Once upon a type*

David N. Turner Philip Wadler
University of Glasgow†
dnt,wadler@dcs.glasgow.ac.uk

Christian Mossin
University of Copenhagen‡
mossin@diku.dk.

Abstract

A number of useful optimisations are enabled if we can determine when a value is accessed at most once. We extend the Hindley-Milner type system with *uses*, yielding a type-inference based program analysis which determines when values are accessed at most once. Our analysis can handle higher-order functions and data structures, and admits principal types for terms.

Unlike previous analyses, we prove our analysis sound with respect to call-by-need reduction. Call-by-name reduction does *not* provide an accurate model of how often a value is used during lazy evaluation, since it duplicates work which would actually be shared in a real implementation.

Our type system can easily be modified to analyse usage in a call-by-value language.

1 Introduction

This paper describes a method for determining when a value is used at most once. Our method is based on a simple modification of the Hindley-Milner type system. Each type is labelled to indicate whether the corresponding value is used at most once, or may possibly be used many times.

Our type system has a number of applications:

Program transformation: If it is determined that a variable is accessed at most once, then one may safely inline the expression bound to the variable without reducing efficiency. In particular, we can determine when it is safe to inline an expression into the body of a function.

Avoiding closure update: Implementations of lazy languages use updates to share the evaluation of closures. If it is determined that a closure is accessed at most once, then there is no need to overwrite the closure with the result of evaluation.

Enabling data structure update: If it is determined that a structure such as a 'cons' cell or an array is accessed at most once, then the structure may safely be updated in place.

The last of these application areas has received considerable attention, the second of these some attention, and the first almost none. This is quite surprising, since expression inlining is central to a wide range of program transformations. In particular, our method provides a sound basis for a number of transformations that were previously dealt with in an ad hoc manner in the Glasgow Haskell Compiler, and a solution to a problem that has bedevilled those attempting to extend Deforestation to higher order.

Previous analyses to determine when a value is used at most once have been based on either call-by-name (Wright and Baker-Finch [WB93], Courtenage and Clack [CC94]) or call-by-need (Launchbury *et al.* [Lau92], Marlow [Mar93]). The call-by-name analyses have been proved sound, but are not well-suited for optimisation of lazy languages. Our analysis is the first call-by-need analysis to be proved sound, and sometimes provides more accurate information than other call-by-need analyses. Our proof of soundness is based on the operational semantics of Launchbury [Lau93] and the call-by-need calculus of Ariola *et al.* [AFMOW95].

The type system presented here is based on ideas taken from the linear logic of Girard [Gir87] and its successor the Logic of Unity [Gir93]. However, it turns out to be convenient to present this work without reference to linear logic. Some of the connections are traced in a companion paper [MOTW95], which relates linear logic to the call-by-need calculus of Ariola *et al.*

We modify the Hindley-Milner type system [Hin69, DM82] by attaching *uses* to types. Type judgements include a constraint set relating uses, similar to the constraint sets relating subtypes in the work of Mitchell [Mit84, Mit91]. As with the Hindley-Milner system, there is an algorithm that determines a principal type for an expression. Representing usage information as type annotations provides a convenient mechanism for communicating usage information across module boundaries, since typed languages such as Haskell already import type information from separately-compiled modules.

A small modification to our analysis enables it to determine when variables are used *exactly* once (as opposed to at most once), making it suitable for use with call-by-value (as opposed to call-by-need) evaluation.

*Appears in the *7'th International Conference on Functional Programming and Computer Architecture*, San Diego, California, June 1995. This work was supported by EPSRC (GR/J 527099), 'Save space with linear types', ESPRIT (6809) 'Semantique', ESPRIT (ECUS025) 'Atlantique', and the Danish Technical Research Council.

†Computing Science Department, University of Glasgow, Glasgow G12 8QQ, Scotland.

‡DIKU, University of Copenhagen, Universitetsparken 1, 2100 Copenhagen Ø, Denmark. This work was performed while Mossin was visiting the University of Glasgow.

1.1 The problem

Some examples may help to illustrate the nature of the problem solved.

We wish to attach *uses* to values. The use 1 indicates that a value is used at most once, while the use ω indicates that a value may be used any number of times.

Consider the following. (Example 1.)

$$\begin{array}{l}\text{let } x = 1+2 \text{ in} \\ \quad \text{let } y = x+3 \text{ in} \\ \quad\quad y+y\end{array}$$

Here it is safe to replace x by $1+2$ within the body of the outer 'let'. But it is not safe to replace y by $x+3$ within the body of the inner 'let', as the resulting program would compute $x+3$ twice rather than once. Our type system attaches use 1 to x and use ω to y.

Our argument depends crucially on the use of call-by-need. Under call-by-name $x+3$ is computed twice regardless. Hence a call-by-name analysis must attach a use equivalent to ω to both x and y, showing why such analyses are not suited for our purpose.

At first glance, it may seem child's play to determine if a value is used at most once under call-by-need. Surely, if a variable appears at most once in a program, then the value it is bound to is used at most once? In fact, this is not the case.

Consider the following. (Example 2.)

$$\begin{array}{l}\text{let } x = 1+2 \text{ in} \\ \quad \text{let } f = \lambda z.\, x+z \text{ in} \\ \quad\quad f\,3 + f\,4\end{array}$$

Even though x appears only once in the body of the outer 'let', replacing x by $1+2$ is unsafe, as the resulting program will compute $1+2$ twice rather than once. Our type system attaches use ω to both x and f.

Consider the following. (Example 3.)

$$\begin{array}{l}\text{let } x = 1+2 \text{ in} \\ \quad \text{let } f = (\text{let } y = x+3 \text{ in } \lambda z.\, y+z) \text{ in} \\ \quad\quad f\,4 + f\,5\end{array}$$

Here, again, one can safely replace the one occurrence of x by $1+2$, although it may require a moment's thought to convince oneself this is the case. Indeed, this example is sufficiently difficult that the analyses proposed by Launchbury et al. [Lau92] and by Marlow [Mar93] are both overly conservative, and in effect attach use ω to x. However, our type system attaches use 1 to x and use ω to y and f.

1.2 Call-by-need

Our work is based on the operational semantics of call-by-need proposed by Launchbury [Lau93], and on the call-by-need lambda calculus of Ariola *et al.* [AFMOW95]. A correspondence between these two approaches has already been shown in the latter work.

Launchbury's rules include an explicit treatment of closure update. By modifying his rules to allow some closures to be non-updating, we verify that our type system can be used to avoid unnecessary closure updates.

The soundness of our type system is established by showing that it satisfies a subject reduction property: applying a call-by-need reduction to a term leaves its type unchanged, including type information regarding usage.

Launchbury restricts functions to be applied to variables, while Ariola *et al.* allow functions to be applied to arbitrary expressions. As we explain in Section 5, the difference between these approaches is significant for our chosen implementation, the Glasgow Haskell Compiler [PHHPW93], which is closely based on the STG-machine of Peyton Jones [Pey92]. Therefore, in this paper we adopt Launchbury's syntax (which was influenced by the Haskell compiler and the STG machine), and adapt the results of Ariola *et al.* to it.

1.3 Program transformation

If we want to use program transformation as the basis of efficient compilation of a functional language, it is not only important that transformation preserves meaning but that the transformed program executes at least as fast as the original.

Consider Church's beta rule:

$$(\lambda x.\, e_0)\, e_1 \Longrightarrow [e_1/x]e_0.$$

This rule is good in that it eliminates one application step, but bad in that it may duplicate some computation. (In particular, computation of e_1 may be duplicated if x is used more than once in e_0.)

The call-by-need calculus of Ariola *et al.* addresses this problem by modifying the above rule:

$$(\lambda x.\, e_0)\, e_1 \Longrightarrow \text{let } x = e_1 \text{ in } e_0.$$

This rule, together with a number of rules for manipulating 'let', allow us to safely transform programs, without the risk of duplicating work.

