction and forcing of thunks explicit (Friedman & Wise [1976]). This is what we do in \( \mathcal{L}_2 \).

Figure 7 gives the syntax and extra type rules for \( \mathcal{L}_2 \). There is now only one monad, \( \text{ST} \); the \( \text{Lift} \text{Int} \) monad is now implicit in the semantics of \( \mathcal{L}_2 \) so that \( \text{let} \) and function application can be eager. There is a new syntactic form \( \langle \perp \rangle \), that suspends the evaluation of \( e \), and a new constant \( \text{force} \), that forces the suspension returned by its argument. There is one new type \( \langle \perp \rangle \), which is the type of \( \perp \) if \( e \) has type \( \rho \). The two new type rules, \( \langle \text{DELAY} \rangle \) and \( \langle \text{FORCE} \rangle \) are just as you would expect.

Another new feature is that types are divided into value types \( \tau \) and computation types \( \rho \). Intuitively, an expression has a computation type, while a variable is always bound to a value type. Another way to say this is that the typing judgement now has the form

\[
\{x_1 : \tau_1; \ldots; x_n : \tau_n\} \vdash e : \rho
\]

The type rules of Figure 1 apply unchanged, because we carefully used \( \tau \) and \( \rho \) in the right places, although they were synonymous in \( \mathcal{L}_1 \). Function arguments and the right-hand sides of \( \text{let}(\text{sec}) \) expressions all have value types, and are evaluated eagerly. This separation of value types from computation types neatly finesse the awkward question of what it means to “evaluate” an argument computation without also “performing” it, which caused us some heart-searching in earlier un-stratified versions of \( \mathcal{L}_2 \). For example, the expression \( (\text{f } \langle \text{read} \ x \rangle) \) is ill-typed, and hence we do not have to evaluate \( \langle \text{read} \ x \rangle \) without also performing its state changes. Indeed, expressions of type \( \text{ST} \tau \) can only occur as
is that \( \mathcal{L}_2 \)'s function type arrow is now interpreted as the CPO of partial functions, denoted \( \dashv \), and the semantic evaluation function \( \mathcal{E} \) takes an expression to a partial function from environments to values. Many of the equations are defined conditionally. For example, the equation for \( \mathcal{E}[e_1][e_2] \) says that if both \( \mathcal{E}[e_1] \) and \( \mathcal{E}[e_2] \) are defined then the result is just the application of those two values; otherwise there is no equation that applies for \( \mathcal{E}[e_1][e_2] \), so it too is undefined.

The \( \mathcal{E} \) type constructor is modeled using lifting: the semantics of force and \( \mathcal{E} \) move to and fro between lifted CPOs and partial functions. It may seem odd that we use two different notations — Lift \( \tau \) in \( \mathcal{L}_1 \) and \( \mathcal{E} \) in \( \mathcal{L}_2 \) — with the same underlying semantic model, namely lifting. The reason is that in \( \mathcal{L}_1 \) we use lifting as a monad (with a bind operation, for example), whereas in \( \mathcal{L}_2 \) we use it to model thunks (with a force operation but no bind).

The entire semantics of \( \mathcal{L}_2 \) could instead be presented in the CPO of total functions, using the isomorphism:

\[
S \to T \cong S \to T_a
\]

Which to choose is just a matter of taste. What we like about our presentation is that each \( \mathcal{L}_2 \) type constructor corresponds directly to a single categorical type constructor, whereas in the alternative presentation the \( \mathcal{L}_2 \) function type gets a more "encoded" translation. Launchbury & Baraki [1996] use partial functions in essentially the same way.

The translation of "ML" into \( \mathcal{L}_2 \) is exactly the same as the translation of \( \mathcal{L}_1 \). The translation of "Haskell" is different, however, because we now have to be explicit about the introduction of thunks (Figure 9). Concerning types, notice the use of the type constructor \( \mathcal{E} \) on the arguments of functions and data constructors. Concerning terms, the thunk-former \( \mathcal{E} \) is used for function arguments and the right-hand side of all let and letrec definitions. Thunks are evaluated explicitly, using force, when returning a variable or the result of \( \text{fst} \) or \( \text{snd} \).

### 4.1 Controlling evaluation in \( \mathcal{L}_2 \)

The main benefit of using \( \mathcal{L}_2 \) is that its semantics permit an eager interpretation of vanilla \( \text{let} \); namely, "evaluate the right-hand side, bind the value to the variable, and then evaluate the body". A consequence is that any variable of type other than \( \mathcal{E} \), or a type variable (which might be instantiated as \( \mathcal{E} \)), is sure to be fully evaluated, just as in any ML implementation.

