Counting Immutable Beans

Reference Counting Optimized for Purely Functional Programming

Sebastian Ullrich
Karlsruhe Institute of Technology
Germany
sebastian.ullrich@kit.edu

Leonardo de Moura
Microsoft Research
USA
leonardo@microsoft.com

ABSTRACT

Most functional languages rely on some kind of garbage collection for automatic memory management. They usually eschew reference counting in favor of a tracing garbage collector, which has less bookkeeping overhead at runtime. On the other hand, having an exact reference count of each value can enable optimizations such as destructive updates. We explore these optimization opportunities in the context of an eager, purely functional programming language. We propose a new mechanism for efficiently reclaiming memory used by nonshared values, reducing stress on the global memory allocator. We describe an approach for minimizing the number of reference counts updates using borrowed references and a heuristic for automatically inferring borrow annotations. We implemented all these techniques in a new compiler for an eager and purely functional programming language with support for multi-threading. Our preliminary experimental results demonstrate our approach is competitive and often outperforms state-of-the-art compilers.

CCS CONCEPTS

• **Software and its engineering** → **Runtime environments**; **Garbage collection**.

KEYWORDS

purely functional programming, reference counting, Lean

ACM Reference Format:
Sebastian Ullrich and Leonardo de Moura. 2020. Counting Immutable Beans: Reference Counting Optimized for Purely Functional Programming. In *Proceedings of International Symposium on Implementation and Application of Functional Languages (IFL'19).* ACM, New York, NY, USA, 12 pages.

1 INTRODUCTION

Although reference counting [Collins 1960] (RC) is one of the oldest memory management techniques in computer science, it is not considered a serious garbage collection technique in the functional programming community, and there is plenty of evidence it is in general inferior to tracing garbage collection algorithms. Indeed, high-performance compilers such as ocamlopt and GHC use tracing garbage collectors. Nonetheless, implementations of several popular programming languages, e.g., Swift, Objective-C, Python, and Perl, use reference counting as a memory management technique. Reference counting is often praised for its simplicity, but many disadvantages are frequently reported in the literature [Jones and Lins 1996; Wilson 1992]. First, incrementing and decrementing reference counts every time a reference is created or destroyed can significantly impact performance because they not only take time but also affect cache performance, especially in a multi-threaded program [Choi et al. 2018]. Second, reference counting cannot collect circular [McBeth 1963] or self-referential structures. Finally, in most reference counting implementations, pause times are deterministic but may still be unbounded [Boehm 2004].

In this paper, we investigate whether reference counting is a competitive memory management technique for purely functional languages, and explore optimizations for reusing memory, performing destructive updates, and for minimizing the number of reference count increments and decrements. The former optimizations in particular are beneficial for purely functional languages that otherwise can only perform functional updates. When performing functional updates, objects often die just before the creation of an object of the same kind. We observe a similar phenomenon when we insert a new element into a pure functional data structure such as binary trees, when we use *map* to apply a given function to the elements of a list or tree, when a compiler applies optimizations by transforming abstract syntax trees, or when a proof assistant rewrites formulas. We call it the *resurrection hypothesis*: many objects die just before the creation of an object of the same kind. Our new optimization takes advantage of this hypothesis, and enables pure code to perform destructive updates in all scenarios described above when objects are not shared. We implemented all the ideas reported here in the new runtime and compiler for the Lean programming language [de Moura et al. 2015]. We also report preliminary experimental results that demonstrate our new compiler produces competitive code that often outperforms the code generated by high-performance compilers such as ocamlopt and GHC (Section 8).

Lean implements a version of the Calculus of Inductive Constructions [Coquand and Huet 1988; Coquand and Paulin 1990], and it has mainly been used as a proof assistant so far. Lean has a metaprogramming framework for writing proof and code automation, where users can extend Lean using Lean itself [Ebner et al. 2017]. Improving the performance of Lean metaprograms was the primary motivation for the work reported here, but one can apply the techniques reported here to general-purpose functional programming languages.

IFL '19, September 25–27, 2019, Singapore, Singapore

ACM ISBN 978-1-4503-7562-7/19/09…$15.00
https://doi.org/10.1145/3412932.3412935

We describe our approach as a series of refinements starting from λ_{pure}, a simple intermediate representation for eager and purely functional languages (Section 3). We remark that in Lean and λ_{pure}, it is not possible to create cyclic data structures. Thus, one of the main criticisms against reference counting does not apply. From λ_{pure}, we obtain λ_{RC} by adding explicit instructions for incrementing (**inc**) and decrementing (**dec**) reference counts, and reusing memory (Section 4). The inspiration for explicit RC instructions comes from the Swift compiler, as does the notion of *borrowed* references. In contrast to standard (or *owned*) references, of which there should be exactly as many as the object's reference counter implies, borrowed references do not update the reference counter but are *assumed* to be kept alive by a surrounding owned reference, further minimizing the number of **inc** and **dec** instructions in generated code.

We present a simple compiler from λ_{pure} to λ_{RC}, discussing heuristics for inserting destructive updates, borrow annotations, and **inc**/**dec** instructions (Section 5). Finally, we show that our approach is compatible with existing techniques for performing destructive updates on array and string values, and propose a simple and efficient approach for thread-safe reference counting (Section 7).

Contributions. We present a reference counting scheme optimized for and used by the next version of the Lean programming language.

- We describe how to reuse allocations in both user code and language primitives, and give a formal reference-counting semantics that can express this reuse.
- We describe the optimization of using borrowed references.
- We define a compiler that implements all these steps. The compiler is implemented in Lean itself and the source code is available.
- We give a simple but effective scheme for avoiding atomic reference count updates in multi-threaded programs.
- We compare the new Lean compiler incorporating these ideas with other compilers for functional languages and show its competitiveness.

2 EXAMPLES

In reference counting, each heap-allocated value contains a reference count. We view this counter as a collection of tokens. The **inc** instruction creates a new token and **dec** consumes it. When a function takes an argument as an owned reference, it is responsible for consuming one of its tokens. The function may consume the owned reference not only by using the **dec** instruction, but also by storing it in a newly allocated heap value, returning it, or passing it to another function that takes an owned reference. We illustrate our intermediate representation (IR) and the use of owned and borrowed references with a series of small examples.

The identity function *id* does not require any RC operation when it takes its argument as an owned reference.

id x = **ret** *x*

As another example, consider the function *mkPairOf* that takes x and returns the pair (x, x).

mkPairOf x = **inc** *x*; **let** *p* = *Pair x x*; **ret** *p*

It requires an **inc** instruction because two tokens for x are consumed (we will also say that "x is consumed" twice). The function *fst* takes two arguments x and y, and returns x, and uses a **dec** instruction for consuming the unused y.

fst x y = **dec** *y*; **ret** *x*

The examples above suggest that we do not need any RC operation when we take arguments as owned references and consume them exactly once. Now we contrast that with a function that only inspects its argument: the function *isNil xs* returns true if the list *xs* is empty and false otherwise. If the argument *xs* is taken as an owned reference, our compiler generates the following code

```
isNil xs = case xs of
  (Nil → dec xs; ret true)
  (Cons → dec xs; ret false)
```

We need the **dec** instructions because a function must consume all arguments taken as owned references. One may notice that decrementing xs immediately after we inspect its constructor tag is wasteful. Now assume that instead of taking the ownership of an RC token, we could borrow it from the caller. Then, the callee would not need to consume the token using an explicit **dec** operation. Moreover, the caller would be responsible for keeping the borrowed value alive. This is the essence of *borrowed references*: a borrowed reference does not actually keep the referenced value alive, but instead asserts that the value is kept alive by another, owned reference. Thus, when *xs* is a borrowed reference, we compile *isNil* into our IR as

isNil xs = **case** *xs* **of** (*Nil* → **ret** *true*) (*Cons* → **ret** *false*)

As a less trivial example, we now consider the function *hasNone xs* that, given a list of optional values, returns *true* if *xs* contains a *None* value. This function is often defined in a functional language as

```
hasNone [] = false
hasNone (None  : xs) = true
hasNone (Some x : xs) = hasNone xs
```

Similarly to *isNil*, *hasNone* only inspects its argument. Thus if *xs* is taken as a borrowed reference, our compiler produces the following RC-free IR code for it

```
hasNone xs = case xs of
  (Nil → ret false)
  (Cons → let h = proj_head xs; case h of
    (None → ret true)
    (Some → let t = proj_tail xs; let r = hasNone t; ret r))
```

Note that our **case** operation does not introduce binders. Instead, we use explicit instructions $\mathbf{proj}_i$ for accessing the head and tail of the *Cons* cell. We use suggestive names for cases and fields in these initial examples, but will later use indices instead. Our borrowed inference heuristic discussed in Section 5 correctly tags *xs* as a borrowed parameter.

When using owned references, we know at run time whether a value is shared or not simply by checking its reference counter. We observed we could leverage this information and minimize the amount of allocated and freed memory for constructor values such

as a list *Cons* value. Thus, we have added two additional instructions to our IR: **let** y = **reset** x and **let** z = (**reuse** y **in ctor**$_i$ $\overline{w}$). The two instructions are used together; if x is a shared value, then y is set to a special reference ■, and the **reuse** instruction just allocates a new constructor value **ctor**$_i$ $\overline{w}$. If x is not shared, then **reset** decrements the reference counters of the components of x, and y is set to x. Then, **reuse** reuses the memory cell used by x to store the constructor value **ctor**$_i$ $\overline{w}$. We illustrate these two instructions with the IR code for the list *map* function generated by our compiler as shown in Section 5. The code uses our actual, positional encoding of cases, constructors, and fields as described in the next section.

```
map f xs = case xs of
  (ret xs)
  (let x = proj_1 xs; inc x; let s = proj_2 xs; inc s;
   let w = reset xs;
   let y = f x; let ys = map f s;
   let r = (reuse w in ctor_2 y ys); ret r)
```

We remark that if the list referenced by *xs* is not shared, the code above does not allocate any memory. Moreover, if *xs* is a nonshared list of list of integers, then *map* (*map inc*) *xs* will not allocate any memory either. This example also demonstrates it is not a good idea, in general, to fuse **reset** and **reuse** into a single instruction: if we removed the **let** w = **reset** xs instruction and directly used xs in **reuse**, then when we execute the recursive application *map f s*, the reference counter for s would be greater than 1 even if the reference counter for xs was 1. We would have a reference from xs and another from s, and memory reuse would not occur in the recursive applications. Note that removing the **inc** s instruction is incorrect when xs is a shared value. Although the **reset** and **reuse** instructions can in general be used for reusing memory between two otherwise unrelated values, in examples like *map* where the reused value has a close semantic connection to the reusing value, we will use common functional vocabulary and say that the list is being *destructively updated* (up to the first shared cell).