However, there are a number of transformation that are useful and safe and which are not part of the call-by-need calculus. The most important of these is that the beta rule

$$(\lambda x.\, e_0)\, e_1 \Longrightarrow [e_1/x]e_0$$

is safe when x has use 1. Another is that the rule

$$\text{let } x = e_0 \text{ in } \lambda y.\, e_1 \Longrightarrow \lambda y. \text{let } x = e_0 \text{ in } e_1$$

is safe when the function $\lambda y.\, e_1$ has use 1. Both of these transformations are used extensively in the Glasgow Haskell Compiler. Until now, their safety was ensured only by ad hoc techniques. And the ad hoc techniques were not adequate – an unsafe version of the second rule was allowed, with the result that the compiler itself (which is bootstrapped) was slowed by as much as one third [SP95].

Another example of program transformation is the deforestation algorithm [Wad90a]. In order to ensure safety, this algorithm requires that variables are used at most once. The definition of 'used at most once' is easy because deforestation applies to a first-order language. Attempts to apply deforestation to higher-order [MW92] have been hindered by the lack of a suitable definition of 'used at most once' at higher-order. This paper provides such a definition.

1.4 Outline

This paper is organised as follows. Section 2 introduces the language used and its semantics. Section 3 describes the fundamentals of the type system. Section 4 discusses principal types and polymorphism. Section 5 summarises the call-by-need reduction rules. Section 6 discusses how to adapt the analysis so that it is appropriate for a call-by-value language. Section 7 describes related work. Section 8 concludes.

$$\text{Var-Once}\ \frac{|x| = 1 \qquad \langle H_0 \rangle\, e \Downarrow \langle H_1 \rangle\, v}{\langle H_0,\ \text{let } x = e,\ H_2 \rangle\, x \Downarrow \langle H_1,\ H_2 \rangle\, v}$$

$$\text{Var-Many}\ \frac{|x| = \omega \qquad \langle H_0 \rangle\, e \Downarrow \langle H_1 \rangle\, v}{\langle H_0,\ \text{let } x = e,\ H_2 \rangle\, x \Downarrow \langle H_1,\ \text{let } x = v,\ H_2 \rangle\, v}$$

$$\text{Var-Rec}\ \frac{}{\langle H_0,\ \text{letrec } x = v,\ H_1 \rangle\, x \Downarrow \langle H_0,\ \text{letrec } x = v,\ H_1 \rangle\, v}$$

$$\text{Abs}\ \frac{}{\langle H \rangle\, \lambda x.\, e \Downarrow \langle H \rangle\, \lambda x.\, e} \qquad \text{App}\ \frac{\langle H_0 \rangle\, e_0 \Downarrow \langle H_1 \rangle\, \lambda x.\, e_1 \qquad \langle H_1 \rangle\, [y/x]e_1 \Downarrow \langle H_2 \rangle\, v}{\langle H_0 \rangle\, e_0\ x \Downarrow \langle H_2 \rangle\, v}$$

$$\text{Int}\ \frac{}{\langle H \rangle\, n \Downarrow \langle H \rangle\, n} \qquad \text{Plus}\ \frac{\langle H_0 \rangle\, e_0 \Downarrow \langle H_1 \rangle\, n_0 \qquad \langle H_1 \rangle\, e_1 \Downarrow \langle H_2 \rangle\, n_1}{\langle H_0 \rangle\, e_0 + e_1 \Downarrow \langle H_2 \rangle\, \underline{n_0 + n_1}}$$

$$\text{Nil}\ \frac{}{\langle H \rangle\, \text{nil} \Downarrow \langle H \rangle\, \text{nil}} \qquad \text{Cons}\ \frac{}{\langle H \rangle\, \text{cons}\ x\ y \Downarrow \langle H \rangle\, \text{cons}\ x\ y}$$

$$\text{Case-Nil}\ \frac{\langle H_0 \rangle\, e_0 \Downarrow \langle H_1 \rangle\, \text{nil} \qquad \langle H_1 \rangle\, e_1 \Downarrow \langle H_2 \rangle\, v}{\langle H_0 \rangle\, \text{case } e_0 \text{ of } \{\text{nil} \to e_1;\ \text{cons}\ x\ y \to e_2\} \Downarrow \langle H_2 \rangle\, v}$$

$$\text{Case-Cons}\ \frac{\langle H_0 \rangle\, e_0 \Downarrow \langle H_1 \rangle\, \text{cons}\ y_0\ y_1 \qquad \langle H_1 \rangle\, [y_0/x_0, y_1/x_1]e_2 \Downarrow \langle H_2 \rangle\, v}{\langle H_0 \rangle\, \text{case } e_0 \text{ of } \{\text{nil} \to e_1;\ \text{cons}\ x_0\ x_1 \to e_2\} \Downarrow \langle H_2 \rangle\, v}$$

$$\text{Let}\ \frac{\text{fresh } x' \qquad \langle H_0,\ \text{let } x' = e_0 \rangle\, [x'/x]e_1 \Downarrow \langle H_1 \rangle\, v}{\langle H_0 \rangle\, \text{let } x = e_0 \text{ in } e_1 \Downarrow \langle H_1 \rangle\, v} \qquad \text{Letrec}\ \frac{\text{fresh } x' \qquad \langle H_0,\ \text{letrec } x' = [x'/x]v \rangle\, [x'/x]e \Downarrow \langle H_1 \rangle\, v}{\langle H_0 \rangle\, \text{letrec } x = v \text{ in } e \Downarrow \langle H_1 \rangle\, v}$$

Figure 1: Natural semantics

2 Language

We now present the syntax and operational semantics of a call-by-need lambda calculus extended with integers, lists, and recursion. For the sake of brevity other constructs have been omitted, but there would be no difficulty in including them.

2.1 Terms

The syntax of the language is given below. We use the syntax of Launchbury [Lau93] where arguments in applications and in cons are restricted to variables.

Variables	x, y, z	
Values	$v ::=$	$\lambda x.\, e \mid n \mid \text{nil} \mid \text{cons}\ x\ y$
Terms	$e ::=$	$v \mid x \mid e\ x \mid e_0 + e_1 \mid$
		$\text{case } e_0 \text{ of } \{\text{nil} \to e_1;\ \text{cons}\ x\ y \to e_2\} \mid$
		$\text{let } x = e_0 \text{ in } e_1 \mid \text{letrec } x = v \text{ in } e$

It is trivial to translate terms with the standard syntax for application and cons into the restricted syntax, for example we can translate '$e_0\ e_1$' to 'let $x = e_1$ in $e_0\ x$'. The syntax closely resembles the STG language [Pey92].

Our syntax differs from Launchbury in three respects. First, we distinguish between non-recursive 'let' and recursive 'letrec' bindings; second, we allow only a single binding in 'letrec', rather than several mutually recursive bindings; and, third, we restrict the definiens of 'letrec' to be a value. The third restriction is required to permit the second, since otherwise reducing a single recursive binding, such as

$$\text{letrec } y = (\text{let } x = e_0 \text{ in } e_1) \text{ in } e_2,$$

may introduce a mutually recursive binding, such as

$$\text{letrec } x = e_0 \text{ and } y = e_1 \text{ in } e_2.$$

The restricted 'letrec' is still powerful enough to define recursive functions and cyclic lists. Without these three changes the typing and reduction rules would need to be rather more complex; for instance, see Ariola *et al.* [AFMOW95] for the reduction rules required for mutual recursion. We feel that the slight loss of expressiveness in the language is justified by the considerably simpler presentation that it makes possible.

2.2 Use annotations

The operational semantics of this section and the reduction rules of Section 5 require that let-bound variables are annotated with uses, and the type rules of Sections 3 and 4 allow us to infer such annotations.

Each let-bound variable x is annotated with a use $|x|$, which is either 1 or ω. If $|x| = 1$, then x is used at most once during evaluation, and if $|x| = \omega$ then x may be used any number of times. No annotation is required for letrec-bound variables, as they always have use ω.