### 4.2 Recursion in \( \mathcal{L}_2 \)

Another advantage of \( \mathcal{L}_2 \) is that we can solve our earlier difficulties with recursion (Section 3.5) without requiring bounded quantification.

Firstly, we more or less have to restrict letrecs to bind only syntactic values, because we cannot eagerly evaluate the right-hand side. (Why not? Because we cannot construct the environment in which to evaluate it.) That in turn means that the meaning of the right-hand side is always defined, which is why there is no side condition in the semantics of letrec.

But Figure 7 further restricts the right-hand side of a letrec to be a particular sort of syntactic value, a pointed value, or
PV alue: The syntactic category of PV alues is chosen so that it can only denote a value from a pointed domain, and hence a letrec definition always has at least a fixpoint. To see this, consider the forms that a PV alue can take:

- A lambda abstraction denotes a partial function, and the CPO of partial functions is always pointed; its least element is the everywhere undefined function.
- A thunk \( \langle e \rangle \), where \( e : T \), is drawn from the pointed CPO \( \mathcal{T}[T] \).

Fortunately, the syntactic restriction of letrec does not lose any useful expressiveness. ML insists that letrec bind only functions (which are PV alues), while Haskell binds thunks (which are also PV alues). So there is no difficulty with translating the recursion arising in both ML and Haskell into \( \mathcal{L}_2 \).

4.3 Why not have just one monad?

Now that we have eliminated the Lift monad, and made vanilla let eager, there is another question we should ask: why not give vanilla let the semantics of letST, and eliminate the latter altogether? To put it another way, we have made eager evaluation implicit in the semantics of let; why not add implicit side effects as well? After all, the code generated for \( \text{letST} \ x = e \) in \( b \) will be something like “the code for \( e \) followed by the code for \( b \)”, and that is just the same as the code we now expect to generate for \( \text{let} \ x = e \) in \( b \).

However, if we have just one form of let we lose valuable optimising transformations. In particular, the sequence of computations in ST must be maintained, whereas let bindings can be reordered freely. Changing the order of evaluation is fundamental to several useful transformations, including common sub-expression, loop invariant computations, and all kinds of code motion (Peyton Jones, Partain & Santos [1996]), inlining, and strictness analysis (remember we may be compiling a lazy language into \( \mathcal{L}_1 \)). To take a simple example, the following transformation is not in general valid for letST, but is valid for vanilla let (assuming there are no name clashes):

\[
\text{let } x_1 = e_1 \text{ in let } x_2 = e_2 \text{ in } b
\]

\[
= \text{let } x_2 = e_2 \text{ in let } x_1 = e_1 \text{ in } b
\]

Of course, one could do an effects analysis to determine which sub-expressions were pure, as good ML compilers do, . . . but that is effectively just what the monadic type system records!

5 Assessment

5.1 \( \mathcal{L}_1 \) vs \( \mathcal{L}_2 \)

What have we lost in the transition from \( \mathcal{L}_1 \) to \( \mathcal{L}_2 \), apart from a somewhat more complicated semantics? One loss is \( \mathcal{L}_1 \)’s ability to describe types whose values are sure to terminate. If a \( \mathcal{L}_1 \) function has type \( \text{Int} \rightarrow \text{Int} \), then a call to the function cannot diverge; but the same is not true of \( \mathcal{L}_2 \). This does not have much impact on a compiler, but it makes programmer reasoning about \( \mathcal{L}_2 \) programs more complicated.
Another important difference is that $L_2$ has a weaker $\beta$ rule. $L_1$ has full $\beta$-conversion. That is, for any expressions $\alpha$ and $\beta$:

$$\text{let } x = \epsilon \text{ in } b = \beta[\epsilon/x]$$

(A similar rule holds for application, of course.) In $L_2$, however, $\beta$ does not hold in general. A particular case of this is that if $x$ is not mentioned in $b$ then in $L_1$ the binding can be discarded; in $L_2$ the binding can only be discarded if the right-hand side is a value.

However $\beta_V$ — a restricted version form of $\beta$ that allows only values to be substituted — is valid in $L_2$. Values are defined in Figure 7, and include variables, constants, and lambda abstractions, as usual. However, values also include thunks. Hence any Haskell $\beta$ reduction has a corresponding $\beta_V$ reduction in its $L_2$ translation. Thus, the restriction to $\beta_V$ will not prevent a Haskell compiler from doing anything it can do in an implicitly lazy language with a full $\beta$ rule.