As another example, a zipper is a technique for traversing and efficiently updating data structures, and it is particularly useful for purely functional languages. For example, the list zipper is a pair of lists, and it allows one to move forward and backward, and to update the current position. The *goForward* function is often defined as

```
goForward ([], bs) = ([], bs)
goForward (x : xs, bs) = (xs, x : bs)
```

In most functional programming languages, the second equation allocates a new pair and *Cons* value. The functions *map* and *goForward* both satisfy our resurrection hypothesis. Moreover, the result of a *goForward* application is often fed into another *goForward* or *goBackward* application. Even if the initial value was shared, every subsequent application takes a nonshared pair, and memory allocations are avoided by the code produced by our compiler.

```
goForward p = case p of
  (let xs = proj_1 p; inc xs;
   case xs of
     (ret p)
     (let bs = proj_2 p; inc bs;
      let c_1 = reset p;
      let x = proj_1 xs; inc x; xs' = proj_2 xs; inc xs';
      let c_2 = reset xs;
      let bs' = (reuse c_2 in ctor_2 x bs);
      let r = (reuse c_1 in ctor_1 xs' bs'); ret r))
```

3 THE PURE IR

Our source language λ_{pure} is a simple untyped functional intermediate representation (IR) in the style of A-normal form [Flanagan et al. 1993]. It captures the relevant features of the actual IR we have implemented and avoids unnecessary complexity that would only distract the reader from the ideas proposed here.

$$
\begin{array}{lll}
w, x, y, z \in \mathit{Var} & & \\
\quad c \in \mathit{Const} & & \\
\quad e \in \mathit{Expr} & ::= c\ \overline{y} \mid \textbf{pap}\ c\ \overline{y} \mid x\ y \mid \textbf{ctor}_i\ \overline{y} \mid \textbf{proj}_i\ x & \\
\quad F \in \mathit{FnBody} & ::= \textbf{ret}\ x \mid \textbf{let}\ x = e;\ F \mid \textbf{case}\ x\ \textbf{of}\ \overline{F} & \\
\quad f \in \mathit{Fn} & ::= \lambda\ \overline{y}.\ F & \\
\quad \delta \in \mathit{Program} & = \mathit{Const} \rightharpoonup \mathit{Fn} &
\end{array}
$$

All arguments of function applications are variables. The applied function is a constant c, with partial applications marked with the keyword **pap**, a variable x, the i-th constructor of an erased datatype, or the special function **proj**$_i$, which returns the i-th argument of a constructor application. Function bodies always end with evaluating and returning a variable. They can be chained with (non-recursive) **let** statements and branch using **case** statements, which evaluate to their i-th arm given an application of **ctor**$_i$. As further detailed in Section 5.3, we consider tail calls to be of the form **let** $r = c\ \overline{x}$; **ret** r. A program is a partial map from constant names to their implementations. The body of a constant's implementation may refer back to the constant, which we use to represent recursion, and analogously mutual recursion. In examples, we use $f\ \overline{x} = F$ as syntax sugar for $\delta(f) = \lambda\ \overline{x}.\ F$.

As an intermediate representation, we can and should impose restrictions on the structure of λ_{pure} to simplify working with it. We assume that

- all constructor applications are fully applied by eta-expanding them.
- no constant applications are over-applied by splitting them into two applications where necessary.
- all variable applications take only one argument, again by splitting them where necessary. While this simplification can introduce additional allocations of intermediary partial applications, it greatly simplifies the presentation of our operational semantics. All presented program transformations

can be readily extended to a system with n-ary variable applications, which are handled analogously to n-ary constant applications.
- every function abstraction has been lambda-lifted to a top-level constant c.
- trivial bindings **let** $x = y$ have been eliminated through copy propagation.
- all dead **let** bindings have been removed.
- all parameter and **let** names of a function are mutually distinct. Thus we do not have to worry about name capture.

In the actual IR we have implemented[1], we also have instructions for storing and accessing unboxed data in constructor values, boxing and unboxing machine integers and scalar values, and creating literals of primitive types such as strings and numbers. Our IR also supports *join points* similar to the ones used in the Haskell Core language [Maurer et al. 2017]. Join points are local function declarations that are never partially applied (i.e., they never occur in **pap** instructions), and are always tail-called. The actual IR has support for defining join points, and a **jmp** instruction for invoking them.

4 SYNTAX AND SEMANTICS OF THE REFERENCE-COUNTED IR

The target language λ_{RC} is an extension of λ_{pure}:

$$\begin{aligned} e \in \mathit{Expr} &::= \ldots \mid \mathbf{reset}\ x \mid \mathbf{reuse}\ x\ \mathbf{in}\ \mathbf{ctor}_i\ \overline{y} \\ F \in \mathit{FnBody} &::= \ldots \mid \mathbf{inc}\ x;\ F \mid \mathbf{dec}\ x;\ F \end{aligned}$$

We use the subscripts $_{pure}$ or $_{RC}$ (e.g., Expr_{pure} or Expr_{RC}) to refer to the base or extended syntax, respectively, where otherwise ambiguous. The new expressions **reset** and **reuse** work together to reuse memory used to store constructor values, and, as discussed in Section 2, simulate destructive updates in constructor values.

We define the semantics of λ_{RC} (Figures 1 and 2) using a big-step relation $\rho \vdash \langle F, \sigma\rangle \Downarrow \langle l, \sigma'\rangle$ that maps the body F and a mutable *heap* σ under a context ρ to a location and the resulting heap. The context ρ maps variables to locations. A heap σ is a mapping from locations to pairs of values and reference counters. A value is a constructor value or a partially-applied constant. The reference counters of live values should always be positive; dead values are removed from the heap map.

$$\begin{aligned} l &\in \mathit{Loc} \\ \rho &\in \mathit{Ctxt} &&= \mathit{Var} \rightharpoonup \mathit{Loc} \\ \sigma &\in \mathit{Heap} &&= \mathit{Loc} \rightharpoonup \mathit{Value} \times \mathbb{N}^+ \\ v &\in \mathit{Value} ::= \mathbf{ctor}_i\ \overline{l} \mid \mathbf{pap}\ c\ \overline{l} \end{aligned}$$

When applying a variable, we have to be careful to increment the partial application arguments when copying them out of the **pap** cell, and to decrement the cell afterwards.[2] We cannot do so via explicit reference counting instructions because the number of arguments in a **pap** cell is not known statically.

[1] https://github.com/leanprover/lean4/blob/IFL19/library/init/lean/compiler/ir/basic.lean

[2] If the **pap** reference is unique, the two steps can be coalesced so that the arguments do not have to be touched.

$$\textsc{Const-App-Full}\quad \frac{\delta(c) = \lambda\ \overline{y_c}.\ F \qquad \overline{l} = \overline{\rho(y)} \qquad [\overline{y_c} \mapsto \overline{l}] \vdash \langle F, \sigma\rangle \Downarrow \langle l', \sigma'\rangle}{\rho \vdash \langle c\ \overline{y}, \sigma\rangle \Downarrow \langle l', \sigma'\rangle}$$

$$\textsc{Const-App-Part}\quad \frac{\delta(c) = \lambda\ \overline{y_c}.\ F \qquad \overline{l} = \overline{\rho(y)} \qquad |\ \overline{l}\ | < |\ \overline{y_c}\ | \qquad l' \notin \mathrm{dom}(\sigma)}{\rho \vdash \langle \mathbf{pap}\ c\ \overline{y}, \sigma\rangle \Downarrow \langle l', \sigma[l' \mapsto (\mathbf{pap}\ c\ \overline{l}, 1)]\rangle}$$

$$\textsc{Var-App-Full}\quad \frac{\sigma(\rho(x)) = (\mathbf{pap}\ c\ \overline{l}, \_) \qquad \delta(c) = \lambda\ \overline{y_c}.\ F \qquad l_y = \rho(y) \qquad [\overline{y_c} \mapsto \overline{l}\ l_y] \vdash \langle F, \mathrm{dec}(\rho(x), \mathrm{inc}(\overline{l}, \sigma))\rangle \Downarrow \langle l', \sigma'\rangle}{\rho \vdash \langle x\ y, \sigma\rangle \Downarrow \langle l', \sigma'\rangle}$$

$$\textsc{Var-App-Part}\quad \frac{\sigma(\rho(x)) = (\mathbf{pap}\ c\ \overline{l}, \_) \qquad \delta(c) = \lambda\ \overline{y_c}.\ F \qquad l_y = \rho(y) \qquad |\ \overline{l}\ l_y\ | < |\ \overline{y_c}\ | \qquad l' \notin \mathrm{dom}(\sigma)}{\rho \vdash \langle x\ y, \sigma\rangle \Downarrow \langle l', \mathrm{dec}(\rho(x), \mathrm{inc}(\overline{l}, \sigma))[l' \mapsto (\mathbf{pap}\ c\ \overline{l}\ l_y, 1)]\rangle}$$

$$\textsc{Ctor-App}\quad \frac{\overline{l} = \overline{\rho(y)} \qquad l' \notin \mathrm{dom}(\sigma)}{\rho \vdash \langle \mathbf{ctor}_i\ \overline{y}, \sigma\rangle \Downarrow \langle l', \sigma[l' \mapsto (\mathbf{ctor}_i\ \overline{l}, 1)]\rangle}$$

$$\textsc{Proj}\quad \frac{\sigma(\rho(x)) = (\mathbf{ctor}_j\ \overline{l}, \_) \qquad l' = \overline{l}_i}{\rho \vdash \langle \mathbf{proj}_i\ x, \sigma\rangle \Downarrow \langle l', \sigma\rangle} \qquad \textsc{Return}\quad \frac{\rho(x) = l}{\rho \vdash \langle \mathbf{ret}\ x, \sigma\rangle \Downarrow \langle l, \sigma\rangle}$$

$$\textsc{Let}\quad \frac{\rho \vdash \langle e, \sigma\rangle \Downarrow \langle l, \sigma'\rangle \qquad \rho[x \mapsto l] \vdash \langle F, \sigma'\rangle \Downarrow \langle l', \sigma''\rangle}{\rho \vdash \langle \mathbf{let}\ x = e;\ F, \sigma\rangle \Downarrow \langle l', \sigma''\rangle}$$

$$\textsc{Case}\quad \frac{\sigma(\rho(x)) = (\mathbf{ctor}_i\ \overline{l}, \_) \qquad \rho \vdash \langle F_i, \sigma\rangle \Downarrow \langle l', \sigma'\rangle}{\rho \vdash \langle \mathbf{case}\ x\ \mathbf{of}\ \overline{F}, \sigma\rangle \Downarrow \langle l', \sigma'\rangle}$$

Figure 1: λ_{RC} semantics: the λ_{pure} fragment

Decrementing a unique reference removes the value from the heap and recursively decrements its components. **reset**, when used on a unique reference, eagerly decrements the components of the referenced value, replaces them with ■,[3] and returns the location of the now-invalid cell. This value is intended to be used only by **reuse** or **dec**. The former reuses it for a new constructor cell, asserting that its size is compatible with the old cell. The latter frees the cell, ignoring the replaced children.