2.3 Heaps

A *heap* abstracts the state of the store at a point in the computation. It consists of a sequence of bindings associating variables with terms.

Heaps	H	$::=$	$B_1, \ldots, B_n$
Bindings	B	$::=$	$\text{let } x = e \mid \text{letrec } x = v$

We distinguish between non-recursive bindings (written 'let $x = e$') and recursive bindings (written 'letrec $x = v$'). A *configuration* pairs a heap with a term, and is written $\langle H \rangle\, e$.

The expression e in the heap 'H_0, let $x = e, H_1$' can only refer to variables bound in H_0. Similarly, the value v in the heap H_0, letrec $x = v, H_1$ can only refer to the recursively-defined variable x and the variables bound in H_0.

2.4 Evaluation rules

Figure 1 presents a natural semantics for lazy evaluation, which closely resembles the one given by Launchbury. The key difference is that the evaluation of a let-bound variable depends on its use annotation.

Evaluation rules have the form $\langle H_0 \rangle\, e \Downarrow \langle H_1 \rangle\, v$, meaning that evaluating expression e in initial heap H_0 returns value v and final heap H_1.

Rule Var-Once evaluates a variable that is used at most once. Look up the expression e that is bound to the variable x in the heap then evaluates e. As the variable will no longer be used, it is removed from the heap.

Rule Var-Many evaluates a variables that may be used many times. Look up the expression e that is bound to the variable x in the heap, evaluates e, and update the heap to bind x to the resulting value. Note that the expression e in the heap H_0, let $x = e, H_2$ can refer only to variables. In practice, the update required by Var-Many may have a significant cost, whereas Var-Once avoids this cost.

Rule Var-Rec evaluates recursively bound variables: look up the value v that is bound to the variable x in the heap. It is simpler because 'let' binds a term, while 'letrec' can only bind a value.

Rule Abs evaluates abstractions: it is trivial since a lambda expression is already a value.

Rule App evaluates applications: evaluate the function to yield a lambda abstraction, then evaluate the body of the abstraction with the argument substituted for the bound variable of the abstraction.

Rules Int and Plus are easy. Rules Nil and Cons are trivial since both 'nil' and 'cons $x\, y$' are already values. The two Case rules are again straightforward.

Rules Let and Letrec are similar, each creating a new binding on the heap. In these rules x' is a fresh name not appearing in the expression or the heap.

Observe that all bindings added to the heap are to fresh variables, so it trivially follows that all bindings in the heap are unique. This property is not quite as trivial in Launchbury's formulation, where renaming occurs during variable access rather than during evaluation of a binding construct.

3 A use type system

We now present a type system which indicates when values will be accessed at most once. For simplicity, the type system is monomorphic; extensions which allow polymorphism are discussed in Section 4.

3.1 Uses

Types will be annotated with *uses*. A type is annotated with use 1 if each value of that type is used at most once, and annotated with ω otherwise. Thus, 1 and ω stand for upper bounds on the number of times a value can be used. To enable type inference, we also allow use variables (ranged over by j, k, l, m). Let κ range over uses.

$$\text{Uses} \quad \kappa \quad ::= \quad j \mid 1 \mid \omega$$

3.2 Constraints

We use Θ to record the constraints generated by our typing rules. We define Θ to be a set of constraints of the form $j \le \{k_1, \dots, k_n\}$.

The following rules define an ordering on uses, parameterised on a constraint set Θ:

$$\text{Omega}\ \frac{}{\kappa \le_\Theta \omega} \qquad \text{One}\ \frac{}{1 \le_\Theta \kappa} \qquad \text{Refl}\ \frac{}{\kappa \le_\Theta \kappa}$$

$$\text{Taut}\ \frac{(j \le \{k_1, \dots, k_n\}) \in \Theta}{j \le_\Theta k_i}$$

3.3 Types

Types include type variables (let a, b, c range over these), function types, integers, and list types.

$$\text{Types} \quad \tau \quad ::= \quad a^\kappa \mid \tau \to^\kappa \tau' \mid \text{Int}^\kappa \mid [\tau]^\kappa$$

The type a^κ indicates that the type variable a ranges over types with use κ. The type $\tau \to^\kappa \tau'$ denotes functions from type τ to type τ' that can be used at most κ times, Int^κ denotes integer values that can be used at most κ times, and $[\tau]^\kappa$ denotes lists with elements of type τ, where the list can be accessed at most κ times.

Write $|\tau|$ for the use attached to type τ, defined as below:

$$|\tau \to^\kappa \tau'| = \kappa \qquad |\text{Int}^\kappa| = \kappa \qquad |[\tau]^\kappa| = \kappa \qquad |a^\kappa| = \kappa$$

We impose the following well-formedness condition on list types:

The type $[\tau]^\kappa$ is well-formed only if $\kappa \le_\Theta |\tau|$.

In other words, if a list can be accessed many times, then its elements also might be accessed many times (through the list). A similar restriction appears in the type systems of Guzmán and Hudak [GH90] and Wadler [Wad90b, Wad91].

3.4 Contexts

A context associates a type with each variable that may appear in a term, and is represented by a list of entries of the form $x : \tau$.

$$\text{Contexts} \quad \Gamma, \Delta \quad ::= \quad x_1 : \tau_1, \dots, x_n : \tau_n$$

Each variable in a context must be distinct. If $x : \tau$ is in Γ, we say that x has use κ if $|\tau| = \kappa$. If Γ and Δ are contexts containing no variables in common, write Γ, Δ to denote the concatenation of the two contexts.

We extend our ordering on uses so that it applies to complete contexts, written $\kappa \le_\Theta |\Gamma|$ and defined as below:

$$\kappa \le_\Theta |x_1 : \tau_1, \dots, x_n : \tau_n| \quad \text{iff} \quad \kappa \le_\Theta |\tau_i| \text{ for all } i$$

Consider the constraint $\kappa \le |\Gamma|$. If $\kappa = 1$, then no constraint is placed on any use in Γ. But if $\kappa = \omega$ then for every entry $x_i : \tau_i$ in Γ, our definition implies that $|\tau_i| = \omega$.

$$\text{Var}\ \frac{}{x : \tau \vdash_\Theta x : \tau} \qquad \text{Exch}\ \frac{\Gamma,\, x : \tau_0,\, y : \tau_1,\, \Delta \vdash_\Theta e : \tau}{\Gamma,\, y : \tau_1,\, x : \tau_0,\, \Delta \vdash_\Theta e : \tau}$$

$$\text{Cont}\ \frac{\Gamma,\, x : \tau,\, y : \tau \vdash_\Theta e : \tau' \qquad |\tau| = \omega}{\Gamma,\, z : \tau \vdash_\Theta [z/x, z/y]e : \tau'} \qquad \text{Weak}\ \frac{\Gamma \vdash_\Theta e : \tau'}{\Gamma,\, x : \tau \vdash_\Theta e : \tau'}$$

$$\text{Abs}\ \frac{\Gamma,\, x : \tau \vdash_\Theta e : \tau' \qquad \kappa \leq_\Theta |\Gamma|}{\Gamma \vdash_\Theta \lambda x.e : \tau \to^\kappa \tau'} \qquad \text{App}\ \frac{\Gamma \vdash_\Theta e : \tau \to^\kappa \tau' \qquad \Delta \vdash_\Theta x : \tau}{\Gamma,\, \Delta \vdash_\Theta e\, x : \tau'}$$

$$\text{Int}\ \frac{}{\vdash_\Theta n : \text{Int}^\kappa} \qquad \text{Plus}\ \frac{\Gamma \vdash_\Theta e_0 : \text{Int}^{\kappa_0} \qquad \Delta \vdash_\Theta e_1 : \text{Int}^{\kappa_1}}{\Gamma,\, \Delta \vdash_\Theta e_0 + e_1 : \text{Int}^\kappa}$$

