The Call-by-need Lambda Calculus, Revisited

Stephen Chang and Matthias Felleisen

College of Computer Science
Northeastern University
Boston, Massachusetts, USA
{ stchang | matthias } @ ccs.neu.edu

Abstract. The existing call-by-need λ calculi describe lazy evaluation via equational logics. A programmer can use these logics to safely ascertain whether one term is behaviorally equivalent to another or to determine the value of a lazy program. However, neither of the existing calculi models evaluation in a way that matches lazy implementations.

Both calculi suffer from the same two problems. First, the calculi never discard function calls, even after they are completely resolved. Second, the calculi include re-association axioms even though these axioms are merely administrative steps with no counterpart in any implementation.

In this paper, we present an alternative axiomatization of lazy evaluation using a single axiom. It eliminates both the function call retention problem and the extraneous re-association axioms. Our axiom uses a grammar of contexts to describe the exact notion of a *needed computation*. Like its predecessors, our new calculus satisfies consistency and standardization properties and is thus suitable for reasoning about behavioral equivalence. In addition, we establish a correspondence between our semantics and Launchbury's natural semantics.

Keywords: call-by-need, laziness, lambda calculus

1 A Short History of the λ Calculus

Starting in the late 1950s, programming language researchers began to look to Church's λ calculus [6] for inspiration. Some used it as an analytic tool to understand the syntax and semantics of programming languages, while others exploited it as the basis for new languages. By 1970, however, a disconnect had emerged in the form of call-by-value programming, distinct from the notion of β and normalization in Church's original calculus. Plotkin [25] reconciled the λ calculus and Landin's SECD machine for the ISWIM language [16] with the introduction of a notion of correspondence and with a proof that two distinct variants of the λ calculus corresponded to two distinct variants of the ISWIM programming language: one for call-by-value and one for call-by-name.

In the early 1970s, researchers proposed call-by-need [12, 14, 28], a third kind of parameter passing mechanism that could be viewed as yet another variant of the ISWIM language. Call-by-need is supposed to represent the best of both worlds. While call-by-value ISWIM always evaluates the argument of a function,

the call-by-name variant evaluates the argument every time it is needed. Hence, if an argument (or some portion) is never needed, call-by-name wins; otherwise call-by-value is superior because it avoids re-evaluation of arguments. Call-by-need initially proceeds like call-by-name, evaluating a function's body before the argument—until the value of the argument is needed; at that point, the argument is evaluated and the resulting value is used from then onward. In short, call-by-need evaluates an argument at most once, and only if needed.

Since then, researchers have explored a number of characterizations of call-by-need [8, 11, 13, 15, 23, 24, 26]. Concerning this paper, three stand out. Launchbury's semantics [17] specifies the meaning of complete programs with a Kahn-style natural semantics. The call-by-need λ calculi of Ariola and Felleisen [2–4], and of Maraist, Odersky, and Wadler [4, 20, 21] are equational logics in the spirit of the λ calculus.

The appeal of the λ calculus has several reasons. First, a calculus is sound with respect to the observational (behavioral) equivalence relation [22]. It can therefore serve as the starting point for other, more powerful logics. Second, its axioms are rich enough to mimic machine evaluation, meaning programmers can reduce programs to values without thinking about implementation details. Finally, the λ calculus gives rise to a substantial meta-theory [5, 7] from which researchers have generated useful and practical results for its cousins.

Unfortunately, neither of the existing by-need calculi model lazy evaluation in a way that matches lazy language implementations. Both calculi suffer from the same two problems. First, unlike the by-name and by-value calculi, the by-need calculi never discard function calls, even after the call is resolved and the argument is no longer needed. Lazy evaluation does require some accumulation of function calls due to the delayed evaluation of arguments but the existing calculi adopt the extreme solution of retaining every call. Indeed, the creators of the existing calculi acknowledge that a solution to this problem would strengthen their work but they could not figure out a proper solution.

Second, the calculi include re-association axioms even though these axioms have no counterpart in any implementation. The axioms are mere administrative steps, needed to construct β-like redexes. Hence, they should not be considered computationally on par with other axioms.

In this paper, we overcome these problems with an alternative axiomatization. Based on a single axiom, it avoids the retention of function calls and eliminates the extraneous re-association axioms. The single axiom uses a grammar of contexts to describe the exact notion of a *needed computation*. Like its predecessors, our new calculus satisfies consistency and standardization properties and is thus suitable for reasoning about behavioral equivalence. In addition, we establish an intensional correspondence with Launchbury's semantics.

The second section of this paper recalls the two existing by-need calculi in some detail. The third section presents our new calculus, as well as a way to derive it from Ariola and Felleisen's calculus. Sections 4 and 5 show that our calculus satisfies the usual meta-theorems and that it is correct with respect to Launchbury's semantics. Finally, we discuss some possible extensions.

2 The Original Call-by-need λ Calculi

The original call-by-need λ calculi are independently due to two groups: Ariola and Felleisen [2, 3] and Maraist, et al. [20, 21]. They were jointly presented at POPL in 1995 [4]. Both calculi use the standard set of terms as syntax:

$$e = x \mid \lambda x.e \mid e\,e \qquad \text{(Terms)}$$

Our treatment of syntax employs the usual conventions, including Barendregt's standard hygiene condition for variable bindings [5]. Figure 1 specifies the calculus of Maraist et al., λ_{mow}, and λ_{af}, Ariola and Felleisen's variant. Nonterminals in some grammar productions have subscript tags to differentiate them from similar sets elsewhere in the paper. Unsubscripted definitions have the same denotation in all systems.

$v_m = x \mid \lambda x.e$	$v = \lambda x.e$
	$a_{af} = v \mid (\lambda x.a_{af})\,e$
$C = [\ ] \mid \lambda x.C \mid C\,e \mid e\,C$	$E_{af} = [\ ] \mid E_{af}\,e \mid (\lambda x.E_{af})\,e \mid (\lambda x.E_{af}[x])\,E_{af}$
$(\lambda x.C[x])\,v_m = (\lambda x.C[v_m])\,v_m \quad (\mathcal{V})$	$(\lambda x.E_{af}[x])\,v = (\lambda x.E_{af}[v])\,v \quad (\mathit{deref})$
$(\lambda x.e_1)\,e_2\,e_3 = (\lambda x.e_1\,e_3)\,e_2 \quad (\mathcal{C})$	$(\lambda x.a_{af})\,e_1\,e_2 = (\lambda x.a_{af}\,e_2)\,e_1 \quad (\mathit{lift})$
$(\lambda x.e_1)((\lambda y.e_2)\,e_3) = \quad (\mathcal{A})$	$(\lambda x.E_{af}[x])\,((\lambda y.a_{af})\,e) = \quad (\mathit{assoc})$
$(\lambda y.(\lambda x.e_1)\,e_2)\,e_3$	$(\lambda y.(\lambda x.E_{af}[x])\,a_{af})\,e$
$(\lambda x.e_1)\,e_2 = e_1, x \notin fv(e_1) \quad (\mathcal{G})$	

Fig. 1. Existing call-by-need λ calculi (left: λ_{mow}, right: λ_{af})

In both calculi, the analog to the β axiom—also called a *basic notion of reduction* [5]—replaces variable occurrences, one at a time, with the value of the function's argument. Value substitution means that there is no duplication of work as far as argument evaluation is concerned. The function call is retained because additional variable occurrences in the function body may need the argument. Since function calls may accumulate, the calculi come with axioms that re-associate bindings to pair up functions with their arguments. For example, re-associating $(\lambda x.(\lambda y.\lambda z.z)\,v_y)\,v_x\,v_z$ in λ_{af} exposes a *deref* redex:

$$(\lambda x.(\lambda y.\underline{\lambda z.z})\,v_y)\,v_x\,\underline{v_z} \overset{\mathit{lift}}{\to} (\lambda x.(\lambda y.\underline{\lambda z.z})\,v_y\,\underline{v_z})\,v_x \overset{\mathit{lift}}{\to} (\lambda x.(\lambda y.\underline{(\lambda z.z)\,v_z})\,v_y)\,v_x$$

The two calculi differ from each other in their timing of variable replacements. The λ_{mow} calculus allows the replacement of a variable with its value anywhere in the body of its binding λ. The λ_{af} calculus replaces a variable with its argument only if evaluation of the function body needs it, where "need" is formalized via so-called evaluation contexts (E_{af}). Thus evaluation contexts in λ_{af} serve the

double purpose of specifying demand for arguments and the standard reduction strategy. The term $(\lambda x.\lambda y.x)\, v$ illustrates this difference between the two calculi. According to λ_{mow}, the term is a $\mathcal{V}$ redex and reduces to $(\lambda x.\lambda y.v)\, v$, whereas in λ_{af}, the term is irreducible because the x occurs in an inner, unapplied λ, and is thus not "needed."