If **reset** is used on a shared, non-reusable reference, it behaves like **dec** and returns ■, which instructs **reuse** to behave like **ctor**. Note that we cannot simply return the reference in both cases and do another uniqueness check in **reuse** because other code between the two expressions may have altered its reference count.

5 A COMPILER FROM λ_{pure} TO λ_{RC}

Following the actual implementation of our compiler, we will discuss a compiler from λ_{pure} to λ_{RC} in three steps:

(1) Inserting **reset**/**reuse** pairs (Section 5.1)
(2) Inferring borrowed parameters (Section 5.2)
(3) Inserting **inc**/**dec** instructions (Section 5.3)

[3] which can be represented by any unused pointer value such as the null pointer in a real implementation. In our actual implementation, we avoid this memory write by introducing a `del` instruction that behaves like `dec` but ignores the constructor fields.

$$\textsc{Inc}\ \frac{\rho \vdash \langle F, \mathrm{inc}(\rho(x), \sigma)\rangle \Downarrow \langle l', \sigma'\rangle}{\rho \vdash \langle \mathbf{inc}\ x;\ F, \sigma\rangle \Downarrow \langle l', \sigma'\rangle} \qquad \textsc{Dec}\ \frac{\rho \vdash \langle F, \mathrm{dec}(\rho(x), \sigma)\rangle \Downarrow \langle l', \sigma'\rangle}{\rho \vdash \langle \mathbf{dec}\ x;\ F, \sigma\rangle \Downarrow \langle l', \sigma'\rangle}$$

$$\mathrm{inc}(l, \sigma) = \sigma[l \mapsto (v, i+1)] \quad \text{if } \sigma(l) = (v, i)$$

$$\mathrm{inc}(l\ \overline{l'}, \sigma) = \mathrm{inc}(\overline{l'}, \mathrm{inc}(l, \sigma))$$

$$\mathrm{dec}(l, \sigma) = \begin{cases} \sigma & \text{if } l = \blacksquare \\ \sigma[l \mapsto (v, i-1)] & \text{if } \sigma(l) = (v, i), i > 1 \\ \mathrm{dec}(\overline{l'}, \sigma[l \mapsto \bot]) & \text{if } \sigma(l) = (\mathbf{pap}\ c\ \overline{l'}, 1) \\ \mathrm{dec}(\overline{l'}, \sigma[l \mapsto \bot]) & \text{if } \sigma(l) = (\mathbf{ctor}_i\ \overline{l'}, 1) \end{cases}$$

$$\mathrm{dec}(l\ \overline{l'}, \sigma) = \mathrm{dec}(\overline{l'}, \mathrm{dec}(l, \sigma))$$

$$\textsc{Reset-Uniq}\ \frac{\rho(x) = l \qquad \sigma(l) = (\mathbf{ctor}_i\ \overline{l'}, 1)}{\rho \vdash \langle \mathbf{reset}\ x, \sigma\rangle \Downarrow \langle l, \mathrm{dec}(\overline{l'}, \sigma[l \mapsto (\mathbf{ctor}_i\ \blacksquare^{|\overline{l'}|}, 1)])\rangle}$$

$$\textsc{Reset-Shared}\ \frac{\rho(x) = l \qquad \sigma(l) = (\_, i) \qquad i \neq 1}{\rho \vdash \langle \mathbf{reset}\ x, \sigma\rangle \Downarrow \langle \blacksquare, \mathrm{dec}(l, \sigma)\rangle}$$

$$\textsc{Reuse-Uniq}\ \frac{\rho(x) = l \qquad \sigma(l) = (\mathbf{ctor}_j\ \blacksquare^{|\overline{y}|}, 1) \qquad \overline{\rho(y)} = \overline{l''}}{\rho \vdash \langle \mathbf{reuse}\ x\ \mathbf{in}\ \mathbf{ctor}_i\ \overline{y}, \sigma\rangle \Downarrow \langle l, \sigma[l \mapsto (\mathbf{ctor}_i\ \overline{l''}, 1)]\rangle}$$

$$\textsc{Reuse-Shared}\ \frac{\rho(x) = \blacksquare \qquad \rho \vdash \langle \mathbf{ctor}_i\ \overline{y}, \sigma\rangle \Downarrow \langle l', \sigma'\rangle}{\rho \vdash \langle \mathbf{reuse}\ x\ \mathbf{in}\ \mathbf{ctor}_i\ \overline{y}, \sigma\rangle \Downarrow \langle l', \sigma'\rangle}$$

Figure 2: λ_{RC} semantics cont.

The first two steps are optional for obtaining correct λ_{RC} programs.

5.1 Inserting destructive update operations

In this subsection, we will discuss a heuristics-based implementation of a function

$$\delta_{reuse} : Const \to Fn_{RC}$$

that inserts **reset**/**reuse** instructions. Given **let** z = **reset** x, we remark that, in every control path, z may appear at most once, and in one of the following two instructions: **let** y = **reuse** z **ctor**$_i$ $\overline{w}$, or **dec** z. We use **dec** z for control paths where z cannot be reused. We implement the function δ_{reuse} as

$$\delta_{reuse}(c) = \lambda\ \overline{y}.\ R(F) \text{ where } \delta(c) = \lambda\ \overline{y}.\ F$$

The function $R(F)$ (Figure 3) uses a simple heuristic for replacing **ctor**$_i$ $\overline{y}$ expressions occurring in F with **reuse** w **in ctor**$_i$ $\overline{y}$ where w is a fresh variable introduced by R as the result of a new **reset** operation. For each arm F_i in a **case** x **of** $\overline{F}$ operation, the function R requires the arity n of the corresponding matched constructor. In the actual implementation, we store this information for each arm when we compile our typed frontend language into λ_{pure}. The auxiliary functions D and S implement the *dead* variable search and *substitution* steps respectively. For each **case** operation, R attempts to insert **reset**/**reuse** instructions for the variable matched by the **case**. This is done using D in each arm of the **case**. Function

$$R : FnBody_{pure} \to FnBody_{RC}$$

$$R(\mathbf{let}\ x = e;\ F) = \mathbf{let}\ x = e;\ R(F)$$

$$R(\mathbf{ret}\ x) = \mathbf{ret}\ x$$

$$R(\mathbf{case}\ x\ \mathbf{of}\ \overline{F}) = \mathbf{case}\ x\ \mathbf{of}\ \overline{D(x, n_i, R(F_i))}$$

where n_i = #fields of x in i-th branch

$$D : Var \times \mathbb{N} \times FnBody_{RC} \to FnBody_{RC}$$

$$D(z, n, \mathbf{case}\ x\ \mathbf{of}\ \overline{F}) = \mathbf{case}\ x\ \mathbf{of}\ \overline{D(z, n, F)}$$

$$D(z, n, \mathbf{ret}\ x) = \mathbf{ret}\ x$$

$$D(z, n, \mathbf{let}\ x = e;\ F) = \mathbf{let}\ x = e;\ D(z, n, F)$$

if $z \in e$ or $z \in F$

$$D(z, n, F) = \mathbf{let}\ w = \mathbf{reset}\ z;\ S(w, n, F)$$

otherwise, if $S(w, n, F) \neq F$ for a fresh w

$$D(z, n, F) = F \quad \text{otherwise}$$

$$S : Var \times \mathbb{N} \times FnBody_{RC} \to FnBody_{RC}$$

$$S(w, n, \mathbf{let}\ x = \mathbf{ctor}_i\ \overline{y};\ F) = \mathbf{let}\ x = \mathbf{reuse}\ w\ \mathbf{in}\ \mathbf{ctor}_i\ \overline{y};\ F$$

if $|\ \overline{y}\ | = n$

$$S(w, n, \mathbf{let}\ x = e;\ F) = \mathbf{let}\ x = e;\ S(w, n, F) \quad \text{otherwise}$$

$$S(w, n, \mathbf{ret}\ x) = \mathbf{ret}\ x$$

$$S(w, n, \mathbf{case}\ x\ \mathbf{of}\ \overline{F}) = \mathbf{case}\ x\ \mathbf{of}\ \overline{S(w, n, F)}$$

Figure 3: Inserting reset/reuse pairs

$D(z, n, F)$ takes as parameters the variable z to reuse and the arity n of the matched constructor. D proceeds to the first location where z is dead, i.e. not used in the remaining function body, and then uses S to attempt to find and substitute a *matching* constructor **ctor**$_i$ $\overline{y}$ instruction with a **reuse** w **in ctor**$_i$ $\overline{y}$ in the remaining code. If no matching constructor instruction can be found, D does not modify the function body.