$$\text{Nil}\ \frac{}{\vdash_\Theta \text{nil} : [\tau]^\kappa} \qquad \text{Cons}\ \frac{\Gamma \vdash_\Theta x : \tau \qquad \Delta \vdash_\Theta y : [\tau]^\kappa}{\Gamma,\, \Delta \vdash_\Theta \text{cons}\, x\, y : [\tau]^\kappa}$$

$$\text{Case}\ \frac{\Gamma \vdash_\Theta e_0 : [\tau]^\kappa \qquad \Delta \vdash_\Theta e_1 : \tau' \qquad \Delta,\, x : \tau,\, y : [\tau]^\kappa \vdash_\Theta e_2 : \tau'}{\Gamma,\, \Delta \vdash_\Theta \text{case}\ e_0\ \text{of}\ \{\text{nil} \to e_1;\ \text{cons}\, x\, y \to e_2\} : \tau'}$$

$$\text{Let}\ \frac{\Gamma \vdash_\Theta e : \tau \qquad \Delta,\, x : \tau \vdash_\Theta e' : \tau'}{\Gamma,\, \Delta \vdash_\Theta \text{let}\ x = e\ \text{in}\ e' : \tau'} \qquad \text{Letrec}\ \frac{\Gamma,\, x : \tau \vdash_\Theta v : \tau \qquad \Delta,\, x : \tau \vdash_\Theta e : \tau' \qquad |\tau| = \omega}{\Gamma,\, \Delta \vdash_\Theta \text{letrec}\ x = v\ \text{in}\ e : \tau'}$$

Figure 2: Type rules

3.5 Typing judgements

Typing judgements take the form $\Gamma \vdash_\Theta e : \tau$, indicating that in context Γ, and under the constraints Θ, the term e has type τ. The type rules are shown in Figure 2. As usual, these consist of zero or more hypotheses above and a conclusion below the line.

The type rules are quite similar to the usual rules for lambda calculus, and so we concentrate on explaining the unusual features: the structural rules, and the constraints on uses.

3.6 Structural rules

The manipulation of contexts is carefully designed so that if any variable is used more than once this will be indicated by the presence of the structural rule contraction (Cont), which introduces the use ω.

Terms that may be evaluated together are typed in different contexts which are then combined, as can be seen in rules App, Plus, Cons, Case, and Let. As all variables in a context must be distinct, the only way for the same variable to be used more than once is via the Cont rule. In this rule, the substitution $[z/x, z/y]e$ replaces all occurrences of the placeholder variables x and y in term e by the variable z. The type of z (and its placeholders x and y) must be annotated with the usage ω. For instance, here is a type tree for the term $z + z$.

$$\dfrac{\dfrac{\dfrac{}{x : \text{Int}^\omega \vdash_\Theta x : \text{Int}^\omega}\,\text{Var} \qquad \dfrac{}{y : \text{Int}^\omega \vdash_\Theta y : \text{Int}^\omega}\,\text{Var}}{x : \text{Int}^\omega,\, y : \text{Int}^\omega \vdash_\Theta x + y : \text{Int}^j}\,\text{Plus}}{z : \text{Int}^\omega \vdash_\Theta z + z : \text{Int}^j}\,\text{Cont}$$

As one would expect, the variable z has use ω. The use variable j on the result type may be instantiated to 1 or ω, depending on how the result of the addition is used.

If a variable is never used, this is indicated by the presence of the structural rule weakening (Weak). This rule places no constraints on the use, since the use 1 (at most once) and the use ω (any number of times) are both compatible with not being used at all. However, the weakening rule may be helpful in devising a type system for strictness analysis, and is certainly important in usage analysis for call-by-value languages (see Section 6).

The last structural rule, exchange (Exch), simply indicates that the order of bindings in a context is irrelevant.

The contraction, weakening, and exchange rules are not syntax directed, but do not pose an impediment to the existence of principal types since it easy to devise an algorithm which determines whether contraction or weakening must be used on each variable, placing these rules as close to the root of the type tree as possible.

A subtlety in the manipulation of contexts is revealed by the Case rule. In the Case rule, the term e_0 is always evaluated, and then either e_1 or e_2 is evaluated. Hence it makes sense to type e_0 in a different context from e_1 and e_2, but to type e_1 or e_2 in the same context. For instance, the following is a valid typing.

$$\begin{array}{l} xs : [\text{Int}^1]^1,\, y : \text{Int}^1 \vdash_\Theta \\ \quad \text{case}\ xs\ \text{of}\ \{\text{nil} \to y;\ \text{cons}\, x\, xs' \to x + y\} : \text{Int}^1 \end{array}$$

Although y appears twice in the term, it is only labelled as being used once, which is correct because only one branch of the 'case' term will be evaluated.

3.7 Term rules

In rule Abs, the constraint $\kappa \leq |\Gamma|$ reflects the fact that if a function abstraction may be accessed more than once, then every free variable of that abstraction may be accessed more than once.

Consider again this example from the introduction.

$$\begin{array}{l} \text{let}\ x = 1 + 2\ \text{in} \\ \quad \text{let}\ f = \lambda z.\, x + z\ \text{in} \\ \qquad f\, 3 + f\, 4 \end{array}$$

Since f appears twice in $f\,3 + f\,4$ it has use ω. Since x is a free variable of a lambda abstraction with use ω, it is in turn forced to have use ω. Thus, despite appearing only once in the term, x must be labelled with use ω, as indeed it should be since it will be accessed twice in the course of evaluation.

Note that the Nil and Cons implicitly include the condition $\kappa \leq_\Theta |\tau|$ because of our global well-formedness condition on list types (see Section 3.3).

The term 'cons x y' does *not* create any closures. It does, however, refer to the variables x and y, which can be thought of as pointers to closures. Consider the following example, where the term 'cons x y' is used twice.

$$\begin{array}{l}\text{let } l = \text{cons } x\ y \text{ in} \\ \quad \text{case } l \text{ of } \cdots \text{ case } l \text{ of } \cdots\end{array}$$

Our typing rules give 'cons x y' the type $[a^\omega]^\omega$ which in turn forces x to have type a^ω and y to have type $[a^\omega]^\omega$ as expected, since both x and y may be accessed twice.

The Plus rule deserves some explanation. Our addition operator is strict, so the result of evaluating $e_0 + e_1$ will simply be an integer constant which will not refer to any part of the results of evaluating e_0 and e_1. Therefore, the usage assigned to the expression $e_0 + e_1$ need not depend at all on the usages κ_0 and κ_1. A similar argument applies to the App rule, since application is strict in its first argument.

3.8 Recursion

Finally, in a recursive definition, even a single access to a variable may allow additional accesses via the recursion (actually the use of the letrec bound variable in both the body and the letrec bound expression requires an implicit contraction). Hence in rule Letrec, the type of the recursively bound variable must have use ω.

Note that this does not mean that whenever recursion is involved that all uses must degenerate to ω. If a function is defined recursively, the argument and result of the function may still have use 1. Here is a function to append two lists.

$$\begin{array}{l}\text{letrec } \mathit{append} = \lambda xs.\,\lambda zs. \\ \quad \text{case } xs \text{ of } \{ \\ \quad\quad \text{nil} \rightarrow zs \\ \quad\quad \text{cons } y\ ys \rightarrow \text{let } as = \mathit{append}\, ys\, zs \text{ in cons } y\ as\} \\ \text{in } \cdots\end{array}$$

It has the (principal) type

$$[a^j]^k \rightarrow^\omega [a^j]^l \rightarrow^m [a^j]^l$$

with the set of constraints

$$k \leq \{j\},\ l \leq \{j\},\ m \leq \{k\}$$

The constraints $k \leq \{j\}$ and $l \leq \{j\}$ are generated by our global well-formedness condition on list types, and indicates that if the argument or result list are accessed more than once, then the elements of those lists may also be accessed more than once. The constraint $m \leq \{k\}$ is generated by the Abs rule, and indicates that if *append* is partially applied and then used more than once, the first argument list may be accessed more than once.