Thus far we have assumed a call-by-name semantics, in which we are content to duplicate arbitrary amounts of work provided we do not change the overall result. In practice no compiler would be so liberal; we desire a call-by-need semantics in which work is not duplicated. As Arjola et al. [1995] describes, we can give a call-by-need semantics to $L_1$ by weakening $\beta$ to $\beta_N$ and adding a garbage-collection rule that allows an unused let binding to be discarded. An analogous result holds in $L_2$: we can obtain call-by-need semantics by replacing $<\epsilon>$ by $<\epsilon>$ in the definitions of values in Figure 7.

5.2 $L_2$ vs Haskell and ML ILs

Our main theme is the search for an IL that can serve for both ML- and Haskell-like languages. However, we believe that a language like $L_2$ is attractive in its own right to either community in isolation, because one might get better code from an $L_2$-based compiler.

For the Haskell compiler writer $L_2$ offers the ability to express in its type that a value is certainly evaluated. This gives a nice way to express the results of strictness analysis: a function argument of unpointed type must be passed by value. Flat arrays and strict data structures also become expressible.

For the ML compiler writer $L_2$ offers the ability to express the fact that a computation is free from side effects, which is a precondition for a raft of useful transformations (Section 4.3). While this information can be gleaned from an effects analysis, maintaining this information for every sub-expression, across substantial program transformations is not easy. In $L_2$, however, local transformations can perform, and record the results of a simple incremental effects analysis. For example, consider the following ML function:

$$\text{fun } f \; x = \text{fst} \; (\text{fst} \; x)$$

If we translate this into $L_2$ we obtain:

$$f = \text{return}_{ST} \; (\lambda x. \text{let } a_2 = x \text{ in } \text{let } a_1 = \text{fst} \; a_1 \text{ in } \text{return}_{ST} \; (\text{fst} \; a_2))$$

Simple application of the rules of Figure 3 allows this expression to simplify to:

$$f = \text{return}_{ST} \; (\lambda x. \text{let } a_1 = x \text{ in } \text{let } a_2 = \text{fst} \; a_1 \text{ in } \text{return}_{ST} \; (\text{fst} \; a_2))$$

Now the return$_{ST}$ can be floated outwards, to give:

$$f = \text{return}_{ST} \; (\lambda x. \text{return}_{ST} \; (\text{fst} \; (\text{fst} \; x)))$$

In this form, the inner return$_{ST}$ makes it apparent that $f$ has no side effects. We have, in effect, performed a sort of incremental effects analysis. The same idea can be taken further. If $f$ is inlined at its call sites, then the return$_{ST}$ may cancel with let$_{ST}$ there, and so on. Even if $f$’s body is big, we can use the “worker-wrapper” technique of Peyton Jones & Launchbury [1991] to split $f$ into a small, inlinable wrapper and a large, non-inlinable worker, $fw$, thus:

$$f = \text{return}_{ST} \; (\lambda x. \text{return}_{ST} \; (\text{fst} \; (\text{fst} \; x)))$$

Blume & Appel [1997] describe a similar technique that they call “lambda-splitting”.

The point of all this is that there is a real payoff for an ML compiler from making the ST monad explicit. Easy, incremental transformations perform a local effects analysis; at each stage the state of the analysis is recorded in the program itself, rather than in some ad hoc auxiliary data structures; and all other program transformations will automatically preserve (or exploit) the analysis.

5.3 Parametricity

Polymorphic functions have certain parametricity properties that may be derived purely from their types (Mitchell & Meyer [1985]; Reynolds [1983]; Wadler [1989]). For example, in the pure polymorphic lambda calculus, a function $f$ with type $\forall \alpha. \alpha \rightarrow \alpha$ satisfies the theorem:

$$\forall A, B : \forall h : A \rightarrow B. \forall x : A. h \; (f \; x \; y) = f \; (h \; x) \; (h \; y)$$

In fact $f$ satisfies something even stronger in which the function $h$ can be an arbitrary relation between $A$ and $B$.

When we add “polymorphic” constants to the pure calculus, the effect is that the choice of functions $h$ becomes restricted. For example, adding a fix point operator $fix : \forall \alpha. (\alpha \rightarrow \alpha) \rightarrow \alpha$ forces the restriction that the $h$ functions be strict (map $\bot$ to $\bot$) and inductive (i.e. continuous). This is the situation in Haskell for example.