As an example, consider the *map* function for lists

```
map f xs = case xs of
  (ret xs)
  (let x = proj1 xs; let s := proj2 xs;
   let y = f x; let ys = map f s;
   let r = ctor2 y ys; ret r)
```

Applying R to the body of *map* , we have D looking for opportunities to **reset**/**reuse** *xs* in both **case** arms. Since *xs* is unused after **let** s = **proj**$_2$ *xs*, S is applied to the rest of the function, looking for constructor calls with two parameters. Indeed, such a call can be found in the let-binding for r. Thus, function D successfully inserts the appropriate instructions, and we obtain the function described in Section 2. Now, consider the list *swap* function that swaps the first two elements of a list. It is often defined as

```
swap [] = []
swap [x] = [x]
```

$$swap\ (x : y : zs) = y : x : zs$$

In λ_{pure}, this function is encoded as

$$\begin{array}{l} swap\ xs = \mathbf{case}\ xs\ \mathbf{of} \\ \quad (\mathbf{ret}\ xs) \\ \quad (\mathbf{let}\ t_1 = \mathbf{proj}_2\ xs; \mathbf{case}\ t_1\ \mathbf{of} \\ \quad\quad (\mathbf{ret}\ xs) \\ \quad\quad (\mathbf{let}\ h_1 = \mathbf{proj}_1\ xs; \\ \quad\quad\ \ \mathbf{let}\ h_2 = \mathbf{proj}_1\ t_1; \mathbf{let}\ t_2 = \mathbf{proj}_2\ t_1; \\ \quad\quad\ \ \mathbf{let}\ r_1 = \mathbf{ctor}_2\ h_1\ t_2; \mathbf{let}\ r_2 = \mathbf{ctor}_2\ h_2\ r_1; \mathbf{ret}\ r_2)) \end{array}$$

By applying R to *swap*, we obtain

$$\begin{array}{l} swap\ xs = \mathbf{case}\ xs\ \mathbf{of} \\ \quad (\mathbf{ret}\ xs) \\ \quad (\mathbf{let}\ t_1 = \mathbf{proj}_2\ xs; \mathbf{case}\ t_1\ \mathbf{of} \\ \quad\quad (\mathbf{ret}\ xs) \\ \quad\quad (\mathbf{let}\ h_1 = \mathbf{proj}_1\ xs; \mathbf{let}\ w_1 = \mathbf{reset}\ xs; \\ \quad\quad\ \ \mathbf{let}\ h_2 = \mathbf{proj}_1\ t_1; \mathbf{let}\ t_2 = \mathbf{proj}_2\ t_1; \\ \quad\quad\ \ \mathbf{let}\ w_2 = \mathbf{reset}\ t_1; \mathbf{let}\ r_1 = \mathbf{reuse}\ w_2\ \mathbf{in}\ \mathbf{ctor}_2\ h_1\ t_2; \\ \quad\quad\ \ \mathbf{let}\ r_2 = \mathbf{reuse}\ w_1\ \mathbf{in}\ \mathbf{ctor}_2\ h_2\ r_1; \mathbf{ret}\ r_2)) \end{array}$$

Similarly to the *map* function, the code generated for the function *swap* will *not* allocate any memory when the list value is not shared. This example demonstrates that our heuristic procedure can avoid memory allocations even in functions containing many nested **case** instructions. The example also makes it clear that we could further optimize our λ_{RC} by adding additional instructions. For example, we can add an instruction that combines **reset** and **reuse** into a single instruction and is used in situations where **reuse** occurs *immediately* after the corresponding **reset** instruction such as in the example above where we have $\mathbf{let}\ w_2 = \mathbf{reset}\ t_1;\ \mathbf{let}\ r_1 = \mathbf{reuse}\ w_2\ \mathbf{in}\ \mathbf{ctor}_2\ h_1\ t_2$,

5.2 Inferring borrowing signatures

We now consider the problem of inferring borrowing signatures, i.e. a mapping $\beta : Const \rightharpoonup \{\mathbb{O}, \mathbb{B}\}^*$, which for every function should return a list describing each parameter of the function as either $\mathbb{O}$wned or $\mathbb{B}$orrowed. Borrow annotations can be provided manually by users (which is always safe), but we have two motivations for inferring them: avoiding the burden of annotations, and making our IR a convenient target for other systems (e.g., Coq, Idris, and Agda) that do not have borrow annotations.

If a function f takes a parameter x as a borrowed reference, then at runtime x may be a shared value even when its reference counter is 1. Thus, we must never mark x as borrowed if it is used by a $\mathbf{let}\ y = \mathbf{reset}\ x$ instruction. We also assume that each $\beta(c)$ has the same length as the corresponding parameter list in $\delta(c)$.

Partially applying constants with borrowed parameters is also problematic because, in general, we cannot statically assert that the resulting value will not escape the current function and thus the scope of borrowed references. Therefore we extend δ_{reuse} to the program δ_β by defining a trivial wrapper constant $c_\mathbb{O} := c$ (we will assume that this name is fresh) for any such constant c, set $\beta(c_\mathbb{O}) := \overline{\mathbb{O}}$, and replace any occurrence of $\mathbf{pap}\ c\ \overline{y}$ with $\mathbf{pap}\ c_\mathbb{O}\ \overline{y}$. The compiler step given in the next subsection will, as part of the general transformation, insert the necessary **inc** and **dec** instructions into $c_\mathbb{O}$ to convert between the two signatures.

$$\begin{array}{l} collect_\mathbb{O} : FnBody_{RC} \to 2^{Vars} \\ collect_\mathbb{O}(\mathbf{let}\ z = \mathbf{ctor}_i\ \overline{x}; F) = collect_\mathbb{O}(F) \\ collect_\mathbb{O}(\mathbf{let}\ z = \mathbf{reset}\ x; F) = collect_\mathbb{O}(F) \cup \{x\} \\ collect_\mathbb{O}(\mathbf{let}\ z = \mathbf{reuse}\ x\ \mathbf{in}\ \mathbf{ctor}_i\ \overline{x}; F) = collect_\mathbb{O}(F) \\ collect_\mathbb{O}(\mathbf{let}\ z = c\ \overline{x}; F) = collect_\mathbb{O}(F) \cup \{x_i \in \overline{x} \mid \beta(c)_i = \mathbb{O}\} \\ collect_\mathbb{O}(\mathbf{let}\ z = x\ y; F) = collect_\mathbb{O}(F) \cup \{x, y\} \\ collect_\mathbb{O}(\mathbf{let}\ z = \mathbf{pap}\ c_\mathbb{O}\ \overline{x}; F) = collect_\mathbb{O}(F) \cup \{\overline{x}\} \\ collect_\mathbb{O}(\mathbf{let}\ z = \mathbf{proj}_i\ x; F) = collect_\mathbb{O}(F) \cup \{x\} \text{ if } z \in collect_\mathbb{O}(F) \\ collect_\mathbb{O}(\mathbf{let}\ z = \mathbf{proj}_i\ x; F) = collect_\mathbb{O}(F) \text{ if } z \notin collect_\mathbb{O}(F) \\ collect_\mathbb{O}(\mathbf{ret}\ x) = \emptyset \\ collect_\mathbb{O}(\mathbf{case}\ x\ \mathbf{of}\ \overline{F}) = \bigcup_{F_i \in \overline{F}} collect_\mathbb{O}(F_i) \end{array}$$

Figure 4: Collecting variables that should not be marked as borrowed

Our heuristic is based on the fact that when we mark a parameter as borrowed, we reduce the number of RC operations needed, but we also prevent **reset** and **reuse** as well as primitive operations from reusing memory cells. Our heuristic collects which parameters and variables should be owned. We say a parameter x should be owned if x or one of its projections is used in a **reset**, or is passed to a function that takes an owned reference. The latter condition is a heuristic and is not required for correctness. We use it because the function taking an owned reference may try to reuse its memory cell. A formal definition is given in Figure 4. Many refinements are possible, and we discuss one of them in the next section. Note that if a call is recursive, we do not know which parameters are owned, yet. Thus, given $\delta(c) = \lambda \overline{y}.\ b$, we infer the value of $\beta(c)$ by starting with the approximation $\beta(c) = \mathbb{B}^n$, then we compute $S = collect_\mathbb{O}(b)$, update $\beta(c)_i := \mathbb{O}$ if $y_i \in S$, and repeat the process until we reach a fix point and no further updates are performed on $\beta(c)$. The procedure described here does not consider mutually recursive definitions, but this is a simple extension where we process a block of mutually recursive functions simultaneously. By applying our heuristic to the *hasNone* function described before, we obtain $\beta(hasNone) = \mathbb{B}$. That is, in an application *hasNone xs*, *xs* is taken as a borrowed reference.

5.3 Inserting reference counting operations

Given any well-formed definition of β and δ_β, we finally give a procedure for correctly inserting **inc** and **dec** instructions.[4]

$$\begin{array}{ll} \delta_{RC}(c) : Const \to Fn_{RC} & \\ \delta_{RC}(c) = \lambda\ \overline{y}.\ \mathbb{O}^-(\overline{y}, C(F, \beta_l)) & \text{where } \delta_\beta(c) = \lambda\ \overline{y}.\ F, \\ & \qquad \beta_l = [\overline{y} \mapsto \beta(c), \ldots \mapsto \mathbb{O}] \end{array}$$

The map $\beta_l : Var \to \{\mathbb{O}, \mathbb{B}\}$ keeps track of the borrow status of each local variable. For simplicity, we default all missing entries to $\mathbb{O}$.

[4] We will tersely say that a variable x "is incremented/decremented" when an **inc**/**dec** operation is applied to it, i.e. the RC of the referenced object is incremented/decremented at runtime.

$$
\begin{aligned}
&C : \mathit{FnBody}_{RC} \times (\mathit{Var} \to \{\mathbb{O}, \mathbb{B}\}) \to \mathit{FnBody}_{RC}\\
&C\,(\texttt{ret}\ x, \beta_l) = \mathbb{O}^+_x(\emptyset, \texttt{ret}\ x, \beta_l)\\
&C\,(\texttt{case}\ x\ \texttt{of}\ \overline{F}, \beta_l) = \texttt{case}\ x\ \texttt{of}\ \overline{\mathbb{O}^-(\overline{y}, C(F, \beta_l), \beta_l)}\\
&\qquad\qquad \text{where } \{\overline{y}\} = \mathrm{FV}(\texttt{case}\ x\ \texttt{of}\ \overline{F})\\
&C\,(\texttt{let}\ y = \texttt{proj}_i\ x; F, \beta_l) = \texttt{let}\ y = \texttt{proj}_i\ x; \texttt{inc}\ y; \mathbb{O}^-_x(C(F, \beta_l), \beta_l)\\
&\qquad\qquad \text{if } \beta_l(x) = \mathbb{O}\\
&C\,(\texttt{let}\ y = \texttt{proj}_i\ x; F, \beta_l) = \texttt{let}\ y = \texttt{proj}_i\ x; C\,(F, \beta_l[y \mapsto \mathbb{B}])\\
&\qquad\qquad \text{if } \beta_l(x) = \mathbb{B}\\
&C\,(\texttt{let}\ y = \texttt{reset}\ x; F, \beta_l) = \texttt{let}\ y = \texttt{reset}\ x; C\,(F, \beta_l)\\
&C\,(\texttt{let}\ z = c\ \overline{y}\,; F, \beta_l) = C_{app}(\overline{y}, \beta(c), \texttt{let}\ z = c\ \overline{y};\ C(F, \beta_l), \beta_l)\\
&C\,(\texttt{let}\ z = \texttt{pap}\ c\ \overline{y}\,; F, \beta_l) = C_{app}(\overline{y}, \beta(c), \texttt{let}\ z = \texttt{pap}\ c\ \overline{y};\ C(F, \beta_l), \beta_l)\\
&C\,(\texttt{let}\ z = x\ y; F, \beta_l) = C_{app}(x\ y, \mathbb{O}\ \mathbb{O}, \texttt{let}\ z = x\ y;\ C(F, \beta_l), \beta_l)\\
&C\,(\texttt{let}\ z = \texttt{ctor}_i\ \overline{y}\,; F, \beta_l) = C_{app}(\overline{y}, \overline{\mathbb{O}}, \texttt{let}\ z = \texttt{ctor}_i\ \overline{y};\ C(F, \beta_l), \beta_l)\\
&C\,(\texttt{let}\ z = \texttt{reuse}\ x\ \texttt{in}\ \texttt{ctor}_i\ \overline{y}\,; F, \beta_l) =\\
&\quad C_{app}(\overline{y}, \overline{\mathbb{O}}, \texttt{let}\ z = \ \texttt{reuse}\ x\ \texttt{in}\ \texttt{ctor}_i\ \overline{y};\ C(F, \beta_l), \beta_l)
\end{aligned}
$$