One instance of the above type is

$$[\mathrm{Int}^\omega]^\omega \rightarrow^\omega [\mathrm{Int}^\omega]^\omega \rightarrow^\omega [\mathrm{Int}^\omega]^\omega$$

indicating that *append* can take two lists to which there may be multiple pointers, and return a list to which there may be multiple pointers. Another instance is

$$[\mathrm{Int}^1]^1 \rightarrow^\omega [\mathrm{Int}^1]^1 \rightarrow^1 [\mathrm{Int}^1]^1$$

indicating that *append* may be applied multiple times to two lists to each of which there is only one pointer, returning a list to which there is only one pointer. (Attaching the use 1 to the second arrow guarantees that one cannot create extra pointers to the first argument list by creating and duplicating a partial application.) For this version of *append* it is possible to generate code that reuses the 'cons' cells of the first argument in producing the result.

3.9 Typeability

The ordinary rules for simply typed lambda calculus can be derived by simply omitting all use annotations and use constraints from the rules given here. It follows that if a term is typeable in this system, it is typeable in simply typed lambda calculus. Conversely, if a term is typeable in simply typed lambda calculus, then it is also typeable in this system (just take all uses to be ω).

4 Principal types and polymorphism

Before discussing what it means for a type to be principal for a given term, we first need to define when a type is an instance of another type. Our definition of instantiation is closely related to Mitchell's definition of instantiation for a type system with simple subtypes [Mit84, Mit91].

4.1 Instantiation

A substitution is a pair of finite maps. One component maps type variables to types, while the other maps use variables to uses.

Type substitutions	$TS ::=$	$\{a_1^{\kappa_1} \mapsto \tau_1, \ldots, a_n^{\kappa_n} \mapsto \tau_n\}$
Use substitutions	$US ::=$	$\{k_1 \mapsto \kappa_1, \ldots, k_n \mapsto \kappa_n\}$
Substitutions	$S ::=$	(TS, US)

Whenever a type variable is replaced by a type, the new type must have the same usage: for each $(a_i^{\kappa_i} \mapsto \tau_i) \in S$ we have $S(\kappa_i) = |\tau_i|$.

We can derive an instance of the typing derivation $\Gamma \vdash_\Theta e : \tau$ by applying a substitution S to Γ and τ, and replacing Θ with a stronger constraint set Θ'. The behaviour of S on types and contexts is defined in the usual way. We define the conditions under which Θ' is stronger than Θ (under the substitution S) below:

$\Theta' \models S\Theta$ iff for each $(j \leq \{k_1, \ldots, k_n\}) \in \Theta$ we have that $S(j) \leq_{\Theta'} S(k_i)$ for each i.

A straightforward induction on the structure of typing derivations proves the following substitution lemma.

Lemma 4.1.1 (Type substitution)
If $\Gamma \vdash_\Theta e : \tau$ *and* $\Theta' \models S\Theta$ *then* $S\Gamma \vdash_{\Theta'} e : S\tau$.

4.2 Unification

It is easy to modify Robinson's unification algorithm so that it unifies types containing usage information. However, whenever we unify a usage variable with another usage, we need to update the current constraint set. Suppose Θ is the following constraint set:

$$j \leq \{l\},\ k \leq \{m\},\ l \leq \{m\},\ m \leq \{\}$$

If we unify the types a^j and Int^k we get the substitution:

$$(\{a^j \mapsto \mathrm{Int}^k\}, \{j \mapsto k\})$$

Since we have unified j and k, we must modify the constraint set Θ so that it merges the constraints for j and k:

$$k \leq \{l, m\},\ l \leq \{m\},\ m \leq \{\}$$

Similarly, if we unify the types a^j and Int^ω we get the substitution:

$$(\{a^j \mapsto \mathrm{Int}^\omega\}, \{j \mapsto \omega\})$$

We have instantiated j to ω, which in turn forces us to also instantiate l and m to ω. The constraint $k \leq \{m\}$ then simplifies to $k \leq \{\omega\}$, which can be eliminated, since ω is the maximal usage.

If we unify usage variables with other usage variables, or with ω, we can always derive a new constraint set, as explained above. We can only fail to produce a new constraint set if we unify use variables with 1 (for instance, we might unify j with ω and l with 1, generating the unsatisfiable constraint $\omega \leq 1$). Fortunately, it is easy to show that, during type inference, we never need to make such constraints.

4.3 Principal types

Every term e has a principal type judgement $\Gamma \vdash_\Theta e : \tau$, of which all other type judgements for e are instances.

Proposition 4.3.1 (Principal types)
If $\Gamma \vdash_\Theta e : \tau$ then there exist Γ', Θ' and τ' such that $\Gamma' \vdash_{\Theta'} e : \tau'$ and for all Γ'', Θ'' and τ' such that $\Gamma'' \vdash_{\Theta''} e : \tau''$ there exists a substitution S such that $S\Gamma' \subseteq \Gamma''$, $S\tau' = \tau''$ and $\Theta'' \models S\Theta'$.

The result is proved, as usual, by exhibiting an algorithm that computes principal types.

4.4 Annotations

The operational semantics of Section 2 and the reduction rules of Section 5 require that each let-bound variable x is annotated with a use $|x|$ which is either 1 or ω. Such annotations may be inferred as follows. First, determine a principal typing for the given term, and a corresponding principal type derivation. The typing will include a constraint set Θ, and we may choose any substitution S of use variables such that $\{\} \models S\Theta$. Naturally, we choose the substitution that maps each use variable to 1, since this yields the best usage information.

Since in the end all of the use variables are set to 1, one might wonder why we bother with constraint sets at all? But a moment's thought will show that we need the constraint information in order to infer the principal typing of a term from the principal typing of its subterms. This is because in general the usage of a subterm depends on the context in which it appears, and the constraints on use variables allow us to propagate this information.

4.5 Polymorphism

The next step is to use 'let' terms to introduce polymorphism in the usual way. There are two possibilities. The first is to allow polymorphism only on type variables. For instance, the polymorphic type for *append* would be:

$$\forall a^j.\, [a^j]^k \to^\omega [a^j]^l \to^m [a^j]^l$$

with the same constraints as before. This allows *append* to be used on lists of different types, but every occurrence of *append* in the program must have the *same* usage labelling. For instance, if the labelling indicated that the first list passed to *append* always had use 1, then the code for *append* could be optimised to reuse the 'cons' cells of that list. Although crude, an analysis of this sort may be suitable for some purposes, such as removing unnecessary closures. An existing analyser for this purpose, based on abstract interpretation, has a similar limitation but has proved reasonably effective [Mar93].

The second possibility is to allow polymorphism on both type and use variables. For instance, the polymorphic type for *append* would be

$$\begin{array}{c} \forall j.\forall k \leq \{j\}.\forall l \leq \{j\}.\forall m \leq \{k\}. \\ \forall a^j.\, [a^j]^k \to^\omega [a^j]^l \to^m [a^j]^l \end{array}$$

In order to maximise the potential for optimisation, the compiler needs to generate different versions of *append* for different instantiations of the use variables. A similar technique is used in Haskell compilers to specialise code involving overloaded functions, and experience to date suggests that this is feasible and does not necessarily lead to an explosion in code size [Aug93, Jon93]. In some situations, instead of specialising *append* for different uses, we might consider having just one version of *append*, and interpret the use variables j, k, l, m as additional arguments to the *append* function, enabling run-time selection of the behaviour of *append*.

The trade-off between these two possibilities is similar to the trade-off between monovariant and polyvariant binding time analysis in partial evaluation. Further experimentation will be necessary to better understand the strengths and weaknesses of each approach.

5 Reduction and subject-reduction

We previously described the semantics of our language using Launchbury's operational semantics of call-by-need. We now give an alternative characterisation of that semantics using a modification of the call-by-need calculus of Ariola *et al.*

Working in the framework of a calculus with reduction rules simplifies our proof of subject-reduction, but more importantly, gives a set of rules which can be used by a compiler to optimise programs without danger of duplicating work (or returning the wrong result). We show how our usage information enables additional "safe" reduction rules to be formulated, allowing more aggresive optimisation when values are known to be used at most once.