Also, λ_{mow} is more lenient than λ_{af} when it comes to re-associations. The λ_{af} calculus re-associates the left or right hand side of an application only if it has been completely reduced to an answer, but λ_{mow} permits re-association as soon as one nested function layer is revealed. In short, λ_{mow} proves more equations than λ_{af}, i.e., $\boldsymbol{\lambda_{af} \subset \lambda_{mow}}$.

In λ_{af}, programs reduce to answers:

$$\texttt{eval}_{af}(e) = \texttt{done} \text{ iff there exists an answer } a_{af} \text{ such that } \boldsymbol{\lambda_{af}} \vdash e = a_{af}$$

In contrast, Maraist et al. introduce a "garbage collection" axiom into λ_{mow} to avoid answers and to use values instead. This suggests the following definition:

$$\texttt{eval}_{mow}(e) = \texttt{done} \text{ iff there exists a value } v_m \text{ such that } \boldsymbol{\lambda_{mow}} \vdash e = v_m$$

This turns out to be incorrect, however. Specifically, let $\texttt{eval}_{name}$ be the analogous call-by-name evaluator. Then $\texttt{eval}_{af} = \texttt{eval}_{name}$ but $\texttt{eval}_{mow} \neq \texttt{eval}_{name}$. Examples such as $(\lambda x.\lambda y.x)\, \Omega$ confirm the difference.

In recognition of this problem, Maraist et al. use Ariola and Felleisen's axioms and evaluation contexts to create their Curry-Feys-style standard reduction sequences. Doing so reveals the inconsistency of λ_{mow} with respect to Plotkin's *correspondence criteria* [25]. According to Plotkin, a useful calculus *corresponds to* a programming language, meaning its axioms (1) satisfy the Church-Rosser and Curry-Feys Standardization properties, and (2) define a standard reduction function that is equal to the evaluation function of the programming language. Both the call-by-name and the call-by-value λ calculi satisfy these criteria with respect to call-by-name and call-by-value SECD machines for ISWIM, respectively. So does λ_{af} with respect to a call-by-need SECD machine, but some of λ_{mow}'s axioms cannot be used as standard reduction relations.

Finally, the inclusion of $\mathcal{G}$ is a brute-force attempt to address the function call retention problem. Because $\mathcal{G}$ may discard arguments even before the function is called, both sets of authors consider it too coarse and acknowledge that a tighter solution to the function call retention issue would "strengthen the calculus and its utility for reasoning about the implementations of lazy languages" [4].

3 A New Call-by-need λ Calculus

Our new calculus, λ_{need}, uses a single axiom, β_{need}. The new axiom evaluates the argument when it is first demanded, replaces all variable occurrences with that result, and then discards the argument and thus the function call. In addition, the axiom performs the required administrative scope adjustments as part of the same step, rendering explicit re-association axioms unnecessary. In short, every reduction step in our calculus represents computational progress.

Informally, to perform a reduction, three components must be identified:

1. the next demanded variable,
2. the function that binds that demanded variable,
3. and the argument to that function.

In previous by-need calculi the re-association axioms rewrite a term so that the binding function and its argument are adjacent.

Without the re-association axioms, finding the function that binds the demanded variable and its argument requires a different kind of work. The following terms show how the demanded variable, its binding function, and its argument can appear at seemingly arbitrary locations in a program:

- $(\underline{\lambda x}.(\lambda y.\lambda z.\underline{x})\, e_y)\, \underline{e_x}\, e_z$
- $(\lambda x.(\underline{\lambda y}.\lambda z.\underline{y})\, \underline{e_y})\, e_x\, e_z$
- $(\lambda x.(\lambda y.\underline{\lambda z}.\underline{z})\, e_y)\, e_x\, \underline{e_z}$

Our β_{need} axiom employs a grammar of contexts to describe the path from a demanded variable to its binding function and from there to its argument.

The first subsection explains the syntax and the contexts of λ_{need} in a gradual fashion. The second subsection presents the β_{need} axiom and also shows how to derive it from Ariola and Felleisen's λ_{af} calculus.

3.1 Contexts

Like the existing by-need calculi, the syntax of our calculus is that of Church's original calculus. In λ_{need}, calculations evaluate terms e to answers $A[v]$, which generalize answers from Ariola and Felleisen's calculus:

$$e = x \mid \lambda x.e \mid e\, e \qquad \text{(Terms)}$$
$$v = \lambda x.e \qquad \text{(Values)}$$
$$a = A[v] \qquad \text{(Answers)}$$
$$A = [\ ] \mid A[\lambda x.A]\, e \qquad \text{(Answer Contexts)}$$

Following Ariola and Felleisen, the basic axiom uses evaluation contexts to specify the notion of demand for variables:

$$E = [\ ] \mid E\, e \mid \ldots \qquad \text{(Evaluation Contexts)}$$

The first two kinds, taken from λ_{af}, specify that a variable is demanded, and that a variable in the operator position of an application is demanded, respectively.

Since the calculus is to model program evaluation, we are primarily interested in demanded variables under a λ-abstraction. This kind of evaluation context is defined using an answer context A:

$$E = \ldots \mid A[E] \mid \ldots \qquad \text{(Another Evaluation Context)}$$

Using answer contexts, this third evaluation context dictates that demand exists under a λ if a corresponding argument exists for that λ. Note how *an answer context descends under the same number of λs as arguments for those λs.* In particular, for any term $A[\lambda x.e_1]\,e_2$, e_2 is always the argument of $\lambda x.e_1$. The third evaluation context thus generalizes the function-is-next-to-argument requirement found in both call-by-name and call-by-value. The generalization is needed due to the retention of function calls in λ_{need}.

Here are some example answer contexts that might be used:

$$A_0 = \underbrace{(\lambda x.[\ ])\,e_x}$$

$$A_1 = \underbrace{(\lambda x.\,\underbrace{(\lambda y.[\ ])\,e_y})\,e_x}$$

$$A_2 = \underbrace{(\lambda x.\,\underbrace{(\lambda y.\,\underbrace{(\lambda z.[\ ])\,e_z})\,e_y})\,e_x}$$

An underbrace matches each function to its argument. The examples all juxtapose functions and their arguments. In contrast, the next two separate functions from their arguments:

$$A_3 = \underbrace{\underbrace{(\lambda x.\lambda y.\lambda z.[\ ])\,e_x}\ e_y}\,e_z$$

$$A_4 = (\lambda x.\underbrace{\underbrace{(\lambda y.\lambda z.[\ ])\,e_y}\,e_x}\,e_z$$

To summarize thus far, when a demanded variable is discovered under a λ, the surrounding context looks like this:

$$A[E[x]]$$

where both the function binding x and its argument are in A. The decomposition of the surrounding context into A and E assumes that A encompasses as many function-argument pairs as possible; in other words, it is impossible to merge the outer part of E with A to form a larger answer context.