$$
\begin{aligned}
&C_{app} : \mathit{Var}^* \times \{\mathbb{O}, \mathbb{B}\}^* \times \mathit{FnBody}_{RC} \times (\mathit{Var} \to \{\mathbb{O}, \mathbb{B}\}) \to \mathit{FnBody}_{RC}\\
&C_{app}(y\ \overline{y'}, \mathbb{O}\ \overline{b}, \texttt{let}\ z = e;\ F, \beta_l) =\\
&\qquad \mathbb{O}^+_y(\overline{y'} \cup \mathrm{FV}(F), C_{app}(\overline{y'}, \overline{b}, \texttt{let}\ z = e;\ F, \beta_l), \beta_l)\\
&C_{app}(y\ \overline{y'}, \mathbb{B}\ \overline{b}, \texttt{let}\ z = e;\ F, \beta_l) =\\
&\qquad C_{app}(\overline{y'}, \overline{b}, \texttt{let}\ z = e;\ \mathbb{O}^-_y(F, \beta_l), \beta_l)\\
&C_{app}([\,], \_, \texttt{let}\ z = e;\ F, \beta_l) = \texttt{let}\ z = e;\ F
\end{aligned}
$$

Figure 5: Inserting inc/dec instructions

In general, variables should be incremented prior to being used in an *owned context* that consumes an RC token. Variables used in any other (*borrowed*) context do not need to be incremented. Owned references should be decremented after their last use. We use the following two helper functions to conditionally add RC instructions (Figure 5) in these contexts:

- $\mathbb{O}^+_x$ prepares x for usage in an owned context by incrementing it. The increment can be omitted on the last use of an owned variable, with V representing the set of live variables after the use.

$$
\begin{aligned}
&\mathbb{O}^+_x(V, F, \beta_l) = F && \text{if } \beta_l(x) = \mathbb{O} \wedge x \notin V\\
&\mathbb{O}^+_x(V, F, \beta_l) = \texttt{inc}\ x;\ F && \text{otherwise}
\end{aligned}
$$

- $\mathbb{O}^-_x$ decrements x if it is both owned and dead. $\mathbb{O}^-(\overline{x}, F, \beta_l)$ decrements multiple variables, which may be needed at the start of a function or **case** branch.

$$
\begin{aligned}
&\mathbb{O}^-_x(F, \beta_l) = \texttt{dec}\ x;\ F && \text{if } \beta_l(x) = \mathbb{O} \wedge x \notin \mathrm{FV}(F)\\
&\mathbb{O}^-_x(F, \beta_l) = F && \text{otherwise}\\
&\mathbb{O}^-(x\ \overline{x'}, F, \beta_l) = \mathbb{O}^-(\overline{x'}, \mathbb{O}^-_x(F, \beta_l), \beta_l)\\
&\mathbb{O}^-([\,], F, \beta_l) = F
\end{aligned}
$$

Applications are handled separately, recursing over the arguments and parameter borrow annotations in parallel (in the case of partial applications, the former list may be shorter); for partial, variable and constructor applications, the latter default to $\mathbb{O}$.

Examples

We demonstrate the behavior of the compiler on two application special cases. The value of β_l is constant in these examples and left implicit in applications.

(1) Consuming the same argument multiple times

$$
\begin{aligned}
&\beta(\mathsf{c}) := \mathbb{O}\ \mathbb{O}\\
&\quad \beta_l := [\mathsf{y} \mapsto \mathbb{O}]\\
&C(\texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z})\\
&\qquad = C_{app}(\mathsf{y}\ \mathsf{y}, \mathbb{O}\ \mathbb{O}, \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ C(\texttt{ret}\ \mathsf{z}))\\
&\qquad = C_{app}(\mathsf{y}\ \mathsf{y}, \mathbb{O}\ \mathbb{O}, \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z})\\
&\qquad = \mathbb{O}^+_{\mathsf{y}}(\{\mathsf{y}, \mathsf{z}\}, C_{app}(\mathsf{y}, \mathbb{O}, \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z}))\\
&\qquad = \mathbb{O}^+_{\mathsf{y}}(\{\mathsf{y}, \mathsf{z}\}, \mathbb{O}^+_{\mathsf{y}}(\{\mathsf{z}\}, C_{app}([\,], [\,], \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z})))\\
&\qquad = \mathbb{O}^+_{\mathsf{y}}(\{\mathsf{y}, \mathsf{z}\}, \mathbb{O}^+_{\mathsf{y}}(\{\mathsf{z}\}, \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z}))\\
&\qquad = \mathbb{O}^+_{\mathsf{y}}(\{\mathsf{y}, \mathsf{z}\}, \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z})\\
&\qquad = \texttt{inc}\ \mathsf{y};\ \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z}
\end{aligned}
$$

Because y is dead after the call, it needs to be incremented only once, *moving* its last token to c instead.

(2) Borrowing and consuming the same argument

$$
\begin{aligned}
&\beta(\mathsf{c}) := \mathbb{B}\ \mathbb{O}\\
&\quad \beta_l := [\mathsf{y} \mapsto \mathbb{O}]\\
&C(\texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z})\\
&\qquad = C_{app}(\mathsf{y}\ \mathsf{y}, \mathbb{B}\ \mathbb{O}, \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ C(\texttt{ret}\ \mathsf{z}))\\
&\qquad = C_{app}(\mathsf{y}\ \mathsf{y}, \mathbb{B}\ \mathbb{O}, \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z})\\
&\qquad = C_{app}(\mathsf{y}, \mathbb{O}, \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \mathbb{O}^-_{\mathsf{y}}(\texttt{ret}\ \mathsf{z}))\\
&\qquad = C_{app}(\mathsf{y}, \mathbb{O}, \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{dec}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z})\\
&\qquad = \mathbb{O}^+_{\mathsf{y}}(\{\mathsf{y}, \mathsf{z}\},\ C_{app}([\,], [\,], \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{dec}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z}))\\
&\qquad = \mathbb{O}^+_{\mathsf{y}}(\{\mathsf{y}, \mathsf{z}\}, \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{dec}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z})\\
&\qquad = \texttt{inc}\ \mathsf{y};\ \texttt{let}\ \mathsf{z} = \mathsf{c}\ \mathsf{y}\ \mathsf{y};\ \texttt{dec}\ \mathsf{y};\ \texttt{ret}\ \mathsf{z}
\end{aligned}
$$

Even though the owned parameter comes after the borrowed parameter, the presence of y in the **dec** instruction emitted when handling the first parameter makes sure we do not accidentally move ownership when handling the second parameter, but copy y by emitting an **inc** instruction.

Preserving tail calls

A *tail call* $\texttt{let}\ r = c\ \overline{x};\ \texttt{ret}\ r$ is an application followed by a **ret** instruction. Recursive tail calls are implemented using *goto*s in our compiler backend. Thus, it is highly desirable to preserve them as we transform λ_{pure} into λ_{RC}. However, the previous example shows that our function for inserting reference counting instructions may insert **dec** instructions after a constant application, and consequently, destroy tail calls. A **dec** instruction is inserted after a constant application $\texttt{let}\ r = c\ \overline{x}$ if $\beta(c)_i = \mathbb{B}$ and $\beta_l(x_i) = \mathbb{O}$ for some $x_i \in \overline{x}$. That is, function c takes the i-th parameter as a borrowed reference, but the actual argument is owned. As an example, consider the following function in λ_{pure}.

$$
\begin{aligned}
&f\,x = \texttt{case}\ x\ \texttt{of}\\
&\quad (\texttt{let}\ r = \texttt{proj}_1\,; \texttt{ret}\ r)\\
&\quad (\texttt{let}\ y_1 = \texttt{ctor}_1\,; \texttt{let}\ y_2 = \texttt{ctor}_1\ y_1\,; \texttt{let}\ r = f\,y_2\,; \texttt{ret}\ r)
\end{aligned}
$$

The compiler from λ_{pure} to λ_{RC}, infers $\beta(f) = \mathbb{B}$, and produces

$$
\begin{array}{l}
f\ x = \texttt{case}\ x\ \texttt{of} \\
\quad (\texttt{let}\ r = \texttt{proj}_1\ x; \texttt{inc}\ r; \texttt{ret}\ r) \\
\quad (\texttt{let}\ y_1 = \texttt{ctor}_1\ ; \texttt{let}\ y_2\ \texttt{ctor}_1\ y_1\ ; \\
\quad\ \texttt{let}\ r = f\ y_2\ ; \texttt{dec}\ y_2\ ; \texttt{ret}\ r)
\end{array}
$$

which does not preserve the tail call $\texttt{let}\ r = f\ y_2;\ \texttt{ret}\ r$. We addressed this issue in our real implementation by refining our borrowing inference heuristic, marking $\beta(c)_i = \mathbb{O}$ whenever c occurs in a tail call $\texttt{let}\ r = c\ \overline{x};\ \texttt{ret}\ r$ where $\beta_l(x_i) = \mathbb{O}$. This small modification guarantees that tail calls are preserved by our compiler, and the following λ_{RC} code is produced for f instead

$$
\begin{array}{l}
f\ x = \texttt{case}\ x\ \texttt{of} \\
\quad (\texttt{let}\ r = \texttt{proj}_1\ x; \texttt{inc}\ r; \texttt{dec}\ x; \texttt{ret}\ r) \\
\quad (\texttt{dec}\ x; \texttt{let}\ y_1 = \texttt{ctor}_1\ ; \texttt{let}\ y_2 = \texttt{ctor}_1\ y_1\ ; \\
\quad\ \texttt{let}\ r = f\ y_2\ ; \texttt{ret}\ r)
\end{array}
$$

6 OPTIMIZING FUNCTIONAL DATA STRUCTURES FOR `reset`/`reuse`

In the previous section, we have shown how to automatically insert **`reset`** and **`reuse`** instructions that minimize the number of memory allocations at execution time. We now discuss techniques we have been using for taking advantage of this transformation when writing functional code. Two fundamental questions when using this optimization are: Does a **`reuse`** instruction now guard my constructor applications? Given a $\texttt{let}\ y = \texttt{reset}\ x$ instruction, how often is x not shared at runtime? We address the first question using a simple static analyzer that when invoked by a developer, checks whether **`reuse`** instructions are guarding constructor applications in a particular function. This kind of analyzer is straightforward to implement in Lean since our IR is a Lean inductive datatype. This kind of analyzer is in the same spirit of the *inspection-testing* package available for GHC [Breitner 2018]. We cope with the second question using runtime instrumentation. For each $\texttt{let}\ y = \texttt{reset}\ x$ instruction, we can optionally emit two counters that track how often x is shared or not. We have found these two simple techniques quite useful when optimizing our own code. Here, we report one instance that produced a significant performance improvement.