5.1 Reduction rules

In the call-by-need calculus of Ariola *et al.* [AFMOW95] a closure is created for the argument of each function application, whereas in the operational semantics of Launchbury [Lau93] a closure is created only by the appearance of

Contexts	$C ::= [\,] \mid C\,x \mid \ldots \lambda x.\,C \mid \text{let } x = C \text{ in } e \mid \text{let } x = e \text{ in } C \mid \text{letrec } x = C \text{ in } e \mid \text{letrec } x = v \text{ in } C \mid$ $C + e \mid e + C \mid \text{case } C \text{ of } \{\text{nil} \to e_1;\ \text{cons } x\ y \to e_2\} \mid \text{case } e_0 \text{ of } \{\text{nil} \to C;\ \text{cons } x\ y \to e_2\} \mid$ $\text{case } e_0 \text{ of } \{\text{nil} \to e_1;\ \text{cons } x\ y \to C\}$	
Let contexts	$L ::= [\,]\,x \mid \text{let } x = [\,] \text{ in } e \mid [\,] + e \mid e + [\,] \mid \text{case } [\,] \text{ of } \{\text{nil} \to e_1;\ \text{cons } x\ y \to e_2\}$	

Figure 3: Call-by-need contexts

$n_0 + n_1$	$\Longrightarrow$	$\underline{n_0 + n_1}$	
$(\lambda x.\,e)\,y$	$\Longrightarrow$	$[y/x]e$	
$\text{case nil of } \{\text{nil} \to e_1;\ \text{cons } x\ y \to e_2\}$	$\Longrightarrow$	e_1	
$\text{case } (\text{cons } y_0\ y_1) \text{ of } \{\text{nil} \to e_1;\ \text{cons } x_0\ x_1 \to e_2\}$	$\Longrightarrow$	$[y_0/x_0, y_1/x_1]e_2$	
$\text{let } x = v \text{ in } C[x]$	$\Longrightarrow$	$\text{let } x = v \text{ in } C[v]$	if $\lvert v \rvert = \omega$
$\text{let } x = e' \text{ in } e$	$\Longrightarrow$	$\{e'/x\}e$	if $\lvert x \rvert = 1$
$\text{letrec } x = v \text{ in } C[x]$	$\Longrightarrow$	$\text{letrec } x = v \text{ in } C[v]$	
$L[\text{let } x = e' \text{ in } e]$	$\Longrightarrow$	$\text{let } x = e' \text{ in } L[e]$	
$L[\text{letrec } x = v \text{ in } e]$	$\Longrightarrow$	$\text{letrec } x = v \text{ in } L[e]$	
$\text{let } x = e' \text{ in } e$	$\Longrightarrow$	e	if $x \notin fv(e)$
$\text{letrec } x = v \text{ in } e$	$\Longrightarrow$	e	if $x \notin fv(e)$

Figure 4: Call-by-need reductions

'let'. This difference is significant: it means that the model of Ariola *et al.* may create many more closures than the model of Launchbury. For example, consider the following.

$$\begin{array}{l} \text{letrec } f = \lambda xs.\,\lambda y. \\ \quad \text{case } xs \text{ of } \{\text{nil} \to y;\ \text{cons } x\ xs' \to f\ xs'\ y\} \\ \text{in let } xs = e_0 \text{ in let } y = e_1 \text{ in } f\ xs\ y \end{array}$$

Here the model of Launchbury only creates closures for the original call, and for each element of the list e_0. In contrast, the model of Ariola *et al.* also creates two two extra closures for each recursive call of f.

Fortunately, it is straightforward to adapt the calculus of Ariola *et al.* to correspond to the model of Launchbury. The required contextual forms are given in Figure 3, and the reduction rules are given in Figure 4.

A context C is a term with a hole. Note a hole cannot appear as the argument of an application or cons, since these are restricted to variables and cannot be replaced by arbitrary terms. A let context L has a hole in a strict position (the function of an application, the selector of a case, or an argument of plus) or in the definiens of a let.

The rules are the compatible closure of the rules shown in Figure 4. That is, if $e \Longrightarrow e'$ then also $C[e] \Longrightarrow C[e']$ for any context C. Capture of free variables is disallowed, so context C should not bind x in the two rules containing C, and context L should not have x as a free variable in the two rules containing L. The rules containing L correspond to the several let-floating rules of Ariola *et al.*, which are necessary to guarantee that every closed term can be reduced to a weak head normal form.

The key change from the work of Ariola *et al.* involves the rule that allows substitution of a value,

$$\text{let } x = v \text{ in } C[x] \Longrightarrow \text{let } x = v \text{ in } C[v], \quad \text{if } |x| = \omega.$$

This rule is safe because, since v is already a value, the substitution cannot duplicate computation. However, in order to guarantee that reductions preserve use types, we must restrict this rule to the case where all free variables of v have use ω; otherwise we may duplicate a variable with use 1 resulting in an ill-typed term. An adequate restriction is to require that x has use ω, since this implies that all free variables of v also have use ω, as the reader may easily check (the two important cases are when v is a function abstraction and when v is a cons).

In the case where x has use 1, we can allow substitution of not just a value but of any expression,

$$\text{let } x = e' \text{ in } e \Longrightarrow \{e'/x\}e, \quad \text{if } |x| = 1.$$

Since we cannot substitute an expression for a variable that appears as the argument of an application or cons, we use a modified form of substitution, written $\{e'/x\}e$. The definition of $\{e'/x\}e$ is standard, except for the following two clauses:

$$\begin{array}{ll} \{e'/x\}(e\ x) & = \text{let } x = e' \text{ in } e\ x, \\ \{e'/x\}(\text{cons } y\ z) & = \text{let } x = e' \text{ in cons } y\ z, \text{ if } x = y \text{ or } x = z. \end{array}$$

Note that the substitution rules together with let-floating rules introduce possible loops in reduction sequences. For example,

$$\begin{array}{l} \text{let } x = e' \text{ in let } z = \text{cons } x\ y \text{ in } e \\ \quad \Longrightarrow \text{let } z = \text{let } x = e' \text{ in cons } x\ y \text{ in } e \\ \quad \Longrightarrow \text{let } x = e' \text{ in let } z = \text{cons } x\ y \text{ in } e. \end{array}$$

As we note below, compilers such as the Glasgow Haskell compiler are often based on sets of reductions containing such loops. Nonetheless, it would be preferable to have a reduction system without such loops, and we note this as an interesting topic for future work.

5.2 Confluence, soundness, and completeness

The following results are straightforward adaptations of the results of Ariola *et al.* The system is confluent, and sound and complete with respect to Launchbury's operational semantics.

Proposition 5.2.1 (Confluence)
If $e_0 \Longrightarrow^* e_1$ and $e_0 \Longrightarrow^* e_2$ then there exists a term e_3 such that $e_1 \Longrightarrow^* e_3$ and $e_2 \Longrightarrow^* e_3$.

If H is a heap as in the operational semantics, and e is a term, then the configuration $\langle H \rangle e$ corresponds to the term 'let $\langle H \rangle$ in e', defined as follows.

$$\begin{array}{ll} \text{let } \langle\,\rangle \text{ in } e & = e \\ \text{let } \langle \text{let } x = e', H \rangle \text{ in } e & = \text{let } x = e' \text{ in } (\text{let } \langle H \rangle \text{ in } e) \\ \text{let } \langle \text{letrec } x = v, H \rangle \text{ in } e & = \text{letrec } x = v \text{ in } (\text{let } \langle H \rangle \text{ in } e) \end{array}$$

Proposition 5.2.2 (Soundness)
If $\langle H \rangle e \Downarrow \langle H' \rangle v'$ *then* let $\langle H \rangle$ in $e \Longrightarrow^*$ let $\langle H' \rangle$ in v'

Proposition 5.2.3 (Completeness)
If let $\langle H \rangle$ in $e \Longrightarrow^*$ let $\langle H'' \rangle$ in v'' *then there exist* H' *and* v' *such that* $\langle H \rangle e \Downarrow \langle H' \rangle v'$ *and* let $\langle H' \rangle$ in $v' \Longrightarrow^*$ let $\langle H'' \rangle$ in v''.