To know which argument corresponds to the demanded variable, we must find the λ that binds x in A. To this end, we split answer contexts so that we can "highlight" a function-argument pair within the context:

$$\hat{A} = [\ ] \mid A[\hat{A}]\,e \qquad \text{(Partial Answer Contexts–Outer)}$$

$$\check{A} = [\ ] \mid A[\lambda x.\check{A}] \qquad \text{(Partial Answer Contexts–Inner)}$$

Using these additional contexts, any answer context can be decomposed into

$$\hat{A}[A[\lambda x.\check{A}[\ ]]\,e]$$

where e is the argument of $\lambda x.\check{A}[\ ]$. For a fixed function-argument pair in an answer context, this partitioning into $\hat{A}$, A, and $\check{A}$ is unique. The $\hat{A}$ subcontext represents the part of the answer context around the chosen function-argument pair; the $\check{A}$ subcontext represents the part of the answer context in its body; and A here is the subcontext between the function and its argument. Naturally we must demand that $\hat{A}$ composed with $\check{A}$ is an answer context as well so that the overall context remains an answer context. The following table lists the various subcontexts for the example A_4 for various function-argument pairs:

	$A_4 = (\lambda x.(\lambda y.\lambda z.[\ ])\, e_y)\, e_x\, e_z$		
$\hat{A} =$	$[\ ]\, e_z$	$(\lambda x.[\ ])\, e_x\, e_z$	$[\ ]$
$A =$	$[\ ]$	$[\ ]$	$(\lambda x.(\lambda y.[\ ])\, e_y)\, e_x$
$\check{A} =$	$(\lambda y.\lambda z.[\ ])\, e_y$	$\lambda z.[\ ]$	$[\ ]$
$A_4 =$	$\hat{A}[A[\lambda x.\check{A}]\, e_x]$	$\hat{A}[A[\lambda y.\check{A}]\, e_y]$	$\hat{A}[A[\lambda z.\check{A}]\, e_z]$

Now we can define the fourth kind of evaluation context:

$$E = \ldots \mid \hat{A}[A[\lambda x.\check{A}[E[x]]]\, E], \quad \text{where } \hat{A}[\check{A}] \in A \qquad \text{(Final Eval. Context)}$$

This final evaluation context shows how demand shifts to an argument when a function parameter is in demand within the function body.

3.2 The β_{need} Axiom and a Derivation

Figure 2 summarizes the syntax of λ_{need} as developed in the preceding section.[1] In this section we use these definitions to formulate the β axiom for our calculus.

$$\begin{aligned}
e &= x \mid \lambda x.e \mid e\, e && \text{(Terms)}\\
v &= \lambda x.e && \text{(Values)}\\
a &= A[v] && \text{(Answers)}\\
A &= [\ ] \mid A[\lambda x.A]\, e && \text{(Answer Contexts)}\\
\hat{A} &= [\ ] \mid A[\hat{A}]\, e && \text{(Partial Answer Contexts–Outer)}\\
\check{A} &= [\ ] \mid A[\lambda x.\check{A}] && \text{(Partial Answer Contexts–Inner)}\\
E &= [\ ] \mid E\, e \mid A[E] \mid \hat{A}[A[\lambda x.\check{A}[E[x]]]\, E], && \text{(Evaluation Contexts)}\\
&\qquad \text{where } \hat{A}[\check{A}] \in A
\end{aligned}$$

Fig. 2. The syntax and contexts of the new call-by-need λ calculus, λ_{need}.

Here is the single axiom of λ_{need}:

$$\hat{A}[A_1[\lambda x.\check{A}[E[x]]]\, A_2[v]] = \hat{A}[A_1[A_2[\check{A}[E[x]]\{x := v\}]]], \qquad (\beta_{need})$$
$$\text{where } \hat{A}[\check{A}] \in A$$

[1] We gratefully acknowledge Casey Klein's help with the A production.

A β_{need} redex determines which parameter x of some function is "in demand" and how to locate the corresponding argument $A_2[v]$, which might be an answer not necessarily a value. The contexts from the previous section specify the path from the binding position (λ) to the variable occurrence and the argument. A β_{need} reduction substitutes the value in $A_2[v]$ for all free occurrences of the function parameter—just like in other λ calculi. In the process, the function call is discarded. Since the argument has been reduced to a value, there is no duplication of work, meaning our calculus satisfies the requirements of lazy evaluation. Lifting A_2 to the top of the evaluation context ensures that its bindings remain intact and visible for v.

Here is a sample reduction in λ_{need}, where $\longrightarrow$ is the one-step reduction:

$$
\begin{aligned}
& ((\lambda x.(\lambda y.\underline{\lambda z.\mathbf{z}\ y\ x})\ \lambda y.y)\ \lambda x.x)\ \underline{\lambda z.z} && (1)\\
\longrightarrow\ & (\lambda x.(\underline{\lambda y.(\lambda z.\underline{z})\ \mathbf{y}\ x})\ \underline{\lambda y.y})\ \lambda x.x && (2)\\
\longrightarrow\ & (\lambda x.((\underline{\lambda z.\mathbf{z}})\ \underline{\lambda y.y})\ x)\ \lambda x.x && (3)\\
\longrightarrow\ & (\underline{\lambda x.(\lambda y.\underline{y})\ \mathbf{x}})\ \underline{\lambda x.x} && (4)
\end{aligned}
$$

The "in demand" variable is in bold; its binding λ and argument are underlined. Line 1 is an example of a reduction that involves a non-adjoined function and argument pair. In line 2, the demand for the value of z (twice underlined) triggers a demand for the value of y; line 4 contains a similar demand chain.

$$
\begin{aligned}
v &= \lambda x.e && \text{(Values)}\\
a_{af} &= A_{af}[v] && \text{(Answers)}\\
A_{af} &= [\ ] \mid (\lambda x.A_{af})\ e && \text{(Answer Contexts)}\\
E_{af} &= [\ ] \mid E_{af}\ e \mid A_{af}[E_{af}] \mid (\lambda x.E_{af}[x])\ E_{af} && \text{(Evaluation Contexts)}
\end{aligned}
$$

$$
\begin{aligned}
(\lambda x.E_{af}[x])\ v &= E_{af}[x]\{x := v\} && (\beta'_{need})\\
(\lambda x.A_{af}[v])\ e_1\ e_2 &= (\lambda x.A_{af}[v\ e_2])\ e_1 && (\mathit{lift}')\\
(\lambda x.E_{af}[x])\ ((\lambda y.A_{af}[v])\ e) &= (\lambda y.A_{af}[(\lambda x.E_{af}[x])\ v])\ e && (\mathit{assoc}')
\end{aligned}
$$

Fig. 3. A modified calculus, $\lambda_{af\text{-}mod}$.

To furnish additional intuition into β_{need}, we use the rest of the section to derive it from the axioms of λ_{af}. The $\lambda_{af\text{-}mod}$ calculus in figure 3 combines λ_{af} with two insights. First, Garcia et al. [13] observed that when the answers in λ_{af}'s *lift* and *assoc* redexes are nested deeply, multiple re-associations are performed consecutively. Thus we modify *lift* and *assoc* to perform all these re-associations in one step.[2] The modified calculus defines answers via answer contexts, A_{af}, and the modified *lift*′ and *assoc*′ axioms utilize these answer contexts to do the multi-step re-associations. Thus programs in this modified calculus reduce to answers

[2] The same modifications cannot be applied to $\mathcal{C}$ and $\mathcal{A}$ in λ_{mow} because they allow earlier re-association and thus not all the re-associations are performed consecutively.

$A_{af}[v]$. Also, the A_{af} answer contexts are identical to the third kind of evaluation context in $\lambda_{af\text{-}mod}$ and the new definition of E_{af} reflects this relationship.

Second, Maraist et al. [19] observed that once an argument is reduced to a value, all substitutions can be performed at once. The β'_{need} axiom exploits this idea and performs a full substitution. Obviously β'_{need} occasionally performs more substitutions than *deref*. Nevertheless, any term with an answer in λ_{af} likewise has an answer when reducing with β'_{need}.