Red-black trees

Red-black trees are implemented in the Lean standard library and are often used to write proof automation. For the purposes of this section, it is sufficient to have an abstract description of this kind of tree, and one of the re-balancing functions used by the insertion function.

$$
\begin{array}{l}
\mathit{Color} = R \mid B \\
\mathit{Tree}\ a = E \mid T\ \mathit{Color}\ (\mathit{Tree}\ a)\ a\ (\mathit{Tree}\ a) \\
\mathit{balance}_1\ v\ t\ (T\ \_\ (T\ R\ l\ x\ r_1)\ y\ r_2) = T\ R\ (T\ B\ l\ x\ r_1)\ y\ (T\ B\ r_2\ v\ t) \\
\mathit{balance}_1\ v\ t\ (T\ \_\ l_1\ y\ (T\ R\ l_2\ x\ r)) = T\ R\ (T\ B\ l_1\ y\ l_2)\ x\ (T\ B\ r\ v\ t) \\
\mathit{balance}_1\ v\ t\ (T\ \_\ l\ y\ r) = T\ B\ (T\ R\ l\ y\ r)\ v\ t \\
\mathit{insert}\ (T\ B\ a\ y\ b)\ x = \mathit{balance}_1\ y\ b\ (\mathit{insert}\ a\ x)\ \mathit{if}\ x < y\ \mathit{and}\ a\ \mathit{is\ red} \\
\ldots
\end{array}
$$

Note that the first two $\mathit{balance}_1$ equations create three T constructor values, but the patterns on the left-hand side use only two T constructors. Thus, the generated IR for $\mathit{balance}_1$ contains T constructor applications that are not guarded by *reuse*, and this fact can be detected at compilation time. Note that even if the result of $(\mathit{insert}\ a\ x)$ contains only nonshared values, we still have to allocate one constructor value. We can avoid this unnecessary memory allocation by inlining $\mathit{balance}_1$. After inlining, the input value $(T\ B\ a\ y\ b)$ is reused in the $\mathit{balance}_1$ code. The final generated code now contains a single constructor application that is not guarded by a *reuse*, the one for the equation:

$$\mathit{insert}\ E\ x = T\ R\ E\ x\ E$$

The generated code now has the property that if the input tree is not shared, then only a single new node is allocated. Moreover, even if the input tree is shared we have observed a positive performance impact using **`reset`** and **`reuse`**. The recursive call $(\mathit{insert}\ a\ x)$ always returns a nonshared node even if x is shared. Thus, $\mathit{balance}_1\ y\ b\ (\mathit{insert}\ a\ x)$ always reuses at least one memory cell at runtime.

There is another way to avoid the unnecessary memory allocation that does not rely on inlining. We can chain the T constructor value from *insert* to $\mathit{balance}_1$. We accomplish this by rewriting $\mathit{balance}_1$ and *insert* as follows

$$
\begin{array}{l}
\mathit{balance}_1\ (T\ \_\ \_\ v\ t)\ (T\ \_\ (T\ R\ l\ x\ r_1)\ y\ r_2) = T\ R\ (T\ B\ l\ x\ r_1)\ y\ (T\ B\ r_2\ v\ t) \\
\mathit{balance}_1\ (T\ \_\ \_\ v\ t)\ (T\ \_\ l_1\ y\ (T\ R\ l_2\ x\ r)) = T\ R\ (T\ B\ l_1\ y\ l_2)\ x\ (T\ B\ r\ v\ t) \\
\mathit{balance}_1\ (T\ \_\ \_\ v\ t)\ (T\ \_\ l\ y\ r) = T\ B\ (T\ R\ l\ y\ r)\ v\ t \\
\mathit{insert}\ (T\ B\ a\ y\ b)\ x = \mathit{balance}_1\ (T\ B\ E\ y\ b)\ (\mathit{insert}\ a\ x)\ \mathit{if}\ x < y\ \mathit{and}\ a\ \mathit{is\ red}
\end{array}
$$

Now, the input value $(T\ B\ a\ y\ b)$ is reused to create value $(T\ B\ E\ y\ b)$ which is passed to $\mathit{balance}_1$. Note that we have replaced a with E to make sure the recursive application $(\mathit{insert}\ a\ x)$ may also perform destructive updates if a is not shared. This simple modification guarantees that $\mathit{balance}_1$ does not allocate memory when the input trees are not shared.

7 RUNTIME IMPLEMENTATION

7.1 Values

In our runtime, every value starts with a header containing two tags. The first tag specifies the value kind: **`ctor`**, **`pap`**, **`array`**, **`string`**, **`num`**, **`thunk`**, or **`task`**. The second tag specifies whether the value is single-threaded, multi-threaded, or *persistent*. We will describe how this kind is used to implement thread safe reference counting in the next subsection. The kinds **`ctor`** and **`pap`** are used to implement the corresponding values used in the formal semantics of λ_{pure} and λ_{RC}. The kinds **`array`**, **`string`**, and **`thunk`** are self explanatory. The kind **`num`** is for arbitrary precision numbers implemented using the GNU multiple precision library (GMP). The **`task`** value is described in the next subsection.

Values tagged as single- or multi-threaded also contain a reference counter. This counter is stored in front of the standard value header. We will primarily focus on the layout of **`ctor`** values here because it is the most relevant one for the ideas presented in this paper. A $\texttt{ctor}_i$ value header also includes the constructor index i, the number of pointers to other values and/or boxed values, and the number of bytes used to store scalar unboxed values such as machine integers and enumeration types. In a 64-bit machine, the **`ctor`** value header is 16 bytes long, twice the size of the header

used in OCaml to implement the corresponding kind of value. After the header, we store all pointers to other values and boxed values, and then all unboxed values. Thus, in a 64 bit machine, our runtime uses 32 bytes to implement a *List Cons* value: 16 bytes for the header, and 16 bytes for storing the list head and tail. The unboxed value support has restrictions similar to the ones found in GHC. For example, to pass an unboxed value to a polymorphic function we must first box it.

Our runtime has built-in support for array and string operations. Strings are just a special case of arrays where the elements are characters. We perform destructive updates when the array is not shared. For example, given the array write primitive

$$Array.write : Array\ \alpha \to Nat \to \alpha \to Array\ \alpha$$

the function application *Array.write a i v* will destructively update and return the array a if it is not shared. This is a well known optimization for systems based on reference counting [Jones and Lins 1996], nonetheless we mention it here because it is relevant for many applications. Moreover, destructive array updates and our **reset**/**reuse** technique complement each other. As an example, if we have a nonshared list of integer arrays *xs*, *map* (*Array.map inc*) *xs* destructively updates the list and all arrays. In the experimental section we demonstrate that our pure quick sort is as efficient as the quick sort using destructive updates in OCaml, and the quick sort using the primitive ST monad in Haskell.

7.2 Thread safety

We use the following basic task management primitives to develop the Lean frontend.

$$Task.mk : (Unit \to \alpha) \to Task\ \alpha$$
$$Task.bind : Task\ \alpha \to (\alpha \to Task\ \beta) \to Task\ \beta$$
$$Task.get : Task\ \alpha \to \alpha$$

The function *Task.mk* converts a closure into a **task** value and executes it in a separate thread, *Task.bind t f* creates a **task** value that waits for t to finish and produce result a, and then starts $f\ a$ and waits for it to finish. Finally, *Task.get t* waits for t to finish and returns the value produced by it. These primitives are part of the Lean runtime, implemented in C++, and are available to regular users.

The standard way of implementing thread safe reference counting uses memory fences [Schling 2011]. The reference counters are incremented using an atomic fetch and add operation with a relaxed memory order. The relaxed memory order can be used because new references to a value can only be formed from an existing reference, and passing an existing reference from one thread to another must already provide any required synchronization. When decrementing a reference counter, it is important to enforce that any decrements of the counter from other threads are visible before checking if the object should be deleted. The standard way of achieving this effect uses a *release* operation after dropping a reference, and an *acquire* operation before the deletion check. This approach has been used in the previous version of the Lean compiler, and we have observed that the memory fences have a significant performance impact even when only one thread is being executed. This is quite unfortunate because most values are only touched by a single execution thread.

We have addressed this performance problem in our runtime by tagging values as *single-threaded*, *multi-threaded*, or *persistent*. As the name suggests, a single-threaded value is accessed by a single thread and a multi-threaded one by one or more threads. If a value is tagged as single-threaded, we do not use any memory fence for incrementing or decrementing its reference counter. Persistent values are never deallocated and do not even need a reference counter. We use persistent values to implement values that are created at program initialization time and remain alive until program termination. Our runtime enforces the following invariant: from persistent values, we can only reach other persistent values, and from multi-threaded values, we can only reach persistent or multi-threaded values. There are no constraints on the kind of value that can be reached from a single-threaded value. By default, values are single-threaded, and our runtime provides a $markMT(o)$ procedure that tags all single-threaded values reachable from o as multi-threaded. This procedure is used to implement *Task.mk f* and *Task.bind x f*. We use $markMT(f)$ and $markMT(x)$ to ensure that all values reachable from these values are tagged as multi-threaded *before* we create a new task, that is, while they are still accessible from only one thread. Our invariant ensures that *markMT* does not need to visit values reachable from a value already tagged as multi-threaded. Thus values are visited at most once by *markMT* during program execution. Note that task creation is not a constant time operation in our approach because it is proportional to the number of single-threaded values reachable from x and f. This does not seem to be a problem in practice, but if it becomes an issue we can provide a primitive *asMT g* that ensures that all values allocated when executing g are immediately tagged as multi-threaded. Users would then use this flag in code that creates the values reachable by *Task.mk f* and *Task.bind x f*.