The soundness and completeness results take an even simpler form for terms of type integer.

Corollary 5.2.4 (Soundness and completeness)
There exists a heap H' *such that* $\langle H \rangle e \Downarrow \langle H' \rangle n$ *if and only if* let $\langle H \rangle$ in $e \Longrightarrow^* n$.

NB: We have checked in detail the proof of soundness; it is a straightforward adaptation of the proof given by Ariola *et al.* We believe the proofs of confluence and completeness should also be straightforward adaptations of their proofs, but we have not checked these in detail.

5.3 Subject Reduction

Use types are preserved by reduction.

Proposition 5.3.1 (Subject reduction)
If $\Gamma \vdash_\Theta e : \tau$ *and* $e \Longrightarrow e'$ *then* $\Gamma \vdash_\Theta e' : \tau$.

The proof is straightforward, verifying each rule in Figure 4 separately, and structural induction over terms for the compatible closure.

Combining Propositions 5.3.1 and 5.2.2 yields a soundness result for our type system with respect to the operational semantics.

Corollary 5.3.2 (Operational subject reduction)
If $\vdash_\Theta$ let $\langle H \rangle$ in $e : \tau$ *and* $\langle H \rangle e \Downarrow \langle H' \rangle v$ *then* $\vdash_\Theta$ let $\langle H' \rangle$ in $v : \tau$.

5.4 Additional transformations

There are many useful reduction rules that we might add to those appearing in Figure 4.

For instance, it is helpful to have the reduction

$$\begin{array}{ll} \text{let } x = e_0 \text{ in } (\text{let } y = e_1 \text{ in } e_2) \Longrightarrow & \\ \text{let } y = (\text{let } x = e_0 \text{ in } e_1) \text{ in } e_2, & \text{if } x \notin fv(e_2), \end{array}$$

which avoids the creation of a closure for e_0 if, say, y is not evaluated at run time. By the way, note that the rule $L[\text{let } x = e' \text{ in } e] \Longrightarrow \text{let } x = e' \text{ in } L[e]$ has as an instance

$$\begin{array}{ll} \text{let } y = (\text{let } x = e_0 \text{ in } e_1) \text{ in } e_2 \Longrightarrow & \\ \text{let } x = e_0 \text{ in } (\text{let } y = e_1 \text{ in } e_2), & \text{if } x \notin fv(e_2), \end{array}$$

which is helpful if, say, further simplification reduces e_1 to a value. The Glasgow Haskell compiler makes extensive use of both reductions [San92]. This explains why we are not too bothered by the existence of loops in our reduction rules, as noted previously.

Some useful additional transformations depend on usage information. The prime example is the reduction

$$\text{let } x = e' \text{ in } (\lambda y.\, e) \Longrightarrow \lambda y.\, (\text{let } x = e' \text{ in } e), \quad \text{if } |\lambda y.\, e| = 1.$$

This requires adding a further use annotation to terms: we assume that each that each abstraction $|\lambda x.\, e|$ is annotated with a use of 1 or ω which we write $|\lambda x.\, e|$. The use information is crucial for safety: the reduction might duplicate computation if $|\lambda y.\, e| = \omega$. Again, this reduction is extensively used in the Glasgow Haskell compiler. It turns out to be particularly important for one form of deforestation [GLP93].

6 Use analysis for call-by-value calculi

Our use analysis can easily be adapted for call-by-value calculi. In such calculi, we are interested in variables which are used *exactly* once (rather than *at most* once). For example, if f is used exactly once, and x is only used in the body of f, then we can safely transform

$$\text{let } x = e \text{ in let } f = \lambda y.\, x + 3 \text{ in } \ldots$$

into

$$\text{let } f = \lambda y.\, e + 3 \text{ in } \ldots$$

(hopefully reducing the maximum amount of storage used by the program). This transformation is clearly unsafe if f is never used, since the expression e might be non-terminating. We would have transformed a non-terminating program into a terminating program.

Changing our type system to determine when a value is used *exactly* once is easy. We simply change the weakening rule as below:

$$\text{Weak}\ \frac{\Gamma \vdash e : \tau' \qquad |\tau| = \omega}{\Gamma,\, x : \tau \vdash e : \tau'}$$

We can now interpret the use 1 as meaning that a value is used *exactly* once. This transforms our type system into something much closer to a *linear* type system. In a companion paper [MOTW95], we elaborate on the connections between linear logic and call-by-value reduction, and affine logic and call-by-need reduction.

We conjecture that usage-based program transformation in the presence of side-effects could be handled by combining our usage analysis with an effect system [Luc87, LG88, JG91]. (Effect systems can distinguish side-effecting computation from purely functional computation.)

7 Related work

7.1 Linear logic

The type system presented here is based on ideas taken from the linear logic of Girard [Gir87] and its successor the Logic of Unity [Gir93]. A companion paper describes the embedding of the call-by-need calculus into linear logic that underlies the type system used here [MOTW95]. Interested readers are referred to that paper for a survey of related work on linear logic.

7.2 Call-by-need analyses

We are aware of two other analyses that attempt to determine when a value is used at most once under call-by-need evaluation. One is a type system due to Launchbury and others [Lau92], the other is an abstract interpretation due to Marlow [Mar93]. We note three points of comparison.

First, unlike ours, neither of the other analyses possess a proof of soundness. Second, our system sometimes derives more precise information than the other two; see Example 3 in Section 1.1. Third, unlike the above analyses, our type system does not detect the case where closures are *never* used (we omitted the zero usage from our analysis so as to simplify our usage constraints).

Our next step is to implement our analysis in the Glasgow Haskell compiler, allowing us to compare it directly with Marlow's. By observing how our analysis performs on real programs we can test whether omitting zero usages has a significantly impact.

7.3 Call-by-name analyses

We also are aware of two analyses that determine usage information for values under call-by-name, one due to Wright and Baker-Finch [WB93], the other due to Courtenage and Clack [CC94]. Both are based on type systems, and both have been argued to be sound. We note three points of comparison.

First, choosing call-by-name evaluation instead of call-by-need prevents even fairly simple optimisations from being discovered. For instance, in Example 1 of Section 1.1, the variable x is used once under call-by-need, as our system discovers, but twice under call-by-name. Experience suggests this difference is significant, as the situation encountered in Example 1 is fairly common.

Second, even if we are satisfied to limit our attention to call-by-name, neither system provides an especially useful analysis. The Wright and Baker-Finch system discovers too much information: function types are annotated with natural numbers which indicate the number of times a function uses its argument, and this level of accuracy renders the type system undecidable. The Courtenage and Clack system discovers too little information: they can determine when an argument is used zero times, exactly once, or at least once; but they do not determine when an argument is used at most once, which is most helpful for the problems we are interested in.

Third, unlike our system, the other systems also provide information about when a value is used at least once, which is useful for strictness analysis. For this question, the distinction between call-by-need and call-by-name is irrelevant, which may explain why the authors of these systems were willing to settle for a call-by-name analysis.

7.4 Data structure update

A number of analyses for in-place update of data structures have been proposed, including those by Schmidt [Sch85], Hudak [Hud86], Baker [Bak90], Hudak and Guzmán [GH90], and Wadler [Wad90b, Wad91]. These systems are not especially well suited for enabling program transformations or eliding closure update in call-by-need languages. Conversely, our system is not the best possible for in-place update, as it can only determine when there is at most one pointer to a structure. For many purposes, it is better to use a weaker criterion which allows multiple pointers when reading a structure but ensures there is at most one pointer when a structure is to be updated in place.