Adding polymorphic sequencing, say through an operator $seq : \forall \alpha, \beta, \alpha \rightarrow \beta \rightarrow \beta$ or by building it into the semantics of function application, forces the restriction that the $h$ functions be bottom-reflecting (i.e. defined on all defined arguments). This is the basic situation in pure ML.

Adding polymorphic equality forces the $h$ functions to be at least one-to-one; and adding polymorphic state operations like $\tau$ seems to remove any last shred of interesting parametricity.

What, then, are the parametricity properties of $L_1$ and $L_2$? If parametricity properties are weakened by claiming various primitives to be more polymorphic than they really are, then by being more cautious in the types we assign them, we may hope to reestablish parametricity.
In $\mathcal{L}_2$, for example, recursion is only done either at a function type, or at a suspension type. Recursion is never permitted as a fully polymorphic type (unlike in Haskell). This has the effect of allowing the strictness side condition to be dropped, though inductiveness (or continuity) is still required. The same is achieved in $\mathcal{L}_1$ through the use of the pointed restriction (see Launchbury & Paterson [1996] for a comparable situation). Furthermore, since all state operations are explicitly typed within the state monad, they also do not interfere with parametricity in a negative way.

The main difference between $\mathcal{L}_1$ and $\mathcal{L}_2$ is to do with forcing evaluation. $\mathcal{L}_1$ has no polymorphic forcing operation, so has no consequent weakening of its parametricity property. $\mathcal{L}_2$ does, however — it is built into its eager function application. Thus for $\mathcal{L}_2$ the parametricity theorem demands the $h$ functions to be everywhere defined.

To see an example of this, consider the function $K : \forall \alpha. \beta \alpha \rightarrow \beta \rightarrow \alpha$ which selects its first argument, discarding its second. The parametricity theorem is

$$\forall A, A', B, B'. \forall h_1 : A \rightarrow A', h_2 : B \rightarrow B'. \forall x : A, y : B. \ h_1 (K x y) = K (h_1 x) (h_2 y)$$

Clearly this holds only if $h_2$ is total (defined everywhere), otherwise the right hand side may not be defined when the left hand side is.

There is a practical implication to this. A class of techniques for removing intermediate lists called fold-build relies on parametricity for its correctness (Gill, Launchbury & Peyton Jones [1993]). While a strictness side condition is not damaging, a totality condition is too restrictive. The technique can no longer rely on the types to provide sufficient guidance for correctness. This is disappointing, although unsurprising. The compiler can still recover the short-cut deforestation technique by refining $\mathcal{L}_2$'s type system to use qualified types along the lines of Launchbury & Paterson [1996].

### 5.4 Side effects and polymorphism

It is well known that the ability to create polymorphic references can lead to unsoundness in the type system (Tofte [1990]). For example, if we are able to create a reference $r$ with type $\forall \alpha. \text{Ref} \alpha$ then we would be able to write the following erroneous code:

```plaintext
let $r$ = (Ref $\alpha$) in
let $f$ = ($\alpha$ -> Int) -> Int in
let $x$ = $f$ $r$ in
```

However in both $\mathcal{L}_1$ and $\mathcal{L}_2$ any expression of type $\forall \alpha. \text{Ref} \alpha$ is undefined in any environment! The way to extract a value of Ref type is with new, which returns a value of type (Ref $\tau$). The only way to strip off the Ref constructor is with letsg. Looking at the typing rule for letsg, we can see that bound variable must have type Ref $\tau$.

SML's so-called "value restriction" conservatively restricts generalisation in let bindings precisely to avoid the construction of such polymorphic references. We conjecture (though we have not proved) that $\mathcal{L}_1$ and $\mathcal{L}_2$ are both sound without any such side conditions.

### 5.5 ML thunks

One of the advantages of a language that supports both strict and lazy evaluation is that it can accommodate source languages that have such a mixture. Indeed, it is quite straightforward to map Haskell's strictness annotations (Peyton Jones et al. [1997]) onto $\mathcal{L}_2$. Coming from the other direction, it has long been known that thunks can be encoded explicitly in a strict, imperative language. For the sake of concreteness we use the notation proposed for ML in Okasaki [1996]. In this proposal delayed ML expressions are prefixed by a "$\llbracket\$", thus:

```plaintext
let val x = $\llbracket$f y$\rrbracket$ in b end
```

Here, assuming ($f y$) has type Int, $x$ is bound to a thunk of type Int susp that, when forced, evaluates ($f y$) and overwrites the thunk with its value.