Next an inspection of the axioms shows that the contractum of a *assoc*′ redex contains a β'_{need} redex. Thus the *assoc*′ re-associations and β'_{need} substitutions can be performed with one merged axiom:[3]

$$(\lambda x.E_{af}[x])\, A_{af}[v] = A_{af}[E_{af}[x]\{x := v\}] \qquad (\beta''_{need})$$

The final step is to merge *lift*′ with β''_{need}, which requires our generalized answer and evaluation contexts. A naïve attempt may look like this:

$$A_1[\lambda x.E[x]]\, A_2[v] = A_1[A_2[E[x]\{x := v\}]] \qquad (\beta'''_{need})$$

As the examples in the preceding subsection show, however, the binding occurrence for the "in demand" parameter x may not be the inner-most binding λ once the re-association axioms are eliminated. That is, in comparison with β_{need}, β'''_{need} incorrectly assumes E is always next to the binder. We solve this final problem with the introduction of partial answer contexts.

4 Consistency, Determinism, and Soundness

If a calculus is to model a programming language, it must satisfy some essential properties, most importantly a Church-Rosser theorem and a Curry-Feys standardization theorem [25]. The former guarantees consistency of evaluation; that is, we can define an evaluator *function* with the calculus. The latter implies that the calculus comes with a *deterministic* evaluation strategy. Jointly these properties imply the calculus is *sound* with respect to observational equivalence.

4.1 Consistency: Church-Rosser

The λ_{need} calculus defines an evaluator for a by-need language:

$$\texttt{eval}_{need}(e) = \texttt{done} \text{ iff there exists an answer } a \text{ such that } \boldsymbol{\lambda_{need}} \vdash e = a$$

To prove that the evaluator is indeed a (partial) function, we prove that the notion of reduction satisfies the Church-Rosser property.

Theorem 1. *$\texttt{eval}_{need}$ is a partial function.*

Proof. The theorem is a direct consequence of lemma 1 (Church-Rosser).

[3] Danvy et al. [8] dub a β''_{need} redex a "potential redex" in unrelated work.

Our strategy is to define a parallel reduction relation for λ_{need} [5]. Define $\rightarrow$ to be the compatible closure of a β_{need} reduction, and $\twoheadrightarrow$ to be the reflexive, transitive closure of $\rightarrow$. Additionally, define $\Rightarrow$ to be the relation that reduces β_{need} redexes in parallel.

Definition 1 ($\Rightarrow$).

$$
\begin{aligned}
e &\Rightarrow e \\
\hat{A}[A_1[\lambda x.\check{A}[E[x]]]\, A_2[v]] &\Rightarrow \hat{A}'[A_1'[A_2'[\check{A}'[E'[x]]\{x := v'\}]]], \\
&\quad \textit{if } \hat{A}[\check{A}] \in A,\ \hat{A}'[\check{A}'] \in A,\ \hat{A} \Rightarrow \hat{A}',\ A_1 \Rightarrow A_1', \\
&\quad\quad A_2 \Rightarrow A_2',\ \check{A} \Rightarrow \check{A}',\ E \Rightarrow E',\ v \Rightarrow v' \\
e_1\, e_2 &\Rightarrow e_1'\, e_2',\ \textit{if } e_1 \Rightarrow e_1',\ e_2 \Rightarrow e_2' \\
\lambda x.e &\Rightarrow \lambda x.e',\ \textit{if } e \Rightarrow e'
\end{aligned}
$$

The parallel reduction relation $\Rightarrow$ relies on notion of parallel reduction for contexts; for simplicity, we overload the relation symbol to denote both relations.

Definition 2 ($\Rightarrow$ for Contexts).

$$
\begin{aligned}
[\ ] &\Rightarrow [\ ] \\
A_1[\lambda x.A_2]\, e &\Rightarrow A_1'[\lambda x.A_2']\, e',\ \textit{if } A_1 \Rightarrow A_1',\ A_2 \Rightarrow A_2',\ e \Rightarrow e' \\
A[\hat{A}]\, e &\Rightarrow A'[\hat{A}']\, e',\ \textit{if } A \Rightarrow A',\ \hat{A} \Rightarrow \hat{A}',\ e \Rightarrow e' \\
A[\lambda x.\check{A}] &\Rightarrow A'[\lambda x.\check{A}'],\ \textit{if } A \Rightarrow A',\ \check{A} \Rightarrow \check{A}' \\
E\, e &\Rightarrow E'\, e',\ \textit{if } E \Rightarrow E',\ e \Rightarrow e' \\
A[E] &\Rightarrow A'[E'],\ \textit{if } A \Rightarrow A',\ E \Rightarrow E' \\
\hat{A}[A[\lambda x.\check{A}[E_1[x]]]\, E_2] &\Rightarrow \hat{A}'[A'[\lambda x.\check{A}'[E_1'[x]]]\, E_2'], \\
&\quad \textit{if } \hat{A}[\check{A}] \in A,\ \hat{A} \Rightarrow \hat{A}',\ A \Rightarrow A', \\
&\quad\quad \check{A} \Rightarrow \check{A}',\ E_1 \Rightarrow E_1',\ E_2 \Rightarrow E_2'
\end{aligned}
$$

Lemma 1 (Church-Rosser). *If $e \twoheadrightarrow e_1$ and $e \twoheadrightarrow e_2$, then there exists a term e' such that $e_1 \twoheadrightarrow e'$ and $e_2 \twoheadrightarrow e'$.*

Proof. By lemma 2, $\Rightarrow$ satisfies a diamond property. Since $\Rightarrow$ extends $\rightarrow$, $\twoheadrightarrow$ is also the transitive-reflexive closure of $\Rightarrow$, so $\twoheadrightarrow$ also satisfies a diamond property.

Lemma 2 (Diamond Property of $\Rightarrow$). *If $e \Rightarrow e_1$ and $e \Rightarrow e_2$, there exists e' such that $e_1 \Rightarrow e'$ and $e_2 \Rightarrow e'$.*

Proof. The proof proceeds by structural induction on the derivation of $e \Rightarrow e_1$.

4.2 Deterministic Behavior: Standard Reduction

A language calculus should also come with a deterministic algorithm for applying the reductions to evaluate a program. Here is our *standard reduction*:

$$E[e] \longmapsto E[e'],\ \text{where } e\ \beta_{need}\ e'$$

Our standard reduction strategy picks exactly one redex in a term.

Proposition 1 (Unique Decomposition). *For all closed terms e, e either is an answer or $e = E[e']$ for a unique evaluation context E and β_{need} redex e'.*

Proof. The proof proceeds by structural induction on e.

Since our calculus satisfies the unique decomposition property, we can use the standard reduction relation to define a (partial) evaluator function:

$$\texttt{eval}^{\text{sr}}_{need}(e) = \texttt{done} \text{ iff there exists an answer } a \text{ such that } e \longmapsto\!\!\!\!\!\rightarrow a$$

where $\longmapsto\!\!\!\!\!\rightarrow$ is the reflexive, transitive closure of $\longmapsto$. Proposition 1 shows $\texttt{eval}^{\text{sr}}_{need}$ is a function. The following theorem confirms that it equals $\texttt{eval}_{need}$.

Theorem 2. $\mathit{eval}_{need} = \mathit{eval}^{sr}_{need}$

Proof. The theorem follows from lemma 3, which shows how to obtain a standard reduction sequence for any arbitrary reduction sequence. The front-end of the former is a series of standard reduction steps.

Definition 3 (Standard Reduction Sequences $\mathcal{R}$).

- $x \subset \mathcal{R}$
- $\lambda x.e_1 \diamond \cdots \diamond \lambda x.e_m \in \mathcal{R}$, *if* $e_1 \diamond \cdots \diamond e_m \in \mathcal{R}$
- $e_0 \diamond e_1 \diamond \cdots \diamond e_m \in \mathcal{R}$, *if* $e_0 \longmapsto e_1$ *and* $e_1 \diamond \cdots \diamond e_m \in \mathcal{R}$
- $(e_1\, e'_1) \diamond \cdots \diamond (e_m\, e'_1) \diamond (e_m\, e'_2) \diamond \cdots \diamond (e_m\, e'_n) \in \mathcal{R}$, *if* $e_1 \diamond \cdots \diamond e_m, e'_1 \diamond \cdots \diamond e'_n \in \mathcal{R}$.