The reference counting operations perform an extra operation to test the value tag and decide whether a memory fence is needed or not. This additional test does not require any synchronization because the tag is only modified before a value is shared with other execution threads. In the experimental section, we demonstrate that this simple approach significantly boosts performance. This is not surprising because the additional test is much cheaper than memory fences on modern hardware. The approach above can be adapted to more complex libraries for writing multi-threaded code. We just need to identify which functions may send values to other execution threads, and use *markMT*.

8 EXPERIMENTAL EVALUATION

We have implemented the RC optimizations described in the previous sections in the new compiler for the Lean programming language. We have implemented all optimizations in Lean, and they are available online [5]. At the time of writing, the compiler supports only one backend where we emit C code. We are currently working on an LLVM backend for our compiler. To test the efficiency of the compiler and RC optimizations, we have devised a number of benchmarks[6] that aim to replicate common tasks performed in compilers and proof assistants. All timings are arithmetic means of 50 runs as reported by the `temci` benchmarking tool [Bechberger

[5] https://github.com/leanprover/lean4/tree/master/library/init/lean/compiler/ir
[6] https://github.com/leanprover/lean4/tree/IFL19/tests/bench

2016], executed on a PC with an i7-3770 Intel CPU and 16 GB RAM running Ubuntu 18.04, using Clang 9.0.0 for compiling the Lean runtime library as well as the C code emitted by the Lean compiler.

- `deriv` and `const_fold` implement differentiation and constant folding, respectively, as examples of symbolic term manipulation where big expressions are constructed and transformed. We claim they reflect operations frequently performed by proof automation procedures used in theorem provers.
- `rbmap` stress tests the red-black tree implementation from the Lean standard library. The benchmarks `rbmap_10` and `rbmap_1` are two variants where we perform updates on shared trees.
- `parser` is the new parser we are developing for the next version of Lean. It is written purely in Lean (approximately 2000 lines of code).
- `qsort` it is the basic quicksort algorithm for sorting arrays.
- `binarytrees` is taken from the Computer Languages Benchmarks Game[7]. This benchmark is a simple adaption of Hans Boehm's `GCBench` benchmark[8]. The Lean version is a translation of the fastest, parallelized Haskell solution, using `Task` in place of the Haskell `parallel` API.
- `unionfind` implements the union-find algorithm which is frequently used to implement decision procedures in automated reasoning. We use arrays to store the *find* table, and thread the state using a state monad transformer

We have tested the impact of each optimization by selectively disabling it and comparing the resulting runtime with the base runtime (Figure 6):

- *-reuse* disables the insertion of **`reset`**/**`reuse`** operations
- *-borrow* disables borrow inference, assuming that all parameters are owned. Note that the compiler must still honor borrow annotations on builtins, which are unaffected.
- *-ST* uses atomic RC operations for all values

The results show that the new **`reset`** and **`reuse`** instructions significantly improve performance in the benchmarks `const_fold`, `rbmap`, and `unionfind`. The borrowed inference heuristic provides significant speedups in benchmarks `binarytrees` and `deriv`.

We have also directly translated some of these programs to other statically typed, functional languages: Haskell, OCaml, and Standard ML (Figure 7). For the latter we selected the compilers MLton [Weeks 2006], which performs whole program optimization and can switch between multiple GC schemes at runtime, and MLKit, which combines Region Inference and garbage collection [Hallenberg et al. 2002]. While not primarily a functional language, we have also included Swift as a popular statically typed language using reference counting. For `binarytrees`, we have used the original files and compiler flags from the fastest Benchmark Game implementations. For Swift, we used the second-fastest, safe implementation, which is much more comparable to the other versions than the fastest one completely depending on unsafe code. The Benchmark Game does not include an SML version. For `qsort`, the Lean code is pure and relies on the fact that array updates are destructive if the array is not shared. The Swift code behaves similarly because Swift arrays are copy-on-write. All other versions use destructive updates, using the *ST* monad in the case of Haskell.

[7] https://benchmarksgame-team.pages.debian.net/benchmarksgame/performance/binarytrees.html
[8] http://hboehm.info/gc/gc_bench/

	base	*-reuse*	*-borrow*	*-ST*
`binarytrees`	$1.00\tilde{}$	$0.98\tilde{}$	$1.14\tilde{}$	$1.22\tilde{}$
`deriv`	$1.00\tilde{}$	1.00	$1.16\tilde{}$	1.42
`const_fold`	1.00	1.64	0.90	1.23
`parser`	1.00	1.00	$1.00\tilde{}$	1.68
`qsort`	1.00	1.00	1.00	1.13
`rbmap`	$1.00\tilde{}$	3.23	$1.07\tilde{}$	1.71
`rbmap_10`	1.49	3.62	1.52	2.43
`rbmap_1`	$4.72\tilde{}$	5.42	4.47	8.02
`unionfind`	$1.00\tilde{}$	$1.41\tilde{}$	$1.00\tilde{}$	2.31
geom. mean	1.24	1.74	$1.27\tilde{}$	1.89

Figure 6: Lean variant benchmarks, normalized by the base run time (`rbmap` for `rbmap_*`). Digits whose order of magnitude is no larger than that of twice the standard deviation are marked by squiggly lines.

While the absolute runtimes in Figure 7 are influenced by many factors other than the implementation of garbage collection that make direct comparisons difficult, the results still signify that both our garbage collection and the overall runtime and compiler implementation are very competitive. We initially conjectured the good performance was a result of reduced cache misses due to reusing allocations and a lack of GC tracing. However, the results demonstrate this is not the case. The only benchmark where the code generated by our compiler produces significantly fewer cache misses is `rbmap`. Note that Lean is 5x as fast as OCaml on `const_fold` even though it triggers more cache misses per second. The results suggest that Lean code is often faster in the benchmarks where the code generated by other compilers spends a significant amount of time performing GC. Using `const_fold` as an example again, Lean spends only 17% of the runtime deallocating memory, while OCaml spends 90% in the GC. This comparison is not entirely precise since it does not include the amount of time Lean spends updating reference counts, but it seems to be the most plausible explanation for the difference in performance. The results for `qsort` are surprising, the Lean and Swift implementations outperforms all destructive ones but MLton. We remark that MLton and Swift have a clear advantage since they use arrays of unboxed machine integers, while Lean and the other compilers use boxed values. We did not find a way to disable this optimization in MLton or Swift to confirm our conjecture. We believe this benchmark demonstrates that our compiler allows programmers to write efficient pure code that uses arrays and hashtables. For `rbmap`, Lean is much faster than all other systems except for OCaml. We imagined this would only be the case when the tree was not shared. Then we devised the two variants `rbmap_10` and `rbmap_1` which save the current tree in a list after every tenth or every insertion, respectively. The idea is to simulate the behavior of a backtracking search where we store a copy of the state before each case-split. As expected, Lean's performance decreases on these two variants since the tree is now a shared value, and the

	Lean 4			GHC 8.8.3			ocamlopt 4.10			MLton 20180207			MLKit 4.4.2			Swift 5.1.1		
	Time	Del	CM	Time	GC	CM	Time	GC	CM	Time	GC	CM	Time	GC	CM	Time	GC	CM
binarytrees	1.00	38%	8	3.38	74%	13	1.31	—	17	—	—	—	—	—	—	5.35	59%	10
deriv	1.00	22%	12	2.23	43%	6	1.51	78%	6	0.98	21%	18	4.22	59%	20	3.73	42%	6
const_fold	1.00	13%	18	2.73	59%	7	5.11	90%	5	1.15	29%	30	4.59	64%	14	6.30	49%	11
qsort	1.00	10%	0	1.76	1%	0	1.37	34%	0	0.54	0%	0	3.17	0%	0	0.64	0%	0
rbmap	1.00	3%	3	2.44	36%	5	1.03	32%	5	3.25	30%	32	6.49	60%	10	8.08	59%	1
rbmap_10	1.49	13%	11	3.88	58%	7	1.91	59%	9	3.48	29%	32	9.34	72%	16	8.67	55%	3
rbmap_1	4.72	28%	22	14.66	87%	8	9.20	89%	13	4.43	37%	30	15.50	84%	32	12.76	48%	14

Figure 7: Cross-language benchmarks. The measurements include wall clock time (normalized by the Lean base run time), GC time (in percent, as reported by the respective compiler), and last-level cache misses (CM, in million per second, as reported by `perf stat`). For Swift, we measure time spent in inc, dec, and deallocation runtime functions as GC time using `perf`. For Lean, the former are always inlined, so we can only measure object deletion time.

time spent deallocating objects increases substantially. However, Lean still outperforms all systems but MLton on `rbmap_1`. In all other systems but MLton and Swift, the time spent on GC increases considerably. Finally, we point out that MLton spends significantly less time on GC than the other languages using a tracing GC in general.

9 RELATED WORK

The idea of representing RC operations as explicit instructions so as to optimize them via static analysis is described as early as Barth [1977]. Schulte [1994] describes a system with many features similar to ours. In general, Schulte's language is much simpler than ours, with a single list type as the only non-primitive type, and no higher-order functions. He does not give a formal dynamic semantics for his system. He gives an algorithm for inserting RC instructions that, like ours, has an on-the-fly optimization for omitting **`inc`** instructions if a variable is already dead and would immediately be decremented afterwards. Schulte briefly discusses how RC operations can be minimized by treating some parameters as "nondestructive" in the sense of our borrowed references. In contrast to our inference of borrow annotations, Schulte proposes to create one copy of a function for each possible destructive/nondestructive combination of parameters (i.e. exponential in the number of (non-primitive) parameters) and to select an appropriate version for each call site of the function. Our approach never duplicates code.