We expected that these "ML thunks" would map directly onto $\mathcal{L}_2$'s thunks, but that turned out not to be the case. The semantics of ML thunks is considerably more complicated than that of $\mathcal{L}_2$'s thunks, because of the interaction with state. Consider the following ML expression:

```plaintext
let val rec x = $\llbracket$let val y = !r - 1 in
r; "y"
if y=0 then 0
else force x + force x
end
in ... end
```

(This defines $x$ recursively, which is not possible in ML, but essentially the same thing can be done using another reference to "tie the knot". We use the recursive form to reduce clutter.) When $x$ is evaluated it decrements the contents of the reference cell $r$; but then, if the new value is non-zero, $x$ evaluates itself! In effect, there can be multiple simultaneous activations of $x$, rather like the multiple activations of a recursive function. (Indeed, a non-memoising implementation of ML thunks can be obtained by representing $\llbracket e$ by $\lambda (x.e)$.) Furthermore, these multiple activations can each have a different value, because they each read the state.

$\mathcal{L}_2$'s thunks have a much simpler semantics. A thunk has only one value, and there can be at most one activation of the thunk7. The key insight is that evaluation of a $\mathcal{L}_2$ thunk has no side effects, unlike the ML thunk above. But what if the contents of the thunk performs side effects? For example:

```plaintext
let x = letsg v : Int <- rd r in wr (v+1) in e
```

Here, if $r$ : Ref Int, then $x$ has type (Ref Int) -> not $. Forcing the thunk (with force) causes no side effects (apart from updating the thunk itself), and yields a computation that, when subsequently performed (by a letsg), will increment the location $r$. The computation $x$ may be performed many times; for example, $e$ might be

```plaintext
letsg al : () <- force x in letsg a2 : () <- force x in ...
```

What this means though, is that the more complicated semantics of ML thunks have to be expressed explicitly in $\mathcal{L}_2$, presumably by coding them up using explicit references.

7 More precisely, if there is more than one then the thunk's value depends on its own value, so its value is undefined. This property justifies the well-known technique of "black-holing" a thunk, both to avoid space leaks and to report certain non-termination (Jones [1992]).
6 Related work

The FLINT language has rather similar objectives to the work described here, in that it aims to serve as a common infrastructure for a variety of higher-order typed source languages (Shao [1997b]). However, FLINT has not (so far) concentrated much on the issue of strictness and laziness, which is the main focus of this paper. The ideas described here could readily be incorporated in FLINT.

Both the Glasgow Haskell Compiler and the TIL ML compiler use a polymorphic strongly-typed internal language, though the latter is considerably more sophisticated and complex (Peyton Jones [1996]; Tarditi et al. [1996]). Neither, however seriously attempt to compile the other’s main evaluation-order paradigm.

7 Further work

In this paper we have concentrated on a core calculus. Some work remains to extend it to a practical IL:

- Recursive data types and case expressions must be added — we anticipate no difficulty here.
- A proof of type soundness is needed. As we note in Section 5.4 its soundness is not obvious.
- We have a simple operational semantics for \( \mathcal{L}_2 \); we are confident that it is sound and adequate, but have yet to do the proofs.
- We are studying whether is it possible to combine \( \mathcal{L}_1 \)’s ability to describe certainly-terminating computations with \( \mathcal{L}_2 \)’s operational model.

Accommodating the ML module system is likely to involve a significant extension of the type system (Harp er & Stone [1997]); we have not yet studied such extensions.

In a separate paper we discuss how to use the framework of Pure Type Systems to allow the language of terms, types, and kinds to be merged into a single language and compiler data type (Peyton Jones & Meijer [1997]). We hope to merge the results of that paper and this one into a single IL.

We have made no attempt to address the tricky problem of how to combine monads. For example, ML includes the monad of state and exceptions. Is it advantageous to separate them into the composition of two monads, or is it better to have a single, combined monad? In the former case, what transformations hold?

An important operational question is that of the representation of values, especially numbers. Quite a few papers have discussed how to use unboxed representations for data values, and it would be interesting to translate their work into the framework of \( \mathcal{L}_2 \) (Leroy [1992]; Peyton Jones & Launchbury [1991]; Shao [1997a]).

Acknowledgements

We would like to thank the POPL referees, Nick Benton, Bob Harper, Andrew Kennedy, Jeff Lewis, Erik Meijer, Chris Okasaki, Ross Paterson, Amr Sabry, and Tim Sheard for helpful feedback on earlier discussion and drafts of this paper. We gratefully acknowledge the support of the Oregon Graduate Institute, funded by a contract with US Air Force Material Command (F19628-93-C-0069).

References