Lemma 3 (Curry-Feys Standardization). $e \twoheadrightarrow e'$ *iff there exists* $e_1 \diamond \cdots \diamond e_n \in \mathcal{R}$ *such that* $e = e_1$ *and* $e' = e_n$.

Proof. Replace $\twoheadrightarrow$ with $\Rightarrow$s, and the lemma immediately follows from lemma 4.

The key to the remaining proofs is a size metric for parallel reductions.

Definition 4 (Size of $\Rightarrow$ Reduction).

$$\begin{array}{ll}
\left|e \Rightarrow e\right| & = 0 \\
\left|(e_1\, e_2) \Rightarrow (e'_1\, e'_2)\right| & = \left|e_1 \Rightarrow e'_1\right| + \left|e_2 \Rightarrow e'_2\right| \\
\left|\lambda x.e \Rightarrow \lambda x.e'\right| & = \left|e \Rightarrow e'\right| \\
\left|r\right| & = 1 + \left|\hat{A} \Rightarrow \hat{A}'\right| + \left|A_1 \Rightarrow A'_1\right| + \left|\check{A}[E[x]] \Rightarrow \check{A}'[E'[x]]\right| + \\
 & \qquad \left|A_2 \Rightarrow A'_2\right| + \#(x, \check{A}'[E'[x]]) \times \left|v \Rightarrow v'\right| \\
\multicolumn{2}{l}{\quad \textit{where } r = \hat{A}[A_1[\lambda x.\check{A}[E[x]]]\, A_2[v]] \Rightarrow \hat{A}'[A'_1[A'_2[\check{A}'[E'[x]]\{x := v'\}]]]} \\
\#(x, e) & = \textit{the number of free occurrences of } x \textit{ in } e
\end{array}$$

The size of a parallel reduction of a context equals the sum of the sizes of the parallel reductions of the subcontexts and subterms that comprise the context.

Lemma 4. *If* $e_0 \Rightarrow e_1$ *and* $e_1 \diamond \cdots \diamond e_n \in \mathcal{R}$, *there exists* $e_0 \diamond e'_1 \diamond \cdots \diamond e'_p \diamond e_n \in \mathcal{R}$.

Proof. By triple lexicographic induction on (1) length n of the given standard reduction sequence, (2) $\left|e_0 \Rightarrow e_1\right|$, and (3) structure of e_0.[4]

[4] We conjecture that the use of Ralph Loader's technique [18] may simplify our proof.

4.3 Observational Equivalence

Following Morris [22] two expressions e_1 and e_2 are observationally equivalent, $e_1 \simeq e_2$, if they are indistinguishable in all contexts. Formally, $e_1 \simeq e_2$ if and only if $\texttt{eval}_{need}(C[e_1]) = \texttt{eval}_{need}(C[e_2])$ for all contexts C, where

$$C = [\ ] \mid \lambda x.C \mid C\,e \mid e\,C \qquad \text{(Contexts)}$$

An alternative definition of the behavioral equivalence relation uses co-induction. In either case, λ_{need} is sound with respect to observational equivalence.

Theorem 3 (Soundness). *If* $\boldsymbol{\lambda_{need}} \vdash e_1 = e_2$*, then* $e_1 \simeq e_2$*.*

Proof. Following Plotkin, a calculus is sound if it satisfies Church-Rosser and Curry-Feys theorems.

5 Correctness

Ariola and Felleisen [3] prove that λ_{af} defines the same evaluation function as the call-by-name λ calculus. Nakata and Hasegawa [23] additionally demonstrate extensional correctness of the same calculus with respect to Launchbury's natural semantics [17]. In this section, we show that λ_{need} defines the same evaluation function as Launchbury's semantics. While our theorem statement is extensional, the proof illuminates the tight intensional relationship between the two systems.

5.1 Overview

The gap between the λ_{need} standard reduction "machine" and Launchbury's natural semantics is huge. While the latter's store-based natural semantics uses the equivalent of assignment statements to implement the "evaluate once, only when needed" policy, the λ_{need} calculus exclusively relies on term substitutions. To close the gap, we systematically construct a series of intermediate systems that makes comparisons easy, all while ensuring correctness at each step. A first step is to convert the natural semantics into a store-based machine [27].

To further bridge the gap we note that a single-use assignment statement is equivalent to a program-wide substitution of shared expressions [10]. A closely related idea is to reduce shared expressions simultaneously. This leads to a parallel program rewriting system, dubbed $\lambda_{\|}$. Equipped with $\lambda_{\|}$ we get closer to λ_{need} but not all the way there because reductions in λ_{need} and $\lambda_{\|}$ are too coarse-grained for direct comparison. Fortunately, it is easy to construct an intermediate transition system that eliminates the remainder of the gap. We convert λ_{need} to an equivalent CK transition system [9], where the program is partitioned into a control string (C) and an explicit context (K) and we show that there is a correspondence between this transition system and $\lambda_{\|}$.

Figure 4 outlines our proof strategy pictorially. The four horizontal layers correspond to the four rewriting systems. While λ_{need} and $\lambda_{\|}$ use large steps to

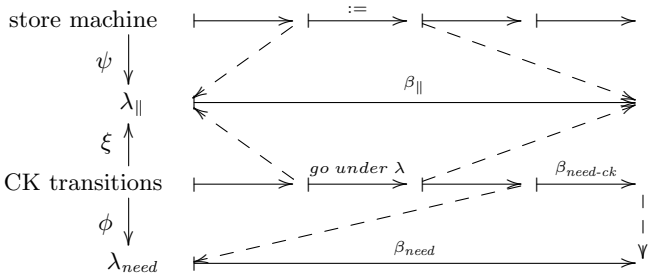

Fig. 4. Summary of correctness proof technique.

progress from term to term, the machine-like systems take several small steps. The solid vertical arrows between the layers figure indicate how mapping functions relate the rewriting sequences and the dashed arrows show how the smaller machine steps correspond to the larger steps of λ_{need} and $\lambda_{\|}$:

- The ψ function maps states from the store-based machine to terms in the $\lambda_{\|}$ world. For every step in the natural-semantics machine, the resulting operation in $\lambda_{\|}$ is either a no-op or a $\beta_{\|}$ reduction, with assignment in the store-machine being equivalent to program-wide substitution in $\lambda_{\|}$.
- Similarly, the ξ function maps states of the CK transition system to the $\lambda_{\|}$ space and for every CK transition, the resulting $\lambda_{\|}$ operation is also either a no-op or a $\beta_{\|}$ reduction, with the transition that descends under a λ-abstraction being equivalent to substitution in $\lambda_{\|}$.
- Finally, the ϕ function maps states of the CK transition system to λ_{need} terms and is used to show that the CK system and λ_{need} are equivalent. For every CK transition, the equivalent λ_{need} operation is either a no-op or a β_{need} reduction.

Syntax

$$
\begin{array}{rr}
S_L = \langle e, F_L s, \Gamma \rangle & \text{(States)} \\
F_L s = F_L, \ldots & \text{(List of Frames)} \\
F_L = (\mathtt{arg}\, e) \mid (\mathtt{var}\ x) & \text{(Frames)} \\
\Gamma = (x \mapsto e, \ldots) & \text{(Heaps)}
\end{array}
$$

Transitions

$$
\begin{array}{rl}
\langle e_1\, e_2, F_L s, \Gamma \rangle \overset{ckh}{\longmapsto} \langle e_1, ((\mathtt{arg}\, e_2), F_L s), \Gamma \rangle & (\textit{push-arg-ckh}) \\
\langle \lambda x.e_1, ((\mathtt{arg}\, e_2), F_L s), \Gamma \rangle \overset{ckh}{\longmapsto} \langle e_1\{x := y\}, F_L s, (\Gamma, y \mapsto e_2) \rangle\,,\ y \text{ fresh} & (\textit{descend-lam-ckh}) \\
\langle x, F_L s, (\Gamma, x \mapsto e) \rangle \overset{ckh}{\longmapsto} \langle e, ((\mathtt{var}\ x), F_L s), \Gamma \rangle & (\textit{lookup-var-ckh}) \\
\langle v, ((\mathtt{var}\ x), F_L s), \Gamma \rangle \overset{ckh}{\longmapsto} \langle v, F_L s, (\Gamma, x \mapsto v) \rangle & (\textit{update-heap-ckh})
\end{array}
$$

Fig. 5. The natural semantics as an abstract machine.