Introducing destructive updates into pure programs has traditionally focused on primitive operations like array updates [Hudak and Bloss 1985], particularly in the functional array languages Sisal [McGraw et al. 1983] and SaC [Scholz 1994]. Grelck and Trojahner [2004] propose an `alloc_or_reuse` instruction for SaC that can select one of multiple array candidates for reuse, but do not describe heuristics for when to use the instruction. Férey and Shankar [2016] describe how functional update operations explicit in the source language can be turned into destructive updates using the reference counter. In contrast, Schulte [1994] presents a "reusage" optimization that has an effect similar to the one obtained with our **`reset`**/**`reuse`** instructions. In particular, it is independent of a specific surface-level update syntax. However, his optimization (transformation $T14$) is more restrictive and is only applicable to a branch of a **`case`** x if x is dead at the beginning of the branch. His optimization cannot handle the simple *swap* described earlier, let alone more complex functions such as the red black tree rebalancing function $balance_1$.

While not a purely functional language, the Swift programming language[9] has directly influenced many parts of our work. To the best of our knowledge, Swift was the first non-research language to use an intermediate representation with explicit RC instructions, as well as the idea of (safely) avoiding RC operations via "borrowed" parameters (which are called "+0" or "guaranteed" in Swift), in its implementation. While Swift's primitives may also elide copies when given a unique reference, no speculative destructive updates are introduced for user-defined types, but this may not be as important for an impure language as it is for Lean. Parameters default to borrowed in Swift, but the compiler may locally change the calling convention inside individual modules.

Baker [1994] describes optimizing reference counting by use of *two* pointer kinds, a standard one and a *deferred increment* pointer kind. The latter kind can be copied freely without adding RC operations, but must be converted into the standard kind by incrementing it before storing it in an object or returning it. The two kinds are distinguished at runtime by pointer tagging. Our borrowed references can be viewed as a static refinement of this idea. Baker then describes an extended version of deferred-increment he calls *anchored* pointers that store the stack level (i.e. the lifetime) of the standard pointer they have been created from. Anchored pointers do not have to be converted to the standard kind if returned from a stack frame above this level. In order to statically approximate this extended system, we would need to extend our type system with support for some kind of *lifetime annotations* on return types as featured in Cyclone [Jim et al. 2002] and Rust [Matsakis and Klock 2014].

Ungar et al. [2017] optimize Swift's reference counting scheme by using a single bit to tag objects possibly shared between multiple threads, much like our approach. However, because of mutability, every single store operation must be intercepted to (recursively) tag

[9] https://developer.apple.com/swift/

objects before becoming reachable from an already tagged object. Choi et al. [2018] remove the need for tagging by extending every object header with the ID of the thread T that allocated the value, and two reference counters: a shared one that requires atomic operations, and another one that is only updated by T. Thanks to immutability, we can make use of the simpler scheme without introducing store barriers during normal code generation. Object tagging instead only has to be done in threading primitives.

10 CONCLUSION

We have explored reference counting as a memory management technique in the context of an eager and pure functional programming language. Our preliminary experimental results are encouraging and show our approach is competitive with state-of-the-art compilers for functional languages and often outperform them. Our resurrection hypothesis suggests there are many opportunities for reusing memory and performing destructive updates in functional programs. We have also explored optimizations for reducing the number of reference counting updates, and proposed a simple and efficient technique for implementing thread safe reference counting.

We barely scratched the surface of the design space, and there are many possible optimizations and extensions to explore. We hope our λ_{pure} will be useful in the future as a target representation for other purely functional languages (e.g., Coq, Idris, Agda, and Matita). We believe our approach can be extended to programming languages that support cyclic data structures because it is orthogonal to traditional cycle-handling techniques. Finally, we are working on a formal correctness proof of the compiler described in this paper, using a type system based on intuitionistic linear logic to model owned and borrowed references.

ACKNOWLEDGMENTS

We are very grateful to Thomas Ball, Johannes Bechberger, Christiano Braga, Sebastian Graf, Simon Peyton Jones, Daan Leijen, Tahina Ramananandro, Nikhil Swamy, Max Wagner and the anonymous reviewers for extensive comments, corrections and advice.

REFERENCES

Henry G. Baker. 1994. Minimizing Reference Count Updating with Deferred and Anchored Pointers for Functional Data Structures. *SIGPLAN Not.* 29, 9 (Sept. 1994), 38–43. https://doi.org/10.1145/185009.185016

Jeffrey M. Barth. 1977. Shifting Garbage Collection Overhead to Compile Time. *Commun. ACM* 20, 7 (July 1977), 513–518. https://doi.org/10.1145/359636.359713

Johannes Bechberger. 2016. Besser Benchmarken. https://pp.ipd.kit.edu/publication.php?id=bechberger16bachelorarbeit https://github.com/parttimenerd/temci.

Hans-J. Boehm. 2004. The Space Cost of Lazy Reference Counting. In *Proceedings of the 31st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '04)*. ACM, New York, NY, USA, 210–219. https://doi.org/10.1145/964001.964019

Joachim Breitner. 2018. A promise checked is a promise kept: Inspection Testing. *arXiv preprint arXiv:1803.07130* (2018).

Jiho Choi, Thomas Shull, and Josep Torrellas. 2018. Biased Reference Counting: Minimizing Atomic Operations in Garbage Collection. In *Proceedings of the 27th International Conference on Parallel Architectures and Compilation Techniques (PACT '18)*. ACM, New York, NY, USA, Article 35, 12 pages. https://doi.org/10.1145/3243176.3243195

George E. Collins. 1960. A Method for Overlapping and Erasure of Lists. *Commun. ACM* 3, 12 (Dec. 1960), 655–657. https://doi.org/10.1145/367487.367501

Thierry Coquand and Gérard Huet. 1988. The Calculus of Constructions. *Inform. and Comput.* 76, 2-3 (1988), 95–120.

Thierry Coquand and Christine Paulin. 1990. Inductively Defined Types. In *COLOG-88 (Tallinn, 1988)*. Lecture Notes in Comput. Sci., Vol. 417. Springer, Berlin, 50–66.

Leonardo de Moura, Soonho Kong, Jeremy Avigad, Floris van Doorn, and Jakob von Raumer. 2015. The Lean Theorem Prover (System Description). In *Automated Deduction - CADE-25 - 25th International Conference on Automated Deduction, 2015, Proceedings*. 378–388.

Gabriel Ebner, Sebastian Ullrich, Jared Roesch, Jeremy Avigad, and Leonardo de Moura. 2017. A Metaprogramming Framework for Formal Verification. *Proc. ACM Program. Lang.* 1, ICFP (Sept. 2017). https://doi.org/10.1145/3110278

Gaspard Férey and Natarajan Shankar. 2016. Code Generation Using a Formal Model of Reference Counting. In *Proceedings of the 8th International Symposium on NASA Formal Methods - Volume 9690 (NFM 2016)*. Springer-Verlag New York, Inc., New York, NY, USA, 150–165. https://doi.org/10.1007/978-3-319-40648-0_12

Cormac Flanagan, Amr Sabry, Bruce F. Duba, and Matthias Felleisen. 1993. The Essence of Compiling with Continuations. In *Proceedings of the ACM SIGPLAN 1993 Conference on Programming Language Design and Implementation (PLDI '93)*. ACM, New York, NY, USA, 237–247. https://doi.org/10.1145/155090.155113

Clemens Grelck and Kai Trojahner. 2004. Implicit memory management for SAC. In *Implementation and Application of Functional Languages, 16th International Workshop, IFL*, Vol. 4. 335–348.

Niels Hallenberg, Martin Elsman, and Mads Tofte. 2002. Combining Region Inference and Garbage Collection. In *ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI'02)*. ACM Press. Berlin, Germany.

Paul Hudak and Adrienne Bloss. 1985. The Aggregate Update Problem in Functional Programming Systems. In *Proceedings of the 12th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '85)*. ACM, New York, NY, USA, 300–314. https://doi.org/10.1145/318593.318660

Trevor Jim, J Gregory Morrisett, Dan Grossman, Michael W Hicks, James Cheney, and Yanling Wang. 2002. Cyclone: A Safe Dialect of C. In *USENIX Annual Technical Conference, General Track*. 275–288.

Richard Jones and Rafael D Lins. 1996. *Garbage Collection: Algorithms for Automatic Dynamic Memory Management*. Wiley.

Nicholas D. Matsakis and Felix S. Klock, II. 2014. The Rust Language. In *Proceedings of the 2014 ACM SIGAda Annual Conference on High Integrity Language Technology (HILT '14)*. ACM, New York, NY, USA, 103–104. https://doi.org/10.1145/2663171.2663188

Luke Maurer, Paul Downen, Zena M. Ariola, and Simon Peyton Jones. 2017. Compiling Without Continuations. In *Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI 2017)*. ACM, New York, NY, USA, 482–494. https://doi.org/10.1145/3062341.3062380

J. Harold McBeth. 1963. Letters to the Editor: On the Reference Counter Method. *Commun. ACM* 6, 9 (Sept. 1963), 575–. https://doi.org/10.1145/367593.367649

James McGraw, Stephen Skedzielewski, Stephen Allan, D Grit, R Oldehoeft, J Glauert, I Dobes, and P Hohensee. 1983. *SISAL: streams and iteration in a single-assignment language. Language reference manual, Version 1*. Technical Report. Lawrence Livermore National Lab., CA (USA).

Boris Schling. 2011. *The Boost C++ Libraries*. XML Press.

Sven-Bodo Scholz. 1994. Single Assignment C - Functional Programming Using Imperative Style. In *In John Glauert (Ed.): Proceedings of the 6th International Workshop on the Implementation of Functional Languages. University of East Anglia*.

Wolfram Schulte. 1994. Deriving Residual Reference Count Garbage Collectors. In *Proceedings of the 6th International Symposium on Programming Language Implementation and Logic Programming (PLILP '94)*. Springer-Verlag, London, UK, 102–116.

David Ungar, David Grove, and Hubertus Franke. 2017. Dynamic Atomicity: Optimizing Swift Memory Management. In *Proceedings of the 13th ACM SIGPLAN International Symposium on on Dynamic Languages (DLS 2017)*. ACM, New York, NY, USA, 15–26. https://doi.org/10.1145/3133841.3133843

Stephen Weeks. 2006. Whole-program Compilation in MLton. In *Proceedings of the 2006 Workshop on ML (ML '06)*. ACM, New York, NY, USA, 1–1. https://doi.org/10.1145/1159876.1159877

Paul R. Wilson. 1992. Uniprocessor garbage collection techniques. In *Memory Management*, Yves Bekkers and Jacques Cohen (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–42.