8 Conclusions

We have presented a simple type system which can determine when a value is used at most once, even in the presence of higher-order functions and data structures. Our analysis is tailored to the precise reduction strategy used in the Glasgow Haskell compiler, and therefore yields more accurate results than analyses which assume call-by-name reduction. We have proved our type system sound with respect to Launchbury's natural semantics of lazy evaluation, and have provided safe reduction rules which the compiler can use to transform programs without risking duplicating work.

A prototype type inference algorithm has already been implemented. Our next step is to incorporate our type system into the Glasgow Haskell compiler [PHHPW93]. This will enable us to measure the effect of our optimisations on large Haskell programs. The Glasgow Haskell compiler uses an explicitly-typed core language to express most of its program transformations. By adding our usage information to the core language type system we can conveniently provide information to the optimiser, enabling additional program transformations, and allowing the code generator to omit unnecessary closure updates.

Our annotated types provide a convenient way of communicating usage information across module boundaries (we simply add usage information to the user-level type information which is already exported from a module).

We intend to further explore how our type system enables in-place update of data structures. An interesting question is how much of this should be done automatically by the compiler, how much should be under the control of the user, and to what extent the type system acts as an effective mechanism to let the user understand and control optimisations.

9 Acknowledgements

The ideas described here grew out of collaboration with the Glasgow functional programming group, especially Simon Peyton Jones, John Launchbury, Andy Gill, Simon Marlow, Will Partain, Patrick Sansom, and Andre Santos. The Semantique II working group provided helpful feedback, especially Patrick Cousot, Sebastian Hunt, Torben Mogensen, Dave Sands, and Yan-Mei Tang.

References

[AFMOW95] Z. Ariola, M. Felleisen, J. Maraist, M. Odersky, and P. Wadler. A call-by-need lambda calculus. *Symposium on Principles of Programming Languages*, ACM Press, San Francisco, California, January 1995.

[Aug93] L. Augustsson, Implementing Haskell overloading. In *Functional Programming Languages and Computer Architecture*, Copenhagen, Denmark, June 1993. ACM Press.

[Bak90] H. Baker, Unify and conquer (garbage, updating, aliasing ...) in functional languages. In *Conference on Lisp and Functional Programming*, ACM Press, Nice, June 1990.

[CC94] S. A. Courtenage and C. D. Clack, Analysing resource use in the λ-calculus by type inference, In *ACM Sigplan Workshop on Partial Evaluation and Semantics-Based Program Manipulation*, 1994.

[DM82] L. Damas and R. Milner, Principal type schemes for functional programs. In *Proceedings 9'th ACM Symposium on Principles of Programming Languages*, Albuquerque, N.M., January 1982.

[Gir87] J.-Y. Girard, Linear logic. *Theoretical Computer Science*, 50:1–102, 1987.

[Gir93] J.-Y. Girard, On the unity of logic. *Annals of Pure and Applied Logic*, 59:201–217, 1993.

[GH90] J. Guzmán and P. Hudak, Single-threaded polymorphic lambda calculus. In *Proceedings 5'th IEEE Symposium on Logic in Computer Science*, Philadelphia, Pa., June 1990.

[GLP93] A. Gill, J. Launchbury, and S. Peyton Jones, A short-cut to deforestation. In *Functional Programming Languages and Computer Architecture*, Copenhagen, Denmark, June 1993. ACM Press.

[Hin69] R. Hindley, The principal type scheme of an object in combinatory logic. *Trans. Am. Math. Soc.*, 146:29–60, December 1969.

[Hud86] P. Hudak, A semantic model of reference counting and its abstraction. In *Proceedings ACM Conference on Lisp and Functional Programming*, Cambridge, Mass., August 1986.

[Jon93] M. Jones, Partial evaluation for dictionary-free overloading. Technical report TR-959, Computer Science Department, Yale University, April 1993.

[JG91] Pierre Jouvelot and D.K. Gifford, Algebraic reconstruction of types and effects. In *ACM Symposium on Principles of Programming Languages*, 1991.

[Lau92] J. Launchbury, A. Gill, J. Hughes, S. Marlow, S. Peyton Jones, and P. Wadler. Avoiding unnecessary updates. In *Glasgow Workshop on Functional Programming*, Ayr, July 1992. Springer Verlag Workshops in Computing Series.

[Lau93] J. Launchbury, A natural semantics for lazy evaluation. In *Proceedings 20'th ACM Symposium on Principles of Programming Languages*, Charleston, S.C., January 1993.

[Luc87] J.M. Lucassen, Types and effects, towards an integration of functional and imperative programming. PhD thesis, MIT Laboratory for Computer Science, 1987.

[LG88] J.M. Lucassen and D.K. Gifford, Polymorphic effect systems. In *ACM Symposium on Principles of Programming Languages*, 1988.

[Mar93] S. Marlow, Update avoidance analysis by abstract interpretation. In *Glasgow Workshop on Functional Programming*, Ayr, July 1993. Springer Verlag Workshops in Computing Series.

[Mit84] J. C. Mitchell, Coercion and type inference (summary). In *Proceedings 11'th ACM Symposium on Principles of Programming Languages*, January 1984.

[Mit91] J. C. Mitchell, Type inference with simple subtypes. *Journal of Functional Programming*, 1(3):245–286, July 1991.

[MOTW95] J. Maraist, M. Odersky, D. Turner, and P. Wadler. Call-by-name, call-by-value, call-by-need, and the linear lambda calculus. Submitted to *11'th International Conference on the Mathematical Foundations of Programming Semantics*, New Orleans, Louisiana, April 1995. (Also available by ftp to `ftp.dcs.glasgow.ac.uk`, directory `pub/glasgow-fp/authors/Philip_Wadler`, files `linearcall.dvi,ps`.)

[MW92] S. Marlow and P. Wadler, Deforestation for higher order functions. In *Glasgow Workshop on Functional Programming*, Ayr, July 1992. Springer Verlag Workshops in Computing Series.

[Pey89] S. Peyton Jones and J. Salkild, The Spineless Tagless G-machine In *Fourth International Conference on Functional Programming Languages and Computer Architecture. Imperial College, London*, pp. 184–201. ACM, Addison-Wesley, September 1989.

[Pey92] S. Peyton Jones, Implementing lazy functional languages on stock hardware: the Spineless Tagless G-machine *Journal of Functional Programming*, 2(2):127–202, April 1992.

[PHHPW93] S.L. Peyton Jones, C.V. Hall, K. Hammond, W.D. Partain, and P.L. Wadler, The Glasgow Haskell compiler: a technical overview. In *Joint Framework for Information Technology (JFIT), Technical Conference Digest*, March, 1993.

[Sch85] D. A. Schmidt, Detecting global variables in denotational specifications. *ACM Trans. on Programming Languages and Systems*, 7:299–310, 1985.

[San92] A. Santos and S. Peyton Jones In *Glasgow Workshop on Functional Programming*, Ayr, July 1992. Springer Verlag Workshops in Computing Series.

[SP95] P. Sansom and S. L. Peyton Jones, Time and space profiling for non-strict higher-order functional languages. In *22'nd ACM Symposium on Principles of Programming Languages*, San Francisco, January 1995.

[Wad90a] P. Wadler, Deforestation: Transforming programs to eliminate trees, *Theoretical Computer Science*, 73, 1990.

[Wad90b] P. Wadler, Linear types can change the world! In M. Broy and C. Jones, editors, *Programming Concepts and Methods*, Sea of Galilee, Israel, April 1990. North Holland, 1990.

[Wad91] P. Wadler, Is there a use for linear logic? In *Conference on Partial Evaluation and Semantics-Based Program Manipulation (PEPM)*, ACM Press, New Haven, Connecticut, June 1991.

[WB93] D. A. Wright and C. A. Baker-Finch, Usage analysis with natural reduction types. In *Workshop on Semantic Analysis*, 1993.