Subsections 5.2 and 5.3 present the store-based machine and the parallel rewriting semantics, respectively, including a proof of equivalence. Subsection 5.4 presents the CK system and subsection 5.5 explains the rest of the proof.

5.2 Adapting Launchbury's Natural Semantics

Figure 5 describes the syntax and transitions of the store machine.[5] It is dubbed CKH because it resembles a three-register machine [9]: a machine state S_L is comprised of a control string (C), a list of frames (K) that represents the control context in an inside-out manner, and a heap (H). The ... notation means "zero of more of the preceding kind of element." An ($\texttt{arg}\, e$) frame represents the argument in an application and the ($\texttt{var}\ x$) frame indicates that a heap expression is the current control string. Parentheses are used to group a list of frames when necessary. The initial machine state for a program e is $\langle e, (), () \rangle$. Computation terminates when the control string is a value and the list of frames is empty.

The *push-arg-ckh* transition moves the argument in an application to a new `arg` frame in the frame list and makes the operator the next control string. When that operator is a λ-abstraction, the *descend-lam-ckh* transition adds its argument to the heap,[6] mapped to a fresh variable name, and makes the body of the operator the new control string. The *lookup-var-ckh* transition evaluates an argument from the heap when the control string is a variable. The mapping is removed from the heap and a new ($\texttt{var}\ x$) frame remembers the variable whose corresponding expression is under evaluation. Finally, when the heap expression is reduced to a value, the *update-heap-ckh* transition extends the heap again.

5.3 Parallel Rewriting

The syntax of the parallel λ-rewriting semantics is as follows:

$$
\begin{array}{rl}
e_{\|} = e \mid {e_{\|}}^{x} & \text{(Terms)} \\
v_{\|} = v \mid {v_{\|}}^{x} & \text{(Values)} \\
E_{\|} = [\ ] \mid E_{\|}\, e_{\|} \mid {E_{\|}}^{x} & \text{(Evaluation Contexts)}
\end{array}
$$

This system expresses computation with a selective parallel reduction strategy. When a function application is in demand, the system substitutes the argument for all free occurrences of the bound variable, regardless of the status of the argument. When an instance of a substituted argument is reduced, however, all instances of the argument are reduced in parallel. Here is a sample reduction:

$$(\lambda x.x\, x)\,(I\ I) \overset{\|}{\longmapsto} (I\ I)^x\,(I\ I)^x \overset{\|}{\longmapsto} I^x\ I^x \overset{\|}{\longmapsto} I$$

[5] To aid comparisons, we slightly alter Launchbury's rules (and the resulting machine) to use pure λ terms. Thus we avoid Launchbury's preprocessing and special syntax.

[6] The notation $(\Gamma, x{\mapsto}e)$ is a heap Γ, extended with the variable-term mapping $x{\mapsto}e$.

The $\lambda_{\|}$ semantics keeps track of arguments via labeled terms ${e_{\|}}^{x}$, where labels are variables. Values in $\lambda_{\|}$ also include labeled λ-abstractions. Reducing a labeled term triggers the simultaneous reduction of all other terms with the same label. Otherwise, labels do not affect program evaluation.

We require that all expressions with the same label must be identical.

Definition 5. *A program $e_{\|}$ is* consistently labeled (CL) *when for any two subterms $e_{\|1}^{x_1}$ and $e_{\|2}^{x_2}$ of $e_{\|}$, $x_1 = x_2$ implies $e_{\|1} = e_{\|2}$.*

In the reduction of $\lambda_{\|}$ programs, evaluation contexts $E_{\|}$ determine which part of the program to reduce next. The $\lambda_{\|}$ evaluation contexts are the call-by-name evaluation contexts with the addition of the labeled ${E_{\|}}^{x}$ context, which dictates that a redex search goes under labeled terms. Essentially, when searching for a redex, terms tagged with a label are treated as if they were unlabeled.

The parallel semantics can exploit simpler evaluation contexts than λ_{need} because substitution occurs as soon as an application is encountered:

$$E_{\|}[((\lambda x.e_{\|1})^{\vec{y}})\, e_{\|2}] \overset{\|}{\longmapsto} \begin{cases} E_{\|}[e_{\|}], & \text{if } [\ ] \text{ is not under a label in } E_{\|} \\ E_{\|}[e_{\|}]\{\!\{z \Leftarrow E_{\|2}[e_{\|}]\}\!\}, & \text{if } E_{\|}[\ ] = E_{\|1}[(E_{\|2}[\ ])^{z}] \\ & \text{and } [\ ] \text{ is not under a label in } E_{\|2} \end{cases} \qquad (\beta_{\|})$$
$$\text{where } e_{\|} = e_{\|1}\{x := {e_{\|2}}^{w}\}, w \text{ fresh}$$

On the left-hand side of $\beta_{\|}$, the program is partitioned into a context and a β-like redex. A term $e^{\vec{y}}$ may have any number of labels and possibly none. On the right-hand side, the redex is contracted to a term $e_{\|1}\{x := {e_{\|2}}^{w}\}$ such that the argument is tagged with an unique label w. Obsolete labels $\vec{y}$ are discarded.

There are two distinct ways to contract a redex: when the redex is not under any labels and when the redex occurs under at least one label. For the former, the redex is the only contracted part of the program. For the latter, all other instances of that labeled term are similarly contracted. In this second case, the evaluation context is further subdivided as $E_{\|}[\ ] = E_{\|1}[(E_{\|2}[\ ])^{z}]$, where z is the label nearest the redex, i.e., $E_{\|2}$ contains no additional labels. A whole-program substitution function is used to perform the parallel reduction:

$$\begin{aligned} {e_{\|1}}^{x}\{\!\{x \Leftarrow e_{\|}\}\!\} &= {e_{\|}}^{x} \\ {e_{\|1}}^{x}\{\!\{y \Leftarrow e_{\|}\}\!\} &= (e_{\|1}\{\!\{y \Leftarrow e_{\|}\}\!\})^{x},\ x \neq y \\ (\lambda x.e_{\|1})\{\!\{x \Leftarrow e_{\|}\}\!\} &= \lambda x.(e_{\|1}\{\!\{x \Leftarrow e_{\|}\}\!\}) \\ (e_{\|1}\, e_{\|2})\{\!\{x \Leftarrow e_{\|}\}\!\} &= (e_{\|1}\{\!\{x \Leftarrow e_{\|}\}\!\}\, e_{\|2}\{\!\{x \Leftarrow e_{\|}\}\!\}) \\ \text{otherwise, } e_{\|1}\{\!\{x \Leftarrow e_{\|}\}\!\} &= e_{\|1} \end{aligned}$$

Rewriting terms with $\beta_{\|}$ preserves the consistent labeling property.

Proposition 2. *If $e_{\|}$ is CL and $e_{\|} \overset{\|}{\longmapsto} e_{\|}'$, then $e_{\|}'$ is CL.*

The ψ function reconstructs a $\lambda_{\|}$ term from a CKH machine configuration:

$$\begin{aligned}
\psi(\langle e, ((\texttt{var}\ x), F_L s), \Gamma\rangle) &= \psi(\langle x, F_L s, (x \mapsto e, \Gamma)\rangle) \quad \boxed{\psi : S_L \to e_{\|}} \\
\psi(\langle e_1, ((\texttt{arg}\, e_2), F_L s), \Gamma\rangle) &= \psi(\langle e_1\, e_2, F_L s, \Gamma\rangle) \\
\psi(\langle e, (), \Gamma\rangle) &= e\{\!\{\Gamma\}\!\}
\end{aligned}$$

The operation $e\{\!\{\Gamma\}\!\}$, using overloaded notation, replaces all free variables in e with their corresponding terms in Γ and tags them with appropriate labels.

Lemma 5 demonstrates the bulk of the equivalence of the store machine and $\lambda_{\|}$.[7] The rest of the equivalence proof is straightforward [9].

Lemma 5. *If* $\langle e, F_L s, \Gamma\rangle \overset{ckh}{\longmapsto} \langle e', F_L s', \Gamma'\rangle$, *then either:*

1. $\psi(\langle e, F_L s, \Gamma\rangle) = \psi(\langle e', F_L s', \Gamma'\rangle)$
2. $\psi(\langle e, F_L s, \Gamma\rangle) \overset{\|}{\longmapsto} \psi(\langle e', F_L s', \Gamma'\rangle)$

5.4 A Transition System for Comparing λ_{need} and $\lambda_{\|}$

The CK layer in figure 4 mediates between $\lambda_{\|}$ and λ_{need}. The corresponding transition system resembles a two-register CK machine [9]. Figure 6 describes the syntax and the transitions of the system.[8]

Syntax

$$\begin{aligned}
S &= \langle e, Fs\rangle && \text{(States)} \\
Fs &= F, \ldots && \text{(List of Frames)} \\
F &= (\texttt{arg}\, e) \mid (\texttt{lam}\, x) \mid (\texttt{bod}\, x\, Fs\, Fs) && \text{(Frames)}
\end{aligned}$$

Transitions

$$\begin{aligned}
\langle e_1\, e_2, Fs\rangle &\overset{ck}{\longmapsto} \langle e_1, ((\texttt{arg}\, e_2), Fs)\rangle && (\textit{push-arg-ck}) \\
\langle \lambda x.e, Fs\rangle &\overset{ck}{\longmapsto} \langle e, ((\texttt{lam}\, x), Fs)\rangle && (\textit{descend-lam-ck}) \\
&\qquad \text{if } \texttt{balance}(Fs) > 0 \\
\langle x, (Fs_1, (\texttt{lam}\, x), Fs_2, (\texttt{arg}\, e), Fs)\rangle &\overset{ck}{\longmapsto} \langle e, ((\texttt{bod}\, x\, Fs_1\, Fs_2), Fs)\rangle && (\textit{lookup-var-ck}) \\
&\qquad \text{if } \phi_F(Fs_1) \in \check{A}[E], \phi_F(Fs_2) \in A, \phi_F(Fs) \in E[\hat{A}], \hat{A}[\check{A}] \in A \\
\langle v, (Fs_3, (\texttt{bod}\, x\, Fs_1\, Fs_2), Fs)\rangle &\overset{ck}{\longmapsto} \langle v, (Fs_1\{x := v\}, Fs_3, Fs_2, Fs)\rangle && (\beta_{need\text{-}ck}) \\
&\qquad \text{if } \phi_F(Fs_3) \in A
\end{aligned}$$

Fig. 6. A transition system for comparing λ_{need} and $\lambda_{\|}$.

States consist of a subterm and a list of frames representing the context. The first kind of frame represents the argument in an application and the second

[7] The lemma relies on an extension of the typical α-equivalence classes of terms to include variables in labels as well.

[8] The CK transition system is a proof-technical device. Unlike the original CK machine, ours is ill-suited for an implementation.

frame represents a λ-abstraction with a hole in the body. The last kind of frame has two frame list components, the first representing a context in the body of the λ, and the second representing the context between the λ and its argument. The variable in this last frame is the variable bound by the λ expression under evaluation. The initial state for a program e is $\langle e, () \rangle$, where $()$ is an empty list of frames, and evaluation terminates when the control string is a value and the list of frames is equivalent to an answer context.

The *push-arg-ck* transition makes the operator in an application the new control string and adds a new `arg` frame to the frame list containing the argument. The *descend-lam-ck* transition goes under a λ, making the body the control string, but only if that λ has a corresponding argument in the frame list, as determined by the `balance` function, defined as follows:

$$\boxed{\texttt{balance} : Fs \to \mathbb{Z}}$$

$$\texttt{balance}(Fs_3, (\texttt{bod}\, x\, Fs_1\, Fs_2), Fs) = \texttt{balance}(Fs_3)$$

$$\texttt{balance}(Fs) = \texttt{\#arg-frames}(Fs) - \texttt{\#lam-frames}(Fs)$$

$$Fs \text{ contains no } \texttt{bod} \text{ frames}$$

The `balance` side condition for *descend-lam-ck* dictates that evaluation goes under a λ only if there is a matching argument for it, thus complying with the analogous evaluation context. The `balance` function employs `#arg-frames` and `#lam-frames` to count the number of `arg` or `lam` frames, respectively, in a list of frames. Their definitions are elementary and therefore omitted.

The *lookup-var-ck* transition is invoked if the control string is a variable, somewhere in a λ body, and the rest of the frames have a certain shape consistent with the corresponding parts of a β_{need} redex. With this transition, the argument associated with the variable becomes the next control string and the context around the variable in the λ body and the context between the λ and argument are saved in a new `bod` frame. Finally, when an argument is an answer, indicated by a value control string and a `bod` frame in the frame list—with the equivalent of an answer context in between—the value gets substituted into the body of the λ according to the $\beta_{need\text{-}ck}$ transition. The $\beta_{need\text{-}ck}$ transition uses a substitution function on frame lists, $Fs\{x := e\}$, which overloads the notation for regular term substitution and has the expected definition.

Figure 7 defines metafunctions for the CK transition system. The ϕ function converts a CK state to the equivalent λ_{need} term, and uses ϕ_F to convert a list of frames to an evaluation context.

Now we can show that an evaluator defined with $\overset{ck}{\longmapsto}$ is equivalent to $\texttt{eval}^{\text{sr}}_{need}$. The essence of the proof is a lemma that relates the shape of CK transition sequences to the shape of λ_{need} standard reduction sequences. The rest of the equivalence proof is straightforward [9].

Lemma 6. *If* $\langle e, Fs \rangle \overset{ck}{\longmapsto} \langle e', Fs' \rangle$*, then either:*

1. $\phi(\langle e, Fs \rangle) = \phi(\langle e', Fs' \rangle)$
2. $\phi(\langle e, Fs \rangle) \longmapsto \phi(\langle e', Fs' \rangle)$

$$\boxed{\phi : S \to e}$$

$$\phi(\langle e, Fs\rangle) = \phi_F(Fs)[e]$$

$$\boxed{\phi_F : Fs \to E}$$

$$\begin{aligned}
\phi_F((\,)) &= [\ ] \\
\phi_F((\mathtt{lam}\, x), Fs) &= \phi_F(Fs)[\lambda x.[\ ]] \\
\phi_F((\mathtt{arg}\, e), Fs) &= \phi_F(Fs)[[\ ]\, e] \\
\phi_F((\mathtt{bod}\, x\, Fs_1\, Fs_2), Fs) &= \\
&\phi_F(Fs)[\phi_F(Fs_2)[\lambda x.\phi_F(Fs_1)[x]]\,[\ ]]
\end{aligned}$$

$$\boxed{\xi : S \to e_\parallel}$$

$$\xi(\langle e, Fs\rangle) = \xi_F(Fs, e)$$

$$\boxed{\xi_F : Fs \times e_\parallel \to e_\parallel}$$

$$\begin{aligned}
\xi_F((\,), e_\parallel) &= e_\parallel \\
\xi_F(((\mathtt{arg}\, e_{\parallel 1}), Fs), e_\parallel) &= \xi_F(Fs, e_\parallel\, e_{\parallel 1}) \\
\xi_F(((\mathtt{bod}\, x\, Fs_1\, Fs_2), Fs), e_\parallel) &= \\
&\xi_F((Fs_1, (\mathtt{lam}\, x), Fs_2, (\mathtt{arg}\, e_\parallel), Fs), x) \\
\xi_F(((\mathtt{lam}\, x), Fs_1, (\mathtt{arg}\, e_{\parallel 1}), Fs_2), e_\parallel) &= \\
&\xi_F((Fs_1, Fs_2), e_\parallel\{x := e_{\parallel 1}^{\,y}\}) \\
&\phi_F(Fs_1) \in A,\ y \text{ fresh}
\end{aligned}$$

Fig. 7. Functions to map CK states to λ_{need} (ϕ) and $\lambda_\parallel$ (ξ).

Finally, we show how the CK system corresponds to $\lambda_\parallel$. The ξ function defined in figure 7 constructs a $\lambda_\parallel$ term from a CK configuration.

Lemma 7. *If* $\langle e, Fs\rangle \overset{ck}{\longmapsto} \langle e', Fs'\rangle$*, then either:*

1. $\xi(\langle e, Fs\rangle) = \xi(\langle e', Fs'\rangle)$
2. $\xi(\langle e, Fs\rangle) \overset{\parallel}{\longmapsto} \xi(\langle e', Fs'\rangle)$

5.5 Relating all Layers

In the previous subsections, we have demonstrated the correspondence between $\lambda_\parallel$, the natural semantics, and the λ_{need} standard reduction sequences via lemmas 5 through 7. We conclude this section with the statement of an extensional correctness theorem, where $\mathtt{eval}_{natural}$ is an evaluator defined with the store machine transitions. The theorem follows from the composition of the equivalences of our specified rewriting systems.

Theorem 4. $\mathit{eval}_{need} = \mathit{eval}_{natural}$

6 Extensions and Variants

Data Constructors Real-world lazy languages come with data structure construction and extraction operators. Like function arguments, the arguments to a data constructor should not be evaluated until there is demand for their values [12, 14]. The standard λ calculus encoding of such operators [5] works well:

$$\mathtt{cons} = \lambda x.\lambda y.\lambda s.s\, x\, y, \quad \mathtt{car} = \lambda p.p\, \lambda x.\lambda y.x, \quad \mathtt{cdr} = \lambda p.p\, \lambda x.\lambda y.y$$

Adding true algebraic systems should also be straightforward.

Recursion Our λ_{need} calculus represents just a core λ calculus and does not include an explicit `letrec` constructor for cyclic terms. Since cyclic programming is an important idiom in lazy programming languages, others have extensively explored cyclic by-need calculi, e.g., Ariola and Blum [1], and applying their solutions to our calculus should pose no problems.

7 Conclusion

Following Plotkin's work on call-by-name and call-by-value, we present a call-by-need λ calculus that expresses computation via a single axiom in the spirit of β. Our calculus is close to implementations of lazy languages because it captures the idea of by-need computation without retaining every function call and without need for re-associating terms. We show that our calculus satisfies Plotkin's criteria, including an intensional correspondence between our calculus and a Launchbury-style natural semantics. Our future work will leverage our λ_{need} calculus to derive a new abstract machine for lazy languages.

Acknowledgments We thank J. Ian Johnson, Casey Klein, Vincent St-Amour, Asumu Takikawa, Aaron Turon, Mitchell Wand, and the ESOP 2012 reviewers for their feedback on early drafts. This work was supported in part by NSF Infrastructure grant CNS-0855140 and AFOSR grant FA9550-09-1-0110.

References

1. Ariola, Z., Blom, S.: Cyclic lambda calculi. In: Proc. Theoretical Aspects of Computer Software 1994. pp. 77–106 (1997)
2. Ariola, Z.M., Felleisen, M.: The call-by-need lambda-calculus. Tech. Rep. CIS-TR-94-23, University of Oregon (1994)
3. Ariola, Z.M., Felleisen, M.: The call-by-need lambda calculus. J. Funct. Program. 7, 265–301 (1997)
4. Ariola, Z.M., Maraist, J., Odersky, M., Felleisen, M., Wadler, P.: A call-by-need lambda calculus. In: Proc. 22nd Symp. on Principles of Programming Languages. pp. 233–246 (1995)
5. Barendregt, H.P.: Lambda Calculus, Syntax and Semantics. North-Holland (1985)
6. Church, A.: The Calculi of Lambda Conversion. Princeton University Press (1941)
7. Curry, H.B., Feys, R.: Combinatory Logic, vol. I. North-Holland (1958)
8. Danvy, O., Millikin, K., Munk, J., Zerny, I.: Defunctionalized interpreters for call-by-need evaluation. In: Proc. 10th Symp. on Functional and Logic Programming. pp. 240–256 (2010)
9. Felleisen, M., Findler, R.B., Flatt, M.: Semantics Engineering with PLT Redex. MIT Press (2009)
10. Felleisen, M., Friedman, D.P.: A syntactic theory of sequential state. Theor. Comput. Sci. 69(3), 243–287 (1989)
11. Friedman, D.P., Ghuloum, A., Siek, J.G., Winebarger, O.L.: Improving the lazy Krivine machine. Higher Order Symbolic Computation 20, 271–293 (2007)
12. Friedman, D.P., Wise, D.S.: Cons should not evaluate its arguments. In: Proc. 3rd Intl. Colloq. on Automata, Languages and Programming. pp. 256–284 (1976)

13. Garcia, R., Lumsdaine, A., Sabry, A.: Lazy evaluation and delimited control. In: Proc. 36th Symp. on Principles of Programming Languages. pp. 153–164 (2009)
14. Henderson, P., Morris, Jr., J.H.: A lazy evaluator. In: Proc. 3rd Symp. on Principles of Programming Languages. pp. 95–103 (1976)
15. Josephs, M.B.: The semantics of lazy functional languages. Theor. Comput. Sci. 68(1) (1989)
16. Landin, P.J.: The next 700 programming languages. Comm. ACM 9, 157–66 (1966)
17. Launchbury, J.: A natural semantics for lazy evaluation. In: Proc. 20th Symp. on Principles of Programming Languages. pp. 144–154 (1993)
18. Loader, R.: Notes on simply typed lambda calculus. Tech. Rep. ECS-LFCS-98-381, Department of Computer Science, University of Edinburgh (1998)
19. Maraist, J., Odersky, M., Turner, D.N., Wadler, P.: Call-by-name call-by-value, call-by-need, and the linear lambda calculus. In: Proc. 11th Conference on Mathematical Foundations of Programminng Semantics. pp. 370–392 (1995)
20. Maraist, J., Odersky, M., Wadler, P.: The call-by-need lambda calculus (unabridged). Tech. Rep. 28/94, Universität Karlsruhe (1994)
21. Maraist, J., Odersky, M., Wadler, P.: The call-by-need lambda calculus. J. Funct. Program. 8, 275–317 (1998)
22. Morris, J.H.: Lambda Calculus Models of Programming Languages. Ph.D. thesis, MIT (1968)
23. Nakata, K., Hasegawa, M.: Small-step and big-step semantics for call-by-need. J. Funct. Program. 19(6), 699–722 (2009)
24. Peyton Jones, S.L., Salkild, J.: The spineless tagless g-machine. In: Proc. 4th Conf. on Functional Programming Lang. and Computer Architecture. pp. 184–201 (1989)
25. Plotkin, G.D.: Call-by-name, call-by-value and the lambda-calculus. Theor. Comput. Sci. 1, 125–159 (1975)
26. Purushothaman, S., Seaman, J.: An adequate operational semantics for sharing in lazy evaluation. In: Proc. 4th European Symp. on Program. pp. 435–450 (1992)
27. Sestoft, P.: Deriving a lazy abstract machine. J. Func. Program. 7(3), 231–64 (1997)
28. Wadsworth, C.P.: Semantics and Pragmatics of the Lambda Calculus. Ph.D. thesis, Oxford University (1971)